International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
©2017 INPRESSCO®, All Rights Reserved Available at http://inpressco.com/category/ijcet

Research Article

Survey on Analysis and Prediction of Road Traffic Accident Severity Levels using Data Mining Techniques in Maharashtra, India

Baye Atnafu[#*] and Gagandeep Kaur[#]

[#]Dept of CS/IT, Symbiosis Institute of Technology, Pune, India

Received 01 Sept 2017, Accepted 01 Nov 2017, Available online 10 Nov 2017, **Vol.7, No.6 (Nov/Dec 2017)**

Abstract

Traffic accidents are the main cause of death as well as serious injuries in the world. India is among the emerging countries where the rate at which traffic accident occurs is more than the critical limit. As a human being, we all want to avoid traffic accidents and stay safe. In order to stay safe, careful analysis of roadway traffic accident data is important to find out factors that are related to fatal, grievous injury, minor injuries, and non-injury. For this purpose, there are a number of classification association rule mining algorithms. From these, the survey paper discusses the algorithms that show better performance in the previous studies and also the survey examines the most widely used data mining tools. The proposed model implements by using the algorithm that shows better performance during the experiment to overcome the shortcomings of previous studies on accident severity prediction. Road traffic accident historical data is obtained from National Highway Authority of India (NHAI).

Keywords: *Road accident, data mining, random tree, J48, Naive Baye's, association rule mining*

1. Introduction

Annually due to road traffic accidents 1.25 million peoples die and 20-50 million peoples hurt non-fatal injuries (WHO report, 2015). According to the road traffic accident data provided by states, Maharashtra records the third highest number of fatal accidents (13,212) (NHAI report, 2016). However, this trend can change in future as it is hard to predict the rate at which road traffic accidents occur as it can occur in any situation. Therefore, we need to investigate the hidden pattern that influences the traffic accident severity levels using data mining techniques.

There are a number of Data Mining classification algorithms available (Like a Random tree, J48, Random forest, CART and Naïve Baye's) to predict the target class by analyzing the training dataset to get better boundary conditions which can be used to determine each target class. After determining the boundary conditions, the subsequent task is to predict the target class based on the boundary conditions.

There is also a number of Data Mining algorithms are available to find out the association between independent variables in a huge data. Association rule mining algorithm is the most popular methodologies to detect the significant associations between the data stored in the large database. There are a number of association rule mining algorithms available. From these Apriori, predictive Apriori and FP-growth algorithm are the most common association rule mining methods to find out the association between various road traffic accident severity factors that influencing the traffic accident severity levels in Maharashtra state, India.

2. Related work

In this study initially authors applied three popular classification algorithms such as a CART, Naïve Bayes, and SVM on PTW power two-wheeler accident data set and compared the results. CART classification algorithm accuracy was found superior to other two algorithms. Hence they have been used CART classification algorithms to find the various factors that influence the accident severity of power two-wheeler accidents in entire Uttarakhand state and its 13 districts separatel. The result shows that each district has different factors associated with power two-wheeler accidents severity (Sachin *et al*, 2017).

K-means clustering algorithm was used to investigate the high and low-frequency accident locations. Further, they have been used association rule mining to recognize the association between the various factors related to road traffic accidents at various places with changeable accident occurrences. The result shows that more accidents occur on

*Corresponding author **Baye Atnafu** is a M.Tech Scholar; ORCID ID: 0000-0001-7853-8386 **and Gagandeep Kaur** is working as Assistant Professor highways, foot-travelers are more vulnerable to road accidents at roads that have intersections, Curve on roads bordered by agriculture land are risky for multi-vehicle crashes and intersections on roads which fall upon marketplaces are more vulnerable to severe accidents (Sachin *et al*, 2016).

Logistic regression model were used to isolate and enumerate the impact of various roadways and environmental factors on the traffic crash severities and predict the accident severity levels. Authors investigated that factors like crash location, road function class, road alignment, light condition, road surface condition, and speed limit have the significant impact on crash severity. The results show that higher crash severity is linked with rural roadways, major arterials, locations without intersection, locations with curves, during night-time, dry roadway conditions, and high-speed limits (Yubian *et al*, 1997).

