Research on Similarity Measurement Algorithm of High-dimensional Data

Jianhua Zhang\textsuperscript{1*}, Xiaojun Meng\textsuperscript{1}, Huidong Huangfu\textsuperscript{1}, Jianglong Zhou\textsuperscript{1}, Haiyan Chen\textsuperscript{1}, Jiachen Shen\textsuperscript{1}, Peng Zhang\textsuperscript{1} and Wei Tao\textsuperscript{1}

\textsuperscript{1}Northwest Institute of Nuclear Technology, Xi’an, Shaanxi, 710024, China
\textsuperscript{*}E-mail: chashuwan@ydgc1009.onexmail.com

Abstract. The similarity measurement of high-dimensional data is an important research content in data mining and other fields. Aiming at the shortcomings of the traditional similarity measurement algorithms, a new measurement algorithm is proposed. The results of comparison with traditional algorithms show that new algorithm can quantitatively and intuitively reflect the similarity of two sets of high-dimensional data, and has strong adaptability and superiority.

1. Introduction

The High-dimensional data similarity analysis is one of the important supporting content in data mining and other fields\cite{1-2}. In recent years, it has been increasingly used in user interest information push, information processing and retrieval, and monitoring data analysis.

Traditional similarity algorithms include the distance function, the correlation coefficient and the hypothesis test, etc. The results obtained by using the distance function and correlation coefficient to calculate the similarity of the two sets of high-dimensional data are difficult to accurately reflect the closeness of the two sets of data, and there is a phenomenon that the calculation results cannot correctly reflect the actual similarity. The similarity discrimination based on the hypothesis test can only qualitatively reflect the fitting degree of the two sets of data, and it is impossible to make further judgments on the degree of similarity.

This paper analyzes the problems of traditional similarity algorithms and proposes a new similarity measurement algorithm. The verification results show that the algorithm can make up for the shortcomings of traditional similarity algorithms and can quantitatively and intuitively reflect the similarity of high-dimensional data.

2. The traditional similarity algorithms

For the two sets of high-dimensional data to be compared, this paper refers the benchmark data set as the theoretical data set, and the other as the actual data set, and suppose that the theoretical data set is $Y=(y_1, y_2, \cdots, y_n)$ and the actual data set is $X=(x_1, x_2, \cdots, x_n)$.

Traditional similarity algorithms include the distance function, the correlation coefficient and the hypothesis test, etc.

2.1. The distance function

Common distance functions include Euclidean Distance, Chebyshev Distance, Manhattan distance and Minkowski Distance, etc. Among them, Euclidean Distance, Chebyshev Distance and Manhattan Distance are a special representation of Minkowski Distance respectively\cite{3-4}.
The distance functions are widely used, and choosing different distance function in different fields can usually achieve the better description effect. However, if the distance function is simply used in the similarity analysis of the actual data set $X$ and the theoretical set $Y$, the result obtained is usually meaningless.

This paper takes the Euclidean Distance as an example. For the three data points $A(2,2)$, $B(4,10)$ and $C(5,1)$ in the plane, the Euclidean distance between $A$ and $B$ is $|AB| = \sqrt{17}$, and the distance between $A$ and $C$ is $|AC| = \sqrt{10}$. It can be seen that the AC distance is relatively short, and it can be concluded that "point $C$ is more similar to point $A$ than point $B$". Assuming that there are only two points $A(2,2)$ and $C(5,1)$, the AC distance is calculated as $|AC| = \sqrt{10}$, and it is difficult to draw a conclusion whether points $A$ and $C$ are similar through $|AC|$. Similarly, for the actual data set $X$ and theoretical data set $Y$, if other constraints are not added, the result calculated by the distance function is meaningless.

2.2. The correlation coefficient

The correlation coefficient is a quantity that reflects the degree of linear correlation between variables. There are many ways to define it according to the different research objects[5]. The correlation coefficient mentioned in this paper is Pearson correlation coefficient which is more commonly used. The calculation formula is as follows.

$$r_{x,y} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(1)

Among them, $\text{Cov}(X, Y)$ is the covariance of $X$ and $Y$, $D(X)$ and $D(Y)$ are the variances of $X$ and $Y$ respectively, and $\bar{x}$ and $\bar{y}$ are the average values of data sets $X$ and $Y$ respectively.

