Analyzing the Road Traffic and Accidents with Classification Techniques

M. Sowmya [1], Dr.P. Ponmuthuramalingam [2]
M. Phil Research Scholar [1]
Department of Computer Science
Government Arts College
Coimbatore, Tamil Nadu, INDIA.
Associate Professor& Head [2]
Department of Computer Science
Government Arts College (Autonomous)
Coimbatore, Tamil Nadu, INDIA.

Abstract - Data mining is the process of extracting data's from the database engines. Now a days the road traffic and accidents are main area for the researchers to discover the new problems behind that. It is commonly used in a marketing, inspection, fraud detection and scientific invention. In data mining, machine learning is mainly focused as research which is automatically learnt to recognize complex patterns and make intelligent decisions based on data. Nowadays road accidents are the major causes of death and injuries in this world. Roadway configurations are useful in the development of traffic safety control policy. Spatial data mining is a research area concerned with the identification of interesting spatial patterns from data stored in spatial databases and geographic information systems (GIS). This paper addresses the analysis of spatial and time stamped data of Slovenian traffic accidents which, together with the GIS data, enabled the construction of spatial attributes and the creation of a time-stamped spatial database.

Key terms: Mining, road traffic, Data mining, spatial databases

1. INTRODUCTION

Ethiopia has the highest rate of Road Traffic Accidents(RTA), due to the fact that road transport is the major transportation system in the country. The costs of fatalities and injuries due to Road Traffic Accidents have a tremendous impact on societal well-being and social economic development. RTAs are among the leading causes of death and grievance worldwide, causing an estimated 1.2 million deaths and 50 million injuries each year. The Ethiopian traffic control system archives data on various aspects of the traffic system, such as traffic volume, concentration, and vehicle accidents. With more vehicles and traffic, the center city of Addis Ababa takes the lion's share of the risk, with an average of 20 accidents being recorded every day and even more going unreported.

The basic hypothesis of this research is that accidents are not randomly scattered along the road system, and that drivers are not involved in accidents at random. There are complex circumstantial relationships between several characteristics (driver, road, car, etc.) and the accident occurrence. As such, one cannot improve safety without successfully relating accident frequency and severity to the causative variables. We will attempt to extend the authors' previous work in this area by generating additional attributes and focusing on the contribution of road related factors to accident brutality in Ethiopia[1]. This will help to identify the parts of a road that are uncertain, thus supporting traffic accident data analysis in decision making processes. Traffic control system is the area, where critical data about the society is recorded and kept. Using this data, we can identify the risk factors for vehicle accidents, injuries and fatalities and to make preventive measures to save the life. The severity of injuries causes an impact on the society. The main objective of the research was to find the applicability of data mining techniques in developing a model to support road traffic accident severity analysis in preventing and controlling vehicle accidents. It leads to death and injuries of various levels. Understanding the patterns of hidden data is very hard due to data accumulation. Organization keeps data on their domain area for maximum usage. Apart from the gathering data it is important to get some knowledge out of it. For effective learning, data from different sources are gathered and organized in a consistent and useful manner.

1.1. DATA MINING

Data mining or Knowledge Discovery in Databases (KDD) is the nontrivial extraction of implied, previously unknown, and potentially useful information from data. Knowledge discovery in databases (KDD) is the process of recognizing a valid, potentially, useful and ultimately understandable structure in data. Data mining is a young and promising field, used to knowledge discovery from data. Data Mining is the process of automatic abstraction of novel, useful, and understandable patterns in very large databases. High-performance accessible and parallel computing is crucial for ensuring system scalability and interactivity as datasets grow inexorably in size and complexity [4]. It is the mechanized process of modeling large databases by means of discovering useful patterns. Consequently, data mining consists of more than collecting and handling data, it includes analysis and prediction

]Financial, communication, and marketing organizations and companies with strong consumer-focus retails primarily use data mining today. It enables these companies and organizations to determine relationships among internal factors such as cost, product positioning, or staff skills, and external factors such as economic gages, competition, and customer demographics. In addition, it enables them to determine the impact on sales, customer gratification, and corporate yields. Finally, it enables them to drill down into summary information to view detail transactional data.

