IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received November 9, 2019, accepted November 24, 2019, date of publication November 29, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956751

Health Big Data Classification Using Improved Radial Basis Function Neural Network and Nearest Neighbor Propagation Algorithm

CONGSHI JIANG AND YIHONG LI
School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430000, China
Corresponding author: Yihong Li (acedemiclyh@163.com)

ABSTRACT Health big data classification can effectively improve the level of medical and health services and management, help medical staff to carry out auxiliary diagnosis, improve the efficiency of doctors and the accuracy of diagnosis. To solve the problem of health big data classification, this paper proposes a Radial Basis Function (RBF) neural network classification algorithm based on manifold analysis and nearest neighbor propagation (AP) algorithm. First, the data set is processed by manifold analysis algorithm. Then, the similarity matrix is adjusted by exponential function, and AP clustering is carried out again. On this basis, RBF neural network classifier is constructed. In order to improve the classification accuracy and shorten the convergence time, an algorithm for constructing the variable basis width neural network model is proposed. This method is based on the subtraction clustering algorithm and K-means algorithm to determine the cluster center. The maximum distance between the sample and the cluster center is selected as the base width. The base width is updated adaptively with the optimization of the cluster center. Finally, three data sets of patients with coronary heart disease, diabetes mellitus and bronchial tuberculosis were collected as test data, and the accuracy of classification and convergence time were compared. Experimental results show that this method can improve the classification accuracy and convergence speed of large data sample set.

INDEX TERMS Health big data classification, radial basis function neural network, neighbor propagation algorithm, manifold analysis, convergence time, similarity matrix.

I. INTRODUCTION

With the rapid development and application of mobile Internet, computer storage technology and cloud computing technology, human society is entering an era of information explosion. With the development and implementation of medical and health information construction, the type and scale of health data are growing rapidly at an unprecedented speed, resulting in massive medical and health data [1]–[4]. According to IDC International Data Corporation, by 2020, the total amount of global medical data will reach 40 trillion GB. The main characteristics of medical record data are relatively high dimension and complex type. The main contents include: The basic personal information of the patient, the examination information of the patient in the process of medical treatment, the inpatient information of the patient and the related expense information etc. The medical record information is rich in doctors' diagnosis information, and it can help doctors to treat diseases better with the help of disease history data.

With the rapid development of medical information and the rapid accumulation of medical records, rich medical records information can provide better medical support for auxiliary diagnosis. Health big data has a variety of complex forms of existence, contains rich medical value, and health big data can make the whole medical resources more efficient. For example, through the method of artificial intelligence, the high-quality doctor resources can be infinitely copied, so that the limited medical resources can be allocated in a more reasonable way, and the hierarchical diagnosis and treatment can be more reasonable and perfect [5], [6]. At the same time, through the analysis of health big data, government agencies can achieve reasonable pricing of drugs, timely detection of epidemic diseases and take relevant preventive measures, health public opinion monitoring, etc., to improve the government's decision-making and early warning control ability. Therefore, in the era of big data, how to effectively

The associate editor coordinating the review of this manuscript and approving it for publication was Yongtao Hao.

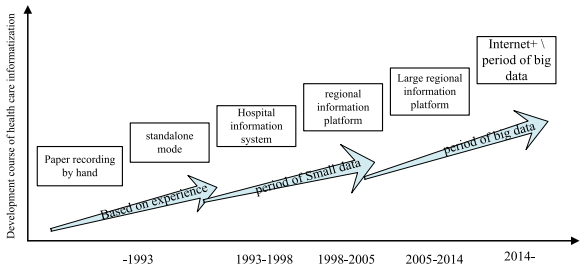


FIGURE 1. Development process of health care informatization.

analyze and deal with healthy big data has gradually attracted great attention of the people and the government. The essence of health big data classification is to merge the health big data with some common attributes or features, and distinguish them through the attributes or features of the categories [7], [8]. Health big data classification can effectively improve the level of medical and health services and management, help medical staff to carry out auxiliary diagnosis, improve the efficiency of doctors and the accuracy of diagnosis, so as to make precision medicine truly possible. Based on this, this paper proposes a new method of health big data classification using radial basis function neural network, and proposes an algorithm of constructing variable basis width neural network model. On the basis of subtracting clustering algorithm and K - means algorithm to determine clustering center, the maximum distance between sample and clustering center is selected as the base width, and the value of base width is adaptive with the optimization of clustering center the new algorithm improves the classification accuracy and shortens the convergence time.

II. RELATED RESEARCH

In the middle of the 20th century, medical informatization began. After that, with the increasing popularity of Internet mobile devices and the demand of social development, machine learning technology has gradually penetrated into the medical industry, and has made some achievements in auxiliary diagnosis, drug development and health management.

With the rapid development of information technology and social changes, traditional medical and health records are gradually transformed from paper documents to digital storage, and gradually enter the era of big data in the medical industry. The development process of health and medical information is shown in Figure 1. Health big data has the characteristics of redundancy, polymorphism, privacy, timeliness and incompleteness, so the analysis and processing of health big data is a new challenge. Common methods of data analysis for health big data include feature analysis, classification, regression analysis, clustering and visualization [9]. Combining machine learning technology to infer and judge different types of test data, medical record data and medical image data is a new way to strengthen medical services, predict disease conditions and reduce medical costs.

