New Clustering Algorithm for Multidimensional Data

Linyuan Fan[1], Jingyang Zhong[2]
[1]School of Statistics, Capital University of Economics and Business
Beijing 100070, China
[2]Department of Mathematics, University of California
Santa Cruz, CA 95064, U.S.A
fanlinyuan@cueb.edu.cn

Journal of Digital
Information Management

ABSTRACT: *Calculating similarity for multidimensional data is one of the key problems that must be addressed in order to promote the development of data clustering algorithms. In this study, we developed and tested a new similarity calculation index to improve the accuracy of multidimensional data clustering. First, the information divergence (ID) and generalized gradient angle (GGZ) were explored in detail. Second, the ID and GGZ were combined to calculate the similarity of multidimensional data, thus enabling a new algorithm for data clustering. Finally, two experiments were conducted to evaluate the performance of the proposed algorithm. The results of the experiments demonstrate that our proposed similarity calculation index for multidimensional data is both accurate and effective, providing better performance as measured by the metrics of accuracy (ACC), normalized mutual information (NMI), and purity (PUR). Based on this research, we conclude that the application of the proposed similarity calculation index is conducive to the improvement of data clustering for multidimensional data.*

Subject Categories and Descriptors
K.2.8 [Database Applications]: Data clustering; **B.2.4 [High-Speed Arithmetic]:** Algorithms

General Terms
Data clustering, Algorithm

Keywords: Data clustering, data clustering algorithm, information divergence, gradient angle, multidimensional data, kernel problem, similarity calculation

Received: 12 March 2015, Revised 2 June 2016, Accepted 12 June 2016

1. Introduction

In recent years, clustering algorithms have attracted an increasing amount of attention. These algorithms have been used widely in data mining, pattern recognition, image processing, parallel computing, and other fields [1-4]. To enable data clustering, first the similarity between various data points must be calculated. With the development of technology, clustering problems now essentially deal with “multidimensional” clustering in which the data structure is highly complex. Most traditional clustering approaches do not work well for complex data because they pay attention only to the global features [5]. For this reason, a new algorithm is needed that takes into consideration both global and local similarity. Such an algorithm would help distinguish data more comprehensively and improve clustering accuracy of multidimensional data.

2. State of the Art

K-means [6] and K-medoids [7] are among the traditional clustering algorithms that are used frequently. The K-means algorithm is based mainly on the Euclidean distance criterion. While this metric is effective when used with certain types of data, the K-means algorithm chooses the initial cluster center randomly, which can influence the clustering result greatly. The K-medoids algorithm takes the data at the center of the class as the cluster center, which allows this method to handle isolated data effectively. However, K-medoids also uses an initial cluster center chosen randomly, which can impact the clustering result as well. Although some improvements [8–9] have been adopted to optimize these classical clustering algorithms, the overall effect has not been optimal.

Currently, new clustering algorithms are beginning to attract more attention. Xuesong Yin [10] developed a novel semi-supervised, metric-based fuzzy clustering algorithm called SMUC by introducing metric learning and entropy regularization simultaneously into the conventional fuzzy clustering method (FCM). SMUC introduces maximum entropy as a regularized term in its objective function. The resulting formulas have clear physical meaning compared with the other semi-supervised FCMs. Experiments on real-world data sets have shown the feasibility and effectiveness of SMUC with encouraging results.

H.Venkateswara Reddy [11] proposed a method for data labeling using Relative Rough Entropy (RRE) for clustering categorical data. In this method, data labeling was performed by integrating entropy with rough sets. The experimental results showed that the efficiency and clustering quality of this algorithm are better than demonstrated by previous algorithms. Lei Zhang [12] proposed a novel ant-based clustering algorithm using Renyi Entropy (NAC-RE). The theoretical analysis conducted by the kernel method showed that the Renyi entropy metric was feasible and superior to a distance metric, demonstrating that NAC-RE can get good results.

