BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LIX No. 1/2007	29 - 34	Seria Matematică - Informatică - Fizică

Data Mining Techniques in Knowledge Management in Academic Environment

Irina Tudor, Mădălina Cărbureanu

Universitatea Petrol-Gaze din Ploieşti, Bd. Bucureşti 39, Ploieşti, Catedra de Informatică
e-mail: tirinelle@yahoo.com

Abstract

During the last years in the academic environment a special attention was placed on the knowledge management in different organizations. Universities represent an environment in which an adequate knowledge management affects their functionality in a positive way. Such management consists both of establishing the good relations between clients (students, other universities and various partners) and institution, and achieving a good knowledge of the client's needs and processing those needs in order to solve them. In this paper an application of data mining techniques is presented so as to highlight the opportunity of using those methods in the academic environment.

Key words: *data mining, simple linear regression, knowledge*

Introduction

Data mining uses a combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge to uncover hidden trends and patterns. These trends and patterns constitute the basis of predictive models that enable analysts to produce new observations from existing data [3, 5].

The paper discusses potential applications of data mining in higher education. The benefits of data mining are its ability to gain deeper understanding of the patterns previously unseen using current available reporting capabilities. Furthermore, prediction from data mining allows the college an opportunity to act before a student drops out or to plan for resource allocation with confidence gained from knowing how many students will transfer or take a particular course.

In this paper an application of data mining techniques is presented to provide information about the graduated number evolution depending on the number of enrolled students.

An Overview of Data Mining Techniques

Several data mining techniques have been reported in literature. They are divided into two classes: traditional techniques (neighbourhoods, clustering, statistics), and next generation techniques (trees, networks and rules).

Decision Trees are analytical tools for developing hierarchical models of behaviour. Compared to other classification models the decision tree model is, like the association rule model, easy to

interpret and it offers a good performance for classification. They are useful when relationships are unknown and when we need to make broad categorical classifications or predictions [4, 10].

K-Nearest Neighbour is one of the traditional methods used in clustering and classification tasks. Analyzing these clusters, either the representative of each cluster is chosen or certain functions are found that divide the attribute space into subspaces only containing elements from one cluster (class). It is important to find a measure of the distance between attributes in the data and then calculate it [1, 7, 10].

Neural networks offer the opportunity to create a model using technology similar to learning patterns of the human brain. Neural net models are useful when large amounts of data need to be modelled and a physical model is not well enough known to use statistical methods [1, 7,10].

Association Rules provide a tool used to look for patterns of coincidence in data. Association rule analysis is useful in discovering patterns of behaviour but does not produce a predictive model [10].

Data Mining in Academic Environment

Data mining is a powerful tool for academic intervention. Through data mining, a university could, for example, predict which students will or will not graduate. The university could use this information to concentrate academic assistance on those students most at risk. In order to understand how and why data mining works, it's important to understand a few fundamental concepts. First, data mining relies on four essential methods [6]: classification, categorization, estimation, and visualization. Classification identifies associations and clusters, and it separates subjects under study. Categorization uses rule induction algorithms to handle categorical outcomes, such as "persist" or "dropout," and "transfer" or "stay." Estimation includes predictive functions or likelihood and deals with continuous outcome variables, such as salary level. Visualization uses interactive graphs to demonstrate mathematically induced rules and scores, and it is far more sophisticated than pie or bar charts. Visualization is used primarily to depict three-dimensional geographic locations of mathematical coordinates.

Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success.

Application Example

The SPSS software is a program package designed for data statistical analysis. Although there is a variety of this kind of programs, like: SAS, GraphPad, MS Excel, the SPSS software has a rigorous structure and is easy to use, even for the beginners.

In order to discover the type of relation between the number of graduated students and the number of enrolled students this paper presents an example of using data mining techniques in higher education, using SPSS software.

The input data were supplied by the report regarding the situation of the graduated student between the years 1990 and 2004, from the official site of the National Institute of Statistics [9].

The simple linear regression represents a procedure in the framework of SPSS software, and it was applied to determine the relation between the number of graduated students during a certain year and the number of enrolled students. The acquired model can also be used to predict the future number of graduate students, knowing the number of enrolled student, e.g. for the year 2004-2005.

The input data are presented in the figure 1.

No.	Period	No.enrolled student	No.graduated student
1	1990-1991	192810	25927
2	1991-1992	215226	29901
3	1992-1993	235669	33366
4	1993-1994	250087	34240
5	1994-1995	255162	47837
6	1995-1996	336141	57360
7	1996-1997	354488	80991
8	1997-1998	360590	67799
9	1998-1999	407720	63622
10	1999-2000	452621	67940
11	2000-2001	533152	76230
12	2001-2002	582221	93467
13	2002-2003	596297	103402
14	2003-2004	620785	110533

Fig. 1. The input data

Two variables were used for this application, *enrollstud* witch represents the number of enrolled students in a certain year and the *graduatedstud* witch represents the number of the graduated students in the same year.

