A Perspective Analysis of Traffic Accident using Data Mining Techniques

S.Krishnaveni
Ph.D (CS) Research Scholar,
Karpagam University,
Coimbatore, India – 641 021

Dr.M.Hemalatha
Asst. Professor & Head, Dept of Software Systems Karpagam University,
Coimbatore, India – 641 021

ABSTRACT
Data Mining is taking out of hidden patterns from huge database. It is commonly used in a marketing, surveillance, fraud detection and scientific discovery. 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 traffic accidents are the major causes of death and injuries in this world. Roadway patterns are useful in the development of traffic safety control policy. This paper deals with the some of classification models to predict the severity of injury that occurred during traffic accidents. I have compared Naive Bayes Bayesian classifier, AdaBoostM1 Meta classifier, PART Rule classifier, J48 Decision Tree classifier and Random Forest Tree classifier for classifying the type of injury severity of various traffic accidents. The final result shows that the Random Forest outperforms than other four algorithms.

Keyword
Data mining, machine learning, Naive Bayes Classifiers, AdaBoostM1, PART, J48 and Random Forest.

1. INTRODUCTION AND RELATED WORKS
Data mining is the extraction of hidden predictive information from large databases and it 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 is as follows,

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

According to a Gartner HPC Research Note, "Due to data capture, transmission and storage, large-systems users have to implement new techniques. They 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"[6].

According to the World Health Organization (WHO) [18], India is leading in the annual reported number of traffic deaths worldwide. In Global Status Report on Road Safety, the WHO revealed that India leads with 105,000 traffic deaths in a year, when compare to China with over 96,000 deaths on road. The survey was conducted in 178 countries, as per the survey 300 Indians die on roads every day. There are two million people have disabilities caused from a traffic accident. This survey is based on data collection for 2006 (Data collection began from March 2008 and completed in September 2008).

A traffic collision occurs when a road vehicle collides with another vehicle, pedestrian, animal, or geographical or architectural obstacle. It 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. 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 is a useful tool to address the need for sifting useful information such as hidden patterns from databases [7].

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.

This proposed work investigates application of Naive Bayes, AdaBoostM1, PART, J48 and Random Forest Classifier and compares these algorithms performance based on injury severity. According to the variable definition 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 [7],

- 70.18 % - output of no injury
- 16.07 % - output of possible injury
- 9.48 % - output of non-incapacitating injury
- 4.02 % - output of incapacitating injury
- 0.25% - fatal injury.

Ossenbruggen [Ossenbruggen et al., 2001][19] are 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 [16] 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.

Abdel-Aty[1] 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.,[2] 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[4] conducted a multivariate population-based statistical analysis to determine the relationship between fatalities and accident notification times. This demonstrated the accident notification time which is an important determinant of the number of fatalities for accidents on rural roadways.

Some of the researchers are studied about the relationship between driver's gender, age, vehicle mass, impact speed or driving speed measure with fatalities [15, 17].

2. CLASSIFICATION MODEL DESCRIPTION

A major focus of machine learning [3, 8] 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 artificial intelligence, adaptive control, statistics, data mining, pattern recognition, probability theory and theoretical computer.

2.1 Naive Bayesian Classifier

A Naive Bayesian classifier [21] is a simple probabilistic classifier based on applying Bayesian theorem (from Bayesian statistics) with strong (naive) independence assumptions. By the use of Bayesian theorem we can write

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

2.1.1 Advantages

- It is fast, highly scalable 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 Naive Bayes can be used for both binary and multi-class classification problems

2.2 J48 Decision Tree Classifier

J48 is a simple C4.5 decision tree, it creates a binary tree. C4.5 builds decision trees from a set of training data which is like an ID3, using the concept of information entropy [20].

2.2.1 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_best, and add those nodes as children of node

2.2.2 Advantages

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

2.3 AdaBoostM1 Classifier

Adaptive Boosting [13] 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.

2.3.1 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.

2.3.2 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.

- Complex resulting classifier can be determined reliably from limited amount of data

2.4 PART (Partial Decision Trees) Classifier

PART is a rule based algorithm [12] and produces a set of if-then rules that can be used to classify data. It is a modification of C4.5 and RIPPER algorithms and draws strategies from both. PART adopts the divide-and-conquer strategy of RIPPER and combines it with the decision tree approach of C4.5.

