Pattern Mining Predictor System for Road Accidents

Sisir Joshi[1], Abeer Alsadoon[1], S. M. N. Arosha Senanayake[2], P. W. C. Prasad[1(✉)], Shiaw Yin Yong[2], Amr Elchouemi[3], and Trung Hung Vo[4]

[1] Study Group Australia, Sydney, Australia
chan999941@yahoo.com, {aalsadoon,cwithana}@studygroup.com
[2] Universiti Brunei Darussalam, Gadong BE 1410, Brunei Darussalam
{arosha.senanayake,shiawyin.yong}@ubd.edu.bn
[3] American Public University System, Charles Town, USA
Amr.elchouemi@ashford.edu
[4] University of Technology and Education, The University of Danang, Da Nang, Vietnam
vthung@ute.udn.vn

Abstract. Road traffic accidents are among the major concerns that are leading for deaths and injuries in the world. Many predictive models use data mining technique to provide semi optimal solutions. Pattern identification and recognition have been used to for road accident predictions based on the critical features extracted depending on the dangerous locations and frequency of occurrences prone for accidents. The aim of this research is to propose a novel predictive model based on the pattern mining predictor which improves the accuracy of accident prediction in frequent accident locations. The proposed system consists of association rules mining technique, which identifies the correlation, frequent pattern and association among the various attributes of the road accident. Clustering technique that discriminates the data based on different patterns and classification technique that classify and predicts the severity of accident. Novel system built leads to an improvement in the accuracy of the accident prediction from 92% to 94%. Furthermore, using selective subset of features decreased the processing time and precision of classification is improved using boosting technique.

Keywords: Road traffic accident prediction · Association rules mining · Road safety · Apriori algorithm · Classification analysis · Data mining · Clustering · Fuzzy logic · Naïve Bayes classifier

1 Introduction

Pattern mining is an emerging area of study of Road Traffic Accident Prediction where prediction is done by identifying the most frequent accident location pattern that are prone to risk. Pattern mining uses different methods and algorithms to find out the relation in very large amount of data set. Mining the data to identify the relation among data is considered as one of the most important technology in previous decades [1]. In the past, various research about road traffic were conducted in many countries and the

M. Hernes et al. (Eds.): ICCCI 2020, CCIS 1287, pp. 605–615, 2020.
https://doi.org/10.1007/978-3-030-63119-2_49

data set from these researches were utilized in this work. The datasets obtained were processed, clustered, classified using data mining algorithms. Association rule mining is one of the popular methodologies which helps to identify the significant relation among the data in large dataset and used for frequent itemset mining [2]. Apriori Algorithm is one of the classical association rule mining technique that we have used to analyse the road traffic accident which yields best rules, but it has major limitation due to lengthy processing time. To overcome this limitation, authors in [3] enhanced the algorithm in order to generate the best rule with the consideration of support count only leading to fast processing time. Classification is used to construct a model from the training dataset to classify the records of unknown class. Naïve Bayes technique is one of the best probability-based method for classification which classifies the data based on the Bayes hypothesis with presumption of independence between pair of variables.

Current studies of data mining in Road Traffic accident uses various technique and algorithm to improve accuracy and processing time to predict the location of accident that are prone to risk, the maximum generate accuracy is 92.45% and processing time of 15 s on average.

The purpose of this research is to develop a novel system to increase the accuracy and precision in the prediction of road traffic accidents and decrease the processing time while predicting to take prevent measures and reduce severity of injuries. The use of single technique is not enough to predict with high accuracy and precision therefore, we use number of data mining techniques that helps in analysis of data and is able to predict road traffic accidents reducing severity of injuries and increase in measures required before, during or after accident.

2 State of the Art

Janani in [4] proposed a system that uses Naïve Bayes classification technique based on Apriori algorithm association mining to identify patterns in order to predict the severity of accident, which further brings out the factor related to the case of accident and a predictive model interfaced with fuzzy based location wise accident frequency. The use of Naïve Bayes Classifier has improved the accuracy of prediction model. The experiment was conducted by clustering dataset generating the rules based on Apriori algorithm, which then decomposed into set of training and test data for predicting the severity of accident as reported in [4]. The outcomes proves an accuracy of 92.45% in predicting accident. This model consists of five stages as illustrated in Fig. 2, i.e. Data Pre-processing, Clustering, Association rules mining, Classification and Prediction.

