doi:10.3772/j. issn. 1006-6748. 2017. 02. 009

# Similarity measurement method of high-dimensional data based on normalized net lattice subspace①

Li Wenfa (李文法)②*, Wang Gongming**, Li Ke*, Huang Su*
(*Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, P. R. China)
(**National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, P. R. China)

**Abstract**

The performance of conventional similarity measurement methods is affected seriously by the curse of dimensionality of high-dimensional data. The reason is that data difference between sparse and noisy dimensionalities occupies a large proportion of the similarity, leading to the dissimilarities between any results. A similarity measurement method of high-dimensional data based on normalized net lattice subspace is proposed. The data range of each dimension is divided into several intervals, and the components in different dimensions are mapped onto the corresponding interval. Only the component in the same or adjacent interval is used to calculate the similarity. To validate this method, three data types are used, and seven common similarity measurement methods are compared. The experimental result indicates that the relative difference of the method is increasing with the dimensionality and is approximately two or three orders of magnitude higher than the conventional method. In addition, the similarity range of this method in different dimensions is [0, 1], which is fit for similarity analysis after dimensionality reduction.

**Key words**: high-dimensional data, the curse of dimensionality, similarity, normalization, subspace, NPsim

## 0 Introduction

A similarity measurement can determine similarity degree between two data, or distance between two points, which is the basis of data-mining methods such as clustering, classification, nearest neighbor search, and association analysis. Conventional similarity measurement methods include Euclidean distance, Jaccard coefficient[1], and Pearson coefficient[2]. These methods can satisfy the similarity measurement requirement in low-dimensional space (less than 16)[3]. However, with the increasing spatial dimensionalities, the distance between a query point and its nearest neighbor point tends to be equal to the distance from the query point to its farthest neighbor point. When the distance between any two points is equal everywhere, the similarity is pointless; this is called the isometrics in high-dimensional space[4]. The root cause of this phenomenon is the curse of dimensionality that is derived from properties of sparsity and empty space in a high-dimensional space. Thus, the performances of many similarity measurements are positively affected in the low-dimensional space, yet decrease sharply in the high-dimensional space.

In recent years, a series of methods have been proposed for similarity measurement of high-dimensional data; these include $Hsim(X,Y)$[5], $HDsim(X,Y)$[6], $Gsim(X,Y)$[7], $Close(X,Y)$[8], and $Esim(X,Y)$[9]. However, these methods ignore the relative difference in property, noise distribution, weight, and are only valid for certain data types[10]. The $Psim(X,Y)$ function considers the above-mentioned factors[10] and is applicable to a variety of data types; however, it is unable to compare similarity under different dimensions because its range depends on the spatial dimensionality.

To solve this problem, a similarity measurement method of high-dimensional data based on normalized net lattice subspace is proposed. The similarity range is no longer limited by the spatial dimensionality.

① Supported by the National Natural Science Foundation of China (No. 61502475) and the Importation and Development of High-Caliber Talents Project of the Beijing Municipal Institutions (No. CIT & TCD201504039).

② To whom correspondence should be addressed. E-mail: liwenfa@buu. edu. cn

Received on Dec. 10, 2016

# 1 Related work

## 1.1 Curse of dimensionality

This is a ubiquitous phenomenon in the application field of high-dimensional data, and occurs because of the sparsity and empty space in high-dimensional space.

### 1.1.1 Sparsity

There is a $d$-dimensional data set $D$ in a hypercube unit $\Psi = [0,1]^d$, and data elements are distributed uniformly. The probability of a point falling into one hypercube with length $s$ is $s^d$, which decreases with the increase of $s$ because $s < 1$. That is, it is very likely that there is no point in a large range[11]. For example, approximately only 0.59% data exists in a hypercube with length 0.95 when dimension $s = 100$.

### 1.1.2 Empty space phenomenon

A normal distribution dataset can be expressed by its center point and standard deviation. The distances between the data points and the center point obey the Gauss distribution; however, their relative orientation can be selected randomly. In addition, the number of possible directions relative to a center point is increased exponentially and the distance between them is increased with the increase of dimensionality. From the viewpoint of the density of a dataset, a maximum value exists at the center point, although there may not be a point close to the center point. This phenomenon of a high-dimensional space is called “empty space.”