PART generates a set of rules according to the divide-and-conquer strategy, removes all instances from the training collection that are covered by this rule and proceeds recursively until no instance remains [5].

To generate a single rule, PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest coverage as the new rule. It is different from C4.5 because the trees built for each rules are partial, based on the remaining set of examples and not complete as in case of C4.5.

2.4.1 Advantages

- It is simpler and has been found to give sufficiently strong rules.

2.5 Random Forest Tree Classifier

A random forest [14] consisting of a collection of tree-structured classifiers (h(x,_k), k = 1, . .) where the _k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

2.5.1 Algorithm

- Choose T number of trees to grow
- Choose m number of variables used to split each node. m<<M, where M is the number of input variables, m is hold constant while growing the forest
- Grow T trees. When growing each tree do
- Construct a bootstrap sample of size n sampled from Sn with the replacement and grow a tree from this bootstrap sample
- When growing a tree at each node select m variables at random and use them to find the best split
- Grow the tree to a maximal extent and there is no pruning
- To classify point X collect votes from every tree in the forest and then use majority voting to decide on the class label

2.5.2 Advantages

- It is unexcelled in accuracy among current algorithms and it runs well on large data bases.
- It can handle thousands of input variables without variable deletion and also the learning is so fast.
- It has an effective method for estimating missing data and maintains accuracy.
- The new generated forests can be saved for future use on other data.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers or give interesting views of the data.

2.6 A Genetic Algorithm for a Feature Selection Problem

Genetic algorithm [9] is a search technique based on Darwin's evolution theory. This technique starts with choosing a set of random plans at a high-dimension space, viewed as a population of chromosomes.

2.6.1 Steps of Genetic Algorithm

- Generates chromosomes which represents a possible solution
- Define a fitness function to measure the solution
- Select parent chromosomes for the next generation (e.g. rank-based and roulette-wheel)
- Apply GA operators for crossover and mutation
- Crossovers spread the advantageous traits to improve the whole population fitness value.
- Mutations reintroduce divergence into a converging population, so the search may explore attribute-values that are not existed in the initial population.
- A new generated population replaces the previous population.

3. DATASET COLLECTION

This study used data which is produced by the Transport department of government of Hong Kong [11]. 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, which are listed in the table 1.

Table 1. 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 hospital for 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

Serious injury	An injury for which a person is detained in hospital as an 'in-patient' for more than twelve hours and the 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 are 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

3.1 Data Preparation

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

3.1.1 Based on 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.

3.1.2 Based on Vehicle

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

3.1.3 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.

4. WEKA TOOL KIT

The Weka Knowledge Explorer is an easy to use graphical user interface that harnesses the power of the Weka software [10]. The major Weka packages are 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 workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling together with graphical user interfaces for easy access to this functionality. It was primarily designed as a tool for analyzing data from agricultural domains. Now it is used in many different application areas, in particular for educational purposes and research.

The main strengths is freely available under the GNU General Public License, very portable because it is fully implemented in the Java programming language and runs on any modern computing platform, contains a comprehensive collection of data preprocessing and modeling techniques. Weka supports several standard data mining tasks like data clustering, classification, regression, preprocessing, visualization and feature selection. These techniques are predicated on the assumption that the data is available as a single flat file or relation. Each data point is described by a fixed number of attributes and an important area is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

4.1 Weka Data Format (ARFF)

In Weka datasets should be formatted to the ARFF format. The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file the Preprocess panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a filtering algorithm. These filters can be used to transform the data and make it possible to delete instances and attributes according to specific criteria.

5. EXPERIMENT 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 is used in this work. The dataset is recorded into three different scenarios;

- Based On Accident Information
- Based On Casualty Information
- Based On Vehicle Information

Totally, this dataset consist of 34,575 record sets. Among them 14576 belongs to accident, 9,628 belongs to vehicle and remaining 10,371 belongs to casualty.

5.1 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 Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers can correctly and incorrectly classified all the attributes. In that District Council District, Weather, Junction Control, Vehicle Movement, Number of Casualties Injured and Type of Collision attributes are used in my work. The values of these attributes are listed in the tables (2 and 3).

