Data Mining Applications in Higher Education and Academic Intelligence Management

Vasile Paul Bresfelean

Babes-Bolyai University, Faculty of Economics and Business Administration

27. June 2008

15

Data Mining Applications in Higher Education and Academic Intelligence Management

Vasile Paul Bresfelean
Babes-Bolyai University of Cluj-Napoca, Faculty of Economics, Romania

1. Introduction

Higher education institutions are nucleus of research and future development acting in a competitive environment, with the prerequisite mission to generate, accumulate and share knowledge. The chain of generating knowledge inside and among external organizations (such as companies, other universities, partners, community) is considered essential to reduce the limitations of internal resources and could be plainly improved with the use of data mining technologies.

Data mining has proven to be in the recent years a pioneering field of research and investigation that faces a large variety of techniques applied in a multitude of areas, both in business and higher education, relating interdisciplinary studies and development and covering a large variety of practice. Universities require an important amount of significant knowledge mined from its past and current data sets using special methods and processes. 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 academic processes.

Higher education institutions have long been interested in predicting the paths of students and alumni (Luan, 2004), thus identifying which students will join particular course programs (Kalathur, 2006), 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 (Vandamme et al., 2007) 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 (Piementel & Omar, 2005).

The distance and on-line education, together with the intelligent tutoring systems and their capability to register its exchanges with students (Mostow et al., 2005) 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 (Myller et al., 2002). Scientists (Anjewierden et al., 2007) 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

Source: Theory and Novel Applications of Machine Learning, Book edited by: Meng Joo Er and Yi Zhou,
ISBN 978-3-902613-55-4, pp. 376, February 2009, I-Tech, Vienna, Austria

educational web-based systems (Minaei-Bidgoli & Punch, 2003). A combination of multiple classifiers led to a significant improvement in classification performance through weighting the feature vectors.

The author's research directions through the data mining practices consist in finding feasible ways to offer the higher education institutions' managers ample knowledge to prepare new hypothesis, in a short period of time, which was formerly rigid or unachievable, in view of large datasets and earlier methods. Therefore, the aim is to put forward a way to understand the students' opinions, satisfactions and discontentment in the each element of the educational process, and to predict their preference in certain fields of study, the choice in continuing education, academic failure, and to offer accurate correlations between their knowledge and the requirements in the labor market. Some of the most interesting data mining processes in the educational field are illustrated in the present chapter, in which the author adds own ideas and applications in educational issues using specific data mining techniques.

The organization of this chapter is as follows. Section 2 offers an insight of how data mining processes are being applied in the large spectrum of education, presenting recent applications and studies published in the scientific literature, significant to the development of this emerging science. In Section 3 the author introduces his work through a number of new proposed directions and applications conducted over data collected from the students of the Babes-Bolyai University, using specific data mining classification learning and clustering methods. Section 4 presents the integration of data mining processes and their particular role in higher education issues and management, for the conception of an Academic Intelligence Management. Interrelated future research and plans are discussed as a conclusion in Section 5.

2. Data mining applications in the large spectrum of education

Data mining is an important data analysis methodology that has been successfully employed in many domains, with numerous applications in educational issues and was identified as one of the ten emergent technologies of the 21st century by the MIT Technology Review. It 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. The educational data mining was defined as "the process of converting raw data from educational systems to useful information that can be used to inform design decisions and answer research questions" (Heiner et al., 2006)

A main concern of each institution of higher education is to predict the paths of students and alumni (Luan, 2004). They would like to identify which students will join particular course programs, and which students will require assistance in order to graduate. At the same time institutions want to learn whether some students more likely to transfer than others, and what groups of alumni are most likely to offer pledges. In addition to this challenge, traditional issues such as enrolment management and time-to-degree continue to motivate higher education institutions to search for better solutions.

The prediction of class configuration based on the course prerequisites and the prior courses taken by the students is another research illustration of educational data mining. Equipped with this information (Kalathur, 2006), the instructor can undertake remedial measures by supplementing the lecture material with the required topics which the students were lacking from their previous courses. After acquiring real data and building the database about the

students and their course history, the model is trained with students' data and the system could provide immediate feedback to the student upon enrolling in a course how their profile fits with the course requirements.

Various studies are based on students' present "knowledge luggage", detecting with realistic accuracy what and how much the students know, in order to deduce specific learning gaps. This set of information could be obtained during an ongoing learning assessment process that makes possible to specify, with reasonable precision, which subject the student is better suited to learn at that moment, and requires automatic or semi-automatic procedures for treatment and analysis for acquisition of new knowledge. Pimentel and Omar (2005) presented a model for organizing and measuring knowledge upgrade in systems of education and learning with the support of data mining tools.

Academic failure among freshmen has long inspired a large number of questions , many psychologists seeking to comprehend and explicate it, and many statisticians have tried to predict it. Vandamme and collaborators (2007) attempted to classify, as early in the academic year as possible, students into three groups: the 'low-risk' students, with a high probability of succeeding; the 'medium-risk' students, who may succeed thanks to the measures taken by the university; and the 'high-risk' students, with a high probability of failing (or dropping out). They present the results of their application of discriminant analysis, neural networks, random forests and decision trees aiming to predict those students' academic success.

Another motivating issue in educational data mining is finding the reasons why students abandon school or certain courses before concluding the subjects required. One methodology (Antunes, 2008) consisted of discovering the frequent sequential patterns among the recorded behaviors, keeping the discovery limited to the sequences that are approximately in accordance to the existing background knowledge. The methodology assumes that the existing background knowledge can be represented by a context-free language, which plays the role of a constraint in the sequential pattern mining process. The curriculum knowledge was represented as a finite automaton, which established the order of subjects that a student should attend to finish his graduation.

Since universities courses progressively oblige students to use online tools in their studies, there are numerous prospects to mine the resultant large quantity of student learning data for hidden valuable information. Several approaches in scientific literature try to classify students in order to predict their final grade based on features extracted from logged data in educational web-based systems (Minaei-Bidgoli & Punch, 2003). A combination of multiple classifiers led to a significant improvement in classification performance through weighting the feature vectors; by using a Genetic Algorithm researchers could optimize the prediction accuracy and got a marked improvement over raw classification.

