Adaptive Data Mining Algorithm under the Massive Data

Weijian Mo

Xinhua College of Sun Yat-sen University

Guangzhou, 510520, China

E-mail: 472252720@qq.com

***Abstract*— In order to solve the problem that Network Reduced accuracy and poor convergence in the existing neural network, which because sample large volumes of data and target data-independent. In response to this phenomenon, this paper put forward a data mining based on compensatory fuzzy neural network. It was optimizing was the Compensative Fuzzy Neural Network. And improve the cutting effect base on calculation algorithm. At the end, it was based on the similarity of each cluster objects to clustering process the system data. Through simulation experiments we can see, algorithm can maintain high precision under different circumstances the amount of data. Compared to other algorithms, we can see that it has a large advantage in terms of both accuracy and time-consuming.**

Keywords—neural networks; data mining; clustering; rule extraction

I Introduction

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [1, 2]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Now has made a lot of data mining algorithms [3, 4, 5]. For example, Data mining algorithm based on fuzzy set, Data mining algorithm based on Clustering algorithm, Data mining algorithm based on neural networks, etc. While these algorithms able to extract useful information in large amounts of data, it will be affected data complexity. Neural network with its good advantage of parallel computing, distributed information storage, fault tolerance capability, with adaptive learning ability, etc.[6,7,8]. So it is favored by the majority of the research scholars. In practical application, neural network takes a long time in the training records and the properties of samples.

In response to these problems, this paper was research the traditional fuzzy neural network [9, 10]. We were using the excellent characteristics of the compensation fuzzy neural network [11, 12]. This paper is proposed a data mining algorithm based on compensation fuzzy neural network. Through improve the compensation fuzzy neural network to make the algorithm has better convergence. It is by establishing the error function to optimization of the computing system. Data using the clustering process based on the similarity of each cluster objects. At the end, corrected the result of the interspersed input.

II Compensatory Fuzzy Neural Network

Traditional fuzzy neural network use the computational methods is optimization fixed and partial, such as minimum and maximum operating [13, 14].

Neural network comes from the neurons structure theory of animals, bases on the M-P model and Hebb learning rule. So in essence it is a distributed matrix structure. Through training data mining, the neural network method gradually calculates (including repeated iteration) the weights the neural network connected. The neural network model can be broadly divided into the following four types. For example: Feed-forward networks, Feedback network, Self-organization networks, and Random neural network.

(1) Feed-forward networks: it has the representative areas, which the perception back-propagation model and the function network. And mainly used in the areas such as prediction and pattern recognition.

(2) Feedback network: it has the representative areas, which Hopfield discrete model and continuous model. And mainly used for associative memory and optimization calculation.

(3) Self-organization networks: it has the representative areas, which adaptive resonance theory (ART) model and Kohonen model. And mainly used for cluster analysis.

(4) Random neural network: it is a special kind of artificial neural network, which has better space for development. As a biological neural mathematical model, it has advantages of associative memory and image processing.

Artificial neural network has the characteristics of distributed information storage, parallel processing, information, and self-organization learning, and has the capability of rapid fitting the non-linear data. Configuration diagram used in this paper is shown below [15, 16].

First layer: it is input layer. Each component node directly connected to each neural input u_i, and passed it to the next layer.

Second layer: it is fuzzy layer. Each second layer node are represents a fuzzy variable value. To calculate Membership function of the fuzzy set in their respective linguistic variables μ_i^j .therein $i = 1,2,\cdots,n$, $j = 1,2,\cdots,m_i$, and n is a dimension, m_i is a fuzzy number of divisions.

Third layer: it is fuzzy Inference Layer. Each neural point represents a fuzzy inference rules, it can be matched the fuzzy rules and calculated fitness for each rule. Let's get the most membership function out of each rule. Therein $N_3 = m$。

Fourth layer: it is complement computing layer. Its role is to complement fuzzy calculation.

Fifth layer: conceptual level. It was using the normalization calculation to obtaining the output network.

