Applying Rough Set Theory to Establish Artificial Neural Networks Model for Short Term Incidence Rate Forecasting

Xiangyu Zhao
School of Computer Science and Technology
Panzhihua University
Panzhihua, China
e-mail: 59599849@qq.com

Liangliang Ma
School of Computer Science and Technology
Panzhihua University
Panzhihua, China
e-mail: mll198684@126.com

***Abstract*—Choosing input variable and networks architecture are key processes for modeling short term incidence rate forecast by artificial neural networks, in this paper a method based on rough set theory is proposed to deal with them. In the proposed approach, the key factors that affect the incidence rate forecasting are firstly identified by rough set theory and then the input variables of forecast model can be determined. On the basis of the process mentioned above a set of influence rules can been obtained through reductive mining process of attributes and attribute values, then a neural networks of incidence rate forecast model is established on the rule set and BP-algorithm is adopt to optimize the networks. The method indicates that incidence rate forecast model can be established according some theoretical principles and avoiding blindness. A practical application is given at last to demonstrate the usefulness of the novel method.**

Keywords- incidence rate forecasting; neural networks; rough set

I. INTRODUCTION

BP neural network and fuzzy neural network are the common tools for disease's incidence rate nonlinear prediction model and prediction model based on various influence factors [1-5]. At present, incidence rate prediction model based on BP neural network or fuzzy neural network (such as the basic problems of selecting input variables and determining the network structure for the model) are all lack of effective theory methods, and are basically determined by the experience and the repeated test.

The fuzzy set theory is a kind of new mathematics theory and method for researching the uncertainty and vague problems, and is one of the important ways for finding the useful information and obtaining knowledge in data mining area, and successful applied in many areas. For the problem of establishing incidence rate neural network prediction model, we try using rough sets theory to choose the input variables and determine the structure of the neural network in this paper.

The rest of this paper is organized as follows: section 2 describes the concept of the decision table and rough set theory, section 3 establishes neural network models based on rough set theory. The established model is applied to forecasting the incidence rate in section 4, and finally the conclusions are discussed in section 5.

Corresponding author: Liangliang Ma, e-mail: mll198684@126.com

II. CONCEPT OF THE DECISION TABLE AND ROUGH SET THEORY

Definition 1[6-8]: Decision table is an information knowledge expression system consisted by (X, R, V, f), where X is the object set, $R = C \cup D$ is the property set, subset C and D are condition attribute set and decision attribute set respectively, $V = \bigcup V_r$ is the set consisted by the attribute value, V_r is the domain for attribute r, $f: X \times R \to V$ is the information function, which specify attribute values for each object for X, $D \neq \varnothing$.

In this paper, we only considering the decision table with single decision attribute, and its general form is shown in table 1, among them, $X = \{x_1, x_2, \cdots, x_n\}$, $C = \{c_1, c_2, \cdots, c_m\}$, $D = \{d\}$, $f(x_i, c_j) = u_{j,i}$, $f(x_i, d) = v_i$.

TABLE I. A GENERAL FORM OF DECISION TABLE

object	condition attribute			decision attribute
X	c_1	$\cdots$	c_m	d
x_1	$u_{1,1}$	$\cdots$	$u_{m,1}$	v_1
x_2	$u_{1,2}$	$\cdots$	$u_{m,2}$	v_2
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_n	$u_{1,n}$	$\cdots$	$u_{m,n}$	v_n

Definition 2: For one equivalence relation T in X, the lower approximation set definition for subset $B \subseteq X$ under this equivalence relation is

$$T\_(B) = \{x : x \in X \wedge [x] \subseteq B\} \quad (1)$$

where $[x]$ is the equivalence class of x.

Assume Φ is the equivalence relation clan which determined by the decision table according to the values of the condition attributes, the intersection of the relationship between the equivalent relations Φ is still an equivalence relation, express as Q. Assume P express the equivalence relation of decision attribute according to the attribute

values. Set $\{X_1, X_2, \cdots, X_s\}$ as the equivalent subset clan for Q, and then we can obtain definition 3 and 4.

