Distributed Clustering Algorithm for Privacy-Preserving Based on Random Projection

Chunhua Cai, Wangcheng Cao

Mudanjiang Normal University, Mudanjiang 157011, Heilongjiang, China

Abstract
In the paper, to solve the problem of Privacy-Preserving in clustering mining, the author presents the data-disturbance method based on centrosymmetric—RPT(Random Projection-based Transformation), which generates the Projection Matrix first, then conducts symmetric disturbance to the data according to the Matrix. The disturbed data has great difference to the original in dimension and shape of clustering, and has high availability, which aims to improve Privacy-preserving and communication fees as well as the resistance to Independent Component Analysis (ICA), on the premise of data availability. Verified by the experiment, this method is better to achieve the expected goals, and showed more extensive application prospects.

Key words: Random Projection, Privacy Protection, Distributed Clustering Algorithm, Data-disturbance.

1. INTRODUCTION

People have put a lot of money, energy, and the technical force into the research of clustering method, and new Privacy-preserving clustering methods emerge in endlessly, which can further the development of the research for privacy protection clustering method both in theory and practical application. Stanley and others proposed GDTMS method, which uses the geometric-transformation method for data-disturbance. The method has high availability, but privacy protection degree is relatively low (Stanley, Olivei and Zaiane, 2010). İnan proposes a data-perturbation method for distributed data in different degree. The method has a wide scope of application and a small amount of calculation, but the disadvantage is its high communication costs (İnan, Kaya, Saygin and Savas, 2007). Xiaokui put forward a kind of ideal method of random perturbation, which supports multiple-level of privacy-protection, but its security protocol is very complicated (Mohammad Ali Kadampur and Somayajulu, 2008). Mohammad put forward a privacy-protection method through clustering inflation on the basis of the scale-transformation. This method is effective for privacy protection, but due to the application of scaling transformation, it will lower the data availability after a disturbance (Sara Hajilan and Mohammad Abdollahi Azgomi, 2008).

After the research of related literature for the existing privacy-preserving clustering methods, we found that these methods all have some shortcomings in some degree, as a result of using different angles and algorithms. Many problems need to be solved, such as the low degree of privacy protection, low data availability, the high cost of communication and malicious attacks. So the author proposes a data-disturbance method based on centrosymmetric-RPT, which generates the Projection Matrix first, then conducts symmetric disturbance to the data according to the Matrix. The disturbed data has great difference to the original in dimension and shape of clustering, and has high availability. Therefore, the method can protect the privacy information and resist the malicious attacks through ICA effectively.

In this paper, the author makes verification to the proposed algorithm and analysis to the experimental results, which shows the superiority and practicability and application value of RPT algorithm. It also makes discussions about the direction for further improvement in the future.

2. THE RELEVANT THEORIES DATA MINING FOR PRIVACY PROTECTION

2.1. Data-disturbance Methods

Data-disturbance Method is a kind of data-deformation process the data holder makes before releasing the data. The purpose of the data deformation, on the one hand, the data holders want to hide the sensitive information included in the released data base through the particular data changes; on the other, the data holder want to protect the attribute information in utmost degree which play an important role to data-mining results, thus meet the requirements of specific data-availability for data mining tasks.

Most of the data perturbation method is a special case of the disturbance of matrix. If the original data set is $X$, then the disturbed data set $Z$ can be calculated by the following formula:

$$Z = AXB + C$$  (1)
Among which A is record-transformation matrix, B is property-transformation matrix, C is substitution matrix (noise).

2.2. Independent component analysis
ICA (Common, 2010; Arinen and Oja, 2000) is a kind of technique for hidden independent factors, which are often hidden in a series of combination of linear or nonlinear unknown variables, and the hybrid system is also mixed set of linear or nonlinear unknown variables. All these unknown variables do not obey Gaussian distribution rules and are independent of each other, which are called independent components (ICs) for data observation. And the independent component can be detected by the means of ICA. The typical example of the ICA is the cocktail party problem. Assuming that at a cocktail party, although a lot of noisy voice mixed together, such as music, the sound of talking and TV news, even the whistling of a passing ambulance, people can still clearly identify their talking voice with others. Although we don’t know how the brain can distinguish these voices, the ICA can do this, the premise is there are enough receiver and continuous sound source.

