A FUZZY BASED APPROACH FOR PRIVACY PRESERVING CLUSTERING

Syed Md. Tarique Ahmad, Shameemul Haque, SM Faizanut Tauhid

Abstract— Extracting previously unknown patterns from massive volume of data is the main objective of any data mining algorithm. In current days there is a tremendous expansion in data collection due to the development in the field of information technology. The patterns revealed by data mining algorithm can be used in various domains like Image Analysis, Marketing and weather forecasting. As a side effect of the mining algorithm some sensitive information is also revealed. There is a need to preserve the privacy of individuals which can be achieved by using privacy preserving data mining. In this paper, fuzzy based data transformation methods are proposed for privacy preserving clustering in database environment. In case one, a fuzzy data transformation method is proposed and various experiments are conducted by varying the fuzzy membership functions such as Z-shaped fuzzy membership function, Triangular fuzzy membership function, Gaussian fuzzy membership function to transform the original dataset. In case two, a hybrid method is proposed as a combination of fuzzy data transformation approach specified in case one and Random Rotation Perturbation (RRP). The experimental results proved that the proposed hybrid method yields good results for all the member functions which are used in case one.

Index Terms— K-Means, S-Shaped Function, Privacy Preservation, Clustering, Fuzzy Membership Function, Random Rotation Perturbation, Data Transformation.

1 INTRODUCTION

DATA mining is the process used to analyze large quantities of data and gather useful information from them. It extracts the hidden information from large heterogeneous databases in many different dimensions and finally summarizes it into categories and relations of data [1]

In order to learn a system in detailed manner, we should be able to decrease the system complexity and increase our understanding about the system. For any application, if the information available is imprecise then fuzzy reasoning provides a better solution [2].

The primary goal of privacy preserving is to hide the sensitive data before it gets published. For example, a hospital may release patient's records to enable the researchers to study the characteristics of various diseases. The raw data contains some sensitive information of individuals, which are not published to protect individual privacy. However, using some other published attributes and some external data we can retrieve the personal identities. Table 1 shows a sample data published by a hospital after hiding sensitive attributes. (Ex. Patients name).

ID	Attributes			
	Age	Sex	Pin	Disease
1	23	M	110025	Cough
2	32	F	110065	Fever
3	28	M	110049	Diabetic
4	33	F	110054	Headache

Table 1: Rawdata

ID	Attributes			
	Name	Sex	Pin	Age
1	Raj	M	110025	23
2	Sona	F	110065	32
3	Salman	M	110049	28
4	Preeti	F	110054	33

Table 2: Voter Registration List

Table 2 shows a sample voter's registration list. If an opponent has access to this table he can easily identify the information about all the patients by comparing the two tables using the attributes like (zip-code, age, sex).These types of attributes are called as Quasi identifier attributes.

This idea of using fuzzy logic is applied to preserve the individual information while revealing the details in public. This paper mainly focuses on converting the sensitive data into modified data by using S – shaped fuzzy membership function. Kmeans clustering algorithm is applied on the modified data and it is found that the relativity of the data is also maintained.

There are a number of methods used for preserving the privacy of the data while clustering. Some of the methods are use of cryptographic algorithms, noise addition, and data swapping. All of these methods introduce a bit of complexity in the algorithm and increase the processing time. Our main aim is to reduce this processing time and at the same time provide an optimum solution to the problem of privacy preserving. For this purpose we are using the concept of fuzzy approach.

Fuzzy sets were introduced by zadeh in 1965 [2] to represent uncertainty, vagueness and provides formalized tools for dealing with the impression intrinsic to many problems. Fuzzy sets perform a gradual assessment of the input dataset by using fuzzy membership function. Fuzzy logic has been

- *Syed Md. Tarique Ahmad is currently pursuing Ph.d in Computer Science in Pacific University Udaipur Rajasthan, India, E-mail: tariquemca@gmail.com*
- *Shameemul Haque is currently working with Dept. of Computer Science, King Khalid University, Abha,Saudi Arabia, E-mail: shameem32123@gmail.com*
- *SM Faizanut Tauhid is currently working with Dept. of Computer Science, Pacific University Udaipur Rajasthan, India. E-mail: tauhidfaiz@gmail.com*

considered as an attractive method for data distortion which can reduce the information loss. In this paper two fuzzy data transformation methods are proposed for privacy preserving clustering in centralized database environment. In method one, a fuzzy data transformation method is proposed which uses fuzzy membership functions to transform the original dataset. In method two, a hybrid method is proposed as a combination of fuzzy data transformation approach specified in case one and Random Rotation Perturbation (RRP). Related works of privacy preserving clustering is discussed in the following section.

2 Literature Survey

In recent year's lot of papers are published to preserve data privacy while releasing the data for various research purposes which adopts various techniques like Data Auditing, Data Modification, Cryptographic methods and k-anonymity.

In Cryptographic methods [3] data is encrypted using protocols like secured multiparty computation (SMC). These protocols do not reveal any private information other than the final result to the data miners. In Noise addition methods [4] we add some random noise (number) to numerical attributes. This random number is usually drawn from a normal distribution with a small standard deviation and with zero mean. Data swapping [5] [6] interchange the attribute values between different records. Similar attribute values are interchanged with higher probability. The unique feature of this approach is all original values are kept back within the data set and only the positions are swapped.

In Aggregation [7] [8] instead of individual values the records are replaced by a group representative .For salary attribute, instead of individual values it can be grouped as {Low, Medium, High}. In Signal Transform methods [9] [10] Wavelet Transformation and Fourier Transformation are used to modify the data. These methods are fast when compared to its predecessors with improved time complexity.

In [11], the authors conducted privacy surveys about general privacy, consumer privacy, medical privacy and created privacy indexes to summarize the results and discussed the trends in privacy. Privacy issues in Hippocratic databases are discussed and identify the technical challenges, problems in designing such databases and suggested some approaches that may lead to solutions in [12]. Authors in [13] addressed the problem of protecting the underlying attribute values when sharing the data for clustering and proposed a novel spatial transformation method called rotation based transformation to achieve privacy. Hybrid data transformation approach for privacy preserving clustering is presented in [14], by adopting geometric transformation methods to modify the sensitive numerical data using translation data perturbation, scaling data perturbation, rotation data perturbation, reflective data perturbation in centralized database environment. Double reflecting data perturbation and rotation data perturbation based hybrid data transformation approach for privacy preserving clustering is proposed in [15]. Privacy preserving clustering approach through cluster bulging has been presented by authors in [16]. The authors in [17] proposed random response method of geometric transformation for privacy preserving clustering in centralized database environment. In [18], a fuzzy based approach is proposed for privacy preserving clustering. The authors used fuzzy membership function to transform the original dataset in order to preserve the privacy of individuals. In [19], random rotation perturbation approach and framework of random rotation perturbation for privacy preserving classification is proposed. The authors also presented a multi-column privacy model to address the problems of evaluating privacy quality for multidimensional perturbation.

Query auditing methods preserve privacy by modifying or restricting the results of a query. [20]. Sweeney [21] introduced the k1-anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least k-other records within the dataset, with respect to a set of quasi-identifier attributes. In this approach for achieving data anonymization methods like generalization and suppression are used. Unlike other privacy protection techniques such as data swapping and adding noise, information in a anonymous table through generalization and suppression remains ingenuous. Table3 shows an example of 2-anonymous generalization for Table1. While k-anonymity prevents identity disclosure, it does not ensure any protection against attribute disclosure. In [22] the clustering operation is performed after applying 2-dimensional transformations to the data.

A different approach for privacy preservation in data mining is given in [23]. This introduces the concept of fuzzy sets which is just an extension to the generic set theory. By using fuzzy sets we can perform a gradual assessment of the data set given to us and this is done by using a fuzzy membership function. Each linguistic term can be represented as a fuzzy set having its own membership function.

ID	Attributes			
	Age	Sex	Pin	Disease
1	2*	M	1100**	Cough
2	3*	*	1100**	Fever
3	2*	M	1100**	Diabetic
4	3*	*	1100**	Headache

Table 3: A 2-Anonymous Table

Fuzzy c-Means (FCM) can be used for clustering. But any element in the set may have membership in more than one category [24].