Intelligent tutoring systems' capability to register its exchanges with students is an important challenge and an opportunity (Mostow et al., 2005). A central matter in mining data from an intelligent tutoring system is "What happened when…?". In comparison to individual observation of live or videotaped tutoring, logs can be more far-reaching in the number of students, more comprehensive in the number of sessions, and exquisite in details, avoiding thus observer effects, costing less to obtain, and easier to analyze. Mostow and collaborators describe (2005) an educational data mining tool to support such case analysis by exploiting three simple but powerful ideas: first, a student, computer, and time interval suffice to specify an event; second, a containment relation between time intervals defines a

hierarchical structure of tutorial interactions; third, the first two ideas make it possible to implement a generic but flexible tool for mining tutor data with minimal dependency on tutor-specific details.

A number of data mining studies were founded on distance and on-line education, which presents some probable information sources for data mining processes. These studies based on collecting and interpreting the information from several courses may assist teachers and students in the web-based learning setting. This information could be applied for assigning varied groups in programming courses or projects and to evaluate the actual learning (Myller et al., 2002).

On-line collaborative discussions have significant role in distance education and web-enhanced courses. Automatic tools for assessing student activities and promoting collaborative problem solving can offer an improved learning practice for students and also offer useful assistance to teachers. Researchers developed a specific mining tool for making the configuration and execution of data mining techniques easier for instructors and in order to be of use for decision making, using real data from on-line courses (Romero et al., 2008). Others (Anjewierden et al., 2007) derived models for classifying chat messages using data mining techniques, tested them and established the reliability of the classification of chat messages comparing the models performance to that of humans. (Ravi et al., 2007) presented an approach that could be used to scaffold undergraduate student discussions by retrieving useful information from past student discussions and an instructional tool that profiled student contributions with respect to student genders and the roles that students play in discussion.

3. Experiments in education based data mining at Babes-Bolyai University

The quality of an institution of higher education is specified among other concerns by its adapting know-how to the continuous varying requirements of the socio-economic background, the quality of the managerial system based on a high level of professionalism and on applying the latest technologies (Bresfelean, 2007).

The author's recent experiments at the Babes-Bolyai University of Cluj-Napoca were focused mostly on two of the main learning methods in data mining: classification learning and data clustering. The main objectives for the data mining practices were to offer the higher education institutions' managers ample knowledge to prepare new hypothesis, in a short period of time, which was precedently rigid or unachievable, in view of large datasets and earlier methods. It was aimed to put forward a way to understand the students' opinions, satisfactions and discontentment in every element of the educational process, and to predict their preference in certain fields of study, the choice in continuing their education and also the understanding, prediction and prevention of the academic failure. Building a profile for the students, and their grouping based on exam failure risk is a very 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.

There are two keys to success in data mining (Edelstein, 1999). First is coming up with an exact formulation of the problem to solve. A focused report usually results in the best payoff. The second key is using the right data. After choosing from the data available, or

perhaps buying external data, one may need to transform and combine it in significant ways.

As presented in the next table (Table 1.), the research was based on questionnaires' collected data from Romanian senior undergraduate students and master degree students from The Faculty of Economics and Business Administration, Babes-Bolyai University of Cluj-Napoca, using on-line and written surveys in order to evaluate their motivation in continuing education and the accomplishment regarding the educational process (Bresfelean et al., 2006), and included multiple choice questions and a few questions requiring written answers. There were also important data collected from faculty's databases, such as: tuition database, students' scholastic situation database, etc. The information collected consists of: general data on the subject (gender etc.), scholastic situation (grades, failed exams), several types of gained scholarships, interruption of study, exams absence, tuition, and also students' opinion (on courses, materials, curricula, research, teachers, laboratories technical novelty, knowledge gained, continuing education) etc.

Weka and RapidMiner workbenches, which are collections of machine learning algorithms and data preprocessing tools, were used to analyze the data, and proved to be valuable tools in order to gain insight into how certain processes are handled within higher education institutions. Weka workbench is an open source software issued under the GNU General Public License, a collection of machine learning algorithms for data mining tasks. It is currently developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. RapidMiner is another important open-source system for knowledge discovery and data mining, providing more than 400 operators for all main machine learning procedures, including input and output, and data preprocessing and visualization. It is written in the Java programming language and therefore can work on all popular operating systems.

3.1 Classification learning experiments

In classification learning, the learning design is offered with a set of classified examples from which it is estimated to learn a way of classifying unseen examples (Witten & Frank, 2005). Decision trees represent a supervised approach to classification and the models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions (Edelstein, 1999). A "divide-and-conquer" approach (Witten & Frank, 2005) to the problem of learning from a set of independent instances leads naturally to a style of representation called decision tree.

There is a significant number of different algorithms that can be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), Quest, C4.5 and C5.0 etc. Decision trees are grown through an iterative splitting of data into discrete groups, where the goal is to maximize the "distance" between groups at each split. Leaf nodes give a classification that applies to all instances that reach the leaf or a set of classifications, or a probability distribution over all possible classifications (Witten & Frank, 2005).

For the classification learning experiments the J48 method was chosen (based on the C4.5 algorithm from the machine learning), for being one of the most used Weka classification algorithms that offers a superior stability between precision, speed and interpretability of results. The basic algorithm for decision tree induction is a greedy algorithm that generates decision trees in a top-down recursive divide-and-conquer manner.

