ijournals International Journal of Software & Hardware Research in Engineering
ISSN No: 2347-4890 Volume 2 Issue 3, March 2014

Efficient Analysis of Traffic Accident Using Mining Techniques

S.Vigneswaran[1]; A.Arun Joseph[2]; E.Rajamanickam[3]

Assistant Professor in Computer Science [1, 2, 3]

K.S.Rangasamy College of Arts & Science, Tiruchengodu, Namakkal District, Tamilnadu

ABSTRACT

Data Mining is the process of extracting patterns from data. Machine Learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to learn based on data, such as from sensor data or databases. A major focus of machine learning research is automatically learn to recognize complex patterns and make intelligent decisions based on data. Engineers and researchers in the automobile industry have tried to design and build safer automobiles, but traffic accidents are unavoidable. Patterns involved in dangerous crashes could be detected if we develop a prediction model that automatically classifies the type of injury severity of various traffic accidents. These behavioral and roadway patterns are useful in the development of traffic safety control policy. We believe that to obtain the greatest possible accident reduction effects with limited budgetary resources, it is important that measures be based on scientific and objective surveys of the causes of accidents and severity of injuries. This paper deals about some classification models to predict the severity of injury that occurred during traffic accidents using two machine-learning approaches. We compared Naïve Bayesian classifier and J48 decision tree Classifier for classifying the type of injury severity of various traffic accidents and the result shows that J48 outperforms Naïve Bayesian.

Keywords

Data Mining, J48 decision tree Classifier, Machine Learning, Naïve Bayesian Classifier, Prediction.

1. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage (0.9 probability)"[1].A traffic collision is when a road vehicle collides with another vehicle, pedestrian, animal, or geographical or architectural obstacle .Traffic collisions [16] can result in injury, property damage, and death.

Road accidents have been the major cause of injuries and fatalities in worldwide for the last few decades. Analyzing, interpreting and making maximum use of the data is difficult and resource demanding due to the exponential growth of many business, governmental and scientific databases. It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100%. This fact shows that we are getting more and more exploded by data/information and yet ravenous for knowledge. Data mining therefore appears as a useful tool to address the need for sifting useful information such as hidden patterns from databases [19].

Traffic control system is one of the various areas, where critical data about the well-being of the society is recorded and kept. Various aspects of a traffic system like vehicle accidents, traffic volumes and concentration are recorded at different levels. In connection to this, injury severities resulted from road traffic accident are one of the areas of concern.

Thus, through this work an attempt has been made to apply data mining tools and techniques in analyzing and determining interesting patterns especially with respect to injury severity, on road accidents data at Addis Ababa Region Traffic Control System. In order to plan and implement effective strategies in reducing the severity of the injury and vehicle accident at large in Ethiopia, there is a need for actionable information, which is obviously a result of a research work. So, in the effort of alleviating the current problem of vehicle accidents, identifying factors leading to accidents through developing a capacity to design and implement an effective traffic information system that can provide timely and accurate traffic information is very crucial. Timely and reliable data collected about vehicle accidents can be used to identify major determinants and risk factors for vehicle accidents, severe injury and fatalities and to take preventive measures so that the effort of improving the quality of life will be enhanced [20].

All the previous researches were conduct by using small proportion of the accumulated data. Besides, in those researches data analysis was conducted by using simple

statistical methods. Since the analysis made by using traditional methods focus on problems with much more manageable number of variables and cases than may be encountered in real world, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional databases . The absence of significant attempt that has been made so far to this level in identifying the major determinants of car accidents and establishing the most important factors influencing the severity of an injury in Addis Ababa region justify the importance of this research. This research work will be groundwork for the effort of reducing vehicle accident in particular and improving the quality of life in general. Moreover, it will also be an input for researches in the same area. The costs of fatalities and injuries due to traffic accidents have a great impact on the society. In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents zero. There are several approaches, which that researchers have employed to study this problem. These include neural network, nesting logic formulation, log-linear model, fuzzy ART maps and so on.

Applying data mining techniques to model traffic accident records data can help to understand the characteristics of drivers' behavior, roadway condition and weather condition that were causally connected with different injury severities. This can help decision makers to formulate better traffic safety control policies.

