doi: 10.3969/j.issn.0490-6756.2018.01.011

基于概率后缀树的股票时间序列预测方法研究

程小林, 郑　兴, 李旭伟
(四川大学计算机学院, 成都 610065)

摘　要: 在时间序列符号化基础上,本文引入概率后缀树 PST 模型,构建基于时间序列符号化和概率后缀树相结合的股票预测模型. 本文选择在沪深 300 的 10 支股票数据上将预测模型与传统的马尔科夫模型 MM 和自回归移动平均模型 ARMA 进行对比,结果显示本文提出的股票预测模型优于 MM 模型和 ARMA 模型,验证了本文所提出的预测模型在投资收益上的有效性.

关键词: 股票数据挖掘; 时间序列符号化; 高斯混合模型聚类; 概率后缀树
中图分类号: TP391　　**文献标识码**: A　　**文章编号**: 0490-6756(2018)01-0061-06

Research of stock time series based on probabilistic suffix tree

CHENG Xiao-Lin, ZHENG Xing, LI Xu-Wei
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: A stock forecasting model was introduced in this paper, which was based on the combination of time series symbolization and Probabilistic Suffix Tree (PST). In addition, the Markov Model (MM) and the Auto Regressive Moving Average Model (ARMA) was compared with the forecasting model of this paper. The stock of 10 CSI 300 indices was used as the experimental sample. The results show that the stock forecasting model proposed in this paper is better than the MM model and the ARMA model, and prove the validity of the forecasting model proposed in this paper.

Keywords: Stock data mining; Time series symbolization; Gaussian mixture modeling; Probabilistic suffix tree

1　引　言

近年来,随着人工智能的飞速发展,学科交叉的日益深入,一些在其他它领域应用的技术不断地运用在股票预测中,如随机过程、混淆理论以及小波分析等. 这些新方法的引入为股票预测研究注入了强大的动力,目前,人工神经网络和支持向量机等计算机技术已经成为研究人员和投资者研究股市的重要方法.

陈曦在 2013 年利用分段线性表示方法(Piecewise Linear Representation ,PLR)和加权支持向量机(Weighted Support Vector Machine, WSVM)相结合来预测股票拐点[1]; Chang 等在 2015 年提出了一个基于核 PCA (Principal Component Analysis)方法提取关键特征来提高股票预测准确率模型[2]; Nair 等人对股票时间序列进行聚类从而产生股票交易决策[3]; Chang 提出了进化趋势反转模型来预测股票交易规则[4]; Liao[5] 等人针对股票交易决策信号问题构建了动态阈值模型来预测未来的交易信号.

收稿日期: 2017-07-08
基金项目: 国家自然科学基金(61173099)
作者简介: 程小林(1993－), 男, 重庆人, 硕士生, 研究方向为金融时间序列. E-mail:1184177375@qq.com
通讯作者: 李旭伟. E-mail:lixuwei@scu.edu.cn

在现有研究的基础上,本文构建了一个基于高斯混合模型聚类符号化和概率后缀树 PST(Probabilistic Suffix Tree)的股票预测模型,实现对股票下一周期趋势进行预测.本文主要的研究工作包含以下几点:

(1) 获取原始交易数据,计算相关股票技术指标,构造模型的特征向量;

(2) 使用高斯混合模型聚类方法将股票序列数据符号化,将同一支股票的特征向量进行多次聚类,获得多组聚类结果,为每次聚类获得的多个簇分配不同的字符,获得多个符号序列,实现股票交易数据符号化;

(3) 使用符号序列构建概率后缀树,通过选择验证集上最佳收益率方法解决高斯混合模型聚类算法的不稳定性;

(4) 通过实验验证本文构建的基于概率后缀树的股票预测模型能有效提高收益率.

