A Model Based Framework for Privacy Preserving Clustering Using SOM

R.Vidyabanu
Department of Applied sciences
Sri Krishna College of Engineering and Technology
Coimbatore, Tamilnadu, India

Dr N.Nagaveni
Department of mathematics
Coimbatore Institute of technology
Coimbatore, Tamilnadu, India

ABSTRACT
Privacy has become an important issue in the progress of data mining techniques. Many laws are being enacted in various countries to protect the privacy of data.This privacy concern has been addressed by developing data mining techniques under a framework called privacy preserving data mining. Presently there are two main approaches popularly used -data perturbation and secure multiparty computation. In this paper we propose a technique for privacy preserving clustering using Principal component Analysis(PCA) based transformation approach. This method is suitable for clustering horizontally partitioned or centralized data sets .The framework was implemented on synthetic datasets and clustering was done using Self organizing Map(SOM). The accuracy of clustering before and after privacy preserving transformation was estimated.

Keywords
PCA, SOM, Rand Index, Transformation matrix.

1. INTRODUCTION
Data mining is a technology for identifying patterns and trends from large quantities of data. Huge volumes of detailed personal data are regularly collected and analyzed by applications using data mining. Such data include shopping habits, criminal records, medical history, credit records, among others[1]. On the one hand, such data is an important asset to business organizations and governments both to decision making processes and to provide social benefits, such as medical research, crime reduction, national security, etc [2]. On the other hand, analyzing such data opens new threats to privacy and autonomy of the individual if not done properly. The ease and transparency of information flow on the Internet has heightened concerns of personal privacy [3]. Web surfing, email, and other services constantly leak information about who we are and what we care about. Many have accepted that some privacy will be lost in exchange for the benefits of digital services. However, in other domains privacy is so important that its protection is federally mandated[4]. Technologies for protecting privacy are emerging in response to these growing concerns [5]. Recently, more emphasis has been placed on preserving the privacy of user-data aggregations, e.g., databases of personal information. Access to these collections is, however, enormously useful. It is from this balance between privacy and utility that the area of privacy preserving data-mining emerged [6].

The threat to privacy becomes real since data mining techniques are able to derive highly sensitive knowledge from unclassified data that is not even known to database holders. Worse is the privacy invasion occasioned by secondary usage of data when individuals are unaware of "behind the scenes" use of data mining techniques [7]. As an example in point, Culnan [8] made a particular study of secondary information use which she defined as "the use of personal information for other purposes subsequent to the original transaction between an individual and an organization when the information was collected." The key finding of this study was that concern over secondary use was correlated with the level of control the individual has over the secondary use. As a result, individuals are increasingly feeling that they are losing control over their own personal information that may reside on thousands of file servers largely beyond the control of existing privacy laws. This scenario has led to privacy invasion on a scale never before possible.

2. RELATED WORK
Some effort has been made to address the problem of privacy preservation in data mining. This effort has been restricted basically to classification and association rules. The class of solutions for this problem relies on data partition, data sanitization, randomization and data distortion. Estivill-Castro and Brankovic [9] introduced a method for ensuring partial disclosure while allowing a miner to explore detailed data. In this approach, one first builds a local decision tree over true data, and then swaps values amongst records in a leaf node of the tree to generate randomized training data. The swapping is performed over the confidential attribute only, where the confidential attribute is the class label. This approach deals with a trade-off: statistical precision against security level, i.e., the closer to the root, the higher the security but lower the precision.

Agrawal and Srikant [6] considered the case of building a decision-tree classifier from training data in which the values of individual records have been perturbed, by adding random values from a probability distribution. The resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution.In [10], the authors proposed a new algorithm for distribution reconstruction which is more effective than that proposed in [6], in terms of the level of information loss. This algorithm, based on Expectation Maximization (EM) algorithm, converges to the maximum likelihood estimate of the original distribution based on the perturbed data, even when a large amount of data is available. They also pointed out that the EM algorithm was in

fact identical to the Bayesian reconstruction proposed in [6], except for the approximation partitioning values into intervals. Evfimievski et al.[11] proposed a framework for mining association rules from transactions consisting of categorical items in which the data has been randomized to preserve privacy of individual transactions. The idea behind this approach is that some items in each transaction are replaced by new items not originally present in this transaction. In doing so, some true information is taken away and some false information is introduced, which seems to have obtained a reasonable privacy protection.

