ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2019.0284

FUZZY BASED PRIVACY PRESERVED K-MEANS CLUSTERING

D. Murugan and S. Selva Rathna
Department of Computer Science, Manonmaniam Sundaranar University, India

Abstract

The aim of this paper is to identify the impact of the fuzzy based privacy preserving method in clustering which is one of the important data mining process. Fuzzy member ship functions like Bell shape, S- Shape and PI shape membership functions are applied on standard database to generate sanitised database. Further, various clustering algorithms are applied on the sanitised database and the results are compared. WEKA tool is used for testing K-Means clustering algorithm on privacy preserved database generated using various fuzzy member ship function. This analysis will help to develop new Fuzzy Based privacy preserving clustering techniques and also lead future researches in Privacy preserved data mining.

Keywords:

Clustering, Fuzzy Membership function, Privacy Preserving data mining, WEKA tool

1. INTRODUCTION

Privacy preserving data mining is an improved data mining which ensures the preservation of sensitive data while performing data mining process. If the entire data is present in a single site as in data warehouse model, privacy will not be a concern while performing data mining in such environment. Because there will be no leakage of data to the third party. However, in real time applications, the data will be distributed in multiple locations and hence, performing data mining along with privacy preservation of sensitive data will be a challenging one Privacy preserving data mining (PPDM) uses several approaches to address privacy concern while sharing the data to third parties. Few of the approaches are:

- An algorithm to modify or alter the data is applied on the data before delivering it to the third party who will run the data mining algorithm on the modified data.
- If the data is partitioned in different locations, then the data mining results are combined globally from the information derived from the individual sites. This will ensure that the data of individual sites are not revealed to other parties.

Cryptography is the most common technique used in PPDM for modifying data before releasing the sensitive data to the third party. The cryptographic techniques are chosen in such a way to ensure accuracy of the data and the performance of the mining process while maintaining the privacy constraints.

Clustering is one of the important data mining process in which the data are partitioned into a set of sub-classes, called clusters. The data objects that are similar are identified using Clustering analysis. The objective of clustering analysis is to identify the clusters which have low inter-cluster similarity and high intra-cluster similarity. Clustering is useful for grouping data and also for anomaly/outlier identification.

In this paper, fuzzy membership functions are applied on the sensitive data to modify them and clustering is applied on the modified data. The performance of the clustering techniques on various fuzzy member ship functions is analyzed. The suitable fuzzy member ship functions and its performance for Privacy preserved data mining are analyzed. It will lead to future researches in designing fuzzy based privacy preserved algorithms with high performance and more accuracy.

2. REVIEW OF RELATED WORKS

2.1 PRIVACY PRESERVING DATA MINING

PPDM techniques and frameworks were reviewed and comparison of advantages and disadvantages of different PPDM technique is made which helps to identify the open issues and future research trends in PPDM. The present scenario of Privacy preserving data mining and some future research directions are proposed in [1]. Privacy preserving data mining techniques presently available are classified in [2], and their advantages and disadvantages are emphasized. The application of cryptographic techniques on secure distributed computation in data mining is demonstrated in [3]. A classification hierarchy is proposed in [4] along with the detailed review.

2.2 FUZZY BASED PRIVACY PRESERVED DATA MINING

A comparison of fuzzy-based mapping techniques in terms of their property of privacy-preservation is presented in [5]. A method of extracting global fuzzy rules from distributed data in a privacy-preserving manner is proposed in [6]. Fuzzy c-regression method [7] is proposed to generate synthetic data which allows third parties to do statistical computations with a limited risk of disclosure of sensitive data. Pattern recognition with privacy preservation is studied in [8] by applying Fuzzy based k-member clustering. A secure framework for privacy preserving fuzzy co-clustering for handling both vertically and horizontally distributed data is presented in [9].

2.3 PRIVACY PRESERVED CLUSTERING

An overview of cluster analysis techniques from a data mining point of view is presented in [10]. Privacy preserving data clustering notably on partition-based and hierarchical methods is focused in [11]. It proposed methods which distort only confidential numerical attributes to meet privacy requirements, while preserving general features for clustering analysis. Most privacy preservation in clustering are developed for k-means clustering algorithm, by applying the secure multi party computation framework.

Clustering on vertical, horizontal, arbitrary partitioned dataset were proposed in [12]. The problem of protecting the underlying

attribute values when sharing data for clustering is addressed in [13]. The challenge is how to meet privacy requirements and guarantee valid clustering results as well. To achieve this dual goal, it proposes a novel spatial data transformation method called Rotation-Based Transformation (RBT). The major challenge of generalizing data for cluster analysis is the lack of class labels that could be used to guide the generalization process. A framework to evaluate the cluster quality on the generalized data is presented in [14]. Its main contribution is to present a general anonymization framework for properly preserving cluster structures and evaluating the resulting cluster solution.

3. PROPOSED FUZZY BASED PRIVACY PRESERVED CLUSTERING

3.1 FUZZY BASED PRIVACY PRESERVED DATA MINING

Several researches were found in privacy-preserving data mining to achieve to privacy preservation. In this paper, fuzzy based privacy preservation is considered. In privacy preservation using fuzzy membership function, the sensitive data is transformed into fuzzy data to achieve the secured sharing of data and the fuzzified data is shared with the party who is performing data mining task. This will ensure that the data mining process can be achieved on the fuzzified data in secured manner. Let D be the database owned by the owner. Fuzzy membership functions are applied on sensitive data to transform the data into fuzzy data D' in such a way that the third party or any other parties who needs to perform the data mining task can use D' instead of original data D. The architecture of the fuzzy based privacy preserving data mining is as shown in Fig 1.

3.2 FUZZY MEMBERSHIP FUNCTION FOR PRIVACY PRESERVED DATA MINING

Based on the study made in Section 2, Fuzzy logic membership functions are used to anonymizing the selective data of the database for maintaining privacy of the sensitive data. The privacy and information loss due to application of privacy preservation process has to be maintained zero or at least minimum. Also, the application of privacy preservation process should not affect the data mining process.

Membership function (MF) is curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The fuzziness of a fuzzy set is determined by its membership function. The shapes of the membership functions are important for a particular problem since they effect on fuzzy inference system. They may have different shapes like Triangular, Trapezoidal, Bell Shape, PI shape etc.

In this paper, the following Fuzzy membership functions are selected for analysis

- S-Membership Function
- Z-Membership Function
- PI-Membership Function
- Bell Membership Function

3.3 FUZZY MEMBERSHIP FUNCTION FOR PRIVACY PRESERVED CLUSTERING

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure.

Few clustering algorithms available are listed below:

- K-Means clustering
- EM clustering
- Farthest First clustering
- Filtered Clustering
- Hierarchical clustering
- COBWEB Clustering
- Make density based clustering

In this paper, K-Means clustering algorithms are applied on anonymized data generated from original data using fuzzy member ship functions and their performance is evaluated. This enables to identify the best fuzzy membership function which will be suitable for the clustering process which will further lead to developing new algorithms for privacy preserved clustering. The Fig.1 shows the architecture of Fuzzy based privacy preserved Clustering

3.4 K-MEANS CLUSTERING

The k-means clustering algorithm attempts to split a given anonymous data set (a set containing no information as to class identity) into a fixed number (k) of clusters. Initially k number of centroids are chosen. A centroid is a data point at the center of a cluster. This algorithm aims at minimizing an objective function knows as squared error function given by:

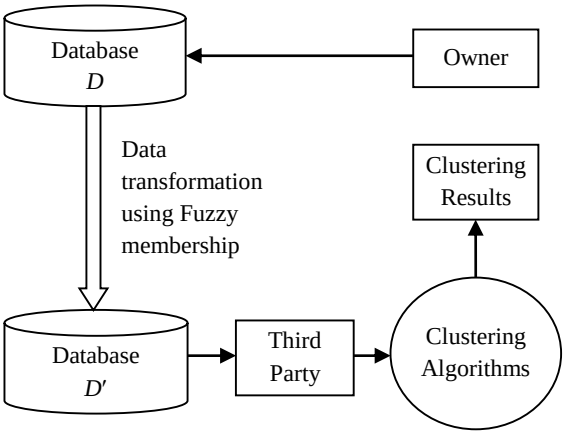

Fig.1. Architecture of Fuzzy based Privacy preserved

Clustering,

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}\left\|x_i - v_j\right\| \quad (1)$$

where,

$\|x_i\text{-}v_j\|$ is the Euclidean distance between x_i and v_j.

c_i is the number of data points in i^{th} cluster.

c is the number of cluster centers.

