Using Privacy Preserving Techniques to Accomplish a Secure Accord

J.Indumathi, Dr.G.V.Uma

Department of Computer Science and Engineering,
Anna University,Chennai – 600 025.
Tamilnadu,India

Summary
Knowledge is supremacy and the more knowledgeable we are about information break-in, we are less prone to fall prey to the evil hacker sharks of information technology. The information technology is perpetually emerging and we are all beyond doubt minuscule gravels in a cosmic marine of information. Knowledge is pre-eminence, but as humble users of the most modern technologies we are pitted with possessions that may even make us paranoid concerning usage of a computer. Don't be discouraged though, as we will tell the techniques to protect ourselves while we are reaching an accord.
Privacy preserving data mining (i.e., is accomplishing valid data mining results devoid of the wisdom of the core data values) has been receiving attentiveness in the research society and beyond.Reaching a pact is defined as negotiation and the process ends in a ultimate contract between the participating parties. Automated Negotiation uses a wide variety of machine learning techniques.One foremost hindrance in automated negotiation is incorporating acumen into a computer system that carries out a negotiation [11], thus enabling the negotiation system to carry out automated negotiations efficiently and shrewdly on behalf of its clients**.** This paper presents an idea about the privacy preserving negotiation that is performed between individuals dealing with negotiating databases of users. It uses association rules and entropy discretization.As the BIRCH algorithms performance is extensive in terms of memory requirements, run time, clustering quality, stability and scalability it is used for data clustering and merging the data without disclosing the private and sensitive details of the user's data. The sensitive data are protected from the other user by using a protocol defined. It has lower communication complexity secure protocol; it will not disclose the private data's of the parties participating and it resists some degree of conspiracy and malevolent attacks. Entropy gives the information required in bits. After splitting of data, the insensitive data are fused and the results are obtained. This negotiation presented is of high order by using a low communication protocol between the users. The protocol used for the negotiation process provides the definite rules and conditions for merging the data tables.

Key Words
BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies),, *Clustering, Entropy-based Splitting, Negotiation.*

1. Introduction

Recently, there has been an escalating accent on exploratory scrutiny of very large datasets to discover useful patterns and/or correlations among attributes which is called as data mining. Privacy preserving data mining is accomplishing valid data mining results devoid of the wisdom of the core data values has been receiving concentration in the research society and beyond.

Negotiation is the process where interested parties resolve disputes, agree upon courses of action, bargain for individual or collective advantage, and/or attempt to craft outcomes which serve their mutual interests. Negotiation is usually regarded as a form of alternative dispute resolution. The first step in negotiation is to determine whether the situation is in fact a negotiation. The essential qualities of negotiation are: the existence of two parties who share an important objective but have some significant difference(s). The purpose of the negotiating conference is to seek compromise and iron out the difference(s). The outcome of the negotiating conference may be a compromise satisfactory to sides, a standoff (failure to reach a satisfactory compromise) or a standoff with an agreement to try again at a later time. Most important, resolving the dilemmas implicated is made up even more knotty by the fact that laws concerning breaches of confidentiality have not done a good job of setting limits on the database development being used to create predictive models.

Automated Negotiation uses a wide variety of machine learning techniques.One foremost hindrance in automated negotiation is incorporating acumen into a computer system that carries out a negotiation [11], thus enabling the negotiation system to carry out automated negotiations efficiently and shrewdly on behalf of its clients**.** A protocol is used to exchange insensitive information for the Negotiation and it defines some rules that are needed as the basics for the negotiation process to get completed without disclosing the private

Manuscript received August 5, 2007.
Manuscript revised August 20, 2007.

details of the users. The existing data clustering methods do not adequately address the problem of processing large datasets with a limited amount of resources (e.g., memory and CPU cycles). So as the dataset size increases, they do not scale up well in terms of memory requirement, running time, and result quality.

This paper presents an idea about the privacy preserving negotiation that is performed between individuals dealing with negotiating databases of users and at the same time, ensures privacy. It serves as a utility that uses a lower communication complexity secure protocol which will not disclose the private data's of the parties participating and to resist some degree of conspiracy and malevolent attacks. It uses association rules and entropy discretization.We aim to use the BIRCH algorithm for user's data clustering and merging the data without disclosing the private and sensitive details of the user's data. BIRCH performance is extensive in terms of memory requirements, run time, clustering quality, stability and scalability. Entropy gives the information required in bits. The sensitive data are protected from the other user by using a protocol defined in this paper. After splitting of data, the insensitive data are fused and the results are obtained. This negotiation is of high order by using a low communication protocol between the users. The protocol used for the negotiation process provides the definite rules and conditions for merging the data tables.

