ISSN 2278-3091

Volume 8, No.5, September - October 2019

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse114852019.pdf

https://doi.org/10.30534/ijatcse/2019/114852019

Road Traffic Accident Case Status Prediction Integrating a Modified C4.5 Algorithm

Jackelou S. Mapa[1] Ariel M. Sison[2], Ruji P. Medina[3]

Technological Institute of the Philippines
mapajackelou1@gmail.com

ABSTRACT

C4.5Algorithm is one of the most popular algorithms for rule base classification. The problem C4.5 algorithms is dealing with continuous attribute and the accuracy of prediction is affected and the inducer will choose an over fitted classifier. In this study, there are 25 variables and 799 accidents record generated from this tree. The case status is the best one that can be used for prediction. based on the total training set of instances are increases the accuracy of 69.83% while the incorrect classified instances are 30.16%. the precision solved 725 while the unsolved is 663, recall 746 solved while the unsolved is 638 and F-measure of the solved case status is 735 while the unsolved is 650. The accuracy of Case status of the Road Traffic Accident are increases. The confusion matrix classified as the solve is 224 while the unsolved is 114. Based on the result the common causes of the accident are that Suspect use alcohol with the 615 records which the highest number record with heavy vehicle in the afternoon and evening during rainy and dry season. The occurrence of accident is in Agusan del Norte with 328 accidents.

Key words: C4.5 Algorithm, Road traffic accident, decision tree, Weka

1. INTRODUCTION

The innovation of technology goes fast, the road accidents implies an unacceptable burden on the community in terms of human injury and economic cost. Road accident has Organization stated that road accident has resulted to 4.8 million of injury worldwide. There will be 1.25 million people killed by the year 2020 [WHO, 2015]. The mortality due to road accident is ranked third after the disease of circulatory system and cancer which are placed at first and second ranks respectively [2]. Road accident severities can be classified into several categories which are no injuries, slight injuries, and serious injuries. In Data Mining, as the method of analyzing different type of data to extract interesting patterns and knowledge [4], is used to discover critical information which can help local authorities detect crime [5] as well as predicting areas which have high probability for occurrence of crime and indicate crime prone areas [6] including the type of crimes [7]. In most countries, the traffic accidents are recorded in accident reports by police officers and subsequently the information is stored in a database. Road accident has also become one of the contributing factors towards death for Caraga Region in the road accident case leaving as the top three nominators respectively [3], [11] analyzed the factors that contributed to road crash severity in Ghana’s Northern Region using binary logistic regression.

The study found that overloading and obstruction were the two most significant factors contributing to road crash severity in Ghana [2]. C4.5 classification algorithm, in particular, represent a set of useful methods for analyzing traffic accident severity because, normally, they are non-parametric methods that do not depend on any functional form and require no prior probabilistic knowledge on the phenomena under study [10]. C4.5 Algorithm have been largely reported in road safety literature. Specifically, the most widely used method in the literature on traffic accident severity is the CART method [8][9] [10]; Among the decision trees, C4.5 Algorithm is one of the well-known algorithms used for data classification that has numerical and categorical attributes. The C4.5 classification algorithm is easy to understand as the derived rules have a very straightforward interpretation. The main problem of c4.5 Algorithm is dealing with continuous attributes and the accuracy of prediction is affected and the inducer will choose an over fitted classifier that is not valid in general and also in optimization, the problem of C4.5.

1.1 Traffic Accident Analysis Based on C4.5 Algorithm in Weka (2019)

Numerous studies have focused on the factors affecting traffic accidents. After statistical analysis, the current mainstream method is to use data mining technology for analysis [18]. Data mining technology is an effective method for analyzing traffic accidents, In-depth information mining of traffic accident data is conducive to accident prevention and traffic safety management. The C4.5 algorithms in WEKA is use to explore the impact of various factors on the accident. Sohn S Y et al. [17] used various algorithms to improve the accuracy of the classification of the severity of two types of road traffic accidents. The algorithms included classifier fusion based on Bayes and logic model; data integration fusion based on arc discharge and bagging, and clustering based on k-means algorithm.

1.2 Decision Tree Ensemble Method for Analyzing Traffic Accidents (April 2019)

Decision tree is a predictive model that can be used for both classification and regression currently; an estimated 1.27 million people die and 20–50 million people are injured in traffic accidents study, every year there are critical need to analyze traffic accidents in order to mitigate their terrible economic and human impact [14]. The information root node variation (IRNV) method (based on decision trees) is use to get a rule set that provides useful information about the most probable causes of fatalities in accidents involving inexperienced drivers in urban areas [17].