With the advancement in networking and multimedia technologies enables the distribution and sharing of multimedia content widely. In the meantime, piracy becomes increasingly rampant as the customers can easily duplicate and redistribute the received multimedia content to a large audience. Insuring the copyrighted multimedia content is appropriately used has become increasingly critical.

3. DATA PERTURBATION ALGORITHM BASED ON RANDOM PROJECTION

3.1. Relevant Definition
In this paper, the data can be seen as a matrix $D_m = \{a_1, a_2, \ldots, a_n\}$, in which each row in line m is an observation value $O_m$, and each observation value contains n the value of the attribute $A_i$. The whole data matrix may contain scope $d < n$, which can be seen as a subspace of Euclidean space, and for any vector quantity $v_i \in V$, we can infer $v_i = (a_1, a_2, \ldots, a_d)$, $1 \leq i \leq d$, and there exist i which makes $a_i$ an example of $A_i$.

Vector space $V$ must be done disturbance transformation to protect the privacy information of data record before the data sets being released for clustering mining. In order to transform vector space $V$ into disturbed vector space $V'$, we should make a projection for all the elements $v_i$ in $V$ through projection matrix to low dimension space. So we define $R$ in the form as the following:

Definition 1 (projection matrix) Assume $R$ as a data matrix composed by k dimension vector space $R = (R_1, R_2, \ldots, R_d)$, and to any element $R_i$, $1 \leq i \leq d$, $1 \leq j \leq k$, we have $R_i \sim N(0, \sigma^2)$.

Given a symmetric center $e$, we can transform vector space $V$ into vector space $V'$ through centrosymmetric transformation.

Definition 2 (projection transformation function) Assume $V$ as a d dimension subspace, for any vector $v_j$ in it, $1 \leq i \leq d$, $v_i = (a_1, a_2, \ldots, a_d)$. And each $a_i$ in $v_i$ is a classified numerical-attribute observation value. Assume $R = (R_1, R_2, \ldots, R_d)$ as a projection matrix, we define a projection function $f$, which is a one-one mapping from d dimension to k dimension and transform $V$ into $V'$ according to $R$. The form of $V'$ is $V' = f(V, R) = \frac{1}{\sqrt{k\sigma}}VR$.

According to above-mentioned, we define the method of centrosymmetric disturbance (RPT) .

Definition 3 (Random Projection Disturbance Method) A d-dimension random projection disturbance method is a ordered pair, we define it as $RPT(V, f)$, and in this expression:

1. $V \subseteq R^d$ is a vector space representative of a data point needed to be transformed.
2. $f$ is a projection transformation function, $f : R^d \rightarrow R^d$.

As for the RPT, its input value is vector $V$, which is composed of classified numerical type attribute and projection matrix $R$, and its output value is the transformed vector $V'$.

3.2. Method and Procedure
When you have two sites for joint clustering mining, each has a set of data $A, B$. Both send their data set to the third party for joint clustering mining. Assume $D_j$ as cascade matrix of all data sets, and the distribution type of data is horizontal portioning, thus:
The third party sends the mining results to all the sites involved in calculation, using clustering mining method to\(D_c\), thus they can share the mining results and obtain mining knowledge favorable to themselves. All the sites including the third party are semi-trustable, which means that they will strictly abide by the agreement, but they will collect any data found so as to find privacy information in the future. And they believe all sites involved will not collude with other.

