Analysis of Classification and Recognition Algorithm for Imbalanced Data Fragment in Large Database

Rongguo Li*, Xiaoning Liu, Jian Xu, Cuilan Zou
Laiwu Vocational and Technical College, Laiwu, 271100, China

Abstract: In order to improve the ability of imbalanced data fragment classification and recognition in large database, and realize the optimization retrieval of large database, an imbalanced data fragment classification and recognition algorithm is proposed based on fuzzy feature clustering. Fuzzy adaptive spectral feature extraction method is used to extract the feature of imbalanced data fragment information in large database, and big data mining is carried out in combination with fuzzy directivity clustering method. The association rule scheduling method is used for the balanced scheduling of imbalanced data in large databases, vector quantization coding is carried out for the fragments of imbalanced data, and the feature extraction and information retrieval are carried out according to the quantization coding results. A fuzzy C-means clustering algorithm is used to classify and recognize the fragments of imbalanced data in large database. The simulation results show that the algorithm has high accuracy and low misclassification rate, and the ability of accessing and retrieving large databases is improved.

Keywords: Large database; Data clustering; Classification and identification; Scheduling; Quantization coding

1. Introduction

In the large digital multimedia information database, a large amount of information, such as sound, text and imbalanced data fragments are stored in large multimedia databases through cloud storage architecture, in order to improve the retrieval ability of multimedia information databases. It is necessary to make accurate data mining and classification of imbalanced data fragments in large database. With the development of processing technology for imbalanced data fragment in large digital database, the methods for quantization coding and feature extraction of imbalanced data fragment in large database are used to compress and classify imbalanced data fragments, so as to improve the ability to process and classify imbalanced data fragment information in large database. In large multimedia database, imbalanced data fragment retrieval and classification are needed. Computer vision reconstruction and modeling of imbalanced data fragment features in large database, accurate accessing to multimedia database information features and analysis of imbalanced data fragments are realized [1].

The key to classify imbalanced data fragments in large database lies in the accurate mining and feature extraction of the key feature points in the imbalanced data fragments. Combined with the segmentation of large database imbalanced data fragments and the corner detection algorithm of large database imbalanced data fragments, data mining technology is used to realize the classification and retrieval of large database imbalanced data fragments. In the traditional methods, the classification methods of large database imbalanced data fragments are mainly based on fuzzy C-means clustering method, K-means clustering method and BP neural network classification method, etc. [2, 3]. By extracting feature points and analyzing information from imbalanced data fragments in large database, the method achieves accurate classification of objects, and good classification effect, but it has a large computational overhead. The real-time performance of the fragmentation feature classification in network multimedia database is not good [4]. In order to solve the above problems, this paper proposes a classification and recognition algorithm for imbalanced data fragment in large database based on fuzzy feature clustering, which encodes imbalanced data fragments by vector quantization. The feature extraction and information retrieval are carried out according to the quantization coding results, and the fuzzy C-means clustering algorithm is used to realize the classification and recognition of imbalanced data fragments. Finally, the simulation results show that the proposed method can improve the classification and recognition ability of imbalanced data fragments in large database.

2. Imbalanced Data Mining and Preprocessing of Large Database

In the imbalanced data mining and preprocessing of large database, the feature extraction method based on fuzzy adaptive spectral is used to extract the feature of fragment information of imbalanced data in large database, and big data mining is carried out by combining fuzzy directivity clustering method [5]. The mean square error $\hat{x}(s_j)$ of each pixel point in the imbalanced data fragment vector quantization coding is obtained. The directivity clustering center of the imbalanced data fragment information in large database is expressed as:

$$\hat{x}(s_j) = \frac{1}{\|s_j\|}\sum_{x_i \in s_j} x_i \quad (1)$$

Where, $\|s_j\|$ represents the similarity of imbalanced data fragments in s_j. The hierarchical matching quantization coding model is established and the feature information output from the vector information fusion center is obtained as follows:

$$F(k_1,k_2) = \sum_{n1n2} f(n_1,n_2) W_{N_1}^{k_1n_1} W_{N_2}^{k_2n_2} = A_F(k_1,k_2) e^{j\theta_F(k_1,k_2)} \quad (2)$$

$$G(k_1,k_2) = \sum_{n1n2} g(n_1,n_2) W_{N_1}^{k_1n_1} W_{N_2}^{k_2n_2} = A_G(k_1,k_2) e^{j\theta_G(k_1,k_2)} \quad (3)$$

