International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 06, June-2015

An Enhanced Clustering Algorithm by Comparative Study on K-Means Algorithm

Sonia Guliani
M.Tech & Department of Computer Science
Lovely Professional University, Phagwara, India

Alpana Vijay Rajoriya
Assistant Professor & Department of Computer Science
Lovely Professional University, Phagwara, India

***Abstract* -Clustering is a term used to divide objects into groups based upon their similarity. This paper proposes an enhanced k-mean clustering algorithm which improves the error and this enhanced algorithm will apply on traffic dataset to analyze the major factors contributing to the accidents. The enhanced algorithm is compared with the existing k-mean algorithm using a software package weka.**

***Keywords**: Clustering, K-mean Clustering, Weka Interface, Database.*

1. INTRODUCTION

Cluster analysis organizes data into groups that is meaningful. To maintain the natural structure of data, resulting clusters should be meaningful. Clustering is a term used to define the grouping of similar data items or objects. Items or objects which are similar in nature are grouped in one cluster and items or objects which are dissimilar in nature are far away are grouped together in another cluster. For example, cluster analysis has been used to identify groups of books according to their topic, author, and their core areas. By using Classification we can distinguish groups or classes of objects. In this the labelling of large set of training Clustering solves many classification problems. But the general question faced by the researchers is that how to classify the data into meaningful clusters. Objects belonging to same groups have maximal degree of association and remain minimal if they belong to different groups. The main objective of the clustering technique is that the objects resides in one group should be close to one another. Clustering is used for analysis of data and also resolves the classification problem. Cluster analysis is an important activity undertaken by humans. Using clustering technique, one is able to classify between cats, rats and various animals. In the machine learning, clustering is very suitable example of unsupervised learning. So we can use clustering used to learn from observations than learning by examples. Several attempts have been taken to find methods for efficient and effective cluster analysis in large number of databases. For example: Consider a library system in which the books related to a huge amount of topics are available. The books are always arranged in the manner that forms the clusters group. The books which are similar in nature are placed in one cluster group and the books which do not have any kind of similarity are placed in another cluster group. For example, the operating system books are placed in one shelf and the Networking books are placed in other shelf, and so on. The complexity can be further minimized by keeping the books which covers same type of topics are placed in that same shelf and then these shelves are given a specified name. Whenever a person wants a book of a particular topic, the user will go to that particular shelf and take the book from that shelf only rather checking in the whole library.

2. K-MEANS CLUSTERING

K-means clustering comes under centroid models of clustering algorithms. The k-means algorithm comes under the family of algorithms called as optimization algorithms of clustering. That is, the examples are divided into clusters groups in this way that the cluster gives good optimal results according to criteria defined. The name of the algorithm has been derived such that the k clusters are formed from the data set where the cluster centre is the arithmetic mean of all objects within that type of cluster. The number of the clusters k is known in advance. The first step is to find the initial centroids for each cluster. The next step is to associate each data object to its nearest centroid. Early grouping is done by assigning each data object to centroid which is so close to it and the first iteration is completed. The algorithm works in iterations until the objects does not change their cluster centres. Centroids move their positions until the convergence criteria have reached. Pseudeocode for algorithm is as follows:

Algorithm 1: K-mean clustering algorithm

Input:
- Dataset containing dataset objects
- Number of clusters k
-

Output:
- A set of number of clusters obtaining from dataset objects

Steps:
1. Randomly choose dataset objects from the dataset as the initial centroids.
2. Repeat

a. Associate each object to its nearest centroid;
b. For each cluster, new mean is calculated;

Until convergence point is reached.

3. LITERATURE REVIEW

Several works have been made to improve the quality of clustering algorithm. The method proposed in [3] to make the algorithm better in terms of number of clusters and execution time. It takes less execution time than both the k-mean and k-medoid algorithms even if the number of cluster size increases. The algorithm proposed in [4] gives a method to eliminate the problem of empty clusters due to bad initialization and provides the modified view of the k-mean algorithm.

