Web Data Mining Framework for Accidents Data

Gowtham Mamidisetti, Nalluri Gowtham, Ramesh Makala

***ABSTRACT*--- Identification of factors associated with large amount of data is the main key challenge in big data analysis. Heterogeneous nature of data is other factor that makes the analysis difficult. Accident occur due to various factors like poor lighting, un controllable speed at curves, hill region with unidentified climate change, fog, vehicle bad condition, driver health status. Data recorded for these above factors are considered under analysis using segmentation and clustering methods. Data analysis is done on the accident data to find differences in traffic conditions, weather conditions and road conditions. A research on reasons behind the accidents and impact of public health on accidents data is presented in this work. Segmentation of accident data is done with k-mode and associate rule mining. Trend Identification with similarity analysis approach is used in analyzing road accident data. This papers focuses on finding best analysis model for accident data analysis and also to find the combination of methods required to predict influenced factors that need to be focused to reduce impact of health care on accidents.**

Keywords—K-modes; Latent Class Analysis; Association Rule Mining; Trend Analaysis.

I. INTRODUCTION

Road accidents are unpredictable incidents so to analyze that it requires factors that influence the accident. Generally the road accidents have set of variables that are discrete in nature. The problem in analysis of accidents data is its heterogeneous nature between the factors. Thus this heterogeneity in data should be considered otherwise some relation between data may be hidden.

One of the important techniques of data mining is clustering of data. This cluster analysis is used to achieve goals of preliminary tasks. Clustering over accident data can produce different clusters and that cluster results are analyzed using negative binomial (NB).

To identify the connection between accidents and the factors influencing them, the most popular technique used is Regression analysis. There are various regression models available to choose. By finding the relation between accidents and factors influencing them, traffic engineers can identify the danger zones of accidents and can provide better facilities to reduce the impact of such factors on accidents. This motivates for the need of better approaches to research on accident data. Data mining gives the scope to research on this area by providing various methods of classification, clustering and analysis. Applying adequate methods of data mining to analyze the road safety can be considered as one of the emerging trends. There is continuous need in this area to improve the analysis and thereby safeguarding the road safety.

This paper proposes to study and analyze the accidents data that helps to predict the reasons behind accidents and its causes. It is done by forming the accident data set into different clusters using K-modes algorithm that internally uses latent class clustering and forming different relations among the accidents using association rule mining. Further trend analysis also performed on the data that assists in identifying of accidents rate in different aspects.

II. PROPOSED FRAMEWORK

The propose approach combines various methods of data mining to analyze accident data in a better way producing some key relationships between the accidents and factors of influence.The proposed framework is processed in four steps.

1. Data preprocessing
2. Clustering
3. Association rule mining
4. Trend analysis

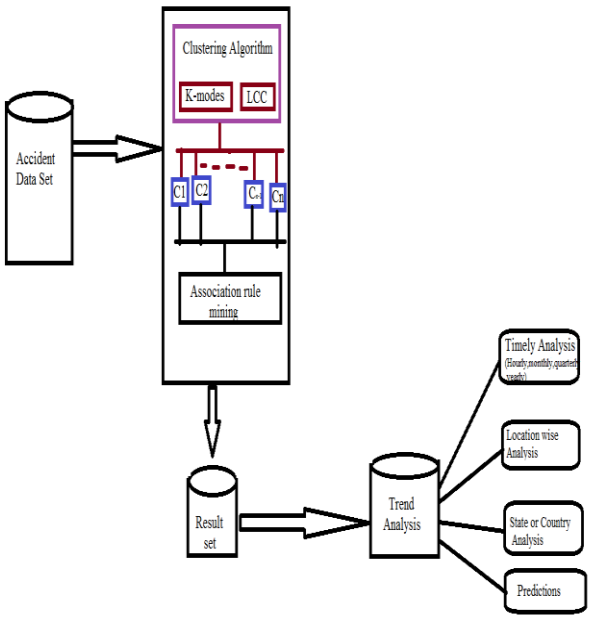

Fig. Proposed Framework

Revised Manuscript Received on February 11 , 2019.

Gowtham Mamidisetti, Department of Computer Science and Engineering,Presidency University,Bangalore,India. (E-mail: mamidisetti.gowtham@gmail.com)

Nalluri Gowtham, Department of Information Technology, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India.

Ramesh Makala, Department of Information Technology, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India. (E-mail: mrameshmailbox@gmail.com)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

1. Data preprocessing

Data preprocessing is starting step in the data mining. Data preprocessing is done to remove noise or outliers present in the dataset. Data preprocessing is also used to handle the missing values in the records and removing irrelevant attributes in the data set. In this step, the aim is to preprocess the accidents dataset by identifying the main attributes that are used for analysis and to handle or replace the values for missing values in the records.

2. Clustering algorithm

The goal of clustering algorithm is to divide the whole data into different groups which are formally known as clusters. Each cluster can be defined as a data of similar items. There can be any number of clusters for a dataset. Any two clusters generally does not have any similar features. In this paper the proposed clustering algorithm is K-Modes algorithm which is an enhanced version of K-Means algorithm. On k-modes algorithm Latent Class Clustering (LCC) is done using Latent Class Analysis (LCA) in order to identify similarity between clusters and possibility of mixing clusters to make single clusters. The latent class analysis (LCA) gives probability of making a cluster.

