Analysis of the Algorithm for the Classification and Recognition of the Unbalanced Data Fragments in Large Database

Yang Jian

College of Artificial Intelligence, Chongqing Three Gorges Vocational College, Wanzhou, Chongqing, 404155, China

email: 39577932@qq.com

Keywords: Big Database, Imbalance, Data, Fragmentation, Classification, Recognition, Convergence Degree, Attribute Value

Abstract: In the traditional algorithm of data fragment recognition, the influence of fragment's own attributes is often ignored, which leads to the reduction of the accuracy of data fragment recognition. Therefore, we propose a large database based on lrfu strategy and association analysis to identify and classify the unbalanced data segments. Algorithm. Based on the resampling algorithm, the unbalanced data segments are upsampled; the zero filling filter coefficients based on the sampling results and the convolution calculation of the unbalanced data segments based on the filter banks, as well as the reconstruction segments used to obtain the segment feature sequences. The similar partition linear method is used to deal with the highly integrated and unevenly distributed data segments with sequences to achieve segment classification. Transform the objective function, and get the suitable function by combining the extended matrix. The matching function uses lrfu strategy to schedule, correlation analysis to determine the attribute value of the integrated unbalanced data segment, and realize the recognition of the unbalanced data segment. The experimental results show that if the improved method is used to classify and recognize the non-uniform data segments, its suitability and recognition accuracy are high, and it has specific advantages.

1. Introduction

With the rapid development of computer science, industry data burst out. The world's industries and industries have entered the era of big data. Data resource has become an important strategic resource of the country and enterprises [1]. However, in the process of big data and system use, it is necessary to analyze several suspicious files or files whose data recovery, computer network security and other fields are often unknown. This leads to an increase in unbalanced data and large pieces of data. The existing big data theory and application technology can basically solve the practical use of big data, but the existing unbalanced data segment recognition method puts forward a new problem that is difficult to solve the data segment distribution [2]. New principles, definitions, tools and algorithms are needed to improve the accuracy of original data fragment recognition. In this paper, a large-scale unbalanced data segment classification algorithm based on the combination of strategy and association analysis is proposed, and the advantages of this method are verified by experiments.

2. Analysis of Fragment Feature Sequence of Unbalanced Data in Large Database

2.1. Data Fragment Sampling

The research of unbalanced data can be divided into data algorithm based classification algorithm and prospect based prediction. The classification algorithm reduces the data fragment of unbalanced data by preprocessing uneven data and adjusting the distribution of training data [3]. At present, the effect comparison and correlation between the two methods are not clear, but generally speaking, the method based on algorithm angle is more accurate, which is a relatively simple and

  DOI: 10.25236/cseem.2020.164

effective research method. Generally speaking, before the training and learning of classification and recognition, it is necessary to resample the sample space according to different evaluation indexes to obtain better classification and recognition results. By transforming the training set D→D1 to resample the nonuniform data, the classifier f is built on the new training set D_1[4]. Sampling is the process of finding samples, and its function is to improve the performance of classifier. The resampling algorithm improves the unbalanced data segment by constructing a new data structure. Small (organic minimum o Fig. 1 shows an example of a small algorithm.

In the graph, black dots represent most samples, triangles represent a few samples, and squares represent a few newly generated composite samples. When dealing with a small number of uneven data fragment sampling points x 1, the same type of K should be calculated first. In HITS algorithm, the value of K is usually chosen as 5 or 10. In the kn UU x 1 set, randomly select a majority of sample points x 2, the difference of attribute J corresponding to x 1 and x 2 is expressed as: Rand [0,1] represents the pseudo-random number between 0 and 1. In this algorithm, the difference diff J is multiplied by the number randomly generated in the interval [0,1], and then the corresponding attribute value in the original attribute vector is added with X 1J to get the new attribute value of the unbalanced data sample. M attribute values $(f_{1j},\cdots, f_{1m})$ .Will be obtained In order to generate new samples of some newly generated unbalanced data segments f_{1j} .

$$f_{1j} = x_{1j} + diff_j \times rand[0,1] = x_{1j} + (x_{2j} - x_{1j}) \times rand[0,1] \tag{1}$$