The remainder of this paper is organized as follows: Section 2 offers an overview of the related works in privacy preserving data mining, the different problems in data mining, the existing solutions, and our solution to the problem of automated negotiation.Section 3 discusses the problem statement, assumptions, notations used etc., for accomplishing a secure accord.Section 4 presents the block diagram, architectural diagram and the work flow architecture. Section 5 discusses the system architecture design, datasets used, User Interface design, and subsystem architecture. Section 6 discusses about the implementation of automated negotiation. Section 7 analyses the results and discusses the results. Section 8 concludes this paper with a brief summary and outlines the future research directions to be carried out.

2. Literature Survey

A prolific trend intended for prospective research in data mining will be the development of techniques so as to slot in privacy concerns [10] distinctively; we concentrate on the primary task of development of models about aggregated data, developing accurate models devoid of access to clear-cut information in individual data records. The fundamental assumption is that an individual will be prepared to selectively disclose information in exchange of value such models can provide [2].

Our central approach to privacy preserving is to allow users offer a modified value for sensitive attributes. Several modification techniques are available and few worth mentioning related to the automated negotiation scheme are: one using the homomorphic public key system of Secure Multiparty Computation-based methods [13], where only the results are revealed. However, it still left a potential privacy breach. Privacy concerns over the proliferation of gathering of personal information by various institutions over the internet led to the development of data mining algorithms that preserve the privacy of those whose personal data are collected and analyzed. A novel approach [11] to such privacy preserving data mining algorithms was proposed where the individual datum in a data set is perturbed by adding a random value from a known distribution. In these applications, the distribution of the original data set is important and estimating it.

Clustering is the process of discovering the groups of similar objects from a database to characterize the underlying data distribution. It has wide applications in market or customer segmentation, pattern recognition, biological studies, and spatial data analysis. Generally, clustering algorithms can be classified into four categories: partitioning-based, hierarchical-based, density-based, and grid-based. BIRCH [15] is a hierarchical clustering method that employs a hierarchical tree to characterize the nearness of data objects. BIRCH first scans the database to construct a clustering-feature (CF) tree to abridge the cluster depiction. Then, a selected clustering algorithm, such as K-means, is applied to the leaf nodes of the CF tree. For a large database, BIRCH can achieve good performance, scalability and effectiveness for incremental clustering of incoming data objects. The aim of these algorithms is the extraction of relevant knowledge from large amount of data, while protecting at the same time sensitive information. Several data mining techniques, incorporating privacy fortification mechanisms, have been developed that permit one to hide sensitive item sets or patterns, before the data mining process is executed.

Privacy Preserving Data Mining (PPDM) classification methods, instead, prevent a miner from building a classifier which is able to predict sensitive data. Additionally, privacy preserving clustering techniques have been recently proposed, which distort sensitive numerical attributes, while preserving general features for clustering analysis. A crucial issue is to determine which ones among these privacy-preserving techniques

better protect sensitive information. However, this is not the only criteria with respect to which these algorithms can be evaluated. It is also important to assess the quality of the data resulting from the modifications applied by each algorithm, as well as the performance of the algorithms. [3]

One of the approaches to achieve privacy preserving learning is to use Secure Multiparty Computation techniques. Yao suggested a general secure two-party function evaluation technique [1].Goldreich et al. extend this to any multiparty function [9, 8]. This generic method is based on representing each function as a Boolean circuit, and then the parties run a protocol for every gate in the circuit. However, as Goldreich points in [9], the communication complexity of this generic method depends on the size of the circuit that expresses the function to be computed, and using the solutions derived by this generic method for special cases of multiparty computation can be impractical. Therefore, recent research work is focused on developing efficient techniques for special cases. These include efficient privacy preserving techniques in secure cooperative statistical analysis [13], naive bayes classifier [7], association rules data mining [6, 4], decision tree [14] and clustering [5].

