Journal of Engineering Science and Technology Review
Special Issue on Telecommunications, Informatics, Energy and Management 2019

Conference Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Knowledge Mining from Student Data

Dimitrios Kazolis[1] and Ioannis Gerontidis[2]

[1]*Dept. of Physics, International Hellenic University, Kavala, Greece.*
[2]*Dept. of Management Science and Technology, International Hellenic University, Kavala, Greece.*

Received 29 September 2019; Accepted 27 February 2020

Abstract

The purpose of this work is to mine knowledge from student data using unattended learning methods. The data were obtained from the Integrated Postsecondary Education Data System (IPEDS) of U.S.A. The methods used are: (a) Factor analysis to reduce the dimension of the problem to a smaller number of derived factors and (b) the k-means clustering algorithm to divide the data into a specific set of clusters. The above methodologies are implemented using the Statistica Data Miner software.

Keywords: Knowledge Mining, Machine Learning, Factor Analysis, k-Means Clustering, IPEDS Database.

1. Introduction

The main objective of all educational institutions is to provide qualitative knowledge and specialization to their students. An important factor in achieving this is not only the full knowledge of the whole educational process but the evaluation of the current and at the same time anticipation of future processes. This will lead to constructive changes and modifications that are necessary in modern society, since the knowledge that is offered must follow, but also determine the evolution. For this reason, in recent years, there has been increased interest in the use of knowledge-extraction techniques from databases containing educational data (Educational Data Mining), [1].

The pursuit of this discipline is, to reveal useful for the educational process conclusions [2], [3]. Thus, for example, there are papers focusing on the extraction of knowledge from student behavior in the final examinations of the semester [4], or from the behavior of a selected sample of students from different colleges [5].

Conclusions from student samples were also documented using linear regression analysis [6], as well as using samples obtained by clustering techniques [7]. In general, many researchers have been engaged with the creation of groups with a high degree of uniformity, [8], [9], [10], [11], while others have concluded that better prediction is given by tree decision models [12]. Also, there are fuzzy logic approaches [13], as well as association rules [14].

In general, all data mining techniques have been mixed in the process of searching for and solving the various problems of the educational process [5].

The present work has the objective to contribute on the previous research to the development of the specific scientific field. The rationale behind this effort is the ability to exploit information contained in the large databases of Educational Data, with the extraction of knowledge.

The procedure described below implements in practice all the steps of extracting knowledge from a database, [15]. Initially, 159 variables were selected that feature 2238 educational institutions in the United States of America. These data were transformed to fit and then fed the process of factor analysis. From this, 5 factors were emerged which led the clustering algorithm in dividing the total sample into 50 groups. The resulting clusters also demonstrated the validity of the procedures followed as they contained human-understood relationships, translated into knowledge. The main conclusion of this paper is that the two methods used can work together perfectly in the process of discovery of knowledge. Their operation is perfectly consistent with the theoretical background that supports them, so they can be applied for further research.

2. The proposed method

2.1. Selecting and Editing Data

The Integrated Postsecondary Education Data System, IPEDS, (http://nces.ed.gov/ipeds/datacenter), includes all US educational institutions. Because the number of these is very large (more than 7500 schools), it was considered appropriate to select part of them using certain criteria, a process provided by the web platform itself.

The institutions were searched by groups according to the years of study, geographical area, sector, diploma etc, characteristics. These search criteria resulted a total number 2238 selected institutions.

Then the process is led to the second step, the selection

* E-mail address: dkazolis2@gmail.com
ISSN: 1791-2377 © 2020 School of Science, IHU.

of the variables. Data related to the characteristics of institutions, such as, records, prerequisites, completion study rates, financial data, facilities for special groups, and much more, were retrieved and saved in CSV format. Although files of this format can be imported directly into Statistical software, for the better implementation of the clustering algorithm, it was considered appropriate to further process the data. This was made by means of MS Office Excel functions in the sense that in addition to the new names given, the data format was changed, since most of them had text formatting. In total, 159 variables were used that feature 2238 educational institutions.

2.2 Factor Analysis Process

The result of the above actions is the creation of a table with dimensions of 2238 rows and 159 columns, with a total of 355,842 entries. Thus, it is reasonable to have a great difficulty in identifying the relationships that link these 159 variables. Furthermore, the table dimension has a definite effect on both the performance and the speed of the implemented clustering algorithm. So, it is necessary to use a methodology to reduce the number of variables available. As such, factor analysis was used.

