ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

Preserve Privacy of Horizontally Distributed Data through Scaling for Clustering

Khatri Nishant P[1], Ms. Preeti Gupta[2], Tusal Patel[3]

M. Tech Scholar, Computer Science and Engineering, Amity School of Engineering and Technology, Jaipur, India [1, 3]

Computer Science and Engineering, Amity School of Engineering and Technology, Jaipur, India [2]

Abstract: Data sharing among organizations is considered to be useful as it offers mutual benefits for effective decision making and business growth. Data mining techniques can be applied on this shared data which can help in extracting meaningful, useful, previously unknown and ultimately comprehensible information from large databases. This paper represents a privacy preserving technique for horizontally distributed data. Procedure stated in this work is based on data matrix scaling operation.

Keywords: Data Mining, Clustering, Data Distribution, Scaling, 2D transformation

I. INTRODUCTION

Data Mining: Data Mining is the technique used by analysts to find out the hidden and unknown pattern from the collection of data. Although the organizations gather large volumes of data, it is of no use if "knowledge" or "beneficial information" cannot be inferred from it. Unlike the statistical methods the data mining techniques extracts *interesting* information. The operations like classification, clustering, association rule mining, etc. are used for data mining purposes.

Confidentiality Issues in Data Mining: This in turn has lead to proliferation of private data by the organizations, which results in the increased concern about privacy of confidential data.Most privacy preserving data mining methods use some form of transformation on data to perform privacy preservation. Typically, such methods reduce the granularity of representation to preserve privacy. This paper presents a technique of privacy preserving clustering where scaling transformation applied on centralized data stored in a data matrix can lead to preserving of confidentiality yet not changing the nature of the data and the relationship existing between the data objects.

Data Distribution: The term data distribution means the manner in which the data has been stored at the sites (DB servers). Primarily there are two types of data distribution i) Centralized Data and ii) Partitioned Data.

In a centralized data environment all data is stored at single site. While in distributed environment all data is distributed among different sites. Distributed data can further be divided in i) Horizontally and ii) Vertically distributed environments. In horizontal distribution the different sites stores the same attributes for different sets of records. In vertical distribution the sites stores different attributes for the same set of records.

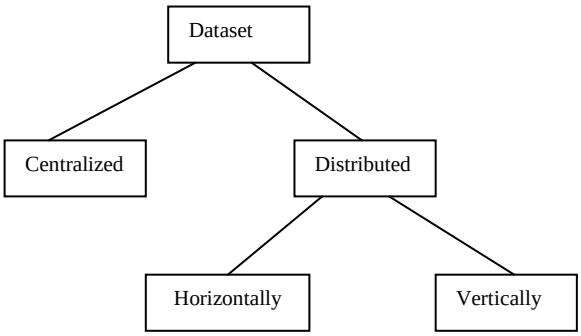
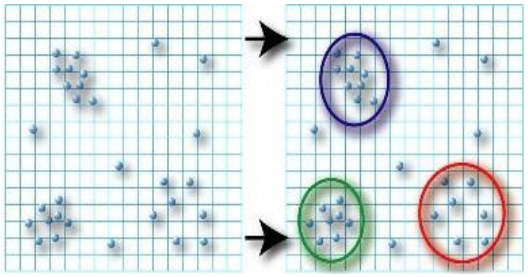


Fig. 1 Classification of Dataset

Clustering: Distributing each data object to the group of similar data objects can be termed as clustering in simple language. Technically it can be defined as the task of grouping a set of objects in such a manner that objects in same group (cluster) are more similar to each other than to those in other groups (clusters). It is main task carried out for machine learning, pattern recognition, information retrieval, etc. Clustering can be partitioned in i) Hierarchical ii) Partition Based iii) Density Based Clustering. Clustering can be graphically shown as:

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013



II. RELATED WORK

[1] presents the k-means technique to preserve privacy of vertically partitioned data. [2] suggests an algorithm for privacy preservation for Support Vector Machines(SVM) based classification using local and global models. Local models are relevant to each participating party that are not disclosed to others while generating global model jointly. Global model remains the same for every party which is then used for classifying new data objects. [3] represents two protocols for privacy preserving clustering to work upon horizontally and vertically partitioned data separately. [4] suggests methods for constructing dissimilarity matrices of objects from different sites in privacy preserving manner. In [5], a procedure is mentioned for securely running BIRCH algorithm over arbitrarily partitioned database. Secure protocols are mentioned in it for distance metrics and procedure is suggested for using these metrics in securely computing clusters. [6] represents various cryptographic techniques for privacy preserving. [7] presents various techniques of privacy preserving for different procedures of data mining. An algorithm is suggested for privacy preserving association rules. A subroutine in this work suggests procedure for securely finding the closest cluster in k-means clustering for privacy preservation. [8] suggests scaling transformation on centralized data to preserve privacy for clustering.