A privacy preserving negotiation learning scheme was proposed to [12] incorporate secure multiparty computation techniques into negotiation learning algorithms to allow negotiation parties to securely complete the learning process on a union of distributed data sets. This kind of privacy preserving learning in negotiation is used in parties to get more reasonable and accurate knowledge in various negotiation environments by borrowing experience from other parties (even if environments are strange to them).

We propose to use a lower communication complexity secure protocol which will not disclose the private data's of the parties participating and it resists some degree of conspiracy and malevolent attacks.

3. Problem Description for Accomplishing a Secure Accord

3.1. Problem Statement

Data clustering is an important technique for exploratory data analysis, and useful in many practical domains such as data classification and image processing. Privacy Preserving Clustering is merging the data of the users without disclosing the private and sensitive details of the users. The goals of Privacy preserving negotiation were to develop functionalities like User friendly negotiation, secure protocol for preserving private data's, Reusability, Portability.

3.2. Problem Description

A limitation of earlier work is that it uses a high cost secure communication protocol. Our work over comes this limitation by providing a low communication complexity protocol. Here we presented a solution for building a decision tree classifier on vertically partitioned private data. Two users have private data sets S_a, and S_b, and they want to build a decision tree classifier on [S_a S_b] (Vertically Partitioned) without disclosing their private data. This product will perform a join between all tables of the database. It will find the tables that are common and perform comparisons between their respective columns to find the differences i.e. which columns are common, which are not and what their differences are and merge them.

We examine privacy preserving learning in negotiation, which is a particular kind of cryptography problem. Data clustering identifies the spare, in large sets of Multi-Dimensional data points and is not uniformly occupied and the crowded places, and hence discovers the overall distribution patterns of the dataset. Besides, the derived clusters can be visualized more efficiently and effectively than the original dataset Permission to make a merging between the users is done without disclosing the sensitive data's. The Goal of BIRCH [1] stands for Balanced Iterative Reducing and Clustering using Hierarchies is to minimize data scans and work under memory constraint. The approach that is associated with BIRCH is to identify data points which are close and dense should be considered collectively instead of individually. These data which are closely related are identified and splitted using the Entropy-based Splitting. Privacy is today an important concern for both users and enterprises. Therefore, intense research is today being carried out on various aspects of privacy-preserving data management systems.

In this paper, an efficient and scalable data clustering method is proposed, based on a new in-memory data structure called CF-tree, which serves as an in-memory summary of the data distribution. We have implemented it in a system called BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and studied its performance extensively in terms of memory requirements, running time, clustering quality, stability and scalability. We also focused on database management systems (DBMS) able to enforce privacy promises encoded in privacy protocol.

This work can go as deep as the length of a column, or can be set to ignore the length, as well as ignore some columns, either by name or data type, that are already known to the user to be different. We are retrieving the metadata information using the catalog function. Some metadata information that are highly secured are held by the DBA's and those users of this product who try to maliciously retrieve data's should be tracked by the organization themselves and here we provide only the login and password for secured retrieval of data's. This product handles other security issues and this is completely organization oriented one.

3.3. Assumptions for Negotiation Scheme

The requirements of the negotiation protocol are as follows:

(i) Only active users will disclose their public details to the other user.

(ii) The users should not reveal their private details.

(iii) The protocol is designed to have less computation complexity.

(iv) It can also reduce the malicious attack.

Conditions to be satisfied by the learning algorithm and allow different negotiation parties to securely conduct the same learning technique on the union of their data sets:

(i) It is a distributed algorithm that can be conducted by multiple parties on distributed data sets;

(ii) All communications do not reveal any party's unnecessary information.

