Road Traffic Accident Analysis using Improving Road Safety using Classification Algorithm

N. Hemalatha[1], C. Radhakrishnan[2], N. Premkumar[3]
[1, 2, 3] Department of Computer Science
[1, 2, 3] Kongunadu college of Engineering and Tecnology, Trichy

Abstract- *Traffic Accidents is a major public health problem, causing approximately 1.2 million deaths and 50 million injuries worldwide each year. In developing countries, interest rates between the main reason for deaths and injuries; in particular, India experienced the highest rate accidents of this type. Therefore, in order to reduce the severity of the accident was of great interest to transportation Institutions and the public. In this paper, we apply data mining techniques to link the characteristics of recording the severity of road accidents in India, and to develop a mechanism through which transit can be used to improve India's security rules.*

This paper deals with the some of classification models to predict the severity of injury that occurred during traffic accidents. In this project I compare Naïve Bayes and KNN (k- nearest neighbor) algorithms for classifying the injury type of various road traffic accidents.

Keywords- Classifier, Data Mining, Distance Measures, KNN, Naïve Bayes.

I. INTRODUCTION

India is still one of the fastest developing nations in the world, with a large population density; because of this density road traffic is also increasing. In recent years, with the growth speed and the displacement volume of road traffic, the number of traffic accidents, especially serious accidents [1] has been increasingly in a hurry annually. The issue of traffic safety has raised great concern worldwide, and has become one of the key issues that require the sustainable development of modern traffic and transport. 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 traffic management can be more accurately informed. Traffic conditions are a multifaceted system because many incidental factors [2], and data from traffic accidents has long been known to be very difficult to process. Many researchers have made in recent years through the application of various methodologies and algorithms.

In the maintenance and management system of city traffic, the traffic system in India is expected structurally into three main departments namely management, accident investigation, safety and control. The main office target city traffic is to serve the manipulation of information. It has been viewed that data very few areas where traffic and the number of vehicles are wide, does not get enough attention to the use as a basis for decision making. The identification of a given data traffic in an office pattern as help the decision makers to decide on future specific activities.

Data mining [3] is a combination of methods, techniques and knowledge discovery processes. In other words, a wide variety of tools ranging from classical statistical techniques and neural networks is required and other new techniques from machine learning and artificial intelligence to improve the promotion of databases and process optimization. The fundamental functions or data mining activities are classified into directed and undirected. Specifically classification, evaluation and prediction target; when the details available are used to build a model that defines a particular variable of importance in terms of the rest available data. The same grouping or association rules, clustering, representation and visualization mode they are not addressed data mining, where the goal is to establish some kind of relationship between all variables.

Data mining in traffic accidents [4] are helping to find hidden knowledge and rules, it has become a key area of research in traffic safety. In recent years, most of the analysis of general traffic information statistical analysis, it is difficult to discover the hiding rules [5] in the traffic accident information are limited. Statistical analysis has the capability to map and showing spatial analysis, and therefore is not able to find the spatial distribution characteristic and the relationship between accidents [6] and road network elements.

Thus, through this research it has made an attempt to apply the tools and techniques of data mining to analyze and identify interesting patterns especially regarding the chances of accident in the accident data traffic control system traffic. In order to plan and implement effective strategies it is in reducing the severity of the accident and traffic accident in the city.

II. RELATED WORKS

It is estimated that the annual cost for the NHS because accidents are more than 10,000 US dollars a year in funding the medical field will be the amount of euro .Tremendous Because of these accidents. In a larger scale, the World Health Organization, in conjunction with the World Bank, the end analysis, more than 100 million people worldwide die each year from road accidents and collisions and injuries result in 2020 road accidents can overcome AIDS and tuberculosis Rank three reasons of premature death and disability worldwide.

[1] To identify statistically significant factors using a logistic regression model that predict the probabilities of crashes and injury crashes aiming at using these models it perform a risk assessment of a given region. This model describes a site by its land use activity, road side design, use of traffic control devices and traffic exposure. Their studies focused on village sites are less hazardous than residential and shopping sites in city.

[2] Classification and regression tree (CART) and negative binomial regression models to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors. This study focused on Automobile industry to help to improve vehicle safety and some environmental organizations are help to reduce the pollution to minimize the traffic accidents and traffic congestion.

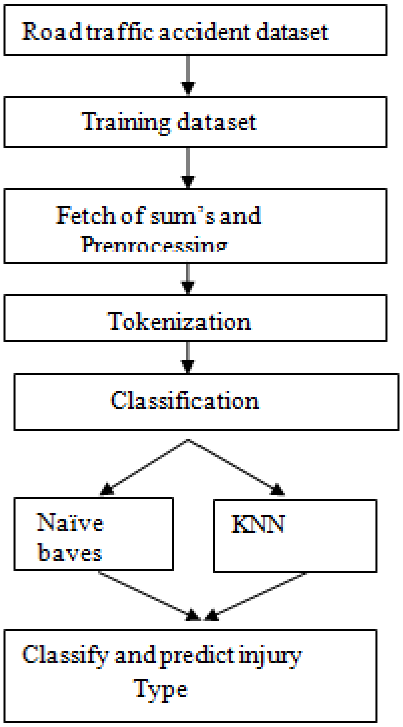
[3] For two RTA severity categories, various algorithm used for improve the individual classifier. Using neural network and decision tree individual classifiers, three different approaches were applied: classifier fusion based on the Dempster– Shafer algorithm, the Bayesian procedure, and logistic model; data ensemble fusion based on arcing and bagging; and clustering based on the *k*-means algorithm. Their empirical results show that a clustering based classification algorithm works optimal for road traffic accident classification in Korea.

[4] The statistical properties of four regression models: two conventional linear regression models and two Poisson regression models in terms of their ability to model vehicle accidents and highway geometric design relationships. Highway Safety Information System (HSIS) has Roadway and truck accident data that have been employed to illustrate the use and the limitations of these models. The Poisson regression models, on the other hand, possess most of the desirable statistical properties in developing the relationships.

[5] A historical RTA data, including 4,658 accident records at the Addis Ababa Traffic Office, the records used to investigate the analysis of accident severity in Addis Ababa, Ethiopia. Using the DT technique and applying the Knowledge SEEKER algorithm of the Knowledge STUDIO data mining tool, the developed model classified the accident data based on accident severity & classified into four classes: fatal injury, serious injury, slight injury, and property damage.

[6] To analyze accident data Non-parametric Classification tree techniques is used from the year 2001 for Taipei, Taiwan. A CART model was developed to establish the relationship between injury severity and driver/vehicle characteristics, Highway/environment variables and accident variables.

System Architecture


Road traffic accident dataset
Training dataset
Fetch of sum's and Preprocessing
Tokenization
Classification
Naïve bayes
KNN
Classify and predict injury Type

III. METHODOLOGY

Traffic Accident Dataset

This data set of traffic accidents is obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the period 1991-2000. More specifically, the data are obtained from the Belgian “Analysis Form for Traffic Accidents” that should be filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. In total, 340.184 traffic accident records are included in the data set. The traffic accident data contain a rich source of information on the different circumstances in which the accidents have occurred: course of the accident (type of collision, road users, injuries …), traffic conditions (maximum speed, priority regulation …), environmental conditions (weather, light conditions, time of the accident …), road conditions (road surface, obstacles …), human conditions (fatigue, alcohol …) and geographical conditions (location, physical characteristics …). In total, 572 different attribute values are represented in the data set. On average, 45 attributes are filled out for each accident in the data set.

Naïve Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

A Naive Bayesian classifier is a simple probabilistic classifier based on applying Bayesian theorem (from Bayesian statistics) with strong (naive) independence norms. By the use of Bayesian theorem we can write

$$p(C|F1 \ldots Fn) = \frac{p(C)p(F1 \ldots Fn|C)}{p(F1 \ldots Fn)}$$

Advantages of Naive Bayes

- The Naive Bayes algorithm affords fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows. The build process for Naive Bayes is parallelized. (Scoring can be parallelized irrespective of the algorithm.)
- Naive Bayes can be used for both binary and multiclass classification problems.

K-Nearest Neighbor

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

- For each training example <x,f(x)>, add the example to the list of training examples.
- Given a query instance x_q to be classified,
- Let x_1,x_2...x_k denote the k instances from training examples that are nearest to x_q .
- Return the class that represents the maximum of the k instances.

Advantages

– Analytically tractable
– Simple implementation
– Nearly optimal in the large sample limit ($N \rightarrow \infty$)

$$P_{Bay}[error] < P1_{NN}[error] < 2P_{Bayes}[error]$$

– Uses local information, which can yield highly adaptive behavior
– Lends itself very easily to parallel implementations

Euclidean Distance Measure

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric.

The Euclidean distance, data vector p and centroid q is computed as

$$d(p, q) = \sqrt{\sum_{k=1}^{n} (q_{ik} - p_{ik})^2}$$

IV. EXPERIMENTAL RESULT

This research is mainly focus on predicting possibilities of road traffic accident in a particular area using machine learning techniques. There are two algorithms are used namely Naïve Bayes and K-nearest neighbor.

A. 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 road traffic accident dataset, which includes 45 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.

B. Results

This research work focus on identifying the possibilities of road traffic accident in a given city. In this work road traffic accident data is taken to consideration with three different accident possibility level like low, medium and high respectively. The machine learning algorithms are implemented in MATLAB. The dataset contains 1500 items with 45 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 city. The following table illustrates the accuracy comparison for each classifier.

Table 1: Comparative result of Classifiers

Classifier	Accuracy
Naïve Bayes	92.8
K-nearest neighbor	94.2

Figure. 1 shows the comparison chat for machine learning classifiers

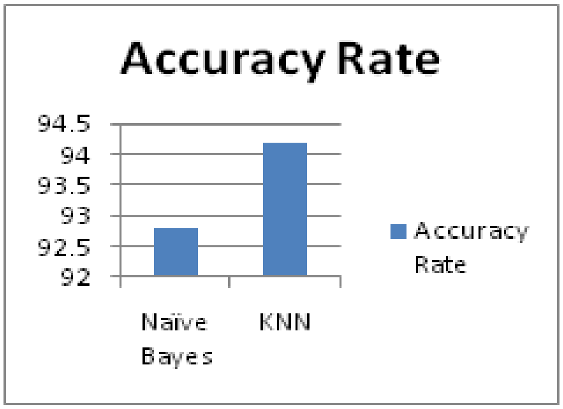


Figure 1 Classification accuracy of the models

V. CONCLUSION

The aim of this study is to explore the development of data mining techniques may sort mode application data in a traffic accident. Classification model can support traffic control activities in transportation decisions. In particular, policy-makers to help understand the behavior, accidents, weather and road conditions and weather driving, causing an accident resulting in death or serious injury problems to develop better policies to control traffic safety. In order to support the city's traffic control system, several models use SMO methods to identify and extract the rules constructed. Selected best performing classification considers generation, but also reduce the number of false negatives, a final evaluation and analysis of the reliability of their rules of prediction accuracy. The results of this study indicate that the likelihood of accidents is high. KNN classification accuracy were tested and showed 94.2% accuracy.

REFERNCES

[1] H.Nabi, L.R.Salmi, S.Lafont, M.Chiron, M.Zins, and E.Lagarde, “Attitudes associated with behavioral predictors of serious road tragic 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] Pramod Anantharam, Krishnaprasad Thirunarayan, Amit Sheth, "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.