Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques

Prayag Tiwari[1(✉)], Sachin Kumar[2], and Denis Kalitin[1]

[1] Department of Computer Science and Engineering,
National University of Science and Technology Misis, Moscow, Russia
prayagforms@gmail.com, kalitindv@gmail.com
[2] Centre for Transportation Systems, Indian Institute of Technology Roorkee,
Roorkee, India
sachinagnihotril6@gmail.com

Abstract. Analysis of road accident is very important because it can expose the relationship between the different types of attributes that contributes to a road accident. Attributes that affect the road accident can be road attribute, environment attributes, traffic attributes etc. Analyzing road accident can provide the information about the contribution of these attributes which can be utilized to overcome the accident rate. Nowadays, Data mining is a popular technique for examining the road accident dataset. In this study, we have performed the classification of road accident on the basis of road user category. We have used Self Organizing map (SOM), K-modes clustering technique to group the data into homogeneous segments and then applied Support vector machine (SVM), Naive Bayes (NB) and Decision tree to classify the data. We have performed classification on data with and without clustering. The result illustrates that better classification accuracy can be achieved after segmentation of data using clustering.

Keywords: Data mining · Accident analysis · Clustering · Classification

1 Introduction

Road accident 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 a traffic or road accident. 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 accident is unpredictable and undetermined. Hence, analysis of traffic accident requires the understanding of circumstance which is influencing them. Data Mining [4, 6, 19, 20, 26–30] 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 transform these data into applicable knowledge and information. These applicable knowledge and information may be utilized to

J.K. Mandal et al. (Eds.): CICBA 2017, Part II, CCIS 776, pp. 398–410, 2017.
DOI: 10.1007/978-981-10-6430-2_31

implement in different areas such as marketing, road accident analysis [11, 12, 15], 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, 23] 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 contrast between clustering and classification is that in classification, every record allocated a pre-defined class in according to an 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 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 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 safety. 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 neural network and decision tree. 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. S. Kuznetsov et al. [21, 22, 24, 25] used an algorithm based on FCA for numerical data mining and provided more efficient results. The organization of the paper is as follows: Sect. 2 will describe the data set used and methodology adopted in the study. Section 3 will present the results and discuss the findings. Finally, Sect. 4 will conclude with a future scope.

2 Materials and Methods

This research work focuses on casualty class based or road user-based classification of a road accident. 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 are used in this paper to classify the accident data.

2.1 Clustering Techniques

Self Organizing Map (SOM)
Self-organizing maps (SOMs) [13] is a method for visualizing data and this method is given 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 vector quantization or data compression 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 networks then their Euclidean distance to final weight vectors are 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

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

In this given formula, s is step index, t is an index in training example, i is an index of BMN for F(t), $\alpha(s)$ is decreasing coefficient and input vector is F(t). $\theta(i, n, s)$ 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).

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 [3] algorithm is a very famous clustering technique for large numerical data analysis. In this, the dataset is grouped into k clusters. There are diverse clustering algorithms available but the assortment of appropriate clustering algorithm relies on type and nature of data. Our major objective of this work is to differentiate the accident location on their frequency occurrence. 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 [14, 16, 20] algorithm can be explained as:

$$d(X_i, Y_i) = \sum\nolimits_{i=1}^{m} \delta(Xi, Yi) \tag{2}$$

$$Where\ \delta(Xi, Yi) = \begin{cases} 1,\ if\ Xi = Yi \\ 0,\ if\ Xi \neq Yi \end{cases} \tag{3}$$

2.2 Classification Techniques

Support Vector Machine

SVM is supervised learning method with an analogous algorithm which analyzes data for regression and classification analysis. SVM works on the basis of decision planes which explain decision boundary. Decision planes are something which differentiates across a set of objects with having distinct classes. It's a classifier technique that executes classification task by making hyper planes 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: $(V^T V/2) + C\sum_{i=1}^{n} \beta_i$

$$\textit{Subjects to the limitations:} Y_i\left(V^T\theta(X_i)+b\right) > = 1-\beta_i, \beta_i > = 0,\ i = 1,2,3,\ldots N \tag{4}$$

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 than more error proscribed so C must be chosen properly.