Data subjects	Instruments of collecting data	Information collected	Data mining method	Results of data mining/ Knowledge obtained
Undergraduate senior students	Questionnaires, Faculty's databases: -tuition database, -students scholastic situation database, etc.	•General information (gender etc.); •Opinions on: -fundamental knowledge gained, -books, course materials, case studies, -curricula, practical activities, -participation to research, grants, -recommending the specialization to future students, -courses teaching methods in each of the years of study; -continuing education; •Gained scholarships; •Parents' material support; •Scholastic situations and degrees. etc.	*Classification Learning (based on C4.5 algorithm)*	☺Prediction of the students' choice in continuing their education with post university studies (master degree, Ph.D. studies etc.); and their preference in certain fields of study. ☺Prediction of students failing to pass their exams
			Data Clustering (based on K-means algorithm)	☺Grouping students in clusters with dissimilar behavior, the students from the same cluster embrace the closest behavior, and the ones from different clusters have the most different one. ☺Drawing up the students profile based on their choice to continue their education. ☺Grouping students in clusters based on the probability of: -passing their exams, -obtaining a scholarship
Alumni – presently master degree students	Questionnaires	•General information (gender etc.); •Opinions on: -contentment in the graduated and in chosen master degree specialization, -fundamental knowledge gained, -books, course materials, case studies, -curricula, practical activities, -participation to research, grants, -recommending specialization, -courses teaching methods in each of the years of study; -continuing education; -undergraduate/master degree courses considered important or outdated; •Details on present job; •Gained scholarships; •Competences obtained; •Parents' material support; •Scholastic situations, degrees etc.	*Classification Learning (based on C4.5 algorithm)*	☺Prediction of the students' choice in continuing their education (Ph.D. studies etc.); and their preference in certain fields of study.
			Data Clustering (based on K-means algorithm)	☺Grouping students in clusters with dissimilar behavior, the students from the same cluster embrace the closest behavior, and the ones from different clusters have the most different one. ☺Drawing up the students profile based on: - their present choice to continue their education. -present job field (Does it correspond to the graduated or the master degree specialization?)

Table 1. Author's experiments in education based data mining at Babes-Bolyai University, after (Bresfelean et al., 2008b)

A general approach[1] to the decision tree algorithm can be summarized as following:

1. Choose an attribute that best differentiates the output attribute values.
2. Create a separate tree branch for each value of the chosen attribute.
3. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
4. For each subgroup, terminate the attribute selection process if:
   a. All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.
   b. The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.
5. For each subgroup created in (3) that has not been labeled as terminal, repeat the above process.

The purpose of the first classification learning experiments (Bresfelean, 2007) was to predict of the students' choice in continuing their education with post university studies (master degree, Ph.D. studies etc.) and their preference in certain fields of study. The results consist in several decision trees generated upon the initial data set, then on a number of filtered instances, corresponding to the students belonging to certain specializations. There were generated values of several performance measures for the classification problems described in the previous table: Kappa statistic, MAE (mean absolute error), RMSE (root mean square error), RAE(relative absolute error, %), RRSE (root relative squared error,%).

An illustration of a decision tree resulted from 409 instances (2007-2008 senior students from all specializations of the Faculty of Economics and Business Administration), has as a central root joint the Curricula attribute (opinions about whether the curriculum was relaxed and gave time to individual studying). The next levels of ramification are based on the futureJob attribute (the confidence in finding a job appropriate to their specialization after graduating) and 1st_year attribute (students' evaluation of courses teaching methods of the 1st year of study, this year being a test for the freshmen).

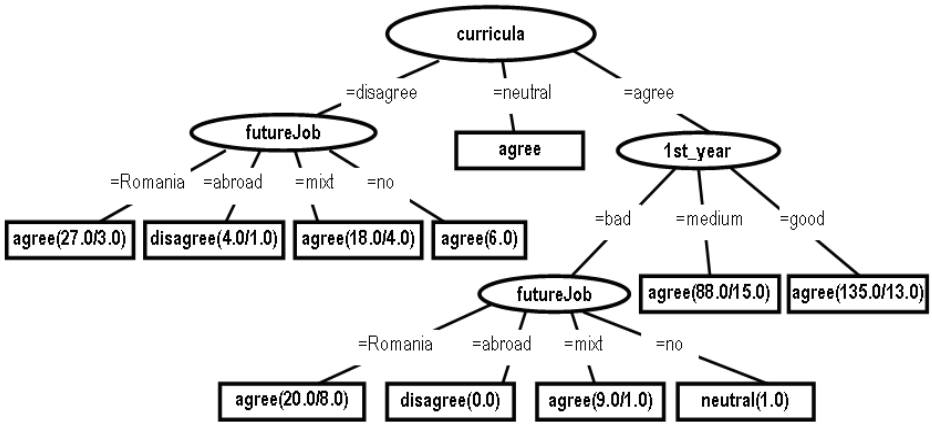

Fig. 1. Decision tree for predicting the students' choice in continuing their education

[1] Minnesota State University, http://grb.mnsu.edu/grbts/doc/manual/J48_Decision_Trees.html

Examples of interpretation of the decision tree's branches:
"If the students agreed they had relaxed curricula which allowed time for individual studying, and had a good opinion about the quality of courses teaching methods in the 1st year of study, then they would agree to continue their education with post university studies".
"If the students disagreed they had relaxed curricula which didn't allow time for individual studying, and were confident to find a job abroad appropriate to their specialization after graduation, then they would not agree to continue their education with post university studies".
The classification learning was also 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, there was a 75,79% success rate (the correctly classified instances), and for the cross-validation experiment we acquired a 72,86% success rate. The Laplace estimator was used with J48, which initiated all numbering starting with 1 as a substitute of 0, a standard technique named after the great mathematician of the 18th century Pierre Laplace. In the next figure (Fig.2) the first ramifications appear at entering_degree numerical attribute (students admittance grade in the Romanian 1-10 numerical grading systems, based on baccalaureate, high school final degree, etc.), and for next levels the ramification is based on expectations attribute (the fulfillment of their prior expectations regarding their present specialization) and the parents_sup attribute (the financial support received from their parents).

