Classifier Prediction Evaluation in Modeling Road Traffic Accident Data

Dr. R. Geetha Ramani[1], S. Shanthi[2]
[1] Department of Information Science and Technology, Anna University, Chennai, India
(rgeetha@yahoocom)
[2]Department of Computer Science and Engineering, Rathinam Technical Campus, Anna University, Chennai, India
(psshanthiselvaraj@gmail.com)

Abstract - **This paper illustrates the research work in exploring the application of data mining techniques to aid in the prediction of road accident patterns related to pedestrian characteristics. It also provides insight into pedestrian accidents by uncovering their patterns and their recurrent underlying characteristics in order to design defensive measures and to allocate resources for identified problems. In this study the Decision Tree algorithms viz. Random Tree, C4.5, J48 and Decision Stump are applied to a database of fatal accidents occurred during the year 2010 in Great Britain. We also used K-folds Cross-Validation methods to measure the unbiased estimate of the four prediction models for performance comparison purposes.**

Keywords – **Road Traffic Accidents, Casualties, Pedestrians, Accident patterns, Decision Tree, Cross Validation**

I. INTRODUCTION

Data mining applications are becoming ever more popular for many applications across a set of very divergent fields [3]. It has become one of the hottest research areas in recent years. Data Mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, rule generation and deduction, regression analysis, genetic algorithms etc. to analyze data and extracting knowledge in different ways [3].

Analysis of crash data is no exception. Applying data mining techniques to analyze traffic accident data records can help us to understand the characteristics of pedestrians, drivers' behavior, roadway condition and weather condition that were related with different accident severity.

Pedestrians are associated with considerable impacts on road safety. The interaction between pedestrian characteristics and the effects of other road safety factors, including roadway, driver, vehicle and intervention variables on road accident frequency is certainly a complex phenomenon that attracts increasing attention by researchers.

In this paper with the help of several data mining techniques we tried to predict road accident patterns related to pedestrian characteristics. This can help decision makers to devise better traffic safety control policies. Decision Tree Classifiers viz. Random Tree, C4.5, J48 and Decision Stump are used to construct our prediction model.

The paper is organized as follows. Section 2 deals with the literature that supports our proposed work. In Section 3 we present a brief introduction to the data set, methodology and decision tree algorithms used in our study. Section 4 summarizes the findings of this research work and Section 5 concludes the paper.

II. LITERATURE SURVEY

The literature shows a great interest in comprehending pedestrian accident patterns in order to plan preventive measures and to allocate resources for identified problems. Several studies included or focused on the effects of pedestrian characteristics on road accidents occurrence and severity, attempting to capture these often complex effects.

In several cases, different groups of drivers such as older drivers [1], motorcyclists [5] and truck drivers [12] are examined. Various studies have addressed the different aspects of Road Traffic Accidents focusing on predicting or establishing the critical factors influencing injury severity [5, 8].

The data mining research focusing on building tree-based models [2] has been conducted to analyze freeway accident frequency.

While evaluating credit risk, Logistic Regression and SVM algorithms give best classification accuracy [4]. It shows the higher robustness and generalization ability compared to the other algorithms. The C4.5 algorithm is sensitive to input data, and the classification accuracy is unstable, but it has the better explanatory [4].

Various feature selection 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 [6, 7, 8, 9, 10]. Also the performance measures such as ROC, Precision and Recall have been used to evaluate the classifiers accuracy [6, 7, 8, 9, 10, and 11].

This paper proposes an analytical and prediction model for predicting road traffic accident patterns related to pedestrian characteristics using data mining decision tree classification algorithms.

III. DATA AND METHODOLOGY

In this paper, we have applied the classification algorithms viz. C4.5, Random Tree, Decision Stump and

978-1-4673-1344-5/12/$31.00 ©2012 IEEE

J48 to predict road traffic accident patterns based on pedestrian characteristics. The performance of these algorithms has been validated using Cross Validation with K folds and the accuracy measures such as precision, recall and ROC. All decision tree approaches investigated are from WEKA and Tanagra. The dataset and the algorithms used in the study are discussed in the following sub sections.

