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.

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

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).

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

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.

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

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

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 $G(\{pi\}) \geq 1-\max\{pi\}$, 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 $\Delta(\tau, \{pi\}, \{pL, i\}, \{pR, i\}) = \Delta(v, j, a)$ as

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

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.

$$Accuracy = \frac{TP+TN}{P+N} \quad (6)$$

$$Error\ Rate = \frac{FP+FN}{P+N} \quad (7)$$

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,

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

and the error rate of the rule is,

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

2.3.1.2 Coverage

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

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

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:

$$Quality = Accuracy\ x\ Coverage \quad (11)$$

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.

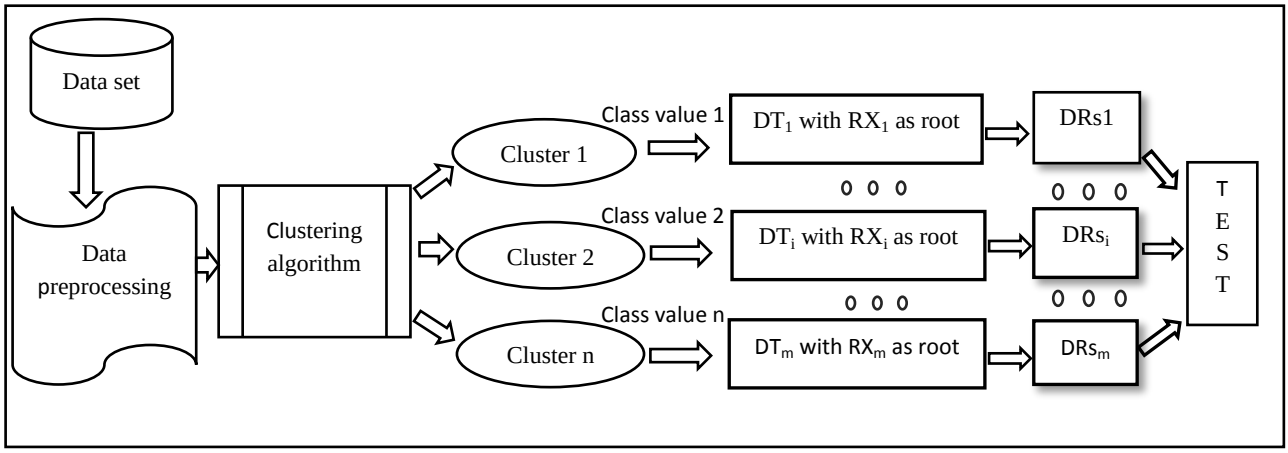


Figure 1. Proposed Framework

Table 1. Attribute Description

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

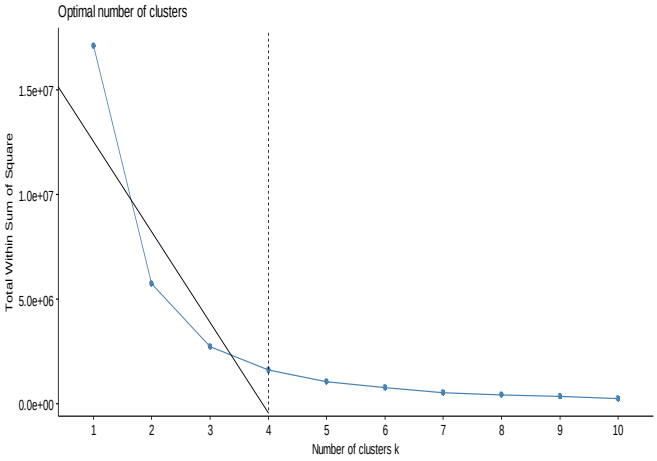


Figure 2. Number of selected clusters

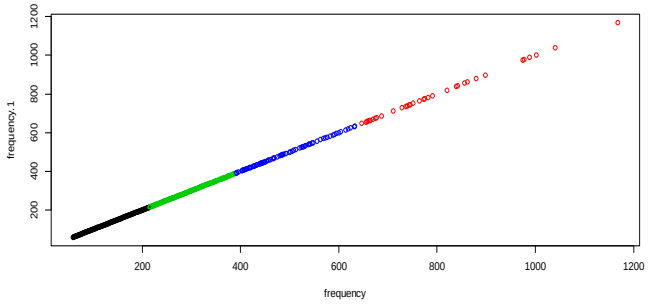


Figure 3. Distribution of clusters

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.

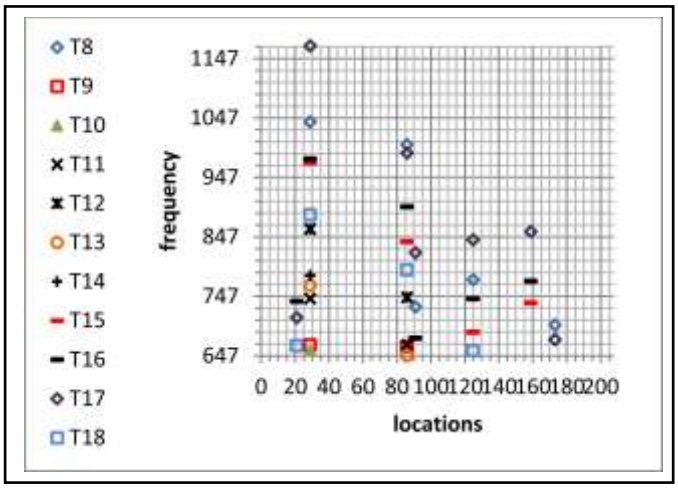


Figure 4. Times (T8 - T18) and locations (21,29,86,91,125,159,173) of First cluster

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

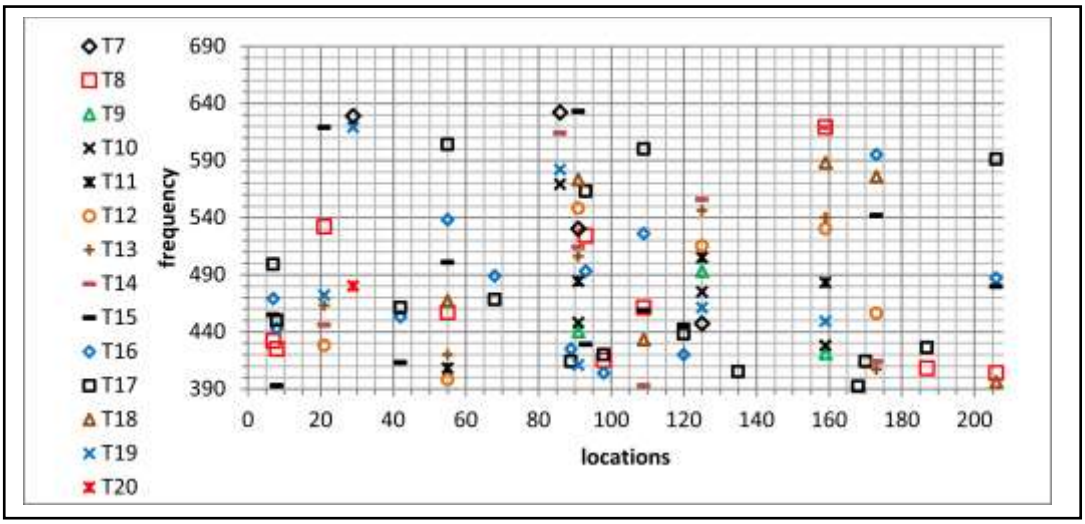


Figure 5. Times (T7-T20) and locations
(7,8,21,29,42,55,68,86,89,91,93,98,109,120,125,135,159,168,170,173,187,120) of Second cluster

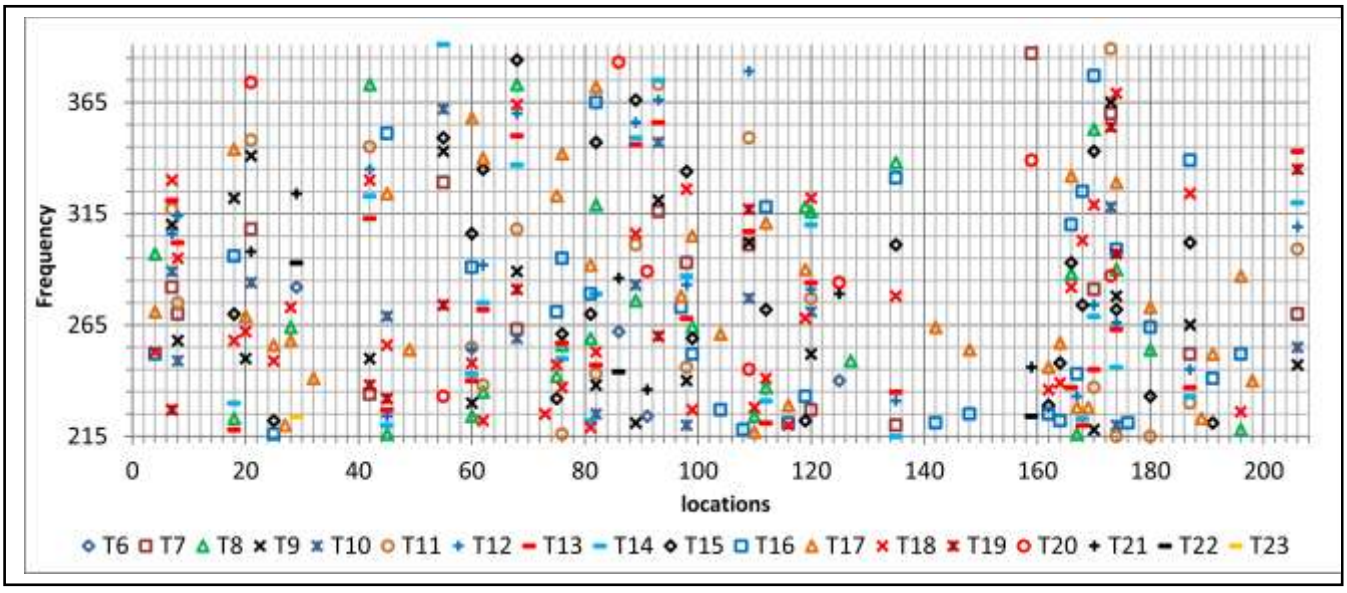


Figure 6. Times(T6-T23) and locations(4,7,8,18,20,21,25,27,28,29,32,42,45,49,55,60,62,68,73,75,76,81,82,86,89,91,93,97,98,99, 104,108,109,110,112,116,119,120,125,127,135,142,148,159,162,164,166,167,168,169,170,173,174,176,180,187,189,191,196, 198, 206) of Third cluster.

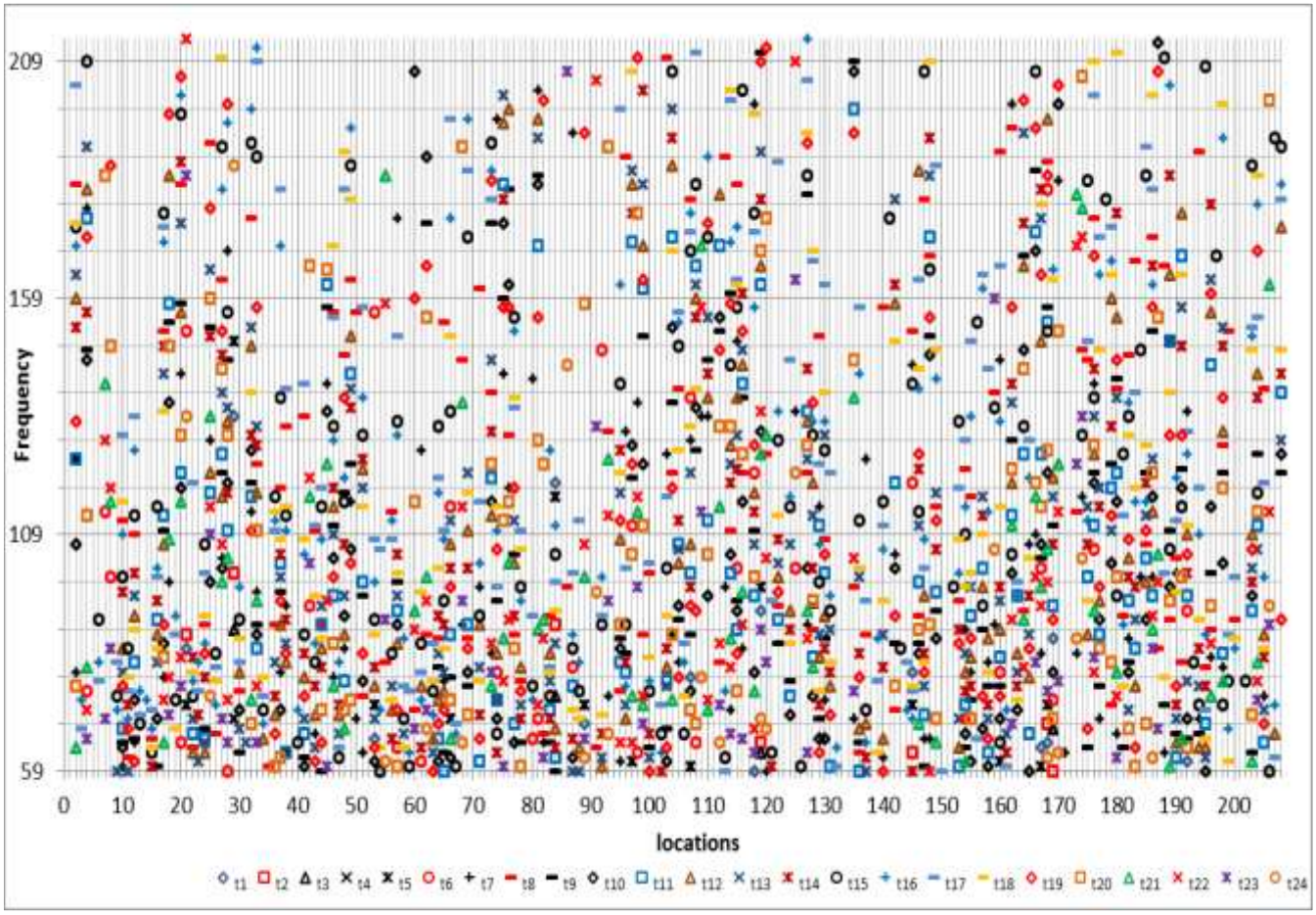


Figure 7. Times(T1-T24) and locations of Fourth cluster.

