Two Wavelet-Based Data Distortion Methods for Privacy Preserving Classification

Mirmorsal Madani

Faculty Member of Computer Engineering Department
Islamic Azad University, Gorgan Branch, Kordkuy Center

Phone Number: +989122864708

*Corresponding Author's E-mail: mt_madani@yahoo.com

Abstract

With rapid development of modern data collection and data warehouse technologies, data mining is becoming more and more a standard practice. Accompanying this trend, preserving privacy in certain data becomes a challenge for data mining applications in many fields, especially in medical, financial and homeland security fields. The main objective of data mining is generalization of information and not to popularization of private data. We can implement Data mining and maintain the privacy preserving capability of data simultaneously. Privacy of preserving data mining algorithms has been recently introduced with the aim of preventing the discovery of sensible information. This paper introduces a privacy-preserving data distortion method in the collaborative analysis situations based on wavelet transformation, which provide an efficient balance between data utilities and privacy preserving beyond its fast run time. In suggested method a dynamic technique was used for engaging between criteria of privacy and accuracy. It provides data owners the ability to make decision for each attribute independently in terms of level of their sensitivity. Wavelet transform and inverse wavelet transform keep Euclidean distance and quality of man data, in this paper KNN algorithm was used for classifying data which are compatible with Euclidean distance for keeping statistical features of data. Experiments on real-life datasets indicate high potential of the suggested method in terms of accuracy, run time and level of privacy.

Keywords: cardiovascular disease, neural network, learning algorithms.

1. Introduction

Data mining is the process of exploring the relations, patterns, and meaningful trends. In other words, it extracts new potentially meaningful information from data sources (databases and texts, web items and pictures, etc.) in order to make utilization of data more possible, as well as to fill the gap between recorded data and available data. Data mining includes different algorithms such as Clustering, Classification and Association Rule Mining, which are traditionally used for extracting information from data which exist in a central location. In this paper we deal with classification. Putting the privacy under consideration can cause restriction of access, because when data is spread in different sites and we need to access them, there is possibility that data owners refuse providing access to data, fearing their private information to be exposed. We should find a way to implement data mining in a way that maintains the privacy of data owners. In this paper we present a method for distortion data based on wavelet transform. Efficiency of the method will be reviewed in testing.
and training phases in terms of accuracy, run time and privacy. In the second section related work in privacy data mining were demonstrated. In the third section KNN classifier are presented. In the fourth section privacy data mining methods, especially wavelet based data distortion are briefly explained. In the next fifth we describe the suggested method and experiment results and in the end of the paper conclusion and future works are presented.

2. Related work

Eufimievski et al in Evfimevski(2003) and Rizvi (2005) expressed a similar method on extracting dependency rules and provided some methods for limiting violation of privacy. In Pinkas(2002) a wide view of SMC framework and its applications in data extracting was provided. In Goldreich (2004) a full paper was provided regarding SMC in which a function with circuit math was established and applied secretly for computations through public protocol of circuit assessment. Hidden subscription was introduced independently by Shamir (2010) and Blakley (1979), where one person has hidden information which is shared among n people and no one can restore information by itself. In fact, the data are shared in a way that information of at least people from n people is necessary for restoring necessary information that should be pre-determined. Muthu et al (2012) provided a method for keeping privacy in extracting dependency rules where data transformation method was applied. Lambodar Jena et al (2011) deal with methods of transforming data such as hiding rules, obscurity level k and generalization of data.

D. ArunaKumari et al (2011) deal with encryption methods, generalization and data corruption in clustering and they provided an efficient method. Md. GolamKaosar et al (2011) provided some solutions for extracting dependency rules in two-player model using symmetric encryption method. Some algorithms were presented for blocking in Saygin(2002) with the aim of reducing loss of non-sensitive rules and establishing fake rules during process of hiding dependency rules. Aggarwal(2006) deal with using methods of blocking in order to prevent from revelation of sensitive inputs from data set in which inputs are classified in form of sensitive data and only the rules relating inputs are hidden. Another method for transforming data is clearing a set of data. Data owners usually sanitize their data and seek to block inference networks so that their opponents cannot penetrate into their key information. As an example, we can mention QIBC which was provided by Ahmed Haj Yasien (2007). Recently, matrix decomposition and factorization techniques have been used to distort numerical valued datasets in the applications of privacy-preserving data mining. In particular, singular value decomposition (SVD) [Xu (2006), Zhang (2005)] and nonnegative matrix factorization (NMF) by Wang (2006) have been shown to be very effective in providing high level data privacy preservation and maintaining high degree data utilities.