S – shaped fuzzy membership function is given by

$$f(x;a,b)=\begin{cases}0, & x\le a\\ 2\left(\frac{x-a}{b-a}\right)^2, & a\le x\le \frac{a+b}{2}\\ 1-2\left(\frac{x-b}{b-a}\right)^2, & \frac{a+b}{2}\le x\le b\\ 1, & x\ge b\end{cases}$$

Where x – is value of the sensitive attribute, a & b – is minimum and maximum value in the sensitive attribute.

3 PROPOSED METHODS

In this section two fuzzy based methods are proposed for privacy preserving clustering. In method one, a fuzzy based data transformation approach is proposed and various experiments are conducted by varying the fuzzy membership functions such as Z-shaped fuzzy membership function, Triangular membership function, Gaussian membership function to transform the original dataset. In method two, a hybrid method is proposed as a combination of fuzzy data transformation with various membership functions specified in case one and Random Rotation Perturbation (RRP). In another experiment novel additive perturbation approach is applied on original dataset to obtain the distorted dataset for comparison purpose.

3.1 Fuzzy based Data Transformation

Data distortion is the process of hiding sensitive data values without loss of information. A fuzzy transformation method distorts the sensitive numerical attributes using built in fuzzy membership functions such as Z-shaped fuzzy membership function (Zmf), Triangular fuzzy membership function (Trimf), Gaussian fuzzy membership function (Gaussmf).

3.2 Hybrid Method

A privacy preserving clustering technique is introduced in order to achieve the dual goal of privacy and utility. A hybrid method combines the strength of existing techniques and gives better results when compared to the single data perturbation method. This method consists of a combination of the two techniques namely fuzzy data perturbation and Random Rotation Perturbation (RRP). The important characteristic of RRP is preserving the geometric properties of the dataset. So the distorted dataset is clustered with similar accuracy when clustering is performed on original dataset. In this method, the original dataset is transformed using fuzzy data transformation method, which will be given as input for RRP method to obtain the final distorted dataset. The following table displays the algorithm for proposed hybrid method.

Input: (a) Original Dataset D of size m x n.
(b) Fuzzy membership functions such as Zmf, Trimf, and Gaussmf.
Output: Distorted datasets D′ of size m x n.
Begin
1. Suppress the identifier attributes.
2. For each Fuzzy membership function
3. For each sensitive attribute in D do
4. Transform the attribute using fuzzy membership function.
5. End For
6. Generate an n × n rotation matrix R randomly.
7. Obtain the final distorted dataset D′ = D × R.
8. End For
9. Release the distorted dataset D′ for clustering analysis.
End

TABLE 4: Algorithm for Hybrid Method

3.3 Novel Additive Perturbation Technique

A review of novel additive perturbation technique [25] is given in this section for privacy preserving data mining. This technique is used to modify the given input dataset in order to hide the highly sensitive information. The additive data perturbation technique is designed for distributed environment where a data owner wants to transform the input data extracted from group of parties. The transformed data is used to perform the data mining operations such as clustering, classification. The following algorithm in Table 5 explains the additive data perturbation.

1. Data owner acquire the input data from multiple parties by giving queries

2. Identify the sensitive data items and perform additive data perturbation on the selected values by adding small amount of noise to protect the values of sensitive data items.

3. To enhance the privacy protection of additive data perturbation, perform swapping on the perturbed dataset obtained in step 2.

4. Release the final distorted dataset to perform data mining operations like classification, clustering.

TABLE 5: Algorithm for Novel Additive Data Perturbation

4 IMPLEMENTATION OF PROPOSED METHODS

The proposed methods are implemented to evaluate the performance of data distortion methods. Experiments are conducted for implementing the proposed fuzzy data transformation approach and hybrid method on three real life datasets obtained from UCI [26]. Housing data set with thirteen attributes and 178 records, Iris data set with four numerical attributes and 150 records, Hayes-Roth dataset with 5 numerical attributes and 1000 instances are considered. Experiment one, a fuzzy data transformation method as given in Table 1 is conducted to transform the original dataset by varying the fuzzy membership functions such as Z-shaped fuzzy membership function, Triangular membership function, Gaussian membership function. Experiment two, a hybrid method as given in Table 2 is conducted by combining fuzzy data transformation approach specified in experiment one and Random Rotation Perturbation (RRP).The well-known k-means clustering algorithm is used to measure the clustering quality. When the data is transformed, the clusters in the original dataset should be equal to the clusters in the distorted dataset. WEKA (Waikato Environment for Knowledge Analysis) software [27] is used to test clustering accuracy of the original and modified data base. The effectiveness is measured by misclassification error [14]. The misclassification error, denoted by M_E, is measured as follows.

