Number 79 pp 285-287 February 2003

Author: José M. Sempere
Title: On some characteristic features of linear grammars and their application to the study of their identification and complexity
Language of presentation: Spanish
Promotor: Dr. Pedro García Gómez
Date of defence: November 21, 2002
Institution granting degree: Universidad Politécnica de Valencia, Spain

Abstract

This PhD thesis is related to three different research areas of theoretical computer science: Formal language theory ,complexity theory and computational (algorithmic-machine) learning theory. We can summarize this work as follows: We are interested in exploring the main dificulties of learning linear languages from different information sources (positive data, complete presentation and structural information). Given that we must fix a hypothesis space representation for the learning methods, we choose a generative approach and we use formal grammars. Then, our attention goes to different properties of linear grammars. These non trivial properties of linear grammars is what we call *characteristic features* We study four different characteristic features of linear grammars: local testability, reversibility, terminal distinguishability and structural distinguishability. We study some aspects of linear grammars with such characteristic features in order to obtain some conclusions for the learning problem. Specially, we focus our attention to the algorithmic difficulties of deciding if a given linear grammar holds a given characteristic feature. Furthermore, while studing the representation of linear grammars we obtain some conclusions about the descriptive complexity of linear languages.

The structure of this thesis is as follows. In chapter 1, we give a brief introduction to the different topics of the work and their relationships. In chapter 2, we give the basic concepts of computational learning theory. We give an overview of the different approaches to solve the learning problem and we define a taxonomy of the different methods based on the learning strategies and the information sources. Then, we explain the inductive inference approach and the grammatical inference paradigm. We summarize the different results of learning grammars and automata from different information protocols: learning from positive data, from complete information and from structural information. Last, we explain the relationship between this PhD work and the learning aspects that we have explained before. In chapter 3, we give the main concepts of formal language theory ,we define different aspects of languages, relations, grammars and automata. This is an introductory chapter devoted to formal languages.

The first step in learning linear grammars is learning even linear ones. In chapter 4, we study the learning problem related to even linear languages. First, we give the definitions, basic results and an overview of some previous works related to this topic. Then, we propose a reduction technique of even linear languages to regular ones and we take advantage of this reduction to solve the learning problem efficiently. Once the learning problem for general even linear languages is solved, we study local testability on even linear languages from different reduction strategies. We define several even linear language subclasses that can be learned from only positive data. We finish this chapter by sum-

marizing the basic results that we have obtained and by proposing new open problems.

In chapter 5, we study the learning problem of linear languages. As in the previous chapter, we start by giving the basic definitions of linear grammars and the most significant previous results related to the learning problem. We show our motivation to use structural information as an information source. This motivation is based on the difficulties of some problems related to linear grammars such as the equivalence and structural equivalence problems, the ambiguity problem, ans so on. Then, we continue this chapter by studing different characteristic features. First, we study terminal and structural distinguishability and we prove the (non) decidability of some of them. From terminal and structural distinguishability we deduce a finite index relation in the structural information and we profit from that to define Terminal and Structural Distinguishable Linear languages (TSDL). We give some relations between TSDL class and some other known language classes (finite, regular, even linear and linear). Then, we propose a learning algorithm to identify any TSDL language in the limit. The previous algorithm can be adapted to work with any finite product of TSDL languages as we show in this section. This chapter continues by studing two more characteristic features, reversibility and local testability. Both characteristics are defined by introducing a new reduction technique from linear grammars to even linear ones. In this case, we define the corresponding language classes to reversible linear languages and local testable linear ones. We finish this chapter by giving the main results that we have obtained and by introducing new open problems.

In chapter 6, we study some aspects about the complexity of linear languages. First, we introduce a family of regular-like expressions to define linear languages which we call *linear expressions*. Then, we propose some equivalence properties for linear expressions. Mainly, we work with permutation and compression equivalence properties. After proving the last properties, we study two different complexity measures for linear grammars. First of them, we study the reversal complexity of linear grammars and we propose an speed-up theorem in order to decrease the reversal complexity of any given grammar within a constant factor. Second, we initiate a preliminar study of the descriptive complexity of linear grammars by using Kolmogorov complexity. We bind the Kolmogorov complexity of any given linear language by using the previous equivalence properties and we relate the Kolmogorov and reversal complexities of linear grammars. We finish this chapter by summarizing the previous results and proposing new open problems.

Finally, chapter 7 is devoted to present the general conclusions of the thesis and propose some research guidelines for future works.

Table of contents

Author's correspondence address
José M. Sempere
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n
46020 Valencia, Spain
email: jsempere@dsic.upv.es