# TESI DOCTORAL

Títol Objectivant per consens la metodologia dels mapes conceptuals

Realitzada per Francesc Teixidó i Navarro

en el Centre Campus La Salle

i en el Departament Departament d'Enginyeria

Dirigida per Dr. Albert Fornells Herrera
Dra. Elisabet Golobardes i Ribé

*C. Claravall, 1-3*
*08022 Barcelona*
*Tel. 936 022 200*
*Fax 936 022 249*
*E-mail: urlsc@sec.url.es*
*www.url.es*

# Agraïments

El present document recull tota una dècada de treball i dedicació (a batzegades) de recerca. Tota aquesta feina no hagués estat possible sense el suport de les persones i aquells grups de recerca i projectes que representen. Grups de recerca que m'han aixoplugat: GRSI, GR-SETAD o HTM. Gràcies a aquests he pogut compartir inquietuds i participar en projectes interessants.

Agrair a tots aquells amics que he anat trobant per aquest camí, aquells que han convertit el camí més costerut en planer. Tots m'han ajudat i espero haver ajudat: Albert O., Álvaro G., Andreu S., Ester B., Guio C., Mireia G., Mónica S., Núria A., Xari R., Xavi V. i Zaida R.

Ricard Santomà per permetre'm formar part de l'equip que representa, per empentar-me i permetre'm participar en la recerca que va iniciar i no ha deixat. Per . . . la confiança.

Amb més que estima, sabeu que sense vosaltres no hagués estat possible. Diuen que un simple gràcies diu moltes coses. . . Gràcies Albert F. i Elisabet G.

Mai es poden concloure sense agrair el suport dels propis; la família. Qui amb estima, temps, dedicació, paciència i consentiment m'han entès i acompanyat sempre. Als meus pares, a la Caro i (als meus *followers*) Cesc i Lluc.

# Resum

Independent de si ens fixem en el sector del turisme, de les ciències de la salut o de l'energia per posar alguns exemples, qualsevol tipus d'empresa o institució ha d'afrontar el repte de prendre decisions a partir de l'anàlisi de grans volums de dades. En aquest sentit, les eines de suport per a la presa de decisió s'han convertit en un element indispensable.

La metodologia dels mapes conceptuals serveix per ajudar a conceptualitzar un pensament abstracte a partir de posar en comú l'opinió subjectiva dels experts. Durant aquest procés es genera un gran volum de dades que analitzar i discriminar de manera individual per part de tots els experts.

Aquesta tesi proposa una millora d'aquesta metodologia per canalitzar la informació d'interès i evidenciar les opinions en consens. Això permet reduir l'espai de solucions dràsticament i, conseqüentment, el procés d'anàlisi es fa més eficient. La proposta s'ha aplicat exitosament en l'àmbit de l'hospitalitat per a respondre la pregunta: "Quins són els factors que contribueixen en l'excel · lència en l'hospitalitat?"

**Paraules clau.** Mapes conceptuals, Consens, Clustering, Aprenentatge iteratiu de regles, Interpretabilitat de les regles, Sistemes de Suport a la Presa de Decisions, Computació evolutiva, Mineria de dades, Hospitalitat.

# Resumen

Independiente de si nos fijamos en el sector del turismo, de las ciencias de la salud o de la energía para poner algunos ejemplos, cualquier tipo de empresa o institución debe afrontar el reto de tomar decisiones a partir del análisis de grandes volúmenes de datos. En este sentido, las herramientas de apoyo para la toma de decisión se han convertido en un elemento indispensable.

La metodología de los mapas conceptuales sirve para ayudar a conceptualizar un pensamiento abstracto a partir de poner en común la opinión subjetiva de los expertos. Durante este proceso se genera un gran volumen de datos que analizar y discriminar de manera individual por parte de todos los expertos.

Esta tesis propone una mejora de esta metodología para canalizar la información de interés y evidenciar las opiniones en consenso. Esto permite reducir el espacio de soluciones drásticamente y, consecuentemente, el proceso de análisis se hace más eficiente. La propuesta se ha aplicado exitosamente en el ámbito de la hospitalidad para responder la pregunta: "¿Cuáles son los factores que contribuyen en la excelencia en la hospitalidad?"

**Palabras clave.** Mapas conceptuales, Consenso, Clustering, Aprendizaje iterativo de reglas, Interpretabilidad de las reglas, Sistemas de soporte a la toma de decisiones, Computación evolutiva, Mineria de datos, Hospitalidad.

# Abstract

Regardless if you look at the tourism sector, health or energy sciences to give only a few examples, any company or institution must meet the challenge of making decisions based on the analysis of large volumes of data. In this regard, support tools for decision making have become indispensable.

The methodology of concept mapping used in this research proposes an improvement in order to help conceptualize abstract thought from pooling the subjective opinion of experts. During this process, a large volume of data is generated which needs to be analyzed and studied by all the experts.

This can dramatically reduce the solution space, and consequently the analysis process becomes more efficient. The proposal has been applied successfully to the field of hospitality to answer the question: "what are the factors that contribute to excellence in hospitality?"

**Keywords.** Concept Mapping, Consensus, Clustering, Iterative rule learning, Rule interpretability, Decision Support Systems, Evolutionary computation, Data mining, Hospitality.

# Tesi doctoral per compendi de publicacions

La present tesi doctoral s'acull a la normativa per a l'elaboració de tesis doctorals per compendi de publicacions de la Universitat Ramon Llull[1]. La normativa consta dels següents punts:

1. Una tesi doctoral per compendi de publicacions estarà formada per un mínim de tres articles sobre una mateixa línia d'investigació.

2. Només s'acceptaran articles de publicacions que disposin d'un sistema d'avaluació per *peer review* i/o que estiguin indexades preferentment en bases de dades científiques internacionals.

3. Només s'acceptaran articles publicats, o acceptats per a la seva publicació, realitzats amb data posterior a la primera matriculació del doctorand als estudis de doctorat o màster oficial.

4. Els coautors dels articles publicats donaran la seva conformitat per escrit a la utilització de l'article com a part de la tesi del doctorand.

5. Els coautors dels articles publicats no formaran part del tribunal de la tesi.

6. Els coautors dels articles publicats i utilitzats en una tesi que no tinguin el grau de doctor renunciaran per escrit a utilitzar l'article en una altra tesi. En el cas que els articles publicats siguin de més d'un equip de recerca, la Comissió de Doctorat del centre podrà considerar excepcions justificades en l'aplicació d'aquesta norma.

7. La tesi comptarà amb una introducció general que presenti els treballs publicats, una justificació de la unitat temàtica, una còpia de cada treball publicat, un resum global dels resultats, la seva discussió i les conclusions finals.

8. Per tot el citat anteriorment, s'haurà de presentar sempre, a l'inici del procés de la tesi, una sol · licitud formal a la Comissió de Doctorat del centre i obtenir la seva acceptació favorable. La Comissió vetllarà per la qualitat de les publicacions que es volen presentar per a la Tesi. A la sol · licitud s'afegirà també un informe del director de la Tesi indicant quina és la contribució específica del doctorand al treball presentat i la de la resta d'autors, si s'escau. S'haurà de presentar l'acta d'aprovació de la Comissió del centre a la Comissió de Doctorat de la URL en el moment de la tramitació ordinària de la Tesi.

[1]Aprovada per la Junta Acadèmica a 18 de setembre de 2008

Aquesta tesi compleix amb tots els punts prèviament citats. Les tres publicacions que formen el compendi són les següents:

1. F. Teixidó-Navarro, A. Orriols-Puig, and E. Bernadó-Mansilla. Hierarchical Evolution of Linear Regressors. In Genetic and Evolutionary Computation Conference (GECCO'07), ACM Press, 1413–1420, ISBN 978-1-60558-131-6, 2008. Publicació en congrés Internacional Core A de Computer Science.

2. A. Garcia-Piquer, A. Sancho-Asensio, A. Fornells, E. Golobardes, G. Corral and F. Teixidó-Navarro. Towards a High Solution Retrieval in Multiobjective Clustering. Information Sciences, vol 320: 12-25, 2015. Publicació indexada al Q1 amb factor d'impacte 3,8.

3. A. Fornells, Z. Rodrigo, R. Santomà, X. Rovira, M. Sanchez, F. Teixidó-Navarro and E. Golobardes. Promoting consensus in the concept mapping methodology: An application in the hospitality sector. Pattern Recognition Letters, doi:10.1016/j.patrec.2015.05.013, 2015. Publicació indexada al Q3 amb factor d'impacte 1,062.

# Índex

# 1 Introducció

En aquest capítol es descriuen els marcs de treball i de recerca de la tesi així com la problemàtica que ha motivat aquesta tesi. Concretament, la tesi proposa una estratègia per a què els experts prenguin decisions més objectives en l'ús de la metodologia dels mapes conceptuals. La metodologia s'ha avaluat satisfactòriament en un cas d'estudi real.

## 1.1 Marc de recerca

Les dades són l'origen del coneixement i la base de la societat de la informació (Castells and Martínez, 2001). Actualment, cada activitat es detecta, registra i analitza amb la finalitat de transformar les dades en coneixement per donar suport als experts d'un determinat àmbit a l'hora de prendre una decisió.

La previsió d'estocs en magatzems (Barak and Modarres, 2015), l'establiment de rutes comercials (Fu et al., 2006), la previsió de la demanda elèctrica (Sung and Ko, 2015; Jurado et al., 2015) o la detecció de càncer de mama (Kourou et al., 2015) són alguns exemples de domini amb naturaleses, requeriments i nivells de criticitat diferents, els quals han de fer front al mateix repte: analitzar grans volums de dades per extreure coneixement que ajudi als experts d'un àmbit a donar resposta a un problema. És justament en aquest escenari on les tècniques de l'àmbit de la intel · ligència artificial (Russell and Norvig, 2009) i la mineria de dades (Witten and Frank, 2011) s'han convertit en la pedra angular que dóna vida i forma als sistemes de suport a la presa de decisió.

John McCarthy va definir per primer cop el terme d'Intel · ligència artificial com "la ciència i l'enginyeria de fer màquines intel · ligents" al 1956 en el marc de la conferència

de Dartmouth College (McCarthy et al., 1955). Una màquina intel · ligent es defineix com aquella màquina que té la capacitat de realitzar processos intel · ligents com aprendre, raonar, solucionar problemes, percebre o comprendre el llenguatge natural. Totes aquestes característiques han fet que la intel · ligència artificial hagi esdevingut una part essencial de la indústria tecnològica capaç de solucionar molts dels problemes més difícils de la informàtica, gràcies a que combina els fonaments de diferents disciplines que aborden el repte d'aprendre i entendre per solucionar problemes i prendre decisions (Russell and Norvig, 2009; Negnevitsky, 2001; Nilsson, 1998).

D'altra banda, les tècniques de la Mineria de dades es centren en extreure coneixement útil i ocult de les dades a partir de l'anàlisi de les seves estructures i relacions. Aquestes tècniques tenen la capacitat d'arribar allà on la ment humana no pot, doncs hom no és capaç de processar ni el seu gran volum ni el nivell de detall que tenen. En general, les tècniques de mineria de dades es poden agrupar en quatre grans tipologies de problemes (Fayyad et al., 1996):

**Clustering.** Identificar patrons en les dades a partir d'agrupar informació en base a un conjunt de criteris. Per exemple, identificar les característiques més rellevants d'un tipus de càncer.

**Sistemes regressors.** Trobar una funció matemàtica capaç de descriure la relació d'un conjunt de variables. Per exemple, modelar el consum elèctric dels usuaris d'una ciutat.

**Classificació.** Assignar una categoria a un exemple prenent com a referència les seves característiques. Per exemple, si una transacció electrònica és fraudulenta o no.

**Regles d'associació.** Descriure un cert comportament a partir d'un conjunt de regles. Per exemple, tots aquells que visiten el producte A i el producte B aleshores compraren el producte C.

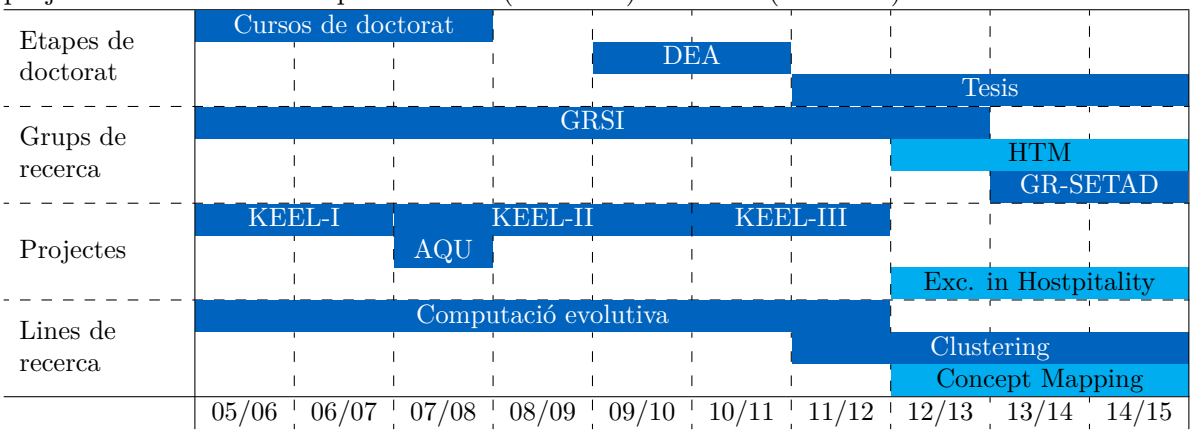
Per tant, la intel · ligència artificial i la mineria de dades són tècniques que ens possibiliten construir sistemes capaços de processar grans volums d'informació a partir de la simulació d'estratègies de raonament similars a la dels humans. Són totes aquestes capacitats tan especials les que van motivar el meu interès i posterior especialització en aquest àmbit.

## 1.2 Marc de treball

La present tesi es desenvolupa dins del programa de doctorat "Tecnologies de la Informació i les Comunicacions i la seva gestió" de La Salle de la Universitat Ramon Llull (URL). Durant el programa s'ha participat com a investigador en tres grups de recerca de dos campus de la URL que, tot i tenir programes i projectes formatius molt diferents, comparteixen la mateixa inquietud de recerca: donar suport a l'expert a la presa de decisions.

A continuació, s'explica l'activitat de recerca realitzada al Campus La Salle i *School of Tourism and Hospitality Management Sant Ignasi* (HTSI), la qual es descriu en el cronograma de la taula 1.1.

TAULA 1.1: L'activitat desenvolupada ha tingut lloc en el marc de treball de grups de recerca de dos centres de la URL. El cronograma temporalitza les línies de recerca emmarcades en els projectes realitzats al campus La Salle (blau fosc) i a HTSI (blau clar).




### 1.2.1 Activitat de recerca en el Campus La Salle - URL

El Campus la Salle és un centre amb més de 100 anys d'història especialitzat en programes formatius d'Enginyeria, Arquitectura i Gestió. La recerca realitzada ha estat vinculada principalment al Grup de Recerca en Sistemes Intel · ligents (GRSI), el qual té com objectiu la creació d'eines de suport a la presa de decisions (intel · ligència artificial) a partir de la detecció i anàlisi de patrons (mineria de dades), ja sigui des d'un punt de vista de recerca base o aplicada. Entre els dominis d'aplicació figuren les ciències de la salut, la telemàtica o l'educació, el sector energètic, entre d'altres. Aquest objectiu és dur a terme a través de les següents línies de recerca:

**Clustering.** Agrupar les dades segons criteris de similitud per identificar patrons. Per exemple: recomanar possibles cerques a partir d'agrupar els usuaris d'Internet segons la seva conducta de navegació (Adeniyi et al., 2014).

**Raonament Basat en Casos / *Case-Based Reasoning* (CBR).** Solucionar nous problemes a partir d'analogies amb problemes prèviament resolts. Per exemple, diagnosticar el càncer a un pacient a partir de les dades de pacients diagnosticats prèviament (el Deen et al., 2013) o diagnosticar errors de modelat de les turbines de vapor (Dendani-hadiby and Khadir, 2014).

**Computació evolutiva.** Conjunt de tècniques especialitzades en la cerca i optimització. S'anomenen tècniques evolutives ja que estan basades en conceptes de l'evolució de les espècies i evolució natural (Goldberg, 1989; Schoenauer and Michalewicz, 1997). S'utilitzen per donar respostes a problemes combinatoris on trobar una solució usant

tècniques de força bruta és difícil o pràcticament impossible en termes temporals i de costos de computació. Per exemple: trobar el conjunt d'accions/regles per conduir un vehicle (Muñoz et al., 2010).

**Sistemes d'aprenentatge híbrids.** Aquests sistemes són una combinació de les diferents tècniques i línies d'investigació citades anteriorment, on la intenció és aprofitar els punts forts de les tècniques que els constitueixen. Per exemple, desenvolupar un sistema per a millorar la predicció del retorn d'estoc i riscs (Barak and Modarres, 2015).

El GRSI és un grup de recerca consolidat per la Generalitat de Catalunya que ha participat en nombrosos projectes d'investigació base i aplicada tant d'àmbit nacional i internacional, ja sigui a través de contractes privats amb empresa o convocatòries públiques. A nivell de recerca base, vaig participar als projectes KEEL-I (04036-03 TIC2002-04036-C05-03), KEEL-II (08386-04 TIN2005-08386-C05-04) i KEEL-III (08386-04 TIN2007-08386-C05-05) els quals estaven emmarcats en la línia de recerca de la Computació Evolutiva. A nivell de recerca aplicada, vaig col · laborar en un projecte per a l'avaluació de les competències en l'àmbit d'enginyeria i arquitectura finançat per l'Agència per a la Qualitat del Sistema Universitari de Catalunya (AQU) (Golobardes and Madrazo, 2009b).

El Grup de Recerca en Sistemes Electronics i Anàlisi de dades (GR-SETAD) és un grup de recerca consolidat per la Generalitat de Catalunya. La motivació del grup és fomentar la col · laboració entre la Universitat i la indústria. Les línies de recerca del grup són: l'adquisició de dades, les telecomunicacions i l'anàlisi de dades o mineria de dades.

Finalment, afegir que la darrera etapa del doctorat he estat vinculat al GR-SETAD, on la recerca s'ha centrat principalment en la línia de recerca de Clustering.

### 1.2.2 Activitat de recerca en HTSI - URL

HTSI és una facultat de recent creació que té com a missió impulsar la docència, la investigació i la difusió del coneixement en els àmbits de la direcció hotelera i la gestió d'empreses turístiques. Sota el marc del projecte *Aristos Campus Mundus* (Calvo-Sotelo, 2010; ACM, 2015), HTSI i la Universitat de Deusto van crear conjuntament un centre distribuït de recerca anomenat Grup de Recerca en Hospitalitat, Turisme i Mobilitat / *Research Group in Hospitality, Tourism and Mobilities* (HTM) l'any 2013. L'objectiu del grup és impactar en el desenvolupament de les activitats del sector per promoure el creixement i l'ocupació intel · ligent, sostenible i integradora a través de dues grans línies de recerca des d'una aproximació de la innovació social i el turisme responsable en consonància amb el marc de l'Horitzó 2020.

**Mobilitat.** Les persones en mobilitat, els visitants o els turistes són consumidors intensius d'espais i temps. El disseny i configuració dels espais i els equipaments és summament important, ja que d'ells dependrà en gran mesura el grau de confort i la satisfacció dels visitants i dels viatgers.

**Hospitalitat.** Des del punt de vista del visitant, el confort està relacionat amb l'hospitalitat, entesa en la seva accepció més àmplia, com acolliment. Per tant, potenciar-la vol dir que no només ens hem de fixar en les infraestructures, també cal incloure estudis d'elements sociològics i psicològics.

Actualment, el grup de recerca està en una etapa inicial de creació i està centrant els esforços en projectes privats amb empreses del sector turístic i hoteler, així com en projectes interns que tenen com a objectiu posicionar els estudis del centre sota el punt de vista de la innovació social i el turisme responsable. En el cas de la recerca realitzada, aquesta s'ha centrat en un projecte intern anomenat *Excellence in Hospitality* que té per objectiu identificar els aspectes que defineixen l'excel · lència en la hospitalitat tenint en compte la globalització i la multiculturalitat.

## 1.3 Objectius

Tot i que la motivació i objectius de la tesi han anat evolucionant a mesura que ha anat canviant el marc de treball iniciat al 2005, aquesta sempre ha tingut una visió: ajudar als experts en el procés de presa de decisions a través de la recerca en tècniques capaces d'extreure coneixement de les dades. Concretament, aquesta tesi s'ha centrat principalment en abordar una problemàtica que es troben els experts a l'hora d'aplicar la metodologia Concept Mapping (Trochim, 1989).

La metodologia del Concept Mapping pretén donar resposta al repte de guiar un grup d'experts en la representació objectiva dels pensaments, idees o conceptes abstractes (Bigné et al., 2002; Trochim, 1989) i s'ha aplicat amb èxit en diferents sectors com en l'educació, els camps d'investigació social o la ciència de gestió entre altres per crear marcs conceptuals basats en aspectes específics (Nabitz et al., 2001). La metodologia defineix sis etapes generals per determinar el mapa conceptual de conceptes interrelacionats (Rosas and Camphausen, 2007a):

1. Donada una temàtica i un concepte/pensament abstracte a definir, es realitza un focus group amb un conjunt d'experts de l'àmbit els quals han de generar idees relacionades amb aquest concepte a través de tècniques de pluja d'idees.

2. Cada expert agrupa les idees en categories en base a la seva similitud i associa un pes de rellevància sota el seu parer.

3. Es representa el coneixement agregat i les relacions entre les idees de tots els experts.

4. Usant l'escalat multidimensional i tècniques clustering (Witten and Frank, 2011) s'identifiquen patrons comuns entre les opinions dels experts. El resultat d'aquesta etapa és un conjunt de $N$ potencials configuracions que els experts han d'avaluar. Una configuració és un resultat de l'algorisme de clustering, el qual està format per un agrupament de grups on cadascun conté un conjunt d'idees inicials de la pluja d'idees inicial.

5. Els experts analitzen i etiqueten els subconceptes de les $N$ configuracions en base al significat que extrapolen dins de cada agrupament.

6. Finalment, seleccionen aquella configuració que representa més fidelment el concepte segons el consens dels experts.

Encara que un dels principals beneficis d'aquest enfocament és la seva flexibilitat i adaptabilitat, la quantitat de dades que cal analitzar dificulta les tasques dels experts ja que (1) la selecció de la millor configuració de l'agrupació no és trivial i (2) aquests han de revisar tots els resultats a través de la premissa subjectiva "té sentit aquesta agrupació?", la qual pot posar en perill l'objectivitat de l'enfocament.

Per tant, l'objectiu de la tesi és definir una estratègia que ajudi als experts a dirigir millor l'avaluació dels resultats del clustering per tal de reduir el volum de dades que han d'analitzar, tot potenciant el consens i l'objectivitat dels experts tenint en compte les valoracions i agrupacions inicials. Aquest objectiu s'ha abordat a través de la participació com a investigador en els projectes dels grups de recerca anteriorment citats, a través dels quals progressivament s'ha anat avançant cap a l'objectiu final. Aquest objectiu s'ha desglossat en els següents subobjectius:

**Subobjectiu 1** Estudi i anàlisi d'estratègies de valoració de resultats de clustering (marc dels projectes KEEL-II i KEEL-III).

**Subobjectiu 2** Proposta d'una estratègia ad hoc a la metodologia dels mapes conceptuals per valorar les agrupacions de subconceptes a partir del consens i perspectiva dels experts (marc del projecte *Excellence in Hospitality*).

**Subobjectiu 3** Avaluació de l'estratègia en un cas real (marc del projecte *Excellence in Hospitality*).

D'altra banda, durant els inicis del doctorat la recerca va estar centrada en ajudar als experts a entendre els resultats en el marc de les tècniques d'extracció de regles i, més concretament, en estudiar com incrementar la seva interpretabilitat tot minimitzant l'impacte en la precisió (marc del projecte KEEL i KEEL-II). Tot i que aquesta línia de recerca es va deixar aturada degut a un canvi de direcció en la recerca de la tesi, és una recerca a tenir present a incorporar en un futur per tal d'ajudar als experts a entendre millor el perquè de les coses tal i com es descriu més endavant.

## 1.4 Estructura de la memòria

La present tesi s'estructura en els següents capítols. Els capítols 2 i 3 emmarquen la recerca dins dels grups de recerca i descriuen les aportacions realitzades en el marc dels projectes KEEL i *Excellence in Hospitality*. El capítol 4 resumeix les conclusions de la recerca realitzada, així com planteja les línies futures. Finalment, els apèndix contenen els acrònims utilitzats i el recull de les publicacions d'aquesta tesi per compendi.

# 2

# Aportacions als projectes KEEL

Els projectes KEEL es centren en l'estudi de tècniques de computació evolutiva així com el desenvolupament d'una eina de codi obert per la comunitat científica i universitària. De les diferents línies del projecte, les meves aportacions s'han centrat en la interpretabilitat de les regles i la valoració de la qualitat dels clústers obtenint resultats d'interès pels projectes.

## 2.1 Marc dels projectes KEEL

KEEL-I (04036-03 TIC2002-04036-C05-03), KEEL-II (08386-04 TIN2005-08386-C05-04) i KEEL-III (08386-04 TIN2007-08386-C05-05) són tres projectes de recerca finançats pel *Ministerio de Educación y Ciencia* (MEC) on van participar els grups de recerca més destacats de les Universitats de l'estat que treballen en computació evolutiva: Universitat de Granada, Universitat de Córdoba, Universitat d'Oviedo i la Universitat Ramon Llull.

L'objectiu d'aquests projectes és investigar en diverses tècniques de l'àmbit de la computació evolutiva i integrar-les en una eina de lliure distribució. Per l'acompliment es contempla:

1. Desenvolupament d'algorismes de mineria de dades evolutius per obtenir nous models d'aprenentatge i millorar els existents en termes de rendiment, equilibri entre exploració i explotació dels resultats i convergència de l'aprenentatge.

2. Implementació i desenvolupament de l'entorn computacional de codi obert sota

llicència de lliure distribució (*GNU General Public License* (GPL)v3) que integri els algorismes .

3. Anàlisi dels algorismes incorporats a KEEL, anàlisi, comparativa i validació de les metodologies i algorismes existents.

A l'octubre del 2015 el projecte inclou 524 algorismes i 908 conjunts de problemes diferents pel seu estudi. Els resultats dels projectes, els resultats científics i l'eina estan disponibles a la Web del projecte http://www.keel.es.

Les meves línies de recerca en aquests projectes han estat la interpretabilitat de les regles i l'avaluació de les solucions generades per les tècniques de clustering, tal com es descriu en els apartats 2.2 i 2.3. Finalment, es conclou amb les conclusions i línies futures.

## 2.2 Interpretabilitat de les regles

Una de les motivacions de la present tesi és donar suport a l'expert en la presa de decisions. Aplicant aquest principi als Sistemes d'Aprenentatge Artificial basats en Algorismes Genètics / *Genetic Based Machine Learning* (GBML), la interpretabilitat de les regles és un factor el qual està condicionat a dos factors: al número de regles generat i a la longitud de la regla. A més a més, el número de regles que genera el sistema serà major o menor depenent de la precisió i del problema. D'altra banda, la longitud de la regla depèn del nombre d'atributs del problema, és a dir, la interpretabilitat del problema ve condicionada per la distribució de les dades, el nombre d'atributs i la precisió requerida del sistema.

Aquest sistemes han de construir regles que cobreixin el màxim espai de cerca amb el mínim error de predicció possible tenint en compte que (1) si hi ha molt poques regles pot haver un increment de l'error de predicció i (2) si hi ha moltes regles poden haver-hi problemes de sobreaprenentatge i, a més a més, la interpretabilitat s'empitjora. Per tant, cal trovar un nombre de regles que sigui un bon compromís

Les aproximacions de GBML més esteses són l'aproximació de Michigan (Holland, 1976) i l'aproximació de Pittsburgh (Carse et al., 1996). L'aproximació de Michigan es caracteritza per evolucionar una única població de regles, on el conjunt representa l'aprenentatge. En canvi, l'aproximació de Pittsburgh evoluciona una població de conjunts de regles, on cada regla representa l'aprenentatge. Sigui quina sigui l'aproximació usada ajustant els paràmetres evolutius es pot aconseguir conjunts de regles més o menys senzills (Bacardit and Garrell, 2003) i interpretables. No obstant, ambdues aproximacions sovint generen gran quantitat de regles, fet que dificulta la interpretació dels resultats.

La recerca en la interpretabilitat de les regles s'ha abordat des de l'aproximació de l'Aprenentatge Iteratiu de Regles / *Iterative Rule Learning* (IRL). Aquesta és una variació de l'aproximació de Michigan que genera un conjunt de regles de mida més petita i, per tant, les regles són més fàcils d'interpretar en comparació amb altres variants.

IRL es caracteritza per explorar l'espai de cerca iterativament (Mitchell, 1997). En cada iteració l'algorisme genètic explora un subconjunt de l'espai de cerca en busca de la

millor regla, posteriorment, descarta els punts coberts per la regla i torna a iterar sobre l'espai de cerca fins que no queda espai per cobrir. Durant aquest procés, l'algorisme en cada iteració guarda la millor regla per ordre. Una vegada finalitza la fase d'entrenament el sistema està preparat per ser explotat. El conjunt de regles resultant s'explota com si d'una llista de decisió(Rivest, 1987) es tractés, en el mateix ordre que s'ha creat. Donat un nou exemple s'explora el conjunt de regles de manera seqüencial fins a trobar una regla que classifiqui l'exemple, finalment, s'usa la regla per inferir la predicció de l'exemple.

Els sistemes IRL creen un conjunt de regles jeràrquic, el qual té un funcionament similar a la llista de decisió. En general, els algorismes IRL són capaços de treballar amb grans volums de dades fent que el conjunt de regles resultant sigui més fàcilment interpretable degut a la seva mida, ja que aquests contenen menys regles. Cada regla $r$ està composada per dues parts:

- Antecedent $A$, el qual representa la condició que cal satisfer per activar la regla.
- Conseqüent $C$ és la inferència que cal aplicar sent $r_k = A \rightarrow C$ on $r_k$ és la regla $k$ del conjunt de regles.

La taula 2.2 mostra a tall d'exemple el resultat d'aplicar un algorisme IRL sobre el problema Iris de l'UCI Repository (A. Asuncion, 2007). Aquest *dataset* és un recull de l'amplada i llargada del petal i l'amplada i llargada del sepal dels espècimens per poder discernir quina és la varietat de la planta Iris: Iris setosa, Iris virgínica i Iris versicolor. La taula 2.2 conté 4 regles jeràrquiques on cada regla defineix un valor mínim i màxim per cadascun dels atributs del problema. Donat un exemple $e$ per predir la varietat on $e = \{6.3, 2.9, 5.6, 1.8\}$ s'explora el conjunt de regles de manera incremental fins a trobar la regla $r$ que classifica cadascun dels atributs, finalment, s'usa $r$ per predir la varietat (en l'exemple donat $e$ iris virgínica).

Taula 2.1: Conjunt de regles jeràrquic resultant per el problema Iris del UCI Repository. Les files són les regles del conjunt, les columnes correspon als atributs del problema i la varietat. Cada regla defineix un valor mínim i màxim per cadascun dels atributs

| | sepal | | petal | | |
|---|---|---|---|---|---|
| Regla | llargada | amplada | llargada | amplada | varietat |
| $R_0$ | 4.30 - 7.90 | 2.00 - 4.40 | 1.00 - 1.90 | 0.10 - 2.50 | setosa |
| $R_1$ | 4.30 - 7.90 | 2.00 - 4.40 | 1.93 - 6.90 | 1.72 - 2.50 | virgínica |
| $R_2$ | 4.30 - 7.90 | 2.00 - 4.40 | 4.85 - 6.90 | 0.10 - 1.70 | virgínica |
| $R_3$ | 4.30 - 7.90 | 2.00 - 4.40 | 1.95 - 4.80 | 0.10 - 1,70 | versicolor |

A l'hora de predir un exemple nouvingut $e$ i un conjunt de regles $H$ de mida $m$ es recórrer $H$ des de $r_0$ fins a $r_i$ on la condició de $r_i$ classifica $e$ i $i < m$.

L'antecedent $A$ tindrà tants cromosomes com atributs tingui el problema d'entrada. El conseqüent $C$ dependrà de l'àmbit de predicció. La predicció és pot fer des de dos punts de vista:

**Predicció de categories.** La classificació consisteix en un conjunt de tècniques de Sistemes d'aprenentatge artificial on els algorismes han d'aprendre a identificar un exemple nouvingut. Per categoritzar els nous exemples el sistema s'entrena/aprèn a partir d'un conjunt de dades característic d'entrenament. Per exemple, discriminar la varietat de flor Iris (*Iris versicolor*, *Iris virginica* o *Iris setosa*) a partir de diferents dades morfològiques de la flor d'Iris (Fisher, 1936; A. Asuncion, 2007).

**Predicció numèrica.** Aquest tipus de sistema ha d'aprendre la funció que descriuen d'un conjunt de variables de l'espai de característiques d'un problema. Donat un conjunt de punts ($X$) les tècniques de Sistemes d'aprenentatge artificial han d'aprendre la funció ($f(..)$) que aquestes defineixen. Una vegada el sistema ha après pot predir la imatge ($f(x')$). A partir de diferents atributs determina el valor d'un habitatge al suburbi de Boston (Belsley et al., 1980; A. Asuncion, 2007).

La recerca realitzada en el camp de la interpretabilitat de les regles s'ha fet en dos tipus de predicció.

### 2.2.1 Interpretabilitat en problemes de predicció de categories

L'algorisme HIDER (Aguilar-ruiz et al., 2000; Aguilar-Ruiz et al., 2003) es caracteritza per ser un sistema IRL (González and Herrera, 1997; Venturini, 1993) per predir categories. La regla manté l'estructura $r_k = A \rightarrow P$ establerta. El conseqüent $P$ indica la categoria que s'inferirà en cas de complir-se les condicions definides a l'antecedent $A$ el qual és un cromosoma que té codificats els atributs del problema. Cal destacar que existeixen diferents tipus de codificacions per cada tipus d'atribut(Aguilar-Ruiz et al., 2007).

Aquesta línia de recerca estudia la capacitat d'aprenentatge que té el HIDER. Es demostra que a l'augmentar la generalització la capacitat d'aprenentatge baixa, degut a que les regles són més genèriques i assumeixen més error. L'avantatge és que es generen menys regles que qualsevol altra aproximació. En cas contrari, al reduir la generalització s'augmenta la precisió, en aquest cas els resultats s'inverteixen: increment del número de regles produïdes i menys error. Analitzant els casos extrems es descobreix la problemàtica: molta generalització introdueix molt d'error, en canvi molta precisió indueix al sobreaprenentatge del problema, a més a més, el cost computacional és proporcional a l'encert/capacitat d'aprenentatge. L'abast de l'estudi inclou diferents tipus de problemes sintètics. Es conclou que el HIDER, algorisme estudiat, genera un conjunt de regles reduït i precís amb resultats significativament robustos, ara bé, cal trobar l'equilibri just entre la generalització i la precisió.

Els resultats es van publicar en l'article "Evolución de modelos jerárquicos de reglas en problemas anidados y no anidados" al congrés *Actas de la primera jornada de algoritmos evolutivos y metaheurísticas (JAEM'07)*. Aquest és un congrés nacional.

### 2.2.2 Interpretabilitat en problemes de predicció numèrica

Durant la recerca es va detectar que no hi havia algorismes de predicció numèrica amb aproximació IRL. Aquesta línia de recerca es centra en l'aplicació dels principis IRL en

l'àmbit de la predicció numèrica.

Es proposa una nova tècnica per aproximar funcions anomenat HIRE-Lin i estudiar-ne el seu comportament respecte altres tècniques competitives. Aquest evoluciona un conjunt de regles jeràrquic, tal i com es proposa a (Rivest, 1987; Aguilar-Ruiz et al., 2003), on cada regla està dividida en un antecedent i un successor. L'antecedent és la condició a complir per a l'activació de la regla, el conseqüent és un regressor lineal calculat aplicant l'estimador d'error quadràtic mig (Glantz and Slinker, 2001; Montgomery and Runger, 2003). L'objectiu és proposar una nova arquitectura per a l'evolució de regressors lineals jeràrquics basat en algorismes genètics i, per tant, volem heretar les capacitats de l'Algorisme Genètic / *Genetic Algorithm* (GA) com la robustesa, la independència del domini, la senzillesa, i la facilitat d'interpretació.

