Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms

S.Shanthi
Senior Lecturer (Ph.D. Research Scholar)
Department of Computer Science & Engineering
Rajalakshmi Institute of Technology
(Affiliated to Anna University, Chennai)
Kuthambakkam, Chennai, India

Dr.R.Geetha Ramani
Professor and Head,
Department of Computer Science & Engineering
Rajalakshmi Engineering College
(Affiliated to Anna University, Chennai)
Thandalam, Chennai, India

ABSTRACT
This paper emphasizes the importance of Data Mining classification algorithms in predicting the vehicle collision patterns occurred in training accident data set. This paper is aimed at deriving classification rules which can be used for the prediction of manner of collision. The classification algorithms viz. C4.5, C-RT, CS-MC4, Decision List, ID3, Naïve Bayes and RndTree have been applied in predicting vehicle collision patterns. The road accident training data set obtained from the Fatality Analysis Reporting System (FARS) which is available in the University of Alabama's Critical Analysis Reporting Environment (CARE) system. The experimental results indicate that RndTree classification algorithm achieved better accuracy than other algorithms in classifying the manner of collision which increases fatality rate in road accidents. Also the feature selection algorithms including CFS, FCBF, Feature Ranking, MIFS and MODTree have been explored to improve the classifier accuracy. The result shows that the Feature Ranking method significantly improved the accuracy of the classifiers.

General Terms
Data Mining, Classification Algorithms, Feature Selection, Accident Data Analysis

Keywords
Classification Algorithms, Feature Selection Algorithms, Manner of Collision, Fatal Severity, Collision Patterns, Prediction

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

1.1 Organization of the paper
The paper is organized as follows. Section 2 provides the summary of related work in this area. In section 3 we investigate the data set and discuss the system model. Section 4 discusses the preparation of the data for analysis and brief about the relevance analysis. Section 5 illustrates the classification algorithms used for the empirical study. The experimental results and observations are discussed in Section 6, and the conclusions and future research directions are presented in Section 7. Section 8 lists the references used in this study and Section 9 gives the authors profile. In next section we discuss the related work carried out in this area.

2. LITERATURE SURVEY
Handan et.al [4] compared logistic regression model with classification tree method in determining social-demographic risk factors which have affected depression status of women in separate postpartum periods. They proposed that Classification tree method gives more information with detail on diagnosis by evaluating a lot of risk factors together than logistic regression model.
Chang et.al [2] applied non-parametric classification tree techniques to analyze Taiwan accident data from the year 2001. They developed a CART model to find the relationship between injury severity and driver/vehicle characteristics, highway/environment variables, and accident variables.

Yong Soo Kim [11] compared the performance of data mining and statistical techniques by varying the number of independent variables, the types of independent variables, the number of classes of the independent variables, and the sample size. The results have shown that the artificial neural network performance improved faster than that of the other methods as the number of classes of categorical variable increased.

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

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

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

3. TRAINING DATASET DESCRIPTION

The accident training data set used in our study is obtained from Fatality Analysis Reporting System (FARS) [13] which is available in Critical Analysis Reporting Environment (CARE) system. FARS was developed by the National Center for Statistics and Analysis (NCSA) of the National Highway Traffic Safety Administration (NHTSA) to provide an overall measure of highway safety, to help identify traffic safety problems, to suggest solutions, and to help provide an objective basis to evaluate the effectiveness of motor vehicle safety standards and highway safety programs.

3.1 Training Data Set: Descriptive Analysis

The objective for this data mining research is the discovery of classification rules based on manner of collision that would find out and differentiate accidents which are serious to those which are potentially not serious in different levels. The data set for the study contains traffic accident records of U.S. country consists of 56 states. It holds the accident details from January, 2007 up to December, 2007 a total number of 37259 cases. This data was in an excel file format with 57 attributes to describe each record. We have taken 26 attributes which are significant for the analysis and classified them into 3 sets: Accident Specific Attributes, Road related attributes and Environment related attributes. Table 1 gives the list of attributes used for the study.

