International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 06 | June -2017 www.irjet.net p-ISSN: 2395-0072

Survey of K means Clustering and Hierarchical Clustering for Road Accident Analysis

Akanksha Mahajan[1], ER. Neena Madan[2]

[1] *Research Scholar, Department of Computer Engineering & Technology, Guru Nanak Dev University Regional Campus, Jalandhar, Punjab, India-143001*
[2] *Assistant Professor, Department of Computer Engineering & Technology, Guru Nanak Dev University Regional Campus, Jalandhar, Punjab, India-143001*

Abstract-*Clustering is popular technique of data mining for analyzing road accidents. This paper presents the clustering techniques to analysis the accident locations and the related data accident at these locations .Accidents locations are divided into three categories high frequency accident location, ,moderate frequency accident location, low frequency accident locations. K-means clustering and hierarchical technique is used on road accident data. After that comparisons are done on the basis of parameter metrics.*

Keywords: **K-means, Hierarchical, Euclidian distance, WEKA Tool.**

1. INTRODUCTION

Clustering is an unsupervised learning technique .In this technique the objects in the data set are grouped into cluster such that groups are very different from each other [1]. Clustering is the grouping of similar objects and the cluster is the set of objects having some similar properties. In this case clusters are not predefined i.e. resulted clusters are not known before the execution of algorithm.

These clusters are made from dataset by grouping the objects in it .In this paper firstly K-means Clustering algorithm used whose results are compared with proposed technique Hierarchical Clustering on data set of road accident.

1.1. K-means Clustering Algorithm

K-means algorithm is a data mining algorithm which performs clustering. It divides the data set into a number of groups such that similar items fall into same groups .K means takes the number of desired clusters such as num cluster=4 and initial means as input by using k means ++ method in WEKA. Euclidean distance is chosen as distance function. It is an iterative process for clustering the dataset. For road accident analysis this data mining approach makes four clusters containing accident locations and the related information for particular accident. The resulted clusters provide the required information about accident. In order to divide the accident location into clusters k-means algorithm[2].It amounts to repeatedly assigning instances to the closest mean thereby using *Euclidean* distance **from attributes instances to a mean** .It is easiest to analyze huge data with clustering.

ALGORITHM

Input:[D,K] // D-> data set of road accidents ,k-> no of clusters ,m>mean for each cluster
Output: k clusters
Method:
1 Initially choose k clusters from data set means clustering the data values into k clusters.
2 Initially assign means to data set.
3 Calculate the distance between each data point and cluster centers using the Euclidean distance metric
4 Assign each value from data set to the k clusters having closet mean for each cluster.
5 Repeat: calculate the new mean from each resulted cluster.
6 Repeat: step 3.
4. until no. of iterations is reached and no changes during assignment of objects into corresponding clusters.

1.2. Data Set of Road Accident

In India, 108 is an emergency ambulance service which provides help to the accident victims. This emergency service is being operated in several states of India. Uttarakhand is an Indian state where this service is running. This service is operated and handled by which provides emergency service to the accident victims and also keeps track of record of every accident. This information is stored at their central server located at one particular place in the state. therefore, these data set provide information about accidents that have occurred in the road area of whole city, district and state. The data for this study have been obtained from GVK EMRI, (Emergency Management Research Institute) Dehradun. The data set consists of 54 records having 326 road accidents for 6 years period from 2009 to 2014, in Dehradun District of Uttarakhand State.

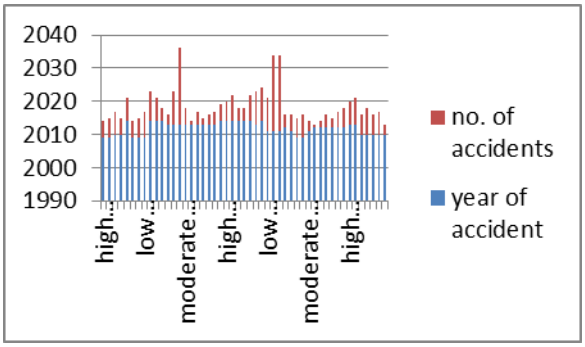

Figure: 1.1 Graphical Representation of Accidents.

1.3 Implementation of K-means Clustering on WEKA Tool.

In implementation choose the .Arff file compatible with WEKA for data set of Road Accident instances.

1. Open the WEKA explorer
2. Select the .ARFF file by open file from preprocess
3. Set class accident location(nom)
4. Visualize all
5. Choose cluster
6. Set num cluster =4
7. Set distance function i.e. Euclidian distance.
8. Select classes to evaluate cluster
9. Select the attribute accident location for class based evaluation
10. Then click on start for clustering.

