Advances and Applications in Statistical Sciences
Proceedings of The IV Meeting on Dynamics of Social and Economic Systems
Volume 2, Issue 2, 2010, Pages 275-285
© 2010 Mili Publications

PATTERN DISCOVERY IN UNIVERSITY STUDENTS DESERTION BASED ON DATA MINING

HORACIO KUNA, RAMÓN GARCÍA MARTÍNEZ and FRANCISCO R. VILLATORO

Departamento de Informática
Facultad de Ciencias Exactas Químicas y Naturales
Universidad Nacional de Misiones, Argentina
E-mail: hdkuna@unam.edu.ar

Laboratorio de Sistemas Inteligentes
Facultad de Ingeniería
Universidad de Buenos Aires, Argentina
E-mail: rgarciamar@fi.uba.ar

Departamento de Desarrollo Productivo y Tecnológico
Área de Ingeniería del Software
Licenciatura en Sistemas
Universidad Nacional de Lanús

Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga, Spain
E-mail: villa@lcc.uma.es

Abstract

Students' desertion, the withdrawal from formal studies at the undergraduate level is a global phenomenon in the whole Argentina Public University System. Preventive retention policies may reduce the part of Universities' budget spent every year because of this problem. The identification of the causes of student's retention and non-retention is a due requirement in order to develop successful retention policies. This paper presents preliminary results using the rule based knowledge discovery based on TDIDT (Top Down Induction of Decision Trees) approach on the academic management database. The approach used allowed an interesting analysis to find behavior rules containing incidence variables in retention the results of which could be used to plan preventive retention policies.

2010 Mathematics Subject Classification: 62-07, 62H30, 62P25, 68U35.

Keywords: data mining, droping out, top-down induction of decision trees.

Received April 30, 2009

1. Introduction

Students' withdrawal is an important problem in the Public University System in Argentina and in other countries in Latin America, as illustrated in Figure 1 [9]. It may be defined when a student gives up his formal studies ofundergraduate studies, [11]. Official figures for Argentina for the year 2006 show that there is a population of about 1.3 million of university students, with0.27 million entering the system and 0.067 million graduates leaving the systemeach year. The University Policies Secretariat (SPU) from the Ministry of Education of Argentina has a deep concern on university students' withdrawals. In 2008, an International Seminar on University Students Withdrawal Problem was organized [21]. The SIU Consortium of Argentine Universities (which congregates 33 Universities) is developing different kinds of studies related to students' withdrawal completing the classical statistical approach with a data mining one.

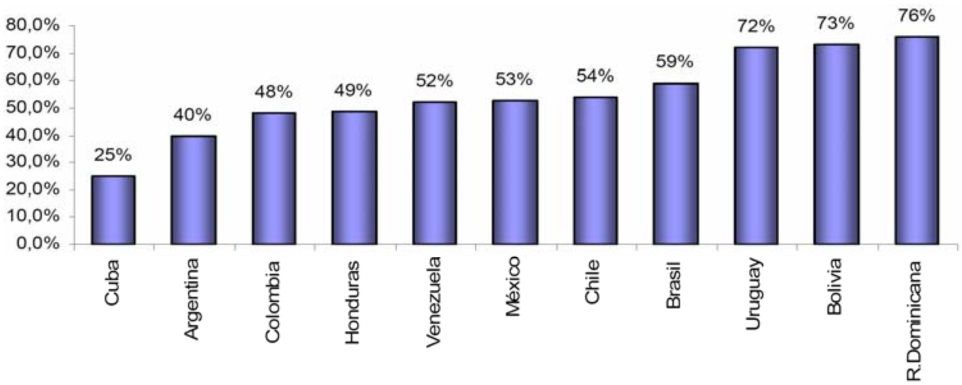

Figure 1. Students' withdrawal in first undergraduate years.

The causes of student retention and non-retention have been studied by several researchers in the past, but their identification is difficult and strongly depending on the specific University System considered, so no consensus about the main reasons have been obtained yet. Spady [20], Tinto [22] and Hackman [8] argue that the proper social integration of the student into academic life and their difficulty to comply with the objectives imposed by university system strongly affects the student's decision of withdrawing or continuing his studies. Bean [2] proposes that withdrawing a degree course strictly depends on factors external to the university. Braxton [4] defines a large set of variables depending on the subject, the subject's family and the subject's social class, and develops a multidimensional analysis based on psychological, economical, sociological, organizational and social interactions. Aparicio and Garzuzi [1] establish vocational processes failures as withdrawal causes.

