Indian Journal of Science and Technology, *Vol 7(7), 975-980, July 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Anonymization by Data Relocation using Sub-clustering for Privacy Preserving Data Mining

V. Rajalakshmi[1*] and G. S. Anandha Mala[2]

[1]Sathyabama University, Chennai, Tamil Nadu, India; rajalakshmi.bala03@gmail.com
[2]Department of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai, Tamil Nadu, India; gs.anandhamala@gmail.com

Abstract

As there are new techniques growing to reveal the hidden information on data, the threat towards those data also increases. Therefore, privacy preservation in data mining is an emerging research area which develops various algorithms to anonymize the data provided for data mining. The existing methodology handles the tradeoff between utility and privacy of data in a more expensive way in terms of execution time. In this paper, a simple Anonymization technique using sub-clustering is specified which achieves maximum privacy and also utility with minimum execution time. The methodology is explained with algorithm and the results are compared with the baseline method.

Keywords: Anonymization, Clustering, Isometric Transformation, Privacy Preservation

1. Introduction

Data are values of qualitative or quantitative variables, belonging to a set of items. In recent years, advances in hardware technology have made an increase in the capability to store and record personal data about consumers and individuals. This has lead to concerns that the personal data may be misused for a various purposes. Data explains a business transaction, a medical record, bank details, educational details etc., Use of technology for data storage and processing has seen an unexpected growth in the last few decades. Such information includes personal details, which the owner doesn't like to disclose. Such data are the input and sources for data mining. Data mining gives us "facts" that are not obviously seen to human analysts of the data. When such private data are given directly for mining, the security and the privacy of the individual is highly affected. So the data are modified and provided for data mining. But the problem is that the modified data should also produce a similar mining result[18, 31]. This has lead to a special research area called privacy preservation in data mining which is an intersection of both data mining and information security. The fact in this area is the additional anonymization task which is used to implement the privacy that degrades the performance of the data mining algorithm, which results in incorrect mining results.

PPDM techniques can be classified into two types[5]. (1) Perturbation methods[13] – which alter the data by generalization[17], suppression, additive or multiplicative factor, fuzzy based, or geometric projections and random number projections. (2) Cryptography based method – they use a public or private key to hide the data and reconstructed when required. Perturbation methods are mainly used with a little compromise on data utility, as the data are altered and or not reversible. Privacy is provided to an extent except closeness attack. For some applications where the data should not be altered at all

*Author for correspondence

does not encourage privacy as it becomes a complete failure. Example, Disease spread in a zip code in the past few years. In this case even if the zip code value is generalized or altered even by single value the result will be a wrong one. Cryptographic methods use a single key value which is vulnerable to privacy threats and the similarity attack is possible in all these methods. The method is also considered to be costly as it is to be applied on each data.

PPDM algorithms can also be implemented for data streams[14]. But in this paper only a static data base is considered.

2. Related Work

Sweeney et al.[28], has started with privacy preservation using k-anonymity. Agarwal[6,23] came up with the technique of perturbation using randomization methods. Sweeney[22] again introduced a methodology of using generalization hierarchy for implementing k-anonymity. A new perturbation technique has been suggested[32] using tree concept. k-anonymity[25] is used for Privacy preservation in data mining using micro-aggregation. Oliviera et al.[27,30] innovated that clustering can be used to group the data for perturbation. The authors[2,3,7–9,15,16,24,26] also concentrates on the success of using clustering technique for implementing k-anonymity or other perturbation techniques. The author[1] specifies various types of clustering with their applications. Other than clustering grouping of data can be done by nearest neighbour[10], decision tree[29] or bucketization[35].

Oliveira have also identified that isometric transformation based rotation can be used for perturbation. There are various advantages for this method, as it maintains the statistical parameters like centroid, variance etc., and also best forwards the correlation between the attributes. The only disadvantage of this method being it reversible and allows similarity attack. K-anonymity has homogeneity problem and hence improvised as l-diversity[4,11,19] (L, α)-diversity[33] and t-closeness[21]. Isometric transformation[9,26,27] is done on clusters, which exhibits the advantages of both of the techniques.

