IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 23, 2020, accepted February 5, 2020, date of publication February 11, 2020, date of current version February 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973304

Classification and Prediction of Tibetan Medical Syndrome Based on the Improved BP Neural Network

SHENGHAO YANG[1], **XIAOLAN ZHU**[1], **LEI ZHANG**[2], **LU WANG**[1], **AND XIAOYING WANG**[1]
[1]State Key Laboratory of Plateau Ecology and Agriculture, Department of Computer Technology and Application, Qinghai University, Xining 810016, China
[2]College of Computer Science, Information Management Center, Sichuan University, Chengdu 610065, China

Corresponding author: Xiaolan Zhu (zxlanscu@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61563044, Grant 81460768, and Grant 61866031, in part by the National Natural Science Foundation of Qinghai Province under Grant 2017-ZJ-902, in part by the Applied Basic Research Programs of Science and Technology Department of Sichuan Province under Grant 2019YJ0110, and in part by the Youth Foundation of Qinghai University under Grant 2018-QGY-7.

ABSTRACT Tibetan medicine has a long history as a traditional ethnic medicine in China. It is playing an important role in the medical system in northwestern of China, and which has attracted more and more attention due to its unique diagnostic system and clinical efficacy. Meanwhile, as the data mining technology has been widely used in traditional Chinese medicine (TCM), its application in the field of Tibetan medicine has also launched preliminarily. In this paper, we are focusing on Chronic Atrophic Gastritis (CAG) which is a typical gastrointestinal disease in the plateau area, and a novel back-propagation (BP) network model is proposed for Tibetan medical syndrome classification and prediction. K-means clustering algorithm was firstly implemented on the diagnostic data which was obtained from the Qinghai Provincial Tibetan Hospital, and then Correlation-based Feature Selection (CFS) method was adopted for feature selection. The selected feature vectors were finally put into the proposed BP network for training and testing. In order to overcome BP network's typical shortcomings including slow convergence and easy to overfit, we use a method based on Gaussian distribution to improve weights initialization, and dynamically adjusted the learning rate using the learning rate exponential decay method. Further, we add regularization to the loss function to prevent overfitting. Ultimately, the experiment achieved an accuracy of 99.09%, which improved significantly after improvement and achieved better result compared with other classification methods.

INDEX TERMS Data mining, learning rate, neural network, Tibetan medicine.

I. INTRODUCTION

Tibetan medicine is a kind of traditional ethnic medicine in the northwest of China. It has a history of more than 3,800 years [1]. With its unique diagnosis and drug administration system, Tibetan medicine has a broad mass base in northwestern China and has obtained good clinical effect. Therefore, it has made important contributions to the survival and development of the local people and protected people's life and health, and has attracted more and more attention.

In recent years, data mining technology based on building data warehouses and processing analytical data has become a hot research topic in the information industry [2]. With the development of hospital information, the acquisition, storage and processing of medical data has become more convenient. The combination of medical big data and data mining technology has been widely used in the clinical medical diagnosis of TCM. However, because Tibetan medicine as a traditional medicine has many differences with TCM during the clinical diagnosis and treatment of Tibetan medical diseases, some data mining methods which were suitable for TCM cannot be directly applied to the field of Tibetan medicine. Therefore, the relevant research of data mining technology in the field of Tibetan medicine is still in its infancy [3].

As a typical digestive tract disease with typical Tibetan medicine characteristics, CAG is mainly characterized by loss of intrinsic glandular caused by loss of gastric mucosa, often accompanied by intestinal metaplasia and inflammatory reactions [4]. It is a high incidence and repeated lingering disease, and the risk of cancer will be increased when it is associated with intestinal metaplasia [5]. In Western medicine, treatment is mainly to protect the gastric mucosa and enhance gastric motility, but the effect is poor [6]. The study of the disease in Tibetan medicine began in the 1990s [5], but because

The associate editor coordinating the review of this manuscript and approving it for publication was Le Hoang Son.

the pathogenesis of the disease is more complicated, there is still disagreement about the etiology and pathogenesis of the disease. At present, there has no clear standard for the clinical classification of the disease [7]. According to TCM syndrome differentiation, An *et al.* [6] divided the disease into four syndrome types: spleen and stomach deficiency, stomach collateral blood, liver and spleen disorder and stomach yin deficiency. However, Lu and Wang [4] divided the syndrome of this into five types according to the ''Chinese and Western medicine diagnosis and treatment plan for chronic gastritis'' [8] and clinical practice. In 2009, Qinghai Province Tibetan Medicine Association first issued the CAG provincial medical treatment standard, named CAG as ''Pu Ru disease'', classified as Long Bacon (cold) Pu Ru disease and Chi Ba Sheng (heat) Pu Ru disease. However, there is still an objection to the classification of the syndrome type within the scope of Tibetan medicine in the country [5]. The irregular and uncertain classification of syndrome of this disease has restricted further research.

Therefore, this paper firstly used the clustering algorithm to objectively classify the syndrome type of this disease from the perspective of data mining to reduce the judgment of subjective experience during the process of the diagnosis and treatment of disease syndrome. Then, the classification and prediction model using the improved BP neural networks was constructed by analyzing the clinical diagnosis and treatment data, which could achieve the symptom type of the disease with high accuracy by inputting relevant symptoms, and provide scientific decision support for Tibetan medicine diagnosis on the disease.

II. RELATED WORKS

Classification prediction refers to learning a classification function or constructing a classification model based on the existing data, and mapping data records in the database to one of the given categories, so that it can be applied to data prediction [9]. Classification prediction has been widely used in medical fields, but its application in Tibetan medicine is still in its infancy. Shiying *et al.* [10] proposed a KNN algorithm based on gray box method using ovel distance discrimination and realized a Tibetan medicine diagnosis and treatment model for the plateau stomach disease (Atrophic Gastritis) by combining the individual characteristics and typical symptoms of patients with an accuracy of 80.1%. In [11], feature selection algorithm was firstly used to select the most useful attributes and reduce redundant features. Then, based on the gray box method, a variety of classical classification algorithms were used to build prediction models. Wang evaluated multiple classification evaluation indicators and found that the Naive Bayesian algorithm model achieved the most ideal prediction effect, whose classification accuracy rate was 87.2%. Zhu *et al.* [3] proposed a classification model based on atomic classification association rules. Firstly, she used the constraint-based Apriori algorithm to mine the strong atomic classification association rules between symptoms and syndrome. Then, established a classification model after pruning and priority rules and applied the model to the classification of the syndrome of Tibetan medicine. It enabled the construction and classification of model in a short period of time, obtaining fewer but more understandable rules and achieving an accuracy of 92.8%.

Most of the classification models in the Tibetan medicine field mentioned above are based on some traditional machine learning algorithms with relatively low accuracy, no comparative experiments have been performed, and no time performance has been studied except the last model. In addition, these works have not analyzed the obtained syndromes of diseases, and the scientific syndrome is an important issue in the field of Tibetan medical.