Recently, the data distortion approach has been applied to boolean association rules[12]. Again, the idea is to modify data values such that reconstruction of the values for any individual transaction is difficult, but the rules learned on the distorted data are still valid.To address privacy concerns in clustering analysis, we need to design specific data transformation methods that enforce privacy without losing the benefit of mining. The proposed data perturbation methods in the literature pertain to the context of statistical databases [13]. They do not apply to data clustering as they have limitations when the perturbed attributes are considered as a vector in the Euclidean space. R.Vaidya and Clifton's algorithm is based on the secure-permutation algorithm of Du and Atallah [14]. However, Vaidya and Clifton's algorithm has to execute Du and Atallah's protocol for every item in the data set. Therefore, their algorithm is not practical for large data sets.There are distributed clustering algorithms where the goal is to reduce communication costs[15]. These distributed clustering algorithms do not consider privacy. However, it will be interesting to investigate whether these algorithms can be made privacy preserving.

3. PRELIMINARIES

3.1. Principle Components Analysis (PCA)

PCA is used for transforming the multidimensional data in to lower dimensions. PCA assumes that all the variability in a process should be used in the analysis therefore it becomes difficult to distinguish the important variable from the less importantA data set $\mathbf{x}_i$, ($i = 1,\ldots,n$) is summarized as a linear combination of orthonormal vectors (called principal components):

$$f(\mathbf{x},\mathbf{V}) = \mathbf{u} + (\mathbf{x}\mathbf{V})\mathbf{V}^T$$

where $f(\mathbf{x},\mathbf{V})$ is a vector valued function, $\mathbf{u}$ is the mean of the data $\{\mathbf{x}_i\}$, and $\mathbf{V}$ is an $d \times m$ matrix with orthonormal columns. The mapping $\mathbf{z}_i = \mathbf{x}_i\mathbf{V}$ provides a low-dimensional projection of the vectors $\mathbf{x}_i$ if $m < d$.

The PCA estimates the projection matrix $\mathbf{V}$ minimizing

$$R_{emp}(\mathbf{x},\mathbf{V}) = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i - f(\mathbf{x}_i,\mathbf{V})\right\|^2$$

Principal components have the following optimal properties in the class of linear functions $f(\mathbf{x},\mathbf{V})$:

The principal components $\mathbf{Z}$ provide a linear approximation that represents the maximum variance of the original data in a low-dimensional projection.They also provide the best low-dimensional linear representation in the sense that the total sum of squared distances from data points to their projections in the space is minimized:If the mapping functions F and G are restricted to the class of linear functions, the composition $F(G(\mathbf{x}))$ provides the best (i.e., minimum empirical risk) approximation to the data.PCA most appropriate for normal / elliptical distributions (where linear PCA approach provides the best possible solution)Consequently, Principle Component Analysis (PCA) replaces the original variables of a data set with a smaller number of uncorrelated variables called the principle components.

3.2. The Self-Organizing Map (SOM)

The self-organizing map is a single layer feedforward network where the output syntaxes are arranged in low dimensional (usually 2D or 3D) grid. Each input is connected to all output neurons. Attached to every neuron there is a weight vector with the same dimensionality as the input vectors. The number of input dimensions is usually a lot higher than the output grid dimension. SOMs are mainly used for dimensionality reduction rather than expansion. The goal of the learning in the self-organizing map is to associate different parts of the SOM lattice to respond similarly to certain input patterns. This is partly motivated by how visual, auditory or other sensory information is handled in separate parts of the cerebral cortex in the human brain.It is trained using unsupervised learning to produce low dimensional representation of the training samples while preserving the topological properties of the input space.

The SOM algorithm

[1]. Randomize the map's nodes' weight vectors

[2]. Grab an input vector

[3]. Traverse each node in the map

[4]. Use Euclidean distance formula to find similarity between the input vector and the map's node's weight vector

[5]. Track the node that produces the smallest distance (this node will be called the Best Matching Unit or BMU)

[6]. Update the nodes in the neighbourhood of BMU by pulling them closer to the input vector

[7]. $Wv(t + 1) = Wv(t) + \Theta(t)\alpha(t)(D(t) - Wv(t))$

There are two ways to interpret a SOM. Because in the training phase weights of the whole neighborhood are moved in the same direction, similar items tend to excite adjacent neurons. Therefore, SOM forms a semantic map where similar samples are mapped close together and dissimilar apart. The other way to perceive the neuronal weights is to think them as pointers to the input space.

They form a discrete approximation of the distribution of training samples. More neurons point to regions with high training sample concentration and fewer where the samples are scarce.

