Indian Journal of Science and Technology, Vol 8(17), DOI: 10.17485/ijst/2015/v8i17/77584, August 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Attribute Segregation based on Feature Ranking Framework for Privacy Preserving Data Mining

R. Praveena Priyadarsini[1*], M. L. Valarmathi[2] and S. Sivakumari[1]

[1]Department of Computer Science and Engineering, Faculty of Engineering, Avinashilingam University, Coimbatore - 641 108, Tamil Nadu, India; Praveena.priya04@gmail.com, prof.sivakumari@gmail.com
[2]Department of Computer Science and Engineering, Government College of Technology, Coimbatore - 641013, Tamil Nadu, India; mlvm@gct.ac.in

Abstract

Attributes in macro-data have to be segregating based on their sensitivity for privacy preservation purposes. Automating this attribute segregation becomes complicated in high dimensional datasets and data streams. In this work, information or correlation of the attribute on the target class attribute is measured using Information Gain [IG], Gain Ratio [GR] and Pearson Correlation [PC] ranker based feature selection methods and this values are used to segregate them as Sensitive Attributes [SA], Quasi Identifiers [QI] and Non-Sensitive [NS] Attributes. Segregated attributes are subjected to various levels of privacy preservation using both the proposed Double layer Perturbation [DLP] and Single Layer Perturbation [SLP] algorithms to form the level-1 perturbed datasets. The level-1 perturbed dataset is further perturbed by applying SLP algorithm to form level-2 and level-3 privacy preserved datasets. Thus, the multiple versions of Adult dataset created are distributed to data seekers based on their trust levels in Multi Trust Level [MTL] environment. The privacy preserved dataset versions created using the proposed algorithms are evaluated based on their utility, distortion and purity metrics. The results show that the ranker methods are able to identify attributes which had sensitive content as either SA or QI automatically and the proposed perturbed datasets have good utility on selected classification and clustering algorithms when compared to original and L-diversified datasets. Also, the distortion values of these datasets signify that they can prevent diversity attacks.

Keywords: Attribute Segregation, Multi-Trust Level Environment, Privacy Preserved Dataset, Ranker Based Feature Selection Methods, Utility and Diversity Attack

1. Introduction

Privacy Preserving Data Mining [PPDM] is a field of research that does tradeoff between privacy protection and utility of the dataset[1]. The new dimension of PPDM is the Multi-Level Trust, where the data at different privacy levels are released to users based on their trust level. The most common attack in this scenario is diversity attack where a miner is able to obtain more than one version of a dataset that may be linked to reconstruct the original data set[2]. A open research problem in a very high dimensional data set is separating the attributes into QI, SA and NS attributes[3]. A lot of analysis is needed on these data before the attributes are classified as SA and QI. Sometimes an attribute may be named as both SA and QI which may lead to problems while treating them with privacy techniques.

In this work it is assumed that all the attributes that are strongly informative and correlated towards the class attributes on which knowledge discovery and statistical analysis is performed are considered to be sensitive. Hence, these attributes are to be subjected to privacy preservation such that they do not expose the sensitive and private information about the individual or organization stored in the dataset. Three attributes evaluation measures IG, GR and PC[4] are used to evaluate the importance of the attributes in the dataset. The attributes are then vertically partitioned based on their rank value. The attributes of very high rank value are classified as SA, the attributes

* Author for correspondence

that have mid-level rank value then the previous partition are classified as QI attributes and the attributes with very low or no rank value are considered as NS attributes. SA attributes are treated with two layered perturbation using proposed DLP algorithm and QI attributes are treated with single layered perturbation using proposed SLP algorithm. These perturbed attributes are combined to form versions of layer-1 perturbed datasets. These layer-1perturbed dataset versions may be further perturbed to form layer-2, layer-3 perturbed derived datasets and distributed among various data seekers based on their trust level.

The utility and the privacy preserving capability the proposed privacy preserved dataset versions are tested using accuracy, purity and distortion metrics. The proposed privacy preserved versions are compared with L-Diversity privacy preserved dataset.