A. The Data

In order to meet the research objectives, a large data set of Road Traffic Accident Casualties information (159417 records, 9 Attributes) was used. The list of attributes and their descriptions is given in the Table 1.

TABLE I
TRAINING DATASET AND ATTRIBUTES DESCRIPTION

S.No.	Attributes Description	
	Attribute Name	Type
1	Accident Index	Identifier
2	Casualty Reference	Identifier
3	Casualty Class(Pedestrian)	Binary
4	Gender	Ordinal
5	Age Band	Ordinal
6	Casualty Severity	Ordinal
7	Car Passenger	Nominal
8	Deprived	Ordinal
9	Casualty Type	Nominal

The dataset is classified based on the dependent variable Casualty Class (Yes-Pedestrian, No-Non Pedestrian) to predict various accident patterns in road accident data. The rest of the attributes are considered as independent variables which influence the dependent variable. As most of the attributes are nominal and ordinal Chi-Square analysis is best [6] to find the statistical significance of the variables related to the study. The attributes Gender, Age Band and Casualty Severity have been selected as best attributes to classify the instances based on Casualty Class. Finally four attributes (3, 4, 5, 6) have been used to proceed further.

B. The Methodology

To predict Casualty Class, various classification models were built using decision tree algorithms viz. Random Tree, C4.5, Decision Stump and J48. Decision trees are easy to build and understand, can manage both continuous and categorical variables, and can perform classification as well as regression. They automatically handle interactions between variables and identify important variables.

Extensive data pre-processing resulted in a clean dataset containing 157463 accidents and four attributes with no missing values. The class label ('Casualty Class') had two values: Yes (Pedestrian), No (Non Pedestrian). During data exploration, different numbers of attributes were selected by different feature selection techniques [6]. In this paper, we collected and cleaned road traffic accident data, attempted to build novel attributes, and tested a number of predictive models. The proposed methodology is depicted in Fig. 1.

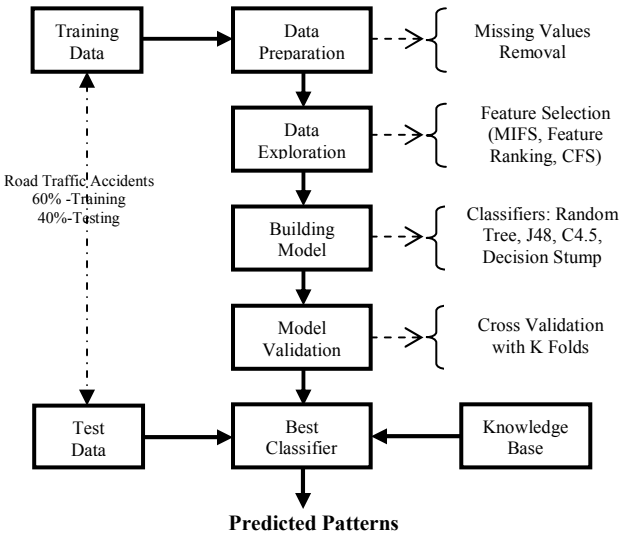

Fig. 1. Predictor Model

The outcome of the classifiers has been validated using Cross Validation with K folds. As the number of instances is large we have applied different values for K in which 7 folds reached its threshold. Also the classifiers accuracy has been evaluated with its misclassification rate, processing time and ROC curve.

C. Decision Tree Algorithms

Decision trees represent a supervised approach to classify the given large dataset. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. By decision tree methods, the classification rules are easily obtained, moreover these rules are clear and easy to understand.

In this paper we have used the decision tree algorithms viz. Random Tree, C4.5, Decision Stump and J48 which have already been discussed in [8].

