Modelling time series with data mining [1]

L.Mora and R.Morales

Dpto. Lenguajes y C.Computación. E.T.S.I.Informática.
Campus de Teatinos. 29071 Málaga, Spain
{llanos,morales}@lcc.uma.es

Abstract. Some of the models obtained in the study of time series using data mining in the group of Málaga are presented. Firstly, two methods to assign discrete values to continuous values from time series are proposed; these methods use dynamic information about the series. The first method is based on a particular statistic which allows us to select a discrete value for a new continuous value from the series. The second one is based on the proposed concept of significant distance between consecutive values from time series. Secondly, the use of probabilistic finite automata to model time series are described. Finally, an algorithm to generate time series with the same statistical properties that the real ones is presented.

1 Introduction

The goal of data analysis by time series is to find models which are able to reproduce the statistical characteristics of the series. Moreover, these models allow us to predict next values of the series from its predecessors.

One of the most detailed analysis of statistic methods for the research of time series has been done by Box & Jenkins [2]. The mathematical model for a time series is the concept of discrete-time stochastic process. It is supposed that the observed value of the series at time t is a random sample of size one from a random variable X_t, for $t \in \{1, ..., n\}$. A time series of length n is a random sample of a random vector like this $(X_1, ..., X_n)$. The random vector is considered as part of a discrete-time stochastic process, and observed values of the random variables are considered as the evolution of the process. The process is completely known if the joint probability distribution function of each random vector is known.

The following steps are pursued in the analysis of data using time series theory: identification of the model, estimation of parameters, diagnosis of the model and prediction of new values. The identification of the model can be achieved either in the time domain, using the sample and partial autocorrelation functions, or in the frequency domain, using spectral analysis. In both cases, a previous selection of the possible models which can be used to fit the data must be

[1] This work has been partially supported by MOISES project number TIC2002-04019-C03-02 of the CICYT, Spain.

done. This can be a restriction in the final results. Another important restriction is the following: once the model has been identified and the parameters have been estimated it is supposed that the relation between the parameters is constant along the time. However, in many time series this can not be true.

On the other hand, the analysis of time series has also been analyzed using machine learning techniques. Some of these techniques have solved successfully the restrictions noted above. Two of the most important works in this line have been developed by D.Ron et al., [14], [15], where the use of probabilistic finite automata is proposed. In [15], a subclass of probabilistic finite automata has been used for modeling distributions on short sequences that correspond to objects such as single handwritten letters, spoken words, or short protein sequences. In [14], another subclass of probabilistic finite automata, called probabilistic suffix automata, has been used to describe variable memory length Markov processes. Other works arise from the work developed by Dagum [3], based on belief network models; in [3], 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. However, almost all models used for time series from machine learning are restricted to input features with known discrete values, not allowing continuous valued features as input. For this reason, before any of this method is used, it is necessary to transform the observed continuous values into discrete values. In Section 2 we describe our two approaches to obtain discrete values. We explain the classical static conversion and the two dynamic conversion that we propose.

In the third section of this paper, we present how the probabilistic finite automata could be used both to represent the relationships observed in stationary time series and to obtain simulated series with the same statistical properties that the real ones.

2 Dynamic Qualitative Discretization

Any method to obtain discrete values must have the following two features: first, it must be known how many different discrete values can appear in the series; second, it must be able to quantify how different two or more consecutive values of the series are.

A possible way to transform into discrete values consists of using fixed-size intervals. We refer to this transformation as static discrete conversion. In the paper [8], we propose a transformation which we will refer to as dynamic discrete conversion. The static discrete conversion methods group a set of items into a hierarchy of subsets whose items are related in some meaningful way. Typically, these algorithms perform the conversion into discrete values according to statistical stationary properties of the values, not taking into account the evolution of these values. This procedure has various problems. Consider, for example, the following time series corresponding to a series of cloudiness index series: {0.71, 0.89, 0.89, 0.91, 0.89}. Using the proposed static discrete conversion and

using a description for each value, we obtain the series: {half overcast, almost completely overcast, almost completely overcast, overcast, almost completely overcast}. However, if we observe the series -or this situation in the real world-, we will probably not consider as different situations those when the cloudiness index take value 0.89 or 0.91. To circumvent this problem, a new approach is developed in the cited paper to transform continuous values into discrete ones. We refered to it as dynamic qualitative discrete conversion: dynamic because it takes into account the evolution of the series; and qualitative because the selection of the discrete value is based on a significant distance which is defined below.

Qualitative models have been used in different areas in order to get a representation of the domain based on properties (qualities) of the systems which, additionally, allows us to avoid the use of complex mathematical models, [4], [5]. One of the objectives which has been pursued is to develop an alternative physics in which the concepts are derived from a far simpler, but nevertheless formal, qualitative basis. Qualitative reasoning can also be used to predict the behavior of systems, [7].