4. Proposed Architecture

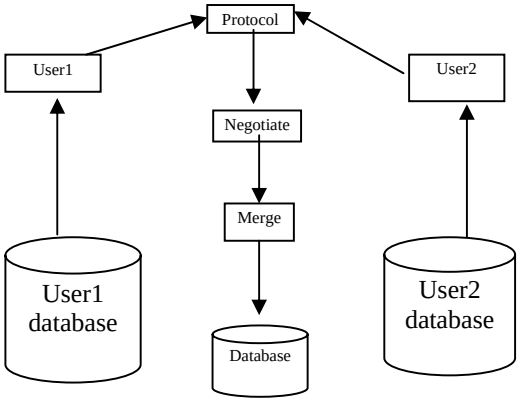


Figure 4.1 Proposed Architecture for accomplishing a secure accord

The proposed architecture as shown in figure 4.1 is for negotiation between two users. This can be extended upto several users. The two users have their own sets of databases. In their databases each one has both data which are private- sensitive and insensitive-public data. The main aim of this proposed product is to negotiate the details of data of the individual users.This is done by merging their data's in their databases without disclosing their private details. The users can access their data from the database after proper validation. After validation their data are clustered using BIRCH algorithm. They are then are splitted using the entropy based splitting. Now the negotiation is done with the help of the Yao rules protocol, after which the merging of the data takes place.

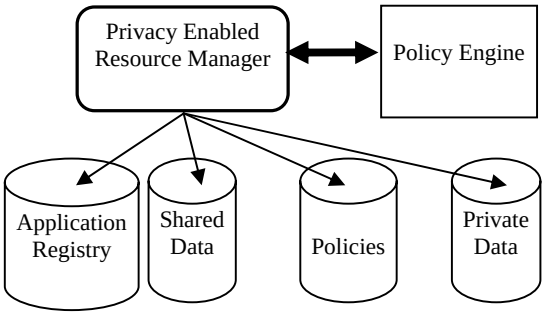


Figure 4.2 Structure of the Protocol

The protocol structure shown in figure 4.2 is defined for the negotiation process is named as Yao rules. It contains the privacy enabled resource manager which enables privacy preservation. It includes the application registry, shared data, policies and private data. The registry keeps track of application of the database that is to be negotiated. The shared data is the data which can be shared between the users. The private data is the data which should be protected.

BIRCH performance is extensive in terms of memory requirements, running time, clustering quality, stability and scalability. Entropy gives the information required in bits. Then the sensitive data are protected from the other user by using the protocol defined in this paper. After splitting of data, the insensitive data are merged and the results are obtained. The negotiation that is done is of high order by using a low communication protocol between the two users. The Yao protocol used for the negotiation process provides the definite rules and conditions for merging the data tables.

5. System Architecture Design

We address the problem of accomplishing a secure accord using privacy preserving technique and show that highly efficient solutions are possible. Our scenario is the following:
Let U_1 and U_2 be parties owning (large) private databases D_1 and D_2. The parties wish to apply a data-mining algorithm to the joint database $D_1[D_2$ without revealing any unnecessary information about their individual databases. That is, the only information learned by U_1 about D_2 is that which can be learned from the output of the data mining algorithm, and vice versa. We do not assume any "trusted" third party who computes the joint output.

5.1. Data structure design and datasets used

The proposed one deals with negotiating two databases about the virus and the disease caused by them in humans, rats and other living beings. This product will perform a join between all tables of the database. It will find the tables that are common and perform comparisons between their respective columns to find the differences i.e. which columns are common, which are not and what their differences are and merge them.

Module: Table Structure

This module will perform a join between all tables of the database. It will find the tables that are common and perform comparisons between their respective columns to find the differences i.e. which columns are common, which are not and what their differences are and negotiate them.

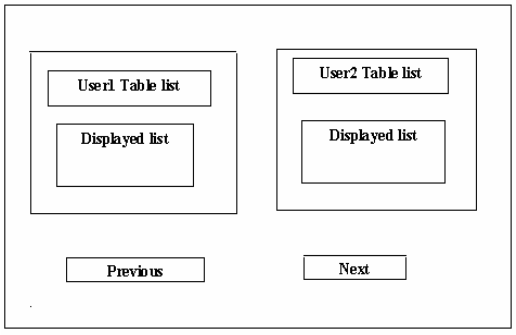


Figure 5.1.GUI

5.3.Subsystem Architecture

The following figure 5.2. shows us the Subsystem Architecture designed.

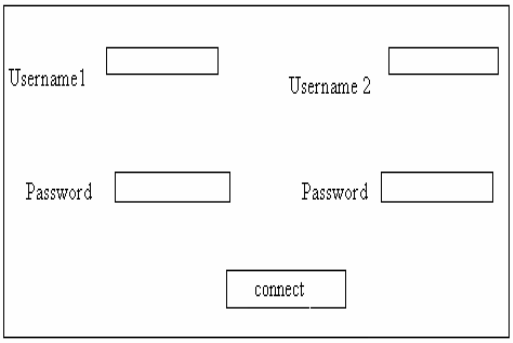
Criticality -This work can go as deep as the length of a column, or can be set to ignore the length, as well as ignore some columns, either by name or data type, that are already known to the user to be different.

The metadata information is retrieved using the catalog function.

Risks -Some metadata information that are highly secured are held by the DBA's and those users of this product who try to maliciously retrieve data's should be tracked by the organization themselves and here we provide only the login and password for secured retrieval of data's. This product does not handle other security issues and this is completely organization oriented one.

5.2. User Interface Design

The following figure 5.1. shows us the designed and used GUI product interface used by the user.



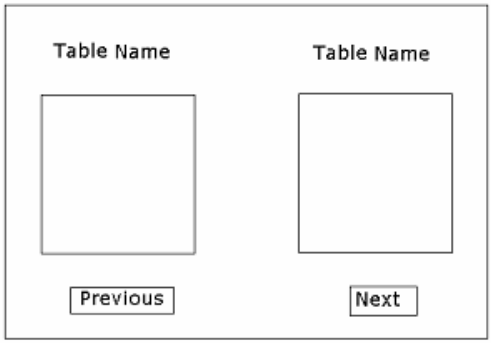


Figure 5.2.(a) Subsystem Architecture

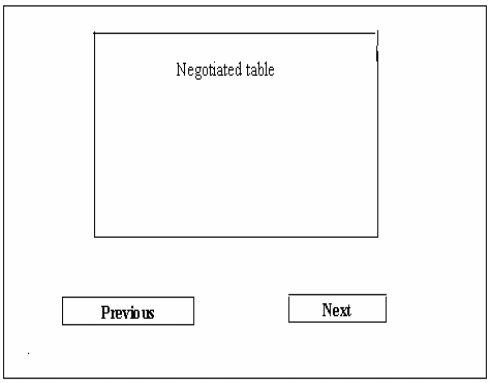


Figure 5.2. (b)Subsystem Architecture

6. Implementation

We implemented using the BIRCH algorithm for user's data clustering and merging the data without disclosing the private and sensitive details of the user's data. BIRCH performance is extensive in terms of memory requirements, run time, clustering quality, stability and scalability. We used a lower communication complexity secure protocol which will not disclose the private data's of the parties participating and it resists some degree of conspiracy and malevolent attacks. It uses association rules and entropy discretization.Entropy gives the information required in bits. The sensitive data are protected from the other user by using a protocol defined in this paper. After splitting of data, the insensitive data are fused and the results are obtained. This negotiation is of high order by using a low communication protocol between the users.

BIRCH algorithm

- Use CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering

Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data into sub-clusters that tries to preserve the inherent clustering structure of the data)

Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree

CF Additive Theorem

- Assume that $CF_1 = (N_1, LS_1, SS_1)$, and $CF_2 = (N_2, LS_2, SS_2)$ are the CF entries of two disjoint sub clusters.

- The CF entry of the subcluster formed by merging the two disjoint subclusters is:

$CF_1 + CF_2 =$

$(N_1 + N_2 , LS_1 + LS_2, SS_1 + SS_2)$

- The CF entries can be stored and calculated incrementally and consistently as subclusters are merged or new data points are inserted.

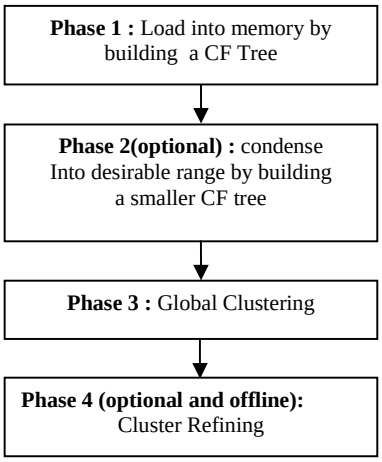
Phases of BIRCH Algorithm

- Phase 1 is to scan all data and build an initial in memory CF Tree using the given amount of memory and recycling disk space.

- Phase 2 is to condense into desirable range by building a smaller CF tree, for applying global or semi global clustering method.

- Phase 3 apply global or semi global algorithm to cluster all leaf entries.

