Improved quantum clustering analysis based on the weighted distance and its application

Fan Decheng, Song Jon, Cholho Pang, Wang Dong, CholJin Won
School of Economics and Management, Harbin Engineering University, Harbin 150001, China
Department of Physics, University of Science, Pyongyang 950003, Democratic People’s Republic of Korea
Department of Material Engineering, Kimchaek University of Technology, Pyongyang 950003, Democratic People’s Republic of Korea
Agriculture and Life Science Department, Pyongyang University of Science and Technology, Pyongyang 950003, Democratic People’s Republic of Korea

Abstract

Cluster analysis is widely used in fields such as economics, management and engineering. The distance and correlation are two of the most important and often used mathematics- and statistics-based similarity measures in cluster analysis. Many studies have been conducted to improve the distance and similarity in high-dimensional and overlapped data. However, these studies do not consider the degree of influence (weight) of different properties on different types of data. In practice, the weight of each property is different, so these methods cannot accurately analyze real data. First, this study proposes a new distance measure that can reflect the weight, so that non-spherical overlapping data in the Euclidean space can be projected onto a weighted Euclidean space to form non-overlapping data. Second, the Fuzzy-ANP method is used to determine the weight of each factor. Then, by applying the Fuzzy-ANP-Weighted-Distance-QC (FAWQC) method to weighted random data, the effectiveness of the method is verified. Finally, the method is applied to the 2015 Economics-Energy-Environment (3E) data for 19 provinces in China for a comparative study of the classification of the system structure and evaluation of the low-carbon economy.
development level. The experiment results show that the FAWQC method can more accurately analyze real-world data than other methods.

Keyword: Economics

1. Introduction

Cluster analysis is an important technique in data mining with growing importance in the fields of machine learning, engineering, neural networks, biology, statistics, social sciences, and economics. Cluster analysis has been developing for several decades, and many clustering algorithms have been published. Traditional clustering algorithms can be divided into several types: partition-based clustering [1, 2], hierarchical clustering [3, 4, 5], density-based clustering [6, 7, 8], grid-based clustering [9, 10], and model-based clustering [11]. Different clustering algorithms have specific applications and shortcomings. For example, some clustering algorithms are more suitable for a particular type of data, while some algorithms may be more suitable for a particular type of special distribution and do not process data of other types of distribution well. Until now, there is no single general clustering algorithm that can be applied to all problems; almost all clustering algorithms have some shortcoming. For example, the partition-based k-means algorithm and fuzzy c-means (FCM) algorithms have a major shortcoming relating to the initial values (determining k, which is the number of clusters, and selecting the set of initial cluster centers).

To overcome the flaws and shortcomings of traditional clustering algorithms, methods such as quantum clustering [12], spectral clustering [13, 14], granularity clustering, probabilistic graph clustering, and synchronous clustering [15, 16, 17, 18, 19] have become popular in recent years. Quantum clustering uses the gradient descent method to solve for the minimum quantum potential at a fixed learning rate and determine the cluster center. Then, a fixed measuring standard is used to cluster the sample. However, quantum clustering has its flaws and shortcomings in terms of the execution time, measuring distance matrix, and learning iterations. To improve upon the shortcomings of QC algorithms, researchers have proposed new clustering methods. Li Zhi-Hua et al. [20] proposed a new distance-based quantum clustering (QC). The method not only has the advantages of the standard quantum clustering algorithms but also does not require sample preprocessing in most cases and has a fixed clustering distance, which significantly improves the efficiency. Casañà-Eslava et al. [21] proposed a method to find suitable length parameters when applying QC in non-spherical data. Li Yang-Yang et al. [22] proposed a new QC method using kernel component analysis.

This method preserves the advantages of the traditional QC, such as the ability to find clusters of any shape and subsequent determination of the cluster center using...
basic information from the data without a priori knowledge about the number of clusters. Moreover, the proposed algorithm is applicable to high-dimensional data sets, particularly to complex datasets. The focus of the above research is on improving the QC algorithm without considering the importance (weight) of each property. Thus, using these methods to analyze standard datasets from the UCI repository yields good accuracy. However, the performance suffers when they are applied to real data because in real data, the distribution of every property is different and complex, and the importance (weight) of every property also varies. For different clustering tasks of different contexts, the weight of each property should be reconsidered.

In this paper, first, we propose an evaluation method based on the Fuzzy-ANP called Fuzzy-ANP-Weighted-Distance-QC (FAWQC). Second, we apply the method to two-dimensional weighted random data and verify the effectiveness of the method. Then, the method is applied in the 2015 Economics-Energy-Environment (3E) data for 19 provinces in China for a comparative study of classification of system structures. Based on this study, the low-carbon economy development level of the provinces in the four categories is evaluated. By verifying the method on weighted random matrices and the 2015 3E data for 19 provinces in China, the effectiveness of the method is proven. The method more accurately reflects real-world situations than the traditional K-means and QC methods.

2. Theory

2.1. Principle of quantum clustering

Quantum clustering is an inversion problem of quantum mechanics. In quantum mechanics, a quantum system is determined in time and space according to the Schrodinger differential equation, which is described by a wave function. In contrast with quantum mechanics, the quantum clustering is the process in which the distribution of particles is estimated on the basis of their potential function when the wave function is given. Therefore, in quantum clustering, it is critical to determine the cluster centers by means of the potential function from the solution of Schrodinger equation (Eq. (1)).

$$H\psi(x) = \left(-\frac{\hbar^2}{2m}\nabla^2 + v(x)\right)\psi(x) = E\psi(x)$$  \hspace{1cm} (1)

where $H$, $E$, $\hbar$ and $m$ denote the Hamiltonian operator, the eigenvalue energy level, the Reduced Plank Constant and the mass of a particle, respectively. The function $\psi(x)$, a so-called wave function, describes the eigenstate of a given quantum system. The function $v(x)$, $\nabla^2$ are the potential function and Laplacian, respectively.
Therefore, the distribution of the particles is finally determined by the potential energy function.

