DOI:10.16208/j.issn1000-7024.2009.21.067

• 人工智能 •

基于隐马尔可夫模型的股票价格预测组合模型

朱嘉瑜[1]， 叶海燕[2]， 高 鹰[3]

(1. 广州大学 数学与信息科学学院，广东 广州 510006；2. 广东商学院 华商学院会计系，广东 广州 511300；
3. 广州大学 计算机科学与教育软件学院，广东 广州 510006)

摘 要：提出了一种用于股票价格预测的人工神经网络(ANN)，隐马尔可夫模型(HMM)和粒子群优化算法(PSO)的组合模型-APHMM 模型。在 APHMM 模型中，ANN 算法将股票的每日开盘价、最高价、最低价与收盘价转换为相互独立的量并作为 HMM 的输入。然后，利用 PSO 算法对 HMM 的参数初始值进行优化，并用 Baum-Welch 算法进行参数训练。经过训练后的 HMM 在历史数据中找出一组与今天股票的上述 4 个指标模式最相似数据，加权平均计算每个数据与它后一天的收盘价格差，则今天的股票收盘价加上这个加权平均价格差便为预测的股票收盘价。实验结果表明，APHMM 模型具有良好的预测性能。

关键词：股票价格预测；隐马尔可夫模型；隐马尔可夫模型优化；粒子群优化算法；人工神经网络

中图法分类号：TP301.6 **文献标识码**：A **文章编号**：1000-7024 (2009) 21-4945-04

Fusion model of hidden Markov model for stock market forecasting

ZHU Jia-yu[1]， YE Hai-yan[2]， GAO Ying[3]

(1. College of Mathematics and Information Sciences， Guangzhou University， Guangzhou 510006， China；
2. Accounting Department of Huashang College，Guangdong University of Business Studies，Guangzhou 511300，China；
3. College of Computer Science and Education Software，Guangzhou University， Guangzhou 510006， China)

Abstract：A fusion model APHMM is proposed by combining the hidden Markov model（HMM)， artificial neural networks（ANN) and particle swarm optimization（PSO) to forecast financial market behavior. In APHMM， use ANN to transform the daily stock price into independent sets of values and become input to HMM. Then draw on PSO to optimize the initial parameters of HMM. The trained HMM is used to identify and locate similar patterns in the historical data. The price differences between the matched days and the respective next day are calculated. Finally， a weighted average of the price differences of similar patterns is obtained to prepare a forecast for the required next day. Forecasts are obtained for a number of securities that show APHMM is feasible.

Key words：stock market forecasting; hidden Markov model; hidden Markov model optimize; artificial neural network; particle swarm optimization

0 引 言

Hassan 和 Nath[1]在 2005 年提出了一种使用隐马尔科夫模型预测股票价格的方法，在这篇文章中，Hassan 和 Nath 通过实验说明了隐马尔科夫模型的预测性能与 ANN 算法的预测性能相近。由于隐马尔科夫模型的传统训练方法有易于陷入局部极值的缺点，而与其它的智能优化算法相比，粒子群优化算法对于数学模型参数的优化具有简单易用、只有少量参数需要调整、具有较好的全局搜索能力、结果数度好等优点，于是本文提出了一种用于股票价格预测的人工神经网络(ANN)，隐马尔可夫模型 (HMM)，粒子群优化算法 (PSO) 组合模型－APHMM 模型。经实验表明，本文提出的 APHMM 模型具有较好的预测结果。

1 隐马尔科夫模型

隐马尔科夫模型(hidden Markov model，HMM)[2-3]包含一个双重随机过程，一重是隐马尔科夫链(描述随机过程状态的转移)，另一重是与隐马尔科夫链中状态相关的随机观察值输出概率函数(描述状态和观值之间的统计对应关系)。在某一个时刻，这个隐马尔科夫链处于某一状态之中，并且由与这个状态相关的随机观察值输出概率函数生成一个观察值。然后，隐马尔科夫链根据状态转移概率转移到下一个状态。这样，站在观察者的角度，只能看到观察值，而不像马尔科夫链模型中的观察值和状态一一对应。因此，不能直接看到状态，而只能通过一个随机过程感知状态的存在及其特性，所以称为“隐”马尔科夫链模型。

