第 28 卷 第 1 期
2002 年 1 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, No. 1
Jan., 2002

基于粗糙集的神经网络建模方法研究[1)]

黎 明 张化光

(东北大学信息科学与工程学院 沈阳 110006)

摘 要 提出了一种基于粗糙集的神经网络模型,该方法利用粗糙集数据分析方法,从数据中提取出规则将输入映射到输出的子空间上,而后在这个子空间上用神经网络进行逼近. 利用这种方法对岩石边坡工程中边坡稳定性进行分析建模,并和传统的神经网络建模方法进行比较,说明了该方法的有效性.

关键词 粗糙集,人工神经网络,粗糙集数据分析

中图分类号 TP18

RESEARCH ON THE METHOD OF NEURAL NETWORK MODELING BASED ON ROUGH SETS THEORY

LI Ming ZHANG Hua-Guang

(*School of Information Science and Engineering, Northeastern University, Shenyang* 110006)

Abstract This paper proposes a new neural network model based on rough sets. The input to the model is mapped into the output subspaces by using rules acquired from rough sets, then the output of the system is approximated by neural network in the subspaces. The method is applied to analysis of the stability of the rock slope. Simulation results show this method is superior to the traditional neural network model.

Key words Rough sets, artifical neural network, rough set data analysis

1 引言

粗糙集(Rough Sets, RS)理论是由波兰学者 Pawlak 于 1982 年提出的[1],它作为一种刻化具有不完整性和不确定性的信息全新的数学工具,已经成为人工智能领域的一个新的学术热点,国外已经利用其取得不少成果,国内研究尚处于起步阶段[2,3].

糙集数据分析(Rough Sets Data Analysis, RSDA)是一种以 RS 为基础的,分析数据之间相关性和依赖性的一种符号方法. 利用 RSDA 可以从数据中提取规则,从而进行预测和决策. 但是粗糙集模型是一种结构化的,非数值化的信息处理方法,适于处理离散数据,对于连续数据的处理能力有限[4]. 目前基于粗糙集的数据分析方法多是针对离散数据的,尤其在

1) 沈阳市科委基金(199951022-00)、国家教委博士点基金资助.

收稿日期 1999-10-25 收修改稿日期 2000-06-10

结论部分只是定性的.

人工神经网络(Artifical Neural Network, ANN),作为人工智能领域的一个重要分支,它最吸引人的一点是其数值逼近能力.它能够处理定量的、数值化的信息,较 RSDA 而言,ANN 能够得出更精细的结果.但通常 ANN 的训练过程是非常复杂和漫长的.文献[5]将 RS 与 ANN 结合起来,利用 RS 来简化 ANN 的训练样本,在保留重要信息的前提下消除冗余的数据,以提高训练速度.本文提出的建模方法是从定性分析和定量分析这个角度,将 RS 和 ANN 结合起来.利用 RSDA 从数据中提取出规则,根据输入确定输出可能的子空间,再在这些子空间上通过 ANN 进行定量的逼近.通过定性和定量的分析建立整个系统的模型.利用这种方法我们从一组实际数据出发建立了边坡稳定性分析的模型,并和文献[6]中的 ANN 模型进行比较,说明了该方法的有效性.

2 粗糙集数据分析

粗糙集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领域.设信息系统 $I=\langle U,\Omega,V_q,f\rangle$, $q\in\Omega$, 其中 U 为全体对象的集合,Ω 为属性集合,V_q 为属性值集合,f 为信息函数,$f_q:U\to V_q$.

定义 1[7]. $x,y\in U$, 对于 $Q\subseteq\Omega$, 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[7]. 设 $P,Q\subseteq\Omega$, 如果等价关系 θ_Q 定义的每一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:$Q\to P$.

依赖关系 $Q\to P$ 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 $t=\{t_1,t_2,\cdots,t_n\}$,其中 $t_i\in V_{q_i}$,唯一决定了属性值集合 $s=\{s_1,s_2,\cdots,s_k\}$,其中 $s_i\in V_{p_i}$,即$(\forall\ x\in U)[(f(x,q_1)=t_1,\cdots,f(x,q_n)=t_n)\Rightarrow f(x,p_1)=s_1,\cdots,f(x,p_k)=s_k)]$.

如何在保持 $Q\to P$ 成立的前提下,得到规则的最小化简,是我们所关心的问题,也正是 RSDA 所要解决的问题.

