International Journal of Computer Applications (0975 - 8887)
Volume 157 - No.1, January 2017

Generic CBTS: Correlation based Transformation Strategy for Privacy Preserving Data Mining

N. P. Nethravathi, Prasanth G. Rao
Visveswaraya Technological University,
Belagavi, India

Chaitra C. Vaidya,
P. Deepa Shenoy, Venugopal K. R.
UVCE, Bangalore University,
Bangalore, India

Indiramma M.
BMS College of Engineering
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 results. 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 proposed technique will work for numerical, ordinal and nominal data. The outcome of CBTS is evaluated on standard datasets against popular data mining techniques with significant success and Information Entropy is also accounted.

Keywords

Transformation Strategy, Privacy Preserving Data Mining, Correlation Analysis, Information Entropy

1. INTRODUCTION

Data mining is efficiently applied to many fields like Customer Relationship Management (CRM), fraud detection (traditional methods are time consuming); improve intrusion detection by adding a level of anomaly detection, analysis in financial banking. However, the increasing habit of using computers for handling huge amount of data and malicious files made data mining a risk to privacy of people and corporations. A sample of privacy problem is as follows. Ashley Madison, a dating website primarily meant to help people have extramarital affairs, was recently threatened by a hacker group called "The Impact Team" When the company did not comply, the hackers publicly released the private information of over 30 million Ashley Madison users online (names, addresses, credit card information and more). Those users now live in fear of being publicly shamed on the web which could have a massive impact on their personal lives. While on the surface this may seem like a deserved day of reckoning for exposed adulterers, it speaks to a much larger privacy issue that concerns everyone. We live our entire lives online, and our actions there hinge on the promise of privacy. We believe that what we buy, where we bank, what we research, and even who we date should be private. We should all be terrified at the idea that a single group or person can decide to compromise that promise for their own personal agenda. Public awareness against data mining has elevated because it is seen as a peril to a person's private data as shown in the example above. On the contrary, data mining is important for productive knowledge discovery. Privacy conservative data mining arise from the demand for carrying out data mining effectively simultaneously protecting private information or knowledge of people's and organizations. It is interpreted as art of data mining that use particular path to prevent leakage of personal data which may include anonymizing private data, deceiving sensitive values, encrypting data, or other mediums to make sure that delicate data is protected. Privacy is needed to protect people's interests in competitive environment to dodge the bad and harmful effects. Individuals have the right to keep the aspects of their lives confidential, as individuals private information can be used to cause harm and humiliation. For instance there are multiple reasons to keep patient records isolated as the social stigma attached to certain grievous diseases can ruin the patient's sentiments and interrupt the treatment. These issues always hinder the quality data analysis as it prevents sincere research activities.

PPDM is an area of data mining that works to protect the privacy for 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 transformed data. In this paper, we also propose a CBTS which can be applied to numerical, ordinal and nominal values. We describe a technique to convert the ordinal and nominal data to numerical data since CBTS works only for numerical data. The CBTS can be applied directly on the numerical data. We measure the Information Entropy values of both Original data and Transformed data and the results are comparable and also we measure Cluster Misclassication Error and prove the error is less in our approach. The paper is organized as follows: Section II describes Related Work. Section III explains Problem Definition. Architecture is presented in Section IV. Result is discussed in Section V. We conclude this paper with future work in Section VI.

2. RELATED WORK

Vassilios S.et al. [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 of raw data. The data miner, while mining the numerical data, should not be able to access data in its original form with the entire confidential nature. Vijayarani S. et al. [2] proposed a technique called modified data transitive technique in which the sensitive numerical data item is to be protected by modifying the original data item. There is a comparison between the modified data transitive technique and the perturbative masking techniques such as additive noise, rounding and micro aggregation and performances are analyzed and results are drawn by concluding with the satisfactory results using the transitive techniques.

The authors of [3] 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 et al. [4] 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 et al. [5] focus on solving the classification problem over encrypted data. The proposed protocol protects the confidentiality of data, privacy of users input query, and hides the data access patterns. The authors of [6] 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 et al [7] analyzed the correlations of numerical and categorical data on the correlation map. Paper [8] explains how Simulation results show that reconstructions achieve high recovery rates, and outperform the reconstructions based on Principal Component Analysis (PCA).

