2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)
19-20 December 2015, BUET, Dhaka, Bangladesh

# CBTS: Correlation Based Transformation Strategy for Privacy Preserving Data Mining

N P Nethravathi, Prasanth G Rao
Visvesvaraya Technological University
Belagavi-590 018,India.
Email: nethravishva123@gmail.com

P Deepa Shenoy, Venugopal K R
University Visvesvaraya College of Engineering,
Bangalore-560001, India.

Indramma M
BMSCE,
Bangalore-560019,India.

***Abstract*—Mining useful knowledge from corpus of data has become an important application in many fields. Data mining algorithms like clustering, classification work on this data and provide crisp information for analysis. As these data are available through various channels into public domain, privacy for the owners of the data is increasing need. Though privacy can be provided by hiding sensitive data, it will affect the data mining algorithms in knowledge extraction, so an effective mechanism is required to provide privacy to the data and at the same time without affecting the data mining algorithms. Privacy concern is a primary hindrance for quality data analysis. Data mining algorithms on the contrary focus on the mathematical nature than on the private nature of the information. Therefore instead of removing or encrypting sensitive data, we propose transformation strategies that retain the statistical, semantic and heuristic nature of the data while masking the sensitive information. The proposed Correlation Based Transformation Strategy (CBTS) combines Correlation Analysis in tandem with data transformation techniques such as Singular Value Decomposition (SVD), Principal Component Analysis (PCA) and Non Negative Matrix Factorization (NNMF) provides the intended level of privacy preservation and enables data analysis. The outcome of CBTS is evaluated on standard datasets against popular data mining techniques with significant success and Information Entropy is also accounted.**

***Index Terms*—Transformation Strategy, Privacy Preserving Data Mining, Correlation Analysis, Information Entropy**

## I. Introduction

Data mining is efficiently applied to many fields like clustering in bioinformatics, association rules in market basket analysis, classification in credit scoring, time series analysis in financial decision supporting. However, the increasing power of computers handling huge amount data and malicious usage made data mining a risk to privacy of individuals and companies. A simple example of privacy problem caused by combining information from different sites is as follows. In this given Zip codes of medical records are anonymized to protect disclosure of patient and information in personal website and address in yellow pages do not cause a privacy problem solely. However, a malicious internal human and hacker may combine the information in different sites and label medical record of patient.

Public sensitivity against data mining increased because it is seen as threat to individual's private information as shown in the example above. On the other hand, data mining is important for efficiently discovering knowledge. Privacy preserving data mining arise from the need for continue performing data mining efficiently but preserving private data or knowledge of individuals and companies. It is defined as data mining techniques that use specialized approaches to protect against the disclosure of private information may involve anonymizing private data, distorting sensitive values, encrypting data, or other means to ensure that sensitive data is protected.

Privacy is necessary to protect people's interests in competitive situations to avoid the negative and harmful consequences. People have the right to keep the aspects of their lives or behaviour confidential as personal information can be used in unethical ways causing much damage and humiliation. For instance there are strong reasons to keep patient records private as the social stigma attached to certain grievous diseases can hurt the patients sentiments and hinder the treatment. These concerns always come in way of quality data analysis as it prevents sincere research activities. PPDM is an area of data mining that works to protect the privacy in sensitive or confidential information from large collections of data. To achieve PPDM, the sensitive data is transformed to some other form, where in privacy is preserved and at the same time, the Data Mining Algorithms like Clustering , Classification and Association rule mining can work effectively on this transformed data.

In this paper, we propose a transformation method based on correlation analysis. The method is based on checking if the sensitive data can be removed and its statistical property can be retained in one or more non- sensitive data and if not transform the sensitive data. Depending on the correlation between attributes in the dataset, the method can give totally sensitive data removed dataset to sensitive data transformed data. We measure cluster Misclassification Error between the original dataset and the transformed dataset and prove the error is less in our approach.

