ANALYZING ROAD ACCIDENT DATA USING CLASSIFICATION TECHNIQUE

Dr T.Jyothirmayi
Department. of CSE
GITAM UNIVERSITY
Visakhapatnam, INDIA

B.Soujanya
Department. of CSE
GITAM UNIVERSITY
Visakhapatnam, INDIA

G L Aruna Kumari
Department. of CSE
GITAM UNIVERSITY
Visakhapatnam, INDIA

***Abstract*—**
Road accidents are uncertain and unpredictable. Many factors includeing type of vehicle, road type, lighting on road, road feature, etc contribute to different types of road accidents. Here in this proposed work the road accident information which has been collected for a period of time in a particular area is observed and analyzes the cause of the accidents by different algorithms possible, and try to reduce the occurrences of the accidents in future. This model also helps traffic engineers to know the information related to accidents and places where they occurred frequently and help them to make necessary changes in the infrastructure. Here one of the classification techniques called Support Vector Machine (SVM) is used in order to classify the data. Support Vector Machine uses kernels, which can build in expert knowledge about the problem via engineering the kernels. The prediction of future instances is done with the help of classification model. This model with some further enhancements will be useful for the road engineers to focus and develop new approaches for reducing the road accidents.

Keywords—Support Vector Machine, classification, Road Traffic Accident

1. INTRODUCTION

In developing countries, the issue of road accidents is a major concern. Increasing road traffic/vehicle occupancy could be the reason behind this. There is an increase in accidents over the years. It is very important to regulate traffic on roads to reduce accidents in accident prone zones. A road traffic accident (RTA) is any injury due to crashes originating from, terminating with or involving a vehicle partially or fully on a public road. Road accidents are uncertain and unpredictable accidents. Many factors like type of vehicle, road type, lighting on road, road feature are responsible for road accidents. Moreover, road accidents are relatively higher in extreme weather and during working hours. In India about 1.3 lakh people died on roads, giving India the dubious honor of topping the global list of fatalities from road crashes. Rapid urbanization, motorization, lack of appropriate road engineering, poor awareness levels, non-existent injury prevention programs, and poor enforcement of traffic laws has exacerbated the situation. Analysis of road accident scenario at state and city level shows that there is a huge variation in fatality risk across states and cities. These risks vary according to different factors like climatic conditions etc. The severity of the accident also depends on the different types of factors. Considering the road safety is important to recognize the worsening situation in road deaths and injuries and to take appropriate action.

The proposed model collects road accident data and analyzes the cause of the accidents by different algorithms, and try to reduce the occurrences of the accidents in future. There exists different classification techniques like Decision Trees, Naïve Bayes, Generalized Linear Models, etc. to analyze road accident information. The proposed model makes use of Support Vector Machine(SVM) as it has advantage that it uses kernels, which can build in expert knowledge about the problem via engineering the kernel which helps in building best model. Support Vector Machines (SVM) is a powerful, state-of-the-art algorithm with strong theoretical foundations based on the Vapnik-Chervonenkis theory.

This model lets traffic engineers to make the necessary changes in the infrastructure of the roads and reduce the chances of occurrence of accidents.

2. LITERATURE REVIEW

Many researchers have carried out research work in the area of road accidents. Some of them have analyzed accident data in different ways. Some of them Identification of Black spot zone. Some of them have developed accident models for forecasting future accident trends. They have also proposed strategies for road safety. Data mining is a powerful tool that can help you find patterns and relationships within your data.

Classification is a learning function that maps a given data item into one of several predefined classes. There exists few techniques in developing scalable and robust classification techniques that are capable of handling large disk-resident data. Classifications are of many types namely Bayesian classification, Artificial Neural Networks, Rule-based mining, Support Vector Machine (SVM), etc. and has numerous applications including trajectory classification, fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis. The performance of the classification techniques is measured by the metrics like accuracy, speed, robustness, scalability, comprehensibility, time and interpretability.

As India is the largest country in the South Asian region with all the problems faced by rapidly developing nations, especially increasing motorization. In spite of such developments, there are limited data in the literature addressing the problem of road traffic injuries. Road traffic injuries are a significant burden on the health care system in India. The most commonly affected group is young males. Pedestrians constitute a large majority of the victims.

