Journal of Applied Sciences

ISSN 1812-5654

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (16): 3251-3256, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.3251.3256

An Integrated RS and ANN Design Method for City's Industry Development Stage Identification

Wang De-Lu, He Xin and Zhao Shen
School of Management, China University of Mining and Technology,
221116, Xuzhou, Jiangsu, China

Abstract: In recent years, research about industry development stage identification is taken seriously increasingly and great achievements have been made. Against the background of urban economy, an identifying method of city's industry development stage is put forward based on integration of Rough Sets (RS) and Artificial Neural Network (ANN). At first, the continuous attribute values are discretized using fuzzy clustering algorithm based on Maximum Discernibility Value (MDV) search method and information entropy. And then the major attributes are reduced by rough sets. At last, the Radial Basis Function (RBF) neural network is trained with training samples and the industry life cycle stages of testing samples are identified. The analysis results taking 669 industries of Dalian city as samples show that the fuzzy clustering algorithm based on MDV and information entropy can improve the discretization performance effectively. Compared with normal fuzzy evaluation method, the predicting precision of integration method is higher and it is an efficient and practical tool to identify development stage of city's industry.

Key words: Rough sets, RBF neural network, city's industry development stage, identifying method

INTRODUCTION

The basic assumption about industry life cycle is that an industry or its links follows four basic stages of development, namely start-up stage, growth stage, maturity stage and decline stage. Industry life cycle theory is of important realistic significance for both the government and enterprises. For the former, scientific and reasonable industrial policy should comply with evolving regulations of industry life cycle without undermining the market mechanism (Gort and Klepper, 1982). For the later, when drawing up and implementing competitive strategy, major effects probably caused by industry life cycle stages must be taken into consideration so as to increase perspectiveness of organization strategy (Jovanovic, 1998). However, the necessary premise for applying life cycle theory efficiently in practice is to identify industry development stage precisely.

In recent years, research about industry development stage identification is taken seriously increasingly and great achievements have been made. But following deficiencies are found in present research through consulting and analyzing numerous pertinent literatures at home and abroad: (1) Study objects are selected at state level or even from a wider range. As a small economic region, city's industry evolution turns out to be regional for intensive action of boundary conditions and influencing factors, so involved conclusions may fail to work. (2) Currently, identifying methods that mainly used are single factor analysis method (Lu, 2002), fuzzy evaluation method (Douglas, 1997) and input-output analysis method (Lave *et al*., 2002), which classifies samples statically according to simple indicators, making it difficult to grasp dynamic evolution trend of the industry, limited in practice.

From these, taking urban economy as background, this paper brings forward an identifying model of city's industry development stage based on integration of rough sets and neural network, which provides a new perspective and cognitive tool for government and enterprise to seize regional industry evolution law.

MODEL DESIGN

Discretization of continuous attributes based on MDV and information entropy: Fuzzy c-mean (FCM) clustering algorithm based on information entropy: The more reasonable the clustering division is, the more definite is the clustering ownership of data points and the smaller is the information entropy value. Therefore we will make it if the specific ownership can be found as possible to get the smallest information entropy. The first step of pedigree

Corresponding Author: Wang De-Lu, School of Management, China University of Mining and Technology, 221116, Xuzhou, Jiangsu, China

method is to define a number range of clustering number $[C_{min}, C_{max}]$ and designate the accuracy threshold ε, which is a decimal value between 0 and 1. Value range of ε is usually from 0.01 to 0.2 and the smaller it is, the more accurate will be the result but the longer it will spend, too. Every cluster number k ranging from C_{max} to C_{min} produces a membership matrix u^k ($k\in[C_{min}, C_{max}]$), which corresponds to information entropy value H_k (x).

If set the column value to data point i and the train value to cluster category j, then:

$$H_k(x) = \sum_{i=1}^{N} H_{ki}(x)$$

$$H_{ki}(x) = -\sum_{j=1}^{k} u_{ij} \bullet \log_2 u_{ij}$$

in which is information entropy of each data point. Select cluster number k of the minimal H_k (x) as the ultimate cluster number C and the result can be finally obtained by FCM algorithm.