Table 1. Attributes and Description

Attributes	Description
ACCIDENT SPECIFIC ATTRIBUTES	
FATAL_SEVERITY	Fatal Severity Level
HIT_RUN	Hit and Run
SCH_BUS	School Bus involved or not
MAN_COLL	Manner of Collision
DRUNK_DR	Drunken Driver
ROAD SPECIFIC ATTRIBUTES	
NHS	National Highway System
ALIGNMENT	Road Alignment
SP_LIMIT	Speed limit
CF1	Crash Related factor
REL_JNC	Related to Junction
REL_ROAD	Related to Road
ROAD_FNC	Road way function
NO_LANES	Number of Lanes
TRA_CONT	Traffic Control Devices
T_CONT_F	Traffic Control Device Functioning
SUR_COND	Surface Condition
PROFILE	Road way Profile
ROUTE	Rural or Urban
TRAF_FLO	Traffic Flow
PAVE_TYP	Pavement Type
SUR_COND	Surface Condition
SP_JUR	Special Jurisdiction
ENVIRONMENT RELATED ATTRIBUTES	
LGT_COND	Light Condition
DAY_WEEK	Day of the week
MONTH	Month in which the accident happened
WEATHER	Weather information
C_M_ZONE	Construction and Maintenance Zone

All the records have been divided into 50 subsets based on the states and we have applied the feature selection and classification algorithms to each and every subset. The distribution of records based on the manner of collision is analyzed in SPSS and statistics is given in the Table 2.

Table 2. Frequency Distribution of Manner of Collision

MANNER OF COLLISION				
Value	**Frequency**	**Cumulative Frequency**	**Percentage**	**Cumulative Percentage**
None	**22699**	**22699**	**60.94%**	**60.94%**
Front-to-Rear	**2314**	**25013**	**6.21%**	**67.15%**
Front-to-Front	**3784**	**28797**	**10.16%**	**77.31%**
Angle - Front-to-Side	**7255**	**36052**	**19.48%**	**96.79%**
Sideswipe - Same Direction	**485**	**36537**	**1.30%**	**98.09%**
Sideswipe - Opposite Direction	**482**	**37019**	**1.29%**	**99.39%**
Rear-to-Side	**74**	**37093**	**0.20%**	**99.58%**
Rear-to-Rear	**83**	**37176**	**0.22%**	**99.81%**
Other	**72**	**37248**	**0.19%**	**100.00%**