2. RELATED WORK

Sachin et, al., (2015), proposed a system for Dehradun, India road mishap (11,574) occurred (amid 2009 and 2014) by utilizing K-modes grouping strategy and affiliation lead mining [15]. The investigation of result utilizing blend of these procedure presume that the outcome will be more powerful if no division has been performed preceding produce affiliation rules [16]. Jin-bao Zhao established urban road traffic accident analysis based on Bayesian network structure, combined with network model and application joint tree inference engine in the accident site under the influence of factors such as the vehicle type, traffic accidents and traffic participants' type of statistical distribution [13], [14].

2.1 Operational Framework

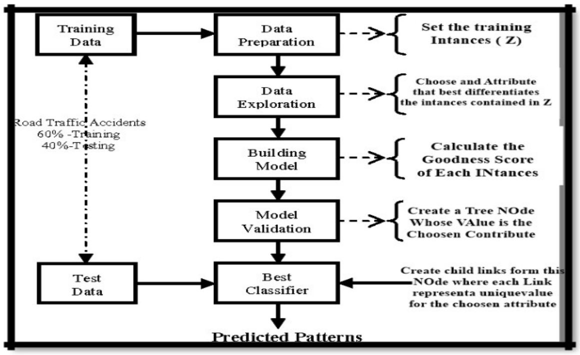

Training Data
Data Preparation
Set the training Intances (Z)
Data Exploration
Choose and Attribute that best differentiates the intances contained in Z
Road Traffic Accidents
60% -Training
40%-Testing
Building Model
Calculate the Goodness Score of Each INtances
Model Validation
Create a Tree NOde Whose VAlue is the Choosen Contribute
Test Data
Best Classifier
Create child links form this NOde where each Link representa uniquevalue for the choosen attribute
Predicted Patterns

Figure 1: Propose Conceptual Framework of the Study

The design is a prediction of the case status of road traffic accident, with prospective data collection, which composed of development and validation phases following suggestions for risk prediction model developments by [12]. The Road Traffic Accident Data is under the Philippine National Police, Caraga Region from January 2018 to February 2019In data preparation, set the (Z) as the training instances. The data Exploration is the chosen attribute that the best differentiates the instances contain to Z, in building model, in this section, calculate the GOODNESS SCORE of each instances. Create a Tree Node which is the highest score in calculating the goodness score and the model validation of the study. The best classifier of the study is to create a Tree child link or subtrees of the ROOT NODE. Where each link represents as an equivalent for the chosen attributes.

2.1 **Data Training of the Study**

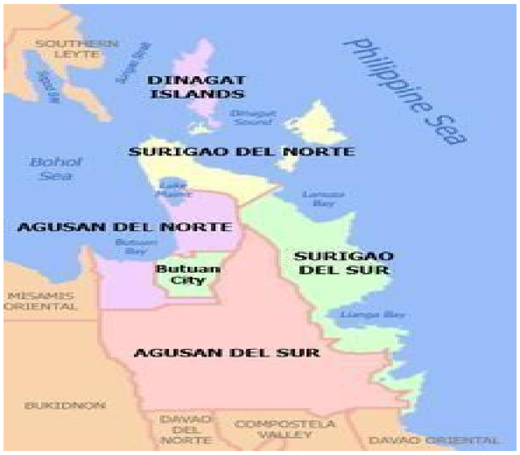


Figure 2: Schematic Road map of the Study

The schematic road of the study is composing of the four provinces, Surigao Del Norte, Surigao Del Sur, Agusan Del Norte and Agusan Del Sur and Dinagat Island. The five Provinces compose of 70 municipalities and total of 1,310 barangays.

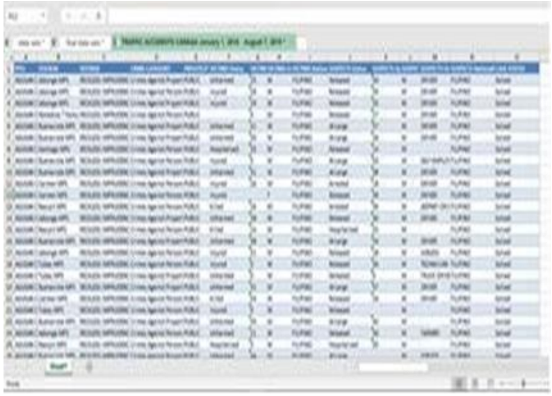


Figure 3: Road Traffic Accident Data Set

In this study, the datasets are from the Philippine National Police Caraga Region. The data are intended to be a nationally representative probability sample from the annual estimated of police accident reports in the Caraga Region. The initial dataset for the study contains traffic accident records from January 2018 to January 2019, a total number of 799 hundred cases. It consists of label- variables: Police Province Office, Victim Nationality, Victim age, Victim Status, Victim Gender, Suspect status, suspect age, Suspect Gender, Suspect Nationality, Suspect drug use, Suspect Alcohol Used, Crime Category, and Incident type. Among these, there were 799 records of head-on collision with fatal injury and all of these records have the impact point categorized as front. Normally, the training sample takes 2/3 and the test sample takes of 1/3.