As per the solution provided in [4], the data was clustered and then rules were generated and further decomposed into training set and test set 70% and 30% respectively. Based on the attribute such as Fatal, Grievous, Damage, and Injury in the data set, class label was created which represent the severity level of accident. Class 0 represents the low severity level and class 1 represents high severity level. Naïve Bayes classifier classified the training data with best accuracy.

Prediction process involves use of fuzzy logic algorithm, which predicts the probability of accident occurrence. The propose model in [4], Fuzzification is done to transform the input to fuzzy values, which is the process in fuzzy domain by inference engine based

on knowledge base supplied by domain expert. Then the processed output is transformed back with the defuzzification method that shows the probability.

The Apriori Algorithm is implemented in Association rules mining to extract set of rules, which defines a set of patterns in [4]. However, the processing time can still be reduced in order to extract rules leading to overall processing time minimum.

This model increases the accuracy of prediction by using Naïve Bayes classifier prediction model, which predicts with 92.25% accuracy and 92 precisions, is achieved compared to 90.25% accuracy of decision tree and 88% accuracy of random forest as reported in [4].

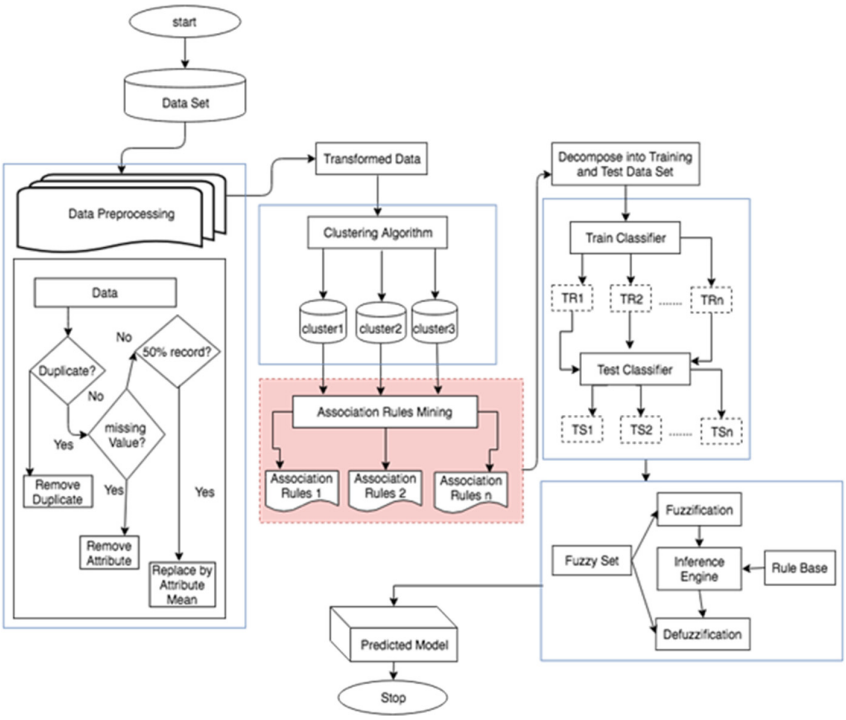

Fig. 1. Naïve Bayes classification technique using Apriori algorithm for prediction the severity of accident proposed as state of the art in [4].

The research work done in [5] investigated classification accuracy on the basic of performance metric to predict the severity of accidents. They proposed a model, which used the Classification and Regression Tree (CART) to analyse road accidents data of Iran and found that not using seat belt, improper overtaking and over speed affect the severity of accidents.

The article [6] discussed the performance of prediction approach in term of predicting model's accuracy. They have discussed that the RTA involved fatal crashes data is directly concerned with nutritional health survey data to analysis of the association of dietary

habit of a motor vehicle driver's to road traffic accident by applying Association rule mining algorithms.

