L'ús de la informació simbólica en l'automatització del tractament estadístic de dominis poc estructurats

Memòria presentada per Karina Gibert Oliveras per a
l'obtenció del títol de Doctor en Informàtica

Director: Dr. Ulises Cortés García

Programa de doctorat: Aplicacions tècniques i informàtiques de
l'Estadística, la Investigació Operativa i l'Optimització
Universitat Politècnica de Catalunya

30 de Novembre de 1994

Resum

Per les seves característiques intrínseques (coexistència de variables quantitatives i qualitatives — amb gran nombre de modalitats —, coneixement addicional sobre l'estructura del domini per part d'un expert), els *dominis poc estructurats* constitueixen problemes difícils de tractar amb les tècniques actuals en Estadística i Intel·ligència Artificial. A grans trets, l'estructura complexa d'aquests dominis fa que la construcció d'una base de coneixement complet sobre el domini, per a un sistema de diagnòstic, esdevingui pràcticament inabordable i que el clustering (basat en distàncies, que se situen en un pla sintàctic) tingui un comportament *pobre*; d'altra banda, la gestió simultània de variables numèriques i quantitats ingents d'informació qualitativa no respon exactament a les situacions per a les que les tècniques de clustering pures han estat pensades.

KLASS és una eina de classificació parametritzable (quant als criteris d'agregació i les mètriques) capaç d'emprar informació semàntica per a dirigir el procés de classificació. Una de les seves característiques més destacables és l'ús conjunt d'informació qualitativa i quantitativa en la descripció dels objectes.

L'objectiu principal d'aquest treball ha estat superar les limitacions de les tècniques de l'Estadística i de la Intel·ligència Artificial en la classificació de dominis poc estructurats. Entre d'altres resultats cal destacar

- Una estratègia mixta de classificació que:
  - incorpora al procés de classificació coneixement *parcial* (ja no serà necessari que sigui complet) i/o no homogeni que tingui l'expert sobre el domini. Aquest coneixement es formalitza en CP_1 (càlcul de predicats de primer ordre), permetent màxima potència expressiva i flexibilitat, i s'utilitza per construir una *partició* inicial sobre el domini *induïda per les regles* — que equival a empaquetar "*troços*" de coneixement en unitats funcionals, introduint informació *semàntica* al sistema — a partir de la qual procedir amb una classificació per veïns recíprocs encadenats (de complexitat quadràtica). Aquesta seria la contribució principal de la present tesi i, entre altres coses, ofereix l'oportunitat de fer intervenir, en les regles, variables derivades de les observades. En aquest punt, el mecanisme és especialment potent i dóna la possibilitat d'estudiar les dades simultàniament en diferents sistemes de coordenades.
  - permet treballar simultàniament amb variables qualitatives i quantitatives, tot mantenint la representació simbòlica d'aquestes darreres i evitant la categorització de les primeres. Això ha suscitat la definició d'un representat de classe per a les variables qualitatives, que constitueix una segona aportació d'aquest treball, i permet un tractament homogeni de:

  - ∗ variables quantitatives i qualitatives
  - ∗ classes i individus (en conseqüència, fa possible la integració de les classes induïdes per les regles en una única jerarquia final).

  Treballar amb aquest tipus de matrius requereix detenir-se en la definició d'una funció que avaluï la distància entre individus. S'ha definit la *família de distàncies mixtes* $d^2_{(\alpha_r,\beta_r)}(i,i')$, estudiant-ne algunes propietats teòriques i aportant una proposta pels valors dels paràmetres (α_r,β_r). No es té notícia que una formulació d'aquest estil hagi estat abans utilitzada.

- i una metodologia de treball iterativa que, partint de les dades observacionals i el coneixement de l'expert, integra el procés anterior amb eines d'ajuda a la interpretació de les classes, de tal forma que s'obtingui finalment una classificació "*satisfactòria*", d'acord amb els objectius de l'expert. Aquesta metodologia resol la dificultat que per ell suposa formalitzar el seu coneixement de forma completa i precisa en dominis poc estructurats, ocasionada per l'existència de regles que aplica inconscientment i que, havent adquirit amb l'experiència, inicialment no és capaç de formalitzar.

En el camp d'ajudes a la interpretació de les classes (actualment poc automatitzada en Estadística) es fan bàsicament dues contribucions

- s'aporta una forma ràpida de mesurar $\delta(\mathcal{P}_1,\mathcal{P}_2)$ la diferència entre dues classificacions, orientada a fer comparacions i, eventualment, avaluar la qualitat d'una classificació. El càlcul d'aquest coeficient es basa en la construcció de la *taula de contingència de la partició producte* i

- un sistema de caracterització de la classificació que detecta les variables *identificadores* d'algunes classes (basant-se en un procés de condicionament progressiu). Aquesta tècnica descansa també sobre el concepte de *representant de classe* definit anteriorment i, a més de dotar al classificador de certa capacitat explicativa, dóna lloc a un procediment de generació automàtica de regles, que podran formar part de la base de coneixement d'un sistema de diagnòstic basat en el coneixement per dominis més complexos que els tractables actualment, la qual cosa suposa una contribució a l'apertura d'un canal d'apropament de l'Estadística i la Intel·ligència Artificial.

Abstract

Because of their intrinsic characteristics (coexistence of quantitative and qualitative variables — the last ones with great number of modalities —, additional expert knowledge on the domain structure), *ill-structured* domains are dificult problems for the actual statistical and artificial intelligence techniques. Briefly, construction of complete knowledge bases of the domain, to be used in diagnostic oriented systems, is almost unreachable due to the complexity of these kind of domains. The clustering (based on distances, which are, in fact, sintactic criteria) has also a *poor* behaviour; actually, standard statistical techniques where not especifically designed for simultaneous treatment of numerical variables and great quantity of qualitative information.

KLASS is a clustering tool that can use semantic information to guide the classification process. It is parameterized on the aggregation criteria and the metrics. One of its most important features is the use of both qualitative and quantitative information in the object descriptions.

The main goal of this work is to overcome the limitations of Statistics and Artificial Intellegence techniques referred to this particular context. Among other results, one may highliht:

- A mixt classification strategy that:
  - incorporates *parcial* (complete knowledge is not necessary anymore) and/or non-homogeneous knowledge that the expert has on the domain. This knowledge is formalized in CP_1, allowing maximum expressivity and flexibility, and it is used to built an initial *induced partition* on the domain — this is equivalent to pack knowledge *pieces* in functional units, introducing *semantics* into the system — from wich a chained reciprocal neighbors classification (of quadratic cost) is performed. This is the main contribution of the thesis and, among others, it offers the oportunity of using, in the rules, variables defined as transformations of the observed ones. In this point, the mecanism is especially powerfull, providing the possiblity to study the data simultaneously in different coordinate systems.
  - manages data matrices composed of both quantitative and qualitative variables, mantaining the symbolic representation of the last ones and evoying the codification of the firsts. This suggests the definition of a class representative for qualitative variables. It constitutes a second contribution of this work, allowing an homogeneous treatement of
    - quantitative and qualitative variables
    - classes and individuals (as a consequence, enables the integration of the classes induced by the rules in an only final hierarchy).

- and an iterative working methodology that, starting with observational data and expert knowledge, integrates the previous process with some tools — oriented to the interpretation of the classes— in such a way that, at the end, a *satisfactory* classification can be obtained, according to the expert goals. This methodology solves the expert difficulties in formalizing its knowledge in a complete and precise way for ill-structured domains. Those difficulties are due to the existence of rules, acquired by experience, that the expert applies inconsciously and, inicially, he was not able to formalize.

Referring to the class interpretation oriented tools (nowadays, at low level of automation in Statistics) two basic contributions are made:

- a quick way of measuring $\delta(\mathcal{P}_1, \mathcal{P}_2)$ the difference between two classifications is provided. This measure is oriented to comparisons and, occasionally, to an evaluation of a classification quality. The coeficient is calculated using the *product partition contingency table* and

- a class characterization system to detect *identifier* variables for some classes (based on a progressive conditioning process). This technique is also founded on the class representative defined before. It gives some explicative capacity to the classifier and it also constitutes a starting point for an automatic rules generation process. The rules generated by this process could be introduced in the knowledge base of a giagnostic oriented system for a domain of greater complexity than those that are manageables at present. This suposes a contribution on the Statistics and Artificial Intelligence approach.

Resumen

Por sus características intrínsecas (coexistencia de variables cuantitativas y cualitativas — de gran número de modalidades —, conocimiento adicional del experto acerca de la estructura del dominio), los *dominios poco estructurados* constituyen problemas difíciles de tratar con las técnicas actuales en Estadística e Inteligencia Artificial. A grandes rasgos, la estructura compleja de estos dominios hace que la construcción de una base de conocimiento completo sobre ellos, para un sistema de diagnóstico, se vuelva prácticamente inabordable y que el clustering (basado en distancias, que se sitúan en el plano sintáctico) tenga un comportamiento *pobre*; por otro lado, la gestión simultánea de variables numéricas y cantidades ingentes de información cualitativa no responde exactamente a las situaciones para las que las técnicas de clustering puras fueron pensadas.

KLASS es una herramienta de clasificación parametrizable (en cuanto a los criterios de agregación y las métricas) capaz de emplear información semántica para dirigir el proceso de classificación. Una de sus características más destacables es el uso conjunto de información cuantitativa y cualitativa en la descripción de los objetos.

El objetivo principal de este trabajo ha sido superar las limitaciones de las técnicas de la Estadística y de la Inteligencia Artificial para la clasificación de dominios poco estructurados. Entre otros resultados, cabe destacar:

- Una estrategia mixta de clasificación que:
  - incorpora al proceso de clasificación conocimiento *parcial* (ya no será necesario que sea completo) y/o no homogéneo que tenga el experto sobre el dominio. Este conocimiento se formaliza en CP_1, permitiendo máxima potencia expresiva y flexibilidad, y se utiliza para construir una *partición* inicial *inducida por las reglas* — equivalente a empaquetar *trozos* de conocimiento en unidades funcionales y que supone introducir información *semántica* al sistema— a partir de la cual proceder a una clasificación por vecinos recíprocos encadenados (de complejidad cuadrática). Esta sería la contribución principal de la presente tesis y, entre otras cosas, ofrece la oportunidad de hacer intervenir, en las reglas, variables derivadas de las observadas. En este punto, el mecanismo es especialmente potente y da la posibilidad de estudiar los datos bajo distintos sistemas de coordenadas simultáneamente.
  - permite trabajar conjuntamente con variables cualitativas y cuantitativas, manteniendo la representación simbólica de estas últimas y evitando la categorización de las primeras. Ello ha suscitado la definición de un representante de clase para las variables cualitativas, que constituye una segunda aportación de este trabajo y permite un tratamiento homogéneo de:

  - variables cualitativas y cuantitativas.
  - clases e individuos (en consecuencia, hace posible la integración de las clases inducidas por las reglas en una jerarquía final).

  Trabajar con este tipo de matrices requiere detener-se en la definición de una función que evalúe la distancia entre individuos. Se ha definido la *familia de distancias mixtas* $d^2_{(\alpha_r,\beta_r)}(i,i')$, estudiando algunas de sus propiedades teóricas y aportando una propuesta para los valores de los parámetros (α_r,β_r). No se tiene noticia de que una formulación de este estilo se haya utilizado con anterioridad.

- y una metodología de trabajo iterativa que, partiendo de datos observacionales y del conocimiento del experto, integra el proceso anterior con herramientas de ayuda a la interpretación de las clases, de tal forma que se obtenga finalmente una clasificación *satisfactoria*, de acuerdo a los objetivos del experto. Esta metodología resuelve la dificultad que para él supone formalizar su conocimiento de forma completa y precisa en dominios poco estructurados, debida a la existencia de reglas que aplica inconscientemente y que, habiendo adquirido con la experiencia, inicialmente no fue capaz de formalizar.

En el campo de ayudas a la interpretación de las clases (actualmente poco automatizada en Estadística) se hacen básicamente dos contribuciones

- se aporta una forma rápida de medir $\delta(\mathcal{P}_1,\mathcal{P}_2)$ la diferencia entre dos clasificaciones, orientada a hacer comparaciones y, eventualmente, evaluar la calidad de una clasificación. El cálculo de este coeficiente se basa en la construcción de la *tabla de contingencia de la partición producto* y

- un sistema de caracterización de la clasificación que detecta las variables *identificadoras* de algunas clases (basándose en un proceso de condicionamiento progresivo). Esta técnica descansa también sobre el concepto de *representante de clase* definido anteriormente y, además de dotar al clasificador de cierta capacidad explicativa, da lugar a un procedimiento de generación automática de reglas, que podrán formar parte de la base de conocimiento de un sistema de diagnóstico basado en el conocimiento para dominios más complejos que los manejables actualmente, lo cual supone una contribución a la apertura de un canal de acercamiento entre Estadística e Inteligencia Artificial.

Contents

1 INTRODUCCIÓ **I**

1.1 Panoràmica històrica i estat actual II

1.2 Presentació de Klass i motivació VII

1.3 Estructura del treball VIII

2 Caracterització del problema **1**

2.1 Els dominis poc estructurats 1

2.2 Limitacions del clustering i de l'aprenentatge 2

2.2.1 L'Estadística 2

2.2.2 La Intel·ligència Artificial 3

2.3 Objectius 4

2.4 Representació de les dades 5

2.4.1 Les variables 5

2.4.2 La matriu de dades 6

2.5 Valors mancants 6

2.6 Resum del capítol 9

3 Algorisme bàsic **11**

3.1 La classificació ascendent jeràrquica 11

3.2 Sobre l'índex de nivell de la jerarquia 15

3.3 La representació de les classes 17

3.3.1 El centre de gravetat d'una variable 17

3.3.2 Objectes compactes i objectes extesos 20

3.3.3 Càlcul del centre de gravetat 21

3.4 Les mètriques 23

3.4.1 Només variables quantitatives 23

3.4.2 Només variables qualitatives 24

3.4.3 Variables de tipus heterogenis 28

3.4.4 La distància mixta 29

3.4.5 Equivalències entre distàncies mixtes 33

3.5 Sobre els valors d'α i β . . . 35
3.6 Criteris d'agregació . . . 37
3.6.1 El criteri del centroide . . . 37
3.6.2 El criteri de Ward . . . 38
3.7 Resum del capítol . . . 49

4 Introducció d'informació declarativa 51
4.1 Representació del coneixement . . . 52
4.2 Classificació amb regles . . . 55
4.2.1 Preliminars . . . 55
4.2.2 La partició induïda per les regles . . . 56
4.2.3 Incorporació de les regles al procés general . . . 57
4.3 Qüestions tècniques . . . 58
4.4 De les inversions . . . 60
4.4.1 Gènesi de les inversions . . . 61
4.4.2 Tractament de les inversions . . . 62
4.5 Proposta d'una metodologia de treball . . . 63
4.6 Resum del capítol . . . 67

5 Ajudes a la interpretació 69
5.1 Caracterització de les classes . . . 69
5.1.1 Variables descriptores . . . 70
5.1.2 Variables caracteritzadores . . . 71
5.1.3 Apunt a la generació automàtica de regles . . . 76
5.2 Comparació entre classificacions . . . 78
5.2.1 La taula de contingència de la partició producte . . . 79
5.2.2 Posicions centrals i posicions de perifèria . . . 81
5.2.3 Una mesura de la diferència entre classificacions . . . 82
5.3 Determinació del tall de l'arbre . . . 83
5.3.1 Eines gràfiques . . . 83
5.3.2 Eines numèriques . . . 84
5.4 Resum del capítol . . . 87

6 Aplicacions (I). Esponges 89
6.1 Introducció . . . 89
6.2 Les dades . . . 91
6.3 Inici de l'estudi . . . 92
6.3.1 Descripció de les classes . . . 92
6.3.2 Estudi comparatiu . . . 94

6.4 Ús d'informació declarativa 97
6.5 Una segona aproximació 100
6.6 Conclusió 101

7 Aplicacions (II). Poblacions estel·lars 103
7.1 Introducció 103
7.2 Marc de treball 104
7.3 Estudi 106
7.3.1 Passa $\xi = 0$ 106
7.3.2 Passa $\xi = 1$ 106
7.3.3 Passa $\xi = 2$ 108
7.3.4 Passa $\xi = 3$ 109
7.3.5 Passa $\xi = 4$ 109
7.3.6 Passa $\xi = 5$ 111
7.4 Conclusions 111

8 Comparació amb altres sistemes i mètodes 113
8.1 Microodinadors 113
8.2 Simulació de cubs 118
8.2.1 Les dades 118
8.2.2 Primera prova 118
8.2.3 Prova segona 120
8.2.4 Prova tercera 122
8.3 Conclusions 123

9 Disseny i implementació 125
9.1 Disseny modular 125
9.1.1 Estructura de fitxers 127
9.1.2 Mòdul de lectura de dades 129
9.1.3 Mòdul de classificació 131
9.1.4 Mòdul d'ajudes a la interpretació 134
9.1.5 Documentació de sortida 137
9.1.6 Mòdul de classificació amb regles 140
9.2 Estudi de la complexitat de l'algorisme 142
9.2.1 Funció veins_rec 143
9.2.2 Funció particio 145
9.2.3 Funció integració 146

10 Conclusions, línies obertes i treball futur 147

Bibliografia **152**

A Resultats addicionals sobre les esponges **161**
A.1 Objectes 161
A.2 Detall d'algunes classificacions 161
A.2.1 Classificació tradicional 161
A.2.2 Resultat de Linneo 161
A.2.3 Classificació proposada per Domingo (1990) 163
A.3 Llista de propietats 163
A.4 Classificació general amb Klass 168
A.5 Variables descriptores (7 classes) 169
A.6 Classificació amb regles 171
A.6.1 El conjunt de regles 171
A.6.2 Partició induïda per les regles 171
A.6.3 Informes 172
A.7 Segona aproximació 177

B Resultats addicionals sobre les estrelles **179**
B.1 Primeres classificacions 179
B.2 Classificació definitiva (Passa $\xi = 4$) 187

C Resultats addicionals d'altres aplicacions **193**
C.1 Microordinadors 193
C.2 Simulació de cubs 194
C.2.1 Prova primera 194
C.2.2 Proves segona i tercera 197

Chapter 1

INTRODUCCIÓ

"La diferència és el principi general de la multitud".
R.LLULL

La classificació és la tècnica d'ús més comú per a separar dades en grups. Moltes vegades, qui fa recerca s'enfronta a grans quantitats d'informació que serien pràcticament intractables a menys que es podessin agrupar en classes que, d'alguna forma, fossin susceptibles de ser considerades com a unitats. En altres ocasions, simplement interessa esbrinar l'estructura interna de les dades per a utilitzar-la com a base de tasques posteriors, bé sigui amb fins predictius[1], bé d'altres naturaleses. Per exemple, classificant un conjunt de malalties en funció dels símptomes que les identifiquen i havent assignat a cada grup un tractament, el diagnòstic d'un malalt pot reduir-se a veure a quina classe pertany aquest. [**?**] fa una relació de les múltiples utilitats que les tècniques de classificació (**clustering**) ofereixen.

En qualsevol cas, l'objectiu d'aquestes tècniques és obtenir una representació esquemàtica i simple d'una matriu rectangular de dades, que en conservi el màxim d'informació. De fet, es pretén descobrir les classes *naturals* del domini en estudi, que abans de començar l'anàlisi són desconegudes; es vol descriure el domini a través d'unes poques classes homogènies, el màxim de diferents entre si, i representatives del total. En aquest sentit, la classificació és una tècnica exploratòria per cercar l'estructura de les dades.

Donada una mostra per classificar, la primera intenció que hom té és d'estudiar totes les classificacions possibles i triar la millor, en base a algun criteri que mesuri com s'ajusta la jerarquia obtinguda a les dades. De fet, donat un grup d'individus, el conjunt de totes les particions que es puguin construir amb aquests individus defineix un espai de cerca que conté la classificació òptima segons un cert objectiu. Però la cerca de l'òptim en aquest espai es submergeix en el món de la complexitat algorísmica, on, naturalment, naufraga, ja que el nombre de particions possibles esdevé intractable fins i tot quan el nombre d'individus a classificar encara no és gaire gran (consulti's [**?**] per un tractament detallat d'aquestes qüestions de caire combinatori).

[1]Seria el cas, per exemple, d'utilitzar els resultats de la classificació per a la generació de regles per un sistema basat en el coneixement.

No podent fer un estudi exhaustiu, es recorre als criteris heurístics, que permeten reduir aquest espai de cerca. Disminuir el nombre de jerarquies a estudiar, porta molt més ràpidament a una solució, que, en contrapartida, pot no ser l'òptima, encara que sí es pot assegurar que sigui *raonable*, és a dir, suficientment propera a l'òptim real com per que no valgui la pena invertir més esforç computacional en calcular-lo. A §**??** s'ofereix una classificació general dels tipus d'heurístics existents.

Així doncs, la base matemàtica dels mètodes de classificació està encara poc desenvolupada i, donat un conjunt qualsevol de dades, no permet elegir un algorisme amb avantatges indiscutibles sobre la resta. Donada aquesta falta de fonament teòric, la classificació automàtica requereix una bona dosi d'empirisme pel que fa a l'elecció d'una mesura de similitud (o distància) sobre l'espai de les variables, així com d'una estratègia de classificació; eleccions de les que depenen en bona part els resultats. No es pot dir, per tant, que existeixi *la* classificació d'un conjunt de dades, sinó *una* classificació, la qualitat de la qual s'hauria de determinar fent una validació *a posteriori* (veure §**??**) dels resultats. En general, la validesa d'una classificació dependrà de la seva utilitat. [**?**] fa tota una dissertació il·lustrant que el concepte de validesa no és absolut, sino relatiu a les condicions de contexte, i presenta exemples de com, en economia, sociologia o comptabilitat, variacions de les condicions de l'entorn, posen en vigència diferents classificacions de les mateixes *coses*. Sense anar més lluny, la classificació de la Societat en *capes socials* ha variat notablement al llarg de la història, en funció de les diferents situacions polítco-ideològiques.

1.1 Panoràmica històrica i estat actual

Classificar gent en tipus és un costum que s'ha vingut practicant des fa molt de temps. Ja els Hindús usaven el sexe, les característiques físiques i el comportament per a classificar la gent en sis tipus que designaven amb noms d'animals. Els primers físics de la Grècia i Roma clàssica desenvoluparen diverses tipologies basant-se en les variacions que diferents combinacions dels quatre humors bàsics de l'organisme humà produïen en les característiques físiques de l'individu. D'entre elles, la més famosa és la de Galen (a.C. 129–199), que definia nou tipus temperamentals relacionats amb la susceptibilitat de les persones, diferents malalties i diferències individuals en el comportament.

Els primers treballs de classificació es desenvolupaven, en la seva majoria, en els camps de la Biologia i la Zoologia, on es sol designar amb el terme *Taxonomia* als arbres de classificació.

Al segle XVIII, un dels més reconeguts classificadors de la història, Linneus, efectua una classificació dels éssers vius encara vigent avui dia en el món de les ciències naturals, per bé que no en la seva compacitat original [**?**]. En la seva obra fonamental, *Genera plantarum*, (1^a*.ed.*, 1737) hi diu:

Tot el coneixement real que tenim depèn dels mecanismes pels quals l'home

> *distingeix la semblança de la diferència. Quan més gran sigui el nombre de distincions que aquest mecanisme sigui capaç de captar, més clara és la nostra idea de les coses. Quant més gran sigui el nombre d'objectes que tractem de reconèixer, més difícil serà establir aquests mecanismes, i també més necessari.*

En principi, la construcció d'una Taxonomia era més aviat un procés artístic, amb una forta intervenció del *sentit comú* del seu autor, i quedava un xic lluny del que habitualment s'entén per treball científic. És a rel de les taxonomies numèriques, basades en les idees d'Andanson (s. XVIII), que comencen a desenvolupar-se tècniques més objectives. Durant la primera meitat del nostre segle, hi ha hagut nombroses temptatives de racionalitzar el procés mental que Linneus va emprar en el seu treball. Entre altres, Zubin o Thorndike [**?**] apliquen tècniques de classificació numèrica a camps que no són les ciències naturals.

Però no és fins que la Informàtica pren cos a l'entorn universitari que apareixen els primers algorismes automatitzant aquest procés. Efectivament, la consolidació d'aquesta ciència ofereix medis per a la còmoda manipulació de les grans quantitats d'informació que es maneguen en el món de l'anàlisi de dades, i permet efectuar de forma ràpida els nombrosos càlculs que requereixen els processos de classificació.

La primera formulació moderna de la taxonomia numèrica és probablement la de [**?**], en la que s'hi exposa

> *(...) l'estudi teòric de la classificació, incloent les seves bases, principis, procediments i regles*

obra que estimulà gran part de la recerca en classificació automàtica dins el camp de la biologia.

Intents de formular models matemàtics més precissos i rigorosos que donguin suport teòric a l'anàlisi de classificació són, entre daltres, els treballs de [**?**], [**?**], [**?**], [**?**], [**?**], [**?**].

Per descansar la majoria d'aquests mètodes en l'elecció d'una distància — o mesura de (dis)similitud — sobre l'espai de les variables, gran part de la recerca s'ha centrat en la definició de diferents mesures de proximitat i el desenvolupament d'algorismes eficients per a utilitzar-les. Compilacions d'aquestes mesures es poden trobar a [**?**], [**?**], [**?**], [**?**], [**?**]

Actualment, i considerant la forma com estructuren el conjunt d'individus, es poden agrupar els heurístics existents en els següents blocs:

Mètodes de particions: Cerquen una partició òptima del conjunt que s'estudia en un nombre prefixat de classes[2] k. En tenim de dos tipus:

- **Mètodes de particions directes:** Les classes que es formin seran disjuntes, i poden ser *aglomeratius* o *divisius*.

[2] Cal que l'usuari sàpiga, de bon principi, quantes classes vol fer.

Els primers, parteixen del núvol de punts inicial i van fen fusions progessives, fins a obtenir una única classe. Entre ells, cal destacar el mètode dels *centres mòbils* [**?**], que és un cas particular dels *núvols dinàmics* degut a [**?**], o el *k-means* de [**?**]. Altres, difereixen en l'elecció dels centres de classe inicials. Així, l'*Isodata* de [**?**] involucra set paràmetres que piloten la partició.

Els mètodes divisius o descendents, inicialment escindeixen el núvol total dels individus, generalment de forma dicotòmica, i procedeixen subdividint cada part, fins a obtenir un nombre suficient de classes. Molt menys utilitzats que els aglomeratius, i sovint faltats de precissió ([**?**]), són sobretot usats en màrketing, on es prima la rapidesa en trobar una solució per sobre de la qualitat d'aquesta. El més conegut és el mètode de [**?**]. [**?**] presenta alguns d'aquests mètodes com el d'Edwards & Carvalli-Sforza, o el de Friedman & Rubin, però anota també que sota certes condicions aquests mètodes poden presentar dificultats operatives.

- **Mètodes de particions en classes solapades:** Les classes poden solapar-se, com el seu nom indica. És a dir, un mateix objecte pot pertànyer simultàniament a més d'una classe. Aquest seria el cas de les paraules polisèmiques com `rosa`, que es podria classificar simultàniament dins els noms de flors i també dins els noms dels colors. Aquests mètodes foren introduïts, sembla ser, per Jones i Needhan a la Cambridge Language Research Unit, i també [**?**] fan un intent de formalització en aquesta línia.

Mètodes de classificació jeràrquica: Busquen l'arbre que reflexa l'estructura jeràrquica de les dades. Segons el nivell pel que es talli l'arbre s'obtindrà una partició més o menys precissa del conjunt objecte d'estudi. L'avantatge d'aquest grup respecte de l'anterior és que no cal avançar el nombre de classes que es volen obtenir al final, sinó que un cop construïda la taxonomia es pot decidir a quin nivell tallar-la horitzontalment, obtenint així una partició del refinament més adient als objectius de l'estudi. És l'usuari qui decideix quantes classes fer, però el mètode li proporciona informació de com són les dades, que l'ajudarà a prendre aquesta determinació.

Es costrueixen les jerarquies fent agregacions successives dels objectes — i després dels grups formats—, en funció de les distàncies entre objectes — i eventualment grups—. Són mètodes de classificació ascendent jeràrquica el del *salt mínim*, degut a Florek, 1951, o el d'*agregació segons la mínima pèrdua d'inèrcia*, [**?**], entre d'altres. [**?**] conté un recull força extens dels diferents mètodes ascendents jeràrquics.

Mètodes de classificació piramidal: Introduïts per Diday, generalitzen el concepte de jerarquia, i permeten que un node pugui tenir dos pares simultàniament. Així, en una piràmide, els *coralls* poden pertànyer alhora al regne *animal* i *vegetal*, atenent a les

seves característiques fisiològiques i morfològiques simultàniament. En aquesta línia, pot consultar-se [?]. A Catalunya hi ha actualment un grup treballant en aquest tema [?].

Mètodes d'arbres additius: Les distàncies entre individus es calculen per addició de les longituds de les arestes que els uneixen. El núvol de punts queda estructurat en uns diagrames en forma d'estrella que poden dificultar la interpretació [?].

Mètodes de classes latents: En aquests mètodes, la variable que indica la classe de cada objecte es considera una variable latent. El mètode consisteix a estimar la probabilitat amb que cada element pot pertànyer a cada classe [?]

A [?], [?], [?] apareixen classificacions alternatives atenent a altres criteris com ara la forma com es va construint l'agregacio i altres.

En un altre ordre de coses es podria citar la *classificació de dades textuals* estudiada per [?], útil en el tractament de preguntes obertes en enquestes, o en la classificació de textos literaris, que s'enfronta al problema de la numerització del *corpus textual* i, en conseqüència, al tractament de matrius esparses.

Sota l'aproximació clàssica, és l'expert el responsable de decidir quines característiques dels individus jugaran en la classificació i indirectament, per tant, les que intervindrien en una possible formulació de regles basada en el resultat de la classificació. La limitació que aquest enfoc presenta, ja assenyalada per alguns autors (ex: [?], [?], [?],...), ha estimulat el desenvolupament de sistemes que poden decidir quan un atribut és potencialment útil o no, alliberant d'aquesta responsabilitat a l'usuari. Un exemple és **COBWEB** [?], sistema que efectua una *classificació conceptual* (veure §**??**) dels individus, basant-se en una mesura d'utilitat dels atributs que deriva de la mesura d'informació clàssica. Aquesta selecció d'atributs rellevants s'ha fet [?] utilitzant mètodes d'inferència inductiva automàtica, que posen de manifest les estructures o relacions existents entre les dades i poden intervenir en la posterior formulació de les regles de classificació. Aquesta aproximació també s'utilitza en el reconeixement de formes (**pattern recognition**), combinant-la, essencialment, amb mètodes estadístics [?].

En l'àrea de l'aprenentatge automàtic (**machine learning**), s'han desenvolupat diversos mètodes *heurístics* d'avaluació d'atributs en termes de la seva utilitat potencial (identificada amb la seva rellevància) que permeten decidir a q¡uin grup pertany cada un, basant-se, en la seva majoria, en la teoria de la informació clàssica (ex: [?], [?], [?], [?], [?], [?] i [?]).

Al voltant dels 80, Michalski s'interessa per formalitzar la classificació des d'un punt de vista totalment diferent. La idea de mimetitzar el procés de classificació realitzat per un humà, de naturalesa inherentment qualitativa, i l'interés per introduir-hi consideracions sobre la rellevància dels atributs — tal com fem els humans —, el porten a presentar a [?] la *classificació conceptual*.

Estem entrant en aquest moment en una altra àrea de la ciència que també fa a la classificació objecte del seu estudi: la Intel·ligència Artificial, i en concret, ens situem en la branca de l'aprenentatge automàtic no supervisat.

L'evolució de l'aprenentatge automàtic dins la Intel·ligència Artificial s'ha articulat en diferents biaixos metodològics. Entre ells, la necessitat en la majoria d'aplicacions d'Intel·ligència Artificial de processar abundant informació simbòlica (qualitativa), l'èmfasi que tradicionalment fa aquesta àrea de la ciència de les representacions basades en la lògica, l'interés de treballar sota una aproximació que imiti els processos de raonament humà, la necessitat de *comunicació* amb l'usuari humà (per exemple, en els sistemes experts) i, per consegüent, l'intent d'automatitzar, almenys parcialment, la interpretació (formant només aquelles classes a les que es pot associar un concepte amb sentit), motiva el desenvolupament de classificadors basats en regles que són comprensibles i inspeccionables per un element utilitzador (i possiblement humà).

En aprenentatge automàtic, els mètodes de classificació automàtica poden constituir la base d'un planificador que seleccioni accions *apropiades* per a desenvolupar una certa tasca, o d'un generador automàtic de regles orientat a la construcció d'entorns basats en el coneixement, o d'un sistema que faci diagnòstics consistents amb el comportament prèviament estudiat de l'expert.

De fet, diversos sistemes experts famosos, com **MYCIN** [**?**], **INTERNIST**, **PLANT/ds** [**?**], no són més que classificadors: Donada una certa entrada, es determina a quina classe pertany aquesta, segons un cert subconjunt de les seves característiques, i les regles corresponents, implementades en forma d'arbre de decisió, i es retorna la resposta associada a aquella classe.

Com ja s'ha dit, doncs, [**?**] inicien la recerca en aquesta línia amb la *classificació conceptual*. A diferència del que fan els mètodes estadístics, la formació de classes vé guiada per criteris aplicats a la descripció conceptual de les dades, no a les dades pròpiament dites.

Actualment es fa molta recerca en classificació sota diferents aproximacions. En contrast amb la representació conjuntiva de les classes pròpia de la *classificació conceptual*, [**?**] i altres autors proposen mètodes en que la formació de les classes es basa en les probabilitats *a priori* i *a posteriori* dels objectes: és la *classificació baiessiana*, que associa a cada objecte graus de pertinença a les diferents classes, i que cal no confondre amb la *classificació difusa*, que aplica la teoria de conjunts difusos, proposada per Zadeh al 1965, al càlcul d'aquests graus de pertinença. En aquest contexte, cal destacar els treballs de *particions difuses* de Ruspini, o la introducció de valors difusos a l'*Isodata* de [**?**] deguda a [**?**].

En relació a la classificació difusa, [**?**] estudia els operadors d'agregació, i a [**?**] o [**?**] es descriu una tècnica per a definir i operar amb els valors difusos que poden prendre les variables presents en la descripció dels individus.

Són també importants els treballs de [**?**] en *formació de conceptes*, basats en la construcció d'un esquema de classificació de forma incremental a partir d'una seqüència d'observacions,

i on la descripció de les classes pot ser conjuntiva (Lebowitz 82) o probabilística [**?**], o fins i tot baiessiana [**?**].

Un altre enfoc és el connexionista. L'ús de xarxes neuronals en la classificació arranca dels treballs del físic finès [**?**]. Si bé les xarxes neuronals són un conegut mètode d'aprenentatge supervisat, els *mapes auto-organitzatius* presentats per Kohonen constitueixen un mètode de classificació dels més recents, utilitzant de diversitat d'aplicacions, incloent l'astronomia [**?**].

Fins i tot els *algorismes genètics*, apareguts al voltant dels 60 (veure [**?**]), s'han aplicat a problemes de classificació. Des del punt de vista que la classificació involucra un problema d'optimització, del que se'n donen detalls a §**??**, es pot abordar sota l'òptica dels genètics ([**?**] conté una descripció extensa d'aquests algorismes).

La classificació ha estat i és, doncs, àmpliament estudiada per diferents disciplines des d'èpoques remotes, i sempre ha respost a un mateix objectiu: trobar les agrupacions pròpies d'un cert núvol d'individus.

Part de la recerca actual [**?**] se situa en una àrea intermèdia combinant Intel·ligència Artificial i Estadística. El present treball, situat en aquest camp, neix de la idea de posar en contacte idees estadístiques i inductives tractant de compaginar, en la mesura del possible, ambdós punts de vista. L'objectiu és oferir una eina prou potent en el tractament d'un tipus de dominis especial, que es descriuen en l'apartat **??**: els dominis poc estructurats.

1.2 Presentació de Klass i motivació

En aquesta secció es presenta breument l'eina informática que constitueix el cos d'aquest treball, i que ha permès posar en pràctica tot el desenvolupament teòric realitzat.

Originalment dissenyat per a dur a terme comparacions entre els mètodes estadístics clàssics i un nou sistema orientat a la classificació difusa basat en el coneixement anomenat **Linneo** [**?**], la seva evolució l'ha convertit en una eina autònoma que permet la classificació de dominis poc estructurats.

La idea inicial fou únicament processar la informació de que s'alimentava **Linneo** utilitzant un mètode de classificació estàndard. **Linneo** implementava un mètode innovador, i la seva posició relativa respecte dels mètodes existents era decisiva per a la seva consolidació . Però aquest sistema presentava una diferència important amb els altres mètodes: creat en el contexte de la intel·ligència artificial tractava quantitats importants d'informació simbòlica. La problemàtica associada a aquesta circumstància, comentada a §**??**, fou la raó per la qual **KLASS** implementa l'*adaptació* d'un mètode clàssic al processament de la informació simbòlica.

Davant del pobre comportament que els mètodes estadístics clàssics presentaven per determinats tipus de dades, sorgí la idea de treballar en una nova direcció: superar les deficiències detectades en aquells mètodes, per tal d'obtenir millores en la classificació

d'aquests dominis *patològics*[3], i es configurà una metodologia específica de *classificació amb constriccions declaratives*.

No obstant, en fer proves reals, es va observar com, en una primera etapa de qualsevol investigació consistent a definir aquestes constriccions, **KLASS** actuava més aviat com una eina d'ajuda a l'adquisició del coneixement (veure §**??**).

Es pot dir, doncs, que **KLASS** cobreix un doble propòsit:

- Per una banda, implementa un mètode de classificació amb constriccions basades en el coneixement.
- Per l'altra, constitueix una eina d'ajuda a l'adquisició del coneixement basada en mètodes estadístics, que obre la porta de la generació automàtica de regles per un sistema de diagnòstic basat en el coneixement.

A §**??** s'hi troben descripcions detallades dels aspectes més tècnics de **KLASS** .

1.3 Estructura del treball

La secció inicial d'aquest treball és un capítol introductori, en el que s'ofereix una ràpida visió del que constitueix la classificació de dades.

A continuació, en un segon capítol es descriu el problema que es vol resoldre, és a dir, el marc de treball i la forma com es representen les dades, les limitacions que els enfocs clàssics dins la intel·ligència artificial o l'estadística suposen en aquest contexte i es descriuen els objectius de la tesi.

Els capítols tercer i quart contenen el desenvolupament teòric que serveix de base per **KLASS**, que és un classificador ascendent jeràrquic per a dades descrites sobre espais de variables de qualsevol naturalesa, tant contínues com categòriques, que accepta constriccions de tipus declaratiu per a guiar el procés de classificació. En el §**??**, després de descriure la classificació ascendent jeràrquica, es detallen les mètriques i criteris d'agregació implementats. En el següent, s'explica com s'introdueix la informació declarativa al sistema per tal de millorar la classificació.

Eines d'ajuda a la interpretació dels resultats de **KLASS** es proporcionen en un cinquè capítol, per completar aquest bloc amb una secció en que es proposa una metodologia general de classificació de dominis poc estructurats i anàlisi de resultats.

El capítol §**??** està dedicat a la implementació. S'hi presenta l'estructura modular del programa, així com l'estudi de la complexitat de l'agorisme utilitzat i els aspectes més rellevants a nivell intern, mentre que la resta de capítols presenten resultats d'algunes aplicacions §**??** i §**??**, així com de contrastació amb altres mètodes §**??**.

[3]Per una caracterització d'aquests dominis en que els mètodes estàndard no tenen massa bon comportament, veure §**??**.

Per acabar hi ha un apartat de conclusions, i les directrius del treball que queda obert com a continuació d'aquest estudi.

En els apèndixs s'hi poden trobar els llistats referents a les aplicaions presentades en els capítols **??** à **??**.

Chapter 2

Caracterització del problema

2.1 Els dominis poc estructurats

S'anomena *domini poc estructurat* aquell en que pràcticament no existeix acord entre els experts a l'hora de donar definicions precisses dels conceptes, o de les relacions entre els objectes que el composen. Són exemples de *dominis poc estructurats* el de les esponges de mar [**?**], el dels transtorns mentals [**?**]...; tots ells dominis prou complexos, susceptibles de ser organitzats de formes diverses per diferents experts que segueixin criteris diversos.

Es pot dir que, i no es tenen referències de cap altra caracterització dels dominis poc estructurats, en aquests dominis concorren, habitualment, les següents particularitats:

- La descripció dels individus (o objectes) se sol fer en termes qualitatius (per bé que no totalment), inclús quan es tracta de conceptes intrínsecament numèrics. Això és degut, en part, al fet que les quantificacions impliquen força més informació, precissió i exactitud. Per exemple, la *longitud del cabell* és una variable clarament quantitativa, i no obstant rarament es tracta com a tal; és més freqüent trobar-la categoritzada en modalitats com *cabell curt*, *mitja cabellera*, *cabellera*...). Tractant dominis poc estructurats, per tant, la presència de variables qualitatives serà important.

- Es pot dir també que, quant més en sap un expert del seu tema, més rica és la seva terminologia, i més freqüent és usar termes diferents per a distingir matitzos més o menys subtils. No cal més que mencionar que el poble *innuït* (els habitants de l'Àrtic, coneguts habitualment per esquimals) té paraules per a distingir fins a **trenta** tipus diferents de **neu**!, depenent de la seva consistència, l'estació de l'any, el gruix... La descripció dels objectes en aquestes condicions es duu a terme fent ús, doncs, d'un ampli espectre de modalitats per cada una de les característiques qualitatives emprades.

- En tercer lloc, ens trobem en un tipus de dominis d'estructura complexa, difícil, sinó impossible, de formalitzar de forma completa i exacta. A més, a major expertesa, més difícil li resulta a l'especialista d'intentar explicitar el seu coneixement.

Malgrat tot, els experts sempre disposen d'una certa informació explícita i fàcilment formalitzable, que de vegades té un caràcter molt general, o pel contrari especifica bé només una part del domini. És a dir, que es tracta d'un tipus de coneixement que, lluny de la completitud, dificulta una aproximació basada en el coneixement, i que, essent valuós, és totalment ignorat sota aproximacions estadístiques. L'existència d'aquest coneixement *explícit* i *parcial* adquirirà un pes important a partir del capítol §**??**.

Aquestes són, a grans trets les principals característiques dels dominis poc estructurats que constitueixen el marc en que s'ha desenvolupat aquest treball.

2.2 Limitacions del clustering i de l'aprenentatge

2.2.1 L'Estadística

Els sistemes habituals de classificació automàtica (des d'un punt de vista estadístic) estan orientats principalment al tractament d'informació numèrica: *strictu sensu* només tracten variables quantitatives. Ara bé, donat que la presència de variables categòriques en la descripció dels individus és important, s'ha intentat de diverses formes aprofitar la informació aportada per aquest tipus de variables en el procés de classificació.

Una proposta [**?**] que gaudeix d'àmplia acceptació és descomposar cada variable categòrica en tantes variables binàries (*dummy*) com modalitats tingui, i tractar la matriu resultant com si d'una matriu numèrica es tractés.

Pel tipus de dominis que ens ocupa, aquesta tècnica produeix unes matrius de dades de gran dimensió, i molt esparses, la qual cosa augmenta artificialment el temps d'execució del procés de classificació. Aquesta serà una de les raons del nostre interés per processar directament la matriu de dades original (X).

Sobre aquesta mateixa transformació de les dades, [**?**] o [**?**] proposen una altra possibilitat: aplicar l'anàlisi de correspondències múltiple sobre la taula de variables categòriques, i fer la classificació usant les coordenades factorials. En aquest cas, la classificació té lloc en un espai fictici que, lluny de l'espai de definició de les variables originals no té una interpretació directa. Caldran, doncs, eines addicionals que permetin a l'analista interpretar les classes resultants.

Ana altra alternativa seria efectuar la classificació només amb les variables numèriques, i utilitzar després les variables categòriques (com a variables il·lustratives) per a caracteritzar les classes. En aquest cas la informació continguda en les variables categòriques no és utilitzada directament pel procés de classificació, desaprofitant-se així una font d'informació que pot tenir dimensions importants quan es tracten dominis poc estructurats com es el cas.

Altres [**?**] proposen, per exemple, la categorització de les variables quantitatives per a passar a treballar amb una taula de variables categòriques en la seva totalitat. Però, a banda

de la pèrdua d'informació que suposa la categorització de les variables numèriques, aquesta tècnica introdueix un grau d'arbitrarietat notable, inherent al propi procés de categorització: La classificació final és, en general, sensible a la forma com es defineixin les modalitats de la variable numèrica corresponent.

Un dels objectius d'aquest treball és, doncs, aprofitar la informació qualitativa que forma part de la descripció dels individus, fent-la intervenir en el procés de classificació amb el mateix paper en que intervenen les variables numèriques, però evitant, alhora, una explosió en les dimensions de la matriu de dades, així com la distorsió que suposa la categorització de les variables contínues.

La gran crítica que alguns autors, com [**?**] fan a aquests mètodes, en general, és que tots ells es basen en la definició d'una distància o funció de (dis)similitud entre individus, i que això pot conduir a resultats faltats d'una interpretació conceptual clara, entre altres coses, perquè aquestes mesures no ténen en compte la rellevància (o irrellevància) de les variables.

L'altra és que els classificadors estadístics no solen produir cap descripció conceptual de les classes, reservant plenament a l'analista la seva interpretació i caracterització [**?**], quan moltes vegades hi ha un interés clar per part de l'usuari de construir aquest tipus de descripcions, i l'ajuda del sistema seria molt apreciable.

2.2.2 La Intel·ligència Artificial

Pel que fa a la Intel·ligència Artificial, els sistemes orientats al diagnòstic tenen darrera, habitualment, un classificador heurístic. La qualitat i capacitat predictiva del sistema depèn enterament de la qualitat de la base de coneixement que el recolza, destinada a emmagatzemar el coneixement que descriu el domini[1].

Les aproximacions tradicionals de l'enginyeria del coneixement a la construcció d'aquestes bases de coneixement estan dirigides a extreure i refinar les *regles* que les componen directament de l'expert humà, i transmetre-les al sistema formalitzades en lògica de primer ordre.

Però, quan es tracta de dominis poc estructurats d'àmbit una mica extens, l'enginyer del coneixement acostuma a tenir dificultats per obtenir aquestes regles. Les raons principals són que l'estructura del domini és, en si mateixa, prou complexa i que els *experts* no disposen de regles precises en les que basar les seves decisions. Fins i tot quan la seva actuació implica l'existència d'aquestes regles, pot ser difícil obtenir-les directament:

Els experts, canalitzen la seva experiència a través de quantitats ingents de coneixement implícit i els sol costar molt explicitar-lo. En aquest sentit, [**?**] diu

> *els experts solen ser millors exercint, o executant les seves regles en la presa de decisions, que verbalitzant o enumerant les que usen en aquest procés.*

[1] El mateix passa amb les bases de coneixement sobre les que descansen els planificadors.

i [**?**], parla del *Representation Trap* que es produeix en escalar sistemes basats en el coneixement petits o mitjans als de gran dimensió, com a conseqüència de

> *representar el coneixement mitjançant predicats amb noms llargs i complexos que comporten molt de coneixement implícit, i d'escriure regles fetes massa a mida del domini d'aplicació.*

D'altra banda, si el sistema de regles no és complet, és a dir, no recull **tot** el coneixement relatiu al domini en estudi, la garantia que faci prediccions de qualitat és molt feble.

El principal requisit per a que un sistema basat en el coneixement, funcioni correctament és doncs obtenir un sistema de regles complet, i s'ha exposat la limitació tan important que això suposa. La conseqüència és que l'ús d'aquestes tècniques queda restringit a dominis prou específics com per a que sigui factible la seva formalització completa i, per tant, discutible, fins a cert punt, la seva utilitat (*i. e. toy problems*).

Aquesta mancança s'ha intentat superar a través de tècniques d'aprenentatge, que transformen una tasca d'enginyeria directa sobre les regles sol·licitades a l'expert, en una enginyeria sobre les dades, seguida d'un procés d'inducció del coneixement [**?**]. S'entra aquí en consideracions relatives a la clausura del coneixement inicial que es transmet al sistema, i a la forma com es realitza la inducció. A [**?**] es detalla els perills de les hipòtesis de món tancat que permeten fer aquesta clausura i les seves conseqüències en les inferències posteriors.

2.3 Objectius

L'objectiu fonamental d'aquest treball és contruir una metodologia d'anàlisi que permeti fer classificacions de dominis poc estructurats. Al principi del present capítol es descriuen les característiques d'aquests dominis. La secció anterior intenta donar perspectives de la problemàtica que a ells s'associa en tractar-los des de l'Estadística o des de la Intel·ligència Artificial independentment.

La metodologia que aquí es presenta intenta apropar ambdues àrees de coneixement per a resoldre la classificació de dominis poc estructurats, superant les deficiències d'una i altra.

D'una banda, cal modificar les tècniques estadístiques per tal que aprofitin la gran quantitat d'informació qualitativa que està disponible, fent-la intervenir en el procés de classificació amb un paper similar al que es dóna a la informació quantitativa.

De l'altra, serà interessant dotar aquests sistemes estadístics d'alguna capacitat per a explicar resultats que els acosti als sistemes més propis de la Intel·ligència Artificial, tot facilitant la interpretació dels resultats, tasca que fins ara requeia, majoritàriament, sobre l'usuari.

Des d'una altra òptica, es pretén donar més flexibilitat als sistemes de classificació basats en el coneixement, de forma que la completitud i homogeneïtat del sistema de regles que

proporciona l'expert no siguin fonamentals per a una bona capacitat predictiva. S'intenta, implícitament, facilitar el procés d'interacció amb l'expert que ha de generar aquestes regles.

En última instància es preten que les tècniques estadístiques contribueixin a la generació automàtica de regles.

2.4 Representació de les dades

L'objectiu de la classificació automàtica és identificar, sobre un conjunt donat d'individus $\mathcal{I} = \{i_1 \dots i_n\}$, grups diferents entre si i formats homogèniament, (veure §**??**). Aquests individus venen descrits per una sèrie de característiques[2] que poden ser de diversos tipus.

2.4.1 Les variables

Poden ser **quantitatives o qualitatives**. Les primeres són sempre conceptes mesurables i s'expressen en forma numèrica, bé en nombres reals (p.ex.: alçada) o naturals (p.ex.: edat, nombre d'habitacions d'una vivenda).

Les segones (també anomenades variables categòriques), en canvi, corresponen a altres tipus de característiques dels individus, que s'expressen mitjançant adjectius (p.ex.: mida — petita, mitjana, gran ...—, color, ...). Es denota $\mathcal{D}_k = \{c_1^k, \dots, c_{n_k}^k\}$ el conjunt de modalitats (o categories) de la variable categòrica X_k i $n_k = card(\mathcal{D}_k)$ representa el nombre de modalitats d'X_k. Així doncs, $c_j^k \in \mathcal{D}_k$, representa la $j-èsima$ modalitat d'X_k. Les variables qualitatives es poden dividir en **ordinals** i **nominals**.

Les **variables qualitatives ordinals** són aquelles per les que existeix una relació d'ordre definida sobre el conjunt $\mathcal{D}_k$ (p. ex.: les modalitats de la variable mida, definides més amunt, estan ordenades de menor a major).

Per contra, les **variables qualitatives nominals** no presenten una ordenació dels seus valors (p. ex.: estat civil — solter, casat, ajuntat, vidu, separat, divorciat — pren uns valors que no es poden ordenar). D'entre elles cal destacar les variables **binàries**, que codifiquen amb 0 i 1 l'absència o presència respectivament d'una certa propietat — (p.ex.: referent a una mostra d'animals, es podria considerar la propietat "*tenir esquelet*") —.

2.4.2 La matriu de dades

Suposi's un conjunt $\mathcal{I} = \{i_1 \dots i_n\}$ d'individus (o objectes) a classificar, els quals venen descrits per K variables $X_1 \dots X_K$ de diferents tipus. Els valors d'aquestes variables per cada un dels individus es representen en una matriu rectangular X de dimensions (n, K), tal que

$$X = (x_{ik}\ (i \in \mathcal{I}), (k = 1 : K))$$

[2] Altrament dites *variables*, o *propietats* o *atributs*.

Nom	*Alçada*	*Cabell*	*Duu ulleres*	*Nom Cònjuge*
Joan	1.60m	Ros	1	?
Maria	1.65m	Negre	0	Jordi
Pep	1.70m	Castany	0	Rosa
⋮	⋮	⋮	⋮	⋮

Table 2.1: Exemple de matriu de dades a classificar.

i on els elements x_{ik} són el valor que, per l'individu $i-èsim$, pren la variable $k-èsima$.

És a dir, que les files de la matriu de dades X contenen informació relativa als individus, mentre que les columnes, fan referència a les variables. En alguns apartats d'aquest document serà útil considerar la descripció d'un individu i com el vector $x_i = (x_{i1}, \ldots, x_{iK})$.

Degut al que s'ha exposat a §**??**, es consideraran matrius de dades heterogènies on els individus es poden caracterizar simultàniament amb variables numèriques i categòriques, aquestes darreres amb un gran nombre de modalitats.

La taula **??** exemplifica l'aspecte que pot oferir una matriu de dades a classificar.

2.5 Valors mancants

S'anomena *valor mancant* (*missing value*) a tota observació no present a la matriu de dades. Un valor mancant sempre és font de files incomplertes de la matriu de dades. L'existència de valors mancants en una matriu de dades constitueix una problemàtica gens despreciable pel que fa a l'anàlisi de tals matrius , en primer lloc, per la reducció d'informació útil que suposen (la qual cosa dificulta el tractament), però també perquè posen en perill la validesa de l'anàlisi realitzada, així com la de les conclusions de l'estudi.

La naturalesa dels valors mancants és variada:

- En gran part de les ocasions, la dada mancant apareix directament a la fase de recollida d'informació:

  Bé sigui perquè l'obtenció de la dada és molt costosa (la determinació del nivell de nicotina en saliva d'un fumador passa per una analítica complicada, i per tant, cara; més que la determinació del nivell de nicotina en sang) o tècnicament impossible (el càlcul de la paral·laxa d'una estrella molt llunyana pel mètode trigonomètric[3]), bé sigui perquè comporta una prova destructiva que redueix la població en estudi (el temps de supervivència, sota una certa dosi de fàrmac, d'un pacient amb determinada malaltia no es pot quantificar fins que ha tingut lloc la mort del pacient); ja es tracti d'un valor inaplicable (tipus de còrtex d'una esponja de mar del gènere Timea, que és un gènere

[3]L'exemple correspon a l'aplicació presentada a §**??**.

d'esponges sense còrtex[4]) ja tingui implicacions ètiques que creïn prejudicis en qui ha de proporcionar la informació (en les enquestes és freqüent que hi hagi reticències, per exemple, en facilitar el nivell d'ingressos mensuals), es generen cel·les buides de la matriu de dades perquè el contingut d'aquestes cel·les ni tan sols ha estat recollit.

- Evidentment, poden aparèixer dades mancants per errors de transcripció en les dades recollides, o per fallida tècnica en la fase de transmissió d'informació.

- I un darrer cas que citarem (sense pretendre ser exhaustius en absolut) serà el de la reutilització d'informació. L'ús de dades recollides prèviament per altres fins pot suposar la incorporació de dades mancants. Per exemple, en estudis arqueològics, no sempre es poden recuperar característiques d'objectes que ja han desaparegut (alçada d'una casa de la que només en resta la planta), o recollir dades històriques per a fer estudis sòcioeconòmics sol involucrar l'aparició de dades mancants en les dades més antigues (p. ex. l'Instituto Nacional de Estadística fa, desde 1985, una Encuesta Continua sobre Presupuestos Familiares similar, pero no idèntica a la que fan altres països de la CEE. Mentre no es completi el procés de normalització que està portant a terme EUROSTAT[5], estudis a nivell europeu generarien valors mancants en aquelles variables que Espanya no recull).

Un dels casos amb conseqüències més nefastes és el de les dades que falten de forma no accidental, les quals solen correspondre a un sector *específic* de la població amb unes determinades característiques (la seva distribució no és uniforme sobre el domini en estudi). Per exemple, una dada mancant en *nivell d'ingressos*, sol correspondre a persones amb ingressos o molt, molt elevats, o molt reduïts. El problema és que, en aquestes condicions, hi ha certes parts del domini (els rics o els pobres, en l'exemple anterior) *infrarepresentades* ! i les conclusions de l'anàlisi poden ser totalment errònies. Un primer pas és intentar, des de la recollida de dades, o fent un seguiment de l'individu, minimitzar tant com sigui possible la presència d'aquest tipus de valors mancants [**?**].

D'altra banda, és important que l'existència de valors mancants en unes poques components no suposi el desaprofitament de la informació útil que conté l'entitat corresponent. Les tècniques per a emprar aquesta informació present poden ser molt sofisticades, però donat que el tractament de valors mancants no és un tema central d'aquest estudi, s'ha optat per una solució de nivell intermedi. Entre les alternatives més senzilles hi ha:

- Obviar totes aquelles files de la matriu de dades que contenen algun valor mancant, amb la consegüent reducció d'informació útil.

[4]L'exemple correspon a l'aplicació de §**??**.
[5]L'Oficina Europea d'Estadística.

- Recodificar tots els valors mancants a alguna constant tractable des del punt de vista de l'algorisme.

- Substituir els valors mancants per algun valor que no pertorbi, estadísticament parlant, la informació proporcionada pels valors presents de cada variable.

En el nostre sistema, **KLASS**, hem optat per la tercera alternativa, substituint els valors mancants de les variables quantitatives pel seu valor mig. En particular, si es treballa amb variables normalitzades segons la distància dels rangs, es substitueixen per 0.5, valor mig de la variable si aquesta ha estat reescalada a l'interval [0,1] (normalització amb la inversa dels rangs).

La substitució del valor mancant per la mitjana té la propietat d'actuar com una mena d'element neutre de la classe, pel que fa al càlcul del seu representant.

En efecte, com es veurà en el capítol **??** el representant d'una classe $\mathcal{C} = \{i_1 \dots i_{n_\mathcal{C}}\}$ per la variable X_k, si aquesta és numèrica, és precisament la mitjana dins la classe d'aquesta variable:

$$\bar{x}_{\mathcal{C}k} = \frac{\sum_{i=1}^{i=n_\mathcal{C}} x_{ik}}{n_\mathcal{C}} \tag{2.1}$$

L'engrandiment de $\mathcal{C}$ amb objectes de valor $\bar{x}_{\mathcal{C}k}$ no altera la mitjana d'X_k dins la classe, i per tant, tampoc el valor d'X_k pel representant de la classe. Suposant η valors mancants que se substituïssin per $\bar{x}_{\mathcal{C}k}$ i s'afegissin a $\mathcal{C}$, la mitjana de la nova classe $\mathcal{C}'$ seria:

$$\bar{x}_{\mathcal{C}'k} = \frac{\sum_{i=1}^{i=n_\mathcal{C}} x_{ik} + \eta\bar{x}_{\mathcal{C}k}}{n_\mathcal{C} + \eta}$$

segons **??**,

$$\bar{x}_{\mathcal{C}'k} = \frac{n_\mathcal{C}\bar{x}_{\mathcal{C}k} + \eta\bar{x}_{\mathcal{C}k}}{n_\mathcal{C} + \eta} = \frac{(n_\mathcal{C} + \eta)\bar{x}_{\mathcal{C}k}}{n_\mathcal{C} + \eta}$$

i simplificant,

$$\bar{x}_{\mathcal{C}'k} = \bar{x}_{\mathcal{C}k}$$

■

Quant a les variables qualitatives, aprofitant la introducció de valors vectorials que es presenta amb motiu de la construcció de representants de classe (veure §**??**), es substitueixen els valors mancants per l'element $(f_\mathcal{C}^{k_1} \dots f_\mathcal{C}^{k_{n_k}})$, on $f_\mathcal{C}^{k_j}$ representa la proporció d'objectes de $\mathcal{C}$ amb valor d'X_k igual a c_j^k. Essent $I_\mathcal{C}^{k_j} = card\{i \in \mathcal{C} : x_{ik} = c_j^k\}$, es pot escriure també

$$f_\mathcal{C}^{k_j} = \frac{I_\mathcal{C}^{k_j}}{n_\mathcal{C}}$$

Amb això es manté la qualitat d'*element neutre* respecte de l'operació d'agregació, també per les variables qualitatives.

En efecte, donada una classe $\mathcal{C}$ de representant[6] $\bar{\imath}_{\mathcal{C}}$, que per la variable k pren el valor

$$\bar{x}_{\mathcal{C}k} = (f_{\mathcal{C}}^{k_1} \dots f_{\mathcal{C}}^{k_{n_k}})$$

L'agregació d'η objectes a $\mathcal{C}$ amb valor mancant a la component k, no altera el representant en X_k si es substitueixen aquests valors mancants per $\bar{x}_{\mathcal{C}k}$. La component k del representant de la nova classe $\mathcal{C}'$ seria:

$$\bar{x}_{\mathcal{C}'k} = (f_{\mathcal{C}'}^{k_1} \dots f_{\mathcal{C}'}^{k_{n_k}}) \quad tq \quad f_{\mathcal{C}'}^{k_j} = \frac{I_{\mathcal{C}'}^{k_j}}{n_{\mathcal{C}} + \eta} \tag{2.2}$$

on $I_{\mathcal{C}'}^{k_j}$ són els individus de $\mathcal{C}'$ que prenen valor c_j^k per la variable X_k. Tenint en compte la composició de $\mathcal{C}'$ i la forma com s'han substituït els valors mancants, $I_{\mathcal{C}'}^{k_j} = I_{\mathcal{C}}^{k_j} + \eta f_{\mathcal{C}}^{k_j}$. I com que $I_{\mathcal{C}}^{k_j}$ és precisament $n_{\mathcal{C}} f_{\mathcal{C}}^{k_j}$, resulta que

$$f_{\mathcal{C}'}^{k_j} = \frac{n_{\mathcal{C}} f_{\mathcal{C}}^{k_j} + \eta f_{\mathcal{C}}^{k_j}}{n_{\mathcal{C}} + \eta} = \frac{(n_{\mathcal{C}} + \eta) f_{\mathcal{C}}^{k_j}}{(n_{\mathcal{C}} + \eta)} = f_{\mathcal{C}}^{k_j}$$

■

La representació que s'ha considerat pels valors mancants és '?' tot i que la possibilitat d'usar altres codis és una qüestió purament tècnica que s'ha deixat oberta en l'algorisme.

2.6 Resum del capítol

En primer lloc s'ha elaborat una caracterització dels *dominis poc estructurats*, els quals, per les seves característiques intrínseques, constitueixen problemes difícils de tractar amb les tècniques existents actualment en Estadística i Intel·ligència Artificial.

En la segona secció s'enuncien les raons per les que això és així. Pel que fa a la Intel·ligència Artificial, l'estructura complexa d'aquests dominis fa que la construcció d'una base de coneixement complet sobre ell, per a un sistema de diagnòstic, es torni pràcticament inabordable: l'expert no sempre sap, o pot, formalitzar explícitament les regles que, havent adquirit amb l'experiència, aplica inconscientment.

Quant a l'Estadística, les actuals tècniques de clustering (que es basen en distàncies, situades en un pla sintàctic) tenen un comportament pobre davant d'aquests tipus de dominis; s'han de gestionar simultàniament variables numèriques i quantitats ingents d'informació qualitativa, la qual cosa no respon exactament a les situacions per a les que les tècniques de clustering pures han estat pensades.

L'objectiu principal d'aquest treball és desenvolupar una metodologia que apropi ambdues ciències per a resoldre la classificació de dominis poc estructurats, superant les deficiències d'una i altra. A la seccio **??** es descriuen els objectius de la tesi amb més detall.

[6] En la notació utilitzada, la barra fa referència a la idea, enunciada en seccions posteriors, que el prototipus és la *mitjana* de la classe.

A continuació es presenta l'estructura general de les matrius de dades que **KLASS** tractarà, de la que cal destacar la coexistència d'informació qualitativa i quantitativa, per acabar detallant el tractament de valors mancants adoptat pel nostre sistema.

Chapter 3

Algorisme bàsic

"És impossible que d'éssers contraris
se segueixi concordància alguna".
R.LLULL

3.1 La classificació ascendent jeràrquica

L'algorisme de base de la metodologia que es presenta en aquest document és un mètode de classificació ascendent jeràrquica. Es podria resumir el principi d'aquests mètodes dient que, a partir d'un conjunt $\mathcal{I}$, $card(\mathcal{I}) = n$, d'individus a classificar, llur classificació ascendent jeràrquica consisteix a construir una successió de particions encaixades $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\}$ sobre $\mathcal{I}$ de tal forma que els elements d'$\mathcal{I}$ queden organitzats com a nodes terminals d'un arbre binari (o arbre jeràrquic)[1].

Aquesta successió és tal que $\mathcal{P}_1$ és la partició més fina possible dins d'$\mathcal{I}$

$$\mathcal{P}_1 = \{\{i\} : i \in \mathcal{I}\},$$

$\mathcal{P}_n$ és la més grollera

$$\mathcal{P}_n = \{\{i : i \in \mathcal{I}\}\},$$

i en un pas intermedi qualsevol ($\xi = 1 : n-1$), en que s'agreguen dos elements de $\mathcal{P}_\xi = \{\mathcal{C}_1 \ldots \mathcal{C}_{n-\xi+1}\}$, siguin[2] $\mathcal{C}_e$ i $\mathcal{C}_d$

$$\mathcal{P}_{\xi+1} = \mathcal{P}_\xi \cup \{\mathcal{C}_e \cup \mathcal{C}_d\} - \{\mathcal{C}_e, \mathcal{C}_d\}$$

Considerant el conjunt $\mathcal{I} = \{1, 2, 3, 4, 5, 6\}$, una possible classificació ascendent jeràrquica vindria descrita per la successió de particions

$$\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_6\}, \quad \text{on} \tag{3.1}$$

[1]Sempre respectant, evidentment, criteris d'homogeneïtat dels individus d'una mateixa classe i heterogeneïtat de les classes entre elles.

[2]Quan es faci la representació en arbre de la classificació, $\mathcal{C}_e$ serà el fill esquerre i $\mathcal{C}_d$ serà el fill dret del node que representi la nova classe.

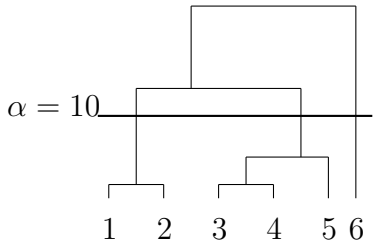


Figure 3.1: Arbre jeràrquic amb α-tall a nivell 10.

$$\begin{aligned}
\mathcal{P}_1 &= \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\} \\
\mathcal{P}_2 &= \{\{1,2\},\{3\},\{4\},\{5\},\{6\}\} \\
\mathcal{P}_3 &= \{\{1,2\},\{3,4\},\{5\},\{6\}\} \\
\mathcal{P}_4 &= \{\{1,2\},\{3,4,5\},\{6\}\} \\
\mathcal{P}_5 &= \{\{1,2,3,4,5\},\{6\}\} \\
\mathcal{P}_6 &= \{\{1,2,3,4,5,6\}\}
\end{aligned}$$

Aquesta relació entre $\mathcal{P}_\xi$ i $\mathcal{P}_{\xi-1}$ permet fer una representació gràfica de $\mathcal{P}$ en forma d'arbre jeràrquic. Un arbre jeràrquic (o dendrograma) té: en els nodes interns, les diferents subagrupacions establertes sobre els individus; a les fulles del corresponent subarbre, els elements que formen part de cada subclasse, i branques de diferents longituds, que situen els nodes interns a diferents nivells respecte de l'horitzontal sobre la que es disposen els individus. El nivell dels nodes, normalment xifrat, indica el grau de semblança (similitud) dels seus fills, i està relacionat directament amb la distància entre ells en l'espai de les variables (veure §**??**). Quant més gran sigui la longitud de la branca que uneix dos fills al seu pare, menys s'assemblen les subclasses que aquests nodes representen.

El tall horitzontal de l'arbre, donat un determinat nivell α, proporciona una certa partició de la mostra en classes (anomenada *α-partició*). Fent variar el nivell del tall, obtenim les diverses classificacions[3] de $\mathcal{P}$. Amb això, hom té classificacions del mateix col·lectiu d'individus a diferents graus d'abstracció, i es podrà elegir la que més s'ajusti als propòsits de l'usuari. És a partir d'un estudi *a posteriori* d'aquest arbre que es tria la classificació definitiva, estudi per al que **KLASS** ofereix algunes ajudes que es tractan a §**??**.

La figura **??** representaria un possible arbre jeràrquic corresponent a la classificació de l'exemple anterior. El tall de la figura **??** particiona la mostra en tres classes: la primera conté els elements 1 i 2, la segona està formada pels elements 3, 4, i 5, i la tercera únicament el 6. Així doncs, la successió de l'expressió (**??**) i l'arbre de la figura **??** són dues representacions alternatives de la classificació ascendent jeràrquica realitzada sobre el

[3]D'aquí endavant, s'usaran indistintament els termes *classificació* i *partició*.

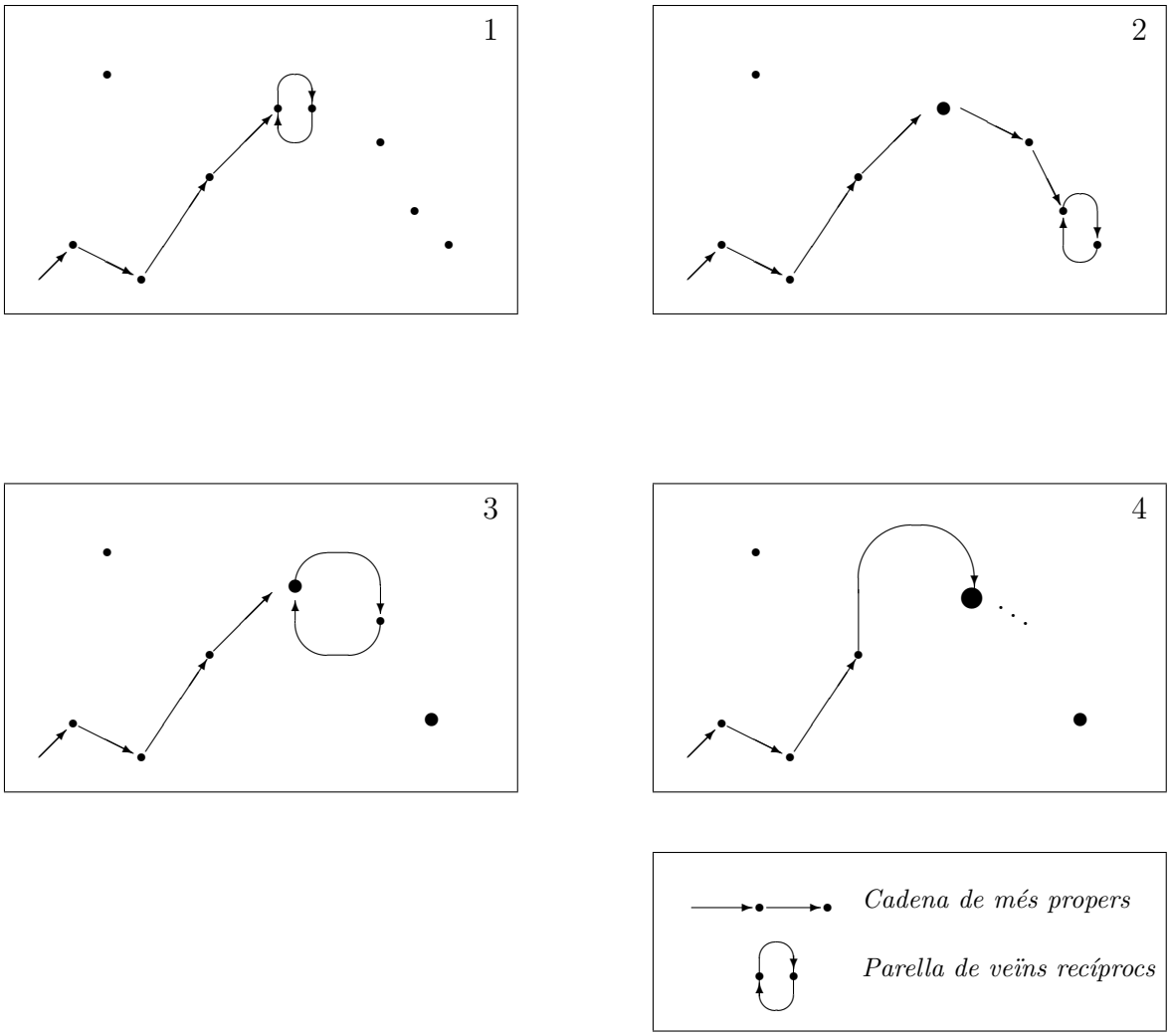


Figure 3.2: El procés dels veïns recíprocs encadenats.

conjunt $\mathcal{I} = \{1, 2, 3, 4, 5, 6\}$.