An important factor to be considered in the student's withdrawal is the behavior pattern followed by students after withdrawing his career. The student may follow one of the following three behaviors: (1) to continue their education either at their original institution but change their degree course, or ask to transfer to another one, (2) to temporaly leave the university system, returning years later, and, finally, (3) to leave permanently university education. Some authors [10], argue that passing through university (although not finishing a degree) may results in the student's personal growth, acquisition of useful knowledge applicable to everyday life among others things. Other authors [21] claim that the Public University System budget is best invested if the non-retention of the students is reduced.

The research presented in this paper was motivated by the high level of dropouts in first and second year of engineering bachelor degree at University of Misiones (Universidad Nacional de Misiones). The Province of Misiones has some particular characteristics: the 90% of its frontiers is international (Brazil and Paraguay); it is located in the heart of Mercosur; its economy is based on regional agriculture products as tea, yerba mate, rain forest woods and tourism (Iguazu Falls); it has a population of 1 million inhabitants; the Public University has 20 thousand students, with 3.6 thousand students entering each year in the university and 600 graduating in the same period.

Intelligent data mining is one of the statistical techniques proposed by the SIU Consortium in order to discover students' withdrawal causes. Data mining is the set of techniques and tools applied to the non-trivial process of extracting and presenting/displaying implicit knowledge, previously unknown, potentially useful and comprehensible, from large data sets, with the objective of predicting automated tendencies and behaviors; and to describe automated form models previously unknown. The term intelligent data mining is the application of automatic learning methods to discover and enumerate present patterns in the data. Currently, data mining techniques have to deal with very large databases facing efficiency and scalability problems. The main advantage of data mining with respect to traditional statistical analysis is the lack of a priori hypotheses on the data to be validated by the analysis. The patterns and the laws behind the data are (semi) automatically extracted from them.

Data mining techniques can be classified into two categories: descriptive and predictive. The most common techniques in descriptive data mining are decision trees (TDIDT), production rules, and self organized maps. Inductive learning is the most common technique for predictive data mining, allowing the development of a model, for example, a Bayesian network, representing the knowledge domain which is accessible to the user. The model allows the determination of the main data dependencies between the variables of the problem and the prediction of the behavior of some unknown variables.

In this paper we present some results on knowledge discovery obtained by induction related to university students desertion in first and second year of engineering bachelor degree. The rest of the paper is as follows: formal aspects of knowledge discovery by induction using decision trees and rule identification and moderation are shown in Section 2, problem variables under consideration in Section 3, some results and their interpretation in Section 4 and conclusions and future research in Section 5.

2. Rule Based Knowledge Discovery Based on TDIDT

Carbonell et al. [7] identify three principal dimensions along which machine learning systems can be classified: the underlying learning strategies used, the representation of knowledge acquired by the system, and the application domain of the system. The product of learning is a piece of procedural knowledge.

Machine Learning based data mining has been addressed as an effective way of discovering new knowledge from data sets of educational processes, data generated by learning systems or experiments, as well as how discovered information can be used to improve adaptation and personalization [3]. Among interesting problems data mining can help to solve: determining which are common learning styles or strategies, predicting the knowledge and interests of a user based on past behavior, partitioning a heterogeneous group of users into homogeneous clusters or detecting misconceptions in learning processes [5, 6, 19].

One of the most common techniques of data mining are the TDIDT algorithms used for discovering knowledge in rule format which constitutes a model that represents the knowledge domain subjacent to the available examples of it. The members of TDIDT family [13] are sharply characterized by their *representation of acquired knowledge* as decision trees. This is a relatively simple knowledge formalism that lacks the expressive power of semantic networks or other first-order representations. As a consequence of this simplicity, the learning methodologies used in the TDIDT family are considerably less complex than those employed in systems which can express the results of their learning in a more powerful language. Nevertheless, it is still possible to generate knowledge in the form of decision trees which is capable of solving difficult problems of practical significance.

