

文章编号: 1001-9081(2005)09-2179-03

一种新的时间序列分析算法及其在股票预测中的应用

周广旭

(清华大学 软件学院, 北京 100084)

(zax02@mails.tsinghua.edu.cn)

摘 要: 分析了股票市场高度非线性特点, 给出了一种改进的时间序列分析算法。新算法利用径向基网络来对序列中的历史信息进行非线性组合, 从而比基于线性组合的时间序列分析算法的基本模型更能有效地挖掘出序列中历史信息之间的相互作用。新算法还利用改进的遗传算法对径向基函数的中心和宽度进行了全局范围的优化选择, 进一步提高了径向基网络的非线性映射能力。运用该算法对股票走势进行了预测, 取得了令人满意的效果。

关键词: 遗传算法; 时间序列分析; 径向基网络; 股票预测

中图分类号: TP183; TP391 **文献标识码:** A

RBF-based time-series forecasting

ZHOU Guang-xu

(School of Software, Tsinghua University, Beijing 100084, China)

Abstract: The nonlinear properties of stock information were analyzed, and a novel time-series forecasting algorithm was provided. The new algorithm introduced radial basis functions into the basic ARMA model to explore the interaction among past information, and then chose optimal parameters for RBFs using an improved genetic algorithm. Then, selected stock prices trend was forecast using the new algorithm and approving results were achieved.

Key words: genetic algorithms; time-series forecasting; radial basis functions; stock prices

0 引言

股票市场, 具有高收益与高风险并存特性, 而预测股市走势, 则一直被普通股民和投资机构所关注。但是由于股票市场的高度的非线性, 众多股市分析方法的应用效果都难如人意^[1]。

近年来, 计算机技术和人工智能技术的迅猛发展, 为股票市场的建模与预测提供了新的技术和方法^[1]。人工神经网络因其广泛的适应能力和学习能力在非线性的预测方面得到广泛应用。径向基神经网络是一种新颖有效的前馈式神经网络^[2], 由于该网络输出层是对隐层的线性加权, 使得该网络避免了反向传播那样繁琐冗长的计算, 具有较高的运算速度和推广能力, 该网络有较强的非线性映射功能。于是, 研究人员在用径向基神经网络预测股市走势方面做了大量的工作^[1,3-6]。在径向基神经网络中, 基函数中心的选取是至关重要的, 也是实现径向基神经网络的难点之一。一般采用聚类算法或神经网络方法求出输入样本的各类中心并将它们作为径向基函数的中心^[6]。另一个需要解决的问题是基函数宽度的选取, 这往往需要根据聚类的结果来确定, 最常见的是令它们等于聚类中心与训练样本之间的平均距离^[7]。但是, 这些方法找到的仅仅是局部最优的径向基函数中心, 而不是全局最优中心。而寻找全局最优的径向基函数中心的问题跟其他许多全局优化问题一样, 目前都没有找到有效的具有多项式时间复杂度的求解算法。

遗传算法是模仿生物进化过程的一种新兴的具有全局寻优能力的概率算法^[8,9]。理论上, 只要种群规模和进化代数足够大, 它是依概率 1 找到全局最优解的; 而且它是一种普适

性的算法, 即对任何一个优化问题 P, 都可以找到适当的编码方案并设计合适的遗传算子, 从而按照遗传算法的基本框架, 设计出求解 P 的遗传算法 GAP。实践上, 遗传算法也确实有许多出色表现, 它不仅广泛用于复杂、多维的非线性优化问题, 而且也成功地应用于许多 NP 困难问题的求解之中^[10]。所以, 有不少学者使用遗传算法来求解径向基函数的中心问题, 取得比聚类算法更优的结果^[11,12]。

时间序列分析是利用序列的历史信息以及历史信息之间的相互作用, 对序列的未来轨迹进行预测的一种数学方法, 它自从一提出就受到统计学和经济学研究领域的高度关注^[13-16]。一方面, 股票等金融信息是一个高度敏感的信号——经济、政治和社会领域里的很多事件都可以引起股票价格的波动; 另一方面, 这些经济、政治和社会领域里的许多“具有影响力”的事件并不像人们一般认为的那样是“突发的”或纯粹偶然的, 而是有一定的“前兆”。总有一些“消息灵通”人士会利用这些信号来指导自己的投资行为, 而这些投资行为反过来就会成为新的“前兆”。从这里可以看出, 如何尽可能充分而准确地挖掘出这些“前兆”信息, 对成功预测股票走势是至关重要的。前兆信息之间的相互作用正好体现了时间序列分析的特点。所以, 许多研究者都尝试了用时间序列分析的方法来预测股票的走势, 并取得了较好的效果^[12,13]。