Com a resultat d'aquesta línia de recerca es va publicar un article a la Conferència de Computació Genètica i Evolutiva / *Genetic and Evolutionary Computation Conference* (GECCO) estenent l'enfocament IRL a la predicció numèrica. En l'article s'avaluen aquestes capacitats i en comparació amb enfocaments clàssics per a la regressió. D'altra banda, també es compara els beneficis de l'enfocament del HIRE-Lin, aproximació IRL, enfront d'enfocaments tipus Michigan i Pittsburgh com: Complexitat de l'espai de cerca acotat i alta interpretabilitat dels resultats.

De la mateixa manera que succeeix amb el HIDER, el HIRE-Lin produex conjunts de regles resultants més grans com més gran és la precisió de la solució, en termes de número de regles. Cal destacar que la generalització i la precisió són els dos objectius de HIRE-Lin els quals s'han de maximitzar (Coello et al., 2002). Com més pressió s'exerceix més probabilitat hi ha de caure en el sobreaprenetatge del problema. Finalment, es demostra estadísticament (Demšar, 2006) que el HIRE-Lin és altament competitiu amb altres tècniques de predicció de funcions com Linear LMS (Rustagi, 1994), Fuzzy Wang-Mendel (Wang and Mendel, 1992), GAP (Sánchez et al., 2001) i XCSF (Drugowitsch and Barry, 2008; Wilson, 2002, 1998) en termes de predicció (aprenentatge) i interpretabilitat de les regles.

Els resultats es van publicar en l'article "Hierarchical Evolution of Linear Regressors" al congrés *In Genetic and Evolutionary Computation Conference* (GECCO'07). Aquest és un congrés internacional indexat al ranking Core A de *Computer Science*.

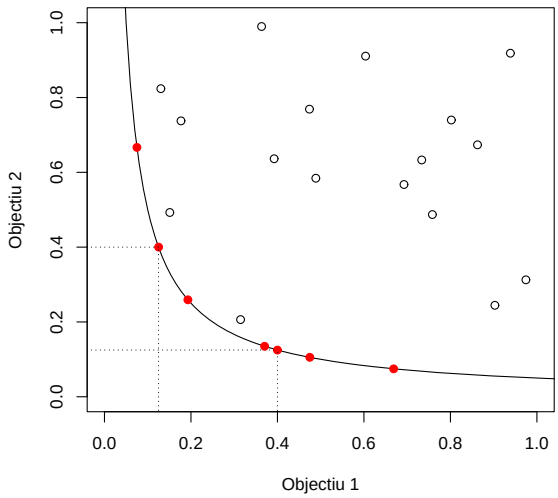
## 2.3 Avaluació de solucions de clustering

Les tècniques de clustering (Duda et al., 2000; Herrera et al., 2010; Kaufman and Rousseeuw, 1990) són tècniques d'agrupament segons tècniques de similitud per identificar patrons. Dos dels seus principals reptes són: definir els criteris de similitud per fer els agrupaments i trobar el nombre òptim de clústers de la solució.

La recerca en aquesta línia de recerca s'ha centrat en estudiar diferents estratègies per avaluar la qualitat dels clústers resultants de les tècniques de clustering multiobjectiu.

Els algorismes de clustering multiobjectiu identifiquen patrons optimitzant diversos objectius simultàniament, les millors solucions en termes d'acompliment d'objectius representen el front de Pareto. La figura 2.1 mostra un exemple d'un espai de solucions

FIGURA 2.1: Resultats del clustering segons dos objectius on cada punt representa una solució. En color vermell el conjunt de Pareto. Valors d'abscisses i ordenades indiquen millors solucions respecte l'objectiu.




valorades segons dos objectius on els punts en vermell identifiquen les millors solucions tenint en compte els dos objectius. Aquests punts determinen el Front de Pareto.

Un dels reptes principals d'aquest àmbit és la selecció de la millor solució del front de Pareto ja que a priori totes les solucions d'aquests són potencialment bones, on a més a més, cadascuna d'elles pot tenir un nombre de clústers diferent. Habitualment, s'usen dues estratègies per seleccionar la millor solució del conjunt de Pareto: considerar l'estructura de Pareto o considerar les característiques morfològiques dels clústers (Garcia-Piquer, 2012). La primera aproximació intenta trobar el colze de Pareto mentre que la segona només té en compte la qualitat de les solucions.

En la primera aproximació, per exemple: donat un problema amb dos objectius disjunts, l'acompliment d'un objectiu perjudica a l'acompliment de l'altre i viceversa, el colze de Pareto esdevé aquell conjunt de solucions on l'assoliment d'ambdós objectius està en equilibri. Només es té en compte l'assoliment d'objectius.

En la segona aproximació només es té en compte la qualitat dels clústers, la qual s'avalua usant els indicadors anomenats índex de validació (Halkidi et al., 2001; Garcia-Piquer, 2012; Gurrutxaga et al., 2011; Handl and Knowles, 2007; Hruschka et al., 2009). Els índexs més utilitzats són:

***Deviation(C)*** (Halkidi et al., 2001) mesura la distància entre els elements d'un clúster i el seu centroide. El rang de valors que pren és [0.. + inf). És preferible obtenir valors propers a 0.

***Connectivity(C)*** (Halkidi et al., 2001) mesura la separació entre els clústers de la

solució. Pren els valor del rang [0.. + inf). Valors alts indiquen major separació entre clústers.

**Davies-Bouldin(C)** (Davies and Bouldin, 1979) avalua els clústers tenint en compte la dispersió, calculant la distància entre les instàncies de cada clúster i el seu centroide. Aquest índex pren valors [0.. + inf). Valors petits indiquen menor dispersió entre els elements dels clústers.

**Dunn(C)** (Dunn, 1974), avalua si els clústers són compactes, penalitzar els clústers amb un gran diàmetre. Aquest índex pren valors [0.. + inf). És preferible valors elevats, aquest indica que els clústers són petits i compactes.

***Silhouette(C)*** (Rousseeuw, 1987; Harrison and Klein, 2007), avalua la cohesió dels grups tenint en compte la distància entre les instàncies de cada clúster. Aquest índex pren valors [−1..1]. És preferible obtenir un valor alt, aquest indica major cohesió i distància entre clústers.

El treball en aquesta línia s'ha fet en el marc d'una línia de recerca del grup on va proposar una tercera estratègia per filtrar les solucions del conjunt de Pareto basada en les estratègies anteriors. L'estratègia contempla un equilibri entre objectius bo i mantenint els índexs de validació dels clústers. Es demostra estadísticament la validesa dels resultats obtinguts fent una comparativa exhaustiva de les tres estratègies amb diversos problemes de diferents repositoris. Els resultats mostren que la precisió i el temps de recuperació es milloren significativament.

Els resultats es van publicar en l'article "Towards a High Solution Retrieval in Multiobjective Clustering. Information Sciences" a la revista *Information Sciences*. Aquesta és una revista indexada al Q1 amb factor d'impacte 3,8.

## 2.4 Conclusions de les aportacions al marc de projectes KEEL

El marc dels projectes KEEL es centra en la computació evolutiva des de tres línies d'actuació important: recerca i desenvolupament d'algorismes d'aquest camp per avaluar, comparar el seu rendiment i desenvolupament de d'una eina de recerca de codi obert pels investigadors. Aquesta tesi s'ha centrat en la interpretabilitat de les regles i l'avaluació de les solucions de clustering. Les aportacions han estat:

- Estudi del comportament de l'IRL. La interpretabilitat de les regles depèn de dos factors: mida del conjunt de regles i la longitud de la regla. L'aproximació IRL explora l'espai de cerca iterativament mentre construeix el conjunt de regles resultant. Aquesta aproximació es caracteritza per crear conjunts de regles de mida menor que altres aproximacions. La regla està constituïda per un antecedent i un conseqüent. L'antecedent és el conjunt de condicions que cal complir per inferir el conseqüent.

- Proposta d'una nova tècnica de predicció numèrica basada en l'aproximació IRL. S'apliquen els principis de l'IRL per proposar una nova tècnica per aproximar funcions. La tècnica es compara amb altres tècniques de predicció numèrica rellevants amb resultats satisfactoris en termes d'ajustament de precisió respecte generalització amb un conjunt de regles més reduït.

- En el marc del grup es va proposar una tercera tècnica per seleccionar els millors resultats d'un algorisme de clustering multiobjectiu. Aquesta tècnica té en compte qualitat, calculada a partir dels principals índexs de validació i l'acompliment dels objectius definits.

Els resultats de la recerca són:

- F. Teixidó-Navarro, and E. Bernadó-Mansilla. Evolución de modelos jerárquicos de reglas en problemas anidados y no anidados. Actas de la primera jornada de algoritmos evolutivos y metaheurísticas (JAEM'07). Publicació en congrés nacional.

- F. Teixidó-Navarro, A. Orriols-Puig, and E. Bernadó-Mansilla. Hierarchical Evolution of Linear Regressors. In Genetic and Evolutionary Computation Conference (GECCO'07), ACM Press, 1413–1420, ISBN 978-1-60558-131-6, 2008. Publicació en congrés Internacional Core A de Computer Science.

- A. Garcia-Piquer, A. Sancho-Asensio, A. Fornells, E. Golobardes, G. Corral and F. Teixidó-Navarro. Towards a High Solution Retrieval in Multiobjective Clustering. Information Sciences, vol 320: 12-25, 2015. Publicació indexada al Q1 amb factor d'impacte 3,8.

# 3

# Aportacions al projecte *Excellence in Hospitality*

Determinar la importància de l'excel · lència en l'hospitalitat és crucial per formar els futurs líders del sector. HTSI ha iniciat una línia de recerca per identificar-los a través de la metodologia dels mapes conceptuals. La meva recerca s'ha centrat en millorar la metodologia. Les millores s'han avaluat en un cas real amb experts del sector de l'hospitalitat obtenint resultats d'interès.

## 3.1 Excel · lència en l'Hospitalitat

Què s'entén per excel · lència de l'hospitalitat? Tot i que la paraula hospitalitat significa el bon acolliment que es fa als estrangers, la literatura de l'àmbit no arriba a definir el terme hospitalitat ja sigui per la falta de consens (Slattery et al., 2002) o perquè sovint es fa una definició limitada d'aquest terme degut a la naturalesa professional del sector (Wood and Brotherton, 2008).

Les persones en mobilitat, els visitants o turistes són consumidors intensius d'espais i temps. Encara que el disseny i la configuració dels espais i els equipaments són summament importants ja que d'ells dependrà en gran mesura el grau de confort i satisfacció dels visitants i dels viatgers, hi ha molts altres aspectes intangibles que intervenen de manera fonamental. Per això, no només cal tenir en compte els aspectes relacionats amb la infraestructura sinó, a més, cal incloure l'estudi dels elements sociològics i psicològics de l'esmentat concepte, així com d'altres aspectes intrínsecs a aquest sector com són la

mobilitat i la internacionalitat. Altres autors suggereixen que el terme hospitalitat ha de fer referència a l'experiència que té un hoste, la qual hauria de ser memorable i anar més enllà en la gestió dels serveis (Hemmington, 2007).

L'objectiu de la línia de recerca de l'excel · lència en l'hospitalitat és doble:

1. Definir què significa l'excel · lència en la Hospitalitat des del punt de vista dels visitants i turistes i la indústria.

2. Identificar els factors clau que contribueixen al fet que l'hospitalitat sigui excel · lent i causin al visitant una experiència memorable.

La consecució d'ambdós objectius ha de permetre anticipar a HTSI a les necessitats del sector per ser capaços de formar els millors professionals que en un futur han de liderar el sector (Vila et al., 2012). Per fer-ho, HTSI ha creat un projecte intern anomenat *Excellence in Hospitality* que es centra en definir aquest concepte a través de l'aplicació de la metodologia del Concept Mapping, on cal definir aquest concepte tant des del nivell global vers la globalitat i la multiculturalitat com des d'un nivell local tenint en compte les particularitats de cada zona.

## 3.2 Representació de les idees en mapes conceptuals

La metodologia dels mapes conceptuals permet obtenir una representació objectiva de pensaments, idees i/o intuïcions d'un grup d'experts en l'àmbit d'estudi (Trochim, 1989). Això s'aconsegueix a través d'un procediment metodològic que integra tècniques qualitatives i quantitatives, les quals inclouen des de la gestió del coneixement d'un grup d'experts fins a l'anàlisi multivariat i la interpretació dels resultants. Aquest procés contempla els passos següents: (vegeu la figura 3.1)

1. **Preparació de l'estudi.** Donat un concepte abstracte a definir, determina l'àmbit d'estudi i fa la selecció del grup d'experts en l'àmbit. El grup d'experts ha de ser heterogeni per assegurar que els resultats siguin representatius (Harrison and Klein, 2007).

2. **Generació d'idees.** S'invita als experts a una sessió de pluja d'idees per a la generació de les idees que millor descriuen l'àmbit d'estudi. Sovint es convida a un expert en dinàmiques de grup per liderar el procés (Bigné et al., 2002; Calvo et al., 2006).

3. **Estructuració dels conceptes.** Cada expert valora entre [1..5] cada idea segons el grau d'afinitat a l'àmbit d'estudi (Trochim, 1989; Travé-Massuyès et al., 2004, 2005) i, posteriorment, cada expert agrupa les idees segons el seu criteri. El darrer pas consisteix en agregar la informació dels diferents experts per tenir una visió global de les perspectives.

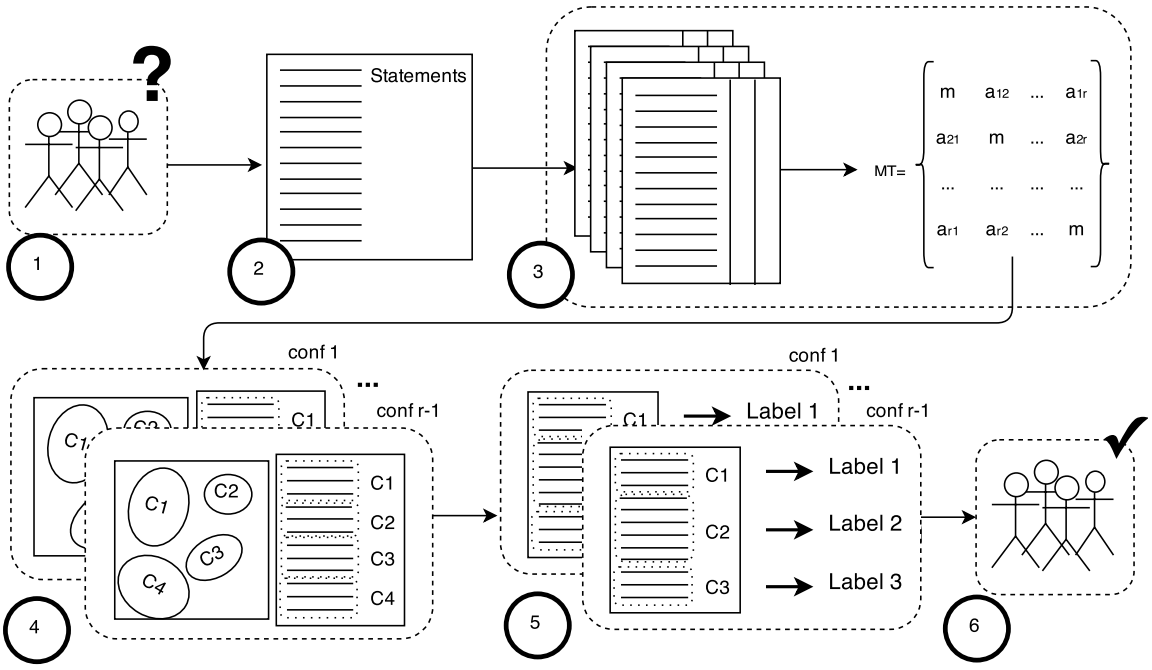
4. **Representació de les idees.** Aquesta fase es realitza en dues etapes a partir d'aplicar el Escalament multidimensional / *Multidimensional Scaling* (MDS) (Borg

and Groenen, 1997) i tècniques de clustering (Jain, 2010). D'una banda, el MDS fa referència al conjunt de tècniques estadístiques utilitzades habitualment en màrqueting i ciències socials per la visualització i exploració de dades amb la finalitat d'identificar preferències i percepcions dels enquestats i poder-ho representar en un diagrama visual. Tot i que quantes més dimensiones hi hagi el resultat serà més fiable, el resultat serà més difícil d'interpretar i per això és habitual treballar en 2 dimensiones (Green, 1975). D'altra banda, les tècniques de clustering permeten identificar les agrupacions d'idees que els experts consideren que estan més relacionades. El resultat de la fase són totes les combinacions de mapes conceptuals resultants.

5. **Interpretació dels mapes de conceptes.** Els experts etiqueten els agrupaments de cada solució generada en el pas anterior.

6. **Ús dels mapes de conceptes.** Decidir per consens de la millor solució i utilització del mapa conceptual com a definició de l'àmbit d'estudi.

FIGURA 3.1: Metodologia del concept mapping on es mostra el detall de l'aportació a la metodologia. Concretament, en el pas 4 i 5 s'apliquen tècniques de validació als resultats per valorar objectivament els resultats.




Un dels reptes que han d'afrontar els experts durant aquesta metodologia és avaluar tots els resultats del pas 4.

La metodologia recomana que es generin prop d'un centenar de conceptes ($n$) (Trochim, 1993) durant la sessió de pluja d'idees (pas 2). Aquest fet provoca que hi hagi la possibilitat

de generar $n$ agrupacions diferents, des d'un grup que inclou tots els conceptes fins a $n$ grups on cada grup inclou un únic concepte. Durant aquest procés s'apliquen tècniques de *clustering* sobre els resultats del MDS. Els agrupaments (solucions candidates) s'envien als experts per a què les etiquetin i decideixin quina entre elles defineix millor l'àmbit d'estudi. Així, doncs, el concept mapping defineix un procediment metodològic que integra tècniques qualitatives i quantitatives per determinar un mapa de conceptes i les relacions que aquests guarden entre si (Rosas and Camphausen, 2007b).

Es proposa aplicar per cada mapa conceptual la tècnica del consens (Bigné et al., 2002). La figura 3.1 il · lustra la metodologia i l'aportació realitzada. La tècnica a partir de l'entropia permet determinar quins són aquells clústers on hi ha conceptes en consens. Per descartar solucions es tenen en compte les puntuacions que prèviament han valorat cadascun dels experts. Les puntuacions indiquen el grau de rellevància del concepte pel cas d'estudi. Finalment, l'índex global de consens (GIC) fa el recompte del nombre de vegades que cada concepte apareix en un clúster en consens. Aplicar la tècnica del consens i l'índex global de consens permet:

1. Descartar clústers sense consens, basant-se en les puntuacions prèvies a l'anàlisi dels experts, aquest punt fa que es puguin descartar grups sense consens dels agrupaments candidats.

2. Havent descartat grups sense consens, facilita la decisió dels experts centrant la seva atenció a agrupaments on els grups són diferents però consensuats.

Ara bé, la dificultat pels experts en la metodologia original radica en decidir quin és el mapa conceptual entre l'espai de solucions que millor descriu el problema formulat. La variació entre els agrupaments és poca i la decisió entre els candidats difícil.

## 3.3 Promoció del consens en la metodologia del concept mapping: una aplicació al sector de l'hospitalitat

Amb la intenció d'esclarir què significa excel · lència en hospitalitat es segueix el procediment de la il · lustració 3.1. Per començar es va convocar a un grup de 11 experts els quals representen la indústria de la hospitalitat a la ciutat de Barcelona. Tots els integrants són directius amb més de 10 anys d'experiència als quals se'ls va formular la següent pregunta: "des del teu punt de vista quins són els principals factors que descriuen l'hospitalitat?" els quals van generar 100 idees. Seguidament, cada expert va valorar entre [1..5] cada concepte i va agrupar les idees. Seguidament, es crea la matriu d'agregació la qual representa la valoració de les idees i la relació entre aquestes.

Una vegada es varen disposar de les dades es va aplicar el MDS de dues dimensions, per posteriorment aplicar l'algorisme de clustering de Ward (Ward, 1963). En aquesta fase es varen generar les solucions des de 1 clúster fins a 99 clústers, és a dir, es varen generar 99 solucions potencials.

FIGURA 3.2: Valors dels índexs qualitatius aplicats a les 99 configuracions del cas d'estudi. Valors alts pels índexs Silhouette i Dunn i valors baixos per Davies-Bouldin indiquen millor qualitat.

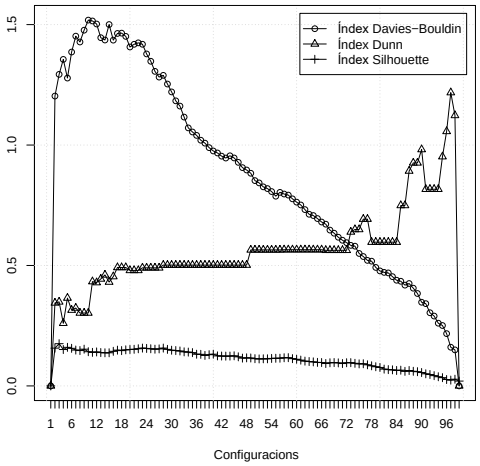
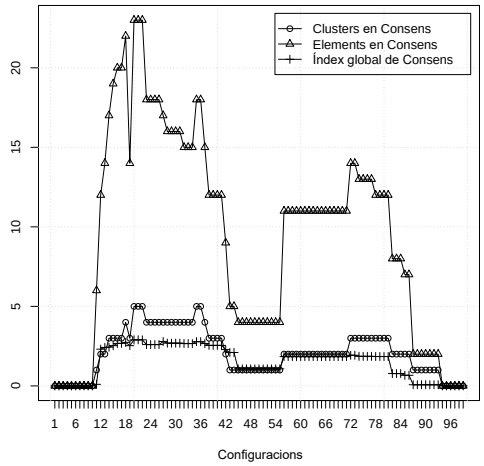



FIGURA 3.3: Valors per l'índex global de consens, així com el número d'elements i clusters per cada configuració. Valors alts per l'índex global de consens indica millor resultat.




Els dos últims passos, el 5 i el 6, estan focalitzats en l'etiquetat i la correctesa de la solució. S'apliquen dues estratègies diferents:

1. Els índexs quantitatius, descrits al punt 2.3. Serveixen per donar una mesura quantitativa de la qualitat dels clústers d'una solució. Només té en compte la geometria. Vegeu la figura 3.2.

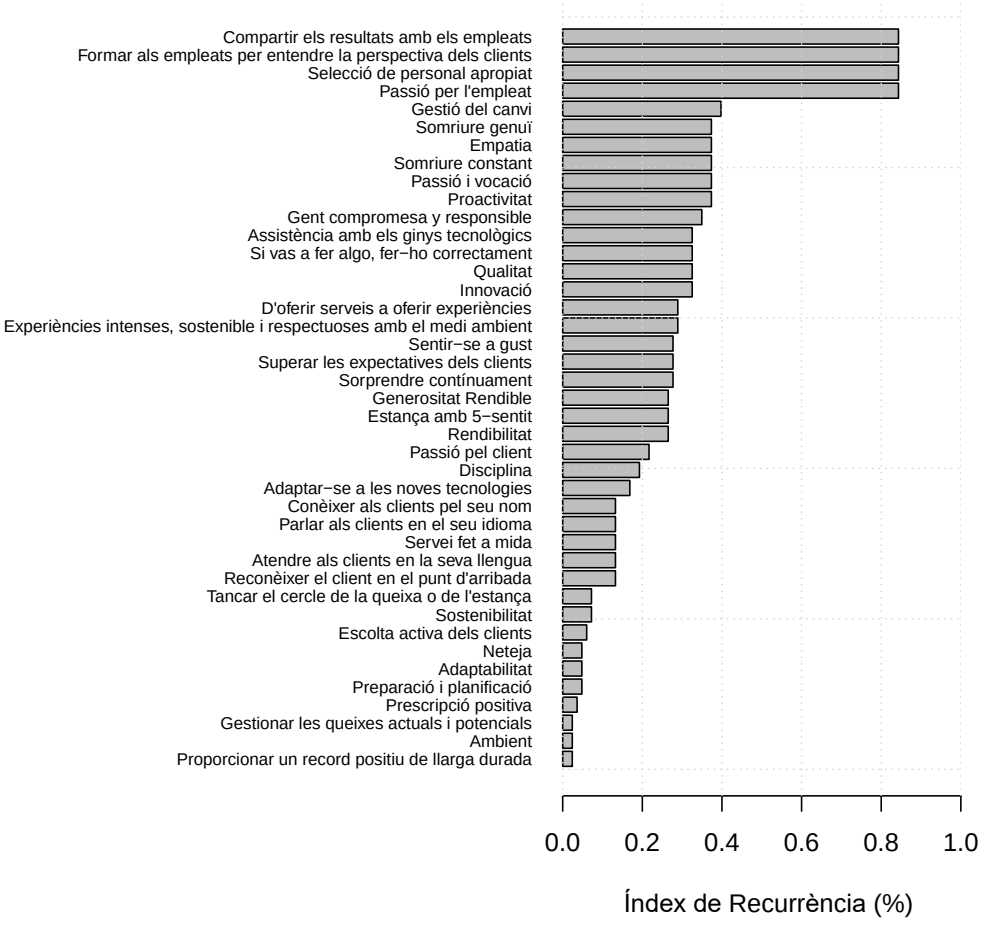
2. L'índex global de consens. En el cas d'estudi es generen 99 solucions potencials, per cada solució es calcula la tècnica del consens, és a dir, es valora cadascun dels clústers si hi ha consens entre les idees agrupades o no. Si no hi ha consens es descarta el clúster. L'índex global de consens es calcula a partir del número de vegades que una idea apareix en un clúster amb consens i el grau de consens del clúster. La figura 3.3 il · lustra els resultats de l'estratègia.

Quan es tenen en compte els índexs de validació (Davies-Bouldin, Dunn i Silhouette) sense tenir en compte cap més detall del domini. Els índexs de Davies-Bouldin i Dunn promouen que els millors resultats es trobin a nombres elevats de clústers, al contrari del que passa amb l'índex Silhouette. En aquest sentit, ninguna de les dues possibilitats és determinant, no es poden escollir solucions amb pocs clústers i totes les idees, ni una solució amb clústers de dues idees. És per aquest motiu que els índexs qualitatius no són suficients per acotar l'espai de solucions.

Aplicar l'índex global de consens permet eliminar de cadascuna de les solucions els clústers sense consens. Els clústers eliminats juntament amb els seus elements incorporen soroll i incertesa a la solució, doncs, es promou els clústers amb idees on els experts estan

d'acord. El número de clústers es redueix dràsticament. Tal com il · lustra la figura 3.3 la solució amb 20 clústers és la millor en termes de qualitat i consens.

Gràcies a les aportacions fetes es va poder determinar que l'excel · lència en l'hospitalitat influeixen els següents factors:




## 3.4 Conclusions de l'aportació a l'excel · lència l'hospitalitat

La metodologia del concept mapping permet, a partir d'una col · lecció de conceptes, generar múltiples mapes conceptuals, cada mapa conceptual agrupa els conceptes en diferents agrupaments, seguidament els experts han de decidir quin d'entre els mapes conceptuals descriu millor, segons la seva expertesa i criteri, l'àmbit d'estudi. La decisió no és trivial ja que cal que hi hagi consens entre els experts. Mitjançant l'índex global de consens s'aconsegueix reduir dràsticament el conjunt de solucions que han d'avaluar els experts sense perjudicar la qualitat de la solució. L'aportació realitzada permet agilitzar

la metodologia ja que permet escollir un nombre determinat de solucions a analitzar pels experts. Els resultats obtinguts permeten continuar la línia de treball de diferents maneres:

- Aplicació de l'índex global de consens en altres tècniques de clustering. Havent demostrat l'eficàcia d'aquest a l'hora de reduir el ventall de possibles solucions caldria estudiar el comportament en altres tècniques.
- Ampliar l'abast de l'estudi a experts de regions geogràfiques diferents. Tot i que l'objectiu de l'aportació és difondre la validesa de la tècnica per aquest estudi, permet ampliar el cas d'estudi. En l'aportació es selecciona els directius dels hotels de 5 estrelles de Barcelona. Permetent definir què s'entén per excel · lència en hospitalitat entre aquest col · lectiu. L'estudi obre una porta a la participació de membres d'altres tipus d'estança i altres regions geogràfiques.

# 4 Cloenda

Els resultats obtinguts durant el desenvolupament d'aquesta tesi han satisfet l'objectiu plantejat. Els resultats de la recerca han produït publicacions d'àmbit nacional i internacional.

## 4.1 Conclusions

La present tesi doctoral s'ha desenvolupat en el marc de dos centres de la URL on s'ha fet i aplicat recerca de l'àmbit de la mineria de dades des d'òptiques diferents però amb una mateixa fita: oferir a l'expert d'un conjunt d'eines per facilitar la presa de decisions.

En el Campus La Salle s'ha format part com a investigador al GRSI i al GR-SETAD on s'ha fet recerca en el marc dels projectes KEEL-I, KEEL-II i KEEL-III, tots ells finançats pel MEC. Aquests projectes tenien com a motivació principal la recerca base en tècniques de la branca de la computació evolutiva per desenvolupar una eina de codi lliure per a la comunitat científica i educativa per tal d'estudiar les tècniques de mineria de dades vigents, especialment aquelles basades en computació evolutiva i incorporar-ne de noves. Respecte la meva recerca, aquesta es va centrar en tècniques d'extracció i explicació de la relació entre les dades d'aproximacions basades en sistemes de regles i clustering fent èmfasi a la interpretabilitat, fiabilitat i error. A més a més, es va tenir l'oportunitat de participar en un projecte per definir una guia d'avaluació de les competències dels àmbits d'enginyeria i arquitectura, on es van aplicar tècniques de mineria de dades per analitzar dades (Golobardes and Madrazo, 2009b,a). D'altra banda, a HTSI s'ha format part del grup de recerca HTM, on la recerca s'ha centrat en aplicar els coneixements adquirits als altres grups de recerca per tal de millorar la metodologia

del mapes conceptuals per tal d'identificar solucions potencialment interessants pels experts en base al consens dels experts, així com avaluar la millora en el marc del projecte *Excellence in Hospitality*. Per tant, durant la primera etapa, la recerca es centra en la recerca base d'algorísmica en computació evolutiva i, durant la segona fase, en l'aplicació de les tècniques adquirides.

Una vegada revisades les aportacions descrites als capítols anteriors, es pot concloure que els objectius de la tesi s'han assolit satisfactòriament:

**Subobjectiu 1** Estudi i anàlisi d'estratègies de valoració de resultats de clustering (marc dels projectes KEEL-II i KEEL-III). La recerca realitzada va permetre conèixer els principals índexs de validació per valorar els resultats de les estratègies de clustering en el marc de la línia de recerca del GRSI de com seleccionar les millors solucions de clustering resultants del clustering multiobjectiu basat en algoritmes genètics (Garcia-Piquer et al., 2015). Aquesta etapa va permetre establir els fonaments per abordar el subobjectiu següent.

**Subobjectiu 2** Proposta d'una estratègia ad hoc a la metodologia dels mapes conceptuals per valorar les agrupacions de subconceptes a partir del consens i perspectiva dels experts (marc del projecte *Excellence in Hospitality*). L'aportació final és l'índex global de consens (GIC), aquest índex s'incorpora a la metodologia dels mapes conceptuals per ajudar els experts a seleccionar l'agrupació més adequada. Les característiques principals d'aquests indicadors són: l'objectivitat i el consens. Per tant, el procés de descobriment de coneixement està millorat dràsticament perquè els experts han de centrar-se només en configuracions útils caracteritzades per contenir les idees en què els experts estan d'acord són similars i amb la mateixa rellevància. Aquest índex es basa en tècniques de raonament qualitatiu i el concepte d'entropia (Shannon, 1948). Raonament qualitatiu és una subàrea de la intel · ligència artificial que busca comprendre i explicar les avaluacions no numèriques dels éssers humans i també permet gestionar amb dades no numèriques preservar el principi de rellevància, és a dir, cada variable pot ser apreciada amb el nivell de precisió requerit (Travé-Massuyès et al., 2004, 2005).

**Subobjectiu 3** Avaluació de l'estratègia en un cas real (marc del projecte *Excellence in Hospitality*). Finalment, GIC s'avalua amb èxit i es compara respecte altres enfocaments per abordar un dels reptes del sector turístic: "quins són els principals factors que condueixen a l'excel · lència en hospitalitat?" (Fornells et al., 2015).

Al marge d'aquests tres subobjectius i com a resultat de la recerca iniciada a la primera etapa de la tesi, es va realitzar recerca base per millorar la interpretabilitat de les regles generades pels sistemes basats en l'aproximació IRL. Aquesta aproximació extreu la millor regla durant la fase d'entrenament iterativament i en construeix un conjunt de regles jeràrquic de manera seqüencial. El resultat de la recerca era conèixer el comportament del sistema HIDER i proposar un nou sistema IRL per fer predicció numèrica, el qual era capaç de ser competitiu amb els sistemes referents de la literatura amb la característica de poder ajustar el llindar precisió-generalització per tal d'obtenir

uns resultats segons els requeriments dels experts i que els ajudessin a entendre millor el perquè de les coses (Teixidó-Navarro and Bernadó-Mansilla, 2007; Teixidó-Navarro et al., 2008).

Resumint, els resultats de la recerca realitzada han generat les següents publicacions:

- F. Teixidó-Navarro, A. Orriols-Puig, and E. Bernadó-Mansilla. Hierarchical Evolution of Linear Regressors. In Genetic and Evolutionary Computation Conference (GECCO'07), ACM Press, 1413–1420, ISBN 978-1-60558-131-6, 2008. Publicació en congrés Internacional Core A de Computer Science.

- A. Garcia-Piquer, A. Sancho-Asensio, A. Fornells, E. Golobardes, G. Corral and F. Teixidó-Navarro. Towards a High Solution Retrieval in Multiobjective Clustering. Information Sciences, vol 320: 12-25, 2015. Publicació indexada al Q1 amb factor d'impacte 3,8.

- A. Fornells, Z. Rodrigo, R. Santomà, X. Rovira, M. Sanchez, F. Teixidó-Navarro and E. Golobardes. Promoting consensus in the concept mapping methodology: An application in the hospitality sector. Pattern Recognition Letters, doi:10.1016/j.patrec.2015.05.013, 2015. Publicació indexada al Q3 amb factor d'impacte 1,062.

- F. Teixidó-Navarro, and E. Bernadó-Mansilla. Evolución de modelos jerárquicos de reglas en problemas anidados y no anidados. Actas de la primera jornada de algoritmos evolutivos y metaheurísticas (JAEM'07). Publicació en congrés nacional.

- Golobardes, E. and Madrazo, L. (2009b). Guia per a l'avaluació de competències en l'àrea d'Enginyeria i Arquitectura. Guies d'avaluació de competències. AQU Catalunya, Barcelona. Participació en un capítol del llibre.

## 4.2 Línies de futur

La recerca de futur apunta en dues direccions íntimament relacionades entre elles. D'una banda, els primers resultats del treball realitzat en el projecte de *Excellence in Hospitality* han determinat un conjunt de factors clau que descriuen l'hospitalitat pels directius principals del sector turístic de Barcelona durant el mes de gener de 2014. Ara bé, la hipòtesi és que el concepte excel · lència en l'hospitalitat està influenciat pel context generacional, el temps, factors socials, nivell econòmic i expectatives, entre d'altres. Actualment, l'estudi s'està realitzant en altres ubicacions geogràfiques com Hong Kong Polytechnic University (Japó), University of San Francisco (EUA), Universidad de Deusto Campus de Donostia/San Sebastián (País Vasc), Universidad de Valparaiso (Xile) i a la Universidad Católica (Uruguay) per tal de veure l'impacte multicultural.

En aquest context la recerca s'ha realitzat usant la metodologia descrita en la memòria, en aquesta l'origen de les dades són els grups de treball de l'estudi. Una alternativa seria complementar l'origen de les dades, sobretot per analitzar l'impacte generacional, local

i temporal, poden ser les dades provinents de xarxes socials. En aquest sentit caldria implementar solucions tenint en compte conceptes del *Big Data* i anàlisi semàntic.

D'altra banda, l'altra línia de recerca està enfocada en introduir noves millores en la metodologia dels mapes conceptuals per tal d'ajudar als experts a prendre decisions amb informació de més qualitat. Algunes d'aquestes estratègies van des de la generació d'explicacions del motiu de les agrupacions, com s'ha fet en altres treballs del GRSI, on s'ha aplicat l'operador d'antiunificació (Fornells et al., 2008), l'aplicació de la tècnica del terme lingüístic difús vacil · lant per calcular la distància entre les opinions dels experts (Agell et al., 2015) o introduir més flexibilitat en les valoracions dels experts, els experts podrien etiquetar els conceptes amb més d'una etiqueta si escaigués.

Una altra via a tenir en compte per millorar l'estratègia es la optimització de la mètrica del consens. De les diferents tècniques de cerca a tenir en compte a priori podrien ser els algorismes genètics.

