Research on Weighted Cluster Analysis Method of Panel Data

Xiqin Ao	extsuperscript{1*}, Yiwen Zhang	extsuperscript{1}, Yinyin Zhou	extsuperscript{1}, and Deyu Xu	extsuperscript{1}

	extsuperscript{1}Institute of big data and artificial intelligence, Anhui Xinhua University, Anhui, 230088, China
	extsuperscript{*Corresponding author’s e-mail: aoxiqin@axhu.edu.cn

Abstract. Panel data is composed of three dimensions: object, index, and time, it has no mature clustering algorithm because of its complexity and dynamic data structure. Two clustering methods for multi-index panel data are proposed in this paper. These two methods are respectively weighted from the index perspective and time perspective, and on this basis, the two-step clustering method is used for clustering. Then these methods are applied to cluster analysis of regional economy. The empirical results show that these two methods can effectively complete the clustering of the research objects, and have their own emphasis on the feature extraction angle of the panel data.

1. Introduction
As a common data form, panel data is composed of three dimensions: object, index, and time, which can be regarded as the expansion form of cross-sectional data in the time dimension[1]. Because the characteristics of panel data include the different properties of each index at different time points of the research object, the comprehensive evaluation based on panel data belongs to dynamic comprehensive evaluation, which needs to consider the weight factors of time and space at the same time, and also needs to solve the problem of multi-index evaluation model[2]. Cluster analysis refers to the process of dividing research objects into several categories according to their similarity.

There are two main problems in the existing panel data clustering analysis: (1) the index weight is not considered. That is, the importance of each index is not distinguished when calculating the distance in the process of sample clustering; (2) Time weight is not considered. It cannot reflect the influence of different time points on the classification of panel data.

In order to fully explore the information contained in the panel data, better reflect the overall dynamic change process of the comprehensive evaluation, and achieve a more scientific quantitative evaluation, it is necessary to study the method of determining the weight of the panel data. Therefore, this paper proposes to weight the data respectively from the perspective of index and time, and on this basis use the two-step method to cluster.

2. Principle and method of weighted clustering

2.1. Clustering method
The essence of clustering analysis is to divide the data into several categories according to the distance, so as to make the difference between the data within the categories as small as possible, and the difference between the categories as large as possible[3]. At present, clustering methods are mainly divided into two categories: traditional clustering methods and intelligent clustering methods. The traditional clustering methods mainly include partitional clustering and hierarchical clustering; The
intelligent clustering methods mainly include two-step clustering, neural network-based clustering, kernel-based clustering and so on[4].

Compared with the traditional clustering method, the two-step clustering method has distinct characteristics. First of all, the variables used for clustering can be continuous variables or discrete variables, which has a wider scope of application; Secondly, based on its algorithm principle, the two-step clustering method occupies less memory resources and the calculation speed is faster; Thirdly, it is really using statistics as a distance index for clustering, and at the same time it can "automatically" recommend or even determine the best number of categories based on certain statistical standards, the correctness of the results is more guaranteed. Based on this, the two-step clustering method is selected for clustering in this paper.

Two-step clustering is accomplished in two steps[5]. The first step is pre-clustering, which is completed by constructing and modifying the cluster feature tree. The clustering feature tree distinguishes nodes into branch nodes and leaf nodes. Each leaf node represents a sub-category, and those branch nodes and the statistics in them are used to guide the newly entered case which leaf node should enter. The information in each entry is the so-called clustering feature. For each case, it is necessary to start from the root to enter the clustering feature tree, and find the closest leaf node according to the guidance of the entry information in the node until it reaches the leaf node. The second step is formal clustering, which takes the pre-clustering results obtained in the first step as input and re-clusters it. This stage can be directly processed by traditional clustering methods.

2.2. Determination of weights

The weight of evaluation index is one of the important factors affecting the accuracy of evaluation results. How to reasonably allocate the weight of evaluation index has always been a hot issue for scholars in related fields. Through the analogy analysis of the existing methods to determine the time weight and index weight, this paper adopts the inverse trigonometric function method to determine the time weight, and the entropy method to determine the index weight.

