Modified Approach of Cluster Algorithm to Analysis Road Accident

Reeta Bhardwaj
Assistant Professor, Department of Computer Application, DAVIET, India

Ridhi
Research Scholar, Department of CSE, DAVIET, India

Rajeev Kumar
Assistant Professor, Department of Information Technology, DAVIET, India

ABSTRACT
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, we have used data mining techniques to analyze the data provided by data.gov.uk in which we first cluster the accident data and further association rule mining technique is applied to identify circumstances in which an accident may occur for each cluster. The results can be utilized to put some accident prevention efforts in the areas identified for different categories of accidents to overcome the number of accidents also the parameters of the proposed approach is compared with the existing approach on the basis of time and accuracy and proves that the proposed technique has better performance.

Keywords
Data Mining; Road Accidents; Association Rule Mining.

1. INTRODUCTION
Data Mining is a process that discovers the knowledge or hidden pattern from large databases. DM is known as one of the core processes of Knowledge Discovery in Database (KDD). It is the process that results in the discovery of new patterns in large data sets. It is a useful method at the intersection of artificial intelligence, machine learning, statistics, and database systems. It is the principle of picking out relevant information from data. It is usually used by business intelligence organizations, and financial analysts, to extract useful information from large data sets or databases DM is use to derive patterns and trends that exist in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

The goal of this technique is to find accurate patterns that were previously not known by us. So, the overall goal of the DM process is to extract information from a data set and transform it into an understandable structure for further use. Data mining area is widely used in number of organizations viz. hospitals, bank sector, retail stores, insurance sector etc.

DM is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics.

Road and accidents are uncertain and unsure incidents. In today's world, traffic is increasing at a huge rate which leads to a large numbers of road accidents. The highway safety is being compromised and there are not enough safety factors by which we can analyse the traffic collisions before it happens. A method is proposed by which we can pre-process the accidental factors. Young drives are more prone to road accident as they pretend to be more courageous after drinking alcohol and this causes them to lose control over the vehicle. Drunk driving will not only risk a person's own life but may also cause an incident life to be lost. Numerous factors that increasethe risk of collision, includes design and manufacture of vehicle, driving speed, road map design, road area and environment, and driver's driving skills, lack of vision due to alcohol or drugs overdose, and behaviour of driver, over speeding and street racing. Vehicular accidents lead to numerous hazards like death, life time disability and monetary loss to individual and society. It was reported that about 54 million people were victim of road injuries in 2013[14] which leads to 1.4 million deaths in 2013 rising from 1.1 million deaths in 1990[15]. About 68,000 of these injuries made children as its victim between age of five years [15]. Almost all developed countries have low death rates, while the majority of under- developing countries have increasing death rates due to vehiclecollisions. [14] Developing countries have the highest rate with 20 deaths per 100,000 inhabitants, 80% of all road mortalities by only 52% of all vehicles. [14] viz. the death rate in Africa is the highest (24.1 per 100,000 inhabitants), the lowest rate is reported in Europe. [16] Road and traffic accidents defined by a set of variable. the major issue is analysis of accidents data is its varied nature. The diverse must be considered analysis of the data. So those researchers are used clustering analysis. The clustering analysis is an important technique. Cluster analysis is useful to various tasks. Dr. R. Geetha Ramani1, S. Shanthi2 [4] used predicated model technique. In this paper technique algorithm are applied on the like random tree c4.3 tree and j4.3. In this paper researcher are discussions about classifier and predication technique to data mining. In this paper predication of road accident patterns related to pedestrian characteristics. This classifier is voided using cross validation with k folds and evaluated using the accuracy measures: precision and recall and roc. In this paper random tree classifier are given to better result as compared to decision stump [13]. Seoung-hun Park and Young-guk Ha is used imbalance technique and map reduce algorithm. Imbalance data means data that have a huge difference between the obverted sizes from one data set. So researcher is solved the problem to used sampling technique. There are two type sampling are: over sampling and under

sampling. over sampling is to use all observation value in a big class and increase size of observation value in a small class and use this value. Under sampling are those who used lost data. Other problem is occurring the researcher are data processing [12] .in this case training set of data make multiple feature .it take time so much so that researcher isto solve the problem in map reduce algorithm. map reduce algorithm are used to big processing technique

