INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.42 Volume 15, 2021

A Matching Method of Heterogeneous Database based on SOM and BP Neural Network

Yongjie Zhu, Shenzhan Feng*
Information Management Center, Xuchang University, Xuchang, 461000, China

Received: September 26, 2020. Revised: March 31, 2021. Accepted: April 15, 2021. Published: April 22, 2021.

***Abstract*—In the process of data integration among heterogeneous databases, it is significantly important to analyze the identical attributes and characteristics of the databases. However, the existing main data attribute matching model has the defects of oversize matching space and low matching precision. Therefore, this paper puts forward a heterogeneous data attribute matching model on the basis of fusion of SOM and BP network through analyzing the attribute matching process of heterogeneous databases. This model firstly matches the heterogeneous data attributes in advance by SOM network to determine the centre scope of attribute data to be matched. Secondly, the accurate match will be carried out through BP network of the standard heterogeneous data various attribute center. Finally, the matching result of the relevant actual database shows that this model can effectively reduce the matching space in the case of complex pattern. As for the large-scale data matching, the matching accuracy is relatively high. The average precision is 89.52%, and the average recall rate is 100%.**

***Keywords*—heterogeneous data attribute matching, SOM network, BP network**

I. INTRODUCTION

Database integration is the basic condition to realize the sharing of information resources. Whether the information between different databases can be integrated without gap will directly affect the comprehensive utilization of data information resources[1-3]. The data integration provides a unified storage and management model and shields the independent operation for all kinds of heterogeneous data[4-5].

Therefore, the integrated heterogeneous data achieves the data sharing between heterogeneous data sources, and effectively utilizes the effective information between different data sets. However, there are various kinds of semantic heterogeneity and data heterogeneity between different databases[6-7], which leads to the unsuccessful integration of heterogeneous database in the matching process.

The academic research on heterogeneous data integration started relatively early in foreign countries with more in-depth researches[8]. However, the corresponding prototype system developed abroad is not suitable for large-scale in commercial field. Although the importance of information integration has been realized in recent years, the research depth is insufficient due to the late start. Most domestic researchers only improved foreign research results on the basis of the theory. The data integration system which could meet the requirements of domestic large-scale utilization hasn't been developed [9]. At the same time, a large number of domestic system integration companies only use foreign data system software for simple information integration and conversion. The integrated data is relatively rough, which can't solve the problem of data conflict fundamentally, and the real data sharing is difficult to be achieved[10].

However, it is the key factor to solve the problem of database integration to describe the data semantic features and to find out the semantic attribute matching and entity matching between heterogeneous databases. Therefore, in view of above problems, this paper focuses on the analysis of the attribute matching problem of heterogeneous databases and puts forward a heterogeneous data attribute matching model on the basis of fusion of SOM and BP network through analyzing the structures and model characteristics. This will provide a practical and effective theoretical guidance for the development of heterogeneous database integration system.

II. SOM NEURAL NETWORK

A. Overview of SOM Neural Network

Self organizing maps (SOM) is an unsupervised competitive learning network proposed by Kohonen et al. It classifies the input data by dividing them into different regions [11-12]. Different from other neural networks, it can not only classify the input vectors in the training process, but also study the

distribution of the input vectors in the network.

Because each neuron in the SOM neural network topology is related to its neighboring neurons, the neurons with similar functions are clustered together by learning the distribution characteristics of input vectors and topology structure, so as to learn and classify them . In the process of learning, competition is unsupervised, that is to say, in the process of learning, we only need to provide some sample data to the network, and do not need to input labels to the sample data, that is, we only need to provide input but not output to the network. In the process of training, the network automatically finds the rules between samples, adaptively adjusts the weights in the process of network connection, and completes the classification of data.

SOM network consists of the following parts.

(1) Processing unit array. The main function is to accept the input data, and then generate a "discriminant function" for the input data processing.

(2) Comparative selection mechanism. The main function is to select the winning neurons by comparing with the "discriminant function".

(3) Local interconnection. The main function is to stimulate the selected neuron and its adjacent neurons.

(4) Adaptive process. The main function is to update the relevant parameters of the stimulated neuron and increase the output value of the "discriminant function" of the input data.

SOM network training process is relatively simple. After the training samples are input into the input layer, each neuron in the layer calculates the distance between the input data and its own weight vector, selects the winning neuron with the smallest distance, and then updates the weight vectors of the winning neuron and its adjacent neurons, so as to reduce the distance between the weight vector of the neuron and the input data, and continuously iterates the process until convergence.

When the network training is over, the corresponding relationship between each neuron in the competition layer and the input mode is determined, and the test data can be used to test. When a pattern is input, the neurons representing the pattern in the competition layer will win in the competition, and the input data will be classified automatically. However, when the input pattern is not seen in training, the network will automatically classify it into the nearest category.

SOM network can automatically classify the input data without a large number of data samples or manual intervention[13]. In fault diagnosis, SOM network has strong real-time performance, but it has long training time and slow convergence. More importantly, the number of nodes and the arrangement of competition layer will affect the fault classification.

B. Algorithm Steps of SOM Neural Network

The learning rules of SOM neural network are obtained from the lateral inhibition of neurons.