### 1.1.3 Isometry

The volume of unit sphere in a $d$-dimensional space is described as follows.

$$V(d) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\Gamma\left(\frac{d}{2}\right)} \tag{1}$$

$V(d)$ decreases gradually with the increase of dimensionality $d$. Fig. 1 shows that $V(d) \to 0$ if $d > 16$.

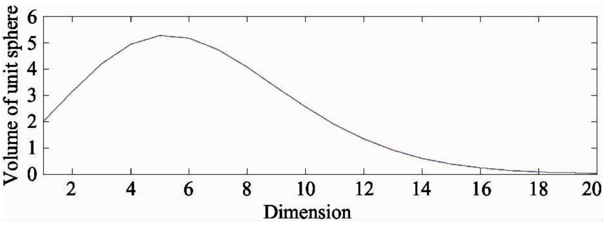


**Fig. 1** Variation trend of unit sphere volume with increasing dimensions

With the increase in dimensionality, the number of corners increases and the volume of unit sphere gradually decreases because the volume of the unit hyperspace does not change. Thus, most of the data will be distributed in the hyperspace corner. This phenomenon is shown in Fig. 2 from left to right; the three subgraphs show the distributions of super-space data with dimensionality of 2, 3, and 8, respectively. In eight-dimensional space, 98% data is distributed in 2^8 = 256 corners. Moreover, the maximum and minimum Euclidean distances between the data and center point are both the same. When the dimensionality tends to infinity, the difference between the maximum and minimum Euclidean distance of the sample points to the center point tends toward 0.

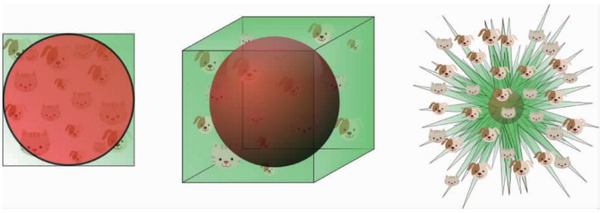


**Fig. 2** Data distribution in different dimensions

Therefore, with the increase in dimensionality, the Euclidean distance between any data tends to remain the same, and no longer has the measurement function. The corresponding data-mining methods, such as clustering, classification, and nearest neighbor, would lose their effect.

## 1.2 Conventional high-dimensional data similarity measurement methods

In recent years, a similarity measurement problem in high-dimensional space has been studied to a certain extent but the research is insufficient. The $Hsim(X,Y)$ function was proposed by Yang[5], which is better than the conventional method but neglects the relative difference and noise distribution. In addition, it is not suitable for measuring the similarity of categorical-attribute data. Next, the $Gsim(X,Y)$ function[7] was proposed according to the relative difference of properties in different dimensions; however, it ignores the weight discrepancy. Zhao introduced the piecewise function $\delta(X,Y)$ into $Hsim(X,Y)$ and proposed the $Hsimc(X,Y)$ function[12], which comprises a function of measuring categorical-attribute data. However, similarity between a pair of points whose components are complementary in every dimension is inconsistent with the actual result. The piecewise function $\delta(X,Y)$ of function Xie modified $Hsimc(X,Y)$ and proposed the $HDsim(X,Y)$ function[6], which can solve the problem derived from a complementary property in every dimension. However, the attribute difference and noise distribution problem are neglected. The $Close(X,Y)$ function[8] based on the monotonous decrease of $e^{-x}$ can overcome the influence from components in some dimensions with large variance but does not consider the relative difference, which would be affected by noise. The $Esim(X,Y)$[9] function was proposed by improving $Hsim(X,Y)$ and $Close(X,Y)$ functions and combining the influence of property on similarity. In every dimension, the $Esim(X,Y)$ component shows a positive correlation to the value in this dimension. All dimensions are divided into two parts: normal and noisy dimensions. In a noisy dimension, the noise occupies majority. When noise is similar or larger than the one in a normal dimension, this method is invalid. The secondary measurement method[13] is used to calculate the similarity by virtue of property distribution, space distance, etc.; however, it neglects the noise distribution and weight. In addition, it is time-consuming. The concept of nearest neighbor projection was proposed by Hinneburg[14], which was combined with dimensionality reduction to solve the problem in high-dimensional space. However, this method complicates the determination of a suitable quality criterion function. Thus, an extension theory was introduced into similarity calculation[15], in which, the high-dimensional data is expressed as an ordered three tuple by virtue of matter element, and the deviation (the interval length of attribute value in every dimension) is added into function $A$. However, this method is too complicated, and the result validation of the high-dimensional data was not described in the corresponding paper. Yi[10] determined that in a high-dimensional space, the difference in a noisy dimension is larger than in a sparse dimension, no matter how similar the data is. This difference occupies a large amount of the similarity calculation, leading to the calculation results of any objects being similar. Therefore, the $Psim(X,Y)$ function[12] was proposed to eliminate the noisy influence by analyzing the difference among all dimensions. The experimental results indicate that this method is suitable for a variety of data. However, its range is $[0,n]$, where $n$ is the dimensionality. Thus, the similarities in different dimensions cannot be compared.