2 股票序列符号化与预测

2.1 数据预处理

与传统的时间序列数据相比,股票交易数据时间序列有自身的一些特性,针对股票交易数据的特征,本文设计了以下方法对股票交易序列进行处理,股票交易数据序列符号化流的具体流程步骤如下.(1) 采集股票日交易数据;(2) 对获取的数据进行复权处理,并以 5 天为周期进行压缩;(3) 计算相应的技术指标;(4) 对部分数据进行归一化处理;(5) 使用高斯混合模型聚类对股票数据进行聚类;(6) 为每个簇分配符号;(7) 对聚类结果进行分析,选取最佳的聚类结果.

为了能在不同角度反映股票趋势变化,研究人员在原始交易数据的基础上提出了技术指标这一概念来反映股票市场的变化,不同的技术指标都是对股票价格趋势的定量分析,以帮助投资者进行投资决策.本文选择了部分技术指标,如:RSI(Relative Strength Index)相对强弱指数、WR(Williams% Rate)威廉指数、MA(Moving Average)移动平均线指标、BIAS 偏离率指标、EMA(Exponential Moving Average)指数平均数指标等.

前文介绍的技术指标中都包含有参数 n(n 表示天数),投资者通常使用不同参数的技术指标相结合的方式对股票进行分析,不同的参数的指标之间的差异非常重要,不仅可以从多个角度来反映股票的走势,而且其相交点往往是股票走势的关键点.例如 RSI 指标在参数取 12 和参数取 6 的差值上就有着明显的意义,当其值小于 0 时,表明目前人们买入的愿望更强烈,其值越小,则买入愿望越强,反之亦然.因此,本文在选择输入特征变量时,不仅使用了技术指标本身的值,还考虑了同一技术指标在不同参数值下的差异和趋势,差异和趋势用式(1)和式(2)来表示.

$$Dif_t(EMA,n_1,n_2)=EMA_t^{(n_1)}-EMA_t^{(n_2)} \tag{1}$$

$$Trd_t^{(\mathrm{index},n_1,n_2)}=(Dif_t^{(EMA,n_1,n_2)}-Dif_{t-1}^{(EMA,n_1,n_2)})*\mathrm{sign}(Dif_{t-1}^{(EMA,n_1,n_2)}) \tag{2}$$

其中,式(1)表示一个技术指标在不同参数下的差异,n_1 ,n_2 表示不同的参数取值天数;式(2)表示技术指标的趋势,其中 sign()为符号函数.

本文选择投资者常用的 WR、RSI、BIAS 三个指标计算其 $Dif_t^{(EMA,n_1,n_2)}$ 和 $Trd_t^{(EMA,n_1,n_2)}$,其中对于 RSI 和 BIAS 指标,n_1 取 12,n_2 取 6,对于 WR 指标,n_1 取 10,n_2 取 6.在此基础上,本文构建的时间序列如下所示.

$$\bar{x}^{\mathrm{new}}=(\bar{x}_1^{\mathrm{new}},\bar{x}_2^{\mathrm{new}},\bar{x}_3^{\mathrm{new}},\ldots,\bar{x}_t^{\mathrm{new}},\ldots,\bar{x}_n^{\mathrm{new}})$$

$$\begin{aligned}\bar{x}_t^{\mathrm{new}}=(&\mathrm{p}_t^{\mathrm{v}},\mathrm{MA}_t^{(10)},\mathrm{RSI}_t^{(12)},\mathrm{RSI}_t^{(6)},\\&\mathrm{Dif}_t^{(\mathrm{RSI},12,6)},\mathrm{Trd}_t^{(\mathrm{RSI},12)},\mathrm{WR}_t^{(10)},\mathrm{WR}_t^{(6)},\\&\mathrm{Dif}_t^{(\mathrm{RSI},10,6)},\mathrm{BIAS}_t^{(6)},\mathrm{Dif}_t^{(\mathrm{BIAS},12,6)},\\&\mathrm{Trd}_t^{(\mathrm{RSI},12)})\end{aligned}$$

2.2 基于高斯混合模型聚类的符号化

聚类算法主要分为四类:层次聚类算法、划分式聚类算法、基于网格和密度的聚类算法以及其他聚类算法[9],其中高斯混合模型聚类算法被广泛使用,本文选择了高斯混合模型聚类算法对子序列进行聚类处理.

