International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 5, October 2018, pp. 4033~4041
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i5.pp4033-4041

4Data Mining Approach of Accident Occurrences Identification with Effective Methodology and Implementation

Meenu Gupta[1], Vijender Kumar Solanki[2], Vijay Kumar Singh[3], Vicente García-Díaz[4]
[1,3]Ansal University, Haryana, India
[2]CMRIT, India
[4]Department of Computer Science, University of Oviedo, Oviedo, Spain

Article Info

Article history:

Received Feb 20, 2018
Revised May 28, 2018
Accepted Jul 10, 2018

Keyword:

Accidents
Bag of words
Classifier
CNB
Data mining
Frequency calculation
Hyperline
Machine learning
SVN
WEKA
Word count

Corresponding Author:

Vicente García-Díaz,
Department of Computer Science,
University of Oviedo,
Oviedo, Spain.
Email: garciavicente@uniovi.es

ABSTRACT

Data mining is used in various domains of research to identify a new cause for tan effect in the society over the globe. This article includes the same reason for using the data mining to identify the Accident Occurrences in different regions and to identify the most valid reason for happening accidents over the globe. Data Mining and Advanced Machine Learning algorithms are used in this research approach and this article discusses about hyperline, classifications, pre-processing of the data, training the machine with the sample datasets which are collected from different regions in which we have structural and semi-structural data. We will dive into deep of machine learning and data mining classification algorithms to find or predict something novel about the accident occurrences over the globe. We majorly concentrate on two classification algorithms to minify the research and task and they are very basic and important classification algorithms. SVM (Support vector machine), CNB Classifier. This discussion will be quite interesting with WEKA tool for CNB classifier, Bag of Words Identification, Word Count and Frequency Calculation.

1. INTRODUCTION

Data mining is the prominent technology to predict or do some analytics on a domain. Traffic management and accident occurrences in different places over the globe and the reason for accident s may vary. But we need to look after some of the things which are related to the mining the most chances of accident occurrences. Let's take a survey on different machine learning classification algorithms which are used on different data sets collected from different region and we can make a decision on which classification rule or association rule have to use for our data set. We have a publicly available data set on which we implemented SVM classifier and CNB classifier with WEKA tool. The required result is to identify which classification algorithm is better for the mining the actual data and predict better with the results. The main motto behind this kind of article is because of more cases being recorded by the regional hospitals as accident cases. The injuries, damages for vehicles and so on can be considered as the main reasons. The main reasons for the deaths on road is traffic accidents [1], that is not following the traffic rules, over taking in a wrong way, over speed, not following safety measures of road. As per WHO (World Health Organization) over 4 million cases have been recorded each year worldwide because of the traffic and road accidents. The main reasons which WHO states is not following traffic rules, not following safety measures like seat belt, hell mate, over speed, wrong crossing, minor driving, lake of literacy on the traffic and road safety rules, drunk

and drive. We can provide the measures to avoid this kind of things with small measures which are discussed by other researchers [2]. Data mining is mainly used to identify the severity of accidents on roads [3].

DMDW (Data Mining Data Warehousing) [4] have all the techniques to be used to predict or identify the severity of accidents on roads. DM is used to extract the semantic things over the data set that is a meaningful extract from the data available [5]. The classification techniques like clustering, anomaly detection, clustering and classification rules [6] are used for most of the DM operations on the road accidents. In this article we would like to share some literature survey on different previous operations done on different data sets and also the current research we would to do on the different data set related to the road accidents and severity. The next section will discuss short literature survey, later current work what this article will speak, experimental results, resources and finally conclude.

2. LITERATURE SURVEY

As we need to consider basics of Support vector machines and CNB classifiers to understand the literature review, let's make a sample collection of knowledge on SVM as it is important in this research scope. In machine learning, SVMs are controlled learning models with related learning counts that separate data used for course of action and backslide examination. Given a course of action of preparing cases, each set apart as having a place with both of two groupings, a SVM arranging check setting up a format). A Support Vector Machine points a delineation of the method as indicates in a plot, pointed or connected with the target that the examples of the instance of classes are disengaged by a sensible manner that is as wide as it could be sensible. New instances are then indentified and connected into that same hypothesis and anticipated to have a place with a class in context of which side of the instance they fall. Not with standing playing out the prompt demand, Support Vector Machines can beneficially act beyond the boundary as a non-straight depiction using the thing what is actually identified as the part-trap, checking and connecting their duties regarding high-instance portion spaces. Right when the data isn't stamped, straight forward things related to learning isn't acceptable, and an un-supervised learning methodology is mandatory, which is leading to identify trademark gathering of the information to get-togethers, and after that guide relevant data to these surrounded social groups. The grouping identifies which leads to a chance of modification to the SVM's is called support vector assembling and it is once in a while used as a bit of mechanical methodology either when the data isn't checked or when just two or three data are named as a pre-processing for a depiction method.