2. OBJECTIVE OF THIS WORK

The costs of fatalities and injuries due to traffic accidents have a great impact on society. In recent years, researchers have paid increasing attention at determining the factors that significantly affect driver injury severity in traffic accidents. The general objective of the research was to investigate the potential applicability of data mining technology in developing a model that can support road traffic accident severity analysis in the effort of preventing and controlling vehicle accident. Road traffic accidents are among the top leading causes of deaths and injuries of various levels.

Analyzing, interpreting and making maximum use of the data is difficult and resource demanding due to the exponential growth of many business, governmental and scientific databases. It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100%. This fact shows that we are getting more and more exploded by data/information and yet ravenous for knowledge.

Data mining therefore appears as a useful tool to address the need for sifting useful information such as hidden patterns from databases. In today's world, where the accumulation of data is increasing in an alarming rate, understanding interesting patterns of data is an important issue to be considered to adjust strategies, to make maximum use of it, and find new opportunities. Organizations keeping data on their domain area takes every record as an opportunity in learning facts. Nevertheless, the simple gathering of data is not enough to get maximum knowledge out of it. Thus, for an effective learning, data from many sources must first be gathered and organized in a consistent and useful manner.

In recent years, researchers have paid increasing attention at determining the factors that significantly affect driver injury severity in traffic accidents. To obtain the greatest possible accident reduction effects with limited budgetary resources, it is important that measures be based on scientific and objective surveys of the causes of accidents and severity of injuries.

This proposed work investigates application of Naive Bayes and j48 and compares both the algorithms performance based on injury severity. According to the variable definitions for the Transport department of government of Hong Kong's traffic accident records of 2008 dataset, this dataset has drivers' only records and does not include passengers' information. It includes labels of severity, district council district, hit and run, weather, rain, natural light, junction control, road classification, vehicle movements, type of collision, number of vehicles involved, number of causalities injured, casualty age, casualty sex, location of injury, degree of injury, role of casualty, pedestrian action, vehicle class of driver or passenger casualty, driver Age, driver sex, year of manufacture, severity of accident and vehicle class. The injury severity has three classes: Based on Accident, Based on Vehicle and based on Casualty. In the original dataset, 70.18% of the cases have output of no injury, 16.07% of the cases have output of possible injury, 9.48% of the cases have output of non-incapacitating injury, 4.02% of the cases have output of incapacitating injury, and 0.25% of the cases have fatal injury [20].

Our task was to develop machine learning based intelligent models that could classify the severity of injuries (5 categories) more accurately. This can in turn lead to greater understanding of the relationship between the factors of driver, vehicle, roadway, and environment and driver injury severity. Accurate results of such data analysis can provide crucial information for the road accident prevention policy. The analysis focused on vehicle accidents that occurred at signalized intersections. The injury severity was dividing into three classes: Based on Accident, Based on Vehicle and based on Casualty.

3. RELATED WORKS

Ossenbruggen [Ossenbruggen et al., 2001][2] used a logistic regression model to identify statistically significant factors that predict the probabilities of crashes and injury crashes aiming at using these models to perform a risk assessment of a given region. Miaou [Miaou and Harry, 1993 [3] studied 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. Roadway and truck accident data from the Highway Safety Information System (HSIS) have been employed to illustrate the use and the limitations of these models.

In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents zero. There are several approaches, which researchers have employed to study this problem. This paper investigates application of Naïve Bayesian algorithm and J48 classifier to find the efficiency of these algorithms based on the several factors involved in traffic accident.

This result could predict injury severity. The performance analysis is based on the labels of year, month, region, primary sampling unit, the number describing the police jurisdiction, case number, person number, vehicle number, vehicle make and model; inputs of drivers' age, gender,

alcohol usage, restraint system, eject, vehicle body type, vehicle age, vehicle role, initial point of impact, manner of collision, rollover, roadway surface condition, light condition, travel speed, speed limit and the output injury severity.

Abdel-Aty[11] used the Fatality Analysis Reporting System (FARS) crash databases covering the period of 1975-2000 to analyze the effect of the increasing number of Light Truck Vehicle (LTV) registrations on fatal angle collision trends in the US [Abdel-Aty and Abdelwahab, 2003]. They investigated the number of annual fatalities that resulted from angle collisions as well as collision configuration (car-car, car-LTV, LTV-car, and LTV-LTV).

Bedard et al.,[12] applied a multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers' fatality risk. They found that increasing seatbelt use, reducing speed, and reducing the number and severity of driver-side impacts might prevent fatalities.