PPDM[34] has been implemented using multi level trust, which combines access control and Anonymization.

A direct rotation on clusters is susceptible to similarity and skewness attacks. Hence to avoid this problem, in our methodology, the clusters are sub-clustered and random angles are generated for each object in an equivalence class.

3. Problem Definition

A novel methodology that anonymize the data without affecting the statistical values and data mining results, and is not susceptible to homogeneity, similarity attack and skewness attack is required. A control towards the Anonymization without compromising on privacy and utility of the data should be provided by a procedure, with less computational time.

4. Data Relocation based on Sub-clustering [DRBS]

An Anonymization technique which maintains the similarity of individual data and the correlation among the data can be implemented by using Isometric transformation. According to the flow diagram provided in Figure 1, the data are grouped using clustering. FCM based clustering is used since it is more efficient compared to k-means algorithm. The problem of homogeneity is solved by sub-clustering each of the clusters and equivalence classes are identified. The sub-clusters are then arranged sequentially based on their positions and Euclidean distance between their centroids. Each record in an equivalence class is anonymized differently using controlled relocation.

In this method, the quality of the data is maintained and the mis-classification error is less. There are some basic terms required to be explored without going in detail about the algorithm. Table 1 specifies the various notations used in the manuscript.

Definition 4.1: Quasi Identifiers - A set of non-sensitive attributes $\{a_1, \ldots, a_m\}$ of a table is called a quasi-identifier if these attributes can be connected with external data to uniquely identify at least one individual record in the whole database.

Definition 4.2: Privacy – A database is privacy preserved if there is minimum probability of associating any single record with its sensitive attribute.

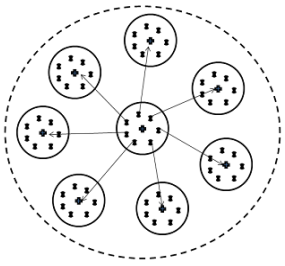

Figure 1. Operation of DRBS.

Table 1. Notations used

Variables used	Explanation
n	Number of records
m	Number of attributes
K	Number of Clusters
C	Centroid
$r_1,r_2,\ldots,r_i\ldots r_n$	Individual record
$a_1,a_2,\ldots,a_j\ldots a_m$	Individual attributes
1…k…K	Individual cluster
$c_1,c_2,\ldots,c_k\ldots c_K$	Individual centroid
$X_1,X_2,\ldots X_k\ldots X_K$	Number of sub-clusters in each Cluster
$S_1,k\ldots S_x,k\ldots S_{Xk,K}$	Individual sub-cluster in cluster k

Definition 4.3: Isometric Rotation - Let T be a transformation in the n-dimensional space, i.e.,

$$F : T_n \rightarrow T_n$$

is said to be an isometric transformation if it preserves distances between any two points p,q satisfying the following constraint: $|F(p)- F(q)| = |p-q|$ for all $p, q \in T_n$.

Definition 4.4: Centroid – It is defined for an attribute as a mean or average of all the transaction values. All the data are concentrated around this point. After Anonymization the centroid of a cluster is expected to be stable. The centroid of cluster k for attribute j is given by,

$$c_{k,j} = \frac{1}{n}\sum_{i=1}^{i=n} r_{i,j} \tag{1}$$

4.1 Algorithm of DRBS

Input: Original data T
Output: Anonymized data T'
Method:

Step 1: T is clustered into K clusters using FCM algorithm

Step 2: Each cluster is grouped into X_k number of sub-clusters using FCM algorithm

Step 3: The sub-clusters are arranged sequentially using Euclidean distance between their centroids.

Step 4: For each cluster k,

Step 4a: For each Sub-cluster $S_{x,k}$,

(i) Determine the number of records in $S_{x,k}$ as Y.
(ii) Identify Y number of adjacent sub-clusters
(iii) Map each of the records with the centroids of selected sub-clusters.
(iv) Perform isometric transformation with respect to the selected centroids.