Fong P.K. et al. [9] 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 data distributed evenly. For uneven distribution, the storage requirement in this approach is high. For uneven distribution, the privacy is at risk. The authors of [10] 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 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 [11]. In this approach, the data is divided into many subsets row wise. The data is transformed by repeated rotation in such a way that 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 [12] 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. After shearing in by distance, rotation transformation is done. But in this approach since the same transformation is applied on all dataset, even if one location is compromised, all the location can be compromised. In paper [13], 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 paper [14], privacy preserving clustering is done and for privacy the private attribute is split into 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. The authors of [15] proposed a cryptographic technique using homomorphic encryption which 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 paper [16], 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. As explained in [17] micro data is usually stored in the form of table where each row represents an individual. Here the table has three types of attributes (1) Identity attribute (To uniquely identify an individual like name) (2) Quasi identifier (which includes demographic attributes) (3) Sensitive attributes (which include confidential information like diseases). Quasi identifier attributes may be merged with other public databases to uniquely identify the individual and their sensitive data (Linking attack).

3. PROBLEM DEFINITION

In this paper, the data is assumed to be a matrix D_{pq}, where each of the p rows is an observation p_i, and each observation contains values for each of the q attributes q_i. The matrix D_{pq} may contain ordinal, nominal and numerical attributes. However, our PPDM rely on numerical, ordinal and nominal attributes. Thus, the pXq matrix is subject to transformation. Transformation has the following two steps.
Step1: CBTS is directly applied on the data if it is numerical data.

Step 2: If the data is ordinal or nominal first we convert the data to numerical and then we apply CBTS.
The objective of this paper is to transform the numerical, ordinal and nominal sensitive data in such a way that the correctly classified instances, the decision trees and Information Entropy of original data and transformed data are almost comparable.

4. ARCHITECTURE

4.1 Numerical Data

CBTS takes input data and the metadata of the data that specifies the private columns in the data. Each private column specified in the metadata is processed individually. We use selected private column to refer the private column which is being processed currently. For each selected private column, subset of highly correlated columns are selected. Selected subset is perturbed using existing perturbation techniques. Component of the perturbed data is substituted in place of the selected private column. CBTS is divided into four major stages; they are Correlation Analysis, Subset Selection, Subset Transformation and Component Substitution.

4.2 Ordinal Data

This is the type of data which contains many categories and these categories have intrinsic ordering. For example, the economic status has three categories low, medium and high which have some intrinsic ordering.

4.3 Nominal/Categorical Data

This is the type of data where there is no intrinsic ordering between the categories. For example, the gender has two categories male and female and there is no intrinsic ordering to the categories. The conversion step has two sub-steps. Initially, the dataset is parsed to extract the unique data values in each column which is given to next step. In the next step, based on the data type of the column conversion is done. If the column has ordinal data values, they are converted to numerical data values given by the user. For the columns with categorical data values, correlation coefficient is calculated. If there exists a strong correlation, then they are converted to random numbers. If the correlation is weak, then the conversion is done by substituting categories with close ranged numbers.
Here we are proposing a CBTS for Ordinal and Nominal Data and the Architecture is shown in Figure. 1. Our method first converts the ordinal and nominal data to equivalent Numerical Data.

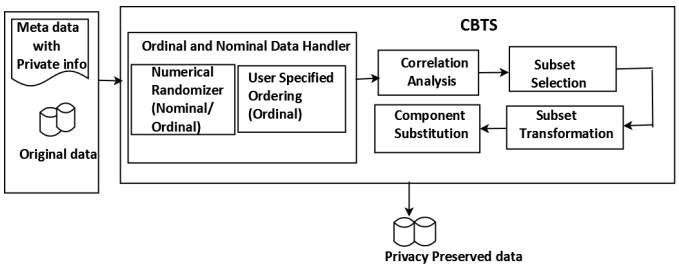

Fig. 1: Architecture of CBTS for Ordinal and Nominal Data

The conversion is based on the type of data.

