Research on Cluster Analysis Method of Heterogeneous Data Resources Based on FCM

Xiaotao Xu

***Abstract*—Cluster analysis of heterogeneous data has a wide range of applications, which is an important basic method of big data mining, and is also a hot research topic in the field of information sharing. In order to improve the Cluster analysis effect of massive heterogeneous data, two pretreatment steps, data preparation and dimensionality reduction, are adopted in this paper. Firstly, irrelevant elements are eliminated to reduce the computation of Cluster analysis of heterogeneous data resources. Then, FCM Cluster analysis concept is introduced to cluster analysis based on membership matrix of Euclidean distance. Finally, the grey prediction model is used to predict and analyze the data. Cluster analysis efficiency of massive heterogeneous data has important reference value for massive data mining in the era of big data.**

***Index Terms*—Heterogeneous data, FCM, cluster analysis.**

I. Introduction

Cluster analysis of heterogeneous data resources is an important research direction in the development of modern big data technology. It is also the latest product of data mining (DM) technology in the era of big data. In the Internet + era, data sources with multiple sources, massive and heterogeneous data can only maximize the efficiency of sharing information through big data mining, give full play to the "long tail" effect of big data, and serve users of all kinds of information. In big data applications, the number and dimension of data resources are unprecedentedly huge, and the heterogeneous characteristics of data are obvious. All these bring great challenges to the analysis and application of big data resources. Cluster analysis of heterogeneous data resources is of great significance to the sharing and application of big data resources.

At present, with the rapid development of big data technology, various data engineering service providers have provided data mining process models to support big data applications, such as the typical SEMMA (Sample, Explore, Manipulate, Model, Assess) process model provided by SAS, the 5A (Assess, Access, Analyze, Act and Automatic) flow provided by SPSS. CRISPDM (Process model and Cross-industry Standard Process for Data Mining) widely used in business. Throughout these data mining models, are based on data preparation and data analysis as the core, the basic process is much the same. In the Internet + information sharing environment, massive heterogeneous data resources, as core resources, are more complex and diverse. They not only reflect the actual situation intuitively through data analysis, but also predict the future development trend in some areas, and even analyze outliers [1]. The "long tail effect" can detect abnormal phenomena in time so as to deal with them in time. The main data mining process models and big data resource sharing activities at home and abroad are analyzed comprehensively. The Cluster analysis method of heterogeneous data resources based on FCM proposed in this paper mainly includes data preparation, dimensionality reduction, Cluster analysis and other steps.

II. Data Preparation

Data preparation is the basic link of Cluster analysis of massive heterogeneous data resources. In the application of traditional data mining technology, data preparation takes the longest time, which directly affects the normal operation of subsequent data analysis models. In the resource sharing activities in the era of big data, data preparation is unprecedentedly heavy and difficult, including data sample selection, data quality analysis and data set preprocessing [2].

A. Data Sample Selection

In the Cluster analysis of massive heterogeneous data, the data capacity of the data source is often very huge. It is not necessary to collect all the data from the data source for analysis. In engineering practice, data sample selection is often used to analyze the data acquisition associated with big data mining objectives. In the Internet + information sharing environment, big data resources are of various types and huge capacity. When choosing data samples, it is difficult to choose the only criteria. Instead, the corresponding sampling methods should be selected for different data types. Typical methods include random sampling, systematic sampling, group sampling and stratified sampling [3]. Among them, random sampling is to extract samples according to random rules, the advantages are simple application, the disadvantages are that large-scale data is difficult to identify, suitable for small-scale data mining; system sampling is also called equidistant sampling, the advantages are uniform rules, simple marking, the disadvantages are difficult to adapt to random change environment, applicable to strong regularity of the knot. Constructive Data Mining; Group Sampling is to divide large-scale data into different data sets and then sample them in the data sets. It is easy to organize, but the disadvantage is that the sampling error of large-scale heterogeneous data is large, and it is suitable for sampling of professional data sources; Stratified sampling is to divide the data into different levels. Source segmentation, and then random sampling process, has the advantage of clear hierarchy, the disadvantage of different levels of data

Manuscript received February 9, 2019; revised May 7, 2019.

