Computing, Performance and Communication systems (2016) 1: 22-27
Clausius Scientific Press, Canada

An Anonymous Algorithm for Calculating Dissimilarity Metric on Clustering

Li Chi[1,a], Sun Yi[1,b]*

[1]Beijing University of Posts and Telecommunications, China
[a]lichi@bupt.edu.cn , *sunyisse@bupt.edu.cn

Keywords: dissimilarity metric; multi-category; interval-scaled attribute; binary attribute; nominal attribute; ordinal attribute; proportional-scaled attribute; symmetric binary attribute; asymmetric binary attribute; clustering

Abstract: By the research on calculating the dissimilarity metric among tuples with many different attributes based on clustering, this paper improves dissimilarity metric algorithm, which can more accurately reflect the differences between tuples. Besides, in terms of various attribute types ,the value of attribute is divided into multi-category. According to the multi-category, we come to the final dissimilarity metric result through analysis. The experimental results show that this algorithm is able to achieve highly accurate dissimilarity metric results.

1. Introduction

Dissimilarity metric [1] is expressed in terms of a distance function and token values in [0.0,1.0], but its dissimilarity metric result depends on various attributed types including interval-scaled [2], binary [2], nominal [2], ordinal ratio [2] and proportional-scaled attributes [3]. Therefore, this article proposes the method of calculating the dissimilarity on various attribute types.The work first divides the various attribute types into two big categories. One category's attributes content only 0 and 1 which is called the binary attribute category. The other category's attributes content numerical values which are called the interval-scaled attribute category. The binary attribute category includes binary attributes and nominal attributes which can be regarded as the extension of binary attributes. For the binary attribute category, this article mainly analyzes symmetric binary attributes and asymmetric binary attributes. By the research on the confidence level [4], this article gets the proportion of the symmetric binary and the asymmetric binary. To symmetric binary attributes, this article uses symmetric binary dissimilarity [2] to calculate the dissimilarity.To asymmetric binary attributes, this article uses asymmetric binary dissimilarity [2] to calculate the dissimilarity. Finally, with the consideration of weights, we achieve the dissimilarity of the binary attribute category. Because ordinal ratio attributes and proportional-scaled attributes both can be transformed into interval-scaled attributes, the interval-scaled attribute category includes interval-scaled attributes, ordinal ratio attributes and proportional-scaled attributes. For the interval-scaled attribute category, this article first converts both ordinal ratio attributes and proportional-scaled attributes with data normalization [2]. Then attributes in the interval-scaled attribute category can all be regarded as interval-scaled attributes. Processing interval-scaled attributes with data normalization and adapt the normalization result to Euclidean distance formula [2] will result in transforming the data to fall within the range [0.0, 1.0] [2]. Finally, we achieve the dissimilarity of the interval-scaled attribute category. With the dissimilarity result of both the binary

attribute category and the interval-scaled attribute category, we finally get the result which can adapt to all kinds of attributes and also have a good performance on accuracy. Experimental results also show that the algorithm can achieve highly accurate dissimilarity metric results.

2. Calculated the multi-category proportion

Assume that the input data is a $m \times n$ dimensional matrix named X, m represents the number of the tuples, n represents the number of attributes. This paper constructs a $m \times n$ dimensional matrix whose contents are 0 named Y_0 and a $m \times n$ dimensional matrix whose contents are 1 named Y_1.

Definition 1 (similarity formula between X and Y):

$$J(X,Y) = \frac{X \bigcap Y}{X \bigcup Y}. \tag{1}$$

Similarity between X and Y_0 is:

$$k_1 = \frac{X \bigcap Y_0}{X \bigcup Y_0}. \tag{2}$$

Similarity between X and Y_1 is:

$$k_2 = \frac{X \bigcap Y_1}{X \bigcup Y_1}. \tag{3}$$

The proportion of the binary attribute category is:

$$k = k_1 + k_2. \tag{4}$$

The proportion of the interval-scaled attribute category is:

$$1-k. \tag{5}$$

3. Calculated the dissimilarity of the binary variable category

Table 1 Contingency Table for Binary Attributes[2]

	Object x_j			
Object x_i		1	0	sum
	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p

If all binary attributes are considered as having the same weight, we have the 2 × 2 contingency table of Table 1, where q is the number of attributes that equal 1 for both objects x_i and x_j, r is the number of attributes that equal 1 for object x_i but equal 0 for object x_j, s is the number of attributes that equal 0 for object x_i but equal 1 for object x_j, and t is the number of attributes that equal 0 for both objects x_i and x_j [2].

