Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques

S. Shanthi, R. Geetha Ramani

Abstract— **This research work emphasizes the significance of Data Mining classification algorithms in predicting the factors which influence the road traffic accidents specific to injury severity. It precisely compares the performance of classification algorithms viz. C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree, applied to modelling the injury severity that occurred during road traffic accidents. Further we applied feature selection methods to select the relevant road accident related factors and Meta classifier Arc-X4 to improve the accuracy of the classifiers. Experiment results reveal that the Random Tree based on features selected by Feature Ranking algorithm and Arc-X4 Meta classifier outperformed the individual approaches. The results have been evaluated using the accuracy measures such as Recall and Precision. In this research work we used the road accident training dataset which was obtained from the Fatality Analysis Reporting System (FARS), provided by the University of Alabama's Critical Analysis Reporting Environment (CARE) system.**

Index Terms— **Road Traffic Accidents, Classification, Feature Selection, Meta Classifier, Accuracy Measures.**

I. INTRODUCTION

The major reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [10]. The information and knowledge [10] gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. Data mining techniques include association, classification, prediction, clustering etc. Classification algorithms are used to classify large volume of data and to provide interesting results.

Application of data mining on social issues has been a popular technique recently. Fatal rates due to road accidents contribute more on the total death rate of the world. Over 1.2 million people [8] die each year on the world's roads and between 20 and 50 million suffer non-fatal [1]injuries. The report by [8] says that around 1.2 million people were killed and 50 million injured in traffic collisions on the roads around the world each year and was the leading cause for death among children 10 – 19 years of age. Many road related factors which increase the death ratio were discussed in various literatures.

In our research work we have classified the road traffic accidents based on injury severity using classification algorithms such as C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree. We have also used feature selection methods viz. CFS, FCBF, MIFS, MODTree and Feature Ranking algorithms to select the relevant features needed for injury severity specific classification and Arc-X4 Meta classifier to improve the accuracies of the classifiers. The focal point of this research work is to compare the accuracies of the classification algorithms with and without using feature selection and Meta classifier.

The literature surveys related to the work are discussed in Section II. The experimental design is presented in Section III. It includes the training data description, system design, classification algorithms, feature selection algorithms, Arc-X4 Meta classifier and accuracy measures. The experimental results have been discussed in Section IV and Section V concludes the paper.

II. RELATED WORK

This section discusses various studies which have been conducted to emphasize the use of classification algorithms, feature relevance algorithms and Meta learners.

In 2009 the accuracy of data mining techniques viz. discriminant analysis, logistic regression, Bayes classifier, nearest neighbor, artificial neural networks, and classification trees has been investigated in analyzing customers' default credit payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods [12]. The results reveal that artificial neural network is the only one that can accurately estimate the real probability of default credit payments.

CART model has been modeled to find the relationship between injury severity and driver/vehicle characteristics, highway/environment variables, and accident variables in Taiwan accident data from the year 2001 [5].

Logistic regression model and classification tree method have been compared in determining social-demographic risk factors which have affected depression status of women in separate postpartum periods [11]. They projected that Classification tree method gives more information than logistic regression model with details on diagnosis by evaluating a lot of risk factors. A comparative study has been conducted between data mining and statistical techniques by varying the number of independent variables, the types of independent variables, the number of classes of

Manuscript received March 13, 2012; revised June, 1, 2012.

S. Shanthi, is with Rathinam Technical Campus, Affiliated to Anna University, Coimbatore, India (phone: 9952001614; e-mail: psshanthiselvaraj@gmail.com).

Dr. R. Geetha Ramani, is with College of Engineering Guindy campus, Anna University, Chennai, India (e-mail: rgeetha@yahoo.com).

the independent variables, and the sample size [21]. 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 variables increased.

FCBF algorithm [1] is designed for high dimensional data and has been shown effective in removing both irrelevant features and redundant features. The limitation of Mutual Information Feature Selector (MIFS) has been analyzed in [16] and proposed a method to overcome this limitation. The basics and implementation of various feature selection algorithms have been discussed in [13].

