A Weighted Hierarchical Clustering Method Based on AHP
Liang HU1 and Ze-qing YAO2

1Graduate School, Army Engineering University of PLA, Nanjing Jiangsu 211101, China
2Department of Basic, Army Engineering University of PLA, Nanjing Jiangsu 211101, China

Keywords: Hierarchical clustering, Weighted distance, Analytic hierarchy process.

Abstract. Under specific circumstances, hierarchical clustering method cannot cluster accurately. This paper redefined the distance, weighted distances using analytic hierarchy process, through weighting made clustering indicators more focused, getting a better clustering consequent. Finally, we analyzed the influence of subjective and objective weights on clustering, and gave the scope of use.

Introduction
As a commonly used technique for data mining, there are many methods for clustering [1], such as hierarchical clustering, K-means clustering, and the neural networks for unsupervised learning in machine learning. All these methods must define distance formulas between two classes. The general clustering algorithms mostly have good effects on some certain types of data, but once the data distribution characteristics and the clustering targets change, the clustering methods will also be different.

Common distance formulas as follow:

Minkowski distance: \( d_q(x, y) = \left( \sum_{i=1}^{n} |x_i - y_i|^q \right)^{1/q} \) (when q=1, it’s Manhattan distance; when q=2, it’s Euclidean distance; when q tends to infinity; it’s Chebyshev distance)

Lance-Williams distance: \( d_{LW}(x, y) = \frac{1}{p} \sum_{i=1}^{p} \left| x_i - y_i \right| \)

Mahalanobis distance: \( d_M(x, y) = \left( x_i - y_i \right) \Sigma^{-1} \left( x_i - y_i \right) \)

Others distances defined by using similarity coefficients, such as angle cosine, correlation coefficient.

However, the above distances all acquiesced that each index of the data has the equal status in the evaluation. In reality. Because of the different choice of clustering target, the weight of each index in analysis should be different, so the distance formulas mentioned above will not be the best choices. Many times, the analytical data we get is quite different from what we want to get, and some are far from what we want. In order to solve this problem, scholars proposed to determine the weight of each index and use weighted distance [2] to cluster.

Improved Weighted Distance
Assumption: There are n samples, and each sample has p indexes, \( X_{ij} \) which denotes the value of the j-th index of the i-th sample. Get the following data matrix:

| NO. | X1  | X2  | ... | Xp  |
|-----|-----|-----|-----|-----|
| 1   | X_{11}| X_{12}| ... | X_{1p}|
| 2   | X_{21}| X_{22}| ... | X_{2p}|
| ... | ... | ... | ... | ... |
| n   | X_{n1}| X_{n2}| ... | X_{np}|
Weighted Distance Formula Definition

The weighted Euclidean distance formula mentioned in many papers is:

\[ d_y = \sqrt{\sum_{j=1}^{p} \omega_j (x_{ij} - x_{jy})^2} \]  

(1)

This formula obviously just adds a weight on the basis of Euclidean distance, but in fact it does not express the actual meaning of weight accurately.

This paper defines the weighted distance formula as follow:

\[ d_y = \left[ \sum_{j=1}^{p} \omega_j (x_{ij} - x_{jy})^q \right]^{\frac{1}{q}} \]  

(2)

\[ \omega_j > 0. \]

Absolutely it satisfies non-negative, identity, symmetry and transitivity.

Simplify \( d_y \)

\[ d_y = \left[ \sum_{j=1}^{p} (w_i x_{ij} - w_j x_{ij})^q \right]^{\frac{1}{q}} \]  

(3)

\[ y_{ij} = w_i x_{ij}. \]

In this way, not only the weight of distance is considered, but also the weight of distance can be equal to the weight relation between indexes, which is more practical and operational.

Weight Determination Method

The commonly weight determination methods are subjective or objective, include expert scoring method, analytic hierarchy process (AHP), entropy method, principal component Analysis (PCA) and Neural Network (Ann). In this paper, we adopted analytic hierarchy process (AHP) [3] to improve the clustering method based on hierarchical clustering.

AHP can make the unsupervised clustering have a certain goal, and can cluster more accurately according to the specific targets of the specific problems or specific requirements. It is helpful to have a clearer understanding of the target data to make a more accurate judgment.

It is divided into five steps to determine the weights:

Step 1: Analysis the relationship among the elements in the system, to establish the hierarchical structure of the system.

![Figure 1. Hierarchy of military forces.](image)
Step 2: Construct a pairwise comparison judgment matrix.

\[ C = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix} \]

\( C_{ij} \) indicates the value how important \( i \)-th factor than \( j \)-th factor. \( C_{ij} > 0 \), \( C_{ii} = 1/\sum_{j \neq i} C_{ij} \), \( C_{ii} = 1 \).

Step 3: Consistency of judgment Matrix

\[ CI = \frac{\lambda_{\text{max}} - n}{n - 1} \]

\[ CR = \frac{CI}{RI} < 0.1 \]

\( RI \) can be obtained by looking up relevant table.

Step 4: Calculate the relative weight \( \omega \) of the compared elements to the criterion from the judgement matrix. The square root method is used in this paper.

Step 5: Weighted distance formula:

\[ d_{ij} = \left( \sqrt{\sum_{l=1}^{n} \omega_l (x_{il} - x_{jl})^2} \right)^{\frac{1}{2}} \] (4)

Weighted Clustering Analysis

Weighted Hierarchical Clustering Based on AHP

Step 1: Data standardization \( x^* \)

Step 2: The experts determine the decision matrix and calculates the weight according to 2.2

Step 3: Data reprocessing \( y^* = \omega x^* \)

Step 4: Hierarchical clustering with SPSS[4].

