ADVANCES in NATURAL and APPLIED SCIENCES

ISSN: 1995-0772
EISSN: 1998-1090
2016 November10(16): pages 1-11

Published BY**AENSI Publication**
http://www.aensiweb.com/ANAS
Open Access Journal

Learning Through Misclassified Instances Using Pipelined Voting Algorithm for Aggregated Feature Selection (P-VAAFS) in the Prediction of Road Accident Severity

[1]Geetha Ramani and [2]Shanthi Selvaraj

[1]Associate Professor, Department of Information Science and Technology, Anna University, Guindy Campus, Chennai, Tamilnadu, India.
[2]Associate Professor, Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore, Tamilnadu, India.

Received 23 August 2016; Accepted 1 November 2016; Published 20 November 2016

Address For Correspondence:
Shanthi Selvaraj, Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore, Tamilnadu, India,

Copyright © 2016 by authors and American-Eurasian Network for Scientific Information (AENSI Publication).
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/

Open Access

ABSTRACT

Background: To formulate a suitable function for measuring the relevance between different factors of the heterogeneous, high dimensional and sparse nature of the road accident datasets. **Objectives:** To predict the influencing features of fatal accidents by computing feature selection and classification based on misclassified samples in pipelined stages. **Results:** The performance of the proposed approach is evaluated using an Indian Road Accident Dataset obtained from the Coimbatore City Traffic Head Quarters, Tamilnadu, India and with an International dataset obtained from Fatality Analysis Reporting System (FARS), USA. Simulation result shows that the proposed approach helps to improve the discriminating power of the feature selection algorithms and to predict influencing factors of fatal accidents with regard to the target class fatal severity using the pipelined misclassified instances. **Conclusion:** The proposed method selects features iteratively by pipelining the misclassified samples to the subsequent stages, rather than the whole sampling space, thus it can exactly measure the relevance between the feature and the class labels by following the selection procedure. Also the predictive performance of the classification algorithms improved significantly.

KEYWORDS: Classification, Data mining, Feature Selection, Road Accident Dataset, Misclassification, Pipelining, Voting

INTRODUCTION

Fatal rates due to road accidents contribute more on the total death rate of the world (WHO, 2009). Generally the accident data are heterogeneous, sparse and the factors which influence the fatal accidents are high dimensional and site specific. Because of the sparse nature, the data are highly imbalanced. In recent years, the class imbalance problem has emerged as one of the great challenges of data mining [27]. The use of ensemble algorithms is one of the solutions to handle imbalanced datasets [12].

Road accident fatality rate depends on many factors and it is a very challenging task to investigate the dependencies between the attributes because of the many environmental and road accident factors. It is crucial to extract these factors so that it can help improve the overall driving safety situation, not only by preventing accidents but also by reducing their severity [9][13]). From the literature, it is very difficult to define the set of factors which influence the fatal accidents because the accidents are site specific in nature.

Application of data mining techniques on social issues has been a popular technique of late. Feature selection and classification algorithms are important methods in analyzing road accidents. Finding an

To Cite This Article: Geetha Ramani and Shanthi Selvaraj., Learning Through Misclassified Instances Using Pipelined Voting Algorithm for Aggregated Feature Selection (P-VAAFS) in the Prediction of Road Accident Severity. **Advances in Natural and Applied Sciences**. 10(16);Pages: 1-11

appropriate Feature selection algorithm and to improve the classifiers accuracy is an active research area in machine learning for high dimensional dataset analysis. Also a learning model designed using the sparse, imbalanced and large set features will have higher computational cost, slower learning process and poorer classification accuracy due to the phenomenon known as the curse of dimensionality [8].

Feature selection algorithms generally fall into two broad categories, the filter model and the wrapper model [26]. The feature selection algorithms select important features by calculating the relevance on the entire sampling space only once and these features will be used for further prediction [8]. It is very difficult to extract the informative features from the sparse and skewed data in a single iteration. Thus the learning using multi stage ensemble algorithms has become a very active area of research [10] [14] [24].

Thus to improve the understanding of the intrinsic features of such dataset, we studied the following important problems in this paper:

i) Usefulness of "insignificant features" to improve the overall accuracy of the learning algorithm.
ii) Using "misclassified samples" to extract the hidden informative features.

In this paper, we proposed a Pipelined Voting Algorithm for Aggregated Feature Selection (P-VAAFS) which is an iterative learning algorithm based on feature selection and combination voting. It extracts the important features, initially using all the samples and these features will be used by a set of classifiers for the prediction. Unlike the traditional way of discarding the insignificant features and misclassified samples, we iteratively build classifiers using misclassified sample spaces. The misclassified samples out of the best classifier in the first iteration will be pipelined for the next iterative process. Such an iterative pipelined process will learn the hidden information in different feature spaces, and automatically stop when the learning algorithm cannot provide better information or the numbers of misclassified samples are less. Simulation results on the Indian and International dataset show this method can improve the classification accuracy and the discriminative power of the feature selection algorithms with certain levels of trade-off of the computational cost.

The structure of rest of the paper is as follows: Section 2 gives a review of application of data mining algorithms in the analysis of road accidents data. Section 3 briefly introduces the method of road accident data profiling. In Section 4, architecture and steps of the proposed algorithm is presented. Details of simulations conducted using an Indian and an International data sets and the results are reported in Section 5. Concluding remarks are given in Section 6.

