Application of Wavelets in Privacy-preserving Data Mining

Jun Zhang
Department of Computer Science
University of Kentucky
Outline

• Privacy-preserving in Collaborative Data Analysis
• Advantages of Wavelets
• Introduction to Wavelet Analysis
• Maintaining Statistical Analysis
• Experimental Analysis
• Summary
Challenges in Data Mining Analysis

- In data mining applications, data privacy-preserving is an issue in need of urgent attention. This is particularly important in collaborative business data analysis, medical record analysis, national security, etc.
Collaborative Data Mining Analysis

- Two or more data owners want to share their data, to obtain useful information from the combined data, and to benefit both.
- In this collaborative data analysis, due to business competition and relevant privacy laws, each party does not want the other party to have his/her original data.
Collaborative Data Mining

- **Aim:** Achieve collaborative data mining analysis

- **Requirements:**
  - Keep each party’s private data
  - Maintain data utilities of the combined data

**Mr. Perturbation:**
I can satisfy your both requirements
Collaborative Data Mining

- **Aim:** Achieve collaborative data mining analysis

- **Requirements:**
  - Keep each party’s private data
  - Maintain data utilities of the combined data

- **Procedure:** Before collaboration, each party perturb his/her own data with certain constraints. The perturbation can hide the original data values, but maintain its data patterns, which will maintain the results of certain data mining algorithms

*Mr. Perturbation:* I can satisfy your both requirements!
Collaborative Data Mining

Procedure: Before collaboration, each party perturb his/her own data with certain constraints. The perturbation can hide the original data values, but maintain its data patterns, which will maintain the results of certain data mining algorithms.

Assumptions:
1. Both datasets are in numerical value
2. Vertical collaborative data analysis: Both have the same customers
3. Horizontal collaborative data analysis: Both have the same data attributes

Mr. Perturbation: I can satisfy your both requirements!
**Vertical Collaborative Data Analysis:**
Both have the same customer set

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Weight</th>
<th>Height</th>
<th>Math</th>
<th>Language</th>
<th>Credit Score</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>User1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Government
- Hospital
- School
- Bank
Horizontal Collaborative Data Analysis: Horizontal Collaborative Data Analysis: Both have the same attribute set

<table>
<thead>
<tr>
<th>Gend</th>
<th>Age</th>
<th>Weig</th>
<th>Heigh</th>
<th>Math</th>
<th>Lanuag</th>
<th>Credit</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>User1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>KY</td>
</tr>
<tr>
<td>User2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>KY</td>
</tr>
<tr>
<td>User3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CA</td>
</tr>
<tr>
<td>User4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CA</td>
</tr>
<tr>
<td>User5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NY</td>
</tr>
<tr>
<td>User6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NY</td>
</tr>
<tr>
<td>User7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NY</td>
</tr>
</tbody>
</table>
Outline

- Privacy-Preserving in Collaborative Data Mining Analysis
- Advantages of Wavelet Analysis
- Introduction to Wavelet Analysis
- Maintaining Statistical Values
- Experimental Analysis
- Summary
Why do we use wavelet for decomposition?

- Lower time complexity, for an n*m data matrix

<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>wavelet</td>
<td>O(n)</td>
</tr>
<tr>
<td>Fourier</td>
<td>O(n log n)</td>
</tr>
<tr>
<td>SVD</td>
<td>O(mn^2)</td>
</tr>
<tr>
<td>PCA</td>
<td>O(mn^2)</td>
</tr>
</tbody>
</table>
Why do we use wavelet for decomposition?

- Low time complexity
- Can simultaneously decompose Frequency and Phase

Frequency: The curve represents frequency. E.g., how many points have y-coordinate greater than $t$?

Frequency analysis: Global analysis
Why do we use wavelet for decomposition?

- Low time complexity
- Can simultaneously decompose

Frequency and Phase

Phase: local frequency analysis in some range

E.g., for $x=[2-\varepsilon, 2+\varepsilon]$, how many points’ $y$-coordinates are greater than $t$?

Frequency analysis: Global analysis
We do we use wavelet for decomposition?

