Continuing education in a future EU member. Analysis and correlations using clustering techniques

VASILE PAUL BREŞFELEAN
Babeş-Bolyai University
Faculty of Economics and Business Administration,
Mihail Kogălniceanu 1, 400084
Cluj-Napoca,ROMÂNIA

MIHAELA BREŞFELEAN
Babeş-Bolyai University
Faculty of Economics and Business Administration,
Mihail Kogălniceanu 1, 400084
Cluj-Napoca,ROMÂNIA

NICOLAE GHIŞOIU
Babeş-Bolyai University
Faculty of Economics and Business Administration,
Mihail Kogălniceanu 1, 400084
Cluj-Napoca,ROMÂNIA

CĂLIN-ADRIAN COMES
Petru Maior University
Nicolae Iorga 1, 540088
Târgu-Mureş, ROMÂNIA

Abstract: - The purpose of the present paper is to highlight the bonds between the university studies, professional refinement, and master degree studies in the premises imposed by our country's integration in EU's structures. The questionnaire was the instrument used in a collaborative fashion circumscribed to the research project whose member we are. The survey was directed to a number of senior undergraduate students and master degree students at the Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca, and the resulting data was processed using data mining clustering techniques through Weka workbench, graphical and percentage representations.

Key-Words: - clustering, cluster, K-means, questionnaire, correlation, analysis

1 Introduction

Clustering is a practice in which a set of data is substituted by clusters, which represent collections of data points that belong together, its success often being measured subjectively in terms of how useful the result appears to be to a human user [15]. It is related to the process of grouping patterns so that the patterns are similar within each group and distant among different groups [9]. The distribution of groups can be defined as a cluster pattern which is valid if clusters cannot reasonably happen by chance or as a beneficial artifact of a clustering algorithm [8]. An optimal cluster configuration is considered as an effect of all potential combinations of groupings, which presents a set of the most "meaningful" associations [9].

This process is often followed by a phase in which a decision tree or rule set is inferred that allocates each instance to the cluster in which it belongs [15]. Then, the clustering operation is just one step on the way to a structural description. Clustering methods apply when there is no class to be predicted but rather when the instances are to be divided into natural groups [16].

Clustering algorithms have been applied to a broad variety of problems, including exploratory data analysis, image segmentation, data mining, and information retrieval [4].

The main purpose of this paper is to emphasize the connections between the university studies, professional refinement, and master degree studies on the basis required by our country's integration in the European Union's structures. The questionnaire was the instrument used in a collaborative fashion, circumscribed to the research projects whose member we are, and the resulting data was processed using clustering techniques through Weka workbench, graphical and percentage representations.

2 The K-means clustering algorithm

In our research we used the clustering method called FarthestFirst which implements the transversal algorithm of Hochbaum and Shmoys, quoted by Sanjoy Dasgupta [15], a simple, fast, approximation method based on K-means algorithm.

The classic clustering technique called K-means is widely used in the domain of partitional clustering.

The algorithm is known to be dependent on its initialization: a poor set of initial positions for the means will cause convergence to a poor final clustering [12].

The K-means algorithm [11] is an iterative method that evolves K crisp, compact, and hyperspheroidal clusters in the data such that a measure

$$J = \sum_{j=1}^{n} \sum_{k=1}^{K} u_{kj} \|x_j - z_k\|^2 \quad (1)$$

is minimized.

Frequently used for large problems, the K-means algorithm integrates the next hypothesis [7]:
1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster's centroid than to any other cluster's centroid.

The K cluster centers are initialized to K randomly chosen points from the data, which is being partitioned based on the minimum squared distance criterion [11]. The cluster centers are then updated to the mean of the points belonging to them. This entire process is repeated until either the cluster centers do not alter or there is no major change in the J values of two successive iterations. At this point, the clusters are stable and the clustering process ends. The K-means algorithm gives good results only when the initial partitioning is close to the optimal solution [7].

3 Analysis and correlations on questionnaires' collected data

Questionnaires are a technical structure to gather data from a potentially large number of respondents [5]. Often they are the only feasible way to reach a number of reviewers large enough to allow statistically analysis of the results. A well-designed questionnaire that is used effectively can gather information on both the overall performance of the test system as well as information on specific components of the system. A questionnaire should be viewed as a multi-stage process beginning with definition of the aspects to be examined and ending with interpretation of the results.

The steps required to design and administer a questionnaire include [5]: defining the objectives of the survey; determining the sampling group; writing the questionnaire; administering the questionnaire; interpretation of the results.

