International Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011
1793-8201

An Improved Prediction Model based on Fuzzy-rough Set Neural Network

Jing Hong

Abstract—**After the brief review of the basic principles and characteristics of BP (back-propagation) neural network and rough set theory, a novel artificial neural network model based on fuzzy-rough set is proposed in this paper, which is suitable for nonlinear regression to achieve the precise prediction result by introducing improved rough set to obtain minimal attribute set to solve the optimized problem of input layer. Compared to the normal prediction model, the feasibility and effectiveness of the proposed model is verified by a practical case study of the mine gas emission prediction.**

Index Terms—fuzzy-rough set, neural network, prediction model

I. Introduction

Inductive learning for rules generation is an important research area in artificial intelligence. Recently, many techniques have been developed to perform inductive learning [1,2,3,4,5]. Among them rough set theory is a new tool to solve vague and uncertain data analysis which was put forward by Polish scholar Z. Pawlark based on the thought of borderline of G.Frege[6],and which to some extent overcomes the limitations of fuzzy sets theory and Dempster Shafer theory while solving uncertainty problems. The basic idea of the theory is to classify the objects of interest into similarity classes (equivalent classes) containing indiscernible objects via the analysis of attribute dependency and attribute reduction [1,2]. The rule induction from the original data model is data-driven without any additional assumptions, namely which requires no external parameters and uses only the information presented in the given data, and enriches the multiple dimension data in the direction of attribute and case, and consistently or inconsistently reduces to find hidden law of the data to be treated. It is widely used in medical diagnosis, pattern recognition, machine learning, expert system and etc.

In practice, however, there are some limitations with the classical rough set. The original rough set can deal with the discrete attributes efficiently, but it cannot deal with the continuous attributes well. For continuous attributes, the traditional decision table is normally transformed into binary table by partitioning the attribute value into several intervals subjectively. However, there is no crisp boundary between the neighboring concepts. On the other hand, the original rough set is based on the indiscernibility relation. The universe is classified into a set of equivalent classes with the indiscernibility relation. The lower and upper approximations are generated in terms of the equivalent classes. So the original rough set classifies the knowledge too fussily, which leads to the complexity of the problem. The fuzzy set theory and rough set theory are generalizations of classical set theory for modeling vagueness and uncertainty [1]. Pawlak proposed that the two theories were not competitive but complementary [7]. Dubois and Prade also proposed that they were related but distinct and complementary theories [8]. Both of the theories model different types of uncertainty. The rough set theory takes into consideration the indiscernibility between objects, whereas the fuzzy set theory deals with the ill-definition of the boundary of a class through the membership functions. Hence, it is possible to combine the two theories to enhance the reasoning power of an intelligent system.

BP network is one of the most widely applied artificial networks [9,10], which is a feed-forward multi-layer mapping net and has powerful mapping capability for nonlinear problems. It converts the problem of input and output of a set of samples to nonlinear optimization, making use of the most common method of optimization namely gradient descent method, iteratively computing net connection weights which correspond to learning and memory problem, and introduces hidden nodes to add adjustable parameter to get more precise solution. So it has incomparable superiority such as arbitrarily approaching a none-linear function and parallel reasoning, and has been widely used in pattern recognition, system identification, prediction, control, image processing, function fitness and etc. However, the prediction model based on normal BP network which applies the general factors as input nodes may result in over-large scale and impair the convergency and accuracy, and very difficult to predict timely and precisely. It is necessary to explore a novel method to effectively solve the problem input node selection of classical neural network input layer which is generally subjective and random.,

In this paper, an improved prediction mothod based on fuzzy-rough set neural network model is presented, which introduces the fuzzy-rough set theory as a complementary tool of BP network to construct the model by using minimal attribute set to solve the optimized problem of input layer. In section 2, the fuzzy-rough set neural network model is discussed in detail. Section 3 describes one practical case study of the mine gas emission to illustrate how to apply the proposed model and demonstrate its effectiveness and advantages over other normal prediction models. Finally, conclusions are drawn in section 4.