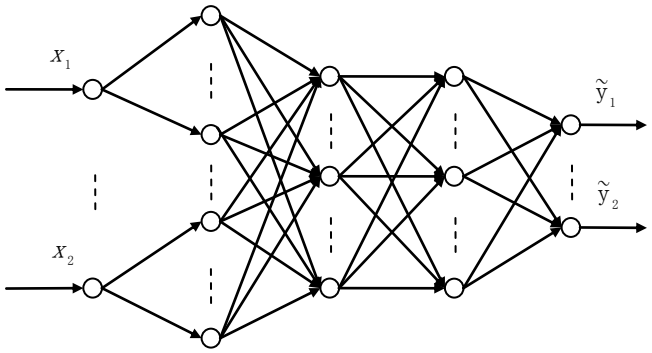

Figure 2. The figure of the network architecture

III Use compensation fuzzy neural network to data mining

A. Clipping algorithm

Network after training, It still will be some redundancy rights. So we are tailored to certain rules. It would be reduce the number of network weights. It is benefit to ensure the smooth progress of the clustering process. The procedure is as follows [17].

Step 1: Make the above error function F substituted into the clipping algorithm.

Step 2: Training network reaches a predetermined accuracy.

Step 3: Right collection of networks w_{ml}

When $\max | v_{pm} w_{ml} | \le 4F$, Then remove this weight w_{ml}.

Step 4: For the right collection of networks v_{pm}

When $| v_{pm} | \le 4F$, Then remove this weight v_{pm}

Step 5: Without satisfies step three and step four.

Then for each of the weights in the network, calculate that $w_{ml} = \max_p | v_{pm} w_{ml} |$, Delete the minimum value in the w_{ml}。

Step 6: Training network again.

If the network classification accuracy rate is below a predetermined value, this algorithm would be stop and use the original network weights. Otherwise, Jump back to step three.

In the above procedure, this paper introduces an error function *F*. Error function expression is:

$$F = -\sum_{i=1}^{k} \mu_{A_i^k}(x_i) \ln A_{A_i^k}(x_i) + (1 - \mu_{A_i^k}(x_i)) \ln(1 - A_{A_i^k}(x_i))$$

Therein $\mu_{A_i^k}(x_i)$ is a fuzzy expected input value, $A_{A_i^k}(x_i)$ is a fuzzy actual input value.

Through the network of the trimming process, it can effectively reduce the redundant network weights. Provide the basis for efficient data mining.

Optimization calculation and training

In practical application, we can see the data in the network is easy to interference by external factors. It must lead some deviation from the actual value of the original value. Therefore, In order to make the algorithm more reliable information, this paper is introducing an error function *F*. Error function expression is as follows [18, 19].

$$F = -\sum_{i=1}^{k} \mu_{A_i^k}(x_i) \ln A_{A_i^k}(x_i) + (1 - \mu_{A_i^k}(x_i)) \ln(1 - A_{A_i^k}(x_i))$$

Therein $\mu_{A_i^k}(x_i)$ is a fuzzy expected input value, $A_{A_i^k}(x_i)$ is the actual input values for the fuzzy.

In order for the error function can be applied to the crop of the network, we make network weights quickly become 0. In training, this paper was combining with the error function and penalty function p [20, 21].Function expression:

$$p = \frac{X}{2}(w_{ml}^2 + v_{pm}^2)$$

Therein, w_{ml} are expressed vague connection weights between the input layer l-th node and the hidden layer m-th node. v_{pm} are expressed vague connection weights between hidden layer m-th node and fuzzy output layer p-th node. X is the penalty function parameters, and X determining a symmetrical interval centered 0. L.e :

$$F = -p\sum_{i=1}^{k} \mu_{A_i^k}(x_i) \ln A_{A_i^k}(x_i) + (1 - \mu_{A_i^k}(x_i)) \ln(1 - A_{A_i^k}(x_i))$$

B. Optimal of the clustering detection

The basic idea of clustering algorithm is assumed that the data set of the object is relatively stable. In addition, a large number of data objects changes are smooth. If there is a sudden jump in the cluster, the abnormal data should be processed.

Cluster detection is calculated based on the similarity of different objects between different clusters of objects. According to the distance formula:

$$d(i,j) = \sqrt{| x_{i1} - y_{i1} |^2 + \cdots + | x_{im} - y_{im} |^2}$$

And satisfies:

$d(i,j) \ge 0$

$d(i, j) = d(j, i)$

$d(i, j) \le d(i, h) + d(h, j)$

Therein i, j is two dimensional data of the objects, h is a random object i to j between.

According to the formula (1) .

$$d(i, j) = \sqrt{diff^2(x_{i1} - y_{i1}) + \cdots + diff^2(x_{im} - y_{im})}$$

Therein x_{i1}, y_{i1}, x_{i2}, y_{i2}, $\cdots$, x_{im}, $diff(x, y)$ is a property of the distance.

Functional properties of computable functions for *snorm*: $X_{norm} = \dfrac{(X_i - \min_i)}{\max_i - \min_i}$

Continuous attribute function expression of the cluster centers as follows.

$$Mod(i) = \frac{\sum_{i=1}^{n} weight_i * value}{\sum_{i=1}^{n} weight_i}$$

Therein, $weight$ is a load data.

And satisfies that $i \ne j$, $R_i = \max_{j, j \ne i}\{R_{i,j}\}$, and calculates an average similarity function: $\overline{R} = \dfrac{1}{M_n}\sum_{i=1}^{M_n} R_i$

When $\overline{R}$ is a minimum value, it can be considered to the cluster is optimization.

This paper hope solve the problem that the classification results inconsistent for a sample at different input times. So we using the modified rules to correction it. First, judgments the size of the membership value in samples. If it was greater than the threshold, then it is qualified. Other cases, it instructions of the sample is not high degree of membership. We can use it as a network node number of wins. But, if we cannot found the winning node, then we should opening up new classes. And we have according to adaptive resonance theory, to correct for the winning node. Amended rules are as follows.

$V_{win} = (1 - \eta^2_{u_{win}})V_{win} + \eta^2_{u_{win}} x$

Therein η is a correction factor, and u_{win} A is a membership values of the winner node.

IV Simulation experiments

This selection of the network is operational, The network data signals 70000 regarded as experimental data. Wherein 60,000 data are sample data, and 10,000 data as test data network. The initial parameters of this algorithm are set as follows.

Assuming the neural network input neurons is 6，utput neurons are 6，wavelet hidden layer neurons set for 15, Learning rate η is set to 0.01，Wavelet translation factor and scale factor are randomly generated in the containing layers. Selection of data mining algorithm base on traditional fuzzy neural network, literature algorithm, using them to compare article algorithm. Using its to test the feasibility of the algorithm.The following graph circles is represent the traditional algorithm, Quadrilateral representative literature algorithm, Triangles represent article algorithm.

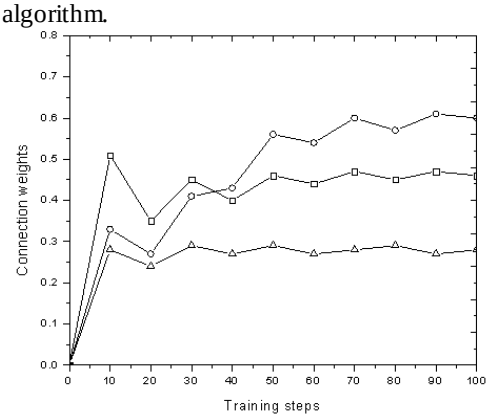

Figure 1. curve of the training steps and connection weights

By observing Figure 1, we can see that this algorithm curve is always at the bottom of other algorithms curve. Prove that it has a strong stability. Article algorithm is earlier than other algorithms into stable state about 40 steps. It is further proof of the convergence speed.