The algorithm proposed by this paper is almost similar to the existing algorithm, but there is no degradation of performance due to any modifications. It can say that modified algorithm handles the empty cluster problem very efficiently. But as the number of clusters varies than specified range, empty clusters may form. The study done in [4] presents an improved mean based algorithm. This paper focuses on making k-mean algorithm globally optimum. For achieving this, initial centroids must be selected carefully. The paper described an improved algorithm to determine initial centroids, which results in better clusters equally for uniform and non-uniform data sets. This paper proposes an enhanced algorithm which focuses on reducing the error function, hence improving the accuracy of the algorithm. To check the error in the proposed algorithm a real traffic dataset is used. The algorithm described in [5] gives a method to identify initial clustering centroid of k-mean clustering. This paper enhances the performance of initialization method over many datasets by taking into consideration different observations, number of clusters, groups and clusters complexity.

4. PROPOSED WORK

The original k-mean algorithm starts by choosing the initial centroid randomly and works in three iterations. The number of clusters used in this algorithm is two. The next step is to associate each object to its nearest centroid. The process is iterated until it reaches to a convergence point. But the algorithm proposed by this paper works as follows:

Algorithm 2: The enhanced k-mean algorithm
Steps:

1. Choose arbitrary point p from a dataset.
2. Find all the points which lie in the neighbourhood of point p.
3. All the points which lie in the same area make a cluster.
4. Continues the whole process until all the remaining points have been processed.
5. At each iteration calculate the standard deviation and probability of each cluster formed.

Until cluster do not reaches its convergence point.

In order to calculate standard deviation of data objects lies in the cluster, calculate the variance of the dataset. The variance is calculated by using the following formula:

$$\frac{\sum(x-\mu)^2}{n}$$

Whereμ represents the mean and n is number of items. By taking the square root of the variance gives the standard deviation. It is denoted by σ notation. It is the square root of the variance and calculated by the formula as:

$$\sigma = \frac{\sqrt{\sum(x-\mu)^2}}{n}$$

5. RESULTS AND DISCUSSION

To test the performance of the enhanced algorithm traffic dataset is taken. For effective performance the experiment result of the enhanced algorithm is compared with the existing k-mean algorithm. The dataset is divided into thirteen attributes. The dataset contains six clusters. Each cluster contains different clustered instances. The total number of instances of dataset is 866. The result of the k-mean is as shown in figure1:

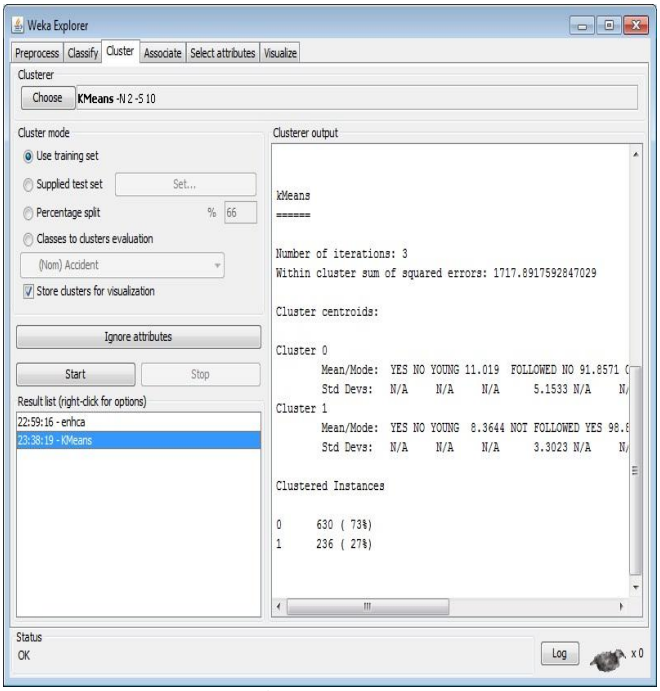

Fig.1. Result of existing k-mean algorithm

The traffic dataset is taken and is used for checking the error of the enhanced k- mean algorithm. The enhanced clustering algorithm is used for mining high dimensional dataset. The same dataset is used for the enhanced k-mean and the existing algorithm. The number of clusters is six. Figure 2 depicts the error obtained from the enhanced algorithm. The error is less in this case. The mean and the standard deviation is calculated in each cluster.