Step 1. Network Initialization. The connection weights of the network are initialized randomly.

Step 2. Input vector. Input the input vector $X=(x_1(t),x_2(t),...,x_m(t))^T$ to the input layer.

Step 3. Find the winning node.

$$d_j=\left\|X-w_j\right\|=\sqrt{\sum_{i=1}^{m}(X_i(t)-w_{ij}(t))^2} \quad (1)$$

$$\left\|X-w_c\right\|=\min\{d_j\} \quad (2)$$

Where w_{ij} is the connection weight between input layer neuron *i* and competitive layer neuron *j*.

Step 4. Adjust the weight and update the formula, such as (3)

$$w_{ij}(t+1)=w_{ij}(t)+\eta(t)h(t)(x_i-w_{ij}(t)) \quad (3)$$

Where $\eta(t)$ is the learning rate function, its value range is $0<\eta(t)<1$, and $h(c)$ is the neighborhood function, which gradually decreases with time. Their learning rules are as follows

$$h(t)=\exp(-\frac{d_{cj}^2}{2r^2(t)}) \quad (4)$$

$$r(t+1)=INT((r(t)-1)\times(1-\frac{t}{T}))+1 \quad (5)$$

$$\eta(t+1)=\eta(t)-\frac{\eta(0)}{T} \quad (6)$$

Where d_{cj} is the distance between neuron c and neuron j , $r(t)$ is the neighborhood radius, *INT* is the integer function, and *T* is the total number of learning.

Step 5. Let $t=t+1$ and return to Step 2 until the maximum number of iterations or learning rate reaches a set value.

The algorithm flow of SOM neural network is shown in the Figure 1.

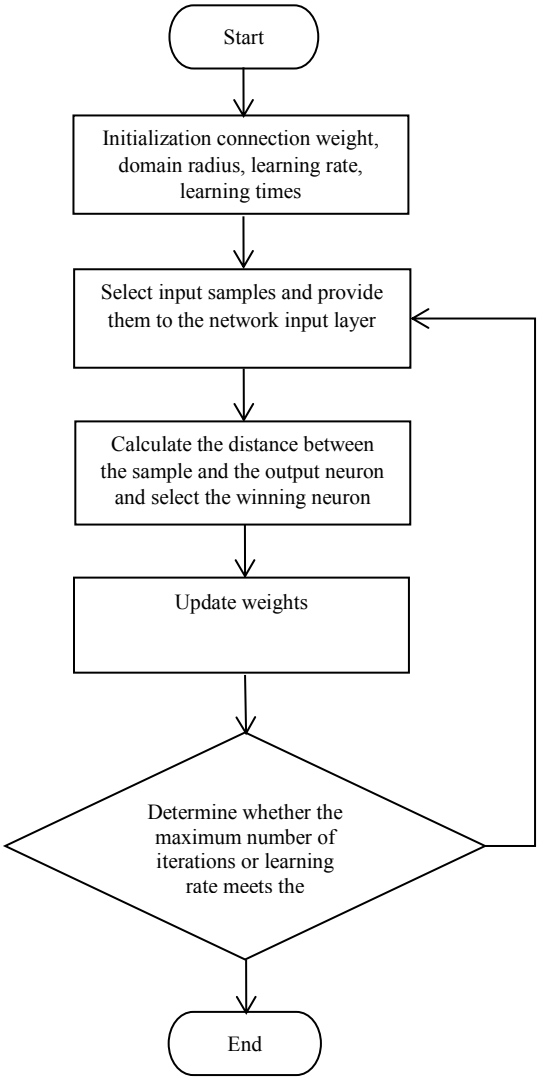


Fig. 1. The flow chart of SOM network algorithm

C. Limitations of SOM Neural Network

SOM network can classify input patterns automatically. It uses multiple neurons to respond to the classification results at the same time. It does not need a large number of sample data, and its network structure is simple, and the algorithm process is easy to implement. However, it also has shortcomings.

(1) Before training, users need to initialize the number of clusters and the initial weight matrix [14], that is to say, the number of clusters and the network structure are fixed and can not be adjusted during training.

(2) In the process of training, some neurons never win in the competition, while some neurons are overused in the process of learning [15], which will affect the training performance of the network.

(3) In training, if you add a new type to the training sample, you must relearn it.

(4) The input order of data will affect the output result, and sometimes determine the output result [16]. This feature is particularly obvious when the amount of data is small.

(5) If the initial value and parameters of the connection weights between the input layer and the competition layer are not suitable, the convergence time of the network will be too long, or even can not reach the convergence state.

III. BP NEURAL NETWORK

A. The Basic Idea of BP Neural Network

Back propagation (BP) network is a kind of multilayer feedforward neural network proposed by McCelland et al. Its learning is divided into two stages: one is the forward propagation stage of the signal, the sample data is input to the input layer, and the middle is processed by each hidden layer [17]. The processed results are transmitted to the output layer, and the output layer outputs the final results. The other is the error back propagation. The error between the expected output and the actual output is transmitted to the input layer through the hidden layer in a certain way. In the process of transmission, the error is divided equally to each unit in the middle. Each neuron obtains its own error and modifies its own weight according to the error. The process of weight modification is the process of neural network learning. The process continues until the number of learning times set before the training of sample data is reached, or the output error reaches the allowable range. Its structure is shown in Figure 2, which is composed of input layer, hidden layer and output layer.

