A Pragmatic Approach for Refined Feature Selection for the Prediction of Road Accident Severity

R. GEETHA RAMANI[1]**, Shanthi SELVARAJ**[2]

[1] Anna University (CEG Campus),
Guindy, Chennai, Tamilnadu, 600 025, India,
rgeetha@yahoo.com

[2] Rathinam Technical Campus,
Coimbatore, Tamilnadu, 641 021, India,
psshanthiselvaraj@gmail.com

Abstract: Road accident analysis is very challenging task and investigating the dependencies between the attributes become complex because of many environmental and road related factors. An exhaustive research is being conducted to identify the optimal factors which influence fatal accidents. In this paper we propose a novel methodology called Voting Algorithm for Aggregated Feature Selection (VAAFS) which selects an optimal number of significant features with majority votes identified by more than one Feature Mining algorithms. The optimal features selected by VAAFS will be then extended to the classifiers over an Indian road accident data set obtained from the Coimbatore City Traffic Head Quarters, Tamilnadu, India and with international datasets obtained from Fatality Analysis Reporting System (FARS), USA, and the STATS19 data collection system, maintained by the of United Kingdom (UK) to model the accident severity. The output from VAAFS shows that type of vehicle, high risk road users like pedestrian and two wheelers, young road users, government holidays, selected week days, manner of collision, seating position etc. are most significant factors in modeling Accident Severity. The proposed method is highly successful in reducing misclassification error rate and to improve the predictive accuracy with optimal features than the previous studies. It seems very promising for observing road accident patterns.

Keywords: Chi-Square, Classification, Ensemble, Fatality, Feature Selection, Majority Voting, Road Accidents.

1. Introduction

The costs of fatalities and injuries due to traffic accidents have a great impact on society. World Health Organization (WHO, 2009) predicts that road collisions will jump from the ninth leading cause of death in 2004 to the fifth in 2030. Road accidents have earned India a dubious distinction. More people die in road accidents in India than anywhere else in the world. According to the report of National Crime Records Bureau (NCRB), India, the total number of deaths every year due to road accidents has now passed the 1,35,000 mark. About 60,000 lives are lost every year in road accidents and this rate is 25 times more than that of the U.S.A. The alarming rate of increase of fatality due to road-accidents in the country warrants a method to understand the causes users and their behavior.

Road traffic accident is under the influence of many factors. With an exponential growth of population, number of vehicles and the need for their use, understanding the multiple causes of road accident fatalities has become more significant especially in the advent of sophisticated technology [20]. In recent years, with the growth of the volume and travel speed of road traffic, the number of traffic accidents, especially severe crashes, has been increasing rapidly on a yearly basis [13]. Identification of these factors can help improve the overall driving safety situation, not only by preventing accidents but also by reducing their severity [14]. It is crucial for engineers to extract useful information from existing data to analyze the causes of traffic accidents, so that traffic administrations can be more accurately informed and better policies can be introduced [19].

The characteristics and availability of fatal road-crash databases have been listed worldwide [17]. Among the listed databases most of the databases are having only summary data rather individual accident information. The ever increasing tremendous amount of data has far exceeded human ability for comprehension without the use of powerful tools [10]. Data mining is the process of analyzing data from various perspectives and summarizing it into useful, meaningful and related information [10, 7]. This information can then be seen as a kind of outline of the input data, and can be used in further investigation or can be applied in the field of machine learning and predictive analytics. There are many data mining algorithms and tools that have been developed for feature selection, classification and clustering. These algorithms are used to discern and uncover knowledge patterns and make out significant and meaningful information associated with the application domain. Though, many studies have dealt this problem, lack of consensus is still visible for analyzing such data sets. This is further augmented with the complex features due to varied geographical, environmental, and social practices. In this paper, a new novel method of traffic accident data mining, based on aggregated voting method and through a comparative analysis of a variety of traffic accident data mining techniques, is put forward to identify the significance of different attributes and their respective values. The proposed method is validated on an Indian Accident data set and a foreign data set. Precisely this work has the following objectives:

1. it makes an attempt to initiate a scientific process through data mining tools effectively and to provide reasonable findings for traffic management for a site-specific purpose;
2. a novel accident severity prediction framework for accident datasets with enhanced prediction accuracy is proposed;
3. a set of optimal and significant features are identified to predict accident severity;
4. the performance of machine learning algorithms, binary class categorization of two accident datasets have been compared and evaluated.

The work has been explicated with the road accidents datasets that are collected from the Traffic Head Quarters, Coimbatore City, Tamilnadu, India for the year 2012 and Fatality Analysis Reporting System (FARS) which is available in the University of Alabama's Critical Analysis Reporting Environment (CARE) system, USA and Road accident training dataset obtained from the STATS19 data collection system, maintained by the government of United Kingdom (UK).

The paper is organized as follows. Section 2 gives a review of previous models studied for the analysis of road traffic accidents. The nature of the input data is described in Section 3 and it provides the necessary details about modeling strategies used in this study. Results are discussed in Section 4 with concluding remarks in Section 5.

2. Application of Data Mining to Road Traffic Accidents

Several researches exhaustively study this issue and a range of interventions have been recommended to improve road safety based on the major factors for crashes such as road, environment, driver and vehicle factors.