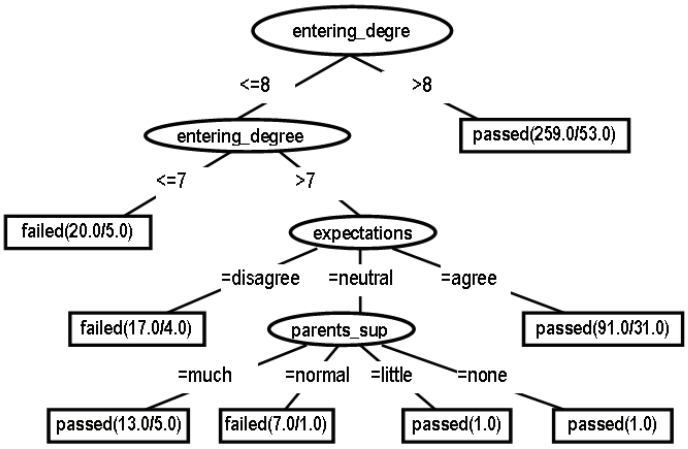

Fig. 2. Decision tree for predicting academic failure/success

Here are examples of interpretation of the decision tree's branches:
"If students' admittance grade was above 8, then they would pass all their exams"
"If students' admittance grade was in the (7,8] interval, were neutral that their expectations regarding the present specialization were fulfilled, believed the financial support from their parents was normal, then they would fail one or more exams".
"If students' admittance grade was in the (7,8] interval, did not agree that their expectations regarding the present specialization were fulfilled, then they would fail one or more exams".

The previous experiments were conducted over significant data collected from senior undergraduate students from all specializations of the Faculty of Economics and Business Administration. Several analogous experiments utilized data also from previous years senior students (currently master degree students or alumni), and were concentrated on certain specializations such as Business Information Systems (IE – Informatica economica), Marketing, Management, or Accounting (CIG – Contabilitate si informatica de gestiune) with different results to some extent from one specialization to another (Bresfelean et al., 2008a), (Bresfelean, 2007).

3.2 Data clustering experiments

The clustering process is a practice in which a set of data is replaced with clusters, which symbolize collections of data points belonging together, its success often being measured subjectively in terms of how useful the result appears to be to a human user (Witten & Frank, 2005). Clustering has been extensively used to partition data into groups so that the level of association is high between members of the same group and low between members of dissimilar groups (Jung et al., 2004).

The clustering algorithms generally follow hierarchical or partitional approaches. Several algorithms have been proposed in the literature for clustering, among which K-means clustering algorithm is the most commonly used because it can be easily implemented (Hung et al., 2005). For the partitional approach, the K-means and its variants, such as the fuzzy c-means algorithm, are the most popular algorithms.

In the recent data mining research (Bresfelean et al., 2006), (Bresfelean et al., 2007), we applied the clustering method called FarthestFirst which implements the transversal algorithm of Hochbaum and Shmoys, a simple, fast, approximation method based on the K-means algorithm. The idea (Dasgupta & Long, 2005) is to pick any data point to start with, then choose the point furthest from it, then the point furthest from the first two (the distance of a point x from a set S is the usual min $\{d(x, y) : y \in S\}$), and soon until k points are obtained. These points are taken as cluster centers and each remaining point is assigned to the closest center. If the distance function is a metric, the resulting clustering is within a factor two of optimal.

One of the main goals in applying the data clustering methods was to group students in clusters with dissimilar behavior; the students from the same cluster embrace the closest behavior, and the ones from different clusters have the most different one (Bresfelean et al., 2006), (Bresfelean et al., 2007). At the same time this process facilitates the drawing up the students profile based on their choice to continue their education, but also on the academic failure risk. We used the analysis based on data clustering with the purpose to classify students founded on their present job field, made a number of correlations, and tried to answer an important question: "Does it correspond to the graduated or to the master degree specialization?". In this way, we tried to get a feed-back from our alumni and/or master degree students, resulting in some important information for the higher education managers, a part of a superior sequence: "Are the current specializations competitive on the labor market?"

The next study was conducted over senior undergraduates and master degree students from one of the last generations of the four-years undergraduate first cycle (one-year Master, 4-years doctorate) during 2006-2008, before the full implementation of Bologna declaration (first degree of three years, two-years Master, 3-years doctorate). Using the FarthestFirst

clustering method based on K-means algorithm (Bresfelean et al., 2006), (Bresfelean et al., 2007), we initialized the k cluster centers to k randomly chosen points from the data, which was partitioned based on the minimum squared distance criterion (Maulik & Bandyopadhyay, 2002). In our experiment, the k parameter is 3, corresponding to students' 3 choices in continuing their post university studies: disagree, neutral, agree. The cluster centers were then updated to the mean or the centroid of the points belonging to them. This entire process was repeated until either the cluster centers did not alter or there was no major change in the J values of two successive iterations. At this point, the clusters were stable and the clustering process ended. The clustering process proved to be particularly useful in dividing the students in segments with different behavioral models, the students from the same segment have the closest behavior, and the ones from different segments have the most different one.

Based on the students' choice to continue their education we divided them into 3 groups (Bresfelean et al., 2006), each presenting specific centroids, with an optimistic result after Weka validation (27.4151 % of the instances were incorrectly clustered):

Group 0: Students agree to continue their post university studies;

Group 1: Students do not agree to continue their post university studies;

Group 2: Students are neutral to continue their post university studies.

As a result of applying the FarthestFirst algorithm, we obtained 3 clusters with the following centroids:

Attributes (*Information/Opinion*):	**Cluster 0 Agree**	**Cluster 1 Disagree**	**Cluster 2 Neutral**
Gender	Male	Female	F
Specialization	IE-Business Information systems	Marketing	Management
Graduated High school	Other (mainly theoretical)	Agricultural profile	Economical profile
Their expectations regarding the specialization were fulfilled	Agree	Disagree	Neutral
Gained important knowledge	Neutral	Disagree	Agree
Were offered high quality courses, materials	Agree	Disagree	Agree
The curriculum was relaxed	Agree	Disagree	Neutral
The faculty had a good technical endowment	Agree	Neutral	Agree
Participated in practical activities	Agree	Disagree	Disagree
Participated in research/grants	Neutral	Disagree	Neutral
Would recommend the specialization	Agree	Disagree	Neutral
Teaching quality in the 1'st year	Good	Very bad	Medium
Teaching quality in the 2nd year	Good	Very bad	Medium

Teaching quality in the 3rd year	Good	Very bad	Excellent
Teaching quality in the 4th year	Good	Very bad	Excellent
Present job	none	Part time	Part time
Benefited from parents' financial support	Very much	No	Much
Expect to find a future job	In Romania	In Romania	No
Failed exams	none	3 or 4 failed	1 or 2 failed

Table 2. FarthestFirst clusters based on students' choice to continue their education, adapted from (Bresfelean et al., 2006)

The needed information is extracted from the clusters' centroids. Following this, it was determined that there were no common values fields for the three clusters, and as a result all the fields contain relevant information for the segmentation process.