$$M_E = \frac{1}{N} + \sum_{i=1}^{K} (|\text{Cluster}_i(D)| - |\text{Cluster}_i(D')|)$$

In the above formula

N - Number of points in the original dataset.
K - Number of clusters.
Clusteri (D) - Number data points of the i^{th} cluster of the original data set.
Clusteri (D') - Number of data points of the i^{th} cluster of the transformed dataset.

The following table shows the ME values obtained for the proposed fuzzy data transformation method using Z-shaped fuzzy membership function (Zmf), Gaussian membership function (Gaussmf), Triangular membership function (Trimf) and novel additive method on three datasets. Higher ME values indicates lower clustering quality where as lower ME show the higher clustering quality. The experiments are conducted 10 times and ME value is taken as an average of 10. The following table displays the misclassification error values of fuzzy data transformation approach.

Data Distortion Methods	Housing	Iris	Hayes-Roth
Zmf	0.0892	0.09746	0.1242
Gaussmf	0.1785	0.1731	0.2164
Trimf	0.17	0.15	0.2223
Novel Additive	0.199	0.196	0.2432

Table 6: Misclassification Error Rates of Fuzzy Data Transformation

When comparing the misclassification error values in Table 6, it is confirmed that the proposed fuzzy data transformation methods which uses the three membership functions gives the lower misclassification error for all the three datasets. Among three membership functions Z-shaped fuzzy membership function gives the lower misclassification error.

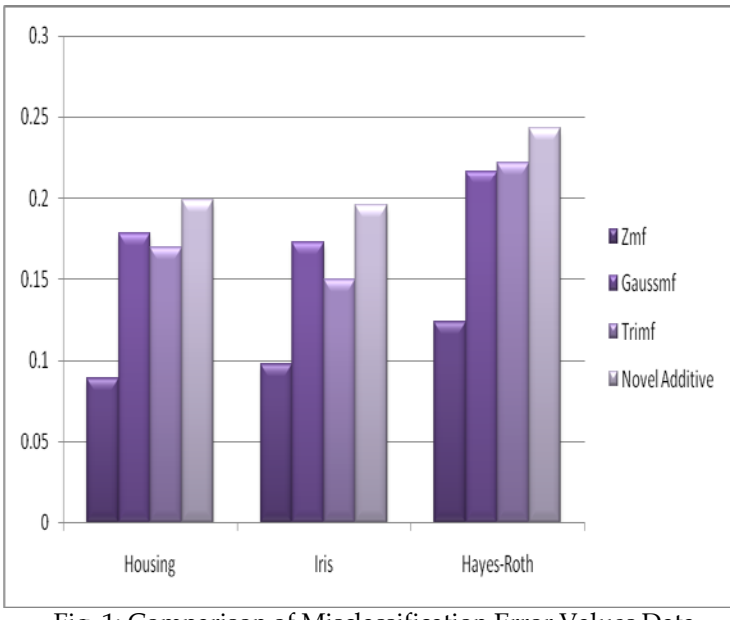

Fig. 1: Comparison of Misclassification Error Values Data Transformation Approaches

As the results shown in Fig. 1, it clearly indicates that for all the three datasets, the proposed fuzzy data transformation method gives the lower misclassification error for all the three member functions. These results proved that fuzzy data transformation methods gives higher utility than the novel additive data perturbation method. The following table shows the misclassification error values of the proposed hybrid method.

