An Application of Data Mining to the Problem of the University Students' Dropout Using Markov Chains

S. Massa[1] and P.P. Puliafito[1]

[1]DIST, Department of Communication, Computer and System Sciences,
University of Genoa, via Opera Pia, 13,
16145 Genova – ITALY
{silviam, ppp}@dist.unige.it

Abstract. A new application of data mining to the problem of University dropout is presented. A new modeling technique, based on Markov chains, has been developed to mine information from data about the University students' behavior. The information extracted by means of the proposed technique has been used to deeply understand the dropout problem, to find out the high-risk population and to drive the design of suitable politics to reduce it. To represent the behavior of the students the available data have been modeled as a Markov chain and the associated transition probabilities have been used as a base to extract the aforesaid behavioral patterns. The developed technique is general and can be successfully used to study a large range of decisional problems dealing with data in the form of events or time series. The results of the application of the proposed technique to the students' data will be presented.

1 Introduction

Data mining represents the core activity of the so-called Knowledge Discovery in Databases (KDD) process, which aims at extracting hidden information from large collections of data.

Data mining techniques can be divided into five classes of methods according to their different goals that is the different kind of knowledge they aim to extract [1]. These methods include predictive modeling (i.e. decision trees [2]), clustering [3], data summarization (i.e. association rules [4]), dependency modeling (i.e. causal modeling [5], [6]) and finally change and deviation detection [7].

The work presented in this paper deals with the application of a new predictive modeling technique to the problem of the University students' dropout. A modeling technique based on Markov chains has been developed in order to mine the students' behavior during their University period and to identify the population at risk.

In the dropout problem, the time represents an important attribute characterizing the available information. The available data about the students' University career can be associated with a time-ordered sequence of events, which can represent, for example, the passed examinations. The analysis of such a sequence could then provide knowledge about the behavior of the system that, at least ideally, has generated the data. The mined knowledge in such cases could successively be used to

J.M. Zytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 51-60, 1999.

predict, with a sort of “black box” pattern matching approach, the evolution of the considered system from the observation of its past behavior.

The addressed application can be considered a good reference for a large range of problems which deal with data available in the form of time ordered sequence of events. As a consequence the proposed technique should be intended as general and can be used in different contexts.

The approach that has been studied in this work tries to exploit a model based on the theory of Markov chains in order to provide a statistical representation of the properties of the observed system. The data are mined in order to extract the probability of transition among the possible states in which the system could evolve and to show implicit correlation between the different elements of a student state (the number of passed exams, the average mark, changes of residence and so on) and the decision to give up studying. In our case the goal consists in the extraction of expected patterns from data ending with dropout or degree, through data mining.

The analysis of such patterns leads to identify the set of students who run the risk of dropping-out and therefore to determine high-risk situations in the students’ careers.

The paper is organized 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 modeling technique is explained step by step with reference to dropouts. Then the results of the application of the proposed technique to the data about the students are discussed. Finally the further developments of the work are presented.

2 Markov Chains

The theory of Markov chains ([8], [9], [10], [11], [12]) is often used to describe the system asymptotic behavior by means of relevant simulation algorithms (Gibbs sampling [13], Metropolis, Metropolis-Hastings [14]). The use of Markov chains simplifies the modeling 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 behavior of systems whose evolution can be modeled 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.

Markov chain methods are considered a standard tool in statistical physics, in biological systems simulation and for performing probabilistic inference in statistical applications. Such methods are also successfully employed in expert systems to carry out probabilistic inferences, in the discovery of latent classes from data and in Bayesian learning for neural networks.

2.1 Definitions and Basic Properties

Let $X^{(k)}$ be a set of possible states of a system, at the k-th step or time instant, for any entity of a considered population. If the state of an entity at a generic k-th step can be expressed through a vector of variables, then $X^{(k)}$ can be written as follows:

$$X^{(k)} = \left\{ \underline{x}_1^{(k)}, \underline{x}_2^{(k)}, \dots, \underline{x}_w^{(k)} \right\} \tag{1}$$

where w is the number of possible states for the considered entity at k-th step.