Table 2. Correctly (Cc) and Incorrectly (Icc) Classified Accident Dataset

Classifier	District Council District		Weather		Junction Control	
	# record	Accuracy%	# record	Accuracy%	# record	Accuracy%

Naive Bayes	**Cc**	2042	14.01	13196	90.53	10804	74.12
	Icc	12534	85.99	1380	9.47	3772	25.88
J48	**Cc**	3036	20.83	13321	91.39	10953	75.14
	Icc	11540	79.17	1255	8.61	3623	24.86
AdaBoost M1	**Cc**	1742	11.95	13051	89.54	10577	72.56
	Icc	12834	88.05	1525	10.46	3999	27.44
PART	**Cc**	3010	20.65	13459	92.34	11111	76.23
	Icc	11566	79.35	1117	7.66	3465	23.77
Random Forest	**Cc**	3743	25.68	13867	95.14	11592	79.53
	Icc	10833	74.32	709	4.86	2984	20.47

Table 3. Correctly (Cc) and Incorrectly (Icc) Classified Accident Dataset

Classifier		**Vehicle Movements**		**Number of Casualties Injured**		**Type of Collision**	
		# record	**Accuracy %**	**# record**	**Accuracy %**	**# record**	**Accuracy %**
Naive Bayes	**Cc**	12259	84.10	11605	79.62	9976	68.44
	Icc	2317	15.90	2971	20.38	4600	31.56
J48	**Cc**	12357	84.78	12378	84.92	10329	70.86
	Icc	2219	15.22	2198	15.08	4247	29.14
AdaBoost M1	**Cc**	11722	80.42	12378	84.92	9187	63.03
	Icc	2854	19.58	2198	15.08	5389	36.97
PART	**Cc**	12532	85.98	12489	85.68	10390	71.28
	Icc	2044	14.02	2087	14.32	4186	28.72
Random Forest	**Cc**	13090	89.81	12992	89.13	10836	74.34
	Icc	1486	10.19	1584	10.87	3740	25.66

5.1.1 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, Rain, Natural Light, Road Classification and Number of Vehicles Involved.

Severity
For Naive Bayes the correctly classified percentage is 84.66, for J48 is 84.64, for AdaBoostM1 is 84.64, for PART 85.18 and for Random Forest 88.25. From this, it concludes that there is no significant difference between them.

Hit and Run
For Naive Bayes the correctly classified percentage is 98.87, for J48 is 98.91, for AdaBoostM1 is 98.91, for PART 98.91 and for Random Forest 99.16. From this, it concludes that there is no significant difference between them.

Rain
For Naive Bayes the correctly classified percentage is 90.13, for J48 is 90.16, for AdaBoostM1 is 89.91, for PART 90.81 and for Random Forest 93.44. From this, it concludes that there is no significant difference between them.

Natural Light
For Naive Bayes the correctly classified percentage is 65.01, for J48 is 65.79, for AdaBoostM1 is 65.66, for PART 69.00 and for Random Forest 75.20. From this, it concludes that there is no significant difference between Naive Bayes, AdaBoostM1 and J48.

Road Classification
For Naive Bayes the correctly classified percentage is 99.39, for J48 is 99.42, for AdaBoostM1 is 99.42, for PART 99.42 and for Random Forest 99.51. From this, it concludes that there is no significant difference between them.

Number of Vehicles involved
For Naive Bayes the correctly classified percentage is 96.84, for J48 is 96.85, for AdaBoostM1 is 90.89, for PART 96.85 and for Random Forest 97.91. From this, it concludes that there is no significant differences between them expect AdaBoostM1.

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, Number of Casualties injured, Weather, Junction Control and District Council District. The correctly classified Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers' percentages of the attributes are listed in the table 4.