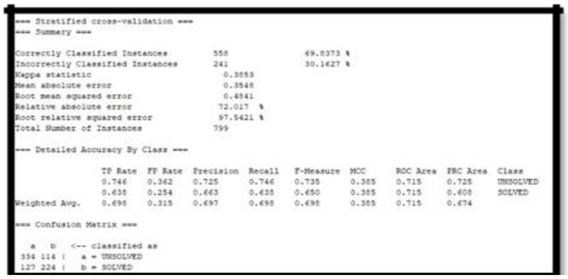
3. THE PROCESS OF INTEGRATING A MODIFIED C4.5 ALGORITH

The road traffic accident in Caraga consists of basic and advanced life support units (BLS, ALS), which are provided by city government and provincial / regional hospitals, respectively, under regulations of provincial and the Department of Health. This paper uses the simple formula to describe the performance of modified C4.5 Algorithm in determining the Status of the road traffic accident of caraga region, if the status is solved or unsolved based on the data collected from the Philippine National Police. In this study, to predict the status of the case using the modified c4.5 algorithms with the mathematical formula where; CS- Case Status

$$\frac{CS= (X+Y) / Z}{T}$$

Figure.4: Formula of Goodness Score

The Formula above, represent the (X+Y) – Sum of the most frequently encountered class in each branches Level in (**Z**). Z – Total number of training set instances T- Total number of branches level



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         558               69.8373 %
Incorrectly Classified Instances       241               30.1627 %
Kappa statistic                          0.3853
Mean absolute error                      0.3548
Root mean squared error                  0.4841
Relative absolute error                 72.017  %
Root relative squared error             97.5421 %
Total Number of Instances              799

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.746    0.362    0.725      0.746    0.735      0.385    0.715     0.725     UNSOLVED
                 0.638    0.254    0.663      0.638    0.650      0.385    0.715     0.608     SOLVED
Weighted Avg.    0.698    0.315    0.697      0.698    0.698      0.385    0.715     0.674

