A Comparative Analysis of Road Accident Data Using Dataclustering Techniques

Bavya.A.S , Gayathri.R , Guhan.K
Dr. N.Shanthi,
Dept. of Computer Science &Engineering,
Kongu Engineering College,
Perundurai, Erode

***Abstract*:-**The provisions of roadways are resulting in convenience for the people but the main problem which is faced by the government of any country is that more frequent road accidents. Road accident data analysis is a very important means to identify various factors which come into account for road accidents that is design of the road, driver's impairment and vehicle design, etc., which can cause serious and most dangerous types of accidents.In this study,we are making use of k-modes, latent class clustering(LCC) and hierarchical clustering on a new road accident data from Leeds, West Yorkshire, England. Further, Frequent Pattern (FP) growth is applied on the clusters formed and entire data set. From the rules generated for latent class clustering, we found that LCC provides more accuracy than the other two clustering methods and hierarchical clustering provides more computational speed for the new data set.Based on the rules generated on each cluster and entire data set, we found that heterogeneity exists and clustering before the analysis gradually reduces heterogeneity. The rules for Leeds city affirms some important factors that can be used to formulate the disciplinary policies to prevent and overcome the accident rate.

*****

1 Introduction

An accident, also known as unintentional injury, is an undesirable, incidental, and unplanned event that could have been prevented had circumstances leading up to the accident been recognized, and acted Upon, prior to its occurrence. Most scientists who study unintentional injury avoid using the term "accident" and focus on factors that increase risk of severe injury and that reduce injury incident and severity. [4]

The reasons for these accidents are the extremely dense road traffic and then relatively great freedom of movement given to drivers. Accident involving heavy goods vehicles occur all too frequently despite calls responsible behavior, for respect of the loading regulations and the highway code, as well as the obligation for drivers to adapt their speed, which affects stopping distances, to the traffic and weather conditions. The prevention of road accident is also extremely important and will be ensured by strict laws, by technical and police controls, ongoing training for drivers and, if need to be, by legal and administrative penalties for those responsible.[4]

Accidents are complex events that involve a variety of human responses to external stimuli, as well as complex interactions between the vehicles, roadway feature, traffic related factors, and environmental conditions. In addition, there are complexities involved in energy dissipation that relate to vehicle design , impact angles, the physiological characteristics of involved humans, and other factors. With such a complex process, it is impossible to have access to all of the datathat could potentially determine the likelihood of a highway accident or its resulting injury severity. The absence of such important data can potentially present serious specifications problems for traditional statistical analyses that can lead to be biased and inconsistent parameter estimates, erroneous inferences and erroneous accident predictions. This problem is called the heterogeneity. Various statistical approaches available to address this unobserved heterogeneity are presented along with their strengths and weaknesses.[4]

In road accident data analysis, it is suggested that prior segmentation is very much useful in producing good results. Previously, Ulfarsson and Mannering (2004) and Islam and Mannering (2004), tried to group the data into homogeneous subgroups based on some methodologies. However Ona et al(2013) suggest that these factors can segment the data into the workable groups but this

cannot be guaranteed that subgroups will comprise of homogeneous groups of accidents. Therefore, data mining techniques likecluster analysis have been used to remove heterogeneity of road accident data. [4]

The authors claimed that clustering algorithms are very much useful to remove heterogeneity in road accident data and providing good results that can be used for accident preventive factors. The next task after prior segmentation of road accident data is to select a representative or target variable in order to perform the classification of the data. Most of the road accident data analysis work selects the severity or criticality of the road accident as a target variable for classification. [4]

In this paper, we are presenting a comparative study of the latent class clustering, k modes clustering and hierarchical clustering on a newly available road accident data of Leeds city of United kingdom. As both authors claim that their clustering technique is highly efficient in dividing the accident data into homogeneous groups. This paper attempts to identify which of the three techniques perform better on our new accident data set. Further, the association rule mining using Frequent Pattern (FP) growth technique is applied on the subgroups identified by k modes, latent class clustering and hierarchical clustering techniques to identify several factors associated with new road accidents in Leeds city.[4]

2 Methodologies

In this section, the paper explains the k modes, latent class clustering and hierarchical clustering technique for cluster analysis. Further, various cluster selection criteria are discussed followed by association rule mining technique using FP growth algorithm.

