Modified Immune Evolutionary Algorithm for Medical Data Clustering and Feature Extraction under Cloud Computing Environment

Jing Yu,1 Hang Li,2 and Desheng Liu3

1Luxun Academy of Fine Arts, No. 19, Miyoshi Street, HePing District, Shenyang P. C 110000, China
2Software College, Shenyang Normal University, Shenyang 110034, China
3College of Information and Electronic Technology, Jiamusi University, Jiamusi 154007, Heilongjiang, China

Correspondence should be addressed to Hang Li; lihangsoft@163.com and Desheng Liu; zdhlds@163.com

Received 9 October 2019; Revised 26 November 2019; Accepted 10 December 2019; Published 20 January 2020

Copyright © 2020 Jing Yu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Medical data have the characteristics of particularity and complexity. Big data clustering plays a significant role in the area of medicine. The traditional clustering algorithms are easily falling into local extreme value. It will generate clustering deviation, and the clustering effect is poor. Therefore, we propose a new medical big data clustering algorithm based on the modified immune evolutionary method under cloud computing environment to overcome the above disadvantages in this paper. Firstly, we analyze the big data structure model under cloud computing environment. Secondly, we give the detailed modified immune evolutionary method to cluster medical data including encoding, constructing fitness function, and selecting genetic operators. Finally, the experiments show that this new approach can improve the accuracy of data classification, reduce the error rate, and improve the performance of data mining and feature extraction for medical data clustering.

1. Introduction

Through the support of existing technologies, relevant medical research organizations only rely on coupled dictionary technology to classify and store medical images [1]. However, with the continuous increase of the number of slices, some images begin to show serious frame rate overlap phenomenon, which not only causes the sharp decline of the original image gray level but also causes a series of image data redundancy problems. It brings great trouble to the mining and scheduling of the following image information. The so-called image data redundancy refers to the phenomenon of uneven or excessive storage caused by data repetition in the process of data imaging that can lead to the real information loss in the image and cause a certain negative impact on the image sharpness. Frame rate overlap is a common image fault problem, which is often associated with image data redundancy. Under certain circumstances [2], a certain degree of frame rate overlap may lead to a small increase of the image sharpness. But excessive frame rate overlap will lead to serious damage to the modal property of the medical image, which will lead to a large increase of the redundant region in the medical image data. Diagnosis in medicine is related to the patient’s medication and treatment. Many diseases are more complex. Data clustering analysis is integrated into the diagnosis of diseases, such as clinical urology and breast cancer, so that doctors can greatly enhance the diagnosis accuracy of patients.

With the fast growth of information science, the research of biological applications has been used for computational science to analyze the intelligent bionic optimization algorithm design and improve the ability of processing big data and analysis [3]. Intelligent bionic algorithms mainly include ant colony algorithm [4], particle swarm optimization (PSO) algorithm [5], and the quantum swarm algorithm [6–8]. Swarm intelligence optimization algorithms have a good application value in artificial intelligence design, data clustering analysis, computer control, and other fields.
Clustering technology is an important part in data mining and machine learning. Domestic researchers mainly focus on the following two aspects: (1) a clustering algorithm dynamically determines the number of clustering centers and (2) a clustering algorithm improves the accuracy of clustering. Zhao et al. [9] presented a new dynamic clustering method based on genetic algorithm; the main idea of the method was that, in order to effectively overcome the sensitivity to the initial state value clustering algorithm, it used the maximum attribute value range partitioning strategy and two stages and dynamic selection method in mutation, which obtained the optimal clustering center.

Clustering analysis is a kind of unsupervised model in pattern recognition. The task of cluster is to divide an unmarked pattern according to the certain criteria into several subsets, which requires that similar samples have the most similar cluster center and dissimilar samples should be divided in different classes. Therefore, it is also called unsupervised classification. Clustering analysis has been extensively used in data mining, image processing, object detection, radar target detection, etc. [10, 11]. Zhang et al. [12] proposed a Geometric-constrained multiview image matching method based on semiglobal optimization. It was obvious that some features had more information than others in a dataset. So it was highly likely that some features should have lower importance degrees during a clustering or a classification algorithm due to their lower information, their higher variances, etc. So, it was always a desire for all artificial intelligence communities to enforce the weighting mechanism in any task that identically used a number of features to make a decision. Parvin and Minaei-Bidgoli [13] proposed a weighted locally adaptive clustering algorithm that was based on the locally adaptive clustering algorithm.