# Bibliografia


A. Asuncion, D. N. (2007). UCI machine learning repository.

ACM (2015). *Proyecto Aristus Campus Mundus 2015. Campus de excelencia*. Ministerio de Educación.

Adeniyi, D., Wei, Z., and Yongquan, Y. (2014). Automated web usage data mining and recommendation system using k-nearest neighbor (knn) classification method. *Applied Computing and Informatics*, pages –.

Agell, N., Sánchez, M., Prats, F., and Ruiz, F. J. (2015). Computing distances between decision makers using hesitant fuzzy linguistic term set. *Proceedings of the 18th Catalonian Conference on Artificial Intelligence*, page n.a.

Aguilar-Ruiz, J. S., Giráldez, R., and Santos, J. C. R. (2007). Natural encoding for evolutionary supervised learning. *IEEE Trans. Evolutionary Computation*, 11(4):466–479.

Aguilar-ruiz, J. S., Santos, J. C. R., and Toro, M. (2000). Data set editing by ordered projection. In *European Conference on Artificial Intelligence*, pages 251–255.

Aguilar-Ruiz, J. S., Santos, J. C. R., and Toro, M. (2003). Evolutionary learning of hierarchical decision rules. *IEEE Trans. on Systems, Man and Cybernetics, B*, 33(2):324–331.

Bacardit, J. and Garrell, J. (2003). Bloat control and generalization pressure using the minimum description length principle for a Pittsburgh approach Learning Classifier System. In *Learning Classifier Systems, Revised Selected Papers of the International Workshop on Learning Classifier Systems 2003-2005*, volume 4399 of *Lecture Notes in Computer Science*, pages 59–79. Springer.

Barak, S. and Modarres, M. (2015). Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems with Applications*, 42(3):1325–1339.

Belsley, D., Kuh, E., and Welsch, R. (1980). *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*. John Wiley, New York.

Bigné, J. E., Manzano, J. A., Küster, I., and Vila, N. (2002). The concept mapping approach in marketing: an application in the travel agencies sector. *Qualitative Market Research: An International Journal*, 5(2):87–95.

Borg, I. and Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer.

Calvo, A., Criado, F., and Periáñez, R. (2006). *Desarrollo de un instrumento para evaluar la idoneidad de los planes docentes: una aplicación a la diplomatura en turismo. Presented in Decisiones basadas en el conocimiento y en el papel social de la empresa*. Academia Europea de Dirección y Economía de la Empresa, Palma de Mallorca.

Calvo-Sotelo, P. C. (2010). *Campus de excelencia internacional. Convocatoria CEI 2010. Presentación de los proyectos seleccionados*. Ministerio de Educación.

Carse, B., Fogarty, T., and Munro, A. (1996). Evolving fuzzy rule based controllers using genetic algorithms. *Fuzzy Sets and Systems*, 80:273–293.

Castells, M. and Martínez, C. (2001). *La era de la información: economía, sociedad y cultura*. La era de la información. Alianza, Madrid.

Coello, C. C., Veldhuizen, D. V., and Lamont, G. (2002). *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers.

Davies, D. and Bouldin, D. (1979). A cluster separation measure. In *IEEE Transactions on Pattern Analysis and Machine Learning*, volume 4, pages 224–227.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *journal of Machine Learning Research*, 7:1–30.

Dendani-hadiby, N. and Khadir, M. (2014). Comparative analysis of case retrieval implementation for knowledge intensive cbr application. In *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*, pages 1107–1114.

Drugowitsch, J. and Barry, A. (2008). A Formal Framework and Extensions for Function Approximation in Learning Classifier Systems. *Machine Learning*, 70(1):45–88.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern classification*. John Wiley and Sons, Inc., New York.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. In *Journal of Cybernetics*, volume 4, pages 95–104.

el Deen, D. A. S., Moawad, I. F., and Khalifa, M. E. (2013). Article: A breast cancer diagnosis system using hybrid case-based approach. *International Journal of Computer Applications*, 72(23):14–20. Full text available.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, Menlo Park, CA, USA.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–188.

Fornells, A., Armengol, E., and Golobardes, E. (2008). Retrieval based on self-explicative memories. In *9th European Conference on Case-Based Reasoning,*, volume 5239 of *LNAI, Springer-Verlag*, pages 210–224. Lecture Notes in Computer Sciencie, Springuer-Verlag.

Fornells, A., Rodrigo, Z., Rovira, X., Sánchez, M., Santomà, R., Teixidó-Navarro, F., and Golobardes, E. (2015). Promoting consensus in the concept mapping methodology: An application in the hospitality sector. *Pattern Recognition Letters*.

Fu, L., Sun, D., and Rilett, L. R. (2006). Heuristic shortest path algorithms for transportation applications: State of the art. *Comput. Oper. Res.*, 33(11):3324–3343.

Garcia-Piquer, A. (2012). *Facing-up challenges of multiobjective clustering based on evolutionary algorithms: Representations, scalability and retrieval solutions*. PhD thesis, Research Group in Intelligent Systems, Campus LaSalle, Universitat Ramon Llull.

Garcia-Piquer, A., Sancho-Asensio, A., Fornells, A., Golobardes, E., Corral, G., and Teixidó-Navarro, F. (2015). Toward high performance solution retrieval in multiobjective clustering. *Information Sciences*.

Glantz, S. and Slinker, B. (2001). *Primer of Applied Regression & Analysis of Variance*. McGraw-Hill.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Company, Inc.

Golobardes, E. and Madrazo, L. (2009a). *Guía para la evaluación de competencias en el área de Ingeniería y Arquitectura*. Guías de evaluación de competencias. AQU Catalunya, Barcelona.

Golobardes, E. and Madrazo, L. (2009b). *Guia per a l'avaluació de competències en l'àrea d'Enginyeria i Arquitectura*. Guies d'avaluació de competències. AQU Catalunya, Barcelona.

González, A. and Herrera, F. (1997). Multi-stage Genetic Fuzzy Systems Based on the Iterative Rule Learning Approach. *Mathware & Soft Computing*, 4:233–249.

Gurrutxaga, I., Muguerza, J., Arbelaitz, O., Pérez, J. M., and Martín, J. I. (2011). Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters*, 32(3):505 – 515.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *journal of Intelligent Information Systems*, 17:107–145.

Handl, J. and Knowles, J. (2007). An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*, 1(1):56–76.

Harrison, D. A. and Klein, K. J. (2007). What's the difference? diversity constructs as separation, variety, or disparity in organizations. *Academy of management review*, 32(4):1199–1228.

Hemmington, N. (2007). From wervice to experience: Understanding and defining the hospitality business. *The Service Industries Journal*, 27(6):747–755.

Herrera, F., Carmona, C., González, P., and del Jesus, M. (2010). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, pages 1–31.

Holland, J. H. (1976). Adaptation. In Rosen, R. and Snell, F., editors, *Progress in Theoretical Biology*, volume 4, pages 263–293. Academic Press.

Hruschka, E. R., Campello, R. J. G. B., Freitas, A. A., and de Carvalho, A. C. P. L. F. (2009). A survey of evolutionary algorithms for clustering. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 39(2):133–155.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Jurado, S., Nebot, n., Mugica, F., and Avellana, N. (2015). Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. *Energy*, 86(C):276–291.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data : an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Wiley, New York. A Wiley-Interscience publication.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8 – 17.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Montgomery, D. and Runger, G. (2003). *Applied Statistics and Probability for Engineers*. John Wiley & Sons.

Muñoz, J., Gutiérrez, G., and Sanchis, A. (2010). A human-like torcs controller for the simulated car racing championship. In *Proceedings 2010 IEEE Conference on Computational Intelligence and Games*, pages 473–480.

Nabitz, U., Severens, P., Brink, W. V. D., and Jansen, P. (2001). Improving the efqm model: An empirical study on model development and theory building using concept mapping. *Total Quality Management*, 12(1):69–81.

Negnevitsky, M. (2001). *Artificial Intelligence: A Guide to Intelligent Systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.

Nilsson, N. J. (1998). *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rivest, R. L. (1987). Learning decision lists. *Machine Learning*, 2(3):229–246.

Rosas, S. R. and Camphausen, L. C. (2007a). The use of concept mapping for scale development and validation in evaluation. *Evaluation and Program Planning*, 30(2):125 – 135.

Rosas, S. R. and Camphausen, L. C. (2007b). The use of concept mapping for scale development and validation in evaluation. *Evaluation and Program Planning*, 30(2):125–135.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In *Journal of Computational and Applied Mathematics*, volume 20, pages 53–65.

Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition.

Rustagi, J. (1994). *Optimization Techniques in Statistics*. Academic Press.

Sánchez, L., Couso, I., and Corrales, J. A. (2001). Combining GP operators with SA search to evolve fuzzy rule based classifiers. *Information Sciences*, 136(1–4):175–191.

Schoenauer, M. and Michalewicz, Z. (1997). Evolutionary computation.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Slattery, P. et al. (2002). Finding the hospitality industry. *Journal of Hospitality, Leisure, Sport and Tourism Education*, 1(1):19–28.

Sung, M. and Ko, Y. (2015). Machine-learning-integrated load scheduling for reduced peak power demand. *Consumer Electronics, IEEE Transactions on*, 61(2):167–174.

Teixidó-Navarro, F. and Bernadó-Mansilla, E. (2007). Evolución de modelos jerárquicos de reglas en problemas anidados y no anidados. In *Actas de la primera jornada de algoritmos evolutivos y metaheurísticas (JAEM'07)*.

Teixidó-Navarro, F., Orriols-Puig, A., and Bernadó-Mansilla, E. (2008). Hierarchical evolution of linear regressors. In *Genetic and Evolutionary Computation Conference, GECCO 2008, Proceedings, Atlanta, GA, USA, July 12-16, 2008*, pages 1413–1420.

Travé-Massuyès, L., Ironi, L., and Dague, P. (2004). Mathematical foundations of qualitative reasoning. *AI Magazine*, 24(4):91–106.

Travé-Massuyès, L., Prats, F., Sánchez, M., and Agell, N. (2005). Relative and absolute order-of-magnitude models unified. *Annals of Mathematics and Artificial Intelligence*, 45(3-4):323–341.

Trochim, W. (1993). The reliability of concept mapping. In *Annual Conference of the American Evaluation Association, Dallas, Texas*.

Trochim, W. M. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning*, 12(1):1 – 16. Special Issue: Concept Mapping for Evaluation and Planning.

Venturini, G. (1993). SIA: A Supervised Inductive Algorithm with Genetic Search for Learning Attribute Based Concepts. In *Proc. European Conf. Machine Learning*, pages 280–296.

Vila, M., Rovira, X., Costa, G., and Santoma, R. (2012). Combining research techniques to improve quality service in hospitality. *Quality & Quantity: International Journal of Methodology*, 46(3):795–812.

Wang, L. and Mendel, J. (1992). Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man and Cybernetics*, 22(6):1414–1427.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Wilson, S. (2002). Classifiers that approximate functions. *Journal of Natural Computing*, 1(2–3):211–234.

Wilson, S. W. (1998). Generalization in the XCS Classifier System. In Koza, J., Banzhaf, W., Chellapilla, K., Deb, K., Dorigo, M., Fogel, D., Garzon, M., Goldberg, D., Iba, H., and Riolo, R., editors, *Genetic Programming: Proc. of the Third Annual Conference*, pages 665–674. San Francisco, CA: Morgan Kaufmann.

Witten, I. H. and Frank, E. (2011). *Data mining: Practical machine learning tools and techniques with java implementations*. Morgan Kaufmann, San Francisco, 3th edition.

Wood, R. and Brotherton, B. (2008). *The SAGE Handbook of Hospitality Management*. SAGE Publications.

# A
# Acrònims

**AGAUR** Agència de Gestió d'Ajuts Universitaris i de Recerca

**AQU** Agència per a la Qualitat del Sistema Universitari de Catalunya

**BI** intel · ligència Empresarial / *Business Intelligence*

**BD** *Big Data*

**CBR** Raonament Basat en Casos / *Case-Based Reasoning*

**CEDI** Congrés espanyol d'Informàtica / *Congreso Español De Informática*

**CM** Mapes de Conceptes / *Concept Mapping*

**CSR** Responsabilitat Social Corporativa / *Corporate Social Responsibility*

**DEA** Diploma d'Estudis Avançats

**EEES** Espai Europeu d'Estudis Superiors

**GA** Algorisme Genètic / *Genetic Algorithm*

**GBML** Sistemes d'Aprenentatge Artificial basats en Algorismes Genètics / *Genetic Based Machine Learning*

**GECCO** Conferència de Computació Genètica i Evolutiva / *Genetic and Evolutionary Computation Conference*

**GP** Programació Genètica *Genetic Programming*

**GPL** *GNU General Public License*

**GR-SETAD** Grup de Recerca en Sistemes Electronics i Anàlisi de dades

**GRSI** Grup de Recerca en Sistemes Intel · ligents

**HSJD** Hospital Sant Joan de Déu

**HTM** Grup de Recerca en Hospitalitat, Turisme i Mobilitat / *Research Group in Hospitality, Tourism and Mobilities*

**HTSI** *School of Tourism and Hospitality Management Sant Ignasi*

**IDSS** Sistemes de Suport a la Presa de Decisions / *Intelligent Decision Support Systems*

**IRL** Aprenentatge Iteratiu de Regles / *Iterative Rule Learning*

**KNN** K-Veí més Proper / *K-Nearest Neighbor*

**KEEL** *Knowledge Extraction based on Evolutionary Learning*

**LCS** Sistemes Classificadors d'Aprenentatge artificial / *Learning Classifier Systems*

**LOU** *Ley Orgánica de Universidades*

**MDS** Escalament multidimensional / *Multidimensional Scaling*

**MEC** *Ministerio de Educación y Ciencia*

**MID-CBR** *Marco Integrador para el Desarrollo de Sistemas de CBR*

**NN** Xarxes Neuronals / *Neural Networks*

**SOM** Mapes Autoorganitzatius / *Self-Organizing Maps*

**UPC** Universitat Politècnica de Catalunya

**URL** Universitat Ramon Llull

# B

# Publicacions

# Evolución de modelos jerárquicos de reglas en problemas anidados y no anidados. *Actas de la primera jornada de algoritmos evolutivos y metaheurísticas*, 2007

F. Teixidó-Navarro, and E. Bernadó-Mansilla. Evolucion de modelos jerárquicos de reglas en problemas anidados y no anidados. Actas de la primera jornada de algoritmos evolutivos y metaheurísticas (JAEM'07). Publicació en congrés nacional.

# Evolución de modelos jerárquicos de reglas en problemas anidados y no anidados

Francesc Teixidó-Navarro, Ester Bernadó-Mansilla

GRSI - Grup de Recerca en Sistemes Intel·ligents

Enginyeria i Arquitectura La Salle - Universitat Ramon Llull

Quatre Camins 2, 08022 Barcelona (Spain)

{fteixido,esterb}@salle.url.edu

## Resumen

Este artículo estudia el comportamiento del sistema HIDER, que se caracteriza por ser un sistema clasificador incremental que evoluciona un conjunto jerárquico de reglas. Mediante un conjunto de problemas artificiales, se analiza la capacidad del algoritmo incremental para evolucionar representaciones en entornos formados por ejemplos distribuidos de forma anidada y no anidada. A la vez, se estudia el efecto de la función de *fitness* en el conjunto de reglas final. Asimismo, se destaca el interés por el sistema HIDER por ser competitivo respecto a otras aproximaciones basadas en algoritmos evolutivos, ya que explora eficientemente el espacio de búsqueda y evoluciona conjuntos reducidos de reglas.

## 1. Introducción

Recientemente ha habido un auge del estudio de los sistemas clasificadores basados en algoritmos genéticos. La mayor parte de estos estudios se han centrado en los sistemas clasificadores de tipo Michigan [5], especialmente motivados por los resultados exitosos de sistemas como XCS [9] y UCS [3]. Los sistemas Michigan se caracterizan por evolucionar una única población de reglas. El papel del algoritmo genético (AG) es el de un algoritmo de búsqueda de reglas. Los sistemas obtienen conjuntos de reglas con elevada precisión. No obstante, algunos problemas detectados son el elevado coste computacional y la obtención de conjuntos con gran cantidad de reglas, lo cual dificulta la interpretación del resultado. Por otra parte, los sistemas Pittsburgh [8] evolucionan una población de conjuntos de reglas, suponiendo un coste computacional todavía más elevado. Cada individuo del algoritmo genético codifica un conjunto de reglas que se comporta como una lista ordenada de reglas.

HIDER [2] se caracteriza por ser un sistema iterativo incremental que evoluciona un conjunto jerárquico de reglas. La ventaja de HIDER es un espacio de búsqueda reducido que implica un menor coste computacional. Este sistema ha sido testeado en problemas reales [2] y ha obtenido buenos resultados comparado con sistemas como el C4.5. El objetivo de este estudio es analizar su comportamiento y si los resultados son prometedores, recuperar su uso para posteriormente poder compararlo con otros esquemas actuales como XCS. Creemos que las características de HIDER lo hacen idóneo para problemas con elevado número de ejemplos, donde los sistemas clasificadores clásicos basados en AGs presentarían un coste computacional demasiado elevado. Incluso con problemas de dimensionalidad pequeña, HIDER puede ofrecer conjuntos de reglas más reducidos y por tanto, más fácilmente interpretables que los de un sistema Michigan. Posiblemente los conjuntos de reglas serían similares a los que obtendrían los sistemas Pittsburgh puesto que éstos también evolucionan un conjunto jerárquico de reglas.

Para analizar el comportamiento de HIDER se diseña un conjunto de problemas artificiales que permiten testear la capacidad expresi-

va de los conjuntos jerárquicos, así como las características internas del algoritmo. La estructuración del artículo es la siguiente. En la sección 2 se describe brevemente el algoritmo HIDER. A continuación, se comenta el diseño de los problemas artificiales y en la sección 4 se analiza el comportamiento de HIDER para estos problemas. Finalmente, se presentan las conclusiones y las líneas de trabajo futuro.

## 2. Algoritmo HIDER

HIDER [2], *HIerarchical DEcision Rules*, es un algoritmo de aprendizaje incremental supervisado basado en reglas jerárquicas. Este sistema explora el espacio de búsqueda utilizando un algoritmo evolutivo que produce reglas ordenadas de forma similar a una lista de decisión [7].

El conjunto de reglas es una secuencia ordenada de reglas del tipo *condición* $\rightarrow$ *clase*. La disposición de las reglas en estructura jerárquica implica una dependencia entre ellas. Es decir, en un conjunto de $n$ reglas, un ejemplo $e$ queda clasificado por la regla $r_k$ $(k \leq n)$ cuando satisface la condición de la regla $r_k$ y no satisface la condición de ninguna de las $(k-1)$ reglas anteriores.

```
Proc HIDER(Instancias I) ret CjtReglas
Regla r;
CjtReglas R;
R := 0;
Mientras I conjunto no vacío Hacer
     r := genético(I);
     R := R + r;
     I := I - InstanciasClassif(r);
Fin Mientras
```

**Algoritmo 1:** Fase de entrenamiento

El algoritmo evoluciona incrementalmente el conjunto de reglas jerárquicas hasta que el espacio de búsqueda queda cubierto (ver algoritmo 1). Dado un conjunto de entrenamiento I, el algoritmo evolutivo devuelve una regla que clasifica los ejemplos de I de la manera más general y precisa posible. La regla evolucionada se archiva en el conjunto final de reglas R y los ejemplos cubiertos por la misma se borran del conjunto I. Este proceso se repite para el conjunto reducido de ejemplos hasta que se satisface la condición de finalización.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **R1** | $A_{11}$ | $A_{12}$ | $A_{13}$ | $A_{14}$ | $A_{15}$ | ... | $A_{1m-1}$ | $C_1$ |
| **R2** | $A_{21}$ | $A_{22}$ | $A_{23}$ | $A_{24}$ | $A_{25}$ | ... | $A_{2m-1}$ | $C_2$ |
| | ... | ... | ... | ... | ... | ... | ... | ... |
| **Rn-1** | $A_{n-11}$ | $A_{n-12}$ | $A_{n-13}$ | $A_{n-14}$ | $A_{n-15}$ | ... | $A_{n-1m-1}$ | $C_{n-1}$ |
| **Rn** | $A_{n1}$ | $A_{n2}$ | $A_{n3}$ | $A_{n4}$ | $A_{n5}$ | ... | $A_{nm-1}$ | $C_n$ |

Figura 1: Conjunto de reglas

A continuación se especifican los detalles de la representación de las reglas y del algoritmo evolutivo.

### 2.1. Conjunto de reglas

Como se ha comentado previamente, el conjunto de reglas de HIDER sigue una estructura jerárquica. Cada regla es del tipo *condición* $\rightarrow$ *clase*. La condición consiste en un conjunto de tests sobre los atributos del ejemplo, del tipo $(T_1, T_2, \ldots, T_n)$, donde $n$ es el número de atributos. Si el ejemplo satisface estos tests, entonces se clasifica como la clase codificada en la regla.

Para problemas con ejemplos de atributos continuos, cada test $T_i$ se codifica mediante un intervalo $[l_i..u_i]$. Dado un ejemplo $e = (a_1, a_2, a_3, \ldots, a_n)$, donde $a_i$ es el atributo $i \leq n$, siendo $n$ el número total de atributos, y una regla $r = (T_1, T_2, \ldots, T_n) \rightarrow C$, el ejemplo satisface la condición de la regla si todos los atributos $a_i$ están dentro del intervalo correspondiente. Es decir $\forall_i : 1 \leq i \leq n : l_i \leq a_i \leq u_i$. De esta forma, la regla definiría una región del espacio de búsqueda mediante un hiperrectángulo.

### 2.2. Algoritmo evolutivo

HIDER evoluciona una población de individuos, donde cada individuo codifica una única regla. El mejor individuo obtenido después de la evolución será el que contendrá la regla que clasifique de forma más general y precisa el conjunto de entrenamiento I. La función de *fitness* juega un papel básico puesto que define

cómo debe ser la regla, y hasta qué punto se pondera la generalización y la precisión de la misma. El algoritmo evolutivo tiene las fases habituales.

#### 2.2.1. Inicialización

En la fase de inicialización se crea la población de reglas. Cada regla de la población es inicializada utilizando un ejemplo seleccionado aleatoriamente del conjunto de entrada. Cada atributo de la regla inicializada cumple $\forall_i : 1 \leq i \leq n : (l_i = a_i - v_{c1}) \wedge (u_i = a_i + v_{c2})$ siendo $v_{c1}$ y $v_{c2}$ valores del intervalo $[0..Covering]$. Los valores de *Covering* son parámetros de configuración del sistema.

#### 2.2.2. Fase de evaluación

La función de evaluación califica la bondad de las reglas evolucionadas. El algoritmo evolutivo debe evolucionar una única regla que clasifique el conjunto de instancias. De hecho, clasificar todas las instancias mediante una única regla es a veces imposible y da lugar a errores de clasificación. Por tanto, el objetivo del algoritmo es encontrar una regla que clasifique **el mayor número de instancias** del conjunto de entrenamiento de la forma **más precisa** posible. El objetivo es doble y a veces contrapuesto. Si intentamos clasificar muchas instancias a la vez, se incrementa el error de clasificación. Nos encontramos pues con un problema multiobjetivo. La función de *fitness* usada por los autores de HIDER en [2] realiza una ponderación de estos objetivos, tal como se ilustra en la fórmula 1.

$$f(\varphi) = 2(N - CE(\varphi)) + G(\varphi) + Coverage(\varphi) \quad (1)$$

$$\text{donde} \begin{cases} \varphi & \text{Individuo} \\ N & \text{Número de ejemplos} \\ CE & \text{Error de clase} \\ G & \text{Ejemplos bien clasificados} \end{cases}$$

$CE(\varphi)$ es el error de clasificación, medido como el número de ejemplos que están cubiertos por la regla pero cuya clase no coincide con la de la regla. $G(\varphi)$ es el número de ejemplos correctamente clasificados. $Coverage(\varphi)$ es una medida del volumen cubierto por una regla. Concretamente, es la fracción del volumen de la región definida por la regla por el volumen total del espacio de búsqueda. Se define $[l_i..u_i]$ como el intervalo continuo asociado al atributo $i$ de la regla y $[L_i..U_i]$ es el rango de un atributo continuo $i$. El *coverage* con atributos continuos se calcula con la fórmula siguiente:

$$\begin{aligned} Coverage(\varphi) &= \prod_{i=1}^{m} \frac{Cov(\varphi, i)}{Range(\varphi, i)} \quad (2) \\ Cov(\varphi, i) &= u_i - l_i \\ Range(\varphi, i) &= U_i - L_i \end{aligned}$$

#### 2.2.3. Fase de selección

El proceso de selección se utiliza para seleccionar entre la población los candidatos a formar la población futura [6]. El algoritmo de selección utilizado es por torneo. Estudios recientes demuestran que este algoritmo es más robusto que el algoritmo de la ruleta [1]. Este método propone que un número $S$ de individuos compita. Para ello se seleccionan aleatoriamente $S$ individuos, donde $S \leq P_S$ y $P_S$ es el número total de individuos de la población. Para hacer un muestreo más equitativo de los candidatos a competir se ha elegido el muestreo sin repetición, *Sampling without repetition*. El individuo con mayor evaluación según la función de *fitness* será el seleccionado. Este proceso se repite hasta tener una población final del tamaño deseado.

#### 2.2.4. Operadores genéticos

La mutación se aplica a nivel de gen. En la representación basada en hiperrectángulos. La mutación consiste en sumar o restar un pequeño valor al gen seleccionado, ya sea este un límite superior o inferior.

El proceso de cruce utilizado es $(2, 2)$, es decir, de dos individuos seleccionados como padres $r_{p1}$ y $r_{p2}$ se crean dos nuevos individuos hijos $r_{h1}$ y $r_{h2}$, los cuales sustituirán a los padres dentro de la población. El operador de cruce está adaptado a individuos que codifican intervalos. Para obtener más detalles, consultar [2].

### 2.2.5. Fase de recuperación

Es posible que individuos potencialmente buenos se pierdan durante el ciclo evolutivo. La fase de recuperación se encarga de reestablecer los individuos desestimados de poblaciones anteriores. Concretamente, se ha usado *steady state*, el cual recupera un porcentaje de los mejores individuos de la población anterior.

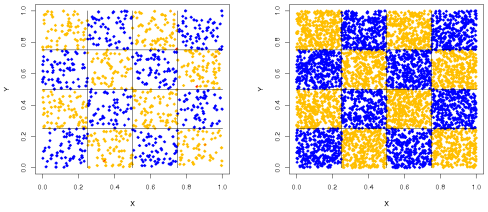
## 3. Síntesis del entorno

Para estudiar el comportamiento de HIDER se han diseñado varios problemas artificiales. El requisito principal es que los problemas diseñados sean fácilmente representables visualmente para, posteriormente, diagnosticar los resultados obtenidos. Para ello, los problemas tienen únicamente dos atributos continuos y dos clases. Para facilitar el estudio, no se ha incluido ruido y no existen valores desconocidos en los datos. No obstante, reconocemos que sería interesante añadir estas características como trabajo futuro para profundizar el estudio del comportamiento de HIDER.

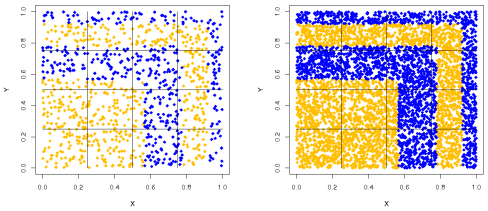
Se han definido dos tipos de problemas: con anidamiento descendente y sin anidamiento. HIDER tiene la particularidad que aprende incrementalmente un conjunto de reglas jerárquicas. Por tanto, es un sistema ideal para aprender problemas en que los ejemplos estén distribuidos de forma anidada. Los problemas diseñados permitirán testear esta capacidad. Asimismo, se han diseñado fronteras de separación entre clases rectas y curvas para evaluar el comportamiento con la representación de hiperrectángulos.

### 3.1. Problema sin anidamiento

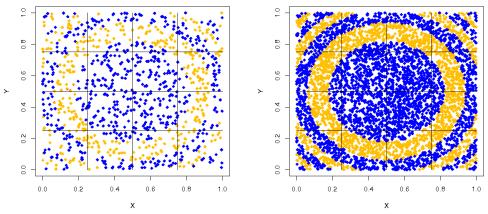
Este problema se caracteriza por tener subconjuntos rectangulares alternados pertenecientes a clases distintas, de forma similar a un tablero de ajedrez (*Checkerboard*). Los ejemplos se distribuyen uniformemente en el espacio de atributos de forma que las dos clases son equiprobables. La figura 2(a) muestra dos conjuntos de entrenamiento, formados por 1024 y 5120 ejemplos respectivamente.



(a) Tablero de ajedrez



(b) Cuadrados anidados



(c) Círculos anidados

Figura 2: Problemas de entrenamiento diseñados (a la izquierda con 1024 instancias y a la derecha con 5120)

### 3.2. Problemas con anidamiento

Este conjunto de problemas está compuesto por dos problemas de anidamiento distintos: uno de regiones cuadradas (*Nested squares*) y otro de regiones circulares (*Nested circles*) anidadas. Los problemas se han diseñado de manera que el área anidada contiene más ejemplos que la contenedora. Las figuras 2(b) y 2(c) muestran la distribución de los ejemplos de entrenamiento en los dos problemas respectivamente (cada uno con 1024 y 5120 ejemplos).

| | |
|---|---|
| Iteraciones del genético: | 300 |
| Tamaño población: | 100 |
| Tamaño selección: | 3 |
| Probabilidad cruce: | 0.5 |
| Probabilidad mutación individual: | 0.2 |
| Probabilidad mutación gen: | 0.1 |
| *Offset* mutación: | 0.25 |
| *Covering* numérico: | 0.25 |
| *Covering* nominal: | 0.5 |
| Porcentaje *Steady state*: | 0.1 |

Cuadro 1: Configuración de los parámetros

## 4. Experimentación

Se ha ejecutado el sistema HIDER con la configuración de parámetros detallada en el cuadro 1. La figura 4 muestra el resultado obtenido por HIDER con los problemas para conjuntos de entrenamiento de 1024 y 5120 instancias. Cada figura representa la clasificación realizada por HIDER de un conjunto intensivo de puntos que están muestreados uniformemente en el espacio de búsqueda.

En los tres problemas se observa que HIDER obtiene mejor rendimiento para conjuntos de entrenamiento más reducidos. Esto significa que HIDER tiene más dificultades para conjuntos de ejemplos mayores, aunque las fronteras de decisión sean las mismas. El motivo es que resulta más difícil evolucionar reglas precisas y a la vez generales, que cubran gran cantidad de ejemplos.

Por lo que se refiere al problema del tablero de ajedrez, HIDER evoluciona casi a la perfección las fronteras de clasificación. La representación con hiperrectángulos es óptima para este tipo de problema. Asimismo, la distribución de ejemplos en cuadros alternos no supone ningún obstáculo para el esquema de aprendizaje incremental de HIDER. El número de reglas evolucionadas para este problema es mínimo (ver cuadro 2).

En el segundo problema el rendimiento de HIDER es prácticamente óptimo (ver la figura 3(b)). Las fronteras de clasificación son correctas y de nuevo, el conjunto de reglas evolucionado es el mínimo posible. En este caso, la distribución de ejemplos se describe muy fácilmente con un conjunto de reglas jerárquico.

| **Problema** | Num. de instancias **1024** | **5120** |
|---|---|---|
| *Checkerboard, B* | 15 | 14 |
| *Nested Squares, NS* | 4 | 4 |
| *Nested Circles, NC* | 29 | 46 |

Cuadro 2: Número de reglas generadas con el sistema normalizado

El algoritmo resulta muy eficiente; es rápido y la solución obtenida es óptima. Si se usara un conjunto de reglas no jerárquico, como por ejemplo una disyunción de reglas, el número de reglas tendría que ser superior. Éste sería el caso de las reglas evolucionadas por sistemas clasificadores como XCS y UCS (del tipo Michigan) que a pesar de ser muy competitivos (en términos de precisión), tienden a obtener conjuntos grandes de reglas.

El tercer problema es el que presenta más dificultades para HIDER, tal como se muestra por las fronteras de clasificación obtenidas (ver figura 3(c)). El motivo de esta dificultad es doble. Por un lado, la representación en hiperrectángulos tiene menos precisión puesto que las fronteras de clasificación son curvas. Esto implica el uso de mayor número de reglas y por tanto, menos generalización de las mismas. Por otro lado, y a raíz del resultado obtenido, se observa que el conjunto de reglas tiende a iniciarse desde los límites del espacio de búsqueda. Las ecuaciones (3)-(5) muestran las matrices de confusión de HIDER para los tres problemas. Cabe notar que el error de clasificación del problema del tablero y de los cuadrados anidados es prácticamente nulo. En cambio, el error en el caso de los círculos anidados es mucho mayor.

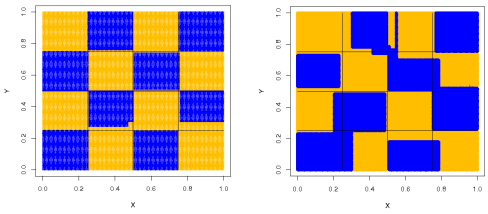
$$CM_B^{1024} = \begin{bmatrix} 535 & 4 \\ 16 & 469 \end{bmatrix} \quad (3)$$

$$CM_{NS}^{1024} = \begin{bmatrix} 591 & 3 \\ 9 & 421 \end{bmatrix} \quad (4)$$

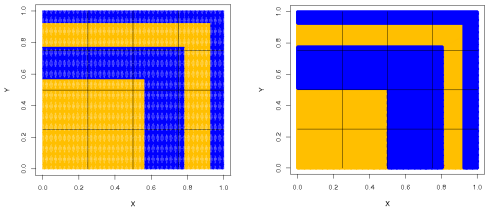
$$CM_{NC}^{1024} = \begin{bmatrix} 295 & 119 \\ 33 & 577 \end{bmatrix} \quad (5)$$

### 4.1. Sistema con *fitness* normalizado

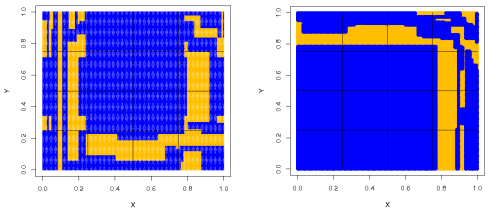
A raíz de los resultados obtenidos, se han estudiado las distintas fases del algoritmo evo-



(a) Tablero



(b) Cuadrados anidados



(c) Círculos anidados

Figura 3: Fronteras de clasificación obtenidas por HIDER con conjuntos de entrenamiento de 1024 instancias (izquierda) y 5120 instancias (derecha).

lutivo. Nuestro estudio se ha centrado principalmente en la función de *fitness*, ya que juega un papel muy importante en el tipo de reglas que se evolucionan. Como se ha comentado anteriormente, el aprendizaje de un conjunto de reglas general y preciso es un problema multiobjetivo complejo. En HIDER, la función de *fitness* simplemente realiza una suma ponderada de los objetivos deseados (ver fórmula 1). Específicamente, la función pondera tres aspectos: 1) el error (número de ejemplos mal clasificados), 2) la bondad (número de ejemplos bien clasificados) y 3) la generalización (medida como la proporción del volumen del espacio de búsqueda cubierto por la regla).

La minimización del error se realiza con el término $2(N - CE(\varphi))$. Por tanto sus valores están comprendidos entre 0 y 2N. La bondad (o precisión) equivale a las instancias bien clasificadas por dicha regla, valor entero del intervalo $[0, N]$. La generalización es la proporción del volumen cubierto y por tanto presenta valores en el intervalo $[0, 1]$.

Se observa que el factor de generalización apenas influye en la función de *fitness*, dado que siempre está comprendido entre 0 y 1. Además, su influencia en la función de *fitness* y la importancia relativa respecto al error y precisión depende del número de ejemplos de entrenamiento N. Por este motivo, y para poder entender más a fondo la contribución de cada factor en el conjunto de reglas final, decidimos normalizar a N el factor de error y de precisión. Con ello, conseguimos que la generalización tenga más importancia en el cómputo final del *fitness*. Además su contribución relativa no depende de N. La fórmula 6 muestra la función de *fitness* normalizada.

$$f(\varphi) = 2 - \frac{2 * CE(\varphi)}{N} + \frac{G(\varphi)}{N} + Coverage(\varphi) \quad (6)$$

La figura 4 muestra el resultado obtenido en los tres problemas. Por brevedad, sólo se muestran los resultados para los conjuntos de entrenamiento de 1024 instancias. Se observa que el rendimiento en este caso es mejor. En el problema del tablero y los círculos anidados, se observan mejores fronteras de clasificación. Asimismo, las ecuaciones (7), (8) y (9) muestran las matrices de confusión con error de clasificación nulo. En cambio, el número de reglas obtenidas es mucho mayor en los tres problemas (ver cuadro 3).