3.1 The first questionnaire

The first questionnaire (Chestionar IV) collected data from senior undergraduate students from all the departments of The Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca, aiming to evaluate the motivation in choosing the specialization, and the satisfactions regarding the educational process and cognitive skills. From all the questions presented in Chestionar IV were extracted the most adequate for data mining studies, in order to acquire a model to predict students' behavior and to evaluate their motivation in continuing their education with post university studies (master degree, Ph.D. studies). The collected data was drawn off in Excel worksheets, resulting 400 articles with 35 attributes. The Weka workbench, a collection of machine learning algorithms and data preprocessing tools, was used to analyze the data.

Weka's native data storage method is ARFF format which consists of a list of the instances, and the attribute values for each instance are separated by commas. We exported the data into a file in comma-separated value (CSV) format as a list of records with commas between items, loaded the file into a text editor or word processor; added the dataset's name using the @relation tag, the attribute information using @attribute, and a @data line; saved the file as raw text and changed its extension to ARFF.

3.1.1 Clustering the data

To extract suggestive knowledge from the collected data we utilized the data mining process named clustering.

We applied the clustering method FarthestFirst based on K-means algorithm which requires several iterations, each involving finding the distance of k cluster centers from every instance to determine its cluster.

First, we specified the number of clusters to be sought: this is the parameter k. In our case, the k parameter is 3, corresponding to students' 3 choices in continuing their post university studies: disagree, neutral, agree (dezacord, neutru, acord). 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, or mean, 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.

By using the clustering process we divided 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. This process will help the higher education institution to elaborate the most efficient strategies for individuals [14], [2] without the need to deal with each individual student. Segmentation on similar behaviors is considered to be a cost-benefit concession with a good investment-result report.

We divided the students into 3 groups:
Group 0: Students agree to continue their post university studies (master degree, Ph.D. studies);
Group 1: Students do not agree to continue their post university studies;
Group 1: 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:

Cluster 0 <--- Acord																		
M	IE	altul	Acord	Neutru	Acord	Acord	Acord	Acord	Acord	Neutru	Acord	bun	bun	bun	bun	nu	fmult	int
sex	sectie	liceu	Asteptari	Cunostinte	Calitate_ma	Programa_r	Dotare	Practica	Particip_cer	Recomanda	Anul_1	Anul_2	Anul_3	Anul_4	Job_Actual	Parinti_susti	Job_viitor	Restante
Cluster 1 <---Dezacord																		
f	Mk	agr	Dezac	Dezac	Dezac	Dezac	Neutru	Dezac	Dezac	Dezac	fprost	fprost	fprost	fprost	parT	deloc	fara	3_4
sex	sectie	liceu	Asteptari	Cunostinte	Calitate_m	Programa_r	Dotare	Practica	Particip_ce	Recomanda	Anul_1	Anul_2	Anul_3	Anul_4	Job_Actual	Parinti_sust	Job_viitor	Restante
Cluster 2 <--- Neutru																		
f	Mag	econ	Neutru	Acord	Acord	Neutru	Acord	Dezac	Neutru	Neutru	mediu	mediu	excel	excel	parT	mult	nu	1_2
sex	sectie	liceu	Asteptari	Cunostinte	Calitate_m	Programa_r	Dotare	Practica	Particip_ce	Recomanda	Anul_1	Anul_2	Anul_3	Anul_4	Job_Actual	Parinti_sust	Job_viitor	Restante

Table 1. The 3 clusters resulted from the FarthestFirst algorithm

Weka workbench automatically validates the model; following this validation, 27.4151 % of the instances were incorrectly clustered (optimistic result).

This paper focuses on the IE department (Informatica Economica – Business Information Systems) students, belonging to Cluster 0. The Cluster 0 students are characterized by the following choices:
-they agree to continue their post university studies;
-gender: male;
-belong to IE specialization;
-graduated a different high school (mostly a Informatics high school);
-agree their expectations regarding the specialization are fulfilled;
-are neutral regarding the fundamental knowledge they obtained;
-agree they were given sufficient books, course materials, case studies of the highest quality;
-agree the curricula were relaxed and gave time to individual studying;
-agree the faculty has a good quality endowment;
-agree to have made contact to specialization's real problems, in curricula's practical activities;
-neutral regarding their participation to grants/research contracts;
-agree to recommend the specialization to future students;
-have a good opinion of courses teaching methods in the years of study;
-do not have a job;
-benefit a great deal of parents' material support;
-believe to find a job in Romania in IE specialization;
-passed all exams at the end of last academic year.

3.1.2 Graphical representation of clusters

To obtain a graphical representation (Fig. 1) on the clusters, we chose 2 of the most significant attributes (*Programa_relax* – opinion on relaxed curricula, and *Anul_4* – opinion on the 4th year of study).