Heuristic search method based on MDV function: Definition 1 Based on the decision table of normalized attributes, in accordance with distinguishability of the objects in rough sets, the definition of MDV is as follows:

$$MDV(a_i) = \sum_{j=1}^{d_{n-1}} \sum_{k=j+1}^{d_n} a_{i(j,k)} \tag{1}$$

$$a_i(j,k) = \begin{cases} 1, & \text{if} \quad a_i(x_j - x_k) = \max(a_{ii}(x_j - x_k)) \quad \text{and} \quad d(x_i) \neq d(x_j) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where, i = 1, ..., m, ii = 1, ..., m, j = 1, ..., n and k = j = 1, ..., n and a_i, m, d_n, n represent respectively a condition attribute, the number of condition attributes in the decision table, the number of decision patterns and quantity of objects on the universe of discourse. And a_i (j, k) denotes the times the absolute value of variation of a_i between relevant objects x_j and x_k is maximum in different decision patterns (j~k). a_{ii} (x_j-x_k) shows the absolute value of variation of a_i from decision pattern j to k.

MDV of original decision table is got by adding 1 to a_i (j, k) along with the maximum absolute value of variation and recording the times. MDV reflects relative dispersion degree of different attributes and the larger it is, the more evident will be the dispersion degree and the larger is the cluster quantity, or vice versa.

FCM discretization steps based on MDV Function and information entropy

Step 1: Calculate MDV function value of each condition attribute and arrange in accordance with the order from the biggest to the smallest

Step 2: Compare information entropy value to get the best cluster number k and discretize the condition attribute with maximum MDV using FCM clustering algorithm during the process of clustering number decreasing from C_{max} to C_{min}

Step 3: Update C_{max} with current k and delete this condition attribute from the sequence

Step 4: Repeat Step 2 and 3, discretizing each condition attribute of the sequence in turn

Attribute reduction of rough sets based on channel capacity: According to information theory, channel capacity between condition attribute and decision attribute in decision table shows the information content provided by the attribute. The maximum value of mutual information that is transmitted by the channel at different input probability distributions is regarded as channel capacity, which is a parameter of the channel. Channel capacity between condition attribute R and decision attribute D is defined as follows:

$$\begin{aligned} &\text{Capacity}(R, D) = \max_{P(R)}\{I(X, Y)\} \\ &\text{SGF}(p, R, D) = \text{Capacity}\left(R\bigcup\{p\}, D\right) - \text{Capacity}(R, D) \end{aligned} \tag{3}$$

Large capacity and vast information go hand in hand, which is more beneficial for decision-making. To decision system S = ⟨U, C∪D⟩, the importance of any attribute p ⊂C-R is defined as follows:

$$\text{SGF}(p, R, D) = \text{Capacity}\left(R\bigcup\{p\}, D\right) - \text{Capacity}(R, D) \tag{4}$$

According to decision attribute D, the universe of discourse U is divided into n categories $\{X_1, X_2, ..., X_n\}$ as input end Xof information transfer system, while it is divided into m categories $\{Y_1, Y_2, ..., Y_m\}$ as output end Y according to condition attribute R⊆C. So the computation formula of mutual information is as follows:

$$I(X, Y) = H(X) - H(X|Y) \tag{5}$$

Where:

$$H(X) = -\sum_{i=1}^{n} p(X_i)\log p(X_i)$$

is information entropy, describing information uncertainty of the system in original state and:

$$H(X \mid Y) = -\sum_{j=1}^{m}\sum_{i=1}^{n} p(Y_j)p(X_i \mid Y_j)\log[p(X_i \mid Y_j)]$$

serving as conditional entropy describes the information uncertainty under condition Y.