Asking for data is a general undertaking in ML. Expect some shown data shows every point as a place either of the available classes and the purpose is to pick exact class alternative Data point will be using. By ideals of SVM's, a data point is identified as a p dimensional vector (a quick overview of p identifiers), and the thing we have to identify is that possible that we can isolate such pointers with a (p-1)- multi-dimensional hyper plane. This can be identified as directed classifier. There are different hyper lines that may total data regarding the points. The one sensitive opinion as the better hyper-plane is the one that tends to the best partition, or point, between the different classes. So we select the hyper-line so the isolation from it to the closest data-point on other side is improved. In such data-point that hyper-line identifies, it is known as the best fitted hyper-line and the quick identifier it portrays is mentioned as a most over the top data classifier; or proportionately, the perceptron of flawless security

All the more generally, a SVM develops a hyper-line or set of hyper-lines in a high-or tremendous dimensional plane, which was used for depiction, fall away from the faith, or various undertakings like irregularities affirmation. Regularly, a mind blowing package is refined by the hyper-line that has the best division to the closest preparing information purpose behind any class (attested accommodating edge), since all around the more prominent the edge the lower the hypothesis spoil of the classifier

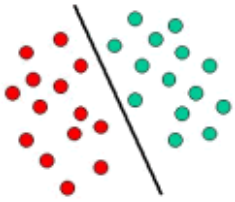

Figure 1. Support Vector Machine Sample plotting

The Figure 1 is a model occurrence of a SVM classifier, i.e., a SVM classifier that limits a strategy of things into their diverse social events (GREEN, RED) indicates a hyper-line. Most assembling undertakings, regardless, are not that crucial, and reliably more fanciful structure is required recollecting the genuine goal to make a flawless separation, i.e., decisively portray new difficulties (test instances) in light of the depictions that are operates (prepare instances). This situation is depicted in the structure below. Emerged from the previous semantic, unmistakably a complete section of the colors Green and also Red indication could be require a wind (which is more puzzling than a hyper-line). The Course of activity assignments in light of attracting hyper lines to see methods of different objects participating are defined as hyper-line classifiers as shown in Figure 2. Support Vector Machines are especially suited to oversee that kind of errands.

The Figure 3 below displays the critical thought behind SVM's. Here we can observe the basic differences (red part of the semantic) connected, i.e., adjusted, using a game-plan of sensible cutoff points specified as sections. The process of modifying the articles is defined as connecting. Make a note that in this new operations, the mapped objects (Green part of the semantic) is straightly unmistakable and, in like manner, instead of building the confusing turn (left semantic), we should just to locate an impeccable line that can disengage the Green and also the Red things.

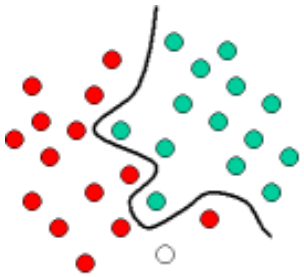

Figure 2. Differentiation between plots

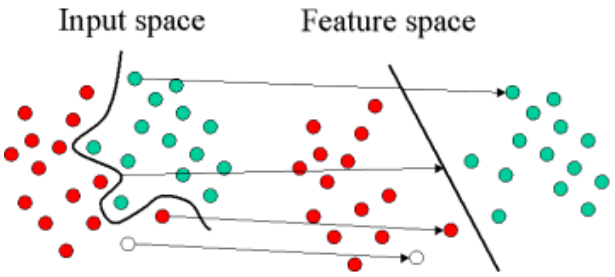

Figure 3. Input and output space differentiation

SVM is perhaps a champion among the most well known and talked about machine learning estimations. They were incredibly standard around the time they were delivered in the 1990s and continue being the go-to system for a high-performing count with a little tuning. In this post, you will discover the SVM machine learning figuring. In the wake of examining this post you will know:

Well ordered guidelines to disentangle the various names used to insinuate help vector machines. The depiction used by SVM when the model is truly secured to the plate. How an informed SVM demonstrate depiction can be used to make desires for new data. Well ordered directions to take in an SVM show from getting ready data. Guidelines to best set up your data for the SVM estimation. Where you may like to get more information on SVM. SVM is a stimulating estimation and the thoughts are by and large direct. This post was created for architects with basically no establishment in estimations and a straight factor based math.