4.2 Experimental Setup

The Adult dataset in UCI data repository is used, which has 30,162 records after pre-processing. Seven attributes form the dataset are chosen. They are age, marital-status, race, sex, hours-per-week, native-country and Salary. In this "Salary" is selected as sensitive attribute and the remaining are quasi-identifying attributes. MATLAB is used to cluster and compare the results with base-line method chosen as a direct rotation using isometric rotation.

4.3 Performance Metrics

The Anonymization alters the data which affects the usability of the data. Algorithms can be measured in terms of some defined metrics[12] as follows:

4.3.1 Information Distortion

Information distortion[20] is a measure that can be calculated from the difference between the original table and the anonymized table. It can also be defined as the distribution of data with respect to the centroid. The information distortion of each cluster can be calculated separately and their sum is calculated using the following equations. The dissimilarity of record i in jth attribute with respect to centroid c_k is given by,

$$\text{diss}\,(r_{i,j}, c_{k,j}) = [r_{i,j} - c_{k,j}]^2 \tag{2}$$

The distortion of all records is given by

$$D = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} \tag{3}$$

where, u_{ik} specifies the membership of ith record in kth cluster.

$$\sum_{k=1}^{K} \tag{4}$$

$$u_{ik} \in \{0,1\} \tag{5}$$

Figure 2 shows the variation of information distortion with respect to the number of sub-clusters. As the number of sub-cluster increases, the size of the sub-clusters and the distance between the sub-cluster decreases. Hence, the increase in number of sub-clusters decreases the amount of movement of data and hence also decreases the amount of information distortion.

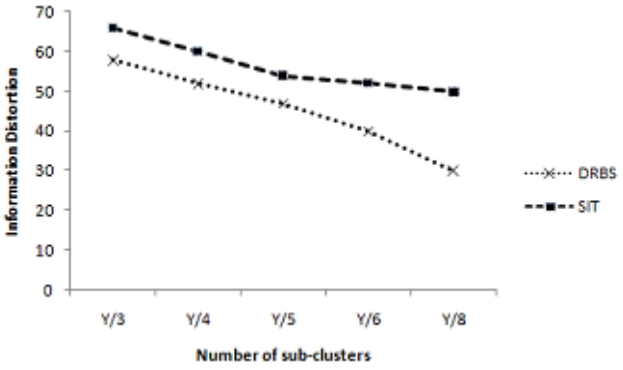

Figure 2. Performance based on Information Distortion.

4.3.2 *Mis-classification Error (ME)*

The number of records wrongly matched to a different cluster due to data modification with respect to the total number of records is termed as mis-classification error. The value should be zero for an efficient system. But computationally only for a direct isometric transformation it is zero. Since they are rotated with respect to the centroid within the cluster the error is minimum compared to a randomization method. As the number of sub-clusters increases, the distance between the centroids decreases resulting in less ME. Figure 3 shows the change of ME with respect to simple Isometric transformation as the size of the sub-cluster increases.

4.3.3 *Amount of Privacy*

This is measured by the number of records altered during Anonymization. If the records remain unaltered, then they are considered to be not protected. If the records are homogenous then SIT results in more number of unaltered records. Figure 4 shows the amount of privacy preserved by DRBS method. As the number of sub-cluster increases, there are more chances for a record to be in equivalence classes and the distance between the centroids become lesser, leading to more number of unaltered records. The problem of homogeneity attack in SIT has been overcome in this method as different sub-clusters are chosen for relocation. The problem of similarity attack is also dealt successfully as no direct heuristic methods are followed and the centroids of sub-clusters before and after Anonymization are different. Skewness attack is handled, as any partial or complete records knowledge will not assist to identify the remaining records.

