Send Orders for Reprints to reprints@benthamscience.ae

2268 *The Open Automation and Control Systems Journal*, 2015, *7*, 2268-2272

Open Access

Data Mining Model Based on the Improved BP Neural Network Algorithm

Shuolei Lin[1,*], Siriguleng Zheng[2] and Shumei Peng[1]

[1]*Beijing Polytechnic College, Basic Education Department, Beijing 100042, China*

[2]*Beijing Polytechnic College, Electrical and Information Engineering Department, Beijing 100042, China*

Abstract: A data mining algorithm based on fuzzy FKM algorithm and BP was proposed to solve the problem of long training time and low efficiency when the sample data contains the attributes unrelated to target data. The attributes of input data was clustered by using FKM clustering algorithm, and the attributes with weak correlation to target data were abandoned, and then there remain the attributes with strong correlation to target data, which reduce the training samples of neural network, and improving the training efficiency of the network. Tests on forecasting the content of Hemoglobin in the body of children show that the proposed algorithm is very practicable and reliable.

Keywords: FKM, BP, data mining, clustering.

1. INTRODUCTION

Data Mining Technology is a nontrivial method which is based on computer, like the new technology, to achieve effective and ultimately understandable patterns with the potential useful value. The application of data mining technology in the data processing can improve the efficiency of data processing and find the relation of the data. It will change the situation that data analysis rely more on their own intuition and experience to judge at present. Then, it will bring benefits for the enterprise, by finding the best solution for scientific research.

2. THE DATA MINING PROCESS

Data mining is a process of finding various summary models and export value from the known data collection. This process mainly includes the following steps:

(i) Data Collection and Cleansing. According to the problem to be solved, we need to obtain the data source and to determine the subset of data for data mining and establishing data mining library. If it satisfies the requirement of data mining, data warehouse can be set as data mining library.

(ii) Data Mining. Because the source data may be incomplete, noisy, random, we need to do the data cleaning. Then, we will select the variables related to data mining, or change the variable.

(iii) Pattern Evaluation. To choose an appropriate model according to the features and the purpose of mining.

(iv) The Interpretation and Evaluation Model. According to the requirements of end-user, we need to evaluate the information of data mining and to select the optimal model. If there are redundant or irrelevant patterns, they should be deleted. If the mode cannot meet the requirement of the user, we need to return to modeling in the previous phase.

3. THE NEURAL NETWORK PRINCIPLE

Neural network is a kind of mathematical model which simulates the behavior characteristics of animal neural network and is based on parallel distributed information processing.

The network depends on the complexity of the system to achieve the purpose of processing information by adjusting the internal relations of a large number of nodes connected. Neural network is proposed for solving complex problems and is a relatively simple method which is more effective. It can easily solve problems that have a hundreds of parameters or even more.

Neural network is often used in two types of problems: classification and regression. A neural network can be divided into input layer, output layer and hidden layer in structure.

Each node of the input layer corresponds to each predictor variable, while the node of the output layer corresponds to the target variable and is allowed to have multiple nodes. Input layer and output layer are both the hidden layers. The complexity of the neural network depends on the layer number and the number of nodes of the hidden layer. It's shown in Fig. (**1**).

4. THE IMPROVED FKM CLUSTING ALGORITHM

Classically, the clustering has two goals. One is that the similarity between objects of the same class is as large as possible, while the other is as small as possible between different classes. In the literature [1], the cost function only satisfies the first goal. Inspired by the literature [1], an improved FKM clustering algorithm is proposed in this paper.

*Address correspondence to these authors at the Beijing Polytechnic College, Basic Education Department, Beijing 100042, China;
Tel: +86-18607874465; E-mail: Shuolei@163.com

1874-4443/15

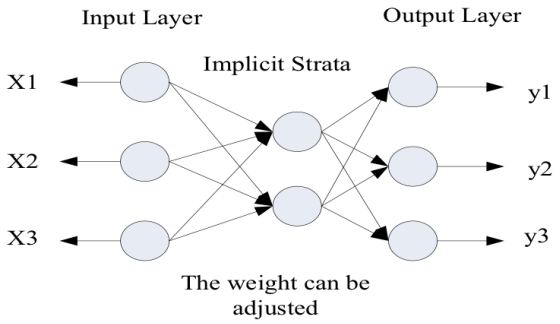

Fig. (1). Three layers BP network structure.