State-of-the-art:

A review of the latest and related work on road accident data analysis is concisely presented in the following section.

The road accident data are collected from heterogeneous sources so that it contains a lot of features which affects injury severity. This high dimensional heterogeneous data may hide the significant variables during feature selection process. The recent literature on road accident data reveals the multitude aspects of data collection mechanisms, methods to analyze and varied conclusions. Many studies have been tried to identify the most significant variables which influence the fatal accidents [3] [4] [11] [25].

Methods varying from log-linear model [4], Poisson regression [13], network models such as artificial neural networks [5] and Bayesian networks [15], CART or its variant [2] [3] [9] [16] have been used to identify the significant factors.

Further studies that apply specific methods related to the accident data analysis can also be found; Decision Tree [1], Ensemble Classification Algorithms [12], Association Analysis [11], bootstrap-based iterative method [14], and K-modes clustering [17].

Most or all these and similar other studies have attempted to highlight the importance of choosing specific methods to analyze the accident data and identify the major causing attributes that are available within the data set considered. This helps to identify the varied nature of data sets that are prevalent in different countries so as to apply relevant techniques.

Conclusions of such varied studies are mainly focusing on the causal agents for accidents and / or severity of accidents; for example, late-night or early-morning driving increases the risk for bus drivers of being severely injured, importance of considering intersections in crash models [4]; vehicle type, Pedestrians, motorcycle and bicycle riders are identified to have higher risks of being injured than other types of vehicle drivers in traffic accidents [3]; exceeding speed limits on rural roads, risky use of cellular phone and sleepy driving [13]; driver's characteristics, vehicle type, and roadway conditions [25]; not using the seat belt, improper overtaking and speeding [9]; accident type, age, atmospheric factors, gender, lighting, number of injured, and occupant [15]; intersections at community areas forest area and agriculture land area that are aside of certain highways are dangerous for two wheeler accidents; Intersections on highways, vehicle roll-over accidents, multi-vehicular and fixed object/divider hit [17].

On the other hand, methodological comparisons also provide useful information to the choice of techniques adopted in different studies; limitation of Poisson regression is addressed earlier studies [2] [3] and recommended CART as a good alternative method for analyzing freeway accident frequencies; yet, subsequent

studies adopt Poisson regression method [13] [25] have observed that when the sample data size is small, with proper prior setting, the Bayesian Ordered Probit (BOP) model can produce more reasonable parameter estimations and better prediction performance.

These kinds of complex mechanism involved in the analysis of accident data provide scope to study the techniques for such analyses in a more comprehensive way. This includes understanding the available data structure and formats, variables involves and their nature, sparseness in the data and / or sample size issues, and computing algorithms and platforms.

Motivated with such array of functionalities, authors of present study have approached various methods related to different data sets. The authors [23] have used feature selection algorithms including CFS, FCBF, Feature Ranking, MIFS and MOD Tree, Classification Algorithms such as C4.5, C-RT, CS-MC4, Decision List, ID3, Naïve Bayes and Random Tree, Arc-X4 Meta classifier [18]; Random Tree using ensemble methods viz. Bagging, Ada Boost and Multi Cost Sensitive Bagging [19]; Random Tree, C4.5, J48 and Decision Stump [20] [21]and C4.5, CART, ID3, Naïve Bayes, Random Tree [22]. Subsequently a procedure, Voting Algorithm for Aggregated Feature Selection (VAAFS), has been developed that works on the basis of an aggregated voting to the problem of feature selections [6]. Also a method has been developed to study the possible association between the important variables involved in a data set related to road accidents together with amount of heterogeneity using Random Effect Model [7].

Therefore, researchers in the field of traffic accidents and specifically in the domain of traffic accident injury severity focused their research on trying to identify the hidden, most significant variables which influence the injury severity [15].

Dataset Description:

Due to the heterogeneous nature of the datasets with two different set of site specific features we have used two datasets to evaluate the model; an Indian road accident dataset obtained from the Coimbatore City Traffic Head Quarters, Tamilnadu, India for the year 2014 and an International road accident dataset of California obtained from Fatality Analysis Reporting System (FARS) which is available in the University of Alabama's Critical Analysis Reporting Environment (CARE) system, USA. The features in both the datasets are different. The features in Coimbatore Dataset are person related which covers both deceased and accused information and the features in the International dataset includes environmental and vehicle specific features.

The Coimbatore dataset for the study contains traffic accident records for the year 2014, a total number of 19 features and 342 fatal cases have been used for the study. The class attribute Fatal_Severity is a multivariate variable which has 5 values: Level_1, Level_2, Level_3, Level_4 and Level_5 in the increasing order of severity. The features and their values of Coimbatore dataset are given in Table 1.