A novel rule-based method to predict traffic accident severity according to user's preferences instead of conventional DTs was proposed in [7]. The novel multi-object and rule-based method in [7] outperforms the classification methods such as ANN, SVM, and conventional DTs according to classification metrics like accuracy (88.2%), and performance metrics of rules like support and confidence (0.79 and 0.74, respectively). Proposed method yielded promising result with an increased accuracy of 4.5% from other methods but the obtained rules from this method are not very much effective. Therefore, feature selection method and extraction method must be used to increase the accuracy and improve the effectiveness of the obtained rules.

Authors in [8] collected and analyzed 398 wrong-way driving (WWD) crashes in Illinois and Alabama States. They employed multiple correspondence analysis (MCA) to define the structure of the crash data set and identify the significant contributing factors to crashes. According to the obtained results, driver age, driver condition, roadway surface conditions, and lighting conditions were among the most significant contributors to WWD crashes.

A learning model was built in [9] for predicting accidents on the road using classification analysis and a data mining procedure. The Hadoop framework was proposed to process and analyze big traffic data efficiently and a sampling method to resolve the problem of data imbalance. Based on this, the predicting system first preprocesses the big traffic data and analyzes it to create data for the learning system. The imbalance of created data is corrected using a sampling method. To improve the predicting accuracy, corrected data are classified into several groups, to which classification analysis is applied.

3 Proposed System

After reviewing a range of method for prediction for road traffic accident, we analysed pros and cons of each method. Accuracy, processing time, precision, recall was the main issues to be considered. Proposed solution is the Enhanced Apriori algorithm for extracting best rule with faster iteration compared to the reported algorithm in [4]. Enhanced Apriori algorithm generates the candidate item sets faster, which reduces the processing time. This algorithm reduces the number of candidate item sets that is needing to be scanned and by reducing the number of candidate item sets. Whenever the k of k-itemset and value of minimum support increases, from view of time consumed enhance Apriori algorithm improves the processing time significantly. The use of breadth-first search strategy to count the support of item sets and use of candidate generation function which exploits the downward closure property of support which is one of the best ways to mine the pattern/association rules as reported in [3]. This is a feature adapted from second-best solution. This information reduces the number of transactions to be scanned making it possible to minimize the generation of candidate item sets.

Novel system implemented here consists of five main stages as illustrated in Fig. 1; Data Pre-processing, Clustering, Feature Extraction, Classification and Prediction.

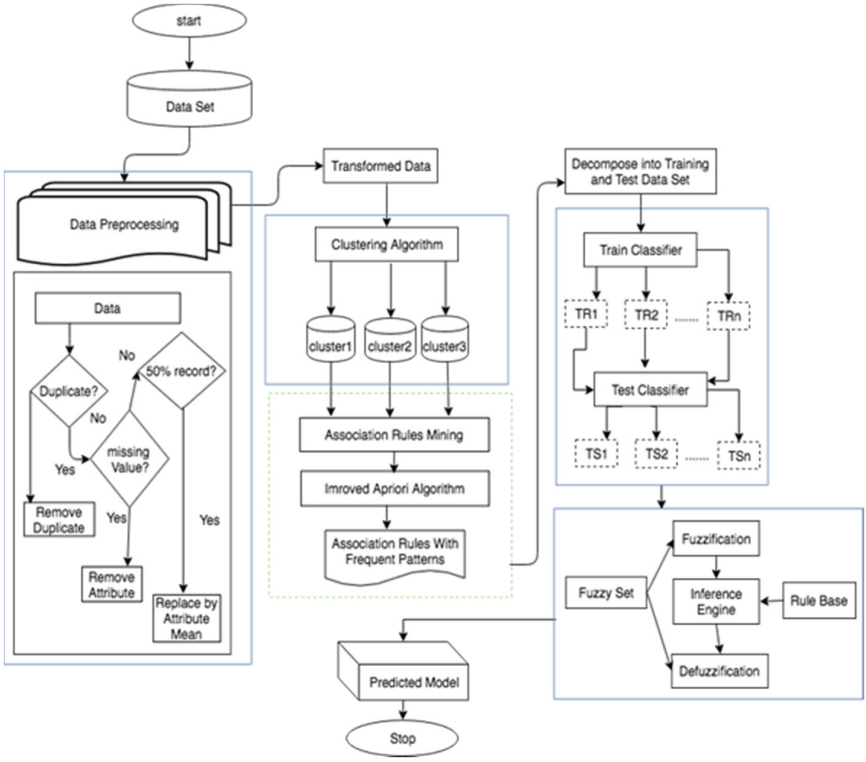

Fig. 2. Flow chart of novel system proposed

Processing time as expressed in [10], Apriori Algorithm is expressed using (1):