高斯混合模型聚类方法中的 n 值是需要预先设定的参数,表示最终输出簇的数量,n 的取值直接关系到聚类结果,需要更具实际情况做相应的选择.本文通过遗传算法多次实验得到当 n 取值为 4 时,聚类效果较为理想,各簇数据之间得到了很好的区分.

聚类完成后,每个数据点都会被归入相应的簇,接下来需要做的工作是为每个点赋予字符.符号化的大致思想是为同一簇中的数据赋予相同的字符,每个簇对应同一个字符,然后按照时间先后顺序将每个符号组合起来,即为符号化序列.本文通过高斯混合模型聚类将数据划分到 4 个簇中,按

照每个簇的数据对应一个符号的原则，最终的符号化序列将由 4 个字符组成，该序列即为概率后缀树的输入信息. 该过程的数学表达如下：

输入时间序列 $\bar{x}^{\text{new}}=(\bar{x}_1^{\text{new}},\bar{x}_2^{\text{new}},\bar{x}_3^{\text{new}},\ldots,\bar{x}_t^{\text{new}},\ldots,\bar{x}_n^{\text{new}})(t=1,2,\cdots,n)$，4 个符号组成的符号集合 $M=\{m_1,m_2,\ldots,m_i,\ldots,m_6\}(i=1,2,\ldots,6)$，输出的符号化时间序列为 $S=\{s_1,s_1,\ldots,s_i,\ldots s_n\}S$ 由 M 中的元素构成，且元素个数与 $\bar{x}^{\text{new}}$ 相同.

在使用概率后缀树预测时，输出信息为下一个可能出现的符号以及概率. 本文的模型中，不同的符号代表不同的交易信号，每一个字符有其特定的含义，因此在分配字符前需要确定不同簇内的数据代表的含义. 对于投资者而言，需要获取的交易信号通常有三种，第一种是买入信号，即下一个周期股价会出现上涨；第二种是持有信号，表示下一个周期的股价不会有大的变化，只有小范围的波动；第三种是卖出信号，表示下一个周期股票价格会出现下跌的情况. 判断以上三种交易信号的最重要特征为交易周期的股票价格变化趋势，即收益率，因此本文使用簇中所有数据的平均收益率来确定每个簇的数据对应的类型.

现以(000002 万科 A)为例，选取时间区间为 2011 年 3 月到 2016 年 9 月，总共 97 个子序列(万科 A 在 2015 年 12 月 8 日到 2016 年 7 月 4 日停牌，因此子序列数量少于正常值)进行聚类和符号化操作并计算平均收益率. 在最终获取的 6 个符号中，有 3 个平均收益率为正，1 个平均收益率为负，每个簇的评价收益率和符号分配如表 1.

表 1　符号统计表

Tab. 1　The statistical resultsofsymbol

对应符号	v	u	a	b
平均收益率	−0.01937	0.00207	0.02348	0.04992
数量	57	56	41	43

2.3　基于概率后缀树的预测

概率后缀树作为一个树形的存储结构，其存储的主要内容为训练序列中的上下文统计概率，运用这些概率即可对未来的序列进行预测[10]. 概率后缀树的预测思想如下：从根节点开始，按照序列倒序的方式匹配树中的各层节点，得到概率最大的序列即为预测的下一个符号，如果无法匹配整个序列，则去掉离当前时刻最远的字符，继续从根节点开始匹配序列，直到匹配成功. 去掉最远的字符就使得当前匹配的阶数发生改变，体现了概率后缀树变阶的思想.

设定一个待匹配序列 $S=\{s_1,s_2,s_3,\ldots,s_n\}$ 规定 $\text{suffix}(S_i)(1\leq i\leq n)$ 表示序列 S 的后缀，可得出 S 最长后缀 $\text{suffix}(S_1)=\{s_1,s_2,s_{n-4},\ldots,s_n\}$，最短后缀为 $\text{suffix}(S_n)=\{s_n\}$，使用 L 阶概率后缀树匹配序列 S 的过程如下.

