A Study and Analysis of Road Accident in Tamilnadu using Data mining Technique

MS.R. Saravanya [1] Ms.Mangayarkarasi[2]

1. *M.Phil Scholar, Department of Computer science Vels University, Chennai.*
2. *Professor, Department of Computer science Vels University, Chennai.*

***Abstract** - In this modern world the usage of automobiles by the people are increasing day by day. As such due to enhanced traffic in urbanized areas such as in highways and roads, Motor vehicle accident and rail accidents are increasing in our state as well as in our country. Though accidents are not wantonly being done ,the causes of the accidents are many such as drunk and drive, violation of traffic rules ,non application of protective appliances, defective roads ,obstacles on the road ,due to workload of continuous driving for hundreds of hours and due to defective mechanism in the motor vehicle. Only few accidents are due to actus reus'. Consequent to the increasing number of accidents there are losses of precious human lives and limbs, loss of properties, Traffic Jam etc., and they are root cause for some social problems. So it is just and necessary to curtail the road accidents. By way of detecting the basic reasons for the occurring accidents it would be easier to prevent the accident in future. It will be useful to the police authorities as well as to the entire society for awareness. So the data analyzing of road accidents being done. This research aims to provide a review to extract useful information by means of Data Mining, in order to find accident hot spots out and predict accident trends for them using data mining techniques.*

Keywords

Data mining, Accident, clustering, Classification

I.INTRODUCTION

The ever increasing tremendous amount of data, collected and stored in large and numerous data bases, has far exceeded human ability for comprehension without the use of powerful tools [3]. Consequently, important decisions are often made based not on the information rich data stored in databases but rather on a decision maker's intuitions due to the lack of tools to extract the valuable knowledge embedded in the vast amounts of data [3]. This is why data mining has received great attention in recent years. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis [3][19]. General data mining principles, including Associations, Sequential Patterns, Classifications, Predictions, and Clustering can be applied to many areas. Classification algorithms give interesting results from a large set of data attributes. The costs of fatalities and injuries due to traffic accidents have a great impact on society. The World Health Organization [14] predicts that road collisions will jump from the ninth leading cause of death in 2004 to the fifth in 2030. Many research works are concentrating on analyzing various crash related factors which increase the death ratio. In relation to this, fatal severities resulted from road traffic accident are one of the areas of concern. Out of all road related factors the manner of collision influences the fatal rate. As the size of these accident databases increases rapidly both spatially and temporally, it is quite a challenge to analyze and extract useful information from them without using advanced data analysis tools. The contribution of classification algorithms in analyzing the road accident factors are discussed in the following sections. The next subsection gives an overview of the paper.

Data Analysis

IRTAD, GLOBESAFE - ACCIDENT RESOURCE

Limitations

Research in respect of whole India is a tedious task.So the research focused here is to do research only about Tamilnadu.

Deployment

Sno	Variable	Description
1	Vehicle Type	Small Cars Heavy Vehicle vehicle body type, vehicle age, vehicle role Vehicle Condition
2	Time of the Day	Morning Afternoon Evening Night/Midnight
3	Season	Wet/Dry/muddy
4	Causes	Wrong Overtaking,Careless Drivivg,Loss of control Tyre Bust,Over Speeding,Obstruction,Pushed by another Vehicle,Broken Shaft,Broken Spring,Break Failure Road Problem,Unknown Causes,Robbery Attack,Alcohol/drug
5,	Person,	Age,Gender,Driver/ passenger,Race/Ethnicity
6	Injury	No injury, Possible injury, Non-incapacitating injury, Incapacitating injury, Fatal injury.
7	Intial Point of Impact	no damage/non-collision, front, right side, left side, back, front right corner, front
8	Accident Information	month, Region, primary sampling unit, the number of the police jurisdiction, case number, person number,vehicle number, vehicle make and model,RoadSeparation , RoadOrientation,RoadSurfaceType,RoadSurfCondition WeatherCondition , LightCondition

II. LITERATURE SURVEY

Handan et.al [4] compared logistic regression model with classification tree method in determining social-demographic risk factors which have affected depression status of women in separate postpartum periods. They proposed that Classification tree method gives more information with detail on diagnosis by evaluating a lot of risk factors together than logistic regression model.

Chang et.al [2] applied non-parametric classification tree techniques to analyze Taiwan accident data from the year 2001. They developed a CART model to find the relationship between injury severity and driver/vehicle characteristics, highway/environment variables, and accident variables. Yong Soo Kim [11] compared the performance of data mining and statistical techniques by varying the number of independent variables, the types of independent variables, the number of classes of the independent variables, and the sample size. The results have shown that the artificial neural network performance improved faster than that of the other methods as the number of classes of categorical variable increased.