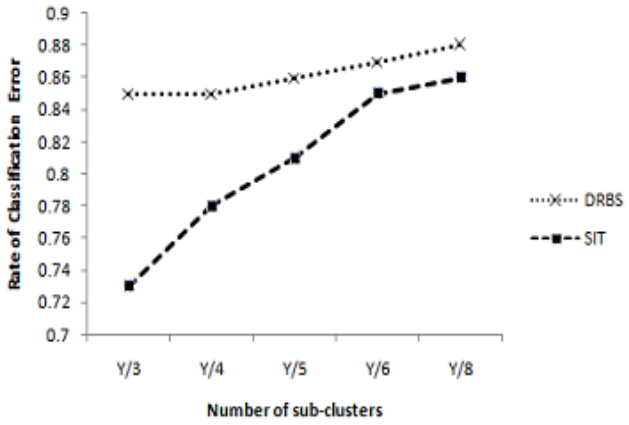

Figure 3. Performance based on Classification Error.

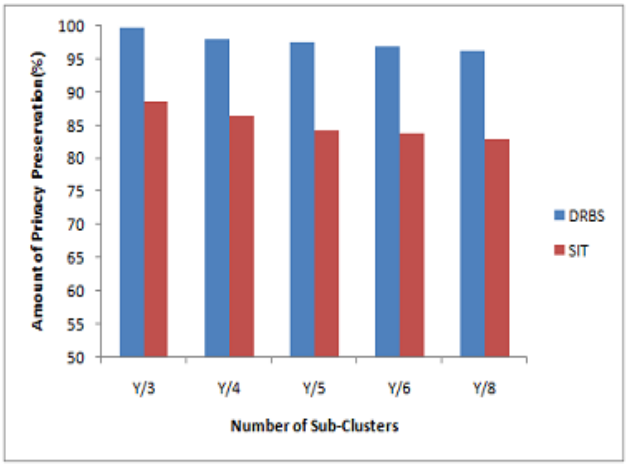

Figure 4. Performance based on Privacy preservation.

The analysis shows that as the number of sub-clusters increases, the information distortion and mis-classification error decreases but the amount of privacy preserved records also decreases. Hence to avoid this tradeoff between utility and privacy a suitable number of sub-clusters should be chosen which is found to be Y/4, where Y is the number of records in a cluster.

5. Conclusion and Future Work

The methodology of data relocation using sub-clustering, successfully anonymize the data, which can be used for data mining with maximum efficiency. The method is computationally irreversible and prevents from reconstruction attack, similarity and skewness attack. Since the procedure uses simple clustering scheme it takes linear amount of time to execute.

The method is implemented only for numeric attributes, which can be extended to categorical attributes. The method is not suitable for data streams, as each data may be present in different clusters which alter the centroids. Hence a suitable method for handling data streams is required.

