International Journal of Research in Engineering, Science and Management
Volume-2, Issue-7, July-2019
www.ijresm.com | ISSN (Online): 2581-5792

Analysis of Classification Algorithms to find the Key Factors Causing Road Accidents

T. Harini[1], A. Jerline Amutha[2]
[1]M.Phil. Student, Dept. of Computer Science and Technology, Women's Christian College, Chennai, India
[2]Assistant Professor, Dept. of Computer Science and Technology, Women's Christian College, Chennai, India

***Abstract*: Road calamities have become the root cause of increasing fatal death count that hands-on with roads, driver, vehicle-condition and weather. Concerning the safe longevity, it entails the data analytic tool to take the measures for diminishing the death toll. Data mining is one such area that identifies the underlying reason behind every accident. Classification technique is the most likely preferable for the prediction steps. At the same pace, the increases in death count stimulates to discover the very fine methodology to handle the huge amount of accidental data in an effective way. This research led to uncover the fact that serves as the root cause for the road calamities using the Decision tree and K-Nearest Neighbor (KNN) algorithm to prevent the human loss.**

***Keywords*: Data mining, Decision Tree, K-Nearest Neighbor, Road accidents.**

1. Introduction

Every day over 3,700 people were dying on the world's roads and which comes around of 1.35 million lives were ends on a year. Tens of millions more are injured or disabled every year, which is making them to do their daily routine as difficult as nothing else. According to World Health Organization (WHO), Road accident is ranked as the eighth leading cause of death globally. Road traffic injury is the leading cause of death for people aged between 5 and 29 years but these crashes were not considered as accidents because they are completely preventable [1]. Deaths and injuries stem from road crashes continues to be serious issues globally and the current trends were predicting that this will continue by beating the developing countries economic growth in the foreseeable future [2].

Mining holds a great potential to identify the best strategy for satisfying the future needs. Data Mining is widely used for Healthcare, Education, Manufacturing sectors, Customer Relationship Management (CRM), Financial Banking, Corporate Surveillance, Criminal Investigation and Bio Informatics [3]. The essence of technological advances leads a way for recording every accidental detail made available as digital data. It made easier to go through a huge dataset searching for the valuable information by applying the Data Mining methods in fraction of seconds. The road crashes are the notable cause for increasing death counts globally where the Classification algorithm can be applied to reveal the hidden factor which pays for fatal death. The key objective is to analysis the road accident dataset to find a factor that influences to the accident and to form best fitting policies to avoid it. The common bottom-line that involves in accident are Drive nature, vehicle nature, road natures and weather nature [4].

In general, the data will be broadly classifies as trained data and testing data. The classifier will be constructed based on the trained data and the obtained classifier would be used for the testing in order to identify the reason behind the accidents. The commonly used Classification algorithms for road accidents are Decision Tree, Neural Network, K-Nearest Neighbor and Navie Bayesian [5]. The aim of this research is to identify the factor that influenced to the road crashes. The literature review has been done to find a best Classification algorithm for the analysis of road calamities.

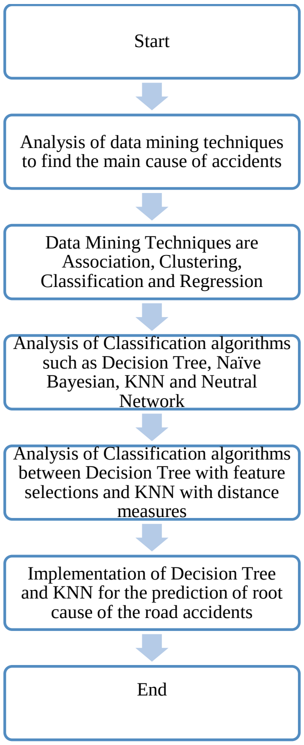
Start
Analysis of data mining techniques to find the main cause of accidents
Data Mining Techniques are Association, Clustering, Classification and Regression
Analysis of Classification algorithms such as Decision Tree, Naïve Bayesian, KNN and Neutral Network
Analysis of Classification algorithms between Decision Tree with feature selections and KNN with distance measures
Implementation of Decision Tree and KNN for the prediction of root cause of the road accidents
End

Fig. 1. Flow of the literature review

The Utilization of Decision Tree, in road crashes is of great importance because of its effectiveness in identifying reason behind every road accident.

