# 基于神经网络和模式匹配的股票预测研究

林倩瑜,冯少荣,张东站
(厦门大学 计算机科学系,福建 厦门 361005)

**摘 要**:BP神经网络是分析股票数据最流行的工具之一。近期对模式匹配算法的研究表明模式匹配简化了股票趋势预测的复杂度并为股票市场预测提供了一种简单有效的方法。文中分别阐述了BP神经网络和模式匹配识别的原理,并提出将两种算法相结合,建立一个基于BP神经网络和模式匹配识别的股票市场分析和预测系统。这个系统克服了神经网络预测系统目标函数存在局部最小和模式匹配识别预测系统缺少股票价格自身变化特性的缺点,具有两种算法在股票预测应用方面的优势。通过对泰山石油的股价进行分析来测试这个系统。实验结果表明此方法不仅收敛速度快、预测精度高,而且易于操作,具有一定应用价值。

**关键词**:股票;预测;反向传播神经网络;模式匹配;非线性
**中图分类号**:TP183 **文献标识码**:A **文章编号**:1673-629X(2010)05-0017-04

# Stock Market Forecasting Research Based on Neural Network and Pattern Matching

LIN Qian-yu, FENG Shao-rong, ZHANG Dong-zhan
(Dept. of Computer Science, Xiamen University, Xiamen 361005, China)

**Abstract**: BP Neural Networks is one of the most popular tools in the analysis of stock data. Recent research activities in Pattern Matching indicate that Pattern Matching just simplify the complexity of stock trend prediction and provide a simple but effective way for the stock market prediction. This paper analyses the theory of BP Neural Networks and Pattern Matching, proposes a method for combining these two algorithms to establish a stock market forecasting system based on BP Neural Networks and Pattern Matching. This system overcomes the shortcomings of the local least in the Neural Networks forecasting system's objective function and Pattern Matching System's lack of stock changing probabilities, takes advantage of the unique strength in stock price forecasting of these two algorithms. Finally, test this system by analyzing and forecasting the Titan Oil's stock price. The experimental results show that not only this method has a quick convergent rate and a precise forecast, but also that it is easy to use and has much application value.

**Key words**: stock; forecasting; back propagation neural networks; pattern matching; nonlinear

## 0 引 言

股票市场从诞生的那天起就牵挂着数以万计投资者的心,它的风险与利润具有巨大的魅力,每一个投资者都想从中获利,因此股票的价格预测在金融数据挖掘方面一直是个比较热门的研究领域。人们采用各种方法,如K线图分析法、点数图分析法、移动平均线法,甚至抛硬币、算卦等方法来预测股票市场的波动[1]。

研究表明,基于多层前向神经网络的时间序列预测方法是目前最好的方法之一,这主要是因为神经网络具有可任意逼近非线性函数的能力和对于信息的综合能力,这是其他方法所不具有的。但它也存在一些缺点,主要表现在普通的BP算法收敛速度慢且容易陷入局部最优,从而影响了模型的建立和可靠性。另外,神经网络一般利用时间序列的最后几个节点来预测下一个节点的值,没有充分利用历史数据的变化规律,而模式匹配识别方法可以弥补这方面的不足。用模式匹配识别方法进行预测具有较好的拟合度,因为它是以一段较长的历史数据作为参考依据进行预测,由于股票历史数据的变化本来就包含多方面的影响因素,如人为因素、市场规律、突发事件等丰富的信息,因此模式匹配识别预测结果也就包含了这些信息。但股价也不仅仅是简单地重复过去,它随着时间的推移有着自身的变化规律。所以将人工神经网络和模式匹配

**收稿日期**:2009-09-07;**修回日期**:2009-12-19
**基金项目**:国家自然科学基金(50604012)
**作者简介**:林倩瑜(1985-),女,福建厦门人,硕士研究生,研究方向为数据挖掘;冯少荣,博士,副教授,研究方向为分布并行数据库、数据仓库、数据挖掘。

识别进行结合，应用模式匹配识别系统来产生神经网络的训练数据，用和预测值比较接近的时间序列来训练神经网络的权值，这样能够扬长避短，取得更好的预测结果。

## 1 BP人工神经网络算法

### 1.1 人工神经网络概述

人工神经网络是由大量类似于神经元的简单处理单元广泛相互连接而成的复杂网络巨型系统。它是在人类对其大脑神经网络认识理解基础上，人工构造的能够实现某种功能的网络。它是理论化的人脑神经网络的数学模型，是实现与模仿人脑神经网络和结构而建立的一种信息处理系统[2]。它实际上是大量的处理单元相互连接组成的复杂的网络，能够进行复杂的逻辑操作和非线性关系的实现。

