Research Journal of
Applied Science
MEDWELL PUBLICATIONS

The Application of Binary k-Means Clustering to Identify Groups of Road Traffic Accident's Factors in United Kingdom

Nur Atiqah Binti Hamzah, Sabariah Binti Saharan and Sie Long Kek
Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia (UTHM), Batu Pahat, Malaysia

Key words: Clustering, k-means clustering, binary data, similarities, road accidents

Corresponding Author:
Nur Atiqah binti Hamzah
Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia (UTHM), Batu Pahat, Malaysia

Page No.: 135-138
Volume: 15, Issue 4, 2020
ISSN: 1815-932x
Research Journal of Applied Sciences

Abstract: Cluster analysis is a formal study of methods and algorithms for natural grouping or clustering of objects according to measured or perceived intrinsic characteristics or similarities in each objects. The pattern of the each cluster and the relationship for each cluster were identified and then relate with the frequency of occurrence in the data set. This study aims to apply one of well-known clustering techniques, k-means clustering into binary data set in order to cluster the factors of road traffic accidents as the number of road accidents is increasing from day to day. Although there might be a list of expected factors that causing the road traffic accidents, none of us known which group of factors that has highest contribution that lead to road accident. By using k-means clustering, the patterns of road traffic accidents factors were identified.

INTRODUCTION

Cluster analysis is a formal study of methods and algorithms for natural grouping of objects according to similarities in each object. Clustering are important techniques used in order to distinguish objects that have many attributes into meaningful disjoint subgroups so that each variables in the group are having similar characteristics to each other. Cluster analysis is used to separate data elements into groups by maximizing the homogeneity within elements of clusters and heterogeneity between clusters[1]. Clustering also known as unsupervised learning algorithm because the true number of clusters are unknown[2].

Specifically, the k-means method was applied in this study. k-means clustering is a data mining machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. K clusters is represented by the mean of the objects which known as centroid. By computing the distance between each pairs of factors, the similarities were measured. As the distance computed, the most importance factors also can be defined.

k-means is one of well-known clustering techniques proposed by MacQueen in 1967. Because of its simplicity in computations, many researchers still used this technique in their research until today. The multiple category attributes can be transformed into binary attributes with 0 as absent and 1 for present which act as numeric in the k-means[3]. To find the cluster solution, the important steps in k-means is to decide the K value and initial points as the center for clustering iteration process. k-means method require a suitable distance measure that

fits with the data used[4]. In this study, the selected distance measurement is Hamming distance as it is suitable for binary data[5].

To choose the best K cluster, cluster validation index is used. Silhouette Index is one of cluster validation method. It was used to evaluate and determine the optimal cluster solution[6]. From range -1 to 1, the highest value of index will be considered as the best cluster solution.

As road accidents was an interesting issue with an increasing trends from year to year in most of developing countries, this study aimed to apply k-means clustering to find factors of road traffic accidents. The roads legislation seems cannot control the increasing trend of accidents. This study can become a contribution by identifying the groups of road accidents factors to help the nations take a wise step to be more careful in order to avoid accidents. Accidents cause a high loss to the victims and also becomes a reason of severity and mortality. Thus, organizing the data related with road accidents is important to do cluster analysis.

k-means clustering method was applied to the United Kingdom road traffic accidents data in order to find the group of factors of road traffic accidents. The December 2014 data was used with n = 11,952 and p = 42 factors. The road accidents data consisted of three severity level which are fatal, serious and minor.

MATERIALS AND METHODS

In order to find the cluster of road accidents factors, there are a few steps that need to be done as shown in Fig. 1.

Data cleaning: From 12,037 data, only 11,952 were selected after cleaning process. The missing data was not included for further analysis to avoid misclassification results of clustering.

Selection of variables: Only related variables which recognized as road accidents factors were chose.

Data standardization: The raw data was in categorical form. The variables were transformed into binary form with representation as accidents observed (Yes = 1) or otherwise (No = 0).