Table 1: Feature Set and their Values of Coimbatore Dataset

Feature#	Attribute	Values
f1	fatal_severity	Level_1, Level_2, Level_3, Level_4, Level_5
f2	ps_region	East, West, Central
f3	govt_holiday	AH, BH, H, WD
f4	season	Autumn, Monsoon, Prevernal, Spring, Summer, Winter
f5	weekday	Sunday, …, Saturday
f6	time_slot	0-3, 15-18, 18-21, 21-24, 12-15, 6-9, 9-12, 3-6
f7	type_of_road	CR, NH, SH
f8	gender_d	M, F
f9	age_group_d	Adult, Minor, Old_Adult, Seniors, Young_Adult
f10	type_of_person_d	Bus_Passenger, Car_Driver, Cyclist, Jeep_Rider, Pedestrian, Two_Wheeler_Rider
f11	direction_d	E-W, N-S, S-N, W-E
f12	accident_type	Fixed_Object_Collision, Govt_Vehicle, Others, Pedestrian, Private_Vehicle, Rollover
f13	direction_a	E-W, N-S, NONE, S-N, W-E, Unknown
f14	angle_of_collision	Opposite_Direction, Same_Direction, Side_Swipe_Left, Side_Swipe_Right, Unknown
f15	hit_run	Y, N
f16	gender_a	Not_Known, M, F
f17	age_group_a	Not_Known, Adult, Minor, Old_Adult, Seniors, Young_Adult
f18	type_of_person_a	Ace_Driver, Bus_Govt_Driver, Bus_Private_Driver, Car_Driver, Crane_Driver, Jeep_Driver, Lorry_Driver, Maxi_Cab_Driver, Not_Known, Tempo_Driver, Two_Wheeler_Rider, Van_Driver
f19	type_of_vehicle_a	Ace, Bus_Govt, Bus_Private, Car, Crane, Jeep, Lorry, Maxi_Cab, Not_Known, Tempo, Two_Wheeler, Van

The California dataset for the study contains traffic accident records for the years 2008-2010, a total number of 25 features and 1680 fatal cases have been used for the study. The class attribute Injury_Severity is a multivariate variable which has 6 values: Fatal, Incapacitate, None, Possible_Injury and Nonincapacitate, Unknown. The features and their values of the dataset are listed in Table 2.

Table 2: Feature Set and their Values of California Dataset

Feature #	Attribute	Values
f1	Injury_Severity	Fatal, Incapacitate, None, Possible_Injury, Nonincapacitate, Unknown
f2	Month	Jan,…,Dec
f3	Day	Sunday, …, Saturday
f4	Time	Night, Noon, Day, MidNight, Early_Morning, Evening, Unknown
f5	Manner_of_Collision	Angle-Front-to-Side,None,Front-to-Front,Front-to-Rear,sideswipe_Opposite_Direction,Other,UnknownSideswipe_Same_Direction,Rear-to-Side,Rear-to-Rear
f6	Person_Type	Driver,Bicyclist,Passenger,Pedestrian,Occupant,Other_Cyclist,Unknown,Persons_In/On_Buildings
f7	Seating_Position	Front_Seat,Second_Seat,Third_Seat,Others,Vehicle_Exterior,Fourth_Seat
f8	Age_Range	20_24_years,35_44_years,45_54_years,75_above,65_74_years,55_64_years,25_44_years,13_19_years,Unknown,10_12_years,0_4_years,5_9_years
f9	Gender	Male, Female
f10	Race	White,None,Black,Mexican,American_Indian,Hispanic,South_American,Puerto_Rican,Others,Asian,Korean,Spanish,Cuban,Chinese,Japanese, Indian, Filipino
f11	Transported_for_Treatment	EMS, None, Others, Unknown, Law_Enforcement
f12	Air_Bag	Unknown, Front, None, Multiple_Directions, Roof, Side, Other
f13	Protection_System	None, Lap_Shoulder_Belt, Unknown, Child_Safety_Seat, Helmet, Lap_Belt, Improper_Use, Shoulder_Belt, Booster_Seat
f14	Ejection	None,Not_Applicable,Totally_Ejected,Partially_Ejected,Unknown,Ejected
f15	Ejection_Path	None,Back_Window,Unknown,Side_Door,Side_Window,Roof,Windshield,Back_Door
f16	Extrication	None, Extricated, Unknown
f17	Dead_on_Arrival	None, At_Scene, At_Route, Unknown
f18	Time_of_Death	Noon, Unknown, Night, Day, Midnight, Early_Morning, Evening
f19	Fatal_Injury_at_Work	No, Not_Applicable, Yes, Unknown
f20	Alcohol_Test	Not_Reported,Observed,Evidential,Breath,Other,Behavioral,Passive_Alcohol_Sensor,
f21	Alcohol_Test_Method	None, Blood, Both, Unknown, Urine, Other
f22	Drug_Test	Notest,Unknown,Depressant,Cannabinol,Narcotic,Stimulant,Other,Hallucinogen,No_Drug,Inhalant,Phencyclidine
f23	Drug_Involvement	No, Unknown, Yes, Not_Reported
f24	Accident_Location	None, Non_Intersection, Intersection, Unknown
f25	Related_Factors	Drowsy/Sleepy/Asleep/Fatigued,Emotional,Ill,Nonmedication,Injured_Before,Others,Drugs/Alcohol,WheelChair,None,Aggressive,Mentally_Challenged,Physical_Impairment,Careless

Though each attribute shows the contribution towards road traffic accident severity, we could not find the relation between attributes. Thus to find the association between variables and to select the significant attributes feature selection algorithms have been applied in iterative pipelined manner.