La representació en arbre, com es pot veure, aporta més informació que la representació en forma de successió de particions (veure **??** sobre les implicacions teòriques dels dendrogrames), donat que indica quina relació de semblances hi ha entre els nodes que s'agreguen d'una partició $\mathcal{P}_\xi$ a la següent $\mathcal{P}_{\xi+1}$.

El pas de $\mathcal{P}_\xi$ a $\mathcal{P}_{\xi+1}$ requereix però, que s'especifiqui com determinar les entitats a fusionar[4] en cada pas, així com la forma d'efectuar aquesta fusió.

L'algorisme dels **veïns recíprocs** [**?**] utilitza un concepte propi per a determinar quins són els individus que s'agreguen:

> S'anomena **veïns recíprocs** als individus i, i' si i és l'objecte més proper a i' en la mostra donada, i i' és també el més proper a i.

[4]En endavant s'utilitzaran indistintament els termes *fusió* i *agregació*.

```
veins reciprocs (nuvol, element, metrica, criteri_d'agregacio)
      vei:=mes_proper(element, metrica)
      vei_reciproc:=mes_proper(vei, metrica)
      Si element=vei_reciproc
      llavors
         classe:=agregar(element, vei, criteri_d'agregacio)
         nuvol_aux:=treure((element, vei), nuvol)
         nuvol':=insertar(classe, nuvol)
         element':=anterior(element)
         veins_reciprocs(nuvol', element', metrica, criteri_d'agregacio)
      sino
         anteriors:=insertar(element, anteriors)
         veins_reciprocs(nuvol, vei, metrica, criteri_d'agregacio)
      fsi
```

Table 3.1: Esquema algorísmic de la classificació per veïns recíprocs encadenats.

En aquest algorisme doncs, sempre s'agreguen parelles de veïns recíprocs. Això requereix la definició d'una mètrica sobre l'espai de les variables que permeti de calcular la distància entre dos individus, per tal d'identificar quines són les parelles d'elements més propers.

Quant al procés d'agregació, hi ha diverses possibilitats, que es detallen en l'apartat **??** i que donen lloc a diferents formacions de les classes. És a dir, segons el criteri d'agregació, la característica rellevant d'una classe variarà[5]. Entre altres, es poden utilitzar com a selectors dels nodes a agregar, el seu centre de gravetat, o la distància entre els dos individus més *separats* d'entre els de la classe, o la dels més *propers*, ..., sempre en termes de la distància definida.

L'estratègia de classificació escollida consisteix, doncs, a detectar les parelles de veïns recíprocs que es poden fusionar i anar construint l'arbre d'agregacions. Per a la implementació es pot aprofitar el fet que la relació de veïnatge és local [**?**], i no queda afectada pel procés d'agregació. Això es contempla en l'algorisme dels **veïns recípocs encadenats**, de complexitat quadràtica, que d'altra banda és de fàcil tractament recursiu, i per tant, adient per a ser implementat en **LISP**[6].

Aquesta versió (veure figura **??**) parteix d'un objecte de la mostra, busca el seu veí més proper, i va saltant de més proper en més proper fins que detecta dos individus per a agregar. En aquest punt s'efectua la fusió i es torna a classificar a partir del darrer element de la *cadena* de més propers que s'ha anat construint (d'aquí el nom de l'algorisme), però considerant ara la nova classe com un altre objecte. Els seus components desapareixen com a individus i ja no s'utilitzaran més en la classificació.

[5]Pròpiament, el que realment varia és com es determina la classe més propera d'una donada, definida la distància entre dos individus.

[6]S'utilitza el SUN-COMMONLISP 4.0.1. Existeix també una versió en VAX-COMMONLISP 2.2.

Per a fer-ho, resulta molt útil de mantenir una matriu de distàncies entre individus, la gestió de la qual es detalla a l'apartat **??**. A **??** es presenta l'esquema algorísmic corresponent als **veïns recíprocs encadenats**.

La funció **agregar** actualitza totes les estructures d'acord amb la formació d'una classe amb *element* i *vei*. S'indica en cursiva quines són les funcions dependents de la distància definida sobre l'espai de les variables o del criteri d'agregació seleccionat. En la resta del capítol es donen detalls de la representació de les classes, les distàncies i criteris d'agregació emprats per **KLASS**.

3.2 Sobre l'índex de nivell de la jerarquia

Puix que ha de donar idea de la distància existent entre els seus fills, l'índex de nivell d'un node de la jerarquia $\mathcal{C} = \{\mathcal{C}_e, \mathcal{C}_d\}$ es defineix habitualment com

$$\nu(\mathcal{C}) = d^2(\mathcal{C}_e, \mathcal{C}_d)$$

i es representa com l'ordenada dels nodes de l'arbre ascendent jeràrquic.

Evidentment, els nodes terminals tenen índex de nivell 0: Tot node terminal representa l'agregació d'un element amb si mateix: $\nu(\mathcal{C}) = d^2(i, i) = 0$

Si $\nu(\mathcal{C})$ satisfà que $\mathcal{C} \subset \mathcal{C}' \Longrightarrow \nu(\mathcal{C}) < \nu(\mathcal{C}')$, llavors es té una *jerarquia indexada*, i construir una jerarquia indexada sobre el conjunt $\mathcal{I}$ és equivalent a definir una ultramètrica sobre $\mathcal{I}$ [**?**]. Les $\alpha-$particions d'una jerarquia indexada tenen una interpretació clara com a classificacions d'$\mathcal{I}$. És a dir, que d'alguna manera, trobar una ultramètrica sobre $\mathcal{I}$ garantitza l'existència d'un dendrograma que permet fer classificacions sobre aquest conjunt.

Una ultramètrica es defineix com una funció d sobre $\mathcal{I} \times \mathcal{I}$ amb les següents propietats:

$$i) \quad d(i, i') = 0 \Longleftrightarrow i = i',\ \forall i, i'$$

$$ii) \quad d(i, i') = d(i', i),\ \forall i,\ i'$$

$$iii) \quad d(i, i') \leq \max\{d(i, i''), d(i', i'')\},\quad \forall i, i', i''$$

essent la tercera propietat més restrictiva que la desigualtat triangular, axioma, aquest darrer, que han de satisfer totes les distàncies. Així, una ultramètrica sempre és una distància, mentre que el recíproc no té perquè ser cert.

Es diu que un arbre presenta *inversions* (veure figura **??**) quan no és una jerarquia indexada; en altres paraules, quan hi ha nodes de l'arbre que tenen un índex de nivell menor que el d'algun dels seus fills:

$$\exists \mathcal{C} \subset \mathcal{C}'\ tq\ \nu(\mathcal{C}) > \nu(\mathcal{C}')$$

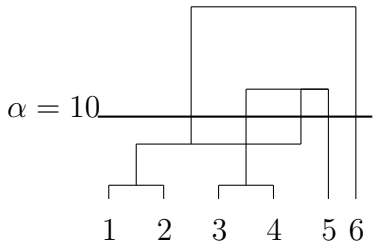


Figure 3.3: Arbre jeràrquic amb inversions.

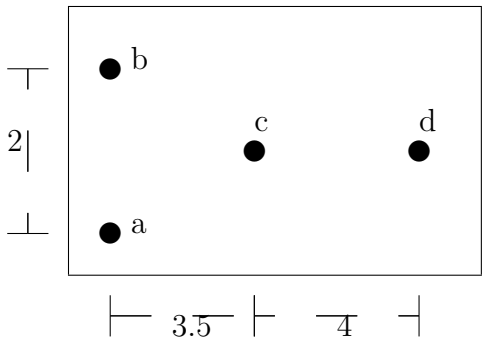


Figure 3.4: Exemple amb inversions i classificació inestable.

Aquest és un cas no desitjable, ja que la interpretació dels talls d'un arbre d'aquestes característiques és, en general, confusa.

Per exemple, l'$\alpha-$partició marcada sobre l'arbre de la figura **??** es podria interpretar com la partició $\{\{1,2,3,4,5\},\{6\}\}$ si es dóna prioritat al nivell dels pares, o bé com $\{\{1,2\},\{3,4\},\{5\},\{6\}\}$, si es considerés que tenen prioritat els nivells dels fills.

D'altra banda, si no es treballa en un espai ultramètric, el resultat de la classificació pot variar segons el punt pel que s'inicïi el procés de classificació. En efecte, amb un núvol de punts com el de la figura **??**, es poden obtenir les jerarquies **??** dreta o esquerra segons es comenci a classificar pel punt a o d respectivament.

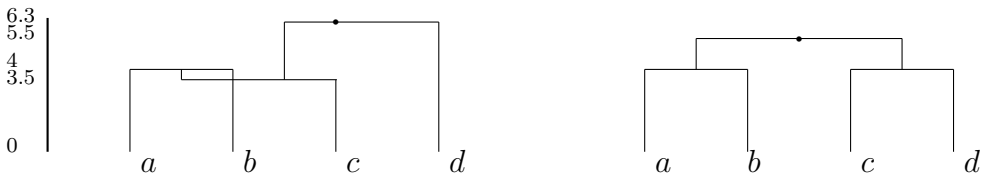


Figure 3.5: Començant per a (esquerra); començant per d (dreta).

3.3 La representació de les classes

L'explicitació d'un element que representi cada una de les classes formades per KLASS permet:

- D'una banda, tractar de forma *homogènia* individus pròpiament dits i les classes, les quals s'identificaran amb el seu representant. Això adquirirà especial rellevància en el capítol **??**.

- Proporcionar una descripció *conceptual* (o *prototípica*) de les classes formades. Aquesta via obre les portes de la generació automàtica de regles de classificació per a un sistema basat en el coneixement, a més de dotar de certa capacitat expressiva al sistema, que facilitarà la interpretació de les classes resultants (veure §**??**).

La idea de prototipus $\bar{\imath}_c$ d'un conjunt d'individus $\mathcal{C} = \{i_1 \dots i_{n_\mathcal{C}}\} \subseteq \mathcal{I}$ porta a pensar un individu, real o no, que sintetitzi les característiques dels seus representats.

Consideri's la descripció d'$\bar{\imath}_\mathcal{C}$, $\bar{x}_\mathcal{C} = (\bar{x}_{\mathcal{C}1} \dots \bar{x}_{\mathcal{C}K})$. El càlcul d'$\bar{x}_\mathcal{C}$ es fa component a component, i cal estudiar per separat el cas de les components numèriques i el de les categòriques.

3.3.1 El centre de gravetat d'una variable

Pel que fa a les components numèriques del vector $\bar{x}_\mathcal{C}$, el concepte de prototipus suggereix ràpidament la idea d'individu intermedi i, per tant, descriptors de tendència central com ara la mitjana o la mediana. Ens ha semblat adient l'ús de la mitjana, fonamentalment per dues raons:

- Per bé que la mediana té un comportament més robust enfront dels valors aberrants, el seu càlcul és considerablement més costós que el de la mitjana, donat que suposa un procés d'ordenació de les observacions.

- L'ús de la mitjana permet un tractament i interpretació anàlegs en variables quantitatives i qualitatives (com es veurà més endavant), la qual cosa mai seria possible si s'usés la mediana: el seu càlcul descansa sobre l'existència d'un ordre definit sobre els valors de les variables, que no existeix per les variables nominals.

Així doncs, i pel que fa a les variables quantitatives,

$$\bar{x}_{\mathcal{C}k} = \frac{\sum_{i \in \mathcal{C}} x_{ik}}{n_\mathcal{C}},\ k \text{ tq } X_k \text{ és quantitativa} \tag{3.2}$$

Convé en aquest punt, pel que es veurà tot seguit, recordar que si en una classe $\mathcal{C}$ hi ha T valors diferents $\{x_1 \dots x_T\}$ per la variable X_k i f_t és la proporció d'objectes de $\mathcal{C}$ de valor

x_t, hi ha una expressió alternativa per **??**, que és:

$$\bar{x}_{\mathcal{C}k} = \sum_{t=1}^{T} f_t x_t \tag{3.3}$$

i que el valor $\bar{x}_{\mathcal{C}k}$ s'interpreta [**?**] com el valor que prendrien les observacions si fossin totes iguals, i iguals al seu centre de gravetat.

Pel que fa a les variables categòriques, la manera més natural[7] de sintetitzar un grup d'individus en aquest cas és detallant la forma com aquests es distribueixen en les diferents modalitats de la variable, bé sigui mitjançant proporcions[8], o percentatges, o contingents absoluts.

Considerant que en tota classificació jeràrquica se succeeixen nivells creixents d'abstracció, i que es pretén que les subclasses siguin tractables com objectes ordinaris, la solució de les proporcions sembla la més adequada, ja que permet fer la següent interpretació:

> *El centre de gravetat d'un conjunt d'individus en un espai de variables qualitatives és* **una** *entitat fictícia que no té una modalitat definida per a cada variable, sinó que es reparteix en les diferents categories de cada una de forma proporcional a com ho fan els objectes que el generen.*

Per exemple:

> Suposi's un núvol de punts definit sobre l'espai de les variables **FORMA** i **COLOR** amb les modalitats
> *(quadrat, triangular, rodó)* i *(groc, blau, vermell, negre)*
> respectivament, i dos individus d'aquest núvol definits de la forma
>
> x_i = *(quadrat, blau),* x'_i= *(quadrat, groc)* .
>
> L'*agregació* d'i i i' en la classe $\mathcal{C}$ seria, seguint aquesta aproximació:
>
> $$\bar{x}_{\mathcal{C}} = \textit{(quadrat, (1/2 blau) (1/2 groc))}$$
>
> Suposi's també que i i i' són taques de llum projectades sobre una paret, i que el concepte d'agregació es materialitza en la superposició dels components de la classe. Llavors $\bar{\imath}$ no seria res més que una taca quadrada, i aproximadament de color verd. La propietat additiva dels colors és particularment apropiada per a interpretar aquest cas i ofereix una idea intuïtiva del significat que té l'agregació de dos individus en termes de variables qualitatives[9]. La figura **??** il·lustra aquesta idea.

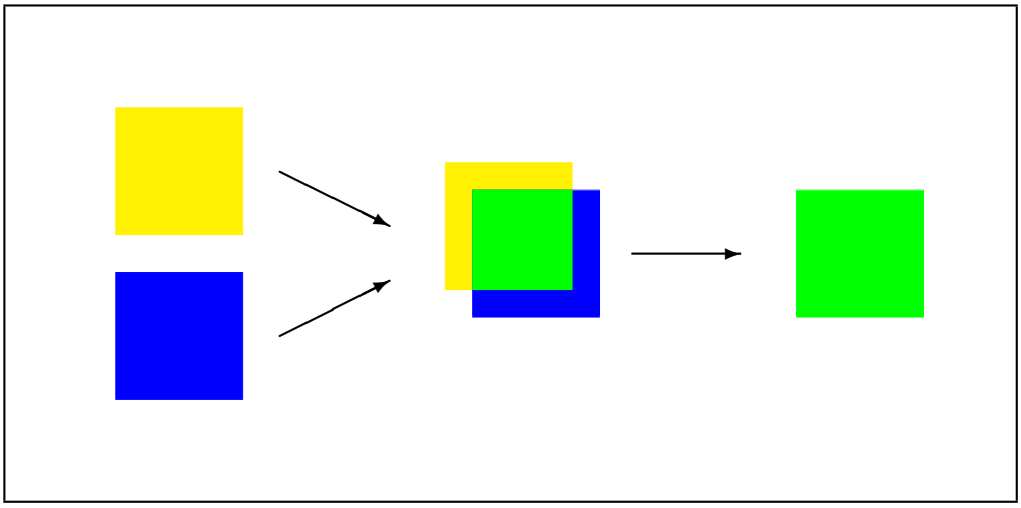


Figure 3.6: Representació del centre de gravetat en variables qualitatives.

La interpretació, a la majoria dels textes, és de tipus conjuntiu amb disjunció interna. Sota aquesta aproximació, la classe de l'exemple s'interpretaria com *La classe* $\mathcal{C}$ *d'objectes de* **forma** $=$ *quadrat* $\wedge$ **color** $=$ *blau* $\vee$ *groc*. La present proposta és encara més informativa que aquestes descripcions conjuntives, pel fet d'incorporar les proporcions amb que cada valor apareix en una certa classe.

És més, $\bar{\imath}$ donaria les característiques que tindrien les taques de la classe $\mathcal{C}$ si haguessin de ser totes iguals.

> *Aquesta és precisament la interpretació de la mitjana per les variables quantitatives!.*

Per tant, donada una classe $\mathcal{C} = \{i_1 \dots i_{n_\mathcal{C}}\}$ i la variable qualitativa X_k que pren valors en el conjunt $\mathcal{D}_k = \{c_1^k \dots c_{n_k}^k, \}$, la $k-$*èsima* component de la descripció del representant de $\mathcal{C}$, $\bar{\imath}_\mathcal{C}$ és[10]:

$$\bar{x}_{\mathcal{C}k} = \left((f_\mathcal{C}^{k_1} \ c_1^k) \dots (f_\mathcal{C}^{k_{n_k}} \ c_{n_k}^k) \right), \quad f_\mathcal{C}^{k_j} = \frac{card\{i \in \mathcal{C} : \ x_{ik} = c_j^k\}}{n_\mathcal{C}} \tag{3.4}$$

[7] Alguns autors utiltitzen la *moda* com representant d'una classe, però com es veurà més endavant, aquest descriptor sintetitza massa la informació de la classe.

[8] En termes més tècnics, quan es parla de proporcions s'està fent referència, ni més ni menys, que a la distribució marginal de cada variable.

[9] En general, es classificaran objectes discrets, i aquesta *confusió* dels integrants de la classe en un únic objecte representant —que recorda el Misteri Trinitari —, no tindrà lloc, però com a il·lustració ens ha semblat molt plàstica. L'agregació, per exemple, de dues comunitats d'igual grandària, una católica i una musulmana, donaria lloc a una nova comunitat, el representant de la qual hauria de ser (o d'actuar com si fos) 50% catòlic i 50% musulmà.

[10] No cal dir que si X_k pren un mateix valor en tota la classe (c_s^k), el representant de classe serà tal que $\bar{x}_{\mathcal{C}k} = c_s^k$ també.

Anomenant $I_{\mathcal{C}}^{k_j}$ al nombre d'individus de la subclasse representada per $\mathcal{C}$ de modalitat c_j^k, es poden denotar les components del vector $\bar{x}_{\mathcal{C}k}$ com

$$f_{\mathcal{C}}^{k_j} = \frac{I_{\mathcal{C}}^{k_j}}{\sum_{j=1}^{n_k} I_{\mathcal{C}}^{k_j}} = \frac{I_{\mathcal{C}}^{k_j}}{n_{\mathcal{C}}} \tag{3.5}$$

Justament, **??** es pot considerar com una generalització de l'expressio **??** on les c_j^k juguen el paper de les x_t i les $f_{\mathcal{C}}^{k_j}$ el de les f_t, en un cas en que, per la naturalesa de les c_j^k, sumes i productes no tenen sentit. Aquesta analogia permet interpretar $\bar{x}_{\mathcal{C}}$ com el *centre de gravetat* de la classe $\mathcal{C}$.

En endavant, i per simplificar la notació, es representaran les components qualitatives del centre de gravetat d'una forma equivalent a l'anterior:

$$\bar{x}_{\mathcal{C}k} = \left((f_{\mathcal{C}}^{k_1}\ c_1^k)\dots(f_{\mathcal{C}}^{k_{n_k}}\ c_{n_k}^k)\right) \equiv \left(f_{\mathcal{C}}^{k_1}\dots f_{\mathcal{C}}^{k_{n_k}}\right) \tag{3.6}$$

Amb aquesta definició de prototipus d'una classe per les variables qualitatives, que fins on nosaltres coneixem és la primera vegada que s'utilitza, apareixerà un nou tipus d'objecte que pot tenir components vectorials en algunes variables categòriques: Un objecte qualsevol i pot tenir, per a la variable categòrica k, un valor del tipus $(f_i^{k_1},\dots,f_i^{k_{n_k}})$ on $f_i^{k_j}, (j = 1 : n_k)$ és la proporció d'objectes elementals de la classe representada per i sobre la categoria c_j^k. Sota aquesta representació, els valors possibles de les variables categòriques seran un subconjunt de $\mathcal{D}_k \cup [0,1]^{n_k}$, donat que les $f_i^{k_j}$ són proporcions, i sempre satisfan

i) $f_i^{k_j} \geq 0\ ,\ \forall i$

ii) $\sum_{j=1}^{n_k} f_i^{k_j} = 1$

3.3.2 Objectes compactes i objectes extesos

S'anomena *valor en forma extesa* a tot valor vectorial d'una variable qualitativa, i *valor en forma compacta* al que és representable mitjançant un símbol. Tot valor en forma compacta té una representació en forma extesa, però el recíproc únicament és cert per als valors extesos $(f_i^{k_1},\dots,f_i^{k_{n_k}})$ que pertanyen a la base canònica de $\Re^{n_k}$, és a dir, aquells que estan totalment concentrats en una única modalitat.

Quan es faci referència a trobar la representació extesa d'un valor compacte, es parlarà d'*extendre el símbol*.

Es considera *objecte compacte* tot aquell que té valors compactes en totes les seves variables categòriques, mentre que els *objectes extesos* presenten valors extesos almenys en una de les variables categòriques. Tots els nodes terminals d'un arbre jeràrquic són objectes compactes. Si almenys una de les variables qualitatives presents en la descripció

dels individus és informativa (no és constant per a tots els individus), l'arrel de l'arbre jeràrquic serà un objecte extès.

En un cas genèric, els nodes interns de l'arbre tenen tanta més tendència a ser extesos quan més alt sigui el seu índex de nivell.

3.3.3 Càlcul del centre de gravetat

Vist com es representa el prototipus d'una classe a partir dels objectes que la integren, es proporcionen formes recurrents de trobar aquest representant, que preservin l'eficiència de l'algorisme. El procés de classificació, encara que va formant classes cada cop més nombroses, mai agrega directament $card(\mathcal{C})$ individus, sinó que ho fa de dos en dos, construint subclasses intermèdies (les que després constituiran la jerarquia).

Per tant, per a evitar el càlcul directe de centres de gravetat d'un nombre creixent de punts, s'ha procedit a adaptar les expressions **??** i **??** de forma que el centre de gravetat de $\mathcal{C}$ es calculi en funció dels centres de gravetat de les dues subclasses que s'agreguen.

Així doncs, en fusionar dues subclasses $\mathcal{C}_e,\ \mathcal{C}_d$ compostes d'$n_{\mathcal{C}_e}$ i $n_{\mathcal{C}_d}$ individus respectivament, i observant la forma en que prèviament s'havien calculat els centres de gravetat d'aquestes subclasses ($\bar{\imath}_{\mathcal{C}_e}$ i $\bar{\imath}_{\mathcal{C}_d}$), es pot demostrar que les components quantitatives del centre de gravetat $\bar{\imath}_{\mathcal{C}}$ de la nova classe $\mathcal{C} = \mathcal{C}_e \cup \mathcal{C}_d$ són

$$\bar{x}_{\mathcal{C}k} = \frac{n_{\mathcal{C}_e}\bar{x}_{\mathcal{C}_e k} + n_{\mathcal{C}_d}\bar{x}_{\mathcal{C}_d k}}{n_{c_e} + n_{c_d}}$$

La demostració és senzilla:

$$\bar{x}_{\mathcal{C}k} = \frac{\sum_{i\in\mathcal{C}} x_{ik}}{n_{\mathcal{C}}} = \frac{\sum_{i\in\mathcal{C}_e} x_{ik} + \sum_{i\in\mathcal{C}_d} x_{ik}}{n_{\mathcal{C}_e} + n_{\mathcal{C}_d}}$$

Com que

$$\bar{x}_{\mathcal{C}_e k} = \frac{\sum_{i\in\mathcal{C}_e} x_{ik}}{n_{\mathcal{C}_e}} \quad \text{i} \quad \bar{x}_{\mathcal{C}_d k} = \frac{\sum_{i\in\mathcal{C}_d} x_{ik}}{n_{\mathcal{C}_d}},$$

llavors

$$\bar{x}_{\mathcal{C}k} = \frac{n_{\mathcal{C}_e}\bar{x}_{\mathcal{C}_e k} + n_{\mathcal{C}_d}\bar{x}_{\mathcal{C}_d k}}{n_{\mathcal{C}_e} + n_{\mathcal{C}_d}} \tag{3.7}$$

■

El cas de les variables qualitatives és menys homogeni. Concentrant-nos en el valor d'X_k qualitativa,

- Si s'agreguen dues entitats que rauen en la mateixa modalitat de la variable, no cal construir la distribució marginal, ja que el centre de gravetat descansarà, en la seva totalitat, sobre aquesta mateixa categoria. És possible llavors mantenir la representació simbòlica

$$\bar{x}_{\mathcal{C}k} = c_s^k, \quad \text{si} \quad \bar{x}_{\mathcal{C}_e k} = \bar{x}_{\mathcal{C}_d k} = c_s^k$$

- Si els objectes no són de la mateixa categoria, però tenen valors compactes $c^k_{s_e}$ i $c^k_{s_d}$ respectivament, la distribució resultant únicament tindrà dues modalitats no nul·les, i les freqüències respectives seran:

$$\bar{x}_{\mathcal{C}k} = (f^{k_1}_{\mathcal{C}} \dots f^{k_{n_k}}_{\mathcal{C}}), \text{ amb } \; f^{k_j}_{\mathcal{C}} = \begin{cases} 0 & \forall j \neq s_e, s_d \\ \frac{n_{\mathcal{C}_e}}{n_{\mathcal{C}_e}+n_{\mathcal{C}_d}} & j = s_e \\ \frac{n_{\mathcal{C}_d}}{n_{\mathcal{C}_e}+n_{\mathcal{C}_d}} & j = s_d \end{cases}$$

- L'agregació d'un valor en forma extesa $\mathcal{C}_e$ (de valor $(f^{k_1}_{\mathcal{C}_e} \dots f^{k_{n_k}}_{\mathcal{C}_e})$) i un de compacte $\mathcal{C}_d$ (de valor c^k_s) resulta en

$$\bar{x}_{\mathcal{C}k} = (f^{k_1}_{\mathcal{C}} \dots f^{k_{n_k}}_{\mathcal{C}}), \text{ amb } \; f^{k_j}_{\mathcal{C}} = \begin{cases} \frac{n_{\mathcal{C}_e} f^{k_j}_{\mathcal{C}_e}}{n_{\mathcal{C}_e}+n_{\mathcal{C}_d}} & j \neq s \\ \frac{n_{\mathcal{C}_e} f^{k_j}_{\mathcal{C}_e}+n_{\mathcal{C}_d}}{n_{\mathcal{C}_e}+n_{\mathcal{C}_d}} & j = s \end{cases}$$

  cas que val la pena implementar, si es considera que serà relativament freqüent, i que estalvia un producte per a cada component i una suma per totes menys 1; més si es pensa que, en definitiva, aquest procés caldrà repetir-lo per a cada una de les variables categòriques que hi hagi.

- Per últim, el centre de gravetat de dos objectes extesos es calcula com:

$$\bar{x}_{\mathcal{C}k} = (f^{k_1}_{\mathcal{C}} \dots f^{k_{n_k}}_{\mathcal{C}}), \quad \text{amb} \quad f^{k_j}_{\mathcal{C}} = \frac{n_{\mathcal{C}_e} f^{k_j}_{\mathcal{C}_e} + n_{\mathcal{C}_d} f^{k_j}_{\mathcal{C}_d}}{n_{\mathcal{C}_e} + n_{\mathcal{C}_d}} \tag{3.8}$$

  Partint de la definició (**??**), $f^{k_j}_{\mathcal{C}} = \dfrac{I^{k_j}_{\mathcal{C}}}{n_{\mathcal{C}}}$.

  Com que $I^{k_j}_{\mathcal{C}} = I^{k_j}_{\mathcal{C}_e} + I^{k_j}_{\mathcal{C}_d}$

$$f^{k_j}_{\mathcal{C}} = \frac{I^{k_j}_{\mathcal{C}_e} + I^{k_j}_{\mathcal{C}_d}}{n_{\mathcal{C}}} = \frac{n_{\mathcal{C}_e} f^{k_j}_{\mathcal{C}_e} + n_{\mathcal{C}_d} f^{k_j}_{\mathcal{C}_d}}{n_{\mathcal{C}_e} + n_{\mathcal{C}_d}}$$

■

Aquestes recurrències fan que la complexitat en càlcul del centre de gravetat d'una classe no augmenti amb la grandària d'aquesta, sino que sempre, sigui quin sigui el cardinal de $\mathcal{C}$, és funció de dos operands. I això resulta especialment interessant en les darreres passes de l'algorisme, on les classes solen comptar amb un gran nombre d'individus.

La implementació d'aquest càlcul en **LISP** és directe si es disposa del nombre d'individus de cada classe; aquesta informació es té en una llista `l_nro_nodes`.

3.4 Les mètriques

En el nucli del classificador, la identificació de parelles de veïns recíprocs passa per la definició d'una distància — o, en el seu defecte, una mesura de (dis)similitud — sobre l'espai de les variables que descriuen als individus a classificar. Un cop definida l'estructura de la matriu de dades a §**??**, es pot passar al detall de com calcular la distància entre dos individus d'$\mathcal{I}$, siguin i i i'.

El problema sovint és difícil i la literatura ofereix nombroses propostes. Consulti's [**?**], [**?**] o [**?**] al respecte.

Dir, abans que res que, al llarg de la classificació, serà necessari calcular distàncies, no només entre individus, sino també entre classes, i que, donades dues classes $\mathcal{C}_e$ i $\mathcal{C}_d$, es defineix[11]

$$d^2(\mathcal{C}_e, \mathcal{C}_d) = d^2(\bar{\imath}_{\mathcal{C}_e}, \bar{\imath}_{\mathcal{C}_d})$$

En les properes seccions es tracta únicament la distància entre individus, entenent que la distància entre classes queda totalment determinada un cop definida la primera.

3.4.1 Només variables quantitatives

Tradicionalment s'ha vingut utilitzant la distància euclídia canònica per a mesurar distàncies entre individus descrits per un conjunt de variables quantitatives X_k, $(k = 1 : K)$ [**?**].

Donada una matriu de dades X com les definides a §**??**, la distància entre i i i' és, si totes les variables són quantitatives:

$$d^2(i, i') = \sum_{k=1}^{K} (x_{ik} - x_{i'k})^2 \tag{3.9}$$

Ara bé, quan les variables són d'ordres de magnitud molt diferents, l'efecte de cada una d'elles sobre la distància pot ser molt desigual. El concepte de proximitat entre individus depèn de les unitats de mesura de les variables. En principi, això no hauria de ser així, i solucions per a evitar-ho són normalitzar el càlcul de les distàncies dividint l'efecte de cada variable, bé per la seva desviació tipus empírica:

$$d^2(i, i') = \sum_{k=1}^{K} \left(\frac{x_{ik} - x_{i'k}}{s_k} \right)^2 \tag{3.10}$$

amb

$$\left(s_k^2 = \sum_{i=1}^{n} \frac{(x_{ik} - \bar{x}_k)^2}{n-1} \quad \text{i} \quad \bar{x}_k = \sum_{i=1}^{n} \frac{x_{ik}}{n} \right),$$

bé pel rang de la variable:

$$d^2(i, i') = \sum_{k=1}^{K} \left(\frac{(x_{ik} - x_{i'k})}{rang_k} \right)^2$$

[11] Evidentment, tot individu pot ser considerat com a cas particular de classe amb un únic element $\mathcal{C} = \{i\}$ i $\bar{\imath}_{\mathcal{C}} = i$, amb la qual cosa, si $\mathcal{C}_e$ i $\mathcal{C}_d$ són individus, $d^2(\mathcal{C}_e, \mathcal{C}_d) = d^2(i_e, i_d)$ com era d'esperar.

essent

$$(rang_k = \max_i\{x_{ik}\} - \min_i\{x_{ik}\},\ k = 1:K)$$

en funció de l'elecció de l'usuari.

3.4.2 Només variables qualitatives

Si les variables són totes categòriques, no es poden fer aquests càlculs ja que els elements de la matriu de dades són les diferents modalitats de la variable, de caràcter, per tant, alfanumèrics.

En estadística, s'acostumen a transformar aquest tipus de matrius de dades a la seva *forma disjuntiva completa* [**?**] descomposant cada variable qualitativa $X_k, (k = 1:K)$ en un paquet de variables binàries $\{Z_1^k, \ldots, Z_{n_k}^k\}$, on cada Z_j^k $(j = 1:K)$ representa la modalitat $c_j^k \in \mathcal{D}_k$ d'X_k. La taula **??** és la forma disjuntiva completa de la taula **??**, presentada a §**??**. Aquesta taula disjuntiva completa de dimensions $(n,\ \sum_{k=1}^K n_k)$ està formada, doncs, pels elements

$$Z = \left(z_{ij}^k = \begin{cases} 1, & x_{ik} = c_j^k \text{ en la matriu original de dades brutes} \\ 0, & \text{altrament} \end{cases} \right), \begin{array}{c} (i = 1:n) \\ (k = 1:K) \\ (j = 1:n_k) \end{array}$$

Per exemple, la matriu

$$\begin{array}{c} \\ i_1 \\ \vdots \\ i_n \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & \ldots & X_K \end{array} \\ \begin{pmatrix} x_{11} & \ldots & x_{1K} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{nK} \end{pmatrix} \end{array} \text{ amb } \begin{cases} x_{11} = c_2^1 \\ x_{1K} = c_2^K \\ \vdots \\ x_{n1} = c_{n_1}^1 \\ x_{nK} = c_{n_K}^n \end{cases}$$

correspon a la taula disjuntiva completa següent:

$$\begin{array}{c} \\ i_1 \\ \vdots \\ i_n \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1^1 & Z_2^1 & \ldots & Z_{n_1}^1 & \ldots & Z_1^K & \ldots & Z_{n_K}^K \end{array} \\ \begin{pmatrix} z_{11}^1 = 0 & z_{12}^1 = 1 & \ldots & 0 & 0 & 1\ldots & & 0 \\ & & & \vdots & & & & \\ 0 & \ldots & \ldots & 1 & 0 & \ldots & & z_{i_n n_k}^k = 1 \end{pmatrix} \end{array}$$

Sota aquesta representació, és la distància de χ^2 se sol utilitzar per a calcular distàncies entre individus [**?**]:

$$d^2(i, i') = \sum_{k=1}^K \sum_{j=1}^{n_k} \frac{\left(\frac{z_{ij}^k}{z_{i.}} - \frac{z_{i'j}^k}{z_{i'.}} \right)^2}{z_{.j}^k} \tag{3.11}$$

En aquesta expressió,

$$z_{i.} = \sum_{k=1}^K \sum_{j=1}^{n_k} z_{ij}^k,\ (i = 1:n),$$

Nom	*Alçada*	*Ros*	...	*Castany*	*Negre*	*Duu ulleres*	*Nom Cònjuge*
Joan	1.60m	1		0	0	1	?
Maria	1.65m	0		0	1	0	Jordi
Pep	1.70m	0		1	0	0	Rosa
⋮	⋮	⋮	⋮	⋮			

Table 3.2: Forma disjuntiva completa de la taula **??**.

i

$$z^k_{.j} = \sum_{i=1}^{n} z^k_{ij} = card\left(i : x_{ik} = c^k_j,\ (k = 1 : K), (j = 1 : n_k)\right)$$

representa el nombre d'individus de la mostra de modalitat c^k_j d'X_k.

Com es pot veure, quantes més modalitats tinguin les variables X_k, majors seran les dimensions de la matriu Z, que sempre és una matriu *esparsa*, donat que un individu i pertany a una única modalitat de cada X_k, i l'expansió d'X_k en $\{Z^k_1, \ldots, Z^k_{n_k}\}$ produeix subfiles de Z d'un sol element no nul:

$$\forall i, k\ (\exists j_0 = \{1 : n_k\}\ : z^k_{ij_0} = 1\ \&\ \forall j \neq j_0,\ z^k_{ij} = 0) \tag{3.12}$$

Aquest increment artificial de les dimensions de Z té com a conseqüència un encariment del procés de classificació, la complexitat del qual depèn de les dimensions de la matriu a tractar. El nostre objectiu és, per tant, evitar aquest desplegament de la matriu X. De la propietat **??** es dedueix que

$$z_{i.} = \sum_{k=1}^{K}\sum_{j=1}^{n_K} K z^k_{ij} = \sum_{k=1}^{K} 1 = K,\ \forall i \in \mathcal{I}$$

Així, l'expressió **??** es pot reescriure de la següent manera:

$$d^2(i,i') = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \frac{\left(\frac{z^k_{ij}}{z_{i.}} - \frac{z^k_{i'j}}{z_{i'.}}\right)^2}{z^k_{.j}} = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \frac{\left(\frac{z^k_{ij}}{K} - \frac{z^k_{i'j}}{K}\right)^2}{z^k_{.j}}$$

$$= \sum_{k=1}^{K}\sum_{j=1}^{n_k} \frac{\frac{(z^k_{ij} - z^k_{i'j})^2}{K^2}}{z^k_{.j}} = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{j=1}^{n_k} \frac{(z^k_{ij} - z^k_{i'j})^2}{z^k_{.j}}$$

i anomenant

$$d^2_k(i,i') = \sum_{j=1}^{n_k} \frac{(z^k_{ij} - z^k_{i'j})^2}{z^k_{.j}}$$

es pot escriure

$$d^2(i,i') = \frac{1}{K^2}\sum_{k=1}^{K} d^2_k(i,i')$$

$d_k^2(i,i')$ representa conjuntament els termes corresponents a les columnes $\{Z_1^k \dots Z_{n_k}^k\}$ que provenen de la descomposició de la variable qualitativa X_k en variables binàries.

La interpretació de $d_k^2(i,i')$ com la **contribució de la $k-$*èsima* variable** a la distància total entre i i i' és clara. Tot seguit s'analitza com calcular $d_k^2(i,i')$ directament sobre la matriu X.

Fent un estudi en una única variable i sense pèrdua de generalitat, cal distingir dos casos:

1. Si i i i' estan en la mateixa categoria es recodificarien, tradicionalment, en el mateix subvector, i per tant $d_k^2(i,i') = 0$.

2. En el cas més general, i i i' tindran l'1 en columnes diferents, siguin j, $j' \in (1 : n_k)$, i la resta de termes seran 0. Llavors
$$d_k^2(i,i') = \frac{(1-0)^2}{z_{.j}^k} + \frac{(0-1)^2}{z_{.j'}^k} = \frac{1}{z_{.j}^k} + \frac{1}{z_{.j'}^k}$$

Definim $I_{k^i} = card(\hat{\imath} : x_{\hat{\imath}k} = x_{ik})$, nombre d'individus de la mostra que, per la $k-$*èsima* variable, són de la mateixa modalitat que i i això permet un lleuger canvi de notació que posa de manisfest com aquest càlcul es pot efectuar *directament* sobre la matriu X, essent prescindible la construcció explícita de Z:

$$d_k^2(i,i') = \begin{cases} 0, & \text{if } x_{ik} = x_{i'k} \\ \dfrac{1}{I_{k^i}} + \dfrac{1}{I_{k^{i'}}}, & \text{en cas contrari} \end{cases} \tag{3.13}$$

Aquesta és la forma com **KLASS** calcula la distància entre dos individus descrits en un espai de variables qualitatives. Pel que fa a la implementació i cares a l'optimització dels càlculs, es manté una estructura amb les I_{k^i} calculades. S'ha vist a §**??** que el mecanisme d'agregació fa que es pugui perdre la representació simbòlica dels valors de les variables categòriques. Les classes es representen de forma prototípica pels seus centres de gravetat (veure §**??**), i els representants de classe no sempre tenen com a valors de les variables categòriques una única modalitat.

Per als objectes en forma extesa, es pot utilitzar la fòrmula general per a calcular distàncies

$$d_k^2(i,i') = \sum_{j=1}^{n_k} \frac{(f_i^{k_j} - f_{i'}^{k_j})^2}{I^{k_j}} \tag{3.14}$$

essent I^{k_j} el nombre d'individus de la mostra de valor c_j^k per la variable X_k (de fet, $I^{k_j} = z_{.j}^k$).

Si un dels dos objectes és compacte i l'altre extès, la distància es pot calcular com a (**??**) prèvia extensió del símbol corresponent. El pas de forma compacte a forma extesa d'una

PROVA	dades	distància	criteri	temps de **CPU** (s)			
				Carregat		Gastat	
				abans	després	abans	després
Compositors	Normalitz.	Euclídia	Centroide	20.33	17.83	1:41.73	36.48
Paisred	No Norm.	χ^2	Centroide i Ward	30.78	34.14	4:31.42	47.42
Paispet	No Norm.	Mixta	Centroide	17.98	19.43	38.32	28.98

Table 3.3: Taula de temps d'execució.

modalitat coincideix amb la recodificació que en estadística s'ha vingut fent tradicionalment [**?**], esmentada a l'inici d'aquest apartat.

Considerant però que el canvi de representació depèn directament del número de categories de la variable i que aquesta és una operació que es pot haver de repetir un nombre força elevat de vegades, val la pena de tractar directament aquest cas particular. Donats dos individus i en la forma extesa $x_{ik} = (f_i^{k_1}, \ldots, f_i^{k_{n_k}})$ i i' en forma compacta amb valor $x_{i'k} = c_s^k$,

$$d_k^2(i,i') = \frac{(f_i^{k_s} - 1)^2}{I^{k_s}} + \sum_{j \neq s}^{n_k} \frac{(f_i^{k_j})^2}{I^{k_j}}$$

Proves empíriques[12] mostren que la introducció d'aquesta modificació en l'algorisme redueix el temps de **CPU** considerablement. Efectivament, l'execució del programa sobre diversos bancs de dades amb o sense aquesta petita optimització genera resultats força diferents. La taula **??** ho mostra.

La primera de les proves fa referència a una matriu de vuit directors d'orquestra [**?**], descrits amb variables contínues totes elles. Les dues restants són extractes d'una matriu més gran que conté dues variables categòriques i dues de contínues descrivint la situació i tendència político-econòmica d'un nombre considerable de països de tot el món [**?**]. En la segona només s'estudien les variables categòriques, mentre que en la tercera es mantenen totes, però només es tracten els països del continent Americà.

En definitiva, la distància entre dos individus i, i', descrits únicament per variables categòriques, es calcula com:

[12] Realitzades amb un VAX-8600.

$$d_k^2(i,i') = \begin{cases} 0, & \text{si } x_{ik} = x_{i'k} \\ \dfrac{1}{I_{k^i}} + \dfrac{1}{I_{k^{i'}}}, & \text{altrament, si } i \text{ i } i' \text{ són compactes en } X_k \\ \dfrac{(f_i^{k_s}-1)^2}{I^{k_s}} + \displaystyle\sum_{j\neq s}^{n_k} \dfrac{(f_i^{k_j})^2}{I^{k_j}}, & \text{si } x_{ik} = c_s^k, \text{ i } i' \text{ és extès en } X_k \\ \displaystyle\sum_{j=1}^{n_k} \dfrac{(f_i^{k_j} - f_{i'}^{k_j})^2}{I^{k_j}}, & \text{per } i \text{ i } i' \text{ extesos en } X_k \end{cases} \tag{3.15}$$

3.4.3 Variables de tipus heterogenis

Si la matriu de dades X conté indistintament variables quantitatives i qualitatives, cal definir una nova distància entre individus. Hi ha, a la literatura, algunes propostes en aquest sentit, com per exemple [**?**] o [**?**], però, mentre la primera no permet treballar amb els representants de classe, tal com els entenem nosaltres i no es tracta d'una distància, sino d'un coeficient de similitud[13], la segona fa la hipòtesi que totes les variables qualitatives són ordinals, que ens sembla excessivament restrictiva.

Sigui $\mathcal{V} = (1 : K)$ el conjunt d'índexos de les variables X_k usades en la descripció dels individus, i $\delta_k^2(i,i')$ la distància entre i i i' en relació a la variable X_k. Les expressions **??** i **??** són casos particulars de

$$d^2(i,i') = \sum_{\forall k \in \mathcal{V}} \delta_k^2(i,i') \tag{3.16}$$

tals que $\delta_k^2(i,i')$ és la mateixa per a totes les variables.

En el cas general, $\mathcal{V}$ es pot particionar en ¡

$$\mathcal{Q} = \{k \in \mathcal{V} : X_k \textit{ és qualitativa}\}, \quad n_{\mathcal{Q}} = card(\mathcal{Q})$$

i

$$\mathcal{Ç} = \{k \in \mathcal{V} : X_k \textit{ és quantitativa}\}, \quad n_{\mathcal{Ç}} = card(\mathcal{Ç})$$

de forma que $\mathcal{V} = \mathcal{Q} \cup \mathcal{Ç}$, i llavors l'expressió **??** es pot generalitzar al cas en que X_k pugui ser indistintament quantitativa o qualitativa. En aquestes condicions, δ_k^2 es calcularà d'acord amb el tipus de cada variable.

$$\delta_k^2(i,i') = \begin{cases} \dfrac{(x_{ik} - x_{i'k})^2}{s_k^2}, & k \in \mathcal{Ç} \\ \dfrac{d_k^2(i,i')}{n_{\mathcal{Q}}^2}, & k \in \mathcal{Q}^{14} \end{cases}$$

[13] En les següents seccions es veurà com la mesura que definim en aquesta tesi, té estructura de distància sota certes condicions molt fàcils de satisfer, fins i tot quan la matriu de dades té estructura heterogènia.

Amb això, es pot reescriure l'expressió **??** com segueix:

$$d^2(i,i') = \sum_{\forall k \in \mathcal{Q}} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \frac{1}{n_{\mathcal{Q}}^2} \sum_{\forall k \in \mathcal{Q}} d_k^2(i,i') \tag{3.17}$$

No obstant, l'addició directa d'aquestes dues distàncies presenta un problema pel que a la magnitud de les variables numèriques fa referència. De fet, la distància de χ^2 es calcula en base a un vector fictici, i en el nostre cas implícit, de 0's i 1's, mentre que la component quantitativa de la distància, malgrat estar normalitzada, varia segons la unitat de mesura de les variables. Per a equilibrar la influència d'ambdós grups de variables, sembla raonable afectar-los d'un factor ponderador, α i β respectivament (amb $\alpha \geq 0, \beta \geq 0$), transformant l'expressió (**??**) en (**??**), a la que s'anomenarà *funció mixta* a partir d'ara.

$$d^2_{(\alpha,\beta)}(i,i') = \alpha \sum_{\forall k \in \mathcal{Q}} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \beta \frac{1}{n_{\mathcal{Q}}^2} \sum_{\forall k \in \mathcal{Q}} d_k^2(i,i') \tag{3.18}$$

En realitat, l'expressió (**??**) és una família de funcions indexada pel parell (α, β)

$$\{d^2_{(\alpha,\beta)}(i,i')\}_{(\alpha,\beta)\ \in(\Re^* \times \Re^* - \{(0,0)\})}, \quad (\Re^* = \Re^+ \cup \{0\}) \tag{3.19}$$

de la qual en formen part, evidentment, les distàncies **??** i **??**, que corresponen precisament a $(\alpha, \beta) = (1, 0)$ i $(\alpha, \beta) = (0, 1)$ respectivament.

3.4.4 La distància mixta

En aquest apartat es fa un estudi de les condicions que doten a l'expressió **??** d'estructura de distància. Tal com s'ha definit, la funció mixta es pot veure com la ponderació de dues components fonamentals i que faciliten la notació de les següents demostracions.

- Per una banda, la distància euclídia normalitzada calculada sobre la submatriu de totes les variables numèriques. A partir d'ara s'hi farà referència com a *subdistància quantitativa*.

$$d^2_{\mathcal{Q}}(i,i') = \sum_{\forall k \in \mathcal{Q}} \frac{(x_{ik} - x_{i'k})^2}{s_k^2}$$

- De l'altra, la distància de χ^2 calculada sobre la matriu de totes les variables de $\mathcal{Q}$, que es pot anomenar *subdistància qualitativa*.

$$d^2_{\mathcal{Q}}(i,i') = \frac{1}{n_{\mathcal{Q}}^2} \sum_{\forall k \in \mathcal{Q}} d_k^2(i,i')$$

[14] Quan totes les variables són qualitatives, el denominador d'aquesta expressió és K^2, ja que $n_{\mathcal{Q}} = K$.

La funció mixta entre dos individus i, i' es denotaria ara com:

$$d^2_{(\alpha,\beta)}(i,i') = \alpha d^2_{\mathcal{C}}(i,i') + \beta d^2_{\mathcal{Q}}(i,i') \tag{3.20}$$

Tota distància ha de ser una funció sempre positiva, que satisfaci les tres condicions següents:

i) Simetria: $d_{(\alpha,\beta)}(i,i') = d_{(\alpha,\beta)}(i',i),\ \forall i,\ i'$

ii) Identitat: $d_{(\alpha,\beta)}(i,i') = 0 \Longleftrightarrow i = i',\ \forall i,i'$

iii) Desigualtat triangular: $d_{(\alpha,\beta)}(i',i'') \leq d_{(\alpha,\beta)}(i',i) + d_{(\alpha,\beta)}(i,i''),\ \forall i$

Partint de la base que $d_{\mathcal{C}}$ i $d_{\mathcal{Q}}$ satisfan les condicions de distància sobre els espais $\{X_k : k \in \mathcal{C}\}$ i $\{X_k : k \in \mathcal{Q}\}$ respectivament, es pot demostrar que la funció mixta també ho és:

Simetria

Es vol demostrar que

$$d_{(\alpha,\beta)}(i,i') = d_{(\alpha,\beta)}(i',i),\ \forall i,\ i'$$

Partint de l'expressio **??** i donat que

$$d_{\mathcal{C}}(i,i') = d_{\mathcal{C}}(i',i),\ \forall i,i'$$

i que

$$d_{\mathcal{Q}}(i,i') = d_{\mathcal{Q}}(i',i),\ \forall i,i'$$

es té que

$$d_{(\alpha,\beta)}(i,i') = \sqrt{\alpha d^2_{\mathcal{C}}(i,i') + \beta d^2_{\mathcal{Q}}(i,i')} = \sqrt{\alpha d^2_{\mathcal{C}}(i',i) + \beta d^2_{\mathcal{Q}}(i',i)} = \sqrt{d^2_{(\alpha,\beta)}(i',i)}$$

■

Desigualtat triangular

En aquest cas es tracta de veure que

$$d_{(\alpha,\beta)}(i',i'') \leq d_{(\alpha,\beta)}(i',i) + d_{(\alpha,\beta)}(i,i''),\ \forall i$$

i es farà servir la propietat que tant $d^2_{\mathcal{C}}(i,i')$ com $d^2_{\mathcal{Q}}(i,i')$ satisfan la desigualtat triangular:

$$d_{(\alpha,\beta)}(i',i'') = \sqrt{\alpha d^2_{\mathcal{C}}(i',i'') + \beta d^2_{\mathcal{Q}}(i',i'')} \leq$$

$$\sqrt{\alpha\left(d^2_{\mathcal{C}}(i',i)+d^2_{\mathcal{C}}(i,i'')\right)+\beta d^2_{\mathcal{Q}}(i',i'')}\leq$$

$$\sqrt{\alpha d^2_{\mathcal{C}}(i',i)+\alpha d^2_{\mathcal{C}}(i,i'')+\beta\left(d^2_{\mathcal{Q}}(i',i)+d^2_{\mathcal{Q}}(i,i'')\right)}=$$

$$\sqrt{\alpha d^2_{\mathcal{C}}(i',i)+\alpha d^2_{\mathcal{C}}(i,i'')+\beta d^2_{\mathcal{Q}}(i',i)+\beta d^2_{\mathcal{Q}}(i,i'')}=$$

$$\sqrt{d^2_{(\alpha,\beta)}(i',i)+d^2_{(\alpha,\beta)}(i,i'')}\leq\sqrt{d^2_{(\alpha,\beta)}(i',i)}+\sqrt{d^2_{(\alpha,\beta)}(i,i'')}=$$

$$d_{(\alpha,\beta)}(i',i)+d_{(\alpha,\beta)}(i,i'')$$

■

Identitat

Cal veure que

$$d_{(\alpha,\beta)}(i,i')=0\Longleftrightarrow i=i'\ ,\ \forall i,i'$$

S'utilitza el fet que

$$d_{\mathcal{C}}(i,i')=0\Longleftrightarrow x_{ik}=x_{i'k},\ (k\in\mathcal{C})\ \ \forall i,i'$$

i

$$d_{\mathcal{Q}}(i,i')=0\Longleftrightarrow x_{ik}=x_{i'k}\ ,\ (k\in\mathcal{Q})\ \forall i,i'$$

propietats derivades del fet que les subdistàncies qualitativa i quantitativa ténen estructura de distància sobre els subvectors respectius.

($\Longleftarrow$): En primer lloc es demostra que

$$i=i'\Longrightarrow d_{(\alpha,\beta)}(i,i')=0,\ \forall i,i'$$

Suposant, doncs $i=i'$,

$$d_{(\alpha,\beta)}(i,i')=\sqrt{\alpha d^2_{\mathcal{C}}(i,i')+\beta d^2_{\mathcal{Q}}(i,i')}=\sqrt{\alpha 0+\beta 0}=0$$

■

($\Longrightarrow$): A continuació es demostra el recíproc:

$$d_{(\alpha,\beta)}(i,i')=0\Longrightarrow i=i',\ \forall i,i' \tag{3.21}$$

Suposar $d_{(\alpha,\beta)}(i,i')=0$ equival a dir que

$$\alpha d^2_{\mathcal{C}}(i,i')+\beta d^2_{\mathcal{Q}}(i,i')=0$$

Com que es té una suma de termes positius, s'ha de complir simultàniament que

$$\alpha d^2_{\mathcal{Q}}(i,i') = 0 \ \wedge \ \beta d^2_{\mathcal{Q}}(i,i') = 0$$

Fent una anàlisi per casos:

1. La suposició $d^2_{\mathcal{Q}}(i,i') = 0$ significa que totes les components quantitatives dels vectors x_i i x'_i coincideixen.

   En aquest cas, per tal que $d^2_{(\alpha,\beta)}(i,i') = 0$ cal que

   (a) o bé $d^2_{\mathcal{Q}}(i,i') = 0$, la qual cosa significa que totes les components qualitatives dels vectors x_i i x'_i també coincideixen i, per tant, $i = i'$ ∎

   (b) o bé $\beta = 0$.
   Sota aquestes hipòtesis $d^2_{(\alpha,\beta)}(i,i') = 0$ sigui quin sigui el valor de $d^2_{\mathcal{Q}}(i,i')$.
   Això significa que pot donar-se el cas $d^2_{\mathcal{Q}}(i,i') \neq 0$, la qual cosa implica que en algunes variables categòriques i i i' són diferents i, per tant, $i \neq i'$. En aquestes condicions, seria possible que objectes diferents tinguessin distància mixta nul·la i no es satifaria l'axioma d'identitat.
   Per tant, *condició necessària* per a que $d^2_{(\alpha,\beta)}(i,i')$ satisfaci l'axioma d'identitat és que

   $$\text{Si } \beta = 0 \Longrightarrow d^2_{\mathcal{Q}}(i,i') = 0 \ \forall i,i'$$

   ∎

   Però $d^2_{\mathcal{Q}}(i,i') = 0 \ \forall i,i'$ només és possible si

   i. $\mathcal{Q} = \emptyset$

   ii. Les variables de $\mathcal{Q}$ són no informatives (totes prenen un únic valor per a tots els individus de la matriu de dades).
   En el context en que treballem no serà així, ja que si una variable s'usa en la descripció dels individus, és perquè aporta alguna informació, i per tant, pren valors diferents en els diferents individus[15].
   En conseqüència, serà interessant imposar *directament* que

   $$\text{Si } \ \beta = 0 \Longrightarrow \mathcal{Q} = \emptyset$$

2. De forma totalment anàloga s'arribaria a la conclusió que cal també que

   $$\alpha = 0 \Longrightarrow d^2_{\mathcal{Q}}(i,i') = 0 \ \forall i,i'$$

   ∎

[15] Òbviament, si es presentés aquest cas, n'hi hauria prou amb eliminar les columnes d'X no informatives, i treballar amb la matriu de dades reduïda, aprofitant que el resultat de la classificació és invariant a la introducció o supressió de columnes constants de la matriu de dades.

per a que **??** satisfaci aquest axioma.

És a dir, que la condició necessària i suficient per a que **??** satisfaci l'axioma d'identitat, i sigui, per tant, una distància és que:

$$\alpha = 0 \Longrightarrow d^2_{\mathcal{C}}(i,i') = 0 \ \forall i,i' \quad \wedge \quad \beta = 0 \Longrightarrow d^2_{\mathcal{Q}}(i,i') = 0 \ \forall i,i'$$

En aquest cas, (**??**) es pot anomenar *distància mixta*, i (**??**) és una família de distàncies.

Tal com s'ha argumentat més amunt, de fet, la condició necessària i suficient a la pràctica serà:

$$\alpha = 0 \Longrightarrow \mathcal{C} = \emptyset \quad \wedge \quad \beta = 0 \Longrightarrow \mathcal{Q} = \emptyset \tag{3.22}$$

3.4.5 Equivalències entre distàncies mixtes

Des del punt de vista de la classificació, la informació proporcionada per alguns parells de distàncies mixtes $d^2_{(\alpha_1,\beta_1)}(i,i')$ i $d^2_{(\alpha_2,\beta_2)}(i,i')$ és equivalent, en el sentit que generen arbres de classificació iguals quant a estructura, excepte variacions en l'escala en que es representen els índexos de nivell. Efectivament, tota classificació ascendent jeràrquica ès invariant per reescalat de l'arbre que la representa (*i. e.* la successió $\mathcal{P}$ de particions associada serà la mateixa).

Utilitzant aquesta idea, es pot definir una relació d'equivalència sobre la família de distàncies $d^2_{(\alpha,\beta)}(i,i')$, de tal forma que distàncies que generin el *mateix* arbre ascendent jeràrquic pertanyin a la mateixa classe d'equivalència.

Això passa sempre que es produeixin les mateixes agregacions en el mateix ordre, o el que és el mateix, sempre que les distàncies relatives entre els individus es mantinguin. Per tant, tot parell de distàncies $d^2_{(\alpha_1,\beta_1)}(i,i')$ i $d^2_{(\alpha_2,\beta_2)}(i,i')$ tals que una pugui ser escrita com un reescalat de l'altra, seran de la mateixa classe d'equivalència.

Una distància $d^2_{(\alpha_1,\beta_1)}(i,i')$ és un reescalat d'una altra $d^2_{(\alpha_2,\beta_2)}(i,i')$ si existeix una constant real $\ell \in \Re$ que satisfà la relació:

$$d^2_{(\alpha_1,\beta_1)}(i,i') = \ell d^2_{(\alpha_2,\beta_2)}(i,i')$$

Fenomen que té lloc si i només si $\alpha_1 = \ell\alpha_2 \quad \wedge \quad \beta_1 = \ell\beta_2$
o, dit d'una altra manera, $\alpha_1\beta_2 = \alpha_2\beta_1$

La relació d'equivalència és, doncs: $d^2_{(\alpha_1,\beta_1)} \equiv d^2_{(\alpha_2,\beta_2)} \Longleftrightarrow \alpha_1\beta_2 = \alpha_2\beta_1$ (3.23)

Per veure que l'expressió **??** defineix efectivament una relaciò d'equivalència, cal demostrar que satisfà les propietats:

i) **Reflexiva** $d^2_{(\alpha,\beta)} \equiv d^2_{(\alpha,\beta)}, \forall \alpha, \beta$, evident $(\alpha\beta = \beta\alpha)$ ∎

ii) **Simètrica** $d^2_{(\alpha_1,\beta_1)} \equiv d^2_{(\alpha_2,\beta_2)} \Longrightarrow d^2_{(\alpha_2,\beta_2)} \equiv d^2_{(\alpha_1,\beta_1)}, \forall \alpha_1, \beta_1, \alpha_2, \beta_2$

L'expressió $d^2_{(\alpha_1,\beta_1)}(i,i') \equiv d^2_{(\alpha_2,\beta_2)}(i,i')$

indica que $\alpha_1\beta_2 = \alpha_2\beta_1$

que es pot escriure com $\alpha_2\beta_1 = \alpha_1\beta_2$

i que no és més que la definició de $d^2_{(\alpha_2,\beta_2)}(i,i') \equiv d^2_{(\alpha_1,\beta_1)}(i,i')$ ∎

iii) **Transitiva** $\left.\begin{array}{l} d^2_{(\alpha_1,\beta_1)} \equiv d^2_{(\alpha_2,\beta_2)} \\ d^2_{(\alpha_2,\beta_2)} \equiv d^2_{(\alpha_3,\beta_3)} \end{array}\right\} \Longrightarrow d^2_{(\alpha_1,\beta_1)} \equiv d^2_{(\alpha_3,\beta_3)},\ \forall \alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3$

De les expressions

$d^2_{(\alpha_1,\beta_1)}(i,i') \equiv d^2_{(\alpha_2,\beta_2)}(i,i') \quad \wedge \quad d^2_{(\alpha_2,\beta_2)}(i,i') \equiv d^2_{(\alpha_3,\beta_3)}(i,i')$

se'n dedueix respectivament que $\alpha_1\beta_2 = \alpha_2\beta_1 \quad \wedge \quad \alpha_2\beta_3 = \alpha_3\beta_2$

A) Suposant $\alpha_1 \neq 0$, $\alpha_2 \neq 0$ i aïllant β_2 i β_3

$$\beta_2 = \frac{\alpha_2\beta_1}{\alpha_1} \quad \wedge \quad \beta_3 = \frac{\alpha_3\beta_2}{\alpha_2}$$

Substituint la primera expressió en la segona,

$\beta_3 = \dfrac{\alpha_3}{\alpha_2}\dfrac{\alpha_2\beta_1}{\alpha_1} = \alpha_3\dfrac{\beta_1}{\alpha_1}$ és a dir $\beta_3\alpha_1 = \beta_1\alpha_3$

que és precisament la definició de $d^2_{(\alpha_1,\beta_1)}(i,i') \equiv d^2_{(\alpha_3,\beta_3)}(i,i')$ ∎

B) Si $\alpha_1 = 0$, llavors forçosament $\beta_1 > 0$ ja que

$(\alpha,\beta) \neq (0,0)\ \forall \alpha, \beta$ per construcció.

D'aquí, $\alpha_1\beta_2 = \alpha_2\beta_1$ només es pot donar si $\alpha_2 = 0$.

Necessàriament, doncs $\beta_2 > 0$, i com que $\alpha_2\beta_3 = \alpha_3\beta_2$

resulta que $\alpha_3 = 0$ també, i sempre, $\alpha_1\beta_3 = \alpha_3\beta_1$ ∎

Donada aquesta relació d'equivalència és possible treballar amb el seu conjunt quocient, ja que el conjunt dels arbres de classificació diferents generats per **KLASS** a partir d'aquest conjunt quocient coincideix amb el corresponent a la família completa $d^2_{(\alpha,\beta)}$. Com que α i β juguen el paper de ponderadors, té sentit elegir com a representant de cada classe d'equivalència la distància:

$$d^2_{(\alpha_0,\beta_0)}(i,i'), \text{ tal que } \alpha+\beta=1$$

Cal només demostrar la **unicitat** d'aquest representat, que es fa veient que si hi ha dos parells (α_1,β_1) i (α_2,β_2) de la mateixa classe d'equivalència, tals que satisfan simultàniament la condició de representant de classe, llavors $(\alpha_1,\beta_1)=(\alpha_2,\beta_2)$.

Suposant doncs dues distàncies equivalents $d^2_{(\alpha_1,\beta_1)}$ i $d^2_{(\alpha_2,\beta_2)}$ es té que:

$$\alpha_1\beta_2=\alpha_2\beta_1 \tag{3.24}$$

Per ser ambdues representants de classe satisfan que

$$\alpha_1+\beta_1=1 \quad \wedge \quad \alpha_2+\beta_2=1$$

és a dir

$$\alpha_1=1-\beta_1 \quad \wedge \quad \alpha_2=1-\beta_2$$

Substituint aquestes expressions a **??** resulta que

$$(1-\beta_1)\beta_2=(1-\beta_2)\beta_1 \Longleftrightarrow \beta_2-\beta_1\beta_2=\beta_1-\beta_2\beta_1 \Longleftrightarrow \beta_2=\beta_1$$

i en conseqüència $\alpha_2=\alpha_1$ també ∎

Treballar amb el conjunt quocient via la condició de representat de classe, correspon a utilitzar la següent família de distàncies:

$$\{d^2_{(\alpha_r,\beta_r)}(i,i')\}_{(\alpha_r,\beta_r)} \text{ és el representant de la classe } {}_{(\bar{\alpha},\bar{\beta})}$$

amb $(\bar{\alpha},\bar{\beta})$ elements del conjunt quocient $[0,1]\times[0,1]-\{(0,0)\}_{|\equiv}$. En altres paraultes, α_r i β_r satisfan la condició **??**.

3.5 Sobre els valors d'α i β

En aquesta secció s'introdueix un heurístic que permet trobar valors acceptables per a les constants de ponderació α i β. Els valors d'α i β representen la importància d'aquestes subdistàncies en la distància total, i estan destinats a equilibrar-la. Com ja s'ha dit, la magnitud de les subdistàncies quantitatives depèn de la unitat de mesura de la variable corresponent. La de les qualitatives, per la seva part, depèn del nombre de modalitats de cada una de les variables qualitatives.

L'equilibri es troba quan les dues subdistàncies estan referides a un interval comú, per exemple $[0,1]$. Una primera aproximació és dividir la subdistància quantitativa pel seu màxim:

$$d^2_{\mathcal{Q}max} = max_{i,i'}\{d^2_{\mathcal{Q}}(i,i')\}$$

i fer el mateix amb la qualitativa

$$d^2_{\mathcal{Q}max} = max_{i,i'}\{d^2_{\mathcal{Q}}(i,i')\}$$

Aquesta operació garanteix que les dues components tindran igual influència en la determinació de $d^2(i,i')$.

No obstant, per adquirir més robustesa, aquests màxims no es calculen pas directament sobre la totalitat de les subdistàncies, sino que prèviament s'elimina el 5% de subdistàncies extremes. Aquests màxims truncats es denotaran respectivament per $d^2_{\mathcal{Q}max^*}$ i $d^2_{\mathcal{Q}max^*}$. Llavors,

$$\alpha \propto \frac{1}{d^2_{\mathcal{Q}max^*}} \quad \& \quad \beta \propto \frac{1}{d^2_{\mathcal{Q}max^*}}$$

Amb això, si hi ha un valor aberrant o *outlier* a gran distància de la resta d'objectes, no es farà servir com a punt de referència, en el qual cas les altres distàncies serien molt petites, i estarien totes concentrades en un subinterval [0,c_0], $c_0 \ll 1$. És més, si no hi ha cap valor aberrant, les distàncies suprimides per al càlcul d'aquests màxims seran totes més o menys de la mateixa magnitud que $d^2_{\mathcal{Q}max^*}$ i $d^2_{\mathcal{Q}max^*}$ respectivament, i l'interval de treball real serà [0,c_0], amb $c_0 \approx 1$, que no suposa un canvi gaire significatiu pel que fa a la classificació.

Un cop trobada aquesta referència comú, sembla raonable donar més importància a la component qualitativa si els individus estan principalment descrits per variables qualitatives i viceversa. Així, definim les constants de ponderació de les subdistàncies proporcionalment al nombre de variables de cada tipus utilitzades en la descripció dels objectes. Fent $n_{\mathcal{Q}} = card(\mathcal{Q})$ i $n_{\mathcal{Q}} = card(\mathcal{Q})$, es proposen com a valors per α i β:

$$\alpha = \frac{n_{\mathcal{Q}}}{d^2_{\mathcal{Q}max^*}} \qquad \beta = \frac{n_{\mathcal{Q}}}{d^2_{\mathcal{Q}max^*}} \tag{3.25}$$

que és condició suficient per garantir que es treballa en un espai mètric.

En efecte, α i β definides d'aquesta manera,

$$\alpha = 0 \implies n_{\mathcal{Q}} = 0 \implies \mathcal{Q} = \emptyset$$

$$\beta = 0 \implies n_{\mathcal{Q}} = 0 \implies \mathcal{Q} = \emptyset$$

■

Sigui (α_r, β_r) el representant d'aquesta classe d'equivalència. A la secció anterior s'ha vist que la condició d'equivalència és:

$$d^2_{(\alpha_r,\beta_r)}(i,i') \equiv d^2_{(\alpha,\beta)}(i,i') \Longleftrightarrow \frac{\alpha_r}{\beta_r} = \frac{\alpha}{\beta}$$

El representant de la classe compleix, a més, la condició que:

$$\alpha_r + \beta_r = 1$$

Combinant ambdues relacions

$$\frac{\alpha}{\beta}\beta_r + \beta_r = 1 \quad ; \quad \left(\frac{\alpha}{\beta} + 1\right)\beta_r = 1 \ ;$$

$$\left(\frac{\alpha+\beta}{\beta}\right)\beta_r = 1 \quad ; \quad \beta_r = \frac{\beta}{\alpha+\beta}$$

De forma anàloga es troba el valor d'α_r. Així doncs, la proposta definitiva pels valors α i β és el representant d'(α, β) en el conjunt quocient:

$$\alpha_r = \frac{\alpha}{\alpha+\beta} \qquad \beta_r = \frac{\beta}{\alpha+\beta} = 1 - \alpha_r \tag{3.26}$$

on α i β queden definits en l'expessió (**??**). Per la forma com s'han construït, α_r i β_r també donen lloc a una distància, satisfent la condicio **??**.

3.6 Criteris d'agregació

El criteri d'agregació és el que seleccionarà en cada pas de la classificació el parell de punts (d'individus o subclasses) que s'han de fusionar en una nova classe. Vista l'estructura de l'algorisme dels veïns recíprocs encadenats, en que sempre s'agreguen parells de veïns recíprocs, i es manega el concepte de màxima proximitat entre objectes, es tractarà de definir com es determina l'objecte més proper d'un donat.

D'entre tots els criteris d'agregació, se n'han tractat dos: El criteri del centroide (del baricentre) que té una base geomètrica molt propera al que fan alguns sistemes basats en el coneixement, com **Linneo**, i el de Ward, que descansa sobre el concepte d'informació. No obstant, com que el criteri d'agregació està parametritzat en tot l'algorisme, se'n poden implementar d'altres amb relativa facilitat.

3.6.1 El criteri del centroide

Segons el criteri del centroide [**?**], el resultat de fusionar dos individus és un nou individu que té com a coordenades les del centre de gravetat dels seus components, que es pot interpretar

com a la caracterització promig de la classe en conjunt, i que no té perquè coincidir amb cap element real.

A la secció **??** ja s'ha presentat amb detall com es calcula el baricentre, o centre de gravetat, d'un conjunt de punts. Per aquest cas, doncs, l'algorisme dels veïns recíprocs encadenats té una implementació directa.