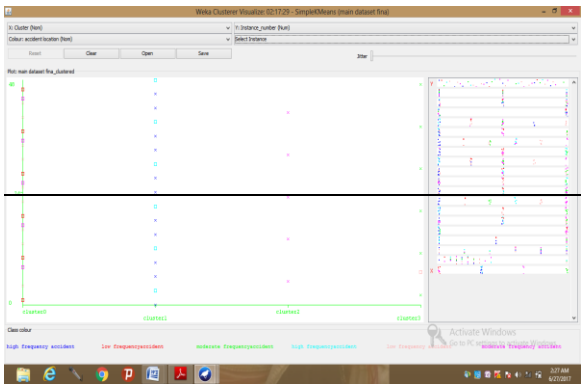

Fig: 1.2 Resulted Graphical Representation of Clusters

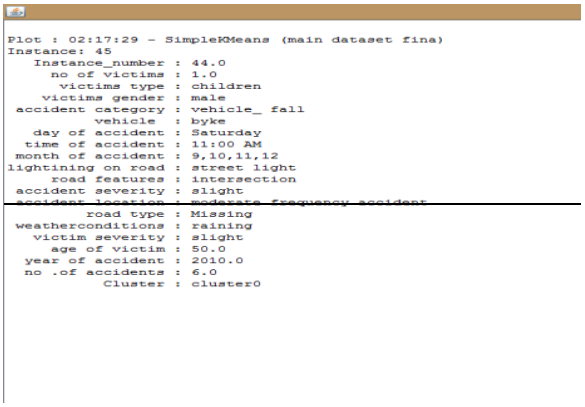

Fig:1.3 Information of Accident for instance 45

1.3. Results of K-means clustering mpl

The resulted cluster provided are accident locations and related information of accident. The correctly cluster instances i.e TP rate is 41%.The incorrectly cluster instances i.e FP false positive rate must be low for better performance of clustering [3] .It is 36.7347%.

2. Hierarchical Clustering Algorithm

In Hierarchical Clustering is a method of **cluster analysis** which build a **hierarchy** of clusters. Hierarchical means to arrange the attributes in order to their importance or rank. The data is not partition into a specific cluster in a single step. At the beginning, every point represents one cluster. The algorithm then finds the most similar cluster pairs and combines them into a single cluster. Similarity between clusters can be calculated by using one of the linkage criteria. There are two methods of hierarchical clustering algorithm .

1. gglomerative approach: where start from the bottom where all the objects are and going up i.e. bottom up approach through merging of objects. We begin with each individual objects and merge the two closest objects. The process is iterated until all objects are aggregated into a single group.
   Link used in agglomerative approach
   Link is used in this approach to sets the method used to measure their distance between two cluster.

Average link: This is the link choose for calculating average distance between items in cluster1 and items in cluster2.

2. ivisive approach: where we start with the assumption that all objects are group into a single group and then we split the group into two recursive until each group consists of a single object.

It Assign each instance to its own cluster. At each step the two clusters that are the most similar are merged the algorithm continues until all the clusters have been merged. Set num cluster=4 , Euclidean distance is chosen as distance function measures the distance between two instances or possibly the distance between instance and the centroid of a cluster depends on link type [4].link type use is average. For road accident analysis this data mining approach makes clusters of accident locations with respect to cluster num and instance number. The resulted clusters provide the required information about accident.

Algorithm of Hierarchical Clustering.

Given a set of items to be clustered, the basic process of hierarchical clustering start by assigning each item to a cluster, so that if there are N items, now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain

1. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now, one cluster less. Closest Distance is measured using average link.
2. Compute distances (similarities) between the new cluster and each of the old clusters.

Repeat steps 2 and 3 until all items are clustered into a single cluster.

2.1 Implementation of Hierarchical Clustering

This will shows the implementation results of Hierarchical Clustering Algorithm .This approach makes the clusters of Accident locations. Accident locations describes the three different locations for accident high frequency, low frequency, moderate frequency. It analysis the factors of road accident happened today[4].The another Clustering technique used for better analysis is hierarchical technique for this same data attributes is taken and loaded the .ARFF file in WEKA tool [12]

1. Open the WEKA
2. select the .ARFF file from open file
3. Set class accident location(nom)
4. Visualize all
5. Choose Hierarchical clustering
6. Take num cluster =4
7. Set the distance function Euclidean distance.
8. Set link type average for finding average distance between two clusters to merge.
9. Select Accident location as classes to cluster evaluation.
10. Then click start.