In the quantum clustering (QC) algorithm that Horn and Gottlieb proposed [12], the wave function is given as a Gaussian Kernel shown in Eq. (2), and thus \( m = h^2/\sigma^2 \), where \( \sigma \) is the width parameter.

\[
\psi(x) = \sum_i e^{-\frac{(x-x_i)^2}{2\sigma^2}}
\]  
(2)

Then, \( \nu(x) \) is calculated by solving Eq. (1) in the following steps. Eq. (3) is the rewritten type of Eq. (1).

\[
H\psi(x) = \left( -\frac{\sigma^2}{2} \nabla^2 + \nu(x) \right) \psi(x) = E\psi(x)
\]  
(3)

Therefore, the \( \nu(x) \) can be solved as Eq. (4):

\[
\nu(x) = E + \frac{\sigma^2}{2} \nabla^2 \psi(x) \psi(x)
\]  
(4)

Furthermore, the first -order derivative of \( \psi(x) \) in the Eq. (1) is solved as Eq. (5):

\[
\psi(x)' = \sum_i e^{-\frac{(x-x_i)^2}{2\sigma^2}} - \frac{e^{(x-x_i)/\sigma^2}}{\sigma^2}
\]  
(5)

In addition, the second-order derivative of \( \psi(x) \) is described as Eq. (6).

\[
\psi(x)'' = \sum_i \left( e^{-\frac{(x-x_i)^2}{2\sigma^2}} \frac{(x-x_i)^2}{\sigma^4} - e^{-\frac{(x-x_i)^2}{2\sigma^2}} \frac{1}{\sigma^2} \right)
\]  
(6)

Thus, the \( \nu(x) \) can be solved as Eq. (7). Here, the eigenvalue energy \( E \), which is considered as a constant, has no any influence on the topological structure of \( \nu(x) \). Therefore, the \( \nu(x) \) is ultimately described as Eq. (8).

\[
\nu(x) = E + \frac{1}{2\sigma^2 \psi(x)} \sum_i (x-x_i)^2 e^{-\frac{(x-x_i)^2}{2\sigma^2}}
\]  
(7)

\[
\nu(x) \approx \frac{1}{2\sigma^2 \psi(x)} \sum_i (x-x_i)^2 e^{-\frac{(x-x_i)^2}{2\sigma^2}}
\]  
(8)

The algorithm uses gradient descent to solve for the minimum quantum potential at a fixed learning rate and determine the cluster center. Then, a fixed measuring standard is used to cluster the sample.
The advantages of the algorithm are as follows:

1. It is a partition-based clustering algorithm with unsupervised learning;
2. It has only one parameter.

The shortcomings of the algorithm are:

1. Under most circumstances, the sample requires preprocessing when clustering, which increases the computation time.
2. The measuring distance is relatively fixed, and different sample distribution features are not differentiated;
3. When one solves for the minima of the potential energy function using the gradient descent, the degree of learning cannot be easily known, which affects the clustering precision.

2.2. Weighted-distance-based quantum clustering algorithm

To overcome these shortcomings the distance-based quantum clustering method was proposed in [20]. In this method, first determined the parameter delta value according to [23] and calculate the potential to determine the center of the cluster. Next, we took the measuring distance ($\beta$) and proceeded to analyze the cluster. Here, the parameter $\beta$, a parameter to perform the cluster analysis after determining the cluster centers, can be determined through training process because it has no any specific decision method and has different values according to each data. The distances among all objects in the distance-based QC method (e.g., samples and cluster centers) are calculated using the following distance formula (Eq. (9)):

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ih} - x_{jh})^2}, \quad (i \neq j, 1 \leq i \leq m, 1 \leq j \leq m)$$

(9)

Although the algorithm has the advantages of the standard QC method, it overcomes the issue of requiring sample preprocessing in most cases and fixed clustering distances, which significantly improves the efficiency. However, it fails to consider the weights and characteristics of the indicator data of each property.

In response to different characteristics of different data categories, researchers have proposed many types of distance functions. Some well-known distance functions are Minkowski distance (Manhattan, Euclidean, Supremum), Canberra distance, Czeknowski distance, Standardized Euclidean distance and cosine distance. However, these distance functions fail to consider the importance of each property. To address these flaws and shortcomings, this paper proposes a distance function (Eq. (10)) that considers the characteristics of the data and the weight of each property, specifically as follows:
where the standard deviation $S$ reflects the characteristics of the data of each property, and weight $W$ reflects the importance of each property. The specific method to determine the weight is described in 2.3. This distance function satisfies the three conditions (non-negativity, symmetry, and subadditivity) so it can be used to calculate the distance between two samples.

Based on the distance formula, the proposed algorithm in this study is:

Step 1: Determine the parameter $\sigma$ and the weight of each attribute according to the characteristics of the data.

In this paper, the weight was determined by using the Fuzzy-ANP method. In addition, the parameter $\sigma$, a parameter to determine the number of cluster centers by calculating the potential in quantum cluster, depends on the dimensions of data, and the detailed description was specified in the literatures [23, 24].

Step 2: Initialize $\beta, c = 0$, where $c$ is the number of clusters, and $\beta$ is the scale of the measurement.

