ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Survey to Identify a Suitable Clustering Algorithm for Estimating the Efficiency of Privacy Preserving Data Mining Techniques.

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

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

ABSTRACT

With the current advancement in information technology we have data warehouses containing abundant data which results in a new dimension of research in data mining arena known as Privacy Preserving Data Mining (PPDM). Many algorithms have been proposed incorporating a privacy mechanism that allows the users to extract the required information without revealing the sensitive data. The efficiency of privacy preserving algorithm is computed using a privacy metric known as misclassification error. The prime objective of this paper is to analyze various clustering algorithms and to identify the one with the minimum misclassification error.

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

Corresponding author

ISSN: 0975-8585

INTRODUCTION

The recent advancement in various technologies allows us to accumulate and store enormous data. Data mining is the leading technique used by various organizations to analyze the stored data and to extract the useful information for arriving at an effective business decision [1]. Data mining techniques results in a security breach if not used in the proper manner. Privacy preserving data mining (PPDM) is an emerging solution to ensure security for the stored data during the mining process by preserving the privacy of the sensitive attribute from disclosure [2]. Efficiency of the PPDM is expressed using various privacy metrics and one among them is misclassification error.

This work analyzes various clustering algorithms and identifies the best clustering algorithm which results in the minimal misclassification error. Clustering is performed with both the bench mark data set and the one obtained after applying a privacy technique [3]. Clustering process is repeated with different 'k' values to identify the best clustering algorithm which results in a minimal misclassification error.

Privacy Preserving Techniques

The idea of privacy preserving data mining is to modify the original data such that the sensitive information remains confidential during the entire course of the mining process. In this section we provide a brief summary of various privacy techniques used to generate the sanitized data.

Fuzzy-membership Functions

Fuzzy membership functions can be used to preserve privacy by generating a sanitized data from the original data set [4-5]. Data in the sanitized data is mapped in the range between 0-1 and it depends on the type of membership function used. Since the entire data set is mapped to a small range it's usage is limited to certain attributes and cannot be generalized to all attributes.

Normalization

Normalization is used to represent the data in a different scale. The most prominent normalization techniques are Decimal Scaling, Min-Max, and Z-Score. Decimal scaling modifies the original data by moving the decimal point. Min-Max normalization generates the modified data linearly [6]. Z-score normalization uses the mean and standard deviation to generate the modified data.

Mutation

Basically mutation is a genetic operator used to maintain diversity within the given population. It generates an entirely different offspring from the parent. The generated offspring depends on the type of mutation used. Mutation is performed when the data is in binary form. Based on the above observation this operator can be used to preserve privacy[7-9]. Uniform mutation is used in our study.

Rule Based Approach

In this approach the original data is altered by adding a noise to the original data. The uniqueness of this approach is that the noise is generated based on the attributes in the given data set. The rules are dynamic and they are formed using heuristics [10-11].

Substitution

This technique is applicable only for the numerical data. In the first step the data in the data set is represented in the binary form. In order to maintain a relationship between the original and the modified data only the LSB in the binary data is flipped. Decimal value of the flipped binary replaces the data in the original data set to preserve privacy [12-13].

ISSN: 0975-8585

Clustering Algorithms

The objective of the clustering algorithm is to group the similar elements in the same cluster. Clustering is an example for unsupervised learning where the elements are first grouped into different clusters and then the clusters are labeled based on the similarity of elements it possess. In this section we provide a brief summary of various clustering algorithms used in this study.

k-Means

This algorithm creates different clusters from the data set by placing the more similar elements in a cluster. Here 'k' denotes the number of clusters to be generated. Several variations exist for this algorithm depending on the initial centroid value selection process. In general, the initial values in the data set are used as the initial mean values for the clusters. This algorithm repeats by selecting a new mean for successive iterations and terminates when there is no deviation in the mean values.

Farthest First

Farthest First clustering algorithm is same as the K-means clustering with a small variation in it. Every cluster center is placed at the point farthest from the existing cluster centers which must mandatorily lie within the instances. The main advantage of this algorithm is the data is clustered in a greater rate since only less adjustments and reassigning of center points are involved.