Nowadays, different clustering methods are being used to resolve several machine learning problems. According to the clustering criterion, different clustering algorithms can be divided into clustering algorithm based on fuzzy relations and clustering criterion, different clustering algorithms can be divided into clustering algorithm based on fuzzy relations and clustering method based on genetic algorithm; the main idea of the method was that, in order to effectively overcome the sensitivity to the initial state value clustering algorithm, it used the maximum attribute value range partitioning strategy and two stages and dynamic selection method in mutation, which obtained the optimal clustering center.

Moreover, deep learning-based methods are used for feature selection. Minaei-Bidgoli et al. [20] proposed an ensemble based approach for feature selection. The results showed that, although the efficacy of the method was not considerably decreased in most of cases, the method became free from setting of any parameter. Some algorithms could not properly represent data distribution characteristics when datasets were imbalanced. In some cases, the cost of wrong classification could be very high in a sample of a special class, such as wrongly misclassifying cancerous individuals or patients as healthy ones. Hu and Du [21] tried to present a fast and efficient way to learn from imbalanced data. This method was more suitable for learning from the imbalanced data having very little data in class of minority. Gao et al. [22] was devoted to the exploration of brain images for early detection of Parkinson’s disease. All brain images were analyzed to extract Gabor 2D features. It was also shown that the models created on Gabor features outperform the ones created without Gabor features. Zhao et al. [23] analyzed the triple-negative breast neoplasm gene regulatory network using gene expression data. We collected triple-negative breast neoplasm gene expression data from the Cancer Genome Atlas to construct a triple-negative breast neoplasm gene regulatory network using least absolute shrinkage and selection operator regression. In addition, it constructed a triple-positive breast neoplasm network for comparison. Nejatian [24] presented that the available additional information at different times and conditions and gold-standard protein complexes was employed to determine fitting thresholds. By doing so, the problem was converted into an optimization problem. Thereafter, the problem was solved using the firefly metaheuristic optimization algorithm.

Hence, we propose a new medical big data clustering algorithm based on modified immune evolutionary method under cloud computing environment to overcome the above disadvantages in this paper. The reminder of this paper is organized as follows: Section 2 presents big data structure analysis in cloud computing environment. Immune evolutionary algorithm is stated in Section 3. Section 4 describes the improved clustering method in detail, Section 5 provides the MapReduce framework, and Section 6 manifests the experiments results. Finally, the conclusion is given in Section 7.

2. Analysis on Storage Mechanism and Structure of Medical Big Data in Cloud Computing Environment

Cloud computing [25–28] is through the Internet to provide dynamic data to extend large storage space and the structure model. In order to evaluate the data clustering and mining in the cloud computing environment, it needs to build a big data storage system architecture in cloud computing environment. Big data storage structure adopts virtualized storage pool and depends on the computer cluster. From top to bottom, these are the I/O (input/output) virtual computer, USB interface layer sequence,
and disk layer, respectively. Enterprise data center through all kinds of terminal accesses the application service, which makes the calculation of distribution on a large number of distributed computers. When all the cloud computing virtual machines are assigned to the physical machine, it uses the following formula to calculate the global optimal solution in this clustering process. And, it also can assign big data feature clustering center BF_M of the cloud computing on the physical machine P_M according to the optimal solution:

\[
N = \frac{1}{n} \sum_{j=1}^{n} \left| U_{x \alpha j} - U_{x \beta j} \right| + \frac{1}{n} \sum_{j=1}^{n} \left| U_{x \alpha j} - U_{x \beta j} \right| + \frac{1}{n} \sum_{j=1}^{n} \left| U_{x \alpha j} - U_{x \beta j} \right|.
\]  
(1)

The sample is collected and analyzed to determine whether the sample belongs to a typical sample. Assuming that data information stream sample \( S = (X_1, X_2, \ldots, X_k) \) makes sampling in time \((T_1, T_2, \ldots, T_k)\). We divide big data set \( X \) in cloud environment into \( c \) clusters, \( 1 < c < n \). The data segmentation can be transformed as space segmentation. Storage structure central vector of big data is obtained:

\[
V = \{v_{ij} | i = 1, 2, \ldots, c, j = 1, 2, \ldots, s\},
\]  
(2)

where \( V_i \) is the \( i \)-th vector of object cluster feature.