As to decision table S = $\langle U, C \cup D\rangle$, $R \subset C$, the importance of attribute $P \in C-R$ is derived from the Eq. 3-5 as follows:

$$SGF(P,R,D) = H(D \mid R) - H(D \mid R \cup \{p\}) \tag{6}$$

The algorithm flow of attribute reduction of rough sets based on channel capacity is as follows:

Step 1: Calculate the core of C to D;
Step 2: Let C_{rest} = C-Core and calculate the importance of each attribute in condition attribute set C_{rest}, which is listed in descending order of importance;
Step 3: Let B = Core and calculate positive region $Pos_C(D)$ and Pos_B (D) of D to C and B separately. If Pos_C (D) ≠ Pos_B (D), then do the following loop:

① Select attribute of the maximum importance as P and let $C_{rest} = C_{rest}$-P
② If $Pos_{B\cup\{P\}}$ (D) = Pos_C (D), B = B∪{P}, then end the loop
③ If $Pos_{B\cup\{P\}}$ (D) = PosC (D), B = B∪{P}, then return to ①

The final result B is a relative reduction of C to D. And core attribute set is solved by discernibility matrix (Wang *et al*., 2007).

Definition 2 With regard to decision table S = $\langle U, C \cup D\rangle$, C is the universe of discourse, C and D are respectively condition attribute set and decision attribute set and a(x) is value of x for attribute a. Discernibility matrix can be defined as follows:

$$C_{ij} = \begin{cases} a \in C; a(x_i) \neq a(x_j), D(x_i) \neq D(x_j) \\ 0; D(x_i) = D(x_j) \\ -1; a(x_i) = a(x_j), D(x_i) \neq D(x_j) \end{cases} \tag{7}$$

In discernibility matrix, attribute sets with attribute combination 1 are referred to as core and other useful attributes can be acquired from those being not.

RBF neural network training: RBF neural network is a three-tier feed-forward network. The number of input layer neurons coincides with the dimensions of input vector and so does that of output layer with output vector. And hidden neuron adopts RBF as its output characteristic (Han and Kamber, 2001; Chen, 2012). The most common RBF is Gaussian kernel function:

$$\ker(\|x-u_j\| = e^{-\frac{\|x-u_j\|^2}{2\delta_j^2}}, j = 1, 2, \ldots l \tag{8}$$

where, u_j and δ_j are respectively the central value and width of the hidden node j, determining the scope of Gaussian kernel function.

The learning algorithm of RBF neural network classifies training process into two stages: In the first place, confirm central point u_j and standard deviation δ_j of basic function and then decide the weight between hidden layer and output layer.

Central parameter of RBF neural network by subtractive clustering algorithm: To determine u_j and δ_j is a key issue in RBF network study. This paper takes subtractive clustering algorithm (Nogales *et al*., 2002) as a guide of cluster study so as to reduce artificial factors and controls the number of cluster to get reasonable central parameter through an automatic termination criterion.

Given P-dimensional space and n sets of sample data ($x_1, x_2, \ldots, x_n$), first normalize them into the hypercube and afterwards calculate dispersion index of each sample x_i:

$$D_i = \sum_{j=1}^{n} \exp\left(-\frac{\|x_i - x_j\|^2}{(\sigma_a/2)^2}\right), (i = 1, 2, \ldots n) \tag{9}$$

where positive number σ_a defines a neighborhood. Choose data point x_{u1}, of which the dispersion index D_{u1} is highest, to be the first clustering center. With modifying each data point x_i, the new dispersion index as follows:

$$D_i = D_i - D_{u1}\sum_{j=1}^{n} \exp\left(-\frac{\|x_i - x_{u1}\|^2}{(\sigma_b/2)^2}\right), (i = 1, 2, \ldots n) \tag{10}$$

where constant σ_b is greater than σ_a. Select the next clustering center x_{u2} after modification and modify again. Just do it over and over again and then stop clustering until there are very few data points in some clustering center. Cluster termination criterion is known as:

$$D_{max}/D_{u1} < \gamma \tag{11}$$

where, D_{max} means the maximum dispersion index while D_{u1} is the initial one and γ is a small constant. Then the width parameter δ_j is confirmed by means of getting average distance of partial points adjacent to x_{ui}.

