Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

# Data mining in temporal sequences: a technique based on MC

S. Massa[1], P.P. Puliafito[1]
[1]*DIST, Department of Communication, Computer and System Sciences, Genoa University, Italy.*

## Abstract

Data Mining is a new research area that develops techniques for knowledge discovery in massive amounts of data. The way such data are collected and modeled is quite important to define which kind of knowledge could be investigated and consequently to develop proper methods. One of the most investigated topics in data mining is the discovery of knowledge about data organized as temporal sequences. This work proposes a new modeling technique to mine information from data that are expressed in the form of temporal sequences of events associated to entities. In particular such a technique aims at extracting non-evident behavioral patterns from data in order to identify different classes of entities in the considered population. In order to represent the behavior of the entities, a Markov chain model is adopted and the transition probabilities for such a model are computed. The information extracted by means of the proposed technique can be used as decisional support in a large range of problems, such as marketing or social behavioral questions. A case study concerning the university dropout problem is presented together with further developments of Markov chain modeling technique in order to improve the prediction and/or interpretation power. The technique based on Markov chains is used for defining a dynamic framework that gives a concise representation of the problem as well as the opportunity to perform local drill down activities to improve the analysis. In particular the proposed interface gives the decision-maker the possibility to make access to a different kind of information derived from the analysis of the individual patterns followed by the students. Experimental results on the data about the students of University of Genova, Faculty of Engineering will be presented in order to show the accuracy of the proposed method and its potential application to the decisional process.

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X


# 1. Introduction

Temporal sequences of collected data are one of the most common ways to describe a phenomenon, which evolves through time. They can concern several topics such as economics, marketing, social behavioural matters and so on. The extraction of new knowledge through the analysis of data organised as temporal sequences, is one of the most investigated topics in data mining. Many authors deal with temporal sequences or time series of data ([1], [2], [3]) in order to identify repeated behavioural patterns and then to improve their knowledge about the dynamics of the observed system.

This work proposes to describe the evolution of a system through the definition of the states, which the system goes across. Then it suggests adopting Markov chain theory to define the transition probabilities among such states. In particular such a technique aims at extracting non-evident behavioural patterns from data in order to identify different classes of entities in the considered population. The case study we present concerns the application of the Markov chain based modelling technique to the problem of dropout at University in order to extract non-evident behavioural patterns from data and then to identify different classes of entities in the considered population. Specifically we consider the data about the students of University of Genova, Faculty of Engineering, who have been examined for the twenty-year period 1978-1998.

A further improvement of the modelling technique will be presented to face the problem of drop of statistical support when the number of attributes characterising the state vector increases. Actually, when we need to include more than a couple of attributes in our analysis, the number of the states of the system increases and then the number of entities per state, namely the support of the state, can become insufficient. The proposed improvement consists in performing the principal components analysis (PCA) to identify the principal components of the observed system and then carry out the Markovian analysis for each component separately. Such an analysis should not have any problem of lack of support, being performed for each component, which is a linear combination of attributes, at a time. Moreover these Markovian analyses, which have been obtained from the uncorrelated attributes resulting from PCA, can be considered as independent. After such Markovian analyses have been computed, thanks to their approximate independence, we can gain information about the joined effect of the various PC on the transition probabilities of the system, simply by multiplying the relative Markov chains.

The paper is organised as follows. It begins with a short introduction to the problem we mean to address. Then the Markov chains are briefly introduced and the proposed modelling technique is explained step by step with reference to dropouts. Then the further development of the work is presented. The PCA is introduced and its application to the dropout problem is discussed.

