Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering pp. 53–56
DOI: 10.15439/2017R44
ACSIS, Vol. 10 ISSN 2300-5963

# A Conjoint Analysis of Road Accident Data using K-modes Clustering and Bayesian Networks (Road Accident Analysis using clustering and classification)

[1]Sachin Kumar, [2]Vijay Bhasker. Semwal,
[3]Vijender Kumar Solanki
[1]Indian Institute of Technology Roorkee
[2]Indian Institute of Information Technology Dharwad
[3]Institute of Technology and Science
Ghaziabad, Uttar pradesh
[1,2]{sachinagnihotri16, vsemwal}@gmail.com
[3]spesinfo@yahoo.com

Prayag Tiwari, Denis Kalitin
Computer Science Department
National University of Science & Technology MISIS
MOSCOW, Russian Federation
{prayagforms, kalitindv}@gmail.com

***Abstract*— Road and traffic accidents are one of the important concerns in today's world. Every country receives a huge damage from road accidents in terms of public health and property loss. Therefore, road accident analysis plays an important role in public health domain. Road accident analysis is performed in order to identify the associated factors that are responsible for road accidents. Knowledge of these factors would be very useful to understand the circumstances of road accidents and can be used to avoid the road accidents. One of the problems in accident analysis is that most of the road accident data is of biased nature. For example, the critical road accidents are very few in comparison to slight/minor injury accidents. Various studies has focused that clustering prior to analysis can increase the efficiency and accuracy of classification. The motive of this study is to perform a conjoint analysis on road accident data, to investigate improvement in the performance of classification of unbiased data after clustering.**

***Index Terms*—Clustering; Road Safety; Bayesian Network; Accident Analysis**

## I. Introduction

Road and traffic accident [1] is one of the biggest harm received from the transportation to the public health. Transportation systems itself is not responsible for these traffic accidents but several other factors [2, 3]. These factors can be defined as environmental factors such as weather and temperature, road specific factors such as road type, road width, and road shoulder width, human factors i.e. wrong side driving, excess driving speed and other factors. Whenever a road accident took place in any road across the world, some of these accident factors are involved. Also, these factors and their influence on road accident are not similar in all countries; but they influenced every road accidents in different countries in different ways. Several studies [4-13] have focused on identification of these factors so that relationship between accident factors and accident severity can be established. This relationship can be utilized to overcome the accident rate by providing some preventive measures [13]. Analysis of road and traffic accidents is widely known as road and traffic safety in which outcome of accident analysis can be utilized for traffic accident prevention. The literature in the traffic safety domain is quite rich as it consists of several research studies [14-20] on road accident data analysis using several techniques such as statistical techniques, mathematical models, data mining and machine learning techniques. It has been observed that classification accuracy is one of the most important parameter to evaluate the performance of the classifier on certain data sets. But, if the data is not balanced or if the distribution of target attribute class values is not uniformly distributed, the classifier accuracy can be biased. In this study, we are using k-modes clustering and Bayesian networks to perform a conjoint analysis on imbalanced road accident data from Leeds, UK in which severe injury accidents and slight injury accidents has a large difference in accident counts. The results reveal that although conjoint analysis on imbalanced data is efficient enough to improve the accuracy of classifier but it is not guarantee that all clusters will achieve a biased classification or improved performance that ca n be achieved without clustering. The organization of the paper is as follows: The section 2 will discuss about the data set used and the methodology adopted for this study. Section 3 will discuss the experimental results and discussion. Finally, we conclude in section 4.

## II. Materials and Methods

### A. Data Set

The data set used for this study is obtained from Leeds, UK [21]. The data set consists of 14 attributes and 1246 accident records over a period of five years from 2011-2015. The

accident attributes in the data are geo-coordinates of the accident locations, number of vehicles, accident date, time, month and year, type of victim, sex of victim, type of accident, severity of accident, type of vehicle, road type, road surface conditions, weather conditions etc.