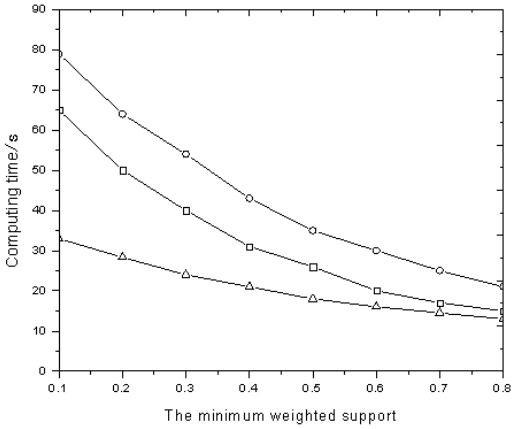

Figure 2. running time curve of transaction volumes

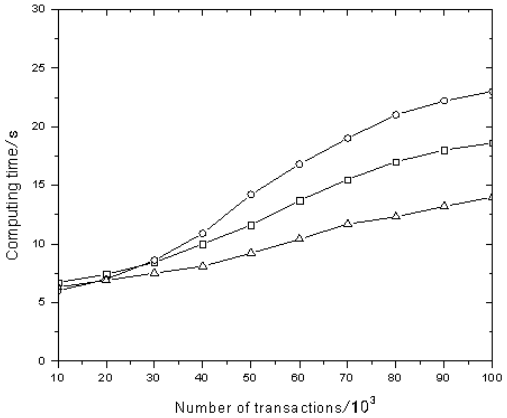

Figure 3. Running curve of the minimum support

Looking at figure 2 we can see that each algorithm gap is gradually narrowing when the increase of the minimum weighted support. The time required for this algorithm is

always less than other algorithms. Therein, the biggest difference with the traditional algorithm has 45s, and the maximum difference with the literature algorithm has 31s. As seen in Figure 3, with the gradual increase in the number of network transactions, the running time of each algorithm showing a linear growth trend. In the beginning, under the transaction is less, this algorithm is not dominant. But with the number of transactions increases, the advantage is gradually reflected of this algorithm.

To further verify the convergence of the algorithm, this paper was compared with the convergence error of each algorithm. Its data table is as follows.

Table 1. Data table of convergence error

Training steps	Convergence error		
	Traditional algorithms	Literature algorithm	Article algorithm
0	5.00	5.00	5.00
10	3.55	3.25	2.65
20	2.63	2.23	2.04
30	2.11	1.63	1.32
40	1.46	1.21	0.80
50	1.22	1.09	0.54
60	1.01	0.93	0.40
70	0.94	0.85	0.34
80	0.85	0.82	0.35
90	0.88	0.80	0.34
100	0.84	0.80	0.34

By observing Figures 4 and 5 can be seen that data of the algorithm with increasing the number of training steps is reduced gradually, finally kept at a lower value to maintain stability. When the training steps are less, we can see that convergence error of each algorithm is essentially the same. But with the increase of the number of steps, the gap are widening between each algorithms. At the end, article algorithm is stable in 60 steps, and other algorithms stable in 80 steps. In addition, error value of the article algorithm is less than the other algorithms in the case of steady-state.

V Summary

This paper presents an adaptive data mining algorithms. According to fuzzy compensation detection algorithm has good characteristics of the global and dynamic. This paper is optimization of computing systems and networks based on compensation fuzzy neural network. And combined with the similarity of each cluster objects to clustering process the data. To provide data relevant to the efficiency and quality of the algorithm. Laboratory results showed that the experimental results is basically consistent with the expected results. It has advantage are search speed and strong convergence.

REFERENCES

[1] Liu Jun, Yan Zheng, Yang Laurence T. Fusion - An aide to data mining in Internet of Things [J]. INFORMATION FUSION. 2015,23:1-2.

[2] Tsai Chieh-Yuan, Huang Sheng-Hsiang. A data mining approach to optimise shelf space allocation in consideration of customer purchase and moving behaviours [J]. INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH . 2015,53(3):850-866.

[3] Chemchem A, Drias H. From data mining to knowledge mining: application to intelligent agents [J]. Expert Systems with Applications. 2015,42(3): 1436-1445.

[4] Nadaban S. Fuzzy Euclidean Normed Spaces for Data Mining Applications [J]. INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL. 2015,10(1):70-77.

