Available online at www.sciencedirect.com
SCIENCE DIRECT®
Environmental Modelling & Software 20 (2005) 753–760
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Modeling time series of climatic parameters with probabilistic finite automata

L. Mora-López[a,*], J. Mora[b], R. Morales-Bueno[a], M. Sidrach-de-Cardona[c]

[a] *Dpto. Lenguajes y C. Computación, E.T.S.I. Informática, Universidad de Málaga, Campus Teatinos, 29071 Malaga, Spain*
[b] *Dpto. Fundamentos del Análisis Económico, Universidad de Alicante, Apdo. Correos 99, E-03080 Alicante, Spain*
[c] *Dpto. Física Aplicada II, E.T.S.I. Informática, Universidad de Málaga, Campus Teatinos, 29071 Malaga, Spain*

Received 1 November 2002; received in revised form 22 October 2003; accepted 20 April 2004

Abstract

A model to characterize and predict continuous time series from machine-learning techniques is proposed. This model includes the following three steps: dynamic discretization of continuous values, construction of probabilistic finite automata and prediction of new series with randomness. The first problem in most models from machine learning is that they are developed for discrete values; however, most phenomena in nature are continuous. To convert these continuous values into discrete values a dynamic discretization method has been used. With the obtained discrete series, we have built probabilistic finite automata which include all the representative information which the series contain. The learning algorithm to build these automata is polynomial in the sample size. An algorithm to predict new series has been proposed. This algorithm incorporates the randomness in nature. After finishing the three steps of the model, the similarity between the predicted series and the real ones has been checked. For this, a new adaptable test based on the classical Kolmogorov–Smirnov two-sample test has been done. The cumulative distribution function of observed and generated series has been compared using the concept of indistinguishable values. Finally, the proposed model has been applied in several practical cases of time series of climatic parameters.

Keywords: Machine learning; Modeling climatic data; Time series

Software availability

Name: SIMULTS
Developer: L. Mora-López
Contact address: llanos@lcc.uma.es. Tel.: +34-95-213-2802; fax: +34-95-213-1397
Year first available: 2002
Hardware required: Pentium II (minimum)
Software required: None
Program language: C
Program size: 314 K
Cost: Free

1. Introduction

The fundamental idea in this paper is the use of probabilistic finite automata (PFA) as a means of representing the relationships observed in stationary time series. PFA are mathematical models used in the machine-learning field. Traditionally, the analysis of stationary time series has been carried out using stochastic process theory. One of the most detailed analyses of statistical methods for time series research was done by Box and Jenkins (1976). The goal of these methods is to find models which are able to reproduce the statistical and sequential characteristics of the series. Usually, the approach is as follows (see Box and Jenkins, 1976): first, the recorded series are statistically analyzed in order to select the best model for the series. Then the parameters of the model are estimated. After this, a new series of values can be generated using the estimated model. For example, for solar radiation series

* Corresponding author. Tel.: +34-95-213-2802; fax: +34-95-213-1397.
E-mail addresses: llanos@lcc.uma.es (L. Mora-López), juan@merlin.fae.ua.es (J. Mora), morales@lcc.uma.es (R. Morales-Bueno), msidrach@ctima.uma.es (M. Sidrach-de-Cardona).

1364-8152/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2004.04.007

this approach has been followed in Brinkworth (1977), Bendt et al. (1981), Aguiar et al. (1988), Aguiar and Collares-Pereira (1992), and Mora-López and Sidrach-de-Cardona (1997). Many of these models can only generate new sequences of values which present normal probability distribution functions. However, the original series do not have this type of probability distribution. Other types of time series which have been analyzed using some class of stochastic models are shown in McMillan et al. (2000) and Anh et al. (1997).