实现时间序列分析技术的关键, 在于如何挖掘历史信息之间的相互作用信息。由于计算和建模方面的复杂性, 许多文献都简单地用历史信息的线性组合来表达这种信息。许多研究人员都意识到, 可以用历史信息的非线性组合来提高预测精度。但是用什么样的非线性组合最优呢? 这首先是一个

理论上没有解决的难题。但是,人工神经网络具有极强的非线性映射能力,这启发人们把人工神经网络技术引入到时间序列分析中来^[12],即不去关心具体用什么样的非线性组合,而是用人工神经网络去逼近那个未知的最佳映射。用什么类型的人工神经网络模型呢?文献[12]使用了径向基神经网络模型,并使用了遗传算法对径向基函数的中心和宽度进行优化选择。文献[12]利用这种新的方法,对1979年~1983年英镑对美元的周平均汇率走势进行预测:共选取了209个数据点,把前109个看作历史信息,而对后100个预测,并和已有的原始数据进行比较,结果较为满意,证实了这种新方法的有效性。

本文借鉴文献[1~7, 11~16]的思想,做了如下几方面的工作:第一,我们对原始信号做了归1化处理,使其更适合作为神经元的输入信号;第二,针对文献[12]的遗传算法部分做了改进,变固定交叉概率和变异概率为自适应的情形,进一步提高了遗传算法种群的多样性,这样,我们的算法就可以找到更优的基函数中心和宽度;第三,我们运用改进的算法对1999年~2004年上海证券公司某支股票(由于数据来源有保密限制,所以我们目前暂不公开该股票的名称)的周平均价格走势进行了预测,取得了很好的效果。

1 径向基神经网络的基本结构

径向基函数(Radial basis function, RBF)神经网络是基于人脑的神经元细胞对外界反应的局部性而提出的,是一种新颖而有效的前馈式神经网络,其具有最佳逼近性能和全局最优的特性^[1~3]。

RBF网络通常由输入层、隐含层和输出层组成,它包括两个阶段的数据处理结构。首先,由网络隐层的基函数对输入数据进行非线性变换;之后,网络的输出由基函数的响应通过输出层的加权组合给出^[4,5,12]。该网络实际上完成的映射可写为^[2,12]:

$$y(x) = \sum_{j=1}^n w_j h_j(\|x - c_j\|) \quad (1)$$

其中, x 是网络的输入向量, y 是网络的输出, w_j 是连续隐含层到输出层的可变权值, $h_j(\cdot)$, ($j = 1, 2, \dots, s$)是一种径向对称的基函数, $c_j \in R^n$ 为这些基函数的对称中心, $\|\cdot\|$ 一般取为欧氏距离。基函数具体形式的选取根据具体问题的背景可以有不同的选择,但是最常用的还是下面的高斯型函数(有的文献给出的高斯型函数与此略有不同,但是在本质上,它们都是等价的,最多相差一个正常数因子):

$$h_j(d) = \exp\left[-\frac{d^2}{r_j}\right] \quad (2)$$

这里的 r_j 称为基函数 $h_j(\cdot)$ 的半径或宽度。一旦 c_j, r_j ($j = 1, 2, \dots, s$)确定下来,那么如果给定了对应于输入序列 $|x_i|_n$ 的理想输出序列 $|y_i|_n$,则可以要求下式达到极小化来构造网络的训练算法:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y(x_i))^2 \quad (3)$$

2 时间序列分析的基本模型和改进

自回归移动平均模型(简称ARMA模型)是时间序列分析方法的基本模型,其一般形式为^[12~14,16]:

$$y(k) = \sum_{j=1}^n a_j y(k - \tau_j) \quad (4)$$

这里, $y(k - \tau_j)$ 表示系统过去的输出, a_j 表示加权系数,参数 n 表示“向回看”(即向过去看)的历史信息的个数(仅当 $\tau_j = j$ 时,表示恰好连续回看前面 n 个历史信息;否则,向回看的方式就不一定是连续的),这个 n 也经常称为自回归模型的“阶数”。