The Maximal-Margin Classifier is a theoretical classifier that best clears up how SVM works eventually. The numeric data factors (x) in your data (the sections) outline an n-dimensional space. For example, if you had two information factors, this would shape a two-dimensional space. A hyperplane is a line that parts the data variable space. In SVM, a hyperplane is bested isolate the concentrations in the information variable space by their class, either class 0 or class 1. In two-estimations, you can picture this as a line and we ought to expect that the larger part of our data centers can be completely segregated by this line. For example:

$$B0 + (B1 * X1) + (B2 * X2) = 0$$

Where the coefficients (B1 and B2) that choose the inclination of the line and the catch (B0) are found by the learning computation, and X1 and X2 are the two data factors. You can take courses of action using this line. By interfacing with entering regards into the line condition, you can process whether another point is above or underneath the line. Over the line, the condition reestablishes a regard more noticeable than 0 and the point has a place with the five star (class 0). Underneath the line, the condition reestablishes a regard under 0 and the point has a place with the beneath normal (class 1). A regard close to the line reestablishes a regard almost zero and the point may be difficult to mastermind. If the span of the regard is generous, the model may have more trust in the desire. The division between the line and the closest data

shows is implied as the edge. The best or perfect line that can separate the two classes is the line that as the greatest edge. This is known as the Maximal-Margin hyperplane. The edge is figured as the contrary detachment from the line to only the closest core interests. Simply these concentrations are pertinent in portraying the line and in the improvement of the classifier. These concentrations are known as the assistance vectors. They support or describe the hyperplane. The hyperplane is picked up from planning data using a streamlining framework that lifts the edge.

When all is said in done, authentic data is disorganized and can't be separated impeccably with a hyperplane. The basic of growing the edge of the line that segregates the classes must be easygoing. This is routinely called the fragile edge classifier. This change allows a couple of demonstrates in the arrangement data manhandle the secluding line. An additional game plan of coefficients are exhibited that give the edge squirm room in every estimation. These coefficients are rarely called slack variables. This grows the multifaceted idea of the model as there are more parameters for the model to fit to the data to give this capriciousness. A tuning parameter is displayed called basically C that portrays the span of the squirm allowed over all estimations. The C parameters describes the measure of encroachment of the edge allowed. A C=0 is no encroachment and we are back to the unbendable Maximal-Margin Classifier depicted already. The greater the estimation of C the greater encroachment of the hyperplane are permitted. The taking of the hyperplane from data, all readiness cases that exist in the division of the edge will impact the circumstance of the hyperplane and are suggested as help vectors. Likewise, as C impacts the amount of events that are allowed to fall inside the edge, C impacts the amount of assistance vectors used by the model.

In this short literature survey we would like to discuss about different approaches worked out by different researchers over the globe. Machine Learning is the base concept behind the mining the severity of accidents. As we discussed previous over 4 million cases are being recorded as road accidents every year. Some of the machine learning algorithms like clustering is used as unsupervised learning technique. We need to consider clusters for a specific function in the data set. The function may be a reason of getting accident. For example over speed might be one reason so will be considering that as one of the function.

ANN (Artificial Neural Networks) [7] will be helping for analyzing the road accidents with different parameters. Tree based analyzing is one other concept [8], if we consider LCC (Latent Class Clustering) it is faster and accurate than k-NN with some functions of the data set. [9]-[13]. let's take a shore review on the data mining techniques which are being used in different domains of research over the globe by different researchers. The reason to know about the other research domains regarding the data mining techniques is to know the main functionality of each and every thing. There are few fundamental operations in the data mining and one among those is to split the data set into different clusters for the better clustering operations. Clustering is unsupervised learning in which we have no specific predicted output based on the available data and past data available we need to perform the operations and obtain the prediction results [14], [15]. If we consider the clustering we need to split the data set to identify the common and same category of the functions in the data set. Suppose if we are considering the accident severity in our case there may be different functions to be considered and some cases we need to consider the combination of the functions from the dataset. Lets take an example regarding the clustering the dataset. Consider the sample Table 2 below which is having some common things in the dataset.