人工神经网络是通过修改连接强度，即权值调整，表现出类似于人脑的分析、归纳的能力。研究它的目的就在于探究人脑加工、储存、处理信息的机制，进而探究将这个原理应用到信号处理等方面的可能性。

神经网络在预测非线性的系统方面有着很大的优势，它通常应用历史数据来训练网络，并利用在时间上最靠近预测数据的几个时间序列数据来预测实际输出。神经网络按网络结构可以分为层次神经网络和互联型神经网络。层次神经网络源于20世纪60年代出现的感知器。其后，在80年代中期，D. E. Rumelhart等人发表了称为反向传播(BP)算法的学习算法，同时给出了使用BP算法的实例。

### 1.2 BP算法

BP神经网络，即多层前馈反向传播神经网络，由一个输入层，一个或多个隐层，一个输出层组成。它可以用来模拟非线性映射模型，用来解决现实世界中的分类、估价、预测等问题[3]。在理论研究和实际应用中，人们最常用的是具有线性输出的单隐层网络，即三层前馈反向传播神经网络。

BP算法的学习过程主要分成两个阶段，即信息的前向传播和误差的反向传播。在前向传播过程中，输入信息由输入层经过隐层单元逐层处理，并传向输出层，第一层神经元的状态只影响到下一层神经元的状态。如果在输出层不能得到期望的输出，则转入反向传播，将误差信号沿原来的连接通路返回，通过修改各层神经元的权值，使得误差信号最小[4]。具体算法描述如下：

(1)前馈阶段。

① 输入结点的输出：$X_j$

② 隐结点的输出：

$$Y_i = f(\sum_i W_{ij}X_j + \theta_i)$$

其中 $W_{ij}$ 为输入层与隐层之间的连接权值，$\theta_i$ 为隐层结点阈值。

③ 输出结点输出：

$$O_l = f(\sum_i T_{li}Y_i + \gamma_l)$$

其中 $T_{li}$ 为隐层与输出层之间的连接权值，$\gamma_l$ 为输出结点阈值。

(2) 反向传播阶段。

沿着误差函数负梯度方向修改权值使得网络收敛。

对输出单元，误差为：

$$\delta_{li} = [O_l'(t) - O_l(t)]\{O_l(t)[1 - O_l(t)]\}$$

对隐层单元，误差为：

$$\delta_{ij} = Y_i(t)[1 - Y_i(t)](\sum_l \delta_l T_{li})$$

输出层与隐层之间的连接权值按下式修正：

$$T_{li}(t+1) = T_{li}(t) + \eta\delta_{li}Y_i(t)$$

隐层与输入层之间的连接权值按下式修正：

$$W_{ij}(t+1) = W_{ij}(t) + \eta\delta_{ij}X_j(t)$$

其中 $O_l'(t)$ 为期望输出，$O_l(t)$ 为神经网络实际输出，$\eta$ 为步长经验值，它的大小关系到学习速度的快慢。$T_{li}(t+1)$ 和 $W_{ij}(t+1)$ 都为当前的权值修正值，而 $T_{li}(t)$ 和 $W_{ij}(t)$ 则为上一学习周期的权值修正值[5]。

## 2 模式匹配预测算法

一个预测系统是否具有科学性，很大程度上依赖于其计算模型的选择，数据的数量和质量，以及其数据挖掘的能力。将模式匹配识别用于金融数据挖掘，在国内外比较少，但也有人取得了不错的成果，例如Sameer Singh的PMRS系统。模式匹配识别最重要的部分就是确定和当前时间序列数据趋势行为最接近的历史时间序列数据，并根据最接近的历史时间序列数据段的最后一个节点的价格升降趋势，来预测将来的价格趋势。算法主要包括时间序列选择，对时间序列编码并寻找最近似匹配时间序列，根据最近似的时间序列数据预测未来的股票价格三个关键步骤。

(1)时间序列选择。

假设所选取的历史时间序列，$X = \{x_1, x_2, \cdots, x_n\}$，根据该时间序列的历史观测值 $x_{n-k+1}, x_{n-k+2}, \cdots, x_n$ 对未来时刻 $x_{n+1}$ 的取值进行估计。由实践得出，$n$ 的取值以250～400最为合适，$k$ 的取值以2～6最佳。后面的算法描述以 $k$ 取值为3举例说明。