In order to protect privacy information included in all the data sets and to make the disturbed data sets have higher privacy-protection degree and data availability, we make data disturbance to data sets by RPT method before sending data sets to the third party. The after-disturbance data set is \(A', B'\). Assume \(D_c'\) as the cascade matrix of all the disturbed data sets, we have:

\[
D_c' = (A', B')
\]

Figure 1 describes the relationship of the k sites for joint clustering with the third party:

\[\text{semi-trustable communication} \quad \text{RPT disturbance} \quad \text{site 1} \quad \text{clustering analysis} \]

\[\text{site 2} \quad \text{RPT disturbance} \quad \text{the third party}\]

\[\text{analysis results}\]

**Figure 1.** The relationship of the k sites for joint clustering with the third party

The RPT method presented in this paper mainly contains three steps:

**Step1:** All the sites involved in calculation jointly select random seeds, elements variance in random matrix, and the number of dimensions after disturbance;

**Step2:** Each site generates a projection matrix respectively, according to the parameters in Step1;

**Step3:** Each site makes data disturbance to its own data set according to projection.

Next we will introduce these three steps respectively in detail.

### 3.2.1 The Variances of Elements in Random Seeds and Random Matrix

To reduce the communication traffic between various sites, RPT do not generate the elements in projection matrix one by one, which will produce huge traffic; instead, all the sites together generate only one random seed, through which they can generate the same random matrix and thus reduce communication traffic. We will introduce the matrix-generation method in 3.2.2.

In the RPT method, any element \(R_{ij}\) in the selected projection matrix obeys the normal distribution with mean value as 0 and variance as \(\sigma^2\):

Given any element \(R_{ij}\) in \(R, 1\leq i \leq d, 1\leq j \leq k\), we have \(R_{ij} \sim N(0, \sigma^2)\)

So before the generation of projection matrix, all the sites need to communicate with each other briefly to select random seed \(r\) and variance \(\sigma\) of projection matrix, and the dimension \(k\) of random matrix. The agreement contents of the communication process are as follows:

Assume there are 2 sites involved in calculation, and all the sites sort according to a kind of rule (and the order does not affect the algorithm results), each site selects \(r, \sigma, k\), randomly, of which \(1 < J < 2\).

Each site send its own numerical value to the next site through information, its format is \(\text{Msg} \{\text{Type}, r, \sigma, k\}\), and \(\text{Type}\) is the current numerical type, its data range is \{\text{Part}, \text{Fin}\}, among which \(\text{Part}\) shows that the current numerical value is not the final value. \(\text{Fin}\) means that the current numerical value is the final value.

### 3.2.2 The Generation of Projection Matrix

In the RPT method, the element \(R_{ij} \sim N(0, \sigma^2)\) in the projection matrix generated by each site, generates normal-distributed random number \(y\) with mean value \(\mu\) and variance \(\sigma^2\), its calculation formula is as follows:
\[ y = \mu + \sigma \frac{\sum_{i=0}^{n-1} \text{rnd}_i - \frac{n}{2}}{\sqrt{n/12}} \]  \hspace{1cm} (2)

Among which \( n \) is large enough, normally, as \( n = 12 \), its degree of approximation is very good. Then we have

\[ y = \mu + \sigma \left( \sum_{i=0}^{n-1} \text{rnd}_i - 6 \right) \]  \hspace{1cm} (3)

Among which \( \text{rnd}_i \) is even-distributed random number from 0 to 1. According to definition 1:

\[ R_{y} = \sigma \left( \sum_{i=0}^{n-1} \text{rnd}_i - 6 \right) , \ 1 < j < d \]  \hspace{1cm} (4)

Among which \( \text{rnd}_i \) is even-distributed random number from 0 to 1.

We can obtain even-distributed random number from 0 to 1 by the following method:

\[ r_i \approx \text{mod}(2053 \times r_{i-1} + 13849, m) \ i = 1, 2, \ldots ; \ p_i = r_i / m \]

In this formula, \( m = 2^{36} \), \( p_i \) is the number \( i \) random number. \( r_i \) is the random-generated seed \( r \) in 3.2.1, and the same random seed generates the same random number.

Communication Process is shown in Figure 2.