By considering the Table 1 we can say that most of the accidents are happening to the car riders, reasons may be over speed, drunk and drive etc. We need to form the clusters based on the most weight reason for the accident.

Table 1. Sample Data from Dataset to implement sample clustering

State	Vehicle Types	Estimated Accident Reason	Estimated count
AP	Cars	Over speed, drunk and drive	150
UP	Cars, bikes	Over Speed, lack of safety measures	120,50
MH	Bikes	Lack of safety Measures	200
Kerala	Cars, Bikes, bus	Over speed, Drunk and drive, Lack of safety measures	50,25,15
Karnataka	Cars	Over Speed, Violating traffic rules, Lack of safety measures	150
TN	Bus, Car, Lorry, Walkers	Using phones on road, Over speed, road issues, drunk and drive	15,120,200,50
TS	Cars, Bus	Over Speed, Road Safety	150,200

3. PROPOSED APPROACH

We have seen some of the classification algorithms [16]-[19] and rules which are based on latest machine learning techniques. Clustering is based on unsupervised learning, K-NN, K-Means [20] is also under unsupervised learning technology. Let us take a time and execute the same data sets which are available in supervised learning. SVM (Support Vector Machines), CNB Classifier are the two classification

algorithms which we are explaining in this article. Based on the three categories we would like to explain our work in accident severity. BOW (Bag of Words), word frequency and word raking. BOW is consisting of the set of pre-defined words which are mostly used to explain the research component in the application. Support if we are having data set with some words like hell mate, seatbelt, speed etc those things will be considered as bag of words. First we need to perform the pre-processing of the data set. We need to identify the missing values in the data set and we need to substitute the missing values with the related values, whether it may be considering the mean or median of the values of that function or object. Lets take a look of the sample table which will consisting of the sample data which might be available with the data set.

This sample data set from Table 2 will be used for pre processing in machine learning technique may be using python or R programming. In this process we need to eliminate or handle the missing values. While handling the missing values we need to identify the text values and need to convert those to numerical format to apply prediction or data mining classification algorithm. Algorithms we are using can't be able to handle the string format in the data set always. There is a sequence to follow to predict the accuracy or to predict the main reason behind these accidents. Lets take a clear look on the flow with Figure 4.

Table 2. Sample Data set with some missing values

State	Number of accidents	Dead Cases	Injured Cases	Reason	Identifications
Andhra Pradesh	150	25	125	Lake of hell mate, over speed, wrong cut	Vehicle damaged severely, wrong cut
Rajasthan	100	50	50	Seat belt, over speed	Wrong cut
Maharastra	100	25			Vehicle damaged

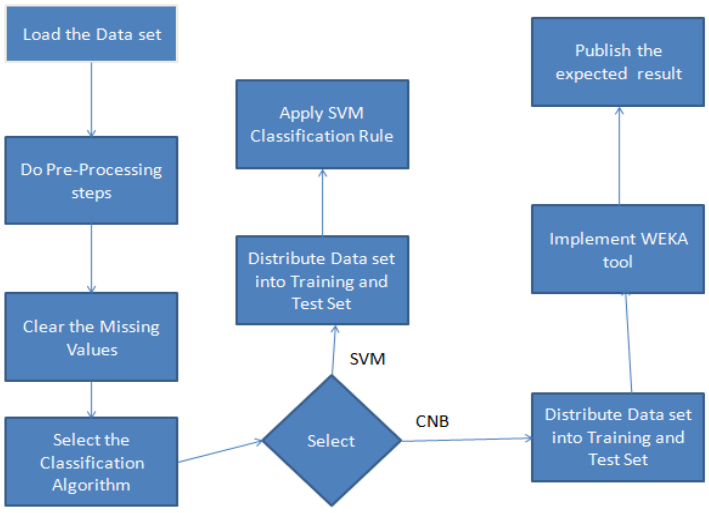

Figure 4. Structure of the mining the data set

First we need to load the data set which we need to process. Later do some pre-processing steps like eliminating the missing values and substituting those with the valid information like mean of the data of median. Then select the classification algorithm with which we need to apply. The missing values cleaned data set must be separated as training and test data set. The training dataset will be used for train the machine or classification algorithm which we are writing; test data set is used to correlate the things with the required result. We need to test the values of the data set with the training set and have to correlate with the previous work or with the training data set [21]-[23].