They have been evaluate the factors that affect the accident severity levels at urban road intersections using back propagation neural network and generalized linear mixed model. Both methods demonstrate that traffic flows have a significant role in predicting severity; this role is not limited to the flow when the crash occurred, but also extends to the other vehicle crash flow data before the crash occurs after the crash occurred (L. Mussone *et al*, 2017).

Logistic regression model they have been used to recognize the various factors contributing to increased vehicle crash risk during fog and investigate the situations in which crash risk are more likely to occur. The analysis results show that drivers will be more careful when fog is present and the chances of increasing crash risk would be more near ramp areas (Yina *et al*, 2017).

The Latent Class clustering and k-Modes clustering algorithms they have been used to form different homogeneous clusters using a heterogeneous road accident data. Further, FP growth algorithm is applied to the clusters formed to find out the algorithm that is better-performing when decreasing the heterogeneity of traffic accident data (Sachin *et al*, 2016). The results prove that there is no any clustering algorithms is superior to others, that means both the clustering techniques perform well when to reduce the heterogeneity nature of accident data (Sachin *et al*, 2016).

Authors investigated road accident severity per vehicle type by using log-normal regression techniques. The result of this study shows that bad weather situations and accidents during nighttime increase accident severity. Furthermore, authors concluded that there is a major impact of crash type while examining accident severity (Yannis *et al*, 2017).

Authors explored the factors of injury severity of truck drivers in the United States using an ordered probit regression model. The outcomes of analysis show bad weather and visibility condition enhance the chances of high-level injury severity in truck drivers'(Wei et.al, 2016).

They have been combined two diverse sources of data and applied random parameters (mixed) ordered logit model to consider the distinct heterogeneity in the data. Results depicted that rain, air temperature, humidity, air temperature, wind speed, and rain were found a factor for injury severity. From these factors, warmer air temperatures and rain were associated with high severe injuries while less severe injuries were associated with higher levels of humidity (Bhaven *et al*, 2016).

Authors were focused on recognizing the person, vehicle, and accident-related risk factors that are significant in building a variance in injury severity levels sustained by a driver in a car crash. The authors used a number of predictive analytics algorithms to find the composite associations between various stages of injury severity and the risk causes related to crash. The authors also found the importance of crash-related risk factors after applying a systematic series of information fusion-based sensitivity analysis on the trained predictive models. Sensitivity analysis results prove that use of a preventive system (i.e., seatbelt), the way of crash and drug usages are the main predictors of the injury severity (Dursun *et al*, 2017).

Authors were applied statistics analysis and data mining algorithms such as, Apriori rule mining, Naive Baye's and k-means clustering algorithm on the FARS Fatal Accident dataset for the purpose of investigating the relationship between fatal rate and other attributes such as weather condition, collision manner, light condition, drunk driver and surface conditions. The analysis result shows that environmental factors like roadway surface, weather, and light conditions do not strongly affect the fatal rate, while the human factors like drunk or not and the collision type have a stronger effect on the fatality rate (Liling a*et al*, 2017).

Using a binomial logistic regression model they have been identified the various causes that influence the motorcycle fatal crashes. Study results show that the chances to occur rear-end, sideswipe and head-on collision are 42 times, 35 times and 25 times more than hit pedestrian for variable "collision type", respectively; the probability of fatal crash increase in single-vehicle crashes than two or more vehicles crashes for variable "number of vehicle", and, the probability fatal crash on two-lane national highway is more than four-lane national highway for variable "number of lane" (Hasan *et al*, 2016).

