Research on Big Data Classification Algorithm Based on Fuzzy Mathematics

Chengyuan Zhao
Zaozhuang Vocational College of Science & Technology Tengzhou, 277599, China

Abstract: In order to improve big data's ability of information fusion and retrieval, it is necessary to optimize the classification design of data. An improved big data classification algorithm is proposed based on fuzzy mathematics. The distributed storage design of database is carried out by using grid topology, and big data's nearest neighbor fuzzy clustering center is calculated by semantic autocorrelation function analysis method, and then the vector quantization feature coding model of big data is constructed. The mass big data characteristic distribution gradient map is extracted, the distributed scheduling input vector value of database is obtained, the self-organization neural network training is used to the cluster coding sample of big data, fuzzy C-means algorithm is used to carry on the big data clustering processing. The method of fuzzy mathematics analysis is used to realize the optimal classification of data. The simulation results show that this method has high accuracy and low error rate in big data classification, and it can improve data fusion and scheduling ability.

Keywords: fuzzy mathematics; big data; classification; database

1. Introduction

With the rapid development of computer technology, especially database technology, as well as the expansion of human activities and the acceleration of life rhythm, people can obtain and store data in a quicker, easier and cheaper way, which increases the amount of data and information in an exponential way. The huge pressure of "explosion" and "Data Glut". If these massive data cannot be effectively used, it will only become "data garbage". Knowledge is a great effect on the progress of human society. Data mining is the method and technology to extract useful knowledge from a large number of data and to extract useful knowledge. There is a lot of content in data processing, one of the most important aspects is classification rule mining[1].

Classification technology has been applied in many fields, for example, we can construct a classification model to evaluate the risk of bank loans through customer classification. A very important feature in current marketing is to emphasize customer segmentation. This is also the function of customer category analysis[2]. By using the classification technology in data mining, customers can be divided into different categories, and the security domain includes intrusion detection based on classification technology and so on. Researchers in the fields of machine learning, expert systems, statistics and neural networks have proposed a number of specific classification and prediction methods. The KNN(K Nearest neighbors) algorithm[3], also known as the K nearest method, it is generally one of the relatively easy to understand algorithms, assuming that each class contains multiple sample data. And each data has a unique class tag to indicate which classification these samples belong to, which is to calculate the distance from each sample data to the data to be classified, to take the nearest K-sample data to the data to be classified. So which category of the K sample data is the majority, then the data to be classified belong to that category. The disadvantage of this method is that it requires a large amount of computation, because for each text to be classified, the distance between it and all known samples must be calculated in order to obtain its K nearest neighbor points[4]. In the field of data classification, neural networks have aroused the interest of many scientists, but with the in-depth analysis of the functions and limitations of the neural networks represented by the perceptron, the research of neural networks has fallen into a low tide. However, some scholars persist in the research and have achieved some results, and the ART model and Kohon of the Grossberg have emerged. EN's SOM model. Artificial neural networks can be used in data mining classification, clustering, feature mining, prediction and pattern recognition. Therefore, artificial neural networks play an important role in data mining[5.6].

In short, data mining technology is a young and promising research field, the powerful driving force of business interests will continue to promote its development. New data mining methods and models come out every year. Nevertheless, data mining technology still faces many problems and challenges: for example, the efficiency of data mining methods needs to be improved, especially the efficiency of data mining in large scale data sets, and the development of mining methods adapted to multiple data types and noise tolerance, in order to solve the problem of data mining in heterogeneous data sets, dynamic data and knowledge mining, data mining in network and distributed environment, and so on. In addition, multimedia databases have developed rapidly in recent years. The mining technology and software for multimedia database will become the research and development hotspot in the future.

Aiming at the problems existing in traditional methods, this paper presents a classification technique of big data based on fuzzy C-means mathematical classification method. The distributed storage design of database is carried out by using the grid topology structure, the self-organizing neural network is trained to the cluster coding samples of big data, and the algorithm of fuzzy C-means is used to deal with big data clustering. The method of fuzzy mathematics analysis is used to realize the optimal classification of data. Finally, the simulation results show that the proposed method can improve the ability of data classification.

