International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 3, May- June 2017 ISSN 2278-6856

Traffic Accidents Analysis with respect to Road Users using Data mining Techniques

Velivela Gopinath[1], K Purna prakash[2] , Challa Yallamandha[3] G Krishna Veni[4] and Dr S Krishna Rao[5]

[1,2,3,4,5]Sir C R Reddy College of Engineering, Department of Information Technology,

Abstract: ***Analysis of road accidents is a very significant because it can expose the association between the diverse types of attributes that contributes to a road accident. Attributes that affect the road accidents can be road attribute, environment attributes, traffic attributes etc. Analyzing road accidents can provide the information about the role of these attributes which can be utilized to defeat the accident rate. There is an enormous range of research work that has already contributed to the road accident analysis. Few studies concentrated on determining the related state of affairs behind the occurrence of road accident as well as few studies concentrated on determining the related circumstances with the brutality of the accident. In these days, Data mining has often utilized a technique for probing the dataset of a road accident. In this study, we applied the classification analysis of data and we achieved some results and later we implemented different clustering techniques such as Self Organizing map (SOM), K-modes and classifications techniques such as Support vector machine (SVM), Naive Bayes (NB) and Decision tree to achieve better accuracy on the basis of victim class and we achieved better results by using SOM instead of using k-modes.***

Keywords: Data mining, SVM, Accident Analysis, clustering and classification techniques

1. INTRODUCTION

Road accidents have been the major reason for untimely death as well as damage to property and economic losses around the world. There are a lot of people die every year in road accidents. Hence, traffic authority devotes substantial endeavor to lessen the road accident but still there is no such reduction in accident rate since in these analyzed years. Road accidents are unpredictable and undetermined occurrence so analysis of road accidents needs the understanding of circumstance which is influencing them. Data Mining [4] has pulled in a lot of consideration in the IT industries as well as in public arena because of the extensive accessibility of vast quantity of data. So, it's necessary to [24] transform these data into applicable knowledge and information. These applicable knowledge and information may be utilized to implement in different areas such as marketing, road accident analysis, fraud detection and so on.Lee C [1] stated that statistical pattern was a better option to determine the connection between traffic, accident, and other geometric circumstances. Data mining [3] is a mutative method which has been utilizing in the area of transportation. Although Barai [2] stated that there is the diverse approach of data mining in the engineering field of transportation such as pavement analysis, road surface analysis and so on. Data mining comprises many techniques such as preprocess, clustering, association, prediction, classification and etc.

Clustering [5] is the errand of categorizing a heterogeneous quantity into various more homogeneous clusters or subgroups. What makes a difference between classification and clustering is that in classification, every record allocated a pre-defined class in according to an [23] enhanced model along with training on the pre-classified examples as well as clustering does not depend on predefined classes. Karlaftis and Tarko [7] utilized analysis to cluster the data and then categorized that dataset of the accident into individual categories and moreover cluster results of analyzed data by utilizing Negative Binomial (NB) to determine the reason of road accident by focusing age of driver which may demonstrate some results. Ma and Kockelman [9] utilized clustering techniques as their initial level to group the dataset into individual division and moreover they utilized Probit model to determine the connection between individual accident features. In this paper, we used Self organizing map (SOM) and k-modes clustering techniques.

Classification comprises of analyzing the characteristics of a recently introduced object and appointing this to one of the predetermined [12] set of classes. The classified objects are to be demonstrated by the record in the table of the file for the database, and the demonstration of classification [22] comprises of including another segment with a class code of some type. To classify the dataset, we used support vector machine (SVM), Naïve bays and J48. Kwon OH [10] utilized decision tree and naive bays classification techniques to analyze aspect dependencies associated with road security.

Young Sohn [17] used a different algorithm to enhance the accuracy of different classifiers for two severity categories
of a traffic accident and each classifier used decision tree and neural network. Tibebe [18] developed a classification model that could assist the traffic officers at Addis Ababa Traffic office for taking the decision to control traffic activities in Ethiopia/Tanzania.

2. ANALYSIS METHODS

This research work focuses on casualty class based or road user-based classification of road accidents. The paper describes the Self Organizing Map (SOM) and K-modes clustering techniques for cluster analysis of the dataset.