Step 3: Using Eq. (9), calculate the weighted measure to obtain the measuring matrix $D$.

Step 4: Estimate parameter $\sigma$ and calculate potential energy $V$ of the sample (identical to that for QC [23]).

Step 5: $c = c + 1$.

Step 6: According to potential energy $V$, find the minima $V_{\text{min}} = \min\{V_n\}$. Let $V(x_k) = V_{\text{min}}$, and $v_c = x_k$ as the $c$th cluster center.

Step 7: According to the calculation for $D$, cluster all samples that satisfy the distance measure $d(x_i, v_c) \leq \beta$ ($1 \leq i \leq n$ and $i \neq k$) in $C$ and remove these samples from the set of samples $X$.

Step 8: If $X$ is empty, the algorithm ends. Else, go to Step 5.

The above algorithm produces $C$ clusters, and the cluster center is represented by $v_i$.

2.3. Determining the weights based on Fuzzy-ANP

In the real world, most data are composed of high dimensions and present a certain degree of hierarchical structure. Within the hierarchy, there are interdependence and a feedback relationship among the properties. The importance of each property is also different. For example, when we establish the evaluation system for the low-carbon economic development level, there are an interdependence and a feedback relationship between each subsystem or between the properties in the subsystem.
The details are described in 3.2. Compared to the AHP method, the evaluation results using the ANP method are more credible and accurate. Using the traditional ANP method to assign weights to the indicators, the decision maker may encounter subjective uncertainty when quantitatively evaluating each indicator. The uncertainty may negatively affect the overall evaluation result. To eliminate this effect, the study introduces triangular fuzzy numbers to the analytical network process to determine the weight of each indicator.

2.3.1. Constructing a fuzzy judgment matrix based on triangular fuzzy numbers

To eliminate as much as possible the negative effect of the expert subjective uncertainty in the process of comparing elements, this paper introduces the triangular fuzzy number judgment method, which calculates the fuzzy comprehensive value from the triangular fuzzy number given by T experts. Set \( X = \{x_1, x_2, ..., x_n\} \) as an object set and \( U = \{u_1, u_2, ..., u_m\} \) as a target set. Then, the degree by which the \( i \)th object satisfies \( m \) targets is Eq. (11):

\[
S_i = \sum_{j=1}^{m} M_{ei}^j \otimes \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} M_{ei}^j \right]^{-1}
\]  

(11)

where \( M_{ei}^j \) is described as the Eq. (12)

\[
M_{ei}^j = \frac{1}{T} \otimes \sum_{k=1}^{T} a_{ei}^k
\]  

(12)

\( a_{ei}^k = [e_{ei}^k, m_{ei}^k, u_{ei}^k] \) is the triangular fuzzy number given by the \( k \)th expert, \( k = 1, 2, ..., T \).

Then, the weight vector is calculated based on the triangular fuzzy number.

The probability that fuzzy number \( M \) is greater than \( k \) fuzzy numbers (Eq. (13)) is defined as Eq. (14):

\[
V(M \geq M_1, M_2, ..., M_k) = \min V(M \geq M_i), i = 1, 2, ..., k
\]  

(13)

\[ \text{let } d'(A_i) = \min V(S_i \geq S_k) \]  

(14)

Therefore, the weight vector can be obtained as \( W' = (d'(A_1), d'(A_2), ..., d'(A_n))T \), which can be simplified as Eq. (15):

\[
W = (d(A_1), d(A_2), ..., d(A_n))T
\]  

(15)
2.3.2. Computing the ANP super matrix

To determine the effect of an element on its standard, the weight vector of the obtained elements must be combined to construct a supermatrix. In the ANP network structure, the control layer has elements $B_1, B_2, \ldots, B_m$, whereas the network layer has elements $C_1, C_2, \ldots, C_m$, where $C_i$ contains elements $e_{i1}, e_{i2}, \ldots, e_{in}, i = 1, 2, \ldots, n$. With $B_s$ as the criterion and $e_{jl}$ ($l = 1, 2, \ldots, nj$) as the sub-criterion, a comparative analysis is done on the effect of the other elements in set $C_i$ on $e_{jl}$ to construct the comparison matrix $W_{ij}$ (Eq. (16)) in element $B_s$, and we separately calculate its weight vector.

$$W_{ij} = \begin{pmatrix} W_{i1j} & W_{i2j} & \cdots & W_{inj} \\ W_{i2j} & W_{i2j} & \cdots & W_{inj} \\ \vdots & \vdots & \ddots & \vdots \\ W_{inj} & W_{inj} & \cdots & W_{inj} \end{pmatrix} \quad (16)$$

The form of the supermatrix of the influence of each element on element set $C_i$ (Eq. (17)) under the above criterion is as follows:

$$W = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1m} \\ W_{21} & W_{22} & \cdots & W_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ W_{m1} & W_{m2} & \cdots & W_{mm} \end{pmatrix} \quad (17)$$

With $B_s$ as the criterion, the importance of every set of elements under $B_s$ to criterion $C_i$ ($i = 1, 2, \ldots, m$) is compared. This yields the weighted matrix as Eq. (18):

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix} \quad (18)$$

The weighted matrix $A$ is multiplied by the unweighted supermatrix $W$ to obtain a weighted supermatrix $\overline{W}$ of $W$, where $= A \ast W$. Then, the steady-state supermatrix $\overline{W}^\infty$ can be calculated after the $n$th iteration, where the row vector is the stable weight of each element in the network layer under criterion $B_s$.

