一种基于流特征模式的股市跟踪预测算法

姚宏亮 杜明超 李俊照 王 浩
(合肥工业大学计算机与信息学院 合肥 230009)

摘 要 由于股市波动的突发性、多变性,且时序数据呈非正态分布,传统的时序预测模型难以有效预测股市。提出了一种基于流特征模式的股市跟踪预测算法(SFM-PG),该算法根据股票之间的相关性构建贝叶斯网络,选取目标股票的马尔科夫毯作为其同辈群体,然后基于同辈群体之间的接近度,给出一种窗口跟踪式预测模型,其通过对同辈群体权重的动态更新进行跟踪式预测,以减少股票数据分布非正态性对预测的影响;进而,使用滑动窗口提取时序数据中的特征并形成流特征,通过与模式知识库的匹配提取流特征模式,并利用与流特征模式对应的知识调整预测结果,以减少由于突变所引入的预测误差。最后,在上证股票板块网络上的实验结果显示了算法的实用性和有效性。

关键词 流特征,流特征模式,同辈群体分析,股市预测
中图法分类号 TP181 **文献标识码** A

Stock Market Tracking Prediction Algorithm Based on Stream Feature Model

YAO Hong-liang DU Ming-chao LI Jun-zhao WANG Hao
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract Because stock market volatility is of mutability and variability, and the distribution of the time series data does not follow the normal distribution, the traditional time series forecast algorithms are difficult to accurate prediction. The stock market tracking prediction algorithm based on Stream Feature Model was proposed (SFM-PG). It is builds the Bayesian networks based on correlation between stocks, selects the Markov Blanket of the target stock as its peer group, and gives a windows tracking prediction model based on the proximity between peer group, through dynamicly updating the weight of peer group to tracking prediction, effective avoids the influence of the non-normal distribution of time series data on prediction. And then, using the sliding window to extract the feature of the time series data to formation stream feature, and extracting the stream feature model by matching with knowledge base of base, using the knowledge of stream feature model to adjust the predicted results, in order to reduce the prediction error introduced by mutability. Finally, the practicability and effectiveness are showed in the experiment on the network of plate of the Shanghai stock.

Keywords Stream feature, Stream feature model, Peer group analysis, Stock market price forecasting

1 引言

股票价格预测对战略投资组合以及经济发展具有重要的作用,但是,非量化因素会严重影响股票价格的走势,如政治事件、经济情况、投资者的情绪、欺诈等,诸多非量化因素导致股票价格波动的突发性和多变性,股票价格时序数据呈非正态分布[1],使股票价格的精确预测是极具挑战的任务。

在早期的研究过程中,一些传统的时间序列模型被用于解决股票价格预测的问题,如 AR 模型、ARMA 模型、ARIMA 模型等,但是这些模型是在假设股票时序数据是正态分布和平稳的情况下进行分析[2]的。因此,这些模型不能有效地对股票数据进行预测[3]。R. F. Engle 和 T. Bollerslev 对传统时间序列模型进行改进,分别提出了 ARCH 模型[4]和 GARCH 模型[5],用于非正态分布时序数据的预测,但是,ARCH 和 GARCH 模型简单,且单一模型预测精度较低[6]。

从数据挖掘的角度,为了提高预测精度,J. V. Hansen 将 ANNs 模型与传统时序模型相结合来预测国家经济,能够避免数据非常态分布对模型的影响,但是,ANNs 模型在预测股票方面具有模型控制参数多、过度学习风险等缺点,容易陷入局部最小陷阱[7]。另一方面,SVR 因为其结构风险最小化原则[8],能够提供全局最优和更好的预测精度[9]。ANNs 模型和 SVR 模型可以处理非正态分布时序数据,但是,这两种模型都没有很好地利用先验知识,尤其是走势出现的一些具有特殊意义的价格波动。

股票价格中包含了很多有价值的信息,有效地分解和聚合信息能够提高总体预测精度[10]。Stephen J. Taylor 对股票

到稿日期:2013-05-21 返修日期:2013-07-11 本文受国家自然科学基金(61175051,61070131,61175033)资助。
姚宏亮(1972—),博士,副教授,主要研究方向为机器学习与数据挖掘,E-mail:dmicyhl@163.com;杜明超(1987—),男,硕士生,主要研究方向为人工智能与数据挖掘;李俊照(1975—),男,博士生,讲师,主要研究方向为机器学习与数据挖掘;王 浩(1962—),男,博士,教授,主要研究方向为人工智能与数据挖掘。

价格波动信息进行分析，提取波动数据中的特征，结合ARCH模型进行预测[11]，与传统单一ARCH模型相比，提高了预测精度。Ling-Jing Kao使用非线性独立成分分析和SVR组合模型预测股票价格[7]，非线性独立成分分析用于特征提取，发现历史数据中的隐藏信息。实证分析表明，这种组合模型由于传统的SVR方法，提高了预测的准确度，同样验证了有效利用历史时序数据中的信息能够提高预测的精度。

因此，针对现有的股票价格预测方法，为了解决股价波动突发性和多变性带来的问题，提出了一种基于流特征模式的股市跟踪预测算法（Stream Feature Model _ Peer Group，SFM-PG），SFM-PG算法分成两个部分。首先，借鉴组别分析[12]的思想，依据股票之间的相关性，构建所有股票之间的贝叶斯网络，从网络中选取目标股票的马尔科夫毯作为其同辈群体，依据同辈群体之间的接近度，使用窗口对同辈群体的权重进行动态更新，从而建立窗口跟踪式预测模型。其次，为了进一步利用股价波动的先验知识，SFM-PG算法提取历史数据中隐藏的知识，构建模式知识库，在预测时，同时使用滑动窗口提取出特征形成流特征，通过与已构建的模式知识库进行匹配提取流特征模式，并更新模式知识库，然后利用与提取的流特征模式对应的知识调整预测结果，减少由于突变所引入的预测误差，提高了预测精度。通过对SFM-PG算法的分析和抽象，文中给出了序列接近度、流特征和流特征模式的定义。

在股票市场中，股票与股票之间是相互关联和相互影响的，一种股票的变化会造成与其相关联的股票发生变化，因此，基于同辈群体跟踪和条件流特征模式研究股票之间的走势相关性具有现实意义。在实证分析中，以股票行业板块网络为例，应用SFM-PG算法对实用性和有效性进行了验证。

2 贝叶斯网络与马尔科夫毯

2.1 贝叶斯网络

贝叶斯网络（Bayes Network，BN）是变量之间概率依赖关系的一种图形表示方法，网络中的结点对应一个变量，结点之间有边说明对应的变量之间有依赖关系，这种依赖的关系和程度可以使用概率参数来表示[13]。贝叶斯网络可以用一个二元函数 $B=(G,\theta)$ 表示，其中 $G=\langle X,E\rangle$ 表示一个有向无环图，$X=\{X_1,X_2,\cdots,X_n\}$ 表示网络中的随机变量集，网络中共有 n 个随机变量，E 是有向边集，E 中的每条边代表网络中结点之间具有直接的依赖关系。参数集 $\theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$ 表示网络的条件概率分布集，在变量取离散值的情况下，θ 表示网络的概率分布表，$\theta_i=P(X_i|Pa(X_i))$ 表示结点 X_i 的联合概率分布，其中 $Pa(X_i)$ 表示结点 X_i 的父结点集。BN结构由如下的一组条件独立性假设决定：

$$P(X_i|X_1,\cdots,X_{i-1})=P(X_i|Pa(X_i)) \tag{1}$$

其中，$i=1,2,\cdots,n$，由式(1)可以看出，对于BN中的每个结点 X_i，在给定 $Pa(X_i)$ 的情况下，X_i 条件独立于其任何非子孙结点。因此，BN中变量集 X 的联合概率分布可表示为：

$$P(X)=\prod_{i=1}^{n}P(X_i|Pa(X_i)) \tag{2}$$

其中，$i=1,2,\cdots,n$，式(2)表明变量集 X 的联合概率分布可表示成各个局部模型的因式形式。图1是一个标准BN网络的实例。

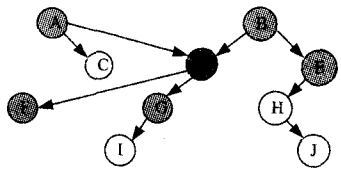


图1 贝叶斯网络

2.2 马尔科夫毯

如2.1节中式(2)所示，对于BN网络 G 和联合概率分布 $P(X)$，在给定网络中任意一个结点 X_i 的父结点 $Pa(X_i)$ 的情况下，该结点与它的非子孙结点独立，我们称 $\langle G,P\rangle$ 满足因果马尔科夫条件。同样，对于一个BN，在给定路径 R 的情况下，路径 R 中包含的结点 X_i 有两条指向它的边，那么结点 X_i 在 R 中可以称为碰撞结点，在给定 X_i 取值的情况下，X_i 的两个父结点条件依赖。假如 X_i 的两个父结点是 T 和 Z，那么 $Ind(T,Z|X_i)$ 表示变量 T 和 Z 在给定变量 X_i 时条件独立；$Dep(T,Z|X_i)$ 表示变量 T 和 Z 在给定 X_i 时条件依赖[14]。

BN中结点 X_i 的马尔科夫毯（Markov Blanket，MB）用 $MB(X_i)$ 表示，$MB(X_i)$ 是 X_i 的最小特征子集。在给定 X_i 和变量子集 S 且 $X_i\notin S$ 的情况下，X_i 的马尔科夫毯 $MB(X_i)$ 存在条件独立 $Ind(S,X_i|MB(X_i))$。图1所示BN中的阴影部分是变量 D 的马尔科夫毯。

在一个忠实的BN中，变量 X_i 的马尔科夫毯 $MB(X_i)$ 是存在且唯一的，$MB(X_i)$ 中的结点是由变量 T 的父结点、子结点和子结点的父结点组成[15]。

由马尔科夫毯的定义和条件独立可以看出，在网络中，一个变量的马尔科夫毯能够屏蔽网络中其他变量对该变量的影响，可用来进行预测、分类和因果发现等。

3 基于窗口跟踪式预测模型

对于一个特定的目标股票，将与目标股票价格波动和特征相似的一系列对象称为目标股票的同辈群体（Peer Group，PG），又称对等组。通过构建同辈群体的跟踪模型可以模拟目标股票价格波动的趋势，进而，可以利用这种相似性预测目标股票的走势。SFM-PG算法分成两部分，一部分是利用窗口跟踪式模型预测股价，另一部分是利用流特征模式动态调整跟踪模型。本部分主要介绍SFM-PG算法的窗口跟踪式预测模型（Windows Tracking _ Peer Group，WT-PG），WT-PG是基于贝叶斯网络、马尔科夫毯特征学习和同辈群体的思想提出的，其使用马尔科夫毯局部结构学习的思想在股票网络中获取目标股票的同辈群体，使用窗口动态更新权重的方法构建同辈群体的跟踪模型来预测目标股票的走势。

3.1 序列窗口和接近度

随着时间的推移，和目标股票具有相似行为的其他股票被称为目标股票的同辈群体。假设同辈群体中有 k 个股票对象，每个对象对应一个时间序列 i，对于任何一个时间序列 i 都有 T 个数值对应于每个时间点 t，其中 $t=1,2,\cdots,T$。因此，时间序列 i 可以表示成长度为 T 的向量 X_i，时间序列 i 在第 t 个时间点的数值可以用 $x_{i,t}$ 表示。

对于所要研究的股票对象 X_i，其对应的日收盘指数是一种流数据，用 $SD=\{\cdots,x_{i,m},x_{i,m+1},\cdots,x_{i,m+n},\cdots\}$ 表示，SD 是一组无穷不断生成的一位数组，在时间点 t 处的时序数据用 $x_{i,t}$ 表示。对于 SD，有如下相关定义：

定义 1(序列窗口,Series Window,SW) 在股票时间序列流数据 SD 中,设 $SW=\{x_{i,m},x_{i,m+1},\cdots,x_{i,m+t}\}$ 是由 T 个连续的不同时序数据组成的集合,其中,m 是窗口开始的时间戳,$m+t$ 是窗口结束的时间戳,$SW\subseteq SD$,窗口的大小为 T。

定义 2(序列的接近度,Proximity of Sequence,PS) 假设两个股票对象在同一个时间段内的非线性时序数据的欧式距离为 ED_{ij},ED_{ij} 的计算如下:

$$ED_{ij}=\sqrt{(X_i-X_j)(X_i-X_j)^{\mathrm{T}}} \tag{3}$$

其中,$X_i=\{x_{i,1},x_{i,2},\cdots x_{i,T}\}$,对于给定的目标股票时间序列 X_i,同辈群体中其他股票的时间序列 X_j 与目标股票时间序列 X_i 之间的欧式距离越大,其相似性越低。

PS_{ij} 表示两个股票在同一个时间段内的接近度,则 PS_{ij} 的计算如下:

$$PS_{i,j}=\exp(-\gamma ED_{ij}) \tag{4}$$

其中,$\gamma>0$,γ 是一个调整两个股票欧氏距离和接近度之间关系的系数,ED_{ij} 是对象 X_i 和 X_j 之间的欧氏距离。

3.2 同辈群体跟踪模式

目标股票的同辈群体是基于股票时间序列之间的相似性构建的,为衡量时间序列的相似性,通过计算时间序列 X_i 和 X_j 的欧氏距离来度量两个向量之间的差异性,欧式距离通过定义 2 中的式(3)计算,按照同辈群体股票与目标股票欧氏距离从小到大的顺序,X_i 的同辈群体 $PG(i)$ 记为:

$$PG(i)=\{X_{i,P(1)},X_{i,P(2)},\cdots,X_{i,P(k)}\} \tag{5}$$

式中,$X_{i,P(j)}$ 是目标股票 X_i 的第 j 个同辈群体成员。在 SFM-PG 算法中,目标股票 X_i 的同辈群体可表示为行业板块网络中目标股票的马尔科夫毯。在一个 NB 中,目标变量 X_i 的马尔科夫毯 $MB(X_i)$ 与 X_i 之间的联系最紧密,$MB(X_i)$ 中变量的股价走势行为会直接影响到目标变量 X_i,并且 $MB(X_i)$ 可以屏蔽网络中其他股票对 X_i 的影响,用于特征选择。因此,可以将 $MB(X_i)$ 作为 X_i 的同辈群体 $PG(i)$,式(5)可表达如下:

$$PG(i)=MB(X_i)=\{X_{i,P(1)},X_{i,P(2)},\cdots,X_{i,P(k)}\} \tag{6}$$

其中,$i=1,2,\cdots,k$,k 是 $MB(X_i)$ 中变量的数量。

目标股票 X_i 的同辈群体确定之后,可以用 $PH(t)$ 来表示同辈群体在时刻 t 的行为,$PH(t)$ 的计算如下:

$$PH(t)=\frac{1}{k}\left(\sum_{j=1}^{k}x_{i,p(j),t}\right) \tag{7}$$

其中,$x_{i,p(j),t}$ 是目标股票时间序列 X_i 的第 j 个同辈在时刻 t 的时间序列值,$1/k$ 可以看成 $PG(i)$ 中为每个同辈群体对象赋予的权重 $w_{i,p(j)}$,此时 $w_{i,p(j)}=1/k$,在一般同辈群体分析中,建立目标股票同辈群体的跟踪模式时,同辈群体股票使用相同的权重值。

3.3 同辈群体权重更新

在建立同辈群体 $PG(i)$ 的跟踪模式时,为同辈群体成员分配相同的权重存在一定的缺陷。同辈群体中,有些股票和目标股票之间的行为相似性高,优于对等组中其他的对象。因此,我们使用加权平均的方法重新计算 $PH(t)$,同辈群体中每个股票分配不同的权重值,权重最大的股票与目标股票之间的相似性最高。在对跟踪模式中引入加权平均法之后,改进的 $PH(t)$ 的计算如下:

$$PH(t)=\sum_{j=1}^{k}w_{i,p(j)}x_{i,p(j),n} \tag{8}$$

式中,$n=1,2,\cdots,T$,$w_{i,p(j)}$ 是目标股票的第 j 个同辈对象的权重,同辈群体中每个股票通过它们和目标股票之间的接近度来确定各自的权重。通过定义 2 中的式(4)计算目标股票的第 j 个同辈股票和目标股票之间的接近度为 $PS_{i,p(j)}$。接近度的计算和两只股票之间的欧氏距离有直接的关系,图 2 显示了欧氏距离和接近度之间的关系。

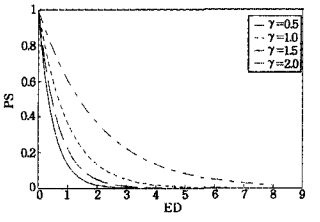


图 2 不同的 γ 下欧氏距离和接近度之间的关系

从图 2 可以看出 γ 越大时,欧氏距离 $d_{i,p(j)}$ 和接近度 $PS_{i,p(j)}$ 之间的灵敏度越大。基于上面定义的接近度的计算方法,目标股票的第 j 个同辈对象权重 $w_{i,p(j)}$ 的计算方法如下:

$$w_{i,p(j)}=\frac{PS_{i,p(j)}}{\sum_{j=1}^{k}PS_{i,p(j)}} \tag{9}$$

从式(6)可以看出,同辈群体中股票和目标股票之间的接近度越大,其权重也就越大,并且目标股票的同辈群体的权重之和为 1。通过引入加权平均法重新计算 $PH(t)$,使同辈群体中的成员会更加紧密地跟踪自己的目标。

3.4 构建跟踪模式描述

在 WT-PG 中,主要通过参数 sw 和 γ 对其进行调整,并建立最佳窗口跟踪式模型。构建 WT-PG 过程描述如下:

1. for all items t∈Datasets
2. Create a BN for Datasets;
3. for target X_i,obtain the MB(X_i) and PG(i) on BN
4. set sw and γ for X_i;
5. for(j=1;j<=k;j++){
6. $ED_{ij}=\sqrt{(X_i-X_{i,P(j)})(X_i-X_{i,P(j)})^{\mathrm{T}}}$;
7. $PS_{i,j}=\exp(-\gamma ED_{ij})$;}
8. for(j=1;j<=k;j++){
9. $w_{i,p(j)}=PS_{i,p(j)}/\sum_{j=1}^{k}PS_{i,p(j)}$;}
10. $PH(t)=\sum_{j=1}^{k}w_{i,p(j)}x_{i,p(j),n}$;
11. return PH(t);

4 SFM-PG 算法

4.1 相关定义

定义 3(流特征,Stream Feature,SF) 在流数据 SD 中,利用窗口捕捉数据特征,当一个特征在窗口出现后,新的窗口将以该特征数据为起点,去捕捉下一个特征。每一个窗口中的特征用 $SF_{id}(number,tr_1,tr_2)$ 表示,其中 id 表示流特征的标识,$id=1,2,\cdots,n$;$number$ 表示该特征第几次出现,$number=1,2,\cdots,k$;tr_1 表示出现该特征后第一天的涨跌幅;tr_2 表示出现该特征后第二天的涨跌幅。

定义 4(流特征模式,Stream Feature Model,SFM) 对于给定区间的流数据 SD,流特征序列为 $SF=\{sf_1,sf_2,\cdots,sf_N\}$,其中,当 $i\neq j$ 时,sf_i 和 sf_j 可以相同;给定特征模式提取策略 σ,流特征模式集合为 $SFM_{SF}=f_\sigma(sf_1,sf_2,\cdots,sf_N)$,其中流特征模式集 $SFM_{SF}=\{sfm_1,sfm_2,\cdots,sfm_M\}$ 中的任意一个流特征模式 sfm_i 是 SF 中特征的一个序列组合。

4.2 构建流特征模式

在股票价格时序数据的K线技术分析中，走势中出现的平台、顶部、底部、巨量长阳、巨量长阴等都将成为流特征，对于股票价格时序数据，依据标准SF_Standard提取流特征，标准SF_Standard描述如下：

1)在股票价格走势中，K线特征无大起大落，趋势为倾斜直线，涨跌幅在－2%和2%之间，成为正常走势，用流特征 SF_1 表示；

2)连续3日及3日以上涨跌幅在－1%和1%之间，起始前后一般均为阳线或阴线，整体趋势为平行直线，称为平台，用流特征 SF_2 表示；

3)在下降通道中，出现一个跳空或中长实体阴线，随后形成一个类似平台的特征，以中长实体阳线结束，称为底部，用流特征 SF_3 表示；

4)在上升通道中，出现一个跳空或中长实体阳线，随后形成一个类似平台的特征，以中长实体阴线结束，称为顶部，用流特征 SF_4 表示；

5)在K线走势中，单日涨幅超过2%，或者出现连续两日的涨幅和超过3%，称为长实体阳线，用流特征 SF_5 表示；

6)在K线走势中，单日跌幅超过－2%，或者出现连续两日的跌幅和超过－3%，称为长实体阴线，用流特征 SF_6 表示；

在SFM-PG算法中，构建特征模式的描述如下：

1)确定流特征提取标准SF_Standard；

2)给定已确定时间段的数据集Datasets；

3)在Datasets上依据SF_Standard提取流特征 SF；

4)依据 SF，构建和更新流特征模式集合库 SFM_{SF}；

5)在 SFM_{SF} 中，依据时间流和提取策略 σ，提取流特征模式 sfm_i。

以2012年11月26日至2013年02月21日的上证综合指数为例，该段数据流提取的流特征模式 $SFM_{SF}=\{SF_2, SF_3, SF_5, SF_2, SF_5, SF_2, SF_5, SF_1(SF_5), SF_5, SF_2, SF_1, SF_4, SF_6\}$，其中 $SF_1(SF_5)$ 表示在正常走势中出现一个长实体阳线特征。

4.3 约束更新

在股票市场，投资者的行为和意图通过表现的数据而体现出来，在分析股票价格数序数据时，随着时间的推移，得到新的数据，序列样本不断增加，原始数据中隐藏的信息则越丰富。因此，有效地挖掘先验数据中的隐藏知识对股票价格预测具有重要作用。在SFM-PG算法中，引入了时序窗口(Time Series Window，TSW)，TSW 中包含 Tsw 个时间序列数据，记为：

$$TSW=\{X_{i,1}, X_{i,2}, \cdots, X_{i,Tsw}\}$$

以 TSW 为单位构建目标股票的 SFM_{tsw}，并维护流特征库 SFB_{tsw}，在 SFM_{tsw} 的条件下更新WT-PG模型，获得SFM-PG模型。

此时，SFM-PG算法中欧氏距离 $ED_{i,p(j)}$ 定义如下：

$$ED_{i,p(j)}=\sqrt{(X_i-X_{i,p(j)})(X_i-X_{i,p(j)})^{\mathrm{T}}}$$

其中，$X_i=\{X_{i,1}, X_{i,2}, \cdots, X_{i,Tsw}\}$，$X_{i,p(j)}=\{X_{i,p(j),1}, X_{i,p(j),2}, \cdots, X_{i,p(j),Tsw}\}$，$ED_{i,p(j)}$ 依据时序窗口 TSW 的大小 Tsw 计算而来。随着时序的向前推移，TSW 的数据进行动态的变化，以保证数据的最新，假设在 Tsw 时刻，TSW 的数据记为 $TSW=\{X_{i,1}, X_{i,2}, \cdots, X_{i,Tsw}\}$，那么在 $Tsw+1$ 时刻，TSW 的数据更新为 $TSW=\{X_{i,2}, X_{i,3}, \cdots, X_{i,Tsw+1}\}$。$TSW$ 更新之后，$ED_{i,p(j)}$ 随之更新，然后根据式(4)和式(9)对权重 $w_{i,p(j)}$ 进行更新。

使用 SFM_{tsw} 可以使同辈群体更紧密地跟踪其目标对象，在SFM-PG中，引入参数 λ，在 SFM_{tsw} 下更新权重 $w_{i,p(j),Tsw+1}$，$w_{i,p(j),Tsw+1}$ 定义如下：

$$w_{i,p(j),Tsw+1}=(1-\lambda)w_{i,p(j),Tsw}+\lambda\frac{PS_{i,p(j),Tsw+1}}{\sum_{j=1}^{k}PS_{i,p(j),Tsw+1}} \quad (10)$$

其中，$PS_{i,p(j),Tsw+1}$ 和 $w_{i,p(j),Tsw}$ 依据式(4)和式(9)计算而来，$\lambda\in[0,1]$。

4.4 算法描述和参数优选

在SFM-PG算法中有3个参数，分别是 Tsw、γ 和 λ，通过分析确定3个参数的值，建立最佳的SFM-PG模型进行股票价格的预测。

SFM-PG算法描述：

1. for all items t∈Datasets
2. Create a BN for Datasets;
3. Datasets is divided into Train－Data and Test－Data;
4. for target X_i and Train－Data, obtain the MB(X_i) and PG(i) on BN
5. set Tsw and γ for X_i;
6. for(j=1;j<=k;j++){
7. $ED_{ij}=\sqrt{(X_i-X_{i,P(j)})(X_i-X_{i,P(j)})^{\mathrm{T}}}$;
8. $PS_{i,j}=\exp(-\gamma ED_{ij})$;}
9. for(j=1;j<=k;j++){
10. $w_{i,p(j)}=PS_{i,p(j)}/\sum_{j=1}^{k}PS_{i,p(j)}$;}
11. $PH(t)=\sum_{j=1}^{k}w_{i,p(j)}x_{i,p(j),n}$;
12. To extract SF and sfm_i, create a SFM_{SF} for Datasets;
13. in Test－Data;
14. if (have a SF in Tsw+1)
15. update SFM_{SF} and sfm_i;
16. set λ for X_i;
17. update $w_{i,p(j),Tsw+1}$ by λ;
18. return 10;
19. return PH(t);

5 实证分析

为了充分验证本文提出的方法的实用性和有效性，选择了31个上证行业板块指数进行实证分析。在31个行业板块的历史日收盘指数数据上应用SFM-PG算法预测板块指数的走势，并分析其相关性，以为投资者的投资决策提供参考。

5.1 行业板块网络

5.1.1 数据预处理

板块数据来源于大智慧软件提供的2008年12月24日至2013年01月10日期间的日收盘指数共31组984个数据，利用这些数据学习行业板块网络结构。实验中采用具有连续交易量的每日收盘指数，计算日收益率：

$$r_i(t)=\ln[p_i(t)/p_i(t-1)] \quad (11)$$

其中，$p_i(t)$ 表示板块 i 在第 t 天的收盘指数，$r_i(t)$ 为板块 i 从第 $(t-1)$ 天到第 t 天的对数收益率。计算出对数收益率之

后，使用下式对其进行标准化处理：

$$R_i(t)=(r_i(t)-\bar{r}_i)/S_i \tag{12}$$

其中，$R_i(t)$表示标准化处理之后的板块 i 在第 t 天的收盘指数，$\bar{r}_i$ 表示第 i 个板块指数的样本均值，S_i 表示第 i 个板块指数的样本标准差。标准化处理之后，为了获得行业板块网络，需要对日收盘指数的对数收益率进行离散化，将连续数据转化为离散的数据，离散化后的数据由“1”、“2”和“3”组成，分别表示板块指数下降、基本不变和上升。

5.1.2 行业板块网络的构建

以 31 个行业板块作为贝叶斯网络的结点，以 5.1.1 节中离散处理后的数据为样本数据，数据离散化和网络结构的学习都是基于贝叶斯工具包进行的。行业板块 BN 结构如图 3 所示，有向边表示行业板块结点之间的影响关系，图中的结点数字对应表 1 中的每个板块。

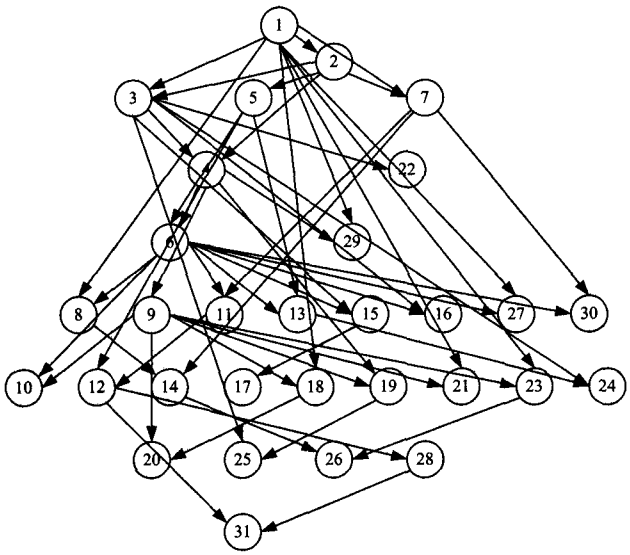


图 3 行业板块贝叶斯网络

表 1 31 个上证行业板块及其代码

1 工程建筑(991002)	12 银行类(991017)	23 外　贸(991031)
2 电　力(991003)	13 旅游酒店(991018)	24 教育传媒(991032)
3 计算机(991004)	14 煤炭石油(991019)	25 仪电仪表(991033)
4 电子信息(991006)	15 酿酒食品(991020)	26 有色金属(991034)
5 房地产(991007)	16 农林牧渔(991021)	27 造纸印刷(991035)
6 纺织服装(991008)	17 商业连锁(991023)	28 券　商(991036)
7 钢　铁(991009)	18 建　材(991024)	29 通　信(991135)
8 供水供气(991010)	19 其他行业(991025)	30 运输物流(991136)
9 化工化纤(991011)	20 交通工具(991026)	31 保　险(991255)
10 电　器(991014)	21 机　械(991027)	
11 交通设施(991016)	22 医　药(991028)	

5.2 算法实证分析

在行业板块网络上，以结点 5(房地产板块)为例对 SFM-PG 算法进行实证分析，验证算法的实用性和有效性。使用 3 个算法与 SFM-PG 进行对比分析，分别是 ARIMA、SKSVR 以及不加入流特征模式的 SFM-PG 算法(WT-PG)。