该模型的缺陷就在于它是历史信息的线性组合,对于一些突发事件缺少必要的预测能力;甚至在高阶情况下,即 n 较大时,也不能使该模型的预测能力有显著的提高。文献[12]给出的改进方案就是,利用径向基函数的非线性映射能力,实现历史信息的非线性组合。改进的时间序列分析模型为:

$$y(k) = \sum_{j=1}^n w_j h_j(y(k - \tau_j)) \quad (5)$$

其中, $w_j, h_j(\cdot)$, ($j = 1, 2, \dots, n$)分别如(1)式和(2)式中定义的权值和径向基函数, n 仍然表示模型的阶数。

接下来需要解决的问题就是:如何优化选择径向基函数的中心和宽度?为此,借鉴文献[12]的思想,我们也采用遗传算法来实现,但是适应值函数设计、种群多样性保持以及各个遗传算子设计等方面,我们对文献[12]中的方案做了必要的改进^[17]。

3 基函数中心和宽度优化选择的遗传算法设计

遗传算法设计的关键是确定遗传评价环境^[8~10],即模拟“优胜劣汰”的自然选择环境。这包括两方面的内容:一是个体的染色体编码如何表示;二是个体的适应值如何计算。

3.1 个体编码方案及适应值函数设计

采用变长染色体编码方案。设个体 a 的染色体形如

$$a = \begin{bmatrix} c_1 & r_1 & w_1 \\ \cdots & \cdots & \cdots \\ c_{n_a} & r_{n_a} & w_{n_a} \end{bmatrix} \quad (6)$$

其中, n_a 为个体 a 所代表的模型(5)的阶数,而 c_j, r_j, w_j ($j = 1, 2, \dots, n$)则分别是该模型中相应基函数的中心、宽度和权值。

给定个体 a 以后,就可以确定模型(5)中的所有参数(我们恒取 $\tau_j = j$),从而可以利用历史信息作为训练序列,按照(3)式计算相应网络输出的最小平方误差,记为 $MSE(a)$ 。现在,个体的适应值函数可以设计为如下形式:

$$f(a) = \frac{1}{1 + MSE(a)} \quad (7)$$

3.2 种群进化策略和遗传算子设计

给定种群规模 P ,初始种群随机生成,每代选择 $0.8P$ 个较优个体进入交配池,另外随机生成 $0.2P - 1$ 个个体直接进入下一代,并且采取“保优”策略,让当代最优个体的一个副本直接进入下一代。

选择算子:采用改进的一次旋转赌盘的算法。

交叉算子:采用两个染色体的“一致交叉算子”来实现。由于染色体不等长,所以此一致交叉是标准一致交叉算子的一个变种。详述如下:设参与交叉的两个染色体为 a 和 b ,则先把染色体 a 和 b 并置在一起构成一个长度为 $(n_a + n_b)$ 行3列的矩阵,对该矩阵的第 s 行,依交叉概率 p_c 进行“搬迁”——如果原来 s 行属于 a ,则搬迁后属于 b ;反之亦然。

变异算子:让染色体的 c, r, w 基因片段分别加上某个随机值(并确保在各自的取值范围内)。

3.3 自适应的交叉概率和变异概率

文献[12]中的交叉概率和变异概率是固定的,分别为0.85和0.01。这样高的交叉概率很容易使优势个体控制种

群,出现“早熟”收敛。本文引入自适应的策略,让交叉概率和变异概率随种群多样性的变换而进行动态调整。具体调整公式如下^[17]:

$$p_c = \frac{1}{2}(1 + \tanh(6 \cdot V(F) - 3)) \quad (8)$$

$$p_m = 0.05(1 - p_c)$$

其中, F 看作是由当代种群中的个体的适应值所定义的随机变量, $V(\cdot)$ 为 F 的方差。这是一个S型的函数,它模拟单个神经元的连续型响应函数,当种群成熟度过高时,降低交叉概率,增大变异概率;当种群成熟度过低时,它又自动增大交叉概率,降低变异概率,从而实现种群多样性的自适应调节^[17]。

