2010 34th Annual IEEE Computer Software and Applications Conference Workshops

Privacy Preserving Outlier Detection using Hierarchical Clustering Methods

Ajay Challagalla
cjajay@gmail.com

S. S. Shivaji Dhiraj
shivajidhiraj@yahoo.com

D.V.L.N Somayajulu
soma@nitw.ac.in

Toms Shaji Mathew
mathew.toms@gmail.com

Saurav Tiwari
srvtiwari@gmail.com

Syed Sharique Ahmad
sharique.co@gmail.com

Department of Computer Science and Engineering
National Institute of Technology, Warangal
India

Abstract— **Data objects which do not comply with the general behavior or model of the data are called Outliers. Outlier Detection in databases has numerous applications such as fraud detection, customized marketing, and the search for terrorism. However, the use of Outlier Detection for various purposes has raised concerns about the violation of individual privacy. Therefore, Privacy Preserving Outlier Detection must ensure that privacy concerns are addressed and balanced, so that the data analyst can get the benefits of outlier detection without being thwarted by legal counter-measures by privacy advocates. In this paper, we propose a technique for detecting outliers while preserving privacy, using hierarchical clustering methods. We analyze our technique to quantify the privacy preserved by this method and also prove that reverse engineering the perturbed data is extremely difficult.**

Keywords— ***Data Mining, Outlier Detection, Privacy Preservation, Data Perturbation, Hierarchical Clustering.***

I. INTRODUCTION

Outliers generally represent anomalous behaviour [1]. By definition, outliers are rare occurrences and hence represent a small portion of the data.

Outlier detection has direct applications in a wide variety of domains such as mining for anomalies to detect network intrusions, fraud detection in mobile phone industry [2] and recently for detecting terrorism related activities [3]. However, the application of Outlier Detection for these purposes has opened new threats to the privacy and autonomy of the individual if not done properly [8].

Clustering is an important concept used for outlier analysis. Several clustering based outlier detection techniques have been developed [16], [11]. Most of these techniques rely on the key assumption that normal objects belong to large and/or dense clusters, while outliers form very small clusters [13].

In [14], a novel method for outlier detection using hierarchical clustering techniques has been proposed. In this method, the hierarchical clustering process stops at a pre-determined level, and the member objects of small clusters (clusters having very few objects) are considered to be outliers. However, [14] does not incorporate privacy concerns related to outlier detection. In general, collection and analysis of data trigger a variety of privacy incursions on our "right to be let alone" [5]. There are certain databases where privacy/security concerns restrict the sharing of data. However, analysis of such data would prove to be an important asset to business organizations and governments for decision-making processes. Therefore, there is a need for the development of Data Mining techniques that incorporate privacy concerns.

In this paper, we propose a technique for Privacy Preserving Outlier Detection in statistical databases through data perturbation. In essence, data perturbation is accomplished by the alteration of an attribute value by a new value. Therefore, by perturbing the data while preserving the stages in the process of aggregation (clustering) we can provide a technique for outlier detection without violating the privacy of individual data objects.

II. MOTIVATION AND RELATED WORK

Outlier detection has recently gained wide attention as an axe against global terrorism. Terrorism fits the description of a rare and anomalous activity.

In the recent past, law enforcement authorities in the UK have used outlier detection techniques by monitoring bank transactions and other patterns of people to distil out possible terrorists from the populace [3]. Though effective, two major concerns with such a technique are the occurrences of false positives and the subsequent loss of privacy of individuals. In a city of 50 million people, even if the algorithm is 99 percent accurate it would still identify about 500,000 people as terrorists which might not be the case. This problem of false positives in an outlier detection technique has to be carefully addressed [3], [9]. Secondly, it is well documented that the limitless explosion of data and mining of an individual's activities raises legitimate privacy concerns as they have a potential for misuse. Considering that the final results of a data mining technique do not result in a privacy concern, it is possible to use privacy preserving data mining techniques for knowledge discovery without compromising sensitive information [19].

The primary task in data mining is the development of models about aggregated data [4]. Hence by developing accurate models without allowing access to precise information we can carefully address the privacy concerns of data mining.

