文章编号: 1001-0920(2005)07-0782-04

粗糙集理论框架下的神经网络建模研究及应用

何 明, 李 博, 马兆丰, 傅向华
(西安交通大学 计算机科学与技术系, 陕西 西安 710049)

摘 要: 为协调决策支持和分类, 引入了一种新的方法, 该方法将粗糙集理论和神经网络有机地结合在一起, 提出了一种基于粗糙集理论的神经网络模型构造方法. 首先, 利用粗糙集理论智能数据分析的能力, 对神经网络进行预处理, 抽取关键成分作为神经网络的输入, 从而确定粗糙神经网络的初始拓扑结构. 在此基础上, 进一步研究和分析了该模型的实现步骤, 并应用原始数据对网络进行训练, 最后将该模型应用于分类规则的抽取. 试验结果比较表明, 该模型可以有效地提高分类的精度.

关键词: 神经网络; 粗糙集; 分类; 数据挖掘
中图分类号: TP18 **文献标识码**: A

On the Neural Network Modeling with Support Rough Set Theory

HE Ming, LI Bo, MA Zhao-feng, FU Xiang-hua
(Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China. Correspondent: HE Ming, E-mail: ming.he1314@sina.com)

Abstract: A method is introduced to cooperate the decision support and classification, in which rough set theory and neural network are formed integrated into a model. A neural network modelling way based on rough set theory is proposed. The neural network is preprocessed by the intelligent data analysis capability of rough set theory, and the key components are extracted as the inputs of the neural network to determine original topology of the rough neural network. Furthermore, the realization steps of the model are analyzed, and the rough neural network is trained with original data. The model constructed is applied to extraction of classification rules. The experimental results show that the model can increase the classification correctness effectively.

Key words: Neural network; Rough set; Classification; Data mining

1 引 言

粗糙集理论[1]是一种刻画不完整和不确定性问题的数学工具, 近年来在模式识别、机器学习、故障诊断、知识获取与发现、决策分析与支持等领域取得了较为成功的应用[2~4].

本文将粗糙集和神经网络结合在一起, 构成粗糙神经网络模型, 并用于分类规则的抽取. 在保留人工神经网络自学习、自组织特性的基础上, 利用粗糙集理论对数据进行预处理, 提取其中关键要素作为网络的输入, 从而简化了神经网络的结构, 提高了分类精度.

2 粗糙集智能数据分析[5](RSIDA)

2.1 基本概念

基于粗糙集的智能数据分析主要是用来分析信息系统中各属性之间的依赖关系, 它是粗糙集理论的一个主要应用技术. 设一个信息系统 S 可以表示为 $S = \langle U, A, V, f \rangle$, 其中: U 是非空有限对象的集合, 即 $U = \{x_1, x_2, \cdots, x_n\}$, 也称为论域; A 是属性集合; $V = \bigcup_{a \in A} V_a$, V_a 表示属性 a 的值域; $f: U \times A \to V$ 是一个信息函数, 它指定 U 中每一个对象 x 的属性值, 即对 $x \in U$, $a \in A$, 有 $f_a(x) \in V_a$. 如果属性集 A 可以分为条件属性集 C 和决策属性集 D, 即 C

收稿日期: 2004-07-14; **修回日期**: 2004-09-09.
基金项目: 国家高技术研究发展计划项目(2003AA1Z2610).
作者简介: 何明(1975—), 男, 陕西礼泉人, 博士生, 从事数据挖掘、机器学习的研究; 李博(1981—), 男, 河南洛阳人, 硕士生, 从事数据挖掘、智能网络等研究.

$\cup D = A$, $C \cap D = \varnothing$, 则该信息系统称为决策系统或决策表, 简记为 $S = (U, C, D)$, 其中 D 一般只含有一个属性, 记为 d.