The opinion on relaxed curriculum plays a substantial part in differentiating the clusters population:

cluster 0 - Agree
cluster 1 - Disagree
cluster 2 - Neutral

The same situation is observed in the case of the following: opinion on expectations' fulfillment regarding the specialization, and the opinion on recommending the specialization to future students.

The failed exams attribute (scholastic situation at the end of last academic year) also contains significant information in differentiating the cluster population:

cluster 0 - none (passed all exams at the end of last academic year)
cluster 1 - 3 or 4 failed exams
cluster 2 - 1 or 2 failed exams

Moreover, the opinion on the quality of courses teaching methods in the 1st and the last years of study (4th year) plays an important role in defining the clusters as seen in Table 2.

To fundament the decisions regarding the managerial strategies the faculty leaders can approach in order to fulfill all students' expectations, and enhance their competiveness on the labor market, it is compulsory to correlate the information extracted from terminal year students' questionnaires with graduate students' data, currently master degree students. Starting from the information mined in the undergraduate and master degree questionnaires, the following correlations and analysis were concluded:

- correlation and percentage relation between the graduated specialization and the master degree specialization;
- correlation and percentage relation between the current job and the graduated specialization;
- correlation and percentage relation between the current job and the master degree specialization.

The following table (Table 3) presents the data extracted from the questionnaires filled up by master degree students from the Faculty of Economics and Business Administration, Cluj-

Napoca, filtered to include only the students from Business Information Systems -IE, Marketing -Mk and Management –Mng master specializations.

Master degree students / Cathegories	IE	Mk	Mng
Total master degree students in the above specialization	11	40	15
First degree specialization similar to master specialization	10	16	6
First degree specialization different from master specialization	1	24	9
Job in other areas than the graduated specialization	4	11	5
Similar job to the graduated specialization	5	9	2
Job in other areas than the master degree specialization	5	13	6
Similar job to the master degree specialization	4	7	1
Unemployed IE master degree students	2	20	8

Table 3. Master degree students - Jobs and specializations, from (Bresfelean et al., 2006)

In the next tables and diagrams I present the correlations and percentage relations between different attributes suggestive to this study.

Master degree specialization	First degree specialization similar to master specialization (%)	First degree specialization different from master specialization (%)
IE	90,9%	9,09%
Mk	40%	60%
Mng	40%	60%

Table 4. Percentage relation between graduated specialization and the master degree specialization

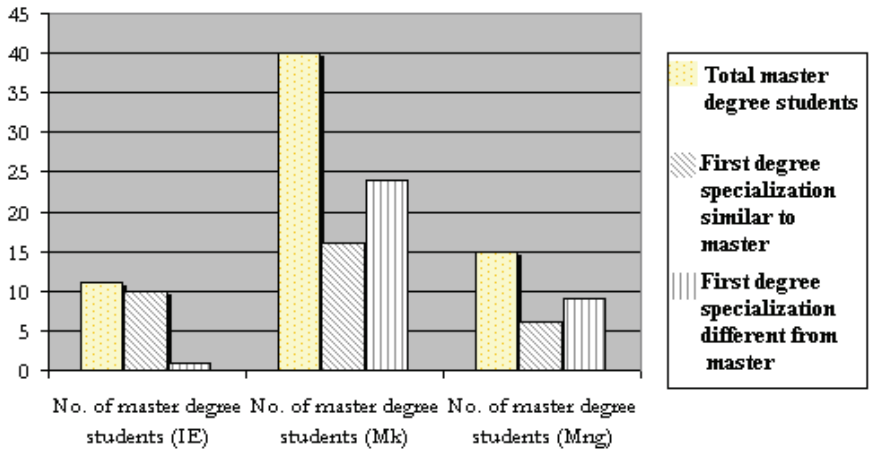

Fig. 3. Correlation between the graduated specialization and master degree specialization

Graduated specialization	Job in other area than the graduated specialization (%)	Similar job to the graduated specialization (%)	Unemployed master degree students of the present specialization (%)
IE	36,36%	45,45%	18,18%
Mk	27,5%	22,5%	50%
Mng	33,33%	13,33%	53,33%

Table 5. Percentage relation between the current job and the graduated specialization

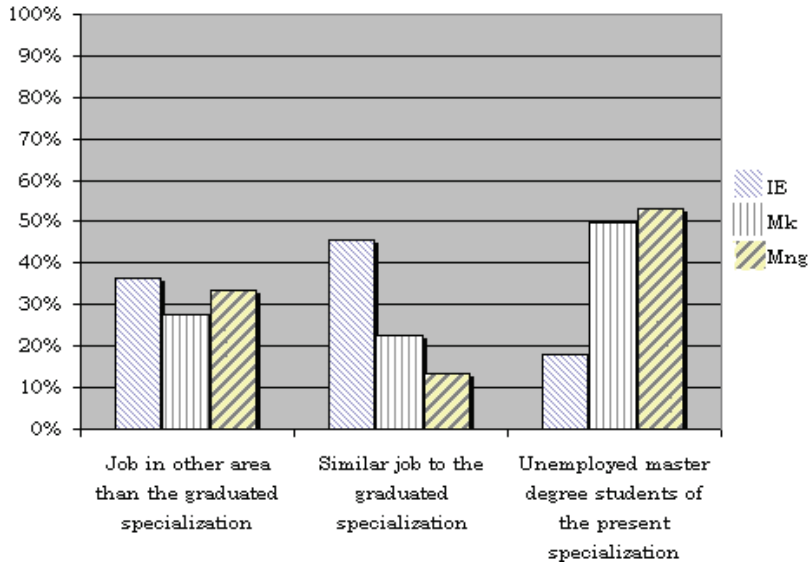

Fig. 4. Graphical representation of the percentage relation between the current job and the graduated specialization

Master degree specialization	Job in other areas than the master degree specialization (%)	Similar job to the master degree specialization (%)	Unemployed IE master degree students (%)
IE	45,45%	36,36%	18,18%
Mk	32,5%	17,5%	50%
Mng	40%	6,66%	53,33%

Table 6. Percentage relation between the current job and the master degree specialization

From the data analysis, correlation and percentage relations presented in this study, and based on other detailed research included in (Bresfelean et al., 2006) and (Bresfelean et al., 2007), we can conclude that:

- The majority of the undergraduate IE students were keen on continuing their education with master degree studies, while the undergraduate Mng students formed an important segment with neutral opinions on continuing education, and the undergraduate Mk students formed another segment with negative opinions on continuing education.