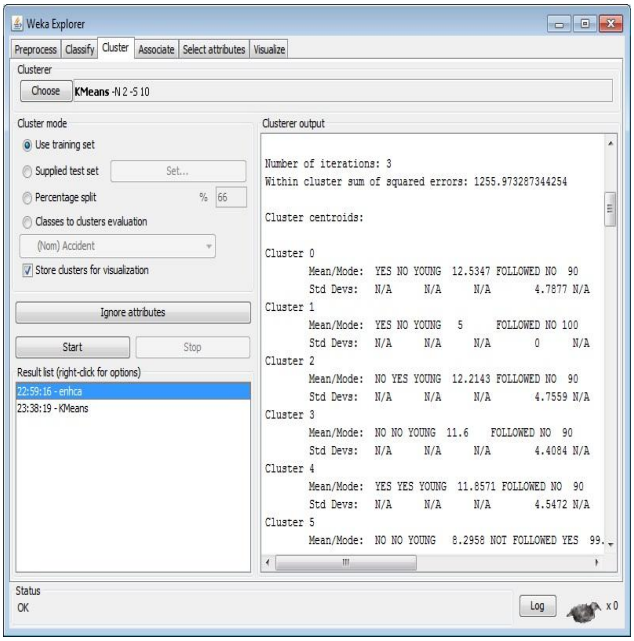

Fig.2. Result of enhanced k-mean algorithm

The results of the experiments are shown in the table 1. Thee error obtained from the enhanced algorithm is less than that of the existing k-mean algorithm.

Table I. Comparison shows error of both algorithms.

	Kmeans	Enhanced algorithm
Error	1717.89	1255.97

Figure 3 shows the error bar graph for the performance of both the algorithms. It is clear from the graph that the enhanced algorithm generates less error as compared to the existing k-mean algorithm.

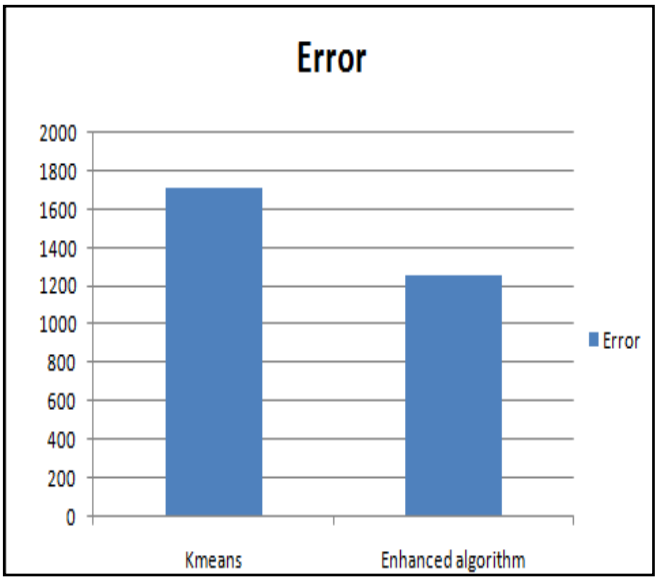

Fig. 3. Error graph shows performance of both the algorithms

It is clear from the table 2 that the numbers of iterations are same in both the cases. Same numbers of iterations are used
for the comparison of both the algorithms.

Table II. Number of iterations

	Kmeans	Enhanced algorithm
No. of iterations	3	3

It is clear from the figure 4 that the number of iterations is same. Same number of iterations is used for comparing the error.