The value range of the correlation coefficient is $[-1, 1]$. The larger the absolute value, the stronger the correlation between the two sets of data. The correlation coefficient can well represent the linear trend of one variable with another variable, but when measuring the similarity of two sets of data, there are the following shortcomings.

(1) It is difficult to accurately reflect the closeness of the two sets of data.

(2) There is a phenomenon that the value cannot correctly reflect the actual degree of similarity.

Example 1: For (1), this paper assumes the actual data sets $X_1$, $X_2$ and the theoretical data set $Y$ as shown in Table 1.

| Item | Data | Correlation Coefficient |
|------|------|-------------------------|
| Y    | 25   | 26 27 28 29 30 31 32 33 34 35 |
| $X_1$| 24   | 25 26 27 28 29 30 31 32 33 34 |
| $X_2$| 23   | 24 25 26 27 28 29 30 31 32 33 |

Among them, the difference between the data in the actual data set $X_1$ and the corresponding data in the theoretical data set $Y$ is 1, and the correlation coefficient between $X_1$ and $Y$ is 1. The difference between the data in the actual data set $X_2$ and the corresponding data in the theoretical data set $Y$ is 2, and the correlation coefficient between $X_2$ and $Y$ is also 1. However, it is generally considered that the data of $X_1$ is closer to the theoretical value which means higher similarity, but the conclusion cannot be obtained by the value of correlation coefficient. If the correlation coefficient is used to characterize the similarity, the similarity between $X_1$ and $Y$ and the similarity between $X_2$ and $Y$ are both 1 (100%), which is obviously unreasonable.

Example 2: For (2), this paper assumes the actual data sets $X_3$, $X_4$ and the theoretical data set $Y$ as shown in Table 2.
Table 2. Data of Example 2.

| Item | Data | Correlation Coefficient |
|------|------|-------------------------|
| $Y$  | 25 26 27 28 29 30 31 32 33 34 35 |                         |
| $X_3$ | 35 26 27 28 29 30 31 32 33 34 35 | 0.569                   |
| $X_4$ | 35 26 27 28 29 30 31 32 33 38 35 | 0.612                   |

Among them, the actual data set $X_3$ and the theoretical data set $Y$ differ only in the first dimension data, and $X_4$ and $Y$ differ in the first dimension data and the tenth dimension data. The correlation coefficient between $X_3$ and $Y$ is 0.569, and the correlation coefficient between $X_4$ and $Y$ is 0.612. In actual analysis, it is obviously that the similarity of $X_3$ is higher than the similarity of $X_4$, but the conclusion drawn by the correlation coefficient is not the case, i.e. the correlation coefficient cannot correctly reflect the actual degree of similarity.

2.3. The hypothesis test
The hypothesis test is a method in mathematical statistics to infer the population from a sample based on certain assumptions.

In the similarity evaluation of two sets of data, the Chi-square test or Kolmogorov-Smirnov test can be used. The theoretical data set is regarded as a known theoretical distribution, and it can be judged whether the actual data distribution is consistent with the theoretical distribution through hypothesis test, so as to judge the similarity of two sets of data. However, the hypothesis test can only qualitatively reflect the fitting degree between the actual data and theoretical data under the selected significance level, and cannot reflect the similarity of the two sets of data quantitatively and intuitively.

Figure 1. The probability density function curve and cumulative distribution function curve of $X_5$ and $X_6$.

Figure 2. The probability density function curve and cumulative distribution function curve of $X_7$ and $X_8$. 
Example 3: This paper generates two sets of random arrays that obey the normal distribution \( N(0,100) \) as \( X_5 \) and \( X_6 \), and generate \( X_7 \) and \( X_8 \) that obey the normal distribution \( N(8,105) \), each with 1000 data.

Using the Kolmogorov-Smirnov test to analyze the similarity of \( X_5 \) and \( X_6 \), \( X_7 \) and \( X_8 \) respectively, the probability density function curve and cumulative distribution function curve are shown in Figure 1 and Figure 2.

When the significance level is 0.05, the analysis results show that the distributions of \( X_5 \) and \( X_6 \) are consistent, and the distributions of \( X_7 \) and \( X_8 \) are consistent, that is, the array \( X_5 \) is similar to the array \( X_6 \), and the array \( X_7 \) is similar to the array \( X_8 \). In the two cases, it is difficult to draw conclusions from the analysis results that which case is more similar or how similar is.

