Determining Students' Academic Failure Profile Founded on Data Mining Methods

Vasile Paul Bresfelean, Mihaela Bresfelean, Nicolae Ghisoiu
Babes-Bolyai University, Faculty of Economics and Business Administration,
Cluj-Napoca, Romania
bresfelean@yahoo.com, miha1580@yahoo.com, nghisoiu@personal.ro

Calin-Adrian Comes
Petru Maior University,
Targu-Mures, Romania
calin.comes@ea.upm.ro

Abstract. *Exams failure among university students has long fed a large number of debates, many education experts seeking to comprehend and explicate it, and many statisticians have tried to predict it. Understanding, predicting and preventing the academic failure are complex and continuous processes anchored in past and present information collected from scholastic situations and students' surveys, but also on scientific research based on data mining technologies.*

In the current article the authors illustrate their experiments in the educational area, based on classification learning and data clustering techniques, made in order to draw up the students' profile for exam failure/success.

Keywords. Data mining, classification learning, clustering, FarthestFirst, J48.

1. Introduction

The ways in which information and knowledge are represented and delivered to the university managers are in a continuous transformation due to the involvement of the information and communication technologies in all the higher education processes. The Bologna Declaration imposed a motivating process of change for a large and diversified number of countries to work together in order facilitate the quality assurance in the creation of a European Higher Education Area. Therefore, an integration of the latest research results in education managerial issues, in terms of their contents and impact, is a matter that should be essential, whilst taking into account the fundamental task of a university as a facilitator of teaching and research.

Our work reveals the research on educational issues based on data mining processes, made in order to provide the higher education managers valuable knowledge on understanding, predicting and preventing the academic failure for decision taking progression. In this article we applied classification learning through and data clustering methods founded on Weka's J48 and FarthestFirst algorithms tested and utilized in order to match the purpose of determining students' academic failure/success profile.

2. General issues on data mining in educational areas

Data mining is an innovative field of research and study which is being implemented in education with several promising areas for data mining suggested and partially put into practice in the academic world.

Higher education institutions have been interested in predicting the paths of students and alumni [12], thus identifying which students will join particular course programs [10], and which students will require assistance in order to graduate. Another important preoccupation is the academic failure among students which has long fuelled a large number of debates. Researchers [22] attempted to classify students into different clusters with dissimilar risks in exam failure, but also to detect with realistic accuracy what and how much the students know, in order to deduce specific learning gaps [16]. This can be attained throughout ongoing learning assessment processes that indicate which subject the student is better suited to study at that moment, and

necessitates automatic or semi-automatic procedures for treatment and analysis for acquisition of new knowledge.

The distance and on-line education, together with the intelligent tutoring systems and their capability to register its exchanges with students [14] present various feasible information sources for the data mining processes. Studies based on collecting and interpreting the information from several courses could possibly assist teachers and students in the web-based learning setting [15]. Researchers [1] derived models for classifying chat messages using data mining techniques, in order to offer learners real-time adaptive feedback which could result in the improvement of learning environments. In scientific literature there are some studies which seek to classify students in order to predict their final grade based on features extracted from logged data in educational web-based systems [13]. A combination of multiple classifiers led to a significant improvement in classification performance through weighting the feature vectors.

3. Classification learning and data clustering

Data clustering is a key process in data mining, often applied when no information is accessible regarding the bond of data items with predefined classes, and is commonly viewed as part of unsupervised learning [8]. It can be described as the process of organizing objects in a database into clusters/groups such that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity [19]. Data clustering has been used for applications in life sciences and over the years has been used in many areas [23] from the analysis of clinical information, phylogeny, genomics, and proteomics etc.

Classification is a well-known and widely used data analysis method [11] that can automatically learn models or rules describing categories of data. Given a set of training data assigned class labels, the learning system first partitions data into two sets [11]: training and testing. In the training phase, obtained models are used to predict class labels of testing data to verify the learning quality. It is frequently applied to obtain knowledge from databases to make business decisions.

Classification learning is sometimes called supervised because, in a sense, the method operates under supervision by being provided with the actual outcome for each of the training examples [24].

3.1. FarthestFirst clustering algorithm

FarthestFirst provides the "farthest first traversal algorithm" by Hochbaum and Shmoys, which works as a fast simple approximate clusterer modelled after simple k-means.