Give the cost function of the improved FKM clustering algorithm, as follows:

$$F(T,W,C)=\sum_{l=1}^{k}\left[\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}\tau_{lj}w_{li}(c_{li}-x_{ji})^2}{\sum_{i=1}^{m}(c_{li}-\overline{x}_i)^2}+\gamma\sum_{i=1}^{m}w_{li}\lg w_{li}\right] \quad (1)$$

At the same time, it will also satisfy the following constrains, as follows:

$$\begin{cases}\sum_{l=1}^{k}\tau_{lj}=1, 1\le j\le n, \tau_{lj}\in\{0,1\}\\ \sum_{i=1}^{m}w_{li}=1, 0\le w_{li}\le 1, 1\le l\le k\end{cases}$$

Where x is the mean of all objects, $\overline{x}_i$ is the ith dimension value of $\overline{x}$, that's, $\overline{x}_i\frac{1}{n}\sum_{j=1}^{n}x_{ij}$. The Eq. (1) only satisfies the situation, that's, $n>1$. Otherwise, when $n=1$, $\sum_{i=1}^{m}(c_{li}-\overline{x}_i)^2=0$, then the cost function can not be obtained.

Theorem 1 Remain T and C unchanged, if the weight meet the following equation, as follows:

$$w_{lt}=\frac{\exp(\frac{-\psi_{lt}}{\gamma})}{\sum_{i=1}^{m}\exp(\frac{-\psi_{lt}}{\gamma})} \quad (2)$$

$$\psi_{lt}=\frac{\sum_{j=1}^{n}\tau_{lj}(c_{lt}-x_{jt})^2}{\sum_{i=1}^{m}(c_{li}-\overline{x}_i)^2} \quad (3)$$

Then, the cost function F is minimum.

Prove.

The Lagrange method [1, 3] is adopted here, we can get the minimum loop problem with no restraint, as follows:

$$\min F(\{w_{lt}\},\{\xi_l\})=\sum_{l=1}^{k}\left[\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}\tau_{lj}w_{li}(c_{lt}-x_{jt})^2}{\sum_{i=1}^{m}(c_{li}-\overline{x}_i)^2}+\gamma\sum_{i=1}^{m}w_{lt}\lg w_{lt}\right]-\sum_{l=1}^{k}\zeta_l(\sum_{i=1}^{m}w_{lt}-1) \quad (4)$$

Where, $[\xi_1,\xi_2,...,\xi_n]$ is a non-negative vector. As it's independent between classes, the minimum of F can also be expressed as:

$$\min F(\{w_{li}\},\{\xi_l\})=\sum_{l=1}^{k}\min F(w_{li},\xi_l) \quad (5)$$

Where,
$$F(w_{li},\xi_l)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}\tau_{lj}w_{li}(c_{li}-x_{ij})^2}{\sum_{i=1}^{m}(c_{li}-\overline{x}_i)^2}+\gamma\sum_{i=1}^{m}w_{li}\lg w_{li}-\xi_l(\sum_{i=1}^{m}w_{li}-1) \quad (6)$$

As $F(w_{li},\xi_l)$ is continuously differentiable, when the derivative of $F(w_{li},\xi_l)$ is set to be zero, we can get the weights of each dimension, as follows:

$$\frac{\partial F(w_{li},\xi_l)}{\partial\xi_l}=\left(\sum_{i=1}^{m}w_{li}-1\right)=0 \quad (7)$$

$$\frac{\partial F(w_{li},\xi_l)}{\partial w_{lt}}=\frac{\sum_{j=1}^{n}\tau_{lj}(c_{lt}-x_{jt})^2}{\sum_{i=1}^{m}(c_{li}-\overline{x}_i)^2}+\gamma(1+\lg w_{lt})-\xi_l=0 \quad (8)$$