3. Experimental

To verify the validity of the method, the paper first takes the mathematical approach and verifies the method on a theoretical two-dimensional weight random data. Then, the method is verified using real data. The method was applied to the 2015 low-carbon economy data from China.
3.1. Verifying on weighted random data

To verify the validity of the algorithm in theory, a cluster analysis was performed on two-dimensional random data. A comparative analysis was performed on the results of the QC algorithm and the traditional k-means method. First, three sets of weighted random test data were generated. Each set of data included 100 random two-dimensional data points of 0–6. When we generated the data, the weights of x and y were 0.8 and 0.2, respectively. The traditional k-means method, QC method, and FAWQC method were used for the cluster analysis. The results are shown in Fig. 1 and Table 1 ($\sigma = 0.7071$, $\beta = 1.3501$).

Fig. 1(a) shows the distribution of the two-dimensional weighted sample in Euclidean space. Overlapping evidently occurs, and it is difficult to determine to which cluster the data belongs during the cluster analysis. When viewed in Euclidean space, they appear to belong to two clusters when they actually belong to one cluster. Fig. 1(b) shows the cluster analysis results when the 2-dimensional sample is projected on the weighted Euclidean space. The overlapping part has evidently been separated, and the method yields better accuracy than the traditional

Fig. 1. Comparison of the cluster analysis results with two-dimensional weighted random data. (a) Distribution of two-dimensional weight random data; (b) FAWQC analysis result in weighted Euclidean space; (c) k-means analysis result in Euclidean space; (d) QC analysis result in Euclidean space.
k-means method and QC method. Fig. 1(c, d) show the cluster analysis results with the k-means method and QC method in Euclidean space, respectively. The overlapped portions cannot be distinguished, which causes a less accurate cluster analysis. The experiment results demonstrate that the method is more accurate than traditional cluster analysis methods that do not consider weights.

3.2. Verification using low-carbon data

To verify the efficacy of the method on real-world data, the study uses the method to evaluate the low-carbon economic development level of provinces in China. To accurately reflect the actual situation, we used a low-carbon development level evaluation indicator system for the 3E system based on past literature and expert opinions. The system includes 3 sub-systems and 16 indicators, as shown in Table 2. The ANP network structure in the study is composed of a control layer and a network layer, as shown in Fig. 2. The control layer is divided into two parts: target and decision criteria. The network layer is composed of elements under each decision criterion. Between each element and collection elements, there is an interdependence and feedback relationship, which forms a complex network structure. Using an expert scoring system on a 1—9 scale, the weight and fuzzy relationship matrices of the primary and secondary indicators were determined. Then, according to the Fuzzy-ANP method in section 2.2, the weight of each indicator was calculated, as shown in Table 3. The weight matrix is: $W = \begin{bmatrix} 0.037 & 0.0775 & 0.0362 & 0.1195 \\ 0.0329 & 0.021 & 0.0632 & 0.0849 \\ 0.0329 & 0.021 & 0.0632 & 0.0849 \\ 0.1195 & 0.0632 & 0.0849 & 0.0463 \\ 0.0329 & 0.021 & 0.0632 & 0.0849 \\ 0.0632 & 0.0849 & 0.0463 & 0.0287 \\ 0.0329 & 0.021 & 0.0632 & 0.0849 \\ 0.1195 & 0.0632 & 0.0849 & 0.0463 \\ 0.0329 & 0.021 & 0.0632 & 0.0849 \\ 0.0632 & 0.0849 & 0.0463 & 0.0287 \\ 0.0329 & 0.021 & 0.0632 & 0.0849 \\ 0.1195 & 0.0632 & 0.0849 & 0.0463 \\ 0.0329 & 0.021 & 0.0632 & 0.0849 \\ 0.0632 & 0.0849 & 0.0463 & 0.0287 \\ 0.0329 & 0.021 & 0.0632 & 0.0849 \\ 0.1195 & 0.0632 & 0.0849 & 0.0463 \end{bmatrix}$.

The data for the 16 indicators are from the 2016 China Statistical Yearbook and Statistical Yearbook of each province. Since the data are incomplete for some provinces in the statistical yearbooks, we only evaluated and analyzed the data for 19 Chinese provinces. Table 4 shows the 16 original data for the 19 provinces.

In general, there are many evaluation indicators in the evaluation indicator system of the 3E system with different properties. Therefore, the threshold method was used to make the indicator data dimensionless. The following formula (Eq. (19)) was used to standardize the positive indicators:

### Table 1. Clustering results for three cluster analysis methods.

| Method  | Number of samples | Correct | Incorrect | Accuracy (%) |
|---------|------------------|---------|-----------|--------------|
| K-means | 300              | 246     | 52        | 82           |
| QC      | 300              | 262     | 38        | 87.3         |
| FAWQC   | 300              | 286     | 14        | 95.3         |
Table 2. Low-carbon economy development level evaluation indicator system.

