发现金融市场预测模型的计算智能方法*

童 频 费良俊

(上海大学计算机科学系 上海 201800)

摘要 文章首先扼要论述了金融市场数据以及计算智能方法学的基本性质及其在数据发掘中的应用前景,提出了一个用遗传算法配合神经网络进行优化训练后,用于发现股票市场价格变化趋势和预测模型的实验系统.文章着重论述了这一系统的设计思想和实现技术.最后,随意选择上海中百一店股票行情为实验研究对象,给出了用所述方法进行预测的实验结果.

关键词 金融市场预测,计算智能,遗传算法,神经网络.

中图法分类号 TP18

(1) 金融市场数据的基本性质

市场价格波动作为一种社会现象,不但受到供求关系这一普遍规律的制约,而且还受到参加交易的人群中流行的主观意念的影响.而参加者的主观意念又受到市场价格波动的影响.设 y 代表市场价格的波动,x 代表参加者的主观意念,便可用以下两个函数表示二者的相互影响:

$$y=f(x);\ x=g(y).$$

由此可导出以下两个递归函数:

$$y=f(g(y));\ x=g(f(x)).$$

这就是关于市场价格波动的"反射"理论(Theory of Reflexivity)[1].这一理论揭示了市场、特别是金融市场价格波动的不稳定性和混沌性的本质:永远是混沌性的涨、跌交替序列(boom/bust sequence)和良性、恶性相继循环(benign/vicious circles).这也说明了为什么经典的数学方法或传统的 AI 技术难以用来对具有这种性质的大量历史数据记录进行发掘、分析,发现其中蕴涵的规律性或预测模型,用来预测市场发展趋势、帮助贸易或交易的优化决策等的原因.而计算智能方法本身就有的自适应优化、自组织学习机制,正好可以在这方面发挥作用.因此,早已引起了以华尔街为代表的金融界的重视,投入大量人力、物力和财力进行相应的应用开发[2].

(2) 计算智能的方法学性质与特点

所谓计算智能(computational intelligence,简称 CI),这里主要是指:在微观层次上模仿神经网络功能的神经计算(neural computing,简称 NC)和模仿生物进化过程的演化计算(evolutionary computing,简称 EC).它又包括在微观层次上模仿生物遗传变异机制的遗传算法(genetic algorithm,简称 GA)和在宏观层次上模仿自然界"物竞天择"进化规律的演化程序设计(evolutionary programming,简称 EP)或演化策略(evolutionary strategy,简称 ES)[3].EP 和 ES 在算法上可谓大同小异,只是前者发源于美国,着眼于不同种群之间的竞争;后者发源于德国,着眼于种群内个体之间的竞争.

从理论上讲,与传统人工智能(AI)相比,CI 的最大特点,也即它的潜力所在是:它不需要建立问题本身的精确(数学或逻辑)模型,也不依赖于知识表示,而是直接对输入数据进行处理,得出结果.以神经网络为例,只要把数据输入到一个已经训练好的神经网络输入端(相当于刺激模式),就可从输出端直接得到预期的结果(反应).这个神经网络本身是一个"黑箱".问题是如何训练神经网络.但这类训练算法并非求解问题本身的算法,而是使

---

* 本文研究得到 NSFC 基金和国家 863 高科技项目基金资助.作者童频,1930年生,教授,博士生导师,主要研究领域为人工智能,数据库与知识工程,计算机体系结构.费良俊,1968年生,讲师,主要研究领域为人工智能,数据库与知识工程.

本文通讯联系人:童频,上海 201800,上海大学(嘉定校区)计算机科学系

本文 1997-11-20 收到原稿,1998-04-03 收到修改稿

神经网络"学会"怎样解决问题的算法,可以说是一种准元算法;类似地,演化算法,无论是模仿生物通过遗传,还是"物竞天择"的自然进化机制来达到优化的目的,都是模仿生物进化方式的算法,而不是根据问题本身的数学/逻辑模型来制订的算法,所以也是一类准元算法.

