IOSR Journal of Research & Method in Education (IOSR-JRME)
e- ISSN: 2320–7388,p-ISSN: 2320–737X Volume 9, Issue 1 Ser. IV. (Jan. – Feb. 2019), PP 61-64
www.iosrjournals.org

The USA Higher Education Institutes Segmentation Using Clustering Technique

Arash Ajam[1]
[1](*Gianforte School of Computing, Montana State University, USA)*
Corresponding Author: Arash Ajam

Abstract: *The huge quantities of data and information about universities, students, faculty members, personnel, material resources, etc., can in most cases contains valuable information and patterns for higher education. American higher education institutions are usually classified according to indicators such cost including tuition fees, living and educational expenses. Also, the percentage of doctoral students and graduates of each faculty, reflecting the scientific level of the faculty of each university and the availability of facilities such as laboratories, libraries and research centers, can provide significant insights and patterns. In the present research, cluster analysis and K-Means techniques are used to categorize and analyze data related to te mentioned indicators.*

Keywords: *Data mining, Clustering, K-means, Patter Recognition*

Date of Submission: 22-02-2019 Date of acceptance: 08-03-2019

I. Introduction

As one of the most dynamic educational systems, the higher education system is responsible for the most important roles, and because of dealing with important data about students, professors, personnel and material resources, etc. is one of the areas where data mining is most used. The present study aims to achieve useful results using the data from US higher education institutions and data mining techniques.

II. Literature Review

Data mining or discovery in the databases refers to non-obvious extraction of potentially useful information from data that has previously been unknown and Data Mining is the process of extracting hidden knowledge from large volumes of raw data. [1, 2]. Pandey et al. proposed the performance of clustering algorithm using heart disease dataset. They evaluated the performance and prediction accuracy of some clustering algorithms. The performance of clusters will be calculated using the mode of classes to clusters evaluation. Finally they proposed Make Density Based Cluster with the prediction accuracy of 85.8086%, as the most versatile algorithm for heart disease diagnosis [3]. Finding such patterns and knowledge in data can be useful in making future decisions such as eco-efficiency [4] and other computational method such as Data Envelopment Analysis (DEA) [5], simulation model [6] and Fuzzy expert system [7].
The stages of the data mining process can be observed in Figure 1.

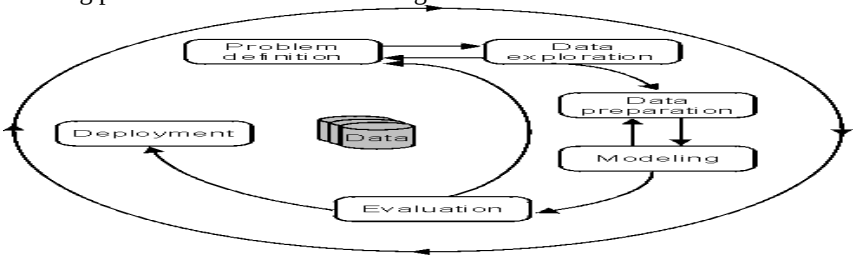

Figure 1 - Data mining process

III. Clustering

Clustering refers to grouping the similar data in a data mass. The basic issue of clustering is the distribution of data to k different groups such that the data of each group being similar and the data of different groups are dissimilar. A good clustering method generates high quality clusters based on the following two criteria:

- high similarity of the internal points of each cluster
- low similarity of the points of different clusters

K-means method is one of the most important methods of clustering algorithm

IV. Data Set

The dataset used in this study has been extracted from the www.ics.uci.edu website which are related to higher education institutions located in the United States until the end of 2007. Table 1 represents the fields of this database.