6. References

1. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. Acm; 2000.
2. Inan A, Saygın Y, Savas E, Hintoglu AA, Levi A. Privacy preserving clustering on horizontally partitioned data. Data & Knowledge Engineering. 2007; 63:646–66.
3. Inan, A, Saygin, Y. Privacy preserving spatio-temporal clustering on horizontally partitioned data. DaWaK. 2006; 459–68.
4. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: privacy beyond k-anonymity. ACM Transactions On Knowledge Discovery From Data. 2007 Mar; 1(1).
5. Wu CW. Privacy preserving data mining with unidirectional interaction. IEEE International Symposium on Circuits and Systems; 2005.
6. Aggarwal CC, Yu PS. A condensation approach to privacy preserving data mining. Advances in Database Technology-Lecture Notes in Computer Science. 2004; 2992; 183–99.
7. Chiu C-C, Tsai C-Y. A k-anonymity clustering method for effective data privacy preservation. Advanced Data Mining and Applications- Lecture Notes in Computer Science. 2007; 4632:89–99.
8. Chiu C-C, Tsai C-Y. Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm. Comput Stat Data Anal. 2008 Jun; 4658–72.
9. Hong D, Mohaisen A. Augmented rotation-based transformation for privacy-preserving data clustering. Etri Journal. 2010 Jun; 32(3).
10. Ghinita G, Kalnis P, Tao Y. Anonymous publication of sensitive transactional data. IEEE Trans Knowl Data Eng. 2011 Feb; 23(2).
11. Tian H, Zhang W. Extending l-diversity for better data anonymization. Nfs Grant Iis-0524612; 2009.
12. Buratović I, Miličević M, Žubrinić K. Effects of data anonymization on the data mining results. Mipro; 2012.
13. Vaidya J, Clifton C. Privacy-preserving data mining: why, how, and when. IEEE Security & Privacy; 2004.
14. Cao J, Carminati B, Ferrari E, Tan K-L. Castle: continuously anonymizing data streams. IEEE Transactions on Dependable and Secure Computing. 2011 May/Jun; 8(3).
15. Byun J-W, Kamra A, Bertino E, Li N. Efficient k-anonymity using clustering technique. Cerias Tech Report; 2006–10.
16. Guo AK, Zhang Q. Fast clustering-based anonymization approaches with time constraints for data streams. Knowledge-Based Systems. 2013; 46:95–108.
17. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems. 2002; 10 (5):571–88.
18. Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. Proc. ACM Sigmod Int'l Conf. Management of Data. 2007; 665–76.
19. Machanavajjhala A, Gehrke J, Kifer D. L-diversity: privacy beyond k-anonymity. Proc of the International Conference On Data Engineering; 2006; Atlanta, GA, USA. p. 24.
20. Adam NR, Wortmann JC. Security-control methods for statistical databases: a comparative study. ACM Computing Surveys. 1989; 21(4):515–56.
21. Li N, Li T, Venkatasubramanian S. T-closeness: privacy beyond k-anonymity and l-diversity. IEEE; 2007.
22. Fong PK, Weber-Jahnke JH. Privacy preserving decision tree learning using unrealized data sets. IEEE Trans Knowl Data Eng. 2012 Feb; 24(2).
23. Agrawal R, Srikant R. Privacy preserving data mining. Proc ACM Sigmod; 2000. p. 439–50.
24. Banu RV, Nagaveni N. Evaluation of a perturbation-based technique for privacy preservation in a multi-party clustering scenario. Information Sciences. 2013; 437–48.
25. Xiangmin R, Jing Y. Research on privacy protection based on k-anonymity. IEEE; 2010.
26. Dhiraj SSS, Khan A, Khan W, Challagalla A. Privacy preservation in k-means clustering by cluster rotation. IEEE, Tencon; 2009.
27. Oliveira SRM, Zaïane OR. Achieving privacy preservation when sharing data for clustering. Proc SDM; 2004. p. 67–82.
28. Samarati P, Sweeney L. generalizing data to provide anonymity when disclosing information. Proceedings Of The Seventeenth ACM Sigact sigmod-Sigart Symposium on Principles of Database Systems, Pods. 1998; Seattle,WA, USA. p. 188.
29. Kisilevich S, Rokach L, Elovici Y, Shapira B. Efficient multidimensional suppression for k-anonymity. IEEE Trans Knowl Data Eng. 2010 Mar; 22(3).
30. Oliveira RMS, Zaiane OR. Data perturbation by rotation for privacy-preserving clustering. Technical Report Tr 04-17; 2004 Aug.
31. Iyengar V. Transforming data to satisfy privacy constraints. Proc ACM Sigkdd Int'l Conf Knowledge Discovery and Data Mining; 2002. p. 279–88.

32. Li X-B, Sarkar S. A tree-based data perturbation approach for privacy-preserving data mining. IEEE Trans Knowl Data Eng. 2006 Sep; 18(9).
33. Sun X, Li M, Wang H. A family of enhanced (l, a)-diversity models for privacy preserving data publishing. (Future Generat Comput Syst. 2011; 27:348–56.
34. Li Y, Chen M, Li Q, Zhang W. Enabling multilevel trust in privacy preserving data mining. IEEE Trans Knowl Data Eng. 2012; 24.
35. Tao Y, Chen H, Xiao X, Zhou S, Zhang D. Angel: enhancing the utility of generalization for privacy preserving publication. IEEE Trans Knowl Data Eng. 2009 Jul; 21(7).