正是CI的上述特点,使它适用于解决那些因为难以建立有效的形式化模型而不存在直接求解算法、或带有"不可言传"性质的问题.特别是对高维非线性随机动态或混沌系统行为的分析和预测,例如,从反映市场价格变动及其相关诸因素的数据中发现预测模型更是如此.

1 发现金融市场预测模型的计算智能方法

针对金融市场数据变化动态的混沌性质,我们研究开发了一个复合运用不同计算智能方法来实现的、预测股市行情变化趋向的辅助决策系统[4].这一系统的主体是一个多层感知器(双层前馈结构的神经网络).通过根据大量股市历史记录数据构造的样本集进行训练后,发现其中隐含的预测模型,同时用遗传算法(GA)对神经网络的结构、性能参数进行优化.经过这样优化训练后的神经网络,将能根据当前,例如本周股市行情变化(输入)来预测下一周的变化趋向(输出).这正是一个数据发掘/知识发现的过程[5].本文将讨论用如图1所示的神经网络和遗传算法复合协同进行优化训练的设计思想、实现机制及其用于发现股市行情变化预测模型的实验结果.

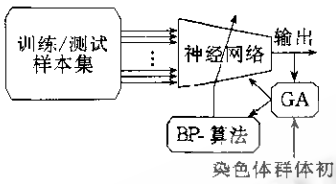


图1 神经网络和遗传算法协同优化训练

1.1 神经网络输入数据结构及训练样本集

一个训练样本的数据结构是一个$n+m$维向量:$(x_1,x_2,\ldots,x_i,\ldots,x_m,y_1,y_2,\ldots,y_j,\ldots,y_n)$.其中$x_i|i=1,\ldots,m$是神经网络的输入数据集;$y_j|j=1,\ldots,n$是神经网络的预期输出数据集.在本文所提出的系统中,输入数据集X主要包含股市本身可量化的技术指数(technical indecise)和影响市场的基本因素(fundamentals).为了压缩输入数据的实际动态变化范围,提高训练效率及训练结果的普适性,必须对输入数据进行变换和规格化.在本文所提出的系统中,最终选定了综合股价指数、总成交量、股价指数上涨、下跌、不变的企业数、某一选定股票(本例中选择了"中百一店")的成交股数、银行利率等共13项指数的周平均值的每周变化量,再加上代表当前周所处月份的季节因素.相应的输出数据子集则包含股票价格涨、跌和成交量升、降(皆取其百分比)两项.

结合用GA优化神经网络结构及训练性能的要求,把按上述数据结构采集的输入样本集分成互不相交的3个子集:一个训练子集和两个测试子集.训练子集和第1个测试子集用于对神经网络的训练;第2个测试子集中包含最能检测网络性能的样本,用于配合遗传算法中适合度函数(fitness function)值的计算、评估不同结构和控制参数的网络性能,同时,也用于检验最终优选出来的神经网络预测模型性能.

1.2 输入数据的时序性

金融市场数据无时无刻不在变化,显然带有明显的时序性.在预测股市行情变化趋势的统计分析中,经常使用以一定时间段为准的移动平均值(moving average)[2].例如,短期的有数日平均、周平均、旬平均,较长的有月平均、季平均、年平均等等.我们采用了周平均值,即每周做一次输入数据采样.为了能恰当地反映时序性质的影响,每一个训练样本中的输入数据X_i都是一个由3个顺序采样的数据$X_i(t),X_i(t-1),X_i(t-2)$组成的子集,其中时间t的单位是周.这样,下一个样本集的同一输入数据子集X_i将是$X_i(t-1),X_i(t-2),X_i(t-3)$等等,依此类推.可见,输入数据集将从当前周开始,向后每间隔一周,以3周为时间窗口,采集3个数据作为同一指数的输入.先后两个样本间,将有两周时间重叠,相应的输入神经元的数目也将扩大为3倍.