El resultado obtenido es en cierta manera sorprendente. Al aumentar la importancia relativa de la generalización se esperaría, en principio, reglas más generales y por tanto, conjuntos de reglas más compactos. Sin embargo, sucede lo contrario: se obtienen conjuntos más numerosos de reglas menos generales y con mejor precisión. Analizando el comportamiento interno de HIDER en este caso, hemos visto que lo que sucede es lo siguiente. Al aumentar el peso de la cobertura de las reglas en el *fitness*, las reglas tienden a ser más genera-

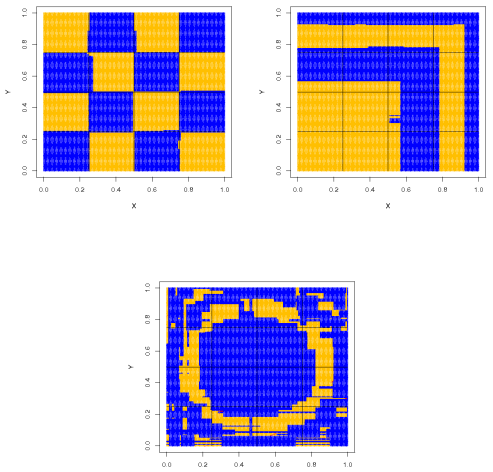


Figura 4: Fronteras de clasificación obtenidas con la función de *fitness* normalizada

| Problema | Num. de instancias 1024 |
|---|---|
| *Checkerboard, B* | 30 |
| *Nested Squares, NS* | 27 |
| *Nested Circles, NC* | 114 |

Cuadro 3: Número de reglas generadas con el sistema normalizado

les. Sin embargo, al aumentar la generalización de las reglas vemos que éstas tienden a cubrir ejemplos pertenecientes a las dos clases. Como se penaliza mucho más el error de clase ($CE$) que los ejemplos bien clasificados ($G$), una regla general que cubre ejemplos de las dos clases tiende a tener un bajo *fitness*. Por tanto, lo que estamos haciendo es presionando implícitamente hacia reglas más precisas (es decir, con menor error). Esto explica que el error obtenido sea nulo.

Este resultado muestra las relaciones implícitas entre los tres factores de la función de *fitness*. Los tres términos están relacionados de forma que si modificamos el peso de uno de ellos también estamos alterando implícitamente la contribución de los otros. Es por eso que es difícil ajustar la función de *fitness*. Un cambio en la misma produce resultados significativamente distintos tanto en términos de error como en el número de reglas obtenido. Estos resultados nos animan a seguir estudiando la función de *fitness*. Una posible línea de trabajo futuro es analizar a fondo cómo ajustar la función de *fitness* en función de los requisitos esperados. Puede que para determinados problemas interese obtener el mínimo error y para otros que prioricemos la interpretación del conjunto de reglas y por tanto, prefiramos conjuntos más compactos aunque menos precisos. Además, se podría codificar una función de *fitness* multiobjetivo para que sea el usuario el que decida el compromiso entre precisión y generalización a posteriori (cuando la evolución ha terminado).

Aunque el sistema con *fitness* normalizado obtiene mayor número de reglas, todavía sigue siendo competitivo con respecto a otros sistemas clasificadores basados en computación evolutiva. Por ejemplo, según [4] el sistema XCS obtuvo en el problema del tablero de ajedrez (usando el mismo conjunto de entrenamiento) 45 reglas de clasificación (representadas también mediante hiperrectángulos). En el caso de los cuadrados anidados, XCS obtuvo 43 reglas. En los dos casos el aprendizaje incremental de HIDER obtiene conjuntos más compactos. El motivo es que XCS no aprende de forma jerárquica: dado un conjunto de entrenamiento evoluciona un conjunto de reglas para todos los ejemplos a la vez. Esto aumenta el espacio de búsqueda y tiende a generar muchas más reglas. Además, en los problemas anidados la representación jerárquica es más sencilla.

$$CM_B^{1024} = \begin{bmatrix} 539 & 0 \\ 0 & 485 \end{bmatrix} \quad (7)$$

$$CM_{NS}^{1024} = \begin{bmatrix} 594 & 0 \\ 0 & 430 \end{bmatrix} \quad (8)$$

$$CM_{NC}^{1024} = \begin{bmatrix} 414 & 0 \\ 0 & 610 \end{bmatrix} \quad (9)$$

## 5. Conclusiones

Este artículo estudia la capacidad de extracción de conocimiento de HIDER el cual aprende de incrementalmente un conjunto de reglas jerárquicas. Se ha estudiado el conjunto de re-

glas obtenido en problemas artificiales definidos de forma anidada y no anidada. En ambos casos HIDER es capaz de obtener fronteras de clasificación prácticamente óptimas. En general, los problemas anidados facilitan el aprendizaje de HIDER, ya que el conjunto de reglas jerárquico es más natural para describir estos problemas.

El estudio realizado también destaca la sensibilidad del resultado a la función de *fitness*. Pequeñas variaciones en la función de *fitness* pueden dar lugar a resultados significativamente distintos. Además, los factores de error, precisión y generalización de las reglas están estrechamente ligados, lo cual dificulta el ajuste de una función de *fitness* óptima. El uso de algoritmos genéticos multiobjetivo podría evolucionar una serie de reglas con distintos compromisos de generalización, precisión y error. El resultado podría ser más robusto, y permitiría que el usuario seleccionase la regla que mejor se adapta a sus intereses.

En resumen, destacamos que HIDER es un sistema que ofrece ventajas respecto a otros sistemas clasificadores basados en algoritmos genéticos. El aprendizaje incremental jerárquico permite evolucionar conjuntos de reglas más compactos que sistemas del tipo Michigan y con menos recursos computacionales.

## Agradecimientos

Los autores agradecen el apoyo de *Enginyeria i Arquitectura La Salle*, Universitat Ramon Llull, y del *Ministerio de Ciencia y Tecnología* con el proyecto TIN2005-08386-C05-04.

## Referencias

[1] Orriols-Puig A., Sastry K., P. L. Lanzi, D. E. Goldberg, and Bernadó-Mansilla E. Modeling Selection Pressure in XCS for Proportionate and Tournament Selection. *IlliGAL Rep. 2007004*, 2007.

[2] Jesús S. Aguilar-Ruiz, José Cristóbal Riquelme Santos, and Miguel Toro. Evolutionary learning of hierarchical decision rules. *IEEE Trans. on Systems, Man and Cybernetics, B*, 33(2):324–331, 2003.

[3] Ester Bernadó Mansilla and Josep M. Garrell. Accuracy-Based Learning Classifier Systems: Models, Analysis and Applications to Classification Tasks. *Evolutionary Computation*, 11(3):209–238, 2003.

[4] Ester Bernadó-Mansilla and Tin K. Ho. Domain of Competence of XCS Classifier System in Complexity Measurement Space. *IEEE Trans. on Evolutionary Computation*, 9(1):82–104, 2005.

[5] John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.* MIT Press, Cambridge, MA, USA, 1992.

[6] George F. Luger. *Artificial Intelligence, Structures and Strategies for Complex Problem Solving Fourth edition.* Addison-Wesley, 2002.

[7] R. L. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, 1987.

[8] S. F. Smith. Flexible Learning of Problem Solving Heuristics through Adaptive Search. In *8th Int. Joint Conf. on Artificial Intelligence*, pages 422–425, 1983.

[9] Stewart W. Wilson. Classifier Fitness Based on Accuracy. *Evolutionary Computation*, 3(2):149–175, 1995.

# Hierarchical Evolution of Linear Regressors. *Genetic and Evolutionary Computation Conference (GECCO'07)*, 2007

F. Teixidó-Navarro, A. Orriols-Puig, and E. Bernadó-Mansilla. Hierarchical Evolution of Linear Regressors. In Genetic and Evolutionary Computation Conference (GECCO'07), ACM Press, 1413–1420, ISBN 978-1-60558-131-6, 2008. Publicació en congrés Internacional Core A de Computer Science

# Hierarchical Evolution of Linear Regressors

Francesc Teixidó-Navarro, Albert Orriols-Puig, and Ester Bernadó-Mansilla
Grup de Recerca en Sistemes Intel·ligents
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
Quatre Camins, 2, 08022 Barcelona, Spain
{fteixido,aorriols,esterb}@salle.url.edu

## ABSTRACT

We propose an algorithm for function approximation that evolves a set of hierarchical piece-wise linear regressors. The algorithm, named HIRE-Lin, follows the iterative rule learning approach. A genetic algorithm is iteratively called to find a partition of the search space where a linear regressor can accurately fit the objective function. The resulting ruleset performs an approximation to the objective function formed by a hierarchy of locally trained linear regressors. The approach is evaluated in a set of objective functions and compared to other regression techniques.

## Categories and Subject Descriptors

I.2.6 [**Learning**]: concept learning, knowledge acquisition

## General Terms

Algorithms

## Keywords

Genetic algorithms, machine learning, function approximation, regression

## 1. INTRODUCTION

Genetic algorithms (GAs) have proved to be valuable in machine learning and data mining applications. Particularly, genetic algorithms have been used in classification problems. Therein, a model is required to describe the relationship between the characteristics of the examples provided in a dataset and their associated class. Some of the benefits offered by genetic algorithms are the domain independence, the ability to evolve several types of representations (e.g., rulesets, trees), and high performance.

Among the current approximations dealing with rulesets, the Michigan approach [14] evolves a set of overlapping classifiers that together approximate the class boundary. The approach evolves a set of individuals that are incrementally evaluated. Since each individual codifies a single rule, the GA has to balance the competition-cooperation tradeoff to achieve a set of optimal classifiers jointly approximating the class boundary. Hence, the Michigan approach uses a genetic algorithm that searches simultaneously for several subsolutions that together can cover the whole search space. The Pittsburgh approach [23, 24] evolves a population of rulesets, where each ruleset usually works as a decision list [20]. The algorithm searches for the best ruleset among the set of possible rulesets. Thus, the search space is larger than in Michigan approaches; however, the evolutionary pressures can be adjusted to obtain simpler rulesets (e.g., see [3]). The so-called iterative rule learning (IRL) [13, 25] approach, also referred to as sequential covering algorithm [18], can alternatively be used to evolve rulesets. The IRL approach allows the GA to search for a single rule in each iteration. Each time a rule is obtained, the region of the search space covered by the rule is removed for the subsequent searches. Search complexity is bounded in each iteration in two respects. First, the evolution of a single rule in each iteration provides less complexity than in the Pittsburgh approach. Second, the search space is progressively reduced in each iteration. The evolved ruleset must be evaluated in order, analogously to a decision list.

Due to the benefits of reduced complexity, IRL algorithms are valuable to address large datasets. Moreover, the rule sets are highly interpretable because they contain fewer rules, possibly comparable to Pittsburgh rulesets.

Recently, some learning classifier systems such as XCS [29, 30] have been extended to deal with numeric prediction [28]. Numeric prediction (also called *regression*) can be seen as a variant of classification learning where the class is a numerical value rather than a category [31]. Herein, the emphasis of the learner is to perform function approximation. Much research has been conducted recently on Michigan approaches, particularly XCSF [28, 15, 8], and also on Pittsburgh approaches [4] for function approximation. In this paper, we extend the IRL approach to numeric prediction applications. The proposed system, named HIRE-Lin, evolves iteratively a set of linear regressors performing piece-wise linear approximations. Our aim is to propose a new architecture for the evolution of hierarchical linear regressors based on genetic algorithms, and thus, we wish to inherit the GA's capabilities such as robustness, domain independence, simplicity, and interpretability. Such capabilities will be evaluated and compared to classical approaches for regression. Moreover, such an approach would offer compound benefits from the Michigan and Pittsburgh approaches: a bounded search space complexity and high interpretable results.


*GECCO'08,* July 12–16, 2008, Atlanta, Georgia, USA.

The remainder of the paper is structured as follows. The next section describes HIRE-Lin for function approximation. Section 3 describes the experimental methodology. Section 4 analyzes HIRE-Lin on a case study and evaluates its behavior with different settings. Next, we compare HIRE-Lin with other types of regression techniques. Finally, we conclude and present future work lines.

# 2. ALGORITHM

HIRE-Lin is an Iterative Rule Learning (IRL) algorithm, inspired by Hider [1], that evolves a hierarchical rule set that approximates functions using linear regressors. At each iteration, the algorithm searches for a single rule that best approximates the data. The rule is added to the final ruleset, and instances covered by the rule are removed. The algorithm continues iteratively until the data set is covered. The final ruleset must be interpreted in the same order as rules were produced. The search is based on a genetic algorithm. The details of HIRE-Lin are as follows.

## 2.1 Rule Representation

HIRE-Lin iteratively evolves a rule set that works as a *decision list* as proposed in [20][1]. Given an example $\mathbf{e}$ and a ruleset $H$, each rule $r_i \in H$ is checked in order until a matching rule is found. That is, rule $r_m$ will predict a given example $\mathbf{e}$ if there is not any preceeding rule $r_i \,|\, i < m$, that covers the example. In such a case, the linear predictor coded in rule $r_m$ will be used to approximate the objective function $f(\mathbf{x})$ at $\mathbf{x} = \mathbf{e}$. Each rule has the form:

$$r_i = X \rightarrow W \tag{1}$$

where $X$ stands for the condition (or antecedent), and $W$ (the consequent) corresponds to the linear regressor applied when the condition is satisfied by the example. The condition defines a hyperrectangle in the search space, represented as a sequence of intervals $(x_1, x_2, \ldots, x_\ell)$, where $\ell$ is the dimension of the feature space. Each interval $x_i$ is defined by its lower and upper bounds $[lb_i, ub_i]$, both real-valued. An example $\mathbf{e} = (e_1, e_2, \ldots, e_\ell)$ satisfies the rule if $\forall_i : 1 \leq i \leq \ell : (lb_i \leq e_i \leq ub_i)$.

The consequent, $W$, represents the parameters of a linear regressor. Given an example $\mathbf{e} = (e_1, e_2, \ldots, e_\ell)$, the linear regressor approximates $f(e)$ by the hyperplane:

$$y_i = w_0 + w_1 e_1 + w_2 e_2 + ... + w_\ell e_\ell \tag{2}$$

where $w_0, w_1, \ldots, w_\ell$ are the regressor parameters. Thus, a rule defines a hyperplane $W$ which is applicable in the attribute domain defined by the hyperrectangle coded by $X$. The number of parameters of the linear regressor is $(\ell + 1)$. The final length of the rule is $(3\ell + 1)$, from where $2\ell$ correspond to the antecendent and the remainder $\ell + 1$ belong to the consequent.

The linear regressor is computed by simple regression [12, 19] according to the least squares criterion. We used the multi-dimensional least squares fitting routine available with GNU Scientific Library [1] (GSL).

## 2.2 Learning Process

The algorithm of HIRE-Lin is depicted in Alg. 1. Given a dataset $E$, the algorithm iteratively evolves a hierarchical ruleset $H$. In each iteration, a genetic algorithm (GA) is fired to search for the best rule covering accurately a high number of instances of the dataset $E$. The best rule returned by the GA is added to $H$. Next, the instances covered by the rule are removed from the dataset. The process is repeated until $E$ is empty. Finally, a default rule is added into ruleset H. Its antecedent covers the entire search space, although it will be only applicable when the previous rules do not match. The consequent is a rough approximation computed as the average of the value of the objective function of all training points.

**Algorithm 1**: Algorithm of HIRE-Lin.

```
setOfExamples E;
Rule r;
setOfHierarchicalRules H;
H := ∅;
while E not empty do
    r := evolve(E);
    H := H + r;
    E := E - CoveredInstances(r);
H := H + defaultRule;
```

## 2.3 Genetic Algorithm

Given a dataset $E$, the GA searches for the best rule that approximates the dataset. The search goal is to find the *best rule that approximates accurately the highest number of instances of* $E$, i.e., to search for the largest hyperrectangle that can be accurately approximated by a linear predictor.

The GA evolves a population $P$ of $N$ individuals, where each individual codifies a rule as described in equation 1. Each rule is a vector of $(3\ell + 1)$ real numbers. The GA only modifies the antecedent of the rule, which is of size $2\ell$. The consequent of the rule are the parameters of the linear predictor which are obtained by least squares.

### 2.3.1 Initialization

In the initialization phase, a population $P$ is created. Each individual contains a rule which is initialized in two steps. First, the antecedent $X$ of each rule is initialized using an example $\mathbf{e}$ randomly selected from the dataset $E$. For each attribute $e_i \,|\, i : 1 \ldots \ell$, an interval $[lb_i, ub_i]$ containing $e_i$ is set according to: $lb_i = e_i - v_c$ and $ub_i = e_i + v_c$, where $v_c$ is a value uniformly distributed in the interval $[0, r_0]$, being $r_0 \geq 0$ a configuration parameter. Both values are limited to the range of the attribute. Then, the consequent $W$ is calculated as follows. A linear regressor is computed by a least squares procedure, considering only the examples enclosed in the hyperrectangle defined by the antecedent $X$.

### 2.3.2 Fitness function

The fitness of the rule codifies the search goals: to maximize the hyperrectangle while minimizing the approximation error of the linear regressor.

Given a rule $r$, a linear regressor is computed using only the instances covered by the hyperrectangle. Then, the fitness of the rule is computed as follows:

$$F(r) = coverage(r) * acc(r)^\gamma \tag{3}$$

where $coverage(r)$ is the portion of the search space covered by the rule, $acc$ is the accuracy of the approximation, and $\gamma$ is a user-defined parameter. $0 \leq coverage(r) \leq 1$ is the

[1] http://www.gnu.org/software/gsl

ratio of the subspace covered by the hyperrectangle divided by the search space defined by the original training dataset:

$$\begin{aligned} coverage(r) &= \prod_{i=1}^{\ell} \frac{coverage(r,i)}{range(i)} \\ coverage(r,i) &= ub_i - lb_i \\ range(i) &= UB_i - LB_i \end{aligned} \quad (4)$$

where $coverage(r,i)$ is the size of the interval of rule $r$ for attribute $i$, and $range(i)$ is the difference between the maximum ($UB_i$) and minimum ($LB_i$) values of attribute $i$. *acc* considers the quality of the linear approximation and is defined as:

$$acc = R^2 \quad (5)$$

where $R^2$ is the coefficient of determination [19] of the linear regressor, computed as follows:

$$R^2 = 1 - \frac{SS_E}{SS_T} \quad (6)$$

where $SS_E$ is the sum of squared errors and $SS_T$ is the total sum of squares of the function value of the corresponding points. $R^2$ is the proportion of variability in the training points that is accounted for by the regressor model. That is, $R^2$ is a statistic that provides information on how well the regression line approximates the real data points. An $R^2$ of 1.0 indicates that the regression line perfectly fits the data. Thus, *coverage* represents the generalization of the rule, while *acc* is the accuracy or quality of the linear regressor trained for that rule. $\gamma \geq 1$ is a parameter that specifies the relevance of the accuracy term with respect to the generalization term. In the remainder of the paper, we will refer to it as *accuracy pressure* parameter.

Note that generalization and accuracy are two objectives that must be maximized. Our fitness function takes an aggregating approach [5], being $\gamma$ a control parameter that specifies the relative weight of these two objectives.

### 2.3.3 Genetic Operators

Selection of individuals is performed via *tournament selection* with tournament size $S$, where $0 < S \leq N$. Mutation is applied with probability $p_m$ per gene. Let $g_i$ be a gene representing either the lower or upper bound of an interval. The value is mutated to a value uniformly distributed in the range $[g_i - m_{off}, g_i + m_{off}]$, where $m_{off}$ is a parameter. The new value is restricted to the range of the corresponding attribute so that the resulting interval is correct. The crossover operator is applied with probability $p_c$. Given two parents, two children are obtained that replace their parents in the population. One point crossover is implemented which chooses a cut point uniformly distributed in the range $[1..2\ell]$, where $\ell$ is the number of attributes. Elitism is applied to preserve the best solution found from one cycle to the next one.

## 3. EXPERIMENTAL METHODOLOGY

To analyze HIRE-Lin, we designed a set of artificial datasets that corresponded to functions of different orders, topologies (concave, convex), and dimensions. Table 1 shows the mathematical formula of each function together with the mnemonic we will use in the paper. Some of these functions have already been used as benchmarks to test regression models (see for example [17]).

**Table 1: Functions test bed**

| Mnemonic | Function |
|---|---|
| $f_{asx}$ | $f(x) = \lvert \sin(10x) \rvert$ |
| $f_{px}$ | $f(x) = 1 + x + x^2 + x^3$ |
| $f_{s4x}$ | $f(x) = \sin(x) + \sin(2x) + \sin(3x) + \sin(4x)$ |
| $f_{scx}$ | $f(x) = \sin(x) * \cos(x)$ |
| $f_{x2}$ | $f(x) = x^2$ |
| $f_{xsx}$ | $f(x) = x * \sin(10x)_{f(x)>0}$ |
| $f_{rxy}$ | $f(x,y) = \sqrt{xy}$ |
| $f_{scxy}$ | $f(x,y) = \sin(xy) * \cos(xy)$ |
| $f_{sxy}$ | $f(x,y) = \sin(3xy)_{f(x,y)>0}$ |
| $f_{x2y}$ | $f(x,y) = x^2 y$ |

To build the datasets, we uniformly sampled 100 instances per dimension. That is, if the function was defined by a single attribute (one dimension) the resulting dataset contained 100 instances. For functions defined by two attributes, the resulting dataset contained 10000 instances.

To analyze the quality of the model evolved by HIRE-Lin, we considered the error and the model size. To estimate the error, each example of the dataset is checked against the ruleset. The first rule that matches provides an approximation which is compared with the value of the objective function. Thus, the error $\varepsilon$ was estimated according to the following formula:

$$\varepsilon = \frac{\sum_{i=1}^{N_E} \left( f(\mathbf{e}_k) - \widehat{f(\mathbf{e}_k)}_{w_0,\ldots,w_\ell} \right)^2}{N_E} \quad (7)$$

where $f(\mathbf{e}_k)$ is the value of the objective function at point $\mathbf{e}_k$, $\widehat{f(\mathbf{e}_k)}_{w_0,\ldots,w_\ell}$ is the function approximation provided by HIRE-Lin, and $N_E$ is the number of examples. The model size was computed as the number of rules evolved.

A 10-fold cross-validation procedure was used to estimate the error of the method. For each fold, the method was trained 10 times with different seeds and the error was averaged. When needed, statistical tests were applied to compare several approaches and test for significant differences among them. Our methodology followed the guidelines provided by Demsar [7] for multiple comparison tests. Briefly, we first tested the null hypothesis that the group of learners performed equivalently by means of a Friedman's test. If this hypothesis could be rejected, then we applied a post-hoc test to compare the learners to the best performer. Specifically, the Bonferroni-Dunn's test was used.

Our study consists of two parts. First, in Sect. 4 we analyze the behavior of HIRE-Lin. By means of a graphical analysis centered on a case study, we investigate the influence of the accuracy pressure parameter $\gamma$. We analyze the quality of the approximation and the number and types of rules evolved. We further extend this study to the whole function test bed to validate the influence of the accuracy pressure parameter. Then, in Sect. 5 we compare HIRE-Lin to three other well-known regression techniques, so that we can place our approach within some of the state-of-the-art methodologies.

## 4. ANALYSIS OF HIRE-LIN

This section analyzes the behavior of HIRE-Lin. First, we use function $f_{xsx}$ as a case study (see Table 1 for the de-

Table 2: Parameters of HiRe-Lin

| Parameter | Description | Value |
|---|---|---|
| ec | Evolutionary cycles | 300 |
| N | Population size | 100 |
| S | Tournament selection size | 10 |
| $p_c$ | Crossover probability | 0.5 |
| $p_m$ | Mutation probability | 0.02 |
| $m_{off}$ | Mutation offset (fraction) | 0.25 |
| $r_0$ | Covering parameter (fraction) | 0.25 |
| $\gamma$ | Accuracy pressure | 1 |

tails). Then, we numerically compare the results obtained with the remaining test bed. In both cases, HIRE-Lin was run with pressures $\gamma$={1,10,100,1000} and parameter settings provided in Table 2.

## 4.1 A Case Study

Figures 1 to 4 plot the results of HIRE-Lin in function $f_{xsx}$ with $\gamma$={1,10,100,1000} respectively. Each figure shows, in the upper part, the training points used to train the algorithm (plotted with points) and the function approximation provided by HIRE-Lin (with solid line). In the lower part, the subspace covered by the each rule is plotted. Note that the ruleset is hierarchical and must be checked in order.

Figure 1 shows the result of the algorithm with $\gamma = 1$. See that the algorithm evolves only two rules. The first one is a large rule covering the domain [0,0.927]. The linear regressor obtained for this subspace cannot fit the shape of the objective function. The second rule applies to the domain (0.927,0.935], which allows for a better approximation because the subspace is small enough for a linear fitting. Note that the domain of the second rule $R_2$ is [0,0.935]. However, since the previous rule already covers the subspace [0,0.927], $R_2$ is only applied in the range (0.927,0.935]. The range codified by $R_2$ which is hidden by the previous rule is plotted in dotted lines, while the effective domain is plotted in solid line. The ruleset adds another rule $R_d$, the default rule, which always covers the range [0,1], although the effective range depends on the previous rules in the list. In this case, $R_d$ would be applied only in the interval (0.935,1]. The reason why rules $R_1$ and $R_2$ have not expanded to cover completely the search space is that there are no training instances defined outside [0,0.935]. Thus, no rules are evolved for the region (0.935,1]. If a test example from this subspace is given, $R_d$ would be applied with the value of the average objective function of training points.

Figure 2 plots the results for $\gamma = 10$. Note that the ruleset contains more rules than with $\gamma = 1$ and each rule covers a smaller subspace. This allows for better fitting than with $\gamma = 1$. By increasing $\gamma$, we change the relative weight of accuracy with respect to generalization (see equation 3). With $\gamma = 1$, generalization was so important that a very general rule with a rough approximation was obtained. With $\gamma = 10$, generalization is decreased in favor of accuracy. Thus, less general but more accurate rules are given. The effect of further increasing $\gamma$ is plotted in Figures 3 and 4 which show even finer approximations to the objective function. Higher $\gamma$ values also produce larger rulesets.

## 4.2 Comparison on Several Datasets

We extended the study on the influence of $\gamma$ parameter to the remaining test bed described in Table 1. We

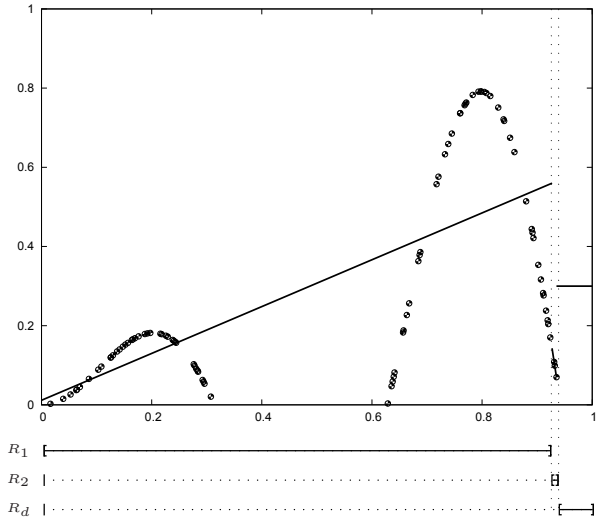



**Figure 1: $f_{xsx}$ approximation using $\gamma = 1$**

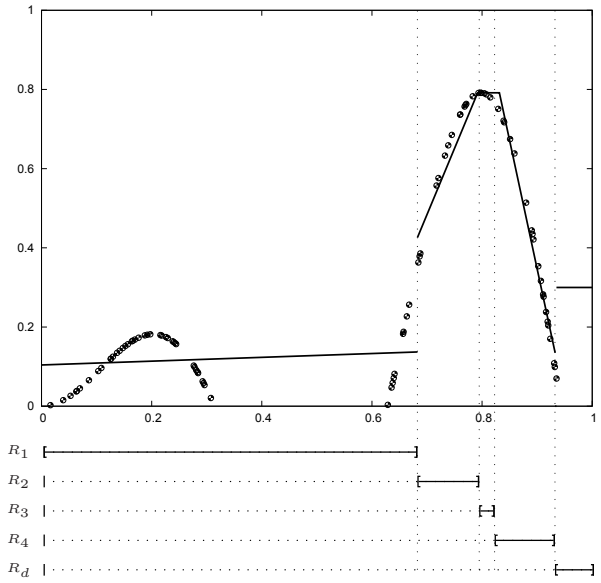



**Figure 2: $f_{xsx}$ approximation using $\gamma = 10$**

studied the training error, the test error, and the number of rules obtained for the different accuracy pressures $\gamma = \{1, 10, 100, 1000\}$. The training error is a measure of the fit of the approximation to the training points, while the test error is an estimate of the generalization capability to unseen points. The number of rules is useful as a measure of interpretability of the final ruleset.

Tables 3 and 4 show the average and standard deviation of HIRE-Lin in the training dataset and test dataset respectively. As mentioned before, these values correspond to the error estimated by a 10-fold cross-validation procedure with 10 random seeds. The best approach giving the minimum average error is marked in bold. Regarding the training errors, larger values of $\gamma$ yield smaller approximation errors. In

**Table 3: Train error of HIRE-Lin with different accuracy pressures. Each cell gives the average and standard deviation of HIRE-Lin for the dataset in the row**

| DS | $\gamma = 1$ | $\gamma = 10$ | $\gamma = 100$ | $\gamma = 1000$ |
|---|---|---|---|---|
| $f_{asx}$ | 9.407e-02±3.60e-03 | 5.094e-02±2.72e-02 | 1.867e-03±1.46e-03 | **1.653e-03±1.72e-03** |
| $f_{px}$ | 3.592e-03±1.85e-03 | 2.916e-03±1.63e-03 | 9.706e-04±1.07e-03 | **1.574e-04±1.10e-03** |
| $f_{s4x}$ | 5.966e-02±5.07e-03 | 8.176e-03±2.92e-03 | 1.284e-03±3.11e-04 | **9.511e-05±1.35e-04** |
| $f_{scx}$ | 2.368e-03±1.03e-04 | 2.610e-03±7.25e-04 | **7.544e-04±8.58e-04** | 1.343e-03±1.03e-03 |
| $f_{x2}$ | 3.385e-03±1.35e-04 | 3.821e-03±2.62e-03 | 8.261e-04±1.67e-04 | **6.935e-05±8.15e-06** |
| $f_{xsx}$ | 2.706e-02±3.00e-03 | 2.639e-03±2.07e-03 | 1.029e-04±4.12e-05 | **7.108e-05±5.07e-04** |
| $f_{rxy}$ | 4.780e-03±1.25e-04 | 4.401e-03±4.19e-04 | 1.105e-03±1.24e-04 | **1.068e-04±1.39e-05** |
| $f_{scxy}$ | 2.368e-03±1.03e-04 | 2.432e-03±9.34e-05 | 5.697e-04±8.79e-05 | **5.090e-05±9.51e-06** |
| $f_{sxy}$ | 2.552e-02±5.58e-04 | 1.602e-02±2.69e-03 | 1.592e-03±3.01e-04 | **1.717e-04±1.95e-04** |
| $f_{x2y}$ | 6.739e-03±7.77e-05 | 5.382e-03±6.37e-04 | 6.181e-04±2.27e-04 | **4.391e-05±6.87e-06** |

**Table 4: Test error of HIRE-Lin with different accuracy pressures. Each cell gives the average and standard deviation of HIRE-Lin for the dataset in the row**

| DS | $\gamma = 1$ | $\gamma = 10$ | $\gamma = 100$ | $\gamma = 1000$ |
|---|---|---|---|---|
| $f_{asx}$ | 1.005e-01±2.54e-02 | 6.983e-02±3.82e-02 | 9.010e-03±2.47e-02 | **6.932e-03±1.92e-02** |
| $f_{px}$ | 3.463e-03±1.77e-03 | 3.272e-03±1.88e-03 | 1.269e-03±7.28e-04 | **1.561e-04±6.03e-04** |
| $f_{s4x}$ | 6.676e-02±2.65e-02 | 1.215e-02±8.18e-03 | 1.784e-03±1.12e-03 | **6.120e-04±3.97e-03** |
| $f_{scx}$ | 2.554e-03±1.22e-03 | 2.899e-03±2.91e-03 | **3.531e-04±1.81e-04** | 4.363e-04±2.71e-03 |
| $f_{x2}$ | 3.694e-03±1.85e-03 | 3.450e-03±1.70e-03 | 1.121e-03±6.05e-04 | **2.293e-04±9.36e-04** |
| $f_{xsx}$ | 3.179e-02±1.17e-02 | 3.505e-03±3.89e-03 | **8.770e-04±6.28e-03** | 1.542e-03±6.13e-03 |
| $f_{rxy}$ | 4.805e-03±1.15e-03 | 4.618e-03±1.19e-03 | 1.413e-03±9.31e-04 | **1.569e-04±7.09e-05** |
| $f_{scxy}$ | 2.483e-03±4.55e-04 | 2.473e-03±4.66e-04 | 6.596e-04±2.42e-04 | **7.590e-05±5.94e-05** |
| $f_{sxy}$ | 2.558e-02±3.26e-03 | 1.765e-02±4.23e-03 | **2.325e-03±1.73e-03** | 2.860e-03±5.04e-03 |
| $f_{x2y}$ | 6.798e-03±1.21e-03 | 5.438e-03±1.19e-03 | 6.698e-04±2.19e-04 | **8.222e-05±3.98e-05** |

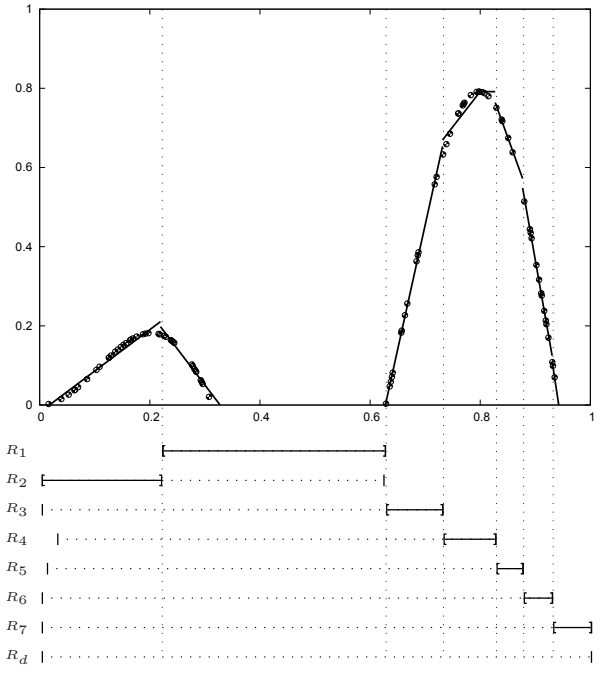



**Figure 3: $f_{xsx}$ approximation using $\gamma = 100$**

all functions, except for $f_{scx}$, the smallest error is obtained for $\gamma = 1000$. Larger values of $\gamma$ also tend to give smaller test errors. However, in functions $f_{xsx}$ and $f_{sxy}$ the largest accuracy pressure ($\gamma = 1000$) gives the smallest training error but this does not correspond to the smallest test error. This indicates that overfitting is occurring in these cases.

Table 5 shows the number of rules of the final ruleset. For a given problem, the number of rules obtained increases with larger values of $\gamma$, as it was already observed in the case study. For $\gamma = 1$ the average ruleset consists of a single rule, two at maximum (the default rule is not counted). This value is too extreme to get a good approximation. Larger pressures provide larger rulesets. Values of $\gamma$ ranging from 100 to 1000 provide fairly good approximations. By adjusting parameter $\gamma$ we can balance the compromise between accurate approximation and interpretability (smaller rulesets are usually more interpretable). Note also that the most complex problems, such as those with two attributes, require larger rulesets. $f_{sxy}$ is the problem that requires the highest number of rules.

We statistically compared the accuracy and size of the models evolved with the different configurations. In Figure 5, each system is placed in the axes according to its average rank regarding the approximation error (x-axis) and its average rank regarding population size (y-axis). The vertical dashed lines delimit the region of the comparison space where the learners perform equivalently to the learner that presented the best performance according to a Bonferroni-Dunn test at a significance level of 0.10. Similarly, horizontal lines determine the region of equivalence to the method that created the smallest models. Note that HIRE-Lin with $\gamma = 100$ and $\gamma = 1000$ evolved the most accurate models of the comparison. On the other hand, HIRE-Lin with $\gamma = 1$ and $\gamma = 10$ built the most reduced rulesets, which went in detriment of the test accuracy.