The combination of the AdaBoost and random forests algorithms were used for constructing a breast cancer survivability prediction model [15]. It was proposed to use random forests as a weak learner of AdaBoost for selecting the high weight instances during the boosting process to improve accuracy, stability and to reduce over-fitting problems [15].

Several voting algorithms, including Bagging, AdaBoost, and Arc-x4, have been studied using decision tree and Naïve Bayes to understand why and when these algorithms affect classification error [6].

The accuracies of simple classification algorithms such as C4.5, C-RT, CS-MC4, Decision List, ID3, Naïve Bayes and Random Tree have been evaluated using the accuracy measures such as Precision, Recall and ROC curve [20].

The accuracy of classifiers using feature selection algorithms has been compared and the results have shown that Random Tree using Feature Ranking algorithm better performs other algorithms in modeling the vehicle collision patterns in road accident data [19].

In this research work we focused on finding accident patterns in road accident data based on injury severity using various classification algorithms. Next section illustrates the methodology used in our research work which includes training dataset description, classification algorithms, feature selection algorithms and Arc-X4 Meta Classifier algorithm.

III. Experimental Setup

This research work focuses on finding road accident patterns specific to injury severity. The existing classification algorithms viz. C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree are adopted for the classification. Among the base classifiers Random Tree algorithm produces classification results with 14.2% misclassification rate.

Since all the selected classifiers gave high misclassification rate, feature selection algorithms have been incorporated with all the classifiers to select the relevant features for classification followed by Meta learning algorithm Arc-X4 to improve the classifier's accuracy. The details of the work are given in the following sub sections.

A. Training Dataset Description

The aim of this research work is to study the classification patterns based on injury severity in road traffic accidents. We carry out the experiment with road accident training dataset obtained from Fatality Analysis Reporting System (FARS) [7] which is provided by Critical Analysis Reporting Environment (CARE) system. This safety data consists of U.S. road accident information for 56 states from the year 2005 to 2009. It consists of 457549 samples and 33 attributes.

To train the classifiers we have selected accident details of the year 2009 which consist of 77125 samples as training dataset with 33 attributes for 56 states. Training data set is used to train the model and to formulate the rules which are used to make decisions. The test dataset is used to validate the rules obtained from trained classifier using new data. Table I provides the list of attributes and their descriptions.

TABLE I
Training Data Attributes Description

Attributes	Description
Key_Value	Identifier to identify an accident
State	State in which the accident occurred
County	County in which the accident occurred
Month	Month in which the accident occurred
Date	Date on which the accident occurred
Day	Day of a week on which the accident occurred
Time	Time at which the accident occurred
Harmful_Event	First harmful event occurred during accidents
Manner_of_Collision	Manner of collision
Person_Type	Driver/Passenger
Seating_Position	Seating Position
Age	Age of the person involved
Age_Range	Age range of the person involved
Gender	Male/Female
Race/Ethnicity	Nationality of People involved in accident
Injury_Severity	Injury Severity
Transported_for_Treatment	Mode of transport used to send the victims for treatment
Air_Bag	Location of the airbag
Protection_System	Type of the protection system used
Ejection	Whether the person is ejected out of the vehicle or not during accident
Ejection_Path	Path from which the persons were ejected
Extrication	Whether the person is extricated
Dead_on_Arrival	Whether dead on arrival for treatment
Year_of_Death	Year of death
Month_of_Death	Month of death
Time_of_Death	Time of death
Fatal_Injury_at_Work	Fatal Injury
Alcohol_Test	Alcohol Test Type
Alcohol_Test_Method	Alcohol Test Method
Drug_Test	Drug Test
Drug_Involvement	Police Reported Drug Involvement
Accident_Location	Accident Location
Related_Factors	Road related factors

In this study we have applied the classification algorithms using Injury_Severity attribute as the class attribute. The class attribute Injury_Severity takes six values as target values. The distribution of the class values of training dataset is given in Table II.

