Educational Data Mining: An Intelligent Approach using Modified K-Means Clustering

[1]A Sumathi, [2] Dr N Sengottaiyan
[1]Ph.D Scholar & Assistant Professor, [2]Director
[1]Department of Computer Science,
[1]R & D Centre, Bharathiar University, Coimbatore, Tamilnadu, India &
Navarasam Arts and Science College for Women, Erode, Tamilnadu, India

Abstract : In the recent years, many research works have been carried out in order to predict the students' performance for their knowledge evaluation, students' weakness and the total number of failures in the end semester examination. The performance of a particular student depends upon various factors existing in the society. When all the factors are analyzed, a clear picture can be given to improve the performance of all the students. In this paper, issues related to women's education creating impact on continuing or discontinuing the course has been analyzed using the modified k-means algorithm with the help of the data set created.

***IndexTerms* - Decision making, Educational Data Mining, Modified K-Means algorithm, Prediction Data Set**

I. Introduction And Background Study

The innovative process of extracting the useful information from large data set is referred as Data Mining. The ultimate goal of data mining process is to extract the information from the available data set into an understandable structure. The educational data mining is categorized under the emerging field of research where the analysis of educational data is carried out for developing various models for improvising the experience in learning and institutional effectiveness. Specifically various EDM methods are developed for the process of decision making in the educational system. Dinesh Kumar et al. [1] presented the C4.5 and ID3 algorithm with feature selection technique ranker analysis in predicting the higher secondary students' performance. Abdulmohsen Algarni [2] epitomizes the survey on datamining techniques related to education, which addresses the various problems faced by the students in improving their education. Analysis on various models representing the data mining process [3] has been addressed by Veeramuthu et al. The justification and capabilities of data mining related to higher educational system was designed and presented. Durairaj et al. [4] addressed the trust model using data mining techniques in prediction of the student performance. The trust model aims at mining the information required in such a way that the present education system might adopt in this management tool. Sreedevi et al. [5] discusses the students' academic performance with the help of k-means and decision tree algorithm, which ensures the quality of the educational system. In the year 2016, Sagar S Nikam [6] presented an extensive survey related to data mining techniques that explains the features and limitations of classification methods. Mansi et al [7] proposed the various tools and its features used in data mining techniques, where the different validation indices are summarized. Suguna et al. [8] presented the detailed literature on data mining techniques related to web documents and services. Lovleen et al [9] discussed bout filling the bridge gap between two different communities in terms of similarities and dissimilarities. Anu Sharma et al. [10] epitomizes the review on data mining techniques and discussed about the challenges related to social network analysis. The validation has been carried out using the web mining techniques. From the above analysis, it is very clear that the general issues related to education has been analyzed using various data mining techniques. and the issues related to the retardation of womens education has been missed out. In this article, using the real time data set of 50 samples, the most important issues related to womens education has been analyzed using the modified k-means clustering algorithm.

II. Clustering

Clustering analysis is a process of discovering the substrate of a data set by grouping into several clusters. The term "Cluster" refers to the method of grouping unlabeled data. In the data analysis, the similar items are grouped in to one partition and in the same way different items are grouped in to different partition. Various applications such as data analysis, market research, image processing and pattern recognition are analyzed using cluster analysis. Based on the values assigned to the attributes the prediction is carried out. Generally the clustering techniques can be classified into different categories such as density-based, model based, partitioning, grid-based and hierarchical based algorithms. Figure 1 explains the simple process of data extraction from huge data set.

