A Clustering Algorithm Based on Information Visualization

[1] Ding Shifei, Qian Jun, Xu Li, Zhao Xiangwei, Jin Fengxiang
[1, First Author] *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116*
dingsf@cumt.edu.cn
Geomatics College, Shandong University of Science and Technology, Qingdao 266510
fxjin@sdust.edu.cn
doi:10.4156/jdcta.vol5. issue1.4

Abstract

This p aper st udies a clustering algorithm b ased on i nformation visulization. In th is alg orithm, through a nonlinear ma pping (NLP), s ome hig h-dimensional and complicated f eature data i s transformed into low-dimensi onal feature d ata, such as o ne, two and three dim ensionality. I ts mai n aim is that the geometry i mage in high-dimensional s pace is ma pped in to one, two a nd three dimensional i mage in low-dimensional s pace, and the inheren t data " structure" is approximat ely preserved after mapping. The simulated results show that the algorithm presented here is feasible and effective wi th d irect ob servation and im age et al. It describes well non linear character for h igh-dimensional feature data.

Keywords: *Information Visualization, High-dimensional Data, Nonlinear Mapping (NLM), Clustering Algorithm*

1. Introduction

Clustering analysis is a multivariate data analysis method which studies "Things of one kind come together". From the angle of pattern recognition, clustering analysis belongs to the nonsupervision pattern recognition problem. Its characteristic is the samples in input space has no expected output. For clustering problem, according to some similarity degree measures, similar samples are clustered into one class, different samples are clustered into different classes, namely the clustering process depends on the features among samples entirely. Classic clustering methods are as follows: Hierarchical Clustering method (HCA) which is the most widely used method in clustering analysis. Due to HCA is based on the distance matrix, merges classes gradually according to some clustering methods and needs to store distance matrix in the process of storage, so when the number of samples is very big, computer must has enough memory space which brings a certain degree of inconvenience for application. Dynamic Clustering method (DCA) solved this problem in technology. The most commonly used methods are c-mean method (CM) and fuzzy c-means method (FCM). These methods don't optimize the characteristics of sample and use them to cluster directly. Moreover, when FCM or CM adjusts a sample category, they calculate the mean values of each sample once again, so they are also called the method of correcting samples gradually. Iterative Self-organizing Data Analysis Algorithm (ISODATA) is called the batch sample correction method. ISODATA Algorithm has the features like heuristic reasoning, analysis supervision, control of clustering analysis structure and human-computer interaction, etc. It is a good method of clustering. Because of the effectiveness of the algorithm DCA depends on the distribution of samples largely, only when the natural distribution is the ball or closer to the ball, it just has better effect[1-5].

These clustering algorithms, in different data distribution and data structure, can achieve different clustering goals. But these clustering methods still exist several shortcomings: firstly, for some certain clustering methods, the classification results depends on a series of artificial determine parameters heavily, such as the sample similarity measures, all sorts of similarity threshold, the choice of clustering method, etc. There has no very good method to evaluate the clustering results. Secondly, cluster pedigree chart can't express the true relationship between samples. When the two groups close, the samples between them has a trend of forming a bridge, and is likely to cause false incorporated[6].

Information visualization (IV) is a technology which can chang the information and data into image expression that people can understand easily in order to provide a more rapid, effective service for computer customers. For the purpose of scientific exploration study, IV technology provides a unique method to explain phenomenon, reveal mechanism, find law and forecast results. Applications are involved in the following areas: mass data processing, human-machine interaction, finite element analysis, geographical information, the flood forecasting, meteorological service, material design, medical diagnosis, farming management, mechanics, physics, chemistry, etc. Experience tells us that a person's intuition has very big effect to classification, if the sample's distribution in the feature space can be shown, we can see which samples can come together directly, thus may belong to which class. But people can observe the space below the 3d or 3d, so we'd better put the original dimensional feature mapping to the planar space and display on this map. This mapping should keep the original sample distribution as much as possible or try to keep the mutual distances between samples remain unchanged.

In many cases, through the Linear mapping (LM), we can put the original dimensional feature space mapping to the two-dimensional plane and display them. Such as the principal component analysis (PCA) and factor analysis (FA) are all LM dimension-reduction methods. That is putting high-dimensional variables integrated into few comprehensive variables (linear combinations of original variables) which make aggregative index can express more original index information to the maximum extent. If you take three or below main component, the original data clustering features can be seeked directly by information visualization technology. The LM method often make the high-dimensional data meet original probability distribution, usually to meet high-dimensional normal distribution. Thus, this limits its application scope.