# 2 Application of Markov chains to the mining of time series

The theory of Markov chains ([4], [5], [6]) is often used to describe the system asymptotic behaviour by means of relevant simulation algorithms (Gibbs

sampling [7], Metropolis, Metropolis-Hastings [8]). The use of Markov chains simplifies the modelling of a complex, multi-variant population by focusing on the information associated with the system state.
This basic property of Markov chains allows to describe easily the behaviour of systems whose evolution can be modelled by a sequence of stochastic transitions from one state to another in a discrete set of possible states, which occur in correspondence of time instants or events.
The class of addressed problems takes the form of time series analysis to extract non-evident behavioural pattern from data.
Time series analysis is a typical subject for data mining that can model a wide range of real cases; examples can be reported from economic-financial problems as the analysis of sale trends and of price and market behaviour, from medical and diagnostics problems and from environmental contexts.
In the next sections a modelling technique aiming at applying Markov chain theory to data mining problems, which can be modelled with time-series, is presented with particular reference to the analysis of University dropouts.

## 2.1 Definition of the problem

In general, given a population made of a finite number of entities, each entity can be associated with a series of successive events that characterise its behaviour.
Let e be a generic entity from the considered population and $<s_1\ s_2\ s_3 \ldots s_n>$ a sequence of successive events. Then the association between the entity e and its relevant series of events can be written as follows:

$$e \leftrightarrow <s_1\, s_2\, s_3 \ldots s_n> \tag{1}$$

where n = n(e) is the number of events of the series.
Let us consider the administrative database of a University [9]. Specifically the students are the entities and the events are their passed examinations whereas time represents the distance (in months) from the matriculation date.
Information about the passed examination can be aggregated to form the student's state vector. In theory such a state vector could include several aggregated variables beyond the number of passed exams, such as the average mark, the average difficulty degree of the passed exams, the number of gained credits and so on. Unfortunately a full description of a student's situation through his state vector is not possible in this way of proceeding because of the already mentioned problem of support.
The concept of support is crucial in data mining in order to give a measure of the statistic importance of probabilistic information resulting from this kind of analysis. In our case the support gives a measure of the statistic importance of the transactions leaving from a state $x_i$ at time k. The support of the considered state that can be defined as follows:

$$\sigma(\mathbf{x}_i^k) = \frac{\mathbf{n}_i^k}{\sum_i \mathbf{n}_i^k} \tag{2}$$

© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X


By increasing the number of states, the number of students for each state ($n_i^k$) decreases and consequently the support of each state drops. Afterwards in the next section we'll propose a technique that aims at getting round the problem.
The transition probability among the feasible states can be computed ignoring the previous history, that is the main advantage of using the Markov chain.
The final result of the process computing transition probabilities is a sequence of matrices $T^{(k)}$ , corresponding to the transition from states in the k-th stage to states in the (k+1)-th. For each stage there are two absorbing states, one associated with degree (G) and the other one with dropout (D).

## 2.2 Preliminary results

In this chapter the students' behaviour is summarised through some graphs. The presented graphs are based on the data relative to the University of Genoa, Faculty of Engineering and they refer to the students that were attending University in the years between 1978 and 1998. The observed sample consists of about 15000 students relative to the twenty-year period of time under consideration.
Figure 1 is based on the dropout probabilities previously computed for each state. Here a white ball represents a state where such a probability is under a fixed threshold while a black ball a state where the same threshold is exceeded.

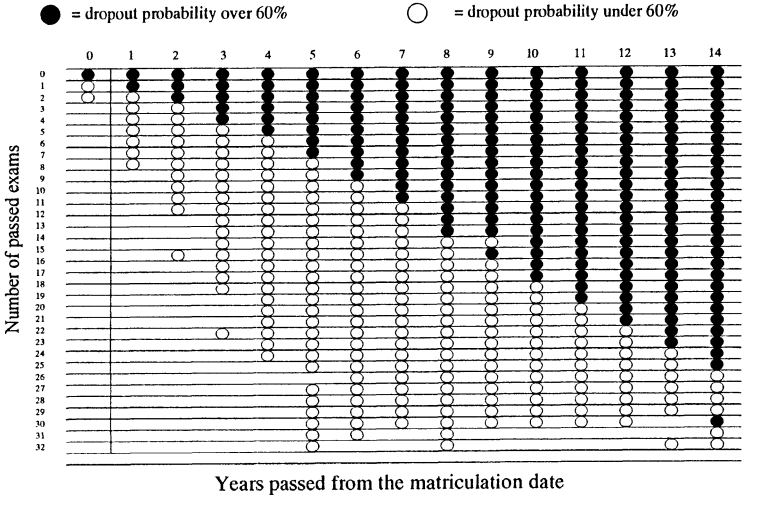



Figure 1: Dropout risk-zone.