To control the most important factors which affect injury severity of drivers involved in traffic crashes on Iran roads the Classification and Regression Tree (CART), has been used to analyze the crash data pertaining to the last three years (2006-2008) [14]. The results have revealed that improper usage of the seat belt, overtaking and speeding are the most important factors associated with injury severity. Many data mining feature selection and classification algorithms have been applied to find the factors influencing the fatal accidents using Fatality Analysis Reporting System (FARS) [22-27].

In [18] the authors have made noticeable attempts at identifying the degree of importance of Information Entropy for road traffic accident analyses. A classification tree based on Gini Index has been presented for analyzing crashes on mid-block segments of multilane arterials of Florida, U.S Route 19 [1]. The study has provided the safety analysis community an additional tool to assess safety without having to aggregate the corridor crash data over arbitrary segment lengths. Adaptive regression trees (CART) have been developed to build a decision support system to handle road traffic accident analysis for Addis Ababa city traffic office [29]. The road traffic accident data of Finland between 2004 and 2008 have been analyzed using descriptive data mining, clustering and association rule mining to create reasonable knowledge [3]. Also, attitude and behavior of driver scores along with other variables such as driving mileage, driver age and personality tend to exhibit statistically significant association with collision involvement [8]. Binomial models have been employed in analyzing hierarchical data structure [15]. The factors involved in motor vehicle crashes have been predicted from two-lane rural intersections in the state of Georgia. The data set has been in a hierarchical structure with respect to driver's characteristics, crash characteristics, site characteristics. An increased risk of accidents has been observed with the consumption of Qat, a locally grown stimulant among road users of Yemen using Smeed's model [2]. Further, literature is available for detailed evaluation of machine learning algorithms such as neural network, decision tree, support vector machines and a hybrid decision tree [4, 6, 9].

Using Feature Mining techniques, irrelevant and redundant features from a dataset will be filtered out so that highly informative features will be provided. CFS method has been used to find the correlation between features and those features have been validated using SVM and ANN models [28]. Various Feature Mining algorithms, classification algorithms such as C4.5, C-RT, CS-MC4, Decision List, ID3, Naive Bayes, Random Tree etc. and ensemble algorithms such as AdaBoost, Arc-X4 etc. have been explored to analyze Road Traffic Accidents data based on road and vehicle specific characteristics [22-27]. Also the Feature Mining algorithms including CFS, FCBF, Feature Ranking, MIFS and MODTree have been explored to improve the classifier accuracy. Interestingly, few influencing factors on road users have also drawn attention to portray the causal relationships. The most effective way to reduce road accident is to better understand the causative road accidents [21]. Hence, data mining literature has a clear road map with an aim of finding causal factors, predicting the future risk with the aid of comparative algorithms. However, a study specific mining has been emphasized in most of the studies because of greater diversity prevail in road accident data. The occurrence of road accidents in India has been considered since very few researches have focused on such studies but with descriptive tools [31]. This study aims on the identification of accident patterns and major accident factors to answer an increasing need of designing preventive measures with the ultimate objective of reducing the number of traffic accidents and fatalities. The present study also emphasizes the use of a newly proposed more comprehensive method that extract information from selected Feature Mining algorithms to identify the significant features and classification algorithms to predict road accident patterns.

3. Methods and Materials

The proposed computational methodology for the prediction of road accident severity is given in the Figure 1.

3.1 Descriptive analysis of Coimbatore accident data set

For this study three datasets have been used; an Indian road accident data set obtained from the Coimbatore City Traffic Head Quarters, Tamilnadu, India for the year 2012, international datasets Fatality Analysis Reporting System