3.2 System Model

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

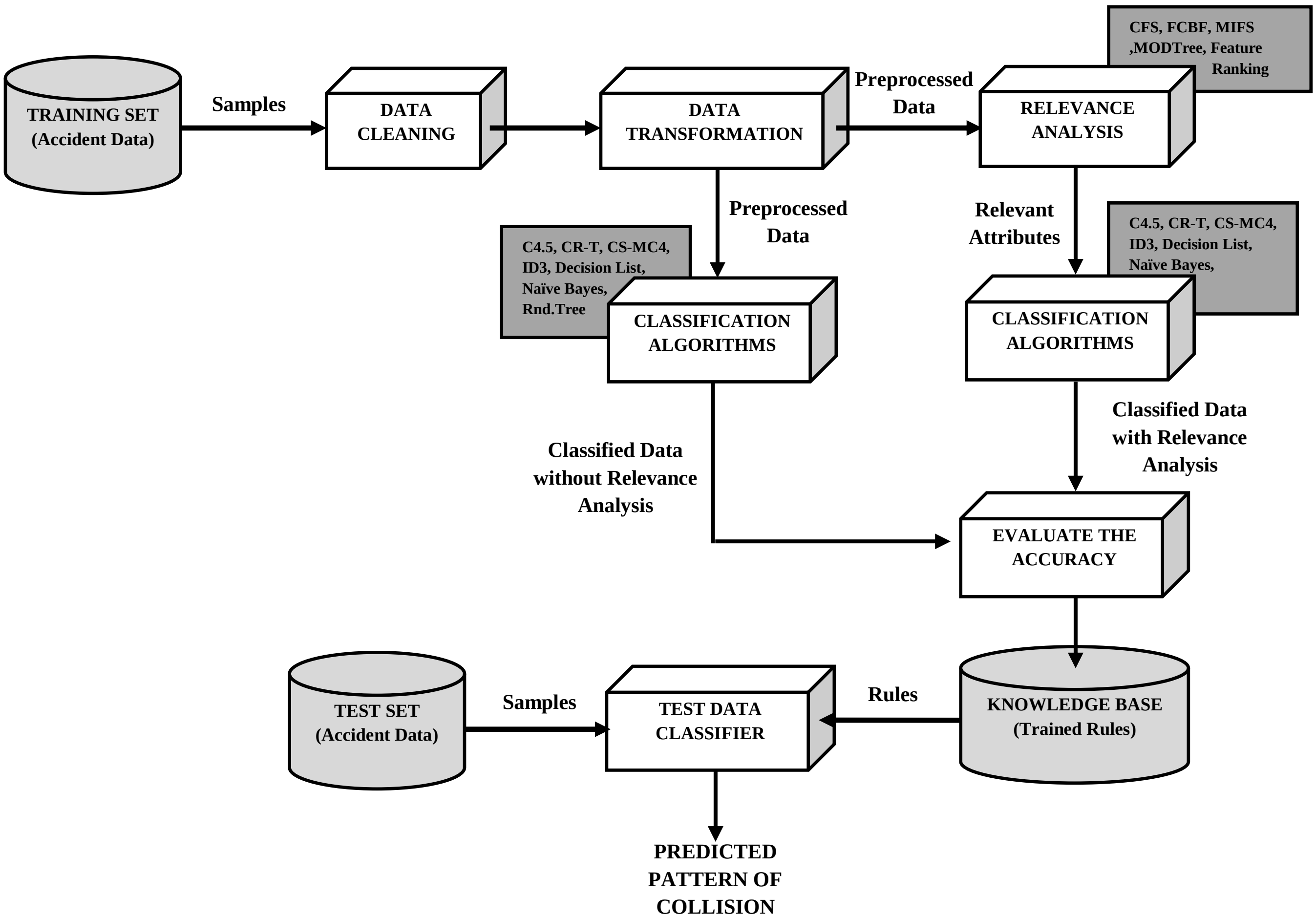


Fig 1: Methodology

The next section discusses the data preparation which is to be done prior to classification to obtain the accurate results.

4. DATA PREPARATION

The data set we would like to analyze may be incomplete, noisy and inconsistent [3]. Thus data preprocessing needs to be completed before applying the algorithms so as to improve the performance of the same. We see the details of preprocessing in the following sections.

4.1 Data Cleaning

It attempts to fill in missing values, smoothing noise data and correct inconsistencies in the data [3]. Upon an in - depth exploration of the data has shown that a good part of the variables were insignificant to our study. Accordingly, based on the observation insignificant attributes like VE_Forms, latitude, longitude, etc. which is a total of 31 were removed and 26 have been included. Specifically records with missing values were excluded in order to avoid compromising the result. Consequently the size of the dataset was reduced to 37248 records.

4.2 Data Transformation

It converts the data into appropriate forms for mining [3]. The data set used in our study contained integer values for the entire attributes. So we have identified categorical variables and coded them by converting integer into text For example *Sp_Limit* is derived to classify the input values between 0 and 30 as Low, 31 and 60 as Medium and greater than 60 as High. Similar transformations have been done to have the categorical variables. When the pre-processing was completed, the final dataset used for modeling had 37248 records described by 26 attributes.

4.3 Relevance Analysis

Dimensionality reduction and feature subset selection are two techniques for reducing the attribute space of a feature set, which is an important component of both supervised and unsupervised classification and regression problems [1]. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [6]. In the following sections we discuss briefly few feature selection algorithms we applied in our study.

4.3.1 CFS

Correlation based Feature Selection is a supervised feature selection algorithms [8] based upon a filtering approach. It processes the selection independently form the learning algorithm it considers the redundancy of the input attributes [12].

4.3.2 FCBF

Fast Correlation Based Filter algorithm [1] is designed for high dimensional data and has been shown effective in removing both irrelevant features and redundant features. Lei et.al [7] in their results suggests that FCBF is practical for feature selection for classification of high dimensional data. It can efficiently achieve high degree of dimensionality reduction and enhance classification accuracy with predominant features [5].

4.3.3 Feature Ranking

It is a univariate feature ranking algorithm [12] using CHI-2 criterion. It ranks the input attributes according to the relevance. It does not allow the redundancy of the attributes.