2.2.1. Determination of time weight based on inverse trigonometric function method. The panel data has time series trend, and the importance and representativeness of index values are different at different time points. Obviously, when clustering panel data samples, the closer the time point, the more important and representative the index value is, the greater the weight should be, and with the time approaching, the weight coefficient should increase slowly, and finally tend to 1. Generally speaking, the time weight function should satisfy the following three conditions[6]:

1. F(t) is a strictly increasing function with time t;
2. The increasing speed of F(t) is gentle;
3. \[ \lim_{t \to \infty} F(t) = 1 \]

Inverse trigonometric function \( F(t) = \frac{2}{\pi} \tan^{-1} t \) \((t = 1, 2, \cdots, n)\) can meet the above basic requirements.

2.2.2. Determination of index weight based on entropy method. Entropy value method is a method to determine the importance weight of index according to the principle of entropy[7]. According to the basic theory of information theory, "information" is used to measure the degree of order of the system, and "entropy" is used to measure the degree of disorder of the system. If the information entropy of an index is smaller, it means that the more information the index provides, the greater the role it plays in clustering analysis, and the greater the index weight should be.

The steps of determining index weight by entropy method are as follows:

1. Calculate the proportion of the value of the i-th sample relative to other samples under the j-th index.
   \[ q_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, j = 1, 2, \cdots, p \]

2. Calculate the entropy value of the index \( x_j \)
\[ e_j = -k \sum_{i=1}^{n} q_{ij} \log(q_{ij}), j = 1,2,\ldots,p \]

In the formula (2), \( k > 0, \quad e_j > 0, \quad k = 1/\log n. \) The larger the \( e_j, \) the smaller the effect of the index \( x_j; \) on the contrary, the smaller the \( e_j, \) the greater the effect of the index \( x_j. \)

(3) The weight of index is determined after entropy normalization, and the weight coefficients of the index \( x_j \) are as follows:

\[ w_{x_j} = \frac{1-e_j}{\sum(1-e_j)}, j = 1,2,\ldots,p \]

3. Empirical analysis

3.1. Data source and collation

In order to verify the effectiveness of the weighted clustering method, this paper applied the method to cluster analysis of the economic development of various provinces and cities in China. From the website of the national bureau of statistics, eight sub-indexes of consumer prices of 31 provinces and cities in China from December 2019 to November 2020 were collected. The data set includes three dimensions: time, sub index and province, which was a typical panel data. Then the collected data was preprocessed, mainly included the detection of missing values and the treatment of abnormal values.

3.2. Determination of weights

3.2.1. Determination of index weight. The eight sub-indexes of consumer prices in the data set are the eight clustering indicators: food, tobacco and alcohol price index (x1); clothing price index (x2); residential price index (x3); daily necessities and services price index (x4); transportation and communication price index (x5); education, culture and entertainment price index (x6); medical and health care price index (x7); other supplies and service price index (x8).

Using entropy method to assign weights which mentioned in the previous section, the weights of the eight indexes calculated by formulas were 0.291, 0.076, 0.049, 0.023, 0.042, 0.076, 0.290 and 0.153.

3.2.2. Determination of time weight. Use the inverse trigonometric function mentioned in the previous section to determine the time weight. \( F(t) = \frac{2}{\pi} \tan^{-1} t \) \( (t = 1,2,\ldots,n), \) the value of n is 12, the time weights of 12 months were calculated as 0.049, 0.069, 0.078, 0.083, 0.086, 0.088, 0.089, 0.090, 0.091, 0.092, 0.092 and 0.093.

3.3. Cluster analysis of regional economy

3.3.1. Clustering analysis based on time weighting. The 12-month data was time-weighted by month, and the eight sub-indexes were used as clustering variables. The two-step clustering method was used for clustering. The clustering results were shown in the table 1, and the clustering comparison charts were shown in figure 1 to figure 4.