Data Distortion Methods	Housing	Iris	Hayes-Roth
Zmf	0.0892	0.09746	0.1242
Zmf & RRP	0.06853	0.08199	0.0951
Gaussmf	0.1785	0.1731	0.2164
Gaussmf& RRP	0.1612	0.1599	0.1765
Trimf	0.17	0.15	0.2223
Trimf & RRP	0.1477	0.1333	0.2
Novel Additive	0.199	0.196	0.2432

Table 7: Misclassification Error Rates Of Hybrid Method

Table 7 illustrates the misclassification error values calculated for the proposed methods and novel additive data perturbation method. When comparing the misclassification error values of the fuzzy data transformation method with hybrid method, it clearly indicates that the proposed hybrid method provides lower misclassification error values for all the three datasets. Hence the hybrid method provides better clustering quality and the transformed dataset that is generated with the hybrid method looks very different from the original dataset, which preserves the privacy of individuals.

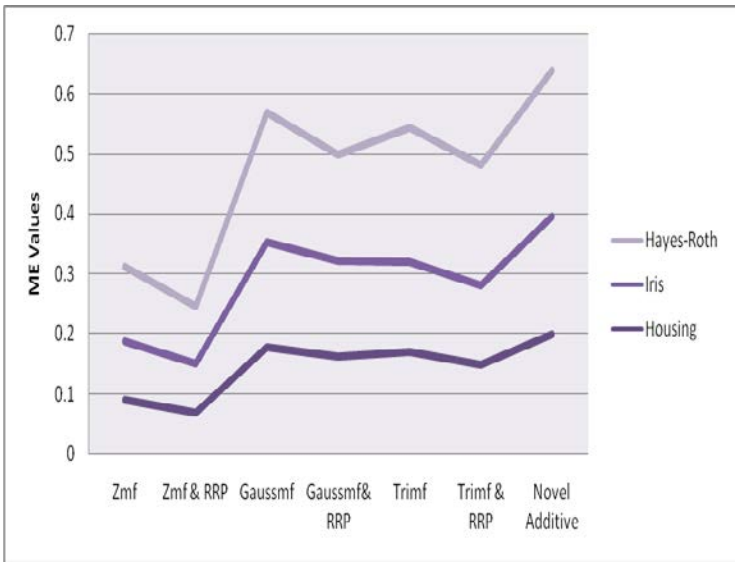

Fig. 2: Comparison of Misclassification Error Values of fuzzy Data Transformation Methods

The misclassification error values of fuzzy data transformation methods and hybrid method for three membership functions such as Z-shaped, triangular, Gaussian on three datasets are shown n Fig. 2. It clearly indicates that the proposed hybrid method gives lower misclassification error for all the three member functions and for all the three datasets.

4 Conclusion

Privacy places a vital role for organizations when the data consists of sensitive information and which is shared among different users. The problem protecting individual privacy while releasing the data for clustering analysis is considered in this paper. Random rotation is one of the popular approaches for data perturbation and it can preserve privacy without affecting the accuracy for clustering analysis. Two methods are proposed in order to address this problem. Method one is a fuzzy based transformation approach that uses Z-shaped fuzzy membership function, Triangular membership function and Gaussian membership functions for data transformation. Experiments are conducted on three real life datasets from UCI and the results proved that the proposed method satisfying the privacy constraints as well as retains the clustering quality. To enhance the privacy preservation, method two which is a hybrid method is proposed by adopting the techniques fuzzy data transformation approach as specified in method one and random rotation perturbation. Experiments on three real life datasets reveal that, hybrid method is efficient for data utilization as well as privacy preservation.

References

[1] Sairam et al "Performance Analysis of Clustering Algorithms in Detecting outliers", International Journal of Computer Science and Information Technologies, Vol. 2 (1) , Jan-Feb 2011, 486-488.

[2] Zadeh L "Fuzzy sets", Inf. Control. Vol.8, PP, 338 – 353, 1965.

[3] Pinkas,."Cryptographic Techniques for Privacy- Preserving Data Mining", ACM SIGKDD Explorations, 4(2), 2002.

[4] Agrawal D, Aggarwal C.C, "On the Design and Quantification of Privacy Preserving Data mining algorithms", ACM PODS Conference, 2002.

[5] Fienberg S.E. and McIntyre J. "Data Swapping:Variations on a theme by Dalenius and Reiss." In Journal of Official Statistics, 21:309-323,2005.