Sachin Kumar and Durga Toshniwal[20] in "A data mining framework to analyze road accident data" analyzed about the road accident data using a classification technique called Support Vector Machine(SVM) for classifying the data. In addition, they applied an Association Rule Mining on the result to generate the rules for future predictions. The best number of classes are obtained by cluster analysis.

Luis F. Miranda-Moreno et al (2005)[3] developed Alternative Risk Models for Ranking Locations for Safety Improvement. The authors has shown their study between performance and practical implications of these models and ranking criteria made comparison when they were used for finding dangerous locations. In their research the relative performance of three alternative models is investigated: the traditional binomial model, the heterogeneous negative binomial model and the Poisson lognormal model. The focus in their work is particularly on the impact of choice of two alternative prior distributions (i.e., gamma versus lognormal) and the effect of allowing variability in the dispersion parameter on the outcome of the analysis.

FajaruddinMustakinetal(2008)[4] used multiple regression linear models to study block spot study and accident prediction model. Federal Route (FT50) BatuPahat - Ayer Hitam was the study area. Following is the regression model

In(APW)0.5=0.0212(AP+0.0007(HTV0.75+GAP1.25)+0.0210(85th PS)

Where,

APW= accident point weight age

AP= number of access points per kilometer

HTV= hourly traffic volume

Gap= amount of time, between the end of one vehicle and the beginning of the next in second.

85th PS= 85th percentile speed

The model has R-square of 0.9987

As per the results it is seen that existence of a large major junction density, an increase in traffic volume and vehicle speed in federal Route 50 contributed to traffic accident. Influential effect on road traffic accident may be seen by reduced vehicle speed, access point, traffic volume and gap.

Cheng-Tao Chu et al (2007) adapted Google's map-reduce [8] paradigm to demonstrate this parallel speed up technique on a variety of learning algorithms including locally weighted linear regression (LWLR), k-means, logistic regression (LR), naive Bayes (NB), SVM, ICA, PCA, gaussian discriminant analysis (GDA), EM, and backpropagation (NN). Their experimental results shown basically linear speedup with an increasing number of processors.

R.R. Dinu, A. Veeraragavan (2011)[18] presented Random Parameter Models for Accident Prediction on Two-Lane Undivided Highways in India.Based on three years of accident history, from nearly 200 km of highway segments, is used to calibrate and validate the models. The results of the analysis suggest that the model coefficients for traffic volume, proportion of cars, motorized two-wheelers and trucks in traffic, and driveway density and horizontal and vertical curvatures are randomly distributed across locations.

E.S.Park et al (2012)[19] studied the safety effect of wider edge lines was examined by analyzing crash frequency data for road segments with and without wider edge lines. The data from three states, Kansas, Michigan, and Illinois, have been analyzed. The consistent findings lend support to the positive safety effects of wider edge lines installed on rural, two-lane highways. In conclusion, their study lends scientific support to the positive safety effects of wider edge

lines installed on rural two-lane highways. Although the magnitudes of crash reductions were somewhat different from state to state, the results point in the samedirection.

3. METHODOLOGY

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next, the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyzes the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is.

3.1 Data Set

The dataset used has a total of 16 valued attributes and one class label. The valued attributes are further divided into numeric, categorical and other attributes.

Categorical Attributes	Numeric Attributes	Other Attributes	Class Label
1 Road Surface	1. Number of Vehicles	1 Reference Number	1 Accident
2 LightingCondition	2. Accident Date	2 Grid Ref: Easting	
3 Weather Conditions	3. Time (24hr)	3 Grid Ref: Northing	
4 Casualty Class	4. Age of Casualty	4 Expr1	
5 Casualty Severity		5 1st Road Class	
6 Sex of Casualty			
7 Type of Vehicle			

Table – 1: Categories of Attributes.

The levels of 7 categorical attributes vary according to the attributes. The Numeric attribute 'Number of Vehicles' and 'Age of Casualty' are simple numbered values. The Numeric attribute 'Accident Date' and 'Time' will be in the universal date and time format. Other attributes are those which are neither numbered nor categorized, but are fixed like reference identities. The class label 'Accident' consists of 2 values 'YES' and 'NO'.