It's the second type to reduce error function for classification type: $(V^T V/2) + v\alpha + \frac{1}{N}\sum_{i=1}^{N} \beta i$

$$\textit{Subjects to the limitations:} Y_i\left(V^T\theta(X_i)+b\right) > = \alpha-\beta_i, \beta_i > = 0,\ i = 1,2,3,\ldots N \textit{ and } \alpha > = 0 \textit{ always} \tag{5}$$

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

$Y = f(x) + \text{some noises}$

$$\textit{For the regression type of SVM}: \left(V^T V/2\right) + C\sum_{i=1}^{n} \beta i + C\sum_{i=1}^{n} \beta' i \tag{6}$$

These reduce subjects to $V^T\theta(X_i) - bY_i = <\varepsilon+\beta_i'$

$$Y_i - V^T\theta(X_i) - b = <\varepsilon+\beta_i$$
$$\beta_I\beta_I' > = 0,\ i = 1,2,3,\ldots N$$

It's the second type to reduce error function for classification type:

$$\left(V^T V/2\right) - C\left(v\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\beta i + \beta' i)\right) \tag{7}$$

$$(V^T\theta(X_i)+b)-Y_i = <\varepsilon+\beta_i$$
$$Y_i-(V^T\theta(X_i)+b) = <\varepsilon+\beta_i'$$

$\beta_I\beta_I' > = 0,\ i = 1,2,3,\ \ldots N, \varepsilon > 0$

Naïve Bayes

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 element vector y_1 through y_n, Bayes' hypothesis expresses the accompanying relationship:

$$P(x|y_1,\ldots\ldots.y_n) = \frac{P(x)P(y_1,\ldots\ldots\ldots\ldots.y_n|x)}{P(y_1\ldots\ldots\ldots\ldots.y_n)} \tag{8}$$

By using the Naive Bayes assumption that

$$\mathrm{P}\,(y_i|x,y_1\ldots\ldots,y_{i-1},\ldots\ldots,y_n) = \mathrm{P}\ (y_i|x) \tag{9}$$

for all i, this relationship is streamlined to

$$P(x|y_1,\ldots\ldots.y_n) = \frac{P(x)\prod_{i=1}^{n}\mathrm{P}(y_i|x)}{P(y_1\ldots\ldots\ldots\ldots.y_n)} \tag{10}$$

Since $P(y_1, \ldots y_n)$ is steady given the information, we can utilize the accompanying classification run the show:

$$\begin{aligned} P(x|y_1,\ldots...y_n) &\propto P(x)\prod_{i=1}^{n}\mathrm{P}(y_i|x) \\ x^{\wedge} &= arg\ max = P(x)\prod_{i=1}^{n}\mathrm{P}(y_i|x) \end{aligned} \tag{11}$$

What's more, we can utilize Maximum A Posteriori (MAP) estimation to gauge P (x) and P (y_i|x); the previous is then the relative recurrence of class x in the preparation set. Regardless of their clearly over-improved suppositions, credulous Bayes classifiers have worked great in some genuine circumstances, broadly record classification and spam separating. They require a little measure of preparing information to gauge the fundamental parameters.

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 overfitting pruning may be 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. Here Entropy $\vec{X}$ is measured by:

$$\text{Entropy } (\vec{X}) = -\sum_{i=1}^{n} \frac{|Xi|}{\overrightarrow{|X|}} \log\left(\frac{|Xi|}{\overrightarrow{|X|}}\right) \tag{12}$$

$$\text{Entropy } (i|\vec{X}) = \frac{|Xi|}{\overrightarrow{|X|}} \log\left(\frac{|Xi|}{\overrightarrow{|X|}}\right) \tag{13}$$

Hence,

$$\text{Total Gain} = \text{Entropy } (\vec{X}) - \text{Entropy } (i|\vec{X}) \tag{14}$$

Here the goal is to increase the total gain by dividing total entropy because of diverging arguments $\vec{X}$ by value i.

2.3 Description of Data Set

The traffic accident data is obtained from the online data source for Leeds UK [8]. This data set comprises 13062 accident that occurred during 2011 to 2015. Initial preprocessing of the data results in 11 attributes that found to be suitable for further analysis. The attributes selected for analysis are a number of vehicles, time of the accident, road surface, weather conditions, lightening conditions, casualty class, sex of casualty, age, type of vehicle, day and month of the accident. The accident data is illustrated in Table 1.

