International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Identifying Road Accidents Severity using Convolutional Neural Networks

L. Yasaswini[1*], G. Mahesh[2], R. Siva Shankar[3], L.V. Srinivas[4]

[1,2,3,4]*Department of CSE, S.R.K.R. Engineering College, Bhimavaram, Andhra Pradesh, India.*

[*]*Corresponding Author: lanka.yasaswini95@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 20/July/2018, Published: 31/July/2018

Abstract: The purpose of the proposed work is to identify the factors contributing to fatal accidents. This is achieved by analysing road accidents using Convolutional Neural Networks by considering appropriate features and effectively clustering the records. Several combinations of attributes of large datasets are analysed to discover hidden patterns that are the root cause for accidents. The chances of accident occurrence could be identified by considering various criteria like speed limit and injury severity, time of accidents and drunk driver, month and weather during the accident, lightness and speed limit, human factors, surface and light conditions. The experimental results on road accident data set FARS (Fatality Analysis Reporting System) generated risk factors that cause fatal accidents which will be helpful in generating safer driving principles.

Keywords: Association Rules, Classification, Convolutional Neural Networks, Traffic Data.

I. INTRODUCTION

Roadway traffic is the major issue these days. Increase in number of vehicles moving on roads accelerated the risk of accidents. Of them, fatal accidents is the major issue where people lose their lives. Also, these accidents are unpredictable that they may occur anywhere, anytime. As a human being we should save the lives of people and avoid these accidents. A secure roadway movement is a major concern for both transportation administering organizations and common nationals. Keeping these facts in mind, the aim of this work is to provide safe driving instructions to people moving on the roads and emergency services to people effected in the accident zone. So, factors like weather conditions, collision manner, surface condition, light condition, speed, drunk driver and so on were considered and examined. Analysed data can be used to give safer driving suggestions and reduce the accident rate. Also, emergency services can be provided to people affected at accident prone area.

Data mining is one of the important mechanisms used in Information Technology from previous times. Data mining techniques best works in processing data and identifying the relationship among data. Association rule mining is a method used for finding interesting patterns among variables from huge databases. To find association among data, support and confidence are calculated by placing a threshold value. Finding related data using association rules helps in frequent itemset mining.

Classification is performed on data using some classification model suitable to the given set of data. The purpose is to find out the frequent itemsets. During classification a model is constructed in which different records of data set with unspecified class labels are separated easily. Naive Bayesian classification is one of the probabilistic methods used to predicate the independence among variable pairs. It strongly assumes and auto correlates the information. Sometimes these assumptions may go wrong. Thus, a better classification technique proposed to efficiently classify the data is Convolution Neural Network. It assumes data based on the locality. The classification technique proposed can be applied on the data to get effective results.

The association rule mining algorithm ever used is Apriori. The algorithm efficiently works based on relevant association rules for frequent itemset mining. It uses a bottom up approach. The property followed by this algorithm says any subset of frequent itemset must be frequent. It uses larger itemsets and can be implemented easily. This algorithm is applied on roadway traffic fatal accident dataset to test the data.

Clustering is an unsupervised learning problem which is considered being important for data mining. The structure of data can be found from a collection of unlabelled data. K-means is a clustering algorithm which is absolute. The procedure follows classification of a given data set into certain number of clusters. K-means is both simple and adaptive algorithm for lot many problems.

FARS dataset is often called as Fatality Analysis Reporting System. It is used for traffic analysis, study of fatality rate and reasons for fatality. FARS, a nationwide census that provides fatal accident data to NHTSA (National Highway Traffic Safety Administration) and the American public. This data is provided by California Polytechnic State University. It has more than 37000 records and 55 attributes. Hence the dataset FARS is proposed to be used in this work.

II. RELATED WORK

Traffic accident dataset had been analysed by Jayasudha. Many data mining techniques, tools and applications have been used to control fatality rate. Functionalities like data characterization, data discrimination, association analysis, classification, prediction and cluster analysis were performed during this survey. Road safety databases like International Road Traffic and Accident Database (IRTAD), GLOBESAFE, ACCIDENT RESOURCE were used. By analysing this data, it is identified that accident rates are more at the intersections than in highways or main roads. Use of sensors with GPS technology helped to gather details of the incident. It was collected in to a data base and then classified using the decision support systems to identify the exact decision. The decision is given to the driver as an alert. This process helped to reduce accident conditions [1].

