UNIVERSITAT POLITÈCNICA DE CATALUNYA

Departament de Llenguatge i Sistemes Informàtics
Ph.D. Programme: Artificial Intelligence

SYMBOLIC AND CONNECTIONIST LEARNING TECHNIQUES FOR GRAMMATICAL INFERENCE

Autor: René Alquézar Mancho
Director: Alberto Sanfeliu Cortés

March 1997

A Ana, a Marc y Jéssica, y a mis padres

Acknowledgements

This doctoral thesis has been developed in the framework of a predoctoral research grant given by the *Generalitat de Catalunya* (by Resolution dated 29/10/91 and published in DOGC-1514-6/11/91) and a CICYT project (CICYT TAP96-0629-C04-02), that have provided economical and infrastructure support for this work. This support is gratefully acknowledged.

I would like to express my thankfulness to all the people who have contributed in one way or another to the work presented in this doctoral thesis. In particular, I would like to mention:

Dr. Alberto Sanfeliu, Professor in the Department of *Enginyeria de Sistemes, Automàtica i Informàtica Industrial* at the *Universitat Politècnica de Catalunya* (UPC) and researcher in the *Institut de Robòtica i Informàtica Industrial* (CSIC-UPC), for his continuous advice and guidance along the realization of this dissertation. He gave me his support in every single moment, something that I will never forget, and he also showed me to enjoy scientific research. My sincere acknowledgements and admiration to him.

Dr. Rafael Huber and *Dr. Pere Caminal*, Heads of the *Institut de Robòtica i Informàtica Industrial* and the *Institut de Cibernètica*, respectively, for their kind support during the evolution of this dissertation. They provided me with all the resources I needed of these research institutions.

Dr. Pere Brunet and *Dr. Fernando Orejas*, who have been successively Head of the Department of *Llenguatges i Sistemes Informàtics* (LSI) at the UPC since the time I entered this department, as well as *Dr. Núria Castell*, Head of the Artificial Intelligence Section, for giving me the necessary support to carry out this thesis. Likewise, I would like to thank the Department of LSI as a whole for offering me some teaching load reductions that have enabled me to devote more time to this work.

Dr. Ton Sales, Professor in the Department of LSI, who has been my tutor during the development of this dissertation, for his guidance in doctoral courses and his enjoyable lectures I have had the chance to attend.

Dr. Josep Maria Sopena, Professor in the Department of *Psicologia Bàsica* at the *Universitat de Barcelona*, for helping me to introduce into the field of neural networks, as well as for their valuable comments and opinions during the numerous meetings that we held together.

Dr. Paco Casacuberta and *Dr. Enrique Vidal*, Professors at the *Universitat Politècnica de València*, and also *Dr. José Oncina* and *Dr. Mikel Forcada*, Professors at the *Universitat d'Alacant*, for their estimable comments and scientific aid.

Dr. Pierre Dupont, currently in the Department of Computer Science at Carnegie Mellon University, for kindly providing me with his benchmark data for my experiments.

María Asunción Castaño (I hope to call her Dr. soon), Professor in the Department of *Informàtica* at the *Universitat Jaume I*, for her friendship and for sharing with me a common interest and activity in the areas of grammatical inference and neural networks.

Miguel Sainz, *Jordi Cueva*, *Martí Font*, and *Daniel Casillas*, all of them students and members of the research group led by Dr. Sanfeliu, for their collaboration in multiple aspects of the work and for their good mood and solidarity.

My sincere gratitude goes to my friend and office mate *Angela Nebot*, who is also Dr. and Professor in the Department of LSI, for her support and for sharing with me a lot of hard and happy moments during these years. I wish her all the best, she truly deserves it.

I would also like to thank the members (including professors, research fellows, administrative and technical staff) of the *Institut de Cibernètica*, the *Institut de Robòtica i Informàtica Industrial*, and the Department of LSI, who have been my companions during the latest years. Their support and friendship is highly appreciated.

Finally, I want to mention all the people that provided moral support in every single moment, especially my wife *Ana*, my parents *René* and *Aurea*, and my mother-in-law *Maria*. They gave me courage to continue with this research during the hard moments that all too often exist during the development of a doctoral thesis. I do not forget my children *Marc* and *Jessica*, whom I must ask their forgiveness for all the time I could not be with them due to this work.

Summary of the doctoral thesis

During the last thirty years, *learning* has become a foremost topic within the field of *artificial intelligence* and a great deal of effort has been devoted to it by researchers. The topic of learning has been divided into several areas as progress in artificial intelligence has developed. One of them is related to *inductive inference*, which can be defined as "the process of inferring general rules from examples, or theories from facts, carried out by an agent or system".

A great part of the work on the theory and methods for inductive inference has been done within Gold's general paradigm of inductive inference, in which an agent tries to identify a concept from a determined class of concepts or object space, subject to a certain success criterion, by collecting data and guessing a description of the concept, which is selected from a hypothesis space. Normally, the input data consist of examples (and maybe also counterexamples) of the target concept. The term *grammatical inference* (GI) has been used in this context to denote the inductive inference of classes of languages described by some appropriate representations like grammars or automata.

Although some theoretical limitations of grammatical inference are known, several practical GI methods have been proposed in the literature for a variety of language classes, typically regular languages or some subclass of regular languages, using both the classical *symbolic* paradigm and the more recent *connectionist* approach, in which some type of *artificial neural network* is used. Since GI can provide automatic methods for learning models of pattern classes in *syntactic pattern recognition* tasks, the interest in the development of powerful and flexible GI tools is clear. Indeed, some applications of GI have been reported in the areas of natural language modelling and translation, speech recognition, and computer vision.

