Research on Mining Cloud Data Based on Correlation Dimension Feature

Jingwen Tu¹, Li Tao²

Abstract

Large hierarchical cloud storage database has distributed of non-continuous massive data, the data has nonlinear characteristics of strong coupling, and using traditional methods for data mining, mining exist difficult problems. This paper proposes mining algorithm based a cloud non continuous layer data, and analyze the overall data mining model. The paper use fuzzy C means clustering algorithm to complete the semantic ontology feature point clustering beam based on semantic feature extraction and quantization encoding, to realize improved data mining algorithm. The experimental results show that the improved algorithm, the non-continuous mining level data have high precision, good performance, anti-interference ability strong, performance is superior to the traditional method.

Keywords: K means algorithm; FCM algorithm; Cloud Computing; Data clustering.

1. Introduction

With the rapid development of network information and large data processing technology, a large number of data are distributed in the network space to constitute the network Web Deep database through the cloud storage model. The big data information processing technology is highly developed today, using cloud computing method for data transmission and scheduling, which can effectively raise the ability to access the ability of information retrieval and the Deep Web database. In a large cloud storage database, which distribute a massive non continuous level data with self-coupling nonlinear characteristics, there is mining difficult in other world environmental disturbance. In order to improve the semantic of network database retrieval and information analysis capabilities, we need the mining method for cloud computing discontinuous hierarchical data based on cloud computing platform to build cloud mining environment data.

Therefore, this paper proposes a mining algorithm for nonhierarchica data semantic ontology feature point beam based on clustering, the algorithm can analyze the overall level and non-continuous data structure model of data mining, but also extract and encode on the non-continuous level data semantic feature. On the basis of quantization coding, this paper uses the fuzzy C means clustering algorithm to realize the semantic ontology characteristic directional beam clustering, and realize the data mining algorithm improvement. Finally, the performance test is carried out by simulation experiment.

¹Nanchang Business College, Jiangxi Agriculture University, Nanchang, Jiangxi, 330044 China
²Jiangxi Radio & TV University, Nanchang, Jiangxi, China
2. Working principle of data mining model

The realization of the overall framework of the non-continuous data mining model based on cloud computing is shown in Figure 1.

![Figure 1. Realization of overall framework of data mining model based on cloud computing.](image)

2.1 Quantization and coding

In a cloud of non-continuous levels of data mining method based on the overall framework, we design a hierarchical data mining model of non-continuous mass distribution of large cloud storage in the database, due to the non-continuous level data have strong self-coupling nonlinear characteristics, it has mining difficult under big disturbance. In the cloud computing environment, this paper proposes a nonhierarchical data mining algorithm based on point semantic ontology feature beam clustering. We extract and encode non-continuous level data semantic feature, by adaptive weighted for nonhierarchical data semantic ontology model of continuous gradient window and the maximum value, we get non-continuous level data association directional weighted vector:

(1)

We used a 1 * N time window to compress feature, and determine data mining time window size N of the non-continuous levels, the time window is divided into many small intervals, the vector quantization encoding, testing function is x (T), the distance of a continuous sliding encoding window is as shows:

(2)

Among them WJ is the maximum gradient difference weighted factor of the non-continuous data, expressed as

(3)

In order to extract directivity of semantic feature through useful text of non-continuous level data, assuming that the data is piecewise smooth, in the linear region of piecewise stationary, they judge non-continuous hierarchical data in narrow time window TLx, TLy, to get the text feature extraction judgment type of non-continuous data:

(4)

The energy density spectrum of the non-continuous layer data is m; at the minimum distance of the window, we obtain the clock sampling Nj*, Among them, the vector space trajectory function of vector quantization coding is:

(5)

The semantic ontology is divided into 3*3 topological structure, and the specific window function is selected to obtain the output vector quantization coding object set Fm (x, y):

(6)

(7)

among: is the estimated value of the frequency resolution;uj(k) is a set of information fusion attributes of non-continuous level data.;uj is adjustable window of Fu Liye transform;Pj is
probability density function of semantic ontology feature. According to the above analysis, we quantify the new code as the gradient direction of the discontinuous data, and we give the non-continuous hierarchical data edge fusion series $N$; initialize the semantic features of non-continuous data threshold level $\varepsilon$; the training sequence for the edge pixels of the given non-continuous hierarchy data is $\{x_j\}$, $j=0, 1, \cdots, m-1$, feature decomposition in an initial $N$ level symbol, $i=1, 2, \cdots, N$. Set $n=0$, $D-1=\infty$. According to the above analysis, based on the vector quantization coding, the semantic ontology characteristic directional beam clustering is carried out.