2. RELATED WORK

Park and Ha [17] describes about imbalance technique and map reduce algorithm. Imbalance data means data that have a huge difference between the obverse sizes from one data set. So researcher is solved the problem to used sampling technique. There are two type sampling are: over sampling and under sampling. Over sampling is to use all observation value in a big class and increase size of observation value in a small class and use this value. Under sampling are those who used lost data. Other problem is occurring the researcher are data processing .in this case training set of data make multiple feature .it take time so much. So that researcher is solving the problem in map reduce algorithm. Map reduce algorithms are used to big processing technique.

Gakis and Tzovaras[18],presents a support vector machine (svm) based approach for detecting traffic network incidents. Support vector machine are performing wide range of classification problems and is fairly robust to irrelevant feature. The evaluation metrics that are commonly used assessing any traffic incident detection system. Drawbacks-the limitation is speed and size, both in training and testing and slow in test phase, a problem addressed in.

Tao and Jian [19], explained the use of time-series model to analyse the problem. The cell transmission model is traffic flow data and transformed into time series and time series method involved it. Cluster analysis was implemented to analysis the traffic flow sequence. This paper researcher is created time series data by converting ternary numbers to decimal number. to detect the Euclidean distance will not consider the linear drift. Temporal data miningis those which can observe data point in statistically independent from observation.

Kashani and Ranjbari [20], has discussed about the severity injury of drivers involved in traffic crashes on the roads. In this paper researcher are overcome the problem of traffic accidents by using CART and variable selection procedure. The Classification and Regression Tree (CART), an important data mining technique, is a without parameter model without any pre-acknowledged relationship between the dependent and independent variables. Variable Importance Measure (VIM) for trees, which may be applied as a criterion to select a subset of variables that have a major importance in predicting the target variable

Changalasetty and Ghribi [21], has discussed about the vehicle classification. in this paper researcher are described the cluster with the technique using k-mean clustering algorithm. In this paper researcher are using cluster like big and small. it also described image processing technique to control traffic accidents. It is used k-modes clustering-means clustering described the size of the vehicle is small or big. According to this technique it provides the information of vehicle. It also tell that how many vehicle are across the particular area. It also provides the information of traffic accidents by using k-modes algorithm.

Tian and Zhang [22],have presented the rough sets and theory association rule. The rough sets are based upon the boundaries. rough sets are used the fuzzy on uncertainly. Association rule are used frequent items on data sets. Rough set theory is used the fuzzy boundaries in particular knowledge. The rough set can express by the extract concept of the upper approximation and lower approximation. It presented the radial and multi-dimensional, three-dimensional structure and multi-level super. It is containing multiple data attributes.

Sriratna [23], presented on this paper association mining. In this paper association mining are help frequent items on traffic data sets. It provides the relationship with traffic feature and injury severity data. In this paper also provide the information about rough set association mining. The rough set are those which provide hidden the relation of data. In this paper also provide interesting indicator. Interesting indicator are those which help measures the data traffic. It used various kinds of measures which provide proper result of according in this paper. It is used various data regarding traffic. Due to this data are helpful to solve the problem of traffic.

Wang [24] has discussed about multidimensional association rule in traffic accident. Multidimensional is involved two or more dimension called as multidimensional association rule. in this paper researcher are discuss about apriori algorithm. Apriori algorithm is bottom-up approach used. It uses frequent subsets are extended one item at a time.