收稿日期：2008-11-14；修订日期：2009-04-21。

作者简介：朱嘉瑜 (1982－)，男，广东茂名人，硕士研究生，研究方向为智能信息处理、智能优化算法等； 叶海燕 (1981－)，女，广东河源人，硕士研究生，研究方向为智能信息分析与处理； 高鹰 (1963－)，男，江苏吴江人，博士后，教授，硕士生导师，研究方向为智能优化算法、盲信号处理等。E-mail：zhujiayu2006@126.com

连续观察值密度 HMM 可以被以下 5 个参数完全描述：

(1)N 是 HMM 中的状态数。

(2)M 是每个状态中混合高斯概率密度函数。

(3)A 是状态转移概率矩阵。$A=(a_{ij})$，a_{ij}为隐马尔科夫链中从当前状态i转移到下一状态j的概率，即

$$a_{ij}=P[q_{t+1j}=j \mid q_t=1] \tag{1}$$

式中：q_t——t时刻所在的状态，且a_{ij}满足约束条件：$a_{ij}\geq 0$，$1\leq i,j\leq N$和$\sum a_{ij}=1$，$1\leq i\leq N$。

(4)B 是观察值输出概率矩阵 $B=\{b_j(o)\}$，$b_j(o)$是状态j的随机观察值输出概率函数。最常见的随机观察值输出概率函数是有限个高斯函数的线性混合

$$b_j(o)=\sum_{k=1}^{M} c_{jk}G(o,\mu_{jk},U_{jk}),1\leq j\leq N \tag{2}$$

式中：o——观察向量，c_{jk}——第j个状态第k个混合高斯函数的权值，G——在第j个状态第k个混合函数中的均值为μ_{jk}和协方差为U_{jk}的高斯分布。混合高斯函数的权值c_{jk}满足约束条件：$c_{jk}\geq 0$，$1\leq j\leq N$，$1\leq k\leq M$和$\sum c_{ij}=1$，$1\leq j\leq N$。

(5)$\prod$是初始状态分布矩阵，$\prod=\{\pi_i\}$，且π_i满足约束条件：$\pi_i=P[q_1=i]$，$1\leq i\leq N$和$\sum \pi_i=1$。

通常为了方便起见，可将 HMM 表示为：$\lambda=(\boldsymbol{A},\boldsymbol{B},\prod)$。

2 粒子群算法

粒子群优化算法[4]是一个基于种群的优化算法，种群称作粒子群，粒子群中的个体被称为粒子。设有N个粒子组成的一个群体，其中第i个粒子表示为一个m维的向量$\boldsymbol{x}_i$ $(i=1,2,\cdots,N)$，第i个粒子的“飞行”速度也是一个m维的向量，记为$\boldsymbol{v}_i$ $(i=1,2,\cdots,N)$。再设$f(x)$为最大化的目标函数，则粒子群优化算法采用下列公式对粒子操作

$$\boldsymbol{v}_i(t+1)=w\boldsymbol{v}_i(t)+c_1r_1(\boldsymbol{p}_i(t)-\boldsymbol{x}_i(t))+c_2r_2(\boldsymbol{p}_g(t)-\boldsymbol{x}_i(t)) \tag{3}$$

$$\boldsymbol{x}_i(t+1)=\boldsymbol{x}_i(t)+\boldsymbol{v}_i(t+1) \tag{4}$$

式中：$\boldsymbol{p}_i(t)$ $(i=1,2,\cdots,N)$——第i个粒子迄今为止搜索到的最优位置，$\boldsymbol{p}_g(t)$——整个粒子群迄今为止搜索到的最优位置，即：$\boldsymbol{p}_g(t)\in\{\boldsymbol{p}_1(t),\boldsymbol{p}_2(t),\cdots,\boldsymbol{p}_N(t)\mid f(\boldsymbol{p}_g(t))=\max\{f(\boldsymbol{p}_1(t)),f(\boldsymbol{p}_2(t)),\cdots,f(\boldsymbol{p}_N(t))\}\}$。$w$为惯性因子，$c_1$和$c_2$称为加速系数；$r_1$和$r_2$是介于$[0,1]$之间的随机数。迭代终止条件根据具体问题一般选为最大迭代次数或(和)粒子群迄今为止搜索到的最优位置满足预定最小适应阈值[5]。