3.3 Rand Index

In order to compare clustering results against external criteria, a measure of agreement is needed. Since we assume that each record is assigned to only one class in the external criterion and to only one cluster, measures of agreement between two partitions can be used.

The Rand index or Rand measure is a commonly used technique for measure of such similarity between two data clusters.

Given a set of n objects S = {O1, ..., On} and two data clusters of S which we want to compare: X = {x1, ..., xR} and Y = {y1, ..., yS} where the different partitions of X and Y are disjoint and their union is equal to S; we can compute the following values:

a is the number of elements in S that are in the same partition in X and in the same partition in Y,

b is the number of elements in S that are not in the same partition in X and not in the same partition in Y,

c is the number of elements in S that are in the same partition in X and not in the same partition in Y,

d is the number of elements in S that are not in the same partition in X but are in the same partition in Y.

Intuitively, one can think of a + b as the number of agreements between X and Y and c + d the number of disagreements between X and Y. The rand index, R, then becomes,

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

The rand index has a value between 0 and 1 with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

4. PROPOSED METHODOLOGY

In the proposed methodology the a PCA based transformation matrix is prepared from some of the randomly selected records from each cluster. We assume that the random samples will contain all kinds of data from the original data set. After preparing a transformation matrix using PCA, the transformation matrix was shifted by multiplying it with a constant. Further, the transformation matrix can be shifted by multiplying it with an arbitrarily selected shifting factor. This will further increase the security against any revere mechanism which can be used to guess the original data by doing some reverse transformation. This is used to transform the original data to a lower dimension. The dimension of the transformed data will be always less than that of original number of dimension and will be increasing with respect to dimensions of the data under consideration. SOM based clustering is applied to both original and transformed data and results are compared using Rand Index.

Fig 1 explains the process of privacy preserving transformation on original data by projecting it in to a lower dimension using the shifted transformation matrix

The Steps involved in Implementation and Evaluation

[1]. Prepare N Number of D dimensional synthetic data which belongs to C Number of classes using Gaussian distribution Function.

[2]. Randomly sample n number of data form N from all the classes.

[3]. Prepare a Transformation Matrix using PCA.

[4]. Shift the Transformation Matrix using a shift factor if necessary

[5]. Project the original data on the Transformation Matrix to produce the d dimensional data of the original N records.

[6]. Cluster the Original records using an unsupervised SOM Neural Network. This will give new class labels L1

[7]. Apply SOM clustering algorithm and classify the reduced dimensional data and this will give a set of new class labels L2.

[8]. Compare the Rand Index of the Class labels L1 and L2 with the original Class labels L and estimate the accuracy of calculation using Rand Index.

[9]. Repeat Evaluation with different parameters from Step 1.

