
Abstract of PhD Thesis

Author: José Luis Triviño-Rodríguez
Title: A new learning model of sequences of symbols
Language: Spanish
Supervisor: Rafael Morales-Bueno
Institute: University of Málaga, Spain
Date: 8 May 2002

Abstract

Many learning models have been developed in order to represent the evolution of a system. An important kind of these models represents the state of the system in every instant by a symbol. So, these models represent the evolution of a system by means of a sequence of symbols.

Markov chains is a widely used model that expresses the evolution of a system in terms of sequences of symbols. Moreover, there exist many models derived from Markov chains. These models represent the behaviour of the system by means of the probability distribution of every symbol given previous symbols in the sequence. *Markov Chains* and its derived models have been applied to model data sequences that exhibit the *short memory* statistical property. If we consider the (empirical) probability distribution on the next symbol given the preceding subsequence of some given length, then exists a length L (the *memory length*) such that the conditional probability distribution does not change substantially if we condition it on preceding subsequences of length greater than L . This feature can be found in many applications related with natural language processing such as speech recognition, and part of speech tagging.

This doctoral dissertation describes a multiattribute variable memory length Markov chain model. This model is called MPSA (Multiattribute Probabilistic Suffix Automata) and it has been developed like a generalization of the PSA model (Probabilistic Suffix Automata) described by Dana Ron. So, the Dana Ron's model can be seen as a MPSA with only one attribute.

A MPSA is hard to learn. So, only the learning of the probability distribution defined by one attribute has been developed. Therefore, the hypothesis model used by the learning algorithm is the MPSG model (Multiattribute Predictive Suffix Graph) instead of the MPSA model. Moreover, it has been proved that a MPSG can be learned from a polynomial size sample generated by a MPSA.

The main feature of MPSGs is that they combine the learning of sequences

of data like Markov chains and the learning based on several attributes like inductive learning. So, in order to learn a MPSG two tasks must be accomplished: to compute the next symbol probability distribution of the target attribute and to compute the needed information (memory length) of the rest of attributes in order to compute this probability distribution.

In order to show the practical application of the MPSG, two applications of this model have been describe in this thesis: part of speech tagging and music prediction and generation.

On the one hand, a spanish part of speech tagger (POS Tagger) has been developed using the MPSG model. This tagger has been trained with the LexEsp corpus of the UPC. The tagging accuracy of the MPSG tagger is similar to the accuracy of the best spanish taggers.

On the other hand, a model of the Bach's chorales has been computed using the MPSG model. Later, several new pieces of music have been generated following this model. It has been proof that these pieces have the same probability distribution that Bach's chorales. Moreover, we performed an auditory test where subjects guessed whether fragments they heard came from the original chorale or from the MPSG simulation. Fifty-two listeners evaluated the music from the MPSG with this test. They correctly classified the fragments about 55% of the time.

Table of Contents

1 Introduction	1
2 Time modeling in AI	11
2.1 Introduction	11
2.2 Time ontology	13
3 The time dimension in ML	17
3.1 Introduction	17
3.2 Behaviour pattern learning	19
3.3 Learning models of sequences of symbols	22
3.4 The time dimension in other models of ML	70
4 MPSA and MPSG models	75
4.1 Introduction	75
4.2 MPSA	79

4.3 Multiattribute Prediction Suffix Graph (MPSG)	85
4.4 MPSG vs Decision Trees	96
5 MPSG learning algorithm	103
5.1 Implementation	103
5.2 Analysis of the learning algorithm	107
6 Applications of the MPSG model	115
6.1 Part of speech tagging	115
6.2 Using MPSG to predict of generate music	125
7 Coclusions and future work	143
A MPSG learning subalgorithms	165
B Algorithm for sequences generation	175
C Doménico Scarlatti	179

Author's correspondence address José Luis Triviño-Rodríguez
E.T.S. Ingeniería Informática
Boulevard de Louis Pasteur, 57
Campus Teatinos
29071 - Málaga
Spain