$$T = \sum_{i=1}^{n} t_s * m_k + t_c + 1_{k+1} * k + \frac{1}{2} * t_s * n_k / A \tag{1}$$

Where

t_s is the time cost of single scan of database
t_c be the time cost for generating candidate itemset
m_k be the amount of time in candidate itemset
n_k is number of item set in frequent Itemset

A denotes the amount of record in the database k is length of itemset.

From the features of (1), we propose (2) to be used in Improved Apriori Algorithm which is modified Apriori:

$$T = \sum_{i=1}^{n} (2E(|x|)) + k)n\tau \tag{2}$$

Where

n denotes the number of items
E is number of data points
|x| is the average complexity of each data point
m_k is the amount of time in candidate itemset
k is the length of itemset
τ is amount of time per non zero element

As per (2), item set is obtained for each number of iterations prolonging the scan of database which was then addressed in [2] and replaced by (3)

$$T = t_s * m_k + \sum_{i=2}^{k+1} (t_c + 2 * t_x * 1_{k+1}) \tag{3}$$

Where

t_s is the time cost of single scan of database
t_c be the time cost for generating candidate itemset
m_k be the amount of time in candidate itemset
k is the length of itemset

From the features of (3), we propose (4). It is done so because the candidate itemset obtained from overall set rather than obtaining in each iteration reduces processing time. If t_s is the time required for each scan of the database, then total processing time can be calculated as:

$$T = t_s * m_k + \sum_{i=1}^{n} (2E(|x|)) + k)\tau \tag{4}$$

Where

t_s is the time cost of single scan of database
t_c be the time cost for generating candidate itemset
m_k be the amount of time in candidate itemset
k is the length of itemset
E is number of data points
|x| is the average complexity of each data point
τ is amount of time per non-zero element

The main purpose of this change is to reduce the scale of database and reduce itemset generate from candidate set. Reduction of candidate set results in less scanning of database which reduces the processing time significantly thus reducing the time for feature extraction.

Data can be considered as sequence of binary vectors and can be represented by binary matrix which has defined number of rows and columns. The novel equation proposed in

(1) states that data are represented in columns because columns are generally in lesser numbers than rows which allows pointer to go through data faster to cluster data of similar behaviour or data which has occurred frequently. With the help of the proposed equation, association rules are applied much faster to the data. This is possible as data are not visited again.