II. Fuzzy-Rough Set Neural Network Model

A. BP neural network model

In 1989, Robert Hecht –Nielson proved that any continuous functions on a close interval can be approximation by BP network with a hidden layer, that is a three-layer BP network can complete the mapping from and n dimension to m dimension space[11]. The model has incomparable superiority in machine reasoning, capacity of fault tolerance, parallel reasoning, arbitrarily approaching a none-liner function etc. The number of hidden layer, the number of nodes of each hidden layer, weights, learning rate, momentum constant and reecho test should be considered when constructing a model. According to some experiences, the principles of setting parameters are respectively shown as follows.

1) Number of hidden layer

Increasing the number of hidden layers can reduce the error rate effectively, but it also brings about some large problems, such as over-large scale network, longer training time so as to impair the system's practicality. In fact, the precision can also be increased by increasing the number of nodes of hidden layer, which is more adjustable than others. The BP network in this paper has three layers.

2) Number of the nodes in hidden layer

To determine the number of the nodes in hidden layer is a complicated problem, which currently does not have analytical expression to use. With the BP network used as function approximation, the number of the nodes of the hidden layer relates to the precision of function approximation and the fluctuation degree of the function. Generally speaking, if the number of nodes in hidden layer is too large, the learning time may be too long and the scale is too over-lager. On the other hand, if the number of nodes in hidden layer is too small, the trained network would not be robust enough to recognize the samples which do not learn before, and has poor performance of fault tolerance. In practice, people reference some experience formula But sometimes, formulas were not proper for some problems. We compare several result of different node number of hidden layer, and add little nodes as initialization, then optimize the model with stepwise regression method. Some simulations prove it feasible.

3) Connection weight

Before the training of a BP network have to determine initial connection weight of the net. Due to the nonlinearity of the system, there are normally several local minimums on error curve. $\delta \to f'(s)$ in standard BP. When the slop of error curve is small ($f'(s) \to 0$, then $\delta \to 0$), the connection weight can not be tuned effectively. In this research, we apply additional momentum to improve standard BP model. The modified formula is $\Delta W_{ji}(t+1) = (1-\alpha)\eta\delta_{pj}O_{pj} + \alpha\Delta W_{ji}(t)$, which takes the influence of last weight transformation by momentum factor into consideration, and adjusting α based on the rate error change in the mean time. The formulas to adjust α are shown as follows;

$$\alpha = \begin{cases} 0.00 & SSE(k) > SSE(k-1) \\ 0.95 & SSE(k) \le SSE(k-1) \end{cases} \quad (1)$$

Where *SSE(k)* is the rate of the *k*th iterative absolute error change.

4) Learning rate and momentum

Learning rate and momentum constant are two important parameters in the training of a BP network, and each time the degree of BP network modifying the net connection weight mainly lies on these parameter. The quantity of weight transformation depends on the learning rate greatly. If the learning rate is too large, the system has poor performance on stability. On the other hand, if the learning rate is too small, the learning time may be too long. Different error curve matching proper learning rate can improve system performance. We adjust the learning rate by the following formula 2 in this research. Simulations prove that the network is convergent very quickly.

$$\eta(t+1) = \eta(t)E(t-1)/E(t) \quad (2)$$

Where $E = \sum_p E^{(p)} / N$, N is capacity of learning sample, E is total average error.

Momentum constant has the function of smoothing learning procedure, which is an increased momentum ratio with arithmetic improvement. Increasing learning rate should properly reduce the value of momentum constant to avoid sharp wave crest or trough.

5) Reecho test and prediction test

Reecho test checks out whether the actual output of the net satisfies the requirements by coefficient of correlation R. R>0.9 usually means that the trained net is acceptable, and the more R approaches 1, the better the net fits the data. Prediction test makes the simulation of the trained BP network with the data of the test set and computes the relative error, and is the little the relative error, the stronger is the ability of generalization of the network. By reecho test and prediction test check the precision and generalization of BP network whether or not to meet the precision requirement. The formula of coefficient of correlation is show as follow,

$$R = \sqrt{1 - \frac{\Sigma(F-E)^2}{\Sigma(E-Y)^2}} \quad (3)$$

Where: R is coefficient of correlation, F is the actual output of the net, E is the expectation of output of the net, Y is mean of the expectation of output of the net.

