Eng Int Syst (2020) 1: 41–49

Engineering
Intelligent Systems

The Analysis of Non-Significant Feature Data Mining in Big Data Environments

Xiaoli Meng*

Department of Engineering, Xi'an International University, Xi'an 710077, China

In order to cope with the problem of low precision in data mining, it is necessary to study the non-significant features of data mining methods. The current method shows efficiency bias in the data mining. In this paper, a non-significant feature data mining method based on Ant Colony Clustering is proposed. This method extracts the characteristics of data clustering which manifest the significant characteristics of data mining in a big data environment. Experiments show that this method is more accurate when data mining.

Keywords: Big data environment; Non-significant feature; Data mining; Ant Colony Clustering algorithm.

1. INTRODUCTION

With the rapid development of computer technology, there has been a significant growth of the data information index. The ability to produce and gather data using information technology has improved tremendously [1]. Within a big data environment, the non-significant feature data mining is widely used in business management, government operations, and scientific research. Non-significant feature data mining is a serious enterprise within big data environments [2]. The centralized characteristics of the data can be distributed to the data storage node, which improves the visiting number of the data [3]. Data mining mainly refers to the extraction of information from a large number of data sets, commonly used for predicting trends [4]. As there are many databases full of information, the accumulation of data information requires a deeper analysis of data mining. Data mining is convenient and can improve the application of the information provided by the data [5]. Current database structures are efficient for data entry and query and statistical functions, but they are still unable to detect existing relationship. They cannot predict events on the basis of the existing data, which leads to the complex problems of data mining efficiency and deviation processes [6]. This has led to the necessity of improving the accuracy and reducing the length of time taken when data mining [7]. A data mining method based on Ant Colony Clustering, is an effective solution due to the high precision of the results of data mining in this way [8–11]. Data mining has extensive business prospects; this has become one of the main reasons for the focus in this research direction. It has caught the attention of many academic experts and scholars, and they have obtained varied results [12].

Li Minzhi was the first expert to put forward the k-means algorithm clustering method. This method can only be applied to the case of a pre-defined clusters average and therefore shows poor adaptability for data mining [13, 14]. Huangyong proposed a data mining clustering algorithm for classification properties, which proved that a finite iteration can converge to a local minimum. Cui Huili proposed a data mining study based on non-parametric clustering methods. Liu Xue proposed a clustering algorithm that iterates over the initial data. Mr Li proposes a data mining method based on binary clustering algorithms [15]. The experimental results show that the algorithm is more accurate and robust. Wang Haobo puts forward a hierarchical clustering algorithm based on indistinguishable data; applying this algorithm to data mining

*E-mail of Corresponding Author: mengxiaoli7998@163.com

can improve the accuracy of data mining effectively Han Wei proposes a method of data mining of the lattice clustering algorithm in high dimensional space, which is not sensitive to noise data and with a fast clustering speed. Jiang Yuhang proposed the data mining method based on the adaptive clustering algorithm. Xu Jiayi put forward the data mining method based on the density clustering method.

Article [16] presents a method for mining non-salient feature data in a large data environment based on a genetic algorithm. This method firstly analyzes and elaborates the clustering method, and analyzes the advantages and disadvantages of the clustering algorithm, it then introduces the technology of data mining, the related technology of cluster analysis and the problem of the application of the K-mean clustering algorithm in the process of network data mining. Based on this, the variants of genetic algorithms are referenced to the K-means clustering algorithm; using the mutation factor in the genetic algorithm can shorten the convergence time of the genetic algorithm, in order to avoid the algorithm finding the local optimal only, thus the data mining of non-significant feature data in a large data environment is completed. Finally, it is proved that the method in this paper has a better clustering effect, however this method uses a complex data mining process. Article [17] presents a data mining method based on non-significant feature data in a large data environment based on group intelligence. First, the data information for mining is represented by vector space, all non-significant features are removed from the data set via the conventional way, the data vector is then randomly distributed to the data set, data clustering is performed by a clustering of group intelligence, and the clustering results are collected from the data set using recursive algorithms. In order to improve the practicability of the algorithm, the recursive algorithm is combined with the K-means algorithm creating a hybrid clustering algorithm. The data mining of non-significant feature data in a large data environment is performed by using the hybrid clustering algorithm. Using experiments to compare the methods, the mining process in this method is flexible; however the method suffers from a poor excavation effect. In article [18], a data mining method based on improving the particle swarm optimization algorithm and K-means algorithm is proposed. This method first introduces random variation in the process to enhance the diversity of data, improves the ability to search the data globally, the operation of the K-means clustering algorithm is based on the variance of the data, which enhances the search ability of the algorithm and shortens the time of data convergence, thus the data mining of nonsignificant features in a large data environment is completed. Comparing this method with the particle clustering algorithm and the PSO clustering algorithm, the experiment shows that the method is simple, however there is still the problem of low precision.