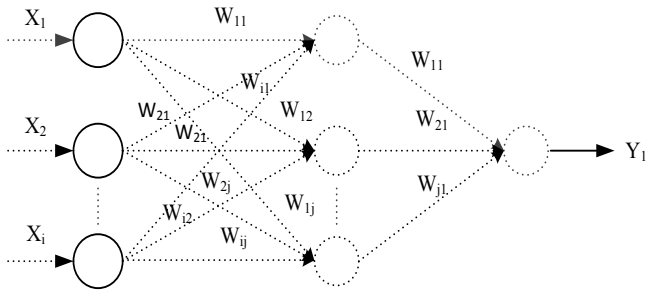


Fig. 2. The structure of BP neural network

In Figure 2, X_1, X_2, ..., X_i is the input, Y_1 is the output, W_{ij} is the weight from the input layer to the hidden layer, W_{jk} is the weight from the hidden layer to the output layer.

B. Algorithm Steps of BP Neural Network

BP neural network algorithm steps are as follows.

Step 1. Network initialization.

Step 2. Calculate the output of the hidden layer.

Let the hidden layer and the output layer adopt *sigmoid* function, that is,

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

$$H_j = f\left(\sum_{i=1}^{n} w_{ij} x_i - \theta_j\right) \tag{2}$$

Step 3. Calculate the output of the output layer.

$$O_k = \sum_{j=1}^{q} H_j w_{jk} - b_k \quad (3)$$

Step 4. Calculation error.

According to the actual output and the expected output, the error e_k and mean square error E_k are calculated

$$e_k = Y_k - Q_k \quad (4)$$

$$E_k = \frac{1}{2}\sum_{j=1}^{l}(Q_k - Y_k)^2 \quad (5)$$

Step 5. Update weights.

$$w_{ih} = w_{ih} + \eta H_j (1 - H_j) x_i \sum_{k=1}^{n} w_{hi} e_k \quad (6)$$

$$w_{hj} = w_{hj} + \eta H_j e_k \quad (7)$$

Where η is the learning rate.

Step 6. Update threshold.

$$\theta_j = \theta_j + \eta H_j (1 - H_j) \sum_{k=1}^{l} w_{jk} e_k \quad (8)$$

$$\gamma_h = \gamma_h + e_h \quad (9)$$

Step 7. It is judged whether it is within the allowable range according to the output result error. If it is, the algorithm ends, otherwise it jumps to step 2.

Figure 3 is the flow chart of BP neural network algorithm.

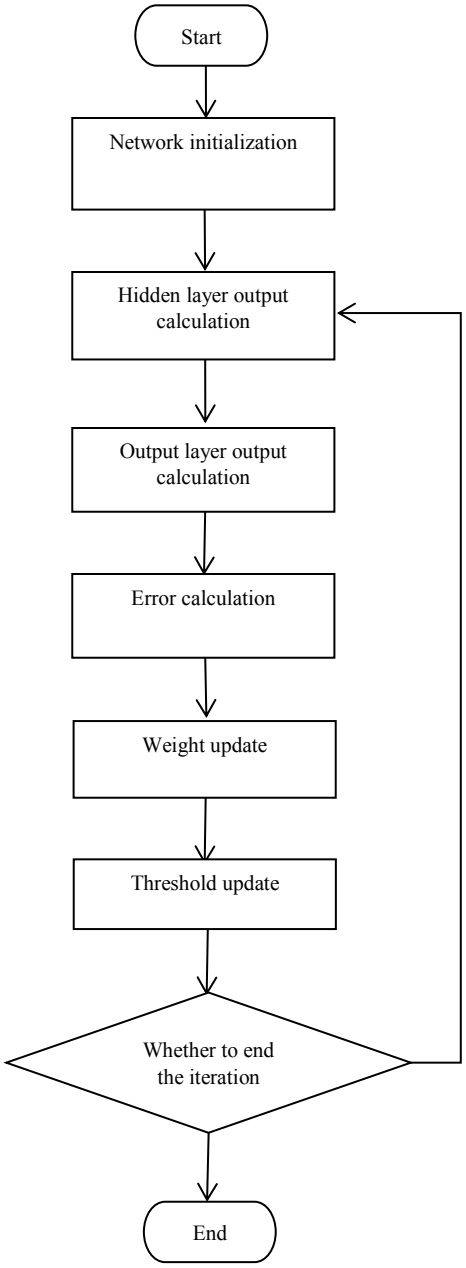


Fig. 3. The flow chart of BP network algorithm

C. Shortcomings of BP Neural Network

BP neural network algorithm can approximate any nonlinear function with any precision, but it also has some shortcomings.

(1) Because BP neural network algorithm is a local search method, and it solves a nonlinear problem, the network connection weight is adjusted along the direction of local improvement in the training process, so it is easy to fall into local minimum [18]. In addition, with different initial weights, the network converges to different local minima in the training process, which makes the performance of the network unstable.