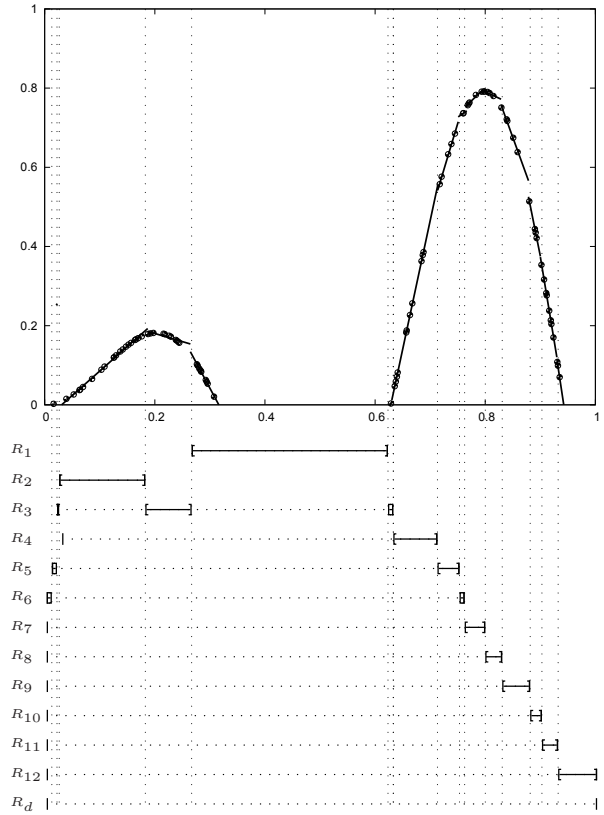



**Figure 4: $f_{xsx}$ approximation using $\gamma = 1000$**

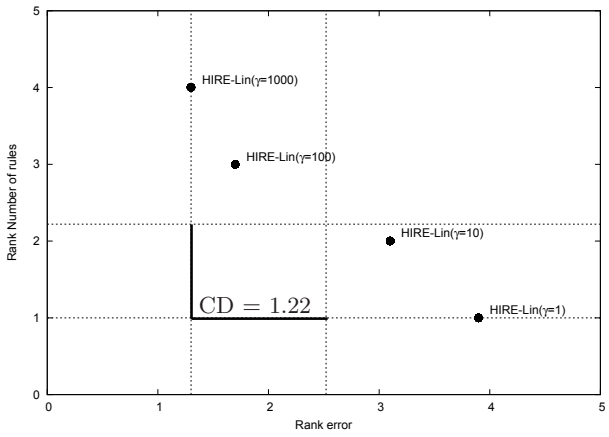



**Figure 5: Average rank of HIRE-Lin with accuracy pressures {1,10,100,1000}. The x-axis plots the rank of HIRE-Lin with respect to the test error (the approach with the smallest error has the smallest rank). The y-axis plots the rank of each approach with respect to the smallest ruleset. The critical distance (CD) delimits the region of equivalence with the best learner in each objective. It is computed according to a Bonferroni-Dunn test at $\alpha = 0.10$.**

**Table 5: Number of rules (average and standard deviation) obtained by HIRE-Lin with different accuracy pressures**

| DS | $\gamma = 1$ | $\gamma = 10$ | $\gamma = 100$ | $\gamma = 1000$ |
|---|---|---|---|---|
| $f_{asx}$ | 1.25±0.44 | 4.13±1.55 | 11.39±0.90 | 16.88±1.35 |
| $f_{px}$ | 1.05±0.22 | 1.86±0.35 | 2.42±0.50 | 4.47±0.50 |
| $f_{s4x}$ | 1.85±0.36 | 2.05±0.22 | 3.12±0.33 | 5.21±0.41 |
| $f_{scx}$ | 1.00±0.00 | 1.32±0.47 | 2.21±0.41 | 3.78±0.48 |
| $f_{x2}$ | 1.00±0.00 | 1.58±0.50 | 3.04±0.20 | 5.75±0.61 |
| $f_{xsx}$ | 1.96±0.20 | 3.71±0.56 | 6.23±0.78 | 11.23±1.21 |
| $f_{rxy}$ | 1.01±0.10 | 1.68±0.67 | 8.96±1.22 | 25.07±2.97 |
| $f_{scxy}$ | 1.01±0.10 | 1.34±0.52 | 9.68±1.32 | 29.81±2.97 |
| $f_{sxy}$ | 1.01±0.10 | 4.14±1.34 | 15.36±3.74 | 43.07±3.79 |
| $f_{x2y}$ | 1.00±0.00 | 2.89±0.60 | 10.88±2.17 | 39.55±4.33 |

**Table 6: Comparison of (a) Linear Least Mean Squares (LMS), (b) Fuzzy Wang-Mendel (WM), (c) GAP, and (d) HIRE-Lin with $\gamma = 1000$ on a collection of eleven artificial problems.**

| DS | LMS | WM | GAP | HIRE-Lin |
|---|---|---|---|---|
| $f_{asx}$ | 0.10019 | 0.10056 | 0.15491 | 0.00693 |
| $f_{px}$ | 0.00443 | 0.00022 | 0.00300 | 0.00016 |
| $f_{s4x}$ | 0.07331 | 0.05779 | 0.00273 | 0.00061 |
| $f_{scx}$ | 0.00359 | 0.00123 | 0.00075 | 0.00044 |
| $f_{x2}$ | 0.00561 | 0.00001 | 0.00101 | 0.00023 |
| $f_{xsx}$ | 0.04227 | 0.04028 | 0.05650 | 0.00154 |
| $f_{rxy}$ | 0.00488 | 0.00000 | 0.00363 | 0.00016 |
| $f_{scxy}$ | 0.00300 | 0.00125 | 0.00047 | 0.00008 |
| $f_{sxy}$ | 0.02891 | 0.02690 | 0.00219 | 0.00286 |
| $f_{x2y}$ | 0.00898 | 0.00003 | 0.00103 | 0.00008 |
| **rank** | *3.73* | *2.27* | *2.64* | *1.36* |

# 5. COMPARISON WITH OTHER REGRESSION TECHNIQUES

So far, we have analyzed the impact of the accuracy pressure in the size and accuracy of the models evolved by HIRE-Lin. In this section, we compare the behavior of HIRE-Lin with three regression techniques: Linear LMS [21], Fuzzy Wang-Mendel [26], and GAP [22]. Linear LMS uses the least mean square algorithm to create a linear approximation of the input data. Fuzzy Wang Mendel builds a set of Mandani fuzzy rules [6] that minimize the error with the covered instances. GAP is a method based on genetic algorithms and genetic programming that evolves a function represented in a tree. All these methods were run using KEEL [2]. We used the default configuration recommended in the software [2] to configure each method. We configured HIRE-Lin with the parameters specified in Table 2; besides, we set $\gamma$=1000, since, as shown in the last section, it yields accurate models of moderate size.

Table 6 provides the test error obtained for each problem and learner. The multi-comparison Friedman's test [10, 11] permitted us to reject the null hypothesis that all learners performed the same on average with $p = 8.22 \cdot 10^{-4}$. To analyze which learners performed significantly differently from HIRE-Lin, we used the post-hoc Bonferroni-Dunn test [9] at $\alpha = 0.10$. Figure 6 ranks the four learners and connects those that perform equivalently according to the Bonferroni-

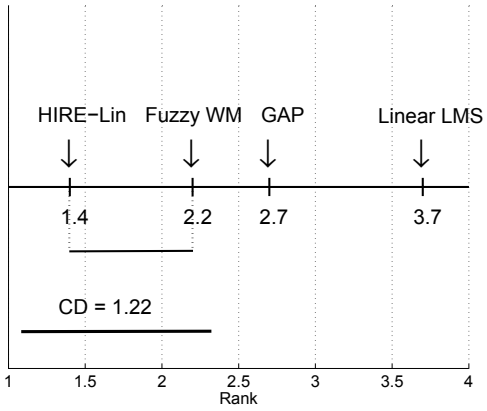



**Figure 6: Comparison of the test performance of HIRE-Lin with the other methods by means of a Bonferroni-Dunn Test at $\alpha = 0.10$. Groups of classifiers that are not significantly different to the best ranked method are connected.**

Dunn procedure. HIRE-Lin is the best ranked method, and outperforms the results obtained by Linear LMS and GAP. Linear LMS uses a least mean square approach to build a linear function that approximates the output. Note that HIRE-Lin uses the same approach to evolve piece-wise linear approximations of the function drawn by the input instances. Therefore, the partition of the feature space promoted by the genetic algorithm allows HIRE-Lin to achieve much better approximations. GAP is an evolutionary approach that evolves a function coded as a tree, which permits to represent more complex, non-linear expressions. Notice that HIRE-Lin significantly outperforms this technique on the collection of tested problems by evolving simple linear functions to approximate the input. It is worth highlighting that both LMS and GAP use a global approximation, while HIRE-Lin evolves an arbitrary number of rules that locally approximate the objective function. The number of rules evolved depends on both the non-linearity of the objective function and the accuracy pressure $\gamma$.

As the Bonferroni-Dunn test is said to be quite conservative, we also performed pairwise comparisons among learners by means of the non-parametric Wilcoxon signed-ranks test [27], assuming the risk of increasing the error of rejecting the null hypothesis when it is actually true. Table 7 provides the approximate p-values. The symbols $\oplus$ and $\ominus$ indicate that the method in the row significantly improves/degrades the performance obtained by the method in the column at a significance level of 0.05. The symbols $+/-$ denote a non-significant improvement/degradation. The pairwise analysis confirms the conclusions extracted from the Bonferroni-Dunn test; moreover, it also detects that HIRE-Lin outperforms Fuzzy Wang Mendel. Therefore, the pairwise analysis supports the conclusion that HIRE-Lin outperforms all the other methods in the comparison.

**Table 7: Pair-wise comparison of the test performance achieved by HIRE-Lin with the accuracy obtained with Linear LMS (LMS), Fuzzy Wang Mendel (WM), and GAP.**

| | LMS | WM | GAP | HIRE-Lin |
|---|---|---|---|---|
| **LMS** | | 0.004 | 0.182 | 0.003 |
| **WM** | $\oplus$ | | 0.657 | 0.026 |
| **GAP** | $+$ | $-$ | | 0.008 |
| **HIRE-Lin** | $\oplus$ | $\oplus$ | $\oplus$ | |

## 6. CONCLUSIONS

In this paper, we proposed a regression algorithm that evolves a hierarchical set of rules performing piece-wise linear approximations. The algorithm is based on an iterative rule learning approach, which consists in evolving a single rule in each iteration. Each rule delimits a subspace where an optimal linear regressor is constructed. The search space of the algorithm is progressively reduced as rules are evolved.

The genetic algorithm is applied to search for the largest hyperrectangular subspace where the optimal linear regressor, trained for the data points enclosed in that subspace, accurately approximates the objective function.

The balance between generalization and fit of the model can be adjusted in the fitness function. We used an aggregation approach, where the relative influence of these objectives could be modified by the so-called accuracy pressure. As the accuracy pressure was increased, the model obtained finer approximations at the cost of evolving larger rulesets, compromising interpretability of the final ruleset and even leading to overfitting. We acknowledge that the search could be formulated as a multiobjective fitness function based on Pareto approaches. A key advantage of such an approach is to let the user to choose among alternative compromises between generalization and model fit. A possible aid to avoid overfitting is to use an additional validation set containing points different from those in the training dataset to evaluate whether the approximation is generalizing to these unknown points. In this sense, a Pareto-based multiobjective approach would be more flexible, because it would allow the user to choose the solution with less overfitting.

Our approach evolves a set of hierarchical piece-wise linear regressors. Similarly, non-hierarchical piece-wise linear regressors are evolved by XCSF, which belongs to the category of Michigan approaches. XCSF searches simultaneously for a set of overlapping piece-wise regressors which together cover the search space. A key point of our approach is that rulesets tend to be smaller than those usually obtained by Michigan approaches. However, this hypothesis must be further investigated. As a future work we aim at comparing the rulesets and model fitting of both approaches. Also XCSF has been trained to evolve other types of regressors such as neural and polynomial regressors [16]. This feature could also be included easily in HIRE-Lin.

The architecture was highly competitive with respect to other regression techniques, such as LMS, Fuzzy Wang Mendel and GAP. In fact, it is not surprising that HIRE-Lin surpasses the behavior of the linear regressor LMS, since our approach is a local approach and LMS a global approach training a single linear regression for the whole search space. HIRE-Lin also improves GAP, a global method evolving a regression function by means of genetic algorithms and genetic programming. Other types of regressors such as locally weighted regression [18] could be more advantegeous

than global methods and compare similarly to HIRE-Lin. Although this particular study remains for further work, we already demonstrated that HIRE-Lin is competitive with respect to Fuzzy Wang Mendel, which is a local approach.

## Acknowledgements

The authors would like to thank the *Ministerio de Educación y Ciencia* for its support under project TIN2005-08386-C05-04, and *Generalitat de Catalunya* for its support under grants 2005FI-00252 and 2005SGR-00302.

## 7. REFERENCES

[1] J. S. Aguilar-Ruiz, J. C. R. Santos, and M. Toro. Evolutionary learning of hierarchical decision rules. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 33(2):324–331, 2003.

[2] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, and F. Herrera. Keel: A software tool to assess evolutionary algorithms to data mining problems. *Soft Computing*, forthcoming.

[3] J. Bacardit and J. Garrell. Bloat control and generalization pressure using the minimum description length principle for a Pittsburgh approach Learning Classifier System. In *Learning Classifier Systems, Revised Selected Papers of the International Workshop on Learning Classifier Systems 2003-2005*, volume 4399 of *Lecture Notes in Computer Science*, pages 59–79. Springer, 2003.

[4] B. Carse, T. Fogarty, and A. Munro. Evolving fuzzy rule based controllers using genetic algorithms. *Fuzzy Sets and Systems*, 80:273–293, 1996.

[5] C. C. Coello, D. V. Veldhuizen, and G. Lamont. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, 2002.

[6] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena. *Genetic fuzzy systems: Evolutionary tuning and learning of fuzzy knowledge bases*, volume 19 of *Advances in Fuzzy Systems—Aplications and Theory*. World Scientific, 2001.

[7] J. Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[8] J. Drugowitsch and A. Barry. A Formal Framework and Extensions for Function Approximation in Learning Classifier Systems. *Machine Learning*, 70(1):45–88, 2008.

[9] O. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64, 1961.

[10] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701, 1937.

[11] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

[12] S. Glantz and B. Slinker. *Primer of Applied Regression & Analysis of Variance*. McGraw-Hill, 2001.

[13] A. González and F. Herrera. Multi-stage Genetic Fuzzy Systems Based on the Iterative Rule Learning Approach. *Mathware & Soft Computing*, 4:233–249, 1997.

[14] J. H. Holland. Adaptation. In R. Rosen and F. Snell, editors, *Progress in Theoretical Biology*, volume 4, pages 263–293. Academic Press, 1976.

[15] P. Lanzi, D. Loiacono, S. Wilson, and D. Goldberg. Generalization in the XCSF Classifier System: Analysis, Improvement, and Extension. *Evolutionary Computation Journal*, 2007.

[16] P. L. Lanzi, D. Loiacono, S. W. Wilson, and D. E. Goldberg. Extending XCSF beyond linear approximation. In *GECCO '05*, pages 1827–1834. ACM, 2005.

[17] P. L. Lanzi, D. Loiacono, S. W. Wilson, and D. E. Goldberg. Prediction update algorithms for XCSF: RLS, Kalman filter, and gain adaptation. In *GECCO '06*, pages 1505–1512. ACM, 2006.

[18] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[19] D. Montgomery and G. Runger. *Applied Statistics and Probability for Engineers*. John Wiley & Sons, 2003.

[20] R. L. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, 1987.

[21] J. Rustagi. *Optimization Techniques in Statistics*. Academic Press, 1994.

[22] L. Sánchez, I. Couso, and J. A. Corrales. Combining GP operators with SA search to evolve fuzzy rule based classifiers. *Information Sciences*, 136(1–4):175–191, 2001.

[23] S. F. Smith. Flexible Learning of Problem Solving Heuristics through Adaptive Search. In *Proc. of the 8th International Joint Conference on Artificial Intelligence*, pages 422–425, 1983.

[24] W. M. Spears and D. F. Gordon. Adaptative Strategy Selection for Concept Learning. In *Proc. of the First International Workshop on Multistrategy Learning (MSL-91)*, pages 231–246. Harpers Ferry, MD, 1991.

[25] G. Venturini. SIA: A Supervised Inductive Algorithm with Genetic Search for Learning Attribute Based Concepts. In *Proc. European Conf. Machine Learning*, pages 280–296, 1993.

[26] L. Wang and J. Mendel. Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man and Cybernetics*, 22(6):1414–1427, 1992.

[27] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[28] S. Wilson. Classifiers that approximate functions. *Journal of Natural Computing*, 1(2–3):211–234, 2002.

[29] S. W. Wilson. Classifier Fitness Based on Accuracy. *Evolutionary Computation*, 3(2):149–175, 1995.

[30] S. W. Wilson. Generalization in the XCS Classifier System. In J. Koza, W. Banzhaf, K. Chellapilla, K. Deb, M. Dorigo, D. Fogel, M. Garzon, D. Goldberg, H. Iba, and R. Riolo, editors, *Genetic Programming: Proc. of the Third Annual Conference*, pages 665–674. San Francisco, CA: Morgan Kaufmann, 1998.

[31] I. H. Witten and E. Frank. *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, 2000.

# Toward High Performance Solution Retrieval in Multiobjective Clustering. *Information Sciences*, 2015

A. Garcia-Piquer, A. Sancho-Asensio, A. Fornells, E. Golobardes, G. Corral and F. Teixidó-Navarro. Towards a High Solution Retrieval in Multiobjective Clustering. Information Sciences, vol 320: 12-25, 2015. (Publicació indexada al Q1 amb factor d'impacte 3,8)

Information Sciences 320 (2015) 12–25

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Toward high performance solution retrieval in multiobjective clustering

Alvaro Garcia-Piquer [a], Andreu Sancho-Asensio [b,*], Albert Fornells [c], Elisabet Golobardes [b], Guiomar Corral [d], Francesc Teixidó-Navarro [c]

[a] Institut de Ciències de l'Espai (IEEC – CSIC), Campus UAB, Facultat de Ciències, Torre C5 – parell – 2a planta, E-08193 Bellaterra, Spain
[b] Research Group in Electronic and Telecommunications Systems and Data Analysis, Ramon Llull University, Quatre Camins 2, 08022 Barcelona, Spain
[c] Research Group in Tourism, Hospitality and Mobilities, School of Tourism and Hospitality Management – Sant Ignasi, Ramon Llull University, Marqués de Mulhacén 40-42, 08034 Barcelona, Spain
[d] Research Group in Internet Technologies and Storage, Ramon Llull University, Quatre Camins 2, 08022 Barcelona, Spain

ARTICLE INFO

*Article history:*
Received 26 September 2014
Received in revised form 4 February 2015
Accepted 18 April 2015
Available online 23 April 2015

*Keywords:*
Soft-computing
Genetic algorithms
Multiobjective optimization
Clustering
Pareto set filtering

ABSTRACT

The massive generation of unlabeled data of current industrial applications has attracted the interest of data mining practitioners. Thus, retrieving novel and useful information from these volumes of data while decreasing the costs of manipulating such amounts of information is a major issue. Multiobjective clustering algorithms are able to recognize patterns considering several objective function which is crucial in real-world situations. However, they dearth from a retrieval system for obtaining the most suitable solution, and due to the fact that the size of Pareto set can be unpractical for human experts, autonomous retrieval methods are fostered. This paper presents an automatic retrieval system for handling Pareto-based multiobjective clustering problems based on the shape of the Pareto set and the quality of the clusters. The proposed method is integrated in CAOS, a scalable and flexible framework, to test its performance. Our approach is compared to classic retrieval methods that only consider individual strategies by using a wide set of artificial and real-world datasets. This filtering approach is evaluated under large data volumes demonstrating its competence in clustering problems. Experiments support that the proposal overcomes the accuracy and significantly reduces the computational time of the solution retrieval achieved by the individual strategies.



## 1. Introduction

Clustering [39,15,32] is a trending data mining technique used in real-world situations to partition a data set into several groups according to some criteria and therefore identifying novel and potentially useful patterns from data. Conventional clustering algorithms are focused on obtaining groups by optimizing a single fitness function. In contrast, it can be difficult to obtain good data partitions in some real-world problems using a single objective function, and it is necessary to define several of them to obtain more accurate clusters [35]. These objective measures can be summarized in a single fitness function if they are disjoint. However, when the defined objectives conflict with each other it is necessary to define a fitness

* Corresponding author.
*E-mail addresses:* agarcia@ice.csic.es (A. Garcia-Piquer), andreus@salleurl.edu (A. Sancho-Asensio), albert.fornells@tsi.url.edu (A. Fornells), elisabet@salleurl.edu (E. Golobardes), guiomar@salleurl.edu (G. Corral), francesc.teixido@tsi.url.edu (F. Teixidó-Navarro).

http://dx.doi.org/10.1016/j.ins.2015.04.041
0020-0255/

function for each objective in order to find a solution which would give acceptable values for all of them [7]. A widely used technique to competently carry out this is multiobjective clustering (MC) [30], which uses the concept of Pareto Optimum with a posteriori approach [8] for simultaneously optimizing a set of mutually confronted objectives in order to promote the definition of clusters. This technique returns a collection that contains a number of Pareto optimal solutions (the so called Pareto set), none of which can be further improved on any objective without degrading another one [12].

There are different strategies for multiobjective optimization such as Simulated Annealing [47] and Ant Colony Optimization [37], but Multiobjective Evolutionary Algorithms (MOEAs) [7] have become one of the most capable strategies to solve this kind of problems [17,51] since they (1) work with a collection of solutions with different trade-offs among objectives, which are improved until a Pareto set with optimal trade-offs is obtained; (2) can be easily adapted to the type of data of the studied domain, due to the flexible knowledge representation used; and (3) are able to optimize different objectives without assuming any underlying structure of the objective functions. However, the performance of MOEAs can be compromised in large databases due to their high computational and memory usage requirements [19]. Moreover, one of the key challenges in Pareto-based MOEAs is the retrieval of the most suitable solution from the final Pareto set. This solution is typically identified by an expert in the domain. Nonetheless this process results in a subjective criterion and in a non trivial and tedious task if there are several solutions in the Pareto set. Thus, automatic methods are strongly required in order to help experts and simplify the identification of the most suitable solution, which can be beneficial in challenging domains such as health, smart networks or education. These are areas in which large volumes of data are generated.

In MC algorithms there are mainly two approaches to retrieve the most suitable solution from the Pareto set: (1) consider the shape of the Pareto set [43] or (2) consider the features related to the morphological properties of clusters [30]. The first method tries to identify the knee of the Pareto set to retrieve the solution with the best trade-off between objectives, but it does not take into account the resulting quality of clusters. The term quality is defined as how useful the solution is for the expert in the domain. Furthermore, quality is directly related to the shape, size and compactness of the clusters and the separation between them, characteristics which can be evaluated using clustering validation indexes [25,26,40]. The second method retrieves the best solution according to clustering validation indexes but its objective values could be unbalanced and the solution may only properly optimize a single objective.

The purpose of this paper is to propose a scalable retrieval filtering method that contemplates both the shape of the Pareto set and the quality of the clusters. The goal is to retrieve explanatory solutions with an acceptable trade-off between objectives in MC based on MOEAs. The proposed retrieval method is based on the observation that solutions with acceptable balance between objectives are placed around the knee of the Pareto front. The aim is to filter clustering solutions with less objective trade-off in order to retrieve the best solution from the remaining ones according to a clustering validation index. Thus, extra computations to evaluate non-interesting solutions are avoided. To test our approach we use the *Clustering Algorithm based on multiObjective Strategies* (CAOS) [10,22], a MC algorithm based on PESA-II [9]. CAOS uses a representation that does not depend on the number of instances of the data set, subsequently it is memory scalable [21]. Moreover, it scales the computational time of the clustering process by dividing the original data set to several subsets that are alternatively used in each generation of the MOEA process, thus it uses less data in each evolutionary cycle. This is performed in this way to avoid biasing the population by using only a single sample, while achieving low penalization in accuracy [2]. More specifically, the approach acts iteratively through the evolutionary cycle, being an automatic, adaptive system, thence fostering objectivity in the filtering parameters.

We compare the proposed method with the retrieval strategies based on (1) the shape of the Pareto set and (2) the morphological properties of clusters. All approaches are compared along a wide set of synthetic data sets [30] and real-world ones from the UCI [18] and KEEL [1] repositories. Furthermore, we carry out another set of experiments in data sets with large amounts of data in order to test the scalability capabilities of the method. Results show that accuracy and retrieval time are improved with this new proposal with a negligible additional cost to the evolutionary cycle. For a comparison between CAOS and other clustering methods, the reader is referred to [21].

The contributions of this paper are the following:

- It explores a filtering method that greatly increases the efficiency in retrieving solutions in two-objective clustering MOEAs.
- It integrates the proposed method in a scalable and flexible clustering framework.
- It tests the filtering method in a massive amount of data sets, including large ones.
- It shows a high performance in solution retrieval in both moderate and large data sets.
- It encourages practitioners to exploit the presented filtering technique to address the problem of retrieving the most suitable solutions from Pareto-based MOEAs.

The remainder of the paper is organized as follows. Section 2 briefly summarizes the related work on retrieving solutions in MC based on MOEAs. Section 3 introduces CAOS and describes the required modifications in order to adapt it to (1) become memory scalable and (2) the new filtering method. Section 4 describes the proposed retrieval method. Section 5 describes the experimentation and discusses the results. Finally, Section 6 ends with conclusions and further work.

## 2. Related work

Despite the huge popularity of MOEAs in the area of optimization due to their capabilities [7], there have been a few generic MC applications published in the literature. True MOEA-based MC algorithms did not appear until Handl and Knowles [28] introduced VIENNA. However, this algorithm needed to know the number of clusters in advance and did not provide any retrieval strategy from the Pareto set. To solve these issues, Handl and Knowles [30] proposed MOCK, the most well-known MC algorithm based on MOEAs. Another appealing approach is CAOS [10,22] which has a flexible configuration and allows a high degree of customization. The main differences of CAOS in respect of MOCK is that the former is scalable and its retrieval step is based on several cluster validation indexes.

A key aspect in Pareto Optimum MOEAs with a posteriori approach [8] lies in the identification of the fittest solution from the Pareto set at the end of the algorithm. An intuitive approach is to aggregate all the objectives into some kind of overall metric to sort the solutions, such as predicting the relative objective weighting [38]. Nevertheless, coming up with exact relative objective weights is a daunting task with complicated ramifications [44]. Other approaches are focused on ad-hoc methods. Those identify the desirable solution according to the specific domain of the problem [41] but they are not useful when the domain is not well-known and, unfortunately, this is the case of most real-world problems. Another kind of strategies not oriented to an specific domain consist in retrieving the solution according to the shape of the Pareto set by identifying the knee region or a solution in it [4,48,45,13].

In the specific case of the MC based on MOEAs, Handl and Knowles [30] proposed the use of the GAP statistic [49] to identify the most suitable solution with a good trade-off between objectives in the knee of the Pareto set. The main drawback of this technique lies in its high computational cost when applied to large data. To overcome this issue, Matake et al. [43] following the work of Branke et al. [4] proposed a technique based on the angle between solutions to find a clustering result in the knee of the Pareto set, and it was demonstrated that this technique improved the previous results. However, these techniques do not have into account the morphological characteristics of clusters, which is related to a poor explanation capacity for each cluster. On the other hand, Handl and Knowles [29] also proposed the use of some clustering validation indexes to retrieve the solution according to the properties of the clusters instead of taking into account the shape of the Pareto set. The main problem of doing this is that the validation indexes can return a solution that only properly optimizes a single objective, so the given result does not have a good trade-off between the desired objectives.

Dealing with the aforementioned issues are of the uppermost importance in many of today's industrial and scientific applications as these have increased dramatically the amount of data used and collected. Therefore, we investigate a reliable, accurate and scalable filtering method that tackles the drawbacks of MOEAs. The proposed retrieval method is based on the observation that solutions with acceptable balance between objectives are placed around the knee of the Pareto front. The aim is to filter clustering solutions with less objective trade-off in order to retrieve the best solution from the remaining ones according to a clustering validation index. Thus, extra computations to evaluate non-interesting solutions are avoided, which is an important aspect when dealing with large data.

Our contribution is focused on obtaining a solution with a balanced trade-off among the objectives to be optimized while getting also high quality clusters. For this reason, our proposal is based on combining the use of clustering validation indexes by filtering the solutions with less balanced objectives in order to obtain competitive clustering results. Notice that we do not propose to use a knee region identification algorithm but a filtering method based on the knee of a Pareto, whose mathematical foundation can be found at [48]. The main advantage of this proposal is that it is not sensitive to the type of Pareto front (concave or convex) and to the number of knee regions. Moreover, our aim is focused on obtaining a process able to improve the performance of the retrieval step when it is applied to large data. To carry out this, we introduce this approach into CAOS by (1) modifying its individual representation with a scalable one, and (2) modifying its learning process to work with data sampling with the aim of using less data in each evolutionary cycle. CAOS, the modifications done in it and the retrieval method are detailed in the following sections.

## 3. CAOS

In order to overcome some limitations of traditional clustering algorithms and to obtain high-quality clustering solutions, multiple criteria optimization is contemplated. It is focused on optimizing several objectives simultaneously by obtaining a collection of non-dominated solutions with different trade-offs among objectives called Pareto set. Recall that, in the field of multiobjective optimization, a solution $S$ is called non-dominated when there is not a single solution better than $S$ in regard to all the objectives. Otherwise the solution is called dominated. Thus, to obtain a final solution it is necessary to retrieve the most suitable solution from the Pareto set according to the problem to be solved. The purpose of this section is to describe CAOS, a multiobjective evolutionary algorithm specifically designed to solve clustering problems [22]. The system evolves a set of mutually non-dominated clustering solutions that correspond to different trade-offs between objectives. CAOS adopts PESA-II [9] as main basis due to its competitiveness and its ability to evolve accurate solutions from domains with complex structures [30].

In what follows, the knowledge representation used by CAOS is detailed. Next, the process organization of the algorithm is reviewed, placing special focus on the genetic operators that manipulate the representation. Finally, the data subsets method for computational scalability is depicted.

### 3.1. Knowledge representation

To successfully apply MOEAs to real-world problems it is important to choose a suitable individual representation according to the problem domain, because it defines the search space where solutions will be looked for. This has motivated many works focused on the analysis and design of several representations that have demonstrated their competitiveness [36]. CAOS uses a prototype-based representation due to (1) its search space exploring capacity and (2) its scale-up capabilities [21]. This representation is made up of real numbers which represent the coordinates of the cluster prototype (centroid) by means of its features. Therefore, each individual consists of $n \cdot t$ genes $\{x_{11}, \ldots, x_{1t}, \ldots, x_{n1}, \ldots, x_{nt}\}$, being $n$ the number of clusters described by the individual, $t$ the number of features of the data set, and $x_{ij}$ the value of the feature $j$ of the cluster centroid $i$. The genotypic representation is transformed into the phenotypic representation by assigning each instance to the cluster with the nearest centroid to it. Notice that CAOS normalizes the attribute values between 0 and 1. Several objective functions are used to validate the quality of individuals in MOEAs. These are detailed in what follows.

### 3.2. Objective functions

Two complementary optimization objective functions are used to measure the quality of a solution: (1) Deviation and (2) Connectivity. These objective functions are the most widely used [30] due to the fact that they indicate how nearby are the elements of each cluster (intra-cluster variance) and how separated are the clusters between them (inter-cluster variance), respectively. Deviation assesses the intra-cluster variance and it is computed as the overall summed distances between data items and their corresponding centroid. On the other hand Connectivity refers to the inter-cluster variance and it considers the degree to which data points that are close in the feature space have been placed in the same cluster. For the sake of brevity the reader is referred to [30,21] for more information about these objective functions.

### 3.3. Evolutionary process

CAOS evolves a population of individuals through a number of generations where individuals are selected, crossed and mutated following the typical evolutionary cycle [23]. Algorithm 1 presents the CAOS algorithm. Four aspects need further explanation to fully understand the genetic process that deals with the prototype-based representation: (1) the population initialization, (2) the selection operator, (3) the crossover operator, and (4) the mutation operator.

**Algorithm 1.** Scheme of CAOS algorithm.

```
Let EP be a external population which stores a maximum of N_EP individuals ;
Let IP be an internal population which stores N_IP individuals where N_IP < N_EP;
Let α_i, ∀i ∈ {1,2} be the angles, set to 0 in the beginning of the run
Initialize IP with N_IP individuals stochastically created;
Initialize the EP individuals with non-dominated clustering results from IP;
Evaluate all the individuals from EP according to the objectives ;
foreach Generation do
    Select N_IP individuals from EP to generate a new IP ;
    while (|IP| != ∅) do
        Select and remove two individuals from IP;
        Cross and mutate them to obtain 2 new individuals: I_New1 and I_New2 ;
        foreach I_Newi do
            Evaluate the I_Newi fitness according to the objectives ;
            if I_Newi dominates any individual from EP then
                Remove the dominated individuals by I_Newi from EP;
                Add I_Newi into EP;
            end
            else if I_Newi is not-dominated and I_Newi not-dominates any individual then
                if EP is full then
                    Remove an individual from the most crowded niche;
                end
                Add I_Newi into EP;
            end
        end
    end
    α_i, ∀i ∈ {1,2} ← ParetoFilter(α_i, ∀i ∈ {1,2}) (see Algorithm 2)
end
Select a individual from EP as a solution;
```

*3.3.1. Population initialization*

The population initialization step is responsible for filling the population with individuals that contain potentially worthy clusters. This approach uses an initialization based on medoids to define the initial prototypes, following the same idea as the *k*-means algorithm [31]. The process for each initial individual is the following:

(a) Select randomly a number $k$ of clusters between a minimum and a maximum value.
(b) Generate the individual by randomly choosing $k$ elements of the data set, where each one represents the prototype of a cluster.

*3.3.2. Selection operator*

This operator selects the individuals of the population used in each iteration. The population objective space is divided into hypercubes of equal size, creating an uniform hyper-grid and so each individual is mapped to one of these hypercubes according to its objective values. To select one individual, it chooses a non-empty niche from the population and selects randomly one of the individuals mapped into the chosen niche [23].

*3.3.3. Crossover operator*

Crossover mixes the genetic information of the selected individuals to obtain new potential solutions. In this case, a one-point crossover operator [23,24] is used to generate two offspring from pairs of parents. One point is selected for each parent and parts of the chromosomes are interchanged, taking into account that individuals should be cut at the same attribute but not necessarily at the same cluster.

*3.3.4. Mutation operator*

Mutation modifies a piece of the genetic information of an individual in order to explore new solutions. The probability $P_\mu$ determines when this operator is applied. To mutate the individuals, a cluster-oriented mutation operator [36] is used to promote the right search. This operator defines three different types of mutations and all of them have the same probability to be applied: (1) merge two clusters, (2) split a cluster, and (3) move the centroid of a cluster. The first mutation type merges a randomly selected cluster $s1$ with its nearest cluster $s2$, adding the new cluster centroid to the individual and erasing both original clusters. The new centroid is calculated with the weighted average between the original cluster centroids and the elements of each one. The second type splits a randomly selected cluster $s$ in two clusters $s1$ and $s2$. $s1$ is equal to $s$ and $s2$ is the most distant element $x$ from $s$ using the Euclidean distance. The last type of mutation moves the centroid of a randomly selected cluster $s$ by adding or subtracting a $delta_j$ value to each attribute. The range of each attribute has to be between the minimum and maximum value of the corresponding attribute in the data set. If it is out of the attribute range, the value is fixed to the corresponding maximum or minimum value.

Therefore, offspring could be inconsistent individuals with empty clusters after crossover and mutation operators. These clusters are eliminated from the individual to obtain a new individual where each cluster has at least one instance assigned.

*3.4. Data sampling*

MC algorithms based on MOEAs are costly in terms of computational time due to the huge amount of calculations required in large data sets, understanding computational time as the elapsed CPU time (in seconds). One way to improve the computational time is by using a subset of the available data to evaluate the individuals [2]. The main idea behind this strategy is to stratify the initial data set into disjoint data subsets (strata) of equal size and with equal class distribution [2,6], where the number of strata is selected by the user. However, in clustering problems the strata cannot be generated because classes are unknown. To avoid this limitation in CAOS, a random strata method is used to randomly assign the instances to each one of the strata. Moreover, in order to avoid the bias produced when only one random stratum is used, strata are alternated in each iteration of the evolutionary algorithm using a round-robin strategy [3]. Thus, if the stratum is changed in each cycle, the final individuals can generalize more than using a single strata.