=== Confusion Matrix ===

   a   b   <-- classified as
 334 114 |   a = UNSOLVED
 127 224 |   b = SOLVED
```

Figure 5: Sample Result of Predicting Case Status

In this study, there are Several works have been carried out by different researchers both on road accident analysis and forecasting, using Decision Tree and Artificial Neural Networks. In this study, the result suggested the classified instances based on the total training set of instances are increases the accuracy of 69.83% while the incorrect classified instances are 30.16%. The precision solved 725 while the unsolved is 663, recall 746 solved while the unsolved is 638 and F- measure of the solved case status is 735 while the unsolved is 650. Based on the result of the study using the WEKA application the accuracy in predicting the Case status of the Road Traffic Accident are increases. The confusion matrix classified or cases solve in 558 while the unsolved is 241 cases in 5 municipalities of the Caraga Region.

3.1 Summary and Findings

There are 25 variables and 799 accidents record generated from this tree. The case status is the best one that can be used for prediction**.** based on the total training set of instances are increases the accuracy of 69.83 while the incorrect classified instances are 30.16. the precision solved 725 while the unsolved is 663, recall 746 solved while the unsolved is 638 and F- measure of the solved case status is 735 while the unsolved is 650. the Case status of the Road Traffic Accident are increases. The confusion matrix classified as the solve is 224 while the unsolved is 114.

4.CONCLUSION

The Road Traffic Accident Case Status is collected from Philippine National Police Caraga Region it is found that c4.5 tree can accurately predict the Case Status cause(s) of accident and accident using alcohol along the road and other roads if relevant data are gathered and analyzed as in this case. In c4.5 algorithm Performance analysis, the dataset is experimented integrating the modified c4.5 algorithms there are 69.83% correctly classified instances as solved and 30.16% incorrectly classified instances as unsolved which represent. Mean absolute error was 0.3548 and Root mean squared error was 0.4841. From the detailed accuracy by class and confusion matrix, c4.5 attained accuracy rate of 0.725 and FT attained accuracy rate of 0.608. Base d on the result the common causes of the accident are that suspect use alcohol with the records which the higher number record with heavy vehicle in the afternoon and evening during rainy season. The occurrence of accident in Agusan del Norte with 328 accident.

REFERENCES

[1.] World Health Organization (WHO). (2015) 10 Facts on global road safety. [Online]. Available: http://www.salute.gov.it/imgs/C_17_pubblicazioni_1662 _ulterioriall egati_ulterioreallegato_0_alleg.pdf

[2] M. T. Obaidat and T. M. Ramadan, "Traffic accidents at hazardous locations of urban roads," Jordan Journal of Civil Engineering, vol. 6, pp. 436-447, Oct. 2012

[3]. A.R.A.Rahman (2015) Malayasia ke-20 rekod Kemlangan Terting (Online) Available : http://m.utusan.com.my/berita/nahasbencana/malaysia-ke-20-tertinggi-catat-kemalangan-jalan-raya

[4] K. Rajalakshmi, S. S. Dhenakaran, and N. Roobini, "Comparative analysis of K-means algorithm in disease prediction," Int. J. Sci. Eng. Technol. Res., vol. 4, no. 7, pp. 2697–2699, 2015.

[5] J. Agarwal, "Crime analysis using K-means clustering," Int. J. Comput. Appl. , vol. 83, no. 4, pp. 975–8887, 2013.
https://doi.org/10.5120/14433-2579

[6] O. Vaidya, S. Mitra, R. Kumbhar, S. Chavan, and R. Patil, Comprehensive Comparative Analysis of Methods for Crime, pp. 715–718, 2018

[7]. Delima, Allemar Jhone 2019/07/09 533- 538 Applying Data Mining Techniques in Predicting Index and non-Index Crimes Volume – 9 DOI - 10.18178/ijmlc.2019.9.4.837 - International Journal of Machine Learning and Computing
https://doi.org/10.18178/ijmlc.2019.9.4.837

[8.] De Oña, J., López, G., & Abellán, J. (2013a). Extracting Decision Rules from police accident reports through Decision Trees. Accident Analysis and Prevention, 50, 1151–1160.
https://doi.org/10.1016/j.aap.2012.09.006

[9.] De Oña, J., López, G., Mujalli, R. O., & Calvo, F. J. (2013b). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. 3.Accident Analysis and Prevention, 51, 1–10.
https://doi.org/10.1016/j.aap.2012.10.016

[10.] De Oña, J., Mujalli, R. O., & Calvo, F. J. (2011). Analysis of traffic accident injury on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention, 43, 402–411.
https://doi.org/10.1016/j.aap.2010.09.010

[11.] Haadi, A., 2014. Identification of factors that cause severity of road accidents in Ghana: acase study of the Northern Region. Int. J. Appl. Sci. Technol. 4, 242–249.

[12.] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskil lP, Steyerberg EW, etal. Transparent Reporting of a multi variable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162: W1–73
https://doi.org/10.7326/M14-0698

[13]. JiajiaLi, JieHe, ZiyangLiu, Hao Zhang , and ChenZhang School of transportation, Southeast, Nanjing 21189, Jiangsun China. MATEC Web of Conference 272, 01035 (2019) ICF MCE 2018
https://doi.org/10.1051/matecconf/201927201035

[14] Serafín Moral-García , Javier G. Castellano, , Carlos J. Mantas , Alfonso Montella and Joaquín Abellán in Decision Tree Ensemble Method for Analyzing Traffic Accidents:
www.mdpi.com/journal/entropy
Entropy 2019, 21, 360; doi:10.3390/e21040360

[15]. Sachin Kumar and Durga Toshniwal, "A data mining framework to analyse road accident data", Journal of Big Data (2015) 2:26 DOI 10.1186/s40537-015-0035-y.

[16]. J.B. Zhao, W. Deng, J. Wang, Bayesian network-based urban road traffic accidents analysis, Southeast University,41(6) (2011)1300-1306

[17]. Sohn S Y, Lee S H. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea [J]. Safety Science, 2003, 41(1):1-14

[18]. Dipo T. Akomolafe, Ak inbola Olutayo Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways, American Journal of Database Theory and App lication 2012, 1(3): 26-38DOI: 10.5923/j.dat abase.20120103.01

[19]. Abellán, J., López, G., & de Oña, J. (2013). Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications, 40(15), 6047 6054.doi: 10.1016/j.eswa.2013.05.027

[20]. Li, J., He, J., Liu, Z., Zhang, H., & Zhang, C. (2019). *Traffic accident analysis based on C4.5 algorithm in WEKA. MATEC Web of Conferences, 272, 01035.* doi:10.1051/matecconf/201927201035

[21]. Selvaraj S. Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques[C]// Iaeng-World Congress on Engineering and Computer Science. 2012.

[22]. Sohn S Y, Lee S H. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea[J]. Safety Science, 2003, 41(1):1-14.
https://doi.org/10.1016/S0925-7535(01)00032-7