Privacy preserving methods have been developed for a wide variety of data mining tasks. Privacy preserving clustering techniques have been previously addressed in [8], [18] and [17]. In particular privacy preserving outlier detection was initially addressed in [9].

In [15], a generic technique for privacy preserving data mining has been proposed by creating a condensed group of

978-0-7695-4105-1/10 $26.00 © 2010 IEEE
DOI 10.1109/COMPSACW.2010.35

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL. Downloaded on November 13,2024 at 06:27:58 UTC from IEEE Xplore. Restrictions apply.

data from a given data set. While forming the condensed group this method randomly chooses an object, finds its nearest (k-1) objects and then put them in the same group. But, such a method cannot be used for finding outliers, since if the random point chosen happens to be an outlier, then the method would force the nearest (k-1) objects to group with the outlier, even though these points may be very much distant from the outlier. Hence, the information about the outlier is hidden in such a group.

Other outlier detection techniques have been addressed in [11] and [16]. In [11], a two-stage outlier detection algorithm has been proposed where the dataset is clustered by a one-pass clustering algorithm in stage one and then outlier clusters are discovered by an outlier factor during the second stage. In contrast [16] presents a technique for outlier detection using hierarchical clustering. This algorithm stops clustering according to the dissimilarity reflected by the detected outliers.

This paper proposes a technique for Privacy Preserving Outlier Detection using Hierarchical Clustering methods. A technique for Outlier Detection using hierarchical clustering methods has been addressed in [14].

The use of hierarchical clustering methods in this technique is motivated by the unbalanced distribution of outliers versus "normal" cases in data sets. In almost all attempts to create the initial clusters, non-hierarchical clustering methods would spread the outliers across all clusters. Given that most of those methods strongly depend on the initialization of the clusters, we expect this to be a rather unstable approach. Therefore, we use hierarchical clustering methods, which are not dependent on the initialization of the clusters.

The technique proposed in [9] assumes that the data is inherently distributed and hence the way in which the data is partitioned results in very different solutions. In contrast the technique proposed here requires that the entire dataset be available to create a perturbed dataset which can be provided to an analyst in such a way that privacy is not compromised.

III. BASIC CONCEPTS

We briefly review the basic concepts that are necessary to understand the technique proposed in this paper. First, we describe the Agglomerative Hierarchical Clustering Algorithm and then we explain the rotational data perturbation technique.

A. Agglomerative Hierarchical Clustering

The process of grouping a set of data points into classes of similar objects is called clustering [1]. The fundamental idea of clustering is that the intra-cluster similarity should be high and the inter-cluster similarity should be low. This bottom-up strategy of clustering starts by treating each data object as a cluster and then recursively merges these clusters based on the minimum distance between them until all objects are in a single cluster or until a condition is satisfied. At each step, a cluster is represented by the mean value of all the objects in the cluster, i.e., the centroid of that cluster.

A tree structure called dendrogram is commonly used to represent the process of hierarchical clustering [1]. It is a step by step diagram of the clustering process where the clustering is represented as the fusion of branches of the tree.

B. Rotational Data Perturbation

Rotational data perturbation is performed by pair-wise attribute distortion of all objects within a cluster using the rotational transformation matrix as proposed in [23]. Every cluster is rotated about its own cluster centroid and thus, the cluster centroids themselves do not undergo any change during cluster rotation. The centroid of a cluster is the attribute-wise arithmetic mean of its objects. The origin is assumed to be at the centroid of every cluster during its rotation.

This technique of rotational data perturbation has been applied for k-means clustering in [6].

IV. PROPOSED APPROACH

Consider the rotational data transformation technique for privacy preservation as presented in [6]. In principle, this technique can be applied for any kind of clustering algorithm.

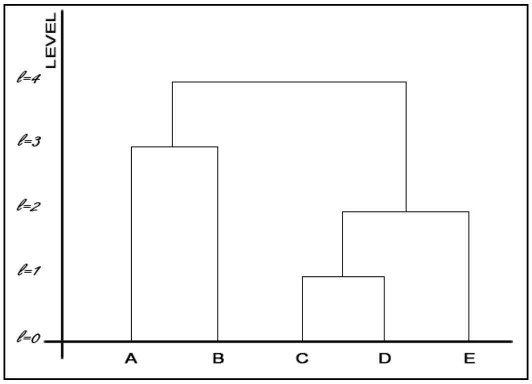


Fig. I Dendogram of the clustering process of a dataset consisting of points A,B,C,D,E.

In this method, given a set of m clusters, we perform geometric transformation through cluster rotation such that every cluster is rotated about its own centroid. The proposed technique is for privacy preserving outlier detection using hierarchical clustering. This technique consists of two phases, viz., cluster displacement and cluster rotation, iteratively at each level of the hierarchical clustering.

Cluster displacement is a translational perturbation performed by first displacing a set of clusters by a displacement factor '*λ*' from the mean of the entire dataset, so that no two clusters overlap after the subsequent cluster rotation process. For this purpose, the displacement vector between every cluster centroid G_i and the mean of the entire dataset G must be multiplied by a factor λ. To get the new value of G_i, this product vector, $\lambda \times GG_i$, must be added to the position vector OG of the mean G, where 'O' represents the origin.

To calculate *λ*, consider the worst case scenario, wherein the perturbed objects A_i' and A_j' lie on the line joining centroid G_i and G_j after rotation of C_i and C_j respectively. Therefore, C_i and C_j must be displaced from the mean of the entire dataset so that,

$$|G_i'G_j'| > 2 \times |A_i''G_i'| \qquad (1)$$

$$|G_i'G_j'| > 2 \times |A_j''G_j'| \qquad (2)$$

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL. Downloaded on November 13,2024 at 06:27:58 UTC from IEEE Xplore. Restrictions apply.

Where A_i'' and A_j'' are the coordinates after cluster displacement and rotation.
Thus, from (1) and (2), it follows that

$|G_i'G_j'| > 2 \times \text{maximum } \{|A_i''G_i'|, |A_j''G_j'|\}$.
$|G_i'G_j'| = \lambda_{ij} \times |G_iG_j|$.

Therefore, λ_{ij} must satisfy the condition

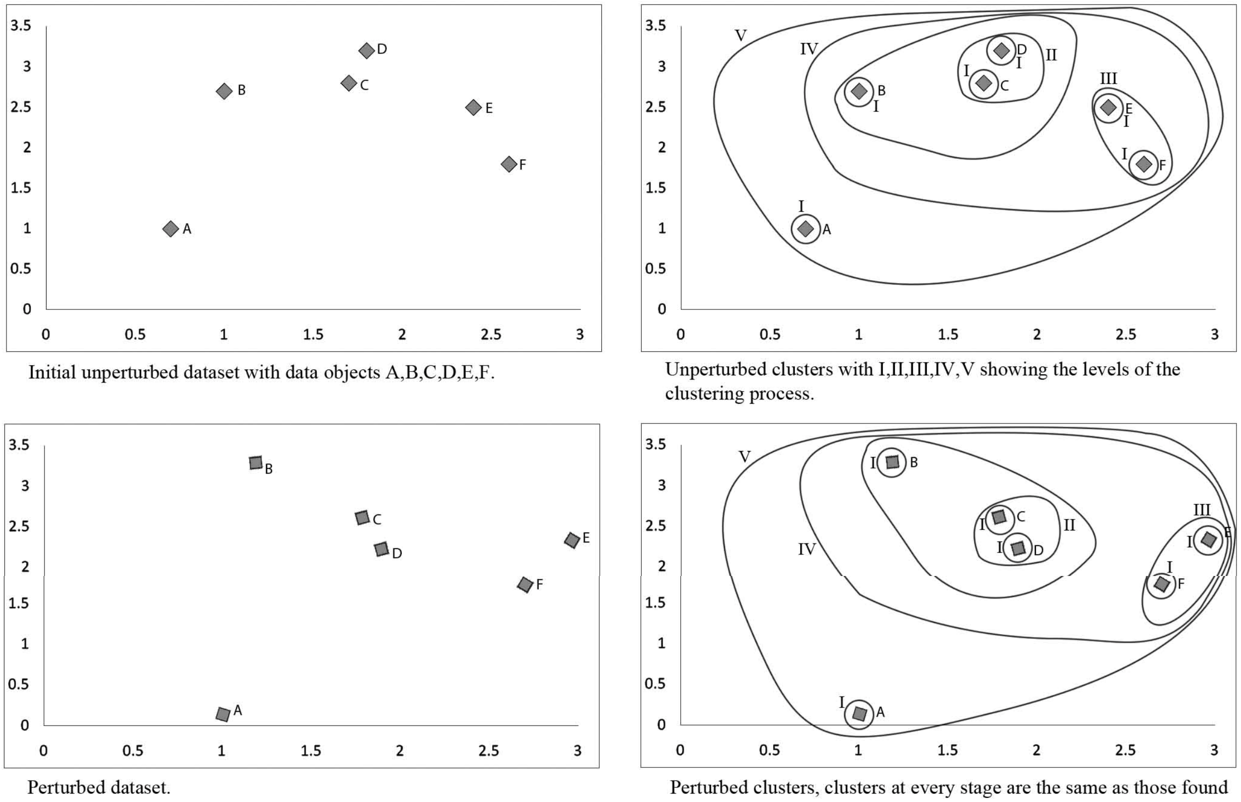
$\lambda_{ij} > 2 \times \text{maximum } \{|A_i''G_i'|, |A_j''G_j'|\} / |G_iG_j|$.