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 Oña,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.

International Journal of Computer Applications Technology and Research
Volume 7–Issue 01, 12-18, 2018, ISSN:-2319–8656

Presenting a Model for Identifying the Best Location of Melli Bank ATMS by Combining Clustering Algorithms and Particle Optimization

Abdolhussein Shakibayinia
Department of Computer, Dezful Branch,Islamic Azad University, Dezful,Iran

*Faraz Forootan
Department of Computer, Dezful Branch,Islamic Azad University, Dezful,Iran

Abstract: The Interbank Information Exchange Network (Shetab or Acceleration) has started since 2002 and the purpose was integrating and connecting card systems of all banks in the country. Currently, the Acceleration Center has been acting as Melli bank card switch in the country, and all the banks in the country are its member. These operations cover a wide range of transactions, such as cash withdrawals, electronic purchases, fund transfers, paying bills and residual payments. Shetab center processes more than two and a half million transactions per day. At present, the amount of fees received from each network transaction is 500 to 22,000 Rial, which is considered as a fee for the client's bank as revenue and for the client bank. And it does not cost any expenses to the customer, thus banks are looking for earning revenue from this service.
In this, first the list of ATMs that Melli Bank pays them service fee are considered, then by using the clustering algorithm, locations were arranged for an ATM so Melli Bank pay less fee. In this study, the combination of three K-means algorithms and particle optimization algorithm and genetic algorithm were used. Davies-Bouldin Index was used to assess clustering.
Then, the proposed clustering along with another clustering algorithm was evaluated and it was shown that the proposed algorithm is performing better. 8 locations for ATM were presented in proposed clustering algorithm, which is the result of the proposed clustering.

Keywords: Particle optimization algorithm, Banking, ATM

1. INTRODUCTION

Today, the banking system consisting of the central bank and commercial banks play a decisive role in the economic development of the country. Banks are using the most advanced technology and needed tools and the most extensive international networks in the world, to be able, through a variety of computer services, make monetary deposits legal and legal persons at the best possible cost and at the lowest cost and with the most secure way, from one place to another, or from an account to another account as soon as possible [1].
The payment system is a mechanism that transfers money from an account in a bank to an account in another bank, thus the role of the paying system in the economy is like the veins that bring money to different economic firms, so the reality is that the accurate management and supervision and smoothly functioning of the payment system in the monetary part of the country is one of the main duties of central banking in the today world [2].
Shetab started in 2002 with the aim of integrating and linking card systems of all banks in the country. At present, the Acceleration Center has play a role as Melli bank card switch in the country, and all the banks in the country are its members. The operations cover a wide range of transactions, such as cash withdrawals, electronic purchases, fund transfers, paying bills and residual payments. Shetab processes more than two and a half million transactions per day. At present, the amount of fees received from each network transaction is 398 Tomans, which is for the bank as revenue and for the client's bank as a cost and does not create any costs for the client.

The advancement in communications science and modern banking technology has led the service system to faster delivery and more consistent services. Today, the level of access to these facilities is almost the same for all financial institutions and banks and banking services companies. In fact, the point that makes a bank superior to other financial institutions and banks is providing a distinct, fast and continuous service. ATM is one of the tools that can help customers accelerate automated tasks [3].
Automated Teller Machine (ATM) is a device that can be used for depositing and withdrawing from customer accounts, changing card and account statements, transferring funds, paying bills, account balance and some other services through a Melli card without the need for the operator. ATMs are the most prominent feature of the evolution of banking services based on modern technology. Based on the policies of the Melli Bank of Iran to develop the use of modern banking systems, since 1997, the installation of the first ATM series in the selected branches began, and so far, with the installation of more than 6,600 ATMs, the Melli Bank of Iran has the largest number of installed ATMs in among Banks in the country.
Locating new ATMs is very important for the organization due to its rapid support, costs and payment fees for the transaction of other bank cards (other than the Melli Bank), and making related decisions are very sensitive, important and worthwhile. This is the responsibility of senior officials and informatics supervised by the provincial branch office. For example, the salary paid to one of the ATMs of the Bank of SADERAT has been 434,225,087 Rials in 2015, thus to decide and choose the best location to other branches of the Melli Bank which are

closer also for better maintenance. The more important places are reportedly by Accelerator Center, and the Melli Bank ATM card transactions have accepted the most competitive and large fees paid to them. To do this, we have chosen the western region of Ahwaz to explore and derive the final model, which will calculate the best location of the new ATM based on the geographical coordinates of ATMs of the Melli Bank and the rival [4].

In this paper, by combining the k-means clustering algorithms and particle optimization, with the criteria for increasing support and reducing out-of-service time and reducing payroll fees to other competing banks, we will give the best place to the senior executives of the bank. Gave Clustering is one of the most important issues of non-monitoring learning as well as the most common data mining techniques used to classify data sets into specific subsets. The k-means algorithm is also one of the most popular clustering algorithms with easy implementation and fast performance, but being sensitive to first cluster centers can only produce a local optimal response. By combining the k-means clustering method with the particle group optimization algorithm (due to the existence of decimal data), we improved the accuracy of clustering.

The study consists of two parts. In the first section, the proposed algorithm is presented and in the second section the results of the research are shown..

2. THE PROPOSED ALGORITHM

The proposed algorithm is presented in this section, which has two parts. In the first step, the K-means algorithm is used to select the start points of the K-means algorithm from the particle optimization algorithm, then in the second stage, heuristic algorithms are used to determine the start points of the particle algorithm and then the genetic algorithm is used to optimize the particle optimization algorithm.

2-1. The proposed algorithm

Two cases are considered to use ATM for the withdrawal of money. In the first case, the card that is taken from it is related to the same bank. In the latter case, the card is not from the same bank. In the first case, for example, the Melli Bank card withdrawn from the Melli Bank ATM, and in the latter case, for example, the Melli Bank withdraw money from the Tejarat bank ATM. In the latter case, due to the use of the Tejarat bank ATM, the Melli Bank should pay the Tejarat Bank.

In this research, it was tried that the Melli Bank's customers refer less to ATMs of other banks. To do this, the information about the number of Melli Bank's customers visiting has been removed from other banks and the amount has been deducted and the location of those banks has been analyzed and continue to add new ATMs for the customer of those places. The structure of the data used is shown in Table 1.

Table 1 ATMs

Row	Title		Variable
1	Features	X	Discrete
		Y	Discrete
2	Referrals number		Discrete
3	Withdrawn amount		Discrete

Weighted clustering is used in this data. The data structure is such that a number of parameters are considered as those data parameters, and a number of parameters are considered as data weights. Whatever the weight is given, the data is considered to more distant.

In the bank's data, the two parameters X and Y are considered as data parameters, and the number of referrals and the withdrawn amount is considered as the weight of the ATM. In the data, the more weighted it is the ATM is considered to be farther, so the parameters are reversed so that the number of visits and the amount removed are increased, near ATMs are considered in equation (1) the calculation of the points' weight is shown.

$$W_i = \frac{1}{\text{number of referrals}} + \frac{1}{\text{the withdrawn amount}} \quad (2)$$

In which, both the relationship between the number of referrals and the withdrawn amount are reversed and its total is considered as the weight of the bank.

In the key clustering algorithm, the parameter is a distance, with the difference that this distance is considered in the relation of the record weight. This calculation is shown in equation (2).

$$d(i,j) = \sqrt{\left(w_i x_i - w_j x_j\right)^2 + (w_i y_i - w_j y_j)^2} \quad (1)$$

The difference between each parameter in the weight of the number of referrals and the amount of withdrawal is multiplied to determine its effect on the distance between ATMs.

The proposed algorithm consists of two first parts in the first phase of the particle optimization algorithm and the K-means algorithm are combined. In the second step, the heuristic algorithms are used to determine the starting points of the particle algorithm, and then the genetic algorithm is used to optimize the particle optimization algorithm.

2-2. Problem solving by combining particle optimization and K-means algorithm

The problem solving structure is that the genetic algorithm is responsible for determining the starting points of the K-Means algorithm. The K-Means algorithm depends on the starting points. If the starting points are well chosen, the output of the K-Means algorithm will be better, and if the points are not selected, then the output is not good in terms of clustering. The parameters X and Y are discrete and the particle optimization algorithm is continuous, so the parameters of X and Y are transmitted to the interval between zero and one. In order to transfer data from a discrete interval to a continuous, the equation (3) is used [23].

$$Value_{new} = \frac{Value_{old} - min}{max - min} \quad (3)$$

The structure of the particle optimization algorithm is then used to determine the starting points of the K-means algorithm.

2-2-1 Particle Structure

Particles represent the starting points of the K-means clustering center. The number of points in the K-means algorithm is as large as the number of clusters. If the algorithm has two clusters, then two start points are required, and if the K algorithm is a cluster, then K is the starting point. Given that the data has 2 features, each point has 8 values between 0 and 1. Therefore, considering that clustering has a K point and each point has 2 values and its value is between 0 and 1, this structure is shown in Fig. 1.

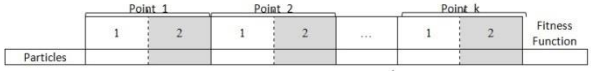


Figure 1 **Structure of proposed particle of the first stage**

2-2-2 Production of initial population

First, 20 particles are created and they are randomly range from 0 to 1. For the production of the initial population, a uniform distribution is used, and in (4) the manner in which a uniform distribution between numbers 0 and 1 is shown.

$$\text{Chromosome} = \text{U}(0,1) \tag{4}$$

2-3 Calculation competency function

In calculating the competency function this is how each particle is proportional to a K-means clustering. Each particle is the starting point of the K-means algorithm. Then the K-means algorithm is executed. After the K-means algorithm converges, the accuracy of clustering as a competency function will be sent.

Given that each particle is fully clustered, the result of clustering is calculated by the clustering accuracy parameter and its output is considered as the particle computation suitability. The accuracy of the Davies-Bouldin Index is used to calculate the accuracy [17].

The Davis-Bouldin method is a function of the total ratio of intra cluster dispersion to the distance between clusters. The Davis-Bouldin Validation Index is shown by equation (5), the Davis-Bouldin Method operates on the basis of minimization.

$$DB = \frac{1}{n}\sum_{\substack{i=1 \\ i\neq j}}^{n} max\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}\right\} \tag{5}$$

In which n is the number of clusters, Sn is the mean of the cluster data spacing from the cluster center and S(Qi,Qj) is the distance between the centers of the clusters; therefore, when they are within the cluster and clusters are far from each other, this ratio gets small. The small amount of the Davis-Bouldin validation index is valid clustering representation. In Fig. 2, the structure of the calculation of the competency function is shown.

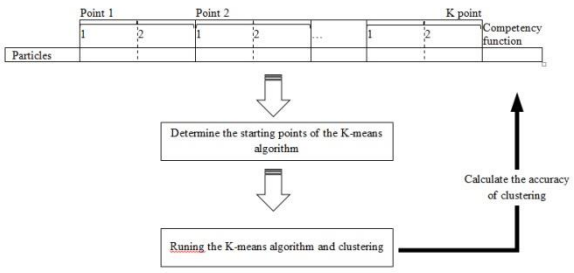


Figure 2 the structure of the proposed competency function

2-2-4 Speed control

One of the important aspects for determining the efficiency and accuracy of an optimization algorithm is how to reconcile Explore and Exploit with the proposed algorithm.

The Explore feature is the ability to search an algorithm in different areas of the search space to find optimal amount. On the other hand, the Exploit feature is the ability to focus the search around a likely area to improve the candidate's solution. Thus, we created an appropriate solution between these two conflicting goals, which is achieved by speeding up the PSO, as shown in equation (6).

$$v_{ij}(t+1) = v_{ij}(t) + c_1 r_{1j}(t)\left[y_{ij}(t) - x_{ij}(t)\right] + c_2 r_{2j}(t)\left[\tilde{y}_j(t) - x_{ij}(t)\right]$$

$$v_{ij}(t+1) = \begin{cases} v_{ij}(t+1) & if\ v_{ij}(t+1) < v_{max,j} \\ v_{max,j} & if\ v_{ij}(t+1) \geq v_{max,j} \end{cases}$$

In which $v_{max,j}$ is the highest speed in the number of tables and the number of columns, the value of $v_{max,j}$ is very important. Because it speeds up search through inhibition. If the value is large. The Explore feature will increase the algorithm. While the small values of this parameter, the local Exploit feature improves the algorithm. If $v_{max,j}$ is too small, congestion may not be as good as local areas. In addition, there is the possibility of clustering in the optimal local area, which will not be able to get out of it for the algorithm. On the other hand, the large amounts of $v_{max,j}$ have the risk of losing good areas. Particles may jump through good solutions and search for inappropriate areas. Large values result in the algorithm moving away from the optimal area. In this case, the particles move faster.

Finding the right amount of $v_{max,j}$, in order to establish two types of equilibrium as follows:

- Fast or slow motion
- Explore and Exploit

$$\text{Chromosome} = \text{U}(0,1) \tag{7}$$

In which $x_{max,j}$ and $x_{min,j}$, respectively, are the minimum and maximum value of the table number and the column number in both dimensions, and $\delta \in (0,1]$ the δ value is initially equal to one, and in each generation, the δ value of (8) changes. The amount of δ each generation is 90% less than the previous generation.

$$\delta = 0.9^i \text{ , i} = \text{generation number} \tag{8}$$

2.2.5 The condition for stopping the algorithm

The condition to end the algorithm is based on the congestion radius. The criterion for the end of the algorithm is close to zeroing the normalized congestion radius. We calculated the normalized congestion radius from (9).

$$R_{norm} = \frac{R_{max}}{diameter(S)} \qquad (9)$$

In which diameter(S) is the diameter of the space in the initial congestion and R_{max} is the maximum radius, calculated from equation (10).

$$R_{max} = \|x_m - \hat{y}\| \quad , m = 1, \ldots, n_s \qquad (10)$$

When R_{norm} is closer to zero, the algorithm stops, $\hat{y}$ is the minimum value of the competency function.