2.Analysis of Data Storage Structure and Feature Coding of Database

2.1. Big data quantization feature coding

In order to realize the optimized classified storage design of mass big data, it is necessary to analyze the data storage structure of large database, and to design distributed storage grid of database by using 3×3 grid topology structure. Four retrieval channels are set up for mass big data access design, then the semantic autocorrelation function analysis method is used to calculate the neighbor fuzzy cluster center, and the vector quantization feature coding model of multimedia data is constructed[7]. In the semantic pheromone extraction of four retrieval channels, a mass of big data feature distribution gradient map is extracted, and the distributed scheduling input vector values x_1, x_2, x_3 and x_4 of the database are respectively:

$$\begin{cases} x_1 = p_1 - m \\ x_2 = p_2 - m \\ x_3 = p_3 - m \\ x_4 = m \end{cases} \tag{1}$$

Where, m is the embedded dimension of the initial cluster center, and the database stores the distributed spatial feature distribution region { $W_j^{(0)}$, $j = 0,1,\cdots,N-1$ }.The vector quantization coding is carried out by using the feature sequence training and reconstructing method of massive big data, and the initialization codebook is set up as $\{c_j^{(0)} = 0, j = 0,1,\cdots,N-1\}$, it sets the weight response of multimedia data distribution to $\{S_j^{(0)} = 1, j = 0,1,\cdots,N-1\}$ .The initial value of vector quantization coding for multimedia data is $t = 0,1,\cdots,n-1$. Let $x(t)$ be the training sequence, and the initial value of frequency count of large database storage channel is set to $t = 0$.

2.2. Data feature extraction and quantization coding

The distributed storage design of database is carried out by using grid topology structure, and big data's nearest neighbor fuzzy clustering center is calculated by semantic autocorrelation function analysis method, and the vector quantization feature coding module of multimedia data is constructed. Type, the input big data is controlled by steady-state periodic decomposition, and the training vector pattern is obtained as $x(t) = (x_0(t), x_1(t), \cdots, x_{k-1}(t))^T$; the distance between big data's time-domain vector $x(t)$ and the weight vector w_j connected by all classified storage nodes is calculated as:

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

Where, $w_j = (w_{0j}, w_{1j,} \cdots, w_{k-1,j})^T$, the output of big data's quantization characteristic coding is as follows:

$$w_{j^*}(t+1) = w_{j^*}(t) + a(c_{j*})[x(t) - w_{j^*}(t)] \tag{3}$$

Where, $j \in (j^*, NE_{j^*}(t))$. When the initial value of clustering center is determined, the efficiency of classification storage is improved by using big data's quantization feature coding. The initial N level symbol $\hat{A}_0 = \{y_i\}$, $i = 1,2,\cdots,N$ of the large database is constructed, and the estimated value of big data's interference information parameter $\hat{A}_n = \{y_i\}$, $i = 1,2,\cdots,N$ is calculated. The best codebook of big data is obtained as follows:

$$c(t,t) = \sum_n a_n(t) e^{-j2p f_c t_n(t)} d(t - t_n(t)) \tag{4}$$

Thus, the regional distribution function of the data storage cluster is obtained as follows:

$$\begin{aligned} E^{cv}(c_1, c_2) = {} & m \cdot Length(C) + n \cdot Area(inside(C)) \\ & + l_1 \int_{inside(C)} |I - c_1|^2 dxdy \\ & + l_2 \int_{outside(C)} |I - c_2|^2 dxdy \end{aligned} \tag{5}$$

Thus, the fuzzy mathematical coding analysis of big data classification is realized.