- Low time complexity
- Can simultaneously decompose Frequency and Phase

Advantage: Seem to be good for collaborative data mining analysis

E.g.: In combined data, fine the average age of all males (global property). For males between 30-45, what is the average age for having children (local property)
We do we use wavelet for decomposition?

- Low time complexity
- Can simultaneously decompose Frequency and Phase
- Flexibility: In wavelet analysis, we can choose different wavelet basis (there are currently more than 20 popular wavelet basis). This implies that there is no need for all data owners to choose the same wavelet basis for data perturbation
Why do we use wavelet decomposition?

- Low time complexity
- Can simultaneously decompose Frequency and Phase
- Flexibility

Independence: Wavelet decomposition does not depend on the statistical distribution of the original data. It does not depend on the data analysis method used by other collaborators.
Outline

- Privacy-Preserving In Collaborative Data Mining
- Advantages of Wavelet Decomposition
- Introduction to Wavelet Decomposition
- Maintaining Statistical Properties
- Experimental Analysis
- Summary
Wavelet Decomposition

- For a 1-D discrete numerical data \( x \) (length \( L \)), wavelet decompose the data into high frequency \( y_{\text{high}} \) components and low frequency \( y_{\text{low}} \) components. High frequency components are also called detail coefficients, low frequency components are called approximation coefficients.

\[
y_{\text{low}}[l] = \sum_{k=-\infty}^{\infty} x[k]g[2l - k]
\]

\[
y_{\text{high}}[l] = \sum_{k=-\infty}^{\infty} x[k]h[2l - k]
\]

Here \( h \) and \( g \) are high and low frequency filters, respectively.
1D Wavelet Decomposition Example (with Haar basis)

Input data

| 39.00 | 33.00 | 3.00 | 6.00 | 23.00 | 24.00 | 24.00 | 1.00 |

0th level

Original data
1D Wavelet Decomposition Example (with Haar basis)

Original data: [39.00, 33.00, 3.00, 6.00, 23.00, 24.00, 24.00, 1.00]

0th level:
- (39+33) / √2 = 50.91

1st level:
- (39-33) / √2 = 4.24

Detail coeff: [50.91, 4.24]
1D Wavelet Decomposition Example (with Haar basis)

Original data: 39.00 33.00 3.00 6.00 23.00 24.00 24.00 1.00

Detail coeff: 6.36

Approx coeff: -2.12
1D Wavelet Decomposition Example (with Haar basis)

0th level:
- 39.00
- 33.00
- 3.00
- 6.00
- 23.00
- 24.00
- 24.00
- 1.00

1st level:
- 50.91
- 6.36
- 33.24
- 17.68
- 4.24
- -2.12
- -0.71
- 16.26

Original data
Detail coeff
Approx coeff
ID Wavelet Decomposition (with Haar basis)

Detail coeffs will not be processed further
ID Wavelet Decomposition Example (with Haar basis)

Original data: 39.00, 33.00, 3.00, 6.00, 23.00, 24.00, 24.00, 1.00

0th level
- 39.00
- 33.00
- 3.00
- 6.00
- 23.00
- 24.00
- 24.00
- 1.00

1st level
- 50.91
- 6.36
- 33.24
- 17.68
- 4.24
- -2.12
- -0.71
- 16.26

2nd level
- 40.5
- 36.0
- 31.5
- 11
- 4.24
- -2.12
- -0.71
- 16.26

3rd level
- 33.24 + 17.68

Approx coeff

Detail coeff

Original data
ID Wavelet Decomposition (with Haar basis)

Detail coeffs will not be processed further
ID Wavelet Decomposition Example (with Haar basis)