| categories | provinces and cities | quantity |
|------------|----------------------|----------|
| first category | Beijing, Henan, Tianjin | three |
| second category | Liaoning, Neimenggu, Ningxia, Shanxi, Shanghai, Xinjiang | six |
| third category | Anhui, Guangdong, Guangxi, Guizhou, Hainan, Heilongjiang, Hubei, Hunan, Jiangxi, Shandong, Sichuan, Yunnan | twelve |
| fourth category | Fujian, Gansu, Hebei, Jilin, Jiangsu, Qinghai, Shanxi, Xizang, Zhejiang, Chongqing | ten |
It could be seen from Table 1 that 31 provinces and cities were divided into four categories. From the cluster comparison charts it could be concluded that the first-category provinces and cities had higher consumer price indexes of other supplies and services, education, culture and entertainment, and health care, while the consumer price indexes of transportation and communication and clothing were relatively low. The second-category provinces and cities had higher consumer price indexes of housing and clothing, while consumer price indexes of other articles and services, food, tobacco and alcohol, daily necessities and services, education, culture and entertainment are lower. The third-category provinces and cities had higher consumer price indexes of food, tobacco and alcohol, while consumer price indices of residential, medical and health care were lower. The fourth-category provinces and cities had higher consumer price indexes of daily necessities and services, transportation and communications. It could be seen that there were obvious differences in the consumption expenditure structure of residents in different provinces and cities. The relevant parts could formulate corresponding economic policies according to the characteristics of consumption expenditure in the region, so as to promote faster and better economic development.

3.3.2. Clustering analysis based on index weighting. The eight sub-indexes are weighted according to their weights, and time was used as the clustering variable. The two-step clustering method was used for clustering. The clustering results were shown in the table 2.
Table 2. Clustering results

| categories   | provinces and cities                                      | quantity |
|--------------|----------------------------------------------------------|----------|
| first category | Gansu, Ningxia, Shanghai, Xinjiang                     | four     |
| second category | Neimenggu, Qinghai, Xizang                            | three    |
| third category | Anhui, Fujian, Guizhou, Hebei, Hunan, Jilin, Jiangsu, Jiangxi, Liaoning, Shanxi, Tianjin, Zhejiang, Chongqing, Guangdong, Hainan, Heilongjiang | sixteen  |
| fourth category | Beijing, Guangxi, Henan, Hubei, Shandong, Shanxi, Sichuan, Yunnan | eight    |

From the clustering results in table 2, it could be found that the provinces and cities belonging to the same category were relatively similar in the overall economic level, and these cities also had a certain correlation in geographical location, indicating that the economy had a certain degree of radiation within a certain range. Therefore, in general, the clustering results with time index as clustering variable can better reflect the overall economic development level and trend of each region.

4. Summary and conclusions

As a common data form, panel data had no mature clustering algorithm because of its complexity and dynamic data structure. This paper proposed two weighted clustering methods based on two-step clustering after weighting panel data from two perspectives of index and time. The clustering effect of the two methods was verified by clustering the regional economy. It had been verified that the two clustering models can effectively complete the clustering classification of research objects, and the focus was different. In practical applications, researchers needed to flexibly choose appropriate clustering methods based on the characteristics of the panel data itself and the focus of the research problem.

Acknowledgements

This research was financially supported by Anhui Province Innovative Training Program for College Students (Grant No. 201912216004); 2021 University Outstanding Young Talents Support Project (Project Master: Xiqin Ao, Project Name: Research on Weighted Cluster Analysis Method of Panel Data); School-level Research Institute (Grant No. yjs201707); 2020 Anhui University Scientific Research Project (Grant No. KJ2020A0784); School-level Research Team (Grant No. kytd201902).

Reference

[1] Zhang L J, Peng H (2017) . Research on weighted clustering analysis method of panel data. Statistics and information forum, 04: 21-26.
[2] Bian W L, Wang X N(2016). Research review on Stochastic Frontier Analysis of panel data [J]. Statistical research, 06: 13-20.
[3] Zhu L L(2016). Ordered cluster analysis of panel data on regional economic development differences. Statistics & Decision, 18:148-150.
[4] Jin J G(2014). Summary of clustering methods . Computer science, S2: 288-293.
[5] Zhang W T (2018), SPSS Statistical Analysis Advanced Course (3rd Edition). higher education press, Beijing.
[6] Liu Y X(2016). Research and application of panel data clustering method based on dynamic time warping . Statistical research, 11: 93-101.
[7] Dang Y G, Hou D Q(2016). Multi index panel data clustering method based on feature extraction [J]. Statistics and Decision, 19: 68-72.