Moreover, Support Vector Machine (SVM), Naïve Bays and Decision tree used in this paper to classify the accident data

2.1 Clustering techniques

i) Self Organizing Map (SOM) Clustering

Self-organizing maps (SOMs) is a method for visualizing data and this method is developed by Professor Teuvo Kohonen, the primary objective of this technique is to convert multidimensional data into the lower dimension data or one or two-dimensional data. It is also known as data compression or vector quantization because it reduces the dimension of vectors. The major goal of this study is to entrench different fragments of the neural network to respond similarly to some identified input pattern. When a training set has been imposed to the neural network then their Euclidean distance to final weight vectors is computed. Now the neuron weight is approximately similar to the weight of input. So, this is called by the winner or Best Matching Neuron (BMN). The neuron and weight of BMN which are adjacent in the lattice of SOM are moved towards the input vector. The weight of changes reduces with distance and time from the BMN. The estimated formula for neuron n with having weight vector $W_n(s)$ is given as

$$W_n(s+1) = W_n(s) + \theta(i, n, s) \cdot \alpha(s) \cdot (F(t) - W_n))$$

In this given formula, s is step index, t is an index in training example, i is an index of BMN for F(t), α(s) is decreasing coefficient and input vector is F(t). θ(i, n, s) is the district function which provides the space between neuron i and n in s step. As upon the execution, t may analyze dataset consistently (t=0, 1, 2, 3, 4 ---------T-1 and T is the size of training example).

ii) K-modes Clustering

Clustering is an unsupervised data mining method whose major objective is to categorize the data objects into a distinct type of clusters in such a way that objects inside a group are more alike than the objects in different clusters. K-means algorithm is a very famous clustering technique for large numerical data analysis. In this, the dataset is grouped into k clusters Let**'s** assume that X and Y is a matrix of m by n matrix of categorical data**.** The straightforward closeness coordinating measure amongst X and Y is the quantity of coordinating quality estimations of the two values. The more noteworthy the quantity of matches is more the comparability of two items. K-modes algorithm can be explained as:

$$d(X_i, Y_i) = \Sigma\, \delta(Xi, Yi)_{mi=1} \quad (1)$$

Where $\delta(Xi, Yi) = \{$

$$1, if\ Xi = Yi$$
$$0, if\ Xi \neq Yi \quad (2)$$

2.2 Classification Techniques

i) *Support Vector Machine(SVM)*

SVM is supervised learning method with an analogous algorithm which analyzes data for regression and classification
analysis. SVM work on the basis of decision planes which explain decision boundary. Decision planes are something which differentiates across a set of objects with having different classes. It's a classifier technique that executes classification task by making hyperplanes in n-dimensional space which differentiates the level of classes. SVM assist classification task as well as regression task also and can manage multiple categorical as well as continuous variables.

For the classification type of SVM, minimize the error function: $(V_T V/2) + C\Sigma\, \beta i_{Ni}$

Subjects to the limitations: $Y_i(V_T\theta(X_i)+b) >= 1- \beta_i, \beta_i >= 0, i=1,2,3,\text{------}N$

Here v is vector coefficient, c which is known as capacity constant, β explain the boundary for managing non separable data which is input data and here b is constant. Here i is the index for level T cases of training set, X_i and Y_i describe the class labels and independent variables. α is generally using for transmuting data from the input data to the space feature. If C is greater then more error proscribed so C must be chosen properly.

It's second type to reduce error function for classification type: $(V_T V/2) - v\alpha + 1/N\Sigma\, \beta i_{Ni}$

Subjects to the limitations: $Y_i(V_T\theta(X_i)+b) >= \alpha - \beta_i, \beta_i >= 0, i=1,2,3,\text{------}N$ *and* $\alpha >= 0$ *always*

You need to evaluate the dependent function of the y dependent factor on an arrangement of independent factors x. It accepts as other regression issues that the connection across the independent and dependent factors is provided by a deterministic function which is f in addition to the expansion of some extra noise

Y=f(x) + some noises

For the regression type of SVM: $(V_T V/2) + C\Sigma\, \beta i_{Ni} + C\Sigma\, \beta' i_{Ni}$

These reduce subjects to $V_T\theta(X_i)+b-Y_i =< \varepsilon+\beta'_i$

$Y_i - V_T\theta(X_i)-b =< \varepsilon+\beta_i$

$\beta_l, \beta'_l >= 0, i=1,2,3$ --------N

ii) *Naïve Bays*

This classifier is on the basis on Bayes' hypothesis with autonomy suspicions across indicators. This model is easier to design, with no astonishing iterative measure approximation which makes it primarily precious for large datasets. Despite its smoothness, this classifier often works very well and which is generally utilized on the grounds that it regularly outflanks more complex order techniques. Given a class variable x and a reliant element vector y_1 through y_n, Bayes' hypothesis expresses the accompanying relationship:

$P(x|y_1, \ldots . . y_n) = P(x)\, P(y_1, \ldots\ldots\ldots\ldots y_n|x) / P(y_1 \ldots\ldots\ldots\ldots y_n)$