(2) 对时间序列编码并寻找最近似匹配时间序

列。

对原始时间序列 $X$ 作如下处理:

若 $x_{i+1} > x_i, x_i' = 1$;若 $x_{i+1} \leq x_i, x_i' = 0$

$x_i'' = x_{i+1} - x_i$

由此得到两个新的时间序列 $X' = \{x_1', x_2', \cdots, x_{n-1}'\}, X'' = \{x_1'', x_2'', \cdots, x_{n-1}''\}$。$X'$ 是股价趋势变化时间序列,$X''$ 是股价变化量时间序列。对历史观测值 $x_{n-2}, x_{n-1}, x_n$ 作相应的处理得到 $x_{n-2}', x_{n-1}'$ 和 $x_{n-2}'', x_{n-1}''$,接着从序列 $X'$ 中寻找与 $x_{n-2}', x_{n-1}'$ 模式最接近的数据段。这样的数据段可能有好几个,因此还要比较这些近似数据段和数据段 $x_{n-2}, x_{n-1}, x_n$ 的差异大小,从而选出一个最近似匹配时间序列。

数据段 $x_{n-2}, x_{n-1}, x_n$ 和数据段 $x_{j-2}, x_{j-1}, x_j$ 的差异计算公式如下:

$$D(X_j, X_n) = | x_{j-2}'' - x_{n-2}'' | + | x_{j-1}'' - x_{n-1}'' |$$

$D(X_j, X_n)$ 越小,说明两个数据段 $X_j$ 和 $X_n$ 的股价振幅度越接近,匹配程度较高;反之,说明两个数据段近似度较小。

(3) 根据最近似匹配时间序列预测未来股价。

假设数据段 $x_{j-2}, x_{j-1}, x_j$ 是数据段 $x_{n-2}, x_{n-1}, x_n$ 的最近似匹配序列,则可按如下公式对 $x_{n+1}$ 的值进行预测。

$$x_{n+1} = x_n + \frac{| x_{n-2}'' | + | x_{n-1}'' |}{| x_{j-2}'' | + | x_{j-1}'' |} \times x_j''$$

## 3 神经网络和模式匹配识别相结合在股票预测中的应用

### 3.1 神经网络结构的设计

网络结构的设计主要包括决定神经网络的层数,每层的神经元个数和网络拓扑的设计。

1) 神经网络层数:规模小的网络推广能力好,同时也易于理解和抽取规则、知识,利于软硬件实现。实践表明,4 层网络的结构比 3 层网络更容易进入局部最小,增加了网络权值的训练时间,因此在股价预测的应用中选用只有一个隐层的 3 层网络[6]。

2) 输入层结点数:每天股价的成交量组成一个时间序列,其走势呈波浪形式展开,具有周期性。我们认为每天的股价和前 $m$ 天的股价有某种函数关系。这个 $m$ 称为分析周期,也就是神经网络输入层的结点个数[7]。分析周期的选择是否恰当,对预测结果有直接影响,通常技术分析周期采用 5 日、10 日、20 日、60 日等,通过实验,选用 5 日。

3) 输出层结点数:考虑到时间、任务量及便于选择等因素,故确定神经网络输出结点数目为 1,即用前 $m$ 天的股价来预测当天股价的走势。

4) 隐层结点数:隐层结点个数的选取是一个非常复杂的问题,尚无理论上的指导。隐层神经元数目与问题的要求和输入、输出单元的多少都直接相关。过多的网络结点会增加训练网络的时间,也会使网络的泛化能力减弱,网络的预测能力下降,但网络结点过少则建模不充分[8]。一般地,隐层结点数在输入层结点和输出层结点之间。本实验隐层节点数目选为 6。

5) 网络拓扑的设计:对 BP 网络来说,结点的激励函数应选取连续可微的。在本系统中,选取 Sigmoid 函数 $f(x) = 1/(1 + e^{-x})$ 作为结点的激励函数。因为 Sigmoid 函数是一个连续可微的函数,其一阶导数存在,且 Sigmoid 函数具有非线性放大系数的功能,它可以把输入从负无穷大到正无穷大的信号,变成 0 和 1 之间的输出。对较大的输入信号,放大系数较小,对较小的输入信号,放大系数则较大,所以采用 Sigmoid 函数可以处理和逼近非线性的输入/输出关系。