## 2 Similarity measurement method based on normalized net lattice subspace

### 2.1 Sparse and noisy dimensions

With increasing dimensionality, the similarities based on the $L_d$ norm between any data become the same. The root cause is that the $L_d$ norm depends on the dimension too much which has largely different components. In other words, when calculating similarity between $X$ and $Y$, the larger the value of $X_i - Y_i$ on the $i$-th dimension, the greater the contribution of the $i$-th dimension to $X$ and $Y$. Although both $X$ and $Y$ are very similar in other dimensions except the $i$-th dimension, the overall similarity of $X$ and $Y$ is very small. This $i$-th dimension is called sparse or noisy dimension.

Owing to the existence of sparsity and noise in the high-dimensional space, no matter how similar the two records are there will always be a different dimension. The difference in these dimensions occupies a large proportion of the whole similarity, leading to any record in the high-dimensional space being dissimilar[16].

To solve this problem, the data range in every dimension can be divided into several intervals, and the components can be mapped onto corresponding intervals. When calculating the similarity between two points, only the dimensions that fall into the same interval are used. The other dimensions are regarded as sparse or noisy dimensions, and are not included in the calculation.

### 2.2 Meshing of high-dimensional data space

Let the dimension of dataset be $d$, and the number of data object be $M$. Then, every data object is expressed as $x_k (1 \leqslant k \leqslant M)$. In addition, every dimension is divided into $n = [\theta d]$ continuous intervals, and $\theta$ is the real number between 0 and 1. Thus, the number of points in every interval is $G = [M/n]$.

In the $i$-th dimension, all components are sorted in an ascending order. The $k$-th sorted value is $Val[k] (1 \leqslant k \leqslant M)$. $R_{ij}$ is the $j$-th interval in the $i$-th dimension, whose lower and upper bounds are $L_{R_{ij}}$ and $U_{R_{ij}}$, respectively. It can be seen that $L_{R_{ij}} = Val[(j-1)G+1]$ and $U_{R_{ij}} = Val[jG]$.

For any two data objects $x_k$ and $y_l$ in the $d$-dimensional space, their components in the i-th dimension are $x_k^i$ and $y_l^i$, respectively, and the serial numbers of the corresponding intervals are $\gamma(x_k^i)$ and $\gamma(x_l^i)$, as follows.

$$\gamma(x_k^i) = j_k,\ L_{R_{ij_k}} \leqslant x_k^i \leqslant U_{R_{ij_k}} \tag{2}$$

$$\gamma(y_l^i) = j_l,\ L_{R_{ij_l}} \leqslant y_l^i \leqslant U_{R_{ij_l}} \tag{3}$$

For $x_k$ and $y_l$, the set of dimensions in which components fall into the same interval is

$$S_1 = \{ i \mid \gamma(x_k^i) = \gamma(y_l^i) \} \tag{4}$$

If the $i$-th components of $x_k$ and $y_l$ fall into the adjacent intervals, and the distance between them is less than the average length of the two adjacent intervals, the two points are regarded as close to each other, and included in the similarity calculation. The set of these dimensions is shown as

$$S_2 = \{ i \mid (\mid \gamma(x_k^i) - \gamma(y_l^i) \mid = 1) \wedge (\mid x_k^i - y_l^i \mid < \mid U_{R_{i\gamma(x_k^i)}} - L_{R_{i\gamma(x_k^i)}} \mid + \mid U_{R_{i\gamma(y_l^i)}} - L_{R_{i\gamma(y_l^i)}} \mid /2) \} \tag{5}$$