3 ANN, HMM, PSO 组合模型

HMM 模式识别的性能与它的参数的估计有很大的联系，本文首先用 ANN 算法把观察序列(历史的股票价格的每日开盘价、最高价、最低价与收盘价)分离为一组独立的值，并作为 HMM 的输入值，使得训练数据更适于用 HMM 来进行识别。PSO 算法则用于寻找更好的 HMM 参数初始值。在应用上述两个算法之后，得到了分离的训练数据和最优的初始参数，这时便用 Baum-Welch[2]算法来对 HMM 的参数进行训练，从而得到最终的 APHMM 模型。

在 APHMM 模型中，我们通过 Viterbi 算法[2]对今天股票价格行为模式进行识别，在训练数据中找到与今天股票价格行为模式相似的一组数据。在这组数据中，计算每个数据与它后一天的价格差，然后对这组价格差进行加权平均。则今天的股票价格加上这个加权平均价格差便为我们预测的（明天的)股票价格。APHMM 模型的示意图如图 1 所示。

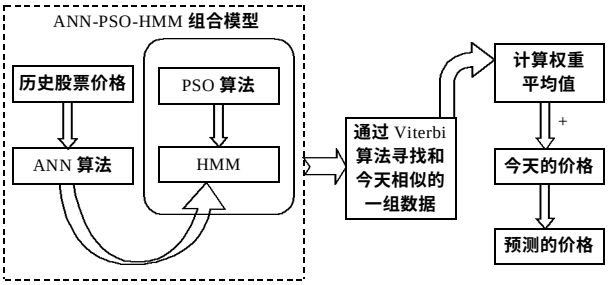


图 1 APHMM 模型

3.1 HMM 的优化

3.1.1 使用 ANN 算法优化 HMM 的输入

HMM 的参数训练与观察序列之间有很强的依赖关系[6]。因此我们用 ANN 将观察序列转换成更适应于 HMM 的形式，令 HMM 的参数训练过程变得更有效率，具体步骤如下：

(1)创建一个 N 个输入与 N 个输出的三层前向型神经网络。

(2)随机初始化 ANN 算法内结点的权重。

(3)将实际观察序列输入 ANN 算法。

(4)得到 ANN 的输出序列，并将这个输出序列作为 HMM 的输入。

在未经 ANN 处理之前，观察序列中股票价格中的开盘价、最高价、最低价、收盘价这 4 组数值之间的相关性是很高的，如图 2(a)所示。经过 ANN 处理后，将观察序列分离为 4 组相互独立的数据，如图 2(b)所示。

3.1.2 使用 PSO 算法优化 HMM 的初始参数

在组合模型中，我们采用 Baum-Welch 算法来对 HMM 的参数进行训练。由 Sankar [7]，Kowng 和 Qianhua [8] 等人提出，Baum-Welch 算法的性能是由与 HMM 的初始值和输入序列决定的。因此，我们引入 PSO 算法的目的就是优化 HMM 的参数初始值，以提高经 Baum-Welch 算法训练后的 HMM 的模式识别性能。

我们用 PSO 算法优化 HMM 中 3 个参数的初始值：初始概率矩阵、状态转移概率矩阵、输出概率矩阵。如果将这 3 个参数作为一个 PSO 算法中的粒子，将使粒子的数据结构变得十分复杂。因此我们一次只优化一个参数：将每一个参数作为一个粒子，分别构造 3 个 PSO 算法来对这 3 个粒子进行优化。而且当一个参数被优化时，另两个参数保持不变。这 3 个参数的具体优化步骤如下：