Notice that the definition of the number of strata will influence in the algorithm performance. As the number of strata increases the computational time of the clustering process decreases but pattern extraction becomes more complex due to the lack of information. Furthermore, it must be emphasized that the retrieval process is computed using the complete data set, and this is an issue when working with large data sets. For this reason, a competent retrieval method is a practical approach.

## 4. Retrieving the most suitable solution

Retrieving a clustering solution based on the shape of the Pareto set can obtain a solution with a good trade-off among objectives but without any warranty about the morphological properties of clusters. This is depicted in Fig. 1, which shows (1) a Pareto front, (2) the solution retrieved according to the Pareto front shape and (3) the most suitable solution to be retrieved. In this figure, the solution identified in the knee of the Pareto front does not properly generalize. Therefore other solutions are more interesting from the point of view of the morphological properties of clusters, in order to provide useful knowledge. On the other hand, methods based only on clustering validation indexes can obtain the desirable solution

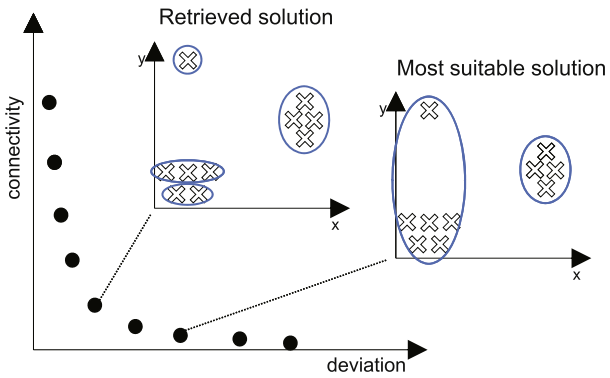



**Fig. 1.** Pareto front representation where the bullets are several non-dominated clustering solutions. The solution retrieved using a method based on identifying the knee of the Pareto front and the most suitable solution are identified. Their corresponding clusters are detailed.

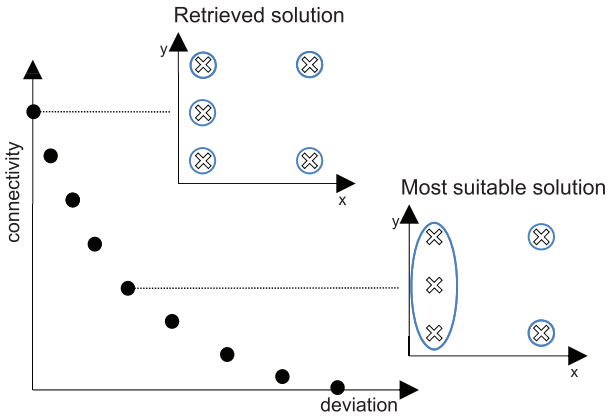



**Fig. 2.** Pareto front representation where the bullets are several non-dominated clustering solutions. The solution retrieved with clustering validation indexes and the most suitable solution are identified. Their corresponding clusters are detailed.

according to the quality of clusters, for instance by means of compactness, but they can be sensitive to outliers and to some specific shape of clusters that are unattractive as a solution. Fig. 2 shows a Pareto front jointly with (1) the solution identified by clustering validation indexes and (2) with the most suitable solution.

In this figure, the indexes select a solution with a bad trade-off between objectives, thus the solution given is not properly optimized and does not add any useful knowledge to experts. It must be emphasized that CAOS does not discard solutions according to the number of clusters that they contain, due to the fact that this is a subjective decision. Taking into consideration these aspects, the combination of both approaches for tackling both drawbacks can be an interesting win–win situation. The proposed hybrid approach is explained in what follows.

### *4.1. Retrieval method*

The proposed retrieval method filters the solutions that are in the boundaries of the Pareto set, because they barely take into account more than a single objective. Thus, solutions characterized by having very large or small clusters are discarded. The objective of the proposed technique is to apply clustering validation indexes to the remaining solutions of the Pareto set. For this reason, the indexes can obtain better results because the solutions with unbalanced objectives are discarded. The difficulty of this approach is to determine the solutions to be omitted. This issue is important because if the regions of solutions to be discarded are very large, some valuable solutions from the point of view of the quality of clusters will not be considered. On the other hand, if the regions are very small, the solutions that are not interesting from the point of view of clustering will be also considered. The identification of the regions to be discarded in a two-objective clustering problem is subsequently detailed.

### *4.2. Identification of the solutions to be discarded in a two-objective clustering problem*

The most useful objectives to promote the compactness and separation among clusters are Deviation and Connectivity [30] as they were described in Section 3. In two-objective optimization problems, the Pareto set can be represented in a two-dimensional graph where each axis correspond to each objective. The proposed method creates a hyperplane per objective to filter the solutions, in such a way that the regions outside the area comprised between the hyperplanes are discarded (that is,

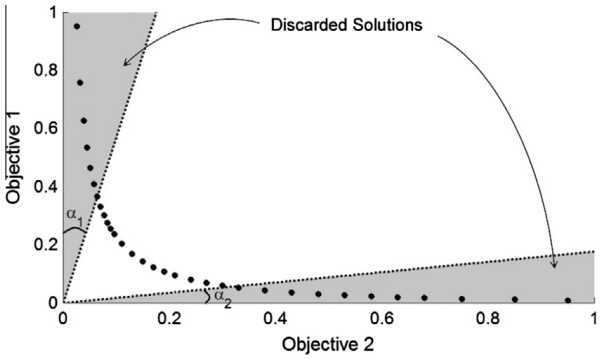



**Fig. 3.** Graphical representation of the regions of solutions to be discarded, so that the solutions (represented by bullets) in the gray areas are discarded. $\alpha_1$ and $\alpha_2$ are the angles that determine the hyperplanes (dashed lines) and consequently the size of the discarded regions.

the gray areas in Fig. 3). Each hyperplane is described by one angle $\alpha_i$ in regard to the corresponding objective axis. Thus, the size of the discarded regions is determined by each angle. If both angles are 0 degrees, no solution is discarded. It is important to highlight that both angles would not be equal or higher than 45 degrees due to the fact that the area between the hyperplanes cannot comprise any solution of the Pareto front. The angles $\alpha_1$ and $\alpha_2$ are calculated in the evolutionary process. Specifically, initially they start with 0 degrees and are adjusted in each iteration.This adjustment is calculated in two steps:

1. $\forall i \in \{1,2\}: \ \alpha'_i \leftarrow \alpha_i + rand(\delta_{min}, \delta_{max})$, where $\delta_{min}$ and $\delta_{max}$ are two user defined parameters in the range [0, 45).
2. Compare the regions between both hyperplanes with $\alpha_i$ and $\alpha'_i$ using a quality measure estimator. If the new angles $\alpha'_i$ define a better region, update $\alpha_i$ with them.

Algorithm 2 shows the complete process of the presented filtering method, which is called in each generation of the GA as Algorithm 1 indicates. The quality of the region delimited by the hyperplanes is averaged from the quality of a random subset of the solutions contained in it. This random subset can have a maximum size of $p_{max}$ solutions and a minimum size of $p_{min}$. These two parameters indicate the proportion of solutions of the Pareto set to be evaluated. If the minimum size is not achieved the current iteration does not update the angles with $\alpha'_i$. An approximative and fast measure to estimate the quality of each solution is proposed. This measure takes into account the overall compactness of the clusters and the overall distance among clusters for evaluating the quality of a solution. The Estimator is calculated as Eq. (1) shows, where $C$ is the clustering obtained; $n$ is the number of clusters; $d(x,y)$ is the Euclidean distance between the element $x$ and $y$; $C_i$ is the cluster $i$ and $v_i$ is its corresponding centroid; $m$ is the number of examples in the training data set; and $t$ is the number of features of the instances.

$$Estimator(C) = \frac{Comp(C)}{Dist(C)}, \ where \tag{1}$$

$$Comp(C) = 1 - \frac{\sum_{i=1}^{n}\sum_{x \in C_i} d(x, v_i)}{m \cdot t},$$

$$Dist(C) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d(v_i, v_j).$$

**Algorithm 2.** A high-level description of the proposed Pareto filter algorithm.

Let $\alpha_i, \forall i \in \{1,2\}$ be the angles that receive as parameters
Let $\delta_{min}$ and $\delta_{max}$ two user defined parameters $\in$ [0, 45)
Let $Q$ be the overall quality estimator of the filtered solution, set to 0 in the beginning of the run
Let $I_{sel}$ be the number of individuals selected between $\alpha_i$ for quality estimation
Let $p_{min}$ be the user-defined minimum population threshold
$\alpha'_i \leftarrow \alpha_i + rand(\delta_{min}, \delta_{max}), \forall i \in \{1,2\}$
$I_{sel} \leftarrow$SelectIndividualsBetweenAngles($\alpha'_i$)
**if** *The number of individuals in* $I_{sel}$ *is greater than* $p_{min}$ **then**
  $Q' \leftarrow$ ComputeEstimation ($I_{sel}$)  //Using Eq. (1)
  **if** $Q' > Q$ **then**
    $Q \leftarrow Q'$
    $\alpha'_i \leftarrow \alpha_i, \forall i \in \{1,2\}$
  **end**
**end**
return $\alpha'_i, \forall i \in \{1,2\}$

This process allows the system to filter the solutions during the evolutionary process without removing them from the population in order not to lose generalization capacity. It is worth noting that the process filters the non-interesting solutions, and it is not focused on identifying the knee region. Thus, the method is not sensitive to the type of Pareto front (concave or convex) and to the number of knee regions in it.

### 4.3. Clustering validation indexes selection

After discarding the non-interesting clustering solutions, it is necessary to select the most suitable one from this region according to cluster properties. Therefore, clustering validation indexes are used to achieve this by using a relative criteria method [25,27,40], which consists in comparing all the solutions among themselves and then selecting the fittest one. In the experimentation, the most known validation indexes were integrated into the framework. Those indexes are the following: (1) Adjusted Rand index [50], (2) Davies-Bouldin index [11], (3) Dunn's index [16], (4) Silhouette index [46] and (5) Calinski-Harabasz index [5]. Adjusted Rand Index is the supervised index of reference used. It retrieves the clustering solution from the Pareto set regarding to the original classes of the problem, it returns values between 0 and 1 and it should be maximized. Specifically, it compares two clustering results (the original one and the proposed as solution) counting the number of pairwise co-assignments of instances between them and introducing a statistically induced normalization in order to yield values close to 0 for random partitions (see Eq. (2)). In the equation, $n$ is the number of clusters of the evaluated solution $C$, $n_o$ is the number of the original classes of the data set $O$, $m$ is the number of instances of the data set, $m_{ij}$ is the number of data items that have been assigned to both class $i$ and cluster $j$, $m_{i.}$ is the number of instances assigned to class $i$ and $m_{.j}$ is the number of instances assigned to cluster $j$. The other four indexes are based on inherent information of the data set in order to obtain a solution with clusters of high quality. Each one of these indexes makes different calculations and they can return a different clustering solution from the collection of potential solutions, so the use of one index or another depends on the point of view of the expert. Having explained in detail the intrinsics of CAOS, in the next section, the different strategies are analyzed in a variety of experiments.

$$R(C,O)=\frac{\sum_{i=1}^{n_o}\sum_{j=1}^{n}\binom{m_{ij}}{2}-\left[\sum_{i=1}^{n_o}\binom{m_{i.}}{2}\cdot\sum_{j=1}^{n}\binom{m_{.j}}{2}\right]\Big/\binom{m}{2}}{\frac{1}{2}\left[\sum_{i=1}^{n_o}\binom{m_{i.}}{2}+\sum_{j=1}^{n}\binom{m_{.j}}{2}\right]-\left[\sum_{i=1}^{n_o}\binom{m_{i.}}{2}\cdot\sum_{j=1}^{n}\binom{m_{.j}}{2}\right]\Big/\binom{m}{2}} \tag{2}$$

## 5. Experiments, results and discussion

This section analyzes the performance of the retrieval strategies to select the most suitable solution using CAOS. First, 35 artificial data sets and 35 real-world data sets are analyzed. Specifically, the proposed filtering method is compared with respect to the technique presented by Matake [43] that is based on the shape of the Pareto set, and with respect to other strategies based on using clustering validation indexes to assess cluster quality. The technique based on adjacent angles proposed by Matake returns a solution in the knee of the Pareto front and has demonstrated a high degree of competitiveness. Moreover, another series of experiments applied to large data are performed using the same methodology in order to analyze the approaches behavior in this kind of data. In what follows, the experimental methodology and the results of the comparison are presented and discussed.

### 5.1. Experimental methodology

This section presents the experimental methodology followed in order to evaluate the performance of the different retrieval strategies to select the most suitable solution from the Pareto set found by CAOS. The analysis enables us to emphasize the benefits and the drawbacks of each one. In the followings, we provide details about (1) the data set collection chosen for the experimentation, (2) the CAOS configuration, and (3) the comparison metrics.

#### 5.1.1. Test bed

The experimentation is divided into two kinds of experiments. The first kind is oriented to non-large data sets and assess the algorithm performance using different typologies of artificial and real-world problems (see Table 1). First, 35 artificial data sets were selected according to different number of instances (from 900 to 2990), attributes (from 2 to 100) and classes (from 2 to 10). They were built using the tool presented by Handl and Knowles [30]. Also, 35 real-world problems were selected according to different number of instances (from 101 to 7494), attributes (from 3 to 60) and classes (from 2 to 11). The second kind of experiments uses large data for assessing the algorithms performance (see Table 2). Specifically, it uses 6 data sets with a number of instances between 19,000 to 581,012, a number of attributes from 9 to 54, and a number of classes between 2 and 26. All these data sets were obtained from the UCI [18], KEEL [1] and KDD [33] repositories.

**Table 1**
Summary of the characteristics of the 35 artificial data sets (left block) and real-world data sets (right block) used. The columns of each block are referred to the number of instances (nI), to the number of attributes (nA) and to the number of classes (nC).

| Data set | nI | nA | nC | Data set | nI | nA | nC |
|---|---|---|---|---|---|---|---|
| 100d-10c | 2198 | 100 | 10 | appendicitis | 106 | 7 | 2 |
| 100d-4c | 1218 | 100 | 4 | balance | 625 | 4 | 3 |
| 10d-10c | 2122 | 10 | 10 | biopn | 1027 | 24 | 2 |
| 10d-4c | 1092 | 10 | 4 | bpa | 345 | 6 | 2 |
| 2d-10c | 2990 | 2 | 10 | contraceptives | 1473 | 9 | 3 |
| 2d-4c | 1261 | 2 | 4 | crx | 690 | 15 | 2 |
| curves1 | 1000 | 2 | 2 | dermatology | 366 | 35 | 6 |
| curves2 | 1000 | 2 | 2 | echocardiogram | 132 | 12 | 2 |
| dartboard1 | 1000 | 2 | 4 | ecoli | 336 | 8 | 8 |
| dartboard2 | 1000 | 2 | 4 | glass | 214 | 9 | 6 |
| donut1 | 1000 | 2 | 2 | haberman | 306 | 3 | 2 |
| donut2 | 1000 | 2 | 2 | heart-statlog | 270 | 13 | 2 |
| donut3 | 999 | 2 | 3 | hepatitis | 155 | 19 | 2 |
| donutcurves | 1000 | 2 | 4 | housevotes | 435 | 16 | 2 |
| long1 | 1000 | 2 | 2 | ionosphere | 351 | 34 | 2 |
| long2 | 1000 | 2 | 2 | iris | 150 | 4 | 3 |
| long3 | 1000 | 2 | 2 | liver-disorders | 345 | 6 | 2 |
| longsquare | 900 | 2 | 6 | mammographic | 961 | 5 | 2 |
| sizes1 | 1000 | 2 | 4 | pendigits | 7494 | 17 | 10 |
| sizes2 | 1000 | 2 | 4 | pim | 768 | 8 | 2 |
| sizes3 | 1000 | 2 | 4 | segment | 2310 | 19 | 7 |
| sizes4 | 1000 | 2 | 4 | sonar | 208 | 60 | 2 |
| sizes5 | 1000 | 2 | 4 | tae | 151 | 5 | 3 |
| smile1 | 1000 | 2 | 4 | thyroids | 215 | 5 | 2 |
| smile2 | 1000 | 2 | 4 | transfusion | 748 | 4 | 2 |
| smile3 | 1000 | 2 | 4 | vehicle | 846 | 18 | 4 |
| spiral | 1000 | 2 | 2 | vertebral | 310 | 6 | 3 |
| spiralsquare | 1500 | 2 | 6 | vowel | 990 | 13 | 11 |
| square1 | 1000 | 2 | 4 | waveform | 5000 | 40 | 3 |
| square2 | 1000 | 2 | 4 | wdbc | 569 | 30 | 2 |
| square3 | 1000 | 2 | 4 | wine | 178 | 13 | 3 |
| square4 | 1000 | 2 | 4 | wisconsin | 699 | 9 | 2 |
| square5 | 1000 | 2 | 4 | wpbc | 198 | 33 | 2 |
| triangle1 | 1000 | 2 | 4 | yeast | 1484 | 9 | 10 |
| triangle2 | 1000 | 2 | 4 | zoo | 101 | 16 | 7 |

**Table 2**
Summary of the characteristics of the 35 large data sets used. The columns of each block are referred to the number of instances (nI), to the number of attributes (nA) and to the number of classes (nC).

| Data set | nI | nA | nC | Data set | nI | nA | nC |
|---|---|---|---|---|---|---|---|
| covtype | 581012 | 54 | 7 | letter | 20000 | 16 | 26 |
| kddcup | 494021 | 41 | 23 | magic | 19022 | 10 | 2 |
| census | 299324 | 41 | 2 | 2d-20c-125m | 16097 | 2 | 20 |
| shuttle | 58000 | 9 | 7 | 5d-20c-175m | 15675 | 5 | 20 |
| 10d-30c-175m | 23898 | 10 | 30 | 20d-20c-125m | 15508 | 20 | 20 |
| 10d-30c-75m | 23471 | 10 | 30 | 2d-20c-75m | 15012 | 2 | 20 |
| 5d-30c-75m | 23234 | 5 | 30 | 20d-20c-175m | 14970 | 20 | 20 |
| 100d-30c-175m | 22788 | 100 | 30 | 10d-20c-75m | 14830 | 10 | 20 |
| 20d-30c-75m | 22470 | 20 | 30 | 20d-20c-75m | 14491 | 20 | 20 |
| 2d-30c-175m | 22229 | 2 | 30 | 5d-20c-125m | 14261 | 5 | 20 |
| 5d-30c-125m | 22038 | 5 | 30 | 10d-20c-175m | 14023 | 10 | 20 |
| 10d-30c-125m | 21974 | 10 | 30 | 10d-20c-125m | 13875 | 10 | 20 |
| 2d-30c-125m | 21846 | 2 | 30 | 100d-20c-75m | 13790 | 100 | 20 |
| 20d-30c-175m | 21491 | 20 | 30 | 100d-20c-125m | 13702 | 100 | 20 |
| 5d-30c-175m | 21129 | 5 | 30 | 100d-20c-175m | 13421 | 100 | 20 |
| 20d-30c-125m | 20986 | 20 | 30 | 2d-20c-175m | 13355 | 2 | 20 |
| 2d-30c-75m | 20370 | 2 | 30 | 5d-20c-75m | 13289 | 5 | 20 |
| 100d-30c-125m | 20156 | 100 | 30 | | | | |

### 5.1.2. CAOS configuration

CAOS was run with 50 different random seeds with the synthetic and the real-world problems and with 20 different random seeds with the large data sets. The system was configured using the following parameters (the author is referred to [22] for notation details): $\ell$ was 5% of the number of data set instances, the maximum size of the initial population was 100, $N_{EP}$

was 1000, $N_{IP}$ was 50, $N_{niches}$ was 5, the number of generations was 400, the probability of crossover ($P_c$) was set to 0.7 and the probability of mutation ($P_\mu$) was set to $1/m$. The filtering method was configured with the next parameters: $\delta_{min}$ was 0.1, $\delta_{max}$ was 0.75, $p_{min}$ was 50% of the numbers of solutions in the Pareto set and $p_{max}$ was 10% of them. As we are interested in robust systems that perform competently on average, the same configuration was used for all the data sets. To set these parameters to their optimal values, the iterated F-Race procedure [42] was followed. Moreover, the experiments done with large data sets use data sampling as Section 3.4 explains. Each data set has been divided in four strata (i.e., each stratum contains a 25% of the instances of the original data set). The reader is referred to [2] for more information about this issue.

#### 5.1.3. Retrieval strategies analyzed

The proposed filtering technique was applied with some of the most used clustering validation indexes such as Davies, Dunn, Silhouette and Calinski-Harabasz. Next, these results were compared with the ones obtained with the same clustering validation indexes and the adjacent angles approach using the overall Pareto set. In addition to these strategies, we also contemplated the best solution from the overall Pareto set according to the Adjusted Rand index [50]. It must be emphasized that the Adjusted Rand index is based on obtaining the best solution according to a prespecified structure of the data set, in our case, the classes assigned to each instance–that are known in benchmark problems–. This strategy is used to compare our proposal with the ideal solution.

#### 5.1.4. Comparison metrics

The accuracy of each solution was quantified using the Adjusted Rand index in order to evaluate them according to the original classes of the problems. The recommendations pointed out by [14] were followed to perform the statistical analysis of the accuracy results, which is based on the use of nonparametric tests. More specifically, the following methodology was employed. First, the Friedman test [20] was applied to contrast the null hypothesis that all the learning algorithms obtained the same results on average. If the Friedman test rejects the null hypothesis, we perform pair-wise comparisons by means of the Holm's step-down procedure [34]. Following this procedure, we distinguish pairs of retrieval strategies that are significantly different in performance.

### 5.2. Massive comparison in non-large data sets

The analysis of the performance among all the strategies using the overall Pareto set and the proposed filtering method was carried out with all the presented data sets. Table 3 shows the results using a pairwise comparison by means of Holm's procedure. In it, the strategy used to retrieve the most suitable solution is indicated by $Dv, Dn, Sl, CH$ for the Davies, Dunn, Silhouette and Calinski-Harabasz indexes respectively. Also, the symbols of each strategy are preceded by an *F* when the filtering method is used and by an *A* when the overall Pareto set is used. Moreover, *AA* indicates the adjacent angles strategy

**Table 3**
Pairwise comparison of all the strategies in non-large data sets in respect of (a) Davies index, (b) Dunn's index, (c) Silhouette index and (d) Calinski-Harabasz index. $Dv, Dn, Sl, CH$ represent the results of the Davies, Dunn, Silhouette and Calinski-Harabasz indexes respectively. Also, the symbols of each strategy are preceded by an *F* when the filtering method is used and by an *A* when the overall Pareto set is used. Moreover, *AA* indicates the adjacent angles strategy and *AR* the supervised solution retrieved with the Adjusted Rand index, which only takes into account the overall data set. The symbols $\oplus$ and $\ominus$ show that the method in the row obtained results that were significantly higher/lower than those obtained with the method in the column at $\alpha = 0.05$. Similarly, the symbols $+$ and $-$ denote a non-significant higher/lower results. The last column shows the Friedman rank, where the minimum value indicates the best rank.

| | AR | AA | ADv | FDv | Friedman |
|---|---|---|---|---|---|
| *(a)* | | | | | |
| AR | | | | | 1.29 |
| AA | $\ominus$ | | | | 3.26 |
| ADv | $\ominus$ | $-$ | | | 3.91 |
| FDv | $\ominus$ | $\oplus$ | $+$ | | 2.54 |
| *(b)* | | | | | |
| AR | | | | | 1.22 |
| AA | $\ominus$ | | | | 3.05 |
| ADn | $\ominus$ | $-$ | | | 3.06 |
| FDn | $\ominus$ | $+$ | $+$ | | 2.65 |
| *(c)* | | | | | |
| AR | | | | | 1.28 |
| AA | $\ominus$ | | | | 3.24 |
| ASl | $\ominus$ | $-$ | | | 2.97 |
| FSl | $\ominus$ | $\oplus$ | $+$ | | 2.51 |
| *(d)* | | | | | |
| AR | | | | | 1.24 |
| AA | $\ominus$ | | | | 3.19 |
| ACH | $\ominus$ | $-$ | | | 2.88 |
| FCH | $\ominus$ | $+$ | $+$ | | 2.70 |

and *AR* the supervised solution retrieved with the Adjusted Rand index, which only takes into account the overall data set. The symbols $\oplus$ and $\ominus$ show that the method in the row obtained results that were significantly higher/lower than those obtained with the method in the column at $\alpha = 0.05$. Similarly, the symbols $+$ and $-$ denote a non-significant higher/lower results. Likewise, the Friedman rank value of each method is indicated, where lower values are better.

Table 3 summarizes the results obtained with the proposed filtering method ($FD\nu, FDn, FSl$ and $FCH$) and the results obtained according only to (1) the morphological characteristics of clusters ($AD\nu, ADn, ASl$ and $ACH$), (2) the shape of the Pareto front ($AA$) and (3) the supervised solution ($AR$). It can be observed that the solutions obtained with the filtered method are better ranked in terms of accuracy than the solutions obtained using the morphological properties of clusters from the overall Pareto set. Moreover, the filtered method solutions are better, and in some cases significantly better (i.e., Davies and Silhouette indexes) than the solution which only takes into account the shape of the Pareto front. In regard to the supervised solution, it is obvious that it is always significantly better than the other strategies due to the fact that it considers the original classes of the data set.

Furthermore, the proposed filtering method improves the computational time of the retrieval step due to the fact that the non-interesting solutions are not analyzed. This is depicted in Fig. 4, where the computational time of the retrieval step is shown for both filtered and overall Pareto set for each data set. We considered as retrieval step the process that selects the solutions according to each one of the strategies for each method. Particularly, the computational time of the retrieval step is calculated as the average time of all the strategies. Recall that filtering methods are always faster than the non-filtering ones and, in some data sets, the speedup can become faster in an order of magnitude. It is worth noting that the time required for computing the hyperplanes of the filtering method is not expensive. The evolutionary cycle only adds in average an extra 4.12% ± 4.73 of computational cost but, in return, the retrieval step time is reduced on average a 89.53% ± 7.79. In respect of the angles needed for building the hyperplanes, on average they take small values ($\alpha_1 = 3.82 \pm 2.86$ and $\alpha_2 = 3.57 \pm 2.94$). Thus, not a huge quantity of solutions is discarded, just only the solutions that are in the extremes of the Pareto front.

An interesting observation lies in the result of applying the filtering technique to Pareto fronts with concave shapes or with discontinuities. Because this method does not assume a particular shape or continuity in the Pareto front, and because it only filters non-interesting solutions, it can be safely applied to any kind of MC problem. Fig. 5 shows four cases of Pareto fronts with these features. It can be observed that the filtered regions do not consider the solutions with a bad trade-off among objectives. Thus, they help clustering validation indexes to avoid the problem of obtaining solutions far from the knee of the Pareto front. It is important to highlight that in the Pareto front there are not solutions with a very high value of the Deviation objective due to the fact that the genetic operators defined in the evolutionary algorithm tend to obtain a reasonable number of clusters. For example, the maximum value of the Deviation is achieved if all the elements of the data set are in a different cluster. Thus, hyperplane in the Deviation objective area filters few solutions in comparison with the hyperplane of the Connectivity area.

### 5.3. Comparison in large data sets

In order to analyze the performance of the presented method with large data, similar experiments to the ones in the previous section are carried out but using large data sets. Table 4 summarizes the results obtained by means of the Holm's procedure using the aforementioned nomenclature. It can be observed that solutions obtained with the filtered method are

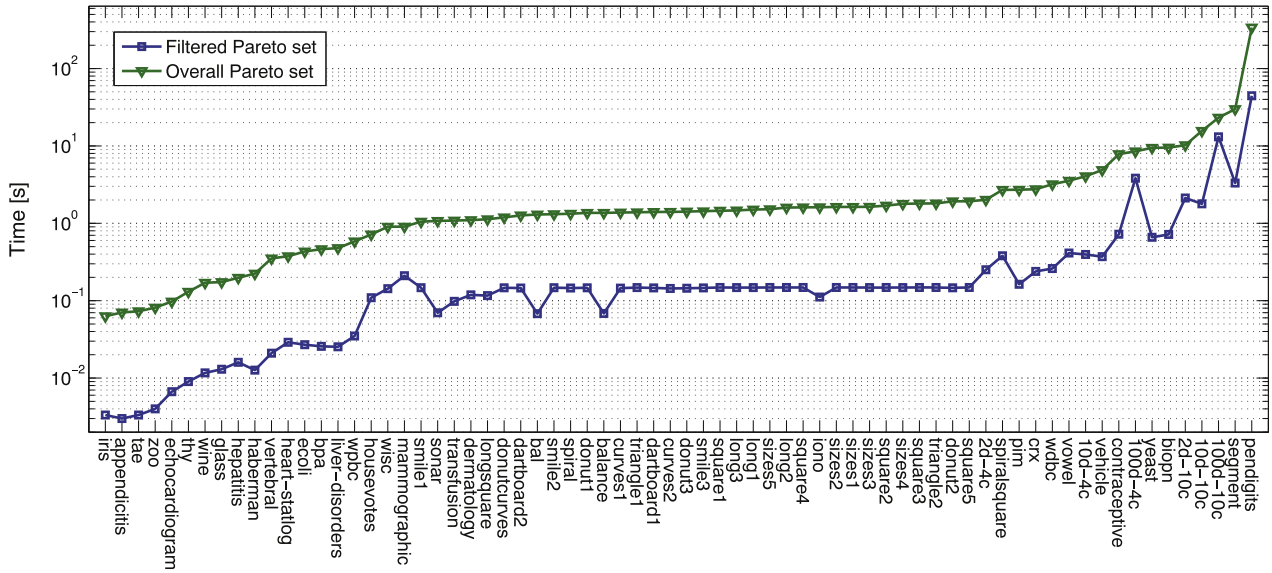



**Fig. 4.** Average computational time of the retrieval step in seconds for each one of the non-large data sets of (1) filtered Pareto sets retrieval strategies and (2) overall Pareto set retrieval strategies. Notice the logarithmic scale of time axis. Results are averages of ten runs.

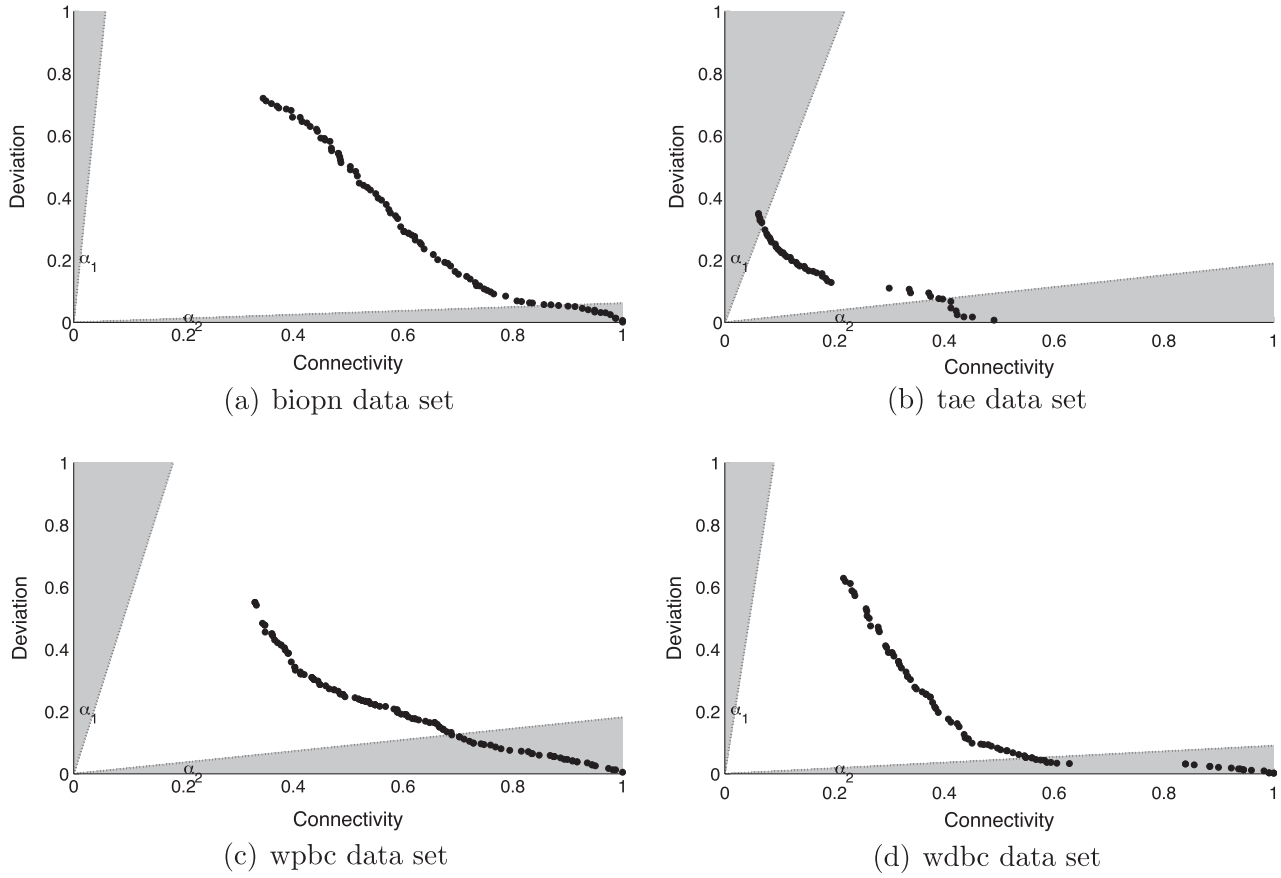



**Fig. 5.** Examples of the filtering method applied to problems with a complex Pareto front. These examples come from (a) the biopsia problem, (b) the tae problem, (c) the wisconsin problem and (d) the wdbc problem.

**Table 4**
Pairwise comparison of all the strategies in large data sets in respect of (a) Davies index, (b) Dunn's index, (c) Silhouette index and (d) Calinski-Harabasz index. $Dv, Dn, Sl, CH$ represent the results of the Davies, Dunn, Silhouette and Calinski-Harabasz indexes respectively. Also, the symbols of each strategy are preceded by an *F* when the filtering method is used and by an *A* when the overall Pareto set is used. Moreover, *AA* indicates the adjacent angles strategy and *AR* the supervised solution retrieved with the Adjusted Rand index, which only takes into account the overall data set. The symbols $\oplus$ and $\ominus$ show that the method in the row obtained results that were significantly higher/lower than those obtained with the method in the column at $\alpha = 0.05$. Similarly, the symbols $+$ and $-$ denote a non-significant higher/lower results. The last column shows the Friedman rank, where the minimum value indicates the best rank.

| | AR | AA | ADv | FDv | Friedman |
|---|---|---|---|---|---|
| *(a)* | | | | | |
| AR | | | | | 1.16 |
| AA | $\ominus$ | | | | 2.94 |
| ADv | $\ominus$ | $-$ | | | 3.37 |
| FDv | $\ominus$ | $+$ | $\oplus$ | | 2.53 |
| *(b)* | | | | | |
| AR | | | | | 1.17 |
| AA | $\ominus$ | | | | 2.93 |
| ADn | $\ominus$ | $-$ | | | 3.31 |
| FDn | $\ominus$ | $+$ | $+$ | | 2.59 |
| *(c)* | | | | | |
| AR | | | | | 1.13 |
| AA | $\ominus$ | | | | 3.04 |
| ASl | $\ominus$ | $-$ | | | 3.17 |
| FSl | $\ominus$ | $+$ | $+$ | | 2.66 |
| *(d)* | | | | | |
| AR | | | | | 1.13 |
| AA | $\ominus$ | | | | 3.04 |
| ACH | $\ominus$ | $-$ | | | 3.17 |
| FCH | $\ominus$ | $+$ | $+$ | | 2.66 |

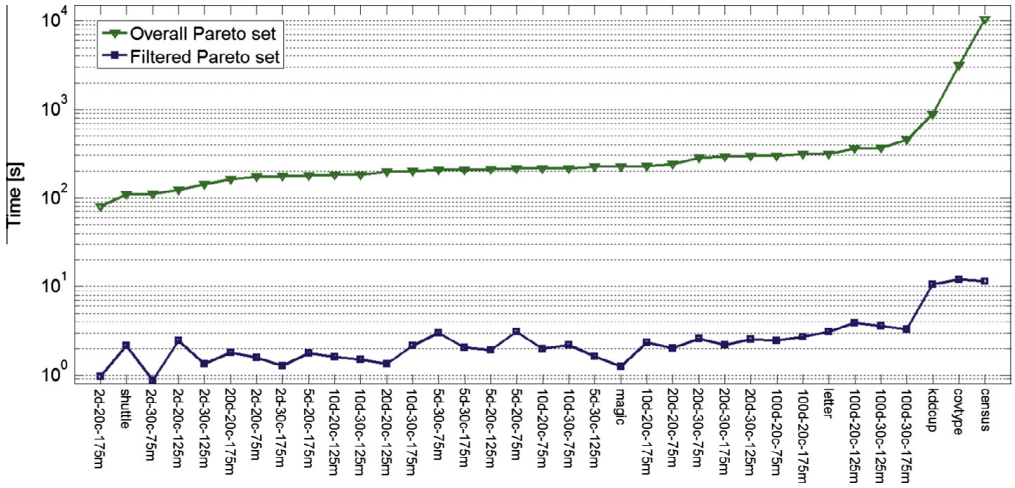



**Fig. 6.** Average computational time of the retrieval step in seconds for each one of the large data sets of (1) filtered Pareto sets retrieval strategies and (2) overall Pareto set retrieval strategies. Notice the logarithmic scale of time axis. Results are averages of ten runs.

better ranked in terms of accuracy than the solutions obtained using the clustering validation indexes from the overall Pareto set. Moreover, the majority of the filtered results are not significantly different than the supervised results. Nevertheless, they are slightly behind the ones of the strategy based only in the shape of the Pareto set.