The underlying strategy is non-incremental learning from examples. The systems are presented with a set of cases relevant to a classification task and develop a decision tree from the top down, guided by frequency information in the examples but not by the particular order in which the examples are given. The example objects from which a classification rule is developed are known only through the values of a set of properties or attributes, and the decision trees in turn are expressed in terms of these same attributes. The examples themselves can be assembled into two ways. They might come from an existing database that forms a history of observation.

The basis of the induction task [12, 13] is a universe of objects that are described in terms of a collection of attributes. Each attribute measures some important feature of an object and will be limited here to taking a (usually small) set of discrete, mutually exclusive values, each object in the universe belongs to one of a set of mutually exclusive classes. The induction task is to develop a *classification rule* which can determine the class of any object from its values of the attributes [14, 15]. The immediate question is whether or not the attributes provide sufficient information to do this. In particular, if the training set contains two objects which have identical values for each attribute and yet belong to different classes, it is clearly impossible to differentiate between these objects with reference only to the given attributes. In such a case attributes will be termed *inadequate* for the training set and hence for the induction task.

As mentioned above, a classification rule will be expressed as a decision tree [16, 17, 18]. Leaves of a decision tree are class names, other nodes represent attribute-based tests with a branch for each possible outcome. In order to classify an object, it starts at the root of the tree, evaluates the test, and takes the branch appropriate to the outcome. The process continues until a leaf is encountered, at which time the object is asserted to belong to the class named by the leaf. Only a subset of the attributes may be encountered on a particular path from the root of the decision tree to a leaf; in this case, only the outlook attribute is tested before determining the class. If the attributes are adequate, it is always possible to construct a decision tree that correctly classifies each object in the training set, and usually there are many such correct decision trees. The essence of induction is to move beyond the training set, to construct a decision tree which correctly classifies not only objects from the training set but other (unseen) objects as well. In order to do this, the decision tree must capture some meaningful relationship between an object's class and its values of the attributes. Given a choice between two decision trees, each of which is correct over the training set, it seems sensible to prefer the simpler one on the grounds that it is more likely to capture structure inherent in the problem. The simpler tree would therefore be expected to classify correctly more objects outside the training set.

3. Problem, Variables, Knowledge Discovery Process

The problem where knowledge discovery process focused was to identify the relevant causes for student retention and student non-retention, the universe of analysis being first and second year student dropouts at the School of Engineering. A data query is applied to the database of the system of academic administration SIU Guaraní and a view with the identified class attribute and related ones is built.

Selecting those variables which may be used as earlier indicators for potential withdrawal of students requires the identification of its causes. In this paper, knowledge discovery techniques using TDIDT are used. The variables considered in the project are shown in Table 1.

Table 1. Variable description.

VARIABLE NAME	TYPE OF VARIABLE	DESCRIPTION	POSSIBLE VALUES
studies funding	attribute	way in which students studies are funded	by family or others by student own work by student own work and family
number of subjects not passed	attribute	number of subjects with practices not passed or incomplete in first year	$<=1$ >1
number of years between school and university	attribute	number of years between end of high school and university entrance	$<=3$ years $>=3$ and $<=7$ years $>=8$ and $<=15$ years $>=16$ years
number of subjects not approved	attribute	number of subjects with final examinations failed or no show	$<=3$ >3
first year subjects with students as full time	attribute	number of subjects attended as full time student	1 to n
high school orientation	attribute	major area orientation of high school studies	Humanities Commerce Technology Construction Others
subsequent studies	class	subjects passed in second year	true false
commuter	attribute	student who lives more than 10 kilometers away from the university campus	yes no

The scheme of knowledge discovery process is presented in Figure 2. A dataset query which contains attributes related to involved variables (shown in Table 1) is obtained from SIU-Academic Management Database. Once the class attribute is identified in the dataset query, the TDIDT based knowledge discovery process is applied and a rules set is obtained. Then each rule from this is discussed with and interpreted by the domain experts.

