第 33 卷 第 5 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.5 Computer Engineering March 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)05—0154—02 文献标识码：A 中图分类号：TP18

基于粗糙集与BP神经网络的多因素预测模型

江洋溢，孟 科，张恒喜，徐 鑫
（空军工程大学工程学院，西安 710038）

摘 要：运用粗糙集方法和信息熵概念，在不改变训练样本分类质量的条件下，按照输入影响因素相对于输出的重要度的大小，对输入参数集进行约简，确定神经网络输入层变量和神经元个数。通过对典型样本的学习，建立粗糙集 BP 神经网络多因素预测模型，将其用于导弹系统研制费用预测。结果表明，该方法减少了网络的训练时间，改善了学习效率，具有较高的预测精度，是可行的、有效的。

关键词：粗糙集；神经网络；信息熵；多因素预测；费用预测

Multi-factor Estimation Model Based on Rough Set and BP Artificial Neural Network

JIANG Yangyi, MENG Ke, ZHANG Hengxi, XU Xin
（Engineering College, Air Force Engineering University, Xi'an 710038）

【Abstract】 In term of the important degree of input influence factor to output, rough set approach and the conception of information entropy are employed to reduce the parameters of the input parameter set with no changing classification quality of samples. Thus, the number of the input variables and neurons is gotten, and the multi-factor estimation model based on rough set and BP artificial network is set by learning from the typical samples. Its application to the cost estimation of missile system is given. It is shown that the approach can reduce the training time, improve the learning efficiency, enhance the predication accuracy, and be feasible and effective.

【Key words】 Rough set; Neural network; Information entropy; Multi-factor estimation; Cost estimation

近年来，国内外学者对多因素预测理论和方法做了大量的研究工作，提出了各具特色的预测方法。如时间序列分析方法、回归分析法以及基于模糊数学、灰色理论、分形理论或人工神经网络的预测方法等。其中，神经网络具有强大的任意函数逼近能力、学习能力、自组织和自适应能力，但当输入参数过多，样本数量过大时，网络收敛速度变慢，需要较长的训练时间，并易陷入局部最优。波兰科学院院士Z.Pawlak教授于1982年提出的粗糙集理论是继概率论、模糊集、证据理论之后的又一个刻画不完整性和不确定性的数学工具，具有很强的实用性，在许多领域（如人工智能、控制与决策、模式识别与故障诊断、冲突分析等）取得了令人鼓舞的成果[1,2]。鉴于此，本文尝试融合粗糙集方法和神经网络技术各自的优势，在不改变样本分类质量的条件下，运用粗糙集方法和信息熵概念约简特征参数，确定网络输入层变量和神经元个数，建立粗糙集BP神经网络（Rough set Back Propagation Networks，RSBPN）多因素预测模型。

1 RSBPN 多因素预测模型的建立

为了利用粗糙集理论与方法确定模型的输入变量，首先应建立粗糙集的决策数据模型，将影响输出结果的特征参数 $x_1, x_2, \ldots, x_n$ 作为条件属性，得条件属性集 $C=\{x_1, x_2, \ldots, x_n\}$，将输出 y 作为决策属性，得到决策属性集 $D=\{y\}$。那么任一系统的输入参数和输出的历史数据构成一个对象 $u_i=(x_{i,1}, x_{i,2}, \ldots, x_{i,n}; y)$，论域为 $U=\{u_i \mid i=1,2,\ldots,m\}$。这样就建立了一个决策数据模型，记为 $S=<U, C\cup D>$。

1.1 训练样本属性特征化[2]

用粗糙集方法进行数据处理时，必须对连续属性值特征化。特征化的方法有很多，如等距离、等频率、Naive Scaler算法、 Semi Naive Scaler 算法以及基于 Rough 集理论相结合的离散化算法等，可根据实际情况选择其中一种方法。

1.2 特征参数约简[2,3]