Getting cluster solution: When the data was ready, the clustering process was taken place. By using Hamming distance, the distance of each data to the closest distance were calculated. Hamming distance was used because it was the most suitable measurement for binary data set. The distance between data points can be calculated by following formula:

$$d(i, j) = q+r$$

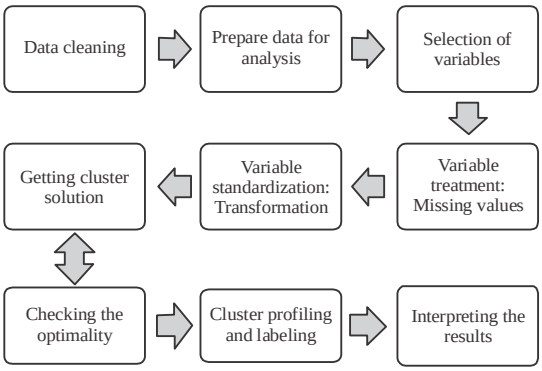

Fig. 1: Analysis process

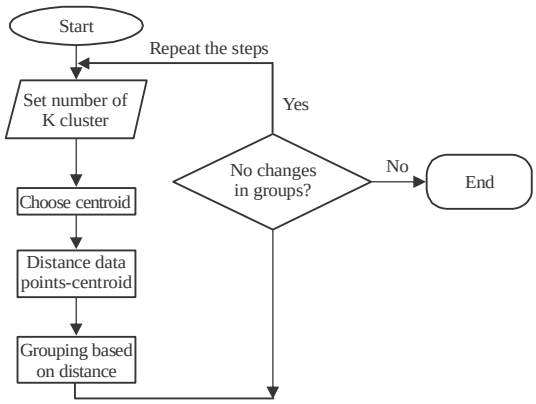

Fig. 2: k-means clustering steps

To apply this Hamming distance formula, the data can be represented in strings form. The Hamming distance can only be calculated between two strings of equal length. For example:

i object: 1001 0010 1101
j object: 1010 0010 0010

The strings above are compared. If i = 1 and j = 0, then q = 1. If i = 0 and j = 1, then r = 1. Based on the string above, q = 4 and r = 2. Thus, the distance between i object and and j object is d (i, j) = 4+2 = 6 (Fig. 2).

To start the clustering process, the number of K need to be set. In this study, the number of K used to be tested was from 2-8. Then, one data point was selected as center. By using the formula of Hamming distance above, the distance between each data points to center was calculated. After the distance was obtained, the data points were grouped into cluster based on the shortest distance. The iterations process were continued until there are no changes in groups as it has achieved its stable conditions. For this study, all the process were done using MATLAB R2015A.

Checking the optimality: To select the best number of K, Silhouette Index was used to compare the validation index of the cluster solutions obtained. The range of Silhouette Index is from -1 to 1. The cluster that has highest Silhouette Index value was chose as the best number of K. Finally, the results can be further interpreted.

RESULTS AND DISCUSSION

The number of cases based on severity level were not balanced. The total number of fatal cases was only 1.49 and 85.19% out of 11,952 cases were for minor accident cases. By using Matlab R2015A on 64-bit Windows 7 operating system, the clusters of related factors were defined. The value of K = 2, …, 8 was used to find the stable clusters for this data set. The Silhouette index for each value of K was defined in order to find out the optimal value of clusters.

From Table 1, the highest Silhouette index was when K = 2 followed by K = 7 which K = 7 was chosen as the optimal number clusters for road accident factors. This index value was chosen as it was closest to 1. Although, K = 2 had the highest Silhouette index, it was not selected as the differences and similarities of each factors were hard to be defined clearly.

From Fig. 3, the second, third, fourth, sixth and seventh cluster had the highest index which were equal to 1.0. This shows that the factors in these five cluster were the main factors of the road accident for year 2014 in United Kingdom. However, there were some misclassifications occurred in the clusters. The misclassification might happened due to imbalanced data according to its severity level. The factors in each cluster had been identified in Table 2.

From the clusters formed, the relationship between each factors were identified. The second, third, fourth, sixth and seventh cluster contain single factor only. This means that these factors were independent.

The five clusters that had a single variable or independent factor were unknown road type, the accidents were not at a junction, accident at the roundabout, mini-roundabout and footbridge. These factors did not related or dependent with the other factors.

Motorway, speed and slip road junction were the factors in the fifth cluster. From 11,952 cases, motorway showed 130 cases, respectively. Although, the number of cases for the factor motorway were very small but it could be related to the factor speed. Motorway is an express-way road which always become a main road in United Kingdom. The speed traffic in motorway could be the main reason how the accident occur on this road. Meanwhile, slip road junction are interchange junction that usually exist in motorway. This factor also can be related to motorway factor which then the factor can be the reason of accident occurrence. The 213 cases occurred on the existing of slip road junction. As the slip road connects two motorways, speed vehicles on the motorway can cause accident. This phenomenon give the same pattern for these factors to be in the same group.