TABLE II
Injury_Severity Attribute Class Values Distribution

Values	Number of Instances	Percentage
Fatal	34175	44.3
Incapacitate	8979	11.6
Nonincapacitate	10329	13.4
Possible_Injury	5829	7.6

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Values	Number of Instances	Percentage
None	17086	22.2
Unknown	727	0.9

From Table II it is clear that approximately 45% of the accidents lead to fatal injury. So it is necessary to find the vital factors which are related to fatal accidents.

The training dataset used to learn the road accident patterns, is preprocessed to handle missing values. All the missing values have been filled with appropriate values. 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 in [7]. Thus the continuous attributes have been converted into categorical attributes. The training dataset after conversion is loaded as text file in Tanagra data mining tool.

In the first experiment we have applied the basic classification algorithms viz. C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree on the preprocessed training data, to predict road accident patterns based on injury severity. The results produced by the classifiers are very high in error rates which are reduced in the subsequent processes.

In the second experiment we have applied the feature selection algorithms such as CFS, FCBF, MIFS, MODTree and Feature Ranking algorithms to select the relevant features related to the study. The classification algorithms have been then applied with the relevant features selected by the feature selection algorithms which result in reduced misclassification rates.

In the third experiment to further improve the classifier's accuracy we have applied the Meta classifier Arc-X4 with the classification algorithms with relevant features resulted in second experiment. The error rates of the classification algorithms are significantly reduced here.

The results from experiment I, II and III have been evaluated using Precision and Recall values. The patterns or rules obtained using training data is evaluated using test data for its correctness.

B. Classification Algorithms

Classification trees are used to predict the classes of a categorical dependent variable from their measurements on one or more predictor or independent variables [10]. Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Decision tree analysis is one of the main techniques used in Data Mining. In our research work we have used the classification algorithms Viz. Iterative Dichotomiser 3 (ID3) [18], C4.5 [17], Classification and Regression Trees (CR-T) [10], Random Tree [14], Naïve Bayes, Cost-Sensitive Classification and Regression Trees (CS-CRT) and Cost-Sensitive Classification using M-Estimate (CS-MC4) which are widely used in the literature.

C. Feature Selection Algorithms

Classification and prediction need to be preceded by relevance analysis which attempts to identify attributes that do not contribute to the classification or prediction process [10]. These attributes can then be excluded. The feature selection algorithms such as Correlation Based Feature Selection (CFS) [9], Fast Correlation Based Filter (FCBF) [22], Mutual Information Feature Selector (MIFS) [2], Feature Ranking and Multi valued Oblivious Decision Tree Filtering (MODTree) have been considered for our study.

D. Meta Classifier: Arc-X4 (Adaptive Resample and Combining Algorithm)

Methods for voting classification algorithms, such as bagging and AdaBoost, have been shown to be very successful in improving the accuracy of certain classifiers [4]. The Arcing term is used to describe the family of algorithms that Adaptively Resample data and Combine the outputted hypotheses [3].

E. Accuracy Measures

In our study we used precision and recall as accuracy measures to evaluate the accuracies of classifiers. It can be calculated from contingency table [10]. The high values of precision and recall denote high accuracy.

IV. Experimental Results

We have used Tanagra tool for our experimental study. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. We have conducted three experiments in our study. The results obtained in each experiment are discussed below.

A. Experiment I: Error Rates of Classifiers without using Feature Selection Algorithms

In this experiment we have used the classifiers viz. C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree without using feature selection methods to find the accident patterns in road accident data based on Injury_Severity. The error rate of C4.5 classifier is depicted in the Fig. 1. The error rate of C4.5 classifier is 15.38% thus the accuracy of the same is 84.62%.