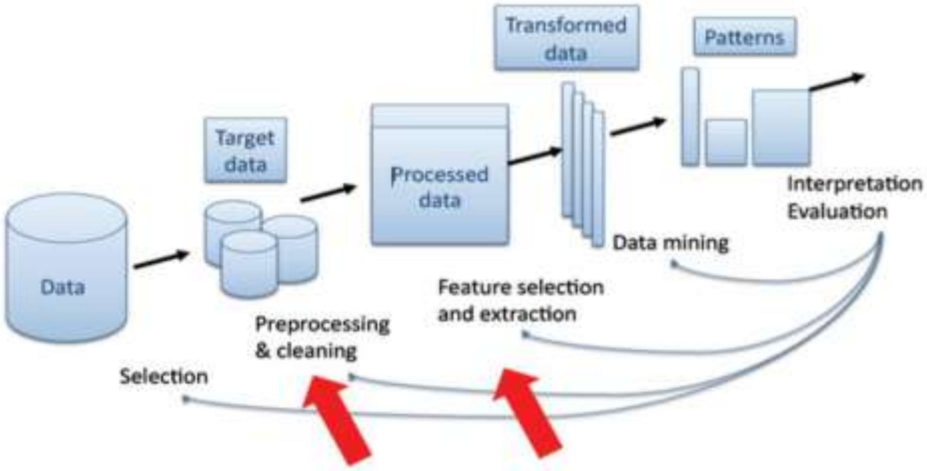

Fig. 1 process of data extraction

III. EXISTING K-MEANS CLUSTERING

K-Means is the simplest method of clustering in the data mining process. Existing KMC involves in partitioning 'n' objects into clusters of 'k' values where a single object belongs to the cluster of nearest mean. The definite number of clusters are produced by this method with some specific characteristics. Initially the best number of clusters leading to greatest distance is not known and the same has to be calculated from the input data. K means test aims at choosing the best cluster center which is to be acted as the centroid. The existing method involves in two different phases. In the first phase, the value of 'k' is chosen randomly where the value of k is chosen in advance. In the second phase, each object is taken near to the center. On bringing all the objects within the cluster, the initial process is over. The iteration gets repeated till the objective function attains the minimum value. The inputs to the K-mean algorithm are the number of clusters (predefined), the value 'k' and the data set with n data objects. The output will be a group of 'k' clusters. The disadvantage of the existing KMC is that it takes high computational time in determining the cluster centroids.

IV. MODIFIED K-MEANS CLUSTERING

In the modified k-means clustering, the recalculation of new centroid takes place. This might be due to the addition of new points in the cluster. Once when the new centroids are determined, a new position is created between the new centroids and the existing data points by a loop creation. With the help of step-by-step process, the k-centroids might change their original position. After sometime there exists a situation in such a way that there is no movement for centroids.

3.1Steps involved in Modified K-Means Clustering

i) For the chosen data set, determine 'range' between the minimum and the maximum value.

i) Column which has the maximum range is determined.

iii)Sort the chosen data set in the increasing order based on the maximum range(column wise)

iv)Divide the sorted data set of 'k' equal parts.

v) Arithmetic mean of each part is obtained in step 4 and assume those mean values as initial centroids.

vi) Repeat the steps until the convergence criterion is met out.

3.2 R-Tool Implementation

The implementation has been carried out using R-tool. The following social and personal related issues are opted for the purpose of analysis. Age when Join the Course, Course Related to High school Major, Social factor (M/Un), Affected by Domestic Issue (Y/N), Parent Mentality, Educational facilities (G/B/W), Participate Awareness related to women's welfare, Receive any Incentives, Worked as a Labour, Facing Financial Difficulty, Physically Challenged (or) Any Health Problem, Attendance, and End semester Result.

Table 1 list of attributes, its description and its possible values for data set

Attributes	Description	Possible values
Age of the student	Age when the student joins the course	18-23
Course	Selected course whether related to high school major	Yes / No
Social factor	Whether the student is married / unmarried	Married / Unmarried
Domestic Issue	Whether the student is affected by social / domestic issue	Yes / No
Parent Mentality	Parents' thought about the education for women	Supportive / Discourage
Educational facilities	Facility available for education in particular area	Good / Bad / Worst
Awareness programme	Participating in Awareness programmes related to womens welfare	Yes / No
Labour	Whether working as a labour	Yes / No
Incentives	Receive any incentives	Yes / No
Financial position	Whether the student is facing financial difficulty	Yes / No
Physically Challenged	Whether the student is affected by health related problems	Yes / No
Attendance	Attendance performance in a particular semester	Students' performance
End semester Result	Percentage of marks obtained in the current semester exam	Good / Bad / Worst

Table 2 Cluster segment centre assigned to each data

S.No.	X1	X2	X3	X4	S.No.	X1	X2	X3	X4
1	2.25	1	1	0.925	**26**	2	2	0.5	0.88
2	1.75	1	0.75	0.9	**27**	2	2	0.5	0.865
3	1.5	3	0	0.85	**28**	2	2	1	0.87
4	1.25	3	0	1.355	**29**	1.25	3	0	0.885
5	1.75	2	0.75	0.865	**30**	1.75	2	0	0.9
6	2.25	1	1	0.935	**31**	2	2	1	0.89
7	1.25	2	0.25	0.85	**32**	2.25	1	1	0.945
8	2	2	0.5	0.86	**33**	2.25	1	1	0.925
9	2	1	0.5	0.865	**34**	1.75	1	0.75	0.9
10	1.25	3	0.25	0.855	**35**	1.25	2	0.25	0.9
11	1.75	2	0.5	0.85	**36**	1.75	2	0.75	0.895
12	2.25	1	1	0.88	**37**	2	2	1	0.87