3. Data Classification Algorithm Optimization Implementation

3.1. Fuzzy C-means clustering

The vector quantization feature coding model of big data is constructed, and the gradient-map of feature distribution is extracted. The input vector value of distributed scheduling of database is obtained, and the feature distribution of Chinese text information in the mass big data retrieval region is assumed. The sequence is $x(t)$, $t=0,1,\cdots,n-1$, the quantization coding information of big data is divided into regions, and the window weight control of storing information is carried out by using template matching method, described as:

$$u=[u_1,u_2,\mathbf{L},u_N]\in R^{mN} \tag{6}$$

In the weighted control of database storage and distribution mentioned above, big data's clustering coding samples are trained by self-organizing neural network in both horizontal and vertical gradients[8]. The results show that the threshold of classification control of data attribute sets is calculated as follows:

$$AVG_X=\frac{1}{m\times n}\sum_{x=1}^{n}\sum_{y=1}^{m}\left|G_X(x,y)\right| \tag{7}$$

The output weight vector is:

$$x(t)=(x_0(t),x_1(t),\cdots,x_{k-1}(t))^T \tag{8}$$

The method of fuzzy mathematics analysis is used to decide the threshold of multi-attribute classification control. The fuzzy C-means clustering processing of data classification is realized.

3.2. Fuzzy mathematics analysis

Let X and Y be the sets of distinguishing attributes of big data's classification features, and describe the fuzzy C-means clustering space matrix of the magnanimous big data as follows:

$$\hat{x}(s_j)=\frac{1}{\left\|s_j\right\|}\sum_{x_i\in s_j}x_i \tag{9}$$

By using fuzzy mathematics analysis method, the optimal decision formula of data classification is obtained as follows:

$$TL_X(x,y)=\begin{cases}Text & ,if\,(GD_X(x,y)>T_X)\\ NonText & ,Otherwise\end{cases} \tag{10}$$

Combined with fuzzy C means clustering and feature extraction, the classification result of large data output is obtained as:

$$\hat{x}(k/k)=\sum_{j}^{m}\hat{x}^i(k/k)u_j(k) \tag{11}$$

$$\begin{aligned}P(k/k)=&\sum_{j}^{m}u_j(k/k)\{P^j(k/k)\\&+[\hat{x}^j(k/k)-\hat{x}(k/k)][\hat{x}^j(k/k)-\hat{x}(k/k)]^T\}\end{aligned} \tag{12}$$

In above, the big data optimization classification is realized, the fuzzy recognition ability of data is improved.

4. Simulation Experiment and Result Analysis

In order to verify the application performance of this method in the realization of big data optimization classification, the simulation experiment is carried out. In the experiment, C++ and Matlab 7 mixed programming are used to realize the mass data of large database. The optimized classification shows that the bandwidth of data sampling is 50~120kHz, the bandwidth of storage interval distribution of database is -10~0dB, the initial frequency of big data access is $f_1=2.1$ Hz, the stop frequency $f_2=0.23$ HZ, the information interference intensity in data storage space is $SNR=-10\text{dB}$, and the maximum fraction is obtained, data class search radius $G_{\max}=30$, according to the above simulation environment and parameter settings, big data classification, get the original data set as shown in figure 1.

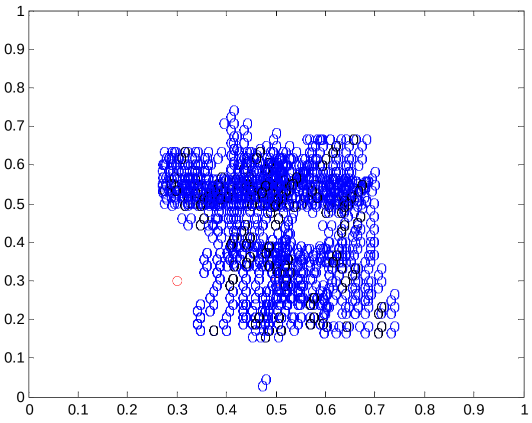

Figure 1. Original data

The proposed method is used to classify the data, the result of the classification output is shown in figure 2.