B. Fuzzy-rough set model

In this section, the fuzzy-rough set model is discussed in detail, including fuzzifying the continuous (numerical) attributes, new definitions based on fuzzy similarity relation and attribute reduction based on the fuzzy similarity relation.

1) Fuzzifying the continuous attributes

A decision table with 4-tuple can be represented as T=<U，C∪D，V，f>, where U is the universe, C and D are sets of condition and decision attributes respectively, V is the value set of the attribute a in A, and f is an information function.

Practically, there are many continuous or numerical attributes in the decision table, such as salary and experience. These attributes need to be fuzzified into linguistic terms, such as high, average and low. In other words, each attribute a is fuzzified into k linguistic values Ti，∀i=1,…,k. The

membership function of Ti can be subjectively assigned or transferred from numerical values by a membership function. The popular triangular membership function is shown in Fig. 1, where μ(x) is membership value and x is attribute value.

The slop of triangular membership functions are selected in the way that adjacent membership functions cross at the membership value 0.5, so the only parameters need to be determined are the set of k centers M={mi, i=1,2,…,k}. The center mi can be calculated through Kohonen's feature-map algorithm [12].

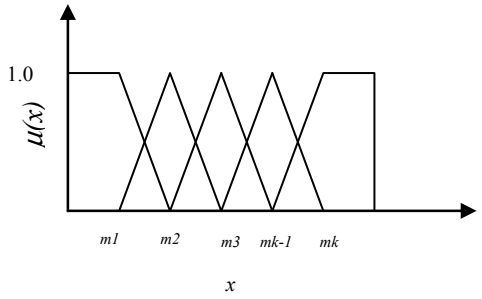

Figure 1 Triangular membership function

2) New definitions based on fuzzy similarity relation

The classical lower and upper approximations are originally introduced with reference to an indiscernibility relation (reflexive, symmetric, and transitive). Practically, it can be extended to fuzzy similarity relation.

Definition 1: Considering U={u1, u2, …, un} is the universe, the fuzzy similarity relation $\widetilde{R} \in \Re_{n\times n}$ on U is called fuzzy similarity matrix, if each element r_{ij}.∈ $\widetilde{R}$ has the two following properties:

Reflexive: $r_{ii} = 1, \forall i = 1,2,\cdots,n;$

Symmetric: $r_{ij} = r_{ji}, \forall i, j = 1,2,...,n.$

In order to constructing the fuzzy similarity relation, the measurement of fuzzy similarity relation should be introduced first, namely the method to calculate the factor r_{ij}. Generally, the max-min method, relational factor method and Minkowski distance based closeness degree method are used.

Definition 2: Considering $\widetilde{R}$ is a fuzzy similarity matrix and λ is the level value, the matrix $\widetilde{R}$ λ is called normal similarity relation matrix with the level value λ after the following operation.

$$\begin{cases} r_{ij} = 1, & r_{ij} \geq \lambda; \\ r_{ij} = 0, & r_{ij} < \lambda. \end{cases} \quad i, j=1,2,\ldots,n$$

Obviously, matrix $\widetilde{R}$ λ has the reflexive property and symmetric property. In order to obtain the classification of U given the fuzzy similarity relation, an algorithm is designed as follows.

Algorithm 1:

Input: fuzzy similarity matrix $\widetilde{R}$ and level value λ

Output: U/IND($\widetilde{R}$ λ), which is a partition of U given fuzzy similarity relation $\widetilde{R}$ and level value λ

Step 1 Calculate normal similarity relation matrix $\widetilde{R}$ λ in terms of definition 2;

Step 2 xi∈ U, X⇐ϕ, Y⇐ϕ;

Step 3 j⇐0;

Step 4 If rij=1 and xj∉ X, then X⇐X∪{xj}, Y⇐Y∪{xj} ;

Step 5 j⇐j+1;

Step 6 If j<n, then GOTO Step 4; otherwise, GOTO next step;

Step 7 If card(Y)>1, then select xi∈ Y and Y⇐Y－{xi}, GOTO Step 3; otherwise, GOTO next step;

Step 8 Output the set X and let U⇐U－X;

Step 9 If U=ϕ, then end; otherwise, GOTO Step 2.

