International Journal of Advance Research in Science and Engineering
Volume No.08, Issue No.03, March 2019
www.ijarse.com

IJARSE
ISSN: 2319-8354

THE COMPARATIVE ANALYSIS OF ROAD ACCIDENT DATA USING DATA MINING TECHNIQUES

Dr. B.G.Geetha[1], Abinaya.N[2], Abirami Sivasakthi.M[3], Aishwarya.T[4]

[1]Professor & Head of the Department, Dept. of Computer Science and Engineering
K.S.Rangasamy College of Technology, Tiruchengode (India)
[2,3,4]Dept. of Computer Science and Engineering, K.S.Rangasamy College of Technology, Tiruchengode (India)

ABSTRACT

Road accidents are one of the most imperative factors that affect the sudden death among people and economic loss of public and private property. Road safety is a term linked with the planning and implementing certain strategy to overcome the road and traffic accidents. Road accident data analysis is a very significant 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 confront in road safety analysis. In this study, we are making use of latent class clustering (LCC) and k-modes clustering technique on a new road accident data. The heart to use both the techniques is to identify which technique performs better. The rules generate for each clusters do not prove any cluster analysis technique superior over other.

Keywords LCC, Apriori, k-means

1 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 analysts to carry out a comprehensive 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.

The major problem with road accident data analysis is its heterogeneous nature. Heterogeneity in road accident data is highly undesirable and unavoidable.This heterogeneous nature of road accident data may lead to less accurate results .

1.1 DATAMINING

In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., tremendous amounts of information were collected. Initially, with the advent of computers and means for mass digital storage, 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 very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems (DBMS). The efficient database management systems have been 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 needed.

The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, there are far more information than can be handled: 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.

1.2 HETEROGENEITY

The concept of "heterogeneity" is much invoked in social science research these days. Though this has long been the case in disciplines like psychology, sociology and anthropology, it was not heard much in mainstream economics until recently. Heterogeneity is now an integral part of economics in sub—disciplines like industrial organization, entrepreneurship, behavioural economics, and similar fields.

Fundamentally, heterogeneity is about the relationship between quantity and quality. When one breaks it down, the difference between quantitative and qualitative change becomes clear. Qualitative change involves the emergence of something new and is not amenable to measurement in quantitative terms. Quantitative and qualitative changes are categorically different.

The social researcher has much in common with the entrepreneur. The entrepreneur is defined by the "production" of novelty – the introduction of new categories of things, as well as simply new instances of old categories. Economic progress is about innovation, not merely accumulation. Thus, entrepreneurship research and research into social science in particular, must deal with this phenomenon of category change and emergence.

Social systems are composed of essentially heterogeneous elements that are related in complex ways. Because social systems are complex there is category dynamism and ambiguity. Received quantitative methods, deriving from probabilistic frames, will not work in investigating much of the social lB world. This is more true the more dynamic and innovative that social world is. Thus, to understand this world, the researcher must try to see it as the entrepreneurs who made it and are making it see it. There is an essential connection between researcher and subject matter that is absent in the case of the natural world. This paper reaffirms this well-known phenomenon.

1.3 FP growth

Data mining has been confronted with new opportunities and challenges. Some limitations are exposed when traditional association rule mining algorithms are used to deal with large-scale data. In the Apriority algorithm, scanning the external storage repeatedly leads to high I/O load and brings about low performance. As for FP-Growth algorithm, the effectiveness is limited by internal memory size because mining process is on the base of large tree-form data structure. What's more, although remarkable achievements have been scored, there are still problems in dynamic scenarios. The paper presents a parallelized incremental FP-Growth mining strategy based on Map Reduce, which aims to process large-scale data. The proposed incremental algorithm realizes effective data mining when threshold value and original database change at the same time. This novel algorithm is implemented on Hadoop and shows great advantages according to the experimental results.