By using the Naive Bayes assumption that

$P(y_i \mid x, y_1 \ldots \ldots . , y_{i-1}, \ldots \ldots , y_n) = P(y_i|x)$
for all i, this relationship is streamlined to

$P(x|y_1, \ldots . . y_n) = P(x)\, \Pi\, P(y_i|x)\, ni / P(y_1 \ldots\ldots\ldots\ldots y_n)$

Since P (y_1, ------, y_n) is steady given the information, we can utilize the accompanying classification run the show:
$P(x|y_1, \ldots . . y_n) \propto P(x)\, \Pi\, P(y_i|x)$

$\hat{x} = arg\ max = P(x)\, \Pi\, P(y_i|x)$

iii) *Decision Tree*

J48 is an augmentation of ID3. The additional elements of J48 are representing missing data. In the WEKA, J48 is a Java platform open source of the C4.5 calculation. The WEKA gives various alternatives connected with tree pruning. If there should arise an occurrence of possible over fitting pruning maybe utilized as a tool for accuracy. In different calculations, the classification is executed recursively till each and every leaf is clean or pure, that is the order of the data ought to be as impeccable as would be prudent. The goal is dynamically speculation of a choice tree until it picks up the balance of adaptability and exactness. This technique utilized the 'Entropy' that is the computation of disorder data.
Hence so total gain = Entropy ($\vec{X}$) - Entropy ($i|\vec{X}$)
Here the goal is to increase the total gain by dividing total entropy because of diverging arguments $\vec{X}$ by value i.

3 DESCRIPTION OF DATASET

The traffic accident data is obtained from online data source for Leeds UK [8]. This data set comprises 13062 accidents which happened since last 5 years from 2011 to 2015. After carefully analyzed this data, there are 11 attributes discovered for this study. The dataset consist attributes which are Number of vehicles, time, road surface, weather conditions, lightening conditions, casualty class, sex of casualty, age, type of vehicle, day and month and these attributes have different features like casualty class has driver, pedestrian, passenger as well as same with other attributes with having different features which was given in data set. These data are shown briefly in table 2.

4 MEASUREMENT OF ACCURACY

The accuracy is defined by different classifiers of provided dataset and that is achieved a percentage of dataset tuples, which is classified precisely by help of different classifiers. The confusion matrix is also called as error matrix which is just layout table that enables to visualize the behavior of an algorithm. Here confusing matrix provides also an important role to achieve the efficiency of different classifiers. There are two class labels given and each cell consist prediction by a classifier which comes into that cell.

TABLE I: CONFUSION MATRIX

Confusion Matrix		
	Correct Labels	
	Negative	**Positive**
Negative	TN (True negative)	FN (False negative)
Positive	FP (False positive)	TP (True positive)

TPR (Accuracy or True Positive Rate) $= (TN+TP)/All$

$$FPR\ (False\ Positive\ Rate) = \frac{FP}{TN+FP}$$

$$Precision = \frac{TP}{FP+TP}$$

$$Sensitivity = \frac{TP}{FN+TP}$$

5 RESULTS AND DISCUSSION

Table 2 describes all the attributes available in the road accident dataset. There are 11 attributes mentioned and their code, values, total and other factors included. We divided total accident value on the basis of casualty class which is Driver, Passenger, and Pedestrian by the help of SQL.

A) Classification Analysis

We utilized different approaches to classify this bunch of dataset on the basis of casualty class and we used these classifier such as SVM (support vector machine), Naïve bays and Decision tree and we attained some result to few level as shown in table 3.

We achieved some results to this given level by using these three approaches and then later we utilized different clustering techniques which are SOM (Self Organizing map) and K-modes.

TABLE 3

Classifiers	Accuracy
SVM	68.4624 %
Naïve Bays	68.5375%
Decision Tree	70.7566%

TABLE 2

S. NO.	Attribute	Code	Value	Total	Casualty Class		
					Driver	Passenger	Pedestrian
1.	No. of vehicles	1	1 vehicle	3334	763	817	753
		2	2 vehicle	7991	5676	2215	99
		3+	>3 vehicle	5214	1218	510	10
2.	Time	T1	[0-4]	630	269	250	110
		T2	[4-8]	903	698	133	71
		T3	[6-12]	2720	1701	644	374