定义 1　$x, y \subseteq U$, 对于 $Q \subseteq A$, θ_Q 是 U 上的一个等价关系, 如果满足 $x\theta_Q y \Leftrightarrow (\forall q \in Q)(f_q(x) = f_q(y))$, 则称 θ_Q 是 x, y 的一个不可分辨关系

定义 2　设 $P, Q \subseteq A$, 如果等价关系 θ_Q 定义的每个等价类都属于等价关系 θ_P 定义的等价类, 则称 P 依赖于 Q, 记作 $Q \rightarrow P$. 依赖关系 $Q \rightarrow P$ 表达了如下规则: 假设 $Q = \{q_1, q_2, \cdots, q_n\}$, $P = \{p_1, p_2, \cdots, p_k\}$, 对每一个 $t = \{t_1, t_2, \cdots, t_n\}$, $t_i \in V_{q_i}$, 唯一决定了属性值集合 $s = \{s_1, s_2, \cdots, s_k\}$, $s_i \in V_{p_i}$, 即 $f(x, q_1) = t_1, \cdots, f(x, q_n) = t_n) \Rightarrow (f(x, p_1) = s_1, \cdots, f(x, p_k) = s_k)\ \forall x \in U$. 通过 RSDA, 在保持 $Q \rightarrow P$ 成立的前提下, 可以得到规则最小化简

定义 3　给定信息系统 $S = \langle U, A, V, f \rangle$, 设 $B \subseteq A$, $X \subseteq U$, 则 X 关于 B 的下、上近似集分别定义为

$$B_-(X) = \cup \{Y \in U/B,\ Y \subseteq X\},$$
$$B^-(X) = \cup \{Y \in U/B,\ Y \cap X \neq \varnothing\}.$$

定义 4　粗糙隶属函数[6] (RMF) 元素 $u \in U$ 在关系 R 下对集合 X 的粗糙隶属函数为

$$\mu_X^R(u): U \rightarrow [0, 1] = \frac{|[u]_R \cap X|}{|[u]_R|}. \tag{1}$$

其中: $|\cdot|$ 表示集合中元素的个数, $[u]_R$ 为包含元素 u 的等价类, $0 \leq [u]_R \leq 1$.

2.2　规则的匹配度和适用度

根据数据本身的信息, 利用 RSDA 对数据进行约简, 从原始数据集中抽取 m 条 $Q \rightarrow P$ 规则, 其中第 i 条规则 R^i 为

$$\text{if } f(x, q_1) = t_1^i, \cdots, f(x, q_n) = t_n^i,$$
$$\text{then } f(x, p_1) = s_1^i, \cdots, f(x, p_r) = s_r^i.$$

其中: $t_j^i \in V_{q_j}$, $s_k^j \in V_{p_k}$, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n$, $k = 1, 2, \cdots, r$. 对于一组输入 Input$\{In_1, In_2, \cdots, In_n\}$, 定义规则最大匹配函数

$$M_i = 1 - \min \frac{\| c_{exp} - c_i \|}{\| c_i \|},\ i = 1, 2, \cdots, m. \tag{2}$$

其中: c_{exp} 是根据输入 Input$\{In_1, In_2, \cdots, In_n\}$ 而构造的条件向量, c_i 是根据粗糙集理论抽取的第 $i (i \leq m)$ 条规则的条件向量

从原始数据中抽取的规则可靠程度是不同的, 可用粗糙隶属函数来表示规则的可靠程度, 并和匹配度相结合得出规则的适用度. 根据粗糙隶属函数的定义, 对于第 i 条规则第 j 个属性值相对于结论等价类 X 的粗隶属函数为

$$\mu_X^{q_j}(t_j^i) = \frac{|X \cap [t_j^i]_{q_j}|}{|[t_j^i]_{q_j}|},\ X = [s_1^i, s_2^i, \cdots, s_r^i]_P,$$
$$i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, n. \tag{3}$$

$\mu_X^{q_j}$ 越大说明由属性 t_j^i 推出结论的可能性越大. 特别地, $\mu_X^{q_j} = 1$ 说明当 $f(x, q_j) = t_j^i$ 时, 结论肯定成立. 输入 Input 对于第 i 条规则的适用度 $\mu_i = \max(\mu_X^{q_j}(t_j^i) M_i)$.

3　粗糙神经网络的结构和学习算法

3.1　粗糙神经网络的基本结构

本文构造的粗糙神经网络的结构如图 1 所示

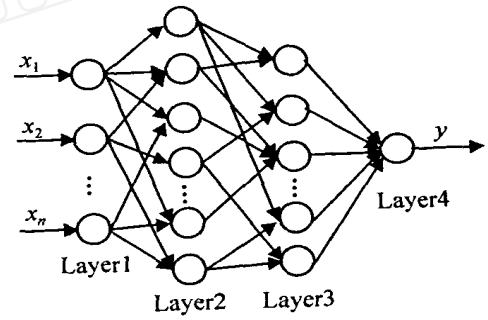

图 1　粗糙神经网络模型

该网络分为 4 层:

第 1 层: 输入层, 它的值为实际的精确值, 表示输入向量 $x = (x_1, x_2, \cdots, x_n)^T$.

