The Application and Analysis of Data Mining in Clustering Data of Petroleum Pipeline

Donghua Yu, Shuangshuang Sun, Yan Shen
College of Science, Applied Mathematics
Harbin Engineering University
Harbin 150001, China;

Wenlu Zhou
Foreign Languages Department, English
Harbin Engineering University
Harbin 150001, China

***Abstract*-Adopting the methods of the K-means and the SOFM neutral network in the data mining and basing on the characteristics of data of petroleum pipeline, a system is built that fits for the data mining and the evaluation on the effects of the clustering data. It is shown that the effects of this data mining are the best by comparison between the two results of the clustering data by using the multiple regression analysis.**

Keywords-Data mining, K-means, SOFM neutral network, Multiple regression analysis.

I. INTRODUCTION

The data mining makes the database technique enter into a higher stage. It can not only query and ergodic the past data but also find out the potential connections between the past data. Thus it can promote the transmission of the information. Today, data mining technology is mainly applied in the classifying, clustering and forecasting areas[1]. This paper adopts the clustering K-means and SOFM neutral networks.

In the oil and gas industry, pipeline transport grows rapidly in the present worldwide. There is an obvious advantage of the pipeline transport over others in the transportation of the oil and gas. The oil transport is an important link in the normal operation of this industry, and it preserves large amount of statistics as well, which provides a great deal of information for the analysis of the whole oil transportation[2]. So, analyzing the data of petroleum pipeline by using the clustering algorithm in the data mining is helpful to the thorough research on the fault identification of the pipeline, diagnosis and forecast, the safety of the transport, the response to accidents, the price making after interconnection reconstruction**.**

II. CLUSTERING METHOD AND REGRESSION ANALYSIS IN DATA MINING

Data mining is a process which reveals meaningful and new relationship, tendency , and pattern through analyzing amount of data carefully. Clustering is one that classifies data items into many clusters in which the data have as few differences as possible, and between which the data have as many differences as possible[3].

A. Clustering method of K-means in data mining

K-means algorithm is one that classifies n data objects into k clusters. k (final classification number) is a parameter of K-means algorithm.Then higher similarity of the same cluster can be ensured, and lower similarity of the different cluster. According to average of data object of one cluster, similarity calculation can be operated.

Algorithm flow of K-Means algorithm

- Select k objects randomly as initial center of clusters;
- Distribute every object to every initial center of clusters;
- Repeat;
- Classify every data object into the most similar cluster, according to the average of data object of one cluster;
- Calculate the average of data object of every cluster, then update the average of every cluster;
- Stop until average error $E \le \varepsilon$ or every cluster dose not change again[4].

B. Clustering method SOFM neural network in data mining

Self-organizing feature map (SOFM) is a self-guiding, self-organizing, self-learning network which is composed of fully connected neurons array [5]. It is the main purpose of SOFM that arbitrary dimensional input signal can be transformed to one-dimensional or two-dimensional discrete mapping, and this transformation can be realized adaptively by the way of topological order[6].

Algorithm flow of SOFM network algorithm

- Initialize the network;
- Accept input vector;
- Find out winning node;
- Define winning neighborhood;
- Adjust weights;
- Stop and check [7-9].

C. Principle of multiple regression analysis

Regression analysis is a method which finds out some regularity from practical data. Certain and uncertain relationship between observable factor variables $(x_1, x_2, \cdots, x_n)$ and dependent variable Y can be set up, that is

$$Y = f(x_1, x_2, \cdots, x_n) + \varepsilon \tag{1}$$

$$y = E(Y) = f(x_1, x_2, \cdots, x_n) \tag{2}$$

Published by Atlantis Press, Paris, France.

In which, ε, Y are random variables , $\varepsilon \sim N(0,\sigma^2)$. Autocorrelation test of residual is also called D-W test[10]. The statistics of D-W test is

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{3}$$

Test for heteroscedasticity of residual is also called Goldfeld-Quant test[10], and the statistics of F is

$$F = \frac{\sum e_2^2 \Big/ \left[\frac{(n-c)}{2} - k\right]}{\sum e_1^2 \Big/ \left[\frac{(n-c)}{2} - k\right]} = \frac{\sum e_2^2}{\sum e_1^2} \sim F\left(\frac{n-c}{2} - k, \frac{n-c}{2} - k\right) \tag{4}$$

III. EMPIRICAL ANALYSIS OF CLUSTERING METHODS----K-MEANS AND SOFM NEURAL NETWORK IN OIL PIPELINE DATA

A. Pretreatment of oil pipeline data

With the development of database technology, storage data have reached thousands mega. Because of database itself with noise interference, data should be pretreated to improve the quality of cluster. There are many methods of pretreatment of data, such as data cleaning, data merging, data protocol, data transformation, data standardization[11].

There are 61 sample data. Delete null value problem directly, thus 3 samples are deleted. Translation is a reversible process which this paper adopts to deal with data, and every sample of negative data pluses vector $\vec{a} = (85,85,85,85,85)$. Select four attribute for clustering. In order to cluster easily , 59 renewedly numbering data should be stored in format files in TXT and EXSEL.

