正则化训练的神经网络与粗集理论相结合的股票时间序列数据挖掘技术 [1]

王晓哗 * ** 王正欧 *

*(天津大学系统工程研究所 天津 300072)
**(河北工业大学自动化系 358# 天津 300130)

摘 要: 论文提出将正则化神经网络与粗集理论相结合应用于股票时间序列数据库的数据挖掘. 首先对时间序列数据库进行预处理, 除去高频干扰信号, 然后将股票时间序列数据按照收盘价的变化趋势分割成一系列静态模式, 每种模式代表股票价格的一种行为趋势 (上涨或下跌), 把决定各种模式的相关属性组成一系列信息, 形成一个适用于粗集方法的信息表. 然后使用正则神经网络对信息表进行学习, 用粗集理论从正则神经网络所存储的知识中抽取规则, 得到的规则可以用于预测时间序列在未来的行为. 该方法融合了正则神经网络优良的泛化性能和粗集理论的规则生成能力, 实验表明, 该方法预测效果比较准确.

关键词: 时间序列, 正则神经网络, 数据挖掘, 粗集理论
中图分类号: TP391, TN-052 **文献标识码**: A **文章编号**: 1009-5896(2004)04-0625-07

Stock Market Time Series Data Mining Based on Regularized Neural Network and Rough Set

Wang Xiao-ye* ** Wang Zheng-ou*

*(*Institute of Systems Engineering, Tianjin University, Tianjin* 300072, *China*)

**(*Automation Dept., Hebei University of Technology, Tianjin* 300130, *China*)

Abstract This paper presents a new method of stock market time series data mining. It combines regularized neural network with rough set. The process includes preprocessing of time series and data mining. The preprocessing cleans and filters time series. Then, the time series are partitionel into a series of static patterns, which is based on the trend (i.e., increasing or decreasing) of closing price. The most important predicting attributes identified from every model form an information table. The regularized neural network is used to learn and predict the data. Rough set can extract rule knowledge in the neural network, which can be used to predict the time series' behavior in the future. This method combines the generalization faculty of regularized neural network and the rule reduction capability of rough set. The experimental results demonstrate the effectiveness of the algorithm.

Key words Time series, Regularized Neural network, Data mining, Rough set

1 引言

股票市场高风险和高收益并存, 因此对于股票数据的知识发现的研究一直受到人们的关注. 近年来, 随着计算机技术的飞速发展和存储能力的大大提高, 使得这一方面的研究有了很大的发展. 目前对股票数据库的数据挖掘方法大致集中在四个方面.

(1) 相似性 [1] 的研究将时间窗口在时间序列上滑动, 通过距离计算从一个时间序列或多个时间序列中寻找相似的时间序列模式进行聚类形成相似组群, 当有一个新的时间序列需要分析时, 可以从相似组群中寻找与它最相似的类来匹配.

[1] 2002-11-13 收到, 2003-04-21 改回
国家自然科学基金 (60275020) 、河北省教委基金 (401023) 资助课题

(2) 关联规则的研究在相似性的基础上做更深入的研究，即试图从相似组群中寻找出关联规则 [2]，类似于 “如果某一天 Microsoft 上涨而且 Intel 下降，则 IBM 第二天上涨”，由于上述两种方法是基于时间窗口的，因此算法的挖掘效果对窗长有很强的依赖性。

(3) 值预测的研究即利用过去和当前的观测值估计未来值，使用比较多的方法是用神经网络来预测 [3]，这种预测方法由于是基于时间序列的具体数值，而这些数据往往含有许多干扰数据，因此值预测方法的抗干扰能力较弱。

(4) 趋势预测的研究 [4] 按照股票数据的发展趋势将时间序列分类，从时间序列数据中抽取出有限数量的对分类有决定性的属性来代替原来的时间序列，从而大大减小了时间序列的数据量。然后将属性组成一个信息表，应用分类算法进行数据挖掘。这种方法既减小了运算量而且由于从时间序列中抽取出决策属性与单个时间序列值无关，因此抗干扰能力较强。本文提出的时间序列数据挖掘方法即属于趋势预测的研究，它是通过极值法以收盘价的变化趋势为依据将股票时间序列数据分割成一系列模式 (各种模式持续时间是不同的)，一种模式代表股票时间序列的一种行为趋势 (上涨或下跌)，把决定各种模式的决策属性组成一系列信息形成一个信息表。采用正则神经网络对信息表进行学习和预测，用粗集理论从神经网络的知识中抽取规则，规则形如 “如果目前股票形势波动比较强，则本次行情持续时间很短”，得到的规则集可以用于对股票数据做行情转折点的预测。这种预测结果清晰易懂，对投资者的投资行为更具指导意义。