In addition to the above documents, signal transformation methods related to Fourier or wavelet transformation have been used as strategies for data perturbation [Pinkas (2002), Mukherjee(2006), and Xu (2006)]. Both transformation based privacy preserving distortion methods seem to have a very good property on privacy protection and data utility preservation. The run time complexity of the wavelet-based transformation O(t) is better than the O(tlogt) run time of the Fourier transformation, where t is the maximum level number of wavelet or Fourier decompositions. Therefore, data analysts may prefer the wavelet-based methods which have a very attractive merit and fast run time, in dealing with very large datasets. However, Bapna (2006) presented the wavelet perturbed dataset in the transformed space has different dimensions from that in the original space. This might create a problem when a third party data miner or the collaborative analyst has data parts from different sources to match each other. There is certainly an advantage to consider the
transformed dataset that keeps the same dimension as the original dataset in the collaborative data analysis situation. Therefore, we propose a different set of data distortion, suppression and reconstruction (transformation back) strategies based on wavelets to keep the dimensions of the original and distorted datasets. In the paper presented by Lian Liu et al (2015). Wavelet-based transformation was used. They transformed dataset in two forms of vertically partitioned data and horizontally portioned data and used Haar wavelet and db4 wavelet.

3. KNN algorithm

Classification is one of data mining methods in which learning is done through supervision and it is conducted in two phases of training and test. Classification methods can be divided into two forms of lazy and eager. In lazy method, training samples are stored but their model will not be established. So it is not efficient for big datasets. But in eager method, before conducting classification operation training samples are modeled. Therefore, training period and test periods are different. K-Nearest Neighbor algorithm (KNN) is a sample of lazy method with high speed process of classification.

KNN algorithm is applied for classifying data and also for estimating and prediction. It is one of the most simple classification algorithms and it is a sample of Instance-based learning where total training data are memorized and Classification is done only if test instance is compatible with features of one of instances of training sets. For the instance we may not be able to find any instance of training set during test. Or we may find some instances which are compatible with them in which target class label is specified based on majority class label.

to perform KNN, following process must be followed:
K-Nearest Neighbor (k: number of nearest neighbor, E: training instances, z: unlabeled instance).
1. Compute the distance or similarity of z to all the training instances.
2. Let E’⊂E be the set of k closed training instances to z.
3. Return the predicted class label for z: class ← Voting (E’).

In this paper we used KNN algorithm for classifying data, which is compatible with Euclidean distance and is appropriate for keeping statistical features of data. We should bear in mind that wavelet transform makes data wavelet and inverse wavelet transform can transform data to main field without considering without losing fundamental data. K optimum value was specified based on several experiments for values k=1...25.

\[
d_{\text{Euclidean}}(x, y) = \sqrt{\sum_{i}(x_i - y_i)^2}
\]

3.1. Assessment method

One of assessment methods is K-Fold Cross-Validation which is used for assessing classifiers where set of data available is divided into K (equal independent set). Each of subsets are used as test set and other ones are used as training set for training classification. This operation is repeated K times so that K accuracy is obtained in K step which is in fact accuracy of final assessment. In fact, for conquering fitting and measuring performance of training, this method is used.

4. Privacy of preserving data mining

The overall goal of data mining process is to extract information from a data set and transform it into an understandable structure for further use that described by Clifton(2010). In privacy preserving
data mining several participants, who are called data owners, attend an operation called *datasharing*. Some of data indicate private information of owners which are called sensitive data, therefore they should not be revealed for other data owners. Once data owners specified sensitive data, they need to keep privacy for preventing from data revelation during using algorithms of data extraction. Various methods of data extraction are different with keeping privacy in terms of accuracy in results, run time, computation complexity, communication overhead and level of revelation where selecting an appropriate method needs engagement between the factors.