为方便起见，样本属性值特征化后所确定的决策数据模型仍记为 $S=<U, C\cup D>$。在此利用信息熵概念定义条件属性重要度，通过属性重要度约简特征参数。

设 $\tilde{C}\subseteq C$，$U/\tilde{C}$ 上的子集组成的 σ 代数上的概率分布为

$$P(X)=card(X)/card(U)，X\in U/\tilde{C} \tag{1}$$

其中 $card(X)$ 表示集合 X 的势，X 为有限集合时，表示 X 所含元素的个数。令

$$H(\tilde{C})=-\sum_{X\in U/\tilde{C}} P(X)\ln(P(X)) \tag{2}$$

称 $H(\tilde{C})$ 为条件属性子集 $\tilde{C}$ 的信息熵。令

$$H(D|\tilde{C})=-\sum_{X\in U/\tilde{C}} P(X)\sum_{Y\in U/D} P(Y|X)\ln(P(Y|X)) \tag{3}$$

其中 $P(Y|X)=\dfrac{card(Y\cap X)}{card(X)}$，称 $H(D|\tilde{C})$ 为条件属性子集 $\tilde{C}$ 对于决策属性 D 的条件信息熵。$H(D|\tilde{C})$ 提供了决策属性 D 对条件属性集 $\tilde{C}$ 的信息依赖性的一个较为合理的度量，$H(D|\tilde{C})$ 越大，依赖性越强，否则越弱。设 $C_j=C-\{x_j\}$，$j=1,2,\ldots,n$，令

基金项目：空军工程大学博士论文创新基金资助项目
作者简介：江洋溢(1979－)，男，博士生，主研方向：飞行器设计，装备系统工程，军用飞机型号发展工程，效费分析；孟 科，博士生；张恒喜，教授、博导；徐 鑫，硕士生
收稿日期：2006-03-23 **E-mail：**j.yy@163.com

$$\lambda(x_j,C,D)=H(D|C_j)-H(D|C)，\quad j=1,2,\ldots,n \tag{4}$$

$\lambda(x_j,C,D)$越大，说明条件属性（特征参数）x_j对决策属性（输出）y的影响越大，因而越重要。若$\lambda(x_j,C,D)=0$，则特征参数x_j是冗余的，应从特征参数集中去掉，参数集C化简为C_j。不断地计算简化后参数集中各个参数的重要度，直至所有参数的重要度大于 0，这时便得到约简的参数集，记为C^*，不妨假设$C^*=\{x_1,x_2,\ldots,x_N\}$。

1.3 BP神经网络模型的建立[4,5]

一个三层前反馈 BP 神经网络是前反馈 BP 神经网络的典型结构，它分为输入层、隐藏层和输出层。同层节点间无关联，异层网络间前向连接。以约简的参数集C^*作为 BP 神经网络的输入变量，N 是输入层神经元的个数，将系统决策属性作为神经网络的输出，设神经元输出特性函数为逻辑 sigmoid 型函数，即$f(u)=(1+e^{-u})^{-1}$。利用训练样本集来训练这个网络，得到训练好的权值和阈值，然后就可用于多因素系统的输出预测。

2 应用实例

装备研制费用预测就是一种典型的多因素预测。对武器装备系统研制费用预测，可为国防预算的节省，新武器装备系统的论证、研制、生产、使用和保障提供一个可靠的依据。影响装备研制费用的因素众多，各项因素之间的相关关系也比较复杂。根据预测所处的阶段和所掌握的信息，费用预测方法大致可分为参数法、类比法、工程法 3 类，但这些方法都有一定的局限性，有时难以取得满意的结果[6]。

现以地空导弹为例，将RSBPN模型应用于其研制费用预测。影响地空导弹系统研制费用的特征参数很多，主要有发射重量G_0、导弹长度L、导弹翼展长度E、导弹的最大飞行速度$M_{\max}$、最大射高$H_{\max}$、导弹最大直径$d_{\max}$、战斗部重量G_z和最大射程$R_{\max}$等[7]。本文采集了 8 种不同导弹系统的研制费用数据作为训练样本（见表 1，研制费用数据已经过处理，并折算到同一财政年度）。

表 1 导弹系统研制费 RSBPN 模型训练样本

序号	型　号	G_0 (kg)	L (m)	E (m)	M_{max}	H_{max} (km)	d_{max} (m)	Gz (kg)	R_{max} (km)	Cost (亿)
1	罗兰特	63.5	2.4	0.5	1.6	5.5	0.16	5.9	9.3	3.2
2	小榭树	86.2	2.9	0.64	2.5	2.5	0.12	4.5	5	3.7
3	响尾蛇	78	2.94	0.55	2.3	3	0.16	4	14.5	4.5
4	海响尾蛇	87	2.94	0.54	2.2	4	0.16	14	10	5
5	霍克	630	5.03	1.2	2.5	11	0.36	50	25	7.2
6	改进霍克	625	5.03	1.2	2.5	18	0.36	50	40	8.9
7	标准 II	1 360	8.23	1.58	3	24	0.34	61.2	127.9	18
8	爱国者	1 000	5.3	0.80	6	24	0.41	100	70	18.7