The algorithm used in the proposed approached follows the following steps

Step1 Generate a random number which represents the clusters to be formed. Same random number can also be represented for identifying the modes. Let the random number be named K.

Step2 Find the dissimilarity between each mode formed and remaining data points.

Step3 For each mode assign the data points having the lowest dissimilarity.

Step 4 Update the modes till the assigning of data points to the modes becomes stable.

Step5 Perform Latent class clustering to minimize the clusters.

The reason for choosing k modes as part of clustering algorithm is that, k modes algorithm is good at handling multiple attributes with categorical values. As accident data consists of large number of attributes which are categorical in nature, k modes helps best in finding the clusters. The reasons to include latent clustering is that it further helps in minimizing the clusters after k modes. As we are dealing with accident data there will be multiple factors of influence, so minimizing the clusters to maximum will help in forming qualitative results.

3. Results & Disscussions

To find the factors which influence the accidents most, there is need of mining algorithm which can extract the relation between various factors in the data. For this purpose association rule mining serves the best. It is capable of uncovering many hidden relations in the data set. By applying this methodology of mining the influence of specific factor on accident data can be found by aiming at the frequency rate of occurrence of that specific factor along with accident. Relationship between factors which occur together can also be extracted.

Apriori algorithm is one of the best among the association rule mining algorithms to extract all the hidden relations in the given data set. Apriori algorithm is used in this approach to the frequently influencing factors set in the accident data and association between various factors. Apriori algorithm works in various stages of finding frequent factors set starting from one factor in the set influencing the accident to set of factors as large as possible which are occurring together and influencing the accidents. If a factors set is to be said as frequently occurring, all its subsets should also be frequent. Such factors sets are finally extracted as frequent factor set influencing the accidents.

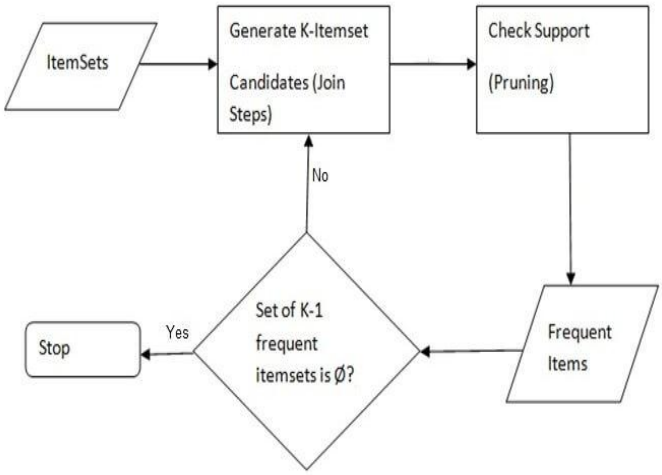

Fig Apriori algorithm

The relationships or associations which are extracted are not accepted as the results needed. The factor sets finally obtained are tested under some interesting measures to find their standard of importance. Support and confidence are two values which are for this purpose. To measure the standard of rules obtained other than confidence and support values a value called Lift can also be calculated which serves the same purpose. The final result after all these association rule mining and calculating the values of standard are the set of factors which are occurring together in accident data set that are indirectly responsible for occurrence of accidents.

4. TREND ANALYSIS

Road accidents are the important reasons for death rate in the countries especially in India. Trend analysis is the concept in which it measures the trend of the data. Trend analysis in proposed approach is done using cophenetic correlation coefficient method and agglomerative hierarchical clustering method.

The rate of accidents can be identified using trend analysis. The accidents rate analysis can be found for a year or a month or a day or a particular location or a particular period of time using trend analysis. Trend analysis plots different relations and rate of accidents in that relations which helps for prediction and prevention of accidents.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

III. CONCLUSION

The proposed approach is the combination of clustering, association rule mining and trend analysis which are performed in sequence of steps in the proposed framework. The whole data set is divided into different clusters using K-modes algorithm and LCC (Latent Class Clustering). Association rule mining is performed on that clusters to identify different associations among data in clusters respectively. On the result data the trend analysis is performed using Cophenetic correlation coefficient (CPCC) and Agglomerative Hierarchical Clustering method. The proposed framework results in extracting the factors which influenced the accidents more. This helps in concentrating on the factors to reduce their influence on accidents. Trend analysis of the framework helps in extracting or predicting the chances of occurrence of accidents due to specific factors. Predications can done based on factors, location or even time of accident. The accidents data is analyzed and came to know respective reasons for accidents and helps to predict and reduce the accidents causes and its rate by using the proposed framework.

IV. REFERENCES

1. Sachin Kumar, DurgaToshniwal ., " A data mining framework to analyse road accident data" , Journal of Big Data , a springer open jounal (2015) 2:26 ,DOI 10.1186/s40537-015-0035-y.
2. Sachin Kumar, DurgaToshniwal ., " Analysis of hourly road accident counts using hierarchial clustering and Cophenetic correlation coefficient (CPCC) " , Journal of Big Data , a springer open jounal (2015) 3:13 ,DOI 10.1186/s40537-016-0046-3
3. " Mining of massive Datasets" by AnandRajaraman and Jeffrey David Ullman , Printed by Cambridge University Press.
4. http://www-01.ibm.com/software/data/bigdata/
5. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication