文章编号: 1000-6788(2003)09-0067-04

基于神经网络集成系统的股市预测模型

张秀艳, 徐立本
(吉林大学商学院, 吉林 长春 130012)

摘要: 基于神经网络集成理论,建立股市预测模型. 其中分别建立"基本数据模型"、"技术指标模型"和"宏观分析模型",最后以简单平均生成集成系统. 实证分析表明,股市预测神经网络集成系统的泛化能力高于各个独立的模型,从而使模型具有更好的稳健性和更好的应用价值.
关键词: 人工神经网络; 神经网络集成; 股市预测
中图分类号: F830 文献标识码: A

The Stock Market Forecast Model Based on Neural Network Ensemble

ZHANG Xiu-yan, XU Li-ben
(Business School, Jilin University, 130012)

Abstract: The technique of artificial neural networks provides a novel and effective method for stock market forecast. The neural network ensemble can heighten the generalization. In this paper, we proved that the generalization of stock market forecast system based on neural network ensemble is superior to the single models and the system is more effective and applicable.
Key words: artificial neural networks; neural network ensemble; stock market forecast

1 引言

股票市场预测是一个非线性函数值估计和外推问题. 应用传统的分析方法(如指数平滑方法、ARMA模型、MTV模型),可以预测一段时间内股指变化的大致走势,但传统方法需要事先知道各种参数,以及这些参数在什么情况下应做怎样的修正. 相比之下,神经网络依据数据本身的内在联系建模,具有良好的自组织、自适应性,有很强的学习能力、抗干扰能力. 它能自动从历史数据中提取有关经济活动中的知识,可以克服传统定量预测方法的许多局限以及面临的困难,同时也能避免许多人为因素的影响,因而为股票市场的建模与预测提供了新的方法. 在实际应用中,网络的泛化能力是最主要的. 而网络的泛化能力往往又决定于问题本身的复杂度、网络结构和样本量大小. 由于缺乏问题的先验知识,往往很难找到理想的网络结构,这就影响了网络的泛化能力的提高. 而神经网络集成(neural network ensemble)方法不仅易于使用,还能够以很小的运算代价显著地提高网络的泛化能力.

2 神经网络集成理论

1996年, Sollich和 Krogh给神经网络集成下了一个定义: 神经网络集成是用有限个神经网络(或其它学习系统)对同一个问题进行学习,集成在某输入示例下的输出由构成集成的各神经网络在此示例下的输出共同决定. 同时,也有一些研究者认为,神经网络集成指的是多个独立训练的神经网络进行学习并共同决定最终输出结果,并不要求网络对同一个问题进行学习[1].

收稿日期: 2002-07-26
作者简介: 张秀艳,女,吉林大学商学院讲师,经济学博士, Email: zhang0081_cn@sina.com; 徐立本,男,吉林大学商学院教授,博士生导师;

1995年, Krogh 等人给出了计算神经网络集成泛化误差的公式.

假设学习任务是对 $f: R^n \to R$ 进行逼近. 集成由 N 个神经网络组成,采用加权平均法,各神经网络分别被赋予权值 k_T,满足:

$$k_T > 0 \text{且} \sum_T k_T = 1 \tag{1}$$

训练集从分布 $p(x)$ 中随机抽取得到. 假设对于输入 X,网络 T的输出为 $V^T(X)$,则神经网络集成的输出为:

$$\bar{V}(X) = \sum_T k_T V^T(X) \tag{2}$$

分别定义神经网络和神经网络集成的泛化误差为:

$$E^T = \int dx p(x)(f(x) - V^T(x))^2 \tag{3}$$

$$E = \int dx p(x)(f(x) - \bar{V}(x))^2 \tag{4}$$

定义神经网络的差异度为:

$$A^T = \int dx p(x)(V(x) - \bar{V}(x))^2 \tag{5}$$

定义各网络泛化误差的加权平均为:

$$\bar{E} = \sum_T k_T E^T \tag{6}$$

定义神经网络集成的差异度为:

$$\bar{A} = \sum_T k_T A_T \tag{7}$$

则神经网络集成的泛化误差为: $E = \bar{E} - \bar{A}$ (8)