Where, $A_F(k_1,k_2)$ and $A_G(k_1,k_2)$ are mutual bit correlation functions of imbalanced data fragments in large database. Thus, the regional distribution functions for retrieval of imbalanced data fragments in multimedia database are obtained as follows:

$$E^{cv}(c_1,c_2) = \mu \cdot Length(C) + \nu \cdot Area(inside(C)) + \lambda_1 \int_{inside(C)} |I - c_1|^2 dxdy + \lambda_2 \int_{outside(C)} |I - c_2|^2 dxdy \quad (4)$$

Wherein, c_1 and c_2 represent the characteristic coefficients of the feature distribution of big data, and $Length(C)$ represents the normalized length of the matched window. $Area(inside(C))$ represents the contour reference point and the gradient modulus of the imbalanced data fragments in the large database. λ_1 and λ_2 denote the correction weight coefficient of the imbalanced data fragment vector quantization coding in the large database, all the constant of greater than 0. After the above processing, the data fragment vector quantization coding for imbalanced data fragment in large database is realized, which provides the data feature input basis for the classification and information fusion of imbalanced data fragment in large database.

The scale values of the edge contour line segment of the imbalanced data fragment in large database are obtained as follows:

$$s(k) = \phi \cdot s(k-1) + w(k) \quad (5)$$

Where

$$\phi = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad w(k) = \begin{pmatrix} N(0,\sigma_{\theta(k)}) \\ 0 \\ N(0,\sigma_{x(k)}) \\ 0 \\ N(0,\sigma_{y(k)}) \end{pmatrix} \quad (6)$$

The median filter is used to detect the correlation feature of big data, initialize the correlation feature matching filter, define K=0, and solve the variable component for the middle point of the t frame of imbalanced data fragment. Data mining method is used to filter corner points of imbalanced data fragments in large database, the matching function of feature points is obtained as follows:

$$s(k \mid k-1) = \phi \cdot s(k-1 \mid k-1) \quad (7)$$

The feature space characteristic locus of imbalanced data fragment in large database (x, y) is obtained by quantization fusion of imbalanced data fragment in large database. Because $s(k) = [\theta(k), \Delta x(k), \Delta y(k)]$, based on the feature matching of imbalanced data fragment output data in large database, the iterative process of large database imbalanced data fragment quantization fusion is described as follows:

$$t(x) = 1 - \min_{c\in\{r,g,b\}} \left(\min_{y\in\Omega(x)} \left(\frac{I^c(y)}{A} \right) \right) \quad (8)$$

$$U(x) = 1 - t(x) = \min_{c\in\{r,g,b\}} \left(\min_{y\in\Omega(x)} \left(\frac{I^c(y)}{A} \right) \right) \quad (9)$$

Where, $I^c(y)$ is the correlation eigenvalue of the fragment corner of imbalanced data, A is the amplitude and $\Omega(x)$ is the neighborhood space of the imbalanced data fragment in large database. The feature points extracted from large database are used as data input to classify and process the fragments of imbalanced data in large database by quantization and fusion feature point data mining.

3. Data Classification and Recognition Algorithm Optimization

The data mining of feature points is taken, the improved design of classification algorithm for imbalanced data fragment in large database is carried out, and the fuzzy C-means clustering algorithm is used to classify and retrieve the imbalanced data fragment feature. Assuming that the time series of feature point data is $x(t)$, $t = 0,1,\cdots,n-1$, the initial window of fuzzy C-means clustering is defined as:

$$u = [u_1, u_2, \mathbf{L}, u_N] \in R^{mN} \quad (10)$$

The maximum gradient difference pixels are obtained when searching matching points for the reference points of fragment classification feature points in large database imbalanced data:

$$AVG_X = \frac{1}{m \times n} \sum_{x=1}^{n} \sum_{y=1}^{m} |G_X(x, y)| \quad (11)$$