According to the algorithm, U/IND($\widetilde{R}^{\{a_i\}}_{\{\lambda_i\}}$), the classification given the attribute a_i∈ A with the level value λ_i, is calculated. The classification of U given the attribute set A with the level value set λ can be defined as follows,

U/IND($\widetilde{R}^{A}_{\lambda}$) =⊗{U/IND($\widetilde{R}^{\{a_i\}}_{\{\lambda_i\}}$): ai ∈ A, λi∈ λ } (4)

Where A and λ are the attribute set and the level value set, respectively, and operator ⊗ is defined as follows,

A⊗B={ $X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \Phi$} (5)

Definition 3: Considering a subset X⊆U and a fuzzy similarity relation $\widetilde{R}^{A}_{\lambda}$ defined on U, the lower approximation of X, denoted by $\widetilde{R}^{A}_{\lambda\_}$（X）, and upper approximation of X, denoted by $\widetilde{R}^{A^-}_{\lambda}$（X）, are respectively defined as follows,

$\widetilde{R}^{A}_{\lambda\_}$（X）=∪{Y:Y∈ U/IND($\widetilde{R}^{A}_{\lambda}$),Y⊆X); (6)

$\widetilde{R}^{A^-}_{\lambda}$（X）=∪{Y:Y∈ U/IND($\widetilde{R}^{A}_{\lambda}$),Y∩X≠∅).(7)

Definition 4: Assuming U/IND($\widetilde{R}^{C}_{\lambda}$) and Y are two partitions on U, where U/IND($\widetilde{R}^{C}_{\lambda}$) = {X1, X2,…,Xk} and Y={Y1,Y2,…, Yr}, the positive region $POS^{\lambda}_{C}(Y)$ is defined as follows,

$$POS^{\lambda}_{C}(Y) = \cup\left\{\widetilde{R}^{C}_{\lambda\_}(Y_i) : Y_i \in Y\right\} \quad .(8)$$

3) Attribute reduction fuzzy based on similarity relation

In decision system, reduction is the dependence and association of the decision attributes on condition attributes. In practice, there is a lot of redundant information in the original data source. Attribute reduction can remove the redundant or noise information successfully. In the attribute reduction, the attribute reduction set is not single. The cardinality of reduction set determines the dimensionality of problem, so it is important to select a minimal reduction. The minimal reduction can be defined as follows,

Definition 5: Considering T=<U,C∪D,V, f> is a decision table, the classification of U with C and D are respectively denoted as $X=\{X_1,X_2,\ldots,X_m\}$and $Y=\{Y_1,Y_2,\ldots,Y_n\}$. The conditional entropy of C with D is defined as follows,

$H(D \mid C) = -\sum^{m} P(X_i) \sum^{n} P(Y_j \mid X_i) \log_2 P(Y_j \mid X_i)$ Wher e $p(X_i) = card(X_i) / card(U), i = 1,2,...,m$ .Definition 6: Considering T=<U,C∪D,V, f> is a decision table, $R \subseteq C$, the significance for each attribute a∈ C-R is defined as follows:

$$SIG\ (a,R,D)=\omega_1(R\cup\{a\},D)\cdot(\gamma(R\cup\{a\},D)-\gamma(R,D))$$
$$+\omega_2(R\cup\{a\},D)\cdot(H(D\mid R)-H(D\mid R\cup\{a\})/\log_2 n)$$
(9)

Where $\omega_1(R\cup\{a\},D)=card(POS_{R\cup\{a\}}(D))/card(U)$; $\omega_2(R\cup\{a\},D)=1-\omega_1(R\cup\{a\},D)$.