Xiaotao Xu is with Institute of Information and Communication, Wuhan, China (e-mail: superxxt@163.com).

doi: 10.18178/ijiee.2024.14.3.1

organic connection is difficult, suitable for different levels of data sample selection [4].

B. Data Quality Analysis

Data quality analysis is the process of evaluating the validity of data, screening available data, and directly deciding whether the sample data can be used in big data mining. Commonly used data quality analysis methods include value analysis, statistical analysis and histogram analysis. Among them, value analysis is the most basic method of data quality analysis, mainly through the analysis of data variable null value proportion, non-zero proportion, unique value and other elements, the process of data validity distribution is obtained; statistical analysis is the process of using mathematical statistics knowledge to analyze data. Statistics values such as mean value, maximum value, minimum value, variance, mode, quantile, median and skewness are usually used to describe the statistical characteristics of data [5]. Histogram analysis is to represent the sample data in the form of a histogram to visually describe the process of data distribution. In the Internet + information sharing environment, in order to maximize the "long tail effect" of unstructured data, we mainly use the method of value analysis to filter the sample data effectively, and use statistical analysis and other in-depth analysis methods in the subsequent big data mining links [6].

C. Data Set Preprocessing

After data quality analysis, the data set can satisfy the data validity problem. Data set pretreatment is also needed to solve the data credibility and interpretability problems, including missing value processing, noise filtering, data transformation and data reduction. Among them, deletion method, mean interpolation method, regression interpolation method and maximum likelihood estimation method can be used to deal with missing values, mainly to meet the integrity requirements of structured data; noise filtering is mainly to filter the random error data of the data set, screening noise data which obviously deviates from the normal threshold range [7]. Generally, regression analysis, mean smoothing and wavelet analysis are used; data transformation mainly solves the problem of explaining data, and usually adopts the methods of data standardization, data discretization and semantic transformation, so that data mining users or data mining systems can accurately understand data information; data reduction is through attribute selection. Selection or sample selection integrates the transformed data to form a relatively small amount of data, but is close to the original data set of refined data sets to improve the implementation efficiency of subsequent big data mining activities.

III. Dimensionality Reduction Processing

In the Internet+ information sharing environment, big data resources are of diverse sources and complex structures. In the process of developing big data resources, data resources are characterized by high-dimensional characteristics, variable quantities and complex relationships. It is necessary to reduce dimension processing so as to facilitate visual analysis and provide support for generating battlefield situation products. Considering the situation that big data resources involve many specialties and the coexistence of common data products and special data products, principal component analysis (PCA) can be used to reduce the dimension of data resources [8]. The basic idea of principal component analysis is to combine the related variables of the data set into independent variables linearly, then analyze the contribution rate of each main component variable, and analyze which main component combination can describe the basic characteristics of the original data set by evaluating the score based on the contribution rate. And realize the dimensionality reduction of complex data sets.

Based on principal component analysis, this paper quantitatively describes the dimensionality reduction process of big data mining, including standardization processing, correlation matrix calculation, component contribution description and main component evaluation.

A. Standardized Treatment

Standardization is the initial step to reduce the dimension of data by using principal component analysis. The main process is to construct the observation matrix M of sample data by using the sample data set provided by the previous data preparation step:

$$M=\begin{bmatrix} m_{11} & m_{12} & \square & m_{1p} \\ m_{21} & m_{22} & \square & m_{2p} \\ \vdots & \vdots & \ddots & \ddots \\ m_{q1} & m_{q2} & \square & m_{qp} \end{bmatrix}$$

The matrix M^* can be obtained by normalization of the sample data observation matrix:

$$M^*=\begin{bmatrix} m^*_{11} & m^*_{12} & \square & m^*_{1p} \\ m^*_{21} & m^*_{22} & \square & m^*_{2p} \\ \vdots & \vdots & \ddots & \ddots \\ m^*_{q1} & m^*_{q2} & \square & m^*_{qp} \end{bmatrix}$$