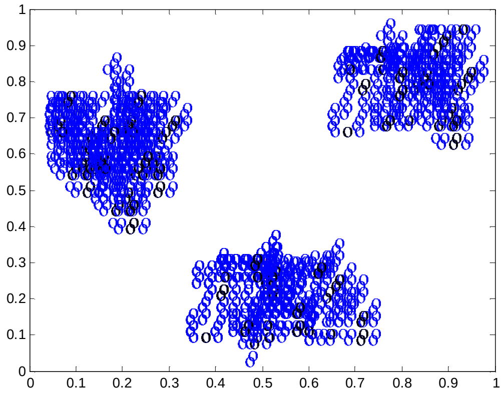

Figure 2. data classification results

Figure 2 shows that the accuracy of using this method to classify data is high. The error rate of different methods for data classification is tested. The result of comparison is shown in figure 3. Figure 3 shows that the method is used in this paper. The bit error rate(BER) of data classification is low and the performance is better.

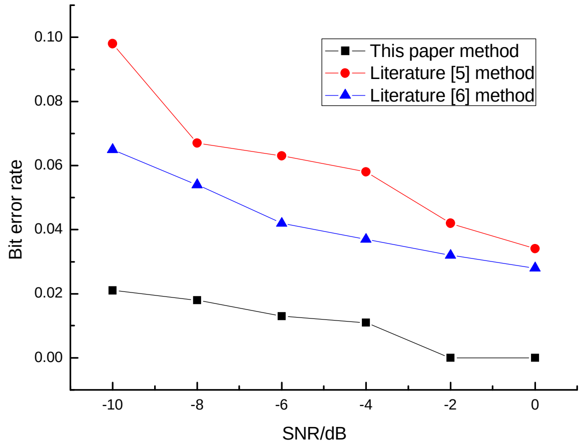

Figure 3.BER performance comparison

5. Conclusions

In this paper, an improved big data classification algorithm is proposed based on fuzzy mathematics. The distributed storage design of database is carried out by using grid topology, and big data's nearest neighbor fuzzy clustering center is calculated by semantic autocorrelation function analysis method, and then the vector quantization feature coding model of big data is constructed. The mass big data characteristic distribution gradient map is extracted, the distributed scheduling input vector value of database is obtained, the self-organization neural network training is used to the cluster coding sample of big data, fuzzy C-means algorithm is used to carry on the big data clustering processing. The method of fuzzy mathematics analysis is used to realize the optimal classification of data. The simulation results show that this method has high accuracy and low error rate in big data classification, and it has good application value in data fusion and data mining.

References

[1] Jie Mi, Peng Zhang, Haipeng Yu. Large Data Clustering Algorithm Based on Particle Swarm Differential Perturbation Optimization[J]. Journal of Henan University of Engineering (Natural Science Edition), 2016, 28(1):63-68.

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

[3] BAE S H , YOON K J. Robust online multiobject tracking with data association and track management[J]. IEEE Transactions on Image Processing, 2014, 23(7): 2820-2833.

[4] X Jiang, K Harishan, R Thamarasa, et al. Integrated track initialization and maintenance in heavy clutter using probabilistic data association[J]. Signal Processing, 2014, 94: 241-250.

[5] L Li and W Xie. Intuitionistic fuzzy joint probabilistic data association filter and its application to multitarget tracking [J]. Signal Processing, 2014, 96: 433-444.

[6] D Svensson, M Ulmke, and L Hammarstrand Multitarget sensor resolution model and joint probabilistic data association[J]. IEEE Transactions on Aerospace and Electronic Systems, 2012, 48(4): 3418-3434.

[7] Tao Wu Lifei Chen Gongde Guo. High-dimensional data clustering algorithm with subspace optimization. Journal of Computer Applications, 2014, 34(8): 2279-2284.

[8] Sen Hou, Xing-guo Luo, Ke Song. A Maximum Entropy Weighted Trust-Analysis Algorithm Based on Sources Clustering[J]. Chinese Journal of Electronics, 2015, 43(5): 993-999.