## II. Related Work

Vassilios S[1] proposed the goal of PPDM to develop algorithms for modifying original data, so that private knowledge remains private even after the mining process. Researchers may use census, medical records, criminal records and it is often released for public welfare, which may threaten the existence of an individual or organization. The main concern of Privacy Preserving Data Mining is the sensitive nature

978-1-4673-8786-6/15/$31.00 ©2015 IEEE

of raw data. The data miner, while mining for numerical data, should not be able to access data in its original form with the entire condential nature. Boora et al. [2 ] proposed more robust techniques in Privacy Preserving Data Mining that intentionally alter the data to preserve sensitive information as well as to protect the inherent statistics of the data which is necessary for mining purpose.

Tianqing Zhu [3] proposed correlated differential privacy solution which enhances the privacy guarantee for a correlated dataset with less accuracy cost. Experimental results show the proposed solution outperforms traditional differential privacy in terms of Mean Square Error on large group of queries and it suggests that correlated differential privacy can successfully retain the utility while preserving the privacy.

Bharath K.Samanthula [4] focus on solving the Classication problem over encrypted data. The proposed protocol protects the confidentiality of data, privacy of users input query, and hides the data access patterns.

Ximeng Liu [5] proposed a new privacy preserving patient centric clinical decision support system, which helps clinician complementary to diagnose the risk of patients disease in privacy preserving way.

Zhiyuan Zhang [6] analyzed the correlations of numerical and categorical data on the correlation map.Yingpeng Sang et al. [7] explained how Simulation results show that reconstructions achieve high recovery rates, and outperform the reconstructions based on Principal Component Analysis (PCA).

Pui K. Fong in his work [8] proposed a method for Privacy Preserving Decision Tree Learning. In this approach original data samples are first converted to unreal datasets. They modified the ID3 decision tree algorithm to learn decision tree from the unreal datasets. The approach performs better only for distributed evenly. For uneven distribution, the storage requirement in this approach is high. For uneven distribution, the privacy is at risk.

Khaled Alotaibi in his work [9] proposed a non metric multidimensional scaling to transform the original dataset to transformed data. The transformed data was used to construct the SVM Classifier and accuracy was good. It was made possible by generation of higher generation feature space so that separation between the positive and negative classes were high. But the rank ordering in the perturbed data is not possible in this solution.

Augmented Rotation-Based Transformation was proposed in [10]. In this approach, the data is divided into many subsets row wise. The data is transformed by repeated rotation in such a way the distance between the data rows are invariant and it allows for clustering. The computation and storage overhead is very high in this approach. This approach can be used for iterative clustering with semi supervised active learning.

AA Hossain [11] proposed a sheer based privacy scheme for spatial dataset. In this method , the spatial transformation is done for location privacy in the dataset by pushing original location to new location with distance based on shearing factor values.

$$\acute{x} = F_x = \acute{A}x_l + (\acute{A}x_h - \acute{A}x_l).\frac{((x + \alpha y) - Ax_l)}{(Ax_h - Ax_l)} \quad (1)$$

$$\acute{y} = F_y = \acute{A}y_l + (\acute{A}y_h - \acute{A}y_l).\frac{((x\beta + y) - Ay_l)}{(Ay_h - Ay_l)} \quad (2)$$

After shearing in by distance rotation transformation is done. But in this method since the same transformation is applied on all dataset, even if one location is comprised, all the location can be compromised.

In [12], authors have proposed a randomized response method for distorting the original data before using frequent item set mining on the data. The data distortion method used here is probabilistic and when the number of attributes to be privacy preserved is higher, then error in item set mining is also higher.

In [13], privacy preserving clustering is done and for privacy the private attribute is split to multiple secrets. The clustering algorithm is then customized to work on these secret shares. By mapping a single attribute to multiple secret shares, the privacy is preserved. But the computation overhead is high in this approach for computing distance every time, the share reassembling by secure function is needed. If the distance computation overhead can be reduced, this method would work well for clustering and classification.

In [14], cryptographic technique using homomorphic encryption is used to transform data and then clustering is done on this private data. For non numeric data attributes the complexity in this approach is very high.

In [15], two additive perturbation algorithms RDD and RACC which combines additive perturbation with matrix multiplicative perturbation is proposed. The computation and reconstruction cost is less in this approach. But the distance distortion is high in this approach, so the clustering and classification accuracy will be affected.

## III. Problem Definition

Given complex data containing sensitive information, our solution must transform the data in such a way private and sensitive information are preserved. The level of privacy must also be fine tuned. Only private and sensitive information must be converted retaining the statistical and correlation structure in the document intact.

We consider two dimensional dataset, where in data set has many attributes and some of attributes are private and sensitive. The data set has many rows. The dataset is used by clustering algorithm like k-means to group the similar rows. The purpose of the project is to transform or remove the sensitive data in such way, the cluster misplacement error between cluster of original data and cluster of transformed data is as low as possible.

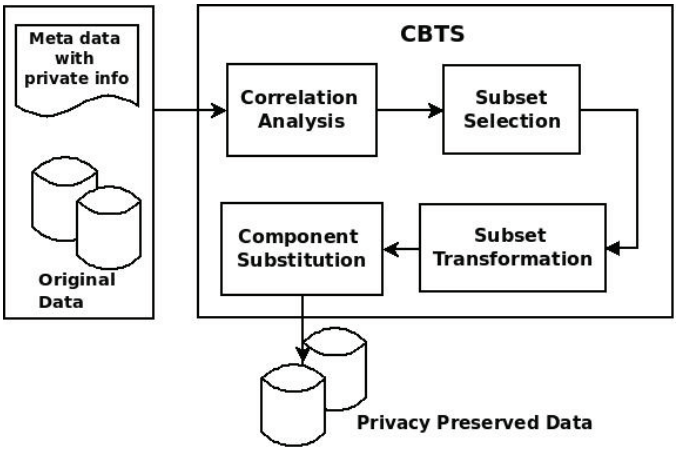


Fig. 1. Architecture of CBTS

## IV. ARCHITECTURE

Given complex data containing sensitive information, CBTS determines the subset of vectors correlated to sensitive data and generates equivalent components as substitutes. Correlation is computed using Pearson's correlation coefficient. The subsets formed are subject to transformation techniques that tend to converge on the observed similarity generating new components. Hence derived components are a mathematical reflection of the sensitive data and used instead of sensitive data for data mining. Existing transformation methods PCA, SVD and NNMF have been used prior in PPDM by [16][17][18] and demonstrate the required property of convergence. Fig.1 gives the overview of proposed algorithm. The framework takes data with private vector to sensitive information and a threshold suggesting the expected level of privacy conservation. The threshold is a normal value between 0 for maximum and 1 for minimum conservation. Our work concentrates on applying perturbation techniques on the correlated subsets of sensitive information.
It has four stages Correlation Analysis, Subset Selection, Subset Transformation and Component Substitution.

*1) Correlation Analysis ($D_c$) :* Correlation Matrix is computed using Pearson Coefficients. The correlation matrix is the fundamental in determining similarity among vectors, especially with the private vector. To do this the sensitive attributes in the data set must be provided to the system and correlation of each attribute X to each private attribute Y is computed using Pearson correlation.

$$\rho_{x,y} = \frac{con(X,Y)}{\sigma_x \sigma_y} \tag{3}$$

The correlation matrix is a $M * N$ vector where
$M$- the number of non private attributes
$N$- the number of private attributes

$$M = \begin{bmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ a_{m1} & a_{m2} & a_{m3} & ... & a_{mn} \end{bmatrix}$$

The matrix values $a_{mn}$ are normalized value from 0 to 1. 0 means no correlation and 1 means high correlation. Let the matrix to be transformed as

$$M = \begin{bmatrix} s_{11} & s_{12} & s_{13} & ... & s_{1y} \\ s_{21} & s_{22} & s_{23} & ... & s_{2y} \\ s_{x1} & s_{x2} & s_{x3} & ... & s_{xy} \end{bmatrix}$$

Where *Y* is the number of attributes and *X* is the total number of rows.