M. Zarinbal [13], concerned with uncertainties, proposed the Interval Type-2 Relative Entropy Fuzzy C-Means (IT2 REFCM) clustering method. The study's results demonstrated that this proposed method has a strong ability to detect noise and assign suitable membership degrees to observations. M. Zarinbal [14] also proposed a novel collaborative fuzzy clustering method. In this approach, relative entropy is used as the communication method, which prompts it to have the highest quality of collaboration and enables it to classify data more efficiently. Thus, the concept of entropy introduced by Shannon with particular reference to information theory proved to be a powerful mechanism for the measurement of uncertainty information and effectively improved the clustering result. This type of method, however, pays more attention to the global similarity between various data with much less focus on local similarity, thereby causing the method to fail for some data sets.

To address this limitation, Niladri Shekhar Mishra [15] studied how fuzzy clustering algorithms incorporate local information to enhance the performance of the algorithms. Tests showed that the proposed technique is less time consuming and does not require any a priori knowledge of distributions of changed and unchanged pixels. Based on a non-Euclidean metric, Xiao-bin Zhi [16] presented a robust local feature weighting hard c-means (RLWHCM) clustering algorithm. The experimental results indicated that this algorithm has a high level of effectiveness. Ruochen Liu [17] proposed a dynamic local search-based immune automatic clustering algorithm (DLSIAC) to automatically evolve the number of clusters as well as a proper partition of data sets. Experimental results showed that this algorithm provides good image segmentation, demonstrating that local information can improve the performance of the original algorithm effectively.

Based on the analysis above, our study proposed a new clustering algorithm that considers both global and local features. Inspired by the entropy theory, first we calculated the information divergence (ID) between the two groups of data to reflect the global feature. Our second step was to characterize the local feature using the tangent of the generalized gradient angle (GGZ). At the same time, we proposed a conversion to change the scope of the angle, which contributes to the algorithm's ability to find subtle differences between the two groups of data. A similarity index was then developed by combining the information divergence and the generalized gradient angle. This index can reflect both the global and local similarity and can be used for data clustering. Finally, two experiments were conducted to show the performance of this algorithm. The remainder of this paper provides details of our study. Section 3 describes the information divergence and cosine similarity, and explains how to use them to calculate the similarity of multidimensional data. Section 4 introduces the evaluation metrics and discusses the performance of the proposed clustering algorithm through some case studies. Section 5 summarizes our conclusions.

3. Methodology

Since the proposed data clustering algorithm is based on our calculation of similarity, we will introduce information divergence and cosine similarity briefly.

3.1 Information divergence

Information divergence is an important measurement in information theory because it can reflect the differences between various sources. Information divergence has been used in various fields in recent years [18] mainly for information identification and fusion. For example, if there are two possible probability density functions $p_1(z)$ and $p_2(z)$ of the random variable z, then the information divergence of $p_1(z)$ to $p_2(z)$ is

$$D(p_1(z),p_2(z))= \int p_1(z) In \frac{p_1(z)}{p_2(z)} dz = E\left[In \frac{p_1(z)}{p_2(z)} \right] \quad (1)$$

Information divergence has the following properties

a) Nonnegative. $D(p_1(z), p_2(z)) \geq 0$ the equality holds if and only if $p_1(z)= p_2(z)$;

b) Asymmetry. $D(p_1(z), p_2(z)) \neq D(p_2(z), p_1(z))$.

Information divergence evaluates the similarity between two probability distributions on the side of the amplitude: the higher the similarity of probability density function, the smaller the information divergence. In order to overcome the asymmetry of the information divergence, some improved information divergences have been adopted. In this paper, the J information divergence was used, defined as follows

$$\begin{aligned} D_J(p_1(z), p_2(z)) &= D(p_1(z), p_2(z)) + D(p_2(z), p_1(z)) \\ &= \int p_1(z)\ln\frac{p_1(z)}{p_2(z)}dz + \int p_2(z)\ln\frac{p_2(z)}{p_1(z)}dz \\ &= \int \left(p_1(z) - p_2(z)\right)\ln\frac{p_2(z)}{p_1(z)}dz \end{aligned} \tag{2}$$

For discrete data, the J information divergence was calculated as follows

$$D_J(p_1(z), p_2(z)) = \sum_i \left[p_1(z(i)) - p_2(z(i))\right]\ln\frac{p_1(z(i))}{p_2(z(i))} \tag{3}$$

where $p(z_i) = z_i / \sum_i^n z_i$.