2.2. Fragment Feature Sequence Extraction

Based on the samples collected from non-uniform data segments, the filter coefficients are added to zero. Then, the convolution calculation is performed on the uneven data segment to eliminate the interference of the lower sampling difference and the reconstruction difference in the decomposition process, so as to maintain the main characteristics of the unbalanced data segment [5]. In order to achieve the purpose of filtering unbalanced data segments. Assuming H (k) is a low-pass filter, the j-th feature of a (n) is shown as follows:

Where n is the unbalanced data quantity and K is the cut-off frequency signal of low-pass filter, and the calculation formula of J filter coefficient $D_J(n)$ can be calculated by high pass filter g (n).

In the equation, δ (n) is a sequence of unbalanced data segments. When δ (n) ≈ 0, the feature sequence of unbalanced data segments cannot be obtained, and must be returned to resampling. $If\delta(n) = 0$ then no filtering is needed and the feature sequence cannot be obtained directly. When δ (n) is 0, even data fragment sequences a and B match. In addition, the feature points are obtained by surfing (synthesizing minority oversampling) algorithm, and the feature points are a pair, which are matched to get the feature sequence of uneven data segments.

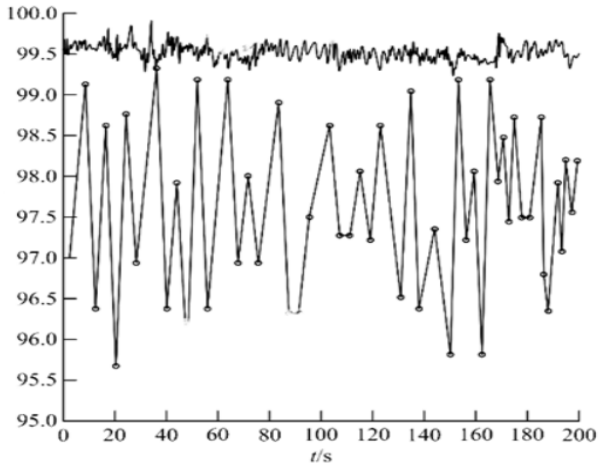

Figure 1 Traditional methods to identify the fragment fitness of unbalanced data

3. Research on Classification and Recognition of Unbalanced Data Fragments

3.1. Fragment Classification

After obtaining an imbalanced feature sequence of data fragments, the fragments need to be classified. In the classification process, we must pay attention to the impact of classification points on fragment classification ie, the independence and convergence of fragment . After confirming two, classify them. In order to classify unbalanced data fragments, it is divided into two intervals, and the independence of the two intervals x^2 is as follows.

$$x^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

In the equation, n_{ij} is the number of objects belonging to the j-th determination class in the i-th interval[6]. Schematic diagram of the gradient distribution. The unevenness of the segmented attributes of the uneven data leads to the diversification of the data gradient. By extracting the characteristics of gradient convergence, the convergence of village data fragments is described. The initial directions of the data gradient arrows are different due to the uneven properties of the uneven data fragments. The arrow indicates the state before the gradient of the data converges. The length of the vector g'i from the data string p is one[7]. In the calculation of the two adjacent features on the segment feature sequence, the inclination and duration of the assassinated straight line segment, the discriminative linear method in high convergence should be used to respond to the straight line segment, the overall feature The distance and e of the sequence will also be calculated.

3.2. Debris Identification

Normalize the classified imbalanced data segments. In this way, the influence of different characteristics of the data segments is eliminated, and all unbalanced data segments are classified as the calculation formula is as follows.

$$x_i = \frac{2x_i - (x_{\min} - x_{\max})}{x_{\max} - x_{\min}} \tag{3}$$

In the equation, x'i is an unbalanced piece of data normalized by the corresponding input value. Based on this, the lrfu strategy is used for scheduling. In order to achieve the identification value of the inconsistent data segments, the correlation analysis method is used to determine the attribute values of the uneven data segments. When the characteristics of the imbalanced data segment are insufficient, a replacement process is required to calculate the attribute value of the imbalanced data segment with the smallest feature crf t base (b).