第 2 层: 隶属度函数层, 分别将 n 个输入分量 $(x_1, x_2, \cdots, x_n)$ 依照某种不可分辨关系进行划分, 确定其与每个相应分类之间的联系; 将每一个输入分量离散化为 r_i 个不同的值, 这些值在 $[0, 1]$ 之间. 与第 t 个输入节点相连结的一组神经元的作用是对输入向量的第 t 个分量进行解释. 本文定义该层神经元的作用函数为粗糙隶属函数. 神经元 N_{jt} 的输出为 $N_{jt}^{out} = \mu_{A_{jt}}(x_t)$. 其中: N_{jt} 是与第 t 个输入节点相连结的第 j 个神经元, 其含义是与第 t 个分量相关的分类中的第 j 个等价类; A_{jt} 为 x_t 与 N_{jt} 间的粗糙隶属函数值连结权, 含义为第 j 个类所代表的等价类

第 3 层: 推理层, 该层的每个节点代表一条规则, 这些规则是通过粗糙集理论得到的. 假设有 $m (m \leq n)$ 条规则, 该层节点的作用函数为

$$\pi_i = \mu_{1i} \cdot \mu_{2i} \cdot \cdots \cdot \mu_{ni} = \prod_{j=1}^{n} \mu_{ji},\ 1 \leq i \leq m. \tag{4}$$

第 4 层: 清晰化层, 这一层的节点代表输出变量. 在多输入单输出系统中, 该层的节点数为 1, 权值 ω 的初始值预先设为各规则粗糙隶属度值, 该层节点的输出为

$$y = \sum_{i=1}^{m} \omega \pi_i \tag{5}$$

3.2 粗糙神经网络的 BP 算法

学习的目的是对网络的连接权值进行调整, 使得对任一输入都能得到所期望的输出. 设 i 与 j 的上下神经元是完全连接方式, $f(x)$ 为传递函数, 那么粗糙神经元 r 的输入、输出可计算如下[7]:

$$\begin{aligned}&\text{input}_j = \sum \omega_j \text{output}_i,\\ &\text{Output}_r^- = \max(f(\text{input}_r^- - \theta_j^-), f(\text{input}_r^+ - \theta_j^+)),\\ &\text{Output}_r^+ = \min(f(\text{input}_r^- - \theta_j^-), f(\text{input}_r^+ - \theta_j^+)).\end{aligned} \tag{6}$$

为简化问题, 只讨论输出层有一个粗糙神经元[7~9] 的情形, 对应于任一输入模式 k 和输出神经元 p 的实际输出为 $o_p = (o_p^-, o_p^+)$. 其中: o_p^-, o_p^+ 分别为粗糙神经元 p 中上、下神经元的实际输出; $\hat{o}_p^-, \hat{o}_p^+$ 为神经元 p 中上、下神经元的期望输出. 为了使学习以尽可能快减小误差的方式进行, 对误差的计算采用广义的 δ 规则. 定义粗糙神经网络误差函数为

$$E^k = \frac{1}{2}[(\hat{o}_p^- - o_p^-)^2 + (\hat{o}_p^+ - o_p^+)^2],$$

如果样本数为 m, 则输入全部样本后的总误差函数定义为

$$E = \frac{1}{2}\sum_{k=1}^{m} E^k = \frac{1}{2}\sum_{k=1}^{m}[(\hat{o}_p^- - o_p^-)^2 + (\hat{o}_p^+ - o_p^+)^2]. \tag{7}$$

连接权值的修改由下式计算:

$$\omega_j(t+1) = \omega_j(t) + \eta \times \text{output}_i \times \text{error}_j \times f'(\text{input}_i).$$

这里: $f'(\text{input}_i)$ 是传递函数的导数且取 $f'(\text{input}_i) = 1/(1+e^{-x})$, $\text{error}_j = \hat{\text{output}}_j - \text{output}_j$, η 为学习速率.

4 建模方法

4.1 模型框架

模型框架如图 2 所示.