### B. Cluster Analysis

Clustering [22] provides homogeneous segments out of the large data set. Usually, clustering is applied on large data set in which class labels are missing. After clustering, homogeneous segments are achieved, this can be assigned with a label after investigating the properties of the data objects in the group. We have used k-mode clustering technique to segment our accident data into homogeneous groups. K-modes algorithm [23] is an enhanced version on traditional k-means algorithm with only difference of the similarity measure that is given as follows.

The distance function of k-modes algorithm can be defined as,

$$d(A,B) = \sum_{i=1}^{x} \delta(A_i, B_i) \quad (1)$$

Where,

$$\delta(A_i, B_i) = \begin{cases} 1, & If(A_i = B_i) \\ 0, & If(A_i \neq B_i) \end{cases} \quad (2)$$

Given a set of categorical data objects D defined by n attributes $A_1$, $A_2$… $A_n$. A mode of D = {$D_1$, $D_2$…., $D_n$} is a vector V= {$v_1$, $v_2$…,$v_3$} that minimize

$$d(D,V) = \sum_{i=1}^{n} d(D_i, V) \quad (3)$$

K-modes algorithm is quite suitable for nominal or categorical data sets. Our accident data consists of categorical attributes; hence we have selected k-modes clustering for road accident analysis. The procedure of k-modes algorithm is given as follows:

*K-mode clustering Algorithm:*

Input: Data set D, k number of cluster to be formed

Output: k clusters

1. Initially select k random objects as cluster centers or modes
2. Find the distance between every object and the cluster centre using k-modes distance measure
3. Assign each object to the cluster whose distance with the object is minimum
4. Select a new center or mode for every cluster and compare it with the previous value of centre or mode; if the values are different, continue with step 2.

### C. Number of Cluster Selection

In order to determine the number of clusters to be formed out of the data, Bayesian information criteria (BIC) is used [24]. The BIC criteria can be defined in Eq.4.

$$BIC = -2\log L + p \log(n)) \quad (4)$$

Where, p is the number of model parameters and n is the sample size.

### D. Bayesian Networks

Bayesian Networks (BNs) have proven track record in the field of data analysis. It is widely applicable to establish relationships between different set of attributes using probabilistic calculations. It has wide applications is bioinformatics, text classification, medicine, information retrieval, gaming and transportation. In BNs, the relationships between different set of variables is represented by arcs or edges in a graph, and variables are represented as nodes.The detailed desciption about Bayesian Networks can be found in [25-26].

### E. Performance Evaluation Parameters

In this paper, several performance paameters [22] have been used to calculate the model fittingfor every clusters made from the data. These indicators/parameters are accuracy, sensitivity, specificity and the HMSS (Harmonic means of sensitivity and specificity) and ROC area. This indicators can be calculates using follwing equations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (5)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (6)$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \quad (7)$$

$$HMSS = \frac{2 \times Sensitivity \times Specificity}{Sensitivity + Specificity} \quad (8)$$

Where, TP-True Positive, TN-True Negative, FP-False Positive, FN-False Negative.

## III. Results and Discussion

This section presents the experimental analysis and discussion on results. Initially, data preprocessing is performed on the road accident data to give it a proper shape required for analysis. Several attributes are transformed into suitable form using data transformtion methods..

### A. Cluster Analysis

After data selection and data preprocessing, the selected data is used for cluster analysis using k-modes clustering algorithm. The number of clusters for k-modes algorithm is determined by observing the BIC values for different cluster models. The Fig 1 illustrates the cluster selection using BIC values.

Based on fact mentioned in previous studies [27-28], a cluster model with 4 clusters is selected. Further, k-modes technique is applied on the data and the four cluster obtained. The description of these four clusters are given in Table 1.