Among,

$$m^*_{ij}=\frac{m_{ij}-\overline{m}_j}{\sqrt{\operatorname{var}(x_j)}}\ ; \quad \overline{m}_j=\frac{1}{q}\sum_{i=1}^{q}m_{ij}\ ;$$

$$\operatorname{var}(m_j)=\frac{1}{q-1}\sum_{i=1}^{q}(m_{ij}-\overline{m}_j)^2$$

$$i=1,2,\square,q\ ; \quad j=1,2,\square,p$$

B. Incidence Matrix Computation

The correlation matrix of the sample data set is mainly used to quantify the related attributes among the variables of the sample data set. The correlation coefficient can be expressed as follows:

$$K=\begin{bmatrix} k_{11} & k_{12} & \square & k_{1p} \\ k_{21} & k_{22} & \square & k_{2p} \\ \vdots & \vdots & \ddots & \ddots \\ k_{p1} & k_{p2} & \square & k_{pp} \end{bmatrix}$$

Among,

$$k_{ij}\ \frac{\operatorname{cov}(m_i,m_j)}{\sqrt{\operatorname{var}(m_1)}\sqrt{\operatorname{var}(m_2)}}=\frac{\sum_{x=1}^{q}(m_{xi}-\overline{m}_i)(m_{xj}-\overline{m}_j)}{\sqrt{\sum_{x=1}^{q}(m_{xi}-\overline{m}_i)^2}\sqrt{\sum_{x=1}^{q}(m_{xi}-\overline{m}_j)^2}}\quad q>1$$

C. *Component Contribution Description*

The corresponding eigenvector can be obtained by calculating the eigenvalue set a\b\c of incidence matrix K.

$$\lambda_i = (\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{ip}) \ , \ i = 1,2,\cdots,p$$

According to the above eigenvector analysis, the P main components can be obtained from the sample data set. Since the variance of the main components presents a decreasing relationship, the corresponding proportion of information contained should also be a decreasing relationship. If the contribution rate is used to describe the importance of each main component, it can be expressed as follows:

$$w_i = \frac{\alpha_i}{\sum_{i=1}^{p} \alpha_i}$$

D. *Evaluation of Major Components*

According to the original data after standardization, the scores of the main components can be obtained by substituting them into the component contribution description formulas, as follows:

$$E = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1x} \\ e_{21} & e_{22} & \cdots & e_{2x} \\ \vdots & \vdots & \ddots & \ddots \\ e_{q1} & e_{q2} & \cdots & e_{qx} \end{bmatrix}$$

Among,

$e_{ij} = a_{j1}m_{i1} + a_{j2}m_{i2} + \cdots + a_{jp}m_{ip}$; $i = 1,2,\cdots,q$; $i = 1,2,\cdots,x$

According to the score, not only the contribution rate of each main component can be obtained, but also the correlation between the original variables can be analyzed, and the dimensionality reduction of the sample data set can be realized.

IV. Cluster Analysis

Cluster analysis is an engineering application of the idea of "clustering things by clusters" in the field of quantitative analysis. It is a multivariate statistical method. In the process of Cluster analysis of massive heterogeneous data, because of its large capacity and complex structure, it is necessary to cluster the data according to its basic attributes, so that the data with relatively large degree of association can be analyzed in the same data set, so as to improve the efficiency and accuracy of data analysis. Typical Cluster analysis methods include neural network clustering method, K-means method, Gaussian mixture clustering method and Fuzzy C-means Method (FCM) [9]. Among them, neural network clustering method model is difficult to build, in the case of less original data accumulation, learning experience is insufficient, easy to produce obvious errors; K-means method and Gaussian mixture clustering method are suitable for small capacity, low dimension data mining applications, not suitable for comprehensive evaluation of the data Cluster analysis; FCM method through the use of data clustering; Membership to determine the clustering of data, simple structure, high operational efficiency, suitable for data mining applications with large differences in data capacity and complex latitude situation. In this paper, we believe that in the information sharing environment of the Internet + era, massive data analysis based on big data can be implemented by FCM method, including the steps of determining objective function, iterative relation description and clustering conclusion analysis.