(2) There is no unified theoretical guidance for the construction of the initial network structure, which is generally set according to experience, and then adjusted according to the output results [19], which will directly affect the approximation

ability and generalization ability of the network, thus affecting the final effect of the algorithm.

(3) The learning ability and generalization ability of BP neural network are closely related to the selection of samples [20]. If the selected samples are contradictory, redundant and unrepresentative, the training of BP neural network is difficult to achieve the desired effect.

IV. Heterogeneous Database Attribute Matching Model Based on SOM Competitive Fusion BP Network

A. The Idea of SOM-BP Algorithm

The main idea of SOM-BP network is to combine unsupervised learning with supervised learning. They are combined in series to form a new network and in parallel to form a new network. In this paper, SOM-BP neural network is composed of SOM neural network and BP neural network in series. When the data is input into SOM network, the competition layer outputs the competition winning neuron, and the network classifies the input samples according to the position of the winning neuron. The generalization ability of SOM network just overcomes the influence of tolerance factor on its classification, so it does not need a lot of data training. In this paper, SOM network and BP network are combined to avoid the defect that BP neural network training needs a lot of data.

SOM-BP neural network adds a competition layer before the hidden layer of traditional BP neural network, which is composed of input layer, competition layer, hidden layer and output layer. First of all, the input samples are classified, and the SOM network automatically carries out linear mapping to cluster the sample data, so as to complete the preliminary classification of samples and reduce the pressure of BP network recognition. Then, the clustering information is transferred from the competitive layer to the hidden layer, and the clustering information is transferred to the BP network The neural network is trained in supervised learning mode to complete the nonlinear mapping between input and output and classify the data.

After the sample data is trained by SOM neural network, the related data will gather together to form a specific data set. After the data sample is clustered by SOM, the classification of the data sample is preliminarily obtained. Then the results of SOM preliminary clustering are normalized, and the processed results are used as the input of BP neural network to train and test the model.

B. SOM-BP Algorithm Steps

The learning steps of SOM-BP neural network are as follows.

Step 1. Select sample data, preprocess the data, take part as training data, and part as test data.

Step 2. SOM neural network is initialized and sample data are input. The preliminary classification results are get after training.

Step 3. The output of competition layer of SOM network is normalized.

Step 4. The BP neural network is initialized, and the normalized SOM output is input into BP network for training again.

Step 5. After many times of training, the SOM-BP network classification model is formed. The test samples are input into the model for testing, and the results are analyzed and compared.

The algorithm flow chart of SOM-BP neural network is shown in Figure 4.

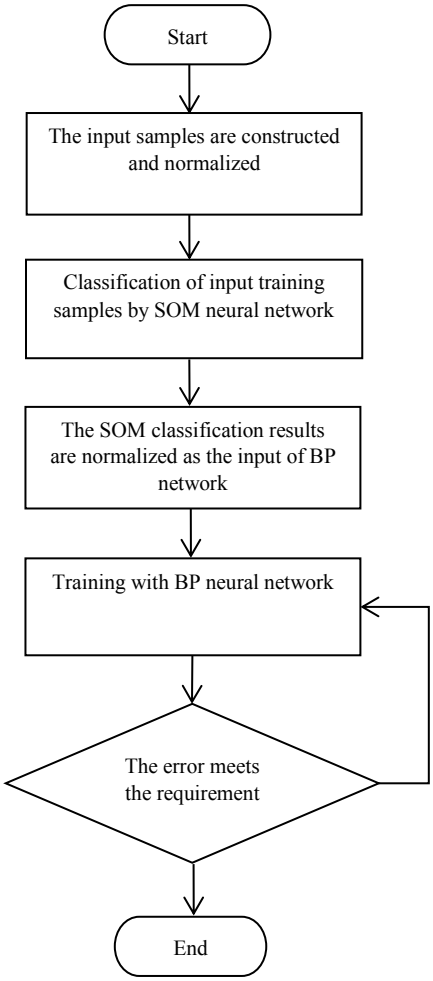


Fig. 4. The flow chart of SOM-BP network algorithm

C. Heterogeneous Database Attribute Matching Model

When the characteristics match of two heterogeneous data attributes need to be carried out, the heterogeneous database attributes to be matched and the standard database SOM network can be classified rapidly through SOM network. That is to say, the attributes characteristic center within the standard database can be quickly found, which will be carried out with detailed and accurate attribute matching. Finally, the attribute matching of two heterogeneous databases will be completed.

Through above analysis of the matching method, the construction steps of heterogeneous database attribute matching mode on the basis of SOM and BP neural network put forward in this paper are as follows:

Step 1. Select the attribute characteristic value of the heterogeneous database, and carry out the normalization treatment for attribute characteristic value of the standard heterogeneous database and the heterogeneous database to be matched.

Step 2. Input each attribute index of the standard heterogeneous database into SOM network, and the index with similar attribute will be classified through the mutual competition of neurons. *N* class attribute center of standard heterogeneous database will be obtained.

Step 3. Extract the N class attribute center sample of standard heterogeneous database, and establish the BP neural network of each attribute center, among which there are N BP networks.

