基于模糊神经网络的粗糙集在股市预测中的应用

叶德谦　马志强　李　帼　姜皇普
（燕山大学中德信息技术合作研究所　秦皇岛 066004）

摘　要　提出在模糊神经网络中使用粗糙集理论进行网络的设计。在模糊神经网络中引入粗糙集理论，不仅可以去除模糊神经网络中输入层的冗余神经元而且可以确定隐含层神经元的数目，从而使模糊神经网络具有更准确的逼近收敛能力和较高的精度。最后应用于股票市场，在股票买卖时机预测中取得了良好的效果。

关键词　粗糙集，模糊聚类，神经网络，股市预测

Rough Set Based on Fuzzy Neural Network for the Application in Prediction of Stock Market

YE De-Qian　MA Zhi-Qiang　LI Guo　JIANG Huang-Pu
(ICDZ - Institute for Information Technology, Yanshan University, Qinhuangdao 066004)

Abstract　A new scheme of knowledge encoding in a fuzzy neural networks using rough set theoretical concepts is described. Introducing the rough theory in the fuzzy neural network can not only remove redundant neurons in the input layers of fuzzy neural, but also construct the neurons in the hidden layers. Thus enable the fuzzy neural network to have more accurate restraining ability and the high precision. Finally applies in the stock market, the stock business opportunity forecast has obtained the good effect.

Keywords　Rough sets theory, Fuzzy clustering algorithm, Neural network, Prediction of stock market

1　引言

股票市场在我国不断发展，逐步成为证券业及至整个金融业的必不可少的组成部分，受到投资者的普遍关注，因而对股票市场的预测研究具有重大的理论意义和诱人的应用价值。

粗糙集理论[1]是波兰数学家 Z. Pawlak 于 1982 年提出的一种数据分析理论，是继概率论、模糊集、证据理论之后的又一个处理不确定性的数学工具。粗糙集理论能有效地分析不精确、不一致不完整等各种不完备的信息，还可以对数据进行分析和推理，从中发现隐含的知识，揭示潜在的规律，其应用范围已拓展至包括机器学习、信号处理、模式识别等较为广泛的领域。

模糊神经网络已被广泛应用于各种领域的预测中，如经济、工业等。模糊神经网络同时具有神经网络和模糊逻辑的优点。神经网络有很好的学习能力，采用并行分布处理方法使得快速进行大量运算成为可能，并且能够同时处理定量、定性知识。而模糊逻辑技术能够处理更高层次的问题。但模糊神经网络不能够去除输入层的冗余神经元，而且也不能够确定隐含层神经元的个数，使其预测精度受到限制。

本文首先采用模糊聚类和粗糙集相结合的属性约简算法对决策表属性进行归约，去掉冗余的属性，并得到决策规则。然后利用粗糙集和模糊神经网络相结合的技术（即模糊粗糙神经网络）对现存的人工神经网络算法进行改进和优化。选择合适的神经网络输入层、隐含层和输出层，使决策输出结果有较高的精度和可信度。实验结果表明，利用粗糙集理论对模糊神经网络的输入数据进行预处理，提取其中的关键成分作为神经网络的输入以及初始规则的组成元素，可以简化神经网络的拓扑结构，提高训练样本的质量，缩短训练时间。

2　结合粗糙集和模糊聚类方法的属性约简算法

常见的基于粗糙集的属性约简算法包括基于可辨识矩阵和逻辑运算的属性约简算法，归纳属性约简算法，基于互信息的属性约简算法，基于特征选择的属性约简算法等。这些方法没有考虑到数据领域知识的特殊性以及用户需求的灵活性，因此在基于粗糙集的属性约简算法中引入模糊集方法，采用了一个结合粗糙集和模糊聚类方法的属性约简算法，用户可以根据实际决策需要和领域知识更改阈值 λ，得到用户满意的属性约简结果。

