Data Mining Algorithm Based on Fuzzy Neural Network

Wu Jianhui [1], Su Yu [2], Shao Hongbo [3], Yin Sufeng [1], Xue Ling[4], Hu Bo[1] and Wang Guoli[*,1]

[1]*Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, Division of Epidemiology and Health Statistics, School of Public Health, Hebei United University, Tang Shan, China*

[2]*Department of Personnel, Hebei United University, Tang Shan, China*

[3]*College of Science, Agricultural University of Hebei, Baoding, China*

[4]*Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, School of Public Health, Hebei United University, Tang Shan, China*

Abstract: In this paper, the fuzzy neural network is selected as the algorithm for data mining (DM), introducing the artificial neural network into the fuzzy logic by treating it as a computing tool, it is a networklized description form by using the artificial neural network as the membership function in a fuzzy system, fuzzy rules and extension principle. The fuzzy neural network (FNN) is selected as the algorithm for data mining. By combining the fuzzy theory with neural network, using the strong nonlinear processing ability of FNN, finding the classification after producing a fuzzy partition neural network training, using thresholds and extracting rules, finally, the validity of this algorithm is verified, comparing with other fuzzy neural network, the neural network is faster learning speed and smaller in size. It owns has a good application prospect.

Keywords: Data mining, error back propagation, fuzzy neural network, fuzzy system, neural network.

1. INTRODUCTION

In this paper, the fuzzy neural network is selected as the algorithm for data mining (DM). The so-called data mining can be used for extracting unknown and useful knowledge through processing the data that is stored in a database system. With the rapid development of computer and communication technology, the data produced by human activities is increasing, thousands kinds of database are used for government affairs, business management, industrial production, scientific research and other aspects. The explosion of data makes it becoming an urgent need for people to find a way to collect data intelligently and process data automatically, so that the useful information can be collected. Hence, data mining becomes more and more important in this area. The obtained knowledge can be used for information management, problem solving, judgment, decision-making, production control and many other aspects [1, 2], there is no doubt that it will produce huge economic and social benefits.

There are many types of data mining algorithms, they mainly includes: the rough set, fuzzy set, clustering, association rules, decision trees, etc. [3, 4], however, the relationship of data in bulk data set is very complex, the nonlinear degree is quite high, besides the noise is commonly existing [5], so these methods are unsuitable on many occasions. By drawing on the neural network technology into data mining, with the aid of nonlinear processing ability and the usage of the noise capacity, this problem can be solved.

The classification of neural network is precise, and the robustness of neural network is good, so this method has certain advantages in solving the problem of data mining [6] Neural network as a classifier is widely used in machine learning. Although it needs long time, but the accuracy for classification is higher than the decision tree method [7]. On the other hand, knowledge is fuzzy sometimes, and the fuzzy knowledge plays a huge role in controlling and decision-making. The fuzzy processing method can be used in the knowledge discovery process to obtain the fuzzy knowledge. It is important to form a complete process by combining the fuzzy knowledge with fuzzy reasoning and fuzzy decision. Fuzzy logic and neural network are relevant and complementary in many ways. So combination of fuzzy logic system and neural network can improve the information processing to a new level. In such system, the topology of the neural network that simulates the brain's function is the "hardware". The fuzzy logic system that simulates the thinking ability of the fuzzy information processing is the "software". A fuzzy neural network is founded by the combination of fuzzy logic and neural network.

There are a lot of combination ways for artificial neural network and fuzzy system. In general there are two basic ways: the first way is introducing the artificial neural network into the fuzzy logic by treating it as a computing tool [8], specifically, it is a networklized description form by using the artificial neural network as the membership function in a fuzzy system, fuzzy rules and extension

principle; The second way is introducing nature principle of the fuzzy into the existing neural network structure [9] for instant: using the training and working process of the neuron as a sign of a fuzzy category, or drawing on the fuzzy theory into each neuron in the neural network to make the neurons conducting a computing operation for the various fuzzy algorithms; this way can also be used for the fuzzy pretreatment of the input data in neural network etc.

