DOI:10.13656/j.cnki.gxkx.2013.02.004

广西科学 Guangxi Sciences 2013,20(2):128～131,136

A Model of Classifier based on Rough Set and Genetic Neural Network*
一种基于粗糙集和遗传神经网络的数据分类器模型

HONG Yue-hua[1],XU Shuang[2],LIANG Jia-rong[3]

洪月华[1],徐 霜[2],梁家荣[3]

(1. Department of Computer Science,Guangxi Economic Management Cadre College,Nanning, Guangxi, 530007, China; 2. Vocational Technical College, Yulin Normal University, Yulin, Guangxi,537000,China;3. School of Computer and Electronics and Information,Guangxi University,Nanning,Guangxi,530004,China)

(1. 广西经济管理干部学院计算机系,广西南宁 530007;2. 玉林师范学院职业技术学院,广西玉林 537000;3. 广西大学计算机与电子信息学院,广西南宁 530004)

Abstract:In order to realize classification and recognition for the high dimensional redundancy and uncertain data monitored by wireless sensor network, a BP neural network data classifier model optimized by genetic algorithm and rough set is proposed, and then a classification algorithm of data mining is formed. In the model, attribute reduction algorithm of rough set theory is used to delete redundant attribute of training samples, and then genetic algorithm is used to optimize weights and threshold values, and carry out neural network learning. The classification algorithm of data mining has better learning speed, and can improve the efficiency of data classification in wireless sensor network.

Key words:rough set theory,genetic algorithm,BP neural network,wireless sensor network,data mining

摘要:为了实现对无线传感器网络监测得到的高维冗余且不确定的数据进行分类识别,提出一种由遗传算法和粗糙集进行优化的BP神经网络数据分类器模型,并形成了数据挖掘分类算法。该模型通过粗糙集理论的属性约简算法删除训练样本的冗余属性,利用遗传算法对神经网络的权值和阈值进行优化,并进行神经网络学习。数据挖掘分类算法学习速度快,能够有效提高无线传感器网络中数据的分类效率。

关键词:粗糙集 遗传算法 BP神经网络 无线传感器网络 数据挖掘

中图法分类号:TP183 **文献标识码**:A **文章编号**:1005-9164(2013)02-0128-04

The monitoring data collected by sensor node is high dimension, redundancy and uncertain. If the monitoring data is transmitted directly to central server,large amounts of data transmission will cost precious network energy,and bandwidth occupation is bigger also,which can′t handle the situation of a lot of sensors [1]. Before transmitting data it is necessary to classify the huge monitoring data,so as to reduce the monitoring data transmission,make use of wireless sensor network precious energy efficiently,and lengthen the life cycle of wireless sensor network.

The purpose of classification is to construct a classification function or classification model (also known as classifier) according to the characteristics of the data set,and this model can put the unknown sample mapping to a given category. Artificial neural network has widely used in the classification field. However,the common BP neural network has the

收稿日期:2013-03-07

作者简介:洪月华(1973-),女,副教授,主要从事人工智能算法在数据挖掘中的应用研究。

Supported by the National Natural Science Foundation of China (No. 61064002), Guangxi Science Foundation Junior Program (2012jjBAG0074),The Open Fund Project of Key Laboratory of Guangxi Hybrid Calculation and Integrated Circuit Design and Analysis (2012HCIC05), The Guangxi Department of Education Project (200103YB195).

problem of slow convergence speed and low accuracy, easily falls into the local minimum value. Especially in high dimensional data characteristics information is huge, which often can′t meet the request of fast convergence and accuracy diagnosis, so its application is restricted in the areas of classification. To solve these problems, many scholars have studied in this area. In literature[2～4], genetic algorithm was used to improved BP network learning algorithm. In literature [5, 6], a algorithm was proposed, which integrated the rough sets and neural network. In the paper, we proposed a BP neural network classifier model, which based on the rough set theory and genetic algorithm.