4.3.4 MIFS

Mutual Information Feature Selector is a supervised feature selection algorithm based on a filtering approach. It allows the redundancy of the input attributes. The selection phase is preceded by a feature transformation step [12] where continuous descriptors are discretized using the MDLPC algorithm.

4.3.5 MODTree Filtering

Multi valued Oblivious Decision Tree feature selection algorithm is a supervised feature selection algorithms based on a filtering approach [12]. It processes the selection independently from the learning algorithm. It considers the redundancy of the input attributes.
The comparison between these feature selections algorithms are discussed in the coming sections.

5. CLASSIFICATION ALGORITHMS

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

5.1 C4.5

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

5.2 ID3

ID3 is a decision tree induction algorithm. In the decision tree each node corresponds to a non-categorical attribute [17] and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute for the records described by the path from the root to that leaf. In the decision tree at each node should be associated the non-categorical attribute which is most informative among the attributes not yet considered in the path from the root. Entropy is used to measure how informative is a node.
The ID3 algorithm takes all unused attributes and counts their entropy concerning test samples. Choose attribute for which entropy is minimum (or, equivalently, information gain is maximum).

5.3 C&RT

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

5.4 CS-MC4

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

5.5 Decision List

The decision list induction is an ordered list of conjunctive rules [12]. It can handle a multi class problem. The obtained classifier gives an ordered set of rules.

5.6 Naïve Bayes

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

5.7 Random Tree

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

- Let the number of training cases be N, and the number of variables in the classifier be M.
- We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.
- Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample).
- Use the rest of the cases to estimate the error of the tree, by predicting their classes.

- For each node of the tree, randomly choose m variables on which to base the decision at that node.
- Calculate the best split based on these m variables in the training set.
- Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).
- For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in.
- This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction [18].

5.8 Rule Induction

Inductive rule learning algorithm [12] based on the separate and conquers principle. The obtained classifier is an unordered set of rules. The algorithm can handle a multi class problem.
The results we obtained are discussed in the following section.

6. EXPERIMENTAL RESULTS

TANAGRA [12] is data mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. It is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license. In our study we used Tanagra to carry out experiments. The results we obtained from our experiment are discussed in further sub sections.

6.1 Phase I: Feature Selection

The data set with 26 attributes was used for the study. We applied the feature selection algorithms viz. CFS, FCBF, Feature Ranking, MIFS and MODTree algorithms. The number of attributes selected by these algorithms for few states is listed in Table 3.

Table 3. Number of Attributes selected by Feature Selection algorithms

STATE	FEATURE SELECTION METHODS				
	CFS	FCBF	Feature Ranking	MIFS	MODTree
Alabama	2	6	18	7	7
Alaska	8	4	3	3	1
Arizona	3	7	20	8	7
Arkansas	2	3	16	8	5
California	2	4	18	7	6
Colorado	2	5	8	8	5
Delaware	2	4	6	4	4
Columbia	4	4	3	1	2
Florida	2	6	18	9	7
Georgia	2	6	17	9	8
Hawaii	2	4	5	1	1
Idaho	3	3	8	4	4
Illinois	2	4	12	7	5

For all the subsets the feature ranking algorithm selected more attributes as relevant attributes. The comparison between the feature selection algorithms is depicted in the Figure. 2.