Error rate			0.1538							
Values prediction			Confusion matrix							
Value	Recall	1-Precision		Fatal	Incapacitate	None	Possible_Injury	Nonincapacitate	Unknown	Sum
Fatal	0.9999	0.0000	Fatal	34173	0	2	0	0	0	34175
Incapacitate	0.6652	0.3695	Incapacitate	0	5973	51	279	2663	13	8979
None	0.9786	0.1289	None	0	36	16721	99	131	99	17086
Possible_Injury	0.2690	0.3892	Possible_Injury	0	908	1470	1568	1866	17	5829
Nonincapacitate	0.6274	0.4227	Nonincapacitate	0	2467	767	598	6480	17	10329
Unknown	0.4773	0.2961	Unknown	0	89	184	23	84	347	727
			Sum	34173	9473	19195	2567	11224	493	77125

Fig.1. Error Rate of C4.5 Classifier

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Similarly the error rates of all the classifiers are portrayed in Fig. 2. It shows that the misclassification rate of Random Tree is very less compared with all other classifiers and that of CS-CRT is very high. Thus Random Tree gives better accuracy among all the classifiers.

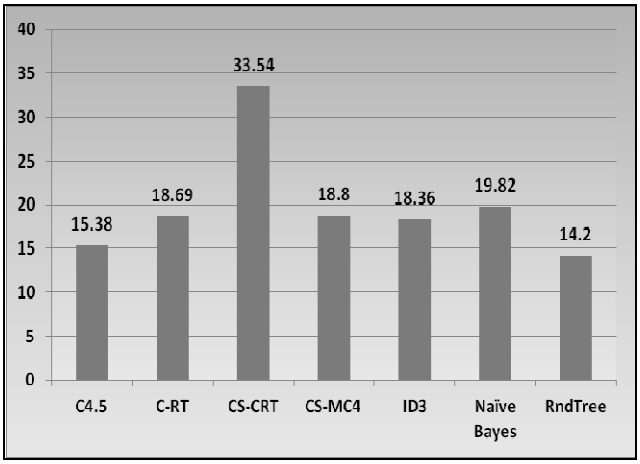

Fig.2. Error Rates of Classifiers without Feature Selection

B. Experiment II: Error Rates of Classifiers with Feature Selection Algorithms

In Experiment-I we have applied the classification algorithms with all the 33 attributes which have produced very high error rates. To reduce the error rates we have to eliminate the irrelevant features [10]. To select the relevant features and to eliminate the irrelevant features we have used feature selection algorithms viz. CFS, FCBF, MIFS, MODTree and Feature Ranking.

For example the attributes selected by Feature Ranking algorithm are shown in Fig. 3. It has selected 29 attributes among 33 attributes as relevant attributes to find road accident patterns based on the attribute Injury_Severity as class attribute.

Feature Ranking algorithm uses CHI-2 criterion. It ranks the input attributes according to the relevance. It does not allow the redundancy of the attributes. It uses p-value to rank the relevant attributes. The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. The irrelevant attributes will be rejected when the p-value is less than the significance level α, which is often 0.05 or 0.01. When the null hypothesis is rejected, the result is said to be statistically significant.

	Before filtering	33
	After filtering	29
N°	Attribute	Values	Statistic
1	Year_of_Death	4	0.51
2	Fatal_Injury_at_Work	4	0.51
3	Time_of_Death	7	0.43
4	Month_of_Death	14	0.35
5	Dead_on_Arrival	4	0.34
6	Transported_for_Treatment	5	0.33
7	Race/Ethnicity	21	0.32
8	Drug_Test	6	0.23
9	Ejection	6	0.22
10	Protection_System	9	0.21
11	Age_Range	12	0.17
12	Extrication	3	0.15
13	Person_Type	9	0.14
14	Accident_Location	4	0.13
15	Air_Bag	7	0.13
16	Drug_Involvement	4	0.12
17	Seating_Position	26	0.11
18	Manner_of_Collision	10	0.11
19	Harmful_Event	57	0.11
20	Ejection_Path	8	0.11
21	Alcohol_Test	7	0.08
22	Related_Factors	13	0.08
23	Drug_Test	11	0.08
24	Gender	2	0.07
25	State	52	0.06
26	Time	7	0.05
27	Day	7	0.02
28	Date	31	0.02
29	Month	12	0.02

Fig.3. Relevant Attributes Selected by Feature Ranking Algorithm

With the selected attributes of different feature selection algorithms we have applied classification algorithms. The error rates of classifiers with Feature Ranking algorithm is given in Table III.