3.2 Data Pre-processing

The purpose of data pre-processing is to reduce heterogeneity of data for better quality. The data collected is converted into understandable format. Real world data is often incomplete, inconsistent and lacking in certain behaviours or trends, and likely to contain many errors. So data cleaning is done to filter, aggregate, and fill in missing values. All the typist errors are rectified so as to keep the compatibility up. Here the raw data is organized for efficient access.

As an initial step all the single valued attributes that are of no use to the SVM (or removing which doesn't affect the result).are separated. Next all the missing valued attributes are marked as 'NA' so that the system could recognise the undefined values. Then spelling mistakes are rectified by renaming the values and replace lengthy names with smaller ones. Lastly, a class label is added as all the records present are YES conditions and we derive some NO conditioned records. The resultant data of this phase will be the input to the SVM Algorithm.

3.3 Classification Algorithm

Support Vector Machine is used as classification algorithm to classify the data. It is important to classify the road accident data in order to study it and get the required results. The dataset is divided into 2 parts training data and testing data. By analyzing the training data, SVM algorithm classifies the data. The more training it gets, the more efficient the model becomes.

4. Experimentation and Results

The proposed model predicts the class label as well as missing values (if required).

i) Internal prediction, which predicts the missing/categorical attribute values of train data.

ii) External prediction, which predicts the class label of test data.

The internal prediction is used to measure the performance of the model, as it is compared between the existing and predicted values shown in figure1.

Predicting Training Data - Internal Prediction

Parameters:

Attribute: Road Surface
SVM-Type: C-classification
SVM-Kernel: linear
cost: 1
gamma: 1

Number of Support Vectors: 1728
Number of Classes: 4
Levels: 'Dry' 'Frost / Ice' 'Snow' 'Wet / Damp'

Figure 1: Internal Prediction

The SVM classification plot for the internal prediction is as follows in figure2 and shows that linear kernel suits the best for the data.

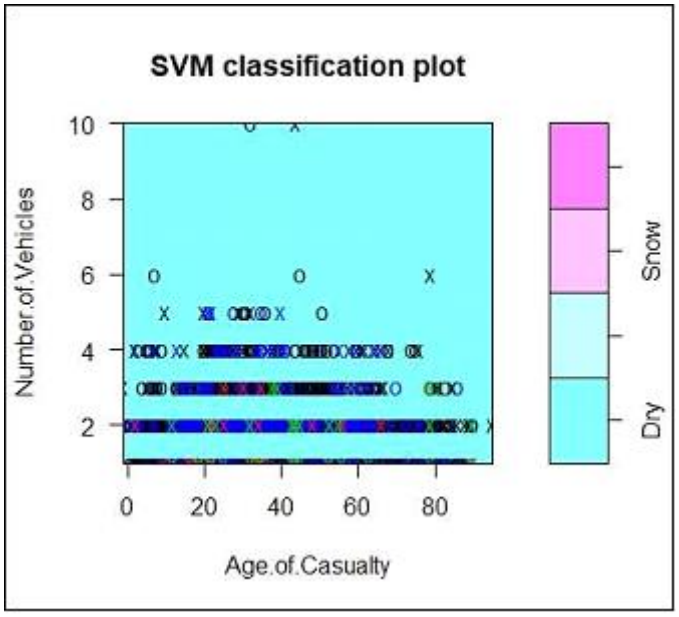

Fig2:SVM Plot with respect to attribute 'road acident'

The external prediction is strictly used for the prediction of test cases, i.e. for future predictions and is shown in fig3 as follows

Predicting Test Data - External Prediction

Parameters:

Attribute: Accident (Class Label)
SVM-Type: C-classification
SVM-Kernel: linear
cost: 1
gamma: 1

Number of Support Vectors: 64
Number of Classes: 2
Levels: 'NO' 'YES'

fig3. External Prediction

The SVM classification plot for the external prediction between Number of Vehicles and Age of Casualty which are classified respect to the class label Accident is as follows in figure4