Definition 2 (symmetric binary dissimilarity between tuples x_i and x_j)

$$d(x_i, x_j) = \frac{r+s}{q+t+r+s}. \tag{6}$$

Definition 3 (asymmetric binary dissimilarity between tuples x_i and x_j)

$$\mathrm{d}(x_i, x_j) = \frac{r+s}{q+r+s}. \tag{7}$$

3.1 Confidence level.

Definition 4 (confidence level represents the proportion of symmetric binary attributes)

E represents the premise of the rule, H represents the conclusion of the rule. $P(H)$ represents the occurrence possibility of H. $P(H \mid E)$ represents the occurrence possibility of H with the premise that E already occurred. $CF(H)(H,E)$ represents the confidence level. The range of $CF(H)(H,E)$ is [0.0,1.0].

$$CF(H)(H,E) = \begin{cases} \dfrac{P(H \mid E) - P(H)}{1 - P(H)} & P(H \mid E) > P(H) \\ 0 & P(H \mid E) = P(H) \\ \dfrac{P(H) - P(H \mid E)}{P(H)} & P(H \mid E) < P(H) \end{cases}. \tag{8}$$

3.2 The dissimilarity of binary variable category.

Set variable α represents the confidence level of the symmetric binary.

$$\alpha = CF(H)(H,E). \tag{9}$$

Considering the symmetric binary and the asymmetric binary, this paper uses the following improved formula:

$$d_{c1}(x_i, x_j) = k\alpha \frac{r+s}{q+t+r+s} + k(1-\alpha)\frac{r+s}{q+r+s}. \tag{10}$$

k is already defined in (4).

4. Calculated the dissimilarity of the interval-scaled attribute category

4.1 Ordinal ratio attributes.

Definition 5 (an ordinal ratio attribute normalization) The value of f for the ith object is x_{if}, and f has M_f ordered states, representing the ranking. Replace each x_{if} by its corresponding rank, $r_{if} \in \{1,2,...,M_f\}$.Since each ordinal attribute can have a different number of states, it is necessary to map the range of each attribute onto [0.0, 1.0] so that each attribute has equal weight. We perform such data normalization by replacing the rank r_{if} of the ith object in the fth attribute by [2] :

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}. \tag{11}$$

4.2 Proportional-scaled attributes.

Definition 6 (a proportional-scaled attribute normalization) x_{if} represents the i th object in the f th attribute, Z_{if} represents the normalization result:

$$Z_{if} = \log(x_{if}). \tag{12}$$

4.3 Interval-scaled attributes.

Regard the normalization result from both ordinal ratio attributes and proportional-scaled attributes as interval-scaled attributes. Then use the same way to deal with them. The normalization steps are as follow [3] :

Definition 7(average value) x_{nf} represents the value of attribute f in object n, m_f represents the average value of attribute f with total n objects.

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}). \tag{13}$$

If the attribute is an ordinal ratio attribute, $x_{nf} = Z_{if}$, Z_{if} is already defined in (11), if the attribute is a proportional-scaled attribute, $x_{nf} = Z_{if}$, Z_{if} is already defined in (12) .

Definition 8(average absolute deviation) S_f represents average absolute deviation.

$$S_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|). \tag{14}$$

Definition 9(final normalization result) Z_{if} is the final normalization result we want. The value of Z_{if} is mapped to [0.0, 1.0].

$$z_{if} = \frac{x_{if} - m_f}{S_f}. \tag{15}$$

Definition 10(Euclidean distance) X_{if} represents the value of attribute $f\ (1 \le f \le p)$ in object i.

$$d(x_i, x_j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2}. \tag{16}$$

$X_{if} = Z_{ifx}$, $Y_{if} = Z_{ify}$, Z_{if} is already defined in (15)

4.4 The dissimilarity of the interval-scaled attribute category.

According to (13) ~ (15). This paper establish the following equations.

$$\begin{cases} m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}) \\ S_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|) \\ Z_{if} = \frac{x_{if} - m_f}{S_f} \end{cases} . \quad (17)$$

According to (17), the dissimilarity of the interval-scaled attribute category is :

$$d_{c2}(Z_x, Z_y) = (1-k)\sqrt{|z_{x1} - z_{y1}|^2 + |z_{x2} - z_{y2}|^2 + ... + |z_{xp} - z_{yp}|^2} . \quad (18)$$

$Z_{xn}(1 \le n \le p)$ represents the value of attribute $n(1 \le n \le p)$ in the final normalization object x.