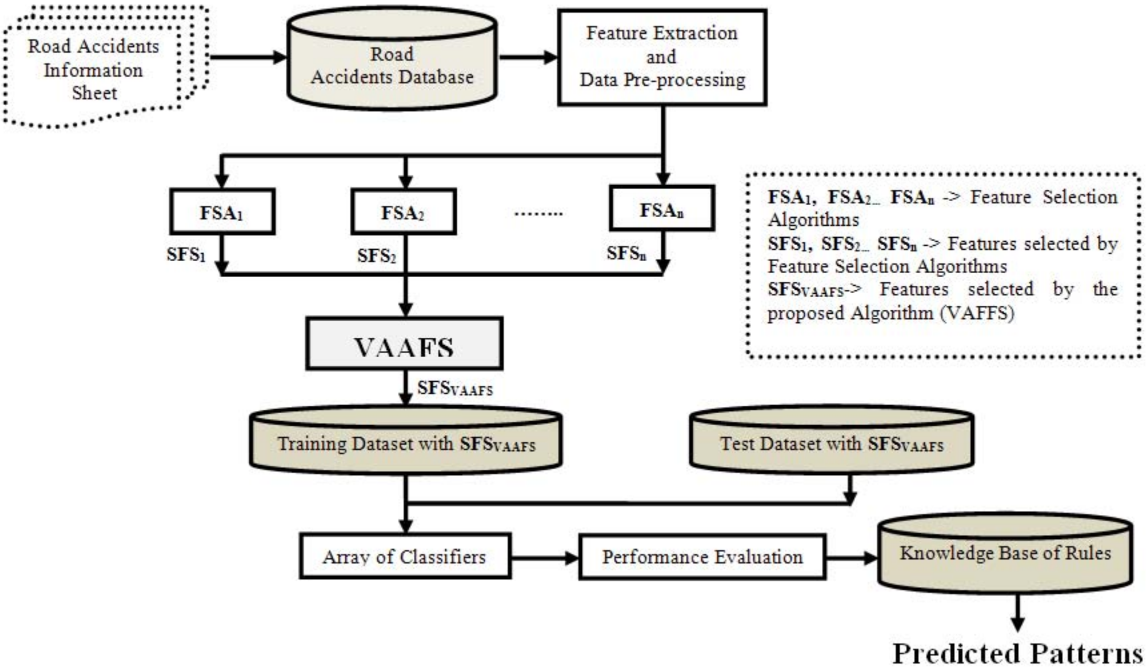

Road Accidents Information Sheet
Road Accidents Database
Feature Extraction and Data Pre-processing
FSA1
FSA2
....
FSAn
SFS1
SFS2
SFSn
FSA1, FSA2_ FSAn -> Feature Selection Algorithms
SFS1, SFS2_ SFSn -> Features selected by Feature Selection Algorithms
SFSVAAFS-> Features selected by the proposed Algorithm (VAFFS)
VAAFS
SFSVAAFS
Training Dataset with SFSVAAFS
Test Dataset with SFSVAAFS
Array of Classifiers
Performance Evaluation
Knowledge Base of Rules
Predicted Patterns

Figure. 1 Generic Computational Approach of the Proposed Methodology

(FARS) which is available in the University of Alabama's Critical Analysis Reporting Environment (CARE) system, USA and Road accident training dataset obtained from the STATS19 data collection system, maintained by the government of United Kingdom (UK). Coimbatore, one of the major cities in Tamilnadu has been identified as a good indicator for different environments and for different geographical regions. The information pertaining to road crashes attributable to Central, East and West regions of Coimbatore city are the main focus of this investigation. The month wise data for the year 2012 has been used to understand the nature and extent of the causes of fatalities and to build models.

The original data set for the study contains traffic accident records for the year 2012, a total number of 438 cases of which 308 proper cases have been identified through the data cleaning exercise. The class attribute, Accident Severity, is a multivariate variable which has three values: Fatal Injury, Grievous Injury and Minor Injury. The values of the variables of the data set are listed in Table 1.

Table 1. Independent Variables and the Respective Values

Variables	Values	Variables	Values
PS_Region $(\mathbf{f}_{1})$	Central Region	Pedestrian $(\mathbf{f}_{14})$	N-No
	East Region		Y-Yes
	West Region	TypeofVehicle $(\mathbf{f}_{15})$	Bus_Private
Month $(\mathbf{f}_{2})$	Jan-Dec		Car_Private
Govt_Holiday $(\mathbf{f}_{3})$	N-No		Cycle
	Y-Yes		Jeep
School_Vacation $(\mathbf{f}_{4})$	N-No		None
	Y-Yes		Two_Wheeler
Addl_Vacation $(\mathbf{f}_{5})$	N-No	Minor $(\mathbf{f}_{16})$	N-No
	Y-Yes		Y-Yes
WeekDay $(\mathbf{f}_{6})$	Monday	Govt_Vehicle $(\mathbf{f}_{17})$	N-No
	Tuesday		Y-Yes
	Wednesday	Type_of_Person $(\mathbf{f}_{18})$	Bus_Govt_Passenger
	Thursday		Bus_Private_Passenger
	Friday		Car_Private_Driver
	Saturday		Car_Private_Passenger
	Sunday		Cyclist
WeekEnd $(\mathbf{f}_{7})$	N-No		Jeep_Passenger
	Y-Yes		Pedestrian
Friday $(\mathbf{f}_{8})$	N-No		Two_Wheeler_Pillion_Rider
	Y-Yes		Two_Wheeler_ Rider
Day $(\mathbf{f}_{9})$	N-No	Self_Accident $(\mathbf{f}_{19})$	N-No
	Y-Yes		Y-Yes
Type_of_Road $(\mathbf{f}_{10})$	NH-National Highway	Vehicle_Hit_By $(\mathbf{f}_{20})$	Bus_Govt
	SH-State Highway		Bus_Private
	CR-Corporation Road		Car_Private
Gender $(\mathbf{f}_{11})$	F-Female		Lorry
	M-Male		Jeep
Age_Coded $(\mathbf{f}_{12})$	0-18: Minor		Two_Wheeler
	19-30: Young		Tata Ace
	31-50:Young Adult		Tempo
	51-60: Adult		Van Private
	>60-Senior Citizen	**Accident_Severity (Class Variable)**	Fatal_Injury
Hit_And_Run $(\mathbf{f}_{13})$	N-No		Grievous_Injury
	Y-Yes		Minor_Injury

It could be observed from the descriptive values of the data that the PS_Region (West region), Weekend(yes), Type_of_Road (National Highways & Corporation Roads), Age_Coded (Young road users in the age range of 19-30), Gender (Male) and Vehicle_Hit_By (the accidents made by Lorry) are more prone to Fatal Accidents. Though the descriptive statistics of the data is showing the contribution of each attributes towards road traffic accidents, we could not find the interrelationship between attributes. Thus to find the association between variables and to select the significant attributes feature selection algorithms have been applied. The details of these algorithms are discussed in the following sections.

3.2 Descriptive analysis of FARS data set and GB data set

We carried out the analysis with road accident dataset obtained from Fatality Analysis Reporting System (FARS). The detailed description of the dataset is available in [22-27, http://www-fars.nhtsa.dot.gov]. GB dataset consists of Great Britain road accident information for the year 2010. It consists of 154410 records and 10 Attributes.