(1) 选择序列 L 阶概率后缀树能够匹配到 S 的最长后缀 $\text{suffix}(S_{n-L+1})$，从根节点开始倒序搜索概率后缀树匹配 $\text{suffix}(S_{n-L+1})$.

(2) 若无法匹配 $\text{suffix}(S_{n-L+1})$，则去掉离 n 时刻最远的数 s_{n-L+1}，此时需要匹配的序列从 $\text{suffix}(S_{n-L+1})$ 变为 $\text{suffix}(S_{n-L+2})$，若匹配成功，则结束对序列 S 的搜索，若 $\text{suffix}(S_{n-L+2})$ 也无法匹配，则去掉 s_{n-L+2}，此时匹配的目标序列为 $\text{suffix}(S)$，以此类推，直到匹配到原始序列 S 的最长后缀序列 $\text{suffix}(S_m)(n-L+1\leq m\leq n)$ 为止.

(3) 根据匹配成功的最长后缀 $\text{suffix}(S_m)$ 定位到其对应节点，获取该节点对应的概率向量，向量中的概率值即为预测的下一个符号的概率选择其中最大概率值对应的字符为预测的下一个字符[11,12].

由于本文构建的股票序列中符号种类较少，各符号出现频率相对较高，会导致某节点的概率向量中概率值完全相等的情况出现，此种情况对未来符号的预测就毫无意义. 针对这一情况，本文规定若某序列在概率后缀树中的匹配结果出现概率向量中概率值完全相等的情况，则认为本次匹配失败，需要将其中历史最久远的字符去除，更改匹配目标，重新获取符合要求的结果，如图 1 所示.

概率后缀树每次预测的是当前时刻 t 之后的下一时刻的符号概率，即 $t+1$ 的符号概率，完成对 $t+1$ 时刻预测后，若需要对 $t+2$ 时刻进行预测，可以继续使用原有概率后缀树完成预测. 但是原有概率后缀树中并未包含有 $t+1$ 时刻的符号信息，对于本文的符号化股票序列而言，当前时刻的符号状态与其上一时刻的符号状态有很大的关系，上一时刻的符号包含影响当前时刻的重要因素. 因此本文在每次预测完成后都对原有后缀树进行更新，将最新的符号信息添加到概率后缀树中，从而确保每一次预测结果的精确.

设下一个周期的符号为 x，更新概率后缀树流程如下.

(1) 更新各个符号在新的符号序列中出现的

概率,即根节点的概率向量;

(2) 搜索所有长度不大于 L 且后缀包含 x 的子序列;

(3) 对每一个子序列按照倒序的方式匹配,并更新其节点概率,若原有树中无此子序列导致匹配失败,则判断该序列在原始序列中出现的概率是否大于 $P_{\min}$,若大于则将其作为新节点添加到概率后缀树中,否则不添加到概率后缀树中;更新后的概率后缀树加入了最新的符号信息,使得对未来预测结果更为可靠. 预测流程图如图 2 所示.