Artificial neural network (ANN), abbreviated as neural network (NN), is one of the data mining classification algorithms. It is a mathematical model or computational model that mimics the structure and function of biological neural networks for precise processing of various types of data [12]. In recent years, with the deepening of research work on neural networks, it has been widely used in many fields and has achieved good results. For the medical field, artificial neural networks are widely used in clinical diagnosis, image analysis and interpretation, signal analysis and interpretation, drug development [13]. Yasumoto *et al.* [14] investigated the feasibility of accurate state classification of autonomic neural activity (ANA) based on power spectrum models of heart rate fluctuations (HRF). The artificial neural network (ANN) was used to classify HRF for clinical diagnosis, revealing the characteristic oscillation of the entropy bandwidth and index derived from blood pressure changes, thereby improving the accuracy of classification. Fujita *et al.* [15] used artificial neural networks to assist radiologists in detecting and classifying coronary artery disease in single-photon emission computed tomography bullseye images, and the results showed that the recognition performance of neural network systems was comparable to that of experienced radiologists (two to ten year). Petroni *et al.* [16] used a neural network architecture to classify three types of baby crying. The input data consisted of 10 consecutive frames of mel-cepstrum coefficients, ranging from 0.75 seconds to 1 second in length. The mel-cepstrum coefficient was extracted from anger, fear and crying in pain. As can be seen from the test results, artificial neural network is a useful tool for cry classification. Liu *et al.* [17] proposed a variety of feature learning models for disease analysis and evaluation. Based on convolutional neural network, a medical text feature learning model was proposed for disease risk assessment, and deep learning method was used for medical data feature analysis. In the aspect of time series feature learning, a multichannel convolutional self-encoding neural network was proposed, in which multi-channel convolutional neural network was used to learn data features, and convolutional self-encoding neural network was used to learn facial image data features. The characteristics were combined with the collected physiological data for emotional fatigue testing. Xu *et al.* [18] used the chronic obstructive pulmonary disease (COPD) as an example to construct a TCM syndrome typing model of the disease using ANN. In addition, the data set was divided into four subsets and a sub-model was built using ANN for each subset, which achieved better performance than the full model. Chen *et al.* [19] designed a medical diagnosis system that combined the expert diagnosis and neural network reasoning. It built an expert system in the form of a rule tree, and used multiple BP neural networks to group diagnosis and reasoning. The system made full use of expert prior knowledge and the numerical reasoning and self-learning ability of neural network to make disease diagnosis more accurate and convenient. Zheng [20] constructed a feedforward multi-layer neural network for the diagnosis of cardiac neurosis. It firstly converted the original data attribute into data features through data preprocessing, and then judged the effect of the feature on the sample differentiation according to whether the feature was divergent to select features, and added a gradient test in the training process to prevent the occurrence of erroneous calculation. Zhou *et al.* [21] designed a palm print diagnosis expert system based on neural network, which divided the palms into different regions and numbered the lines with different features to distinguish the complex palm print features. There were 113 neurons in the input layer of the neural network, which correspond to the intensity of different palmprint pathological features. It used an improved algorithm of learning rate factor self-tuning for the problem that the back-propagation algorithm converges slowly.

However, the above work is just the application of neural networks in the field of conventional medicine, the application of neural network in the field of Tibetan medicine is still in its infancy, and there is no substantive result. Due to the limitation of Tibetan medicine research, there are problems such as difficulties in determining the syndrome scientifically and too few data samples. At the same time, problems such as the slow convergence rate of the neural network itself and the tendency to overfit when there are fewer data samples needing to be resolved. Therefore, this paper proposed to use BP neural network to classify and predict the Tibetan medical syndrome type of CAG, and studied the related issues mentioned above.

III. BACK-PROPAGATION NETWORK

Back-propagation network is a multi-layer feedforward network trained by a back-propagation method [22], whose core algorithm is error gradient descent method. The training process consists of two parts: forward propagation and back propagation [23]. In the case of forward propagation, the sample data starts from the input layer, passes through the hidden layer and the output layer in turn, and finally the output layer outputs the result. Then compare the output value with the expected value and calculate the error using the loss function [24], and then proceed the back-propagation phase of the error. As for back propagation, the output error is inversely calculated from the output layer through the hidden layer to the input layer, and the error is assigned to each unit in each layer. Once all the units obtain the error information, the weights of each unit will be corrected according to the error. The BP network is a hierarchical structure, generally consists of an input layer, a hidden layer, and an output layer [23]. The number of hidden layers can be more than one, and each layer is composed of multiple neurons, which are called nodes or units. McCulloch and Pitts [25] first proposed the M-P model of neural network neurons, and the model is shown in Fig 1.

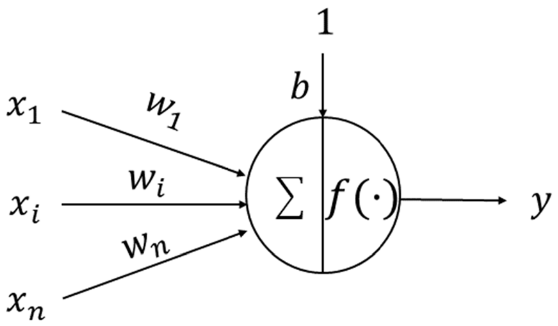


FIGURE 1. M-P model of neural network neurons.

Where, x_i represents the input of the current neuron, w_i represents the weight of the current neuron for the input, b is the bias of the neuron, f is the activation function, y is the output of the neuron. As shown in (1):

$$y = f(\sum\nolimits_{i=1}^{n} x_i \times w_i + b) \tag{1}$$

In this experiment, the activation function $f(\cdot)$ uses the sigmoid function, which has the following formula:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

A neural network structure is connected by multiple neurons, as shown in Fig. 2:

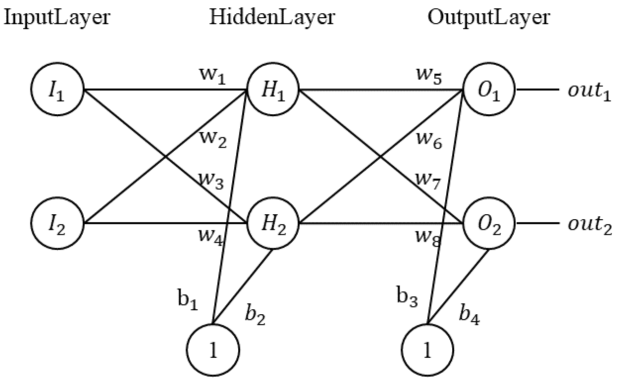


FIGURE 2. Neural network structure.

We use the neural network of Fig. 2 as an example to introduce the back-propagation algorithm briefly. This is a three-layer neural network with two neurons in each layer. The forward propagation process starts from the input layer. For each neuron, use (1) to calculate the output, which would be used as the input to the next layer until the output layer gets the final output.

After the end of a forward propagation, the loss function is used to calculate the error between the actual output and the expected output. We use the cross-entropy error function as an example to calculate the error. It can be described as (3):

$$\mathrm{E}=\frac{1}{m}\sum\nolimits_{i=1}^{m}\left(O_i\times\ln out_i+(1-O_i)\times\ln\left(1-out_i\right)\right)\qquad(3)$$

where, O_i represents the desired output, out_i denotes the actual output, and m represents the number of outputs.

In the process of back propagation, the weight and bias are updated based on the error. The updating process is to determine the influence of the parameter by calculating the partial derivative of the error for the parameter. Then update the value of the parameter using (4):

$$w_{new}=w-\alpha\times\frac{\partial E}{\partial w}\qquad(4)$$

where, w represents the parameter, α represents learning rate, $\frac{\partial E}{\partial w}$ is the partial derivative of the error E for the parameter.

The chain rule is used to calculate the partial derivative. Take the update of w_5 as an example, we use the following formula:

$$\frac{\partial E}{\partial w_5}=\frac{\partial E}{\partial out_1}\times\frac{\partial out_1}{\partial sum_{O1}}\times\frac{\partial sum_{O1}}{w_5}\qquad(5)$$

where, $\frac{\partial E}{\partial out_1}$ is the partial derivative of the error E for final outputout_1, that is the partial derivative of the loss function (as shown in (3)), $\frac{\partial out_1}{\partial sum_{O1}}$ is the partial derivative of the final output out_1 for the input weighted sum sum_{O1} of neurons O_1, that is the partial derivative of the activation function (as shown in (2)), $\frac{\partial sum_{O1}}{\partial w_5}$ is the partial derivative of the input weighted sum sum_{O1} of neurons O_1 for weightw_5, and it is easy to calculate that the value is the input of neuron O_1 from neuron H_1.

After a round of forward propagation and back propagation, once iteration is completed. After several iterations, the error will gradually decrease and we consider the model tend to converge.