As seen from this definition, the information divergence can reflect the global similarity between two groups of multidimensional data effectively, but it lacks a description of the local information.

3.2 Cosine similarity

The main principle of cosine similarity is to take two groups of multidimensional data as a two-dimensional vector space, and then to characterize the degree of similarity by calculating their generalized angle. The smaller the angle, the higher the similarity [19]. If we suppose two groups of multidimensional data $x \in R^n$ and $y \in R^n$, respectively, then the inverse cosine form of their generalized angle is

$$S(x, y) = \arccos\frac{\langle x, y\rangle}{|x|\cdot|y|} = \arccos\left(\frac{\sum_{i=1}^n x_i \cdot y_i}{\sqrt{\sum_{i=1}^n x_i}\cdot\sqrt{\sum_{i=1}^n x_i}}\right) \tag{4}$$

The smaller the value of S, the closer the cosine value of the two groups of multidimensional data is to 1.

Obviously, the method mentioned above evaluates the similarity between two groups of multidimensional data from the vector geometry feature, which can reflect the local feature change of the data effectively.

3.3 Similarity calculation based on ID and GGZ

It can be seen from the descriptions of information divergence and cosine similarity that information divergence can reflect the global similarity between two groups of multidimensional data, while the generalized angle can characterize local features. By using a calculation that combines ID and GGZ, similarity can be described comprehensively, which should enable more effective clustering.

Although the generalized angle can characterize local features, it is not sufficient. Therefore, an improved form called the generalized gradient angle was adopted for this research. We first needed to perform a derivation of the two groups of multi- dimensional data, and then calculate the generalized gradient angle using the same method as the generalized angle. For the data offered above $x \in R^n$ and $y \in R^n$, the gradient vector was

$$SG(x) = \left(x_2 - x_1, x_3 - x_2, \cdots, x_n - x_{n-1}\right) \ , \tag{5}$$

$$\dot{S}G(y) = (y_2 - y_1, y_3 - y_2, \ldots y_n - y_{n-1}) \tag{6}$$

The generalized gradient angle was calculated as

$$\dot{S}G(x,y) = \arccos\frac{\langle SG(x), SG(y)\rangle}{|SG(x)|\cdot|SG(y)|} \tag{7}$$

The formula (7) showed that the generalized gradient angle could reflect the local features of the data in a manner especially sensitive to the changes of the geometric slope.

In order to design an effective similarity calculation index, we first calculated separately the information divergence and generalized gradient angle, and then performed a product operation for them. It should be noted that the information divergence cannot multiply the generalized gradient angle directly, so the tangent of the generalized gradient angle was calculated here. In theory, the scope of the generalized angle between the two groups of multidimensional data is $[0, \pi/2]$, and the scope of the generalized gradient angle is the same. Obviously, the tangent of the generalized gradient angle is a monotone increasing function. Its demarcation point is $\pi/4$, with the value increasing slowly when the angle smaller than $\pi/4$. Otherwise, it will increase quickly if the angle is larger than $\pi/4$. According to this characteristic, a conversion was used to change the scope of the angle to $[\pi/4, \pi/4]$. Thus, if the angle between two groups of multidimensional data changes slightly, the tangent value will change obviously, which is beneficial for finding the similarity changes between them. The final similarity calculation formula was

$$\dot{S}IM(x,y) = D(x,y) \times \tan\left[\frac{SG(x,y) + \pi/2}{2}\right] \tag{8}$$

3.4 Implementation of the proposed algorithm

For our proposed algorithm for multidimensional data clustering, our approach is to calculate first the information divergence and the generalized gradient angle, and then calculate the *SIM*. If *SIM* $\leq \theta$, the two groups of data belong to the same class; otherwise, they belong to two different classes separately. For SIM, θ is a threshold for which the initial value is determined by experience or some criterion and will gradually change in the algorithm running process. The implementation steps for the algorithm were as follows

Step 1: Initialization: Includes the data $X = \{x_1, x_2, \cdots, x_m\}$ to be clustered, where $x_j = \{x_{j1}, x_{j2}, \ldots x_{jn}\}$, the number N of the class, the cluster center c_i, $i = 1,2,\ldots,N$, and the threshold $\theta \in [1,\infty)$.

Step 2: Calculate the information divergence between the cluster center C_i and other data by the formula (3).