The research in this paper belongs to the first way; this paper proposes a data mining method that is based on the classification pattern of the fuzzy neural network (FNN). The method by combining fuzzy set theory and neural network together, improves the abilities of online adaption, rules expression, pattern classification, learning, and interpretation; according to the relationship between fuzzy set and pattern classification, this method firstly studies the structure and the learning algorithm of fuzzy neural network, then it carries on a training, finally a result of classification paten of customers' satisfaction for goods, which is based on the fuzzy neural network is obtained. The validity of the data mining method that is based on fuzzy neural network is further confirmed by the analysis and comparison of the actual situation. Comparing with other fuzzy neural network, the neural network is faster learning speed and smaller in size. It owns has a good application prospect.

2. THE FUZZY NEURAL NETWORK SYSTEM

Fuzzy system is a theory that is based on the fuzzy set, it processes the following treatment for a complicated issue: fuzzy measurement, fuzzy recognition, fuzzy reasoning and fuzzy control. In the classical theory, the characteristic function is only allowed to take two values of 0 and 1, and the fuzzy set adopts the concept of membership function, the characteristic function is extended to (0,1).

Fuzzy set: supposing $A:U \rightarrow [0,1], U \infty U(x)$ is a map on universe U, then A is the fuzzy set of U, $U(x)$ is the membership function of A. $U(x)$ reflects the membership degree of x on the fuzzy set A.

Fuzzy system is good at expressing and interpreting knowledge; it is a description of fuzziness based on precise mathematical language. The basic computing unit of neural network is an artificial neuron, which simulates the structure and function of neurons in the brain and has a strong self-learning ability. Through the continuous learning of adjustment, the artificial neuron can be finally presented by the specific structure of network.

Fuzzy neural network transform knowledge into the neural network by the usage of fuzzy model, then by using an algorithm to train neural network, the accuracy of the system mode is improved, the structure of fuzzy system's principle is shown in the Fig. (**1**).

3. THE DATA MINING BASED ON FUZZY NEURAL NETWORK

For a multiple attribute database, each properties of the database is continuous variable, if you want to extract the rules of the relationship between n properties and the other m properties, then the number of inputs is n which can be

expressed as $x_1, x_2, \cdots x_n$, and the corresponding m outputs, can be expressed as a $y_1, y_2, \cdots y_m$. In this kind of problem, by using the membership degree function, they can be translated into the membership degree value of their corresponding fuzzy inputs and fuzzy outputs. Using these as training samples, the following form is expected to obtain:

The steps of using fuzzy neural network system for mining are as follows.

3.1. The Standardization of Sample Data

This refers to the preprocessing of sample data, and characteristic values are mapped to the interval of (0,1). Supposing there are f samples $x_1, x_2, \cdots x_f$, each sample x_i has n sample indicators $z_1, z_2, \cdots, z_n$, x_{ij} refers to the number j indicator of the number i sample, then the n indicators of f samples are shown in Table **1**.

Table 1. Sample data.

Indicator	z_1	z_2	$\cdots$	z_n
x_1	x_{11}	x_{12}	$\cdots$	x_{1n}
x_2	x_{21}	x_{22}	$\cdots$	x_{2n}
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_f	x_{f1}	x_{f2}	$\cdots$	x_{fn}

The mean and standard deviation of the number j indicator in f samples are as followings:

$$x_j = \frac{1}{f} \sum_{i=1}^{f} x_{ij} \tag{1}$$

$$s_j = \sqrt{\sum_{i=1}^{f} (x_{ij} - x_j)^2} \tag{2}$$

The standardization of the original data is:

$$x_{ij}^{'} = (x_{ij} - x_j) / s_j \tag{3}$$

Compressing the standardize data within (0,1) by using the extreme value standardization formula, that is:

$$x_{ij} = \frac{x_{ij}^{'} - x_{j\min}^{'}}{x_{j\max}^{'} - x_{j\min}^{'}} \tag{4}$$

In the formula $x_{j\min}^{'}$ and $x_{j\max}^{'}$ are the smallest value and the biggest value in $x_{1j}^{'}, x_{2j}^{'}, \cdots x_{fj}^{'}$, x_{ij} is the standardized indicator.

3.2. The Structure of Fuzzy Neural Network

The structure of fuzzy neural network is shown in Fig. (**2**). The first layer is input layer, it can send the input value into the next layer, and the nodes' number in this layer is n.