Figure 1 refers to a dropout threshold of 60% but such a threshold may be defined to consider possible deviations in the dropout rate throughout the time or to perform special advising policies.

© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X


The performed Markovian analysis can have two main applications in the present context.
First, it helps at positioning a single student on the basis of his "state" in a framework, which gives him a measure of his dropout risk level. This first screening can be followed by further more accurate analysis of individual behaviour, obtained, for example, through drill down activities.
Second, Markovian analysis can lead to the construction of clusters of students combined by the dropout risk level. Such clusters can be useful to identify appropriate actions to try to influence the behaviour of the dropout risk students and, as a consequence, the evolution of their careers.
Being $T^{(k)}$ dependent on those actions, the fact that the transition probabilities $T^{(k)}$ could significantly change in the long run, can be inferred. In this case the model based on Markov chains could be used to the purposes of planning and control.
Both these applications of Markovian analysis would take advantage from a better description of the system's state. In particular the analysis of individual behaviour could be more accurately performed if more attributes characterising the student were included in the analysis, but, of course, this would imply a larger state vector and consequently a lower support.
In the following section a technique based on PCA will be proposed, in order to minimise the effects of the above-mentioned problem.

## 3. Principal components analysis

The method of principal components is primarily a data-analytic technique that makes linear transformations of a group of correlated variables so that the transformed variables are uncorrelated.
There have been a lot of applications of PCA over the years concerning several research area, such as psychology and education [10], quality control ([11], [12]), chemistry ([13], [14]), market research ([15], [16]), economics ([17], [18]), etc.

### 3.1 The method of PC

Let us suppose to have a data set consisting of $p \times 1$ vectors of observations on the $p$ original variables of the problem. The starting point for PCA is the sample covariance matrix **S**. For a $p$-variable problem,

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{1p} & s_{2p} & \cdots & s_p^2 \end{bmatrix} \tag{3}$$

where $s_i^2$ is the variance of the i-th variable, $x_i$, and $s_{ij}$ is the covariance between the i-th and the j-th variables. Being **S** a $p \times p$ symmetric, non singular matrix, it

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X


may be reduced to a diagonal matrix **D** by pre-multiplying and post-multiplying it by a particular orthonormal matrix **U** so that

$$\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U}=\mathbf{D} \tag{4}$$

The diagonal elements of **D**, $d_1$, $d_2$... $d_p$ are called the characteristic roots or eigenvalues of **S**. The columns of **U**, $\mathbf{u}_1$, $\mathbf{u}_2$,..., $\mathbf{u}_p$ are called the characteristic vectors or eigenvectors of **S**. The characteristic roots may be obtained from the solution of the following characteristic equation:

$$|\mathbf{S}-d\mathbf{I}|=0 \tag{5}$$

where **I** is the identity matrix. This equation produces a p-th degree polynomial in d from which the values $d_1$,..., $d_p$ are obtained.
The characteristic vectors may then be obtained by the solution of the equations:

$$[\mathbf{S}-d\mathbf{I}]\cdot\mathbf{t}_i=0 \tag{6}$$

$$\mathbf{u}_i=\frac{\mathbf{t}_i}{\sqrt{\mathbf{t}_i\mathbf{t}_i}}$$

These characteristic vectors make up the matrix

$$\mathbf{U}=[\mathbf{u}_1\,|\,\mathbf{u}_2\,|\ldots|\,\mathbf{u}_p] \tag{7}$$

which is orthonormal.
The principal axis transformation will transform p correlated variables $x_1$, $x_2$,...,$x_p$ into p new uncorrelated variables $z_1$, $z_2$, ... ,$z_p$. The co-ordinate axes of these new variables are described by the characteristic vectors $\mathbf{u}_i$ which make up the matrix **U** of direction cosines used in the transformation:

$$\mathbf{z}=\mathbf{U}^{\mathrm{T}}\mathbf{x} \tag{8}$$

Here **x** is the $p\times 1$ vector of observations on the original variables.
The i-th principal component is

$$\mathbf{z}_i=\mathbf{u}^{\mathrm{T}}\mathbf{x} \tag{9}$$

### 3.2 Application of PCA to Markovian analysis

PCA can be applied to Markovian analysis in order to minimise the problem of lack of support when more than one characteristic attribute is considered. As we have already underlined, when we would like to better describe the state of an entity by using several attributes, the problem of lack of support may arise. Therefore PCA can be applied to transform the state space of the considered problem into a space where the attributes are uncorrelated; then Markovian analysis can be performed for each transformed attribute separately. Moreover these Markovian analyses, which have been obtained from the uncorrelated attributes resulting from PCA, can be considered approximately independent. After such Markovian analyses have been computed, thanks to their approximate independence, we can gain information about the joined effect of the various PCs on the transition probabilities of the system, simply by multiplying the relative Markov chains.

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X


Let us consider the state space of the observed system at the k-th step;

$$\mathbf{X}^{(k)} = \left\{\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \ldots, \mathbf{x}_w^{(k)}\right\} \tag{10}$$

$$\mathbf{x}_i^{(k)} \in \mathbf{X}^{(k)}, i = 1, \ldots, w \text{ gen.state vector at } k\text{-th step}$$

$$\mathbf{x}_i^{(k)} = \begin{bmatrix} a_{1i}^{(k)} \\ \vdots \\ a_{pi}^{(k)} \end{bmatrix} \tag{11}$$

where $a_{ji}^{(k)}$ is the j-th attribute of the state vector $\mathbf{x}_i$ at k-th step. The covariance matrix for these w observations, that are the feasible states at k-th step, is a $p \times p$ matrix that can be obtained from:

$$\mathbf{S}^{(k)} = \frac{1}{w-1} \sum_{i=1}^{w} \left(\mathbf{x}_i^{(k)} - \mathbf{m}^{(k)}\right)\left(\mathbf{x}_i^{(k)} - \mathbf{m}^{(k)}\right)^T \tag{12}$$

where

$$\mathbf{m}^{(k)} = E\left(\mathbf{x}^{(k)}\right) = \frac{1}{w} \sum_{i=1}^{w} \mathbf{x}_i$$

For each k, k=1,.., K, a $p \times p$ covariance matrix is computed and then the corresponding transformation matrix $\mathbf{U}^{(k)}$ is obtained. Therefore, for each k-step, a state space transformation is performed, which aims at uncorrelating the relative attributes:

$$\mathbf{y}_i^{(k)} = \mathbf{U}^{(k)^T} \mathbf{x}_i^{(k)} \tag{13}$$

with

$$\mathbf{y}_i^{(k)} = \begin{bmatrix} \alpha_{1i}^{(k)} \\ \vdots \\ \alpha_{pi}^{(k)} \end{bmatrix}$$

where $\alpha_{ji}^{(k)} \in A_j^{(k)}$, which is the set of values assumed by the j-th attribute, j=1,..., p. Here $\alpha_{ji}^{(k)}$ are the transformed attributes, which are uncorrelated since the covariance matrix for the transformed set of variables is diagonal:

$$\mathbf{S}_y^{(k)} = \mathbf{U}^{(k)^T} \mathbf{S}^{(k)} \mathbf{U}^{(k)} = \begin{bmatrix} d_1^{(k)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_p^{(k)} \end{bmatrix} \tag{14}$$

Afterwards, we can repeat the Markovian analysis for the transformed state spaces, considering each transformed attribute separately (Figure 2).

