Improving Road Safety in India Using Data Mining Techniques

Gaurav(✉) and Zunaid Alam(✉)

SGT University, Gurugram, India
gauravsingla31@gmail.com, zunaid.aalam6194@gmail.com

Abstract. Road accidents are very common in India. World Health Organization (WHO) has revealed that India the worst road traffic accident rate worldwide. According to the report, poor driving pattern, drunk driving, badly maintained roads and vehicles are the main triggering factors to road casualties. Statistics shows that one serious road accident in the country occurs every minute. The national capital, Delhi is among the deadliest. To achieve aim of reducing road accidents, novel and robust prevention strategies for improved road severity have to be developed. In this work, we propose use of data mining frame work to analyze traffic on National Highways of India. Using real data set of National Highway of India, we will mine important patterns for accidental data on National Highways of India and, identify key causes to road casualties. The discovered knowledge can be used by Ministry of road transport & highways of India to take effective decisions to reduce road severity.

Keywords: Data mining · Association rules · Road safety

1 Introduction

Road accident is one of the undesirable events that are uncertain and unpredictable. Road accident is one of the major causes of unnatural deaths, disability and property damage [2]. Many people die daily due to road accidents. World Health Organization has revealed India the worst road traffic accident worldwide. Statistics shows that one serious road accident in the country occurs every minute and 16 die on Indian roads every hour. According to the WHO report, poor driving pattern, drunk driving, low use of helmets, badly maintained roads and vehicles are the main contributing factors to road casualties. Road safety improvements can be achieved within the three components of the road safety system through changes in infrastructure design, vehicle safety and road user behavior [11]. However the main cause of road accidents also depends on the area in which it happens but still there may be some other reasons. So we can't be assure of road safety by considering only one or two reasons for road accidents in an area. It's very important to be accurate in case of finding reasons for road accidents in an area. India has 1% of vehicles in the world; but it accounts for about 6% of the total cases of unintentional injuries. According to the survey, male victims involved in the accidents are 78% and rest are females victims [3].

Data mining algorithms are extremely useful in prediction. Once we have the accurate accidental data, we can use data mining approach to find the main causes of

B. Panda et al. (Eds.): REDSET 2017, CCIS 799, pp. 187–194, 2018.
https://doi.org/10.1007/978-981-10-8527-7_17

road accidents in a specific area. In this paper, we have taken the accidental data from National Highway Authority of India (NHAI). As a sample, we applied the algorithm on accidental data from Panipat (Haryana) to Jalandhar (Punjab). In dataset, we had valuable information such as Time, Accident Location, Nature of Accident, Road feature, Road condition, Vehicle responsible for accident, Accident Type etc. We applied data mining methods such as association rule to determine the major factors for the accident.

2 Methodology and Tools

The following figure shows the general path/way which we have followed for mining association rules (Fig. 1):

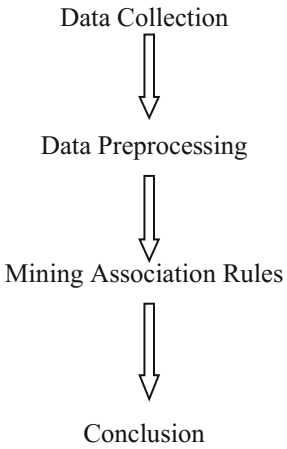

Fig. 1. Steps for mining association rules

As preprocessing of the data set takes 70% of the time to make data set useful and according to as per requirement of algorithms. The goal of data preprocessing is to choose cardinal features then remove irrelevant information and finally transform raw data into sessions. To achieve its goal Data preprocessing is divided into Data Cleaning, user identification, and Session Identification and Path Completion steps [6]. Noisy and erroneous data makes algorithms useless which we apply to mine data. So before processing, data needs to be investigated and preprocessed. Then only it makes it useful [7]. The most common issue that comes during the process of knowledge discovery through data mining includes the missing values. A dataset with 1–5% missing values may not impact as much whereby from 5–15% range requires sensitive algorithms to apply [9] (Fig. 2).

We applied preprocessing on almost every feature of the dataset to make it as per requirement, e.g. we extracted only year from the feature Date. Then we converted the time feature into six zones like early morning, morning, afternoon, evening, night, late night on the basis of timing of accidents.