Methodology:

A novel methodology, 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 has been implemented [6]. The optimal features selected by VAAFS will be then passed to the classifiers over an Indian road accident data set obtained from the Coimbatore City Traffic Head Quarters, Tamilnadu, India for the year 2013 and with international dataset obtained from Fatality Analysis Reporting System (FARS), USA, to model the accident severity. The pseudo code 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.*
Pseudo code:
(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
{
Iff$_i$ 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
{
IfV_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}

VAAFS calculates the relevance on the entire samples in a single iteration which cannot exactly identify the informative features. Thus P-VAAFS has been proposed to handle misclassification instances using iterative and pipelined feature selection and classification algorithms. In this work optimal features have been selected based on VAAFS and the selected optimal features are pipelined to a set of classifiers to validate the model. The misclassified instances out of the best classifier have been passed to the feature selection phase for further iterative and pipelined process. This process is repeated until the number of misclassified instances is equal to α which is equal to the number of folds used for cross validation or there is no improvement in the classification accuracy. The performance of the proposed approach is evaluated using 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. Experimental results show that the proposed method selects highly informative features and produces comparable classification accuracy than the other approaches. The proposed computational methodology for the prediction of road accident severity is given in the Figure 1. The terminologies used in Figure 1 is defined as follows.

RADS - Road Accidents Data Set
D_{Train} - Training Samples
D_{Test} - Testing Samples
FSA_i – Array of Feature Selection Algorithms
SFS_i - Significant Feature Set Selected by FSA_i
SFS_{VAAFS} = Significant Features selected from SFS_i based on voting method.
CA_i- Array of Classification Algorithms
Acc_i – Accuracy of CA_i
CA_b- Best Classifier Selected Based on Acc_i
D_m - Misclassified Samples out of CA_b
D_c - Classified Samples out of CA_b
$SFS_{P\text{-}VAAFS}$ - Significant Features Set selected from iterative P-VAAFS
α - number of folds used for cross validation.

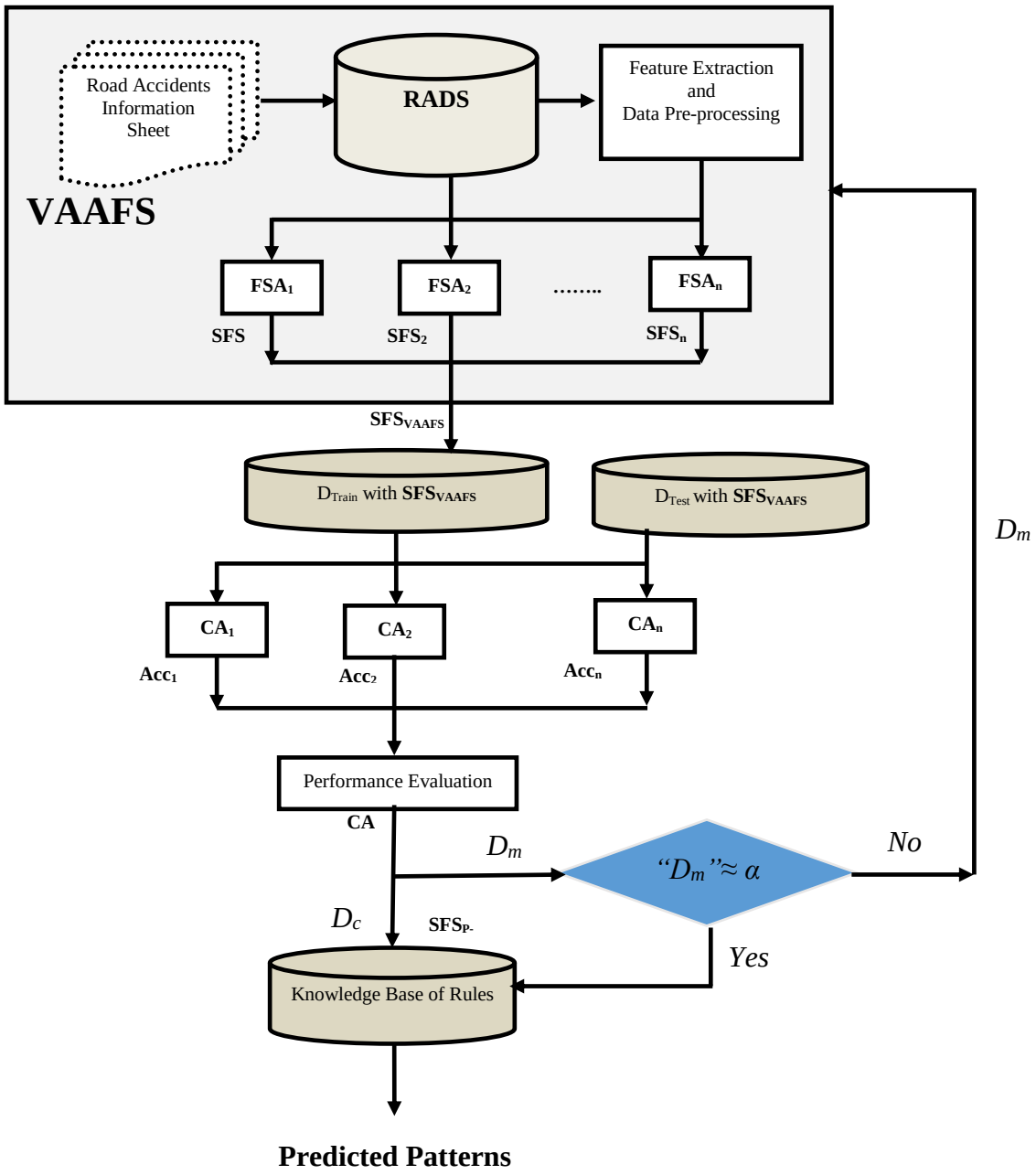

Fig. 1: Proposed Methodology

The pseudo code of P-VAAFS is given below.