Definition 3: The compatibility degree (classification quality) of decision table is

$$d_p(Q) = \sum_{k=1}^{s} | P\_(X_k) | \vee | X | \quad (2)$$

where $| X |$ the element number of X, $P\_(X_k)$ is the approximation set of subset X_k under the equivalence relation P, $0 \le d_p(Q) \le 1$, when $d_p(Q) = 1$, we cam say the decision table is compatible or coordination.

Definition 4: Assume the compatibility degree is $d_p(Q-\{c_i\})$ after removing one condition attribute c_i for the decision table, and then the importance definition of condition attributes is

$$\Gamma(c_i) = d_p(Q) - d_p(Q - \{c_i\}) \quad (3)$$

where P and Q have same meaning with definition 3.

If the importance of an attribute is zero, which can show when removing this attribute the compatibility degree of the decision table will not change, namely this property is redundant.

III. NEURAL NETWORK MODELS ESTABLISHED BASED ON ROUGH SET THEORY

Aim to incidence rate prediction, the amount of predictors is the nonlinear functions of influencing factors, so when establishing forecasting model, we should determine the appropriate input variables and the input-output nonlinear function. For the nonlinear function defined in a bounded area, if we divided the input space into several areas, and made each portion of the function is convex function, then we can use these convex function forms to fit the anti-derivative. If we can try to reduce the number of such a division area, we still can obtain a concise fitting form. Approach to the nonlinear function, if we first appropriate division the output space, then decided the input space division by those divisions, that the function determine of each area are on convex functions. Set the nonlinear function in the interval $[a,b]$ in graph 1 as an example, y_1 and y_3 were the minimum and maximum of the function, the point y_1 of the output space confirmed three points x_1, x_2 and of the input space, and the corresponding relations between the intervals as follows:

$$\begin{cases} [a, x_1) \to [y_1, y_2) \\ [x_1, x_2) \to [y_2, y_3] \\ [x_2, x_3) \to [y_1, y_2) \\ [x_3, b] \to [y_2, y_3] \end{cases} \quad (4)$$

Each of these interval are all correspondingly determined a convex function, and for these given division of the output space, the number of the corresponding interval is the minimum. These corresponding intervals can be seeing as the characterization information of the nonlinear function structure.

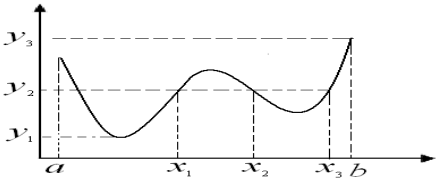

Figure 1. Output space partition producing input space partition for nonlinear function.

According to the above analysis, when establishing multidimensional models on the known data sample, how to choose the least and the most effective variables division, and determine the least corresponding relations will be the key resolving problems after made the division of the output space. Based on the rough set theory, we using the continuous variable discretization algorithm to select the input variables division, and using the reduction algorithm of decision table obtain the reasoning rules set, these reasoning rules set is the least corresponding area which represent the corresponding relations between the output space division and the original data.

The procedure of the proposed neural network incidence rate prediction model based on the rough set theory in this paper is as follows:

- Construct the decision table on the incidence rate of history data and related information historical data.
- Initialization, set $n = 2$.
- Discriminating the decision table according to the decision attribute variables into n regions.
- Choose condition attribute variables and the breakpoint (point) based on the rough set breakpoints importance discretization algorithm[6-8], and calculating the consistence degree, when the consistence degree is 1 or no longer increases, then the choose process end.
- Construct new decision table by choose condition attribute variables and their sample discrete values, and reduction, and obtain reasoning rules set[6-13].
- Using the reasoning rules set establishing neural network model.
- Training several times for the neural network.
- If the fitting error of the neural network satisfied the requirements, then end; otherwise increase n, turn step 3).