Table 4. Accident Dataset Classification

Classifier	District Council District	Weather	Junction Control	Vehicle Movement	No. of Casualties Injured	Types of Collision
Naive Bayes	14.01	90.53	74.12	84.10	79.62	68.44
J48	20.83	91.53	75.14	84.78	84.92	70.86
AdaBoostM1	11.95	89.54	72.56	80.42	84.92	63.03
PART	20.65	92.34	76.23	85.98	85.68	71.28
Random Forest	25.68	95.14	79.53	89.81	89.13	74.34

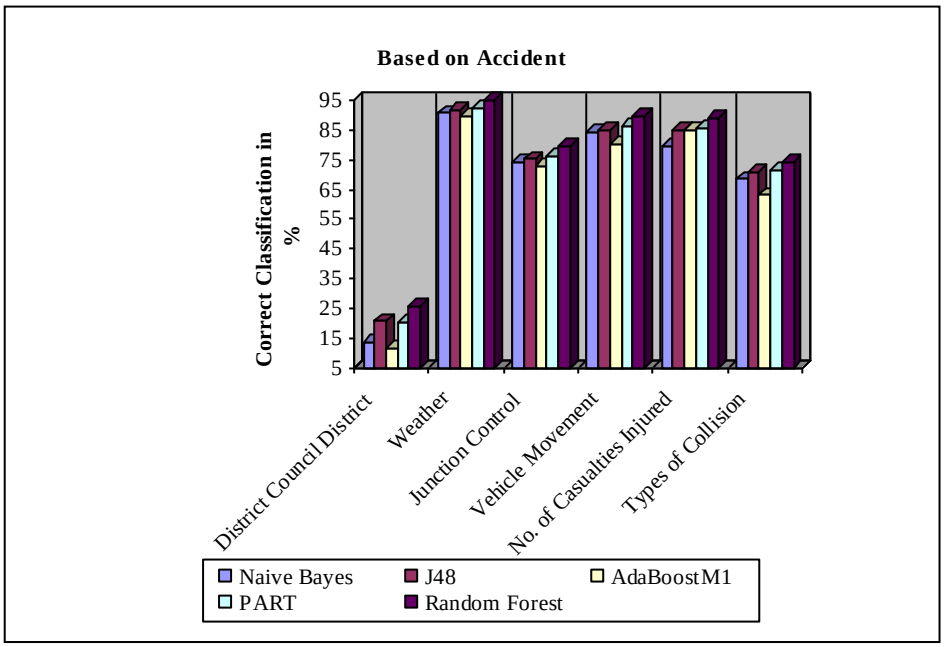

Figure 1. Comparison of Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers based on accident dataset

Figure 1 shows the Accident dataset graph. In this, The District Council District attribute takes 14.01% for Naive Bayes, 20.83% for J48, 11.95% for AdaBoostM1, 20.65% for PART and Random Forest classifier takes 25.68%. The Weather attribute takes 90.53% for Naive Bayes, 91.53% for J48, 89.54% for AdaBoostM1, 92.34% for PART and Random Forest classifier takes 95.14%. The Junction Control attribute takes 74.12% for Naive Bayes, 75.14% for J48, 72.56% for AdaBoostM1, 76.23% for PART and Random Forest classifier takes 79.53%. The Vehicle Movement attribute takes 84.10% for Naive Bayes, 84.78% for J48, 80.42% for AdaBoostM1, 85.98% for PART and Random Forest classifier takes 89.81%. The Number of Casualties Injured attribute takes 79.62% for Naive Bayes, 84.92% for J48, 84.92 for AdaBoostM1, 85.68% for PART and Random Forest classifier takes 89.13%. The Types of Collision attribute takes 68.44% for Naive Bayes, 70.86% for J48, 63.03% for AdaBoostM1, 71.28% for PART and Random Forest classifier takes 74.34%. 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 Accident dataset.

5.2 Based on Casualty

The total record set used is 10,371. 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 10,371 records Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers can correctly and incorrectly classified all the attributes. In that Casualty Age, Casualty Sex, Role of Casualty and Location of Injury attributes are used in this work. The values of these attributes are listed in the table 5.