13	1.25	2	0	0.865	**38**	2.25	1	1	0.91
14	2	1	1	0.9	**39**	1.25	3	0	1.35
15	2.25	1	1	0.915	**40**	1.75	2	0.25	0.88
16	2.25	2	0.75	0.9	**41**	2	1	0.5	0.9
17	1	3	0.25	0.85	**42**	2.25	1	1	0.95
18	1.5	2	0	1.36	**43**	1.25	3	0	0.88
19	1.5	3	0	0.865	**44**	1.25	3	0	0.88
20	1.5	2	0	0.86	**45**	2	2	0.5	0.88
21	1.5	3	0	0.85	**46**	1.5	3	0	0.865
22	2.25	2	1	0.87	**47**	1.5	2	0	0.86
23	2.25	1	1	0.95	**48**	1.5	3	0	0.85
24	1.25	3	0	0.88	**49**	2.25	2	1	0.87
25	1.25	3	0	0.88	**50**	2.25	1	1	0.95

Table 1 and Table 2 epitomize the list of attributes selected, its possible values and the cluster segment centre assigned to each data for the attributes chosen. It can be seen from table 2 that four different clusters re grouped and are denoted as X1, X2, X3 and X4. Here X1 cluster carries the "social factor and parent mentality", X2 holds the "End semester result", X3 carries "Affected by domestic issue, Participating in awareness programme, Receive any incentives and worked as labour", X4 holds "Physically challenged and Attendance". In data mining, the method of interpretation and consistency validation within the clusters is referred as Silhouette. A plot showing the silhouette average width as the cluster center varies is represented in figure 2.

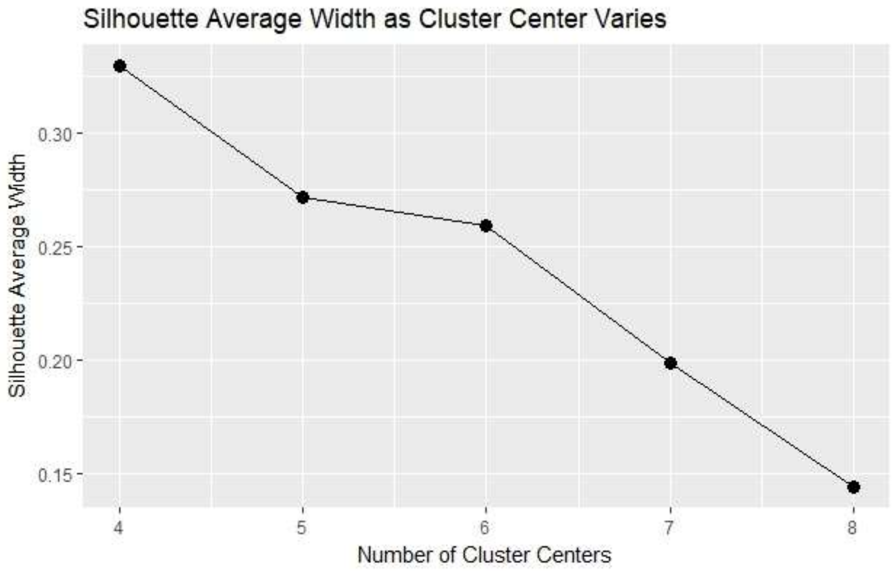

Fig. 2 Silhouette Average Width

The above figure shows the representation of data points consistency within the same cluster. It is observed that minimum number of cluster centers are formed and also it changes according to the changes in the data set.

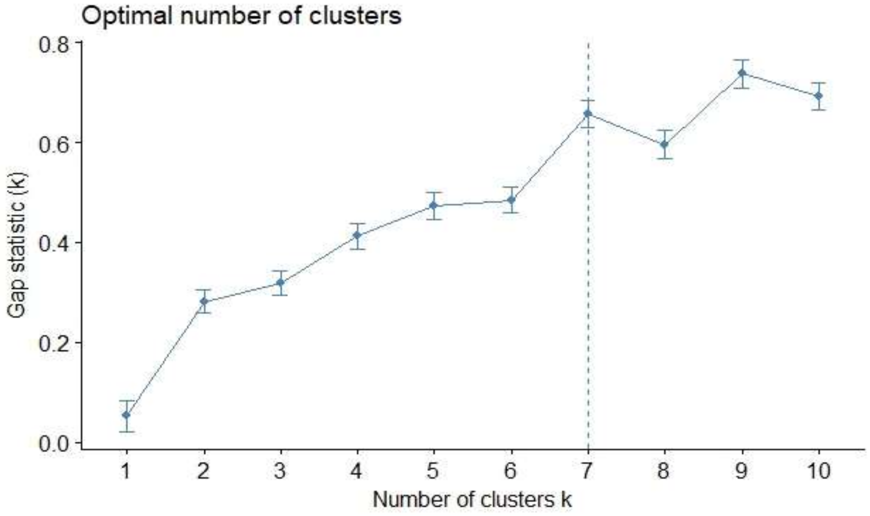

Fig. 3 Optimal number of Clusters

From the figure 3, it is observed that when the data set is processed in the R-tool, there arises the formation of many cluster points within the data set. The optimal number of clusters is measured with the number of clusters and gap statistics. Figure 4 epitomizes the k-mean cluster formation. The entire data set is categorized into 3 various clusters with specific distance between each cluster center.