After selecting the classification algorithm, if we select the SVM algorithm, we need to select how main columns or rows we need to use for the test set to correlate, then submit the values. The result will be in three types. It will do BOW collection, word count and word frequency. Based on the word frequency we can estimate that which is the main reason behind the sever road accidents. The same follows with CNB classifier, but the thing will change here is we need to give sample count of columns and rows to process, it will take entire dataset without missing values and imply WEKA tool on it and produce the estimated result.

In the later part of the section we will discuss the experimental results with related to the sample data set we are using for the processing of the data. To be precise there are three types of results we acquire and we have already discussed the types of results we are going to get with this experiment.

As we discussed the proposed approach to identify the accident severity using two classification algorithms it worth to know about the whether these two will completely satisfy our requirement or anything need to be included. Coming to pros of these two approaches is we need not include every function into the algorithm or the model which we are using. The entire thing we need is limited model data or functions to be implemented in the algorithm. These two will give quick results than other algorithms. As these two are oldest algorithms and classification models the expected results may be vary as we predicted. As we use limited number of functions we cannot get the complete analysis of the predicted things required.

The better way to solve the problem regarding the accidents severity we can make use of the clustering algorithms, K-Means, ANN etc. So that we can get the apt results we required predicted results.

4. EXPERIMENTAL RESULTS

The results we acquire here have three types and the first thing is bag of words collection (BOW). Based on the number of values we assigned we can calculate the accuracy of the algorithm. Figure 5 Describes the graph of predicted results which describes the main reason for the accidents in those areas. Accuracy is based on the time taken and the number of rows or columns processed with the given classification algorithm using Data Mining or Machine Learning [24]-[26].

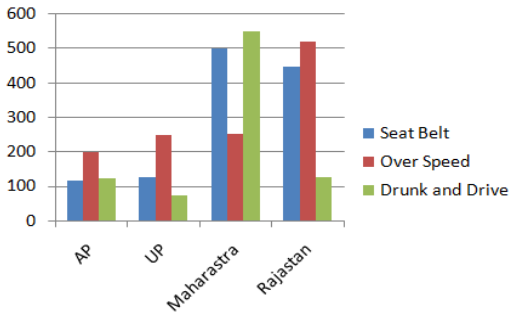

Figure 5. Graph of predicted result

By this graph we can predict the main reason for the severity of accidents in different locations. Classification problems are more related to the Machine Learning technique with which we need to train the machine with an algorithm [27]. Using ML the result we got here is classified into some of the functions. Let the Function be Reason type behind the accident. Let the City 1 may have 200 cases and out of that 100 are drunk and drive remaining are over speed, and for city 2 the total cases may be 300 and drunk and drive cases are 150 and remaining are over speed, no traffic rules are followed etc. [28], [29]. Therefore we can get the result that drunk and drive is the major function which is common in all the aspects.

We need to use Decision Trees [29], ANN from the machine learning community [30] for better prediction models for the domain of research. ANN here may be used to predict the future cause of the accidents and to identify the ratio of happening of the accident to the specific reason. That means we need to predict the reason which may cause and effect in future and how much ratio the cause may take part in the happened effect like accident in a specific region.

In This research we are planning to implement some of the advanced algorithms like ANN, Decision trees, Regression algorithms like SVR (Support Vector Regression) to design better prediction algorithm with the available data sets. We collected the public data set available from the government research web site which will give the brief information about the different reasons behind the accidents and how many number of cases are recorded region wise in the span of years .The reasons will be clear with a picture that the main reason may be not following the traffic rules and over speed are the main reasons for the accidents severity in every region. The following image Figure 6 will explain the sample about the coefficient and standard deviation levels in our algorithm related to the domain of research.