1.3 GA_BP复合协同训练/学习算法

单纯采用反传(BP)算法对这一双层感知器网络进行训练时,训练的收敛速度在很大程度上取决于其结构和控制参数的选择.然而,诸如学习率、动量系数、隐单元数目及初始权值范围等参数,一般都是通过实验或直接凭经验来选择的.本文提出把遗传算法与神经网络的训练/学习算法相互配合,对神经网络的结构、性能进行优化的方法.所采取的复合协同演化训练/学习算法,大体上可表述如下:

```
GA_BP(Crm_NN):
  P(i):=⋃_{j=1}^{n} Crm_NN(j)        //构成染色体群体,每个染色体对应于神经网络的一组参数.群体的规模限定为 n.i 代表群体繁殖的"代"数,i=1,即 P(1)为初始化群体//
  for i=1 to N do                    //从 P(1)→P(N),对每一代 P(i)中的每一个 Crm_NN(j),用第 2 组测试样本集按 BP-算法进行训练,并计算其拟合度函数值//
    for j=1 to n do
      training_bp (Crm_NN(j),sample_test(2))      //调用 BP-算法//
      evaluate (φ(fitness),Crm_NN(j))
    P(i+1):=apply (GA,P(i))          //调用遗传算法,产生新一代群体 P(i+1)//
  select (Crm_NN(x)∈P(N)|φ(fitness) for x⇒max/admissible)
                                     //从繁衍的最后一代 P(N)中,选出拟合度函数值为最大或可接受的 Crm_NN(x),作为所发现的神经网络预测模型//
```

1.3.1　遗传算法的数据结构

大家知道,遗传算法是对生物在繁殖过程中,通过染色体的交配和遗传基因的适应性变异而达到优化目的(改良品种)的一种模拟.GA 的基本数据结构,就代表一个"染色体".在这里,每一个"基因"代表一项神经网络的结构或训练控制参数.共选择了双层感知器及其 BP 训练算法涉及的 5 项参数:双层感知器网络中的隐单元数目、学习率、动量系数(momentum)、输出的允许误差和初始权值的取值范围.图 2 就是一个"染色体"的例子.它代表了神经网络的一组相应的结构和学习性能参数.

基因1	基因2	基因3	基因4	基因5
72	0.7	0.9	0.05	0.4
隐单元数目	学习率	动量系数	允许误差	初始权值范围

图2　一个染色体及其基因

1.3.2　拟合度函数

在染色体较少的 GA 中,可能存在一些适合度比平均适合度高得多的个体.经过少数几代演化后,这类个体在种群中占很大比例,不利于引进新的个体,容易引起过早收敛[6].为避免这种现象,我们设计了一个能够随 GA 执行过程的进展自动调整的拟合度函数(fitness function):

$$\varphi(fitness) = (\bar{\varepsilon} \times c_{bp})^{-g/G}. \tag{1}$$

式(1)中 c_{bp} 是用 BP 算法训练神经网络,使其输出误差达到允许范围时所进行的反传迭代次数;g 代表 GA 算法"繁殖"到当前代时的"代"数;G 是预期要"繁殖"的总"代"数.$\bar{\varepsilon}$ 是用第 2 组测试集作为输入时,某个染色体所对应的神经网络输出的平均误差,可计算如下:

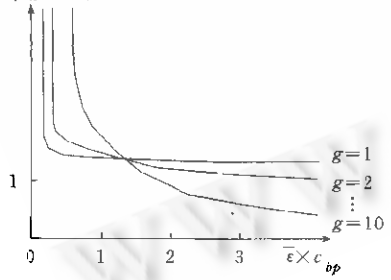


图3　$\varphi(fitness)$,$\bar{\varepsilon}\times c_{bp}$,$g$ 三者间的关系

$$\bar{\varepsilon} = \prod_{l=1}^{n} \sqrt{\frac{\sum_{k=1}^{n} (e_{kl} - y_{kl})^2}{S}}. \tag{2}$$

其中 e_{kl} 是关于第 k 个输入测试样本集的第 l 个输出的期望值;y_{kl} 是相应的实际输出值;n 是输出神经元总数;S 是测试样本集的总数.

