A COMPARATIVE STUDY OF CLASSIFICATION ALGORITHM USING ACCIDENT DATA

A. Priyanka
M.Phil Research scholar
PSGR Krishnammal College for Women
Coimbatore, India
priyait1991@gmail.com

K. Sathiyakumari
Assistant Professor
GR Govindarajulu School of Applied Computer Technology
Coimbatore, India
sathiyakumari@grgsact.com

Abstract— Road traffic accidents are the majority and severe issue, it results death and injuries of various levels. The traffic control system is one of the main areas, where critical data regarding the society is noted and kept as secured. Various issues of a traffic system like vehicle accidents, traffic volumes and deliberations are recorded at different levels. In connection to this, the accident severities are launched from road traffic accident database. Road traffic accident databases provide the origin for road traffic accident analysis. In this research work, Coimbatore city road traffic databases is taken to consideration, the city having higher number of vehicles and traffic and the city having higher number of vehicles and traffic and the cost of these loss and accidents has a great impact on the socio-economic growth of a society. Traditional machine learning algorithms are used for developing a decision support system to handle road traffic accident analysis. The algorithms such as SMO, J48, IBK are implemented in Weka version 3.7.9 the result of these algorithms were compared. In this work, the algorithms were tested on a sample database of more than thousand five hundred items, each with 29 accident attributes. And the final result proves that the SMO algorithm was accurate and provides 94%.

Keywords- Road Traffic Accident; SMO; J48; IBK; NCRB; MLP;

I. INTRODUCTION

India is being one of the fastest developing nation in the world with a vast population density, because of these the road traffic density is also increasing. In recent years, with the growth of the volume and travel speed of road traffic, number of traffic accidents, especially severe crashes [1], has been increasing hurriedly on a yearly basis. The issue of traffic safety has lifted great concerns across the world, and it has become one of the key issues demanding the sustainable development of modern traffic and transportation. For that reason, it is essential for engineers to be able to extract useful information from existing data to analyze the causes of traffic accidents, so that traffic administration can be more exactly informed. Traffic conditions are a multifaceted system due to many incidental factors [2], and traffic accident data has long been known to be very difficult to process. Many researchers have been made in recent years through applying various methodologies and algorithms.

In maintaining and managing the city's traffic system, the Coimbatore traffic office is structurally expected under three major departments' namely administration, accident investigation, security and control. The main aim of the Coimbatore city traffic office is to serving information handling. It has been visited that, the data extremely in some areas where the traffic and number of vehicles are wide, does not get enough attention to use it as a basis for decision-making. Identifying a given pattern of data in a given traffic office will help the decision makers to deciding the specific future activities.

Data mining [3] is a combination of methods, techniques and process in knowledge discovery. In other words, it requires a wide variety of tools ranging from classical and statistical techniques to neural networks and other new techniques originating from machine learning and artificial intelligence for improving databases promotion and process optimization. Six fundamental functions or activities of data mining are classified into directed and undirected. Specifically classification, evaluation and prediction are directed, when the available details are used to build a model that define one particular variable of significance in terms of the rest of available data. Similarly grouping or association rules, clustering, depiction and visualization are undirected data mining where the goal is to establish some relationship among all variables.

Data mining in traffic accidents [4] are which helps to find the hidden knowledge and rules, has become an essential research area in traffic safety. In recent years, most of the traffic information analyses are limited to general statistical analysis, which is hard to discover the rules hiding [5] in traffic accident information. Statistical analysis does not have the capability of map displaying and spatial analysis, and hence it is not able to find the spatial distribution characteristic and relationship between traffic accidents [6] and road network elements.

Thus, through this research work an attempt has been made to apply data mining tools and techniques in analyzing and determining interesting patterns especially with respect to possibilities of road traffic accident, on road accidents data at Coimbatore region traffic control System. In order to plan and implement effective strategies in reducing the accident severity and vehicle accident at Coimbatore city.

The remaining paper is organized as follows; Section II describes background study about the road traffic accident. Section III converse the related works behind in road traffic accidents prediction. Section IV focus on experimental results comparison. Finally, section V discuss about the conclusion and feature work.