Businesses employing data mining quickly see a return on venture, but also they recognize that the number of predictive models can quickly become very large. Rather than one model to foretell which customers will agitate, a business could develop a separate model for each region and customer type. Then instead of carriage an offer to all people that are likely to agitate, it may only want to send offers to customers that will likely take to offer. Finally, it may also want to determine which customers are going to be profitable over a window of time and only send the offers to those that are likely to be lucrative. In order to maintain this measure of models, they need to manage model versions and move to automated data mining.

Data mining in customer relation management applications can contribute significantly to the bottom line. Rather than contacting a outlook or customer through a call center or sending mail, only panorama that are predicted to have a high likelihood of responding to an offer are contacted. Methods that are more sophisticated may be used to optimize across campaigns so that we can predict which channel and which offer an individual is most likely to respond to - across all latent offers. Finally, in cases where many people will take an action without an offer, uplift modeling can be used to regulate which people will have the greatest increase in responding if given an offer.

Classification and clustering are two important techniques in data mining. Classification means classify the data or grouping the data based on some criteria. In Writings, there are different models to compute the classification tasks, but the only model, which is most popular, is the decision tree. Decision trees have been found very effective for classification especially in Data Mining A decision tree is a tree structure in which each internal node denotes a test on an attribute, each branch signifies an outcome of the test and leaf nodes represent classes.

Clustering is another important technique in data mining. Clustering is the classification of objects into different clusters, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure.Data clustering algorithms can be hierarchical or partitioned. Hierarchical algorithms find consecutive clusters using previously established clusters, whereas partitioned algorithms regulate all clusters once. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms instigate with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms initiate with the whole set and proceed to divide it into successively smaller clusters.

Scope of Data mining derives its name from the similarities between searching for valuable business information in a large database for instance, discovering linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes needs either sifting through an immense amount of material, or intelligently inquiring it to find exactly where the value resides. Given databases of adequate size and quality, data mining technology can breed new business opportunities by providing these capabilities:

1.2. ADVANTAGES OF DATAMINING

Data mining automates the process of finding predictive information in large databases. Data mining tools sweep through databases and identify previously hidden patterns in one step. Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be executed on new systems as existing platforms are upgraded and new products established. When data mining tools are implemented on high performance parallel processing systems, they can analyze vast databases in minutes. Rapid processing means that users can automatically experiment with more models to understand complicated data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield superior predictions. Databases can be more in both depth and breadth:

More columns. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are rejected because they seem unimportant may carry information about unknown form of datas.

More rows. Larger samples yield lower estimation errors and variance, and allow users to make interpretation about small but important segments of a population.

1.3 APPLICATIONS OF DATA MINING

1.3.1 POLITICS

Data mining has been cited as the method by which the U.S. Army unit Able Danger had recognized the September 11, 2001 attacks head, Mohamed Atta, and three other 9/11 hijackers likely members from an Al Qaeda cell operating in the U.S. more than a year beforehand the attack.

It has been suggested that both the Central Intelligence Agency and their Canadian counterparts, Canadian Security Intelligence Service, have put this method of interpreting data to work for them as well, although they have not said how.

1.3.2 GAMES

Since the early 1960s, with the accessibility of oracles for certain combinatorial games, also called table bases with any beginning formation, small-board dots-and-boxes, small-board-hex, and assured endgames in chess, dots and boxes, a new area for data mining has been opened up. This is the extraction of human working strategies from these oracles. Current pattern detection approaches do not seem to fully have the required high level of abstraction in order to be applied successfully. Instead, wide experimentation with the table bases combined with an intensive study of table base-answers to well designed problems and with knowledge of prior art i.e. pre-table base information is used to yield insightful patterns[5].

1.3.3 CRM (Customer Relationship Management)

Data mining is a necessary part of CRM (Customer Relationship Management) systems. CRM has many goals, which does not just know clients, but understanding them, it will have to deal with customer analysis. A CRM deal with the profit of the company is obvious: spending less cost on marketing operations, receiving more responses and clients.

1.3.4 BUSINESS LOGIC

The main object for Data Mining in business is improving the business logic. Inspection of high-value opportunities is a key to future success in potential deals. For Example, Some opportunities lead to expenses, while others turn to a huge profit. How can distinguish a profitable one? The best way is to analyze the data gathered. This data describes business objectives in quantifiable ways. It is better if the data contains manageable values ("Increase profit" is less manageable than "reduce the cost of shipment of the TV"). Subsequently select the factors that impact the objective (How can you tell if someone would be interested in a given product?). In addition, the last step- processes the analysis and obtains the answers.

