ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Survey on the role of k-means Clustering Algorithm in Privacy Preserving Data Mining.

Keerthika K, Manikandan G*, Harish V, and Nooka Saikumar.

School of Computing, SASTRA University, Thanjavur, Tamilnadu, India.

ABSTRACT

With the rapid development in data storage technologies it is possible for the organizations to store ample data. Data mining is a familiar method used for gathering unseen patterns from a large quantity of data. One of the drawbacks with data mining is that it also reveals some private information. Privacy preserving data mining is an emerging research area which aims at preserving the private data from accidental disclosure. The foremost intention of privacy preserving data mining is to devise proficient algorithms that can extract significant information from the heterogeneous data sources without revealing sensitive information. Misclassification error is used as a metric to evaluate the accuracy of the PPDM technique. This paper analyzes the use of different variations of k-means algorithm in computing the misclassification error.

Keywords: Data Privacy, Misclassification Error, Normalization, Clustering, Mutation.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

The recent progress in several technologies permits the organizations to store a lot of data. Data mining is one of the most prominent approaches implemented by many administrations to examine the stored data to arrive at an efficient business conclusion. Data mining techniques leads to a security breach if not implemented in the right way. Privacy preserving data mining (PPDM) is the only possible technique to assure security for the stored data throughout the mining process by conserving the privacy of the sensible attribute from disclosure. Efficiency of the PPDM is conveyed by employing different privacy metric functions and one among them is misclassification error.

This work proposes variations to the basic k-means clustering algorithm and identifies the best one with maximum efficiency. Clustering is implemented with the original data and also with the modified data after implementing a privacy technique. Experiments were done by varying the number of clusters i.e 'k'.

PRIVACY PRESERVING TECHNIQUES

The theme of privacy preserving data mining is to change the original data so that the data privacy is maintained throughout the complete track of the mining process. In this section we put forward a brief sum-up of several privacy techniques used to yield the sanitized information.

Fuzzy-membership Functions

Fuzzy membership approach is implemented for preserving privacy by yielding sanitized information from the raw data set [1]. Information from the sanitized data is represented in the scale of 0-1 and it also depends on the type of membership function. Since the complete data set is mapped to a small range it's utilization is bounded to certain attributes and cannot be inferred to all attributes.

Normalization

Normalization is employed to map the data in a different scale. The most eminent normalization techniques are Decimal Scaling, Z-Score and Min-Max. Decimal scaling changes the original data by displacing the decimal point. Mean and standard deviation is used to generate the modified data in Z-score normalization [2]. Min-Max normalization yields the modified data linearly.

Mutation

Mutation, a genetic manipulator is used to maintain diversity in the given population. It yields an exclusively dissimilar offspring from the parent. The generated offspring depends on the type of mutation applied. Mutation is implemented when the data is available in the binary format. Based on the above observance this manipulator can be utilized to conserve privacy [3-5].

Rule Based Approach

In this approach the original data is falsified by adding up a noise to the original data. The unicity of this approach is that the noise is generated based on the attributes present in the given data set. The approach employs heuristics for generating dynamic rules [6-7].

Flipping / Substitution

This technique can be implemented for the numerical values. In the beginning of this approach, the information in the data set is interpreted in the binary form. The LSB in the binary data is flipped to maintain a relationship between the original and the sanitized data. Decimal value of the flipped binary replaces the data in the original data set to conserve privacy [8-9].

ISSN: 0975-8585

CLUSTERING ALGORITHMS

The aim of the clustering algorithms is to place the similar elements in a cluster. Clustering is a good example for unsupervised learning. Firstly the similar elements are grouped in a cluster and a label for the cluster is then assigned based on the contents of the cluster. In this discussion we provide a brief sum-up of different variations of k-means clustering algorithms namely Traditional, Random and Division. These three approaches differ only in the selection of the initial centroid values.

Traditional Approach

In this approach, the initial centroid values are selected in such a way that, they are the first most non-recurring elements of the dataset. Based on the number of clusters, the initial centroid values are selected.

Random Approach

In this method of implementation, the initial centroid values are selected in a random manner. The execution of the algorithm starts with a different initial centroid values for each iteration which makes the clustering of the sanitized data very unusual.

Division Approach

In this mechanism, the initial centroid values are selected in a different fashion. Initially, the whole Dataset is divided into equal parts, based on the number of clusters required. Later, the initial value of each group is taken as the initial centroid value.