Case

Different countries and regions have different levels of military forces. Only by facing up to the power gap and finding out the weak links can we enhance our own strength in the future military development. Taking the South China Sea as an example, analysis the military data[5][6] of China, the United States, Russia, South Korea, Japan, India, Malaysia, Singapore, Cambodia, Thailand, Indonesia, Vietnam, the Philippines and Taiwan, construct the determination matrix.

Table 2. Military data of relevant countries and regions in the South China Sea.

| Area       | Population | The Arm force | The Air force | The Navy force | Petroleum reserve | defense budget | Nuclear strike capability |
|------------|------------|---------------|---------------|----------------|-------------------|----------------|---------------------------|
| USA        | 0.5868     | 0.3490        | 1.0000        | 1.0000         | 0.4565            | 1.0000        | 0.9125                    |
| Russia     | 0.3067     | 1.0000        | 0.2745        | 0.9527         | 1.0000            | 0.0756        | 1                         |
| China      | 1.0000     | 0.5256        | 0.2135        | 0.8530         | 0.3125            | 0.2749        | 0.3125                    |
| Japan      | 0.0787     | 0.0365        | 0.1144        | 0.5613         | 0.0068            | 0.0742        | 0.0426                    |
| India      | 0.5703     | 0.3893        | 0.1514        | 0.3045         | 0.0709            | 0.0865        | 0.1623                    |
| South Korea| 0.2444     | 0.3467        | 0.1059        | 0.2869         | 0.0000            | 0.0742        | 0                         |
| Indonesia  | 0.1785     | 0.0051        | 0.0305        | 0.2070         | 0.0462            | 0.0114        | 0                         |
| Vietnam    | 0.1516     | 0.1964        | 0.0186        | 0.1380         | 0.0550            | 0.0054        | 0                         |
| Taiwan     | 0.1008     | 0.0938        | 0.0603        | 0.1239         | 0.0001            | 0.0179        | 0                         |
| Thailand   | 0.1052     | 0.0312        | 0.0388        | 0.1098         | 0.0050            | 0.0088        | 0                         |
| Singapore  | 0.0000     | 0.0070        | 0.0148        | 0.0685         | 0.0000            | 0.0162        | 0                         |
| Malaysia   | 0.0168     | 0.0012        | 0.0153        | 0.0595         | 0.0450            | 0.0077        | 0                         |
| Philippines| 0.0654     | 0.0000        | 0.0092        | 0.0522         | 0.0013            | 0.0048        | 0                         |
| Cambodia   | 0.0234     | 0.0196        | 0.0000        | 0.0000         | 0.0000            | 0.0000        | 0                         |
According to the square root method, the weight is obtained:

\[ \omega = (0.03, 0.05, 0.23, 0.37, 0.08, 0.15, 0.09) \]

From the weights, we can see that on the issue of the South China Sea, the navy and the air forces have absolute advantages over the military forces, which is in line with the reality.

From the clustering results, we find that the United States is the strongest military country, China and Russia followed, but still have a gap. South Korea, Japan and India are key military powers in the South China Sea. Combined with the results, it is not difficult to find that Japan is a relatively strong military power in these three countries, and cannot be ignored in the problem in the South China Sea. The strength of Thailand, Indonesia, Vietnam and Taiwan in the South China Sea is stronger than that of Malaysia, Singapore, Cambodia and the Philippines, but the latter have more interests in the South China Sea. How to maintain their stability is the top priority.

Compared with Fig 3, we can see the difference is China belongs to which camp, the first echelon or the second. There is no doubt that China belongs to the first echelon, and the conventional clustering ignores the weight of naval and air forces, causing the clustering results to deviate. At the same time we can get from the two graphs that the speed of weighted clustering is faster than that of conventional clustering. That can be attributed to the increase of distance difference caused by weighting.

Therefore, the weighted clustering is better than the ordinary clustering in clustering effect, and fits the research problem, and can get the relatively professional clustering, instead of blindly unsupervised conventional clustering, ignoring the important indicators.

**Conclusion**

In summary, we can conclude that the weighted clustering method has better clustering effect than the traditional clustering method. The objective weighting method can obtain more accurate clustering results on the data itself. When the subjective weighting method is applied to the clustering method, a more accurate clustering result can be obtained for the problem itself. Many weighting methods can be applied to weighted clustering. Different weighting methods have different emphases, and the purpose of clustering is different. Therefore, in the future, we need to choose weighted clustering methods according to specific requirements in order to obtain more accurate clustering results.
References

[1] He Ling, Wu Ling-da, Cai Yi-chao. A Survey of clustering algorithms in data Mining, J. Computer Application Research 01(2007) 10-13.

[2] Yuan Zhong-hu, Li Jia, Zhang Bo. A weighted hierarchical clustering method and its application, J. Journal of Shenyang University (Natural Science Edition). 2014, 26(03) 201-207.

[3] Saaty T L. How to Make a Decision: The Analytic Hierarchy Process, J. European Journal of Operational Research, 1990, 48(1) 9-26.

[4] He Xiao-qun. Multivariate statistical analysis, second ed., Renmin University of China press. Beijing, 2015.

[5] GFP. 2017 Military Strength Ranking on http://www.globalfirepower.com/countries-listing.asp

[6] Information on http://mil.huanqiu.com/observation/2014-06/5022028.html