III. SCALING BASED PRIVACY PRESERVING ALGORITHM

A. Terms Used

1. Data Matrix

Objects (e.g. individuals, patterns, events) are usually represented as points (vectors) in a multidimensional space. Each dimension represents a distinct attribute describing the object. Thus, an object is represented as an m x n matrix D, where there are m rows, one for each object, and n columns, one for each attribute. This matrix is referred to as a data matrix, represented as follows:

$$\begin{bmatrix} a_{11} & . & a_{1k} & . & a_{1n} \\ a_{21} & . & a_{2k} & . & a_{2n} \\ . & . & . & . & . \\ a_{m1} & . & a_{mk} & . & a_{mn} \end{bmatrix}$$

B. Assumptions

a) The work in this paper concentrated on securing numeric attributes only with an assumption that numeric data is the most sensitive data that needs to be secured, such as salary, phone number, etc.

b) All transformations carried out in this work is assumed to be 2D transformations only.

C. General Approach

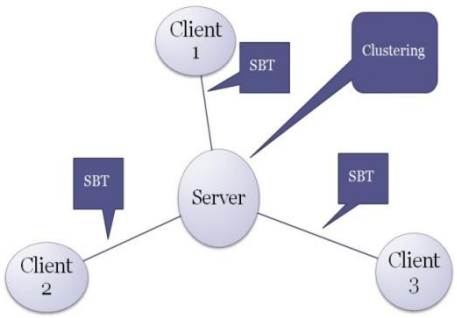
Let D_{mxn} be a data matrix, where each row represents an object, and each object contains values for each of n numerical attributes. The scaling based method of dimension n is an ordered pair, defined as (D,f_s) where,

1. $D \in R_{mxn}$ is a normalized data matrix of objects to be clustered
2. f_s is scaling based transformation function.

As the procedure is completely dependent on the data matrix scaling it is necessary to select proper scaling factor (s).As only 2D transformations are considered here the scaling factor must be kept equal in both x and y directions.

D. SBT Application Scenario

The following diagram shows that the proposed SBT algorithm must be applied on the participating site (Clients) before sending it for clustering. The operating site (Server) performs the clustering and returns the output to all the participating sites.



ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

E. Proposed Algorithm

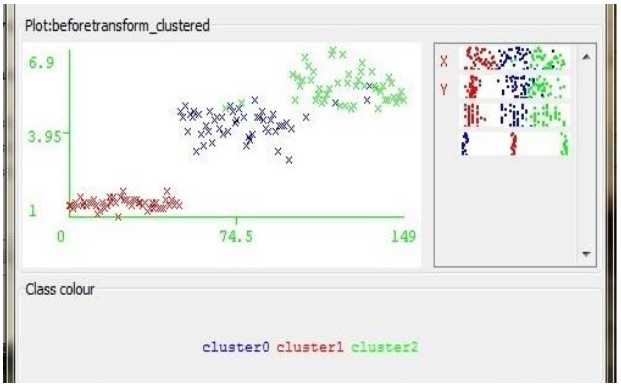
SBT_Algorithm

Input : D_{mxn} // D_{mxn} is normalized data matrix
Output: D'_{mxn}
1. $k \leftarrow \lceil n/2 \rceil$
2. $P_k \leftarrow$ k pairs(A_i,A_j) in D such that $1 \le i,j \le n$ and $i \ne j$
3. Decide scaling factor s.
4. For each selected pair P_k in pairs(d) do
   a. $V(A'_i,A'_j) \leftarrow S \times V(A_i,A_j)$ // S is scaling matrix with s as scaling factor
   End for
End

F. Results

For implementing the suggested SBT algorithm, iris dataset , containing 150 records. We have used Weka 3.6 for performing clustering operation. The proposed SBT algorithm is clustering algorithm independent. Here we have used k-means clustering algorithm.
a) Cluster distribution before transformation.