Finally to assign λ_{ij} any value greater than right hand side, we consider

$\lambda_{ij} = 2 \times \text{maximum } \{|A_i''G_i'| + 1, |A_j''G_j'| + 1\} / |G_iG_j|$.

We find λ_{ij} for all pairs of clusters and take λ as the maximum of all these λ_{ij}.

Cluster rotation is performed by pair-wise attribute distortion of all the objects within a cluster by rotating them about their cluster centroid. These rotations are isometric transformations; hence the distances between the objects within a cluster are preserved after each rotation.



Initial unperturbed dataset with data objects A,B,C,D,E,F.

Unperturbed clusters with I,II,III,IV,V showing the levels of the clustering process.

Perturbed dataset.

Perturbed clusters, clusters at every stage are the same as those found in the unperturbed dataset.

Fig. 2: Illustration of the proposed data perturbation technique

The rotational transformation matrix is applied to every pair of attributes for each object within a particular cluster, i.e., to perform this transformation, a pair of attributes must be distorted at a time. Hence, to distort the values of all the attributes of an n-dimensional object the minimum number of distortions required would be $\frac{n}{2}$ if n is even and $\frac{n+1}{2}$ if n is odd. The clustering process depends on the relative position of the clusters, and if we show that the new positions of the clusters after cluster displacement are relative to the initial distances between them, then we can conclude that cluster displacement has no effect on the clustering process.

Consider two clusters C_i and C_j, with position vectors of each object in the clusters as OA_{ir} and OA_{jr} respectively. The relative distance between the two objects belonging to two clusters is given by,

Relative distance, $\beta = OA_{ir} - OA_{jr}$

Now, consider the effect of cluster displacement using λ as cluster displacement factor. As shown in [6] the new coordinates (OA_{ir}' and OA_{jr}') of objects belonging to i^{th} and j^{th} clusters are,

$$OA_{ir}' = OA_{ir} + (\lambda - 1) \times GG_i$$
$$OA_{jr}' = OA_{jr} + (\lambda - 1) \times GG_j$$

where GG_i and GG_j are vectors, which join the mean of the entire dataset, G to the centroid G_i of cluster C_i and centroid G_j of cluster C_j respectively.
The relative distance in this case would be,

$$\begin{aligned} \beta' &= OA_{ir}' - OA_{jr}' \\ &= (OA_{ir} - OA_{jr}) + (\lambda - 1)(G_i - G_j) \\ &= \beta + (\lambda - 1)\,(G_i - G_j) \end{aligned}$$

Hence β' depends on β, λ and G_i-G_j, of which G_i-G_j and λ are constant for a given pair of clusters. Thus β' depends only on β.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL. Downloaded on November 13,2024 at 06:27:58 UTC from IEEE Xplore. Restrictions apply.

Therefore, we conclude that cluster displacement has no effect on the clustering process.

In hierarchical clustering, two clusters are merged based on the distance between their centroids [1].The rotation of a cluster with respect to its centroid does not affect the position of its centroid [6]. Hence we can conclude that cluster rotation will not affect the clustering at the next stage of the hierarchical clustering process.