2. Background and Related Work

Xia okui and Xiao[5] discussed about Multi-level perturbation scenario in privacy preservation where dataset with various level of perturbation is released to the users with various levels of trust. Multi-level trust setting where the most trusted data miner is given a least perturbed dataset was proposed by Yaping Li et al.[2]

Kun Liu et al.[6] have explored the problem in computing statistical aggregates like the inner product matrix, correlation co-efficient matrix and Euclidean distance matrix for distributed privacy of sensitive data, possibly owned by multiparty. The collaborative filtering system algorithm based on randomization perturbation techniques for secure multi-party computing in distributed environment was presented by Sangyie Gong [7].

Feature selection methods are commonly used to identify sensitive attributes and privacy preservation is applied on them. Vidya Banu and Nagavani[8] proposed a principal component analysis based feature selection technique for preserving privacy of sensitive attributes. The use of feature selection techniques for privacy preservation was proposed by Peng Peng Lin[9], where Sparsified Singular value decomposition is used for data distortion and filter based feature selection method is used for feature selection.

K. Kisilevich, et al.[10] proposed a technique where less influence attributes are identified using decision tree method and suppressed. This work proved that by suppressing less influential attributes and anonymizing more influential attributes yields high utility. Attributes with high entropy values, when privacy preserved can increase the utility of the privacy preserved dataset. Ling Guo and Xiaowei[11] proved that randomization technique preserves the correlation between the sensitive attributes and quasi identifiers across various privacy thresholds.

Weiwei Ni and Zhihong Chang[12] used information entropy theory to differentiate attribute and apply different obfuscation techniques to attributes based on the entropy values. It was proved that the above process had increased the data utility.

Lihiu et al.[13] proposed a individually adaptable perturbation model where the individuals choose their own privacy level and random noise is added to sensitive attributes. Likun Liu et al.[14] proposed two noise addition algorithm that was applied to obtain perturbation model that ensured privacy in health care data. Identifying quasi attributes and sensitive attributes and performing privacy preservation is a challenge in high dimensional dataset. Charu C. Aggarwal[15] discussed that in a high dimensional data in attribute combinations within a record have a powerful reeling effect.

Xiao Xun Sun et al.[16] proposed finer level anonymization scheme where attributes are prioritized and anonymization based on the priority level of the attribute and have built a link between trust level and degree of data anonymization. Islam, M. Z. and Brankovic, L.[17] proposed noise addition technique namely leaf innocent attribute perturbation technique, leaf influential attribute perturbation technique for numerical attributes and Detective clustering perturbation technique for categorical attributes to protect the privacy of the private attributes of the macro data. Rajalakshmi and Anandha Mala[18] have clustered the data using k-means clustering. The sub-clusters are formed from the clusters based on their distance from centroid and arranged sequentially. The equivalence class of each record in the sub-cluster are anonymized using controlled relocation which yields good utility to the perturbed dataset. Data mining privacy by decomposition algorithm was proposed by Nissim Matato et al.[19] that protects sensitive data using K-Anonymity. Attributes are projected using genetic algorithms and each projection is anonymized such that when rejoined they comply with K-Anonymity.

3. Methodology and Definitions

Multi Trust Level [MTL] is a scenario in privacy preservation where versions of privacy preserved data are distributed to the users based on their trust levels. The

most likely attack in this scenario is the diversity attack. This work, automatic selection and projection of SA, QI and NS attributes are performed by using three attribute ranking methods namely Information Gain, Gain Ratio and Pearson Correlation. Privacy Preservation algorithms are then applied on these projected attributes based on their information content. DLP algorithm is applied on numerical and categorical values of the SA projected attributes. SLP algorithm is then applied on attributes classified as QI. Combining the privacy preserved SA and QI with NS attributes will produce perturbed base version of the dataset. The various derived versions of dataset are then generated from the base datasets by applying SLP algorithm on the base perturbed datasets.

In the proposed framework new version of data set is produced on every request as per the user trust level using the following steps:

For every request

Step-1: Rank the attributes based on any one of the following attribute ranker method: 1. Pearson Correlation, 2. Gain Ratio and 3. Information Gain.

Step-2: Based on the rank generated by the Ranker method on the attributes, divide them into high priority and low priority attributes.

If Attribute Rank Value >= Threshold Ranke Value then

Project the attributes as High priority attributes and set as SA of the dataset.

If Attribute Rank Value <= Threshold Ranke Value then

Project the Attributes as Low priority attributes and set as QI attribute of the dataset.

Step-3: For the high priority attributes set as SA apply DLP Algorithm.