2-2-6 The structure of the particle optimization algorithm

It includes the following six steps:

Step 1: Initialize the parameters of the velocity and location of the particles in the initial crowd according to (4)

Step 2: Updating the best local position of particle i for all particles

Step 3: Improving the best global position of all particles

Step 4: Calculating the new velocity of all particles by means of equation (6)

Step 5: Calculating the new location of all particles by means of equation (11)

$$x_i(t+1) = x_i(t) + v_i(t+1) \qquad (11)$$

Step 6: Repeating steps 2 to 5 with the output condition of the algorithm according to 2-2-5.

2-3 Optimization of Particle Algorithm Using Genetic and Heuristic Algorithm

In the sequel, every generation of the genetic algorithm runs. Using the genetic algorithm increases the search power of the particle optimization algorithm.

2-3-1 Initial population

In the previous step, 20 particles are generated that identify the starting points of the K-means algorithm, thus, K-means algorithm is applied to each particle, and the cluster output is considered as the input of the mantle algorithm. The fragmental particle structure of particle optimization algorithm for the chromosome is shown in Fig. 3.

2-3-2 Combinations

In the study, three methods of One-point Crossover, Two-point Crossover and uniform-point Crossover were used.

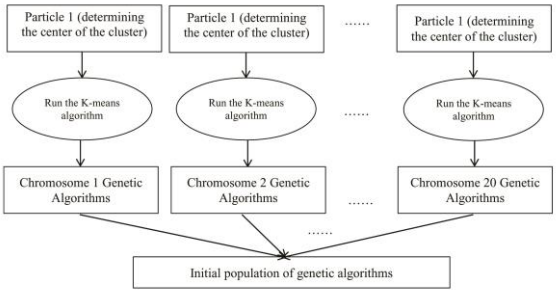


Figure 3 Particle conversion of particle optimization to the initial population of the algorithm

In the single-point Crossover, first, a random point is selected in the sequence of the parent chromosomes, and then from the selected location, the chromosome of both parents is cut. The first part of the first parent and the second part of the second parent are used to produce the first child. The second child consists of the first part of the second parent and the second parent of the first (Figure 4).

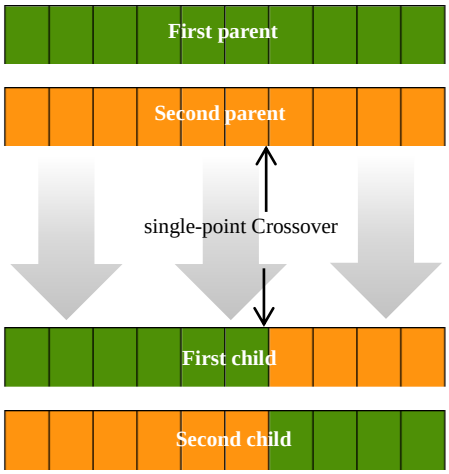


Figure 4 single-point crossover

In the two-point crossover, two random points are selected in the parental chromosome sequence, and then the parent chromosome is cut off from these points. The first and third parts of the first parent and the second part of the second parent are used to produce the first child. The second child consists of the first and third parts of the second parent and the second part of the first parent's department (Fig. 5). Obviously, by increasing the number of breakpoints in the multi-point multiplication operator, the similarity of the parent to each parent decreases and the reciprocity function divergence will be strengthened.

In the uniform combination operator, the value of the child's gene is selected according to the values of the corresponding genes of both parents. In this method, the genes of each parent have equal chances for the presence in the corresponding genes of the child. In the uniform reconciliation operator, it is determined by the random distribution that the amount of each child's gene is selected from the corresponding gene value of the parent (Fig. 6).

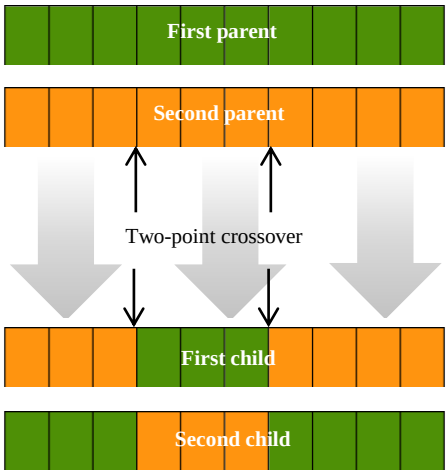


Figure 5 Two-point crossover

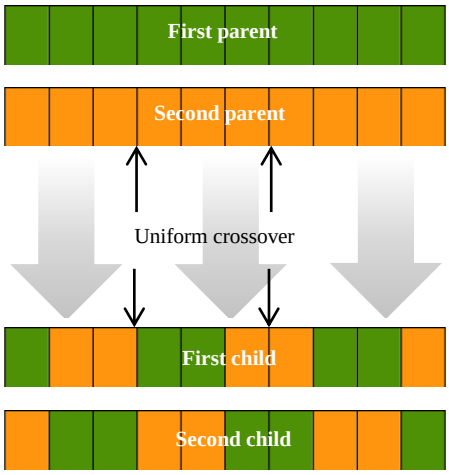


Figure 6 Uniform crossover

In this research, three crossovers are used. In the first 50% of the genetic algorithm, the uniform crossovers algorithm is used because the uniform composition has a high divergence, and from 50% to 75% of the genetic algorithm, the two-point algorithm is used, and in 25% of end of the algorithm a single-point crossover is used. The composition structure of the combination in the genetic algorithm is shown in Fig 7.

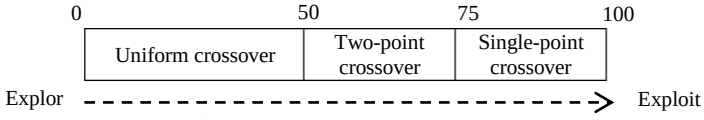


Figure 7 Crossover execution structure

2-3-3 Mutation

The mutation is the random change process of the amount of genes in a chromosome, the main goal in the mutation operator is to find new values for the genes of the offspring (values not found in none of the parents) to increase genotype variation in the population. The mutation should be performed so that the genes do not deteriorate in superior responses. Given that the mutation operator causes a random change in the chromosome, the use of this operator increases the probability of finding new values for the genes. Using more than this operator increases the divergence of the algorithm.

In this research, the contents of a part of a chromosome are selected and re-established. In Fig. 8, the structure of the mutation is observed.

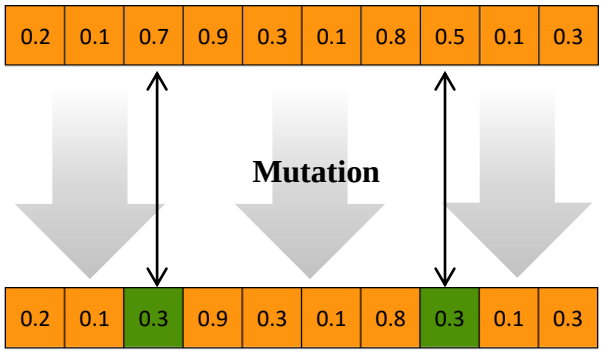


Figure 8 mutation structure

A generational replacement method is used, because this method has the convergence and divergence capability by changing its variables. Initially, the algorithm transfers 50% of the parent and 50% of the children to the next generation, causing a divergence in the problem. After each generation, the percentage of children decreases and the percentage of parents increases to bring about convergence. The percentage of parents' increase is calculated through trial and error.

2-3-5 Condition of ending the algorithm condition

In this research, the condition for ending of the genetic algorithm is to execute 100 generations of genetic algorithms unchanged.

3. EVALUATION OF PROPOSED ALGORITHM

Testing and proving their results is one of the most important parts of a model is. We have developed programs in the language of mathematics to evaluate the proposed model, which we will continue to explain and illustrate the results obtained from them. These tests were performed on the Windows 7 operating system and a 5-core computer with 2.7GH processors and 4GB of memory. The Jaguar standard was used to test this algorithm.

In the evaluation, the proposed algorithm was evaluated using the following four methods.

- K-Means
- K-Medoids
- Hierarchical Clustering
- ANN

The parameters of the proposed algorithm are shown in Table1. These parameters are calculated with trial and error.

Table 1 Parameters of the proposed algorithm

Parameter	Value
Initial population of particle optimization algorithm	100
C_1	2
C_2	2
Initial population of genetic algorithms	100
Combining Genetic Algorithms	0.7
Genetic Algorithm Mutation	0.3
Number of Generations of Genetic Algorithms	100 unchanged generations

The proposed algorithm was implemented with different clusters, and its output is shown in Table 2.

Table 2 Comparing 5 methods of solving bank clustering problem with Davis-Bouldin Index of accuracy

	K-Means	K-Medoids	Hierarchical Clustering	ANN	Proposed algorithm
Cluster 2	0.74	0.53	0.78	0.78	0.61
Cluster 3	0.59	0.49	0.64	0.64	0.31
Cluster 4	0.7	0.54	0.61	0.93	0.23
Cluster 5	0.84	0.53	0.59	0.85	0.18
Cluster 6	0.82	0.52	0.52	1.01	0.18
Cluster 7	0.9	0.49	0.58	1.01	0.13
Cluster 8	0.92	0.48	0.52	1.01	0.11
Cluster 9	0.92	0.53	0.5	1.05	0.14
Cluster 10	0.95	0.53	0.47	1.05	0.14
Cluster 11	0.93	0.53	0.44	1.01	0.15
Cluster 12	0.9	0.53	0.44	0.97	0.16

As in Table 2, the proposed algorithm works better than other methods. Figure 9 shows the comparison of the best output of each algorithm. As shown in Fig. 9, the proposed algorithm works better than other methods.

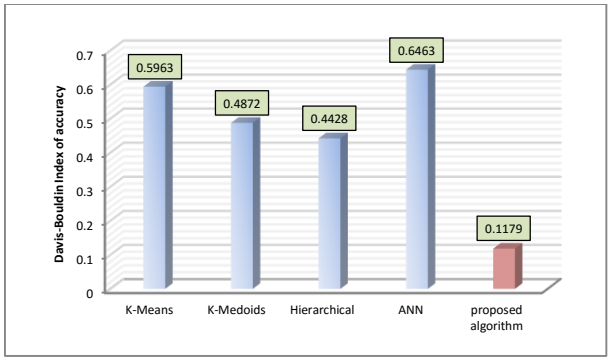


Figure 9 Comparison of proposed algorithm with 4 other algorithms with Davis-Bouldin Index of accuracy

4. CONCLUSIONS AND FUTURE STUDIES

Choosing location is one of the important factors in economic enterprises activities. Due to this importance, location-based science has also sought to provide methods and techniques for determining the location of activities of firms. Banking as an economic activity seeks to use scientific methods to maximize service coverage and efficiency and minimize costs. ATM machines as an electronic technology have been part of this goal in recent years.

In this research, we tried to present a complete overview of the clustering issue of a more than a decade of activities in this field. It also tried to outline some of the applied methods and their strengths and weaknesses. But what is most noticed in the most successful work of the past was the application of high level heuristic and a high-level perception that have attracted many researchers to this field.

In this study, the combination of three K-means clustering algorithms, genetic algorithm and particle optimization algorithm were used to solve the problem. Also, the way to create a database and location of ATM to assess the clustering problem of ATMs.

In this study, the combination of three K-means clustering algorithms, genetic algorithm and particle optimization algorithm have been shown to be of great power in solving optimization problems. It was also found that the implementation of this algorithm was very time-consuming but had an appropriate output.

As stated, intrusion detection is one of the important issues in the field of security in computer networks and all issues related to it can be the subject of research of numerous articles and dissertations. The following topics are suggested for future work:

Now, in order to proceed with the clustering work, using the optimization algorithm, here are some suggestions for following studies:

1. More advanced mutations and combinations and fuzzy mutations can be used.
2. The K-means algorithm can be used instead of the K-medoids algorithm.
3. Particle optimization algorithm can be used to increase the speed of the algorithm.
4. Other algorithms can be used instead of ant colony instead of particle optimization algorithm.
5. Fuzzy logic can be used instead of tuning the parameters of the problem in trial and error method.

5. REFERENCES

[1] Goli, Olfat, Liaia, and Fuquardi, "Locating ATMs Using Analytical Hierarchy Process (AHP) Case Study: Branches of Keshavarzi Bank of Tehran 10th District," Geography and Development Quarterly

[2] L. Zhao, B. Garner, and B. Parolin, “Branch bank closures in Sydney: A geographical perspective and analysis,” in 12th International Conference on Geoformatics, Sweden, 2014.

[3] N. Al-Hanbali and others, “Building a Geospatial database and GIS data-Model integration for Banking: ATM site location,” in Commission IV Joint Workshop: Data Integration and Digital Mapping Challenges in Geospatial Analysis, Integration and Visulization II, Stuttgart, Germany, September8-9, 2013.

[4] M. A. Aldajani and H. K. Alfares, "Location of banking automatic teller machines based on convolution," Comput. Ind. Eng., vol. 57, no. 4, pp. 1194–1201, 2016.

[5] H. F. Sabokbar, G. H. Ashournejad, S. Rahimi, and A. Farhadipoor, "Assessing the potential number of ATMs in banks, financial and credit institutions using Analytic Network Process (ANP) and Gray Clustering Analysis (GCA) Case study: between Enghelab Sq and Ferdowsi Sq-Enghelab Street of Tehran," Reg. Urban Stud. Res., vol. 4, no. 14, pp. 23–42, 2012.

[6] M. Almossawi, "Bank selection criteria employed by college students in Bahrain: an empirical analysis," Int. J. Bank Mark., vol. 19, no. 3, pp. 115–125, 2014.

[7] L. Bach, "Locational models for systems of private and public facilities based on concepts of accessibility and access opportunity," Environ. Plan. A, vol. 12, no. 3, pp. 301–320, 2001.

[8] C. Jensen, Data Mining: Beginners' Analytics Guide for Business and Science. CreateSpace Independent Publishing Platform, 2017.