$\varphi(fitness)$ 的值将随 $\bar{\varepsilon}\times c_{bp}$ 的增大而减小,又受到繁殖"代"数(参数 g 和 G)的制约,从而在 GA 优化过程中能自动调节每一代染色体群体间 $\varphi(fitness)$ 值之差异:在最初几代($g \ll G$),差别较小,使各种染色体都有参加竞争的机会;随着 $g \to G$,差别逐渐拉大,使拟合度值较大的染色体脱颖而出(如图 3 所示).这将有助于找到全局最优的染色体,从而获得一个训练周期短而性能好的神经网络预测模型.

1.3.3　交配和变异策略

当染色体群体初始化后,首先计算出每个染色体的适合度,算出它相对于全部染色体适合度总和的百分比.使用"转轮盘"(spinning a roulette wheel)的方法随机选择双亲染色体.对于被选为"双亲"的两个染色体中每个对位基因,产生一个随机数 $r=CRand(0,1)$,若 $r<0.5$,则复制"父"染色体中对应的基因;否则,复制"母"染色体中对应的基因.对两个双亲染色体中每一个基因进行这样的复制操作后,便形成了一个"后代"染色体,接着对这

个染色体进行变异操作.我们采用的是"蠕变"(creep)策略,即对其中的每一个基因,产生一个随机数(即变异概率)$q=CRand(0,1)$,若 $q>0.5$,就对这个基因执行变异操作;否则,不执行.变异操作的方法是在该基因原有值上,加上一个不大的、平均值为 0 的随机实数,

$$\Delta_n=CRand(-1,+1)\times\Delta_{max}. \tag{3}$$

其中 $CRand(-1,1)$是 Cauchy 分布随机数发生器产生的值;Δ_{max}是变异的限度.Δ_m 的取值将因"基因"所代表的意义不同而不同.例如,基因是隐单元数目,则 Δ_m 必须取整数;基因是学习率,则 Δ_m 必须取小于 1 的实数等等.如果由式(3)得出的值超出了相应基因允许的取值范围,则自动恢复该基因原值.

1.3.4　群体更新策略

从总体上来说,将按优胜劣汰的自然选择规律,在新一代群体中,保留 $\varphi(fitness)$值高的染色体而淘汰 φ 在临界值以下的染色体,以保证最终达到优化目标.为此,有多种不同的实现途径.在本系统中,群体中的个体是按照其适合度减小的顺序排列的.当添加一个新的个体时,对群体进行二分查找,找到要插入新个体的位置,这样,在保持群体大小不变的前提下,排在最后的、适合度最小的个体就自动被淘汰.

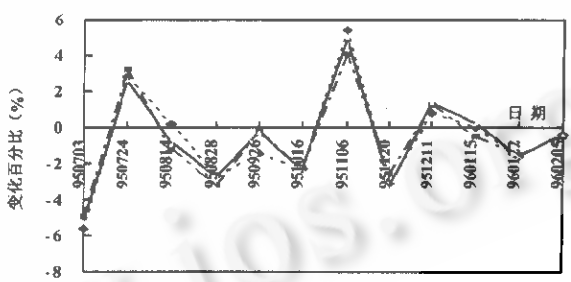
2　实验结果

在执行本文提出的 GA_BP(Crm_NN)算法后,种群中的第 1 个染色体(拟合度函数值最大)所对应的参数,就是所发现的优化神经网络预测模型.我们选择了上海第一百货有限公司的股票价格和成交量作为实验对象.数据的主要来源是上海《解放日报》每周公布的一周股市行情.用 GA_BP 复合协同算法训练的结果,获得了两个优化性能接近的神经网络预测模型,其结构和学习性能参数对应于以下两个染色体:

72　0.7　0.9　0.05　0.4　(染色体 1)