*2) Subset Selection (Private-vectors, Threshold) :* The idea in this method is that if a particular non sensitive attribute *X* is highly correlated with a sensitive attribute *Y* , the *Y* can be removed as *X* can compensate *Y* in case of classification or clustering tasks as the correlation and statistical property is still satisfied.

To find the best level or threshold for correlation we will start with a lower value and proceed till a best threshold for correlating the non sensitive and sensitive attribute is found.

Correlated vector subset satisfying the threshold is formed analysing the Correlation Matrix for each of the private vector. There are three possibilities that arise in the subsets.

1) The vectors are highly correlated in which case one of the non private vector can substitute the data
2) The vectors are correlated within the threshold bounds in which case the transformation can proceed.
3) No vectors are found for the threshold in which case threshold is lower incrementally till a subset is formed.

By this process highly correlated sensitive attributes are removed and now low correlated sensitive attributes are waiting for transformation which is done in next steps. As a result of this step if values $s_{xy} > T$ then attribute value *X* column is replaced for *Y* column in the original matrix as

$$M = \begin{bmatrix} s_{11} & s_{12} & s_{13} & ... & s_{1y} \\ s_{21} & s_{22} & s_{23} & ... & s_{2y} \\ s_{x1} & s_{x2} & s_{x3} & ... & s_{xy} \end{bmatrix}$$

Here in this matrix $s_{13} > T$ and 1 is non private and 3 is private

*3) Subset Transformation (Subset, Transform):* From the remaining sensitive non correlated data *Y*, the subsets of sensitive data is formed using methods like PCA, SVD or NNMF. Any of these three methods can be used as our work is not dependent of any method. The result of this step is set of components forming the candidates for substitution. In an ideal scenario there shall be exactly one component of convergence.

*4) Component Substitution (Private-vector, Component):* Private Vector is substituted with the most similar component derived from subset transformation and Entropy is computed for the perturbed data set against the original data. Entropy computed is given by Shannon Information Entropy. So if *Y* is private attribute but we cannot find a single non private attribute but able to find a subset of attributes $[a_{1b}a_{1c}a_{1d}]$ to

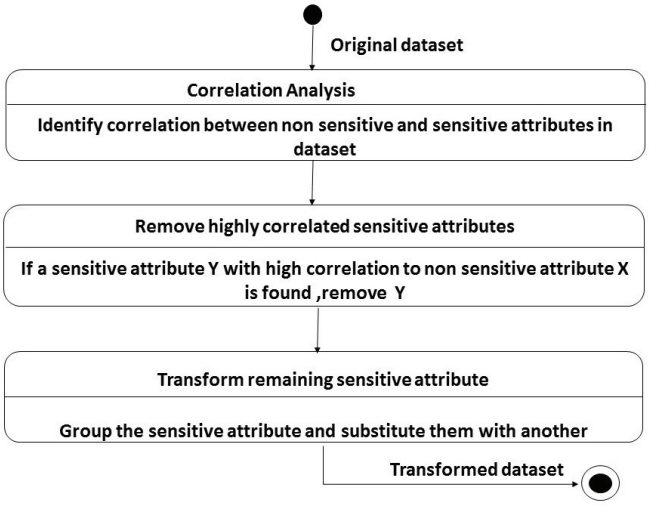


Fig. 2. Transformation Method

be correlated to *Y*, the *Y* will be replaced with a composite of $[a_{1b}a_{1c}a_{1d}]$

$$M = \begin{bmatrix} & & \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \\ s_{x1} & s_{x2} \end{bmatrix} \\ s_{11} & s_{12} & \\ s_{21} & s_{22} & \\ s_{x1} & s_{x2} & \end{bmatrix}$$

The overall process in our transformation method is given in Fig.2.Transformation Method.