Eq. (9) can be obtained from the Eq. (8), as follows:

$$w_{lt}=\exp(\frac{-\psi_{lt}+\xi_l-\gamma}{\gamma}) \quad (9)$$

Where,

$$\psi_{lt}=\frac{\sum_{j=1}^{n}\tau_{lj}(c_{lt}-x_{jt})^2}{\sum_{i=1}^{m}(c_{li}-\overline{x}_i)^2}$$

Eq. (9) is substituted in Eq. (7), we can get Eq. (10), as follows:

$$\exp(\frac{\xi_l-\gamma}{\gamma})=\frac{1}{\sum_{i=1}^{m}\exp(\frac{-\psi_{lt}}{\gamma})} \tag{10}$$

Eq. (10) is substituted in Eq. (9), to get Eq. (11), as follows:

$$w_{lt}=\frac{\exp(\frac{-\psi_{lt}}{\gamma})}{\sum_{i=1}^{m}\exp(\frac{-\psi_{li}}{\gamma})} \tag{11}$$

Refer to the solution of w_{lt}, W and C remain unchanged, then there is

$$\tau_{lj}=\begin{cases}1, if\sum_{i=1}^{m}w_{li}(c_{li}-x_{ij})^2\le\sum_{i=1}^{m}w_{zi}(c_{zi}-x_{ij})^2\\0, others\end{cases} \tag{12}$$

Where, $\tau_{lj}=1$ expresses the object j belonging to the class l. Otherwise, it expresses that j doesn't belong to the class l.

T and W remain unchanged, C can be solved by the means of solving the average value in Mathematics, as follows:

$$c_{li}=\sum_{j=1}^{n}\tau_{lj}x_{ij}/\sum_{j=1}^{n}\tau_{lj} \tag{13}$$

Where, $1\le l\le k, 1\le i\le m$.

The process of solving the minimization of a cost function is shown in Fig. (**2**).

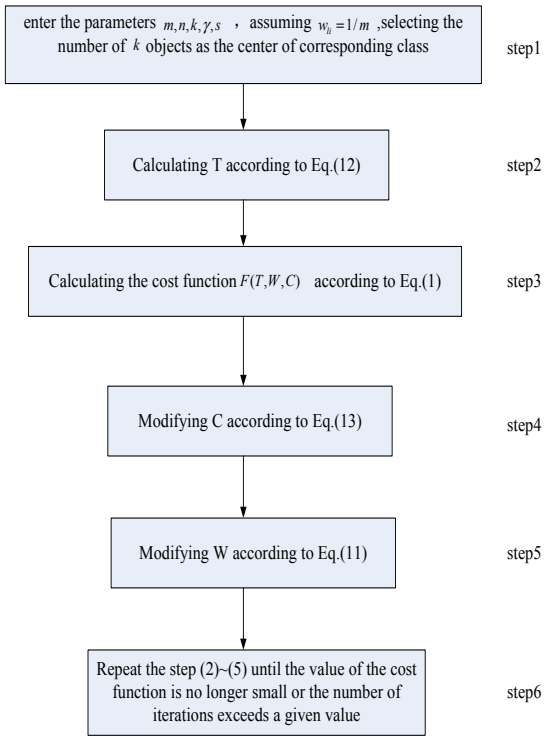

Fig. (2). The process of solving the minimization O A cost function.

5. THE DATA MINING ALGORITHM BASED ON FKM AND BP NEURAL NETWORK

After the samples are selected, start fuzzy clustering processing, we can keep the attributes strongly related to decision attributes by cutting those that are weak or redundant according to the specific circumstances. Using the filtered samples after fuzzy clustering as neural network's training samples, it can greatly reduce the network training time and improve the efficiency of network training. In literature [2-6], the algorithm's cost function is the extension of traditional k-means algorithm.

In this paper, the data mining algorithm based on FKM and BP neural network is adopted. Therefore, the structure of BP neural network with one storey of hidden layer is adopted here. For the activation function, it usually adopts sigmoid function in hidden layer, as follows [7, 8],

$$f^1=\frac{1}{1+e^{-x}} \tag{14}$$

The activation of output layer is usually due to linear function, as follows:

$$f^2=ax+b \tag{15}$$

The learning process of BP neural network used to adopt the standard BP neural network' process. After clustering in this paper, the output layer has only one neuron. The network output is calculated in the forward propagation during training.