On the other hand, the analysis of time series and stochastic process has also been analyzed from machine-learning techniques. When a time series presents a probabilistic behavior, some machine-learning models could be very useful to study it. In these series the recorded variables are insufficient to exactly determine the future values, due to the random nature of these variables. The systems in which these models can be used must have the following properties:

To present probabilistic behaviour or uncertainty. This uncertainty can be due to several factors. For example, for the prediction of climatic variables the number of parameters which affect them is very high.
Although there is uncertainty in these systems, there is always some structure within this uncertainty.

For example, the machine-learning models based on probabilistic finite automata have been used to model several types of natural sequences. Examples of such applications are: universal data compression (Rissanen, 1983), analysis of biological sequences, for DNA and proteins (Krog et al., 1993), analysis of natural language, for handwriting and speech (Nadas, 1984; Rabiner, 1994; Ron et al., 1998), etc. Different classes of automata have been developed. For instance, acyclic probabilistic finite automata have been used for modeling distributions on short sequences (Ron et al., 1998); probabilistic suffix automata, based on variable order Markov models, have been used to construct a model of the English language (Ron et al., 1994). All these automata allow us to take into account the temporal relationships in a series. In a different way, other approaches from machine learning have been used to model climatic parameters. For instance, neural networks have been used by Mohandes et al. (1998) and Kemmoku et al. (1999) on the characterization of one climatic parameter (daily global solar radiation). The main problem of this approach is that the obtained models are 'black boxes', and no significant information can be obtained from them.

Other works arise from the ideas developed by Dagum, based on belief network models: it is proposed the use of dynamic network models, which are a compromise between belief network models and classical models of time series. They are based on the integration of fundamental methods of Bayesian analysis of time series.

This paper describes how to use certain models from the machine-learning field in the analysis and prediction of climatic parameters. The model we propose is based on the Probabilistic Finite Automata (PFA) theory. Our goal is to use PFA to represent all the relationships observed in stationary climatic time series and to use these PFA to predict new values of the series. The use of this model allows us to represent and generate time series with non-normal probability distribution functions and to obtain information about the nature of the analyzed series. Moreover, an adaptable test based on the classic Kolmogorov–Smirnov two-sample test has been used to check the proposed model. Finally, preliminary results of the model obtained for climatic parameters are shown.

2. Probabilistic finite automata

We propose using a mathematical model called probabilistic finite automata (PFA). We propose the use of this mathematical model to represent a stationary univariate time series. Formally, a PFA is a 5-tuple $(\Sigma,Q,\tau,\gamma,q_0)$ where (see for instance, Ron et al. (1998) or Mora-López and Sidrach-de-Cardona (2003)):

Σ is a finite alphabet; that is, a set of discrete symbols corresponding to the different continuous values of the analyzed parameter. The different symbols of Σ will be represented by x_i. For a series, the values observed can be $x_5x_3,\ldots x_3$ To represent the different observable series for a period t_1 to t_m we will use the symbols $y_1y_2\ldots y_m$. So, in the series $x_5x_3\ldots x_3$, the symbol y_1 corresponds to the value x_5, the symbol y_2 to x_3 and so on.

Q is a finite collection of states. Each state corresponds to a subsequence of the discretized time series. The maximum size of a state—number of symbols—is bounded by a value N fixed in advance. This value is related to the number of previous values which will be considered to determine the next value in the series and depends on 'memory' of the series. The 'memory' of one series could be estimated both empirically and using information about the parameter which is analyzed. In the first case, an iterative process could be used: a small value is selected at the beginning; this value is increased in each iteration, and the results obtained with the previous value of memory and the actual value are compared; the process continues until the models do not improve when the memory increases. In the second case, the previous information about the analyzed parameter could be used to initialize the value of the memory in the iterations.

τ: $Q\times\Sigma\rightarrow Q$ is the transition function
γ: $Q\times\Sigma\rightarrow[0,1]$ is the next symbol probability function
$q_0\in Q$, is the initial state

The function γ satisfies the following requirement: For every $q\in Q$ and for every $x_i\in\Sigma$, $\Sigma_{xi}\in\Sigma\gamma(q,x_i)=1$. Moreover, the following conditions are required: