# 深度计算的同辈群体股市态势预测算法*

姚宏亮 洪竞帆 王 浩

(合肥工业大学 计算机与信息学院 合肥 230009)

**摘 要** 针对同辈群体的同辈群体算法(PG)的不足,提出深度计算的同辈群体生成算法. 首先计算目标股票和候选股票之间的波段相似性,然后通过对亲密度、相关性和活跃度的深度计算,生成目标股票的同辈群体,并证明深度计算生成的同辈群体质量优于 PG 算法. 针对 PG 算法不具有预测功能,通过结合自回归(AR)模型与同辈群体算法,提出基于同辈群体的自回归股价态势预测算法(DPG-AR). DPG-AR 利用深度计算生成同辈群体,实现同辈群体权重的动态更新,并利用 AR 模型预测目标股票态势. 上海证券综合指数及对应个股的对比实验证明 DPG-AR 的优越性.

**关键词** 同辈群体算法, 深度计算, 自回归模型, 态势预测

**中图法分类号** TP 181 **DOI** 10.16451/j.cnki.issn1003-6059.201601007

**引用格式** 姚宏亮,洪竞帆,王 浩. 深度计算的同辈群体股市态势预测算法. 模式识别与人工智能, 2016, 29(1): 54-62.

# Peer Group Stock Market Trend Prediction Algorithm Based on Deep Computing

YAO Hongliang, HONG Jingfan, WANG Hao

(*School of Computer and Information*, *Hefei University of Technology*, *Hefei* 230009)

## ABSTRACT

Aiming at the deficiency of peer group (PG) algorithm, a peer group generation algorithm based on depth computing is proposed. Firstly, the band similarity between target stock and candidate stock is calculated. Then, grounded on depth calculation of intimacy, correlation and liveness, peer group of the target stock is generated, and it is proved that the quality of peer group generated by depth calculation is superior to that of PG algorithm. Since PG algorithm is unable to predict, autoregressive stock market trend prediction algorithm based on peer group (DPG-AR) is proposed by combining peer group algorithm and autoregression model. Peer group is generated by deep computing. Thus, the weights of peer group members are updated. The target stock trend is prodicted by autoregression model. The effectiveness of DPG-AR is verified in the experiment on Shanghai composite index and the corresponding stock.

**Key Words** Peer Group Algorithm, Deep Computing, Autoregression Model, Trend Prediction

* 国家重点基础研究进展计划项目(No. 2013CB329604)、国家自然科学基金项目(No. 61175051,61175033)资助
Supported by the National Basic Research Program of China (No. 2013CB329604) and National Natural Science Foundation of China (No. 61175051,61175033)
收稿日期:2015-05-13;修回日期:2015-10-13;录用日期:2015-12-02
Manuscript received May 13, 2015; revised October 13, 2015; accepted December 2, 2015

**Citation**　YAO H L, HONG J F, WANG H. Peer Group Stock Market Trend Prediction Algorithm Based on Deep Computing. Pattern Recognition and Artificial Intelligence, 2016, 29(1): 54-62.

股票市场变化莫测[1],股票的价格态势受到众多因素的影响,如经济因素、政治因素、企业因素、心理因素和技术指标等[2],这些因素的变化使股票不断出现新波动模式[3],导致基于有监督学习的股票态势预测方法难以适应波动模式的不断变化.

很多研究者从监督学习的角度研究股票价格波动的多变性问题. 处理股票市场突变性的主要算法如下. Kirkos 等[4]讨论基于人工神经网络和决策树的方法,利用股票市场的历史数据训练生成股票市场状态描述模型,判断股票市场状态的异常行为. Wheeler 等[5]提出基于案例学习的股票市场异常行为监测方法. Chang[6]提出人工神经网络(Artificial Neural Networks, ANN)模型、决策树模型和 ANN 与决策树混合模型的股票价格态势预测方法,但只能获得局部最优解,且当股票价格出现新模式之后,效果快速退化. Yi 等[7]依据股票之间的相关性构建贝叶斯网络,当某种股票发生异常时,贝叶斯网络自动调整参数,从而适应多变的市场环境. 但股票与股票之间的相关性由很多因素决定[8],股票市场具有多变性,当某种属性分布发生变化,贝叶斯网络的表达能力受到影响.

为了更好解决多变性和突变性难题,研究者提出结合机器学习理论与统计学理论预测股票市场. Chen 等[9]提出基于粒子计算的混合模糊的时间序列,Tian 等[10]提出基于信息分类的时间序列模型,Sarvan 等[11]提出模拟传统经济的缩放分析. 上述算法都是基于有监督学习的思想,通过统计过去状态以分析未来态势,并做出判断,故对新模式不敏感.

为了适应股票价格态势中不断产生新模式的情况,无监督学习的方法受到关注. Ferdousi 等[12]提出基于同辈群体(Peer Group)的无监督学习思想,但未考虑目标股票的同辈群体亲密度是不断变化的. 进而,Kim 等[13]提出的同辈群体算法(Peer Group Algorithm, PG),依据同辈群体之间的接近度,使用窗口动态更新同辈群体权重,建立窗口跟踪式预测模型. 然而,PG 算法在同辈群体选择上只考虑股票之间的欧氏距离,欧氏距离较小的股票不能保证与目标股票态势相同,这种情况极大弱化同辈群体的参照作用,且 PG 算法只有跟踪功能,不具有预测功能.

针对上述 PG 算法的不足,基于深度计算思想,本文提出同辈群体分析方法与自回归(Auto Regression, AR)模型结合的股票价格态势预测算法. 通过聚类分析[14]确定目标股票的同辈群体,计算同辈群体与目标股票的亲密度,并更新每个窗口的同辈群体权值,更好描述目标股票的实际态势.

深度计算基本过程[15]如下. 首先通过波段涨跌幅指标筛选与目标股票走势相同的股票. 在保证待选股票与目标股票走势相同的基础之上,通过相关系数指标和股票亲密度指标进一步筛选,使得到的股票是与目标股票走势相同、相关性强、亲密度大的股票. 为了保证目标股票同辈群体波动幅度平缓,引入活跃度指标进一步筛选候选股票. 通过层层筛选,最终得到的股票是与目标股票走势相同、相关性强、亲密度大且波段幅度小的股票. 与原有的同辈群体选取算法相比,该算法有效避免选择亲密度大但走势不同的股票,防止因同辈群体选取不当导致参照作用被极大弱化的情况出现.

通过结合 AR 模型与同辈群体算法,本文提出基于同辈群体的自回归股价态势预测算法(Autoregressive Stock Market Trend Prediction Algorithm Based on Peer Group, DPG-AR). DPG-AR 利用深度学习生成的同辈群体,动态更新同辈群体权重,然后利用 AR 模型预测目标股票态势,并优化模型参数. 在实验分析中,以中小板指、上海证券综合指数及各自包含的个股为例,验证 DPG-AR 的实用性和有效性.

# 1　相关知识

## 1.1　同辈群体

同辈群体是指在给定时间段中,将与目标股票波动特征相似的多个股票称为目标股票的同辈群体. 通过对目标股票与其同辈群体股票价格波动的跟踪对比,发现目标股票的异常波动.

以股票的收盘价为研究对象,对每只股票的收盘价分别进行归一化,归一化后的收盘价全部映射至(0,1)内. 使用 min-max 的归一化方法:

$$x^* = \frac{x - min}{max - min},$$

其中,$max$ 为指定时间段内的股票收盘价最大值,$min$ 为指定时间段内的股票收盘价最小值,$x$ 为股票收盘价,$x^*$ 为归一化后的收盘价,将归一化后的收盘价简称为收盘价.

设有 $m(m>2)$ 只股票,每只股票都有 $D$ 个交易日,$x_i$ 表示第 $i$ 只股票收盘价的时间序列. 将每只股票划分为 $N$ 个不重叠的窗口,计算每个窗口的平均长度为 $D/N$,第 $i$ 只股票的某个窗口 $n$ 的平均收盘价 $y_{in}$ 可表示为

$$y_{in}=\frac{1}{D/N}\sum_{j=1}^{D/N}x_{i(n-1)D/N+j}.$$

根据上式,股票 $i$ 和股票 $j$ 对应的窗口向量分别为 $\boldsymbol{y}_i$ 和 $\boldsymbol{y}_j$,两个窗口向量之间的欧氏距离可表示为

$$d_{ii'}=\sqrt{(\boldsymbol{y}_i-\boldsymbol{y}_{i'})(\boldsymbol{y}_i-\boldsymbol{y}_{i'})^{\mathrm{T}}}.$$

根据到目标股票的欧氏距离的大小,从小到大排序 $m-1$ 只股票,得到目标股票 $i$ 的同辈群体 $p_i$ 可表示为

$$p_i=\{y_{i\pi(1)},y_{i\pi(2)},\cdots,y_{i\pi(m-1)}\}.$$

目标股票 $i$ 的同辈群体确定之后,同辈群体在时间窗口 $n$ 内,平均收盘价 $p_{in}$ 可表示为

$$p_{in}=\frac{1}{m-1}\sum_{j=1}^{m-1}y_{i\pi(j)n},$$

其中,$y_{i\pi(j)n}$ 是目标股票 $i$ 的第 $j$ 个同辈群体在时间窗口 $n$ 内的时间序列值.

设 $prox_{i\pi(j)}$ 表示两个股票在同个时间窗口内的亲密度:

$$prox_{i\pi(j)}=\exp(-\gamma d_{ii'}),\tag{1}$$

其中,$\gamma>0$ 是调整两个股票欧氏距离和接近度之间关系的系数.

在目标股票的同辈群体成员中,亲密度越高的股票,分配权重越大. 使用加权平均的方法重新计算 $p_{in}$,改进后的 $p_{in}$ 的计算公式如下:

$$p_{in}=\sum_{j=1}^{k}w_{i\pi(j)}y_{i\pi(j)n},\tag{2}$$

其中,$w_{i\pi(j)}$ 为目标股票的第 $j$ 个同辈对象的权重,利用式(2)计算目标股票的第 $j$ 个同辈股票和目标股票之间的亲密度 $prox_{i\pi(j)}$. 目标股票的第 $j$ 个同辈对象权重 $w_{i\pi(j)}$ 为

$$w_{i\pi(j)}=\frac{prox_{i\pi(j)}}{\sum_{j=1}^{k}prox_{i\pi(j)}}.\tag{3}$$

### 1.2 自回归模型

设 $\{x_t\}$ 为零均值的平稳过程,关于 $x_t$ 的合适模型为

$$x_t=\varphi_1x_{t-1}+\varphi_2x_{t-2}+\cdots+\varphi_px_{t-p}+a_t,\tag{4}$$

其中,$a_t$ 为白噪声序列,$x_t$ 遵循一个 $p$ 阶自回归随机过程,即 $AR(p)$ 模型.

使用多维最小二乘法进行无偏参数估计,模型可表示为如下方程组:$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\varphi}+\boldsymbol{a}$,其中

$$\boldsymbol{Y}=[x_{n+1},x_{n+2},\cdots,x_N]^{\mathrm{T}},$$

$$\boldsymbol{\varphi}=[\varphi_1,\varphi_2,\cdots,\varphi_n]^{\mathrm{T}},$$

$$\boldsymbol{a}=[a_{n+1},a_{n+2},\cdots,a_N]^{\mathrm{T}},$$

$$\boldsymbol{X}=\begin{bmatrix}x_n & x_{n-1} & \cdots & x_1\\ x_{n+1} & x_n & \cdots & x_2\\ \vdots & \vdots & & \vdots\\ x_{N+n-1} & x_{N+n-2} & \cdots & x_N\end{bmatrix}.$$

根据多维最小二乘法,

$$\boldsymbol{\varphi}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.\tag{5}$$

## 2 基于深度计算的同辈群体生成算法

### 2.1 相关定理及证明

Kim 等[13]提出的 PG 算法是以亲密度为参考指标,计算待选股票与目标股票的亲密度,按照亲密度降序排列,选取排序靠前的股票,将这些股票作为目标股票的同辈群体. 这种构建方法忽视一种很重要情况,即亲密度大的股票,态势与目标股票不同. 这种情况将在定理 1 中予以证明.

**定理 1** 设 2 只待选股票 $A$、$B$ 与目标股票 $C$ 的亲密度分别为 $prox_{AC}$ 和 $prox_{BC}$,态势分别为 $T_A$、$T_B$、$T_C$,则存在

$$prox_{AC}>prox_{BC},$$

但

$$T_A\neq T_C,\ T_B=T_C.$$

**证明** 由图 1 所示,从左向右顶点依次是 $a$、$b$、$c$、$d$、$e$,实线波段代表目标股票的态势图,点线波段表示态势不同但欧氏距离小的股票,虚线波段表示态势相同但欧氏距离大的股票.

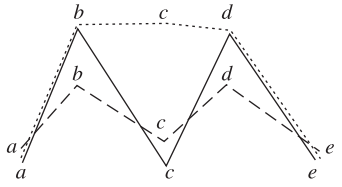



图 1 定理 1 证明示意图

Fig. 1 Sketch map of theorem 1

设点线波段和实线波段每个顶点间的距离分别为:$D_{aa}, D_{bb}, D_{cc}, D_{dd}, D_{ee}$.

设虚线波段和实线波段每个顶点间的距离分别为:$E_{aa}, E_{bb}, E_{cc}, E_{dd}, E_{ee}$.

首先由点线波段,有

$$D_{aa}=0,\ D_{bb}=0,\ D_{dd}=0,\ D_{ee}=0,$$

所以

$$Dis_1^2 = D_{cc}D_{cc},$$

再由虚线波段,有

$$Dis_2^2 = E_{aa}E_{aa} + E_{bb}E_{bb} + E_{cc}E_{cc} + E_{dd}E_{dd} + E_{ee}E_{ee},$$

令

$$E_{aa}^2 + E_{bb}^2 + E_{cc}^2 + E_{dd}^2 + E_{ee}^2 > D_{cc}^2,$$

则 $Dis_2 > Dis_1$.

如果同辈群体里包含与目标股票态势不同的股票,在跟踪目标股票的态势时会存在偏差,影响预测效果(将在定理 2 中予以证明). 设股票相邻两个交易日态势增幅为 $\Delta P$,目标股票 $C$ 有两组同辈群体 $P_1$、$P_2$,$P_1$ 中股票态势与目标股票相同,即 $T_{P_1}=T_C$,$P_2$中部分股票态势与目标股票不同,即 $T_{P_2}\neq T_C$.

**定理2**　设同辈群体 $P_1$模拟目标股票 $C$ 态势的增幅$\Delta P_{P_1C}$,同辈群体 $P_2$模拟目标股票 $C$ 态势的增幅 $\Delta P_{P_2C}$,目标股票实际态势增幅为 $\Delta P_C$,则

$$\Delta P_C - \Delta P_{P_2C} < \Delta P_C - \Delta P_{P_1C}.$$

**证明**　设 $w_1, w_2, \cdots, w_n$ 表示同辈群体 $P_1$ 中 $n$ 只股票的权值,$X_{t_1}, X_{t_2}, \cdots, X_{t_n}$ 表示第 $t$ 天 $n$ 只股票的收盘价,$X_{(t-1)_1}, X_{(t-1)_2}, \cdots, X_{(t-1)_n}$ 表示第 $t-1$ 天 $n$ 只股票的收盘价;$w_1', w_2', \cdots, w_n'$表示同辈群体 $P_2$ 中 $n$ 只股票的权值,$X'_{t_1}, X'_{t_2}, \cdots, X'_{t_n}$表示第 $t$ 天 $n$ 只股票的收盘价,$X'_{(t-1)_1}, X'_{(t-1)_2}, \cdots, X'_{(t-1)_n}$表示第 $t-1$ 天 $n$ 只股票的收盘价.

1) 同辈群体 $P_1$ 中存在与目标股票态势不同的股票,第 $t$ 天的同辈群体可表示为

$$P_t = w_1X_{t_1} + w_2X_{t_2} + \cdots + w_nX_{t_n},$$

第 $t-1$ 天的同辈群体可表示为

$$P_{t-1} = w_1X_{(t-1)_1} + w_2X_{(t-1)_2} + \cdots + w_nX_{(t-1)_n},$$

$$\begin{aligned}\Delta P_1 &= (w_1X_{t_1} + w_2X_{t_2} + \cdots + w_nX_{t_n}) - \\ &\quad (w_1X_{(t-1)_1} + w_2X_{(t-1)_2} + \cdots + w_nX_{(t-1)_n}) \\ &= w_1(X_{t_1} - X_{(t-1)_1}) + w_2(X_{t_2} - X_{(t-1)_2}) + \cdots + \\ &\quad w_n(X_{t_n} - X_{(t-1)_n}) \\ &= w_1\Delta X_1 + w_2\Delta X_2 + \cdots + w_n\Delta X_n.\end{aligned}$$

由于同辈群体 $P_1$ 中 $n$ 只股票至少存在 1 只股票的态势与目标股票不同,可设第 2 只股票的态势与目标股票态势不同. 假设目标股票态势上升,则 $\Delta X_2 < 0$,下面分 2 种情况讨论.

(1) 情况 1.

$$|w_2\Delta X_2| > w_1\Delta X_1 + w_3\Delta X_3 + \cdots + w_n\Delta X_n,$$

$$\Delta P_1 = w_1\Delta X_1 + w_2\Delta X_2 + \cdots + w_n\Delta X_n.$$

设

$$Z = -\Delta X_2,$$

则

$$\begin{aligned}\Delta P_1 &= w_1\Delta X_1 + w_2\Delta X_2 + \cdots + w_n\Delta X_n \\ &= w_1\Delta X_1 - w_2(-\Delta X_2) + \cdots + w_n\Delta X_n \\ &= w_1\Delta X_1 - w_2Z + \cdots + w_n\Delta X_n \\ &= w_1\Delta X_1 + w_3\Delta X_3 + \cdots + w_n\Delta X_n - w_2Z \\ &< 0.\end{aligned}$$

这种情况下,当目标股票态势上升时,同辈群体描述值的态势下降.

(2) 情况 2.

$$|w_2\Delta X_2| < w_1\Delta X_1 + w_3\Delta X_3 + \cdots + w_n\Delta X_n,$$

则有

$$Z = -\Delta X_2,$$

$$\begin{aligned}\Delta P_1 &= w_1\Delta X_1 + w_2\Delta X_2 + \cdots + w_n\Delta X_n \\ &= w_1\Delta X_1 - w_2(-\Delta X_2) + \cdots + w_n\Delta X_n \\ &= w_1\Delta X_1 - w_2Z + \cdots + w_n\Delta X_n \\ &= w_1\Delta X_1 + w_3\Delta X_3 + \cdots + w_n\Delta X_n - w_2Z.\end{aligned}$$

这种情况下,原先上升态势被削弱.

2) 同辈群体 $P_2$ 中所有股票与目标股票态势相同. 同理得

$$\Delta P_2 = w_1'\Delta X_1' + w_2'\Delta X_2' + \cdots + w_n'\Delta X_n'.$$

设目标股票态势增幅为 $\Delta P$, 令

$$0 < \Delta P_2 < \Delta P,$$

分如下 2 种情况考虑.

(1) 情况 1.

$$|w_2\Delta X_2| > w_1\Delta X_1 + w_3\Delta X_3 + \cdots + w_n\Delta X_n,$$

此时

$$\Delta P_1 < 0,\ \Delta P_2 > 0,$$

则

$$\Delta P - \Delta P_1 > \Delta P - \Delta P_2.$$

(2) 情况 2.

$$|w_2\Delta X_2| < w_1\Delta X_1 + w_3\Delta X_3 + \cdots + w_n\Delta X_n,$$

此时

$$0 < \Delta P_1 < \Delta P_2 < \Delta P,$$

则

$$\Delta P - \Delta P_1 > \Delta P - \Delta P_2.$$

证毕.

为了避免态势不同导致同辈群体跟踪目标股票

态势效果不佳,首先保证待选股票与目标股票态势相同,在态势相同的基础上,选取亲密度大、形态相似和活跃度低的股票.

### 2.2 相关定义

**定义 1** 波段涨跌幅(Band Rang, BR) 在某个股票价格态势波段中,存在若干个波峰和波谷,将这些波峰和波谷对应的时间作为时间节点,将波段划分为若干个子波段,子波段包含的交易日的平均涨跌幅记为波段涨跌幅,波段涨跌幅 $BR$ 可表示为

$$BR = \frac{cp_l - cp_s}{cp_s},$$

其中,$cp_l$ 表示子波段最后一天的收盘价,$cp_s$ 表示子波段第一天的收盘价.

**定义 2** 股票活跃度(Liveness, LN) 将待选股票和目标股票的同一波段按照相同的时间节点进行划分,得到 $n$ 个子波段. 计算待选股票 $s$ 与目标股票 $g$ 对应子波段的欧氏距离,子波段欧氏距离的集合定义为

$$dis = \{dis_{sg_1}, dis_{sg_2}, \cdots, dis_{sg_n}\},$$

对比各个子波段的欧氏距离,得出最大值

$$dis_{sg_i} = \max(dis)$$

和最小值

$$dis_{sg_j} = \min(dis),$$

则活跃度为

$$LN = dis_{sg_i} - dis_{sg_j}. \tag{6}$$

**定义 3** 相关系数 设股票 $i(i = 1,2,\cdots,N)$ 在第 $t$ 个分时收盘价为 $Y_i(t)$,对应的目标股票 $s$ 在时间 $t$ 的收盘指数为 $P_s(t)$,符号 $S_i(t)$ 表示第 $i$ 支股票的收盘价,符号 $Z_s(t)$ 表示目标股票 $s$ 的收盘价.

股票 $i$ 和股票 $j$ 之间引入时间相关系数 $R_{ij}$,包含目标股票对其潜在的影响,定义如下:

$$R_{ij} = \frac{\langle S_i S_j\rangle - \langle S_i\rangle\langle S_j\rangle}{\sqrt{(\langle S_i^2\rangle - \langle S_i\rangle^2)(\langle S_j^2\rangle - \langle S_j\rangle^2)}}.$$

### 2.3 深度计算的同辈群体生成算法

以 Kim 等[13] 提出的 PG 算法为基础,提出基于深度计算的同辈群体生成算法(Peer Group Generation Algorithm Based on Depth Computing, DPG). 算法基本思想如下:按照目标股票研究波段的波峰和波谷对应的时间节点,对待选股票进行波段划分,划分之后的波段包含多个子波段.

计算每个子波段的波段涨跌幅,与目标股票波段态势完全一致的股票保留. 根据股票与目标股票之间的欧氏距离,将股票按照欧氏距离升序排列,取前 $k_1$ 只股票. 根据股票与目标股票之间的相关系数,将股票按照相关系数降序排列,取前 $k_2$ 只股票. 根据股票的活跃度,将股票按照活跃度升序排列,取前 $k_3$ 只股票,将这 $k_3$ 只股票组成的股票集合作为目标股票的同辈群体.

首先定义 3 个函数.

1) $SignS(X_1, X_2)$: 若 $G_1, G_2$ 符号相同,$SignS(G_1, G_2) = 1$,否则 $SignS(G_1, G_2) = 0$.

2) $Asce(G_1, k, G_2)$:将集合 $G_1$ 按升序排列,取前 $k$ 个加入集合 $G_2$.

3) $Desc(G_1, k, G_2)$:将集合 $G_1$ 按降序序排列,取前 $k$ 个加入集合 $G_2$.

下面给出 DPG 的具体步骤.

**算法 1** DPG

**输入** 数据集 *Dategather*

**输出** 目标股票的同辈群体

step 1 对数据集 *Dategather* 中的个股波段 $B$ 按照目标股票的波峰和波谷对应的时间作为时间节点进行划分,划分之后的波段包含多个子波段:

$$DB = \{B_1, B_2, \cdots, B_n\},$$

$$TDB = \{TB_1, TB_2, \cdots, TB_n\},$$

其中,$DB$ 表示待选股票子波段的集合,$TDB$ 表示目标股票子波段的集合.

step 2 根据定理 1 和定理 2,由于亲密度大的股票走势可能与目标股票不同,导致模拟目标股票走势出现偏差,所以首先筛选走势与目标股票相同的股票,计算 $B_t(t = 1,2,\cdots,n)$ 的波段涨跌幅和 $TB_t(t = 1,2,\cdots,n)$ 的波段涨跌幅:

$$BR_{1t} = \frac{CP_{B_tL} - CP_{B_tS}}{CP_{B_tS}},$$

$$BR_{2t} = \frac{CP_{TB_tL} - CP_{TB_tS}}{CP_{TB_tS}},$$

其中,$BR_{1t}$ 表示待选股票在第 $t$ 个波段的波段涨跌幅,$BR_{2t}$ 表示目标股票在第 $t$ 个波段的波段涨跌幅;若 $SignS(BR_{1t}, BR_{2t}) = 1$,对应个股加入数据集 *Dategather*_1.

step 3 初始化数据集 *Dategather*_2,由于走势相同的股票同样存在与目标股票偏离程度较大的情况,例如,股票都是上涨走势,但上涨幅度不同,所以引入欧氏距离进行控制,计算数据集 *Dategather*_1 中的个股与目标股票之间的欧氏距离. 所有个股与目标股票之间的欧氏距离组成集合:

$$Dis = \{d_1, d_2, \cdots, d_n\},$$

将 $C_1 = Asce(Dis, k, G_2)$ 对应的个股加入数据集 *Dategather*_2.

step 4　初始化数据集 *Dategather_3*,相关系数表示待选股票与目标股票之间的相关程度,由于波段涨跌幅只能保证待选股票与目标股票之间的走势相同,无法表示相关性,所以引入相关系数,从整体上筛选相关性大的股票,计算数据集 *Dategather_2* 中的个股与目标股票之间的相关系数. 所有个股与目标股票之间的相关系数组成集合:

$$R = \{R_{1g}, R_{2g}, \cdots, R_{ng}\},$$

将 $C_2 = Desc(R, k, G_2)$ 对应的个股加入数据集 *Dategather_3*.

step 5　初始化数据集 *PeerGroup*,由于在保证走势相同、亲密度大的情况下,如果待选股票的波动十分强烈,会对同辈群体选择造成误差,所以提出股票活跃度的概念,控制待选股票的波动幅度与目标股票一致,根据式(6) 计算数据集 *Dategather_3* 中的个股的活跃度. 所有个股的活跃度组成集合:

$$LN = \{L_1, L_2, \cdots, L_n\},$$

将 $C_3 = Asce(LN, k, G_2)$ 对应的个股加入数据集 *PeerGroup*.

算法中 step 1 按照目标股票的波峰和波谷对应的时间节点对待选股票波段进行划分,得到多个子波段,分别对比每个子波段与目标股票对应子波段的态势是否相同. 根据定理 1 和定理 2,与目标股票态势不同的股票构成的同辈群体对于描述目标股票的态势存在偏差,在此算法保证选取的股票态势全部与目标股票态势相同,并且 step 5 ~ step 7 保证选取的股票是亲密度大、形态相似和活跃度低的股票.

# 3　基于同辈群体的自回归模型价格态势预测算法

文献[13] 提出的 PG 算法只能对股票的异常进行监测,无预测功能,本文的 DPG-AR 在深度计算的同辈群体生成算法基础之上,加入 $AR(p)$ 模型,实现预测. 同时在 PG 算法中加入 $AR(p)$ 模型(PG-AR),与 DPG-AR 进行对比.

## 3.1　算法思想

1) 使用 DPG 生成目标股票的同辈群体.

2) 对生成的同辈群体赋权值,使用式(2) 得到每个时间窗口的同辈群体的行为描述值.

3) 同辈群体的行为描述值构成一个时间序列,将这个时间序列代入 AR 模型中,构建价格态势预测模型,进而实现预测功能.

## 3.2　算法过程

1) 通过 2.3 节得到的数据集 *PeerGroup*,使用式(3),计算目标股票的同辈群体的权值序列.

2) 由于在每个时间点都需计算目标股票与待选股票之间的亲密度,重新构建同辈群体,花费大量的计算时间,因此为了使同辈群体能更紧密地跟踪目标对象,提出同辈群体权值更新算法,得到的权值代入式(2),得到每个时间窗口的同辈群体的收盘价

$$P = \{p_1, p_2, \cdots p_n\},$$

带入 $AR(p)$ 模型,使用多维最小二乘法,代入式(5),计算 $AR(p)$ 模型参数. 具体算法过程如下.

初始化:$w_{i\pi(j),1}$

$$w_{i\pi(j),t+1} = (1-\lambda) w_{i\pi(j),t} + \lambda \frac{prox_{i\pi(j),t}}{\sum_{j=1}^{k} prox_{i\pi(j),t}},$$

由上式和式(2),得到同辈群体的收盘价序列

$$P = \{p_1, p_2, \cdots, p_n\},$$

将序列 $P$ 代入式(4),得

$$\begin{aligned}
&\phi_0 p_1 + \phi_1 p_2 + \cdots + \phi_p p_p = p_{p+1},\\
&\phi_0 p_2 + \phi_1 p_3 + \cdots + \phi_p p_{p+1} = p_{p+2},\\
&\vdots\\
&\phi_0 p_{n-p} + \phi_1 p_{n-p+1} + \cdots + \phi_p p_{n-1} = p_n.
\end{aligned}$$

通过式(5) 得到 $\boldsymbol{\phi}$ 矩阵如下:

$$\begin{pmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_p\end{pmatrix} = \left(\begin{pmatrix} p_1 & p_2 & \cdots & p_p\\ p_2 & p_3 & \cdots & p_{p+1}\\ \vdots & \vdots & & \vdots\\ p_{n-p} & p_{n-p+1} & \cdots & p_{n-1}\end{pmatrix}^{\mathrm{T}} \cdot \begin{pmatrix} p_1 & p_2 & \cdots & p_p\\ p_2 & p_3 & \cdots & p_{p+1}\\ \vdots & \vdots & & \vdots\\ p_{n-p} & p_{n-p+1} & \cdots & p_{n-1}\end{pmatrix}\right)^{-1} \cdot \begin{pmatrix} p_1 & p_2 & \cdots & p_p\\ p_2 & p_3 & \cdots & p_{p+1}\\ \vdots & \vdots & & \vdots\\ p_{n-p} & p_{n-p+1} & \cdots & p_{n-1}\end{pmatrix}^{\mathrm{T}} \cdot \begin{pmatrix} p_{p+1}\\ p_{p+2}\\ \vdots\\ p_n\end{pmatrix}.$$

而 $prox_{i\pi(j),t}$ 由式(1) 得到,$\lambda \in (0,1)$.

## 3.3　算法性能优化

### 3.3.1　根据波段性质构建自回归模型

在这里讨论的波段包括上涨波段和下降波段,对于一只股票的态势来说,一个大的波段总会包括多个上涨子波段和下降子波段. 由于 $AR(p)$ 模型对于波段转折点的检测较困难,所以根据不同的波段性质,构建两类预测模型:第 1 类预测模型对应于上

涨波段,第 2 类模型预测对应于下跌波段. 在如下情况中,采用第 2 类模型:

1) 波段处于下行趋势;

2)$K$ 线脱离上轨,寻求中轨或下轨的支撑;

3)$K$ 线刚刚触底反弹.

### 3.3.2 $AR(2)$ 模型参数调节

定义偏离值 $\sigma$,根据先验知识,设定偏离值 $\sigma$ 的大小.

参数调节过程如下.

step 1 计算每只股票通过 $AR(p)$ 模型得到的预测值和实际值的差 $PA$:

$$PA = |CP_{\text{实际值}} - CP_{\text{测试值}}|.$$

step 2 计算所有股票 $PA$ 值的和,并且计算其平均数 $PA_{ave}$.

step 3 如果

$PA_{ave} > \sigma$,

$\Phi_0 = \Phi_0 - 0.0001, \cdots, \Phi_p = \Phi_p - 0.0001$,

代入 $AR(p)$ 模型重新计算,跳至 step 1,如果 $PA_{ave} \leqslant \sigma$,跳至 step 4.

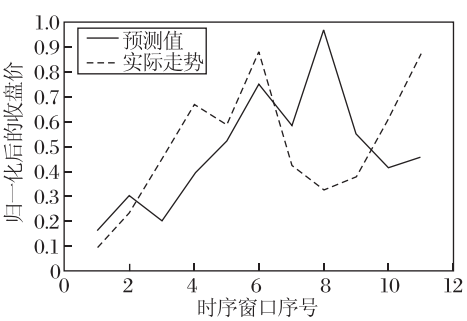
step 4 模型参数调节完毕.

## 4 实验及结果分析

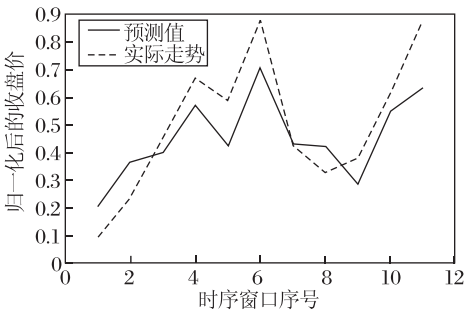
### 4.1 预测结果对比

实验数据来源于国元领航软件,抽取股票中任意一个波段,本次抽取 2014 年 11 月 14 日至 2015 年 1 月 16 日的中小板指综合指数和中小盘对应的所有个股数据、2014 年 11 月 24 日至 2015 年 1 月 26 日的上海证券综合指数和上海证券 $A$ 股所有个股数据.

实验 1 中将 2014 年 11 月 14 日至 2015 年 1 月 16 日的交易日划分成 11 个窗口序列,针对该时间波段中中小板指综合指数和中小盘对应的所有个股进行实验,结果见图 2.




(a) PG-AR



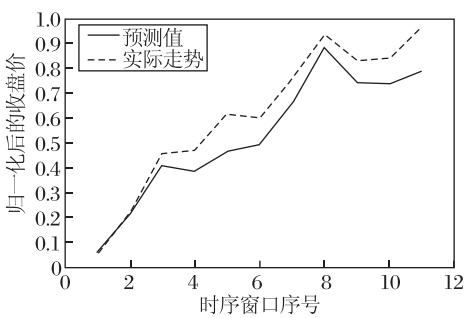

(b) DPG-AR

图 2 实验 1 中 2 种算法预测态势对比

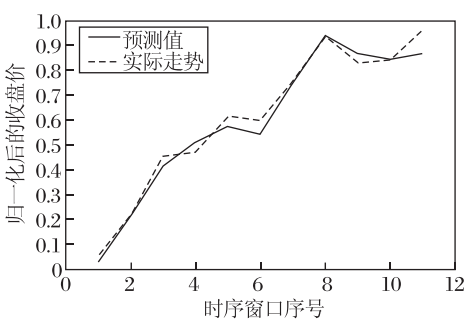
Fig. 2 Prediction trend comparison of 2 algorithms in experiment 1

由图 2(a) 可知,在第 2 个窗口、第 4 个窗口、第 7 个窗口 ~ 第 11 个窗口中,PG-AR 都与实际态势不符. 由(b) 可知,DPG-AR 仅在第 9 个窗口中态势与实际态势具有一定差异. 由此可见,DPG-AR 对于目标股票的态势描述更准确.

实验 2 选取 2014 年 11 月 24 日至 2015 年 1 月 26 日的上海证券综合指数和上海证券 A 股里的所有个股数据,划分成 11 个时间窗口,结果如图 3 所示.




(a) PG-AR




(b) DPG-AR

图 3 实验 2 中 2 种算法预测态势对比

Fig. 3 Prediction trend comparison of 2 algorithms in experiment 2

从图3中可见,DPG-AR对于目标股票的价格态势模拟与实际态势接近,明显优于PG-AR.

### 4.2　误差分析

均方根误差:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}(y - y')^2},$$

平均绝对误差:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y - y'|,$$

其中,$N$为预测集的样本个数,$y$为真实值,$y'$为预测值.$RMSE$、$MAE$用来表示预测值偏离实际值的大小,值越小,偏离度越小,表明预测结果的精确度越高.

实验1将中小板指综合指数作为目标股票,中小板指对应个股作为待选股票,实验2将上海证券综合指数作为目标股票,上海证券综合指数对应个股作为待选股票.通过PG-AR和DPG-AR在2个实验上进行测试,结果见表1.