Table 5. Correctly (Cc) and Incorrectly (Icc) Classified Casualty Dataset

Classifier		Casualty Age		Casualty Sex		Role of Casualty		Location of Injury	
		# record	Accuracy %	# record	Accuracy %	# record	Accuracy %	# record	Accuracy %
Naive Bayes	Cc	3201	30.86	7492	72.24	6748	65.07	3737	36.03
	Icc	7170	69.14	2879	27.76	3623	34.93	6634	63.97
J48	Cc	3265	31.48	7672	73.01	6793	65.50	3828	37.01
	Icc	7106	68.52	2799	26.59	3578	34.50	6533	62.99
AdaBoostM1	Cc	2840	27.38	7434	71.68	6666	64.28	3652	35.21
	Icc	7531	72.62	2937	28.32	3705	35.72	6719	64.79
PART	Cc	3281	31.64	7642	73.69	6902	66.55	3915	37.75
	Icc	7090	68.36	2729	26.31	3469	33.45	6456	62.25
Random Forest	Cc	3311	32.89	7667	74.89	7023	67.72	4134	39.86
	Icc	7060	67.11	2704	25.11	3348	32.28	6237	60.14

5.2.1 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, Degree of Injury, Pedestrian Action and Vehicle Class of Driver or Passenger Casualty.

Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers' 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 10,371 and the attributes are Casualty Age, Casualty Sex, Role of Casualty and Location of Injury. The correctly classified Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers' percentages of the attributes are listed in the table 6.

Table 6. Casualty Dataset Classification

Classifier	Casualty Age	Casualty Sex	Role of Casualty	Location of Injury
Naive Bayes	30.86	72.24	65.07	36.03
J48	31.48	73.01	65.50	37.01
AdaBoostM1	27.38	71.68	64.28	35.21
PART	31.34	73.69	66.55	37.75
Random Forest	32.89	74.89	67.72	39.86

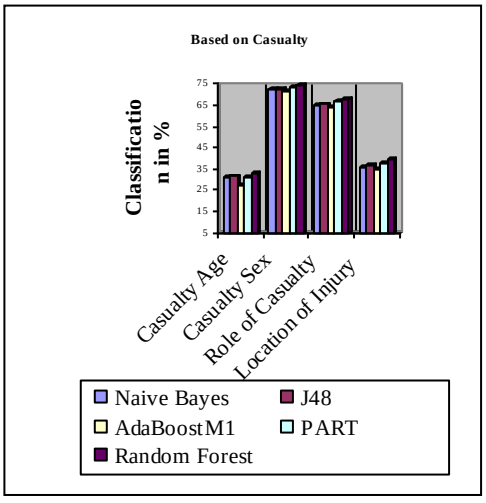

Figure 1. Comparison of Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers based on casualty dataset

Figure 2 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 takes 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 takes 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.

5.3 Based on Vehicle

The total record set used is 9,628. The attributes involved in this case are Driver Age, Drive Sex, Vehicle Class, Year of Manufacture and Severity of accident.

Out of 9,628 records Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers can correctly and incorrectly classified all the attributes. In that Driver Age, Vehicle Class and Year of Manufacture attributes are used in my work. The values of these attributes are listed in the table 7.

Table 7. Correctly (Cc) and Incorrectly (Icc) Classified Vehicle Dataset

Classifier		Driver Age		Vehicle Class		Year of Manufacture	
		# record	Accuracy %	# record	Accuracy %	# record	Accuracy %
Naive Bayes	Cc	2687	27.91	4349	45.17	2509	26.06
	Icc	6941	72.09	5279	54.83	7119	73.94
J48	Cc	2810	29.19	4437	46.08	2640	27.42
	Icc	6818	70.81	5191	53.92	6988	72.58
AdaBoost M1	Cc	2183	22.67	3584	37.22	2206	22.91
	Icc	7445	77.33	6044	62.78	7422	77.09
PART	Cc	2950	30.64	4537	47.12	2733	28.39
	Icc	6678	69.36	5091	52.88	6895	71.61
Random Forest	Cc	3053	31.71	4686	48.67	2800	29.08
	Icc	6575	68.29	4942	51.33	6828	70.92

5.3.1 Applying Genetic Algorithm for Feature Selection in Vehicle Dataset

From the vehicle dataset not all five 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 Driver Sex and Severity of Accident.

Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers' results conclude that there is no significant difference between them. Therefore, the Genetic Algorithm eliminates these attributes, which are not potential for classification.