2. Related works

A. Data mining techniques for an accidental data

Data Mining brought a great revolution in decision making by replacing the earlier statistical methods[6]. It is used for extracting the hidden pattern or information from the huge dataset[5]. It has also been quick in collaborating with ideas from other areas, including optimization, evolutionary computing, information theory, signal processing, visualization and information retrieval [7]. It is also called as knowledge extraction [8]. Data mining has two major categories depends on the process that undergoes for the discovery of information. The major categories are supervised and unsupervised techniques [9].

- Supervised or Predictive techniques: They are characterized by learning the thing before from the training data, then observing the test data and figure out the class to which it belongs. It is subdivided into Classification and Regression [10].
- Unsupervised or Descriptive techniques: They are characterized only by an observation. Since the data are unlabeled, there won't be any training data. Here they will derive the pattern that reveals the underlying relationships among the data. It is subdivided into Clustering and Association Rules[10].

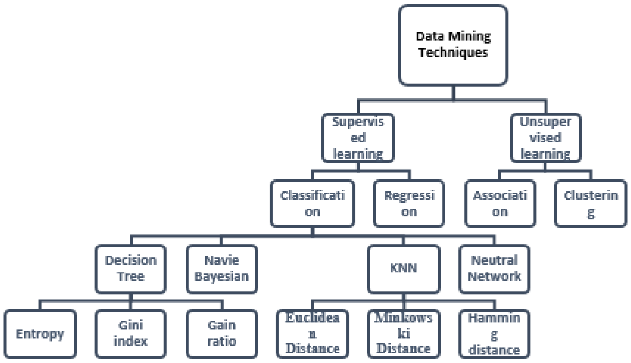

Fig. 2. Data mining techniques

Predictive modeling, clustering and association rules are the branches of data mining, which are used by researchers in order to solve the issues that raise the fatal death count. In Predictive model, the value for target events are known beforehand and the task of building the model will be done. With that model, the future values will be predicted for the unknown events [11]. There are two types of predictive modeling: classification and regression [8]. Classification will explore the features of unknown events and categories it with the predefined class of the known events [11]. Regression will be preferred for the prediction of variables in case of continuous values (range of numeric values). Clustering will form a group that has closely related objects. So the objects within the group will have a high similarities and behavior than objects in some other group [12]. Association rules, identifies the pattern or relationship among the objects that make them strongly associated together [8].

B. Impact of Classification algorithms on an accidental data

In accidental dataset, the first and foremost step is to find major causes that lead to road crashes. Here the four factors are categorized as major contribution to calamities such as drive condition, vehicle condition, road condition and weather. The classification methods will have depth insight into it and identifies the label for the classes like drunken drive, nerve issues, climatic impact, road condition, vehicle brake and eye issues [13]. The classification follows the mechanism of dividing the huge dataset into two groups. One group is for training purpose and other one for testing in the ratio of 3: 1 manner [14]. The training data are enclosed with the predetermined classes and desired outputs in order to build a classifier that train the machine to make prediction on class label on the future data [15]. The dataset which is chosen may hold incomplete, noisy or irrelevant data entries. The data preprocessing steps are needed to have a consistent and trustworthy decision making [16].