2.1 K-Modes Clustering:-

K-means clustering fail to handle datasets with categorical data since it minimize the cost function by calculating means. The K-modes clustering algorithm is based on k-mean pattern other than remove the numeric data limitation even as preserve its effectiveness.The k-modes algorithm used a simple matching similarity measure criterion for clustering of categorical data. [6]

2.2Latent Class clustering.-

A latent class model or generally a finite fixed model can be a thought of as probabilistic model for clustering. The goal of this algorithm is generally the same to identify homogeneous groups within a larger population. The main differences between latent class models and algorithmic approaches to clustering are that the former obviously lends itself to more theoretical speculation about the nature of the clustering, and because the latent class model is probabilistic, it gives additional alternatives for assessing model fit via likelihood statistics, and better captures in the classification.[3]

2.3 Hierarchical clustering:-

The general concept of the hierarchical fuzzy clustering is the partitioning of the data items into a collection of clusters. The data points are allocated membership values for each of the clusters. Many available clustering techniques have problems in managing extreme outliers but fuzzy clustering algorithms tend to give them very small membership degree in adjacent clusters. This algorithm is an improvement of fuzzy relational clustering algorithms. An expectation-maximization (EM) algorithm is an iterative process, in which the model mainly depends on some undetected latent/unseen variables. This algorithm is mainly used in finding maximum likelihood estimates of parameters. The EM algorithm's iteration alternates between performing an expectation (E). This step creates a function to compute the cluster membership probabilities and maximization (M) step, in which these probabilities are then used to re-estimate the parameters. These parameter estimates are then used to find out the distribution of the latent variables in the next E step. [7]

2.4 Cluster Selection:-

Cluster analysis is the process of segmenting the data set in to homogeneous groups of clusters. The primary requirements for any cluster analysis task are to find the number of clusters to form. Various approaches are exists in literature to identify the number of clusters e.g. gap statistic and different information criteria such as AIC, BIC and CAIC (Akaike, 1987; Raftery 1986; Fraley and Rafetery 1998). We have used both gap statistics and different information criterion to identify the number of clusters in our dataset and also to verify and validate the results off all approaches

The AIC, BIC, CAIC criterions can be calculated as follows,

AIC = − 2logL + 2p, ---------- (1)

BIC = −2logL + p log (n), ---------- (2)

CAIC = −2logL + p (log (n) + 1) ---------- (3)

Where p is the number of model parameters (Akaike 1987), n is the sample size. The values of AIC, BIC, CAIC are compared across severable possible cluster values.The gap statistic (Tibshiraniet al.2001), is a cluster identification method that can be used with any clustering technique. [5]

2.5 FP Growth techniques:-

The FP-Growth Algorithm, proposed by Han in, is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed information about frequent patterns named frequent-pattern (FP-tree). The FP-Growth Algorithm is an alternative way to find frequent item sets without using candidate generations, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree, which retains the item set association information. The frequent-pattern tree is a compact structure that stores quantitative information about frequent patterns in a database. After constructing the FP-Tree it's possible to mine it to find the complete set of frequent patterns. The biggest advantage found in in FP-Growth is the fact that the algorithm only needs to read the file twice, as opposed to apriori which reads it once for every iteration. [2]

2.6 Dataset:-

The data set used for this study is provided byhttp://data.gov.uk/,United Kingdom (2016).The data set covers data from Leeds, United Kingdom .In this paper,we haveselected 2500 road accident records with 15 accident attributes of Leeds district for data analysis.

Table 1k-modes clustering

clusters	Road class	Road surface	Lighting conditions	Weather conditions	Casualty severity	sex	Type of vehicle
C1	Unclassified	Dry	Daylight	Fine	Slight	Male	Car
	Unclassified	Dry	Darkness	Fine	Slight	Female	Car
C2	A	Dry	Daylight	Fine	Serious	Male	Pedal cycle
	A	Wet	Daylight	Fine	Slight	Male	Pedal cycle
C3	Motorway	Wet	Daylight	Fine	Slight	Female	Car
	Motorway	Wet	Daylight	Raining	Serious	Female	Car

Table 2 Latent Class clustering

clusters	Road class	Road surface	Lighting conditions	Weather conditions	Casualty severity	sex	Type of vehicle
C1	A	Dry	Darkness	Fine	Serious	Male	Motorcycle
C2	Motorway	Wet	Daylight	Fine	Serious	Female	Car
	Unclassified	Wet	Darkness	Raining	Slight	Female	Car
C3	A	Wet	Daylight	Fine	Slight	Male	Pedal cycle
	A	Dry	Daylight	Fine	Serious	Male	Pedal cycle