I-Cheng et.al [5] investigated the accuracy of data mining techniques viz. discriminate analysis, logistic regression, Bayes classifier, nearest neighbor, artificial neural networks, and classification trees in analyzing customers' default credit payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. Their results reveal that artificial neural network is the only one that can accurately estimate the real probability of default credit payments.

Weimin et.al [10] demonstrated that the hybrid SVM technique having better capability of capturing nonlinear relationship among variables and had best classification rate than CART, MARS and SVM while analyzing the credit card data.

Nojun et.al [9] analyzed the limitation of Mutual Information Feature Selector (MIFS) and proposed a method to overcome this limitation. Isabelle et.al [6] discussed the basics of feature selection and summarized the steps to solve a feature selection problem. The implementation of various feature selection algorithms have been discussed in [15]. Next section summarizes the details about the training data set.

III.SYSTEM ARCHITECTURE

In this paper we have compared few classification algorithms with and without using feature selection algorithms. The steps carried out in our study are depicted in Figure 1. The data set is divided into training set which consists of 60% of total records and test set which consists of 40% of total records. Training set is used to build the model and test set is used to validate the model for correctness.

IV.CLASSIFICATION ALGORITHMS

Classification trees are used to predict membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables. Classification tree analysis is one of the main techniques used in Data Mining [19]. Next subsections deals with the basic classification algorithms we used in our study.

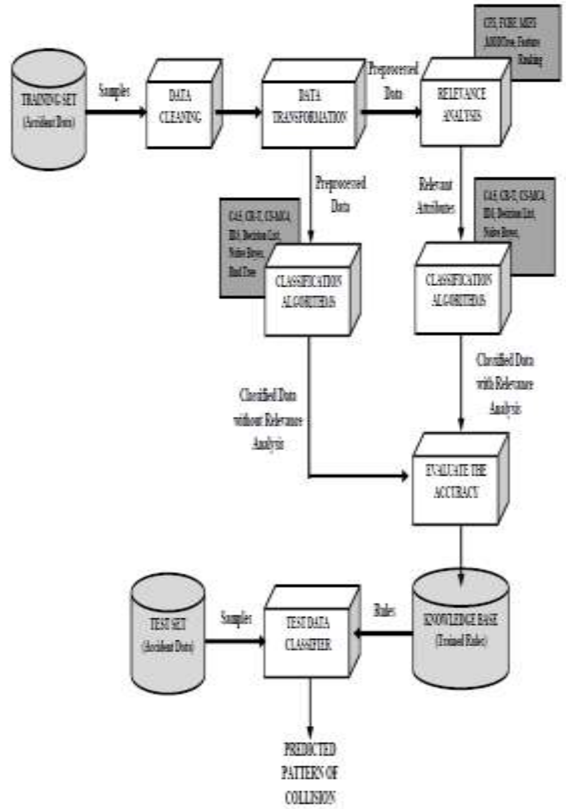
TRAINING SET
(Accident Data)
Samples
DATA CLEANING
DATA TRANSFORMATION
Preprocessed Data
RELEVANCE ANALYSIS
CFS, FCBF, MIFS
Feature Ranking
Preprocessed Data
Relevant Attributes
C4.5, CR-T, CS-MC4, ID3, Decision List, Naive Bayes, Rnd Tree
CLASSIFICATION ALGORITHMS
C4.5, CR-T, CS-MC4, ID3, Decision List, Naive Bayes,
CLASSIFICATION ALGORITHMS
Classified Data without Relevance Analysis
Classified Data with Relevance Analysis
EVALUATE THE ACCURACY
TEST SET
(Accident Data)
Samples
TEST DATA CLASSIFIER
Rules
KNOWLEDGE BASE
(Trained Rules)
PREDICTED PATTERN OF COLLISION

Fig 1: Methodology

5.1 C4.5

C4.5 starts with large sets of cases [16] belonging to known classes. The cases, described by any mixture of nominal and numeric properties, are scrutinized for patterns that allow the classes to be reliably discriminated. These patterns are then expressed as models, in the form of decision trees or sets of if-then rules that can be used to classify new cases, with emphasis on making the models understandable as well as accurate.

5.2 ID3

ID3 is a decision tree induction algorithm. In the decision tree each node corresponds to a non-categorical attribute [17] and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute for the records described by the path from the root to that leaf. In the decision tree at each node should be associated the non-categorical attribute which is most informative among the attributes not yet considered in the path from the root. Entropy is used to measure how informative is a node.