For better understanding of the decision trees and decision algorithms, and data mining techniques we can take any health care example like cancer [31]. We apply some of the data mining knowledge on that to predict the cancer percentage and the functional life time of that patient and the severity of the disease.

[32]-[34]. Data mining and Machine Learning are the two areas which are used for the further research of the domains like predicting the accident prone areas and types of reasons based on the locality in the future. The future of data mining is machine learning.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.0575279 -0.0163589 -0.0008483  0.0168662  0.0718922

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1570203  0.2324673   0.675   0.5058
PRICE       -0.1636906  0.7438870  -0.220   0.8277
INC          0.0012301  0.0012133   1.014   0.3208
TEMP         0.0028231  0.0004171   6.769 5.31e-07 ***
PRICEINCi   -0.2786003  0.1344397  -2.072   0.0491 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03094 on 24 degrees of freedom
Multiple R-squared: 0.7411, Adjusted R-squared: 0.698
F-statistic: 17.18 on 4 and 24 DF,  p-value: 8.968e-07
```

Figure 6. Coefficients and the Standard Error explanation

Figure 7 explains the count of accidents totally in one location. Let it be one city or state. So that these are the total number of accidents done in one month and we can make a conclusion that because of Lorries more accidents are happening. Whether it may be because of the over speed or drunk and drive. We can see the combination of those in Figure 5. In Figure 5 we will get the combination of the reason of accidents in one state for one month.

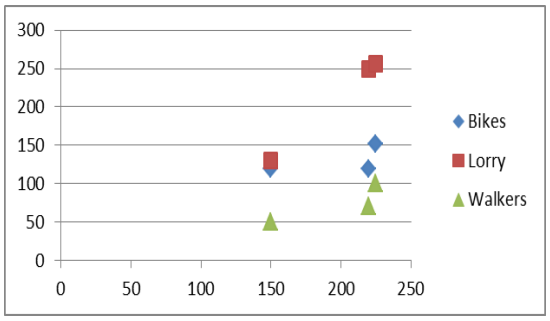

Figure 7. Plotting accidents severity

Based on Figure 8 the major reason of accidents in one state in one month is Drunk and Drive and Not Following the Traffic Rules. Like this we can consider few many conditions based on the requirement of the prediction model and its architecture

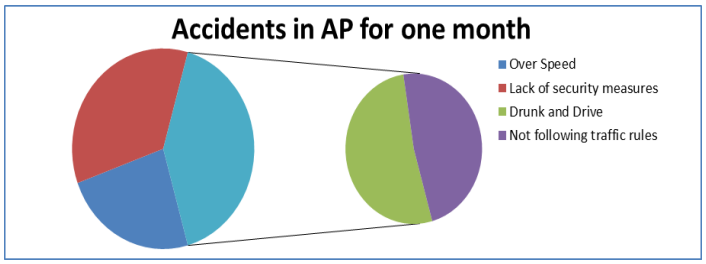

Figure 8. Predicting majority of the reason for accidents

5. CONCLUSION

The data mining and machine learning are the things we need to be considered to identify any unprocessed thing using datasets. In this article we tried to implement SVM and CNB classifiers with which we are predicting the main reason for the severity of accidents and we also predict the main reason on overall results. For example we can consider each state in india and we can predict both the things like main reason for the accidents in individual state and also main reason in overall country. For some cases SVM is showing more accuracy of 97% and some cases CNB is showing accuracy of 98%. With the obtained results both the algorithms are working well with all the conditions considered.

REFERENCES

[1] S. Kumar and D. Toshniwal, "A novel framework to analyze road accident time series data," *Journal of Big Data*, vol/issue: 3(8), pp. 1-11, 2016.

[2] M. Karlaftis and A. Tarko, "Heterogeneity considerations in accident modeling," *Accid. Anal. Prev*., vol. 30, no. 4, pp. 425-433, 1998.

[3] S. Kumar and D. Toshniwal, "Analysis of Hourly road Accident Counts using Hierarchical Clustering and Cophenetic Correlation Coefficient (cpcc)", *Journal of Big Data*, vol. 3, no. 13, pp. 1-11, 2016.

[4] P. N. Tan, *et al.*, "Introduction to Data Mining", Boston, Pearson Addison-Wesley, p. 769, 2006.

[5] S. Kumar and D. Toshniwal, "Analysing road Accident Data using Association rule Mining", *International Conference on Computing Communication and Security (ICCCS-2015), Kanyakumari, India*, 2015.

[6] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", United States, Morgan Kaufmann Publishers, 2001.