Table 1: Data mining algorithm's reviews

Authors	Title	Methods	Algorithm Performance	Objective	Result
(Dursun et.al, 2017)	Investigating injury severity risk factors in automobile crashes	ANN SVM C5 LR	85.77% 90.41% 86.61% 76.96%	To identify the person, vehicle, and accident-related risk factors in automobile crashes.	Use of seat belt, the manner of collision and drug usage are the top predictors of the injury severity.
(Wei *et al*,2016)	Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States.	Ordered probit model	-	To analyze the effect of various factors on injury severity.	Bad weather and road visibility found to increase the probability high level of injury severity.
(L. Mussone *et al*,2017)	Analysis of factors affecting the severity of crashes in urban road intersections	BPNN & GLMM	BPNN model shows better performance than the GLMMs.	To identify factors that influence crash severity level at urban road intersections.	Flows have a significant role in predicting severity.
(Yina *et al*, 2017))	Crash risk analysis during fog conditions using real-time traffic data	CRII & LR model	CRII were identified the dangerous traffic status under fog conditions	To investigate the situations in which crash risk are more likely to increase.	Investigate the change of traffic flow and traffic safety under fog conditions.
(Sachin *et al*, 2016)	A comparative analysis of heterogeneity in road accident data using data mining techniques	LC clustering k-modes clustering FP growth technique.	They did not find any difference.	To identify the most influential factors that affect the road accident severity.	The rules generated for each cluster do not prove any cluster analysis technique superior over other.
(Alexander *et al*, 2015)	The relationship between age with driving attitudes & behaviors among older Americans	LR model	It helps to identify the association between dependent and independent variables.	To identify the relationship between age with driving attitudes & behaviors.	Younger drivers engage more in unsafe traffic safety compared to older drivers.
(Liling *et al*, 2017)	Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques	Apriori algorithms Naive Bayes k Means clustering	67.9%	To find out variables that are closely related to fatal accidents.	The environmental factors were do not strongly influence the fatal rate, while the human factors have been a stronger influence on the fatal rate.
(Hasan *et al*, 2016)	Factors Contributing to Motorcycle Fatal Crashes on National Highways in India	LR models	It helps to explore various factors	To identify the factors that influence the motorcycle fatal crashes.	Collision type, number of vehicles, number of lanes, and time of the crash were observed to have a significant effect on motorcycle fatal crash.
(Yubian *et al*, 1997)	Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities	LR model	P-values for all the coefficients were are less than 0.0001	To identify the impacts of roadway and environmental factors to the traffic crash severities.	Road function class, crash location, road alignment, light condition, road surface condition, and speed limit has significant impacts on traffic crash severity.
(Yannis *et al*, 2017)	Investigation of road accident severity per vehicle type	LR model	It helps to explore the accident severity levels	To investigate the road accident severity per vehicle type.	Good weather conditions and crashes during the night were found to be increased accident severity.
(Sachin *et al*, 2016)	Data mining approach to characterize road accident locations	Association Rule Mining and k means clustering	Both technique's help to identify locations at which the accident occurs frequently	To identify various factors that affect road accidents	They have been identified high, moderate and low-frequency accident locations.

3. Methodology

1) Classification Algorithms

The classification algorithm is one of the data analysis methods that used to construct models to predict the future data. The type of classification algorithms used varies according to the target variable. The target variable for this survey paper is represented as a category variable with four possible outcomes (fatal, grievous injury, minor injury, and non-Injury) severity. Accordingly, the analyzing problem is characterized as a nominal classification problem and based on the extant literature there are various data mining techniques available that can handle this type of classification problem such as a Random tree, j48, Naïve Bayes etc.

A. Random Tree

A random tree (RndTree) is a group of distinct decision trees, which means that operator of random tree works just like the decision tree operator except, for each split, only a random subset of attributes is accessible. A RndTree is a tree haggard at a chance from the collection of achievable trees. In this perspective "at random" signifies that every tree has an equal chance of being selected from the set of trees.

B. J 48

J48 is an advanced version of ID3; it decides target value of a new test data with respect to diverse attribute values of training dataset (Sewaiwar *et al*, 2015). The inner nodes of a decision tree are represented by different attributes while the branches tell the achievable values of these attributes. The internal nodes denote the dependent variable values (Sewaiwar *et al*, 2015). Escalating the count of trees provides a more intelligent learner just as having a large varied group is capable of reaching intelligent conclusion (Zhao *et al*, 2008).

C. NAÏVE BAYE'S

The naïve Bayesian classifier is one of the most effective and widely used supervised learning algorithms to classify the road accident data. It is a statistical model that predicts class membership probabilities based on Bayes' theorem. The Naive Baye's classification algorithm is one of the probability-based methods used for classification and prediction based on the Bayes' hypothesis with the assumption of independence between each pair of variables.

2) Association Rule Mining Algorithms

Association rule mining algorithm is the most popular methodologies used to detect the significant associations between the data stored in a huge database. For this purpose there are a number of association rule mining algorithms present, from these Apriori, predictive Apriori and FP-growth association rules mining algorithm are the most unusually used algorithms in the area of road traffic accident analysis, to generate the best rules that show the association between various attributes in large datasets.