Where, m and n are the maximum series of windows and the width of the time window. The beam directivity information of the classification feature points of imbalanced data fragments in large database is extracted [6]. According to the sub-pixel offset information of imbalanced data fragments to be matched, the weighted vector of output is obtained:

$$x(t) = (x_0(t), x_1(t), \cdots, x_{k-1}(t))^T \quad (12)$$

A 1 × N window is used to search the clustering center of the large database imbalanced data fragment classification. The space distance of the weighted vector ω_j is calculated, which is expressed as:

$$d_j = \sum_{i=0}^{k-1} (x_i(t) - \omega_{ij}(t))^2, \quad j = 0,1,\cdots,N-1 \quad (13)$$

Taking the extracted feature points as the data input, the LGB vector quantization coding is used to divide the cluster center of the imbalanced data fragments in large database, and the $l_{\max}$ level matching window is expressed as:

$$U = \{\mu_{ik} \,|\, i = 1,2,\mathbf{L},c, k = 1,2,\mathbf{L},n\} \quad (14)$$

The priori knowledge filtering model between two matching windows is calculated. The initial state of imbalanced data fragment retrieval is $x^i(0) = \hat{x}^i(0)$. By fuzzy C-means clustering, the objective function of imbalanced data fragment optimization classification is obtained:

$$J_m(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}{}^m (d_{ik})^2 \quad (15)$$

According to the feature points extracted from the imbalanced data mining model in the database, the measure distance $(d_{ik})^2 = \|x_k - V_i\|^2$ of the fragment data sample V_i is obtained when the clustering center satisfies:

$$\sum_{i=1}^{c} \mu_{ik} = 1, k = 1,2,\mathbf{L},n \quad (16)$$

At this point, in multimedia database, the maximum value of search objective function is:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (d_{ik} / d_{jk})^{\frac{2}{m-1}}} \quad (17)$$

$$V_i = \frac{\sum_{k=1}^{m} (\mu_{ik})^m x_k}{\sum_{k=1}^{n} (\mu_{ik})^m} \quad (18)$$

With the design of the above algorithms, the extracted feature points are taken as the data input, and the fuzzy C-means clustering algorithm is used to realize the data mining and the imbalanced data fragment classification.

4. Simulation Experiment and Result Analysis

In order to test the application performance of the proposed algorithm in the classification and recognition of imbalanced data fragments in large database, the simulation experiment is carried out. The simulation experiment is based on the Matlab 2010 programming platform, and the imbalanced data fragment in the large database is to be classified. The initial resolution of the chip is 520 × 308, the width of the sliding window normalization time is 1.4 s, and the scale of the characteristic decomposition of the fragment data is $\sigma^{(n)}$ (1, 2,…,n)=0.345. According to the above simulation environment and parameter setting, the classification and simulation analysis of imbalanced data fragments in large database are carried out. The results of 2D and 3D feature mining of imbalanced data fragments are shown in figure 1.

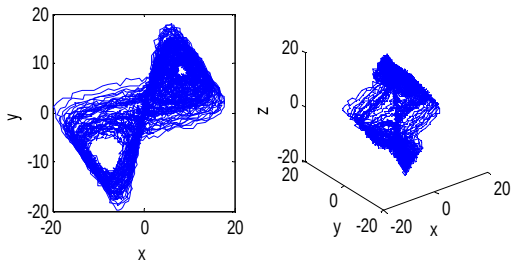
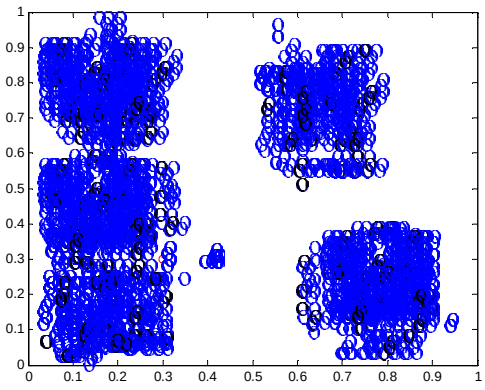
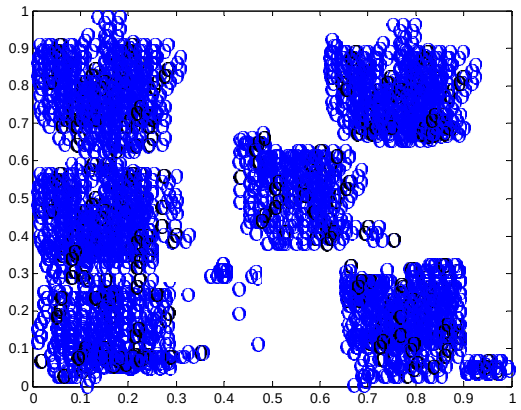