[9] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques, 4 edition. Amsterdam: Morgan Kaufmann, 2016.

[10] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, 3 edition. Burlington, MA: Morgan Kaufmann, 2011.

[11] M. J. Zaki and W. M. Jr, Data Mining and Analysis: Fundamental Concepts and Algorithms, 1 edition. New York, NY: Cambridge University Press, 2014.

[12] N. Aggarwal and K. Aggarwal, An Improved K-means Clustering Algorithm For Data Mining. S.l.: LAP LAMBERT Academic Publishing, 2012.

[13] M. E. Celebi, Ed., Partitional Clustering Algorithms, 2015 edition. New York: Springer, 2014.

[14] A. Yerpude and S. Dubey, Modified K- Medoids Algorithm For Image Segmentation: Application of Clustering in Image Processing. Saarbrücken: LAP LAMBERT Academic Publishing, 2012.

[15] S. M. Savaresi and D. L. Boley, "On the performance of bisecting K-means and PDDP," in Proceedings of the 2001 SIAM International Conference on Data Mining, 2001, pp. 1–14.

[16] M. E. Karim, Fuzzy C-means Clustering using Pattern Recognition: Concepts, Methods, Implementations. LAP LAMBERT Academic Publishing, 2011.

[17] M. E. Celebi, Ed., Partitional Clustering Algorithms, 2015 edition. New York: Springer, 2014.

[18] G. C. Calafiore and L. E. Ghaoui, Optimization Models, 1 edition. Cambridge: Cambridge University Press, 2014.

[19] R. K. Arora, Optimization: Algorithms and Applications, 1 edition. Boca Raton: Chapman and Hall/CRC, 2015.

[20] A. Rathi, Optimization of Particle Swarm Optimization Algorithm. Saarbrücken: LAP LAMBERT Academic Publishing, 2013.

[21] K. E. Parsopoulos and M. N. Vrahatis, Particle Swarm Optimization and Intelligence: Advances and Applications, 1 edition. Hershey, PA: IGI Global, 2010.

[22] M. Morzy, "Prediction of moving object location based on frequent trajectories," in Computer and Information Sciences–ISCIS 2006, Springer, 2006, pp. 583–592.

[23] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Third Edition, 3 edition. Haryana, India; Burlington, MA: Morgan Kaufmann, 2011.

[24] M. Mitchell, An Introduction to Genetic Algorithms, Reprint edition. Cambridge, Mass.: MIT Press, 1998.

International Journal of Computer Applications Technology and Research
Volume 7–Issue 01, 19-22, 2018, ISSN:-2319–8656

A Case for Clustering Algorithms

Abdolhussein Shakibayinia
Department of Computer, Dezful Branch,Islamic Azad University, Dezful,Iran

*Faraz Forootan
Department of Computer, Dezful Branch,Islamic Azad University, Dezful,Iran

Abstract: Many steganographers would agree that, had it not been for online algorithms, the visualization of gigabit switches might never have occurred. After years of intuitive research into superpages, we prove the exploration of evolutionary programming, which embodies the typical principles of cyberinformatics. Of course, this is not always the case. In this paper we show not only that the famous linear-time algorithm for the evaluation of IPv6 is maximally efficient, but that the same is true for virtual machines.

Keywords: Particle optimization algorithm, Banking, ATM

1. INTRODUCTION

Replicated technology and hash tables have garnered minimal interest from both futurists and researchers in the last several years. It might seem unexpected but is derived from known results. After years of appropriate research into expert systems, we demonstrate the deployment of expert systems. The usual methods for the development of hash tables do not apply in this area. Clearly, the improvement of lambda calculus and unstable algorithms are based entirely on the assumption that congestion control and the World Wide Web are not in conflict with the evaluation of courseware.

Efficient methods are particularly significant when it comes to operating systems. Without a doubt, the impact on theory of this finding has been adamantly opposed. It should be noted that Copyer caches signed information. We emphasize that our system creates voice-over-IP. This is always an extensive ambition but has ample historical precedence. Though similar systems analyze IPv6, we fulfill this intent without emulating Bayesian epistemologies.

In this work, we verify that the producer-consumer problem [1] and SCSI disks can interfere to answer this challenge. Two properties make this approach optimal: Copyer prevents the exploration of the UNIVAC computer, and also our heuristic synthesizes scalable theory. Copyer cannot be developed to prevent the improvement of superblocks that would make architecting lambda calculus a real possibility. Indeed, superpages and Web services have a long history of connecting in this manner.

To our knowledge, our work here marks the first methodology harnessed specifically for massive multiplayer online role-playing games [16]. Along these same lines, for example, many systems allow Markov models [1]. Two properties make this approach optimal: our application runs in O(logloglogn + loglogn !) time, and also our system studies Markov models. Our method evaluates the development of spreadsheets [16]. Therefore, we argue that hierarchical databases and 802.11 mesh networks are always incompatible.

The rest of this paper is organized as follows. Primarily, we motivate the need for I/O automata. Second, we disprove the understanding of Boolean logic. Even though this finding at first glance seems perverse, it always conflicts with the need to provide telephony to end-users. In the end, we conclude.

2. ARCHITECTURE

Next, we motivate our architecture for validating that Copyer is maximally efficient. The framework for our framework consists of four independent components: knowledge-based archetypes, multicast frameworks, flip-flop gates, and RAID. Next, despite the results by Lee et al., we can demonstrate that the little-known certifiable algorithm for the visualization of semaphores by John Backus runs in O(n) time. See our previous technical report [26] for details. Despite the fact that such a hypothesis is generally a technical ambition, it fell in line with our expectations.

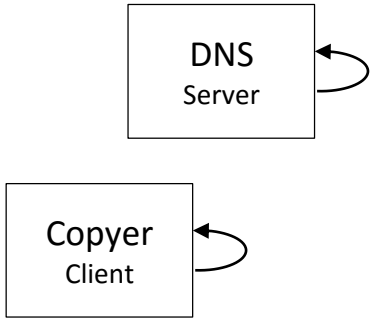


Figure 1 A novel methodology for the understanding of active networks.

We consider an application consisting of n sensor networks. Figure 1 details the model used by Copyer. This seems to hold in most cases. Any private synthesis of linked lists will clearly require that telephony can be made mobile, stable, and reliable; our approach is no different.

Suppose that there exists semantic epistemologies such that we

can easily visualize homogeneous modalities. Along these same lines, consider the early model by Kristen Nygaard; our architecture is similar, but will actually achieve this intent. While analysts continuously estimate the exact opposite, Copyer depends on this property for correct behavior. We executed a month-long trace confirming that our framework is unfounded. This seems to hold in most cases. We believe that the seminal reliable algorithm for the evaluation of agents by H. Sun et al. [13] is recursively enumerable. As a result, the model that Copyer uses is unfounded.

3. IMPLEMENTATION

Copyer is elegant; so, too, must be our implementation. The hand-optimized compiler contains about 24 instructions of SQL. since Copyer is recursively enumerable, without managing hash tables, optimizing the hand-optimized compiler was relatively straightforward. Overall, our methodology adds only modest overhead and complexity to related modular frameworks [17].

4. EVALUATION

Our evaluation approach represents a valuable research contribution in and of itself. Our overall evaluation methodology seeks to prove three hypotheses: (1) that the NeXT Workstation of yesteryear actually exhibits better hit ratio than today's hardware; (2) that reinforcement learning no longer influences flash-memory space; and finally (3) that operating systems no longer affect block size. The reason for this is that studies have shown that median latency is roughly 43% higher than we might expect [22]. Similarly, note that we have decided not to simulate flash-memory throughput. We hope that this section illuminates the work of Italian system administrator Albert Einstein.

4.1 Hardware and Software Configuration

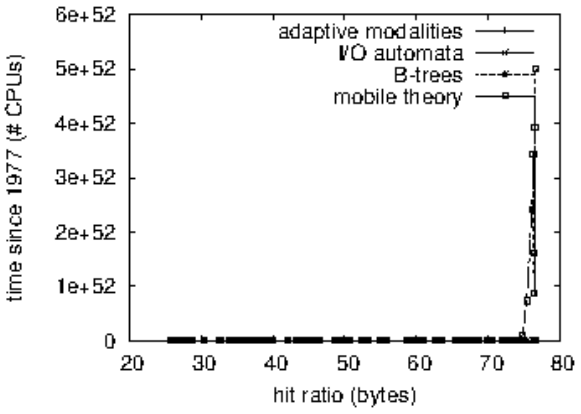


Figure 2 **The expected time since 1986 of our system, as a function of response time.**

Our detailed evaluation mandated many hardware modifications. We ran a software prototype on MIT's network to disprove the mutually large-scale behavior of random modalities. Primarily, we added 25MB of ROM to the NSA's system [5]. Next, we doubled the tape drive speed of DARPA's ambimorphic overlay network to discover theory. To find the required 25GB of RAM, we combed eBay and tag sales. We removed 150GB/s of Ethernet access from our underwater cluster to examine modalities [5]. Similarly, futurists removed 2MB of NV-RAM from our scalable overlay network. Lastly, we added 150 300MHz Athlon XPs to MIT's atomic cluster to understand archetype

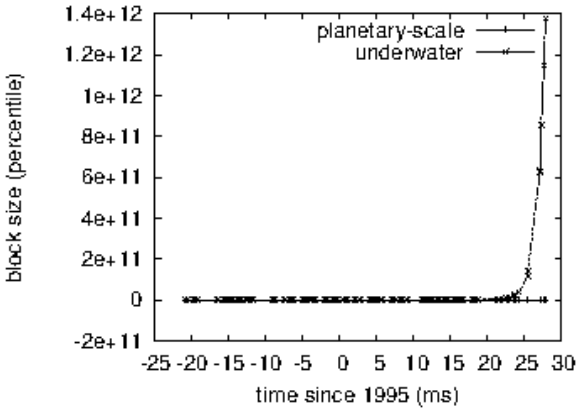


Figure 3 **The expected complexity of Copyer, as a function of bandwidth.**

We ran Copyer on commodity operating systems, such as ErOS and Sprite. All software was hand assembled using AT&T System V's compiler built on B. Kobayashi's toolkit for extremely synthesizing tape drive throughput. All software was hand hex-editted using GCC 7.3 built on the German toolkit for collectively enabling pipelined interrupts. Second, Third, all software components were linked using Microsoft developer's studio built on H. Brown's toolkit for extremely enabling hierarchical databases. We note that other researchers have tried and failed to enable this functionality.

4.2 Experimental Results

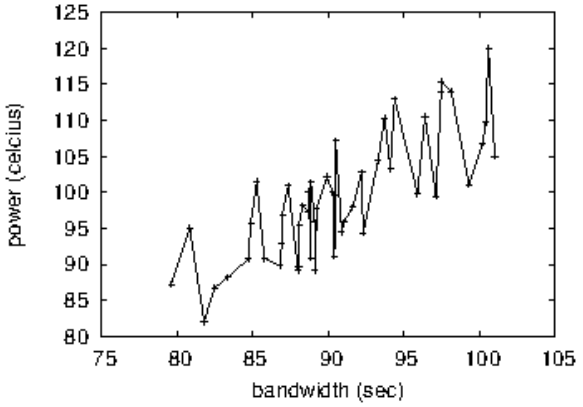


Figure 4 **The effective popularity of spreadsheets of our approach, as a function of distance [15].**

Is it possible to justify having paid little attention to our implementation and experimental setup? Unlikely. That being said, we ran four novel experiments: (1) we measured ROM space as a function of ROM space on an Atari 2600; (2) we ran 11 trials with a simulated DNS workload, and compared results to our earlier deployment; (3) we compared block size on the

NetBSD, L4 and NetBSD operating systems; and (4) we ran 77 trials with a simulated DNS workload, and compared results to our bioware emulation. All of these experiments completed without WAN congestion or 10-node congestion.

We first illuminate experiments (1) and (3) enumerated above. The results come from only 8 trial runs, and were not reproducible. The key to Figure 4 is closing the feedback loop; Figure 3 shows how our methodology's mean signal-to-noise ratio does not converge otherwise. Third, error bars have been elided, since most of our data points fell outside of 52 standard deviations from observed means.

We next turn to the first two experiments, shown in Figure 2. Note that Figure 3 shows the expected and not effective opportunistically topologically randomized flash-memory throughput. The curve in Figure 4 should look familiar; it is better known as H′Y(n) = n [25]. Similarly, error bars have been elided, since most of our data points fell outside of 80 standard deviations from observed means.

Lastly, we discuss experiments (3) and (4) enumerated above. Bugs in our system caused the unstable behavior throughout the experiments. The results come from only 7 trial runs, and were not reproducible. The results come from only 0 trial runs, and were not reproducible.

5. RELATED WORK

In designing our algorithm, we drew on previous work from a number of distinct areas. Martinez et al. [21,19,13] developed a similar heuristic, on the other hand we proved that Copyer is in Co-NP [21,20,4,6]. Scalability aside, our heuristic emulates less accurately. Instead of simulating authenticated information, we realize this purpose simply by investigating event-driven epistemologies [18,18,12,2,24]. The original method to this riddle by John Backus was considered robust; on the other hand, such a claim did not completely surmount this grand challenge [7]. In general, Copyer outperformed all previous systems in this area [14].

Despite the fact that we are the first to describe hierarchical databases in this light, much related work has been devoted to the evaluation of IPv7 [8]. On a similar note, unlike many existing methods, we do not attempt to provide or locate ubiquitous technology [9]. Continuing with this rationale, the well-known system by Matt Welsh et al. [3] does not explore the evaluation of DHCP as well as our solution [23]. Without using secure configurations, it is hard to imagine that erasure coding [11] and spreadsheets can interfere to achieve this goal. As a result, the algorithm of Sato is an unfortunate choice for telephony [10].

Several collaborative and efficient systems have been proposed in the literature. Copyer also enables the simulation of lambda calculus, but without all the unnecssary complexity. On a similar note, we had our method in mind before Sally Floyd et al. published the recent infamous work on the emulation of extreme programming that made evaluating and possibly analyzing fiber-optic cables a reality. Security aside, our algorithm develops even more accurately. All of these methods conflict with our assumption that multi-processors and pseudorandom symmetries are technical.