Table 1: Silhouette value for each number of clusters, K

Value of K	Silhouette Index
2	0.8974
3	0.4623
4	0.4259
5	0.4227
6	0.5264
7	0.5477
8	0.3377

Table 2: Road accident factors based on respective clusters

Cluster	Factors
1	Other road accident factors
2	Unknown road type
3	Roundabout
4	Not at junction or within 20 m
5	Motorways
	Speed
	Slip road junction
6	Mini roundabout
7	Footbridge

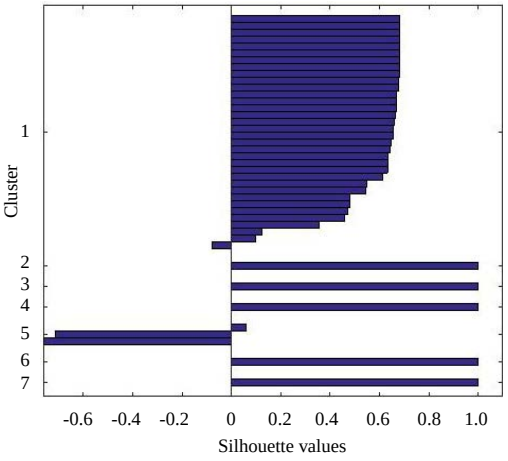

Fig. 3: Silhouette figure for K =7

CONCLUSION

As a conclusion, it shows that all 42 factors had its own strength as road accidents factors. By using the k-means clustering approach, the patterns of factors within clusters were defined. The main factors of road accidents had been recognized based on each cluster index value. From the Silhouette figure, cluster 2, 3, 4, 6 and 7 gave the highest index value which is equal to one. All this five clusters contained only single factor. Thus, five factors can be categorized as the main road accident factors in year 2014 were unknown road type, roundabout, not at junction, mini roundabout and footbridge.

From the results, it was proved that k-means clustering can be applied to binary data set. The used of suitable distance measure and cluster validation index according to the type of data were very useful in order to achieve the best clusters for the data. The Silhouette index played important roles in this study to identify best number of k cluster and groups of main road accident's factors. The patterns of factors of road accident were successfully identified. Thus, prevention methods can be carried out to decrease the number of accidents occurred in future.

RECOMMENDATIONS

As the data was imbalanced according to its severity level, the clustering lead to a misclassification on the clusters formed. Future research can be implemented with new approaches in order to avoid misclassification throughout clustering procedure.

Besides that, this clustering technique was an unguided approach. The number of K for this study need to be decided by our own and number of iterations to achieve stabilization of the cluster is unknown. Future researcher can propose a new method to choose value of K based on number of data and number of iterations. In this study, the limitation was the efficiency of standard computer operating system to carry out the iteration process for the whole year data. Thus, future research also can be done using the whole year data to see the difference of groups of factors that will be formed.

In addition, the authority can take an action in order to reduce the number of road traffic accidents based on the results of the research that had been done. The groups of factors that had been identified are very useful information, thus, the authority can make sure their actions are suitable with the current factors of road traffic accidents.

REFERENCES

01. Hair, Jr. J.F., R.L. Tatham, R.E. Anderson and W.C. Black, 1998. Multivariate Data Analysis. 5th Edn., Prentice Hall International, Englewood Cliffs, NJ., USA., ISBN-13: 978-0138948580, pp: 169-215.
02. Vermunt, J.K. and J. Magidson, 2002. Latent Class Cluster Analysis. In: Applied Latent Class Analysis, Hagenaars, J. and A. McCutcheon (Eds.). Cambridge University Press, Cambridge, UK., ISBN-13: 9781139439237, pp: 89-106.
03. Gowda, K.C. and E. Diday, 1992. Symbolic clustering using a new similarity measure. IEEE. Trans. Syst. Man Cybern., 22: 368-378.
04. Finley, T. and T. Joachims, 2008. Supervised k-means clustering. Masters Thesis, Cornell University, Ithaca, New York, USA.
05. Wagstaff, K., C. Cardie, S. Rogers and S. Schrodl, 2001. Constrained k-means clustering with background knowledge. Proceedings of the 18th International Conference on Machine Learning, June 28-July 1, 2001, San Francisco, CA., USA., pp: 577-584.
06. Deborah, L.J., R. Baskaran and A. Kannan, 2010. A survey on internal validity measure for cluster validation. Int. J. Comput. Sci. Eng. Surv., 1: 85-102.