L'element i' més proper d'un donat i és aquell que es troba a mínima distància d'ell en la mètrica donada, i és indiferent que i i i' siguin objectes elementals o representants de classe. En una passa intermèdia de l'algorisme es té una certa partició del conjunt d'individus $\mathcal{P}_\xi = \{\mathcal{C}_1 \dots \mathcal{C}_{n-\xi+1}\},\ (\xi = 1 : n)$, i es forma una nova classe amb els nodes $\mathcal{C}_e$ i $\mathcal{C}_d \in \mathcal{P}_\xi$ que satisfan la condició d'agregació:

$$d^2(\mathcal{C}_e, \mathcal{C}_d) = \min_{\substack{\mathcal{C},\mathcal{C}' \in \mathcal{P}_\xi \\ \mathcal{C} \neq \mathcal{C}'}} d^2(\mathcal{C}, \mathcal{C}') \text{ o el que és equivalent: } d(\mathcal{C}_e, \mathcal{C}_d) = \min_{\substack{\mathcal{C},\mathcal{C}' \in \mathcal{P}_\xi \\ \mathcal{C} \neq \mathcal{C}'}} d(\mathcal{C}, \mathcal{C}') \qquad (3.27)$$

Es tracta, per tant, de mantenir una matriu de distàncies entre individus

$$\mathcal{D} = \begin{pmatrix} & & \\ & d_{ed} & \text{, amb } d_{ed} = d(\mathcal{C}_e, \mathcal{C}_d) \\ & & \end{pmatrix}$$

i identificar-ne el mínim a cada passa, realitzant, evidentment, les actualitzacions pertinents després de cada agregació.

Succeeix, però, que ni la mètrica euclídia, ni la de χ^2, i en conseqüència tampoc la distància mixta, satisfan l'axioma de reductibilitat, i són, per tant, mètriques però no ultramètriques. Això fa que puguin aparèixer inversions en els arbres ascendents jeràrquics, i significa que l'espai podria col·lapsar amb les successives agregacions del procés de classificació. Per aquesta raó, no s'utilitza gaire aquest criteri.

3.6.2 El criteri de Ward

Previs

El criteri de Ward [**?**], a diferència de l'anterior, no decideix la fusió de dues classes basant-se en una noció de distància, sinó en el concepte d'inèrcia. La inèrcia total d'un conjunt de punts $\mathcal{I}$ respecte del seu centre de gravetat $\bar{\imath}$ és constant:

$$M^2(\mathcal{I}/\bar{\imath}) = \sum_{i \in \mathcal{I}} m_i \, d^2(i, \bar{\imath}) \qquad (3.28)$$

essent m_i la massa del punt i, o efectiu de la subclasse que representa.

Donada una partició qualsevol d'$\mathcal{I}$ en ξ classes, sigui $\mathcal{P} = \{\mathcal{C}_1 \dots \mathcal{C}_\xi\}$, aquesta inèrcia es pot descomposar en dos termes independents:

- La inèrcia *intra-classes*, corresponent a l'acumulació de les inèrcies internes a cada una de les classes:

$$Mi^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath}) = \sum_{\mathcal{C}\in\mathcal{P}} M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) = \sum_{\mathcal{C}\in\mathcal{P}}\sum_{i\in\mathcal{C}} m_i\, d^2(i,\bar{\imath}_{\mathcal{C}}) \tag{3.29}$$

- La inèrcia *inter classes* (entre classes), corresponent a la inèrcia del núvol format pels representants de les classes, afectats de la massa corresponent:

$$Me^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath}) = \sum_{\mathcal{C}\in\mathcal{P}} m_{\mathcal{C}}\, d^2(\bar{\imath}_{\mathcal{C}},\bar{\imath}) = \sum_{\mathcal{C}\in\mathcal{P}} m_{\mathcal{C}}\, d^2(\mathcal{C},\bar{\imath}), \quad m_{\mathcal{C}} = card(\mathcal{C}) \tag{3.30}$$

La inèrcia d'un conjunt de punts està directament relacionada amb la dispersió d'aquest conjunt respecte del seu centre de gravetat. Quan menor sigui la inèrcia d'un núvol de punts, més apinyats estaran els seus integrants al voltant del seu centre de gravetat, i per tant, més homogènia serà la classe, perquè els elements seran molt semblants. Pel contrari, a major inèrcia, major heterogeneïtat del núvol de punts.

Ja s'ha dit que l'objectiu de la classificació jeràrquica és identificar sobre $\mathcal{I}$ classes diferents entre si, compostes d'individus homogenis. Resulta, doncs, que fer màxima $Me^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath})$ significa trobar classes molt distanciades entre elles, en altres paraules, grups clarament diferenciats; mentre que fer mínima la $Mi^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath})$ significa trobar classes d'individus molt semblants, molt homogenis. Com que $M^2(\mathcal{I}/\bar{\imath})$ és constant[16] per qualsevol partició d'$\mathcal{I}$ i

$$M^2(\mathcal{I}/\bar{\imath}) = Mi^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath}) + Me^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath}), \quad \forall\mathcal{P}, \quad \text{partició d'}\mathcal{I}$$

l'objectiu de la classificació es podria expressar com trobar la partició

$$\varrho \;\; tq \;\; Me^2_{\varrho}(\mathcal{I}/\bar{\imath}) = \max_{\mathcal{P}}\{Me^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath})\} \Longleftrightarrow Mi^2_{\varrho}(\mathcal{I}/\bar{\imath}) = \min_{\mathcal{P}}\{Mi^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath})\}$$

Davant la impossibilitat, per explosió combinatòria (veure §**??**), de resoldre amb exactitud aquest problema d'optimització, Ward proporciona un heurístic que porta, a baix cost, a un òptim local a cada passa de l'algorisme.

Reprenent la caracterització de la classificació ascendent jeràrquica donada a **??**, la partició inicial d'$\mathcal{I}$ és $\mathcal{P}_1 = \{\{i_1\},\ldots,\{i_n\}\}$. Per tant, a l'inici de l'algorisme, la descomposició de la inèrcia total és:

$$M^2(\mathcal{I}/\bar{\imath}) = Mi^2_{\mathcal{P}_1}(\mathcal{I}/\bar{\imath}) + Me^2_{\mathcal{P}_1}(\mathcal{I}/\bar{\imath}) = 0 + M^2(\mathcal{I}/\bar{\imath})$$

i la inèrcia intra-classes, com que no s'ha format cap classe encara, és 0.

Al final del procés, la partició és $\mathcal{P}_n = \{\{i_1,\ldots,i_n\}\}$, i la inèrcia es descomposa com:

$$M^2(\mathcal{I}/\bar{\imath}) = Mi^2_{\mathcal{P}_n}(\mathcal{I}/\bar{\imath}) + Me^2_{\mathcal{P}_n}(\mathcal{I}/\bar{\imath}) = M^2(\mathcal{I}/\bar{\imath}) + 0$$

[16] Queda fixada per la matriu de dades.

on la inèrcia inter classes, com que només n'hi ha una, és nul·la.

Es presenta a continuació, la descomposició que la inèrcia total experimenta en cada passa del procés ascendent jeràrquic:

Suposi's la passa ξ en que s'agreguen els elements $\mathcal{C}_e, \mathcal{C}_d \in \mathcal{P}_\xi$. Llavors,

$$\mathcal{P}_{\xi+1} = \mathcal{P}_\xi \cup \{\mathcal{C}_e \cup \mathcal{C}_d\} - \{\mathcal{C}_e, \mathcal{C}_d\}$$

La inèrcia intra-classes de $\mathcal{P}_{\xi+1}$ serà

$$Mi^2_{\mathcal{P}_{\xi+1}}(\mathcal{I}/\bar{\imath}) = Mi^2_{\mathcal{P}_\xi}(\mathcal{I}/\bar{\imath}) + M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) - M^2(\mathcal{C}_e/\bar{\imath}_{\mathcal{C}_e}) - M^2(\mathcal{C}_d/\bar{\imath}_{\mathcal{C}_d})$$

Per aplicació del Teorema de Huygens,

$$M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) = M^2(\mathcal{C}_e/\bar{\imath}_{\mathcal{C}_e}) + M^2(\mathcal{C}_d/\bar{\imath}_{\mathcal{C}_d}) + m_{\mathcal{C}_e} d^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) + m_{\mathcal{C}_d} d^2(\mathcal{C}_d, \bar{\imath}_{\mathcal{C}})$$

D'aquí que

$$Mi^2_{\mathcal{P}_{\xi+1}}(\mathcal{I}/\bar{\imath}) = Mi^2_{\mathcal{P}_\xi}(\mathcal{I}/\bar{\imath}) + m_{\mathcal{C}_e} d^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) + m_{\mathcal{C}_d} d^2(\mathcal{C}_d, \bar{\imath}_{\mathcal{C}})$$

Es pot assegurar que l'augment de la inèrcia intra-classe corresponent a l'agregació del pas ξ satisfà sempre,

$$\Delta_\xi = Mi^2_{\mathcal{P}_{\xi+1}}(\mathcal{I}/\bar{\imath}) - Mi^2_{\mathcal{P}_\xi}(\mathcal{I}/\bar{\imath}) = m_{\mathcal{C}_e} d^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) + m_{\mathcal{C}_d} d^2(\mathcal{C}_d, \bar{\imath}_{\mathcal{C}}) > 0 \tag{3.31}$$

En efecte, la creació o engrandiment d'una classe suposa afegir inèrcia a l'existent, i la inèrcia sempre és positiva[17]. Considerant que Δ_ξ és deguda a l'agregació de $\mathcal{C}_e$ i $\mathcal{C}_d$, es denotarà també com $\Delta(\mathcal{C}_e, \mathcal{C}_d)$, i s'interpreta com l'augment d'inèrcia que es produïria en fusionar-se $\mathcal{C}_e$ i $\mathcal{C}_d$ en la nova classe $\mathcal{C}$.

La proposta de Ward, d'agregar a cada passa aquell parell de punts que generin el Δ_ξ més petit possible, significa preservar al màxim la homogeneïtat de les classes formades, que és precisament un dels objectius pretesos, i equivalent també a preservar al màxim l'heterogeneïtat entre les classes (que és l'altre objectiu enunciat ja al capítol **??**).

En aquest cas, la condició d'agregació és triar $\mathcal{C}_e$ i $\mathcal{C}_d \in \mathcal{P}_\xi$ (siguin subclasses o individus) de tal forma que:

$$\Delta(\mathcal{C}_e, \mathcal{C}_d) = \min_{\substack{\mathcal{C},\mathcal{C}' \in \mathcal{P}_\xi \\ \mathcal{C} \neq \mathcal{C}'}} \Delta(\mathcal{C}, \mathcal{C}') \tag{3.32}$$

A cada pas s'agregaran les dues classes més homogènies (o *properes*), i això ens situa molt a prop del que s'ha vingut fent fins ara. L'expressió **??** és equiparable a **??**, la qual cosa permet considerar que l'augment d'inèrcia juga el paper d'una pseudodistància entre classes.

[17] Observi's que és funció de masses i distàncies, termes sempre positius.

Des del punt de vista dels **veïns recíprocs** doncs, no cal més que substituir la matriu de distàncies entre classes $\mathcal{D}$ per una matriu on s'hi recull l'augment d'inèrcia que es produiria si s'efectués la fusió de cada dues classes, i operar en la forma ordinària. Sota el criteri de Ward, doncs, $\mathcal{D} = (d_{ed})$, amb $d_{ed} = \Delta(\mathcal{C}_e, \mathcal{C}_d)$.

La matriu inicial es construeix considerant que cada punt representa una classe d'un únic individu[18] — i per tant, de massa 1 —. L'element ed de la matriu representa l'augment de dispersió que es tindria si s'efectués la fusió dels objectes $\mathcal{C}_e$ i $\mathcal{C}_d$. Com que a $\mathcal{P}_1$ la inèrcia intra-classes és nul·la, l'augment d'inèrcia a la primera agregació serà la inèrcia interna de la classe que es formi. Per tant, aplicant la fòrmula (**??**) per la fusió $\mathcal{C} = \{\mathcal{C}_e, \mathcal{C}_d\}$ i massa unitària es té:

$$d_{ed} = M^2(\mathcal{C}, \bar{\imath}_{\mathcal{C}}) = d^2(i, \bar{\imath}_{\mathcal{C}}) + d^2(i', \bar{\imath}_{\mathcal{C}}) \tag{3.33}$$

Desenvolupant aquesta expressió s'arriba a la inicialització de la matriu de distàncies $\mathcal{D}$

$$\mathcal{D} = (d_{ed}) \text{ , amb } d_{ed} = \frac{1}{2}\, d^2(\mathcal{C}_e, \mathcal{C}_d)$$

En un primer pas, s'agreguen els punt i, i' que tinguin la menor pseudodistància entre si. En les passes successives, s'identificarà la nova agregació després d'actualitzar la matriu $\mathcal{D}$ en la forma convenient.

Desenvolupament de la recurrència

S'ha vist com en una passa qualsevol ξ, la hipotètica agregació de dos objectes $\mathcal{C}_e$ i $\mathcal{C}_d$ en una classe $\mathcal{C}$, bé siguin individus o subclasses, generaria l'augment de dispersió:

$$d_{ed} = \Delta(\mathcal{C}_e, \mathcal{C}_d) = m_e d^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) + m_d d^2(\mathcal{C}_d, \bar{\imath}_{\mathcal{C}}) \tag{3.34}$$

que es pot posar en funció dels centres de gravetat de cada classe i els seus efectius únicament:

$$d_{ed} = \frac{m_e m_d}{m_e + m_d} d^2(\mathcal{C}_e, \mathcal{C}_d) \tag{3.35}$$

Demostració : El càlcul[19,20] de $d^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}})$:

$$d^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) = \sum_{k=1}^{K} \delta_k^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) = \alpha \sum_{k \in \mathcal{Q}} \left(\frac{\bar{x}_{\mathcal{C}_e k} - \bar{x}_{\mathcal{C} k}}{s_k} \right)^2 + \beta \sum_{k \in \mathcal{Q}} \frac{d_k^2(\mathcal{C}_e, \mathcal{C})}{n_{\mathcal{Q}}^2} \tag{3.36}$$

Anomenant $m_e = card(\mathcal{C}_e)$, $m_d = card(\mathcal{C}_d)$ i $m_{\mathcal{C}} = card(\mathcal{C})$, en el cas general, la $k-$*èsima* component dels centres de gravetat de cada classe és de la forma:

$$\bar{x}_{\mathcal{C}_e k} = \frac{\sum_{i \in \mathcal{C}_e} x_{ik}}{m_e} \quad \text{i} \quad \bar{x}_{\mathcal{C}_d k} = \frac{\sum_{i \in \mathcal{C}_d} x_{ik}}{m_d}, \quad k \in \mathcal{Q}$$

[18] Aquesta condició serà reconsiderada en el capítol **??**.

[19] Per facilitar la lectura de la demostració, s'utilitza la propietat que tenen tots els valors compactes (§**??**) de ser representats en forma extesa, i es treballarà, únicament, amb valors d'aquest darrer tipus.

[20] Noti's que en la definició de $d^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}})$ s'hi inclouen ja els paràmetres α i β.

(3.37)

$$\bar{x}_{\mathcal{C}_e k} = (f_{\mathcal{C}_e}^{k_1} \dots f_{\mathcal{C}_e}^{k_{n_k}}),\ f_{\mathcal{C}_e}^{k_j} = \frac{I_{\mathcal{C}_e}^{k_j}}{m_e} \ \text{ i } \ \bar{x}_{\mathcal{C}_d k} = (f_{\mathcal{C}_d}^{k_1} \dots f_{\mathcal{C}_d}^{k_{n_k}}),\ f_{\mathcal{C}_d}^{k_j} = \frac{I_{\mathcal{C}_d}^{k_j}}{m_d},\quad k \in \mathcal{Q}$$

i la del centre de gravetat de la classe englobant seria

$$\bar{x}_{\mathcal{C}k} = \frac{\sum_{i\in\mathcal{C}} x_{ik}}{m_{\mathcal{C}}} = \frac{\sum_{i\in\mathcal{C}_e} x_{ik} + \sum_{i\in\mathcal{C}_d} x_{ik}}{m_e + m_d},\quad k \in \mathcal{Q}$$

(3.38)

$$\bar{x}_{\mathcal{C}k} = (f_{\mathcal{C}}^{k_1} \dots f_{\mathcal{C}}^{k_{n_k}}),\ \text{ on } f_{\mathcal{C}}^{k_j} = \frac{I_{\mathcal{C}}^{k_j}}{m_{\mathcal{C}}} = \frac{I_{\mathcal{C}_e}^{k_j} + I_{\mathcal{C}_d}^{k_j}}{m_e + m_d},\quad k \in \mathcal{Q}$$

- En primer lloc s'estudia el càlcul de $\delta_k^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}})$, per una variable numèrica:

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) = \alpha \left(\frac{\bar{x}_{\mathcal{C}_e k} - \bar{x}_{\mathcal{C}k}}{s_k}\right)^2 = \frac{\alpha}{s_k^2}\left(\frac{\sum_{i\in\mathcal{C}_e} x_{ik}}{m_e} - \frac{\sum_{i\in\mathcal{C}_e} x_{ik} + \sum_{i\in\mathcal{C}_d} x_{ik}}{m_e + m_d}\right)^2$$

fent denominador comú i desenvolupant el producte:

$$\begin{aligned}\delta_k^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) &= \frac{\alpha}{s_k^2}\left(\frac{1}{m_e(m_e+m_d)}\ (\ (m_e+m_d)\textstyle\sum_{i\in\mathcal{C}_e} x_{ik} + \right.\\ &\qquad \left. - m_e\left(\textstyle\sum_{i\in\mathcal{C}_e} x_{ik} + \sum_{i\in\mathcal{C}_d} x_{ik}\right)\right)^2\\ &= \frac{\alpha}{s_k^2}\left(\frac{1}{m_e(m_e+m_d)}\ (\ m_e\textstyle\sum_{i\in\mathcal{C}_e} x_{ik} - m_e\sum_{i\in\mathcal{C}_e} x_{ik} + \right.\\ &\qquad \left. + m_d\textstyle\sum_{i\in\mathcal{C}_e} x_{ik} - m_e\sum_{i\in\mathcal{C}_d} x_{ik}\right)\right)^2\end{aligned}$$

anul·lant termes, utilitzant **??**, i factoritzant:

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) = \frac{\alpha}{s_k^2}\left(\frac{m_e m_d}{m_e(m_e+m_d)}(\bar{x}_{\mathcal{C}_e k} - \bar{x}_{\mathcal{C}_d k})\right)^2$$

és a dir,

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) = \frac{(m_e m_d)^2\,\delta_k^2(\mathcal{C}_e, \mathcal{C}_d)}{(m_e\,(m_e+m_d))^2}$$

ja que la distància entre els centres de les dues classes per la component k és:

$$\delta_k^2(\mathcal{C}_e, \mathcal{C}_d) \ = \ \alpha\left(\frac{\bar{x}_{\mathcal{C}_e k} - \bar{x}_{\mathcal{C}_d k}}{s_k}\right)^2$$

Així doncs, per X_k quantitativa,

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_{\mathcal{C}}) = \frac{m_d^2\,\delta_k^2(\mathcal{C}_e, \mathcal{C}_d)}{(m_e+m_d)^2} \text{ i anàlogament } \delta_k^2(\mathcal{C}_d, \bar{\imath}_{\mathcal{C}}) = \frac{m_e^2\,\delta_k^2(\mathcal{C}_e, \mathcal{C}_d)}{(m_e+m_d)^2} \tag{3.39}$$

■

- El càlcul de $\delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C})$, per una variable categòrica es desenvolupa d'una forma molt semblant. Partint de la definició i utilitzant les expressions **??** i **??**:

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \frac{\beta}{n_\mathcal{Q}^2} \sum_{j=1}^{n_k} \frac{(f_{\mathcal{C}_e}^{k_j} - f_\mathcal{C}^{k_j})^2}{I^{k_j}}$$

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \frac{\beta}{n_\mathcal{Q}^2} \sum_{j=1}^{n_k} \frac{1}{I^{k_j}} \left(\frac{I_{\mathcal{C}_e}^{k_j}}{m_e} - \frac{I_{\mathcal{C}_e}^{k_j} + I_{\mathcal{C}_d}^{k_j}}{m_e + m_d} \right)^2$$

fent denominador comú i desenvolupant el producte

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \frac{\beta}{n_\mathcal{Q}^2} \sum_{j=1}^{n_k} \frac{1}{I^{k_j} m_e^2 (m_e + m_d)^2} (m_e I_{\mathcal{C}_e}^{k_j} + m_d I_{\mathcal{C}_e}^{k_j} - m_e I_{\mathcal{C}_e}^{k_j} - m_e I_{\mathcal{C}_d}^{k_j})^2$$

Anul·lant termes

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \frac{\beta}{n_\mathcal{Q}^2} \sum_{j=1}^{n_k} \frac{1}{I^{k_j} m_e^2 (m_e + m_d)^2} (m_d I_{\mathcal{C}_e}^{k_j} - m_e I_{\mathcal{C}_d}^{k_j})^2$$

usant **??**

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \frac{\beta}{n_\mathcal{Q}^2} \sum_{j=1}^{n_k} \frac{1}{I^{k_j} m_e^2 (m_e + m_d)^2} (m_d m_e f_{\mathcal{C}_e}^{k_j} - m_e m_d f_{\mathcal{C}_d}^{k_j})^2$$

i fent factor comú

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \frac{\beta}{n_\mathcal{Q}^2} \sum_{j=1}^{n_k} \frac{(m_e m_d)^2 (f_{\mathcal{C}_e}^{k_j} - f_{\mathcal{C}_d}^{k_j})^2}{I^{k_j}\ m_e^2 (m_e + m_d)^2} = \frac{\beta}{n_\mathcal{Q}^2}\ \frac{m_d^2}{(m_e + m_d)^2} \sum_{j=1}^{n_k} \frac{(f_{\mathcal{C}_e}^{k_j} - f_{\mathcal{C}_d}^{k_j})^2}{I^{k_j}}$$

$$\delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \frac{\beta}{n_\mathcal{Q}^2} \frac{m_d^2}{(m_e + m_d)^2}\ d_k^2(\mathcal{C}_e, \mathcal{C}_d) = \frac{m_d^2}{(m_e + m_d)^2} \delta_k^2(\mathcal{C}_e, \mathcal{C}_d)$$

■

Igualment es deduïria que

$$\delta_k^2(\mathcal{C}_d, \bar{\imath}_\mathcal{C}) = \frac{m_e^2}{(m_e + m_d)^2}\ \delta_k^2(\mathcal{C}_e, \mathcal{C}_d)$$

■

que és el *mateix* resultat obtingut a **??** per les variables numèriques.

En definitiva,

$$\begin{aligned} d^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \ & \textstyle\sum_{k=1}^K \delta_k^2(\mathcal{C}_e, \bar{\imath}_\mathcal{C}) = \sum_{k=1}^K \frac{m_d^2}{(m_e+m_d)^2} \delta_k(\mathcal{C}_e, \mathcal{C}_d) = \\ & \frac{m_d^2}{(m_e+m_d)^2} \textstyle\sum_{k=1}^K \delta_k(\mathcal{C}_e, \mathcal{C}_d) = \frac{m_e^2}{(m_e+m_d)^2} d^2(\mathcal{C}_e, \mathcal{C}_d) \end{aligned} \tag{3.40}$$

i de forma anàloga $d^2(\mathcal{C}_d, \bar{\imath}_{\mathcal{C}}) = \dfrac{m_e^2}{(m_e+m_d)^2} d^2(\mathcal{C}_e, \mathcal{C}_d)$

Substituint a (**??**) les expressions (**??**)

$$\begin{aligned}\Delta(\mathcal{C}_e, \mathcal{C}_d) &= m_e \frac{m_d^2\, d^2(\mathcal{C}_e, \mathcal{C}_d)}{(m_e+m_d)^2} + m_d \frac{m_e^2\, d^2(\mathcal{C}_e, \mathcal{C}_d)}{(m_e+m_d)^2} \\ &= \frac{d^2(\mathcal{C}_e, \mathcal{C}_d)}{(m_e+m_d)^2}\,(m_e m_d^2 + m_d m_e^2) \\ &= \frac{d^2(\mathcal{C}_e, \mathcal{C}_d)}{(m_e+m_d)^2}\, m_e m_d (m_d + m_e)\end{aligned}$$

d'on, finalment, s'obté

$$d_{ed} = \Delta(\mathcal{C}_e, \mathcal{C}_d) = \frac{m_e m_d}{(m_e+m_d)}\, d^2(\mathcal{C}_e, \mathcal{C}_d) \tag{3.41}$$

■

Ara bé, amb això, es necessita la distància entre les dues classes que s'agreguen per a calcular la nova matriu, la qual cosa implica, o bé mantenir una estructura auxiliar amb les distàncies entre individus, de dimensions considerables, o bé guardar aquestes distàncies i modificar la forma de calcular l'element més proper d'un donat, que ja no és el mínim de la corresponent fila de la matriu sinó que cal fer els càlculs de l'expressió (**??**). El que realment seria útil des del punt de vista de l'algorisme és una forma recurrent de calcular les noves pseudodistàncies en funció de les actuals, sense utilitzar aquesta distància entre classes que no interessa guardar.

L'augment d'inèrcia que produïria la fusió de $\mathcal{C} = \{\mathcal{C}_e, \mathcal{C}_d\}$ amb qualsevol altre punt i es pot expressar com a funció de l'augment d'inèrcia que s'hauria produït agregant i a qualsevol dels components de $\mathcal{C}$. La recurrència buscada és:

$$d_{\mathcal{C}i} = \frac{1}{m}\Big(\ (m_e+m_i)d_{ei} + (m_d+m_i)d_{di} - m_i d_{ed}\ \Big)\ ,\ m = m_i + m_e + m_d \tag{3.42}$$

Demostració: En un pas intermedi de l'algorisme, les noves pseudodistàncies que cal calcular són les corresponents a la darrera classe que s'ha format — suposi's $\mathcal{C} = \{\mathcal{C}_e, \mathcal{C}_d\}$ —, amb cada un dels objectes restants del núvol de punts — sigui i —. Partint de l'expressió (**??**) i denotant per m el terme $m_e + m_d + m_i$:

$$\begin{aligned}d_{\mathcal{C}i} = \Delta(\mathcal{C}, i) = &\ \frac{m_{\mathcal{C}} m_i}{(m_{\mathcal{C}}+m_i)} d^2(\mathcal{C}, i) = \\ &\frac{(m_e+m_d)m_i}{(m_e+m_d+m_i)} d^2(\mathcal{C}, i) = \frac{(m_e+m_d)m_i}{m}\left(\sum_{k=1}^{K} \delta_k^2(\mathcal{C}, i)\right)\end{aligned} \tag{3.43}$$

- S'analitza en primer lloc el càlcul de $\delta_k^2(\mathcal{C}, i)$ per les variables quantitatives. Conegudes les coordenades del centre de gravetat de la nova classe (expressió **??**) i

$$\delta_k^2(\mathcal{C}, i) = \frac{\alpha}{s_k^2}\left(\frac{m_e\bar{x}_{\mathcal{C}_e k} + m_d\bar{x}_{\mathcal{C}_d k}}{m_e + m_d} - x_{ik}\right)^2$$

fent denominador comú i factoritzant

$$\delta_k^2(\mathcal{C}, i) = \frac{\alpha}{s_k^2(m_e + m_d)^2}\left(m_e\bar{x}_{\mathcal{C}_e k} + m_d\bar{x}_{\mathcal{C}_d k} - m_e x_{ik} - m_d x_{ik}\right)^2$$

Reagrupant termes i desenvolupant el quadrat

$$\delta_k^2(\mathcal{C}, i) = \frac{\alpha}{s_k^2(m_e + m_d)^2} \quad (m_e^2(\bar{x}_{\mathcal{C}_e k} - x_{ik})^2 + m_d^2(\bar{x}_{\mathcal{C}_d k} - x_{ik})^2 + \\ + \ 2m_e m_d(\bar{x}_{\mathcal{C}_e k} - x_{ik})(\bar{x}_{\mathcal{C}_d k} - x_{ik}))$$

Sumant i restant els termes $m_e m_d(\bar{x}_{\mathcal{C}_e k} - x_{ik})^2$ i $m_e m_d(\bar{x}_{\mathcal{C}_d k} - x_{ik})^2$ a dins del parèntesi:

$$\begin{aligned}\delta_k^2(\mathcal{C}, i) = \frac{\alpha}{s_k^2(m_e + m_d)^2}(\quad & m_e(m_e + m_d)(\bar{x}_{\mathcal{C}_e k} - x_{ik})^2 \\ & + m_d(m_e + m_d)(\bar{x}_{\mathcal{C}_d k} - x_{ik})^2 + \\ & + 2\ m_e m_d(\bar{x}_{\mathcal{C}_e k}\bar{x}_{\mathcal{C}_d k} - \bar{x}_{\mathcal{C}_e k}x_{ik} - \bar{x}_{\mathcal{C}_d k}x_{ik} + x_{ik}^2) + \\ & - m_e m_d(\bar{x}_{\mathcal{C}_e k}^2 - 2\bar{x}_{\mathcal{C}_e k}x_{ik} + x_{ik}^2 + x_{ik}^2 - 2\bar{x}_{\mathcal{C}_d k}x_{ik} + \bar{x}_{\mathcal{C}_d k}^2))\end{aligned}$$

Distribuint algunes constants

$$\begin{aligned}\delta_k^2(\mathcal{C}, i) = \frac{1}{(m_e + m_d)^2}(\quad & m_e(m_e + m_d)\,\delta_k^2(\mathcal{C}_e, i)^2 + m_d(m_e + m_d)\,\delta_k^2(\mathcal{C}_d, i)^2 + \\ & + \frac{\alpha}{s_k^2} m_e m_d\ (\quad 2\bar{x}_{\mathcal{C}_e k}\bar{x}_{\mathcal{C}_d k} - \ 2\bar{x}_{\mathcal{C}_e k}x_{ik} - 2\bar{x}_{\mathcal{C}_d k}x_{ik} + \\ & + 2x_{ik}^2 + 2\bar{x}_{\mathcal{C}_e k}x_{ik} - \bar{x}_{\mathcal{C}_e k}^2 - 2x_{ik}^2 + \\ & - \bar{x}_{\mathcal{C}_d k}^2 + 2\bar{x}_{\mathcal{C}_d k}x_{ik}))\end{aligned}$$

$$\delta_k^2(\mathcal{C}, i) = \frac{m_e(m_e + m_d)\delta_k^2(\mathcal{C}_e, i)^2 + m_d(m_e + m_d)\delta_k^2(\mathcal{C}_d, i)^2 - m_e m_d\delta_k^2(\mathcal{C}_e, \mathcal{C}_d)^2}{(m_e + m_d)^2} \tag{3.44}$$

- Pel que fa a les variables categòriques, utilitzant l'expressió **??**:

$$\delta_k^2(\mathcal{C},i) = \frac{\beta}{n_{\mathcal{Q}}^2}\sum_{j=1}^{n_k}\frac{\left(f_{\mathcal{C}}^{k_j}-f_i^{k_j}\right)^2}{I^{k_j}} = \frac{\beta}{n_{\mathcal{Q}}^2}\sum_{j=1}^{n_k}\frac{1}{I^{k_j}}\left(\frac{m_e f_{\mathcal{C}_e}^{k_j}+m_d f_{\mathcal{C}_d}^{k_j}}{m_e+m_d}-f_i^{k_j}\right)^2 =$$

$$\frac{\beta}{n_{\mathcal{Q}}^2}\sum_{j=1}^{n_k}\frac{1}{I^{k_j}(m_e+m_d)^2}\quad\left(m_e f_{\mathcal{C}_e}^{k_j}+m_d f_{\mathcal{C}_d}^{k_j}-m_e f_i^{k_j}-m_d f_i^{k_j}\right)^2 =$$

$$\frac{\beta}{n_{\mathcal{Q}}^2}\sum_{j=1}^{n_k}\frac{1}{I^{k_j}(m_e+m_d)^2}\quad\left(m_e(f_{\mathcal{C}_e}^{k_j}-f_i^{k_j})+m_d(f_{\mathcal{C}_d}^{k_j}-f_i^{k_j})\right)^2 =$$

$$\frac{\beta}{n_{\mathcal{Q}}^2}\sum_{j=1}^{n_k}\frac{1}{I^{k_j}(m_e+m_d)^2}(\quad m_e^2(f_{\mathcal{C}_e}^{k_j}-f_i^{k_j})^2+m_d^2(f_{\mathcal{C}_d}^{k_j}-f_i^{k_j})^2+$$
$$+\ 2\ m_e m_d\ (f_{\mathcal{C}_e}^{k_j}-f_i^{k_j})(f_{\mathcal{C}_d}^{k_j}-f_i^{k_j}))$$

Distribuint el sumatori

$$\delta_k^2(\mathcal{C},i) = \frac{1}{(m_e+m_d)^2}(\ m_e^2\delta_k^2(\mathcal{C}_e,i)+m_d^2\delta_k^2(\mathcal{C}_d,i)+$$
$$+\frac{\beta}{n_{\mathcal{Q}}^2}\sum_{j=1}^{n_k}\frac{2m_e m_d}{I^{k_j}}\left(f_{\mathcal{C}_e}^{k_j}f_{\mathcal{C}_d}^{k_j}-f_{\mathcal{C}_e}^{k_j}f_i^{k_j}-f_i^{k_j}f_{\mathcal{C}_d}^{k_j}+f_i^2\right)\Bigg) =$$

$$\frac{1}{(m_e+m_d)^2}(\ m_e^2\delta_k^2(\mathcal{C}_e,i)+m_d^2\delta_k^2(\mathcal{C}_d,i)\pm$$
$$\pm m_e m_d\delta_k^2(\mathcal{C}_e,i)\pm m_d m_e\delta_k^2(\mathcal{C}_d,i)+$$
$$+\ \frac{\beta m_e m_d}{n_{\mathcal{Q}}^2}\sum_{j=1}^{n_k}\frac{2}{I^{k_j}}\left(f_{\mathcal{C}_e}^{k_j}f_{\mathcal{C}_d}^{k_j}-f_{\mathcal{C}_e}^{k_j}f_i^{k_j}-f_i^{k_j}f_{\mathcal{C}_d}^{k_j}+f_i^2\right)\Bigg) =$$

$$\delta_k^2(\mathcal{C},i) = \frac{1}{(m_e+m_d)^2}(\ m_e(m_e+m_d)\delta_k^2(\mathcal{C}_e,i)+m_d(m_e+m_d)\delta_k^2(\mathcal{C}_d,i)+$$
$$\tfrac{\beta}{n_{\mathcal{Q}}^2}m_e m_d\textstyle\sum_{j=1}^{n_k}\frac{1}{I^{k_j}}(\ 2f_{\mathcal{C}_e}^{k_j}f_{\mathcal{C}_d}^{k_j}-2f_{\mathcal{C}_e}^{k_j}f_i^{k_j}-2f_i^{k_j}f_{\mathcal{C}_d}^{k_j}+$$
$$+\ 2{f_i^{k_j}}^2-{f_{\mathcal{C}_e}^{k_j}}^2-2{f_i^{k_j}}^2+$$
$$+\ \ 2f_i^{k_j}f_{\mathcal{C}_e}^{k_j}-{f_{\mathcal{C}_d}^{k_j}}^2+2f_{\mathcal{C}_d}^{k_j}f_i^{k_j}\Big)\Big) =$$

$$\delta_k^2(\mathcal{C},i) = \frac{1}{(m_e+m_d)^2}(\ m_e(m_e+m_d)\delta_k^2(\mathcal{C}_e,i)+m_d(m_e+m_d)\delta_k^2(\mathcal{C}_d,i)+$$
$$+\ \tfrac{\beta}{n_{\mathcal{Q}}^2}m_e m_d\textstyle\sum_{j=1}^{n_k}\frac{1}{I^{k_j}}\left(2f_{\mathcal{C}_e}^{k_j}f_{\mathcal{C}_d}^{k_j}-{f_{\mathcal{C}_e}^{k_j}}^2-{f_{\mathcal{C}_d}^{k_j}}^2\right)\Big) =$$

$$\delta_k^2(\mathcal{C},i) = \frac{1}{(m_e+m_d)^2}(\quad m_e(m_e+m_d)\delta_k^2(\mathcal{C}_e,i) + m_d(m_e+m_d)\delta_k^2(\mathcal{C}_d,i) +$$

$$- m_e m_d \delta_k^2(\mathcal{C}_e,\mathcal{C}_d))$$

i de nou es troben expressions idèntiques per les variables qualitatives i quantitatives.

Reprenent **??**

$$d_{\mathcal{C}i} = \frac{(m_e+m_d)m_i}{m}\left(\sum_{k=1}^{K}\delta_k^2(\mathcal{C},i)\right) =$$

$$\frac{(m_e+m_d)m_i}{m}\sum_{k=1}^{K}\frac{1}{(m_e+m_d)^2}(\quad m_e(m_e+m_d)\delta_k^2(\mathcal{C}_e,i) +$$
$$+ m_d(m_e+m_d)\delta_k^2(\mathcal{C}_d,i) - m_e m_d \delta_k^2(\mathcal{C}_e,\mathcal{C}_d)) =$$

$$\frac{1}{m(m_e+m_d)} \; (m_e(m_e+m_d)m_i d^2(\mathcal{C}_e,i) + m_d(m_e+m_d)m_i d_k^2(\mathcal{C}_d,i) +$$
$$+ m_e m_d m_i d_k^2(\mathcal{C}_e,\mathcal{C}_d) +$$

■

Fent factor comú i simplificant

$$d_{\mathcal{C}i} = \frac{1}{m}\left(m_e m_i d^2(\mathcal{C}_e,i) + m_d m_i d^2(\mathcal{C}_d,i) - m_i \frac{m_e m_d}{m_e+m_d} d^2(\mathcal{C}_e,\mathcal{C}_d)\right)$$

Multiplicant i dividint per factors constants

$$d_{\mathcal{C}i} = \tfrac{1}{m}(\quad (m_e+m_i)\tfrac{m_e m_i}{m_e+m_i} d^2(\mathcal{C}_e,i) +$$
$$+ (m_d+m_i)\tfrac{m_d m_i}{m_d+m_i} d^2(\mathcal{C}_d,i) - m_i \tfrac{m_e m_d}{m_e+m_d} d^2(\mathcal{C}_e,\mathcal{C}_d))$$

i s'obté la fòrmula recurrent buscada

$$d_{\mathcal{C}i} = \frac{1}{m}\left(\; (m_e+m_i)\Delta(\mathcal{C}_e,i) + (m_d+m_i)\Delta(\mathcal{C}_d,i) - m_i\Delta(\mathcal{C}_e,\mathcal{C}_d) \; \right)$$

Amb aquesta relació s'actualitzaran les pseudodistàncies d'un pas a l'altre, i a la matriu $\mathcal{D}$ s'hi tenen les quantitats que ens determinen *directament* quins nodes agregar en el següent pas.

L'axioma de reductibilitat

Per últim, comprovar que la recurrència trobada satisfà *l'axioma de reductibilitat*, la qual cosa fa lícit l'ús d'aquesta distància en l'algorisme dels veïns recíprocs encadenats.

Es tracta de veure que, en una passa ξ, l'agregació d'una parella de veïns recíprocs $\mathcal{C}_e$, $\mathcal{C}_d$ en una nova classe $\mathcal{C}$ és tal que, donat un punt qualsevol i extern[21] a $\mathcal{C}$:

$$d_{\mathcal{C}i} \geq \min\{d_{\mathcal{C}_e i}, d_{\mathcal{C}_d i}\}$$

[21] Recordi's que, en aquest cas, a la matriu de distàncies hi ha l'increment d'inèrcia $d_{\mathcal{C}i} = \Delta(\mathcal{C},i)$.

Pel fet que $\mathcal{C}_e$ i $\mathcal{C}_d$ siguin veïns recíprocs se satisfà que

$$d(\mathcal{C}_e, \mathcal{C}_d) < d(\mathcal{C}_e, i) \quad \text{i} \quad d(\mathcal{C}_e, \mathcal{C}_d) < d(\mathcal{C}_d, i)$$

Suposant $\min\{d(\mathcal{C}_e, i), d(\mathcal{C}_d, i)\} = d(\mathcal{C}_e, i)$ resulta que

$$d_{\mathcal{C}i} = \frac{1}{m}\Big((m_e + m_i)d_{\mathcal{C}_e i} + (m_d + m_i)d_{\mathcal{C}_d i} - m_i d_{\mathcal{C}_e\mathcal{C}_d} \Big) >$$

$$\frac{1}{m}\Big((m_e + m_i)d_{\mathcal{C}_e i} + (m_d + m_i)d_{\mathcal{C}_e i} - m_i d_{\mathcal{C}_e i} \Big) =$$

$$\frac{1}{m} d_{\mathcal{C}_e i}(m_e + m_i + m_d + m_i - m_i) = \frac{1}{m} d_{\mathcal{C}_e i} m$$

En definitiva $d_{\mathcal{C}i} > d_{\mathcal{C}_e i}$ ∎

De forma anàloga es faria per a demostrar el cas que

$$\min\{d(\mathcal{C}_e, i), d(\mathcal{C}_d, i)\} = d(\mathcal{C}_d, i)$$

El criteri de Ward es pot utilitzar amb l'estratègia dels veïns recíprocs per qualsevol distància que satisfaci la relació de recurrència **??**, ja que, com es pot veure, l'augment d'inèrcia satisfà l'axioma de reductibilitat independentment de quina sigui l'expressió de la distància, si aquesta satisfà **??**.

Amb aquesta propietat es garantitza que l'espai no col·lapsa amb les successives agregacions, i que la relació de veïnatge és local, quedant, per tant, assegurat que el punt pel qual s'inicïi el procés d'agregació no altera els resultats de la classificació, i que sempre s'obté una mateixa jerarquia indexada: de fet, l'axioma de reductibilitat assegura que si dos punt són veïns recíprocs, no deixaran de ser-ho siguin quines siguin les agregacions que es facin en el seu entorn i que, tard o d'hora es fusionaran en una classe, passant a formar part de la jerarquia.

Finalment, es comprova que, en efecte, els arbres així construïts mai generen inversions. Sota el criteri de Ward, l'índex de nivell d'un node $\mathcal{C} = \{\mathcal{C}_e, \mathcal{C}_d\}$ és l'augment d'inèrcia corresponent a la formació d'aquest node:

$$\nu(\mathcal{C}) = \Delta(\mathcal{C}_e, \mathcal{C}_d) = \frac{m_e m_d}{m_e + m_d} d^2(\mathcal{C}_e, \mathcal{C}_d) = d_{ed}$$

i la inèrcia interna a cada node es calcula recursivament com

$$M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) = \begin{cases} 0, & \mathcal{C} \text{ és fulla} \\ \nu(\mathcal{C}_d) + \nu(\mathcal{C}_e), & \mathcal{C}_d \text{ és el fill dret de } \mathcal{C} \text{ i } \mathcal{C}_e \text{ és el fill esquerre de } \mathcal{C} \end{cases} \tag{3.45}$$

és a dir, com la suma d'índexos de nivell de tots els seus fills.

Cal veure que en una agregació qualsevol $\mathcal{C}' = \{\mathcal{C}, i\}$ on $\mathcal{C} = \{\mathcal{C}_e, \mathcal{C}_d\}$, se satisfà sempre que $\nu(\mathcal{C}) \leq \nu(\mathcal{C}')$ o el que és equivalent $d_{\mathcal{C}_e\mathcal{C}_d} = d_{\mathcal{C}i}$.

Els fills de $\mathcal{C}$ són veïns recíprocs, altrament no s'haurien agregat en una classe. Per tant $d_{\mathcal{C}_e\mathcal{C}_d} < d_{\mathcal{C}_e i}, \forall i$ i també $d_{\mathcal{C}_e\mathcal{C}_d} < d_{\mathcal{C}_d i}, \forall i$. En particular, $d_{\mathcal{C}_e\mathcal{C}_d} < \min_i\{d_{\mathcal{C}_d i}, d_{\mathcal{C}_e i}\}$. Aplicant l'axioma de reductibilitat és clar que $d_{\mathcal{C}_e\mathcal{C}_d} < d_{\mathcal{C}i}$, i que l'arbre no té inversions ∎

3.7 Resum del capítol

En aquest capítol es presenten primerament les generalitats de l'algorisme de base que composa **KLASS** i s'aprofundeix en els elements bàsics de hi intervenen i algunes propietats teòriques.

Es descriu el principi dels veïns recíprocs encadenats i es formalitzen les propietats teòriques dels dendrogrames, sobre les que es tornarà al capítol **??**.

A continuació es desenvolupa la representació de les classes per les variables qualitatives. El sistema que s'utilitza constitueix una de les aportacions d'aquest treball i permet per les variables qualitatives.

- el tractament homogeni de les variables qualitatives i quantitatives
- el tractament homogeni de les classes i dels individus pròpiament dits, propietat que tindrà especial interés a §**??**.

Es desenvolupen expressions de càlcul recurrent del centre de gravetat d'una classe en funció dels centres de gravetat dels seus dos fills que permeten una implementació eficient.

Aquesta representació i la consideració que variables qualitatives i quantitatives es presenten simultàniament en les matrius de dades que estudiem obliga a detenir-se en el càlcul de les distàncies entre individus. Amb aquest esperit, s'aporta una forma de calcular distàncies sobre components qualitatives inspirada en χ^2, que evita la construcció de la taula disjuntiva completa i que s'extén a matrius heterogènies, donant lloc a la família de distàncies mixtes $d^2_{(\alpha_r,\beta_r)}(i,i')$.

En següent terme s'estudien algunes propietats teòriques de $d^2_{(\alpha_r,\beta_r)}(i,i')$: es formula la condició necessària i suficient per a que els elements d'aquesta família tinguin estructura de de distància; es defineix una relació d'equivalència entre els (α_r,β_r) en funció de les classificacions jeràrquiques a que donen lloc les distàncies associades; s'aporta també un valor pels paràmetres (α_r,β_r) que garantitza l'estructura mètrica i representa un ús equilibrat de la informació quantitativa i qualitativa. No es té notícia que una formulació d'aquest estil hagi estat abans utilitzada.

En un següent pas, es tracten dos criteris d'agregació, fent fonamentalment incidència sobre el criteri de Ward. Formalitzant ambdós criteris de manera que el tractament sigui homogeni, l'augment d'inèrcia entre classes considerat en el criteri de Ward fa el mateix paper que la distància entre individus en el del centroide.

Es desenvolupa una fòrmula recurrent pel càlcul dels augments d'inèrcia usant distàncies mixtes. El formalisme de representació de les classes desenvolupat per les variables categòriques fa que la recurrència sigui comuna a totes les variables. En última instància es demostra que l'augment d'inèrcia amb distàncies mixtes satisfà l'axioma de reductibilitat,

essent lícit el seu ús en l'algorisme dels veïns recíprocs encadenats i finalment es comprova que els arbres així generats no presenten inversions.

Per tant, es pot dir que en aquest capítol es presenta un sistema de representació de les classes adient per matrius heterogènies, la qual cosa suscita la definició d'una nova distància i es profunditza en les propietats teòriques de la distància mixta i les seves implicacions en l'algorisme dels veïns recíprocs encadenats.

Chapter 4

Introducció d'informació declarativa

"Scientific endeavors like hypothesis formation and experimentation are dependent upon a sound classification of the relevant objects".

NORD HAUSEN & LANGLEY

L'aplicació de mètodes estadístics estàndards a dominis poc estructurats i amb moltes dades és molt costosa, i generalment *cega*. En moltes ocasions, malgrat desconèixer amb exactitud les classes que conformen l'estructura del domini, l'expert en té suficient coneixement com per a organitzar part d'ell en entitats que tenen sentit. Els mètodes estadístics clàssics pràcticament ignoren aquesta informació. La idea fonamental de la nostra proposta és recollir aquest coneixement en forma declarativa per a construir regles que subdivideixin l'espai de classificació en *entorns* coherents i respectar aquesta primera estructuració sorgida de les indicacions de l'especialista. Aquest tipus de biaix fou suggerit per primer cop per [**?**] per a ser emprat en el contexte dels classificadors inductius simbòlics i desenvolupat, en aquell contexte, per [**?**]. La seva utilització en un sistema de classificació constitueix l'aportació més important d'aquesta tesi i al nostre entendre, és una forma raonable de proporcionar informació *semàntica* al sistema. No es tenen referències de treballs que intentin corregir una classificació amb informació declarativa.

En efecte, els classificadors estadístics reposen habitualment en la definició d'una distància (o similitud) sobre l'espai de les variables, la qual cosa forma part del nivell *sintàctic*, i fa que, en algunes ocasions, la validesa *semàntica* de la classificació proporcionada sigui molt pobra. En altres paraules, la *interpretació* dels resultats, per part dels experts (des del punt de vista semàntic, doncs), pot ser massa forçada, pel fet que una distància no té en compte les possibles rellevàncies (o irrellevàncies) de certes variables amb respecte als objectius de la classificació [**?**] ni tampoc reflexa els objectius i necessitats de l'expert a l'hora de classificar. Si bé aquesta informació (que actuarà com un biaix de la classificació) només ha de ser considerada com una *ajuda* a la classificació, sens dubte la seva utilització millora la qualitat dels resultats del procés.

Incorporar informació de caire semàntic a un sistema és una tasca complexa, i no solament

en estadística, sino que també suposa un problema de representació del coneixement difícil en el món de la intel·ligència artificial.

Es presenta aquí una primera aproximació orientada a explotar aquest tipus d'informació, en la mesura del possible, sense caure en pretensions que dificultin massa l'esforç de formalització per part de l'expert. Una formalització dificultosa revertiria en un distanciament sistema-usuari i, en conseqüència, en un baix rendiment del sistema.

A la secció §?? es detalla el formalisme sota el que es representa aquesta informació, i les raons per les que s'ha fet així. Només dir que, amb això, es pretén cobrir un triple objectiu:

- Incorporar a la classificació informació abans ignorada, com ara relacions entre variables, restriccions ...
- Recollir, cas que sigui rellevant, els objectius de la classificació que s'està demanant
- *Garantir* plausibilitat en la interpretació de la classificació obtinguda.

4.1 Representació del coneixement

Ja s'ha dit que es preten recollir qualsevol tipus d'informació semàntica que li pugui semblar rellevant a l'expert, pel que fa a la classificació.

Incloure relacions entre variables, condicions de pertinença a una classe, restriccions d'incompatibilitat de grups d'objectes, qüestions d'interpretació ... en un únic *formalisme*, condueix a buscar un *model* de representació molt genèric, d'elevada capacitat expressiva, amb prou potència per tractar amb tot això, i per tant, de tipus *declaratiu*. Aquesta és la raó per la que s'ha plantejat treballar amb regles lògiques de primer ordre.

L'estructura d'una regla r és senzilla des del punt de vista sintàctic. Està composta d'una part dreta indicant alguna classe $\mathcal{C}$ i una part esquerra amb la condició $\mathcal{B}$ que ha de satisfer un objecte i per a formar part de la classe $\mathcal{C}$.

$$\mathcal{B} \rightarrow \mathcal{C} \tag{4.1}$$

Direm que un objecte i és *seleccionat* per una regla $r = (\mathcal{B} \longrightarrow \mathcal{C})$ si

$$\mathcal{B}(x_i) = cert$$

Llavors s'interpreta que l'objecte i *pertany* a la classe $\mathcal{C}$, i s'identifica $\mathcal{C}$ amb el conjunt d'individus d'$\mathcal{I}$ que pertanyen a $\mathcal{C}$. Per contra, si $\mathcal{B}(x_i) = fals$ i no és seleccionat per r.

$\mathcal{C}$ pot ser una classe amb entitat real, és a dir, que existeixi (com a exemple, la classe de les Cliones en les esponges de mar §??, o la de les estrelles del Disc §??), o un simple identificador destinat a especificar quins objectes té sentit que s'agrupin en una mateixa classe o no.

Entendre — $\mathcal{B}$ — la *condició* de pertinença a $\mathcal{C}$ en sentit molt ampli, és el que dota les regles de la seva potència expressiva. Solament es demana un requisit a $\mathcal{B}$: ser una expressió booleana avaluable amb la informació disponible en una única fila de la matriu de dades X (és per això que no s'especifica cap forma concreta per $\mathcal{B}$). $\mathcal{B}$ pot, doncs, presentar formes ben diverses segons el que es vulgui expressar[1]. Entre d'altres:

- Una condició de la forma

$$\mathcal{B} = \left(\bigwedge_{k=1}^{K} (X_k \in \mathcal{D}'_k) \right), \begin{cases} \mathcal{D}'_k \subset \mathcal{D}_k, & X_k \text{ qualitativa} \\ \mathcal{D}'_k \subset \Re & X_k \text{ quantitativa} \end{cases} \tag{4.2}$$

  serveix per a especificar els valors (o conjunts de valors) que han de prendre les variables (o part d'elles) pels objectes de $\mathcal{C}$.

  Com a exemple es presenten regles corresponents a una de les aplicacions reals que s'han realitzat (veure §**??**). El domini poc estructurat en qüestió és el de les `esponges de mar`. A més de la matriu de dades, els espongiòlegs han proporcionat un conjunt de 4 regles per a ser considerades en el procés de classificació (veure [**?**]). Aquestes regles tenen el següent aspecte:

$$\begin{aligned} r_1 : \quad & ((tipo\ espicula\ principal\ tilostilo = 1) \wedge \\ & \wedge (tipo\ microsclera = aster) \wedge (tipo\ de\ aster = espiraster) \wedge \\ & \wedge (forma\ final \in \{masivo\ irregular, incrustante\}) \\ & \longrightarrow \mathcal{C}_1) \end{aligned} \tag{4.3}$$

  Aquesta formalització inclou altres relacions funcionals més complexes que hagin de satisfer els valors dels atributs per als objectes de la classe $\mathcal{C}_1$, $(\leq, <, >, \geq \ldots)$. En aquesta línia, una altra particularització possible de **??** seria, per exemple,

$$((X_1 = 3) \wedge (X_2 \leq 4) \wedge (X_3 \in [-6, 6]) \wedge (X_4 \in \{a, b\}) \wedge (X_5 > 7) \wedge (X_5 < 14)) \longrightarrow \mathcal{C}_1$$

- La forma

$$\mathcal{B} = \bigwedge_{k=1}^{K} (X_k \notin \mathcal{D}'_k)$$

  indicaria justament les combinacions de valors *prohibits* per als objectes de la classe $\mathcal{C}$.

  Continuant amb l'exemple **??**, per expressar que en la classe $\mathcal{C}_1$, l'atribut $tipo - microsclera$ **no** pot prendre els valors $microestilos - y - microxas$, i $sin - microscleras$, la regla corresponent seria:

$$((\ tipo\text{-}microsclera \neq (sin\text{-}microscleras \quad microestilos\text{-}y\text{-}microxas\ )\ )) \longrightarrow \mathcal{C}_1)$$

[1] No es posa l'expressió regular de $\mathcal{B}$ perquè, en primer lloc l'interés no és la sintàctica i, a més, la seva forma és prou general i oberta com per a que tal expressió no resultés massa aclaridora pel lector.

- La relació que ha d'existir entre algunes de les variables dins una classe $\mathcal{C}$, entenent, per tant, que les relacions poden ser locals a cada una de les classes, també es pot expressar segons aquest model:

  $$\mathcal{B} = \bigwedge_b (h_b(X_1, \ldots, X_K)) \quad h_b \text{ funció booleana}$$

  Per exemple

  $$(x_1 > x_2) \longrightarrow \mathcal{C}_1$$

  $$(x_1 \leq x_2) \longrightarrow \mathcal{C}_2$$

  indicaria que hi ha dues classes en les que s'inverteix la relació d'ordre entre X_1 i X_2.

- El formalisme que es proposa, fins i tot permet tractar amb variables transformades, és a dir, imposar condicions no sobre les variables observades, sino sobre transformacions d'aquestes. Només cal definir a l'intèrpret la forma com efectuar la transformació.

  Aquest és un cas que s'ha presentat, per exemple, en tractar de classificar les `estrelles` de la nostra galàxia (veure **??**). Una de les regles proporcionades per l'expert diu que

  $$(|\vec{v}| > 75) \longrightarrow Halo$$

  però la velocitat d'una estrella qualsevol no és cap de les variables usades en la seva descripció. Ara bé, existeixen relacions que permeten calcular-la a partir d'altres variables que sí estan presents a la matriu de dades. A §**??** es detallen.

  Doncs bé, si es defineixen aquestes funcions convenientment, ni tan sols és necessari que les variables implicades en una regla estiguin presents a la matriu de dades.

És evident, doncs, que, quant a les regles, hi ha una àmplia gamma de possibilitats que permet a l'usuari expressar condicions molt diverses. La introducció d'aquest tipus de variables augmenta considerablement la potència del sistema, ja que permet avaluar l'estructura de les dades sobre un espai de representació *diferent* del que defineixen les variables del fitxer de propietats, amb el consegüent guany informatiu. D'alguna forma, s'està combinant la informació que es deriva del núvol de punts observant-lo desde la seva representació simultània en diferents sistemes de coordenades, i això no es pot fer amb *cap* sistema estadístic actual, més que duplicant les variables a la matriu de dades i utilitzant-les totes en la classificació. La idea d'utilitzar variables derivades en les regles se m'apareix, per primer cop, arran de la classificació de les estrelles que es presenta a §**??** i degut a l'especial idiosincràsia de les dades que es tracten en aquell contexte. La naturalesa del propi **LISP** fa que no hi hagi diferències en la implementació d'un avaluador de regles construïdes sobre les variables observades i la d'aquell que admeti variables transformades també. Així, **KLASS** i

LINNEO comparteixen aquesta característica, que és una de les que més potència donen al sistema.

A la §**??** es discuteix sobre el tractament que es fa en **LISP** d'aquest tipus de regles.

És important remarcar que la *simplicitat* d'aquest tipus de regles suggereix un cert grau de generalitat: quan més termes conjuntius tingui una regla, o més variables hi estiguin implicades[2], menys general és i menys objectes la satisfaran. Una regla *complexa*, sol ser molt específica, però també més precissa; és a dir, el conjunt d'individus que la satisfan és petit, però molt homogeni.

A la propera secció s'explica com utilitzar aquestes regles per a millorar la classificació estadística estàndard. Caldrà induir sobre $\mathcal{I}$ una partició inicial en funció d'aquestes regles. De fet, la *qualitat* d'aquesta partició, i en conseqüència, la de la classificació final, depèn directament de l'estructura de les regles. Per això és important que l'expert seleccioni correctament els atributs (o variables) que s'hi fan intervenir, treballant, amb preferència, amb els més *rellevants*, segons l'objectiu de la classificació.

En general, els experts saben prou bé el que és rellevant i el que no ho és en el seu domini. De tota manera, hi ha algunes eines automàtiques que poden ajudar-los, si és que fa falta, a quantificar la rellevància dels atributs que apareixen en la descripció dels individus, així com a identificar aquells que tenen una rellevància significativa (per una proposta en aquesta línia, veure [**?**] o [**?**]).

4.2 Classificació amb regles

4.2.1 Preliminars

Un cop obtingudes les regles de l'expert $\mathcal{R} = \{r_1 \dots r_R\}$ es tracta de fer-les intervenir en el procés de classificació de forma que la informació que contenen sigui tinguda en compte pel classificador.

A primera vista, hom podria pensar que, amb tota probabilitat, els experts proporcionaran regles basades en les característiques més rellevants per la classificació. I efectivament, així sol ser. Per l'aplicació **??**, com a exemple, la tècnica proposada a [**?**] identifica 18 atributs *rellevants*, la qual cosa representa el 40% del total. Les regles proporcionades pels experts inclouen únicament informació referent a aquestes 18 variables.

Denotant per $\mathcal{A}^{\mathcal{R}}$ el conjunt de variables (atributs) utilitzades en el conjunt de regles $\mathcal{R}$, i $\mathcal{A}$ al conjunt de totes les variables X_k, resulta que generalment $\mathcal{A}^{\mathcal{R}} \subset \mathcal{A}$.

I ja que els propis experts diuen què és rellevant i què no ho és, es podria prescindir directament dels atributs $\mathcal{A} - \mathcal{A}^{\mathcal{R}}$ i reduir notablement la dimensió de la matriu a classificar.

[2]Una variable es considera implicada en una regla, si hi apareix directa o indirectament. En aquest sentit, una sola variable transformació d'un conjunt important de variables originals, donaria lloc a una regla d'elevada complexitat.

Proves empíriques mostren [**?**] que no es pot fer aquesta simplificació, que suposa fer una hipòtesi ingènua de món tancat sobre $\mathcal{R}$ [**?**] — suposant que conté tot el coneixement sobre el domini —, que no és complet! És més, en realitat, no es pot fer cap suposició sobre la rellevància dels atributs d'$\mathcal{A} - \mathcal{A}^{\mathcal{R}}$.

Si bé és cert que els atributs d'$\mathcal{A}^{\mathcal{R}}$ són rellevants, hi ha, en les regles, informació molt més rica que encara es pot utilitzar, com són per exemple les relacions entre atributs (o variables) locals a cada classe. Aquesta informació té capacitat, com ja s'ha dit amb anterioritat, per donar una estructura preliminar al domini en estudi, i una forma de posar-la de manifest és el que es presenta en les següents seccions.

4.2.2 La partició induïda per les regles

Sia $\mathcal{R} = \{r_1 \dots r_R\}$ amb $(r_\rho = \mathcal{B}_\rho \rightarrow \mathcal{C}_\rho),\ (\rho = 1 : R)$ el conjunt de regles proporcionades per l'expert.

Anomeni's $\mathcal{P}_{\mathcal{R}}$ al conjunt de classes que apareixen com a part dreta d'alguna regla d'$\mathcal{R}$ (amb o sense existència real).

$$\mathcal{P}_{\mathcal{R}} = \{\mathcal{C}_\rho : \exists(\rho = 1 : R),\ r_\rho = (\mathcal{B}_\rho \rightarrow \mathcal{C}_\rho)\}$$

Els objectes d'$\mathcal{I}$ poden ser seleccionats :

1. Per cap regla d'$\mathcal{R}$.
2. Per una o més regles d'$\mathcal{R}$ amb una mateixa part dreta.
3. Per diverses regles d'$\mathcal{R}$ amb diferents parts dretes.

Els objectes que no satisfan cap regla no són pas motiu de preocupació: ja s'ha dit que l'expert proporciona coneixement *parcial* sobre el domini. Si un objecte no satisfà cap regla, simplement està en aquella porció del domini de la que no es té cap informació addicional. El conjunt d'objectes que estan en aquesta situació forma part del que es denota com $\mathcal{C}_0$ o *classe residual*.

Pel que fa als objectes seleccionats per regles contradictòries, (de diferents parts dretes), s'assumeix que corresponen a parts del domini per a les que l'expert ha proporcionat informació inconsistent, i s'ha preferit tractar-los prescindint d'aquesta informació. S'inclouen, per tant, en la classe residual. Al capítol **??** es descriu un procés per a reduir el nombre d'objectes en aquesta situació[3].

És a dir que

$$\mathcal{C}_0 = \left\{ i \in \mathcal{I} : \begin{array}{l} i \text{ no és seleccionat per cap regla d'}\mathcal{R} \\ i \text{ és seleccionat per almenys dues regles d'}\mathcal{R} \text{ amb part dreta diferent} \end{array} \right\}$$

[3]Amb freqüència el conjunt de regles presenta casos d'inclusió (una de les classes és subclasse de l'altra) que es poden resoldre afegint alguna condició a la més general de les regles implicades en la contradicció. Aquest procés desemboca en un conjunt de regles amb poques seleccions contradictòries molt ràpidament.

i la classe residual es pot interpretar com la classe dels objectes pels quals o bé no es disposa d'informació addicional, o bé la que hi ha no és consistent.

La resta d'objectes pertanyen a alguna de les classes de $\mathcal{P}_{\mathcal{R}}$ sense cap contradicció. Així doncs, resulta que $\mathcal{P}_{\mathcal{R}^*} = \mathcal{P}_{\mathcal{R}} \cup \{\{i\}\ tq\ i \in \mathcal{C}_0\}$ no és més que una partició d'$\mathcal{I}$ en $card(\mathcal{C}_0)+R$ classes, que satisfà les restriccions imposades per l'expert. Definim $\mathcal{P}_{\mathcal{R}^*}$ *partició induïda per les regles d'*$\mathcal{R}$ *sobre* $\mathcal{I}$.

La raó per la que no es treballa directament amb $\mathcal{P}_{\mathcal{R}} \cup \mathcal{C}_0$ és que $\mathcal{C}_0$ no representa cap agrupació d'objectes amb significat semàntic. Simplement aglutina aquells objectes pels que no s'usa cap informació addicional, per les raons que sigui.

4.2.3 Incorporació de les regles al procés general

A la secció §**??** s'havia caracteritzat el procés de classificació com una successió $\mathcal{P}$ de particions encaixades que partien de la més fina possible $\mathcal{P}_1$ i s'anaven engrossint fins arribar a $\mathcal{P}_n = \{\mathcal{I}\}$.

Per tal de respectar l'estructura del domini suggerida per l'expert a través d'$\mathcal{R}$, cal que $\mathcal{P}_{\mathcal{R}^*}$ constitueixi un dels elements de $\mathcal{P}$.

De primer antuvi, es podria pensar a considerar el núvol de punts format per la classe residual més els representants de les classes induïdes per les regles

$$\mathcal{I}^* = \{\bar{\imath}_{\mathcal{C}} : \mathcal{C} \in \mathcal{P}_{\mathcal{R}}\} \cup \mathcal{C}_0,$$

i construir l'arbre jeràrquic d'$\mathcal{I}^*$ segons el mètode presentat a les seccions anteriors, fent $\mathcal{P}_1 = \{\mathcal{I}^*\}$. Això és perfectament factible gràcies a la possibilitat de tractar aquestes representacions prototípiques de cada classe com a objectes ordinaris §**??**.

En aquest cas, la partició inicial induïda per les regles quedaria en el nivell terminal de l'arbre de classificació, la qual cosa vindria a indicar que el grau d'abstracció (o de generalitat) de les classes identificades *a priori* per l'expert és *homogeni* !, i això no correspon en absolut a la realitat. És important remarcar el fet que es vol aprofitar *qualsevol* coneixement que del domini tingui l'especialista[4]. Per això s'ha eliminat la restricció de completitud del sistema de regles. També es voldria eliminar la d'homogeneïtat.

És freqüent que un metge pugui fer una primera caracterització de totes les malalties a nivell general, però ben segur que podrà ser molt més precís per les de la seva pròpia especialitat! El nivell d'abstracció d'unes i altres classes no té per què coincidir. I per a que això es posi de manifest en la classificació final, únicament es pot considerar $\mathcal{P}_{\mathcal{R}^*}$ com un element *intermedi* de la successió de particions d'$\mathcal{I}$, en que els representants de classe tenen *diferents* índexos de nivell no nuls.

Per tant, es proposa una estratègia mixta de classificació en que els nodes de la partició inicial induïda per les regles siguin al seu torn organitzats en forma d'arbre jeràrquic seguint

[4]L'única condició que s'imposa a $\mathcal{R}$ és que conceptualment no contingui inclusions (veure §**??**).

un procés idèntic al que els ha d'integrar amb els elements de la classe residual en una única estructura.

En primer lloc es realitza una classificació local a cada un dels elements de $\mathcal{P}_\mathcal{R}$. Això genera els primers $(\sum_{\rho=1}^{R} card(\mathcal{C}_\rho)) - 1$, elements de la successió de particions $\mathcal{P}$ que caracteritzarà la classificació. Utilitzant la descripció conceptual $\bar{\imath}_\mathcal{C}$, el contingent $m_\mathcal{C}$ i l'estructura arbòrea de de cada classe $\mathcal{C} \in \mathcal{P}_\mathcal{R}$, es fa una darrera classificació del conjunt $\tilde{\mathcal{I}} = \{(\bar{\imath}_\mathcal{C}, m_\mathcal{C}) : \mathcal{C} \in \mathcal{P}_\mathcal{R}\} \cup \{(i, 1);\ i \in \mathcal{C}_0\} \subset \mathcal{I}^* \times \mathcal{M}$, per a integrar tots els elements d'$\mathcal{I}$ en un únic arbre ascendent jeràrquic. Noti's que la classificació tradicional que no incorpora regles, constitueix un cas particular d'aquesta estructura en que el conjunt de masses de cada punt és sempre la unitat.

Això permet aprofitar qualsevol informació parcial sobre el domini: l'expert no necessita donar un conjunt ni complet ni homogeni de regles per a millorar la classificació, sinó només aquelles regles que realment sap, deixant que el sistema busqui l'estructura d'aquelles parts del domini per a les que no s'ha proporcionat informació. Si hi havia classes més generals que d'altres en $\mathcal{R}$, l'índex de nivell d'aquestes serà també més alt (estaran formades per objectes menys homogenis), i per talls inferiors a aquest índexos, elles quedaran subdividides en altres classes més específiques.

A nivell tècnic, la construcció d'una única jerarquia requereix que:

- Es treballi amb una distància sempre comuna a les classificacions locals i a la residual[5]. De no fer-ho així, no serien comparables els índexos de nivell dels subarbres provinents de diferents regles.

- Els $\bar{\imath}_\mathcal{C}$ van dotats d'una massa $m_\mathcal{C}$. Altrament, no s'estarien tractant com objectes intermedis de la jerarquia, ni se'ls consideraria com a representants de classes de diferents grandàries. La incidència que això té en l'algorisme és a nivell d'inicialitzacions, fonamentalment.

  Amb el criteri de Ward, la inicialització de la matriu de pseudodistàncies ha de tenir l'augment d'inèrcia intra classes que es produïria en fusionar i i i', i ara ja no es pot suposar que $m_i = 1,\ \forall i$. Per tant,

$$\mathcal{D}_0 = \left(d_{ii'} = \frac{m_i m'_i}{m_i + m'_i} d^2(i, i') \right)$$

4.3 Qüestions tècniques

Ja s'ha dit que la implementació del classificador s'ha dut a terme en **LISP**. En aquesta secció es descriu la forma com l'usuari ha d'especificar les regles per tal que **KLASS** les pugui

[5] Això té importància en el càlcul d'α i β quan es treballa en distància mixta, o en el de s_j^2 per a normalitzar la distància euclídia.

processar.

En realitat, el sistema és totalment obert i flexible, i no s'ha volgut dissenyar una sintaxi específica associada a una gramàtica d'expressions per a no restringir les seves possibilitats. La sintaxi d'una regla seria

$$((< funcio\ booleana\ LISP >) -> \quad classe)$$

El darrer element de la llista és la part dreta de la regla (*i.e.* l'identificador de la classe), mentre que el primer representa la condició $\mathcal{B}$. L'única *restricció* que hauria de satisfer $\mathcal{B}$, en principi, és que sigui una funció **LISP** que retorni `cert` o `fals`.

Amb el fitxer de regles, **KLASS** determinarà la partició induïda sobre $\mathcal{I}$. El procés suposa l'avaluació de totes les regles per tots els objectes de la mostra. Aquesta avaluació passa per la instanciació de les variables que apareixen en les regles, pels valors que aquestes prenen amb cada un dels objectes. Procés, aquest, que, *a priori*,evident requereix una anàlisi sintàctica i lèxica de les regles, amb la detecció dels ítems que representen variables, i la seva substitució.

La regla **??** tindria una representació:

```
((and (= c.tipo_espicula_principal_tilostilo 1)
 (equal i.tipo_microsclera  'aster )
 (equal j.tipo_de_aster 'espiraster)
 (member k.forma_final '(masivo_irregular incrustante) ))
-> C1)
```

Aprofitant el cicle d'avaluació recursiva i la capacitat de generar codi dinàmicament de **LISP**, es pot *eliminar* completament aquesta etapa d'anàlisi de les regles introduint una norma sintàctica molt senzilla en les regles:

Els nom de les variables que apareixen en les regles han d'anar parentitzats.