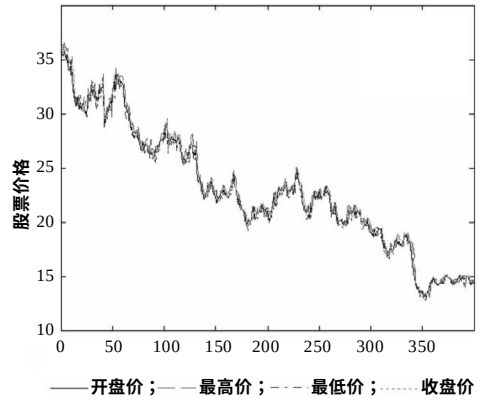
(1)随机选取这些参数的初始值。

(2) 对于初始概率矩阵执行 PSO 而其它两个参数保持第(1)步的值。

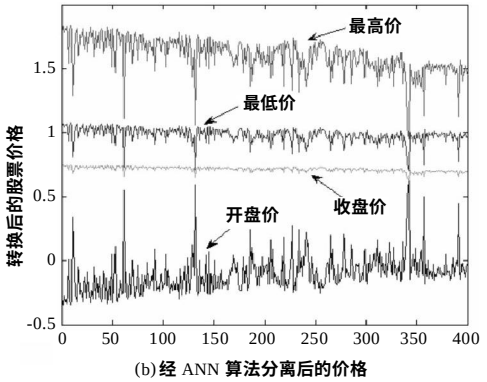
(3)对于状态转移概率矩阵执行 PSO，初始概率矩阵保持第(2)步的值，输出概率矩阵保持第(1)步的值。

(4)对输出概率矩阵执行 PSO，而初始概率矩阵保持第(2)步的值，状态转换概率矩阵保持第(3)步的值。

(5)如果达到终止条件则退出，否则转到第(2)步。



(a) 实际股票价格



(b) 经 ANN 算法分离后的价格

图 2 实际观察序列及其分离

PSO 算法的目标函数是“绝对平均误差”—Mean Absolute Percentage Error(MAPE)。对于优化参数的每一个 PSO 算法具体的执行步骤为：

(1)初始化种群。

(2)进化种群。

(3)用进化的种群作为 HMM 的初始参数并用 Baum-Welch 算法对 HMM 进行训练。

(4)用完成训练的 HMM 对于测试数据集中的每一个数据进行识别(用 Viterbi 算法计算 HMM 对每一数据的输出概率)，然后在训练数据中找出若干相似数据 (即找出在训练数据中输出概率与测试数据输出概率相近的数据)，并用这些相似数据分别计算测试集中的每一个数据预测值。

(5)根据测试数据的预测值计算 PSO 算法的 MAPE。

(6)如果终止条件没达到转到下一步否则退出。

(7)用速度和位置公式对种群进行进化。

(8)转到第(2)步。

3.2 基于加权平均的股票价格预测

在 20 世纪 50 年代学者提出了时间序列的长记忆性问题，但是直到 80 年代长记忆性模型才被用于研究金融市场的时间序列[9]。长期记忆是指每一个观测都带有在它之前所发生的所有事件的“记忆”。而且近期的发生事件的影响比远期的大。

而本文提出的基于加权平均的股票价格预测正是建立在股票价格的时间序列是具有长记忆性这个基础之上的。具体来说就是：假设当前日期为 t，在历史数据中第 m 日的股票价格走势与当前日期 t 的走势相似，则第 t+1 日的股票价格走势就与第 m+1 日的走势相似。而且，若第 m 日与 t 的日期相差越小，则第 t+1 日的走势与 m+1 日走势就越相似，反之，两者差别就越大。

使用加权平均预测股票价格具体方法是，对某一个日期 t，先用 APHMM 模型在历史数据中找出一组数据(历史数据集 Th)。这就意味着，历史数据集 Th 中每个元素的股票价格行为模式与今天 t 股票价格行为模式是相似的。接着我们计算历史数据集中每个元素的收盘价与它下一交易日的收盘价的价格差，对于这组价格差使用加权平均(与日期 t 越接近的价格差权重越大)得到 t 日与 t+1 日收市价格差。具体公式如下

$$wd = \frac{\sum_{Th} w_i * diff_i}{\sum_{Th} w_i} \tag{5}$$

式中：wd ——第 t 日与 t+1 日的收市价格差；i ——历史数据集 Th 中的第 i 个数据；w_i ——第 i 个数据的价格差对 wd 权重，$w_i = \frac{\sum_{Th}(D_{i-t}) - D_{i-t}}{\sum_{Th}(\sum_{Th}(D_{i-t}) - D_{i-t})}$，$D_{i-t}$ ——第 i 个数据的日期与第 t 日日期相差的天数，也就是若 D_{i-t} 越小，第 i 个数据对价格差 wd 影响的越大。$diff_i$ ——第 i 日与 i+1 日的收盘价格差。