公式(8)右边的第二项度量了神经网络集成中各网络的相关程度. 若神经网络集成是高度偏置的,即对于相同的输入,集成中各个网络会给出相同或类似的输出,则神经网络集成的差异度会接近于零,于是其泛化误差接近于各神经网络泛化误差的加权平均. 若集成中各网络的响应是相互独立的,则神经网络集成的差异度较大,集成的泛化误差将远小于各网络泛化误差的加权平均. 因此,要增强神经网络集成的泛化能力,就应该尽量使集成中各网络的误差互不相关.

当集成用于回归估计时,集成的输出由各网络的输出简单平均或加权平均产生. Perrone等人认为采用加权平均法比采用简单平均法得到更好的泛化能力,并给出了指导权值选择的公式. 但也有的研究者认为,对权值的优化过程会导致过拟合,从而使集成的泛化能力降低,因此提倡使用简单平均.

在生成集成中的个体网络方面,最重要的技术是 Boosting和 Bagging方法. 此外,有的研究者利用遗传算法产生神经网络集成中的个体,有的使用不同的目标函数、隐层神经元数和权空间初始点来训练不同的网络,从而获得神经网络集成的个体.

实验和应用成果表明,神经网络集成是一种非常有效的方法. 即使在对神经网络集成的原理不清楚的情况下,也可以通过对一组网络进行简单的投票或平均,提高学习系统的处理能力[1]. 基于这一思想,本文构造了"股市预测神经网络集成系统".

3 股市预测神经网络集成系统

笔者建立"基本数据模型"、"技术指标模型"和"宏观分析模型",构成股市预测神经网络集成系统,进一步提高股市预测模型的泛化能力,强化神经网络应用于股市预测的实效性.

3.1 基本数据模型

选取 2000年 8月 23日至 2001年 6月 28日的沪市上证综合指数做原始数据(时间序列),采用滑动窗技术,实现通过序列的前 3个时刻的值预测后 1个时刻的值.

为了满足网络输入输出对数据的要求,在学习之前首先对数据按下式进行归一化处理:

$$x_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \quad i = 1, 2, \cdots, m \tag{9}$$

取网络输入节点个数为 $p = 3$,输出节点个数为 $t = 1$,即用沪市上证综合指数的前天、昨天和今天的收盘价,预测明天的收盘价.

建立三层带有附加动量项和自适应学习速率的 BP网络,经过 10次试验对比分析,设定输入节点为 3个,输出为 1个节点,隐含层为 5个节点. 训练样本为 100个,测试样本为 100个. 误差精度设为 0.01(误差平方和),初始学习速率为 0.01,最大迭代次数设为 5000.

结论　学习训练至 5000次后的平均最小误差为 0.00210,预测误差为 0.00308. 对上证指数的数据拟合效果较好. 带有附加动量项和自适应学习速率的 BP网络,具有较快的运算速度和最佳的逼近性能,同时可以克服陷入局部极小值. 可见,人工神经网络在处理诸如股票数据这种非线性时间序列的预测方面,具有很好的学习、映射和泛化能力和应用价值,模型的输出对于股市的短期趋势的研判具有参考价值.

3.2　技术指标模型

技术指标是按照事先定好的固定方法对证券市场的原始数据(开盘价、最高价、最低价、收盘价、成交量和成交金额,简称 4价 2量)进行处理,处理后的结果是某个具体的数字,即技术指标值. 每一个技术指标都是从某个特定的方面对市场进行观察,通过一定的数学公式产生技术指标,这个指标反映了市场某一方面深层的内涵,这些内涵仅仅通过原始数据是很难看出来的. 技术指标可以进行定量分析,使得具体操作的精度大大提高.