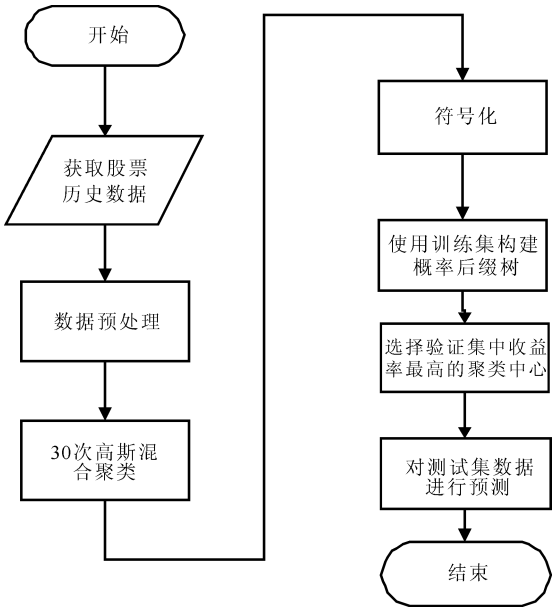

图 1 基于概率后缀树的股票预测模型流程图

Fig. 1 The process of stock data forecasting based on PST

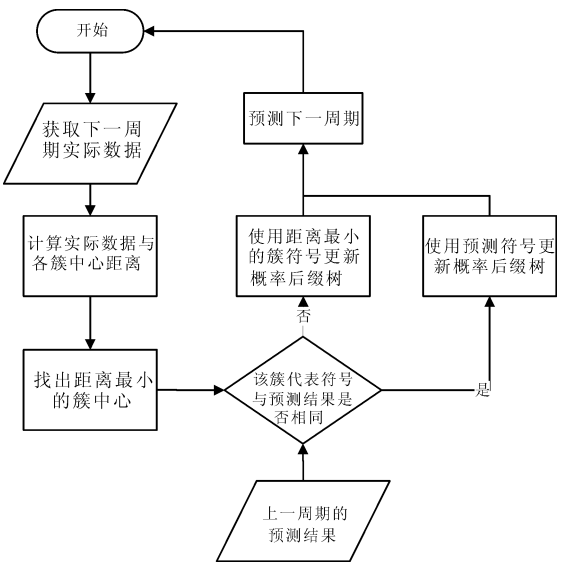

图 2 预测流程图

Fig. 2 The process of forecasting

3 实验与分析

3.1 实验方案

实验方案包括实验数据选取,实验参数选择和投资者策略.

3.2 实验数据选取

为验证预测模型的有效性,本文选择使用沪深 300 指数包含的 10 支股票对模型进行验证,选择数据的时间跨度从 2014 年 4 月 9 日到 2017 年 4 月 7 日.

日交易数据的压缩方式 5 日为周期,数据按照 80%、10%、10%的比例划分为训练集、验证集、测试集.

3.3 实验参数选择

由于实验中涉及到部分参数,现对所有参数进行统一规定,聚类中的参数 n 设为 4,构建概率后缀树中的参数设定如表 2 所示.

表 2 概率后缀树参数选择

Tab. 2 The parameters of PST

参数	L	r	$P_{\min}$	$\gamma_{\min}$
数值	5	0.07	0.06	0

3.4 投资策略

本文提出的模型最终预测得到的是股票下一个周期的符号,根据符号的含义得到该周期的涨跌信号,进而可以得到交易信号,如预测的下一个周期的符号代表的收益率大于 0 时,即为买入信号,以当前周期的收盘价买入,并且在下一个周期结束时以该周期的收盘价卖出. 为了简化,本文采用简单的固定买入策略:当出现买入信号时,买入固定金额的股票,在周期结束时全部卖出. 在实际的股票交易中,还需要考虑交易手续费,因此在计算最终收益率时需要考虑交易手续费. 固定策略的投资收益率计算公式如式(3)所示.

$$P = \sum_{i=1}^{m} \left(\frac{(1-F_s) * p_{i+1}^c - (1+F_b) * p_i^c}{(1+F_b) * p_i^c} \right) \tag{3}$$

其中,F_b 为买入交易的手续费比率;F_s 为卖出交易的手续费比率;根据目前的交易规则,$F_b = 0$;$F_s = 0.1\%$,m为交易次数.

为防止预测结果导致的大幅亏损,本文还使用了相应的止损策略,规定当某个周期买入股票后出现下跌超过 5%的情况时,将买入的股票立即卖出,以达到减少亏损的目的.

3.5 评价指标

为评价模型的预测效果，本文选择了预测正确率、趋势预测正确率、持有收益率、收益率、胜率共 6 个指标来分析最终结果. 其中趋势预测正确率是指每次预测符号代表的收益率与实际收益率正负情况是否一致，若一致，则表明预测正确，反之则预测错误；胜率是指所有交易次数中，获利的交易次数所占比例，表示本文的预测模型每次投资的获利情况.