IV. CONSTRUCTING THE CLASSIFICATION AND PREDICTION MODEL

In this paper, the flow chart of the construction the classification and prediction model of Tibetan medical syndrome with the improved BP neural network the is shown in Fig.3.

A. PREPROCESSING

CAG is a digestive tract disease with typical Tibetan medicine characteristics. At present, scholars are still researching the classification of CAG in Tibetan medicine, and have not yet formed a standardized Tibetan medical classification system, which lacks corresponding objective evidence. In the original dataset of Tibetan medical diagnosis and treatment of this paper, the decision-making attribute "syndrome" is lacked. Therefore, we need to use the clustering algorithm to classify the types of syndrome objectively. At the same time, some symptom attributes are incomplete, there are noises and inconsistencies. Before starting to construct a neural network, the original dataset is processed in the following ways.

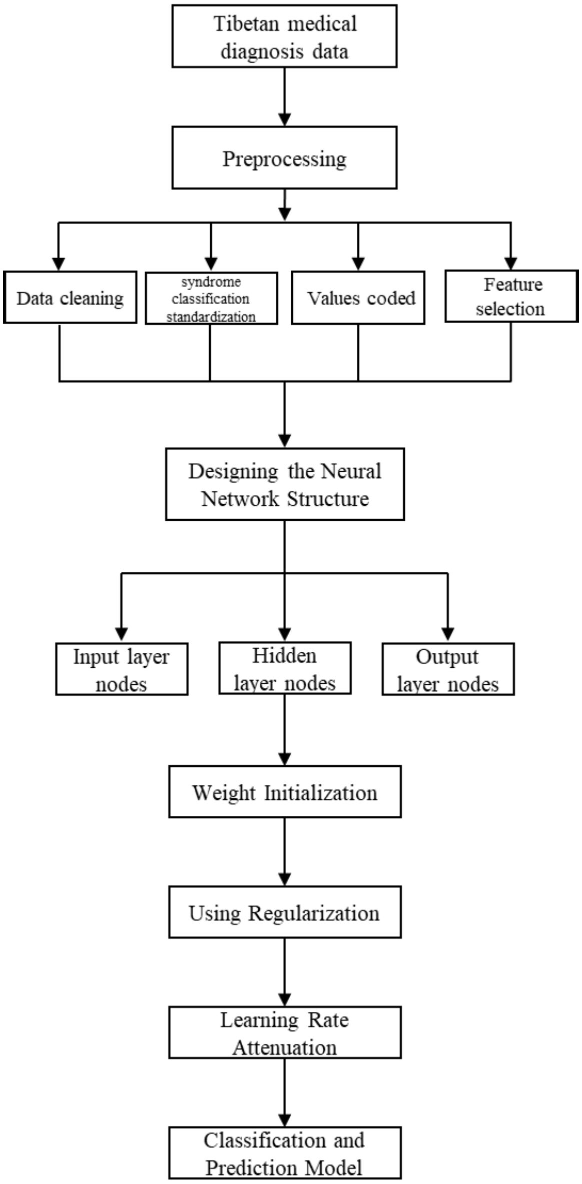

Tibetan medical diagnosis data
Preprocessing
Data cleaning
syndrome classification standardization
Values coded
Feature selection
Designing the Neural Network Structure
Input layer nodes
Hidden layer nodes
Output layer nodes
Weight Initialization
Using Regularization
Learning Rate Attenuation
Classification and Prediction Model

FIGURE 3. Constructing the Classification and Prediction Model.

Firstly, fill in the incomplete attributes and smooth the noise data. For the decision-making attribute "syndrome" in the dataset, the K-means clustering algorithm is used to standardize the classification of syndrome, and then the standardized dataset is obtained. Secondly, the attribute values are coded into discretized numbers by using number from 1 to n according to the attribute content. Finally, in order to improve the accuracy of classification, the feature selection method is used to reduce the redundant attributes.

The original dataset includes 81 diagnostic items, but too many features usually make the structure of the neural network too complicated. In order to reduce the redundant features and extract the feature subsets that are most beneficial to the model effect, we use the feature selection method.

Feature selection, also known as attribute selection, refers to the selection of some effective features from a set of features to reduce the dimension of the feature space [26]. When constructing a classification prediction model, too many features will increase the computational overhead and affect the performance of the model. Especially in the case of limited samples, it is more necessary to make correct and effective feature selection [27].

The feature selection method used in this paper is CFS [28], which is based on the importance of a feature to the classification system and its correlation with other features, and the feature selection and dimension reduction are achieved by removing redundant features those are not related to the classification system and highly correlated with other features. The evaluation formula for this method is defined as follows:

$$F_S = \frac{k\overline{r_{zi}}}{\sqrt{k + k\,(k-1)\,\overline{r_{ii}}}} \tag{6}$$

where, S is a feature set containing k features. $\overline{r_{zi}}$ represents the average of the correlation between feature i and classification system z. $\overline{r_{ii}}$ is the average of the correlation between the features. The numerator part represents the prediction ability of the feature set S to the classification system z, and the denominator part represents the redundancy between the attributes in S.F_S represents the classification ability of the classification system z after removing the redundant features from S.

B. DESIGNING THE NEURAL NETWORK STRUCTURE

The structure of the input layer and the output layer of the neural network is relatively easy to be determined. Generally, the number of input layer nodes is the number of sample features, and the number of output layer nodes is the number of categories of classification [29]. We encode and standardize the values of each input feature as the input value of the neural network, and the output values are encoded according to the type of syndrome that obtained with K-means algorithm, that is, (1, 0) and (0, 1) correspond to syndrome 1 and syndrome 2, respectively.

For the structure of the hidden layer, in general, a 3-layer network with only one hidden layer is enough to be able to approximate all nonlinear mappings. From the simple and practical point of view, it is enough to choose a hidden layer. On the other hand, there is no unified conclusion for the determination of the number of hidden layer nodes. Too much and too little may affect the performance of the network. If the number of nodes is large, the learning time will be too long but the error will not be the smallest. If the number of nodes is small, the network fault tolerance is poor, and more local optimum is generated.

Zhang and Li [29] presented an empirical formula for calculating the number of hidden layer nodes, as shown below:

$$N_H = \frac{N_I + (N_O, N_C)_{max}}{2} \tag{7}$$

where, N_H, N_I, N_O are the number of nodes in hidden layer, input layer and output layer, respectively. N_C is the number of classification categories.

Cai [30] introduced an empirical formula based on the least squares method, as shown below:

$$N_H = \left| \sqrt{0.43N_IN_O + 2.54N_I + 0.77N_O + 0.35 + 0.12N_O^2} + 0.51 \right| \tag{8}$$

Since the number of hidden layer nodes is an integer, the calculation result is rounded.

It is pointed out in [31] that using $(2\times N_I + 1)$ hidden layer nodes can achieve a good compromise between network capacity and training time.

In this paper, we use the above three methods as a reference to find the number of hidden layer nodes that can get the best results. The details will be introduced in Section V.B.

C. WEIGHT INITIALIZATION

Jin *et al.* [32] pointed out that the weight initialization is one of the important factors affecting the network training process, and the weight is generally initialized to a small random number uniformly distributed around 0. In this paper, we can use a Gaussian distribution based on zero mean and standard deviation as the initial value of the weight. In addition, according to the idea of the literature [33], we divide each weight by the square root of the number of input data based on the above method. It avoids the problem that the distribution variance of the output data of randomly initialized neurons increases as the amount of input data increases.