II. BACKGROUND STUDY

Coimbatore is the third leading city in Tamil Nadu with a population of more than 15 lakhs. The city includes more than 30,000 of tiny, medium and large industries and textile mills are running popularly in Coimbatore. The city is also known entrepreneurship of its residents. The climate is too comfortable around the year. Road accident and injuries are now a growing and serious problem in the world.Maintaining the Integrity of the Specifications.

A. Accident Statistics

Coimbatore is ranked 23rd proportionate to its population in the number of fatalities in road accidents. A report on "Accidental Deaths and Suicides in India 2010" released by National Crime Records Bureau (NCRB) given that a total of 1,131 accidents on the city roads. Ranking is not done on the basis of number of accidents but on the basis of number of accidents proportionate to the population. The number of deaths in 2009 was 515. The following table (Table I) shows the road accident details includes Total number of accidents, number of persons injured and number of persons killed during the year 2012 to 2013.

TABLE I. ROAD ACCIDENT

Year	City	Total No. of Accidents	No. of persons injured	No. of persons killed
2012-2013	Coimbatore	3726	3434	1059

B. Vehicle Statistics

In the year 2013 alone, the Coimbatore Circle of the Transport Department is comprising Coimbatore, Tirupur and the Nilgiris districts find a registration of 2,17,785 new vehicles, both 2-wheelers and 4-wheelers, as against the registration of 2,15,627 vehicles of the previous year. The table (Table II) describes the number of vehicles registered in Coimbatore city.

TABLE II. NUMBER OF VEHICLES REGISTERED

Vehicle type	No. of registered vehicle
Auto- rickshaw	501
Motor Cab	788
Maxi Cab	364
School Bus	202
Ambulance	30
Light commercial vehicle	3254
Lorry	966
Motor Cycle	64192
Scooter	20743
Moped	7235
Motor Car	15130

III. RELATED WORK

In recent years, many researchers have been carried out through applying various methodologies and algorithms. In managing and controlling the city's traffic system.

T. Tesema, A. Abraham, and C. Grosan[7] proposed a rule mining of road traffic accident. They applied a clustering algorithm to the dataset to divide it into subsets, and then used each subset of data to train the classifiers. CART is the most advanced decision-tree technology for data analysis, pre-processing and predictive modeling. When the pre-processing was completed, the final dataset used for modeling had 4,658 records described by 16 attributes (13 base and 3 derived). Levenberg-Marquardt algorithm was used for the MLP training and achieved 65.6 and 60.4 percent classification accuracy for the training and testing phases, respectively. They compared the performance of Multi-layered Perceptron (MLP) and Fuzzy ARTMAP, and found that the MLP classification accuracy is higher than the Fuzzy ARTMAP. The FuzzyARTMAP achieved a classification accuracy of 56.1 percent.

Static approaches: M. Hirasawa [8] developed a model for traffic accident analysis system. It is an Urgent task to reduce these accidents by performing analyses and taking appropriate Countermeasures. Traffic accident data accumulated for more than ten years with digital map data indicating accident locations. The ultimate goal of this research is to establish a GIS-based system to analyze factors contributing to traffic accidents in Hokkaido, and to devise accident countermeasures. Toward developing a flexible system that achieves this goal, we have combined Arc View GIS Ver.3.2, a GIS software application from ESRI, with Access, a database management software application from Microsoft Corp. It is possible to customize menus by using Visual Basic, for enhanced user-friendliness. GIS visually displays the results of analyses, thus enabling sophisticated analysis and quick decision making. We combined the weather data with the accident data to create a new database. The relevant accident data were extracted.

H. Nabi, L. R. Salmi, S. Lafont, M. Chiron, M. Zins, and E.Lagarde [9] are proposed the behavioural predictors of serious road traffic crashes. The best predictors of serious RTCs were: "exceeding speed limits on rural roads", "risky use of cellular phone", and "sleepy driving". Adjusted Rate Ratio (RR) ranged from 1.47 to 2.16. Our study supports the view that individuals with a high propensity for driving behaviors associated with an increased risk of RTCs were more likely to have negative attitudes towards traffic safety. Changing drivers' negative or distorted opinions of traffic "enforcement" as well as "speed limitations" and "alcohol prohibition on roads" could improve their compliance with road traffic rules.