## V. Result

TABLE I
Comparison of Information Entropy

| Types of data | Original Entropy | Information Entropy(IE) (Using CBTS Method) / (Using existing methods) | | |
|---|---|---|---|---|
| | | PCA | SVD | NNMF |
| Ionosphere (351x35) | 9.8042 | 10.2250/ 10.236 | 10.2196/ 13.583 | 10.1828/ 2.0047 |
| Cancer (699x11) | 2.5663 | 2.9818/ 6.3399 | 2.9174/ 2.0807 | 2.7794/ 0.7634 |
| Vehicle (846x18) | 7.9660 | 8.3148/ 8.3399 | 8.1252/ 13.8944 | 7.8624/ 4.3333 |
| Letter (20,000x16) | 3.5403 | 3.9585/ 8.0001 | 3.6655 / 8.2666 | 3.4617/ 1.8046 |

Information Entropy of original data against perturbed data using CBTS with transformation methods is summarized in Table I. We can infer from the table that deviation in Information Entropy is minimum using proposed CBTS method against using transformation techniques alone. Table II gives the comparison of classifier accuracies for various machine learning algorithms using CBTS against original data. It is clearly observable from the results the classifier performance is comparable to the original data. Table III shows the Misclassification Error $M_E$values with k-means clustering. Higher $M_E$ values indicates lower clustering quality where as Lower $M_E$ values indicate the higher utilization of the data.

TABLE II
Comparison of various Machine Learning Algorithms Using CBTS

| Dataset | Machine Learning Algorithm | Observed Classifier Accuracy (%) | | | |
|---|---|---|---|---|---|
| | | Actual Data | Transforming using CBTS | | |
| | | | PCA | SVD | NNMF |
| Ionosphere | Decision Tree | 99.71 | 98.86 | 98.86 | 97.72 |
| | Multilayer Perceptron | 99.71 | 99.43 | 99.14 | 99.71 |
| | Naive Bayes | 82.9 | 78.91 | 79.77 | 83.76 |
| Breast Cancer | Decision Tree | 97.99 | 97.99 | 97.42 | 98.28 |
| | Multilayer Perceptron | 99.14 | 98.56 | 98.99 | 98.71 |
| | Naive Bayes | 96.13 | 96.28 | 96.28 | 96.28 |

TABLE III
Cluster Misclassification Error ($M_E$)

| Types of data | Clusters (K) | $M_E$ (with CBTS) | | | $M_E$(without CBTS) | | |
|---|---|---|---|---|---|---|---|
| | | PCA | SVD | NNMF | PCA | SVD | NNMF |
| IONOSPHERE (351x35) | 2 | 0.573 | 0.011 | 0.006 | 0.011 | 0.182 | 0.217 |
| | 3 | 0.028 | 0.0798 | 0.017 | 0.519 | 0.387 | 0.325 |
| | 4 | 0.573 | 0.091 | 0.051 | 0.593 | 0.558 | 0.279 |
| | 5 | 0.04 | 0.068 | 0.023 | 0.558 | 0.792 | 0.342 |
| BREAST CANCER (699x11) | 2 | 0.009 | 0 | 0.003 | 0.037 | 0.009 | 0.023 |
| | 3 | 0.037 | 0.009 | 0.006 | 0.26 | 0.266 | 0.532 |
| | 4 | 0.063 | 0.006 | 0.057 | 0.718 | 0.243 | 0.389 |
| | 5 | 0.069 | 0.132 | 0.069 | 0.741 | 0.04 | 0.252 |

$$M_E = \frac{1}{N}\sum(|Cluster_i(D)| - |Cluster_i(D')|) \quad (4)$$

The clustering quality provided by the proposed CBTS is higher compared to existing methods.

## VI. Conclusion and Future Work

The present work explores CBTS for numerical data. The proposed method was able to remove the highly correlated sensitive data and transform the non correlated sensitive data.Through experiment analysis we have proved that our proposed dataset transformation method has low clustering misplacement error. In future a more generic CBTS needs to be devised to address complex and ordered datasets. Private vector instances can be identified from the rich semi structured metadata and entity relations in databases but requires Semantic Transformation Strategies (STS) to intelligently sense private vectors based on context inputs from the DB administrator. STS with Generic CBTS can provide in future a simplified way to anonymize and publish quality data for productive mining.