The set of dimensions included in the similarity calculation is the union of $S_1$ and $S_2$:

$$S = S_1 \mathrm{U} S_2 \tag{6}$$

### 2.3 Similarity measurement

The $Psim(X,Y)$ function proposed by Yi is suitable for a variety of data types[10]; however, its range is dependent on the spatial dimensionality, and thus the comparison of similarity in different dimensions is not possible. Under the circumstance of maintaining effects, $Psim(X,Y)$ is corrected as

$$NPSim(X,Y) = \sum_{j=1}^{d} \frac{1}{d} \cdot \delta(X_j,\ Y_j) \cdot \left(1 - \frac{\mid X_j - Y_j \mid}{U_{R_j} - L_{R_j}}\right) \cdot \frac{E(X,Y)}{d} \tag{7}$$

where $X$ and $Y$ are any two points in the $d$-dimensional space, and $X_j$ and $Y_j$ are components in the $i$-th dimension. Moreover, $\delta(X_j,\ Y_j)$ is the discriminant function. If $X_j$ and $Y_j$ are in the same interval $[L_{R_j},\ U_{R_j}]$, $\delta(X_j,\ Y_j) = 1$, otherwise $\delta(X_j,\ Y_j) = 0$. $E(X,Y)$ represents the number of intervals in which components of $X$ and $Y$ are all the same. The range of $NPsim(X,Y)$ is observed to be in $[0,\ 1]$. The above is the outline of $NPsim$, and the detailed introduction can be found in Ref. [10].

## 3 Experiment

To validate this method, three data types with different distributions were generated through Matlab. Next, the similarities in different dimensions were calculated using the proposed method, and were compared with the result obtained from calculating Manhattan distance, Euclidean distance, $Hsim(X,Y)$, $Gsim(X,Y)$, $Close(X,Y)$, $Esim(X,Y)$, and $Psim(X,Y)$.

### 3.1 Data description

The following three data types were used in the experiment[10].

(1) Independent and identically distributed (IID): Here, all variables obey the same data distribution function but are independent of each other. The IID data $Z$ is generated by $Z = (Z_1, \cdots, Z_M)$, and $Z_i$ follows the distribution of $Z_i \sim F(0,1)$.

(2) Relevant and identically distributed (RID): The data in every dimension are generated independently but are related to each other. The generation method is as follows. First, the $d$-dimension random variables $W_1, \cdots, W_M$ are generated, and $W_i$ follows the distribution of $W_i \sim F(0, \sqrt{i})$. Then, $Z_1$ is generated as $Z_1 = W_1$. Next, $Z_1 (2 \leqslant i \leqslant M)$ is generated according to $Z_i = W_i + Z_{i-1}/2$. Finally, the RID data $Z$ is produced as follows: $Z = (Z_1, \cdots, Z_M)$.

(3) Dependent and identically distributed (DID): All variables obey the same data distribution but are not independent. In addition, two dimensions are independent of each other called "free dimensions"; the other dimensions are related to them. The DID data $Z$ is generated as follows. First, two $d \times 1$ random variables $A$ and $B$ obeying the distribution of $F(0,1)$ are generated. Second, two $1 \times M$ random variables $U$ and $V$ obeying the distribution of $F(-1,\ 1)$ are produced. Third, $Z_1 (2 \leqslant i \leqslant M)$ is generated through $Z_i = A \times U_i + B \times V_i$. At last, the DID data $Z$ is produced as $Z = (Z_1, \cdots, Z_M)$.