结合粗糙集和模糊聚类对决策表属性进行约简，并获得决策规则。具体实现为：

步骤 1　将待分析数据表中的各属性取值离散化[5]。

步骤 2　计算每个属性 α_i 的一组数字表征。设属性 α_i 有 m 个不同的属性取值，根据这个不同属性取值，确定对论域 U 的划分：$U/IND(\alpha_i)$，$U/IND(\alpha_i)=\{E_{1i}, E_{2i}, \cdots, E_{mi}\}$。则属性 α_i 的一组数字表征为 $\{card(E_{1i}), card(E_{2i}), \cdots, card(E_{mi})\}$。其中运算符 $card()$ 是求集合中元素的个数。

步骤 3　根据属性 α_i，α_j 的数字表征，采用绝对值减数法，求出它们之间的相似矩阵 $[R]=(r_{ij})$。这里 r_{ij} 用来刻画对象 x_i，x_j 之间的相关程度。其中绝对值减数法为：

$$r_{ij}=\begin{cases}1 & \Leftrightarrow i=j\\ 1-c\sum_{k=1}^{m}|x_{ik}-x_{jk}| & \Leftrightarrow i\neq j\end{cases}$$

式中 $c>0$ 为常数，可根据实际情况选定，使 r_{ij} 取值范围在[0，

叶德谦　博士，教授，硕士生导师，研究方向：数据仓库，决策支持系统及并行处理；马志强　硕士研究生，主要研究方向为决策支持系统，人工智能及应用。

1]。

步骤4　用平方法求出相似矩阵[R]的传递闭包[t(R)]。[R]→[R]²→[R]⁴→…→[R]ⁿ，其中 $n=2^k$。

步骤5　对矩阵作分析，依据属性之间关联强度，适当选取阈值λ得主条件属性集。

$[t(R)]\lambda=(r_{ij}(\lambda))_{m\times n}$

其中：$r_{ij}(\lambda)=\begin{cases}1 \Leftrightarrow r_{ij}\geqslant\lambda\\0 \Leftrightarrow r_{ij}<\lambda\end{cases}$

3　模糊-粗神经网络的构造

3.1　网络结构

模糊-粗神经网络的结构如图1所示。该网络共分为4层。它是根据粗糙集理论的工作过程设计的，可以说是一个模糊神经网络实现的粗糙集推理系统。

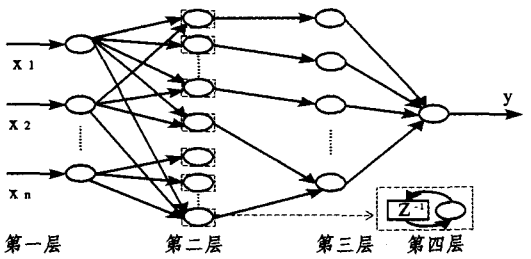


图1　模糊-粗神经网络的结构

第一层是输入层，表示输入 $x=(x_1,x_2,\cdots,x_n)^T$，n 为输入变量的个数。

第二层称为隶属度函数层，将 n 个输入量$(x_1,x_2,\cdots,x_n)$按照一种不可分辨关系进行划分，使得每一个输入分量离散化为 r 个不同的值，这些值在[0,1]之间，可定义该层神经元的作用函数为 Gauss(高斯)函数：

$$\mu_{ij}=\exp\left(-\frac{(x_i-m_{ij})^2}{\sigma_{ij}^2}\right)$$

$i=1,2,\cdots,n$；$j=1,2,\cdots,r$，r 为离散分割数；其中，μ_{ij}，m_{ij}，σ_{ij} 与第一隐含层的各节点相对应。从上至下各节点相对应的输出 μ_{ij} 表示为 $\mu_{11},\mu_{12},\cdots,\mu_{1r};\mu_{21},\mu_{22},\cdots,\mu_{2r};\cdots;\mu_{n1}\ \mu_{n2},\cdots,\mu_{nr}$，且 μ_{ij} 与 m_{ij} 的下标完全相同。