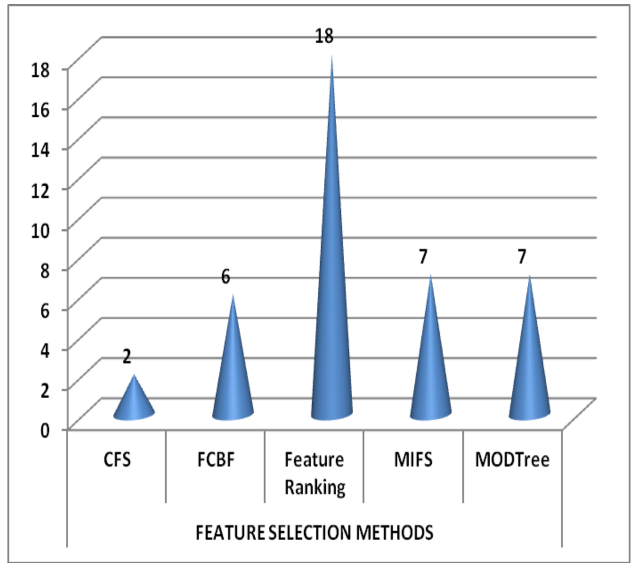
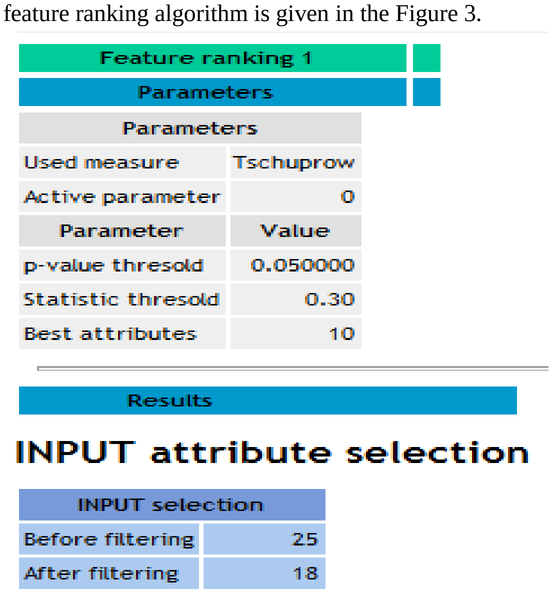


Fig 2: Comparison of Feature Selection Algorithms

Total number of 25 attributes has been used for the study. The Feature ranking algorithm selected 18 attributes as relevant attributes to classify the manner of collision. Without feature selection (with 25 attributes) the accuracy of the Random Tree classifier was 87.3%. After doing feature selection (18 attributes) the accuracy of the Random Tree classifier was 94.38 which is a significant improvement of 7.08%. Sample result produced by the feature ranking algorithm is given in the Figure 3.



Feature ranking 1

Parameters

Parameters	
Used measure	Tschuprow
Active parameter	0
Parameter	**Value**
p-value thresold	0.050000
Statistic thresold	0.30
Best attributes	10

Results

INPUT attribute selection

INPUT selection	
Before filtering	25
After filtering	18

Fig 3: Sample result produced by Feature Ranking

The relevant attributes selected by Feature Ranking algorithm is listed in the Figure 4.

Attributes	
1	REL_ROAD
2	REL_JUNC
3	TRA_CONT
4	T_CONT_F
5	ROUTE
6	ALIGNMNT
7	NHS
8	ROAD_FNC
9	NO_LANES
10	TRAF_FLO
11	CF1
12	LGT_COND
13	PAVE_TYP
14	MONTH
15	SP_LIMIT
16	HIT_RUN
17	DRUNK_DR
18	C_M_ZONE

Fig 4: Relevant Attributes selected by Feature Ranking

The feature ranking algorithm selects the variables whose p-value<=0.05. The selection criteria followed by feature ranking algorithm is given in the Figure 4.

N°	Attribute	Values	Statistic	Statistic (Histogram)	p-value
1	REL_ROAD	7	0.288078		0.000000
2	REL_JUNC	7	0.249779		0.000000
3	TRA_CONT	5	0.240389		0.000000
4	T_CONT_F	4	0.239375		0.000000
5	ROUTE	6	0.153017		0.000000
6	ALIGNMNT	3	0.146518		0.000000
7	NHS	3	0.142831		0.000000
8	ROAD_FNC	15	0.137442		0.000000
9	NO_LANES	6	0.126717		0.000000
10	TRAF_FLO	5	0.124123		0.000000
11	CF1	6	0.122354		0.000001
12	LGT_COND	6	0.118224		0.000008
13	PAVE_TYP	5	0.110533		0.000082
14	MONTH	12	0.106521		0.037034
15	SP_LIMIT	4	0.103267		0.000448
16	HIT_RUN	2	0.102571		0.000210
17	DRUNK_DR	2	0.096440		0.000806
18	DAY_WEEK	7	0.094231		0.050030

Fig 4: Calculation details of Feature Ranking Algorithm

Similarly the feature selection algorithms have been applied to all the subsets, of which Feature Ranking significantly improved the performance of classifiers.