Table 1. Road accident attribute description

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

(*continued*)

Table 1. (*continued*)

S. No.	Attribute	Code	Value	Total	Casualty class		
					Driver	Passenger	Pedestrian
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		7657	0	0
		PSG	Passenger		0	3542	0
		PDT	Pedestrian		0	0	1862
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	Goods 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	Jul-Sep	3376	2021	948	406
		Q4	Oct-Dec	3452	2018	884	549

2.4 Measurement of Accuracy

The classification accuracy is one of the important measures of how correctly a classifier classifies a record to its class value? The confusion matrix is an important data structure that helps in calculating different performance measures such as precision, accuracy, recall and sensitivity of classification technique on some data.

Table 2. Confusion matrix sample

	Negative	Positive
Negative	TN (True negative)	FN (False negative)
Positive	FP (False positive)	TP (True positive)

Table 2 provides a sample confusion matrix table and Eqs. 1–4 illustrates the formulas to calculate different performance measures.

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \tag{15}$$

$$False\ Positive\ Rate = \frac{FP}{TN + FP} \tag{16}$$

$$Precision = \frac{TP}{FP + TP} \tag{17}$$

$$Sensitivity = \frac{TP}{FN + TP} \tag{18}$$

3 Results and Discussion

In this section, the experimental analysis and the obtained results will be discussed.

3.1 Classification Analysis

We utilized different approaches to classifying this bunch of dataset on the basis of casualty class using SVM (support vector machine), Naïve bays and Decision tree. The classification accuracy achieved is shown in Fig. 1. It can be seen that decision tree obtained the highest accuracy of 70.7% in comparison to other two classifiers.

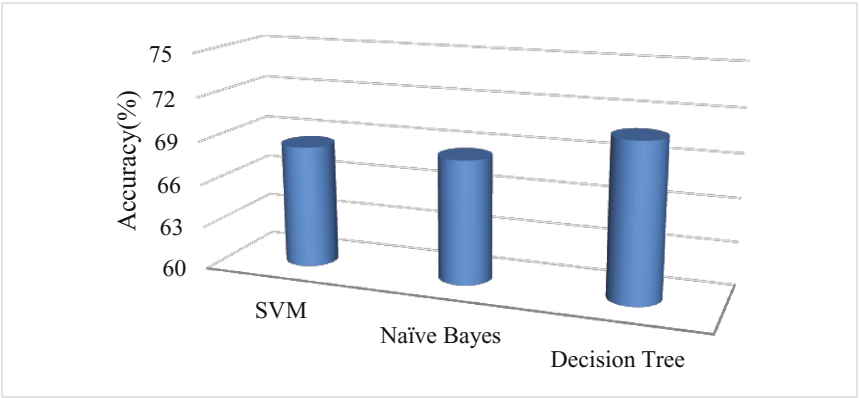

Fig. 1. Classification accuracy of different classifiers on accident data

3.2 Classification Followed by Clustering of Accident

In this analysis, we utilized two clustering techniques which are SOM (Self organizing map) and K-modes techniques. We achieved better results by using k-modes as compared to SOM technique and therefore we are describing the performance of classifiers on clusters obtained by k-modes only.

Performance evaluation of SVM

In this study, we applied SVM to classify dataset on the basis of casualty class and this classifier classified data into 3 classes. The output of this classifier are determined on the basis of their precision, recall, error rate and other factors and we achieved accuracy which is *75.5838%* and it's increased approximately 7% and it is better than earlier when we analyzed our dataset without clustering. Table 3 provides the performance of SVM on clusters obtained from k-modes.

Table 3. Performance of SVM

Rate of error = 0.1628								
Predicted values					Confusion matrix			
Class	Precision	Recall	TPR	FPR	Class	DR	PSG	PDT
DR	0.779	0.909	0.90	0.36	DR	6958	153	546
PSG	0.824	0.375	0.37	0.03	PSG	1828	1330	384
PDT	0.630	0.851	0.85	0.083	PDT	146	132	1584

Performance evaluation of Naïve Bayes

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 *76.4583%* which is approximately better than earlier without clustering as we achieved *68.5375%*. Table 4 provides the performance of Naïve Bayes on clusters obtained from k-modes.