The proposed technique is summarized in Algorithm-I below.

Algorithm-I: Data Perturbation by Iterative Cluster Rotation for Hierarchical Clustering

Input: Dataset D containing 'm' n-dimensional objects.

Output: Perturbed dataset consisting of 'm' n-dimensional objects.

Method:

```
(1)Create 'm' clusters corresponding to each data object.
(2)Repeat
(3)    Distance($C_p$,$C_q$)=minimum(distance($C_i$,$C_j$)), for all
       i , j forming cluster pairs.
(4)    Merge the clusters $C_p$ and $C_q$.
(5)    Update the cluster centroid of merged cluster.
(6)    For each cluster $C_i$, calculate the maximum distance
       di of an object in $C_i$ from its centroid $G_i$.
(7)    For each pair of clusters $C_i$ and $C_j$, calculate $\lambda_{ij} > 2 *$
       $\max \{d_i,d_j\} / (G_i - G_j)$;
(8)    Displace the clusters by $\lambda$ around the mean of the
       entire dataset if $\lambda>1$, where $\lambda=\max\{\lambda_{ij}\}$, for all pair
       of clusters $C_i$ and $C_j$.
(9)    for each object in cluster $C_i$, apply rotational
       transform about centroid $G_i$ to each pair of
       attributes p and q , where $\theta_{pq}$ is constant for all
       objects in $C_i$;
(10)Until, only 1 cluster remains.
```

The input to the perturbation technique is a dataset of 'm' n-dimensional objects. The output is a perturbed dataset which gives the same hierarchical clustering dendrogram as the original dataset.

Fig. 2 depicts the data perturbation process and the hierarchical clustering process for a sample dataset.

A. Normalization of the Perturbed Dataset

It is possible that as a result of the perturbation process the attribute values of the objects get transformed to undesirable domains (e.g., negative value for a non-negative attribute such as 'age'). Though this does not interfere with our proposed approach, the meaning of the data may be lost as a result of the undesirable domain.

The dataset can be normalized by scaling it by a suitable factor around the mean of the dataset. It must be noted that this normalization process does not affect the quality of the clusters formed as the entire dataset is being scaled by the same factor around the mean of the dataset.

B. Computational Costs

The proposed data perturbation technique has a complexity of the order O (k × m_{avg} × n × l) where 'k' is the number of clusters, 'm_{avg}' is the average number of objects in a cluster, 'n' is the number of attributes and 'l' is the number of levels in hierarchical clustering. Considering that only one new cluster is formed at each level of clustering, l = M where 'M' is the number of data objects. So the computational cost becomes O (k × m_{avg} × n × M). The technique proposed in [11] for finding outliers has a complexity of O (k x M x n), where M is the number of data objects, 'k' is the number of clusters and 'n' is the number of attributes. Thus, the proposed technique is computationally more expensive than other comparable methods. However, these additional computational costs are effectively offset by the fact that this technique gives the data analyst the freedom to set the parameters for stopping the hierarchical clustering at any level and offers a zero misclassification error thereby increasing the usability of the perturbed dataset.

V. Experiments and Analysis

The aim of privacy preserving data mining algorithms is to extract relevant knowledge from large amounts of data while protecting at the same time sensitive information [12]. However, due to the variety of characteristics of these algorithms, it is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. Rather, an algorithm may perform better than another one on specific criteria like hiding failure, data usability, etc. Therefore, we analyse the performance of our algorithm with respect to the following parameters.

A. Hiding Failure

Hiding failure is the portion of sensitive information that is not hidden by the application of a privacy preservation technique [12]. The percentage of sensitive information that is still discovered, after the data has been sanitized gives an estimate of the hiding failure parameter. In [21], Oliveira and Zaiane define the hiding failure (HF) as the percentage of restrictive patterns that are discovered from the sanitized database. Most of the developed privacy preserving algorithms are designed with the goal of obtaining zero hiding failure. HF is measured as follows:

$$HF = \frac{\#R_p(D')}{\#R_p(D)}$$

where #Rp (D) and #Rp(D′) denote the number of restrictive patterns discovered from the original data base D and the sanitized database D′ respectively. Ideally, the HF must be 0.
In the data perturbation technique proposed in this paper, the value of every object in the dataset is modified because every cluster is rotated about its centroid. Therefore, no sensitive information is discovered after the application of the data perturbation method. Hence, the Hiding Failure for the proposed data perturbation technique is 0.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL. Downloaded on November 13,2024 at 06:27:58 UTC from IEEE Xplore. Restrictions apply.