4 股票走势预测结果及分析

基于前面的分析和设计,我们用Matlab编写了相应的计算程序,并选取了上海证券公司某支股票1999年12月21日~2004年12月3日共约6年期的周平均价格数据进行试验:用前半数据作为历史数据,对后半数据点进行预测。在我们的算法中,种群规模和最大进化代数均为100。多次运行都得到比较好的效果。图1~图2分别给出了两次比较好的结果(图中虚线为原始数据,共705点;实线前352点为网络训练和拟合过程,从353点开始为预测数据)。

可以看出,总体预测效果还是很令人满意的。尤其是在353点到400点之间,预测信号和原始信号还是吻合得非常好。

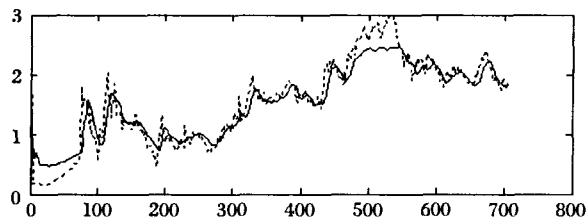


图1 某次比较好的预测效果(1)

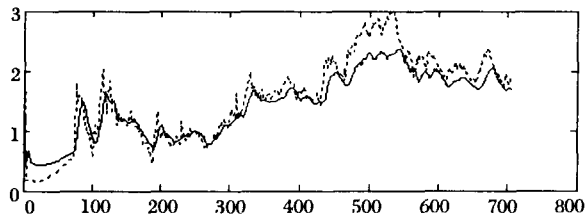


图2 某次比较好的预测效果(2)

5 结语

融合了遗传算法和径向基神经网络技术,设计了一种新的时间序列分析算法:(1)使用径向基神经网络来挖掘序列中历史信息之间的相互作用;(2)利用改进的遗传算法,对径向基函数的中心和宽度进行全局优化选择;(3)给出了完整的遗传算法设计,并实现了相关的软件包。最后,运用该计算软件包对股票走势进行了预测,取得了较为满意的结果。另外,从预测结果还可以看到一个有用的信息:如果每次不是用前半数据去预测后半数据,只是用历史数据预测很少几个新的数据点,效果会更好。这也是进一步改进算法的方向之一。

在设计、调试和反复测试该算法的过程中,还总结出以下经验:

1) 理论上,只要能够确定基函数中心的取值范围,然后在该范围内随机生成,使用GA算法,只要种群规模和进化代

数足够大,将依概率1找到全局最优的中心。而至于取值范围的确定,可以通过对输入样本做归1化处理,将其确定为 $[0,1]$ 区间;

2) 关于 c 的考虑:以初始样本作为基函数中心的初始种群而不是随机在 $[0,1]$ 区间去产生,相当于首先让这几个样本点关于中心所引起的误差降到最小,然后考虑对所有训练样本的偏差,在交叉、变异等遗传算子的操作下,仍然可以认为——从理论上,基函数的中心是从 $[0,1]$ 区间上进行搜索的;

3) 关于归1化的理由是:神经元的输入约定都是 $[0,1]$ 之间的数,必须做这样的归1处理,才能为神经网络所接受;

4) 关于 w 的考虑:首先可以考虑的是取 w 为 $[0,1]$ 之间的实数,这就相当于让所有的基函数做凸组合。如果简单考虑线性组合,那么相当于假定某些基函数之间是起互相抵消作用的(取值正负相抵),这实际上是不可能的,因为高斯型的基函数恒为正值,所以只需要考虑凸组合即可。再进一步,假定所有 M 个基函数的贡献都相等,那么 w 的取值都应当为 $1/M$;如果他们的贡献不等,贡献大的基函数在取值上已经较大,没有必要让它在权重上再大,所以,设定的 w 的取值范围为 $[0,1/M]$ 。当然了,从理论上,在GA进化机制作用下,设置 w 的区间为 $[0,1]$ 也完全可以,但是缩小搜索空间有利于加快算法收敛。而且,通过多次预测试算,发现 w 的取值区间对算法性能影响非常大,试算的结果是取 $[0,1/M]$ 较好;

