ISSN No. 0976-5697

Volume 6, No. 2, March-April 2015

International Journal of Advanced Research in Computer Science

CASE STUDY AND REPORT

Available Online at www.ijarcs.info

Knowledge Discovery to Analyze Student Performance using k-mean Clustering depend upon various mean values input methods: A Case Study

Mrs. Biradar Usha
Assistant professor P. G. Department Of Computer Science,
Alva's College Moodbidri,India

Abstract: - The main objective of educational institutions is to provide high quality of education. Providing a high quality of education depends on predicting the unmotivated students before they entering in to final examination. One way to achieve quality to higher education system is by discovering knowledge of student in particular subject.

Data Clustering is used to extract meaningful information and plays a vital role in data mining. Its main job is to group the similar data together based on the characteristic they possess. The mean values are the centroids of the specified number of cluster groups. The centroids, though gets changed during the process of clustering, are calculated using several methods. In this paper, the k-mean clustering algorithm, depending upon various mean values input methods is used to discover knowledge that describes student performance. This study will help the teachers to reduce the drop out ratio, improve the performance of students and help identifying students who need special attention.

Keywords: Data Mining, Knowledge discovery, Cluster techniques, K-mean Method.

I. INTRODUCTION

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories. It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, signal processing, and many application fields, such as business, economics, and bioinformatics.

Data clustering is a process of extracting previously unknown, valid, positional useful and hidden patterns from large data sets. The amount of data stored in educational databases is increasing rapidly. Clustering technique is most widely used technique for future prediction. The main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities. These applications can help both instructor and student to enhance the education quality.

In this paper is to find out group of student who need special attention in their studies. The students who are low in their studies are found using k-means depend upon various mean value input method by using three clusters and then compare result of mean value input method. The three cluster of the object with Good, Medium and Low marks. The distance is computed using Euclidean distance. Based on these distance each student allocated to nearest cluster. The distance is recomputed using new cluster means. When the cluster shows that the object have not changed that cluster are specified as final cluster.

II. DATA MINING TASKS

Generally different classes of tasks can be achieved by exercising DM [3][5][6][7][8]:

a. ***Prediction*****:** this task aims at forecasting what might happen in the future by estimating the likelihood of a certain event's occurrence.

b. ***Classification***: it is usually exercised to identify group membership in a population instances. Popular classification techniques use Neural Networks (NN) and Decision Trees.

c. ***Clustering***: it is applied to position elements of a database into specific groups according to some attributes. The most frequently modi operandi are k-means and expectation maximization.

d. ***Association*****:** this area of DM aims at analyzing data to identify consolidated occurrence of events and uses the criteria of support and confidence. It is known to be applied in customer behaviour and machine learning. A popular procedure used is the *Apriori* algorithm.

e. ***Sequential Analysis***: this task targets the occurrence of special sequence of events where time plays a key role. It leads to the Identification of the events that most likely will lead to later ones with a specified minimum support or percentage.

III. K-MEANS CLUSTERING ALGORITHM DEPENDING UPON VARIOUS MEAN VALUES INPUT METHODS

K-Means is one of the simplest unsupervised learning algorithms used for clustering. K-means partitions "n" observations in to k clusters in which each observation belongs to the cluster with the nearest mean.

The Basic algorithm and flow-chart of K-means clustering is given see fig1[4]…

Algorithm 1 Basic K-means Algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Fig.1 Traditional K-Means Algorithm [4].

From the algorithm it is easily seen that, initially we have only the raw data. So, it is clustered around a single point. If the cluster number K is fixed then we need to cluster around that point. If the cluster is not fixed then it is continued until the cantered is not changed. Initially the students are all in a same group. But when K-means clustering is applied on it then it clusters the student's into three major categories, one is good, one is medium, and the other is low standard student. The flow chart of the k-means algorithm that means how the k-means work out is given see fig 2 [1] .