- The great majority of IE master degree students (approximate 90,9%) were formed by former IE graduate students, and only a small percent of other than IE graduates (approx. 9,09%);
- The majority of Mk and Mng master degree students (approx. 60%) were formed by students that didn't previously graduate the same specialization;
- An important percent (45,45%) of the IE master degree students found a similar job to the graduated specialization, and 36,36% of IE master degree students had an occupation similar to the master specialization;
- A small percent (22,5%) of the Mk master degree students found a similar job to the graduated specialization, and 17,5% of Mk master degree students had an occupation similar to the master specialization;
- A very small percent (13,33%) of the Mng master degree students found a similar job to their graduated specialization, and 6,66% of Mng master degree students had an occupation similar to the master specialization;
- Only a small percent (18,18%) of the IE master degree students were unemployed, but half of Mk (50%) and the majority of Mng (53,33%) master degree students were unemployed for different reasons, not mentioned in the questionnaires.
- Due to the financial support obtained from different companies, banks etc. we observed an increased number of students to other than IE master degree specializations; Mk and Mng master degree specializations attracted a large number of graduate students from other areas.

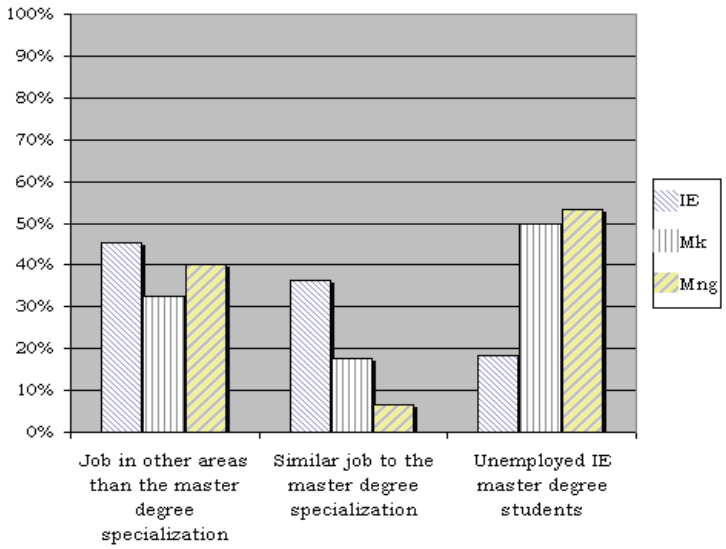

Fig. 5. Graphical representation of the percentage relation between the current job and the master degree specialization

4. The integration of data mining processes in higher education topics

The higher education institutions represent dissimilar and complex environments which involve links to communication and collaboration among its various departments and the

society. In the situation of a state-financed higher education, the society manifests as the main partner of a university, and is represented at the purveyor-client interface by central or local governmental institutions, companies and organizations, labor management institutions etc. (Rusu & Bresfelean, 2006).

The European Council stated in Lisbon 2000 that the Europe should became by 2010, "the most competitive and dynamic knowledge-based economy in the world, capable of sustainable economic growth with more and better jobs and great social cohesion". The continuous change in European educational expectations due to Bologna Process and the demands for an EU area of educational collaboration have been gradually replacing the old-fashioned routine management with ICT-based knowledge management, leading to the emergence of an Academic Intelligence Management.

The Academic Intelligence Management includes all higher education institution's processes utilized to acquire, generate and spread knowledge in order to accomplish its objectives and strategies, based on the latest ICT (Information and Communication Technologies) and collaborative practices. It is tied to organizational goals such as improved performance, competitive advantage, innovation, research and development, which derive from technologies and applications providing historical, present, and predictive analysis of all academic activities. 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. Consequently, an integration of the latest research results involving these technologies, in terms of their contents and impact, is an issue that should be vital, while taking into account the fundamental role of a university as a knowledge creator and facilitator of teaching and research. Such is the case of all Romanian higher education institutions involved in complex processes of evaluation and accreditation: the traditional universities, which aim to develop their activities to include new areas, or the recently established private institutions aspirant to achieve university status.

The university is progressively regarded as a collaborative organization whose mission is to foster knowledge creation and knowledge diffusion among communities of students, scholars and researchers (Rodrigues & Barrulas, 2003). In the case of higher education institutions, the designing circuits of a managerial system (Rusu & Bresfelean, 2006) aiming an academic intelligent/intelligence management (Fig. 6) must be closely tied to:

- educational process activities: structure design and curricula content must be token of labor market requests and institution's capability; quality and freshness of courses information; adequate teaching/learning and evaluation methods; appropriate performance of educational processes etc.
- scientific research activities: thematic originality and opportunity; consistency of results; scientific probity; ethic experimenting; ways of utilizing the results etc.
- internal organization: authority and responsibility delegation; transparency and efficiency in the utilization of human and material resources; equity and performance encouragement in personnel promotion; continuous personnel training etc.
- external relations with local, national and international community; relations with other educational institutions and companies from different activity sectors; alumni etc.