For the low priority attributes set QI apply SLP Algorithm.

Step-4: Level-1perturbed datasets **Adult-v1, Adult-v2, Adult-v3**, are obtained based on attribute segregation using three ranker methods are formed using equation (1)

$$B_n = \text{DLP[SA]} \cup \text{SLP[QI]} \cup \text{NSA} \quad (1)$$

//Combine the perturbed attributes to form base version of perturbed dataset//.

Step-5: Various levels of Derived datasets $D[V_n]$ from the level -1 perturbed dataset B_n are generated by using equations 2&3.

$$DV_n = \text{SLP}\ [B_n] \quad (2)$$

$$DV_{n.n} = \text{SLP[DVn]} \quad (3)$$

Step-6: Evaluate the utility and distortion for various privacy preserved versions of dataset.

This framework aims to build sequential release of privacy preserved dataset such that each release contains attributes that are perturbed at various level of privacy.

3.1 Assumption and Definitions

3.1.1 Assumption-1

Let $\{A_1, A_2,..A_n \subseteq D\}$ be the attributes of dataset D. Let $\{A_{x1}, A_{x2},... .A_{xn} \subseteq D\}$ be the attributes with high information/correlation value on class attribute than are projected as privacy high or SA. Let $\{A_{y1}, A_{y2},... .A_{yn} \subseteq D\}$ be the attributes having mid-level information/correlation values, hence these attributes are projected as QI. Attributes with very low level Information/Correlation values are considered as NS.

Attributes in the dataset can partitioned based on the privacy content of the attributes as Sensitive Attributes, Quasi Identifiers, and Non_Sensitive Attributes which are defined as follows:

The attacks that are common in a Multi Trust Level [MLT] Environment are Diversity attacks which is defined as follows:

3.1.2 Diversity Attack

Let be the original dataset and $V(D_1),V(D_2),..V(D_n)$ are privacy preserved versions of the dataset when published these datasets should obey the following condition (1)

$$V(D_n) \cup V(D_{n+1}) \neq D \quad (1)$$

That is two perturbed version should not combined to detect the original dataset[20].

3.1.3 Definition 1

Let $\{a_1, a_2, a_3.....a_n \subseteq D\}$ where $a_1, a_2,...a_n$ are the attributes of the dataset D. If $\{a_{x1}, a_{x2},.....a_{xn} \subseteq D\}$ are the attributes which contain sensitive information about an individual or an organization then $\{a_{x1}, a_{x2},.....a_{xn} \subseteq D\}$ are termed as SA attributes.

3.1.4 Definition 2

Let $\{a_1, a_2, a_3.....a_n \subseteq D\}$ where $a_1, a_2,...a_n$ are the attributes of the dataset D. If $\{a_{y1}, a_{y2},.....a_{yn} \subseteq D\}$ are the attributes

that contain information which may be linked to identify a particular individual or organization then $\{a_{y1}, a_{y2}.....a_{yn} \subseteq D\}$ is termed as Quasi Identifiers.

3.1.5 Definition 3

The attributes that do not contain any information which may expose the privacy of an individual or organization is called Non-Sensitive Attributes.

4. Attribute Partitioning using Ranker Methods

4.1 Information Gain [IG] Attribute Ranker Method

Symmetric measures within the attributes of the dataset can be identified using information gain value. Information Gain is an information measure based on entropy[21]. Uncertainty in a system can be measured using entropy value. If X, Y are the attributes of the dataset D, the entropy of an attribute Y is calculated using equation (4)

$$H(Y) = -\Sigma_{y \in Y} p(y) log_2 p(y) \quad (4)$$

This decrease in entropy value of Y due to the additional information about Y provided by X is termed as Information Gain [IG] or mutual information and is given by equation 5, 6 and 7:

$$Gain = H(Y) - H\left(\frac{Y}{X}\right) \quad (5)$$

$$H(X) - H\left(\frac{Y}{X}\right) \quad (6)$$

$$H(Y) + H(X) - H(X, Y) \quad (7)$$

Information Gain is also called symmetrical measure since the amount of information gained about Y after observing X is equal to the amount of information gained about X after observing Y. Symmetry is a measure of feature inter-correlation.