Evanco[13] conducted a multivariate population-based statistical analysis to determine the relationship between fatalities and accident notification times. The analysis demonstrated that accident notification time is an important determinant of the number of fatalities for accidents on rural roadways.

Some researchers studied the relationship between drivers' age, gender, vehicle mass, impact speed or driving speed measure with fatalities [14, 15]

4. CLASSIFICATION MODEL DESCRIPTION

A major focus of machine learning [9] [10] research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Hence, machine learning is closely related to fields such as statistics, probability theory, data mining, pattern recognition, artificial intelligence, adaptive control, and theoretical computer science.

Naive Bayesian Classifier

A Naive Bayesian classifier [7] is a simple probabilistic classifier based on applying Bayesian' theorem (from Bayesian statistics) with strong (naive) independence assumptions. Using Bayesian' theorem, we write

$$p(C \mid F1 \ldots Fn) = \frac{p(C)\,p(F1 \ldots . Fn \mid C)}{p(F1 \ldots . Fn)}$$

Advantages

- Fast, highly scalable model building and scoring.
- It scales linearly with the number of predictors and rows.
- The build process for Naive Bayesian is parallelized.
- Induced classifiers are easy to interpret.
- Robust to irrelevant attributes.
- Uses evidence from many attributes.
- Naive Bayesian can be used for both binary and multi-class classification problems.

4.1 J48 Decision Trees

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable. The J48 Decision tree classifier follows the following simple algorithm [4]. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly.

J48 Classifier is a simple C4.5 decision tree for classification. It creates a binary tree. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy [17].

4.2 A Genetic Algorithm for a Feature Selection Problem

The first phase of the algorithm deals with isolating the very few relevant features from the large set. This is not exactly the classical feature selection problem known in Data mining as, for example, around 50% of features are selected. Here, have the idea that less than 5% of the features have to be selected. Nevertheless, this problem is close from the classical feature selection problem, and we will use a genetic algorithm as saw they are well adapted for problems with a large number of features. Present here the main characteristics and adaptations made to deal with this particular feature selection problem. Genetic Algorithm has different phases. It proceeds for a fixed number of generations. A chromosome here is a string of bits whose size corresponds to the number of features. A 0 or 1, at position i, indicates whether the feature i is selected (1) or not (0).

4.3 The genetic operators

These operators allow Genetic Algorithms [18]to explore the search space. However, operators typically have destructive as well as constructive effects. They must be adapted to the problem.

- **Crossover**

Use a Subset Size-Oriented Common Feature Crossover Operator (SSOCF), which keeps useful informative blocks and produces offspring, which have the same distribution than the parents. Offspring are kept, only if they fit better than the least good individual of the population.

Features shared by the two parents that are kept by offspring and the non-shared features are inherited by offspring. That is corresponding to the i^{th} parent with the probability $(n_i - n_c/n_u)$ where ni is the number of selected features of the i^{th} parent, n_c is the number of commonly selected features across both mating partners and n_u is the number of non-shared selected features.

- **Mutation**

The mutation is an operator that allows diversity. During the mutation stage, a chromosome has a probability pmut to mutate. If a chromosome is selected to mutate, we choose

randomly a number n of bits to be flipped then n bits are chosen randomly and flipped. In order to create a large diversity, we set pmut around 10% and n [1, 5].

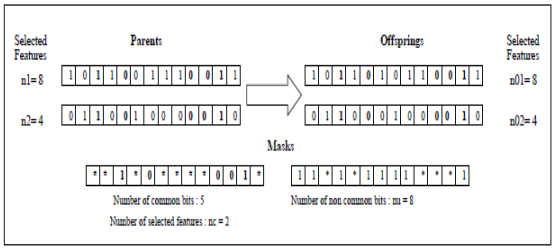

Figure 1 Genetic Algorithm for Crossover

- **Selection**

Implement a probabilistic binary tournament selection. Tournament selection holds n tournaments to choose n individuals. Each tournament consists of sampling two elements of the population and choosing the best one with a probability p [0.5, 1].

5. DATA SET COLLECTION

For This study used data produced by the Transport department of government of Hong Kong [8]. 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 definitions for dataset, this dataset has drivers' records only and does not include passengers' information. It includes labels.