Subjectant a l'expert a aquesta norma (no gens restrictiva), es pot dissenyar una estructura que permet avaluar les regles directament, sense cap tractament previ: Es tracta d'associar, en el procés de lectura del fitxer de propietats, una macro sense paràmetres a cada variable X_k, que retornarà, quan es llegeixi la matriu de dades, el valor que aquesta variable pren per l'objecte en curs. L'accés a aquest element de la matriu de dades es fa per posició i és de cost constant. La definició de la macro també té cost constant. A partir d'aquí, l'element `(X`$_k$`)` de qualsevol regla, avaluarà en la instanciació de la variable per l'objecte en curs i el propi intèrpret podrà gestionar aquest procés perquè ja no hi ha cap diferència entre les *crides* a variables i les crides a qualsevol altra funció **LISP**.

D'una forma elegant, s'ha aconseguit obviar el farregós preprocés d'anàlisi de les regles. Amb aquest mecanisme, l'avaluació de la regla es redueix a un simple

```
(eval condicio)
```

solució molt més còmoda que el recorregut seqüencial i exhaustiu de la condició, discernint si cada un dels elements que la composen són noms de variables o no, i fent les substitucions necessàries. L'aspecte definitiu de la regla **??** és:

```
((and (= (c.tipo_espicula_principal_tilostilo) 1)
 (equal (i.tipo_microsclera)  'aster )
 (equal (j.tipo_de_aster) 'espiraster)
 (member (k.forma_final) '(masivo_irregular incrustante) ))
-> C1)
```

Paral·lelament, la variable `k.forma_final`, que ocupa la posició 35 de la matriu de dades, es defineix com una macro, de cos:

```
(elt 35 objecte\_en\_curs)
```

I aquesta tècnica és extrapolable a qualsevol proces que requereixi una anàlisi sintàctica per tal d'instanciar alguns elements d'una expressio als valors concrets.

D'igual forma, s'ha vist com en les regles s'hi poden fer intervenir variables transformades. Únicament cal introduir-hi crides a funcions que calculin, en funció de les variables definides explícitament, les noves variables (veure §**??**).

Finalment, en relació al cost del procés de classificació pròpiament dit, dir que la partició induïda per les regles està formada d'agrupacions que contenen menys elements que el conjunt d'objectes original. S'ha observat que la classificació separada de cada agrupació requereix menys temps en conjunt que la classificació de l'espai global. En general, es podrà considerar que els costos computacionals de les classificacions locals als $\mathcal{C} \in \mathcal{P}_{\mathcal{R}}$ no són significatius, ja que les seves grandàries acostumen a ser molt petites comparades amb la grandària de la mostra total. Encara més, aquestes classificacions locals es poden executar en paral·lel, ja que les $\mathcal{C} \in \mathcal{P}_{\mathcal{R}}$ són conjunts d'individus disjunts.

Per tant, des d'un punt de vista computacional, la classificació d'$\tilde{\mathcal{I}}$ és l'únic procés rellevant. $\tilde{\mathcal{I}}$ té menys elements que $\mathcal{I}$, i això redueix considerablement el cost computacional total. Al capítol **??** hi ha resultats concrets que il·lustren aquests darrers paràgrafs.

4.4 De les inversions

Una classificació ascendent jeràrquica construïda a partir del criteri de Ward, en condicions normals, no presenta inversions de cap mena. Això significa que l'augment d'inèrcia produït per les agregacions és una ultramètrica, o el que és el mateix, defineix una jerarquia indexada sobre el conjunt de punts a classificar.

Per construcció, les agregacions es fan en un ordre tal que

$$\Delta_\xi \geq \Delta_{\xi-1},\ \forall\,(\xi = 1 : n-1)$$

Efectivament, a cada passa s'agrega aquell parell de punts que produeixen el mínim increment d'inercia intra-classe i, per tant, a mesura que es puja en l'arbre jeràrquic, les branques que uneixen un node al seu fill s'allarguen més i més.

Doncs bé, fent una classificació amb regles es pot perdre la condició d'ultramètrica de Δ.

4.4.1 Gènesi de les inversions

Si bé és cert que l'augment d'inèrcia de tota agregació és *sempre* positiu, el que ja no es preserva, usant les regles, és que Δ sigui monòtonament creixent i això és inevitable.

En el cas estàndard, Δ és ultramètrica, és a dir, el dendrograma associat no té inversions, com a conseqüència que primer s'agreguin els punts que estan molt acostats entre ells — que generen uns Δ's petits —, i cada cop s'ajuntin classes més separades entre si — amb més inèrcia —, amb la qual cosa Δ es fa cada cop més gran (veure §**??**).

La introducció de les regles suposa una discontinuïtat en aquest procés deguda a l'existència de la classe residual i a les connotacions associades a la partició induïda per elles. És a dir, els objectes són seleccionats per les regles en base a consideracions semàntiques (p. ex. un objectiu de classificació), i no pas perquè mètricament estiguin lluny o a la vora, malgrat que puguin esser-ho — i justament aquesta és la raó de ser de les regles. Si la distància tota sola fos prou potent per a no contradir la semàntica, les regles no serien ni necessàries, ni útils—.

Suposi's una partició $\mathcal{P}_\mathcal{R}$ induïda sobre $\mathcal{I}$ per un conjunt de regles $\mathcal{R}$. Suposi's $\mathcal{C} \in \mathcal{R}$ i un element $i \in \mathcal{C}_0$ de la classe residual que quedi dins de l'espai mètric de la classe $\mathcal{C}$ ($\mathcal{C}_5$ en la figura **??**), malgrat no haver estat seleccionat per cap de les regles.

És perfectament possible que i quedi més a prop del centre de gravetat de la classe que altres punts que foren seleccionats per la regla.

La classificació local a $\mathcal{C}$ es fa prescindint dels objectes de la classe residual, i en l'etapa d'integració, l'augment d'inèrcia produït per la fusió d'$\overline{\imath}_\mathcal{C}$, i i és inferior al que correspon a la darrera agregació de $\mathcal{C}$. L'element i està més a la vora d'$\overline{\imath}_\mathcal{C}$ que els dos fills de l'arrel de l'arbre, i l'augment d'inèrcia depèn de la distància entre els punts que s'agreguen. I apareix la inversió (Δ deixa de ser monòtona creixent).

D'haver classificat $\mathcal{C} \cup \{i\}$, i hauria quedat immers en la jerarquia, formant part d'alguna classe de nivell intermedi, i generant un arbre sense inversions. Però quan s'estructuren les classes induïdes per les regles no es tenen tots els objectes que hi pertanyen. Aquesta és la raó de les inversions. En conclusió, es pot dir que el criteri de Ward presenta inversions si objectes *interns* al núvol de punts no estan presents a l'inici de la classificació, que és exactament el que succeeix quan s'usen les regles.

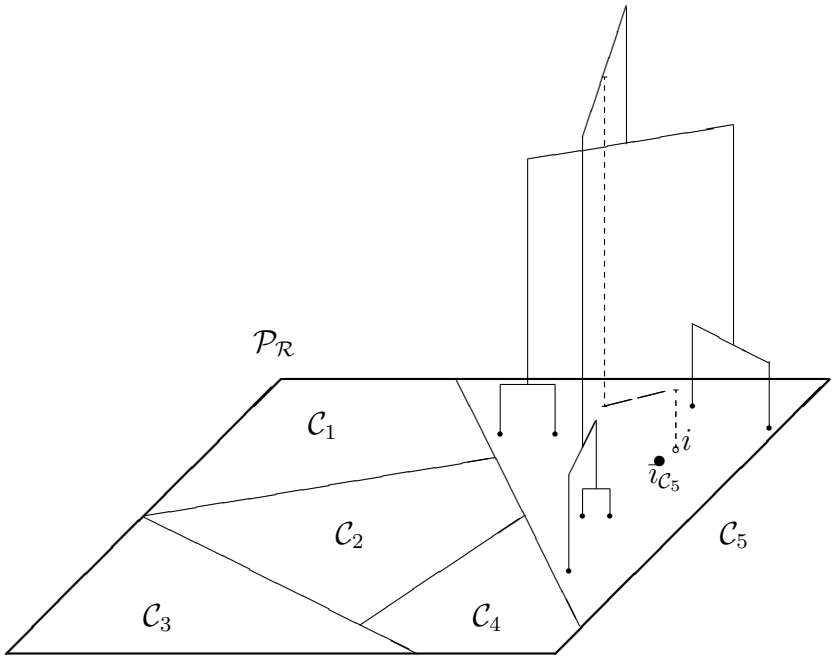


Figure 4.1: Naturalesa de les inversions.

Aquest problema es resoldria si cada cop que es detecta una inversió en agregar dues subclasses $\mathcal{C}$ i $\mathcal{C}'$ es bloquegés el procés i es classifiqués $\mathcal{C} \cup \mathcal{C}'$ localment. Però aquesta estratègia presenta alguns inconvenients:

- Computacionalment és caríssima
- No respecta l'estructura de base suggerida per l'expert, que és un dels nostres objectius.

Per bé que resol el problema des del punt de vista mètric, no sempre conserva la semàntica capturada en $\mathcal{R}$.

4.4.2 Tractament de les inversions

Donat doncs que el criteri de Ward presenta inversions si s'usa amb les regles, i que aquestes porten alguns problemes en la interpretació de les $\alpha-$particions de l'arbre de classificació (veure §**??**), es proposa una altra forma d'indexar la jerarquia.

Una funció que sí és sempre creixent és la inèrcia interna d'una subclasse. Llavors, fent $\nu(\mathcal{C}) = M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}})$ es resol definitivament el problema de les inversions, perquè per

$$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \implies M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) > M^2(\mathcal{C}_1/\bar{\imath}_{\mathcal{C}_1}) + M^2(\mathcal{C}_2/\bar{\imath}_{\mathcal{C}_2})$$

$$\text{i així } M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) > M^2(\mathcal{C}_1/\bar{\imath}_{\mathcal{C}_1}) \ \wedge \ M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) > M^2(\mathcal{C}_2/\bar{\imath}_{\mathcal{C}_2})$$

per petita que sigui la diferència. L'expressió **??** ja expressa que l'augment d'inèrcia és sempre positiu.

Hom podria plantejar-se llavors l'ús de $\nu(\mathcal{C}) = M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}})$ com a criteri d'agregació en el procés general i agregar en cada passa els nodes

$$\mathcal{C} = \{\mathcal{C}_e, \mathcal{C}_d\} \; tq \; M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) = \min_c M^2(c/\bar{\imath}_c)$$

però amb això es perd la condició d'òptim local per la inèrcia intra classe dels successius $\alpha-$talls. En efecte, aquest criteri, anomenat de la variància interna mínima [**?**], prioritza les classes d'inèrcia interna petita, sigui quin sigui l'augment que això representa per la inèrcia intra-classe total, i es poden agregar nodes que estan més a prop d'altres classes, només perquè aquestes ja són prou grosses i ténen una inèrcia interna gran.

Utilitzant

$$\nu(\mathcal{C}) = M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}})$$

com a índex de nivell i mantenint el criteri de Ward com a criteri d'agregació, la interpretació del dendrograma (o arbre ascendent jeràrquic) és clara:

- Les $\alpha-$particions donen, a cada nivell, la partició que minimitza la inèrcia intra-classe
- Les longituds de les branques de l'arbre indiquen l'augment d'inèrcia corresponent a la fusió de cada node
- El nivell de cada node és la seva inèrcia interna

A diferència del que s'ha exposat a **??** la inèrcia intra-classe d'una $\alpha-$partició ($\mathcal{P}$) serà ara

$$M_i^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) = \sum_{\mathcal{C}\in\mathcal{P}} \nu(\mathcal{C}) \text{ i } \nu(\mathcal{C}) = \Delta(\mathcal{C}_e, \mathcal{C}_d) + \nu(\mathcal{C}_e) + \nu(\mathcal{C}_d)$$

on $\mathcal{C}_e$ és el fill esquerre de $\mathcal{C}$ i $\mathcal{C}_d$ el fill dret.

Val a dir que l'aspecte del dendrograma pràcticament no varia usant com a índex l'augment d'inèrcia si es disposa de **tots** els elements a classificar a l'inici del procés, i que òbviament, en aquest cas, les agregacions que es fan són les mateixes i, tanmateix, la posterior aparició de punts per classificar genera jerarquies indexades.

4.5 Proposta d'una metodologia de treball

En aquesta secció es proposa una metodologia interactiva d'anàlisi que condueixi a una classificació final satisfactòria. De fet, s'estructura l'anàlisi en un cicle de dues fases, una d'adquisició del coneixement de suport i una de classificació pròpiament dita, que condueixen a una estructuració final del núvol de punts en un conjunt de classes presumiblement interpretables.

Al llarg del procés, hi ha un intercanvi d'informació entre el sistema i l'usuari, de forma que, usant la informació que KLASS produeix, l'expert omple les llacunes que té sobre

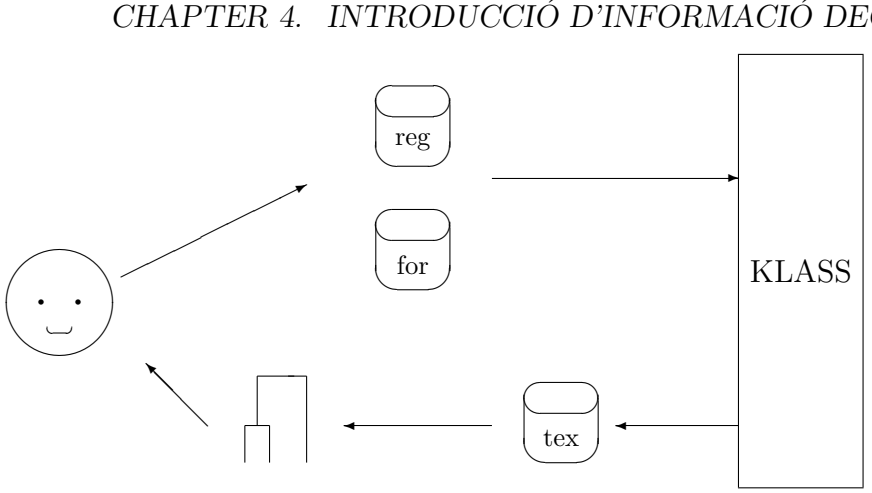


Figure 4.2: Fluxe d'informació.

l'estructura del domini, esdevé capaç de verbalitzar allò que inconscientment ja sabia, i pot explicitar-ho, transmetent més informació al sistema que desembocarà en la generació de nous resultats. Procés al final del qual, el sistema ha adquirit el coneixement necessari per a estructurar correctament el domini, i l'expert ha lograt fer explícit el coneixement que tenia d'una forma relativament còmoda.

El punt de partida és el conjunt $\mathcal{I} = \{i_1 \dots i_n\}$ d'objectes a classificar. Les passes a seguir, es detallen a continuació:

1. En primer lloc, classificar $\mathcal{I}$ per obtenir un arbre jeràrquic inicial τ^0.

2. Amb ajuda de les eines presentades al capítol **??**, es determina el tall de l'arbre $\mathcal{P}_0$ i es procedeix a la caracterització de la classificació resultant, que ha de facilitar la seva interpretació.

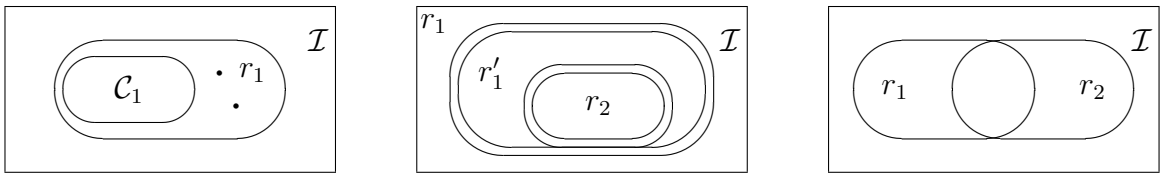
3. L'anàlisi d'aquesta partició i de les seves deficiències, ajuda a determinar un conjunt de regles inicial expressables en lògica de primer ordre $\mathcal{R}^0$, que recullin part del coneixement del(s) expert(s) sobre el domini en estudi.

4. S'entra a la passa ξ de l'algorisme ($\xi = 0$)

5. **Fase d'adquisició de coneixement de suport**:

   - Utilitzant la tècnica presentada al capítol **??**, determinació de la partició $\mathcal{P}^{\xi}_{\mathcal{R}}$ induïda per les regles d'$\mathcal{R}^{\xi}$. La part del domini sobre la que no es té informació constitueix una classe residual $\mathcal{C}^{\xi}_0$.
   - A continuació cal fer un estudi detallat dels objectes seleccionats per regles contradictòries.

     S'estableix aquí una primera fase iterativa de resolució de conflictes en que l'expert ha d'examinar quines regles contradictòries satisfan els diferents objectes, així com



(a) (b) (c)

Figure 4.3: Conflictes de les regles.

quines són les característiques d'aquests objectes que originen la contradicció. El resultat d'aquest estudi pot respondre a una de les següents situacions (que es representen a la figura **??**):

(a) Una de les regles (o ambdues) està mal especificada i selecciona alguns objectes erròniament. És precís, en aquest cas, que l'expert corregeixi la regla (o regles) incorrecta(es).

(b) Una de les regles (r_1) *implica* l'altra (r_2).
En aquest cas es té una regla general (r_1) que identifica una classe i una altra que apunta a un subconjunt d'aquesta classe (r_2). Per tal de conservar la potència d'r_2 que determina una classe més precissa que r_1, cal especificar aquesta última. De forma sistemàtica, i a falta de millor suggerència de l'expert, es pot corregir r_1 com

$$r_1' = (\mathcal{B}_1 \wedge \neg\mathcal{B}_2 \longrightarrow \mathcal{C}_1)$$

En l'aplicació de les esponges, la regla

$$\begin{aligned} r_0: \quad & ((papilas = 1) \wedge (tipo\ espicula\ principal\ tilostilo = 1) \wedge \\ & \wedge (perforante = 1) \wedge \\ & \wedge (tipo\ microsclera \in \{aster, microestilos\ y\ microxas\}) \quad \longrightarrow \mathcal{C}_0) \end{aligned}$$

identifica els objectes `C.carteri, C.schmidti, C.viridis`, mentre que la regla

$$\begin{aligned} r_1: \quad & ((tipo\ espicula\ principal\ tilostilo = 1) \wedge \\ & \wedge (tipo\ microsclera = aster) \wedge \\ & \wedge (tipo\ de\ aster = espiraster) \wedge \\ & \wedge (forma\ final \in \{masivo\ irregular, incrustante\}) \quad \longrightarrow \mathcal{C}_1) \end{aligned}$$

selecciona aquestes mateixes esponges, a més de `S.minax` i `S.cunctratix`. De fet, $r_1 \longrightarrow r_0$ i la correcció automàtica d'aquesta regla donaria lloc a una nova regla

$$\begin{aligned} r_1': \quad & ((tipo\ espicula\ principal\ tilostilo = 1) \wedge \\ & \wedge (tipo\ microsclera = aster) \wedge \\ & \wedge (forma\ final \in \{masivo\ irregular, incrustante\}) \wedge \\ & \wedge \neg ((papilas = 1) \wedge (perforante = 1)) \quad \longrightarrow \mathcal{C}_1') \end{aligned}$$

que es podria incorporar al sistema de regles en substitució d'r_1. En l'aplicació s'ha suprimit la regla r_0, que identifica un conjunt molt petit d'esponges.

(c) Les dues regles intersequen.
Aquest cas és el més difícil de resoldre, perquè suposar que la intersecció és una classe amb entitat pròpia és una hipòtesi de món tancat, i ja s'ha dit que $\mathcal{R}$ no descriu el domini de forma completa. És de nou necessària la intervenció de l'expert, que ha d'interpretar el sentit de la intersecció, i decidir com reformular $\mathcal{R}$.
En l'aplicació de les esponges, la regla r_0 també intersecava amb

$$\begin{array}{rl} r_2: & ((tipo\ espicula\ principal\ tilostilo = 1) \wedge \\ & \wedge (tipo\ microsclera = aster) \wedge \\ & \wedge (forma\ final = incrustante) \qquad\qquad \longrightarrow \mathcal{C}_2) \end{array}$$

que selecciona les esponges `C. carteri`, `C. schmidti`, `D.ornata`, `D. bistellata`, `S.cunctratix, S.minax, T.hallezi, T.mixta, T.stellata` i `T.unistellata`. No s'ha considerat generar una classe amb la intersecció `C.carteri` i `C.schmidti` que no formen un grup apart i donada la problemàtica associada a r_0 es prescindeix d'ella.

- Si l'estudi és satisfactori es passa a la fase de classificació.
- Sino es reformula $\mathcal{R}$ i es torna a iniciar la fase d'adquisició del coneixement de suport.

6. **Fase de classificació**:

- Un cop efectuada una partició ${\mathcal{P}_{\mathcal{R}}}^{\xi}$ que compleixi, *a priori*, amb les exigències de l'expert, per a cada element d'aquesta partició $\mathcal{C} \in \mathcal{P}^{\xi}_{\mathcal{R}}$ s'efectua una classificació ascendent jeràrquica i es determinen:
  - els corresponents arbres jeràrquics $\tau^{\xi}_{\mathcal{C}}$,
  - llurs prototipus $\bar{\imath}^{\xi}_{\mathcal{C}}$ i masses associades $m^{\xi}_{\mathcal{C}}$.
- Els prototipus obtinguts en la fase anterior s'afegeixen a la classe residual, segons es detalla a §**??** i es construeix un nou núvol de punts:

$$\tilde{\mathcal{I}}^{\xi} = \left\{(\bar{\imath}^{\xi}_{\mathcal{C}}, m^{\xi}_{\mathcal{C}}) : \mathcal{C} \in \mathcal{P}^{\xi}_{\mathcal{R}}\right\} \cup \left\{(i,1) : i \in \mathcal{C}^{\xi}_{0}\right\}$$

- La classificació de $\tilde{\mathcal{I}}^{\xi}$ integrarà tots els $\tau^{\xi}_{\mathcal{C}}$, $(\mathcal{C} \in {\mathcal{P}_{\mathcal{R}}}^{\xi})$ en un únic arbre de classificació τ^{ξ}.
- A partir de τ^{ξ}, utilitzar l'heurístic de §**??** per a determinar el tall de l'arbre que constituirà la partició $\mathcal{P}^{\xi+1}$ i el mètode presentat a §**??** per a caracteritzar-la. Amb aquesta informació interpretar la classificació resultant.

- L'expert haurà de confirmar que les regles que havia donat corregeixen en la forma que es pretenia la classificació anterior $\mathcal{P}^{\xi}$.

  En aquest punt, l'estudi de les Tcpp's (§**??**) de les diferents classificacions, i dels termes que més contribueixen a la *distància* entre elles, pot ser una valuosa ajuda per determinar

  - Si la classificació és útil pels objectius de l'expert, en el qual cas es pot continuar el procés, o
  - Si, pel contrari, no ho és, entrant així en un procés d'anàlisi de la classificació obtinguda, destinat a reformular el conjunt de regles i construir un $\mathcal{R}^{\xi+1}$, per tornar seguidament, fent $(\xi = \xi + 1)$, al punt 5, fase d'adquisició del coneixement de suport.

7. Un cop construïda una classificació satisfactòria, es genera l'informe definitiu i es passa a la fase de producció de resultats.

La metodologia presentada conforma una nova tècnica per a efectuar classificacions de matrius de dades heterogènies i produeix especialment bons resultats quan s'analitzen dades de dominis poc estructurats, per bé que és generalitzable a altres tipus de dominis.

4.6 Resum del capítol

Es presenta el formalisme sota el que es recull el coneixement addicional que l'expert té sobre el domini. El càlcul de predicats és el que satisfà millor els nostres objectius de màxima potència expressiva, màxima flexibilitat.

Les regles poden expressar diferents tipus de relacions entre variables i una de les possibilitats importants que ofereixen és que vinguin en funció de variables que no s'han recollit directament a la matriu de dades. En aquest cas cal adjuntar al sistema de regles la relació entre aquestes variables i les observades realment. La potència d'aquest pas radica en el fet que, d'aquesta forma, es poden combinar en un mateix sistema diferents definicions dels objectes que permeten analitzar l'estructura del núvol de punts sota diferents perspectives.

Es proposa una estratègia mixta de classificació en la que l'avaluació de les regles determina una partició inicial en el núvol de punts, a partir de la qual es procedeix a una classificació per veïns recíprocs clàssica.

La representació de les classes desenvolupada a §**??** permet la combinació dels diferents representants de la partició induïda per les regles en una única estructura final.

S'omet la hipòtesi d'homogeneïtat o completitud de les regles, aquesta darrera clàssica en els sistemes basats en el coneixement, i s'integren tots els elements en una sola jerarquia.

El mecanisme de partició-integració que es proposa pot ocasionar inversions en la jerarquia, si s'utilitza la representació estàndard. Es proposa una transformació en els

índexos de nivell de la jerarquia que eviten aquestes inversions, tot respectant el contingut semàntic que les regles han donat a la jerarquia.

A la §**??** es discuteixen alguns aspectes tècnics relatius a la implementació, els quals permeten eliminar el procés d'anàlisi sintàctica associat a la instanciació de les variables que apareixen en les regles.

Per acabar, i a modus de visió compacta de sistema d'anàlisi que es proposa, es descriu una metodologia de treball iterativa que, d'una banda, integra els elements descrits fins ara amb les eines d'ajuda a la interpretació dels resultats que es presenten en el proper capítol; de l'altra, facilita la tasca de transmissió de coneixement de suport al sistema per part de l'expert. Com ja s'ha dit, la problemàtica associada als dominis poc estructurats radica en la dificultat per a explicitar de forma completa i precissa el coneixement relatiu al domini. Amb la metodologia que es proposa, aquest procés és progressiu i resulta relativament còmode per l'expert.

Chapter 5

Ajudes a la interpretació

"L'explicació científica no consisteix en el pas de la complexitat a la simplicitat, sino en la substitució d'una complexitat menys intel·ligible per una de més intel·ligible".

LÉVI-STRAUSS

Ja en la introducció s'enuncia que les tècniques de classificació són de caràcter exploratori i tracten d'esbrinar una estructura en les dades desconeguda *a priori*, almenys parcialment, per l'usuari. Com que s'està treballant en classificació jeràrquica, no és sino a partir d'un estudi *a posteriori* de l'arbre corresponent, que s'obté la classificació definitiva.

D'acord amb la idea que la millor mesura de la qualitat d'una classificació és la seva facilitat per ser interpretada per un humà o una màquina, en aquest capítol es presenten algunes eines que ajuden a la interpretació de la classificació final, a part, està clar, de les més habituals, com pot ser la producció d'informes, que es detalla a §**??**.

En estadística, sistemes com **SPAD** permeten calcular les contribucions de cada modalitat de les variables categòriques a les diferents classes, i aquelles que resulten significatives permeten interpretar-les. També es poden calcular les contribucions dels individus a cada classe, i observar com són els individus més típics d'aquestes. Es presenta tot seguit un enfoc de la caracterització més inspirat en tècniques pròpies de la Intel·ligència Artificial.

5.1 Caracterització de les classes

Donada una partició de l'arbre jeràrquic és interessant oferir eines que ajudin a l'usuari a interpretar-la. Una d'elles és la caracterització de les classes formades. Partint, doncs, d'una classificació $\mathcal{P} = \{\mathcal{C}_1 \dots \mathcal{C}_\xi\}$ dels elements d'$\mathcal{I}$, un objectiu important és la identificació de les variables més rellevants en cada una de les classes formades; dit d'una altra manera, les que han resultat més decissives en la constitució d'aquestes i, eventualment, permetrien detectar la pertinença d'un objecte a una classe determinada, excloent-lo de les restants. En les següents seccions es fa ús de les descripcions conceptuals de les classes formades per

extreure informació de la classificació, que permeti interpretar-la.

5.1.1 Variables descriptores

S'entén que X_k és una *variable descriptora de la classe $\mathcal{C}$, en sentit estricte* per les dades estudiades, si pren un únic valor per a tota la classe:

$$x_{ik} = \begin{cases} \ell \in \Re, & X_k \text{ quantitativa} \\ c_s^k, & X_k \text{ qualitativa} \end{cases}, \forall i \in \mathcal{C}$$

Proporcionar per a cada classe $\mathcal{C} \in \mathcal{P}$ la llista de variables descriptores en sentit estricte i els valors que prenen aquestes variables, permet descriure les classes, i inclús batejar-les (en el cas de les esponges, per exemple, es pot parlar de la classe de les esponges que no tenen còrtex, o de la classe de les esponges que tenen la capa interna del còrtex de tipus microsclera... — a l'annex **??** es poden trobar les variables descriptores de les diverses classificacions de les esponges). Noti's que una mateixa variable pot ser descriptora en sentit estricte d'una classe, però no ser-ho en una altra de la mateixa partició. Es pot dir que aquest sistema de descripció està més aviat en la línia dels *weak-methods* de la Intel·ligència Artificial.

La detecció d'aquestes variables és immediata: X_k és descriptora en sentit estricte de la classe $\mathcal{C} = \{i_1 \dots i_{n_\mathcal{C}}\} \in \mathcal{P}$ si:

$$\begin{cases} s^2_{\mathcal{C}k} = 0 & , X_k \text{ és quantitativa} \\ \bar{x}_{\mathcal{C}k} \text{ és un valor compacte} & , X_k \text{ és qualitativa} \end{cases}$$

entenent per $s^2_{\mathcal{C}k} = \sum_{i \in \mathcal{C}} \frac{(x_{ik} - \bar{x}_{\mathcal{C}k})^2}{n_\mathcal{C} - 1}$ la variància dins la classe, i per $\bar{x}_{\mathcal{C}k} = \frac{\sum_{i \in \mathcal{C}} x_{ik}}{n_\mathcal{C}}$ la mitjana d'X_k dins la classe.

No obstant, en alguns casos reals de gran nombre d'elements, s'observa que per alts nivells de la jerarquia, en que les classes són molt nombroses, pràcticament no s'identifiquen variables descriptores en sentit estricte[1]. Per aquesta raó, s'ha relaxat aquest concepte en el de *variables descriptores* (en sentit ampli), que correspon a variables que, si bé no prenen un valor constant dins la classe que descriuen, hi experimenten variacions suficientment pertites com per a no ser considerades. La detecció de les variables descriptores (en sentit ampli) és a través de la variància local a cada classe de les variables numèriques, o la concentració de proporció sobre una determinada categoria per les categòriques. Així, X_k és descriptora (a nivell $(1 - \varepsilon)$) d'una classe $\mathcal{C} \in \mathcal{P}$ si

$$\begin{cases} s^2_{\mathcal{C}k} \leq \varepsilon^2 \bar{x}^2_{\mathcal{C}k}, & X_k \text{ quantitativa}^2 \\ \exists s = (1 : n_k) : f_\mathcal{C}^{k_s} \geq (1 - \varepsilon), & X_k \text{ qualitativa}^3 \end{cases}$$

[1] Això dependrà de la variabilitat de les variables observades, i no sempre serà així. A §**??** es mostra un cas real en que, malgrat la grandària de la mostra, amb particions de només cinc classes, s'identifiquen fins a un 20% de variables descriptores en sentit estricte.

i tota variable descriptora en sentit estricte d'una classe $\mathcal{C}$, és també descriptora de $\mathcal{C}$ a nivell $(1-\varepsilon)$, per qualsevol valor d'ε positiu.

Si s'agafa $\varepsilon = 0.1$, s'està considerant que un 10% de variació en el valor de la variable dins la classe és negligible, i que la variable en qüestió es pot considerar fortament determinant de la classe.

En aquest cas, el valor descriptor d'X_k per la classe $\mathcal{C}$ seria

$$v_{\mathcal{C}k} = \begin{cases} f_{\mathcal{C}}^{k_s} & X_k \text{ quantitativa} \\ \bar{x}_{\mathcal{C}k} & X_k \text{ qualitativa} \end{cases}$$

Si $\mathcal{A} = \{X_1 \dots X_k\}$ és el conjunt de variables o atributs utilitzats en la matriu de dades, es denotarà $\mathcal{A}_{\mathcal{C}}^{\varepsilon} = \{X_k \in \mathcal{A} : X_k \text{ descriptora de la classe } \mathcal{C} \text{ a nivell } \varepsilon\}$. Amb aquest procés es poden construir descripcions conceptuals de les classes de caràcter conjuntiu, del tipus

$$\bigwedge_{X_k \in \mathcal{A}_{\mathcal{C}}^{\varepsilon}} (X_k = v_{\mathcal{C}k})$$

estructures que tenen una interpretació clara, i proporcionen una idea més o menys precissa (en funció d'ε) de la *composició* de les classes.

5.1.2 Variables caracteritzadores

Ara bé, si l'objectiu de la caracterització és extreure la informació que permeti *identificar* les classes amb un mínim d'informació, llavors l'enfoc descriptiu resulta insuficient.

Que es trobi un $\mathcal{A}_{\mathcal{C}}^{\varepsilon}$ amb molts elements per una classe determinada, informa clarament sobre l'homogeneïtat d'aquesta, però no ajuda necessàriament a distingir-la de les altres classes. Seria el cas, per exemple de les classes `c56` i `c57` de la classificació amb regles per les esponges (§**??**), les variables descriptores de les quals, o bé coincideixen — és el cas de la variable *cortex*, per exemple, que es descriptora de totes dues amb un mateix valor 1 —, o bé són disjuntes — la variable *Microscleras* és descriptora de `c56` amb el valor 0, però no de `c57`, mentre que variables com *Color* descriuen `c57` amb el valor `otros`, i no `c56`—.

Per a fer *identificació*, cal estudiar els valors que pren una variable dins una classe *en relació* a les altres. Això porta a introduir alguns *conceptes nous* que es presenten tot seguit: Donada una partició $\mathcal{P} = \{\mathcal{C}_1, \dots, \mathcal{C}_\xi\}$,

Es defineix, com a *valor d'X_k propi de la classe* $\mathcal{C}$ al valor $c_s^k \in \mathcal{D}_k$ si[4]

[3]El coeficient de variació de la variable k dins la classe $\mathcal{C}$ es calcula com $\frac{s_{\mathcal{C}k}}{\bar{x}_{\mathcal{C}k}}$ i s'interpreta com la variació de la variable en relació a la seva magnitud.

[3]S'assumeix la represent¡ació extesa del valor. Evidentment, qualsevol valor compacte és descriptor per qualsevol valor d'ε.

[4]La generalització de tots aquests conceptes a variables quantitatives no s'ha dut a terme encara. La idea bàsica és fer exensives aquestes definicions a intervals de valors enlloc de categories i treballar amb els valors màxims i mínims de cada classe.

$$\left(\exists i \in \mathcal{C} : x_{ik} = c_s^k\right) \wedge \left(\forall i \notin \mathcal{C} : x_{ik} \neq c_s^k\right)$$

Una variable X_k és *parcialment caracteritzadora* de la classe $\mathcal{C} \in \mathcal{P}$ si

- té almenys un valor propi de la classe $\mathcal{C}$, per bé que en pugui compartir algun amb altres classes; s'anomenarà llavors $V_{\mathcal{C}}^k$ al conjunt de valors parcialment caracteritzadors de $\mathcal{C}$: $V_{\mathcal{C}}^k = \{c_j^k : c_j^k$ és valor propi d'X_k per la classe $\mathcal{C}\}$

Una variable X_k és *totalment caracteritzadora* de la classe $\mathcal{C}$, si

- tots els valors que pren X_k a la classe $\mathcal{C}$ són propis de $\mathcal{C}$. En aquest cas, es denota per $\Lambda_{\mathcal{C}}^k$ el conjunt d'aquests valors, els quals caracteritzen totalment la classe $\mathcal{C}$:

$$\Lambda_{\mathcal{C}}^k = \{c_j^k : c_j^k \in V_{\mathcal{C}}^k \wedge \ \forall \mathcal{C}' \neq \mathcal{C},\ c_j^k \notin V_{\mathcal{C}'}^k\}$$

Per tant, la condició de variable caracteritzadora (total o parcial) d'una classe és relativa a una partició de referència, i es pot donar o no, depenent de la distribució d'X_k en les altres classes.

Si una classe té una o més variables totalment caracteritzadores, qualsevol d'elles, per si sola, pot identificar-la.

La caracterització parcial és una generalització de la total, i resulta d'interés quan hi ha manca de caracteritzadors totals per una classe, fet que es dóna amb freqüència, com ja s'ha dit, en alts nivells d'una jerarquia construïda sobre un conjunt gran d'individus.

El procediment consistent en caracteritzar totes les classes que es pugui amb una única variable totalment caracteritzadora, genera un sistema de caracterització basat exclusivament en informació positiva, mínim per les classes totalment caracteritzades, però que pot no ser complet per les restants. El resultat d'aquest procés és un *sistema de caracterització de* $\mathcal{P}$, que es representa com:

$$\mathcal{S}_{\mathcal{P}} = \{(\mathcal{C}, X_k, \Lambda_{\mathcal{C}}^k) : \mathcal{C} \in \mathcal{P} \wedge \Lambda_{\mathcal{C}}^k \neq \emptyset\}$$

Es diu que un conjunt $\mathcal{S}_{\mathcal{P}}$ és un *sistema de caracterització completa* de $\mathcal{P}$ si

$$\forall \mathcal{C} \in \mathcal{P}\ \exists k : (\mathcal{C}, X_k, \Lambda_{\mathcal{C}}^k) \in \mathcal{S}_{\mathcal{P}}$$

I $\mathcal{S}_{\mathcal{P}}$ és mínim si donats $\mathcal{C}, \mathcal{C}' \in \mathcal{P},\ \exists k, k'$ tals que $(\mathcal{C}, X_k, \Lambda_{\mathcal{C}}^k), (\mathcal{C}', X_{k'}, \Lambda_{\mathcal{C}'}^{k'}) \in \mathcal{S}_{\mathcal{P}}$, aleshores $\mathcal{C} \neq \mathcal{C}'$.

En l'aplicació dels microordinadors (§**??**) el procés de caracterització definit és suficient, generant un sistema de caracterització complet i mínim. Pel cas de les esponges, en canvi,

el procés anterior permet caracteritzar totalment 4 de les 7 classes. Per les restants, només es poden trobar variables parcialment caracteritzadores.

El següent objectiu és completar les caracteritzacions incompletes, a fi de poder identificar les classes que falten. De fet, en el camp de l'aprenentatge automàtic supervisat, del que **ID3** i els seus successors són un exemple característic, un objectiu és aconseguir mètodes robustos que permetin la identificació de la classe a la qual pertany un objecte sabent, tan sols, el valor que prenen alguns atributs. La construcció d'aquests arbres de decisió sol tenir cost exponencial i existeixen tècniques que escurcen raonablement l'espai de cerca. Aquí s'explora una possiblitat més barata, que és l'ús d'informació condicionada de forma incremental. D'un sistema de caracterització $\mathcal{S}_\mathcal{P}$ es pot generar informació negativa també. Si $\Lambda^k_\mathcal{C}$ caracteritza totalment la classe $\mathcal{C}$, significa que en cap altra classe de $\mathcal{P}$ hi ha individus que prenguin per X_k valors de $\Lambda^k_\mathcal{C}$. D'alguna manera, $\overline{\Lambda^k_\mathcal{C}}$ es podria interpretar, doncs, com un caracteritzador parcial de qualsevol altra classe de $\mathcal{P}$. En realitat, tal afirmació involucra una hipòtesis de món tancat amb tot el risc que això comporta, però es compta amb l'expert per calibrar la validesa d'aquelles[5]

Tot parell $(\mathcal{P}, \mathcal{S}_\mathcal{P})$ tal que $\mathcal{S}_\mathcal{P}$ no és complet, determina un conjunt $\mathcal{P}' \subset \mathcal{P}$

$$\mathcal{P}' = \{\mathcal{C} \in \mathcal{P} : \ \forall k \ (\mathcal{C}, X_k, \Lambda^k_\mathcal{C}) \notin \mathcal{S}_\mathcal{P}\}$$

que és el de totes les classes que no s'han pogut caracteritzar totalment[6].

Per la partició restringida $\mathcal{P}'$ es pot calcular igualment el sistema de caracterització associat $\mathcal{S}_{\mathcal{P}'}$. Si $\mathcal{S}_{\mathcal{P}'} \neq \emptyset$, significa que hi havia solapament entre els valors d'algunes variables per les classes de $\mathcal{P}'$ i alguna classe de $(\mathcal{P} - \mathcal{P}')$. La caracterització de $\mathcal{P}'$ es pot fer utilitzant com a condicionant el fet que estem en el conjunt $(\mathcal{P} - \mathcal{P}')$, i això és possible tractant informació d'$\mathcal{S}_\mathcal{P}$ en sentit negatiu. Per exemple, en l'aplicació dels microordinadors (§**??**), la classificació proposada per Michalski té associada un sistema de caracterització que no és complet:

$$\mathcal{S}_\mathcal{P} = \{(\mathcal{C}_1, \ MP, \ \{6502, 6502A, 6502C\}), (\mathcal{C}_2, \ MP, \ HP), (\mathcal{C}_4, MP, \{Z80, 8080A\})\}$$

perquè no hi ha una caracterització total de la classe $\mathcal{C}_3$. Ara bé, fent la hipòtesi de món tancat que qualsevol ordinador amb un microprocessador que no sigui `HP`, no és de la segona classe, es pot caracteritzar totalment $\mathcal{C}_3$ a partir de la variable *Display*, que és totalment caracteritzadora, si no es considera $\mathcal{C}_2$. S'està doncs utilitzant la informació negativa "*Si* $(\ MP \neq HP)$ *aleshores el microodinador no és de la classe* $\mathcal{C}_2$" per a caracteritzar la classe $\mathcal{C}_3$, i el resultat és l'ampliació del sistema de caracterització amb l'element $(\mathcal{C}_3, \ Display, \{Terminal, Built-in\})$, subjecte, està clar, a l'esmentada hipòtesi.

[5] A la següent secció es fan algunes consideracions sobre l'àmbit d'aquests tipus d'hipòtesis, i de les precaucions necessàries per a anar més enllà de la simple descriptiva.

[6] Simètricament, $\mathcal{P} - \mathcal{P}'$ és el conjunt de totes les classes de $\mathcal{P}$ que han estat totalment caracteritzades.

Aquest mateix procés es pot repetir per totes les classes que no han estat totalment caracteritzades. El resultat és una *possible* ampliació del sistema de caracterització de $\mathcal{P}$ subjecta a l'ús d'algunes condicions en sentit negatiu, el conjunt de les quals es denota per $\mathcal{S}^{Cond}_{\mathcal{P}'}$ i es representa com:

$$\mathcal{S}^{Cond}_{\mathcal{P}'} = \{(\mathcal{C}', X_k, \overline{\Lambda^k_{\mathcal{C}}})\ tq\ \exists \mathcal{C}, k : (\mathcal{C}, X_k, \Lambda^k_{\mathcal{C}}) \in \mathcal{S}_{\mathcal{P}} \wedge \exists k', (\mathcal{C}', X_{k'}, \Lambda^{k'}_{\mathcal{C}'}) \in \mathcal{S}_{\mathcal{P}'}\}$$

En un segon pas, es determinaria $\mathcal{P}''$ (classes de $\mathcal{P}'$ que no han quedat totalment caracteritzades). L'$\mathcal{S}^{Cond}_{\mathcal{P}''}$ hauria d'heretar els condicionants d'$\mathcal{S}_{P'}$ també, i així successivament:

$$\mathcal{S}^{Cond}_{\mathcal{P}''} = \{(C'', X_k, \neg\Lambda^k_{\mathcal{C}}) : \exists \mathcal{C}, k : \quad (\mathcal{C}, X_k, \Lambda^k_{\mathcal{C}}) \in \mathcal{S}_{\mathcal{P}} \wedge$$

$$(\exists k', (\mathcal{C}', X_{k'}, \Lambda^{k'}_{\mathcal{C}'}) \in \mathcal{S}_{\mathcal{P}} \vee (\mathcal{C}, X_k, \neg\Lambda^k_{\mathcal{C}}) \in \mathcal{S}^{Cond}_{\mathcal{P}'})\}$$

En l'exemple, $\mathcal{S}^{Cond}_{\mathcal{P}'} = \{(\mathcal{C}_3, MP, \overline{HP})\}$ i $\mathcal{S}_{\mathcal{P}'} = \{(\mathcal{C}_3,\ Display,\ \{Terminal,\ Built\text{-}in\})\}$.

La caracterització de $\mathcal{P}$, es planteja, per tant, en una primera aproximació, com un procés iteratiu de la forma:

- Trobar un $\mathcal{S}_{\mathcal{P}}$ mínim.
- Determinar el conjunt de les classes que no s'han caracteritzat totalment $\mathcal{P}'$ (si $\mathcal{P}' = \emptyset$ finalitza el procés: s'ha obtingut una caracterització completa).
- Construir $\mathcal{S}^{Cond}_{\mathcal{P}'}$ i determinar $\mathcal{S}_{\mathcal{P}'}$ subjecte a $\mathcal{S}^{Cond}_{\mathcal{P}'}$.
- Si $\mathcal{S}_{\mathcal{P}'} = \emptyset$, determinar el conjunt de variables parcialment caracteritzadores de $\mathcal{P}'$ i finalitzar. En cas contrari, reduir el conjunt de classes a les que no s'han caracteritzat totalment i repetir el procés.

La caracterització final de les classes es fa amb els successius sistemes de caracterització trobats. Tampoc així es té la garantia que la caracterització sigui completa, però, incloent pel darrer $\mathcal{P}'$ totes[7] les variables caracteritzadores parcials, s'extreu el màxim d'informació sobre les classes que es pot obtenir consultant variables d'una en una. I amb això no es vol pas dir que la caracterització tingui lloc amb una sola variable en cada classe. De fet, la imbricació de condicionants donarà lloc a conjuncions tant llargues com nivells d'imbricació s'efectuin.

El resultat d'aplicar el procés descrit a la classificació de les esponges no genera millores, degut fonamentalment a la presència d'alguns valors comuns a moltes classes que, en realitat, estan representant valors il·legals[8]. Per contra, l'aplicació dels microordinadors

[7]Depenent de l'aplicació, pot tenir sentit ignorar aquelles que facin referència a percentatges massa petits (menys d'un 30%, per exemple) de la classe en qüestió.

[8]Es reserva l'ús d'aquesta informació per a treballs futurs, però es pot dir de moment que, per les esponges, la caracterització de les classes millora notablement si es prescindeix de les categories `sin_cortex`, `sin_megascleras_intermediarias`, i similars.

té una resposta molt favorable al mecanisme que es presenta: l'algorisme permet obtenir caracteritzacions completes de totes les classificacions que es presenten. Val a dir, no obstant, que aquest és un procediment basat en criteris sintàctics, i el sistema de caracterització resultant no té perquè tenir una utilitat pràctica directa, per bé que la seva interpretació semàntica pugui ser clara. De vegades, la utilitat dependrà del cost d'observació de les variables caracteritzadores. En aquest sentit esmentar el cas de les β-bandes en colesterols que, malgrat ser uns excel·lents caracteritzadors d'aquestes cadenes i tenir plè significat semàntic, s'observen amb tanta dificultat, que no es poden utilitzar com a tals. En altres ocasions, i aquest és el cas de la classificació dels microordinadors presentada per [**?**] el que es posa en dubte és el propi significat de la classificació. Si la caracterització no és interpretable, la classificació pot ser carent de sentit.

Pel que fa a la implementació, la detecció de variables caracteritzadores passa per l'estudi de les interseccions de modalitats de cada variable en les diferents classes. Això suposa, en principi, un cost $O(nmods \times nclasses)$, essent $nmods_k$ el nombre de modalitats de cada una de les variables qualitatives.

Ara bé, amb la llista de variables i la partició a caracteritzar es poden construir les estructures auxiliars:

$$X_k \longrightarrow \left(\begin{array}{c} \overset{c_j^k}{\downarrow} \\ codi_{kj} \end{array} \right) \qquad X_k \longrightarrow \left(\begin{array}{c} \overset{classe_c}{\downarrow} \\ mods_{kc} \end{array} \right)$$

on $codi_{kj}$ és un codi numèric que compacta una tercera dimensió corresponent a les classes, i que es construeix com segueix:

- Si la classe c conté elements que per la variable X_k prenen el valor c_j^k, llavors $codi_{kj}$ té un 1 a la posició c.
- En cas contrari, $codi_{kj}$ té un 0 a la posició c.

i $mods_{kc}$ representa una mena de marginal dels codis, indicant el nombre de modalitats no buides de la variable X_k a la classe $\mathcal{C}$, la complexitat del cap pitjor es redueix sensiblement.

La construcció es pot fer en paral·lel amb un cost[9]. $O(nmods \times nclasses)$. La mera consulta d'aquestes estructures permet detectar de forma molt econòmica, una variable caracteritzadora d'una classe (el seu codi és potència exacta de 2) i distingir si la caracterització és parcial o total (si el nombre de modalitats caracteritzadores de la classe per la variable en qüestió coincideix amb $mods_{kc}$ la caracterització és total, i sino és parcial).

Amb aquesta representació de la informació, s'aconsegueix que el cost del procés de detecció de les classes caracteritzadores sigui $O(nmods)$. Haver representat desplegada la

[9] $nmods = \sum_{k \in \mathcal{Q}} card\ \mathcal{D}_k$

tercera dimensió de la matriu hauria portat a un cost molt superior, ja que per cada modalitat s'hauria hagut d'estudiar a quines classes apareixia, la qual cosa és $O(nmods \times nclasses^2)$. És a dir que l'estalvi tant en espai com en temps és significatiu, especialment si es treballa amb particions de baixos nivells de l'arbre.

El procediment general és un cicle que realitza caracteritzacions condicionades successives, fins que no troba cap classe caracteritzable totalment. En aquest punt, proporciona les variables parcialment caracteritzadores de les classes restants. En el pitjor cas, a cada pas es caracteritzaria totalment una única classe i s'hauria de fer *nclasses* iteracions. El cost total del cas pitjor es calcula considerant la repetició $(nclasses - 1)$ cops de les següents passes:

- Obtenció de la descripció conceptual de cada classe: $O(nclasses)$.
- Construcció de les estructures auxiliars: $O(nmods \times nclasses)$.
- Detecció de caracteritzadores totals: $O(nmods)$
- Restringir les estructures de treball a les classes que encara no s'han caracteritzat totalment: $O(nclasses - \xi)$ en el pitjor cas són totes excepte una.

I això porta a un cost total de l'ordre d'$O(nclasses^2, nmods \times nclasses^2)$.

Fet això, únicament resta l'intent d'obtenir caracteritzacions completes a través de l'estudi de tuples de variables de grandària major que la unitat. El problema que es planteja és combinatori i, per sobre d'interaccions de més de 3 variables, es fa molt difícil d'estudiar.

5.1.3 Apunt a la generació automàtica de regles

La secció anterior preten cobrir objectius purament descriptius que extreguin el màxim d'informació útil de les dades. Ara bé, el càlcul de variables caracteritzadores pot ser el punt de partida d'un sistema de generació automàtica de regles de classificació, que es situa ja en tot un altre ordre de coses.

En aquesta àrea de recerca cal citar tres direccions de treball, que són les més actives actualment: d'una banda, la generació automàtica de regles a partir de la descripció intensional i extensional de les classes, una certa teoria del domini i l'ús de tècniques de *dividir i conquerir* [**?**]; d'una altra, la generació de regles a partir de tècniques d'**IBL** del tipus *conquerir sense dividir* (aprenentatge inductiu) [**?**]; finalment, l'aprenentatge algorísmic (PAC-learning) [**?**].

El nostre objectiu és obrir un canal que apropi l'estadística a aquest tipus de tècniques, sense pretendre, en absolut, ésser exhaustius. La nostra proposta es basa en l'ús de les variables caracteritzadores de les classes (definides a la secció anterior) en la generació automàtica de regles. En concret, de la detecció d'una variable totalment caracteritzadora

de $\mathcal{C}$ amb els valors propis de $\mathcal{C}$, $\Lambda^k_{\mathcal{C}}$, se n'extreu una regla que identifica la classe amb el mínim d'informació, de la forma:

$$(X_k \in \Lambda^k_{\mathcal{C}}) \longrightarrow \mathcal{C}$$

Els objectes seleccionats per aquesta regla coincideixen exactament amb $\mathcal{C}$ i també amb els seleccionats per la regla associada a qualsevol altra variable totalment caracteritzadora de la classe $\mathcal{C}$. D'aquesta forma, la nostra proposta es situaria en un punt intermig entre les dues primeres línies de recerca mencionades.

Si es parteix d'un sistema de caracterització completa de $\mathcal{P}$, $\mathcal{S}_{\mathcal{P}}$, el conjunt de regles que se'n deriven permetrà identificar exactament les classes de $\mathcal{P}$.

Si la caracterització completa està subjecta a un $\mathcal{S}^{Cond}_{\mathcal{P}} \neq \emptyset$, llavors la generació de regles ha d'incloure la informació condicionant:

Donats $\mathcal{S}^{Cond}_{\mathcal{P}'}$ i $\mathcal{S}_{\mathcal{P}'}$, es pot generar un conjunt de regles de la forma

$$r : \bigwedge_{[k \;\; tq(\mathcal{C},X_k,\Lambda^k_{\mathcal{C}})\in\mathcal{S}_{\mathcal{P}'}]} (X_k \in \Lambda^k_{\mathcal{C}}) \bigwedge_{[k' \;\; tq(\mathcal{C},X_{k'},\overline{\Lambda^{k'}_{\mathcal{C}}})\in\mathcal{S}^{Cond}_{\mathcal{P}'}]} (X_k \notin \Lambda^k_{\mathcal{C}}) \longrightarrow \mathcal{C}$$

Pel cas dels microordinadors, una possible regla seria:

$$(Display \in \{Terminal, Built-in\}) \wedge (MP \neq HP) \longrightarrow \mathcal{C}_3$$

Regles similars es poden construir per les variables parcialment caracteritzadores de les classes, les quals servirien per a identificar parts d'una classe. Per bé que, en aquest cas, el sistema de regles no seria complet, es pot assegurar que no genera mai errors de selecció dels objectes de la mostra. Si la mostra representa bé la població en estudi, la qualitat de selecció del sistema de regles així construït serà molt elevada. En aquest sentit, el sistema es comporta de forma molt *conservadora* en aquesta fase, generant un conjunt de regles que no siguin massa generals.

El sistema de variables totalment caracteritzadores i sense condicionants que s'obté de la classificació amb regles de les esponges, permet identificar la classe de pertinença d'un 10% de les esponges. Completant el sistema de caracterització al màxim, utilitzant informació condicionada i caracteritzadors parcials, es passa a identificar la classe del 75% de les esponges, guany substancial, malgrat que seguim tenint un sistema de caracterització que no és complet.

Pel que fa als microordinadors, la classificació que genera **KLASS** és l'única que dóna lloc a un sistema de caracterització complet sense condicionaments. La de Michalski, sense usar informació condicionada, permet identificar la classe del 76% dels objectes, i és completa si s'hi afegeix la informació condicionada. La primera de les que presenta [**?**] construïda utilitzant l'algorisme del salt mínim, passa d'un 7% al 100% incorporant la informació

condicionada al sistema de regles, mentre que la que aquests mateixos autors obtenen amb els veïns recíprocs, passa del 7% al 76% aprofitant tota la informació disponible.

I no hi ha assignacions errònies per la forma tan *conservadora* com s'han generat el sistema de caracterització i les regles associades.

A partir de la informació mostral es genera una descripció de les classes en forma de regles, que permet identificar els objectes observats. Des d'un punt de vista estadístic, no es pretén pas, de moment, anar més enllà d'aquesta descriptiva. Tot pas d'inferència que permetés el salt al món dels sistemes experts, hauria de tenir en compte que aquest procés es realitza amb informació parcial i validar les hipòtesis de món tancat que es construeixen, així com quantificar l'error d'assignació que es pot produir per mor de la qualitat de la mostra estudiada, fent ús de les proves estadístiques pertinents. Aquesta ja és tasca del procés inferent i resta fora de l'àmbit d'aquesta tesi, en la que es planteja la caracterització de les classes com una simple eina d'extracció de la informació útil continguda en la classificació, amb tal de contribuir a la seva interpretació.

5.2 Comparació entre classificacions

La comparació objectiva entre diferents classificacions d'un mateix conjunt d'individus $\mathcal{I}$ és un camp al que es dediquen molts esforços ja que tals comparacions permetran avaluar els resultats d'un mètode respecte d'altres, així com mesurar la *bondat* o *qualitat* d'una classificació davant d'un conjunt de dades.

Entre els treballs previs en aquesta direcció cal destacar [**?**], on es proposa una distància entre arbres ascendents jeràrquics, que permet fer comparacions entre classificacions d'aquest tipus o [**?**], qui proposa una distància entre particions basada en la informació mútua aportada per les classificacions a comparar.

Pel que fa a la primera, el seu ús queda reduït a l'àmbit de les classificacions jeràrquiques. La comparació de classificacions realitzades amb mètodes diferents, planteja la necessitat de definir una mesura de les diferències entre classificacions[10] pròpiament dites, prescindint de la forma com s'han generat aquestes.

La segona proposta estaria en aquesta situació, però considera que dues particions encaixades, una més fina que l'altra, són iguals i estan a distància nul·la. Des del punt de vista de la informació això és cert. Seguint aquest enfoc, doncs, la distància existent entre dos talls d'un mateix arbre jeràrquic seria nul·la sempre i, en el present contexte, interessa avaluar les diferències en el resultat que implicades pel fet de tallar a un o altre nivell de la jerarquia.

KLASS ofereix una mesura de les diferències entre dues classificacions que no depèn del mètode de classificació utilitzat, i que només és nul·la davant la igualtat estricta de les

[10]Recordi's que els termes *classificació* i *partició* d'$\mathcal{I}$ són sinònims en aquest contexte.

classificacions que es comparen. És possible, així efectuar estudis comparatius entre diferents classificacions que poden o no provenir d'un mateix arbre.

5.2.1 La taula de contingència de la partició producte

Suposant dues classificacions d'un conjunt $\mathcal{I}$, $\mathcal{P}_1 = \{\mathcal{C}_1^1 \dots \mathcal{C}_{p_1}^1\}$ i $\mathcal{P}_2 = \{\mathcal{C}_1^2 \dots \mathcal{C}_{p_2}^2\}$, es pretén definir una funció $\delta(\mathcal{P}_1, \mathcal{P}_2)$ que creixi quant més diferents siguin $\mathcal{P}_1$ i $\mathcal{P}_2$, i que valgui 0 pel cas de la igualtat.

Detectar les diferències entre les dues classificacions passa per l'estudi dels elements assignats a classes diferents en una i altra, o el que és el mateix, a l'estudi dels elements classificats d'igual forma en $\mathcal{P}_1$ i $\mathcal{P}_2$. I això representa treballar amb la partició producte:

$$\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2 = \{\mathcal{C}_{ij} : \mathcal{C}_{ij} = \mathcal{C}_i^1 \cap \mathcal{C}_j^2\}$$

El cardinal dels elements $\mathcal{C}_{ij}$, que no és més que la grandària de les interseccions entre les classes de $\mathcal{P}_1$ i de $\mathcal{P}_2$ es pot representar en forma de matriu, disposant en files la informació relativa a $\mathcal{P}_1$ i en columnes la de $\mathcal{P}_2$. Aquesta matriu, que es denota per $\mathcal{T}$, és la *taula de contingència de la partició producte*, que es denotarà abreviadament com *Tcpp*, i dóna idea de la relació existent entre les dues classificacions:

$$\mathcal{T} = \begin{array}{c} \begin{matrix} \mathcal{C}_1^2 \dots \mathcal{C}_j^2 \dots \mathcal{C}_{p_2}^2 \end{matrix} \\ \left(\begin{array}{c} \vdots \\ \vdots \\ p_{ij} = card(\mathcal{C}_{ij}) \\ \vdots \\ \vdots \end{array} \right) \end{array} \begin{array}{l} \\ \mathcal{C}_1^1 \\ \vdots \\ \mathcal{C}_i^1 \\ \vdots \\ \mathcal{C}_{p_1}^1 \end{array}$$

Fent provisionalment la hipòtesis que $p = p_1 = p_2$, $\mathcal{T}$ satisfà les següents propietats:

- $\mathcal{T}$ és una matriu quadrada
- En total, a $\mathcal{T}$ s'hi distribueixen n elements:

$$\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} p_{ij} = n \tag{5.1}$$

- Si $\mathcal{P}_1 = \mathcal{P}_2$, o bé $\mathcal{T}$ és una matriu diagonal, o existeix una reordenació de files i columnes que la fan diagonal, i a més

$$p_{ij} = \begin{cases} 0, & \mathcal{C}_i^1 \neq \mathcal{C}_j^2 \\ card(\mathcal{C}_i^1\} = card(\mathcal{C}_j^2), & \text{altrament} \end{cases}$$

- Suposar que $\mathcal{P}_1 = \mathcal{P}_2$, i $\mathcal{T}$ és diagonal comporta que

$$\sum_{i=1}^{p_1} p_{ii} = n \tag{5.2}$$

o el que és equivalent

$$\sum_{i=1}^{p_1} \sum_{\substack{j=1 \\ i \neq j}}^{p_2} p_{ij} = 0 \tag{5.3}$$

- Si $\mathcal{P}_1 = \mathcal{P}_2$, tant la taula condicionada per files com la condicionada per columnes són una permutació de la matriu identitat[11].

Si el cardinal de $\mathcal{P}_1$ i $\mathcal{P}_2$ no coincideixen, la taula de contingència de la partició producte $\mathcal{T}$, de dimensions (p_1, p_2), $p_1 \neq p_2$ es construeix de la mateixa manera, però deixa de ser quadrada i admet les següents interpretacions:

- Si per addició de grups de columnes es pot trobar $\mathcal{T}'$ que sigui diagonal o una permutació d'una matriu diagonal, llavors $\mathcal{P}_2$ és una partició més fina que $\mathcal{P}_1$ i $\mathcal{P}_1$ i $\mathcal{P}_2$ estan encaixades. Per exemple, suposant que es volen comparar dues particions d'un conjunt d'esponges de mar tals com

$$\begin{aligned}
\mathcal{P}_1 = \quad \{\mathcal{C}_1^1 = \{ \quad & Cliona_carteri, Cliona_celata, Cliona_labyrinthica, \\
& Diplastrella_ornata, Diplastrella_bistellata\}, \\
\mathcal{C}_2^1 = \{ \quad & Tylexocladus_joubini, Oxycordyla_pellita\}, \\
\mathcal{C}_3^1 = \{ \quad & Timea_chondrilloides, Timea_hallezi, \\
& Timea_mixta, Timea_unistellata\}, \\
\mathcal{C}_4^1 = \{ \quad & Tethya_aurantium, Tethya_citrina, Aaptos_aaptos, \}\}
\end{aligned}$$

$$\begin{aligned}
\mathcal{P}_2 = \{ \quad \mathcal{C}_1^2 = \{ \quad & Diplastrella_ornata, Diplastrella_bistellata\}, \\
\mathcal{C}_2^2 = \{ \quad & Cliona_carteri, \\
& Cliona_celata, Cliona_labyrinthica\}, \\
\mathcal{C}_3^2 = \{ \quad & Timea_chondrilloides, Timea_hallezi, \\
& Timea_mixta, Timea_unistellata\}, \\
\mathcal{C}_4^2 = \{ \quad & Tethya_aurantium, Tethya_citrina\}, \\
\mathcal{C}_5^2 = \{ \quad & Tylexocladus_joubini, Oxycordyla_pellita\}, \\
\mathcal{C}_6^2 = \{ \quad & Aaptos_aaptos, \}\}
\end{aligned}$$

la taula de contingència de la partició producte és:

$$\mathcal{T} = \begin{pmatrix} 2 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 \end{pmatrix}$$

[11] Per tant, quantificar les diferències entre $\mathcal{P}_1$ i $\mathcal{P}_2$ podria passar per a mesurar les distàncies entre aquestes matrius condicionades i la matriu identitat, si no fos perquè en el cas general $p_1 \neq p_2$ no queda gaire clar quina ha de ser la matriu de referència.

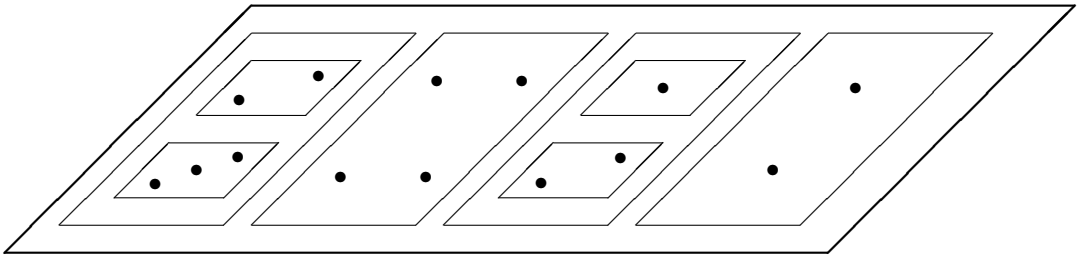


Figure 5.1: Exemple de relació entre classificacions.

L'agregació de les columnes 2 i 3, d'una banda i la de 3 amb 5, de l'altra, dóna lloc a la matriu

$$\mathcal{T}' = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \tag{5.4}$$

que esdevé diagonal per permutació de les dues columnes centrals.

La figura **??** il·lustra la relació existent entre les classificacions de l'exemple.

- Simètricament, si per addició de grups de files es pot trobar $\mathcal{T}'$ que sigui diagonal o una permutació d'una matriu diagonal, llavors $\mathcal{P}_1$ és una partició més fina que $\mathcal{P}_2$ i $\mathcal{P}_1$ i $\mathcal{P}_2$ estan encaixades.

5.2.2 Posicions centrals i posicions de perifèria

La taula de contingència de la partició producte de dues classificacions $\mathcal{P}_1$ i $\mathcal{P}_2$ idèntiques és una matriu diagonal en que els únics elements no nuls (que ocupen les posicions $(\imath, \imath)$) representen les classes coincidents en una i altra.

Suposant que $\mathcal{P}_2$ és una permutació de $\mathcal{P}_1$, les posicions ocupades pels elements no nuls de la matriu ja no són les de la diagonal pròpiament dita $(\imath, \imath)$, sino que estan permutades. És el cas de la matriu **??**, que representa dues classificacions iguals. L'equivalent de les posicions diagonals serien: (1,1), (2,3), (3,2), i (4,4), que són precisament les ocupades pels únics elements no nuls de la matriu.

En una matriu d'aquestes característiques, anomenem *posició central* a $(\imath, \jmath)$ si $p_{\imath\jmath} \neq 0$, encara que no estigui situat en la diagonal principal de la Tcpp, i *posicions de perifèria* a les restants. Les posicions centrals de la Tcpp venen a representar els nuclis comuns entre les dues classificacions que s'estudien, mentre que les de perifèria acumularien els objectes que s'han classificat de forma diferent en una i altra.

Anant més enllà es pot veure com, si no hi ha igualtat entre les dues classificacions, les posicions de perifèria no tenen perquè ser nul·les. En aquestes condicions, la identificació de les posicions centrals de la matriu s'ha de generalitzar. I és més, en el cas més general

de tots, $\mathcal{P}_1$ i $\mathcal{P}_2$ ni tan sols tindran el mateix nombre de classes. És el cas de comparar diferents talls d'un mateix arbre, o dues classificacions qualssevol, provinguin o no d'un mateix classificador.

Reprenent l'exemple de la figura **??**, i depenent del qrup que es consideri estable, els objectes desplaçats de classe entre $\mathcal{P}_1$ i $\mathcal{P}_2$ són, els 2 de $\mathcal{C}_1^2$ cap a $\mathcal{C}_2^2$, o bé els 3 de $\mathcal{C}_2^2$ cap a $\mathcal{C}_1^2$, ja que $\mathcal{C}_1^1 = \mathcal{C}_1^2 \cup \mathcal{C}_2^2$; i el mateix passa amb $\mathcal{C}_4^2$ i $\mathcal{C}_6^2$.

Sembla raonable conferir la condició de posicions centrals a les que contenen major contingent i considerar que les de perifèria (que s'interpreten com les que contenen els objectes que s'han desplaçat de classe) són les més petites. D'aquesta manera, les posicions centrals de la Tcpp aglutinen el màxim d'objectes possible que es classifiquen igual en una i altra partició.

Així les posicions (1,2) i (5,5) de $\mathcal{T}'$ serien *de perifèria*, mentre que (1,1), (2,5), (3,3) i (4,4) juguen el paper de *posicions centrals* identificant-se, aquestes últimes, pel fet que contenen el màxim de llurs files i columnes.

En general, no totes les files i columnes de la Tcpp allotjaran posicions centrals. Seria el cas de les columnes 1 i 6 de l'exemple. Aquest fenomen apareix, o bé quan dues files tenen el màxim a la mateixa columna (la qual cosa està representant que una classe de $\mathcal{P}_2$ està dividida en dues a $\mathcal{P}_1$), fet que pot donar lloc a columnes sense posicions centrals, o bé dues columnes tenen el màxim a la mateixa fila (que representa el fet contrari), amb la qual cosa poden aparèixer files sense cap posició central.

No hi ha res que impedeixi que ambdues condicions es puguin donar simultàniament en una mateixa Tcpp. D'aquesta manera, en una matriu $\mathcal{T}$ hi hauria unes poques *posicions centrals*, que correspondrien als elements més nombrosos de la partició producte.

Per tant, generalitzar la condició que permet detectar les posicions centrals de $\mathcal{T}$ és trivial i correspon a definir $(\imath, \jmath), \imath \in [1, p_1], \jmath \in [1, p_2]$ com a *posició central* de la matriu $\mathcal{T}$ si $p_{\imath\jmath} = \max_{\jmath} p_{\imath\jmath} = \max_{\imath} p_{\imath\jmath}$ i com a *posició de perifèria* a qualsevol que no satisfaci aquesta condició.

Des del punt de vista algorísmic, el procediment és extremadament senzill: es tracta d'identificar en quines posicions de la matriu hi ha elements que alhora són màxim de la seva fila i de la seva columna.

5.2.3 Una mesura de la diferència entre classificacions

A [**?**] es proposa l'ús d'un test de χ^2 per a contrastar la independència entre $\mathcal{P}_1$ i $\mathcal{P}_2$. En aquesta secció es fa una proposta més barata, molt senzilla, per a quantificar les diferències entre les dues classificacions. Resulta que l'expressió **??** és nul·la quan les dues classificacions són iguals, i creix a mesura que augmenten les diferències entre elles. Doncs bé, si es defineix

$$\wp(\mathcal{P}_1, \mathcal{P}_2) = \sum_{(\imath,\jmath)\notin\mathcal{N}} p_{\imath\jmath}$$

amb $\mathcal{N} = \{(\imath, \jmath) : (\imath\jmath) \text{ és posició central de } \mathcal{T}\}$ llavors $\delta(\mathcal{P}_1, \mathcal{P}_2)$ mesura el nombre d'objectes classificats de forma diferent a $\mathcal{P}_1$ i a $\mathcal{P}_2$; en altres paraules, quants objectes s'han de moure, com a mínim, d'una classe a l'altra per a que ambdues classificacions coincideixin. Val 0 si $\mathcal{P}_1 = \mathcal{P}_2$ i creix en augmentar les diferències entre elles. És, per tant, interpretable com una mesura de les diferències entre $\mathcal{P}_1$ i $\mathcal{P}_2$.

Recollint les propietats **??** i **??** es pot descomposar l'expressió **??** de la forma:

$$\sum_{(\imath,\jmath)\in\mathcal{N}} p_{\imath\jmath} + \sum_{(\imath,\jmath)\notin\mathcal{N}} p_{\imath\jmath} = n$$

$$\text{i per tant, } \delta(\mathcal{P}_1, \mathcal{P}_2) = n - \sum_{(\imath,\jmath)\in\mathcal{N}} p_{\imath\jmath}$$

També es pot treballar en termes relatius:

$$\delta(\mathcal{P}_1, \mathcal{P}_2) = \frac{n - \sum_{(\imath,\jmath)\in\mathcal{N}} p_{\imath\jmath}}{n} = 1 - \frac{\sum_{(\imath,\jmath)\in\mathcal{N}} p_{\imath\jmath}}{n} \tag{5.5}$$

En aquest cas $\delta(\mathcal{P}_1, \mathcal{P}_2)$ mesuraria la *proporció* d'objectes classificats de formes diferents entre $\mathcal{P}_1$ i $\mathcal{P}_2$ i seria un coeficient:

- Sempre positiu
- Sempre menor que la unitat
- Adimensional (característica sempre interessant en estudis comparatius).

