International Journal of Computer Applications Technology and Research
Volume 7- Issue 01, 001-011 , 2018, ISSN:-2319-8656

Analysis of Accident Times for Highway Locations Using K-Means Clustering and Decision Rules Extracted from Decision Trees.

Ali Moslah Aljofey
Department of Computer Science
University of Thamar
Thamar, Yemen
ali.moslh@gmail.com

Khalil Alwagih
Department of Information Technology
University of Thamar
Thamar, Yemen
khalilwagih@gmail.com

Abstract: Analyzing of traffic accident data play an important role in identifying the factors that affecting the repeated accidents and trying to reduce them. Accidents frequencies and their causes are different from one location to another and also differ from time to time in the same location. Data mining techniques such as clustering and classification are widely used in the analysis of road accident data. Therefore, this study proposes a framework to analyze times of accident frequencies for highway locations. The proposal framework consists of clustering technique and classification trees. The k-means algorithm is applied to a set of frequencies of highway locations accidents within 24 hours to find out when and where accidents occur frequently. These frequencies were extracted from 358,448 accident records in Britain between 2013 and 2015. As a result of clustering technique, four clusters were ranked in descending order according to the accidents rate for location within the cluster. After that, the decision tree (DT) algorithm is applied to the resulting clusters to extract the decision rules as the cluster name represents the class value for all tuples contained. However, extracting decision rules (DRs) from the DT is restricted by the DT's structure, which does not allow us to extract more knowledge from a specific dataset. To overcome this problem, in our study, we develop an ensemble method to generate several DTs in order to extract more valid rules. The DRs obtained were used for identifying the causes of road accidents within each cluster.

Keywords: Accident Severity Analysis, Road Accident Frequencies, Clustering, Decision Tree, Decision Rules.

1. INTRODUCTION

Road accident data are classified as big data. They include many attributes that belong to the accident such as driver attributes, environmental causes, traffic characteristics, vehicle characteristics, geometric characteristics, location nature and time of day. Road accidents data are taken for a long period of time and available in the form of datasets, statistical tables and reports or even GPS data. Most studies used statistical techniques [40, 31] and data mining techniques [19] to analyze the road accident data.

Many researchers studied the causes of accident severity in several ways, depending on different accident factors. For example, De oña et al. [14] used Latent class clustering and Bayesian networks in analysis of traffic accidents to identify the main factors of accident severity. The combined use of both techniques is very interesting as it reveals further information. In other work, K-modes clustering technique and association rule mining have been used as framework to analyze road accident data [27]. Another important factors of accident severity such as vehicle type, driving behavior, collision type, and pedestrian crash have been analyzed [45, 5, 46, 10, 35, 28, 41, 8, 21]. It is possible to discover the influencing factors of bicycle accident severity, and prevent the occurrence of its accidents by comparing participants riding on-road with riding in the simulator [36, 15, 34]. Other studies have investigated the correlation between values of severity level (i.e., no-injury, injury and fatality) and values of other road attributes by using the full Bayesian and multivariate random parameters models [4, 6]. Similarly, Huang et al. [18] employed multivariate spatial model to find the correlation between different modes of accidents (i.e., vehicle, bicycle and pedestrian) at individual intersections and adjacent intersections. While Xie et al. [43] presented crash model to find the correlation of crash occurrence between neighboring intersections.

Rather than defining factors that are closely related to accidents severity, some studies focused on defining factors that are related to accidents occurrence and their locations. Time series data from several locations has been used to synthesize the districts with similar accident pattern by using hierarchical clustering technique [22, 23]. The same authors provided data mining framework to clustering similar locations in accident frequencies together, and discovering the characteristics of these locations [24]. Clustering accident frequencies locations is useful to know the most frequent accident locations. However, this is not enough to know the most frequent accident times for locations. Because if we take accident frequencies for each location within 24 hours, where each hour represents number of accidents for several years, the number of maximum frequencies for each location will be 24 frequencies. Therefore, we need to clustering accident frequencies times for locations rather than clustering accident frequencies locations.

Decision tree DT algorithms is non-linear and non-parametric data mining techniques for supervised classification and regression problems [7] which is a useful way to classify accidents and find factors that influence in frequency of accidents. Classification and Regression Tree (CART), is one of the most famous algorithms that has been used widely to analyze and identify factors that affect the severity of accidents [9, 29, 39, 44, 13]. In order to gain better understanding of crash characteristics, classification trees analysis and rules discovery were performed on two-wheeler (PTW) crashes data [33].

Furthermore, risk factors of accident severity that identified with the decision tree classifier and the Naive Bayes classifier were compared [25].

One of the problems of DT is how to improve its accuracy and produce the largest number of valid rules when the used data are huge. Ensemble methods(i.e., Bagging, boosting, and random forests) generate a set of classification models in order to Increasing classifier accuracy[19]. Abellán and Masegosa [1] have introduced an ensemble decision trees method in which each decision tree differs by the root variable trees. The procedure to build these DTs was based on the Abellán and Moral method [3] which estimates the probabilities by mean of a specific kind of convex sets of probability distributions (also called credal sets) and uncertainty measures. Based on it, Abellán et al. [2] have proposed information root node variation method for extracting rules from DTs, which applied on traffic accident data from rural roads. In this method the root of tree only changes according to the number of road accident attributes and the rest of tree is built by Abellán and Moral producer [3] that can easily adapted to be used with precise probabilities; for example, via the Gini index [7] or gain ratio split criteria [38]. However, this producer to build the rest of DT deal only with small data. There are other procedures to build DTs that can be used with a massive data for example, rxDTree algorithm is an approximate decision tree algorithm with horizontal data parallelism [11] inspired by the algorithm proposed by Haim and Tov [20].

In this paper, the K-means algorithm is used to divide the times of road accident frequencies associated with different locations into several clusters. Furthermore, a particular method for extracting DRs from DTs is applied on these clusters to understand the characteristics of road accidents. The main characteristics of this method is that different DTs are built by varying the root node. The procedure to build DTs, which we will use here, it is based on the procedure proposed by Haim and Tov [20].

The paper is structured as follows: Section 2 shows the methodology of the K-means clustering and decision tree classifier. It also describes the method used to obtain DRs, and the accident data used in this study. In Sections 3, the results of the analysis are displayed and we discuss them. Finally, the last section presents the conclusions.

2. METHODOLOGY

2.1 Clustering Algorithm

Clustering is one of the most data mining techniques used in unsupervised learning, the result of clustering is a group of clusters contain data objects that are similar within the same cluster and are dissimilar to the objects in other clusters. There are many of clustering algorithms [42 ,19] such as k-means, and k-modes. K-means algorithm is a centroid based technique, it deals with the numeric data while k-modes algorithm deals with the nominal data.

K-means algorithm needs a parameter K to determine the number of clusters. At first, the clusters are initiated with random values of data objects as cluster centers. These cluster centers are the centers around which the data objects centered, data objects are assigned to the clusters by calculating the distance between each object and all other centers based on Euclidean distance and is given by Eq. (1), then the nearest distance is chosen. Cluster center is updated by the mean value of objects in the cluster. The process of updating the centers and reassigning the cluster objects are an iterative process until the assignment is stable.

$$d(i,j)= \sqrt{(xi1-xj1)^2+(xi2-xj2)^2+\cdots+(xip-xjp)^2} \qquad (1)$$

where i and j two objects described by *p* numeric attributes.