D. USING REGULARIZATION TO PREVENT OVERFITTING

The over-fitting phenomenon refers to the fact that the classification accuracy of the model in the training sample is increasing, but the classification accuracy in the test sample may be reduced, which is more serious when the sample is smaller [34]. Regularization is commonly used to prevent overfitting, which adds a regular penalty term to the loss function, which is a weight-related function. There are two commonly used regular items, L1 regularization and L2 regularization. The role of L1 regularization is to generate a sparse weight matrix, that is, there are many zeros in the weight matrix, so that some features will have no contribution in the model because the coefficient is 0, so it can be used for feature selection. Since we have used CFS for feature selection in this experiment, we use L2 regularization, the formula is $\frac{\lambda}{2n}\sum_{i=1}^{N} w_i^2$. Then, we use E_{new} to represent the loss function after adding the regular term, which has the following formula:

$$E_{new} = E + \frac{\lambda}{2n}\sum_{i=1}^{N} w_i^2 \tag{9}$$

Correspondingly, the parameter gradient formula will also change after adding the regular term, which is shown in (10):

$$\frac{\partial E_{new}}{\partial w} = \frac{\partial E}{\partial w} + \frac{\lambda}{n}w \tag{10}$$

Furthermore, the parameter update formula also changes, which is shown in (11):

$$w_{new} = \left(1 - \frac{\alpha\lambda}{n}\right) w - \alpha \frac{\partial E}{\partial w} \tag{11}$$

It can be noticed that compared with the (4), the parameter w is preceded by a coefficient smaller than 1, so that the value of w is continuously reduced during the iterative process. This shows that the L2 regularization will punish the larger parameters, and finally the model will get smaller parameters. In general, models with small parameter values are more adaptable to different data sets, and over-fitting can be avoided to some extent.

E. LEARNING RATE ATTENUATION

Learning rate is an important hyperparameter in the neural network. We can see from (4) that it directly affects the update of the weight, so it will directly affect the training process of the model. For the selection of the learning rate value, if the value is large, the model learning speed will be faster, but if it is too large, the error value will stop falling and the oscillation will occur. If the value is small, the model converges slowly and may fall into local optimum. In this paper, we use the method that the learning rate will be automatically attenuated during the training process, so that the error value will decrease faster at the beginning of training because of the higher learning rate. In the later stage of training, using a smaller learning rate helps the model to gradually close to the optimal solution. The commonly used learning rate attenuation method is exponential decay, and the formula is as follows:

$$learn_n = learn_0 \times decay^{n/step} \tag{12}$$

where, $learn_0$ represents the initial learning rate, n represents the number of iterations, *decay* represents the attenuation coefficient, and *step* represents the attenuation speed, that is, the initial learning rate will decrease a *decay* times after each *step* iterations.

On the other hand, for the selection of the initial learning rate, the trial and error method is generally used, but this method is time consuming. Smith [35] proposed a method to find the optimal initial learning rate. At beginning, a very small learning rate is selected, which is continuously increased in the iteration. At the same time, the change of the error value is recorded, and the value of the learning rate is observed when the error value reaches the lowest point, thereby obtaining a relatively reasonable initial learning rate.

V. EXPERIMENT

A. DATA SET

1) SYNDROME CLASSIFICATION STANDARDIZATION

The experimental data in this paper was derived from the clinical diagnosis and treatment data of 223 patients with CAG provided by National Natural Science Foundation project cooperation unit - Qinghai Provincial Tibetan Hospital. The Fig.4 below depicts some information about the patient samples in this dataset.

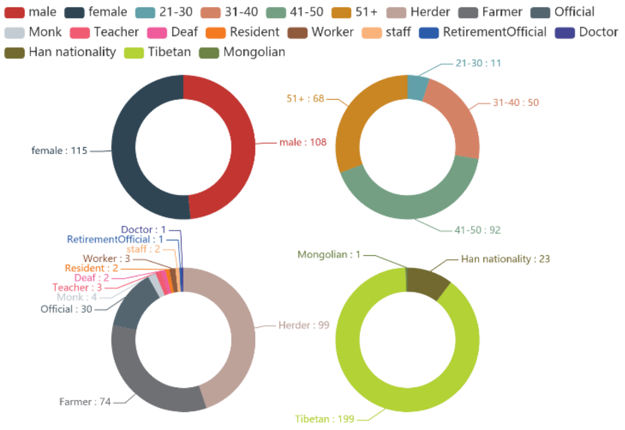


FIGURE 4. From top to bottom, from left to right are gender, age, occupational and ethnic distribution of patient samples, respectively.

Because the dataset is lack of the decision-making attribute Tibetan medicine "syndrome", we first used the K-means clustering algorithm to normalize the classification of syndrome based on the elbow rule and the average contour coefficient as the evaluation indicators of the clustering results.

a: SYNDROME CLASSIFICATION BASED ON ELBOW RULE

The measure of the elbow rule is the sum of squared errors (SSE), and the point at which the SSE value drops the fastest is usually the best type of syndrome. In the elbow rule, as the K value increases, the SSE value decreases. When the optimal K value is reached, as the K value increases, the SSE value decreases significantly, and the SSE value changes the fastest. Then, as the K value increases, the SSE value drops slowly. At this time, there is an "elbow" corresponding to the K value of the best syndrome type. When the K value varies from 2 to 10, the corresponding SSE value is as shown in Fig.5:

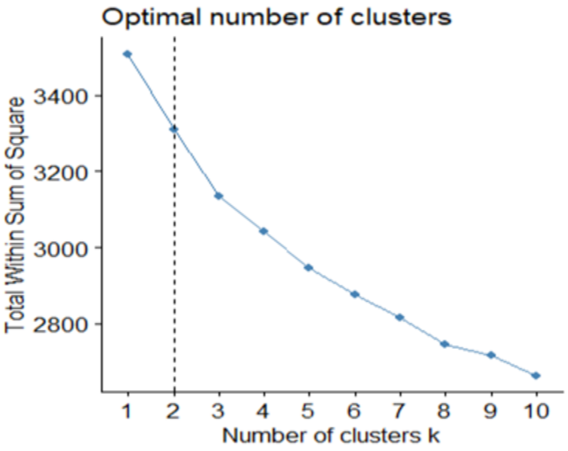


FIGURE 5. Syndrome classification based on elbow rule.

It can be seen from Fig.5 that when K = 2, the SSE drops the fastest, and then as the K value increases, the SSE changes more and more slowly. Therefore, the optimal number of clusters for Tibetan medical syndrome is 2.

b: SYNDROME CLASSIFICATION BASED ON AVERAGE CONTOUR COEFFICIENT

The average contour coefficient is used to evaluate the quality of the clustering result. If the clustering result is good, its average contour coefficient value will be higher. According to the given clustering algorithm, the average contour coefficient can directly obtain the optimal number of syndrome types, and the K value corresponding to the highest contour coefficient value is the optimal K value. When the K value varies from 2 to 10, the corresponding average contour coefficient is shown in Fig. 6:

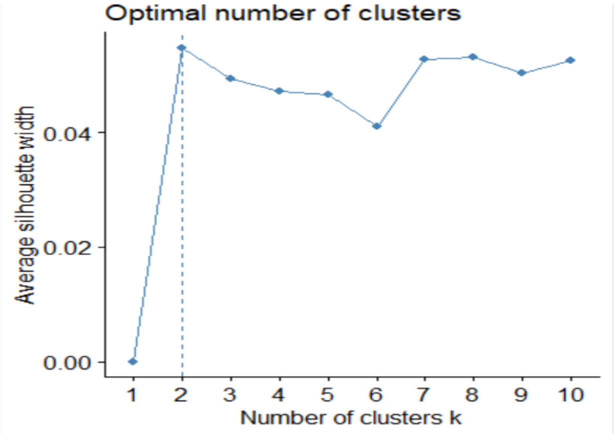


FIGURE 6. Syndrome classification based on average contour coefficient.

It can be seen from Fig.6 that when the K value is 2, the average contour coefficient is the largest. Therefore, the optimal cluster number of the syndrome type is 2, which is consistent with the result obtained by the elbow rule, and the final syndrome type is determined to be 2 clusters. Therefore, we divided the syndromes of dataset in this paper into two types: syndrome 1 and syndrome 2. Among the 219 samples, there were 77 samples of syndrome 1 and 142 samples of syndrome 2. The following Table 1 shows several attributes that are significantly different in the number ratio in tow syndromes:

TABLE 1. Comparison of number ratio of several attributes.