In order to obtain the minimal condition attribute set, an algorithm is designed as follows,

Algorithm 2:

Input: decision table

Output: the minimal attribute set R;

Step1 classify condition attributer set C with fuzzy level value;

Step 2 calculate *Core*(C,D) by discernible matrix and let $R\Leftarrow Core(C,D)$;

Step 3 calculate SIG（a,R,D）for each attributer;

Step 4 select attributer a∈C－R with maximum SIG（a,R,D）and let R⇐R∪{x};

Step 5 if r(C,D)==r(R,D) ,then GOTO step 6;. otherwise, GOTO step 3;

Step 6 return R.

Obviously, the computational complexity of the algorithm is O(m2), where m is the number of condition attribute set in decision table. In terms of the algorithm implementation, attribute reduction can be treated as a tree traversal. Each node of the tree represents the condition attribute. Calculating the minimal reduction can be transformed to picking the best path based on some heuristic information. On the other hand, operator ⊗ can reduce the computation by using results from the previous levels.

4) Prediction procedure based on fuzzy-rough set neural network

Step 1 select proper fuzzy level value, then fuzzy classify the original data in terms of the algorithm 1;

Step 2 reduce the condition attributers in terms of the algorithm 2;

Step 3 select minimal attributers as input nodes of neural network, then set proper hidden layer and construct model;

Step 4 train and test model;

Step 5 if the error meets limits, then GOTO end; otherwise, GOTO step 1.

III. Experiment Result And Analysis

Recently, Artificial neural network has been applied in mine gas emission prediction [14,15].In this section, we use this practical engineering case to illustrate the prediction process and discuss the effectiveness of the proposed method.

A. Experiment result

With the research of the gas emission prediction of a coal mine[15], the nonlinear fitness of gas emission data including 18 cases. According to geographical structure and exploitation conditions of the coal mine condition determine the components of the input vector are depth, gas content of coal seam, space between coal seam, the daily output of working face, thickness of coal seam, and circulatory style, and the output vector is absolute gas emission. We select eighteen months' samples to evaluate the constructed model performance, among which the first sixteen months' samples are used for training, and the seventeenth month and eighteenth month samples are used for testing. On account of each sample dimensionality changes largely, the sample data should be normalized.

The prediction model is realized in VC6.0++. In the model, the three-layer BP network is applied and average system error is 0.0001, and both weight and value are initialized by random functions. The performance of different optimized models based on several minimal attribute set given different fuzzy classification are shown in Fig. 2 and Tab. 1, respectively.

TABLE 1 COMPARISON OF THE PREDICTED RESULTS

Number of input nodes	Iterative time	17th month flooding quantity m^3/min		18th month flooding quantity m^3/min	
		Prediction value	Relative error%	Prediction value	Relative error %
2	1801	5.215	5.89	8.25	2.48
3	2542	5.028	2.19	7.89	1.87
6	4033	5.035	2.34	7.88	1.86

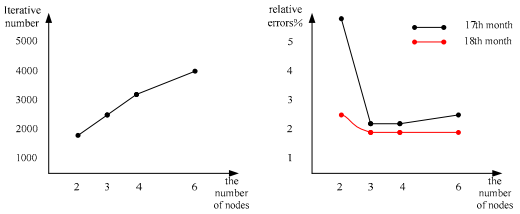

Figure 2 Analysis of the predicted results

B. Experiment analysis

As we can see from Fig.2, because the fuzzy credibility level value is different, prediction model based on fuzzy-rough set neural network has several input sets. In [15], all the factors have been used as input nodes (namely, the number of input nodes is six). Due to the redundant attributes both the precision and efficiency are not satisfied. It is shown from Tab. 1 that when the number of reduced attributes is minimal, the model is convergent very quickly while the accuracy is low; as the number of input nodes reaches three, the model performs better. Another attribute with most importance among the rest attributes is continually added intentionally and the number of input nodes reaches four. After 3222 times of iterative, the gas emissions of 17th month and 18th month are 5.03m3/min and 7.86m3/min, respectively. Compared to practical results, the corresponding relative errors are 2.29% and 1.84%, which shows that the precision of prediction is not obviously improved despite the longer training time.