定义 3[2]. 粗糙隶属函数(Rough Membership Function). 元素 α 在关系 R 下对集合 X 的粗糙隶属函数为

$$\mu_X^R(\alpha)=\frac{|X\cap[\alpha]_R|}{|[\alpha]_R|}\tag{1}$$

其中 $|\cdot|$ 表示集合中元素的个数,$[\alpha]_R$ 为包含元素 α 的等价类.显然,$0\leqslant\mu_X^R(\alpha)\leqslant 1$.

例. 有 7 个物体,可以用颜色(C)、形状(S)、体积(L)和重量(W)来描述. 则可用如下信息系统来描述. $I=\langle U,\Omega,V_q,f\rangle$, $q\in\Omega$, 其中 $U=\{x_1,x_2,\cdots,x_7\}$,$\Omega=\{\mathrm{C},\mathrm{S},\mathrm{L},\mathrm{W}\}$, $V_C=\{1(红),2(黄),3(绿)\}$, $V_S=\{1(三角形),2(方形),3(圆形)\}$,$V_L=\{1(大),2(小)\}$, $V_W=\{1(轻),2(适中),3(重)\}$, f 完成了用语言描述的属性值到数值描述的关系,如把“黄色”用数值“2”表示.表 1 为该信息系统的决策表.由等价关系 θ_R 所决定的等价类集合用 U/R 表示.

表 1 决策表

	C	S	L	W
1	1	1	1	1
2	2	2	2	3
3	1	3	1	1
4	2	3	1	2
5	3	1	1	2
6	3	2	1	1
7	1	3	2	3
8	3	3	2	3

$$U/C = \{\{x_1,x_3,x_7\},\{x_2,x_4\},\{x_5,x_6,x_8\}\};$$
$$U/S = \{\{x_1,x_5\},\{x_2,x_6\},\{x_3,x_4,x_7,x_8\}\};$$
$$U/L = \{\{x_2,x_7,x_8\},\{x_1,x_3,x_4,x_5,x_6\}\};$$
$$U/W = \{\{x_1,x_3,x_6\},\{x_4,x_5\},\{x_2,x_7,x_8\}\};$$

由于 U/W 中的等价类属于 U/L 中的等价类,由定义 2 知,$W \to L$,直观的理解为该信息系统中物体的体积和重量有关,重量重的物体体积大.

现给定一个集合 $X=\{x_i \mid f_L(x_i)=2\}=\{x_2,x_7,x_8\}$,重量属性值为 3 的集合$[3]_w=\{x_2,x_7,x_8\}$,则重量属性值为 3 的对象属于集合 X 的隶属度为

$$\mu_X^w(3)=\frac{|X \cap [3]_w|}{|[3]_w|}=1,$$

这表明重量重的物体体积一定大.

若设集合 $X=\{x_i \mid f_w(x_i)=1\}=\{x_1,x_3,x_6\}$,体积属性值为 1 的集合$[1]_L=\{x_1,x_3,x_4,x_5,x_6\}$,则体积属性值为 1 的对象属于集合 X 的隶属度为

$$\mu_X^L(1)=\frac{|X \cap [1]_L|}{|[1]_L|}=0.6.$$

这表明体积小的物体并不一定属于重量轻物体的集合,也就是说物体体积小并不能说明其重量轻.由此,可以看出粗糙隶属度为决策提供了一定的依据,且它是完全来源于数据本身,不需要做人为假设.

3　规则的匹配度和适用度

利用 RSDA 的化简方法,从原始数据中提取出 m 条 $Q \to P$ 规则,其中第 i 条规则 R^i 为

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

其中 $t_j^i \in V_{q_j}, s_k^i \in V_{p_k}$, $i=1,2,\cdots,m$, $j=1,2,\cdots,n$, $k=1,2,\cdots,r$. 对于一组输入 $IN=\{in_1,in_2,\cdots,in_n\}$,定义函数

$$g_i(k)=\begin{cases}1, & in_k=t_k^i\\ 0, & \text{others}\end{cases}, \quad k=1,2,\cdots,n,\ i=1,2,\cdots,m.$$

定义输入 IN 与第 i 条规则的匹配度为

$$C_i=\frac{\sum_{k=1}^{n} g_i(k)}{n}, \quad i=1,2,\cdots,m,$$

它表明了输入模式与第 i 条规则的匹配程度.

规则是从数据中提取出来的,但各个规则的可靠程度是不一样的,这里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得适用度.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

$$\mu_X^{q_j}(t_j^i)=\frac{|X \cap [t_j^i]_{q_j}|}{|[t_j^i]_{q_j}|} \tag{2}$$

其中 $i=1,2,\cdots,m$, $j=1,2,\cdots,n$, $X=[s_1^i,s_2^i,\cdots,s_r^i]_P$. $\mu_X^{q_j}(t_j^i)$越大说明由属性值 t_j^i 推出结论的可能性越大,$\mu_X^{q_j}(t_j^i)=1$ 说明当 $f(x,q_j)=t_j^i$ 时,结论肯定成立.