With the development of information technology, many platforms about health big data have been opened for researchers to use, so more and more researchers focus on the classification application of health big data, and have made some achievements. For example, reference [10] developed an Intelligent Heart Disease Prediction System (IHDPS) using traditional machine learning technology. Based on decision tree, naive Bayes, neural network and other algorithms, the system selects 15 vital signs such as age, gender, serum, blood pressure and so on to predict the probability of heart disease, and proves the accuracy of machine learning algorithm in heart disease prediction through experiments. In reference [11], Bayesian network, k-nearest neighbor algorithm, support vector machine and other algorithms are integrated, and the method of random subspace division is used for automatic diagnosis of breast cancer. Experiments show that the classifier can effectively distinguish breast cancer patients. In reference [12], naive Bayesian classifier was used to diagnose diabetic retinopathy. Through the combination of morphological coarse segmentation and naive Bayesian fine segmentation to process the diabetic retinopathy image, 18 kinds of micro aneurysm feature attributes are extracted as the input of naive Bayesian classifier, and then the micro aneurysm classifier is established to complete the detection of diabetic retinopathy. The sensitivity, specificity and accuracy of naive Bayesian classifier are 85.68%, 99.99% and 83.34% respectively.

In reference [13], the improved random forest algorithm is used to assist diagnosis of breast cancer. The results of 5-fold cross validation show that the average accuracy rate of detecting breast cancer in 683 patients is 96.93%. Reference [14] used support vector machine, decision tree and naive Bayes to establish classifiers for sarcoidosis and tuberculosis data respectively. Through experiments on 106 cases of sarcoidosis and tuberculosis data collected, it was found that SVM has the highest classification ability and accuracy, and can effectively carry out clinical differential diagnosis for sarcoidosis and tuberculosis. Reference [15] Based on the patient's magnetic resonance image (MRI), artificial neural network (ANN) was used to assist the diagnosis of DMD, which alleviated the pain brought by the traditional diagnosis and detection scheme. The experimental results show that the sensitivity, specificity and accuracy of ANN algorithm are 98.5%, 97.3% and 97.9% respectively.

In a word, through the research of health big data classification technology, it can not only detect potential diseases, assist doctors in diagnosis, but also serve people, government, medical institutions and scientific research institutions. Therefore, the classification technology in the field of health is an area that needs to be further studied.

III. DATA SET PREPROCESSING BASED ON MANIFOLD NANLYSIS AND AP ALGORITHM

A. MANIFOLD NANLYSIS ALGORITHM

Classification is to use the known attribute data of data set to speculate some unknown discrete attribute data. The key to accurate speculation is to build an effective model between the known attribute information and the unknown

discrete attribute, that is, the classification model. Therefore, classification problem includes two processes: learning and classification. The health big data has the characteristics of redundancy, polymorphism and incompleteness, so the traditional supervised learning algorithm must carry on the preprocessing and feature selection before training the classifier. Manifold analysis is to analyze the data from the perspective of observation space. In fact, it is to cluster the data set based on the idea of neighborhood and the threshold of gap between classes, and then delete the clusters with less samples in the class, which will eliminate the influence of some isolated points on AP algorithm to a certain extent. The main steps are as follows:

a. For datasets

$$D_1 = \{x_i | i = 1, 2, \ldots, n\} \tag{1}$$

The threshold χ of inter class gap and the threshold δ of isolated cluster decision are initialized, and the similarity matrix $\boldsymbol{S}$ is calculated.

b. Any sample x_i from dataset D_1 is added to class C_1, and all samples whose distance from sample x_i is less than χ are added to class C_1, and then all samples whose distance from sample x_i is less than x_i are added to class C_1 in all non class C_1 samples, and so on. Finally, all samples in class C_1 are related to the reachable neighbor sample set, and class C_1 samples are counted The total number of books is recorded as C_{1_num}.

c. For all samples in dataset D_1 except for class C_1, obtain Class C_2 according to the method in the previous step, record the total number of Class C_2 samples as C_{2_num}, and so on, until all samples in dataset D_1 are processed, a total of K classifications $\{C_1, C_2, \ldots, C_K\}$ are obtained.

d. Each class is judged as an isolated cluster, if the number of samples is

$$C_{i\_num} \leq \delta \quad (i = 1, 2, \cdots, K) \tag{2}$$

all the samples in the class will be deleted from the dataset, and a new dataset will be formed. A cluster is called a data manifold.