Table 3 Hierarchical clustering

clusters	Road class	Road surface	Lighting conditions	Weather conditions	Casualty severity	sex	Type of vehicle

C1	Unclassified	Dry	Daylight	Fine	Slight	Male	Car
	Unclassified	Dry	Darkness	Fine	Slight	Female	Car
C2	Motorway	Wet	Daylight	Fine	Serious	Female	Car
	A	Wet	Daylight	Fine	Slight	Male	Pedal cycle
C3	Unclassified	Wet	Darkness	Fine	Slight	Female	Car
	Motorway	Wet	Daylight	Fine	Slight	Female	Motor cycle

3 Results and Analysis

3.1 Cluster Analysis:-

The primary task of cluster analysis is to determine the number of clusters that can be formed in the data set. We generated the different models for cluster 1 to cluster 4 using gap statistic and AIC, BIC and CAIC Criteria. [5]

3.2 Comparison of k-modes, Latent class and hierarchical clustering

In this study, we attempted to use k-modes, latent class clustering and hierarchical clustering technique to segment the real world road accident data. The clusters formed by the clustering techniques have different number of data instances. LCC uses maximum likelihood technique to measure the probability of data objects (Table 1), k-modes technique uses similarity (Table 2) whereas hierarchical uses entropy measure (Table 3) in order to form clusters.

3.3 Rule Mining:-

This paper uses an association rule mining technique. This is a variation of an algorithm for mining all association rules from a database. Traditionally, association rue mining is performed by using two interestingness measured named the support and confidence to evaluate rules. The input we given here is a road database and three thresholds named minsup (a Value between 0 and 1), minsconf (a value between –infinity to +infinity). The output of this algorithm is a set of all association rules that have a support, confidence and lift respectively higher than minsup, minsconf and minlift. [1]

4. Conclusion and suggestion

This paper presents a comparative study of k-modes, LCC and Hierarchical clustering on a new road accident data from Leeds, West Yorkshire, England. This study uses 2500 road accident records of Leeds city from data.gov.uk, United Kingdom (2016). The number of attributes that has been used in the analysis was 15 which were analyzed. Based on the results obtained from k-modes, LCC, and Hierarchical clustering methods the clusters were selected. On analyzing the clusters selected by clustering techniques, it is found that different number of road accidents was grouped in each cluster. In order to generate association rules, FP growth technique is applied on each selected cluster and entire dataset to find the relationship between the different values in the data. The rules generated by FP growth technique have different confidence and lift values but there are no deviations among the rules. For Leeds city's accident data we found that the purity of k-modes and hierarchical clustering were same but lesser than the purity of Latent class Clustering. On the other hand, the computational speed of Hierarchical clustering is faster than the other two clustering methods. Also the association rules generated provides information and reveals the factors associated with the various types of road accidents. With the help of these rules, we can also predict that causes of road accident in one area may be similar for the adjacent areas. In addition to that, it can be used to formulate the disciplinary policies to prevent and overcome the accident rate.

References

[1] R.Agarwal and R.Srikanth. Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994. Fast algorithms for mining large association rules in large datasets.
[2] Han, Data Mining Algorithms In R/Frequent Pattern Mining/The FP-Growth Algorithm proposed by Han

[3] KalalGayatriPratip, D.R.patil, Department of Computer Science, SES's, R.C.P.I.T, Shirpur, Maharashtra, India. Clustering vs. Latent class clustering Analysis- Differences
[4] Road Accident-Causes and Prevention techniques www.icdo.org/en/disasters/man-made-disasters/transport-accidents/road

[5] Sachinkumar, DurgaToshnwal, Manoranjanparida. A Comparative analysis of heterogeneity in road accident data using data mining techniques
[6] Samrat Ashok Technological Institute, Department of IT, Vidisha, India. K-modes Clustering Algorithm for Categorical Data Neha Sharma and Nirmal Gaud
[7] Summarization of sentences using Fuzzy and Hierarchical clustering
[8] Unobserved Heterogeneity and Statistical Analysis of Highway Accident data Fred Mannering (flm@usf.edu)