输入 IN 对于第 i 条规则的适用度 μ_i 为

$$\mu_i = \max\{\mu_X^{q_1}(t_1^i)g_i(1), \mu_X^{q_2}(t_2^i)g_i(2), \cdots, \mu_X^{q_n}(t_n^i)g_i(n)\}.$$

4 建模方法

该建模方法是基于系统的输入输出数据，建模过程分为三步：

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量化. 对于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. 这里采用模糊 C-mean 聚类方法，对输入数据进行聚类，从而实现连续数据的离散化，经过定性化处理的数据可以用粗糙集数据分析方法进行分析.

这里需要注意的一点是，假设所获得的数据是合理的，本身是不矛盾的，即没有输入完全相同而输出相差很远的. 因此聚类以后，形成的决策表应该是一致的. 如果发生不一致的情况，则是由于聚类区间划分不当造成的. 这时需要分析不一致规则产生的原因，并对相应的区间进行再划分或合并，直至所形成的规则完全一致为止.

2) 基于粗糙集的数据分析

a 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，为进一步化简作准备. 文献[4]给出了一种基于二值信息系统的数据过滤方法，并证明了该方法能在不改变信息系统依赖性的前提下，增强规则的统计特性. 本文引用了这种数据过滤方法.

b 化简决策表消除冗余的属性和属性值，得到规则[1].

c 计算各条规则中的粗隶属度.

3) 人工神经网络子网络的训练. 在第一步中，将数据离散化后，按结论可以分为 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，训练算法可采用 BP(Back Propagation)算法，训练结果得到 m 个子网. 其中第 i 个子网的输入输出关系为

$$out_i = fnet_i(in), \quad i = 1, 2, \cdots, m,$$

其中 in 为输入数据，out_i 为神经网输出.

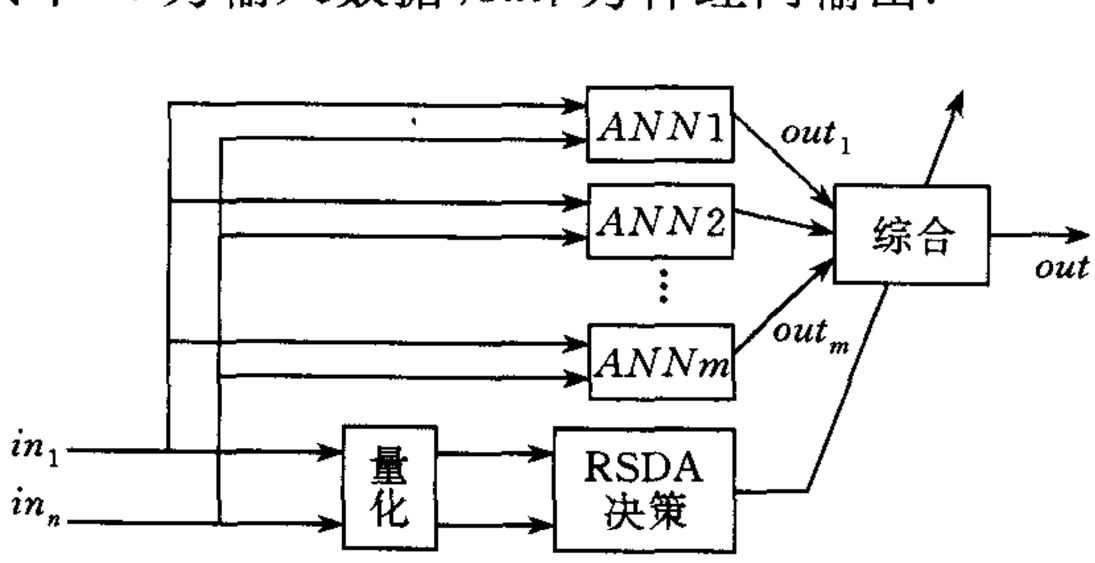

图 1 基于粗糙集的神经网络模型

以上建立了一个基于粗糙集的神经网络模型，其结构如图 1 所示. 利用该模型进行预测和决策过程分为以下三步：

1) 根据建模过程中得到的区间划分对输入数据进行离散化.