B. ADJUSTMENT OF SIMILARITY MATRIX

The AP algorithm is based on the similarity matrix of the data set, but the calculation of the similarity matrix does not consider the spatial characteristics of the data set. If the distance between two data points on different data manifolds is relatively close, then the distance between these two data points will be smaller than the distance between most data points in the same first-class shape. From the perspective of data space, the data points near the cluster center and their neighbors should be in the same manifold, which is likely to become the same cluster in the later clustering results; Therefore, if we can increase the similarity between data points on the same data manifold (actually shorten the distance between them), and reduce the similarity between data points on different data manifolds, then we can effectively improve the clustering accuracy of AP algorithm. Therefore, considering the concave curve characteristics of the exponential function (when the independent variable is small, the small change will lead to the sharp change of the value), the following formula can be used to adjust the distance between samples. Therefore, considering the concave curve characteristics of index function $y = a^x$ $(a > 1)$ (when the independent variable $x > 1$, a small change of x will cause a sharp change of y), the distance between samples can be adjusted by the following formula.

$$D'\left(x_i, x_j\right) = \begin{cases} \theta^{D\left(x_i, x_j\right)/\tau_{ij}-1} \\ \left(x_i \text{ and } x_j \text{ are not reachable with respect to} \chi\right); \\ \theta^{\chi/\left[\tau_{ij} \max\left(d\left(x_i, x_j\right)\right)\right]-1} \\ \left(x_i \text{ and } x_j \text{ about } \chi \text{ reachable}\right) \end{cases} \tag{3}$$

where: $\theta > 1$; $D\left(x_i, x_j\right)$ is the distance between samples before adjustment, τ_{ij} is the amplitude limit parameter, $\max\left(d\left(x_i, x_j\right)\right)$ is the maximum distance between samples in the same manifold. In this way, the data points in the same data manifold are transformed to an approximate hypersphere centered on x_i. The improved AP algorithm is as follows: according to the similarity matrix $\boldsymbol{S}$ of N data points, the manifold analysis is carried out. After removing the isolated clusters, the new data set D_2 containing m data manifolds is obtained. Then, the similarity matrix is adjusted according to formula (1), which is recorded as $\boldsymbol{S}_2$, and then the AP clustering is carried out again with $\boldsymbol{S}_2$ as the object.

IV. THE CONSTRUCTION OF VARIABLE BASE WIDTH RADIAL BASIS NEURAL NETWORK CLASSIFIER

A. RADIAL BASIS FUNCTION NEURAL NETWORK

Radial Basis Function Neural Network (RBFNN) is a special three-layer feedforward network with a single hidden layer, which not only has global approximation performance, but also has the best local approximation performance [16]–[18]. Its application has penetrated into various fields. RBFNN is used in sample classification, how to improve the accuracy and convergence speed of RBF neural network model has been a problem that many scholars are committed to. When building the neural network model, many scholars combine the fuzzy clustering algorithm with the artificial neural network, but this method needs a large number of training samples in the training process, and the application of this method to the classification of large data sample set may lead to the instability of classification accuracy. There are also many scholars applying genetic algorithm and particle swarm optimization algorithm to the construction of radial basis function neural network structure, but the convergence time of classification may be longer.

So far, there is still no very reliable algorithm to solve this problem, and in the existing RBF neural network model construction, the basis width of Gaussian function is fixed, this method of determining the basis width makes the network in the training process, with the shortcomings of slow convergence speed. In this paper, a radial basis function neural

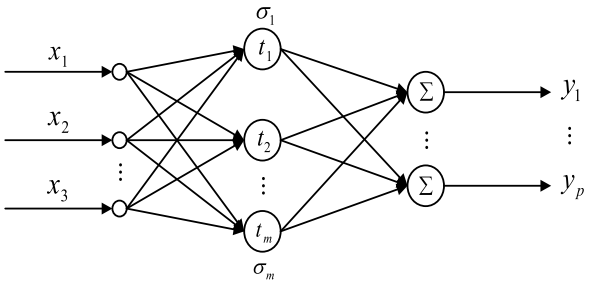


FIGURE 2. RBF neural network structure of Gaussian function.

network method with variable basis width factor is proposed, which can greatly reduce the requirements of sample parameters and improve the accuracy of neural network model in classification.

B. RADIAL BASIS FUNCTION NEURAL NETWORK MODEL STRUCTTURE

The radial basis function neural network model consists of three layers: (1) the input layer is composed of input vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$; (2) there are m neurons in the hidden layer, and the transfer function of each neuron is h_i; (3) the output layer is composed of the output information of the hidden layer of RBF neural network. The relationship between input and output of RBF neural network is a kind of mapping relationship $f(x) : \boldsymbol{R}_n \rightarrow \boldsymbol{R}$. The connection weight between the base function of the i-th neuron in the hidden layer and the j-th output unit in the output layer is w_{ji}, and the center of the base function of the i-th neuron in the hidden layer is t_i, and the width of the base function of the i-th neuron is σ_i. The structure of RBF neural network is shown in Figure 1.

Taking the structure of multi input and single output as an example, $\boldsymbol{X} = [x_1, x_2, x_3, x_4, \cdots\cdots, x_m]^T$ represents the input vector in the structure of RBF neural network. The corresponding RBF neural network approximation principle is shown in Figure 3.