1.3.5 BALANCING EXPLORATION & EXPLOITATION

There is always the trade off between exploration (learning more and gathering more facts) and exploitation (taking immediate advantage of everything that is currently known). This theme of exploration versus exploitation is echoed also at the level of collecting data in a targeted marketing system: from a limited population of prospects/customers to choose from how many to you sacrifice to exploration (trying out new promotions or messages at random) versus optimizing what you already know.

There was for instance no reasonable way that Barnes and Noble bookstores could in 1995 look at past sales figures and foresee the impact that Amazon books and others would have based on the internet sales model. Compared to historic sales and marketing data the event of the internet could not be predicted based on the data alone. Instead perhaps data mining could have been used to detect trends of decreased sales to certain customer sub-populations - such as to those involved in the high tech industry that were the first to begin to buy books online at Amazon. So caveat emptor uses the data mining tools well but strike while the iron is hot. The performances of predictive model provided by data mining tools have a limited half life of decay. Unlike a good bottle of wine they do not increase in value with over a period of time.

II. RELATED WORKS

Road traffic accidents are complicated to analyze as they cross the boundaries of engineering, geography, and human behavior. It is not desirable to reduce the dimensionality of data through aggregation as statistical trends can be covered, or even reversed, a problem formally known as Simpson’s paradox. It is also possible that attempting to join too many disparate datasets may introduce errors.

Data mining can significantly help improving traffic safety, and has been used in many traffic related studies. In this study, a time-stamped database of Slovenian traffic accidents was analyzed by means of selected data visualization methods, an approach to short time sequence clustering, as well as clustering and spatial cluster visualization with GIS tools and the Google Earth facility. A first step in understanding and mining the data for relevant information is a good choice of data visualization techniques, graphs and diagrams. In traffic data, traffic density patterns on hourly, weekly, and monthly scales can be achieved from density plots. Such plots recognize traffic peaks, and can be of help to traffic specialists in planning routes and safety measures, as well as to different drivers.

Various trends, for example in the total number of accidents and in the number of accidents with serious injuries, presented on graphs, allow an overall analysis of traffic safety. Temporal analysis of the traffic accidents is performed using methods of short time series analysis. A time series is a very common type of data, and there are many available algorithms and methods for time series analysis. In this work, time series clustering was used to identify groups of similar time series obtained for all Slovenian municipalities. We analyze two types of time series: the number of accidents by month and the number of accidents over the 11 years included in the database. Both types of time series measured are of the type referred to as short time series, where a different approach to clustering is required than the approaches used for clustering of long time series. Short time series clustering has newly become popular due to its practical applications in biology and economics. An approach to qualitative clustering of short time series of traffic accidents is one of the contributions of this paper. The analysis of spatial aspects of the traffic accident database was performed by spatial clustering and cluster visualization in GIS and GoogleEarth. GoogleEarth4 is a program, freely obtainable for users on the internet, whose main functionality is watching satellite images of the whole Earth surface. In addition it can display peripheral layers of data on satellite images. GoogleEarth picturing of traffic accidents cluster centroids is another methodological contribution of this paper.

2.1 ROAD ACCIDENT DATA

In this research two traffic accident datasets were used: (a) the database of Slovenian traffic accidents, with data on accidents in Slovenia between the years 1995 and 2005, which is in text file format and is openly available on the internet5, and (b) the database obtained directly from the Slovenian police, which include also the data for year 2006. The openly available data has no geographic attributes; this data was used for short time series clustering. On the other hand, the police database of traffic accidents includes also geographic locations of accidents; this data was used for spatial clustering.

2.2 METHODOLOGIES

A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Hence, machine knowledge is closely related to fields such as artificial cleverness, adaptive control, statistics, data mining, pattern finding, probability theory and theoretical computer.

A .Naive Bayesian Classifier

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

It is fast, highly accessible model building and scoring Scales linearly with the number of predictors and rows Build process for Naive Bayes is parallelized Induced classifiers are easy to interpret and robust to irrelevant attributes Uses evidence from many attributes, the Naïve Bayes can be used for both binary and multiclass classification problems.