4.2 Gain Ratio [GR] Attribute Ranker Method

Gain Ratio (GR) supports those attributes which have unequal distribution of values. Hence gain ratio measures are more attractive than Information Measures[22,23]. Since information gain value is a symmetric measure GR is also an symmetric measure and capable of measuring the linking of the attributes with its class[21]. This can be measured by calculating the ratio of the gain in information (IM) from using the attribute to the information value of the attribute itself. Thus Quinlan proposes a gain-ratio measure as given in equation (8)

$$GR(A) = \frac{IM(A)}{IV(A)} \quad (8)$$

Where IV (A) is the information measure of A and is calculated using the equation (9)

$$IV(A) = -\sum \frac{x_i}{N} \log \frac{x_i}{N} \quad (9)$$

Where the value of the attribute A and N is is the number of values in the attribute.

4.3 Pearson Correlation [PC] Attribute Ranker Method

The benefit of this measure is that it helps in identifying the linear correlation of the attribute A with its class C. The linear or Pearson Correlation ($\rho(A,C)$)[24] is measured using the equation (10).

$$\rho(A,C) = \frac{\sum_i (a_i - \bar{a_l})(c_i - \bar{c_l})}{\sqrt{\sum_i (a_i - \bar{a_l})^2 \sum_i (c_i - \bar{c_l})^2}} \quad (10)$$

Where $\bar{a}_1$ is the mean value of the attribute A and c_1 is the mean value of the attribute C. Value of $\rho(A,C)$ will vary from -1 to +1. If the attribute A has a linear dependency on the class C than $\rho(A,C)$ will be highly positive towards +1. If the attribute does not have linear dependency then $\rho(A,C)$ will have either zero or negative values nearing -1. In this work using weka (Mark Hall et al., (2009)), the attributes in the adult dataset are ranked by decreasing order of $\rho(A,C)$. Since $\rho(A,C)$ can be calculated only for numeric values, Weka simulator considers the nominal attributes in the dataset on value by value basis. Here each nominal value of attributes are treated as indicators and weighted average is used for calculating the overall correlation of the attribute.

5. Proposed Multi-Level Perturbation Frame Work

5.1 Double Layer Perturbation Algorithm [DLP]

In the proposed DLP algorithm, the two layers of independent noise is added to the numeric attributes

of the dataset generated using an exponential equation. For the categorical attributes, the swap function where a percentage of values of most occurring values of SA values are swapped with the other values of the same attribute. The algorithm for the same is given in Figure 1:

Double layer Perturbation Algorithm [DLP]
Input: Attributes of dataset set as sensitive attributes (SA);
Output: SA attributes with two level perturbations;
If SA(n) = Numeric; then
For (i=1 ; i<=n; n++)
{
$SA^n = SA(i) + Y$ //Y = independent noise of SA domain
$SA^{n+1} = SA^n + Y$ //Y = independent noise of SA domain
}
If SA[n]=categorical ;then
for (i=1 ; i<=n; n++)
{
$SA^n = RankSwapN_1$ % *of SA*
$SA^{n+1} = RankSwapN_2$ % *of* SA^n
Where $N_1 > N_2$
}

Figure 1. ProposedDouble Layer Perturbation Algorithm.

5.2 Single Layer Perturbation Algorithm [SLP]

In SLP algorithm single layer of noise is added to the numeric values of the dataset. Whereas a percent of values of most occurring attributes are swapped with other values using rank swapping method and the algorithm for the same is given in Figure 2.

As a result of applying DLP and SLP algorithms on the attributes partition the generated level-1 perturbed datasets are

- Adult-info-gain-perturbed-Base-dataset [AIGBV1],
- Adult-gain-ratio-perturbed-Base-dataset [AGRBV1]
- Adult-correlation-perturbed-Base-dataset [APCBV1].

These datasets will obey the condition (1). Since every attribute in level-1 perturbed dataset will under goes different level of perturbation based on the ranker method used to set the SA, QI attributes, it is difficult to link the values of two perturbed datasets to obtain the real values of the original dataset. Thus, diversity attack is prevented.

Single Layer Perturbation Algorithm [SLP]
Input: Attributes of dataset set as Quasi Identifiers (QI);
Output: QI attributes with single level perturbations;
If A_n= Numeric
Then
$A_n = QI_{n-1} + Y$ // Y= independent noise of domain A
if A_n= categorical
then
$A_n = RankSwapN_1$ %$(Domain(A_1))$
then base version perturbed dataset
B[*n*] = *DLP*[*SA*] ∪ SLP[QI] ∪ NON_*SA*
End;

Figure 2. Proposed Single Level Perturbation algorithm.