4. STEPS IN FUZZY BASED PRIVACY PRESERVED CLUSTERING

In this analysis, the original data set is transformed to fuzzified data using the fuzzy member ship function to generate sanitized data. The sanitized data is used for applying k-means clustering algorithm. Hence, the fuzzy based privacy preserved clustering involved two major steps as below.

Step 1: Apply Fuzzy membership algorithm on original data to generate sanitized data.

In this step, various fuzzy membership algorithm is applied on the original data. This will convert the original data with crisp value to fuzzy values. In this step, S-membership function, Z-membership function, PI-membership function, Bell membership function are considered for converting the original data into fuzzy value.

Step 2: K mean clustering is applied on Sanitized data

In this step, the original data which is converted to fuzzy data using the step 1, is taken for clustering. On the fuzzy data, K-Means clustering is applied for evaluating the performance of K-Means clustering.

5. RESULTS

K-Means clustering process with fuzzy membership functions is tested with IRIS and GLASS data set using WEKA tool. IRIS is a data set with 5 attributes of 150 items and glass is a data set with 10 attributes of 214 items.

Initially the test data is converted to fuzzified data set. Further, K-Means clustering algorithm is applied to fuzzified data set and the results are compared based on the within cluster mean error.

Table.1. Comparison of within Clustering Error

Fuzzy Membership Functions	Within Clustering Error	
	IRIS	GLASS
Original Data set	62.14	118.20
Bell Function	314	178.43
PI Function	284	161.60
S Function	275	150.50
Z Function	329	199.20

Within Clustering error is one of the parameter which is used for evaluating the clustering accuracy. It shows the difference in similarity of the point within the cluster. Hence, within clustering error is taken for consideration to evaluate the result.

The comparison of within clustering error for various member ship functions for K-Means clustering algorithm is tabulated in Table.1 and plotted in Fig.2.

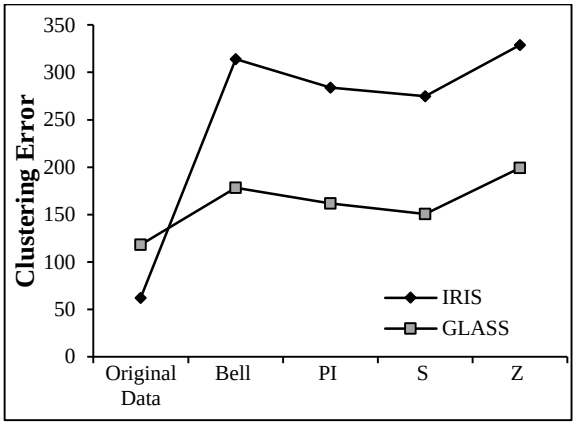

Fig.2. Comparison of Clustering Error using IRIS and GLASS dataset

The comparison of mean square error for various member ship functions for K-means clustering algorithm is tabulated in Table.2 and plotted in Fig.3.

Table.2. Comparison of Mean Square Error

Fuzzy Membership Functions	Mean square error	
	IRIS	GLASS
Original Data set	0.43	0.34
Bell Function	0.76	0.86
PI Function	0.83	0.65
S Function	0.64	0.59
Z Function	0.75	0.69