6.2 Phase II: Classification Algorithms

The data set is analyzed using Random Tree, C4.5, CS-MC4, C&RT, Decision List, Naïve Bayes, Rule Induction and ID3 classifier models by having MAN_COLL as dependent variable and all others were set as independent variables. Accuracy is measured using confusion matrix. A sample confusion matrix is given in the Figure 5.

Supervised Learning 15 (Rnd Tree)

Parameters

Nb att for split = -1

Results

Classifier performances

Error rate	0.1119

Values prediction

Value	Recall	1-Precision
None	0.9730	0.0968
Front-To-Front	0.7027	0.2041
Front-To-Rear	0.4833	0.2927
Angle-Front-To-Side	0.9189	0.0377
SidesWipe-Opposite-Direction	0.2500	0.0000

Confusion matrix

	None	Front-To-Front	Front-To-Rear	Angle-Front-To-Side	SidesWipe-Opposite-Direction	Sum
None	504	10	3	1	0	518
Front-To-Front	27	78	5	1	0	111
Front-To-Rear	20	9	29	2	0	60
Angle-Front-To-Side	5	0	4	102	0	111
SidesWipe-Opposite-Direction	2	1	0	0	1	4
Sum	558	98	41	106	1	804

Fig 5: Classification Result produced by Random Tree Algorithm

The results obtained from various classification algorithms is given in Table 4.

Table 4. Comparison of Classifier Accuracy Based on Feature Ranking

CLASSIFIER ACCURACY		
ALGORITHM	**ATTRIBUTES**	
	All Attributes	**Relevant Attributes**
C4.5	80.99	80.59
C-RT	76.24	76.24
CS-MC4	71.09	71.09
Decision List	67.92	67.92
ID3	75.54	75.54
NaiveBayes	73.27	72.28
RndTree	87.30	94.38
Rule Induction	75.64	75.54

Boost up in the accuracy of the classifiers using Feature Ranking algorithm is depicted in the Figure 6.

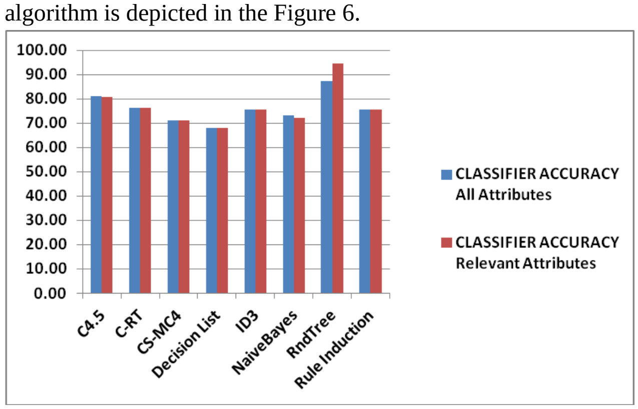


Fig 6: Classifiers Accuracy with and without Feature Ranking Algorithm

Table 5 gives the accuracy of all the classifiers experimented in all the sub sets.

Table 5. Classifier Accuracy for all the States Based on Feature Ranking Algorithm