TABLE III
ERROR RATES OF CLASSIFIERS WITH FEATURE RANKING ALGORITHM

Classifier	Error Rates (%)
C4.5	13.06
C-RT	18.08
CS-CRT	21.5
CS-MC4	17.29
ID3	17.3
Naïve Bayes	18.65
Random Tree	5.17

For instance the error rate of CR-T algorithm with relevant attributes selected by Feature Ranking algorithm is given in Fig. 4.

Without feature selection algorithm the misclassification rate CS-CRT algorithm in Experiment-I is 33.54%. But with relevant attributes selected by Feature Ranking algorithm the misclassification rate of CS-CRT algorithm is reduced to 21.54% in Experiment-II. The misclassification rate of CS-CRT algorithm is reduced by 12% using feature selection algorithm. Using this algorithm the misclassification rate increases when classifying the tuples which belong to the class value “Incapacitate”.

Error rate			0.2154						
Values prediction			Confusion matrix						
Value	Recall	1-Precision		Fatal	Incapacitate	None	Possible_Injury	Nonincapacitate	Unknown	Sum
Fatal	0.9999	0.0000	Fatal	34173	0	2	0	0	0	34175
Incapacitate	0.0000	1.0000	Incapacitate	0	0	55	0	8924	0	8979
None	0.9867	0.1488	None	0	0	16858	0	228	0	17086
Possible_Injury	0.0000	1.0000	Possible_Injury	0	0	1662	0	4167	0	5829
Nonincapacitate	0.9180	0.5904	Nonincapacitate	0	0	847	0	9482	0	10329
Unknown	0.0000	1.0000	Unknown	0	0	381	0	346	0	727
			Sum	34173	0	19805	0	23147	0	77125

Fig.4. Error Rate of CS-CRT Algorithm with Relevant Attributes Selected by Feature Ranking Algorithm

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The error rates of classification algorithms using all feature selection algorithms are given in the Fig.5.

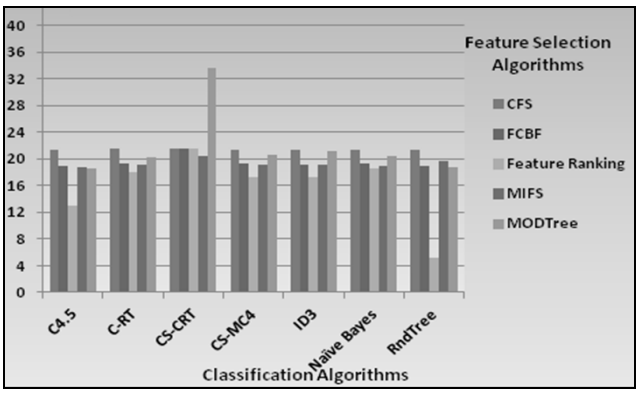

Fig.5. Error Rates Classification Algorithms with Feature Selection Algorithms

Among feature selection algorithms attributes selected by the Feature Ranking Algorithm improves the classifiers' accuracy. From Fig. 5 it is clear that Random Tree classification algorithm using relevant features selected by Feature ranking algorithm gives very less misclassification rate of 5.17%.

The comparison between the misclassification rates of the classifiers with and without using Feature Ranking algorithm is given in the Fig. 6. From Fig. 6 it is clear that Random Tree with Feature Ranking algorithm gives better accuracy with very less misclassification rate. From the results of Experiment-II we conclude that Random Tree with Feature Ranking algorithm gives less misclassification rate of 5.17% while classifying the road accident data based on Injury_Severity.