IV. RESULTS AND DISCUSSION

In this paper we have used Tanagra and WEKA for applying few decision tree algorithms viz. C4.5, Random Tree, Decision Stump and J48 to predict road traffic accident patterns. The results of these classifiers are validated using Cross Validation with K folds and evaluated using the accuracy measures: Precision, Recall and ROC curve. The accuracies and the time taken by the decision tree algorithms to predict the road accident

patterns based on pedestrian characteristics are given in the Table. 2.

TABLE II

PREDICTION RESULTS OF DECISION TREE ALGORITHMS

Classifiers	Validation Using Training and Test Dataset		Validation Using Cross Validation (7 Folds)	
	Accuracy (%)	Time (Milli Seconds)	Accuracy (%)	Time (Milli Seconds)
Random Tree	88.97	203	88.96	647
C4.5	88.97	327	88.92	779
Decision Stump	87.61	300	87.49	750
J 48	88.77	281	88.73	676

From Table 2 it is clear that the accuracy of the all the algorithms are same with slight variation. Thus we have included the time taken by each algorithm to complete the process. By considering the accuracy and the execution time we could able to find Random Tree gives better accuracy (88.97%) with less time (203 milli seconds). The accuracies of the classifiers are validated through the Cross Validation with K folds technique. The value K=7 converged with the predicted accuracy. The sample rules generated by C4.5 algorithm (88.97% accuracy) based on Casualty Class accident patterns is given in the Fig. 2.

- AgeBand in [16-24] then Pedestrian = No (91.61 % of 39697 examples)
- AgeBand in [25-39] then Pedestrian = No (92.70 % of 44001 examples)
- AgeBand in [55-69] then Pedestrian = No (89.78 % of 16196 examples)
- AgeBand in [40-54] then Pedestrian = No (92.85 % of 34523 examples)
- AgeBand in [0-15]
  - CasualtySeverity in [Slight] then Pedestrian = No (63.84 % of 12454 examples)
  - CasualtySeverity in [Serious] then Pedestrian = Yes (65.52 % of 1795 examples)
  - CasualtySeverity in [Fatal] then Pedestrian = No (50.00 % of 34 examples)
- AgeBand in [>70] then Pedestrian = No (82.11 % of 8763 examples)

Fig. 2. Rules generated by C4.5 Classifier

The error rate, precision and recall values of the C4.5 classifier are given in Confusion Matrix as given in the Fig. 3.

Error rate	0.1103
Values prediction	Confusion matrix

Value	Recall	1-Precision
No	0.9956	0.1076
Yes	0.0656	0.3448

	No	Yes	Sum
No	138914	619	139533
Yes	16754	1176	17930
Sum	155668	1795	157463

Fig. 3. Confusion Matrix generated by C4.5 Classifier

The decision tree generated by the Random Tree algorithm (88.97% accuracy) is depicted in Fig. 4.