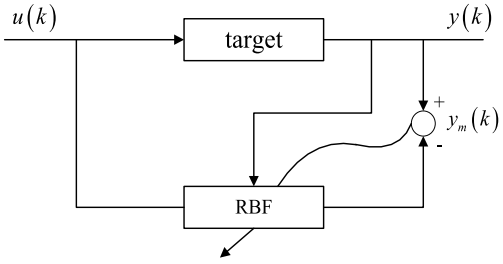


FIGURE 3. Approximation principle of RBF neural network.

Let the radial basis vector $\boldsymbol{H} = [h_1, h_2, h_3, h_4, \cdots\cdots, h_m]^T$ and the Gaussian function $h_i = \exp\left(-\left\|X - C_j\right\|^2 / 2b_j^2\right)$, $j = 1, 2, \cdots, n$ of RBF network. Where: $\boldsymbol{C} = \left[c_{j,1}, c_{j,2}, c_{j,3}, c_{j,4}, \cdots\cdots, c_{j,m}\right]^T, j = 1, 2, \cdots, n$, the center vector of the j-th node in the network, has

$$\boldsymbol{C} = \begin{bmatrix} c_{1,1} & c_{2,1} & \cdots & c_{j,1} & \cdots & c_{n,1} \\ c_{1,2} & c_{2,2} & \cdots & c_{j,2} & \cdots & c_{n,2} \\ \vdots & \vdots & & \vdots & & \vdots \\ c_{1,m} & c_{2,m} & \cdots & c_{j,m} & \cdots & c_{n,m} \end{bmatrix} \tag{4}$$

Let the base width vector be $\boldsymbol{B} = [b_1, b_2, b_3, b_4, \cdots\cdots, b_n]^T$, where $b_j > 0$ is the base width parameter value of node j. If $\boldsymbol{W} = [\omega_1, \omega_2, \omega_3, \omega_4, \cdots\cdots, \omega_n]^T$ is the network weight vector, the corresponding network output is

$$y_m(k) = \boldsymbol{W}^T \boldsymbol{H} = \omega_1 h_1 + \omega_2 h_2 + \cdots + \omega_n h_n \tag{5}$$

The process of determining the structure of RBF neural network is the process of determining the parameters of the three networks: the connection weight ω, the radial basis center t and the base width σ. A large number of experiments show that the size of the base width σ determines the complexity of the network.

C. DETERMINATION OF ADAPTIVE BASE WIDTH BASED ON SUBTRACTIVE CLUSTERING ALGORITHM AND K-MEANS ALGORITHM

Base width σ is the norm of vector $\boldsymbol{x} - \boldsymbol{t}_i$, which usually represents the distance between x and t_i, and determines the radius of the center of the ith base function, so the value of σ should be adjusted adaptively according to the range of network clustering. When the input distribution of the network is sparse, in order to generate the appropriate number of clusters, the value of σ should be larger, while when the distribution of the network is dense, the value of σ should be smaller. The density of the network is D. After calculating the density of K initial clustering centers by formula (6) and formula (7), K - means algorithm is used to classify all input vectors, so that the value of clustering center can be finally determined, and the base width σ should take the maximum value from clustering samples to clustering center.

$$D_i = \sum_{j=1}^{n} \exp\left(-\frac{\|x_i - t_i\|^2}{(r_a/2)^2}\right) \tag{6}$$

where, r_a represents a neighborhood radius of x_i, x_i and x_j are two n dimensional input samples

$$D_i = D_i - D_{ts} \exp\left(-\frac{\|x_i - x_{ts}\|^2}{(r_b/2)^2}\right) \tag{7}$$

where, r_b is represented as a neighborhood with significantly reduced density index function. In order to avoid clustering centers with similar distance, $r_b = 1.5_{ra}$ is generally selected.

D. THE STEPS OF VARIABLE BASIS WIDTH RBF NEURAL NETWORK CLASSIFIER

In fact, the process of building RBF neural network model is to determine the three core parameters of the network, that is, to determine the output weight ω, cluster center t and base width σ of the network [19]–[22]. The core idea

of the algorithm is to constantly adjust the value of σ in the process of determining parameters ω and t.In the whole learning process of the algorithm, the input samples of the input layer are represented by $x_1, x_2, \cdots, x_N$, in which x_i is a n dimensional vector, that is $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})$, the output of the corresponding hidden layer is $y_1, y_2, \cdots, y_N$, and the transfer function of the jth node in the hidden layer is h_j.The specific steps of the algorithm are as follows.,

Step 1 First, the cluster center should be determined. According to the idea of subtractive clustering algorithm, k clustering centers are determined according to formula (6) and formula (7). At this time, the determined cluster centers are $t_1, t_2, \cdots, t_k$, and the corresponding cluster domain of these cluster centers is $c_1, c_2, \cdots, c_k$.

Step 2 Since there is only one sample in each cluster domain, the base width of cluster center is initialized to $\sigma_i = 0$

Step 3 For all the input sample data, calculate the distance $l_{ji} = \|x_j - t_i\|$ between all the sample inputs and the cluster center according to the distance formula (8) between the two points, where $i = 1, 2, \cdots, k; j = 1, 2, \cdots, N$

$$l_{ji} = \sqrt{(x_{j1} - x_{i1})^2 + (x_{j2} - x_{i2})^2 + \cdots + (x_{jn} - x_{in})^2} \quad (8)$$

Step 4 According to the principle of minimum distance, x_j is classified into the nearest class. K-means then, according to the algorithm idea, the cluster center t_i' of class i is updated by calculating the mean value of all samples contained in class i.

$$t_i' = \frac{1}{m_i} \sum_{x \in c_i} x \ (1 \le j \le k) \quad (9)$$

where m_i is the number of samples contained in the i-th cluster domain c_i.