- Phase 4 is optional and entails additional passes over data to correct inaccuracies and refines the cluster further.



Entropy-based Splitting

(i) Given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the entropy after partitioning is calculated.

(ii) The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.

(iii) The process is recursively applied to partitions obtained until some stopping criterion is met

Entropy splitting

Let

D = Database

Di = Sub-database or partition created through a decision (Example Age < 25)
P_{Di} = Probability of a transaction within D_i (or relative size of the partition). $P_{Di} = |D_i| / |D|$
B = Number of sub databases or partitions
i = Generic Counter variable
C = A set of all possible Classes. Example : C = {High risk = H, Low Risk = L }

K = Number of classes. Example : K = 2

C_i = Class i within the classification set C.

$$\text{Entropy (D)} = -\sum_{i=1}^{k} P_i \text{ Log } (P_i)$$

Where K = Number of classes.

P_i = Probability of a transaction being of a class 'i'

P_i = (Number of transactions in class 'i') / (Total number of entries)

Example : Entropy of database D

Given : Sample database D
D contains two types of classification C = {H , L}

$$\text{Weighted Entropy ("split")} = \sum_{i=1}^{B} \frac{|Di|}{|D|} \text{ entropy } (D_i)$$

Where: |D| = Number of transactions in Database D

|Di| = Number of transactions in sub database Di

B = Number of sub databases the Decision Criteria Splits the database 'D' into
"Split" = the decision criteria for how to subdivide the database 'D'

7. Results and Analysis

The main aim of the algorithm is to shun the security breaches and never allow a sequence of queries that compromises the data, regardless of the actual data. A simple classical cryptography mechanism is used for preserving the sensitive data from security breaches .A 'BRICH algorithm' provides a clustering work of data. Entropy-based splitting is used to reduce the data's. The two users have two different databases and depending upon the entropy splitting the attributes is categorized and the data's are reduced. The server hides the privacy of the sensitive data of the users. The private data's are checked and merged. The protocols used are well-defined Yao rules protocol. This protocol helps in less malicious attack and they exhibit less communication loss.

This work shows a scheme to achieve privacy preserving negotiation learning on the union two user's databases. This helps in easy usage in negotiating two databases with the help of a server The user types the username and password of the programmer (whose table are compared), and it gets validated .The user can chose either of the users created Table to list and their corresponding attributes (fields).The user can compare the table of the two users. Then we negotiate the tables as seen from figure given in appendix 10.1,10.2,10.3.

Figure 7.1 shows an increase in the performance of the data after negotiation The sensitive data that is to be protected in the negotiation process is given higher weight-age and is protected. The private data are not disclosed to other users,except the owner of the data. Figure 7.2 shows the performance that can be obtained using the protocol for the negotiation. The sensitive data that is to be protected in the negotiation process is said to be in large amount and also helps in providing less communication complexity.