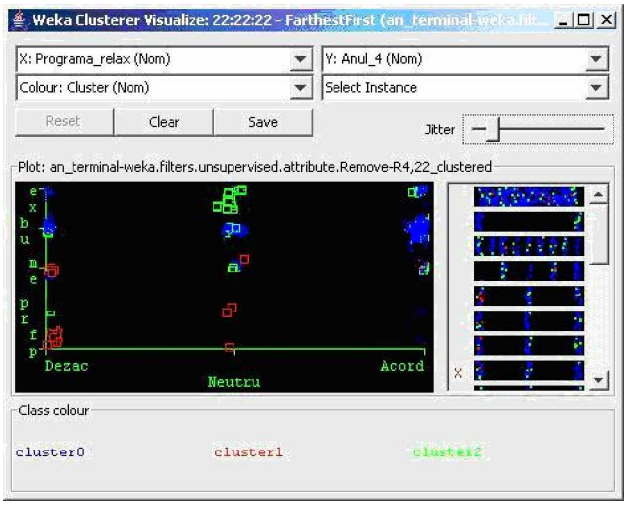

Fig. 1. Cluster graphical representation -dependent upon *Programa_relax* and *Anul_4* attibutes.

We noticed the grouping of the in the down-left corner of the screen, representing Cluster 1, with attribute Programa_relax having *Dezac (do not agree)* as label, and attribute Anul_4 with *fprost (very bad)* label.

In the top-center part of the screen there is represented Cluster 2, with attribute Programa_relax having *Neutru (neutral)* as its label, and attribute Anul_4 with *excel (excellent)* label.

In the top-right corner we detect Cluster 0, with attribute Programa_relax having *Acord (agree)* as its label, and attribute Anul_4 with *bun (good)* label.

There are some areas in which clusters engage over each other and are not so well defined for some middle values of axe Y (*bun* –good, *prost* –bad) due to the large number of data, even though they were filtered and clustered through FarthestFirst algorithm.

3.2 The master degree questionnaires

To fundament the decisions regarding the managerial strategies the faculty leaders can approach in order to fulfill all students' expectations, 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 master degree questionnaires, we made the following correlations and analysis, divided in three categories:

- 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 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 IE master specialization.

Cathegories	**Number of students**
Total IE master degree students	11
Total IE specialization graduates	10
Total other than IE graduates	1
Job in other areas than the graduated specialization	4
Similar job to the graduated specialization	5
Job in other areas than the master degree specialization	5
Similar job to the master degree specialization	4
Unemployed IE master degree students	2

Table 2. IE master degree students

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

Total IE specialization graduates(%)	**Total other than IE graduates(%)**
90,9%	9,09%

Table 3. Percentage relation between graduated specialization and the master degree specialization

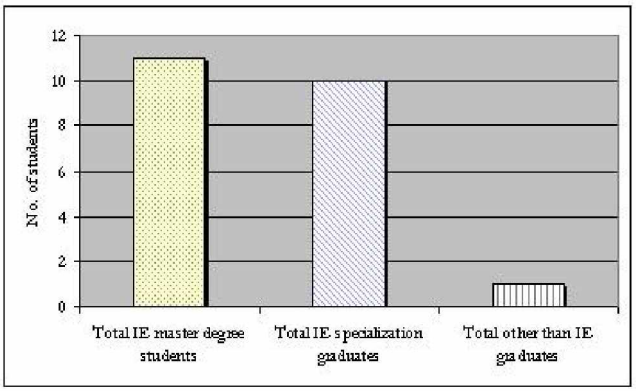

Fig. 2. Correlation between the graduated specialization and IE master degree specialization

Job in other area than the graduated specialization (%)	Similar job to the graduated specialization (%)	Unemployed IE master degree students (%)
36,36%	45,45%	18,18%

Table 4. Percentage relation between the current job and the graduated specialization

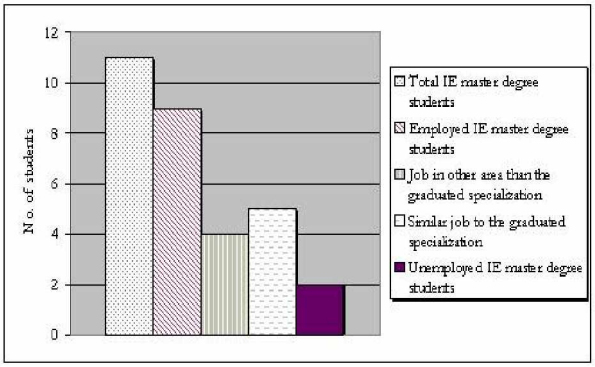

Fig. 3. Correlation between the current job and the graduated specialization

Job in other areas than the master degree specialization (%)	Similar job to the master degree specialization (%)	Unemployed IE master degree students (%)
45,45%	36,36%	18,18%