神经网络与粗集理论相结合的方法有多种，其中主要常用以下方法：

采用粗集理论来进行神经网络的结构优化，其中文献 [5] 主要是通过粗集理论对前馈网络输入节点的选取进行指导，粗集理论根据各个输入节点对预测精度的贡献大小对输入节点赋予不同的权重，从而提高预测精度。文献 [6] 将粗集理论的上、下界语义概念加入前馈神经网络的网络结构中，粗集神经单元可以存储某个输入值或输出值的上、下边界，排除了普通神经元只能处理精确值的缺陷，从而解决了神经网络对不可分辨实例的学习问题。但是上述两种神经网络学习得到的知识仍然是只存在于网络的权值及其神经元的连接当中，仍然类似于暗箱，不容易被人们所理解。

正则神经网络 [7] 由于容错能力高，泛化性能强等优点，目前已经成为公认的数据挖掘工具。在股票数据中有许多干扰数据，因此通过正则网络的学习，可以有效去除干扰。但由于神经网络的规则生成能力较低，可读性较差，因此不适于单独使用。而粗集应用于数据挖掘的主要目标是从信息系统表示的数据中抽取规则，所以本文将正则神经网络与粗集理论相结合，用粗集理论从正则神经网络的知识中抽取规则，得到最终的易于人们理解的知识。算法结构如图 1 所示，这种结合方法与以往算法相比，既提高了系统的泛化性能，同时所得结果又容易理解。

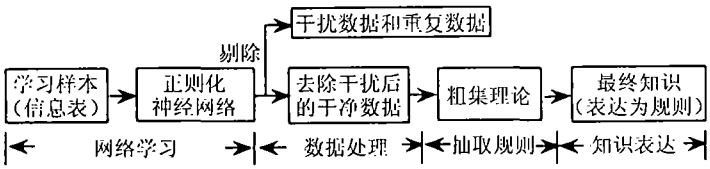

图 1 算法结构

2 股票时间序列数据库的预处理 [4]

定义 1 时间序列可以看作是通过时间排列的一系列数据。

股票价格时间序列是指由股票每天的收盘价按时间排序的一系列数据。

时间序列预处理包括数据清洗、模式分割、属性抽取和离散化四部分。

2.1 数据清洗

在股票时间序列中由于人为因素或消息影响，存在有许多噪声，因此每天的收盘价包括随机波动和长期趋势数据，在分析这些数据之前必须进行清洗，除去随机波动。

假设原始数据为

$$a_{\mathrm{raw}}(n) = a(n) + e(n), \qquad n = 1, 2, 3, \cdots$$

其中 $a(n)$ 为长趋势， $e(n)$ 为噪声，清洗数据是指生成 $\hat{a}(n)$ 来近似描述 $a(n)$ 。相比较而言，$a(n)$ 是一个稳定的信号，而噪声信号是一个随机的受各种因素影响的信号。如果采用傅里叶变换，可以看到 $a(n)$ 是一个低频信号， $e(n)$ 是一个高频信号。为了滤去噪声，本文采用了信号处理技术中的低通滤波器，其中最简单的一种滤波方法是有限脉冲响应 (FIR) 法，算式如下：

$$\hat{a}(n)=\sum_{i=0}^{N-1}a_{\mathrm{raw}}(n-i+[N/2])\cdot \boldsymbol{c}(i)$$

其中 $a_{\mathrm{raw}}(n)$ 是原始数据， $\hat{a}(n)$ 是清洗后的结果， $\boldsymbol{c}(i)$ 是含 N 维系数的向量。 N 根据具体情况来定， $\boldsymbol{c}(i)$ 是设计 FIR 的重点，由脉宽和精度来确定，可以用 Matlab 信号处理工具箱中的有关函数来得到。

2.2 模式分割

模式分割的原则是使每个模式内数据的变化趋势一致。因为对于广大投资者来说，最关心的是行情的持续时间，从而做出买入还是卖出的决定，因此在模式分割时应使每个模式内部收盘价的变化趋势是不变的。模式分割主要是寻找数据中行为趋势改变的转折点。

