Research on Unbalanced Data Classification Algorithm for Complex Network Data Streams

Yingxue Cai, Jia Chen, Song Hu, Hui Hu, Sibo Huang, Zhaoquan Cai*
Huizhou University, Huizhou, 516000, China

Abstract: In order to improve the ability of fast processing and recognition of unbalanced data in complex network data flow, it needs to carry out fast classification and analysis of data. A fast classification algorithm based on association mining and fuzzy C-means clustering is proposed for unbalanced data flow in complex networks. Non-equilibrium data flow model of complex network data flow is constructed by nonlinear time series analysis method, and the delay scale characteristic parameters of non-equilibrium data flow in complex network data flow are extracted. Taking the extracted association rules as the feature quantity, the data clustering is processed, the fuzzy C-means classification algorithm is used to realize the optimal classification of the non-equilibrium data in the complex network data flow, and the convergence control of the classification center is carried out with the differential evolution method. The global convergence of the classification process is improved. The simulation results show that the proposed method has good accuracy and low error rate in the classification of unbalanced data in complex network data streams.

Keywords: Complex network; Data; Fuzzy C-means; Clustering; Classification

1. Introduction

In the age of non-balanced data flow of complex network data flow, it is necessary to deal with and analyze non-equilibrium data of complex network data flow regularly, and extract useful information features from unbalanced data flow of complex network data flow. The non-equilibrium data information of complex network data flow is analyzed for users, and the related information parameter services are mined from the unbalanced data of complex network data flow. With the rapid development of computer information processing technology, the speed and precision of non-balanced data processing of complex network data flow are required higher, in the complex network data flow non-balanced data information processing, the key link is to classify the data. The data classification is to analyze the data by mining the characteristic parameters with the same attributes in the unbalanced data of complex network data flow[1]. On the basis of data classification, an expert system and a non-balanced database of complex network data flow are established. In order to carry out related pattern recognition and diagnosis and analysis services, the research of data optimization and classification technology will be carried out in fault diagnosis, target recognition, etc[2]. Cloud storage database model construction and information retrieval have high application value.

Non-equilibrium data classification of complex network data flow is taken based on the analysis of time series of data flow. The characteristic parameters of non-equilibrium data flow in complex network data flow are extracted. The attribute feature selection and search of data classification, the calculation of data classification center, the automatic classification and recognition of data are carried out. The related data classification algorithms mainly include data classification method based on grid technology and fuzzy C-means classification method[3]. The fuzzy K-means classification method, the classification algorithm based on adaptive beamforming and so on, the fuzzy C-means and fuzzy K-means classification algorithms need to adjust the classification results repeatedly to optimize the classification with the expansion of data scale, the initial classification center is more sensitive[4].

In order to solve the above problems, this paper proposes a fast classification algorithm based on association mining and fuzzy C-means clustering for unbalanced data in complex network data streams. Non-equilibrium data flow model of complex network data flow is constructed by nonlinear time series analysis method, and the delay scale characteristic parameters of non-equilibrium data flow in complex network data flow are extracted. The fuzzy C-means classification algorithm is used to realize the optimal classification of the non-equilibrium data in the complex network data flow. Finally, the experimental analysis is carried out. The advantages of this method in improving the ability of data classification and reducing the rate of misclassification are demonstrated.

2. Nonlinear Time Series Analysis Model and Feature Parameter Extraction of Non-equilibrium Data in Complex Network Data Flow

2.1. Non-balanced data nonlinear time series analysis model for complex network data flow

In order to realize the automatic classification of non-equilibrium data in complex network data flow, the information flow model of non-equilibrium data in complex network data flow is constructed by using nonlinear time series analysis method[5]. The feature attribute analysis and classification center search are carried out by feature parameter extraction. The time series of non-equilibrium data in complex network data flow is a group of nonlinear time series. The method of nonlinear time series analysis can be used to analyze and classify the time series of non-equilibrium data in complex network data flow, and the univariate time series of non-equilibrium data series of complex network data flow can be constructed as follows. The data sample length is N, the data distribution is a scalar time series in the data sampling time period, and the classification characteristic attribute category of the data stream is set and set. The phase space reconstruction analysis method is used to deal with the nonlinear mapping of non-equilibrium data in complex network data flow. The information flow model of the time series of non-equilibrium data in complex network data flow is as follows:

$$x_n = x(t_0 + n\Delta t) = h[\mathbf{z}(t_0 + n\Delta t)] + w_n \tag{1}$$

Where, $h(.)$ is the similarity characteristic quantity contained in each sample of the unbalanced data time series of complex network data flow, the phase space reconstruction method is used[6]. The characteristic space distribution trajectory expression of the nonlinear time series of non-equilibrium data in complex network data flow is obtained as follows:

$$\begin{aligned}\boldsymbol{X} &= [\boldsymbol{x}(t_0), \boldsymbol{x}(t_0+\Delta t), \mathbf{L}, \boldsymbol{x}(t_0+(K-1)\Delta t)] \\ &= \begin{bmatrix} x(t_0) & x(t_0+\Delta t) & \mathbf{L} & x(t_0+(K-1)\Delta t) \\ x(t_0+J\Delta t) & x(t_0+(J+1)\Delta t) & \mathbf{L} & x(t_0+(K-1)\Delta t+J\Delta t) \\ \mathrm{M} & \mathbf{L} & \mathbf{O} & \mathbf{L} \\ x(t_0+(m-1)J\Delta t) & x(t_0+(1+(m-1)J)\Delta t) & \mathbf{L} & x(t_0+(N-1)\Delta t) \end{bmatrix}\end{aligned} \tag{2}$$

Where, $\boldsymbol{x}(t)$ is the sampling time series, J is the similarity correlation coefficient, m is the embedded dimension.

2.2. Feature extraction of association rule

The nonlinear time series analysis model of non-equilibrium data in complex network data flow is constructed, and the association rule feature extraction is carried out. $R_{u,v}$ represent the fuzzy set of non-equilibrium data attribute set of complex network data flow, $R_{u,v}$ is the cross-correlation function between the data characteristic vectors, then the cross-distribution model of the non-equilibrium data attribute set of complex network data flow is expressed as follows:

$$x_n = a_0 + \sum_{i=1}^{M_{AR}} a_i x_{n-i} + \sum_{j=0}^{M_{MA}} b_j h_{n-j} \tag{3}$$

Where, a_0 is the sample amplitude of the unbalanced data time series of the initial complex network data flow, and x_{n-i} is the scalar time series of the non-equilibrium data time series of the complex network data flow with the same mean value and variance[7]. The division of decision attribute values under homomorphism conditions may be repeated. The method of random frequency estimation can be used to calculate the expression of association rule eigenvalue as follows:

$$\begin{aligned}\hat{w}(n) = \arg\min_{k(n)\in \mathrm{K}} [\sum_{n=n_1}^{n_2} g(k(n), k(n+1)) \\ + \sum_{n=n_1}^{n_2} f(WVD(n, k(n)))]\end{aligned} \tag{4}$$

The following normalization methods are given:

$$m_i(d_l) = \frac{x_{c_i}^{d_l}}{2\sum_{l=1}^{k} x_{c_i}^{d_l} - 1} \quad m_i(\Theta) = \frac{\sum_{l=1}^{k} x_{c_i}^{d_l} - 1}{2\sum_{l=1}^{k} x_{c_i}^{d_l} - 1}, \; l = 1, 2, \mathbf{L}, k \tag{5}$$

The delay scale characteristic parameters of the unbalanced data flow in complex network data flow are extracted, and the extracted association rules are used as the feature quantities for data clustering.

3. Data Classification Optimization

3.1. Fuzzy C-Means clustering

The fuzzy c-means clustering model of non-equilibrium data in complex network data flow is constructed as:

$$\begin{cases} \dot{m}(t) = -Am(t) + Wg(p(t-s)) + u, \\ \dot{p}(t) = -Cp(t) + Dm(t-t), \end{cases} \tag{6}$$

Wherein:

$A = diag\{3,3,3\}, C = diag\{2.5,2.5,2.5\}, D = diag\{0.8,0.8,0.8\}$,

$W = \begin{bmatrix} 0 & 0 & -2.5 \\ -2.5 & 0 & 0 \\ 0 & -2.5 & 0 \end{bmatrix}$, Doppler frequency shift fuzzy search is used to smooth the non-equilibrium data of complex network data flow, and nonlinear autoregressive sliding time window is used to construct multi-layer spatial fuzzy classification center. The fuzzy C-means classification algorithm is used to search the initial classification center, and the finite data set is assumed as:

$$X = \{x_1, x_2, \mathbf{L}, x_n\} \subset R^s \tag{7}$$

According to the classification of attribute sets, x_i $i = 1, 2, \mathbf{L}, n$, samples are found in the data set, where the information gain of sample FG is as follows:

$$x_i = (x_{i1}, x_{i2}, \mathbf{L}, x_{is})^T \qquad (8)$$

Thus, the autocorrelation coefficient of the delay scale of the non-equilibrium data flow in complex network data flow is expressed as follows:

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(y)}} \qquad (9)$$

Where, r_{xy} is a dimensionless quantity in the upper expression.

3.2. Convergence control

The fuzzy C-means classification algorithm is used to realize the optimal classification of non-equilibrium data in complex network data flow. The convergence control of the classification center is carried out by using the differential evolution method to improve the global convergence of the classification process[8]. The Logistic chaotic mapping is defined as:

$$x_{n+1} = m x_n (1 - x_n) \qquad (10)$$

Where, $n = 1, 2, 3, \mathbf{L}$ $x \in [0,1]$ $m \in [0,4]$.

The training function is used to adjust the scale of the fuzzy classification center of the unbalanced data in complex network data stream. The delay scales in t and $t+t$ in the retrieval of the classification center are as follows:

$$V = \{v_{ij} \,|\, i = 1, 2, \mathbf{L}, c, j = 1, 2, \mathbf{L}, s\} \qquad (11)$$

Where, V_i is the weight of adjacent data points to the classification center disturbance, for the first vector of the non-equilibrium data series of the complex network data flow, the Logistics chaotic map is used for the difference perturbation, and each data point is regarded as a possible classification center. The stable periodic solution of the classification center is obtained as follows:

$$U = \{m_{ik} \,|\, i = 1, 2, \mathbf{L}, c, k = 1, 2, \mathbf{L}, n\} \qquad (12)$$

$$\sum_{i=1}^{c} m_{ik} = 1, k = 1, 2, \mathbf{L}, n \qquad (13)$$

Combined with complex network data flow unbalanced data classification target function. The Logistics chaotic differential evolution method is used to suppress the classification center disturbance:

$$D(x_i, A_j(L)) = \min\{D(x_i, A_j(L))\} \qquad (14)$$

Where, $x_i \in w_k$, At this time, the classification center obtains the optimal solution. The chaos perturbation is introduced into the example of evolutionary classification cluster, and the initial membership matrix is calculated, and the average value is taken as the average value of the new classification attribute feature vector:

$$C(l) = \sum_{j=1}^{k} \sum_{k=1}^{n_j} (\| x_k^j - A_j(L) \|)^2 \qquad (15)$$

Thus, the optimal classification and recognition of the unbalanced data in complex network data flow are realized.

4. Simulation Experiment and Result Analysis

In order to verify the performance of this algorithm in the realization of fast and automatic classification of non-equilibrium data in complex network data flow, the simulation experiment is carried out, and the experiment is based on Matlab simulation software. The experimental data comes from two complex network data streams, the KDDP201 large network database simulation data set, including two 120MB partitions, which are actual data sets, including 60MB scale partitions. The original data waveform is shown in figure 1.

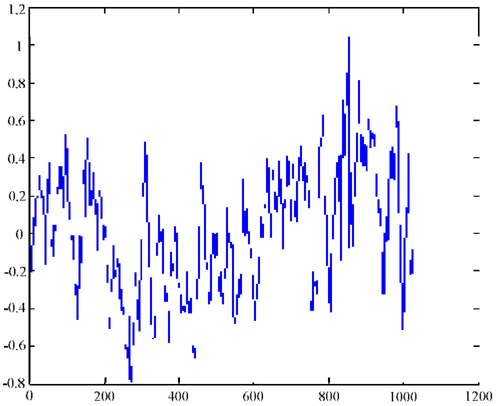

Figure 1. Complex network data waveform

This method is applied to the classification of unbalanced data in complex networks, and the classification results are shown in Figure 2.