Table 1 – the database fields

Field name	Field type	Description
University name	Nominal	University name along with state name
State code	Numerical	Each state has a unique code
Public / Private university	Numerical (1/2)	Public / private university
admission application	Numerical	Number of admissions applications at the university
Accepted application	Numerical	Number of applications accepted at university
New entrance	Numerical	Number of new students
Elite entrance	Numerical	Number of elite entrance students
Graduates	Numerical	Number of students on graduation
Tuition	US currency (dollars)	The tuition fee of each university
Number of rooms	Numerical	Number of rooms per university
Tuition	US currency (dollars)	Expensive study fees like dormitory and ...
Living expenses	US currency (dollars)	personal expenses
PhD students	Numerical (percent)	Number of PhD students
College students	Numerical (percent)	Student rate of each faculty
Graduates percent	Numerical (percent)	The graduation rate of each university

V. Preprocessing and Data Preparation

The data preparation, is the most important and time consuming stage in the data mining project. As the data are the input into this project, the more accurate the inputs, the more accurate the output. There are several preprocessing methods introduced in the literature. In [8] several data mining clustering algorithms were evaluated to find the most accurate one in heart disease prophecy and a particular preprocessing filtering method are applied to find the superior algorithms. In [9], normalization technique is used as an effective preprocessing method before training a data-mining model. This technique is very efficient especially for a large dataset [10]. Another strong preprocessing method is called principle components analysis (PCA) which reduces the dimension of the dataset. In [11], PCA is used for cell recognition.

The data was coded in order to prepare data, so that they can be easily used in the Clementine12 software. Considering the numerous fields in this database, we selected some of its fields to be used in the data mining process. Table 2 shows the fields used in data mining and how they are encoded.

Table 2 - The fields used in data mining and coding

Field name (attribute)	Type	Coding	Feature Type
Tuition	Less than $ 5,000 10,000 $ - 5000 $ 15,000 $ - 10,000 $ 20,000 $ - 15,000 $ More than $ 20,000	Less than $ 5000 = 1 10,000 $ - 5000 $=2 15000 $ - 1000 $=3 20,000 $ - 1500 $=4 More than $ 20,000 = $ 5	Numerical
Education secondary costs	Less than $ 1,000 1,500 $ - 1,000 $ 2000 $ - 1500 $ More than $ 2,000	Less than $ 1000 = 1 $ 1500 - $ 1000=2 $ 2000 - $ 1500=3 More than 2000 $ = 4	Numerical
Living expenses	Less than $ 100 2000 $ - 100 $ 2000 $ - 1500 $ More than $ 2,000	Less than $ 100 = 1 $ 2000 - 100 $ $ 2000 - $ 1500=3 More than 2000 $ = 4	Numerical
Number of doctoral students	Less than 25% %50 - %25 %75 - %50 %100 - %75	Less than 25% = 1 %50 - %25=2 %75 - %50=3 %100 - %75=4	Numerical
Graduates percent	Less than 25% %50 - %25 %75 - %50 %100 - %75	Less than 25% = 1 %50 - %25=2 %75 - %50=3 %100 - %75=4	Numerical

VI. Problem Statement

Cluster analysis refers to a group of multivariate techniques whose primary purpose is to cluster objects (respondents, products, and other entities) so that every object of the group is very similar to the other members of the cluster (based on predetermined criteria).

The observations in the final clusters, should have a high homogeneity (within clusters) and external heterogeneity (between clusters). This analysis is also helpful when the researcher intends to test hypotheses about the nature of data or to develop a new hypothesis in this regard. The number of clusters in this analysis can be predefined or determined based on the maximum variance within clusters (or between the clusters).

The variable concept is a fundamental issue in this technique, however it has completely different methods from the other techniques. The clustering variable is a set of variables that show the features used to compare objects in cluster analysis. There are always three main objectives of "describing empirical groups", "data simplification" and "determining relationships" in this technique along with the inclusion of two assumptions on the Representativeness and Multicollinearity between the independent variables in question.

The K-means clustering technique has been used in this study to classify US higher education institutions.

As can be seen from Table 2, the fields, used in the data mining process, were introduced and after applying the K-means technique, five clusters were obtained, which are shown in Table 3 of the features of each cluster.

Table 3. Features of each cluster

Cluster	Number of records	Percent of total
First cluster	259	20%
Second cluster	139	11%
Third cluster	429	33%
Fourth cluster	195	14%
Fifth cluster	280	22%

Considering the clustering mentioned in Table 3, the majority of universities are in the third cluster. Table 4 provides the information on the third cluster with the highest frequency

Table 4. The third cluster features

Cluster	Tuition	Education secondary costs	Living expenses	The number of doctoral students	Graduation Rate	Percent of total
Third cluster	$ 10,000 to $ 150,000	$ 1500 to $ 2000	$ 2000 to $ 4000	50% to 75%	50% to 75%	32%

Accordingly, most of the US higher education institutions have the features noted in Table 4.
Now, the indicators are individually examined within other clusters which results are presented in Table 5 (tuition, education secondary costs, living expenses, doctoral and graduate students' rates).