寻找转折点的最简单的方法是求取曲线的极值点，即 $\left.\dfrac{\mathrm{d}\hat{a}(t)}{\mathrm{d}t}\right|_{t=t_x}=0$ ，由 t_x 组成的一系列时间点 $T_e=\{t_0,\cdots t_{N_e}\}$ ，将时间序列分割成了 N_e 个模式。

2.3 属性抽取

经过 2.2 节的模式分割，使得每个模式内部数据的行为趋势是不变的，因此可用一个直线方程近似代替原来曲线，则近似直线方程的斜率以及模式区间的长度是模式的重要属性。

2.3.1 模式长度　当 $T_e=\{t_0,\cdots,t_{N_e}\}$ 中某个区间的宽度 $t_{i+1}-t_i\le d$ 时 (d 为设计者设定的阈值)，从 T_e 中削去 t_i 然后插入 $t_i=(t_{i+1}-t_{i-1})/2$ 。

2.3.2 模式斜率　对于上述分割得到的每个区间，求取其近似斜率 α_i

$$\alpha_i=[\hat{a}(t_{i+1})-\hat{a}(t_i)]/(t_{i+1}-t_i)$$

2.3.3 信噪比 (SNR)　信噪比是时间序列的另一个重要特征，它显示了时间序列的波动情况。信噪比越高则时间序列越不稳定，受各种因素影响越多。计算信噪比用如下公式：

$$\mathrm{SNR}_i=\sqrt{\int_{t_i}^{t_{i+1}}\frac{\varepsilon^2(t)}{\hat{a}^2(t)}\Big/(t_{i+1}-t_i)}$$

式中 $\varepsilon(t)=|a(t)-\hat{a}(t)|$, $a(t)$ 是原始数据。

2.4 属性离散化

若希望挖掘得到的知识是以规则的形式表示，则要求信息表中的值用离散 (如整型，字符串型和枚举型) 数据表达，因此，必须将上述连续属性进行离散化处理。

由于上述连续属性的值域范围较宽，数值比较分散，因此首先采用 Naive Scaler 算法 [8] 初步减小断点个数，形成断点集。然后设定离散区间个数，采用信息熵评价函数从断点集中寻找离散区间的端点。

Naive Scaler 算法对每一个连续属性 A 进行如下的处理过程：

设 $a(x)$ 为属性 A 在实例 x 中的取值， D 为决策属性， $d(x)$ 是实例 x 的决策值。

步骤 1　根据属性值 $a(x)$ 由小到大的顺序对信息表中的实例排序。

步骤 2　从上到下扫描，设 x_i 和 x_j 代表两个相邻的实例：

如果 $a(x_i)=a(x_j)$ ，则继续扫描；

如果 $d(x_i)\neq d(x_j)$ ，即决策相同，则继续扫描；

否则，得到一个断点 c, $c=(a(x_i)+a(x_j))/2$ 。

由于上述连续属性的数值比较分散，由 Naive Scaler 算法离散化得到的断点集个数仍然很多，一般为几十个，由几十个值表达的属性用于信息表显然不合适。因此再由人为设定离散区间个数，采用基于信息熵 [9] 的离散化方法从断点集中寻找离散区间的端点。

由 Naive Scaler 算法离散化得到的属性 A 在断点集中的每个值都可以认为是一个潜在的区间边界，属性 A 的取值 a_j 可以将实例集 S 划分成满足条件 $a(x) < a_j$ 和 $a(x) > a_j$ 的两个子集，这样就创建了一个二元离散化。

定义 2 给定信息表中的实例集 S，实例集个数为 s，假定实例集有 m 个类，设 S 包含 s_i 个 C_i 类实例，$i = 1, \cdots, m$，一个任意实例属于类 C_i 的可能性是 s_i/s，对一个实例集分类所需的期望信息函数 I 是

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

定义 3 具有值 $\{a_1, a_2, \cdots, a_v\}$ 的属性 A 可以将 S 划分为子集 $\{S_1, S_2, \cdots, S_v\}$，其中 S_j 包含 S 中 A 值为 a_j 的那些样本。设 S_j 包含 C_i 的 s_{ij} 个样本，根据 A 的这种划分的期望信息称为 A 的信息熵。

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s} I(s_{i1}, s_{2j}, \cdots, s_{mj})$$