Expectation Maximization

Every instance of the dataset is assigned a probability distribution which shows the probability of that instance belonging to each of the clusters. EM can decide the number of clusters to be created by the cross validation or the user can specify the number of clusters that is to be generated.

X-Means

This is K-means algorithm extended by an Improve-Structure part. The centers are attempted to be split in its region in this part of the algorithm. The Bayesian-Information Criterion value is computed for the two structures and the decision between the children of each cluster and itself is done.

RESULTS

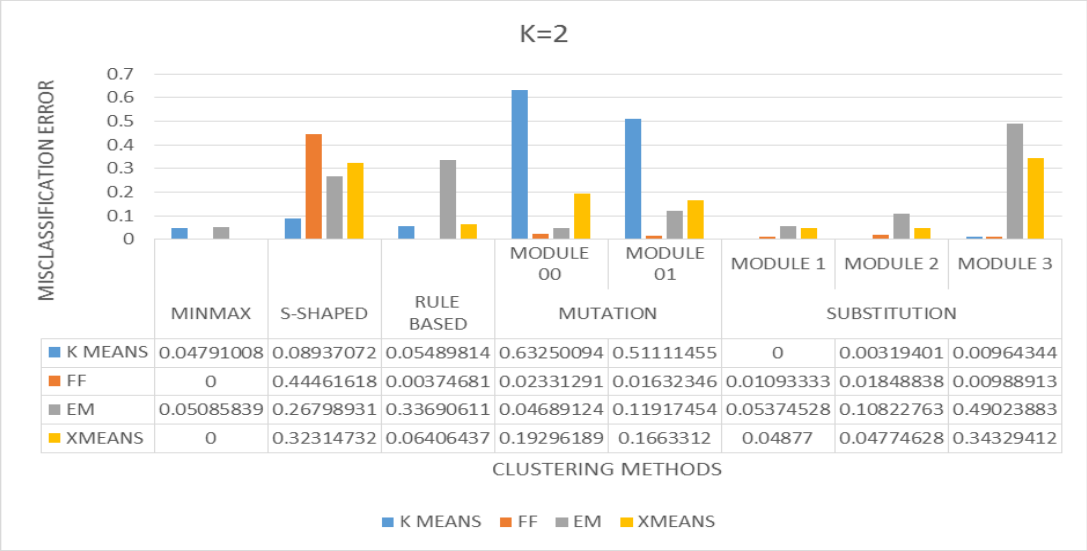

Fig. 1 – Misclassification error for clustering methods with 2 clusters.