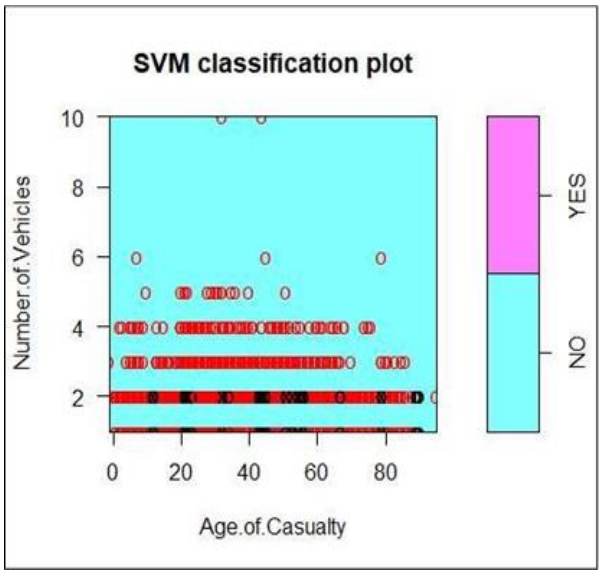

Fig 4. SVM Plot for Class Label

5. Performance Evaluation

SVM models of different attributes in trained data are evaluated. The whole data is classified according to each categorical attributes during training and the error rates and efficiencies are calculated for knowing the performance of each model.

The success conditions while testing the training data are shown in table 1.

Model Using the Attribute	Error rate In %	Efficiency In %
Road Surface	0.07	99.93
Lighting Conditions	1.53	98.47
Weather Conditions	0.31	99.69
1st Road Class	0.67	99.33
Casualty Class	9.33	90.67
Casualty Severity	2.67	97.33
Sex of Casualty	9.72	90.28

Table 1: Performance of successful cases.

Each model has a specific error rate but the model with least error rate (or the highest efficiency), i.e. Model generated with the attribute 'Road Surface' is chosen as the best model for predicting the test data.

Failure Conditions for training data is shown in table 2 as

Model Using the Attribute	Error rate In %	Efficiency In %
Number of vehicles	99.06	0.04

Table 2: Failure cases

As the attribute 'Number of Vehicles' is numeric but not categorical, it has more levels. So, the model fails to predict such specific values.

Predicting the class values of the test data

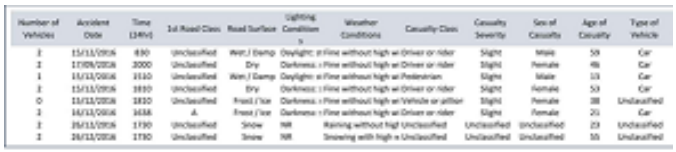

Fig 5. Input for test data

The records show only selective attributes as all the remaining attribute values are not assigned. The class label Accident is left not assigned so that the model can predict the values. Among the 8 records, first 6 records are possibilities of YES conditions and the last 2 records are NO conditions. The output of the model is the list of predicted class labels columns value and is shown in following table3

Ref No	2666	2667	2668	2669	2670	2671	2672
Predicted Value	yes	Yes	yes	yes	yes	yes	no

Table 3: outputs of test records

6.CONCLUSION

Current system is manual where government sector make use of ledger data and analyze the data manually, based on the analysis they will take the precautionary measures to reduce the number of accidents. Proposed system uses SVM classification technique for predicting road accidents.

There has been a drastic increase in the economic activities and consumption level, leading to expansion of travel and transportation. The increase in the vehicles, traffic lead to road accidents. Considering the importance of the road safety, government is trying to identify the causes of road accidents to reduce the accidents level. Using proposed system it is easy to predict accidents with its most probable conditions..

REFERENCES

[1] Neuma and Glenn on (1982)," A Theoretical model that relates accident on crest curves to available sight distance ", Transportation Research Record 923

[2] Glennon, J., 1985. Effect of Alignment on Highway Safety, Relationship between Safety and Key Highway Features. SAR 6, TRB Ltd., Washington, D.C., pp: 48-63.