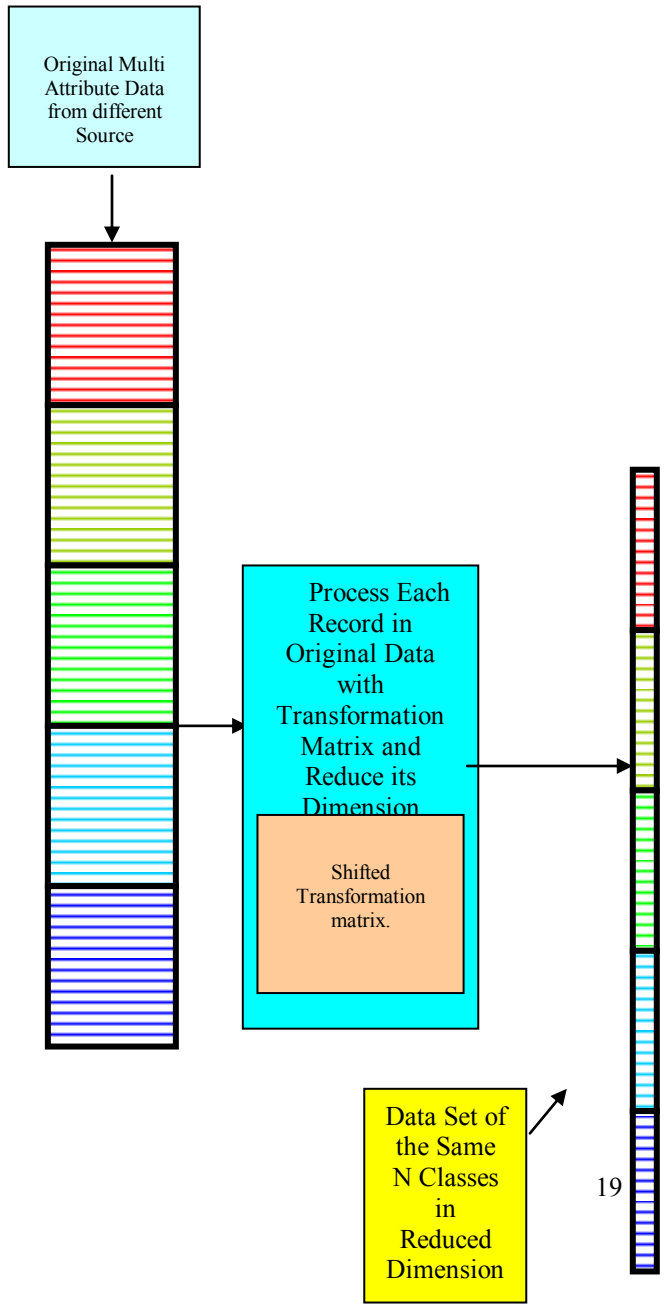

Figure 1: The Proposed Transformation Method

5. IMPLEMENTATION RESULTS

To evaluate the accuracy of the proposed privacy preserving transformation a large multidimensional data set is needed. Since the number of dimension is varied during evaluation, we proposed to use a synthetic data set in a very controlled manner for the creation of very fine well defined data clusters using Gaussian Distribution function.

Few sample records of the original dataset and the corresponding transformed records is shown below

3 records of the Original Dataset (dim=5)

52.00	59.00	42.00	67.00	28.00
54.00	52.00	58.00	68.00	22.00
52.00	51.00	51.00	61.00	19.00

3 records of the Transformed Dataset (dim=4)

-2.03	-0.93	0.98	-1.54
-3.44	1.02	-0.78	-1.33
-2.99	0.32	0.04	-1.66

After transformation dimension is reduced to 4.

Results with Synthetic Student Exam Result Datasets

The following table summarizes the results with respect to different number dimension of input records. During this evaluation, only 10% of the original records were used as a model to prepare the transformation matrix.

Total Number of groups/Clusters : 6 Nos

Total Number Students per Clusters : 200 Records/Cluster

Dimension/Attributes of Data : 2,3,4,5,6 & 7

Total Student Records : 1200 Records

Table 1 summarizes the results of clustering for Different Dimensions.Figure 2 shows the accuracy of results for different dimensions.

Table 1. Clustering with different dimensions

Sl No	Number of Dim of Input Data	Accuracy of SOM Based Clustering (Rand Index)	
		Original Dataset	Transformed Dataset
1	2	0.97	0.96
2	3	0.98	0.97
3	4	0.99	0.98
4	5	1.00	0.99
5	6	1.00	1.00
6	7	1.00	1.00

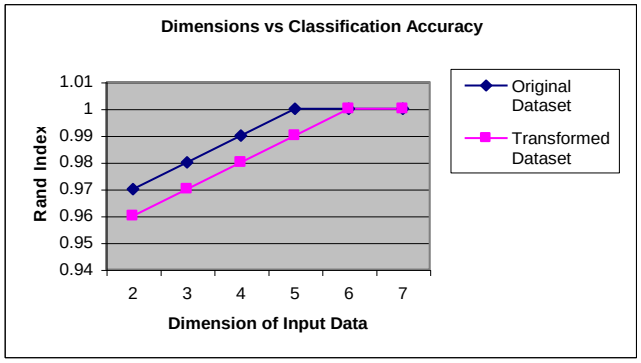

Figure 2: Accuracy of Clustering Before and After Transform

The accuracy increases in proportion to the number of dimensions.

Table 2 summarizes the results of clustering with different number of records.Figure 3 depicts the transformation time for different number of records.

Performance With Different number of Records

The following table shows the time taken for Transformation of Different number of Records only 10 % of the original Records were only used for Transformation.