Big data refers to a collection of datasets which is so huge and complicated that it is infeasible to process by using traditional methods and available technologies. Even if some analytical approach can barely finish the work, it still takes a long time and the outcome might not be satisfactory. Data mining , using existing data to analyze the overall trend or predict a problem that may arise in the future, is undoubtedly the core area of big data research. Association rule mining , one kind of data mining algorithms, becomes more and more popular these years. It intends to identify strong rules between no less than two items in database through different measures of interestingness. In a market analysis, association rules like “the customers who buy beer are likely to get diapers” might be generated according to the processing results. And these rules could be really helpful in making market plans. In addition to this typical application, association rules are also employed in Web usage mining, intrusion detection and continuous production.

After years of study, association rule mining algorithms are well established and effective in general cases. However, when it comes to big data, related algorithms are not mature and need further research. In a practical situation, database is updated periodically and threshold value often changes with needs of mining. It is clearly inefficient that the whole mining process has to be restarted from the beginning every time new data is inserted into database or mining parameter is reset. Furthermore, to deal with the issues resulted from large scale data, algorithm parallelization has become inevitable. This paper presents a parallelized incremental FP-Growth (PIFP-Growth) mining strategy. The proposed algorithm successfully solves the incremental issue brought by the dynamic threshold value and database at the same time, which avoids repeated computation. This parallel mining strategy based on Map Reduce framework is implemented on Apache Hadoop . The experimental results have proved the effectiveness and advantages of PIFP-Growth. Traditional data mining methods and algorithms have limitations when dealing with big data. For instance, Apriority algorithm needs to scan the data from external storage repeatedly so as to obtain the frequent item sets, which brings heavy I/O load with low performance. Moreover, existing incremental algorithms cannot be applied in situations when both threshold value and database change. This paper presents an incremental FP-Growth mining strategy, which is parallelized under the Map Reduce framework. Experimental results indicate that this improved algorithm is effective in reducing time of duplicated work.

2 METHODOLOGIES

- k-Modes clustering
- Latent class clustering
- Number of cluster selection
- FP growth technique

2.1 k-modes clustering

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.The k-mode algorithm is quite efficient in handling large categorical data.

Let A and B be two qualitative data objects categorized by x categorical attributes.

The distance function of k-modes algorithm can be defined a $d(A,B) = \sum \delta(A_i ,B_i)$

where, $(A_i, B_i) = 1$, If $(A_i = B_i)$; $(A_i, B_i) = 0$, If $(A_i = B_i)$

Given a set of categorical data objects D defined by n attributes A1, A2, …, An. A mode of D= {D1, D2, …, Dn } is a vector V = {v1, v2, …, v3} that minimize $d(D, V) = \sum d(Di, V)$

The iterative process of k-modes algorithm is similar to k-means algorithm.

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.

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 both. LCC does not require any prior standardization that affects the results.

The basic form of LC cluster model is given below,

$$F(Yi| \theta) = \sum \pi j\ Fj\ (Yi| \theta\ j)$$

where, Yi indicates the score of an object on a set of variables under observation.

N is the number of clusters to be formed

πj is the prior probability of an object's membership to a cluster j

θj is the model parameters of cluster j

$Fj\ (Yi| \theta\ j)$ is the mixture probability density.

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.

2.4 FP growth technique

Association rule mining is a popular data mining technique that is based on market basket data analysis.Previous studies, used association rule mining using Apriori algorithm for road accident data analysis. 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. Hence, 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 item set association information.

CLUSTER ANALYSIS

The primary task of cluster analysis is to determine the number of cluster that can be formed in the data set.On the basis of the similarities in the features, locations, vehicle type, etc., and many attributes are considered into account.

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.

3 PROPOSED SYSTEM

In this proposed system consider the dynamic travel time prediction (DTTP) problem in three different situations. In the first case, the problem of predicting the travel time of a vehicle was addressed when the pickup location and the drop-off coordinates are both known. In the second case,the more difficult situation of predicting the travel time was considered when only the pickup location coordinates is known. In the third and final case, the prediction of travel time at different points on the trajectory of the vehicle was addressed when the drop-off coordinates are known. Two different types of problems were explored 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. To the best of our knowledge, there has been no work reported in the literature that takes into account the spherical nature of the data while solving the travel time prediction problem for GPS enabled taxis in streaming data context.

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 like our own. 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 ADVANTAGES:

- Booking goals for a query by clustering user booking seasons.
- Different user search goals can be obtained conveniently after feedback sessions are clustered
- Better accuracy.
- Less in memory usage.