| Primary indicator | Secondary indicator | Tertiary indicator | Name (Unit) | Description | Indicator sign |
|-------------------|---------------------|--------------------|-------------|-------------|---------------|
| Socioeconomic subsystem (B₁) | GDP growth (%) (C₁₁) | Tertiary industry product/GDP | + |
| Industry structure (%) (C₁₂) | + |
| Employment rate (%) (C₁₃) | + |
| GDP per capita (10,000 Yuan) \((C₁₄)\) | + |
| Urbanization (%) (C₁₅) | City population/total population | + |
| Energy subsystem (B₂) | Total energy consumption \((10^7 \text{ tce}) (C₂₁)\) | - |
| Energy efficiency (C₂₂) | GDP growth/energy consumption growth | + |
| % Renewable (%) (C₂₃) | Renewable energy/total energy consumption | + |
| % Fossil fuels (%) (C₂₄) | Fossil fuel energy/total energy consumption | - |
| Environmental subsystem (B₃) | Energy consumption per capita \((\text{Ton/Person}) (C₂₅)\) | Total energy consumption/total population | + |
| Carbon emission per capita \((\text{Ton/Person}) (C₃₁)\) | Total carbon emission/total population | - |
| Carbon intensity \((\text{Ton/Yuan}) (C₃₂)\) | Carbon dioxide emission per unit GDP | - |
| Forestation (%) (C₃₃) | Green area/total area | + |
| Pollution control cost \((100 \text{ million Yuan}) (C₃₄)\) | - |
| Carbon dioxide emission \((\text{Ten thousand Yuan}) (C₃₅)\) | - |
| Soot emission \((\text{Ten thousand Ton}) (C₃₆)\) | - |

\[ U(x_i) = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad (19) \]

Where \(U(x_i)\) is the standardized data; \(x_i\) is the raw data; \(x_{\text{max}}\) is the data maximum; \(x_{\text{min}}\) is the data minimum. The following formula (Eq. (20)) was used to standardize the negative indicators:

\[ U(x_i) = \frac{x_{\text{max}} - x_i}{x_{\text{max}} - x_{\text{min}}} \quad (20) \]

Matlab was used to perform the cluster analysis. The clustering result divides the 19 provinces into four categories, as shown in Table 5\((\sigma = 0.9372, \beta = 5.4703)\). Here, the close-value method [25] was used to evaluate the development degrees of 4 clusters produced after the cluster analysis. The close-value method is an excellent method of multi-objective decision making in the field of system engineering. The detail content is as follows:
Fig. 2. Structural model of the ANP network of the low-carbon economy development level evaluation system.

Table 3. Weight of each indicator in the low-carbon economy development level evaluation system.

| Primary indicator                                      | Secondary indicator | Tertiary indicator |
|--------------------------------------------------------|---------------------|--------------------|
| Low-carbon economy 3E system coordination evaluation (A) | Socioeconomic subsystem (B₁) | GDP growth (C₁₁) 0.1221 0.0370 |
|                                                        | Industry structure (C₁₂) 0.2557 0.0775 |
|                                                        | Employment rate (C₁₃) 0.1194 0.0362 |
|                                                        | GDP per capita (C₁₄) 0.3942 0.1195 |
|                                                        | Urbanization (C₁₅) 0.1086 0.0329 |
|                                                        | Energy subsystem (B₂) | Total energy consumption (C₂₁) 0.0860 0.0210 |
|                                                        | Energy efficiency (C₂₂) 0.2589 0.0632 |
|                                                        | Renewable (C₂₃) 0.3478 0.0849 |
|                                                        | Fossil fuel (C₂₄) 0.1897 0.0463 |
|                                                        | Energy consumption per capita (C₂₅) 0.1176 0.0287 |
|                                                        | Environmental subsystem (B₃) | Carbon emission per capita (C₃₁) 0.1654 0.0794 |
|                                                        | Carbon intensity (C₃₂) 0.3450 0.1562 |
|                                                        | Forestation (C₃₃) 0.1513 0.0685 |
|                                                        | Pollution control cost (C₃₄) 0.1378 0.0624 |
|                                                        | Carbon dioxide emission (C₃₅) 0.1318 0.0597 |
|                                                        | Soot emission (C₃₆) 0.0687 0.0311 |
1) Construct an initial decision matrix of original data:

\[
A = \begin{pmatrix}
a_{11} & \cdots & a_{1n} \\
\vdots & \ddots & \vdots \\
a_{m1} & \cdots & a_{mn}
\end{pmatrix}
\]

where \( m \) and \( n \) denote the numbers of data and attributes, respectively.

2) Construct the normalization of initial decision matrix:

\[
R = (r_{ij})_{m \times n}
\]

\[
r_{ij} = \frac{a_{ij}}{\left( \sum_{i=1}^{m} a_{ij}^2 \right)^{1/2}} \quad j \text{ is the positive direction decision index}
\]

\[
r_{ij} = -\frac{a_{ij}}{\left( \sum_{i=1}^{m} a_{ij}^2 \right)^{1/2}} \quad j \text{ is the negative direction decision index}
\]

3) Determine the optimal or pessimal decision sample can be searched from \( R \)

The optimal decision sample:

\[
A^+ = (r_1^+, r_2^+, \cdots, r_n^+)
\]

The pessimal decision sample:

\[
A^- = (r_1^-, r_2^-, \cdots, r_n^-)
\]

Here, \( r_j^+ = \max_{1 \leq i \leq m} \{ r_{ij} \} \quad r_j^- = \min_{1 \leq i \leq m} \{ r_{ij} \} \quad j = 1, 2, \cdots, n \)

4) Calculate \( d_i^+ \) and \( d_i^- \), the weighted-Euclidean distance of each decision sample to the optimal decision sample \( A^+ \) and the pessimal decision sample \( A^- \).

\[
d_i^+ = \left[ \sum_{j=1}^{n} \left( w_j r_{ij} - w_j r_j^+ \right)^2 \right]^{1/2}
\]

\[
d_i^- = \left[ \sum_{j=1}^{n} \left( w_j r_{ij} - w_j r_j^- \right)^2 \right]^{1/2}
\]

where \( w_j \) is the weight of each attribute.
Table 4. Original data.