### 3.2 训练样本集的准备

当前所得到的股票价格历史观测值构成了历史时间序列 $X = \{x_1, x_2, \cdots, x_n\}$,目的是根据该时间序列去预测值 $x_{n+1}$。一个大小为 $k$ 的模式 $\rho = (x_{n-k+1}, x_{n-k+2}, \cdots, x_n)$ 是由序列 $X$ 中的最后 $k$ 个值所组成的。被匹配的模式 $\rho$ 的大小 $k$ 的具体取值对预测准确度有着直接影响。因此,为了得到最好的预测准确度,必须优化匹配模式大小 $k$ 的取值。对 $k$ 的取值从 2 开始进行尝试,每次尝试增加一个单位直到实验所允许的最大值并记录所得到的误差值,最终选择可以得到最低误差值的 $k$。经过试验,选定 $k$ 的取值为 5[9]。

利用前面所述的模式匹配识别算法在历史时间序列 $X = \{x_1, x_2, \cdots, x_n\}$ 中选择出 10 组与时间序列 $x_{n-4}, x_{n-3}, \cdots, x_n$ 最近似模式匹配的序列。以这 10 组序列作为神经网络的训练样本的输入向量,每个序列在 $X$ 中的下一个值作为相应样本的教师数据。由此得到大小为 10 的训练样本集[10]。

### 3.3 利用神经网络进行股价预测

神经网络结构建立起来以后,就可以用模式匹配识别算法挑选出来的训练样本集来训练网络权值[11]。其中输入层到隐层,隐层到输出层的学习效率均为 0.7,当全局误差小于 0.005 时,训练结束。

训练完成后就可以进行股价的实际预测。把时间序列 $x_{n-4}, x_{n-3}, \cdots, x_n$ 作为训练好的神经网络的输入,得到的输出就是预测值 $x_{n+1}$。

## 4 实验测试结果分析对比

先用三层 BP 神经网络对泰山石油的综合指数进

行了分析和预测，以2006年8月29日至2007年5月27日共165个连续交易日作为训练样本，2007年5月28日起的100个交易日作为检验样本，进行预测效果检验，学习和预测的拟合曲线见图1。图中实线为实际曲线，虚线为神经网络对股票价格的预测值，横轴表示从2006年8月29日到2007年10月19日265个交易日，纵轴表示具体的股价大小。

图2是用BP神经网络和模式匹配识别相结合的混合预测系统对2007年5月28日起的100个交易日进行预测的拟合曲线。图中实线为实际曲线，虚线为混合系统对股票价格的预测值。