Step 4. In the matching process, rapidly classify each attributes characteristic value of the heterogeneous database to be matched through SOM network established in Step 2, and carry out the accurate matching through the attribute center BP network established in Step 3 (The threshold value is set as 0.9.) And input the complete matching result.

V. RESULT ANALYSIS

A. Heterogeneous Database Sample Index

This paper put forward that the model information use of attributes can't effectively classify the attributes. Therefore, in order to recall rate of attribute matching, the data index system which describes the attributes should include the mode information, data content information and other limitation semantic information.

However, through sensitivity analysis result of each attribute data, the main indicators which play important roles should include: data type, length of data type, whether allow for null or not, data precision, decimal digits, minimum value, maximum value, average value, difference coefficient, standard deviation and numeric character ratio. Therefore, the 13 data indexes used in attribute matching in this paper include: character type, value type, rare type, length of data type, whether allow for null or not, precision, decimal digits, minimum value, maximum value, average value, difference coefficient, standard deviation, and numeric character ratio. When the attribute data type is described as a character type, minimum value, maximum value, average value, difference coefficient and standard deviation respectively refer to the minimum string length, the maximum string length, average string length, corresponding difference coefficient and standard deviation. As for the date type attributes, the minimum value, maximum value, average value, difference coefficient and standard deviation of attribute value will be calculated with the function of year and month. The attributes normalization method adopted in this paper includes

(1) The information of true/false with the binary system evaluation characteristic is expressed as 0 or 1 respectively.

(2) For the class information with determined value, if is required to be quantified as a numerical value and normalized to the interval of [0,1]. The adopted normalization formula is f(length)=2(1/(1+1.01-length)-0.5).

(3) As for the data type of Int,Decimal,Bit,Char,Money and Datetime, they are expressed as the integer of 20，30，40，100，150 and 400. The differences of these data types are obvious. The adopted normalization formula is f(length) =2(1/(1+1.01-length)-0.5).

(4) For other data types, the same standard formula will be adopted: f(length)=2(1/(1+1.01-length)-0.5)

The experimental database selected in this paper is from the SQL Server2000 database. The attribute data of the standard heterogeneous database are shown in Table 1, which contains 8 attribute characteristics, including 2 numeric type attributes, 4 character type attributes, and 2 rare type attributes. The heterogeneous database to be matched is shown as Table 2, which contains 10 attribute characteristics, including 1 numeric type attributes, 7 character type attributes, and 2 rare type attributes.

By observing and comparing the data in Table Employees and Table Orders, there are totally three pairs of pairs of attributed could be matched, which are (E.EmployeeID, O.EmployeeID), (E.City, O.ShipCity) and (E.HireDate, O.OrderDate). Since the correct matching results have been known in advance, the validity of the attribute matching method put forward in this paper can be examined.

Table 1. Standard Orders data sheet attribute information (after normalization)

Attribute	Data type	Attribute value
Order ID	Numeric type	[0,0.0992,0.0199,0,0.0497,0,1,1,1, 0.0001,0.831,1]
Employee ID	Numeric type	[0,0.0992,0,0.0199,1,0.0497,0,0.00 5,0.0448,0.0199,0,0.012,1]
Customer ID	Character type	[0.46,0,0,0.0249,1,0,0,0.0249,0.02 49,0.0249,0,0,0]
Ship Name	Character type	[0.46,0,0,0.196,1,0,0,0.0398,0.168, 0.0878,0.0014,0.0248,0]
Ship City	Character type	[0.46,0,0,0.0745,1,0,0,0.0199,0.07 45,0.0402,0.0018,0.0148,0]
ShipPostal Code	Character type	[0.46,0,0,0.0497,1,0,0,0,0.0447,0.0 273,0.0016,0.0088,0.97]
		[0,0,0.963,0.0398,1,0,0,1,1,1,0,0.0 035,0.778]

Order Date	Rare type	[0,0,0.633,0.0398,1,0.094,0.0199,0.001,0.9999,0.371,0.0074,0.5234,1]
Freight	Rare type	

Table 2. Employees data attributes information to be matched (after normalization)

Attribute	Data type	Attribution center category
Employee ID	Numeric type	[0,0.0992,0,0.0198,0,0.0497,0,0.005,0.0447,0.0249,0.0027,0.0136,1]
Last name	Character type	[0.46,0,0,0.0990,0,0,0,0.0199,0.0447,0.0354,0.0011,0.008,0]
First name	Character type	[0.46,0,0,0.0497,0,0,0,0.0199,0.0398,0.0287,0.001,0.006,0]
Title	Character type	[0.46,0,0,0.1482,1,0,0,0.0646,0.119,0.0981,0.0007,0.0142,0]
Courtesy	Character type	[0.46,0,0,0.1237,1,0,0,0.0149,0.0199,0.0155,0.0005,0.0017,0]
Address	Character type	[0.46,0,0,0.2899,1,0,0,0.0745,0.148,0.105,0.0012,0.0251,0.1257]
City	Character type	[0.46,0,0,0.0745,1,0,0,0.0298,0.0398,0.0326,0.0006,0.0036,0]
Country	Character type	[0.46,0,0,0.0745,1,0,0,0.001,0.0149,0.0127,0.001,0.0026,0]
BirthData	Rare type	[0,0,0.963,0.0396,1,0,0,1,1,1,0,0.0451,0.778]
HireData	Rare type	[0,0,0.963,0.0396,1,0,0,1,1,1,0,0.0043,0.778]

B. Experimental results of heterogeneous database attribute matching

Construction of standard heterogeneous data attribute center and center BP network. The standard orders data sheet attribute value are taken and each sample consists of 13 dimensionality input information, in which the structure diagram of the competitive neural network can be seen in Figure 3.