以 PAC－1 为测试样本，采用 RSBPN 费用预测模型对其研制费用进行估算，具体步骤如下：

(1)利用表 1 建立决策数据模型，并利用阈值法（对每个属性设定一个阈值，1 表示达到标准，0 表示没有达到标准）将模型中的属性特征化，得到一个决策表（见表 2）。

表 2 二维信息表

U	C								D
	G_0	L	E	M_{max}	H_{max}	d_{max}	Gz	R_{max}	
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	1
5	1	1	1	0	0	1	1	1	1
6	1	1	1	0	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1
8	1	1	0	1	1	1	1	1	1

(2)利用式(3)计算各特征参数相对于费用的重要度，可知发射重量G_0、导弹长度L、导弹最大直径$d_{\max}$、战斗部重量G_z和最大射程$R_{\max}$相对于研制费的重要度为 0，因而是冗余的，将其从特征参数集中剔除掉，得到约简后的决策表（见表 3）。

表 3 约简后的二维信息表

U	C			D
	E	M_{max}	H_{max}	
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	1
5	1	0	0	1
6	1	0	1	1
7	1	1	1	1
8	0	1	1	1

(3)以筛选出的 3 个费用特征参数，翼展长度E、最大飞行速度$M_{\max}$、最大射高$H_{\max}$作为输入，采用三层 BP 网络（隐层神经元数为 6，各层神经元的传递函数分别取 tansig、tansig、purelin 函数，BP 网络的训练函数取 trainlm）进行计算，达到确定的精度要求后，输入测试样本参数进行费用估算。在训练中，设定初始学习率为 0.06，动量常数为 0.9，训练目标误差 0.001，网络迭代终止平均误差平方值 0.000 555 19，迭代终止步数 76 步，误差变化曲线见图 1。

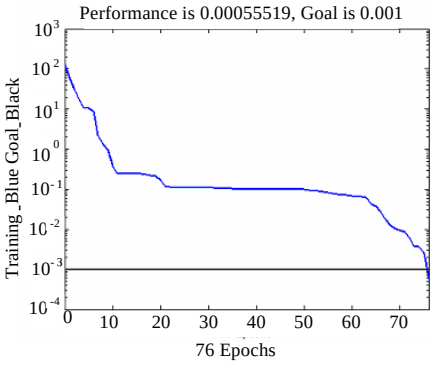

图 1 误差变化曲线

(4)对 PAC－1 研制费用进行预测，测试结果见表 4。

表 4 样本 RSBPN 测试结果

型号	E	M_{max}	H_{max}	实际研制费	RSBPN 预测值	相对误差
PAC－1	0.87	6	24	19.2618	18.6998	2.92%

3 结论

本文利用粗糙集的数据分析与处理能力以及信息熵概念，按照输入影响因素相对于输出的重要度的大小，对输入参数集进行约简，简化神经网络输入变量的个数，减少了网络的训练时间，改善了学习效率，提高了预测精度，其基本思想具有很强的推广性，适用于其它大样本多因素预测问题。

参考文献

1 Pawlak Z. Rough Set[J]. International Journal of Computer Information Sciences, 1982, 11(5): 342-356.

2 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.

3 Pawlak Z. Rough Set Theory and Its Applications to Data Analysis[J]. Cybernetics and Systems, 1998, 29(7): 661-688.

4 许　东, 吴　铮. 基于 MATLAB 6.X 的系统分析与设计——神经网络[M]. 西安: 西安电子科技大学出版社, 2002.

5 焦李成. 神经网络系统理论[M]. 西安: 西安电子科技大学出版社, 1992.

6 陈学楚. 装备系统工程[M]. 北京: 国防工业出版社, 1995-03.

7 徐品高. 防空导弹体系总体设计[M]. 北京: 宇航出版社, 1996.