Table 4. Performance of naive bayes

Rate of error = 0.2352								
Predicted values					Confusion matrix			
Class	Precision	Recall	TPR	FPR	Class	DR	PSG	PDT
DR	0.788	0.86	0.86	0.33	DR	6649	515	493
PSG	0.697	0.43	0.43	0.07	PSG	1624	1535	383
PDT	0.742	0.828	0.828	0.078	PDT	170	151	1541

Performance evaluation of Decision Tree

In this study, we used Decision Tree classifier which improved the accuracy better than earlier which we achieved without clustering. We achieved accuracy *81%* which is better than earlier. Table 5 provides the performance of decision tree on clusters obtained from k-modes.

Table 5. Performance of decision tree

Rate of error = 0.1628								
Predicted values					Confusion matrix			
Class	Precision	Recall	TPR	FPR	Class	DR	PSG	PDT
DR	0.784	0.893	0.893	0.348	DR	6841	422	394
PSG	0.724	0.457	0.457	0.065	PSG	1649	1620	273
PDT	0.683	0.770	0.770	0.060	PDT	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 Tables 3, 4 and 5 and as we can observe from these tables classification accuracy increased for each classification technique after doing clustering.

Figure 2 illustrates the classification accuracy of SVM, Naïve Bayes and decision tree on clusters obtained from k-modes and SOM. It can be seen that classification accuracy is better for clusters obtained from k-modes clustering rather than obtained from SOM. It can be concluded that k-modes clustering technique provides better clustering than SOM on data with categorical road accident attributes.

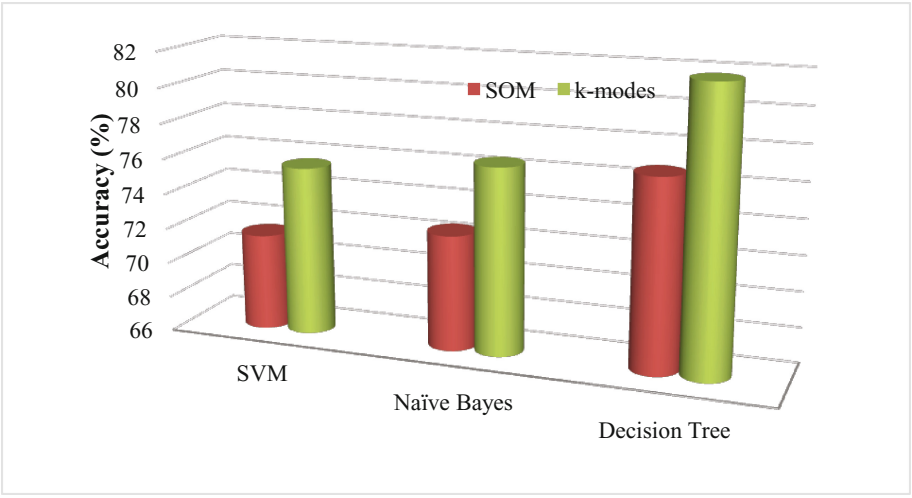

Fig. 2. Classification accuracy on clusters obtained from k-modes and SOM

4 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, k-modes provided us better clustering results as compared to SOM and classification accuracy of SVM, Naïve Bays, and Decision Tree is found better on clusters obtained from k-modes. It indicates that clustering certainly improves the classification accuracy of classifiers and k-modes clustering would be a better option to cluster road accident data with categorical attributes.

References

1. Lee, C., Saccomanno, F., Hellinga, B.: Analysis of crash precursors on instrumented freeways. Transp. Res. Rec. (2002). doi:10.3141/1784-01
2. Barai, S.: Data mining application in transportation engineering. Transport **18**, 216–223 (2003). doi:10.1080/16483840.2003.10414100
3. Kumar, S., Toshniwal, D.: A data mining approach to characterize road accident locations. J. Mod. Transp. **24**(1), 62–72 (2016)
4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Academic Press, San Francisco (2000). ISBN 1-55860-489-8