The second layer is the membership function layer; the membership degreed of each input component on the corresponding linguistic variable fuzzy set can be calculated in the in this layer the membership degree function, each node stands for the linguistic variable value. The third layer is the fuzzy rule of matching. The application degree of each rule is calculated, the node number in this layer is m; the third layer and the fourth layer are fully connected. The data of the third layer is normalized in the fourth layer; the node number in this layer is m. The fifth layer is an output layer; in this layer the output data is obtained by weighted coefficient, and then through the membership function, the jurisdiction of the output under the fuzzy language variables can be determined.

For the input data, a membership function is selected as the normal distribution function:

$$u_{ij}(x) = e^{\frac{(c-x_{ij})^2}{\sigma_{ij}^2}} \tag{5}$$

$i = 1, 2, \cdots n$, n is the dimension of the input variables, $j = 1, 2, \cdots m$, m is the division number of fuzzy classification,, x_{ij} and σ_{ij} are the center and width of the membership function, respectively, i.e. translating the input value into the linguistic variable (big, medium, small, etc.) in the second layer.

a   layer is a input layer, which can translate $x = (x_1, x_2, \cdots x_n)$ into the next layer.

b   layer can calculate the membership degree function μ_{ij} of the jurisdiction of the input values under the fuzzy set.

Each node in *c* layer stands for a fuzzy rule. The application degree of each rule can be calculated, here is the minimized calculation:

$$\alpha_j = \min\{\mu_{1j}, \mu_{2j}, \cdots \mu_{nj}\}, \quad j = 1, 2, \cdots m \tag{6}$$

The normalized calculation is realized in *d* layer:

$$\alpha_j = \alpha_j \bigg/ \sum_{j=1}^{m} \alpha_j, \quad j = 1, 2, \cdots m \tag{7}$$

The sharpness calculation can be realized in *e* layer:

$$y_i = \sum_{j=1}^{m} \omega_{ij} \bar{\alpha}_j, \quad i = 1, 2, \cdots r \tag{8}$$

Besides the output data can be translated into the fuzzy linguistic variables through membership function. ω_{ij} is the connection weight of the $\bar{\alpha}_j$ in the fourth layer and the y_i in the fifth layer.

3.3. Network Learning

The learning process of fuzzy neural network is a multivariable optimization process. In the fuzzy neural network, the parameters that need to be learnt are the connection weight ω_{ij} in the fifth and the center value c_{ij} and width σ_{ij} of the fuzzy membership function.

The error function for the system is:

$$E = \frac{1}{2} \sum_{i=1}^{r} (Y_i - y_i)^2 \tag{9}$$

Y_i and y_i are the expected output and real output respectively, r is the number of output.

In the calculation from the input layer to the output layer is essentially a multilayer feed-forward back propagation network, by simulating the BP neural network and using the error back propagation method, the adjusting ω_{ij}、c_{ij} and σ_{ij} can be designed.

For each sample data, the first step is to calculate the output value of each node before the start of each input node, then starting from the output node using the back propagation algorithm, to calculate all the partial derivatives of nodes in the hidden layer.

Using the one-step-degree optimization algorithm for adjustment, finally learning algorithm of parameter adjustment can be obtained:

$$\omega_{ij}(k+1) = \omega_{ij}(k) - \beta \frac{\partial E}{\partial \omega_{ij}} \tag{10}$$

$$x_{ij}(k+1) = x_{ij}(k) - \beta \frac{\partial E}{\partial c_{ij}} \tag{11}$$

$$\sigma_{ij}(k+1) = \sigma_{ij}(k) - \beta \frac{\partial E}{\partial \sigma_{ij}} \tag{12}$$

$\beta > 0$ is the learning rate.

When system error reaches the given minimum error range, the learning of the fuzzy neural network is finished, which means the data mining can be used.

3.4. The Selection of Rules

After the success of the network training, the fuzzy rules are needed to summarize from the fuzzy neural network. According to the center value of the output values of membership function, and through the conversion of fuzzy language value of the membership function on the output value, the fuzzy language value, which has the largest membership degree, is selected as the output value. So, the rule extraction like can be like If $(x_1 = \upsilon_1)$ and $(x_2 = \upsilon_2)$ and $\cdots$ and $(x_n = \upsilon_n)$ then $(y = \zeta_i)$ can be conducted.