The original data set was in MS-Excel format. All the values of attributes have been represented as continuous values. These values have been replaced with its equivalent categorical information as per the guidelines given by FARS and GB datasets. Thus the continuous attributes have been converted into categorical attributes. These two previous studies have been used to substantiate the proposed model.

3.3 Voting algorithm for aggregated feature selection (VAAFS)

Many factors affect the success of machine learning algorithms on a given task [11]. Most of the studies have exploited the features that are extracted from different Feature Selection algorithms in an ad-hoc manner; such features are subsequently used for classification purposes using appropriate algorithms. Also in such procedures, it has been observed that the misclassification rates of the classifiers are increased than that of the classifiers with original variables. However, the present study aims to improve the classification by a method through which Feature Mining procedures could be optimally utilized and the results can subsequently be extended to the classification algorithm.

In order to identify the significant features the appropriate algorithms viz. CFS[11], Feature Ranking[5], Info Gain [10, 18], Gain Ratio[10], Correlation Attribute Evaluation (CAE)[10], MIFS [10, 22-27] and Relief [16, 33] which performed better [22-27] on the foresaid databases have been used in this study. Also four classification algorithms viz. Random Tree [12,13], Naive Bayes [10, 22-27], J48 [http://weka.waikato.ac.nz] and C4.5 [30] which have been identified as best algorithms from the previous studies [22-27] have been used in this study to model road accident severity.

The new approach, VAAFS is based on an optimum choice of results of more than one Feature Mining algorithm. Also, it is perceived that the common features identified by majority of such algorithms have more impact on the study objective. Hence, voting method is observed as a better amiable alternative to form a set of features which are identified as significant by most of the Feature Selection algorithms. If the votes received by a feature is more than $\lceil n/2 \rceil$ where 'n' is the number of feature selection algorithms, then that feature will be selected by VAAFS as an optimal significant feature. The entire process of prioritizing attributes is based on the pre assigned significant values which is generally 5% wherever applicable. The pseudocode of VAAFS is given below.

```
Algorithm: Voting Algorithm for Aggregated Feature Selection
Variables:
RADS - Road Accidents Data Set
$D_{Train}$ - Training Samples $D_{Test}$
$D_{Test}$ - Testing Samples
FSA - Feature Selection Algorithm
FS - Feature Set
SFS - Significant Feature Set
$V_i$-Vote of the feature $f_i$
Input:
Feature Set FS = $\{ f_1, f_2,...., f_m\}$, m=number of features
Feature Selection Algorithms FSA = $\{ FSA_1, FSA_2, ......., FSA_n\}$, n=number of algorithms
MINVOTE = $\lceil n/2 \rceil$
Intermediate Output:
Significant Feature Set $SFS_i = \{ SFS_1, SFS_2,......., SFS_n\}$ selected by $FSA_i$
Output:
Significant Feature Set $SFS_{VAAFS}$ = Significant Features selected from $SFS_i$ based on voting method.
```

Algorithm: *Voting Algorithm for Aggregated Feature Selection* ***(continued)***

Pseudocode:

```
(1)  Load Training Samples, $D_{Train}$
(2)  Set Cycle = 1
(3)  Repeat
(4)  For each $FSA_i$ give FS as Input
     {
        Apply best feature selection process of $FSA_i$
        $SFS_i = \phi$
        For each $f_i$ in FS
        {
           If $f_i$ is best feature to classify $D_{Train}$
           Then $SFS_i = SFS_i \cup \{f_i\}$
           Increment the vote of $f_i$ , $V_i = V_i + 1$
        }
     }
(5)  cycle=cycle+1
(6)  until cycle=n
(7)  $SFS_{VAAFS} = \phi$
(8)  Set Cycle = 1
(9)  Repeat
(10) For each $SFS_i$
     {
        If $V_i$ is greater than or equal to MINVOTE
        Then $SFS_{VAAFS} = SFS_{VAAFS} \cup \{f_i\}$
     }
(11) cycle=cycle+1
(12) until cycle=n
(13) Perform Classification using $SFS_{VAAFS}$
(14) Evaluate using $D_{Test}$
```

The features that are selected by this way can then be the set of inputs for an array of classification algorithms. To evaluate the performance of the classifiers accuracy, confusion matrix, Precision and Recall are used [10, 22-27].

4. Results and Discussion

After preprocessing the training datasets have been loaded in Weka machine learning software, SPSS statistical package and Tanagra data mining tool in the specified format. The multivariate attribute Accident Severity has been used as the class attribute. The results are discussed in two subsections:

1. Feature Significance Analysis
2. Performance Evaluation of the Classifiers.

The analysis has been conducted on road accidents dataset of Coimbatore city, FARS and Great Britain.

4.1 Feature significance analysis

The algorithm and method presented in the previous section have been applied on road accident dataset of Coimbatore city which has been primarily divided by four regions (All, East, West and Central). Feature Ranking algorithm identified the subset of significant features which has 5% significant level of Chi Square statistics[5], CFS algorithm has chosen the significant features which has highest merit [11], and other algorithms extracted all features which gives the best weight [16, 33]. On ranking the attributes by the recommendations of each algorithm, the set of significant attributes which have been selected by VAAFS from all the regions is depicted in Figure 2.