A. *Determining Objective Function*

Taking the observation matrix M of sample data as an example, Cluster analysis is to divide Q samples into C classes and describe the center of these C classes by $V = (v_1, v_2, \cdots v_c)$. Each sample is divided by membership matrix $U = (u_{ik})_{c \times q}$. The objective function of Cluster analysis can be defined as follows:

$$J(U,V) = \sum_{k=1}^{q}\sum_{i=1}^{c} u_{ik}^{iq} d_{ik}^{2}$$

Among, $d_{ik} = \| m_k - v_i \|$ denotes the Euclidean distance between the first cluster center and the K cluster center, $J(U,V)$ denotes the sum of the weighted square distances from the sample data to the cluster center, and the weight coefficient is the x power of the class I membership degree of the data m_k. Therefore, the Cluster analysis based on FCM method is essentially the case of obtaining the minimum $J(U,V)$ value.

B. *Iterative Relation Description*

The original membership matrix $U^{(0)} = (u_{ik}^{(0)})$ is constructed by uniformly distributed random numbers in [0,1] interval. After the Y th iteration, the membership matrix is $U^{(l)}$, and the clustering center can be expressed as follows:

$$v^{(y)} = \frac{\sum_{k=1}^{q} (u_{ik}^{(y-1)x} m_k)}{\sum_{k=1}^{q} (u_{ik}^{(y-1)})^{x}} \ , \ i = 1,2,\cdots,c$$

After the Y iteration, the membership matrix can be expressed as:

$$u_{ik}^{(y)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^{y}}{d_{jk}^{y}} \right)^{\frac{2}{x-1}}} \ , \ i = 1,2,\cdots,c \ ; \ k = 1,2,\cdots q$$

After the Y iteration, the objective function can be expressed as:

$$J^{(y)}(U^{(y)}, V^{(y)}) = \sum_{k=1}^{q}\sum_{i=1}^{c} (u_{ik}^{(y)})^{x} (d_{ik}^{(y)})^{2}$$

Among, $d_{ik}^{(y)} = \left\| m_k - v_i^{(l)} \right\|$

C. *Cluster analysis*

The termination threshold of membership degree $\sigma_u > 0$, when the maximum iteration step $\max\left\{ \left| u_{ik}^{(y)} - u_{ik}^{(y-1)} \right| \right\} < \sigma_u$, stops the iteration process, then the objective function $J(U,V)$ is the smallest. According to the membership degree matrix U and the clustering center V, all sample data sets can be distinguished from each other. When $u_{jk} = \max_{1\le i\le c}\{u_{ik}\}$, data m_k belongs to class J data set.

V. Data Prediction

Data forecasting is to predict the development trend of the things to which the data belong according to the objective law of the sample data. It is an important development direction of the application of large data technology. It develops very rapidly in the field of shared economy. Many Internet e-commerce platforms use large data forecasting function to predict the future development based on the operation data of existing e-commerce platforms. The trend prediction has received very good prediction results. Grey prediction model and Markov prediction model are widely used nowadays [10]. Grey prediction model has simple structure and good applicability. It has no special requirement for data rules, and can analyze and forecast data sets with weak correlation. The state transfer feature is suitable for data analysis and prediction with strong regularity. The state transition characteristics of massive heterogeneous data resources are not obvious, so the most typical single-sequence first-order linear differential equation model in the grey prediction model will achieve better results. It mainly includes the construction of the prediction model equation, the solution of the prediction model equation, the accuracy test of the prediction model and the data grey prediction analysis.