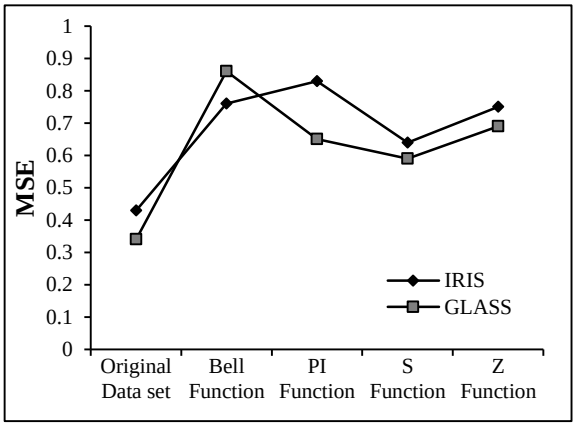

Fig.3. Comparison of Mean Square Error using IRIS and GLASS dataset

6. CONCLUSIONS

As per the result data, it is noticed that K-means clustering process shows less within cluster error while applying S-membership function for both IRIS and GLASS data set. Hence,

original data can be converted to fuzzified data using S-Membership function and further clustering shall be applied to achieve Fuzzy based Privacy Preserved K-Means Clustering. This analysis will give idea to develop new privacy preserved clustering algorithms with better performance with negotiated within cluster error.

REFERENCES

[1] M.B. Malik, M.A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", *Proceedings of 3rd International Conference on Computer and Communication Technology*, pp. 26-32, 2012.

[2] Y.A. Alsahib, S. Aldeen, M. Salleh and M. Razzaque, "A Comprehensive Review on Privacy Preserving Data Mining", SpringerPlus, Vol. 4, pp. 694-705, 2015.

[3] Benny Pinkas, "Cryptographic Techniques for Privacy-preserving Data Mining", *ACM SIGKDD Explorations Newsletter*, Vol. 4, No. 2, pp. 12-19, 2002.

[4] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin and Yannis Theodoridis, "State-of-the-Art in Privacy Preserving Data Mining", *ACM SIGMOD Record*, Vol. 33, No. 1, pp. 50-57, 2004.

[5] R. Mukkamala and V.G. Ashok, "Fuzzy-based Methods for Privacy-Preserving Data Mining", *Proceedings of 8th International Conference on Information Technology: New Generations*, pp. 348-353, 2011.

[6] J. Jiang and M. Umano, "Privacy Preserving Extraction of Fuzzy Rules from Distributed Data with Different Attributes", *Proceedings of Joint 7th International Conference on and Advanced Intelligent Systems and 15th International Symposium on Soft Computing and Intelligent Systems*, pp. 1180-1185, 2014.

[7] I. Cano and V. Torra, "Generation of Synthetic Data by Means of Fuzzy C-Regression", *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1145-1150, 2009.

[8] H. Kasugai, A. Kawano, K. Honda and A. Notsu, "A Study on Applicability of Fuzzy K-Member Clustering to Privacy Preserving Pattern Recognition", *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1-6, 2013.

[9] D. Tanaka, T. Oda, K. Honda and A. Notsu, "Privacy Preserving Fuzzy Co-Clustering with Distributed Cooccurrence Matrices", *Proceedings of International Conference on Soft Computing and Intelligent Systems*, pp. 700-705, 2014.

[10] J. Grabmeier and A. Rudolph, "Techniques of Cluster Algorithms in Data Mining", *Data Mining and Knowledge Discovery*, Vol. 6, pp. 303-360, 2002

[11] Stanley R.M. Oliveira and Osmar R. Zaiane, "Privacy Preserving Clustering by Data Transformation", *Journal of Information and Data Management*, Vol. 1, No. 1, pp. 53-56, 2010.

[12] M. Fathima and S.N. Bahloul, "Privacy Preserving K Mean Clustering-A Survey Research", *The International Arab Journal of Information Technology*, Vol. 9, No. 2, pp. 194-200, 2012.

[13] S.R.M. Oliveira and O.R. Zaiane, "Achieving Privacy Preservation When Sharing Data for Clustering", *Proceedings of International Workshop on Secure Data Management in a Connected World*, pp. 67-82, 2004.

[14] Benjamin C.M. Fung, Ke Wang, Lingyu Wang and Mourad Debbabi, "A Framework for Privacy-Preserving Cluster Analysis", *Proceedings of International Conference on Intelligence and Security Informatics*, pp. 17-20, 2008.

[15] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods", *Journal of the American Statistical Association*, Vol. 66, No. 336, pp. 846-850, 1971.

[16] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, pp. 1-38, 1977.