![Communication Process](image)

**Figure 2.** The flow diagram of dimension after disturbance and the selection of random seeds and element variance in random matrix

### 3.2.3 Random Projection Transformation

Assume \( X, Y \) are the data sets owned by the two organization \( P_1, P_2 \) involved in distributed computing respectively. Among which \( X \) is matrix \( m_1 \times n \), \( Y \) is matrix \( m_2 \times n \). \( R \) is the random matrix \( n \times k \ (k < n) \) generated by \( P_1, P_2 \) together. For each element \( r_{ij} \) in \( R \) is independently identically distributed, and \( r_{ij} \sim N(0, \sigma^2) \). According to definition 2, we have:

\[ U = \frac{1}{\sqrt{k} \sigma} X R, \ V = \frac{1}{\sqrt{k} \sigma} Y R \]

Figure 3 and 4 show the changes of data matrix after projection:
Figure 3. Coordinate Relations between the Two Points symmetric to Point P.

Figure 4. Coordinate Relations between the Two Points symmetric to Point P.

Not hard to see, the disturbed data sets have not only changed greatly in shape, but also, its number of attribute has been reduced accordingly.

3.3 Algorithm Description

----------

Algorithm name: centrosymmetric data disturbance algorithm
Input: disturbance-needed attribute set A
Output: transformed attribute set $A'$
begin //data disturbance
begin park1 // Generate the projection matrix $k$, $r$ and $\sigma$, and common used for each site
Site I: Generate $k$, $\sigma$ and random vector $r$;
Send $\langle Part, r, \sigma, k \rangle$ to site 2;
Site II: Receive $\langle Type, r, \sigma, k \rangle$;
If $Type = Part$ then
Generate vector $r_1, \sigma_1$ and $k_1$;
Send $\langle Fin, (r_1 + r_2)/2, (\sigma + \sigma_2)/2, (k_1 + k_2)/2 \rangle$ to site 1;
Else
Type = Fin;
End if
End park1 // Output $\sigma$, rand $k$
Begin park2 // Generate own projection matrix respectively
Park2 Input $r$, $\sigma$ and $k$;
Site I, II: Generate random number $rnd_i, r_1 = r$
$r_i = \text{mod}(2053 \times r_{i-1} + 13849, m)$ $i = 1, 2, \ldots$
$\text{rnd}_i = r_i/m$;
Generate element $r_{ij}$ of R in random matrix according to $\text{rnd}_i$
$r_{ij} = \sigma(\sum_{i=0}^{\text{rnd}_j}(-6), 1 < i < k, 1 < j < n$
End park2 // Output R
Begin park3 // Each site makes disturbance according to projection matrix
Part 2 Input data set A owned by the site, the generated projection matrix R;
3.4 Algorithm Analysis

Data disturbance method presented in this paper is designed for privacy protection of distributed clustering mining, so whether it can ensure the accuracy of clustering mining is very important. Assume disturbed data adopt K-MEANS clustering algorithm. We can estimate clustering results according to square error criterion, its definition is as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$  \hspace{1cm} (5)

Here, $E$ is the sum total of square error of all the objects in data base, $m_i$ is the center of clustering $C_i$, $p$ is data object in data sets, namely vector in space.

To the disturbed data set $U$, $V$, the inner product operation results are:

$$TT E RR k I$$

Assume $r_{ij}, e_{ij}$ are the number $i, j$ elements of $R, RR^T$ respectively, and:

$$E[e_{ij}] = \sum_{i=1}^{k} r_{ij}, E[e_{ij}] = E[\sum_{i=1}^{k} r_{ij}] = \sum_{i=1}^{k} E[r_{ij}]$$

For the elements in $R$ are independent and its distribution is known:

$$E[e_{ij}] = \left\{ \begin{array}{ll}
0 & \text{if } i \neq j; \\
\frac{1}{k\sigma^2} & \text{if } i = j;
\end{array} \right.$$  \hspace{1cm} (6)

Also $E[r_{ij}] = 0$ and $E[r_{ij}^2] = \sigma^2$, so:

$$E[e_{ij}] = \left\{ \begin{array}{ll}
0 & \text{if } i \neq j; \\
\frac{1}{k\sigma^2} & \text{if } i = j;
\end{array} \right.$$