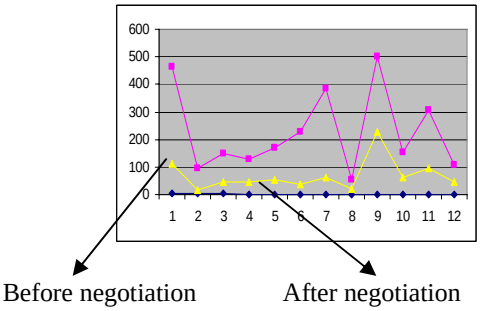


Figure 7.1 Time Vs. Performance of data before and after negotiation

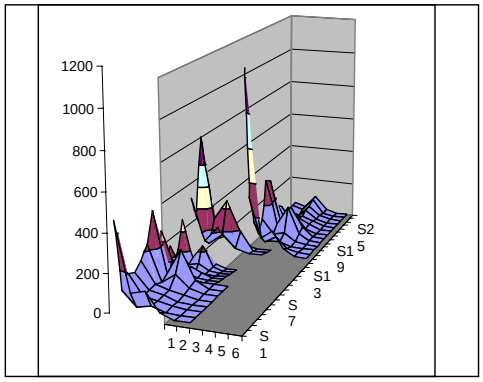


Figure 7.2 Performance of the protocol

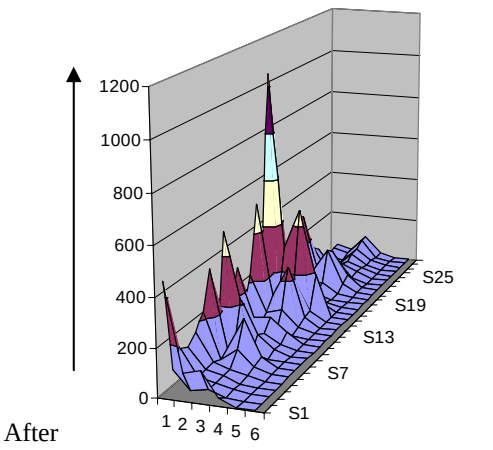


Figure 7. 3 Utility

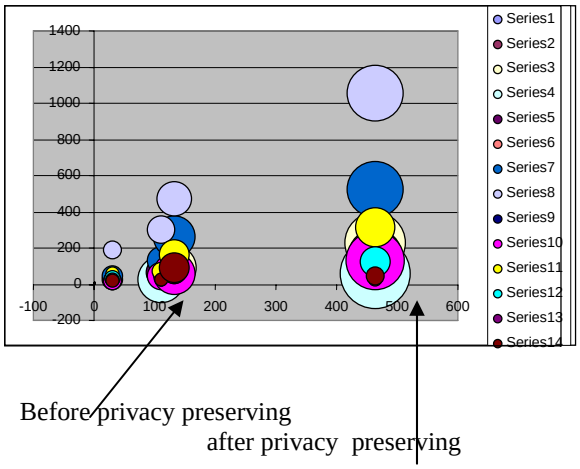


Figure 7.7 Sensitivity

The sensitivity and the utility of the data shows to increase after preserving the sensitive data which can be concluded from the graphs shown above after and before privacy preserving. With reference to the above figures 7.5, 7.6 and 7.7 we can visualise that there is an increase in the performance, sensitivity and utility of the preserved data. Thus we claim that this scheme provides a good way for preserving the private data and it shows that it may reduce data size and improve classification accuracy..This work is a utility that uses a lower communication complexity secure protocol which will not disclose the private data's of the parties participating.

8. Conclusion and Future Work

The future work is to develop a utility that is relatively having a lower communication complexity secure protocol designed with less communication loss for multiparty participating in the negotiation process and that which will not disclose the private data's of the multiple parties participating in the workspace. Designing the protocol to provide QoS along with the protection against the malicious attacks can be thought about. New algorithms can be designed and used for intrusion detection.

References

[1] A. Yao. How to generate and exchange secrets. In Proceeding of 27th IEEE Symposium on Foundations of Computer Science, 1986.

[2] A.F. Westin. Freebies and privacy: What net users think? Technical report, Opinion Research Corporation, July 1999. Available from http://www.privacyexchange.org/iss/surveys/sr990714.html.

[3] Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza,A Framework for Evaluating Privacy Preserving Data Mining Algorithms. Data Mining and Knowledge Discovery , Vol -11, pp- : 121 – 154,September 2005

[4] J. Vaidya and C. Clifton. Privacy Preserving Association Rule Mining In Vertically Partitioned Data.In Proceedings of the Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining,2002.

[5] J. Vaidya and C. Clifton. Privacy-Preserving K-Means Clustering Over Vertically Partitioned Data. In Proceedings of the Ninth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2003.

[6] M. Kantarcioglu and C. Clifton. Privacy-Preserving Distributed Mining Of Association Rules On Horizontally Partitioned Data. In the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 2002.

[7] M. Kantarcioglu and J. Vaidya. Privacy Preserving Naive Bayes Classifier For Horizontally Partitioned Data. In the IEEE ICDM Workshop on Privacy Preserving Data Mining,

[8] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game. In Proceedings of the 19th ACM conference on Theory of computing, 1987.

[9] O.Goldreich. Secure multi-party computation.Working Draft, 2000.

[10] Rakesh Agrawal. Data Mining: Crossing the Chasm. In 5th Int'l Conference on Knowledge Discovery in Databases and Data Mining,San Diego, California, August 1999. Available http://www.almaden.ibm.com/cs/quest/papers/kdd99chasm.ppt.

[11] S. Y. Su, C. Huang, J. Hammer, Y. Huang, H. Li, L.Wang, Y. Liu, M. Lee, and H. Lam. An Internet-Based Negotiation Server Fore-Commerce. VLDB, Vol-10, No.1, pp-72-90, 2001.