Chi-squared test is done to determine the correlation between categorical data values. The value of the test-statistic is

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{1}$$

where,
O_k - Observed frequency
E_k - Expected frequency
Figure. 1. shows the architecture of CBTS for ordinal and nominal data. Here the input to the CBTS block is the numerical attribute received by the conversion of Ordinal and Nominal data to numerical. This method was able to remove the highly correlated sensitive data and transform the non correlated sensitive data. The CBTS is applied to datasets which has numerical values and the information entropy values are calculated for the original data and the transformed data and the results are obtained. Through the experiment analysis it was proved that proposed dataset transformation method has low clustering misplacement error.
Given complex data D_{pq} 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 [18][19][20] and demonstrate the required property of convergence. CBTS 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.

4.3.1 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 dataset must be provided to the system and correlation of each attribute *X* to each private attribute *Y* is computed using Pearson correlation.

$$\rho_{x,y} = \frac{con(X,Y)}{\sigma_x \sigma_y} \tag{2}$$

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

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

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

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

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

4.3.2 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*, then *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

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

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

4.3.3 Subset Transformation (Subset, Transform). From the remaining sensitive non correlated data *Y*, the subset 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.

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

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

Algorithm 1 for conversion of categorical data to numerical data.

INPUT: Original dataset DS with ordinal,nomi nal and numerical data.
OUTPUT: Converted dataset DS' with only numerical data.
(1) Read the dataset
(2) Parse the file column wise and extract unique values from each.
(3) Repeat step 4 for each column.
(4) Based on the type of data in each column, convert the data. If the column under consideration has ordinal data then take appropriate inputs from the user. Else if the column has categorical data, then replace the categorical data in that column with random numbers. If the column has numerical data, then the values in that are retained.
(5) The result of the above steps is the converted dataset DS' which is given to CBTS algorithm for transformation.

Algorithm 2 CBTS

INPUT: Original dataset DS with ordinal and numerical data.
OUTPUT: Transformed data DS" generated from converted data DS'.
(1) Construct the Correlation matrix (Dc).
(2) Normalize the original data.
(3) Repeat the following steps from 3 to 6 for each private column $p_i \in P$.
(4) Calculate threshold coefficient for selected column p_i which separates the highly correlated data of size separation factor. Select columns whose correlation coefficient with selected private column is greater than the threshold correlation value to obtain the subset.
(5) Transform subset using required transformation technique or perturbation method.
(6) Substitute corresponding component of transformed data in place of p_i in normalized original data.
(7) Denormalize the original data.
(8) Return *DS*" transformed data.

5. RESULT

Information Entropy of numerical data against perturbed data using CBTS with transformation methods is summarized in Table 1. We can infer from the table that deviation in Information Entropy is minimum using proposed CBTS method against using transformation techniques alone. Table 2 gives the comparison of classifier accuracies for various machine learning algorithms using CBTS against original data. It is clearly observable from the results that

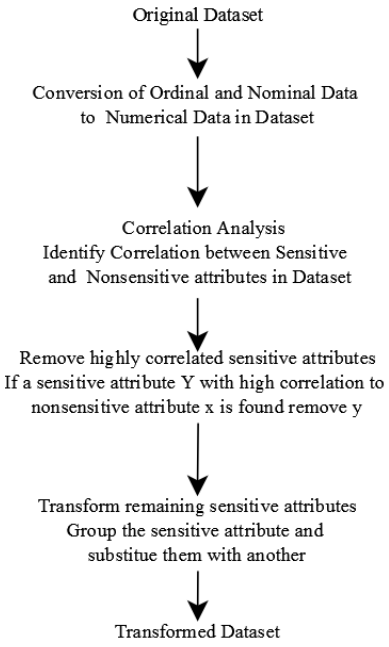

Fig. 2: Transformation Method

the classifier performance is comparable to the original data. Table 3 shows the Misclassification Error M_E values with k-means clustering. Higher M_E values indicate lower clustering quality where as Lower M_E values indicate the higher utilization of the data.

$$M_E = \frac{1}{N}\sum(|Cluster_i(D)| - |Cluster_i(D')|) \qquad (3)$$

The clustering quality provided by the proposed CBTS is higher compared to existing methods. The datasets used in this paper are Soybean and Breast Cancer. Both the datasets are taken from UCI Machine Learning Repository. Soybean dataset is a dataset with 307 instances and 35 attributes. Among 35 attributes some are ordinal and some are nominal. Breast Cancer is another dataset with ordinal, nominal and numerical attributes. There are 286 instances and 10 attributes in this dataset. This dataset contains two classes and among 286 instances, 201 belong to one class and the other 85 belong to another class.