Bring the results into the formula:

$$E[UV^T] = XX^T = E[UV^T - XY^T] = 0$$

And for any $i$, there is $E[e_{ii}] = k\sigma^2$, $Var[e_{ii}] = 2k\sigma^4$, for any $i, j\neq i$, there is $E[e_{ij}] = 0$, $Var[e_{ij}] = k\sigma^4$. It’s easy to prove the following formula:

$$Var[uv^T - xy^T] = \frac{1}{k} \left( \sum_i x_i^2 \sum_j y_j^2 + (\sum_i x_i y_i)^2 \right)$$

In practical use, as $x$, $y$ have been both standardized, $\sum_i x_i^2 \sum_j y_j^2 = 1$ and $(\sum_i x_i y_i)^2 \leq 1$, we have:

$$Var[uv^T - xy^T] \leq \frac{2}{k}$$

Next let’s calculate the variation of Euclidean distance after disturbance $\|x - y\|$ , namely $\|x - y\| = (x - y)(x - y)^T$, based on the above-mentioned formula:

$$E[\|u - v\|^2 - \|x - y\|^2] = 0 Var[\|u - v\|^2 - \|x - y\|^2] \leq \frac{8}{k}$$

According to definition 2 in document (Arriaga and Vempala, 2008):

$$Pr \left\{ (1 - \varepsilon)\|x - y\|^2 \leq \|u - v\|^2 \leq (1 + \varepsilon)\|x - y\|^2 \right\} \geq 1 - 2e^{-\varepsilon^2 \|x - y\|^4 / k}$$

Here, $0 < \varepsilon < 1$, $Pr(A)$ shows the probability of event $A$. 

$$\text{Site I, II: } \quad A' = f(A, R) = \frac{1}{\sqrt{k\sigma}} AR$$

End park3

End
Through the derivation above, obviously, as the number of dimensions \( k \) increases, the effect of projection disturbance becomes better; and at the same time, after disturbance, the distance change between vectors becomes smaller, and data’s availability higher.

As to the security of algorithm, under certain circumstances, ICA constitutes a threat to RPT method, which is based on multiplicative model. But as long as the number of dimensions of projection matrix meets certain qualifications \( (m > 2k - 1) \) is nonevent number, etc.), there still not a new method that can restore the original data even if the data is completely in conformity with decomposition model.

### 3.5 Algorithm Evaluation

Usually, the complex rates of algorithm includes the complex rates of time and space, they have measured the execution time needed for algorithm and the processing resources for data processing. So to speak, it’s a criterion directly relevant to computational efficiency. It is an important goal for algorithm design to make the complex rate as low as possible.

In the data transformation method based on random projection, the total time complex rate of algorithm is \( O(\text{mmk})(k \ll n) \). Compared to all the privacy protection methods using matrix multiplication disturbance, the method’s time complex rate is not high.

### 4. SIMULATION EXPERIMENT

#### 4.1. Evaluation Criterion for Experimental Results

The experiment adopts misclassification rate and relative error as indexes to estimate the validity and effectiveness of data disturbance method.

Misclassification rate is used to show the percentage of total of the number of misclassified clustering objects, which is expressed as \( M_k \):

\[
M_k = \frac{1}{N} \sum_{i=1}^{k} | \text{Cluster}_i(D) | - | \text{Cluster}_i(D') |
\]  

(6)

In the formula, \( N \) is the number of data objects in the original data set, \( k \) is the number of clustering. \( |\text{Cluster}_i(D)| \) signifies the number of points of number \( i \) clustering in the original data set, and \( |\text{Cluster}_i(D')| \) signifies the number of points of number \( i \) clustering in the disturbed data set.

In addition, the specific value of the number of dimensions for disturbed data set to the number of dimensions before disturbance is called projection rate. The lower the projection rate is, the higher the disturbed data reduces, the higher for the degree of privacy protection, and the corresponding data availability will lower, and vice versa.