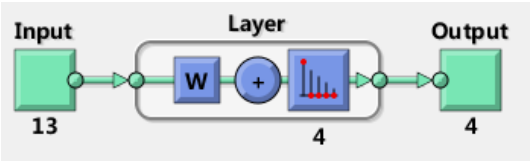


Fig. 3 SOM neural network model structure diagram

In which the standard Orders data table attribute center category number is set as 4. When Trainru (unsupervised random) function is used to train the network, both the initial learning rate and the adjusted learning rate are 0.3, and the maximum iteration number is 150. Through the training of the above SOM network, the center of the standard Orders data table attribute is shown in Table 3.

Table 3. the center of the standard Orders data table properties

Attribute	data type	Attribution center category
Order ID	Numeric type	4
Employee ID	Numeric type	2
Customer ID	Character type	1
Ship Name	Character type	1
Ship City	Character type	1
Ship Postal Code	Character type	2
Order Date	Rare type	4
Freight	Rare type	3

The already-built neural network is trained and simulated. The distribution results of the gradient of various indexes obtained through convergence after the end of network training can be seen in Figure 4.

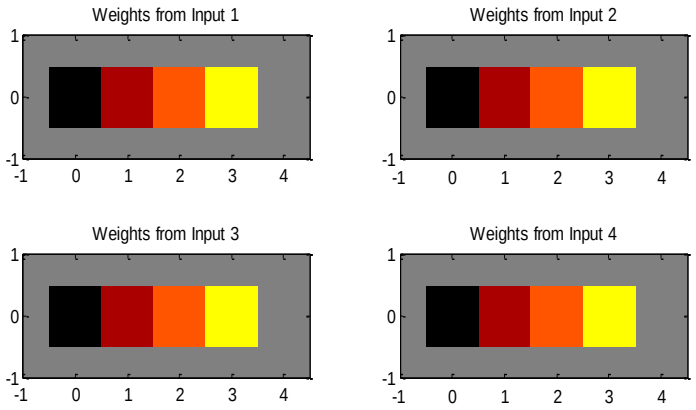


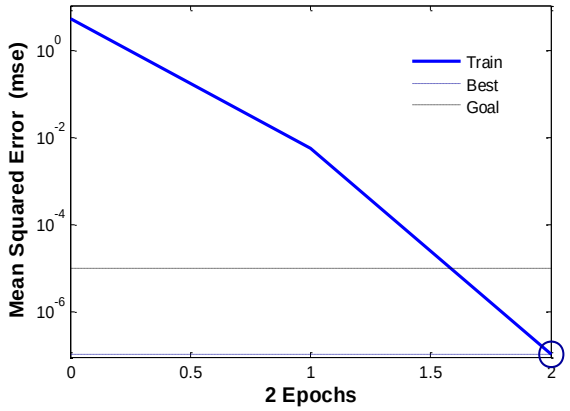
Fig. 4 Distribution results of the gradient of different index weights during network convergence

Extract the sample of the each attribute center, and take them as the input information of the BP neural network. The results can be seen in Table 4 and Figure 5.

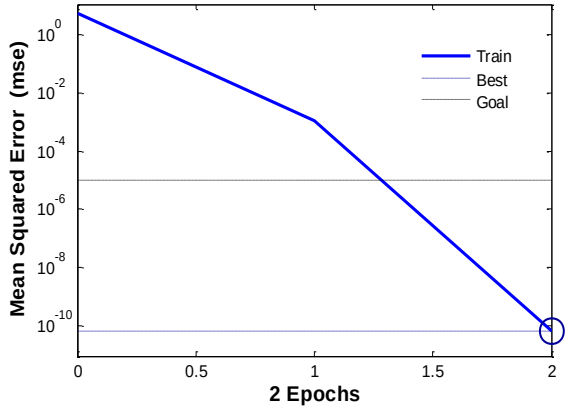
Table 4. The parameters values

BP network attribute index	Parameter values
number of layers	2
Quantity of the node in the 1th layer	20
Quantity of the node in 2th layer	sample number of attribute center

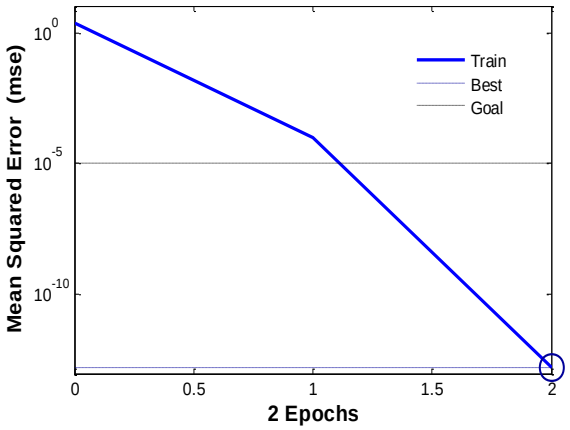
Network target error	0.00001
Network training function	trainlm
Network learning rate	0.3