第三层的每一个节点代表一条规则，这些规则的获取和描述的理论依据是前期粗糙集和模糊聚类的属性约简算法获得的规则。该层节点的作用函数取决于对规则的适用度：

$$H_i=\mu_{1i}\cdot\mu_{2i}\cdots\mu_{ni}=\prod_{j=1}^{z}\mu_{ji}\quad 1\leqslant i\leqslant z$$

（z 为第三层的节点数，即规则条数）

第四层表示输出，即为清晰化层。在多输入单输出系统里，权值 ω_i 的初始值预设为各规则粗糙隶属度值，然后用 BP 算法迭代，该层的输出函数：

$$y=\sum_{i=1}^{z}\omega_i H_i$$

3.2　网络的学习算法

令 $e^k(n)=d^k(n)-y^k(n)$，$E^k(n)=\frac{1}{2}(e^k(n))^2$。其中 $y^k(n)$为第 n 次迭代中输入第 k 个样本后输出层的实际输出，而 $d^k(n)$是所对应的期望值，$E^k(n)$是误差函数。

则学习过程对 m_{ij}，σ_{ij}，ω_{ij} 的调整可以如下计算：

$$m_{ij}(n+1)-m_{ij}(n)=\frac{1}{N}\sum_{k=1}^{N}(-\eta\partial E^k/\partial m_{ij})$$

$$\sigma_{ij}(n+1)-\sigma_{ij}(n)=\frac{1}{N}\sum_{k=1}^{N}(-\eta\partial E^k/\partial\sigma_{ij})$$

$$w_{ij}(n+1)-w_{ij}(n)=\frac{1}{N}\sum_{k=1}^{N}(-\eta\partial E^k/\partial w_{ij})$$

其中，η 为学习速率。以上各式的推导如下：

$$m_{ij}(n+1)-m_{ij}(n)=\frac{1}{N}\sum_{k=1}^{N}(-\eta\partial E^k/\partial m_{ij})$$

$$=-\frac{1}{N}\sum_{k=1}^{N}\eta(n)(d^k(n)-y^k(n))\cdot\partial y^k/\partial m_{ij}$$

$$=-\frac{1}{N}\sum_{k=1}^{N}\eta(n)(d^k(n)-y^k(n))$$

$$\cdot w_j\prod_{i=1j\neq i}^{n}\mu_{ij}\cdot 2\exp\left(-\frac{(x_i-m_{ij})^2}{\sigma_{ij}^2}\right)\cdot\frac{(x_i-m_{ij})^2}{\sigma_{ij}^3}$$

同理有

$$\sigma_{ij}(n+1)-\sigma_{ij}(n)=\frac{1}{N}\sum_{k=1}^{N}(-\eta\partial E^k/\partial\sigma_{ij})$$

$$=-\frac{1}{N}\sum_{k=1}^{N}\eta(n)(d^k(n)-y^k(n))\cdot w_j\prod_{i=1j\neq i}^{n}\mu_{ij}\cdot$$

$$2\exp\left(-\frac{(x_i-m_{ij})^2}{\sigma_{ij}^2}\right)\cdot\frac{(x_j-m_{ij})^2}{\sigma_{ij}^3}$$

以及

$$w_{ij}(n+1)-w_{ij}(n)=\frac{1}{N}\sum_{k=1}^{N}(-\eta\partial E^k/\partial w_{ij})$$

$$=-\frac{1}{N}\sum_{k=1}^{N}\eta(n)(d^k(n)-y^k(n))H_i$$

$$=-\frac{1}{N}\sum_{k=1}^{N}\eta(n)(d^k(n)-y^k(n))\cdot\prod_{j=1}^{n}\mu_{ij}$$

4　股票交易时机的预测

4.1　指标体系

按照技术分析的思想，股指的未来走势与现行的股市行情相关。因此，可以利用现有的有关收盘价与成交量的数据，对未来股市的走势进行预测。对于股票交易买入卖出点的判断分析，虽然已存在众多的方法，但投资者依然难以把握。本章中我们将应用神经网络模型，为选择股票交易时机提供决策支持。为此，我们编制了两个技术指标

$$O_3=\frac{\mathrm{Max}(R_T,\cdots,R_{T+3})-S_T}{\mathrm{Max}(R_T,\cdots,R_{T+3})-\mathrm{Min}(R_T,\cdots,R_{T+3})}$$

$$O_6=\frac{\mathrm{Max}(R_T,\cdots,R_{T+6})-S_T}{\mathrm{Max}(R_T,\cdots,R_{T+6})-\mathrm{Min}(R_T,\cdots,R_{T+6})}$$