PramodAnantharam et al., [10] presented a method for traffic analysis Probabilistic Graphical Models Enhanced with Knowledge Bases. They presented an approach to leverage such \top-down" domain knowledge to enhance \bottom-up" building of graphical models. The proposed three operations on the graphical model structure to enrich it with nodes, edges, and edge directions. We illustrate the enrichment process using traffic data from 511.org and declarative knowledge from ConceptNet. Graphical models have been to deal with uncertainty, incompleteness, and dynamism within many domains. These models built from data often ignore pre-existing declarative knowledge about the domain in the form of ontologies and Linked Open Data (LOD) that is increasingly available on the web. The resulting enriched graphical model can potentially lead to better predictions of traffic delays.

S.Krishnaveni and Dr.M.Hemalatha [11] proposed a method for Perspective Analysis of Traffic Accident using Data Mining Techniques. They deal with the classification models to predict the severity of injury that occurred during traffic accidents. They have compared Naive Bayes Bayesian classifier, AdaBoostM1 Meta classifier, PART Rule classifier, J48 Decision Tree classifier and Random Forest Tree classifier for classifying the type of injury severity of various traffic accidents. The final result shows that the Random Forest outperforms than other four algorithms.

TibebeBeshah, Shawndra [12] Hill implemented a model to investigate the role of road-related factors in accident severity, using RTA data from Ethiopia and predictive models. Our three specific objectives include, exploring the underlying variables (especially road-related ones) that impact car accident severity and predicting accident severity using different data mining techniques, various related works have been analyzed here and the classification is done using decision tree, naïve bayes and knn algorithms. Decision tree and knn obtained an accuracy of about 80% and naïve bayes obtained an accuracy of about 79%.

IV. EXPERIMENT AND RESULTS

This research is mainly focus on predicting possibilities of road traffic accident in a particular area using machine learning techniques. There are three algorithms are used namely SMO, J48,IBK.

A. Support Vector Machine

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is important one which separates between a set of objects having different class memberships. Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports two different tasks like regression and classification and can handle multiple continuous and categorical variables.

To construct an optimal hyper plane SVM uses an iterative training algorithm, for minimize the error function. According to error function, SVM models can be classified into two distinct groups like Classification SVM and Regression SVM.

1) Classification SVM

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

subject to the constraints:

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \ldots, N$$

where C refers the capacity constant, w refers the vector of coefficients, b is a constant, and ξ_i refers parameters for handling non separable data (inputs). The index i label the N training cases. Note that $y \in \pm 1$ represents the class labels and xi represents the independent variables. The kernel function is used to transform data input (independent) to the feature space. It should be renowned that it is larger than C, then the error is penalized. So the C must be chosen with carefully to avoid over fitting problem. The SMO is based on the Support Vector Machine process.

B. Decision Tree

A decision tree [13] is a simple flowchart that selects labels for input values. This flowchart contains the decision nodes, which will check feature values, leaf nodes, and assign labels. To select the label for an input value, begin at the flowchart's initial decision node, called as its root node. This node contains a condition for checks one of the input value's features, and also selects a branch based on that feature's value. Following the branch that describes the input value, attain at a new decision node, condition on the input value's features. Then continue following the branch selected by each node's condition, until arrive at a leaf node which provides a label for the input value.

Once have a decision tree, it is straightforward to use it to assign labels to new input values. What's less straightforward is how build a decision tree that models a given training set. But before look at the learning algorithm for building decision trees, consider a simpler task: picking the best "decision stump" for a corpus. A decision stump is also a decision tree with a single node that decides how to classify inputs based on a single feature. It encompasses of one leaf for each possible feature value, identifying the class label that should be assigned to inputs whose features have that value. In order to build a decision stump, first decide which feature should be used. The simplest method is to build a decision stump for each possible feature, and get which one achieves the highest accuracy on the training data. Once feature has been picked, build the decision stump by assigning a label to each leaf based on the most frequent label for the selected examples in the training set. The information gain is used a criteria. Information gain is an impurity-based criterion that uses the entropy measure as the impurity measure.

