International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015 Page No.67-69
ISSN: 2278-2419

Predicting Students Performance using K-Median Clustering

B. Shathya
Asst. Professor, Dept. of BCA, Ethiraj College for Women, Chennai, India
Email: Shathya80@yahoo.co.in

Abstract— The main objective of education institutions is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge of students in a particular course. The knowledge is hidden among the educational data set and it is extractable through data mining techniques. In this paper, the K-Median method in clustering technique is used to evaluate students performance. By this task the extracted knowledge that describes students performance in end semester examination. It helps earlier in identifying the students who need special attention and allow the teacher to provide appropriate advising and coaching.

Keywords—Data Mining, Knowledge, Cluster technique, K-Median Method

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases. Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects. The aim of cluster analysis is to find the optimal division of m entries into n cluster. The aim of this paper is to find out group of students who needs special attention in their studies. The students' who are below average in their studies are found by using K-Median method by using three seeds. The three seeds the students' (objects) with lowest, average and highest marks. The distance is computed using the attributes and sum of differences. Based on these distance each student is allocated to nearest cluster. The distance is recomputed using new cluster means. When the cluster shows that the objects have not change that clusters are specified as the final cluster.

II. DATA MINING TECHNIQUES

A. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples.

B. Association rule

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

C. Clustering Analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task

D. Outlier Analysis

A database may contain data objects that do not comply with the general behaviour of the data and are called outliers. The analysis of these outliers may help in fraud detection and predicting abnormal values.

III. TYPES OF CLUSTERS

A. Well-separated clusters

A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in the cluster. . Even with the Manhattan-distance formulation, the individual attributes may come from different instances in the dataset;

thus, the resulting median may not be a member of the input dataset.

B. Center-based clusters

A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the "center" of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid.

C. Contiguous clusters

A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.

D. Density-based clusters

A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density.

IV. K-MEDIAN CLUSTERING

In data mining, K-Median clustering is a cluster analysis algorithm. It is a variation of *k*-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median. The K-Median method is the simplest and popular clustering method that is easy to implement. One of the commonly used distances metric is the Manhattan distance or the L_1 norm of the difference vector. In the most cases, the results obtained by the Manhattan distance are similar to those obtained by using Euclidean distance. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

$$\mathbf{D(x , y) = \sum \left| x_i - y_i \right|}$$

Although the largest valued attribute can dominate the distance not as much as in the Euclidean distance. This has the effect of minimizing error over all clusters with respect to the 1-norm distance metric, as opposed to the square of the 2-norm distance metric. The median is computed in each single dimension in the Manhattan-distance formulation of the K-Median problem, so the individual attributes will come from the dataset. This makes the algorithm more reliable for discrete or even binary data sets. In contrast, the use of means or Euclidean-distance medians will not necessarily yield individual attributes from the dataset. Even with the Manhattan-distance formulation, the individual attributes may come from different instances in the dataset; thus, the resulting median may not be a member of the input dataset.

V. EXPERIMENTAL RESULT

The k-means method uses the Euclidean distance measure which appears to work well with compact clusters. If instead of the Euclidean distance, the Manhattan distance is used then the method is called k-Median method. The K-Median method is less sensitive to outliers.

TABLE I. Sample students' Data

Student	T1	T2	Q	A
S1	13	16	12	18
S2	7	6	9	13
S3	4	1	10	8
S4	10	12	13	16
S5	6	3	16	10
S6	15	18	18	19
S7	2	7	10	9
S8	12	12	12	12
S9	18	17	14	18
S10	4	7	11	14
S11	10	11	11	13
S12	16	15	17	19
S13	7	5	14	11
S14	5	7	14	9

T1 – Continuous Assessment1
T2 – Continuous Assessment2
Q – Quiz
A – Assignment

In this study the students' data has been collected from a reputed college. The college offers many courses in both shifts (I & II). The data are taken from BCA Department. The class contains 50 students. From the 50 students', 14 objects has taken as sample. There are various components used to assess the internal marks. The various components include two continuous assessment tests, Quiz and Assignment. Each component carries twenty marks.

TABLE II. The three seeds

Student	T1	T2	Q	A
S3	4	1	10	8
S11	10	11	11	13
S6	15	18	18	19

Let the three seeds be the students' with lowest, average and highest marks. These seeds are found by finding the sum of all attributes values of the student. Now the distance is computed by using the four attributes and using the sum of absolute differences. The distance values for all the objects are given with the distances from the three seeds. Based on these distances, each student is allocated to the nearest cluster given in table 3.The first iteration leads to two students in the first cluster and eight students in the second cluster. There are four students in third clusters.