STATE	CLASSIFIER ACCURACY							
	C4.5	C&RT	CS-MC4	Decision List	ID3	Naïve Bayes	Rnd Tree	Rule Induction
Alabama	81	76	71	68	76	72	94	76
Alaska	77	53	60	64	53	75	97	70
Arizona	77	71	68	71	72	70	94	74
Arkansas	79	71	70	69	70	69	88	71
California	79	73	67	73	75	68	90	75
Colorado	78	73	64	73	73	75	96	78
Delaware	77	75	60	79	60	79	96	75
Columbia	81	78	78	78	78	86	92	78
Florida	75	69	60	69	69	66	90	71
Georgia	79	71	65	72	72	70	94	74
Hawaii	82	66	70	70	66	78	94	70
Idaho	83	74	74	82	74	83	97	81
Illinois	79	73	67	74	75	73	94	75
Indiana	77	71	56	69	74	71	92	76
Lowa	84	77	70	77	73	77	95	77
Kansas	77	64	61	68	71	72	94	71
Kentucky	82	75	62	69	79	75	95	80
Louisiana	79	72	68	65	76	74	95	79
Maine	83	67	67	79	67	85	97	82
Maryland	77	71	60	72	68	71	93	73
Massachusetts	80	69	69	73	70	72	93	76
Michigan	79	74	57	73	74	70	95	75
Minnesota	82	71	50	73	72	77	97	76
Mississippi	87	81	74	81	84	80	89	83
Misouri	82	76	67	68	75	75	95	77
Montana	84	80	76	76	72	82	96	83
Nebraska	83	78	70	80	78	76	97	81
Nevada	80	78	67	77	68	76	95	80
New Hampshire	81	61	61	70	61	81	93	80
New Jercy	77	73	63	68	65	69	95	73
Mexico	77	75	73	76	70	78	97	77
New York	75	69	61	70	65	65	95	70
North Carolina	80	76	61	74	76	71	92	76
North Dakota	86	82	79	87	72	83	99	83
Nohio	79	72	69	71	74	71	92	74
Oklahoma	80	69	58	64	75	78	94	75
Oregon	81	75	66	69	70	75	91	78
Pennsylvania	80	72	68	71	79	69	93	75
Rhode Island	80	80	80	77	80	88	98	80
South Carolina	83	78	74	75	79	74	91	79
South Dakota	82	68	76	78	68	84	93	79
Tennessee	81	76	62	69	76	72	95	77
Texas	79	73	59	69	74	71	90	72
Utah	73	63	63	70	64	72	95	73
Virginia	78	72	64	73	72	69	93	74
Washington	75	72	60	64	67	68	92	74
West Virginia	81	73	67	75	69	77	96	78
Wisconsin	78	72	58	69	72	74	94	74
Wyoming	84	75	75	77	75	87	96	80

The error rates of all the classifiers with and without using feature selection algorithms are given in Table 6.

Table 6. Comparison of Classifier Error Rates Based on Feature Ranking

CLASSIFIERS	FEATURE SELECTION ALGORITHM					
	NONE	CFS	FCBF	Feature Ranking	MIFS	MOD Tree
C4.5	0.1901	0.270	0.2436	0.1941	0.2525	0.2267
C-RT	0.2376	0.270	0.2703	0.2703	0.2703	0.2465
CS-MC4	0.2891	0.289	0.2891	0.2891	0.2891	0.2891
DECISION LIST	0.3208	0.331	0.3307	0.3208	0.3208	0.3307
ID3	0.2446	0.270	0.2505	0.2446	0.2545	0.2446
NaiveBayes	0.2673	0.272	0.2436	0.2772	0.2941	0.2594
RndTree	0.127	0.270	0.2436	0.0562	0.2624	0.2317
Rule Induction	0.2436	0.290	0.2901	0.2446	0.2941	0.2911

Compared with all the feature selection algorithms Feature Ranking algorithm significantly boosts the accuracy of the classifiers. It is clearly given in the Figure 7.

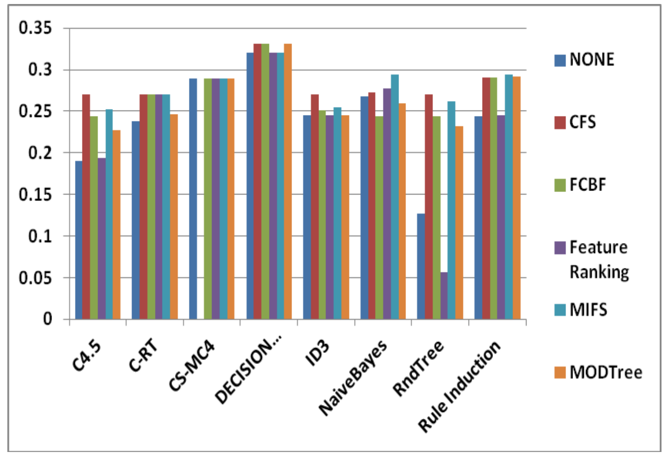


Fig 7: Influence of the Relevance Analysis to the accuracy of the Classifiers

Next section shows few sample rules derived from the classifier.

6.3 Sample Rules

Sample rules obtained from the Decision Tree are given in Figure 8.