Approach to the above modeling steps, step 3) is very important, although as long as the output of space points fine enough in theory, we can get the hope input space division, but this can inevitable led to excessive area corresponding, and make the final network structure too

complex, affects the generalization ability, therefore, how to division the output space to make the resulting regional corresponding number at least is the problem worth further research. Inspired by the example of figure 1, we using the method of frequency in this paper to division the output space. Set the number of sample data set is N , divided it into n small zones. We first division the sample data from childhood in this article, then choose the cent point in turn, make the number of the fall sample data of each interval is N / n .

In step 7), we use the BP learning algorithm with momentum as the learning algorithm of the neural network.

IV. PERFORMANCE ANALYSIS

In this paper, approach to the short-term prediction model of the incidence rate analysis and modeling, we only discuss the biggest incidence rate prediction model modeling for an example. The original data is obtained from Qinghai Haixizhou first people's hospital, incidence rate data and meteorological data from January 2003 to December 2009. When modeling we take the data between January 2003 and December 2008, including month biggest incidence rate, average temperature, average monthly air pressure, average humidity. When establishing the initial decision table, we set the biggest prediction month incidence rate as the decision attribute variables, and select average temperature, average month air pressure and month average humidity of the prediction month as the condition attribute variables, the number of the decision attribute variables is 1, the number of the condition attributes variables is12, all attributes in the decision table are continuous attributes.

First of all, we set $n = 2$ and division the decision attribute variables, based on this division through the rough set discretization algorithm we choose 4 condition attribute variables, the results are listed in table 2. By these four properties we can obtain a new decision table, and calculate the consistency degree (see definition 3) is 1, further calculate the importance of these four attributes (see definition 4), the results are also listed in table 2. At this time, because the importance of each attribute is not equal to zero, so each attribute is not redundant.

After reduction, we get five rules, and we can determine the structure of the neural network by these rules, that the number of input variables is 4, the number of the hidden layer of neurons is 5, each reasoning rules corresponding to a hidden layer and a neurons connected weight. Finally, we using the original historical data to train the network, and set the learning times as 1000, the learning curve is shown in figure 2. The study result is shown in table 3, and at this time, the fitting precision of the neural network is more ideal, and the modeling process is over.

From the above results, we can concluded that although the first choice of the input variables is 12, but through the rough set theory analysis only retained 4 variables as final input values, and based on the rough set theory analysis results, and the constructed neural network has more reasonable coverage for the input variable space, making the learning process and learning results of the network all ideal.

TABLE II. THE DISCRETIZING RESULT OF CONDITION ARTTRIBUTES

attribute code	average temperature	average pressure	average humidity	month biggest incidence rate	Attribute importance
C_1		prediction month			0.2384480
C_2	prediction month				0.3450367
C_3			prediction month		0.2216430
C_4				prior seven months	0.1948723

TABLE III. THE LEARNING RESULT OF ROUGH SET NEURAL NETWORS

learning number	learning time/ms	average relative error/%
1000	16	0.9253716

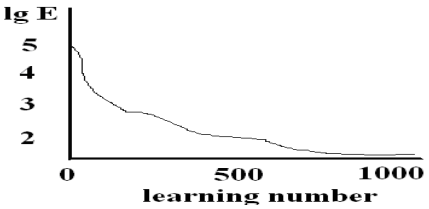

Figure 2. Learning curve of rough set neural networks.

For comparison, according to the above data, we using the Sugeno fuzzy forecasting model in reference [2-4], first select the 3 variables which related to the biggest month calculation: average temperature of the prior month, average pressure and average humidity, and use (u_1, u_2, u_3) as the corresponding property variable, and also as the fuzzy reasoning if-then former domain. And controlling the fuzzy reasoning rules and input variables by the corresponding 2 tolerances ρ_1 and ρ_2 , tolerance ρ_2 is 3.5×10^{-4} , the fuzzy inference rule 2-4 are calculated, and the calculation results is shown in table 4.