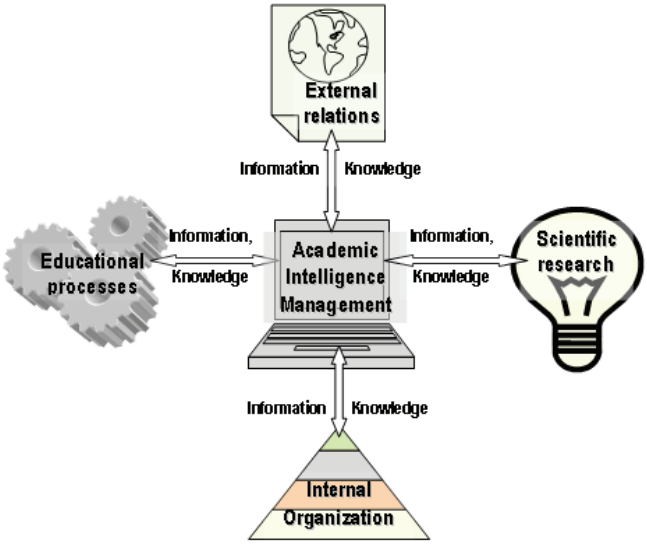

Fig. 6. Academic Intelligence Management, the hub of higher education activities

The higher education institutions' objectives answer the needs of the society and of the labor market and can surpass them, playing an important part in demand's generation (Rusu & Bresfelean, 2006). They serve as reference values for the adjusting circuits between the university-external partners' levels. Starting from this point, university managers set the strategic objectives, which would supply reference values for the lower levels, playing a parameter part in levels' correction. Based on it, in a designing stage, the strategies for developing a coherent academic offer would be implemented for quality management, strategic management, scientific research management and the "must" values for the lower levels: elaboration of education curricula and analytic programmes; scholastic management; school taxes management; accounting; human resource administration etc. The results, amount features of lower levels, are transmitted to the superior ones in a continuous communication within a managerial collaborative system (Fig.7), based on the latest ICT and knowledge management.

The data mining process integrated in the managerial system has two objectives: knowledge discovery, for offering explicit information, and a prediction objective for the forecast of future evolutions and events. By contrast with the normal interrogations addressed to current databases, using an interrogation language, the data mining process classifies and groups different systems data, eventually incompatible, searching for new associations. The decision support (DS) provides a variety of data analysis, simulation, visualization and data modeling techniques, and software such as decision support systems, executive support systems, databases and data warehouses.

Data mining and decision support are two disciplines aimed at solving difficult practical problems, and in many ways they are complementary (Bohanec & Zupan, 2001). To solve a particular problem, DS tends to rely on knowledge acquired from experts, while data mining attempts to extract it from data. Their combination would result in important benefits in solving real-life decision and data-analysis problems:

- Data mining has the prospective of solving decision support problems, when earlier decision support answers was recorded as analysis data to be used with mining tools.
- DS methods typically products a decision model, proving the expert knowledge of decision makers.

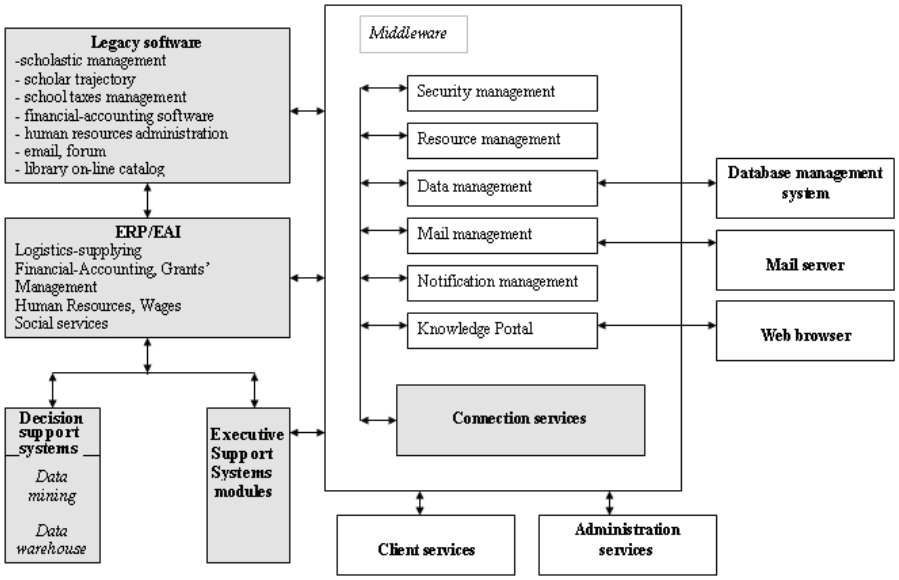

Fig. 7. Conceptual scheme of a collaborative managerial information system, from (Rusu & Bresfelean, 2006), (Dustdar, 2004)

An important step for successful combination will be the switch from the current data mining software tools to a data mining application systems approach (Rupnik et al, 2006) which introduces the possibility to develop decision support systems which use data mining methods and do not demand expertise in data mining for business users. It is an approach which focuses on users and decision makers, enabling them to view data mining models which are presented in a user-understandable manner through a user friendly and intuitive GUI using standard and graphical presentation techniques. Through the use of data mining application systems approach, data mining can become better integrated in business environments and their decision processes.

The exploitation of data mining processes for decision support is based on the CRISP European standard[2] (CRoss Industry Standard Process for Data Mining) proposed in mid 90's by a European consortium of companies as an industry- and tool-neutral data mining process model. The CRISP methodology provides guidelines and a sequence of steps to be followed in the applied knowledge discovery process. The life cycle of a data mining project consists of six phases:

1. Business Understanding (in our case – academic understanding) phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition.

[2] CRISP, http://www.crisp-dm.org/Process/index.htm

2. Data Understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets.
3. Data Preparation phase covers all activities to construct the final dataset from the initial raw data. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.
4. In the Modeling phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.
5. At the Evaluation stage in the project the user has built a model that appears to have high quality, from a data analysis perspective. At the end of this phase, a decision on the use of the data mining results should be reached.
6. Deployment phase. Creation of the model is generally not the end of the project. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process, expanding the obtained model and its results at the level of managerial information system of the higher education institution.