EXPERIMENTAL RESULTS

In this paper, we have considered various attributes such as Age, Gender and Income from the adult data set available in the UCI repository [10]. The data set consists of about 32561 records. The above proposed approaches have been implemented using JAVA Programming language and the resulting observations are tested in Intel core i3 processor with 4GB RAM and Windows 10 operating system. From our experimental results it is evident that the original data cannot be inferred from the modified data by homogeneity attack and background knowledge attack. Figure 1, 2 and 3 shows the outcome of our experimental study.

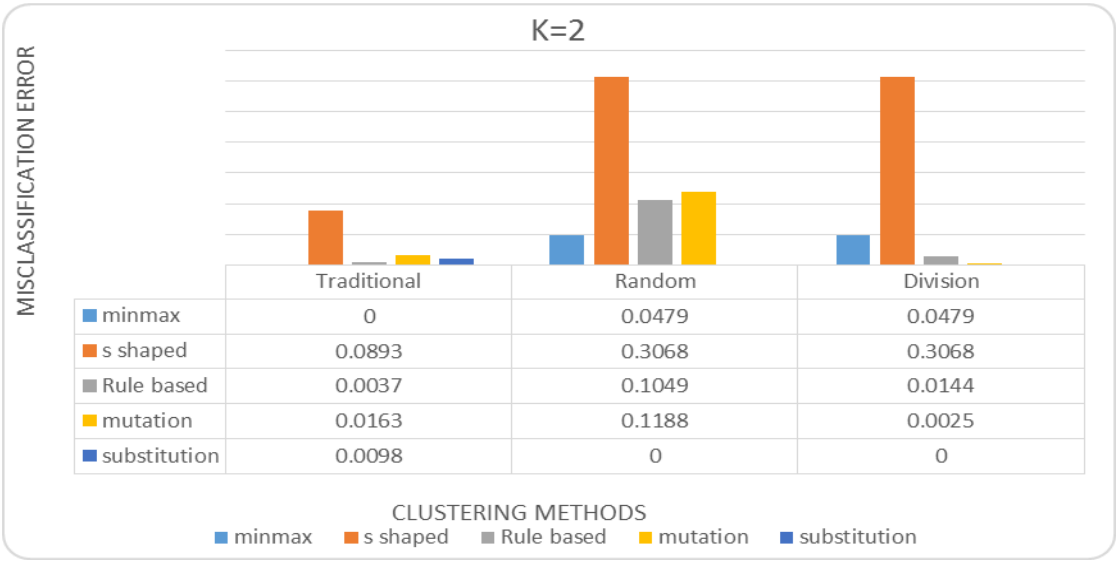

Fig. 1 – Misclassification error for k-means with 2 clusters.

ISSN: 0975-8585

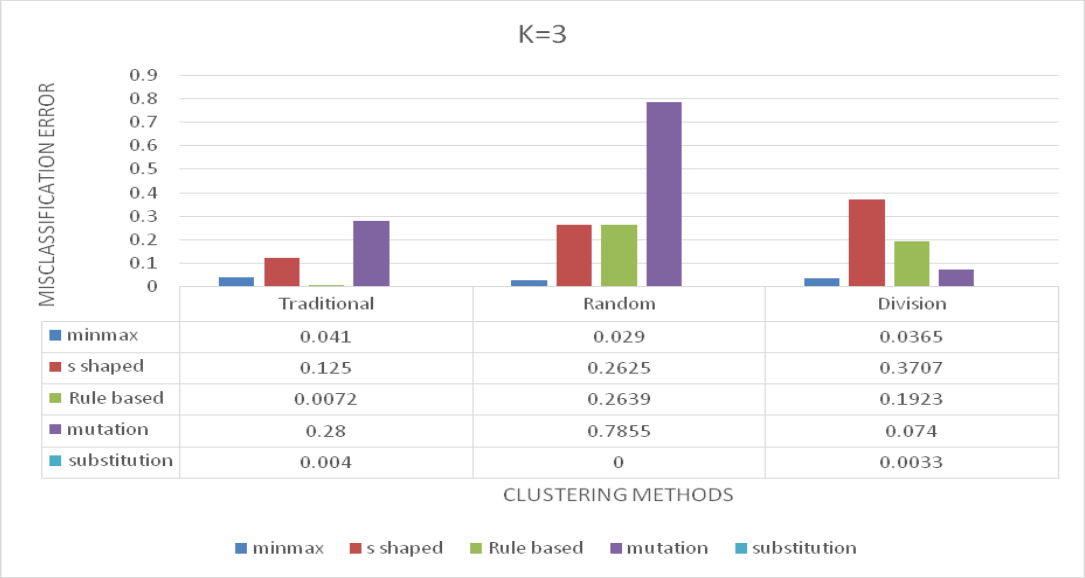

Fig. 2 – Misclassification error for k-means with 3 clusters.