Step 3: Calculate the generalized gradient angle between the cluster center C_i and other data by the formula (7).

Step 4: Calculate the similarity between the cluster center C_i and other data by the formula (8).

Step 5: For all data, if $SIM(C_i, X_j) \le \theta$, $j = 1,2,\ldots,m$ holds, then the data x_j belongs to the class X_{ci}.

Step 6: Set $i = i+1$, $\theta = \theta + \Delta\theta$, if $i < N$, and jump to the step 2. Otherwise, the algorithm stops.

4. Analysis of Results and Discussion

In this section, we present two experiments: the first experiment used 2-dimensional Gaussian random data sets with added noise points, and the second used four well-known real-world data sets. The proposed ID-GGZ algorithm was applied in these experiments, and the results were then compared with four well-known clustering methods including K-means, K-medoids, fuzzy clustering method (FCM), and the RRE method proposed in [11].

4.1 Brief description of evaluation metrics

Three metrics were used to evaluate the clustering performance of all methods: accuracy (ACC), normalized mutual information (NMI), and purity (PUR).

ACC is the percentage of correctly predicted labels. If the real label of each cluster was unknown, the Hungarian algorithm [20] was used to get the best map to the real label.

If C denotes the ground truth label, and C' denotes the label obtained from a clustering algorithm, the mutual information *(MI)* is defined as

$$MI(C,C') = \sum_{c_i \in C_j' \in C'} p(c_i, C_j') \log \frac{p(c_i, C_j')}{p(c_i) \times p(c_j')} \qquad (9)$$

In the above formula, $p(c_i)$ and $p(c_j')$ are the probability of an arbitrarily selected sample belonging to cluster c_i and c_j', respectively. $p(c_i, C_j')$ is the probability of an arbitrarily selected sample belonging to both cluster c_i and c_j'.

NMI is the normalized MI as follows

$$NMI(C,C') = \frac{MI(C,C')}{\max(H(C), H(C'))} \qquad (10)$$

where $H(C)$ and $H(C')$ are the entropies of c and C' respectively.

PUR is computed by assigning the label of a cluster to the most frequent class. More formally, it is defined as

$$PUR(C,C') = 1 \sum_{j} \max_{i} (c_j' \cap c_i) \qquad (11)$$

4.2 Case study A: Toy datasets

This experiment involved three clusters of 100 points each with 100 noise points added to this noise-free data set. The first 100 points were generated based on the multivariate normal distribution with parameters μ_1=[1,2], $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0.1 \\ 0.1 & 2 \end{bmatrix}$. The parameters used for generating the second 100 points were μ_2=[-1,-2], $\Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$.

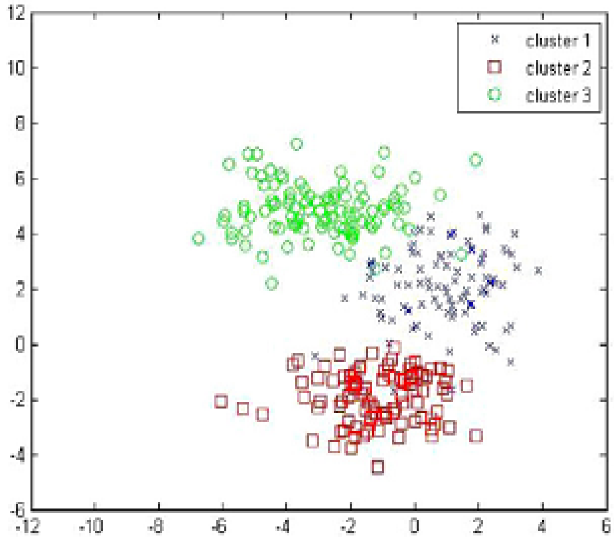
The third cluster's parameters were μ_3=[-3,4], $\Sigma_3 = \begin{bmatrix} 3 & 0.1 \\ 0.1 & 1 \end{bmatrix}$. These three clusters are demonstrated in Fig.1(a). The 100 noise points were then generated based on μ_N=[-3,5], $\boldsymbol{\Sigma}_N = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$. and were added to the noise-free data set (Fig.1(b)).