[7] L. Mussone, *et al*., "An Analysis of urban Collisions using an Artificial Intelligence Model", *Accident Analysis and Prevention*, vol. 31, pp. 705-718, 1999.

[8] L. Chang and W. Chen, "Data Mining of Tree based Models to Analyze Freeway Accident Frequency", *Journal of Safety Research*, vol. 36, pp. 365- 375, 2005.

[9] J. D. Oña, *et al.*, "Analysis of Traffic Accidents on Rural Highways using Latent Class Clustering and Bayesian Networks", *Accid Anal Prev*, vol. 51, pp. 1-10, 2013.

[10] S. Kumar and D. Toshniwal, "A Data Mining Framework to analyze road Accident Data", *Journal of Big Data*, vol. 2, no. 1, pp. 1-18, 2015.

[11] V. K. Solanki and V. K. Singh, "A Novel Framework to Use Association Rule Mining for Classification of Traffic Accident Severity".

[12] M. Gupta, "Analysis of Datamining Technique for Traffic Accident Severity Problem: A Review".

[13] M. Gupta, "Performance Evaluation of Classification Algorithms on Different Data Sets".

[14] Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining".

[15] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values".

[16] N. Dogan and Z. Tanrikulu, "A Comparative Analysis of Classification Algorithms in Data Mining for Accuracy, Speed and Robustness".

[17] Maimon O. and Rokach L., "The Data Mining and Knowledge Discovery Handbook", Springer, Berlin, 2010.

[18] Han J. and Kamber M., "Data Mining Concepts and Techniques", 2nd edn. Morgan Kaufmann, USA, 2006.

[19] Dunham M. H., "Data Mining: Introductory and Advanced Topics", Prentice Hall, New Jersey, 2002.

[20] T. N. Phyu, "Survey of Classification Techniques in Data Mining".

[21] Putten P., *et al.*, "Profiling Novel Classification Algorithms: Artificial Immune System", *Proceedings of the 7thIEEE International Conference on Cybernetic Intelligent Systems (CIS 2008), London, UK*, pp. 1-6, 2008.

[22] Hergert F., *et al.*, "Improving Model Selection by Dynamic Regularization Methods", in Petsche T., *et al*., "Computational learning theory and natural learning systems: selecting good models," MIT Press, Cambridge, pp. 323-343, 1995.

[23] Kaelbling L. P., "Associative methods in reinforcement learning: an emprical study," in Hanson S. J., *et al.*, Computational Learning Theory and Natural Learning Systems: Intersection between Theory and Experiment, MIT Press, Cambridge, pp. 133-153, 1994.

[24] Ge E., *et al.*, "Data Mining for Lifetime Prediction of Metallic Components", *Proceedings of the 5th Australasian Data Mining Conference (AusDM2006), Sydney, Australia*, pp. 75-81, 2006.

[25] Chiarini T. M., *et al.*, "Identifying fall-related Injuries: Text Mining the Electronic Medical Record", *Inf Technol Manage*, vol. 10, no. 4, pp. 253-265, 2009.

[26] Breiman L., *et al.*, "Classification and Regression tree", Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, 1984.

[27] R. Agrawal, *et al.*, "Database Mining: A Performance Perspective", *IEEE Trans. Knowledge and Data Engineering*, vol. 5, no. 6, pp. 914-925, 1993.

[28] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann, 1993.

[29] Y. Bengio, *et al.*, "Introduction to the Special Issue on Neural Networks for Data Mining and Knowledge discovery," *IEEE Trans. Neural Networks*, vol. 11, pp. 545-549, 2000.

[30] D. Michie, *et al.*, "Machine Learning, Neural and Statistical Classification," Ellis Horwood Series in Artificial Intelligence, 1994.

[31] "Comparative Analysis of Classification Algorithms for the Prediction of Leukemia Cancer."

[32] S. Vijayarani, "Comparative Analysis of Bayes and Lazy Classification Algorithms."
[33] "A Novel Design Specification Distance (DSD) Based K-Mean Clustering Performace Evaluation on Engineering Materials' Database."
[34] "A Survey on Decision Tree Based Approaches in Data Mining."