And relative error is a common index in scientific experiments, it shows the degree of error for measured value, and is signified as \( \delta \):

\[
\delta = \Delta / L \times 100\
\]

(7)

Here, \( \Delta \) is absolute error, \( L \) is for truth-value.

#### 4.2. Experiment Design and Data Measuring

In this paper, we design two different experiments aiming at two possible applications to test the presented method, verifying the superiority of the method through the experimental data.

1. **The calculation of inner product for distributed data and estimation of Euclidean distance.**

   Based on the above introduction, both inner product and Euclidean distance are very important parameter indexes for clustering analysis, so the relative error between inner product and Euclidean distance calculated with disturbed data and the original data can affect the accuracy of clustering analysis directly. The RPT method enables the third party be able to calculate the inner product and Euclidean distance of data sets using disturbed data. The experiment procedure is as follows:

   1. The organization \( A \) and \( B \) involved in calculation generate random seeds together through which they generate a random matrix \( R \) of \( k \times m \).
   2. \( A \) and \( B \) project data sets onto \( k \) dimension space \( R^k \), and send the disturbed data sets \( U = \frac{1}{\sqrt{k} \sigma} RX \) and \( V = \frac{1}{\sqrt{k} \sigma} RY \) to the third party.
   3. The third party obtain inner matrix using \( U \), \( V \)
Similarly, the third party can calculate Euclidean distance at the disturbed data sets.

The experimental data in this part adopts the adult data base in UCI Machine Learning Repository, which is derived from the census data in 1994. To ensure the generality, we choose the former 10000 rows data and use only two attributes (fnlwgt, education-num), and make experiments 20 times repeatedly.

(2) K-means clustering of the disturbed data
The data in this part of experiment comes from data sets of Synthetic Control Chart Time Series of UCI KDD. The data set has 600 examples of control chart, each example has 60 attributes. These control charts can be divided into 6 types: normal, periodic, increasing trend, downward trend, upward movement, downward movement. The elements in the projection matrix all abide by the normal distribution with mean value 0 and variance 4. The clustering algorithm is the classic K-means algorithm. First, we make disturbance to the data sets using RPT method, then using K-means algorithm to disturbed cascading matrix, analyze the accuracy and degree of privacy protection of data clustering after disturbance. The experimental procedure is as follows:

1. The organization A and B involved in calculation generate random seeds together, and through which they generate a random matrix $R$ of $n \times k$.

2. A and B project data sets onto $k$ dimension space $R^k$, and send the disturbed data sets $U = \frac{1}{\sqrt{k\sigma}} XR$ and $V = \frac{1}{\sqrt{k\sigma}} YR$ to the third party.

3. The third party uses k-means clustering algorithm at data set $\begin{pmatrix} U \\ V \end{pmatrix}$.

4.3. Experiment and Result Analysis
Table 1 shows the relative error of inner product calculation and Table 2 shows the relative error Euclidean distance calculation. The results are acquired by doing the first experiment 20 times repeatedly.

**Table 1. The relative error of inner product calculation**

<table>
<thead>
<tr>
<th>k</th>
<th>Mean value (%)</th>
<th>Variance (%)</th>
<th>minimum (%)</th>
<th>maximum (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100(1%)</td>
<td>9.91</td>
<td>0.41</td>
<td>0.07</td>
<td>23.47</td>
</tr>
<tr>
<td>500(5%)</td>
<td>5.84</td>
<td>0.25</td>
<td>0.12</td>
<td>18.41</td>
</tr>
<tr>
<td>1000(10%)</td>
<td>2.94</td>
<td>0.05</td>
<td>0.03</td>
<td>7.53</td>
</tr>
<tr>
<td>2000(20%)</td>
<td>2.69</td>
<td>0.04</td>
<td>0.01</td>
<td>7.00</td>
</tr>
<tr>
<td>3000(30%)</td>
<td>1.81</td>
<td>0.03</td>
<td>0.27</td>
<td>6.32</td>
</tr>
</tbody>
</table>