The level-2 perturbed datasets derived from level-1 perturbed datasets by applying SLP algorithm are as follows:

- Adult-info-gain-perturbed-Derived - dataset [AIG-DV1].
- Adult-gain-ratio-perturbed-Derived-dataset [AGRDV1].
- Adult-correlation-perturbed-Derived-dataset [APC-DV1].

The level-3 derived datasets are than obtained by applying SLP algorithm on the numeric and categorical attributes on the level-2 perturbed datasets are as follows:

- Adult-info-gain-perturbed-Derived-dataset [AIGDV2].
- Adult-gain-ratio-perturbed-Derived-dataset [AGRDV2].
- Adult-correlation-perturbed-Derived-dataset [APC-DV2].

Based on trust level the level 1, 2 and 3 perturbed datasets may be distributed to the users in MLT environment.

6. Performance Metrics

The base and derived datasets are evaluated for its utility, attribute distortion and ability to preserve privacy based on the following metrics:

6.1 Classification Accuracy and RMSE

The utility and the amount of information lost due to the application of privacy preservation technique is measured

using classification accuracy, Root Mean Squared Error[24] that represents the outliers in the data.

6.2 Purity Measure

Purity[20] is a metric used to measure the quality of clustering solution. Purity is calculated by dividing the correctly assigned objects to the total number of objects in the dataset and the equation for the same is given in equation (11):

$$Purity\,(C,P) = \frac{1}{n}\sum_{j}\max_{i}|C_j \cap P_i| \tag{11}$$

Where $\{C_1 \ldots C_J\}$ denotes the partition built by the clustering algorithm on objects, and $P = \{P_1 \ldots P_i\}$ denotes the partition inferred by the original classification. J and I are respectively the number of clusters |C| and the number of classes' |P|. The value of n denotes the total number of objects.

7. Change of Rank of Features (CP)

The metric CP measures the difference in the rank of the attributes before and after privacy preservation. If the rank of the attribute before privacy preservation is denoted as RAVi and after privacy preservation is denoted as the difference in their rank is measured using the metric CP as given in equation (12):

$$CP = \frac{\sum_{i=1}^{m}|Rav_i - Rav_i^*|}{m} \tag{12}$$

Where RAV_i is the rank of the average value of attribute i, while RAV^*_i denotes its rank after the distortion[25].

7.1 Maintenance of Rank of Features (CK)

CK is the metric that measures the percentage of the attributes that keep their ranks after the privacy preservation[25] and is calculated using equation (13 and 14).

$$CK = \frac{\sum_{i=1}^{m} CK}{m} \tag{13}$$

$$CK = \begin{cases} 1 & if\ \ Rav_i - Rav_i^* \\ 0 & otherwise \end{cases} \tag{14}$$

Where m is the total number of attributes in the dataset D.

8. Dataset Description

The dataset used in this work is Adult dataset available on UCI Machine Learning Repository[26]. The dataset contains 14 attributes including class attribute, with values, ">50K" and "<=50K". The dataset is resampled and 8410 instance are taken for experimentation. The dataset contains 14 attributes including class attribute, with values, ">50K" and "<=50K".

9. Results and Discussions

9.1 Attribute Segregation using Ranker Methods

The attributes in the dataset are vertically partitioned based on their rank value using any one of the IG, GR or PC attribute selection measure. From the segregated attributes, using the three ranker methods it is known that Education Number, Marital Status, Relationship, Age that are set as sensitive attributes are common for all the three ranker methods while one attribute Work class is commonly partitioned as QI by all the ranker methods. Also, except the attribute Race the proposed rank based partitioning method has correctly partitioned all other attributes as SA and QI.

9.2 Utility Measurement

The experiments were conducted using Weka software (Mark Hall et al., (2009)). The utility of these datasets are compared based on classification accuracy and Root mean squared error values of Ripper Algorithm[27], C4.5[23] and Naive Bayes[28] algorithms and shown in Figure 3.