On the other hand, any process of discretization has some psychological plausibility since in many cases humans apparently perform a similar preprocessing step representing temperature, weather, speed, etc., as nominal (discrete) values. Following [16], the desirable attributes for a discretization method are:

- Measure of classification "goodness"
- No specific closeness measure
- No parameters
- Globality rather than locality
- Simplicity
- Use of feedback
- Use of *a priori* knowledge
- Higher order correlations
- Fast

When using a static discrete conversion method, the continuous values are transformed into s discrete values through s intervals of same length. Specifically, the width w_X of a discretized interval is given by:

$$w_X = \frac{\max\{X_t\} - \min\{X_t\}}{s}, \tag{1}$$

where, hereafter, max and min are always considered for $t \in \{1, ..., n\}$. The discrete value v_i corresponding to a continuous value X_i of the series is an integer from 1 to s which is given by:

$$v_i = discretize(X_i) = \begin{cases} s & \text{if } X_i = \max\{X_t\} \\ [(X_i - \min\{X_t\})/w_X] + 1 & \text{otherwise} \end{cases}, \tag{2}$$

where $[A]$ means the integer part of A. After deciding upon s and finding w_X, it is straightforward to transform the continuous values into discrete ones using this expression.

We proposed two different procedures to obtain time series with discrete values taking into account the preceding values for the discretization of each value.

2.1 Using a t statistic

The idea behind this method is to use statistical information about the preceding values observed from the series to select the discrete value which corresponds to a new continuous value of the series. A new continuous value will be associated to the same discrete value as its preceding values if the continuous value belongs to the same population. Otherwise, the static discrete conversion method will assign a new discrete value to this new continuous value. To decide if a new continuous value belongs to the same population as the previous ones, a statistic with Student's t distribution is computed. The method is formally described below.

Given a set of observations, $X_1, ..., X_n, X_{n+1}$, it is possible to examine whether X_{n+1} belongs to the same population as the previous values using the statistic:

$$t_{observed} = \frac{X_{n+1} - \bar{X}}{\sqrt{\hat{\sigma}^2(1 + 1/n)}}, \tag{3}$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, and $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. As it is proved in [8], if certain statistical conditions are met, when X_{n+1} comes from the same population as the previous values the statistic $t_{observed}$ has Student´s t distribution with $n-1$ degrees of freedom. The algorithm can be found in [8]

2.2 Using qualitative reasoning

This method is based on the ideas of qualitative reasoning. In order to characterize the evolution of the system and select discrete values, we proposed to use distance functions. These distance functions measure the relationship between consecutive values. They have been used in Instance-Based learning, to determine how close a new input vector is to each stored instance, and use the nearest instance or instances to predict the output class. Therefore, distances are often normalized by dividing the distance for each attribute by the range (i.e. the difference between maximum and minimum) of that attribute, so that the distance for each attribute is in the approximate range $[0, 1]$. It is also common to use standard deviation instead of range in the denominator. Domain knowledge can often be used to decide which method is most appropriate.

We defined the concept of significant distance between values of the series: two consecutive continuous values correspond to the same discrete value when the distance between them is smaller than a threshold significant distance. This significant distance can be absolute (ASD) -the same for all the sequence- or relative (RSD) to the values which are being compared. We proposed the use of the following expressions for these two distance functions:

$$ASD = \frac{|X_i - X_j|}{range\{X_t\}}, \tag{4}$$

$$range\{X_t\} = \max\{X_t\} - \min\{X_t\}, \tag{5}$$

$$RSD = \frac{|X_i - X_j|}{|X_i|}. \tag{6}$$

The proposed expression for the ASD is based on the euclidean metric distance, [17]. The new discrete value is determined depending on how far it is from the preceding values. Changes above the threshold involve changes in the discrete value. When this procedure is used, smooth changes may not be detected, especially if the time series evolves slowly but always in an increasing or decreasing way. For instance, in the time series {0.87, 0.88, 0.89, 0.90, 0.91, 0.92, 0.93, 0.94, 0,95, 0.96} all continuous values would be assigned to the same discrete value. To solve this problem we propose to consider only the most recent values of the series to estimate the significant distance.

The first continuous value of the time series is used as reference value. The next values in the series are compared with this reference. When the distance between the reference and a specific value is greater than the threshold (there is a significant difference between them), the comparison process stops. For each value between the reference and the last value which has been compared, the following distances are computed: distance between the value and the first value of the interval, and distance between the value and the last value of the interval. If the former one is lower than the latter one, the discrete value assigned is the one corresponding to the first value; otherwise, the discrete value assigned is the one corresponding to the last value. The algorithm can be found in [8].

3 Probabilistic finite automata and time series

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.
- Although there is uncertainty in these systems, there is always some structure within this uncertainty.

The models based on probabilistic finite automata have been used to model several types of natural sequences. Examples of such applications are: universal data compression, [12], analysis of biological sequences, for DNA and proteins, [6], analysis of natural language, for handwriting and speech, [10], [11] and [15], etc. Different classes of automata have been developed. For instance, acyclic

probabilistic finite automata have been used for modeling distributions on short sequences, [15]; probabilistic suffix automata, based on variable order Markov models, have been used to construct a model of the English language, [14] and to model climatic parameter [9]. All these automata allow us to take into account the temporal relationships in a series. Moreover, in [9] a method to predict new values for a time series is presented.