C1 → S3,S7
C2 → S2,S4,S5,S8,S10,S11,S13,S14
C3 → S1,S6,S9,S12

Now the new cluster means are used to recomputed the distance of each object to each of the means, again allocating each object to the nearest cluster Now the sample is clustered. Each remaining objects are then assigned to the nearest cluster obtained from the sample.Finally three clusters are formed using three seeds. The cluster obtained by using object with highest mark as starting seed is the above average students group.

TABLE III. First Iteration – allocating each object to the nearest cluster

C1	4	1	10	8	Distance from clusters			Allocation to nearest clusters
C2	10	11	11	13				
C3	15	18	18	19	C1	C2	C3	
S1	13	16	12	18	36	14	11	C3
S2	7	6	9	13	14	10	31	C2
S3	4	1	10	8	0	22	47	C1
S4	10	12	13	16	28	6	19	C2
S5	6	3	16	10	12	10	35	C2
S6	15	18	18	19	47	25	0	C3
S7	2	7	10	9	5	17	42	C1
S8	12	12	12	12	25	3	22	C2
S9	18	17	14	18	44	22	3	C3
S10	4	7	11	14	13	9	34	C2
S11	10	11	11	13	22	0	25	C2
S12	16	15	17	19	44	22	3	C3
S13	7	5	14	11	14	8	33	C2
S14	5	7	14	9	12	10	35	C2

.

TABLE IV. New Seeds

Student	T1	T2	Q	A
SEED1	3	4	10	8.5
SEED2	7.6	7.9	12.5	12.3
SEED3	16	17	15.3	18.5

TABLE V. Second Iteration - allocating each object to the nearest cluster

C1	3	4	10	8.5	Distance From Clusters			Allocation to nearest clusters
C2	7.6	7.9	12.5	12.3				
C3	16	17	15.3	18.5	C1	C2	C3	
S1	13	16	12	18	33.5	19	6.8	C3
S2	7	6	9	13	9.5	5.3	30.8	C2
S3	4	1	10	8	2.5	17	42.8	C1
S4	10	12	13	16	25.5	11	14.8	C2
S5	6	3	16	10	9.5	5.3	30.8	C2
S6	15	18	18	19	39.5	30	4.2	C3
S7	2	7	10	9	2.5	12	37.8	C1
S8	12	12	12	12	22.5	7.7	17.8	C2
S9	18	17	14	18	41.5	27	1.2	C3
S10	4	7	11	14	10.5	4.3	29.8	C2
S11	10	11	11	13	19.5	4.7	20.8	C2
S12	16	15	17	19	41.5	27	1.2	C3
S13	7	5	14	11	11.5	3.3	28.8	C2
S14	5	7	14	9	9.5	5.3	30.8	C2

After the second iteration, the number of students in C1, C2 and C3 remains same.
C1 → S3,S7
C2 → S2,S4,S5,S8,S10,S11,S13,S14
C3 → S1,S6,S9,S12

The cluster obtained by using object with average mark as starting seed is the average students group. The cluster obtained by using object with lowest mark as starting seed is the below average students group.The cluster C1 contains below average students' group. The objects S3 and S7 in this sample are considered as weak students in studies. They have to concentrate more on their studies. Otherwise they may fail in their final examinations. This study helps the teachers to provide extra coaching to the particular cluster of students to reduce the failure percentage in end semester examinations.

VI. CONCLUSION

In this paper, the Clustering technique is used on student database to predict the students division on the basis of previous database. Information like Class tests, Quiz and Assignment marks were collected from the students' previous database to predict the performance at the end of the semester. This study will help to the students and the teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reduce fail ratio and taking appropriate action for the end semester examination.

REFERENCES

[1] Heikki, Mannila, Data mining: machine learning, statistics, and databases, IEEE, 1996.

[2] U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, ISBN 0–262 56097–6, 1996.

[3] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[4] Alaa el-Halees, Mining students data to analyze e-Learning behavior: A Case Study, 2009.

[5] Z. N. Khan, Scholastic achievement of higher secondary students in science stream, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[6] U. K. Pandey, and S. Pal, A Data mining view on class room teaching language, (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[7] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, Data mining model for higher education system, Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.

[8] Ali Buldua, Kerem Ucgun, Data mining application on students data. Procedia Social and Behavioral Sciences 2 5251–5259, 2010.

[9] Singh, Randhir. An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education, 2013.

[10] Baradwaj, Brijesh Kumar, and Saurabh Pal. Mining Educational Data to Analyze Students' Performance.