RESEARCH ARTICLE OPEN ACCESS

Information Confidentiality Using Fuzzy Based Data Transformation Method

[1]Mrs. TamilSelvi. S. [2]Mrs. Anette Regina I.

[1]M.phil Research Scholar, Department of computer Science Muthurangam Government Arts College (Autonomous), Vellore, TamilNadu, India.
[2]Associate Professor, Department of Computer Science Muthurangam Government Arts College (Autonomous), Vellore.

----------**********----------

Abstract:

The freedom and transparency of information flow on the Internet has heightened concerns of privacy. Given a set of data items, clustering algorithms group similar items together Knowledge extraction process poses certain problems like accessing sensitive, personal or business information. Privacy invasion occurs owing to the abuse of personal information. Hence privacy issues are challenging concern of the data miners. Privacy preservation is a complex task as it ensures the privacy of individuals without losing the accuracy of data mining results. In this paper, fuzzy based data transformation methods are proposed for privacy preserving clustering in centralized 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.

Keywords — **Clustering, Fuzzy mean Clustering, Random Rotation Perturbation.**

----------**********----------

I. INTRODUCTION

Data storage is growing at a phenomenal rate. Data users are interested to extract useful information from these large amounts of data. Data mining is becoming an increasingly important tool that can transform the data into useful knowledge. Data mining tasks provide accurate information for decision making. Some of the data mining tasks are classification, prediction, association rules and clustering. Clustering is a well known problem in statistics and engineering and widely used in various applications including biology, medicine, marketing etc. Clustering is the process of grouping the set of items by finding similarities between the data according to characteristics found in the data. There are two types of privacy concerns when mining the databases. Firstly, disclosure of personal or sensitive data existed in the databases. Secondly, sensitive patterns should not be disclosed when mining is performed on the shared data. New technologies are required for protecting sensitive information in electronic commerce [1]. Privacy preserving data mining is a new research area, which allows mining useful information while preserving privacy of individuals. Privacy issues are considering in two situations. They are centralized database environment and distributed environment. In centralized environment, database is available in single location. In this environment, privacy preserving data mining techniques are used to hide sensitive data of individuals. In distributed database environment, data is distributed to multiple sites. In this environment privacy preserving data mining techniques are applied for integrating the data from multiple sites, without revealing the privacy of individuals.

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 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

Environment. Double reflecting data perturbation and rotation data perturbation based hybrid data transformation approach for privacy preserving clustering is proposed in [7]. Privacy preserving clustering approach

through cluster bulging has been presented by authors in [8]. The authors in [9] proposed random response method of geometric transformation for privacy preserving clustering in centralized database environment. In [10], 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 [11], 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.

II. RELATED WORK

In [3], 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 [4]. Authors in [5] 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 [6], 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

As already pointed out in the introduction, our paper considers clustering for horizontally-partitioned data. Vaidya and Clifton's algorithm is based on the secure-permutation algorithm of Du and Atallah [13]. However, Vaidya and Clifton's algorithm has to execute Du and Atallah's protocol for every item in the data set. Therefore, their algorithm is not practical for large data sets. Moreover, Vaidya and Clifton did not perform an experimental evaluation of their algorithm. By contrast, the complexity of our algorithm only depends on the number of steps taken by the K *m*eans algorithm and the dimension of the data items. There are distributed clustering algorithms where the goal is to reduce communication costs [12, 3]. These distributed clustering algorithms do not onsider privacy. However, it will be interesting to investigate whether these algorithms can be made privacy preserving.

III. PREVIOUS IMPLEMENTATIONS

Evaluated three clustering algorithms. The simple scheme is used throughout as a baseline for our experiments. This protocol implements the k-means clustering algorithm as described in section 3. This algorithm does not use any privacy-preserving protocols. This represents the nominal cost of clustering, and will be present in any k means clustering approach, independent of if and how privacy is implemented. Throughout this section features refer to the dimension of the vectors being clustered and each iteration of the k-means algorithm is referred to as round. Our first privacy-preserving protocol (referred to as OPE) uses oblivious polynomial evaluation. A congestion window manager that continually tracks the network congestion status of the multiple paths that have been setup for Access data transport. A real-time scheduler that schedules packets over the multiple paths based on the inputs from the congestion window manager. Similarly, the receiver must be equipped with the ability to aid the sender by informing it of Packets that are arriving late on particular paths Packets that have not shown up at all within a reasonable time limit [2].