Initially people of Washington were allowed to increase the speed limit form 55mph to 65mph. The results were conflicting. The incidence of fatal crashes more than doubled. The fatality rate increased up to 27% which was 10% lower when compared to states that did not increase their speed limit. Later on, Eric and Peter have studied crash data from a single state to examine fatal crash incidence after speed limits were increased. Specific aims of this study were to estimate the effect of the speed limit on the incidence of fatal crashes. Also, conditions like total crashes on rural freeways, average vehicle speeds and speed variance were examined. Data was collected by the Washington State Department of Transportation (WADOT). The study concluded that fatal crashes occurred even when there is no increase in speed limit in Washington. Road conditions, roadway and vehicle design, age of driver, and alcohol use were also some factors for fatal crashes [2].

Providing emergency services to people affected in accidents may reduce the fatality rate as discovered by William M. Evanco. An express emergency system was introduced to provide basic medical services to the victims at accident prone area. Accident notification time and EMS notification were considered to save lives of victims. People with trauma disease are to be treated immediately to help them out of danger. First 10 minutes is so precious for them. The person who treated at 8.4th minute survives rather a person who has been treated at 12.3rd minute lost his life. A four-minute difference changed a lot. At rural areas an active and immediate EMS response affected people's lives. Also factors like vehicle kilometres travelled, alcohol consumption, driver age distribution, accident notification time, personal income per capita and so on affect the fatality rate [3].

Solaiman ET. Al. provided complete and accurate information source regarding accidents. A clear picture of accident rate and accident trend was incorporated. Software and a website were designed for identifying the data to be analysed and the information to be provided. In AVRA BANGLADESH a system was implemented that automatically generates different visualization patterns of accident data and corresponding factors in terms of various related variables. Various parameters like junction type, collision type, location, month, time of occurrence and vehicle type are visualized with in a particular time strap to find how these parameters change and behave with respect to time. Type of accidents could be effectively classified based on these attributes. A map API can be used along with the system to find the safest and dangerous roads [4].

A study is carried out on roadway accident information by Sachin Kumar and Durga Toshniwal. EMRI is the source that provides accident records of entire state. The dataset is processed during data pre-processing step. Both density and partition-based clustering algorithms are applied to categorize the data. During the cluster analysis, k-mode was used. 6 numbers of clusters formed after clustering the data. Association rules were applied on each cluster for finding the correlation among different attributes in the data. WEKA 3.6 tool was used to perform association rule mining. In a country like India, working with traffic and road accidents data is not as per mark. This data provides information that was not available with the previous data collection process by police officers. The analysis of data has shown that frequent accident occurrence is at places like markets, hospitals, local colonies. Also, two wheelers met with accidents at various intersections resulting in fatal accidents. It is also identified that the rate of accidents was more during 4 to 6 in the morning and 8 to 4 in the evening [5].

Krishnaveni and Hemalatha worked with large amount of data and several data mining approaches. Various clustering algorithms like Bayesian classifier, Meta classifier, Rule classifier, Decision Tree classifier, Tree classifier were used. Naive Bayes, AdaBoostM1, PART J48 and Random Forest are the clustering algorithms that belong to different clustering techniques. All these algorithms are applied on road accident dataset and concluded that random forest best performs the classification of data. A genetic algorithm for feature selection was also used for data selection process [6].

Amira ET. AL. analysed traffic and accident data of Dubai. The association rule mining algorithms apriori and predictive apriori were applied to analyse the data. From the analysis performed, reasons for accidents were identified. A comparison is carried out between the two algorithms apriori association rule mining and predictive apriori association rule mining. Of these two algorithms apriori best found the relationship among data variables and generated appropriate results [7].

Liling Li ET.AL. analysed FARS data set using data mining techniques. Apriori, Naive Bayes and K-means are the algorithms used for association rule mining, classification and clustering. The study aimed at providing safer driving principles to people based on the statistics [8].