In summary, by using a similar partitioning linear method to handle uneven data fragments with high density, classification of uneven data fragments can be achieved[8]. The fitness function is obtained by transforming the objective function and combining the expansion matrix. The LRFU strategy is used to schedule the matching function, and the correlation analysis method is used to determine the attribute values of the uneven data fragments, so as to realize the identification of the data fragments.

3.3. Experimental Results and Analysis

In order to verify the effectiveness and feasibility of the improved algorithm of unbalanced data segment recognition in large-scale database, the matching degree of index recognition and the accuracy of unbalanced data segment are compared. The accuracy is calculated as follows.

4. Experimental Environment Setting

In order to analyze the effectiveness of the algorithm of large database unbalanced data segment recognition, an experimental environment is needed. The comprehensive replication information database of the company was selected. In the database, China's replication information data is complex, various types and huge amount of data[9]. 600 data groups were selected for the experiment. Data grouping has multiple data segments, distribution and different data gradient

directions. According to the above experimental environment, the fitness of bump data and the accuracy of segment classification and recognition of uneven data are tested.

With the increase of recognition time, there is a big error between the matching degree and the best matching value. The minimum value is about 95. The highest value of 5 is 99. 2. None of them reach the optimal 99. 6. The improved method, when the recognition time increases, its fitness value gradually increases, close to the best fit value, in the highest place, it is close to the best fit value 99. 4 is a minimum of about 98. 5. Compared with the traditional method of cognition, it has specific advantages.

When the fitness value is uncertain, in the existing methods, with the increase of fitness, the recognition accuracy will gradually improve, but the rise speed is slow, because of the rise or fall and other changes, it is not stable, up to 60. After using the improved method, with the increase of fitness, the recognition accuracy is improved, without too much rise and fall, achieving high-speed, up to 97. 1%, 36.3% higher than the traditional method.

5. Concluding

In this study, to reduce the low fitness problem and the accuracy of the previous recognition methods, a large data segment classification and recognition algorithm based on the combination of lrfu strategy and association analysis is proposed. The experimental results show that, in order to classify and recognize the non-uniform data, the improved method is more suitable and accurate than the existing method.

Acknowledgements

This research has been financed by Major projects of teaching reform in Chongqing Higher Education in 2019 of the Chongqing Education Commission "reform and practice of talent training mode of integration of production, competition and education" of Vocational Education in the Three Gorges Reservoir Area "(191042).

References

[1] Solmaz Bagherpour, Angela Nebot, Francisco Mugica. (2018). Wrapper-based Fuzzy Inductive Reasoning model identification for imbalance data classification. 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE.

[2] Andrea Mannini, Mary Rosenberger, William L Haskell,. (2017). Activity Recognition in Youth Using Single Accelerometer Placed at Wrist or Ankle. Med Sci Sports Exerc, vol. 49, no. 4, pp. 801-812.

[3] Holton, Sarah, McCann, Fergal. (2017). Sources of the small firm financing premium: evidence from euro area banks. Social Science Electronic Publishing.

[4] C. Yao, Y. Zhang, H. Liu. (2017). APPLICATION OF CONVOLUTIONAL NEURAL NETWORK IN CLASSIFICATION OF HIGH RESOLUTION AGRICULTURAL REMOTE SENSING IMAGES. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2/W7, pp. 989-992.

[5] Eisenschmidt, Jens, Kedan, Danielle, Schmitz, Martin,. (2017). The Eurosystem’s asset purchase programme and TARGET balances. Social Science Electronic Publishing.

[6] Kyle Lavin, Dimitry S. Davydow, Lois Downey. (2017). Effect of Psychiatric Illness on Acute Care Utilization at End of Life From Serious Medical Illness. J Pain Symptom Manage, vol. 54, no. 2.

[7] Van Nghia Luong, Van Son Le, Van Ban Doan. (2018). Fragmentation in Distributed Database Design Based on KR Rough Clustering Technique. Context-Aware Systems and Applications, and

Nature of Computation and Communication.

[8] Isaac Fernández-Varela, Elena Hernández-Pereira, Diego Álvarez-Estévez,. (2017). Combining Machine Learning Models for the Automatic Detection of EEG Arousals. Neurocomputing.

[9] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu,. (2017). Clustering-based undersampling in class-imbalanced data. Information Sciences, vol. 409.