Table 5 – individual frequency of each indicator

Index	Tuition	Education secondary costs	Living expenses	The number of doctoral students	Graduation Rate
Cluster	Second cluster	Fourth cluster	First cluster	Second cluster	Second cluster
The range with the most frequency	$ 5000 to $ 10,000	More than $ 2,000	Less than $ 100	25% - 50%	25% - 50%

VII. Conclusion

The abundance of higher education institutions in the United States and their extraordinary spread throughout the country has led to a variety of unique indicators such as living expenses, training, university tuition, the number of doctoral students and graduation rates. The present study was conducted with the purpose of segmentation of these institutions using cluster analysis and K-means technique, which resulted in five clusters so that highest frequency was belonged to the third cluster. Also, the indices of tuition, the number of doctoral students and the graduation rate in the second cluster have the most frequencies in the individual review of each field; in terms of education secondary expenses, the fourth cluster has the most frequency, and finally, the living expenses index has the most frequency in the first cluster.

References

[1]. A. Shahrabi, and M. Hazari, (2010), *Application of decision making techniques and cluster analysis to prioritize improvement projects*, Fourth Iranian Data Mining Conference, Tehran, Iran.

[2]. A. Khayrabadi A., and B. Minaei Bidgoli (2010), Investigation and clustering of the results of evaluation of university professors using data mining methods, 4th Iranian Data Mining Conference, Tehran, Iran.

[3]. A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "Data Mining Clustering Techniques in the Prediction of Heart Disease usingAttribute Selection Method," International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2277798, Vol 2, Issue10, October 2013.

[4]. M. Mirmozaffari, (2018), Eco-Efficiency Evaluation in Two-Stage Network Structure: Case Study: Cement Companies, Iranian Journal of Optimization. Volume 11, Issue 2.

[5]. M. Mirmozaffari and A. Alinezhad, "Ranking of heart hospitals using cross-efficiency and two-stage DEA," 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, 2017, pp. 217-222.

[6]. A. Goodini, M. Torabi, M. Goodarzi, R. Safdari, M. Darayi, M. Tavassoli, and M. Shabani, "The simulation model of teleradiology in telemedicine project," The Health Care Manager, Vol. 34- Issue 1, p 69- 78, January/March 2015.

[7]. M. Mirmozaffari, "Developing an Expert System for Diagnosing Liver Diseases", *EJERS*, vol. 4, no. 3, pp. 1-5, Mar. 2019.

[8]. M. Mirmozaffari, A. Alinezhad, and A. Gilanpour, "Heart Disease Prediction with Data Mining Clustering Algorithms," Int'l Journal of Computing, Communications & Instrumentation Engg (IJCCIE), ISSN 2349-1469 EISSN 2349-1477, Vol.4, Issue1, Jan 2017.

[9]. O. A. Gashteroodkhani, M. Majidi, M. Etezadi-Amoli, A. F. Nematollahi, B. Vahidi "A hybrid SVM-TT transform-based method for fault location in hybrid transmission lines with underground cables", Electric Power Systems Research, vol. 170, pp. 205-214, 2019.

[10]. S. Aznavi, P. Fajri, M. Benidris, and B. Falahati,"Hierarchical droop controlled frequency optimization and energy management of a gridconnected microgrid," in 2017 IEEE Conference on Technologies for Sustainability, Phoenix, AZ, USA, 2017, pp.1-7.

[11]. X. Long, W. L. Cleveland, Y. L.Yao "A new preprocessing approach for cell recognition". IEEE Transactions on Information Technology in Biomedicine; 9(3):407-12, Sep 2005.

Arash Ajam. "The USA Higher Education Institutes Segmentation Using Clustering Technique". IOSR Journal of Research & Method in Education (IOSR-JRME) , vol. 9, no. 1, 2019, pp. 61-64.