Algorithm: *Pipelined Voting Algorithm for Aggregated Feature Selection (P-VAAFS)*
Variables:
RADS - Road Accidents Data Set
D_{Train} *- Training Samples*
D_{Test} *- Testing Samples*
CA-Classification Algorithm
CA_b*-Best Classifier*
D_m *- Misclassified Samples*
D_c *- Classified Samples*
Input:
Feature Selection Algorithms FSA = $\{FSA_1, FSA_2, \ldots\ldots, FSA_n\}$, *n=number of algorithms*
Classification Algorithms CA = $\{CA_1, CA_2, \ldots\ldots, CA_n\}$, *n=number of algorithms*
Intermediate Output:
Significant Feature Set $SFS_i = \{SFS_1, SFS_2, \ldots\ldots, SFS_n\}$ *selected by* FSA_i
Significant Feature Set SFS_{VAAFS} *= Significant Features selected from* SFS_i *based on voting method.*
Output:

Significant Feature Set $SFS_{P\text{-}VAAFS}$ = Significant Features selected from iterative P-VAAFS.
Pseudocode:
(15) Load RADS, Split the samples into D_{Train} and D_{Test}..
(16) For each CA_i
{
Classify the samples using D_{Train} and D_{Test}.
Observe Classified and Misclassified Samples.
Calculate and Display the Accuracy of the i^{th} Classifier, Acc_i
}
(17) Set Stage = 1
(18) $D_m = D_{Train}$
(19) Repeat
(20) For each FSA_iandwith all Samples in D_m
{
Call VAAFS( );
For each CA_i
{
Train CA_i using SFS_{VAAFS}.
Observe Classified and Misclassified Samples.
Calculate and Display the Accuracy of the i^{th} Classifier, Acc_i
}
}
(21) $SFS_{P\text{-}VAAFS} = SFS_{P\text{-}VAAFS} \cup SFS_{VAAFS}$
(22) Evaluate the performance of the Classifiers CA_i and select the best classifier CA_b.
(23) D_c = Classified samples from CA_b
(24) D_m = Misclassified samples from CA_b
(25) Store the patterns predicted from the classified samples D_c
(26) Pipeline $D_m$for the next stage of iteration
(27) stage=stage +1
(28) until$D_m \approx \alpha$ or there is no significant improvement in the classification accuracy

Note: Here α is the number of folds used for cross validation.

RESULTS AND DISCUSSION

The proposed approach is implemented using Java and executed in a PC with Intel Core i5 processor with 1.8 GHz speed and 6GB of RAM. Experiments are conducted in Java Eclipse IDE to demonstrate the effectiveness of the proposed approach in selecting the significant features which are prone to fatal accidents.

1.1 Road Traffic Accidents Dataset:

This section presents the details of simulations carried out using an Indian Road accident dataset and an international dataset. The details of the datasets used for the study are described in Section 3.

1.2 Classification Accuracies Without Feature Selection:

A set of Decision Tree classification algorithms viz. Random tree, J48, NaiveBayes, and Decision Stump have been applied on both the Indian and International datasets. These algorithms have used the entire feature set of both the datasets. The accuracies of the decision tree algorithms without feature selection are listed in Table 3.

Table 3: Classification Accuracy without Feature Selection Algorithms

Dataset	Classifier	# Samples	# CI	# MCI	Accuracy
Indian (Coimbatore) Road Accident Dataset	Random Tree	342	164	178	47.95
	J48	342	169	173	49.42
	Naïve Bayes	342	152	190	44.44
	Decision Stump	342	198	144	57.89
International (California) Road Accident Dataset	Random Tree	1680	1331	349	79.23
	J48	1680	1498	182	89.17
	Naïve Bayes	1680	1453	227	86.49
	Decision Stump	1680	1103	577	65.65

From Table 3 it could be observed that both J48 and Decision Stump algorithms have given better accuracy than the other algorithms with all the features. The features selected by these decision tree algorithms are listed in Table 4.

Table 4: Features Selected by Decision Tree Algorithms

Decision Tree Algorithm	Indian (Coimbatore) Road Accident Dataset	International (California) Road Accident Dataset
Random Tree	f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f17, f18, f19	f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f17, f18, f19, f20, f21, f23
J48	f4, f5, f7, f8, f9, f10, f11, f12, f14, f18	f2, f3, f4, f5, f6, f7, f8, f9, f11, f12, f13, f19, f21, f23
NaiveBayes	f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f17, f18, f19	f2, f3, f4, f5, f6, f7, f10, f11, f12, f13, f14, f15, f16, f17, f18, f19, f20, f21, f22, f23, f24, f25
DecisionStump	f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f17, f18, f19	f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f19, f20, f21, f23, f24, f25

All the decision tree algorithms except J48 have selected most of the features to evaluate the model. However, it is not possible to focus on the significant features so as to reduce the fatal accidents. Thus to select the significant features which are influencing the fatal accidents, we have explored on feature selection algorithms.