Using these five methods, we tried to cluster the data and eliminate the noise points from the data set. Note that a data point was considered as a noise point if it had a low membership degree in all clusters. The obtained new de-noised data sets based on these clustering methods were plotted in Fig.2(a)–(e), respectively. It is clear that among these methods, ID-GGZ was the most effective in clustering data and detecting noise points.

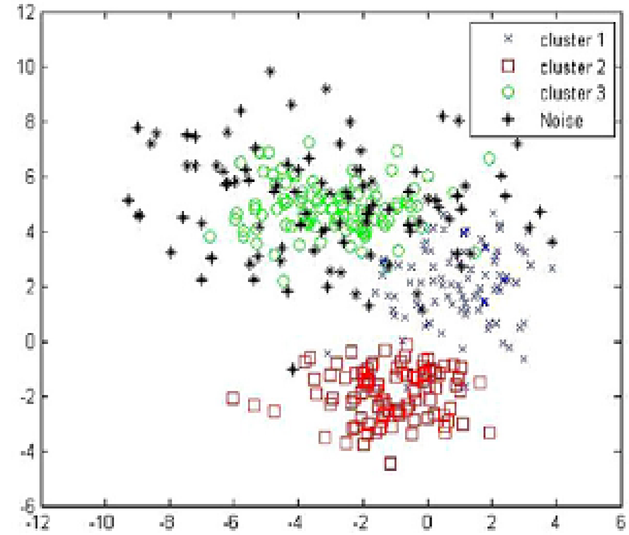
4.3 Case study B: Real-world data sets

In this experiment, four real-world data sets from the UCI database were used to validate the effectiveness of these methods. Important statistics for the data sets are summarized in Table 1. For each data set, we ran different methods 40 times, with the comparison based on the average performance.

Experiments were performed on the whole real-world data set. Tables 2, 3, and 4 show the clustering performance of the aforementioned methods evaluated by the ACC, NMI, and PUR measures. The mean and standard deviation values were both provided as percents. The bold digit in each line indicates that the corresponding best mean value was obtained from the corresponding method. The average running times (ART) of all the methods are