If there are n input variables, which are all converted into fuzzy language variables m, then $f = m \times n$ rules can be extracted from the to the n inputs, removing the same rules, there are still $mn - h$ rules. However, we are not interested in the $mn - h$ rules, in other words these rules are not all meaningful, we need to find the redundant rules, extracting
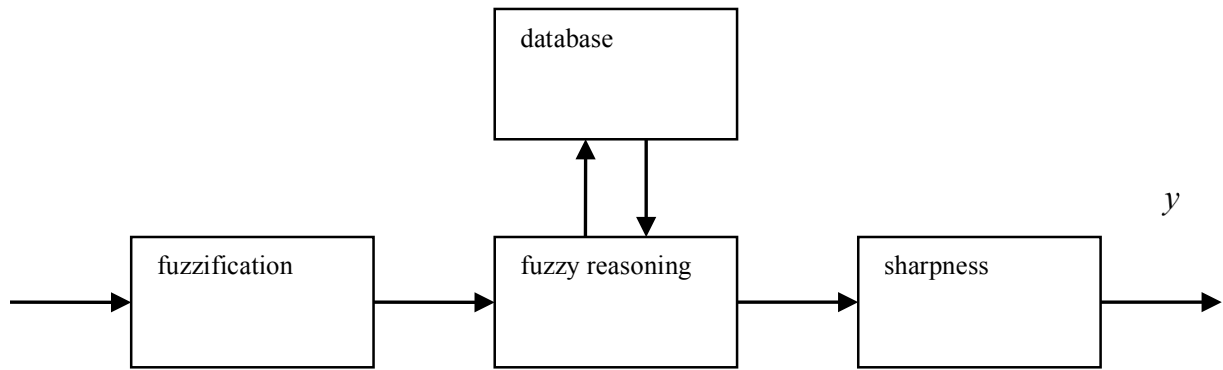
Fig. (1). Structure of fuzzy system.

rules that we are interested in. by set a threshold λ the rules that we need and of practical significance can be selected.

$$\lambda_k = \frac{n}{N}, \quad k = 1, 2, \cdots mn - h \qquad (13)$$

The rules which are smaller than the threshold can be used as redundant rules, so the rules that we are interested in and have practical significance can be selected from a mass rules. According to the different needs, the values can be different. Thus the number of the extracted rules is different. Based on the needs, the finally extracted rules are seen as the valid conclusions of data mining.

4. EXAMPLES

In order to verify the validity of the data mining algorithm, this paper collected 500 customer survey data (Table **2**) for some brand goods of a mall. By the usage of the data mining method, factors that cause different customers satisfactions on goods are hound and analyzed, and then a quantitative associated mining is conducted. For the correctness of the objectivity of data description and rules, in the process of investigation considering three state variables (price P, Q, quality after-sales service S) and an outcome variable (satisfied degree A) are taken into consideration.

According to the situation of the state of each attribute, the defined values are shown in Table **3**.

According to Tables **2** and **3**, the rule (P_3, Q_1, S_1, A_1) refers to: If $(x_1 = P_3)$ and $(x_2 = Q_1)$ and $(x_3 = S_1)$ then $(y = A_i)$, which means that customers will be satisfied with the goods of low price, good quality, and excellent service. So there are $4 \times 4 \times 3 \times 4 = 192$ rules in total. However in the real application, only a small part of these rules can be used.

According to experts' opinion, a series of initial values are confirmed at first. The membership function's center values c_{11}, c_{12}, c_{13} of price are set as 90%, 70% and 50% respectively; The membership function's center values of quality are set 80%, 60%, 30% respectively; he membership function's center values of service are set as 80% and 40%; and the width σ of the membership function are all set as 0.0001.

Table 2. The satisfied degree of customers towards goods with different attributes.