The SPSS database structure is presented in figure 2.

	no	enrollstud	graduatedstud
1	1	192810	25927
2	2	215226	29901
3	3	235669	33366
4	4	250087	34240
5	5	255162	47837
6	6	336141	57360
7	7	354488	80991
8	8	360590	67799
9	9	407720	63622
10	10	452621	67940
11	11	533152	76230
12	12	582221	93467
13	13	596297	103402
14	14	620785	110533

Fig. 2. The SPSS database

The regression analysis is an application of the correlation, used in prediction purposes. The regression analysis finality is the determination of the following coefficients: a (the regression right line origin), b (the regression right line gradient). Using these coefficients we can estimate the future number of graduated students having the current number of enrolled students. To obtain the regression we apply the following procedure: *Statistics-Regression-Linear*. The results obtained are presented in figure 3 and figure 4.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,946[a]	,895	,886	9278,510

a. Predictors: (Constant), the number of enrolled students

b. Dependent Variable: the number of graduated students

Fig. 3. The result of the Statistics-Regression-Linear procedure

The Model Summary table supplies the regression coefficient value (*R*), which has the same values like the correlation coefficient (*r*), 0.94. *R Square* is the determination coefficient of *R*.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2833,766	7051,147		-,402	,695
	the number of enrolled students	,173	,017	,946	10,089	,000

a. Dependent Variable: the number of graduated students

Fig. 4. The result of the Statistics-Regression-Linear procedure (coefficients)

R Square values show that 89% from the graduated student's variation is explained by the enrolled student's variation. *Adjusted R Square* is an *R Square* correction in accordance with the predictors and subjects number.

The Coefficients table contains the *B* coefficients (non-standardized) and the *Beta* coefficient (standardized) which can be used in the prediction equation. If the number of enrolled students has the value 192.810, the model can supply the number of graduated students in the period 1990-1991 using the following relation:

$$graduated\_stud = 0.173 * enroll\_stud - 2833.766 \quad (2)$$

where the value -2833.766 represents the outset and the value 0.173 is the gradient of the regression right line.

As a result of using this procedure the SPSS database was modified by adding new variables presented in figure 5.

	no	enrollstud	graduatedstud	PRE_1	RES_1	LICI_1	UICI_1
1	1	192810	25927	30497,48383	-4570,48383	8373,21202	52621,75564
2	2	215226	29901	34372,55923	-4471,55923	12505,67506	56239,44339
3	3	235669	33366	37906,56023	-4540,56023	16248,86634	59564,25411
4	4	250087	34240	40399,01370	-6159,01370	18873,80063	61924,22677
5	5	255162	47837	41276,33382	6560,66618	19794,73743	62757,93021
6	6	336141	57360	55275,25115	2084,74885	34269,51649	76280,98580
7	7	354488	80991	58446,91462	22544,08538	37489,80921	79404,02002
8	8	360590	67799	59501,77320	8297,22680	38555,90183	80447,64457
9	9	407720	63622	67649,18147	-4027,18147	46706,62567	88591,73726
10	10	452621	67940	75411,26036	-7471,26036	54334,77638	96487,74434
11	11	533152	76230	89332,73150	-13102,73150	67690,38969	110975,07331
12	12	582221	93467	97815,33655	-4348,33655	75634,62548	119996,04763
13	13	596297	103402	100248,66816	3153,33184	77888,19482	122609,14150
14	14	620785	110533	104481,93219	6051,06781	81783,10987	127180,75450

Fig. 5. The SPSS database after procedure applying

The variable *PRE_1*, contains the predicted values generated by the linear regression model. The variable *RES_1*, contains the difference between the real value and the predicted value. The variables *LICI_1* and *UIC1* represent the margins of confidence interval for each value.