Fuzzy division matrix can be presented as

\[
U = \{\mu_{ik} | i = 1, 2, \ldots, c, k = 1, 2, \ldots, n\}.
\]  
(3)

Redundant data reduction is processed for a single data source. In the process of multichannel QoS demand virtual machine clustering, some parameters are defined as virtual machine set \( V_{MS} = \{V_{M_1}, V_{M_2}, \ldots, V_{M_n}\} \) and physical machine set \( P_{MS} = \{P_{M_1}, P_{M_2}, \ldots, P_{M_n}\} \). Inspiring factor is \( \alpha \), and the expect of inspiring factor is \( \beta \). Biggest mining number is \( I_{\text{max}} \). As a result, uploaded data blocks provide a fixed size of data blocks, which is beneficial to analyze the cloud clustering. Through the big data storage mechanism analysis in cloud computing environment, it provides the accurate data for big data clustering.

Supposing that the time series of information stream is \( \{x(t_0 + i\Delta t)\}, i = 1, 2, \ldots, N - 1 \). \( X \) and \( Y \) are attribute sets. The vector expression of big data clustering space in the cloud computing environment is

\[
\Re = \{r(t_0), r(t_0 + \Delta t), \ldots, r(t_0 + (K - 1)\Delta t)\},
\]  
(4)

where \( r(t) \) is information stream time series of big data clustering in cloud computing environment and \( \Delta t \) is data sampling interval. The spectral characteristic \( X_p(u) \) of discrete samples of big data can be calculated as

\[
X_p(u) = s_\alpha(t) e^{j2\pi f_t u} = \frac{1}{\sqrt{T}} \text{rect} \left( \frac{u}{T} \right) e^{j2\pi (f_t + Kf^2)/2},
\]  
(5)

where \( s_\alpha(t) \) is the characteristic scalar time series of big data, \( e^{j2\pi f_t u} \) is the discrete sample center of big data clustering, and \( (F, Q) \) is sample data high-order Bessel function statistics of data set \( \{X_1, X_2, \ldots, X_N\} \). So, we can get the confidence and confidence interval:

\[
x_{i,d}^{k+1} = x_{i,d}^{k} + F \cdot \left(x_{i,d}^{k} - x_{i,d}^{k}\right),
\]

\[
u_{i,d}^{k+1} = \begin{cases} x_{i,d}^{k+1}, & f_{\text{fin}} < f_{\text{fit}}' \\ x_{i,d}^{k+1}, & f_{\text{fit}}' \geq f_{\text{fin}}' \end{cases},
\]  
(6)

Suppose the information flow time series in the cloud computing environment is \( \{x(t_0 + i\Delta t)\}, i = 0, 1, \ldots, N - 1 \). Let \( X \) and \( Y \) be the set of properties. The expression of clustering space vector of big data in cloud computing environment is as follows:

\[
X = [x(t_0), x(t_0 + \Delta t), \ldots, x(t_0 + (K - 1)\Delta t)]
\]

\[
= \begin{bmatrix} x(t_0) & x(t_0 + \Delta t) & \cdots & x(t_0 + (K - 1)\Delta t) \\ x(t_0 + J\Delta t) & x(t_0 + (J + 1)\Delta t) & \cdots & x(t_0 + (K - 1)\Delta t + J\Delta t) \\ \vdots & \vdots & \ddots & \vdots \\ x(t_0 + (m - 1)J\Delta t) & x(t_0 + (1 + (m - 1)J)\Delta t) & \cdots & x(t_0 + (N - 1)\Delta t) \end{bmatrix},
\]  
(7)

\[
X_p(u) = s_\alpha(t) e^{j2\pi f_t u} = \frac{1}{\sqrt{T}} \text{rect} \left( \frac{u}{T} \right) e^{j2\pi (f_t + Kf^2)/2},
\]  
(8)

where \( s_\alpha(t) \) is the characteristic scalar time series of big data and \( e^{j2\pi f_t u} \) is the center of discrete sample of big data clustering.

The data set is \( \{X_1, X_2, \ldots, X_n\} \). \( (F, Q) \) is the high-order Bessel function statistics of the sample data to determine the confidence of node data packets and establish the confidence
interval. The obtained confidence and confidence intervals are
\[
\begin{align*}
Z_{(t,d)}^{(k+1)} &= x_{r_1}^k + F \cdot (x_{r_2}^k - x_{r_1}^k), \\
H_{(t,d)}^{(k+1)} &= \begin{cases} 
Z_{(t,d)}^{(r+1)}, & f_1^{\text{fitness}} < f_2^{\text{fitness}}, \\
H_{(t,d)}^{(k+1)}, & f_1^{\text{fitness}} \geq f_2^{\text{fitness}}.
\end{cases}
\end{align*}
\]

3. Immune Evolutionary Algorithm (IEA)

IEA consists of crossover and mutation operator which represent two strategies with group search and information exchange. It provides optimization opportunities for each individual. However, this inevitably produces the degradation phenomenon in some cases, and the degradation phenomenon is quite obvious.