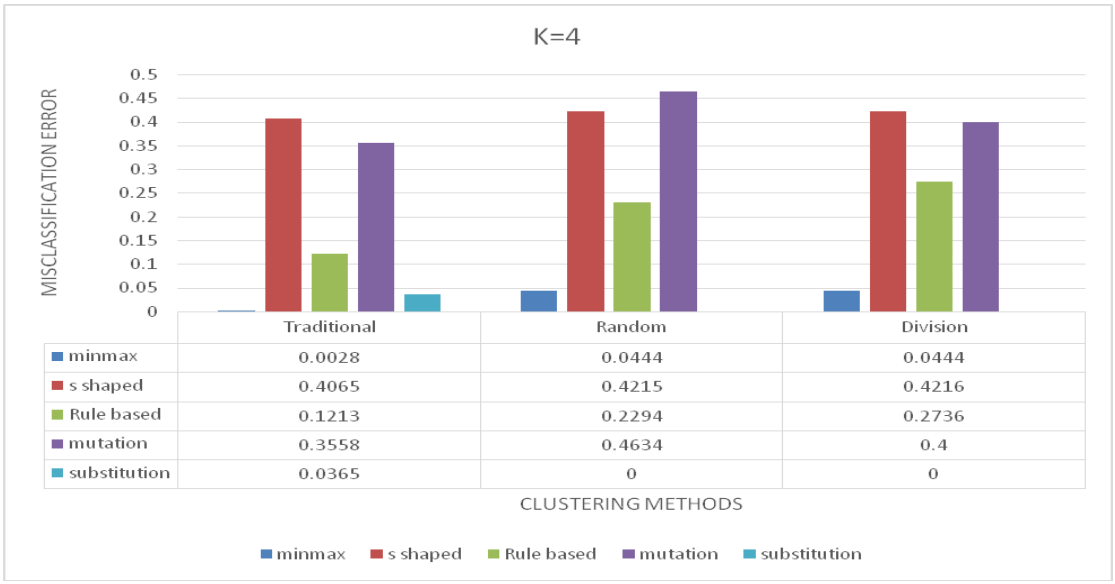

Fig. 3 – Misclassification error for k-means with 4 clusters.

CONCLUSION

One of the major ambitious tasks in data mining is to maintain data privacy. Diverse methodologies have been implemented for this purpose. An approach is averred to be an effective one if it maintains very low misclassification error. In this paper we have calculated misclassification error for various approaches with variations in k-means clustering algorithm. From the study it is apparent that the traditional k-means algorithm is the best option for calculating misclassification error since it results in an optimal value when compared with the other approaches.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Science and Technology, India for their financial support through Fund for Improvement of S&T Infrastructure (FIST) programme SR/FST/ETI-349/2013.

ISSN: 0975-8585

REFERENCES

[1] B.Karthikeyan,G.Manikandan,Dr.V.Vaithiyanathan. Journal of Theoretical and applied information Technology 2011; 32:118-122.

[2] G.Manikandan,N.Sairam,S.Sharmili,S.Venkatakrishnan. Indian Journal of Science and Technology 2013; 6: 4268-4272.

[3] [G.Manikandan,N.Sairam,S.Jayashree,C.Saranya. Middle East Journal Of Scientific Research 2013;14:107-111.

[4] G.Manikandan,N.Sairam,C.Akshaya,S.Venkatakrishnan. Journal of Applied Engineering Research 2014; 9:589-597.

[5] G.Manikandan, N.Sairam, S.Rajarajeswari, H.Ramya. International Journal of Applied Engineering Research 2014; 9:755-761.

[6] Manikandan G, Sairam N, Rajendiran P, Balakrishnan R, Rajesh Kumar N, Raajan NR.; Journal of Computation and Theoretical Nanoscience 2015 ;12: 5463-5466.

[7] G Manikandan, N Sairam, M.SathyaPriya, Sree Radha Madhuri, V Harish, and Nooka Saikumar.Journal of Engineering and Applied Sciences 2016; 11:8063 – 8066.

[8] Manikandan G, Sairam N, Harish V, Nooka S.Research Journal of Pharmaceutical, Biological and Chemical Sciences 2016; 7: 1136-1139.

[9] Manikandan G, Sairam N, Harish V, Nooka S. Research Journal of Pharmaceutical, Biological and Chemical Sciences 2016; 7: 344-348.

[10] UCI Data Repository http://archive.ics.uci.edu/ml/datasets.html.