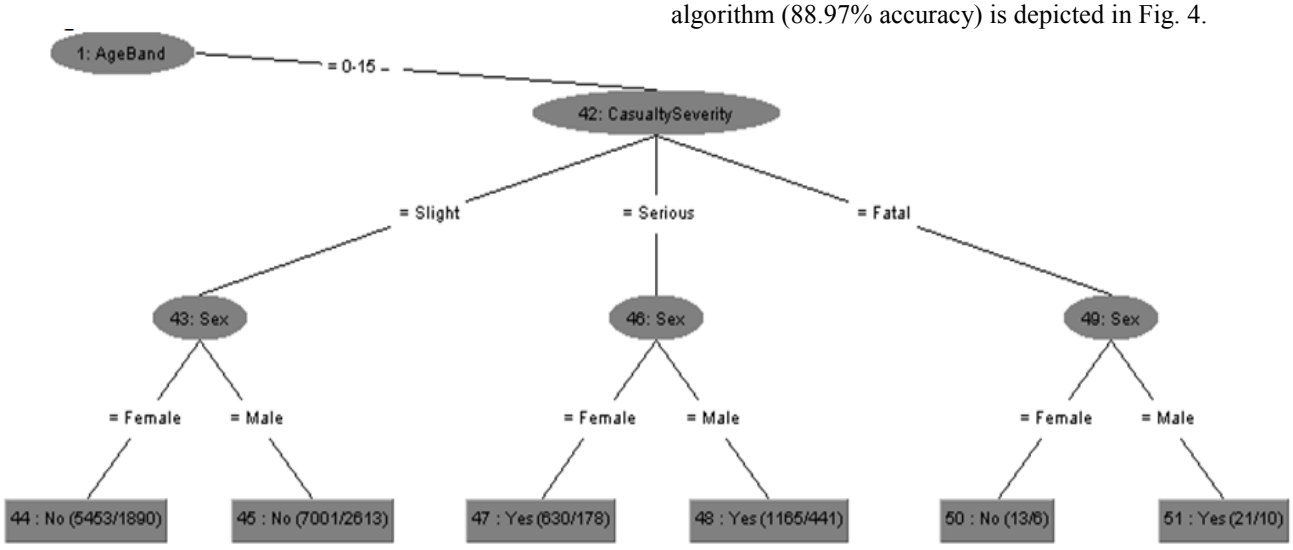

Fig. 4. Decision Tree generated by Random Tree Classifier

From the rules and the decision tree it is clear that the male pedestrian within the age range 0-15 is likely to get fatal injury. The pedestrians within the age range 0-15 irrespective of the gender are more prone to get serious injury. It shows that the children are most likely to involve in serious accidents.

The performance of the classifiers has also been evaluated using ROC curve. ROC curve is a plot of TPR (True Positive Rate) against FPR (False Positive Rate) which depicts relative trade-offs between true positives and false positives [3]. If the curve is closer to the diagonal line then the model is less accurate [3]. Area under receiver operating characteristic curve (AUC) was calculated to assert the prediction accuracy besides the sensitivity, specificity and accuracy. An area of 0.5 represents a random test; values of AUC<0.7 represent

poor predictions; AUC>0.8 represents good prediction [3]. The ROC curve for Decision Stump and Random Tree algorithms are given in Fig. 5. (a) and (b) respectively.