Methods available on privacy preserving data mining are classified into two main groups. They include secure multiparty computation and data distortion methods. Secret sharing methods form blocks of secure multiparty computation. Methods of rule hiding, data swapping, data sanitization, group based anonymization are all methods of data transformation such as wavelet based data perturbation and they belong to data distortion. Secure multiparty computation has higher level of security and accuracy compared to data distortion. Since data distortion methods transform main data into noisy data so that the data will not be accessed by other data owners. But on the other hand, in data distortion level of complexity, overhead and run time is less compared to secure multiparty computation. Generally in most of data distortion methods, the distance between data values are not far which reduces accuracy and performance of methods of data extraction based on distance. Also, in these methods reduction of data set is not considered. But methods of data transform which are based on distance seek to keep privacy. They also seek to transform data set so it can be used for algorithm of data mining available. We can use methods of geometric data transformation through method of transform, ratio transform method, period method, and wavelet-based transform. In this paper we used wavelet-based transform.

### 4.1. Wavelet-based data transformation

One wavelet is function of $L^2(R)$ which is defined as follows:

$$\psi_{j,k}(x) = 2^{-j/2}\psi(2^j x - k), \quad j, k \in \mathbb{Z}$$  \hspace{1cm} (2)

Wavelet means small wave. And it means it has unlimited length and wave indicates fluctuation of function which is unlimited and this is an advantage. Transforming wavelet provides the condition for analysis of data, functions or pragmatic for different frequency elements and then every single elements can be studied. Wavelets have features such as compact support, vanishing moment, Hierarchical and Multiresolution decomposition, timing and spatial complexity of linear transformation. They have the maximum level of number of wavelet decompositions in dilating relation and correlated coefficients and provide many basis function.

Compact support guarantees wavelets being local. Vanishing moment guarantees that wavelets have the ability of recognizing important information from unnecessary information (coefficients).

Coefficients make wavelet capable of time correlation reduction, hence wavelet can be transformed for reducing complexity of processes in time field to more simple processes. According to features of wavelet, we can consider them right tool for data extraction operation.

Another feature is Parseval’s Theorem expressing that the energy, which is defined based on squares of values $L^2$, can be maintained under trusted transformation wavelet.

$$\|e\|^2_2 = \sum_i |\langle e, \psi_i \rangle|^2 \quad \text{(3)}$$
From the theory we can conclude that wavelet transformation can provide data in a form that data extraction process can be conducted with high level of accuracy. Secondly, wavelets can be combined with core of many data extraction algorithms. Therefore, we don’t need to create new algorithms after transformation.

Discrete wavelet transformation (DWT) divides input single into two elements in each section. Elements produced are coefficients and details indicating low frequencies and high signal, respectively. Coefficients can be analyzed to low-level and high-level sub signals and multiresolution analysis is possible. In many signals coefficients are significant, since main data are maintained with low frequency and they should not be removed. As it can be observed in figure, main signal is analyzed in the first level to two elements of \( c_{A1} \) and \( c_{D1} \) indicting coefficients and detail, respectively. Noise is cleaned up from the main signal.

![Figure1: Analysis levels in wavelet transform](image)

5. Suggested method

The Suggested method (WBT) has two sections. The algorithm presented in the first section for distortion dataset are applied for maintaining privacy of sensitive data and are wavelet-based transform. Higher level of accuracy leads to reduction of privacy and it is effective on run time factor. Hence, we need to make tradeoff between factors of accuracy, run time and privacy preserving. Second part is for data extraction process where KNN classifier is used for classification of input data and predicting the class related. In KNN classifier the accuracy in prediction of target class depends on parameter \( k \), so in this section the method suggested was performed for finding optimum value of \( K \) several times for values of \( k=1...25 \).

The method suggested for vertically partitioned data can be applied in both two-participant state and multiparty state. In vertically partitioned datasets, datasets of all the participants including transactions, are equal with Tid but each one has their own attributes. Combining them with each other can lead to dataset where combination operation in two-participant state is done by participant A and B in both sides. In multiparty state we used semi honest third-party with the aim of reducing communication overhead. Semi honest is a participant who follows protocol but it may be curious and it does not have any information except what participants send for him.

5.1. WBT method

Several participants have their own dataset but they are related to same transactions (Tid), with the aim of obtaining better results. But they seek for perturbation of their sensitive data, since they are afraid of revelation of their private data. This is called heterogeneous collaboration which is shown in figure 2.
5.1.1. Multi party state:

Hypotheses:
1. We use semi honest third-party for communication between participants.
2. All participants don’t colluding with each other or third person.