Table 5. Percentage relation between the current job and the master degree specialization

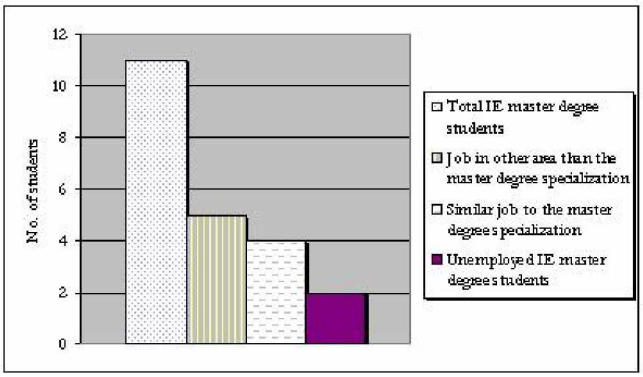

Fig. 4. Correlation between the current job and the master degree specialization

4 Conclusion

From the data analysis, correlation and percentage relations presented in this study, we can conclude that:

- The majority of the undergraduate IE students are keen on continuing their education with master degree studies;
- The majority of IE master degree students (approximate 90,9%) are formed by former IE graduate students, and only a small percent of other than IE graduates (aprox.9,09%);
- 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 have occupation similar to the master specialization.
- A small percent (18,18%) of the IE master degree students are 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.

Acknowledgments*:* This paper was partially supported by the CNCSIS Consortium Grant 8/2005, "Collaborative Information Systems in the Global Economy" and by the Babeş-Bolyai University Priority Themes Grant 2/2005, "Collaborative Decision Support Systems in Academic Environments".

References:

[1] Bodea, C.; Bodea, V.; Tudor, C.A. (2006) Data mining in higher education, *The 3rd*

International Workshop IE&SI, Timişoara, 26-27 May 2006, Editura Mirton

[2] Breşfelean, V.P. (2006) Development Strategies for The Universities' Management Using Information And Communication Technologies, *InfoBUSINESS'2006 The International Conference on Business Information Systems*, "Alexandru Ioan Cuza" University of Iaşi, Iaşi, Romania, 2006

[3] Enăchescu, D. (2003) *Tehnici statistice de Data Mining*, Ed. Univ. Bucureşti, 2003

[4] Folino, G.; Forestiero, A.; Spezzano, G., Decentralized Clustering through a Swarm of Autonomous Agents, *WSEAS Transactions on Systems*, Issue 2, Volume 3, April 2004

[5] Georgia Tech, The College of Computing, Atlanta, Georgia, *Questionnaire Design*, http://www-static.cc.gatech.edu/ classes/cs6751_97_winter/Topics/quest-design/

[6] Ghişoiu, N. ; Breşfelean, V.P. ; Faur,G. ; Vereş, O. (2006) Collaborative Software Systems, *The 3rd international Workshop IE&SI*, 26-27 Mai 2006 Timişoara, Editura Mirton, 2006

[7] Hourani, M.; Ozden, M.; Moore, F.; Maynard-Zhang, P. Genetic Algorithm Application to Clustering Problems, *WSEAS Transactions on Systems*, Issue 3, Volume 3, May 2004

[8] Jain, A.K.; Murty, M.N. ; Flynn, P.J. (1999), Data clustering: a review, *ACM Computing Surveys 31*, 1999

[9] 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

[10] Lipai, A. (2003) Finding Costumer Profile using Data Mining, *The Sixth International Conference On Economic Informatics*, Academy of Economic Studies, Bucharest, Romania, May 8-11, 2003

[11] 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

[12] Pena, M. ; Fyfe, C. (2005) Tight Clusters and Smooth Manifolds with the Harmonic Topographic Map, *Proceedings of the 5th WSEAS Int. Conf. on Simulation, Modeling and Optimization*, Corfu, Greece, August 17-19, 2005

[13] Ullman, J.D., Stanford University. *Data Mining (Note de curs)*, Carti.ss.pub.ro - Biblioteca ta electronică, info.cs.pub.ro/books/ullman/mining/

[14] Universitatea Babeş-Bolyai Cluj-Napoca, România (2003), *Plan Strategic de dezvoltare a Universităţii Babeş-Bolyai pentru perioada 2004-2007*, Cluj-Napoca 2003

[15] Witten, I. H.; Eibe, F. (2005) *Data mining : practical machine learning tools and techniques*, 2nd ed., Morgan Kaufmann series in data management systems, Elsevier Inc., 2005

[16] Witten, I. H. (2004) *Text mining*. In M. P. Singh, editor, Practical handbook of Internet computing. Boca Raton, FL: CRC Press, 2004