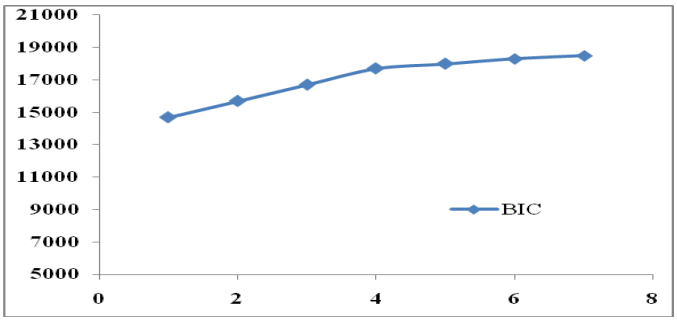

Fig. 1: Cluster selection using Bayesian Informantion Criteria

Table1: Cluster description

| Cluster Id | Description | No. of accidents | Size (%) |
|---|---|---|---|
| C1 | Two wheeler accidents in bad weather | 196 | 15 |
| C2 | Other accidents in bad weather | 299 | 24 |
| C3 | Other accident in clear weather | 247 | 20 |
| C4 | Two wheeler accidents in bad weather | 504 | 41 |

Table 2: Bayesian network performance on each cluster and whole data set

| Subset | Accuracy | Sensitivity | Specificity | HMSS | ROC |
|---|---|---|---|---|---|
| C1 | 0.88 | 0.68 | 0.91 | 0.778 | 0.776 |
| C2 | 0.85 | 0.59 | 0.94 | 0.724 | 0.874 |
| C3 | 0.78 | 0.40 | 0.93 | 0.559 | 0.796 |
| C4 | 0.85 | 0.78 | 0.86 | 0.818 | 0.708 |
| WD | 0.84 | 0.59 | 0.90 | 0.712 | 0.856 |

### *B. Performance Evaluation of Bayesian Network*

Further, Bayesian Networks (BNs) are used to investigate the responsible factors that contribute to accident severity. Therefore, several BNs were built for each clusters and the whole data set.

The main objective of this study is to identify if some new findings are there after performing a conjoint analysis (k-modes and BN). Further, these BNs that were built for 4 clusters and whole data set were compared using performance indicators and complexity to validate the goodness of model fitting. Table 2 illustrates the accuracy, sensitivity, specificity and HMSS and ROC for each cluster and whole data (WD).

It can be seen from Table 2 that minimum accuracy is achieved in C3 and the highest accuracy is achieved in C1. As the accident data was imbalanced data, ROC values are also taken into consideration. The ROC values indicate that performance of classification is better in C2 whereas in other clusters, the ROC values are lower than the ROC value of WD. It simply indicates that although more accuracy can achieved as a result of clustering process but is the data is of imbalanced nature, it is not guarantee that efficient classification results can be achieved.

## IV. Conclusion

The paper presents a conjoint analysis using k-modes clustering and Bayesian Networks on an imbalanced road accident data from Leeds, UK. The main objective of this study was to validate the performance of classification before and after the clustering process. Initially, the k-modes algorithm is used to cluster the data into 4 homogeneous groups and further these clusters and the whole data set is analyzed using Bayesian Networks. Different Bayesian Networks are built for each cluster and the entire data. Further, these Bayesian Networks are evaluated on the basis of performance indicators. The result indicates the classification accuracy is slightly improved as a result of clustering process but the ROC values are slightly decreased for some clusters. This indicates that performance of the classifier in terms of accuracy is biased towards one class value which has comparatively large number of instances. The future work will comprise of detailed analysis of these Bayesian Networks to establish the relationships between different road accident attributes to identify which attributes have higher impact on severity of accidents.

## References

[1] World Health Organization. Global Status Report on Road Safety 2015. Available online: http://www.who.int/violence _injury_ prevention/ road_safety_status /2015/GSRRS2015_ Summary_ EN_final2.pdf?ua=1 (accessed on 01.07.2016).

[2] Mussone, L., Ferrari, A. and Oneta, M. An analysis of urban collisions using an artificial intelligence model. *Accid Anal Prev* 1999, vol. 31, pp. 705-718.

[3] Kumar, S. and Toshniwal, D. A data mining framework to analyze road accident data. *Journal of Big Data, vol. 2, No. 26, pp. 1-18.*

[4] Chang, L.Y. and Chen, W.C. Data mining of tree based models to analyze freeway accident frequency. *J Saf Res Elsevier.* 2005; vol. 36.