**Table 2. The relative error Euclidean distance calculation**

<table>
<thead>
<tr>
<th>k</th>
<th>Mean value (%)</th>
<th>Variance (%)</th>
<th>minimum (%)</th>
<th>maximum (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100(1%)</td>
<td>10.44</td>
<td>0.67</td>
<td>1.51</td>
<td>32.58</td>
</tr>
<tr>
<td>500(5%)</td>
<td>4.97</td>
<td>0.29</td>
<td>0.23</td>
<td>18.32</td>
</tr>
<tr>
<td>1000(10%)</td>
<td>2.70</td>
<td>0.05</td>
<td>0.11</td>
<td>7.21</td>
</tr>
<tr>
<td>2000(20%)</td>
<td>2.59</td>
<td>0.03</td>
<td>0.31</td>
<td>6.90</td>
</tr>
<tr>
<td>3000(30%)</td>
<td>1.80</td>
<td>0.01</td>
<td>0.61</td>
<td>3.91</td>
</tr>
</tbody>
</table>

In the table, k signifies the number of dimensions of disturbed matrix, as well as the percentage of the number of dimensions of disturbed data to the number of dimensions of original data.

Figure 5 shows how the original data be disturbed.
Figure 5. The random projection ratio of the original data and disturbed data

Combined with Figure 5, from the experimental results, we can see that the disturbance degree is moderate when the projection ratio is 30%, and the mean value of relative error for inner product and Euclidean distance calculation is around 1.8%. It can be assumed the projection ratio about 30% is suitable for practical use.

Figure 6. The 60 attributes distribution diagram of original data

Figure 6 and Figure 7 demonstrate the distribution diagrams of clustering center for the experiment data before and after disturbance respectively, the data is aiming at the second experiment.

When using projection ratio of 50%, 33% and 50% respectively for the data sets disturbance and clustering of the disturbed data sets, we can see in Table 3: when projection ratio is 50%, the 60 attributes of the original data sets are reduced to 30, its clustering result is perfect—the misclassification rate is nearly 0.17%, namely only 2 in 600 data objects have been classified into the wrong clustering. When projection ratio is 33%, the 60 attributes of the original data sets are reduced to 20, the misclassification rate of clustering is nearly 2.5%, and namely 15 data objects have been classified into the wrong clustering. And when the attributes are reduced to10, the projection ratio is 17%, the misclassification rate is still relatively low, only 4.33%.
Figure 7. The 10 attributes distribution diagram of disturbed data

This shows that the RPT method can protect privacy information and compress data well, meantime the availability of disturbed data is also high.

Table 3. The clustering results after disturbance with different projection ratio

<table>
<thead>
<tr>
<th>attributes</th>
<th>Clustering 1</th>
<th>Clustering 2</th>
<th>Clustering 3</th>
<th>Clustering 4</th>
<th>Clustering 5</th>
<th>Clustering 6</th>
<th>misclassification rate $M_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 (original)</td>
<td>187</td>
<td>25</td>
<td>41</td>
<td>34</td>
<td>117</td>
<td>196</td>
<td>0.00%</td>
</tr>
<tr>
<td>30 (50%)</td>
<td>188</td>
<td>25</td>
<td>40</td>
<td>34</td>
<td>117</td>
<td>196</td>
<td>0.17%</td>
</tr>
<tr>
<td>20 (33%)</td>
<td>182</td>
<td>29</td>
<td>36</td>
<td>32</td>
<td>128</td>
<td>193</td>
<td>2.50%</td>
</tr>
<tr>
<td>10 (17%)</td>
<td>182</td>
<td>19</td>
<td>65</td>
<td>36</td>
<td>108</td>
<td>190</td>
<td>4.33%</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

With the data-disturbance algorithm based on random projection, we can protect the privacy information and ensure the security level in the process of distributed data mining; additionally, the availability of disturbed data is maintained at a high level while compressing data. Whereas the range of application is limited slightly—the method can’t process non-numerical type data yet. It will be a main concern for our further research on how to provide the compatibility and support for various data types.

ACKNOWLEDGEMENTS

This work was supported by research project of Mudanjiang normal university with grant no.YB201601.

REFERENCES