B. Misclassification Error

The quality of the perturbed data obtained in this method, with respect to cluster analysis, can be measured by the misclassification error percentage. Misclassification error is measured in terms of the percentage of data objects that are not well-classified in the distorted database. A lower misclassification error is desirable as it signifies a smaller change in the object membership of clusters. Ideally, the misclassification error should be 0%. The misclassification error, denoted by M_E, is measured as

$$M_E = \frac{1}{k} \times \sum_{i=1}^{k}(|C_i| - |C_i'|)$$

where k is the number of clusters in the dataset, and C_i□ is the perturbed cluster corresponding to the unperturbed cluster C_i.
In [22], a set of hybrid transformations has been introduced to ensure privacy of categorical data in clustering. The misclassification errors obtained after applying the hybrid data transformation techniques for various noise levels are computed and they are found to be the least for a noise level of 75%.

In [8], a family of Geometric Data Transformation methods (GDTMs) is introduced for Privacy Preserving Clustering. Datasets perturbed by GDTMs have a nonzero misclassification error.

However, in the proposed data perturbation technique, the clusters are displaced in such a way that they do not overlap after the subsequent rotation process, i.e., the object membership of clusters does not change due to the perturbation. Therefore, the misclassification error is always 0%.

C. Data Usability

The term data usability refers to the ability of a data perturbation technique to provide accurate aggregate information. An ideal data perturbation technique is one that preserves both statistical as well as clustering information [7]. Data randomization techniques [4], which obfuscate data by the addition of random noise to the original data, can be tailored to preserve statistical information. However, the inherent clusters in the original data are distorted because of the addition of random noise.

The technique proposed in this paper perturbs the data by cluster rotation. The misclassification error is zero because clusters do not rearrange after the perturbation, i.e., the dendrogram of hierarchical clustering (as shown in Fig. 1) for the perturbed dataset is similar to that of the original dataset. Therefore, the perturbed dataset can still be used for outlier detection. Since the misclassification is zero, an outlier in the perturbed dataset would correspond to an outlier in the original dataset also. Therefore, the problem of false positives, as discussed in [3] and [9], does not arise in this case.

In this technique, cluster displacement and cluster rotation are iteratively performed at every stage in the hierarchy. This is done so that the clusters obtained in the perturbed dataset are consistent with those obtained for the original dataset at every stage, i.e., the data analyst has the freedom to set the parameter to stop the hierarchical clustering at any level based on the number of clusters obtained at that level, and to set the threshold to determine which clusters can be classified as outliers (i.e., objects within small clusters).
Thus, the perturbed dataset obtained in this method yields outliers consistent with those in the original dataset, thereby making the perturbed data more useful.

D. Experimental Results

We implemented the proposed technique on an open source LINUX platform using C++. We tested the technique on standard datasets, each having ‘m’ n-dimensional tuples and evaluated the value of Misclassification Error, M_E, for each dataset. The number of levels in the hierarchical clustering dendrogram for each dataset is denoted by ‘l’.

Table I lists the experimental results for four statistical datasets (i.e., datasets having only numerical attributes), obtained from the UCI Machine Learning Repository. The values of Misclassification Error and Hiding Failure for all datasets perturbed by this technique are found to be 0.

TABLE I: RESULTS FOR SAMPLE DATASETS

Dataset	M	n	l	M_E (%)	HF
D1	768	8	768	0	0
D2	2000	649	2000	0	0
D3	4601	57	4601	0	0
D4	5473	10	5473	0	0

D1. Pima Indians Diabetes Data Set.
D2. Multiple Handwritten Features Data Set.
D3. Spam base classifying email as spam and non-spam.
D4. Page Blocks Classification Data Set.