在得到第 t 日与 t+1 日的收市价格差 wd 之后，第 t+1 日的收市价格为：$P = P_t + wd$，其中 P_t 为当日 t 的收盘价。

4 实验仿真及结果

4.1 实验数据

为了测试本文提出的 APHMM 模型的性能，在本节我们用 APHMM 模型对美股中 Apple Computer Inc.,International Business Machines Corporation(IBM)和 Dell Inc.这 3 个股票的价格进行测试。这 3 个股票的数据来源于 www.finace.yahoo.com。

表 1 显示了实验数据的详细情况。训练数据中的每个交易日的开盘价、最高价、最低价和收盘价组作为 APHMM 模型的训练数据集，而对测试数据进行的预测则作为检验 APHMM 模型性能的指标。

表 1 实验数据情况

股票名称	训练数据		测试数据	
	开始日期	截止日期	开始日期	截止日期
Apple Computer Inc.	2003-2-10	2004-9-10	2004-9-13	2005-1-21
IBM Corporation	2003-2-10	2004-9-10	2004-9-13	2005-1-21
Dell Inc.	2003-2-10	2004-9-10	2004-9-13	2005-1-21

4.2 实验各算法的参数设置

ANN 算法为 4 个输入与 4 个输出的三层前向型神经网络。它包含一个输入层，一个隐含层，一个输出层。每层均有 4 个结点，隐含层的激励函数为 Sigmodi 函数 tanh(·)，而输出层激励函数为线性函数。经过实验发现，表 2 的两组权重对于训练数据的分离效果较好。

因为实验数据每个向量有 4 个元素，所以 HMM 的状态数

表 2 ANN 各层的权重

W1(隐含层)				W2(输出层)			
0.483 3	0.674 9	−0.442 9	−0.132 5	0.601 6	0.758 1	0.197 3	0.865 4
−0.153 3	0.340 0	−0.541 6	−0.477 2	−0.959 9	−0.157 3	0.971 4	−0.368 6
−0.681 2	0.948 5	0.175 1	0.859 6	−0.634 8	−0.738 0	0.335 2	−0.190 3
0.897 9	−0.127 1	0.717 4	−0.203 7	−0.751 1	0.602 0	0.289 1	0.069 7

设置为 4。由于训练数据集是随着时间变化的连续序列，所以采用“左－右模型”[2]的 HMM。随机观察值输出概率函数采用连续的概率密度函数。HMM 的参数训练采用多观察序列 Baum-Welch 算法，进化代数为 10 次，HMM 识别算法采用 Viterbi 算法。HMM 的 3 个参数 $(A, B, \prod)$的初始化均采用随机数。在训练数据中寻找与测试数据最相似的天数为 15 天。

PSO 算法采用的粒子子维数为 4，种群个数为 8，最大进化代数为 25，c_1,c_2=2。

4.3 性能指标——绝对平均误差（MAPE）

本文提出的组合模型的性能是通过绝对平均误差(MAPE)来衡量的。MAPE 的计算方法如下

$$\text{MAPE} = \frac{\sum_{i=1}^{r}\left(\frac{\text{abs}(y_i - p_i)}{y_i}\right)}{r} \times 100\% \quad (6)$$

式中：r ——测试数据中用于测试的数据个数；y_i ——第 i 日的实际收盘价；p_i ——第 i 日的预测收盘价。

4.4 实验结果

本文提出的 APHMM 模型对测试集的预测结果对表 3 所示。由表 3 可以看出，本文提出的 APHMM 模型对测试集的预测结果的绝对平均误差是比较小的，说明了 APHMM 模型有效性。