Weight adjustment of RBF network with least square method: Suppose there are n sets of input samples x_i, of which d_i is the desired output. The objective error function is specified as follows:

$$E = \frac{1}{2}\sum_{i=1}^{n}\left\|d_i - y_i\right\|^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(d_{ij} - y_{ij}\right)^2 \qquad (12)$$

where, y_i is output vector of the network along with x_i. The weight training from hidden layer to output layer is conducted when u_j and δ_j is determined. Because the relationship between hidden layer and output layer is a linear system of equations, least square method is adoptable to solve the problem of linear optimization. The output neuron with greatest function value is the sort result owing to the correspondence between output neuron and linear combination.

EMPIRICAL ANALYSIS

Sample selection and variable design: Taking into account the actual state of industrial structure of Dalian city and roughly the same samples in different development stages, 669 middle class industries are chosen as samples whose development stage is available from 1986 to 2006 and uses 1, 2, 3 and 4 to separately represent start-up stage, growth stage, maturity stage and decline stage. Using the software SPSS13.0 446 sets of samples are selected randomly as training samples and the remainder are testing samples. There are 74, 138, 183 and 51 industries respectively at start-up stage, growth stage, maturity stage and decline stage in training samples, while the numbers in testing samples are respectively 41, 68, 87 and 27.

In the meantime, getting each index variable on the basis of former research or adjustment combining with the research content in order to make the scale scientific and authoritative, considering availability of index data, this paper ultimately identifies 15 indicators as proposed index variables of this study from aspects of industry evolution trend at macroscopic level, city's industry competitiveness and resource constraints, as shown in Table 1.

Discretization performance: Matlab 7.0 is applied to discretize in terms of algorithms provided by this paper. Discretization results are finally acquired on the basis of discretization interval adjustment according to expert advice, as shown Table 1. Compared with conventional equifrequent and equidistance methods, etc., discretization results of FCM clustering algorithm is more in line with objective reality. What's more, heuristic search method based on MDV makes cluster seeking times reduce from 105 to 39 and discretizaiton efficiency be significantly improved.

Reduction results of condition attributes: Attribute reduction of discretized data is implemented by rough sets, gaining 39 reductions, of which the two smallest condition attribute sets, namely $\{I_2, I_3, I_5, I_{13}, I_{14}, I_{15}\}$ and $\{I_3, I_6, I_8, I_{13}, I_{14}\}$, are originally formed into new learning samples to train RBF network.

RBF neural network learning outcomes: There are 16 neurons in hidden layer and 4 in output layer according to subtractive clustering algorithm. Reduction 1

Table 1: Evaluating index system and discretization interval

Factors	Evaluating Indicator	Symbol	Discretization Interval			
Resource constraints	Emissions of unit added value (ton billon yuan^{-1})	I_1	[0, 25.3)	[25.3, 39.5)	[39.5, +8)	-
	Unit area productivity (billon yuan km^{-2})	I_2	[0, 2.7)	[2.7, 8.8)	[8.8, 12.7)	[12.7, +8)
	Freshwater consumption of unit added value (m^3 million yuan^{-1})	I_3	[0, 21)	[21, 80)	[80, 154)	[154, +8)
Industry evolution trend	Growth rate of ratio of output to GDP (%)	I_4	(-8, 0)	[0, 0.93)	[0.93, 6.6)	[6.6, +8)
	Growth rate of ratio of added value to GDP (%)	I_5	(-8, 0)	[0, 0.68)	[0.68, 4.7)	[4.6, +8)
	Growth rate of sales revenue (%)	I_6	(-8, 0.1)	[2.1, 8.5)	[8.5, +8)	-
	Growth rate of industrial profit rate (%)	I_7	(-8, 0)	[0, 7.3)	[7.3, 22)	[22, +8)
Industry competitiveness	Index of personnel with university degree and intermediate and senior title	I_8	[0, 0.61)	[0.61, 1.00)	[1.00, 1.27)	[1.27, +8)
	Weight index No. of technology funds to sales revenue	I_9	[0, 0.51)	[0.51, 1.12)	[1.12, +8)	-
	Weight index No. of digestive absorption funds to sales revenue	I_{10}	[0, 0.36)	[0.36, 1.23)	[1.23, +8)	-
	Weight index No. of R and D staff	I_{11}	[0, 0.24)	[0.24, 1.15)	[1.15, 3.07)	[3.07, +8)
	Weight index No. of R and D expenses to sales revenue	I_{12}	[0, 0.50)	[0.50, 0.91)	[0.91, 2.92)	[2.92, +8)
	Weight index No. of new products' sales revenue	I_{13}	[0, 0.13)	[0.13, 1.00)	[1.00, +8)	-
	Labor productivity index	I_{14}	[0,1.00)	[1.00,+8)	-	-
	Average output value index	I_{15}	[0, 1.00)	[1.00, 2.89)	[2.89, +8)	-