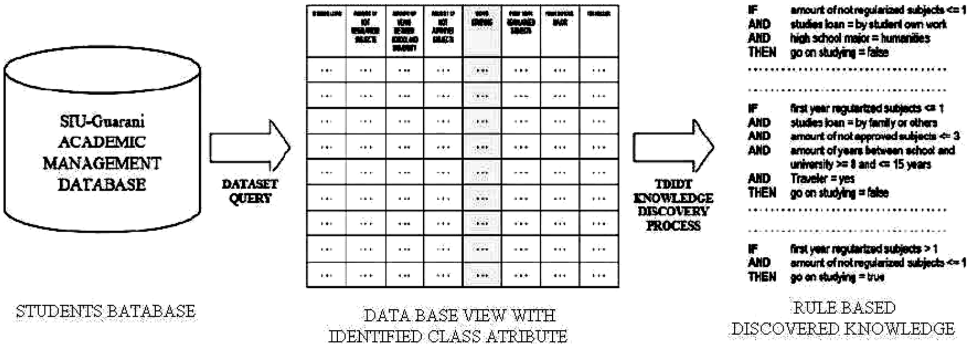
SIU-Guarani
ACADEMIC
MANAGEMENT
DATABASE
DATASET
QUERY
TDIDT
KNOWLEDGE
DISCOVERY
PROCESS
IF amount of not regularized subjects <= 1
AND studies loan = by student own work
AND high school major = humanities
THEN go on studying = false
IF first year regularized subjects <= 1
AND studies loan = by family or others
AND amount of not approved subjects <= 3
AND amount of years between school and university >= 8 and <= 15 years
AND Traveler = yes
THEN go on studying = false
IF first year regularized subjects > 1
AND amount of not regularized subjects <= 1
THEN go on studying = true
STUDENTS BATABASE
DATA BASE VIEW WITH
IDENTIFIED CLASS ATRIBUTE
RULE BASED
DISCOVERED KNOWLEDGE

Figure 2. Scheme of knowledge discovery process.

4. Presentation of Results

Multiple and significant results were obtained and are presented in this section in IF-THEN rule format. We also use a percentage to show the amount of the Guarani data base records that support each condition involved in the rules.

The rule shown in Table 2 casts light on the central cause of students' withdrawal, that is, students that passed one or no subjects during first year, students who afford their own studies and students who attended Humanities at high school do not go on studying in second year.

The attributes "*number of subjects not approved*", "*number of years between school and university*" and "*commuter*" appear together with the attributes with highest relevance in students', withdrawal shown in rule of Table 2.

Table 2. Rule of students who goes on studying.

RULE		SUPPORT
IF	number of subjects not passed < = 1	62, 38%
AND	studies funding = by student own work	58, 82%
AND	high school orientation = Humanities	83, 33%
THEN	subsequent studies = false	28, 59%

The rule shown in Table 3 brings others knowledge of the characteristics of students who do not go on studying.

The rule shown in Table 4 shows the characteristics of students who go on studying in second year of the degree course, establishing that this kind of students have more than one first year subjects attended as full time and equal or less than one first year subject not attended as full time. This rule has a support of 71, 41% of the cases under study.

Table 3. Rule of students who do not go on studying.

RULE		**SUPPORT**
IF	first year subjects attended as full time < = 1	62, 38%
AND	studies funding = by family or others	31, 01%
AND	number of subjects not approved < = 3	35, 14%
AND	number of years between school and university > = 8 and < = 15 years	44, 44%
AND	commuter = yes	100%
THEN	subsequent studies = false	28, 59%

Table 4. Rule of students that go on studying.

RULE		**SUPPORT**
IF	first year subjects with students as full time > 1	88, 94%
AND	number of subjects not passed < = 1	96, 47%
THEN	subsequent studies = true	71, 41%

5. Conclusions and Future Research

The approach addressed in this paper, based on using rule based knowledge discovery based on TDIDT approach on the SIU-Guarani academic management database, allowed an interesting analysis finding behavior rules containing incidence variables in withdrawal as: who funds students university studies, the number of years from the end of secondary school to university entrance. This rule has called the *attention of educational experts* focusing their attention in first year subjects attended as full time students and orientation of high school degree of students.