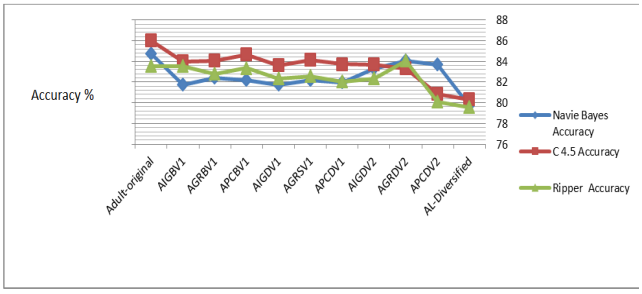

Figure 3. Comparison of accuracies values of various perturbed dataset versions.

From Figure 3 it is incurred that accuracy of base and derived versions of datasets exhibit a very small decrease in accuracy than the original dataset. On Ripper

algorithm, there is a decrease in accuracy by 2% for all the proposed perturbed dataset. The result also indicates that as the perturbation level increases there is small decrease in accuracy percentage.

On C 4.5 algorithm there is a reduction of accuracy of about 2% in all the proposed perturbed dataset except for the APCDV2 dataset. This shows that attribute perturbed using IG and GR attribute selection methods are able to give good utility than PC based attribute segregation and perturbation method. Also, the proposed perturbed dataset accuracies are comparable with original and L-Diversified Adult dataset.

When the accuracy of the Naive Bayes algorithm on the proposed privacy preserved datasets are compared with original and L-Diversified Adult dataset, the accuracy of the perturbed datasets decreases in the second level and starts to increase in the next level. All the perturbed datasets have a better accuracy than Adult L-Diversified dataset on all the three classification algorithms.

The Root Mean Squared Error [RMSE] of the perturbed datasets are compared on Ripper Naive Bayes and C4.5 classification algorithms and shown in Figure 4.

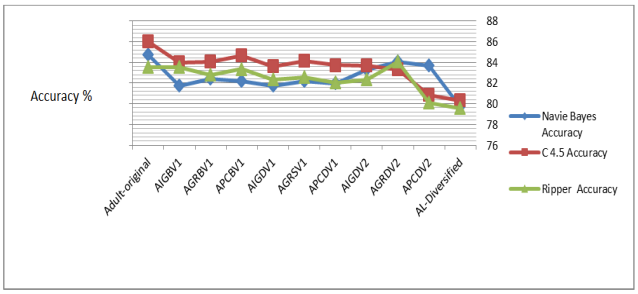

Figure 4. Comparison of RMSE values on various perturbed dataset versions.

From Figure 4 it is incurred that RMSE for Naive Bayes algorithm for all the perturbed versions increases from 0.35 to 0.39 at level one perturbation and remains almost the same for second level perturbation versions. On the third level perturbed datasets the error rate decreases except for APCDV2 dataset. On Ripper algorithm, RMSE values increases from 0.35 to 0.39 for level-1 perturbed dataset and remains at 0.39 for all level-2 perturbed datasets and decreases in level-3 perturbed dataset except for APCDV2. Also, for The RMSE values of all the proposed perturbed dataset are comparable with RMSE value of L-Diversified adult dataset on the both classification algorithms.

9.3 Privacy Preservation Measurements

The purity values of the various perturbed datasets are compared on K-Means and EM clustering algorithms[24] and given in Figure5.

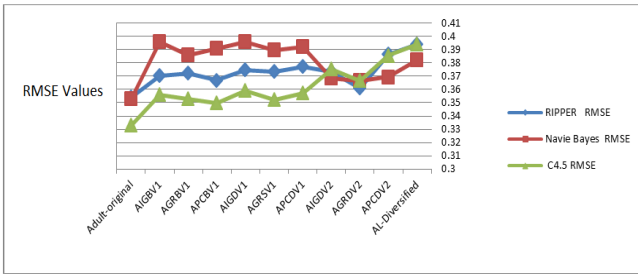

Figure 5. Purity value comparisons of the perturbed datasets with L-diversity dataset.

From Figure 5 it is incurred that perturbed dataset AGRBV1 has the lowest purity values of the all proposed perturbed datasets. While all the other proposed perturbed datasets have the same or higher purity values than the original dataset. Adult L-diversified [AL-Diversified] dataset has the lowest purity value. The accuracy and RMSE values of the all proposed perturbed datasets on both the classification algorithms, Ripper, C4.5 and Naïve Bayes have a decrease 2-3% from the original dataset except for APCDV2 dataset. When the purity values of the all proposed perturbed datasets on K-Means and EM are observed except for the level-1 perturbed datasets other datasets have comparable and higher purity than original and L-Diversified datasets.