- Clustering using DPE is two orders of magnitude more bandwidth efficient than OPE, and executes in 4.5 to 5 times less time. This is largely due to bandwidth and computational costs associated with the oblivious transfers used by OPE.
- Our protocols clustering with perfect fidelity; that is, the clusters resulting from our algorithms are identical to those reported by a k-means algorithm with no privacy for reasonable parameter choices.
- Small, medium, and large data-sets can be clustered efficiently.

- Costs scale linearly with feature and rounds. The number of samples affects runtime only inasmuch as it increases the number of rounds toward convergence.

- Protocol parameters affect bandwidth usage by constant factor. Moreover, exponential increases in security or supported message space result in linear increases in execution run-times.

IV. SYSTEM IMPLEMETNATION

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.

A. 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). Table 1 shows the algorithm for fuzzy based data transformation method. The input to the fuzzy transformation method is a dataset D consists of sensitive attribute data in m rows and n columns. The input data D is transformed by initially suppressing the identifier attributes and distort the dataset using fuzzy membership function

Algorithm1: Algorithm For Fuzzy Based Data Transformation Method

Input: (a) Dataset D consists of sensitive attribute data in m rows and n columns.

(b) Fuzzy membership functions such as Zmf, Trimf, and Gaussmf.

Output: Distorted Datasets and each D' consist of m rows and n columns.

Begin

(1) Suppress the identifier attributes

(2) For each membership function (Zmf, Trimf, Gaussmf)

(3) For each sensitive attribute in D do

(4) Transform the attribute using selected fuzzy membership

function.

(5) End For

(6) End For

(7) Release the all distorted datasets for clustering analysis.

End

B. 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 described in the table 1, 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.

Algorithm 2:

Input: (a) Original Dataset D consists of sensitive attribute data of size m x n.

(b) Fuzzy membership functions such as Zmf, Trimf, and Gaussmf.

Output: Distorted datasets and each D" consist of size m x n.

Begin

(1) Suppress the identifier attributes.

(2) For each membership function (Zmf, Trimf, Gaussmf)

(3) For each sensitive attribute in D do

(4) Transform the attribute using fuzzy membership function.

(5) End For

(6) Generate an n x n rotation matrix R randomly.

(7) Obtain the final distorted dataset D" = D x R.

(8) End For

C. Novel Additive Perturbation Technique

A review of novel additive perturbation technique [12] 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.

- Data owner acquire the input data from multiple parties by giving queries
- 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.
- To enhance the privacy protection of additive data perturbation, perform swapping on the perturbed dataset obtained in step 2.
- Release the final distorted dataset to perform data mining operations like classification, clustering.

Table: Data Distortion Method

Data Distortion Methods	Iris	Wine	Credit-g
Zmf	0.09746	0.0892	0.1242
Trimf	0.15	0.17	0.2223
Gaussmf	0.1731	0.1785	0.2164
Novel Additive	0.196	0.199	0.2432

Implementation

Precision is used to measure the degree to which the approximated clustering diverge from those reported by a simple *k*-means algorithm, and is calculated as follows.

Let $X = \{x1, \ldots, xn\}$ be the sample data set to be clustered. $C1$ $2X$ is the clustering of X by the simple algorithm, and $C2$ $2X$ is the clustering returned by the OPE algorithm (the DPE metric is defined similarly in the obvious manner). For each pair (xi, xj) such that $1 \leq i < j \leq n$ an error occurs if

1. xi and xj are in the same cluster in $C1$, but in $C2$ they are in different clusters.
2. xi and xj in the same cluster in $C2$, but in $C1$ they are in different clusters.

The total number of errors is denoted E. The maximum number of errors is $N = n(n-1)/2$. The precision P is given by $(N - E)/N$.

The size of the message space in DPE and the finite-field in OPE are chosen to achieve the desired precision. In Benaloh's encryption scheme r denotes the size of the message space. For efficiency reasons we choose $r = 3k$ (see [5] for details). Two crucial parameters in the oblivious polynomial evaluation protocol of Naor and Pinkas are D, the degree of the masking polynomial and M, the total number of points used (details of this algorithm can be found in [40]). The sender's masking polynomial D has degree $k.d$, where d is the degree of the polynomial P being evaluated and k is the security parameter. Since in our algorithm the polynomial being evaluated is always linear, the security parameter is simply D. Increasing D strengthens the sender's security. Only D+1 points are needed to interpolate, but the receiver sends $(D+1).M$ pairs of values to the sender. Out of each set of M pairs, one of them is related to α (the point the polynomial is being evaluated on), and the other $M - 1$ values are random. The 1-out of M oblivious transfer protocol (denoted as $OT\ M\ 1$) is repeated D+1 times to learn the required value. So, increasing M strengthens the receiver's security. Unless otherwise specified, we selected $D = 7$ and $M = 6$. For brevity, we do not consider D or M further.