Participants Phase:
- for each attribute (i): Convert attributes to signal and add to signal matrix using db6 wavelet.
- for each attribute (i) without target class: converted array (CAi, CDi) = Transform (attribute (i)).
  - Determining sensitive attributes and level of their sensitivity by each participant.
- for each sensitive attribute (i):
  - Determining damping factor value in (0...1). (The more the level of sensitivity, the bigger the level of damping factor).
  - Reduce the high frequency "noise" (CDi) using the Reduced technique.
- In IDWT phase: final converted array (i) = CAi + CDi.
  - Use the normal technique on final converted array (i) and send to third party.

Third party phase:
- Receiving distortion datasets from participants
- Combining datasets and establishing main dataset
- Using KNN for classification and predicting target class
  - Determining optimum k value and level of accuracy of classification based on error percentage using resublossand k-fold loss with several times of testing
- Sending results to participants

Reduced technique and signal normalization phase:
- assume β=damping_factor + (0.1*(rand (1)).
- for each CDi: (i is the number of dataset records)
  - Calculate Local_avg(i) using average of 4 nearest neighbors of CDi
  - dis(i)=CDi - Local_avg(i)
  - Δ = β * dis(i)  
    (Local_avg(i) is local thresholding value for each CDi)
  - CDi = CDi – Δ         CDi>Local_avg (i)
  - CDi = CDi + Δ         CDi<Local_avg (i)
    Delete CDi  o.w
- Final converted array (i) = final converted array (i) + min of attribute (i).
5.1.2. Two-participant state:

The method in two-participant state is like multiparty state. For increasing privacy, combination operation and data mining process are conducted independently by each participant.

Alice and Bob established their distortion dataset using WBT after determining sensitive attributes and level of sensitivity and then they send it to each other.

Alice has access to original dataset A and distortion dataset B. Bob has access to original dataset B and distortion dataset A. As we see in figure 3, each of two participants are able to establish main dataset by combining these datasets which is as an input for data mining algorithm.

There is no certain framework for recognizing sensitivity of data. But data owners can recognize attributes which directly lead to revelation of personal information of individuals such as ID, level of salary, they can also recognize insensitive attributes by which we may find sensitive data of individuals indirectly such as zip code and sex. As we know, criteria of accuracy and privacy are against each other and rise of them leads to reduction of the other one. Therefore, we prioritize one of them in terms of main objective of partnership and this way we will obtain an efficient engagement. By giving privacy priority in WBT method, we considered all attributes in both datasets sensitive attributes. It is clear that this can reduce accuracy and increase privacy. If we consider priority of accuracy more than privacy, then we need to add follow pre-processing phase in WBT method in which accuracy of prediction in process of clustering will increase. If we call this new method WBTperim, we will have risen of accuracy to 7 % in WBC dataset and 13 % in WDBC dataset, compared to WBT method. It is clear that values of criteria of privacy will reduce. Based on spearman correlation coefficient, attributes which are not related to target class or have low level are considered sensitive attributes.

- Preprocess Phase:
  - Calculating Spearman correlation between all attributes and target class.
  - If corr (att (i), target class) < mean then add to insensitive attribute (i) (mean is the average value of all of calculated correlation)

6. Experimental Results

6.1. Data Privacy Measures

The five data distortion privacy measure metrics, VD, RP, RK, CP and CK, are defined in Wang (2006), and then Xu (2006), to evaluate the proposed data distortion methods. The
The objective of these measure metrics is to evaluate the possibility of estimating and predicting the true values and range of the original data from the distorted data [6].

\[
VD = \frac{||A - \tilde{A}||_F}{||A||_F}
\]  

(4)

Where \(A\) is the original dataset and \(\tilde{A}\) is the perturbed version of \(A\), and \(||A||_F\) is the Frobenus norm of the matrix \(A\). The Frobenus norm is a matrix norm.

The RP value presents the ratio of the average change of ranks for all attributes to the number of total elements of the matrix. Its definition is as follows:

\[
RP = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{\sum_{j=1}^{n} |Rank_{ij} - Rank_{ij}^*|}{n} \right)
\]  

(5)

Where for the \(n \times m\) dataset \(A\), \(Rank_{ij}\) denotes the rank in the ascending order of the \(j\)-th element in the attribute \(i\), and \(Rank_{ij}^*\) denotes the rank in ascending order of the perturbed version \(\tilde{A}\).