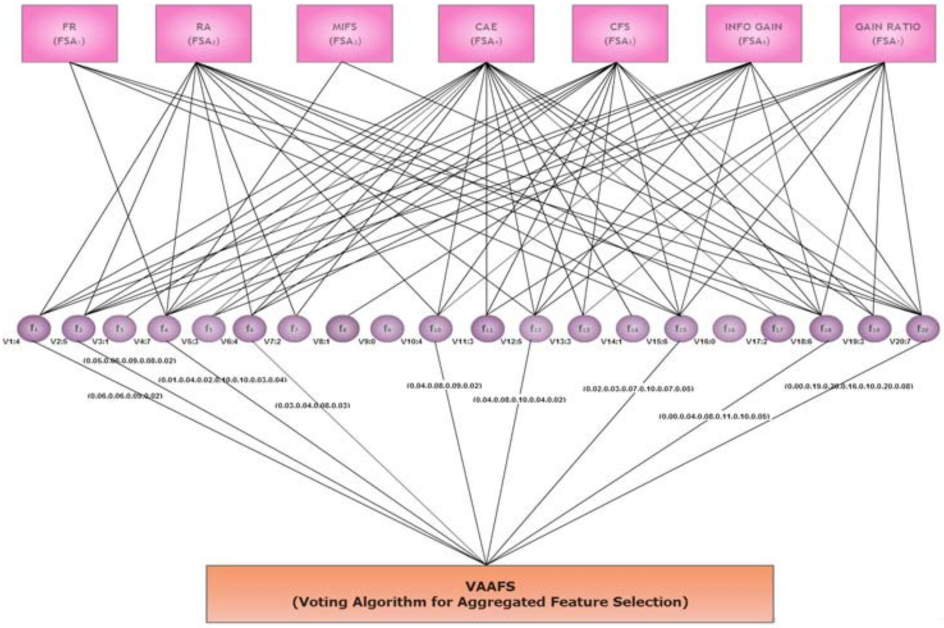


Figure 2. Features Selected by VAAFS

Table 2. Features Selected in Site Specific Data

Variables (f_i)	Feature Selection Algorithms (Number of Features) and Weights								
	FR (4)	RA (11)	MIFS (2)	CAE (17)	CFS (13)	Infogain (10)	Gain ratio (10)	Vote V_i	VAAFS (9)
PS_Region (f_1)	-	0.058	-	0.050	0.087	0.022	-	4	√
Month (f_2)	-	0.049	-	0.050	0.089	0.075	0.021	5	√
Govt_Holiday (f_3)	-	-	-	0.040	-	-	-	1	-
School_Vacation (f_4)	0.005	0.042	0.023	0.103	0.099	0.026	0.038	7	√
Addl_Holiday (f_5)	-	-	-	0.039	0.084	-	0.033	3	-
WeekDay (f_6)	-	0.027	-	0.040	0.083	0.030	-	4	√
WeekEnd (f_7)	-	0.041	-	0.024	-	-	-	2	-
Friday (f_8)	-	-	-	-	-	0.013	-	1	-
Day (f_9)	-	-	-	-	-	-	-	0	-
Type_of_Road (f_{10})	-	0.043	-	0.082	0.086	0.020	-	4	√
Gender (f_{11})	-	-	-	0.057	0.085	-	0.021	3	-
Age_Coded (f_{12})	-	0.038	-	0.083	0.095	0.043	0.021	5	√
Hit_And_Run (f_{13})	-	-	-	0.087	0.088	-	0.023	3	-
Pedestrian (f_{14})	-	-	-	0.087	-		-	1	-
TypeofVehicle (f_{15})	0.019	0.027	-	0.073	0.091	0.069	0.048	6	√
Minor (f_{16})	-	-	-	-	-	-	-	0	-
Govt_Vehicle (f_{17})	-	0.029	-	0.077	-	-	-	2	-
Type_of_Person (f_{18})	0.002	0.043	-	0.079	0.105	0.096	0.049	6	√
Self_Accident (f_{19})	-	-	-	0.079	0.084	-	0.045	3	-
Vehicle_Hit_By (f_{20})	0.000	0.186	0.195	0.158	0.100	0.195	0.077	7	√

The Proposed VAAFS has selected the best features using majority voting method; it has selected the features which have been selected by any four of the above seven algorithms i.e. The minimum number of votes should be equal to $\lceil n/2 \rceil = 4$. If any attribute gets votes >= $\lceil n/2 \rceil$ that will be selected as the significant attribute by VAAFS. In Figure 2 the features selected by VAAFS have been coupled with the weights and each feature is assigned with its votes.

The attributes which have been selected by these Feature Selection algorithms and their optimal measures in each region are listed in Table 2. The symbol '-' indicates that the feature is not selected by either of the algorithms. The symbol '√' indicates that the feature is selected as an optimal feature.

From the results it could be observed that each feature selection algorithms gave different subset of features. Thus it could be understood that the factors influencing accident severity differ geographically.