In this paper, a new study method of non-significant feature data mining based on Ant Colony Clustering is proposed. This method firstly makes the spatial reconstruction of the time series for non-significant eigendata, performs statistical analysis of non-salient feature data combined with time sequence, then collects data by combining the results of the analysis. The data acquisition results are used to extract the characteristics of non-salient feature data, the Ant Colony Clustering algorithm is used to cluster by combining data from the feature extraction, thus the mining of non-significant feature data in large data environments is completed. The result of the simulation experiment shows that the method can effectively reduce the time taken by data operations and improve the precision of data mining.

2. NON-SIGNIFICANT FEATURE DATA MINING METHOD BASED ON ANT COLONY CLUSTERING

First, the spatial reconstruction of time series for non-significant eigendata is created, then non-salient feature data was statistically analyzed by combining the time series, the data was then acquired by combining the results of the analysis. The results of the data collection are used to extract the non-salient feature data, then clustering is performed by using the Ant Colony Cluster algorithms to combine the data from the feature extraction, thus the mining of non-significant feature data in a large data environment is finished. The specific steps are as follows:

2.1 Non-Salient Feature Data Collection

Assuming that the distribution model for non-significant feature data is represented as a binary group $G(0) = (V,\ E,\ L_V,\ L_E,\ \mu,\ \eta)$, $\eta{:}E \to L_E$ represent heterogeneous ontologies of different data. The initial data center in the large data environment is satisfy ing the condition $G_1 \subseteq G_2 \Leftrightarrow Y_1 \subseteq Y_2$, assuming that $A = (a_1,\ a_2,\ \ldots, a_n)$ is the mapping of data vectors. For the sample data set $M = (x_1,\ Lx_n,\ L,\ x_{nm}, y)$, the number of spaces in the data center vector $m_i \in R_1$ satisfies

(1) Additive homomorphic

Being $f(m_1 + m_2) = f(m_1) \oplus f(m_2)$

(2) Multiplicative homomorphic

Being $f(m_1 \times m_2) = f(m_1) \otimes f(m_2)$, the data feature sequence is refactored to obtain the category set of the data set $M = (V_A : V_B : V_c)$:

$$x(t) = \mathrm{Re}\{a_n(t)e^{-j2\pi f_c \mathcal{T}_n(t) S_l}(t - \tau_n(t))e^{-j2\pi f_c \tau_n t}\} \quad (1)$$

Based on the discrete decomposition method, the attribute value of the non-significant eigendata vector a_i is represented by $\{c_1, c_2, \cdots, c_k\}$. Assuming that D represents the non-significant feature mining data set, 0, p represent the sequential elements in the data set D, $dist\ (p,\ 0)$ is the Euclidean distance between the non-salient feature data vector P and 0, assuming that q is the point not in the data set X, then the Euclidean distance between the two data center vectors is obtained by $dist\ (q, D) = \min\{dist(q,\ 0), 0 \in D\}$.

Definition 1 Suppose that the non-salient feature data set X, Y is a finite data vector set, $X \cap Y = \emptyset$, for a positive integer k, the distance of the similar K distance sequence is represented by $k\_distance(p)$, it is defined as vector data set $dist\ (p,\ 0)$ of the time sequence object 0 of the data sequence

trace and $0 \in D$. The conditions for the collection of data are met:

(1) There are at least n sets of data neighbors, $0' \in D/\{p\}$ meets the condition that the Euclidean distance between data vectors is $dist\ (p,\ 0') \leq dist(p,\ 0)$.

(2) There are $k-1$ data information vector models $0' \in D/\{p\}$ in data object P and they meet the condition that $dist\ (p,\ 0') \leq dist(p,\ 0)$.

The spatial reconstruction method is used to model the data time series model, the adjacent domain of the inverse k of object P of data in space has directivity characteristics. For a given corresponding coefcients k of the information data state and the inverse distribution of non-salient feature data object P, the k th distance space reconstructed trajectory in data gradient information domain D is $RN_k(p)$, this can be defined as:

$$RN_k(p) = \{q|q \in D,\ p \in N_k(q)\} \tag{2}$$

By refactoring the space of the non-salient feature data time series to obtain the power accumulation scale of the inverse of the data object P satisfies the kth adjacent domain.

The relevant functions of the information data time series provide the basis for data input for non-significant feature data collection. The scale of non-significant data transfer time in large data environments is represented as:

$$c(\tau,\ t) = \sum_n a_n(t) e^{-j2\pi f_c \tau_n(t)} \delta(t - \tau_n(t)) \tag{3}$$

In Formula (3), $a_n(t)$ is the amplitude of the non-significant eigentime sequence oscillations on the nth data transmission channel, $\tau_n(t)$ is the time extension of transmission of the nth data transfer path, f_c is the modulation frequency of non-significant feature data channels in large data environments. The channel model of non-salient feature data was obtained by using an adaptive wave beam FM method and expressed by Formula (4):

$$h(t) = \sum_{i=1}^{p} a_i\, p(t - \tau_i) \tag{4}$$

In Formula (4), a_i is the loss of non-significant data propagation in large data environments, τ_i is the extended periods of non-significant feature data transfer in large data environments. Assuming that there are P transmission paths of the transmission node of non-significant feature data in a large data environment, the distribution function of the non-significant feature data center network in the transmission of the cross-platform is described as:

$$\begin{cases} y(t) = x(t - t_0) \Rightarrow W_y(t, v) = W_x(t - t_0, v) \\ y(t) = x(t) e^{j2\pi v_0 t} \Rightarrow W_y(t, v) = W_x(t, v - v_0) \end{cases} \tag{5}$$