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X



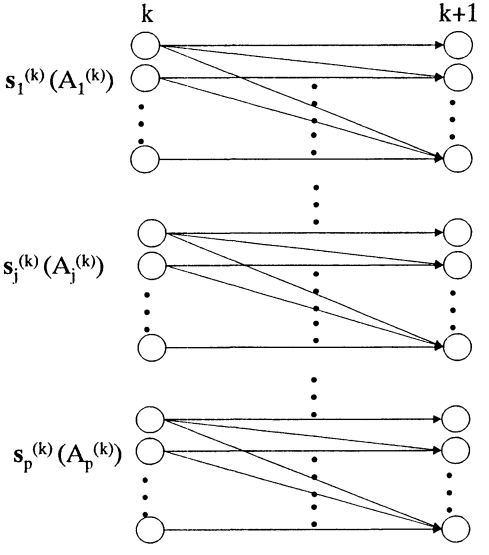



Figure 2: Markov chains computed on uncorrelated attributes

We obtain p Markov chains computed on p uncorrelated attributes. Such chains should result approximately independent and then some information about the joined effect of several attributes should be gained by multiplying the relative chains.

Let us suppose, for example, to need information about the joined effect of the attributes $\alpha_{ji}^{(k)}$ and $\alpha_{hi}^{(k)}$ on the transition probabilities between the aggregated states $\mathbf{z}_i^{(k)}$ and $\mathbf{z}_w^{(k+1)}$, where:

$$\mathbf{z}_i^{(k)} = \left[\alpha_{ji}^{(k)}, \alpha_{hi}^{(k)}\right] \qquad (15)$$

$$\mathbf{z}_w^{(k+1)} = \left[\alpha_{jw}^{(k+1)}, \alpha_{hw}^{(k+1)}\right]$$

Such a probability should be obtained simply by multiplying the transition probabilities corresponding to the relative single-attribute chains:

$$p\left(\mathbf{z}_i^{(k)}, \mathbf{z}_w^{(k+1)}\right) = p\left(\alpha_{ji}^{(k)}, \alpha_{jw}^{(k+1)}\right) \cdot p\left(\alpha_{hi}^{(k)}, \alpha_{hw}^{(k+1)}\right) \qquad (16)$$

Let us suppose, for example, to have a data set consisting of n *2×1* vectors of observations, which are randomly and equally distributed on three possible values (0,1,2). Then the transition probabilities between two contiguous temporal stages and among the nine feasible states, which correspond to any possible combinations of the two 3-value variables, should tend to 1/9. By increasing the number of samples, the values of such probabilities come up to 1/9 more and more. Transition probabilities can be obtained either by computing a single Markov chain on two attributes or using PCA, as described above, to uncorrelate the attributes and then performing MC analysis on the transformed attributes separately. In the second case transition probabilities are obtained by multiplying the probabilities corresponding to the relative single-attribute chains. Figure 3

shows the comparison between the mean absolute error (MAE), computed for the two methods described above, versus the number of samples. Here MAE represents the mean absolute value of the distance between the computed transition probabilities and 1/9. The value of MAE for the method based on PCA converges faster to zero as the number of sample increases. Such a rapid convergence can be taken as a first proof of the efficacy of the method.