### 3.2 Relative difference

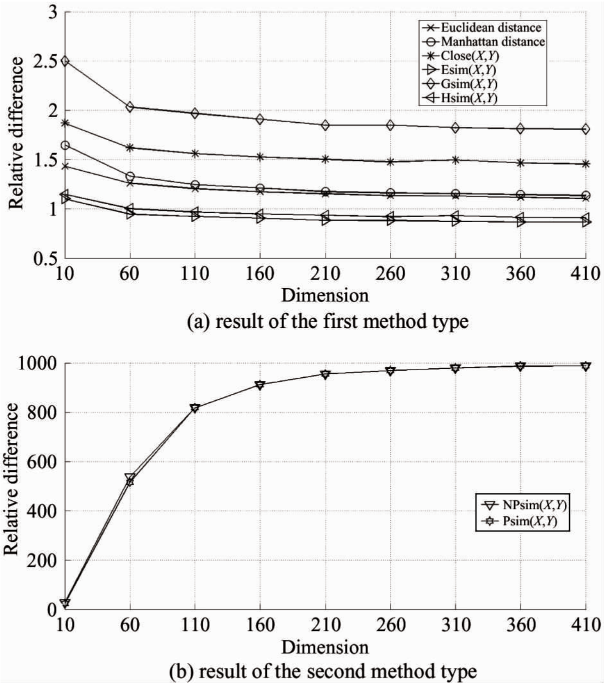
To validate this method, IID, RID, and DID data are generated using a normrnd ( ) function of Matlab[10]. The dimension of every data type is as follows: 10, 60, 110, 160, 210, 260, 310, 360, and 410. The number of data in every dimension is 1000. In addition, the relative difference between the farthest and nearest neighbors is calculated as follows[17]:

$$v = \frac{D_{\max n} - D_{\min n}}{D_{\mathrm{avg} n}} \tag{8}$$

where $D_{\max n}$, $D_{\min n}$, and $D_{\mathrm{avg} n}$ are maximal, minimal, and average similarities in the $d$-dimensional space, respectively. The relative difference results are shown in Figs 3 ~ 5.

According to the characteristics of the results, similarity measurement methods are divided into two types: the first includes Manhattan distance, Euclidean distance, $Hsim(X,Y)$, $Gsim(X,Y)$, $Close(X,Y)$, and $Esim(X,Y)$; and the others include $Psim(X,Y)$ and $NPsim(X,Y)$. The relative difference of the second type of methods is two or three magnitudes larger than that of the first type of methods. Therefore, the performance advantage of the second method type is obvious.

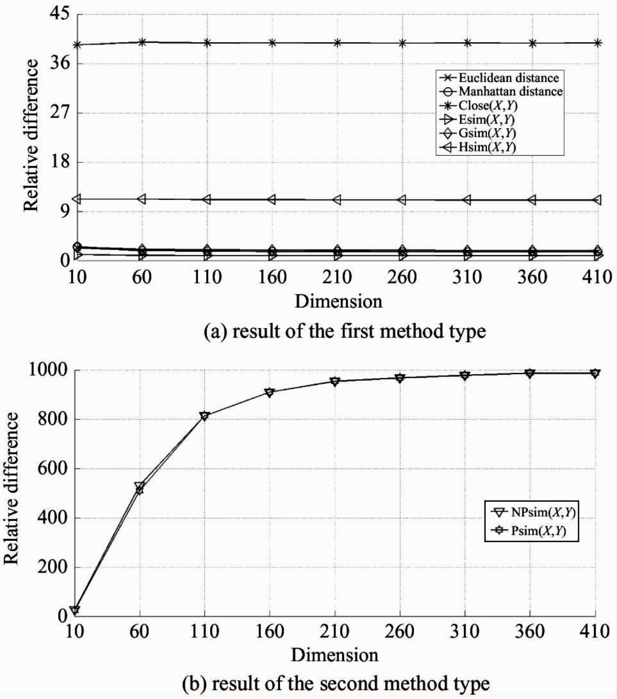
The relative difference of $Psim(X,Y)$ and $NPsim(X,Y)$ has no differentiation degree. Thus, the statistical analysis needs to be studied further.