Clustering techniques are classically divided into two broad categories: hierarchical and partitional algorithms [18]. Other authors use a different approach to classification [9]: Partition-based; Density-based algorithms; Grid-based algorithms; Model-based algorithms; Fuzzy algorithms.

K-means clustering algorithm and its successors [7] have been a very popular technique for partitioning large data sets with numerical attributes. Originally devised as an online clustering technique, most people refer to it as a batch algorithm which provides a "hard" partition of the data as opposed to its "fuzzy" counterpart called fuzzy C-means [18].

The general algorithm was introduced by Cox in 1957, and it was first named k-means by Ball and Hall, and MacQueen in 1967, and since then it has become widely popular and is classified as a partitional or non-hierarchical clustering method [19]. The original online algorithm [2] as presented by MacQueen, is as follows:

```
Let
k be the predefined number of centroids
n be the umber of training patterns
X be the set of training patters x1,
x2,...xn
P be the set of k initial centroids m1,
m2,...mk taken from X
η be the learning rate initialized to a
value in ]0,1[
1 Repeat
2   For i=1 to n
3       Find centroid mjϵP that is closer
to xi
4       Update mj by adding to it Δmj= η(xi
- mj)
5   Decrease η
6   Until η reaches 0
```

The FarthestFirst algorithm starts by randomly selecting an instance to be a cluster centroid [24]. The distance between each remaining instance and its nearest centroid is computed. The instance that is farthest away

from its closed centroid is selected as a cluster centroid. This process is repeated until the number of clusters is greater than a specified threshold. In the FarthestFirst in order to find k cluster centers, is required to follow the steps:

1. randomly choose one point as the first center
2. for i = 2 up to k

next center = point with maximal min-distance to current centers.

3.2. Weka J48 class

In Weka, the implementation of a particular learning algorithm is encapsulated in a class. For example, the J48 class builds a C4.5 decision tree [24]. Each time the Java virtual machine executes J48, it creates an instance of this class by allocating memory for building and storing a decision tree classifier. The algorithm, the classifier it builds, and a procedure for outputting the classifier are all part of that instantiation of the J48 class [24].

The J48 class does not actually contain any code for building a decision tree [24]. It includes references to instances of further classes that do a large amount of the work. When there are a lot of classes—as in Weka—they become difficult to comprehend and navigate. Java allows classes to be organized into packages. Packages are organized in a hierarchy that corresponds to the directory hierarchy: a tree is a subpackage of the classifiers package, which is itself a subpackage of the overall Weka package [24].

C4.5 is an algorithm developed by Ross Quinlan used to create decision trees, an expansion of his prior ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. It constructs decision trees of arbitrary depth in a top-down recursive divide-and-conquer strategy with splits maximizing the Gain Ratio [17]. It is biased, however, in favor of continuous attributes, a weakness partly addressed by later improvements [17],[20]. C4.5 employs a pruning technique that replaces subtrees with leaves, thus reducing overfitting, and in numerous of datasets the accuracy achieved by C4.5 was comparatively high [17].

4. Students' exam failure/success profile

The understanding, prediction and prevention of the academic failure among students have long been a preoccupation for each higher education institution. Building a certain profile for the students, predicting their choices and also their grouping [3]-[6], in this case based on exam failure risk, are important elements of our research and a motivating approach which could help both institution and students. Universities could learn students content/discontent regarding its education processes, curricula, courses, endowment, and figure out specific learning gaps and which students might require assistance in order to graduate. The student, which is the main focus of a student-based education, could benefit from the institution's know-how and support.

In the present paper we try to build up a student's exams failure profile, based on information extracted from on-line surveys filled out by the students of the Faculty of Economics and Business Administration Cluj-Napoca, and also from our faculty's databases, such as the students' scholastic situation database. The information collected consists of: general data on the subject (gender, high school etc.), scholastic situation, types of gained scholarships, interruption of study, exams absence, tuition, and students' opinions (on courses, materials, curricula, research, teachers, laboratories technical novelty, knowledge gained, continuing education) etc.