IEA uses some features or knowledge in original problems to suppress the degradation phenomenon appeared in the process of optimization. The key operation of IEA is to construct the structure of immune operator that is finished through vaccination and immune selection. The immune evolutionary algorithm can improve the fitness of the individual and prevent the group degradation, so as to reduce the original wave phenomenon in the late evolutionary algorithm and improve the convergence speed. The main steps for immune evolutionary algorithm are as follows, and the detailed information can be obtained from [29, 30].

1. Randomly generate the initial parent group \( A_1 \).
2. Extract the vaccine according to prior knowledge.
3. If the current group contains the best individual, it stops running the process and outputs the result.
4. Otherwise, the procedure continues to work.
5. Cross operation of the current \( k \)-th group \( A_k \) is conducted, and it obtains the population \( B_k \).
6. It makes mutation operation for \( B_k \) and obtains the population \( C_k \).
7. It executes vaccination for \( C_k \) and gets group \( D_k \).
8. It executes immune selection for \( D_k \) and obtains new parent group \( A_{k+1} \). Then back to step 3.

4. Modified Immune Evolutionary Algorithm for Data Clustering

Fuzzy clustering is regarded as one of the commonly used approaches for data analysis. The Fuzzy C-means (FCM) algorithm is the most well-known and widely used method for fuzzy clustering and provides an optimal way to construct fuzzy information granules [31]. Cluster prototypes and membership values of data across all clusters can be developed by optimizing the FCM clustering model. Basically, the FCM is a steepest-descent algorithm with variable step length that is adjusted according to the majorization principle for the step length, showing the simplicity and efficiency of the algorithm. Therefore, we combine immune evolutionary algorithm and FCM to optimize the cluster result [32, 33]. The detailed improved data clustering processes are as follows:

The objective function of FCM is
\[
J(X;U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m D_{ik}^2,
\]
where \( D_{ik} \) is the distance from \( k \)-th data point to \( i \)-th cluster center, \( V = (v_1, v_2, \ldots, v_c) \) denotes the cluster center of each class, and \( v_j \in R \) and \( m \in (1, \infty) \) are fuzzy index:
\[
X = (x_1, x_2, \ldots, x_n) \subset R,
\]
\[
U = \left\{ U \in R^{c \times n} | u_{ik} \in [0, 1]; \sum_{i=1}^{c} v_{ik} = 1; 0 < \sum_{k=1}^{n} v_{ik} < n \right\}.
\]

4.1. Encoding. According to \( J(X;U,V) \), the aim of cluster is to obtain fuzzy division matrix \( U \) and cluster prototype \( V \) of sample \( X \). \( U \) and \( V \) are associated with each other. So we have two encoding methods. First, we encode \( U \). Suppose that \( n \) samples need to be divided into \( c \) clusters. Gene cluster \( a = \{a_1, a_2, \ldots, a_l\} \) denotes one clustering result; \( a_i \in \{1, 2, \ldots, c\} \). When \( a_i = k (1 \leq k \leq c) \), then \( x_i \) belongs to \( k \)-th cluster. Its search space is \( c^n \). If the data samples are bigger, the search space of this encoding is very big too. Therefore, we adopt the second encoding method for \( V \). The quantized values are encoded into strings according to their respective values. \( a = \{a_1, a_2, \ldots, a_l\} \), \( l = c \times p \). The former \( p \) quantized values denote the first \( p \) dimension cluster center. But it does not change with the data sample \( n \).

4.2. Constructing Fitness Function. According to \( J(X;U,V) \), if the clustering effect is better, the object function value is smaller. The formula (10) is used for constructing fitness function \( f \):
\[
f = \frac{1}{J(X;U,V) + 1}.
\]

4.3. Genetic Operator Selection. Genetic operator has a point crossover, two-point crossover, and multipoint crossover methods. The immune operator inverts the selected individual genes based on certain probability. We can also adopt a reverse genetic mutation operator, namely, it randomly generates a gene in the parent group and the gene is reverted. It basically prevents premature phenomenon. In genetic selection methods, it adopts the roulette wheel selection method and ranking selection. Crossover probability \( p_c \in [0.75, 0.95] \), \( p_m \in [10^{-3}, 10^{-2}] \).