(a) result of the first method type

(b) result of the second method type

**Fig. 3** Relative difference of various similarity measurement methods for IID data



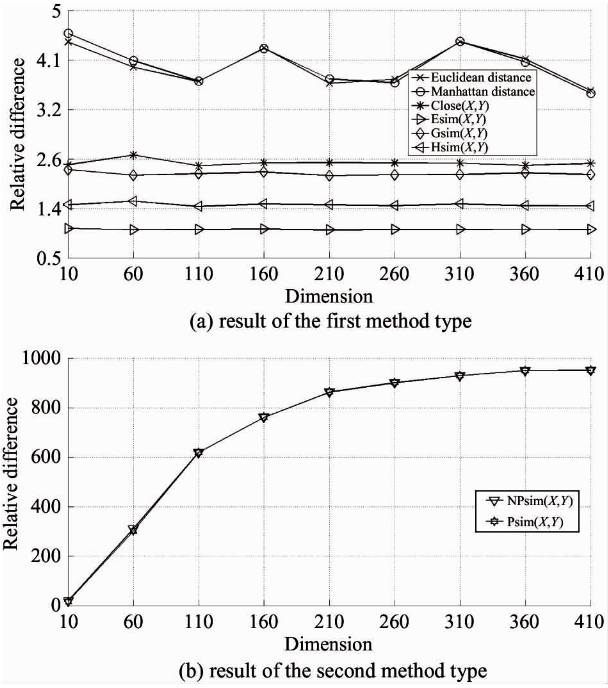
(a) result of the first method type

(b) result of the second method type

**Fig. 4** Relative difference of various similarity measurement methods for RID data

### 3.3 Statistical analysis

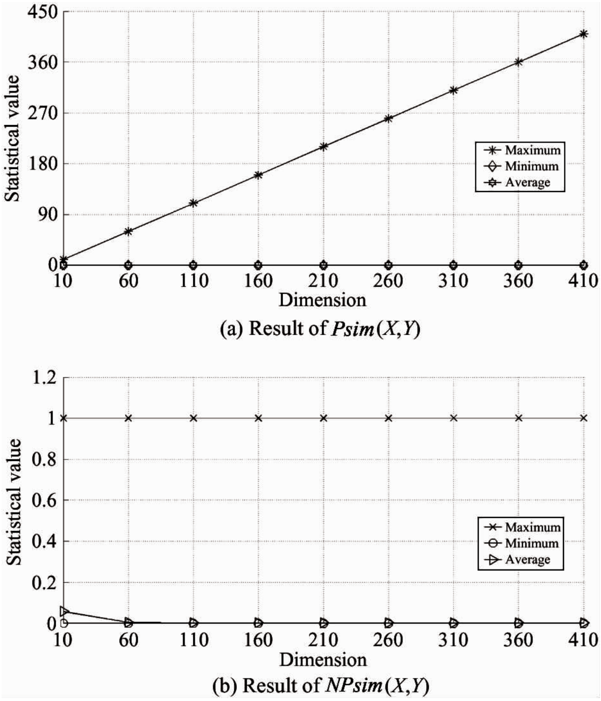
To compare the effect of $Psim(X,Y)$ and $NPsim(X,Y)$, the maximum, minimum, and average of DID data in different dimensions are calculated, as shown in Fig. 6. The experimental results indicate that the similarity range of $Psim(X,Y)$ increases with the dimension. Thus, the function is not suitable for the similarity comparison in different dimensions. However, the problem does not exist in $NPsim(X,Y)$. Table 1 lists the numbers of $Psim(X,Y)$ whose value is greater than 1 in different dimensions. The number of



(a) result of the first method type

(b) result of the second method type

**Fig. 5** Relative difference of various similarity measurement methods for DID data



(a) Result of *Psim*(X,Y)

(b) Result of *NPsim*(X,Y)

**Fig. 6** Statistical value of various similarity measurement methods for DID data

Table 1 Number of *NPsim* $(X, Y) >= 1$ in different dimensions

| Dimension | 10 | 60 | 110 | 160 | 210 |
|---|---|---|---|---|---|
| Number | 168604 | 120373 | 113248 | 10452 | 84672 |
| Dimension | 260 | 310 | 360 | 410 | 260 |
| Number | 98429 | 63024 | 72015 | 58851 | 98429 |

$Psim(X,Y)$ in every dimension is $1000 \times 1000 = 1,000,000$. In addition, the 5% ~ 17% result is more than 1, and thus the comparison of similarity in different dimensions is not possible. Therefore, $NPsim(X, Y)$ can satisfy the requirement of similarity comparison in different dimensions.