[5] Kumar, S. and Toshniwal, D. A novel framework to analyze road accident time series data. *Journal of Big Data*, vol. 3, No. 8, pp. 1-11.

[6] Kumar, S and Toshniwal, D. A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, vol. 24, Issue-1, pp. 62-72.

[7] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (San Mateo, CA: Morgan Kaufmann) **2** (12): 1137–1143.

[8] Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme". *Biochimica et Biophysica Acta (BBA) - Protein Structure,* 1975, **405** (2): 442–451.

[9] Fawcett, Tom (2006). "An Introduction to ROC Analysis". *Pattern Recognition Letters,* 2006, **27** (8): 861 – 874.

[10] Kumar, S.; Toshniwal, D. Analysis of road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). *Journal of Big Data*,3, 13:1-11.

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

[12] Kumar S and Toshniwal D, Analysing road accident data using association rule mining, *Proceedings in IEEE International Conference on Computing, Communication and Security(ICCCS2015) held in Mauritius. India:* IEEE Xplore; 2015

[13] Kumar S, Toshniwal D and Parida M. A comparative analysis of heterogeneity in road accident data using data mining techniques. Evolving Systems. Springer, 2016. DOI: 10.1007/s12530-016-9165-5.

[14] Geurts K, Wets G, Brijs T, Vanhoof K (2003). Profiling of high frequency accident locations by use of association rules. Transportation Research Record-1840. Doi: 10.3141/1840-14.

[15] Tesema TB, Abraham A, Grosan C, Rule mining and classification of road accidents using adaptive regression trees. Int J Simulation, vol. 6, 2005, pp. 80–94.

[16] Abellan J, Lopez G, Ona J, Analysis of Traffic Accident Severity using Decision Rules via Decision Trees. Expert System with Applications. Vol. 40, 2013, pp. 6047-6054.

[17] Kashani T, Mohaymany AS, Rajbari A, A Data Mining Approach to Identify Key Factors of Traffic Injury Severity. Promet-Traffic & Transportation. Vol. 23, 2011, pp. 11-17.

[18] Kwon O H, Rhee W, Yoon Y, Application of Classification Algorithms for Analysis of Road Safety Risk Factor Dependencies. Accident Analysis and Prevention. Vol. 75, 2015, pp.1-15. Doi:10.1016/j.aap.2014.11.005.

[19] MacQueen J, Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967. http://projecteuclid.org/euclid.bsmsp/1200512992.

[20] Tibshirani R, Walther G, Hastie T, Estimating the Number of Clusters in a Data set via the Gap Statistic, J. R. Statist. Soc. B. vol. 63, 2001, pp.411-423. Doi: 10.1111/1467-9868.00293.

[21] Data Source: https://data.gov.uk/dataset/road-traffic-accidents, accessed on 01.Oct.2016.

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

[23] Chaturvedi A, Green P, Carroll J, K-modes Clustering, *Journal of Classification*, vol. 18, 2001, pp. 35-55.

[24] Raftery A E. A note on Bayes factors for log-linear contingency table models with vague prior information. *J Roy Stat Soc B*, vol. 48, 1986; pp. 249–50.

[25] M. G. Madden, "On the classification performance of TAN and general Bayesian networks", *Journal of Knowledge-Based Systems*, vol. 22, 2009, pp. 489–495.

[26] H. Helai, C.H. Chor, M.M. Haque, Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis, *Accident Analysis and Prevention*, vol. 40, 2008, pp. 45–54

[27] J. M. Pardillo-Mayora, C. A. Domínguez-Lira, R. Jurado-Piña, "Empirical calibration of a roadside hazardousness index for Spanish two-lane rural roads", *Accident Analysis and Prevention,* vol. 42, 2010, pp. 2018–2023.

[28] Depaire B, Wets G, Vanhoof K, Traffic Accident Segmentation by means of Latent Class Clustering. Accident Analysis and Prevention. Vol. 40, 2008, pp.1257-1266. Doi:10.1016/j.aap.2008.01.007.