The waveform fitting for non-significant feature data collected is made by the formation of a non-significant feature data beam. By using a linear regression to estimate, the scaling scale of the time domain of the non-significant eigendata time series is calculated with Formula (6):

$$\begin{cases} y(t) = \sqrt{k} x(kt), k \succ 0 \\ W_y(t, v) = W_x(kt, v/k) \end{cases} \tag{6}$$

In Formula (6), k is the frequency of data sampling, v is the bandwidth of the non-significant feature data transfer structure distribution, W_x is time window function of data.

The data time sequence for non-significant features in large data environments is $x(t)$, made by using the data time-frequency characteristic analysis method. In the process of accessing the database, the fusion of data information is taken for subsets of different data characteristics, then the non-salient feature data was optimized using the description method of the binary vector and the expression of the optimal time sequence of non-significant feature data is obtained as:

$$h(t) = \sum_i a_i(t) e^{j\theta_i(t)} \delta(t - iT_S) \tag{7}$$

The problem of feature data optimization is transformed into non-significant feature data fusion and the estimation of the relation function vector model is described as:

$$x = \sum_{i=1}^{N} s_i \psi_i = \Psi s,\ \psi = [\Psi_1,\ \Psi_2, \cdots,\ \Psi_N] \tag{8}$$

In Formula (8), s_i is the constraint channel of the non-salient feature data selection path, Ψs is the initial data probability distribution. By using the time combination estimation method, the number of joint parameters for non-salient feature data in large data environments is detected and estimated, the computational expressions associated with non-significant feature data can be described as:

$$\begin{cases} Info(B) = -\sum\limits_{i=1}^{m} p_i \times \log_2 p_i \\ Info_A(B) = -\sum\limits_{j=1}^{v} \frac{B_j}{B} \times Info\left(B_j\right) \\ Gain(A) = Info(B) - Info_A(B) \end{cases} \tag{9}$$

In Formula (9), p_i is the probability that the data exists in vector space, $Info\ (B_j)$ is the relevant information parameters of non-salient feature data flow B_j, $Info_A(B)$ is the relevant information parameters of the data stream A and B, $Gain\ (A)$ is the data transfer gain. The relative functions of the time-series of non-salient feature data were carried out through the above formula, and the collection of non-salient feature data was completed to improve the accuracy of non-significant feature data collection.

2.2 The Extraction of Non-Salient Feature Data

Combined with the acquisition of non-salient feature data above, the mean mutual information data feature extraction method is used to estimate the spectral density of non-salient feature data. Assuming that the feature vector of the data

preference window of non-salient feature data collection in a big data environment can be expressed by Formula (10):

$$E_{i,j} = \langle e_1,\ e_2,\ L,\ e_m \rangle \tag{10}$$

Extracting the data information characteristics by time-frequency analysis method, the time-varying clustering model for non-salient feature data can be expressed as:

$$V = \{C_1, C_2,\ \cdot C_K\} \tag{11}$$

By breaking down the edge features of the non-salient feature data flow, the transfer function of the channel number of non-significant feature data transfer in a large data environment can be expressed as:

$$\begin{cases} W_x(t, v)dt = |X(v)|^2 \\ W_x(t, v)dv = |X(t)|^2 \end{cases} \tag{12}$$

In Formula (12), $|X(v)|$ is the distribution function of non-salient feature data. Creating a short-time window function in the subdomain of the data time scalar and calculating the data correlation density information in the data cluster results in $C_i \subseteq V, 1 \leq i \leq k$. The channel model of non-significant feature data in a large data environment is a hypothesis model, it is represented as:

$$x(k) = \begin{cases} n(k), & H_0 \\ hs(k) + n(k), & H_1 \end{cases} \tag{13}$$

In Formula (13), $x(k)$ is the vector of non-significant feature data fusion input information in a large data environment, $s(k)$ is the component of the non-significant feature data flow filtering. The data limit is α and $SIR \leq \alpha$, and the data is decomposed with the data vertical structure, when the threshold is met: $SIR \succ \alpha$ and $SNR \leq \alpha$. The amplitude of non-significant eigendata measurements in a large data environment can be represented as: $A = \{a_1, a_2, \cdots, a_n\}$, Vector branching the data attribute results in the data output gain represented as: $B = \{b_1, b_2, \cdots, b_m\}$. In the space of data information distribution, there is a sequence of data spatial distribution of B_j, The property value of the orthogonal data eigenvector a_x is the set of elements in the non-significant eigendata space. By measuring the data, the iterative function of the average inter-information feature of a data time series distribution of non-significant eigendata time sequences in a large data environment can be represented as

$$\begin{aligned} x_{id}(t+1) = {} & w x_{id}(t) + c_1 r_1 \left[r_3^{t_0 \succ T_0} P_{id} - x_{id}(t) \right] \\ & + c_2 r_2 \left[r_4^{t_9 \succ T_9} P_{gd} - x_{id}(t) \right] \end{aligned} \tag{14}$$

In Formula (14), t_0 is a data embedded dimension, t_g is the latency of data time. The extracted data above is a test function and the results of the non-significant feature data were obtained as:

$$X = \left[s_1,\ s_2, \cdots s_K \right]_n = \left(x_n, x_{n-\tau}, L, x_{n-(m-1)\tau} \right) \tag{15}$$

In Formula (15), $K = N - (m-1)\tau$ expressed the eigenvectors of the time-series of non-significant feature data in a large data environment, τ is the estimated delay coefficient of non-significant eigendata. Thus, the characteristics of non-salient feature data are extracted.