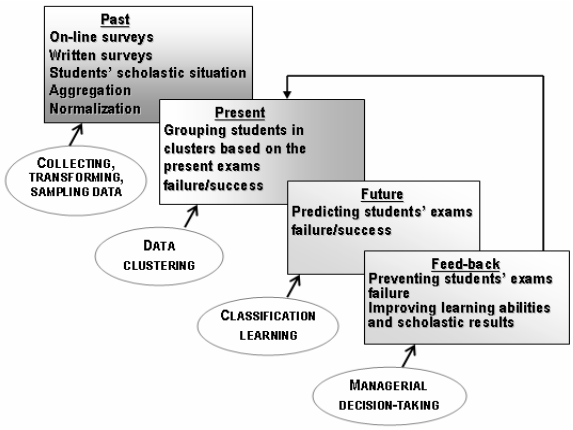

Figure 1. Our research pursue

For the application of data mining processes to improve managerial decisions, seen as a Past-Present-Future model (Figure 1) we used the University of Waikato's Weka software. The initial step was to modify the initial format of the raw data in Excel spreadsheets, extract relevant data from the table fields using search and replace methods. We codified the attributes using only one or two linked words, which were suggestive to each questionnaire item,

eliminating blank spaces between words, special characters and symbols, diacritics. Then, we substituted the numerical instances of some attributes with suggestive character values, through one word or analytic abbreviations. We also eliminated a series of attributes where the answer was very difficult to be codified, for the students had the freedom of certain written answers, and not chosen, resulting in a variety of non-homogeneous instances.

The data mining processes were applied after an attribute selection based on Gain Ratio feature evaluator through the Ranker method. The experiments were conducted over data gathered from hundreds of students, and for the present article we present the investigation based on the CIG (accounting specialization) senior students, with 50 instances. We essentially used classification learning and data clustering methods founded on algorithms tested in order to suit best the purposes of the research.

4.1. The data clustering process

For the clustering process we utilized the FarthestFirst method based on K-means algorithm. We specified the parameter k, the number of clusters to be sought. For this theme the k parameter was 2, corresponding to the two groups of students we were interested in building the exam failure profile: the ones who passed all exams and the ones who failed one or more exams. Then k points were chosen at random as cluster centers. All instances were assigned to their closest cluster center according to the ordinary Euclidean distance metric. Next the centroid of the instances in each cluster was calculated, and these centroids were taken to be new center values for their respective clusters. Finally, the whole process was repeated with the new cluster centers. Iteration continued until the same points were assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and would remain the same.

We separated the students in segments with dissimilar profiles, the students from the same segment have the closest profile, and the ones from different segments have the most different one. The students were divided into 2 clusters with the current specific distinctiveness:

Cluster 0: Students who passed all exams in the last academic semester (Figure 2); a number of 42 instances belong to this cluster (84%);

Cluster 1: Students who failed one or more exams in the last academic semester (Figure 3); a number of 8 instances belong to this cluster (16%);