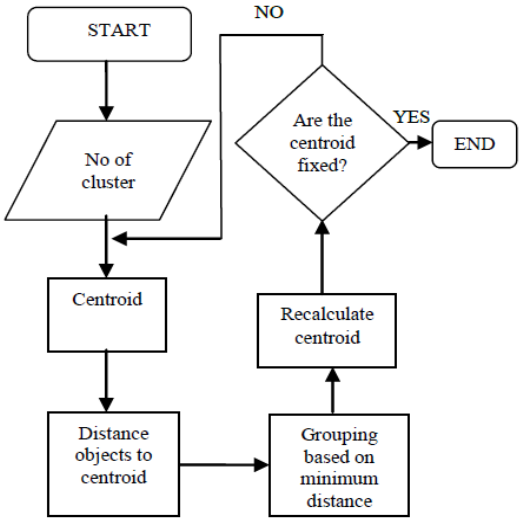

Fig.2 Flow-Chart Of K-Means Clustering.

The clustering is done by using shortest distance method to cluster the given input set, depending upon the centroids values initialized.
The centroids initially are specified by the user with one of the described methods below. Further processing of centroids are calculations by finding the mean value of every cluster, leading to name the algorithm as k-mean clustering. K denotes the total number of clusters. The three methods for initializing the centroids are[2]:

☐Taking the first 'k' values as centroids.

☐ Random centroids generation.

☐ User specified centroids.

A. Algorithm (1st method):

1. Assign first k - values as initial centroids m1,m2….mk from the given set of inputs.
2. Assign each item to the cluster which has nearest mean.
3. Calculate new mean for each cluster until the centroids do not change.

B. Algorithm (2nd method):

1. Assign initial centroids m1, m2….mk randomly.
2. Assign each item to the cluster which has nearest mean.
3. Calculate new mean for each cluster until the centroids do not change

C. Algorithm (3rd method):

1. Assign initial centroids by getting user inputs manually m1, m2….mk
2. Assign each item to the cluster which has nearest mean.
3. Calculate new mean for each cluster until the centroids do not change

IV. RESULT

In this the student data has been collected from reputed college. There are various components include two Assessment Test, Seminar and Assignment. Each component carries 30 marks see table I. TABLE I Sample dataset1

SN	TEST1	TEST2	SEMINAR	ASSIG.
s1	29	9	21	23
s2	17	17	21	23
s3	25	16	30	26
s4	30	30	27	25
s5	16	26	30	30
s6	21	7	24	23
s7	14	17	27	27
s8	18	27	30	28
s9	21	8	18	21
s10	15	13	24	21
s11	21	28	27	27
s12	16	22	30	30
s13	30	21	24	25

s14	29	22	30	26

Method 1:- k=3

In first method assign first k value as initial centroids from given set of input see table II.

Table II the three seeds

SN	TEST1	TEST2	SEMINAR	ASSIGNMENT
s1	29	9	21	23
s2	17	17	21	23
s3	25	16	30	26

Now the distance is computed by using four attributes and using the sum of absolute difference. The distance values of all objects are given with the distances from the three seeds.

Table III First Iteration

C1 29 9 21 23 C2 17 17 21 23 c3 25 16 30 26					Distance From cluster s			Allocation To nearest cluster
					C1	C2	C3	
s1	29	9	21	23	0	20	23	c1
s2	17	17	21	23	20	0	21	c2
s3	25	16	30	26	23	21	0	c3
s4	30	30	27	25	30	34	23	c3
s5	16	26	30	30	46	26	23	c3
s6	21	7	24	23	13	17	22	c1
s7	14	17	27	27	33	13	16	c2
s8	18	27	30	28	43	25	20	c3
s9	21	8	18	21	14	18	29	c1
s10	15	3	24	21	25	21	34	c2
s11	21	28	27	27	37	25	20	c3
s12	16	22	30	30	42	22	19	c3
s13	30	21	24	25	18	22	17	c3
s14	29	22	30	26	25	29	10	c3

Based on distances each object is allocated to nearest cluster see table III. The first iteration