In terms of retrieval time, as it is depicted in Fig. 6, the proposed filtering method highly improves the computational time, on average, in three orders of magnitude, being reduced in a $98.79\% \pm 0.03$. It is worth mentioning that the time required for computing the hyperplanes slightly increments the time of the evolutionary algorithm in $8.43\% \pm 6.07$. Regarding to the angles needed for building the hyperplanes, they take small values on average ($\alpha_1 = 4.51 \pm 2.00$ and $\alpha_2 = 4.98 \pm 2.14$), so only the solutions with a bad trade-off between objectives are discarded.

## 6. Conclusions and further work

The solution returned by a Pareto-based MOEA is a Pareto set of non-dominated solutions in which none of those solutions can be further improved on any objective without degrading the other ones. Although there is not a winner solution according to all the optimizing objectives, the most suitable solution to solve a specific problem can be manually retrieved with the help of an expert. This has motivated the necessity of proposing methods for automatically retrieving the most suitable solution, specially in the case of large volumes of data. In the case of MC, these methods usually select the solution in regard to (1) the shape of the Pareto set, which correspond to the value of the objectives to optimize, or (2) the quality of the solutions conforming to specific characteristics of the problem. The main drawback of the first method is that it retrieves a solution without taking into account the morphological characteristics of clusters and it can return a solution with a good trade-off between objectives but with poor quality clusters (i.e., non useful for expert proposals). On the other hand, the second method retrieves a solution according to the quality and shape of clusters using clustering validation indexes but it does not consider the value of the objectives, so it can return a solution with an inadequate trade-off between them. For these reasons, we proposed the combination of both methods to obtain a new hybrid mechanism which filters and selects a solution according to a clustering validation index from the region of the Pareto set where all the solutions with a good trade-off between objectives are placed. Moreover, this filtering technique can be applied to any kind of Pareto-based MOEA.

The proposed filtering method was analyzed using several clustering validation indexes in both large and non-large data sets. Traditional approaches were also included in the analysis in order to compare the results. To carry out the experimentation, CAOS algorithm was used to build the Pareto set with clustering solutions. Experiments show that, in the case of non-large data sets, the proposed filtering technique is the most accurate and the one that requires less computation. Furthermore, the proposed method can obtain solutions that are not significantly different to the solutions retrieved by a supervised method, so they work as well as a method that uses the classes of the problem to retrieve the best solution. In the case of large data, the results highlight a huge improvement in the retrieval step time without losing generalization capacity, demonstrating that the proposed technique is memory scalable and useful to tackle large data sets. It must be emphasized that the performance of the filtering technique does not depend on the way that the Pareto set is built, so the obtained results are not influenced by the CAOS algorithm.

Clustering is not focused on classifying a data set according to a specified structure and, consequently, the morphological properties of the obtained clusters are key to understand the proposed patterns. It is for this reason that the solutions retrieved with clustering validation indexes consider this issue and if they are obtained from the filtered Pareto set, the solutions consider an acceptable trade-off between objectives–the aim of MC.

As future work we are working on analyzing the effects of using other retrieval strategies and the application of the filtering technique to Pareto sets with more than two objectives.

## References


[1] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework, Multiple-Valued Logic Soft Comput. 17 (2–3) (2011) 255–287.
[2] J. Bacardit, Pittsburgh genetic-based machine learning in the data mining era: representations, generalization, and run-time, Ph.D. thesis, Enginyeria i Arquitectura La Salle, Universitat Ramon Llull, Barcelona, Spain, 2004.
[3] J. Bacardit, X. Llorà, Large scale data mining using genetics-based machine learning, in: Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers, GECCO '09, ACM, 2009, pp. 3381–3412.
[4] J. Branke, K. Deb, H. Dierolf, M. Osswald, Finding knees in multi-objective optimization, in: 8th Conference on Parallel Problem Solving from Nature (PPSN VIII), Lecture Notes in Computer Science, Springer-Verlag, 2004, pp. 722–731.
[5] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. Simul. Comput. 3 (1) (1974) 1–27.
[6] J.R. Cano, S. García, F. Herrera, Subgroup discovery in large size data sets preprocessed using stratified instance selection for increasing the presence of minority classes, Pattern Recogn. Lett. 29 (2008) 2156–2164.
[7] C.A. Coello, A comprehensive survey of evolutionary-based multiobjective optimization techniques, Knowl. Inform. Syst. 1 (1999) 269–308.
[8] C.A. Coello, G.B. Lamont, D.A.V. Veldhuizen, Evolutionary Algorithms for Solving Multi-objective Problems, Springer-Verlag New York, Inc, 2007.
[9] D.W. Corne, N.R. Jerram, J.D. Knowles, M.J. Oates, PESA-II: region-based selection in evolutionary multiobjective optimization, in: Proceedings of the Genetic and Evolutionary Computation Conference, Morgan Kaufmann Publishers, 2001, pp. 283–290.
[10] G. Corral, A. Garcia-Piquer, A. Orriols-Puig, A. Fornells, E. Golobardes, Analysis of vulnerability assessment results based on CAOS, Appl. Soft Comput. J. 11 (2011) 4321–4331.
[11] D. Davies, D. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Machine Intell. 1 (4) (1979) 224–227.
[12] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms, John Wiley and Sons, Ltd, England, 2001.
[13] K. Deb, S. Gupta, Understanding knee points in bicriteria problems and their implications as preferred solution principles, Eng. Optim. 43 (11) (2011) 1175–1204.
[14] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Machine Learn. Res. 7 (2006) 1–30.
[15] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley and Sons, Inc., New York, 2000.
[16] J.C. Dunn, Well separated clusters and optimal fuzzy partitions, J. Cybernet. 4 (1974) 95–104.
[17] C.M. Fonseca, P.J. Fleming, An overview of evolutionary algorithms in multiobjective optimization, Evol. Comput. 3 (1995) 1–16.
[18] A. Frank, A. Asuncion, UCI machine learning repository, 2010. <http://archive.ics.uci.edu/ml>.
[19] A.A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2002.
[20] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Annals Math. Stat. 11 (1940) 86–92.
[21] A. Garcia-Piquer, A. Fornells, J. Bacardit, A. Orriols-Puig, E. Golobardes, Large-scale experimental evaluation of cluster representations for multiobjective evolutionary clustering, IEEE Trans. Evol. Comput. 18 (1) (2014) 36–53.
[22] A. Garcia-Piquer, A. Fornells, A. Orriols-Puig, G. Corral, E. Golobardes, Data classification through an evolutionary approach based on multiple criteria, Knowl. Inform. Syst. 33 (1) (2012) 35–56.
[23] D.E. Goldberg, Genetic Algorithm in Search, Optimization, and Machine Learning, Addison-Wesley, Inc., Boston, MA, USA, 1989.
[24] D.E. Goldberg, The Design of Innovation, Kluwer Academic Publishers, Massachusetts, US, 2002.
[25] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inform. Syst. 17 (2001) 107–145.
[26] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Cluster validity methods: Part i, SIGMOD Record 31 (2) (2002) 40–45.
[27] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Clustering validity checking methods: Part ii, SIGMOD Record 31 (3) (2002) 19–27.
[28] J. Handl, J. Knowles, Evolutionary multiobjective clustering, Lecture Notes Comput. Sci. (2004) 1081–1091.
[29] J. Handl, J. Knowles, Improvements to the scalability of multiobjective clustering, in: The 2005 IEEE Congress on Evolutionary Computation, vol. 3, September 2005, pp. 2372–2379.
[30] J. Handl, J. Knowles, An evolutionary approach to multiobjective clustering, IEEE Trans. Evol. Comput. 1 (1) (2007) 56–76.
[31] J. Hartigan, M. Wong, A k-means clustering algorithm, in: Applied Statistics, 1979, pp. 28:100–108.
[32] F. Herrera, C. Carmona, P. Gonzlez, M. del Jesus, An overview on subgroup discovery: foundations and applications. Knowledge and Information Systems, 2010. doi: http://dx.doi.org/10.1007/s10115-010-0356-2.
[33] S. Hettich, S.D. Bay, The UCI KDD archive university of california, 1999. <http://kdd.ics.uci.edu>.
[34] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (1979) 65–70.
[35] E. Hruschka, R.J.G.B. Campello, A. Freitas, A.C.P.L.F. De Carvalho, A survey of evolutionary algorithms for clustering, IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev. 39 (2) (2009) 133–155.
[36] E.R. Hruschka, R.J.G.B. Campello, A.A. Freitas, A.C.P.L.F. de Carvalho, A survey of evolutionary algorithms for clustering, IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev. 39 (2) (2009) 133–155.
[37] S. Iredi, D. Merkle, M. Middendorf, Bi-criterion optimization with multi colony ant algorithms, in: Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization (EMO 2001), Springer, 2000, pp. 359–372.
[38] E.M. Kasprzak, K.E. Lewis, Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method, Struct. Multidisci. Optim. 22 (3) (2001) 208–218.
[39] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990.
[40] C. Le.gny, S. Juh. sz, A. Babos, Cluster validity measurement techniques, in: Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, 2006, pp. 388–393.
[41] Y. Liu, M. Yoshioka, K. Homma, T. Shibuya, Efficiently finding the 'best' solution with multi-objectives from multiple topologies in topology library of analog circuit, in: Proceedings of the 2009 Asia and South Pacific Design Automation Conference, ASP-DAC '09, IEEE Press, 2009, pp. 498–503.
[42] M. López-Ibáñez, J. Dubois-Lacoste, T. Stützle, M. Birattari, The irace Package: Iterated Racing for Automatic Algorithm Configuration, Tech. Rep. TR/IRIDIA/2011-004, IRIDIA, Université Libre de Bruxelles, Belgium, 2011.
[43] N. Matake, T. Hiroyasu, M. Miki, T. Senda, Multiobjective clustering with automatic k-determination for large-scale data, in: GECCO '07: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, ACM, 2007, pp. 861–868.
[44] A. Messac, G.J. Sundaraj, R.V. Tappeta, J.E. Renaud, Ability of objective functions to generate points on nonconvex Pareto frontiers, AIAA J. 38 (6) (2000) 1084–1091.
[45] L. Rachmawati, D. Srinivasan, Multiobjective evolutionary algorithm with controllable focus on the knees of the pareto front, IEEE Trans. Evol. Comput. 13 (4) (2009) 810–824.
[46] P.J. Rousseew, Silhouttes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[47] S. Saha, S. Bandyopadhyay, A new multiobjective clustering technique based on the concepts of stability and symmetry, Knowl. Inform. Syst. 23 (2010) 1–27.
[48] O. Schütze, M. Laumanns, C.A. Coello, Approximating the knee of an mop with stochastic search algorithms, in: Proceedings of the 10th International Conference on Parallel Problem Solving from Nature: PPSN X, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 795–804.
[49] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a dataset via the gap statistic, J. Roy. Stat. Soc. 63 (2000) 411–423.
[50] K. Yeung, W. Ruzzo, Details of the adjusted rand index and clustering algorithms. supplement to the paper "an empirical study on principal component analysis for clustering gene expression data", Science 17 (9) (2001) 763–774.
[51] E. Zitzler, K. Deb, L. Thiele, Comparison of multiobjective evolutionary algorithms: empirical results, Evol. Comput. 8 (2) (2000) 173–195.

# Promoting consensus in the concept mapping methodology: An application in the hospitality sector. *Pattern Recognition Letters*, 2015

A. Fornells, Z. Rodrigo, R. Santoma, X. Rovira, M. Sanchez, F. Teixido and E. Golobardes. Promoting consensus in the concept mapping methodology: An application in the hospitality sector. Pattern Recognition Letters, doi:10.1016/j.patrec.2015.05.013, 2015. (Publicació indexada al Q3 amb factor d'impacte 1,062)

Pattern Recognition Letters 000 (2015) 1–10

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Promoting consensus in the concept mapping methodology: An application in the hospitality sector ☆

Albert Fornells [a,*], Zaida Rodrigo [a], Xari Rovira [c], Mónica Sánchez [d], Ricard Santomà [a], Francesc Teixidó-Navarro [a,b], Elisabet Golobardes [b]

[a] Research Group in Hospitality, Tourism and Mobilities, School of Tourism and Hospitality Management Sant Ignasi, Universitat Ramon Llull. Av. Marquès de Mulhacén 40–42, Barcelona 08034, Spain
[b] GR-SETAD, La Salle, Universitat Ramon Llull. Av. Quatre Camins 30, Barcelona 08022, Spain
[c] ESADE Business School, Universitat Ramon Llull. Av. Pedralbes 62, Barcelona 08034, Spain
[d] Universitat Politècnica de Catalunya. UPC-Barcelona Tech., Jordi Girona, 1–3, Barcelona 08034, Spain

ARTICLE INFO

*Article history:*
Available online xxx

*Keywords:*
Concept mapping methodology
Qualitative reasoning techniques
Consensus measures
Excellence in hospitality

ABSTRACT

The concept mapping methodology aims to respond to the non trivial task of conceptualising abstract thoughts by means of a focus group composed by experts from the studied domain. The approach defines a set of general steps that allow experts to lead the generation of ideas, group the ideas in a conceptual map of interrelated concepts using clustering multidimensional scaling and clustering techniques, analysing the quality of the conceptual maps and deciding on a final interpretation. In this sense, this final decision is not trivial because clustering techniques provide a set of potentially conceptual maps so experts must select the one that fits best according to their opinion. For this reason, we present the global index of consensus as an indicator for filtering the most suitable clustering solutions using qualitative reasoning. It promotes the consensus of experts opinions and ensures objectivity in the final interpretation. The index outperforms three of the most well-known clustering validation indexes in a case study focused on the meaning of excellence in the hospitality industry.

This work presents the global index of consensus as an indicator for filtering the most suitable clustering solutions using qualitative reasoning that promotes the consensus of experts' opinions, which is one of the key aspects in the concept mapping methodology. The index outperforms three of the most well-known clustering validation indexes in a case study focused on the meaning of excellence in hospitality.



## 1. Introduction

The concept mapping methodology aims to respond the challenge of guiding a group of experts in the objective representation of thoughts, ideas or abstract concepts based on promoting their agreement regarding what they consider most relevant in consensus [1,35]. Thus, this method is used to offer clarity and develop a model or specify a conceptual framework and it has been successfully applied in education, social research and management science fields to create conceptual frameworks based on specific aspects [26]. The methodology defines a set of general steps using qualitative and quantitative data to determine a conceptual map of interrelated concepts [27]. Giving a specific topic study through a set question, a focus group composed of experts in this domain generate ideas related to this topic using brainstorming. Next, the focus group have to group and weight the ideas in categories based on their point of view. This information is converted into knowledge using data mining techniques [37], which are applied to identify shared patterns between the opinion of the experts using multidimensional scaling and clustering techniques. It is important to highlight that clustering techniques often return more than one possible solution where each one represents a clustering configuration that groups elements in a specific way. Therefore, the last step is to validate and select the most suitable clustering configuration based on the criteria of the group of experts. Although one of the main benefits of this approach is its flexibility and adaptability, the amount of data that has to be analysed may hinder the tasks of experts because the selection of the best clustering configuration is non trivial and they have to review all the results following the subjective premise "does it make sense to you?" [35], which may compromise the objectivity of the approach.

This paper presents the global index of consensus (GIc) to help experts in selecting the most suitable clustering configuration based

☆ This paper has been recommended for acceptance by Lledó Museros.

* Corresponding author. Tel.: +34 932522890.
*E-mail address:* albert.fornells@tsi.url.edu (A. Fornells).

http://dx.doi.org/10.1016/j.patrec.2015.05.013
0167-8655/

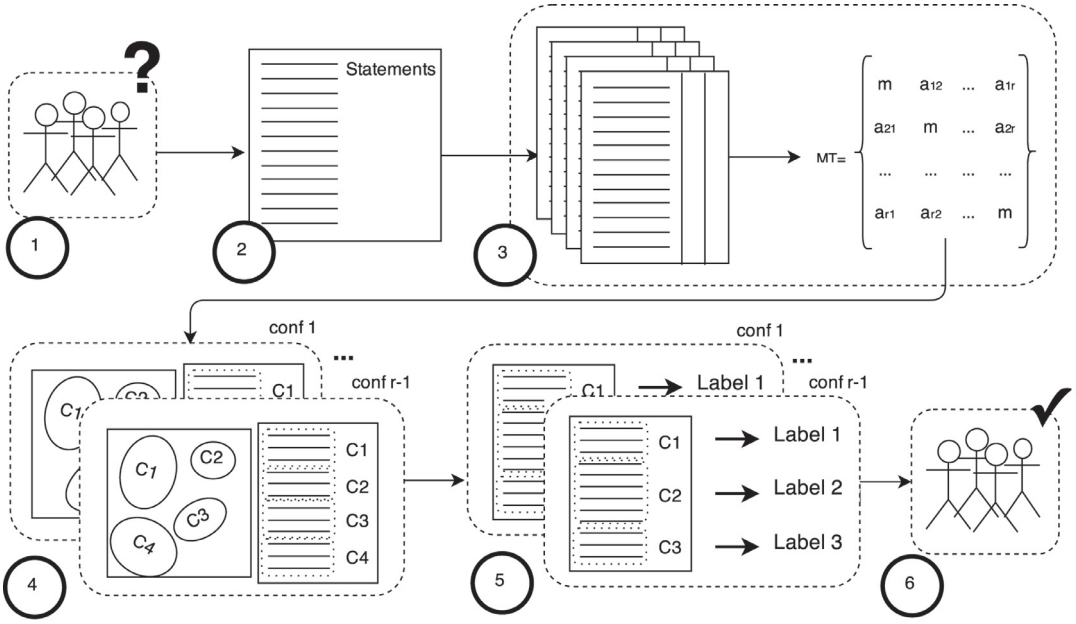



**Fig. 1.** The concept mapping process is split into 6 steps: (1) A set of experts is selected for finding out the meaning of a specific concept; (2) A list of $r$ ideas is generated through a Brainstorming process; (3) Ideas are evaluated by the experts and this information is used to build a matrix $MT_{r\times r}$; (4) A multi dimensional scaling and a clustering technique are applied over the matrix to project the information in a 2D space. The result is a set of $r$-1 possible cluster configurations; (5) Experts analyse all the configurations and they label each one of the clusters based on their items; (5) Experts agree on selecting the best concept representation based on the subset of concepts identified (6).

on two of the main premises of the concept mapping methodology: objectivity and consensus. Thus, the knowledge discovery process is drastically improved because experts have to focus only in useful configurations characterised for containing ideas in which the experts agree are similar and with the same relevance. This index is based on qualitative reasoning techniques and the concept of entropy [30]. Qualitative reasoning is a sub area of artificial intelligence that seeks to understand and explain human beings' non numerical evaluations and it also permits to handle with non numerical data preserving the principle of relevance, i.e., each variable can be valued with the level of precision required [33,34]. Finally, GIc is successfully evaluated and compared with respect to other approaches for tackling one of the challenges of the tourism sector: 'what are the main factors that lead to excellence in hospitality?'

This article is structured as follows. Section 2 describes the concept mapping methodology and how quantitative validation index can be used for selecting the most suitable patterns. Section 3 proposes the global index of consensus and describes its bases. Section 4 applies the concept mapping methodology for discovering the meaning of excellence in hospitality, and it also show how this index outperforms the results provided by some of the most well-known quantitative index. Finally, Section 5 ends with the conclusions and further work.

## 2. Framework

This section summarises the concept mapping methodology and some of the most well-known validation indexes used for the selection of the most suitable clustering solution.

### 2.1. Concept mapping methodology

Concept mapping was developed by [35] to respond to the conceptualisation needs based on the objectification of opinions and ideas from a group of experts. It uses a methodology which incorporates statistical techniques, such as multidimensional scaling and cluster analyses, and its applications are based on six main steps as indicated in Fig. 1: preparation, generation of statements, structuring statements, representation of statements, interpretation of maps and utilisation of maps.

1. **Preparation**. The aim of the preparation step is twofold. On one hand, clarify the construct for research. On the other, choose the focus group members who will participate in the process. For the best results, the group should contain up of 8–15 participants who are as diverse as possible in order to have different points of view reflected [19].
2. **Generation of statements**. Upon finishing the first step, the participants are invited to offer their ideas regarding on the main topic at hand using a brainstorming session. The development of this phase usually counts on the collaboration of an expert team specialised in group dynamics in order to obtain the best results possible [1,4].
3. **Structuring of statements**. The purpose of this step consists in determining how the different statements raised in the previous process are related. For this, the participants are asked to, firstly, evaluate each of the statements mentioned during the brainstorming session and, secondly, each member had to group the different statements according to their own criteria. For the first part, the focus group members are asked to rate the list of statements using a Likert scale from 1 to 5 [1] according to the degree of adjustment of the statement towards the concept set out. For the second part, participants have to group the list items according to their own criteria and a label that represents the main concept of each group they considered. Once the participants have rated the items and grouped them, a similarity matrix $S_{rxr}$ for each of the participants is created where $r$ represents the total number of statements generated during the brainstorming session. In each intersection, a '1' was introduced if a person in the group had put both items in the same group; otherwise, a '0' was introduced. Next, each individual's matrix is added to create a general grouping matrix as shown in Fig. 1. In the central diagonal there is the total number of participants, and the number of each intersection shows the number of people who put both corresponding statements in the same group independently of their meaning or the criteria used [4].

Please cite this article as: A. Fornells et al., Promoting consensus in the concept mapping methodology: An application in the hospitality sector, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.013

4. **Representation of statements**. Two statistical techniques are applied to objectify the results obtained during the previous phase. The data gathered in the grouping matrix in the previous step is processed using a multidimensional scaling (MDS) technique as agreed by [2] to project the original *r* dimensional space into a two dimensional space where each dot represents an statement. The closer together two dots are, the greater the number of people is who feel that these statements were in the same group. Once the map is constructed, a clustering algorithm [22] is applied using the same distance coordinates as those obtained by the MDS [2]. Clustering algorithms are able to group data from different points of view, and their suitability mainly depends on the application domain. Thus, the elements in a cluster are similar among them and different from the elements of other clusters and it provides the experts a possible classification or categorisation of the elements. One of the main challenges that experts have to tackle when they apply clustering techniques is the selection of the most suitable cluster solution as these techniques usually offer more than one possible solution, and there is no exact method to determine the definitive number of clusters as stated by [14]. Thus, the expert needs to evaluate the different configurations generated by the approach in order to select the most suitable representation of the statements.
   For this reason, the usage of a cluster validation index has become crucial to facilitate the data analysis in order to score or sort the possible solutions based on indicators (also called validation index) that promotes the separability and/or the compactness of the clusters [11,12]. These indicators allow experts to reduce the range of potentially valid solutions because they are based only on quantitative values without taking into account information related to the problem, which limits their capabilities. In this sense, the definition of a specific index based on the domain characteristics is the best way to help experts to select the most suitable solution [5]. This is exactly what we do in this article by means of the GIc, which allows experts to rank the clustering solutions based on what experts agree.
5. **Interpretation of maps**. After the statements had been mapped via the cluster analysis, a name is given to represent the key statements in each group as shown in Fig. 1. To carry out this step, interviews are held with experts in the field to evaluate the content of each group. The participating experts analyse the obtained results so that the global result corresponds to reality [35].
6. **Utilisation of maps**. Lastly, the maps are used as a graphic representation of the experts' opinion regarding the concept under study [1].

One of the biggest benefits of this methodology is its flexibility and capability for being applied to any kind of domain because it defines a framework that can be easily adapted to fit the problem that has to be faced. In contrast, this flexibility becomes a challenge because there is no one single way of doing things and this may hamper analysis. A clear example occurs in step 4 where the application of the clustering technique offers a set of potentially valid configurations. In that case, the selection of the configuration will set the difference between the success or failure and that is where the expert can be aided to select the most suitable configuration using the key element: consensus.

### 2.2. Looking for the most suitable patterns

The use of unsupervised learning approaches such as clustering algorithms is an essential step in almost any data analysis problem. However, two independent steps are needed before hand: analysts have to select (1) the clustering approach and (2) select the most appropriate solution from the whole set of possible solutions.

There is not a single criteria to classify the clustering algorithms, so they can be classified according to many criteria [7,9,37]: (1) the search strategy to find the clusters (centre-based, graph-based, model-based, search-based, density-based and subspace clustering), (2) the relationships between the clusters (partitional and hierarchical), (3) the instances distribution into the clusters (hard clustering or fuzzy clustering), and (4) the optimisation of the clusters (conventional clustering, ensemble clustering, or multiobjective clustering). For this reason, it is important to select the algorithm according to the data typology and the features of the application domain [22].

On the other hand, clustering techniques usually offer more than one possible cluster solution because they are not able to automatically identify the optimal number of groups to discover or the results are conditioned by some initial parameters that need to be tuned among other reasons [10]. These facts often hinders the data analysis step because experts must evaluate all the different solutions generated by the algorithm, which is highly time consuming and quite arbitrary because the selection will depend on the subjectivity of the expert due to the fact that all of them are potentially valid. For this reason, the application of evaluation functions for automatically scoring the clustering solutions has become the key for helping experts to select the best [13]. These evaluation functions define metrics that measure the cluster quality by using the same features included in the data set. Therefore, the challenge is define what quality means as the following subsections describe.

#### 2.2.1. Deviation and connectivity

The simplest evaluation metric for measuring the quality of a cluster solution is to evaluate how close the elements of each cluster are and how separated the clusters between them are. Consequently, the smaller their values the better is the solution. This is exactly what Deviation and Connectivity measures do respectively [10,13,17,21].

The *deviation* (Dev) measures the compactness of the clusters. It is computed as the overall summed distances between data items and their corresponding cluster centre as Eq. 1 shows, where $C$ is the clustering obtained, $C_i$ is the set of instances belonging to cluster $i$, $\nu_i$ is the centroid of cluster $i$, and $d(x, \nu_i)$ is the Euclidean distance between the element $x$ and $\nu_i$.

The *connectivity* (Conn) refers to the cluster connectedness. It takes into account the degree to which data points that are close in the feature space have been placed in the same cluster as Eq. 2 shows, where $r$ is the number of examples in the training data set, $C$ is the clustering obtained, $nn(x, i)$ returns the $i$th nearest element of $x$ using the Euclidean distance and $\ell$ is the amount of nearest elements taken into account. Note that, for each instance $i$, the metric computes a weighted sum of the $\ell$ nearest neighbours that belong to a different cluster from that of $i$ (the weight is decreased according to how far instances $i$ and $j$ are).

Although the information provided by the Deviation and the Connectivity allows to obtain insights from the cluster solution analysed, this information is not enough to select the best configuration. Thus, it is necessary to define more powerful indicators to evaluate the clustering solutions.

$$Dev(C) = \sum_{i \in C} \sum_{x \in C_i} d(x, \nu_i) \tag{1}$$

$$Conn(C) = \sum_{x=1}^{r} \left( \sum_{i=1}^{\ell} \chi(x, nn(x, i), i) \right), \quad \text{where} \tag{2}$$

$$\chi(x, y, i) = \begin{cases} \frac{1}{i} & \text{if } \neg\exists j : x \in C_j \wedge y \in C_j, \\ 0 & \text{otherwise.} \end{cases}$$

Please cite this article as: A. Fornells et al., Promoting consensus in the concept mapping methodology: An application in the hospitality sector, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.013

$$DB(C) = \frac{1}{k}\sum_{i=1}^{k}\max_{\substack{j=1\\ j\neq i}}^{k}\left\{\frac{S_k(C_i)+S_k(C_j)}{d(v_i,v_j)}\right\} \quad (3)$$

$$S_k(C_i) = \frac{1}{|C_i|}\sum_{x\in C_i} d(x,v_i)$$

$$Dn(C) = \min_{i=1}^{n}\left\{\min_{j=i+1}^{n}\left\{\frac{d(C_i,C_j)}{\max_{k=1}^{n}\{diam(C_k)\}}\right\}\right\} \quad (4)$$

$$diam(C_k) = \max_{x,y\in C_k}\{d(x,y)\}$$

$$d(C_i,C_j) = \min_{x\in C_i, y\in C_j}\{d(x,y)\}$$

$$Sil(C) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{|C_i|}\sum_{x\in C_i}\left(\frac{b(C,i,x)-a(C_i,x)}{max\{a(C_i,x),b(C,i,x)\}}\right)\right)$$

$$a(C_i,x) = \frac{1}{|C_i|-1}\sum_{y\in C_i} d(x,y)$$

$$b(C,i,x) = \min_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{|C_j|}\sum_{y\in C_j} d(x,y) \quad (5)$$

*2.2.2. Cluster validation index*

Cluster validation index [10,15] is an objective function that evaluates the clustering results. Validation techniques can be based on comparing clusters to the original classes of the problem if classes are known (supervised approach), or by validating clusters according to their quality based on the compacting and separation between them when classes are unknown (unsupervised approach).

Regarding the characteristics of the clustering technique used in the concept mapping (partitional and hard clustering), there are three possible clustering validation approaches [16,25]. The reader is referred to [32] for hierarchical and fuzzy clustering algorithms. The first one is called external criteria and the idea is to evaluate a clustering result comparing it with a structure of the data set obtained without applying any clustering algorithm. The second approach is called internal criteria and the objective is to evaluate a clustering result comparing it with only quantities and features inherent to the data set. The third approach is called relative criteria and it is based on comparing a clustering result with other results obtained from the application of the same clustering algorithm with different parameter values, or of other clustering algorithms. The cluster validation methods based on external or internal criteria are based on statistical hypothesis testing, and their major drawback is their high computational cost. Moreover, these two approaches measure the degree to which a data set confirms an a-priori specified scheme that can be inherent to the data set or an intuitive structure of the data. On the other hand, relative criteria methods find the hypothetical best clustering scheme from several clustering results obtained with different parameters or clustering algorithms without using statistical tests, so they are less computationally expensive. Therefore, this last approach is what we really need for our purposes because we want to compare the performance of several configuration in an unsupervised process.

The Davies–Bouldin index [6] (see Eq. (3)), the Dunn index [8] (see Eq. (4)) and the Silhouette index [29] (see Eq. (5)) are three of the most well known validation strategies. The range of the first two indexes is $[0,+\infty]$ and $[-1,1]$ for the last one. Silhouette and Dunn indexes have to be maximised, and Davies–Bouldin index has to be minimised. In the three equations $C$ is the clustering obtained; $k$ is the number of clusters; $C_i$ is the set of instances belonging to cluster $i$; $|C_i|$ is the number of elements in $C_i$; $v_i$ is the centroid of $C_i$; $d(x,y)$ is the Euclidean distance between $x$ and $y$ elements; $nn(x,i)$ returns the $i$th nearest element of $x$ according to $d(x,y)$; and $\ell$ is the amount of nearest elements taken into account. The main difference of these three indexes is the calculation of the quality of the shape of each cluster. Davies–Bouldin index evaluates the clusters taking into account if they are scattered, calculating the distance between the instances of each cluster and their respective centroid. Dunn index evaluates the clusters calculating if they are compact, penalising the clusters with a long diameter. Silhouette index calculates the tightness of the clusters, taking into account the distance between the instances of each cluster.

## 3. Methodological approach

The approach proposed in this paper helps the experts to evaluate in the step 4 the different configurations generated in the previous steps, and to select de optimal clustering solution, as described in Section 2.1. To this end, the proposed methodology aims to identify and emphasise the clusters and the ideas in which the experts agree that are the most important ones. This is done in two phases. First a degree of consensus of clusters is defined to measure the agreement among the members of the focus group with respect to each cluster. Second a recurrence index of each idea, highlighting the ideas that are more in consensus, is used as a weight coefficient to define a global index of consensus of configurations in order to promote the ideas in which experts agree. This methodology relies on the use of qualitative labels belonging to a qualitative absolute order-of-magnitude model. This allows dealing with the focus group members evaluations to improve Concept Mapping processes through the new global index.

### 3.1. Qualitative reasoning

The one-dimensional absolute order-of-magnitude model [33,34] works with a finite number of qualitative labels corresponding to an ordinal scale of measurement. The number of labels chosen to describe a real problem is not fixed, but depends on the characteristics of each represented variable.

Let us consider an ordered finite set of *basic* labels $\mathbb{S}_* = \{B_1,\ldots,B_n\}$, being $n$ the number of labels chosen each one of them corresponding to a linguistic term, in such a way that $B_1 < \cdots < B_n$.

**Example 1.** To illustrate the expression of a set of ordered linguistic evaluations, an example of the *basic* labels for $n = 5$ is given by: $B_1$ = "not important at all" $< B_2$ = "of little importance" $< B_3$ = "important" $< B_4$ = "very important" $< B_5$ = "extremely important".

The complete universe of description for the order-of-magnitude space is the set $\mathbb{S} = \mathbb{S}_* \cup \{[B_i,B_j]\,|B_i,B_j \in \mathbb{S}_*, i<j\}$, where the label $[B_i,B_j]$ with $i<j$ is defined as the set $\{B_i,B_{i+1},\ldots,B_j\}$, with the convention $[B_i,B_i] = \{B_i\} = B_i$.

Consistent with Example 1, the linguistic evaluation "very or extremely important" can be represented by the non-basic qualitative label ["very important", "extremely important"], i.e., $[B_4,B_5]$. The label "unknown" is represented by ["not important at all", "extremely important"], i.e., $[B_1,B_5]$.

The order in the set of basic labels $\mathbb{S}_*$ induces a partial order $\leq$ in $\mathbb{S}$ defined as: $[B_i,B_j] \leq [B_r,B_s] \Leftrightarrow B_i \leq B_r$ and $B_j \leq B_s$. This relation is trivially an order relation in $\mathbb{S}$, but a partial order, since there are pairs of non-comparable labels. For instance, in Example 1, the relation $[B_i,B_j] \leq [B_r,B_s]$ expresses that $[B_i,B_j]$ is "less or equal important than" $[B_r,B_s]$.

There is another partial order relation $\leq_P$ in $\mathbb{S}$ "to be more precise than", given by $[B_i,B_j] \leq_P [B_r,B_s]$ iff $[B_i,B_j] \subset [B_r,B_s]$, i.e. $r \leq i$ and $j \leq s$. The less precise label is $? = [B_1,B_n]$. This structure permits working with all different levels of precision from the basic labels to the ? label (see Fig. 2).

Two different binary operations are defined in the complete universe of description $\mathbb{S}$, called the connex union and the intersection, introduced in a more general context as the mix and the common operations in [28].

Please cite this article as: A. Fornells et al., Promoting consensus in the concept mapping methodology: An application in the hospitality sector, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.013

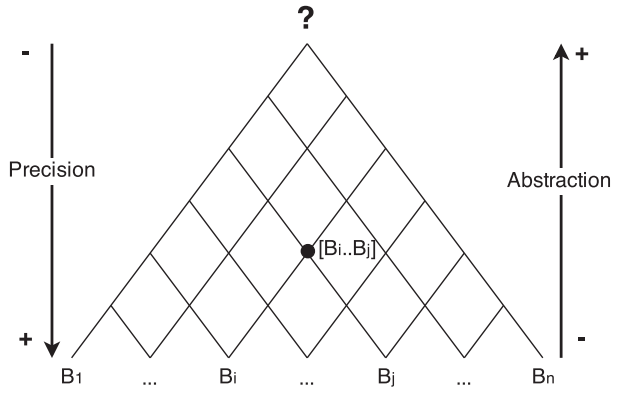



**Fig. 2.** The complete universe of description $\mathbb{S}$ [28].

**Definition 1.** Given two qualitative labels $[B_{i_1}, B_{j_1}]$, $[B_{i_2}, B_{j_2}] \in \mathbb{S}$, their *connex union* is the qualitative label $[B_{i_1}, B_{j_1}] \sqcup [B_{i_2}, B_{j_2}] = [B_{\min(i_1,i_2)}, B_{\max(j_1,j_2)}]$.