### 2.2 Data mining implementation

The improved algorithm mainly aimed at FCM algorithm cannot automatically obtain the optimal cluster number and sensitive to initial cluster. According to a great deal of information display: The optimal cluster number usually meet $c^* \leq \sqrt{n}$. Therefore, the range of optimal cluster number is set to: $(1, \sqrt{n})$. The basic idea of the algorithm is: first get the $\sqrt{n}$ clustering results by using K means algorithm is improved, the results of the $\sqrt{n}$ cluster is as the cluster center of FCM algorithm, corresponding calculated the effectiveness function based on principle of granularity value when the $C=\sqrt{n}$, using the method of clustering center to get the $C-1$ clustering center, the clustering number minus 1 again, and calculate results of FCM algorithm and the effectiveness of the function value when $C=\sqrt{n}-1$ clustering, order cycle, until $C$ reaches the lower clustering number range. Finally the validity of all function value, effectiveness evaluation to obtain the best clustering. The basic flow chart is as shown in figure 2 based on K algorithm and the principle of mean grain size.

![Figure 2. The improved algorithm flow chart based on the K mean and the principle of granularity.](image)

According to the basic algorithm flow chart based on K mean and the principle of granularity, can get the specific steps of the algorithm:

1. InitialParameter() // initialize each parameter
2. $V(0)=$Improved_KMeans_Centers($c_{\text{max}}, n$) // using improved K means algorithm, get the $C_{\text{max}}$ clustering center, as the initial clustering center of FCM algorithm
3. For $c=\sqrt{n}$ to 2 {
    
    - Calculate effective function $GD(c)$
    - Output the optimal cluster number and cluster centers
    - Select minimum value of $GD(c)$
    - Refresh the clustering center value
    - Calculate the new classification matrix composed of membership value
    - Combine with the clustering center, set cluster $c-1$
    - $b=b+1$
    - $\|V_{b}^{(b)}-V_{b}^{b-1}\|<\varepsilon$

Finish

136
4. Repeat // Repeat FCM algorithm clustering process
5. \( U^{(b)} = \text{UpdateMatrix}(c, n) \) // Update the classification matrix
6. \( V^{(b)} = \text{UpdateCenters}(c, n) \) // updating the cluster center
7. \( Until \|V^{(b+1)} - V^{(b)}\| < \varepsilon \) // until to satisfy the convergence condition
8. \( \text{Dis} \tan ce_{\text{min}}(i, j) \) // Get the between class distance, calculate the distance of two kinds of min
9. \( \text{MergerCenters}(c) \) // Combined with the cluster center, the distance between classes of two kinds of minimum, merged into one category, the clustering number is C-1
10. \( M = \text{CountEffectiveFunction}() \) // Calculation effective function value, save it to the M collection
11. } End For
12. \( c^* = \min(M) \) // to find out the best clustering validity function minimum value corresponding to the number in the M collection
13. Output \( (C^*) \) // output clustering

3. Simulation experiment and performance test

We can see from Figure 3, the original data flow is distribute in the cloud storage database by self-coupling nonlinear strong interference, which result the system has the low accuracy of data mining, poor performance. In order to quantitatively analyze the performance of mining, using this algorithm and the traditional method in data mining accuracy as test index, the experiment achieved across 10000 Monte-Carlo the output of data mining, the root mean square error of RMSE, the results shown in Figure 4.

![Figure 3. Time domain waveform of primary data info flow.](image1)

From Figure 4, we can see that this algorithm is the non-continuous data mining, the root mean square error of the data mining is low, which shows that the data mining accuracy is higher than the traditional method, and the anti-jamming performance is strong.

![Figure 4. Performance comparison of data mining.](image2)
4. Conclusions

According to the traditional data mining method mining with low precision, big error problem. This paper presents the cloud data mining algorithm based on non-continuous level, to analyze the overall data mining model, extract the non-continuous level data semantic feature and quantization encoding, using the fuzzy C mean clustering complete algorithm based on encoding, semantic ontology feature directional beam clustering, improved data mining algorithm is realized. The experimental results show that the algorithm of data mining has high precision, good performance, semantic ontology feature a directional beam has better clustering effect, strong anti-interference ability.

Reference