Step 5 After calculating the ith cluster center t_i, the base width σ_i also changes. Then, according to the distance formula (8) between the two points, calculate the distance from all samples of the cluster domain to the cluster center.

Step 6 Select the maximum distance calculated in step 5 as the base width of the cluster center. That is, update the base width σ_i according to formula (10), that is: $\sigma_i = \sigma_i'$

$$\sigma_i = \sigma_i' = \max \|x - t_i\|, x \in c_i \quad (10)$$

Step 7 Whether the conditions $t_i'/t_i < \gamma$ and $\sigma_i' < \sigma_i$ are true or not, γ is a pre-set minimum constant, $\sigma_i' < \sigma_i$ means that the clustering change is very small. If the condition is correct, the clustering will be ended and turn to step 8. Otherwise, turn to step 1.

Step 8 The gradient descent method is used to determine the network weight of the output unit. Firstly, the small random number is used to initialize w_i, the network training error is e, and the weighted error objective function $E(k)$ at k time is defined

$$E(k) = \frac{1}{2} \sum_{i=1}^{k} \lambda^{k-1} \sum_{i=1}^{m} (z_i(t) - y_i(t))^2 \quad (11)$$

In the formula, λ is the weighted forgetting factor, usually the value range of λ is $0 < \lambda < 1$, $z_i(t)$ is the expected output corresponding to the input of the ith hidden layer node network in the output layer, and $y_i(t)$ is the actual output of the corresponding network.

Step 9 If $E(k) < e$, the algorithm stops running, otherwise, modify the weight according to formula (12) and formula (13), and return to step 8 to continue.

$$w_i' = w_i - \eta \frac{\partial E(k)}{\partial w_i} \quad (12)$$

$$\frac{\partial E(k)}{\partial w_i} = \frac{1}{2} \sum_{i=1}^{k} (z_i(t) - y_i(t))^2 \cdot y_i(t) \quad (13)$$

In equation (13), η is the learning step, which is usually a very small constant value.

E. CLASSIFICATION EFFECTIVENESS EVALUATION

In the process of classification algorithm research, the commonly used evaluation criteria is based on external criteria, that is, using the data with clear classification structure as the input sample of the algorithm, comparing the classification results of the algorithm with the classification results of the original data to evaluate the classifier [23]–[25]. In this study, the FMI index is used to evaluate the classification accuracy of RBF neural network [26]–[28]. Suppose the raw data classification is

$$U = \{U_1, U_2, \cdots, U_k\} \quad (14)$$

The result of classification algorithm is

$$U' = \{U_1', U_2', \cdots, U_k'\} \quad (15)$$

a is the number of data pairs that belong to the same category in U and also belong to the same category in U', b is the number of data pairs that belong to the same category in U but do not belong to the same category in U', c is the number of data pairs that do not belong to the same category in U but belong to the same category in U', then the calculation method of FMI indicator f is

$$F = a\sqrt{[1/(a+b)][1/(a+c)]} \quad (16)$$

Its value is between 0 and 1. The larger the value is, the higher the classification accuracy is.

V. EXPERIMENT

A. DATA SET AND PARAMETER SETTING

In order to verify the classification effect of the model, we collected 3231 cases of coronary heart disease, 9628 cases of diabetes, 5628 cases of bronchial tuberculosis data set provided by a city health and Family Planning Bureau and some municipal hospitals as test data. The experiment is carried out in MATLAB 7.13. 600 samples were randomly selected from three data sets to ensure that the number of samples in each data set is equal when classifying different data sets in the experiment. 300 samples were selected as training samples and 200 samples as test samples. Based on a large number

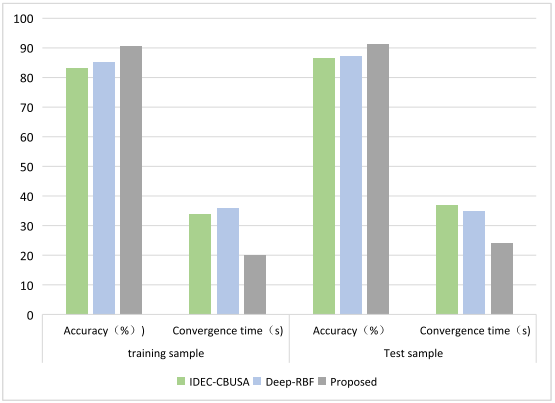


FIGURE 4. Comparison of classification of data sets of patients with coronary heart disease using different algorithms.