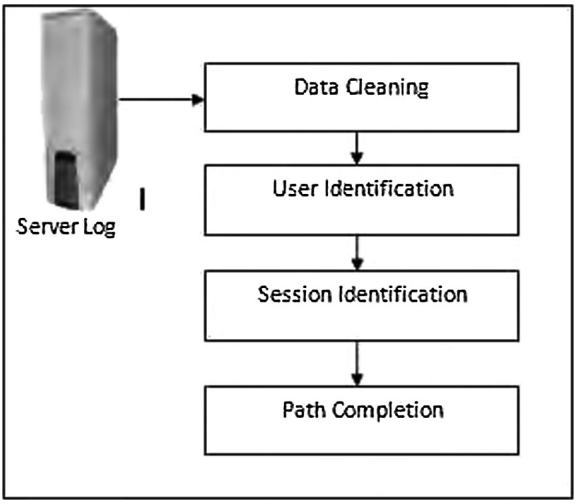

Fig. 2. Steps involved in data preprocessing [6]

Then there are some features such as Nature of accident, Causes of accident, Road features, Road conditions, weather conditions etc. which were in numeric form. We converted them according to their actual values. It's important because while mining association rules, without conversion, it's not possible to determine the actual value of the antecedent and consequent. Then we worked on the feature Accident type. Originally the accident type feature was partitioned into four features such as fatal, grievous, minor, injured. We converted these four features into one to indicate the kind of accident. Here we applied data preprocessing for data mining purpose. But it can also used be used for constructing on data warehouses, www etc. [8].

Once the data was prepared, we mined association rules. Association rule mining method is the most efficient data mining approach to find out patterns among the high volume of data. It is responsible to find correlation relationships among various data attributes in a huge set of items in a database [10]. The data set of association rules is denoted by T, T is the transaction database, T = {t1, t2… tk… tn}, tk = {il, i2,… im… ip}, tk (k = 1,2,… n) is called a transaction, im(m = l, 2,… p) is called a item. In the transaction database, item is the name of goods or services, transaction also includes other information such as date, customer number and so on. Simply put, transaction is a collection of items [1]. While mining association rules, we prioritize the most important parameters of association rules and these are support and confidence. However there are some other parameters as well but we have focused on support and confidence. *Support* is an indication of how frequently the items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true.

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. The major factor of accidents are shown in the result section for the data which we have uses for mining association rules. As far as tools are concerned we used Rstudio for preprocessing of the data. Initially we imported the data

from excel to Rstudio and then we preprocessed the data using R Language as per the requirement. Once the data was prepared, then we mined association rules.

The rules which we have mined are in the form X- > Y, where X is antecedent and Y is consequent. We have used all combination of support and confidence for mining association rules such as HS (High support) and LC (Low confidence), LS (Low support) and HC (High confidence), HS (High support) and HC (High confidence). Using different combination of support and confidence, the result which have been mined are accurate and effective for future use.

3 Results

As we mentioned earlier, we have applied different combination of support and confidence to mine association rules. So here in this section, we will see the rules which are mined from the data. First result/table shows the rules for high support and low confidence. Second result/table shows the rules for low support and high confidence. Third result shows the result/table for high support and high confidence. We will see the major factors involved in road accidents. As far as percentage for support and confidence is concerned, we have taken low support as 5–10% and high support as nearby 40%. In case of confidence, we have taken low confidence as 5–10% but high confidence as 80–90% that we can find some closely related data. As there is different role of support and confidence, we have taken all kinds of values so that we can check each angle of data and best association rules can be mined.

3.1 Following is the Table for High Support and Low Confidence

See Fig. 3.

Data	
# Transactions in Input Data	2869
# Columns in Input Data	8
# Items in Input Data	138
# Association Rules	10
Minimum Support	1100
Minimum Confidence	10.00%

Row ID	Confidence %	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Lift Ratio
1	81.2284334	Injured	Petrol Vehicle	1449	1502	1177	1.551560422
2	78.36218375	PetrolVehicle	Injured	1502	1449	1177	1.551560422
3	76.16511318	Petrol Vehicle	Punjab	1502	2032	1144	1.07538243
4	56.2992126	Punjab	Petrol Vehicle	2032	1502	1144	1.07538243
5	93.44375431	Injured	Four lanes	1449	2659	1354	1.008236672
6	50.92139902	Four lanes	Injured	2659	1449	1354	1.008236672
7	93.06102362	Punjab	Four lanes	2032	2659	1891	1.004107096
8	71.11696126	Four lanes	Punjab	2659	2032	1891	1.004107096
9	51.93681835	Four lanes	Petrol Vehicle	2659	1502	1381	0.992055472
10	91.94407457	Petrol Vehicle	Four lanes	1502	2659	1381	0.992055472

Fig. 3. Result for High support and Low confidence

Here we can see that mostly accidents are injured based accidents and happened by petrol vehicle. As the accidental data is from Panipat to Jalandhar wherein two states are involved (Haryana and Punjab), so we can clearly see that the high percentage of accidents are happened in Punjab state as compared to Haryana. On the highway, there are four lanes and at some place there are two lanes but statistics shows that the accident rate is high on four lanes. Statistics shows that there are more number of average accidents in Haryana as compared to other states [5]. So here it's important look over this matter.