Figure2 Highway of Hong Kong

Data Preparation

The variables are already categorized and represented by numbers. The manner in which the collision occurred has three categories:

Based on the Accident:

The attributes used are severity, district council district, hit and run, weather, rain, natural light, junction control, road classification, vehicle movements, type of collision, number of vehicles involved and no of causalities injured.

Based on Vehicle:

The attributes used are Driver Age, Drive Sex, Year of manufacture, Severity of accident and vehicle class.

Based on casualty:

The attributes used are casualty Age, Casualty sex, Degree of injury, role of casualty, location of casualty, pedestrian action and vehicle class of driver or passenger casualty.

Table 1. Variable Definitions used in data set captions

Casualty	A person killed or injured in an accident in which there may be more than one casualty.
Fatal accident	Traffic accident in which one or more persons dies within 30 days of the accident.
Serious accident	Traffic accident in which one of more persons injured and detained in hospital for more than twelve hours.
Slight accident	Traffic accident in which all persons involved are 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.
Serious injury	An injury for which a person is detained in hospital as an 'in-patient' for more than twelve hours. Injuries causing death 30 or more days after the accident are also included in this category.
Slight injury	An injury of a minor character such as a sprain, bruise or cut not judged to be severe, or slight shock requiring roadside attention and detention in hospital is less than 12 hours, or not required.
Road users	Pedestrians and vehicle users, which include all occupants (i.e. driver or rider and passengers, including persons injured while boarding or alighting from the vehicle).
Vehicles involved	Vehicles whose drivers or passengers are injured, which hit a pedestrian, or another vehicle whose driver or passengers are injured, or which contributes to the accident.

6. Open source software package Weka tool kit

Weka Knowledge Explorer

The Weka Knowledge Explorer [6] is an easy to use graphical user interface that harnesses the power of the weka software. Each of the major weka packages Filters, Classifiers, Clusters, Associations, and Attribute Selection is represented in the Explorer along with a Visualization tool, which allows datasets and the predictions of Classifiers and

Clusters to be visualized in two dimensions.

The Weka workbench[1] contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modelling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains,[2][3] but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. The main strengths of Weka are that it is freely available under the GNU General Public License, very portable because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform, contains a comprehensive collection of data preprocessing and modeling techniques, and is easy to use by a novice due to the graphical user interfaces it contains.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka[4]. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The Explorer interface has several panels that give access to the main components of the workbench. The Preprocess panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria. The Classify panel enables the user to apply classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree). The Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data. The Cluster panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions. The next panel, Select attributes provides algorithms for identifying the most predictive attributes in a dataset. The last panel, Visualize, shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators

The Weka Data Format (ARFF)

Datasets for WEKA should be formatted according to the ARFF format. (However, there are several converters included in WEKA that can convert other file formats to ARFF. The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file.) Examples of ARFF files can be found in the "data" subdirectory.

7. Experimental Results

This work deals with performance of two classification algorithms namely Naive Bayesian & J48 classifiers. The Transport Department of Government of Hon Kong produces Dataset for the year 2008 [5] is used in this work.

The dataset is recorded into three different scenarios;

(a) Based On Accident Information

(b) Based On Casualty Information

(c) Based On Vehicle Information

Totally, this dataset consist of 34,575 record sets. Among them 14576 belongs to accident, 10,000 belongs to vehicle and remaining 9,999 belongs to casualty.

(a) Based On Accident

The total record set used is 14,576. The attributes involved in this case are severity, district council district, hit and run, weather, rain, natural light, junction control, road classification, vehicle movements, type of collision, number of vehicles involved and number of causalities injured.

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, Severity is 12,343 records (84.68%) and incorrectly classified are 2,233 records (15.32%). The J48 classifier can able to correctly classified 12,337 records (84.64%) and incorrectly classified are 2,239 records (15.36%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, District Council District is 2,041 records (14%) and incorrectly classified are 12,535 records (86%). The J48 classifier can able to correctly classified 3,042 records (20.87%) and incorrectly classified are 11,534 records (79.13%).

Out of 14,576 records the Naive Bayesian classifier can able to correctly classified the attribute, Hit and Run is 14,410 records (98.86%) and incorrectly classified are 166 records (1.14%). The J48 classifier can able to correctly classified 14,417 records (98.91%) and incorrectly classified are 159 records (1.09%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, Weather is 13,197 records (90.54%) and incorrectly classified are 1,379 records (9.46%). The J48 classifier can able to correctly classified 13,321 records (91.39%) and incorrectly classified are 1,255 records (8.61%).