Per acabar, dir que si l'expert proporciona una classificació correcta $\mathcal{P}_e$, llavors $1 - \delta(\mathcal{P}_e, \mathcal{P}) = \frac{\sum_{(\imath,\jmath)\in\mathcal{N}} p_{\imath\jmath}}{n}$ és una mesura de la *qualitat* de la classificació $\mathcal{P}$, on la unitat representa la qualitat màxima. En general, però, no existirà una $\mathcal{P}_e$, ja que, de ser-hi, segurament no caldria fer una classificació d'$\mathcal{I}$, exceptuant assajos experimentals en que es vulgui, per exemple, validar un conjunt de regles a partir d'uns pocs exemples d'estructura coneguda. A la §**??** es presenta una aplicació en aquesta línia.

5.3 Determinació del tall de l'arbre

5.3.1 Eines gràfiques

A [**?**] es proposen proves per a identificar el nivell apropiat del tall en una jerarquia. [**?**], [**?**], en canvi, s'inclinen per la construcció d'un gràfic que visualitzi l'evolució dels índexos de nivell de les successives agregacions. Les discontinuïtats pronunciades en aquest gràfic s'interpreten com agregacions "*forçades*": de classes diferents (que generen grans augments de la inèrcia intra-classe). Es recomana tallar en un d'aquests salts sempre i quant la partició resultant admeti la interpretació. De fet, el gràfic no fa més que representar un fenomen

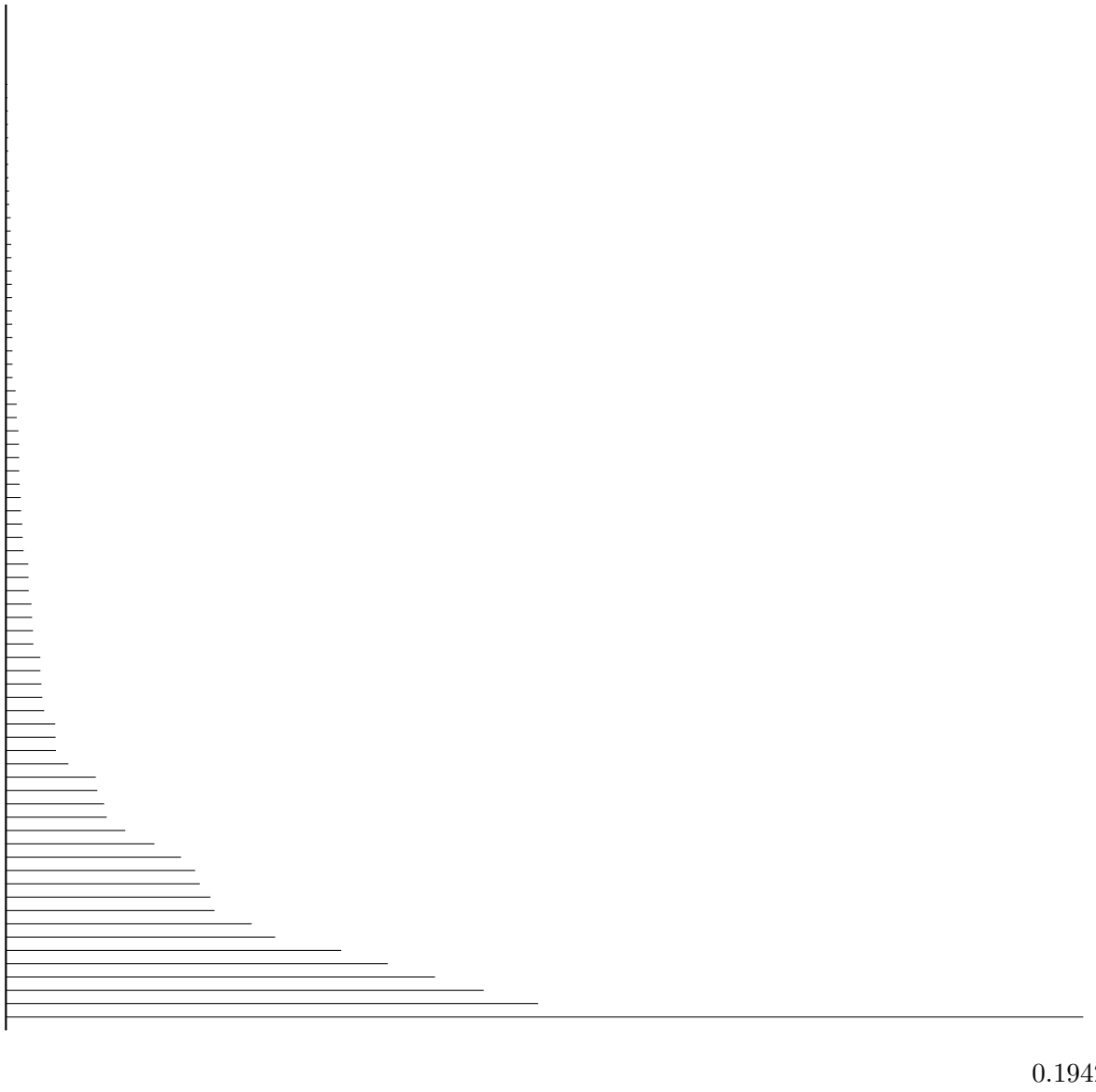


Figure 5.2: Gràfic d'índexos de nivell en agregacions successives per la classificació de les esponges utilitzant les regles.

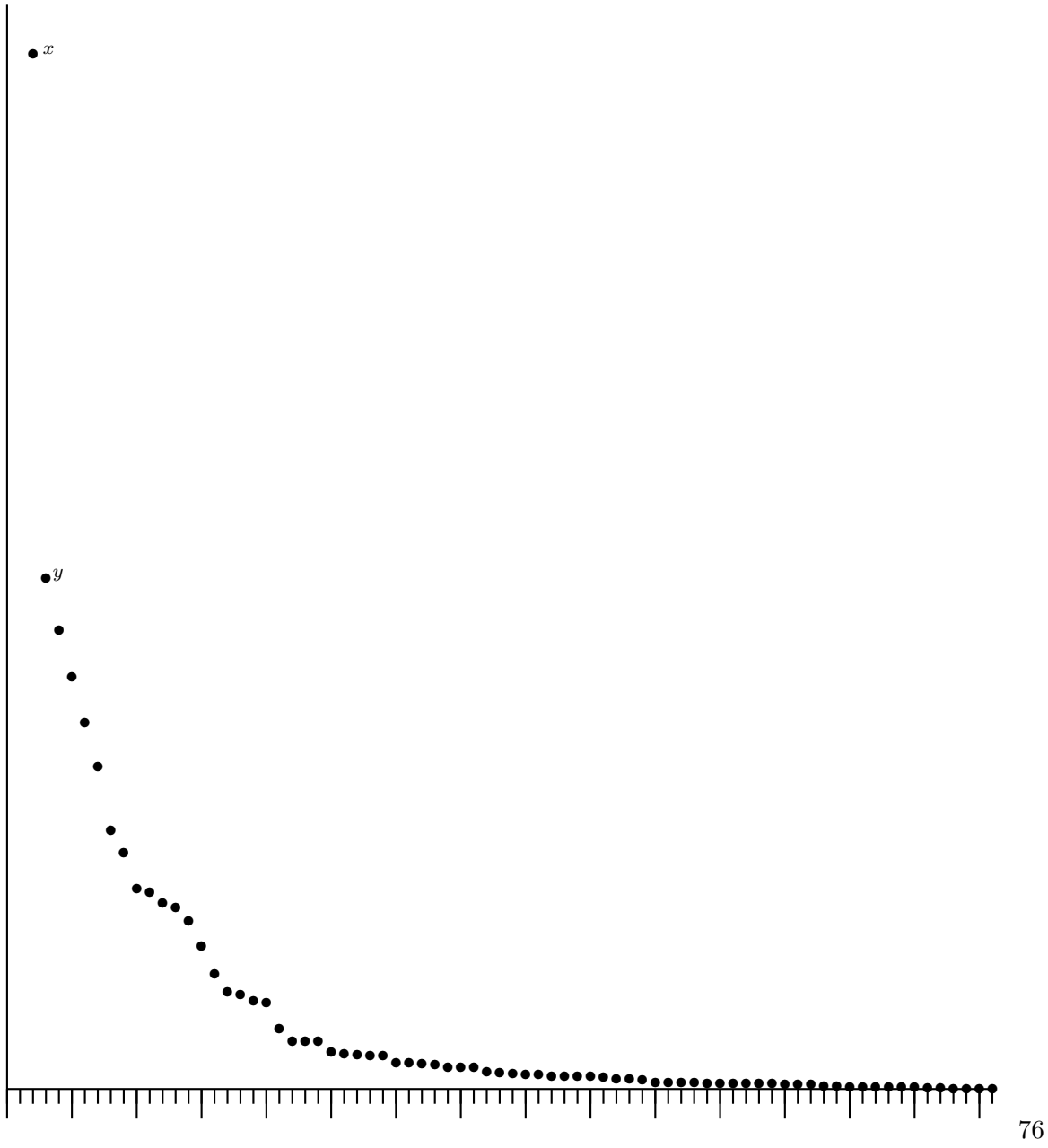


Figure 5.3: Gràfic transposat.

que també és perceptible, de forma més velada, en el dendrograma. **KLASS** adopta també aquesta solució. La figura **??** mostra el gràfic corresponent a la classificació de les esponges amb regles, i sembla suggerir un tall en 7 classes.

Considerant, però, que l'eix d'abcisses està sotmès a una compressió major que el d'ordenades, donades les proporcions d'una fulla de paper, es proposa una representació alternativa i equivalent, que consisteix bàsicament a transposar els eixos. Amb això s'accentuen encara més les discontinuïtats i canvis de tendència en els índexos. La figura **??** conté la representació equivalent del gràfic **??**. El punt x té com a ordenada l'índex de nivell de l'arrel del dendrograma (el qual es pot trobar a l'annexe **??**) `cR62`, el punt y té per ordenada l'índex de nivell del node `cR61` i així successivament.

Representant l'índex de nivell de l'arrel amb una abcissa igual a 2, i incrementant d'una unitat les abcisses de punts correlatius, l'abcissa corresponent al punt de tall és directament el número de classes en que s'ha de particionar el conjunt d'individus, la qual cosa estalvia anar comptant les línies que s'atravessen — com en la representació estàndard — o els punts que queden a l'esquerra del nivell de tall per averiguar el número de classes.

Noti's que el tall en 7 classes és molt evident en la representació que es proposa, mentre que en la clàssica podria haver-hi un lleu titubeig inicial entre els talls amb 7, 6, 5 ó 4 classes. Segons això, doncs, s'adoptaria aquesta nova representació per una major claredat.

5.3.2 Eines numèriques

D'altra banda, s'ha implementat una heurística per a determinar el nivell del tall de forma automàtica. Això permet generar informes complets sense la intervenció de l'usuari, i és purament orientatiu. Si l'expert no és capaç d'interpretar la partició que de forma automàtica genera **KLASS**, sempre pot tallar l'arbre per un altre punt[12].

L'heurística és molt senzilla, i consisteix en les següents passes:

- Es parteix d'una llista amb els índexos de nivell de cada node, ordenats de major a menor (és suficient treballar amb el 40% dels nodes, ja que per regla general no es busquen talls amb moltes classes).

- Es calcula la diferència entre l'índex de nivell d'un node i el següent i s'ordenen aquestes diferències de major a menor, associant a cada una el número de classes del tall que determinen. Això es pot fer amb un cost $O(0.08n^2)$ perquè es disposa d'una estructura amb els nodes ordenats per índex de nivell. Es prescindeix del primer element d'aquesta

[12]En el futur es preten que l'heurística involucri també criteris relatius a la caracterització automàtica de les classes. Atès que es busca major facilitat en la interpretació de la classificació resultant, un criteri per a decidir quin dels talls s'ofereix de forma automàtica seria agafar el que suposi una millor caracterització de les classes, mesurable, per exemple, com el nombre de classes totalment caracteritzades (més la proporció de les classes caracteritzades parcialment).

ordenació, si aquest recomana un tall en dues classes, ja que per norma general se'n voldran més de dues.

- Es fa el tall determinat pel major increment d'inèrcia, però s'indiquen els successors immediats per tal que l'expert pugui optar per un altre si és necessari.

Proves empíriques mostren que l'heurístic es comporta de forma excel·lent en la majoria dels casos, i que és millor que el que consisteix a calcular quocients enlloc de diferències. La raó és que, en les agregacions de baix nivell, increments absoluts d'índex de nivell molt petits, representen, en termes relatius a l'increment anterior, increments molt grans i l'ús del quocient tendeix a donar prioritat a talls de molt baix nivell en la jerarquia.

En el cas de les esponges, per exemple, l'anàlisi visual del gràfic d'índexos de nivell fa pensar, exceptuant el tall en dues classes, que un tall en 7 classes seria raonable. La recomanació resultant d'aplicar aquesta heurística és de fer 7, 3, 4, 5, 6, 9 ó 15 classes. Pel cas de la classificació sense regles, el resultat de l'heurística és que es facin 7, 4, 6, 5, 10, 12 ó 16 classes. En ambdós casos, la classificació determinada pel gràfic clàssic és la que té 7 classes.

5.4 Resum del capítol

En aquest capítol s'han presentat algunes eines d'ajuda a la interpretació de la classificació resultant.

En primer lloc s'ofereixen medis per a obtenir una representació gràfica dels índexos de nivell de l'arbre, a través de la qual es pot decidir el nivell per on tallar-lo. Amb l'intent de contribuir a la presa d'aquesta decisió, s'ha dissenyat una heurística que selecciona un nombre prefixat de talls i els ordena de millor a pitjor en funció de la "*resistència*" d'aquests talls a reduir el nombre de classes.

Com a contribució en la interpretació de les classes es presenten eines de comparació i caracterització d'aquestes.

Pel que fa a la comparació entre classificacions, en la secció **??** es desenvolupa una possible mesura de la diferència entre dues classificacions, basada en la construcció de la *taula de contingència de la partició producte* de $\mathcal{P}_1$ i $\mathcal{P}_2$. Aquesta mesura, $\delta(\mathcal{P}_1, \mathcal{P}_2)$ s'interpreta com la proporció d'objectes que es classifiquen d'igual forma en una i altra partició, i requereix el càlcul dels nuclis estables (o *posicions centrals*) de la Tcpp corresponent.

En darrer lloc, la caracterització de les classes s'aborda des de dos punts de vista. D'una banda es descriu un sistema per a identificar les *variables descriptores*, fixat un cert nivell ε d'una classe, les quals permeten fer-se una idea de la *composició* d'aquesta i corresponen a un enfoc maximalista, en el sentit que la descripció d'una classe es fa amb el màxim nombre de variables possible de valor constant (o *quasi-constant*) a ella. Aquest mecanisme aniria en

la línia dels *weak-methods* propis de la Intel·ligència Artificial, i es basa en un principi similar al que fan els sistemes estadístics clàssics, que consideren el màxim nombre de modalitats amb contribució significativa a una classe per a poder-la interpretar.

Plantejada la insuficiència d'aquest enfoc per a generar descripcions que permetin *identificar* les classes, es defineix un procediment de *caracterització* que permet detectar conjunts mínims, ara sí, de variables que permetin distingir una classe de les altres, mecanisme que pot incidir directament en processos de generació automàtica de regles.

S'introdueix, amb aquest fi, el concepte de *variable totalment caracteritzadora* d'una classe $\mathcal{C}$, que és la que pren, a la classe $\mathcal{C}$, valors específics i diferents de les altres classes. A partir d'aquí i analitzant la completitud de les caracteritzacions, s'arriba a un procediment que utilitza informació condicional de forma imbricada, amb un cost $O(nmods \times nclasses)$ i pot generar descripcions conjuntives de llargada variable, formades, sempre, d'un i només un terme positiu, acompanyat d'un seguit de condicionants negatius.

Donada l'estructura d'aquestes descripcions, no es pot garantir la completesa de la caracterització resultant, per a la qual cosa caldria repetir el procés estudiant parelles, ternes, en general n-tuples, de variables que poguessin caracteritzar totalment les classes restants. El problema es torna combinatori i la seva complexitat creix molt. No obstant, la implementació que es proposa redueix notablement el cost de calcular aquestes caracteritzadores (es passa de $O(nvars^n \times nmods_k^{2n} \times n_classes)$ a $O(nvars^n \times nmods_k^n \times nclasses)$).

La generació automàtica de regles a partir dels sistemes de caracterització definits és directa i és una contribució, de la que no en coneixem antecedents, a l'apropament de l'estadística cap a aquesta àrea de recerca.

Chapter 6

Aplicacions (I). Esponges

"Perquè la diferència és principi general,
fa del gènere moltes espècies".
R.LLULL

6.1 Introducció

Les esponges marines presenten greus problemes a l'hora de ser organitzades en una taxonomia i, actualment, hi ha diverses classificacions d'aquests éssers vius acceptades per diferents escoles.

La classificació d'un conjunt qualsevol d'espècies ja és, en si mateixa, objecte d'estudi per aquells que es dediquen a la sistemàtica, però l'interés d'una classificació estable per a les esponges radica principalment en el fet que el tàxon constitueix el marc comú a través del qual es pot establir la comunicació entre les diferents disciplines que treballen sobre un mateix grup, i a través del qual es poden fer estudis comparatius.

El creixent interés per la identificació d'espècies del bentos marí en la producció de substàncies d'aplicació diversa, és una altra de les raons que fan que una classificació estable de les esponges resulti útil i necessària.

L'any 1989, el departament de **Llenguatges i Sistemes Informàtics** de la **UPC** inicià una estreta col·laboració amb el **Grup de Biologia Marina** del **Centre d'Estudis Avançats de Blanes (CSIC)**, interessat en estudiar aquests organismes tant des d'un punt de vista sistemàtic, com pel que fa a la seva organització en comunitats, o com a font productora de substàncies químiques actives, com a models biològics, ... i un dels aspectes de la col·laboració fou tractar d'obtenir una classificació d'aquests Porífers per medis informàtics, en principi, més objectius que els purament humans.

El primer pas d'aquest procés va ser restringir l'estudi a espècies pertanyents a l'ordre *Hadromerida* de la classe *Demospongiæ* a partir d'una mostra representativa d'aquest Ordre. La classificació que en aquell moment s'admetia en la comunitat biològica, a nivell de Família, era la següent (als annexes **??** es troba més informació sobre aquesta classificació):

F.*Tethyidæ* : `Tethya, Aaptos.`

F.*Polymastiidæ* : `Polymastia`, `Quasilina`, `Weberella`, `Tentorium`, `Ridleya`, `Trichostema`, `Trachyteleia, Tylexocladus, Spinularia, Proteleia`, `Sphærotylus`

F.*Suberitidæ*: `Terpios`, `Rhizaxinella`, `Suberites`, `Prosuberites`, `Laxosuberites`, `Pseudosuberites`

F.*Spirastrellidæ* : `Spirastrella`

F.*Clionidæ* : `Cliona`

F.*Timeidæ* : `Timea, Diplastrella`

F.*Stylocordylæ* : `Oxycordyla, Stylocordyla`

Classificació que despertava certa controvèrsia pel que fa a la ubicació d'algunes espècies (o gèneres) com per exemple `Suberites caminatus`, espècie conflictiva que al llarg de la bibliografia ha anat canviant de gènere[1] perquè és força diferent de la resta d'espècies de F.*Suberitidæ*. O com el cas de la F.*Tethyidæ* que conté dos gèneres conceptualment molt distants (altres punts de desacord es poden trobar a [**?**]).

En una primera etapa de la investigació, les dades foren tractades per una eina d'adquisició del coneixement (**LINNEO**) donant lloc a la classificació que es troba a l'annex **??**. Com a conseqüència de l'anàlisi d'aquests resultats, [**?**] presenta una nova proposta, de la qual se'n dóna més detall a l'annex **??**:

F.*Semipolymastiidæ*: `Tentorium, S.caminatus, Spinularia, Trachyteleia, Tylexocladus`

F.*Spirastrellidæ* : `Spirastrella, Diplastrella, Cliona, Alectona`

F.*Suberitidæ*: `Suberites`, `Laxosuberites`, `Prosuberites`, `Pseudosuberites`, `Rhizaxinella`, `Terpios`

F.*Stylocordylæ* : `Oxycordyla, Stylocordyla`

F.*Polymastiidæ* : `Polymastia, Sphærotylus, Proteleia`

F.*Quasilinidæ*: `Quasilina, Weberella, Ridleya`

F.*Tethyidæ*: `Tethya`

F.*Timeidæ* : `Timea`

F.*Trichostematidæ*: `Trichostema`

F.*Aaptidæ* : `Aaptos`

[1] En algun moment apareix amb el nom de `Tentorium caminatus`.

Les diferències entre els resultats de **LINNEO** i la proposta [**?**] són degudes a un raonament *a posteriori* sobre els resultats de la primera per part d'experts en el tema. Com es pot veure, aquesta proposta comporta algunes modificacions respecte de la classificació tradicional com, per exemple, la composició de la F.*Timeidæ*, entre d'altres. Partint de la hipòtesi que l'espícula principal de les `Diplastrella` correspon a la fusió de dues del tipus de les `Timea`, la classificació tradicional presenta en F.*Timeidæ* els gèneres `Timea i Diplastrella`, mentre que `Spirastrella` forma una Família apart. A [**?**] s'argumenta com es pot considerar que l'espícula principal de les `Diplastrella` té relació amb la de `Spirastrella`, i es proposa una família per aquests dos gèneres juntament amb els de la F.*Clionidæ* tradicional, i una altra per `Timea`[2].

6.2 Les dades

La matriu de dades a analitzar conté 76 espècies pertanyents a 27 gèneres de l'Ordre *Hadromerida* de la Classe Demospongiæ. El criteri de selecció de la mostra a estudiar fou que les files de la matriu de dades final continguin el màxim d'informació útil, és a dir, s'inclogueren amb preferència espècies amb pocs valors mancants en llurs descripcions.

L'elecció de les variables a utilitzar és, pel cas que ens ocupa, més difícil i subjectiva que en altres grups d'invertebrats degut, sobretot, al fet que les esponges tenen una resposta altament plàstica a canvis en el medi, i moltes característiques morfològiques poden variar, per una mateixa espècie, en relació a ell.

Pel grup concret que s'està estudiant, els experts han descrit les espècies seleccionant un conjunt de 45 característiques morfològiques que fan referència, a grans trets, a còrtex, esquelet, característiques externes i característiques ecològiques.

Aquestes variables són de naturalesa diversa, és a dir, quantitatives i qualitatives, essent algunes d'elles binàries. Als annexes apareixen detallades (**??**).

Les esponges són un exemple clar de la necessitat d'incloure tractament de valors mancants en els mòduls estadístics: No és que el color sigui una característica massa important d'aquests animals. De fet, segons el lloc on es trobi i altres factors, una mateixa espècie pot canviar de color, però els espongiòlegs utilitzen aquesta característica, juntament amb d'altres, per a identificar l'espècie a la qual pertany una certa esponja. Algunes de les espècies incloses a la mostra, o bé estan extingides, o bé viuen prou amagades com per a no trobar-ne gaires, i les que es troben no mantenen les seves característiques massa temps fora del seu hàbitat natural. Tot això fa que, d'algunes espècies, els pocs exemplars que

[2]Aquesta modificació es recolza en el fet que **LINNEO**, executat amb radi 12, presenta una classe aïllada per les esponges del gènere `Timea`, però si s'estudia la distància que presenten les espècies d'aquest gènere a altres classes, es veu que la classe que conté `Spirastrella i Diplastrella` inclou totalment la de les `Timea`. Cal recordar que **LINNEO** no genera una partició dels objectes, sinó que les classes que proposa poden intersecar entre elles.

en queden romanen conservats dins pots de formol en museus, laboratoris i altres llocs on els estudiosos recórren per saber d'aquests animals. El formol ha alterat el color original d'aquests exemplars, de manera que, en els casos en que no es disposa de prou bibliografia suplementària, no hi ha forma de definir aquesta característica de l'esponja. La casella corresponent de la matriu de dades contindrà un interrogant (**?**).

6.3 Inici de l'estudi

Pel que fa a **KLASS**, aplicant la metodologia proposada a **??** es comença efectuant una primera classificació de les dades utilitzant la distància mixta i el criteri de Ward. A l'annex **??** hi ha l'extracte de l'informe corresponent. L'arbre jeràrquic suggereix, i així ho confirma el gràfic dels índexos de nivell, un tall bé en 2, bé en 7 classes.

Donat que es pretén obtenir una caracterització a nivell de Família i, presumiblement, n'hi haurà més de 2, s'opta per estudiar la segona d'elles, tenint en compte, sempre, que la interpretació de la classificació proposada finalment és el criteri de validació acceptat actualment per la comunitat científica.

Cap, aquí, fer una anàlisi a diferents nivells. Sense pretendre fer un estudi profund des del punt de vista biològic, la qual cosa és força complexa, i resta fora de l'àmbit d'aquest estudi, una primera passa fóra caracteritzar les classes resultants, per a després comparar-les amb els resultats de **LINNEO** , amb la classificació tradicional o amb la proposta [**?**].

6.3.1 Descripció de les classes

Referent a la descripció de les classes i utilitzant el mecanisme descrit a §**??**, és remarcable l'elevat nombre de variables descriptores en sentit estricte que es detecten a cada classe, prova clara de l'homogeneïtat de les classes per una banda, i de la relació entre alguns atributs per l'altra. No és, doncs, necessari recórrer a descripcions amb $\varepsilon > 0$, i es pot treballar de forma estricta. Tal com s'ha exposat a **??**, el sistema proporciona la llista de variables descriptores en sentit estricte de les classes de la partició donada. D'acord amb això, les variables descriptores de la classe |classe64|, per exemple, són a la taula **??**.

Amb aquesta llista s'obté informació suficient per a interpretar la classe que ens ocupa: Es tracta d'una classe formada per esponges amb còrtex, espícula principal sempre de tipus *estil-tilòstil, sense microscleres*, amb *papilles* i que es fixen sobre substrat *dur*. Hi ha tota una sèrie de variables que donen informació en sentit negatiu sobre la classe, com pot ser que no són esponges perforants ni ténen pseudoarrels; el còrtex no és només d'espícules tangencials, ni allotja cossos extranys, ni té feixos d'espícules en pompó; no presenten mai microscleres, i en conseqüència, no té àster ni cap de les seves variants (ni esferàster, ni diplàster).

Descripcions similars poden fer-se per amb les variables descriptores de les classes restants, les quals es relacionen a l'apèndix.

	Variable	valor típic
$\|C_{64}\|$	Córtex	1
	Córtex sólo de espículas tangenciales	no
	Cuerpos extraños en el córtex	no
	Haces de espículas principales en pompón en el córtex	0
	Tipo espícula principal diactina tuberculada	0
	Tipo espícula principal estilo	0
	Tipo espícula principal estilos 2 tamaños	0
	Tipo espícula principal estilo tilostilo	1
	Tipo espícula principal estrongiloxa	0
	Tipo espícula principal oxas	0
	Tipo espícula principal tilostilo	0
	Microscleras	0
	Tipo microsclera	sin microscleras
	Áster	0
	Diámetro esferáster	sin esferáster
	Tipo de áster	sin áster
	Tipo de diplaster	sin diplasteres
	Tipo de esferáster	sin esferáster
	Papilas	1
	Aloja cangrejo ermitaño	0
	Perforante	0
	Pseudoraíces	0
	Sustrato	duro

Table 6.1: Variables caracteritzadores de la |classe64|.

6.3.2 Estudi comparatiu

La comparació dels resultats de **KLASS** amb les altres classificacions no es pot fer sinó en base a les descripcions extensionals de les classes, ja que únicament **KLASS** i **LINNEO** proporcionen descripcions a un altre nivell. Es disposa, doncs, de quatre classificacions sobre el domini en estudi: la tradicional, la proposta [**?**], la resultant de **LINNEO** i l'obtinguda amb **KLASS** . La taula **??** indica les diferències existents entre aquestes classificacions dos a dos utilitzant la mesura exposada a **??** i a la taula **??** apareixen les Tcpp's d'aquestes quatre classificacions comparades entre elles.

Observant la matriu **??** es pot veure com la classificació obtinguda amb **KLASS** s'assembla més a la proposta [**?**] que a la tradicional. I encara més als resultats de **LINNEO** .

D'altra banda, **LINNEO** està també més a prop de [**?**] que de la classificació tradicional, però els experts encara van moure un 19% dels objectes de les classes que proporcionava, per a constituir [**?**].

Comparant els resultats de **KLASS** amb la classificació tradicional es veu que hi ha entre elles un 0.36% de diferències. És a través de les Tcpp's de la taula **??** que es poden localitzar aquestes diferències:

- `Aaptos aaptos` no apareix a la mateixa classe que el gènere `Tethya`, que és un dels punts apuntats per [**?**], però surt classificada amb espècies de la F.*Polymastiidæ*.
- La F.*Suberitidæ* es detecta per complet, a excepció de `Suberites caminatus`. El conflicte associat a aquesta espècie ja s'ha comentat anteriorment. És precisament amb espècies del gènere `Tentorium` que s'uneix aquesta esponja.
- Si bé hi ha una subdivisió clara entre els gèneres `Cliona` i `Alectona` per una banda, `Timea` per l'altra i `Spirastrella` i `Diplastrella` per una tercera, en la línia de la proposta [**?**], **KLASS** no arriba a fer una divisió clara entre ells per alts nivells de la jerarquia i, en fer el tall de l'arbre en set classes, existeix confusió entre tots ells que es consideren com una sola. Com ja s'ha dit, són també, aquests, gèneres la ubicació dels quals està subjecta a discusió (veure [**?**]).
- Un següent punt de difícil interpretació és la forma com s'han distribuït les espècies de la F.*Polymastiidæ*, en les classes `c43, c64` i `c62`. Els experts no veuen clar amb quin criteri s'ha efectuat aquesta subdivisió.

Respecte de la classificació tradicional, la classificació ascendent jeràrquica resulta, doncs, no massa satisfactòria.

Tampoc hi ha un apropament clar a la proposta de [**?**]. Els punts de divergència venen produïts bàsicament pels mateixos objectes:

- La unió en una sola classe de les F.*Spirastrellidæ* i F.*Timeidæ*, la primera de les quals agrupa, per la classificació [**?**], els gèneres `Spirastrella, Diplastrella, Cliona` i `Alectona`.

Tradicional

Klass	T	P	S	Sp	C	Ti	St
c72	0	0	19	0	0	0	0
c71	0	0	0	2	6	7	0
c61	2	0	0	0	0	0	0
c62	0	18	0	0	0	0	0
c43	1	10	1	0	0	0	0
c38	0	0	0	0	0	0	2
c64	0	8	0	0	0	0	0

KLASS

Domingo	T	P	S	Sp	C	Ti	St
Spo	0	5	1	0	0	0	0
Sp	0	0	0	2	6	2	0
S	0	0	19	0	0	0	0
St	0	0	0	0	0	0	2
P	0	23	0	0	0	0	0
Q	0	6	0	0	0	0	0
T	2	0	0	0	0	0	0
Ti	0	0	0	0	0	5	0
Tr	0	2	0	0	0	0	0
A	1	0	0	0	0	0	0

	c72	c71	c61	c62	c43	c38	c64
Spo	0	0	0	0	6	0	0
Sp	0	10	0	0	0	0	0
S	19	0	0	0	0	0	0
St	0	0	0	0	0	2	0
P	0	0	0	16	1	0	6
Q	0	0	0	0	4	0	2
T	0	0	2	0	0	0	0
Ti	0	5	0	0	0	0	0
Tr	0	0	0	2	0	0	0
A	0	0	0	0	1	0	0

Linneo	T	P	S	Sp	C	Ti	St
C1	0	0	0	2	4	7	0
C2	0	0	0	0	0	0	2
C3	0	18	0	0	0	0	0
C4	0	5	0	0	0	0	0
C5	0	0	17	0	0	0	0
C6	0	0	2	0	0	0	0
C7	2	0	0	0	0	0	0
C8	0	0	0	0	2	0	0
C9	1	5	1	0	0	0	0
C10	0	2	0	0	0	0	0
C11	0	6	0	0	0	0	0

	c72	c71	1	c62	c43	c38	c64
C1	0	13	0	0	0	0	0
C2	0	0	0	0	0	2	0
C3	0	0	0	15	1	0	2
C4	0	0	0	1	0	0	4
C5	17	0	0	0	0	0	0
C6	2	0	0	0	0	0	0
C7	0	0	2	0	0	0	0
C8	0	2	0	0	0	0	0
C9	0	0	0	0	7	0	0
C10	0	0	0	2	0	0	0
C11	0	0	0	0	4	0	2

Domingo 1990

Linneo	Spo	Sp	S	St	P	Q	T	Ti	Tr	A
C1	0	8	0	0	0	0	0	5	0	0
C2	0	0	0	2	0	0	0	0	0	0
C3	0	0	0	0	18	0	0	0	0	0
C4	0	0	0	0	5	0	0	0	0	0
C5	0	0	17	0	0	0	0	0	0	0
C6	0	0	2	0	0	0	0	0	0	0
C7	0	0	0	0	0	0	2	0	0	0
C8	0	2	0	0	0	0	0	0	0	0
C9	6	0	0	0	0	0	0	0	0	1
C10	0	0	0	0	0	0	0	0	2	0
C11	0	0	0	0	0	6	0	0	0	0

Table 6.2: Tcpp's de les quatre classificacions a comparar.

	Tradicional	**Domingo 1990**	**LINNEO**
KLASS	.36	.27	.21
Tradicional		.25	.39
Domingo 1990			.19

Table 6.3: Diferències entre les quatre classificacions.

- La dispersió poc interpretable dels gèneres de les F.*Quasilinidæ*, F.*Polymastiidæ*, F.*Semipolymastiidæ* i F.*Trichostematidæ*
- La ubicació d' `Aaptos aaptos`

Cal dir, no obstant, que tampoc amb **LINNEO** milloren els resultats pel que fa al primer i tercer punt.

Pel que fa al segon, en la classificació tradicional, els gèneres `Polymastia, Sphærotylus`, `Weberella`, `Tentorium, Quasilina, Ridleya, Spinularia, Trichostema, Trachiteleia, Tyxelocladus` i `Proteleia` formen la família F.*Polymastiidæ*. Aquest és un grup molt ampli en el que hi ha gèneres que s'assemblen realment molt, i també alguns (`Weberella, Tentorium, Quasilina i Trichostema`) que, compartint gran part de les seves característiques amb les del grup, són un xic més diferents del conjunt.

Adduint raons d'aquest tipus, la proposta [**?**] presenta una escissió de la Família *Polymastiidæ* clàssica en les Famílies F.*Polymastiidæ*, F.*Semipolymastiidæ*, F.*Quasilinidæ* i F.*Trichostematidæ*, mentre que, per la seva banda, les classes **C3** i **C4** (gèneres `Polymastia` i `Sphærotylus`), **C9** (gèneres `Tylexocladus, Trachiteleia, Tentorium` i `Aaptos`), **C10** (`Trichostema`) i **C11** (`Weberella, Quasilina` i `Ridleya`) de **LINNEO** es reparteixen els gèneres d'aquesta família.

Segons els experts, `Weberella` i `Quasilina` presentarien un còrtex més senzill que la resta de gèneres d'aquest grup, i juntament amb `Ridleya`, constitueixen F.*Quasilinidæ* (que **LINNEO** detecta amb aquesta mateixa forma — **C11** —), mentre que `Tentorium` i `Trachiteleia` tenen per papil·les simples protuberàncies de l'estructura externa, a diferència de la resta, les papil·les dels quals són d'estructura més el·laborada. D'aquí que [**?**] opti per presentar aquests gèneres en la F.*Semipolymastiidæ*, juntament amb `Tylexocladus, Spinularia` i `Suberites caminatus`. **LINNEO** detecta una classe que conté exactament aquests gèneres més l'espècie `Aaptos aaptos` (**C9**).

KLASS, en canvi, presenta una dispersió notable en les espècies del gènere *Polymastia*. De fet, aquest és un dels gèneres més ben estudiats a la literatura i, per tant, un dels que compta amb més informació disponible. D'una banda, les espècies d'aquest gènere determinen la presència de gran part dels descriptors utilitzats com a variables de l'anàlisi, els quals, únicament resulten rellevants per aquesta Família. De l'altra, aquest també és el gènere amb més representació en la mostra d'estudi. Tot això pot estar exercint un pes que deformi els resultats. Així doncs, **KLASS** presenta tres classes contenint espècies de F.*Polymastiidæ* en una divisió poc clara, que als experts els resulta difícil d'interpretar.

Com a darrer comentari, dir que l'arbre enregistra dues espècies amb distància 0 entre elles, és a dir, idèntiques. No hi ha consciència d'haver introduït dues files d'iguals components a la matriu de dades. Observant les files de la matriu corresponents a `Quasilina brevis` i `Quasilina ricardii`, s'aprecia com aquestes dues esponges es distingeixen únicament pel color, mancant en el primer cas, i de la categoria *altres* en el segon. La

distància mixta entre aquestes dues espècies és, únicament la distància en la component del *color*, i es manté despreciable, malgrat la substitució que es fa del valor mancant.

6.4 Ús d'informació declarativa

En la classificació general, executada sota el criteri de Ward i la distància mixta, s'identifiquen les principals famílies distingides pels experts però, tot i obtenint resultats millors que amb altres mètriques, algunes espècies i gèneres resten mal classificats. És el cas de les `Quasilina`, per exemple, o altres gèneres citats a l'apartat anterior.

D'acord amb la metodologia presentada a §**??**, és possible aprofitar coneixement parcial que els espongiòlegs tenen sobre les Demospongiæ, si s'expressa mitjançant regles. En aquesta etapa de l'estudi, es procedeix a fer ús d'aquesta informació addicional. Els experts proporcionen un conjunt $\mathcal{R}$ de quatre regles que es llisten a l'annex **??**, i que tenen aspectes com el següent:

$$\begin{aligned} r_1: \quad & ((tipo\ espicula\ principal\ tilostilo = 1) \wedge \\ & \wedge (tipo\ microsclera = aster) \wedge (tipo_de_aster = espiraster) \wedge \\ & (forma_final \in \{masivo_irregular, incrustante\}) \qquad \longrightarrow \mathcal{C}_1) \end{aligned}$$

la qual cosa indica que

> *esponges amb espícula principal de tipus tilòstil i microscleres de tipus "*espiraster*" (que és un tipus d'*aster*), i que en estat adult presenten forma de "*massiu irregular*" o "*incrustant*".*

pertanyen totes a una mateixa classe.

Les quatre regles d'$\mathcal{R}$ determinen un $\mathcal{P}_\mathcal{R} = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$, i indueixen una partició sobre la mostra estudiada $\mathcal{P}^*_\mathcal{R} = \{\mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_4\} \cup \mathcal{C}_0$, on la classe residual conté 54 esponges que no han satisfet cap regla, i 5 que satisfan regles contradictòries. S'observa que $\mathcal{R}$ conté algunes regles que no seleccionen cap esponja (seria el cas de $\mathcal{C}_2$)[3], però això no representa cap dificultat pel desenvolupament de l'estudi. A l'annex referit es pot trobar també la partició induïda per les regles. Les classes de $\mathcal{P}_\mathcal{R}$ contenen espècies dels següents gèneres:

C_1 `Cliona`

C_3 `Timea` i `Diplastrella`

C_4 `Rhizaxinella` i `Suberites`

Fet això es pot ara passar a la classificació local a les classes induïdes per les regles generant una descripció conceptual de cada una i la corresponent estructura jeràrquica (de les que se'n troben fragments als annexes).

[3]De fet, no és que r_2 no seleccioni cap objecte, sinó que tots els que selecciona van a la classe residual per satisfer simultàniament la regla r_3.

De fet, les descripcions conceptuals representen objectes *intermedis* de les classes, que poden ser ficticis. Per exemple, la classe $\mathcal{C}_3$ inclou les espècies `D.bistellata, D.ornata, T.hallezi, T.mixta, T.stellata, T.unistellata`.

La seva descripció conceptual és el centre de gravetat d'aquests sis elements i té el següent aspecte:

Capas del córtex = Sin córtex
⋮

Microsclera = 1

Tipo microsclera = Àster
⋮

Superficie= ((Lisa, 0.333), (Rugosa, 0.1667), (Totalmente híspida, 0.5))

Aquestes components poden ser interpretades en el següent sentit: cap dels objectes de la classe $\mathcal{C}_3$ té còrtex [...]; tots ells presenten microscleres de tipus àster. [...] Quant a la superfície, un 33% de les esponges de la classe és de superfície llisa, la meitat la té totalment híspida, i la resta és de superfície rugosa.

Les descripcions conceptuals d'aquestes classes, que són nous objectes extesos ($\bar{\imath}_{\mathcal{C}_1}$, $\bar{\imath}_{\mathcal{C}_3}$ i $\bar{\imath}_{\mathcal{C}_4}$), més les esponges de la classe residual, afectats dels seus corresponents pesos, formen el conjunt, $\tilde{\mathcal{I}} = \{(\bar{\imath}_{\mathcal{C}_1}, 1), (\bar{\imath}_{\mathcal{C}_3}, 3), (\bar{\imath}_{\mathcal{C}_4})\} \cup \{(i, 1) : i \in \mathcal{C}_0\}$. Construït aquest conjunt es pot passar a classificar-lo seguint el procediment ordinari. La figura de l'informe 4 (annex **??**) representa l'arbre ascendent jeràrquic de la classe residual ampliada amb els representants de les classes induïdes per les regles[4]. Un primer cop d'ull d'aquesta classificació mostra com els grups principals detectats en la classificació sense regles es mantenen, amb la particularitat que les espècies mal classificades pràcticament han desaparegut.

L'estudi del gràfic d'índexos de nivell aconsella un tall en 7 classes que contenen les següents espècies:

C50 `S.carnosus. v.incrustans, S.carnosus. v.typicus, S.carnosus.v.ramosus, S. gibbosiceps, S.domuncula, S.ficus, R.biseta, R.uniseta, R.elongata, R.pyrifera, Ps.sulfureus, Ps.hyalinus, L.ferrerhernandezI, L.ectyonimus, L.rugosus, Pr.epiphytum, Pr.rugosus, Pr.longispina, T.fugax`

C47 `C.celata, C.labyrinthica, C. carteri, C. schmidti, A. millari, C. viridis, S. cunctatrix, S.minax, D.bistellata, D.ornata, T.stellata, T.hallezi, T. unistellata, T.mixta, T.chondrilloides`

C45 `T.aurantium, T.citrina`

C57: `T.hemisphæricum, T. sarsi, T. papillatus, T.semisuberites, T.joubini, S. caminatus, S. spinularia, W. bursa, W. verrucosa , P. invaginata, T.stephensi`

[4] $\mathcal{C}_1$, per ser una classe d'un únic element, té com a descripció les components de la pròpia `C.viridis`

D \ c50	c47	c45	c57	c56	c39	c44	
Spo	0	0	0	6	0	0	0
Sp	0	10	0	0	0	0	0
S	19	0	0	0	0	0	0
St	0	0	0	0	0	0	2
P	0	0	0	1	22	0	0
Q	0	0	0	2	0	4	0
T	0	0	2	0	0	0	0
Ti	0	5	0	0	0	0	0
Tr	0	0	0	2	0	0	0
A	0	0	0	0	1	0	0

Table 6.4: Tcpp de la proposta Domingo 1990 versus Klass amb regles.

C56: `P. infrapilosa`, `P.littoralis`, `P. inflata`, `P. radiosa`, `P. ectofibrosa`, `P.uberrima`, `P.spinula`, `P.grimaldi`, `P.fusca`, `P.martæ`, `P.tissieri`, `P.agglutinaris`, `P. corticata`, `P.mammillaris`, `P.robusta`, `P.hirsuta`, `P.polytylota`, `P.tenax`, `P.conigera`, `A.aaptos`, `P.sollasi`, `S.antarcticus`, `S.capitatus`

C39 `Q.brevis`, `Quasilina.richardii`, `Quasilina.intermedia`, `R.oviformis`

C44 `O.pellita`, `S.borealis`

Efectuant aquest tall, doncs, s'obté una classificació que difereix en un 0.14 de la proposta [**?**], xifra substancialment inferior al 0.27 obtingut per la classificació sense regles. També s'ha reduït, si bé d'una forma menys espectacular, la *distància* a la classificació tradicional, que ha passat d'un 36% a un 31%, i respecte de **LINNEO** , s'ha passat d'un 21% a un 19%. Per tant, ara, els resultats de **KLASS** són més a prop de la proposta [**?**] que de qualsevol altra i, a més, són els que més s'hi acosten de tots. La Tcpp que permet analitzar on és el 14% de diferències entre una i altra és la taula **??**.

En primer lloc, notar que les F.*Suberitidæ*, F.*Tethyidæ* i F.*Stylocordilæ* es detecten amb tota exactitud.

Potser la part que experimenta una major reestructuració és la corresponent als gèneres de la F.*Polymastiidæ* que ja no apareixen dispersos en totes les classes; tot al contrari, la F.*Polymastiidæ* de la proposta [**?**] s'ha organitzat notablement respecte del que s'obtenia amb una classificació sense regles, malgrat que incorpora l'espècie `Aaptos aaptos`, que a [**?**] forma Família apart. En aquest sentit, dir que tampoc amb **LINNEO** es segrega una classe aïllada per aquesta espècie.

D'altra banda, amb **KLASS** , els gèneres `Quasilina` i `Ridleya` formen una classe apart en la línia de [**?**], encara que no inclouen `Weberella` com la F.*Quasilinidæ* fa.

De fet, la classe `C56` conté tots els altres gèneres de la F.*Polymastidæ* clàssica que [**?**] escindeix en F.*Trichostematidæ* i F.*Semipolymastiidæ*, i es pot observar una certa subestructura en que `Trichostema` s'aparta força de la resta.

Per últim, citar el fet que les F.*Spirastrellidæ* i F.*Clionidæ* es mantenen confoses en una única classe per alts nivells de la jerarquia, separant-se en dues ja per α-particions de nivells intermedis. El fet que `Diplastrella` no formi part del nucli `Cliona-Spirastrella-Alectona`, sinó de `Timea`, no és més que una *conseqüència directa* de la regla r_3, que selecciona pràcticament la totalitat de les `Timea` juntament amb les `Diplastrella`, la qual cosa força la unió d'aquests dos gèneres ja a baixos nivells de la jerarquia. Sense un canvi en $\mathcal{R}$, cap millora és possible en aquest sentit[5]. La reformulació del conjunt de regles es reserva per a un treball futur.

Pel que fa a la caracterització de les classes, es poden caracteritzar totalment tres de les classes, de forma mínima, amb el sistema

$$\begin{aligned}\mathcal{S}_{\mathcal{P}} = \{ \quad & (\mathcal{C}_{45}, \textit{capa interna del cortex}, \textit{microscleras}),\\ & (\mathcal{C}_{44}, \textit{tipo megasclera ectosomica}, \textit{oxas estrongiloxas}),\\ & (\mathcal{C}_{39}, \textit{disposicion espicular esqueleto}, \textit{reducido})\}\end{aligned}$$

i es detecten, a més, 26 caracteritzadors parcials per les altres classes, que permeten generar un conjunt de regles que selecciona el 76% dels objectes, consultant, com a màxim, dues variables.

Una discussió en termes biològics més profunda del problema de la classificació taxonòmica de les esponges, en particular de l'Ordre *Hadromerida*, es pot trobar en el treball ja referit [**?**].

6.5 Una segona aproximació

En el capítol **??** s'apuntava la possibilitat d'utilitzar la informació proporcionada per les regles fent una hipòtesi de món tancat relacionada amb la rellevància dels atributs. D'acord amb això, es tractaria de classificar les esponges utilitzant només els atributs mencionats a les regles. Pel cas que ens ocupa, el conjunt de propietats rellevants seria

$\mathcal{A}^{\mathcal{R}}$ = {Tipo espícula principal estilo, Tipo espícula principal tilostilo, Tipo microsclera, Tipo de áster, Forma final, Papilas, Disposición espicular esqueleto }

La classificació resultant apareix a l'apèndix **??**. Segons el gràfic d'índexos de nivell i l'heurística dissenyada, es pot fer un tall en 3, 5, ó 7 classes. Independentment del nivell al que es talli l'arbre, s'observa clarament com el resultat és poc menys que caòtic. No sembla haver-hi cap possibilitat d'interpretar la classificació obtinguda, en la que espècies de totes les famílies es barregen en qualsevol classe (vegi's al respecte `C57` que conté simultàniament espècies de `Tylexocladus, Aaptos, Cliona, Tethya` i `Alectona`).

És evident, doncs, que no resulta en absolut equivalent utilitzar les regles segons es presenta en aquesta secció o fent el tractament booleà exposat en la secció anterior. La

[5]Ni tan sols **LINNEO+** és capaç de dissociar unes d'altres amb les regles proporcionades.

informació addicional proporcionada per les regles fa referència a les relacions entre les variables, la qual cosa és una informació molt rica, que es reflexa en les diferències de resultats.

De fet, la informació proporcionada pels experts a través de les regles no és necessàriament completa (avenç important, atès que els experts quasi mai poden formalitzar tot el seu coneixement) i característiques importants poden ometre's si només es treballa amb $\mathcal{A}^{\mathcal{R}}$.

6.6 Conclusió

Es pot dir que la incorporació de les regles al procés de classificació augmenta la interpretabilitat dels resultats i, en aquesta aplicació concreta, situa la classificació de **KLASS** prop de la recent proposta [**?**], fins i tot més a prop (en el sentit, òbviament, de la mesura proposada a §**??**) que la classificació de **LINNEO**, que fou la classificació que s'utilitzà com a referència en l'elaboració de la proposta [**?**].

Chapter 7

Aplicacions (II). Poblacions estel·lars

"La distància de la Terra al Sol [...]
és com res quan se la compara
amb l'esfera de les estrelles fixes."

COPÈRNIC

7.1 Introducció

L'Astronomia presenta algunes dificultats comparada amb altres ciències experimentals. Primerament, cal fer notar que els astrònoms no són experimentadors pròpiament dits, sinó mers observadors. En segon lloc, les magnituds que s'observen poden ser variables respecte del temps, com posicions, moviments propis, velocitats radials i colors d'una estrella determinada, entre d'altres. I encara hi ha una dificultat addicional si es considera la qüestió dels biaixos inherents al fet que totes les observacions venen lligades a una base d'observació concreta i en moviment: la Terra. D'altra banda, els mètodes utilitzats en l'anàlisi d'aquestes observacions són més aviat antics, i convé una actualització urgent que permeti aprofitar el material d'alta qualitat que produeixen els observatoris orbitals [**?**]. Per tot això, l'Astronomia s'interessa per les noves eines i tendències en el camp de l'Estadística i la Teoria de la Informació que li puguin resultar útils.

Un aspecte rellevant dels estudis efectuats a partir de catàlegs d'estrelles és la segregació d'aquestes en poblacions segons criteris espectrals, fotomètrics o cinemàtics. Aquesta és una qüestió interessant en si mateixa, però també per les seves conseqüències en l'estudi cinemàtic del fluïd estelar de l'entorn solar, i s'ha abordat utilitzant aproximacions diverses (estadística, numèrica, dinàmica, connexionista ...). Ens trobem, doncs, davant d'un problema de classificació en un domini poc estructurat susceptible, per tant, de ser tractat per **KLASS** .

7.2 Marc de treball

En els darrers quinze anys hi ha hagut certa controvèrsia sobre el nombre de poblacions necessàries per a representar la composició de la nostra Galàxia, la proporció d'estrelles de cada població i llur naturalesa. La divisió entre *disc* i *halo* presentada per [**?**], per exemple, contrasta amb la proposta d'alguns autors, com [**?**], que distingeixen entre *disc prim*, *halo* i *disc gruixut*[1].

En contacte amb membres del **Departament de Matemàtica i Telemàtica** de la **UPC**, ens estem introduint en aquesta àrea d'investigació, i el primer pas ha estat realitzar una prova inicial utilitzant dades simulades per tal d'avaluar la qualitat del classificador.

L'equip amb el que treballem disposa d'un generador d'estrelles sintètiques que utilitza, entre altres coses, el diagrama de Hertzsprung-Russell (conegut per diagrama HR) en la versió de [**?**][2], i ha proporcionat una mostra de 100 estrelles, descrites per vuit variables que es detallen més avall. L'avantatge de treballar amb dades simulades és que hi ha un coneixement inicial de les classes que es volen distingir i, d'aquesta manera, es pot valorar perfectament el comportament del sistema. Aquest és un dels casos en que es disposa d'una classificació *correcta a priori*, que servirà de referència — equivalent d'aquesta tècnica, des del punt de vista de la Intel·ligència Artificial, és l'aprenentatge automàtic supervisat —, i permetrà calibrar el bon comportament del sistema. Un cop superada aquesta fase, es podria passar al tractament de dades reals, de les que no es coneix, *a priori*, la població de pertinença.

En segon lloc, l'estudi de les estrelles sintètiques confereix a **KLASS** el paper d'eina d'ajuda a l'adquisició del coneixement de suport que posteriorment es pot utilitzar en la classificació de les estrelles reals. Així, el processament de no massa estrelles sintètiques permet perfilar les regles amb les que després es classificarà un conjunt gran d'estrelles reals.

Les variables utilitzades en la descripció de les estrelles sintètiques són:

Paral·laxa: es denota per π i es mesura en milisegons d'arc. De fet, $\pi = \frac{1}{r}$, amb r distància al Sol[3] en *pc*.

Velocitat radial de l'estrella (v_r), en Km/s.

Longitud galàctica: (l), en graus.

Latitud galàctica: (b), en graus.

[1] A títol orientatiu dir que l'*halo* és la part més antiga d'una galàxia i es distribueix esfèricament entorn del seu centre. El disc, en canvi, està format per estrelles més recents, que segueixen, bàsicament, un moviment de rotació dins del pla galàctic.

[2] El diagrama HR presenta la relació entre els tipus espectrals i la brillantor real de l'estrella. [**?**] donen diagrames separats per les diferents poblacions.

[3] La distància des de la que el Sol i la Terra tenen una separació aparent d'1 *arcsec* és 1 *pc* i 1 Kpc equival a 1000 *pc*.

Moviment propi en longitud galàctica: $(\mu_l \cos b)$, en $''/any$.

Moviment propi en latitud galàctica: (μ_b), en $''/any$.

Magnitud aparent: (m_v) o brillantor aparent, en magnituds.

Tipus espectral: segons la categorització utilitzada a [**?**] en el diagrama HR. Pot prendre com a valors les categories

> **Tipus primerencs:** o7-o9, b0-b4, b5-b9, a0, a2-a3, a4-a9
> **Tipus tardans:** f0-f4, f5-f9, g0-g7, g8-g9, k0-k4, k5-k6, k7-k9, m0-m4, m5+

totes elles, característiques directament observables (en realitat, (r, l, b) són les coordenades esfèriques de l'estrella, i $(v_r, \mu_l \cos b, \mu_b)$ són les coordenades esfèriques de la seva velocitat), i coincidents amb les que s'enregistren en el catàleg d'entrada de l'**Hipparcos**[4], d'on s'extreurien les estrelles reals en una segona etapa de l'estudi.

Amb aquest mateix generador, [**?**] estan estudiant el comportament dels mapes autoorganitzatius de Kohonen — un tipus de xarxa neuronal — en la classificació. La limitació que les xarxes neuronals presenten és, fonamentalment, que només poden tractar informació numèrica, mentre que el tipus espectral és una variable de tipus categòric de la qual no es pot prescindir perquè conté molta informació. La solució adoptada per [**?**] és continuïtzar aquesta variable, utilitzant un model probabilístic.

L'objectiu d'aquesta primera prova és separar les dues poblacions *halo* i *disc*[5]. Per a fer-ho, s'aplica la metodologia d'anàlisi proposada al capítol **??** utilitzant el criteri de Ward, i la mètrica mixta.

7.3 Estudi

7.3.1 Passa $\xi = 0$

En la classificació inicial, en que no intervé cap regla, s'observa un 42% d'objectes mal classificats. La taula **??** representa la taula de contingència de la partició producte (Tcpp)

[4] L'**Hipparcos** és un satèl·lit, llençat a l'espai amb l'objectiu d'observar unes 100000 estrelles, per tal de millorar l'estimació dels seus moviments propis, paral·laxes, etc. En ser fora de l'atmosfera terrestre s'espera un guany qualitatiu en les mesures. El catàleg que s'està utilitzant és el més fiable actualment, i constitueix la base de dades que el satèl·lit disposa per a localitzar les estrelles que ha d'estudiar.

[5] No es planteja la distinció de les tres poblacions perquè les estrelles de l'*halo* que són observables des de la Terra presenten unes posicions, brillantors aparents i altres propietats que es confonen fàcilment amb les de la població intermèdia. Només el pas d'uns quants milions d'anys permetria distingir unes d'altres. Però fins i tot en el supòsit que es superés la barrera temporal, aquestes estrelles de l'*halo* s'haurien allunyat suficientment de la Terra com per a quedar fora del nostre abast.

És amb l'edat o la metal·licitat de les estrelles que es podrien distingir bé les tres poblacions. Però aquestes característiques, que no són directament observables, comporten grans problemes d'estimació, sobre els que la literatura és extensa, i reservem aquesta possibilitat per a treballs futurs.

KLASS \ Real	**Disc**	**Halo**
Disc	52	29
Halo	13	6

Table 7.1: Tcpp de les classes reals *versus* les de **KLASS** sense usar regles.

de la classificació generada per **KLASS** *versus* l'autèntica, i el coeficient de diferències entre ambdues classificacions, segons la proposta de la §**??**, és de 0.42. Pel fet de disposar de la classificació real, es pot interpretar 1-0.42 = 0.58 com una mesura de la qualitat de la classificació obtinguda, que no resulta gaire satisfactòria. Pràcticament la meitat dels objectes de la mostra surten mal classificats (a l'annex **??** s'hi poden trobar els detalls d'aquesta classificació).

7.3.2 Passa $\xi = 1$

A la vista dels resultats tan pobres de la classificació tradicional, es comença a treballar en l'elaboració d'un conjunt inicial de regles que condueixin a millors resultats. La primera dificultat a afrontar és el fet que els experts no proporcionin regles sobre les característiques observables, sinó sobre altres variables que no s'han recollit a la matriu de dades, i que són transformació de les observades. Per exemple, se sap que les estrelles més *ràpides* i amb moviment excèntric són de l'*halo*, mentre que les més *lentes*[6] són del *disc*.

Per tal d'obtenir més informació sobre les direccions del moviment, és aconsellable projectar el vector velocitat en un espai de coordenades cartesianes. Cal, doncs, transformar les coordenades de la velocitat, la qual cosa és possible a partir dels moviments propis de l'estrella, i de la pròpia v_r.

Així, les coordenades de la velocitat, en Km/s, es poden calcular com:

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix} = {}_3C^{3'} \begin{pmatrix} v_r \\ kr\mu_l \cos b \\ kr\mu_b \end{pmatrix}$$

on

- $r = \frac{1}{\pi}$ és la distància al Sol,
- k és una constant de transformació d'unitats ($k = 4741\ Km\ s^{-1}$ ($"/any)^{-1} Kpc^{-1}$)
- i

$${}_3C^{3'} = \begin{pmatrix} \cos l \cos b & -\sin l & -\cos l \sin b \\ \sin l \sin b & \cos l & -\sin b \sin l \\ \sin b & 0 & \cos b \end{pmatrix}$$

[6]En aquest contexte, entendrem velocitat heliocèntrica, és a dir, referida al Sol.

Amb tot això s'obté que

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix} = \begin{pmatrix} v_r \cos l \cos b - kr\mu_l \sin l \cos b - kr\mu_b \cos l \sin b \\ v_r \sin l \sin b + kr\mu_l \cos l \cos b - kr\mu_b \sin b \sin l \\ v_r \sin b + kr\mu_b \cos b \end{pmatrix}$$

i es pot utilitzar la regla que ens proposa l'especialista que diu:

$$r_1 : (|W| \geq 40) \longrightarrow Halo$$

Aquest tipus de transformacions són possibles quan es disposa d'un coneixement de suport . Hi ha, a la literatura, treballs que indiquen solucions per aprofitar aquest tipus de coneixement en la millora de tasques cognitives complexes [**?**].

La partició induïda per $\mathcal{R}^1 = \{r_1\}$ sobre $\mathcal{I}$ només té una classe, l'*halo*, que conté 15 objectes. La classificació ascendent jeràrquica d'aquesta classe dóna lloc a una classe de 15 elements descrita per $\bar{\imath}^1_{Halo}$. La resta d'objectes que no han estat seleccionats per r_1 conformen la classe residual $\mathcal{C}^1_0$.

Variable	$\bar{\imath}^1_{Halo}$	
	$\bar{x}$	s
PAR	6.6	4.8960
VR	-22.8866	113.3569
GALON	201.2167	103.3539
GALAT	-8.5305	47.5703
MLCB	0.06587	0.1362
MB	-0.0481	0.1837
MV	7.4611	0.6433

$$; \quad \text{SP} = \begin{pmatrix} G0 - G7 = 1/15 \\ G8 - G9 = 1/15 \\ K0 - K4 = 3/5 \\ K5 - K6 = 4/15 \end{pmatrix}$$

Generar un tall d'una única classe sobre l'arbre jeràrquic de l'*halo* permet veure quines són les components del seu centre de gravetat. És aquesta descripció que, associada a un pes de 15 unitats i conjuntament amb els elements de la classe residual $\mathcal{C}^1_0$ es classifica seguint el procés ordinari. A l'annexe hi ha l'arbre resultant de classificar

$$\tilde{\mathcal{I}}^1 = \{(\bar{\imath}_{Halo}, 15)\} \cup \{(i, 1) : i \in \mathcal{C}^1_0\} \subset \mathcal{I}^{*1} \times \mathcal{M}^1$$

Ara, la taula de contingència de la partició producte és diferent de l'anterior (veure taula **??**) i hi ha un 43% de diferències entre la classificació obtinguda i la real. Això significa que la qualitat de $\mathcal{P}^1_{\mathcal{R}}$ és de 1-0.43=0.57, i que s'experimenta una lleugera millora, però no suficient.

Si bé la regla selecciona correctament les estrelles de l'*halo* (no hi ha cap estrella del disc que satisfaci r_1), no és prou potent per a separar en dues classes les estrelles de tota la mostra.

KLASS \ Real	**Disc**	**Halo**
Disc	47	32
Halo	8	3

KLASS \ Real	**Disc**	**Halo**
Disc	65	20
Halo	0	15

Table 7.2: Tcpp de classificar utilitzant $\mathcal{R}^1$ (*esq.*). $\mathcal{R}^1$ + HMT (*dret.*)

L'observació de l'arbre jeràrquic (τ^1) suggereix que $\mathcal{R}^2$ ha de contenir informació també sobre les estrelles del *disc*.

Aquest és un exemple clar del fet que no es poden fer hipòtesis de món tancat com la de dir que $\neg(|w| \geq 40) \longrightarrow Disc$, que donaria lloc a la classificació que apareix en els annexes, amb la Tcpp de la taula **??** (*dret.*), de qualitat 1-0.20 = 0.8, sensiblement superior a l'obtinguda, però que està identificant com a estrelles del disc estrelles com E41-2 o E11-2, que són de l'*halo* malgrat no ser seleccionades[7] per r_1.

7.3.3 Passa $\xi = 2$

Veient que calen regles que caracteritzin al Disc, els experts suggereixen que es formalitzi en Càlcul de Predicats el mateix model que hi ha darrera del seu simulador. Es tracta del diagrama HR, que relaciona, localment a l'*halo* i al *disc*, la brillantor real de l'estrella i el seu tipus espectral.

Novament, trobem que la magnitud real (M_v) és transformació de les variables observades. La relació entre una i altres és la següent:

$$M_v = m_v + 5 - \left(5 \log \frac{1}{\pi} + 0.15 \frac{1 - e^{-|\sin b| \frac{1}{\pi} \frac{1}{100}}}{|\sin b|}\right) \tag{7.1}$$

La formalització del diagrama genera un conjunt nombrós de regles (que es poden trobar en els annexes) que juntament amb r_1 constitueixen $\mathcal{R}^2$. L'aspecte d'aquestes regles és el següent[8]:

$$\begin{array}{ll} (\ (and & (M_v < 5) \wedge \\ & (Sp \in \{o7 - o9, b0 - b4, b5 - b9, a0, a2 - a3, a4 - a9, f0 - f4, f5 - f9\}) \\ ) \rightarrow Disc) & \end{array}$$

Sorprenentment, part del coneixement que ha servit per ha generar les dades no és capaç de seleccionar més que 4 estrelles, totes elles del disc, i que no apareixen pas massa allunyades a τ^1. La resta quedarien totes incloses a la classe residual, de no ser per la potència d'r_1, que captura algunes estrelles de l'Halo. Però la grandària de $\mathcal{C}_0^2$ no es deu pas a estrelles que no satisfan cap regla, ans al contrari, a estrelles que satisfan regles contradictòries.

Per observació d'aquestes estrelles, els experts se n'adonen que el diagrama HR no té gaire capacitat de separació de classes perquè la majoria de les estrelles se situen en el que

[7] Recordi's que $\mathcal{R}^\xi$ no té per què ser complet.
[8] A l'annex **??** hi ha la descripció total d'$\mathcal{R}^2$.

KLASS \ Real	**Disc**	**Halo**
Disc	59	33
Halo	6	2

Table 7.3: Tcpp de les classes reals *versus* les de KLASS usant el model HR.

KLASS \ Real	**Disc**	**Halo**
Disc	60	5
Halo	35	0

Table 7.4: Tcpp de les classes reals *versus* les de KLASS usant metal·licitats i coordenades de la velocitat.

es coneix com la *seqüència principal*, que és la zona del diagrama en que s'ubiquen gran part de les estrelles, i on les característiques observades d'aquestes són similars per totes les poblacions.

La qualitat d'aquesta classificació es xifra en un 59% resultant també poc satisfactòria. La taula **??** mostra la relació entre els resultats obtinguts i la classificació real (que es presenta a l'informe nro 4. de §**??**).

7.3.4 Passa $\xi = 3$

En una tercera fase s'utilitza informació sobre la relació existent dins cada població entre metal·licitat – velocitat. Les regles es recolzen en la caracterització que [**?**] fan de les relacions entre la metal·licitat i algunes components de la velocitat, concretament W i V, (veure annex **??**). La Tcpp resultant és a la taula **??** i la qualitat és del 0.6.

La selecció que fan les regles és correcta, però no prou potent per a definir nuclis per l'*halo* i el *disc* de forma que els objectes de la classe residual s'agreguin a un d'ells....

7.3.5 Passa $\xi = 4$

Per inspecció dels objectes que satisfan regles contraditòries, els experts *recorden*[9] una relació entre les components transformades de la velocitat interessant. Es tracta de les regles:

> **Si** el *mòdul de la velocitat* (en relació a la velocitat del Sol, per ser exactes) és *alt* **llavors** l'estrella és de l'*halo*
>
> **Si** el *mòdul de la velocitat* (en relació a la velocitat de Sol, per ser exactes) és *baix* llavors l'estrella és del *disc*

De fet, aquestes dues regles són expressables en els següents termes

$$r_{18} = \left(\sqrt{\bar{U}^2 + \bar{V}^2 + \bar{W}^2} > 75\right) \longrightarrow halo$$

[9] I aquest és un cas clar en que el sistema contribueix a posar de manifest coneixement *implícit* sobre el domini.

K \ R	**Disc**	**Halo**
Disc	64	1
Halo	35	0

R \ K	**Halo**	**Disc**	**C3**	**C4**	**C5**	**C6**
Disc	5	56	1	1	1	1
Halo	32	3	0	0	0	0

Table 7.5: Tcpp de les classes reals *versus* les de **KLASS** usant el mòdul de la velocitat respecte del Sol. Talls en 2(*esq.*) i 5 (*dret.*) classes.

$$r_{19} = \left(\sqrt{\bar{U}^2 + \bar{V}^2 + \bar{W}^2} < 60\right) \longrightarrow disc$$

on $\bar{U}, \bar{V}, \bar{W}$ representen les coordenades de la velocitat amb respecte a la velocitat del Sol. En concret, $\bar{U} = U - 9$, $\bar{V} = V - 12$ i $\bar{W} = W - 7$, ja que la velocitat del Sol és $(U_s, V_s, W_s) = (9, 12, 7)\ Km/s$.

Fent $\mathcal{R}^4 = \{r_{18}, r_{19}\}$ s'obté una $\mathcal{P}^4_{\mathcal{R}}$ amb una classe residual de només 5 estrelles, cap de les quals satisfà regles contradictòries. La classificació de les estrelles seleccionades com a *halo* dóna lloc a l'informe nº 6.H i a l'arbre τ^4_{Halo} dels annexes i la de les estrelles seleccionades per r_{19} dóna lloc a τ^4_{Disc}. Els centres de gravetat d'aquestes dues classes, afectats de la massa corresponent són classificats conjuntament amb les estrelles de la classe residual i conformen l'arbre τ^4, de l'informe nro. 6. Un tall en dues classes permet construir la Tcpp de la taula **??** i valorar aquesta classificació en una qualitat de 0.64, que no ens anima massa pel que fa als resultats de les iteracions anteriors.

Ara bé, observant que el diagrama d'índexos de nivell suggereix un tall clar en 6 classes, resulta una Tcpp tal com la de la taula **??** (*dret.*), que s'avalua en una qualitat del .88, sensiblement superior a totes les altres!

La descripció de les classes per aquest tall es mostra a continuació:

Informe de classificació nº 6 (τ^4)

Dades: data/resta02
Nombre d'objectes: 100
Nombre de propietats: 8
Criteri de classificació: WARD
Mètrica utilitzada: MIXTE
Valors de les constants de ponderació: $\alpha = 0.103$ i $\beta = 0.897$

Descripció de les 6. classes

Classe	Halo		Disc		E61-1	E58-1	E13-1	E7-1
Var.	$\bar{x}$	s	$\bar{x}$	s				
PAR	5.5139	5.0984	14.5932	14.1032	6.0	6.0	2.0	34.0
VR	8.1486	92.5322	-3.6559	18.0381	62.5	-27.9	4.3	-0.8
GALON	207.6789	117.7034	209.8683	88.0957	163.64	328.07	72.65	27.14
GALAT	-12.5908	44.8745	-1.6064	33.5009	-7.97	9.26	-66.51	-63.9
MLCB	0.0191	0.1369	0.0022	0.06532	0.004	-0.05	0.013	-0.073
MB	-0.024375	0.15212	-0.0129	0.0846	0.023	-0.025	-0.017	-0.33
MV	7.2176	0.9898	7.2237	1.3343	7.63	6.935	6.987	7.928
SP	A0 = 1/72 A4-A9 = 1/36 F0-F4 = 1/36 G0-G7 = 1/36 G8-G9 = 1/72 K0-K4 = 23/72 K5-K6 = 41/72		B0-B4 = 11/59 B5-B9 = 2/59 A0 = 7/59 A2-A3 = 8/59 A4-A9 = 3/59 F0-F4 = 5/59 F5-F9 = 9/59 G0-G7 = 8/59 G8-G9 = 1/59 K0-K4 = 5/59		A0	K0-K4	B0-B4	G0-G7
n	37		59		1	1	1	1

7.3.6 Passa $\xi = 5$

En una darrera passa i per a arrodonir l'estudi es construeix un cinquè conjunt de regles $\mathcal{R}^5 = \mathcal{R}^2 \cup \mathcal{R}^3 \cup \mathcal{R}^4$. Els resultats són idèntics als que s'obtenen en la iteració anterior, la qual cosa posa de manifest el fet que si una regla és potent, no és necessari ampliar la base de coneixement de suport, i que la qualitat dels resultats no és pas funció de la grandària d'R^{ξ}, sino de la potència dels seus components.