2) 利用 RSDA 提取出的规则对输入进行判决. 按照匹配度最大原则，选出最合适的规则，这时规则可能不止一个，设其下标集合 F 为 $F=\{r_1, r_2, \cdots, r_k\}$，对应的决策为 $\{D_{r_1}, D_{r_2}, \cdots, D_{r_k}\}$，分别计算其适用度为 $\{\mu_{r_1}, \mu_{r_2}, \cdots, \mu_{r_k}\}$.

3) 计算系统输出

$$out = \frac{\sum_{i \in F} \mu_i \cdot fnet_{D_i}(in)}{\sum_{i \in F} \mu_i}.$$

5　仿真研究

为了验证本文提出的方法的有效性,我们对一组岩石边坡工程中的历史数据[6]进行分析,建立了边坡稳定性分析模型.该数据收集了世界上 82 个边坡工程中关于边坡稳定性的实例,其中包括 44 个不稳定实例和 38 个稳定实例.经典的结论是由极限平衡法得出的,文献[6]提出了一种基于 ANN 的建模方法.

由数据表可得该信息系统有 6 个条件属性,$Q=\{\gamma,c,\phi,\varphi_f,H,\gamma_u\}$,决策属性有 2 个 $P=\{F,S\}$,$\Omega=\{Q,P\}$.其中各符号含义,γ:岩石容重(吨/米^3),c:粘结力(千帕),ϕ:摩擦角(度),φ_f:边坡角(度),H:边坡高度(米),γ_u:孔隙压力比,F:安全系数,S:边坡状态.

除了 S 是用稳定(Stable)和不稳定(Failure)两个离散值来表示,其余均为连续数据.特别是,结论中安全系数也为定量的连续属性.这在传统的 RSDA 中是不能处理的.为了和文献[6]的方法相比较,这里采用了相同的仿真条件,取前 71 条数据作为训练数据,剩余的 11 条数据作为检验数据.具体步骤如下

1) 离散化处理

首先,对各个连续连续属性进行聚类分析,这里采用 Matlab 5.2 工具箱中的 stepfcm.m 函数,聚类结果得到等价类的区间划分见表 2.经规则的一致性检验后,发现存在矛盾规则,因此需要在聚类的基础上再进一步划分区间,最终在消除矛盾规则后得到较为理想的区间划分.表 2 中用黑体表示的区间为产生矛盾规则的区间和重新划分后的区间.通过区间映射 f 将连续的定量值映射到一个整数集合上,构成了属性值集合 V_q,其元素为属性值,由此形成了一个信息系统的决策表.下面利用粗糙集数据分析对这个决策表进行化简,从中提取出有用的规则.

表 2　各连续属性的等价类划分区间

	聚类后的等价类区间	消除不一致规则后的等价类区间
γ	[0,17.2),[17.2,21),[21,24.5), [24.5,26.5),[26.5,30),>=30	[0,17.2),[17.2,21),[21,24.5), [24.5,26.5),[26.5,30),>=30
c	[0,18.2),[18.2,43),[43,80),>=80	[0,3),[3,18.2),[18.2,43),[43,80),>=80
ϕ	[0,8),[8,22),[22,34),[34,43),>=43	[0,8),[8,22),[22,34),[34,43),>=43
φ_f	[5,27),[27,39),>=39	[5,27),[27,39),>=39
H	[0,160),[160,390),>=390	[0,7.8),[7.8,25),[25,160),[160,390),>=390
γ_u	[0,0.05),[0.05,0.32),>=0.32	[0,0.05),[0.05,0.32),>=0.32
F	[0,0.95),[0.95,1.32),[1.32,1.53),>=1.53	[0,0.95),[0.95,1.32),[1.32,1.53),>=1.53

2) 粗糙集数据分析

首先进行数据过滤,经过过滤消除了对所有规则都不必要的属性值.过滤后条件属性仍为 6 个,属性值由原来的 27 个减少到 14 个.接着按照传统的粗糙集分析方法消除完全相同的规则,再对每条规则化简不必要的属性值.最终得到 54 条规则,这里规则数量还是较多

的,但是由于属性值经过了化简,其规则的条件大大减少了,同时由于所分析的系统是经过专家们精心挑选的,其本身冗余较少,因此这个结果也是合理的. 至此,我们已经从系统中提出了关键的规则,再根据式(2)计算各条规则的粗隶属函数.

3) 训练神经子网络

在第 1 步中结论被划分为四类,这样系统可以用四个子网络来表示,每个子网络结构为 6-10-1,用 Matlab 5.2 中的 train.m 函数来训练,训练精度为 0.000 1.

以上建立了基于粗糙集的岩石滑坡稳定性分析模型. 下面用剩余的 11 组数据进行检验,并和文献[6]的结果进行比较,其结果见表 3 和表 4.