Out of 14,576 records, the Naive Bayesian classifier can able

to correctly classified the attribute, Rain is 13,139 records (90.14%) and incorrectly classified are 1,437 records (9.86%). The J48 classifier can able to correctly classified 13,144 records (90.18 %) and incorrectly classified are 1,432 records (9.82%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, Natural Light is 9,478 records (65.02%) and incorrectly classified are 5,098 records (34.98%). The J48 classifier can able to correctly classified 9,589 records (65.78%) and incorrectly classified are 4,987 records (34.22%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, Junction Control is 10,804 records (74.12%) and incorrectly classified are 3,772 records (25.88%). The J48 classifier can able to correctly classified 10,953 records (75.14%) and incorrectly classified are 3,623 records (24.86%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, Road Classification is 14,482 records (99.36%) and incorrectly classified are 94 records (0.64%). The J48 classifier can able to correctly classified 14,491 records (99.42%) and incorrectly classified are 85 records (0.58%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, Vehicle Movements is 12,261 records (84.12%) and incorrectly classified are 2,315 records (15.88%). The J48 classifier can able to correctly classified 12,357 records (84.78%) and incorrectly classified are 2,219 records (15.22%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, Type of Collision is 9,975 records (68.43%) and incorrectly classified are 4,601 records (31.57%). The J48 classifier can able to correctly classified 10,329 records (70.86%) and incorrectly classified are 4,247 records (29.14%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classified the attribute, Number of Vehicles Involved is 14,115 records (96.84%) and incorrectly classified are 461 records (3.16%). The J48 classifier can able to correctly classified 14,117 records (96.85%) and incorrectly classified are 459 records (3.15%).

Out of 14,576 records, the Naive Bayesian classifier can able to correctly classify the attribute, Number of Casualties Injured is 11,605 records (79.62%) and incorrectly classified are 2,971 records (20.38%). The J48 classifier can able to correctly classified 12,378 records (84.92%) and incorrectly classified are 2,198 records (15.08%).

Applying Genetic Algorithm for Feature Selection in Accident Dataset

From the above dataset, not all the twelve attributes are involved in classification. Using Genetic Algorithm, it scrutinizes the potential attributes, which leads to better classification.

The attributes that are insignificant for classification are as follows: Severity, Hit and Run, Weather, Rain, Natural Light, Junction Control, Road Classification and Number of Vehicles Involved.

Severity

For J48 the correctly classified percentage is 84.64 and for Naive Bayesian 84.68. From this, it concludes that there is no significant difference between them.

Hit and Run

For J48 the correctly classified percentage is 98.91 and for Naive Bayesian 98.86. From this, it concludes that there is no significant difference between them.

Weather

For J48 the correctly classified percentage is 91.39 and for Naive Bayesian 90.54. From this, it concludes that there is no significant difference between them.

Rain

For J48 the correctly classified percentage is 90.18 and for Naive Bayesian 90.14. From this, it concludes that there is no significant difference between them.

Natural Light

For J48 the correctly classified percentage is 65.78 and for Naive Bayesian 65.02. From this, it concludes that there is no significant difference between them.

Junction Control

For J48 the correctly classified percentage is 75.14 and for Naive Bayesian 74.12. From this, it concludes that there is no significant difference between them.

Road Classification

For J48 the correctly classified percentage is 99.42 and for Naive Bayesian 99.36. From this, it concludes that there is no significant difference between them.

Number of Vehicles involved

For J48 the correctly classified percentage is 96.85 and for Naive Bayesian 96.84. From this, it concludes that there is no significant difference between them.

Therefore, the Genetic Algorithm eliminates some attributes that are not potential for classification.