5.2.1 数据集

在实证分析中，使用两个数据集对算法的有效性和实用性进行验证。两个数据集分别选取上证指数中上升渠道和下降渠道中的一部分，标记为 DS-Ⅰ和 DS-Ⅱ，数据形式是上证指数和行业板块的日收盘指数。将每个数据集分成两个部分，分别是训练数据集和测试数据集，详细见表 2 所列。

表 2 数据集

Datasets	Train-Data	Test-Data
DS-Ⅰ	2009/3/10-2009/9/3	2009/9/4-2009/10/16
DS-Ⅱ	2011/9/1-2012/7/20	2012/7/23-2012/8/24

在数据集 DS-Ⅰ和 DS-Ⅱ上，将 SFM-PG 算法与 ARIMA、SKSVR、WT-PG 算法进行对比分析，使用 RMSE 评价算法性能，RMSE 定义如下：

$$RMSE=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(p_i-\hat{p}_i)^2} \tag{13}$$

其中，p_i 表示第 i 个板块的原始收盘数据，$\hat{p}_i$ 表示第 i 个板块的预测数据，T 表示评价窗口天数。

5.2.2 算法参数优选

取结点 5(房地产板块)为目标板块，从 BN 网络(见图 3)中得到结点 5 的 $MB(X_5)$，如图 4 所示，$MB(X_5)=\{X_1,X_2,X_4,X_6,X_{11},X_{12},X_{13}\}$，目标板块的同辈群体 $PG(5)=MB(X_5)$。在数据集 DS-Ⅰ和 DS-Ⅱ上，分别构建目标板块的 SFM-PG 模型并进行参数优选，同时，在这两个数据集上对对比算法(ARIMA、SKSVR 和 WT-PG)进行建模。

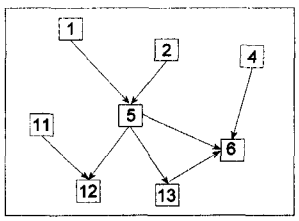


图 4 结点 5(房地产板块)的马尔科夫毯

对于 ARIMA 算法，考虑 25 个模型，即 ARIMA$(m,1,n)$，其中 $m\in\{1,2,3,4,5\}$，$n\in\{1,2,3,4,5\}$，在 DS-Ⅰ和 DS-Ⅱ测试数据集中学习，得到最佳模型，如图 5 所示，对应的最小 RMSE 见表 4。对于 SKSVR 模型，有惩罚参数 C、约束参数 ε 和核函数参数 η，确定参数 $C=1$、$\varepsilon=0.001$、$\eta=0.05$，得到最佳 SKSVR 模型，最小 RMSE 见表 4。

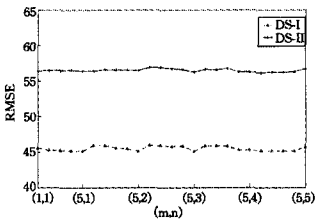


图 5 不同参数的 ARIMA 模型的预测精度

SFM-PG 算法包括两个过程，分别是窗口跟踪式动态更新预测过程和流特征提取并约束更新过程。在构建 SFM-PG 模型时，使用单一的窗口跟踪式动态更新预测算法(WT-PG)与其进行对比。

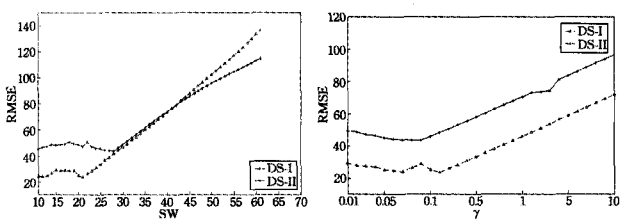
SFM-PG 算法有参数 Tsw、γ 和 λ，WT-PG 算法有参数 sw 和 γ，其中，Tsw 和 sw 取值相同，即两个算法共用参数 Tsw 和 γ。在数据集 DS-Ⅰ和 DS-Ⅱ上，首先对参数 Tsw 和 γ 进行优选，构建跟踪模式和 WT-PG 模型。对于参数 Tsw，固定参数 γ 和 λ 不变，尝试 $10\leqslant Tsw\leqslant 60$ 下预测 RMSE，结果显示，当 $10\leqslant Tsw\leqslant 30$ 时，能够获得较小的 RMSE，如图 6(a)所示。

对于参数 γ，固定参数 Tsw 和 λ 不变，尝试 28 个不同的取值，取值范围为 $0.01\leqslant\gamma\leqslant 0.09$、$0.1\leqslant\gamma\leqslant 0.9$、$1\leqslant\gamma\leqslant 9$，当 $0.05\leqslant\gamma\leqslant 0.09$、$0.1\leqslant\gamma\leqslant 0.3$ 时，预测 RMSE 较小，如图 6

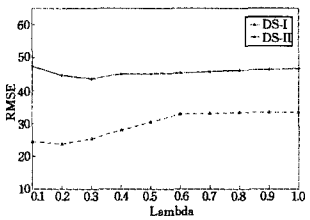
(b)所示。

在对测试数据集进行预测时，窗口跟踪式预测模型的结构不变，而当流特征模式加入新的流特征时，流特征模式得到更新，再使用参数 λ 对跟踪模式进行约束更新，获取 SFM-PG 模型。对于参数 λ，固定参数 Tsw 和 γ 不变，尝试 $0.1 \leqslant \lambda \leqslant 1.0$ 的预测 RMSE，结果显示，当 $0.1 \leqslant \lambda \leqslant 0.4$ 时，能够获得较小的 RMSE，如图 6(c)所示。