2.2 Determining The Number of Clusters

One problem of clustering techniques is how to determine the best number of expected clusters. K-means algorithm requires the user to enter the number of clusters k. In our framework we have used k-means algorithm to divide the accident frequency times associated with locations into different groups depending on the Elbow method [30] to determine the number of clusters K. The Elbow method is one of the optimal methods that depends on both the measure of similarities within a cluster and the parameters that used for partitioning. The idea of partitioning is to create clusters where the variation within a cluster is minimized. The quality of cluster can be measured by summing the squared distances between each object within the cluster and its center by using Eq. (2).

$$E=\sum_{i=1}^{k}\sum_{p\in Ci} dist(p,Ci)^2 \qquad (2)$$

where E is the sum of squared error of all data objects; p is the point of an object; and Ci is the cluster center.

The optimal number of clusters can be defined as following [26]:

1. compute the clustering algorithm (i.e., k-means) for different values of K, k = 2 to k = 15.
2. for each cluster k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

2.3 Classification and Decision Trees

Classification is supervised data mining technique whose main task is to predict class (categorical) labels, where a class label for each tuple of dataset is predefined [19]. A tuple X consists of several attributes represented by an n-dimensional attribute vector, X=(x1,x2,…,xn) , each tuple X is related to class label. We can rely on the clustering technique to determine the values of class variable. Data classification process consists of two stages : learning and classification. In the learning phase, classification model is constructed based on training data. In the classification phase, Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

The decision tree DT consists number of levels and branches that starts with a root node and end with leaves, which each internal node indicates a test on the attribute, each branch represents the result of the test, and each leaf holds a class label. Within a DT, each path that starts from the root node and ends with a leaf node called decision rule DR, and this rule is assigned to most probable value of the class variable.

The DT does not require any setting parameters or a prior underlying relationship between target (dependent) variable and predictors (independent variables), however during training it requires measures(i.e., information gain and the Gini index) to select the best attribute for dividing tuples into distinct classes.

According to the amount of data used in the training and testing, there are many DT algorithms that conform to small data such as ID3, C4.5, and CART. ID3 algorithm uses information gain as its attribute selection measure to determine how the tuples at a given node are to be split [37], C4.5 algorithm used gain ratio [38] and CART algorithm used Gini index measure [7]. In contrast, there are several more scalable algorithms capable of handling large data for example, RainForest [17] and BOAT [16] algorithms. One of the tools provided by Microsoft Corporation is rxDTree algorithm inspired by the algorithm proposed by [20]. The rxDTree tool is a parallel external memory of DT algorithms directed for very large datasets. It uses a histogram to approximate data instead of storing it entirely on processors, and the approximated data is used to improve the classifier over time. The rxDTree algorithm constructs DT in breadth-first mode using horizontal parallelism based on the node's impurity. Impurity of node is a function that measures the homogeneity of labels in samples reaching the node. The most popular impurity functions are the Gini index criterion, and the entropy function.

$$Gini(D) = 1 - \sum_{i=1}^{m} p^2 i, \tag{3}$$

$$Info(D) = -\sum_{i=1}^{m} pi \log_2(pi)\,, \tag{4}$$

where D is a set of training tuples, pi is the probability that a tuple in D belongs to class Ci and is estimated by |Ci,D|/|D|, the sum is computed over m classes. Gap function G is continuous and satisfy G({pi}) ≥ 1−maxi{pi}, and it holds the proprieties of Gini and entropy functions. Suppose that an attribute j and a threshold a are chosen, so that a node v is split according to the rule x(j) < a. denote by τ the probability that a sample reaching v is directed to v's left child node. Denote further by pL, i and pR, i, the probabilities of label i in the left and right child nodes, respectively. The notation Δ, represents the gap in the impurity function before and after splitting, for every candidate split, Δ can be calculated precisely, as in equation (5). The function Δ(τ, {pi}, {pL, i}, {pR, i}) = Δ(v, j, a) as

$$\Delta = G(\{pi\}) - \tau G(\{pL, i\}) - (1 - \tau)\, G(\{pR, i\}). \tag{5}$$

2.3.1 Metrics for Evaluating Decision Rules

The decision rules DRs have to be valid when the required conditions have been achieved during the training and testing sequentially. Training data are used to construct the classifier while the testing data are used to test the classifier. Test data are a set of records which associated with a class label are not used to train the classifier, they used to estimate the accuracy of the classification model. Accuracy and coverage are a set of evaluation measures can be used to verify the effectiveness of DRs [19].

2.3.1.1 Accuracy

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

$$Accuracy = \frac{TP+TN}{P+N} \tag{6}$$

$$Error\ Rate = \frac{FP+FN}{P+N} \tag{7}$$

where TP refers to the positive tuples that are classified as positive, whereas TN refers to the negative tuples that are classified as negative , FN refers to the positive tuples that are classified as negative, FP refers to the negative tuples that are classified as positive, P=TP+FN refers to the total number of positive tuples, and N=FP+TN refers to the total number of negative tuples.

The accuracy of a rule R is,

$$Accuracy\ (R) = \frac{N\ correct}{N\ covers} = \frac{TP}{TP+FP} \tag{8}$$

and the error rate of the rule is,

$$Error\ (R) = \frac{N\ negative}{N\ covres} = \frac{FP}{TP+FP} \tag{9}$$

2.3.1.2 Coverage

A rule's coverage is the percentage of tuples that are covered by the rule R.

$$Coverage\ (R) = \frac{N\ covers}{|D|} \tag{8}$$

where N covers are the number of tuples covered by R, and |D| is the number of tuples in test set.

2.3.1.3 Rule Quality

The accuracy measure of a rule on its own is not a reliable estimate of rule quality and the coverage measure of a rule on its own is not useful [19]. Thus, we can integrate aspects of the accuracy and coverage measures for evaluating rule quality by the multiplication of accuracy and coverage as follow:

$$Quality = Accuracy\ x\ Coverage \tag{11}$$

2.3.2 Model Evaluation and Class-imbalanced Data

When the data is huge, it is better to use a part of data to derive the classification model and the other part to predict the accuracy of the model. The Holdout method is one of evaluation methods , which divides the data into independent parts, typically two parts of data are taken for training set, and the remaining part is taken for test set [19].

Class-imbalanced data is one of the biggest problems associated with the classification of DT. The class-imbalance problem occurs when the main class of interest is represented by only a few tuples. Some strategies for addressing this problem include oversampling, under-sampling, and hybrid-sampling [19]. Oversampling works by resampling the positive tuples so that the resulting of training set contains an equal number of positive and negative tuples. The oversampling has advantage to keep the information, however requires more processing time and space. Under-sampling works by decreasing the number of negative tuples, it randomly eliminates tuples from the majority (negative) class until there is an equal number of positive and negative tuples. Hybrid-sampling combines both of oversampling and under-sampling methods.