[5] Rezaei-Darzi E, Farzadfar F, Hashemi-Meshkini A, etc. Comparison of Two Data Mining Techniques in Labeling Diagnosis to Iranian Pharmacy Claim Dataset: Artificial Neural Network (ANN) Versus Decision Tree Model [J]. Archives of Iranian medicine. 2014,17(12): 837-843.

[6] Jajroudi M, Baniasadi T, Kamkar L, etc. Prediction of Survival in Thyroid Cancer Using Data Mining Technique [J]. TECHNOLOGY IN CANCER RESEARCH & TREATMENT. 2014,13(4): 353-359.

[7] Brito P. Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics [J]. WILEY INTERDISCIPLINARY REVIEWS-DATA MINING AND KNOWLEDGE DISCOVERY. 2014,4(4):281-295.

[8] Montano-Moreno J, Gervilla-Garcia E, Cajal-Blasco B. etc. Data mining classification techniques: an application to tobacco consumption in teenagers [J]. 2014,30(2): 633-641.

[9] Tahat A, Marti J, Khwaldeh A. Pattern recognition and data mining software based on artificial neural networks applied to proton transfer in aqueous environments [J]. CHINESE PHYSICS B. 2014,23(4).

[10] Mozafary V, Payvandy P. Application of data mining technique in predicting worsted spun yarn quality [J]. JOURNAL OF THE TEXTILE INSTITUTE. 2014,105(1):100-108.

[11] Samadianfard Saeed, Sattari Mohammad Taghi, Kisi Ozgur, etc. Determining Flow Friction Factor in Irrigation Pipes Using Data Mining and Artificial Intelligence Approaches [J]. APPLIED ARTIFICIAL INTELLIGENCE. 2014,28(8):793-813.

[12] Jajroudi M, Baniasadi T, Kamkar L, etc. Prediction of Survival in Thyroid Cancer Using Data Mining Technique [J].2014,13(4):353-359.

[13] Afify A. A. A novel algorithm for fuzzy rule induction in data mining [J]. PROCEEDINGS OF THE INSTITUTION OF MECHANICAL ENGINEERS PART C-JOURNAL OF MECHANICAL ENGINEERING SCIENCE. 2014,228(5): 877-895.

[14] Simeunovic V, Preradovic L. Using Data Mining to Predict Success in Studying [J]. CROATIAN JOURNAL OF EDUCATION-HRVATSKI CASOPIS ZA ODGOJ I OBRAZOVANJE. 2014,16(2): 491-523.

[15] Chi-Sen Li, Mu-Chen Chen. A data mining based approach for travel time prediction in freeway with non-recurrent congestion [J]. Neurocomputing. 2014,133:74-83.

[16] Wang Jing, Zhao Sheng-hui, Xie Xiang, etc. Mapping methods for output-based objective speech quality assessment using data mining [J]. JOURNAL OF CENTRAL SOUTH UNIVERSITY. 2014,21(5):1919-1926.

[17] Maragoudakis Manolis, Loukis Euripides. Heart sound screening in real-time assistive environments through MCMC Bayesian data mining [J]. UNIVERSAL ACCESS IN THE INFORMATION SOCIETY. 2014,13(1): 73-88.

[18] Wu Jia-Rui, Tang Shi-Huan, Guo Wei-Xian, etc. [Comment on applications of data mining used in studies of heritage of experiences of national medical masters][J]. China journal of Chinese materia medica . 2014,39(4):614-617.

[19] Govindarajan M. Ensembles of classification methods for data mining applications [J]. International Journal of Information Engineering and Electronic Business. 2013,5(6):6-21.

[20] Pate Pratik C, Singh Upasna. A novel classification model for data theft detection using advanced pattern mining [J]. DIGITAL INVESTIGATION. 2013,10(4):385-397.

[21] Lijuan Zhou, Yuyan Chen, Shuang Li. Improved data mining algorithms based on an early warning system of college students [J]. Journal of Software. 2013,8(9):2352-2359.