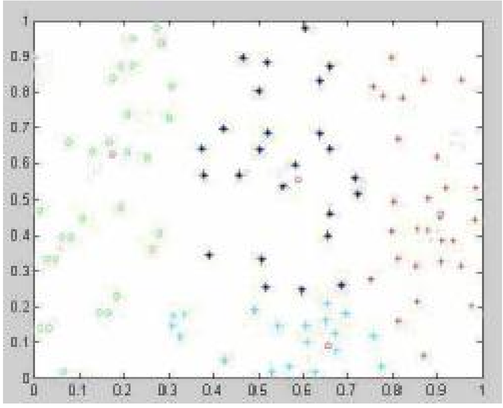

Figure 2. Data classification output

Figure 2 shows that this method is used to classify data, it can effectively realize big data clustering recognition of different attributes and improve the accuracy of data classification. The accuracy of different classification methods is tested, and the comparison results are obtained as shown in figure 3.

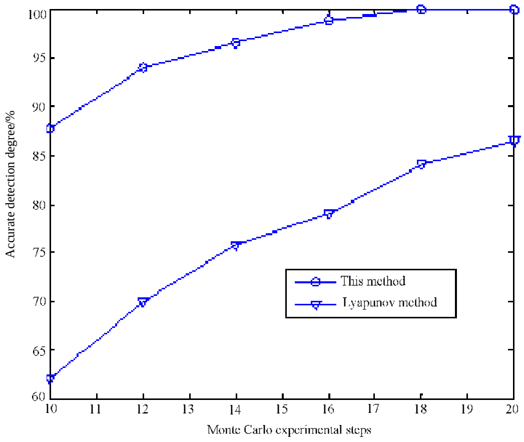

Figure 3. Accuracy comparison

Figure 3 shows that the proposed method has high accuracy and low error rate in the classification of unbalanced data over complex network data streams.

5. Conclusions

In this paper, a fast classification algorithm based on association mining and fuzzy C-means clustering is proposed for unbalanced data flow in complex networks. Non-equilibrium data flow model of complex network data flow is constructed by nonlinear time series analysis method, and the delay scale characteristic parameters of non-equilibrium data flow in complex network data flow are extracted. The data clustering is processed, the fuzzy C-means classification algorithm is used to realize the optimal classification of the non-equilibrium data in the complex network data flow, and the convergence control of the classification center is carried out with the differential evolution method. The global convergence of the classification process is improved. The simulation results show that the proposed method has good accuracy and low error rate in the classification of unbalanced data in complex network data streams, the data clustering performance is improved.

4. Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61772225);
the Foundation for Distinguished Young Talents in Higher Education of Guangdong (No. 2015KQNCX153); Science and Technology Program of Huizhou (No. 2015B010002002, No. 2016X0431046, No. 2016X0434049, No. 2016X0432047, No. 2017c0406022, No. 2017c0407023, No. 2017c0414030).

References

[1] LING Conggang, WANG Hongzhang. Optimization research on differential evolution algorithm and its application in clustering analysis[J]. Modern electronic technology, 2016, 39 (13): 103-107.

[2] LI Mu-dong, ZHAO Hui, WENG Xing-wei, HAN Tong. Differential evolution based on optimal Gaussian random walk and individual selection strategies[J]. Control and Decision, 2016, 31(08): 1379-1386.

[3] PATEL V M, NGUYEN H V, and VIDAL R. Latent space sparse and low-rank subspace clustering[J]. IEEE Journal of Selected Topics in Signal Processing, 2015, 9(4): 691-701.

[4] Sun Li-juan, Chen Xiao-dong, Han Chong, Guo Jian. New Fuzzy-Clustering Algorithm for Data Stream[J]. JEIT, 2015, 37(7): 1620-1625.

[5] XING Changzheng, LIU Jian. Evolutionary data stream clustering algorithm based on integration of affinity propagation and density[J]. Journal of Computer Applications, 2015, 35(7): 1927-1932.

[6] BI Anqi, WANG Shitong. Transfer Affinity Propagation Clustering Algorithm Based on Kullback-Leiber Distance[J]. JEIT, 2016, 38(8): 2076-2084.

[7] LIU Jun, LIU Yu, HE You, SUN Shun. Joint Probabilistic Data Association Algorithm Based on All-neighbor Fuzzy Clustering in Clutter[J]. JEIT, 2016, 38(6): 1438-1445.

[8] WU Hong-hua,MU Yong,QU Zhong-feng,DENG Li-xia. Similarity and nearness relational degree based on panel data[J]. Control and Decision, 2016, 31(03): 555-558.