For some complex data, LM usually can not meet the mentioned requirements which maintain the distribution unchanged and also can't overcome the cluster analysis's shortcomings. At this time, we can use the nonlinear mapping (NLM) to reduce dimension. NLM overcomes these disadvantages largely and it uses geometric method to carry on dimensionality reduction. Through some nonlinear transform, it changes the geometric image in high-dimensional space into a low-dimensional (one-dimensional, two-dimensional or three-dimensional) space of the image, and the conversion can still keep the original geometric relationship. This method is direct-viewing and vivid, so that people could see some similar images about relationship among samples in high-dimensional space[7]. So studying clustering algorithm based on information visualization has important theoretical and practical significance.

2. Clustering algorithm based on information visualization

2.1. Algorithm theory of NLM

Supposed there are n samples, each sample contains p observation index, every sample point is equivalent to a p-dimensional space point: $X_i = (x_{1i}, x_{2i}, \cdots, x_{pi})$ (i=1,2,…,n). Now we will put n samples in the high-dimensional space R^p mapping into a low-dimensional space R^L (L<p). Namely, through the nonlinear mapping, we will change n samples in R^p into n samples in R^L: $Y_i = (y_{1i}, y_{2i}, \cdots, y_{Li})$ (i=1,2,…,n). Generally, in order to draw and be show in a computer screen, L is taken 1,2, or 3. Though this mapping, the distance of the n samples Y_i in space R^L must still approximates the distance of the n samples in high-dimensional space R^p. To achieve the goal, we will introduce nonlinear mapping restriction condition which change high dimension into low dimension.

$$K = \frac{1}{\sum_{i<j} d_{ij}^*} \sum_{i<j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} = \frac{1}{nF} \sum_{i<j} w_{ij} (d_{ij}^* - d_{ij})^2$$

$$nF = \sum_{i<j} d_{ij}^* = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^*$$

Where nF is called standard factor. $w_{ij} = 1/d_{ij}^*$, w_{ij} is called weight coefficient, d_{ij}^* is the distance between X_i and X_j in high-dimensional space R^p, d_{ij} is the distance between Y_i and Y_j in new space R^L. The meaning of K is that it will solve new space's geometric configuration when the difference's quadratic sum between the distance of new space and the distance of old space reaches the minimum.

2.2. The steps of NLM algorithm

Step 1 Data preprocessing: For the to be analyzed observational data matrix *X*, first of all, we shall make data conversion, the mathods are as follows: centralization, standardization. The specific methods can refer to the data conversion processing section in clustering analysis.

Step 2 Calculating the Euclidean distance between any two points X_i and X_j in P-dimensional space is.

$$d_{ij}^* = \sqrt{\sum_{k=1}^{p} (x_{ki} - x_{kj})^2}$$

Where $i, j = 1,2,\cdots,n$. The distance matrix is

$$D^* = \begin{pmatrix} d_{12}^* & d_{13}^* & \cdots & d_{1n}^* \\ & d_{23}^* & \cdots & d_{2n}^* \\ & & \cdots & \cdots \\ & & & d_{(n-1)n}^* \end{pmatrix}$$

Step 3 Seeking the value of *n* samples in new space R^L when K achieves the minimum value: taking any n initial points in R^L: $Y_1=(y_{11}, y_{21}, ..., y_{L1})$, $Y_2=(y_{12}, y_{22}, ..., y_{L2})$, ... , $Y_n=(y_{1n}, y_{2n}, ..., y_{Ln})$, and then putting them into formula K, so K is the function of $L \times n$ variables. If we use Euclidean distance, we can get $d_{ij}(m) = \sqrt{\sum_{k=1}^{L} (y_{ki}(m) - y_{kj}(m))^2}$

Where $i, j = 1,2,\cdots,n$, m is the number of iteration.