7.4 Conclusions

En aquesta aplicació cal notar especialment la interacció amb els experts; en l'explicació s'ha posat l'accent en els aspectes cognitius que precedeixen a la formació de regles (explicitació de relacions). A més, ja la naturalesa sintètica de les dades constitueix un tipus de biaix que només és compensable amb una "*acurada*" explicitació del coneixement addicional.

Un altre aspecte important és que els atributs interaccionen entre si de forma molt acusada i les relacions entre ells creen "*característiques*" imprescindibles per a la correcta classificació dels objectes del domini.

En resum es pot dir que

- Les regles que millors resultats han donat, són les que expressaven un coneixement qualitatiu del fenòmen que s'està estudiant (iteració quarta), superant l'actuació d'aquelles que deriven directament de models quantitatius (iteracions segona i tercera).

- Utilitzar relacions entre variables (passa 4) és més potent que descriure el comportament d'una variable aïllada en cada classe (passa 3). L'ús de regles que descriuen

ξ	**Qualitat**	**Regles**
0	0.52	$\mathcal{R}$ =0
1	0.57	r_1
2	0.59	r_1 + HR
3	0.60	r_1 + Carney
4	0.64	r_{18}, r_{19} (2 classes)
	$\boxed{0.88}$	r_{18}, r_{19} (6 classes)

Table 7.6: Evolució de la *qualitat* dels resultats en iteracions successives.

aquestes relacions aporta al sistema una informació que el procediment de classificació per si sol és incapaç de captar.

- L'ús d'una mostra simulada dota al classificador d'una funció similar a la d'un sistema d'aprenentatge supervisat, i s'estableix un procés interactiu expert-sistema, al final del qual s'ha aconseguit formalitzar un coneixement de suport sobre el domini transmissible al sistema i que inicialment no estava ben perfilat. La taula **??** mostra com en les successives iteracions la qualitat dels resultats millora com a conseqüència d'aquesta interacció.

Chapter 8

Comparació amb altres sistemes i mètodes

En primer lloc es compara els resultats obtinguts per **KLASS** fent servir la distància mixta amb els que produeixen altres distàncies o mètodes de classificació. A la secció **??**, la comparació es fa sobre la base d'una matriu de dades de la que es disposa de bibliografia suficient.

Tot seguit, es presenta una comparació entre **KLASS** i alguns sistemes comercials coneguts i extesos a àmbits empresarials com són **SPSS** [**?**] i **SPAD** [**?**] (aquest darrer de caire un xic més experimental, potser). En aquest cas es treballa una mostra simulada sobre el cub unitari, que facilita les representacions gràfiques.

8.1 Microodinadors

Es presenta l'aplicació de **KLASS** a un conjunt de dades que apareix a [**?**] i que ha estat estudiat a [**?**], referents a microordinadors. Si bé el conjunt de dades és reduït i fa referència a uns ordinadors un tant antics, resulta un conjunt interessant, fonamentalment per dues raons:

- es tracta d'una matriu de dades no homogènia que permetrà estudiar bé el comportament de la distància mixta
- constitueix un joc de dades ben estudiat a la literatura, que permetrà comparar el comportament de la classificació usant la distància mixta respecte d'altres mètodes, sobre la base d'un conjunt de dades comú.

La matriu de dades descriu 12 microordinadors americans a través de 5 variables, tres de les quals són qualitatives: *Display, MP, Keys* (veure la matriu a la taula **??**).

Per aquestes dades, Michalski, utilitzant classificació conceptual, dóna la classificació de la taula **??**. Per la seva banda, Gowda i Diday presenten les de les taules **??** i **??**. Incloent els

Display	TV	Black & White	Microprocessor	Motorola	6502
		Color			6502A
	Built in				6502C
	Terminal			Intel (and similar)	Z80
					8080A
				Hewlett Packard	

Table 8.1: Estructura de les variables categòriques des del punt de vista de l'expert.

Clust.	Objectes
1	APPLE-II ATARI-800 COMMODORE-VIC-20-A COMMODORE-VIC-20-B
2	EXIDI-SORCERER OHIO-SC.-CHALLENGER OHIO-SC.-II-SERIES TRS-80-I
3	ZENITH-H8 ZENITH-H89 HP-85 TRS-80-III
4	HORIZON

Table 8.2: Una de les classificacions proposades per l'expert.

resultats de **KLASS**, hi ha, doncs, quatre classificacions de les dades realitzades per mètodes automàtics.

S'ha volgut, a més, disposar de l'opinió d'un expert en el tema, i s'ha demanat a un especialista local que classifiqui les dades[1].

En primer lloc, cal remarcar que aquest expert realitzà la classificació dels 12 microordinadors en base a les variables més rellevants. Des del seu punt de vista, les variables *ROM* i *Keys* eren molt poc importants per a fer una caracterització dels microordinadors. No obstant, la rellevància de les variables és una informació que, fins al moment, no s'ha tingut present en el nostre sistema.

A continuació, l'expert determinà com es podien agrupar els valors de les variables categòriques. De fet, cercava una estructura en les variables qualitatives (veure taula **??**), basant-se en el seu coneixement sobre el tema, i en la seva experiència.

Fet això, va proposar tres classificacions generals d'acord amb els valors de les variables *Display, MP* i *RAM* respectivament, i va manifestar que qualsevol combinació d'aquestes classificacions bàsiques podria considerar-se acceptable. Cap d'elles tenia quatre classes, exceptuant la que correspon a utilitzar com a criteri de classificació el tipus de pantalla, la qual es mostra a la taula **??**.

D'altra banda, la classificació de les dades amb **KLASS** utilitzant la distància mixta amb $\alpha = 0.014$ i $\beta = 0.986$, tal com resulta de la fòrmula **??**, dóna lloc al dendrograma de la figura **??**.

El diagrama dels índexos de nivell suggereix un tall en 2 ó 4 classes. La taula **??** proporciona una descripció extensional i conceptual de les classes corresponents a una α-partició en quatre elements. Observant-la, és clar com els resultats de la classificació amb

[1]Es tracta del Dr. Miquel Barceló, del Departament de **LSI** de la **UPC**.

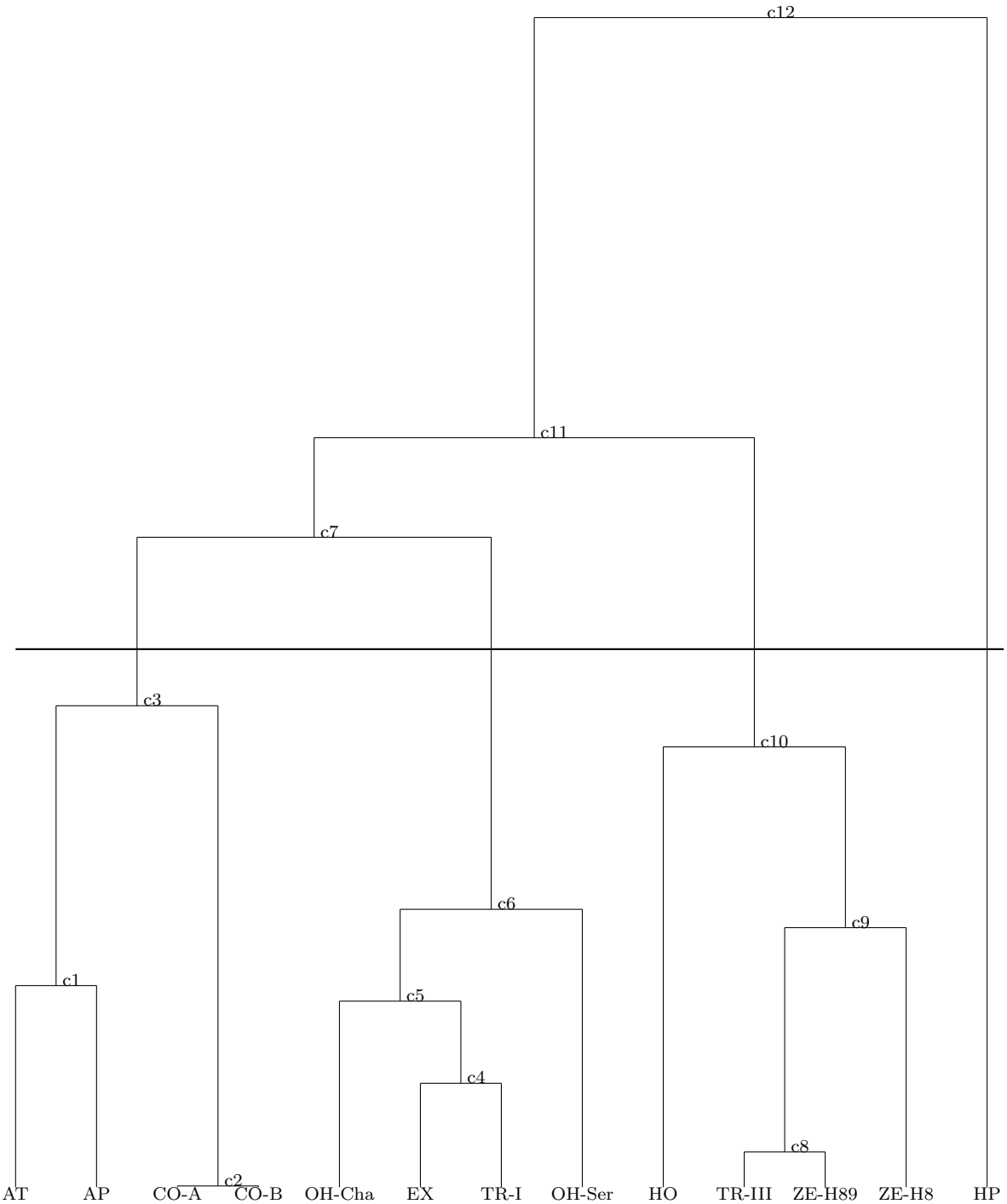


Figure 8.1: Dendrograma dels microordinadors. Criteri de Ward i mètrica mixta ($\alpha = 0.014$ and $\beta = 0.986$).

Classe			1	2	3	4
Proto-	**Display**		COLOR-TV	B-&-W-TV	Built-in 3/4 Terminal 1/4	TERMINAL
	RAM		40	44	60	64
	ROM		47/4	9	31/4	8
tipus	**Micro-** **proces-** **sador**	6502	1/2	1/4		HP
de		6502A	1/2	1/2		
		Z80			3/4	
la		8080A			1/4	
		6502C		1/4		
classe	**Teclat**	52	1/4			92
		57–63	1/4	1/4	1/4	
		64–73	1/2		3/4	
		53–56		3/4		
Descripció extensional			APPLE-II ATARI-800 COMMODO- RE-VIC-20-A COMMODO- RE-VIC-20-B	EXIDI- SORCERER TRS-80-I OHIO-SC.- CHALLEN. OHIO-SC.- II-SERIES	ZENITH- H8 ZENITH- H89 HORIZON TRS-80-III	HP-85

Table 8.3: Descripció conceptual i intensional de les classes proposades per **KLASS**.

la distància mixta són d'una interpretació directa: El criteri de classificació és el tipus de pantalla (*Display*), seguit, amb menys influència, del tipus de microprocessador (*MP*). Així, es podria fer la següent descripció de les classes:

C1 Ordinadors amb pantalla de televisió en color i microprocessador Motorola

C2 Ordinadors amb pantalla de televisió en blanc i negre i microprocessador Motorola

C3 Ordinadors amb pantalla que no és de televisió, i microprocessador Intel (o similar)

C4 HP.

De fet, es poden utilitzar els mecanismes presentats a la secció **??** per a efectuar, de forma automàtica, una caracterització de les classes. El resultat és el següent sistema complet de caracterització, que és extraordinàriament semblant a la interpretació que l'expert va donar d'aquesta classificació:

$$\begin{aligned}\mathcal{S}_{\mathcal{P}_4} = \{ \quad & (\mathcal{C}_1,\ Display,\ Color - TV), (\mathcal{C}_2,\ Display,\ B\&W - TV), \\ & (\mathcal{C}_3,\ MP, \{Z-80, 8080A\}), (\mathcal{C}_4,\ MP,\ HP)\}\end{aligned}$$

No s'obtenen interpretacions tan clares si es classifica usant altres valors d'α i β que no siguin trivials — diferents de (0,1) ó (1,0)—.

Considerant que, pel moment [**?**], la validació de les classificacions no és, en general, una qüestió massa ben resolta, es pot utilitzar com a eina de comparació la capacitat de les diferents propostes per ser interpretades. Malgrat ésser una mesura subjectiva, la

Expert\ KLASS	C1	C2	C3	C4
C1	4	0	0	0
C2	0	4	0	0
C3	0	0	3	1
C4	0	0	1	0

Table 8.4: Tcpp classificació de **KLASS** *versus* classificació de l'expert.

interpretabilitat d'una classificació augmenta la seva utilitat. En aquest sentit, i segons l'opinió de l'expert consultat, **KLASS** i la classificació conceptual de Michalski tindrien millor consideració que les restants. Recolzant aquesta idea, es pot veure com aquesta darrera també es presta a una bona caracterització automàtica:

$$\mathcal{S}_{\mathcal{P}_2} = \{ \quad (\mathcal{C}_1,\ MP, \{6502, 6502A, 6502C\}), (\mathcal{C}_2,\ MP, HP), (\mathcal{C}_4,\ MP, Z80)\}$$

i en una segona passa $S_{\mathcal{P}_2'} = \{(\mathcal{C}_3,\ Display,\ Built-in)\}$ amb el sistema condicionant $\mathcal{S}^{Cond}_{\mathcal{P}_2} = \{(\mathcal{C}_3,\ MP, \overline{HP})\}$

De fet, buscant més arguments objectius, la classificació que més s'assembla amb la que proporciona l'expert, en el sentit de la mesura proposada a la secció **??**, d'entre totes les que s'estan estudiant, és la de **KLASS**, amb un 15% de diferències. La taula **??** mostra on són aquestes diferències.

En realitat, hi ha només discrepància en la ubicació de l'`HP` i de l'`Horizon`, i es pot dir que el resultat de **KLASS** s'ajusta prou bé a la classificació de l'expert. En efecte, si es considera el tipus de pantalla, l'`HP` s'ha de classificar amb els altres ordinadors que tenen monitor. Però és raonable mantenir-lo en una classe apart, des del moment que totes les seves característiques són força diferents dels altres ordinadors. Pel contrari, l'`Horizon` s'assembla força als ordinadors de la classe 3, exceptuant que és l'únic que té monitor incorporat.

La classificació de Michalski dista més de l'anterior (un 46%), però també resulta interpretable. De fet, és la que més s'assembla a la que l'expert donava utilitzant el tipus de microprocessador com a criteri de classificació (un 23% de diferències contra un 30% que enregistra **KLASS**. Les diferències es deuen al fet que l'expert agrupa en una sola classe els grups 3 i 4 de Michalski, amb l'argument que tenen, tots ells, processador de tipus Intel (o similar).

Segons l'opinió de l'especialista, els resultats de [**?**], en canvi, són menys entenedors en termes de trobar un criteri de classificació clar. I el procés de caracterització automàtic ho corrobora. Per la primera d'elles, únicament es logra caracteritzar totalment la classe formada per l'HP. La caracterització total d'una classe amb un únic element és trivial i no aporta cap informació. La del salt mínim és totalment caracteritzable en dues passes:

$$\mathcal{S}_{\mathcal{P}_2} = \{ \quad (\mathcal{C}_2,\ MP,\ HP), (\mathcal{C}_3,\ MP,\ 6502A)\}$$

i per la segona passa

$$\mathcal{S}_{\mathcal{P}_2'} = \{(\mathcal{C}_1,\ Display, \{Color - TV, B\&W - TV\}), (\mathcal{C}_4,\ Display, \{Built - in, Terminal\})\}$$

condicionat per $\mathcal{S}_{\mathcal{P}_2^{Cond}} = \{(\mathcal{C}_1,\ MP, \overline{6502A}), (\mathcal{C}_4,\ MP,\ \overline{HP})\}$.

però no hi ha cap raó per la que els ordinadors amb microprocessador `6502A` apareixin segregats d'altres de la seva família, els quals surten barrejats amb ordinadors de processador Motorola. Des d'un punt de vista pràctic ($\mathcal{C}_1$ i $\mathcal{C}_3$ no semblen gaire ben estructurades).

Un detall curiós és que, fent $\alpha = 0.05$, $\beta = 0.95$ i fixant un tall a nivell 2.5, **KLASS** genera aquesta mateixa classificació no gaire afortunada. Els valors proposats automàticament per les constants de ponderació fan que es dongui més importància a les variables qualitatives, d'acord amb el fet que la informació qualitativa representa un 40% de la informació disponible, i els resultats milloren.

Per últim, val a dir que, amb aquestes mesures, no es pretén quantificar la qualitat de les diferents classificacions, donat que la proposta de l'expert és únicament una de les possibles, però no la millor, ni una referència absoluta.

8.2 Simulació de cubs

En aquesta secció es presenten els resultats d'estudiar algunes mostres del cub unitari amb tres sistemes diferents: **KLASS**, **SPSS** i **SPAD**.

8.2.1 Les dades

En primer lloc s'ha suposat una distribució uniforme sobre 3 cubs iguals i situats sobre la diagonal principal d'un cub unitari. Les variables considerades són precisament les coordenades cartesianes dels punts. La restricció a tres variables és interessant pel fet que permet representacions gràfiques entenedores del núvol de punts.

La figura **??** mostra el domini dels punts estudiats

8.2.2 Primera prova

Es simula una mostra de 150 punts de la població anterior, 50 d'ells en cada un dels cubs diagonals, que es representa a la figura **??**. Els punts estan enumerats de l'1 al 150.

La classificació amb **SPAD** (seguint el mètode dels veïns recíprocs i el criteri de Ward) es fa a partir de les components principals del núvol de punts. L'anàlisi en components principals indica que la primera acumula prou inèrcia, i es podria fer la classificació només amb aquesta component. En efecte, la primera component principal s'interpreta com la diagonal principal del cub, que recull molta informació sobre la disposició dels punts a classificar. Una classificació usant només aquesta component duria a resultats similars als

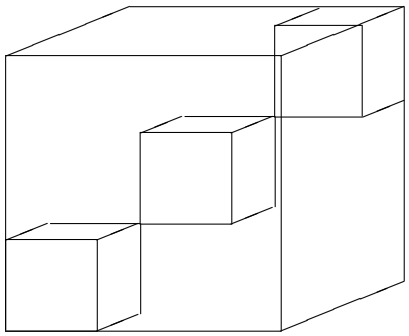


Figure 8.2: Població d'estudi

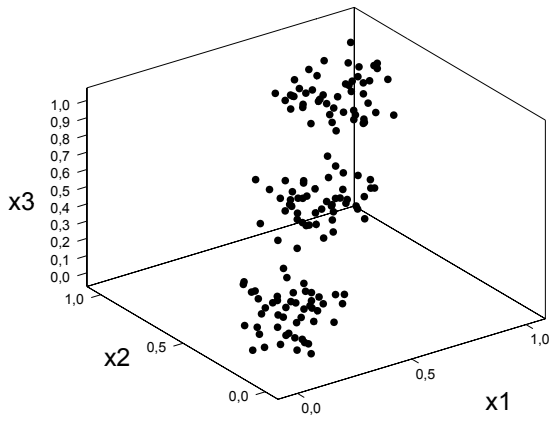


Figure 8.3: Mostra estudiada. Primera prova

Classe	X_1		X_2		X_3		**Elements**
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	
classe147	0.82903	0.10527	0.83743	0.10215	0.80769	0.09213	50
classe143	0.50071	0.10763	0.48819	0.09464	0.51179	0.10053	50
classe82	0.15738	0.08975	0.18455	0.09821	0.17996	0.10913	50

Table 8.5: Descripció de les classes.

presentats, que corresponen a fer la classificació utilitzant només les tres primeres components principals.

El gràfic dels índexos de nivell recomana 3 classes, que són:

$$\mathcal{P} = \{\mathcal{C}_1 = \{i_1 \dots i_{50}\}, \mathcal{C}_2 = \{i_{51} \dots i_{100}\}, \mathcal{C}_3 = \{i_{101} \dots i_{150}\}\}$$

justament es corresponen amb els punts dels tres cubs simulats. El reconeixement de les classes existents és, per tant, correcte com era d'esperar.

L'execució en **SPSS** , sobre les variables inicials, dóna resultats molt semblants, amb una partició en tres classes idèntica a $\mathcal{P}$.

Finalment, **KLASS** executat amb distància mixta[2] i criteri de Ward, detecta també els tres cubs existents, produint l'informe de l'annexe **??**. La taula **??** descriu les classes resultants.

Els arbres formats per **SPSS** i **KLASS** coincideixen plenament, malgrat les diferències de presentació, mentre que l'arbre generat per **SPAD** no és exactament igual: Si bé l'estructura general és força acord amb els altres paquets, que aquest darrer treballi sobre l'espai de les components principals produeix algunes variacions.

Per comprovar que efectivament aquesta és la raó de les diferències entre **KLASS-SPSS** i **SPAD** , s'extreuen les tres primeres components principals amb que **SPAD** realitza la classificació, i es procedeix a fer una classificació en aquest nou espai de definició dels punts amb **KLASS** i **SPSS**. Com era d'esperar, i pel fet que tots tres paquets implementen el mateix algorisme (encara que en versions diferents), els resultats són idèntics.

Es detecten, no obstant, variacions sense importància en l'ordre com s'efectuen les agregacions, que repercuteix en el nom que es dóna a les classes creades, però les agregacions que es fan són les mateixes en tots tres paquets.

La figura **??** representa gràficament la partició en tres classes identificada pels tres classificadors.

8.2.3 Prova segona

Es categoritzen les variables, a fi de tenir una matriu de dades qualitatives en la seva totalitat. Això permet treballar en la mètrica de χ^2.

[2]L'ús de la distància mixta, en aquest cas, és equivalent al de la distància euclídia normalitzada, que és precisament la que utilitzen **SPAD** i **SPSS**.

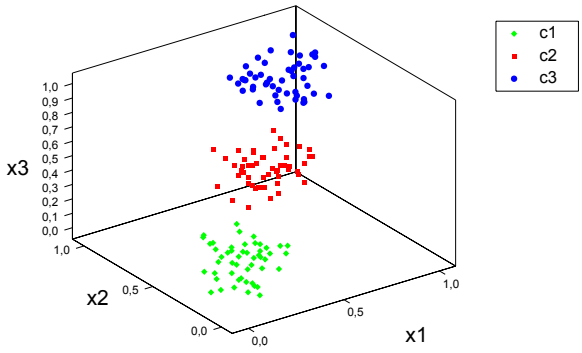


Figure 8.4: Classificació en tres grups.

X_1	
Modalitat	Interval
a	[0.0,0.2)
b	[0.2,0.4)
c	[0.4,0.6)
d	[0.6,0.8)
e	[0.8,1.0]

X_2	
Modalitat	Interval
a	[0, 0.5)
b	[0.5, 1]

X_3	
Modalitat	Interval
a	[0.0,0.25)
b	[0.25,0.50)
c	[0.50,0.75)
d	0.75,1.00]

Table 8.6: Categorització de les variables.

Les classes que s'han fet per les variables són a la taula **??**.

La classificació amb **KLASS** dóna lloc a un tall en 5 classes, en que la variable X_1 és totalment caracteritzadora de totes elles, i identificadora, per tant, de la partició[3]:

A l'annexe **??** es mostra la descripció de les classes (taula **??**) i el dendrograma que proporciona **KLASS** .

Els altres dos paquets no poden operar sobre la matriu de dades categòriques. Cal, en primer lloc, codificar les modalitats en claus numèriques i després optar per una de les següents vies:

- Fer una anàlisi de correspondències múltiples i realitzar la classificació per veïns recíprocs en mètrica euclídia de les components principals resultants.
- Desdoblar totes les variables qualitatives en paquets de variables binàries, passant a classificar la taula disjuntiva completa en la mètrica de χ^2.

Amb **SPAD**[4], de l'anàlisi de correspondències múltiples en resulten 5 components

[3]Valors característics d'X_1 per les diferents classes: $v^1_{\texttt{c145}} = e,\ v^1_{\texttt{c144}} = d,\ v^1_{\texttt{c143}} = c,\ v^1_{\texttt{c140}} = b,\ v^1_{\texttt{c142}} = a.$

[4]Per aquest cas, **SPSS** operaria exactament igual que **SPAD**.

principals. Amb aquestes components es realitza la classificació, i sembla clar un tall en 5 classes. La partició resultant és idèntica a la que s'obté amb **KLASS**, salvant petites diferències pel que fa al detall de l'estructura jeràrquica. Els arbres no són idèntics pel fet que s'està treballant en un espai reduït. Si s'utilitzen totes les components principals, llavors únicament es produeixen variacions en l'ordre com s'agreguen objectes idèntics, formant classes d'índex de nivell 0, i això és només qüestió de la implementació.

Quant a la segona alternativa, presentaria un clar desavantatge inicial: L'expansió de la matriu de dades a la seva forma disjuntiva completa, d'entrada, fa que enlloc de treballar amb una matriu 150×3 es passi a una matriu major (150×7), augment que, en aquest cas, no és crític donades les dimensions del problema, però que acostuma a ser-ho en dominis poc estructurats. Cal dir també que ni **SPSS** ni **SPAD** permeten fer aquesta classificació en mètrica de χ^2 directament. Els resultats utilitzant la mètrica euclídia ordinària no tenen res a veure, essent aquesta una comparació carent de tot sentit.

No obstant, es pot fer anàlisi de correspondències múltiples de la taula disjuntiva completa i classificar les components principals resultants, la qual cosa equival a fer una classificació en mètrica de χ^2. Aquest procediment, executat en **SPAD**, genera un arbre jeràrquic totalment coincident amb el que genera **KLASS**, tal com s'esperava.

8.2.4 Prova tercera

Es treballa amb la versió categòrica d'X_1 i amb les contínues per X_2 i X_3. Això conforma un núvol de punts definit per una variable categòrica i dues de contínues, que permetrà treballar amb la distància mixta pròpiament dita.

L'aplicació de l'heurístic proposat a §**??** als resultats de **KLASS** recomana talls en 3, 5, 7, 6 . . . classes. La segona determinada, sens dubte, per la categorització que s'ha fet de la primera variable i coincident, per tant, amb la de la segona prova.

De fet, fent el tall tant en 5 com en 3 classes, X_1 és una variable totalment caracteritzadora de totes elles i identificadora de la partició realitzada. Les particions de **KLASS** en 5 i 3 classes apareixen descrites a les taules **??** i **??** respectivament. A l'annexe **??** es mostra també el dendrograma generat per **KLASS** .

Pel que fa als altres dos paquets, la primera qüestió a comentar és novament que ni **SPAD** ni **SPSS** poden fer una classificació sobre la matriu de dades que s'ha definit; en tot cas, cal primer efectuar algun tipus de transformació sobre ella, com:

> categoritzar totes les variables contínues i tractar la matriu resultant de forma anàloga a com es feia en la prova segona

alternativa que suposa la introducció de certa arbitrarietat en el procés, el resultat del qual depèn enterament de la forma com es realitzi la categorització d'X_2 i X_3. I no hi ha eines per saber *a priori* com s'ha de fer la *categorització* per tal que resulti adequada. Es produeix,

per tant, inestabilitat en el sistema, la qual cosa no és massa desitjable. Com a exemple, la classificació de **KLASS** es podria comparar amb els resultats que, per la prova segona, proporciona **SPAD**.

Els talls en 3 i 5 classes es mantenen idèntics en les dues classificacions, però amb qualsevol altre tall per sota d'aquest nivell ja no hi ha coincidència. Per exemple, amb talls en 6 classes, s'enregistren diferències en el 3.3% dels casos; un 8% en talls de 7 classes o un 23% si es fan 10 classes.

És a dir, si bé, per aquest cas concret, els talls de nivell superior coincideixen, — la qual cosa es deu, probablement, a l'estructura tant marcada que tenen les dades — l'ús de la distància mixta és una possibilitat diferent a la de la classificació sobre taules disjuntives completes.

8.3 Conclusions

Pel que fa a la primera de les palicacions presentades, es pot dir que el mecanisme de caracterització de les classes proposat permet detectar els trets més rellevants de les diferents classes, contribuint a una fàcil interpretació de les mateixes.

La distància mixta genera classificacions de més qualitat que altres mesures o sistemes de classificació. En l'aplicació del microordinadors, s'observa com la classificació utilitzant la distància mixta amb els valors dels paràmetres proposats a §**??** genera una caracterització automàtica realment propera, per no dir idèntica, a la un expert faria de la classificació.

A partir de l'estudi dels cubs, es poden treure les següents conclusions: Si totes les variables són contínues, **SPSS** i **KLASS** coincideixen totalment. **SPAD** tracta amb components principals, i el seu grau de coincidència dependrà de quant bé representin el núvol de punts les components seleccionades per a fer la classificació. Si es tracten totes, l'arbre pot variar però a nivells baixos.

Si totes les variables són categòriques, ni **SPAD** ni **SPSS** poden tractar la matriu de dades qualitatives directament. La classificació en mètrica euclídia de totes les components principals que es detecten en una anàlisi de correspondències múltiples, equival a la classificació en mètrica de χ^2 de la taula disjuntiva completa i es manté l'equivalència dels diferents procediments. Únicament **KLASS**, però, opera sobre la matriu original.

La classificació de matrius de dades no homogènies dóna lloc a diferents resultats per un paquet i altre. Mentre que **KLASS** dóna un resultat totalment determinat per la matriu de dades, en els altres paquets és necessari un pas previ de categorització de les variables contínues, font indubtable d'arbitrarietat, de la qual depenen els resultats.

Chapter 9

Disseny i implementació

"Ven, cosa, yo te diré tu nombre."
E.LIZALDE

Aquest capítol parla del disseny de KLASS profunditzant en l'estructura i particularitat de cada mòdul en el que el programa s'organitza.

9.1 Disseny modular

Com ja s'ha dit, KLASS és un sistema modular; s'ha dissenyat un mòdul per cada etapa de l'anàlisi de classificació. La comunicació entre els diferents mòduls, té lloc, per regla general, a través de fitxers externs, encara que també és possible treballar directament sobre memòria. Amb això, l'usuari pot decidir si efectua una anàlisi en una sola sessió, o si la realitza en etapes independents. L'arquitectura de mòduls del sistema es mostra a la figura **??**. En primer lloc, dir que KLASS pot funcionar a dos nivells diferents:

- Com un sistema que efectua un procés de classificació determinat a partir d'una matriu de dades i la metainformació associada.

- Com un sistema de classificació amb regles.

Després de detallar l'estructura dels fitxers que KLASS necessita per a treballar, es passa a descriure, cada un dels mòduls involucrats en el procés de classificació sense regles —en alguns casos s'indicarà la complexitat cas pitjor dels processos involucrats —. En darrer lloc, es presenta el mòdul de nivell superior[1].

[1]En les properes seccions podria sonar la sensació que hom s'ha distret de gran quantitat d'accents en descriure les diferents funcions que **KLASS** implementa. L'oblit només és aparent: **KLASS** no accepta accents, com tantes altres eines informàtiques, i els noms de les funcions de crida i dels seus paràmetres d'han transcrit en aquest texte *exactament* en la mateixa forma com **KLASS** els identifica i, per tant, sense accents.

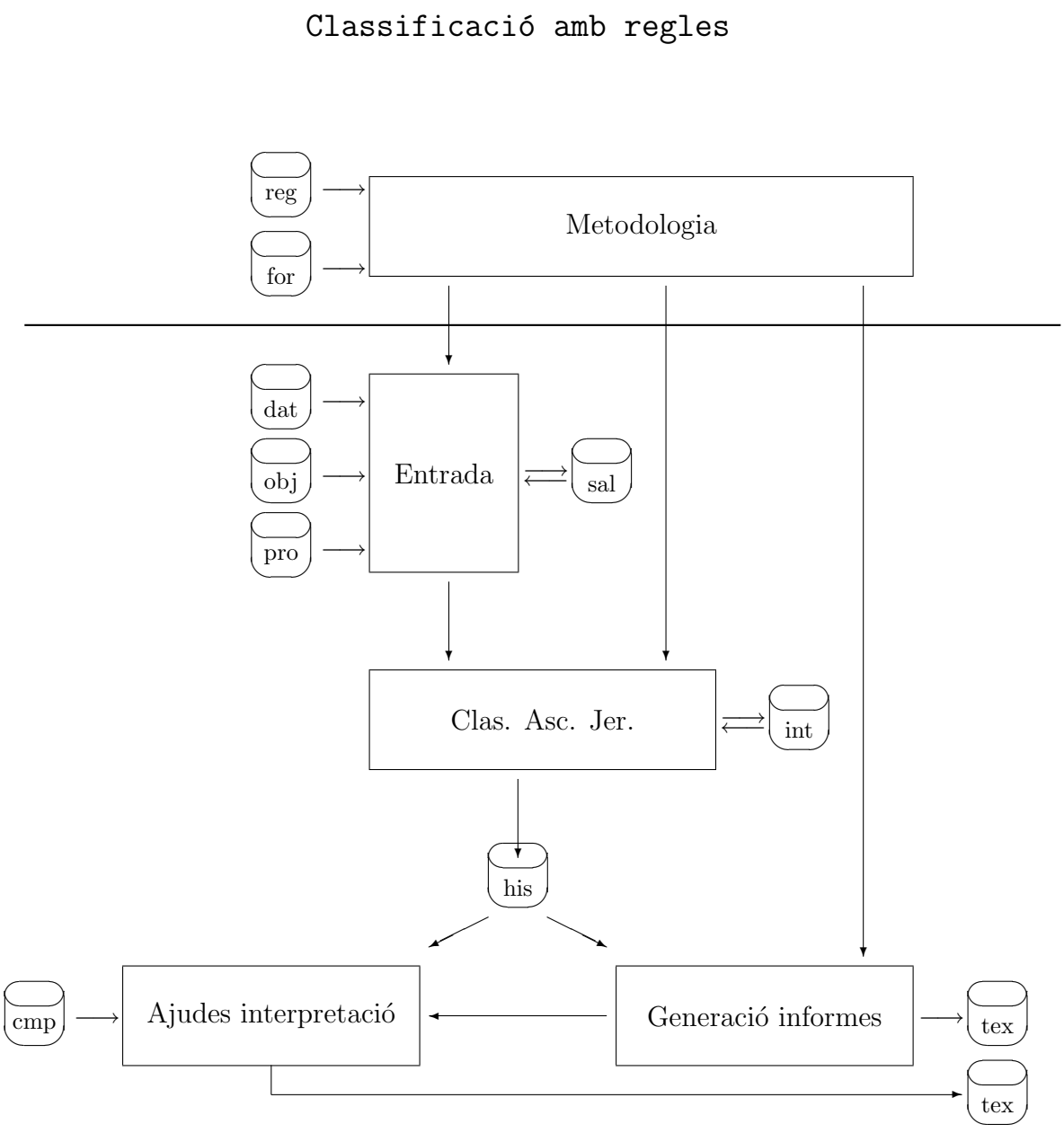


Figure 9.1: Arquitectura de **KLASS**.

9.1.1 Estructura de fitxers

Tal com es mostra a la figura **??** hi ha una sèrie de fitxers d'entrada que contenen la informació necessària per a que el classificador pugui treballar. En aquesta secció s'especifica l'estructura que han de tenir aquests fitxers. Ja s'ha comentat com, inicialment, **KLASS** complementava un entorn ja existent (**LINNEO**). En l'actualitat, si bé **KLASS** és un sistema totalment autònom, manté una estructura de fitxers majoritàriament compatible amb l'actual **LINNEO**$^+$, almenys pel que fa a les classificacions sense regles.

<**nom_fitxer.dat**> Conté la matriu de dades per files. Per a cada objecte hi ha una llista amb les coordenades que el defineixen en cada variable.

<**nom_fitxer.pro**> Conté metainformació referent a les variables, a les *propietats* segons les quals s'ha descrit els individus. Per a cada variable es té, apart del seu nom i un índex numèric associat, les següents dades:

1. Tipus de la variable: **Q :** Qualitativa, **C :** Numèrica
2. Pes: Per a estudis en els que calgui ponderar les propietats[2].
3. Nombre d'objectes que comparteixen aquesta propietat[3]. Per si hi ha propietats no aplicables a tots els individus. Per exemple, tractant persones, no té sentit el nombre de fills d'un nen.

A més, per a cada variable

qualitativa es té la llista de modalitats de la variable.

quantitativa es té el rang en que es mouen els valors de la variable.

<**nom_fitxer.obj**> Caracterització dels objectes de la mostra. Al igual que passa amb les propietats, hi ha un identificador de l'objecte, i un índex associat, així com també la llista de propietats que el descriuen, que sempre és un subconjunt de les que formen la matriu de dades[4].

En una classificació amb regles cal afegir els centres de les classes induïdes per aquelles al fitxer que conté els objectes de la classe residual. La informació addicional que fa possible la integració de les diferents subestructures en una única jerarquia és, per cada centre de classe:

1. el nombre d'objectes que s'hi aglutinen.
2. l'índex de nivell de l'arrel de la classe.

[2]En previsió de treballs futurs.
[3]**KLASS** no ho utilitza.
[4]Aquest paràmetre es manté per simple compatibilitat amb **LINNEO**$^+$.

3. l'estructura jeràrquica associada a ella.

<**nom_fitxer.reg**> L'estructura de les regles és totalment oberta i general — en aquest punt deixa d'haver-hi compatibilitat amb **LINNEO** que treballa amb altres sintaxis —. Essent:

$$(< funcio\ booleana\ avaluable\ en\ LISP >) - > < identificador\ de\ la\ classe >)$$

L'escriptura de les regles està subjecta a una única restricció: que els noms de les variables que apareixin a la condició estiguin parentitzats.

<**nom_fitxer.for**> Funcions en codi **LISP** que permetin fer les transformacions de les dades observades necessàries per a avaluar les regles. El fitxer <estrelles.for>, per exemple, conté les funcions per a calcular les coordenades cartesianes de la velocitat i la magnitud real de l'estrella a partir de les variables observades. Se'n mostra un exemple: El càlcul de la component

$$V = v_r \sin l \sin b + kr\mu_l \cos l \cos b - kr\mu_b \sin b \sin l$$

dóna lloc a la funció **LISP**:

```
(defun v (galon galat vr par mlcb mb)
        (let ((cosl (cos (rad galon))) (sinl (sin (rad galon)))
              (cosb (cos (rad galat))) (sinb (sin (rad galat)))
              (k 4741))
        (if (= par 0) 0
              (- (+ (* sinl sinb vr) (* cosl k (/ 1 par) mb cosb))
                 (* sinl sinb k (/ 1 par) mb)
              )
        ))
)
```

on els paràmetres simbolitzen variables que apareixen a la matriu de dades i que són necessàries per a la transformació. Al fitxer de regles, la crida a aquesta funció es fa seguint les normes generals de crides a funcions en **LISP**, utilitzant com arguments de la crida els identificadors que aquestes variables tenen en el fitxer propietats, convenientment parentitzats[5].

[5]L'usuari pot definir aquestes funcions com a *macros* sense paràmetres que es reescrivissin directament en funció de les variables involucrades. Llavors el propi mecanisme de lectura de **LISP** generaria el codi necessari per a susbtituir `(v)` per la seva expressió en totes les regles, i el procés d'execució instanciaria les variables implicades amb els valors dels diferents objectes.

9.1.2 Mòdul de lectura de dades

És el mòdul de càrrega de les dades. Ofereix bàsicament dues possibilitats:

- Càrrega d'un conjunt de dades i processament de la metainformació associada, creant i inicialitzant les estructures que conformen l'entorn adequat per a que el classificador comenci a treballar (funció **carrega**).
- Salvat (funcio **salva_est**) i recuperació (funció **carrega_rapid**) d'aquestes estructures.

Funció carrega

La primera alternativa se serveix de la funció
(carrega fitxer [modus metrica criteri obj_a_usar props_a_usar limit])
i el significat dels paràmetres és el que s'especifica a continuació:

fitxer És el nom, i eventualment directori, en relatiu o en absolut)[6] de la família de fitxers (<**fitxer**.dat>, <**fitxer**.obj>, <**fitxer**.pro>), que conté les dades a analitzar i la metainformació associada (veure §**??**).

modus paràmetre opcional[7] amb valor per defecte `no`. Indicarà si cal llegir les dades de la matriu amb algun tipus de transformació de normalització o no. De fet, pot ser útil de definir alguna mètrica sobre l'espai dels individus per tal que s'uniformitzi la importància que cada variable pugui tenir en l'anàlisi. Els valors que pot prendre aquest paràmetre són:

- `no` : No es fa cap transformació de les dades.
- `continuo`: Normalitza la matriu de dades utilitzant la inversa dels rangs. L'ús de la mètrica euclídia ordinària sobre una matriu així transformada, on les distàncies entre individus es calcularien com $\sum_{\forall k}\left(\frac{x_{ik}}{r_k}-\frac{x_{i'k}}{r_k}\right)^2$, equival a definir sobre la matriu original la mètrica normalitzada per la inversa dels rangs, on la distància entre individus es calcula com $\sum_{\forall k}\frac{(x_{ik}-x_{i'k})^2}{r_k^2}$, i suposa una solució alternativa a la utilització d'una distància euclídia normalitzada quan totes les variables són quantitatives.
- `binario`: Es transformen totes les variables en binàries en funció d'un cert límit. Les observacions inferiors al límit es codifiquen amb 0, i les restants amb un 1. El resultat és una matriu de variables 0/1.

metrica *p.o.* amb valor per defecte `indiferent` que té implicacions en el càlcul de la matriu de distàncies. Determina la mètrica de treball i pot prendre quatre valors:

[6] Segurament, la possibilitat d'indicar el camí relatiu depengui de la plataforma sobre la que corri el programa, i potser també de la implementació de **LISP** que s'utilitzi.

[7] En endavant s'utilitzarà la indicació (*p.o.*) per significar que un paràmetre és opcional.

- `euclidia`
- `eucl_norm` (normalitza amb la inversa de la desviació tipus).
- `chi_2`
- `mixte`

criteri *p.o.* amb valor per defecte `Ward` que indica el criteri d'agregació amb que es voldrà treballar. Només té implicacions pel que fa al càlcul de la matriu de distàncies. Pot prendre dos valors, acords amb els criteris d'agregació que s'han implementat (veure §**??**):

- `centroide`
- `ward`

obj_a_usar (*p.o.*) amb valor per defecte `tots`. Indica quins objectes, dels que apareixen en el fitxer d'entrada, intervindran en la classificació.

props_a_usar (*p.o.*) amb valor per defecte `totes`, que indica les variables que han de ser considerades en el procés de classificació. Amb l'anterior defineix una submatriu de la matriu de dades inicial.

limit *p.o.* S'ha d'especificar, *obligatòriament*, quan es demana una normalització `binario` de les dades, i correspon al punt de tall per la codificació en 0's i 1's.

Precondició Si la família de fitxers que contenen les definicions, o algun d'ells, no és al directori de treball, no es poden instanciar les dades. A la secció **??** s'indica quina estructura han de tenir aquests fitxers.

Funcionament Aquesta funció s'encarrega de llegir els fitxers on hi ha definides les dades, <**fitxer**.obj>, <**fitxer**.pro> i <**fitxer**.dat>, comprovar la coherència dels paràmetres de la crida; crear la submatriu de dades especificada per **obj_a_usar** i **props_a_usar**, procés que s'aprofita per a construir totes les estructures auxiliars associades[8] $[O(nobj \times nprops)]$; tractar els valors mancants $[O(npropsC) + npropsQ \times nmods_k + nmiss)]$, seguint el que s'ha presentat a §**??**; i, finalment, calcular la matriu de distàncies entre individus[9]. Citar que, per aquest darrer càlcul i, considerant que la matriu de distàncies és simètrica, s'optimitza

[8]Totes les constants que apareixen en les expressions de la complexitat de les diferents funcions al llarg d'aquest capítol venen definides a §**??**.

[9]La raó per la que es calcula en fase de lectura de dades la matriu de distàncies és doble: D'una banda, fer-ho en el procés de classificació podria carregar, excessivament la pila. D'una altra, hi ha un estalvi en temps substancial si **carrega_rapid** ja llegeix la matriu de distàncies calculada. En un entorn **VAX**, per exemple, el càlcul de la matriu de distàncies per les esponges i la classificació en un mateix procés generen el desbordament de la pila, mentre que la divisió d'aquestes dues fases capacita **KLASS** per a tractar matrius de dades més grans.

l'espai ocupat treballant únicament amb la submatriu triangular superior que té $\frac{nobj(nobj-1)}{2}$ elements, menys de la meitat de la matriu completa. El cost de construir aquesta matriu és $O(\frac{nobj(nobj-1)}{2} \times nprops)$, essent aquest el subprocés més car de l'etapa de lectura. D'altra banda, si es treballa en mètrica mixta, la construcció de la matriu de distàncies passa pel càlcul de les constants de ponderació, que es realitza segons la secció §**??**, a menys que s'especifiqui el contrari (a [**?**] es descriu un algorisme eficient pel càlcul d'aquestes constants). Els valors d'α i β queden definitivament instanciats per a etapes posteriors de l'anàlisi.

Funció salva_est

Sintaxi: **(salva_est f_sal)**

Únicament requereix un nom de fitxer **f_sal** (pot ser útil donar el mateix que identifica la família de fitxers que conté les dades) i crea un fitxer <**f_sal**.sal> que conté les estructures creades per la funció **carregar**. Permet interrompre la sessió per a prosseguir més tard i només té sentit executar-la després d'executar **carrega**.

Funció carrega_rapid

Sintaxi: **(carrega_rapid f_sal)**

Utilitza la informació del fitxer <**f_sal**.sal> per a reproduir l'estat en que es troba l'intèrpret despés d'executar la funció **carrega**.

9.1.3 Mòdul de classificació

Es pot dir que aquest és el cor del sistema: es tracta del mòdul de classificació pròpiament dita. Pot efectuar classificacions per veïns recíprocs encadenats utilitzant diferents mètriques i criteris d'agregació (funció **veins_rec**).

Tenint en compte que es pot haver de tractar conjunts de dades de grandària considerable, el sistema està preparat per a executar un nombre determinat de passes de classificació i interrompre el procés, per a continuar en una segona etapa (funció **seguir_classificant**).

Tot procés interromput genera un fitxer intermedi d'extensió <.int> on es guarda l'estat de les estructures de treball en el moment de la interrupció, el qual serà font d'alimentació del procés **seguir_classificant**. La raó per la que KLASS ofereix la possibilitat de classificar a passes té a veure amb la dimensió dels problemes que es tracten. Si el núvol de punts té molts objectes, la recursió generada pel procés de classificació pot desbordar la pila. Partir el procés en blocs, dóna més capacitat al sistema.

Funció veins_rec

Sintaxi: **(veins_rec l_objectes [matriu_distancies metrica alfa beta criteri n f_int])**

i la semàntica dels paràmetres es detalla tot seguit:

l_objectes és la llista d'objectes a classificar

matriu_distancies (*p.o.*). És la submatriu triangular superior de distàncies entre individus, calculada segons el valor dels paràmetres **metrica**, **criteri** i **alfa**, **beta** si es treballa en mètrica mixta. Si no s'indica altra cosa es pren la que calcula el procés **carrega**.

metrica (*p.o.*) amb la mateixa semàntica que a **carrega**.

alfa, beta (*p.o.*) amb valor per defecte el proposat a §**??**. A especificar només si es treballa en mètrica mixte. Són les constants de ponderació de les subdistàncies. Han de ser reals positius i sumar 1.

criteri (*p.o.*) amb valor per fecte el que s'ha indicat a **carrega**. Determina la forma com s'efectua l'agregació.

n_passes *p.n.* amb valor per defecte `tots`, que pot prendre valors enters entre 1 i $(nobj-1)$. Indica el nombre de passes de classificació que s'han de fer. Si no s'especifica, es fa la classificació completa. Si **n_passes**$< (nobj-1)$ es fan **n_passes** agregacions i es salva l'estat de l'intèrpret en el fitxer <**f_int**.int>.

f_int *p.n.* amb valor per defecte `klass` que només s'utilitza si no es fa la classificació completa. Indica el nom del fitxer extern en el que es salva l'estat intermedi de la classificació i que permetrà continuar el procés *a posteriori*. Es recomana utilitzar com a nom el mateix identificador de la família (<.obj>, <.pro> i <.dat>) de fitxers que defineixen les dades. Si no s'especifica, el resultat es guarda al fitxer <**klass**.int>.

Aquest funció crea, si s'interromp el procés abans de finalitzar, el fitxer intermedi <**f_int**.int>. Si, pel contrari, s'arriba al final de la classificació retorna l'identificador de l'arrel de l'arbre jeràrquic i crea un fitxer <**f_int**.his> que conté la informació associada a la classificació ascendent jeràrquica. Degut a l'estructuració de les dades que s'ha fet, a partir d'aquest únic símbol es pot accedir a tot l'arbre, ja que el valor assignat al propi símbol és la llista dels seus fills. Aquesta és la forma més condensada de representar un arbre, i mitjançant una simple avaluació del node, s'obtenenen els fills, en estructura de llista directament, que és la més pròpia per a fer tractaments recursius, seqüencials o exhaustius en LISP. Això ha fet que no s'optés per la implementació clàssica d'arbre binari via *apuntador* a fill dret i esquerra, que requereix un accés individualitzat a cada un dels fills d'un node. Paral·lelament, no obstant, i amb tal de facilitar determinades tasques, es manté una estructura amb tots els nodes de l'arbre ordenats per índex de nivell. A part de la funció principal, que és la que porta el control de l'execució, es pot considerar que aquesta funció té dos subprocessos fonamentals:

- Un de cerca de veïns recíprocs, que es basa principalment en la construcció de la cadena de veïns més propers, que és bàsicament una operació de consulta i
- Un d'agregació dels nodes identificats en la fase anterior, que és, en definitiva, un procés de modificació i actualització d'estructures.

A nivell de disseny, cal només definir les inicialitzacions d'aquestes estructures i els mecanismes d'actualització davant d'una agregació per a tenir el programa completament especificat.

Pel que fa a l'agregació, la creació d'un nou node aprofita l'existència del generador de símbols de LISP , creant un identificador per la nova classe de la forma |**classe**`Xn`|, on `X` és la inicial de l'identificador de les dades i té plè sentit quan es classifica amb regles, permetent distingir les classes induïdes per elles; `n` és un enter que s'assigna seqüencialment a cada classe començant per 1.

El procés d'agregació suposa bàsicament

- L'eliminació de les estructures de dades de la informació referent als nodes que s'estan agregant.
- L'addició de la informació corresponent al node que s'acaba de crear.

Com que totes les agregacions (s'està construint un arbre binari) agrupen dos nodes en un de nou, aquest procés suposa la disminució de les estructures de dades en un element. L'espai ocupat disminueix a mesura que s'avança en la classificació, propietat d'utilitat immediata pel fet que tots els processos de cerca són cada cop menys costosos, ja que treballen amb estructures més petites, i que resulta també molt interessant en processos de classificacions per blocs.

Funció seguir_classificant

Sintaxi: **(seguir_classificant f_int [n_passes])**

Aquesta és la funció que reanuda el procés de classificació interromput per una crida a les funcions anteriors amb un **n_passes**$< (nobj - 1)$.

f_int significa que a <**f_int**.int> s'hi troba la informació generada pels processos anteriors, que permetrà reproduir l'estat de l'intèrpret i reanudar la classificació.

n_passes *p.o.* amb valor per defecte `tots` que permet fer un nombre determinat de passes de classificació. Si no es volen fer tots, es generarà un nou fitxer <**f_int**.int> com a conseqüència d'una segona interrupció, i així successivament.

L'avantatge de fer classificacions per passes es basa en la propietat, ja esmentada, que les dimensions de les estructures de dades amb que es treballa es van reduint, i **f_int**.int

successius són cada cop més i més petits. En ser recuperats per l'intèrpret carreguen menys i menys recursos, permetent realitzar cada cop un nombre més gran de passes de classificació seguides. Així, la limitació en la grandària del núvol de punts ve determinada per la capacitat de KLASS en executar la primera agregació. Si això és possible, hi ha *garantia* que es pot realitzar la classificació completa, ja que les iteracions subsegüents mai necessitaran més recursos que la primera.

9.1.4 Mòdul d'ajudes a la interpretació

Un cop construït l'arbre ascendent jeràrquic, l'usuari ha de decidir un nivell per a efectuar el tall de l'arbre, i interpretar la partició resultant. En aquest mòdul, s'ofereixen algunes eines que el poden ajudar a desenvolupar aquestes tasques:

- Determinació del nivell del tall. S'ofereixen ajudes a nivell numèric i gràfic (funcions **dona_millors_talls, grafic_idn** respectivament).

- Realització d'un tall de l'arbre. Hi ha, en principi, dues formes d'especificar-lo: indicant el nombre precís de classes que es vol obtenir, o donant pròpiament el nivell al qual es vol practicar el tall. Algunes vegades se sabrà exactament *quantes classes* es vol (funció **k_classes**). D'altres, el dendrograma donarà idea de l'*alçada* a la que cal tallar (funció **alfa_tall**).

  Aquestes particions estan totalment definides pels identificadors de les classes que les constitueixen. Pot ésser d'interés, però, una descripció extensional, possibilitat coberta per la funció **extendre_classes**.

- Caracterització d'una partició de l'arbre (funcions **troba_descriptores** i **troba_caracteritzadores**).

- Comparació entre classificacions. En aquest aspecte, KLASS està preparat per a rebre altres classificacions del mateix conjunt de dades, provinents d'altres entorns — automàtics o no —, i efectuar les comparacions amb les que ell mateix pugui produir (funcion¡s **fer_tcpp** i **d_ck**).

Tot seguit es descriu breument cada una de les funcions que integren aquest mòdul.

Funció dona_millors_talls

Sintaxi: **(dona_millors_talls arbre [n_talls])**

Implementa l'heurística presentada a la §**??**, retornant la llista dels **n_talls** millors talls que es poden fer en l'arbre **arbre** [$O(n_talls \times nobj)$].

arbre conté l'estructura jeràrquica construïda pel classificador. Conté la llista de tots els nodes de la jerarquia, ordenats per índex de nivell i començant per l'arrel.

n_talls *p.o.* amb valor per defecte `5`. Indica el nombre de talls que es vol i dóna els 5 millors si no s'especifica el contrari.

Funció grafic_idn

Sintaxi: **(grafic_idn f_tex arbre [tipus])**

Des del punt de vista gràfic, el que fa **KLASS** és generar codi font de LaTeX. El processament d'aquest codi permet visualitzar les diferents representacions dels índexos de nivell de l'arbre, amb l'objectiu de decidir per on tallar-lo [$O(nodes_de_l'arbre) = O(nobj - 1)$].

f_tex nom del fitxer on es guardarà el codi en LaTeX del gràfic. Caldrà processar el fitxer <**f_tex**.tex> per a obtenir la representació gràfica.

arbre llista ordenada per índex de nivell de tots els nodes de l'arbre ascendent jeràrquic.

tipus *p.o.* amb valor per defecte `:punts` que indica quin tipus de gràfic es desitja. Pot prendre dos valors:

- `:punts` presenta el gràfic dels índexos de nivell successius, segons s'introdueix a l'apartat §**??**. Anomenant a_1 a l'arrel de l'arbre, i $idn(a)$ a l'índex de nivell d'un node qualsevol, el gràfic queda definit en l'espai $[0, nobj + 1] \times [0, idn(a_1)]$ i representa, per cada node a, el punt $(i + 2, idn(a'))$, on i és la posició que ocupa a en **arbre**.
- `:barres` fa la representació estàndard en *histograma* dels índexos de nivell ordenats de major a menor. En aquest cas per a cada node a es representa un línia horitzontal de coordenades $(i, idn(a))$ i les dimensions del gràfic són $[0, idn(a_1)] \times [0, nobj - 1]$.

Funció k_classes

Sintaxi: **(k_classes n_classes arbre)**

Donat un **arbre** retorna la partició corresponent a fer **n_classes** d'aquest arbre.

n_classes nombre de classes que es vol. Ha de ser un enter entre 1 i $nobj$.

arbre estructura jeràrquica generada pel procés de classificació pròpiament dit. Conté la llista de tots els nodes de la jerarquia, ordenats per índex de nivell i començant per l'arrel.

El fet que **arbre** contingui tots els nodes de la jerarquia ordenats per índex de nivell facilita notablement la construcció de la partició. Partint de l'arrel de l'arbre cal anar dividint el node de major índex, fins obtenir una partició en **n_classes** elements. Correspon a fer un recorregut en amplada de l'arbre jeràrquic, eliminant el node que es desdobla en cada passa $[O(\frac{n_classes(n_classes-1)}{2})]$.

Funció alfa_tall

Sintaxi: **(alfa_tall nivell arrel)**

Aquesta funció practica un tall de nivell **nivell** sobre l'arbre que té per arrel **arrel**. A partir de l'arrel es va baixant en l'arbre, fent un recorregut en profunditat, que s'atura en cada branca quan s'arriba a un node d'índex de nivell inferior a **nivell**. En el cas pitjor, **nivell**=0 i caldrà recórrer tot l'arbre $[O(nobj)]$.

nivell ha de ser un valor entre 0 (que donaria una classe diferent per cada objecte) i l'índex de nivell de l'arrel de l'arbre (que donaria un tall en una sola classe).

arrel identificador del node arrel de la jerarquia.

Funció extendre_classes

Sintaxi: **(extendre_classes particio)**

particio representa un conjunt de classes. És una llista d'identificadors de classes $(\mathcal{C}_1\ \mathcal{C}_2 \ldots \mathcal{C}_\xi)$.

Retorna una llista de la forma $((\mathcal{C}_1(i \in \mathcal{C}_1))(\mathcal{C}_2(i \in \mathcal{C}_2)) \ldots (\mathcal{C}_\xi(i \in \mathcal{C}_\xi)))$. Simplement fa un recorregut en profunditat del subarbre representat per cada classe, fins obtenir-ne totes les fulles $[O(2nobj - n_classes + 1)]$. En endavant es farà referència a aquesta forma de representar les classificacions com *classificació en format extès*.

Funció troba_descriptores

Sintaxi: **(troba_descriptores particio)**

Retorna la llista de variables descriptores de cada una de les classes i els valors descriptors. La crida únicament requereix el paràmetre:

particio: llista d'identificadors de les classes que es vol descriure.

epsilon: *p.o.* amb valor per defecte `0`. A menys que s'indiqui el contrari es fa una descripció en sentit estricte. Ha de ser un real de [0,1].

Havent calculat els estadístics bàsics de cada classe durant la classificació, si $\varepsilon = 0$ el procés és $O(n_classes \times nprops)$, si $\varepsilon > 0$, és $O(n_classes \times nmods + nprops)$

Funció troba_caracteritzadores

Sintaxi: **(troba_caracteritzadores particio l_propietats)**

Detecta les classes carateritzades totalment amb o sense condicionants per una variable, tot indicant quines són les variables caracteritzadores i els valors característics, i caracteritza parcialment les classes restants, seguint el procediment presentat a §**??**.

particio tall horitzontal de l'arbre ascendent jeràrquic que es vol caracteritzar.

l_propietats llista de variables emprades en el procés de classificació.

Funció fer_tcpp

Sintaxi: **(fer_tcpp classif1 classif2)**

Aquesta funció retorna la taula de contingencia de la particio **classif1**×**classif2** tal com s'ha definit a la §**??** $[O(nobj^2)]$.

classif1, classif2 són les dues classificacions que es vol comparar, en format extès.

Precondició Cal que **classif1, classif2** facin referència *exactament* al mateix conjunt d'individus.

Funció d_ck

Sintaxi: **(d_ck classif1 classif2 nobj)**

Dóna una mesura de les diferències existents entre dues classificacions del mateix conjunt d'objectes, d'acord amb el mètode presentat a §**??** $[O(nobj^2, classif1 \times classif2)]$.

classif1, classif2 tenen el mateix significat que a la funció anterior.

nobj indica el nombre d'objectes que s'està classificant.

Sobre la Tcpp de `classif1` i `classif2`, es calcula el màxim de cada fila i de cada columna. Mitjançant un procés de marcatge simultani s'identifica les posicions centrals de la Tcpp, que són alhora màxim de la seva fila i de la seva columna, i es resten del total $nobj$.

9.1.5 Documentació de sortida

Aquest mòdul ofereix la possibilitat d'extreure fitxers de resultats en codi LaTeX (funció **fer_informe**) que, tractats amb aquest processador de textes, produeixen documents de qualitat netament superior a altres paquets comercials com **SPAD** o **SPSS** que generen fitxers ASCII amb les gràfiques i els dendrogrames presentats en resolució de caràcter. Existeix també l'alternativa de demanar un informe estàndard que es construeix automàticament (funció **informe_automatic**).

La decisió de generar directament codi font en LaTeX obeeix principalment a dues raons:

- El resultat de LaTeX és un fitxer postscript que genera informes de qualitat molt superior a un ASCII[10].

- Utilitzar LaTeX com a interfície de sortida dóna altes prestacions gràfiques que no es poden assolir treballant amb resolució de caràcter. **KLASS** pot generar les primitives gràfiques de LaTeX directament, havent d'assegurar simplement la seva correcció sintàctica. És aquest darrer el que gestiona qüestions més tècniques com la definició de les unitats de treball.

Hom és conscient que la utilització de **KLASS** queda, d'aquesta forma, condicionada al fet que l'usuari disposi d'un compilador i un visualitzador (o una impressora làsser en el seu defecte) de LaTeX, condició que no sempre es donarà. Ja que la generació de sortides ASCII no té més dificultat que la de sortides LaTeX, s'ha volgut donar prioritat a aquestes darreres, per les raons exposades. De tota manera, es preveu la generació de sortides en pantalla en un futur proper.

Funció fer_informe

Sintaxi: **(fer_informe f_tex tipus_class alfa-n_clas [prefixes idn_graf])**

f_tex L'informe es generarà en el fitxer <**f_tex**.tex>.

tipus_class indica si el tall de l'arbre es demana en funció del nombre de classes o del nivell de tall i determina la interpretació del següent paràmetre de la crida. Pot prendre dos valors:

- `idn` Es demana un tall indicant ell nivell d'aquest.
- `kclas` Es demana un tall indicant el nombre de classes.

alfa-n_class Si el paràmetre anterior era `idn`, indica el nivell al que es vol fer el tall de l'arbre. En aquest cas, ha de ser un real entre 0 i l'índex de nivell de l'arrel.

Si **tipus_class** val `kclas`, llavors **alfa-n_class** està indicant el nombre de classes en que es vol particionar el conjunt d'individus. El seu valor ha de ser, llavors, un enter entre 0 i *nobj*.

prefixes *p.o.* amb valor per defecte `:si` que indica si es vol un dendrograma amb identificadors de les classes en els nodes interns o no. Pot prendre dos valors:

- `:si` el dendrograma es representa amb el nom de les classes. A §**??** es pot observar l'aspecte d'aquests dendrogrames.

[10]És preferible la impressió en impressora làsser, tot i que el postscript es pot enviar fins i tot a una impressora de línies, amb el consegüent empobriment de l'aspecte de l'imprés.

- `:no` el dendrograma queda net, exceptuant els identificadors de les fulles de l'arbre.

idn_graf *p.o.* amb valor per defecte `:no` que indica si l'informe ha d'incloure o no una representació gràfica dels índexos de nivell de l'arbre. Pot prendre tres valors que, a excepció del primer, s'interpreten igual que els de la funció **idn_graf**.

- `:no` l'informe no inclou cap gràfic d'índexos de nivell.
- `:punts`
- `:barres`

Funció informe_automatic

Sintaxi: **(informe_automatic f_tex)**

Únicament cal indicar en quin fitxer es vol el resultat (<**f_tex**.tex>). L'estructura de l'informe és l'habitual amb la particularitat que és el propi paquet qui decideix quantes classes es fan, en base a l'heurístic plantejat a §**??**. S'inclou, a més, gràfic d'autocorrelació dels índexos de nivell i una llista amb els 5 millors talls segons l'heurística, pel cas que l'usuari no accepti el que automàticament s'ofereix.

Elements de l'informe

El document conté la informació bàsica per a interpretar els resultats de la classificació.

En primer lloc hi ha una descriptiva de la matriu de dades tractada: fitxer on es troba i dimensions, seguit dels paràmetres de la classificació pròpiament dita (criteri, mètrica — si és la mixta, indica els valors d'α i β).

- Cas que hi hagi valors mancants, se'n dóna el percentatge.
- Ja sia utilitzant l'heurística o a petició de l'usuari, es disposa d'una partició de l'arbre de classificació. Després de la informació inicial es descriuen les classes resultants. Per a cada classe s'imprimeix el seu centre de gravetat, la seva grandària i finalment una descripció extensional completa.

  Pel que fa a la primera, a més de les pròpies coordenades del centre de gravetat, es proporciona, per les variables contínues, la desviació dins la classe. Les variables descriptores de cada classe, es marquen amb lletra tipogàfica, a diferència de les restants, que apareixen amb cursiva.

  Quant a la presentació de les descripcions dels centres de gravetat de les classes, es podrien construir automàticament taules com les de §**??**, però això requereix precaucions quant a les dimensions d'aquestes i el fet que es puguin encabir en una única plana. El cas de les esponges, és un exemple clar que això no és possible, donat el gran nombre de variables que intervenen en la classificació. S'ha optat, doncs, per

una representació seqüencial que no depengui del nombre de variables, ni tampoc del de classes.

- Un altre element de la sortida de **KLASS** és el gràfic dels índexos de nivell, que pot ser el clàssic o el que es proposa en aquesta tesi. La seva aparició en l'informe és opcional i està orientada a facilitar la determinació del nivell del tall que convé fer. **KLASS** ho pot decidir de forma automàtica i, si és així, proporciona també una llista d'altres talls recomanables pel cas que l'usuari no accepti el que automàticament ha ofert.

- Finalment apareix l'arbre de classificació.

  L'arbre s'ubica en l'espai $[0, idn(arrel)] \times [1, nobj]$ convenientment reescalat. Es representa un eix vertical en el lateral esquerra que marca les unitats per tal que hom tregui, a simple vista, idea dels índexos de nivell corresponents a cada node. El codi que identifica els diferents objectes de les fulles de l'arbre està posat en vertical. Les coordenades dels nodes interns són funció de l'índex de nivell del node, en sentit vertical, i del nombre de fills que li queden a l'esquerre, en sentit horitzontal. En els apèndixs es trobaran diverses representacions d'aquests arbres.

  S'aprofiten les capacitats gràfiques de LaTeX i les facilitats recursives de **LISP** per a dibuixar el dendrograma de forma vertical, a diferència del que fan molts paquets comercials, que els presenten en horitzontal. Si la mostra és molt gran, això suposa un problema de solapament dels nodes terminals i dificulta la lectura de l'arbre.

  No obstant, es pot demanar un arbre general sense noms de classes, i partir després dels nodes intermedis que interessi per a generar fitxers amb els subarbres corresponents ampliats. Això és possible cridant directament la funció que pinta l'arbre (funcio **pinta-arbre**).

No cal dir que les operacions de reescalat de tots els gràfics es fan en la forma adequada per a encabir-los en fulls Din-A4.

9.1.6 Mòdul de classificació amb regles

Per damunt de tot això, hi ha encara un mòdul que gestiona el procés de classificació amb regles.

Es distingeixen, en aquest procés, dues fases:

- Lectura de les regles <.reg>i càlcul de la partició que aquestes indueixen sobre el conjunt d'individus. La definició de les funcions que relacionen les variables transformades (que intervenen en les regles) amb les observades fa ús d'un fitxer extern d'extensió <.for>, l'estructura del qual s'ha detallat a §**??**.

- Procés de classificació local a les classes i integració en una única estructura global (funció **integracio**).

Funció particio

Sintaxi: **(particio fitxer)**

Després de carregar normalment les estructures de dades, i abans de calcular la matiru de distàncies entre individus, utilitza els fitxers <**fitxer**.for> i <**fitxer**.reg> per a particionar el conjunt de dades i genera, per cada una de les classes, una subfamília de fitxers <.dat>, <.pro>, <.obj> en el format estàndard que es podrà classificar normalment. A part, també produeix la llista que conté la partició induïda per les regles en el format extès habitual.

fitxer família de fitxers que identifica les dades a analitzar (<**fitxer**.obj>, <**fitxer**.pro>, <**fitxer**.dat>, <**fitxer**.reg>, <**fitxer**.for>).

A més, aquesta funció té, com a paràmetres opcionals, exactament els mateixos que la funció **carrega**.

Funció integracio

Sintaxi: **(integracio particio [situacio ponderacio f_pare])**

Procediment que realitza classsificacions locals a totes les classes induïdes per les regles. Fet això, afegeix els centres de gravetat d'aquestes classes al fitxer que conté les dades de la classe residual. Finalment, efectua la classificació de la classe residual, enllaçant la informació dels arbres locals a cada classe en una única estructura global.

particio llista de classes induïdes per les regles.

situacio *p.o.* amb valor per defecte `:inicial` que indica l'estat del procés. Pot prendre tres valors:

- `:inicial` el procés s'ha d'iniciar des del principi
- `:estruct_creades` les dades ja han estat carregades prèviament i es pot recuperar el fitxer <**f_pare**.sal>
- `:enmig` en sessions anteriors s'han executat algunes passes de la classificació, i cal recuperar el fitxer <**f_pare**.int>

ponderacio *p.o.* amb valor per defecte `:local` que indica, cas de treballar en mètrica **mixta**, si es volen utilitzar valors per α i β comuns en totes les classificacions locals o, pel contrari, es vol classificar independentment cada una de les classes induïdes per les regles amb les seves pròpies constants de ponderació. Des d'un punt de vista teòric, la integració posterior en

un únic arbre de classificació té sentit si les constants són les mateixes per totes les classes. No obstant, s'ha mantingut aquesta possibilitat en vistes a recerques posteriors. Aquest paràmetre pren dos valors:

- `:local` cada classe utilitza constants pròpies que no tenen perquè coincidir entre elles.
- `:global` s'utilitzen els mateixos valors d'α i β en les classificacions locals i en la global.

f_pare cas que **ponderacio** sigui `:global`, cal calcular els valors de les constants sobre la matriu de dades que conté tots els objectes. Aquí cal especificar el nom de la família de fitxers on es defineixen les dades.

Precondicions És imprescindible que el primer element de la particio identifiqui la classe residual, i que els identificadors de les diferents classes tinguin inicials diferents. De no ser així es produïrien repeticions en la generació de símbols per les classes creades pel procés de classificació, els quals utilitzen aquestes inicials, amb la conseqüent ambigüitat en la identificadió d'aquestes.

9.2 Estudi de la complexitat de l'algorisme

En aquesta secció es fa un estudi de la complexitat cas pitjor de l'algorisme dels veïns recíprocs per matrius heterogènies i també del procés de classificació amb regles, que són la base de KLASS. La complexitat cas pitjor de les funcions restants no presenta particularitats i vé indicada en les descripcions de la secció anterior. En qualsevol cas, els processos més cars són els que s'estudien tot seguit. Es veurà que la classe de complexitat cas pitjor de l'algorisme és $O(n^2)$, la qual cosa no representa dificultats en el processament de conjunts nombrosos de dades.

Per tal de facilitar la lectura, es defineixen les següents constants:

nobjstots: nombre d'objectes descrits a <**fitxer**.obj>.

npropstots: nombre de propietats descrites a <**fitxer**.pro>.

nobj: nombre d'objectes que es classifiquen ($\leq nobjstots$).

nprops: nombre de propietats que s'utilitzen en la classificació ($\leq npropstots$).

npropsC: nombre de propietats quantitatives que s'usen en la classificació.

npropsQ: nombre de propietats qualitatives que s'usen en la classificació.