1 BP neural network structure and rough set

1.1 BP neural network structure

The BP neural network we used has three layers: input layer, hidden layer and output layer, and the node number of the three layer is n, h and m . Assuming $X=(x_1,x_2,\cdots,x_n)^T$, $Z=(z_1,z_2,\cdots,z_h)^T$, $Y=(y_1,y_2,\cdots,y_m)^T$ denote input layer, hidden layer and output layer vector, respectively, and $W=(w_{11},w_{12},\cdots,w_{hn})^T$ and $V=(v_{11},v_{12},\cdots,v_{mh})^T$ denote weights between nodes from input layer to hidden layer and from hidden layer to output layer, respectively. $\hat{Y}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_m)^T$ denotes expected output value of output layer. The incentive function $f(\star)$ of output nodes adopts the Sigmoid function, the expression is:

$$f(x)=\frac{1}{1+e^{-x}}. \qquad (1)$$

1.2 Basic knowledge of rough set

Attribute reduction is a core topic of rough set, and its main function is to delete the unimportant or irrelevant attributes under the condition of keeping information system classification, and then make the classification more effective. Here are some concepts and define of rough set.

1.2.1 Information system[6]

Information system is denoted by the four tuple $S=(U,A,V,f)$, among them $U=\{x_1,x_2,\cdots,x_n\}$ is domain, which is the nonempty finite set of study object. $A=\{a_1,a_2,\cdots,a_q\}$ is finite nonempty set of object properties. $V=\bigcup_{\forall a\in A} v_a$, v_a denotes the range of attribute a; $f=U\times A\rightarrow V$ is information function, namely, $\forall x\in U$, $\forall a\in A$, $f(x,a)\in v_a$ denotes the value of objects x on attribute a . If D, C denote decision attribute set and conditional attribute, respectively, and $A=C\cup D$, $C\cap D=\varnothing$, then the information system is also called the decision system.

1.2.2 Equivalence relation

In information system S, $B\subseteq A$ is any subset of attributes, the equivalence relation of domain U is denoted by the statement:

$$IND(B)=\{(x_j,x_k)\in U\times U:\forall b\in B, f(x_j,x_b)=f(x_k,x_b)\}. \qquad (2)$$

Where $x_j,x_k\in U$ and $j\neq k$. The equivalence $IND(B)$ partition of universe U denoted as $U/IND(B)$, abbreviated as U/B :

$$U/B=\{X_1,X_2,\cdots,X_m\}. \qquad (3)$$

X_i is the set got by B , and satisfied the statement: $U=\bigcup_{i=1}^{m}X_i$, $X_i\cap X_j=\varnothing$, $X_i,X_j\neq\varnothing$, $i,j\in[1,m]$.

1.2.3 Attributes reduction and core[7]

In information system S there is a equivalent relation family B, $b\in B$, if $IND(B)$ is equal to $IND(B-\{b\})$, then the attribute b is redundant in B , otherwise b is necessary. All the necessary attributes of B is formed the kernel of B , denoted as core (B) , and the nucleus is the only. For $C\subseteq B$, if each attribute in C is necessary, and the $IND(C)=IND(B)$, then C is a reduction of B , denoted as $C\in$ core(B) . All reduction of B is denoted red (B) , and attribute reduction is not unique. The kernel of B is intersection of all reduction, core(B) $=\bigcap$ red(B) .

1.2.4 Positive region

$IND(B)$ is an equivalence relation on the domain U , for any $X\subseteq U$, the B positive regions of X is defined as

$$\mathrm{POS}_B(X)=\bigcup\{X_i\mid X_i\in U/B, X_i\subseteq X\}. \qquad (4)$$

2 Rough set genetic neural network classifier model

Genetic algorithm is used to train neural network to form genetic neural network, and then the global optimization weights and threshold value are got; the attribute reduction is used to reduce data of inputting by genetic neural network, and then the representative decision attributes are chosen to learn and predict genetic neural network algorithm. The trained neural network algorithm is integrated in the sensor on the base station, so as to shorten the data processing time and improve the accuracy of classification. Rough set genetic neural network classifier model is shown in figure 1.