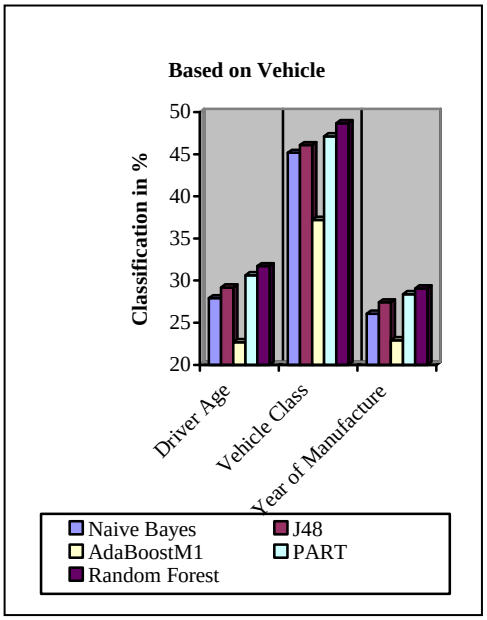

Figure 2 Comparison of Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers based on casualty dataset

Finally overall records used for vehicle is 9,628 and the attributes are Driver Age, Vehicle Class and Year of Manufacture. The correctly classified Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers' percentages of the attributes are listed in the table 8.

Table 8. Vehicle Dataset Classification

Classifier	Driver Age	Vehicle Class	Year of Manufacture
Naive Bayes	27.91	45.17	26.06
J48	29.19	46.08	27.42
AdaBoostM1	22.67	37.22	22.91
PART	30.64	47.12	28.39
Random Forest	31.71	48.67	29.08

Figure 3 shows the Vehicle dataset graph. In this, Driver Age attribute takes 27.91% for Naive Bayes, 29.19% for J48, 22.67% for AdaBoostM1, 30.64% for PART and Random Forest classifier takes 31.71%. The Vehicle Class attribute takes 45.17% for Naive Bayes, 46.08% for J48, 37.22% for AdaBoostM1, 47.12% for PART and Random Forest classifier takes 48.67%. The Year of Manufacture attribute takes 26.06% for Naive Bayes, 27.42% for J48, 22.91% for AdaBoostM1, 28.39% for PART and Random Forest classifier takes 29.08%. 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 Vehicle dataset.

6. 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.

Random Forest outperforms than other classification algorithms 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.

7. REFERENCES

[1]. 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.

[2]. 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.

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

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

[5]. E. Frank and I. H. Witten. Generating accurate rule sets without global optimization. In Proc. of the Int'l Conf. on Machine Learning, pages 144–151. Morgan Kaufmann Publishers Inc., 1998.

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

[7]. Gartner Group Advanced Technologies & Applications Research Note 2/1/95

[8]. Data Mining and Data Warehousing available at: http://databases.about.com/od/datamining/g/Classification.htm

[9]. Genetic algorithm available at: http://en.wikipedia.org/wiki/Genetic_algorithm

[10]. Road Traffic Accident Statistics available at: http://www.td.gov.hk/en/road_safety/road_traffic_accident_statistics/2008/index.html

[11]. Statistical Analysis Software, Data Mining, Predictive Analytics available at: http://www.statsoft.com/txtbook/stdatmin.html

[12]. Data Mining: Bagging and Boosting available at: http://www.icaen.uiowa.edu/~comp/Public/Bagging.pdf

[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]. 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

[15]. 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.

[16]. National Highway Traffic Safety Administration, Traffic Safety Facts 2005, 2007, P. 54. http://www-nrd.nhtsa.dot.gov/Pubs/TSF2006.PDF

[17]. 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.

[18]. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[19]. Rish, Irina. (2001). "An empirical study of the naive Bayes classifier". IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.

AUTHOR

Dr..M.Hemalatha completed MCA, M..Phil., Ph.D. in Computer Science and currently working as a Asst. Professor and Head, Dept. of software systems in Karpagam University. Ten years of experience in teaching and published twenty seven paper in International Journals and also presented seventy papers in various National conferences and one international conference. Area of research is Data mining, Software Engineering, Bioinformatics and Neural Network also reviewer in several National and International journals.

S.Krishnaveni completed MCA, M.Phil, and currently pursuing Ph.D in computer science at Karpagam University under the guidance of Dr.M.Hemalatha, Head, Dept. of Software System, Karpagam University, Coimbatore. Area of Research is Data Mining.