2.4 Method to Extract Decision Rules from Decision Trees

The ensemble methods combine multiple votes to classify new tuple by using set of classification models M1, M2, …, Mk. The benefit of the ensemble method is to increase classifier accuracy. The ensemble method that we used is ensemble DTs method [1] to generate several DTi (i=1,…,n) by changing the root node RXi of the tree (see Fig.1) according to each variable under study (see Table 1). When DT is built, the root node only is selected directly, and the rest of tree is constructed in the Streaming Parallel Decision Tree (SPDT) algorithm proposed by [20]. Thus, we obtain m trees and m rules, DTi and DRi (i=1,...,m), respectively. Each DRi is checked in the test set to obtain the final rule set.

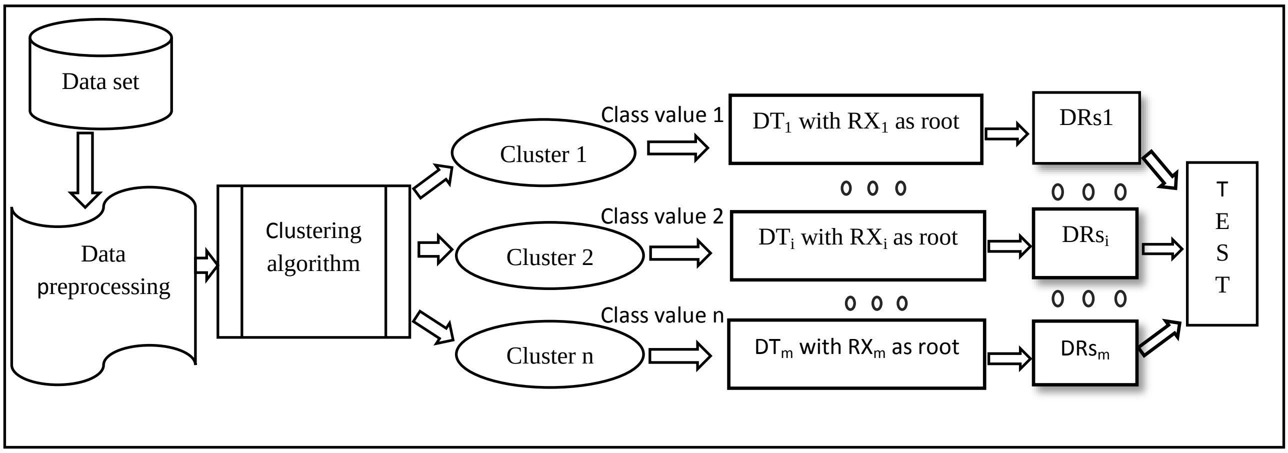

Figure 1. Proposed Framework

Table 1. Attribute Description

Attribute Name	Description: Code	Cluster				
		Count	First	Second	Third	Fourth
Accident Severity: **sev**	Fatal : **f** Serious: **s** Slight: **sl**	3338 50117 301777	19.5% 25.7% 24.9%	27.4% 27.6% 24.6%	28.2% 25.2% 24.9%	24.7 % 21.9% 25.5%
Number of Casualties : **noc**	1 injury: **1** 2 injury: **2** >2 injury: **>2**	275798 55458 23976	24.6% 26.1% 25.9%	24.5% 26.3% 27.3%	25.1% 24.4% 24.8%	25.6% 22.9% 21.9%
1st_Road_Class : **rc**	Motorway: **m** A slope : **a** Express way : **b** Curve : **c** Unclassified: **u**	16481 156312 46895 29927 105617	37.4% 23.3% 26.2% 20.9% 26.3%	24.8% 23.3% 28.6% 23.3% 26.3%	16.5% 26.1% 25.3% 28.9% 23.3%	21.1% 27.1% 19.6% 27.6% 23.9%
Road Type : **rt**	Single carriageway : **sn** Dual carriageway **: du** Roundabout **: ro** One way street **: ow** Unknown **: un**	258232 60878 27016 7437 1673	24% 28.9% 24.7% 24% 35.6%	25.6% 22.85% 24.7% 21.7% 19.3%	26.4% 20.4% 21.5% 25.2% 23.9%	23.8% 27.7% 28.9% 28.9% 21%
Speed limit : **sl**	Less than 40km/h :**less40k** Greater than or equal 40km/h:**more40k**	223810 131422	23.3% 27.8%	22.7% 28.8%	24.9% 25%	28.9% 18.2%
Light Conditions : **lc**	Day light : **dl** Darkness - lights lit :**sl** no light : **nl**	289514 46298 19420	25.6% 21.4% 23.7%	25.2% 21.1% 29.7%	25.1% 24.3% 24.7%	23.9% 33% 21.7%
Weather Conditions : **wc**	Fine : **fi** Raining : **ra** Snowing : **sno** Fog or mist: **fog** Other : **oth**	294156 43858 2437 1557 13224	24.6% 25.1% 22.4% 19.4% 33.2%	25.1% 24.6% 25.1% 24.2% 22.9%	25.1% 24.7% 26.6% 26.6% 20.8%	25% 25.3% 25.8% 29.6% 22.8%
Road Surface Conditions : **rsc**	Dry : **dr** Wet : **w** Icy or slippery : **ic** Other : **ot**	253281 95160 4522 2269	25.2% 24.7% 20.2% 18.8%	24.8% 25.4% 23.7% 24.3%	24.9% 24.9% 26.4% 31.2%	24.9% 24.8% 29% 25.6%
Urban or Rural Area : **ura**	Urban : **ur** Rural : **ru**	217763 137469	23% 28.1%	22% 29.7%	24.1% 26.3%	30.8% 15.8%
Carriageway Hazards : **ch**	None : **no** Other object on road: **ob** Any animal in carriageway: **an** Pedestrian in carriageway : **p**	349711 3457 1242 822	24.9% 29.8% 19.2% 20 .9%	24.9% 28.2% 28.9% 16.4%	25% 21.8% 24.9% 22.7%	25% 19.9% 26.8% 39.9%

2.5 Data Set

The data for this study have been obtained from data.gov.uk Department for Transport [47]. The dataset consists of 425,041 road accidents for 3 years period from 2013 to 2015, in Great Britain of 208 highway locations. After preprocessing, 358,448 accidents records have been considered for this research. The attributes taken from the original dataset were only those related to accident circumstances which represented on Table 1. As a clustering result, accident frequencies between 59 and 1168 for 208 locations with their times were clarified as four clusters. The DT algorithm was applied to the resulting clusters as the class variable values were determined by cluster names (Cluster in Table 1). The Holdout method was applied to divide the dataset into training set and test set. After oversampling and under-sampling, the distribution of tuples was 248,668 for training and 106,568 for testing.

3. RESULT AND DISCUSSION

3.1 Cluster Analysis

The k-means algorithm was applied to 413 unique frequencies of 359,204 frequencies using R statistical software. These frequencies were extracted from 358,448 accident records using PostgreSQL software for 208 highway locations, and each location may have a maximum 24 frequencies associated with 24 hours. The k-means algorithm needs the value of k to determine the number of clusters. Therefore, we used Elbow method which mentioned in Sect. 2.2. We defined the number of clusters which shown in Fig. 2 by using the Elbow method. The distribution of these clusters for frequencies is illustrated in Fig. 3. As a result of clustering, four clusters were ranked in descending order according to the accidents rate for location within cluster, the clusters have been renamed as first, second, third, and fourth.