Formally, a PFA is a 5-tuple $(\sum, Q, \tau, \gamma, q_0)$ where (see for instance, [15]):

- $\sum$ 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 $\sum$ will be represented by x_i.
- Q is a finite collection of states. Each state corresponds to a subsequence of the discretized time series.
- $\tau : Q \times \sum \rightarrow Q$ is the transition function
- $\gamma : Q \times \sum \rightarrow [0, 1]$ is the next symbol probability function
- $q_0 \in Q$, is the initial state

Once the PFA is built, it can be used as a mechanism for generating finite sequences of values in the following manner:

- Start from an initial value selected from the alphabet, called the initial state.
- If q_t is the current state, labeled by the sequence $Y = y_1...y_t$, then the next symbol is chosen (probabilistically) according to $\gamma(q_t, \cdot)$.
- If $x \in \sum$ is the chosen symbol, then the next state, q_{t+1}, is $\tau(q_t, x)$.
- The label of this new state, Y, will be the longest final subsequence of Yx in the PFA.

The process continues until the length of the required sequence is reached. This algorithm is very useful to simulate the long term behaviour of a time series.

With the PFA and the generation method described, new values for the time series can be generated. In order to compare the simulated series to the real ones, several statistical tests can be used. The hypothesis that both series have the same mean and variance can be checked. The frequency histograms of the recorded and simulated series can be also analyzed. To make this comparison, in [9] we have proposed the use of an adaptable goodness-of-fit test, which is based on the two-sample Kolmogorov-Smirnov test, described in [13]; we have used this test instead of an ANOVA test for the mean or Levenes test for the variance-because it captures both the differences in mean and variance and the differences in any characteristic of the probability distribution function. The objective of this adaptable test is to determine if two distribution functions $F_Y(.)$ and $F_Z(.)$ are the same, except for possible changes in location and scale. Specifically, we have checked the null hypothesis that there exist two unknown values m and s such that Z_i and $m + sY_j$ have the same distribution. Replacing unknown parameters m and s by estimates introduces additional random terms in the statistic and traditional critical values cannot be used. Therefore, to obtain the critical values that must be used in the test, we propose using a bootstrap procedure. This procedure is described in [9].

References

1. AGUIAR, R.J., COLLARES-PEREIRA, M., CONDE J.P. Simple procedure for generating sequences of daily radiation values using a library of Markov Transition Matrix. *Solar Energy,* 40, 269-279, 1988.
2. BOX, G.E.P., JENKINS, G.M. *Time Series Analysis forecasting and control.* USA: Prentice Hall, 1976.
3. DAGUM, P., GALPER, A. Time series prediction using belief network models. *Int.Journal Human-Computer Studies*, 42, 617-632, 1995.
4. FORBUS, K.D. Qualitative Process Theory, *Artif. Intell.* 24: 85-168, 1984.
5. KLEER, J. BROWN, J.S. A qualitative physics based on confluences, *Artif. Intell.* 24: 7-83, 1984.
6. KROG, A, MIAN, S.I., HAUSSLER, D. A hidden Markov model that finds genes in E.coli DNA. *Technical report UCSC-CRL-93-16*, University of California at Santa-Cruz, 1993.
7. KUIPERS, B. Commonsense reasoning about causality: deriving behavior from structure, *Artif. Intell.* 24: 169-203, 1984.
8. MORA LOPEZ, L., FORTES, I., MORALES BUENO, R., TRIGUERO, F. Dynamic discretization of continuous values for time series. *Lecture Notes in Artifical intelligence*, Vol. 1810, 280-291, 2000.
9. MORA LOPEZ, L., MORA, J., MORALES BUENO, R., SIDRACH DE CARDONA, M., Modelling time series of climatic parameters with probabilistic finite automata, *Environmetal Modelling and Software*, (in press).
10. NADAS, A.. Estimation of probabilities in the language model of the IBM speech recognition system. *IEEE Trans. on ASSP*, 32(4), 859-861, 1984.
11. RABINER, L.R.. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the Seventh Annual Workshop on Computational Learning Theory*, 1994.
12. RISSANEN, J., 1983. A universal data compression system. *IEEE Trans. Inform. Theory*, 29(5), 656-664.
13. ROHATGI, V.K., 1976. "An Introduction to Probability Theory and Mathematical Statistics". John Wiley & Sons, USA.
14. RON, D., SINGER, Y., TISHBY, N. Learning Probabilistic Automata with Variable Memory Length. *Proceedings oof the Seventh Annual Workshop on Computational Learning Theory*, 1994.
15. RON, D., SINGER, Y., TISHBY, N. On the Learnability and Usage of Acyclic Probabilistic Finite Automata. *Journal of Computer and System Sciences*, 56, 133-152, 1998.
16. VENTURA, D., MARTINEZ, T.R. BRACE: A Paradigm for the Discretization of Continuously Valued Data. *Proceedings of the Seventh Florida Artificial Intelligence Research Symposium*, 117-21, 1994.
17. WILSON, D.R., MARTINEZ, T.R. Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, 6, 1-34, 1997.