Eleyon
SOUTH ASIAN JOURNAL OF ENGINEERING AND TECHNOLOGY

Full Length Article

A Comparative Analysis of Heterogeneity In Road Accident Data Using Data Mining Techniques

Anitha.S[a], Keerthana.R[a], Pavithra.B[a], Dr.M.Sangeetha[b]

[a] *Student, Department of IT, K.S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India*
[b] *Assistant Professor, Department of IT, K.S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India*

***Corresponding Author**
M.Sangeetha

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. In this study, we are making use of K-means clustering on a new road accident data from Alabama, America. The main focus to use these techniques is proving those identify technique and can perform better than other, retrieve all the information about road accident which classify on basis of attributes like state, city, weather etc., This will be useful in reducing the road accidents.

Keywords: K means clustering

1. Introduction

The age often referred to as the information age .In this information age, we believe that information leads to power and success, and sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures became overwhelming. This initial chaos leads to the creation of structured Databases and Database Management Systems (DBMS). The efficient database management system is very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever we want. The proliferation of database management systems is also being contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making.

Confronted with huge collections of data, we have now created new needs to help us make better managerial choices.

These needs are automatic summarization of data, extraction of the “essence” of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as Knowledge Discovery in Databases refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

2.Literature Survey

2.1 modeling traffic accident occurrence and involvement.

Abdel-Aty MA and Radwan AE (2000) (1) has proposed modeling traffic accident occurrence and involvement. In this paper, they introduced a macroscopic model for road traffic accidents along highway sections. They discussed the motivation and the derivation of such a model, and present its mathematical properties. The results are presented by means of examples where a section of a crowded one-way highway contains in the middle a cluster of drivers whose dynamics are prone to road traffic accidents. They discussed the coupling conditions and present some existence results of weak solutions to the associated Riemann Problems. Furthermore, we illustrate some features of the proposed model through some numerical simulations.

2.2 Data mining application in transportation engineering

Barai S (2) has proposed Internet survey which may be one of the effective means to collect big data from the real world. Collected data may realize meaningful analysis of targeted field. Intelligent Transportation is one of the smart city applications which bring us safety driving as well as comfortable driving by mitigation of the traffic congestion. This study proposes an example of vehicle infrastructure cooperative function which would be incorporate into vehicle safety system for smart city application.

2.3 K-modes clustering

Chaturvedi A, Green P and Carroll J (3) has proposed a new pixel unsupervised hyper spectral image (HSI) segmentation method. It relies on a binary in coding of spectral reflectance curve variations of pixels that allows considering HSI segmentation as a clustering problem in the feature set of binary strings. Using a generalized Hamming distance, a k-modes algorithm is applied to obtain a

2.4 Method of Identifying Factors Contributing To Driver-Injury Severity In Traffic Crashes

In this work et.al(4) Chen W, Jovanis P has proposed The objective of this study is to evaluate a set of variables that contribute to the degree of injury severity sustained in traffic crashes of Korean expressways. To this end, we examined three statistical models – ordered probit, ordered logit, and multinomial logit – to determine the most appropriate model for crash records that were collected from the entire network of Korean expressways in 2008. Interpretation of the estimated coefficients in the selected model provides relative risks of significant influential factors for injury severity. The findings from this study are expected to help transportation planners and engineers understand which risk factors contribute more to the injury severity in Korean expressways such that they can efficiently allocate resources and effectively implement safety countermeasures.

2.5 Profiling Of High Frequency Accident Locations By Use Of Association Rules.

In this work et.al(5) Geurts K, Wets G, Brijs T, Vanhoof K has proposed In belgium, traffic safety is currently one of the government's highest priorities. Identifying and profiling black spots and black zones in terms of accident related data and location characteristics must provide new insights into the complexity and causes of road accidents which, in turn, provide valuable input for government actions. In this paper, association rules are used to identify accident circumstances that frequently occur together at high frequency accident locations.

2.6 Mining frequent patterns without candidate generation. in: proceedings of the conference on the management of data

In this work et.al(6) Han J, Pei H and Yin Y has proposed Mining frequent patterns in transaction databases, time- series databases, and many other kinds of databases has been studied popularly in data mining research. Most of the previous studies adopt an Apriority-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there exist prolix c patterns and/or long patterns. In this study, we propose a novel frequent pattern tree (FP-tree) structure, which is an extended pre x- tree structure for storing compressed, crucial information about frequent patterns.

2.7 Estimating truck accident rate and involvements using linear and poisson regression models

In this work et.al (7) Joshua SC, Garber NJ has proposed The goal of this study was to conduct a comparative evaluation of accident rates and patterns for male and female passenger automobile drivers. Two sections of road in Israel, one urban and one rural, were selected for the study. Counts of passenger automobiles by sex of driver were carried out on each section of road. The relative accident rates for male and female drivers on the two roads were assessed by estimating the relative exposure of the two groups and matching it with relative accident frequencies. Accident patterns In terms of severity and type were also compared for the two sexes.