定义 4 A 上该划分的信息增益定义为 $\mathrm{Gain}(A) = I(s_1, s_2, \cdots, s_m) - E(A)$。

确定离散区间端点的过程递归地应用于所得到的每个区间，直到满足 $\mathrm{Gain}(A)$ 大于某个值，如 0.1。

将上述抽取的特征组成一个适用于神经网络进行数据挖掘的信息数据库，信噪比和斜率作为输入属性，区间长度作为输出 (预测属性)，从而采用神经网络就可以学习信息数据库，进行预测。

3 正则神经网络

神经网络目前已经成为公认的数据挖掘工具，前馈神经网络用于数据挖掘存在的一个重要问题是泛化能力。造成前馈网络泛化性能差的重要原因是过拟合。一种有效的限制网络过拟合的方法是正则化法 [7]，而使正则化有效的条件是在保证能够完成学习任务的前提下，保持最小的网络结构，到目前为止，隐节点数目的最优选择还没有理论上的依据，因此在设计网络的结构时，先使用较多的隐节点，以便使网络快速达到训练精度，当训练成功后再逐个删除不必要的隐节点 [10]，将正则化与节点删除技术结合在一起，将有效地提高前馈网络的泛化性能。

3.1 正则前馈网络学习算法

设计一个三层前馈正则网络，将信息表的条件属性作为网络输入，决策属性作为网络输出，基于单个权值的局部化正则最小二乘算法 [11] 如下：

$$\hat{w}_{ji}(k) = \hat{w}_{ji}(k-1) - p_{ji}(k)[(1-\lambda)v\hat{w}_{ji}(k-1) - \sum_{q=1}^{M} F_q^{ji}(k)(d_q(k) - f_q(\hat{w}(k-1), i(k)))]$$

$$p_{ji}(k) = p_{ji}(k-1)(\lambda + p_{ji}(k-1))(F^{ji}(k)F^{ji^T}(k) + (1-\lambda)v)^{-1}$$

$d_q(k)$ 是第 q 个输出节点的期望输出；$f_q(\hat{w}(k-1), i(k))$ 是第 q 个节点的实际输出；w_{ji} 为网络权值，$j = 1, 2, \cdots, N$，$i = 1, 2, \cdots, n_j$，n_j 为与神经元 j 相连的权值数，N 为网络中神经元的个数 (除输入层神经元)；M 为输出层神经元个数；$F_q^{ji}(k)$ 是第 q 个输出节点对 w_{ji} 的梯度；$0 < \lambda < 1$ 为遗忘因子；v 为正则因子。

3.2 隐节点删除

本文采用启发式搜索策略，先设计较大数目的隐节点，训练成功后依次删除每个隐节点，若在训练集上的精度下降不大，则重新训练删除后的网络，成功后若在验证集上的精度下降在允许范围内，则真正删除该隐节点，否则保留，直到网络中所有的隐节点都不能删除为止。

算法如下：

步骤 1　将信息表 S 中的实例分成训练集 $S1$ 和验证集 $S2$，令 ΔA 为 $S2$ 上精度的最大允许下降幅度。

步骤 2　使用正则最小二乘算法训练网络 ANN，使其在 $S1$ 上达到最大训练精度，记录下 ANN 在 $S2$ 上的精度 $A2$。

步骤 3　删除 ANN 中的第 H_n 个隐节点及其与它相连的所有权值，得到网络 ANNnew。

步骤 4　重新训练网络 ANNnew，当其在 $S1$ 上达到训练精度后，记录在 $S2$ 上的精度 $A2'$。

步骤 5　若 $|A2 - A2'|/A2 < \Delta A$，则用 ANNnew 代替 ANN，从隐节点集中删除第 Hn 个隐节点。

若 H_n 不是最后一个隐节点，则 $H_n = H_n + 1$，返回步骤 3。

4 基于粗集理论的规则抽取

经过训练后的神经网络其知识分布于网络的各个连接上，从网络中获取知识是相当困难的。粗集理论有很强的规则生成能力，因此将两种方法融合，神经网络经过充分的训练，信息表 S 中的知识将分布于网络的连接权上，对于 S 中不能被网络学习的实例则可认为是错误实例，应该从信息表中删除，同时删除 S 中的重复实例，得到新的信息表 S_n，用粗集理论从 S_n 中抽取得到的规则集，可以认为是神经网络中知识的一个近似表达。