$$Information\ Gain(a_i, S) = Entropy(y, S) - \sum_{v_{i,j}\ Edom(a_j)} \frac{\left|\sigma_{a_i = v_{i,j} S}\right|}{|S|} . Entropy\ (y, \sigma_{a_i = v_{i,j}}, S)$$

Where

$$Entropy(y, S) = \sum_{c_j\ Edom(a_i)} \frac{\left|\sigma_{a_i = c_j S}\right|}{|s|} . \log 2 \frac{\left|\sigma_{a_{i=c_j S}}\right|}{|S|}$$

The J48 is based on the Decision Tree process.

C. K-Nearest Neighbor

The k-nearest neighbor algorithm is amongst the simplest of all classification algorithms. In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. It can also be used for regression.

If k=1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose k to be an odd number. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its *k* nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

The neighbors are taken from a set of objects for which the correct classification or, in the case of regression, the value of the property is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidean distance, though other distance measures, such as the Manhattan distance could in principle be used instead. The *k*-nearest neighbor algorithm is sensitive to the local structure of the data

The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good *k* can be selected by various heuristic techniques, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (when $k = 1$) is called the nearest neighbor algorithm.

The accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling. Another popular approach is to scale features by the mutual information of the training data with the training classes. The IBK is based on the K-Nearest Neighbor process.

D. Data Collection

Road traffic accident is under persuading of many factors, which make it a complicate and as far as information is concerned, there are different databases of traffic accident in different countries. At present, roughly 1500 items of information are collected manually from the Coimbatore city traffic police, which includes 29 different attributes like Longitude, Latitude, Police Force, Number of Vehicles, Date, Day of Week, Time, Weather Conditions, Urban or Rural Area, Accident Severity, etc. These attributes can be used to rebuild the whole process of the accident in a relatively full and objective manner. It provides more than sufficient information and references for road traffic accident analyses.

E. Results

This research work focus on identifying the possibilities of road traffic accident in a given city. In this work Coimbatore city traffic data is taken to consideration with three different accident possibility level like low, medium and high respectively. The machine learning algorithms are implemented in Weka Version 3.7.9. The dataset contains 1500 items with 29 attributes respectively as mentioned above. For each classifier the dataset is given as two types training and testing, the training set contains 80% of data out of 1500 records and the testing set contains 20% of data out of 1500 items. K-fold cross validation is used to test the model and accuracy.

The experimental result shows that J48 classifier gives above 90% of accuracy when compare to other classifier and also it will proven that the possibility for road traffic accident is high in Coimbatore city. The following table (Table III) illustrates the accuracy comparison for each classifier. Fig. 1 shows the comparison chat for machine learning classifiers.