Figure 1. Mining results of 2D and 3D features of imbalanced data fragments in large databases

The data mining result of figure 1 are taken as the training set, the imbalanced data fragment classification in large database is carried out, and the classification result is shown in figure 2.



(a) Database1



(b) Database2

Figure 2. Data classification and recognition results

Figure 2 shows that the proposed method has better feature clustering and higher classification accuracy. The error rate of different classification methods for big data classification is tested, and the comparison results are shown in figure 3.

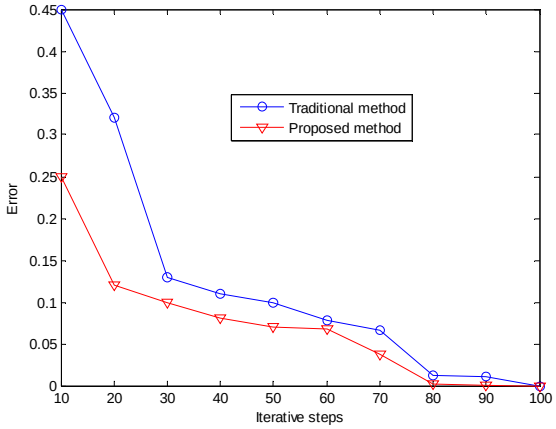


Figure 3. Shows that the classification error rate of this method is low, indicating that the classification and recognition of data fragments is more accurate

5. Conclusions

In this paper, a recognition algorithm for imbalanced data fragment in large database based on fuzzy feature clustering is proposed. Fuzzy adaptive spectral feature extraction method is used to extract feature of imbalanced data fragment information in large database. Combining fuzzy directivity clustering method, big data mining is carried out, and association rule scheduling method is used for large database. Imbalanced data in large database is made balanced scheduling, and vector quantization coding for imbalanced data fragments in large data base is carried out. Based on quantization coding results, feature extraction and information retrieval is made, and fuzzy C-means clustering algorithm are used to implement the algorithm. The simulation results show that the algorithm has high accuracy and low misclassification rate, and the ability of accessing and retrieving large databases is improved. It has good application value in database management.

6. Acknowledgement

Shandong Vocational Education Teaching Reform Research Project "Ai empowerment, professional resonance, fused and reconstruction" the construction path of the Future Generation of information technology specialty group in Higher Vocational Education and Practice, Project No. 2019035.

References

[1] Pan Ying, Tang Yong, and Liu Hai. Access control in very loosely structured data model using relational databases. Acta Electronica Sinica. 2012, 40(3), 600-606.

[2] Suzuki Taizo and Kudo Hiroyuki. Two-dimensional non-separable block-lifting structure and its application to M-channel perfect reconstruction filter banks for lossy-to-lossless image coding. IEEE Transactions on Image Processing. 2015, 24(12), 4943-4951.

[3] Jiang Junzheng and Zhou Fang. Iterative design of two-dimensional critically sampled MDFT modulated filter banks. Signal Processing. 2013, 93(11), 3124-3132.

[4] Arjuna Madanayake and Leonard Bruton. 2D space-time wave-digital multi-fan filter banks for signals consisting of multiple plane waves. Multidimensional Systems and Signal Processing. 2014, 25(1), 17-39.

[5] Omar Rafik Merad Boudia, Sidi Mohammed Senouc, Mohammed Feham. A novel secure aggregation scheme for wireless sensor networks using stateful public key cryptography. Ad Hoc Networks. 2015, 32(C), 98-113.

[6] Wang Jie, Lu Jianzhu, Zeng Xiaofei. Data aggregation scheme for wireless sensor network to timely determine compromised nodes. Journal of Computer Applications. 2016, 36(9), 2432-2437.