|                | GDP growth | Industrialization rate | Employment rate | GDP per capita | Urbanization | Total energy consumption | Energy efficiency | Fossil fuel | Renewables | CO₂ consumption per capita | Energy consumption per capita | Carbon emission per capita | Carbon intensity | Forestation | Pollution | CO₂ | Soot |
|----------------|------------|------------------------|-----------------|----------------|--------------|--------------------------|------------------|------------|------------|----------------------------|-----------------------------|---------------------------|-------------------|------------|----------|-----|------|
| Liaoning       | 3.131      | 0.462                  | 96.6 65354.0    | 67.35          | 20.522       | 1396.982                 | 95.800           | 4.200      | 0.005      | 0.883                       | 0.130                       | 38.240                    | 44.909            | 0.023      | 0.024    |      |      |
| Jilin          | 2.051      | 0.388                  | 96.5 51086.0    | 48.40          | 8.028        | 1751.826                 | 90.800           | 9.200      | 0.003      | 0.520                       | 0.098                       | 40.380                    | 45.529            | 0.014      | 0.017    |      |      |
| Heilongjiang   | 2.007      | 0.507                  | 95.5 39462.0    | 58.80          | 12.126       | 1243.893                 | 96.424           | 3.576      | 0.003      | 5.868                       | 1.483                       | 48.300                    | 50.733            | 0.007      | 0.017    |      |      |
| Shanxi         | 1.853      | 0.532                  | 96.5 34919.0    | 55.03          | 72.489       | 176.116                  | 99.400           | 0.600      | 0.020      | 3.736                       | 1.072                       | 18.030                    | 103.364           | 0.031      | 0.040    |      |      |
| Inner Mongolia | 3.174      | 0.380                  | 96.3 78000.0    | 60.30          | 21.556       | 858.273                  | 93.200           | 6.800      | 0.009      | 1.514                       | 0.206                       | 21.030                    | 174.146           | 0.049      | 0.035    |      |      |
| Jiangsu        | 4.273      | 0.486                  | 97.0 87995.0    | 66.52          | 30.235       | 2319.024                 | 94.738           | 5.262      | 0.004      | 0.682                       | 0.078                       | 15.800                    | 77.949            | 0.010      | 0.008    |      |      |
| Zhejiang       | 12.994     | 0.498                  | 97.0 77644.0    | 65.80          | 19.610       | 2186.970                 | 98.000           | 2.000      | 0.012      | 2.223                       | 0.085                       | 60.960                    | 356.801           | 0.033      | 0.019    |      |      |
| Anhui          | 1.344      | 0.380                  | 97.0 38000.0    | 50.50          | 13.000       | 1658.119                 | 99.432           | 0.568      | 0.002      | 0.403                       | 0.114                       | 28.670                    | 29.418            | 0.007      | 0.009    |      |      |
| Fujian         | 2.812      | 0.416                  | 96.5 67966.0    | 62.60          | 12.180       | 2132.995                 | 80.100           | 19.900     | 0.003      | 0.483                       | 0.071                       | 65.950                    | 90.365            | 0.009      | 0.009    |      |      |
| Jiangxi        | 1.436      | 0.391                  | 96.6 36724.0    | 51.62          | 8.440        | 1981.420                 | 86.800           | 13.200     | 0.002      | 0.306                       | 0.083                       | 60.010                    | 32.463            | 0.012      | 0.011    |      |      |
| Shandong       | 2.898      | 0.453                  | 96.6 64168.0    | 57.01          | 36.759       | 1713.920                 | 95.509           | 4.491      | 0.004      | 0.677                       | 0.106                       | 16.730                    | 96.063            | 0.015      | 0.011    |      |      |
| Henan          | 1.427      | 0.346                  | 97.1 40000.0    | 46.85          | 23.000       | 1695.674                 | 94.200           | 5.800      | 0.002      | 0.435                       | 0.106                       | 21.500                    | 34.899            | 0.012      | 0.009    |      |      |
| Hubei          | 2.177      | 0.431                  | 97.3 8132.72    | 56.85          | 13.828       | 2137.018                 | 86.000           | 14.000     | 0.002      | 0.386                       | 0.076                       | 38.400                    | 26.998            | 0.009      | 0.008    |      |      |
| Hunan          | 1.764      | 0.442                  | 95.9 42754.0    | 50.89          | 15.469       | 1868.443                 | 78.270           | 21.730     | 0.002      | 0.318                       | 0.080                       | 47.770                    | 27.852            | 0.008      | 0.006    |      |      |
| Guangdong      | 3.442      | 0.504                  | 97.5 67555.0    | 68.71          | 26.333       | 2831.224                 | 78.000           | 22.000     | 0.002      | 0.358                       | 0.052                       | 51.260                    | 31.815            | 0.006      | 0.003    |      |      |
| Guangxi        | 1.359      | 0.388                  | 97.0 35190.0    | 47.06          | 9.761        | 1721.516                 | 64.000           | 36.000     | 0.002      | 0.247                       | 0.071                       | 56.510                    | 51.533            | 0.009      | 0.007    |      |      |
| Sichuan        | 1.600      | 0.437                  | 95.9 36775.0    | 47.69          | 17.680       | 1699.826                 | 84.900           | 15.100     | 0.002      | 0.348                       | 0.095                       | 35.220                    | 14.415            | 0.009      | 0.005    |      |      |
| Yunnan         | 1.296      | 0.451                  | 96.0 28806.0    | 43.33          | 10.357       | 1315.028                 | 57.200           | 42.800     | 0.002      | 0.237                       | 0.083                       | 50.030                    | 45.527            | 0.012      | 0.007    |      |      |
| Shaanxi        | 2.643      | 0.548                  | 96.6 47626.0    | 53.92          | 11.716       | 1538.246                 | 95.450           | 4.550      | 0.003      | 0.568                       | 0.118                       | 41.420                    | 74.864            | 0.020      | 0.016    |      |      |
5) Calculate the 'close value' of each evaluation sample, and arrange the order of good and bad.