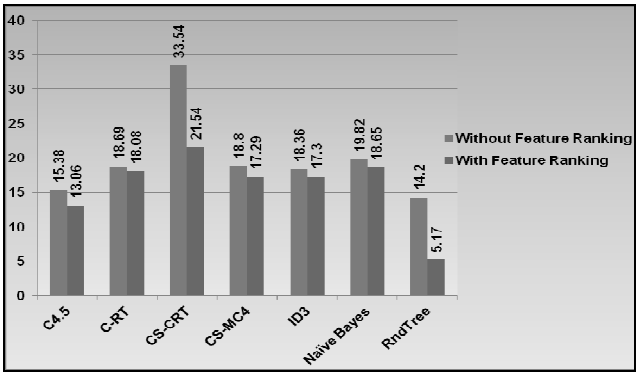

Fig.6. Error Rates of Classifiers with and without Relevant Attributes Selected by Feature Ranking Algorithm

C. Experiment III: Error Rates of Classifiers with Feature Selection Algorithms and Arc-X4 Meta Classifier

Though we have significant reduction in error rates in Experiment –II the best classifier of our study, Random Tree, gives 5.17% misclassification rate. So further to improve the accuracy of the classifiers we have used the Arc-X4 Meta classifier with the classification algorithms.

The Arcing term is used to describe the family of algorithms that Adaptively Resample data and Combine the outputted hypotheses [3]. Adaptively reweighting the training set, growing a classifier using the new weights, and combining the classifiers constructed to date can significantly decrease generalization error.

The error rates of all the classifiers using Arc-X4 Meta classifier are given in the Fig. 7.

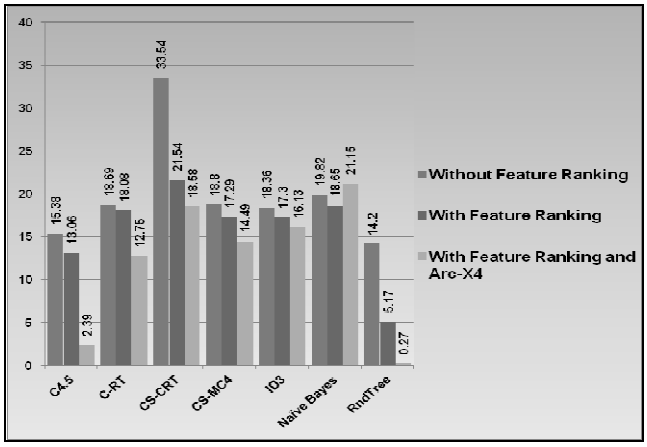

Fig.7. Error Rates of Classifiers without Feature Ranking, with Feature Ranking and ArcX4 Meta Classifier

From Fig. 7 it is clear that the Random Tree classifier using relevant attributes selected by Feature Ranking algorithm and Arc-X4 Meta classifier gives 0.27% error rate i.e. 99.73% accuracy which is higher than that of other classifiers.

The sample result of the Random Tree using Arc-X4 Meta classifier with relevant attributes selected by Feature Ranking algorithm is given in the Fig. 8.

The result reveals that the error rate of Random Tree classifier with relevant attributes and Arc-X4 Meta classifier is excellently reduced from 5.17% to 0.27%. For other classifiers also it shows a notable reduction in the misclassification rates.

Error rate			0.0027						
Values prediction			Confusion matrix						
Value	Recall	1-Precision		Fatal	Incapacitate	None	Possible_Injury	Nonincapacitate	Unknown	Sum
Fatal	1.0000	0.0000	Fatal	34175	0	0	0	0	0	34175
Incapacitate	0.9952	0.0055	Incapacitate	0	8936	3	9	31	0	8979
None	0.9995	0.0054	None	0	0	17078	6	2	0	17086
Possible_Injury	0.9864	0.0040	Possible_Injury	0	9	60	5750	10	0	5829
Nonincapacitate	0.9929	0.0043	Nonincapacitate	0	40	25	8	10256	0	10329
Unknown	0.9917	0.0000	Unknown	0	0	5	0	1	721	727
			Sum	34175	8985	17171	5773	10300	721	77125