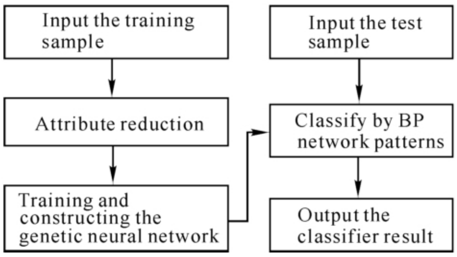
Input the training sample
Attribute reduction
Training and constructing the genetic neural network
Input the test sample
Classify by BP network patterns
Output the classifier result

Fig. 1 Classifier model of rough set and genetic neural network

3 Data mining classification algorithm

When the classifier model is constructed, then data mining algorithm based on rough set neural network is formed. The algorithm is formed by two important parts, namely rough set attribute reduction module and genetic algorithm to optimize the neural network module.

3.1 Rough set attribute reduction

Before the neural network learning, it is necessary to attribute reduction high dimensional data by rough sets algorithm, eliminate impracticably information, simplify the neural network input, and improve the real-time and detection efficiency.

3.1.1 Continuum attribute discretization algorithm

The discretization method that reduces the number of attribute values of sensor data can reduce the complexity of the problem and improve the fitness of knowledge. We use universal unsupervised equidistance partition algorithm[6] to make continuous data better converted into discrete data.

3.1.2 The rough set attribute reduction algorithm

The division of the class merging rough set attribute reduction algorithm[8,9], improved algorithm procedure, is as follows:

Step 1 Seek the domain condition attribute and decision-making classification:

$U/C = \{X_1, \cdots, X_m\}, U/D = \{Y_1, \cdots, Y_n\}$;

Step 2 Ask out positive region of decision attribute D (or $POS_D(X)$ is positive region): $\{X_1, \cdots, X_m\}$, where $r \leqslant m$;

Step 3 For $\{X_1, \cdots, X_m\}$, according to whether it belong to positive region of the decision to classify, and to find out all kinds of representative cell $\{x'_1, \cdots, x'_m\}$;

Step 4 The representative cells $\{x'_1, \cdots, x'_m\}$ compose new decision information systems $S' = \{U', C \cup D, V, f\}$;

Step 5 In the positive region of all decision, seek the min properties subset where interior representative cells can merge and exterior ones can not merge. In the minus region, seek the min properties subset where interior representative cells can merge while positive region representative cells cannot. Above two kinds of attributes subset is all reduction attribute of conditions attributes D.

3.2 Genetic algorithm to optimize the neural network

We use the genetic operation (selection, crossover and mutation) to find the optimal BP neural network weights and threshold value, and then use BP neural network model for local optimization, so as to get global optimum weights and threshold of the BP neural network.

Here the following steps to achieve the optimization of genetic algorithm for neural network connection weights and threshold value.

Step 1 BP neural network weights, threshold value coding and initial population.

All weights and threshold value of optimization of the neural network are as a group of chromosomes of genetic algorithm, and using the floating-point coding to code them respectively. Each individual encoded is a chromosome, which length is $L = n * h + h * m + h + m$, where h is hidden layer node number, m is output number, and n is input number.

Step 2 The determination of fitness function.

One important performance of BP network performance is the error square sum of the output value of the network and the expected output value, if the error square sum is the minimum then it said the network performance is good. So fitness function can be defined as equation (5):

$$f = C - e = C - \sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \qquad (5)$$

Where C is a constant, $\hat{y}_i$ is the expected output of the sample i. y_i is the actual output of sample i, and N is the number of learning sample.

Step 3 Selection, crossover and mutation operations to get new population.

Selections are the operations of selecting individuals, whose fitness is high, and eliminating individuals of poor quality from the population. Selection operators use classical " roulette" algorithm, which is shown as equation (6).

$$p_i = f_i / \sum_{i=1}^{n} f_i \ , \qquad (6)$$

where n is the population scale, f_i is the fitness of the ith individual.

This operation uses the method of real linear cross because chromosome individual is coded by real data. Let α_1 and α_2 be the parent individuals, so the son individuals after cross β_1, β_2 are generated by equation (7).

$$\begin{cases} \beta_1 = \lambda\alpha_1 + (1-\lambda)\alpha_2 , \\ \beta_2 = \lambda\alpha_2 + (1-\lambda)\alpha_1 , \end{cases} \qquad (7)$$

where $\lambda \in [0,1]$ is a parameter and it is changing with evolution generations.