RK denotes the percentage of elements which keep their ranks of values in each column after the distortion.

\[
RK = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{\sum_{j=1}^{n} RK_{ij}^*}{n} \right)
\]  

(6)

Where \(RK_{ij} = 1\) if \(Rank_{ij} = Rank_{ij}^*\) and \(RK_{ij} = 0\) otherwise.

CP stands for change of ranks of the average values of the attributes.

\[
CP = \frac{\sum_{i=1}^{m} |RAVi - RAVi^*|}{m}
\]  

(7)

Where \(RAVi\) (resp. \(RAVi^*\)) is the rank in the ascending order of the average value of the \(i\)-th attribute at \(A\) (resp. \(\tilde{A}\)).

CK is defined to evaluate the percentage of the attributes that keep their ranks of average values after the distortion.

\[
CK = \frac{\sum_{i=1}^{m} CK_{i}}{m}
\]  

(8)

Where \(CK_{i} = 0\) if \(RAVi = RAVi^*\), and \(CK_{i} = 0\) otherwise.

According to their definitions, it is clear that a larger \(VD\), \(RP\) and \(CP\), and a smaller \(RK\) and \(CK\) value refers to a better privacy-preserving level.
6.2. Distortion Experiments

For experiment section, two real-life datasets are obtained from Machine Learning Repository in Newman (2007) at the University of California, Irvine (UCI). They are the Wisconsin breast cancer original dataset (WBC) donated by Olvi Mangasarian in which 699 instances with 9 features are in 2 classes, and the Wisconsin breast cancer diagnostic database (WDBC) donated by Nick Street where 599 instances with 30 features also belong to 2 classes. The attributes both of the WDBC and WBC datasets only have numerical values. There is no missing value in WDBC dataset, but in the original WBC database, there are a few missing values in the sixth column. These missing values are replaced by 1 if the object belongs to the malignant class and 2 if the object is in the benign class, according to the standard classification provided by the UCI Repository. The properties of these datasets are shown in tables 1 and 2.

<table>
<thead>
<tr>
<th>Table 1: Properties of WBC dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set Characteristics: Multivariate</td>
</tr>
<tr>
<td>Attribute Characteristics: Integer</td>
</tr>
<tr>
<td>Number of Instances: 699</td>
</tr>
<tr>
<td>Number of Attributes: 10</td>
</tr>
<tr>
<td>Area: Date Donated</td>
</tr>
<tr>
<td>Life: 1952-07-15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Properties of WDBC dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set Characteristics: Multivariate</td>
</tr>
<tr>
<td>Attribute Characteristics: Real</td>
</tr>
<tr>
<td>Number of Instances: 599</td>
</tr>
<tr>
<td>Number of Attributes: 32</td>
</tr>
<tr>
<td>Area: Date Donated</td>
</tr>
<tr>
<td>Life: 1995-11-01</td>
</tr>
</tbody>
</table>

6.3. Experimental

6.3.1. WBT method on WBC dataset:

Original signal, coefficients (low frequency) and details (high frequency) of the first attribute on WBC dataset using WBT method are brought in figure 4.
Figures 5 and 6 are for performing reduced technique and signal normalization for WBC dataset. For simplicity level of sensitivity of attributes in both datasets were considered equal (impact factor=0.01). For calculating Local_avg (i) is used average value of 4-Nearest Neighbors for each detail (i).

![Figure 4: Original, Coefficients and details signal of att1 of WBC](image)

![Figure 5: Details vs Local_avg of att1 of WBC](image)
In figure 7 original data and transformed data that obtained using IDWT method in participant phase were brought in WBC dataset.

6.3.2. WBT method on WDBC dataset

Original signal, coefficients (low frequency) and details (high frequency) of the first attribute for WDBC dataset using WBT method are brought in figure 8.
Figures 9 and 10 are for performing reduced technique and signal normalization for WDBC dataset. For simplicity level of sensitivity of attributes in both datasets were considered equal (impact factor=0.01). For calculating Local_avg (i) is used average value of 4-Nearest Neighbors for each detail (i).