**Definition 2.** Given two qualitative labels $[B_{i_1}, B_{j_1}]$, $[B_{i_2}, B_{j_2}] \in \mathbb{S}$, such that $[B_{i_1}, B_{j_1}] \cap [B_{i_2}, B_{j_2}] \neq \emptyset$, their *intersection* is the qualitative label $[B_{i_1}, B_{j_1}] \cap [B_{i_2}, B_{j_2}] = [B_{\max(i_1,i_2)}, B_{\min(j_1,j_2)}]$.

Following Example 1, the connex union of $B_1$ and $[B_3, B_4]$ is $B_1 \sqcup [B_3, B_4] = [B_1, B_4]$, and their intersection is empty. In the case of the pair of labels $[B_1, B_3]$ and $[B_3, B_4]$ their connex union is $[B_1, B_3] \sqcup [B_3, B_4] = [B_1, B_4]$ and their intersection is $[B_1, B_3] \cap [B_3, B_4] = B_3$.

### 3.2. Entropy and consensus

A definition of a consensus index, able to synthesise the focus group members' evaluations, is used in the proposed methodology. The consensus index involves the notion of entropy of a qualitative label defined in $\mathbb{S}$, inspired by the Shannon entropy concept in information theory [30]. This notion requires a normalised measure on the set $\mathbb{S}$, i.e., a measure such that $\mu(B_i) > 0$ for all $i = 1, \ldots, n$ and $\sum_{B_i \in \mathbb{S}^*} \mu(B_i) = 1$. Then: $\mu([B_i, B_j]) = \sum_{k=i}^{j} \mu(B_k)$, $\mu([B_i, B_j]) \leq 1$ for all $i, j \in \{1, \ldots, n\}$, and $\mu([B_1, B_n]) = 1$.

In the context of qualitative absolute order-of-magnitude models, the concepts of information of a label and entropy of a qualitative description were introduced in [28]. Following [28], in this paper we define the entropy of a qualitative label by a positive continuous real function of the measure of this label as follows:

**Definition 3.** The *entropy of a qualitative label* $Q \in \mathbb{S}$ is defined as:

$$H(Q) = \ln \frac{1}{\mu(Q)},$$

where $\mu$ is the measure considered in $\mathbb{S}$.

Note that for all $Q \in \mathbb{S}$, if $Q \neq ?$ then $\mu(Q) \in (0, 1)$ and, consequently, $H(Q) > 0$. Moreover, $H$ decreases with respect to $\leq_P$: $Q \leq_P Q' \Rightarrow Q \subset Q' \Rightarrow \mu(Q) \leq \mu(Q') \Rightarrow \ln \frac{1}{\mu(Q)} \geq \ln \frac{1}{\mu(Q')}$. In addition, $H(?) = \ln 1 = 0$.

The definition of the degree of consensus of a set of qualitative labels is as follows:

**Definition 4.** Given $m$ qualitative labels $Q_1, \cdots, Q_m \in \mathbb{S}$, such that $\cap_{j=1}^{m} Q_j \neq \emptyset$, their *degree of consensus* is:

$$Dc(Q_1, \cdots, Q_m) = \frac{H(\sqcup_{j=1}^{m} Q_j)}{H(\cap_{j=1}^{m} Q_j))} = \frac{\ln(\mu(\sqcup_{j=1}^{m} Q_j))}{\ln(\mu(\cap_{j=1}^{m} Q_j))}$$

In the case that $\cap_{j=1}^{m} Q_j = \emptyset$, their *degree of consensus* is $Dc(Q_1, \cdots, Q_m) = 0$.

**Example 2.** Let us consider the concept $C =$ "Customer Oriented" and a focus group of 3 members $\mathbb{E} = \{e_1, e_2, e_3\}$. Let us assume that the opinions of the three members with respect to $C$ are represented by three qualitative labels defined as: $Q_1(C) = [B_4, B_5]$, $Q_2(C) = B_4$, $Q_3(C) = [B_3, B_5]$ using the linguistic evaluations corresponding to basic labels $B_1, \ldots, B_5$ given in Example 1. Finally, let us define $\mu(B_i) = 1/5, i = 1, \ldots, 5$.

Then, since $\sqcup_{k=1}^{3}(Q_i(C)) = [B_4, B_5] \sqcup B_4 \sqcup [B_3, B_5] = [B_3, B_5]$ and $\cap_{k=1}^{3}(Q_i(C)) = [B_4, B_5] \cap B_4 \cap [B_3, B_5] = B_4$, the degree of consensus is:

$$Dc(Q_1, Q_2, Q_3) = \frac{H(\sqcup_{k=1}^{3}(Q_i(C)))}{H(\cap_{k=1}^{3}(Q_i(C))))} = \frac{H([B_3, B_5])}{H(B_4)} = \frac{\ln 3/5}{\ln 1/5} = 0.32$$

### 3.3. The proposed global index of consensus

The structure of qualitative absolute order-of-magnitude models allows us to deal with the focus group members' evaluations of ideas and concepts in a concept mapping process. To this end, we work in a one-dimensional absolute order-of-magnitude model with $n$ basic labels corresponding to the $n$ ordered responses of the Likert scale used by the members of the focus group.

In the following, let us consider a focus group consisting of $m$ members, that after a brainstorming process have generated and evaluated a set of $r$ different ideas. For each member of the focus group $j$, with $1 \leq j \leq m$, and for each idea $X$, the *opinion of member j with respect to X* is an element of $\mathbb{S}_*$, which is denoted by $V_j(X)$.

Note that the concept mapping method provides $r - 1$ different cluster solutions. Each cluster solution provides exactly $k$ clusters or groups, being $1 \leq k < r$. For instance, on the one hand, the cluster solution with 2 clusters groups $r - 1$ ideas in one cluster and the remaining most discordant idea alone in the other cluster. On the other hand, the cluster solution with $r - 1$ clusters, groups the most similar two ideas in a cluster meanwhile the remaining $r - 2$ ideas are each one in a different cluster.

From now on, for each $1 \leq k < r$, clusters belonging to the configuration with $k$ groups will be denoted by $C_i^k$, with $1 \leq i \leq k$.

**Definition 5.** Fixed a configuration with $k$ clusters, let $j$, with $1 \leq j \leq m$, be a member of the focus group, let $C_i^k$, $1 \leq i \leq k$, be a cluster, and let $\{X_1^i, \ldots, X_{s_i}^i\}$ be the set of ideas in cluster $C_i^k$. The *opinion of member j with respect to* $C_i^k$ is defined as:

$$Q_j(C_i^k) = V_j(X_1^i) \sqcup \cdots \sqcup V_j(X_{s_i}^i).$$

Note that $Q_j(C_i^k)$ belongs to the complete universe of description for the order-of-magnitude space. In the case that $V_j(X_1^i) = \cdots = V_j(X_{s_i}^i)$, $Q_j(C_i^k)$ is a basic label, otherwise $Q_j(C_i^k)$ is a non-basic label.

Intuitively speaking, $Q_j(C_i^k)$ is the result of mixing the evaluations of member $j$ of all ideas in cluster $C_i^k$ in a new one that includes all of them.

The entropy $H$ and the degree of consensus $Dc$ introduced in Section 3.2 allows us to define a measure of consensus among the members of the focus group with respect to each cluster:

**Definition 6.** Fixed a configuration with $k$ clusters, let $C_i^k$, $1 \leq i \leq k$, be a cluster, and let $Q_j(C_i^k)$ be the opinion of member $j$ with respect to $C_i^k$, $1 \leq j \leq m$. The *degree of consensus of cluster* $C_i^k$ is

$$Dc(C_i^k) = \begin{cases} \dfrac{H(\bigsqcup_{j=1}^{m} Q_j(C_i^k))}{H(\bigcap_{j=1}^{m} Q_j(C_i^k))}, & \text{if } \bigcap_{j=1}^{m} Q_j(C_i^k) \neq \emptyset \\ 0, & \text{otherwise.} \end{cases}$$

In order to reflect if a configuration has any cluster with nonzero degree of consensus, we define:

Please cite this article as: A. Fornells et al., Promoting consensus in the concept mapping methodology: An application in the hospitality sector, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.013

**Definition 7.** Fixed a configuration with $k$ clusters,

$$N(k) = \begin{cases} 1, & \text{if } \exists C_i^k \text{such that } Dc(C_i^k) \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

The above definition allows us to introduce a recurrence index of an idea. The index is defined via these numbers $N(k)$, $k$=1 to $r$-1 and the characteristic functions of clusters. The recurrence index of an idea takes into account the expert's perception of the relevance of the idea, which is extracted from the evaluations of the focus group members.

**Definition 8.** Let $X$ be an idea, The *recurrence index of X* is:

$$RI(X) = \frac{\sum_{k=1}^{r-1} \sum_{i=1}^{k} 1_{C_i^k}(X)}{\sum_{k=1}^{r-1} N(k)}$$

where $1_{C_i^k}$ is the characteristic function that indicates membership of an element in $C_i^k$.

In this way, RI($X$) provides a normalisation of the number of times the idea X is in consensus among all the configurations.

To be able to select the most suitable cluster solution as an improvement of a concept mapping process, we propose a global index of consensus of each configuration, to measure the cluster solutions quality. The global index of consensus of a configuration is constructed as a weighted sum of the degrees of consensus of clusters in the configuration. Each weight captures the importance of the corresponding cluster by considering the addition of the recurrence index of ideas belonging to it.

**Definition 9.** The *global index of consensus of a configuration with k clusters* is

$$GIc(k) = \sum_{i=1}^{k} w(C_i^k) \cdot Dc(C_i^k),$$

where $w(C_i^k) = \sum_{X \in C_i^k} RI(X)$, i.e. the addition of the recurrence index of ideas in cluster $C_i^k$.

## 4. Case study: what are the main factor that describe excellence in hospitality?

This section tackles the challenge of analysing the meaning of the excellence in hospitality by means of the application of the concept mapping methodology. First, excellence in hospitality is introduced. Next, the application of the first four steps for the statement generation is described. Thus, the selection of the most suitable clustering configuration is analysed by means of the application of quantitative and qualitative index. Finally, experts label and assess the discovered concept through the last two steps.

### 4.1. Excellence in hospitality

There is an increasing concern over the term and meaning of hospitality as well as a search among academics to identify a globally understood an accepted conceptual term that defines this concept. Slattery et al. [31] refers to [24] who posit that the understanding of hospitality has been impaired by an industrial myopia and propose to improve the understanding by; reflecting insights into the study of hospitality that encompass the commercial provision of hospitality and the hospitality industry, yet at the same time it is recognised that hospitality needs to be explored in a private domestic setting and studies the concept of hospitality as a social phenomenon involving relationships between people. Along similar lines of thought, Wood and Brotherton [38] affirm the statement that the conceptual development for the hospitality concept is limited, and the academic literature that does exist is scattered. This current issue and underlying challenge to find the meaning of hospitality can be rooted by the essence of the discipline. The hospitality sector has always been a professional one and from a commercial and management point of view it has evolved into more areas where an interaction between a host and a guest takes place. Even though a holistic approach to the term hospitality is increasingly accepted, it is still important to define the key aspects that define such discipline, not only for academic purposes but also to help develop better suited professionals [36]. Hospitality has always been about relationships. The word itself means friendliness to strangers. [20] suggests that by exploring and defining hospitality as an experience, new perspectives emerge that have important implications for hospitality in commercial contexts. These implications take hospitality beyond services management to a place where hospitableness, a sense of theatre and generosity are central. Hemmington advocates that hospitality businesses must focus on the guest experience and stage memorable experiences that stimulate all five senses. By achieving this, hospitality organisations that are able to capture this sense of theatre and generosity will gain competitive advantage by providing their guests with experiences that are personal, memorable and add value to their lives. Linking this concept to the values in the hospitality concept, [18] posit that hospitality, because of its connection to values and ways of thinking is not an art, a science, or a business, but a philosophy. It further proposes that those motivated by a desire to serve are strongly and deeply attracted to work in commercial hospitality where they can express themselves and find meaning through their work. In those places where there is an interaction between hosts and guests, loyalty is the critical factor for a sustainable business. To achieve loyalty in hospitality, an attitude and philosophy towards excellence is necessary. Business excellence is important in creating sustainable and continuous quality improvement of business processes, that may bring strong financial performance, high customer demand, goal achievement, successful employee recruitment and admission, desired product and service outcome, and outstanding staff [23].

### 4.2. Generation of the possible maps

The steps 1, 2, 3 and 4 of the concept mapping methodology were done in the following way for creating all the set of potentially groups of concepts.

1. **Preparation**. A focus group composed by 11 experts representing the hospitality industry in Barcelona was held in January 2014. All participants were senior managers with more than 10 years of relevant experience in international companies. It is important to highlight that Barcelona is the 10th-most-visited city in the world and the third most visited in Europe after London and Paris, with 8 million tourists every year since 2012 [3]. Barcelona is a internationally renowned tourist destination with numerous recreational areas, historical monuments, including eight UNESCO world heritage sites, many good-quality hotels, and developed tourist infrastructure. The participants were asked to answer the question: From your point of view, what are the main factors that describe excellence in hospitality?
2. **Generation of statements**. A list of 100 ideas were generated following the brainstorming session leaded by an expert in group dynamics (see Table 1 in the supplementary material).
3. **Structuring of statements**. The aggregation matrix was built by means of the addition of the matrix of each focus group member. Each individual's matrix represents the evaluation of ideas from 1 to 5 and how the ideas are related among them.
4. **Representation of statements**. The multidimensional scaling (in accordance with the Alscal Method) was applied to project

Please cite this article as: A. Fornells et al., Promoting consensus in the concept mapping methodology: An application in the hospitality sector, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.013

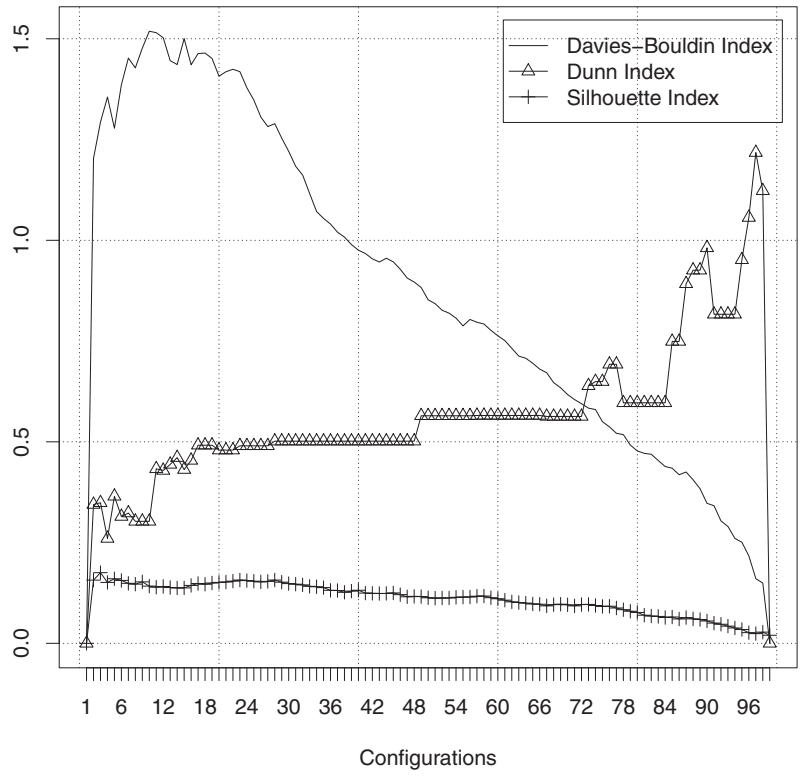



**Fig. 3.** The Davis–Bouldin, Dunn and Silhouette indexes are applied over the 99 configurations generated in the step 4. Silhouette and Dunn indexes have to be maximised, and Davies–Bouldin index has to be minimised. They are optimising only the cluster geometry without taking into account the most key element: the consensus of the experts.

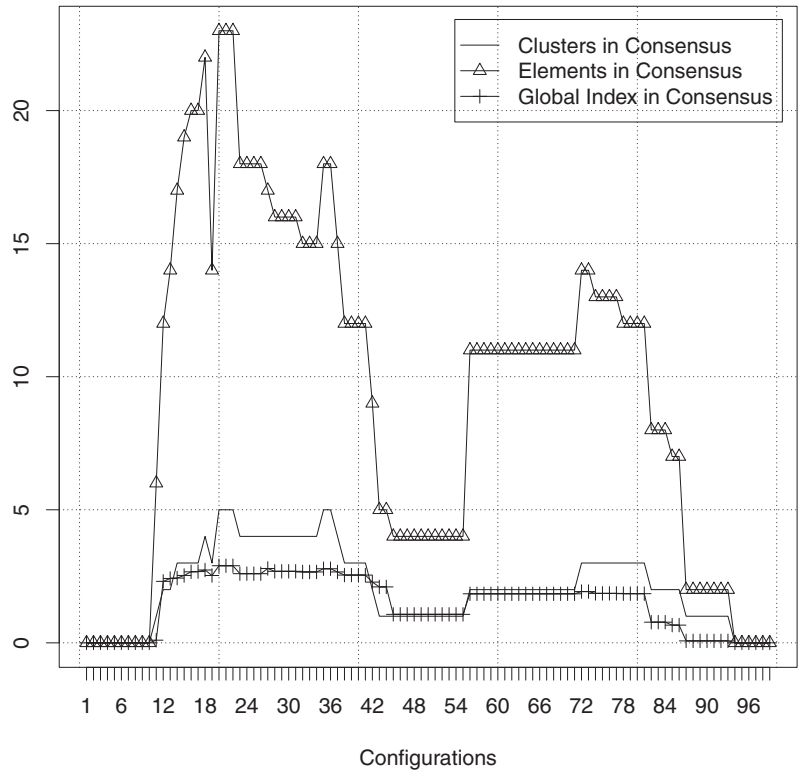



**Fig. 4.** The graphic shows the global index of consensus and how many clusters (and how many items) have a degree of consensus different from zero for each one of the 99 configurations. This approach highlights the most relevant and useful data, which are the ones with consensus. Thus, the global index of consensus is a qualitative index that promotes the selection of configurations based on the elements with the highest degree of consensus.

Please cite this article as: A. Fornells et al., Promoting consensus in the concept mapping methodology: An application in the hospitality sector, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.013

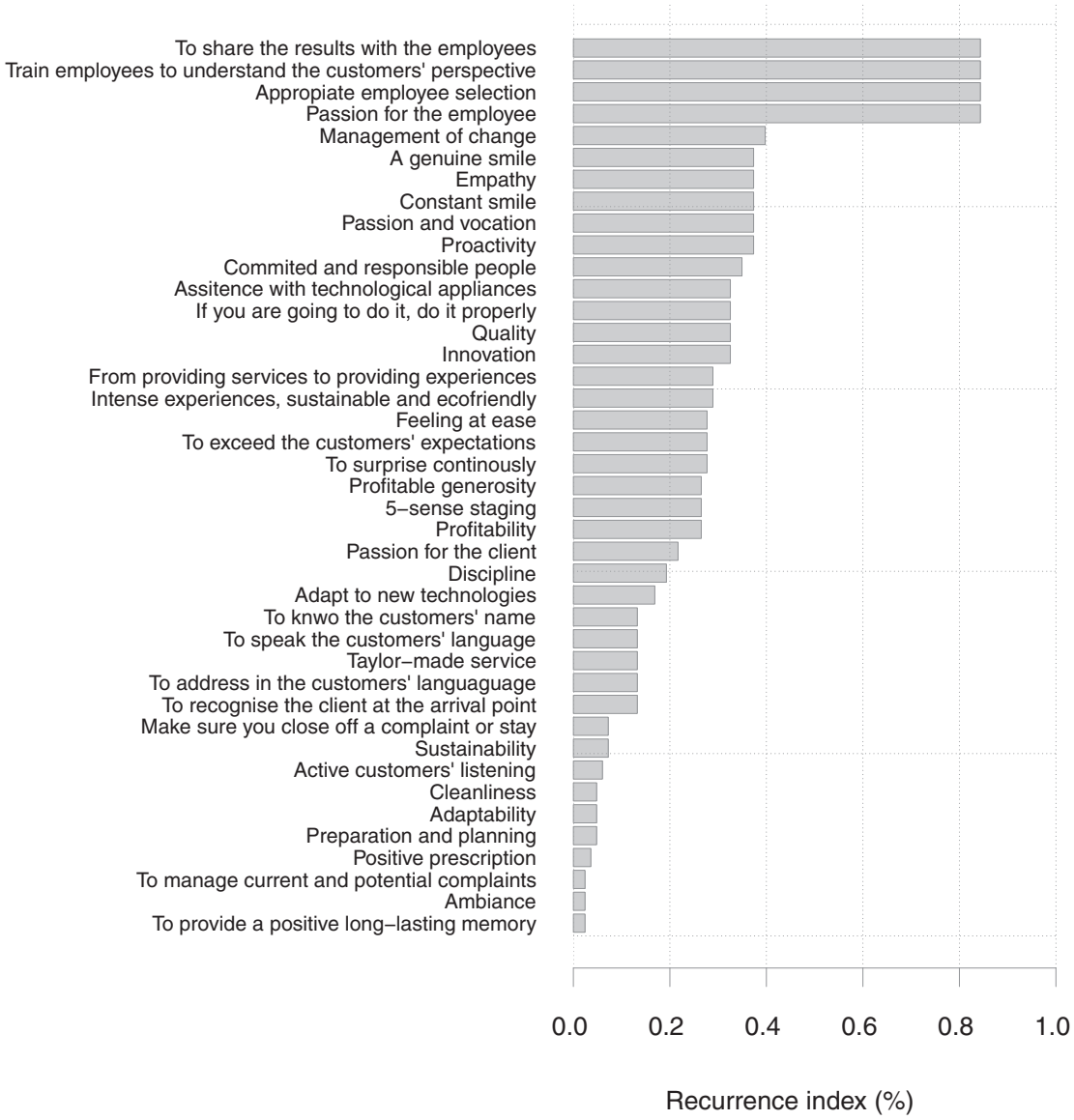



**Fig. 5.** Ranking of the ideas based on a recurrence index greater than zero for each one of the 99 configurations.

the information in two dimensions. Next, a clustering method (according to the Ward Method) was executed to generate the potentially statement, representations that were from only 1 cluster ($k = 1$) since 99 clusters ($k = 99$).

### 4.3. Selection of the most suitability map representation

Steps 5 and 6 are focused on labelling and analysing the correctness of the most suitable clustering configuration from all the $r$ potentially configurations discovered in the step 4 and, consequently, this decision will condition the results. We analyse this selection from three different perspectives using: (1) Quantitative indexes, (2) The proposed global index of consensus and (3) Combining the mentioned quantitative indexes and the consensus concept, using the degree of consensus ($Dc$) introduced in Definition 6.

Fig. 3 shows the results of applying the Davis-Bouldin, Dunn and Silhouette indexes described in Subsection 2.2.2 which are only based on the cluster geometric data without taking into account additional knowledge from the domain. The figure shows that Davis–Bouldin and Dunn indexes obtain the best results when the $k$ is higher. In contrast, Silhoutte index promote the solutions with fewer clusters. Therefore, all index are promoting extreme solutions that will hinder the task of experts because the concept has to be defined using few clusters with almost all ideas or many clusters with one or two ideas. For this reason, quantitative indexes are not enough for guiding the selection process. If the expert is looking for concepts that should represent the consensus between the different focus group members, why do not introduce this domain characteristic into the problem?

The application of the global index of consensus requires three steps: (1) Preprocess the 99 clustering configurations in order to remove the clusters and elements without consensus (Definition 4); (2) Calculate the recurrence index (Definition 8) as Fig. 5 shows; (3) Apply the global index of consensus (Definition 9) over the preprocessed clustering configurations as Subsection 3.3 describes. Fig. 4 shows the global index of consensus and the impact of applying the consensus concept over the clusters and their elements into the 99 configurations. For each one of the configurations, the clusters (and their items) without consensus are removed because they represent information that introduces noise and uncertainty. As seen, the number of clusters and elements is drastically reduced because this preprocessing operation promotes the ideas in which experts agree. Thus, the configuration $k = 20$ is selected as the best because it has the highest global

Please cite this article as: A. Fornells et al., Promoting consensus in the concept mapping methodology: An application in the hospitality sector, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.013

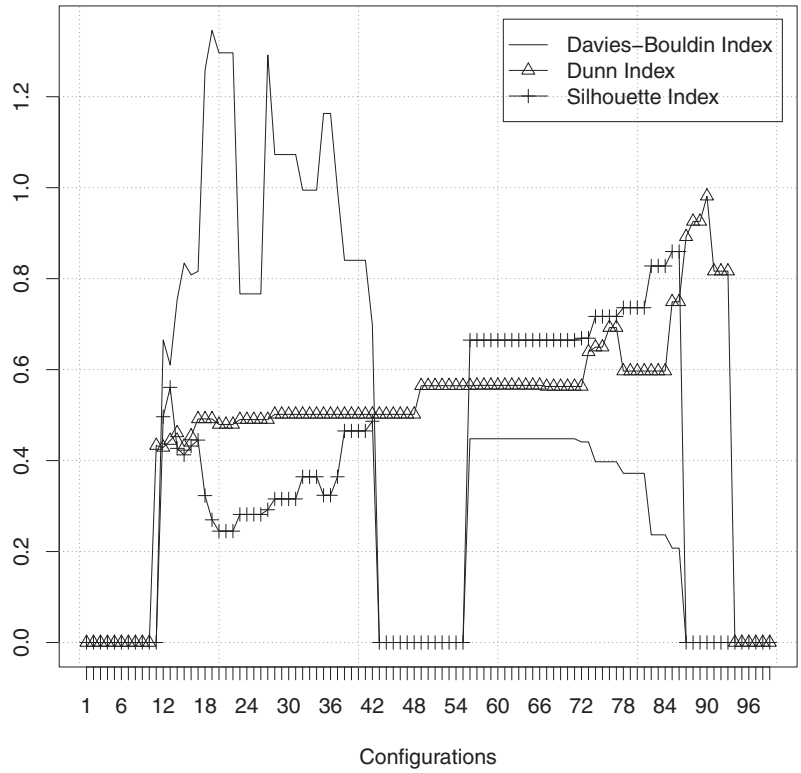



**Fig. 6.** The Davis-Bouldin, Dunn and Silhouette indexes are applied over the 99 preprocessed configurations generated in the step 4. The preprocess consists in removing the clusters (and its elements) with degree of consensus equal to zero. The value of the index is 0 for those configurations that are too small and it is not possible to compute the index values.

index of consensus and the highest number of clusters and elements in consensus.

Finally, Fig. 6 shows the application of the Davis–Bouldin, Dunn and Silhouette indexes described in Subsection 2.2.2 over the 99 preprocessed configurations. The most significant effect that can be observed from the application of that consensus concept is that the linearity of the scoring is broken because some empty configurations are rejected, which is reducing the original scope of configurations. Nevertheless, each index promotes different configurations as it happened before and $k = 85$ is selected as the best cluster configuration based on the three indexes because it is the intersection between them.

### 4.4. Analysis and assessment of the discovered concept

The best configurations selected for the global index of consensus and for the combination of the consensus concept with the quantitative indexes were presented to the experts in order to label the clusters and assess its meaning regarding to the initial question. After the analysis of both configurations, the experts agreed that the configuration selected using the quantitative indexes was too small for extracting any conclusion because it had only two clusters and seven ideas in consensus. In contrast, the configuration provided by the proposed global index of consensus allowed them to label the resulting clusters as:

- Innovation, quality, successful service performance and technology assistance ideas represent service quality. An excellent hospitality means to deliver an innovative and excellent service to the customer.
- To acknowledge the customer upon arrival, To address to the customer in his/her language, tailor-made service, speak the customer in his or her language and to identify the customer by his or her name represents customer oriented. Description: An excellent hospitality means to have a company, which understands and fulfils customer needs and expectations.
- To surprise continuously, exceeding the customers expectations, feel at home and positive prescription represents loyality. An excellent hospitality means to create a long-lasting relationship with the customer by exceeding his or her expectations and making the customer feel special.
- Passion for the employee, change management, discipline, an appropriate employee selection, to train the employees to understand the customers needs and to share the quality results with the employees represents human resource management. An excellent hospitality is delivered by the excellent people at the firm, so an excellent human resource management is the key to achieve it.
- 5 senses setting, an intense, sustainable and eco-friendly experiences, to adapt to new trends and to shift from service delivery to experience creation represents creating an experience. An excellent hospitality means providing memorable experiences rather than a plain service.

Finally, we also reviewed the worst scenarios to figure out why they scored with a low value. In this sense, experts analysed the configurations $k = 2$, 3 and 11 selected by the Silhoutte, Dunn and Davies–Bouldin respectively and their conclusion was that they were not useful. In fact, Fig. 4 shows that the consensus for these configurations is 0.

## 5. Conclusion and further work

One the biggest challenges when applying the concept mapping methodology is to decide the suitability or the relevance of a configuration. In this work a new method that substantially improves the clustering determination in this methodology has been proposed. A global index of consensus is defined based on the objectivity and the

Please cite this article as: A. Fornells et al., Promoting consensus in the concept mapping methodology: An application in the hospitality sector, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.05.013

consensus, which are two of the main premises of the concept mapping methodology. The proposed global index helps experts to select the most suitable clustering configuration improving the knowledge discovery process, since experts have to focus only on useful configurations. In that sense, the global index helps experts to identify the configurations containing concepts and ideas in which experts agree for being the most important ones. The global index presented in this paper is based on qualitative reasoning techniques and permits extract valuable and useful information from experts, which is crucial to select the most suitable configuration.

The proposed methodology has been applied to analyse the meaning of excellence in hospitality in a case study framed in the Barcelona hospitality industry in January 2014. The case study experimental results proved that this method achieves much better results compared with current state-of-the-art approaches, based only on quantitative data.

From a theoretical point of view, future work includes the adaptation of the proposed global index to other unsupervised clustering methodologies to determine the most suitable configuration. On the other hand, understanding what excellence in hospitality means from a cross-cultural perspective should be very interesting because the mobility in this century is a key element in tourism.

## Acknowledgements

This research was partially supported by the SENSORIAL Research Project (TIN2010-20966-C02) funded by the Spanish Ministry of Science and Information Technology and by the Excellence in Hospitality Research Project (URL/R6/2014) funded by Ramon Llull University.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patrec.2015.05.013.

## References

[1] J.E. Bigné, J.A. Manzano, I. Küster, N. Vila, The concept mapping approach in marketing: an application in the travel agencies sector, Qual. Market Res. Int. J. 5 (2) (2002) 87–95.
[2] I. Borg, P. Groenen, Modern Multidimensional Scaling, Springer, 1997.
[3] C. Bremmer, Euromonitor internationals top city destinations ranking, Euromonitor Int. (2011). http://blog.euromonitor.com/2014/01/euromonitor-internationals-top-city-destinations-ranking.html.
[4] A. Calvo, F. Criado, R. Periez, Desarrollo de un instrumento para evaluar la idoneidad de los planes docentes: una aplicacin a la diplomatura en turismo. Presented in Decisiones basadas en el conocimiento y en el papel social de la empresa, Academia Europea de Direccin y Economa de la Empresa, Palma de Mallorca, 2006.
[5] G. Corral, A. Garcia-Piquer, A. Orriols-Puig, A. Fornells, E. Golobardes, Analysis of vulnerability assessment results based on {CAOS}, Appl. Soft Comput. 11 (7) (2011) 4321–4331. Soft Computing for Information System Security.
[6] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 224–227.
[7] R. Duda, P. Hart, D. Stork, Pattern Classification, John Wiley and Sons, Inc, 2000.
[8] J. Dunn, Well separated clusters and optimal fuzzy partitions, in: J. Cybernet., 4, 1974, pp. 95–104.
[9] G. Gan, M. Chaoqun, J. Wu, Data Clustering Theory, Algorithms, and Applications, ASA-SIAM, Philadelphia, 2000.
[10] A. Garcia-Piquer, Facing-up challenges of multiobjective clustering based on evolutionary algorithms: representations, scalability and retrieval solutions, Research Group in Intelligent Systems, Campus LaSalle, Universitat Ramon Llull, 2012 Ph.D. thesis.
[11] A. Garcia-Piquer, A. Fornells, J. Bacardit, A. Orriols-Puig, E. Golobardes, Large-scale experimental evaluation of cluster representations for multiobjective evolutionary clustering, IEEE Trans. Evol. Comput. 18 (1) (2014) 36–53.
[12] A. Garcia-Piquer, A. Fornells, A. Orriols-Puig, G. Corral, E. Golobardes, Data classification through an evolutionary approach based on multiple criteria, Knowl. Inf. Sys. 33 (1) (2012) 35–56.
[13] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J.M. Prez, J.I. Martn, Towards a standard methodology to evaluate internal cluster validity indices, Pattern Recognit. Lett. 32 (3) (2011) 505–515.
[14] J.F. Hair, R.L. Tatham, R.E. Anderson, W. Black, Multivariate Data Analysis, 6th edition, Prentice Hall, 2006.
[15] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Sys. 17 (2001) 107–145.
[16] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Cluster validity methods: part I, SIGMOD Rec. 31 (2) (2002) 40–45.
[17] J. Handl, J. Knowles, An evolutionary approach to multiobjective clustering, IEEE Trans. Evol. Comput. 1 (1) (2007) 56–76.
[18] T. Harkison, J. Poulston, J.-H. G. Kim, Hospitality graduates and managers: the big divide, Int. J. Contemp. Hospitality Manag. 23 (2011) 377–392.
[19] D.A. Harrison, K.J. Klein, What's the difference? diversity constructs as separation, variety, or disparity in organizations, Acad. Manag. Rev. 32 (4) (2007) 1199–1228.
[20] N. Hemmington, From service to experience: understanding and defining the hospitality business, Serv. Ind. J. 27 (6) (2007) 747–755.
[21] E.R. Hruschka, R.J.G.B. Campello, A.A. Freitas, A.C.P.L.F. de Carvalho, A survey of evolutionary algorithms for clustering, IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev. 39 (2) (2009) 133–155.
[22] A.K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666.
[23] G. Kanji, Measuring business excellence, Routledge Advances in Management and Business Studies, 2002.
[24] C. Lashley, A. Morrison, et al., In search of hospitality: theoretical perspectives and debates, In Search of Hospitality: Theoretical Perspectives and Debates, 2000.
[25] C. Legány, S. Juhász, A. Babos, Cluster validity measurement techniques, in: Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, 2006, pp. 388–393.
[26] U. Nabitz, P. Severens, W.V.D. Brink, P. Jansen, Improving the EFQM model: an empirical study on model development and theory building using concept mapping, Total Qual. Manag. 12 (1) (2001) 69–81.
[27] S.R. Rosas, L.C. Camphausen, The use of concept mapping for scale development and validation in evaluation, Eval. Program Plan. 30 (2) (2007) 125–135.
[28] L. Roselló, F. Prats, N. Agell, M. Sánchez, Measuring consensus in group decisions by means of qualitative reasoning, Int. J. Approx. Reason. 51 (4) (2010) 441–452.
[29] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, in: J. Comput. Appl. Math., 20, 1987, pp. 53–65.
[30] C. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, 623–656.
[31] P. Slattery, et al., Finding the hospitality industry, J. Hospitality Leis. Sport Tourism Educ. 1 (1) (2002) 19–28.
[32] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 4th edition, Academic Press, Burlington, USA, 2008.
[33] L. Travé-Massuyès, L. Ironi, P. Dague, Mathematical foundations of qualitative reasoning, AI Mag. 24 (4) (2004) 91–106.
[34] L. Travé-Massuyès, F. Prats, M. Sánchez, N. Agell, Relative and absolute order-of-magnitude models unified, Ann. Math. Artif. Intell. 45 (3–4) (2005) 323–341, doi:10.1007/s10472-005-9002-1.
[35] W.M. Trochim, An introduction to concept mapping for planning and evaluation, Eval. Program Plann. 12 (1) (1989) 1–16. Special Issue: Concept Mapping for Evaluation and Planning.
[36] M. Vila, R. X., G. Costa, R. Santom, Combining research techniques to improve quality service in hospitality, Qual. Quant. 46 (2012) 795–812.
[37] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 3rd edition, Morgan Kaufmann, San Francisco, 2011.
[38] R. Wood, B. Brotherton, The SAGE Handbook of Hospitality Management, SAGE Publications, 2008.

Universitat Ramon Llull

Aquesta Tesi Doctoral ha estat defensada el dia ____ d _____ de ____

al Centre _____

de la Universitat Ramon Llull

davant el Tribunal format pels Doctors sotasignants, havent obtingut la qualificació:

President/a

_____

Vocal

_____

Vocal

_____

Vocal

_____

Secretari/ària

_____

Doctorand/a



C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es