From the Scree-plot criterion, the maximum number of the factors chosen, and checking if the Kaiser criterion is met, is 5. The next step in the process is to obtain the factor loadings that essentially correspond to the extent that each factor interprets each variable. A load is considered significant when it is greater than 0.3. The final stage of the factor analysis is the calculation of the factor scores that is, the calibration of each object on each factor.

The file obtained is now ready for the next processing stage, that of clustering. This will follow after the interpretation of the 5 obtained factors.

2.2.1 Interpretation of the Factors

The interpretation of the derived factors is based on their relationship with the loadings of the various variables. The larger these loadings are, the more the factors interpret the respective variables. So:

The first factor interprets the variables related to the bachelor's degree. For this it is necessary to have a secondary education degree. It also expresses, to a large extent special educational opportunities offered, as well as the studies abroad.

The second factor interprets the variables related to the Master degree, both the opportunities offered and the award rates of these degrees. It also interprets the baccalaureate, as well as part of the Ph.D. degrees.

The third factor interprets the variables related to the two-year education degree (associate degree). It also reflects the adult education as well as the professional orientation of the institutions.

The fourth factor reflects the degree of difficulty or, more generally, the entry requirements in institutions. It also includes the admission rates according to the student's gender.

Finally, the fifth factor reflects the degree of equal opportunity for admission to higher education institutions, as well as the additional benefits that may be available.

At the end of the factor analysis, the 144 columns of the original data were reduced to 5 factor columns thus contributing to the drastic reduction of the data dimensionality. This facilitates the following process.

2.3 Cluster Analysis Process

After extracting the key features of the data selected, the process can continue to the next step, i.e. to find clusters of institutions with common features.

Here is the use of Generalized k-Means Cluster Analysis, one that can handle very well categorical variables. In these, the highest frequency category becomes the center of the corresponding cluster, and all distances can have zero and one.

In the k-Means Clustering process, the number of k cluster centers is defined by the user before the algorithm is applied [3]. In our case the number of clusters are chosen to be k=50. The selection was made on the basis of the number of faculties that will occur in each group. Also, the number of iterations for determining group centers is set to a large number, e.g. 500, so as to avoid interrupting the process from this selection before properly completing it. Also, the Manhattan method is defined to measure the distances of non-uniformity. This method is simply the average of the difference between the elements. This is done to determine the distances in order to avoid influencing the analysis from different gradient variables, and in particular extreme values.

From the above procedure a new file is created, where each institution is grouped into a cluster and its distance from the center is recorded. The histogram in "Figure 1", shows the number of members in each cluster.

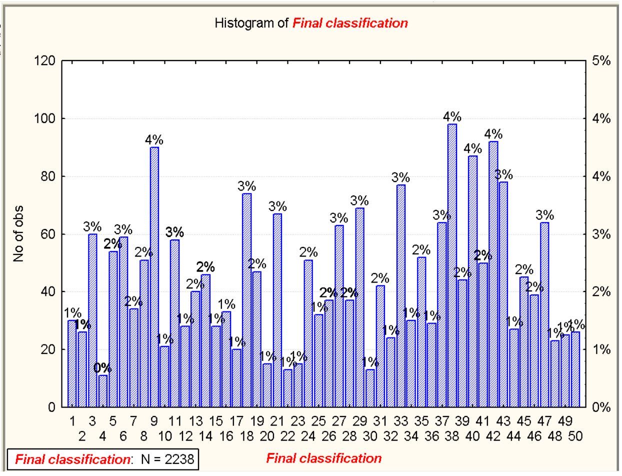

Fig 1. Histogram of institutions by cluster

At this point the whole process has been completed and can be followed by the examination of the uniformity of the clusters. It should be emphasized however, that the k-Means algorithm is sensitive to the selection of the random initial k centers. As a result, whenever the clustering process takes place from the beginning, although it involves the same manipulations and data, it will give similar but not exactly the same results. This is also one of the characteristics of unsupervised learning that is being considered.

3. Interpretation of Results and Validity Control

After the completion of the analysis, it is reasonable to check the created clusters. This will prove the validity of the employed method and will show its ability to extract knowledge. It is worth mentioning however that during the interpretation stage some practical restrictions stemmed mainly from the lack of pre-existing information on the institutions under consideration.