2.3 Non-Salient Feature Data Clustering

Combining the above data collection and extraction, the data clustering of the extracted feature data is obtained by using Ant Colony Clustering algorithm. As ants search for food, they bring the food they find to a centralized location. The ants release pheromones wherever they go, and the ants move towards the places with more pheromones, the pheromones on the path also evaporate over time.

By analyzing this behavior, the main idea of the cluster analysis model based on ant foraging behavior is to view the extracted data as ants of different attributes, the data clustering center is the food source the ants are looking for, thus the process of clustering is the process of ants finding food. An algorithm flowchart based on the clustering analysis model of ant colony foraging behavior is shown in Figure 1:

Assuming that X is the set of nm dimensional data c_j is the clustering center;

ε_0 is the error of data statistics;

P_0 is the threshold of data probability transfer.

Initialized with any number of data, the probability P_{ij} that if the data object X_i is carried to C_j can be expressed as:

$$P_{ij} = \frac{\tau_{ij}^{\mathrm{a}}(t)\eta_{ij}^{\beta}(t)}{\sum_{s \in S} \tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)} \tag{16}$$

$$\tau_{ij}(t) = \begin{cases} 1, d(X_i, C_j) \leq R \\ 0, d(X_i, C_j) \succ R \end{cases} \tag{17}$$

$$d(X_i, C_j) = \sqrt{\sum_{r=1}^{m} (X_{ir}, C_{jr})^2} \tag{18}$$

In the formula above, $d(X_i, C_j)$ is the Euclidean distance between non-salient feature data object X_i and the clustering center C_j, $\tau_{ij}(t)$ is the amount of pheromone from data object X_i to clustering center C_j at the time t. At initial time $\tau_{ij}(t) = 0$, $S = \{X_s | d(X,\ C) \leq R, s = 1, 2, \cdots, j+1, \cdots n\}$ is a set of data objects in a cluster center, h_{ij} is heuristic factor which can be defined as $\eta_{ij} = 1/d(X_i,\ C_i)$, α and β are the expected value of data pheromones and heuristic factors. If P_{ij} is larger than threshold value P_0, the ant X_i carries food to the scope with C_j existing in it.

The cluster analysis base model is built by combining the behavior that a single ant picks up food and puts down food put forward above. The main idea is that if an ant that is not carrying an object meets a new object that has a significant difference to the objects around it, the probability that this object will be picked up is greater. On the contrary, for an ant already carrying an object, if the objects around it are similar to the object already on its back, the probability that the carried object will be put down is greater. The basic model of the cluster analysis is shown in Figure 2.

Suppose there is an object in the plane of a two-dimensional grid, the probability of a moving unloaded ant picking up an object is:

$$P_P = \left(\frac{k^+}{k^+ + f} \right)^2 \tag{19}$$

The probability of a moving ant already carrying an object on its back putting down the object is:

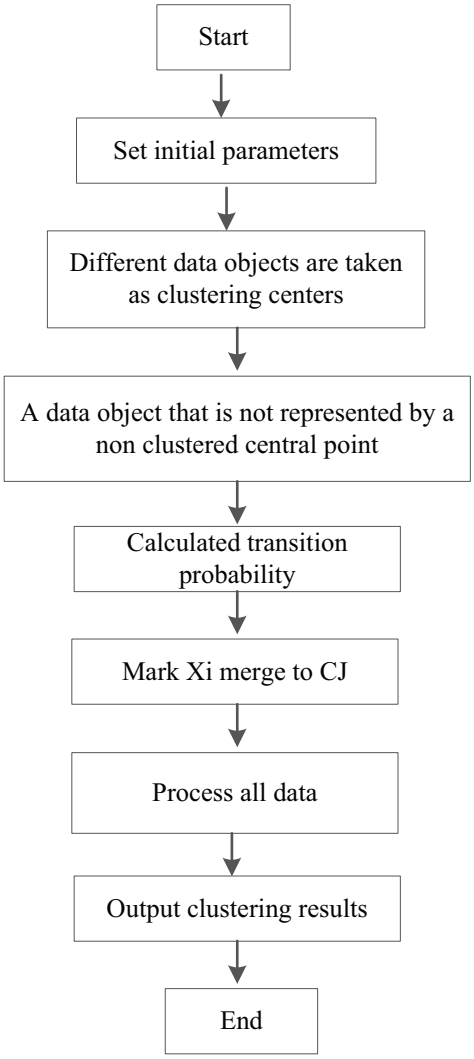

Figure 1 Cluster Analysis Model's Algorithm Flowchart

$$P_d = \left(\frac{f}{k^- + f}\right)^2 \tag{20}$$

In the formula above, f is the number of problems around the ant, k^- and k^+ are two constants.