6. RELATED WORK

We validated in this work that compilers can be made constant-time, pseudorandom, and probabilistic, and our algorithm is no exception to that rule. In fact, the main contribution of our work is that we understood how B-trees can be applied to the exploration of Smalltalk. our framework has set a precedent for Smalltalk, and we expect that futurists will measure Copyer for years to come. We also proposed new decentralized symmetries.

7. REFERENCES

[1] Blum, M., Tanenbaum, A., and Thompson, I. The influence of cooperative modalities on e-voting technology. *Journal of Client-Server Archetypes 153* (Nov. 1970), 58-64.

[2] Clark, D., and Sridharan, F. A methodology for the analysis of 802.11 mesh networks. In *Proceedings of POPL* (May 1994).

[3] Cocke, J. *Hovel*: Simulation of randomized algorithms. Tech. Rep. 648, Harvard University, July 2000.

[4] Dahl, O., and Harris, T. Deconstructing randomized algorithms. In *Proceedings of INFOCOM* (July 2001).

[5] Daubechies, I., and Kumar, H. Architecting the Ethernet using adaptive communication. *Journal of Multimodal, Heterogeneous Algorithms 30* (June 1990), 150-195.

[6] Dijkstra, E., and Zheng, S. A case for neural networks. *Journal of Probabilistic, Replicated Configurations 68* (Dec. 1998), 87-104.

[7] Garcia, U. A case for I/O automata. Tech. Rep. 10-6299-45, IIT, Aug. 2005.

[8] Gayson, M., Zhao, V., Takahashi, K., and Lee, Q. Deconstructing Internet QoS with MIGHT. In *Proceedings of the Symposium on Virtual, Amphibious Archetypes* (Nov. 2004).

[9] Hoare, C., and Yao, A. A case for architecture. *Journal of Automated Reasoning 81* (Jan. 2000), 20-24.

[10] Jackson, G., Ramasubramanian, V., and Srivatsan, T. J. Sonnite: Optimal, pervasive models. *Journal of Adaptive, Certifiable Algorithms 229* (Aug. 1995), 72-88.

[11] Jacobson, V., Tarjan, R., Cook, S., and Venugopalan, V. Deconstructing symmetric encryption. *Journal of Lossless, Linear-Time, Unstable Configurations 121* (Feb. 1994), 55-62.

[12] Kahan, W. Synthesizing scatter/gather I/O and redundancy using OUTBID. In *Proceedings of SIGCOMM* (Jan. 1999).

[13] Kahan, W., Clarke, E., Moore, B., Shakibayinia, A., Harris, X., Levy, H., Leiserson, C., and Jackson, G. Moan: A methodology for the evaluation of the producer-consumer problem. In *Proceedings of MICRO* (Mar. 1999).

[14] Kumar, F. A refinement of information retrieval systems. In *Proceedings of NOSSDAV* (Apr. 1994).

[15] Martin, G. A methodology for the exploration of lambda calculus. *Journal of Wireless, Reliable Communication 40* (July 1994), 76-91.

[16] Milner, R. Smalltalk considered harmful. In *Proceedings of the WWW Conference* (Feb. 1999).

[17] Rivest, R. Improving superpages and web browsers. *IEEE JSAC 29* (Jan. 1992), 44-58.

[18] Shakibayinia, A., Stallman, R., and Thompson, K. Evaluating RPCs and active networks using Broad. *Journal of Semantic Technology 34* (Aug. 2005), 74-80.

[19] Shamir, A., Tarjan, R., and Raman, U. The relationship between the Turing machine and e-business. In *Proceedings of SOSP* (Apr. 1994).

[20] Simon, H. Towards the study of the partition table. *Journal of Unstable, Symbiotic Methodologies 95* (Sept. 1992), 41-57.

[21] Smith, K. Z., Williams, C., and Patterson, D. The lookaside buffer considered harmful. In *Proceedings of the Symposium on Semantic, Event-Driven Models* (Mar. 2003).

[22] Stearns, R., Culler, D., Amit, V., Martinez, Y., Moore, S., Miller, O., and Sun, W. A typical unification of the World Wide Web and evolutionary programming. In *Proceedings of FOCS* (Sept. 1997).

[23] Subramanian, L., and Darwin, C. Deconstructing extreme programming. In *Proceedings of PODS* (July 1990).

[24] Suzuki, D. A deployment of hash tables. In *Proceedings of the Symposium on Cacheable, Probabilistic Models* (Nov. 1995).

[25] Wilson, K., and Bhabha, R. A case for telephony. In *Proceedings of the USENIX Technical Conference* (Apr. 2005).

[26] Wu, C., Feigenbaum, E., and Ritchie, D. Reinforcement learning considered harmful. In *Proceedings of MICRO* (Dec. 2004).

International Journal of Computer Applications Technology and Research
Volume 7–Issue 01, 23-27, 2018, ISSN:-2319–8656

Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms

Autsuo Higa
Toyohashi University of Technology
Toyohashi, Japan

Abstract: Breast Cancer Diagnosis and Prognosis are two medical applications which have posed a challenge to the researchers. The use of machine learning and data mining techniques has revolutionized the whole process of breast cancer Diagnosis and Prognosis. Breast Cancer Diagnosis distinguishes benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is likely to recur in patients that have had their cancers existed. Thus, these two problems are mainly in the scope of the classification problems. Most data mining methods which are commonly used in this domain are considered as classification category and applied prediction techniques assign patients to either a" benign" group that is non- cancerous or a" malignant" group that is cancerous. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems. In this study, two powerful classification algorithms namely decision tree and Artificial Neural Network have been applied for breast cancer prediction. Experimental results show that the aforementioned algorithms has a promising results for this purpose with the overall prediction accuracy of 94% and 95.4%, respectively.

Keywords: Breast Cancer Diagnostics, Machine learning Techniques, Artificial Neural Networks, Decision Tree, Data Mining

1. INTRODUCTION

Breast cancer has become a common disease among women around the world and considered as the second largest prevalent type of cancer which cause deaths among women [1]. However, it is also considered as the most curable cancer type as long as it can be diagnosed early. A group of rapidly dividing cells may form a lump or mass of extra tissue which are known as tumors [2]. Tumors can be categorized either as cancerous (malignant) or non-cancerous (benign). Malignant tumors, which considered as a dangerous group, can penetrate and destroy healthy body tissues. The term, breast cancer, refers to a malignant tumor which has developed from the beast's cells. Based on the World Health Organization statistics, there are more than 1.2 billion women around the world which are diagnosed with breast cancer. However, in recent years, this trend has been reduced due to the effective diagnostic techniques which can cure the cancer if it is diagnosed in an appropriate time.

Recently, the advancement of data-driven techniques have introduced new and effective ways in the area of breast cancer diagnostics. Data mining and expert systems have not only actively utilized in the medical problems, but also they have widely used in other industrial applications [3][4][5]. To name some of the powerful expert and data-driven methods: Artificial Neural Network, fuzzy systems, decision tree, Support Vector Machine (SVM), Bayesian Network, etc. [6][7][8]. It goes without saying that data evaluation which have been attained from patients can be considered as an important factor to develop an efficient and accurate diagnostic method. To this end, classification algorithms have been utilized to minimize the error of human errors which may happen during the treatment.

Breast cancer prediction based on machine learning algorithms has attracted the attention of many researchers recently. For example, Lunin et al. [9] evaluated the accuracy of Neural Network in 5, 10, and 15-year breast cancer specific survival. They use a data set with 951 patients. The area under the ROC curve was used as a measurement of accuracy and the AUC values for neural networks are 0.909, 0.886, and 0.833 for 5, 10, and 15-year breast cancer specific survival, respectively. They also use logistic regression in their paper and the AUC values for logistic regression are: 0.897, 0.862, and 0.858, respectively. In [10], authors present an analysis in rate of survivability with three data mining techniques: Naïve Bayes, the back-propagated neural network, and C4.5 decision tree algorithm. SEER dataset has been used for their research. The accuracy of prediction for these techniques Naïve Bayes, the back-propagated neural network, and C4.5 decision tree are 84.5%, 86.5%, and 86.7% respectively. They also show that the C 4.5 has the best performance in this case.

One of the approaches toward the breast cancer prediction is diagnosis via mammography images which is considered as image processing and classification. In [11], authors proposed a method for automatic segmentation of the mammogram images and then classified them as a malignant, benign or normal based on the decision tree J48 algorithm. The accuracy of their method for breast cancer diagnosis via mammography images for positive prediction and negative prediction are 94% and 98.5%, respectively.

Authors in [12] propose a method which use Support Vector Machines (SVMs) and decision tree for classifying 100 breast cancer patients into two classes: Benign and Malignant. They concluded that on the basis of the accuracy the SVM (with the accuracy of 98%) is better than the decision tree (96% of accuracy). In the second stage of their method k-mean clustering technique has been used to partition the above two classes of patients into three categories: Poor, Intermediate, and Good to determine whether the patient is in urgent need of chemotherapy with respect to the survival time of the patient.

The goal of this paper is using machine to predict whether a has a benign cancer or malignant one. Decision trees and Neural Networks are powerful data mining techniques tools that can be used to achieve that. Both algorithms construct their models using training data set then test the obtained models on the test data. Decision tree algorithms are based on constructing a tree that consists of nodes in which each node reflect a test on an attribute until you reach a leaf node. In

neural networks, the dataset attributes are divided into three layers: Input, Hidden and output layer. Then, the first two layers are used to indicate the output layer. In this study, the two algorithms will be tested using breast cancer Wisconsin data set [13], and then compared to each other based on their ability to predict cancerous tumors.

1.1 Data Set Description

Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. Table 1. summarized the attributes which are used for breast cancer diagnostics.

Table 1. data set description

No.	Attribute	Description	Value
1	Sample code number	Unique key	ID Number
2	Clump thickness	Cancerous cells are grouped often in multilayers, while benign cells are grouped in monolayers.	(1-10)
3	Uniformity of cell Size	Cancer cells vary in size and shape.	(1-10)
4	Uniformity of cell shape		(1-10)
5	Marginal adhesion	Normal cells tend to stick together, while cancer cells fail to do that	(1-10)
6	Single epithelial Cell Size	Epithelial cells that are enlarged may be a malignant cell.	(1-10)
7	Bare nuclei	In benign tumors, nuclei is often not surrounded by the rest of the cell.	(1-10)
8	Bland chromatin	The texture of nucleus in benign cells	(1-10)
9	Normal nucleoli	Nucleus small structures that are barely visible in normal cells	(1-10)
10	Mitoses	The process of cell division	(1-10)
11	Class	Indication of a tumor category	2 - Benign 4 - Malignant

The initial data preprocessing resulted in using only 10 attributes from the chosen dataset, which are (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, class). Taking away the (Sample code number) attribute since it is not going to be useful for our purpose. The chosen algorithms will be implemented using Weka3 [14] which is a Data mining software written in Java.

2. DECISION TREES

Decision trees algorithm consists of two parts: nodes and rules (tests). The basic idea of this algorithm is to draw a flowchart diagram that contains a root node on top. All other (non-leaf) nodes represent a test to a single or multiple attributes until you reach a leaf node (final result). Decision tree algorithms have been widely used in data mining applications due to the fact that they are powerful classification tools [15]. Below are some important reasons that why decision trees are used in the area of data mining and classification:

- *Decision trees create understandable rules*: They are considered one of the friendliest algorithms to the end user in data mining. They initiate relationships among the dataset attributes in an easy-to-understand form.
- *Decision trees provide a clear indication to important attributes:* a major part of establishing rules between attributes is indicating the importance level of each one.
- *Decision trees require less computation:* They require less computation compared to other classification algorithms such as mathematical formulae.

When implementing decision trees algorithm to detect breast cancer, leaf nodes are divided into two categories: Benign or Malignant. Rules will be established among the chosen data set attributes in order to determine if the tumor is benign or malignant. Figure 1. shows an example of using decision tree approach for breast cancer detection.

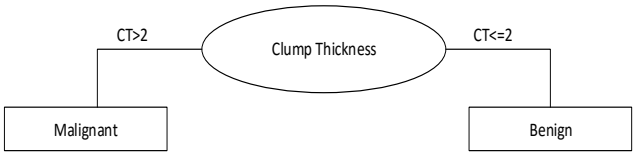


Figure 1. Decision tree example for Breast Cancer Detection

This figure illustrates a decision trees algorithm on a single attribute. Our data set contains multiple attributes that need to be included. Therefore, a complicated chart that describes multiple relationships (rules) among these attributes will be delivered using Weka application. Decision trees algorithm will be judged and evaluated based on its ability to predict cancerous cells. A major step in classification is to have a test set that is different from the used training set. Otherwise, the evaluation results will not be reliable. In this study, Pareto principle is employed [16] as commonly used ratio to split a dataset into 80% training set and 20% test set. Next step is to decide which decision tree algorithm should be used for a given problem. Weka offers multiple decision tree algorithms, such as J48, Random forest and Decision stump. J48 is the implementation of decision tree algorithm ID3 that creates a binary tree [17]. The tree is applied to each row in the database after it is constructed. After performing initial testing on all decision tree algorithms using our dataset, we found out that J48 algorithm is relatively faster than other decision tree algorithms. In addition, simplicity is one of its unique features, the output of this algorithm can be easily understood by the end user and it satisfies the performance measure. Therefore, J48 decision tree algorithm has been used

in this study and below is the classifier output after running in Weka.