Sami Ayramo and Pasi Pirtala performed research on data collected from Finnish Road Administration among 2004 and 2008. By applying several data mining methods like robust clustering, association rule mining, frequent item sets and visualisation, it is found that these methods can produce understandable patterns from raw data. Thus, detailed comprehensive datasets can be obtained [9].

A comparison among classification algorithms like C4.5, C-RT, CS-MC4, Decision List, ID3, Naive Bayes and RndTree have been performed by S. Shanthi and Dr.R. Geetha Ramani during 2011. The result specified that RndTree outperformed of all [10].

Also, some feature selection algorithms such as CFS, FCBF, Feature Ranking, MIFS and MODTree for improving classifier accuracy. Of these algorithms, feature ranking best performed. Later on, the study extended by using Meta classifier Arc-X4 to improve the classifier accuracy. Thus, the combination of RndTree with feature ranking and Arc-X4 Meta classifier came up with best results [11].

All the studies classified road accident data and found various factors that cause accidents. The aim is to propose an efficient method that best identifies these factors and avoid accidents.

III. DATA COLLECTION AND CONSTRUCTION

A huge set of attributes form the input dataset. They are of different types such as accident-specific attributes, driver-specific attributes, circumstance-specific attributes and so on. The work mainly carried based on this data.

Data Construction is also known as Data Preparation. Initially data will be cleaned by removing noise, missing values, and inconsistencies. Missing values are replaced by NULL values. Data will be selected and transformed based on the requirement. It may be either in numerical or nominal form. Data will be in nominal form in the dataset, it can be converted into numerical data while performing operations on it. Also, each attribute data is discretized in order to make it appropriate for further analysis.

Major attributes selected in this study are accident conditions like, manner of collision, light condition, weather condition, roadway surface condition, speed limits and drunk driver. Also, these attributes hold some values listed below.

Manner of Collision: Not Collision with Motor Vehicle, Front-to-Rear (Includes Rear-End), Front-to-Front (Includes Head-On), Angle-Front-to-Side (Same Direction), Angle-Front-to-Side (Opposite Direction), Angle-Front-to-Side (Right Angle (Includes Broadside)), Angle-Front-to-Side (Angle-Direction Not Specified), Sideswipe (Same Direction), Sideswipe (Opposite Direction), Rear-to-Side, Rear-to-Rear, Other, Unknown.

Light Condition: Daylight, Dark, Dark but lighted, Dawn, Dusk, Unknown.

Weather Condition: Blowing sand soil dirt, clear cloud no adverse condition, Fog smog smoke, Other, Rain mist, severe crosswinds, Snow or blowing snow, unknown.

Roadway Surface Condition: Dry, Ice frost, oil, Other, Sand dirt mud gravel, Snow or slush, unknown, Water, Wet.

Speed Limits: 5, 10, 15, 20, 25, 35, 45, 55, 60, 65, 70, 80, 85, 95, 99 (kmph).

Drunk Driver: It has two conditions either yes or no.

All these factors affect the rate of accident occurrence and were used to determine whether it is hazardous for the people moving on roads during these conditions.

IV. IMPLEMENTATION OF CONVOLUTIONAL NEURAL NETWORKS

A Convolutional Neural Network is a class of deep learning, feed-forward artificial neural networks, and most commonly useful for several analyses. They visualize metaphors and Numerical data. It can use a difference of multilayer perceptron designed to need minimal pre-processing. It is very similar to normal Neural Networks. They are made up of neurons that have learned weights and biases. Each neuron receives several inputs, performs a dot product and optionally follows it with a non-linearity. The complete network at rest articulates to achieve the function from the raw input data on one end to achieve the class at the other end. It can make a clear hypothesis that the inputs allow us to encode certain possessions into the CNN process and then, make the forward function more efficient to implement. They very much reduce the number of parameters in the network. Neural Networks consider an input and transform it through a series of hidden layers. Each hidden layer is made up of a set of neurons, where each neuron is fully connected to all neurons in the previous layer. Neurons in a single layer function in a completely separate manner and do not share any relations. The very last fully-connected layer is called the output layer and it represents the class achieved.