When the network error satisfies the given requirement, the weight matrix and threshold of each layer can be saved and then the whole training is finished.

A data mining algorithm based FKM and BP neural network is adopted in this paper.

6. THE EXAMPLE ANALYSIS

Early data [9] was used as an example to test the fitting effect of the standard neural network and the neural network based on fuzzy clustering in this paper. Thus, we could judge which of them is good and bad. It's the determination results of 26 children's hemoglobin and trace elements in Table **1**. $\{a_1,a_2,a_3,a_4,a_5\}$ that make the set of condition attributes, while $\{a_6\}$ is the set of target attributes.

As shown in Fig. (**4**), first of all, we should normalize the sample data by eliminating the dimension between attributes. Then, clustering the attributes of the samples data. Finally, we will find that it only remains four condition attributes as follows $\{a_1,a_2,a_3,a_5\}$. Since the attribute a_4 is too weak to be the target attribute, so it can be ignored here. Here, we use the neural network algorithm tool to fit in Matlab. The results of using neural network directly are shown in Fig. (**3**) and Fig. (**4**), while the corresponding results are shown in Fig. (**5**) and Fig. (**6**) which used the algorithm based on FKM and BP neural network.

Table 1. The determination results of 29 children's hemoglobin and trace elements.

Numbered	Trace Elements					a6 HGB
	a1	a2	a3	a4	a5	/10-2g.mL-1
1	54.89	30.65	447.9	0.013	1.02	13.49
2	72.49	42.61	467.3	0.008	1.65	12.99
3	53.81	38.70	469.5	0.006	1.21	13.75
4	64.74	38.86	456.5	0.004	1.22	14.00
5	58.79	37.46	395.5	0.005	1.22	14.25
6	43.67	26.86	447.5	0.012	1.01	12.75
7	54.87	23.86	444.5	0.016	0.59	12.50
8	86.11	30.26	422.5	0.001	1.01	12.25
9	60.32	32.86	425.5	0.008	1.77	12.00
10	54.01	22.16	455.5	0.031	1.14	11.75
11	61.22	22.06	395.5	0.001	1.3	11.50
12	60.14	38.86	445.5	0.012	1.29	11.25
13	69.50	38.86	439.5	0.000	0.91	11.00
14	72.26	37.56	394.4	0.018	1.35	10.75
15	55.12	26.86	405.5	0.004	1.20	10.50
16	70.09	23.46	384.1	0.024	0.92	10.25
17	63.05	30.26	425.5	0.012	0.82	10.00
18	48.75	32.86	342.9	0.016	1.02	9.75
19	54.0	22.86	325.6	0.048	0.90	9.50
20	52.32	22.86	388.5	0.006	1.19	9.25
21	49.74	38.86	331.1	0.006	1.32	9.00
22	61.02	38.86	258.9	0.016	1.04	8.75
23	53.68	37.86	295.5	0.000	1.03	8.50
24	50.22	26.86	295.5	0.006	1.35	8.25
25	63.54	23.86	325.5	0.022	0.69	8.00
26	56.39	29.29	283.0	0.001	1.35	7.80

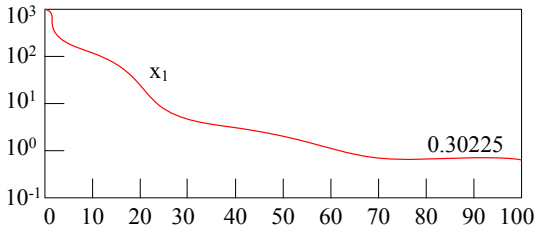

Fig. (3). The neural network training.