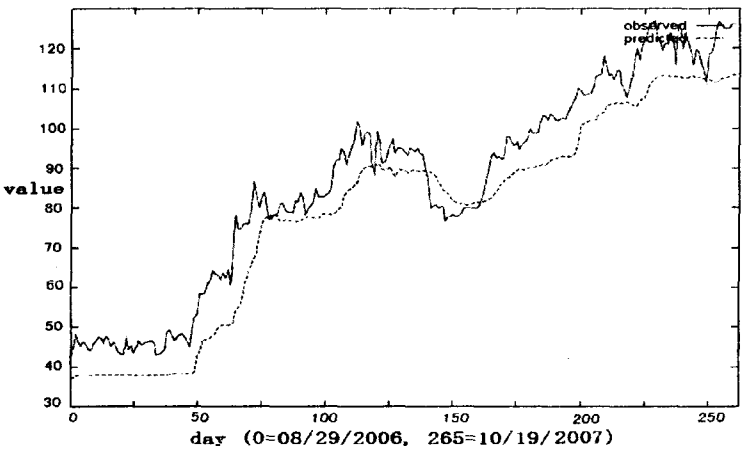


图1 神经网络预测系统预测结果

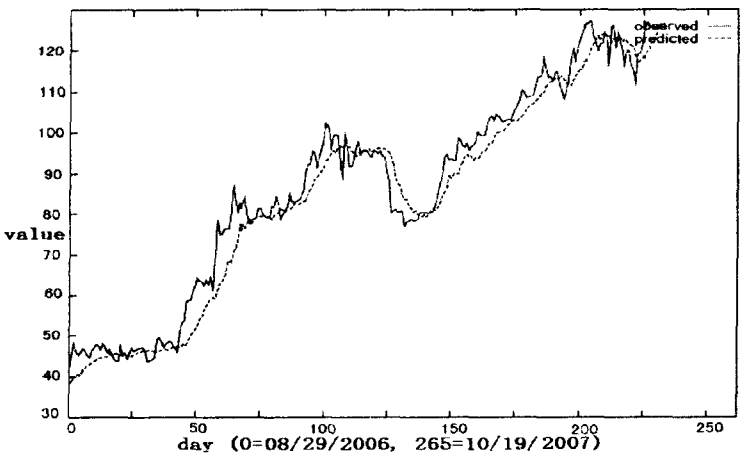


图2 混合预测系统预测结果

这两种方法都是通过股价的历史数据对未来的股票价格进行预测，并且都取得了较好的预测结果。但通过两张预测效果图，可以得出结论，相比较而言，BP神经网络和模式匹配识别相结合的混合预测模型得到的结果更加准确[12]。

分析其原因，神经网络是根据股市环境的不同来建立它的模型参数。它建立模型的依据就是我们所提供的训练样本，因而在这个基础上建立起来的模型所适合的环境即是训练样本所处的股市环境[13]。经由模式匹配识别算法筛选出的样本与检验样本的模式匹配程度高，其数据之间的内在联系比其他样本更加相似。以经过筛选得到的样本去训练神经网络，它所得到的模型更接近于检验样本所处的股市环境，从而得到更高的预测精确度。

## 5 结束语

文中将BP神经网络和模式匹配识别相结合运用于股价预测，预测精度较高，有很高的应用价值。相对于遗传算法和神经网络相结合，这还算是一种比较新的方法，虽然已经有Sameer Singh，Jonathan Fieldsend等人先行一步，但还有很多地方值得去探索研究。例如：该混合预测系统比原有的BP神经网络模型运行时间更长；在输入层增加数据量，反映影响股价的外在因素，如人为因素、经济因素等。总之，要在股票的预测研究上得到更好的准确度，还需要更多的深入研究[14]。

**参考文献：**

[1] 胡守仁，沈 清. 神经网络应用技术[M]. 长沙：国防科技大学出版社，1993.

[2] 钟 颖，汪秉文. 基于遗传算法的BP神经网络时间序列模型预测[J]. 系统工程与电子技术，2002，24(4)：9－11.

[3] 王上飞，沈 谦. 径向基神经网络在股票预测中的应用[D]. 中国科学技术大学电子技术部，1998.

[4] 周春光，梁艳春. 计算智能：人工神经网络－模糊系统－进化计算[M]. 长春：吉林大学出版社，2001.

[5] YANG Yi－wen, YANG Chao－jin. Short term forecasting of stock market based on R/S analysis and fuzzy neural networks[C]//System Man and Cybernetics, 2003[s. l.]: MIS Press, 2003.

[6] Chow Chi Kin, Lee Tong. Construction of multi－layer feed forward binary neural networks by a genetic algorithm[C]//Neural Networks. 2002, IJCNN'02. Proceedings of the 2002 International Joint Confernece. Honolulu, HI, USA: [s. n.], 2002.

[7] 陈向光，裴旭东. 人工神经网络技术及其应用[M]. 北京：中国电力出版社，2003.

[8] 阎平凡，张长水. 人工神经网络与模拟进化计算[M]. 北京：清华大学出版社，2003.