4.4. Immune Vaccine Selection. The immune vaccine selection properly describe two ways. It is not clear. Specifically, the first method, after collecting information,
executes the immune vaccine. The other is an adaptive method, namely, in the process of group evolution from the best individual genes. It extracts useful information and then executes the vaccine. The former is restricted due to two reasons. The first one is it is difficult to form a mature approach for a prior knowledge. It cannot get effective immune vaccine. The second is, to extract the vaccine, the work costs too much. Therefore, in the clustering algorithm based on immune evolution, we adopt the adaptive method to extract the vaccine.

Therefore, we get the new cluster algorithm as follows (Figure 1).

**Step 1.** Fix cluster class number \(c, 1 \leq c \leq n - 1\). Set fuzzy index \(m \in (1, +\infty)\), stop condition \(\tau\), total population number \(p_n\), crossover probability \(p_c\), mutation probability \(p_m\), vaccination probability \(p_v\), and vaccine update probability \(p_u\).

**Step 2.** Randomly generate group \(P(k)\) with \(p_n\) individuals.

**Step 3.** Compute fitness of every individual.

1. Each individual is decoded to calculate each prototype parameter \(v_i, 1 \leq i \leq c\).
2. Use \(v_i\) and (8) to calculate \(D_{ik}^2\).
3. Calculate \(U = \left[u_{ik}\right]_{c \times n}\).

   If \(I_k = \emptyset\),
   \[
   u_{ik} = \frac{1}{\sum_{j=1}^{c} \left[\frac{d_{ik}^2}{\sum_{j=1}^{c} d_{jk}^2}\right]^{\frac{1}{(1/m-1)}}}.
   \tag{14}
   
   If \(I_k \neq \emptyset\),
   \[
   u_{ik} = 0, \quad \forall i \in \sim I_k,
   \sum_{i \in I_k} u_{ik} = 1,
   \tag{15}
   
   where \(I_k = \{i \mid 1 \leq i \leq c, d_{ik} = 0\}\) and \(\sim I_k = \{1, 2, \ldots, c\} - I_k\).
4. Use \(U, D_{ik}^2\) and (7) to calculate object function \(f(X; U, V)\), and then it can get \(f\) for each individual.

**Step 4.** Make statistics for parent group, determine the best individual, then decompose the best individual, and extract immune vaccine \(H = \{h_i \mid i = 1 - m\}\).

**Step 5.** Use \(p_c\) and \(p_m\) to make crossover, mutation operation for \(P(k)\), and get group \(P'(k)\).

**Step 6.** Execute vaccination and immunization selection for \(P(k)\) and get group \(P(k + 1)\).

**Step 7.** If it satisfies \(\tau\), return to Step 8. Otherwise, return to Step 3.

**Step 8.** Then, it decodes the best individual, the clustering prototype \(v_i\) is calculated, the classification results of each sample are calculated, and this classification result is the clustering result of data set \(X\).

5. MapReduce Framework

In order to improve the efficiency of modified immune evolutionary algorithm (MIEA) in processing large datasets, this paper designs the implementation scheme of MIEA in the MapReduce model. There are two main operations in the mechanism processing big data clustering tasks: updating the center of the class and fitness evaluation. Class center is updated based on MIEA. Fitness evaluation is to calculate the sum of Euclidean distance between each object and the center of mass and then find the global optimal value. The clustering program divides data objects into clusters, minimizes the sum of Euclidean distances between all objects and the center of mass, and takes it as the fitness function of MIEA. The data clustering process based on MIEA is shown in Figure 2.

6. Experiments and Analysis

In order to verify the performance of clustering and data mining in cloud computing environment, we conduct abundant experiments. Medical data are taken from http://archive.ics.uci.edu/ml/. The database is constantly updated. Donations of data are also accepted. The database type involves life, engineering, science, etc.; the record number is from several to hundred thousand pieces. The data selected in this paper are Breast Cancer Wisconsin (Original) Data Set. These data sets are from the clinical case reports of the university of Wisconsin hospital in the United States, and each data has 11 attributes.