The aim of the work developed in this dissertation was to study both symbolic and connectionist learning techniques for grammatical inference and their relationships, trying to develop new GI methods and/or improvements on the existing ones. The research on GI methods was restricted to the scope of two classes of languages, the *regular languages* and (some subclass of) the *context-sensitive languages* (CSLs), and to the case of learning from examples (positive and negative samples).

This thesis is structured in four parts for a total of ten chapters. The first part, *introduction and review* (Chapters 1 to 4), presents an extensive state-of-the-art review of both symbolic and connectionist GI methods, that serves also to state most of the basic material needed to describe later the contributions of the thesis. These contributions constitute the contents of the rest of parts (Chapters 5 to 10).

An introduction to the topic and objectives of the work is given in Chapter 1. The known theory and (symbolic) methods for regular GI are described in Chapter 2, while the symbolic methods proposed for the inference of *context-free languages* (CFLs), CSLs, and *pattern languages* are reviewed in Chapter 3. Chapter 4 introduces the neural network architectures, learning algorithms, and training procedures that have been used in the connectionist works on GI reported previously.

The second part, *contributions on symbolic and connectionist techniques for regular grammatical inference* (Chapters 5 to 7), describes the contributions related to the theory and methods for regular GI, which include other lateral subjects such as the representation of *finite-state machines* (FSMs) in *recurrent neural networks* (RNNs).

In Chapter 5, the so-called *unbiased finite-state automata* (or UFSAs) are defined and proposed as hypotheses for the problem of learning regular languages from both positive and negative examples. The theory of regular GI is reformulated according to this representation, and both non-incremental and incremental methods using UFSAs are presented and evaluated on a benchmark test. It must be noted that until very recently, there has been a lack of incremental methods for regular GI from both positive and negative examples. Moreover, in contrast with other methods, the consistency analysis provided by the use of UFSAs is symmetric and might easily be extended to the problem of learning recognizers of several non-overlapping classes (represented by regular languages) from examples belonging to these classes.

In Chapter 6, two (heuristic) connectionist approaches to regular GI based on RNNs are discussed, which involve training RNNs to learn the next-symbol prediction task (from a stochastic presentation of positive examples) or the string classification task (from a given set of positive and negative examples), respectively, and extracting afterwards an UFSA from the trained net. These approaches have been also validated on some benchmark tests used by different researchers, and the effects of varying the network architecture and activation functions have been assessed.

The proposed UFSA extraction methods improve other methods reported previously to extract an automaton from a trained RNN in the sense that they guarantee the obtaining of a consistent deterministic automaton in a single run. Another nice property is that they use an inductive scheme which is similar to the one used by most of the symbolic methods for regular GI (i.e. state merging from a canonical automaton). On the other hand, it is shown that by using some activation functions in RNNs different from the commonly employed sigmoid function the learning performance of these networks can be improved considerably, and a theoretical justification of this empirical result is given.

In Chapter 7, an algebraic framework to represent FSMs in RNNs is presented that unifies and generalizes some of the previous proposals and serves to explain the different representational capabilities of first-order and higher-order single-layer RNNs. This framework can be used to insert symbolic knowledge in RNNs prior to learning from examples and to keep this knowledge while training the network, and it has given rise to a hybrid semi-automated GI methodology, called *active grammatical inference* (AGI), which is described. In contrast with previous approaches, the insertion method proposed is valid for a wide range of activation functions, whenever some stability conditions are met.

The third part of the thesis, *augmented regular expressions and their inductive inference*, comprises Chapters 8 and 9. The *augmented regular expressions* (or AREs) are defined and proposed as a new representation for a subclass of CSLs that does not contain all the context-free languages but a large class of languages capable of describing patterns with symmetries and other (context-sensitive) structures of interest in pattern recognition problems.

In Chapter 8, AREs and their components are formally defined, and an efficient method to recognize a string as belonging or not to the language represented by an unambiguous ARE is described in detail. It is also demonstrated that AREs cover the class of pattern languages defined by Angluin. AREs constitute a better representation than *context-sensitive grammars* (CSGs) for the associated subclass of CSLs, since the language description is more intelligible, the recognition method is more efficient (at least for unambiguous AREs), and to the contrary of CSGs, practical learning algorithms to infer AREs from examples can be devised.

In Chapter 9, the inductive inference of AREs from string examples is discussed. A general method to infer AREs from examples is proposed that is based on a regular GI step followed by a constraint induction process. In addition, a specific method for learning AREs from positive examples, in which a connectionist regular GI technique is used, is presented, and the results of the application of this method to the inference of a set of eight test CSLs are reported.

It must be emphasized that work on CSL learning has been extremely scarce in the literature of GI and there is a lack of methods to infer descriptions of CSLs. However, there is a need for GI methods capable of inferring CSLs, specially for pattern recognition tasks in computer vision, where objects usually contain structural relationships that are not describable by regular languages or CFLs.

The fourth part of the thesis just includes Chapter 10: *conclusions and future research*. Chapter 10 summarizes the main results obtained and points out the lines of further research that should be followed both to deepen in some of the theoretical aspects raised and to facilitate the application of the developed GI tools to real-world problems in the area of computer vision.

The research effort documented in this dissertation has focused on methodological and theoretical issues. Therefore, the different empirical studies presented in this doctoral thesis serve only to demonstrate the feasibility and validity of the proposed methods. Nevertheless, prototype systems that use the GI techniques presented here have already been reported for some computer vision applications (e.g. recognition of traffic signs in outdoor scenes) by other members of our research group.

Index