Tab.4 The modeling result of Sugeno fuzzy system based on OLS-algorithm

TABLE IV. THE MODELING RESULT OF SUGENO FUZZY SYSTEM BASED ON OLS-ALGORITHM

rule number	ρ_1	input variable number	average relative error/%
2	0.00132791	12	0.0164232
3	0.0014261	10	0.0174108
4	0.0025238	8	0.0144537

From table 4, we can conclude that when choose different tolerance ρ_1 , the number of rules different, and the input variables of the corresponding number is also different, that are all more than the method in the paper. Due to minority input variables nonlinear, other variables is in

linear form, so the nonlinear recognition and fitting ability between input and output relationship is limited, the above results showed this point.

V. CONCLUSION

Due to many factors influenced the accuracy of short-term incidence rate prediction, how to determine the model's input variables and the structure of the network is very important when establishing neural network model, the proposed method based on the rough set theory in this paper is an effective method to solve this problem, its features are as follows:

a) Determine the input variables of the prediction model through the history sample data, which can avoid the blindness;

b) Determine the structure of the set the neural network through the reasoning rules taken from data samples, the model is explicable;

c) The whole modeling process can be completed by the algorithm, which is feasible for practical application.

REFERENCES

[1] Tamimi M and Egbert R. Sneddon, "Short term electric load forecasting via fuzzy neural collaboration," Electric Power Systems Research, vol. 56, pp. 243–248, April 2000.

[2] Mastorocostas P A, Theocharis J B, and Bakirtizis A G, "Fuzzy modeling for short term load forecasting using the orthogonal least squares method," IEEE Transaction on Power Systems, vol.14, pp. 29–36, January 1999.

[3] Xie Hong, Chen Zhiye, and Niu Dongxiao, "The research of daily load forecasting model based on wavelet decomposing and climate influencing," Proceeding of the CSEE, vol.21,pp. 5–10, May 2001.

[4] Wang Feng, Yu Erkang, and Yan Chengshan, "Study of short term load forecasting based on influencing factors," Proceeding of the CSEE, vol.19,pp. 54–57, Auguest 1999.

[5] Xie Hong, Cheng Haozhong, Zhang Guoli, Niu Dongxiao, and Yang Jingfei, "Applying rough set theory to establish artificial neural networks for short term load forecasting," Proceeding of the CSEE, vol.23,pp. 1–4, November 2003.

[6] Wang Guoying, Rough set theory and knowledge acquisition. China, Xi'an: Xi'an Jiaotong University Press, 2001.

[7] Fen Chunshan, Wu Jiachun, Jiang Fu. Combined Prediction of Oil Prices.Journal of the University of Petroleum, China (Edition of Social Sciences), vol.1, pp.12-14, January 2004.

[8] Li Xiguo, Tan Dingshan, Shao Jinhua.Research of Precipitation in Yantai Based on ARIMA Model. Journal of Shandong water conservancy, vol.2, pp.37-39, September 2006.

[9] Sheng Yanbo. Forecast of Domestic Product Per Capita in Zhejiang Province Based on Combined Model of BP Neural Network and ARIMA Model.Journal of Business research, vol.23.pp.49-50, April 2006.

[10] Wang Ping.Forecasting of Tourism Demand Theoretical and Empirical Study Based on Artificial Neural Network-Case of Qingdao City.Lanzhou: Northwest Normal University, vol.3, pp.37-39, October 2004.

[11] Ma Liangliang, Tian Fupeng.Research of Nephritis Morbidity Situation in Haixizhou Region Based on Season Model.Journal of Beijing Union University(Natural Sciences),vol.23,pp.68-69,March 2003.

[12] Ma Liangliang, Tian Fupeng. Research of cerebral hemorrhage morbidity situation in Haixizhou region based on PDL Model.Journal of Hunan University of Arts and Science (Natural Science Edition), vol.21, pp.17-19, March 2009.

[13] Ma Liangliang, Tian Fupeng.Research of Hypertension Morbidity Situation in Haixizhou Region Based on ADL Model.Journal of Zhejiang Wanli University, vol.22,pp.6-9,May 2009.

Published by Atlantis Press, Paris, France.