We proceed by examining the clusters created, in order to validate the reliability of the overall process. Initially, cluster 5 is selected (see "Fig. 2"). The 5th cluster is chosen because it includes a well-known institution, the Harvard University. As expected, this university is associated with 53 other major institutions, all of which have a very high research activity. Most are located in large cities and have a large number of graduates in both the academic and professional fields. These institutions offer of full, four-year course and the students must attend in the university campus and therefore, are not suitable for distance learning or for a short-term degree. They are also private institutions, they can award top-level degrees, do not have religious restrictions, and do not apply an open import policy.

Instead, they have a number of requirements for accepting them. For example, high degrees of secondary education, preparatory programs, knowledge of English, letters of recommendation, and performance in entrance examinations. They are generally selective and are not suitable for specific groups of students, such as for adults and for part-time students.

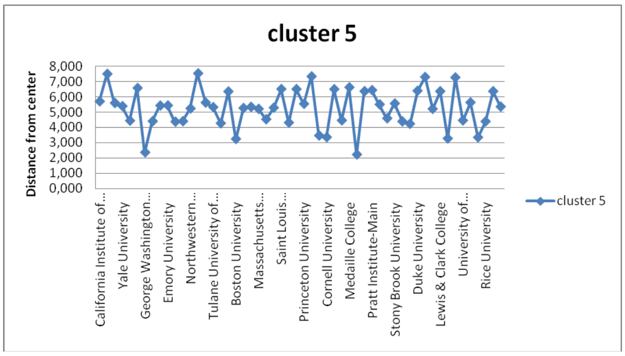

Fig 2. The number 5 cluster.

Next cluster 32, (see "Fig. 3") was then selected. The choice here was made because, at first glance, cluster 32 had the same characteristics as the previous cluster. Therefore, we need to find out why the clustering algorithm differentiated them.

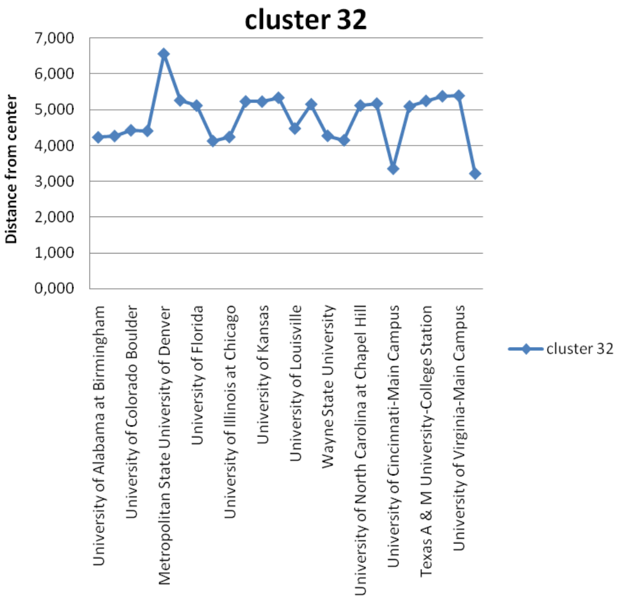

Fig 3. The number 32 cluster.

This cluster consists of 24 institutions. They also have very high research activity they are located in large cities and have a high percentage of graduates, both in the academic and professional fields. They can provide top-level degrees, do not have religious restrictions, and do not apply an open import policy. They require a secondary education degree, entrance exam, English language knowledge and preparatory courses.

In contrast to cluster 5, they are public foundations, and they are not so selective in accepting students, e.g. constituent letters are not considered necessary. Moreover, at least during the first period of study, they do not require as much presence of attendance, while they provide many ancillary services to their students. As an example, we will mention the variable "servCh", which states that they provide childcare services for their students during the day.

Also, within this group, an outlier can be observed, which is: Metropolitan State University of Denver. This is also evidenced by the distance 6,571 of this case from the cluster center, which is the highest value among the distances of other cluster members. Probably this institution has been placed here because it has similarities with other members at all points except that it belongs to baccalaureate colleges and provides education to the postgraduate degree. It is generally concluded that this cluster of institutions, if we exclude the fact that they provide substantial full-time courses, covers a large part of the requirements of specific student groups.

A similar examination in the rest of the clusters created, it is noted that their members, have many common features among them, while different clusters have several differences so that the algorithm can separates them.