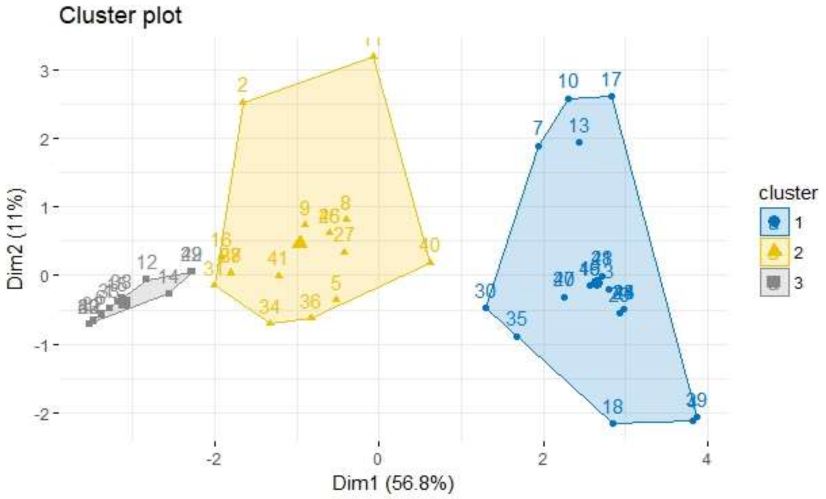

Fig. 4 K-mean cluster formation

From the above analysis and interpretation, it is observed that applying the K-means partitioning the entire data set is divided into various clusters with non-movable centroids after repeated iterations. Also the computational time taken to achieve the cluster center is very less when compared to basic KMC algorithm.

V. Conclusion

In this study we implement the data mining process using K-Means partioning method by taking the students data set as the input. The reason for choosing the aforesaid mentioned attributes is to predict the factor that stops women from discontinuing the degree course. The managements can use some techniques to retard the course discontinuation. The information generated after the implementation of data mining and data clustering technique may be helpful for the management to take necessary steps to get rid of mentioned issues. Following this many public awareness programmes may be conducted to educate the uneducated people and improve the percentage of women continuing the education. Refining the technique to achieve the accurate and valuable outputs may be considered for future work.

REFERENCES

[1] Dinesh Kumar and Radhika. V. 2016. Mining Educational Data to Predicting Higher Secondary Students Performance. International Journal of Computational Intelligence and Informatics, 6(2): 1-6.

[2] Abdulmohsen. A. 2016. Data Mining in Education. International Journal of Advanced Computer Science and Applications, 7(6).

[3] Veeramuthu. P. Periyasamy. R. and Sugasini. V. 2014. Analysis of Student Result Using Clustering Techniques. International Journal of Computer Science and Information Technologies, 5 (4): 5092-5094.

[4] Durairaj. M. and Vijitha. C. 2014. Educational Data mining for Prediction of Student Performance Using Clustering Algorithms. International Journal of Computer Science and Information Technologies, 5 (4): 5987-5991.

[5] Sreedevi. K. and Chandra Srinivas Potluri. 2014. Analyzing the Student's Academic Performance by using Clustering Methods in Data Mining. International Journal of Scientific & Engineering Research, 5(6): 198-202.

[6] Sagar S Nikam. 2015. A Comparative Study of Classification Techniques in Data Mining Algorithms. Oriental Journal of Computer Science & Technology, 8(1): 13-19.

[7] Mansi. G and Shivani. G. 2015. Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity. International Journal of Computer Applications, 113(18): 22-29.

[8] Suguna. K. and Nandhini. K. 2015. Literature Review on Data Mining Techniques. International Journal of Computer Technology & Applications, 6 (4): 583-585.

[9] Lovleen Kumar. G. and Rajni. M. 2017. The Lure of Statistics in Data Mining. Journal of Statistics Education, 16 (1): 1-8.

[10] Anu Sharma, Sharma. M. K. and Dwivedi. R. K. 2017. Literature Review and Challenges of Data Mining Techniques for Social Network Analysis. Advances in Computational Sciences and Technology, 10(5): 1337-1354.