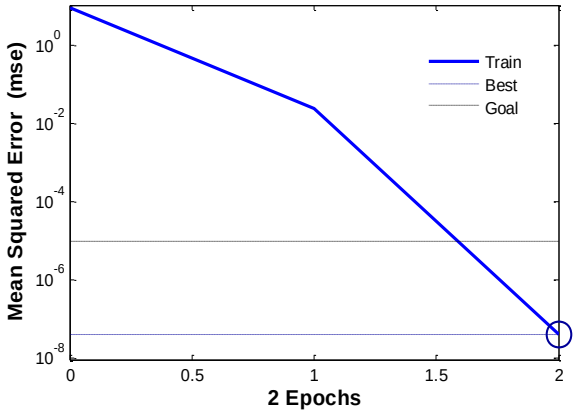
(a) Best training performance is 1.0224e-007



(b) Best training performance is 6.5271e-011



(c) Best training performance is 1.475e-013



(d)Best training performance is 3.9969e-008

Fig. 5 the standard Orders data attribute center BP neural network training error results

Figure a, b, c, and d are respectively the BP network training error variation of the attribute center 1,2,3, and 4. The heterogeneous data attributes to be matched. Classify the attribute value of above Employees data table with the SOM network established through the standard Orders data table attribute values, and the results can be seen in Table 5.

Table 5. the classification results of Employees data table attribute center to be matched

Attribute	data type	Attribution center category
Employee ID	Numeric type	2
Last name	Character type	1
First name	Character type	1
Title	Character type	1
Title of Courtesy	Character type	1
Address	Character type	1
City	Character type	1
Country	Character type	1
Birth Data	Rare type	4
Hire Data	Rare type	4

As shown in Table 5, the attribute value of Employees data table to be matched basically find the matched centre sample through the rapid search of SOM network. The results show that the attributes of E.EmployeeID and O.EmployeeID are classified into the second type, the attributes of E.City and O.ShipCity are classified into the second type, the attributes of E.HireDate and O.OrderDate are classified into the fourth type.

Carry out accurate matching for the Employees data table

attribute values to be matched after rapid classification of the matching through the BP network of the class center. The results show that the model can be matched exactly and accurately, and the results can be seen in Table 6.

Table 6. The results of the accurate matching of standard Orders and Employees data sheet attribute

Orders data sheet		Employees data sheet	
Attribute	Actual output value	Attribute	Predicted output value
		Last name	0.0001, 0.0421, 0.9974
		First name	1.0000, -0.0001, 0.0004
Customer ID	1,0,0	Title	-0.0421, 0.9921, 0.0544
Ship Name	0,1,0	Courtesy	-0.0041, 0.8914, 0.0044
Ship City	0,0,1	Address	-0.0001, 0.9991, 0.0004
		City	0.0037, 0.0042, 0.9874
		Country	-0.0054, 0.9541, 0.1247
Employee ID	1,0	Employee ID	0.9584, 0.0022
ShipPostalCode	0,1		
Order ID	1,0	Birth Data	0.8852, -0.1257
Order Date	0,1	Hire Data	-0.0014, 0.9752

In order to better reflect the model matching condition put forward in this paper, apply above model to the large scale database test and carry out test on sample database Northwind of SQL Server 2000 and Access 2002. There are totally 126 attributes in the tested data, of which 38 numerical attributes, 76 character type properties, 12 rare type attributes, 62 similar attributes in SQL Server 2000 and Access 2002. Take Northwind of SQL as the standard database and Northwind of Access as the database to be matched.

To compare the validity of SOM-BP neural network model during the model test, Table 7 gives the prediction results of raw data by directly using Logistic model, BP neural network model and SVM model algorithms.

Table 7. Comparison of accuracy between models

Model algorithm	Accuracy	Error rate
Logistic model	83.3%	16.7%
BP neural network model	73%	27%
SVM model	87.2%	12.8%
Present model	95.5%	4.5%

As can be seen from Table 7, the model algorithm proposed herein has high recognition accuracy, as SOM model excludes poor training samples and automatically matches the samples number during sample data extraction process. Meanwhile, the model still has a high validity when the amount of data is small since the algorithm proposed extracts the relatively optimal samples in advance. In comparison other models present affected training and predictive abilities as the training sample objects contain many confused samples, which reduces their prediction accuracy.

VI. Conclusion

This paper introduces the significance of heterogeneous database integration research, analyzes the difficulties of heterogeneous database attribute matching, and puts forward the heterogeneous data attribute matching model. Firstly, the model standardizes the index of heterogeneous data attribute matching process, and achieves all kinds of attributes center of the standard heterogeneous database through SOM network. Then it extracts all kinds of attribute center samples of the standard heterogeneous database, to establish the BP neural network matching system of various attribute centers respectively. In the matching process, the attribute value of heterogeneous database to be matched is classified rapidly through the SOM network which is established by the standard heterogeneous database. Then carry out accurate match through the BP neural network established by the various attribute center samples of standard data. Then carry out the matching calculation of related actual database. The result shows that when the mode is relatively complex, this method can effectively reduce the matching space. As for the matching of large scale database, the matching precision is high.