1.3 Feature Selection Using all the Samples:

Initially, feature selection algorithms viz. CFS Subset, Chi Squared, Gain Ratio, Info Gain, CFS, Relief and Machine Learning Ensemble have been applied on all the samples of road accidents datasets. Each feature selection algorithm selects different set of features. The optimal number of most significant features are selected using voting method. If any attribute gets votes >= $\lceil n/2 \rceil$ (where n is the number of feature selection algorithms) that will be selected as the significant attribute by VAAFS [6]. The optimal features are selected and stored in SFS_{VAAFS}. The features selected by the feature selection algorithms are listed in Table 5.

Table 5: Optimal Feature Set Selected by VAAFS

Feature Section Algorithm	Indian (Coimbatore) Road Accident Dataset	International (California) Road Accident Dataset
CFS Subset	f2, f3, f4, f5, f6, f7, f9, f10, f11, f12, f13,f17, f18	F10, f11, f16, f18, f19
ChiSquared	f4, f5, f12, f18, f19	f10, f11, f17, f18, f19
GainRatio	f3, f4, f5, f10, f11, f12,f18, f19	f10, f11, f16, f17, f18, f19, f21, f24
InfoGain	f4, f5, f6, f10, f12, f18, f19	f10, f11, f13, f14, f17, f18, f19
CFS	f2, f5,f8,f10, f11, f13, f14,f18, f19	f10, f11, f13, f14, f17, f18, f19, f21, f24
Relief	f5, f11,f12, f13, f18, f19	f10, f11, f17, f18, f19
ML Ensemble	f2, f3, f4, f5, f6, f16, f17, f18	f2, f3, f4, f5, f6, f17, f18
VAAFS	f4, f5, f12, f18, f19	f10, f11, f17, f18, f19

From the Table 5 it could be observed that the attributes season, weekday, accident type, type of the accused person and the type of vehicle used by the accused person from the Indian dataset and the human race, the mode of transportation to treat the victim after the accident, how quick the impact of the fatal accident, Time of death and Fatal Injury at Work from the International dataset are the significant features which influence the fatal accidents. The features selected by the VAAFS are the optimal features which could represent the whole sampling space.

1.4 Classification Using all the Samples:

The samples with the selected features in SFSVAAFS are evaluated using the decision tree classification algorithms viz. Random tree, J48, Naïve Bayes, and Decision Stump. 10 Folds Cross Validation is used to evaluate the model. The accuracies of the feature selection algorithms with feature set is given in Table 6.

Table 6: Predictive Accuracies of the Classifiers with SFS_{VAAFS}

Dataset	Classifier	# Samples	# CI (D_c)	# MCI (D_m)	Accuracy
Indian (Coimbatore) Road Accident Dataset	Random Tree	342	175	167	51.2
	J48	342	174	168	50.9
	Naïve Bayes	342	137	205	40.1
	Decision Stump	342	198	144	57.9
International (California) Road Accident Dataset	Random Tree	1680	1489	191	88.6
	J48	1680	1489	191	88.6
	Naïve Bayes	1680	1486	194	88.5
	Decision Stump	1680	1103	577	66.7

The classification accuracies of the classifiers with the features selected by VAAFS have been improved compared to the model without feature selection. However, the features selected and the classification accuracies at this stage are based on the evaluation of the algorithms on the whole sampling space. This cannot exactly represent the amount of information of features over misclassified samples. Thus the information metric should be re-estimated only on misclassified samples, rather than the whole sample space.

1.5 Pipelined Feature Selection and Classification Using Misclassified Samples:

Now, as discussed in Section 4, the misclassified samples, D_m from the best classifiers are pipelined for the next iteration. Once again the feature selection algorithms are applied on only misclassified samples, D_m with all the features. Based on the features selected, again the classification algorithms are applied to evaluate the model. From the classification results produced, the road accident samples are again divided into classified (D_c) and misclassified samples (D_m) and this process is repeated until there are "α" number of samples in the misclassified category or there is no further improvement in the classification accuracy. The number of misclassified samples in each stage is observed for both the road accident datasets. The performance of the classifiers in each stage is represented in Table 7 and Table 8.

Table 7: Performance of the P-VAAFS Using Coimbatore Road Accident Dataset

Classifier	Stage 1			Stage 2			Stage 3		
	# CI (D_c)	# MCI (D_m)	Acc (%)	# CI (D_c)	# MCI (D_m)	Acc (%)	# CI (D_c)	# MCI (D_m)	Acc (%)
Random Tree	175	167	51.2	103	41	71.5	28	10	73.7
J48	174	168	50.9	106	38	73.6	23	15	60.5
Naïve Bayes	137	205	40.1	97	47	67.4	26	12	68.4
Decision Stump	198	144	57.9	106	38	73.6	29	9	76.3

Table 8: Performance of the P-VAAFS Using California Road Accident Dataset

Classifier	Stage 1		
	# CI (D_c)	# MCI (D_m)	Acc (%)
Random Tree	1489	191	88.6
J48	1489	191	88.6
Naïve Bayes	1486	194	88.5
Decision Stump	1103	577	65.7

From the Table 7 we could observe that the proposed approach reaches the minimum number of misclassified samples in three stages using Coimbatore road accident dataset. Also there is a gradual increase in the accuracy. Except J48 the accuracies of all other algorithms improved at third stage whereas the improvement of the accuracy of J48 stopped at stage 2. In Table 8, the proposed approach left with minimum of 191 misclassified samples for the California road accident dataset. This is because even if the feature selection is done in many stages it has not given a considerable improvement after the first stage in the classification accuracy. So the process is stopped with 2 stages. Except J48 the accuracies of all other algorithms improved at first stage whereas there is no improvement in the accuracy of J48. As the number of misclassified samples are getting reduced in every stage, it is confirmed that the proposed methodology of selecting optimal feature selection by P-VAAFS using the misclassified samples effectively represent the significance between the features and the injury severity.

