International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 1– No.5, February 2012 – www.ijais.org

Use of Clustering to Improve the Standard of Education System

Shishu Pal Singh
N. K. Sharma
Assistant Professor, Computer Science Department
Institute of Advanced Management & Research
Ghaziabad, Uttar Pradesh, India
Ph.D. Research Scholar – Singhania University
Rajasthan, India

B. K. Sharma
Sr. Scientific Officer & Head
Software Development Centre
Northern India Textile Research Association
(Linked to Ministry of Textiles, Govt. of India)
Ghaziabad, Uttar Pradesh, India

ABSTRACT
This paper deals with the application of Data Mining in the education sector. Generally the benefits of Data Mining are taken in the commercial fields. The study given, proposed a quiet different field where we can use the Data Mining and enhance the quality of education. In the given paper the performance of an institute students were studied. The study takes the performance of students in their examinations and their presence in the classrooms into consideration and finds a relation in them. The observed relation helps in identifying the group of students where the extra are required.

The study was carried out using K – Means method of cluster analysis – a technique of Data Mining for finding the relevant records.

KEYWORDS

Data Mining, K – Means, Cluster, Distance, Education.

1. INTRODUCTION

1.1 Requirement of Data Mining
The amount of data that has to be analyzed and processed for making decisions has significantly increased in the recent years of fast technological development. It has been estimated that every year a million of terabytes of data are generated, a large amount of which is in digital form. This means that more data will be generated in the next three years than in the whole recorded history of humankind. The data is recorded because people believe it to be a source of potentially useful information. This is a common occurrence in all areas of human activity, from collection of everyday data (such as telephone call details, credit card transaction data, governmental statistics, etc.) to more scientific data collection (such as astronomical data, genome data, molecular databases, medical records, etc.). These databases contain potentially useful but as yet undiscovered information and knowledge. The discipline concerned with extracting this information is data mining (Hand et al. 2001, Ye 2003).

Data Mining (DM) includes the concept of all those techniques that permit the extraction of knowledge from a mass of data which would otherwise remain hidden in large database. In the first stage of DM we have pre-processing, in which data are collected, loaded and 'cleaned'. In order to do this successfully, it is necessary to know the database, which involves understanding its data, the cleaning process and preparation data in order to avoid duplication of content as a result, for example, of typing errors, different forms of abbreviation or missing values.

Fig 1: Diagram of data mining technique.

1.2 Definition of Clustering
A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".

A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Cluster analysis is a collection of methods that assists the user in putting different objects from a collection of objects into different groups. In some ways one could say that cluster analysis is best used as an exploratory data analysis exercise when the user has no hypothesis to test. Cluster analysis, therefore can be used to uncover hidden structure which may assist further exploration [1].

Clustering is the process of grouping a set of physical or abstract objects into groups of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters (Han & Kamber, 2001) [2].

The clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items. These distances (similarities) can be based on a single dimension or multiple dimensions, with

each dimension representing a rule or condition for grouping objects. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances. As in the presented paper a multidimensional problem is discussed, so we have utilizes the Euclidian distance measure.

1.3 Distance Methods

The cluster analysis methods are based on measuring similarity between objects by computing the distance between each pair. There a number of methods for computing distance in a multidimensional environment.

Let the distance between two points x and y (both vectors) be D(x,y). We now define two important distance measures.

1.3.1 Euclidean Distance

The Euclidian distance measure is frequently used as a distance measure, and is easy to use in two dimensional planes. It simply is the geometric distance in the multidimensional space. As the number of dimensions increases, the calculability time also increases [2]. It is computed as:

$$\boldsymbol{D(x,y) = \sqrt{\sum (x_i - y_i)^2}}$$

The formula defines data objects x and y with a number of dimension equal to j. The distance between the two data objects d(x, y) is expressed as given in above formula. xij: is the measurement of object x in dimension j [7].

The Euclidean distance is usually computed from raw data, and not from standardized data. This method has certain advantages (e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers). However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed. For example, if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean or squared Euclidean distances (computed from multiple dimensions) can be greatly affected (i.e., biased by those dimensions which have a larger scale), and consequently, the results of cluster analyses may be very different. Generally, it is good practice to transform the dimensions so they have similar scales.

