Academic Performance Model Through the Use of Data Mining

Claudio Gutiérrez-Soto, Patricio Oliva, and Angélica Paredes

Information Systems Department, Faculty of Business Sciences
Universidad del Bío-Bío, Concepción, Chile
cogutier@ubiobio.cl, patoliva@alumnos.ubiobio.cl,
anpade@ubiobio.cl

Abstract Data Mining is used in different disciplines for search of patterns and hidden models in databases. It is usually applied in business and marketing areas. This paper presents a Data Mining application in the superior education area. The main contributions of this paper are, at first, a set of standard variables with effect in the academic performance of students, secondly, obtaining predictive model based on Bayesian Network which determines with 96,55% the probability of successful semester for students from Department of Information Systems, Universidad del Bío-Bío, Chile.

Key Word: Data Mining, Survey, Data Base.

1 Introduction

Nowadays, of the superior education institutions generate a large amount of information related with their students. This information corresponds to data used in the administration of superior education organizations, as well as to academic backgrounds and, information of courses among other subjects. This information is relevant in the strategic decision taking of universities. However, there is a large uncertainty about depth of knowledge as well as factors which directly affect the academic performance of students [1]. The formerly mentioned makes complex to take policies intended to improve teaching/learning and so as reduce the student dropout rate. In fact, it's known that 40% of Chilean university students do not complete their studies, which is mainly attributed to lack of maturity, social and economical problems among other factors [2]. Even more, statistical data have been found in Chile and Latin America which demonstrate that careers such as engineering, architecture and laws present a graduation index under or equal to 30%. Careers like medicine, Dentistry, basic level education and special education show a graduation index over 69% [3].

On the other hand, the incorporation of Data Mining in the education area is recent. However, most of the Data Mining applications in the education don't consider all psycho-social variables which affect the student teaching/learning process. Through Data Mining application on Information Systems Department students from Universidad del Bío-Bío, factors affecting teaching/learning process are possible to be obtained.

The main contributions of this paper are: first, to obtain set of standard variables with impact in the academic success and, secondly, a predictive model by using

© A. Gelbukh, M. Adiba. (Eds.)
Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 213-220

Received 25/06/08
Accepted 01/08/08
Final version 10/09/08

Bayesian Networks. Our work is based on CRISP-DM methodology, and SPSS Clementine and Weka are used to obtain predictive model.

This paper is organized as follow. Section 2 makes factorial analysis of surveys to students in order to establish the variables with impact on academic success of students. Section 3 presents predictive model through cluster analysis. Finally, section 4 presents conclusions and future works.

2 Obtain Variables with Impact in Academic Success

Data Mining is a tool applied in different areas, such as, DNA Analysis and biomedical applications [4][5], industrial sale and marketing [4][6], telecommunications [4][7], banking processes [4], financial industry [4] and medicine [8][9], among other areas. Recently, the education area has been favored with the use of Data Mining [1][10][11][12][13][14].

In Data Mining applications of education area, it is possible to find relationships between admission tests and the academic success [15]. In [12] a global classification model is obtained, which assigns the best tutor according to student profiles. In [1] taxonomy of processes where the students are involved and the specific tools supposed to support these processes is presented. On the other hand, in [5] university success predictive tool for students entering to the university is presented. This predictive tool is based on the use of neural networks.

However, varied factors have impact on academic success and desertion rate. In [16] a categorization composed by five significant factors affecting desertion is presented. These factors are classified as psychological, sociological, economical, organizational aspects being interactive between student and the institution. In [17] it is also possible to find four groups of factors that impact desertion rate of Caribbean and Latin America students. These are: external factors to the superior education system and own factors of system and institutions, academic performance and personal problems of students. This categorization includes all the factors with impact in the student desertion rate and their academic success.

Nevertheless, the works exposed in [1][10][11][12][13][14] do not consider all the factors affecting the academic performance and desertion rate of students formerly mentioned in [16] and [17].

One of the main contributions of this paper is to have a number of significant variables (factors), which impact the performance and academic success of Information System Department students of the Universidad del Bío-Bío, Chile. However, to reach this purpose, it was necessary to look for information from the University corporate databases.

2.1 Obtain Non-available Data

Surveys were carried out to obtain information non available in databases. Five dimensional groups of data were established which were later standardized through the analysis of main components (also well-known as factorial analysis). These dimensions included *student data* (sex, age, university entrance year, number of subject matters renounced, number of failed subject matters and others); *Learning and study techniques of students* (data about their individual learning styles, number of hours dedicated to the study, use of the bibliographical material and relevance of group learning and others); socio-economic aspects (the student's socio-economic information), professor and used techniques (information regarding the styles and educational models applied by professors in their classes); and the use of technological tools for learning (information about impact of technological tools in the student's learning).