4. Conclusions - Proposals

From the above analysis, we first conclude that factor analysis is a fundamental and effective statistical method. With it, the processing difficulties stemming from very large dimensional data can be treated with relative ease. The factors created can accurately replace the majority of initial variables and reveal their essential meaning. Factor analysis, therefore, is a useful and reliable method, the use of which, in such cases, is considered necessary.

Secondly, the k-Means clustering algorithm has emerged as a very effective method of classifying data, especially in cases where learning examples are not feasible. The above conclusion results from the findings given in its present application, where logical relationships could be extracted and consequently the knowledge from data sets that were initially unintelligible.

The overall conclusion of this paper is that the two methods used can perfectly cooperate in the process of knowledge discovery. Their behavior consistently follows the theoretical background that supports them and they can be applied in a predictable way for further research.

As a future work continuation of this effort, one could investigate the behavior of a model in which iterative algorithms would be applied in order to more accurately determine knowledge, especially during clustering with the V-fold cross validation method, where very large clusters of institutions are created.

Another idea would be to combine different methodologies to always improve accuracy, correctness and understanding for the interpretation of the conclusions. Thus, several ways of analysis can be combined with different knowledge mining algorithms in common data to discover the most efficient combination of these.

A topic with many future extensions could be choosing different types of variables and applying them to the same model to investigate whether the particular model is sensitive to the types of variables in question. Also, the analysis of a specific set of variables with different methods, with the purpose of benchmarking and drawing conclusions for each method.

References

1. C. Romero and S. Ventura (2007). Educational data mining. A survey from 1995 to 2005. *Expert Systems with Applications*, vol. 33, pp. 135-146.
2. J. Han and M.Kamber (2001). *Data Mining: Concepts & Techniques*. Morgan Kaufmann Publishers, San Francisco USA.
3. F. Trainer (2008). The role of institutional research in conducting comparative analysis of peers. *New Directions for Higher Education,* vol.141, pp. 21-30.
4. K. Brijesh and P. Saurabh (2011). Mining Educational Data to Analyze Students Performance. *International Journal of Advanced Computer Science and Applications,* vol. 2, pp. 63-69.
5. F. Castro, A. Vellido, A. Nebot and F.Mugica (2007). Applying Data Mining Techniques to e-Learning Problems. *Evolution of Teaching and Learning Paradigms in Intelligent Environment.* vol. 62, pp. 183-221.
6. S. Hijazi and R. Naqvi (2006). Factors affecting students performance: A Case of Private Colleges. *Bangladesh e-Journal of Sociology*, vol. 3, no. 1.
7. Z. Khan (2005). Scholastic achievement of higher secondary students in science stream. *Journal of Social Sciences*. vol. 1, pp. 84-87.
8. S. Ayesha, T. Mustafa, A. Sattar and M. Khan (2010). Data mining model for higher education system. *Europen Journal of Scientific Research*, vol.43, pp.24-29.
9. S. Erdogan and M.Timor. (2005). A data mining application in a student database. *Journal of aeronautics and space technologies*, vol. 2, pp. 53-57.
10. G. Hurley (2002). Identification and assessment of community college peer institution selection systems. *Community College Review*. vol.29, Issue 4, pp. 1.
11. A.J. Medwick (2007). Characteristics of Student Populations In Schools and Universities of the United States, in: Nisbet, R., Elder, J. and Miner, G. (2007). *Handbook of Statistical Analysis and Data Mining Applications.* Academic Press, Canada.
12. Q. AI-Radaideh, E. AI-Shawakfa and M. AI-Najjar, (2006). Mining student data using decision trees. *International Arab Conference on Information Technology*. Yarmouk University, Jordan.
13. G. Hwang, T.Huang and C. Tseng. (2004). A Group-Decision Approach for Evaluating Educational Web Sites. *Computers & Education*. vol. 42, pp. 65-86.
14. B. Dogan and A. Camurcu (2008). Association Rule Mining from an Intelligent Tutor. *Journal of Educational Technology Systems*, vol. 36, pp 433 – 447.
15. U. Fayyad, G. Piatetsky-Shapiro and P.Smyth (1996). From data mining to knowledge discovery in databases, *AI Magazine,* vol. 17, No. 3, pp. 37-54.
16. I. Blaxavas, P. Kefalas, N. Basiliadis, F. Kokoras and I. Sakelariou (2006). *Artificial Intelligence. 3rd Edition, Giourdas Publishing, Thessaloniki.*