Table 2: Misjudgment comparison of testing samples

	1			2			3			4		
Method type	2	3	4	1	3	4	1	2	4	1	2	3
RS-ANN	1 (2.44%)	1 (2.44%)	3 (7.32%)	6 (8.82%)	4 (5.88%)	3 (4.41%)	2 (2.30%)	5 (5.75%)	5 (5.75%)	1 (3.70%)	1 (3.70%)	2 (7.41%)
Fuzzy evaluation	3 (7.31%)	3 (7.31%)	5 (12.20%)	9 (13.24%)	6 (8.82%)	7 (10.29%)	5 (5.75%)	8 (9.20%)	5 (5.75%)	3 (11.11%)	2 (7.41%)	2 (7.41%)

Figures out of brackets are No. of samples that are mistakenly assigned to the stages and bracketed figures are ratios of misjudged samples to the total of the stages

Table 3: General prediction accuracy comparison of testing samples

Method type	1	2	3	4	Accuracy rate
RS-ANN method	36 (87.8%)	55 (80.88%)	75 (86.21%)	23 (85.19%)	189 (84.75%)
Fuzzy evaluation method	30 (73.17%)	46 (67.65%)	69 (79.31%)	20 (74.07%)	165 (74.00%)

Figures out of brackets are No. of samples that are correctly assigned to the stages and bracketed figures are ratios of accurate samples to the total of the stages

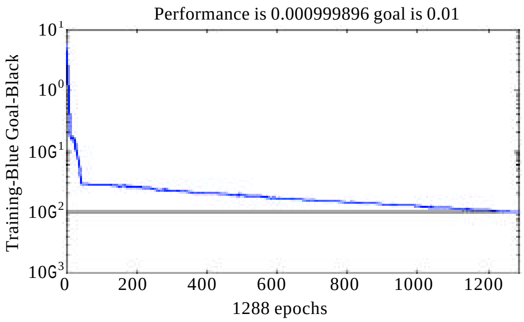

Fig. 1: RBF training results of reduced indexes (1)

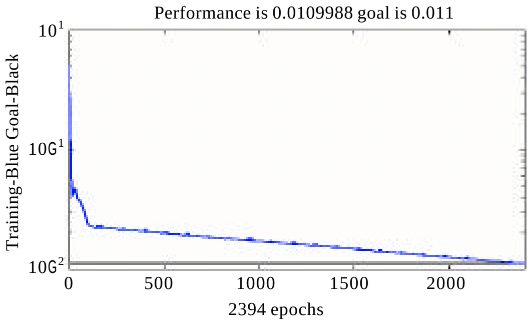

Fig. 2: RBF training results of reduced indexes (2)

$\{I_2, I_3, I_5, I_{13}, I_{14}, I_{15}\}$, which is obtained from reduction of 446 training samples, being input of the RBF network, target error reaches 0.01 after 1288 times of training, as shown in Fig. 1; while target error reaches 0.011 after 2394 times of training by reduction 2 $\{I_3, I_6, I_8, I_{13}, I_{14}\}$, as shown in Fig. 2. It is observed that performance of Fig. 1 is superior to that of Fig. 2 in both training frequency and target error.