Fig 8. Error Rate of Random Tree classifier using Arc-X4 Meta classifier with Relevant Attributes Selected by Feature Ranking Algorithm

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. Conclusion

In this paper we analyzed road accident training dataset using classification algorithms (C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree), feature ranking algorithms (CFS, FCBF, MIFS, MODTree and Feature Ranking), classifiers using Arc-X4 Meta classifier to find road accident patterns using injury severity based classification. Among the algorithms Random Tree classifier using Arc-X4 Meta gives high accuracy of 99.73% with 0.27% misclassification rate. The accuracy is evaluated based on precision and recall values. The results showed that the Random Tree classifier using Arc-X4 Meta classifier based on relevant features selected by Feature Ranking algorithm significantly improved the accuracy from 85.8% to 99.73%. We conclude that Random Tree classifier using Arc-X4 Meta classifier based on relevant features selected by Feature Ranking algorithm outperforms other classifiers to find road accident patterns based on injury severity. Also we observed from rules, generated by the Random Tree using Arc-X4 Meta classifier, is that the factors such as Manner_of_Collision, Seating_Position, Harmful_Event, Protection_System, Age_Range and Drug_Involvement play very important role in deciding Injury_Severity. So these factors need to be focused to reduce the fatality rate.

References

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

[2] R. Battiti, "Using mutual information for selecting features in supervised neural net learning", IEEE Transactions on Neural Networks, vol.5 (4), pp.537-550, 1994.

[3] Breiman, L., "Arcing classifiers", The Annals of Statistics, vol. 26, pp.801-849, 1998.

[4] Breiman, L., "Bagging predictors", Machine Learning, Vol.24, pp.123–140, 1996.

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

[6] Eric Bauer, Ron Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants", Machine Learning, Vol.36, pp. 105-139, 1999.

[7] "FARS analytic reference guide", www.nhtsa.gov.

[8] "Global status report on road safety", World Health Organization, Geneva, 2009.

[9] M. Hall, S. Lloyd, "Feature subset selection: a correlation based filter approach", International Conference on Neural Information Processing and Intelligent Information Systems, pp.855-858. Springer, 1997.

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

[11] 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", Expert Systems with Applications, vol. 32, pp. 987–994, 2007.

[12] 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, vol.36, pp.2473-2480, 2009.

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

[14] M. James, "Classification Algorithms", John Wiley, 1985.

[15] Jaree Thongkam, Guandong Xu and Yanchun Zhang, "AdaBoost algorithm with random forests for predicting breast cancer survivability", International Joint Conference on Neural Networks, 2008.

[16] Nojun Kwak and Chong-Ho Choi, "Input Feature Selection for Classification Problems", IEEE Transactions on Neural Networks, vol. 13, No. 1, 2002.

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

[18] Quinlan, J. R, "Induction of Decision Trees", Machine Learning, vol.1, pp.81-106, 1986.

[19] S.Shanthi, Dr.R.Geetha Ramani, "Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms", International Journal of Computer Applications, Vol.35, No.12, pp.30-37, 2011.

[20] S.Shanthi, Dr.R.Geetha Ramani, "Classification of Seating Position Specific Patterns in Road Traffic Accident Data through Data Mining Techniques", Proceedings of Second International Conference on Computer Applications, ICCA 2012, Vol.5, pp. 98-104, January, 2012.

[21] 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, vol. 34, pp. 1227–1234 2008.

[22] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", In Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03), pp.856-863, 2003

Dr. R. Geetha Ramani is working as Associate Professor in the Department of Information Science and Technology, Anna University, Chennai, 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 8 years of teaching experience. Presently she is working as Associate Professor in the Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore and pursuing her Ph.D (Part Time) in the Department of 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 two papers in international journals and presented many papers at National and International Conferences.

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)