在笔者所建立的技术指标模型中,考虑到指标对股市预测的重要性和指标间的独立性及中国证券市场的广泛使用程度,分别引用移动平均线 MA(5)、随机指标 K、相对强弱指标 RSI、乖离率 BIAS、人气指标 AR、能量潮 OBV、心理线 PSY及前日收盘价、昨日收盘价和今日收盘价.

仍然建立三层带有附加动量项和自适应学习速率的 BP网络,输入节点为 10个(分别是 MA(5)、随机指标 K、相对强弱指标 RSI、乖离率 BIAS、人气指标 AR、能量潮 OBV、心理线 PSY及前日收盘价、昨日收盘价和今日收盘价),输出为 1个节点(明日收盘价). 经过试验比较隐含层取为 8个节点. 样本区间同样为 2000年 8月 23日至 2001年 6月 28日的沪市上证综合指数,共 200个数据,训练样本为 100个,测试样本为 100个.

结论　学习经过 5000次迭代后的平均最小误差为 0.00214,预测误差为 0.00300. 通过一些股市重要技术指标的引入,使得“技术指标模型”增加了反映市场各方面深层内涵的信息,这些内涵信息通过原始数据是很难反映出来的. 因此可以说,“技术指标模型”有更多的“含金量”,同时使股市神经网络模型更有说服力和应用价值. 在与“基本数据模型”同样的试验条件下,模型的复杂度也并没有太多的增加,只是由于问题的限定增加了 7个输入节点,隐层只增加了 3个节点,由于原始数据量的增加,训练时间增加到 1686秒. 网络的泛化能力有所提高,预测误差由“基本数据模型”的 0.00315下降到 0.00300.

3.3　宏观分析模型

众所周知,影响股市行情变化的主要因素有经济因素、政治因素、上市公司自身因素、行业因素、市场因素和投资者的心理因素.

笔者下面所建立的“宏观分析模型”,在分析股市基本数据的同时,考虑到模型的完备性,从理论上应该引入影响股市行情变化的经济因素、政治因素、上市公司自身因素、行业因素、市场因素和投资者的心理因素.但这些因素中的很多指标无从获得,或者无法量化,故只引入汇率和香港恒生指数两项指标,借以分析国际金融环境对我国股市的影响;引入 GDP、通货膨胀率和利率,借以分析国家宏观经济景气对我国股市的影响,此三项数据均以每一个月(或每一季度)中每天相同的数值代替日值.

同样建立三层带有附加动量项和自适应学习速率的 BP网络,输入节点为 8个,分别为今日收盘价、昨日收盘价、前日收盘价、汇率、香港恒生指数、GDP、通货膨胀率和利率,输出为 1个节点,隐含层为 6个节点. 样本区间同样为 2000年 8月 23日至 2001年 6月 28日的沪市上证综合指数,共 200个数据,训练样本为 100个,测试样本为 100个.

结论　学习经过 5000次迭代后的平均最小误差为 0.00220,预测误差为 0.00316. 此模型对上证指数

的数据拟合效果较好,汇率、香港恒生指数、GDP 通货膨胀率、利率 5项指标的引入,使得"宏观分析模型"包含了宏观经济基本面的更多信息,强化了股市神经网络模型的应用价值.更值得一提的是,在此模型中,季度值指标、月值指标和日值指标同时使用,进一步突破了传统统计分析方法对指标时点的限定① ,充分显示了人工神经网络模型对传统统计分析方法的可替代性和应用价值.

3.4 集成系统

将"基本数据模型"、"技术指标模型"和"宏观分析模型",构成股市预测神经网络集成系统,集成系统的输出采用简单平均法,如下式:

$$y = \sum_{i=1}^{3} g_i y_i, \quad i = 1, 2, 3 \tag{10}$$

其中 y 为集成系统的输出,y_i 为第 i 个模型的输出,g_i 为第 i 个模型的加权值,这里取 $g_1 = g_2 = g_3 = \frac{1}{3}$.