2.8 Heterogeneity considerations in accident modeling.

In this work et.al(8)Karlaftis M, Tarko A has proposed Clustering and classification approaches have been commonly applied in reducing the heterogeneity in accident data. As part of an effort to understand the features of the heterogeneity, this study assessed accident data from the perspective of accident occurrences. Using the rule-based classification method, rough set theory,

rules were derived which consisted of indispensable factors to certain accident outcomes and reflected the process of accident occurrences. The occurring frequency of each derived rule was then adopted as the basis for grouping accidents for further analyses. Empirical results showed that rules with high occurring frequencies were largely related to drivers with high-risk characteristics.

2.9 A data mining framework to analyze road accident data

In this work et.al(9) Kumar S, Toshniwal Dhas has proposed a data mining framework to analyze road accident data .Road accident is one of the crucial areas of research in India. A variety of research has been done on data collected through police records covering a limited portion of highways. The analysis of such data can only reveal information regarding that portion only; but accidents are scattered not only on highways but also on local roads. A different source of road accident data in India is Emergency Management research Institute (EMRI) which serves and keeps track of every accident record on every type of road and cover information of entire State's road accidents. In this paper, they have used data mining techniques to analyze the data provided by EMRI in which they first clustered the accident data and further association rule mining technique is applied to identify circumstances in which an accident may occur for each cluster.

2.10 A data mining approach to characterize road accident locations

In this work et.al(10) Kumar S, Toshniwal D (2016) has proposed Data mining that has been proven as a reliable technique to analyze road accidents and provide productive results. Most of the road accident data analysis use data mining techniques, focusing on identifying factors that affect the severity of an accident. However, any damage resulting from road accidents is always unacceptable in terms of health, property damage and other economic factors. Sometimes, it is found that road accident occurrences are more frequent at certain specific locations.

2.11 Analysis of traffic accidents on rural highways using latent class clustering and bayesian networks.

In this work et.al(11)Ona JD, Lopez G, Mujalli R, Calvo FJ (2013)has proposed Traffic accidents are contingent events and analyzing them requires awareness of the particularities that define them. In general, accidents are defined by a series of variables generally discrete variables that explain them. Once the nature of the variables is known, researchers select the method that is most appropriate for developing and implementing the best statistical models for analyzing the data in each case one of the main problems of accident data and their modeling process is their heterogeneity.

2.12 Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes.

In this work et.al(12) Savolainen P, Mannering F has proposed Crash injury severities are recorded in ordinal scales with fatal crashes ranking highest in the scale and property damage minor crashes ranking lowest in the scale. Various researchers have attempted to model injury severity outcomes by taking injury severity levels as either simple categorical variables or ordered categorical variables. The severity level of crashes vary with the collision partners, crash time, roadside activity characteristics and road inventory characteristics. Collision partner is the key attribute in determining the severity outcome of crashes, but severity of crashes vary even if the collision partners remain same.

2.13 Difference in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents.

In this work et.al(13) Ulfarsson GF, Mannering FL has proposed Mass of the striking vehicle is a factor in the prediction of accident severity. The accident data used in this study did not allow distinguishing whether this 'mass' aspect contains hidden stiffness and geometrical aspects suchas bonnet height and bumper height, due to high correlation between mass, stiffness and geometrical aspects. With respect to fatality there is tendency towards a slightly better crash protection for the Sports Utility Vehicles (SUV) driver and his passengers, than for the driver and passengers of a 'normal' passenger car. Sports Utility Vehicles (SUV) occupants seem to be more frequently not injured in a crash. This might indicate a safer environment for the Sports Utility Vehicles (SUV) occupant, but it is most probably due to the higher vehicle mass, less absorbed energy and resulting intrusions in a crash.

3. Methodology

The Dynamic Travel Time Prediction (DTTP) problem is defined in three different situations. In the first case, we address the problem of predicting the travel time of a vehicle when the pickup location and the drop-off coordinates are both known. In the second case, we consider the more difficult situation of predicting the travel time when only the pickup location coordinates is known. In the third and final case, we address the prediction of travel time at different points on the trajectory of the vehicle when the drop-off coordinates are known. We explore two different types of problems here. The first one is the continuous prediction of remaining travel time at each point in the trajectory for a trip and the second one is dynamic updating of the total travel time at each point in the trajectory for a particular trip. The

motivation behind using this method is that the predictor variables i.e. the pickup and drop-off location coordinates (or just the pickup location coordinates) are points on the surface of earth which can be taken approximately as a sphere.