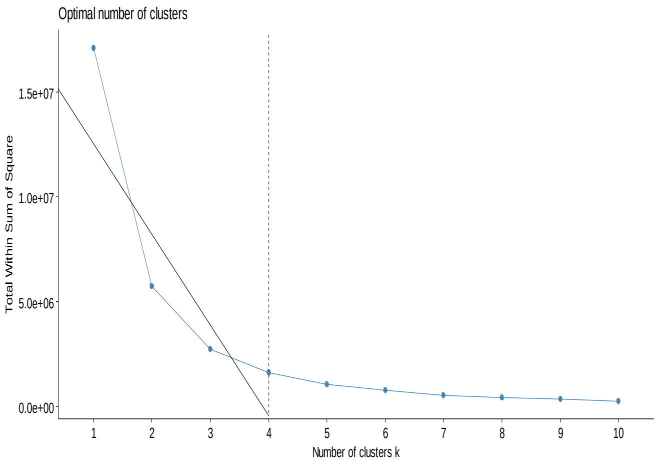

Figure 2. Number of selected clusters

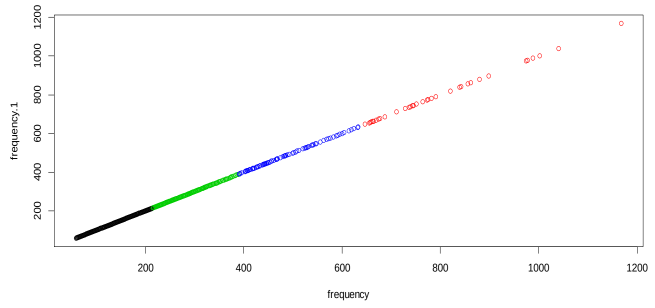

Figure 3. Distribution of clusters

In first cluster, there are 7 locations with frequency range between 647 and 1168, while the number of distinct frequencies is 35 and the total number of accidents is 28,498, therefore the accidents rate for location is 14% . Similarly, the remaining clusters are described in the same manner in Table2.

Table 2 Description of clusters

Cluster id	1	2	3	4
Category name	First	Second	Third	Fourth
Rang of frequency	647-1168	392-633	215–391	59-214
Size of cluster	35	81	141	156
Total of accidents	28498	46569	89066	195071
Number of locations	7	22	61	178
Location accidents ratio	14%	4%	1%	0.5%

The times of accident frequencies for locations within each cluster have been represented in Figs. 4,5,6,and 7, so that each location can have more than one frequency according to the time, and it can be repeated in more than a cluster at a different time, in contrast, the time can be repeated in more than a cluster with different locations. Figure 4 shows that highest frequency value is associated with location 29 at 5 pm, location 29 has 11, 3, 4, and 6 frequencies within clusters first, second, third, and fourth respectively while location 4 has only 4 and 14 frequencies within the third and fourth clusters. The first cluster includes accident times from 8 am to 6 pm, the second cluster includes the times from 7 am to 8 pm, the third cluster refers to the times from 6 am to 11 pm, and the fourth cluster refers to the times from 1 am to 12 pm.

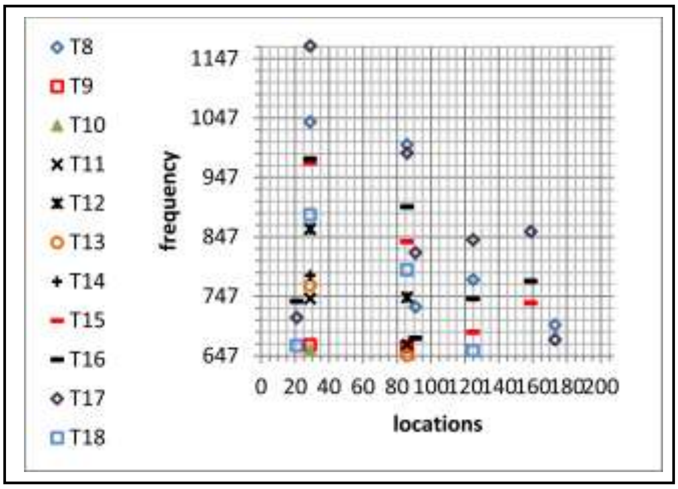

Figure 4. Times (T8 - T18) and locations (21,29,86,91,125,159,173) of First cluster

3.2 Extracting Decision Rules

In our framework the locations associated with times of accident frequencies have been divided into four clusters namely first, second, third and fourth, all tuples within a cluster have the cluster name as class label. Then the method exposed in section 2.4 has been used in order to generate DRs

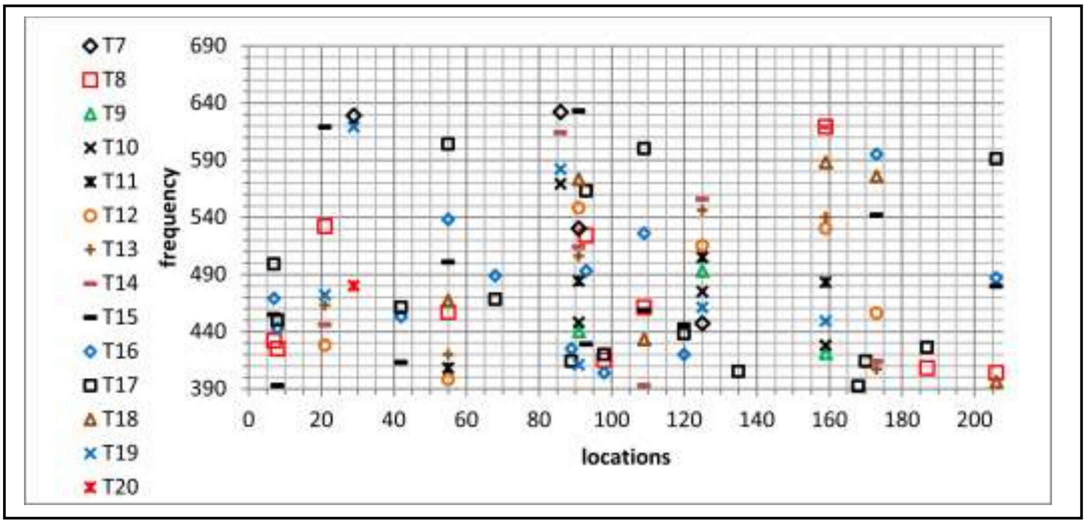

Figure 5. Times (T7-T20) and locations (7,8,21,29,42,55,68,86,89,91,93,98,109,120,125,135,159,168,170,173,187,120) of Second cluster