[6] Muralidhar K. and Sarathy R. " Data Shufflinga new masking approach for numerical data", Management Science, forthcom-ing, 2006.

[7] Y.Li,S.Zhu,L.Wang, and S.Jajodia " A privacyenhanced micro-aggregation method", In Poc. Of 2nd International Symposium on Foundations of Information and Knowledge Systems, pp148-159, 2002.

[8] V.S. Iyengar, "Transforming data to satisfy privacy constraints", In Proc. of SIGKDD'02, Edmonton, Alberta, Canada,2002.

[9] Shuting Xu.,Shuhua Lai, "Fast Fourier Transform based data perturbation method for privacy protection", In Proc. of IEEE conference on Intelligence and Security Informatics, New Brunswick New Jersey, May 2007.

[10] Shibanth Mukharjee, Zhiyuan Chen, Arya Gangopadhyay,"A privacy preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms", The VLDB journal 2006.

[11] A.F.Westin, Freebies and privacy: what net users think Technical report, Opinion Research Corporation, July1999 Available from http://wwwprivacyexchange.org/iss/surveys/sr990714.htm?

[12] R. Agrawal, J. Kiernan, R. Srikant and Y Xu, "Hippocratic Databases", Proc of the 28th Int'l Conf. on Very Large Databases (VLDB'02), 2002, pp. 143-154.

[13] S. R. M. Oliveira, and O. R.Zaiyane, "Achieving Privacy Preservation When Sharing Data for Clustering", In Proceedings of the Workshop on Secure Data Management in a Connected World, in conjunction with VLDB'2004.Toronto, Ontario, Canada, pp. 67–82, 2004a.

[14] S.R. M. Oliveira, O.R. Zaiyane (2003), "Privacy Preserving Clustering by Data Transformation", in proceedings of 18th Brazilian Conference on Databases.

[15] Liming Li, Qishan Zhang, "A Privacy preserving Clustering Technique Using Hybrid Data Transformation Method", In proceedings of 2009IEEE international conference of grey systems and intelligent services.

[16] Mohammad Ali Kadampur, D.V.L.N Somayajulu, S.S. Shivaji Dhiraj and Shailesh G.P. Satyam, "Privacy preserving clustering by cluster bulging for information sustenance", of the 4th International Conference on Information and Automation for Sustainability(ICIAfS,,08), Colombo, Sri Lanka, December 2008.

[17] Jie Liu,Yifeng XU,Harbin, "privacy preserving clustering by Random Response Method of Geometric Transformation", In proceedings of Fourth international conference on internet computing for science and engineering 2009.

[18] B. Karthikeyan, G. Manikandan, V.Vaithiyanathan, A Fuzzy based approach for Privacy Preserving Clustering, Journal of Theoretical and Applied Information Technology, 31st October 2011. Vol. 32 No.2.

[19] K.Chen, and L.Liu, A Random Rotation Perturbation Approach to Privacy Data Classification, In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp.589-592, 2005.

[20] Nabar S. Marthi B, Kenthapadi K, Mishra N, Motwani R., "Towards Robustness in Querry Auditing" VLDB Confer-ence, 2006.

[21] L. Sweeney, "^-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledgebased Systems, 2002, pp. 557-570.

[22] R.R.Rajalaxmi, A.M.Natarajan "An Effective Data Transformation Approach for Privacy Preserving Clustering", Journal of Computer Science 4(4): 320-326, 2008.

[23] V.Vallikumari, S.Srinivasa Rao, KVSVN Raju, KV Ramana, BVS Avadhani "Fuzzy based approach for privacy preserving publication of data", IJCSNS, Vol.8 No.1, January 2008.

[24] Timothy J. Ross "Fuzzy Logic with Engineering Applications", McGraw Hill International Editions, 1997.

[25] P. Kamakshi, A.Vinaya Babu, " A Novel Framework to Improve the Quality of Additive Perturbation Technique", In proceeding of International Journal of Computer Applications, Volume 30, No. 6, September 2011.

[26] UCI Machine Learning Repository http://archive .ics.uci.edu/ml/datasets.html.

[27] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.