[3] Luis F. Miranda-Moreno, Liping Fu,Frank F. Saccomanno, and Aurelie Labbe 2005 Alternative Risk Models for RankingLocations for Safety Improvement transportation Research Record: Journal of the Transportation Research Board, No. 1908,Transportation Research Board of the National Academies, Washington,D.C,pp. 1–8.

[4] Fajaruddin Mustakim (2008)," Black Spot Study and Accident Prediction Model Using Multiplication Linear Regression".Advancing and intergrating construction education, research and Practice, August 4-5, 2008

[5] . LI Chi(Department of Computer Science and Software Engineering of Jincheng College,Sichuan University,Chengdu 611731,China);Review of support vector machine and its applications in welding process[J];Electric Welding Machine;2011-10

[6] BirgulEgeli, MeltemOzturan, BertanBadur, Stock Market Prediction Using Artificial Neutal Networks,Bogazici University,Hisar Kampus,34342,Istanbul, Turkey.

[7] Mahdi PakdamanNaeini et.al Stock Market Value Prediction Using Neural Networks2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM)

[8] Cheng-Tao Chu, Sang Kyun Kim,Yi-An Lin, Map-reduce for machine learning on multi core CS. Department, Standford University 353 Serra Mall, Standford University, Stanford CA 94305-9025.

[9] Jimmy Lin and Michael Schatz Design Patterns for Efficient Graph Algorithms Map Reduce University of Maryland, College Park,MLG_10 Proceeding of the Eighth Workshop on Mining and Learning with Graph Pages 7885

[10] The general inefficiency of batch training of gradient descent learning, D. Randall , Volume 16, Issue 10, December 2003, Pages 1429–1451,ACM Digital Library, Elsevier Science Ltd. Oxford, UK, UK

[11] QI Bing-juan1,DING Shi-fei1,2(1 School of Computer Science and Technology,China University of Mining and Technology,Xuzhou 221116,China;2 Key Laboratory of Intelligent Information Processing,Institute of Computing Technology,Chinese Academy of Science,Beijing 100190,China);Support Vector Machine Based on Memberships of FCM[J];Microelectronics & Computer;2011-10.

[12] 100022);Feature Selection for Cancer Classification Based on Support Vector Machine[J];Journal of Computer Research and Development;2005-10.

[13] CHEN Nian-yi, LU Wen-cong, YE Chen-zhou, LI Guo-zheng (1. Laboratory of Chemical Data Mining, Department of Chemistiy, School of Science, Shanghai University, Shanghai, 200436, China; 2.Institute of Image and Pattern Recognition, Jiaotong University, Shanghai, 200030, China);Application of support vector machine and kernel function in chemometrics[J];Computers and Applied Chemistry;2002-06.

[14] Bureau,Dongying 257017,China;3.Dongxin Oil Production Factory,Shengli Oilfield Company of Sinopec,Dongying 257094,China);A new sanding prediction method and application[J];Journal of Guangxi University(Natural Science Edition);2011-02

[15] Saba B, Usman Q, Farhan HK, M. Younus J. MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble. Arab J Sci Eng, 2014; 39(11): 7771-7783.

[16] Jesmin N, Tasadduq I, Kevin ST, Yi-Ping Ph Ch. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Application, 2013; 40(4): 1086-1093.

[17] Kyle E. Walker*, Sean M. Crotty. Classifying high-prevalence neighborhoods for cardiovascular disease in Texas. Applied Geography, 2014; 57: 22-31, 2014. [18] K.Rajeswari, V.Vaithiyanathan, T.R. Neelakantan. Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks. International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012), Procedia Engineering, 2012; 41: 1818-1823

[18]. R.R. Dinu, A. Veeraragavan "Random parameter models for accident prediction on two-lane undivided highways in India", Journal of safety Research 42(2011) 39-42, 2011.

[19] E.S. Park et al. ,"safety effects of wider edge lines on rural, two-lane highways", Accident Analysis and prevention vol-48,317-325, 2012

[20]. Sachin Kumar and Durga Toshniwal,"A data mining framework to analyze road accident data" Journal of big data 2015 2:26