ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.7

Performance Evaluation of K-Means and EPBC (Enhanced Partitioning Based Clustering) on National Highway Authority

Er. RoopLal[1], Chetna[2]
HOD[1], M.Tech Scholar[2]
Department of Computer Science and Engineering
St. Soldier Institute of Engg. & Technology, Near NIT, Jalandhar, Punjab, India

Abstract
The data involved in this research is extracted from National Highway Authority-1. The objective of this research By utilizing data mining techniques to obtain the possibility distribution of the issues that would cause various kinds of traffic accidents to provide decision-making support for traffic management department, commercial insurance department, driver training department and the drivers. This research presents a data mining model using self developed clustering algorithm to analyze the traffic collision data. The experiment results from this study will show that the developed data mining model using decision tree can effectively classify the major contributing factors to traffic collisions and their collision severity for different groups of people with good accuracy.

Introduction
It is a process of finding patterns in data. Its goal is to find patterns that we repreviously unknown. Datamining is sometimes called Knowledge Discovery Process. Datamining is a powerful tool to find relevant information. Once you have right information, all you will need to do is apply it in the right manner. It is very easy to get information these days. But it is not so easy to obtain the relevant information that can help us to achieve a required goal. Hence the data mining becomes a powerful tool. It will give the power to predict certain behaviors within environment. It is used for a various kind of purposes in both the private and public sectors. The banking insurance and medicine industries use datamining techniques for reducing costs and increases sales. For example-Both the insurance and banking sectors use datamining applications for risk assessment and detecting frauds. Data mining is also used by medical fraternity for predicting the effectiveness of medicine. Data mining is used by various pharmaceutical firms to do research on treatments for various diseases. Datamining applications can be applied in the field of healthcare, immigration sectors and in business applications to solve specific problems. Various algorithms and techniques like clustering, classification, decision trees, artificial intelligence, neural networks etc. are used for knowledge discovery from databases. But here we are going to explain clustering.

K-means clustering
This type of clustering comes under centric models of clustering. This algorithm belongs under the family of algorithms which is known as centric based clustering. The K-mean clustering takes Kasan input parameter and divides the set of objects into number of clusters such that the result should be high in intracluster similarity and low in intercluster similarity. In this, the examples are partitioned into various clusters in such a way that these clusters are optimal by following some criteria. The name has been derived from the various factors that will form the kclusters. In this cluster the center part is the arithmetic mean of all items encloses that cluster. The k-mean when compare to SOM has a disadvantage that it cannot perform vector quantization, which means naturally it, is not in a form that can be easily visualized. K-mean has an advantageover SOM is that itismore computationallyefficient.

Advantages and Disadvantages of K-mean Clustering
1. If we keep k smalls and variables arelarge, then manytimes K-mean isfasterthan he hierarchicalclustering.
2. If the clusters are globalised in nature, K-Means produce tighter clusters than hierarchical clustering.
3. It will best results when the data set aredistinct.
4. It gives good results than classification.
5. It is fast and easy to understand.

Problem Formulation
In real life, number of accidents rate are increasing day by day. So, Traffic safety is highly important and analysis of the major factors contributing to the accidents also became an important issue. These factors are age, season, experience and so on. So there is need of study to focus on these factors such that measure can be taken in order to overcome these. By considering these factors also helps the traffic management to take effective decision. In order to do the analysis of traffic, there is a need of enhancement of clustering algorithm. So the problem can be stated as enhancement of clustering algorithm using a traffic dataset.

Objective
The objective of this research focuses on the data analysis for important attributes of accidents on highways in National Highway by implementing a clustering model. Thus, the problems taken for this research work is divided into some objectives which are as follows:

1. The accuracy and the efficiency are increased by comparing self developed program with the existing algorithm.
2. The objective is to show the advantage of clustering approach for examining the accidental analysis.

3. All the data are categorized into junior, adult and senior according to the driver's age. By validation, the clusters generated are tested accurate.
4. The error rate is reduced.
5. The objective of this project was to preprocess highway accidental data taken from traffic authority by doing surveys.

Proposed Algorithm

Specify the number of clusters to generate. Specify random number seed. Replace missing values in training instances holds the cluster centroids Holds the standard deviations of the numeric attributes in each cluster For each cluster, holds the frequency counts for the values of each nominal attribute attribute min values attribute max values

Keep track of the number of iterations completed before convergence Generates a clusterer. Has to initialize all fields of the clusterer that are not being set via options update centroids clusters an instance that has been through the filters Calculates the distance between two instances Returns the number of clusters.

Results

Table1 Comparison among K-Means and EPBC algorithms with Number of Clusters on National Highway Dataset

	EPBC Clustered Instances	Simple Kmeans Clustered Instances
Cluster 1	308	331
Cluster 2	112	117
Cluster 3	111	28
Cluster 4	174	100
Cluster 5	98	77
Cluster 6	63	213