The scatterplot image presented in figure 6, highlights the linear dependence between the two variables analyzed, *enrollstud* and *graduatedstud.*

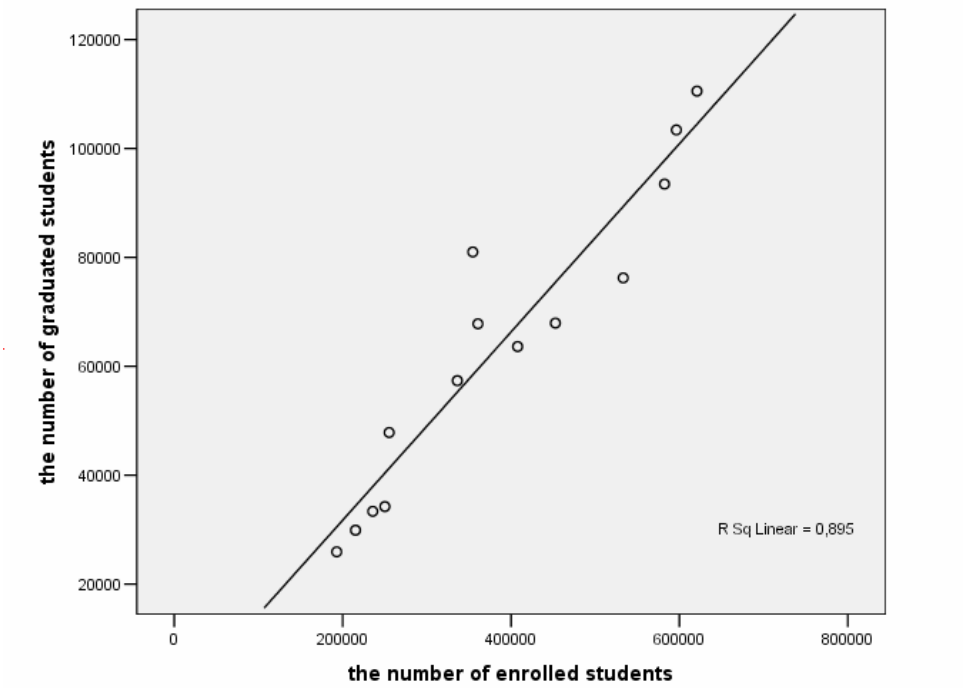

Fig. 6. The regression right line

Knowing that the number of enrolled students in the public education institutions for the academic year 2004-2005 is 650.335, applying the linear regression model previous we estimated the number of the graduated students as:

$$graduated\_stud = 0.173 * 650.335 - 2833.766 = 109.673 \quad (3)$$

In concordance with statistical data supply by National Statistical Institute [5], the number of graduate students was 108.475, showing a minimal error between the calculated value (109.673) and the predicted value (109.325,6).

Conclusion

The application of data mining techniques in knowledge management systems for universities can improve their efficiency. The paper presented an example of data mining use for the prediction of student professional evolution. The tests result indicates the importance of using machine learning algorithms in the educational decision making activity of a university.

References

1. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. - From Data Mining to Knowledge Discovery: An Overview, *Advances in Knowledge Discovery and Data Mining*, eds.

U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, CA: AAAI Press, pp. 1-34, 1996
2. Hand, D., Mannila, H., Smyth, P. - *Principles of Data Mining*, MIT Press, 2001
3. Kidwell, J.J., Vander Linde, K.M., Johnson, S.L. - Knowledge Management Practices in Higher Education, *Educause Quarterly*, 4/2000, pp.28-33, 2000
4. Kittler, R., Wang, W. - The Emerging Role for Data Mining, *Solid State Technology*, 42(11), pp. 45-58, November 1999
5. Luan, J. - Data Mining as Driven by Knowledge Management in Higher Education, *SPSS Public Conference*, UCSF, 2001
6. Luan, J. - Data Mining as Driven by Knowledge Management in Higher Education - Potential Applications, *AIR Forum*, Toronto, Canada, 2002
7. Piatetsky-Shapiro, G. - *Knowledge Discovery in Databases*, AAI/MIT Press, 1991
8. *** - *Correlations and Regression*, http://www.psych.utoronto.ca/courses/c1/spss/page5.htm, accessed 28 May 2007
9. *** - Institutul Național de Statistică, http://www.insse.ro, accesed 25 May 2007
10. Two Crows Corporation - *Introduction to Data Mining and Knowledge Discovery*, Third Edition, tutorial booklet, ISBN 1-892095-02-5, 2006

Tehnici de Data Mining în Managementul Cunoaşterii în Mediul Universitar

Rezumat

În ultima perioadă s-a simțit nevoia accentuată a focalizării atenției asupra cunoaşterii în cadrul diferitelor tipuri de organizații. Universitățile reprezintă acel mediu în care un bun management al cunoaşterii influențează în mod pozitiv funcționarea în ansamblu a acestor instituții. Acest management constă în stabilirea unor bune relații între clienți (studenți, alte universități, diferite firme colaboratoare, etc.) şi instituție, precum şi o mai bună cunoaştere a nevoilor acestor clienți şi încercarea satisfacerii acestor cerințe. În această lucrare este prezentată o aplicație ce subliniază oportunitatea utilizării acestor metode în mediul universitar.