1.3.2 Manhattan Distance

Another commonly used distance metric is the Manhattan distance or the L_1 norm of the difference vector. In most cases, the results obtained by the Manhattan distance are similar to those obtained by using the Euclidean distance. Once again the largest valued attribute can dominate the distance although not as much as in the Euclidean distance [7].

$$\boldsymbol{D(x,y) = \sum_i |x_i - y_i|}$$

2. CLUSTERING ALGORITHM

2.1 K – Means Algorithm

K-means is one of the well-known algorithms for clustering, originally known as Forgy's method (Forgy, 1965), and it has been used extensively in various fields including data mining, statistical data analysis and other business applications. Thus, this study proposes the K-means algorithm to build clusters by attributes (i.e. R–F–M attributes). The K-means algorithm for partitioning is base on the mean value of the objects in the cluster. MacQueen (1967) suggested the term K-means for describing an algorithm that assigns each item to the cluster with the nearest centroid (mean) (Mac- Queen, 1967).

The k-means method will produce exactly k different clusters of greatest possible distinction. It should be mentioned that the best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data.

The program will start with k random clusters, and then move objects between those clusters with the goal to 1) minimize variability within clusters and 2) maximize variability between clusters. In other words, the similarity rules will apply maximally to the members of one cluster and minimally to members belonging to the rest of the clusters. In k-means clustering, the program tries to move objects (e.g., cases) in and out of groups (clusters) to get the most significant results.

Based on the concept above, the computing process for K-means is presented as follows:

Step 1: Partition the items into K initial clusters. Firstly, partition the items (m objects) into K initial clusters.

Step 2: Proceed through the list of items. Assign an item to the cluster whose centroid is nearest (distance is computed by using Euclidean distance with either standardized or un-standardized observations) and re-calculate the centroid for the cluster receiving the new item or for the cluster losing the item.

Step 3: Repeat Step 2 until no more reassigning. Rather than starting with a partition of all items into K preliminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2. The final assignment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points. Experience suggests that most major changes in assignment occur with the first reallocation step.

3. CASE PROBLEM

The study presented here is carried out on the students' data. The data is of their examination performance and their presence in the classrooms. The data is analyzed using the Data Mining clustering technique, so that we can find out some relation in the presence of the students in the classrooms and their performance, so that some solution can be derived to improve their quality.

3.1 Data Set

The data used has been taken from IAMR, an institute of Choudhary Charan Singh University situated in Ghaziabad district of Uttar Pradesh, a state of India. The data were gathered in 2011 and include the records of 485 students of BCA, BBA and Integrated courses.

3.2 Software Tools Used

For data storage purpose Microsoft Excel 2007 is used and for the programming purpose, the Waikato Environment for Knowledge Analysis Version 3.6.5 (c) 1999 – 2011 of the

University of Waikato Hamilton, New Zealand is used. The Weka tool is used for the Data Mining problems. The K – Means algorithm is implemented as one of its function in the tool.

3.3 Data Mining Process

The process of data mining involves the following steps: data preparation, data selection and transformation and then application of data to the clustering algorithm to generate the clusters and then finally presentation of the result.

3.3.1 Data Preparation

This step includes the collection of data required for this study which was done by joining the various tables together using the common attribute. There were three main tables which were taken into consideration, named student_info, student_marks and student_attendance. All these three tables had a common attribute named roll_number on behalf of which the three tables joined. Then the projection operations get performed on the result obtained by joining the three tables to get only desired attributes from the database.

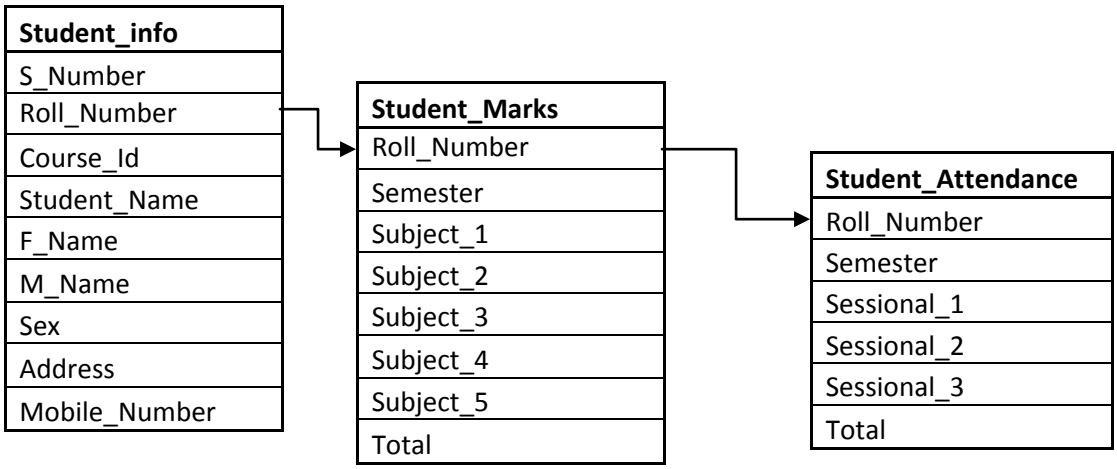

Fig 2: Student_info, Student_Marks and Student_Attendance Tables.