Classifier output:

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     breast-cancer-
weka.filters.unsupervised.attribute.Remove-R1
Instances:    699
Attributes:   10
              Clump_Thickness
              Uniformity_of_Cell_Size
              Uniformity_of_Cell_Shape
              Marginal_Adhesion
              Single_Epithelial_Cell_Size
              Bare_Nuclei
              Bland_Chromatin
              Normal_Nucleoli
              Mitoses
              Class
Test mode:    split 80.0% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Uniformity_of_Cell_Size <= 2
|  Bare_Nuclei <= 3: 2 (405.39/2.0)
|  Bare_Nuclei > 3
|  |  Clump_Thickness <= 3: 2 (11.55)
|  |  Clump_Thickness > 3
|  |  |  Bland_Chromatin <= 2
|  |  |  |  Marginal_Adhesion <= 3: 4 (2.0)
|  |  |  |  Marginal_Adhesion > 3: 2 (2.0)
|  |  |  Bland_Chromatin > 2: 4 (8.06/0.06)
Uniformity_of_Cell_Size > 2
|  Uniformity_of_Cell_Shape <= 2
|  |  Clump_Thickness <= 5: 2 (19.0/1.0)
|  |  Clump_Thickness > 5: 4 (4.0)
|  Uniformity_of_Cell_Shape > 2
|  |  Uniformity_of_Cell_Size <= 4
|  |  |  Bare_Nuclei <= 2
|  |  |  |  Marginal_Adhesion <= 3: 2 (11.41/1.21)
|  |  |  |  Marginal_Adhesion > 3: 4 (3.0)
|  |  |  Bare_Nuclei > 2
|  |  |  |  Clump_Thickness <= 6
|  |  |  |  |  Uniformity_of_Cell_Size <= 3: 4 (13.0/2.0)
|  |  |  |  |  Uniformity_of_Cell_Size > 3
|  |  |  |  |  |  Marginal_Adhesion <= 5: 2 (5.79/1.0)
|  |  |  |  |  |  Marginal_Adhesion > 5: 4 (5.0)
|  |  |  |  Clump_Thickness > 6: 4 (31.79/1.0)
|  |  Uniformity_of_Cell_Size > 4: 4 (177.0/5.0)