**表1　2个实验中的算法精度对比**

Table 1　Precision comparison in 2 experiments

| | 算法 | 均方根误差 | 平均绝对误差 |
|---|---|---|---|
| 实验1 | PG-AR | 0.3511 | 0.2818 |
| | DPG-AR | 0.1231 | 0.1066 |
| 实验2 | PG-AR | 0.1134 | 0.0937 |
| | DPG-AR | 0.0441 | 0.0358 |

由表1可见,DPG-AR的均方根误差和平均绝对误差小于PG-AR的均方根误差和平均绝对误差,算法性能得到提高.同时,DPG-AR已应用于宏大数据网(www.ihdsj.com)的智能个股诊断、选股和个股预警系统中.

## 5　结束语

采用深度计算的思想,结合原有的同辈群体模型,对候选股票进行层层筛选,将上一层得到的股票作为下一层的输入,继续筛选.这种方法优化目标股票的同辈群体构建,优化后的同辈群体算法筛选的股票全部与目标股票态势相同,进而构建AR模型进行后市预测,有效提高预测精度.以中小板指综合指数、上海证券综合指数及各自包含的个股为例,对本文算法和原始算法进行对比分析,验证算法的实用性和有效性.如何优化预测模型,融入股票市场的技术指标,提高预测精度,都将是下一步重点研究的方向.

## 参　考　文　献


[1] 张秀云,郭 树.股票基本面分析在实战中的应用.中国证券期货,2011(5):28-29.
(ZHANG X Y, GUO S. The Application of Stock Fundamental Analysis in the Real. Securities and Futures of China, 2011(5): 28-29.)

[2] LI Q, WANG T J, LI P. The Effect of News and Public Mood on Stock Movements. Information Sciences, 2014, 278: 826-840.

[3] MARSZALEK A, BURCZYNSKI T. Modeling and Forecasting Financial Time Series with Ordered Fuzzy Candlesticks. Information Sciences, 2014, 273: 144-155.

[4] KIRKOS E, SPATHIS C, MANOLOPOULOS Y. Data Mining Techniques for the Detection of Fraudulent Financial Statements. Expert Systems with Applications, 2007, 32(4): 995-1003.

[5] WHEELER R, AITKEN S. Multiple Algorithms for Fraud Detection. Knowledge-Based Systems, 2000, 13(2/3): 93-99.

[6] CHANG T S. A Comparative Study of Artificial Neural Networks, and Decision Trees for Digital Game Content Stocks Price Prediction. Expert Systems with Applications, 2011, 38(12): 14846-14851.

[7] Zuo Y, KITA E. Stock Price Forecast Using Bayesian Network. Expert Systems with Applications, 2012, 39(8): 6729-6737.

[8] 王双成,杜瑞杰,刘 颖.连续属性完全贝叶斯分类器的学习与优化.计算机学报,2012,35(10):2129-2138.
(WANG S C, DU R J, LIU Y. The Learning and Optimization of Full Bayes Classifiers with Continuous Attributes. Chinese Journal of Computers, 2012, 35(10): 2129-2138.)

[9] CHEN M Y, CHEN B T. A Hybrid Fuzzy Time Series Model Based on Granular Computing for Stock Price Forecasting. Information Sciences, 2015, 294: 227-241.

[10] TIAN Q, SHANG P J. Financial Time Series Analysis Based on Information Categorization Method. Physica A: Statistical Mechanics and Its Applications, 2014, 416: 183-191.

[11] SARVAN D, STRATIMIROVIĆ D, BLESIĆ S. Scaling Analysis of Time Series of Daily Prices from Stock Markets of Transitional Economies in the Western Balkans. The European Physical Journal B, 2014. DOI: 10.1140/epjb/e2014-50655-5.

[12] FERDOUSI Z, MAEDA A. Unsupervised Outlier Detection in Time Series Data // Proc of the 22nd International Conference on Data Engineering Workshops. Atlanta, USA, 2006: 51-56.

[13] KIM Y, SOHN S Y. Stock Fraud Detection Using Peer Group Analysis. Expert Systems with Applications, 2012, 39(10): 8986-8992.

[14] WANG W N, LIU X D. Fuzzy Forecasting Based on Automatic Clustering and Axiomatic Fuzzy Set Classification. Information Sciences, 2015, 294: 78-94.

[15] 余 滨,李绍滋,徐素霞,等.深度学习:开启大数据时代的钥匙.工程研究——跨学科视野中的工程,2014,6(3):233-243.
(YU B, LI S Z, XU S X, *et al*. Deep Learning: A Key of Stepping into the Era of Big Data. Journal of Engineering Studies, 2014, 6(3): 233-243.)

## 作者简介

**姚宏亮**,男,1972 年生,博士,副教授,主要研究方向为机器学习、数据挖掘. E-mail:dmicyhl@163.com.

(**YAO Hongliang**, born in 1972, Ph. D., associate professor. His research interests include machine learning and data mining.)

**洪竞帆**(通讯作者),男,1993 年生,硕士研究生,主要研究方向为人工智能、数据挖掘. E-mail:1049284985@qq.com.

(**HONG Jingfan**(Corresponding author), born in 1993, master student. His research interests include artificial intelligence and data mining.)

**王 浩**,男,1962 年生,博士,教授,主要研究方向为人工智能、数据挖掘. E-mail:jsjxwangh@hfut.edu.cn.

(**WANG Hao**, born in 1962, Ph. D., professor. His research interests include artificial intelligence and data mining.)