Finally overall records used for accident is 14,576 and the attributes are Vehicle Movement, Type of Collision, No of Casualties injured and District Council District. The correctly classified Naive Bayesian classifier and J48 classifier percentages of the attributes are listed in the below table 2

Table 2 Accident Dataset Classification

Classifiers	Vehicle Movement	Type of Collision	No of Casualties injured	District Council District
Naïve Bayesian	84.12	68.43	79.62	14
J48	84.78	70.86	84.92	20.87

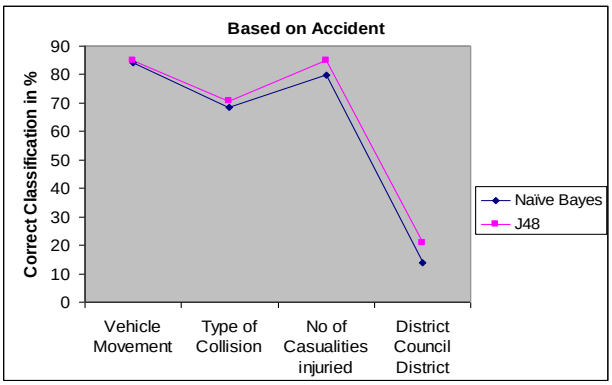

Figure 3 Comparison of Naive Bayesian and J48 classifiers Based on Accident Dataset

The figure 3 shows the Accident dataset graph. In this, the vehicle movements attribute takes 84.12% for Naive Bayesian classifier and J48 takes 84.78%. The type of collision attribute takes 68.43% for Naive Bayesian classifier and J48 takes 70.86%. The number of casualty-injured attribute takes 79.62% for Naive Bayesian classifier and J48 takes 84.92%. The district council district attribute takes 14% for Naive Bayesian classifier and J48 takes 20.87%. Among these J48 classification algorithm took highest percentage when compared with Naive Bayesian classification algorithm. Finally, it gives the result that the overall J48 outperforms Naive Bayesian in Accident dataset.

(b)Based on Casualty

The total record set used is 9,999. The attributes involved in this case are Casualty Age, Casualty Sex, Location of Injury, Degree of Injury, Role of Casualty, Pedestrian Action and Vehicle Class of Driver or Passenger Casualty.

Out of 9,999 records, the Naive Bayesian classifier can able to correctly classified the attribute, Casualty Age is 2,650 records (26.50%) and incorrectly classified are 7,349 records (73.50%). The J48 classifier can able to correctly classified 2,954 records (29.54%) and incorrectly classified are 7,045 records (70.46%).

Out of 9,999 records, the Naive Bayesian classifier can able to correctly classified the attribute, Casualty Sex is 7,025 records (70.26%) and incorrectly classified are 2,974 records (29.74%). The J48 classifier can able to correctly classified 7,186 records (71.87%) and incorrectly classified are 2,813 records (28.13%).

Out of 9,999 records, the Naive Bayesian classifier can able to correctly classified the attribute, Location of Injury is 3,023 records (30.23%) and incorrectly classified are 6,976 records (69.77%). The J48 classifier can able to correctly classified 3,150 records (31.50%) and incorrectly classified are 6,849 records (68.50%).

Out of 9,999 records, the Naive Bayesian classifier can able to correctly classified the attribute, Degree of Injury is 8,776 records (87.77%) and incorrectly classified are 1,223 records (12.23%). The J48 classifier can able to correctly classified 8,768 records (87.69%) and incorrectly classified are 1,231 records (12.31%).

Out of 9,999 records the Naive Bayesian classifier can able to correctly classified the attribute, Role of Casualty is 6,535 records (65.36%) and incorrectly classified are 3,464 records (34.64%). The J48 classifier can able to correctly classified 6,598 records (65.99%) and incorrectly classified are 3,401 records (34.01%).

Applying Genetic Algorithm for Feature Selection in Casualty Dataset

From the casualty dataset not all, the seven attributes are involved in classification. Using Genetic Algorithm, it scrutinizes the potential attributes, which leads to better classification.

The attributes, which are insignificant for classification, are as follows: Pedestrian Action and Vehicle Class of Driver or Passenger Casualty

J48 and the Naive Bayesian classifier results conclude that there is no significant difference between them. Therefore, the Genetic Algorithm eliminates these attributes, which are not potential for classification.