To measure the distortion within the datasets after privacy preservation, the rank value of the attributes in the perturbed versions are compared with the attribute rank values of the original dataset. The CK and CP value of Information Gain perturbed dataset versions are calculated and shown in Table 1.

Table 1. Rank comparisons of Datasets perturbed based on IG ranking method

Dataset versions	Information Gain Rank Vector	Ck Value	CP Value
AIGBV1	[5,11,1,14,8,6,7,4,12,10,13,2,9,3]	0.14	4.2
AIGDV1	[5,11,1,14,8,6,7,4,12,10,13,2,9,3]	0.14	4.2
AIGDV2	[5,8,14,6,1,7,4,11,13,12,2,9,10,3]	0.21	2.5

Higher the CP values, more the level of distortion in the datasets, whereas lower the Ck value indicates more distortion[25]. The table shows that the versions of AIGVD datasets are equally distorted, while AIGDV 2 version considerably less distorted than other two versions. The dataset where the SA and QI attributes are set using Gain Ration ranker method is compared on the level of distortion using CK and CP value measure and shown in Table 2.

Table 2. Rank comparisons of datasets perturbed based on GR ranking method

Dataset versions	Gain Ration Rank Vector	Ck Value	CP Value
AGRBV1	[12,13,7,8,1,5,10,4,14,7,13, 2,9,3]	0.14	0.35
AGRDV1	[12,13,5,1,8,14,4,6,7,10,13, 2,9,3]	0.28	3.1
AGRDV2	[12,13,1,6,8,4,13,7,10,2,9,1 1,12,3]	0.21	3.35

The CK and CP values in Table 2 shows that the first level Gain Ration based perturbed dataset has the least amount of distortion. The second and third level perturbed datasets have higher level of distortion.

The dataset where the SA and QI attributes are set using Pearson Correlation ranker method is compared on the level of distortion using CK and CP value measure and shown in Table 3.

Table 3. Rank comparisons of datasets perturbed based on PC ranking method

Dataset versions	Pearson correlation Rank Vector	Ck Value	CP Value
APCBV1	[12,13,7,8,1,5,10,4,14,7,1 3,2,9,3]	0.071	4.7
APCDV1	[12,13,5,1,8,14,4,6,7,10,1 3,2,9,3]	0.071	4.7
APCDV2	[12,13,1,6,8,4,13,7,10,2,9, 11,12,3]	0.14	3.8

The Ck value of the AGRBV1 and AGRDV1 show that they are the most distorted datasets out of all the perturbed versions.

Thus, the utility of the IG and GR perturbed datasets remain consistent in all the level of perturbation on both classification and clustering algorithms. Also, the utility of the proposed datasets are better than the accuracy of L-Diversified adult dataset. On the distortion metric CP and CK, Information Gain and Pearson Correlation perturbed dataset have the highest distortion rate than Gain Ratio perturbed datasets.

10. Conclusions

This work proposes an attribute segregation and perturbation frame work for Multi-Trust Level scenario. The proposed frame work uses the information or linearity of the attributes with respect to its class attribute to identify the sensitivity among the attributes. The segregated attributes are privacy preserved using DLP and SLP algorithms to form base datasets. Different versions of derived datasets are parallel generated from the base datasets using the proposed frame work and are distributed to different users based on their trust level. The results show that the ranker methods are able to identify attributes which had sensitive content as either SA or QI automatically. Also, when perturbed versions of datasets are evaluated based on distortion metrics, all the perturbed versions have good distortion values thus preventing diversity attacks. When compared for its utility with the original dataset and L-Diversified Adult dataset there is a very small variation in accuracy in all the proposed perturbed datasets. Thus, the experiments show that perturbation of high ranked attributes does not have much effect on the utility of the datasets. As future enhancement vertical and horizontal partitioning and perturbation can be implemented to enhance privacy of the dataset.