结论 集成系统学习训练 5000的平均最小误差为 0.00215,预测误差为 0.00308. 拟合曲线如图 1. 集成系统的泛化能力高于单个独立的模型,这种模型间的融合使得股市集成系统包含更广泛的输入信息,既有基本数据信息、技术指标信息,又包含较多的宏观经济信息,这必然使模型具有更好的稳健性和更好的应用价值. 同时人工神经网络模型突破指标时点的限制,更为实际经济建模另辟蹊径.

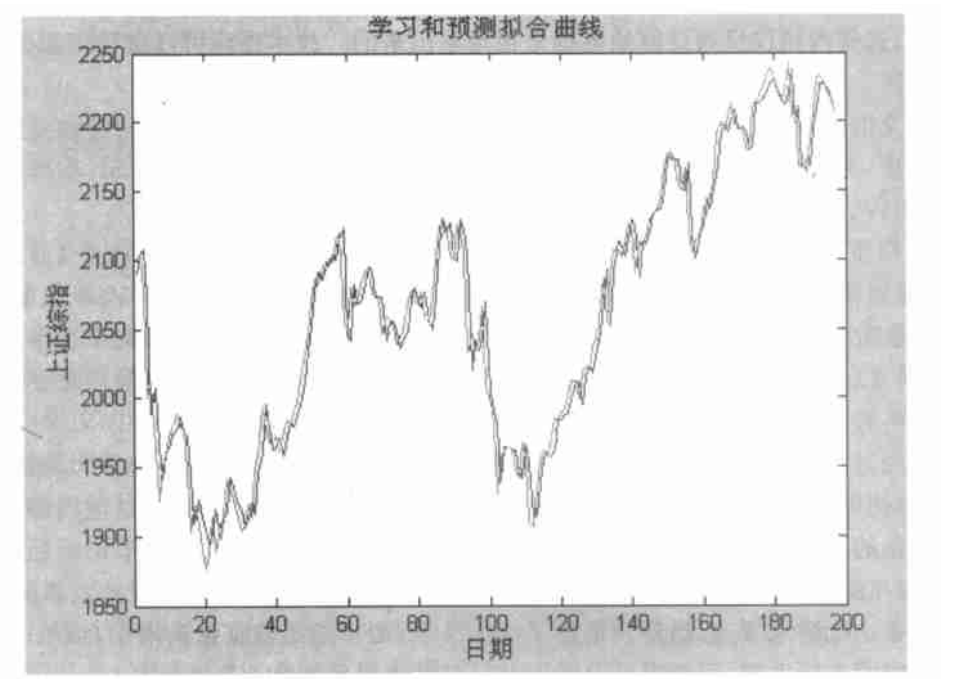

图 1 集成系统的拟合曲线(浅线为实际值,深线为学习和预测的拟合值)

参考文献:

[1] 吴建鑫,周志华,陈世福.神经网络集成综述 [A].中国人工智能学会. 中国人工智能学会第九届全国学术年会论文集 [C], 北京:北京邮电大学出版社,2001. 455- 458.

[2] 陈兴, 孟卫东, 严太华. 基于 T-S模型的模糊神经网络在股市预测中的应用 [J]. 系统工程理论与实践, 2001, 21(2): 66- 72.

[3] 王上飞, 周佩玲, 吴耿峰,等. 径向基神经网络在股市预测中的应用 [J].预测, 1998, (6): 44- 46.

[4] Algis Garliauskas. Neural network chaos and computational algorithms of forecast in finance[J]. Proceedings of the IEEE International Conference on System, Man and Cybernetics, 1999, 2 638- 643.

[5] Burgess A N, Bumm D W, Refenes A-P N. Neural networks with error feedback terms for financial time series modelling [A]. Prceedings of the Neural Network Conference [C], 1997, IOP Publishing Ltd and Dxford University Press, 1997. 65- 75.

① 经济指标往往有日值、月值、季度值和年值等,在使用传统统计分析方法建立经济模型时,这种指标间时点的差异,或者限制了指标被引入到经济模型中,或者很大程度上降低了所建立模型的精度.