A. Constructing Prediction Equation

First-order linear differential equation is the basic model of grey prediction theory applied to the field of data prediction. The data set generated by clustering analysis can be substituted into the equation and the original data set can be accumulated once. The grey prediction model of data can be constructed. The main steps are as follows:

The original data set of data prediction can be expressed as follows:

$$X^{(0)} = \left\{x^{(0)}(i) \ge 0, i = 1,\cdots n\right\}$$

According to the typical first order linear differential equation model, the accumulation of the original data set can be expressed as:

$$x^{(1)}(i) = \sum_{m=1}^{i} x^{(0)}(m)$$

The background values of first order linear differential equations for data prediction can be expressed as:

$$z(k) = 0.5\times(x^{(1)}(k) + x^{(1)}(k-1)),\ k = 2,\cdots,n$$

The first order linear differential equation for data prediction can be expressed as:

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = b$$

B. Solution of Prediction Equation

In the first order linear differential equation described above, the parameters a and B are undetermined coefficients, in which a is the development coefficient and B is the grey action quantity. Assuming the grey parameter $a = [a,b]^T$ as the parameter sequence, the cumulative data can be averaged to generate B and constant term vector Y, which can be expressed as:

$$B = \begin{bmatrix} -z(2) & 1 \\ -z(3) & 1 \\ \cdots & \\ -z(n) & 1 \end{bmatrix} \quad Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \cdots \\ x^{(0)}(n) \end{bmatrix}$$

According to the least square method, the grey parameters of the prediction model equation can be calculated:

$$a = (B^T B)^{-1} B^T Y$$

The grey parameter is substituted into the data prediction model equation, and the solution of the equation can be obtained as follows:

$$x^{(1)}(k+1) = \left[x^{(1)}(0) - b/a\right]e^{-at} + b/a \qquad i = 1,\cdots n$$

When $x^{(1)}(0) = x^{(0)}(0)$ is taken, the restore value of the model is:

$$x^{(0)}(k+1) = x^{(1)}(k+1) - x^{(1)}(k) = (1-e^{a})(x^{0}(1) - \frac{b}{a})e^{-ak} \qquad k = 0,\cdots n$$

Among them, $x^{(0)}(k)$ (k = 1, 2,…,n) is the original data sequence, its fitting value can be expressed as: $x^{(0)}(k)$ (k = 1, 2,…, n) ; and the original data sequence $x^{(0)}(k)$ (k >n) prediction value can be expressed as: $x^{(0)}(k)$ (k >n)

C. Accuracy Test of Equation

When the grey theory is used for data prediction, the accuracy of the equation of the data prediction model should be checked, and the precision data of the model should be obtained, so as to judge the validity of the prediction model. The accuracy of the prediction model equation is as follows:

Let the original sequence of data prediction be represented by $X^{(0)} = \left\{x^{(0)}(i), i = 1,\cdots n\right\}$, the corresponding analog sequence by $\overline{X}^{(0)} = \left\{\overline{x}^{(0)}(i)\right\}$, the residual sequence by $\varepsilon^{(0)} = \left\{\varepsilon(i), i = 1,\cdots n\right\} = \left\{x^{(0)}(i) - \overline{x}(i)\right\}$, and the relative error sequence by:

$$\Delta = \left[\left|\frac{\varepsilon(1)}{x^{(0)}(1)}\right|, \left|\frac{\varepsilon(2)}{x^{(0)}(2)}\right|, \ldots, \left|\frac{\varepsilon(n)}{x^{(0)}(n)}\right|\right] = \{\Delta_k\} \qquad k = 1, 2, \cdots n$$

, The mean $\overline{x}$ and variance S_1^2 of the original sequence can be expressed as:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x^{(0)}(k)$$

$$S_1^2 = \frac{1}{n}\sum_{k=1}^{n}(x^{(0)}(k) - \bar{x})^2$$

The mean $\bar{\varepsilon}$ and S_2^2 variance of the residuals are:

$$\bar{\varepsilon} = \frac{1}{n}\sum_{k=1}^{n}\varepsilon^{(0)}(k)$$

$$S_2^2 = \frac{1}{n}\sum_{k=1}^{n}(\varepsilon^{(0)}(k) - \bar{\varepsilon})^2$$