Table 1. 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

Table 2. 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 3. Misclassification Error M_E values with and without CBTS

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

5.1 Analysis based on decision trees

The overall process in our transformation method is given in Figure.2. The results in the form of decision trees before and after conversion of datasets are shown in Figure 3 and 4. The decision trees of dataset Breast cancer before and after conversion are considered. The root node remains same in both decision trees by which it is evident that our conversion method is correct. Even the sub-trees are almost the same. The structure of the original decision tree is almost maintained in the converted decision tree.

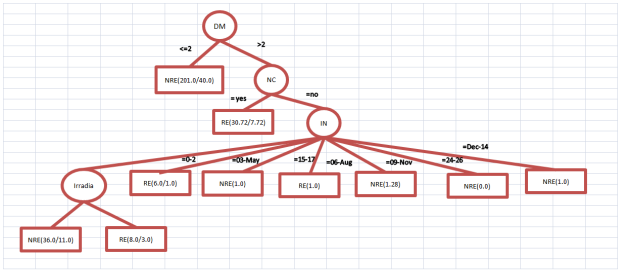

Fig. 3: Decision tree of dataset Breast Cancer before conversion of orinal and nominal data
In decision trees
MP→ MenoPause
TS→Tumour Size
IN→Inv-Nodes
NC→Node Caps
DM→Deg-Malig
BS→Breast Side
BQ→Breast Quad

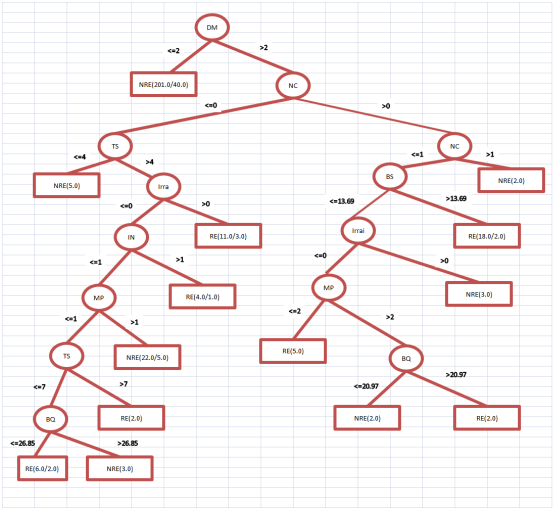

Fig. 4: Decision tree of dataset Breast Cancer after conversion

RE→Reccurence Events
NRE→Non- Recurrence Events

5.2 Analysis based on Information Entropy

The proposed method was also tested using Information Entropy. Information Entropy of original data, Transformed data (using existing methods and CBTS method) is tabulated in Table 4 using Shannons Entropy. Information Entropy of original data against perturbed data using CBTS for Ordinal and Nominal data with transformation methods is summarized in Table 4. We can infer from the table that deviation in Information Entropy is minimum using proposed CBTS method against using transformation techniques alone. Table 5 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. We can infer from the table that deviation in Information Entropy is minimum using proposed CBTS method compared to original Entropy. From this we can conclude that the loss of information in the transformed data is minimum compared to the original data.

Table 4. Comparison of Information Entropy

Types of data	Original Entropy	Information Entropy(IE) (Using CBTS Method) / (Using existing methods)		
		PCA	SVD	NNMF
Soybean (683x36)	3.317	3.30/ 10.25	3.41/ 5.12	3.25/ 9.26
Car (1729x6)	2.31	2.28/ 5.16	2.37/ 2.58	2.22/ 9.14
nursery (12960x7)	1.88	1.88/ 4.6	1.95/ 2.8	1.86/ 11.0
Breast Cancer (286x9)	3.02	3.7/ 6.39	3.5/ 3.8	3.39/ 8.04

5.3 Analysis Based On Machine Learning Algorithms

Different machine classifiers like Decision Tree, Multilayer Perceptron and Naive Bayes are run on the original and transformed data. Classifier accuracy is measured for each machine learing algorithm and tabulate in Table 5. It gives the comparison of various machine learning algorithms using CBTS. The table shows the number of correctly classified instances in the original data set and the transformed dataset on various machine learning algorithms namely Decision Tree, Multilayer Perceptron and Naive Bayes Classifiers. From the table we can infer that the results obtained are comparable with the original data.