5. Berry, M.J.A., Linoff, G.S.: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd edn. Wiley, New York (1997)
6. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, San Francisco (2005)
7. Karlaftis, M., Tarko, A.: Heterogeneity considerations in accident modeling. Accid. Anal. Prev. **30**(4), 425–433 (1998)
8. Data source: https://data.gov.uk/dataset/road-traffic-accidents. Accessed 24 Oct 2016
9. Ma, J., Kockelman, K.: Crash frequency, and severity modeling using clustered data from Washington state. In: 2006 IEEE Intelligent Transportation Systems Conference, Toronto, Canadá (2006)
10. Kwon, O.H., Rhee, W., Yoon, Y.: Application of classification algorithms for analysis of road safety risk factor dependencies. Accid. Anal. Prev. **75**, 1–15 (2015). doi:10.1016/j.aap.2014.11.005
11. Geurts, K., Wets, G., Brijs, T., Vanhoof, K.: Profiling of high-frequency accident locations by use of association rules. Transp. Res. Rec. (2003). doi:10.3141/1840-14
12. Kumar, S., Toshniwal, D.: A novel framework to analyze road accident time series data. J. Big Data **3**(8), 1–11 (2016). Springer
13. Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer, Heidelberg (1995)
14. Kumar, S., Toshniwal, D.: A data mining framework to analyze road accident data. J. Big Data **2**(26), 1–18 (2015)
15. Kumar, S., Toshniwal, D.: Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). J. Big Data **3**(13), 1–11 (2016)
16. Kumar, S., Toshniwal, D., Parida, M.: A comparative analysis of heterogeneity in road accident data using data mining techniques. Evol. Syst. **8**(2), 147–155 (2016)
17. Sohn, S.Y., Lee, S.H.: Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. Saf. Sci. **41**(1), 1–14 (2003)
18. Tibebe, B.T., Abraham, A., Grosan, C.: Rule mining and classification of road traffic accidents using adaptive regression trees. Int. J. Simul. **6**, 10 (2005)
19. Tiwari, P., Mishra, B.K., Kumar, S., Kumar, V.: Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. Int. J. Knowl. Discov. Bioinform. (IJKDB) **7**(1), 30–41 (2017). doi:10.4018/IJKDB.2017010103
20. Kumar, S., Toshniwal, D.: Analyzing road accident data using association rule mining. In: ICCCS-2015, Mauritius. IEEE-Xplore (2015). doi:10.1109/CCCS.2015.7374211
21. Kaytoue, M., Kuznetsov, S.O., Napoli, A., Duplessis, S.: Mining gene expression data with pattern structures in formal concept analysis. Inf. Sci. Int. J. **181**(10), 1989–2001 (2011). Information Science, Special Issue on Information Engineering Applications Based on Lattices. Elsevier, New York (2011)
22. Tiwari, P., Dao, H., Nguyen, G.N., Kumar, S.: Performance evaluation of lazy, decision tree classifier and multilayer perceptron on traffic accident analysis. Informatica **41**(1), 39 (2017)
23. Poelmans, J., Kuznetsov, S.O., Ignatov, D.I., Dedene, G.: Formal Concept Analysis in knowledge processing: A survey on models and techniques. Expert Syst. Appl. **40**(16), 6601–6623 (2013)
24. Tiwari, P.: Comparative analysis of big data. Int. J. Comput. Appl. **140**(7), 24–29 (2016). Foundations of Computer Science (FCS)
25. Kuznetsov, Sergei O.: Fitting pattern structures to knowledge discovery in big data. In: Cellier, P., Distel, F., Ganter, B. (eds.) ICFCA 2013. LNCS, vol. 7880, pp. 254–266. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38317-5_17
26. Kuznetsov, S.O., Poelmans, J.: Knowledge representation and processing with formal concept analysis. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **3**(3), 200–215 (2013)

27. Sachin, K., Semwal, V.B., Tiwari, P., Solanki, V., Denis, K: A conjoint analysis of road accident data using K-modes clustering and bayesian networks. Ann. Comput. Sci. Inf. Syst. 10, 53–56 (2017)
28. Tiwari, P.: Improvement of ETL through integration of query cache and scripting method. In: 2016 International Conference on Data Science and Engineering (ICDSE). IEEE (2016)
29. Tiwari, P.: Advanced ETL (AETL) by integration of PERL and scripting method. In: 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, pp. 1–5 (2016). doi:10.1109/INVENTIVE.2016.7830102
30. Tiwari, P., Kumar, S., Mishra, A.C., Kumar, V., Terfa, B.: Improved performance of data warehouse. In: International Conference on Inventive Communication and Computational Technologies (ICICCT-2017), Coimbatore, 10–11 March 2017. Proceeding will be published IEEE-Xplore