## 4 Conclusion

The similarity measurement is the basis of data-mining algorithms, such as clustering, classification, and nearest neighbor. However, owing to the curse of dimensionality, the measurement always fails in high-dimensional space. A similarity measurement method of high-dimensional data based on a normalized net lattice subspace is proposed. In this method, data range of each dimension is divided into several intervals, and the components are mapped onto the corresponding intervals. During similarity calculation, only the component in the same or adjacent interval is used. This method can avoid the similarity effect that generated from the sparse or noisy dimension. To validate the proposed algorithm, three types of distribution data are used in the experiment, and another seven method types are compared. The experimental results show that the proposed method is suitable for similarity measurement in high-dimension data.

In the future, the weight calculation in different dimensions, and the automatic updating strategy of a spatial grid will be studied. In addition, the proposed method will apply a related data-mining algorithm, such as clustering or correlation analysis.

**References**

[1] Tan P N, Michael S, Vipin K. Introduction to Data Mining. Boston: Addison-Wesley Publishing Company, 2005. 25-36

[2] Chen J N. The Research and Application of Key Technologies in Knowledge Discovery of High-dimensional Clustering. Beijing: Publishing House of Electronics Industry, 2011. 120-128(In Chinese)

[3] Aggarwal C C. Re-designing distance functions and distance based applications for high dimensional data. *ACM SIGMOD Record*, 2001, 33(1):117-128

[4] Warren B P. Approximate Dynamic Programming: Solving the Curses of Dimensionality (2nd Edition). Hoboken, New Jersey: John Wiley & Sons Press, 2011. 124-161

[5] Yang F Z, Zhu Y Y. An efficient method for similarity search on quantitative transaction data. *Journal of Computer Research and Development*, 2004, 41(2):361-368

[6] Xie M X, Guo J Z, Zhang H B, et al. Research on the similarity measurement of high dimensional data. *Computer Engineering and Science*, 2010, 32(5):92-96(In Chinese)

[7] Huang S D, Chen Q M. On clustering algorithm of high dimensional data based on similarity measurement. *Computer Applications and Software*, 2009, 26(9):102-105 (In Chinese)

[8] Shao C S, Lou W, Yan L M. Optimization of algorithm of similarity measurement in high-dimensional data. *Computer Technology and Development*, 2011, 21(2):1-4(In Chinese)

[9] Wang X Y, Zhang H Y, Shen L Z, et al. Re-search on high dimensional clustering algorithm based on similarity measurement. *Computer Technology and Development*, 2013, 23(5):30-33(In Chinese)

[10] Yi L H. Research on clustering algorithm for high dimensional data:[Ph. D dissertation]. Qinhuangdao: Institute of Information Science and Engineering, Yanshan University, 2011. 28-30(In Chinese)

[11] Ericson K, Pallickara S. On the performance of high dimensional data clustering and classification algorithms. *Future Generation Computer Systems*, 2013, 29(4):1024-1034

[12] Zhao H. Study on Some Issues of Data Clustering in Data Mining:[Ph. D dissertation]. Xi'an: School of Electronic Engineering, Xidian University, 2005. 35-42(In Chinese)

[13] Jia X Y. A high dimensional data clustering algorithm based on twice similarity. *Journal of Computer Applications*, 2005, 25(B12):176-177

[14] Alexander H, Charu A C, Keim D A. What is the nearest neighbor in high dimensional spaces? In: Proceedings of the 26th International Conference on Very Large Data Bases, Cairo, Egypt, 2000. 506-515

[15] Yuan R P, Shi M R. Research on the similarity of high dimensional big data based on extenics. *Operations Research and Management Science*, 2015, 24(5):184-188

[16] Kriegel H P, Kröger P, Zimek A. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 2009, 3(1):1-58

[17] Charu C, Aggarwal, Yu P S. The IGrid index: reversing the dimensionality curse for similarity indexing in high dimensional space. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, USA, 2000. 119-129

**Li Wenfa**, born in 1974. He received his Ph. D. degree in Graduate University of Chinese Academy of Sciences in 2009. He also received his B. S. and M. S. degrees from PLA Information Engineering University in 1998 and 2003 respectively. His research interests include information security, data analysis and mining, etc.