<table>
<thead>
<tr>
<th>Original data</th>
<th>Detail coeff</th>
<th>Approx coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>39.00</td>
<td>4.24</td>
<td>16.26</td>
</tr>
<tr>
<td>33.00</td>
<td>-2.12</td>
<td>-0.71</td>
</tr>
<tr>
<td>3.00</td>
<td>-0.71</td>
<td>16.26</td>
</tr>
<tr>
<td>6.00</td>
<td>4.24</td>
<td>-0.71</td>
</tr>
<tr>
<td>23.00</td>
<td>-2.12</td>
<td>16.26</td>
</tr>
<tr>
<td>24.00</td>
<td>-0.71</td>
<td>16.26</td>
</tr>
<tr>
<td>24.00</td>
<td>4.24</td>
<td>-0.71</td>
</tr>
<tr>
<td>1.00</td>
<td>-2.12</td>
<td>16.26</td>
</tr>
<tr>
<td>50.91</td>
<td>4.24</td>
<td>-0.71</td>
</tr>
<tr>
<td>6.36</td>
<td>-2.12</td>
<td>16.26</td>
</tr>
<tr>
<td>33.24</td>
<td>-0.71</td>
<td>16.26</td>
</tr>
<tr>
<td>17.68</td>
<td>4.24</td>
<td>-0.71</td>
</tr>
<tr>
<td>54.09</td>
<td>4.24</td>
<td>-0.71</td>
</tr>
<tr>
<td>3.18</td>
<td>-2.12</td>
<td>16.26</td>
</tr>
<tr>
<td>31.5</td>
<td>-0.71</td>
<td>16.26</td>
</tr>
<tr>
<td>11</td>
<td>4.24</td>
<td>-0.71</td>
</tr>
<tr>
<td>(40.5 + 36.0)</td>
<td>(40.5 - 36.0)</td>
<td></td>
</tr>
</tbody>
</table>

(40.5 + 36.0) / \sqrt{2} = 54.09

(40.5 - 36.0) / \sqrt{2} = 3.18
**1D Wavelet Decomposition (with Haar basis)**

### Input data of wavelet decomposition

<table>
<thead>
<tr>
<th>39.00</th>
<th>33.00</th>
<th>3.00</th>
<th>6.00</th>
<th>23.00</th>
<th>24.00</th>
<th>24.00</th>
<th>1.00</th>
</tr>
</thead>
</table>

**0th level**

### Output data of wavelet decomposition

<table>
<thead>
<tr>
<th>54.09</th>
<th>3.18</th>
<th>31.5</th>
<th>11</th>
<th>4.24</th>
<th>-2.12</th>
<th>-0.71</th>
<th>16.26</th>
</tr>
</thead>
</table>

**3rd level**

<table>
<thead>
<tr>
<th>Original data</th>
<th>Detail coeff</th>
<th>Approx coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXX</td>
<td>XXX</td>
<td>XXX</td>
</tr>
</tbody>
</table>
2D Wavelet Decomposition (with Haar basis)

Two-Dimensional DWT

Decomposition Step

\[ cA_j \rightarrow \]

where

- Downsample columns: keep the even indexed columns.
- Downsample rows: keep the even indexed rows.
- Convolve with filter X the rows of the entry.
- Convolve with filter X the columns of the entry.

Initialization \[ CA_0 = s \] for the decomposition initialization.
2D Wavelet Decomposition (with Haar basis)

Two-Dimensional DWT

Decomposition Step

- **Lo_D**
- **Hi_D**

Downsample columns: keep the even indexed columns.
Downsample rows: keep the even indexed rows.
Convolve with filter X the rows of the entry.
Convolve with filter X the columns of the entry.

Initialization \( CA_0 = s \) for the decomposition initialization.

\[
\begin{bmatrix}
16 & 10 \\
0 & 8
\end{bmatrix}
\begin{bmatrix}
\frac{16 + 0}{\sqrt{2}} & \frac{10 + 8}{\sqrt{2}} \\
\frac{16 - 0}{\sqrt{2}} & \frac{10 - 8}{\sqrt{2}}
\end{bmatrix}
= \begin{bmatrix}
11.32 & 12.73 \\
11.32 & 1.414
\end{bmatrix}
\]
2D Wavelet Decomposition (with Haar basis)

Two-Dimensional DWT

Decomposition Step

- **rows**
  - Lo_D
  - Hl_D

- **columns**
  - Lo_D
  - Hl_D

\[ cA_{j+1} \]
\[ cD_{j+1}^{(h)} \]
\[ cD_{j+1}^{(v)} \]
\[ cD_{j+1}^{(d)} \]

**where**
- Downsample columns: keep the even indexed columns.
- Downsample rows: keep the even indexed rows.
- Convolve with filter X the rows of the entry.
- Convolve with filter X the columns of the entry.