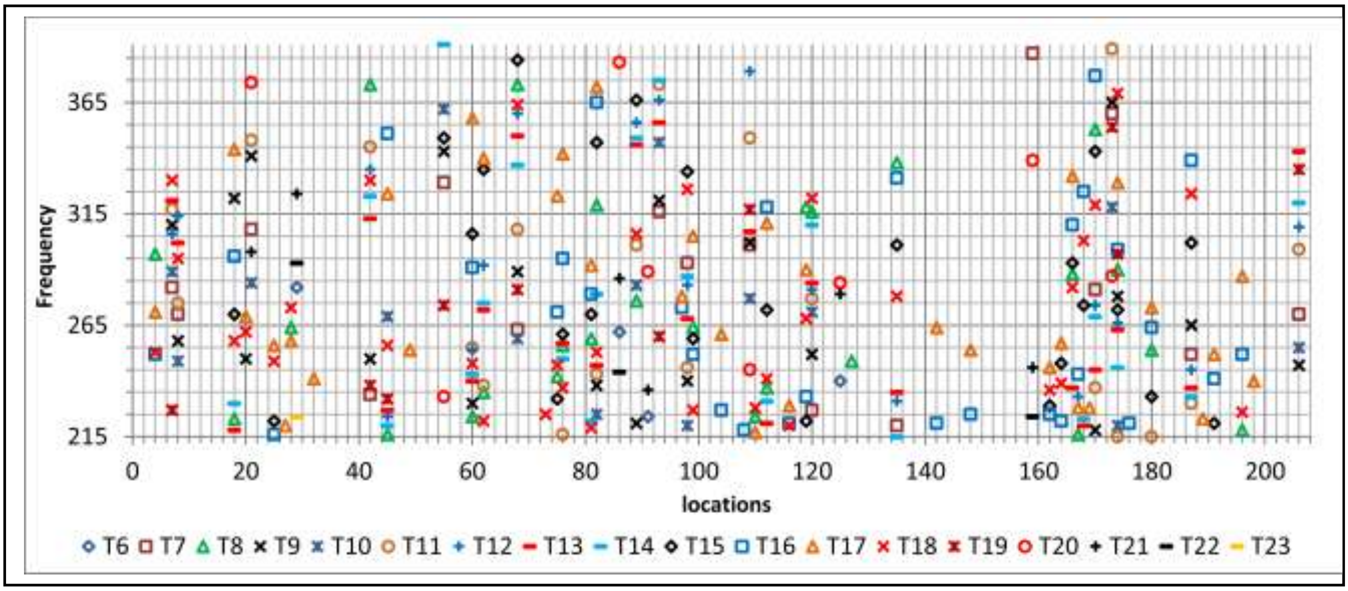

Figure 6. Times(T6-T23) and locations(4,7,8,18,20,21,25,27,28,29,32,42,45,49,55,60,62,68,73,75,76,81,82,86,89,91,93,97,98,99, 104,108,109,110,112,116,119,120,125,127,135,142,148,159,162,164,166,167,168,169,170,173,174,176,180,187,189,191,196, 198, 206) of Third cluster.

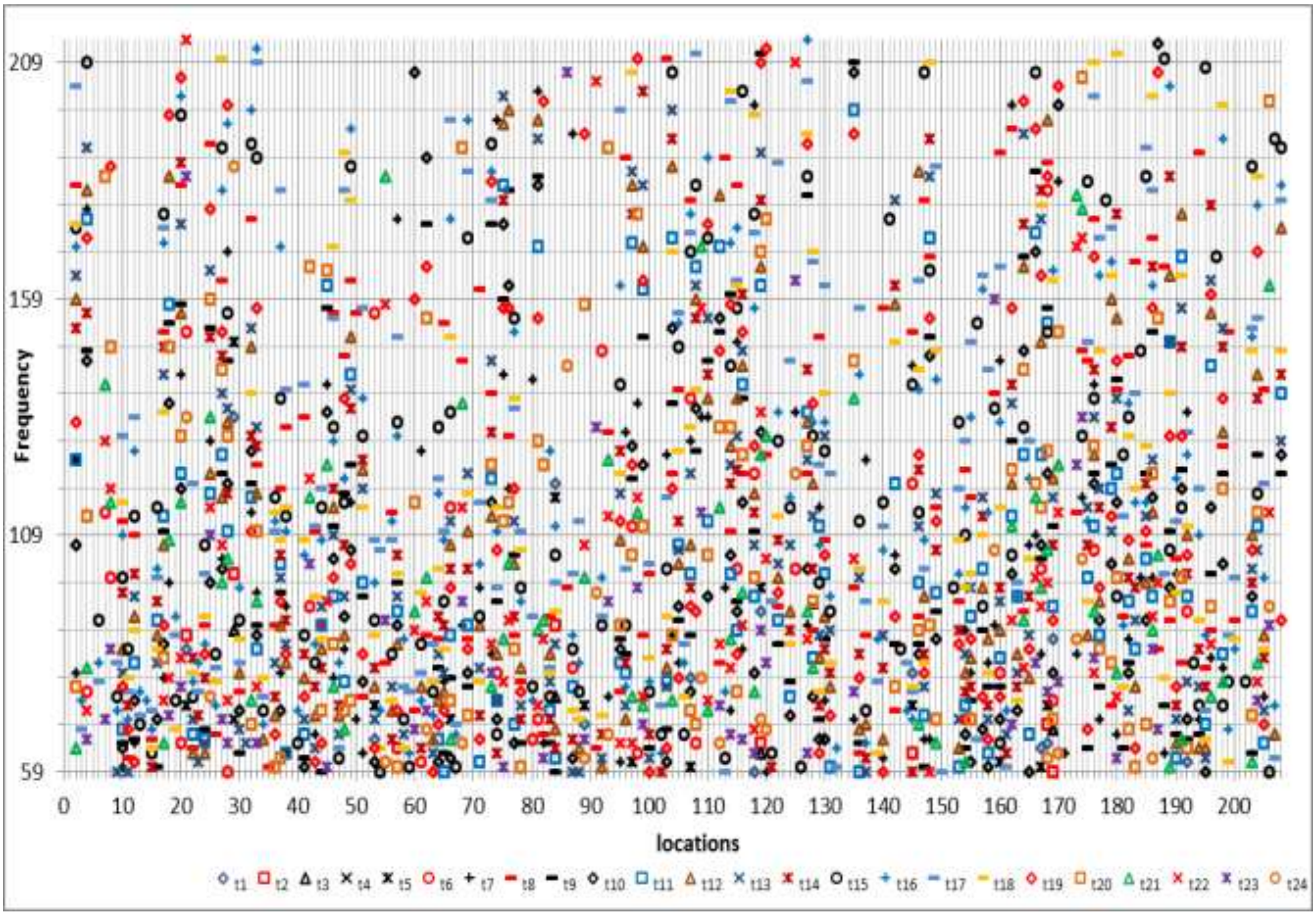

Figure 7. Times(T1-T24) and locations of Fourth cluster.

using rxDTree algorithm in Microsoft R Client software. we used four levels to build DTs in order to obtain easy and useful DRs. Previous studies [32, 33, 2] used the same number of levels.

Table 3 shows a sample of DRs for each level of DTs in ascending order. In addition, the DRs are arranged in descending order within each level according to the quality of rule. For example, the rules 64 and 65 are composed of four levels and belong to the fourth and first class values, while the rule 14 belongs to the fourth class value and is composed of three levels.

3.3 Analyzing Decision Rules

The DRs for each cluster of accident frequencies for locations(highway) associated with times(24 hours) were discussed in order to clarify the most important accident characteristics(see Table1) within clusters. The DRs reflect the relationship between the attributes of road accident and the associated class label, and thus can be used to predict the class of an accident simply.