5) 关于 r 的考虑:同样,首先可以考虑的是取 r 为 $[0,1]$ 之间的实数,但是取0会引起基函数计算异常,因为 r 要作为分母出现;另外,考虑到如果 r 太小时,对应的基函数的取值接近于0,即它对于拟合没有什么贡献。 r 既然作为基函数的作用半径,那么 r 平均值应在0.5左右,最大为1。所以,建议用 $[1/M/M, 1]$, $[1/M, 1]$ 甚至 $[0.5 - 1/M, 0.5 + 1/M]$ 作为 r 的取值区间。但是在多次试算后,发现算法对这几种取法的表现都大致一样。所以,关于 r 的这种考虑还不能成为已经被证实的经验使用。

综上所述,采用的做法是:中心 c 初始是从输入样本中随机选取,但是在变异时不再局限于输入样本,而是在 $[0,1]$ 上考虑; w 在 $[0,1/M]$ 上随机取,变异时仍然确保它在 $[0,1/M]$ 上; r 在 $[0,1]$ 上随机取,变异时仍然确保它在 $[0,1]$ 内(至今运行还没有碰到 r 取0的情况。但是理论上,这种除0引起的异常是存在的,应该避免,推荐使用 $[1/M/M, 1]$)。

参考文献:

- [1] 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.
- [2] HLAVACKOVA K, NERUDA R. Radial basis function network [J]. Neural Network World, 1993, 3(1): 93 - 101.
- [3] 叶东毅, 刘文标. 径向基函数神经网络在股票走势模式分类中的应用[J]. 运筹与管理, 1999, 8(3): 46 - 50.
- [4] 郑丕涛, 马艳华. 基于RBF神经网络的股市建模与预测[J]. 天津大学学报, 2000, 33(4): 483 - 486.
- [5] 徐绪松, 熊保平, 龙虎. 用RBF神经网络确定上海股市的分形维数[J]. 武汉大学学报(理学版), 2003, 49(3): 309 - 312.
- [6] 朱赞, 王行愚. RBF神经网络在股市趋势预测中的应用[J]. 华东理工大学学报, 2002, 28(5): 547 - 550.
- [7] 叶东毅, 刘文标. 个股走势模式分类的RBF神经网络方法[J]. 福州大学学报(自然科学版), 2000, 28(4): 12 - 15.
- [8] HOLLAND JH, MILLER JH. Genetic Algorithms[J]. Scientific American, 1992, 267(1): 44 - 50. (下转第2184页)

支持。

信息集成服务作为用户和元数据目录服务的桥梁,负责提交用户数据访问申请,按照定义规范访问请求,并将从不同数据源中获取的数据结果返回用户。

下面简要介绍地震信息网资源访问的过程:

(1) 用户在提供的信息网格交互界面中申请访问地震信息;

(2) 系统将带有查询条件的访问请求传递给信息集成服务,信息集成服务将请求解析后发送给指定层元数据目录服务进行处理;

(3) 如果查询的元数据不在本层节点上,需要根据数据备份策略,从相关层选取所需的数据,然后运用数据备份管理从相关层备份数据;

(4) 如果查询的元数据在本层节点上,元数据目录服务根据请求中所引用的虚拟视图名查询虚拟视图目录,获取虚拟视图元数据值,系统根据虚拟视图中的引用情况和其他访问需求分解出相关的业务对象、条件约束等相关访问信息;

(5) 元数据目录服务查找到所有相关业务对象信息,在业务对象可调用实体集中选取合适的数据集元数据;

(6) 元数据目录服务获取所需数据集元数据以及相关存储系统元数据的所有信息,并将该信息传送到信息集成服务;

(7) 信息集成服务根据获取的数据集元数据以及相关存储系统元数据,构建本次访问请求的查询任务序列,并通过相应存储系统的适配器将每一个查询任务传递到存储系统中执行;

(8) 信息集成服务负责收集所有查询结果,根据组合规则对信息进行集成加工后返回给应用程序或最终用户。

地震信息网元数据目录管理体系结合了地震应用领域的特点和元数据模型定义,很好地把元数据的概念运用在实际应用中。通过多层数据定义、转换,将地震海量数据有规律地组织起来,使用户有效地访问所需数据。