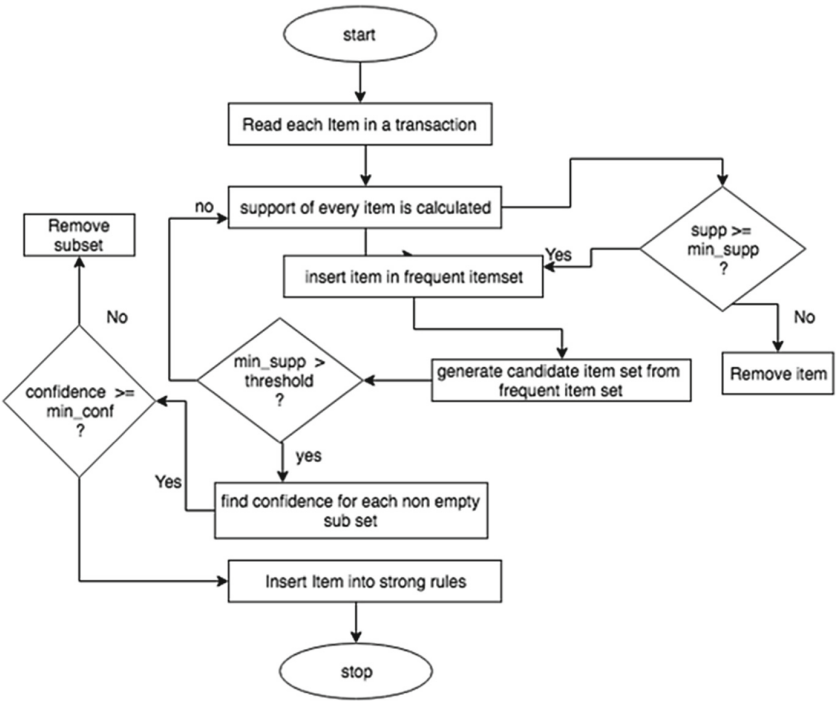

Fig. 3. Improved Apriori algorithm

With general association rules, the data are categorized and grouped with similar behaviour and attributes, which are passed into classifier to derive a prediction model. This proposed system derives a prediction model for the most accident occurring location based on the data supplied. The state of art does not provide the amount of time taken for to apply association rules to generate sets of data, which has similar behaviour, and attribute.

Time is one variant that must be taken under consideration while generating a set of records, which has similar behaviour, and attribute hence, the processing time must be minimized to its very best. With use of Improved Apriori Algorithm, it scans whole database to generate frequently occurring first item set which are then used to find second item set and so on until k number of item sets is reached. Thus, minimizes processing time as previously created cluster is used to create new one.

Current solutions have used association rules without consideration of time to process to form set of records. Our proposed solution has reduced processing time using Apriori algorithm as illustrated in Fig. 3.

Table 1 illustrates the proposed improved Apriori algorithm.

Table 1. Sample implementation of improved Apriori algorithm

Algorithm: Improved Apriori algorithm extracting rules to identify patterns

Begin
Step (1) first item set Lk= find frequent 1 item sets (T) ;
Step (2) For loop k from 2 to while Lk-1!= null;
//Generate the Candidate item set from the Lk-1
Step (3) Ck = candidates generated from L;
//get the item with minimum support in Candidate item set using L1, (1≤w≤k).
Step (4) x = Get items minimum support(Ck, L1);
// get the target transaction IDs that contain item x.
Step (5) TID = get Transaction ID(x);
Step (6) For each transaction t in TID Do
Step (7) Increment the count of all items in Candidate item that are found in TID;
Step (8) Lk= items in candidate itemset ≥ minimum support;
End;

4 Results

Java and IDE Net beans were used for the implementation. A code base program was designed to implement Improved Apriori Algorithm with the sample of 10 dataset. The dataset that was taken as sample to test have different number of data that varies from 500 to 4500 and different number feature, which vary from 45–300. We have specifically taken the sample that have different attributes which covers most of the requirement like accident severity, time and date of the accident, location characteristics etc. The dataset has been taken from different site like data.gov, (Menzies, 2018), which is free online resource and is available on the internet for the student whose focuses on data science. All these datasets are available in CSV and ARFF file format that can be read by many language using data analysis libraries. This library allows for creation of object that can access the method to read the CSV or ARFF file. The data were extracted from java and all of them have been considered. The extracted data were balanced using the XLMiner and then passed through the code, which produces the results as illustrated in Table 2.

The result is compared based on the number of items in the dataset and the number of features extracted along with the amount of time for single scan of dataset. In the dataset, generation of the candidate set is minimized which leads to a smaller number of scanning of dataset therefore decreases the processing time significantly. As shown in Table 2 the number of features extracted is 62, which is significantly less than the number of items. Improved Apriori extracts the candidate item in first scan only and process ahead, which decrease the average processing time extracting the feature as depicted in Table 2. The processing time of proposed solution is 2.119 s, which is 60% less than state of art.

Table 2. Sample implementation of improved Apriori algorithm

Dataset from data.gov	Samples (n)	Items in candidate set (m_k)	State of art		Proposed solution	
			Processing time	Accuracy	Processing time	Accuracy
Road traffic accident data	**915**	**62**	6.001	87.23	2.119	90.33

Samples were compared the state of art and the proposed solutions with the help of graphs and the data reports that are shown in Figs. 4 and 5. The results divided according to number of features extracted and the total amount of time to scan a dataset a single time Here, the results from the sample is presented in terms of average processing time. Processing time is the time in seconds to extract the feature, which is programmed well in java. We have done comprehensive test for 10 datasets. Then the result has calculated by taken the average for all datasets.