A. Apriori Algorithm

Apriori rule mining algorithm is the naive method of finding the frequent item-sets in a huge database by generate a set of all possible combination of items and then compute the support for them. However, the number of possible combinations increases exponentially as the number of items in item-set increases making this method impractical (Kenneth *et al*, 2001).

B. Predictive Apriori Algorithm

The predictive Apriori algorithm is also used for discovering hidden and novel patterns in a large database. It varies from Apriori algorithm in that both confidence and support measures are joined into a unique measure called as predictive accuracy (Sunita *et al*, 2001).

C. FP-Growth Algorithm

Frequent-pattern growth association rule algorithm is the enhanced version of the Apriori rule mining algorithm present by Jiawei *et al*. It compresses data sets to an FP-tree, scans the database twice, does not produce the candidate itemsets in the rule mining process, and greatly improves the mining efficiency (Jun *et al*, 2013). But FP-Growth algorithm needs to create an FP-tree which contains all the datasets. This FP-tree has high requirement on memory space (Zeng *et al*, 2015).

3) Data Mining Tools

Data Mining allows discovering novel patterns that are not discovered yet by using various open source data mining tools. Currently, there are many tools are available for data mining, Such as WEKA, RAPID MINER, R, KNIME…etc.

A. WEKA

Weka is one the most widely used tool for finding hidden patterns established by University of Waikato (Rangra *et al*, 2014). Weka offers three means to use the tool: the Java API, a GUI, and a command line interface (CLI). WEKA contains classification, clustering, association rules mining algorithms and data preprocessing tools (Rangra *et al*, 2014).

B. Rapid Miner

Rapid Miner is one of the open source tools in data mining developed by Ingo Mierswa and Ralf Klinkenberg. Rapid Miner also knew as YALE (Yet another Learning Tool) based on XML used to implement numerous machine learning and data mining classification and clustering algorithms (Rangra *et al*, 2014).

C. R

R is open source data mining tool based on C and FORTRAN programming language recognized by Ross Ihaka and Robert Gentleman for statistical computing and charts. R provides less support to data mining algorithms as compared to Rapid Miner and Weka, it does implement a few data mining algorithms (Bhinge *et al*, 2015).

4) KNIME

KNIME is an easily operable data mining tool contains platform for data integration, data processing, data analysis, and exploration that runs inside the IBM's Eclipse (Patel *et al*, 2015). KNIME is easy to extend and to add plugins (Patel *et al*, 2015).

Table 2: Comparison of most widely used data mining tools in accident analysis

Tools		R	WEKA	RAPID MINER	KNIME
1	Memory Usage	More memory	Less Memory	more memory	-
2	Language	C, Fortran and R	Java	Language Independent	Java
3	speed	Works faster on any machine	Works faster on any machine	Requires more memory to operate	-
4	Usage	Complicated to use	Most easiest to use	Easy to use	Easy to use
5	Interface Type Supported	CLI	GUI/CLI	GUI	GUI

5) Data Preprocessing

The dataset used in this study is obtained from National Highways Authority of India (NHAI) which covers accident historical data from September 2014 to July 2017. The dataset contains 19,167 accident records and 10 attributes after preprocessing. We planned to use WEKA 3.8 data mining tools from various data mining tools based on the future shows during review, for the purpose of classification, prediction, model evaluation, attribute selection, data cleaning, data integrating and managing road accident data obtained from National Highways Authority of India. Figure 1 shows the general block diagram of the proposed work. The first task is data preprocessing which include tasks such as data cleaning, integration, transformation, and reduction. After once the data is preprocessed, the next step is to apply the data mining techniques on the data.

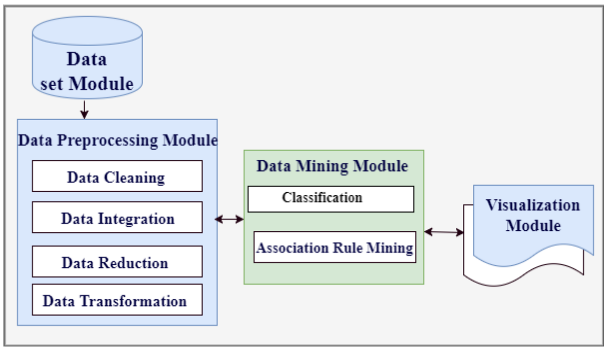

Figure 1: Block Diagram of Proposed work

There are a number of Data mining techniques available but the proposed method uses three Classification algorithms: Random tree, J48 and Naïve Baye's for predictions and Association Rule Mining algorithms to detect the significant associations between the data stored in the large database. After applying these algorithms, the next step is to visualize the outcomes obtained from experiments. The detailed process of the above-mentioned tasks is shown in figure 2.