2.2 Results of Hierarchical Clustering

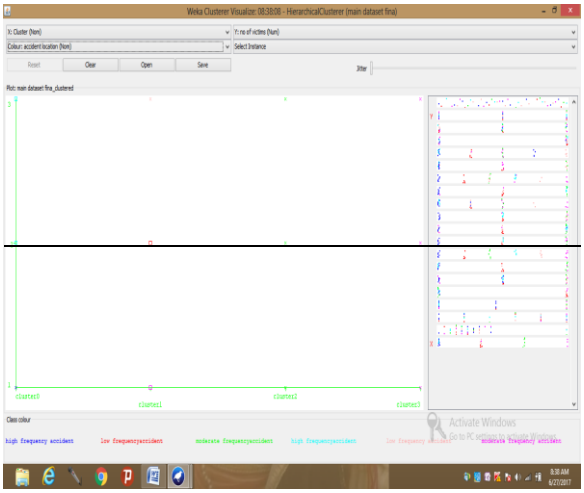

Fig: 2.1 Graphical Representation of Clusters

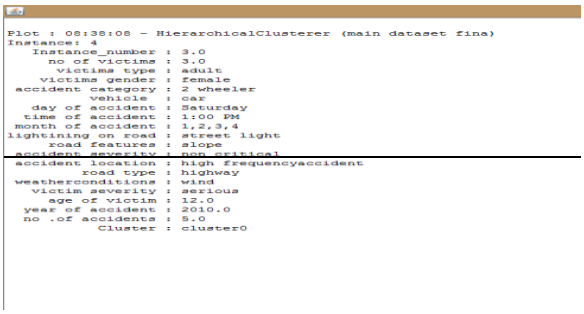

Fig:2.2 Information of accident for instance 4

The resulted cluster provides accident information with respect to accident locations. The correctly cluster instances i.e TP rate is 43%.The incorrectly cluster instances i.e FP false positive rate is 34.6939%.

3. CONCLUSIONS

Both the algorithms are compared on the basis of their performance analysis. Now, moving on performance analysis done basis of their parameters and check that which one provided better option for analysis of road accident with Clustering approach. For this comparison table is made to easily understand the better approach. The TP rate and FP rate of hierarchical clustering provides good results as compared to k means clustering for better analysis on data.

Table 3.1: Comparison table for approaches

Approaches	TP rate	FP rate
Simple K means Clustering	47%	36.7347%
Hierarchical clustering	43%	34.6939%

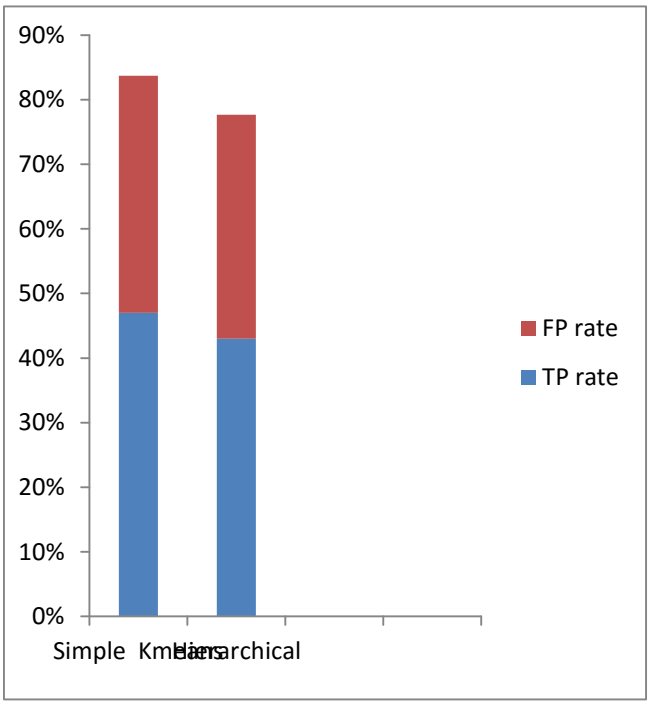

Fig: 3.1 Graphical Representation of Approaches.

4. FUTURE SCOPE

There are number of ways to the work in road Accident data by clustering. There can be improvement in work by implementing another techniques like DB scan for better results of accident analysis .It improves the cluster rates and accuracy further and determine information more accurately of the required data.

REFERENCES

[1] Jiawei Han," Data mining: concepts and techniques"(MorgaKaufmanPublishers, (2006).

[2] SachinKumar, Durga Toshniwal,"Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC)"Journal Big Data (2016) 3:13 DOI 10.1186/s40537-016-0046-3.

[3] Harish kumar Sagar, Varsha Sharma," Error Evaluation on K- Means and Hierarchical Clustering with Effect of Distance Functions for Iris Dataset."International Journal of Computer Applications (0975 – 8887) Volume 86 – No 16, January 2014.

[4] Sachin Kumar .Durga Toshniwal," A data mining approach to characterize road accident locations. J. Mod. Transport. (2016) 24(1):62–72 DOI 10.1007/s40534-016-0095-5.

[5] Ryan Tibshirani,"Clustering 2 Hierarchical Clustering". January 29 2013.

[6] Bharat Chaudhari, Manan Parikh."Comparative Study of clustering algorithms Using weka tools International Journal of Application or Innovation in Engineering&Management (IJAIEM) Volume 1, Issue2, October 2012 ISSN 2319 – 4847.