B.Decision Tree Classifier

J48 is a simple C4.5 decision tree, it produces a binary tree. C4.5 builds decision trees from training data set which is like an ID3, by the concept of information entropy Algorithm Check for base cases For each attribute „a“ find the normalized information gain from splitting on „a Let a best be the attribute with the highest normalized information gain Create a decision node that splits on a best Recourse on the sub lists obtained by splitting on a greatest, and add those nodes as children of node

Advantages

- Gains a balance of flexibility and correctness
- Limits the number of likely decision points
- It had a higher accuracy

C.AdaBoostM1 Classifier

Adaptive Boosting is a meta-algorithm in the sense that it improves or boosts an existing weak classifier. Given a weak classifier (error close to 0.5), AdaBoostM1 algorithm improves the performance of the classifier so that there are fewer classification errors.

Algorithm

All instances are equally weighted A learning algorithm is applied The weight of incorrectly classified example is increased and correctly decreased The algorithm concentrates on incorrectly classified “hard” instances Some “had” instances become “harder” some “softer” A series of diverse experts are generated based on the reweighed data.

Advantages

Simple and trained on whole (weighted) training data Over-fitting (small subsets of training data) protection Claim that boosting “never over-fits” could not be maintained.

III. EXPERIMENTS

This study used data which is produced by the Transport department of government of Hong Kong. This datasets are intended to be a nationally representative probability sample from the annual estimated 6.4 million accident reports in the Hong Kong. The dataset for the study contains traffic accident records of 2008, a total number of 34,575 cases. According to the variable meanings for dataset, this dataset has drivers records only and does not include passengers information. It includes labels, which are listed in the table 1.

To predict accident severity, various classification models were built using decision tree, naive Bayes, and K-nearest neighbor classifiers. Decision trees are easy to build and understand, can manage both continuous and categorical variables, and can perform classification as well as regression. They automatically handle interactions between variables and identify important variables. After assessing the data and selecting the predictive models to be used, a series of experiments were achieved. Extensive data pre-processing resulted in a clean dataset containing 18,288 accidents with no missing values. The class label (‘Accident Severity’) had four nominal values: ‘Fatal,’ ‘Severe injury,’ ‘Slight injury,’ or ‘Property loss.’ During data exploration, different numbers of attributes were selected by different feature selection techniques. Since WEKA’s explorer generally chooses reasonable defaults, the J48 decision tree algorithm was achieved using its default parameters: a confidence interval of 0.25, pruning allowed, and a minimum number of objects for a leaf of 3. Training and testing were done using ten-fold cross-validation.

I.Table:Variable definitions used in data set

Variable	Description
Casualty	A Person killed or injured in an accident ,there may be more than one casualty
Fatal accident	In Traffic Accident one or more persons dies within 30 days of the accident
Serious accident	In Traffic accident, one or more persons injured and detained in hospitals or detained for not more than twelve hours
Slight accident	In Traffic accident, all persons involved either not detained in hospitals or detained for not more than twelve hours
Killed casualty	Sustained injury-causing death within 30 days of the accident

Figure 1 shows the Casualty dataset graph. In this, The Casualty Age attribute takes 30.86% for Naive Bayes, 31.48% for J48, 27.38% for AdaBoostM1, 31.34% for PART and Random Forest classifier yields 32.89%. The Casualty Sex attribute takes 72.24% for Naive Bayes, 73.01% for J48, 71.68% for AdaBoostM1, 73.69% for PART and Random Forest classifier takes 74.89%. The Role of Casualty attribute takes 65.07% for Naive Bayes, 65.50% for J48, 64.28% for AdaBoostM1, 66.55% for PART and Random Forest classifier takes 67.62%. The Location of Injury attribute takes 36.03% for Naive Bayes, 37.01% for J48, 35.21% for AdaBoostM1, 37.75% for PART and Random Forest classifier yields 39.86%.

Among these Random Forest classification algorithm took highest percentage when compared with other classification algorithms. Finally, it gives the result that the overall Random Forest outperforms other algorithms in Casualty dataset.

