Study of Clustering of Data Base in Education Sector Using Data Mining

Umamaheswari.R[1] S.Saravana Mahesan[2] Meenakshi.P[3]
[1,2,3]Assistant Professor
[1,2,3]Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College Avadi, Chennai

Abstract— Data mining is a technology used in different disciplines to search for significant relationships among variables in n number of data sets. Data mining is frequently used in all types' areas as well as applications. In this paper the application of data mining is attached with the field of education. The relationship between student's university entrance examination results and their success was studied using cluster analysis and k-means algorithm techniques.

Key words: Data mining, cluster, k-means

I. INTRODUCTION

The amount of data maintained in an electronic format has seen a dramatic increase in recent times. The amount of information doubles every 20 months, and the number of databases is increasing at an even greater rate. The search to determine significant relationships among variables in the data has become a slow and subjective process. In order to give a complete solution to this type of problem, the Knowledge Discovery in Databases – KDD has introduced. The process of the formation of significant models and assessment within KDD is referred to as data mining.

Data mining is used to uncover hidden or unknown information that is not apparent, but potentially useful Cluster analysis is a technique used in data mining. Cluster analysis involves the process of grouping objects with similar characteristics and each group is referred to as a cluster.

Cluster is nothing but grouping of elements together, which is used in various fields, such as marketing, image processing, geographical information systems, biology, and genetics. In this study, university students were grouped according to their characteristics, forming clusters. The clustering process was carried out using a Means algorithm.

II. CLUSTER ANALYSIS

Cluster analysis is a multivariate analysis technique where individuals with similar characteristics are determined and classified (grouped) accordingly .Through cluster analysis, dense and sparse region can be determined and different distribution patterns may be achieved. The concepts of similarities and differences are used in clustering. Different types of measures are used in determining similarities and differences. This study utilizes the Euclidian distance measure.

A. Euclidian Distance Measure:

The Euclidian distance measure is frequently used as a distance measure, and is easy to use in two dimensional planes. As the number of dimensions increases, the calculability time also increases The formula defines data objects i and j with a number of dimension equal to p. The distance between the two data objects d(i,j) is expressed as given in formula. xip: is the measurement of object i in dimension p.

B. Algorithm;

The K-means algorithm is a cluster analysis algorithm used as a partitioning method. K-means is the most widely used used and studied clustering algorithm. Given a set of n data points in real d-dimensional space, RD and an integer k, the problem is to determine a set of k points in real dimensions, called centers, to minimize the mean squared distance from each data point to its nearest center.

The K means algorithm is used to define random cluster centroids according to the initial parameters .Each consecutive case is added to the cluster according to the proximity between the mean value of the case and the cluster centroids. The clusters are then re-analyzed to determine the new centroids point. This procedure is repeated for each data object. The algorithm is composed of the following steps:

1) Place K points into the space represented by the objects that are being clustered. It represents initial group centroids.
2) Assign each object to the group that has the closest centroids.
3) When all objects have been assigned, remanipulate the positions of the K centroids.
4) Repeat Steps 2 and 3 until the centroids no longer move. This produces a splitting of the objects into groups from which the metric to be minimized can be calculated.

III. APPLICATION

In which the data gathered from the students was analyzed using a k-means algorithm cluster analysis technique.

A. Set Of Data:

Information that is collected are the data. Group of data that are collected are called data. Here the student's data are collected.

B. Database:

The data base system was used for two reasons:

- The software used in analysis was compatible and efficient to use with the database management system
- The data to be analyzed was maintained in the database prior to the study.

C. Data Mining Process:

The data exploration and presentation process consisted of various steps. These steps were data preparation, data selection and transformation, data mining and presentation.

D. Preparation Of Data:

In these steps, the data that was maintained in different tables was joined in a single table. The' students' and 'student's log' tables were joined using the Student's ID field as the key field. After the joining process errors in the data were corrected

STUDENT NAME
ROLL NO
DEPARTMNET
COLLEGE
PERCENTAGE

Fig. 1:Student and student detail table

E. Data Selection And Transformation:

After the data preparation, the data selection and transformation process was performed. In this step the fields used in the study were determined and transformed if necessary. For example, the fields in which the responses were yes/no were transformed to 1/0.

F. Data Mining:

The prepared data was then put through the data mining process. The K-means algorithm was used in this step. The number of clusters was determined as an external parameter. Different cluster numbers were tried, and a successful partitioning was achieved with 5 clusters. The cluster centroids are given in table 1.

Cluster	Name	College Id	Faculty Id
1	Paru	1112	145
2	Bhuvan	1123	134
3	Giri	1167	166
4	gayu	1134	189

Table 1: Cluster Centroids

G. Presentation:

The results of the data mining step are presented in this step. For graphs and tables, the Map Toolbox plug-in of the Mat lab software was used. The resulting clusters are shown in figure 2.

1	2	3	4
****	///////	……	?????
****	///////	……	?????
****	///////	……	?????
****	///////	……	?????

Fig. 2: Each number represents different cluster

The figure above shows the University Entrance Exam percentiles of the x axis, and the grades on the y access. The graph shows that the 1st cluster is more successful in regard to grades while the 5th group is the least successful. The distribution of faculties in these two clusters is shown in figures 3 and 4.

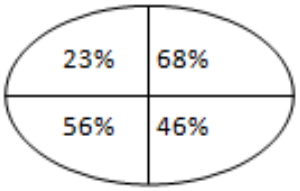

Fig 3: Distribution of Cluster 1

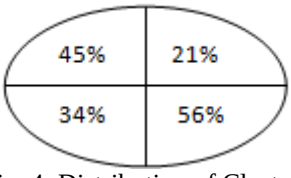

Fig. 4: Distribution of Cluster 4

The majority of the students in cluster 1 are from the Faculty of Arts and Sciences. The reason is that most students in the faculty have high success grades and scholarships. They study hard to keep their scholarship, and therefore have good grades. Cluster 4, however, is mainly made up of Faculty of Communication and Faculty of Business Sciences students. These students have lower grades and lower results in the university entrance exam.

IV. CONCLUSION

This study utilizes data mining in the field of education. Cluster analysis and K-means analysis were used as data mining techniques. The steps of the data mining process were carried out and explained in detail. The application areas used here was education, different from the usual data mining studies. Data mining technique usage in the field of education may provide us with more varied and significant findings, and which may lead to the increase in the quality of education.

REFERENCES

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/ The MIT Press, 1996.
[2] J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
[3] Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.
[4] Sholom M. Weiss and Nitin Indurkhya, "Predictive Data Mining: A Practical Guide", Morgan Kaufmann Publishers, 1998.