Figure 1- Cluster Distribution before transformation

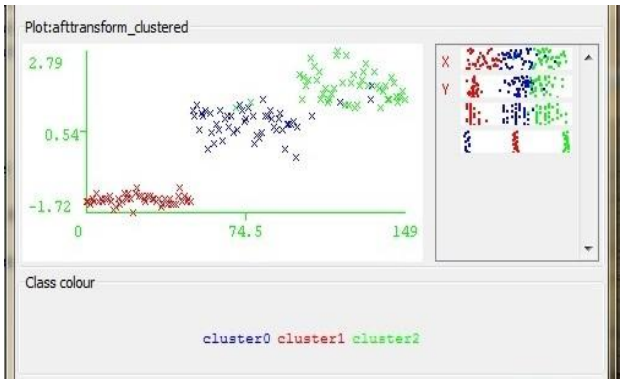


This output shows that 100 records belong to first cluster (cluster 0) and rest of 50 records belong to second cluster (cluster 1).

After this the transformed data set is supplied to Weka for k-Means clustering and the visualized output is as shown below.

b) Cluster distribution after transformation.

Figure 2- Cluster Distribution after transformation



Comparing Figure 1 and Figure 2 it is clear that the cluster distribution before and after transformation remains the same. Hence our procedure works effectively to maintain privacy for the confidential numeric data.

G. Security

The above stated procedure provides security to the numeric data. It means even if the standard deviation and mean of the numeric dataset is published then also the original numeric data of dataset before transformation cannot be interpreted correctly. This is accomplished mainly in two steps:

1) Data Camouflage: First we try to conceal raw data by normalization. Obviously it is not secure but it is beneficial in two ways a) It gives an equal weight to all attributes and b) It makes difficult the re-identification of objects with other datasets.

2) Attribute Distortion: By scaling two attribute values at a time attribute distortion is achieved. Doing so shifts the points to the new scale, which in turn will preserve the clusters of the points. Hence the clustering results will be similar before and after the application of SBT algorithm.

IV. CONCLUSION

In this paper , a scaling based transformation method has been introduced for Privacy Preserving Clustering on Horizontally Distributed Data. The proposed method is designed to preserve privacy only for numeric confidential data. This procedure also ensures the similar cluster distributions before and after transformation. This method is clustering algorithm independent. Moreover unsuccessful attempt is also made to recover original data from normalized data which ensures the security of data after transformation without changes in cluster distribution.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

Nowadays whatever data is required at particular site only that data is stored locally. So the complete dataset is stored in distributed manner. Doing so maintains the availability of data and also reduces the load of data server. In this context the proposed SBT algorithm will prove to be beneficial procedure for performing data mining operations along with preserving privacy of confidential data.

As a part of future work a secure algorithm can be developed for preserving privacy of arbitrarily distributed data. A general model needs to be developed which can be applied to every kind of data for preserving privacy and application of any data mining operations.

V. REFERENCES

[1] "Privacy Preserving KMeans Clustering over Vertically Partitioned Data" JaydeepVaidya, Chris Clifton in SIGKDD 2003.
[2] "Privacy Preserving SVM Classification on Vertically Partitiones Data " Hwanjo Yu, Jaideep Vaidya, Xiaqian Jiang.
[3] "Privacy Preserving Distributed DBSCAN Clustering" Jinfei Liu, Jun Luo, Joshua Zhexue Huang, Li Xiong in PAIS 2012, Berlin , Germany.
[4] "Privacy Preserving Clustering on Horizontally Partitioned Data" Ali Inan, Selim Kaya, Yucel Saygin, Erkay Savas, Ayca Hintoglu, Albert Levi , PDM 2006.
[5] "Privacy Preserving BIRCH Algorithm for Clustering over Arbitrarily Partitioned Databases" P. Krishna Prasad, C. Pandu Rangan, Springer-Verlag Berlin, Heidelberg 2007.
[6] "Cryptographic Techniques for Privacy Preserving Data Mining" , Benny Pinkas.
[7] A thesis on "Privacy Preserving Data Mining over Vertically Partitioned Data" , Jaideep Shrikant Vaidya.
[8] Privacy Preserving Clustering on Centralized Data through Scaling Transformation, Khatri Nishant P., Ms. Preeti Gupta, Tusal Patel.

BIOGRAPHY

Khatri Nishant is pursuing final semester of M.Tech. (Comp. Sci.) from Amity School of Engg & Tech,Amity University Rajasthan, Jaipur. His area of interest includes Database Systems, Data Mining.

Ms. Preeti Gupta is working as Assistant Professor in Computer Science & Engineering Department, Amity School of Engg & Tech, Amity University Rajasthan, Jaipur. She is pursuing her Ph.D. Her areas of interest includes Data Warehousing & Data Mining, DBMS, Compiler Design.

Tusal Patel is pursuing his final semester of M.Tech. (Comp. Sci.) from Amity School Of Engg & Tech, Amity University Rajasthan, Jaipur. His interest includes Database Systems, Data Mining, SAP HANA databases.