Data Description

Target Attribute	MAN_COLL(8 Values)
#descriptors	18

Number of Rules = 3

Knowledge-based System

Antecedent	Consequent	Distribution
IF REL_ROAD in [On Roadway] -- REL_JUNC in [Non Junction] -- TRAF_FLO in [Two Way Not Physically Divided] -- SP_LIMIT in [Medium] -- HIT_RUN in [NO] – NO_LANES in [Two Lanes] – CF1 in [None]	**MAN_COLL in [Front-To-Front]**	(44; 79; 28; 7; 1; 0; 8; 0)
IF REL_JUNC in [Intersection] – T_CONT_F in [Functioning Properly] – REL_ROAD in [On Roadway] – HIT_RUN in [NO] – PAVE_TYP in [Black Top]	**MAN_COLL in [Angle-Front-To-Front]**	(2; 5; 88; 0; 0; 0; 0; 0)
IF REL_ROAD in [On Roadway] – LGT_COND in {Day Light] – ALIGNMENT in [Straight] – TRA_CONT in {None] – PAVE_TYP in [Black Top]	**MAN_COLL in [Angle-Front-To-Front]**	(31; 37; 55; 23; 3; 1; 3; 0)
(DEFAULT RULE)	**MAN_COLL in [None]**	(554; 20; 36; 29; 2; 1; 1; 1)

Fig 8: Sample Rules Obtained from Decision Tree

From the study we could observe that Feature Ranking algorithm is significantly improving the accuracy of the classifiers. Also the results show that the Random Tree algorithm gives accurate results than other classification algorithms in classifying the records based on manner of collision.

7. CONCLUSION

The objective of this research undertaking was to explore the possible application of data mining technology for mining vehicle collision patterns in road accident training data set. The results are validated by testing the model with the test data. In our study we employed classification algorithms on 37248 samples. The results reveal that in all the cases the Random Tree outperforms of all the other classifiers. Also it is observed that the classifier accuracy seems to be increasing when we apply Feature Ranking algorithm. The classification accuracy of the algorithms was tested, and it showed that the classifiers with proper relevance analysis give high accurate results.

8. REFERENCE

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

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

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

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

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

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

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

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

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

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

[11] Yong Soo Kim, “Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size”, Expert Systems with Applications, 2008, Vol. 34, pp. 1227–1234.

[12] Tanagra Data Mining tutorials, http://data-mining-tutorials.blogspot.com

[13] www.nhtsa.gov – FARS Analytic Reference Guide.

[14] World Health Organization, Global status report on road safety: time for action, Geneva, 2009.

[15] Feature Selection Algorithm, http://featureselection.asu.edu

[16] J. Rose Quinlan, “Programs for machine learning”.

[17] Building Classification Models ID3 and C4.5, http://www.cis.temple.edu/~ingargio/cis587/readings/id3-C4.5.html

[18] Random Tree Algorithm, http://www.answers.com

[19] Classification Algorithms, http://www.statsoft.com

9. AUTHORS PROFILE

Dr.R.Geetha Ramani is working as Professor & Head in the Department of Computer Science and Engineering, Rajalakshmi Engineering College, India. She has more than 15 years of teaching and research experience. Her areas of specialization include Data mining, Evolutionary Algorithms and Network Security. She has over 50 publications in International Conferences and Journals to her credit. She has also published a couple of books in the field of Data Mining and Evolutionary Algorithms. She has completed an External Agency Project in the field of Robotic Soccer and is currently working on projects in the field of Data Mining. She has served as a Member in the Board of Studies of Pondicherry Central University. She is presently a member in the Editorial Board of various reputed International Journals.

Mrs.S.Shanthi completed her M.C.A. from Madurai Kamaraj University and M.E. in Computer Science and Engineering at Arunai Engineering College, affiliated to Anna University, Chennai, India. She has 7 years of teaching experience. Presently she is working as Senior Lecturer in the Department of Computer Science and Engineering, Rajalakshmi Institute of Technology, Chennai and pursuing her Ph.D (Part Time) in Computer Science and Engineering at Rajalakshmi Engineering College, affiliated to Anna University, Chennai. Her areas of interest include Data Mining, Data Structures and Analysis of Algorithms and Network Security. She has published one paper in international journal and presented many papers at National and International Conferences.