Attribute	Number ratio in syndrome 1	Number ratio in syndrome 2
Hp Helicobacter pylori level = 1	59%	38%
Atrophy gland reduction level = 1	25%	63%
Atrophy gland reduction level = 2	49%	29%
Atrophy gland reduction level = 3	24%	4%
Intestinal metaplasia level = 0	0%	56%
Intestinal metaplasia level = 1	11%	42%
Intestinal metaplasia level = 2	57%	1%
Intestinal metaplasia level = 3	31%	0%

As can be seen from Table 1, the two syndromes are mainly different in the two attributes of atrophy gland reduction level and intestinal metaplasia. As stated in Section I, CAG is mainly characterized by loss of intrinsic glandular and often accompanied by intestinal metaplasia [4]. It shows that these two attributes are closely related to the CAG's syndrome, which further explained that our experimental results coincide with the theoretical arguments.

In our dataset, 42 samples have been labeled by a professional physician. There were 4 kinds of syndromes in the marking result, which were recorded as syndrome 1, 2, 6 and 7 respectively. We compared our clustering results with the labeled result, and found that the samples with syndrome 1 of the clustering result were highly consistent with the samples labeled with syndrome 1 and 2, and the samples with syndrome 2 of the clustering result were highly consistent with the samples labeled with syndrome 6 and 7. The comparison result is shown in the following table:

TABLE 2. Comparison of labeled results with clustering results.

Classification of labeled result	Number of samples	Whose clustering result is 1	Whose clustering result is 2	Compliance
1	8	5	3	64.3%
2	6	4	2	
6	16	6	10	78.6%
7	12	0	12	

Therefore, we believed that our clustering results had certain rationality in medicine. After using this method to complete the "syndrome" attribute of the dataset, the rationality of the dataset was guaranteed to a certain extent.

2) DATASET DESCRIPTION

Through data preprocessing of the original dataset of Tibetan medicine diagnosis and treatment, the original dataset was finally converted into a standard dataset consisting of 219 complete medical records, with a total of 81 symptom attributes. Among them, the symptoms associated with CAG in Tibetan medicine included pulse diagnosis, urinary examination, tongue diagnosis, symptoms and signs, among which, there were 18 kinds of pulse diagnosis, 16 kinds of urinary examination, 7 kinds of tongue diagnosis, and 20 kinds of symptoms and signs. The specific symptom types and frequency of occurrence are shown in Fig. 7,8,9 and 10 respectively:

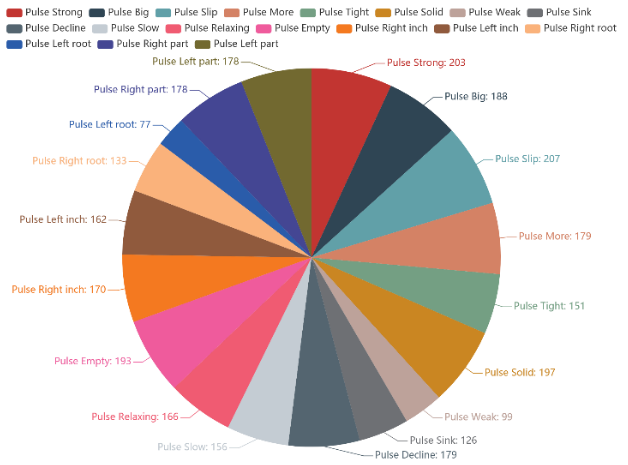


FIGURE 7. The specific symptom types and frequency of occurrence of pulse diagnosis.

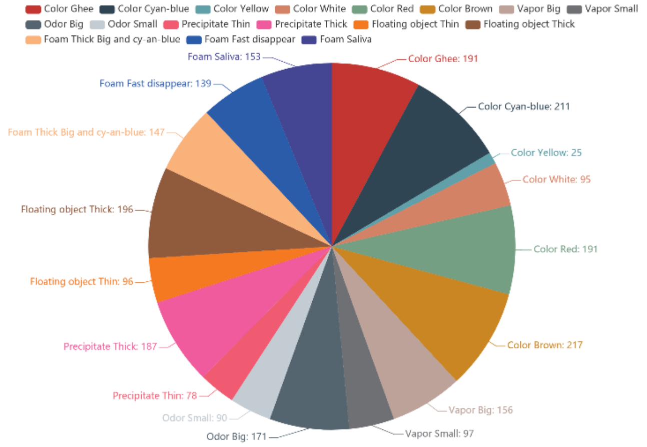


FIGURE 8. The specific symptom types and frequency of occurrence of urinary examination.

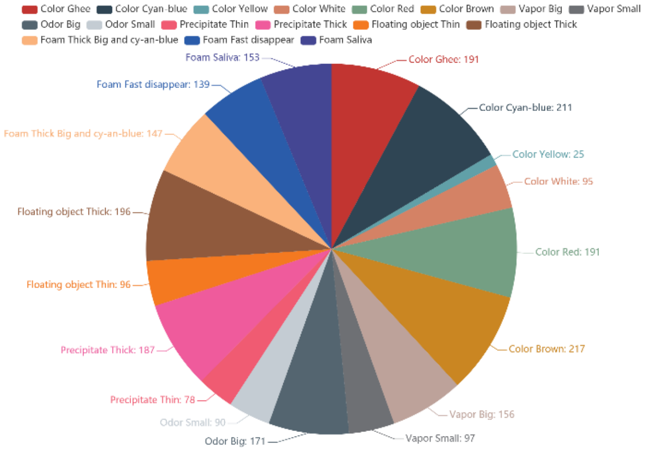


FIGURE 9. The specific symptom types and frequency of occurrence of tongue diagnosis.

Furthermore, we used the CFS method to select the feature set of the data set, and finally got a feature subset with 9 features, and the symptom types and frequency of occurrence in feature subset are shown in Fig. 11.

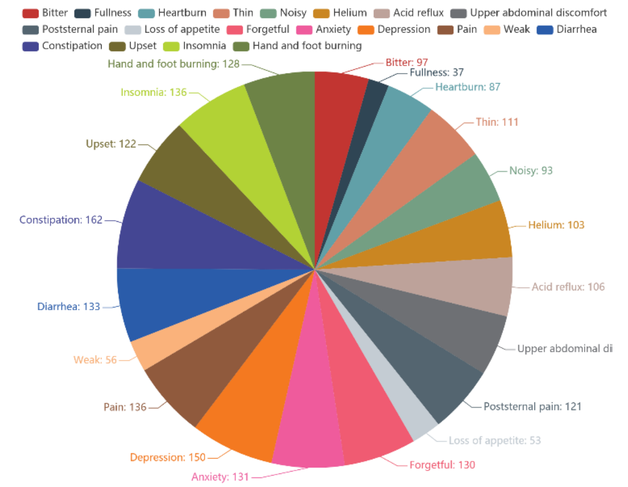


FIGURE 10. The specific symptom types and frequency of occurrence of symptoms and signs.

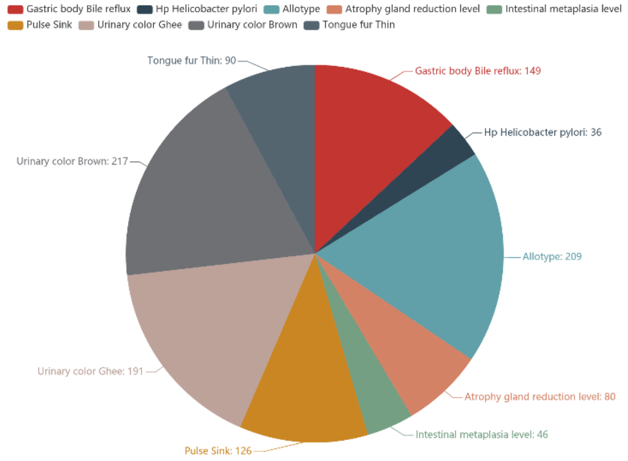


FIGURE 11. The specific symptom types and frequency of occurrence of feature subset.