There are three main Layers that build Convolutional Neural Networks. They are Convolution Layer, Pooling Layer, and Fully-Connected Layer

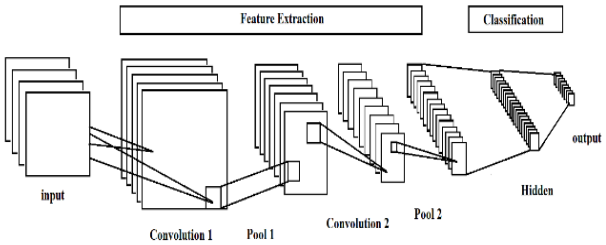

Fig.1 Flow of Work in CNN

The input [HxWxD] initially will hold the input values and in this case an input data of width (rows), height (columns), and depth 1 are considered.

CONV layer will compute the output of neurons that are connected to local regions in the input. It has independent set of filters that work with the input and a small region is connected to in the input volume. The output volume will be as [32x32x12] and 12 is the number of filters that were used. RELU layer applies an element wise activation function, such as max (0, x) where zero is threshold value. Here the volume size remains unchanged ([32x32x12]).

POOL layer performs down sampling operation throughout the spatial dimensions (width, height). It resultant volume is[16x16x12]. Its function is to reduce the spatial size of the representation progressively. Therefore, it reduces the number of parameters and computation in the network. Fully-connected layer will compute the class scores. It results in volume of size [1x1x10], where each of the 10 numbers corresponds to a class achieve. As with ordinary Neural Networks and as the name implies, each neuron in this layer will be connected to all the numbers in the earlier volume.

In this fashion, Convolutional Neural Networks transform the input raw data, layer by layer from the input raw values to the final class scores. Particularly, the Convolution and Fully Connected layers perform transformations that are a function of not only the activations in the input volume, but also of the parameters (the weights and biases of the neurons). On the other hand, the RELU layer and Pooling layer will implement a fixed function. The parameters in the Convolutional and Full-Connected layers will be qualified with gradient descent so that the class scores that the Convolution Neural Networks computes are consistent with the labels in the training set for each input.

To implement convolutional neural networks a training dataset is considered as input. The dataset contains numerical values that correspond to some nominal data. Data corresponds to specified number of rows and columns, according to the training dataset taken. Each column defines an attribute that points to several possible values. A row has values that define possible set of conditions for an accident to occur.

The basic steps of Convolutional Neural Networks on traffic data can be explained using the training dataset. Initially a training data set with numerical values that corresponds to several nominal values is considered. These numerical values may belong to different attributes taken. All the numerical values of a row are counted and stored. These values are sorted from minimum to maximum. According to pooling operation the maximum value or greater value than maximum is considered. This value defines the possibility of risk. If the obtained value is greater than or equal to maximum the risk factor is high for the calculated row of attributes, otherwise it can be considered as low.

Previous studies calculated the risk factor based on Naïve Bayes classification Technique. The proposed method, Convolutional Neural Network is a probabilistic approach that measured the results accurately. The risk factor could be effectively identified using this technique than previous methods.

V. EXPERIMENTAL RESULTS AND ANALYSIS

Table 1. Risk factor Identification using Convolution Neural Networks

collision type	light	Weather	surface	speed	drunk driver	Risk
angle-front-to-side	dark but lighted	Unknown	unknown	95	no	High
not collision with vehicle in transport	dawn	fog smog smoke	snow or slush	60	yes	High
not collision with vehicle in transport	daylight	fog smog smoke	wet	15	no	High
front-to-rear	dawn	clear cloud no adverse condition	sand dirt mud gravel	55	no	High
sideswipe-opposite direction	daylight	blowing sand soil dirt	oil	5	no	High
angle-	dark	Unknow	ice frost	10	no	High

front-to-side		n				
angle-front-to-side	dawn	Other	ice frost	45	no	High
not collision with vehicle in transport	dusk	Unknown	wet	5	yes	High
front-to-front	unknown	clear cloud no adverse condition	snow or slush	25	yes	High
angle-front-to-side	dark but lighted	Other	snow or slush	65	no	High
front-to-rear	dark but lighted	blowing sand soil dirt	oil	99	yes	High
angle-front-to-side	dawn	snow or blowing snow	dry	60	yes	High
front-to-rear	dawn	snow or blowing snow	oil	65	yes	High
sideswipe-opposite direction	dawn	Other	unknown	85	no	High
sideswipe-same direction	dawn	snow or blowing snow	water	5	no	high