Initialization: \( CA_0 = s \) for the decomposition initialization.

Illustration

\[
\begin{bmatrix}
16 & 10 \\
0 & 8 \\
\end{bmatrix}
= \begin{bmatrix}
11.32 & 12.73 \\
11.32 & 1.414 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
11.32 + 12.73 & 11.32 - 12.73 \\
11.32 + 1.414 & 11.32 - 1.414 \\
\end{bmatrix}
= \begin{bmatrix}
17 & -1 \\
9 & 7 \\
\end{bmatrix}
\]
2D Wavelet Decomposition (with Haar basis)

Two-Dimensional DWT

Decomposition Step

Input data

Output data

Initialization $CA_0 = s$ for the decomposition initialization.
Wavelet for Data Perturbation

- Fixed perturbation parameter

After decomposing the original data matrix with wavelet, we can perturb the wavelet coefficients

\[
\tilde{y}_{\text{high}[i]} = \begin{cases} 
    0, & \text{if } |y_{\text{high}[i]}| < \delta \\
    y_{\text{high}[i]} + \delta, & \text{if } |y_{\text{high}[i]}| > \delta \text{ and } y_{\text{high}[i]} < 0 \\
    y_{\text{high}[i]} - \delta, & \text{if } |y_{\text{high}[i]}| > \delta \text{ and } y_{\text{high}[i]} > 0
\end{cases}
\]

\(\delta\) is a user specified perturbation parameter. Different people can use different values of \(\delta\).
Wavelet Perturbation

- Automatic perturbation parameter
  After decomposing the original data matrix with wavelet, each party can perturb the wavelet coefficients

$$
\hat{y}_{high}[i] = \begin{cases} 
0, & \text{if } |y_{high}[i]| < \delta \\
y_{high}[i] + \delta, & \text{if } |y_{high}[i]| > \delta \text{ and } y_{high}[i] < 0 \\
y_{high}[i] - \delta, & \text{if } |y_{high}[i]| > \delta \text{ and } y_{high}[i] > 0
\end{cases}
$$

$\delta$ is an automatic perturbation parameter chosen as the $(k+1)$ maximum absolute values of the detail coefficients
Wavelet Reconstruction

- We can use inverse discrete wavelet transformation (iDWT) to reconstruct the data matrix corresponding to the original data matrix, which will be our perturbed matrix.

\[
\tilde{A} = \text{iDWT}(y_{\text{low}}, \tilde{y}_{\text{high}})
\]
Wavelet Reconstruction

- We use inverse discrete wavelet transformation (iDWT) to reconstruct a data matrix corresponding to the original data matrix

\[ \tilde{A} = \text{iDWT}(y_{\text{low}}, \tilde{y}_{\text{high}}) \]

- For any party in a collaboration with data \( A_i \), he/she can choose any wavelet basis (different high and low frequency filters) and different perturbation parameter \( \delta \) (flexibility)
Data Utilities

Theorem 1: For any selected automatic perturbation parameter $\delta$,

$$
\|A_i^j - A_i^j\| \leq \alpha \cdot \delta \cdot k^{1/2}
$$

Here, $A_i^j$ is the $j$-th perturbed data of the $i$-th party, $\alpha$ is a sufficient small constant.

*W. Wang et al. "Distributed Sparse Random Projects for Refinable Approximation". IPSN'07*
Theorem 1: For any selected automatic perturbation parameter $\delta$, 

$$\| \widetilde{A}_i^j - A_i^j \| \leq \alpha \delta \sqrt{k}$$

Here, $\widetilde{A}_i^j$ is the $j$-th perturbed data of the $i$-th party, $\alpha$ is a sufficient small constant.