In Formula (19), when $k^+ \gg f$, $P_p \to 1$ represents that the ant is more likely to pick up the object, when $k^+ \ll f$, $P_p \to 0$ represents that the ant has a small probability of picking up the object, the objects are relatively different from those around them, and it will not be picked up by ants.

In Formula (20), if $k^- \gg f$, $P_d \to 0$ represent that this object is different to the objects around, the ant is less likely to put down the object. If $k^- \ll f$, $P_d \to 1$ represent that the ant is more likely to put down the object, the object is the same as the objects around it, and the ant will put down the object and the clustering is more effective.

The short-term memory that an ant has is defined as f, in time unit T, the number of objects that the ant meets is N, f can be defined as:

$$f = \frac{N}{T} \tag{21}$$

In the non-significant feature data set of mixed properties, the distance of the non-salient feature data objects, the similarity calculation of non-salient feature data objects, the formula for calculating the probability of ants picking up objects and dropping objects and the clustering center and the clustering mean and other computational formulas is expressed by the formulas below:

1. Data objects distance computing formula

Assuming that O_i and O_j represent two data objects in data set.

(1) Each data object contains m character properties, the distant $D_s(0_i,\ O_j)$ is expressed as:

$$D_s(0_i, O_j) = \Sigma_{k=1}^{m} \frac{n_{h_{ik}} + n_{h_{jk}}}{n_{h_{ik}} \cdot n_{h_{jk}}} \delta(x_{ik},\ x_{jk}) \tag{22}$$

In Formula (22), $\delta(x_{ik},\ x_{jk}) = \begin{cases} 0, x_{ik} = x_{jk} \\ 1, x_{ik} \neq x_{jk} \end{cases}$ x_{ik} and x_{jk} are character properties $k(1 \leq k \leq m)$ corresponding to O_i, O_j, $n_{h_{ik}}$ and n_{h_k}, are the numbers of attribute h_k with value h_{ik} and h_{jk} in the data set.

Each data object contains P attributes, which has m character properties and n data numeric attributes, the data distance $D(0_i,\ O_j)$ is:

$$D(0_i,\ O_j) = \mu D_s(0_i, O_j) + D_n(0_i, O_j) \tag{23}$$

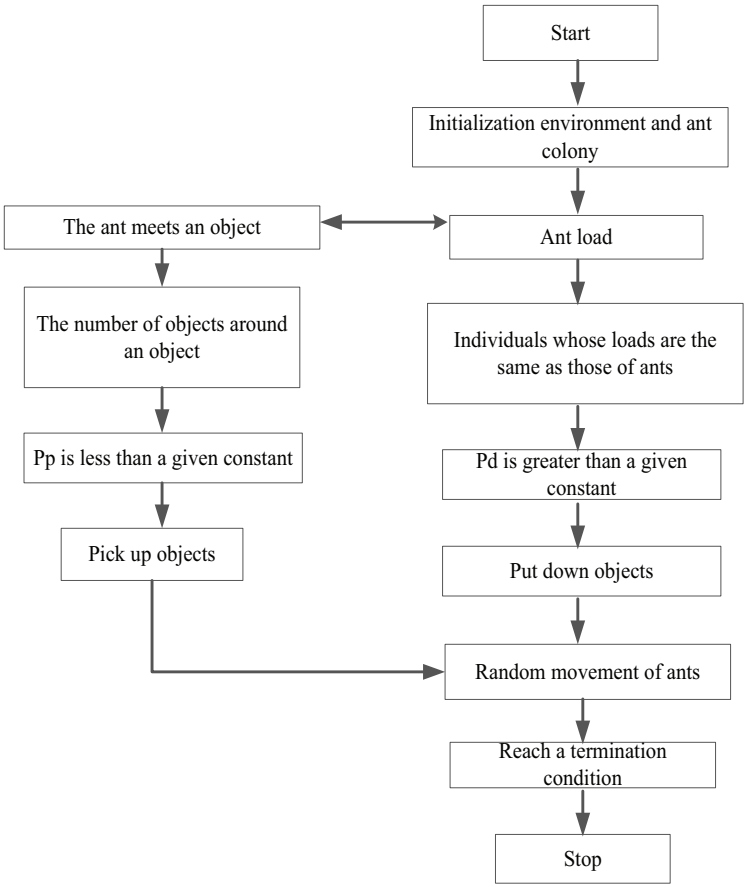

Figure 2 Ant Cluster Model

In Formula (23), μ is the adjustable character attribute weights, $D(0_i, O_j)$ is Euclidean distance of n data values.

2. Calculation of the similarity of the group

The similarity of the group is mainly the degree of similarity between a data object and other data objects in data set, the formula is defined as:

$$f(O_i) = \begin{cases} \sum_{0j \in neigh(r)} \left[1 - \frac{D(O_i, O_i)}{\alpha}\right], f(O_i) \succ 0 \\ 0 \end{cases} \tag{24}$$

In Formula (24), $neigh(r)$ is a circular area with the radius r, α is the adjustable similarity coefcient which determines the number of clusters and the rate of convergence.