The processing of training data set using Convolutional Neural Networks resulted in calculating the risk factor in an efficient manner. Various factors contributed for fatal accidents were identified using this classification technique in a most probabilistic approach. The results obtained were used to specify fatal conditions for an accident. Thus, safety measures can be provided to people moving on roads in such conditions.

During the process, various attributes like collision type, light conditions, weather conditions, surface conditions, speed, drunk driver or not were taken into consideration to find out the risk factor. The risk factor specified the possibilities of fatal accidents at different areas. The results obtained in calculating the risk factor using Convolutional Neural Networks can be shown below.

Efficiency of the classification algorithm, Convolutional Neural Networks can be known by calculating accuracy, precision, recall and f-measure on resulted data. Accuracy defines the trueness of occurred result. The factors precision and recall specify the occurrence of relevant instances over retrieved instances and total number relevant instances respectively. These three measures can be calculated based on following factors.

- **True positive (TP):** If the given combination matches with at least one record in base dataset along with result, that particular result corresponds to TP value.
- **True Negative (TN):** If the given combination matches with at least one record in base dataset but the rate doesn't match, it defines TN.
- **False positive (FP):** If the given combination doesn't match with any record in base dataset but the fatality rate is High, it comes under FP.
- **False negative (FN):** If neither the record matches nor the rate is High, its FN.

The above values TP, TN, FP and FN are compared and incremented on matching basis. Final counts of every case are jotted and following formulae calculates the values respectively.

Accuracy = (TP+FP)/(TP+TN+FP+FN)
Precision = TP/ (TP + FP)
Recall = TP/ (TP + FN)
F-Measure = (2*Precision*Recall)/(Precision+Recall)

Efficiency based Result for the classification techniques Naïve Bayes And Convolutional Neural Networks can be given in the below tabular for based on above calculations.

Table 2. Efficiency calculation for Naïve Bayes and CNN

Efficiency Measures	Naïve Bayes	CNN
Accuracy	**0.72**	**0.68**
Precision	**0.86**	**0.53**
Recall	**0.68**	**1**
f-measure	**0.76**	**0.69**

The results specified in above table shown that CNN performed very efficiently in calculating recall value. Recall value is the measure that specifies defective and unsafe conditions. The results of recall can provide safer guidelines. Hence the conditions obtained using CNN found the conditions that caused accidents. Graphical representation of above table for the classification techniques Naïve Bayes and Convolution Neural Networks can be represented as follows.

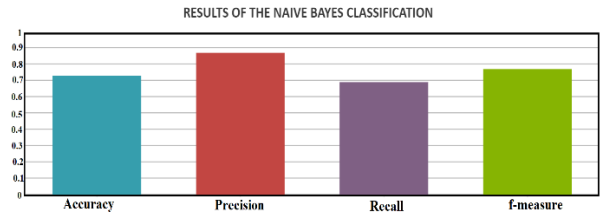

Fig.2 Performance based analysis of Naïve Bayes Classification

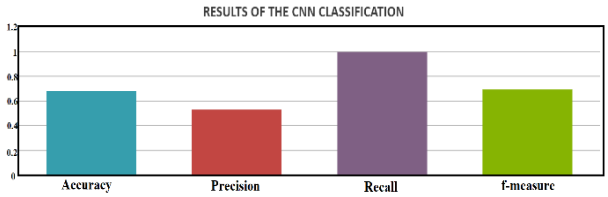

Fig.3 Performance based analysis of Convolutional neural networks

The graph indicated that CNN (convolutional neural networks) best performed in finding probabilistic results by calculating maximum number of instances that cause fatal accidents using given dataset.