3. PROBLEM FORMULATION

There are several major data mining techniques which have been developed and used in various data mining projects. In the proposed work, k –means performance will be enhanced by using hybrid approach for better result. To show the effect of noise on the performance of various clustering techniques. Clustering may be applied on database using various approaches, based upon distance, density hierarchy and partition. Clustering is being widely used in many application including medical, finance etc. our purpose is to study how a particular clustering technique is responsive to the noise in the term of time. Apriori algorithm minimum support is needed to generate the large item set from candidate set in which not so required candidate item sets are pruned by utilizing user defined minimum support threshold. Moreover, in Apriori Algorithm, if the frequencies of items vary a great deal, we will encounter two problems. First of all, if minimum support is set too high, those rules that involve rare items will not be found. Secondly, to find rules that involve both frequent and rare items, min support has to be set very low.

This may cause combination eruption because those frequent items will be linked with one another in all possible ways. So, Apriori is utilizing hit and trial method to find the required number of rules

4. PROPOSED WORK

The main objective of this proposed work is to enhance the existing machine learning techniques in analyzing the road accidents and improve the performance. In this, a hybrid model of clustering with association rule mining is proposed. In clustering Hybridized K mode algorithm is on the formatted road accidents data and does the clustering. Then improved FP tree association rule mining algorithm is applied to determine the rules from the clustered data. These rules are used to predict the accidents analysis.

Collection of data: Dataset was made by referring various traffic accidents analyzers websites It had a considerable measure of respectability issues. It was gathered and arranged

by website -data.gov.uk .it respectability issues in the dataset had been determined. It contains all sorts of traffic accidents. This dataset contains these attributes i.e. attention, drink, physical, inexperience, rule, mistake, speed, vehicle, whether, road and sight.

Preprocessing and filtering: The collected data can be preprocessed and filtered using replace missing value and numeric to nominal data conversion.

Hybrid Clustering algorithm: The main focus of algorithm is to set two simple data structures as collectors to retain the labels of cluster and the distance for the data objects to the nearest mean cluster during the repeated form of iterations, that ordered be used as in next iteration, we calculate the distance between the recent data object and the newest cluster's center, and if the actual computed distance is smallest than or equal to the distance to the previous center of the data object stays in its clusters . Therefore, there is no need to calculate the distance again and again from the data object to the other k clustering centers, saving the calculative time to the k cluster centers. We have combined two approaches i.e. first dividing the dataset into sub samples and then applying the reduced iteration approach to each subsample to find the clusters. This will save lot of time and improve the performance for the large datasets. Split the whole data into various subsamples and Then apply the algorithm that will reduce the number of iterations.

Improved Rule mining algorithm: It allows frequent item set discovery without candidate itemsets generation. Two step approach. Build a compact data structure called the FP-tree and Built using 2 passes over the data-set. Extracts frequent itemsets directly from the FP-tree Because Node sets are based on a Pre order-tree: Here we are using Node sets instead of n-list. Pattern order tree is a tree structure: It consists of one root labeled as "null", and a set of item prefix sub trees as the children of the root. Each node in the item prefix sub tree consists of: item-name, count, and children-list, pre-order.

Proposed method FPM consists of:

(1) Construct the Pre order -tree and identify all frequent 1-itemsets;
(2) scan the Pattern-tree to find all frequent 2-itemsets and their Node sets;
(3) Mine all frequent k (>2)-item sets. For enhance the efficiency of mining frequent item sets, FPM adopts promotion, which is based on superset equivalence property, as pruning strategy

Pre order-tree definition

Because Node sets are based on a Pre order-tree: Here we are using Node sets instead of n-list.

Pattern order tree is a tree structure:

(1) It consists of one root labeled as "null", and a set of item prefix sub trees as the children of the root.
(2) Each node in the item prefix sub tree consists of: item-name, count, and children-list, pre-order.

Item name = which item this node represents.

Count = the number of transactions presented by the portion of the path reaching this node.

Children-list = all children of the node.

Preorder is the pre-order rank of the node.

