Research on Analysis and Improvement of Classification Algorithm of Data Mining Based on Neural Network

Yaze Yuan[1,], Mengmeng Su[2]*

1. Northwestern Polytechnical University, Xi'an, 710000, China

2 . Shanghai Zhizi Network Technology Co., Ltd., Shanghai, 200000, China

**Corresponding Author: Yaze Yuan*

Abstract: Data mining is an effective tool to dig out hidden information from the huge data, which is applied to every field of modern social life. Neural network is an important method of classification algorithm of data mining. This paper analyzes the basic process and common deficiencies of classification algorithm of data mining based on neural network, and gives the improved scheme to provide some references for the relevant researchers.

Keywords: Classification algorithm, Data mining, Neural network

1. INTRODUCTION

Data mining methods can usually be divided into two broad categories. One category is statistical method, and the common techniques include probability analysis, correlation analysis, cluster analysis and discriminant analysis. The other category is the machine learning in artificial intelligence. Many sample sets are trained to obtain patterns or parameters that need to be learned. As various methods have their own functional characteristics and applicable areas, the actual application process is usually combined with a variety of technologies to form complementary advantages. Neural network is one of the most commonly used data mining techniques. It was first proposed by psychologists and neuroscientists to seek the development and testing of neural computing simulations. It is like the method of repetitive learning in the human brain. First, a series of samples are given for learning and training to distinguish the different features and patterns among different samples. The sample set should be as representative as possible, and to accurately fit a variety of sample data, through hundreds of, or even thousands of times of training and learning, the system finally draws the potential model. When it encounters new sample data, the system automatically predicts and classifies the results based on training results. Neural network is composed of basic components of many simple interconnected neurons, simulating the human brain information. The biggest characteristic is that it is difficult to understand, that is, it cannot explain how to get the result and what rule to use. It takes a long training time, requires many parameters, and less explanatory. The advantage of this algorithm is that it can predict complex problems very well, and have higher tolerance to noise data and its ability to classify data without training. The neural network can be subdivided into feedforward, feedback and self-organizing neural networks, which has the functions of optimization calculation, clustering and prediction, and has been widely used in the commercial world. Financial markets use neural networks to establish credit card and currency trading models, which are used to identify credit customers, stock forecasts and securities market analysis.

2. BASIC PROCESS OF CLASSIFICATION ALGORITHM OF DATA MINING BASED ON NEURAL NETWORK

(1) Data Preparation

Data preparation is to define, process and represent the data being excavated so that it can be adapted to specific data mining methods. Data preparation is the first important step in the data mining process, and plays an important role in the whole data mining process. Data cleaning is filling the vacancy values in data, eliminating noise data and correcting inconsistent data in data. Data selection is to select data columns and rows for this mining. Data preprocessing is to enhance the processing of clean data after selection. Data representation is the form of data mining which can be transformed into neural network based data mining algorithms in acceptable form. Data mining based on neural networks can only deal with numeric data, so it is necessary to convert symbolic data into numerical data. The selection process of data is mainly carried out from two

perspectives. We want to select all the data needed in data mining, because the data provided to data mining will be many. On the other hand, we should select some columns in the database according to the features required by the data mining, as the columns in the database represent the attributes of the data. We further process the selected clean data, analyze the actual needs of the mining, merge or delete certain columns according to the actual needs of the selected data. In some cases, a combination of columns in a database can correspond to a requirement, or perhaps a field itself is redundant, and the field should be deleted. Because the neural network data training, the data format has certain requirements. Data can also be processed in groups, and it is generally to select arrays or vectors to allow data to be regularized in batches.

