Application of data mining technology in regional economic analysis

Siyu Zhu
School of Economics Shanghai University, Shanghai, 200444, China.
zhusiyu710@sina.com

Abstract. China's economy is developing rapidly at present, but the uncoordinated development among different regions still exists. Meanwhile, as the technology of data mining, a pair of data distribution characteristics and change rules, is gradually mature, it is also being applied in more fields. Therefore, this paper will study the specific application of data mining technology in economic analysis of folk art from the two aspects of k-means clustering algorithm and weighted CADD algorithm respectively.

1. Application of k-means clustering algorithm in regional economic analysis

1.1. Features of k-means clustering algorithm
K-means clustering algorithm is a division algorithm based on prototype. In the process of operation, the algorithm will take the specified number of users as the representative of the center of mass, so as to continue the cycle until reaching the maximum number of cycles. Therefore, its calculation is small and it is easy to find outliers. The principle of finding outliers is that they rarely appear in observation categories. However, in the search for minimum value, k-means clustering algorithm can only be applied in the search for local minimum value, which can be made up by multiple clustering. Finally, the k-means clustering algorithm is not applicable when the data distribution shape is in a non-convex category [1].

1.2. Sample selection
When studying the application of k-means clustering algorithm in regional economic analysis, samples should be selected. In the selection of samples, certain criteria should be followed, including comprehensiveness criteria, objectivity criteria, independence criteria and periodicity criteria. By abiding by the above four criteria, the samples can be displayed at different levels of content, and the influence of subjective thoughts on results can be reduced to the maximum extent [2]. Only in this way can the information content be independent from each other and relatively stable in a certain period of time. For this reason, this paper will display its sample selection and classification in the form of charts, details of which are shown in chart 1.

| Content of regional economic development | Natural quality | Pollution |
|----------------------------------------|----------------|----------|
|                                        |                 | Energy loss |
|                                        |                 | Energy storage |
| Economic attribute                     |                 | Financial situation |
|                                        |                 | Employment status |
|                                        |                 | Economic growth |
1.3. Modeling

Its modeling process requires a fixed process, which will be clearly shown in the flow chart below, as shown in figure 1 below.

![Flow chart of k-means clustering algorithm modeling](image)

Figure 1. Flow chart of k-means clustering algorithm modeling

In the above process, database analysis is carried out when the yearbook data meets the standards, and database analysis is carried out after the pre-processing results are obtained when the data does not meet the standards. In this paper, natural attributes are taken as an example. In the process of modeling, the influencing factors of natural attributes are input, and the clustering objects are combined to reduce dimensions to improve accuracy. Finally, the K value is confirmed and the correct amount of inter-class deviation is obtained [3].

1.4. Result analysis

The specific formula used in the process of natural attribute clustering algorithm is:

\[
\text{Magnitude} (i, j) = \frac{s^2(i, j)}{s^2(i)}
\]

From this formula, the attribute weight value of energy consumption can be deduced, whose specific attribute weight content is shown in chart 2.

![Attribute weight table](image)

Table 2. Attribute weight table

| Attribute                                | Weight1 | Weight5 | Weight8 | Weight10 |
|-----------------------------------------|---------|---------|---------|----------|
| The total energy consumption            | 0.24    | 0.86    | 0.14    | 0.87     |
| Energy consumption of value-added       | 0.54    | 0.02    | 0.50    | 0.05     |
| The gross power consumption             | 0.25    | 0.17    | 0.35    | 0.09     |

In the formula, I can represent the attributes of class I and their weight, and represent the fixed distance between the attributes. Its weight data is also enough to analyze the total energy consumption, value-added energy consumption and total power consumption of each region.
2. Application of weighted CADD algorithm in regional economic analysis

2.1. CADD algorithm concept
In the implementation of weighted CADD algorithm, the data obtained by calculation should be selected first, and its density value should be calculated, so as to obtain the average distribution point of its density, and select the largest point as the cluster center. After the above content is completed, the cluster center point is taken as the radius to get the reachable cluster value, and the above content is repeated until all the points meeting the requirements are divided. Those that do not meet the requirements are called outliers, divided separately. The specific process details are shown in figure 2 below [4].

2.2. Improvement of CADD algorithm
The normal operation of CADD algorithm mainly depends on the similarity of objects, so when it comes to the inherent properties of clustering objects are different, the results may be inaccurate. Therefore, the weighted method is adopted to improve the application of CADD algorithm in regional economy, so as to overcome its shortcomings in the processing of distribution density clusters and solve the shortcomings of complex cluster shapes in the case of artificial formulation of Numbers. The clustering time is greatly reduced if the cluster is divided according to the correct idea [5].

2.3. Specific application
The specific application of weighted CADD algorithm in regional economic analysis also takes natural attributes as an example, and its density parameter value is set as 1, while its adjustment parameter is set as 0.2. After specific calculation of the corresponding values, the maximum distance is obtained (accurate to 15 decimal places): 0.647505614564114; the average distance coefficient is 0.578248268915805 (accurate to 15 decimal places). Then, by excluding the weight, the attribute dimension can be reached to the correct amount.

3. Comparative analysis of different algorithms

3.1. Comparison of basic principles
Compared with the weighted CADD algorithm, k-means clustering algorithm is more suitable for the case of large computation amount, and its processing speed is relatively fast. Meanwhile, k-means clustering algorithm also has more advantages in finding outliers. Weighted CADD algorithm is more suitable for the case of relatively small number of samples, and it is calculated in different ways of dividing density clusters [6].

3.2. Accuracy of weighted CADD algorithm
The formula for calculating the accuracy of weighted CADD algorithm is:
\[ c = \frac{x}{y} \]

Where, \( x \) in the formula represents the correct number of objects in the cluster, and \( y \) represents the number of objects in the cluster. By dividing the two, the accuracy rate of \( c \) can be obtained. According to the actual data above, it can be concluded that the clustering accuracy will decrease as the dimension increases. Only after dimension reduction can the clustering result call me accurate.

3.3. application results

The application results of these two methods are also compared with the same selection of natural attributes mentioned above. When using weighted CADD algorithm, it can be found that it divides the regions rich in mineral resources in China into the same category, while k-means clustering algorithm is more general, but the clustering effect is also correct and has scientific basis.

4. Conclusion:

Through the discussion in this paper, the specific applications of k-means clustering algorithm and weighted CADD algorithm in regional economic analysis are analyzed respectively. In this process, relevant influencing factors are analyzed, hoping that the research results of this paper can provide some help for the sustainable development of regional economy, and play a role in the decision-making of subsequent development.

References:

[1] Mingjun huang. Analysis of spatial-temporal evolution characteristics and driving factors of economic development quality in the Yangtze River economic belt [D]. Zhejiang University of technology, 2019.
[2] Fang zhang. Theory and method of data scale division for multi-scale data mining [D]. Hebei normal university, 2019.
[3] Ruihan liu. User behavior analysis based on vehicle trajectory data [D]. University of Chinese academy of sciences (shenzhen institute of advanced technology, Chinese academy of sciences), 2019.
[4] Lingfeng Lin. Research on key technologies of performance data mining and algorithm optimization based on mobile processors [D]. University of Chinese academy of sciences (shenzhen institute of advanced technology, Chinese academy of sciences), 2019.
[5] Yuyan wang. Coupling degree analysis of coordinated development of ecological environment and regional economy in underdeveloped areas [D]. Tarim University, 2019.
[6] Zhongping cui. Empirical research on the innovation-driven role of regional economic development [D]. Liaoning University, 2019.