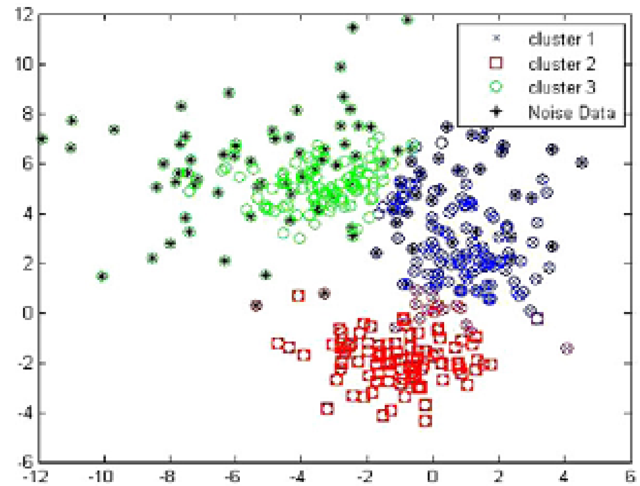


(a) without noise points

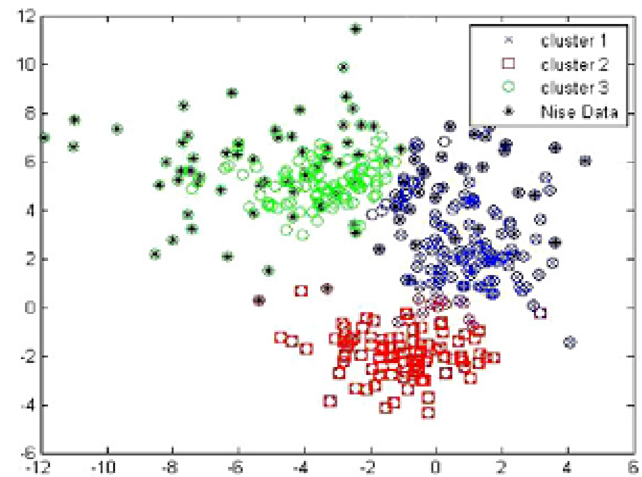


(b) with noise points

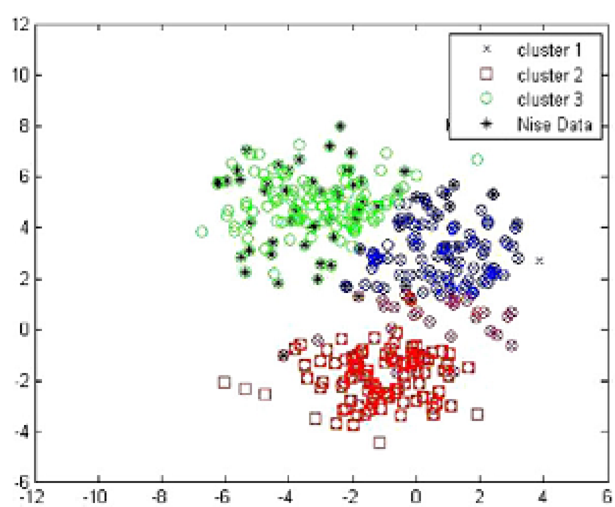
Figure 1. Two-dimensional data set.



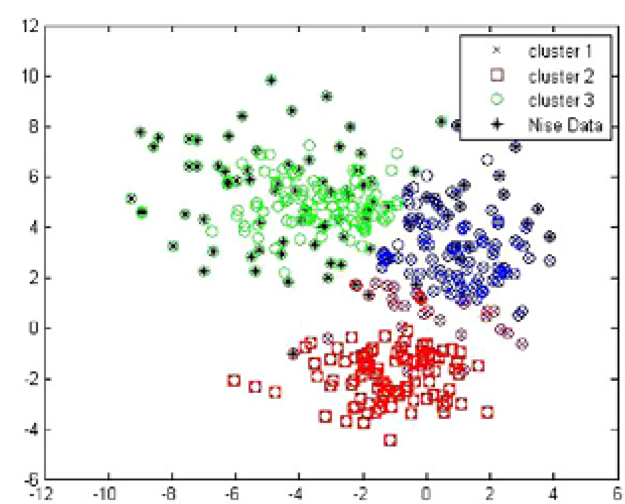
(a) K-means



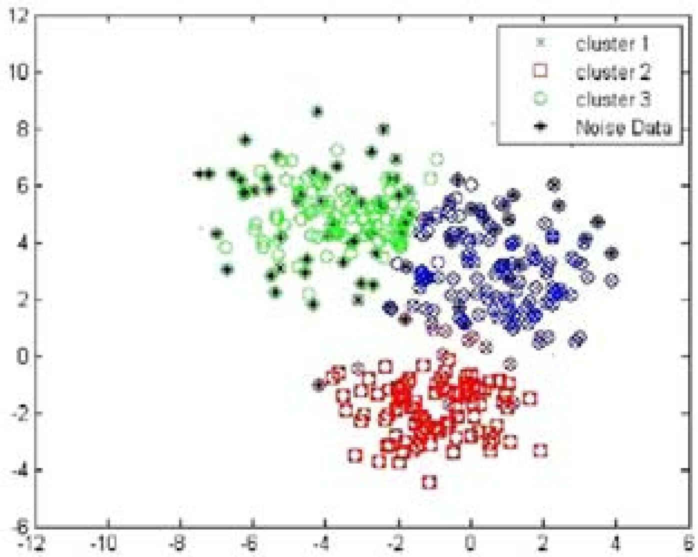
(b) K-medoids



(c) FCM



(d) RRE



(e) ID-GGZ

Figure 2. The clustering results

Data sets	sample	feature	cluster
Wine	178	13	3
Sonar	208	60	2
Vehicle	846	18	4
Pen digits	3498	16	10

Table 1. Data set descriptions

Data sets	K-means	K-medoids	FCM	RRE	ID-GGZ
Wine	56.47±3.73	58.75±4.24	78.46±2.32	89.28±1.27	91.17±1.14
Sonar	51.32±3.81	49.78±4.53	66.58±3.79	72.28±3.43	74.95±3.37
Vehicle	46.92±4.76	48.77±4.26	54.17±4.54	62.48±4.38	59.39±4.17
Pen digits	58.79±4.02	61.12±3.83	65.53±3.46	71.36±3.58	74.69±3.24