Description of the Proposed Algorithm:

Aim of the proposed algorithm is to get approximate equal accuracy of modified dataset as per original dataset. The algorithm is given below. For each attribute of G(X), let R be random rotation, X be a original dataset, T be a translation and D be a Gaussian noise then the value of attribute G(X) is calculated using following formula.

$$G(X) = R*X + T + \Delta$$

Procedure: Geometric Transformation Based Multiplicative Data Perturbation.

Input: Data Stream D, Sensitive attribute S. Intermediate Result: Transformed data stream D'.

Output: Clustering results R and R' of Data stream D and D' respectively.

- Given input data D ith tuples size n, extract sensitive attribute $[S]_{n\times1}$.
- Rotate $[S]_{n\times1}$ into 180o clock-wise direction and generate $[R_n]_{n\times1}$.
- Multiply elements of [S] with [Rs] , transformed sensitive attribute values will be $[X]_{s1} = [S]_{n\times1} \times [R_s]_{s\times1}$.
- Calculate translation T as mean of sensitive attribute $[S]_{n\times1}$.
- Generate transformation $[St]_{n\times1}$ by applying translation T to $[S]_{n\ x\ 1}$.
- Crate transformed dataset D' by replacing sensitive attribute $[S]_{n\times1}$ in original dataset D with $[Gs]_{n\ x\ 1}$.

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. 104th Congress. *Public Law 104-191: Health Insurance Portability and Accountability Act of 1996*, August 1996.
2. N.R. Adam and J.C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM computing Surveys*, 21, 1989.
3. R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, TX, May 2000.
4. P.S. Bardley and U.M. Fayyad. Refining initial points for k-means clustering. In *Proceedings of 15th International*

Conference on Machine Learning (ICML), pages 91–99, 1998.
5. J. Benaloh. Dense probabilistic encryption. In *Workshop on Selected Areas of Cryptography*, pages 120–128, May 1994.
6. D. Boneh and M. K. Franklin. Efficient generation of shared RSA keys. *Journal of the ACM (JACM)*, 48(4):702–722, 2001.
7. R. Canetti. Security and composition of multi-party cryptographic protocols. *Journal of Cryptology*, 13(1):143–202, 2000.
8. Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. *The Platform for Privacy Preferences 1.0 (P3P1.0) Specification*. W3C Recommendation,
16 April 2002.
9. Lorrie Faith Cranor. Internet privacy. *Communications of the ACM*, 42(2):28–38, 1999.
10. D.E. Denning. A security model for the statistical database problem. *ACM Transactions on Database Systems (TODS)*, 5, 1980.
11. I.S. Dhillon, E.M. Marcotte, and U. Roshan. Diametrical clustering for identifying anticorrelated gene clusters. *Bioinformatics*, 19(13):1612–1619, 2003.
12. I.S. Dhillon and D.S. Modha. A data-clustering algorithm on distributed memory multiprocessors. In *Proceedings of Large-scale Parallel KDD Systems Workshop (ACM SIGKDD)*, August 15-18 1999.
13. W. Du and M. J. Atallah. Privacy-preserving cooperative statistical analysis. In *Annual Computer Security Applications Conference ACSAC)*, pages 102–110, New Orleans, Louisiana, USA, December 10-14 2001.
14. R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, 2001.
15. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–228, Edmonton, Alberta, Canada, July 23–26 2002.
16. J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M. Strauss, and R. N. Wright. Secure multiparty computation of approximations. In *28th International Colloquium Automata, Languages and Programming (ICALP 2001)*, Crete, Greece, July 8-12 2001.
17. Niv Gilboa. Two party rsa key generation. In *Advances in Cryptology (CRYPTO '99)*, Santa Barbara, California, USA, August 15-19 1999.
18. Ian Goldberg, David Wagner, and Eric Brewer. Privacy-enhancing technologies for the internet. In *Proc. of 42nd IEEE Spring COMPCON*. IEEE Computer Society Press, February 1997.
19. O. Goldreich. *Foundations of Cryptography: Volume 1, Basic Tools*. Cambridge University Press, May 2001.
20. O. Goldreich. *Foundations of Cryptography: Volume 2, Basic Applications*. Cambridge University Press, 2004.

BIOGRAPHIES

1. Ms. Tamilselvi ., M.phil Research Scholar, Department of computer Science Muthurangam Government Arts College(Autonomous), Vellore, TamilNadu, India.
2. Mr. Saravanan A. M., Assist Prof Department of Computer Science Muthurangam Government Arts College(Autonomous), Vellore, TamilNadu, India.