3.2 实现方式

地震信息元数据目录管理系统采用 Globus 开发平台, Java 开发语言,并以图 4 为框架构建系统。

在上述地震减灾仿真网络系统结构图中,元数据服务器用于对地震元数据信息的建立和维护。地震信息的元数据存储于服务器的元数据库中。

元数据交换服务器用于连接、管理不同层次的元数据服务器,组成分布式结构。元数据交换服务器负责将地震信息资源的交换请求转发到相关元数据服务器,并实现数据的备份。

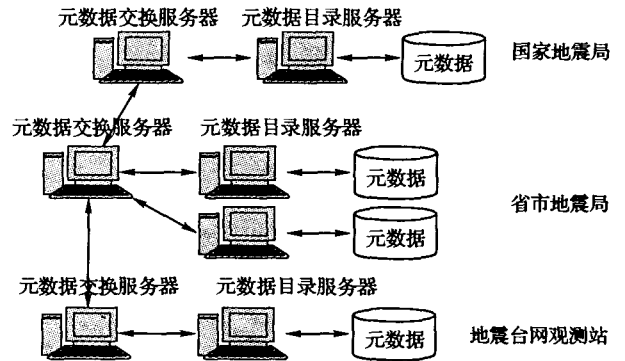


图4 地震减灾仿真网络系统结构

4 结语

目前网络技术还处于发展阶段,国内已有的网络应用系统有限,地震减灾科学计算网络系统是地震应用领域首例网络应用系统,它的意义重大,对我国网络应用技术的发展起到了重要作用。

通过分析地震应用领域的实际情况,运用虚拟元数据模型和元目录服务构建地震减灾仿真网络系统的元数据服务管理体系,实现了分布异构数据访问,为在地震应用领域的数据库管理提出了可行性方案。下面的工作是进一步完善元数据管理平台(包括备份策略管理方案等)。

参考文献:

- [1] FOSTER L, KESSELMAN C. 网络计算[M]. 金海,袁平鹏,石柯,译. 北京:电子工业出版社,2004.
- [2] 都志辉,陈渝,刘鹏. 网络计算[M]. 北京:清华大学出版社,2002.
- [3] 廖华明,程伯羽,刘新周,等. 信息网格中元数据层次化结构模型的研究和应用[J]. 计算机研究与发展,2003,(12).
- [4] 徐志伟,冯白明,李伟. 网络计算技术[M]. 北京:电子工业出版社,2002.
- [5] 王意洁,肖依,任浩,等. 数据网格及其关键技术研究[J]. 计算机研究与发展,2002,(8).
- [6] 肖依,任浩,徐志伟,等. 基于资源目录技术的网格系统软件设计与实现[J]. 计算机研究与发展,2002,(8).
- [7] FOSTER I, KESSELMAN C, NICK J, et al. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration[EB/OL]. Globus Project. USA, 2002.
- [8] CHERVENAK A, FOSTER L, KESSELMAN C, et al. The Data Grid: Towards an Architecture for Distributed Management and Analysis of Large Scientific Datasets[EB/OL]. Globus Project. USA, 2002.

(上接第2181页)

- [9] DENNING PJ. Genetic Algorithms[J]. American Scientist, 1992, 80(1): 12 - 19.
- [10] WINTER G, et al. An Introduction on Global Optimization by Genetic Algorithms[J]. Algorithms for Large Scale Linear Algebra Systems, 1998: 343 - 367.
- [11] 周佩玲,陶小丽,傅忠谦,等. 基于遗传算法的RBF网络用于股票短期预测[J]. 数据采集与处理, 2001,16(2): 249 - 252.
- [12] SHETA AF, JONG KD. Time-series forecasting using GA-tuned radial basis function[J]. Information Science, 2001, 133(3-4): 221 - 228.
- [13] 周家. 利用时间序列分析股票价格和会计盈利的动态关系[J].

现代财经——天津财经学院学报, 2004, 24(169): 35 - 40.

- [14] HOLGER K, THOMAS S. Nonlinear time series analysis[M]. Beijing: Tsinghua university publishing company, 2001. 233 - 234.
- [15] MULLOY BS, RIOLO RL, SAVIT RS. Dynamics of genetic programming and chaotic time series prediction[A]. Proceedings of the first Annual Conference on Genetic Programming[C]. Stanford: Stanford University, 1996. 166 - 174.
- [16] 许国辉,余春林. 时间序列分析方法的研究[J]. 广州大学学报(自然科学版), 2003, 2(6): 556 - 559.
- [17] 王励成. 人工神经网络和遗传算法在数学优化中的应用[D]. 南京: 南京大学硕士学位论文, 2001.