11. References

1. Lindell Y, Pinkas B. Secure multiparty computation for privacy-preserving data mining. Journal of Privacy and Confidentiality. 2009; 1(1):59–98.
2. Li Y, Chen M, Li Q, Zhang W. Enabling multi-level trust in privacy preserving data mining. IEEE Computer Society. 2011 Jun.
3. Xiao Xun S, Huawang LJ, Zhang Y. Injecting purpose and trust into data anonymization. Computers and Security. 2011; 30:332–45.
4. Hall M. Correlation based feature selection for machine learning (Doctoral dissertation). Department of Computer Science, University Waikato; 1999.
5. Xia Okui X. Optional random perturbation at multiple privacy levels. VLDB 09. 2009 Aug 24–28.
6. Kun L, Kargupta H. Random projection based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transaction on Knowledge and Data Engineering. 2006 Jan; 18(1).

7. Sangyie G. Privacy preserving collaborative filtering based on randomized perturbation techniques and secure multiparty computation. International Journal of Advancements in Computing Technology. 2011; 3(4).
8. Vidya Banu R, Nagaveni N. Evaluation of a perturbation based techniques for privacy preservation in a multi party clustering scenario. Information Sciences. 2013 May 20; 232:437–48.
9. Lin P. A comparative study on data perturbation with feature selection. Proceedings of International Multi Conference of Engineers and Computer Scientist (IMECS'2011); 2011 Mar. p. 1–168.
10. Kisilevich K, Rokach L, Elovici Y, Shapira B. Efficient multidimensional suppression for K-Anonymity. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(3):334–47.
11. Guo L, Y Xiaowei. Attribute disclosure in randomization based privacy preserving data publishing. IEEE International Conference on Data MiningWorkshop; 2010.
12. Ni W, Zhihong C. Clustering-oriented privacy preserving data publishing. Knowledge based Systems. 2012; 35:264–70.
13. Liu L, Kantarcioglu M, Thuraisigham B. The applicability of the perturbation based privacy preservation datamining for Real-World Data. Data and Knowledge Engineering. 2008; 65:5–21.
14. Liu L, Kexin Y, Hu H, Li L. Using noise addition method based on pre-mining to protect healthcare privacy. Journal of Control Engineering and Applied Informatics. 2012; 14(2):58–64.
15. Charu CA. On K-anonymity and the course of dimensionality. Proceedings of 31st VLDB Conference; Trondheim, Norway; 2005.
16. Sun X, Wang H, Li J, Zhang Y. Injecting purpose and trust into data anonymization. Computers and Security. 2011; 30:332–45.
17. Islam MZ, Brankovic L. Privacy preserving data mining: Noise addition to categorical values using a novel clustering technique. Proceedings of IEEE Transactions on Industrial Informatics.
18. Rajalakshmi V, Anandha Mala GS. Anonymization by data relocation using sub-clustering for privacy preserving data mining. Indian Journal of Science and Technology. 2014 Jul; 7(7):975–80.
19. Nissim M, Lior R, Oded M. Privacy-preserving data mining: A feature set partitioning approach. Information Sciences. 2010; 180:2696–720.
20. Ienco D, Pensa RG, Meo R. From context to distance: Learning dissimilarity for categorical data clustering. ACM Transactions on Knowledge Discovery from Data; 2010.
21. Yao Y. Information-theoretic measures for knowledge discovery and data mining. In: Karmeshu, editor. Entropy measures, maximum entropy and emerging applications. Berlin: Springer; 2003. p. 115–36.
22. Quinlan JR. Learning efficient classification procedures and their application to chess end games. In: Michalski RS, Carbonell JG, Mitchell TM, editors. Machine learning: An artificial intelligence approach; Los Altos: Morgan Kaufmann; 1983.
23. Quinlan JR. Induction of decision trees. Machine Learning. 1986; 1:81–106.
24. Han J, Micheline K, Pei J. Data mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann; 2011.
25. Xu ST, Zhang J, Han D, Wang J. A singular value decomposition based data distortion strategy for privacy protection. KAIS Journal; 2006.
26. Frank A, Asuncion A. UCI machine learning repository. Irvine, CA: School of Information and Computer Science, University of California; 2010. Available from: http://archive.ics.uci.edu/ml
27. Cohen WW. Fast effective rule induction. 12th International Conference on Machine Learning; 1995. p. 115–23.
28. Mozafari B, Zaniolo C. Publishing naive Bayesian Classifiers: Privacy without accuracy loss. 35th International Conference on Very Large Data Bases (VLDB); 2009.