This proposed work investigates the following classification algorithms like decision tree, K-Nearest Neighbor, Naïve Bayesian and Support Vector Machine. Decision Tree algorithm is a divide and conquers approach to the breaks down a dataset into several subsets and generates decision tree in top-down manner [17]. K-Nearest Neighbor (KNN) algorithm is well known for pattern recognition, follows non parametric and also known as lazy learning algorithm. Naive Bayesian algorithm is a probabilistic classifier that incorporates strong independence assumptions [18]. Support vector Machine (SVM) is based on statistical learning theory of both linear and nonlinear data by a separating hyper-plane [19], [20]. The best suitable algorithm for accidental data will be identified and implemented for further process to achieve a high accuracy.

C. Understanding the classification algorithms decision tree and KNN

1) Analysis of decision tree

Based on the accuracy of the various classification algorithms that carried out the researches on an accidental data stated in table 1, shows that decision tree and KNN are the best suited ones among them. Decision tree is a simple and fast to use in any domain to discover the solution for a problem [21]. Decision tree presents the discovered outcomes in tree like a structure that composed of root node (starting node), internal node (act as a decision node to perform a test on the attributes to find a best split for the dataset) and leaf or terminal node (holds the outcomes of the decision tree) [22]. It doesn't require any prior assumptions that would satisfy the distribution of classes and attributes. So, it would be more preferable as classification methods [23].

The diverse types of decision tree algorithms that provides a potential supports to handling any types of data are CHID, CART, ID3 and C4.5 [24]. Chi-squared Automatic Interaction Detection (CHID) was developed in 1980[25] that are best fitted for all varieties of categorical and continuous variables. Classification and Regression Tree (CART) was developed in 1984 [25]. It produces binary tree known as Hierarchical Optimal Discriminate Analysis (HODA) that generates either classification or regression tree that depends on the types of data which it intakes [24]. ID3 was developed by Ross Quinlan which is used to generate tree in top down manner with given datasets as its inputs and without back tracking [21], [22].The attribute with highest information gain will be the next test attribute for the further building of decision tree [24]. C4.5 was developed by Ross Quinlan and to overcome the drawbacks of ID3 algorithms. Since it is an extension of ID3, it can handle continuous and categorical data with minimum level of missing values and based on the created threshold values, it is splits the dataset [21].

The construction of Decision tree need to undergo the feature selection step in order to find out the most prominent attributes among the others which gives more accurate supports in identifying the cause for the road accidents. By this way the most relative attributes or parameters will be picked and used for building the classifier which grants more explicit results. This research will proceed with comparing the Entropy, Gini index and Gain ratio for the feature selection process and best suited will be used the identifying the reason behind every accident.

2) Analysis of KNN

K Nearest Neighbor is a simple algorithm that classifies the new cases with similarity among the given inputs based on distance functions. It follows the non-parametric technique that has high expectation on pattern recognition because of its easy understanding of its working protocols. Since from its foundation, it undergoes many revised modifications to have a flawless mechanism that adapts to all specifications made. It is known for its lazy learner because it doesn't make any prior assumptions from the training dataset instead it make use of this training data for the calculation of distance measure [26]. It handle high dimensional data well [8].

The basic mechanism of KNN is to randomly select k points which are treated as initially center. The left out points will be assigned to the cluster of similar features [20]. This step will be repeated until there is no change with points that assigned to cluster. The most popular distance measure is Euclidean distance, which calculates as the distance between two objects in Euclidean space [5]. The other distance measures are hamming distance, Manhattan distance and Minkowski distances [27]. This research will be proceeds by finding the best distance measures among the Euclidean, Manhattan and Minowski measures that suited for handling the road calamities data.

3. Results and discussion

The Literature review reveals that the Classification algorithms such as Decision tree and KNN are constantly maintaining their contribution in handling the accidental data in an effective way. While discussing the decision tree construction, the classifier will not do its work properly if the attribute is not selected according to the research preference.