		T4	[12-16]	3342	1812	1027	502
		T5	[16-20]	3976	2387	990	598
		T6	[20-24]	1496	790	498	207
3.	Road Surface	OTR	Other	106	62	30	13
		DR	Dry	9828	5687	2695	1445
		WT	Wet	3063	1858	803	401
		SNW	Snow	157	101	39	16
		FLD	Flood	17	11	5	0
4.	Lightening Condition	DLGT	Day Light	9020	5422	2348	1249
		NLGT	No Light	1446	858	389	198
		SLGT	Street Light	2598	1377	805	415
5.	Weather Condition	CLR	Clear	11584	6770	3140	1666
		FG	Fog	37	26	7	3
		SNY	Snowy	63	41	15	6
		RNY	Rainy	1276	751	350	174
6.	Casualty Class	DR	Driver				
		PSG	Passenger				
		PDT	Pedestrian				
7.	Sex of Casualty	M	Male	7758	5223	1460	1074
		F	Female	5305	2434	2082	788
8.	Age	Minor	<18 years	1976	454	855	667
		Youth	18-30 years	4267	2646	1158	462
		Adult	30-60 years	4254	3152	742	359
		Senior	>60 years	2567	1405	787	374
9.	Type of Vehicle	BS	Bus	842	52	687	102
		CR	Car	9208	4959	2692	1556
		GDV	Good Vehicle	449	245	86	117
		BCL	Bicycle	1512	1476	11	24
		PTV	PTWW	977	876	48	52
		OTR	Other	79	49	18	11
10.	Day	WKD	Weekday	9884	5980	2499	1404
		WND	Weekend	3179	1677	1043	458
11.	Month	Q1	Jan-March	3017	1731	803	482
		Q2	April-June	3220	1887	907	425
		Q3	July-Septem.	3376	2021	948	406
		Q4	Oct- Dec.	3452	2018	884	549

A) Clustering Analysis

In this analysis, we utilized two clustering techniques which are SOM (Self organizing map) and K-modes techniques. We achieved better results by using SOM as compared to K-modes techniques.

TABLE 4: SVM RESULTS

Rate of error					0.1628			
Predicted values					Confusion Matrix			
Value	Precision	Recall	TPR	FPR		Driver	Passenger	Pedestrian
Driver	0.779	0.909	0.909	0.365	Driver	6958	153	546
Passenger	0.824	0.375	0.375	0.030	Passenger	1828	1330	384
Pedestrian	0.630	0.851	0.851	0.083	Pedestrian	146	132	1584

In this study, we applied Naïve Bays to classify our dataset on the basis of casualty class and this classifier classified dataset into 3 classes. Here again, we can see that our output are determined on the basis of precision, recall, error, error rate, TPR and other various factors which play a really important role. Our accuracy reached to *74.4583%* which is approximately better than earlier without clustering as we achieved *68.5375%.*

TABLE 5: NAÏVE BAYS RESULTS

Rate of error					0.2352			
Predicted values					Confusion Matrix			
Value	Precision	Recall	TPR	FPR		Driver	Passenger	Pedestrian
Driver	0.788	0.868	0.868	0.332	Driver	6649	515	493
Passenger	0.697	0.433	0.433	0.070	Passenger	1624	1535	383
Pedestrian	0.742	0.828	0.828	0.078	Pedestrian	170	151	1541

In this study, we used Decision Tree classifier which improved the accuracy better than earlier which we achieved without clustering. We achieved accuracy *75.7599 %* which is almost more than 5% earlier without clustering.

TABLE 6 : DECISION TREE RESULT

Rate of error					0.1628			
Predicted values					Confusion Matrix			
Value	Precision	Recall	TPR	FPR		Driver	Passenger	Pedestrian
Driver	0.784	0.893	0.893	0.348	Driver	6841	422	394
Passenger	0.724	0.457	0.457	0.065	Passenger	1649	1620	273
Pedestrian	0.683	0.770	0.770	0.060	Pedestrian	231	197	1434

We achieved error rate, precision, TPR (True positive rate), FPR (False positive rate), Precision, recall for every classification techniques as shown in given tables and also achieved different confusion matrix for different classification techniques and we can see the performance of different classifier techniques by the help of confusion matrix.

Here in the next table, we have shown the overall accuracy of analysis with clustering with the help of table 7 and as we can compare this table from the previous table that our accuracy increased in each classification techniques after doing clustering.

TABLE 7

Classifiers	Accuracy
SVM	75.5838 %
Naïve Bays	74.4583 %
Decision Tree	75.7599 %

We have shown accuracy level of table 7 in given figure 2 with the help of chart and we can see from the chart that it's improved after doing clustering in accuracy chart also. Here in the next chart, we used accuracy values of table 3 and table 7 where we can see that accuracy value is improved in table no 7 after doing clustering so we compared these values and shown these value in figure 3.