Algorithm:

1) Initializes F, used to store frequent item sets, by setting it to be null.
2) Constructs the Pattern tree and finds F1, the set of all frequent 1-itemset
3) Initializes F2, to store frequent 2-itemsets, by setting it to be null.
4) Insert all candidate frequent 2-itemsets in F2, by scanning the Pattern tree with the pre-order traversal.
5) delete all infrequent 2-itemsets from F2,
6) Generate all frequent k-itemsets ($k \geqslant 3$) by using efficient pruning strategy to generate all frequent k-it

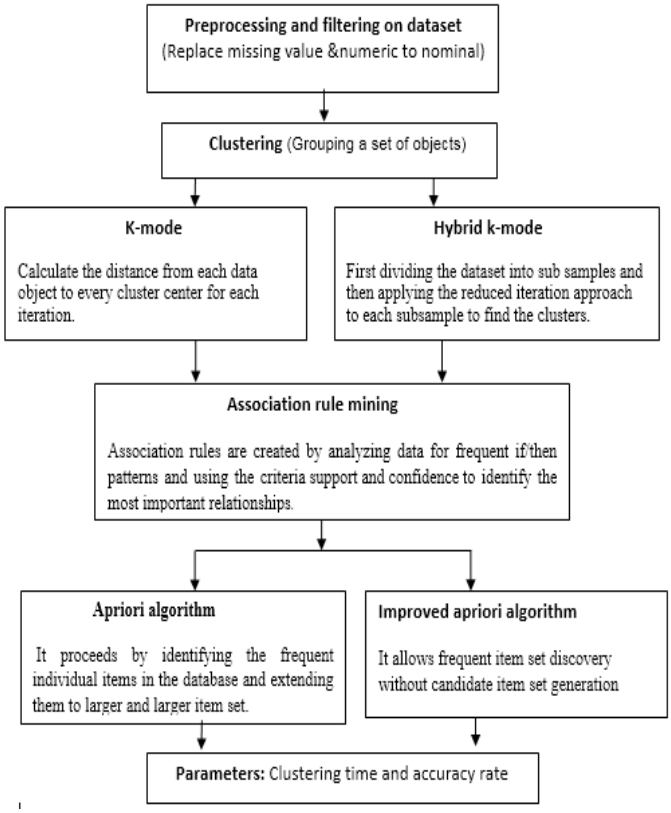

Fig. Flowchart of proposed methodology

5. RESULTS

This paper chooses the data sets of traffic accident to test the efficiency of the hybrid k-mode algorithm and the standard k-mode. Results been carried out to illustrate the performance efficiency of the hybrid k-mode algorithm in this paper. This algorithm been implemented to the clustering of real datasets. In two experiments conducted, time taken and accuracy for each experiment is analyzed. Furthermore, it also computes number frequency item. The same data set is given as input to the k-mode algorithm and the hybrid algorithm. The resulted Experiments compare hybrid k-mode algorithm with the standard k-mode algorithm in terms of the total execution time of clusters and their accuracy.

Table 1: Execution Time comparison of K-Modes and Proposed Hybrid Clustering

No. of clusters 'k'	K-modes clustering time ms	Hybrid clustering time ms
2	380 ms	160 ms
3	192 ms	120 ms
4	150 ms	100 ms
5	240 ms	140 ms
6	160 ms	110 ms

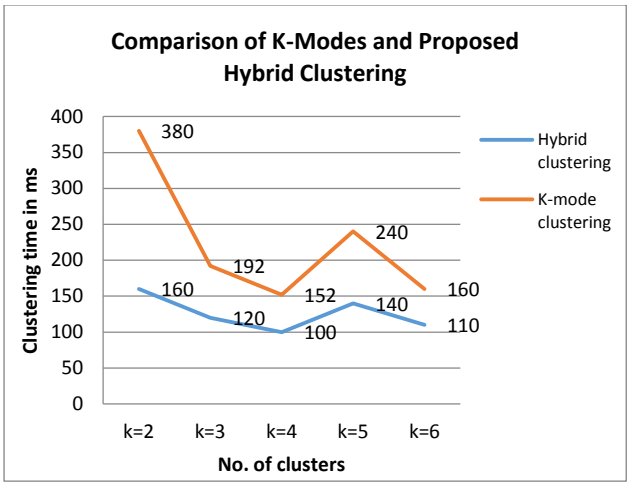

Figure 1: showing the comparison of clustering time between k-modes and hybrid clustering algorithms