$nmods_k$: màxim nombre de modalitats que presenta una variable categòrica ($\max_k \mathcal{D}_k$).

nmods: nombre total de modalitats que intervenen en la classificació($\sum_{k\in\mathcal{Q}} card(\mathcal{D}_k)$).

nmiss: nombre de valors mancants de la matriu de dades.

nreg: nombre de regles del fitxer <**fitxer**.reg>

9.2.1 Funció veins_rec

L'algorisme crea $nobj - 1$ nodes interns de l'arbre binari a base de repetir un mateix procediment sobre estructures de dades cada cop més petites. Les estructures sobre les que es treballa es dimensionen, en una passa ξ, de la següent forma:

l_objectes llista d'$(nobj - \xi)$ (a cada passa s'esborren de la llista d'objectes a classificar dos fills i s'hi afegeix un pare).

l_mes_proper dimensiona i evoluciona en paral·lel a l'anterior.

matriu_distancies submatriu triangular superior de dimesió $(nobj - \xi) \times (nobj - \xi)$. En total té $\frac{(nobj-\xi)(nobj-\xi+1)}{2}$ elements.

En una passa qualsevol de l'algorisme ξ tenen lloc, bàsicament, les següents etapes:

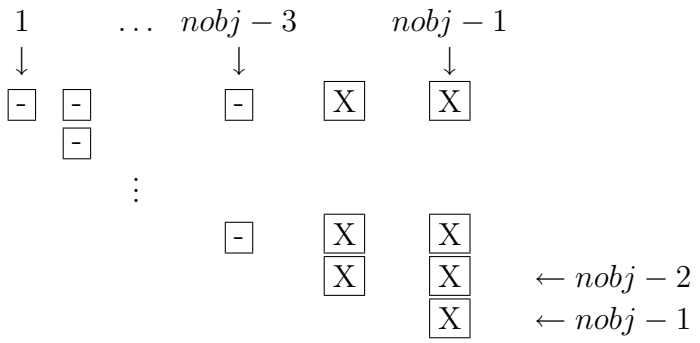
- Construir la cadena de veïns més propers fins localitzar dos nodes que siguin veïns recíprocs. En el pitjor cas, caldrà recórrer tota la **l_objectes**: $O(nobj - \xi)$. Sobre aquest punt s'insistirà una mica més endavant.

- Crear un nou node que els representi. Procés de cost constant en totes les iteracions, degut al disseny d'operacions recursives per a calcular el centre de gravetat de la classe: $O(nprops)$.

- Eliminar-los de les estructures de dades. Sota una implementació directa de la submatriu triangular superior, ja sia com a taula bidimensional o com a llista de subllistes, les tasques més cares d'aquest subprocés serien:

  - Eliminar les files i columnes corresponents de la **matriu_distancies**. En el pitjor dels casos, caldria recórrer tota la matriu, per a eliminar les dues darreres fila i columnes (si la implementació fos amb llistes) o les dues primeres (si la implementació fos via taula): $O\left(\frac{(nobj-\xi-3)(nobj-\xi-2)}{2}\right)$. En el millor, serien les dues primeres (per les llistes) o les dues darreres (per les taules), en el qual cas no

s'hauria d'iterar (cost constant).



Com que aquest és un procés que cal repetir $n-2$ vegades, a la llarga, generaria un factor cúbic en la complexitat total ($O(\frac{n^3}{6})$, per ser precissos).

Fent un abús d'espai i prescindint d'eliminar la informació obsoleta d'aquesta matriu, es treballaria cada cop amb files més llargues i més esparses, però la complexitat en cas pitjor disminuïria en total a $O(4n^2)$.

Considerant, però, que l'agregació de dos nodes deixa *sempre* dues files i dues columnes de la matriu inservibles, es pot actualitzar aquesta matriu seguint un procés de reescriptura que:

* redueixi l'espai ocupat en cada iteració
* el cost de fer-ho sigui *constant* (això rebaixa *dos* graus la complexitat final!).

Per assolir aquests objectius cal concebir la matriu de distàncies, no com una estructura compacta, sino com una entitat fragmentada i associada als diferents objectes (organització que fa ús de la llista de propietats que **LISP** permet associar a tots els símbols).

Mantenint el càlcul explícit de la submatriu triangular superior només, es pot dir que la clau de la qüestió està en els següents punts:

* associar les *subfiles* de la matriu de dades a cada un dels objectes que representen
* a cada agregació, l'eliminació de les files dels nodes que s'agreguen va implícita al fet d'esborrar aquests objectes de la llista d'objectes (la qual cosa té un cost constant): $O(n)$.
* l'eliminació de les columnes no es porta a terme en sentit estricte: la més curta de les dues serà reescrita amb les distàncies dels diferents objectes al nou node creat, el qual ocuparà també la fila connectada amb aquesta columna. L'altra, serà simplement marcada com a columna buida per tal que la cerca de veïns més propers a un donat (en les passes subsegüents de l'algorisme) no la consideri més. Per a optimitzar les cerques posteriors és convenient mantenir

un marcador amb la darrera columna útil de la matriu. Amb això, l'espai *mai creix* i, sota certes circumstàncies (s'ha eliminat la darrera columna de la matriu), es pot arribar a reduir. I el cost d'aquestes operacions també és *constant*.

Per tant, el procés que, en principi, resultaria el més car de tots ha passat a un cost *constant*, tot mantenint la reducció successiva d'espai al llarg del procés.

– Actualitzar la llista *l_mes_propers*. En el pitjor dels casos, tots els objectes que estan per classificar tenen per veí més proper algun dels nodes que s'acaben d'agregar. En aquest cas caldria recalcular el veí més proper de tots els nodes. El procés passa pel càlcul del mínim d'una fila de la **matriu_distàncies**: $O((nobj - \xi)(nobj - \xi - 2))$ i ens portaria a un cost global $O(n^3)$. Aprofitant l'estructura encadenada de la implementació i el fet que les agregacions s'efectuen en ordre a aquesta cadena, el que es fa en aquesta etapa és simplement invalidar aquests elements, per a recalcular-los només quan sigui necessari. Amb això, l'operació redueix el seu cost a $O(nobj - \xi - 2)$, en detriment de la construcció de la cadena de veïns, durant la qual pot ser necessari un càlcul complet del veí més proper d'un element donat.

  Succeeix, però, que, per anar-se reduint el núvol de punts en cada iteració, i classificar de forma encadenada (veure [**?**] per a una demostració al respecte), el cost d'aquesta operació es manté lineal a cada iteració, i quadràtic en tot el procés.

- Incloure el nou node en aquestes estructures, té com a únic subprocés rellevant el càlcul de la nova fila de **matriu_distancies**: $O(nobj - i - 2)$, ja que les distàncies es calculen també de forma recurrent.

El procediment es repeteix per les $nobj - 1$ agregacions que donen lloc a l'arbre binari i la complexitat cas pitjor global és:

$$O(nobj^2, nobj \times nprops)$$

Per tant, l'algorisme té una complexitat, en cas pitjor, quadràtica en el nombre d'elements a classificar.

9.2.2 Funció particio

Per a cada objecte s'avaluen totes les regles: $O(nobj \times nreg \times K)$, on K és el cost d'avaluació màxim d'una regla i probablement dependgui del nombre de variables que hi vinguin implicades. També hi intervindria el cost de calcular les possibles variables

transformades amb ajuda del fitxer <**fitxer**.for>, aquest últim terme fora, ja, del control intern del programa.

La implementació que s'ha realitzat (veure §**??**) suposa un cost constant (el d'accés per posició al vector de dades) en la instanciació dels valors de les variables per cada objecte.

9.2.3 Funció integració

Per a cada classe es fa una classificació completa. Després es classifica, seguint el procediment habitual la classe residual incrementada amb els centres de les classes induïdes per les regles. Les classificacions locals a les classes tenen el cost de:

- Carregar els $m_{\mathcal{C}}$ objectes de la classe i construir les estructures associades:

$$O(\quad npropstot,\ m_{\mathcal{C}},\ m_{\mathcal{C}} \times nprops,\\ nprops\mathcal{C} + nprops\mathcal{Q} \times nmods_k + nmiss_{\mathcal{C}}, \frac{m_{\mathcal{C}}(m_{\mathcal{C}}-1)}{2}, nmods)$$

- Efectuar la classificació: $O(4m_{\mathcal{C}}^2, m_{\mathcal{C}} \times nprops)$

El cost total és:

$$\left(\sum_{\mathcal{C}\in\mathcal{P}_{\mathcal{R}}} (classificacio\ local\ a\ \mathcal{C})\right) + (classificacio\ classe\ residual\ ampliada)$$

Com que la suma de quadrats sempre és menor que el quadrat de la suma, el cost de totes les classificacions locals a les classes és del mateix ordre que la classificació sense regles. Tampoc la classificació de la classe residual fa canviar de classe de complexitat, donat que la grandària d'aquesta classe serà com a molt $nobj - nclasses + nclasses = nobj$, suposant que les classes induïdes per les regles són de grandària mínima, contenint un únic element.

Chapter 10

Conclusions, línies obertes i treball futur

Conclusions

De la presentació de KLASS es poden fer algunes reflexions sobre els problemes que s'han intentat resoldre i la forma com s'han abordat.

Un primera característica del nostre enfoc és la integració de tècniques habituals en Estadística i Intel·ligència Artificial, que s'han descrit respectivament als capítols §?? i §??, en una única eina de classificació capaç de treballar amb dominis poc estructurats. En definitiva, s'ha aplicat un mètode de classificació ascendent jeràrquica (concretament el dels veïns recíprocs encadenats) amb i sense informació de suport. Alguns dels resultats més destacables es comenten tot seguit.

D'una banda, la definició (§??) de la família de distàncies mixta

$$d^2_{(\alpha_r,\beta_r)} = \alpha \sum_{\forall k \in \mathcal{Q}} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \beta \frac{1}{n^2_{\mathcal{Q}}} \sum_{\forall k \in \mathcal{Q}} d^2_k(i,i')$$

que permet calcular distàncies entre individus descrits parcialment amb informació qualitativa i parcialment cuantitativa, tot obviant les transformacions prèvies de la matriu de dades, com categoritzacions o codificacions. D'aquesta forma s'evita la introducció de qualsevol arbitrarietat que puguin produir les categoritzacions o l'augment significatiu en les dimensions de la matriu que produïria el desdoblament de les variables en el contexte dels dominis poc estructurats o, fins i tot, l'ús, previ a la classificació pròpiament dita, de tècniques auxiliars com l'anàlisi de correspondències múltiple.

Després de fer un estudi teòric sobre les propietats d'aquesta família s'ha trobat la condició necessària i suficient per a que $d^2_{(\alpha_r,\beta_r)}$ mantingui l'estructura de distància (§??) i es proposen uns valors pels paràmetres (α_r, β_r) que, d'una banda, garantitzen la satisfacció d'aquesta condició i, de l'altra, representen un ús equilibrat de la informació qualitativa i quantitativa en la distància (§??). També en el pla teòric, s'ha estudiat si la distància mixta satisfà tots els requeriments necessaris per a ser utilitzada amb el criteri de Ward

en l'algorisme dels veïns recíprocs encadenats. Això és possible si l'augment d'inèrcia amb distàncies mixtes satisfà l'axioma de reductibilitat, propietat que es comprova a §**??**.

La definició d'un representant de classe en un entorn en el que s'han utilitzat variables qualitatives per a la descripció dels individus és un altre consecució significativa del treball, per diverses raons. Destacar que la forma com es construeix el representant d'una classe:

- és prou informativa per a generar una caracterització posterior de les classes satisfactòria. A la secció §**??** es descriuen formes de caracteritzar una classificació utilitzant els representants de les classes formades, com es comentarà més endavant.

- permet fer un tractament idèntic de les variables qualitatives i quantitatives pel que fa al procés de classificació, ja que la forma de les recurrències en el càlcul del representant de classe o de les distàncies d'un nou node als restants coincideix per unes i altres (§**??** i §**??** respectivament).

- identificar una classe amb el seu representant permet fer un tractament homogeni de classes i individus pròpiament dits, característica especialment rellevant en el procés que incorpora informació declarativa al sistema.

Quant a la implementació, dir que s'han desenvolupat expressions de càlcul recurrent tant de les distàncies entre individus §**??**, com dels representants de classe §**??**. Els càlculs corresponent a un nou node de la jerarquia es fan amb la informació associada als dos fills que s'agreguen. Això fa que les actualitzacions de les estructures de dades corresponents tinguin cost constant en les successives agregacions i no depenguin de la grandària de les classes que es formen, optimització apreciable en les darreres passes de l'algorisme, en que les classes inclouen gran nombre d'individus. Tot plegat condueix a un procediment de complexitat cas pitjor $O(n^2)$.

En un aspecte més aplicat, el mètode de classificació desenvolupat suposa una alternativa realment diferent a les tècniques de clustering existents. Com s'ha pogut veure a la secció §**??**, les classificacions amb distància mixta produeixen resultats propis, diferents dels que puguin produir altres procediments (l'aplicació sobre els microordinadors ho mostra) o altres sistemes estadístics (veure §**??** on es compara **KLASS** amb **SPSS** i **SPAD**) que no poden fer classificacions amb matrius heterogènies directament. Les classificacions de **KLASS** amb la distància mixta no requereixen cap mena de transformació prèvia sobre la matriu de dades. En particular, s'evita el desplegament de les variables categòriques, causa d'augment en les dimensions de la matriu de dades, així com la categorització de variables numèriques, font indubtable d'arbitrarietat, amb implicacions sobre els resultats de l'anàlisi. En contrast amb aquesta darrera situació, els resultats de **KLASS** , fixats uns paràmetres (α_r, β_r), són estables.

D'altra banda, a la vista de les interpretacions que es poden fer de les classificacions de **KLASS** (l'aplicació §**??** n'és un clar exemple) quan s'usen els valors d'α_r i β_r proposats a §**??**,

es podria dir que l'ús de la distància mixta amb aquests paràmetres millora la qualitat dels resultats.

Malgrat tot, quan es tracta de dominis poc estructurats ni tant sols la classificació usant la distància mixta resulta suficient (seria el cas de les esponges). Considerant les característiques que es donen en aquest tipus de dominis — descrites a §**??**—, s'introdueix una tècnica de classificació amb informació de suport, inspirada en mètodes d'Intel·ligència Artificial, que és novedosa en el camp de l'Estadística. Aquesta informació de suport es materialitza en forma de regles, la incorporació de les quals al procés de classificació (§**??**) suposa la introducció d'informació *semàntica* al sistema, revertint això en un *biaix* de la classificació. L'estructura primària que les regles indueixen sobre les dades i la possibilitat de tractar aquestes primeres unitats estructurals com a nous individus — deguda a la forma com es representen les classes —, permeten realitzar classificacions amb resultats més precissos, segons els objectius continguts en el biaix, que augmenten la seva utilitat i/o interpretabilitat. Així, en les dues aplicacions reals que es presenten a §**??** i §**??**, es pot veure com la incorporació de regles al procés de classificació millora clarament la interpretabilitat de la classificació. En resum es pot dir que

- Les regles que millors resultats han donat, com s'ha vist a §**??**, són les que expressen un coneixement qualitatiu del fenòmen que s'està estudiant, superant l'actuació d'aquelles que deriven directament de models quantitatius.

- Utilitzar regles que expressin relacions entre variables és més potent que descriure el comportament d'una variable aïllada en cada classe. Les regles que descriuen aquestes relacions aporten al sistema una informació que el procediment de classificació per si sol és incapaç de captar (§**??**).

- La intervenció de variables transformades, o derivades, en les regles dota al sistema de la capacitat d'estudiar l'estructura de les dades sota diferents sistemes de coordenades simultàniament, possibilitat que, fins al moment, no havia ofert cap sistema de clustering i de la utilitat de la qual se'n té una mostra en la classificació de les estrelles. En ella s'utilitza la posició d'aquestes en coordenades cartesianes per les regles, però en coordenades esfèriques per a la matriu de dades (aquestes darreres són les que proporcionen directament els observatoris).

Finalment, es proposa una metodologia d'anàlisi §**??** que facilita enormement la difícil tasca d'explicitació del coneixement de l'expert, i permet prescindir de les restrictives hipòtesis de completitud i homogeneïtat del sistema de regles. El cpítol §**??** mostra com l'ús d'aquesta metodologia condueix a un sistema de regles que dirigeix la posterior classificació de forma satisfactòria.

Tradicionalment, la interpretació definitiva de les classes resultants d'un procés de clustering ha recaigut sobre l'usuari, malgrat les eines que alguns paquets estadístics, com **SPAD**, poden oferir al respecte. Amb la intenció d'ajudar-lo en aquesta tasca s'ha implementat amb èxit:

- un heurístic per a decidir el nivell del tall de l'arbre ascendent jeràrquic molt simple §**??**, però que ha resultat tenir un comportament excel·lent en la majoria d'aplicacions.

- una mesura ràpida de la diferència entre particions §**??** que facilita la comparació entre resultats de diferents sistemes o talls diferents d'un mateix arbre i que pot donar idea de la *qualitat* d'una classificació si es disposa d'informació relativa a les classes que cal detectar. Aquesta mesura representa una eina objectiva per a decidir si dues classificacions s'assemblen molt o poc.

- un sistema de caracterització de les classes §**??** en conjuncions de predicats que contenen exactament un conjuntand positiu. L'ús d'informació condicional imbricada permet detectar de forma eficient els trets més rellevants de les diferents classes, contribuint a una fàcil interpretació de les mateixes i constitueix un punt de partida a la generació automàtica de regles de classificació per un sistema basat en diagnòstic. Als capítols §**??** i §**??** es presenten alguns resultats en aquesta línia.

El conjunt d'aquestes eines suposa una contribució a la validació de les classes, que en l'actualitat es fa considerant el grau d'interpretabilitat (i/o utilitat) d'aquestes sense cap més criteri que el d'un especialista que es miri els resultats del classificador.

Treball futur

Com a treball futur immediat es planteja el disseny d'una interfície interactiva "*user-friendly*" per **KLASS**, l'ampliació del mòdul de producció de resultats per tal que es puguin obtenir resultats en codi ASCII i la incorporació d'un mòdul de definició automàtica de noves mètriques o criteris d'agregació que converteixin **KLASS** en un sistema definitivament obert.

Actualment, **KLASS** permet fer classificacions definint només unes quantes *variables actives*. Es pretén que en el futur es pugui també definir un conjunt de *variables il·lustratives* i que es pugui estudiar la seva projecció sobre una certa classificació.

A més, hi ha en marxa un treball que té per objectiu estudiar la sensibilitat de **KLASS** a la forma com es defineixen les variables (és conegut que algunes variables categòriques es poden considerar numèriques discretes i, recíprocament, sempre es poden categoritzar les variables numèriques) i modificar la distància mixta, si és necessari, per a fer un tractament més precís de les variables qualitatives ordinals.

En una propera fase es pretén: introduir la definició de valors il·legals per les variables, entenent per valor il·legal aquell que no té sentit. Per exemple, en les esponges que no ténen

còrtex, hi hauria valors il·legals a les variables `capas_del_cortex`, `Cortex_fibroso` i d'altres per totes les esponges que no tinguin còrtex. Això repercutiria en el procés de caracterització.

També es planteja la recerca en algunes de les línies que aquesta tesi ha deixat obertes.

Línies obertes

- Aprofondir en la caracterització completa de classificacions. D'una banda, extenent el procediment presentat a variables quantitatives també; de l'altra, recorrent a l'estudi d'interaccions entre tuples de variables de grandària igual o superior a 2 i perfeccionant el pas d'inferència implícit en la generació de regles[1], per un sistema orientat al diagnòstic.

- Estudiar formes de quantificar la significació estadística del valor obtingut per $\delta(\mathcal{P}_1, \mathcal{P}_2)$ en una aplicació concret. *A priori*, sembla que un enfoc no paramètric pot donar resultats en aquesta direcció.

- Estudiar com el condicionament a files o columnes de la Tcpp entre dues classificacions pot donar lloc a noves mesures de la distància entre particions.

- Automatitzar l'actualització del conjunt de regles en la metodologia iterativa proposada: en aquests moments el pas $\mathcal{R}^{\xi}$ a $\mathcal{R}^{\xi+1}$ requereix una forta intervenció de l'expert encara, per bé que se li ha facilitat molt la tasca. La detecció d'incoherències internes a una classificació de forma automàtica permetria generar, a partir d'elles, les rectificacions del conjunt de regles pertinents, explicitant un coneixement que el sistema no havia lograt captar de bon principi.

- Per últim, explorar la possibilitat d'utilitzar la metodologia presentada com un sistema d'aprenentatge supervisat, basant-se, fonamentalment, en la classificació d'un conjunt d'exemples. En aquesta línia, una primera passa ha estat l'ús d'una mostra simulada en les estrelles. Al final de la interacció expert-sistema s'ha aconseguit formalitzar un coneixement de suport relatiu al domini que inicialment no s'havia sabut transmetre al sistema.

[1] Un apunt en aquest sentit és l'associació de probabilitats a les regles generades, les quals es calcularien en funció de la mostra analitzada.

Bibliography

[Aguilar & Gibert 91] Aguilar-Martin, J., Gibert, K., "*Sobre variables lingüísticas difusas, paradigmas parmenidianos y lógicas multivaluadas*", proc. 1^{er} Congreso español sobre tecnologías y lógica fuzzy. Granada: Proyecto Sur de Eds. S. A. L., 1991, pág. 185–192.

[Anderberg 73] Anderberg, M.R., "**Cluster analysis for applications**". New York: Academic Press Inc., 1973.

[Anderson 91] Anderson, J.R., Matessa, M., "*An interative bayessian algorithm for categorization*" in Fisher & Pazzani (eds), **Concept formation: Knowledge and experience in unsupervised learning**. San Mateo, CA: Morgan and Kauffmann, 1991.

[Baim 88] Baim, P. W., "*A method for attribute selection in inductive learning systems*". IEEE Trans. on PAMI 10, n. 6, pàg. 888–896.

[Bahcall & Soneira 80] Bahcall, J. N., Soneira, R. M. Astrophisical Journal Suplement 84, pàg. 57–67.

[Ball & Hall 70] Ball, G. H., Hall, D. J., "*Some implications of interactive graphic computer systems for data analysis and statistics*". Technometrics 12, pàg. 17–31.

[Ball & Hall 65] Ball, G. H., Hall, D. J., "*Isodata, a novel method of data analysis and pattern classification*". Tec. Rep. 5 *RI* Project 5533, Stanford Research Inst, Menlo Park, CA, 1965.

[Bécue 93] Bécue, M., *Un sistema informático para el análisis estadístico de datos textuales*. Tesi doctoral. Dept. EIO, UPC, 1993.

[Béjar & Cortés 92] Béjar, J., Cortés, U., "LINNEO+:*Herramienta para la adquisición de conocimiento y generación de reglas de clasificación en dominios poco estructurados*", proc. *IBERAMIA-92*. México: Noriega Eds., 1992. pág.471–481.

[Belanche & Cortés 91] Belanche, L., Cortés, U., "*The nought attributes in knowledge based systems*", proc. *EUROVAV-91*. pàg. 77–103.

[Benzécri 80] Benzécri, J.P., "**L'analyse des données. Tome 1: La Taxinomie, Tome 2: L'analyse des correspondances**". 1^a ed., 1973. Paris: Dunod, 1980.

[Benzécri 76] Benzécri, J.P, "*Histoire et prehistoire de l'Analyse des données*". Les cahiers de l'Analyse des données 1, n. 1à 4.

[Bezdek 87] Bezdek, J., "*Some non-standard clustering algorithms*" in P. Legendre and L. Legendre (eds), **Developments in Numerical Ecology**. Berlin: Springer-Verlag, 1987.

[Capdevila & Arcas 93] Capdevila Marquès, C., Pons Arcas, A., "*Some aspects of statistical inference on pyramidal trees*", proc. *IFCS*-93. Paris, 1993, pàg. 97–98.

[Carney 89] Carney, B. W., Latham, D. W., Laird, J. B., "*A survey of proper-motion stars. VIII. on the galaxy's third population*". The astron. journal 97, n. 2, pàg. 423–430.

[Cheeseman 90] Cheeseman, P., "*On finding the most probable model*" in J. Shrager and P. Langley (eds), **Computational Models of Discovery and Theory Formation**. San Mateo, CA: Morgan Kauffmann, 1990.

[Clancey 85] Clancey, W. J. "*Heuristic Classification*". AI 27, pàg. 289-350.

[Clancey 81] Clancey W. J., *The epistemology of a rule-based expert system: A framework for explanation*. Rep. *STAN-CS-81-896*. Dep. Comput. Sci., Stanford Univ., 1981.

[Cuadras 91] Cuadras, C. M., "**Métodos de análisis multivariante**". Barcelona: Eunibar, 1991.

[Dillon & Goldstein 84] Dillon, W. R., Goldstein, M., "**Multivariate analysis. Methods and applications**". USA: John Wiley and Sons, 1984.

[de Rham 80] de Rham, C., "*La classification hierarchique selon la méthode des voisins réciproques*". Cah. An. des Données V, n. 2, pàg. 135–144.

[Diday, Brito & Mfoumou 93] Diday, E., Brito, P., Mfoumou., M., "*Modelling probabilistic data by conceptual pyramidal clustering*", proc. of 4^{th} Int'l Work. on AI&Stats. Florida, 1993, pàg. 213–218.

[Diday 79] Diday, E. *et coll.* "**Optimisation en classification automatique. Tome 1 et Tome 2**" Versailles: INRIA, 1979.

[Diday & Simon 76] Diday, E., Simon, "*Clustering analysis*" in **Communications & Cybernetics 10**. New York: Springer-Verlag, 1976, pàg. 47–92.

[Diday 71] Diday, E., "*La méthode des nuées dynamiques*". Stat. App. 19, n. 2, pàg. 19–34.

[Domingo 90] Domingo, M., *Aplicació de tècniques d'IA (*LINNEO*) a la classificació sistemàtica. O. HADROMERIDA (DEMOSPONGIÆ- PORIFERA)*. Tesi de llicenciatura. Departament d'Ecologia, Univ. Barcelona, 1990.

[Domingos 94] Domingos, P., *Design and evaluation of the RISE 1.0 learning system*. Tech. report 94-34. Dept. ICS, UIC, Irvine, CA, 1994.

[Everitt 81] Everitt, B., "**Cluster analysis**". London: Heinemann Ed. Books Ltd, 1981.

[Fisher 93] Fisher, D., "*Machine learning. Including a discussion on neural networks*", 4^{th} Int'l Work. on AI&Stats. Florida, 1993.

[Fisher, Pazzani & Langley 91] Fisher, D., Pazzani, M., langley (eds), "**Concept formation: Knowledge and experience in unsupervised learning**". San Mateo, CA: Morgan Kauffmann, 1991.

[Fisher 87] Fisher, D., "*Knowledge acquisition via incremental conceptual clustering*", ML 2, pàg. 139–172.

[Forgy 62] Forgy, E. W., "*Cluster analysis of multivariate data: Efficiency versus interpretability of classifications*". Biometrics 21, pàg. 768–769.

[Gavaldà 94] Gavaldà, R., *Kolmogorov randomness and its applications to structural complexity theory*. Tesi doctoral. Dep. LSI, UPC, 1992.

[Gibert & Cortés 94] Gibert, K., Cortés U., "*Combining a knowledge-based system and a clustering method for a construction of models in ill-structured domains*" in P. Cheeseman and R. W. Oldford (eds), **Selecting models from Data: Artificial Intelligence and Statistics IV**, Lect. Not. in Stats. 89. New York: Springer-Verlag, 1994, pàg. 351–360.

[Gibert 93c] Gibert, K., Cortés, U., "*On the uses of the expert knowledge for automatic biaising of a clustering method*", proc. *ITI'93*. Pula: University Computing Center of Croatia, 1993, pàg. 219–224.

[Gibert & Cortés 93a] Gibert, K., Cortés, U., "*Combining a knowledge based system with a clustering method for an inductive construction of models*", proc. 4^th Int'l Work. on AI&Stats. Florida, 1993, pàg. 243–248.

[Gibert & Cortés 92] Gibert, K., Cortés, U., "KLASS: *Una herramienta estadística para la creación de prototipos en dominios poco estructurados*", proc. *IBERAMIA-92*. México: Noriega Eds., 1992, pàg. 483–497.

[Gibert 91] Gibert, K., KLASS*: Estudi d'un sistema d'ajuda al tractament estad. de grans bases de dades.* Tesi de llicenciatura. Dep. LSI, UPC, 1991.

[Gilmore & Reid 83] Gilmore, G., Reid, N. Monthly Notices Royal Ast. Soc. 202, pàg. 1025.

[Goldberg 89] Goldberg, E., **Genetic algorithms in search, optimization and machine learning**. Addisson Wesley, 1989.

[Gordon 87] Gordon, A. D., "*A review on hierarchical classification*". J. R. Statist. Soc. A. 150, part 2, pàg. 119–137.

[Gowda & Diday 92] Gowda, K. C., Diday, E. "*Symbolic clustering using a new similarity measure*". IEEE Trans. on systems, man., and cib. 22, n. 2, pàg. 368–378.

[Gower 71] Gower,J. C., "*A general coefficient for similarity and some of its properties*". Biometrics 27, pàg. 857–872.

[Hartigan 75] Hartigan, J. A., "**Clustering algorithms**". New York: Wiley & Sons, 1975.

[Haye 84] Hayes-Roth, F., "*The knowledge-based expert system: A tutorial*". Computer 17, n.9.

[Hernández Pajares & Floris 94] Hernández Pajares, M., Floris, J., "*Classification of the Hipparcos input catalogue using the Kohonen network*". Mon. Not. R. Astron. Soc. 268, pàg. 444–450.

[Henández Pajares, Monte, Comellas & Floris 93] Hernández-Pajares, M., Monte, E., Comellas, F., Floris, J., "*Application of pattern recognition algorithms to the classification of stars*". Astronomy and Astrophysics,

[Hernández Pajares, Cubarsi & Monte 92] Hernàndez Pajares, M., Cubarsi, R., Monte, E., "*Study of the self organizing maps of an astronomical catalogue using the stellar velocity moments*". Neural network world 5, pàg. 463–474.

[Holland 92] Holland. Scientific american 267, n. 1, pàg. 44.

[Jardine & Sibson 71] Jardine, N., Sibson, R. "**Mathematical Taxonomy**". New York, London: Wiley & Sons, 1971.

[Jardine & Sibson 68] Jardine, N., Sibson, R. "*The construction of hierarchic and non-hierarchic classifications* ". Comp. J. 11, pàg. 117–184.

[Jain & Dubes 88] Jain, A. K., Dubes, R. C., "**Algorithms for cluster analysis**". Englewood Cliffs, NJ: Prentice Hall, 1988.

[Johnson 67] Johnson, S. C., "*Hierarchical clustering schemes*". Psychometrika 32, pàg. 241–254.

[Kohonen 82] Kohonen, T., "*Self-Organized Formation of Topologically Correct Feature Maps*". Biological Cybernetics 43, pàg. 59–69.

[Lebart *et al.* 85] Lebart, L., Morineau, A., Fenelon, J. P., "**Tratamiento estadístico de datos**". Barcelona: Marcombo, 1985.

[Lebowitz 87] Lebowitz, M., "*Experiments with incremental concept formation*" UNIMEM. ML 2, pàg. 103–138.

[Lerman & Peter 86] Lerman, I. C., Peter, Ph., "*Organization et consultation d'une banque de "petites annonces" a partir d'une méthode de classification hierarchique en parallele*", proc. Data analysis and informatics, IV. North Holland: Elsevier Science Publishers BV, 1986, pàg. 121–135.

[López de Màntaras 91] López de Mántaras, R., "*A distance based attribute selection measure for decision tree induction*" Machine Learning 6, pàg. 81–92.

[López de Mántaras 91b] López de Mántaras, R., "*trends in knowledge engineering*",, proc *TECCOMP-91*. México, 1991, pàg. 20-21.

[López de Mántaras & Crespo 90] López de Mántaras, R., Crespo, J. J., "*El problema de la seleción de atributos en el aprendizaje inductivo: nueva propuesta y estudio experimental*", proc. *IBERAMIA-90*. México: Limusa-Noriega, 1990, pág. 259–271.

[Martín, Sangüesa & Cortés 91] Martín, M. , Sangüesa, R., Cortés, U., "*Knowledge acquisition combining analytical and empirical techniques*". ML91, pàg. 657–661.

[Martín 90] Martín, M., **LINNEO** *eina per a l'ajut en la construcció de bases de coneixement en dominis poc estructurats*. Tesi de llicenciatura. Departament d'LSI, UPC, 1991.

[MacQueen 67] MacQueen, J., "*Some Methods for Classification and Analysis of Multivariate Observations*", proc. 5^{th} Berkeley Symp. Berkeley, 1965, pàg. 281–297.

[McCutcheon 87] Mccutcheon, A., "**latent class analysis**", quantitative applications in the social sciences 64. CA: SAGE public, 1987.

[Michalski & Stepp 83] Michalski, R. S., Stepp, R. E., "*Automated construction of classifications: Conceptual clustering versus numerical taxonomy*". IEEE Trans. on PAMI-5, pàg. 396–410.

[Michalski, Davis, Bisht & Sincler 82] Michalski, R. S., Davis, J. H., Bisht, V. S., Sincler, J. B., "*PLANT/ds: An expert consulting system for the diagnosis of soybean diseases*", proc. of Eur. Conf. of AI. Orsay (France), 1982.

[Michalski & Chilanski 80] Michalski, Chilanski, "*Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean desease diagnosis*". Int. J. Policy Anal. Inform. Syst. 4, n. 2, pàg. 125–161.

[Milligan & Cooper] Milligan, G.W., Cooper, M.C., "*An examination of procedures for determining the number of clusters in a data set*". Psychometrika 50, pàg. 159–179.

[Musen 89] Musen, M. A., "*Automated support for building and extending expert systems*". ML 4, n. 3, pàg. 347–376.

[Núñez *et al.* 91] Núñez, G., Alvarado, M., Cortés, U., Belanche, Ll., "*About the attribute relevance's nature*", proc. *TECCOMP*. México, 1991, pàg. 21–41.

[Núñez 90] Núñez, G., *Caracterización no monótona de la inferencia inductiva y su aplicación al aprendizaje basado en similitudes (SBL)*. Tesi doctoral, Dept. LSI, UPC, Barcelona, 1990.

[Peña 91] Peña Sánchez de Rivera, D., "**Estadística. Modelos y métodos. Tomo 1. Fundamentos**". Madrid: Alianza Universidad, 1991.

[Piera 87] Piera i Carreté, N., *Connectius de lògiques no estàndard com a operadors d'agregació en classificació multivariable i reconeixement de formes*. Tesi doctoral. UPC-FIB, Barcelona, 1987.

[Quevedo, Gibert & Aguilar 93] Quevedo, J., Gibert, K., Aguilar-Martin, J., "*Fuzzy semantics in expert process control*" in B. Bouchon-Meunier, L. Valverde and R.R. Yager (eds), **IPMU'92 – Advanced Methods in Artificial Intelligence**, Lecture Notes in Computer Science 682. Berlin: Springer-Verlag, 1993.

[Quinlan 89] Quinlan, J. R., "*Induction of desition trees*" in **ML 1**, n.1. Boston: Kluwer Academic Publishers, MA(89), pp. 81–106.

[Quinlan 84] Quinlan, J. R., "*Learning efficient classification procedures and their application to chess and games*", Michalski, R.S., Carbonell, J.G. and Mitchel, T.M. (eds) in **ML: An AI Approach**. PA, Calif.: Tioga, 1984, pàg. 463–482.

[Quinlan 83] Quinlan, J. R., "*Learning Efficient Classification Procedures*" in **ML: an AI perspective**. Palo Alto, Calif.: Tioga, 1983.

[Riaño 94] Riaño, D., *Automatic knowledge generation from data in classification domains*. Tesi de llicenciatura. Dept. LSI, UPC, 1994.

[Rizzi 87] Rizzi, A., "*Cluster analysis: hierarchical-type methods*" in **Methods for Multidimensional Data Analysis**, European courses in advanced statistics. Napoli: Universita di Napoli, 1987.

[Robin & Crézé 86] Robin, A., Crézé, M., "*Stellar populations in the Milky way: a synthetic model*". Astronomy and Astrophisics 157, pàg 71–90.

[Rojo 93] Rojo Rodes, E., *Aplicación del software* **LINNEO** *a la classificación de trastornos mentales: Mental-IA*. Tesi doctoral. Facultad de Medicina, U.B., Barcelona, 1993.

[Romesburg 84] Romesburg, "**Cluster analysis for researchers**". Belmont CA: Lifetime Learning Publications, pàg. 463–482.

[Rotnitzky 94] Rotnitzky, A., "*Análisis con datos faltantes* ". Actes sem. Barcelona, 1994.

[Roux 85] Roux, M., "**Algorithmes de classification**". Paris: Masson, 1985.

[Roux 85b] Roux, M., "*Representation d'une distance par un arbre aux arêtes additives*", proc. 4èmes Journées Internationales Analyse de données et Informatique. Versailles: Centre de Rocquencourt, 1985, pàg. 3–15. Tome 1.

[Schlimer & Fisher 86] Schlimer, J. C., Fisher, D., "*A case study of incremental concept induction*", proc. 5^{th} Nat. conf. on AI. Morgan Kauffmann, 1986, pàg. 496–501.

[Sierra 89] Sierra, C., **MILORD II**: *Arquitectura multi-nivell per a sistemes experts en classificació*. Tesi doctoral, Dep. LSI, UPC. Barcelona, 1989.

[Sneath & Sokal 73] Sneath, P.H.A., Sokal, R.S., "**Numerical Taxonomy**". San Francisco: W.H. Freeman & Co., 1973.

[Shortlife 76] Shortlife, E. H., **MYCIN**:*A rule-based computer program for advising physicians regarding antimicrobial therapy selection*. Tesi doct. Standford, Univ. USA, 1976.

[Sokal & Sneath 63] Sokal, R. R., Sneath, P. H. A., "**Principles of numerical taxonomy**". San Francisco: W. H. Freeman & Co., 1963.

[Sokal & Michener 58] Sokal, R. R., Michener, C. D., "*A statistical method for evaluating systematic relationships*". Univ. Kansas Sci. Bull. 38, pàg. 1409–1438. USA.

[Spad 87] Lebart, L., Morineau, A., Lambert, *Système portable pour l'analyse des donées*. Sèvres: *CISIA*, 1987.

[Spss 90] Norusis, M.j., *SPSS base system user's guide*. Chicago: *SPSS* Inc., 1990.

[Steele 90] Steele, G., **Common LISP: The Language**. Second Ed. Digital Press, 1990.

[Torra 94] Torra i Raventós, V., *Contribució a l'estudi de funcions de síntesi per a la Intel·ligència Artificial*. Tesi doctoral. Dept. LSI, UPC, 1994.

[Thorndike 53] Thorndike, R., L., "*Who belongs in a family?*". Psycometrika 18, pàg. 267–273.

[Utgof 86] Utgof, P., "*Shift of bias for inductive concept learning*" in Michalsky, R. S., Carbonell, J. G., Mitchel, T. M. (eds), **ML: An AI approach**. PA, CA: Tioga, 1984.

[van de Velde 89] Van de Velde, W., "*Incremental Learning of optimal decision trees*", proc. of the 4^{th} european working session on learning. Morgan Kaufmann, 1989, pàg. 211–225.

[Volle 85] Volle, M., "**Analyse des données**". Paris: Economica, 1985.

[Ward 63] Ward, J. H., "Hierarchical grouping to optimize an objective function". J. Am. Statist. Ass. 58, pàg. 236–244.

[Williams & Lambert 59] Williams, W. T., Lambert, J. M., "*Multivariate methods in plant ecology*". J. Ecol. 47, pàg. 83–101.

[Wolfe 70] Wolfe, J., H., "*Pattern clustering by multivariate mixture analysis.*". Multiv. Behav. Res. 5, pàg. 329–350.

Appendix A

Resultats addicionals sobre les esponges

A.1 Objectes

Les espècies que s'han classificat apareixen codificades a la taula **??**

A.2 Detall d'algunes classificacions

A.2.1 Classificació tradicional

F.*Tethyidæ* : T.au, T.cit, A.aa

F.*Polymastiidæ* : P.ube, P.spi, P.rad, P.inv, P.poly, P.lit, P.inf, P.cor, P. agglu, P.tenax, P.rob, P. mam, P. infr, P.hir, P.grim, P.coni, P.tis, P.ecto, P.fusca, P.mar, S.capi, S. antarc, W. bursa, W. ver, T. semis, T. papil, Q. inter, Q.bre, Q.ric, R.ovi, S.spi, T.sarsi, T.hemis, T.step, T.jou, P.sol

F.*Suberitidæ* : S.gib, S.c.v.inc, S.c.v ty, S.c.v.ra, S.ficus, S. dom, S. cami, P. sul, P. hya, Pr.rug, Pr.longi, Pr.epi, L.rugosus, L.ferrerhernandezI, L.ecty, T.fugax, R.pyri, R.bise, R.uni, R.elo

F.*Spirastrellidæ* : S.minax, S.cunc,

F.*Clionidæ* : A.mil, C.laby, C.schmidti, C.car, C.ce, C.vir

F.*Timeidæ* : T.uni, T.stel, T.mixta, T.hal, T.chon, D.biste, D.ornata

F.*Stylocordylæ* : O.pel, S.bor

A.2.2 Resultat de Linneo

C1 : T.chon, A.mil, S.cunc, S.minax, C.laby, C.ca, C.schmidti, T.uni, T.stel, T.mixta, T.hal, D.biste, D.ornata

C2 : O.pel, S.bor

C3 : S.capi, P.sol, P.spi, P.rad, P.lit, P.cor, P.ube, P.inv, P.infl, P.agglu, P.tenax, P.rob, P.poly, P.mam, P.infr, P.hir, P.grim, P.coni

C4 : S.antarc, P.tis, P.ecto, P.fusca, P.mar

Espècie	codi	Espècie	codi
Aaptos_aaptos	A.aa	Alectona_millari	A.mil
Cliona_carteri	C.car	Cliona_celata	C.ce
Cliona_labyrinthica	C.laby	Cliona_schmidti	C.schmidti
Cliona_viridis	C.vir	Diplastrella_bistellata	D.bistel
Diplastrella_ornata	D.ornata	Laxosuberites_ectyonimus	L.ecty
Laxosuberites_ferrerhernandezi	L.fer	Laxosuberites_rugosus	L.rug
Oxycordyla_pellita	O.pel	Polymastia_agglutinaris	P.agglu
Polymastia_conigera	P.coni	Polymastia_corticata	P.cor
Polymastia_ectofibrosa	P.ecto	Polymastia_fusca	P.fusca
polymastia_grimaldi	P.grim	Polymastia_hirsuta	P.hir
Polymastia_inflata	P.infl	Polymastia_infrapilosa	P.infr
Polymastia_invaginata	P.inv	Polymastia_littoralis	P.lit
Polymastia_mammillaris	P.mam	Polymastia_martæ	P.mar
Polymastia_polytylota	P.poly	Polymastia_radiosa	P.rad
Polymastia_robusta	P.rob	Polymastia_spinula	P.spi
Polymastia_tenax	P.tenax	Polymastia_tissieri	P.tis
Polymastia_uberrima	P.ube	Prosuberites_epiphytum	Pr.epi
Prosuberites_longispina	Pr.longi	Prosuberites_rugosus	Pr.ru
Proteleia_sollasi	P.sol	Pseudosuberites_hyalinus	Ps.hya
Pseudosuberites_sulfureus	Ps.sul	Quasilina_brevis	Q.bre
Quasilina_intermedia	Q.inter	Quasilina_richardii	Q.ric
Rhizaxinella_biseta	R.bise	Rhizaxinella_elongata	R.elo
Rhizaxinella_pyriphera	R.pyri	Rhizaxinella_uniseta	R.uni
Ridleya_oviformis	R.ovi	Sphærotylus_antarcticus	S.antarc
Sphærotylus_capitatus	S.cap	Spinularia_spinularia	S.spin
Spirastrella_cunctatrix	S.cunc	Spirastrella_minax	S.minax
Stylocordyla_borealis	S.bor	Suberites_caminatus	S.cami
Suberites_carnosus_v.incrustans	S.c.v.inc	Suberites_carnosus_v.ramosus	S.c.v.ra
Suberites_carnosus_v.typicus	S.c.v.ty	Suberites_domuncula	S.domu
Suberites_ficus	S.ficus	Suberites_gibbosiceps	S.gib
Tentorium_papillatus	T.papil	Tentorium_semisuberites	T.semis
Terpios_fugax	T.fugax	Tethya_aurantium	T.au
Tethya_citrina	T.cit	Timea_chondrilloides	T.chon
Timea_hallezi	T.hal	Timea_mixta	T.mixta
Timea_stellata	T.stel	Timea_unistellata	T.uni
Trachyteleia_stephensi	T.step	trichostema_hemisphæricum	T.hemi
Trichostema_sarsi	T.sarsi	Tylexocladus_joubini	T.jou
Weberella_bursa	W.bursa	Weberella_verrucosa	W.ver

Table A.1: Índex d'espècies per ordre alfabètic.

C5 : R.pyri, Ps.sul, P.hya, Pr.ru, Pr.epi, Pr.longi, L.ru, L.fer, L.ectyonimus, T.fugax, S.gib, S.c.v.ty, S.c.v.inc, S.c.v.ra, R.bise, R.uni, R.elo

C6 : S.ficus, S.dom

C7 : T.au, T.cit

C8 : C.cel, C.vir

C9 : T.jou, T.semis, T.papil, S.cami, A.aa, S.spi, T.step,

c10 : T.hemis, T.sarsi

c11 : W.bursa, W.ver, R.ovi, Q.inter, Q.bre, Q.ric

A.2.3 Classificació proposada per Domingo (1990)

F.*Semipolymastiidæ*: T.semis, T.papil, S.cami, T.step, T.jou, S.spi,

F.*Spirastrellidæ* : S.minax, S.cunc, D.biste, D.ornata, C.schmidti, C.car, C.laby, C.cel, C.vir, A.mil

F.*Suberitidæ*: S.ficus, S.dom, S.c.v.ty, S.c.v.inc, S.c.v.ra, S.gib, T.fugax, L. ru, L.fer, L.ecty, Pr.ru, Pr.longi, Pr.epi, Ps.sul, P.hya, R.pyri, R.bise, R.uni, R.elo

F.*Stylocordylæ*: O.pel, S.bor

F.*Polymastiidæ* : P.spi, P.rad, P.lit, P.inv, P.cor, P.u be, P.infl, P.agglu, P.tenax, P.rob, P.poly, P.mam, P.infr, P.hir, P.grim, P.coni, P.tis, P.ecto, P.fusca, P.mar, S.capi, S.antarc, P.sol

F.*Quasilinidæ* : Q.inter, Q.bre, Q.ric, W.ver, W.bursa, R.ovi

F.*Tethyidæ* : T.au, T.cit

F.*Timeidæ* : T.stel, T.mixta, T.hal, T.uni, T.chon

F.*Trichostematidæ* : T.hemis, T.sarsi

F.*Aaptidæ* : A.aa

A.3 Llista de propietats

Utilitzant criteris morfològics [**?**], es poden diferenciar tres tipus de caràcters: estructurals, funcionals o filogenètics — aquests darrers fan referència a semblances evolutives entre espècies —. D'acord amb això, la classificació es podria enfocar des de diferents punts de vista. Actualment es tendeix a l'enfoc filogenètic, però si es vol tractar un grup divers d'esponges en que hi intervinguin espècies de diferents gèneres, cal basar-se en criteris morfològics, els quals integren la major part de la informació, donada la mancança de dades en gran nombre de caràcters funcionals.

D'entre tots els caràcters utilitzats per a constituir la classificació tradicional de les esponges, o considerats per un expert en la identificació d'una espècie, els tractats més exhaustivament són els que es relacionen amb l'esquelet. A [**?**] s'hi pot trobar una discusió detallada sobre la selecció de les variables.

Les variables que finalment ens proporcionaren els experts per a treballar són les següents:

Referents al còrtex :

Cortex : Diu si l'esponja té còrtex o no. Contínua (binària). Rang de variació: [0,1]

Capas del cortex : Qualitativa. Modalitats:

- 1 capa
- 2 capas
- 3 capas
- Sin cortex[1]

Capa interna del cortex : De les capes del còrtex, diferenciables a nivell microscòpic, es dóna la caracterització de la més interna. Qualitativa. Modalitats:

- Bandas de espiculas cruzadas
- Microscleras
- Perpendicular
- Tangencial
- Tangencial y Perpendicular
- Sin capa interna del cortex

- Propietats específiques del còrtex que hom troba en una o poques espècies.

  Cortex fibroso : Qualitativa. Modalitats:

  - No
  - Sí
  - Sin cortex

  Cortex solo de espiculas tangenciales : Qualitativa (binària). Modalitats:

  - No
  - Sí
  - Sin còrtex

  Cuerpos extraños en el cortex : Qualitativa. Modalitats:

  - No
  - Sí
  - Sin còrtex

  Haces de espiculas principales en pompon en el cortex : Contínua. Variació: [0,1]

Grosor del cortex : Malgrat ser conceptualment contínua, la informació està disponible en forma d'intervals. Resolent, doncs, un cert problema de representació, s'ha normalitzat la mesura, identificant cada interval amb el seu punt mig (veure [**?**] per una justificació al respecte). Contínua. Rang de variació: [0,4]

Estudi de l'esquelet :

- Per cada **tipus d'espícula principal** s'ha definit una variable binària. Es pot així indicar que una única espècie pot presentar espícules de més d'un tipus. Les variables estan totes definides de la mateixa forma i són:

  Tipo espicula principal diactina tuberculada : Contínua. Rang de variació: [0,1]

  Tipo espicula principal estilo .

  Tipo espicula principal estilos 2 tamaños .

  Tipo espicula principal estilo tilostilo .

  Tipo espicula principal estrongiloxa .

  Tipo espicula principal oxas .

  Tipo espicula principal tilostilo .

- Quant a la **forma de l'espícula**, s'ha fet referència a les modificacions més aparents, i a la forma de la base:

  Espicula principal estilo : Qualitativa. Modalitata:

[1] De fet aquest serien valors inaplicables. No té sentit estudiar quantes capes té el còrtex d'una esponja sense còrtex. Una situació semblant es trobarà repetida en altres variables.

  - Fusiforme
  - Normal
  - Politilota
  - Sin espicula principal estilo

Espicula principal tilostilo : Qualitativa. Modalitats:
  - Flexuosa
  - Fusiforme
  - Normal
  - Politilota
  - Sin espicula principal tilostilo

Forma base tilostilo principal : Qualitativa. Modalitats:
  - Pomo
  - Subesferica alargada ovoide
  - Sin tilostilos
  - Trilobulada

- Sobre les **megascleres**

Numero de tipos de megascleras : Qualitativa. Modalitats:
  - 1 tipo
  - 2 tipos
  - 3 tipos

Disposicion megascleras ectosomicas en el ectosoma : Qualitativa. Modalitats:
  - Empalizada
  - Ramilletes
  - Tangenciales
  - Sin megascleras ectosomicas

Tipo megasclera ectosomica : Qualitativa. Modalitats:
  - Estilo
  - Estilo tilostilo
  - Oxas estrongiloxas
  - Tilostilo
  - Sin megasclera ectosomica

Forma megasclera ectosomica : Qualitativa. Modalitats:
  - Curvada
  - Recta fusiforme
  - Recta no fusiforme
  - Sin megasclera ectosomica

Forma base tilostilo ectosomico : Qualitativa. Modalitats:
  - Subesferica alargada ovoide
  - Trilobulada con vesicula axial
  - Sin tilostilo ectosomico

Tipo megasclera intermediaria : Qualitativa. Modalitats:
  - Estilo
  - Estilo tilostilo
  - Tilostilo
  - Sin megasclera intermediaria

Forma megasclera intermediaria : Qualitativa. Modalitats:
  - Fusiforme
  - Normal

- Politilota
- Sin megasclera intermediaria

Longitud megascleras : Es recull la longitud màxima que pot presentar una espècie, assumint que en estat adult la variabilitat en cada espècie és, en general, reduïda. Contínua. Rang de variació: [0,3]

- Sobre les **microscleres**, variables pels diferents tipus de microscleres:

Microscleras : Contínua. Rang de variació: [0,1]

Tipo Microsclera : Qualitativa. Modalitats:
- Aster
- Microestilos y microxas
- Microestrongilos centrotilotes
- Tricodragma
- Sin microscleras

Aster : Contínua. Rang de variació: [0,1]

Tipo de aster : Qualitativa. Modalitats:
- Anfiaster
- Diplaster
- Esferaster
- Espiraster
- Quiaster
- Quiaster y esferaster
- Sin aster

Tipo de esferaster : Qualitativa. Modalitats:
- Oxiesferaster
- Oxiesferaster y tiloesferaster
- Sin esferaster

Diametro esferaster : Qualitativa. Modalitats:
- [10 12]
- [20 25]
- [27 44]
- [40 110]
- [5 6] y [15 35]
- Sin esferaster

Tipo de diplaster : Qualitativa. Modalitats:
- Con actinas complejas
- Con actinas simples
- Sin diplasteres

- Altres variables referents a l'esquelet:

Tilostilos adicionales coanosoma : Qualitativa. Modalitats:
- Ectosomicos dispersos
- Ectosomicos en ramilletes
- Intermediarios
- Intermediarios y ectosomicos
- Sin tilostilos adicionales

Tipo de exostilo : Qualitativa. Modalitats:
- Anatrienas
- Cladotilostilos
- Esferotilos

- Tilostilos espinosos en la punta
- Tilostilo liso
- Sin exostilos

Referents a les característiques externes :

Forma final : Ha presentat problemes d'expressió, ocasionats per l'ambigüitat semàntica d'algunes categories. Inconvenient addicional és que algunes espècies varien de forma a llarg de la seva vida. Es considera la forma adulta final. Qualitativa. Modalitats:

- De cono invertido
- De revestimiento
- Globulosa
- Hemisferica
- Incrustante
- Masivo irregular
- Mazas pedunculadas 1/pedunculo
- Mazas pedunculadas varias/pedunculo
- Ramificada

Papilas : Contínua. Rang de variació: [0,1]

Numero de papilas : Contínua. Rang de variació: [0,4]

Color : Una mateixa espècie pot presentar diversos colors. Només s'han indicat els colors clarament típics d'alguna espècie. Qualitativa. Modalitats:

- Amarillo palido
- Azul o anaranjado intensos
- Otros

Superficie : Recull informació sobre l'aspecte i consistència al tacte de la superfície de l'esponja. Qualitativa. Modalitats:

- Aterciopelada
- Con areas poligonales abultadas
- Franja basal de espiculas en fleco
- Hispidez mayor hacia la base
- Lisa
- Rugosa
- Solo papilas lisas
- Totalmente hispida

Disposicion espicular esqueleto : Qualitativa. Modalitats:

- Axial
- Confusa
- Hymedesmoide
- Plumosa
- Radial
- Radial en periferia
- Reducido

Referents a característiques ecològiques :

Aloja cangrejo ermitaño : Contínua (binària). Rang de variació: [0,1]

Perforante : Contínua (binària). Rang de variació: [0,1]

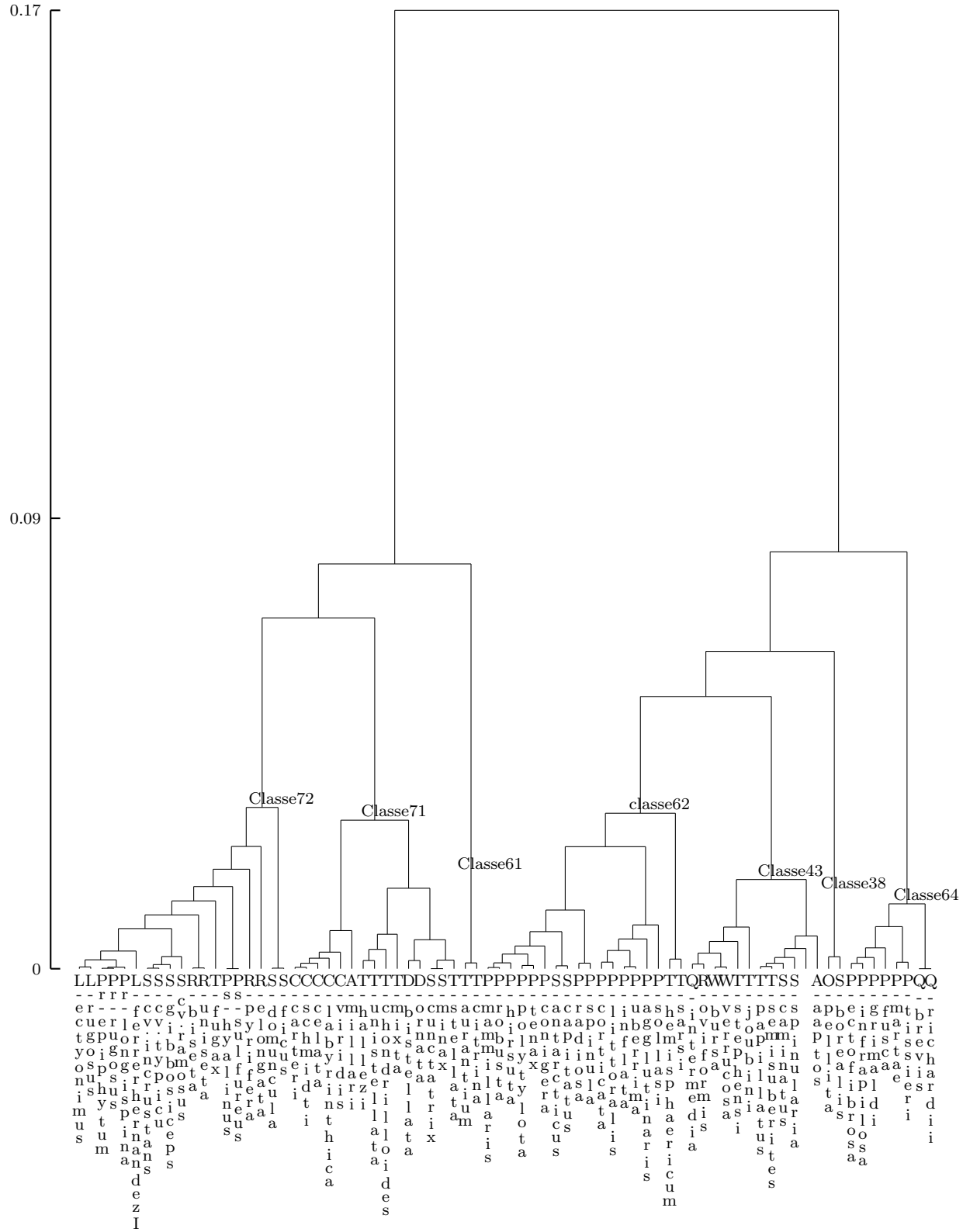
Pseudoraíces : Contínua (binària). Rang de variació: [0,1]

Sustrato : Qualitativa. Modalitats:

- Ambos
- Blando
- Duro

A.4 Classificació general amb Klass

Figure A.1: Arbre ascendent jeràrquic de la classificació sense regles.



A.5 Variables descriptores (7 classes)

	Valor típic						
Variable \ Classe	c72	c71	c62	c61	c43	c38	c64
Cortex	0		1	1	1	1	1
Capas del cortex	Sin còrtex					1	
Capa interna del cortex	Sin	Sin		Microscleras		Sin	
Cortex fibroso	Sin cortex		No			No	
Cortex solo de espiculas tangenciales	Sin		No	No	No		No
Con cuerpos extraños en el cortex	Sin còrtex			No	No	No	No
Grosor del cortex	0			4		1	
Haces de espiculas principales en pompon en el cortex	0	0		0	No	0	0
Tilostilos adicionales coanosoma	Sin	Sin		Sin		Sin	
Tipo espicula principal diactina tuberculada	0		0	0	0	0	0
Tipo espicula principal estilo				0	0		0
Tipo espicula principal estilos 2 tamaños		0	0	0	0	0	0
Tipo espicula principal estilo tilostilo	0	0	0	0	0	0	1
Tipo espicula principal estrongiloxa	0	0		1			0
Tipo espicula principal oxas			0	0	0		0
Tipo espicula principal tilostilo	1			0		0	0
Espicula principal estilo				Sin	Sin		
Espicula principal tilostilo				Sin		Sin	
Tipo de exostilo	Sin	Sin		Sin		Sin	
Numero de tipos de megascleras				1		3	

	Valor típic						
Variable \ Classe	c72	c71	c62	c61	c43	c38	c64
Disposicion megascleras ectosomicas en el ectosoma		Sin		Sin			
Forma megasclera ectosomica		Sin		Sin			
Tipo megasclera ectosomica		Sin	Tilostilo	Sin		Oxas estrongiloxas	
Forma base tilostilo ectosomico		Sin	Subesferica alargada ovoide	Sin	Subesferica alargada ovoide	Sin	
Forma base tilostilo principal				Sin		Sin	
Forma megasclera intermediaria	Sin	Sin		Sin		Sin	
Tipo megasclera intermediaria	Sin	Sin		Sin		Sin	
Longitud megascleras				3		3	
Microscleras			0	1		1	0
Tipo Microsclera			Sin	Aster		Microestilos y microxas	Sin
Aster	0		0	1	0	0	0
Diametro esferaster	Sin		Sin	40-110	Sin	Sin	Sin
Tipo de aster	Sin		Sin		Sin	Sin	Sin
Tipo de diplaster	Sin		Sin	Sin	Sin	Sin	Sin
Tipo de esferaster	Sin		Sin	Oxiesferaster	Sin	Sin	Sin
Numero de papilas	0			0		0	
Papilas	0		1	0		0	1
Perforante	0		0	0	0	0	0
Aloja cangrejo ermitaño		0	0	0	0	0	0
Color				Otros			
Pseudoraíces		0	0	0			0

	Valor típic						
Variable \ Classe	c72	c71	c62	c61	c43	c38	c64
Forma final				Globulosa		Mazas pedunculadas/ 1-pedunculo	
Sustrato		Duro		Duro	Duro		Duro
Superficie				Con areas poligonales			
abultadas		Lisa					
Disposicion especular esqueleto				Radial		Radial	

A.6 Classificació amb regles

Apareix primerament el conjunt $\mathcal{R}$ de regles proporcionades pels experts, seguit de la partició que, sobre la mostra estudiada, indueixen aquestes regles. Finalment els informes resultants d'efectuar la classificació corresponent[2].

A.6.1 El conjunt de regles

$$r_1 : (\ (tipo\ espicula\ principal\ tilostilo = 1) \wedge \\ \wedge (tipo\ microsclera = aster) \wedge (tipo\ de\ aster = espiraster) \wedge \\ \wedge (forma\ final \in \{masivo\ irregular, incrustante\}) \longrightarrow \mathcal{C}_1)$$

$$r_2 : (\ (tipo\ espicula\ principal\ tilostilo = 1) \wedge \\ \wedge (tipo\ microsclera = aster) \wedge (forma\ final = incrustante) \longrightarrow \mathcal{C}_2)$$

$$r_3 : (\ (tipo\ espicula\ principal\ tilostilo = 1) \wedge (papilas = 0) \wedge \\ (T.microsclera \in \\ \{microestrongilos\ centrotilotes, sin\ microscleras\}) \\ \wedge (disposicion\ espicular\ esqueleto \in \{axial, radial\ en\ periferia\}) \longrightarrow \mathcal{C}_3)$$

$$r_4 : (\ (tipo\ espicula\ principal\ estilo = 1) \wedge (papilas = 0) \wedge \\ (T.microsclera \in \\ \{microestrongilos\ centrotilotes, sin\ microscleras\}) \\ \wedge (disposicion\ espicular\ esqueleto = (axial, radial\ en\ periferia)) \longrightarrow \mathcal{C}_3)$$

A.6.2 Partició induïda per les regles

c1: `C.viridis`

c3 : `D.biste`, `D.ornata`, `T.hal`, `T.mixta`, `T.stel`, `T.uni`

c4 : `R.bise`, `R.elo`, `R.uni`, `S.c.v.inc`, `S.c v ra`, `S.c. v ty`, `S.dom`, `S.ficus`, `S.gib`

[2]No s'adjunta l'informe de la classe $\mathcal{C}_1$ que, per contenir un sol element, no aporta cap informació nova.

residual ($\mathcal{C}_0$): P.coni, P.fusca, P.grim, P.hir, P.infra, P.mar, P.mam, P.poly, P.rob, P.tenax, O.pel, S.bor, C. car, C. schmidti, S. cunc, S. minax, A. aa, A.mil, C.cel, C.laby, L.ecty, L.fer, L.ru, P.agglu, P.cor, P.ecto, P.infl, P. inv, P.lit, P.rad, P.spi, P.tis, P.ube, Pr.epi, Pr.longi, Pr.ru, P. sol, P. hya, Ps.sul, Q.bre, Q.inter, Q.ric, R.pyri, R.ovi, S.antarc, S.capi, S. spi, S. cami, T.papil, T.semis, T.fugax, T.chon, T.step, T.hemis, T.sarsi, T.jou, W.bursa, W.ver, T.au, T.cit

A.6.3 Informes

En primer lloc apareixen els informes de les classes induïdes per les regles, de les que se n'omet el dendrograma, que es pot observar en la classificació global.

Informe de classificació n^o 2 ($\tau_{\mathcal{C}_3}$)

Dades: DATA/C3
Nombre d'objectes: 6
Nombre de propietats: 45
Criteri de classificació: WARD
Mètrica utilitzada: MIXTE
Valors de les constants de ponderació: $\alpha = 0.000$ i $\beta = 1.000$

Descripció de les 1 classes ($\bar{\imath}_{\mathcal{C}_3}$)

classeI5: Coordenades promig:

- A.CAPAS-DEL-CORTEX: SIN-CORTEX
- A.CAPA-INTERNA-DEL-CORTEX: SIN-CAPA-INTERNA-DEL-CORTEX
- A.CORTEX: 0.0
- A.CORTEX-FIBROSO: SIN-CORTEX
- A.CORTEX-SOLO-DE-ESPICULAS-TANGENCIALES: SIN-CORTEX
- A.CUERPOS-EXTRAÑOS-EN-EL-CORTEX: SIN-CORTEX
- A.GROSOR-DEL-CORTEX: 0.0
- A.HACES-DE-ESPICULAS-PRINCIPALES-EN-POMPON-EN-EL-CORTEX: 0.0
- A.TILOSTILOS-ADICIONALES-COANOSOMA: SIN-TILOSTILOS-ADICIONALES
- B.NUMERO-DE-TIPOS-DE-MEGASCLERAS: 1-TIPO
- C.TIPO-ESPICULA-PRINCIPAL-DIACTINA-TUBERCULADA: 0.0
- C.TIPO-ESPICULA-PRINCIPAL-ESTILO: 0.0
- C.TIPO-ESPICULA-PRINCIPAL-ESTILOS-2-TAMAÑOS: 0.0
- C.TIPO-ESPICULA-PRINCIPAL-ESTILO-TILOSTILO: 0.0
- C.TIPO-ESPICULA-PRINCIPAL-ESTRONGILOXA: 0.0
- C.TIPO-ESPICULA-PRINCIPAL-OXAS: 0.0
- C.TIPO-ESPICULA-PRINCIPAL-TILOSTILO: 1.0
- D.ESPICULA-PRINCIPAL-ESTILO: SIN-ESPICULA-PRINCIPAL-ESTILO
- D.ESPICULA-PRINCIPAL-TILOSTILO: NORMAL
- D.FORMA-BASE-TILOSTILO-PRINCIPAL: SUBESFERICA-ALARGADA-OVOIDE
- E.DISPOSICION MEGASCLERAS ECTOSOMICAS EN EL ECTOSOMA: SIN MEGASCLERAS ECTOSOMICAS
- E.FORMA-BASE-TILOSTILO-ECTOSOMICO: SIN-TILOSTILO-ECTOSOMICO

- `E.FORMA-MEGASCLERA-ECTOSOMICA:` SIN-MEGASCLERA-ECTOSOMICA
- `E.TIPO-MEGASCLERA-ECTOSOMICA:` SIN-MEGASCLERA-ECTOSOMICA
- `F.TIPO-DE-EXOSTILO:` SIN-EXOSTILOS
- `G.FORMA-MEGASCLERA-INTERMEDIARIA:` SIN-MEGASCLERA-INTERMEDIARIA
- `G.TIPO-MEGASCLERA-INTERMEDIARIA:` SIN-MEGASCLERA-INTERMEDIARIA
- *H.LONGITUD-MEGASCLERAS:* 0.5555555555555555 , desviació dins la classe 0.27217
- `I.MICROSCLERAS:` 1.0
- `I.TIPO-MICROSCLERA:` ASTER
- `J.ASTER:` 1.0
- *J.DIAMETRO-ESFERASTER:*
  - 10-12 = 0.16667
  - 20-25 = 0.16667
  - 5-6-Y-15-35 = 0.16667
  - SIN-ESFERASTER = 0.50000
- *J.TIPO-DE-ASTER:*
  - DIPLASTER = 0.33333
  - ESFERASTER = 0.50000
  - QUIASTER = 0.16667
- *J.TIPO-DE-DIPLASTER:*
  - CON-ACTINAS-COMPLEJAS = 0.16667
  - CON-ACTINAS-SIMPLES = 0.16667
  - SIN-DIPLASTERES = 0.66667
- *J.TIPO-DE-ESFERASTER:*
  - OXIESFERASTER = 0.33333
  - OXIESFERASTER-Y-TILOESFERASTER = 0.16667
  - SIN-ESFERASTER = 0.50000
- `K.FORMA-FINAL:` INCRUSTANTE
- `L.NUMERO-DE-PAPILAS:` 0.0
- `L.PAPILAS:` 0.0
- *M.COLOR:*
  - AMARILLO-PALIDO = 0.16667
  - OTROS = 0.83333
- *N.SUPERFICIE:*
  - LISA = 0.33333
  - RUGOSA = 0.16667
  - TOTALMENTE-HISPIDA = 0.50000
- `O.DISPOSICION-ESPICULAR-ESQUELETO:` HYMEDESMOIDE
- `P.ALOJA-CANGREJO-ERMITAÑO:` 0.0
- `P.PERFORANTE:` 0.0
- `P.PSEUDORAICES:` 0.0
- `P.SUSTRATO:` DURO

Elements de la classe:

6

(Diplastrella bistellata Diplastrella ornata Timea stellata Timea hallezi Timea unistellata Timea mixta)

Nota: Les propietats que apareixen amb lletra tipogràfica són descriptores de la classe.