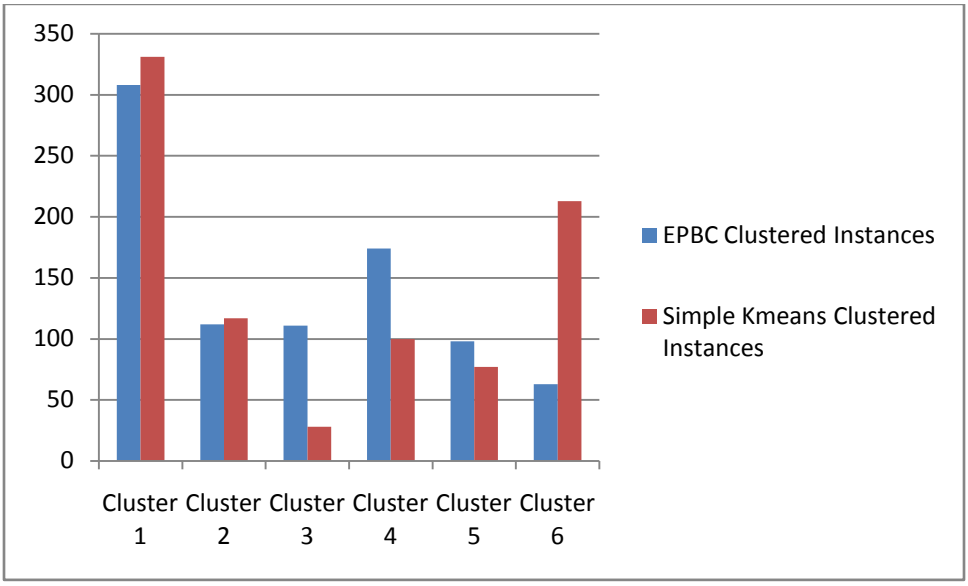

Figure 1 Graphical representation of inter-cluster similarity

Table2 Shows comparison of EPBC AndKMeans algorithms on error rate & No. of iterations

	EPBC	KMeans
Error Rate/Accuracy	73.793	125.597
No. of Iterations	20	32

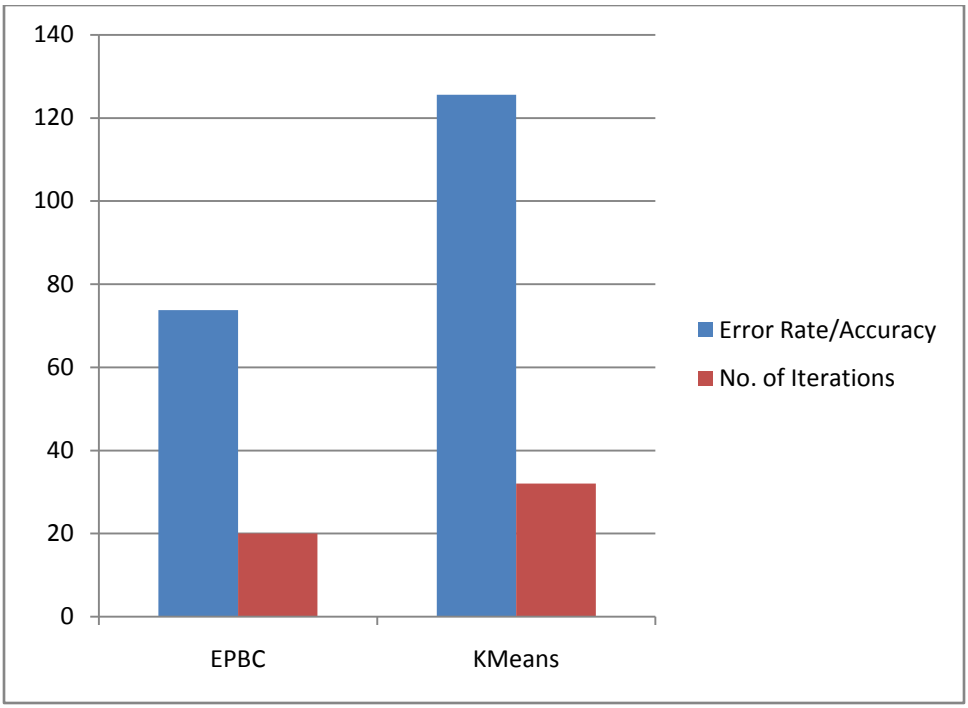

Figure 2 Graphical representation of Accuracy and No. of iterations between EPBC and KMeans

Conclusion

The proposed approach is used to analyze the causes and factors of highway accidents in NH-1 using data mining techniques. The objective of this project is to pre- process the data collected from traffic management by doing surveys during highway accidents and to show the advantage of using clustering approach for accidental analysis. A self-developed program is proposed which is based on clustering algorithm for accidental analysis, whose result is compared with the result obtained by original clustering algorithm. The accuracy and the reliability of a proposed program can be demonstrated by the results obtained from the comparison of the proposed program and the existing program. The analysis is carried through different aspects with regarding age, season and gender. This approach will given following advantages: First, it is developed for highway traffic accidents analysis that makes it easy to operate and highly sufficient while keep the size of the program small. Besides, since the program is self-developed, it can be easy to maintain and further enhanced according to users requirements.

References

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

[2] Mann, Amandeep Kaur, and Navneet Kaur. "Survey Paper on Clustering Techniques." *International Journal of Science, Engineering and Technology Research* 2.4 (2013): pp-0803.

[3] Shah, Saurabh, and Manmohan Singh. "Comparison of a time efficient modified K-mean algorithm with K-mean and K-medoid algorithm." *Communication Systems and Network Technologies (CSNT), 2012 International Conference on*. IEEE, 2012.

[4] Pakhira, Malay K. "A modified k-means algorithm to avoid empty clusters."*International Journal of Recent Trends in Engineering* 1.1 (2009).

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

[6] Agarwal, Jyoti, Renuka Nagpal, and Rajni Sehgal. "Crime Analysis using K-Means Clustering." *International Journal of Computer Applications* 83.4 (2013): 1-4.

[7] Nazeer, KA Abdul, and M. P. Sebastian. "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm." *Proceedings of the World Congress on Engineering*. Vol. 1. 2009.

[8] Goyal, M., and S. Kumar. "Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability." *Journal of the Institution of Engineers (India): Series B*: 1-6.