Table 1 It shows that the total execution time of the clusters, with diverse cluster and time hybrid k-mode we have reduced time and the clustering time hence is reduced with the reduction in the execution time and the formation of clusters are of good quality because in here we don't have to calculate the distance again and again as in case of K-mode and they don't change the clusters again as data iterations does not move .In hybrid K-mode algorithm ,If this distance is less than Dist[i], the data object stays in the initial cluster, Else for every cluster centre cj(1<=j<=k), compute the distance d(di, cj) of each data object to all the centre, assign the data object di to the closet centre cj.

Table2: Accuracy comparison of K-Modes and Proposed Hybrid Clustering

No of clusters' k'	K-Modes accuracy rate percentage	Hybrid accuracy rate percentage
2	81.370 %	87.296%
3	72.575%	76.094%
4	65.708%	67.467%
5	56.695%	59.0987%
6	51.076%	52.3605%

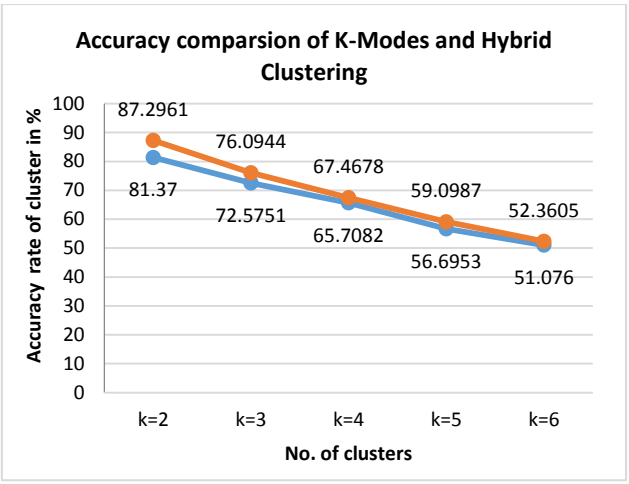

Figure 2: showing the comparison of clustering accuracy between k-mode and hybrid clustering algorithm

Table 2. The line graph depicts the comparison between k-mode and Hybrid clustering with diverse clusters. if the value of k=2 then accuracy rate of k-mode is 81.37 % but hybrid clustering is 87.296%. After that again we select the value of cluster is 3 then hybrid cluster accuracy rate increased come to76.094% as compared to k-mode. Each time accuracy rate of hybrid clustering is incremented.

6. CONCLUSION

To overcome the above issues, in this paper, we proposed a model for the analysis of road accident patterns for various types of accidents occurring on the road which makes use of Hybrid-clustering and improved-association rule mining algorithm. Comparing and analysing the results of proposed technique with K means clustering and Apriori algorithm on the basis of clustering time, accuracy and association rule mining time.

7. ACKNOWLEDGMENT

The paper has been composed with the kind assistance, guidance and support of my department who have helped me in this work. I would like to thank all the people whose encouragement and support has made the fulfillment of this work conceivable.

8. REFERENCES

[1] R. Patel Nimisha, Sheetal Mehta "A Survey on Mining Algorithms" International Journal of Soft Computing and Engineering , vol. 2, issue 6, pp 460-463, January 2013.

[2] Sotiris Kotsiantis, DimitrisKanellopoulos, "Association Rules Mining: A Recent Overview" GESTS International Transactions on Computer Science and Engineering, vol.32 (1), pp 71-82, 2006.

[3] Rakesh Kumar Soni1, Neetesh Gupta, AmitSinhal, "An FP-Growth Approach to MiningAssociation Rules" International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 2, February 2013, pp 1 – 5.

[4] JaiWeiHan, Jian Pei, Yiwen Yin &Runying Mao, "Mining frequent patterns without candidate generation: A Frequent pattern tree approach" Data mining and knowledge discovery, Netherlands, pp 53-87, 2004.