Table 1
Accuracy of the classification algorithms on accidental data

No.	Authors	Title	Algorithm used	Accuracy in %
1	Lun zhang et al. (2013)	An Improved K-nearest Neighbor for Short-term Traffic Flow Prediction	KNN	90
2	Roop Kumar R et al. (2018)	Data Analysis in Road Accidents using ANN and Decision Tree	ANN Decision Tree	79 79.8
3	Dheeraj Khera et al. (2015)	Prediction and analysis of Injury Severity in Traffic System using Data Mining Techniques	Naïve Bayes ID3 Random Forest Naïve Bayes ID3 Random forest	50.7 25.35 45.07 67.6 57.74 92.95
4	Elfadil A. Mohamed (2014)	Predicting Causes of Traffic Road Accidents using Multi-class Support Vector Machines	SVM	75
5	Velivela Gopinath et al. (2017)	Traffic Accidents Analysis with respect to Road Users using Data mining techniques	SVM Naïve Bayes Decision Tree	75.58 74.46 75.76
6	Miao M. Chong et al. (2004)	Traffic Accident Analysis Using Decision Trees and Neutral Networks	Decision Tree ANN	89.46 75.51
7	Tibebe Beshah et al.	Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia	Decision Tree Naïve Bayes K-Nearest Neighbors	80.22 79.99 80.81
8	Sharaf Alkheder et al. (2016)	Severity Prediction Of Traffic Accident using an Artificial Neural Network	ANN	74.6
9	A. Priyanka et al.	A comparative study of classification Algorithm using accident data	SMO J48 IBK	94 92.8 88.3
10	B. Lavanya et al. (2017)	Predictive Analytics on Accident data using rule based & discriminative	SVM DT	98.67 90.67

This results in formation of clumsy or large decision tree. To overcome this problem, the research will undergo the feature selection process to avoid the misleading formation of tree. Likewise, the KNN also has few flaws in randomly selecting and grouping of points. To avoid this issue, the different distance measures are taken in count to find out the best suited measures for accidental data. Finally, the results from both Decision tree and KNN will be compared in further enhancement.

4. Conclusion

Generally, the road accident analyses are based on driver condition, road conditions and weather status but vehicle condition also consider in this research because it is also one of the factor that increases the death count. The objective of this review is to scale the performances of different algorithms that are so for used in the analyses of factor influencing to the road injuries. The review has found that K-Nearest Neighbor (KNN) and Decision tree are the two algorithms acts effectively with different factors that produce the reliable results so far. The further enhancement of the road calamities research would be carried out with the two classification algorithms such as Decision tree and K-Nearest Neighbor (KNN) with different attribute and distance measures.

References

[1] W. H. Organization, "Global Status Report on Road Safety 2018," 2018.

[2] W. Bank, "The High Toll of Traffic Injuires: Unacceptable and Preventable," 2017.

[3] R. P, "14 useful applications of data mining." [Online]. Available: https://bigdata-madesimple.com/14-useful-applications-of-data-mining/.

[4] R. K. R and R. B, "Data Analysis in Road Accidents Using," *Int. J. Civ. Eng. Technol.*, vol. 9, no. 4, pp. 214–221, 2018.

[5] A. Gupta, S. Gupta, and D. Singh, "A Systematic Review of Classification Techniques and Implementation of ID3 Decision Tree Algorithm," pp. 144–152, 2015.

[6] S. E. Madnick, Y. W. Lee, and R. Y. Wang, "Data and Information Quality Research: Its Evolution and Future 16.1," *semanticsholar.org*, p. 16.1-16.15, 2014.

[7] S. Liao, P. Chu, and P. Hsiao, "Expert Systems with Applications Data mining techniques and applications – A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, 2012.

[8] P.-N. Tan, M. SteinBach, and V. Kumar, *Introduction to Data Mining*. Pearson, 2014.

[9] S. Weiss and N. Indurkhya, *Predictive Data Mining A Practical Guide*, First Edit. The Morgan Kaufmann.

[10] L. Martín, L. Baena, L. Garach, G. López, and J. De Oña, "Using data mining techniques to road safety improvement in Spanish roads .," *Procedia - Soc. Behav. Sci.*, vol. 160, no. Cit, pp. 607–614, 2014.