The classification performance produced by the classifiers along with the optimal features in the final stage is reported in Table 9.

Table 9: Predictive Accuracies of the Classifiers with $SFS_{P\text{-}VAAFS}$

Dataset	# Stages	# Optimal Features	Optimal Features ($SFS_{P\text{-}VAAFS}$)	Accuracy (%)		
				Without Feature Selection	VAAFS	P-VAAFS
Indian (Coimbatore) Road Accident Dataset	3	9	govt_holiday (f3), Season (f4), Weekday (f5), time_slot (f6), direction_d (f11), accident_type (f12), age_group_a (f17), type_of_person_a (f18), type_of_vehicle_a (f19)	57.9	73.6	76.3
International (California) Road Accident Dataset	1	10	month (f2), Manner_of_Collision (f5), Seating_Position (f7), Age_Range (f8), Race (f10), Transported_for_Treatment (f11), Protection_System (f13), Dead_on_Arrival (f17), Time_of_Death (f18), Fatal_Injury_at_Work (f19)	79.2	88.6	88.6

Experimental results in Table 9 show the significant features extracted from the misclassified samples out of best classifiers pipelined through various stages. Also the proposed method is significantly good and achieves better classification accuracy (from 57.9% to 76.3% for Indian dataset and from 79.2% to 88.6% for the International dataset) with optimal significant features (9 & 10) and in minimum number of stages (3 & 1). From this, it can be asserted that the proposed method is able to obtain significant features in stages and reliable as the classifier is able to deliver the best performance. Also the study proves that the proposed method is highly successful in reducing misclassification error rates of the classifiers and improve the predictive accuracy with optimal features.

Comparison of Proposed Method with the Existing Methods is given in Figure 2 (i) and Figure 2 (ii).

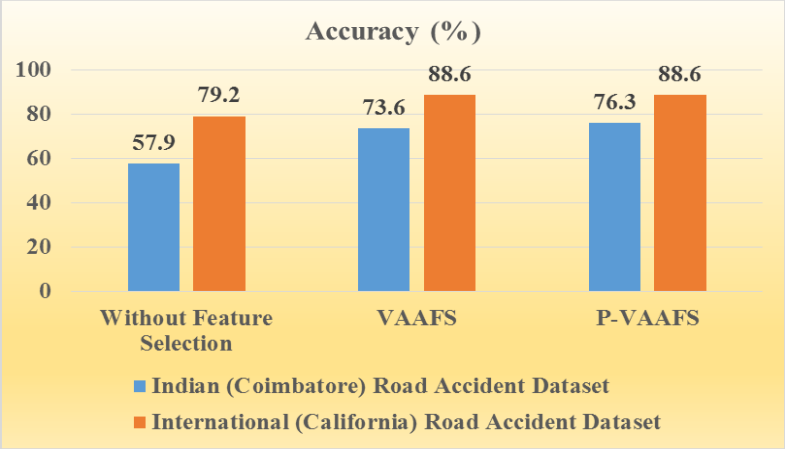

Fig. 2(i): Accuracies of Proposed Method and Existing Methods

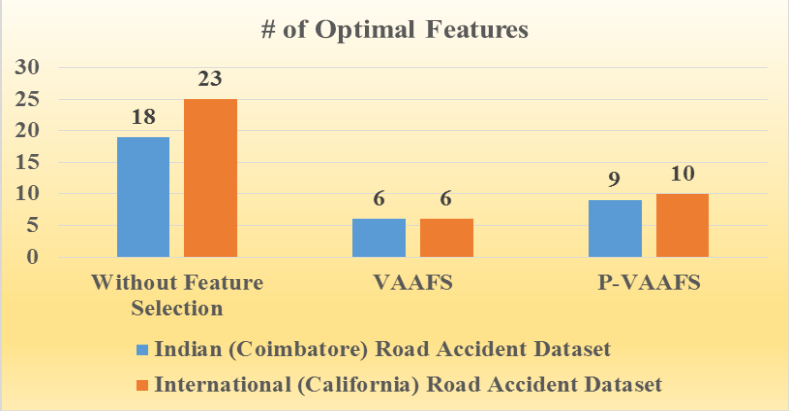

Fig. 2(ii): Optimal Features Selected by the Proposed Method and Existing Methods

Conclusion:

The bottleneck of analyzing road accident data is the identification of significant features because of its high dimensional and sparse nature. This paper proposed a novel approach for highly significant feature selection iteratively in stages based on the classification performance produced by various classification algorithms. Since the proposed method selects features iteratively by pipelining the misclassified samples to the subsequent stages, rather than the whole sampling space, it can exactly measure the relevance between the feature and the class labels by following the selection procedure. The effectiveness of the proposed approach has been demonstrated using two heterogeneous, high dimensional and sparse data sets obtained from the Coimbatore City Traffic Head Quarters, Tamilnadu, India and from Fatality Analysis Reporting System (FARS), USA. In all the cases the proposed approach selects a highly significant, optimal features, which contribute to fatal accidents. These features could not be identified in single iteration with all the samples but with misclassified samples using the pipelined approach the hidden informative features could be identified. Also the classification accuracies of the classifiers are significantly improved (from 57.9% to 76.3% for Indian dataset and from 79.2% to 88.6% for the International dataset). Though the proposed method identifies hidden, significant features with improved accuracies, the complexity of the proposed method is little high compared with the traditional methods. This could be considered as future area of study.