[12] Sheng Zhang and Fillia Makedon, Privacy Preserving Learning in Negotiation, 2005 ACM Symposium on Applied Computing, pp-821-825.

[13] W. Du and M. J. Atallah. Privacy-Preserving Statistical Analysis. In Proceedings of the 17th S. Matwin, T. Szapiro, and K. Haigh. Genetic algorithms approach to a negotiation support system. IEEE transaction on System, Man, and Cybernetics, Vol-21, No.-1, pp:102-114, 1991.

[14] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. Journal of Cryptology, 15(3):177{206, 2002.

[15] Zhang, T., R. Ramakrishnan, and M. Livny, BIRCH: An Efficient Data Clustering Method For Very Large Databases, SIGMOD, June 1996.

APPENDIX

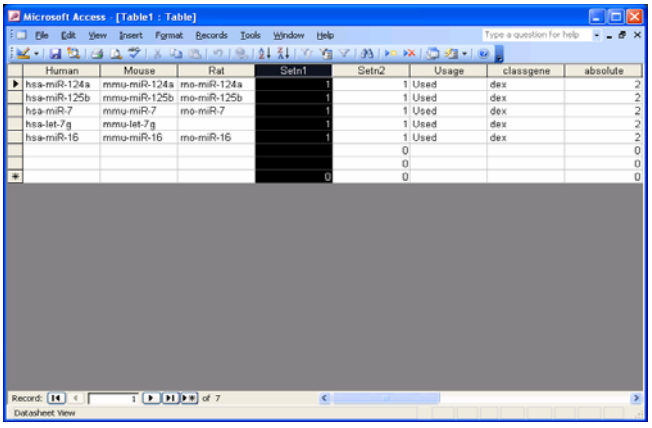


Figure 1 Dataminer1

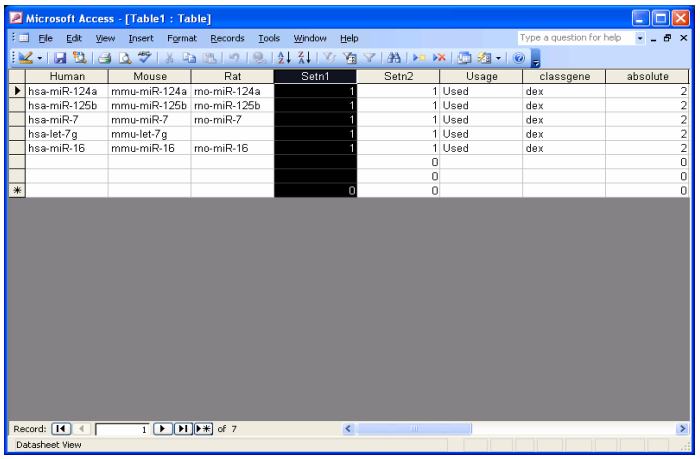


Figure 2 Dataminer2

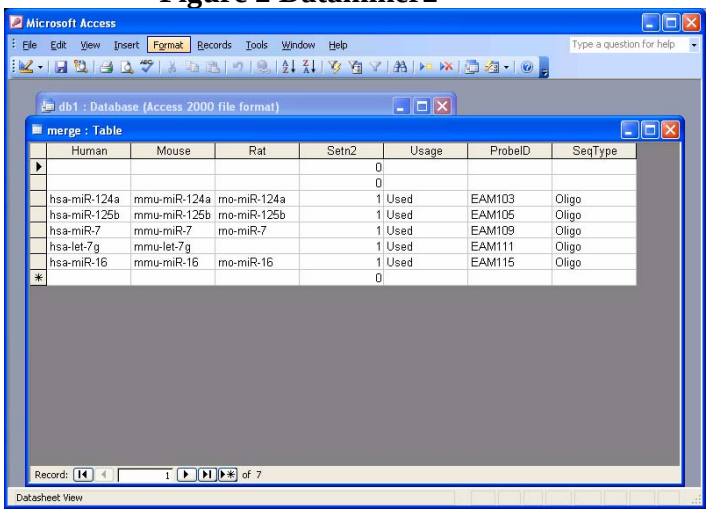


Figure 3 Merged dataset