4.2 Performance evaluation of the classifiers

The performance of the proposed method is substantiated using four classification techniques. The Random Tree, C4.5, Naive Bayes and J48 classifiers are used to evaluate

Table 3. Predictive Accuracy of Classifiers on Coimbatore Accident Dataset

Classification Algorithms	**Without FSA (20)**	**Feature Selection Algorithms (Number of Features)**							
		FR (4)	RA (11)	MIFS (2)	CAE (17)	Infogain (9)	CFS (13)	Gain ratio(10)	VAAFS (9)
Random Tree	92.86	71.10	91.23	68.18	92.86	91.56	92.21	88.96	95.45
C4.5	70.78	67.53	73.05	67.21	70.45	67.53	73.05	69.16	73.05
Naive Bayes	68.83	65.26	69.48	66.23	69.16	69.16	70.78	69.81	73.05
J48	71.75	39.29	71.75	67.53	73.10	71.75	72.40	68.83	75.97

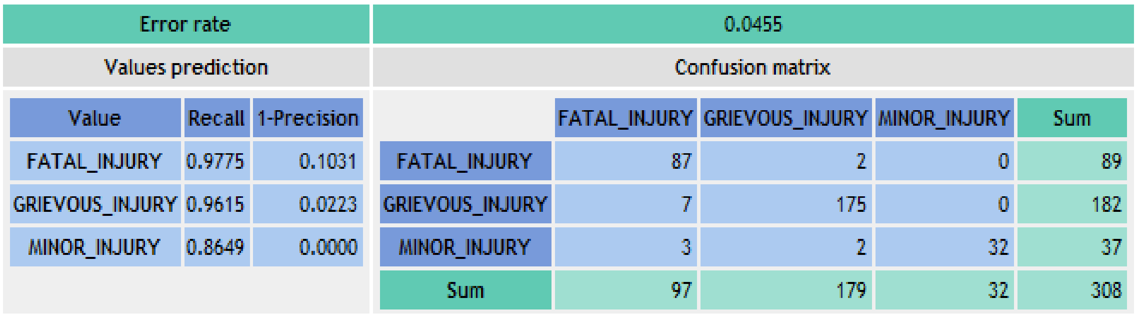
the classification accuracy of the feature selection algorithms. The study results revealed that Random Tree algorithm with the features selected by VAAFS algorithm exhibited better performance in modeling accident severity than other algorithms. The accuracies of the classification algorithms with features selected by the feature selection algorithms are given in the Table 3.

From Table 3 it is clear that in all the regions Random Tree accuracies are high and the features selected by the VAAFS algorithm increase the predictive accuracy of all the classifiers with less number of optimal features set. The contingency table from Random tree algorithm with features selected by VAAFS in the Coimbatore accident dataset and the precision and recall values are is given in the Figure 3.

Classification rules are popular alternative to decision trees in representing the structures that learning methods produce, partly because each rule seems to represent an independent “nugget” of knowledge [32]. Based on such observations this work has extended the feature selection algorithm to find a refined classification rule and following observations have been mode from the generated rules.

During night time most of the young (age group 19-30) male road users are hit by Lorry on national highways which cause fatal accidents. It is a rule which is common for all the regions in the Coimbatore city. Similar other rules have indicated that in the east region during weekend young (age group 19-30) two wheelers are likely to involve in fatal accidents on highways. Whereas in the case of West region most persons involved in the fatal accidents are male two wheel riders and pedestrians who have a higher chance of hit by lorry and car. In central region persons involved in the fatal accidents is more likely between 19 and 50 (young and young adult) during week end and possible hit by bus. These and similar other rules have depicted the distinct pattern of road accidents and likelihood of fatalities when other factors like geography is controlled in the analysis.

The accuracy of feature selection algorithms using classifiers have been evaluated with the international road accident training datasets obtained from Fatality Analysis Reporting



Error rate	0.0455

Values prediction

Value	Recall	1-Precision
FATAL_INJURY	0.9775	0.1031
GRIEVOUS_INJURY	0.9615	0.0223
MINOR_INJURY	0.8649	0.0000

Confusion matrix

	FATAL_INJURY	GRIEVOUS_INJURY	MINOR_INJURY	Sum
FATAL_INJURY	87	2	0	89
GRIEVOUS_INJURY	7	175	0	182
MINOR_INJURY	3	2	32	37
Sum	97	179	32	308

Figure 3. Contingency Matrix of Random Tree algorithm with VAAFS

Table 4. Comparison of Predictive Accuracy of VAAFS with Previously Reported Work

Reference & Dataset	# of Samples	# of Features	Previous Work			Proposed Work (VAAFS+RT)	
			Algorithm Found to be Best	*# of Features*	*Accuracy (%)*	*# of Features*	*Accuracy (%)*
[27], FARS	37259	57	FR+RT	20	94.38	16	**95.12**
[26], FARS	77125	16	RT	16	95.05	15	**95.24**
[25], FARS	63327	17	RT+Adaboost	17	95.59	12	**96.41**
[24], GB	154410	10	FR+RT	10	86.51	8	**86.62**
[22], FARS	277125	33	FR+RT	29	94.83	18	**97.73**

System (FARS) and Great Britain. Table 4 details and compares the predictive accuracy obtained from the proposed method and with that of the previous works. From Table 4 it could be observed that Random Tree classifier works better with VAAFS in modelling road traffic accidents. It effectively identifies the significant features from an aggregated selection of feature selection algorithms and yields a reduced misclassification rate and high accuracy. Figure 4 gives the comparison between the existing work and the present work carried out on various accident data sets.

Thus the proposed algorithm could be used effectively to improve predictive accuracy of the classification techniques with a set of optimal significant features while modeling road traffic accidents.