4 MODULE DESCRIPTION

The Road Accident Analysis dataset have following modules,

- Preprocessing
- Hit Factor Analysis
- Area Wise Stage Factor Analysis
- Match Point Prediction
- Density Based Clustering

4.1 DATA PRE-PROCESSING:

In this module data preprocessing module helps to describes taxi dataset processing performed on raw data to prepare it for another processing procedure. The preliminary data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

4.3 HIT FACTOR ANALYSIS:

The score it receive on a Stage is your total points (minus any penalties) divided by your time to complete that stage. This is referred to as your Hit Factor for that stage and it is what determines your place when scoring that stage.

4.4 AREA WISE STAGE FACTOR ANALYSIS

This module helps to find the highest Hit Factor for a stage earns 100% of the points available for that stage. Everyone else determines the number of points they earned as a percentage of that high hit factor. If it shot 68.36% of the top shooter for stage 3 then it would earn 68.36% of the points available for that stage. This is referred to as your Stage Points. Remember that it only compete against those in your Division so the high hit factor for a shooter in another division doesn't make any difference on your stage points earned

K-Means density-based clustering module helps to find given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors).

The marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).All points within the cluster are mutually density-connected.If a point is density-reachable from any point of the cluster, it is part of the cluster as well.

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

4.6 K-MEANS DENSITY BASED CLUSTERING:

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 Java with Netbeans.

The accident places are divided 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.

5 CONCLUSION

An analysis is done by a comparative study of k-modes clustering and LCC on a new road accident data set. The number of attributes that has been used in the analysis was 11 which were associated with road accidents. The information criteria (AIC, BIC and CAIC) 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 both the 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. Although Chaturvedi et al. (2014), provided that k-modes are better than LCC on categorical data, no differences found that shows that k-modes are better than LCC especially in road accident data except computational speed. There is no doubt that both the cluster analysis technique performs well in reducing the heterogeneity of road accident data. Also the association rules generated is providing information about various types of road accidents and their associated factors. Also, these results are quite similar to Dehradun district which is adjacent to Haridwar district.

ACKNOWLEDGEMENT

We acknowledge DST-File No.368. DST – FIST (SR/FIST/College – 235/2014 dated 21-11-2014) for financial support and DBT – STAR – College – Schee - ref. no: BT/HRD/11/09/2018 for providing infrastructure support.

REFERENCES

[1] Abdel-Aty.MA and Radwan.AE (2013), “Modeling traffic accident occurrence and involvement”, Accident Analysis and Prevention, Vol.no:32(5) ,pp:633-642.

[2] Barai.S (2013), “Data mining application in transportation engineering”, Transport, Vol.no:18, pp:216–223.

[3] Chaturvedi.A, Green.P and Carroll J (2015), “k-Modes clustering”, Classification, Vol.no:18, pp:35–55

[4] Chen.W and Jovanis.P (2013), “Method of identifying factors contributing to driver-injury severity in traffic crashes”, Accident Analysis and Prevention, Vol.no:32(4), pp:600-612.

[5] Depaire.B, Wets.G and Vanhoof.K (2014), “Traffic accident segmentation by means of latent class clustering”, Accident Analysis and Prevention, Vol.no:40, Issue.No:4, pp:1257–1266

[6] Fraley.C and Raftery.AE (2013), “Model-based cluster analysis”,Clustering, Vol.no:41, pp:578–588

[7] Geurts.K, Wets.G, Brijs.T and Vanhoof.K (2016), “Profiling of high frequency accident locations by use of association rules”, Accident Analysis and Prevention, Vol.no:32, Issue.No:3, pp:224-230.

[8] Han.J and Kamber.M (2014), “Data mining: concepts and techniques”, Transportation, Vol.no:16, pp:30-35.

[9] Han.J, Pei.H and Yin Y (2015), “Mining frequent patterns without candidate generation”, In Proceedings of the journals on the management of data, Vol.no:2, pp:213-220

[10] Islam.S and Mannering.F (2016), “Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence”, Accident Analysis and Prevention, Vol.no:37, Issue.No:2 , pp:267–276