Table 5. Comparison of various Machine Learning Algorithms Using CBTS

Dataset	Machine Learning Algorithm	Observed Classifier Accuracy (%)				
		Categorical Data	Numerical Data	Transforming using CBTS		
				PCA	SVD	NNMF
Soybean	Decision Tree	97.0	96.3	97.6	97.0	97.0
	Multilayer Perceptron	99.8	93.3	94.8	95.0	95.0
	Naive Bayes	93.7	82.1	82.5	81.8	81.8
Breast Cancer	Decision Tree	81.4	81.4	81.4	81.4	81.4
	Multilayer Perceptron	84.6	84.6	84.2	84.6	84.6
	Naive Bayes	73.4	73.4	73.4	73.4	73.4

5.4 Analysis based on Clustering Quality

Table 6 shows the Misclassification Error (M_E) values. Higher M_E values indicates lower clustering quality whereas Lower M_E values indicate the higher utilization of the data. The computed M_E values for PCA, SVD and NNMF based CBTS methods for all the three datasets are shown in the following table. The clustering quality provided by proposed CBTS is higher compared to existing methods.

Table 6. Cluster Misclassification Error (M_E)

Types of data	Clusters (K)	M_E (with CBTS)			M_E(without CBTS)		
		PCA	SVD	NNMF	PCA	SVD	NNMF
SOYBEAN	2	0.253	0.455	0.248	0.999	0.999	1.0
	3	1.22	0.88	0.74	1.09	2.6	0.9
BREAST CANCER	2	0.017	0.7	0.7	1.3	1.60	1.50
	3	0.7	0.74	0.74	0.5	1.91	1.54

6. CONCLUSION AND FUTURE WORK

The present work explores CBTS for Numerical data and Categorical 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 transformation method has low clustering Misplacement Error. The proposed work can be extended by use of Vector Marking techniques where these techniques help in increasing the efficiency by avoiding unauthorised access to the information.

7. 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] Vijayarani S. and A. Tamilarasi. *"An efficient masking technique for sensitive data protection."* Recent Trends in Information Technology (ICRTIT), 2011 International Conference on. IEEE, 2011.

[3] 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

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

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

[6] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin,*"Privacy-preserving patient-centric clinical decision support system on naive bayesian classification"*, IEEE Journal of Biomedical and Health Informatics, pp.1-1, 2015.

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

[8] Y. Sang, H. Shen, and H. Tian, *"Effective reconstruction of data perturbed by random projections"*, IEEE Transactions on Computers, vol.61, no.2, pp.101-117, Jan 2012.

[9] Fong, P.K. and Weber-Jahnke, J.H., *" Privacy preserving decision tree learning using unrealized data sets"*, IEEE Transactions on knowledge and Data Engineering 2012, 24(2), pp.353-364.

[10] Alotaibi, Khaled, and Beatriz De La Iglesia. *"Privacy-preserving SVM classification using non-metric MDS."* (2013): pp. 30-35.

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

[12] 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.

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

[14] Upmanyu, Maneesh, Anoop M. Namboodiri, Kannan Srinathan, and C. V. Jawahar.*"Efficient privacy preserving k-means clustering"*. In Pacific-Asia Workshop on Intelligence and Security Informatics, pp. 154-166. Springer Berlin Heidelberg, 2010.

[15] Zekeriya Erkin : *"Privacy-preserving distributed clustering"*. EURASIP Journal on Information Security pp.1-5, 2013(1),.

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

[17] Guo, Ling. *"Randomization Based Privacy Preserving Categorical Data Analysis"* Diss. The University of North Carolina at Charlotte, 2010.

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

[19] 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.

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