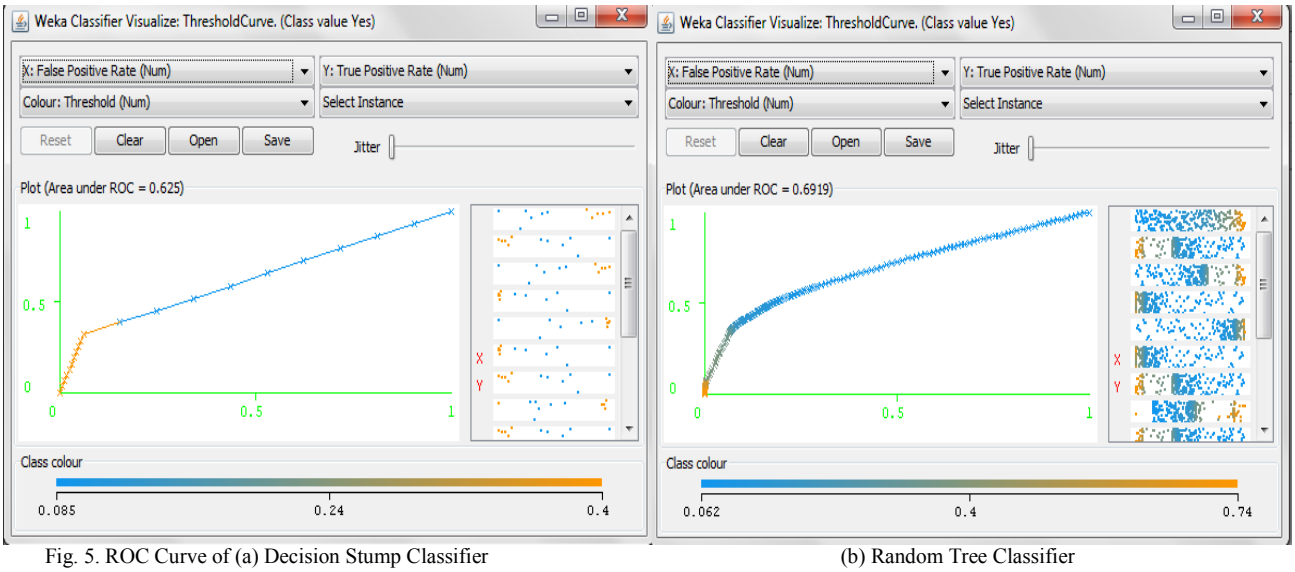

Fig. 5. ROC Curve of (a) Decision Stump Classifier (b) Random Tree Classifier

From Fig. 5 it is clear that the AUC value of Random Tree classifier is more than that of other classifiers.

V. CONCLUSION

The above results indicated that the Random Tree is the best predictor with 88.96% accuracy on the Cross Validation sample, C4.5 and J48 came out to be the second and third with 88.92% and 88.73% respectively accuracy and the Decision Stump model came out to be the least of the four with 87.49% accuracy.

Also from the results we can infer that Young pedestrians, at the age of 0 to 15 cause fatal or serious as well as minor accidents. From the perspective of fatality reduction measures, results suggest the necessity of designing education campaigns for parents, promoting information campaigns for road users.

REFERENCES

[1] Baker, T.K., Falb, T., Voas, R. and Lacey, J., "Older women drivers: Fatal crashes in good conditions", *Journal of Safety Research,* vol. 34, pp. 399-405, 2003.

[2] Chang, L. and W. Chen, "Data mining of tree-based models to analyze freeway accident frequency", *Journal of Safety Research*, vol. 36: pp. 365-375, 2005.

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

[4] Hong Yu, Xiaolei Huang, Xiaorong Hu, Hengwen Cai, "A comparative study on data mining algorithms for individual credit risk evaluation", in Proc. International Conference on Management of e-Commerce and e-Government, 2010.

[5] Pai, C-W. and Saleh, W., "Modeling motorcyclist injury severity by various crash types at T-junctions in the UK", *Safety Science*, vol. 46, pp. 1234-1247, 2008.

[6] S. Shanthi and R. Geetha Ramani, "Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques", in Proc. IAENG-World Congress on Engineering and Computer Science, (WCECS 2012), October 24-26, San Francisco, USA, vol. 1, pp. 122-127, 2012.

[7] S. Shanthi and 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,* in Proc. the Second International Conference on Advances in Computing and Information technology (ACITY 2012), July 13 - 15, Chennai, vol.177, pp. 433-443, 2012.

[8] S. Shanthi and R. Geetha Ramani, "A Comparative evaluation of Classification Methods in the Prediction of Road Traffic Accident Patterns", in Proc. The International Conference on Future Communication and Computer Technology (ICFCCT 2012), Beijing, China, May 19-20, ISBN: 978-988-15121-4-7, 2012.

[9] S. Shanthi and R. Geetha Ramani, "Gender Specific Classification of Road Accident Patterns through Data Mining Techniques", in Proc. IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM -2012), E.G.S. Pillay Engineering College, Nagapattinam, pp.359-369, ISBN: 978-81-909042-2-3, March 30-31, 2012.

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

[11] S. Shanthi and 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, December 2011.

[12] Young, R.K., Liesman, J., "Estimating the relationship between measured wind speed and overturning truck crashes using a binary logit model", *Accident Analysis and Prevention*, vol. 39, pp. 574-580, 2007.