Finally overall records used for casualty is 9,999 and the attributes are Casualty Age, Casualty Sex, Location of Injury, Degree of Injury and Role of Casualty. The correctly classified Naive Bayesian classifier and J48 classifier percentages of the attributes are listed in the below table 3

Table 3 Casualty Dataset Classification

Classifier	Age	Sex	Location of Injury	Degree of Injury	Role of Casualty
Naïve Bayesian	26.5	70.26	30.25	87.77	65.36
J48	29.54	71.87	31.5	87.69	65.99

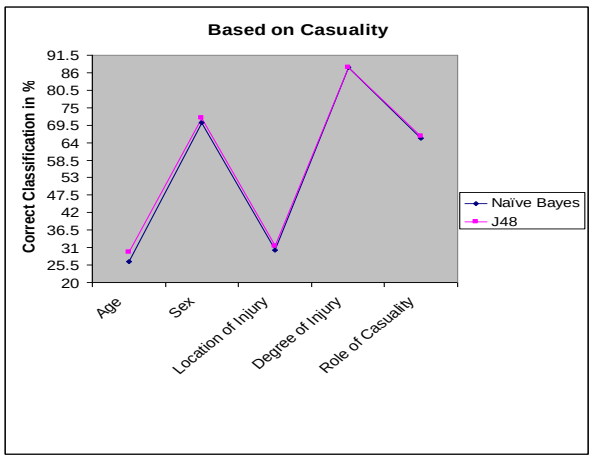

Figure 4 Comparison of Naive Bayesian and J48 classifiers Based on Casualty Dataset

The figure 4 shows the Casualty dataset graph. In this, the age attribute takes 26.5% for Naive Bayesian classifier and J48 takes 29.54%. The sex attribute takes 70.26% for Naive Bayesian classifier and J48 takes 71.87%. The location of injury attribute takes 30.25% for Naive Bayesian classifier and J48 takes 31.5%. The degree of injury attribute takes 87.77% for Naive Bayesian classifier and J48 takes 87.69%. The role of casualty attribute takes 65.36% for Naive Bayesian classifier and J48 takes 65.99%. Among these J48 classification algorithm took highest percentage when

compared with Naive Bayesian classification algorithm. Finally, it gives the result that the overall J48 outperforms Naive Bayesian in Casualty dataset.

(c)Based on Vehicle

The total record set used is 10,000. The attributes involved in this case are Driver Age, Drive Sex, Year of Manufacture, and Severity of accident and vehicle class.

Out of 10,000 records, the Naive Bayesian classifier can able to correctly classified the attribute, Driver Age is 3,653 records (36.53%) and incorrectly classified are 6,347 records (63.47%). The J48 classifier can able to correctly classified 3,823 records (38.23%) and incorrectly classified are 6,177 records (61.77%).

Out of 10,000 records, the Naive Bayesian classifier can able to correctly classified the attribute, Driver Sex is 8,726 records (87.26%) and incorrectly classified are 1,274 records (12.74%). The J48 classifier can able to correctly classified 9,055 records (90.55%) and incorrectly classified are 945 records (9.45%).

Out of 10,000 records the Naive Bayesian classifier can able to correctly classified the attribute, Year of Manufacture is 7,609 records (76.09%) and incorrectly classified are 2,391 records (23.91%). The J48 classifier can able to correctly classified 7,635 records (76.35%) and incorrectly classified are 2,365 records (23.65%).

Out of 10,000 records, the Naive Bayesian classifier can able to correctly classify the attribute, Severity of Accident is 8,630 records (86.30%) and incorrectly classified are 1,370 records (13.70%). The J48 classifier can able to correctly classified 8,630 records (86.30%) and incorrectly classified are 1,370 records (13.70%). In this both the classifiers have the same values.

Out of 10,000 records, the Naive Bayesian classifier can able to correctly classified the attribute, Vehicle Class is 4,321 records (43.21%) and incorrectly classified are 5,679 records (56.79%). The J48 classifier can able to correctly classified 4,373 records (43.73%) and incorrectly classified are 5,627 records (56.27%).

Applying Genetic Algorithm for Feature Selection in Casualty Dataset

From the casualty dataset, all five attributes are involved in classification. Using Genetic Algorithm, which is the potential, attributes that leads to better classification.

Finally overall records used for Vehicle is 10,000 and the attributes are Driver Age, Drive Sex, Vehicle Class, Severity of accident and Year of Manufacture.