## References

[1] Vassilios S. Veryhios, Elisa Bertino, Igor Nai Fovino Loredana Parasiliti Provenza, Yucel Saygin, Yannis eodoridis, *"State-of-the-art in Privacy Preserving Data Mining"*, SIGMOD Record, Vol. 33, No.1, March 2004.

[2] R. K. Boora, R. Shukla, and A. K. Misra, *"An Improved Approach to High Level Privacy Preserving Itemset Mining"* ,USA, no. arXiv:1001.2270. Volume 6. NO. 3.pp. 216-223, ISSN 1947 5500, Jan 2010. [Online]. Available: http://cds.cern.ch/record/1233468

[3] T. Zhu, P. Xiong, G. Li, and W. Zhou,*"Correlated differential privacy: Hiding information in non-iid data set"*,Information Forensics and Security, IEEE Transactions on, vol. 10, no. 2, pp. 229-242, Feb 2015.

[4] B. Samanthula, Y. Elmehdwi, and W. Jiang, *"k-nearest neighbor classification over semantically secure encrypted relational data"*, Knowledge and Data Engineering, IEEE Transactions on, vol. 27, no. 5, pp. 1261-1273, May 2015.

[5] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin,*"Privacy-preserving patient-centric clinical decision support system on naive bayesian classification"*,Biomedical and Health Informatics, IEEE Journal of, vol. PP, no. 99, pp. 1-1, 2015.

[6] Z. Zhang, K. McDonnell, E. Zadok, and K. Mueller, *"Visual correlation analysis of numerical and categorical data on the correlation map"*,Visualization and Computer Graphics, IEEE Transactions on, vol. 21, no.2, pp. 289-303, Feb 2015.

[7] Y. Sang, H. Shen, and H. Tian, *"Effective reconstruction of data perturbed by random projections"*,Computers, IEEE Transactions on, vol. 61, no. 1, pp. 101-117, Jan 2012.

[8] *"Privacy Preserving Decision Tree Learning Using Unrealized Data Sets"* Pui K. Fong IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 2, February 2012.

[9] *"Privacy-Preserving SVM Classification using Non-metric MDS SECURWARE 2013 "*: The Seventh International Conference on Emerging Security Information, Systems and Technologies.

[10] Dowon Hong and Abedelaziz Mohaisen *"Augmented Rotation-Based Transformation for Privacy-Preserving Data Clustering"* ETRI Journal, Volume 32, Number 3, June 2010

[11] AA Hosain *"Shear-based Spatial Transformation to Protect Proximity Attack in Outsourced Databae"* IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2013

[12] Chongjing Sun, Yan Fu, Junlin Zhou, and Hui Gao*"Personalized Privacy-Preserving Frequent Itemset Mining Using Randomized Response"* The Scientific World Journal March 2014.

[13] Maneesh Upmanyu, Anoop M. Namboodiri, Kannan Srinathan, and C.V. Jawahar *"Efficient Privacy Preserving K-Means Clustering"* Springer-Verlag Berlin Heidelberg 2010

[14] Zekeriya Erkin : *"Privacy-preserving distributed clustering"*. EURASIP Journal on Information Security 2013.

[15] Likun Liu *"Using Noise Addition Method Based on Pre-mining to Protect Healthcare Privacy CEAI"*, Vol.14, No.2, pp. 58-64, 2012

[16] S. Patel and K. R. Amin, *"Privacy Preserving Based on PCA Transformation using data perturbation technique"*,International Journal of Computer Science Engineering Technology, vol. 4, no. 35, pp. 477-484, 2013

[17] S. Xu, J. Zhang, D. Han, and J. Wang,*"Singular Value Decomposition based data distortion strategy for privacy protection"*, Knowledge and Information Systems, vol. 10, no. 3, pp. 383-397, 2006.

[18] J. Wang, W. Zhong, J. Zhang, and S. Xu, *"Selective data distortion via structural partition and ssvd for privacy preservation"*, in IKE. Citeseer, 2006, pp. 114-120.