In the research and experiment, it is found that the model still has some shortcomings and needs further study. In the algorithm, the parameters are not calculated by theoretical formula, but determined by empirical function, that is, the optimal value is obtained through a large number of experiments, in which there are many uncertainties, which will affect the final optimization effect. Therefore, other effective algorithms can be used to optimize in order to find the optimal solution. In this paper, the training and testing of the model are based on historical data, not real-time. So we can update the sample data online, train and learn autonomously, and improve the generalization ability and adaptability of the system.

REFERENCES

[1] Asghari A , Sohrabi M K , Yaghmaee F . Online scheduling of dependent tasks of cloud's workflows to enhance resource utilization and reduce the makespan using multiple reinforcement learning-based agents. Soft Computing, 2020, 24, pp. 1-23.

[2] Prasad V K , MD Bhavsar. Monitoring IaaS Cloud for Healthcare Systems: Healthcare Information Management and Cloud Resources Utilization. International Journal of E-Health and Medical Communications, 2020, 11, pp. 122-131.

[3] Skrami E , Carle F , Villani S , et al. Availability of Real-World Data in Italy: A Tool to Navigate Regional Healthcare Utilization Databases. International Journal of Environmental Research and Public Health, 2020, 17, pp. 65-72.

[4] Xu X . Material database management system based on heterogeneous multi-processor and computer embedded system. Microprocessors and Microsystems, 2021, 82, pp.103926.

[5] Asfand-E-Yar M, Ali R. Semantic Integration of Heterogeneous Databases of Same Domain Using Ontology. IEEE Access, 2020, 8, pp. 77903-77919.

[6] Mollero R , X Pennec, Delingette H , et al. Population-based priors in cardiac model personalisation for consistent parameter estimation in heterogeneous databases. Communications in Numerical Methods in Engineering, 2019, 35(2), pp. e3158.1-e3158.25.

[7] Yu H K , Min B L , Nam S H , et al. Enhancing the Accuracies of Age Estimation with Heterogeneous Databases Using Modified CycleGAN. IEEE Access, 2019, 99, pp.1-12.

[8] Yanni Zhao, Hualei Guo. Research on Heterogeneous Database Migration Technology Based on XML. Computer and Digital Engineering, 2018, 46, pp.129-133.

[9] Yan Z, Chen X, Tang X. A Novel Linear Model Based on Code Approximation for GNSS/INS Ultra-Tight Integration System. Sensors, 2020, 20, pp. 3192.

[10] Aa A , Fm A , Kg A , et al. XRepo - Towards an information system for prognostics and health management analysis. Procedia Manufacturing, 2020, 42, pp. 146-153.

[11] Sousa M , Pires R , Del-Moral-Hernandez E . SOMprocessor: A high throughput FPGA-based architecture for implementing Self-Organizing Maps and its application to video processing. Neural Networks, 2020, 125, pp. 349-362.

[12] Junior P O, Conte S, D 'Addona D M, et al. An improved impedance-based damage classification using Self-Organizing Maps. Procedia CIRP, 2020, 88, pp. 330-334.

[13] J Feng, Wang F , Wang Q , et al. Intraseasonal variability of the equatorial Pacific Ocean and its relationship with ENSO based on Self-Organizing Maps analysis. Journal of Oceanology and Limnology, 2020, 38, pp. 1108-1122.

[14] Oleg Milder, Dmitry Tarasov, Andrey Tyagunov, The Artificial Neural Network Structure Selection Algorithm in the Direct Task of Spectral Reflection Prediction, WSEAS Transactions on Systems and Control, Volume 14, 2019, Art. #9, pp. 65-70.

[15] Ashkan Tashk, Jrgen Herp, Esmaeil Nadimi, Automatic Segmentation of Colorectal Polyps based on a Novel and Innovative Convolutional Neural Network Approach, WSEAS Transactions on Systems and Control, Volume 14, 2019, Art. #47, pp. 384-391.

[16] Fan Z, Alley A, Ghaffari K, et al. MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. Metabolomics, 2020, 16, pp. 104.

[17] Wang J , Zhao Z , Liu Y , et al. Research on the Role of Influencing Factors on Hotel Customer Satisfaction Based on BP Neural Network and Text Mining. Information (Switzerland), 2021, 12, pp. 99.

[18] Wu L, J Zhou, Li Z. Applying of GA-BP Neural Network in the Land Ecological Security Evaluation. IAENG Internaitonal Journal of Computer Science, 2020, 47, 11-18.

[19] Wan P , Zou H , Wang K , et al. Research on hot deformation behavior of Zr-4 alloy based on PSO-BP artificial neural network. Journal of Alloys and Compounds, 2020, 826, pp. 154047.

[20] Panda S , Panda G . Performance Evaluation of a New BP Algorithm for a Modified Artificial Neural Network. Neural Processing Letters, 2020, 51, pp. 330-337.