(2) Data Classification

We create a model that describes the categories or concepts of known data sets. For example, a model can be obtained from the analysis of the contents of the data rows in the database. Each row of data can be considered to belong to a determined data class, and its class value has a property description. The data set used in the classification learning method is called the training sample set, so classification learning can be called supervised learning. It builds the corresponding model by learning in the known training sample category, while unsupervised learning is conducted under the condition that the number of training samples is unknown and the number of classes is unknown. Usually, the model obtained by classification learning can be expressed as the form of classification rules, the form of decision tree or the form of mathematical formula. Given a customer credit information database, the classification rules obtained by learning can be used to identify whether a customer has a good credit rating or a general credit rating. Classification rules can also be used to identify unknown data. At the same time, it can also help users better understand the contents of the database. We study the mathematical formulation of classification using neural networks. It uses a set of samples with categories to perform classification tests. The accuracy of the model constructed for a given set of data can be obtained by the ratio of the number of test data that is correctly classified by the model to the total test sample. For each test sample, its known category is compared with the prediction category of the learned model. If the accuracy of the model is obtained by testing the learning data set, the learning model tends to approach the training data too much, which makes the estimation accuracy of the model too optimistic. Therefore, we need to use a test data set to test the accuracy of the acquired model.

(3) Data Prediction

Compared with the classification learning, the prediction can be considered as the class value of unknown class data rows and objects, and it is predicted by the model obtained by learning. The neural network has a good adaptability to the problem. A general forward network consists of an input layer and an output layer, with several hidden units. The hidden units can be layered or layered. If layered, it is called multilayer forward network. The input and output neurons of a network are usually linear functions, while implicit ones are nonlinear functions. An arbitrary forward network is not necessarily a hierarchical network or a fully connected network. The basic decision tree algorithms are memory resident algorithms that usually assume a small amount of data. It is very effective for relatively small data sets. When these algorithms are used for mining large and real-world databases, the effectiveness and scalability of these algorithms have become a concern. Most decision tree algorithms restrict the memory retention of training samples, and in the data mining applications, very large training samples containing tens of thousands of samples are very common. The restriction of the memory of the data set limits the scalability of these algorithms. If the training samples are swapped in and out of memory, the construction of the decision tree may become inefficient. The early strategy for large data to construct decision tree for discretization of continuous attributes, or sample of the massive data, so that only the external data traversed the whole one, while the learning algorithm only needs a stable data access memory. However, these still assume that the entire training data set used by the algorithm can be stored in memory. Although this method can be used for the prediction of large data sets, the accuracy of the prediction is not as good as that of using all the data at one time.

(4) Data Mining.

The correlation analysis is to discover the interdependencies between data or features. Data dependencies can represent a relatively important class of discovered knowledge, which can be used by other pattern extraction algorithms. A dependency relationship that usually exists between two elements. If the value of another element can be derived from the value of the element, we call the element dependent on the element. Data correlation analysis has been widely used in data mining, and the results of data correlation analysis can sometimes be provided directly to users. Collect the original data in the whole data mining process in the proportion of not less than other work less, and the original data collected by the researchers must fully to make the performance and results of data mining can meet our requirements. After gathering data, start sampling and cleanup. The result of sampling and cleaning is that data samples can be obtained, and data sets can be used for training and learning. To improve the

accuracy and efficiency of data mining, data mining must be done before necessary. At this point, the data form is still not up to our requirements. It needs to be converted to data. If the following result is not satisfactory, it needs to return to the previous stage and re sample the raw data. If the result is ideal, you can proceed to the next step. In the study of data storage, we find that data warehouse is a very effective form, and it is very conducive to data mining. After this process is over, it is time for various data mining algorithms to work. Using data analysis tools, we can obtain some valuable knowledge and information, and the content can be widely used in various other fields.

3.DEFICIENCIES OF CLASSIFICATION ALGORITHM OF DATA MINING BASED ON NEURAL NETWORK

Although the neural network algorithm has obvious advantages and is widely used, it also has some shortcomings. As the network structure is more complex, each learning process should dynamically modify the weight of each layer, so that the amount of calculation is quite complex. Therefore, the training speed will be very slow, and the training process is also difficult to master. When training reaches a certain level, the training effect is not obvious. The algorithm used is the steepest descent method, which is approximated downward along the slope of the error surface. When the network into the local minimum value, according to the general method is difficult to make the jump to continue to move forward, but we can choose the appropriate method to avoid getting into the local minimum value, such as the initial value and to modify the weights of training samples. When the network is in the process of learning, the weights may become large, and lead to a numerical input weighted by neurons will be great, so the value of the first derivative of this point will be very small, resulting in network learning step is also very small, network convergence. This phenomenon is network paralysis. Therefore, the choice of weights should be the minimum pseudo random number. At present, the selection of hidden layer node number is not a clear theory or method can be given, but the number of hidden nodes for performance and the convergence speed of the network is very important. Therefore, to choose a proper number of nodes is very important for the improvement of neural network.