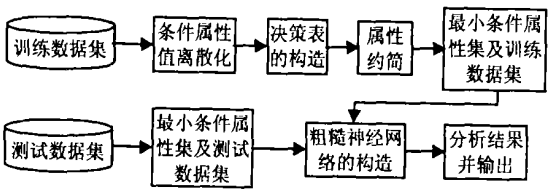

图 2 粗糙神经网络系统框架

4.2 建模过程

1) 数据离散化处理

系统的输入输出数据可能是连续的或是离散的, 在进行 RSDA 之前首先应将连续的数据离散化. 对连续的数据进行适当的区间划分, 并将划分结果用 1, 2… 表示. 本文采用 Rosett[10] 系统对训练数据集进行离散化处理, 经过离散化处理的数据可以用粗糙集数据分析方法进行分析.

2) 基于粗糙集的数据分析

① 数据过滤. 数据过滤作为 RSDA 的准备工作, 它的目的是过滤掉对所有规则都不必要的属性值, 在保持原有知识完备的前提下, 消除冗余的属性和属性值, 减少特征空间.

② 决策表属性约简, 消除冗余的属性和属性值, 得到最优规则集[11].

③ 计算各条规则中的粗隶属度.

④ 粗糙神经网络的构造和训练: 将 ① 中离散化数据按结论分为 m 类, 即 $\{D_1, D_2, \cdots, D_m\}$. 设 X_i 是所有结论为 D_i 的对象的集合 $X_i \subset U$, 这些对象构成了 m 个子网的输入输出, 利用这些数据进行子网训练, 分析并输出结果.

5 试验结果

为验证本文提出模型的有效性, 实验中采用 UCI 机器学习数据库[12] 中的 3 个数据集, 表 1 给出了这些样本集的概要信息.

表 1 数据集概要信息

样本集	属性数目	数值属性数目	类别数目	样本数目
Iris	4	4	3	150
Glass	9	9	6	214
Zoo	16	1	7	101

对于数据集 Iris, 首先采用 C5.0[13] 来产生规则集, 然后与本文提出的粗糙神经网络模型(MRNN) 产生的规则集进行比较. MRNN 生成的规则与 C5.0 具有相同的形式, 所不同的是: 对于 Rule_2, 属性 Petal-Length 右边界的值变为 4.87; 对于 Rule_4, 属性 Petal-Length 左边界的值变为 4.75. 属性值的改变导致了规则分类精度的提高.

表 2 是 C5.0 和 MRNN 规则在训练集和测试集上的精度. 表 2 的结果表明, 采用本文提出的 MRNN 获取的规则精度较 C5.0 要高. 图 3 是采用 C5.0 和 MRNN 分别对 3 个数据集抽取分类规则得到实验结果. 从图 3 可以看出, 将 MRNN 应用于这 3 个数据集, 得到的结果优于 C5.0.

表 2 在 Iris 数据集上的 C5.0 和 MRNN 抽取规则的精度比较

	C5.0 规则集		MRNN 规则集	
	训练集	测试集	训练集	测试集
Virginica	25/25	24/25	25/25	25/25
Setosa	25/25	25/25	25/25	25/25
Versicolor	24/25	23/25	24/25	24/25
总计	74/75	72/75	74/75	74/75
	98.67%	96%	98.67%	98.67%

注: 分母为所用例子的总数, 分子为分类正确的例子数.

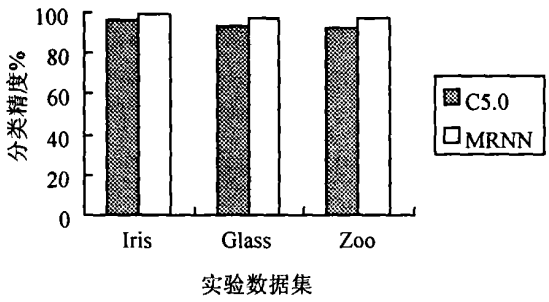

图 3　分类结果比较

6　结　　语

本文提出了一种基于粗糙集的神经网络模型, 该模型综合了粗糙集理论在知识获取方面的能力和神经网络在数值逼近上的优势. 通过粗糙集智能数据分析, 消除初始决策表中的冗余信息和噪声数据的干扰, 从而抽取针对原始数据的最简规则, 减少了粗糙神经网络中输入层和代表规则层的神经元个数, 简化了神经网络的拓扑结构, 提高了系统的速度. 另外, 神经网络训练结果在数值上逼近原系统, 但它依然是一个"黑箱系统", 各参数的物理意义不明确, 训练结束后, 我们对系统的了解依然很少. 而本文提出的方法在输入输出逼近的同时还得出了一些有效的规则, 使我们对系统本身有了一定的认识, 这一点是ANN 建模所不能完成的. 最后通过试验结果对比验证了该模型的有效性.