Step 4 Using the steepest descent method to search mapping minimum error E : the iteration formula is:

$$y_{ij}(m+1) = y_{ij}(m) - MF\,\Delta_{ij}(m)$$

Where MF is a magic factor which generally can be taken 0.3 or 0.4 by experience. And $\Delta_{ij}(m)$ is expressed as follows.

$$\Delta_{ij}(m) = \frac{\partial K(m)}{\partial y_{ij}(m)} \Big/ \left| \frac{\partial^2 K(m)}{\partial y_{ij}^2(m)} \right|$$

Where

$$\frac{\partial K(m)}{\partial y_{ij}(m)} = -\frac{2}{nF} \sum_{\substack{\alpha=1 \\ \alpha \neq i}} \left[\frac{d_{i\alpha}^* - d_{i\alpha}}{d_{i\alpha} \cdot d_{i\alpha}^*} \right] (y_{ij} - y_{i\alpha})$$

$$\frac{\partial^2 K(m)}{\partial y_{ij}^2(m)} = -\frac{2}{nF}\sum_{\substack{\alpha=1 \\ \alpha\neq i}} \frac{1}{d_{i\alpha}d_{i\alpha}^*}\left[(d_{i\alpha}^* - d_{i\alpha}) - \frac{(y_{ij} - y_{\alpha j})^2}{d_{i\alpha}}(1 + \frac{d_{i\alpha}^* - d_{i\alpha}}{d_{i\alpha}^*})\right]$$

For drawing easily, we often take L= 2. In order to reduce computation time, firstly, the original data are analysed by principal component analysis, secondly, identifying the first two principal components and constituting a factor surface, then putting the score points $Y_1, Y_2, \cdots, Y_N$ which are in n sample points $X_1, X_2, \cdots, X_N$ of the factor surface as the initial configuration of to carry on the iterative analysis.

3. Experiment

The original data comes from Bureau of Agriculture, Zibo City, Shandong Province[8]. According to local natural conditions, we choose 17 sets of land quality evaluation models associated with land productivity, denoted by *P(1),P(2),...,P(17)*, affecting characteristic indexes are as follows.

x_1 is soil organic matter content (%);

x_2 is total nitrogen content (%);

x_3 is rapid available phosphorus (ppm);

x_4 is rapid available potassium (ppm);

x_5 is cation exchange capacity (ml/100g);

x_6 is PH value;

x_7 is quality;

x_8 is ground water level (m).

y is acre-yield, the raw data are omitted.

Using NLM algorithm, the original eight-dimensional sample data are compressed into two-dimensional and shown in two-dimensional plane. We can look through the spatial distribution of the original eight-dimensional sample data so we can carry on further feature extraction and pattern classification. The compression results are shown in Figure 1.

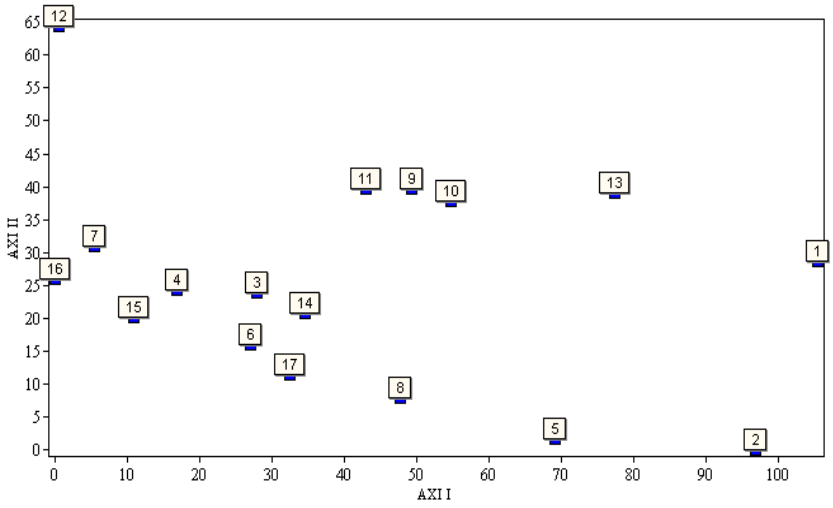

Figure 1. 2-dimensionality display based on NLM compression

In order to compare the compression effect of NLM, now these sample data will be compressed by PCA again, the PCA compression results are shown in Figure 2.