B. EXPERIMENT AND ANALYSIS

According to the preprocessed dataset, the number of input layer nodes was set to 9, the number of output layer nodes was 2, and the numbers of hidden layer nodes were set to 19, 5, and 6 by the methods in Section IV.B. Furthermore, we used the 10-fold cross-validation method to divide the dataset, that is, the whole dataset was divided into 10 parts, and took one of them as the test set in turn, and the other 9 parts as the training set. After 10 rounds of experiments, the average of the evaluation indicators will be calculated.

1) INITIAL WEIGHT

We used the Gaussian distribution based method to initialize the weights of the network to make them small random numbers uniformly distributed around 0. The specific method was to obtain a matrix of m*n using the numpy.random.randn(m,n) function of the numpy library in python, and the elements were subject to a Gaussian distribution with a mathematical expectation of 0 and a standard deviation of 1. Then for each weight divided by the square root of the input data amount, the entire code was np.random.randon(m, n) / np.sqrt(n).

2) LEARNING RATE SETTING

Use the method in Section IV.E to find a suitable initial learning rate. At first, set the learning rate to 0.00001 and increase it exponentially to 3, observing the change of the error value. The situation of the error value changes with the learning rate as shown in the Fig.12 below:

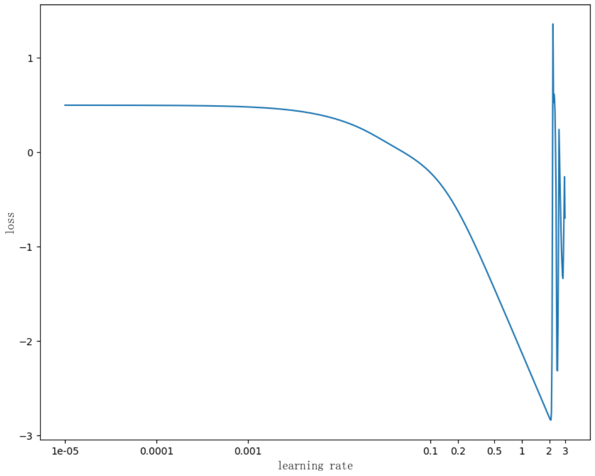


FIGURE 12. The situation of the error value changes with the learning rate (Iteration number: 1500).

From Fig.12, it can be seen that the learning rate is 2 when the error value reaches the lowest point, so the initial learning rate is set to 2 reasonably. In the training process, the exponential decay method was used to adjust the learning rate. The parameters of the exponential decay, attenuation coefficient and attenuation speed were 0.9, 50 respectively.

3) HIDDEN LAYER NODES SETTING

According to the number of input and output layer nodes, the numbers of hidden layer nodes were calculated as 5, 6, 19 respectively using the empirical formula in Section IV.B. The evaluation indicators of the classification results of the model in the case of different number of hidden layer nodes are shown in the following Fig.13:

From Fig.13, when the number of hidden layer nodes is 5 or 6, the accuracy, TP rate and precision of the classification results of the model are higher, which can reach 99.09%, 98.07% and 98.07% respectively. The setting of hidden layer nodes is somewhat contingent, and the method of finding suitable values based on empirical formulas can save a lot of time.

4) REGULARIZATION

According to the definition of over-fitting, as the number of iterations increases, the error gradually decreases, but the accuracy does not increase all the time, which proves that the model leads over-fitting.

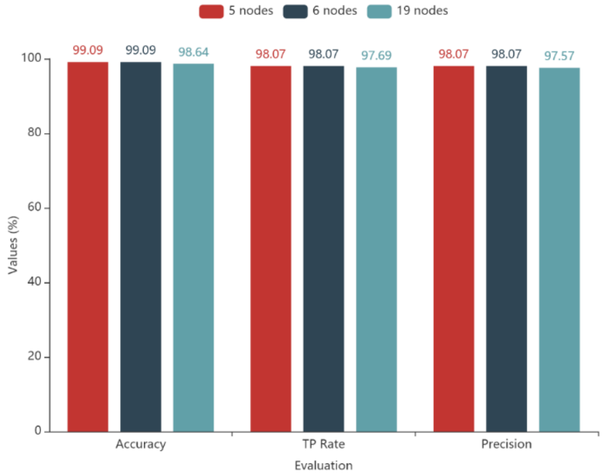


FIGURE 13. Evaluation indicators under different hidden layer nodes (Iteration number: 1500, Initial learning rate: 2).

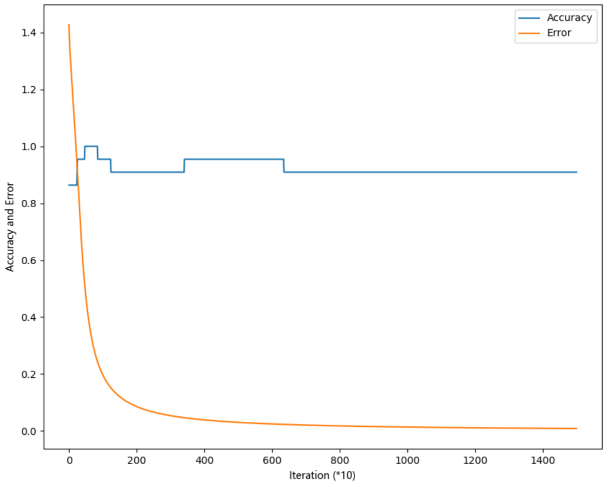


FIGURE 14. The accuracy decreased as the error decreased.

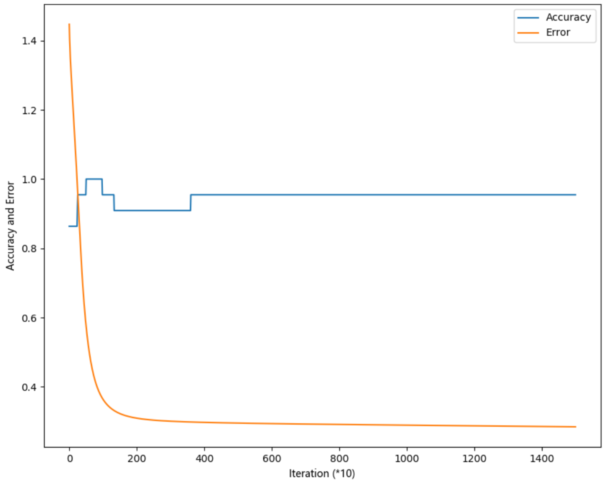


FIGURE 15. The accuracy did not decrease after adding regularization.

For example, in a round of testing shown in the Fig.14 below, the accuracy did not increase as the error decreased, which indicates that an overfitting occurred.