5. Conclusions

The data mining experiments from this chapter are a component of a larger research which is to be used to make several correlations, analysis in order to integrate data mining process in the managerial system for optimal decision support. I offered an insight of how data mining processes are being applied in the large spectrum of education by presenting recent applications and studies published in the scientific literature, considered to be relevant to the development of this emerging science. I presented my work through a number of experiments conducted over questionnaires data and scholastic databases at Faculty of Economics and Business Administration Cluj-Napoca, using classification learning and data clustering methods. In the last part of the chapter I introduced the concept of Academic Intelligence Management, and illustrated the integration of data mining processes and their particular role in higher education management and decision support.
The studies will continue with deeper mining of educational topics, such as performance in scientific research, correlations between the students' knowledge and the competences demanded on the labor market, academic failure, to perceive what and how much the students know, to realize learning gaps, and also improve teaching methods and educational management processes.

6. Acknowledgements

The research included in the present article is a part of Romanian CNCSIS TD-329 Grant "Contributii la perfectionarea managementului institutiilor universitare prin aplicarea de tehnologii informatice moderne" (Contribution to improving universities' management using modern IT technologies).

7. References

Antunes C., Acquiring Background Knowledge for Intelligent Tutoring Systems, Proceedings of Educational Data Mining 2008, The 1st International Conference on Educational Data Mining Montreal, Quebec, Canada, June 20-21, 2008 pp.18-27

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. pp. 27-36.

Bohanec, M., Zupan, B., Integrating decision support and data mining by hierarchical multi-attribute decision models, IDDM-2001: ECML/PKDD-2001 Workshop Integrating Aspects of Data Mining, Decision Support and Meta-Learning, Freiburg, 2001, pp. 25-36.

Bresfelean, V.P., Bresfelean, M., Ghisoiu, N., Comes, C.-A., Determining Students' Academic Failure Profile Founded on Data Mining Methods, 30th International Conference Information Technology Interfaces, ITI 2008, 23-26 June 2008 Cavtat, Croatia (a)

Bresfelean, V.P., Bresfelean, M., Ghisoiu, N., Comes, C.-A., Development of universities' management based on data mining researches, INTED 2008, International Technology, Education and Development Conference, March 3-5 2008 Valencia, Spain (b)

Bresfelean V.P., Analysis and predictions on students' behavior using decision trees in Weka environment, 29th International Conference Information Technology Interfaces, ITI 2007, Cavtat, Croatia, June 2007, pp. 51-56

Bresfelean V.P, Bresfelean M, Ghisoiu N, Comes C-A., Data mining clustering techniques in academia, 9th International Conference on Enterprise Information Systems, 12-16, June 2007, Funchal, Portugal, pp. 407-410

Bresfelean V.P, Bresfelean M, Ghisoiu N, Comes C-A., Continuing education in a future EU member, analysis and correlations using clustering techniques, Proceedings of EDU'06 International Conference, Tenerife, Spain, December 2006, pp. 195-200

Dasgupta S., Long P.M., Performance Guarantees for Hierarchical Clustering, Journal of Computer and System Sciences, Volume 70 , Issue 4, June 2005, Special issue on COLT 2002, pp. 555 - 569

Dustdar, S., Caramba—A Process-Aware Collaboration System Supporting Ad hoc and Collaborative Processes in Virtual Teams, Distributed and Parallel Databases, 15, Kluwer Academic Publishers, 2004

Edelstein H., Introduction to Data Mining and Knowledge Discovery. Third Edition. Two Crows Corporation, Potomac, MD, USA, 1999

Heiner, C., Baker, R., Yacef, K.: Preface. In: Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), Jhongli, Taiwan. 2006

Hung, M. C., Wu, J., Chang, J.H., Yang, D. L., 2005. An Efficient K-Means Clustering Algorithm Using Simple Partitioning. Journal of Information Science and Engineering 21, 1157-1177, 2005

Jung, Y.; Park, H.; Du, D.Z.; Drake, B. (2003) A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering, Journal of Global Optimization 25: 91–111, Kluwer Academic Publishers 2003

Kalathur S. An Object-Oriented Framework for Predicting Student Competency Level in an Incoming Class, Proceedings of SERP'06 Las Vegas , 2006, pp. 179-183

Luan Jing, Data Mining Applications in Higher Education, SPSS Exec. Report, 2004. http://www.spss.com/home_page/wp2.htm

Maulik, U., Bandyopadhyay, S., 2002. Performance Evaluation of Some Clustering Algorithms and Validity Indices, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 12, December 2002

Minaei-Bidgoli B., Punch W.F., Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System, GECCO 2003 Conference, Springer-Verlag, Vol 2, Chicago, USA; July 2003. pp. 2252-2263.

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. pp.15-22.

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. pp.959 – 963.

Pimentel E.P., 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. pp. 56-60

Ravi S., Kim J., Shaw E., Mining On-line Discussions: Assessing Technical Quality for Student Scaffolding and Classifying Messages for Participation Profiling, Workshop of Educational Data Mining, Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education. Marina del Rey, CA. USA. July 2007, pp. 70-79

Rodrigues, J.P.C., Barrulas, M. J. (2003), Towards Web-Based Information and Knowledge Management in Higher Education Institutions, Lecture Notes in Computer Science, Volume 2720, Sep 2003, pp. 188-197

Romero C., Ventura S., Espejo P. and Hervas C., Data Mining Algorithms to Classify Students, Proceedings of Educational Data Mining 2008, The 1st International Conference on Educational Data Mining Montreal, Quebec, Canada, June 20-21, 2008 pp. 8-17

Rupnik R., Kukar, M., Bajec M., Krisper, M., DMDSS: Data mining based decision support system to integrate data mining and decision support, 28th International Conference Information Technology Interfaces, ITI 2006, Cavtat, Croatia, June 2006, pp.225-230

Rusu, L., Breşfelean, V.P., Management prototype for universities. Annals of the Tiberiu Popoviciu Seminar, Supplement: International Workshop in Collaborative Systems, Volume 4, 2006, Mediamira Science Publisher, Cluj-Napoca, Romania, pp. 287-295

Universitatea Babes-Bolyai Cluj-Napoca, Romania. Programul Strategic al Universitatii Babes-Bolyai (2007-2011), Nr.11.366; 1 august 2006.

Vandamme J.P., Meskens N., Superby J.F., Predicting Academic Performance by Data Mining Methods, Education Economics, Volume 15, Issue 4 December 2007 , pp. 405 - 419

Witten I.H., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann series in data management systems, Elsevier Inc., 2005.