Following is the Table for Low Support and High Confidence:

See Fig. 4.

Data	
# Transactions in Input Data	2869
# Columns in Input Data	8
# Items in Input Data	138
# Association Rules	25
Minimum Support	500
Minimum Confidence	85.00%

Row ID	Confidence %	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Lift Ratio
1	98.87429644	Left turn collision	Four lanes	533	2659	527	1.066830976
2	98.42406877	Fault of other driver & Punjab	Four lanes	698	2659	687	1.061973123
3	97.98994975	Fault of other driver	Four lanes	995	2659	975	1.057289078
4	97.34345351	Fault of other driver	Four lanes	527	2659	513	1.050313532
5	93.82488479	Injured & Punjab	Four lanes	1085	2659	1018	1.012348983
6	93.48127601	Car	Four lanes	721	2659	674	1.008641523
7	93.44375431	Injured	Four lanes	1449	2659	1354	1.008236672
8	93.37016575	Injured & Petrol Vehicle & Punjab	Four lanes	905	2659	845	1.007442668
9	93.06102362	Punjab	Four lanes	2032	2659	1891	1.004107096
10	92.97597043	Overturning	Four lanes	541	2659	503	1.003189391
11	92.83088235	LN1	Four lanes	544	2659	505	1.001623924
12	92.65734266	Petrol Vehicle & Punjab	Four lanes	1144	2659	1060	0.999751471
13	92.60832625	Injured & Petrol Vehicle	Four lanes	1177	2659	1090	0.999222595
14	92.17809868	M	Four lanes	831	2659	766	0.994580538
15	91.94407457	Petrol Vehicle	Four lanes	1502	2659	1381	0.992055472
16	91.89655172	M & Punjab	Four lanes	580	2659	533	0.991542711
17	91.7562724	Haryana	Four lanes	837	2659	768	0.990029129
18	91.58415842	Ambulance & Punjab	Four lanes	606	2659	555	0.988172059
19	91.40708915	Ambulance	Four lanes	931	2659	851	0.986261522
20	90.99249374	Rear end Collision	Four lanes	1199	2659	1091	0.981788133
21	90.98765432	Punjab & Rear end Collision	Four lanes	810	2659	737	0.981735917
22	90.28960818	Minor & Punjab	Four lanes	587	2659	530	0.974204159
23	89.33649289	Minor	Four lanes	844	2659	754	0.963920264
24	85.10452962	Overspeeding	Four lanes	1148	2659	977	0.918258351
25	85.07078507	Overspeeding & Punjab	Four lanes	777	2659	661	0.917894255

Fig. 4. Result for low support and High confidence

Here we have found some more interesting features involved in accidents for given specific data. This result shows that almost all accidents happened on four lanes but antecedents are different. As far as timing is concerned, Late night (LN) and morning (M) sessions are more dangerous time for accidents. Over speed is a dangerous factor for accidents and it can be seen here clearly. Other major reasons can be seen that while overturning or because of fault of other drivers there have been many accidents. Rear end collision is also an important feature that can be seen from the statistics. Unlike in previous statistics, here both states can be seen in statistics. Here one interesting feature can be seen that mostly vehicles are petrol vehicles as in previous case. Here we have taken low support but high confidence, it means whenever a rule occur, it occurs very strongly.

Following is the Table for High Support and High Confidence:

See Fig. 5.