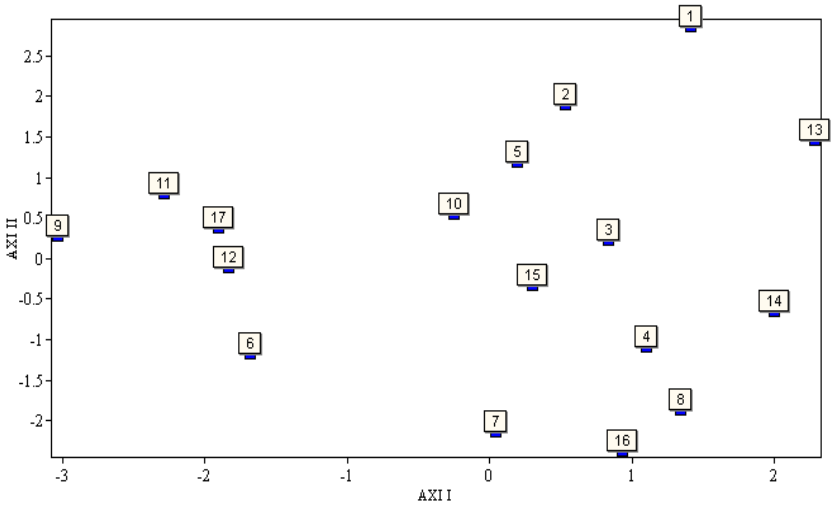

Figure 2. 2-dimensionality display based on PCA compression

Figure 1 and Figure 2 show that information feature compression based NLM can look through a number of approximate images which have the high-dimensional sample data interconnected feature, namely, the raw data have obvious classified characteristics and the original data have nonlinear characteristics. PCA-based information feature compression, the paper selected two principal components, then the two principal components payload 54.3194% of the original data information. The cumulative contribution rate ACR equals to 54.3194%. So it's hard to look the original high-dimensional data's classification features and the compression is less effective.

4. Conclusions

Based on researching the method of clustering analysis, pointed out the deficiencies of various clustering methods, for these deficiencies, we put forward a new clustering algorithm based on information visualization. By nonlinear mapping(NLM), this algorithm put the geometric image in high-dimensional space transforming into geometric image in low dimentional space, such as one-dimension or two-dimension. This method is taking the first two principal components factor scores of PCA as the initial configuration to carry on iterative analysis, visually displayed on a two-dimensional plane and can look through some approximate images which have the high-dimensional sample data interconnected feature. The PCA analysis is a linear dimension reduction method which through a linear transformation, integrate some high-dimensional variables to a few integrated variables. It makes the aggregative index expressing the original index information utmostly. In general, for two-dimensional display easily, people usually choose the first two principal components and make the two principal components as axes for two-dimensional display. It may loss more information, difficult to reflect the nonlinear characteristics of the high-dimensional original data and the compression effect is often poor.

5. Acknowledgements

This work is supported by the Basic Research Program (Natural Science Foundation) of Jiangsu Province of China (No.BK2009093), and the National Natural Science Foundation of China (Nos. 41074003, 60975039).

6. References

[1] Duda R.O., Hart P.E., Pattern Classification and Scene Analysis, Wiley, New York 1973.
[2] Bian Zhaoqi, Zhang Xuegong, Pattern Recognition, Tsinghua University Press, Beijing, 2000.
[3] Ding Shifei, Shi Zhongzhi, Liang Yong, et al., "Information Feature Analysis and Improved Algorithm of PCA," In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 2005, 1756-1761.
[4] Sun Jixiang. Modern Pattern Recognition, The Defence University of Science and Technology Press, Changsha, 2002
[5] Ding Shifei, Shi Zhongzhi, Jin Fengxiang, et al., "A Direct Clustering Algorithm Based on Generalized Information Distance," Journal of Computer Research and Development, Vol.44, No.4, pp.674-679, 2007.
[6] Tang Qiyi, Feng Mingguang, Practical Statistics and DPS Data Processing System, China Agricultural Press, Beijing, 1997.
[7] Sammon J.W., "A Nonliear Mapping for Data Structure Analysis," IEEE Transactions on Computers, Vol.18, No.5, pp.401-409, 1969.
[8] Ding Shifei, Xu Li, Zhu Hong, "Research and Progress of Cluster Algorithms based on Granular Computing," International Journal of Digital Content Technology and its Applications, Vol.4, No.5, pp.96-104.