在数据集 DS-Ⅰ和 DS-Ⅱ上确定 SFM-PG 模型和 WT-PG 模型，从图 6(a)、(b)和(c)可以看出，最佳模式的参数见表 3。



(a) Tsw 与 RMSE 之间的关系　　(b) Gamma(γ)与 RMSE 之间的关系



(c) Lambda(λ)与 RMSE 之间的关系

图 6

表 3　最佳参数表

Datasets	Tsw	γ	λ
DS-Ⅰ	21	0.2	0.2
DS-Ⅱ	28	0.09	0.3

5.2.3　算法预测分析

在数据集 DS-Ⅰ和 DS-Ⅱ上，分别确定 SFM-PG 算法及其对比算法的最佳模型，随后在测试数据集上对股票走势进行预测。

在构建 SFM-PG 算法中的流特征模式时，通过大盘 K 线特征看出，DS-Ⅰ和 DS-Ⅱ中的测试数据集分别是在上升通道和下降通道，通过训练数据集，分别得到如下的条件流特征模式：

$$SFM_{DS-Ⅰ}=\{SF_1, SF_4(SF_6), SF_1, SF_6, SF_6, SF_2, SF_6, SF_3(SF_5), SF_1\}$$

$$SFM_{DS-Ⅱ}=\{SF_1, SF_4, SF_1, SF_2, SF_5, SF_2, SF_6, SF_2, SF_6, SF_2, SF_6, SF_1\}$$

在 DS-Ⅰ的测试数据集中，2009 年 9 月 21 日前形成一个顶部特征 SF_4，更新 $SFM_{DS-Ⅰ}$。同理，在 DS-Ⅱ的测试数据集中，2012 年 8 月 6 日前形成一个顶部特征 SF_3，更新 $SFM_{DS-Ⅱ}$。更新的 SFM_{DS-I} 和 $SFM_{DS-Ⅱ}$ 如下：

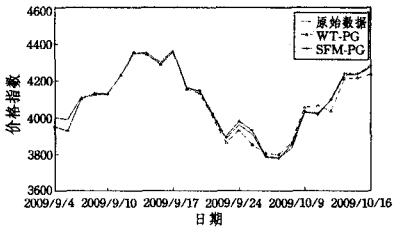
$$SFM_{DS-Ⅰ}=\{SF_1, SF_4(SF_6), SF_1, SF_6, SF_6, SF_2, SF_6, SF_3(SF_5), SF_1, SF_4\}$$

$$SFM_{DS-Ⅱ}=\{SF_1, SF_4, SF_1, SF_2, SF_5, SF_2, SF_6, SF_2, SF_6, SF_2, SF_6, SF_1, SF_3\}$$

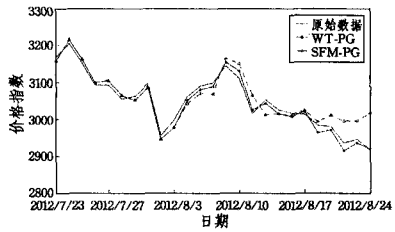
更新 $SFM_{DS-Ⅰ}$ 和 $SFM_{DS-Ⅱ}$ 后，通过表 3 的最佳 γ 对跟踪模型进行约束更新，获取 SFM-PG 模型进行预测，图 7 是 SFM-PG 算法在数据集 DS-Ⅰ和 DS-Ⅱ下的预测结果。表 4 是 ARIMA、SKSVR、WT-PG 和 SFM-PG 算法的预测结果的 RMSE。

表 4　4 种算法最优模型的预测 RMSE 比较

算法	RMSE	
	DS-Ⅰ	DS-Ⅱ
ARIMA	45.106	56.142
SKSVR	25.168	46.346
WT-PG	23.735	43.593
SFM-PG	21.273	41.435



(a) SFM-PG 算法的预测结果(DS-Ⅰ)



(b) SFM-PG 算法的预测结果(DS-Ⅱ)

图 7

结束语　提出了一种基于条件流特征模式的股市跟踪预测算法——SFM-PG 算法，其利用贝叶斯网络、马尔科夫毯、同辈群体分析、流特征模式的思想构建股票价格预测模型。利用贝叶斯网络可以构建股票之间的网络，从而可以快速找到目标对象的马尔科夫毯，确定目标对象的同辈群体；同辈群体是波动和特征相似的一类股票，通过欧氏距离和接近度确定同辈群体的跟踪式预测模型，能够有效地消除数据非正态分布对预测的影响；流特征模式能够有效地捕捉历史数据中的隐藏知识，提取流特征，组合成特征模式，从而指导跟踪模型的更新，进一步提高算法的预测精度。应用 SFM-PG 算法对上证股票板块网络中的房地产板块指数进行预测，并与 ARIMA 和 SKSVR 算法进行对比分析，验证了算法的实用性和有效性。如何优化流特征模式以及将算法应用于其他领域将是下一步的研究课题。

参 考 文 献

[1] Hadavandi E, Shavandi H, Ghanbari A. Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting[J]. Knowledge-Based Systems, 2010, 23(8): 800-808

[2] Kazem A, Sharifi E, Hussain F K, et al. Support vector regression with chaos-based firefly algorithm for stock market price forecasting[J]. Applied Soft Computing, 2013, 13(2): 947-958

[3] Yi Zuo, Kita E. Stock price forecast using Bayesian network[J]. Expert Systems with Applications, 2012, 39(8): 6729-6737

[4] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation[J]. Econometrica, 1982, 50(4): 987-1008

[5] Bollerslev T. Generalized autoregressive conditional heteroskedasticity[J]. Journal of Econometrics, 1986, 31(3): 307-327

[6] Wang Ju-jie, Wang Jian-zhou, Zhang Zhe-george, et al. Stock index forecasting based on a hybrid model[J]. Omega, 2012, 40 (6): 758-766

[7] Kao Ling-jing, Chiu Chih-chou, Lu Chi-jie, et al. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting [J]. Neurocomputing, 2013, 99(1): 534-542

[8] Vapnik V. The nature of statistical learning theory[M]. New York, USA: Springer-Verlag, 1995

[9] Yeh Chi-yuan, Huang Chi-wei, Lee S-J. A multiple-kernel support vector regression approach for stock market price forecasting[J]. Expert Systems with Applications, 2011, 38(3): 2177-2186

[10] Cai C X, Kyaw K, Zhang Q. Stock index return forecasting: The information of the constituents[J]. Economics Letters, 2012, 116 (1): 72-74

[11] Taylor S J, Yadav P K, Zhang Yuan-yuan. The information content of implied volatilities and model-free volatility expectations: Evidence from options written on individual stocks[J]. Journal of Banking & Finance, 2010, 34(4): 871-881