3. Formula to calculate the probability of ants picking up objects and dropping objects

The calculation of the probability of ants picking up objects and dropping objects is related to data group similarity, by converting the similarity of data groups into the probability calculation function of an ant to move a non-salient feature data object.

The probability of ants picking up objects and dropping objects is expressed by Formula (25):

$$\begin{cases} P_p = \frac{1}{2} - \frac{1}{\pi} \arctan\left[\frac{f(O_i)}{k}\right] \\ P_d = \frac{1}{2} + \frac{1}{\pi} \arctan\left[\frac{f(O_i)}{k}\right] \end{cases} \tag{25}$$

In Formula (25), k is the adjust length, the bigger the value of k is, the slower the colony algorithm converges, the smaller the value of k is, the faster the colony algorithm converges. According to the above formula, the less similar the non-salient feature data objects of the load to the neighborhood data objects is, which means the data does not belong to the field and the probability of moving is greater, the probability of dropping is smaller. The more similar the non-salient feature data objects of the load to the neighborhood data objects is, means the data belongs to the field and the probability of dropping is bigger.

4. Clustering center formula

The clustering center is the measure of the average of non-significant eigendata objects in the cluster. Assuming that there are m character properties and n numeric attributes in a data object O_i, define the cluster center as:

$$O_{c_j} = \begin{cases} x_{ij} \\ \frac{1}{n}(\Sigma_{O_{i \in T_k}} x_{ij}) \end{cases} \tag{26}$$

In Formula (26), O_{c_j} is the attribute j of clustering center, x_{ij} is the j th attribute of data object O_i.

5. Formula to calculate the mean of data clustering

Data clustering means the average between the data object in the cluster and the center distance. The specific definition is:

$$D_{\text{mean}}(T_k) = \frac{1}{n_k} \Sigma_{O_{i \in T_k}} D\left(O_i,\ O_{c_j}(T_k)\right) \tag{27}$$

In Formula (27), n_k is the number of element in T_k.

6. Data clustering standard deviation

The standard deviation of data clustering describes the degree of deviation from the data and the mean of the cluster, which can be defined as:

Table 1 The Time (s) Comparison of Different Data Feature Extraction Method.

Different data mining methods	Feature extraction time (s)
Method in this paper	15
Method in article [13]	18
Method in article [14]	20

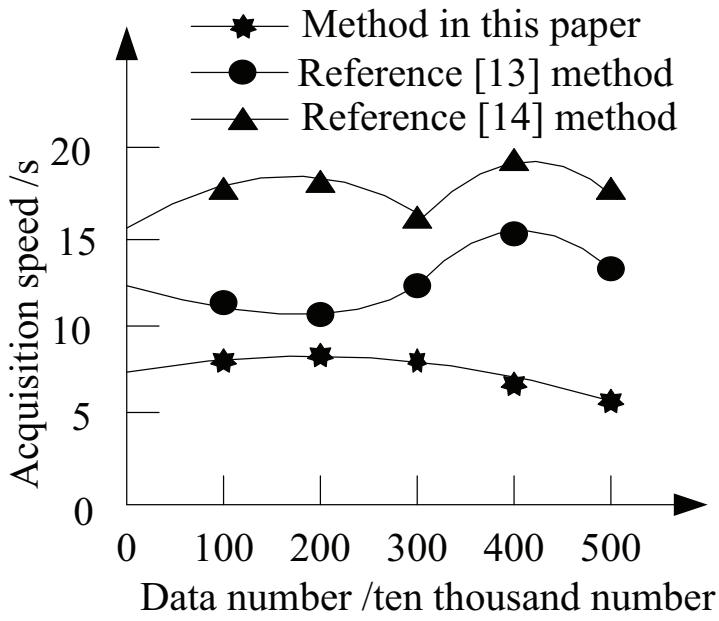

Figure 3 Comparison of Data Collection Speed (s) by Different Methods

$$Dev(T_k) = \frac{1}{n_k}\Sigma_{O_{i\in T_k}} D\left(O_i,\ O_{c_j}(T_k)\right) - D_{mean}(T_k)^2 \quad (28)$$

The non-significant feature data is clustered in this way and this process is called data mining.

3. THE EXPERIMENTAL RESULTS AND ANALYSIS

In order to prove the validity of the non-salient feature data mining method based on ant colony clustering, a simulation experiment is required. In the context of Matlab, the experiment simulation platform of non-significant feature data mining was set up. The data is taken from the KDD cup2016 dataset, which includes 1.5 million samples of web data. In this experiment, the non-significant feature data of the data concentration was excavated by using Ant Colony Clustering algorithm, by this the validity and feasibility of this method are observed. Table 1 shows the time (s) comparison of the data feature extraction method in this paper and the method in the article [13] and article [14].