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. Abellan, J., G. Lopez, J. Ona, 2013. Analysis of traffic accident severity using decision rules via decision trees. Expert System with Applications: Elsevier, 40: 6047-6054.
2. Chang, L.Y., W.C. Chen, 2005. Data mining of tree based models to analyse freeway accident frequency. Journal of Safety Research, Elsevier, 36(4): 365-375.

3. Chang, L.Y. and H.W. Wang, 2006. Analysis of traffic injury severity: An application of nonparametric classification tree techniques. Accident Analysis and Prevention, 38: 1019-1027.
4. Chen, W.H, and P.P. Jovanis, 2000. Method for identifying factors contributing to driver-injury severity in traffic crashes. Transportation Research Record, 1717: 1-9.
5. Delen, D., R. Sharda and M. Bessonov, 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention, 38: 434-444.
6. Geetha Ramani and Shanthi Selvaraj, 2014. A Pragmatic Approach for Refined Feature Selection for the Prediction of Road Accident Severity. Studies in Informatics and Control, 23(1): 41-52.
7. Geetha Ramani and Shanthi Selvaraj, 2016. A Novel Approach to Analyze a Combination of IxJ Categorical Data for Estimating Road Accident Risk. Asian Journal of Information Technology, 15(12): 2005-2015.
8. Han, J. and Kamber, M., 2012. Data Mining: Concepts and Techniques, Academic Press.
9. T. Kashani, A. Shariat-Mohaymany, A. Ranjbari, 2011. A Data Mining Approach to Identify Key Factors of Traffic Injury Severity.Promet – Traffic & Transportation, 23(1): 11-17.
10. Kittler, J., F. Roli, 2000. Multiple Classifier Systems, Proceedings of First International Workshop. Lecture Notes in Computer Science, Springer.
11. Kopelias, P., F. Papadimitriou, K. Papandreou and P. Prevedouros, 2015. Urban Freeway Crash Analysis. Transportation Research Record: Journal Transportation Research Board, pp: 123-131.
12. Lopez, A., S. Fernandez, V. Garcia, Palade and F. Herrera, 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250: 113-141.
13. Nabi, H., L. RachidSalmi, S. Lafont, M. Chiron, M. Zins, E. Lagarde, 2007. Attitudes associated with behavioural predictors of serious road traffic crashes: results from the GAZEL cohort. Injury Prevention, 13(1): 26-31.
14. Oliveira, D.V.R., T.N. Porpino, G.D.C. Cavalcanti, Tsang IngRen, 2015. A bootstrap-based iterative selection for ensemble generation. 2015 International Joint Conference on Neural Networks (IJCNN), 1-7.
15. Randa Oqab Mujalli and Juan de Ona, 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of safety research, 42: 317-326.
16. Rovsek, V., M. Batista, B. Bogunovic, 2014. Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree. Transport. UK: Taylor and Francis.
17. Sachin Kumar and Durga Toshniwal, 2015. A data mining framework to analyze road accident data. Journal of Big Data, 2: 26.
18. Shanthi, S., and R.G. Ramani, 2012(a). 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, 1: 122-127.
19. Shanthi, S., and R.G. Ramani, 2012 (b). Vehicle Safety Device (Airbag) Specific Classification of Road Traffic Accident Patterns through Data Mining Techniques. Proceedings of the Second International Conference on Advances in Computing and Information technology, Chennai, Springer Publications, 177: 433-443.
20. Shanthi, S., and R.G. Ramani, 2012 (c). 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.
21. Shanthi, S., and R.G. Ramani, 2012 (d). Gender Specific Classification of Road Accident Patterns through Data Mining Techniques, IEEE-International Conference on Advances in Engineering, Science and Management, Nagapattinam, 359-369, ISBN: 978-81-909042-2-3.
22. Shanthi, S., and R.G. Ramani, 2012 (e). Classification of Seating Position Specific Patterns in Road Traffic Accident Data through Data Mining Techniques. Second International Conference on Computer Applications, Pondicherry, 5: 98-104.
23. Shanthi, S., and R.G. Ramani, 2011. Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms. International Journal of Computer Applications, 35(12): 30-37.
24. Wozniak, M., M. Grana and E. Corchado, 2014. A survey of multiple classifier systems as hybrid systems. Information Fusion, 16: 3-17.
25. Xie, Y., Y. Zhang and F. Liang, 2009. Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. Journal of Transportation Engineering ASCE, 135(1): 18-25.
26. Xing, E., M. Jordan, R. Karp, 2001. Feature selection for high-dimensional genomic microarray data. Brodely, C.E., Danyluk, A.P. (Eds.). Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 601-608.
27. Yang, Q. and X. Wu, 2006. 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making, 5(4): 597-604.