Figure 8: Original, Coefficients and details signal of att1 of WDBC

Figure 9: Details vs Local_avg of att1 of WDBC
In figure 11 original data and transformed data that obtained using IDWT method in participant phase were brought in WDBC dataset.

Figures 12 and 13 show accuracy rate and loss of classification predicting using KNN based on values k=1…25 in first run the WBT in WBC and WDBC datasets. The average values of accuracy and loss were illustrated in tables 3 and 4.
In WBTperim method after recognizing sensitive attributes in WBC and WDBC datasets, correlation of all the attributes was calculated through target class using Spearman correlation. Attributes with correlation less than average are insensitive attributes and their values do not change. Since these attributes have weak or even reverse correlation with target class and can be effective on accuracy. In WBC and WDBC the WBTperim technique averagely led to rise of accuracy in classification results (7% and 13% respectively). In figures 14 and 15 Spearman correlation and also average value were presented in WBC and WDBC datasets.

**Figure 13.** Accuracy and loss rate of classification predicting in WBT method on WDBC in first run

6.3.3. WBTperim method on WBC and WDBC dataset

In WBTperim method after recognizing sensitive attributes in WBC and WDBC datasets, correlation of all the attributes was calculated through target class using Spearman correlation. Attributes with correlation less than average are insensitive attributes and their values do not change. Since these attributes have weak or even reverse correlation with target class and can be effective on accuracy. In WBC and WDBC the WBTperim technique averagely led to rise of accuracy in classification results (7% and 13% respectively). In figures 14 and 15 Spearman correlation and also average value were presented in WBC and WDBC datasets.

**Figure 14:** Spearman correlation between WBC attributes and target class

**Figure 15:** Spearman correlation between WDBC attributes and target class

In WBC and WDBC the WBTperim technique averagely led to rise of accuracy in classification results (7% and 13% respectively). Figures 16 and 17 show the accuracy of WBTperim in WBC and WDBC datasets.
Figure 16: Accuracy and loss rate of classification predicting in WBTperim method on WBC in first run

Figure 17: Accuracy and loss rate of classification predicting in WDBTperim method on WDBC in first run

In table 3 and table 4 average values of privacy measure metrics and performance obtained from 30-times run of suggested methods were shown.

Table 3. Performance comparison of suggested methods and other methods on WBC

<table>
<thead>
<tr>
<th>Database</th>
<th>VD</th>
<th>RP</th>
<th>RK</th>
<th>CP</th>
<th>CK</th>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96.0%</td>
</tr>
<tr>
<td>SVD[Xu (2006)]</td>
<td>0.2080</td>
<td>239.4</td>
<td>0.006358</td>
<td>1.556</td>
<td>0.4444</td>
<td>0.07882</td>
<td>95.9%</td>
</tr>
<tr>
<td>Wavelet(S) [ Liu (2015)]</td>
<td>0.2557</td>
<td>238.6</td>
<td>0.004769</td>
<td>1.333</td>
<td>0.5556</td>
<td>0.03081</td>
<td>96.0%</td>
</tr>
<tr>
<td>Wavelet(VP) [ Liu (2015)]</td>
<td>0.3526</td>
<td>247.1</td>
<td>0.005564</td>
<td>1.556</td>
<td>0.333</td>
<td>0.06362</td>
<td>95.6%</td>
</tr>
<tr>
<td>Wavelet(HP) [ Liu (2015)]</td>
<td>0.3140</td>
<td>239.1</td>
<td>0.005087</td>
<td>2.000</td>
<td>0.333</td>
<td>0.05153</td>
<td>96.1%</td>
</tr>
<tr>
<td>Improved SVD by G.Li, Y.Wang[Li(2012)]</td>
<td>0.31</td>
<td>77.20</td>
<td>0.01</td>
<td>1.04</td>
<td>0.44</td>
<td>Na</td>
<td>Na</td>
</tr>
<tr>
<td>NMF[Wang (2006)]</td>
<td>0.1228</td>
<td>228.42</td>
<td>0.01</td>
<td>0.22</td>
<td>0.78</td>
<td>Na</td>
<td>96.7%</td>
</tr>
<tr>
<td>Algorithm by Ma. Kadampur[Kadampur(2010)]</td>
<td>0.10</td>
<td>239.0</td>
<td>0.0051</td>
<td>2.00</td>
<td>0.333</td>
<td>Na</td>
<td>0.70</td>
</tr>
<tr>
<td>WBT</td>
<td>0.8755</td>
<td>339.8</td>
<td>0.001412</td>
<td>307.333</td>
<td>0</td>
<td>0.05734</td>
<td>87.8%</td>
</tr>
<tr>
<td>WBTperim</td>
<td>0.5612</td>
<td>188.72</td>
<td>0.4352</td>
<td>174.666</td>
<td>0.4444</td>
<td>0.05242</td>
<td>95.1%</td>
</tr>
</tbody>
</table>
Table 4: Performance comparison of suggested methods and other methods on WDBC