ISSN: 0975-8585

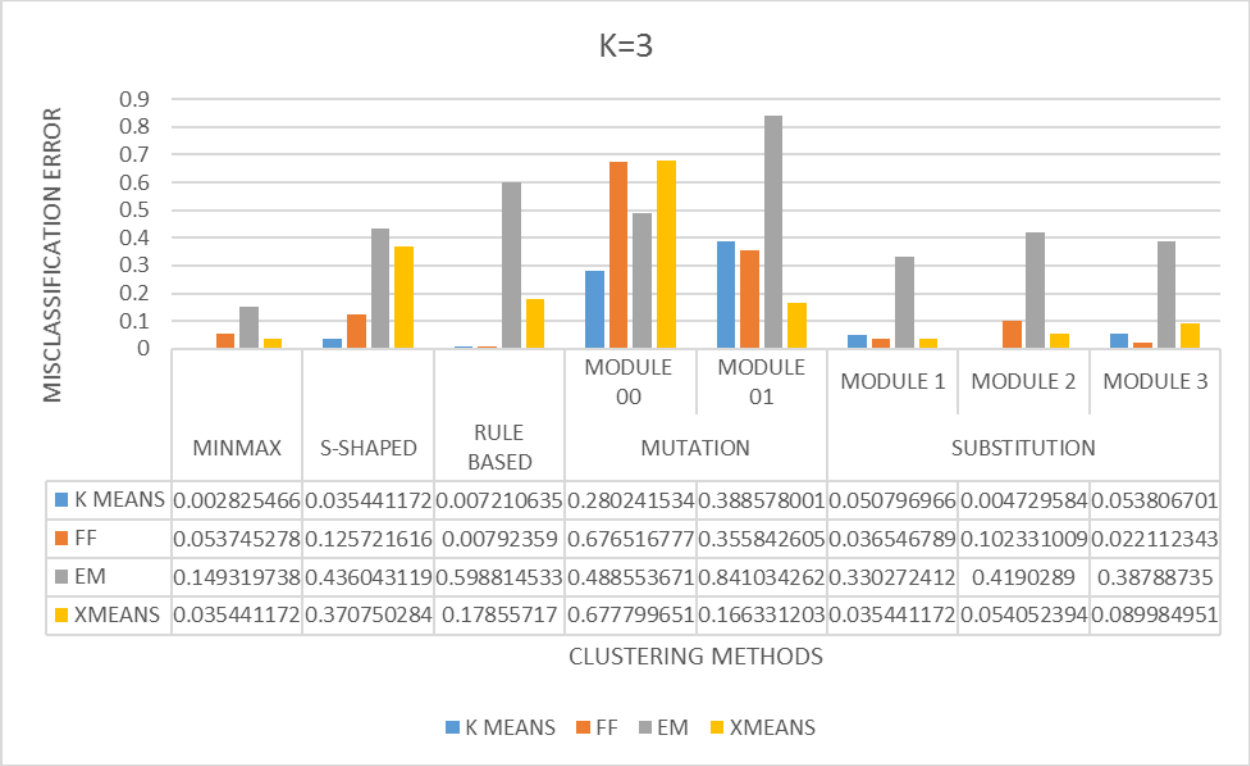

Fig. 2 – Misclassification error for clustering methods with 3 clusters.

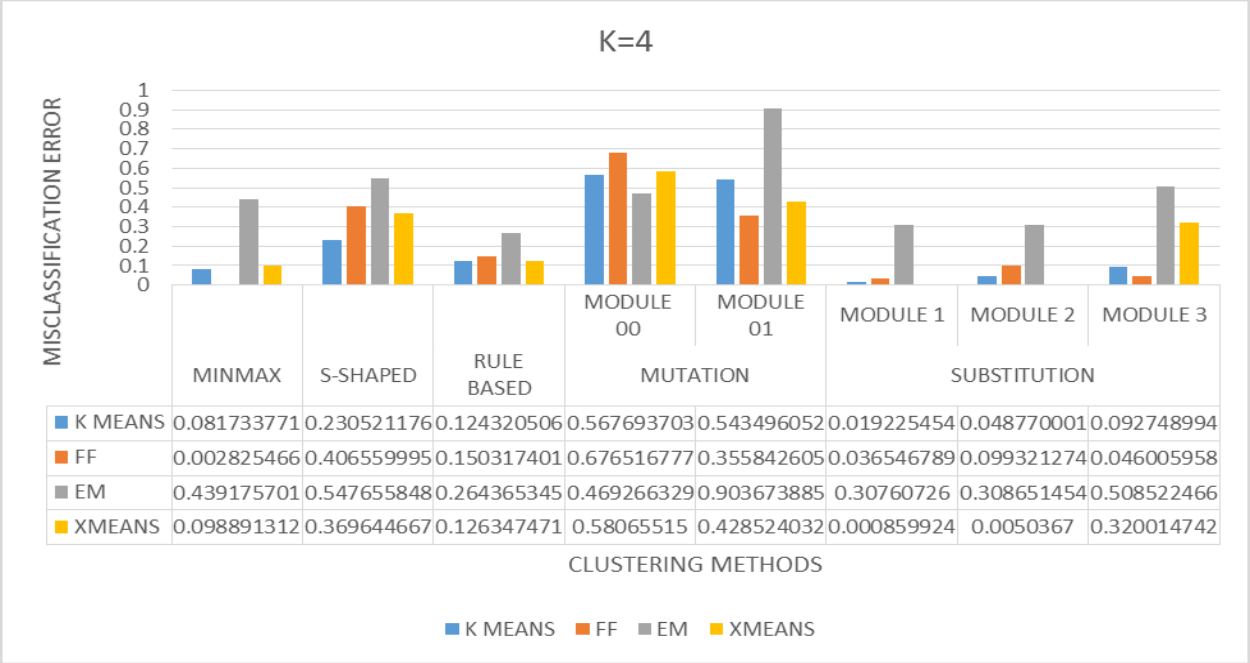

Fig. 3 – Misclassification error for clustering methods with 4 clusters.

