Analysis of Large Database Unbalanced Data Fragment Classification and Recognition Algorithm

Peng Mei, Tan Zerong, Hu Bibo

Guangzhou College of Technology and Business, Guangzhou, Guangdong, China

Keywords: Big data era, Data processing, Unbalanced data fragment classification and recognition algorithm, Accuracy.

Abstract: In view of the current development of big data era and the data processing of various industries, this paper proposes an unbalanced data fragment classification and recognition algorithm suitable for the current big data background. Firstly, the process of classification identification is explained in detail, and the problems existing in it are analyzed. The algorithm is described in detail, and the unbalanced data fragments are classified and identified. Experiments show that the classification and recognition algorithm can effectively improve the accuracy and lay a foundation for further research and development in this field.

1. The Research Background

1.1 Literature review

In traditional data fragmentation algorithms, people tend to ignore the influence of the fragments themselves on the algorithm, which will reduce the accuracy of data recognition. Therefore, the classification analysis of association analysis and LRFU analysis strategies is proposed. algorithm. The algorithm is mainly divided into three steps: firstly, the data fragments are further up-sampled by resampling, and then the coefficients are complemented by zeros according to the sampling result, and then convolution calculation is performed. Finally, the fragment feature sequence is obtained according to the calculation result. Experiments have shown that this method has a high fitness value and accuracy (Wen and Lan, 2018). In addition, the data fragmentation technology of traditional ships cannot achieve the ideal data recovery tracking effect. Therefore, a new method for classification and identification of unbalanced data fragments is obtained. The information acquisition and data visualization methods can be used to visualize ship information. This method effectively improves the tracking level of reverse engineering recovery data (Liu and Yang, 2018). The fragment classification and recognition algorithm can also infer the attribute of the user's network behavior, which is of great significance in the fields of platform service promotion, personalized recommendation and marketing. At present, the main work is to infer attribute estimation by tracking the user's network behavior through social behaviors, browsing behaviors, and the like. However, users of realistic commentary websites are mostly anonymous, which results in fragmentation of behavioral data, uneven information content, and low value. According to environmental information, object information, and other auxiliary user feature modeling, the problem of sparse data and imbalance is reduced to some extent, and an improved algorithm of algorithm is proposed (Liu and Sun, 2017). Data fragmentation forensic analysis has become an important way to obtain data evidence in the computer field. A data fragment type analysis is proposed for data fragmentation forensics, and it is analyzed in turn, which proves that this algorithm has higher recognition rate and accuracy, and achieves good recognition effect (Fu and Jing, 2015). Han Ying proposed a data defragmentation algorithm that can improve the data recovery speed in the deduplication backup system. It turns out that in this system, the use of this data sorting algorithm speeds up data recovery and improves system performance (Han, 2018).

1.2 Research purposes

With the continuous development of economy and technology, computer science has also

  DOI: 10.25236/scmc.2019.072

developed rapidly. The data of various industries has proliferated, and all countries in the world have entered the era of big data. Data resources have become a significant one for both enterprises and the country. Strategic resources. However, in the era of big data, in the process of its use, data recovery, data forensics and computer network security often need to analyze some unknown files or suspicious files, which exacerbates the growth of unbalanced data. A lot of data fragmentation. At present, the problem that needs to be solved in this field is how to classify and identify these unbalanced data. The existing big data application technology and theory can basically solve some problems in the practical application of big data, but these traditional methods can not solve some emerging problems. The complexity of unbalanced data fragmentation requires constant definitions of definitions, principles, algorithms, and tools to deal with new problems that actually arise. In this paper, the large database unbalanced data fragment classification algorithm based on the association analysis method and LRFU strategy is analyzed in detail, and the practicability of this method is further confirmed.