3. The improved similarity algorithm

In order to better judge the similarity of the actual data set and the theoretical data set, the algorithm should at least meet the following two conditions.

(1) The similarity result should have a uniform range. When the value range is uniform, the similarity result can directly reflect the similarity of the two sets of data, which is convenient for comparing the analysis results of different groups. The value range in this paper is (0,1], the larger the value, the more similar the two sets of data, the closer to 0, the greater the difference between the two sets of data, and the 1 indicates that the two sets of data are completely the same.

(2) Under the premise of the same degree of attention to each data, the contribution of each dimension of data should be equivalent and independent. Equivalence means that the impact of each dimension of data on the overall similarity is consistent and limited, and independence means that the similarity of single-dimensional data in a certain dimension should not have an impact on the similarity of single-dimensional data in other dimensions.

3.1. Analysis of new algorithm

For the \( i \)-th dimension data \( x_i \) and \( y_i \) in the actual data set \( X \) and the theoretical data set \( Y \) respectively, define the similarity of the single-dimensional data as \( \rho_i \), which represents the similarity of the two data \( x_i \) and \( y_i \). The value range of \( \rho_i \) is set as (0,1]. The greater the difference between \( x_i \) and \( y_i \), the closer \( \rho_i \) is to 0, which means the less similar it is. The smaller the difference between \( x_i \) and \( y_i \), the closer \( \rho_i \) is to 1, which means the more similar it is. When \( \rho_i \) is 1, it means the two numbers are equal.

In order to effectively avoid the inconsistency of the contribution of each dimension data, this paper uses the absolute value of the ratio of the difference between \( x_i \) and \( y_i \) to the theoretical value \( y_i \) (hereinafter referred to as the difference ratio), as the calculation parameter, and it is easy to know \( |\frac{x_i - y_i}{y_i}| \in [0, +\infty) \).

The single-dimensional data similarity algorithm needs to meet the following conditions.

(1) The value range of the difference ratio \( |(x_i - y_i)/y_i| \) can be mapped to the value range of \( \rho_i \) one by one, and the larger \( |(x_i - y_i)/y_i| \) is, the closer \( \rho_i \) is to 0.

(2) When \( |(x_i - y_i)/y_i| \) is small, the algorithm curve should be in a slow downward trend to avoid faster filtering of similar data. When \( |(x_i - y_i)/y_i| \) is greater than a certain value, the algorithm curve should be in a flat state, and the abscissa axis is the asymptote of the curve. In the value range of \( |(x_i - y_i)/y_i| \), the overall trend of the curve is "stable in the front, flat in the back end and fast in the middle".

Based on the above considerations, this paper is inspired by the normal distribution curve, and the single-dimensional data similarity algorithm used is as follows.

\[
\rho_i = e^{-5\left|\frac{x_i - y_i}{y_i}\right|} \quad (y_i \neq 0)
\]

The curve of the algorithm is shown in Figure 3.
Figure 3. The curve of the algorithm.

Since the probability that the value in the theoretical data set $Y$ is equal to 0 on a large scale is extremely low, for the case of $y_i=0$, two solutions are considered as follows.

1) You can eliminate the dimensions with a value of 0 in the theoretical data set $Y$, and at the same time eliminate the corresponding dimension in the actual data set $X$, and then analyze the $X$ and $Y$ after eliminating the value of 0.

2) For all dimensions with a value of 0 in the theoretical data set $Y$, you can find the corresponding data in the actual data set $X$, and assign the corresponding $\rho_i$ through data analysis and manual decision-making.

This paper defines the overall similarity of the actual data set $X$ and the theoretical data set $Y$ as $\rho$, and the calculation formula is as follows.

$$\rho = \frac{\sum_{i=1}^{n} \rho_i}{n}$$

The similarity $\rho$ has the following properties.

1) The value range of $\rho$ is $(0,1]$, the larger the value, the more similar the actual data set $X$ and the theoretical data set $Y$.

2) The minimum value of $\rho$ tends to 0 infinitely, which means that the difference ratio between each dimension of the actual data set $X$ and the theoretical data set $Y$ is close to infinity, and the similarity of $X$ and $Y$ is the smallest.

3) The maximum value of $\rho$ is 1, which means that the actual data set $X$ and the theoretical data set $Y$ have the same value in each dimension. In this case, $X$ and $Y$ are exactly the same, and the similarity is the largest.