Then, such a system could be modeled through Markov chains only if the probability distribution of a generic $x_{i^{k+1}}^{(k+1)} \in X^{(k+1)}$ depends entirely on the value of the state vector assumed at the k-th step, i.e., $x_{i^k}^{(k)} \in X^{(k)}$.

Formally:

$$p(\underline{x}_j^{(k+1)} \mid \underline{x}_i^{(k)}, \underline{x}_v^{(k-1)}, \dots, \underline{x}_w^{(0)}) = p(\underline{x}_j^{(k+1)} \mid \underline{x}_i^{(k)}) \qquad \forall\, i, v, \dots, w \tag{2}$$

of course, equation (2) is verified for any step k.

To define the Markov chain we need to know the initial probability of a generic state $\underline{x}_j^{(0)}$, $p_{\underline{x}_j}^{(0)}$, $\forall$j and the transition probability for any possible state $\underline{x}_j^{(k+1)}$ to follow the state $\underline{x}_i^{(k)}$ that is denoted by matrix $T_{\underline{x}_i, \underline{x}_j}^{(k)}$.

If the transition probability does not depend on the step k (e.g. for stationary systems), the Markov chain is said homogeneous and the transition probability could be written as $T_{\underline{x}_i, \underline{x}_j}$. Using the transition probabilities, the probability for the state $\underline{x}_j^{(k+1)}$ at time k+1, denoted by $p_{\underline{x}_j}^{(k+1)}$ can be easily computed from the correspondent probabilities at time k as follows:

$$p_{\underline{x}_j}^{(k+1)} = \sum\nolimits_i p_{\underline{x}_i}^{(k)} T_{\underline{x}_i \underline{x}_j}^{(k)} \tag{3}$$

Given the vector of initial probabilities, $\underline{p}^{(0)}$, equation (3) determines the behavior of the chain for all the time instant. The probabilities at step k can be viewed as a row vector, $\underline{p}^{(k)}$, and the transition probabilities at step k as a matrix, $T^{(k)}$, or simply T if the chain is homogeneous. Equation (3) can be expressed as:

$$\underline{p}^{(k+1)} = \underline{p}^{(k)} T^{(k)} \tag{4}$$

For a homogeneous chain, T^k , that is the k-th power of the matrix T, gives the transition probabilities at k step to obtain:

$$\underline{p}^{(k+1)} = \underline{p}^{(0)} T^{(k)} \tag{5}$$

3 Application of Markov Chains to the Mining of Time Series

The class of addressed problems takes the form of time series analysis to extract non-evident behavioral 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 behavior, from medical and diagnostics problems and from environmental contexts.

In the next sections a modeling technique aiming at applying Markov chain theory to data mining problems, which can be modeled with time-series, is presented with particular reference to the analysis of University dropouts.

3.1 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 characterize its behavior.

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:

$$e \leftrightarrow <s_1\ s_2\ s_3 \ldots s_n> \qquad \textbf{(6)}$$

where n = n(e) is the number of events of the series.

Table 1. An example of exam database

Student's number	Exam_data	Exam_ID	Exam_mark
1	20/1/98	10	5
1	20/1/98	12	2
2	21/1/98	2	3
3	22/1/98	22	6
…	…	…	…

Let us consider the administrative database of a University where the various entities are represented by the University students and the events by their passed examinations (Table 1).

The students, who are univocally identified by their matriculation number, are observed for a period of twenty years (from 1978 to 1998). Therefore the data set includes either students who have already left the University or students who are still attending in 1998.

The students' personal data are inserted in a table (Table 2) that reports, for each student, the matriculation date, the date of degree or the date of the first "non-enrollment" that can be considered as the dropout date.