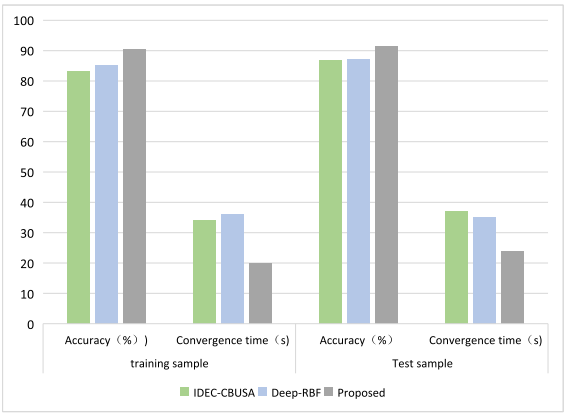


FIGURE 5. Comparison of different algorithms used in data sets of diabetic patients.

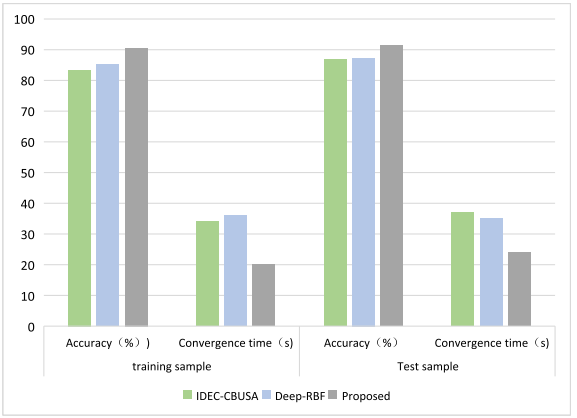


FIGURE 6. Comparison of different algorithms used in data sets of patients with bronchial tuberculosis.

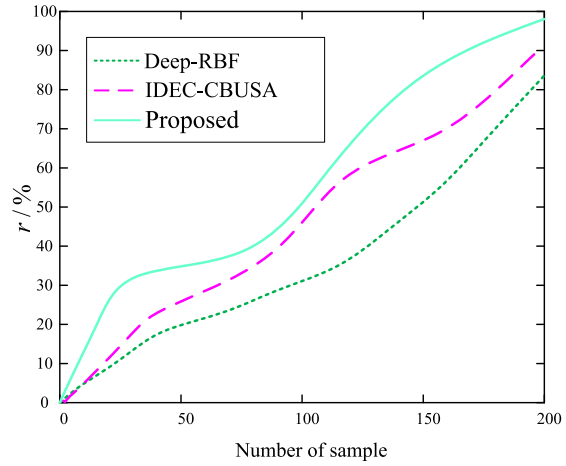


FIGURE 7. Comparison of classification accuracy of three methods for data sets of patients with coronary heart disease.

of previous experimental experience, during the experiment, the relevant parameters are set as follows: $r_a = 0.5$, $\varepsilon = 0.7$, $\eta = 0.002$, $\lambda = 0.9$, $e = 0.1$, $\gamma = 0.8$.

For the model proposed for each data set, the algorithm of Imbalanced data ensemble classification based on the cluster-based under-sampling algorithm (IDEC-CBUSA) in [29] and the Deep-RBF model clustering algorithm which automatically adds hidden layer nodes in the training process in [30].

B. EXPERIMENTAL RESULTS AND ANALYSIS

Compare the classification accuracy and convergence time of training samples and test samples respectively. The experimental results are shown in Figure 4-6.

From the comparison data shown in Figure 4-6, it can be seen that for the same data set, the proposed model is used for classification, not only the classification accuracy is higher than that of IDEC-CBUSA classification method and Deep-RBF classification method, the average classification accuracy is higher than that of other methods by more than 5%, but also the classification accuracy of the three data sets is more than 85%.

From the comparison of the convergence time in the table, it can be seen that the convergence speed of the proposed model is about 40% higher than that of the other two methods. This is because the adaptive variable base width neural network model can adjust the value of the base width adaptively according to the changes of the input samples of the data set at any time, and can flexibly adjust the structural parameters of the neural network model, so that the method can obtain higher classification accuracy and faster classification speed for this kind of large data samples.

Figure 7 shows the comparison between the number of training samples and the accuracy of data classification in the data set of coronary heart disease patients using the proposed model, IDEC-CBUSA method and Deep-RBF method. From this figure, it can be seen that IDEC-CBUSA has higher classification accuracy than Deep-RBF method, because IDEC-CBUSA algorithm uses the idea of combining two methods to classify unbalanced data based on the original under sampling method and integrated learning method based on clustering. In the data processing stage, the cluster based under sampling method is used to form the balanced data set, and then the AdaBoost integration algorithm is used to train the new data set. In the algorithm integration process, the weight weight is used to distinguish the contribution of

the minority data and the majority data to the calculation of the learning error rate of the integration, so that the algorithm pays more attention to the minority data classes. However, Deep-RBF only has enough data, the region with enough feature points is endowed with high confidence, and the structural parameters of neural network model cannot be adjusted flexibly. In this paper, the classification accuracy of the proposed method, which can greatly reduce the requirements of sample parameters and improve the accuracy of neural network model in classification, for data sets of patients with coronary heart disease is generally higher than that of the other two methods.