**Meaning**: $\widetilde{A}_i^j$ and $A_i^j$ are the perturbed and original data of the $i$-th party. Regardless of choice of wavelet basis, the perturbed data and the original data will be close to each other within a certain interval.
Corollary 2: For any selected automatic perturbation parameter $\delta$,

$$\|A_t - A_p\| \leq \|A_t^i - A_p^i\| + \beta \cdot \delta \cdot k^{1/2}$$

Here, $t \neq p$, $i \neq j$, and $\beta$ is a sufficiently small constant.
Data Utilities

- Corollary 2: For any selected automatic perturbation parameter \( \delta \),

\[
\| \widetilde{A}^i_t - A^i_p \| \leq \| A^i_t - A^i_p \| + \beta \times \delta \times k^{1/2}
\]

Here, \( t \neq p \), \( i \neq j \), and \( \beta \) is a sufficiently small constant

**Meaning:** \( \widetilde{A}^i_t \) and \( A^i_i \) are the \( t \)-th perturbed and original data of the \( i \)-th party, they belong to the \( i \)-th party

Regardless of wavelet basis, before and after the perturbation, the distance of data from different parties with not differ by very much
Outline

- Privacy-Preserving in Collaborative Data Mining
- Advantages of Wavelet Decomposition
- Introduction to Wavelet Decomposition
- Maintaining Statistical Properties
- Experimental Analysis
- Summary
Why Need to Maintain Statistical Properties?

- One set of perturbed data can be used for data mining and statistical analysis.
- For example, basic statistical quantities (average, standard deviation, covariance, etc.).
Notations

• For an $n \times m$ original data matrix $A$, $\tilde{A}$ is the corresponding reconstructed matrix. $A_{i,j}$ and $\tilde{A}_{i,j}$ are the $i$-th row and $j$-th column of the matrices $A$ and $\tilde{A}$, we define

\[
\mu_j = \frac{1}{n} \sum_{i=1}^{n} A_{ij} \\
\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} \tilde{A}_{ij} \\
\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} (A_{ij} - \mu_j)^2}{n-1}} \\
\tilde{\sigma}_j = \sqrt{\frac{\sum_{i=1}^{n} (\tilde{A}_{ij} - \tilde{\mu}_j)^2}{n-1}}
\]
Data Normalization

Using the original data matrix with wavelet, we can have the reconstructed perturbed matrix $\tilde{A}$. We can generate a new matrix $\bar{A}$ that maintains the statistical properties:

$$\bar{A}_{ij} = \left( \tilde{A}_{ij} + \frac{\tilde{\sigma}_j}{\sigma_j} \mu_j - \tilde{\mu}_j \right) \frac{\sigma_j}{\tilde{\sigma}_j} \quad i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, m$$
Maintaining Statistical Properties

- **Theorem 3:**

  1. $\mu_{AJ} = \mu_{\bar{A}J}$
  2. $\sigma_{AJ} = \sigma_{\bar{A}J}$

**Meaning:** After wavelet perturbation and data normalization, every attribute in $\bar{A}$ (the column of the matrix) maintains the basic statistical properties (average and standard deviation).
Maintaining Statistical Properties

**Theorem 3:**

1. $\mu_{A_j} = \mu_{\overline{A}_j}$
2. $\sigma_{A_j} = \sigma_{\overline{A}_j}$

- We define a cost function

$$f(j) = \sum_{i=1}^{n} (A_{i,j} - \overline{A}_{i,j})^2, \quad i = 1, \ldots, n; j = 1, \ldots, m.$$  \hspace{1cm} (2)

Function (2) the cost of data perturbation and data normalization.
Minimize Cost Function

- Theorem 4: Under the conditions of Theorem 3, the normalization formula (1) minimizes the cost function (2).

\[
A_{ij} = \left( \bar{A}_{ij} + \frac{\bar{A}_{ij} - \mu_j}{\sigma_j} \right) \cdot \frac{\sigma_j}{\sigma_j} \quad i = 1, \ldots, n; j = 1, \ldots, m. \quad (1)
\]

\[
f(x) = \sum_{i=1}^{n} (A_{ij} - \bar{A}_{ij})^2, \quad i = 1, \ldots, n; j = 1, \ldots, m. \quad (2)
\]