3.3.2 Data Selection & Transformation

After the data preparation, data selection and transformation steps were performed. In this step, the fields used in the study were determined and transformed if necessary. For example, as in the attendance records the student attendance were marked as A and P, A for absent and P for present in the classroom. These A and P got transformed into 0 and 1, 0 for absent and 1 for present. After this transformation, the total present attendance were added together to get the total attendance of the student throughout the semester.

Similarly the marks obtained in the separate subject were added together to get the total marks and then total marks of each sessional examination further added together to get the students marks for the semester means performance of the students in the semester.

```
Run information ===
Scheme:weka.clusterers.SimpleKMeans -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 5
Relation:     Student Database
Instances:485
Attributes:3
            ROLL No                    MARKS                    ATTENDANCE
Test mode: evaluate on training data
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 16
Within cluster sum of squared errors: 490.67073286312655
Missing values globally replaced with mean/mode
Cluster centroids:                    Cluster#
Attribute    Full Data        0        1        2        3
             (485)     (171)     (54)     (168)     (92)
=================================================================
ROLL No         BCA001    BCA001    BCA005    BCA004    BCA002
MARKS          91.6392   69.1988   240.537   56.4881   110.1413
ATTENDANCE     71.6103   79.5088   82.6111   26.3869   133.0543
Clustered Instances
0      171 ( 35%)          1      54 ( 11%)          2      168 ( 35%)          3      92 ( 19%)
```

Fig 3: Run Information of Clusters.

3.3.3 *Data Mining*

The prepared data was then put through the data mining process. The K – Means algorithm was used in this step and the distance measure used was Euclidean distance. The number of clusters was determined as an external parameter. The number of centroid chosen was 5 which are given in table.

Table 1. Cluster Centroids.

ROLL No	MARKS	ATTENDANCE
BCA001	56	34
BCA002	35	101
BCA003	85	85
BCA004	53	48
BCA005	124	33

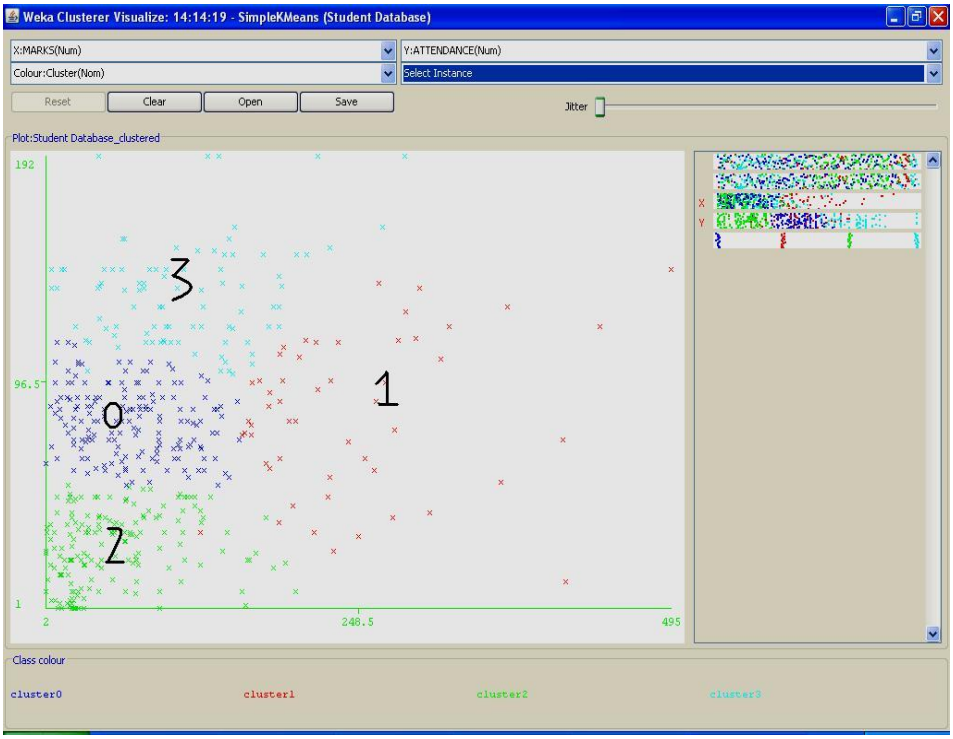

Fig 4: Each number shows different cluster.