These results were compared during different number of transactions. The proposed solution has reduced the processing time by separating the candidate items generating for first scan of dataset as all the transactions are scanned as first it generates the item set with all the ids and support for all the transaction. If another candidate set is to be generated its will not scan all the transaction again but it will look at the item sets that was generated previously in first scan which holds all the id and support, so it can scan for only specific transaction where item exists.

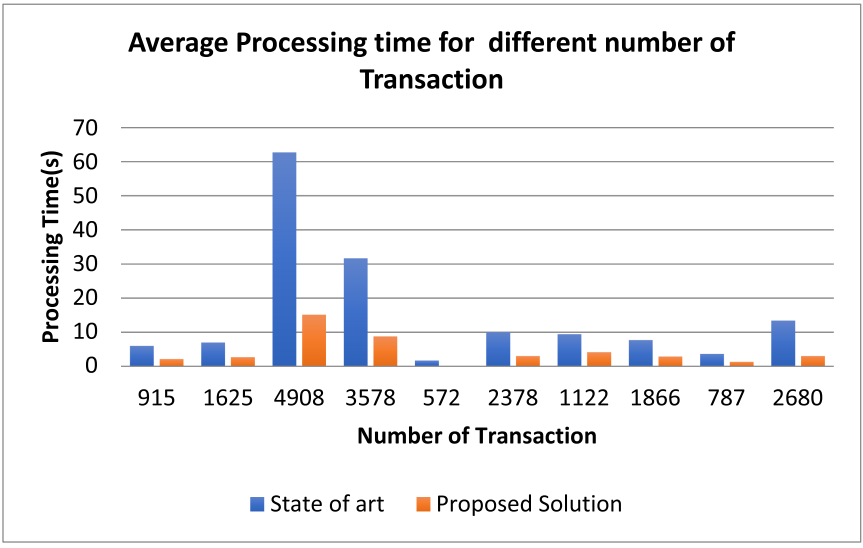

Fig. 4. Average processing time for proposed and current solutions for different transactions.

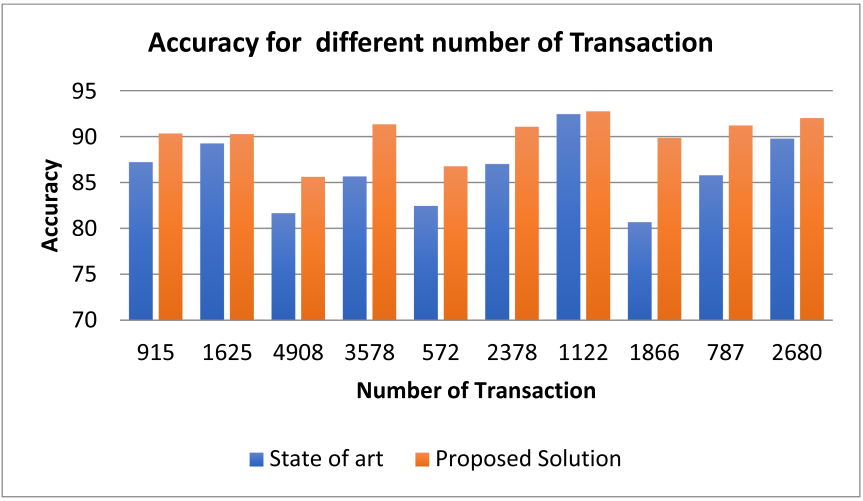

Fig. 5. Accuracy of proposed and current solutions for different number of transactions.

5 Discussion

Results show improved accuracy and decrease in time processing between the current solution and the proposed solution during the generation of prediction model as well as after prediction model are created. An accuracy of 92.75% and improvement in processing time by 67% was achieved, as the cardinal item sets were achieved before iteration rather than on individual iterations. Processing time decreased from 10 s to predict the model for over 50 records was reduced to 3.37 s. Both factors' accuracy and improvement in processing time was calculated and simulated with the help of Java Programming language. Improvement in processing time was calculated with the help of Java for both the state of art for current as well as proposed algorithms.