Data	
# Transactions in Input Data	2869
# Columns in Input Data	8
# Items in Input Data	138
# Association Rules	9
Minimum Support	1000
Minimum Confidence	80.00%

Row ID	Confidence %	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Lift Ratio
1	81.2284334	Injured	Petrol Vehicle	1449	1502	1177	1.551560422
2	80.50221566	Four lanes & Injured	Petrol Vehicle	1354	1502	1090	1.537688793
3	93.82488479	Injured & Punjab	Four lanes	1085	2659	1018	1.012348983
4	93.44375431	Injured	Four lanes	1449	2659	1354	1.008236672
5	93.06102362	Punjab	Four lanes	2032	2659	1891	1.004107096
6	92.65734266	Petrol Vehicle & Punjab	Four lanes	1144	2659	1060	0.999751471
7	92.60832625	Injured & Petrol Vehicle	Four lanes	1177	2659	1090	0.999222595
8	91.94407457	Petrol Vehicle	Four lanes	1502	2659	1381	0.992055472
9	90.99249374	Rear end Collision	Four lanes	1199	2659	1091	0.981788133

Fig. 5. Result for High support and Low confidence

This statistics is for high support and high confidence. This result is almost as like derived from two previous results. Road feature is four lanes, vehicle involved are petrol vehicles. Accidents happened are injured based. There are some rear end collisions. As far as state is concerned, mostly accidents happened in Punjab instead of Haryana. As far as exact black spot is to be identified we have many spatial analysis methods. Different spatial analysis methods have been used in recent years for pedestrian safety studies. But KDE (Kernel Density Estimation) is the most used method for finding the Black spot and is effectively visualize and helps analysing the accident events spread across the Geography [4].

4 Conclusion

On the basis of association rules which are mined, we can conclude that on the four lanes, the major accidents are during late night and morning sessions. As far as vehicles are concerned, petrol vehicles are more involved in accidents as compared to other vehicles. Many of the accidents are due to fault of other drivers but it's not possible to control driving of others. Rear end collisions have occurred in many cases. So as per result, while using petrol vehicles, we should avoid over speed as this is also one of the major factors especially in late night or morning sessions. However most of the accidents are minor and injured based and very less serious accidents but still these may be reduced if we have prior intimation or information about such happenings. Now some algorithms such as Support vector machine can be used to build a model for traffic accident detection [12]. By using the model based on historical data, prediction can be performed on future data. Now this predictive information can be used by passengers as well as government to check if such are the cases then accident may occur. Govt. can apply some rules or protocols for such cases, in the conclusion of that there will definitely be less accidents.

References

1. Zhong, R., Wang, H.: Research of commonly used association rules mining algorithm in data mining. In: IEEE International Conference on Internet Computing and Information Services (2011). ISBN 978-0-7695-4539-4/11
2. Kumar, S., Toshniwal, D.: Analysing road accident data using association rule mining. In: IEEE International Conference on Computing, Communication and Security, December 2015. ISBN 978-1-4673-9354-6/15
3. Shruthi, P., Vanketesh, V.T., Viswakanth, B., Ramesh, C., Sujatha, P.L., Dominic, I.R.: Analysis of fatal road traffic accidents in a metropolitan city of South India. J. Indian Acad. Forensic Med. **35**(4) 2013
4. Raut, U.M., Nalawade, D.B., Kale, K.V.: Mapping and analysis of accident black spot in Aurangabad city using geographic information system. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **6**(1) (2016)
5. Shah, K.D., Sachdeva, S.N.: A comparative study of accident safety on Haryana road. IOSR J. Mech. Civil Eng., 13–16. e-ISSN 2278-1684, p-ISSN 2320-334X
6. Prince Mary, S., Baburaj, E.: An efficient approach to perform pre-processing. Indian J. Comput. Sci. Eng. **4**(5) (2013). ISSN 0976-5166
7. Aubrecht, P., Koub, Z.: A universal data preprocessing system. In: Popelínský, L. (ed.) DATAKON 2003, Brno, pp. 1–3, 18–21 October 2003
8. Pandey, K.K., Pradhan, N.: An analytical and comparative study of various data preprocessing method in data mining. Int. J. Emerg. Technol. Adv. Eng. **4**(10) (2014). ISSN 2250-2459
9. Aleem, A.S., Asif, K.H., Ali, A., Awan, S.M.: Pre processing methods of data mining. In: IEEE/ACM 7th International Conference on Utility and Cloud Computing (2014). ISBN 978-1-4799-7881-6/14
10. Somkunwar, R.: A study on various data mining approaches of association rules. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **2**(9) (2012)

11. Beshah, T., Ejigu, D., Abraham, A., Snasel, V., kromer, P.: Mining pattern from road accident data: role of road user's behavior and implications for improving road safety. Int. J. Tomography Simul. **22**(1) (2013). ISSN 2319-3336
12. Liang, G.J.: Automatic traffic accident detection based on Internet of things and Support vector machine. Int. J. Smart Home **9**(4), 97–106 (2015)