粗集理论进行规则抽取的过程正是对信息表进行值约简的过程。本文采用文献 [12] 的算法，其实质是从决策表的每个实例中寻找对得出决策影响最大的属性。主要步骤如下：

步骤 1　对决策表中的条件属性进行逐列考察。如果删除该属性列后，若产生冲突实例 (即两个实例的所有属性值都相同但决策值却不相同)，则保留冲突实例的原该属性值；若为产生冲突但含有重复实例，则将重复实例的该属性值标为“*”；对其它实例，将该属性值标为“？”。

步骤 2　删除可能产生的重复实例，并考察每条标记“？”的实例。若仅由未被标记的属性值即可判断出决策，则将“？”标记为“*”，否则，修改为原属性值；若某个实例的所有条件属性均被标记，则将标有“？”的属性项修改为原属性值。

步骤 3　删除所有条件属性均被标为“*”的实例及可能产生的重复实例。

步骤 4　如果两个实例仅有一个条件属性值不同，且其中一个实例该属性被标为“*”，那么，对该实例如果可由未被标记的属性值判断出决策，则删除另外一个实例；否则，删除本实例。

5 实验结果

股票时间序列数据库的记录包括股票代码，股票名称和每日开盘价和收盘价。投资者最为关注的是收盘价的变化趋势，因此只将收盘价从中提出进行分析，每日收盘价受各种干扰比较多，因此在分析之前必须进行数据清洗。

对于股票投资者来说，最关心的是股票行情的转变，譬如当股票由升变为降时抛出，而由降变为升时买入。因此，如何能够准确预测行情的转折点，是投资者最关心的。在本节，将上述描述的时间序列的数据挖掘过程应用于上海证券自 1998 年 1 月至 2000 年 12 月 3 年之间的数据，分析行情的转折点。

5.1 数据预处理

首先采用 2.1 节提到的低通滤波器清洗数据，其原始数据是上海证券和深圳证券分别取 100 支股票自 1998 年 1 月至 2000 年 12 月 3 年之间的数据，经过多次实验，脉宽取为 0.1/ 天，N 取为 30。通过 2.3 节的属性抽取，将决定某个模式的属性重组成一系列实例，形成一个信息表 S，共得到 1220 个实例，每个实例的属性包括：条件属性有前一个模式区间的斜率 (Slope1)、本模式区间的斜率 (Slope2) 和本模式区间的信噪比 (SNR2)，决策属性 (即预测属性) 为本模式

区间的长度 (Length2)，即本行情的转折点。因为这些属性都是连续数据，若用规则来表达信息表内在的规律，必须将连续属性离散化，其离散结果见表 1 。

表 1 属性离散化结果

	离散结果	范围
模式长度 (Length)	Short	$[0,30]$
	Medium	$[31,60]$
	Long	$[61,+\infty]$
模式斜率 (Slope)	Negative-large	$[-\infty,-2.95\times10^{4}]$
	Negative- little	$[-2.95\times10^{4},0]$
	Position-little	$[0,3.6\times10^{4}]$
	Position-large	$[3.6\times10^{4},+\infty]$
信噪比 (SNR)	Low	$[0,4.2\times10^{-5}]$
	Medium	$[4.2\times10^{-5},6.3\times10^{-5}]$
	High	$[6.3\times10^{-5},+\infty]$

5.2 网络学习

将上述得到信息表 S 中的 1220 个实例分成两部分，一部分为训练集 $S1$ ，占用其中的 900 个实例，一部分为验证集 $S2$ ，占用其中 320 个实例。因为神经网络只处理数值型数据，在网络训练之前需对离散化后的字符型数据进行标准化处理，因为离散值是以某种方式 (增加或减少) 相互关联的，所以可采用温度计码 [13] 来表示，如属性 Length 的 Short 值被表示为 001 ，Medium 值被表示为 011 ，而 Long 值被表示为 111 。条件属性为网络输入，决策属性为网络输出，因此输入节点个数所有条件属性温度计码位数的和 $4+4+3=11$ 个，输出节点数为决策属性温度计码的个数为 3 个，网络的初始结构定为 11-20-3 ，权值初始化为 $[0,1]$ 上均匀分布的随机数，$\lambda=0.999$, $v=0.1$, $p_{ji}(0)=1000$ 。使用上述第 3 节正则最小二乘前馈网络学习算法和加入隐节点删除的算法训练网络，运行结果如表 2 所示。