E. Adversary Attack on Data Privacy

Existing work on analysis of privacy preserving perturbation techniques primarily considers an adversary that correlates publicly available data to reconstruct sensitive information [7] and [12]. The worst case scenario would be an attacker that has prior information about the perturbation technique as well as some of the original data objects. The attacker could then proceed to reconstruct the entire original dataset. The effectiveness of a perturbation technique, therefore, would be the ability of the perturbed dataset to withstand such an attack.

To withstand such an attack the dataset must be perturbed by several randomized parameters so that a brute force attack would be impossible even if the adversary can correlate from publicly available data.

The proposed technique considers ‘n’ data objects as clusters initially, and then data perturbation is performed on each cluster. At each level, clusters having minimum distance are merged together. So, the maximum number of times an object is rotated in this method is ‘n’.

From the adversary point of view, it would almost be impossible to use brute force to reverse engineer these angles, since these angles are randomly distributed for each data object, as at each level the data object may be present in a different cluster. So the sequences of angles through which these objects are rotated are all different for different objects. Moreover, there is cluster displacement in between each rotation angle sequence and thus the prediction of λ at each

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL. Downloaded on November 13,2024 at 06:27:58 UTC from IEEE Xplore. Restrictions apply.

step would make it difficult for the adversary. The value of λ depends on the orientation of clusters before rotation, which the adversary cannot obtain without actually displacing the clusters towards centroid by a value λ. Even if the adversary, using brute force, figures out one such sequence (only if adversary has the correlation between that object and its original value), adversary would be able to trace the original object corresponding only to A_1. In the worst case, the same sequence of reverse engineering would be required to find another object A_2, which merges with object A_1 at the first level of hierarchical clustering. But it can be shown mathematically, that such reversal is not possible.

We consider two attribute values a_{irp} and a_{irq} of A_{ir} in cluster C_i.

After cluster rotation (θ) the coordinates are given by,

$$a'_{irp} = a_{irp}.\cos\theta - a_{irq}.\sin\theta$$
$$a'_{irq} = a_{irp}.\sin\theta + a_{irq}.\cos\theta$$

Now, final values of coordinates after cluster displacement are,

$$a_{firp} = a'_{irp} + (\lambda\text{-}1)\, g_{ipd} = (a_{irp}.\cos\theta - a_{irq}.\sin\theta) + (\lambda\text{-}1)\, g_{ipd}$$
$$a_{firq} = a'_{irq} + (\lambda\text{-}1)\, g_{iqd} = (a_{irp}.\sin\theta + a_{irq}.\cos\theta) + (\lambda\text{-}1)\, g_{iqd}$$

where g_{ipd} and g_{iqd} are the changes in cluster centroid coordinates.

For, a system of linear equations to yield a particular solution, the number of equations should be greater than or equal to the number of unknown variables. In this case, since the number of unknown variables, viz., $a_{irp,}$ $a_{irq,}$ $a'_{irp,}$ $a'_{irq,}$ λ and θ, is greater than the number of equations, clearly the adversary would not be able to recover the original data values. Hence, data perturbed by this technique cannot be easily reverse engineered.

VI. CONCLUSION AND FURTHER WORK

In this paper, a technique for privacy preserving outlier detection using hierarchical clustering is proposed. The data at every stage of a hierarchical clustering is perturbed such that their values are modified, but, the perturbed dataset will yield the same outliers as the original dataset. Moreover, the hierarchical clustering dendrogram of the perturbed dataset is the similar to that of the original dataset. This gives the data analyst the freedom of setting the parameters for stopping the hierarchical clustering at any stage. The perturbed dataset obtained in this method has a zero hiding failure and we show that it is very difficult to reverse engineer such a dataset. Thus, this technique results in the increased usability of the perturbed dataset while offering a good security measure against attacks on data privacy.

The data perturbation technique proposed here is very robust, with zero misclassification error and zero hiding failure. The possibility of using the perturbed dataset obtained in this method for other data mining tasks needs to be explored. Further work also lies in the application of this technique to detect outliers using other clustering algorithms.

REFERENCES

[1] Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques ,* 2nd edition, Morgan Kaufmann, 2006.

[2 Jaideep Vaidya, Chris Clifton, “Privacy-Preserving Outlier Detection*”*, ICDM, pp.233-240, Fourth IEEE International Conference on Data Mining (ICDM'04), 2004.