A risk group of students who do not have academic activity in second year is characterized by those who attended no more than one subject in first year as full time students, have not passed less than three subjects, whose studies are afforded by family or others, a significant time elapsed between school and university (more than eight and less than fifteen) and live far away from the university campus. A common denominator of students dropping out in second year is "number of not subjects not attended as full time students", becoming an earlier indicator of potential withdrawal in students on whom academic authorities have to focus to decrease students´ withdrawal. As a counterpart, students who attend more than one subject as full time students, go on with their studies; this imples that continuing studying depends strongly on this variable.

During data preprocessing, different problems related to data quality emerged. One of future research work proposals focuses on improving the knowledge discovery process addressing the data quality problem in social contents data bases. Here again, data mining appears as a way to find auditory clues related to missing, inconsistent and noisy data (so typical in that sort of data bases).

References

[1] M. Aparicio and V. Garzuzi, Dinámicas identitarias, procesos vocacionales y su relación con el abandono de los estudios, un análisis en alumnos ingresantes a la universidad, Revista de Orientación Educacional 20 (2006), 15-36.

[2] J. Bean, Student attrition, intentions and confidence, In: Research in Higher Education 17 (1980), 291-320.

[3] J. Beck, T. Calders, M. Pechenizkiy and S. Viola, Workshop on educational data mining, ICALT'05 (2007), 933-934.

[4] J. Braxton, Reworking the Student Departure Puzzle, Vanderbilt University Press, 2000.

[5] P. Britos, Z. Cataldi, E. Sierra and R. García-Martínez, Pedagogical protocols selection automatic assistance, Lecture Notes on Artificial Intelligence 5027 (2008a), 331-336.

[6] P. Britos, E. Jiménez Rey and E. García-Martínez, Work in Progress: Programming Misunderstandings Discovering Process Based on Intelligent Data Mining Tools, Proceedings 38th ASEE/IEEE Frontiers in Education Conference, Session F4H: Assessing and Understanding Student Learning, ISBN 978-1-4244-1970-8, (2008b).

[7] J. Carbonell, R. Michalski and T. Mitchell, An Overview of Machine Learning, In R. Michalski, J. Carbonell and T. Mitchell, (Editors), Machine Learning, An Artificial Intelligence Approach, Tioga Publishing Company, 1983.

[8] J. Hackman and W. Dysinger, Commitment to college as a factor in student attrition, Sociology of Education 43(3) (1970), 311-324.

[9] IESALC-UNESCO, Datos Para Colombia: SNIES, Ministerio de Educación Nacional, 2005.

[10] C. Mansky, Anatomy of the selection problem, Journal of Human Resources 24 (1989), 343-360.

[11] M. Parrino, De la Reflexión a la Acción Política Para Disminuir los Procesos de Deserción Universitaria, IV Coloquio Internacional sobre Gestión Universitaria en America del Sud. Florianopolis, 2004.

[12] J. Quinlan, Discovering Rules by Induction from Large Collections of Examples, In Expert Systems in the Micro Electronic Age, D. Michie, (Ed.), Edinburgh University Press, 1979.

[13] R. Quinlan, Induction of decision trees, Machine Learning 1 (1986), 81-106.

[14] J. Quinlan, Learning logic definitions from relations, Machine Learning 5 (1990), 239-266.

[15] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[16] J. Quinlan, Improved use of continuous attributes in C4.5, Journal of Artificial Intelligence Research 4 (1996a), 77-90.

[17] J. Quinlan, Learning decision tree classifiers, ACM Computing Surveys 28(1) (1996b), 71-72.

[18] J. R. Quinlan, Simplifying decision trees, Int. J. Man-Machine Studies 51(2) (1999), 497-510.

[19] C. Schulte and J. Bennedsen, What do teachers teach in introductory programming?, ICERW'06, (2006), 17-28.

[20] W. Spady, Dropouts from higher education: An interdisciplinary review and synthesis, Interchange 1 (1970), 64-85.

[21] SPU, Memorias del Seminario Internacional sobre la Deserción, Secretaría de Políticas Universitarias Ministerio de Educación de la República Argentina, 2008.

[22] P. Tinto, Dropout from higher education: A theoretical synthesis of recent research, Review of Educational Research 45 (1975), 89-125.