2. The Classification Characteristics and Algorithms of Unbalanced Data Fragments under the Background of Big Data

2.1 Sampling method of data fragments

At the processing perspective level, the study of unbalanced data includes data algorithms and classification algorithms. Among them, the data algorithm is a new algorithm which is improved by the defect of the working principle in the traditional algorithm, so as to improve the recognition accuracy of the unbalanced data fragments; the classification algorithm first preprocesses the unbalanced data, and then Adjust the distribution of data to reduce and ultimately eliminate unbalanced data (Qin, 2017). Nowadays, the correlation and effect comparison between the two algorithms is not clear, but in general, the algorithm based on the algorithm is more accurate and is an efficient and simple research method. Therefore, spatial resampling must be performed based on different evaluation criteria, and then learning in classification and recognition training can be achieved to achieve the desired recognition effect (Chen, 2017). Resampling is performed by the conversion $B \to B'$ of the training set, and a classifier B' is constructed on the new training set f. Sampling is a process of discovering samples to improve the performance of the classifier. The resampling algorithm builds a new mechanism that improves unbalanced fragmentation and provides more global distribution performance than balanced data distribution. The SMOTE algorithm is based on the synthesis of a small number of samples, and finally balances the unbalanced data samples by controlling the number and distribution of new samples. When processing a small amount of unbalanced data, the fragment sample point is x_1, and the same kind K should be calculated first. In the SMOTE algorithm, K is usually 5 or 10. In the set KN_x_1, the randomly selected majority type sample points are x2, and the corresponding attributes j of the expressions x1 and x2 are: $diff_j = x_{2j} - x_{1j}$. A pseudo-random number between 0-1 is denoted as $rand[0,1]$. According to the SMOTE algorithm, the random number generated in the interval $[0,1]$ is multiplied by the difference $diff_1$, and the corresponding attribute value x_{1j} in the original attribute vector is added to obtain a new sample. Property value. Then, the obtained m attribute values are combined, and finally a new synthetic sample f'_{1j} is generated for a small number of newly generated unbalanced data fragments:

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

To solve the imbalance of minority unbalanced data fragments, the formula for calculating the sub-data fragmentation is:

$$N^{j}{}_{\min}=\left\{\frac{\frac{1}{size^{j}{}_{\min}}}{\sum_{j=1}^{S_{\min}}\frac{1}{size^{j}{}_{\min}}}\right\}\times N_{\min} \tag{2}$$

Where: the sample size of a few sub-data fragments is denoted as $size^{j}{}_{\min}$; the number of sub-classes in a minority sample is denoted as $S_{\min}$; $N_{\min}$ represents the minimum number of samples.

In most types of unbalanced data fragments, the number of sub-data fragment samples is proportional to the number of downsampling. To ensure downsampling accuracy, the sub-data fragment downsampling quantity is calculated as:

$$N^{j}{}_{maj}=\left\{\frac{\frac{1}{size^{j}{}_{maj}}}{\sum_{j=1}^{S_{maj}}\frac{1}{size^{j}{}_{maj}}}\right\}\times N_{maj} \tag{3}$$

Where: $size^{j}{}_{maj}$ represents the number of majority subsamples of data; S_{maj} represents the number of subclasses in the majority sample; N_{maj} represents the maximum number of samples.

Finally, according to (2), (3), combined with the previously set number of ups and downs, the calculation is performed multiple times, and a new sample of data fragments is continuously added, and finally a sample of unbalanced data fragments is obtained.

2.2 Extraction of fragment feature sequences

On the basis of the acquisition of the unbalanced data samples, the filter coefficients are intermediately padded to eliminate the difference generated during the downsampling process and the interference during the reconstruction process. Let the low pass filter be $h(k)$, then the feature of number j in the unbalanced data fragment $a(n)$ is expressed as:

$$a_{j}(n)=\sum\nolimits_{k}h(k)a_{j-1}\left(n-2^{j-1}k\right),\ j=1,2,\cdots,N \tag{4}$$

k represents the signal of the low pass filter cutoff frequency; n represents the amount of unbalanced data. The filter coefficient $g(n)$ is calculated by the high pass filter $d_{1}(n)$, namely:

$$d_{j}(n)=\sum\nolimits_{k}g(n)a_{j-1}\left(n-2^{j-1}k\right) \tag{5}$$

When convolution calculation is performed on the unbalanced data fragments by the orthogonal filter, the filters $h(n)$ and $g(n)$ are reconstructed:

$$\delta(n)=\overline{h}(n)*h(n)+g(n)*g(n) \tag{6}$$

Where: * is a discrete convolution; $\delta(n)$ is an unbalanced fragment sequence. When $\delta(n)<0$, the feature sequence cannot be obtained, and a new round of sample sampling should be returned. When $\delta(n)=0$, the feature sequence can be directly obtained without filtering; when $\delta(n)>0$, the unbalanced data fragment sequence A, B should be performed. Rematching, and obtaining feature points by SURF algorithm and pairing them, can obtain the characteristic sequence of unbalanced data fragments:

$$F_{1}=\left\{f_{1i}\,f_{1i},\cdots,f_{1k}\,f_{1t}\right\} \tag{7}$$

$$F_{2}=\left\{f_{2m}\,f_{2n},\cdots,f_{2x}\,f_{2t}\right\} \tag{8}$$

Where: f_{1t} is the feature point of sequence A; f_{2t} is the feature point of sequence B; and the feature points in F_1 and F_2 are classified by two pairs; for example, f_{1k} and f_{2m}, ..., f_{1k} and f_{2x} are sequentially corresponding.

3. Classification and Identification of Unbalanced Data Fragments

3.1 Classification of fragments

On the basis of obtaining the feature sequence of unbalanced data fragments, data fragments need to be classified and processed. In the process of classification, attention should be paid to the impact of classification points on the classification of fragments, and the independence and aggregation of fragments should be confirmed before classifying and processing. When dealing with unbalanced data fragments, unbalanced data fragments can be divided into two intervals, the degree of independence of these two intervals X^2 is: $x^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}$ Among them, n_{ij} denotes the number of objects belonging to j decision class in I interval, while E_{ij} denotes the interval between unbalanced data fragments. The imbalance of data fragmentation attributes makes the direction of data gradient diversified. On the basis of extracting the characteristics of gradient aggregation degree, this paper describes the aggregation degree of unbalanced data fragments. Because of the imbalance of data fragmentation attributes, the initial direction of data gradient is different. The data gradient points to scatter, representing the state before the data gradient converges. The data gradient points to the same direction, which represents the state after its aggregation. The formula for calculating the degree of convergence is as follows: $GC(p)=\frac{1}{D}\sum_{i=1}^{D}\max_{l_{\min}\le n\le l_{\max}}\left\{\frac{1}{n+1}\sum_{l=0}^{D}\cos\theta_{il}(p)\right\}$ Among them, D represents the number of vectors d' formed by unbalanced data fragments P. $\cos\theta_{il}(p)$ can be calculated by vector d' and vector $g_i^{'}$.

3.2 Identifying fragments

After classifying and processing the unbalanced data fragments, it is necessary to normalize the data fragments so as to eliminate the influence of different data fragmentation characteristics and merge the unbalanced data fragments into [－1,1]. The calculation formula is as follows: $x_i^{'}=\frac{2x_i-(x_{\min}+x_{\max})}{x_{\min}-x_{\max}}$ Among them, $x_i^{'}$ represents data fragments after normalization. $X_i\in[-1,1]$, x_i represents the original data fragments corresponding to the input, $x_{\max}$ and $x_{\min}$ represent the maximum and minimum values of data fragments corresponding to the input.

By transforming the objective function and combining the expansion matrix, the fitness function is obtained. The objective of the optimal feature subset is to select a set of data fragments consisting of elements “r” from an unbalanced data fragment set. If the optimal feature subset is selected $B\subseteq A$, The fitness function is calculated as follows: $f(A)=\frac{\sum_{i=1,j=k}r_{ij}}{card(B)}$ Among them, $card(B)$ represents the cardinality of unbalanced data fragment set B, and r_{ij} represents the elements of row I and column J. After classifying the unbalanced data fragments, the LRFU strategy and the correlation analysis method are combined to determine the unbalanced data fragment attribute values, thereby identifying the unbalanced data fragments. If there is a missing feature in the unbalanced data fragment, it needs to be replaced and the attribute value of the smallest data fragment is calculated. The calculation formula is as follows: $CRF_{t_{base}}(b)=\sum_{i=1}^{k}F(_{t_{base}}-_{t_{b1}})$ Among them: t_{base} represents the contribution value of unbalanced data fragments, and t_{b1} represents the contribution value of unbalanced data fragments for one recognition.