Table 2. An example of the students' personal data

Student's number	Matriculation_date	Degree_date	Dropout_date
1	1/11/88	20/4/94	
2	1/11/89	25/7/95	
…	…	…	…

The examinations passed by the students are considered as the "events" that characterize their curriculum; therefore a student's state is given by a three aggregated variables vector (Table 3): the number of passed exams, the average mark and the student's condition (attending/graduate/dropped-out) as coded in Table 4.

Time here represents the distance (months) from the matriculation date and it is sampled non homogeneously to reflect only specific instants that are particularly significant during an academic year.

Table 3. The state variable values for the students' data

ID_code	Sampling_time	Num_exams	Average_mark	Condition_code
1	5	1	25	A
1	12	2	27	A
…	…	…	…	…
1	72	28	26	D
2	5	2	20	A
…	…	…	…	…
2	24	3	22	G
…	…	…	…	…

Table 4. The condition codes

Description	Condition code
Attending	A
Graduate	G
Dropout	D

The expression (6) for the dropout case study and a given time horizon turns to be:

$f(e_{j,}\, t_i) = \underline{x}$, $j = (1,\ldots,n)$ n number of observed students;
$t_i \in T$, T= <5, 12,…, 180>, time instants vector; **(7)**
$\underline{x} \in X$, X = <number of passed exams, average mark, attending/graduate/drop out>

The students are observed for a time horizon of 15 years, corresponding to 180 months, by which the 95% of them end his University career and the remaining 5% can be disregarded.

The description of the student's state as the combination of the number of passed exams, the average mark and the student's condition produce an excessive number of states and a consequent lowering of the support. The concept of support is crucial in data mining to give a measure of the statistic importance of the 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:

$$\sigma(\underline{x}_i^k) = \frac{n_i^k}{\sum_i n_i^k} \qquad \mathbf{(8)}$$

As the number of states increases, the number of students for each state (n_i^k) decreases and consequently the support of each state drops. Then a discretization step is now needed to reduce the number of states that is here achieved by considering suitable ranges for the average mark values to avoid an excessive state scatter and to maintain a sufficient support level.

Table 5. Classes for the average mark

Average mark	Code
27-30	High
23-26	Medium
18-22	Low

Let k and k+1 be two generic successive stages; then, let N, F and S represent the components of the state, respectively the number of passed exams, the average mark range and the condition code. Each state component obviously has values depending on the stage.

The transition probability can be computed as follows using the Markov chain basic property (2).

Considering a pair of states that are contiguous in time, i.e. which are associated with two successive sampling times, the transition probability can be expressed as:

$$p_{i,j}^{(k,k+1)} = \frac{n_{i,j}^{(k,k+1)}}{n_i^{(k)}} \qquad \sum_j p_{i,j} = 1 \tag{9}$$

The computation of the transition probabilities is performed through (9) and then is summarised in Table 6.

The final result of the computing transition probabilities process 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 one. For each stage there are two absorbing states, one associated with degree (G) and the other with dropout (D).

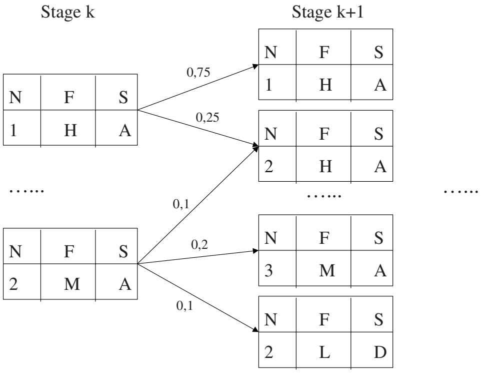
For students still attending in a generic state $\underline{x}_i^{(k)}$ it is possible to calculate the possibility to reach each of the absorbing states, $p_{i,D}^{(k)}$ and $p_{i,G}^{(k)}$. The way to get the two probabilities is similar. Taking for example $p_{i,D}^{(k)}$:

$$p_{i,D}^{(k)} = \sum_{w=k+1}^{h_f} \underline{e}_i^T \cdot \prod_{z=k}^{w-1} T^{(z)} \cdot \underline{e}_D \tag{10}$$

where $\underline{e}_D$ has 1 in correspondence to the absorbing state D and h_f is the final stage of the time horizon considered.