According to the classical grey forecasting theory, the precision of grey forecasting model can be judged accurately by residual evaluation method in the forecasting application based on the first order linear differential equation. If the average simulation relative error $\bar{\Delta} = \frac{1}{n}\sum_{k=1}^{n}\Delta_k$, the mean square deviation ratio $C = S_2 / S_1$ and the small error probability $p = P(\left|\varepsilon(k) - \bar{\varepsilon}\right| < 0.6785 S_1)$ are assumed, the more the average simulation relative error and the mean square deviation ratio are. The smaller the better, the smaller the probability of error is the bigger the better. In the threshold judgment, the time series can be moved forward, and the original data can be used to simulate the calculation. The critical values of $\bar{\Delta}$, C and p, which meet the minimum requirements of the actual situation, can be determined as the precision threshold of the prediction model, so as to determine whether the grey prediction model can meet the accuracy requirements of the data prediction. The prediction results are effective.

VI. Conclusion

In the era of big data, heterogeneous data resources are expanding rapidly, and data mining technology is developing rapidly. However, it is still difficult to adapt to the growth rate of massive heterogeneous data. Based on Euclidean distance and multi-layer data preprocessing, the method proposed in this paper uses FCM to cluster heterogeneous data resources, which can effectively improve the efficiency of data mining and reduce the demand for computing resources. In order to further improve the quality of data resources mining, the next step will also be based on the conclusion of Cluster analysis of massive heterogeneous data for outlier diagnosis, in order to give full play to the application value of big data resources.

References

[1] L. Shen and W. Li, “Data synchronization model for collaborative system integration,” *Application Research of Computers*, vol. 4, no. 53, 2012.

[2] G. Matei and R. C. Bank, “Column-oriented databases, an alternative for analytical environment,” *Database Systems Journal*, vol. 1, no. 2, pp. 3-16, 2010.

[3] L. Tang, “A variational level set model combined with FCMS for image clustering segmentation,” *Mathematical Problems in Engineering*, vol. 2014, no. 2, pp. 1-24, 2014.

[4] D. H. Shin, “Ecological views of big data: Perspectives and issues,” *Telematics and Inforatics*, vol. 32, vol. 311-320, 2015.

[5] C. K. S. Leung, “Mining frequent patterns from uncertain datawith mapreduce for big data analytics,” *Database Systems for Advanced Applications Lecture Notes in Computer Science*, vol. 7825, pp. 440-455, 2013.

[6] J. Jiang and H. Zhang, “Dietary intake of human essential elements from a total diet study in Shenzhen, Guangdong Province, China,” *Journal of Food Composition and Analysis*, vol. 39, no. 2015, pp. 1-7, 2014.

[7] S. S. SivathaSindhu, S. Geetha, and A. Kannan, “Decision tree based light weight intrusion detection using a wrapper approach,” *Expert Systems with Applications*, vol. 39, no. 1, pp. 129-141, 2012.

[8] J. Ji, W. Pang, C. Zhou, *et al.*, “A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data,” *Knowledge-Based Systems*, vol. 30, pp. 129-35, 2012.

[9] A. Ghorbel, M. Ghorbel, and M. Jmaiel, “Privacy in cloud computing environments: A survey and research challenges,” *Journal of Supercomputing*, pp. 1-38, 2017.

[10] S. K. Sood and R. Sandhu, “Matrix based proactive resource provisioning in mobile cloud environment,” *Simulation Modelling Practice&Theory*, vol. 50, pp. 83-95, 2015.

Xiaotao Xu was born in 1981, in Hubei Province of China. He graduated from NUDT, and obtained doctor degree in 2018, work as an associate professor at Information and Communication Institute of NUDT, research mainly on information and communication security.