According to analysis above, we can draw the following conclusions.

1) We can get some key factors or attributes by using the proposed fuzzy-rough set model.
2) There is a balance between the number of input nodes and model performance. The convergency and accuracy of the prediction model will be improved when the balance reaches.
3) We can select different prediction model to meet

different requirements, so the model is of great flexibility.

IV. Conclusion

The classical rough set theory and BP network have respectively superiority and shortage. Hence, it is possible to combine the two theories to enhance the reasoning and predicting power of an intelligent system. The analysis of the gas emission prediction model based on fuzzy-rough set neural network, which using minimal attribute set to solve the optimized problem of input layer, proves that the model can perform better than the traditional neural network in both accuracy and convergency. Furthermore, the model can also be applied in other complicated engineering problems.

Acknowledgement

The paper is supported by the Youth Foundation of Science, Shanghai, P. R. China (05XPYQ45)

References

[1] Z. Pawlak, AI and intelligent industrial applications: the rough set perspective. Cybernetics and Systems: An International Journal, 31(4), pp. 227-252, 2000
[2] Q. Shen and A. Chouchoulas, FuREAP: A fuzzy-rough Estimator of algae populations, Artificial Intelligence in Engineering, 15, pp.13-24, 2001
[3] Wang Shyue liang and Tsai Jenn Shing. Discovery of approximate dependencies from proximity-based fuzzy databases. The Third International conference on Knowledge-based Intelligent Information Engineering System, Adelaide, Australia, IEEE Inc.pp. 234-237, 1999
[4] Y. Yuan and M. J. Shaw, Induction of fuzzy decision trees, Fuzzy Set and System, 69(2), pp. 125-139, 1995.
[5] Y. Y. Yao, A comparative study of fuzzy sets and rough sets, Journal of Information Sciences, 109(1-4), pp. 227-242, 1998
[6] Z.Pawlark, Routh sets, Theoretical Aspects of Reasoning about Data. Nowowiejska 15/19,Warsaw.poland,10
[7] Z. Pawlak, Rough sets and fuzzy sets. Fuzzy Sets and Systems, 17(1), pp.88-102, 1985
[8] D. Dubois and H. Prade, Rough fuzzy sets and fuzzy rough sets[J]. Int. J. General Systems, 17, pp.191-208,1990
[9] Lippmann P R. Pattern Classification using Neural Networks. IEEE Comn., Magazine, April 1987, 4-22
[10] Zhang Zhen-yu, Xie Xiao-yao, Application research of BP neural network in telecom planning prediction, Computer Engineering and Applications，2008,44(20):245-248
[11] R. Hecht-Nielsen, Theory of the back-propagation of neural network, IEEE INNS International Conference on Neural Network, 1, pp. 593-605, 1989
[12] T. Kohonen, Self-Organization and Associative Memory, Springer, Berlin, 1988
[13] Hong Jing, Lu Jing-gui, Shi Feng, Combining Fuzzy Set and Rough Set for Inductive Learning. Intelligent Information Processing II, America: Springer, 2004.143-146
[14] Xue Peng-qian, Wu Li-feng, Li Hai-jun, Predicting the Amount of Gas Emitted Based on Wavelet Neural Network, China Safety Science,2006,2:22-25
[15] Yang Zhi-yi, Xiong Ya-xuan, Zhang Qian-lin, Research on the prediction of gas emission in working face based on neural network Coal Engineering, 2004,10:73-75
[16] R. Slowinski and D. Vanderpooten, A generalized definition of rough approximations based on similarity, IEEE Transactions on Knowledge and Data Engineering, 12(2), pp. 331-336, 2000

Jing Hong was born in Kunming City, Yunnan Province, in 1973. She received the master diploma in Computer Application from Nanjing Univ. of Technology. Her current research interests include research & application of algorithms of data mining and intelligent computation. She is working as a lecturer in College of electronic engineering, Shanghai University of Engineering Science. Email:nj_jinghon00@163.com.