3.3.1 Decision Rules for First Cluster

The DRs show that the road type feature is either two-way or a single road and is associated with rural or urban areas with a speed of vehicles more than 40 km and the number of casualties is more than two. The road category also is a slope and the severity of accidents is serious. When the road class is an expressway in the urban areas and the speed of vehicles is less than 40km, the severity of accidents is fatal. The motorway road is related to day light condition with one injury or more than one injury and a speed more than 40 km. There is no carriageway hazards in this cluster, and the weather is raining or fine, as well as the road surface is dry or wet.

3.3.2 Decision Rule for Second Cluster

In second cluster, DR shows that accident severity is serious to the absence of light, and the number of injuries is often more than 2. Road hazards include animals and pedestrian hit that occur in rural areas when the weather is fogy, the type of road is one-way, the speed of vehicles is less than 40km in the expressway, and the road surface is slippery or wet. Here the severity of accidents is fatal.

3.3.3 Decision Rules for Third Cluster

DRs indicate that a road class is a slope or curved with night light condition. The focus was on the slope road and single road type with a slight severity of accidents in the urban areas. When the road class is an expressway and there is no light, the severity of accidents is fatal. The speed of vehicles is less than 40 km with a single road type, and the number of injuries is often greater than 2.

Table 3. Classification rules for clusters

No	Rules (IF…)	Class	Accuracy	coverage	Quality
Level 2					
1	"lc"='sl' and "rc"='b'	Fourth	38.93%	1.60%	0.62%
2	"rsc"='ot' and "rc"='a'	Third	32.37%	0.13%	0.04%
3	"rt"='ow' and "sl"='more40k'	Second	42.30%	0.04%	0.02%
Level 3					
4	"rc"='c' and "ura"='ur' and "sl"= 'less40k'	Fourth	43.39%	7.06%	3.06%
5	"ura"='ur' and "rt"= 'sn' and "rc"='b'	Fourth	27.84%	9.80%	2.73%
6	"sl"='less40k' and "rt"='sn' and "rc"='b'	Fourth	27.23%	9.61%	2.61%
7	"rsc"='dr' and "sl"='less40k' and "lc"='sl'	Fourth	35.58%	5.47%	1.94%
8	"wc"='fi' and "rt"='du' and "ura"='ur'	Fourth	27.02%	6.07%	1.64%
9	"sl"='more40k' and "rt"='du' and "ura"='ru'	First	49.85%	2.84%	1.42%
10	"lc"='dl' and "rt"='du' and "ura"='ru'	First	49.7%	2.30%	1.14%
11	"sev"='sl' and "rt"='du' and "ura"='ru'	First	48.58%	2.31%	1.12%
12	"noc"='>2' and "rt"='du' and "sl"='more40k'	First	40.82%	2.12%	0.86%
13	"rt"='sn' and "rc"='b' and "sl"='more40k'	Second	34.33%	2.28%	0.77%
14	"wc"='oth' and "lc"='dl' and "rsc"='dr'	First	40.01%	1.82%	0.72%
15	"rc"='u' and "sl"='less40k' and "wc"='oth'	First	37.34%	1.63%	0.60%
16	"rc"='m' and "lc"='dl' and "noc"='>2'	First	47.86%	1.16%	0.55%
17	"wc"='ra' and "rt"='du' and "ura"='ru'	First	56.71%	0.34%	0.19%
18	"noc"='2' and "rt"='du' and "ura"='ru'	First	54.71%	0.34%	0.19%
19	"rc"='m' and "lc"='nl' and "noc"='>2'	First	44.37%	0.15%	0.06%
20	"ch"='an' and "sev"='sl' and "lc"='nl'	Fourth	35.61%	0.06%	0.02%
21	"rsc"='ic' and "lc"='nl' and "rc"='a'	Fourth	32.35%	0.03%	0.01%
22	"rsc"='ic' and "lc"='dl' and "rc"='m'	Second	66.66%	0.01%	0.007%
23	"rsc"='ot' and "rc"='a' and "lc"='nl'	Fourth	36.84%	0.01%	0.006%
24	"ch"='an' and "sev"='f' and "noc"='>2'	Second	66.66%	0.002%	0.0001%
Level 4					
25	"ch" = 'no' and "lc" = 'dl' and "rt" = 'sn' and "rc" = 'a'	Third	28.33%	16.92%	4.79%
26	"sl"='less40k' and "rt"= 'sn' and "rc" = 'a' and "ura" = 'ur'	Third	31.24%	12.07%	3.74%
27	"noc" = '1' and "lc" = 'dl' and "rt" = 'sn' and "rc" = 'a'	Third	30.78%	12.05%	3.71%
28	"sl"='more40k' and "rt" = 'sn' and "lc"='dl' and "ura"= 'ru'	Third	30.46%	7.78%	2.73%
29	"sl"='more40k' and "rt"='sn' and "lc"='dl' and "ura"='ru'	Third	30.46%	7.78%	2.37%
30	"wc"='fi' and "rt"= 'sn' and "rc"= 'b' and "sl" = 'less40k'	Fourth	28.14%	8.20%	2.29%
31	"wc"='fi' and "rt"='du' and "ura"='ru' and "sl" = 'more40k'	First	48.72%	2.39%	1.16%
32	"rsc"='dr' and "sl"='more40k' and "rt"= 'du' and "ura"='ru'	Fourth	49.55%	2.20%	1.09%
33	"ch"='no' and "lc" = 'dl' and "rt" = 'du' and "ura" = 'ru'	First	50.54%	2.13%	1.07%
34	"sl"='more40k' and "rt"='sn' and "lc" ='dl' and "ura "='ur'	First	33.24%	3.09%	1.02%
35	"rsc"='dr' and "sl"='more40k' and "rt"='du' and "ura" ='ur'	Fourth	30.42%	3.18%	0.96%
36	"lc" = 'sl' and "rc" = 'a' and "rt" = 'sn' and "ura" = 'ur'	Third	36.09%	2.26%	0.81%
37	"rsc"='dr' and "sl"='more40k' and "rt"='sn' and "ura" ='ur'	First	29.36%	2.78%	0.81%
38	"rc"='a' and "rt"='sn' and "sl"= 'more40k' and "ura" ='ur'	First	34.37%	2.13%	0.73%
39	"rt"='sn' and "rc"='a' and "sl" ='more40k' and "ura" ='ur'	First	34.37%	2.13%	0.73%
40	"ura"='ur' and "rt"='sn' and "rc"='a' and "sl"='more40k'	First	34.37%	2.13%	0.73%
41	"lc"='dl' and "rt" ='sn' and "rc"= 'b' and "sl" ='more40k'	Second	34.31%	1.91%	0.65%
42	"rc"='b' and "lc" ='dl' and "sl"= 'more40k' and "wc" = 'fi'	Second	35.46%	1.80%	0.64%
43	"wc" = 'fi' and "rt"= 'sn' and "rc"='b' and "sl"='more40k'	Second	34.65%	1.84%	0.62%
44	"ura"='ru' and "rt" = 'sn' and "rc" = 'b' and "lc" = 'dl'	Second	34.71%	1.74%	0.60%
45	"wc"='ra' and "rt"='sn' and "rc"='a' and "noc"='1'	Third	32.16%	1.87%	0.59%

3.3.4 Decision Rule for Fourth Cluster

DR indicates the light condition is street light or day light with a dry road surface and a speed is less than 40 km in the expressway. The severity of accidents is fatal especially when the road class is a slope. There are Pedestrian accidents in the urban areas, even the road type is dual carriageway or one-way street.