其中，$\mathrm{Max}(R_T,R_{T+1},R_{T+2},R_{T+3})$是第 T 天到第 $T+3$ 天之间出现的最大股票指数；$\mathrm{Min}(R_T,R_{T+1},R_{T+2},R_{T+3})$是第 T 天到第 $T+3$ 天之间出现的最小股票指数，S_T 是第 T 天收盘价指数，其他符号的含义类似。O_3 与 O_6 的编制类似于威廉指数，值域都是[0,1]。若 O_3 或 O_6 的取值接近于1，则说明相对于未来几天的股市行情，现在的收盘价处于较低的位置，股价呈上升趋势，可以作为理想的买入点；若 O_3 与 O_6 的取值接近于0，则现在的收盘价相对于未来几天的股价处于较高点，股价呈回落趋势，应见机卖出。

4.2　网络构建

根据有关专家的分析及相关文献的研究成果，表1为本文所构造的决策表的条件属性。

表1　条件属性

日期	威廉指数	DIF	DSY	BIAS
开盘价	最高价	最低价	收盘价	成交量

以上共计10个指标。利用粗糙集和模糊聚类的属性约

（下转第183页）

如表1所示。

表1 测试结果比较

方法	窗口长度	Stem-loop预测准确率	平面伪结预测准确率
SVM	15	80.13%	71.80%
BP网络[4]	13	70.93%	无法预测

表1中的stem-loop预测准确率定义为：预测正确的stem-loop标记数目占基本二级结构标记(不包含伪结结构标记)总数的百分比。而平面伪结预测准确率则为正确预测的平面伪结标记数目占实际平面伪结标记数目的百分比。

与文[4]所使用的BP神经网络相比，支持向量机的方法对stem-loop结构的预测准确率有了较大的提高，同时还可以较理想地预测平面伪结结构。与传统的基于确定性的动态规划算法(SCFG、最小自由能等模型)比较起来，SVM成功地解决了时间复杂度的问题，并且利用滑动窗的输入方式将不受序列长度的影响。

总结 使用SVM模型来预测RNA二级结构是一种全新的尝试。本文利用SVM算法的优势，结合RNA二级结构的特征，扩展了文[4]中的NSSEL标记的内涵，成功地实现了包含平面伪结结构的RNA二级结构预测。实验表明，该算法能够得到理想预测精度，并且能有效解决传统算法中存在的计算复杂性和长链分子的预测问题。由于未能发现非平面伪结结构的有效标记，因此本算法暂时不能实现非平面伪结结构的预测，作者将会继续深入研究，期待有进一步的发现。

参考文献

1 Zuker M. Optimal computer folding of large RNASequence using thermodynamics and auxiliary information, Nucl. Acids Res., 1981, 9:133～148
2 SakakibaraY, Brown M, Hughey R, et al. Stochastic context-free grammars for tRNA modeling[J]. Nucleic Acids Research, 1994, 22(23):5112～5120
3 Shapiro B A, Navetta J. A massively parallel genetic algorithm for RNA structure prediction. J. SuperComput 1994, 8:195～207
4 张秀苇，邓志东，宋丹丹. RNA二级结构预测的神经网络方法. 清华大学学报(自然科学版)，2006(10)
5 Gorodkin J, Stricklin S L, Stormo G D. Discovering common stem-loop motifs in unaligned RNA sequences. Nucleic Acids Res., 2001, 29:2135～2144
6 HolleyL H, KarplusM. 基于神经网络的蛋白质二级结构预测[J]. 生物物理学，1989, 86:152～156
7 Eddy S R, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Research, 1994, 22(1):2079～2088
8 VAPNIKVN著. 统计学习理论的本质[M]. 张学工，译. 北京：清华大学出版社，2000