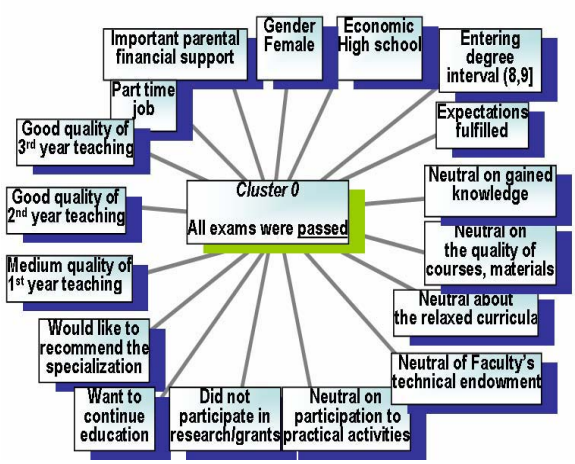

Figure 2. Cluster 0

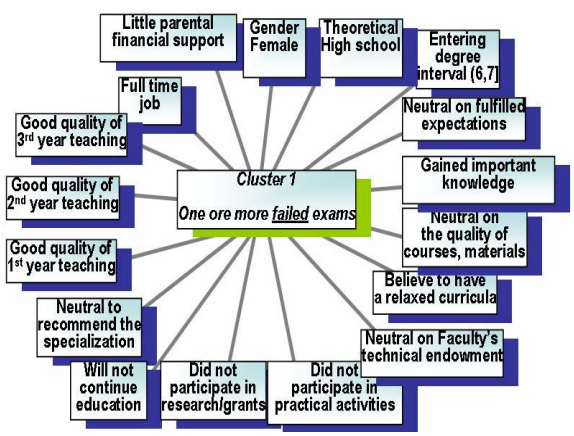

Figure 3. Cluster 1

As resulted from the Gain Ration attribute evaluator and from the previous figures, the highest ranked attributes with an important role in data clustering were the students desire to continue or not their education with post university studies, or try a different specialization, as well as the fulfillment of their prior expectation regarding their present specialization, the parental financial support, job, their view upon the academic curricula, and of the knowledge gained.

4.2. The classification learning process

The classification learning was used to predict the students' failure/success to pass the academic exams based on their present behavioral profile.

For the J48 classification learning based on the training set, we obtained a 96% success rate (the correctly classified instances), and for the cross-validation experiment we acquired a 76% success rate. With J48 we utilized the Laplace estimator, which initiate all numbering starting with 1 as a substitute of 0, a standard technique named after the great mathematician of the 18th century Pierre Laplace.

The classification model is a decisional tree (Figure 4) both in textual and graphic founded on the Scholastic attribute (exams success/failure). As presented in the next figure, the first ramification appears at ContinueEd (students choice in continuing their education with post university studies or other specialization) attribute, and for the second level, the ramification is based on the entering_degree attribute (students admittance grade based on baccalaureate, high school final degree, etc.) and expectations attribute (the fulfillment of their prior expectation regarding their present specialization). The decision tree is represented by a number of leaves (6) and a number of nodes stating its size (9).

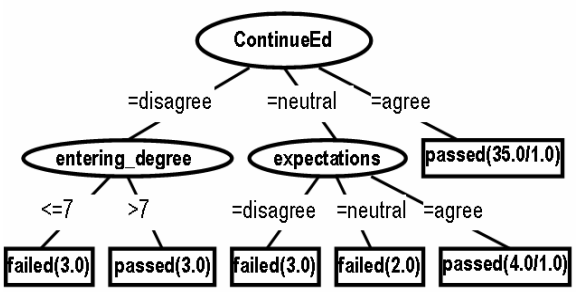

Figure 4. The J48 decision tree

Here are some suggestive examples of interpretation of the decision tree's branches:

"If students were strongly focused on continuing their education after graduating with post university studies or a different specialization, then they would pass all their exams"

"If students were neutral about continuing their education after graduating with post university studies or a different specialization, and believe all their expectations regarding the present specialization were fulfilled, then they would pass all their exams"

"If students did not want to continue their education after graduating with post university studies or a different specialization, and their admittance grade was equal or below 7 (Romania uses a numerical grading system of 1-10), then they would fail one or more exams"

5. Conclusions

The results presented in this article are a part of a larger research which is to be used to make numerous correlations, analysis and to be presented to the higher education institution managers, to offer a better knowledge of students' present scholastic situations, their opinions regarding the each component of the educational process, and to predict some important aspects of their future scholastic situation. The purpose is to contribute to optimal managerial decision taking, in preventing students' exams failure, improving learning abilities and scholastic results.

Our studies continue with deeper mining of academic failure, to detect with pragmatic exactness what and how much the students know, to understand precise learning gaps, and also improve teaching methods and educational management processes.

6. Acknowledgements

The studies included in the present article are an integrated part of TD-329 Grant "Contribution to improving universities' management using modern IT technologies", of the Romanian PN II plan. The authors would also like to thank the students and managers of the Faculty of Economics and Business Administration Cluj-Napoca, for their support in the research, and also the Romanian National University Research Council CNCSIS.

7. References

[1] Anjewierden A, Kollöffel B, Hulshof C. Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. ADML 2007, Crete; September 2007. p. 27-36.

[2] Bação F, Lobo V, Painho M. Clustering Census Data: Comparing the Performance of Self-Organising Maps and K-means Algorithms. KDNet Symposium, Bonn, Germany; June 2004.