C1→s1, s6, s9

C2→ s2, s7, s10

C3→s3, s4, s5, s8, s11, s12, s13, s14

Now calculate new cluster centroids are used to recompute the distance of each object to each of the means, again allocating each object to nearest cluster. After second iteration

C1→s1, s2, s6, s9

C2→ s10

C3→s3, s4, s5, s7, s8, s11, s12, s13, s14

After third iteration the no. of student in c1, c2, c3 remains same

C1→s1, s2, s6, s9

C2→ s10

C3→s3, s4, s5, s7, s8, s11, s12, s13, s14

Finally three clusters formed using three seeds using method 1. The cluster c1 contain medium students group. In this the object s10 consider weak in second test.

Method 1I:- k=3

In second method assign randomly k value as initial centroids from given set of input see table III.

Table III the three seeds

S3	25	16	30	26
S7	14	17	27	27
S12	16	22	30	30

After fourth iteration

C1→s3, s4, s13, s14

C2→ s1, s2, s6, s9, s10

C3→s5, s7, s8, s11, s12

The cluster c1 contain good student group, c2 contain weak group in this s1,s6 and s10 performance low in test2 and c3 contain medium group.

Method 1II:- k=3

In third method Assign initial centroids by getting user inputs manually from given set of input see table III.

Table III the three seeds

S9	21	8	18	21
S1	29	9	21	23
S14	29	22	30	26

After second iteration

C1→s2, s6, s9, s10

C2→ s1

C3→s3, s4, s5, s7, s8, s11, s12, s13, s14

The cluster c1 contain weak student group, c2 contain medium group and c3 contain good group.

V. CONCLUSION

In this paper the clustering techniques used on student database to discover helpful knowledge. Information like test, seminar and assignment marks collected from student previous database to predict the performance at the end of semester. All above methods are help full in analysis student performance in different view. These methods have their own applications and they may help for further research. This study helps to student and teachers to improve performance of student.

VI. REFERENCES

[1] Md. Hedayetul Islam Shovon, Mahfuza Haque " An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree" International Journal of Advanced Computer Science and Applications, Vol.3,2012, No. 8

[2] Sridhar. A, Sowndarya. S "Efficiency of K-Means Clustering Algorithm in Mining Outliers from Large Data Sets Efficiency of outlier detection using several centroid initialization methods" International Journal on Computer Science and Engineering Vol. 02, No. 09, 2010, 3043-3045

[3] Eui-Hong (Sam) Han, AnuragSrivastava and Vipin Kumar "Parallel Formulations of Inductive Classification Learning Algorithm" (1996).

[4] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar & M.Inayat Khan, "Data Mining Model for Higher Education System" European Journal of Scientific Research, ISSN 1450-216X Vol.43 No.1 (2010), pp.27.

[5] Agrawal, R. Srikant. "Fast Algorithms for Mining Association Rules". *Proc. of the 20th Int'l Conference on Very Large Databases.* Santiago, Chile, Sept. 1994.

[6] J. Han, J. Chiang, S. Chee, J. Chen, Q. Chen, S. Cheng, W.Gong, M. Kamber, K. Koperski, G. Liu, Y. Lu, N.Stefanovic, L. Winstone, B. Xia, O. R. Zaiane, S. Zhang, H. Zhu, *DBMiner*. "A System for Data Mining in Relational Databases and Data Warehouses". *Proc. CASCON'97:Meeting of Minds.* Toronto, Canada, November 1997.

[7] Cheung, J. Han, V. T. Ng, A. W. Fu and Y. Fu. "A Fast Distributed Algorithm for Mining Association Rules". *Proc. of 1996 Int'l Conf. on Parallel and Distributed Information Systems (PDIS'96).* Miami Beach, Florida, USA, Dec. 1996.

[8] R. Agrawal, R. Srikant."Mining Sequential Patterns".*Proc. of the Int'l Conference on Data Engineering (ICDE).* Taipei, Taiwan, March 1995.