[12] Kim Y, Sohn S Y. Stock fraud detection using peer group analysis[J]. Expert Systems with Applications, 2012, 39(10): 8986-8992

[13] Daly, Ronan. Learning Bayesian Networks: Approaches and Issues[J]. Knowledge Engineering Review, 2011, 26(2): 99-157

[14] Bui A T, Jun C H. Learning Bayesian network structure using Markov blanket decomposition[J]. Pattern recognition Letters, 2012, 33(16): 2134-2140

[15] Pearl J. *Probabilistic Reasoning in Intelligent Systems* [M]. Morgan Kaufmann, 1988

(上接第 18 页)

加 1/4 左右的硬件开销，即可实现四倍精度 FMA 运算。

表 4 F_SIMDFMA 综合结果

设计	面积(um²)	百分比
F_SIMDFMA	809070.61	100.00%
DPFMA_3	161080.80	19.91%
DPFMA_2	160904.40	19.89%
DPFMA_1	160993.20	19.90%
DPFMA_0	161045.40	19.90%
SIMD_QPFMA	163279.80	20.18%

从上述逻辑综合结果来看，基于 64 位×4 的双精度浮点 SIMD FMA 部件设计 QPFMA 可以显著减少硬件开销。对于已有的双精度浮点 SIMD FMA 部件来说，只需要增加少量硬件开销，即可实现 4 倍精度 FMA 运算。

结束语 本文完成了一种基于 SIMD 乘加部件的 QPFMA 的原型设计、验证与逻辑综合，并与其它设计进行比较分析，主要贡献在于：设计了基于 64 位×4 的双精度浮点 SIMD FMA 部件的 QPFMA 结构，7 级流水线，面积与一个双精度 FMA 部件基本相当，显著减少了实现四倍精度 FMA 运算的延迟和硬件开销。本文的研究也为浮点 SIMD 部件的设计提供了一条重要思路，以少量的硬件开销实现 SIMD 部件功能的扩展，进一步发挥 SIMD 部件的作用。下一步将尝试进一步对设计进行流水线时序优化，重点对现有设计的关键路径进行优化研究，以实现更高的频率。

参 考 文 献

[1] Bailey D H. High-precision floating-point arithmetic in scientific computation [J]. Computing in Science and Engineering, 2005, 7 (3): 54-61

[2] IEEE Computer Society. IEEE Standard for Floating-Point Arithmetic[S]. IEEE Standard 754-2008, 3 Park Avenue New York, NY 10016-5997, USA, August 2008

[3] 黎铁军，李秋亮，徐炜遐. 一种 128 位高性能全流水浮点乘加部件[J]. 国防科技大学学报，2010，32(2)：56-60

[4] Akkas A, Schulte M J. Dual-Mode Floating-Point Multiplier Architectures with Parallel Operations [J]. Journal of Systems Architecture, 2006, 52: 549-562

[5] Akkas A. Dual-Mode Quadruple Precision Floating Point Adder [C] // 9th Euromicro Conference on Digital System Design. 2006: 211-220

[6] Akkas A. A Dual-Mode Quadruple Precision Floating-Point Divider[C] // Fortieth Asilomar Conference on Signals, Systems and Computers. 2006: 1697-1701

[7] Gok M, Ozbilen M M. Multi-functional floating-point MAF designs with dot product support [J]. Microelectronics Journal, 2008, 39(1): 30-43

[8] Huang Li-bo, Ma Sheng, Shen Li, et al. Low-Cost Binary128 Floating-Point FMA Unit Design with SIMD Support[J]. IEEE Transactions on Computers, 2012, 61(5): 745-751

[9] 张峰，黎铁军，徐炜遐. 一种 128 位高精度浮点乘加部件的研究与实现[J]. 计算机工程与科学，2009，31(2)：93-103

[10] 雷元武，窦勇，郭松. 基于 FPGA 的高精度科学计算加速器研究[J]. 计算机学报，2012，35(1)：112-122

[11] Yu Xiao-yan, Chan Yiu-Hing, Curran B, et al. A 5GHz+ 128-bit Binary Floating-Point Adder for the POWER6 Processor[C] // Proceedings of the 32nd European Solid-State Circuits Conference. 2006: 166-169

[12] Intel Company. Intel Compilers and Libraries [EB/OL]. http://soft-ware. intel. com/en-us/articles /intel-cimpilers/, 2012, 12/24

[13] Fousse L, Hanrot G, Lefevre V, et al. Mpfr: A multiple-precision binary floating-point library with correct rounding [J]. ACM Transactions on Mathematical Software (TOMS), 2007, 33(2): 1-14

[14] Hida Y, Li X S, Bailey D H. Quad-double arithmetic: Algorithms, implementation, and application[R]. LBL-46996. Lawrence Berkeley National Laboratory, Berkeley, CA, 2000

[15] Firasta N, et al. Intel AVX: New Frontiers in Performance Improvements and Energy Efficiency[M]. White paper, 2008

[16] IBM Corporation. PowerPC Microprocessor Family: Vector/SIMD Multimedia Extension Technology Programming Environments Manual [M]. 2005

[17] Trong S D, Schmookler M, Schwarz E M, et al. POWER6 Binary Floating-Point Unit[C] // Proceedings of the 18th IEEE Symposium on Computer Arithmetic. Montpellier, France, 2007: 77-86

[18] Boersma M, Kroener M, Layer C, et al. The POWER7 Binary Floating-Point Unit[C] // Proceedings of IEEE Symposium on Computer Arithmetic. Tübingen, Germany, IEEE Computer Society, 2011

[19] Haring R A, Ohmacht M, Fox T W, et al. The IBM Blue Gene/Q Compute Chip [M]. IEEE Micro, March/April 2012: 48-60

[20] TOP500. TOP500 supercomputing sites [EB/OL]. http://www. top500. org/lists/2012/06, 2012

[21] Maruyama T, Yoshida T, Kan R, et al. SPARC64 VIIIfx: a New-Generation Octocore Processor for Petascale Computing[M]. IEEE Micro, March/April 2010: 30-40