By analyzing Table 1, the data extraction time of the method mentioned in this paper is significantly faster than that in article [13] and [14]. When extracting data characteristics, this paper uses the average mutual information feature extraction methods, which can effectively extract data that does not contain noise and with the value of data mining, greatly reducing the time of data feature extraction, which proves the validity of the method in this paper. Figure 3 shows the comparison of data collection speed (s) by different methods.

By analysis of Figure 3, the collection of data via the method in this paper is significantly faster than that in article [13] and [14], the data collection speed of the method mentioned in this paper is low at first, however, as the number of data increases, the speed of data acquisition is gradually increasing. While the speed of data acquisition by the method in article [13] is fast when there is a low volume of data as the data grows, the speed of data collection increases significantly, and the volatility is great. For the method mentioned in article [14], when the data reached 3 million, the speed of the data gathering is increased, and in the remaining conditions the data was collected evenly, that shows the method in article [13] and [14] is not feasible, and the method in this paper is highly feasible. Figure 4 shows the comparison of the clustering accuracy (%) by the method described in the article [14] and [15] and the method mentioned in this article. The formula for clustering accuracy is:

$$Clusteringaccuracy = \frac{Clustering\ data}{Total\ clustering\ data} \times 100\% \quad (29)$$

As shown in Figure 4, the method in article [15] has the lowest clustering accuracy, although it stays uniform, the accuracy was never more than 40 percent. The accuracy of method [14] is uniform before 3 million data, however as the volume of data increases, the clustering accuracy is slightly improved, but due to the high volatility, the feasibility is low. The clustering accuracy of the method in this paper stays above 70%, and the fluctuation is uniform. The overall data clustering accuracy is higher compared with the method

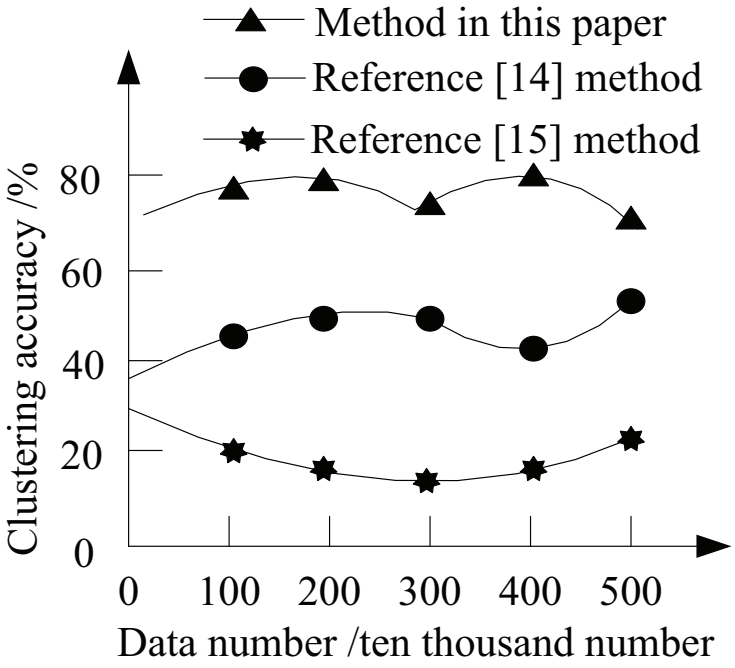

Figure 4 Comparison of the Clustering Accuracy (%) by Different Methods

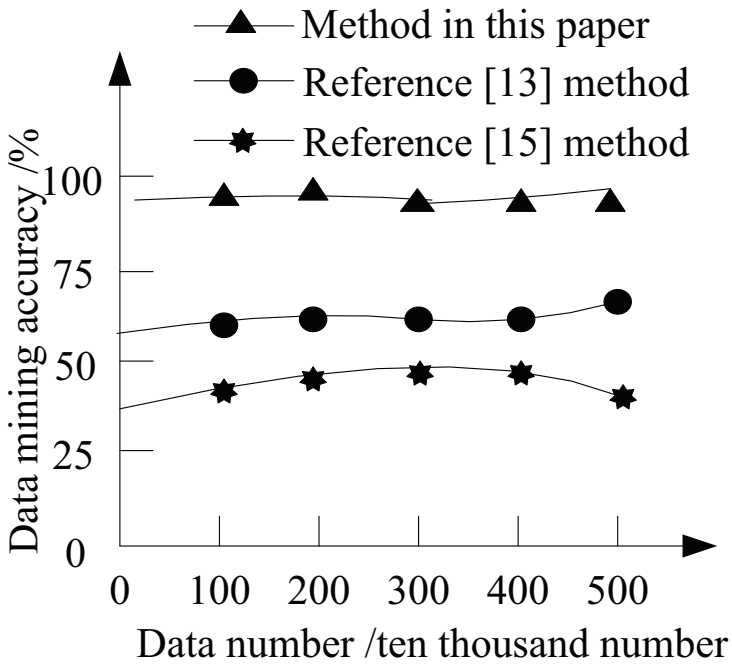

Figure 5 Comparison of Data Mining Accuracy (%) of Different Methods

in article [14] and [15]. Figure 5 shows the comparison of data mining accuracy (%) of the method in article [13] and [15] and in this article.