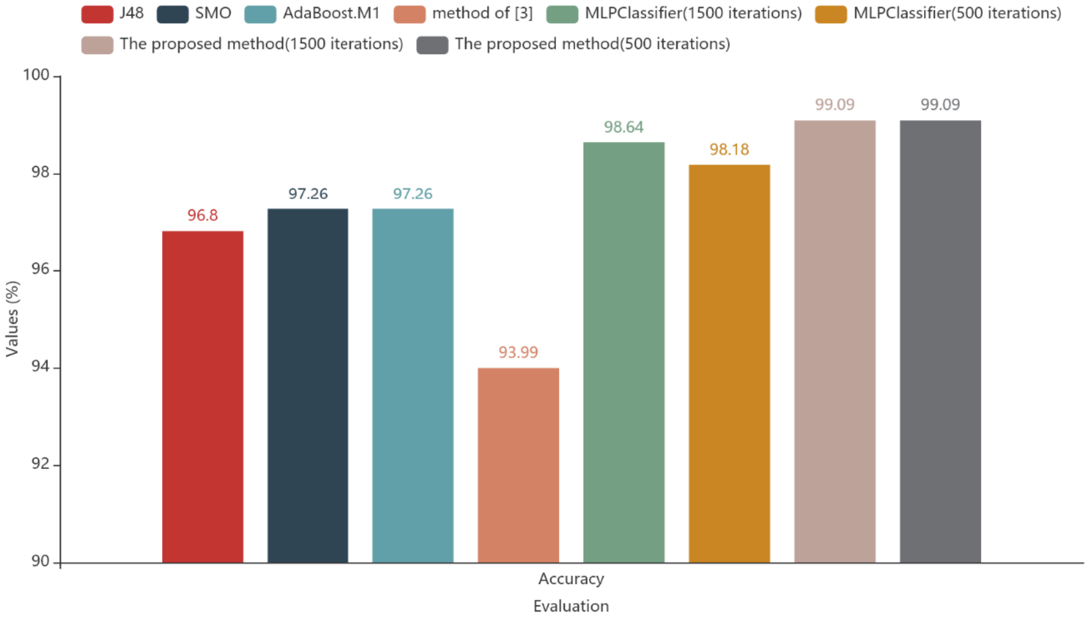


FIGURE 16. Comparison of the accuracy with other methods.

Therefore, we used the method in Section IV.D to prevent over-fitting by adding regularization. Then we found that for the same round of testing, the accuracy did not decrease again, as shown in Fig.15 below:

5) COMPARISON WITH OTHER METHODS

We compare our method with some other traditional machine learning algorithms. The algorithms used are J48 [36], SMO [37], Ada-BoostM1 [38], MLPClassifier [39]. Among them, the J48 algorithm is a decision tree model, which uses the information gain rate as a metric for selecting the current optimal decision attributes when building a decision tree. SMO is an optimization algorithm of SVM. Its basic idea is that if the solutions of all variables meet the necessary conditions for the optimal solution of this optimization problem, then the solution of this optimization problem is obtained. AdaBoostM1 is an integrated learning method that will use multiple classifiers. It will adjust the sample weight distribution according to the classification error rate of the current classifier to ensure that the weight of the incorrectly classified samples does not change, and the weight of correctly classified samples decreases, and adjust the weight of the current classifier in the final decision. MLPClassifier is a classifier of the same type as the BP neural network used in this article. We added it into comparison to show our optimization effect.

Based on the same dataset and experimental environment, we compared our experimental results with above traditional algorithms and the method of [3] from the aspects of the accuracy and the time consumption to build classification model and predict new instances in the dataset under 10-folds cross validation. The experimental results are as shown below:

TABLE 3. Comparison of the time consumption to build classification model and predict new instances with other methods.

Methods	Time(s)
J48	0.023
SMO	0.03
AdaBoostM1	0.639
Method of [3]	0.347
MLPClassifier(1500 iterations)	5.596
MLPClassifier(500 iterations)	4.78
The proposed method(1500 iterations)	3.197
The proposed method (500 iterations)	0.931

For the comparison of the accuracy, it can be seen from Fig.16 that our method achieved the highest accuracy with 99.09%, which is about 5 percentage points higher than methods with the lowest accuracy. In Table 3, in terms of the time consumption to build classification model and predict new instances in the dataset. J48, SMO and AdaBoostM1 these three traditional machine learning algorithms had lower time consumption. In addition, method of [3] used the atomic association rules classification method, which aimed to realize the construction and classification of the classification model in a shorter time with less classification association rules, so the good time performance of 0.347 seconds was also obtained. However, the accuracy of these classification methods was relatively low. Compared with these methods, the proposed BP neural network did not dominate the time performance because it needed to repeat the process of forward propagation and back propagation in multiple iterations when building a high-precision neural network classification model. Therefore, in view of the perspective of neural networks, we compared the time consumption with MLPClassfier, a neural network implemented in the scikit-learn library, and when the number of iterations was 1500, the time consumption reaches 5.596s, which was higher than the proposed BP neural network of 3.197s, and the accuracy was also lower than our method. In addition, because we used the learning rate attenuation method, the network convergence speed was accelerated, so that our network could reach convergence in 500 iterations. With the same 500 iterations setting, the time consumption of the proposed BP neural network in this paper was reduced to 0.931s, which had reached the same level as other non-neural network methods while maintaining an accuracy of 99.09%, but MLPClassfier still had a higher time consumption and its accuracy was reduced, which illustrates the effectiveness of the proposed BP neural network in this paper.

VI. CONCLUSION

This paper used an improved BP network to classify and predict Tibetan syndrome of CAG. Firstly, for the original dataset of Tibetan medical diagnosis and treatment, we used the K-means clustering algorithm to divide the syndrome of CAG scientifically from the perspective of data mining, and then coded the attribute values into discretized numbers. Besides, the CFS method was used for selecting the most useful features. Secondly, the input layer and output layer of the neural network were designed according to the number of features and classes in dataset. In addition, three empirical formulas were used to calculate the number of nodes in the hidden layer, and the optimal values were selected by comparison in the experiment. Furthermore, in order to prevent the over-fitting phenomenon, the L2 regular term was added to the loss function, and the learning rate attenuation method was used for the problem that the network convergence speed was slow. Finally, the experimental results showed that the improved BP neural network can obtain a classification accuracy of 99.09%, which had the best accuracy on various evaluation indicators compared with the method in [3] and some other traditional machine learning algorithms, and had a relatively good time performance. We thought that the reason for the good performance was that the two types of syndromes obtained are more scientific and reasonable by using the K-means clustering algorithm based on the elbow rule and the average contour coefficient. In addition, the number of new syndromes was reduced from four to two, and the number of features was reduced to nine with CFS method. The lower number of classifications and lower feature dimensions contribute to the improvement of accuracy. Simultaneously, we used the Gaussian distribution-based weight initialization method and the learning rate exponential decay method, which speed up the convergence rate of the neural network, which allowed us to achieve faster running time than other neural network methods. In addition, the L2 regularization method allowed us to avoid overfitting even with very small dataset. The proposed classification model in this paper can achieve excellent performance on various evaluation indicators, so it is expected to be extended to practical applications to provide clinical auxiliary decision-making support for the scientific diagnosis and treatment of plateau common diseases in Tibetan medicine.

ACKNOWLEDGMENT

The authors are grateful to the reviewers for valuable comments that have greatly improved the paper. They would like to thank the cooperation unit Qinghai Provincial Tibetan Hospital for providing the original dataset of common plateau disease of Chronic Atrophic Gastritis and the discussion and verification of experimental results.

REFERENCES

[1] D. Dradül and Z. Mei, "A brief history of Tibetan medicine development," *Herald Med.*, vol. 38, no. 4, pp. 456–460, 2019.

[2] Y. Liu and E. Liu, "A survey of medical data mining," *Guangming J. Chin. Med.*, vol. 33, no. 12, pp. 1714–1716, 2018.

[3] X. Zhu, L. Zhang, Y. Zhang, L. Wang, S. Wang, and P. Liu, "Research on classification of tibetan medical syndrome in chronic atrophic gastritis," *Appl. Sci.*, vol. 9, no. 8, p. 1664, Apr. 2019, doi: 10.3390/app9081664.

[4] X. Lu and C. Wang, "Analysis on chronic atrophic gastritis TCM syndrome type and treatment," *J. Liaoning Univ. Traditional Chin. Med.*, vol. 15, no. 6, pp. 140–141, 2013.

[5] D. Renqing, C. Douzhou, D. Gazang, J. Duo, J. Zhou, Q. Hua, and Y. Zhang, "Syndrome classification and distribution of chronic atrophic gastritis in Tibetan medicine," *China J. Traditional Chin. Med. Pharmacy*, vol. 31, no. 3, pp. 841–843, 2016.