[9] Singh S. A Long Memory Pattern Modeling and Recognition

（下转第25页）

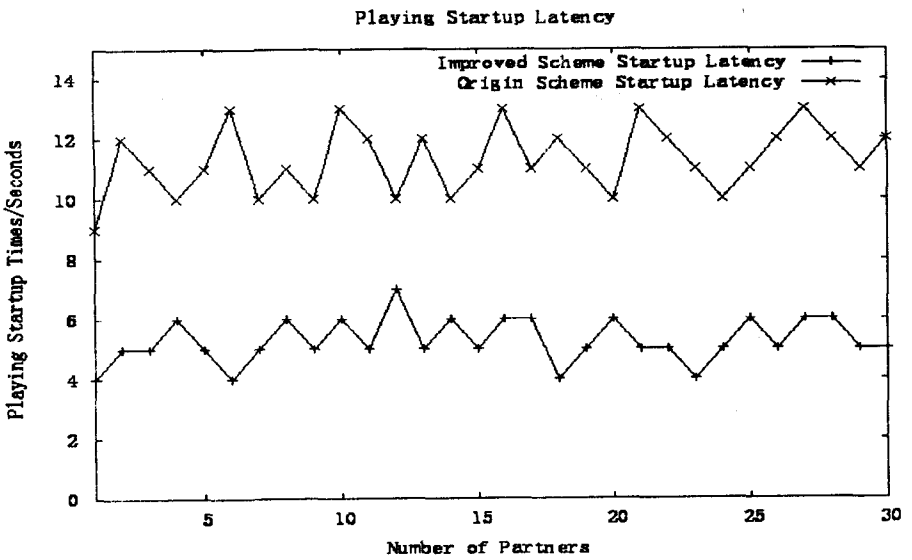


图 3　改进方案的播放启动时延和原方案的播放启动时延对比关系

而改进后方案的播放启动时延大约在 5s 左右,是原来的 1/2,这有效地改善了用户体验感受。

## 4　结束语

通过对基于 BitTorrent 协议的改进,提出了基于自适应大小的滑动窗口模型解决和优化 P2P－VoD 系统部署过程中所遇到的启动时延较长和系统负载均衡较差等问题。试验结果表明所提出的伙伴节点选择策略、分块获取策略、分块服务策略等方案能够在一定程度上解决所提出的问题。当然 P2P－VoD 系统的应用和完善还有诸如版权控制、系统冗余控制、带宽管理、QoS 保证、可扩展性、可靠性、鲁棒性等诸多问题,将对这些方面做进一步的研究。

**参考文献:**

[1] Cheng B, Liu X, Zhang Z, et al. A Measurement Study of a Peer－to－Peer Video－on－Demand System[M]// IPTPS. Bellevue, WA:[s.n.],2007.

[2] Jiang X, Dong Y, Xu D, et al. GnuStream: a P2P Media streaming system prototype[C]// In Proceedings of the 4th International Conference on Multimedia and Expo. Baltimore, Maryland:[s.n.],2003.

[3] Luo J, Zhang Q, Tang Y, et al. A Trace－Driven Approach to Evaluate the Scalability of P2P－Based Video－on－Demand Service[J]. IEEE Transactions on Parallel and Distributed Systems, 2009,20(1):59－70.

[4] Shah P, Páris J －F. Peer－to－Peer Multimedia Streaming Using BitTorrent[C]// In IPCCC 2007. New Orleans, USA:[s.n.],2007.

[5] Cohen B. Incentives build robustness in BitTorrent[C]// In Proc. of First Workshop on Economics of Peer－to－Peer Systems. Berkeley, CA:[s.n.], 2003.

[6] Lu Z, Zhang S, Wu J, et al. Design and Implementation of a Novel P2P－Based VOD System Using Media File Segments Selecting Algorithm[C]// In 7th IEEE Intern. Conf. on Computer and Information Technology (CIT 2007). Washington DC, USA: IEEE Computer Society, 2007:599－604.

[7] Huang Y, Fu T T J, Chiu D M, et al. Challenges Design and Analysis of a Large－scale P2P VoD System[C]// In Proceedings of ACM SIGCOMM 2008. Seattle, Washington, USA:[s.n.], 2008.

[8] Cui Y, Li B, Nahrstedt K. oStream: Asynchronous Streaming Multicast in Application－layer Overlay Networks[J]. IEEE Journal on Selected Areas in Communications. Special Issue on Recent Advances in Service Overlays, 2004, 22(1): 91－106.

[9] Janardhan V, Schulzrinne H. Peer assisted VoD for set－top box based IP network[C]// In Workshop of Proc. of ACM SIGCOMM, P2P－TV'07. Kyoto, Japan:[s.n.], 2007.

(上接第 20 页)

System for Financial Forecasting[J]. Pattern Analysis and Applications, 1999, 2(3):264－273.

[10] Zhang G P. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model[J]. Neuron－computing, 2003, 50(1):185－198.

[11] 张立明. 人工神经网络的模型及其应用[M]. 上海:复旦大学出版社,1993.

[12] Leigh W, Hightower R, Modani N. Forecasting the New York Stock exchange composite index with past price and invest rate on condition of volume spike[C]//Expert System with Applications, 2005. Cambridge: Cambridge University Press, 2005.

[13] Yam J Y F, Chow T W S. Feed forward networks training speed enhancement by optimal initialization of the synaptic coefficients[J]. Neural Networks, IEEE Transactions, 2001, 34(5):73－85.

[14] Alan M S. The application of neural networks to predict abnormal stock returns using insider trading data[C]//Applied Stochastic Models in Business and Industry, 2002. [s.l.]: MIT Press, 2002.