Price	Quality	Service	Accept
high	bad	-	Unacceptable
medium	medium	bad	Unacceptable
medium	excellent	good	very satisfied
-	excellent	good	Satisfied
medium	excellent	good	Satisfied
medium	medium	good	Acceptable
high	medium	good	Acceptable
high	bad	-	Unacceptable
low	medium	-	Acceptable
medium	excellent	good	Satisfied
low	excellent	bad	Acceptable
low	medium	-	Acceptable
high	excellent	good	Satisfied
high	medium	bad	unacceptable
medium	medium	good	Satisfied
high	medium	good	acceptable
low	excellent	good	very satisfied
high	medium	good	acceptable
low	medium	good	very satisfied
low	medium	-	acceptable
medium	medium	bad	unacceptable
low	medium	-	Acceptable
high	bad	bad	unacceptable
high	medium	bad	unacceptable
low	medium	-	Acceptable
low	excellent	bad	Acceptable
-	excellent	good	acceptable
medium	medium	good	satisfied
……	……	……	……

300 groups of data are selected from the 500 groups of data for training the network, the rest of the 200 groups of data are used for verifying the accuracy of the algorithm. First of all, standardizing the data, then inputting the standardized characteristic values into the fuzzy neural network, at last, using the error back propagation algorithm to adjust and correct c_{ij}, σ_{ij}, ω_1, ω_2 and ω_3. Training sample data after 615 iterations, network convergence, and error iteration steps shown in Fig. (**3**) is as follows.
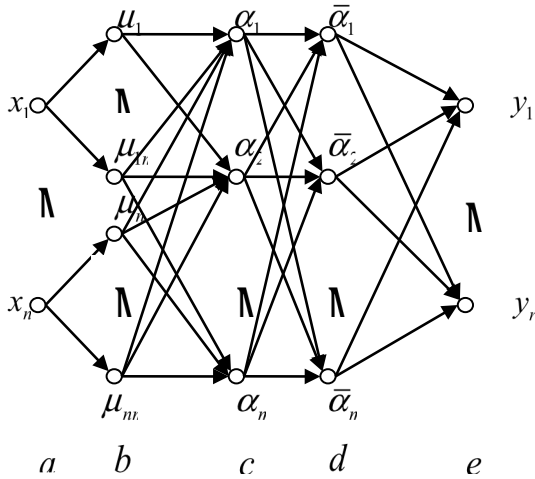

Fig. (2). Fuzzy neural network structure.

After the training, the revised ω_1, ω_2 and ω_3 are 0.544, 0.512 and 0.451, respectively. Setting 0.1 as the threshold λ of the ratio of the samples on the total samples, the following important rules are obtained as shown in Table **4** below.
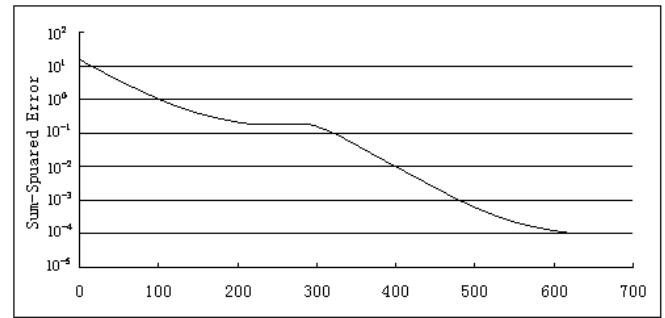

Fig. (3). The sum of squared errors of fuzzy neural network.

Explanations:

Rule 1: if the price is medium and the service is excellent then acceptable.

Rule 2: if the quality is good then acceptable.

Rule 3: if the service is bad then unacceptable.

Rule 4: if the quality is good and the service is excellent then satisfied.

Rule 5: if the quality is medium and the service is excellent then acceptable.

Rule 6: if the price is medium and the quality is medium then acceptable.

Rule 7: if the price is low then acceptable.

Rule 8: if the quality is medium and the service is bad then unacceptable.

Rule 9: if the quality is medium and the service is excellent then satisfied.

By comparing rule 4 with rule 9, it is can be known that

Table 3.   Attribute state of goods.

Price		Quality		Service		Accept	
State	**Value**	**State**	**Value**	**State**	**Value**	**State**	**Value**
high	P_1	excellent	Q_1	good	S_1	Very satisfied	A_1
medium	P_2	medium	Q_2	bad	S_2	satisfied	A_2
low	P_3	bad	Q_3	-		acceptable	A_3
-	P_0	-	Q_0	-	S_0	unacceptable	A_4

Table 4.   Extraction rules.