Close value:

\[ c_i = \frac{d^+_i - d^-_i}{d^+ - d^-} \]

where \( d^+_i = \min_{1 \leq i \leq m} \{d^+_i\} \), \( d^- = \max_{1 \leq i \leq m} \{d^-_i\} \)

The smaller the close value, the higher the quality. When \( c_i = 0 \), the best quality and the best point. Using the osculating value method, the osculating value \( C \) was obtained: \( C = (0.3111, 0, 0.1978, 0.7028) \). We ranked the low-carbon economy development level from good to bad based on the osculating value calculations as follows: \( 2 > 3 > 1 > 4 \). To more intuitively display the regional characteristics of the clustering result, the region distribution diagram is shown in Fig. 3. The light pink sections are regions that were not analyzed because of the lack of data from the 2016 China Statistical Yearbook and Statistical Yearbook of each province.

In order to prove the objectivity of the weight-distance quantum clustering experiment results, the traditional k-means algorithm and QC algorithm and the close value method are applied to cluster analysis in 19 provinces. The results of the k-means method are shown in Table 6. The low carbon economy development level by category is \( 2 > 4 > 1 > 3 \). The results of the QC method are shown in Table 7. The low carbon economy development level by category is \( 2 > 3 > 1 > 4 \).

The best areas for low carbon economic development achieved by the three methods are Jiangsu and Guangdong. Using the k-means method, the third category includes Fujian, Guangxi, and Yunnan, while the fourth category includes Inner Mongolia and Shandong. Using the QC method, the first category includes Shandong and Inner Mongolia and Shanxi, while the second category includes Zhejiang, Fujian and Heilongjiang.

Liaoning, Zhejiang, Fujian, Guangdong, Jiangsu, and Shandong (the highest areas, group A) are mainly located in the East and South coast areas of China, in which economic and environmental indicators are higher than ones in other regions (see

---

Table 5. Clustering result (weighted-distance quantum clustering).

| Category | Province | Osculating Value |
|----------|----------|------------------|
| 1        | Guangxi, Yunnan | 0.3111 |
| 2        | Liaoning, Jiangsu, Zhejiang, Fujian, Shandong, Guangdong | 0 |
| 3        | Jilin, Heilongjiang, Anhui, Jiangxi, Henan, Hubei, Hunan, Sichuan, Shaanxi | 0.1978 |
| 4        | Shanxi, Inner Mongolia | 0.7028 |
Fig. 3. Cluster analysis distribution.

Table 6. Clustering Result (k-means algorithm).

| Category | Province                                      | Osculating Value |
|----------|-----------------------------------------------|-------------------|
| 1        | Heilongjiang, Shanxi, Henan, Hubei            | 0.7966            |
| 2        | Jiangsu, Zhejiang, Guangdong                  | 0                 |
| 3        | Jilin, Anhui, Jiangxi, Fujian, Hunan, Sichuan, Guangxi, Yunnan | 1.0723          |
| 4        | Liaoning, Inner Mongolia, Shandong, Shaanxi  | 0.7930            |

Table 7. Clustering Result (QC algorithm).

| Category | Province                                      | Osculating Value |
|----------|-----------------------------------------------|-------------------|
| 1        | Shanxi, Jilin, Sichuan, Shandong, Anhui, Inner Mongolia, Shaanxi, Hubei | 0.8788            |
| 2        | Jiangsu, Guangdong                           | 0                 |
| 3        | Henan, Heilongjiang, Jiangxi, Hunan, Zhejiang, Fujian, Liaoning | 0.2978          |
| 4        | Guangxi, Yunnan                              | 1.1643            |
The reason is that the third industry, the technical capacity and the foundation of industry and agriculture are superior due to the geographical advantages of these areas. In addition, and Pollution energy consumption is relatively small (see Table 4). Secondly, Heilongjiang, Anhui, Jiangxi, Henan, Hubei, Hunan, Sichuan, and Shaanxi (relatively high areas, group B) are mainly located in the inland area. These areas have the advantages such as high energy efficiency, rich renewable energy and wide area of forest. However, these areas have some disadvantages such as the lack of talents and a low level of science and technology. Moreover, the high-energy consumption industries moved from the East and South coast area to these areas: therefore, there is a high ratio of fossil energy (see Table 4). Thirdly, in Guangxi and Yunnan (relatively low regions, group C), the environment and energy indicators are relatively high, but the economic indicators (i.e., GDP per Capita, Urbanization) are very low because these areas are located in the West and South areas with relatively low development (see Table 4). Finally, in the Inner Mongolia and Shanxi Province (lowest region, group D), the environment and energy indicators are lower than ones in other areas and the economy is also on the developing stage. Therefore, Shandong can not belong to the same category with Guangxi and Yunnan. In Addition, Heilongjiang Province can not belong to the same category with Guangdong, Fujian, Zhejiang, Jiangsu, Shandong and Liaoning.

Results shows, the weighted-distance version of quantum clustering can obtain a clustering result that better objectively reflects the actual situation.