Use of Apriori Algorithm switched its performance from low to high by the reduction of processing time and helps create prediction model for the traffic road accident. Despite voluminous data, Improved Apriori Algorithm helps in search of most frequent item just once rather than in every repetition. Thus, use of Improved Apriori Algorithm for the prediction improved the processing time and accuracy for the traffic road accident prediction.

For road traffic accident prediction model, Apriori Algorithm has been implemented as a major component, which helped in achieving desirable processing time and accuracy. This research successfully provides a prediction model for road traffic accident with greater accuracy in lesser time. The state of art solution provides prediction model with 92.25% accuracy without ever mentioning the processing time. However, the proposed solution provides accuracy of 92.75% with the significant drop in processing time by 67%. This is verified and validated with the pseudo code for the use of Apriori Algorithm for a faster processing of data.

6 Conclusion

The proposed solution with all the necessary data for traffic road accident prediction model is made easier and faster with the use of Apriori Algorithm. It reduces processing time by 67% even on a large data. the Improved Apriori Algorithm has been tested to produce prediction model with high accuracy by reducing processing time. Information as obtained from the model also helps to plot reasons behind all road traffic accidents and helps in understanding them better.

Improved Apriori algorithm has proven the accuracy 92.75% with decrease in processing time with minimum support. Thus, it performs more efficiently as it finds the frequent occurring item in the database to build up the first item set to generate second and so on in which the item with low minimum support is removed. This process is known to as sampling and doing such a way that the efficiency is increased as processing time is decreased. Furthermore, the algorithm introduces the database mapping which avoids the repetition of database scan.

References

1. Amira, A., Vikas, P., Abdelaziz, A.: Applying association rules mining algorithms for traffic accidents in Dubai. Int. J. Soft Comput. Eng. **5**(4), 2231–2307 (2015)
2. Gu, X., Li, T., Wang, Y., Zhang, L., Wang, Y., Yao, J.: Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization. J. Algorithms Comput. Technol. **12**(1), 20–29 (2017). https://doi.org/10.1177/1748301817729953
3. Arwa, A., Mourad, Y.: Hybrid approach for improving efficiency of apriori algorithm on frequent itemset. Int. J. Comput. Sci. Netw. Secur. **18**(5), 7–10 (2018)
4. Janani, G., Devi, N.: Road traffic accidents analysis using data mining techniques. JITA: J. Inf. Technol. Appl. **14**(2) (2018). https://doi.org/10.7251/jit1702084j
5. Tavakoli Kashani, A., Shariat-Mohaymany, A., Ranjbari, A.: A data mining approach to identify key factors of traffic injury severity. PROMET Traffic Transp. **23**(1) (2011). https://doi.org/10.7307/ptt.v23i1.144
6. Mulay, P., Mulatu, S.: What you eat matters road safety: a data mining approach. Indian J. Sci. Technol. **9**(15) (2016). https://doi.org/10.17485/ijst/2016/v9i15/92119
7. Hashmienejad, S., Hasheminejad, S.: Traffic accident severity prediction using a novel multi-objective genetic algorithm. Int. J. Crashworthiness **22**(4), 425–440 (2017). https://doi.org/10.1080/13588265.2016.1275431
8. Jalayer, M., Pour-Rouholamin, M., Zhou, H.: Wrong-way driving crashes: a multiple correspondence approach to identify contributing factors. Traffic Inj. Prev. **19**(1), 35–41 (2017). https://doi.org/10.1080/15389588.2017.1347260
9. Park, S., Kim, S., Ha, Y.: Erratum to: highway traffic accident prediction using VDS big data analysis. J. Supercomput. **72**(7), 2832 (2016). https://doi.org/10.1007/s11227-016-1655-5
10. Krishnaveni, S., Hemalatha, M.: A perspective analysis of traffic accident using data mining techniques. Int. J. Comput. Appl. **23**(7), 40–48 (2011). https://doi.org/10.5120/2896-3788