Table 2. Comparison of ACC (%)

Data sets	K-means	K-medoids	FCM	RRE	ID-GGZ
Wine	53.64±5.58	56.41±5.82	68.42±6.37	77.68±5.74	81.43±4.52
Sonar	58.61±6.51	56.35±6.63	60.81±7.44	66.92±6.25	68.75±5.69
Vehicle	43.53±6.58	44.83±6.84	49.84±7.72	57.76±5.23	57.67±6.30
Pen digits	42.56±9.65	44.64±9.11	58.69±9.03	63.61±8.55	66.95±7.43

Table 3. Comparison of NMI (%)

Data sets	K-means	K-medoids	FCM	RRE	ID-GGZ
Wine	64.76±4.63	65.55±4.74	71.81±3.42	78.27±3.73	82.09±3.94
Sonar	53.59±4.35	51.41±4.86	62.37±5.20	69.62±5.87	71.57±5.26
Vehicle	45.88±5.14	47.72±5.34	52.58±6.73	61.25±5.49	61.17±6.15
Pen digits	46.28±7.33	49.68±7.57	61.43±8.52	66.95±6.27	69.84±6.09

Table 4. Comparison of PUR (%)

Data sets	K-means	K-medoids	FCM	RRE	ID-GGZ
Wine	0.21	0.19	0.24	0.64	0.73
Sonar	0.28	0.26	0.32	0.79	0.92
Vehicle	1.69	1.57	1.88	2.86	3.14
Pen digits	5.07	5.14	5.26	9.29	12.47

Table 5. Comparison of ART (s)

reported in Table 5.

The results evaluated respectively by three different evaluation metrics are inconsistent on these data sets. For instance, according to the ACC metric, for the data set Vehicle the RRE method performs slightly better than ID-GGZ. However, according to the NMI and PUR metrics, the RRE and ID-GGZ methods perform comparably. This finding indicates that using more than one evaluation metric is important for evaluating the performance of a clustering algorithm.

We can see from Tables 2, 3 and 4 that ID-GGZ can give relatively high and consistent clustering accuracies when evaluated by ACC, NMI and PUR as compared with the other clustering methods. In addition, it can be seen from Table 5 that the ART of the proposed ID-GGZ method is much higher than the others. This higher running time mainly reflects the need for this method to calculate the information divergence and generalized gradient angle simultaneously. Nonetheless, the global level of the ART still satisfies the requirements of practical applications.

5. Conclusion

In this study, based on our analysis of information divergence and generalized gradient angle, we designed a new similarity index, which we used to establish a new data clustering algorithm. Our study employed Matlab language to compile the corresponding calculation program. Simulation analyses were conducted to evaluate the ACC, NMI, and PUR of the proposed algorithm. The effect of the new algorithm on ART was assessed as well. The main conclusions of this research are the following:

(1) The ID and GGZ were combined for the first time to calculate the similarity between various data. According to the results of the field dynamic test, this proposed method for calculating similarity is both accurate and effective. Local features are also reflected very well, demonstrating that the new index is more sensitive to data changes and its adjusted scope is larger than found in previous methods.

(2) Our new clustering algorithm for multidimensional data based on the proposed similarity index is shown to be simple and easy to use. It is well suited for multidimensional data clustering with good performance.

(3) The case studies considered in this research verified that the proposed data clustering algorithm can reflect both the global and local similarity of the multidimensional data. Compared with some other commonly used clustering algorithms, the average ACC, NMI, and PUR of this new algorithm was 75.82%, 68.72%, and 71.19% respectively. These results agree with the test conclusion that the clustering performance is much better than the commonly used algorithms.

This proposed similarity calculation index is highly suitable for the demand of multidimensional data clustering with high ACC. However, the ART of the proposed algorithm is poor. Moreover, real-world vagaries may cause uncertainties in the algorithm's parameters, making calculation of an exact degree of membership for each data point impossible. These findings suggest that improving the proposed method using fuzzy logic would be the focus of potential future work resulting from this study.

Acknowledgements

This study was supported by the Natural Science

Foundation of China (NSFC) (No. 11426158) and the research funding by Capital University of Economics and Business (CUEB)(No. 2014XJQ011).