Variation is the operation change of the gene value of some individual string in chromosome group. Individual α is generated by equation (8):

$$\alpha = \begin{cases} \alpha + r_2(\alpha - \alpha_{\max})(1 - g/g_{\max}), \theta \geqslant 0.5 , \\ \alpha + r_2(\alpha_{\min} - \alpha)(1 - g/g_{\max}), \theta < 0.5 , \end{cases} \qquad (8)$$

where θ is a random number and the value is between 0 and 1, r_2 is a random number too, $\alpha_{\max}$ and $\alpha_{\min}$ are up bound and low bound of gene value of chromosome. g is the current generations and $g_{\max}$ is the max generations.

After the operations of selection, crossover and mutation, new generation of population is achieved and the specified genetic algebra is turned to Step 4, or turned to Step 2.

Step 4 The individual genes value got by GA is coded as initial weights and threshold of BP neural network.

Step 5 BP network spreads positively, Positive calculation hidden units and the output of the output layer were made. If the error of output unit meets the demand of precision, then it prove that the initialized weights and threshold value are the best, so finish that the network training, or turns to Step 6.

Step 6 BP network spread k back propagation, Reverse adjust hidden layer to the output layer, input layer to hidden connection weights and threshold were made, then turns to Step 5.

According to the above algorithm process, the flow chart of genetic algorithm (GA) of the BP neural network is shown as figure 2.

4 Simulation experiment

Based on literature [10] 20 mine environmental sample data were as an original data sets, using MATLAB genetic algorithm packets and neural network relevant function programming tool kit realized BP algorithm. And the rough set BP neural network algorithm was based on rough sets and genetic neural network algorithm (Table 1). The BP neural network has three layers of error back propagation prior BP neural network, its structure is 8-5-1; GA parameter is that weight initialization space is [1, 1], the initial population size is 50, genetic algebra is 180, studying accuracy is 10-6, and other parameters are default values. In literature[10] original data set containd 500 samples data through discrete. Then the 500 sample data set is divided into 300 training data sets and 200 test data sets.

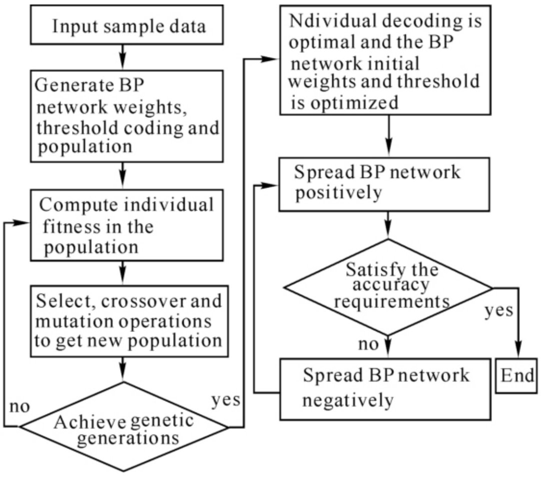
Input sample data
Generate BP network weights, threshold coding and population
Compute individual fitness in the population
Select, crossover and mutation operations to get new population
Achieve genetic generations
no
yes
Ndividual decoding is optimal and the BP network initial weights and threshold is optimized
Spread BP network positively
Satisfy the accuracy requirements
yes
no
Spread BP network negatively
End

Fig. 2 Flow chart of optimized BP neural network by genetic algorithm

Table 1 Result of experiments

Performance	Samples identificated	Times of network traning	Correct rate (%)
BP network	30	98	75
Rough set BP network	36	51	90
Genetic neural network Based on Rough set	38	25	95

After a period of time of MATLAB engine starting, BP network can identify wireless sensor data, because the convergence speed of the BP network is slow, and it needs a long training time. Using rough set simplified the BP neural network can reduce training time greatly. And the BP neural network based on rough set not only contract with the dimensions of input data, and combined with the genetic algorithm, so the time needed for training is short, and responsed immediately for test data.