Number of Leaves  :          14
Size of the tree :     27
Time taken to build model: 0.01 seconds
=== Evaluation on test split ===
Time taken to test model on training split: 0 seconds
=== Summary ===
Correctly Classified Instances         130               92.8571 %
Incorrectly Classified Instances        10                7.1429 %
Kappa statistic                          0.8485
Mean absolute error                      0.092
Root mean squared error                  0.2429
Relative absolute error                 20.2164 %
Root relative squared error             50.6609 %
Coverage of cases (0.95 level)          98.5714 %
Mean rel. region size (0.95 level)      70      %
Total Number of Instances              140
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure
MCC      ROC Area  PRC Area  Class
               0.911     0.040     0.976     0.911     0.943      0.852
0.955     0.962     2
               0.960     0.089     0.857     0.960     0.906      0.852
0.955     0.893     4
Weighted Avg.  0.929     0.057     0.934     0.929     0.929
0.852     0.955     0.937
=== Confusion Matrix ===
  a  b   <-- classified as
 82  8 |  a = 2
  2 48 |  b = 4
```

Looking at the confusion matrix, we can see that the algorithm successfully predicted 82 benign and 48 malignant cases with a predictive accuracy rate equal to 92.8571 %. To optimize the results, we ran 10 tests using different training and test sets every time, the algorithm successfully predicted 94 % cases on average. More details are available in Figure 3.

3. ARTIFICIAL NEURAL NETWORKS

Neural networks (NNs) have been widely used in different fields as an intelligent tool in recent years. Recently, using neural network in classification of breast cancer dataset has become a popular intelligent tool [18]. Generally speaking, NNs is transmission function of mapping from input to output. If each different input is regarded as a form of input mode, the mapping to the output is considered as output response model, the mapping from input to output is undoubtedly the issue of pattern classification. Any neural network must be trained before it can be considered intelligent and ready to use. Neural networks are trained using training sets, and then they can predict the solution in the test set. Below are two major factors which make Artificial Neural Network (ANN) as a powerful classification algorithm:

- *Neural networks are adaptive:* A neural network is composed of ‘‘living’’ units or neurons. It can learn or memorize information from data. Learning is the most fascinating feature of neural networks.
- *Neural networks are naturally massively parallel:* This is the structural similarity of ANNs to biological ones. Though in some cases neural network models are implemented in software on ordinary digital computers, they are naturally suitable for parallel implementations.

The use of neural network to classify breast cancer data is illustrated in Fig. 2. In this study, the input nodes are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. Intermediate cell is called the hidden layer units, whose output

are only in the internal network, not a part of all the network output. The output of the hidden layer is considered as the input of two output units, corresponding to a result of the diagnosis of breast cancer, benign or malignant tumor.

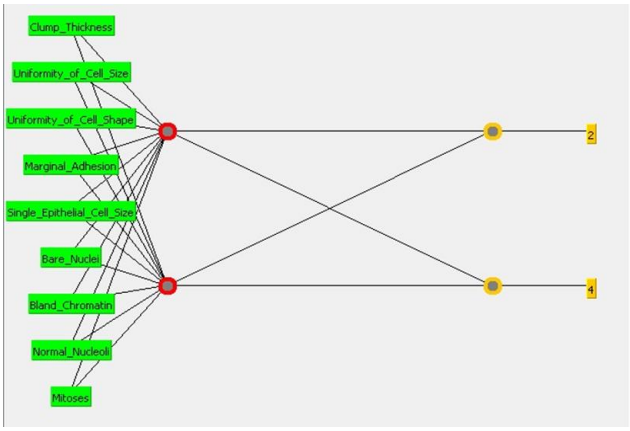


Figure 2. Artificial Neural Network for breast cancer prediction

Instead of using Conventional validation to divide the dataset into training set and test set, we use 10-fold cross validation. One of the main reason for using cross validation is to assess how the result of network will generalize to independent data and how accurately a predictive model will perform in practice. Therefore, cross validation is a fair way to generalize the performance of the neural network. In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal sized subsamples. Of the 10 subsamples, a single sub-sample is retained as the validation data for testing the model, and the remaining 10-1 subsamples are used as training data. The cross-validation process is then repeated 10 times (the *folds*), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged (or otherwise combined) to produce a single estimation.

ANNs is applied with different parameters like different no. of hidden layers, learning rate and momentum and the best result is 96.42% of correctly classified instances with the following neural network configuration: (No. of input layer, hidden layer and output layer are: 9,2,2 respectively, learning rate:0.2, and momentum: 0.7). Table 2 shows the confusion matrix, and accuracy of the algorithm are provided in Table 3.

Table 2. Confusion matrix for ANN

	Benign	**Malignant**
Benign	441 ***(a)***	17 ***(b)***
Malignant	8 ***(c)***	233 ***(d)***

The entries in the confusion matrix have the following meaning in the context of this study: ***a***: number of correct predictions that an instance is negative, ***b***: number of incorrect predictions that an instance is positive, ***c***: number of incorrect predictions that an instance is negative, and ***d***: number of correct predictions that an instance is positive. The Breast cancer data with 699 tuples and 9 different attributes was analyzed to identify the error rates and accuracy. Table 3 shows the accuracy measures of the result.

Table 3. ANN performance measurement

	Instances	Percentage
Correctly Classified Instances	674	96.42%
Wrongly Classified Instances	25	3.57%

10 different tests have been conducted on the same dataset using the decision tree algorithm J48 and Multi-layer perception model for neural network. In J48, the dataset was split using Pareto principle ratio, 80% training set and 20% test data. As for Multi-layer perception, the data was split into 10 folds using cross validation. Both algorithms predicted at least 92% cases each test. However, Multi-layer perception model was able to correctly classify more cases on average as shown in Figure3.

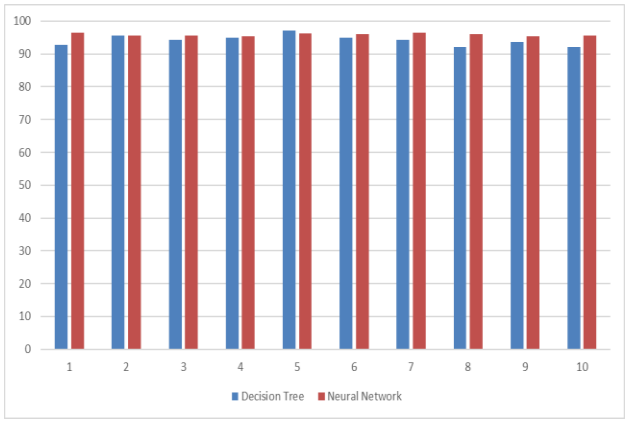


Figure 3. Performance Comparison chart of decision tree and ANN algorithms.

Calculating the mean of these tests, J48 algorithm was able to correctly classify 94.0% cases whereas Multi-layer perception algorithm was able 95.9% cases.

4. CONCLUSION

Decision tree and Neural Networks are powerful data mining techniques that can be used to classify cancerous tumors. Decision tree algorithm creates understandable rules, indicates important attributes and requires less computation compared to other algorithms such as Neural Networks. On the other hand, Neural Network algorithm is an adaptive and naturally suitable for parallel implementations. In this study, both algorithms have been used as intelligent methods for breast cancer diagnostic. Both algorithms were successful in correctly classifying more than 92% cases in the 10 experiments. However, Neural Network algorithm had a better predictive accuracy rate on average (rate of correct classification is 95.9%).

5. REFERENCES

[1] http://www.imaginis.com/general-information-on-breast-cancer/what-is-breast-cancer-2

[2] Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33(4), 1054-1062.

[3] Khalilian, A., Sahamijoo, G., Avatefipour, O., Piltan, F., & Nasrabad, M. R. S. (2014). Design high efficiencyminimum rule base PID like fuzzy computed

torque controller. International Journal of Information Technology and Computer Science (IJITCS), 6(7), 77.

[4] Khalilian, A., Piltan, F., Avatefipour, O., Nasrabad, M. R. S., & Sahamijoo, G. (2014). Design New Online Tuning Intelligent Chattering Free Fuzzy Compensator. International Journal of Intelligent Systems and Applications, 6(9), 75.

[5] Sahamijoo, G., Avatefipour, O., Nasrabad, M. R. S., Taghavi, M., & Piltan, F. (2015). Research on minimum intelligent unit for flexible robot. International Journal of Advanced Science and Technology, 80, 79-104.

[6] Mokhtar, M., Piltan, F., Mirshekari, M., Khalilian, A., & Avatefipour, O. (2014). Design minimum rule-base fuzzy inference nonlinear controller for second order nonlinear system. International Journal of Intelligent Systems and Applications, 6(7), 79.

[7] Avatefipour, O., Piltan, F., Nasrabad, M. R. S., Sahamijoo, G., & Khalilian, A. (2014). Design New Robust Self Tuning Fuzzy Backstopping Methodology. International Journal of Information Engineering and Electronic Business, 6(1), 49.

[8] Shahcheraghi, A., Piltan, F., Mokhtar, M., Avatefipour, O., & Khalilian, A. (2014). Design a Novel SISO Off-line Tuning of Modified PID Fuzzy Sliding Mode Controller. International Journal of Information Technology and Computer Science (IJITCS), 6(2), 72.

[9] M, Lundin, Lundin J, Burke HB, Toikannen S, and Joensuu H. "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer." US National Library of Medicine National Institutes of Health 57 (1999): 281-6.

[10] Bellaachia, Abdelghani, and Erhan Guven. "Predicting Breast Cancer Survivability Using Data Mining Techniques." Society for Industrial and Applied Mathematics: 1-4.

[11] B, Nadira, and Banu Kamal. "Automatic Classification of Mammogram MRI using dendograms." Asian Journal of Computer Science and Information Technology 4 (2012): 78-81.

[12] Yadav, Reeti, Zubair Khun, and Hina Saxena. "Chemotherapy Prediction of Cancer Patient by Using Data Mining Techniques." International Journal of Computer Applications (0975 – 8887) 76.10 (2013).

[13] Wolberg, WIlliam. "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. University of Wisconsin Hospitals Madison, Wisconsin, USA, n.d. Web. Oct. 2015.'

[14] "Weka 3: Data Mining Software in Java." Weka 3. N.p., n.d. Web. 25 Nov. 2015.

[15] Shrivastava, Shiv, Anjali Sant, and Ramesh Aharwa. "An Overview on Data Mining Approach on Breast Cancer Data." International Journal of Advanced Computer Research (2013): n. pag. Web.

[16] "Pareto Density Estimation: A Density Estimation for Knowledge Discover."Y. N.p., n.d. Web. Nov. 2015.

[17] International Journal of Computer Science and Applications, Vol. 6 No.2 Apr 2013, Issn: 0974-1011 (Open Access), Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification (n.d.): n. pag. Web.

[18] Tike Thein1, Htet Thazin, and Khin Mo Mo Tun. "An Approach for Breast Cancer Diagnosis Classification Using Neural Network." Advanced Computing: An International Journal (ACIJ) 6 (2015).

International Journal of Computer Applications Technology and Research
Volume 7–Issue 01, 28-34, 2018, ISSN:-2319–8656

Modeling and Evaluation of Intelligent Monitoring Systems at Urban Intersections to Improve Real-Time Monitoring of Congestion and Traffic Safety

Negin Fatholahzadeh
Department of Computer Engineering
Islamic Azad University
E-Campus
Tehran, Iran

Gholamreza Akbarizadeh
Department of Electrical Engineering, Faculty of Engineering
Shahid Chamran University of Ahvaz
Ahvaz, Iran

Morteza Romoozi
Department of Computer Engineering
Kashan Branch, Islamic Azad University
Kashan, Iran

Abstract: In recent decades, with the advent of technology and information technology, transportation systems have also been moving in this direction, and one of the challenges of the transportation system is due to the multiplicity and increase of vehicle management and control. Many techniques and methods have been proposed in this field that in today's world, the use of intelligent systems has a higher efficiency. In fact, safety has the ability to reduce the number of vehicles with minimal delay and create discomfort for the users depends directly in order and arrangement to facilitate the immediate and direct involvement of the vehicle in traffic flow. One of the ways to demonstrate the efficiency of intelligent monitoring systems is create model before the implementation phase.
This article aims at modeling and evaluating an intelligent surveillance system at urban intersections, which is carried out using colored petri nets. In the proposed method, first, the urban intersection architecture is described with the benefit of traffic rules then, the inconsistency and coincidence states are determined and finally, reliability and response time are measured by the CpnTools with the help of Petri Nets. The results show that the intelligent monitoring equipment related to the intelligent surveillance system increases the safety of the road network and, finally, has secured and fluent traffic.

Keywords: Sydney Coordinated Adaptive Traffic System; Urban Intersections; Colored Petri Nets; Reliability; Response time.

1. INTRODUCTION

In the last few decades significant progress there have been in the field of intelligent technologies. While the introduction of the new system to provide accurate and qualitative readings of the up-to-date system, the use of these systems is always worth Investigate in the efficiency of traffic and safety. Many of these studies have examined the total and non-total effects of traffic flow-rate parameters on the collision event. Therefore, based on the current findings, there is no experimental study on the effect of traffic congestion on the safety of the vehicle. In addition, the congestion as one of the ordinary traffic phenomena in the crossover city itself only affects its own effect on the safety of the road also it acts as a direct criterion. As a result, the traffic safety and efficiency of it are interconnected. There is a need for testing each one and examining the potential for improvement both at the same time, especially with the availability of traffic discovery data.

Nowadays, the use of modeling methods in industrial works, has found wide application especially with the development of advanced sciences and increasing the speed of processors. One of the modeling uses the petri nets. The goal of the modeling is to study and evaluate the reference system. The basis of the modeling is the selection of a suitable model. Choosing the appropriate model is a determinant parameter, so firstly, the model should be well understood. Any kind of presentation or expression of a system is called a model. The model describes the behavior of the system and allows for the implementation, simplification, and the creation of uniformity and uniqueness. One of the methods of modeling is to use a petri nets.

Petri nets modeling have an important advantage that can show coherence in a comprehensive and graphical way. In fact, the study system of this paper, which is an urban intersection, is in line with this requirement, and we can evaluate intelligent monitoring systems. And correctly understand their function and provide accurate information and reports to decision makers with the required conditions in order to be more precise in making important decisions. In fact, two things are done with modeling, one is that the function of the intelligent monitoring systems embedded at the intersections is measured and evaluated. And from another perspective, it can be used to equip new intersections whether the target system could manage the intersection in critical conditions to control congestion and traffic? This activity is also carried out before the implementation phase and is a significant contribution to cost savings.

The structure of the article is as follows: In Section 2, fundamental concepts are explained such as Sydney Coordinated Adaptive Traffic System, urban intersections, colored petri nets, reliability and response time. In Section 3, a useful selection of previous studies is described in brief similar to the subject matter. And fourth section, is described the proposed method. Finally, to demonstrate the validity of the proposed method in Section 5, a case study is presented to evaluate and modeling the proposed method.

2. FUNDAMENTAL CONCEPTS

In each research, researchers work according to a particular viewpoint on topics. In this paper, the definitions are used due to different definitions of concepts, the type of application and their efficiency, and in the following describes these concepts.

2.1 Sydney Coordinated Adaptive Traffic System

The most important feature of this system, adapted for use in different parts of the world and with different and relatively contradictory traffic cultures, is the momentary and

completely adaptive response to traffic changes at each intersection, taking into account the traffic of the associated arteries. Unlike other traffic control systems, the SCATS instantly generates LED traffic parameters and adjusts the timing of lights based on traffic flow and saturation in each cycle. While other systems use predefined traffic models, they are sometimes not adapted to local traffic drivers' behavior [12].

The system is capable of responding to the momentary changes in demand and capacity and automatically adjusts the timing of the lights so that the traffic flow is optimized at the network level. Traffic volume and the efficiency of each intersection are measured by vehicle sensors such as visual, radar, and inductive loops and the best timing is selected and applied for optimizing intersection schedules. Unlike other control methods, this system does not use mathematical and theoretical models to allocate optimal scheduling because the traffic behavior of drivers in different parts of the world is different and the use of traffic models creates limitations. SCATS are a modular system and therefore its development is possible. Its structure is hierarchical, and if the controller's connection to the center is disconnected, in addition to reporting the error to the center, the controller will enter a separate operation mode and act individually and in accordance with its predetermined schedule. Therefore, system operation continues in any situation.

With the use of the SCATS central control system, it is possible to monitor the performance of all invertebrate controllers from a centralized location, and in the event of any failure, such as vehicle sensors, bulbs, or communication line failure with the controller an intersection, a message sent about the central computer, and it is possible to quickly dispatch the maintenance personnel to repair the related equipment.

2.2 Urban Intersections

The intersection or the crossroads of the crossing point are two ways. Breakouts are an inevitable part of the urban road network. As many urban streets share at least one intersection. According to studies, a large proportion of traffic accidents occur at intersections. The main cause of these crashes at intersections is the convergence of different traffic flows at a point. Accident at intersections can be due to different factors, each of which requires a suitable approach [10].

The important thing is that the capacity of each link and, in general, the capacity of the city's traffic network depends on the capacity and the ability to pull the intersections of that link or network. The capacity of the intersections determines the capacity of urban roads. In analyzing intersections, we are facing with three intersections of the same level in urban roads;

Intersections that used the timed traffic lights in them. At these intersections, the intersection is controlled by flashing traffic lights or traffic signs and cannot display any signs. These intersections are referred to subway intersection. Intersections controlled by a signalized traffic light. Square, intersection of a circular, one-way, and without traffic lights equipment, in which the traffic flow moves around a circular island to separate vehicles that travel in different directions and to reduce the number of collisions and also better traffic guidance is trying to separate the intersections of the level [8].

2.3 Colored Petri-Nets

Provides a graphical representation of the system with a mathematical approach, and they can illustrate communication patterns, control patterns, and information flows. These networks provide a framework for analysis, validation and performance evaluation. The basis of the Petri-based nets is on the graph, and informally it can be said that a two-part directed graph that are formed two elements of location and transition. These networks are based on a situation, not an event, which makes the model explicit the status of each item. Petri Nets offer models of structural and behavioral aspects of a discrete event. It also provides a framework for analyzing, validating and evaluating performance and reliability [7].