[3] Steven D. Levitt & Stephen J. Dubner, *Superfreakonomics: Global Cooling, Patriotic Prostitutes, and why Suicide Bombers Should Buy Life Insurance*, William Morrow, October 2009.

[4] R. Agrawal and R. Srikant, “Privacy preserving data mining”, ACM SIGMOD Conf. Management of Data, pp. 439-450, May 2000.

[5] Mary J. Cronin, "Privacy and Electronic Commerce", the new Hoover Press book Public Policy and the Internet: Privacy, Taxes, and Contract, edited by Nicholas Imparato.

[6] S. S. Shivaji Dhiraj, Ameer M. Asif Khan, Wajhiulla Khan, Ajay Challagalla, “Privacy preservation in k-means clustering by cluster rotation”, TENCON 2009 - 2009 IEEE Region 10 Conference

[7] Rupa Parameswaran and Douglas M. Blough, “Privacy preserving data obfuscation for inherently clustered data”, International Journal of Information and Computer Security, v.2 n.1, p.4-26, January 2008.

[8] S. Oliveira and O. R. Za¨iane, “Privacy preserving clustering by data transformation”, the 18th Brazilian Symposium on Databases, pages 304–318, 2003.

[9] Jaideep Vaidya, Christopher Wade Clifton, Michael Zhu, *Privacy Preserving Data Mining*, Springer, December 2009.

[10] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar. “A comparative study of anomaly detection schemes in network intrusion detection”, in SIAM International Conference on Data Mining (2003), San Francisco, California, May 1-3 2003.

[11] Sheng-yi Jiang, Qing-bo- An, “Clustering-Based Outlier Detection Method”, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD'08, 2008.

[12] Elisa Bertino , Dan Lin and Wei Jiang, “A Survey of Quantification of Privacy Preserving Data Mining Algorithms”**,** in Privacy-Preserving Data Mining (Models and Algorithms), Charu C. Aggarwal and Philip S. Yu (Eds.), Springer-Verlag, 2008.

[13] An Effective Clustering-Based Approach for Outlier Detection, Moh’d Belal Al- Zoubi, European Journal of Scientific Research, ISSN 1450-216X Vol.28 No.2 (2009), pp.310-316, 2009

[14] Antonio Loureiro, Luis Torgo, and Carlos Soares, “Outlier Detection Using Clustering Methods: a data cleaning application”, in Proceedings of the Data Mining for Business Workshop, 2009.

[15] Aggarwal, C. C. and Yu, P. S., “A condensation approach to privacy preserving data mining”, in proceedings of the International Conference on Extending Database Technology (EDBT) , 2004.

[16] Tian-yang Lv, Tai-xue Su, Zhengxuan Wang, Wanli Zuo, “An Auto-stopped Hierarchical Clustering Algorithm Integrating Outlier Detection Algorithm”, WAIM 2005: 464-474

[17] J. Vaidya and C. Clifton, “Privacy preserving k-means clustering over vertically partitioned data”, the 9th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining. ACM Press, 2003.

[18] 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.

[19] Murat Kantarcio□lu, Jiashun Jin and Chris Clifton, “When do data mining results violate privacy?”,in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

[20] Muralidhar, K., R. Parsa and R. Sarathy, “A general additive data perturbation method for database security”, J. Mgmt. Science, 1999, Vol: 45, pp: 1399-1415.

[21] S.R.M. Oliveira, O. Za¨iane, “Privacy preserving frequent itemset mining”, IEEE ICDM Workshop on Privacy, Security and Data Mining, vol. 14, pp. 43–54 (2002)

[22] R.R. Rajalaxmi and A.M. Natarajan, “An effective data transformation approach for privacy preserving clustering”, Journal of Computer Science 4(4): 320-326, 2008, Science Publications.

[23] Stanley R.M. Oliveira and Osmar R. Za¨ıane “Achieving privacy preservation when sharing data for clustering”, the Workshop on Secure Data Management in a Connected World (SDM’04) in conjunction with VLDB’2004, Toronto, Canada, pg. 67-82.

Authorized licensed use limited to: NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL. Downloaded on November 13,2024 at 06:27:58 UTC from IEEE Xplore. Restrictions apply.