(下转第 136 页 Continue on page 136)

```
        x = a11 * v_cu.x + a12 * v_cu.y + a13 * v_cu.z
        y = a21 * v_cu.x + a22 * v_cu.y + a23 * v_cu.z
        z = a31 * v_cu.x + a32 * v_cu.y + a33 * v_cu.z
        centroid = vector3d.Vector3d(x,y,z)
        return centroid
```

函数参数 *a*,*b*,*r*,*v_cu* 分别表示欧拉角(*a*,*b*,*r*)和残基重心在球坐标下的直角坐标．函数返回残基重心的直角坐标，这个坐标和 PDB 坐标同型，但原点不同．

将所得的残基重心坐标 centroid 经坐标平移后即可与 PDB 数据的坐标一致．另外，代码中使用到 python 的常用的数学函数标准库 math，在源代码的最开始处必须先导入：import math．

本文只给出蛋白质残基重心计算方法在 python 语言下的实现过程，还可以有其他的实现方法，甚至于不同的编程语言，如 C、C++等．一般来讲，我们可根据不同的要求进行选择．

参考文献：

［1］ Dehouck Y，Gilis D，Rooman M. A new generation of statistical potentials for proteins［J］. Biophys J，2006，90(11)：4010-7.

［2］ Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding［J］. J Mol Biol，1976，104(1)：59-107.

［3］ Bernstein F C，Koetzle TF，Williams GJB，et al. The Protein Data Bank：a computer-based archival file for macromolecular structures［J］. J Mol Biol，1977，112(3)：535-42.

［4］ Kocher J P A，Rooman M J，Wodak S J. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches［J］. Journal of Molecular Biology，1994，235(5)：1598-1613.

［5］ Betancourt M R. A reduced protein model with accurate native－structure identification ability［J］. Proteins：Structure，Function and Genetics，2003，53(4)：889-907.

［6］ Lutz M. Programming python［M］. 4th Edition. O'Reilly Media，2010.

(责任编辑：尹　闯)

(上接第 131 页 Continue from page 131)

To sum up，BP network based on rough set genetic algorithm has high accuracy identification，its speed is fast，and generalization ability is strong，therefore，it is a promising classification method.

References：

［1］ Zhang J M，Song Y Q，Zhou S W，et al. Survey on data aggregation techniques in wireless sensor networks［J］. Computer Applications，2006，26(6)：1273-1283.

［2］ Yuan J B，Pu H C. E-mail information classifier of neural network based on gennetic algorithm Optimization［J］. Nanjing University of Science and Technology：Natural Science，2008，32(1)：78-81.

［3］ Li S，Liu L J，Xie Y L. Chaotic prediction for short-term traffic flow of optimized BP neural network based on genetic algorithm［J］. Control and Decision，2011，26(10)：1581-1585.

［4］ Wang Z W. Improved BP neural network-based license plate character recognition［J］. Microelectronics and Computer，2011，28(9)：66-69.

［5］ Wang H T，Huang Z H，Fang X G，et al. The application of rough sets － neural network theory to mine ventilation system evaluation［J］. Chongqing University ，2011，34(9)：90-94.

［6］ Hong Y H. Stdudy on distributed data mining algorithm in wireless sensor networks［J］. Computer Simulation，2012，29(12)：167-170.

［7］ Nie Y，Chen F J. Research of K-mean clustering algorithm based on rough set［J］. Engineering Journal of Wuhan University，2011，44(2)：257-260.

［8］ Zheng W，Zhou Z Q ，Ma Y L. A feature selection approach based on rough set theory［J］. Hebei North University：Natural Science Edition，2009，5(1)：56-59.

［9］ Luo G Z，Yang X J. Rough attributes reduction algorithm based on partitioning strategy merge［J］. Statistics and Decision，2009(20)：146-148.

［10］ Cao L. Research on the colliery safety wireless monitoring system basedon multisource information fusion teehnique［D］. Xian：Xian University of Architecture and Technology，2008.

(责任编辑：尹　闯)