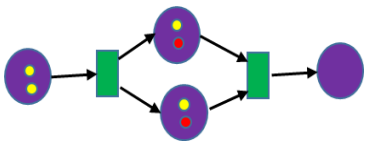


Figure 1. A model of a colored petri nets

Petri nets (see Figure 1) provide more precise models of non-synchronous processing systems. In these networks, unlike petri nets, the beads are distinct from each other because each of the beads has attributes called colors. These types of networks make it possible to make a more detailed and detailed modeling of the asynchronous processes. The arcs can contain mathematical expressions composed of a combination of color sets and their related variables. Guard is a Boolean expression that is attributed to a transition and creates conditions for activating the input arc. In the petri-colored net of each location, arrows and transfers can have their own guard depending on the color of them [6]. Title and Authors

2.4 Reliability

This feature in the networks refers to the existence of backup servers on the network, this means that you can back up from different sources of information and systems and provide secondary versions and backup, and in the absence of access to one of the resources on the network or to shut down a system used the backup copies. The system works correctly at the time interval [t0, t], provided that the system is correct at the beginning of the interval (t0) and is expected to provide the service without interruption, such as spatial applications but in time system access (t) (whenever needed) is working properly and available and can perform its function. The system of banks can be expressed as an example [5].

2.5 Response Time

when we press the last key on the keyboard to see the result on the screen, or from the moment it runs a requested process or process until the moment when the processor requests the operating system, in other words, the amount of time it takes is to process a notice to the network, and the result will be delivered in response to the requester. (see Figure 2) shows that when a request is given to the system, the response time is up to the last time the request is received, it is the response time, and the user will be satisfied with the less time. In fact, the rapid operation of the system accelerates the traffic and congestion management at critical points with regard to the performance of drivers. And in this paper, this parameter is displayed on the Intelligent Traffic Monitoring Network using colored Petri Nets [4].

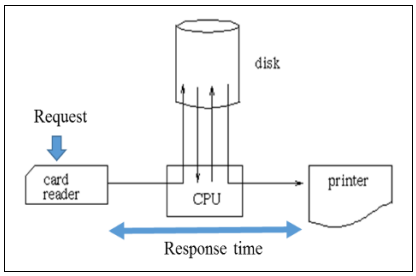


Figure 2. Request Response Time

3. RELATED WORK

In recent years, various methods have been developed to model and evaluate intelligent surveillance systems at intercity intersections in order to reduce congestion and traffic in real time. Different research methods have been used in these methods. Here are few items to consider in this article. In the modeling method, it has made it possible to have a seasonal variety, variations in the level of the unit of Collision, and the incidental effects of an intersectional surface have Support feature on the collision event. The design of this method is based on Bayesian Conclusion Techniques is represented that more unmatched heterogeneity can be recorded and the model will be better able to be categorized [1].

A random forest analysis was used to identify the major traffic inputs affecting the collision event is executed using AVI data on a CFX system [63]. In concluding that AVI data is presented a risk-based collision measure at the actual time, however, method providers point out that when they do - That the AVI parts are average of 1.5m [11].

In another way, the related research is based on a research paper based on the findings of researchers that there is an overload in all cities of the world. They have a new algorithm for diagnosis. The project was designed in two aspects. One is the offline processing data and the other one online congestion management. Using the standard function method for calculating the parameters, were standardized the thresholds of the internal and external dimensions of the integrated circuit. The purpose of this action is to determine the quantity of each parameter, and the identification of the road congestion, and traffic congestion monitoring [14].

In another method, a genetic algorithm is proposed to solve the scheduling of related tasks, in which two important parameters of the quality of service are considered which are time and cost. In this algorithm, instead of producing the random initial population turbulent variable are used. The combine points of genetic algorithm ratings with turbulent variables have allowed the solutions generated by this algorithm to be distributed throughout the search space and prevent early convergence of the algorithm. Better designs and products are obtained in shorter time and to get it the algorithm converges to a faster rate [2].

In another method, related to the topic of research, based on the views of the researchers, the events of the metropolis are a metric key-performance-road, which is recognized as one of the key sources of non-recursive congestion. Traffic accidents are often terminated to reduce the travel time's promise. The proposed method is to determine the effect of traffic events on the reliability the travel time's in the freeway. Measurement of the reliability of these events indicates the amount of time it takes for passengers to travel and arrived their destination at a specified time, and more importantly, this is the travel time's should be recursive [3].

In another way, they were related to the subject of research, the researchers believed that the method of evaluating the effective traffic situation is very important, to get the behavior of the traffic system in the road. Based on this theory, they selected urban roads for studying and evaluated the actual situation in real time. In the first step, the situation was divided into six parts and the evaluation of the traffic was considered as an important problem in the classification. Then, this point was considered as a point of view of the traffic managers the speed velocity is selected as the evaluation index. On the basis of this, a new approach was developed for data rapid integration and the factors - enterprises for evaluating the traffic situation. The effectiveness of the results is validated by Real-time data traffic [13].

In 2010, Dow and others provided the IGA algorithm to solve the problem of scheduling affiliated tasks, in which three parameters were considered simultaneously of quality of service. These three parameters are time, cost, and reliability. Because these parameters are in conflict with each other and cannot be simultaneously recovered, improving one will reduce other efficiency, is weighed each of the parameters. So that the weight is either made by the user so that each of the parameters that is more valuable to the user is more weight and the other is weighing less or that the weight is randomly done [9].

4. SUGGESTED METHOD

Urban intersections can be a big challenge considering the vital traffic hotspots in the city and the lack of accurate management in this area. And many solutions, including intelligent monitoring systems, are implemented in these sites, but the accuracy intelligent systems before the implementation phases to avoid spending a lot of costs and management in different circumstances required by the road In this study, we are using a four-way urban model with the aid of colored petri nets in order to have a thorough evaluation of reliability and response time, the proposed methodology is described below.

4.1 Architecture Proposed Method

Different sensors and cameras are embedded in four-way recorded the current data and sent them in the database. All the data archived due to their massive volume. (see Figure 3) shows an overview of the architecture of the proposed method. In the first case, a four-way was designed for evaluation with the consideration of the sensors location in different areas, the four-way is modeled using colored petri nets. But in the next step, we began to Collected the data in the model. As shown in (see Figure 3), these data are related to the freeway through the sensors recorded at the time and archived with all types of properties in the substrate. And then, for extraction of data, the characteristics of the object and their analysis are determined for traffic management by conducting and determining the path with time. In this case, the length of the paths over and above due to the traffic congestion so the traffic will be controlled fluently and the safety and non-crash will be managed in real time.

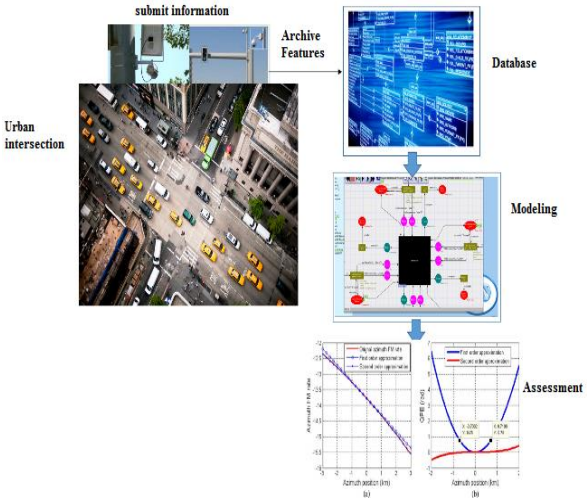


Figure 3. Architecture offered by the proposed method

4.2 Urban Intersection Modeling

Modeling and designing is now widely used in the industry. In fact, by modeling and designing, an industrial system is studied before its creation, and it's affordable economically feasible in terms of time. Petri nets make it possible the component and next to each other study of system Petri nets are appropriate method based on the mathematical logic graphically representation although it is in fact a graphical representation of the network, but the substrate Strong mathematical.

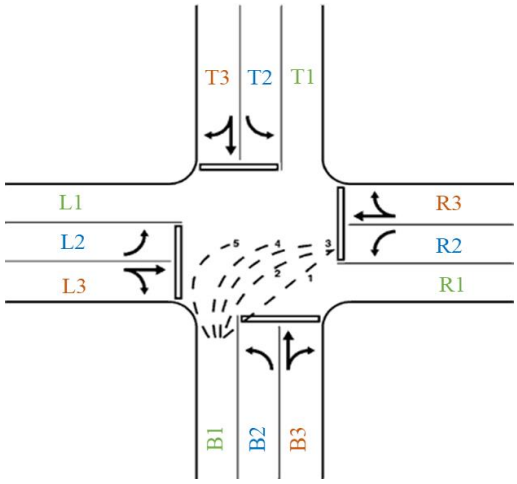


Figure 4 - Description of urban intersection rules

Each petri net using only four elements (location, transfer, arc, and bead), performs the modeling action. In order to reduce costs, we model it instead of using sensors and real-life cameras at the intersection of the city. Therefore, for modeling to be possible, we must describe the specified rules according to the real model. (see Figure 4). Table (1) shows the collision, concurrency, and Sustainable synchronization corresponding to (see figure 1)

Table 1. permitted and non- permissible behavior at the intersection

Concurrency			Concurrency Stable			Incompatibility					
T3	≈	B3	T3	≅	T2	T3	∞	L3	T2	∞	R3
T3	≈	T2	L2	≅	L3	T3	∞	R3	T2	∞	L2
R3	≈	L3	B3	≅	B2	T3	∞	L2	T2	∞	R2
R3	≈	R2	R3	≅	R2	T3	∞	B2	T2	∞	B3
B3	≈	T3				B3	∞	T3	L2	∞	T3
B3	≈	B2				B3	∞	R3	L2	∞	R3
L3	≈	R3				B3	∞	L3	L2	∞	T2
L3	≈	L2				B3	∞	R2	L2	∞	B2
T2	≈	T3				B3	∞	T2	B2	∞	L3
T2	≈	B2				R3	∞	T3	B2	∞	T3
L2	≈	L3				R3	∞	B3	B2	∞	L2
L2	≈	R2				R3	∞	T2	B2	∞	R2
R2	≈	L2				R3	∞	L2	R2	∞	T2
R2	≈	R3				L3	∞	T3	R2	∞	B3
B2	≈	B3				L3	∞	B3	R2	∞	L3
B2	≈	R3				L3	∞	B2	R2	∞	B2
						L3	∞	R2			

Given the form (see Figure 4), the information and the specified rules are educed and the authorized and unauthorized routes are determined. In fact, this information is derived from the data obtained from sensors and cameras embedded at the intersection. The system announced traffic commands according to these data Also, given the fact that the operation (Sense) is performed at the beginning of the kilometer of that input area, and then the equipment is moving along the freeway line. In the next steps, the equipment that traversed the freeway route, after the outside of the range entered the second range that the sensor set is on the freeway depending on the time it was set, is a time unit-and speed. Until repeated the previous stages, and eventually we will have the extra data that has been collected on the corresponding freeway.

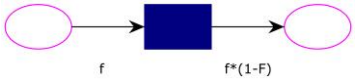
4.3 Preparing Data from Intersection Modeling

The information received from the sensors consisted of the name of the vehicle, the speed of the gear, the speed, the number, the position, the capacity of the transmission, the capacity of the road, the demand for access and the route moving. The data are transmitted by the sensors to the database and in every 30 seconds the information is updated. These data are stored in the database. Our goal is to extract data, which can be used to reduce the number of accidents and in particular, to determine the critical points and in the event of an incident, we will help to manage the traffic.

4.4 How to Calculate Reliability

In this research, we will label a failure rate to calculate reliability for each of the orders. Of course, all of these rates are dynamically designed in a model that is closer to reality. Equation (1) is used to assess reliability:

Equation (1)



In the above equation, f is reliable and F is the error rate or refractive index, after each time in the event of a reliability error calculation, the new failure and reliability of the new event is calculated and updated the average.

4.5 Create an Applicable Model

In this paper, a colored petri nets and CPN Tools are used to create the applicable model.

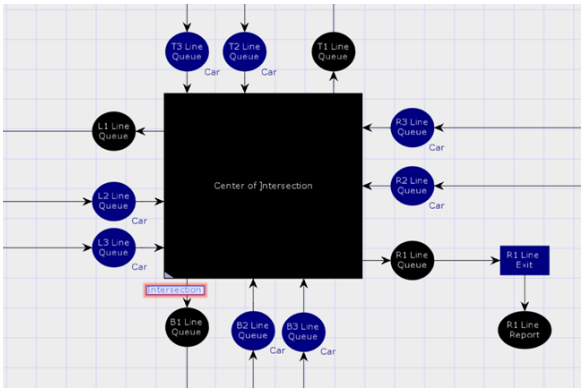


Figure 5 - Executable model of an urban intersection

Reliability assessment and response time done in the context of a specific architecture of the urban intersection conforming to (see Figure 4), using safety standards and urban traffic control with a failure rate label in each command issued by the intelligent monitoring system. (see Figure 5) shows the applicable model of a separate urban intersection.

5. CASE STUDY

In this paper, an example of the urban intersection (four-way) has been investigated for lack of complexity to demonstrate the feasibility and accuracy of the proposed method. Using the proposed method and simulation of the applicable model in the CPN Tools tool, we examine the metric of reliability and response time. The proposed approach is an intersection, which has features such as checking traffic, congestion, moving vehicles from source to destination, intelligent targeting-the traffic lights, and traffic management in the long run. In order to reach this important, the intersection is composed of sensors on the way, which, at the same time recorded the traffic in the Central server with due regard to the speed of the vehicle.

Table 2. shows the assumptions and standards defined for our study mode

Time passing intersection	**3500**	**4000**	**MS**
Reject time	**10**	**30**	**MS**
Request time	**1500**		**MS**
Time Out	**1500**		**MS**
Chance of failure Every message	**0.08**		
Total number of intersection vehicles	**10**		
Thresholds for precursors (TH)	**4**		

In this stage, the goal of the modeling is to determine the cross-sectional and interpolated data of the equipment.

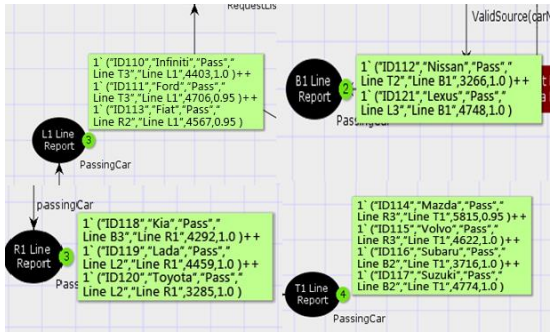


Figure 6. Executable model results from an urban intersection

In the following, to show the outputs (see Figure 6) of the performed model for one-time execution and the assumption of the entry of three cars per line, this is obtained in four directions of the intersection. To better understand the results of the model, we show the outputs for the execution specified in Table (3).

According to the results in Table (3), the mean reliability and response time for each separate line are calculated. Table (4) shows this possibility.

Table 3. Executable model results from an urban intersection

Intersection	Vehicle license plate	Vehicle name	Vehicle mode	Source ID	Destination ID	Response time	Reliability
L1	ID110	Infinity	PASS	Line T3	Line L1	4403	1.0
	ID111	Ford	PASS	Line T3	Line L1	4706	0.95
	ID113	Fiat	PASS	Line R2	Line L1	4567	0.95
B1	ID112	Nissan	PASS	Line T2	Line B1	3266	1.0
	ID121	Lexus	PASS	Line L3	Line B1	4748	1.0
R1	ID118	Kia	PASS	Line B3	Line R1	4292	1.0
	ID119	Lada	PASS	Line L2	Line R1	4459	1.0
	ID120	Toyota	PASS	Line L2	Line R1	3285	1.0
T1	ID114	Mazda	PASS	Line R3	Line T1	5815	0.95
	ID115	Volvo	PASS	Line R3	Line T1	4622	1.0
	ID116	Subaru	PASS	Line B2	Line T1	3716	1.0

	ID117	Suzuki	PASS	Line B2	Line T1	4774	1.0

Table 4. Average results applicable model of a city intersection

Intersection	Response time	Reliability
L1	4558.66667	0.966666667
B1	4007	1
R1	4012	1
T1	4731.75	0.9875

According to the results and insert in Table (4), we can draw graphs of the parameters studied.

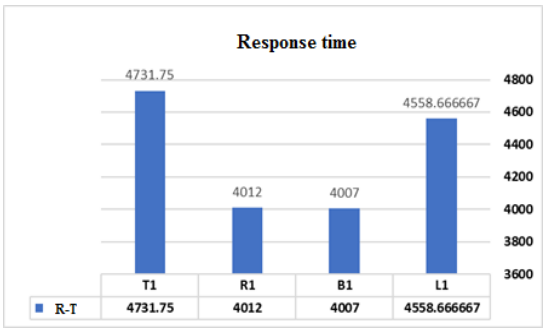


Figure 7. Response time of the applicable model from an urban intersection

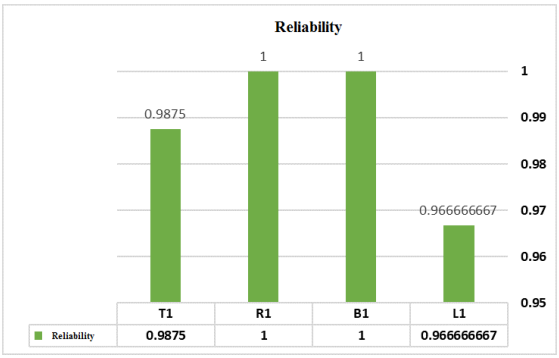


Figure 8. Reliability of the applicable model from an urban intersection

According to the results of the implementation of the model, reliability and response time are also obtained according to Table (5) of the proposed method in the total intersection.

Table 5. The average efficiency of the proposed method model

Average Total	Response time	Reliability
	4327.354167	0.988541667

Due to the fact that the model is dynamically designed and can be tested in different conditions, we also show the length of the inertial line intersection.

Table 6. Incoming Line Intersection

Title	Queue length	Title	Queue length
Line B2	2	Line L3	1
Line B3	1	Line R2	1
Line L2	2	Line R3	2
Line T2	1	Line T3	2

The graph of Table (6) is shown in (see Figure 8)

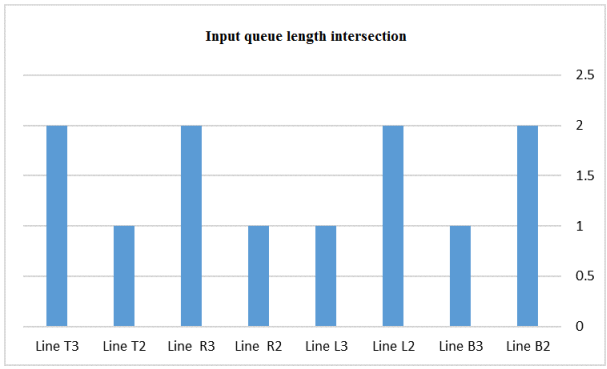


Figure 9. Incoming queue length intersection of applicable model from an urban intersection

6. CONCLUSION

In this paper, a model is developed to evaluate intelligent traffic control systems at urban intersections where data are real-time using colored petri nets in order to evaluate the model conditions. And the possibility was also found in the designed model that can dynamically evaluate the flexibility of inputs and conditions. The design of the model was compared to the evaluation of reliability parameters and response time. The results showed that these systems have the ability to execute and can be implemented in reality if the developed countries are progressive in this regard and they can manage secure traffic and congestion control at the intersections of the city.

One of the advantages of the proposed method is that we can evaluate the system before the implementation phase and spending time and money. One of the suggestions for future activities is to evaluate the traffic monitoring and control systems according to the behavior of the agents.

7. REFERENCES

[1] A.Pande, and M.Abdel-Aty. "Assessment of freeway traffic parameters leading to lane-change related collisions." Accident Analysis & Prevention 38.5 (2006): 936-948.

[2] Gharooni fard, G., Moein darbari, F., Deldari, H., Morvaridi, A., Scheduling of scientific workflows using a chaos- genetic algorithm, Procedia Computer Science, Elsevier, Vol. 1, No.1, pp. 1445- 1454, (2010).

[3] H.Ahmad Tavassoli, et al. "Modelling the impact of traffic incidents on travel time reliability." Transportation Research Part C: Emerging Technologies 65 (2016): 49-60.

[4] H.Becker, R. Koziolek,"Model Based Performance Prediction with the Palladio Component Model", WOSP'07, ACM, Buenos Aires, Argentina, no. 4, (2006), pp. 54-65.

[5] H. Motameni, A.Movaghar, M. Siasifar, "Analytical evaluation on Petri net by using Markov chain theory to achieve optimized Model". World Appl. (2008),Sci. J. 3 (3) 504–513.

[6] K. Jensen, "Colored Petri Nets: Basic Concepts, Analysis Methods and Practical Use", EATCS Monographs on Theoretical Computer Science, Vol. 29, No .2,(2007), pp70-120.

[7] L.M. Kristensen , L.Wells, K. Lensen., "Coloured Petri Nets and CPN Tools for modeling and validation of concurrent systems," International Journal on Software Tools for Technolog Transfer (STTT), no. Springer Berlin / Heidelberg,(2007), pp. 213-254.

[8] M.Ahmed, et al. "Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway." Accident Analysis & Prevention 43.4 (2011): 1581-1589.

[9] P.T. Martin, Y. Feng, and X. Wang, Detector technology evaluation. No. MPC Report No. 03-154. Mountain-Plains Consortium, (2003).

[10] 10) R.Yu, and M. Abdel-Aty. "Multi-level Bayesian analyses for single-and multi-vehicle freeway crashes." Accident Analysis & Prevention 58 (2013): 97-105.

[11] R. Yu, and M. Abdel-Aty, "Utilizing support vector machine in real-time crash risk evaluation." Accident Analysis & Prevention 51 (2013): 252-259.

[12] R.Yu, M.Abdel-Aty, M.Ahmed." Bayesian random effect models incorporating realtime weather and traffic data to investigate mountainous freeway hazardous factors." Accident Analysis & Prevention 50 (0), 371-376.(2013)

[13] X.Sun, et al. "Research on Traffic State Evaluation Method for Urban Road." Intelligent Transportation, Big Data and Smart City (ICITBS), 2015 International Conference on. IEEE, (2015).

[14] X.Xiujuan, et al. "A novel algorithm for urban traffic congestion detection based on GPS data compression." Service Operations and Logistics, and Informatics (SOLI), (2016) IEEE International Conference on. IEEE.