TABLE III. COMPARATIVE RESULTS OF CLASSIFIERS

Classifier	**Accuracy (%)**
SMO	94
J48	92.8
IBK	88.3

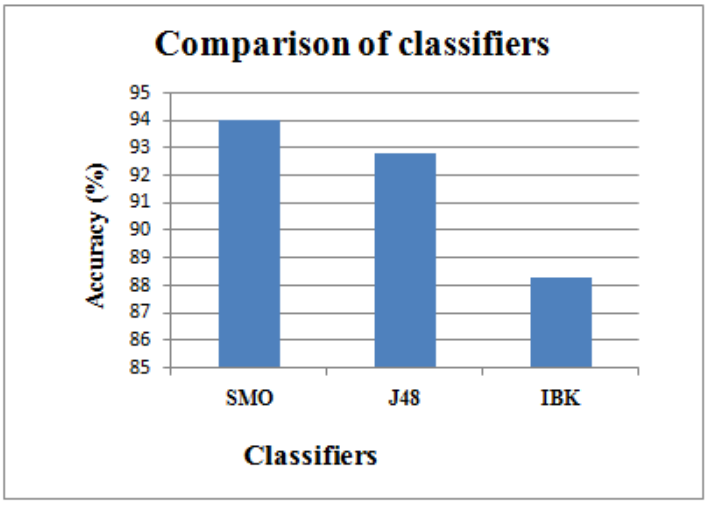

Fig. 1 Classification accuracy of the models

V. CONCLUSION

The objective of this research work is to explore the possible application of data mining technology at Coimbatore city road traffic accident data for developing a classification model. The classification model could support the traffic officers at Coimbatore city traffic office for making decisions in traffic control activities. Specifically it helps decision makers to understand driver's behavior, accident mode, time, road and weather conditions and other related issues which causing accidents resulting in fatalities or serious injuries so as to formulate enhanced traffic safety control policies. In support to traffic control system of Coimbatore City, several models were built by employing SMO approaches for identifying and extracting rules. The most best performing SMO classifier was chosen taken into account the reliability of the rules it generated and also the number of false negatives it reduced, finally its predictive accuracy was evaluated and analyzed. The outcome of this work was proven that possibilities of road traffic accident in Coimbatore city are high. The classification accuracy of the SMO was tested, and it showed an accuracy of 94%.

REFERENCES

[1] H.Nabi, L.R.Salmi, S.Lafont, M.Chiron, M.Zins, and E.Lagarde, "Attitudes associated with behavioral predictors of serious road traic crashes: results from the GAZEL cohort," Injury Prevention, vol.13,no.1, pp.26–31 ,2007.
[2] B.Yu, W.H.K.Lam, and M.L.Tam, "Bus arrival time prediction at bus stop with multiple routes," Transportation Research Part C , vol. 19, no. 6, pp. 1157–1170, 2011.
[3] L.-Y. Dong, G.-Y. Liu, S.-M. Yuan, Y.-L. Li, and Z.-H. Wu, "Applications of data mining to traffic accidents analysis," Journal of Jilin University Science Edition, vol.44, no.6, pp.951–955, 2006.
[4] D.-H. Lee, S.-T. Jeng, and P. Chandrasekar, "Applying data mining techniques for traffic incident analysis, "Journal of the Institution of Engineers, vol.44, no.2, pp.90–101, 2004.
[5] Marie-France Joly, Robert bourbeau and Jacques Bergeron, "What Can We Learn from the Experience of Risk Location Identification?", Proceedings of International Conference on Traffic Safety, New Delhi, India, January 1991.
[6] Babkov, V.F, Road Conditions and Traffic Safety; Mir Publishers; Moscow.
[7] T. Tesema, A. Abraham, and C. Grosan, "Rule mining and classification of road traffic accidents using adaptive regression trees. I," Journal of Simulation, vol. 6, no. 10, pp. 80–94, 2005.
[8] M. Hirasawa, "Development of traffic accident analysis system using GIS," Proceedings of the Eastern Asia Society for TransportationStudies, vol. 10, no. 4, pp. 1193–1198, 2005.
[9] H. Nabi, L. R. Salmi, S. Lafont, M. Chiron, M. Zins, and E.Lagarde, "Attitudes associated with behavioral predictors of serious road traffic crashes: results from the GAZEL cohort,"Injury Prevention, vol. 13, no. 1, pp. 26–31, 2007.
[10] PramodAnantharam, KrishnaprasadThirunarayan, AmitSheth, "Tra_c Analytics using Probabilistic Graphical Models Enhanced with Knowledge Bases" fpramod, tkprasad,Kno.e.sis - Ohio Center of Excellence in Knowledge-enabled Computing Wright State University, Dayton, USA.
[11] S.Krishnaveni, Dr.M.Hemalatha, "A Perspective Analysis of Traffic Accident using DataMining Techniques" in International Journal of Computer Applications (0975 – 8887) Volume 23– No.7, June 2011.
[12] Tibebe Beshah1, Shawndra Hill2," Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia".
[13] Lior Rokach, Oded Maimon, "Decision Trees", Department of Industrial Engineering, Tel-Aviv University.
[14] Yang Song, Jian Huang, DingZhou, Hongyuan Zha, and C. Lee Giles, "IKNN: Informative K-Nearest Neighbor Pattern Classification", Springer-Verlag Berlin Heidelberg, PKDD 2007, LNAI 4702, pp. 248–264, 2007.
[15] H. Wan-Jo Yu, "Data Mining via Support Vector Machines: Scalability, Applicability, and Interpretability", Research work.