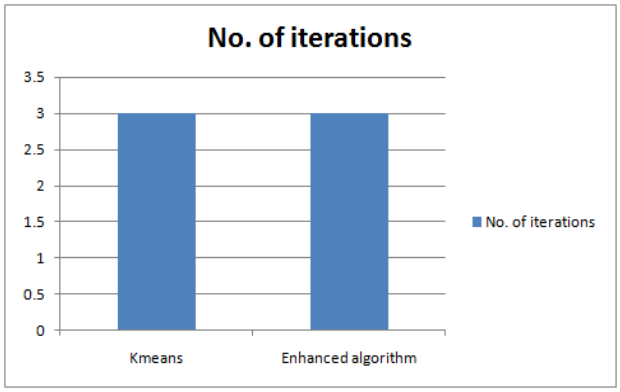

Fig. 4. Bar graph shows performance comparison using same number of iterations

6. CONCLUSION

The k-means clustering algorithm is used for mining high dimensional dataset. But the existing algorithm produces more error as compared to the enhanced algorithm. Thus it does not provide good accuracy results. The accuracy obtained by enhanced algorithm is much better than that of existing one. The paper provides an enhanced algorithm which performs better in number of clusters and the method for finding the centroid. The analysis is done by taking the traffic dataset which considers several attributes. Thus listing various attributes, which are the main reasons for the accidents.

ACKNOWLEDGMENT

The paper is written under guidance and support of computer science department who helped me in completion of the work. I would like to thank all the people who helped and encourage me by which the work is made possible.

REFERENCES

[1] Suman, Pooja Mittal "A Comparative Study on Role of Data Mining Techniques in Education", International Journal of Emerging Trends & Technology in Computer Science, Vol 3, Issue 3, pp. 65-69, May – June 2014.

[2] Amandeep Kaur Mann, Navneet Kaur "Survey Paper on Clustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, pp. 803-806, April 2013.

[3] Saurabh Shah, Manmohan Singh "Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm" International Conference on Communication Systems and Network Technologies pp. 435, 2012.

[4] Malay K. Pakhira," A Modified k-means Algorithm to Avoid Empty Clusters", International Journal of Recent Trends in Engineering, Vol 1, No. 1, pp. 220-226, May 2009.

[5] Raed T. Aldahdooh, Wesam Ashour "Distance-based Initialization Method for K-means Clustering Algorithm‖",I.J. Intelligent Systems and Applications, 02, pp . 41-51, 2013.

[6] Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal" Crime Analysis using K-Means Clustering", International Journal of Computer Applications (0975 – 8887) Volume 83 – No4, pp. 1-4,December 2013.

[7] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress on Engineering 2009 Vol IWCE 2009, July 1 - 3, London, U.K, 2009.

[8] M. Goyal, S. Kumar "Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability", IJ. Inst. Eng. India Ser. B DOI 10.1007/s40031-014-0106-z), 2014.

[9] Narendra Sharma, Aman Bajpai, and Ratnesh Litoriya "Comparison the various clustering algorithms of weka tools" International Journal of Emerging Technology and Advanced Engineering" ISSN 2250-2459, Volume 2, Issue 5, pp. 73-80, May 2012.

[10] Sapna Jain, M. Afshar Aalam, M. N DOJA "K-means Clustering Using Weka Interface" Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi, pp. February 25 – 26, 2010.

[11] Ritu Sharma, M. Afshar Alam, Anita Rani "K-Means Clustering in Spatial Data Mining using Weka Interface" International Conference on Advances in Communication and Computing Technologies (ICACACT) Proceedings published by International Journal of Computer Applications, pp. 26, 2012.

[12] H.S. Behera, Abhishek Ghosh., Sipak ku. Mishra "New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining" International Journal of Advanced Research Computer Science and Software Engineering"Volume 2, Issue 4, pp. 287-292, April 2012.