K-Means Density clustering is a well-known methodology to both model and forecast univariate time series data such as traffic flow data, electricity price and other short-term prediction problems. The K-Means main advantages when compared to other algorithms are two:

1) It is versatile to represent very different types of time series: the autoregressive (AR) ones, the moving average ones (MA) and a combination of those two test and training datasets
2) On the other hand, it combines the most recent samples from the series to produce a forecast and to update itself to changes in the model.

3.1 Module Description

The Road Accident Analysis dataset have following modules,

- Preprocessing
- Hit Factor Analysis
- Area Wise Stage Factor Analysis
- Match Point Prediction

3.1.1 Data pre-processing:

In this module data preprocessing helps to describe a particular dataset performing a processing on raw data to prepare it for required information that is converting the numerical values into understandable language by removing the noise . The preliminary data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

3.1.2 hit factor analysis:

In this module hit factor analysis helps to determine the fatal severity. The fatal severity is based on three categories and they are high, medium and low. It is determined for each record.

3.1.3 Area wise stage factor analysis

This module describes the accident in which it occurred in area wise rather than finding whole. This module is much useful to filter the accident like how we need and it gives the entire data of the given area and retrieves all the information of the entire set of area like type of vehicle, cause and persons.

3.1.4 Data match point prediction:

In this Data Matching prediction module a dataset can be a massive undertaking where all possible patterns are systematically pulled out of the data and then an accuracy and significance are added to them that tell the user how strong the pattern is and how likely it is to occur again.

In general, these rules are relatively in our Road Accident dataset number of accidents appear in a U.S Traffic data's that might find interesting correlations in U.S fatal Accident Datasets database such as:

If two wheeler vehicles got accident then the cause of accident can be predicted by the time and this pattern occurs related to the instance by other accident record.

4.Proposed Sysytem

4.1 k-means density based clustering:

This approach makes the clusters of Accident locations. Accident locations describe the three different locations for accident high frequency, low frequency, and moderate frequency. It analysis the factors of road accident happened today. The another Clustering technique used for better analysis is hierarchical technique for this same data attributes is taken and loaded the .ARFF file in Java with Net beans.We initially divide the accident places into k clusters depends on their accident frequency with K-Means algorithm. Next, parallel frequent mining algorithm is apply on these clusters to disclose the association between dissimilar attributes in the traffic accident data for realize the features of these places and analyzing in advance them to spot different factors that affect the road accidents in different locations. The main objective of accident data is to recognize the key issues in the area of road safety.

The efficiency of prevention accidents based on consistency of the composed and predictable road accident data using with appropriate methods. Road accident dataset is used and implementation is carried by using Weka tool. The outcomes expose that the combination of K-Means and parallel frequent mining explores the accidents data with patterns and expect future attitude and efficient accord to be taken to decrease accidents.

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. Calculate distance function using 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. Execute

5. conclusion

This paper presents a comparative study of k-means algorithm on the road accident dataset on a new accident data set from Alabama, America with 1000 road accident records. Hence with the use o K-means road accident by this we can reduce the accident rate.

References

[1]Abdel-Aty MA, Radwan AE (2000) Modeling traffic accident occurrence and involvement. Accid Anal Prev 32(5):633-642.

[2] Barai S (2003) Data mining application in transportation engineering. Transport 18:216–223.

[3] Chaturvedi A, Green P, Carroll J (2001) k-Modes clustering. J Classif 18:35–55.

[4] Chen W, Jovanis P (2002) Method of identifying factors contributing to driver-injury severity in traffic crashes. Transp Res Rec 1717

[5] Geurts K, Wets G, Brijs T, Vanhoof K (2008) Traffic accident segmentation by means oflatent class clustering. Accid Anal Prev 40(4):1257–1266

[6] Han J, Pei H, Yin Y (2000) Mining frequent patterns without candidate generation. In: Proceedings of the conference on the management of data (SIGMOD'00, Dallas, TX). ACM Press, New York

[7] Joshua SC, Garber NJ (2003) Profiling of high frequency accident locations by use of association rules. Transp Res Rec 1840.

[8] Karlaftis M, Tarko A (2001) Data mining: concepts and techniques. Morgan Kaufmann, New York.

[9] Kumar S, Toshniwal Dhas Mining frequent patterns without candidate generation. In: Proceedings of the conference on the management of data (SIGMOD'00, Dallas, TX). ACM Press, New York.

[10] Kumar S, Toshniwal Dhas(2016) Mining frequent patterns with reliable technique.. In: Proceedings of the conference on the management of data (SIGMOD'00, Dallas, TX). ACM Press, New York

[11] Ona JD, Lopez G, Mujalli R, Calvo FJ (2013) Traffic accidents are contingent events and analyzing them requires awareness of the particularities that define them.

[12] Savolainen P, Mannering F Crash injury severities are recorded in ordinal scales with fatal crashes ranking highest in the scale and property damage minor crashes ranking lowest in the scale.

[13] Ulfarsson GF, Mannering FL has Mass of the striking vehicle is a factor in the prediction of accident severity.