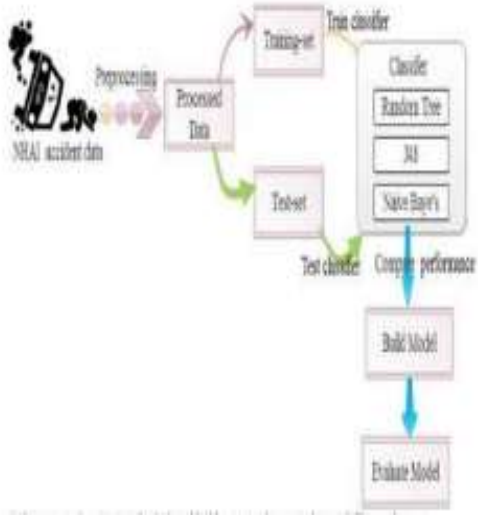
A Statistical Analysis of Road Accident Data -Using Data Mining Techniques

Ram Prasanth T*, Spanglar Diaz V*, Surendran N*, Udhayavel V*, Dr. C. Anand**,Mrs. N. Vasuki***
*Final year, Department of Computer Science and Engineering
**Associate Professor, Department of Computer Science and Engineering
K.S.R College of Engineering , Tiruchengode.
***Assistant Professor, Department of Computer Science and Engineering IRTT , Erode.

1.Abstract:- **Road accidents are one of the most imperative factors that affect the untimely death among people and economic loss of public and private property. Road safety is a term associated with the planning and implementing certain strategy to overcome the road and traffic accidents. Road accident data analysis is a very important means to identify various factors associated with road accidents and can help in reducing the accident rate. The heterogeneity of road accident data is a big challenge in road safety analysis. In this, we are making use of latent class clustering (LCC) and k-modes clustering technique on a new road accident data. Initially,the LCC and k-modes clustering technique are applied on road accident data to form different clusters. Frequent Pattern (FP) growth technique is applied on the clusters formed and entire dataset (EDS). However, in this certain techniques these are well suited to remove heterogeneity of road accident data. The generated results for each cluster and EDS proves that heterogeneity exists in the entire dataset and clustering prior to analysis certainly reduces heterogeneity fromthe dataset and provides better solutions.**

SYSTEM ARCHITECTURE



Keywords

Road accident analysis
Heterogeneity
Data mining
Clustering
FP growth

2.INTRODUCTION

The increasing number of road and traffic accidents is a challenging issue to the transportation systems. It not only concern with health issues but also associated with economic burden on the society. Therefore, it is an important task for the safety analysis to carry out a comparative study of road accidents to identify the factors that causes an accident to happen, so that preventive actions can be taken to overcome the accident rate and severity of accidents consequences. Therefore, an comparative study of road accident data is required to identify the several factors associated with road accidents. The main concern with road accident data analysis is to identify the most influential factors affecting road accident frequency and accident severity. The major problem with road accident data analysis is its heterogeneous in nature.

Heterogeneity in road accident data is highly undesirable and unavoidable. The major disadvantage of heterogeneity of road accident data is that certain relationships may remain hidden such as certain accident factors associated with particular vehicle type may not be significant in entire data set the enormity of the effect of certain accident related factors may be different for various conditions severity levels for an accident contributing factors may be different for different accident types. In order to get more accurate results this heterogeneity of road accident data must be removed to deal with this heterogeneous nature of road accident data, divide the data into groups based on some exogenous attributes e.g. accident location, road condition, cause of accident and analyzed every group separately to identify several influential factors associated with road accidents in each group. However, this method is unrealistic as grouping the data based on certain attributes may results in less important groups. Some subgroups can have large number of samples and some can have very low number of samples and thus restricting the application of accident severity models. Other choice is to use some approach such as data mining in order to remove the heterogeneity of the road accident data.

Data mining is an evolutionary technique which has been used in the field of transportation systems. Transportation safety is one of the important areas of transportation systems which are actively involved in safety from road and traffic accidents on road. Various techniques on this issue have been done previously using traditional statistical techniques.

However, traditional statistical techniques have their own assumptions regarding dependent and independent data attributes. Any wrong selection of these attributes can lead to incorrect outcomes. Data mining techniques include clustering, classification, association rule mining and detection. Clustering techniques such as k-means clustering, k-modes clustering and hierarchical clustering are very popular algorithms in several domains.