Due to limited space, we display only few results in here. The computing platform is configured with Intel Core i7 4.0 GHz CPU, 16G Memory, and NVIDIA GTX 780 GPU. The algorithm is compiled by Apache Hadoop platform. The sampling frequency of big data is \(f_s = 20\) kHz. The time center of big data clustering is \(t_0 = 20s\). Size of the data is from 50 MB to 2 GB. Cross probability \(p_c = 0.95\), variation probability \(p_v = 0.3\), and fuzzy index \(m = 2\). We also select three state-of-the-art clustering methods to make comparisons including HGM [34], WPC [35], and ACCH [36].

6.1. Result 1. Table 1 is the description of 11 attributes of this dataset.

In this paper, the proposed algorithm is adopted to calculate the weight of each feature. Features with the weight less than a certain threshold will be removed. According to the actual situation in this paper, 2 and 3 with the smallest weight will be removed. In the process of the algorithm, we will randomly select sample \(R\). Different random numbers will lead to certain discrepancy in the weight of the result. Therefore, this paper adopts the average method by running it for 20 times. Then, we summarize the results to calculate the average value of each weight as shown in Figure 3.

By analyzing the data set, the importance of attribute weight can be obtained, which has some reference values for clinical diagnosis and can be used for the analysis of actual cases. This can avoid misdiagnosis as far as possible and
improve the diagnosis speed and accuracy. According to the attributes, we obtain the object function optimal value as given in Table 2.

6.2. Result 2. To evaluate the performance of proposed algorithm, the composite data sets given in Table 3 are adopted. The four public data sets are assembled into a large data set, all of which are from UCI Machine Learning Repository with different attributes. Four data sets are randomly copied into several backups to form a large data set with $10^7$ records.

$F$-measure is adopted as the evaluation index of clustering quality. $F$-measure is calculated from two information indexes, precision, and recall rate, defined as

$$F(i, j) = \frac{2 \cdot r(i, j) \cdot p(i, j)}{r(i, j) + p(i, j)},$$

(16)
where $j$ represents the class generated by the cluster method, $i$ denotes the class label of original dataset, and $r$ and $p$ represent recall rate and precision, respectively. Recall rate is defined as $r(i, j) = \frac{n_{ij}}{n_i}$. Precision is defined as $p(i, j) = \frac{n_{ij}}{n_j}$. Here, $n_{ij}$ represents the divided class number of class $i$. $n_i$ and $n_j$ are the data sizes of class $i$ and
class \( j \), respectively. For the data set with size \( n \), the calculation formula of \( F \)-measure is

\[
F = \frac{\sum_i n_i \max_j (F(i, j))}{\sum_i n_i}
\]  (17)

where the upper bound of \( F \) is 1. If the \( F \)-measure value is larger, then the clustering quality will be higher as shown in Table 4. With the increase of dataset number, the \( F \) value is slightly on the whole. However, the value of 0.817 of the proposed method is still higher than that of HGM, WPC, and ACCH.

The following experiments are for the feature extraction under cloud computation.

The original big data feature distribution is random as shown in Figure 4, and it is difficult to achieve feature extraction in the two-dimensional space regularity. We use the proposed algorithm for feature extraction and processing data clustering to build big data feature extraction model.
Figure 9: ACCH method.

Figure 10: Proposed method.
The obtained feature extraction results are shown in Figure 5.

As can be seen in Figure 5, the proposed algorithm can effectively evaluate the feature extraction of big data in cloud computing; the beam focusing performance is good, which provides accurate basis for the data optimal clustering. Using different big data clustering optimization algorithms, we get the clustering center optimal performance curve as shown in Figure 6.

We also get the best value and mean value within 200 iterations as shown in Figures 7–10. And, we can know that the best value is with our proposed method.

7. Conclusions

In cloud computing environment, vast amounts of data need to be scheduled and accessed aiming at achieving the goal of medical data mining. This paper presents a new medical big data clustering algorithm based on modified immune algorithm. It firstly analyzes the big data structure model in the cloud computing environment to build big data feature extraction and information model. Designing immune optimization algorithm for clustering, it achieves the goal of optimization clustering for big data. Simulation results show that the proposed algorithm improves the clustering performance of big data in cloud computing environment. The new algorithm is used for IoT data clustering, which reduces the error rate and exhibits better performance. In the future, we will research the deep learning methods and apply them into actual engineering projects.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Heilongjiang Province Science Found for Returnees (grant number: LC2017027), Jiamusi University Science and Technology Innovation Team Construction Project (grant number: CXTDZY-2016-3), and Basic Research Project of Heilongjiang Province Department of Education (grant number: 2016-kyyw-0547).

References