(上接第169页)

简算法，我们确定了16条规则数和主条件属性。主条件属性有日期，威廉指数，DSY，开盘价，最高价，最低价，收盘价以及成交量。

根据上述的数据处理方法得到相应的神经网络模型，即：第1层的节点数为8个，第2层的节点数为16个，第3层的节点数为16个，第4层的节点数为1个。

4.3 网络的训练及测试的效果评价

4.3.1 网络的训练

网络的训练样本为上证A指数从2004.07.08到2005.06.30的相关数据，共计240个交易日。各个技术指标从指数的原始数据中用程序处理得到。训练结果的统计分析如表2所示。

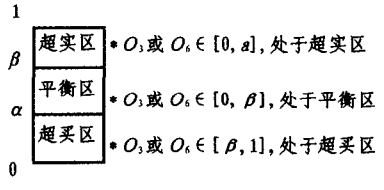
表2 网络训练结果

	最大误差	平均误差	训练周期
NO_3的网络训练结果	0.10125	0.07232	420233
NO_6的网络训练结果	0.10124	0.06412	502500

4.3.2 网络测试的效果评价

测试使用2005.07.01～2006.05.28之间的数据，共221个交易日。将样本值输入，经训练后NO_3的最大误差为0.1024，平均误差为0.0642，NO_6的最大误差为0.1140，平均误差为0.0624。

现根据若O_3和O_6两指标编制的含义，将区间[0,1]按一定的界限划分成如下图所示的三个区域



为了进一步评价所建立的网络模型对指数交易指导意义，假定在测试样本的221个交易日实行如下的交易方式：

1)投资者持有指数股票

1. O_3或$O_6\in[\alpha,1]$，处于平衡区或超卖区，说明股市行情处于盘整或见涨趋势，投资者可以继续持有股票；

2. O_3或$O_6\in[0,\alpha]$，处于超买区，说明行情即将回落，应将指标看成是卖出的信号，投资者出售股票。

2)投资者尚未持有指数股票

1. O_3或$O_6\in[0,\beta]$，行情位于平衡或买区，股市调整或回落，投资者应继续观望，不应马上介入；

2. O_3或$O_6\in[\beta,1]$，行情处于超卖区，股市即将反弹，会有一段上涨趋势，是买入的信号，投资者应买入股票。

α与β的具体数值可根据投资的不同倾向来确定。由于股市存在着交易费用问题，如经纪人佣金，印花税等，因此在一定的时间内，交易的次数很多并不意味着能给投资者带来更大的收益，具体操作要视实际情况来抉择。

结论 从大量的观察和实验数据获取知识，表达知识，推理决策规则是智能信息处理的重要任务。本文基于模糊粗糙集理论构建了新型的模糊-粗神经网络(FRNN)，该模型综合了粗糙集理论在知识获取方面的能力和模糊神经网络在数值逼近上的优势。通过粗糙集智能数据分析，可消除初始决策表中的冗余信息和噪声数据的干扰，减少了模糊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，减少了训练所需的计算量和时间，提高了模型的正确率。

参考文献

1 Pawalk Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5):341～356
2 Hu X H, Cercone N. Learning in Relational Database: a Rough Set Approach [J]. Computational Intelligence, 1995, 11(2):323～338
3 Lingras P J. Rough Neural Networks[A]. In: Proceedings of Sixth International Conference on Information. Processing and Management of Uncertainty in Knowledge_Based Systems[C]. 1996. 1445～1450
4 Fayyad U, Shapiro G, Smyth P. The KDD Process for Extracting Usefl Knowledge From Volumes of Data. Communications of the ACM, 1996, 39(11):27～34
5 苗夺谦. 粗糙集理论中连续属性的离散化方法[J]. 自动化学报，2001, 27(3):296～302
6 刘清. Rough集及Rough推理. 科学出版社，2001
7 张文修，吴伟志. 粗糙集理论与方法. 科学出版社，2001