VI. CONCLUSION AND FUTURE

In this work, a classification technique named Convolution Neural Networks has been used that effectively identified the conditions contributing to fatal accidents. Using these conditions, the public could identify dangerous zones and take measures to avoid accidents. Experimental results have shown that CNN is more efficient than Naïve Bayes classifier in identifying the risk factor. CNN out performed many other techniques used in previous works. In the future it could be planned to make analysis on road accident dataset by considering more features and more clusters and to use deep learning techniques.

REFERENCES

[1] K.Jayasudha and C.Chandrasekar, "*An overview of data mining in road traffic and accident analysis",* Journal of Computer Applications, Vol.2, Issue.4, pp:32–37, 2009.

[2] Eric M Ossiander and Peter Cummings, "*Freeway speed limits and traffic fatalities in Washington state",* Accident Analysis & Prevention, Vol.34, Issue.1, pp:13–18, 2002.

[3] William M Evanco, "*The potential impact of rural mayday systems on vehicular crash fatalities",* Accident Analysis & Prevention, Vol.31, Issue.5, pp:455–462, 1999.

[4] KMA Solaiman, Md Mustafizur Rahman and Nashid Shahriar, "*AVRA Bangladesh collection, analysis & visualization of road accident data in Bangladesh*", In Proceedings of International Conference on Informatics, Electronics & Vision, pp.1–6, IEEE, 2013.

[5] Sachin Kumar and Durga Toshniwal, *"Analysing road accident data using association rule mining",* In Proceedings of International Conference on Computing, Communication and Security, pp.1–6, 2015.

[6] S.Krishnaveni and M.Hemalatha, "*A perspective analysis of traffic accident using data mining techniques",* International Journal of Computer Applications, Vol.23, Issue.7, pp:40–48, 2011.

[7] Amira A El Tayeb, Vikas Pareek, and Abdelaziz Araar, "*Applying association rules mining algorithms for traffic accidents in Dubai",* International Journal of Soft Computing and Engineering, 2015.

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

[9] Sami Ayramo, Pasi Pirtala, Janne Kauttonen, Kashif Naveed and Tommi Karkkainen, "*Mining road traffic accidents",* Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering, University of Jyvaskyla, pp.1-53, 2009.

[10] S. Shanthi and R. Geetha Ramani, "*Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms",* International Journal of Computer Applications (0975–8887), Vol.35, Issue.12, pp:30-37, 2011.

[11] S. Shanthi, R. Geetha Ramani, "*Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques",* Proceedings of the World Congress on Engineering and Computer Science, Vol 1, 2012.

[12] Dharmendra Sharma and Suresh Jain, "*Evaluation of Stemming and Stop Word Techniques on Text Classification Problem",* International Journal of Scientific Research in computer Science and Engineering, Vol.3, Issue.2, 2015.

[13] Mohnish Patel, Aasif Hasan, Sushil Kumar, *"Preventing Discovering Association Rules For Large Data Base",* International Journal of Scientific Research in computer Science and Engineering Vol.1, Issue.3, 2013.

[14] Neeraj Chhabra, *"Comparative Analysis of Different Wireless Technologies",* International Journal of Scientific Research in Network Security and Communication, Vol.1, Issue.5, 2013.

[15] V. Kapoor, *"A New Cryptography Algorithm with an Integrated Scheme to Improve Data Security",* International Journal of Scientific Research in Network Security and Communication Vol.1, Issue.2, 2013.

Authors Profile

Ms L N S Yasaswini pursued Bachelor of Technology in Sri Vishnu Engineering College for Women in year 2016. She is currently Pursuing Master of Technology in Computer Science Department of SRKR Engineering College Affilated under andhra university.

Dr G Mahesh is presently working as Associate Professor in Department of CSE, SRKR Engineering College, Bhimavaram. He has a total teaching experience of 28years. He guided many M.Tech. students for their project work. He has 6 journal papers and 5 conference papers in his credit.

R Shiva Shankar is presently working as an assistant professor in Department of CSE in SRKR Engineering college. He is having an experience of 10 Years.

L V Srinivas is presently working as an assistant professor in Department of CSE in SRKR Engineering college. He is having an experience of 4 Years.