The ID3 algorithm takes all unused attributes and counts their entropy concerning test samples. Choose attribute for which entropy is minimum (or, equivalently, information gain is maximum).

5.3 C&RT

Classification and Regression Trees is a classification method [3] which uses historical data to construct decision trees. Decision trees are then used to classify new data. It works like ID3 except it results in binary decision tree.

5.4 CS-MC4

Cost sensitive decision tree algorithm uses m-estimate smoothed probability estimation [12]. It minimized the expected loss using misclassification cost matrix for the detection of the best prediction with in leaves.

5.5 Decision List

Decision Trees are simple knowledge representation and they classify examples to a finite number of classes, the nodes are labeled with attribute names, the edges are labeled with possible values for this attribute and the leaves labeled with different classes. Tree shaped structures represents set of decisions which generate rules for the categorization of dataset. Decision trees produce rules that are mutually exclusive and collectively exhaustive with respect to the training database. Particular decision trees methods consist of Classification and Regression Tree (CART) and Chi Square Automatic Interaction Detection (CHAID).

.

5.6 Naive Bayes

The Naive Bayes Classifier [19] technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

5.7 Random Tree

Random tree [18] can be applied to both regression and classification problems. The method combines "bagging" idea and the random selection of features in order to construct a collection of decision trees with controlled variation. Each tree is constructed using the following algorithm:

- Let the number of training cases be N, and the number of variables in the classifier be M.
- We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.
- Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample).
- Use the rest of the cases to estimate the error of the tree, by predicting their classes.
- For each node of the tree, randomly choose m variables on which to base the decision at that node.
- Calculate the best split based on these m variables in the training set.
- Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).
- For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in.
- This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction [18].

5.8 Rule Induction

Rule Induction is the process of extracting valuable if-then condition rules from data based on statistical and mathematical consequence. Rule induction on a database can be massive undertaking in which all possible patterns are systematically pulled out of the data and then accuracy and significance calculated, telling users how strong the pattern is and how likely it is to occur again.

CONCLUSION

The aim of this study was to show the applications of data mining techniques in the field of accident investigation. It was done by reviewing various papers. We are currently enhancing it by considering several issues; variation in crash occurrence may have some consequence for traffic safety measures in some places in Tamilnadu. The modeling will be to combine road-related factors with driver information for better predictions, and to find interactions between the different attributes. From the variation we've seen among the different datasets, we believe that some sort of standardization should be enforced among the different police departments in order to make automatic parsing of accident reports more reliable. It will be useful to the police authorities as well as to the entire society for awareness. So the data analyzing of road accidents being done.

8. REFERENCE

[1] Andreas G.K., Janecek, Wilfried N. Gansterer, Michael A. Demel Michael, Gerhard F. Ecker, “On the Relationship Between Feature Selection and Classification Accuracy”, 2008, JMLR: Workshop and Conference Proceedings, pp.90-105.

[2] Chang L. and H. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques Accident analysis and prevention", 2006, Accident analysis and prevention, Vol. 38(5), pp 1019-1027.

[3] Han, J. and Kamber, M., “Data Mining: Concepts and Techniques”, Academic Press, ISBN 1- 55860-489-8.

[4] Handan Ankarali Camdeviren, Ayse Canan Yazici, Zeki Akkus, Resul Bugdayci, Mehmet Ali Sungur, “A Comparison of logistic regression model and classification tree: An application to postpartum depression data”, 2007, Expert Systems with Applications, Vol. 32 ,pp. 987–994.

[5] I-Cheng Yeh, Che-hui Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients”, Expert Systems with Applications, 2009, Vol.36, pp. 2473–2480.

[6] Isabelle Guyon, Andr´e Elisseeff, “An Introduction to variable and Feature Selection”, Journal of Machine Learning Research, 2003, Vol. 3, pp. 1157-1182.

[7] Lei Yu, Huan Liu, “Feature Selection for high-Dimensional Data: A Fast Correlation-Based Filter Solution”, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[8] Mark A. Hall, "Correlation Based Feature Selection for Machine Learning”, Ph.D. Thesis, Department of Computer Science, Waikato University, Hamilton, NZ, 1999.

[9] Nojun Kwak and Chong-Ho Choi , “Input Feature Selection for Classification Problems”, IEEE Transactions On Neural Networks, Vol. 13, No. 1, January 2002.

[10] Weimin Chen , Chaoqun Ma, Lin Ma , “Mining the customer credit using hybrid support vector machine technique”, Expert Systems with Applications, 2009, Vol. 36, pp. 7611–7616.