The result of the application of Markov chains can be used, in the present context, to discover the sets of students with different risk degree of dropout, but it can also constitute the basis for further more accurate analysis of individual behavior.

Table 6. The computed transition probabilities



The second goal comes from the analysis of dropout probability for each intermediate state and then from 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 behavior 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 range, can be inferred. In this case the model based on Markov chains could be used to the purposes of planning and control.

Another use of the proposed Markov chain based modeling technique is the possibility to forecast the single student's behavior that is his position in the state space at time k+1, knowing his position in the state space at time k.

This kind of application needs a sufficient number of variables to be included in the state space to provide the appropriate description and, as a consequence, the possibility of having local lack of support generally increases.

In the following sections the first results of the application of the proposed technique to dropouts will be presented together with further improvements of the model, in order to minimize the effects of the above mentioned problem.

4. Results

In this chapter the students' behavior is summarized 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 years' period of time under consideration.

Figure 1 compares two histograms for each year. The black one represents the local dropout probability, that is the probability to dropout over years from the matriculation date, while the white one represents the cumulative dropout probability.

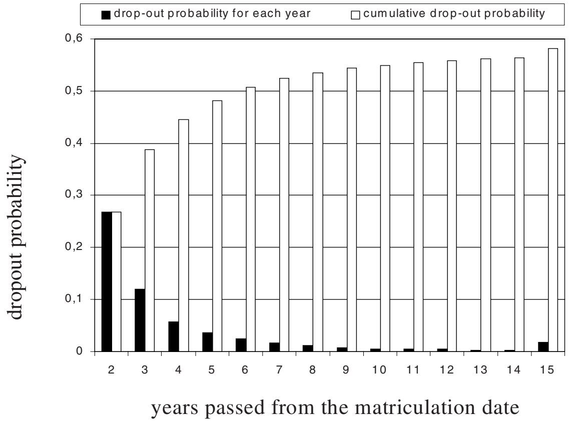


Fig. 1. Dropout probability over the years from the matriculation date

Now a graph, which is based on the probabilities previously computed for each state, is provided. This representation takes into account only the time passed from the matriculation date and the number of passed exams. These two variables define the state to which a dropout probability is associated. A white ball represents a state where the dropout probability is under a fixed threshold while a black ball represents a state where the same threshold is exceeded.

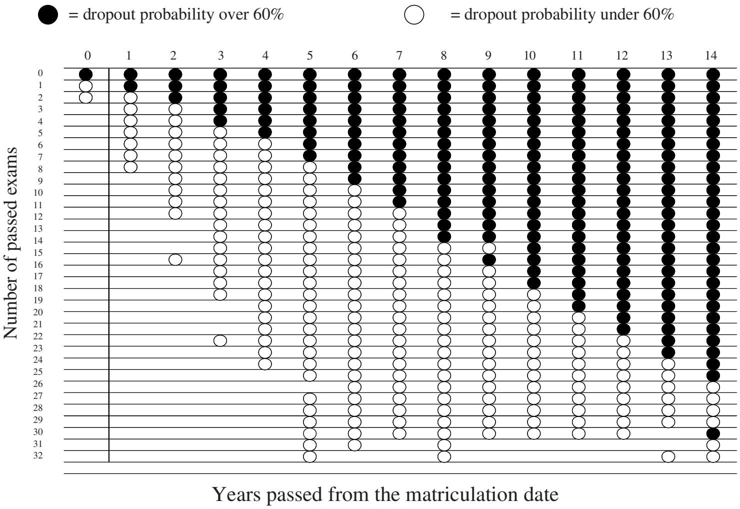


Fig. 2. Dropout risk-zone

Figure 2 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.

Figure 3 represents the dropout probability after 4 years from the matriculation date, with reference to the passed exams, and the relevant shape of the support (see (8)).