References

[1] Tarel Elguebalya, Nizar Bouguila (2015). Model-based approach for high-dimensional non Gaussian clustering and feature weighting. *Digital Signal Processing,* 40(3) 63-79.

[2] Efendi Nasibov, Can Atilgan, Murat Ersen Berberler, et al (2015). Fuzzy joint points based clustering algorithms for large data sets. *Fuzzy Sets and Systems.* 270(7) 111-126.

[3] Yang Tang, Ryan P. Browne, Paul D. McNicholas (2015). Model based clustering of high-dimensional binary data. *Coputational Statistics and Data Analysis,* 87(7) 84-101.

[4] Huang Z (1998). Extensions to K-means algorithm for clustering large data sets with categorical values, *Data Ming and Knowledge Discovery,* 2(3) 283-304.

[5] Caiming Zhonga, Mikko Malinenb,Duoqian Miao (2015). A fast minimum spanning tree algorithm based on K-means. *Information Sciences,* 295(2) 1-17.

[6] Mohamed Walid Ayech, Djemel Ziou (2015). Segmentation of Terahertz imaging using k-means clustering based on ranked set sampling, *Expert systems with Applications,* 42(6) 2959-2974.

[7] Hae-Sang Park, Chi-Hyuck Jun (2009). A simple and fast algorithm for K-medoids clustering, *Expert Systems with Applications,* 36(2) 3336-3341.

[8] Abdolreza Hatamloua, Salwani Abdullahb, Hossein Nezamabadi-pour (2012). Abombined approach for clustering based on K-means and gravitational search algorithms, *Swarm and Evolutionary Computation,* 6(10) 47-52.

[9] B. Bemabe-Loranca, R. Gonzalez-Velazquez, E. Olivares-Benitez, et al (2014). Extensions to K-Medoids with Balance Restrictions over the Cardinality of thePartitions, *Journal of Applied Research and Technology,* 12(3) 396-408..

[10] Xuesong Yin, Ting Shu, Qi Huang (2012). Semi-supervised fuzzy clustering with metric learning and entropy regularization, *Knowledge-Based Systems,* 35(12) 304-311.

[11] H. Venkateswara Reddy, S. Viswanadha Raju, Pratibha Agrawal (2013). Data Labeling method based on Cluster Punity using Relative Rough Entropy for Categorical Data Clustering. In: *IEEE International Conference on Advances in Computing, Communications and Informatics* India: IEEE, pages 500-506.[12] Lei Zhang, Qixin Cao, Jay Lee (2013). A novel ant-based clustering algorithm using Renyi entropy, *Applied Soft Computing,* 13(5) 2643-2657.

[13] M.Zarinbal, M.H. Fazel Zarandi, I.B. Turksen (2014), Interval Type-2 Relative Entropy Fuzzy C-means Clustering, *Information Sciences,* 272(7) 49-72.

[14] M.Zarinbal, M.H. Fazel Zarandi, I.B. Turksen (2015), Relative Entropy collaborative fuzzy clustering method, *Pattern Recognition,* 48(3) 933-940.

[15] Niladri Shekhar Mishra, Susmita Ghosh, Ashish Ghosh (2012), Fuzzy clustering algorithms incorporating local information for change detection in remotely sensed images, *Applied Soft Computing,* 12(8) 2683-2692.

[16] Xiao-bin Zhi, Jiu-lun Fan, Feng Zhao (2014), Robust local feature weighting hard c-means clustering algorithm, *Neurocomputing,* 134(6) 20-29.

[17] Ruochen Liu, Beinbin Zhu, Reny Bian (2015). Dynamic local search based immune automatic clustering algorithm and its applications, *Applied Soft Computing,* 27(2) 250-268.

[18] Changki Lee, Gary Geunbae Lee (2006). Information gain and divergence based feature selection for machine learning-based text categorization, *Information Processing & Management,* 42(1) 155-165.

[19] Peipei Xia, Lia Zhang, Fanzhang Li (2015). Learning similarity with cosine similarity ensemble, *Information Sciences,* 307(6) 39-52.

[20] Kuhn, H.W (1955). The Hungarian method for the assignment problem, *Naval research logistics quarterly,* 2(1-2) 83-97.