Theorem 3:
1. \( \mu_{\bar{A}_j} = \mu_j \)
2. \( \sigma_{\bar{A}_j} = \sigma_j \)

Proof: Omitted, use Laplace multiplier
Outline

- Privacy-Preserving in Collaborative Data Mining
- Advantages of Wavelet Decomposition
- Introduction Wavelet Decomposition
- Maintaining Statistical Properties
- Experimental Analysis
- Summary
Experimental Data

- We used two real life datasets
- WBC (Wisconsin breast cancer dataset)
- WDBC (Wisconsin breast cancer diagnosis dataset)

<table>
<thead>
<tr>
<th>Database</th>
<th># Records</th>
<th># Attributes</th>
<th># Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBC</td>
<td>699</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>WDBC</td>
<td>569</td>
<td>30</td>
<td>2</td>
</tr>
</tbody>
</table>

*obtained from the University of California, Irvine (UCI), Machine Learning Repository*
How to Measure Privacy-Preserving

- We defined 5 different quantities to measure of effects of privacy-preserving VD, RP, RK, CP and CK

\[ VD = \frac{\| \bar{A} - A \|_F}{\| A \|_F} \]
How to Measure Privacy-Preserving

We defined 5 different quantities to measure of effects of privacy-preserving VD, RP, RK, CP and CK

\[ RP = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} |Rank^i_j - rank^i_j|}{m \times n} \]

Here, \( Rank^i_j \) and \( rank^i_j \) are the ranks of the data \((l,j)\) of the \(i\)-th attribute in the original and perturbed data.
How to Measure Privacy-Preserving

- We defined 5 different quantities to measure effects of privacy-preserving VD, RP, RK, CP and CK

\[
RK = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} Rk^i_j}{m \times n}
\]

Here, if the ranks of the data \((i,j)\) of attribute \(i\) are the same in the original and perturbed data, then \(Rk^i_j = 1\), otherwise, it is 0
How to Measure Privacy-Preserving

- We defined 5 different quantities to measure effects of privacy-preserving VD, RP, RK, CP and CK

\[ CP = \frac{\sum_{i=1}^{m} |RAV_i - rav_i|}{m} \]

Here, \( RAV_i \) is the order of the average value of the \( i \)-th attribute in the original data, \( rav_i \) is the order of the average value of the \( i \)-th attribute in the perturbed data.
How to Measure Privacy-Preserving

- We defined 5 different quantities to measure of effects of privacy-preserving VD, RP, RK, CP and CK

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

Here, \( C_{k_i} = 1 \), if \( RAV_i = rav_i \), otherwise \( C_{k_i} = 0 \).
How to Measure Privacy-Preserving

- We define 5 quantities to measure the quality of privacy-preserving (VD, RP, RK, CP and CK)

- It is easy to see that the larger the values of VD, RP, and CP, and the smaller of values of RK and CK, the better the privacy-preserving quality
Comparison of 3 Collaborative Data Mining Methods

- Wavelet(S) uses one wavelet basis and perturbation parameter for decomposing, reconstructing, and normalizing data
- Wavelet(V) uses different wavelet basis and perturbation parameters for different parts of the vertically partitioned data
- Wavelet(H) uses different wavelet basis and perturbation parameters for different parts of the horizontally partitioned data
### WBC