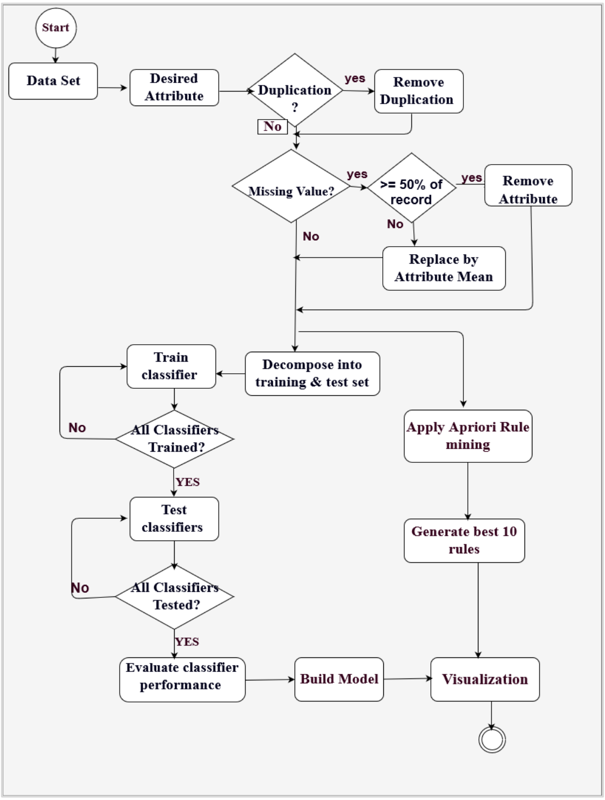

Figure 2: Flowchart diagram for proposed work

The flow chart shows each step followed throughout the study starting from data collection up to prediction of the road accident severity levels. After collecting the dataset from National Highways Authority of India feature selection method is applied to select the desired attributes. After this, selected desired attributes are checked for duplication, missing values, and outliers. After preprocessing, the dataset is decomposed into two sets: training and testing sets.

Next step is applying the classification algorithms on the data set and test whether all classifiers are trained or not. If all the classifiers are trained then test the classifier and generate results. Then accident severity prediction is done, further, we applied the Apriori algorithm to discover the relationship between various factors that frequently influence the severity of an accident. Finally, the result is interpreted for both classification and association rule mining technique

Conclusion

Paper, Survey on Analysis and Prediction of Traffic Accident Severity Levels Using Data Mining Techniques in Maharashtra, India discusses the latest work in the field of road accident analysis and prediction. Road traffic accident severity keeps on changing over time and increase endlessly. The changing and increasing road traffic accident severity leads to the issues of not understanding the accident behavior, factors influencing the traffic accident severity, and managing large volumes of data obtained from various sources properly. Many researchers have tried to solve these issues but still, there are gaps in the road accident severity prediction and finding the contributory factors such as season and time of the accident in which the accident frequently occurred. This leads to the challenges in the field of accident analysis and prediction. Some of the challenges include modeling of accidents for finding suitable algorithms to detect the accident severity levels, data preparation, transformation, and processing time. Therefore, in order to fill some of the gaps, this study identifies the suitable algorithms, tools, review of recent studies and models on accident severity analysis and prediction, which helps to extract hidden road traffic accident patterns for future.

References

WHO. Road traffic safety. Available from http://www.who.int/mediacentre/factsheets/fs358/en/ Accessed on 21 September 2017.

400 road deaths per day in India; up 5% to 1.46 lakh in 2015 Available from http://timesofindia.indiatimes.com/life style/health-fitness/health-news/400-road deaths-per-day-in-India-up-5-to-1-46-lakh-in-2015/articleshow/51920988.cms accessed on 21 September 2017.

Kumar, S., & Toshniwal, D, (2017), Severity analysis of powered two-wheeler traffic accidents in Uttarakhand, India, *European Transport Research Review*, 9(2), 24.