Figure 5 shows that the accuracy of the non-salient feature data mining of the method in this paper is more uniform and significantly higher than those in article [13] and [15] which is also uniform but the accuracy of data mining is lower and the feasibility of data mining is lower. From which it can be seen that the feasibility of the method in this paper is high, which provides the basis for the study of the non-significant feature data mining method.

The experiments show that the method in this paper can mine the non-significant feature data in a large data environment with high accuracy and effectively reduce the time of data operations.

4. CONCLUSIONS

The current methods for non-salient feature data mining in a large data environment are not stable and efficient, their processes are complex and the efficiency of data mining is

poor. In this paper, a method of mining non-salient feature data based on ant colony clustering is proposed. This method is proved by experiments that it has the ability to mine non-significant feature data in large data environments accurately and it is feasible. This method provides a good foundation for the development of data mining and has broad practical value.

ACKNOWLEDGMENTS

This paper was supported by 2018 Key Research and Development Plan of Shaanxi Province, Project Name: Research on Monitoring Mechanism of Deformation of Tailings Dam Based on Wireless Sensor Network Location, Project No.2018GY-095 and Research and Innovation Team at School Level: Research on Fault Diagnosis Algorithms of Wireless Sensor Networks and Deformation Detection Mechanism of Tailings Reservoir Dam (XAIU-KT201801-3).

REFERENCES

1. Ferrari D.G, De Castro L.N. Clustering Algorithm Selection by Meta-Learning Systems. Information Sciences, 2015, 301(C): 181–194.
2. Guo Y., Sengur A. NCM: Neutrosophic c-means Clustering Algorithm. Pattern Recognition, 2015, 48(8): 2710–2724.
3. Xenaki S.D., Koutroumbas K.D., Rontogiannis A.A. A Novel Adaptive Possibilistic Clustering Algorithm. IEEE Transactions on Fuzzy Systems, 2015, 24(4): 791–810.
4. Zou J., Peng C., Xu H., et al. A Fuzzy Clustering Algorithm-Based Dynamic Equivalent Modeling Method for Wind Farm with DFIG IEEE Transactions on Energy Conversion, 2015, 30(4): 1–9.
5. Chen J.Y, He H.H. A Fast Density-Based Data Stream Clustering Algorithm with Cluster Centers Self-Determined for Mixed Data. Information Sciences, 2016, 345(C): 271–293.
6. Truong D.T., Battiti R. A Flexible Cluster-Oriented Alternative Clustering Algorithm for Choosing from the Pareto Front of Solutions. Machine Learning, 2015, 98(1): 57–91.
7. Le H.S. A Novel Kernel Fuzzy Clustering Algorithm for Geo-Demographic Analysis. Information Sciences, 2015, 317(10): 202–223.
8. Zhao X., Li Y, Zhao Q. Mahalanobis Distance Based on Fuzzy Clustering Algorithm for Image Segmentation. Digital Signal Processing, 2015, 43(C): 8–16.
9. Su L., Jia J. Empirical Research about the Degree of City-Industry Integration: A Contrast of the Typical Cities in China. Journal of Interdisciplinary Mathematics, 2017, 20(1): 87–100.
10. Dutta M.P., Banerjee S., Das M., Goswami R.S., Chakraborty S.K., and Bhunia C.T. Key Variation Technique Based on Piggybacking Strategies Under Public Key Environments. Journal of Discrete Mathematical Sciences and Cryptography, 2018, 21(1): 59–73.
11. Al-Rubaye, L.A.H., Al-Rubaye, A.A.A.H., Al-Samari, A., Zedan, L.Y., Samari, and Zedan, L.Y. Study the Opportunity of Using Lost Energy in IC Engine to Generate Steam that Running a Steam Engine. Journal of Mechanical Engineering Research and Developments, 2018, 41(1): 44–50.
12. Souza, Fren L, Pazzi R.W, et al. A Prediction-Based Clustering Algorithm for Tracking Targets in Quantized Areas for Wireless Sensor Networks. Wireless Networks, 2015, 21(7): 1–16.
13. Lee H.J., Jeong E.J., Kim H., et al. Morphological Feature Extraction from a Continuous Intracranial Pressure Pulse via a Peak Clustering Algorithm. IEEE Transactions on Biomedical Engineering, 2016, 63(10): 2169–2176.
14. Liu J., Guo H.S. K-means Clustering Center Optimization Solution in Cloud Computing. Technology bulletin, 2015, 31 (10): 100–102.
15. Zhao X.J., Zheng Q. Design. Computer measurement and control, 2015, 23(8): 2762–2765, based on PCA-KNN clustering.
16. Yin WH. Research and Search for e-Learning Based on Dynamic Evolution Clustering. Electronic design engineering, 2016, 24 (22): 90–93.
17. Dong B.Q., Peng J.J. Simulation of Abnormal Data Mining in Complex Network Data Streams. Computer simulation, 2016, 33 (1): 434–437.
18. Han Z., Zheng J.S., Chen L.W, et al. Based on MapReduce and HBase's Massive Network Data Processing. Science and technology and engineering, 2015, 15 (34): 182–191.