[3] Bresfelean VP, Bresfelean M, Ghisoiu N, Comes CA. Data mining clustering techniques in academia, ICEIS 2007 9th International Conference on Enterprise

Information Systems, Funchal, Portugal; 12-16, June 2007. p. 407-41.
[4] Bresfelean VP, Bresfelean M, Ghisoiu N, Comes CA. Data mining in Continuing Education. INTED 2007, Valencia, Spain; March 7-9 2007.
[5] Bresfelean VP, Analysis and predictions on students' behavior using decision trees in Weka environment, ITI 2007, Cavtat, Croatia; June 2007. p. 51-56.
[6] Bresfelean VP, Bresfelean M, Ghisoiu N, Comes C-A. Continuing education in a future EU member, analysis and correlations using clustering techniques. Proceedings of EDU '06, Tenerife, Spain; December 16-18, 2006. p.195-200.
[7] Bresfelean VP. K-means Based Data Clustering. The Eighth International Conference on Informatics in Economy IE 2007 -Informatics in Knowledge Society. Bucharest, Romania; May 17-18, 2007.
[8] Grira N, Crucianu M, Boujemaa N. Unsupervised and Semi-supervised Clustering: a Brief Survey. Report of the MUSCLE European Network of Excellence (FP6); August 15, 2005.
[9] Ivancsy R, Kovacs F. Clustering Techniques Utilized in Web Usage Mining. Proceedings of the AIKED 2006, Madrid, Spain;February 15-17, 2006. p. 237–242.
[10] Kalathur S. An Object-Oriented Framework for Predicting Student Competency Level in an Incoming Class, Proceedings of SERP'06 Las Vegas , 2006, p. 179-183
[11] Leondes C. (Ed.) Intelligent Knowledge-Based Systems Volume 1: Knowledge-based Systems, Kluwer Academic; 2005.
[12] Luan Jing, Data Mining Applications in Higher Education, SPSS Exec.Report; 2004. http://www.spss.com/home_page/wp2.htm
[13] Minaei-Bidgoli B, Punch WF, Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System, GECCO 2003 Conference, Springer-Verlag, Vol 2, Chicago, USA; July 2003. p. 2252-2263.
[14] Mostow J, Beck J, Cen H, Cuneo A, Gouvea E, Heiner C. An educational data mining tool to browse tutor-student interactions: Time will tell! Proceedings of the Workshop on Educational Data Mining, Pittsburgh, USA; 2005. p.15-22.
[15] Myller N, Suhonen J, Sutinen E. Using data mining for improving web-based course design, Proceedings ICCE'02 of the International Conference on Computers in Education, Auckland, New Zealand vol.2; December, 2002. p. 959-963.
[16] Pimentel EP, Omar N. Towards a model for organizing and measuring knowledge upgrade in education with data mining, The 2005 IEEE International Conference on Information Reuse and Integration, Las Vegas, USA; August 15-17, 2005. p. 56-60
[17] Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
[18] Samson V, Bouthemy P. Learning classes for video interpretation with a robust parallel clustering method. Proceedings of ICPR'04, Cambridge, UK; August 2004.
[19] San OM, Huynh VN, Nakamori Y. An Alternative Extension of The K-Means Algorithm for Clustering Categorical Data. Int. J. Appl. Math. Comput. Sci., Vol. 14, No. 2, 2004. p. 241–247.
[20] Tjortjis C, Keane J. T3: A Classification Algorithm for Data Mining. Lecture Notes in Computer Science, Vol.2412. Proceedings of IDEAL 2002 conference, Manchester, UK; August 2002. p. 50-55.
[21] Universitatea Babes-Bolyai Cluj-Napoca, Romania. Programul Strategic al Universitatii Babes-Bolyai (2007-2011), Nr.11.366; 1 august 2006.
[22] Vandamme JP, Meskens N, Superby JF. Predicting Academic Performance by Data Mining Methods, Education Economics, Vol. 15, Issue 4; December 2007. p. 405-419
[23] Zhao Y, Karypis G. Data Clustering in Life Sciences. Molecular Biotechnology, Vol. 31, Humana Press Inc.; 2005. p. 55-80.
[24] Witten IH, Frank E. Data mining: practical machine learning tools and techniques, 2nd ed., Morgan Kaufmann series in data management systems, Elsevier Inc.; 2005.