In this paper, we have considered various attributes such as Age, Gender and Income from the adult data set available in the UCI repository. 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 i5 processor with 4GB RAM and Windows 8 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. The experimental outcomes are represented in the graphical form as shown in figure 1, 2 and 3.

CONCLUSION

One of the most challenging tasks in data mining is preserving privacy. Various methods have been used for this purpose. A method is said to be an efficient one if it has low misclassification error. In this paper we have computed misclassification error for different methods using different clustering algorithms. From the experimental results it is evident that Min-Max normalization is an efficient one in preserving privacy. Among the clustering algorithms k-means algorithm is the best choice for computing misclassification error since it generates an optimal value when compared with the other clustering algorithms.

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.

REFERENCES

[1] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kauffman Publishers, 2006.
[2] G.K.Gupta, Introduction to Data Mining with Case Studies, Prentice Hall of India, 2008.
[3] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu; Philip S. Yu, Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, Chapman and Hall, 2010
[4] B.Karthikeyan,G.Manikandan,Dr.V.Vaithiyanathan, "A Fuzzy Based Approach for Privacy Preserving Clustering", Journal of Theoretical and applied information Technology , Vol 32(2), 2011,118-122.
[5] G.Manikandan, N.Sairam, R.Sudhan,Vaishnavi, "Shearing Based Data Transformation Approach for Privacy Preserving Clustering", In Proceedings of 3rd IEEE International Conference on Computing, Communication and Networking Technologies, ICCCNT 2012
[6] G.Manikandan,N.Sairam,S.Sharmili,S.Venkatakrishnan , "Achieving Privacy in Data Mining Using Normalization" , Indian Journal of Science and Technology, Vol 6(4) , 2013,4268-4272.
[7] G.Manikandan,N.Sairam,S.Jayashree,C.Saranya , "Achieving Data Privacy in a Distributed Environment Using Geometrical Transformation", Middle East Journal Of Scientific Research , Vol 14(1) , 2013,107-111.
[8] G.Manikandan,N.Sairam,C.Akshaya,S.Venkatakrishnan ,"An Innovative Approach for Classifying Binary Data",International Journal of Applied Engineering Research, Vol 9(5) , 2014,589-597.
[9] G.Manikandan,N.Sairam,S.Rajarajeswari,H.Ramya,"A New Genetic Approach for Data Masking", International Journal of Applied Engineering Research, Vol 9(7) , 2014,755-761.
[10] Manikandan G, Sairam N, Rajendiran P, Balakrishnan R, Rajesh Kumar N, Raajan NR. Random Noise based Perturbation approach using Pseudo Random Number Generators for achieving Privacy in Data Mining. J Comput Theor Nanosci 2015; 12: 5463-5466.
[11] G Manikandan, N Sairam, M.SathyaPriya, Sree Radha Madhuri, V Harish, and Nooka Saikumar , "A Rule Based Approach for Ensuring Data Privacy in Data Mining " ,Journal of Engineering and Applied Sciences, Vol 11(13) , 8063 – 8066.
[12] Manikandan G, Sairam N, Harish V, Nooka S. A Substitution based Approach for Ensuring Medical Data Privacy. Res J Pharma Biol Chem Sci 2016; 7: 1136-1139
[13] Manikandan G, Sairam N, Harish V, Nooka S. Survey on the Use of Fuzzy Membership Functions to Ensure Data Privacy. Res J Pharma Biol Chem Sci 2016; 7: 344-348.
[14] UCI Data Repository http://archive.ics.uci.edu/ml/datasets.html