4. IMPROVEMENT DIRECTIONS OF CLASSIFICATION ALGORITHM OF DATA MINING BASED ON NEURAL NETWORK

The traditional algorithm lacks the theoretical guidance for the number of hidden layer neurons. If the number of hidden layer nodes is not properly chosen, the results of the entire network will deviate greatly. In practical applications, the selection of the number of nodes in the hidden layer is important. If it is too small, the network learning process cannot achieve convergence. If it is too large, the entire network structure will be quite complex, and the amount of computation in the sample learning process will increase. This increases the mapping power of the network and reduces the local minimum. This allows the network to get the global optimal solution. Because of the complexity of the network structure, the learning time is greatly increased, and network training may be excessive. Although the network can remember the general characteristics of the sample after learning, it also remembers the individual features of the sample, such as noise data, which leads to a decrease in the overall network fault tolerance. Because the method avoids the local minimum point, it cannot guarantee to avoid other minimum points. Therefore, we must choose a more appropriate method, so that when the network reaches the minimum value, we can successfully jump out and continue to move forward and get the global optimal solution. In view of this deficiency of neural network, we propose an improved method to make the network dynamically increase the number of nodes in the hidden layer, so that the network has a better output. The algorithm first sets the number of nodes in the hidden layer to a relatively small value, and then according to the learning state of the network. If the actual error does not meet the set requirements, then the network will dynamically increase the number of hidden layer nodes, until the actual error of the network has not been significantly reduced. This method does not increase the number of nodes, because the number of nodes increases too much, which makes the whole network more complex. The training time of the network increases rapidly, and it even makes the network go into infinite cycle, resulting in paralysis of the network. We mainly use adding hidden layer nodes to make the network have more appropriate network model dynamically. The algorithm can effectively avoid the minimum output value of the network. The principle of the algorithm is to continuously improve the network accuracy, so the minimum value is avoided in nature. We verify it by experimental results. For the experiment, the number of nodes in the traditional algorithm with the layer is predetermined and cannot be changed. It has the limitations for further improvement of network accuracy. The improved algorithm can change the number of nodes in the hidden layer according to the output of the network, therefore, the accuracy of the network is greatly improved.

After a large number of repeated experiments, it can be found that the improved algorithm effect is much better than that of traditional algorithm, which is shown in Table 1. After comparison, we can see that

the convergence rate of the improved algorithm is much faster than before, mainly reflected on the number of iterations. The recognition rate of the improved algorithm is also much higher than that of the traditional algorithm.

Table 1. Experimental comparison between the traditional algorithm and the improved algorithm

	Traditional algorithm	Improved algorithm
Convergence speed	Slow	Fast
Whether there is a minimum value	Yes	No
Iterative number	About 300	About 100
Recognition rate	About 80%	About 95%

CONCLUSION

The neural network algorithm of data mining has some shortcomings, such as slow convergence rate and minimum value. In view of these deficiencies, this paper gives the corresponding solutions. The current algorithm is not perfect either. With the development of science and technology and the progress of society, more and more scholars will be involved in the research of the data mining algorithm to strive for better treatment methods.

REFERENCES

[1] Wang Chunmei. Research on data mining algorithm based on neural network [J]. Modern Electronics Technique, 2017, 40(11): 111-114.

[2] Wang Lei, Wang Ruliang. Data Mining Based on Improved BP Neural Network Method [J]. Journal of Guangxi Teachers Education University (Natural Science Edition), 2016, 33(1): 79-83.

[3] Ding Hao. Analysis and Comparison of Commo nClassification Algorithms in Data Mining [J]. Journal of Heze University, 2015, 37(5): 47-50.

[4] Hao Xiaoli, Zhang Jing. Design and Realization of RBF Neural Network Classifier Based on Advanced Self-adaptive Clustering Algorithm [J]. Computer Science, 2014, 41(6): 260-263.

.