[11] S. Mukherjee, R. Shaw, N. Haldar, and S. Changdar, "A Survey of Data Mining Applications and Techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 5, pp. 4663–4666, 2015.

[12] M. Singh and A. Kaur, "A Review on Road Accident in Traffic System using Various Data Mining Techniques," *Int. J. Sci. Res.*, vol. 100, no. 3, pp. 17–22, 2013.

[13] O. V.A and E. A.A, "Traffic Accident Analysis Using Decision Trees and Neural Networks," *Int. J. Inf. Technol. Comput. Sci.*, vol. 6, no. 2, pp. 22–28, Jan. 2014.

[14] A. J. Amutha and R. Padmajavalli, "A review on the performance of classification and prediction algorithms on cardiology data for the prediction of treadmill test through a mobile application," *Int. J. Appl. Pattern Recognit.*, vol. 5, no. 4, p. 293, 2018.

[15] S. Shanthi and D. R. G. Ramani, "Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms," *Int. J. Comput. Appl.*, vol. 35, no. 12, pp. 30–37, 2011.

[16] S. Alkheder, M. Taamneh, and S. Taamneh, "Severity Prediction of Traffic Accident Using an Artificial Neural Network," *J. Forecast.*, vol. 36, no. 1, pp. 100–108, 2017.

[17] L. J. Muhammad *et al.*, "Using Decision Tree Data Mining Algorithm to Predict Causes of Road Traffic Accidents , its Prone Locations and Time along Kano – Wudil Highway," vol. 10, no. 1, pp. 197–206, 2017.

[18] S. Krishnaveni and M. Hemalatha, "A Perspective Analysis of Traffic Accident using Data Mining Techniques," *Int. J. Comput. Appl.*, vol. 23, no. 7, pp. 40–48, 2011.

[19] B. Lavanya and B. Divya, "Predictive Analytics on Accident Data Using Rule Based and Discriminative Classifiers," *Adv. Comput. Serv. Technol.*, vol. 10, no. 3, pp. 461–469, 2017.

[20] B. N. Patel, S. G.Prajapati, and D. kamaljit I. Lakhtaria, "Efficient Classification of Data Using Decision Tree," *Bonfring Int. J. Data Min.*, vol. 2, no. 1, pp. 06-12, 2012.

[21] P. Gulati, A. Sharma, and M. Gupta, "Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review," *Int. J. Comput. Appl.*, vol. 141, no. 14, pp. 19–25, 2016.

[22] B. Gupta, P. Uttarakhand, and I. A. Rawat, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining," *Int. J. Comput. Appl.*, vol. 163, no. 8, pp. 975–8887, 2017.

[23] Q. Dai, C. Zhang, H. Wu, and S. Vocational, "Research of Decision Tree Classification Algorithm in Data Mining," *Int. J. Database Theory Appl.*, vol. 9, no. 5, pp. 1–8, 2016.

[24] M. B. R. Patel and M. K. K. Rana, "A Survey on Decision Tree Algorithm For Classification," *Ijedr*, vol. 2, no. 1, pp. 1–5, 2014.

[25] B. N. Lakshmi, T. . Dr.Indumathi, and D. N. Ravi, "An Empherical Study on Decision Tree Classification Algorithms," *Int. J. Sci. Eng. Technol. Res.*, vol. 4, no. 11, pp. 3705–3709, 2015.

[26] L. Zhang, Q. Liu, W. Yang, N. Wei, and D. Dong, "An Improved K - nearest Neighbor Model for Short-term Traffic Flow Prediction," *Procedia - Soc. Behav. Sci.*, vol. 96, no. Cictp, pp. 653–662, 2013.

[27] P. Mulak and N. Talhar, "Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset," *Int. J. Sci. Res.*, vol. 4, no. 7, pp. 2101–2104, 2015.