6. CONCLUSION

In this research work, we analyzed accident dataset by using clustering techniques which are SOM (Self Organizing Map), K-modes as well as classification techniques which are Support Vector Machine (SVM), Naïve Bays and Decision Tree to find pattern on road user specific and we achieved better accuracy by using clustering techniques.

We achieved better accuracy from this way on the basis of casualty class so we can see clearly that what circumstances affect and who is involved more in an accident between the driver, passenger or pedestrian. In this result, SOM provided us better results as compared to K-modes clustering when we classified dataset by using SVM, Naïve Bays, and Decision Tree.

Acknowledgment

I sincerely thank the management and principal of sir c r reddy college of engineering for supporting to lead my research work. I also convey my deep thanks to HOD and faculty of my home department information technology.

REFERENCES

[1] Lee C, Saccomanno F, Hellinga B (2002) Analysis of crash precursors on instrumented freeways. Transp Res Rec. doi: 10.3141/1784-01.

[2] V.Gopinath,DS Bhupal,Y.AkhilaJ Pavani "e-HDAS:e-healthcare diagnosis & advisory system" International journal of research in Engineering & Technology,ISSN NO:2319-1163,Volume 2 issue-11

[3] Barai S (2003) Data mining application in transportation engineering Transport 18:216–223. doi:10.1080/16483840.2003. 10414100.

[4] Kumar S, Toshniwal D (2016) A data mining approach to characterize road accident locations. J Mod Transp 24(1):62–72.

[5] V.Gopinath,ch yallamanda,k purna prakash,dr s Krishna rao, "a reinforce parallel k means clustering using openMp:OM K means",international of scientific research &engineering research,ISSN NO:2229-5518,volume 7, issue

[6] Han, J., and Kamber, M. , "Data mining: concepts and techniques", Academic Press, ISBN 1- 55860- 489-8

[7] Berry, Michael J. A. Data mining techniques: for marketing, sales, and customer relationship management Michael J.A. Berry, Gordon Linoff.— 2nd ed.

[8]Data mining: practical machine learning tools and techniques.—3rd ed. /Ian H. Witten, Frank Eibe, Mark A Hall.

[9]Karlaftis M, Tarko A. Heterogeneity considerations in accident modeling. Accid Anal Prev. 1998;30(4):425–33.

[10] Data source: https://data.gov.uk/dataset/road-traffic-accidents accessed on 24 October 2016

[11] Ma J, Kockelman K. Crash frequency, and severity modeling using clustered data from Washington state.In: IEEE Intelligent Transportation Systems Conference. Toronto Canadá; 2006.

[12]V.Gopinath, dr s Krishna rao ,ch yallamanda,k purna prakash "a journey of big data:3V's to 32 V's"international journel of research in computer technology,ISSN No:2278-5841,volume 5 ,issue 3

[13] Kwon OH, Rhee W, Yoon Y (2015) Application of classification algorithms for analysis of road safety risk factor dependencies. Accid Anal Prev 75:1–15. doi:10.1016/j.aap.2014.11.005

[14] Geurts K, Wets G, Brijs T, Vanhoof K (2003) Profiling of high-frequency accident locations by use of association rules. Transp Res Rec. doi:10.3141/1840-14

[15] Marc M. Van Hulle Laboratorium Voor Neurofysiologie K.U. Leuven, Belgium marc@neuro.kuleuven.be

[16] Kohonen T (1995) Self-organizing maps, 2nd edn.Springer, Heidelberg

[17] Kumar and Toshniwal, A data mining framework to analyze road accident data, Journal of Big Data (2015) 2:26

[18] Kumar and Toshniwal, Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC), Journal of Big Data (2016) 3:13

[19] Sachin Kumar, Durga Toshniwal, Manoranjan Parida, A comparative analysis of heterogeneity in road accident data using data mining techniques,Evolving Systems

[20] So Young Sohn, Sung Ho Lee, Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea, Safety Science.

[21] Tibebe Beshah Tesema, Ajith Abraham, Crina Grosan, Rule mining and classification of road traffic accidents using adaptive regression trees , I. J. of simulation vol. 6 no 10 and 11

[22] V.Gopinath, DS.Bhupal naik "A Novel Approach towards parallel K-means" International Journal of computer engineering & science, volume-3 issue-2, ISSN NO:2231-6590.

[23] V.Gopinath, dr s Krishna rao ,ch yallamanda,k purna prakash "cloud sharing for economic

benifits"international journel of computer science and information securityISSN No:1947- 5500,volume 14

[24] E.Shailini ,V.Gopinath "Multimedia sensor networks with clustered hierarchical data de-duplication",International journal of current research and academic review",ISSN NO:2347- 3215.