In road accident data analysis, it is suggested that prior segmentation is very much useful in getting good results. However these factors can segment the data into workable groups but this cannot be guaranteed that the subgroups will comprise of homogeneous group of accidents.

2.1 METHODOLOGY

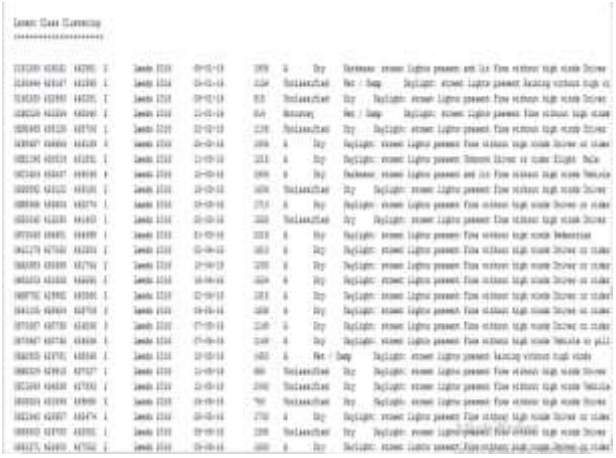
In this methodology,it explains the k-modes and latent class clustering technique for cluster analysis. Further, various cluster selection criteria are discussed followed by association rule mining technique using FP-growth algorithm.

K-Modes clustering approach is an enhanced version of traditional k-means algorithm with an amendment of distance measure, iteration process and cluster center representation. k-Mode clustering is mainly proposed to analyze categorical dataset. The k-modes algorithm uses a simple matching similarity measure criterion for clustering of categorical data. Let A and B be two qualitative data objects categorized by x categorical attributes. The simple similarity matching criterion between A and B is the number of matching attribute values of the two data objects. The greater the number of matches is, more the similarity of two objects. Unlike k-means algorithm, k-modes algorithm uses modes instead of means for clustering purpose. The k-mode algorithm is quite efficient in handling large categorical data.

2.2 Latent class clustering

LCC technique is a cluster analysis techniques widely used technique for the segmentation of road accident data.LCC is a probability based cluster analysis technique. The objects in each cluster formed by LCC are assigned to that cluster based on probability measures with maximum likelihood technique.

LCC is different from other clustering techniques as it is available to be used with any type of data variables such as qualitative, quantitative or mixtures of each. LCC does not require any prior standardization that affects the results. Several statistical criteria such as Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and Consistent AIC (CAIC) are available to be used with LCC.



2.3 Number of cluster selection

Cluster analysis is a process of segmenting the data set into homogeneous groups of clusters. The primary requirements for any cluster analysis task to find the number of clusters to form. Various approaches are exists in literature to identify the number of clusters.

2.4 FP growth technique

Association rule mining is a popular data mining technique.The major problem with Apriori algorithm is that it uses candidate item set generation and then tests whether these item sets are frequent or not. Apriori algorithm is computationally expensive as it requires multiple database scans in order to generate candidate sets. The another association rule mining technique is FP-growth algorithm. The difference between FP growth and Apriori is that it is computationally faster than Apriori as it does not require candidate generation. FP growth algorithm uses a special data structure known as FP tree, which preserve the itemset association information.Like, support, confidence and lift interesting measures to extract strong rules from data set.

2.5 Data set

Data set is a collection of data most commonly a data set corresponds to the contents of a single database table or a single statically data matrices, where ever column of the table

represents a practical variable and each row corresponds to a given number of data sets in questions. The data set lists values for each of the variables should as the height and weight of the object for each number of the data sets each value is known as a datum.

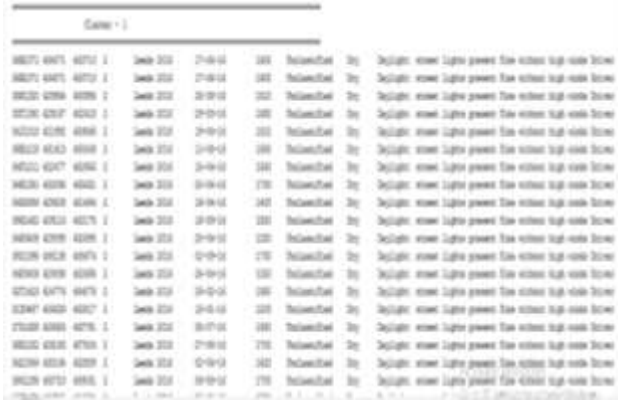
3. CLUSTER ANALYSIS AND RULE MINING

3.1 Cluster analysis

The primary task of cluster analysis is to determine the number of cluster that can be formed in the dataset.