<table>
<thead>
<tr>
<th>Database</th>
<th>VD</th>
<th>RP</th>
<th>RK</th>
<th>CP</th>
<th>CK</th>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>85.4%</td>
</tr>
<tr>
<td>SVD[Xu (2006)]</td>
<td>0.0003</td>
<td>5</td>
<td>121.3</td>
<td>0.3454</td>
<td>0</td>
<td>1.0000</td>
<td>0.1388 0</td>
</tr>
<tr>
<td>Wavelet[S][Liu (2015)]</td>
<td>0.00084</td>
<td>3</td>
<td>165.3</td>
<td>0.1083</td>
<td>4.800</td>
<td>0.4000</td>
<td>0.0516 6</td>
</tr>
<tr>
<td>Wavelet(VP) [Liu (2015)]</td>
<td>0.00101</td>
<td>1</td>
<td>168.6</td>
<td>0.1041</td>
<td>4.733</td>
<td>0.4667</td>
<td>0.0927 4</td>
</tr>
<tr>
<td>Wavelet(HP) [Liu (2015)]</td>
<td>0.00096</td>
<td>2</td>
<td>165.5</td>
<td>0.1141</td>
<td>3.267</td>
<td>0.4667</td>
<td>0.0515 3</td>
</tr>
<tr>
<td>Improved SVD by G.Li, Y.Wang[Li (2012)]</td>
<td>0.07</td>
<td>90.9</td>
<td>0.01</td>
<td>0.2</td>
<td>0.8</td>
<td>Na</td>
<td>Na</td>
</tr>
<tr>
<td>WBT</td>
<td>0.3727</td>
<td>23.3</td>
<td>0.0807</td>
<td>60.133</td>
<td>0</td>
<td>0.0801 3</td>
<td>82.8%</td>
</tr>
<tr>
<td>WBTperim</td>
<td>0.0193</td>
<td>16.4</td>
<td>0.5636</td>
<td>25.566</td>
<td>7</td>
<td>0.5333</td>
<td>0.0700 1</td>
</tr>
</tbody>
</table>

- The method with better performance was highlighted in bold font. As it can be seen in tables 3 and 4, in both datasets, run time of both suggested methods is less than SVD and Wavelet (VP) and it is a few more than the other methods.
- In terms of privacy, WBT method in both datasets have higher level of confidentiality compared to other methods. But accuracy level is less than other methods. Also WBTperim has better performance compared to SVD in terms of privacy in both datasets and in compared to other methods in WBC dataset.
- Accuracy level in WBTperim in dataset of WBC is equal with other methods, but in WDBC it is significantly higher than other methods.

Conclusion and future works

In this paper the two methods were presented based on wavelet-based transform. Using wavelet keeps statistical features of primary data, accuracy of the results has higher potential compared to other data distortion method. In normalizing phase of details in reduced technique, some coefficients are removed, hence in both methods we witness reduction of scalability which will have significant effect, especially when data size is big. As it was observed before, both methods has fine position compared to other methods in terms of privacy. The WBTperim in WDBC has better accuracy in forecasting target class compared to other method in which features have reverse correlation with target class. In the method suggested data owners can determine level of privacy for each attribute based on their recognition which adds dynamism to the method and leads to rise of privacy. Hence, if an individual has access to one attribute, he/she will not be able to recognize values for other attributes. As we know, there are other methods for data mining such as association rule mining and clustering. We may present some efficient algorithms in terms of accuracy and privacy for the methods. We can also use other methods of privacy preserving such as methods secure multi party computation or other methods based on data distortion such as randomization, k-anonymity and so on independently and in form of a mixture with wavelet-based transform.
References