Thus, using a segmentation-like linear approach to deal with unbalanced data fragments with a

high degree of convergence, it is possible to classify and identify unbalanced data fragments. By transforming the objective function and combining the expansion matrices, a fitness function can be obtained. Combined with the LRFU strategy and the correlation analysis method, the attribute values of the unbalanced data fragments can be determined to identify the unbalanced data fragments.

4. Experiment and Result Analysis

In order to verify the feasibility and effectiveness of the improved algorithm in the classification and identification of unbalanced data fragments in large databases, the recognition accuracy of unbalanced data fragments and the recognition fitness are compared and analyzed. The formula for calculating the accuracy rate is as follows: $precision = \frac{TP}{TN + FP}$ Among them, TP represents the number of positive unbalanced data fragments under correct recognition, TN represents the number of negative unbalanced data fragments under correct recognition, and FP represents the number of positive unbalanced data fragments under false recognition.

4.1 Setting up experimental environment

Experiments are carried out to analyze the effectiveness of the recognition algorithm for unbalanced data fragments, and the experimental environment needs to be set up. In the experiment, 500 data packets were selected from a company's comprehensive document information database to carry out the experiment. The document information data in the company's database is more complex, the data types are diversified, and the number is huge. There are a large number of uneven data fragments in the data packet, and the distribution is scattered. The direction of the data gradient is also different. In this experimental environment, the accuracy of identifying unbalanced data and the fitness of unbalanced data are tested.

4.2 Analyze experimental results

When using traditional methods for recognition, along with the increase of recognition time, the optimal fitness and fitness values have larger errors, and have larger fluctuations. The maximum value is about 99.1, and the minimum value is about 95.4. Both of them do not reach the optimal value of 99.5. When using the improved method to identify, along with the increasing recognition time, the fitness value is also increasing, and it is closer to the optimal fitness value. The maximum value is about 99.4, and the minimum value is about 98.5. Both of them are closer to the optimal fitness value. Compared with the traditional recognition method, the improved recognition method has some advantages.

The traditional method and the improved method are used to compare the recognition accuracy. In the case of uncertain adaptation value, with the use of traditional methods, along with the increasing of the adjustment value, the accuracy of recognition is gradually increasing, but the growth rate is slower. In addition, the recognition accuracy has a variety of changes, such as rising and falling, which is not stable enough, and its maximum value is about 60.3%. After using the improved method, along with the continuous increase of fitness value, the recognition accuracy is also gradually improved, and its speed of improvement is faster, there is no repeated rise and fall, the maximum value is about 97.2%, compared with the traditional method, it has a lot of improvement, with certain advantages.

Acknowledgements

The Industry-University Cooperation Educational Project in 2018 of Ministry of Edication ”The Exploration and Practice on the School - enterprise Cooperation Practical Computer professional’s Development”(201801193133)

References

[1] Wen A. H. and Lan Y. (2018). Analysis of Large Database Unbalanced Data Fragment Classification and Recognition Algorithm, Mechanical Design and Manufacturing Engineering, 15 (6): 86-90.

[2] Liu L. M. and Yang D. X. (2018). Classification and Identification of Unbalanced Data Fragments under Visual Processing of Large Ship Database, Ship Science and Technology, 40 (20): 149-151.

[3] Liu Y. and Sun Y. Q. (2017). Property Inference for Social Media User Comment Behavior, Journal of Computer, 40 (12): 130-144.

[4] Fu D. S. and Jing Z. J. (2015). Data Fragment Classification and Recognition Algorithm Based on PCA-LDA and KNN-SMO, Software, 32 (7): 21-25.

[5] Han Y. and Shan W. F. (2018). Data Defragmentation Algorithm in Deduplication Backup System, Science Bulletin, 34 (6): 23-24.

[6] Qin J. D. and Peng H. F. (2017). Space Debris Size Estimation Algorithm Based on Solar Phase Angle Function, Telecommunications Technology Research, 54 (4): 1-8.

[7] Chen H. J. and Luo F. Q. (2017). Optimization Identification of Network Resource Information Loss under Big Data, Computer Simulation, 34 (9): 358-361.