Conclusion

It is crucial for engineers to be able to extract useful information from existing data to analyze the causes of traffic accidents as it causes great impact on society. The main purpose of this study is to identify the optimal set of features for modeling road accidents severity using feature selection and classification techniques. This research work investigated the performance of the classification algorithms by suitably optimized information obtained from a list of feature

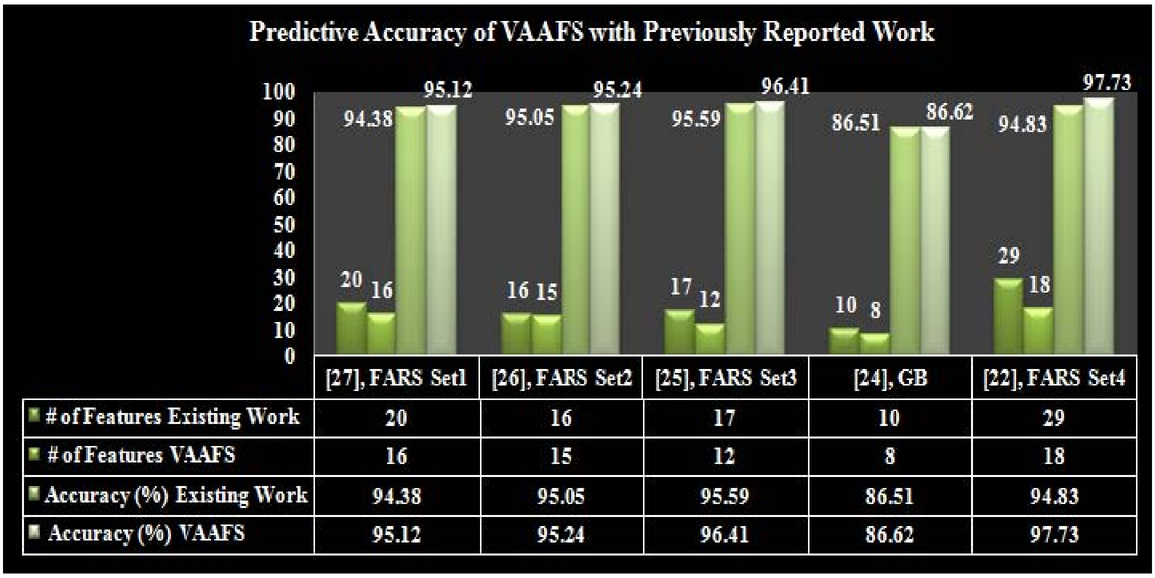


Figure 4. Predictive Accuracy of VAAFS with Previously Reported Work

mining algorithms. The motivation of such attempt is to alleviate the comparison based decision of various methods for obtaining most appropriate study features in a typical data mining application. The widely applicable voting method has been used to succeed in this attempt by selecting most involved features in the classification algorithm to improve the predictive accuracy. Such a multitude search motivates the proposed algorithm called VAAFS, Voting Algorithm for Aggregated Feature Selection has been applied to understand the accident severity and most important factors which influences fatal accidents happened on the roads of Coimbatore City, Tamilnadu, India and on the roads of US and Great Britain, so that by eliminating or controlling such factors an overall safety improvement can be accomplished. This study could be considered as an evidence for the effective use of newly proposed algorithm for its simplicity in applying and to obtain more efficient output as a data mining tool.

Acknowledgement

The authors are grateful to the Coimbatore City Traffic Head Quarters, Coimbatore, Tamilnadu India for supporting this research and offering all the resources that are available to them.

REFERENCES

1. PANDE, A., M. ABDEL-ATY, A. DAS, **A Classification Tree Based Modelling Approach for Segment Related Crashes on Multilane Highways,** Journal of Safety Research, vol. 41, 2010, pp. 391-397.
2. AMEEN, J. R. M., J. A. NAJI, **Causal Models for Road Accident Fatalities in Yemen,** Accident Analysis and Prevention, Elsevier, vol. 33, 2001, pp.547-561.
3. AYRAMO, S., P. PIRTALA, J. KAUTTONEN, K. NAVEED, T. KARKKAINEN, **Mining Road Traffic Accidents**, Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering, No. C. 2/2009.
4. PRATO, C. G., V. GITELMAN, S. BEKHOR, **Mapping Patterns and Characteristics of Fatal Road Accidents in Israel Prato**, Proceedings of the 12th WCTR Conference, July 11-15, Lisbon, Portugal, 2010.
5. CHERNOFF, H., E. L. LEHMANN, **The Use of Maximum Likelihood Estimates in $\chi 2$ Tests for Goodness of Fit,** Annals Mathematical Statistics, vol. 25, issue 3, 1954, pp. 579-586.
6. CHONG, M., A. ABRAHAM, M. PAPRZYCKI, **Traffic Accident Analysis Using Machine Learning Paradigms,** Informatica, vol. 29, 2005, pp. 89-98.
7. CIOS, K., W. PEDRYCZ, R. SWINIARSKI, **Data Mining Methods for Knowledge Discovery,** Boston: Kluwer Academic Publishers, 1998.
8. DARBY, P., W. MURRAY, R. RAESIDE, **Applying Online Fleet Driver Assessment to Help Identify, Target and Reduce Occupational Road Safety Risks,** Safety Science, Science Direct, vol. 47 2009, pp. 436-442.
9. DURDURAN, S. S., **A Decision Making System to Automatic Recognize of Traffic Accidents on the Basis of a GIS Platform,** Expert Systems with Applications, vol. 37, 2010, pp. 7729-7736.
10. HAN, J., M. KAMBER, **Data Mining; Concepts and Techniques,** Morgan Kaufmann Publishers, 2000.
11. HALL, M. A., L. A. SMITH, **Feature Selection for Machine Learning Comparing a Correlation based Filter Approach to the Wrapper**, Proceedings of the 12th International Florida Artificial Intelligence Research Society Conf., 1998.
12. JAMES, M., **Classification Algorithms,** John Wiley, 1985.
13. XI, J., Z. GAO, S. NIU, T. DING, G. NING, **A Hybrid Algorithm of Traffic Accident Data Mining on Cause Analysis,** Mathematical Problems in Engineering, 2013, dx.doi.org/10.1155/2013/302627.
14. KASHANI, A. T., A. SHARIAT-MOHAYMANY, A. RANJBARI, **A Data Mining Approach to Identify Key Factors of Traffic Injury Severity,**