参考文献(References)

[1] Pawlak Z. Rough Sets [J]. *Int J of Computer and Information Science*, 1982, 11(5): 341-356.

[2] Pawlak Z. *Rough Sets: Theoretical Aspects of Reasoning About Data* [M]. Boston: Kluwer Academic Publishers, 1991.

[3] Skowron A. *Rough Setsand Boolean Reasoning* [M]. New York: Physica Verlag, 2001: 95-124.

[4] Skowron A, Peters J. Rough Sets: Trends and Challenges [A]. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* [C]. Berlin: Springer-Verlag, 2003: 25-34.

[5] Pawlak Z. Rough Sets and Intelligent Data Analysis [J]. *Information Science*, 2002, 11(147): 1-12.

[6] Wu Y W, Zhang Chang-N. A Rough Neural Network for Material Proportioning System [A]. *Communications, Circuits and Systems and West Sino Expositions, IEEE 2002 International Conference* [C]. Chendu, 2002, 2: 1189-1193.

[7] Lingras P J. Rough Neural Networks [A]. *Proc of Sixth Int Conf on Information, Processing and Management of Uncertainty in Knowledge-Based Systems* [C]. Granada, 1996: 1445-1450.

[8] Perters J F, Han L, Ramanna S. Rough Neural Computing in Single Analysis [J]. *Computational Intelligence*, 2001, 17(3): 493-513.

[9] Pedrycz W. *Computational Intelligence: An Introduction* [M]. Florida: CRC Press, Boca Raton, 1998.

[10] Rosetta. *A rough Set Toolkit for Analyzing Data* [EB/OL]. http://www.idi.ntnu.no/~aleks/rosetta, 2002.

[11] 何明, 冯博琴, 马兆丰, 等. 基于增量式遗传算法的粗糙集分类规则挖掘[J]. *西安交通大学学报*, 2004, 38(6): 579-582.
(He M, Feng B Q, Ma Z F, et al. Incremental Genetic Algorithm Based Data Mining Method for Discovering Rough Set Classification Rules [J]. *J of Xi'an Jiaotong University*, 2004, 38(6): 579-582.)

[12] Blake C, Keogh E, Merz C J. *UCI Repository of Machine Learning Data Tables* [EB/OL]. http://www.ics.uci.edu/~mlearn, 2004-08-16.

[13] Rule Quest Research. C5.0: Effective Data Mining Tools [EB/OL]. http://www.rulequest.com. 2004-08-16.

(上接第 781 页)

[4] 金波, 俞亚新. 一种自适应CMAC 神经元网络控制器及其在水轮机调速器中的应用[J]. *控制理论与应用*, 2002, 19(6): 905-908.
(Jin B, Yu Y X. Adaptive CMAC Controller for Hydraulic Turbine Speed Governor [J]. *Control Theory and Application*, 2002, 19(6): 905-908.)

[5] 周旭东, 王国栋, 李淑华. 小脑模型控制系统的遗传算法最优设计[J]. *信息与控制*, 1997, 26(6): 455-458.
(Zhou X D, Wang G D, Li S H. Genetic Algorithm Based Optimal Design for CMAC Controller [J]. *Information and Control*, 1997, 26(6): 455-458.)

[6] 李世敬, 王解法, 冯祖仁. 层叠CMAC 补偿的并联机器人变结构控制研究[J]. *系统仿真学报*, 2002, 14(8): 1045-1048.
(Li S J, Wang J F, Feng Z R. Variable Structure for Parallel Manipulators Based on Cascaded CMAC [J]. *J of System Simulation*, 2002, 14(8): 1045-1048.)

[7] 李辉, 杨顺昌. 可调速双馈水轮发电机组控制系统的稳定性分析[J]. *中国电机工程学报*, 2004, 24(6): 152-154.
(Li H, Yang S C. Stability Analysis of Control System of Adjustable Speed Hydroelectric Generating Units with Doubly Fed Generators [J]. *Proc of the CSEE*, 2004, 24(6): 152-154.)