<table>
<thead>
<tr>
<th></th>
<th>VD</th>
<th>RP</th>
<th>RK</th>
<th>CP</th>
<th>CK</th>
<th>Time(S)</th>
<th>Correct rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0.2080</td>
<td>239.4</td>
<td>0.00636</td>
<td>1.556</td>
<td>0.4444</td>
<td>0.07882</td>
<td>96.0%</td>
</tr>
<tr>
<td>SVD</td>
<td>0.2557</td>
<td>238.6</td>
<td>0.00477</td>
<td>1.333</td>
<td>0.5556</td>
<td>0.03081</td>
<td>95.9%</td>
</tr>
<tr>
<td>Wavelet (S)</td>
<td>0.3526</td>
<td>247.1</td>
<td>0.00556</td>
<td>1.556</td>
<td>0.333</td>
<td>0.06362</td>
<td>95.6%</td>
</tr>
<tr>
<td>Wavelet (V)</td>
<td>0.3140</td>
<td>239.1</td>
<td>0.00509</td>
<td>2.000</td>
<td>0.333</td>
<td>0.05153</td>
<td>96.1%</td>
</tr>
<tr>
<td>Wavelet (H)</td>
<td>0.3140</td>
<td>239.1</td>
<td>0.00509</td>
<td>2.000</td>
<td>0.333</td>
<td>0.05153</td>
<td>96.1%</td>
</tr>
</tbody>
</table>

We used SVM (support vector machine) to classify the original and perturbed data.
<table>
<thead>
<tr>
<th></th>
<th>VD</th>
<th>RP</th>
<th>RK</th>
<th>CP</th>
<th>CK</th>
<th>Time(S)</th>
<th>Correct rate</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</td>
<td>0.00004</td>
<td>121.3</td>
<td>0.3454</td>
<td>0</td>
<td>1.0000</td>
<td>0.13880</td>
<td>85.4%</td>
</tr>
<tr>
<td>Wavelet (S)</td>
<td>0.00084</td>
<td>165.3</td>
<td>0.1083</td>
<td>4.800</td>
<td>0.4000</td>
<td>0.05166</td>
<td>85.4%</td>
</tr>
<tr>
<td>Wavelet (V)</td>
<td>0.00101</td>
<td>168.6</td>
<td>0.1041</td>
<td>4.733</td>
<td>0.4667</td>
<td>0.09274</td>
<td>85.4%</td>
</tr>
<tr>
<td>Wavelet (H)</td>
<td>0.00096</td>
<td>165.5</td>
<td>0.1141</td>
<td>3.267</td>
<td>0.4667</td>
<td>0.08177</td>
<td>85.4%</td>
</tr>
</tbody>
</table>
Outline

- Privacy-Preserving in Collaborative Data Mining
- Advantages of Wavelet Decomposition
- Introduction to Wavelet Decomposition
- Experimental Analysis
- Summary
Comparison of Several Methods

- Data perturbation based on wavelet decomposition is similar to that based on SVD in terms of perturbation quality.

<table>
<thead>
<tr>
<th></th>
<th>WBC</th>
<th>Correct rate</th>
<th></th>
<th>WDBC</th>
<th>Correct rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>96.0%</td>
<td></td>
<td>Original</td>
<td>85.4%</td>
</tr>
<tr>
<td></td>
<td>SVD</td>
<td>95.9%</td>
<td></td>
<td>SVD</td>
<td>85.4%</td>
</tr>
<tr>
<td></td>
<td>Wavelet (S)</td>
<td>96.0%</td>
<td></td>
<td>Wavelet (S)</td>
<td>85.4%</td>
</tr>
<tr>
<td></td>
<td>Wavelet (V)</td>
<td>95.6%</td>
<td></td>
<td>Wavelet (V)</td>
<td>85.4%</td>
</tr>
<tr>
<td></td>
<td>Wavelet (H)</td>
<td>96.1%</td>
<td></td>
<td>Wavelet (H)</td>
<td>85.4%</td>
</tr>
</tbody>
</table>
Comparing Parameters (Data I)

Regardless of collaborative data mining patterns, the results of privacy preserving seem to be good. Need larger VD, RP, and CP values, and smaller RK and