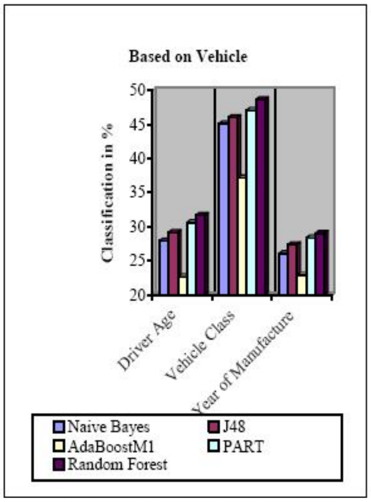

Figure1:Comparison of Naïve Bayes, J48, AdaBoostM1, PART and Random Forest classifiers.

IV. RESULTS

All these classifiers performed similarly well with respect to the number of correctly classified instances The priors on the Property Loss class was approximately 75%, Slight Injury occurred approximately 10% of the time, Severe Injury occurred 8% of the time, and very few accidents were Fatal. Compared our prior (on Property Loss), we perform better than without having a model with respect to accuracy.

However, accuracy alone does not completely describe the prediction efficiency, and other means of assessing our predictive models are necessary. The receiver operating characteristics (ROC) curve, also known as the relative operating characteristic curve, is a evaluation of two operating characteristics as the criterion changes. It can be represented by plotting the fraction of true positives (TPR = true positive rate) versus the fraction of false positives (FPR = false positive rate). An ROC analysis delivers tools to select possibly optimal models and to discard suboptimal ones independent from (and prior to specifying) the cost context or class distribution.

The ROC analysis is directly and naturally related to the cost/benefit analysis of diagnostic decision making. The area under the ROC curve (AUC) quantifies the overall discriminative ability of a test. An entirely random test (i.e., no better at identifying true positives than flipping a coin) has an AUC of 0.5, while a perfect test (i.e., one with zero false positives or negatives) has an AUC of 1.00. Since the accuracies of the above models were almost identical, we used ROC curves to further evaluate our models, using 20% (3,657) of the instance data. Some of the visualizations of the threshold curves are presented below, followed by a summary of the AUCs for each class value of the target class for each model.

V. CONCLUSION

The aim of this paper is to detect the causes of accidents. The dataset for the study contains traffic accident records of the year 2008 produced by the transport department of government of Hong Kong and investigates the performance of Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers for predicting classification accuracy. The classification accuracy on the test result reveals for the following three cases such as accident, vehicle and casualty. Literature review revealed a gap in published studies on the relationship between road characteristics and RTA harshness in Ethiopia. In this paper, we collected and cleaned traffic accident data, tried to construct novel qualities, and tested a number of predictive models. The outputs of the models were offered for analysis to domain experts for feedback. The RTA is keen to continue the study to identify areas of interest that should be given resources for traffic safety.

REFERENCES

[1] Beshah, T. and S. Hill. Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia. 2010.

[2] Lavrac, N., et al., Mining spatio-temporal data of traffic accidents and spatial pattern visualization. Metodoloski zveski, 2008. 5(1): p. 45-63.

[3] Tran, T.N., R. Wehrens, and L.M.C. Buydens, KNN-kernel density-based clustering for high-dimensional multivariate data. Computational Statistics & Data Analysis, 2006. 51: p. 513-525.

[4] Inselberg, A. and B. Dimsdale. Parallel coordinates: A tool for visualizing multi-dimensional geometry. in IEEE Visualization. 1990.

[5] Fukunaga, K. and L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition. Information Theory, IEEE Transactions on, 1975. 21(1): p. 32-40.

[6] Das, S., M. Lazarewicz, and L.H. Finkel. Principal Component Analysis of Temporal and Spatial Information for Human Gait Recognition. in The 26th Annual International Conference of IEEE EMBS. 2004. San Francisco, CA, USA: IEEE.

[7] Guo, D., et al., Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. Cartographic and Geographic Information Science, 2005. 32(2): p. 113-132.

[8] Skupin, A., The world of geography: Visualizing a knowledge domain with cartographic means. 2004, PNAS. p. 5274-5278.

[9] Gorricha, J. and V. Lobo, Improvements on the visualization of clusters in geo-referenced data using Self-Organizing Maps. Computers and Geosciences, 2012. 43: p. 177-186.

[10] Rousseeuw, P.J., Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987. 20(C): p. 53-65.