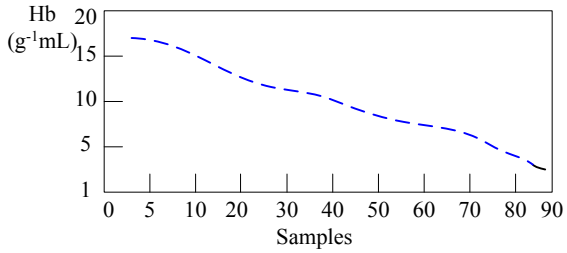

Fig. (4). The fitting effect of neural network.

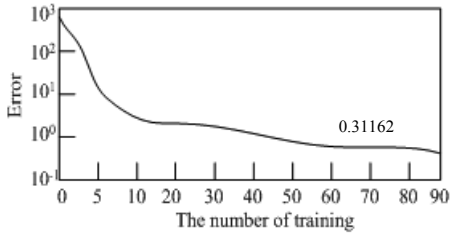

Fig. (5). The process of training based on fuzzy clustering.

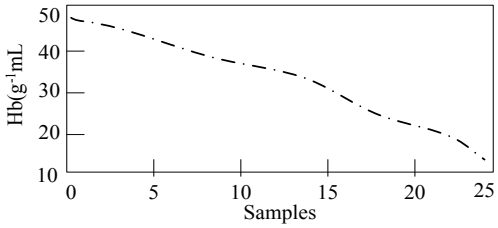

Fig. (6). The fitting effect of neural network based on fuzzy clustering.

The Fig. (**3**) shows the error curve of the training based on neural network; the value is 0.30225, and the training samples contain all attributes. While the figure 5 shows the error curve of the training based on the neural network after clustering; the value is 0.31162, and the training samples remove those attributes which are weak with the target attributes. We can find that the algorithm based on FKM and BP neural network reduce the number of training samples and the training time. Therefore, it can greatly improve the efficiency of the neural network training.

CONCLUSION

In the paper, the characteristic that fuzzy clustering algorithm can be used to cluster attributes is used to established the Data Mining method based on FKM and BP neural network. The clustering is used as the front-end processor in neural network and then the clustering algorithm is improved. The improved clustering algorithm overcomes the disadvantage which only satisfies the distance between classes as small as possible, while it also satisfies the distance between classes as large as possible. The improved algorithm is used to cluster the attributes of input data. It cuts the attributes which are weak with the goal and retains those attributes strongly related to the goal. After clustering, these data are used as the samples of neural network. Finally, we can find

that the algorithm can reduce the number of network training data and improve the rate of neural network training. Therefore, it can improve the training efficiency.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the Youth Elite Project Foundation of Beijing China under Grant No. YETP1781.

REFERENCES

[1] L. Jing, M.K. Ng, J.Z. Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data[J]". *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 8, pp. 1026-1041, 2007.

[2] J.Z. Huang, M.K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 1-12, 2005.

[3] H. Friguiand, and O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognition*, vol. 37, no.3, pp. 567-581, 2004.

[4] E.Y. Chan, W.K. Ching, M.K. Ng, and J.Z. Huang, "An optimization algorithm for clustering using weighted dissimilarity measures," *Pattern Recognition*, vol. 37, no. 5, pp. 943-952, 2004.

[5] C. Domeniconi, "*Locally Adaptive Techniques for Pattern Classification*," New York: George Mason University, 2002.

[6] C. Domeniconi, D. Papadopoulos, Gunopulosd, and S. Ma. "Subspace clustering of high dimensional data," In: *Proceedings of the Fourth SIAM International Conference on Data Mining. Florida*, USA: 2004, pp. 517-521.

[7] S. Yang, Y. Li, X. Hu. "Optimization study on k value of k-means algorithm," *Sys-terrns Engineering-Theory & Practice*, vol. 2, pp. 97-101, 2006.

[8] Q. Hu, S. Zhang, and D. Wang, "Application of neural network to project evaluation," *Journal of Northeastern University: Natural Science*, vol. 8, no. 2, pp. 169-171, 2007.

[9] Q. Song, and J. Shen, "Fuzzy clustering algorithms for web pages and customer segments," *Mini-micro Systems*, vol. 22, no. 2, pp. 229-232, 2001.

Received: February 21, 2015 Revised: April 29, 2015 Accepted: May 20, 2015