The classification rules for first, second, third and fourth clusters are slightly similar with different interesting measures. There are some common factors of accidents between class values (i.e., first, second, third, and fourth cluster) such as, severity of accidents is slight for all clusters and fatal for second and third clusters. All clusters involve the day light condition without the fourth cluster. The number of accident injuries is often equal to 1 or 2 for all clusters while it is often more than 2 for the first and second clusters. The vehicles speed that is greater than 40km concentrated in the first and second clusters , and the speed that is less than 40km concentrated in the third and fourth clusters. The single road type and the rural areas are common within all clusters except the fourth cluster. The weather condition is either fine or rainy for all clusters except the weather for fourth cluster is just fine.

Table 3. Continued

No	Rules (IF…)	Class	Accuracy	coverage	Quality
Level 4					
46	"rc"='a' and "rt"='du' and "ura"='ru' and "sl"='more40k'	First	48.72%	0.99%	0.48%
47	"ura"='ru' and "rt"='du' and "rc"='a' and "sl"='more40k'	First	48.72%	0.99%	0.48%
48	"noc"='1' and "lc"='dl' and "rt"='du' and "ura"='ru'	First	54.93%	0.70%	0.38%
49	"sl"= 'more40k' and "rt" = 'sn' and "lc"='nl' and "rc" = 'a'	Second	27.77%	0.95%	0.26%
50	"rc"='m' and "lc"='dl' and "noc"='1' and "rsc"='dr'	First	38.33%	0.65%	0.25%
51	"rsc"='w' and "lc"='dl' and "rt"='du' and "ura"='ru'	First	52.10%	0.38%	0.19%
52	"rsc"='w' and "lc"='nl' and "rc"='a' and "rt"='sn'	Second	30.18%	0.60%	0.18%
53	"ura"='ur' and "rt"='du' and "rc"='b' and "lc"='dl'	Second	32.72%	0.56%	0.18%
54	"rc"='b' and "lc"='dl' and "sl"='less40k' and "wc"='oth'	First	40.74%	0.35%	0.14%
55	"sev"='s' and "sl"='more40k' and "rc"='m' and "lc"='dl'	First	47.52%	0.28%	0.13%
56	"noc"='2' and "rt"='sn' and "rc"='a' and "ura"='ur'	Third	27.03%	1.87%	0.50%
57	"rt"='ro' and "sl"='more40k' and "rc"='a' and "ura"='ru'	First	42.60%	0.20%	0.08%
58	"rc"='c' and "ura" ='ur' and "sl"= 'more40k' and "rt"= 'du'	Fourth	41.26%	0.05%	0.03%
59	"rt"='un' and "wc"='fi' and "rc"='a' and "ura"='ur'	First	61.40%	0.05%	0.02%
60	"rsc"='ot' and "rc"='a' and "lc"='dl' and "ch"='no'	Third	32.58%	0.08%	0.02%
61	"noc"='2' and "rt"='sn' and "rc"='b' and "lc"='nl'	Second	39.34%	0.05%	0.02%
62	"sev"='f' and "lc"='nl' and "rc"='a' and "noc"='1'	Fourth	57.40%	0.05%	0.02%
63	"noc"='1' and "lc"='nl' and "rc"='a' and "rt"='du'	Second	34.52%	0.07%	0.02%
64	"wc"= 'oth' and "lc" = 'nl' and "rsc" = 'dr' and "rc" ='c'	Fourth	41.26%	0.05%	0.02%
65	"wc"='oth' and "lc"='nl' and "rsc" = 'dr' and "rc"='a'	First	42.85%	0.03%	0.01%
66	"rt"='ro' and "sl"='less40k' and "rc"='b' and "wc"='oth'	First	47.36%	0.03%	0.01%
67	"rsc"='w' and "lc"='nl' and "rc"='a' and "rt"='du'	Second	39.47%	0.03%	0.01%
68	"sl"='less40k' and "rt"='du' and "rc"='b' and "wc"='oth'	First	58.06%	0.02%	0.01%
66	"ch"='p' and "ura"='ur' and "rc"='a' and "rt"='du'	Fourth	53.84%	0.01%	0.006%
70	"rc"='m' and "lc"='dl' and "noc"='1' and "rsc"='ic'	Second	100%	0.05%	0.005%
71	"wc"='oth' and "lc"='dl' and "rsc"='ic' and "rt"='du'	Second	75%	0.007%	0.005%
72	"ch"='p' and "ura"='ur' and "rc"='a' and "rt"='ow'	Fourth	100%	0.001%	0.003%
73	"rc"='u' and "sl"='more40k' and "wc"='fog' and "noc"='>2'	Third	42.85%	0.006%	0.002%
74	"rt"='du' and "ura"='ur' and "rc"='b' and "sev"='f'	First	80%	0.004%	0.003%
75	"rt"='ro' and "sl"='more40k' and "rc"='m' and "ura"='ur'	First	66.66%	0.005%	0.003%
76	"wc"='sno' and "lc"='nl' and "rc"='a' and "noc"='>2'	Third	50%	0.001%	0.0009%

4. CONCLUSION

Data mining techniques such as clustering and classification are widely used in the analysis of road accident data, because these techniques have the ability to extract knowledge from very large data without relying on a prior underlying relationships between data variables. In this study, we proposed a framework for analyzing times of road accident frequencies that uses k-means clustering and DT algorithm. This is the first time that both approaches have been used together. The K-means algorithm was used to identify four clusters(C1-C4) based on accident frequencies for each location within 24 hours over the 3-year period. The DT algorithm is non-linear and non-parametric data mining techniques for supervised classification and regression problems. However, extracting DRs from the DT is restricted by the DT's structure, which does not allow us to extract more knowledge from a dataset. A particular method to increase the number of valid rules that are extracted from DT is used, in this method many DTs are generated for each variable under study (variables that describe the data) by using root node variation for the same tree. Also, this ensemble DTs method can be integrated with DT algorithms that have the ability to handle very large data. Therefore, the Streaming Parallel Decision Tree (SPDT) algorithm is used to build DTs. We applied the previous ensemble method to all clusters at once to obtain unique DRs instead of discovering rules for each cluster independently. Although the data used in our approach was very large in addition to the class-imbalance problem, the extracted DRs were more than 76 valid rules that used for identifying the causes of road accidents within each cluster. Finally, this data mining approach can also be reused on other accident data with more attributes to cover more information.

5. ACKNOWLEDGMENTS

We are thankful to data.gov.uk Department for Transport to provide data for our research

6. REFERENCES

[1] Abellán,J., Masegosa, R., A., 2010. An ensemble method using credal decision trees. European Journal of Operational Research. 205, 218–226.

[2] Abellán, J., López, G., Oña, J., D., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications. 40, 6047–6054.

[3] Abellán, J., & Moral, S. 2003. Building classification trees using the total uncertainty criterion. International Journal of Intelligent Systems, 18(12), 1215–1225.

[4] Barua, S., Basyouny, K., E., Islam, M.,T., 2014. A Full Bayesian multivariate count data model of collision severity with spatial correlation. Anal. Meth. Accid .Res. 3-4 , 28-43.