To show the learning efficiency of RBF network induced by attribute reduction using rough sets, 15 unreduced evaluating indicators are directly input into RBF network and proved to be unqualified on account that target error fails to reach 0.01 after 5000 times of practical training, as shown in Fig. 3.

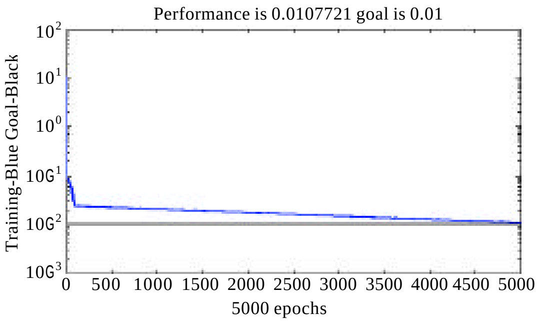

Fig. 3: RBF training results of original indexes

To sum up, RBF neural network in Fig. 1 is chosen to identify city's industry development stage of testing samples on the basis of maximum output.

Predicting accuracy test: Two hundred and twenty three testing samples are Input into trained RBF neural network of Fig. 1 for identification, aiming at verifying practicality of integration of rough sets and RBF neural network proposed by this paper in identification of city's industry development stage. At the same time, the prediction accuracy of integration method is compared with fuzzy evaluation method and misjudgment and general prediction accuracy of testing samples are illustrated in Table 2 and 3. The results indicate that misjudgment rate of integrated identification model is lower than fuzzy evaluation method and its prediction precision reaches up to 84.75%, apparently higher than fuzzy evaluation method's 74.00%, which certifies validity and practicality of the model.

CONCLUSION

Taking urban economy as background, an identifying method of city's industry development stage is put forward based on the integration of rough sets and RBF neural network and 669 industries of Dalian city is analyzed empirically. The results demonstrate that taking advantage of rough sets theory to reduce condition attributes, six evaluating indicators are obtained without information loss, namely unit area productivity, freshwater consumption of unit added value, growth rate of ratio of added value to GDP, weight index number of new products' sales revenue, labor productivity index and industry average output index, which helps to simplify neural network structure effectively and increase its learning efficiency. And yet actual error cannot achieve 0.01 after training neural network 5000 times by adopting 15 unreduced indicators. Furthermore, RBF neural network is applied to train and identify reduced samples in view of its characteristics of high fault-tolerance and excellent extensibility. The results show that the prediction accuracy of integration method is obviously higher than common fuzzy evaluation method, thus it's an efficient and practical tool to city's industry development stage.

ACKNOWLEDGMENT

This study is supported by National Natural Science Foundation of China under Grant No. 71202112.

REFERENCES

Chen, S.G., 2012. Efficient spatial association rule mining algorithm based on region. Int. J. Adv. Comput. Technol., 4: 211-218.

Douglas, N., 1997. Applying the life cycle model to Melanesia. Ann. Tourism Res., 24: 1-22.

Gort, M. and S. Klepper, 1982. Time paths in the diffusion of product innovations. Econ. J., 92: 630-653.

Han, J.W. and M. Kamber, 2001. Data Mining Concepts and Techniques. Academic Press, New York.

Jovanovic, B., 1998. Michael cort's contribution to economics. Rev. Econ. Dyn., 1: 327-337.

Lave, L., C. Hendrickson and A. Horvath, 2002. Economic input-output models for environment life-cycle assessment. Environ. Sci. Technol., 32: 184-191.

Lu, G.Q., 2002. Declining Industry. Nanjing University Press, Nanjing, China.

Nogales, F.J., J. Contreras, A.J. Conejo and R. Espinola, 2002. Forecasting next-day electricity prices by time series models. IEEE Trans. Power Sys., 17: 342-348.

Wang, J.Y., A. Luo and S.Q. Chen, 2007. Interval reduction of variable precision rough sets model. J. Sys. Eng., 22: 621-626.