$1-k$ is already defined in (5).

5. Calculated the dissimilarity metric

Combine the dissimilarity of the binary variable category and the dissimilarity of the interval-scaled attribute category, then this paper achieve the dissimilarity metric:

$$d(x_i, x_j) = k\alpha\frac{r+s}{q+t+r+s} + k(1-\alpha)\frac{r+s}{q+r+s} + (1-k)\sqrt{|z_{x1} - z_{y1}|^2 + |z_{x2} - z_{y2}|^2 + ... + |z_{xp} - z_{yp}|^2} \quad (19)$$

6. Experimental results

In order to verify the accurate dissimilarity metric results of the improved algorithm, this paper carried out experiments. The experiment uses user dataset in the web logs. We use the top 600 users and the top 600 web pages which are accessed frequently for analysis.

There are 600 tuples and 600 columns in the user-page 600×600 dimensional matrix(after deleting the data within missing value).The dataset contains 34216 non-zero value. The experiment establishes user's dissimilarity matrix on the basis of user's statistics information.Some of the data fragments are as follows:

(1.0, 1.0), (0.0, 0.0), (0.0, 0.0), (1.0, 1.0), (0.0, 0.0), (0.0, 1.0), (1.0, 0.08), (0.0, 0.060606060606)…

(0.0,0.0),(1.0,1.0),(0.0,0.0),(1.0,1.0), (0.0, 0.0), (0.0, 0.0), (0.0855487257912, 0.0684150513112),…

(0.0, 0.0), (0.0, 0.0), (1.0, 1.0), (0.222222222222, 0.270228582622),(0.0, 0.0), (0.35 , 0.336),…

(1.0,1.0),(0.0,0.0),(0.222222222222,0.270228582622),(1.0,1.0),(0.2857142714,0.2742905780),…

The experiment gets the confidence level by sampling. After the multiple sampling, the confident level of the symmetric binary in the binary attribute category is 0.47. k value is 0.56. The dissimilarity metric between users can be depended on the visited times to the web page and the the time staying at the web page.The following calculation is based on the visited times to the web page, then we get the dissimilarity between x_i and x_j.

$$d(x_1, x_2) = 0.56*0.47*\frac{26}{15+50} + 0.56*(1-0.47)*\frac{26}{50} + (1-0.56)*\sqrt{(0.0 - 0.08554)^2 + (0.65721 - 0.35295)^2 + ... + (0.23493 - 0.31954)^2} = 0.398$$

Establish dissimilarity matrix:

$$d(x_i, x_j) = \begin{bmatrix} (0,0) & & & & \\ (0.398,0.209) & (0,0) & & & \\ (0.652,0.759) & (0.738,0.779) & (0,0) & & \\ (0.72,0.788) & (0.786,0.728) & (0.201,0.364) & (0,0) & \\ (0.233,0.272) & (0.337,0.122) & (0.739,0.795) & (0.807,0.806) & (0,0) \end{bmatrix}$$

$d(x_i, x_j)$ can achieve highly accurate dissimilarity metric results.

7. Summary

In the study of dissimilarity metric, the cost-sensitive attribute [5] can result in inaccurate dissimilarity metric results.This paper presents the improved dissimilarity metric calculation algorithm and this algorithm can greatly improve the accuracy of dissimilarity metric results in clustering.

Acknowledgements

This work was financially supported by the Research Innovation Fund for College Students of Beijing University of Posts and Telecommunications.

References

[1] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M.B.Vitanyi. IEEE Transactions on Information Theory, December 2004, Vol. 50. No. 12.

[2] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques Third Edition, 2012: 65-73

[3] Zirong Yang, Research on domain oriented high performance information retrieval based on Web data mining

[4] Thomas Junk, Confidence level computation for combining searches with small statistics, Nuclear Instruments and Methods in Physics Research Section A:Accelerators, Spectrometers, Detectors and Associated Equipment, 21 September 1999, Volume 434, Issues 2-3, pp. 435-443.

[5] Yu Fang, Zhong-Hui Liu, Fan Min, Multi-objective cost-sensitive attribute reduction on data with error ranges, International Journal of Machine Learning and Cybernetics, October 2016, Volume 7, Issue 5, pp. 783-793.