[5] Behnood, A., Roshandeh, M., A., Mannering , L., F., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Anal. Meth. Accid .Res. 3-4, 56-91.

[6] Barua, S., Basyouny, K., E., Islam, M.,T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Anal. Meth. Accid .Res. 9, 1-15.

[7] Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

[8] Behnood, A., Mannering, L., F., 2016. An empirical assessment of the effects of economic recessions on pedestrian injury crashes using mixed and latent-class models. Anal. Meth. Accid .Res. 12,1-17.

[9] Chang, L., Y., Chen, C., W., 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research. 36,365–375.

[10] Chonga, L., S., Tyebally, A., Chew, Y., S., Lim, Y., C., Feng, X., Y., Chin, T., S., Lee, L., K., 2017. Road traffic injuries among children and adolescents in Singapore – Who is at greatest risk?. Accid. Anal. Prev. 100, 59-64.

[11] Calaway, R., 2016. Estimating Decision Tree Models. Microsoft. Developer Network. https://github.com/richcalaway.

[12] Cehrke, J., Ramakrishnan, R., Ganti, V., 2000. RainForest-A framework for fast decision tree construction of large datasets. Data mining and Knowledge discovery. 4, 127-162.

[13] Chang, L., Y., Chien , J., T., 2013. Analysis of driver injury severity in truck-involved accidents using a non -parametric classification tree model. Safety Science. 51, 17–22.

[14] De Oňa,J., López, G., Mujalli, R., Calvo, F.J., 2013. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accid. Anal. Prev.51, 1-10.

[15] Elvik, R., 2017. Exploring factors influencing the strength of the safety-in-numbers effect. Accid. Anal. Prev. 100, 75–84.

[16] Gehrke, J., Ganti, V., Ramakrishnan, R., 1999. BOAT-Optimistic Decision Tree Construction. CiteCeer^{x}. 114-2.

[17] Gehrke, J., Ramakrishnan, R., Ganti, V., 2000. RainForest-A Framework for Fast Decision Tree Construction of Large Datastes. Data Mining and Knowledge Discovery. 4, 127-162.

[18] Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. Anal. Meth. Accid .Res. 14, 10-12.

[19] Han, J., Kamber, M., Pei, J., 2012. Data mining concepts and techniques. The Morgan Kaufmann Series in Data Management Systems. Third ed. Morgan Kaufmann Publishers. Waltham. MA.

[20] Haim, Y., Tov, E., 2010. A Streaming Parallel Decision Tree Algorithm. Journal of Machine Learning Research 11, 849-872.

[21] Kim, M., Kho, Y., S., Kima, K., D., 2017. Hierarchical ordered model for injury severity of pedestrian crashes in South Korea. Journal of Safety Research. xxx, xxx–xxx.

[22] Kumar, S., Toshniwal, D., 2016a. A novel framework to analyze road accident time series data. Journal of Big Data. 3:8,DOI 10.1186/s40537-016-0044-5.

[23] Kumar, S., Toshniwal, D., 2016b. Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). Journal of Big Data. 3:13, DOI 10.1186/s40537-016-0046-3.

[24] Kumar, S., Toshniwal, D., 2016c. A data mining approach to characterize road accident locations. J. Mod. Transport. 24(1):62–72, DOI 10.1007/s 40534-016-0095-5.

[25] Kwon, O., H., Rhee, W., Yoon, Y., 2015. Application of classification algorithms for analysis of road safety risk factor dependencies. Accid. Anal. Prev. 75, 1–15.

[26] Kassamara, A., 2015. determining the optimal number of clusters: 3 must known methods – unsupervised Machine learning. STHDA. http://www.sthda.com/english/wiki/determining-the optimal-nubmer-of-clusters-3-must-known methods-unsupervised-machine-learning

[27] Kumar, S., Toshniwal, D., 2015. A data mining framework to analyze road accident data. J. Big. Data. 2(1), 1–26.

[28] Kidd, D.,G., Buonarosa, M., L., 2017. Distracting behaviors among teenagers and young, middle-aged, and older adult drivers when driving without and with warnings from an integrated vehicle safety system. Journal of Safety Research. 61, 177–185.

[29] Kashani, A., T., Mohaymany, A., S., Ranjbari, A., 2010. A Data Mining Approach to Identify Key Factors of Traffic Injury Severity. Promet–Traffic&Transportation. Vol. 23. No. 1, 11-17.

[30] Kaufman, L., Rousseeuw, P., J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

[31] Mannering, F.L., Shankar, V., Bhat,C.R , 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Meth. Accid. Res. 11, 1-16.

[32] Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F. , 2011. Data mining techniques for exploratory analysis of pedestrian crashes. Transportation Research Record. 2237, 107–116.

[33] Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid. Anal. Prev. 49, 58-72.

[34] O'Herna, S., Oxleya, J., Stevenson, M., 2017. Validation of a bicycle simulator for road safety research. Accid. Anal. Prev. 100, 53-58.

[35] Plant, B., R., C., Irwin, J., D., Chekaluk, E., 2017. The effects of anti-speeding advertisements on the simulated driving behaviour of young drivers. Accid. Anal. Prev. 100,65-74.

[36] Prati, G., Pietrantoni, L., Fraboni, F., 2017. Using data mining techniques to predict the severity of bicycle crashes. Accid. Anal. Prev. 101, 44–54.

[37] Quinlan, J., R., 1986. Induction of decision trees. Mach. Learn. 1, 1, 81-106.Analysis and Prevention. 49, 58–72.

[38] Quinlan, J. R., 1993.C4.5: Programs for machine learning. San Mateo. California: Morgan Kaufmann Publishers.

[39] Rovšek, V., Batista, M., Bogunović, B., 2014. Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a nonparametric classification tree. TRANSPORT. ISSN 1648-4142 print. ISSN 1648-3480 online. First. 1–10, doi:10.3846.16484142.915581.

[40] Savolainen, T.,P. , Mannering, L., F., Lord, D. , Quddus, A., M., 2011. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. Accid. Anal. Prev. 43,1666-1676.

[41] Sarwar, T., M., Anastasopoulos, P., C., Golshani, N., Hulme, K., F., 2017. Grouped random parameters bivariate probit analysis of perceived and observed aggressive driving behavior: A driving simulation study. Anal. Meth. Accid .Res. 13, 52-64.

[42] Tan, P., N., Steinbach, M., Kumar, V., 2006. Introduction to data mining. Pearson Addison-Wesley.

[43] Xie, K., Wang, X., Ozbay, .K, Yang, H., 2014. Crash frequency modeling for signalized intersections in a high-density urban road network. Anal. Meth. Accid .Res. 2, 39-51.

[44] Xu, X., Šarić, Z., Kouhpanejade, A., 2014. Freeway Incident Frequency Analysis Based on CART Method. Promet Traffic&Transportation. Vol. 26. No. 3, 191-199.

[45] Yasmin, S., Eluru, N., Bhat, R., C., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Anal. Meth. Accid .Res. 1, 23–38.

[46] Zeng, Q., Wen, H., Huang, H., 2016. The interactive effect on injury severity of driver-vehicle units in two-vehicle crashes. Journal of Safety Research. xxx , xxx–xxx.

[47] [dataset] Department for Transport, 2016. Road safety data. https://data.gov.uk/dataset/road-accidents-safety data.