Informe de classificació n^o 3 ($\tau_{\mathcal{C}_4}$)

Dades: DATA/C4
Nombre d'objectes: 9
Nombre de propietats: 45
Criteri de classificació: WARD
Mètrica utilitzada: MIXTE
Valors de les constants de ponderació: $\alpha = 0.000$ i $\beta = 1.000$

Descripció de les 1 classes ($\bar{\imath}_{\mathcal{C}_4}$)

classeS8: Coordenades promig:

- `A.CAPAS-DEL-CORTEX:` SIN-CORTEX
- `A.CAPA-INTERNA-DEL-CORTEX:` SIN-CAPA-INTERNA-DEL-CORTEX
- `A.CORTEX:` 0.0
- `A.CORTEX-FIBROSO:` SIN-CORTEX
- `A.CORTEX-SOLO-DE-ESPICULAS-TANGENCIALES:` SIN-CORTEX
- `A.CUERPOS-EXTRAÑOS-EN-EL-CORTEX:` SIN-CORTEX
- `A.GROSOR-DEL-CORTEX:` 0.0
- `A.HACES-DE-ESPICULAS-PRINCIPALES-EN-POMPON-EN-EL-CORTEX:` 0.0
- `A.TILOSTILOS-ADICIONALES-COANOSOMA:` SIN-TILOSTILOS-ADICIONALES
- *B.NUMERO-DE-TIPOS-DE-MEGASCLERAS:*
  - 1-TIPO = 0.44444
  - 2-TIPOS = 0.33333
  - 3-TIPOS = 0.22222
- `C.TIPO-ESPICULA-PRINCIPAL-DIACTINA-TUBERCULADA:` 0.0
- *C.TIPO-ESPICULA-PRINCIPAL-ESTILO:* 0.33333, desviació dins la classe 0.50000
- *C.TIPO-ESPICULA-PRINCIPAL-ESTILOS-2-TAMAÑOS:* 0.1111111111111111 , desviació dins la classe 0.33333
- `C.TIPO-ESPICULA-PRINCIPAL-ESTILO-TILOSTILO:` 0.0
- `C.TIPO-ESPICULA-PRINCIPAL-ESTRONGILOXA:` 0.0
- *C.TIPO-ESPICULA-PRINCIPAL-OXAS:* 0.22222 , desviació dins la classe 0.44096
- `C.TIPO-ESPICULA-PRINCIPAL-TILOSTILO:` 1.0
- *D.ESPICULA-PRINCIPAL-ESTILO:*
  - NORMAL = 0.33333
  - SIN-ESPICULA-PRINCIPAL-ESTILO = 0.66667
- *D.ESPICULA-PRINCIPAL-TILOSTILO:*
  - FUSIFORME = 0.33333
  - NORMAL = 0.66667
- *D.FORMA-BASE-TILOSTILO-PRINCIPAL:*
  - SUBESFERICA-ALARGADA-OVOIDE = 0.55556
  - TRILOBULADA = 0.44444
- *E.DISPOSICION-MEGASCLERAS-ECTOSOMICAS-EN-EL-ECTOSOMA:*
  - EMPALIZADA = 0.66667
  - SIN-MEGASCLERAS-ECTOSOMICAS = 0.33333
- *E.FORMA-BASE-TILOSTILO-ECTOSOMICO:*
  - SIN-TILOSTILO-ECTOSOMICO = 0.88889

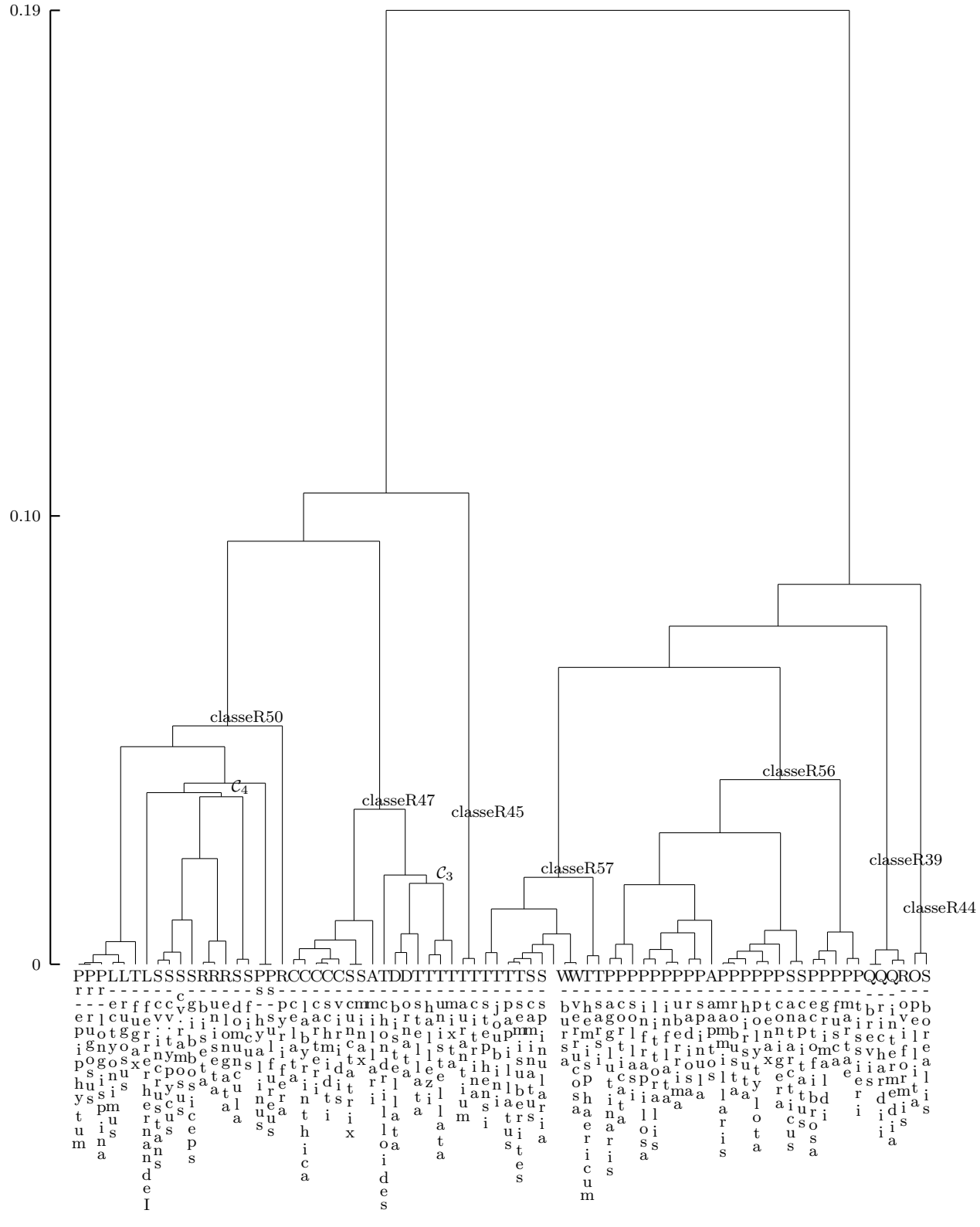
    - SUBESFERICA-ALARGADA-OVOIDE = 0.11111
- *E.FORMA-MEGASCLERA-ECTOSOMICA:*
    - CURVADA = 0.11111
    - SIN-MEGASCLERA-ECTOSOMICA = 0.88889
- *E.TIPO-MEGASCLERA-ECTOSOMICA:*
    - SIN-MEGASCLERA-ECTOSOMICA = 0.88889
    - TILOSTILO = 0.11111
- `F.TIPO-DE-EXOSTILO:` SIN-EXOSTILOS
- `G.FORMA-MEGASCLERA-INTERMEDIARIA:` SIN-MEGASCLERA-INTERMEDIARIA
- `G.TIPO-MEGASCLERA-INTERMEDIARIA:` SIN-MEGASCLERA-INTERMEDIARIA
- *H.LONGITUD-MEGASCLERAS:* 0.66666 , desviació dins la classe 0.28868
- *I.MICROSCLERAS:* 0.22222 , desviació dins la classe 0.44096
- *I.TIPO-MICROSCLERA:*
    - MICROESTRONGILOS-CENTROTILOTES = 0.22222
    - SIN-MICROSCLERAS = 0.77778
- `J.ASTER:` 0.0
- `J.DIAMETRO-ESFERASTER:` SIN-ESFERASTER
- `J.TIPO-DE-ASTER:` SIN-ASTER
- `J.TIPO-DE-DIPLASTER:` SIN-DIPLASTERES
- `J.TIPO-DE-ESFERASTER:` SIN-ESFERASTER
- *K.FORMA-FINAL:*
    - GLOBULOSA = 0.33333
    - INCRUSTANTE = 0.11111
    - MAZAS-PEDUNCULADAS-1/PEDUNCULO = 0.44444
    - RAMIFICADA = 0.11111
- `L.NUMERO-DE-PAPILAS:` 0.0
- `L.PAPILAS:` 0.0
- *M.COLOR:*
    - AMARILLO-PALIDO = 0.55556
    - OTROS = 0.44444
- *N.SUPERFICIE:*
    - ATERCIOPELADA = 0.33333
    - LISA = 0.22222
    - RUGOSA = 0.11111
    - TOTALMENTE-HISPIDA = 0.33333
- *O.DISPOSICION-ESPICULAR-ESQUELETO:*
    - AXIAL = 0.44444
    - RADIAL-EN-PERIFERIA = 0.55556
- *P.ALOJA-CANGREJO-ERMITAÑO:* 0.22222 , desviació dins la classe 0.44096
- `P.PERFORANTE:` 0.0
- *P.PSEUDORAICES:* 0.33333 , desviació dins la classe 0.50000
- *P.SUSTRATO:*
    - AMBOS = 0.22222
    - BLANDO = 0.11111
    - DURO = 0.66667

Elements de la classe:

9 (Suberites carnosus v. incrustans Suberites carnosus v. typicus Suberites carnosus v. ramosus Suberites gibbosiceps Rhizaxinella biseta Rhizaxinella uniseta Rhizaxinella elongata Suberites domuncula Suberites ficus)

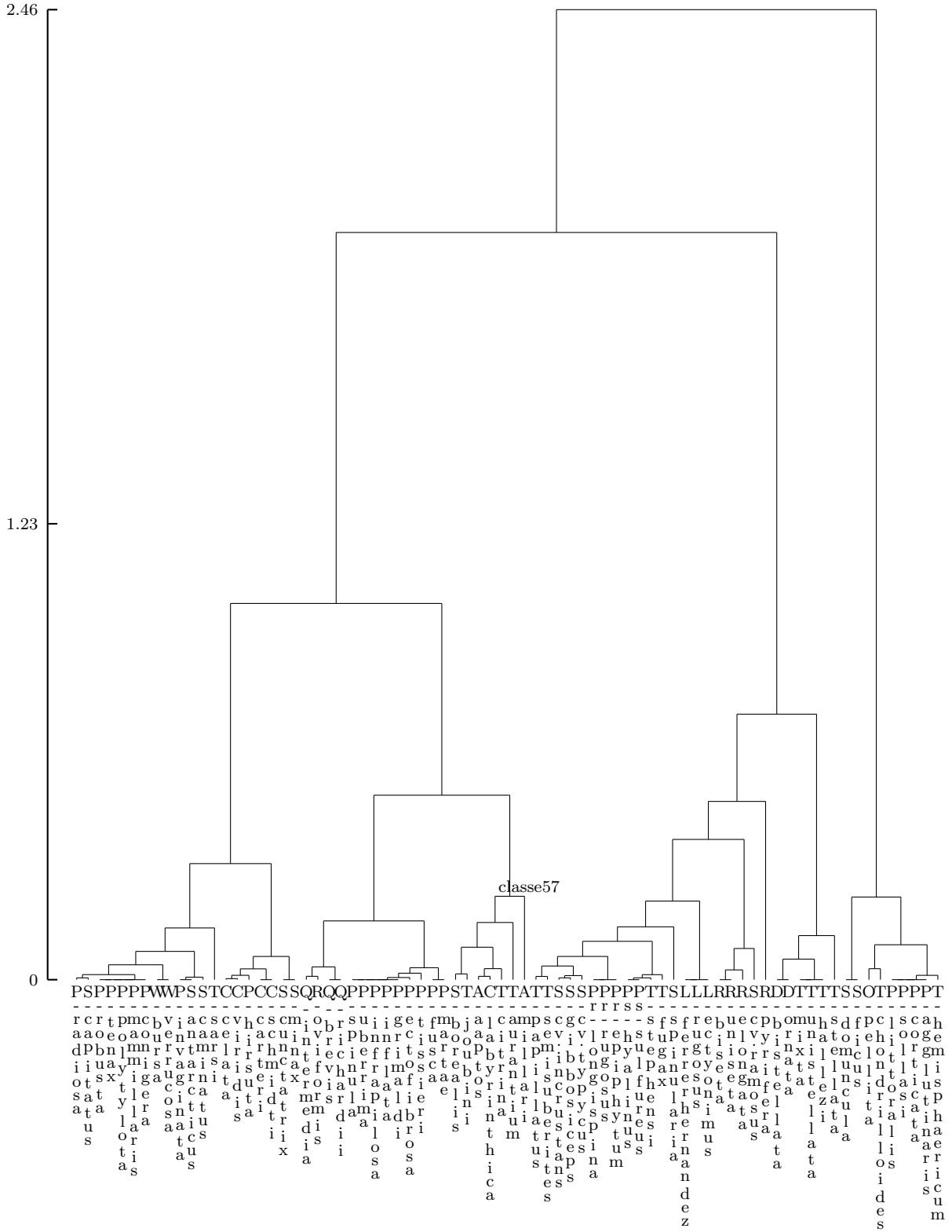
Nota: Les propietats que apareixen amb lletra tipogràfica són descriptores de la classe.

Figure A.2: **Informe de classificació n^o 4 (τ) [...]**



A.7 Segona aproximació

Figure A.3: Classificació amb els atributs mencionats a les regles, només.



Appendix B

Resultats addicionals sobre les estrelles

En aquest annexe s'inclouen alguns resultats de l'anàlisi de les estrelles presentat a §**??** que no figuren en el text per tal de facilitar-ne el seguiment.

Com que les dades són simulades, s'identifica cada estrella amb un codi alfanumèric, i la seva darrera xifra indica a quina classe pertany realment l'estrella en qüestió. Les estrelles acabades amb "-1" són del disc. La resta són de l'halo.

B.1 Primeres classificacions

Passa $\xi = 0$ Primerament hi ha resultats de classificar la totalitat de les estrelles sintètiques sense utilitzar cap regla.

Descripció de les 2 classes

Classe	C98		C97	
Variable	$\bar{x}$	**s**	$\bar{x}$	**s**
PAR	**7.3457**	**6.3799**	**30.0526**	**13.3145**
VR	**-8.2346**	**59.2783**	**-19.0737**	**35.5483**
GALON	**206.0172**	**104.5798**	**137.3853**	**51.2937**
GALAT	**-6.8243**	**37.7693**	**3.3989**	**40.1891**
MLCB	**-0.0101**	**0.0729**	**0.0857**	**0.1440**
MB	**-0.01404**	**0.0605**	**-0.0267**	**0.2415**
MV	**7.4912**	**0.8330**	**6.2319**	**1.8149**
SP	**B0-B4 = 11/81** **A0 = 7/81** **A4-A9 = 5/81** **F5-F9 = 4/81** **K0-K4 = 26/81**	**B5-B9 = 2/81** **A2-A3 = 7/81** **F0-F4 = 2/27** **G0-G7 = 7/81** **K5-K6 = 2/27**	**B0-B4 = 1/19** **A2-A3 = 1/19** **F5-F9 = 5/19** **G8-G9 = 2/19**	**A0 = 2/19** **F0-F4 = 1/19** **G0-G7 = 4/19** **K0-K4 = 3/19**
Grandària	**81**		**19**	

Figure B.1: **Passa** $\xi = 0$: Arbre general de classificació (τ^0, $\alpha = 0.103$ i $\beta = 0.897$)

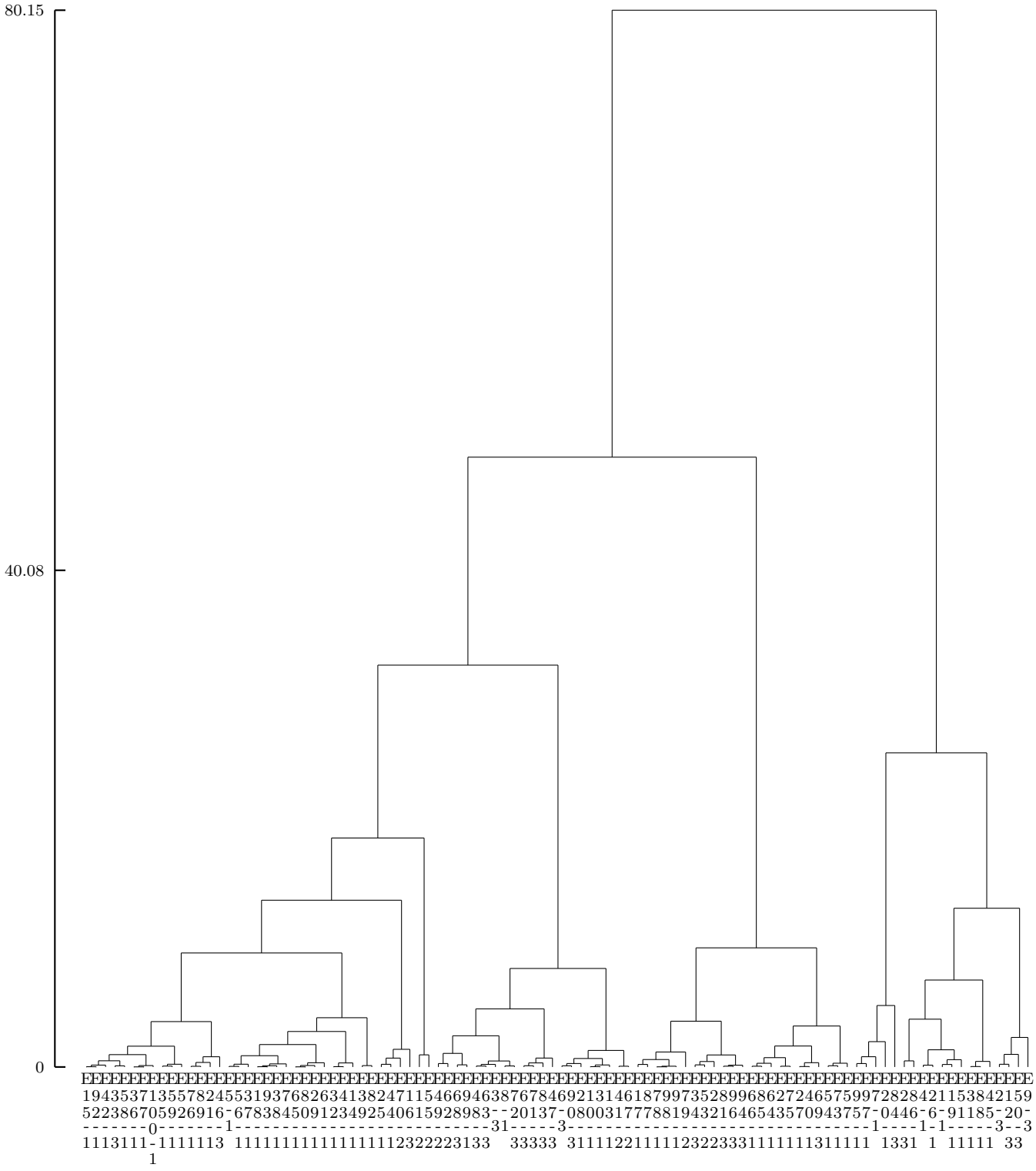


Figure B.2: **Passa** $\xi = 1$: Arbre general de classificació (τ^1)

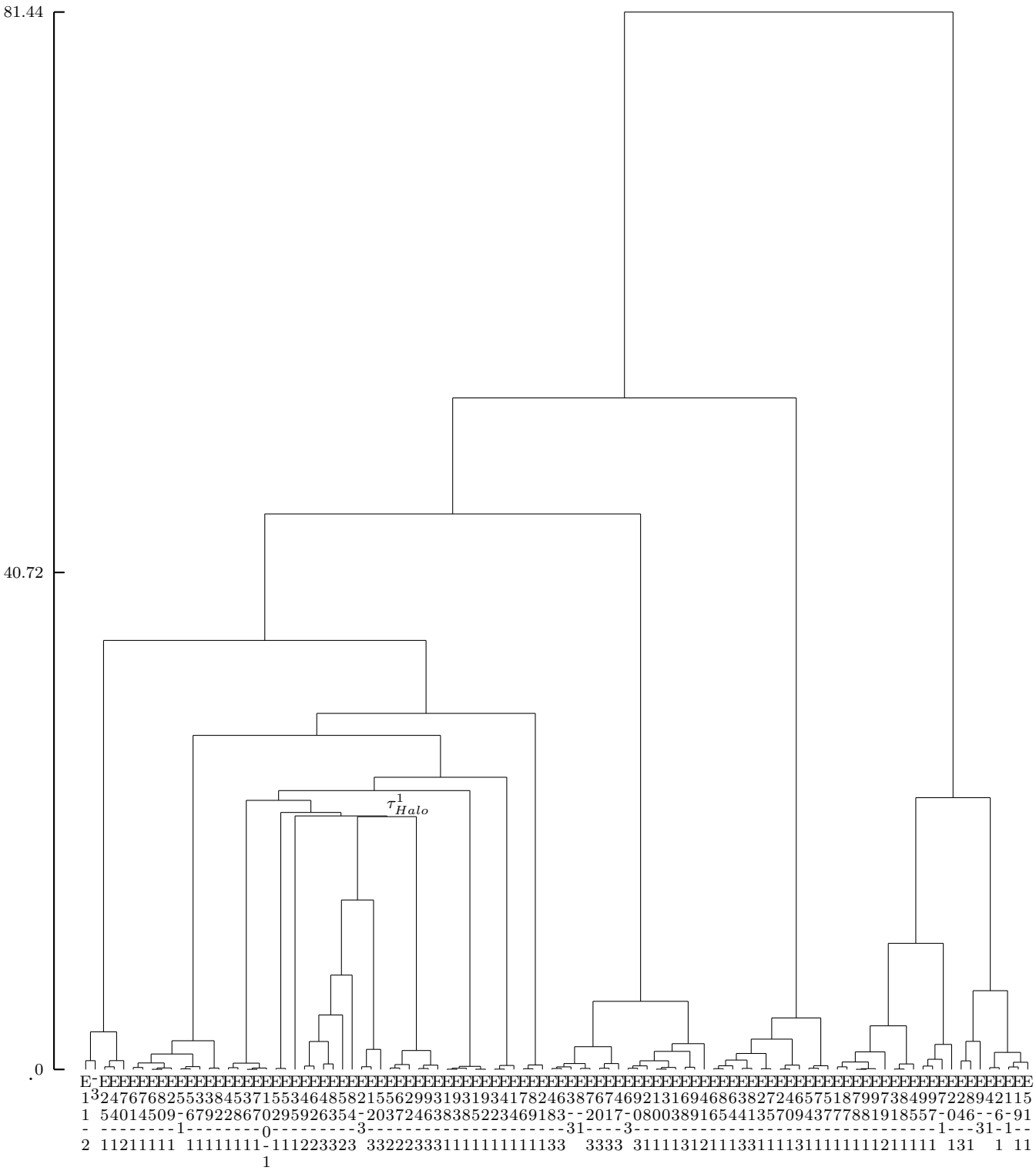
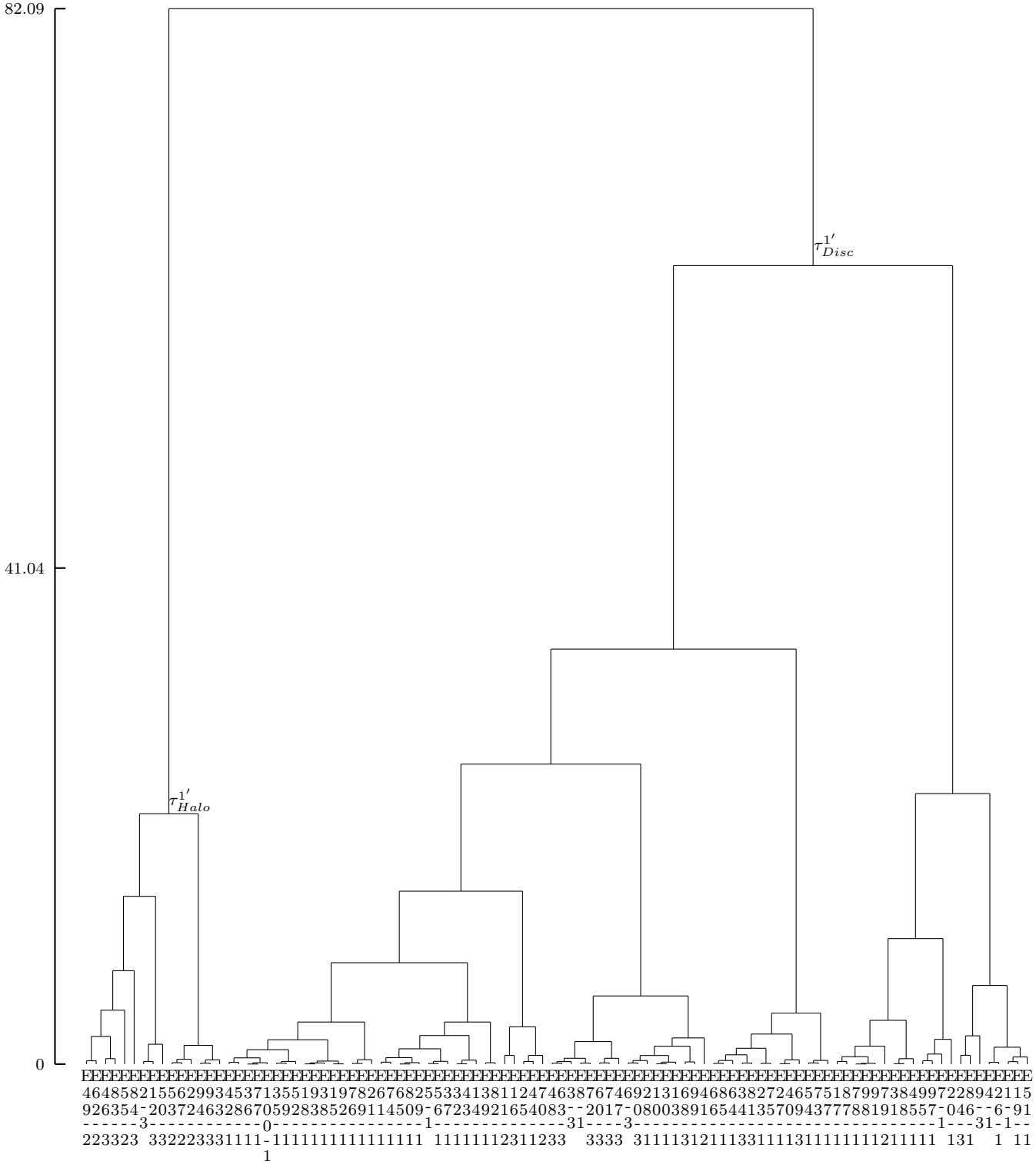


Figure B.3: **Passa** $\xi = 1$ Clausura d'$\mathcal{R}^1$ ($\tau^{1'}$).



Passa $\xi = 2$ Regles d'$\mathcal{R}^2$ en format **LISP**, les d'$\mathcal{R}^1$ ampliat amb les regles derivades del diagrama HR.

```
r1=((rel (> (abs w) 60))-> halo)

; provenen del diagrama H-R

r2=((rel (and (< ( magnitud_absoluta  mv  par galat)
                 5)
           (member (quote sp) '(o7-o9 b0-b4 b5-b9 a0 a2-a3 a4-a9 f0-f4 f5-f9
))))\longrightarrow disc)

r3=((rel (and (> ( magnitud_absoluta  mv  par galat) 9)
           (< ( magnitud_absoluta  mv  par galat) 11.5)
           (equal (quote sp) 'M5+))) \longrightarrow halo)

r4= ((rel (and (>  ( magnitud_absoluta  mv par galat) -2.25)
           (<=  ( magnitud_absoluta  mv par galat) 0.25)
           (equal (quote sp) 'M0-M4))) \longrightarrow disc)

r5= ((rel (and (> (magnitud_absoluta mv par galat) 9.75)
           (<=  (magnitud_absoluta mv par galat) 11.75)
           (equal (quote sp) 'M0-M4))) \longrightarrow disc)

r6= ((rel (and (> (magnitud_absoluta mv par galat) -1.25)
           (<=  (magnitud_absoluta mv par galat) -0.75)
           (equal (quote sp) 'k7-k9))) \longrightarrow halo)

r7= ((rel (and (> (magnitud_absoluta mv par galat)  7.25)
           (<=  (magnitud_absoluta mv par galat) 7.75)
           (equal (quote sp) 'M0-M4))) \longrightarrow halo)

r8= ((rel (and (> (magnitud_absoluta mv par galat) 8.75)
           (<=  (magnitud_absoluta mv par galat) 9.25)
           (equal (quote sp) 'M0-M4))) \longrightarrow disc)

r9= ((rel (and (>  (magnitud_absoluta mv par galat) -0.75)
           (<=  (magnitud_absoluta mv par galat) -0.25)
           (equal (quote sp) 'M0-M4))) \longrightarrow disc)

r10=((rel (and (>  (magnitud_absoluta mv par galat) 2.75)
           (<=  (magnitud_absoluta mv par galat) 3.25)
           (equal (quote sp) 'g0-g7))) \longrightarrow disc)

r11=((rel (and (>  (magnitud_absoluta mv par galat) 0.25)
           (<=  (magnitud_absoluta mv par galat) 2.75)
           (equal (quote sp) 'k0-k4))) \longrightarrow halo)
```

En totes elles la funció **LISP** `(magnitud_absoluta mv par galat)` implementa l'expressió **??**.

A continuació, la classificació de les estrelles usant $\mathcal{R}^2$, seguida de la descripció de les classes.

Node	$\bar{\imath}^2_{Halo}$		Disc		**classeR81**		**classeR75**	
Var.	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
PAR	6.6	4.8961	29.5	17.3109	9.2280	9.5303	36.625	10.0276
VR	-22.8866	113.3569	-3.075	23.6934	-7.3821	57.2181	-18.2625	33.0428
GALON	201.2167	103.3539	185.1025	51.4479	196.8372	102.9692	145.3225	38.2173
GALAT	-8.5305	47.5704	-1.7875	34.78099	-6.5962	38.1754	16.415	34.3115
MLCB	0.06587	0.1362	0.03	0.0636	-0.0149	0.0936	0.1145	0.0946
MB	-0.0481	0.1837	-0.088	0.0507	-0.0191	0.1032	0.0904	0.2003
MV	7.4611	0.6433	3.5885	0.9824	7.4302	1.1805	6.5885	1.1899
SP	G0-G7 = 1/15 G8-G9 = 1/15 K0-K4 = 3/5 K5-K6 = 4/15		B0-B4 = 1/4 A0 = 1/2 A2-A3 = 1/4		B0-B4 = 3/20 B5-B9 = 2/75 A0 = 1/10 A2-A3 = 29/300 A4-A9 = 1/15 F0-F4 = 7/75 F5-F9 = 1/15 G0-G7 = 121/1125 G8-G9 = 1/1125 K0-K4 = 98/375 K5-K6 = 34/1125		F5-F9 = 1/2 G0-G7 = 1/4 G8-G9 = 1/8 K0-K4 = 1/8	
n	15				92		8	

Passa $\xi = 3$ Les regles utilitzades a $\mathcal{R}^3$, referents a la relació entre W i la població que s'infereix a partir de la metal·licitat, són les següents:

```
r12= ((rel (<= (abs w) 10)) -> disc)
r13= ((rel (and (< (abs w) 20) (>= (abs w) 60)))-> halo)
r14= ((rel (> (abs w) 60)) \longrightarrow halo)
r15= ((rel (and (> v -100) (<= v -60))) -> halo)
r16= ((rel (and (> v -150) (<= v -100))) -> halo)
r17= ((rel (<= v -150)) \longrightarrow halo)
```

La partició induïda per les regles conté una classe residual amb 3 objectes que satisfan regles contradictòries, i 44 que no són seleccionats per cap regla.

Les classificacions locals a les classes induïdes per $\mathcal{R}^3$ són als informes 5.H i 5.D. L'arbre resultant és

Figure B.4: CAJ. Arbre general de classificació tallat en 2 classes (τ^2).

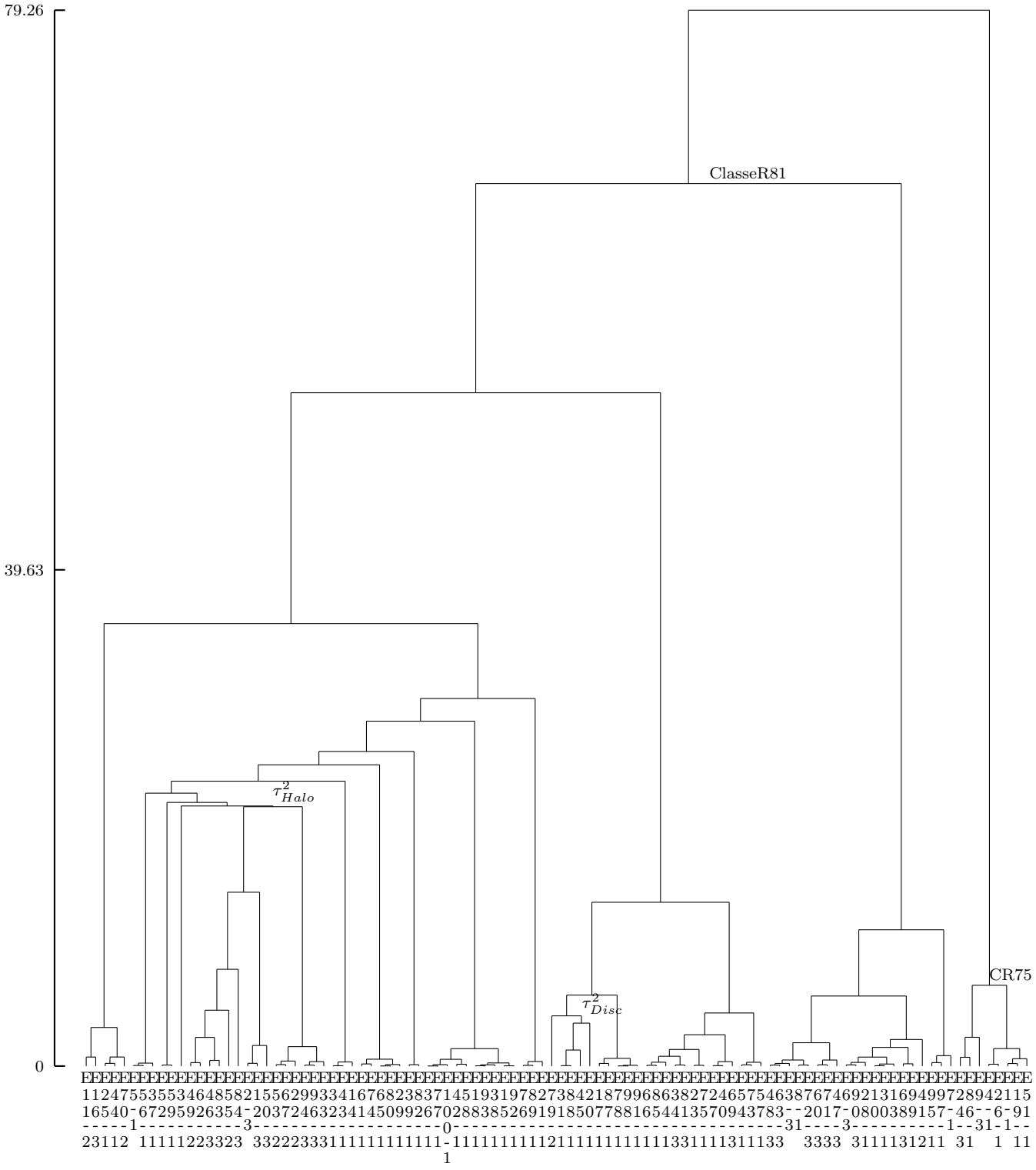
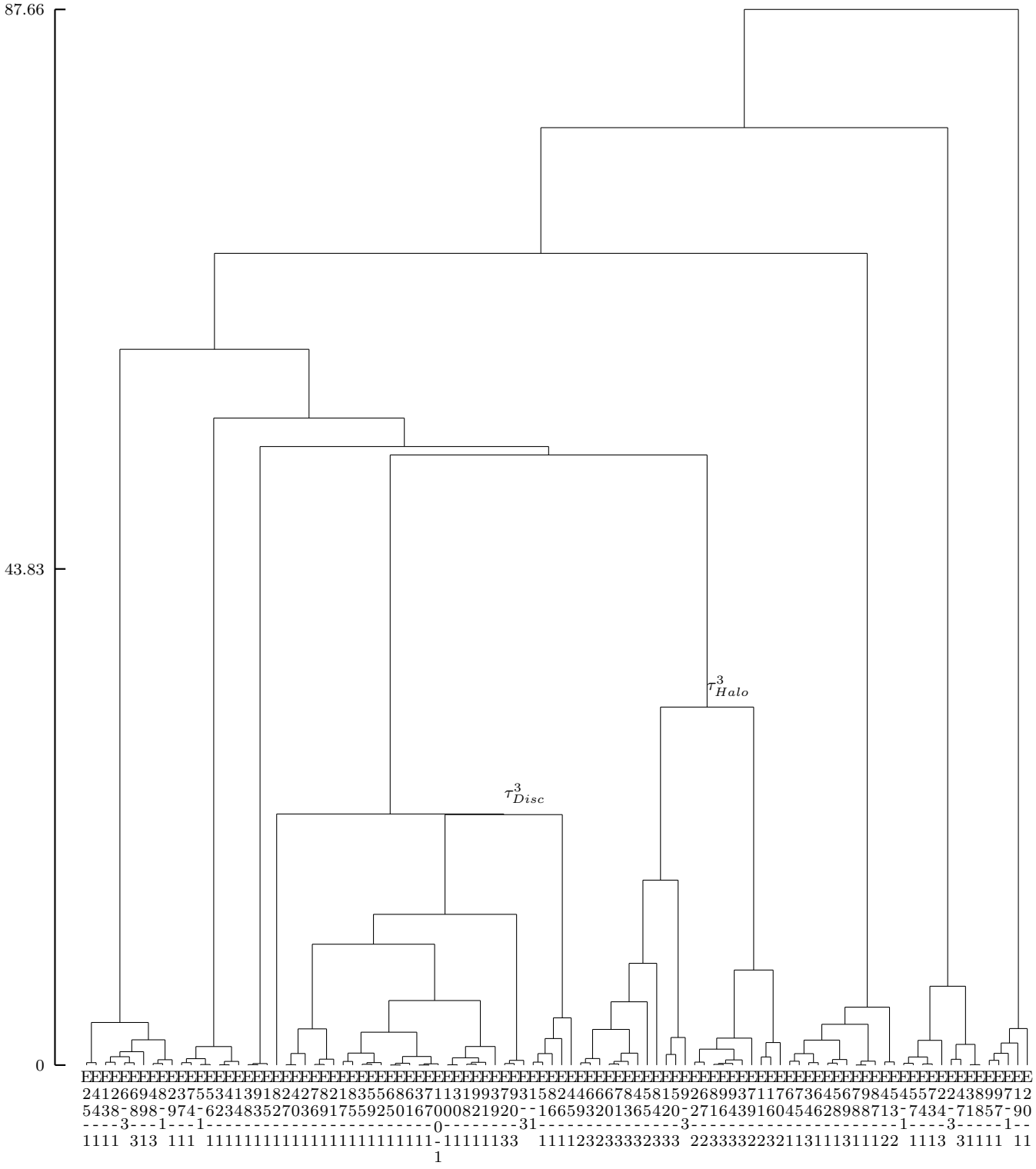


Figure B.5: CAJ. **Passa** $\xi = 3$. Classificació amb $\mathcal{R}^3$: (τ^3).



B.2 Classificació definitiva (Passa $\xi = 4$)

En primer lloc apareix informació sobre la partició induïda per les regles d'$\mathcal{R}^4$. Tot seguit, un tall en dues classes de l'arbre final que integra en una única estructura tots els elements.

Informe de classificació n^o 6.H (τ^4_{Halo})

Dades: DATA/HALO
Nombre d'objectes: 36
Nombre de propietats: 8
Criteri de classificació: WARD
Mètrica utilitzada: MIXTE
Valors de les constants de ponderació: $\alpha = 0.103$ i $\beta = 0.897$

Descripció de les 1 classes ($\bar{\imath}^4_{Halo}$)

classeH35: Coordenades promig:

PAR: 7.0278 , desviació dins la classe 4.9309
VR: -24.8028 , desviació dins la classe 87.5157
GALON: 168.8177 , desviació dins la classe 112.4875
GALAT: -6.8317 , desviació dins la classe 45.1242
MLCB: 0.0213 , desviació dins la classe 0.1388
MB: -0.0137 , desviació dins la classe 0.15389
MV: 7.2882 , desviació dins la classe 1.0012
SP:
A0 = 1/36
A4-A9 = 1/18
F0-F4 = 1/18
G0-G7 = 1/18
G8-G9 = 1/36
K0-K4 = 23/36
K5-K6 = 5/36

Elements de la classe:

36

("E6-3" "E68-3" "E41-2" "E67-2" "E29-1" "E44-1" "E99-1" "E49-2" "E62-2" "E48-3" "E63-3" "E3-3" "E60-3" "E71-3" "E83-3" "E47-3" "E14-1" "E46-3" "E16-3" "E70-2" "E84-3" "E11-2" "E55-2" "E2-3" "E12-3" "E50-3" "E9-3" "E33-3" "E34-3" "E81-3" "E96-3" "E94-3" "E22-2" "E79-2" "E69-3" "E75-1")

Nota: Les propietats que apareixen amb lletra tipogràfica són descriptores de la classe.

Informe de classificació n^o 6.D (τ^4_{Disc})

Dades: DATA/DISC0
Nombre d'objectes: 59
Nombre de propietats: 8
Criteri de classificació: WARD
Mètrica utilitzada: MIXTE
Valors de les constants de ponderació: $\alpha = 0.103$ i $\beta = 0.897$

Descripció de les 1 classes ($\bar{\imath}^4_{Disc}$)

classeD58: Coordenades promig:

PAR: 14.5932 , desviació dins la classe 14.1032
VR: -3.6559 , desviació dins la classe 18.0381
GALON: 209.8683 , desviació dins la classe 88.0957
GALAT: -1.6064 , desviació dins la classe 33.5009
MLCB: 0.0022 , desviació dins la classe 0.06532
MB: -0.0129 , desviació dins la classe 0.0846
MV: 7.2237 , desviació dins la classe 1.3343
SP:
B0-B4 = 11/59
B5-B9 = 2/59
A0 = 7/59
A2-A3 = 8/59
A4-A9 = 3/59
F0-F4 = 5/59
F5-F9 = 9/59
G0-G7 = 8/59
G8-G9 = 1/59
K0-K4 = 5/59

Elements de la classe:

59

("E18-1" "E93-1" "E38-1" "E15-1" "E92-1" "E42-1" "E35-1" "E59-1" "E52-1" "E36-1" "E77-1" "E100-1" "E32-1" "E43-1" "E23-1" "E39-1" "E82-1" "E76-1" "E89-1" "E21-1" "E17-1" "E64-1" "E66-1" "E85-1" "E27-1" "E40-1" "E54-1" "E73-1" "E57-1" "E65-1" "E80-1" "E10-1" "E30-1" "E28-1" "E74-1" "E90-3" "E8-1" "E72-3" "E5-1" "E56-1" "E37-1" "E25-1" "E78-1" "E98-1" "E91-1" "E87-1" "E31-1" "E88-1" "E45-1" "E95-1" "E97-1" "E20-1" "E24-3" "E86-1" "E4-1" "E26-1" "E1-1" "E19-1" "E51-1")

Nota: Les propietats que apareixen amb lletra tipogràfica són descriptores de la classe.

Informe de classificació n^o 6 (τ^4)

Dades: data/resta02
Nombre d'objectes: 100
Nombre de propietats: 8
Criteri de classificació: WARD
Mètrica utilitzada: MIXTE
Valors de les constants de ponderació: $\alpha = 0.103$ i $\beta = 0.897$

Descripció de les 2 classes

classeR5: Coordenades promig:

PAR: 6.6035 , desviació dins la classe 12.8526
VR: 8.5902 , desviació dins la classe 59.0089
GALON: 198.2643 , desviació dins la classe 99.4022
GALAT: -15.3347 , desviació dins la classe 39.5601
MLCB: 0.0013 , desviació dins la classe 0.0979
MB: -0.0135 , desviació dins la classe 0.1128
MV: 7.2018 , desviació dins la classe 1.1907
SP:

B0-B4 = 35/177
B5-B9 = 1/177
A0 = 2435/12744
A2-A3 = 4/177
A4-A9 = 113/6372
F0-F4 = 149/6372
F5-F9 = 3/118
G0-G7 = 203/6372
G8-G9 = 95/12744
K0-K4 = 3661/12744
K5-K6 = 41/216

Elements de la classe:

99

("E53-2" "E6-3" "E68-3" "E41-2" "E67-2" "E29-1" "E44-1" "E99-1" "E49-2" "E62-2" "E48-3" "E63-3" "E3-3" "E60-3" "E71-3" "E83-3" "E47-3" "E14-1" "E46-3" "E16-3" "E70-2" "E84-3" "E11-2" "E55-2" "E2-3" "E12-3" "E50-3" "E9-3" "E33-3" "E34-3" "E81-3" "E96-3" "E94-3" "E22-2" "E79-2" "E69-3" "E75-1" "E18-1" "E93-1" "E38-1" "E15-1" "E92-1" "E42-1" "E35-1" "E59-1" "E52-1" "E36-1" "E77-1" "E100-1" "E32-1" "E43-1" "E23-1" "E39-1" "E82-1" "E76-1" "E89-1" "E21-1" "E17-1" "E64-1" "E66-1" "E85-1" "E27-1" "E40-1" "E54-1" "E73-1" "E57-1" "E65-1" "E80-1" "E10-1" "E30-1" "E28-1" "E74-1" "E90-3" "E8-1" "E72-3" "E5-1" "E56-1" "E37-1" "E25-1" "E78-1" "E98-1" "E91-1" "E87-1" "E31-1" "E88-1" "E45-1" "E95-1" "E97-1" "E20-1" "E24-3" "E86-1" "E4-1" "E26-1" "E1-1" "E19-1" "E51-1" "E61-1" "E58-1" "E13-1")

E7-1: Coordenades promig:

`PAR:` 34.0
`VR:` -0.8
`GALON:` 27.14
`GALAT:` -63.9
`MLCB:` -0.073
`MB:` -0.33
`MV:` 7.928
`SP:` G0-G7

Elements de la classe:

1

("E7-1")

Nota: Les propietats que apareixen amb lletra tipogràfica són descriptores de la classe.

Figure B.6: Índex de nivell d'agregacions successives

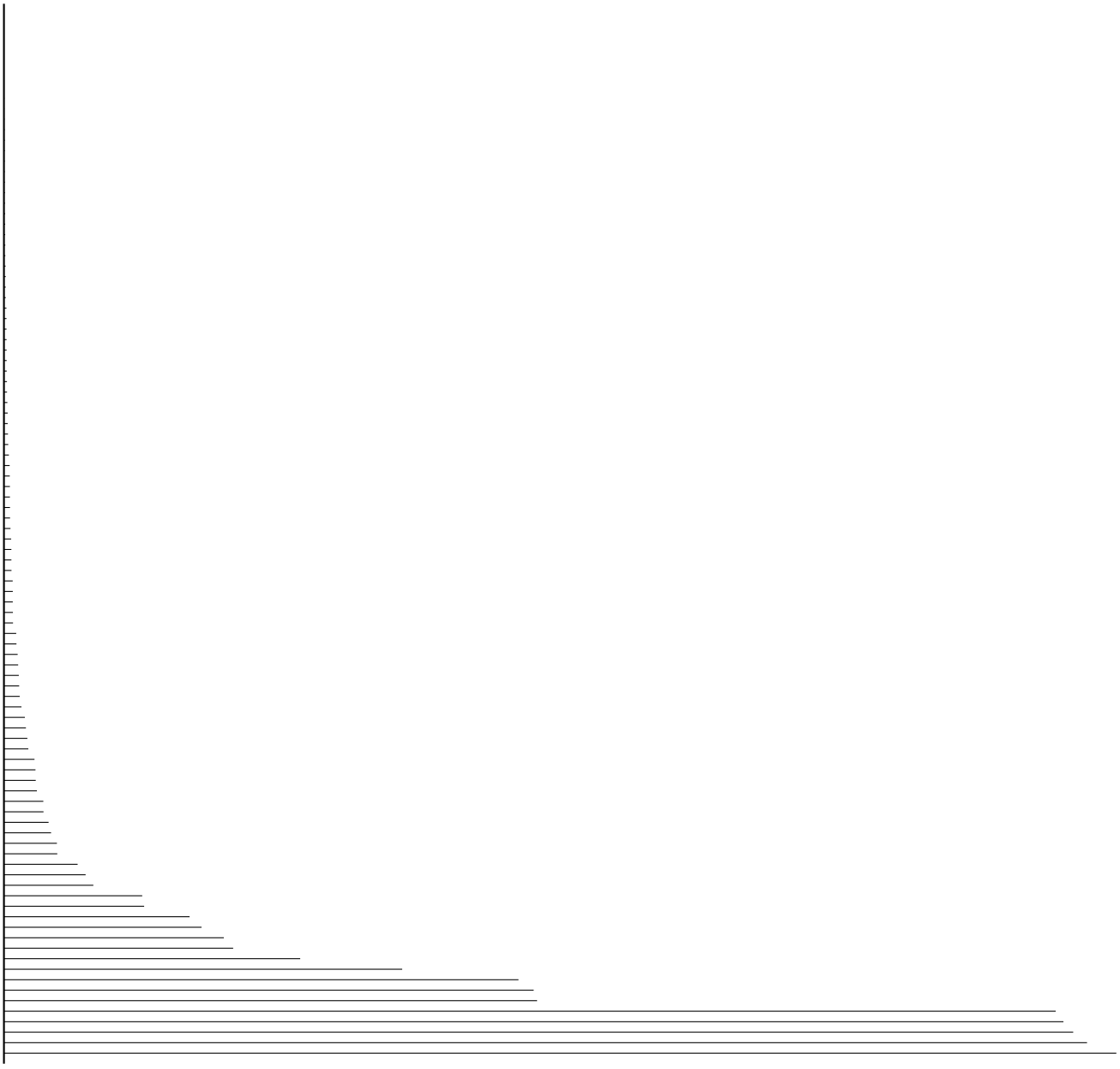
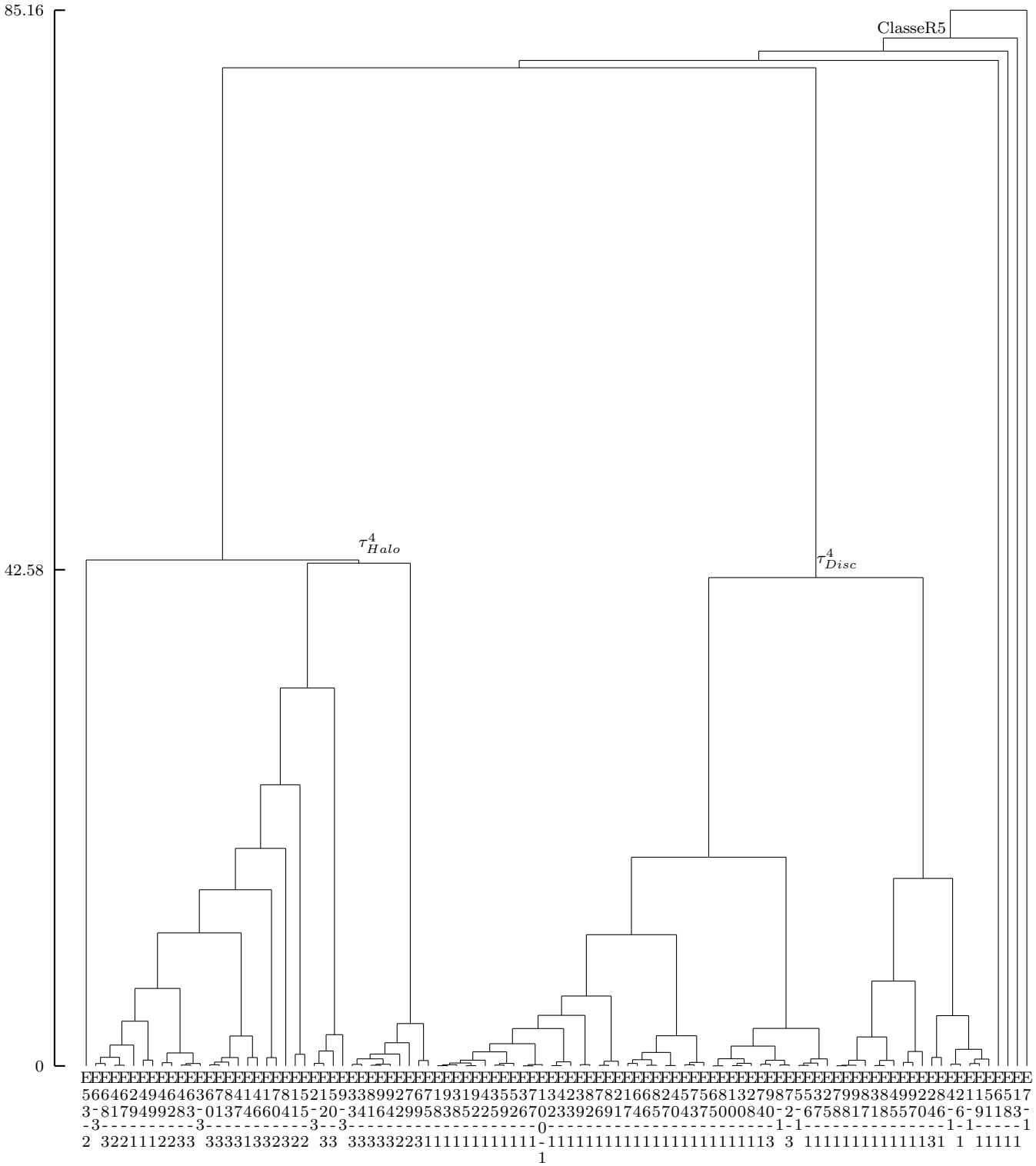


Figure B.7: **Passa** $\xi = 4$: (τ^4).



Appendix C

Resultats addicionals d'altres aplicacions

C.1 Microordinadors

Clust.	Objectes
1	APPLE-II ATARI-800 COMMODORE-VIC-20-A COMMODORE-VIC-20-B OHIO-SC.-CHALLENGER OHIO-SC.-II-SERIES
2	HP-85
3	ZENITH-H8 ZENITH-H89 TRS-80-III
4	EXIDI-SORCERER HORIZON TRS-80-I

Table C.1: Classificació proposada per [MICH83] amb classificació conceptual.

Clust.	Objectes
1	APPLE-II ATARI-800 OHIO-SC.-II-SERIES TRS-80-I
2	HP-85
3	COMMODORE-VIC-20-A COMMODORE-VIC-20-B OHIO-SC.-CHALLENGER
4	EXIDI-SORCERER ZENITH-H8 ZENITH-H89 HORIZON TRS-80-III

Table C.2: Classes presentades a [CHID92] amb la classificació per veïns recíprocs i una mesura de similitud pròpia.

Clust.	Objectes
1	APPLE-II ATARI-800 OHIO-SC.-II-SERIES TRS-80-I EXIDI-SORCERER OHIO-SC.-CHALLENGER
2	HP-85
3	COMMODORE-VIC-20-A COMMODORE-VIC-20-B
4	ZENITH-H8 ZENITH-H89 HORIZON TRS-80-III

Table C.3: Classificació presentada a [CHID92] usant el salt mínim i la mateixa mesura de similitud que a la taula anterior.

Objectes	Display	RAM	ROM	MP	Keys
APPLE-II	COLOR-TV	48	10	6502	52
ATARI-800	COLOR-TV	48	10	6502	57-63
COMMODORE-VIC-20-A	COLOR-TV	32	11	6502A	64-73
COMMODORE-VIC-20-B	COLOR-TV	32	16	6502A	64-73
EXIDI-SORCERER	B-&-W-TV	48	4	Z80	57-63
ZENITH-H8	BUILT-IN	64	1	8080A	64-73
ZENITH-H89	BUILT-IN	64	8	Z80	64-73
HP-85	BUILT-IN	32	80	HP	92
HORIZON	TERMINAL	64	8	Z80	57-63
OHIO-SC.-CHALLENGER	B-&-W-TV	32	10	6502	53-56
OHIO-SC.-II-SERIES	B-&-W-TV	48	10	6502C	53-56
TRS-80-I	B-&-W-TV	48	12	Z80	53-56
TRS-80-III	BUILT-IN	48	14	Z80	64-73

Table C.4: Descripció de les dades classificades.

C.2 Simulació de cubs

En aquesta secció es presenten resultats de la classificació dels cubs.

C.2.1 Prova primera

Informe de classificació

Dades: data/cub
Nombre d'objectes: 150
Nombre de propietats: 3
Criteri de classificació: WARD
Mètrica utilitzada: EUCL-NORM

Descripció de les 3. classes

classe147: Coordenades promig:

- *X1:* 0.82903 , desviació dins la classe 0.10527
- *X2:* 0.83743 , desviació dins la classe 0.10215
- *X3:* 0.80769 , desviació dins la classe 0.09213

Elements de la classe:

50

("I129" "I143" "I121" "I132" "I134" "I128" "I137" "I140" "I131" "I110" "I114" "I102" "I119" "I148" "I107" "I142" "I127" "I123" "I106" "I130" "I118" "I149" "I147" "I111" "I122" "I117" "I136" "I115" "I146" "I150" "I116" "I135" "I103" "I139" "I141" "I126" "I145" "I109" "I104" "I144" "I101" "I112" "I113" "I105" "I133" "I125" "I138" "I120" "I124" "I108")

classe143: Coordenades promig:

- *X1:* 0.50071 , desviació dins la classe 0.10763
- *X2:* 0.48819 , desviació dins la classe 0.09464
- *X3:* 0.51179 , desviació dins la classe 0.10053

Elements de la classe:

50

("I69" "I97" "I63" "I64" "I84" "I59" "I87" "I56" "I58" "I60" "I94" "I62" "I66" "I65" "I90" "I68" "I79" "I51" "I81" "I92" "I78" "I98" "I75" "I80" "I70" "I77" "I88" "I71" "I86" "I91" "I96" "I53" "I95" "I61" "I54" "I73" "I67" "I82" "I100" "I99" "I55" "I74" "I52" "I89" "I72" "I76" "I83" "I57" "I85" "I93")

classe82: Coordenades promig:

- *X1:* 0.15738 , desviació dins la classe 0.08975
- *X2:* 0.18455 , desviació dins la classe 0.09821
- *X3:* 0.17996 , desviació dins la classe 0.10913

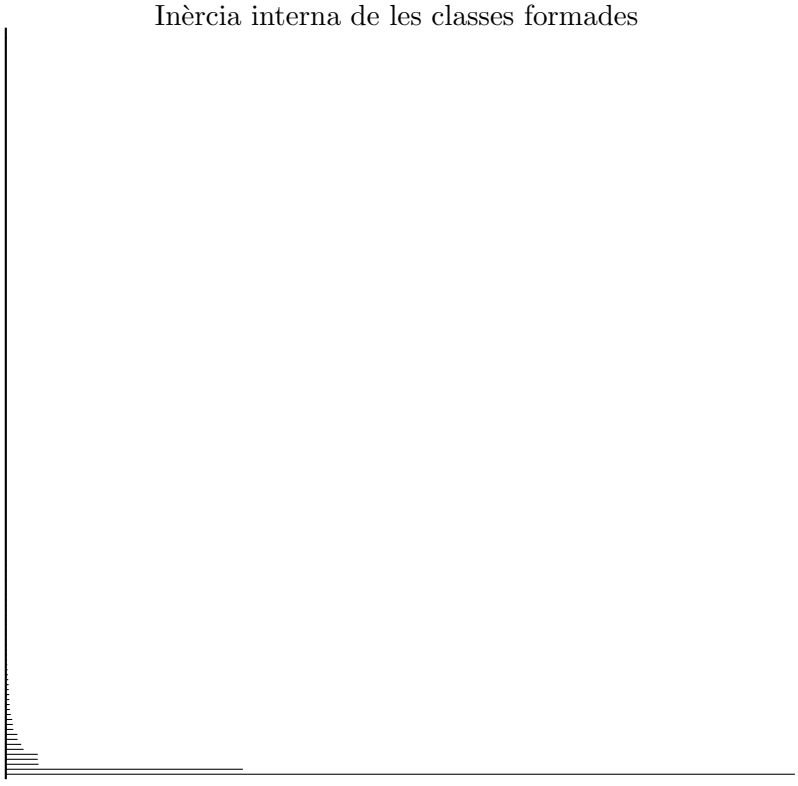
Elements de la classe:

50

("I6" "I24" "I10" "I3" "I20" "I12" "I44" "I9" "I35" "I14" "I37" "I8" "I2" "I5" "I29" "I23" "I32" "I27" "I7" "I28" "I19" "I39" "I26" "I31" "I17" "I30" "I4" "I47" "I36" "I38" "I49" "I1" "I18" "I16" "I41" "I43" "I46" "I34" "I40" "I13" "I22" "I33" "I11" "I45" "I21" "I42" "I48" "I50" "I15" "I25")

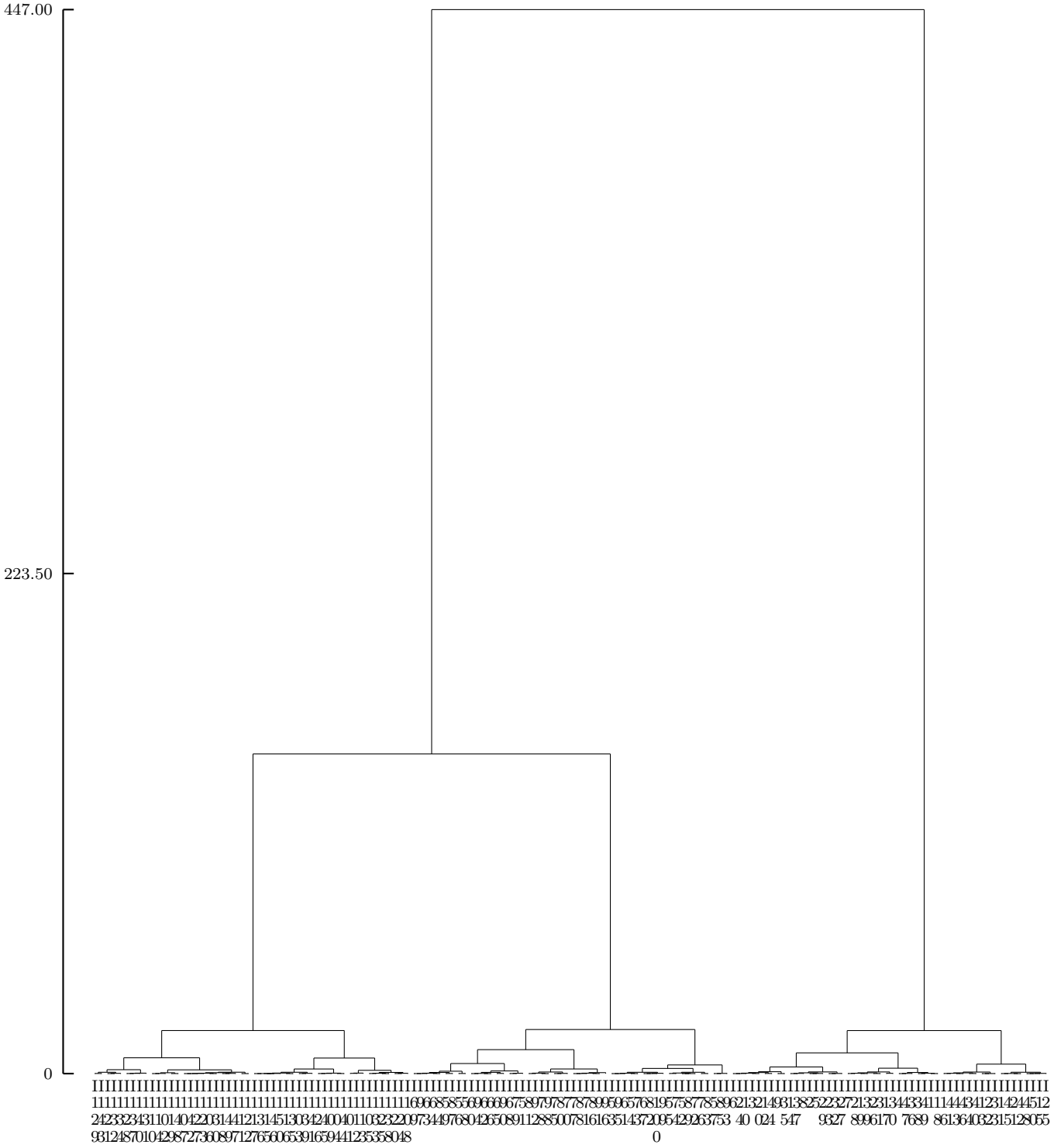
Nota: Les propietats que apareixen amb lletra tipogràfica són descriptores de la classe.

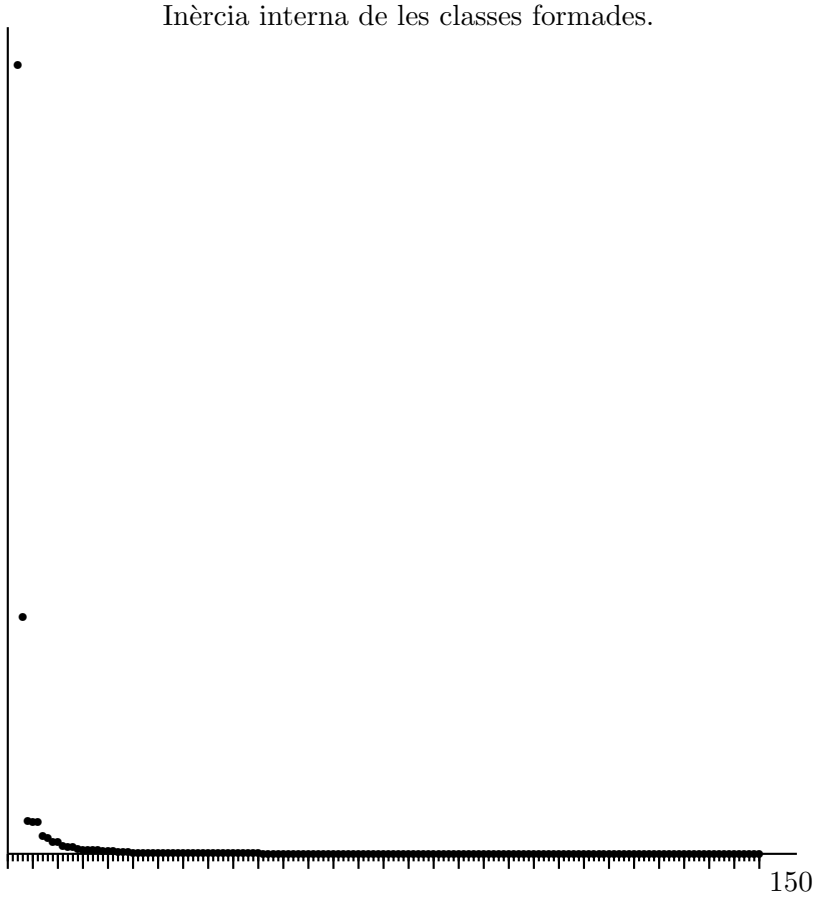
Inèrcia interna de les classes formades



447.0000

Figure C.1: CAJ. Prova primera.





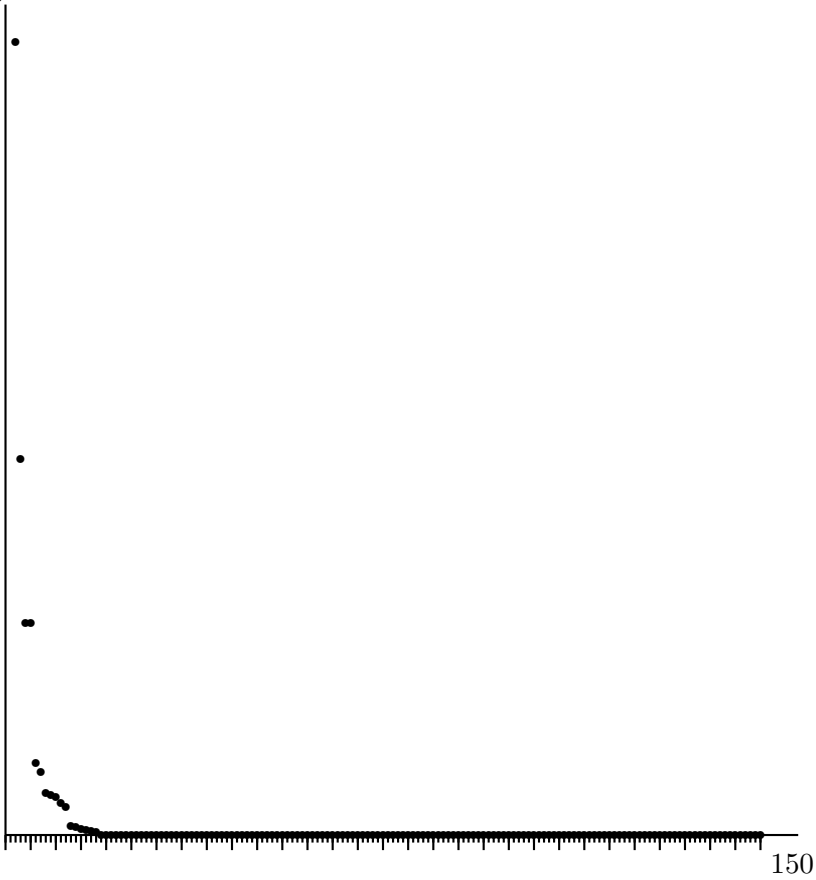
C.2.2 Proves segona i tercera

A continuació es mostra el gràfic d'autocorrelació dels índexos de nivell, seguit de la descripció d'un tall en 5 classes i el dendrograma que els tres paquets proporcionen per la segona prova.

Classe	X_1	X_2	X_3	**Elements**
classe145	E	A	C = 0.46429 D = 0.53571	28
classe144	D	A = 0.77143 B = 0.22857	B = 0.14286 C = 0.37143 D = 0.48571	35
classe143	C	A = 0.41667 B = 0.58333	B = 0.58333 C = 0.41667	24
classe143	B	A = 0.20000 B = 0.80000	A = 0.33333 B = 0.43333 C = 0.23333	30
classe142	A	B	A = 0.66667 B = 0.33333	33

Table C.5: Descripció de les classes (prova segona).

Figure C.2: Inèrcia interna de les classes formades. Prova tercera.



$$\begin{aligned}\mathcal{P} = \{\mathcal{C}_1 = \{ \quad & i_1, i_7, i_{13}, i_{16} - i_{19}, i_{22}, i_{26}, i_{28}, i_{30}, i_{31}, i_{33}, i_{38}, i_{39}, i_{47}, i_{49}, i_{56}, i_{58}, i_{60}, i_{62}, \\ & i_{63}, i_{65}, i_{68}, i_{69}, i_{79}, i_{84}, i_{90}, i_{96}, i_{97}\}, \\ \mathcal{C}_2 = \{ \quad & i_2 - i_6, i_8 - i_{12}, i_{14}, i_{15}, i_{20}, i_{21}, i_{23} - i_{25}, i_{27}, i_{29}, i_{32}, i_{34} - i_{37}, i_{40} - i_{46}, \\ & i_{48}, i_{50}\}, \\ \mathcal{C}_3 = \{ \quad & i_{51}, i_{53}, i_{54}, i_{59}, i_{61}, i_{64}, i_{66}, i_{70}, i_{71}, i_{73} - i_{78}, i_{80}, i_{83}, i_{86} - i_{88}, i_{91}, i_{94}, i_{95}, i_{98}\}, \\ \mathcal{C}_4 = \{ \quad & i_{52}, i_{55}, i_{57}, i_{67}, i_{72}, i_{81}, i_{82}, i_{85}, i_{89}, i_{92}, i_{93}, i_{99}, i_{100}, i_{102}, i_{106}, i_{107}, i_{110}, i_{114}, \\ & i_{118}, i_{119}, i_{121}, i_{122}, i_{127}, i_{128}, i_{130} - i_{132}, i_{134}, i_{137}, i_{140}, i_{142}, i_{143}, i_{147} - i_{149}\}, \\ \mathcal{C}_5 = \{ \quad & i_{101}, i_{103} - i_{105}, i_{108}, i_{109}, i_{111} - i_{113}, i_{115} - i_{117}, i_{120}, i_{123} - i_{126}, i_{129}, i_{133}, \\ & i_{135}, i_{136}, i_{138}, i_{139}, i_{141}, i_{144} - i_{146}, i_{150}\}\}\end{aligned}$$

Classe	X_1	X_2		X_3		**Elements**
		$\bar{x}_{\mathcal{C}}$	$s_{\mathcal{C}}$	$\bar{x}_{\mathcal{C}}$	$s_{\mathcal{C}}$	
classe148	D = 0.55556 E = 0.44444	0.76412	0.17807	0.75015	0.14733	63
classe146	B = 0.55556 C = 0.44444	0.39868	0.16535	0.41017	0.17317	54
classe140	A	0.17699	0.09538	0.16859	0.11439	33

Table C.6: Descripció de les classes (prova tercera).

Figure C.3: CAJ. Arbre general de classificació (prova tercera).