<table>
<thead>
<tr>
<th></th>
<th>VD</th>
<th>RP</th>
<th>RK</th>
<th>CP</th>
<th>CK</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD</td>
<td>0.2080</td>
<td>239.4</td>
<td>0.00636</td>
<td>1.556</td>
<td>0.4444</td>
</tr>
<tr>
<td>Wavelet(S)</td>
<td>0.2557</td>
<td>238.6</td>
<td>0.00477</td>
<td>1.333</td>
<td>0.5556</td>
</tr>
<tr>
<td>Wavelet(V)</td>
<td>0.3526</td>
<td>247.1</td>
<td>0.00556</td>
<td>1.556</td>
<td>0.333</td>
</tr>
<tr>
<td>Wavelet(H)</td>
<td>0.3140</td>
<td>239.1</td>
<td>0.00509</td>
<td>2.000</td>
<td>0.333</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>VD</th>
<th>RP</th>
<th>RK</th>
<th>CP</th>
<th>CK</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD</td>
<td>0.00004</td>
<td>121.3</td>
<td>0.3454</td>
<td>0</td>
<td>1.0000</td>
</tr>
<tr>
<td>Wavelet(S)</td>
<td>0.00084</td>
<td>165.3</td>
<td>0.1083</td>
<td>4.800</td>
<td>0.4000</td>
</tr>
<tr>
<td>Wavelet(V)</td>
<td>0.00101</td>
<td>168.6</td>
<td>0.1041</td>
<td>4.733</td>
<td>0.4667</td>
</tr>
<tr>
<td>Wavelet(H)</td>
<td>0.00096</td>
<td>165.5</td>
<td>0.1141</td>
<td>3.267</td>
<td>0.4667</td>
</tr>
</tbody>
</table>
Comparing Parameters (Data 2)

- Need larger values for VD, RP and CP, and smaller values for RK and CK

<table>
<thead>
<tr>
<th></th>
<th>VD</th>
<th>RP</th>
<th>RK</th>
<th>CP</th>
<th>CK</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD</td>
<td>0.2080</td>
<td>239.4</td>
<td>0.00636</td>
<td>1.556</td>
<td>0.4444</td>
</tr>
<tr>
<td>Wavelet(S)</td>
<td>0.2557</td>
<td>238.6</td>
<td>0.00477</td>
<td>1.333</td>
<td>0.5556</td>
</tr>
<tr>
<td>Wavelet(V)</td>
<td>0.3526</td>
<td>247.1</td>
<td>0.00556</td>
<td>1.556</td>
<td>0.333</td>
</tr>
<tr>
<td>Wavelet(H)</td>
<td>0.3140</td>
<td>239.1</td>
<td>0.00509</td>
<td>2.000</td>
<td>0.333</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>VD</th>
<th>RP</th>
<th>RK</th>
<th>CP</th>
<th>CK</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD</td>
<td>0.00004</td>
<td>121.3</td>
<td>0.3454</td>
<td>0</td>
<td>1.0000</td>
</tr>
<tr>
<td>Wavelet(S)</td>
<td>0.00084</td>
<td>165.3</td>
<td>0.1083</td>
<td>4.800</td>
<td>0.4000</td>
</tr>
<tr>
<td>Wavelet(V)</td>
<td>0.00101</td>
<td>168.6</td>
<td>0.1041</td>
<td>4.733</td>
<td>0.4667</td>
</tr>
<tr>
<td>Wavelet(H)</td>
<td>0.00096</td>
<td>165.5</td>
<td>0.1141</td>
<td>3.267</td>
<td>0.4667</td>
</tr>
</tbody>
</table>
Comparison of Run Time

- Wavelet seems to be faster

<table>
<thead>
<tr>
<th></th>
<th>WBC</th>
<th>WDBC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time(S)</td>
<td>Time(S)</td>
</tr>
<tr>
<td>Original</td>
<td></td>
<td>Original</td>
</tr>
<tr>
<td>SVD</td>
<td>0.07882</td>
<td>SVD</td>
</tr>
<tr>
<td>Wavelet (S)</td>
<td>0.03081</td>
<td>Wavelet (S)</td>
</tr>
<tr>
<td>Wavelet (V)</td>
<td>0.06362</td>
<td>Wavelet (V)</td>
</tr>
<tr>
<td>Wavelet (H)</td>
<td>0.05153</td>
<td>Wavelet (H)</td>
</tr>
</tbody>
</table>
Questions

Thank You