3.3.4 *Presentation*

The results of data mining step are shown in this step. The given figures are generated using the Weka Environment and Paint. Figure 3 gives the algorithm used and the distance method used and all other information related to the clusters calculations.

The resulting clusters are shown in the figure 4. There are four clusters displayed in the given figure named cluster – 0, cluster – 1, cluster – 2 and cluster – 3. All clusters are shown using different colors. In the given figure numbers are mentioned showing the region of each cluster.

4. FINDINGS & OUTCOMES

The clusters obtained through the analysis of provided data can be displayed using the following clusters:

i) Cluster – 0, Students with high marks percentage and with low class attendance.

ii) Cluster – 1, Students with high marks percentage and with high attendance in the class rooms.

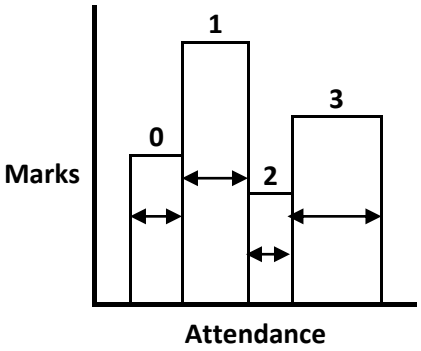

Fig 5: Each bar indicate a cluster as numbered.

iii) Cluster – 2, Students with low marks percentage as well as low attendance in the class rooms.

iv) Cluster – 3, Students with low marks percentage and having high attendance in the class rooms.

Out of these four group of students we need not required to put extra efforts to first and second group of students, the third group of student need not required extra faculty or time but it required the high attention and senior faculty who can analyze the problem of students and overcome the lacking points so that the students can perform better than what they are performing after attending most of the usual classes. Now, the fourth group of students requires proper attention of the faculty which motivates them to attend the classes and time to time mentoring. As once they will come to the class room they will automatically perform better or we can identify that where we should put them out of these groups which we have already observed through the analysis. As, until and unless, these students of the fourth group will come to the class room we cannot improve the quality of the students. The fourth group of students also needs extra time as well as some one to one personal discussion with faculty to find and resolve the problems with them due to which they are not attending the classes regularly.

5. CONCLUSION

On behalf of the study presented in the paper, we can utilize the resource in efficient manner. In academics members are the most important resource. So it is a quite typical task to decide that how to use this resource in an optimize way. As if we put our maximum faculty or junior faculty members to the first cluster of the students then it is not the efficient use of faculty members. In such case the students' already performing better will not improve much as they can if we provide them with the senior faculty member. By using senior faculty we can get the university toppers.

This study helps us to decide that where to use the maximum number of faculty and where the senior faculty members should be deputed. As faculty members in large number play totally different role than the senior faculty in less number.

So this paper helps the academic systems to use their human resource – faculty members in optimized manner and improve the standard of the education.

6. REFERENCES

[1] SP Singh et. al / VSRD International Journal of CS & IT Vol. 1 (7), 2011, 501 – 510.

[2] C.-H. Cheng, Y.-S. Chen / Expert Systems with Applications 36 (2009) 4176–4184.

[3] ERDOGAN, TIMOR, "A data mining application in students database", Journal of Aeronautics and Space technologies, July 2005 Vol. 2 No. 2 (53-57).

[4] Han, J., Kamber, W., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 5-10, 2001.

[5] Shi Na, Liu Xumin, Guan yong, "Research on k-means Clustering Algorithm" Proc. of the Third International Symposium on Intelligent Information Technology and Security Informatics, pp-63-67.

[6] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques"- A reference book.

[7] G.K. Gupta, "Introduction to Data Mining with Case Studies", A reference book, Third edition, PHI publications.

[8] Singh Vijendra, Kelkar Ashwini, Sahoo Laxman, "An Effective Clustering Algorithm for Data Mining" Proc. of the 2010 International Conference on Data Storage and Data Engineering, pp-250-253.

[9] Waikato Environment for Knowledge Analysis Version 3.6.5 (c) 1999 – 2011 of the University of Waikato Hamilton, New Zealand.