B. Application of clustering methods of data mining in oil pipeline

1) Experimental analysis of clustering result by K-Means of data mining

Result of table I can be get by the operation of program which has already been programed during environment of DEV—C++ and reading data in format files in TXT:

2) Experimental analysis of clustering result by SOFM network of data mining

Result of table II can be get by the operation of program which is called neural network box in MATLAB. In which, data in format files in EXSEL are introduced into, and self-organizing feature map network with topological structure of hextop and link distance can be set up.

C. Based on the comparison and analysis between the two cluster results of regression principle

1) Based on the analysis of regression model of residual

From the table I and table II, there is a cluster containing quite a few data points. Now A stands for the third one of the four clusters of K-means, B stands for the fourth cluster of SOMF neural network.

the related concepts of residual has been introduced. Now take the residual and related statistics to evaluate regression model (1).

As for A, carry on linear regression with MATLAB, then regression equation can be obtained:

$$y = -148.1315 + 0.9985x_1 + 0.0243x_2 + 0.858x_3 + 0.8861x_4 \tag{5}$$

Figure1 is the residual of cluster A.

There is a singular point in the Figure.1, and the rest of data points are random distributed at both sides of the mean value of zero. In order to obtain a better regression equation than the equation (5), delete the tenth data point, namely （614， 115,85,63,629）, then make regression again. Regression equation can be obtained as followed,

$$y = -168.7704 + 1.0004x_1 + 0.0041x_2 + 0.9832x_3 + 0.9983x_4 \tag{6}$$

The coefficient of determination of equation (6), $r_1'^2$=0.9999851. Figure 2 is the residual of cluster A except the tenth point.

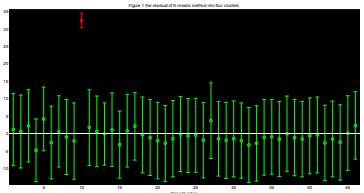

Figure 1. the residual of K-means method into four clusters (left)

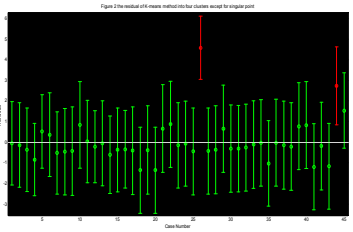

Figure 2. the residual of K-means method into four clusters except for singular point

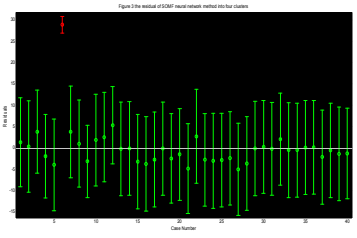

Figure 3. the residual of SOMF neural network method into four clusters (right)

From the coefficient of determination, the more changeable of the independent variable is the more closed relation between it and the dependent variable of equation (6) is. According to the Figure 2, the data points are random distributed at both sides of the mean value of zero. According to the formula (4), delete dozen points in the

middle of data points, then calculate the statistics of heteroscedasticity which is yfl_1 = 2.6561. Consulting F distribution table, $F(12,12) = 2.69$ can be get. $yfl < F(12,12)$. There dose not exist hetero- scedasticity of the residual. According to formulate (3), $dw_1 = 2.0040$ can be get. Consulting D-W distribution table, $dl = 1.29, du = 1.78$ can be get when n=45 , m=5. $du < dw < 4 - du$ .There dose not exist autocorrelation of the residual. Combined with the result of upper analysis, the equation (6) is more adaptable to cluster A.

As to cluster B, carrying on linear regression with MATLAB, regression equation can be obtained,

$$y = -135.6401 + 1.0002x_1 + 0.0988x_2 + 0.7424x_3 + 0.75x_4 \quad (7)$$

The Figure 3 is the residual of cluster B. There is a singular point in the Figure 3, namely （ 614 ， 115,85,63,629）, and the rest of data points are random distribution at both sides of the mean value of zero. Similarly, according to formulate (2.4), test heteroscedasticity and autocorrelation on the residual. Delete 10 points in the middle of data points and calculate statistics of heteroscedasticity which is yfl_2 = 4.2078. Consulting F distribution table, $F(15,15) = 2.40$ can be get. Because of $yfl_2 > F(15,15)$, there exists heteroscedasticity of the residual. According to formulate (4), $dw_2 = 1.8537$ can be obtained. Consulting D-W distribution table, $dl = 1.23$, $du = 1.79$ can be get when n=40，m=5. There dose not exist autocorrelation of the residual.