表 3 仿真结果比较

序号	边坡状态	极限平衡法		神经网络方法						
		边坡状态	安全系数	边坡状态	基于粗糙集的神经网络			文献[6]中的神经网络		
					安全系数	绝对误差	相对误差(%)	安全系数	绝对误差	相对误差(%)
72	*F*	*F*	0.96	*F*	1.018 3	0.058 3	6.07	1.08	0.12	12.50
73	*F*	*F*	1.15	*F*	1.201 4	0.051 4	4.47	1.09	0.06	5.22
74	*S*	*S*	1.34	*S*	1.443 5	0.103 5	7.73	1.25	0.09	6.72
75	*F*	*F*	1.20	*F*	1.198 5	0.001 5	0.13	1.16	0.04	3.33
76	*S*	*S*	1.55	*S*	1.443 5	0.106 5	6.87	1.46	0.09	5.81
77	*S*	*S*	1.45	*S*	1.430 6	0.019 4	1.34	1.49	0.04	2.76
78	*S*	*S*	1.31	*S*	1.277 4	0.032 6	2.49	1.50	0.19	14.5
79	*S*	*S*	1.49	*S*	1.443 5	0.046 5	3.12	1.55	0.06	4.03
80	*F*	*F*	1.20	*F*	1.198 5	0.001 5	0.13	1.29	0.09	7.5
81	*S*	*S*	1.52	*S*	1.443 5	0.076 5	5.03	1.40	0.12	7.89
82	*F*	*F*	1.20	*F*	1.198 5	0.001 5	0.13	1.29	0.09	7.5

表 4 主要性能指标比较

	平均误差	最大相对误差	边坡状态判别失误率	神经网训练时间	计算时间
基于粗糙集的神经网络	3.41%	7.73%	0%	四个神经网每个训练不足 200 步	4 分钟
文献[6]神经网络	7.07%	14.5%	0%	2985 步	48 分钟

计算时间是在 586/100MHz/16MB 的计算机上,采用 Matlab 5.2 进行计算的时间,本文方法的计算时间包括除了人工调整区间划分所需的时间外的整个建模过程所耗时间,文献[6]中方法的计算时间是采用文献[6]方法重新训练网络的总时间.

从表中可以看出,本文提出的基于粗糙集的神经网络模型在训练时间和预测精度上较文献[6]中采用单一神经网络模型要优越.

6 结论

本文提出了一种基于粗糙集的神经网络模型,该模型的特点是综合了粗糙集的在知识获取方面的优点和神经网络在数值逼近上的长处,同时克服了粗糙集数据分析方法不适于处理连续数据的不足,提高了神经网络训练的速度. 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. 训练完成后,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,这一点是神经网络建模所不能完成的. 通过仿真对比也说明了该

方法的有效性.

但是,该方法在数据离散化时,采用了聚类算法,但结果可能产生矛盾规则,其本质是区间划分不当造成的,需要进行调整. 如何实现自动调整,容另文讨论.

致谢　鞍山钢铁学院数理系何希勤副教授和东北大学资源与土木工程学院冯夏庭教授为作者提供了仿真所需的原始数据资料,使得本文能顺利完成,在此致以诚挚的谢意!

参　考　文　献

1　Pawlak Z. Rough Set——Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic Publishers, 1991. 9～30

2　王珏,苗夺谦,周育健. 关于 Rough Set 理论与应用的综述. 模式识别与人工智能, 1996, **9**(4):337～344

3　韩祯祥,张琦,文福栓. 粗糙集理论及其应用综述. 控制理论与应用, 1999, **16**(2):153～157

4　Duntsch I,Gediga G. Simple data filtering in rough set systems. *International Journal of Approximate Reasoning*, 1998, **18**(1-2):93～106

5　Jelonek J. Rough set reduction of attributes and their domains for neural networks. *Computational Intelligence*, 1995, **11**(2):339～347

6　Feng Xia-Ting, Wang Yong-Jia *et al*. Rock slope stability analysis based on neural network. *Journal of the Mining and Material Processing Institute of Japan*, 1996, **112**:25～32

7　Duntsch I, Gediga G. Statistical evaluation of rough set dependency analysis. *International Journal of Human-Computer Study*, 1997, **46**(5):589～604

黎　明　东北大学信息科学与工程学院博士研究生. 研究方向为 Rough sets 理论及其在控制中的应用.

张化光　东北大学信息科学与工程学院教授、博士生导师. 研究方向为模糊控制、智能控制、非线性控制.