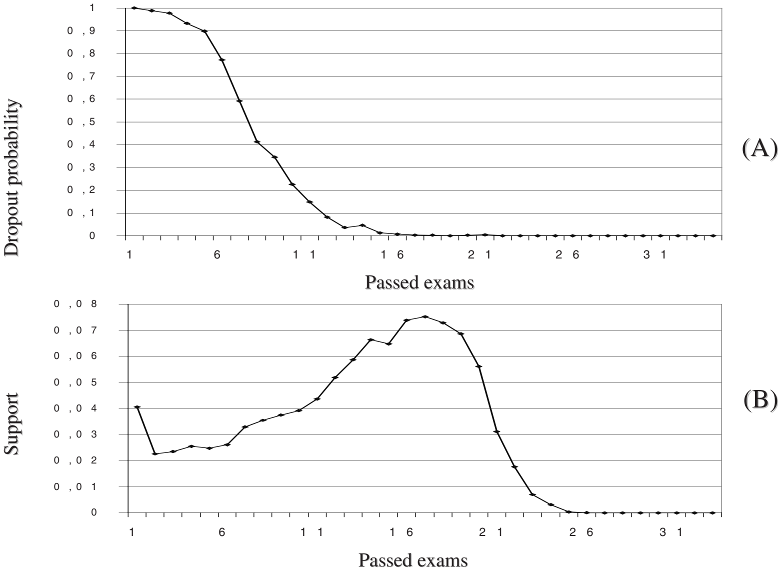


Fig. 3. Dropout probability (A) and support (B) with reference to the passed exams after 4 years from the matriculation date

5. Conclusions

The behavior and the choices of an individual can often be referred to the behavior of the groups of people that statistically represent them. This paper defines an approach based on Markov chains to define clusters of people with a homogeneous behavior and to identify individual pattern that represent the behavior of the single component of the cluster. Such behaviors can be described through Markov chains as a series of transitions characterized by time. The proposed method has been applied to a case study concerning the problem of University dropouts. In such a context the proposed modeling technique can be used in order to define clusters of students associated with different dropout risk degree. Another use of the method concerns the analysis of the individual patterns in order to identify possible policies aimed at lowering the dropout risk levels. Then, in this sense, the proposed method can be used for planning and control activities.

References

1. Geman S., Geman D. (1984) "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.6, pp. 721-741

2. Gelfand A.E., Smith A.F.M. (1990) "Sampling based approaches to calculating marginal densities", Journal of the American Statistical Association, vol.85 pp.398-409

3 Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller E. (1953) "Equation of state calculations by fast computing machines", Journal of Chemical Physics, vol.21 pp.1087-1092

4. Hastings W.K. (1970) "Monte Carlo sampling method using Markov chains and their applications", Biometrika, vol.57 pp. 97-109

5. Agrawal R., Srikant R. (1994) "Fast algorithms for mining association rules in large databases" in Proc. of VLDB Conference, Santiago, Chile.

6. Mannila H., Toivonen H., Verkamo I. (1994) "Efficient Algorithms for discovering association rules". In KDD-94: AAAI Workshop on Knowledge Discovery in Databases.

7. Agrawal R., Srikant R. (1995) "Mining Sequential Patterns". IBM Research Report.

8. Howard R.A. (1960) “Dynamic programming and Markov processes”. John Wiley.

9. Neal, R. (1993) Probabilistic Inference using Markov Chain Monte Carlo Methods. Dept. of Computer Science, University of Toronto.

10. Diaconis, P., Stroock, D. (1991) Geometric Bounds for eigenvalues of Markov Chains. Annals of Applied Probability, 1, 36-61.

11. Hamdj A. Taha (1971) Operation Research, Macmillan Publisher, 400-406.

12. Howard R.A. (1960) Dinamic Programming and Markov Processes, Wiley.

13. Arnold S. F. (1993) Gibbs Sampling. In Handbook of Statistics, 9, 599-626.

14. Chib S., Greenberg E. (1995) Understanding the Metropolis-Hastings Algorithm. The American Statistician. 49, #4, 329-335.