Promet – Traffic & Transportation, vol. 23, no. 1, 2011, pp. 11-17.

15. KIM, D., Y. LEE, S. WASHINGTON, K. CHOI, **Modelling Crash Outcome Probabilities at Rural Intersections: Application of Hierarchical Binomial Logistic Models,** Accident Analysis and Prevention, Elsevier, vol. 39, 2007, pp. 125-134.

16. KIRA, K. L. A. RENDELL, **A Practical Approach to Feature Selection**, Proceedings of the 9th International Conference on Machine Learning (ICML 1992), 1992.

17. LUOMA, J., M. SIVAK, **Characteristics and Availability of Fatal Road-Crash Databases Worldwide**, The University of Michigan, Transportation Research Institute, 2006.

18. R. MARUKATAT, **Structure-based Rule Selection Framework for Association Rule Mining of Traffic Accident Data**, Computational Intelligence and Security, vol. 4456, 2007, pp. 231-239.

19. NABI, H., L. R. SALMI, S. LAFONT, M. CHIRON, M. ZINS, E. LAGARDE, **Attitudes Associated with Behavioural Predictors of Serious Road Traffic Crashes: Results from the Gazel Cohort**, Injury Prevention, vol. 13, no. 1, 2007, pp. 26-31.

20. SINGH, R. K., S. K. SUMAN, **Accident Analysis and Prediction of Model on National Highways**, International Journal of Advanced Technology in Civil Engineering, ISSN: 2231–5721, vol. 1, Issue 2, 2012, pp. 25-30.

21. ARYANI SOEMITRO, R. A., Y. S. BAHAT, **Accident Analysis Assessment to the Accident Influence Factors on Traffic Safety Improvement Case: Palangka Raya Tangkiling National Road,** Proceedings of the Eastern Asia Society for Transportation Studies, Vol. 5, 2005, pp. 2091-2105.

22. SHANTHI, S., R. GEETHA RAMANI, **Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques,** Proceedings of IAENG-World Congress on Engineering and Computer Science, San Francisco, USA, vol. 1, 2012, pp. 122-127.

23. SHANTHI, S., R. GEETHA RAMANI, **Vehicle Safety Device (Airbag) Specific Classification of Road Traffic Accident Patterns through Data Mining Techniques,** Springer Publications: Advances in Intelligent Systems and Computing, Proceedings of the Second International Conference on Advances in Computing and Information technology, Chennai, vol.177, 2012, pp. 433-443.

24. SHANTHI, S., R. GEETHA RAMANI, **A Comparative evaluation of Classification Methods in the Prediction of Road Traffic Accident Patterns,** Proceedings of the International Conference on Future Communication and Computer Technology, Beijing, China, ISBN: 978-988-15121-4-7, 2012.

25. SHANTHI, S., R. GEETHA RAMANI, **Gender Specific Classification of Road Accident Patterns through Data Mining Techniques**, IEEE International Conference on Advances in Engineering, Science and Management, March 30-31, 2012, pp. 359-369, ISBN: 978-81-909042-2-3.

26. SHANTHI, S., R. GEETHA RAMANI, **Classification of Seating Position Specific Patterns in Road Traffic Accident Data through Data Mining Techniques,** Second International Conference on Computer Applications, ICCA 2012, vol. 5, 2012, pp. 98-104.

27. SHANTHI, S., R. GEETHA RAMANI, **Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms,** International Journal of Computer Applications, vol. 35, No. 12, 2012, pp. 30-37.

28. SOHN, S. Y., S. H. LEE, **Data Fusion, Ensemble and Clustering to Improve the Classification Accuracy for the Severity of Road Traffic Accidents in Korea,** Safety Science, vol. 4. issue 1, 2003, pp. 1-14.

29. TESEMA, T. B., A. ABRAHAM, C. GROSAN, **Rule Mining and Classification of Road Traffic Accidents**

using Adaptive Regression Trees, Journal of Simulation, vol. 6, issues 10 & 11, 2005.

30. QUINLAN, R, **C4.5: Programs for Machine Learning,** Morgan Kaufmann Publishers: San Mateo, CA., 1993.

31. VALLI, P. P., **Road Accident Models for Large Metropolitan Cities of India,** IATSS Research, vol. 29, issue 1, 2005, pp. 57-65.

32. WITTEN, I. H. E. FRANK, **Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations,** San-Francisco, Morgan Kaufmann Publishers, 2000.

33. SUN, Y., **Iterative Relief for Feature Weighting: Algorithms, Theories, and Applications,** IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, issue 6, 2007.