The sample corresponds to stratified random sample. The population size corresponded to 614 students from Computing Civil Engineering and Information Computing Engineering. The number of students interviewed corresponded to 87. Given the population sample, reliability level used corresponded to 5%, with a 10% of error.

In order to verify feasibility of factorial analysis, Kaiser-Meyer-Olkin (KMO) and Barlett tests were carried out. KMO test obtained value of 0,6145. On the other hand, Barlett test, specifically chi- square, the value obtained was 760,303; in freedom degrees the value obtained was 253 and, for significance the value obtained was 4308E-52. These values assure feasibility to carry out analysis [18].

Factorial analysis was made with SPSS Clementine software Eight components with correlated variables from surveys were found in this factorial analysis. These eight components represent 70% from total variables, being significant percentage. These components are in the Table 1 associated with the variables of the dimensional groups. The components founded are standardized according to the correlation of variables, that is, from highest to lowest. However, components Personal Information, and Works are available the University database. Therefore, these components won't be used to obtain models. However, Personal Information component is fundamental to carry out mapping between existing data in database and non-available data in databases in order to join both data sources.

Tabla 1. Name of the components related with the dimensional groups

Name of component	Dimensional Set
Personal Information	*Student data*
Time devoted to study	*Learning and study techniques of students*
Work	*socio-economic aspects*
Team study or Internet	*the student's learning and their study techniques* *Use of technological learning tools*
To understand the professor's classes (to learn in the classes)	*Professor and used techniques*
Interest in the subject matter	*Learning of Students and their study techniques* *Professors and techniques used*
Attendance	*the student's learning and their study techniques*
Study of guides and from works given by professor	*Professor and used techniques*

Other important item obtained from personal surveys was the definition of academic success. Here, a percentage of 95,34% determined that academic success corresponds to the non failure of subject matters in academic semester. Nevertheless, it is not possible to appreciate the incidence of the eight components in the definition of academic success since components only reflect correlations among variables. In order to standardize the incidence of the eight components on the academic success definition, a second personal survey was carried out. Where preference 1 (the most relevant component to achieve academic success) was to understand the professor's classes (P1), the second preference corresponds to the component *Attendance* (P2), the third preference corresponds to *time devoted to study* (P3), the fourth preference corresponds to *Study of guides (work sheets) and works given by professor* (P4), the fifth preference corresponds to *Interest in subject matter* (P5), and the last preference corresponds to *Study in group or by Internet* (P6).

2.2 Data Joining

One of the most effective classifiers, in the sense that its performance is competitive with state-of-the-art classifier, is so-called Naïve Bayes (NB)[21]. NB supposes that all the attributes are independent known the value of class variable value. Although this supposition is not very realistic the classifier NB is one of the most usedand competitive classifiers. An example of use of this classifier is for the use against the spam or mail garbage [22].

The information existing on database corresponds both to Computing Civil Engineering and Information Computing Engineering students. These data correspond to the years 1998 and 2006. These data correspond to Admit Year, Birth Year, Number of Applications for credit, Number of applications for failed subject matters, Accumulated transcript's average, Accumulated Credits, failed subject matters and Renounced subject matters.

In order to meet data from personal surveys with databases information, we proceeded to select a group of similar data among Personal Information, Admit Year, Birth Year and Failed Subject Matters. This allowed a subset of 180 registrations. Afterwards, 87 registrations were randomly selected to generate predictive model.

3 Obtain Predictive Model

3.1 Bayesian Networks

Classification is one o the basic tasks in the data analysis patterns recognition. Classification requires the construction of classifier, which is a function that assigns a class label to instances described by a set of attributes. The induction of classifiers from data sets of preclassified instances is a central problem in machine learning. Numerous approaches to this problem, which are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs and decision rules [20].

The Bayesian Networks allow the NB performance improvement as well as manage the independency assumptions among variables [20]. Bayesian Networks represent the qualitative knowledge of a model by means acyclic graph. This knowledge is articulated in the dependence/independence definition among the variables that compose the model. Graphic representation for the model specification makes the Bayesian Networks to be a very attractive tool for the knowledge representation, where the representation of knowledge is important aspect of Data Mining [22]. Within the most popular classifiers based on Bayesian Networks, we find the Tree Augmented Naïve Bayes (TAN). TAN is a NB extension classifier, which seeks to maintain the NB computing simplicity but trying to improve the success rate during the classification. So, instead of supposing all the independent variables, certain dependences are admitted among attributes. Therefore it is supposed that attributes constitute a Bayesian Networks with tree form. The advantage of restricting the topology from the net to a tree, is that this structure can memorize easily [20][22]. Another popular classifier is the Bayesian Network Augmented Naïve Bayes (BAN). BAN possesses the same philosophy of the TAN, it proceeds learning a Bayesian Networks for the attributes excluding the class and later on by adding the C class variable and edges from C toward all the attributes is increased.