[6] H. An, B. Zhang, Y. Guo, and J. Xu, "Study on chronic atrophic gastritis syndrome: An analysis of 172 cases," *J. Liaoning Univ. Traditional Chin. Med.*, vol. 17, no. 02, pp. 156–158, 2015.

[7] D. Renqing, C. Douzhou, J. Duo, J. Zhou, Q. Hua, D. Gazang, and Y. Zhang, "Analyzing etiology of chronic atrophic gastritis from perspective of tibetan medicine," *Liaoning J. Traditional Chin. Med.*, vol. 42, no. 12, pp. 2297–2299, 2015.

[8] W. Zhang, Y. Chen, B. Wei, and D. Li, "Schedule for diagnosis and treatment of chronic gastritis with in-tegrative chinese and western medicine (draft)," *Chin. J. Integr. Traditional Western Med. Digestion*, vol. 12, no. 5, 2004, Art. no. 314317.

[9] H. Tan, W. Wang, and Y. Li, "Review of classification algorithm in data mining," *Microcomputer Appl.*, vol. 2005, no. 2, pp. 4–6 and 9, 2005.

[10] W. Shiying, Z. Lei, W. Lu, C. Nanjia, Z. Xiaolan, L. Hong, and W. Xiaoying, "Research on syndrome classification prediction model of tibetan medicine diagnosis and treatment based on data mining," in *Proc. 12th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, 2016, pp. 497–502.

[11] S. Wang, "Research on the technology of decision support of tibetan medicine treatment of the common diseases of the plateau based on medical data mining," M.S. thesis, Dept. Comput. Technol. Appl. Eng., Qinghai Univ., Xining, China, 2018.

[12] X. Yu, "Comparison and thinking of data mining classification algorithm based on neural network," *Microcomput. Appl.*, vol. 34, no. 7, pp. 94–96, 2018.

[13] Y. Li, J. Peng, and Y. Zhang, "Application of artificial neural networks in medicine," *China J. Modern Med.*, vol. 2003, no. 13, pp. 8–11 2003.

[14] Y. Yasumoto, S. Yagi, K. Yana, M. Nozawa, and T. Ono, "State classification of heart rate variability by an artificial neural network in frequency domain," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 1401–1404, doi: 10.1109/iembs.2010.5626720.

[15] H. Fujita, T. Katafuchi, T. Uehara, and T. Nishimura, "Neural network approach for the computer-aided diagnosis of coronary artery diseases in nuclear medicine," in *Proc. IJCNN Int. Joint Conf. Neural Netw.*, Jan. 2003, pp. 215–220, doi: 10.1109/ijcnn.1992.227168.

[16] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "A comparison of neural network architectures for the classification of three types of infant cry vocalizations," in *Proc. 17th Int. Conf. Eng. Med. Biol. Soc.*, Nov. 2002, doi: 10.1109/iembs.1995.575380.

[17] W. Liu, C. Qin, K. Gao, H. Li, Z. Qin, Y. Cao, and W. Si, "Research on medical data feature extraction and intelligent recognition technology based on convolutional neural network," *IEEE Access*, vol. 7, pp. 150157–150167, 2019, doi: 10.1109/access.2019.2943927.

[18] Q. Xu, W. Tang, F. Teng, W. Peng, Y. Zhang, W. Li, C. Wen, and J. Guo, "Intelligent syndrome differentiation of traditional chinese medicine by ANN: A case study of chronic obstructive pulmonary disease," *IEEE Access*, vol. 7, pp. 76167–76175, 2019, doi: 10.1109/access.2019.2921318.

[19] B. Chen, J. Zhao, and S. Jiang, "The medical diagnosis expert system based on BP net," *Med. Inf.*, vol. 2007, no. 10, pp. 1743–1745, 2007.

[20] R. Zheng, "Auxiliary diagnosis of cardiac neurosis based on feedforward neural network," *Fujian Comput.*, vol. 34, no. 10, pp. 93–95, 2018.

[21] J. Zhou, J. Xue, C. Han, F. Xiao, and L. Sun, "Research on student performance prediction method based on BP neural network," *Comput. Era*, vol. 2018, no. 12, pp. 71–74, 2018.

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[23] R. Hecht-Nielsen, "Theory of the backpropagation neural network," *Neural Netw.*, vol. 1, p. 445, Jan. 1988, doi: 10.1016/0893-6080(88)90469-8.

[24] S. J. Hanson and L. Pratt, "Comparing biases for minimal network construction with back-propagation," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 177–185, 1989.

[25] W. S. Mcculloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bltn Mathcal Biol.*, vol. 52, nos. 1–2, pp. 99–115, Jan. 1990, doi: 10.1007/bf02478259.

[26] X. Yao, X. Wang, Y. Zhang, and W. Quan, "Summary of feature selection algorithms," *Control Decis.*, vol. 27, no. 02, pp. 161–166 and 192. 2012.

[27] J. Wang, L. Ci, and K. Yao, "A survey of feature selection," *Comput. Eng. Sci.*, vol. 2005, no. 12, pp. 72–75, 2005.

[28] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci. Eng., Waikato Univ., Hamilton, New Zealand, 1999.

[29] Q. Zhang and X. Li, "A new method to determine hidden note number in neural network," *J. Jishou Univ., Natural Sci. Ed.*, vol. 2002, no. 1, pp. 89–91, 2002.

[30] B. Cai, "Design method on structure of implicit strata in BP neural network," *J. Tonghua Normal Univ.*, vol. 2007, no. 2, pp. 18–19, 2007.

[31] P. Jin and Y. Rui, "Research and application of various improved algorithms of BP algorithm," *J. Nanjing Univ. Aeronaut. Astronaut.*, vol. 1994, no. S1, pp. 201–205, 1994.

[32] W. Jin, S. Fang, S. Yan, and H. Li, "Methods to improve BP network," *J. Shenyang Jianzhu Univ. Natural Sci.*, vol. 2001, no. 03, pp. 197–199 and 205, 2001.

[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[34] Q. Deng, J. Cheng, H. Wang, and L. Deng, "Dynamical regularization method for neural network optimization based on maturity of knowledge," *J. Electron. Meas. Instrum.*, vol. 32, no. 2, pp. 113–118, 2018.

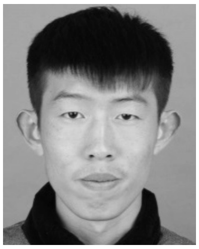
[35] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 464–472.

[36] J. Quinlan, *C4.5—Programs for Machine Learning*. San Mateo, CA, USA: Kaufmann, 1992.

[37] J. Platt. (2019). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Microsoft Research. [Online]. Available: https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/

[38] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proc. ICML*, 1996, pp. 148–156.

[39] B. W. White and F. Rosenblatt, "Principles of neurodynamics: Perceptrons and the theory of brain mechanisms," *Amer. J. Psychol.*, vol. 76, no. 4, p. 705, Dec. 1963, doi: 10.2307/1419730.



SHENGHAO YANG is currently pursuing the degree with the Department of Computer Technology and Applications, Qinghai University, China. His main research interests include neural networks and data mining.

XIAOLAN ZHU received the master's degree from the College of Computer Science, Sichuan University, China. She is currently a Teacher with Qinghai University, China. Her research interests include data mining, deep learning, and medical bigdata analysis.

LEI ZHANG is currently an Associate Professor with the College of Computer Science, Sichuan University, China. His research interests include domain data mining and mobile computing.

LU WANG is currently a Teacher with Qinghai University, China. Her research interests include data mining and bigdata analysis.

XIAOYING WANG received the Ph.D. degree from Tsinghua University. She is currently a Professor with the Department of Computer Technology and Applications, Qinghai University, China. Her research interests include cloud computing and parallel computing.