These k-modes and LCC have been used to form homogeneous segments from the road accident data.

The clusters identified and their size is executed. We formed all four clusters using both LCC and k-modes clustering algorithm. The number of data instances assigned to their respective clusters was different for LCC and k-modes. The cluster 1 consists of two wheeler accidents that were happened at intersection on both highway and local roads.



3.2 Rule mining

Association rule mining is a popular technique that is used to identify the correlation between values of different attributes for a data set. FP growth algorithm is an association rule mining technique which is computationally efficient than Apriori algorithm.

The rules generated of each cluster and EDS using FP growth algorithm were analyzed and it was found the methods for clusters obtained from LCC and k-modes was though similar with different interest values.

4.CONCLUSION AND FUTURE WORK

4.1 CONCLUSION

These methods used to classify the road accidents based on the severity of accidents using the clustering techniques. The information criteria and gap statistic are used to identify the number of clusters to be made. Based on the results obtained from cluster selection criteria four clusters C1–C4 were identified by k-modes and LCC. The clusters identified by these techniques have different number of road accidents in each cluster. Further, FP growth technique is applied to each cluster and EDS to generate association rules which can define the correlation between the values of different attributes in the data. There is no major difference found in the association rules generated by FP growth algorithm except that the rules have different confidence and lift value for the clusters formed by k-modes and LCC.

4.2 FUTURE WORK

By means of this technique we provided a conclusion that k-modes are better than LCC on categorical data, we do not found any differences that shows that k-modes are better than LCC especially in road accident data except computational speed. There is no doubt that the cluster analysis technique performs well in reducing the heterogeneity of road accident data. So this type of analysis can be very useful to overcome the accident rates in various places at the same time. It acts as an effective means to make the preventive measures for the road safety and to save human lives. This kind of research is conducted to improve the safety and security of humans. There exists a good trade-off between the security and efficiency. Further research could be conducted in this domain to make this system more efficient. However,we cannot call a system that can be 100 percent secure .

REFERENCES

1. Abdel-Aty MA, Radwan AE.(2000) Modeling traffic accident occurrence and involvement.
2. Abellan J, Lopez G, Ona J,(2013) Analysis of traffic accident severity using decision rules via decision trees.
3. Chang LY, Chen WC,(2005) Data mining of tree based models to analyze freeway accident frequency.
4. Chen W, Jovanis P.(2002) Method of identifying factors contributing to driver-injury severity in traffic crashes.
5. Depaire B, Wets G and Vanhoof K,(2008) Traffic accident segmentation by means of latent class clustering, accident analysis and prevention.
6. Han J, Kamber M.(2001) Data Mining: Concepts and Techniques. USA: Morgan Kaufmann Publishers.
7. Joshua SC, Garber NJ.(1990) Estimating truck accident rate and involvements using linear and poisson regression models.
8. Karlaftis M, Tarko A,(2015) Heterogeneity considerations in accident modeling.
9. Kashani T, Mohaymany AS, Rajbari A,(2012) A data mining approach to identify key factors of traffic injury severity, pro-metraffic & transportation.
10. Ma J, Kockelman K,(2006) Crash frequency and severity modeling using clustered data.
11. Maher MJ, Summersgill I.(1996) A comprehensive methodology for the fitting of predictive accident models.
12. Ona JD, Lopez G, Mujalli R, Calvo FJ (2013) Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks.
13. Raftery AE (1986) A note on Bayes factors for log-linear contin- gency table models with vague prior information.
14. Rovsek V, Batista M, Bogunovic B,(2016) Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree.
15. Sasidharan L, Wu KF, Menendez M (2015) Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland.
16. Savolainen P, Mannering F, Lord D, Quddus M,(2011) The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives.
17. Tan PN, Steinbach M, Kumar V. (2006) Introduction to data mining. Pearson Addison-Wesley.
18. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic.
19. Ulfarsson GF, Mannering FL (2004) Difference in male and female injury severities in sport-utility vehicle, minivan, pickup and pas- senger car accidents.
20. Vermunt JK, Magidson J (2002) Latent class cluster analysis. In: Hagenaars JA, Mc Cutcheon AL Advances in latent class analysis.