3.2 Predictive Model

In order to predict if a student will reach academic success, nominal variable disapproved subject matter has been selected. This variable allows us to predict the disapproval of a student per semester. To obtain this predictive model Bayesian Networks has been selected. Bayesian Networks possess several advantages, like the generation of a simple analysis model, even more, is one of the most solid theoretical focuses [19]. On the other hand, Bayesian Networks provide more exact model, that is; more classification tests more robust model in the time.

In order to generate tests for model, cross validation has been selected since the tuplas amount to validate model is not big. However, the amount is significant to total population. On the other hand, to validate model the parameters delivered by weka have been considered, these parameters are Correclty Classified Instantes, Mean Absolute Error, Root mean squared error, Relative absolute error and Root relative squared error.

Algorithm selected for model construction corresponds to the algorithm K2 with two parents. K2 is TAN algorithm extension which allows generation of classifier based on Bayesian Networks. On the other hand, K2 has been selected by the understanding that gives to the model, because TAN generates a classification tree where a node only has a predecessor, while K2 generates a graph, where a node possesses two predecessors.

Table 2 and Table 3 allow a view of results for K2 algorithm with one father, K2 with two parents, and the TAN algorithm. Here, we can observe that the K2 algorithm with a single father possesses higher percentage of classified instances. It also possesses the lowest absolute relative error and the smallest square relative error. Nevertheless, this algorithm has been discarded and we have selected the K2 with two parents since resulting model is more descriptive for analysts. Besides, it must be noticed that in obtained model, nodes correspond to all the variables affecting the academic performance. These are the same variables in Table 1, by adding them the variable Admit Year, Age, Sex, N^{c} of Disapprove Courses Last Semester, Preference 4, Preference 5, and Preference 6. The nodes number reaches the 19 variables.

By checking models, 84 cases were correctly classified from a total of 87 instances, and 3 cases were wrongly classified.

On the other hand, classifications summary with 87 cases is possible to be obtained from confusion matrix. Total of 34 classified instances obtained in this matrix were positive. It means a higher probability to fail in these 34 instances. On the other hand, a total of 50 instances classified obtained were negative, which indicates the probability of not failing.

Tabla 2. Weka Results for the Data Mining Process

Algorithm	Description	Correctly Classified Instances	Mean absolute error
Hill Climber	Maximum Parents 1	97.7011%	0.0547
Hill Climber	Maximum Parents 2	93.1034%	0.0828
K2	Maximum Parents 1	97.7011%	0.0547
K2	Maximum Parents 2	96.5517%	0.0547
TAN		93.1034%	0.0764
K2	Maximum Parents 2, simulate BAN	95.4023%	0.0606

Tabla 3. Weka Results for the Data Mining Process

Algorithm	Root mean squared error	Relative absolute error	Root relative squared error
Hill Climber	0.1485	11.3012 %	30.2492 %
Hill Climber	0.2083	17.1831 %	42.4263 %
K2	0.1363	11.3562 %	27.7728 %
K2	0.1645	11.9203 %	33.5125 %
TAN	0.2103	15.8646 %	42.8381 %
K2	0.1756	12.58 %	35.7676 %

4 Conclusions and Future Work

In this paper, we have presented the development of a Data Mining Application in the Information System Department of the Universidad del Bío-Bío, Chile. Here, we have tried to include most of the variables that have influence on the academic success and in the student desertion rates. For this purpose, we have appealed to non-existing data in the university database through personal surveys. Later on, this information has been joined to the University database data. Joining these data, a group of eight correlated variables denominated components has been obtained. It is important to delimit the existing error margin in this procedure for joining data from personal surveys and databases. However, this procedure is justified since there is no Data warehouse with every data which affect the student performance and the student retention rate

Nevertheless, we think it is a good approach of variables affecting the academic performance of students. Moreover, one of the main contributions of this work is to have a group of eight standardized variables that have impact in academic success of Information System Department Students.