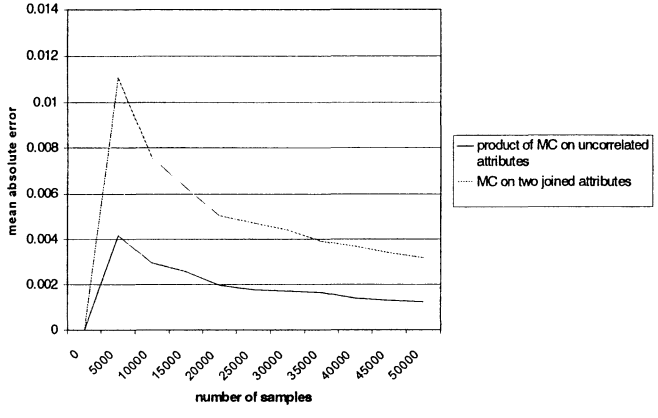



Figure 3: Comparison between MAE in two methods

## 4. Conclusions

This paper studies an approach based on Markov chains in order to define clusters of people with a homogeneous behaviour and to identify individual pattern that represent the behaviour of the single component of the cluster. Such behaviours can be described through Markov chains as a series of transitions characterised by time. The proposed method has been applied to a case study concerning the problem of University dropouts. A further improvement to such a method is presented in order to overcome the problem of lack of support, which emerges when several attributes characterising the state vector need to be considered. The proposed improvement consists in performing the principal components analysis (PCA) to identify the principal components of the observed system and then carry out the Markovian analysis for each component separately. After such Markovian analyses have been computed, thanks to their approximate independence, we can gain information about the joined effect of the various PC on the transition probabilities of the system, simply by multiplying the relative Markov chains. Thanks to this method the state space of the system can be extended including more than a couple of attributes and minimising the problem of lack of support.

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X


## References


[1] Berndt, D.J. & Clifford, J., Finding Patterns in Time Series: A Dynamic Programming Approach (Chapter 9). *Advances in Knowledge Discovery and Data Mining*, ed. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, AAAI Press, 1996

[2] Srikant, R. & Agrawal, R., Mining Sequential Patterns. *Proc. of the 11th Int'l Conference on Data Engineering*, Taipei, Taiwan, March 1995.

[3] Mannila, H., Toivonen, H. & Verkamo, I., Discovery of Frequent Episodes in Sequences. *Proc. of the Int'l Conference on Knowledge Discovery in Database and Data Mining (KDD'95)*, Montreal, Canada, August 1995

[4] Howard R.A. *Dynamic programming and Markov processes*,. John Wiley, 1960.

[5] Diaconis, P. & Stroock, D. Geometric Bounds for eigenvalues of Markov Chains. *Annals of Applied Probability*, 1, 36-61, 1991

[6] Taha, H.A., *Operation Research*, Macmillan Publisher, 400-406, 1971

[7] Arnold, S. F., Gibbs Sampling. *Handbook of Statistics*, 9, 599-626, 1993.

[8] Chib S. & Greenberg E. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*. 49, #4, 329-335, 1995.

[9] Massa, S. & Puliafito P.P. An application of data mining to the problem of the University students' dropout using Markov chains. *Proc. of Third European Conference Pkdd'99*, eds. J.M. Zitkow & J.Rauch, Springer: Prague 15 - 18 September 1999

[10] Harmon, H.H., *Modern Factor Analysis*, University of Chicago Press: Chicago, 1976.

[11] Woodhall, W.H. & Ncube, M.M.. Multivariate cosum quality control procedures. *Technometrics*, 27, 285-292, 1985

[12] Crosier, R.B., Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30, 291-303, 1988.

[13] Brown, S.D. et al. Chemometrics. *Anal Chem.*, 60, 252R-273R, 1988.

[14] Carey, R.N. et al. Principal Components Analysis: an alternative to "referee" methods in method comparison studies. *Anal. Chem.*, 47, 1824-1829, 1975.

[15] Vavra, T.G. Factor analysis of perceptual change. *J. Mark. Res.*., 9, 193-199, 1972.

[16] Jensen, R.E. Scenario probability scaling: an eigenvector analysis of elicited scenario odds ratios. *Futures*, 13, 489-498, 1981.

[17] Chow, G.C. Analysis and control of dynamic economic systems. John Wiley and Sons, Inc., New York, 1975.

[18] Kloek, T., Mennes, L.B.M., Simultaneous equations estimation based on principal components of predetermined variables. *Econometrica*, 28, 45-61, 1960.