4. Conclusion

In this paper, quantum clustering using the weighted distance by the Fuzzy-ANP method is studied. This paper consists of two parts. First, we have improved the cluster method by introducing a new weighing distance in the quantum cluster. Next, the improved method was mathematically and practically proven to be superior to the existing methods by using two-dimensional random data and 2015 3E data for 19 provinces in China. The method maintains the advantages of the traditional quantum clustering (QC) methods, for example the clustering center for arbitrary shapes can be detected according to the basic information of the data itself, and the samples do not need to be preprocessed. In addition, the importance of each attribute to the system can be more accurately reflected. The improved algorithm can be separated into two stages: weight determination and cluster analysis. However, the determination method of weight depends on the characteristics of each data. The reason why we applied the Fuzzy-ANP method to China’s 19 provinces’ 3E system data is as following: Firstly, ANP method is suitable for the weight determination because China’s 19 provinces’ 3E system data have a hierarchical structure, in which the relations between subsystems and between properties in one subsystem or different subsystems exist. Secondly, Fuzzy method was used to eliminate the influence of
subjectivity on the weight determination. The method is more suitable for real-world data with high dimensions and a certain degree of hierarchical structure data, where there is an interdependence and feedback relationship among the properties in the hierarchy. In this paper, our proposed method (Fuzzy-ANP-Weighted—QC) was firstly compared with k-means and QC methods as regard to two-dimensional random data (mathematical data), and then we performed a sound comparison with K-means and QC methods as regard to 2015 3E data (real data) for 19 provinces in China. The method provides satisfactory results in terms of accuracy and processing time, but the unavoidable shortcomings of quantum clustering methods remain.

Declarations

Author contribution statement

Fan Decheng: Conceived and designed the experiments; Analyzed and interpreted the data.

Song Jon: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Cholho Pang, Wang Dong, CholJin Won: Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by National Natural Science Foundation of China (71373059) and Central University Fundamental Fund for Fundamental Research Funds (HEUCF180901).

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[2] J.B. Macqueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of Berkeley Symp on Mathematical Statistics and Probability, 1967, pp. 281–297.
[3] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clustering algorithm for clustering large databases, in: ACM SIGMOD International Conference on Management of Data, 1998, pp. 73–84.

[4] G. Karypis, E.H. Han, V. Kumar, CHAMELEON: a hierarchical clustering algorithm using dynamic modeling, Computer 32 (8) (1999) 68–75.

[5] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficient data clustering method for very large databases, in: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1996.

[6] M. Ester, H.P. Kriegel, J. Sander, et al., A density-based algorithm for discovering clusters in large spatial data sets with noise, in: International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.

[7] M. Ankerst, M.M. Breunig, H.P. Kriegel, et al., OPTICS: ordering points to identify the clustering structure, ACM Sigmod Rec. 28 (2) (1999) 49–60.

[8] S. Roy, D.K. Bhattacharyya, An Approach to Find Embedded Clusters Using Density Based Techniques, ICDCIT 2005, Distributed Computing and Internet Technology, 2005, pp. 523–535.

[9] R. Agrawal, J. Gehrke, D. Gunopulos, et al., Automatic subspace clustering of high dimensional data for data mining application, in: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, 1998, pp. 94–105.

[10] W. Wang, J. Yang, R.R. Muntz, STING: a statistical information grid approach to spatial data mining, in: Proceedings of International Conference on Very Large Data Bases, 1997, pp. 186–195.

[11] S. Theodoridis, K. Koutroumbas, Pattern Recognition, fourth ed., Academic Press, Pittsburgh, 2008.

[12] D. Horn, A. Gottlieb, Algorithm for data clustering in pattern recognition problems based on quantum mechanics, Phys. Rev. Lett. 88 (1) (2002) 018702.

[13] U.V. Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.

[14] B. chölkopf, A. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (10) (1998) 1299–1319.

[15] C. Böhm, C. Plant, J. Shao, et al., Clustering by synchronization, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 583–592.
[16] J.B. Huang, J.M. Kang, J.J. Qi, et al., A hierarchical clustering method based on a dynamic synchronization model, Sci. China Series F Inf. Sci. 43 (5) (2013) 599–610.

[17] J. Shao, X. He, C. Böhm, et al., Synchronization inspired partitioning and hierarchical clustering, IEEE Trans. Knowl. Data Eng. 25 (4) (2013) 893–905.

[18] J. Shao, X. He, C. Plant, et al., Robust Synchronization-based Graph Clustering, Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, 2013, pp. 249–260.

[19] J. Shao, Z. Ahmadi, S. Kramer, Prototype-based learning on concept-drifting data streams, in: ACM SIGKDD, 2014, pp. 412–421.

[20] L.I. Zhi-Hua, S.T. Wang, Improved algorithm of quantum clustering, Comput. Eng. 33 (23) (2007), 189-189.

[21] R. Casaña-Eslava, I.H. Jarman, P.J.G. Lisboa, et al., Quantum clustering in non-spherical data distributions: finding a suitable number of clusters, Neurocomputing 268 (2017).

[22] Y. Li, Y. Wang, Y. Wang, L. Jiao, Y. Liu, Quantum clustering using kernel entropy component analysis, Neurocomputing 202 (2016) 36–48.

[23] D. Horn, A. Gottlieb, The method of quantum clustering, in: International Conference on Neural Information Processing Systems: Natural and Synthetic, MIT Press, 2001, pp. 769–776.

[24] Z.H. Li, S. Wang, Parameter-estimated quantum clustering algorithm, J. Data Acquis. Process. 23 (02) (2008) 211–214.

[25] R. Hu, Z. Jing, Q. Wu, et al., Application of modified close value method in power quality evaluation, in: International Conference on Materials Science, Machinery and Energy Engineering, 2017.