表 2 神经网络训练结果

		训练精度 (%)	测试精度 (%)
隐单元数目 (个)	删除前 (20)	97.79	92.58
	删除后 (15)	97.70	94.77

由表中数据可以发现，删除冗余隐节点后的网络，虽然训练精度和删除之前是大致相同，但是测试精度却高于未删除网络。因此删除冗余隐节点可以与正则化相结合提高系统的泛化能力。

5.3 规则抽取

将信息表 $S1$ 中上述网络不能正确学习的实例删除掉，同时删除重复实例得到新的信息表 Sn ，对于信息表 Sn 采用第 4 节中提到的规则抽取算法，共得到 14 条规则，形式如下：

Rule1: If SNR2 is Low and Slope2 is Positive-large then Length2 is medium

Rule2: If SNR2 is High then Length2 is Long

将 $S2$ 的实例对规则集进行测试，测试精度为 85% ，其预测精度高于文献 [1] 的预测结果 (64.9%) 。分析所得结果，由于粗集理论对规则中的值约简，因此抽取得到的规则中含有较少的条件属性个数，从而使规则集更加精练且更具概括性，其预测精度相对较高。

有待研究的问题如下。

(1) 影响股票价格变化趋势的因素除内在因素以外，还有其他外在因素如物价指数，银行利率等对股市具有大范围影响的经济指标，因此研究外在因素对股票价格变化趋势的影响将是未来研究的一个难点。

(2) 最后得到的规则个数越少，其对投资者的指导意义更大，因此如何减少规则个数将是研究的方向之一。

(3) 我国股市发展还处于初级阶段，人为因素及国家政策干预有时会使股市产生不规则的突变，因此技术预测的难度很大。

(4) 对于连续属性值值域较宽而且数值较分散的情况，其离散化方法有待研究。

6 结论

本文融合了神经网络和粗集理论两种数据挖掘技术，充分利用了正则神经网络的泛化能力和粗集理论的规则生成能力，得到的知识以规则的形式表达出来，易读易懂，应用于股票数据的行情预测取得良好效果。

参 考 文 献

[1] Das G, Gunopulos D. Finding similar time series. In Proc. of the Conference on Principles of Knowledge Discovery and Data Mining, Trondheim, Norway, 1997: 124–135.

[2] Das G, Lin K, Mannila H, Renganathan G, Smyth P. Rule discovery from time series. In Proc. of the 4th Int. Conference on Knowledge Discovery and Data Mining, New York, NY, Aug. 27–31, 1998: 179–183.

[3] Hansen V J, Nelson R D. Data mining of time series using stacked generalizers. *Neurocomputing*, 2002, 43(1): 173–184.

[4] Last M, Klein Y. Knowledge discovery in time series databases. *IEEE Trans. on System, Man and Cybernetics-part B*, 2001, 31(1): 160–169.

[5] Qin Zh, Mao Z. A new algorithm for neural network architecture study. In Proc. of the 3rd World Congress on Intelligent Control and Automation, Hefei, China. June 2000: 795–799.

[6] Lingras P. Comparison of neofuzzy and rough neural networks. *Information Sciences*, 1998, (110): 207–215.

[7] Girosi F, Jones M. Regularization theory and neural networks architectures. *Neural Computation*, 1995, 7(2): 219–269.

[8] 王国胤. Rough 集理论与知识获取. 西安：西安交通大学出版社，2001: 99–104.

[9] Jiawei H, Micheline K. Data Mining: Concepts and Techniques. San Meteo, CA: Morgan Kaufmann Publishers, Inc., 2001, Chapter 5.

[10] Reed R. Pruning algorithms—a survey. *IEEE Trans. on Neural Networks*, 1993, 4(5): 740–747.

[11] 李冬梅，王正欧. 提高前向神经网络泛化性能和实时性能的新方法. 电机与控制学报，2002, 6(3): 241–244.

[12] 常犁云，王国胤等. 一种基于 Rough Set 理论的属性约简及规则提取方法. 软件学报，1999, 10(11): 1206–1211.

[13] 宋擒豹，沈钧毅. 神经网络数据挖掘方法中的数据准备问题. 计算机工程与应用，2000, 36(12): 102–104.

王晓哗：　女，1972 年生，讲师，主要研究方向为数据挖掘、神经网络.

王正欧：　男，1938 年生，教授，博士生导师，主要研究方向为数据挖掘、神经网络、人工智能.