VI. CONCLUSION

Health big data refers to all data closely related to human health and public health, including data generated from birth, infant health care, vaccine injection, school physical examination, work physical examination, medical treatment, hospitalization, exercise, sleep, death and other life cycles. Health big data refers to all data closely related to human health and public health, including data generated from birth, infant health care, vaccine injection, school physical examination, work physical examination, medical treatment, hospitalization, exercise, sleep, death and other life cycles. The test data mainly refers to the data generated by routine blood examination, cell examination and pathological examination of various organs. Medical record data refers to the data related to personal medical diagnosis information records and health status, which are usually managed in electronic form. Medical image data refers to the internal organization structure and other images obtained from the human body, through which the health status of human body can be more detailed understood.

In this study, the manifold analysis algorithm is used to process the data set, and then the similarity matrix is adjusted by exponential function to conduct AP clustering again. On this basis, the RBF neural network classifier is constructed, and an algorithm for constructing the variable basis width neural network model is proposed. The adaptive variable basis width neural network model can adjust the basis width adaptively at any time according to the changes of the input samples of the data set It can adjust the structural parameters of the neural network model flexibly, thus improving the classification accuracy and shortening the convergence time.

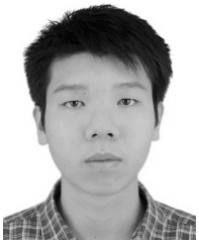
Although this paper has made some achievements in the research of health big data classification model, but due to the limitation of time and the limited level of the author, there are still many shortcomings in this paper. The technology and algorithm of health big data classification model in this paper have great optimization and improvement space. In this paper, the number of different categories of health data samples is basically the same, but in real life, most of the health big data are non-equilibrium data, so follow-up research can be conducted on the non-equilibrium health big data classification model.

REFERENCES

[1] M. M. Hassan, K. Lin, X. Yue, and J. Wan, "A multimedia healthcare data sharing approach through cloud-based body area network," *Future Gener. Comput. Syst.*, vol. 66, pp. 48–58, May 2017.
[2] C. Esposito, A. De Santis, G. Tortora, H. Chang, and K.-K. R. Choo, "Blockchain: A panacea for healthcare cloud-based data security and privacy?" *IEEE Cloud Comput.*, vol. 5, no. 1, pp. 31–37, Jan./Feb. 2018.
[3] J. M. Franklin, W. Eddings, P. C. Austin, E. A. Stuart, and S. Schneeweiss, "Comparing the performance of propensity score methods in healthcare database studies with rare outcomes," *Statist. Med.*, vol. 36, no. 12, pp. 1946–1963, 2017.
[4] M. S. Nawaz, M. Bilal, M. I. Lali, R. Ul Mustafa, W. Aslam, and S. Jajja, "Effectiveness of social media data in healthcare communication," *J. Med. Imag. Health Informat.*, vol. 7, no. 6, pp. 1365–1371, 2017.
[5] V. Thanigaivasan, S. J. Narayanan, and N. C. S. N. Iyengar, "Analysis of parallel SVM based classification technique on healthcare using big data management in cloud storage," *Recent Patents Comput. Sci.*, vol. 11, no. 3, pp. 169–178, 2018.
[6] K. F. Huybrechts, B. T. Bateman, and S. Hernández-Díaz, "Use of real-world evidence from healthcare utilization data to evaluate drug safety during pregnancy," *Pharmacoepidemiology Drug Saf.*, vol. 28, no. 7, pp. 906–922, 2019.
[7] S. Yang, W. D. Leslie, S. N. Morin, and L. M. Lix, "Administrative healthcare data applied to fracture risk assessment," *Osteoporosis Int.*, vol. 30, no. 2, pp. 565–571, 2018.
[8] Y. Sakata, T. Matsuoka, S. Ohashi, T. Koga, T. Toyoda, and M. Ishii, "Use of a healthcare claims database for post-marketing safety assessments of Eribulin in Japan: A comparative assessment with a prospective post-marketing surveillance study," *Drugs Real World Outcomes*, vol. 6, no. 1, pp. 27–35, 2019.
[9] K. Donegan, R. Owen, H. Bird, B. Burch, A. Smith, and P. Tregunno, "Exploring the potential routine use of electronic healthcare record data to strengthen early signal assessment in UK medicines regulation: Proof-of-concept study," *Drug Saf.*, vol. 41, no. 5, pp. 1–12, 2018.
[10] S. Palaniappan and R. Awang "Intelligent heart disease prediction system using data mining techniques," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl.*, Mar./Apr. 2008, pp. 108–115.
[11] A. Onan, "On the performance of ensemble learning for automated diagnosis of breast cancer," in *Proc. Artif. Intell. Perspect. Appl.* Springer, 2015, pp. 119–129.
[12] A. Sopharak, B. Uyyanonvara, and S. Barman, "Simple hybrid method for fine microaneurysm detection from non-dilated diabetic retinopathy retinal images," *Comput. Med. Imag. Graph.*, vol. 37, nos. 5–6, pp. 394–402, 2013.
[13] X. F. Quan, "Computer-aided diagnosis of breast cancer based on random forest," *Comput. Eng. Softw.*, vol. 38, no. 3, pp. 57–59, 2017.
[14] A. X. Chen and Z. F. Chen, "ADST: Using machine learning method to distinguish sarcoidosis and tuberculosis," *Comput. Sci.*, vol. 41, no. s1, pp. 103–109, 2014.
[15] M. H. Zhang, "Classification and identification of magnetic resonance image of neuromuscular disease DMD with wavelet transform and artificial neural network," *Opt. Techn.*, vol. 42, no. 4, pp. 342–346, 2016.
[16] P. S. Bharti, "Process modelling of electric discharge machining by back propagation and radial basis function neural network," *J. Inf. Optim. Sci.*, vol. 40, no. 2, pp. 263–278, 2019.
[17] Y. Yueneng and Y. Ye, "Backstepping sliding mode control for uncertain strict-feedback nonlinear systems using neural-network-based adaptive gain scheduling," *J. Syst. Eng. Electron.*, vol. 29, no. 3, pp. 580–586, Jun. 2018.
[18] Q. Zhang and F. Sepulveda, "RBFNN-based modelling and analysis for the signal reconstruction of peripheral nerve tissue," in *Proc. 8th ACM Int. Conf. Bioinf., Comput. Biol., Health Inform.*, 2017, pp. 474–479.
[19] R. M. Omari and M. Mohammadian, "Rule based fuzzy cognitive maps and natural language processing in machine ethics," *J. Inf. Commun. Ethics Soc.*, vol. 14, no. 3, pp. 231–253, 2016.
[20] H.-F. Shao, X.-Y. Zhao, W.-X. Huang, D.-L. Li, L. Fan, Z.-J. Xiao, and Z.-C. Xu, "Prediction model of flue-cured tobacco sensory quality based on clustering and generalized radial basis function neural network," *J. Comput. Theor. Nanosci.*, vol. 13, no. 9, pp. 6081–6087, 2016.
[21] Z.-Y. Yang, G.-P. Wu, W. Wang, L. Guo, S.-D. Yang, Q. Cao, Y.-J. Zhang, and P. Hu, "Energy saving control method of downslope speed for high-voltage transmission line inspection robot," *J. Jilin Univ. (Eng. Technol. Ed.)*, vol. 47, no. 2, pp. 567–576, 2017.