On the other hand, predictive model has been obtained, based on Bayesian Networks which predicts 96,5517% of probability if a student will reprove some subject matter in the semester. Nevertheless, we think it is a good approach of variables affecting the academic performance of students. Moreover, one of the main contributions of this work is to have a group of eight standardized variables that have impact in academic success of Information System Department Students.

On the other hand, predictive model has been obtained, based on Bayesian Networks which predicts 96,5517% of probability if a student will reprove some subject matter in the semester .

References

1. Naeimeh, D.: Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System. In 6th Information Technology Based Higher Education and Training., pp. F4B/1 - F4B/6. (2005)
2. Díaz, J. P.: Los por qué de la deserción universitaria. consultado 08-09-2006. http://www.universia.cl/portada/actualidad/noticia_actualidad.jsp?noticia=58803
3. González, L., Uribe, D.: Estimaciones sobre la repitencia y deserción en la educación superior chilena. Consideraciones sobre sus implicaciones. consultado 21-01-2007. http://www.cse.cl/public/Secciones/seccionpublicaciones/publicaciones_revista_calidad_detalle.aspx?idPublicacion=35
4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Simonv Fraser University, Morgan Kaufmann publishers. Second Edition. San Francisco, United States. Volume 1. pp. 5-15. (2001)
5. Han, J.: How can Data Mining Help Bio-Data Analysis. BIOKDD02 Workshop on data mining in Bioinformatics. Volume 2. pp.1-2. (2002)
6. Edelstein, H.: Building Profitable Customer Relationships with Data Mining. Two Crows Corporation, SPSS white paper-executive briefing. pp. 1-13. (2000)
7. Chang, W., Lee, H. Y.: Telecommunications Data Mining for Target Marketing. Journal of Computers, Volume 12 No. 4. pp.60-74. (2000)

8. Baylis. P.: Better Health Care with Data Mining. Two Crows Corporation, SPSS white paper-executive briefing. pp. 1-9. (1999)
9. Brossette. S., Sprague. E., Hardin. P., Waites. J. M., Jones. K. B., Moser. T.: Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. Journal of the American Medical Informatics Association (JAMIA), Volume 5. pp.373-381. (1998)
10. Luan. J.: Data mining and Knowledge Management, A System Analysis for Establishing a Tiered Knowledge Management Model (TKMM). Proceedings of Air Forum, Volume 5. pp. 373-381. (2001)
11. Gabrilson. S.: Data Mining with CRCT Scores. Office of information technology, Geogia Department of Education. (2003)
12. Waiyamai. K.: Improving Quality of Graduate Students by Data Mining. consultado 11-02-2008. http://www.ku.ac.th/icted2003/document/kritsana.ppt#280,1,Improving quality of graduate students by data mining
13. Luan. J.: Data Mining Application in Higher Education. consultado 21-12-2007. http://www.pse.pt/Documentos/Data%20mining%20in%20higher%20education.pdf
14. Luan, J.: Data Mining and Knowledge Management in Higher Education- Potential Applications. consultado 12-12-2007. http://www.cabrillo.edu/services/pro/oir_reports/DM_KM2002AIR.pdf.
15. Erdoğan, Ş., Timor. M.: A Data Mining Application in a Student DataBase. Journal of Aeronautics and Space Technologies. Volume 2. pp. 53-57. (2005)
16. Braxton, J., Jonson, R.,. Shaw-Sullivan, A.: Appraising Tinto´s theory of college student departure. In Smart. J. C.(Ed) Higher Education Handbook of theory and research, Agathon Press. Volume. 12. pp. 107-164. (1997)
17. Miel, E.: Modelos De análisis de la deserción estudiantil en la educación superior. consultado 21-01-2007. http://www.cse.cl/public/Secciones/seccionpublicaciones/publicaciones_revista_calidad_detalle.aspx?idPublicacion=35
18. ATS, UCLA. "Annotated SPSS output proncipal components análisis". consultado. 22-1-2007. http://www.ats.ucla.edu/stat/SPSS/output/principal_components.htm
19. J, Hernández. C, Ferri. JP, Ramírez. "Introducción a la minería de datos", Pearson Educación S.A . Primera Edición. Madrid, España. ISBN: 84-205-4091-9. pp. 257-278. 2004.
20. Friedman N., Gieger D., Goldszmidt M., "Bayesian Network Classifiers". *Machine Learning* 29. pp.131-163. (1997)
21. Duda, R.O., Hart P. E., "Pattern Classification and Scene Analysis. New York : John Wiley & Sons. (1973).
22. Hernández J, Ferri C. y Ramírez JP (2004). "Introducción a la minería de datos". Pearson Educación, Madrid. (2004).