3.2. Algorithm comparison

The improved similarity algorithm takes the difference ratio as the calculation parameter, and unifies the similarity value range to $(0,1]$, which can avoid the problem of meaningless parameters generated by simply applying the distance function, and can also solve the problem that the hypothesis test method cannot quantitatively and intuitively reflect the similarity of the two sets of data.

The following analyzes the data in the two table in section 2.2 to compare the improved algorithm and the correlation coefficient method.

As shown in Table 3, using the improved algorithm to calculate the similarity of $X_1$ and $Y$, $X_2$ and $Y$ respectively, the results are 0.994 and 0.977. Compared with the correlation coefficient method, the improved algorithm eliminates the disadvantage that are difficult to accurately reflect the closeness of
the two sets of data. The results obtained by the improved algorithm is consistent with the logical thinking of judging whether the two sets of data are similar, and more objectively reflects the similarity.

Table 3. Results of the correlation coefficient and improved algorithm of Example 1.

| Item | Data | Correlation Coefficient | Improved Algorithm |
|------|------|--------------------------|--------------------|
| Y    | 25 26 27 28 29 30 31 32 33 34 35 |               |                  |
| \(X_1\) | 24 25 26 27 28 29 30 31 32 33 34 1 | | 0.994 |
| \(X_2\) | 23 24 25 26 27 28 29 30 31 32 33 1 | | 0.977 |

It can be seen from Table 4 that for \(X_3\) and \(Y\), the correlation coefficient is 0.569, and the similarity calculated by the improved algorithm is 0.950. If the correlation coefficient 0.569 is used as the similarity measure, the influence of the first dimension is enlarged, and the result obviously does not conform to the conventional judgment logic. The improved algorithm can ensure that the influence of each data on the similarity is equivalent and limited, and the calculation result 0.950 can better reflect the similarity of \(X_3\) and \(Y\).

Table 4. Results of the correlation coefficient and improved algorithm of Example 2.

| Item | Data | Correlation Coefficient | Improved Algorithm |
|------|------|--------------------------|--------------------|
| Y    | 25 26 27 28 29 30 31 32 33 34 35 |               |                  |
| \(X_3\) | 35 26 27 28 29 30 31 32 33 34 35 | 0.569       | 0.950             |
| \(X_4\) | 35 26 27 28 29 30 31 32 33 38 35 | 0.612       | 0.944             |

For \(X_4\) and \(Y\), the correlation coefficient is 0.612, and the similarity calculated by the improved algorithm is 0.944. Taking the theoretical data set \(Y\) as the standard, the actual data \(X_4\) has one more different data than \(X_3\), but the correlation coefficient "does not fall but rises", which obviously does not conform to the conventional judgment logic, and the improved algorithm solves the problem.

4. Conclusion

Aiming at the shortcomings of traditional similarity algorithms, this paper proposes a new high-dimensional data similarity algorithm. The test results show that the algorithm has strong adaptability, and can avoid the problems that arise when the distance function, the correlation coefficient and the hypothesis test are applied to the similarity evaluation of high-dimensional data. The next step of research can consider introducing a weighting operator and a penalty factor to control the trend of the algorithm curve, so that the algorithm is suitable for similarity evaluation applications with different degrees of attention to each dimension of data.

References

[1] SHAO C, LOU W, Yan L. (2011) Optimization of Algorithm of Similarity Measurement in High-dimensional Data. COMPUTER TECHNOLOGY AND DEVELOPMENT, 21(2): 1-4.
[2] HE L, CAI Y, YANG Z. (2011) A Meshing Method for High-dimensional Data Space. COMPUTER ENGINEERING AND APPLICATIONS, 47(5): 152-153.
[3] HE L, WU L, CAI Y. (2006) Research on the Similarity Measurement of Data in High-dimensional Space. MATHEMATICS IN PRACTICE AND THEORY, 36(9): 189-194.
[4] XIA Y, ZHU X, ZHOU C. (2007) Research on Similarity Query of Spatial Data Based on Feature. COMPUTER ENGINEERING AND APPLICATIONS, 43(25): 15-17+47.
[5] XIE M, GUO J, ZHANG H, CHEN K. (2010) Research on the Similarity Measurement of High Dimensional Data. COMPUTER ENGINEERING & SCIENCE, 32(5): 92-96.