[22] Y. Pan and H. Yu, "Biomimetic hybrid feedback feedforward neural-network learning control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 6, pp. 1481–1487, Jun. 2017.
[23] L. Kupková, L. Cervená, R. Suchá, L. Jakešová, B. Zagajewski, S. Brezina, and J. Albrechtová, "Classification of tundra vegetation in the Krkonoše Mts. National Park using APEX, AISA dual and Sentinel-2A data," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 29–46, 2017.
[24] W. Mingchang, Z. Xinyue, Z. Xuqing, W. Fengyan, N. Xuefeng, and W. Hong, "GF-2 image classification based on extreme learning machine," *J. Jilin Univ.*, vol. 48, no. 2, pp. 373–378, 2018.
[25] L. Yuan, W. Wu, C. Tian, W. Song, X. Cao, and L. Liu, "Intelligent identification of ocean parameters based on RBF neural networks," *Int. J. Performability Eng.*, vol. 14, no. 2, pp. 269–279, 2018.
[26] F. Pan, L. Laslett, L. Blizzard, F. Cicuttini, T. Winzenberg, C. Ding, and G. Jones, "Associations between fat mass and multisite pain: A five-year longitudinal study," *Arthritis Care Res.*, vol. 69, no. 4, pp. 509–516, 2017.
[27] M. H. Esfe, "Designing an artificial neural network using radial basis function (RBF-ANN) to model thermal conductivity of ethylene glycol–water-based TiO_2 nanofluids," *J. Therm. Anal. Calorimetry*, vol. 127, no. 3, pp. 2125–2131, 2017.
[28] Q.-L. Zhang, W.-J. Liu, W. Yang, X.-M. Wang, and D.-M. Zhang, "Experiment study of mining technology improvement based on RBF neural network," *Blasting*, vol. 35, no. 11, pp. 1641–1645, 2017.
[29] S. Wu, L. Liu, and D. Lu, "Imbalanced data ensemble classification based on cluster-based under-sampling algorithm," *J. Univ. Sci. Technol. Beijing*, vol. 2017, no. 8, pp. 1244–1253, 2017.
[30] P. H. Zadeh, R. Hosseini, and S. Sra, "Deep-RBF networks revisited: Robust classification with rejection," Tech. Rep., 2018.

CONGSHI JIANG received the Ph.D. degree in photogrammetry and remote sensing from the Wuhan Technical University of Surveying and Mapping, in 1992. He is currently the Ph.D. Supervisor with Wuhan University. His research interests include software engineering, GIS, smart city location service, and cloud computing.



YIHONG LI is currently pursuing the Ph.D. degree in software engineering with the School of Remote Sensing and Information Engineering, Wuhan University. His works focus specifically on big data, edge intelligence, and machine learning.