Total Number of groups/Clusters : 6 Nos

Total Number Students per Clusters : 50-250 Records/Cluster

Dimension/Attributes of Data : 2,3,4,5,6 & 7

Total Student Records : 300, 600, 900, 1200 & 1500 Records

Table 2 : The Results Different Number of Records

Sl No	Total Number of Records	Time Taken For Transformation (sec)
1	300	0.047
2	600	0.093
3	900	0.156
4	1200	0.187
5	1500	0.219
6	1800	0.250

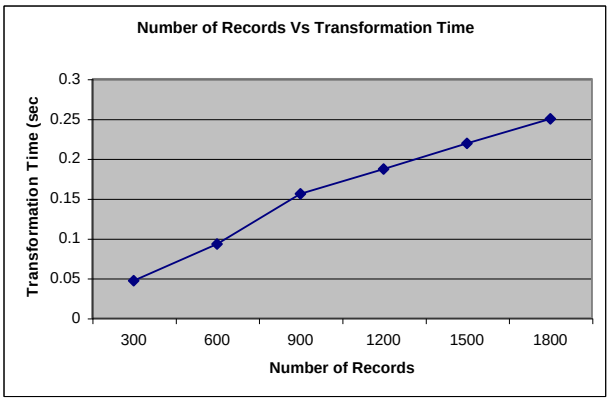

Figure 3: Performance with Different Number of Records

The above graph shows a linear increase of CPU time with respect to the increase of number of records used for transformation. But in a practical application, all the records will not be used for creating transformation matrix. Only a very small fraction of the original record set will be used. So we need not consider the performance lag with high number of records.

6. CONCLUSION

The proposed method has been successfully implemented using Matlab with synthetic datasets under windows xp. The results shows that the proposed method can be used to hide sensitive information. Some of the results of earlier works have shown, accuracy sometimes suffers as a result of security. But in the proposed method, the accuracy was almost equal to that of original data set. Further, if the input data is very noisy, then we may expect little bit improved accuracy with transformed data than the original data, since only the principal components are used to classify the data. Further, the proposed model can be used to multi party collaborative clustering scenario.

When presenting data as important as medical information that could potentially be used in the future to help save people's lives, it would seem logical that data should be mined as accurately as possible. These are issues that need to be worked out in the future. Privacy preserving data mining is by every means, a work in progress, and it will be interesting to see where new research on it leads in the following years.

7. REFERENCES

[1]. L. Brankovic and V. Estivill-Castro. Privacy Issues in Knowledge Discovery and Data Mining. In Proc. of Australian Institute of Computer Ethics Conference(AICEC99), Melbourne Victoria,Australia, July 1999.

[2]. P. Jefferies. Multimedia, Cyberspace & Ethics. In Proc. of International Conference on Information Visualisation (IV2000),pages 99-104,London, England, July 2000.

[3]. Lorrie Faith Cranor. Internet privacy. Communications of the ACM, 42(2):28–38, 1999. Stanley R.M Olivereira, Osmar R. Zailane, "Towards Standardization in privacy preserving data mining", The Privacy-preserving Data Mining: http://www.cs.ualberta.ca/~oliveira/psdm/psdm index.html.

[4]. 104th Congress. Public Law 104-191: Health Insurance Portability and Accountability Act of 1996, August 1996.

[5]. Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. W3C Recommendation,16 April 2002.

[6]. R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, pages 439–450, Dallas, TX, May 2000.

[7]. G. H. John. Behind-the-Scenes Data Mining. Newletter of ACM SIG on KDDM, 1(1):9–11, June 1999.

[8]. M. J. Culnan. How Did They Get My Name?: An Exploratory Investigation of Consumer Attitudes Toward Secondary Information. MIS Quartely, 17(3):341–363, September 1993.

[9]. Estivill-Castro and L. Brankovic. Data Swapping: Balancing Privacy Against Precision in Mining for Logic Rules. In Proc. of Data Warehousing and Knowledge Discovery DaWaK-99, pages 389–398, Florence, Italy, August 1999.

[10]. D. Agrawal and C. C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In Proc. of ACM SIGMOD/PODS, pages 247–255, Santa Barbara, CA, May 2001.

[11]. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy Preserving Mining of Association Rules. In Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining, pages 217–228, Edmonton, AB, Canada, July 2002.

[12]. S. J. Rizvi and J. R. Haritsa. Privacy-Preserving Association Rule Mining. In Proc. of the 28th International Conference on Very Large Data Bases, Hong Kong, China, August 2002.

[13]. K. Muralidhar, R. Parsa, and R. Sarathy. A General Additive Data Perturbation Method for Database Security. Management Science, 45(10):1399–1415, October 1999.

[14]. W. Du and M. J. Atallah. Privacy-preserving cooperative statistical analysis. In Annual ComputerSecurity Applications Conference ACSAC), pages 102–110, New Orleans, Louisiana,USA, December 10-14 2001.

[15]. I.S. Dhillon and D.S. Modha. A data-clustering algorithm on distributed memory multiprocessors.In Proceedings of Large-scale Parallel KDD Systems Workshop (ACM SIGKDD),August 15-18 1999.