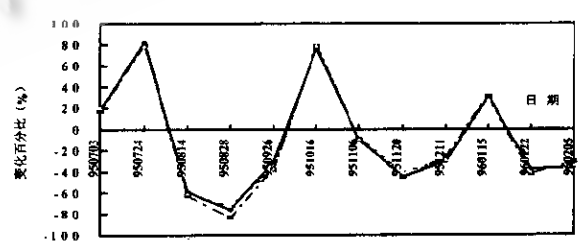
30　0.4　0.3　0.1　0.8　(染色体 2)

它们的 BP 迭代次数 c_{bp}和拟合度函数 $\varphi(fitness)$值分别是 374、0.493 1 和 419、0.507 9.分别用它们来预测下一周的价格变化百分比和成交量变化百分比,其结果与实际值的对比分别如图 4、5 所示.



—·—·■ 实际价格　……◆ 染色体 1 结果　——▲ 染色体 2 结果

图 4　价格预测值和实际价格变化的比较



—·—·■ 实际成交量　……◆ 染色体 1 结果　——▲ 染色体 2 结果

图 5　成交量预测值和实际成交量变化的比较

由图可见,使用本文提出的方法所发现的神经网络预测模型,对股票市场下一周的价格和成交量变化率进

行预测，得到了与实际数据十分接近的结果. 它可以用来发现全局最优的预测模型. 这种模型与未经 GA 优化的神经网络模型相比，具有收敛速度快、预测误差小等优点[7]，并可避免训练过程中的踏步、振荡、过度等现象. 只是由于 GA 中各代的每个染色体都代表一个结构参数和学习参数不同的神经网络，所以都要用 BP 算法进行训练，并计算其适合度. 因此，要最终得出一个优化的神经网络预测模型，所需要的计算时间很长. 我们将进一步研究改进方法，其结果将另文发表.

参考文献

1 Soros George. The Alchemy of Finance——Reading the Mind of the Market. New York: John Wiley & Sons, Inc., 1984～1985

2 Hoptroff R G. The principles and practice of time series forecasting and business modelling using neural nets. Neural Computing and Applications, 1992, I: 59～66

3 Zurada J M *et al*. Computational Intelligence: Imitating Life. New York: IEEE Press, 1994

4 费良俊，童颁. 计算智能技术在金融市场分析中的应用. 见：李乃奎等编. 第4届中国人工智能联合学术会议(CJACI'96)论文集. 北京：清华大学出版社，1996. 353～360
(Fei Liang-jun, Tong Fu. Computational Intelligence Application for Financial Market Analysis. In: Li Nai-kui *et al* eds. Proceedings of the 4th Chinese Joint Conference on Artificial Intelligence. Beijing: Tsinghua University Press, 1996. 353～360)

5 Evangelos Simoudis. Reality check for data mining. IEEE Expert, 1996, 11(5): 26～33

6 Guo Z, Uhrig R E. Using Genetic Algorithm to Select Inputs for Neural Networks. In: Proceedings of COGANN'92. IEEE CS Press, 1992. 252～261

7 Adeshiyan S A. Stock martket price prediction using neural networks [MS. Thises]. Shanghai: ShangHai University, 1994

Computational Intelligence Approach for Discovering the Prediction Model of Financial Market

TONG Fu　FEI Liang-jun

(*Department of Computer Science　Shanghai University　Shanghai*　201800)

Abstract　The nature of the stock market data is briefly discussed first. It follows a brief discussion on the nature of the methodology of computational intelligence and their application perspectives for data mining. An experimental system for discovering the trend of market price changing and the prediction model from the stock exchange data records are proposed, in which, the training parameters of a neural network are optimized and defined by running the genetic algorithm along with the training procedure of the neural network. The ideas of design and the techniques of implementation are described in detail in this paper. Taking arbitrary the Shanghai First Department Store for case study, the experimental results are given finally.

Key words　Financial market prediction, computational intelligence, genetic algorithm, neural network.