表 3 APHMM 模型对测试数据的实验结果

股票名称	测试数据中 91 个数据的绝对平均误差(MAPE)
Apple Computer Inc.	1.749 8
IBM Corporation	0.645 31
Dell Inc.	0.759 28

5 结束语

本文提出了一种用于股票价格预测的人工神经网络(ANN)，隐马尔可夫模型(HMM)，粒子群优化算法(PSO)组合模型－APHMM 模型。在 APHMM 模型中，首先确定 ANN 算法的各层权重，用 ANN 算法将股票的每日开盘价、最高价、最低价与收盘价转换为相互独立的量并作为 HMM 的输入。然后，利用 PSO 算法对 HMM 的参数初始值进行优化，并用 Baum-Welch 算法对 PSO 算法优化后的初始参数进行训练。经过训练后的 HMM 在历史数据中找出一组与今天股票的上述 4 个指标价模式最相似数据，加权平均计算每个数据与它后一天的收盘价格差。则今天的股票收盘价加上这个加权平均价格差便为预测的股票收盘价。经过实验仿真表明，APHMM 模型对选定的测试数据有良好的预测效果。同时也说明了 PSO 算法对于优化 HMM 参数的初始值也是行之有效的。

参考文献：

[1] Hassan M R, Nath B. Stock market forecasting using hidden Makov model: a new approach[C]. Proceedings of 5th International Conference on Intelligent Systems Design and Applications, 2005:192-196.

[2] Lawrence R Rabiner.A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proc IEEE, 1989,77(2):257-286.

[3] 于江德,樊孝忠,尹继豪.隐马尔可夫模型在自然语言处理中的应用[J].计算机工程与设计,2007,28(22):5514-5516.

[4] 章万国,周驰,高海兵,等.粒子群优化算法[J].计算机应用研究,2003,20(12):7-12.

[5] 高鹰.一种自适应扩展粒子群优化算法[J].计算机工程与应用,2006,42(15):16-19.

[6] 齐爱学, 王洪刚. 基于 HMM/ANN 混合模型的带噪语音识别[J].杭州电子科技大学学报,2007,3:17-21.

[7] Sankar A. Experiments with Gaussian merging-splitting algorithm for HMM training for speech recognition [CP/OL]. Proceedings of DARPA Speech Recognition Workshop.www.nist.gov/speech/publications/darpa98/html/am10/am10.htm.

[8] Kwong S,Qianhua He.The use of adaptive frame for speech recognition [J]. EURASIP Journal on Applied Signal Processing, 2001,2:82-88.

[9] 王春峰,张庆翠.中国股票市场收益的长期记忆性研究[J].系统工程,2003,21(1):22-29.

(上接第 4837 页)

[3] Larson S,Snow C,Shirts M,et al.Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology [J]. Computational Genomics,2002:634-654.

[4] Anderson D P.BOINC: A system for public-resource computing and storage[C].Proc of the Fifth International Workshop of Grid Computing,2004:4-10.

[5] Agrawal A, Casanova H. Clustering hosts in p2p and global computing platforms [C]. Proceeding of the IEEE/ACM CCGrid,2003.

[6] Sarmenta L F G.Sabotage-tolerance mechanisms for volunteer computing systems [J]. Future Generation Computer Systems, 2002,18:561-572.

[7] Ratnasamy S, Handley M, Karp R, et al. Topologically-aware overlay construction and server selection [C]. Proc of INFOCOM,2002.

[8] Fedak G,Germain C,Neri V,et al.XtremWeb:A generic global computing system[C].Proceedings of the 1st IEEE/ ACM International Symposium on Cluster Computing and the Grid,2001:582-587.

[9] Hewgill G, RC5 and Java Toys 2009[S]. http://www.hewgill.com/rc5/index.html.

[10] Shudo K,Tanaka Y,Sekiguchi S.P3:P2P-based middleware enabling transfer and aggregation of computational resources [C]. Proc of the IEEE International Symposium on Cluster Computing and the Grid,2005:259-266.