	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 7	Rule 8	Rule 9
P	P_2	P_0	P_0	P_0	P_0	P_2	P_3	P_0	P_0
Q	Q_0	Q_1	Q_0	Q_1	Q_2	Q_2	Q_0	Q_2	Q_2
S	S_1	S_0	S_2	S_1	S_1	S_0	S_0	S_2	S_1
A	A_3	A_3	A_4	A_2	A_3	A_3	A_3	A_3	A_2
Sample Rate	0.3735	0.3302	0.2500	0.2104	0.2084	0.1905	0.1652	0.1312	0.1105

if a shop want to raise the satisfaction of customers for products, on way is to provide good after-sales service and goods with the best quality, another way is to improve the service quality when the quality of the goods is in general conditions. Comparing rule 5 with rule 8 it is can be known that if a shop want to improve customer acceptance rate of goods with medium quality, one way is to provide a good after-sale service, for the improvement of the quality of products is not very easy. As wen can see that price does not affect very much, the main factors that influence the satisfaction are the quality and service, so as long as the quality of the goods and services can be improved, the acceptance rate of customers' can be increased.

Putting the other 200 data as a test set to verify the extracted rules, the accuracy of the test is 95.2%, then effectiveness of the rules can be confirmed.

Based the analyses above, the model can dig up the potential connection between commodity price, quality, after-sales service and customer satisfaction. It provides a powerful decision support tool for the analysis of a goods satisfaction, improvement of commodity's market share and business efficiency.

CONCLUSION

In this paper the fuzzy neural network (FNN) is selected as the algorithm for data mining. By using strong nonlinear processing ability of FNN, finding the classification after producing a fuzzy partition neural network training, using thresholds, extracting rules, and the sample verifying, three inputs one output fuzzy neural networks are formed, nine rules are extracted, and validity of the method is verified. This model also can be used in marketing, finance, credit risk analysis, logistics supply analysis, traffic accident analysis, mine accident analysis, and other areas, it is valuable in research and application.

CONFLICT OF INTEREST

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

REFERENCES

[1]   U.R. Acharya, O. Faust, S.V. Sree, F. Molinari, R. Garberoglio, and J.S. Suri, "Cost-effective and non-invasive automated benign & malignant thyroid lesion classification in 3d contrast-enhanced ultrasound using combination of wavelets and textures: a class of thyro scan?," *Algorithms Technology on Cancer Research and Treatment*, vol. 10, no. 4, pp. 371-380.

[2]   M. Gangwar, R.B. Mishra, and R.S. Yadav, "Application of data mining models in the diagnosis of neuropsychiatric diseases," *International Journal of Data Mining And Emerging Technologies*, vol. 4, no. 2, pp. 72-82, 2014.

[3]   B. Laurent, R. Gilles, B. Mohammed, "Semi-physical neural modeling for linear signal restoration," *Neural Networks*, vol. 38, pp. 90-101, 2013.

[4]   Y.Y. Lin, J.Y. Chang, and C.T. Lin, "Identification and prediction of dynamic systems using an interactively recurrent self-evolving fuzzy neural network," In: *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 2, pp. 310-321, 2013.

[5]   F.J. Lin, Y.C. Hung, and J.C. Hwang, "Fault-tolerant control of a six-phase motor drive system using a takagi–sugeno–kang type fuzzy neural network with asymmetric membership function," In: *IEEE Transactions on Power Electronics*, vol. 28, no 7, pp. 3557-3572, 2013.

[6]   E. McIlvaine, A. Borzabadi-Farahani, C.J. Lane, S.P. Azen, and S.L. Yen, "Apriori feasibility testing of randomized clinical trial design in patients with cleft deformities and Class III malocclusion," *International Journal of Pediatric and Otorhinolaryngology,* vol. 78, no. 5, pp. 725-730, 2014.

[7]   S. Pushpinder, "Some new distance measures for type-2 fuzzy sets and distance measure based ranking for group decision making problems," *Frontiers of Computer Science*, vol. 8, no. 5, pp. 741-752, 2014.

[8]   E. Spyropoulou, T.D. Bie, and M. Boley, "Interesting pattern mining in multi-relational data," *Data Mining and Knowledge Discovery,* vol. 28, no. 3, pp. 808-849, 2013.

[9]   Z.W. Xue Ni, "Research of large scale manifold learning based on MapReduce," *Systems Engineering-Theory & Practice*, vol. 34, no. s1, pp. 151-157, 2014.