The correctly classified Naive Bayesian classifier and J48 classifier percentages of the attributes are listed in the below table 4

Table 4 Vehicle Dataset Classification

	Driver Age	Driver Sex	Vehicle Class	Year of Manufacture	Severity
Naïve Bayesian	36.53	87.26	43.21	76.09	86.3
J48	38.23	90.55	43.73	76.35	86.3

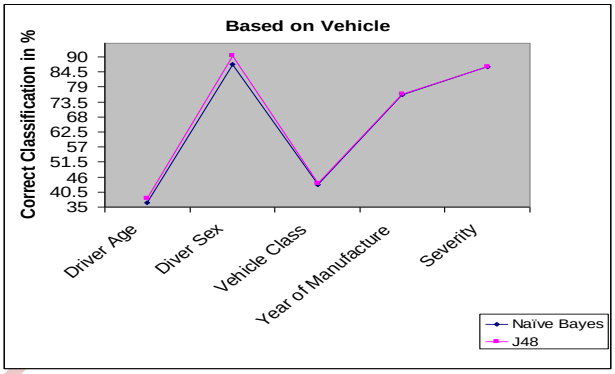

Figure 5 Comparison of Naive Bayesian and J48 classifiers Based on Vehicle Dataset

The figure 5 shows the Vechicle dataset graph. In this, the driver age attribute takes 36.53% for Naive Bayesian classifier and J48 takes 38.23%. The driver sex attribute takes 87.26% for Naive Bayesian classifier and J48 takes 90.55%. The vehicle class attribute takes 43.21% for Naive Bayesian classifier and J48 takes 43.73%. The year of manufacture attribute takes 76.09% for Naive Bayesian classifier and J48 takes 76.35%. The severity attribute takes 86.3% for Naive Bayesian classifier and J48 takes 86.3%. Among these J48 classification algorithm took highest percentage when compared with Naive Bayesian classification algorithm. Finally, it gives the result that the overall J48 outperforms Naive Bayesian in Vehicle dataset.

8. CONCLUSION AND FUTURE WORK

The aim of the thesis work 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 J48 and Naive Bayesian for predicting classification accuracy. The classification accuracy on the test result reveals for the following three cases such as accident, vehicle and casualty.

J48 outperforms than Naïve Bayesian Classification algorithm instead of selecting all the attributes for classification. Genetic Algorithm is used for feature selection to reduce the dimensionality of the dataset. In this work, we extended the research to three different cases such as Accident, Casualty and Vehicle for finding the cause of

accident and the severity of accident. In future, research work will be focused on classification of injury based on Fuzzy Logic.

9. REFERENCES

[1] Gartner Group High Performance Computing Research Note 1/31/95

[2] Ossenbruggen, P.J., Pendharkar, J. and Ivan, J. 2001, "Roadway safety in rural and small urbanized areas". *Accidents Analysis and Prevention,* 33 (4), pp. 485–498.

[3] Miaou, S.P. and Harry, L. 1993, "Modeling vehicle accidents and highway geometric design relationships". *Accidents Analysis and Prevention*, (6), pp. 689–709.27. Desktop Reference for Crash Reduction Factors Report No. FHWA-SA-07-015, Federal Highway Administration September,2007http://www.ite.org/safety/issuebriefs/Desktop%20Reference%20Complete.pdf

[4] S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007

[5] http://www.td.gov.hk/filemanager/en/content_2015/08pubdb.xls

[6] http://www.lri.fr/~pierres/donn%E9es/save/these/weka-3-4/README

[7] http://databases.about.com/od/datamining/g/Classification.htm

[8] http://www.td.gov.hk/en/road_safety/road_traffic_accident_statistics/2008/index.html

[9] Domingos, Pedro & Michael Pazzani (1997) "On the optimality of the simple Bayesian classifier under zero-one loss". Machine Learning, 29:103–137. (also online at CiteSeer: [1])

[10] Rish, Irina. (2001). "An empirical study of the naive Bayes classifier". IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. (available online: PDF, PostScript)

[11] Abdel-Aty, M., and Abdelwahab, H., Analysis and Prediction of Traffic Fatalities Resulting From Angle Collisions Including the Effect of Vehicles' Configuration and Compatibility. Accident Analysis and Prevention, 2003.

Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J. P., The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. Accident analysis and Prevention, Vol. 34, pp. 717-727, 2002.

[12] Evanco, W. M., The Potential Impact of Rural Mayday Systems on Vehicular Crash Fatalities. Accident Analysis and Prevention, Vol. 31, 1999, pp. 455-462.

[13] Kweon, Y. J., & Kockelman, D. M., Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models. Accident Analysis and Prevention, Vol. 35, 2003, pp. 441-450.

[14] Martin, P. G., Crandall, J. R., & Pilkey, W. D., Injury Trends of Passenger Car Drivers In the USA. Accident Analysis and Prevention, Vol. 32, 2000, pp. 541-557.