Kumar, S., & Toshniwal, D, (2016), A data mining approach to characterize road accident locations, *Journal of Modern Transportation*, 24(1), 62-72.

Wang, Y., & Zhang, W, (2017), Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities, *Transportation Research Procedia*, 25, 2124-2130.

Mussone, L., Bassani, M., & Masci, P, (2017), Analysis of factors affecting the severity of crashes in urban road intersections, *Accident Analysis & Prevention*, 103, 112-122.

Wu, Y., Abdel-Aty, M., & Lee, J, (2017), Crash risk analysis during fog conditions using real-time traffic data, *Accident Analysis & Prevention*.

Sachin Kumar, Durga Toshniwal, Manoranjan Parida,(2016), A comparative analysis of heterogeneity in road accident data using data mining techniques, *Evolving Systems*.

George, Y., Athanasios, T., & George, P., (2017), Investigation of road accident severity per vehicle type, *Transportation Research Procedia*, 25, 2081-2088.

Hao, W., Kamga, C., Yang, X., Ma, J., Thorson, E., Zhong, M., & Wu, C., (2016), Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States, *Transportation research part F: traffic psychology and behavior*, 43, 379-386.

Naik, B., Tung, L. W., Zhao, S., & Khattak, A. J., (2016), Weather impacts on single-vehicle truck crash injury severity, *Journal of safety research*, 58, 57-65.

Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E., (2017), Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods, *Journal of Transport & Health*, 4, 118-131.

Li, L., Shrestha, S., & Hu, G., (2017), Analysis of road traffic fatal accidents using data mining techniques, *In Software Engineering Research, Management and Applications (SERA)*, IEEE 15th International Conference on (pp. 363-370). IEEE.

Naqvi, H. M., & Tiwari, G., (2017), Factors Contributing to Motorcycle Fatal Crashes on National Highways in India, *Transportation Research Procedia*, 25, 2089-2102.

Classification, prediction, and clustering. Available from http://shareengineer.blogspot.in/2012/09/classification-and-clustering.html accessed on Accessed on 22 September 2017.

Sewaiwar, Purva, and Kamal Kant Verma,(2015), Comparative Study of Various Decision Tree Classification Algorithm Using WEKA. *International Journal of Emerging Research in Management &Technology* 4: 2278-9359.

Rapidminer.Random tree. Available from https://docs.rapidminer.com/studio/operators/modeling /predictive/trees/random_tree.html accessed on 22 September 2017.

Zhao, Yongheng, and Yanxia Zhang,(2008), Comparison of decision tree methods for finding active objects, *Advances in Space Research* 41.12 (2008): 1955-1959.

Lai, K., & Cerpa, N., (2001), Support vs. confidence in association rule algorithms. In Proceedings of the OPTIMA Conference, Curicó.

Aher, S. B., & Lobo, L. M. R. J., (2012), A comparative study of association rule algorithms for course recommender system in e-learning, *International Journal of Computer Applications*, *39*(1), 48-52.

L. Zhichun and Y. Fengxin, (2008), an improved frequent pattern tree growth algorithm, *Applied Science and Technology*, vol. 35, no. 6, pp. 47–51.

C. Jun and G. Li, (2013), an improved FP-growth algorithm based on item head table node, *Information Technology*, vol. 12, pp. 34–35.

B. Zheng and J. Li, (2008), an improved algorithm based on FP-growth," *Journal of Pingdingshan Institute of Technology*, vol. 17, no. 4, pp. 9–12.

S.K. David, Amr T.M. Saeb, K.A. Rubeaan,(2013), Comparative Analysis of Data Mining Tools and

Classification Techniques using WEKA in Medical Bioinformatics, *Computer Engineering and Intelligent System*, 4(13).

Rangra et al. (2014), comparative Study of Data Mining Tools". *In Proceedings of International Conference on Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 6, pp. 216-223.

Bhinge, A. V., (2015), A Comparative Study on Data Mining Tools (Doctoral dissertation, California State University, Sacramento).

Patel, P. S., & Desai, S. G., (2015), A Comparative Study on Data Mining Tools, *International Journal of Advanced Trends in Computer Science and Engineering*, 4(2).

National Highway Authority of India. Regional Office – Mumbai. Available from http://nhai.org.in accessed on July 30, 2017.