[5] Huan Wu, Zhigang Lu, Lin Pan, RongSeng XU and Wenbaojiang "An improved Apriori based algorithm for association rule mining" IEEE Sixth international

conference on fuzzy systems and knowledge discovery, pp 51-55, 2009.

[6] Badri Patel, Vijay K Chaudhari, Rajneesh K Karan, YK Rana "Optimization of Association Rule Mining Apriori Algorithm using ACO" International Journal of Soft Computing and Engineering vol 1, issue 1, pp 24-26, March 2011.

[7] K.Saravana Kumar, R.ManickaChezian,"A Survey on Association Rule Mining using Apriori Algorithm" International Journal of Computer Application, vol. 45, no. 5, pp 47-50, May 2012.

[8] Rafael S. Parpinelli, Heitor S. Lopes, Alex A. Freitas, "Data Mining With an Ant Colony Optimization Algorithm" IEEE Transactions on evolutionary computing, vol. 6, no. 4, pp 321-332, August 2002

[9] SuhaniNagpal "Improved Apriori Algorithm using logarithmic decoding and pruning" International Journal of Engineering Research and Applications, vol. 2, issue 3, pp. 2569-2572, May-Jun 2012.

[10] Fernando E. B. Otero, Alex A. Fretas and Colin G. Johnson "A new sequential covering strategy for inducing classification rules with ant colony algorithms" IEEE transaction on evolutionary computation, vol. 17, no. 1, pp 64-76, February 2013.

[11] Sang Jun Lee, KengSiau"A review of data mining techniques" Industrial Management and Data Systems, University of Nebraska-Lincoln Press, USA, pp41-46, 2001

[12] S. Shanthi, R.geethaRamani "feature relevance analysis and classfication of road traffic accident data through data mining techniques",2012pg no 24-26

[13] Dr. R. Geetha Ramani1, S. Shanthi2."Classifier Prediction Evaluation in Modeling Road Traffic Accident Data",IEEE International Conference on Computational Intelligence and Computing Research,pp 1-4,2012

[14] Global Burden of Disease Study 2013, Collaborators (22 August 2015). "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.". Lancet (London, England) 386 (9995): 743–800.PMID 26063472

[15] GBD 2013 Mortality and Causes of Death, Collaborators (17 December 2014)."Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013."

[16] Global status report on road safety 2013: Supporting a decade of action (PDF) (in English and Russian). Geneva, Switzerland: world health organization WHO. 2013.ISBN 978 92 4 156456 4. Retrieved 3 October2014.

[17] Seoung-hun Park andYoung-guk Ha, "Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction",Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing,pp45-49,2014

[18] EvangelosGakis, DionysiosKehagias and DimitriosTzovaras,"Mining Traffic Data for Road Incidents Detection",IEEE 17th international conference on Intelligent Transporatation System(ITSC) October 8-11,2014 Qingdao, Chine.

[19] An Shi,Zhang Tao, Zhang Xinming, Wang Jian ,"Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining ", Fifth International conference On Intelligent Sytem Design And Engineering Apllication,2014, pp- 454-457

[20] A. T. Kashani, A. Shariat-Mohaymany, A. Ranjbari," A Data Mining Approach To Identify Key Factors Of Traffic Injury Severity", 2009 Traffic&Transportation, Vol. 23, 2011, No. 1, 11-17

[21] Suresh BabuChangalasetty, LalithaSarojaThota, Ahmed Said Badawy, Wade Ghribi," Classification of Moving Vehicles using K-modes Clustering",2015,pp 1-6

[22] RuiTian and Zhaosheng Yang and Maolei Zhang, "Method of Road Traffic Accidents causes Analysis based On data Minning", 2010, pp 1-4

[23] PannawatSriratna, PakornLeesutthipornchai,"Interesting-based Association Rules for Highway Traffic Data", 2015, pp1-6

[24] He song-bai ,wangYa-jun Sun yue-kun gao wen-weichenqiang An ya_qin "The research of multidimensional assocation rule in traffic Accident",2008. pg no 1-4