2) Based on the fitness of model analysis of these two cluster

Through the moderate analysis of the upper regression model, there is a singular point in Figure 3. It is deviation from the mean value of zero larger. Now from the reality aspect to consider this point and the mode, the sixth point residual value is 28.8298, and compared with the actual value, the relative error is 4.5%, less than 5%, which is acceptable in real engineering application. Therefore, in this paper, under the four cluster properties, the simulation results of K-means and SOFM neural network are not only feasible, but also good, and the following conclusions can be obtained:

- As for K-means method, there is more information of the 46 points in cluster A than the 40 points of SOFM neural network method in cluster B.
- It is the characteristics of K-means that fast calculating speed, less time-consuming and the classifier which need not to be trained under the C++ programming language. However, the classifier should be trained before clustering by SOFM neural network. The SOFM neural network training costs more time for the massive data.
- According to the cluster result (table 1), the oil quantity transported by oil pipeline is rarely centralized, the clustering results of K-means show that most of delivery of the total oil are almost concentrated in interval (397, 1400). We should pay attention to the delivery of the total oil out of the interval (397, 1400). According to the coefficient of equation (6), the storage variation at the end of month has less influence on the total oil. The total oil can be influenced most by the self-using oil and wastage. If the record data is found abnormal, you should check the oil input, wastage and self-using oil.

IV. CONCLUSIONS

According to clustering the data of oil pipeline in this paper, the evaluation system of clustering effect is built based on the regression analysis, which expounds the oil relation between the transportation and the fate, also expounds that the delivery of the oil is almost concentrated in interval (397, 1400). Some reference can be as reference to the pipeline maintenance, checking data and oil using quantity in this area.

K-means algorithm has been widely applied, and it has clear geometric significance and statistical significance[11].Through adjusting weight ratio of geometric distance and property distance, SOFM neural network not only can serve different cluster, but also will be more flexible [12].

Based on the above advantages, this paper adopts K-means and SOFM neural network to realize the cluster of oil pipeline data under four properties. The two kinds of clustering results are regressed by multiple regression. Finally the clustering results are evaluated by analyzing the fitness of model with residual. The result shows the cluster well. This paper only adopts single clustering algorithm to cluster the data. If adopting some clustering algorithm to realize complementary integration, whether the result will be better need further research.

ACKNOWLEDGMENT

This research was supported by: Special Fund of Basic Scientific Research Operating Expenses of Harbin Engineering University (HEUCF20111118), and NNSF of PR China under Grant #11002037 of Harbin Engineering University

REFFERENCES

[1] CHEN Lei. The Research and Application of Data Mining Technology for Dealing with Traffic Flow Data, Chang'an University [D],2009.

[2] MENG Qing-min; ZhANG Shi-cheng; WANG Li; HUANG Xia; ZhANG Jun-jing. Application of Data Mining to Postfracture Response Evaluation in Gas Field, Journal of China University of Petroleum (Edition of Natural Science) [J], 2008. 5, PP 165-169

[3] WANG Guang-hong; JING Ping.Survey of Data Mining, Journal of Tongji University [J], 2004.2, PP 246-252

[4] ZHU Yu-quan; YANG He-biao; SUN Lei. Data Mining Technology, Nanjing: Southeast University Press [M], 2006.

[5] Chinese forum of MATLAB. 30 Cases Analysis of Neural Network by MATLAB, Beijing: Beihang University Press [M], 2010.

[6] Haykin.S. Translation by YE Shi-wei and SHI Zhong-zhi. Neural Networks: A Comprehensive Foundation,2nd Edition, Beijing: China Machine Press[M],2004.

[7] HAN Li-qun. Theory, Design and Application of Artificial Neural Network,2nd Edition, Beijing: Chemical Industry Press [M],2007.

[8] LIAO Qin; HAO Zhi-feng; CHEN Zhi-hong. Data Mining and Mathematical Modeling, Beijing: National Defense Industry Press [M], 2010.

[9] LV Xiao-ling; XIE Bang-chang. Method and Application of Data Mining, Beijing: China Renmin University Press [M], 2009.

[10] ZHANG Xiao-di. Applied Regression Analysis, Zhejiang: Zhejiang University Press [M], 1991.

[11] JIANG Yuan; ZHANG Zhao-yang; CHOU Pei-liang; ZHOU Dong-fang. Clustering Algorithms Used in Data Mining, Journal of Electronics & Information Technology [J], 2005.4,PP 655-662

[12] JIAO Li-ming; LIU Yao-lin; REN Zhou-qiao. Spatial Points Clustering Based on Self-organizing Neural Networks and Its Application, Geomatics and Information Science of Wuhan University [J], 2008.2, PP 168-171

TABLE I. FOUR CLUSTERING RESULTS OF K-MEANS

category	firth	second	third	forth
The number of data	5	5	46	3
number of data	48 、 50 、 51 、 55 、 56、	52、 53、54、 57、58	1、2、3、4、5、6、7、 8、9、10、11、12、13、14、 15、16、17、18、19、20、 21、22、23、24、25、26、 27、28、29、30、31、32、 33、34、35、36、37、38、 39、40、41、42、43、44、 45、46、	47、49、59

TABLE II. FOUR CLUSTERING RESULTS OF SOFM NETWORK

category	firth	second	third	forth
The number of data	12	4	3	40
number of data	48、49、 50、51、 52、53、 54、55、 56、57、 58、 59	4、15、 46、47	5、 6、7	1、2、3、8、9、10、 11、12、13、14、16、17、 18、19、20、21、22、23、 24、25、26、27、28、29、 30、31、32、33、34、35、 36、37、38、39、40、41、 42、43、44、45