3.6 实验结果分析

为了能更好的验证本文模型的有效性，本文引入自回归移动平均模型 ARMA(Auto Regressive Moving Average Model)和马尔科夫模型 MM(Markov Model)与本文构建的最优概率后缀树 PST 模型进行对比. 因为本文的序列为符号序列预测，不适宜选用支持向量机 SVM(Support Vector Machine)或者循环神经网络 RNN(Recurrent Neural Networks)等方法，同时隐马尔科夫模型 HMM (Hidden Markov Model)的假设是当前时刻只与序列前一状态有关，而本文的假设是下一时刻与序列的前面多个状态相关，因此隐马尔科夫模型与本文不符.

从已经进行实验 10 支股票作为 ARMA 和 MM 模型的测试样例. ARMA 模型预测的下一个周期的价格，因此最后只统计了收益率一项指标，MM 的实验统计指标类别与 PST 一致，MM 模型和 ARMA 模型预测结果如表 3，时间为 100 天.

表 3　10 支股票预测结果

Tab. 3　The result of forecasing stock data

股票代码	模型类别	准确率(%)	趋势准确率(%)	收益率(%)	胜率(%)	持有收益率(%)	ARMA 收益率(%)
000630	PST	50.00	73.00	15.80		19.40	1.18
	MM	42.60	62.70	13.30	71.50		
000970	PST	40.00	68.00	24.20	45.50	13.10	12.94
	MM	35.50	59.70	23.60	65.60		
000002	PST	36.00	63.00	6.50	60.00	4.60	12.24
	MM	23.30	41.50	0.70	75.00		
300002	PST	55.00	73.00	15.40	60.00	11.00	3.88
	MM	11.20	56.70	−9.00	67.80		
600547	PST	62.00	63.00	−5.80	28.60	15.60	0.55
	MM	44.50	65.80	−9.90	70.00		
600804	PST	45.00	73.00	17.10	50.00	12.10	0.00
	MM	23.30	56.70	7.20	73.20		
601898	PST	62.00	65.00	15.60	69.20	8.30	0.00
	MM	26.40	56.70	17.00	70.00		
600690	PST	35.00	63.00	−2.70	55.60	−4.60	8.76
	MM	24.70	58.80	−3.10	60.00		
600104	PST	41.50	68.00	13.10	42.90	8.90	7.89
	MM	26.40	51.50	−5.50	67.10		
601633	PST	60.00	83.00	−11.20	42.10	−29.30	−12.13
	MM	24.50	55.80	−5.70	52.90		
平均值	PST	48.65	69.2	8.8	67.31	5.91	3.53
	MM	28.4	56.59	2.86	49.56		

从表 3 可以看出，MM 模型预测平均收益率为 2.86%，ARMA 模型预测收益率为 3.53%，本文构建的 PST 模型平均收益率达到 8.8%，超过其余两个模型收益率，同时也远超过买入持有收益率 5.99%. 在预测准确率和趋势预测准确率方面，PST 模型均高于 MM 模型. 在相同周期长度的情况下，MM 模型的收益率明显低于 PST. 其原因是 MM 模型只匹配了前一个周期的序列，使用较少的信息进行预测无法准确的找到序列中隐藏的规律，而 PST 模型使用了概率后缀树这一变阶马尔科夫模型，能够动态的匹配不同的长度，因此使用 PST 可以获得更高的预测准确率.

4 结 论

本文首先对原始股票交易信息进行处理,构建一个全新的时间序列作为高斯混合模型聚类的输入信息,然后使用高斯混合模型聚类方法实现时间序列符号化,将符号化的结果作为概率后缀树的输入信息,通过选择验证集上最佳收益率方法解决高斯混合模型聚类算法的不稳定性.最后在测试集得到实验结果,证明本文方法的有效性.

参考文献:

[1] 陈曦. 基于分段线性表示和支持向量机的拐点预测[D]. 厦门:厦门大学, 2013.

[2] Chang P C, Wu J L. A critical feature extraction by kernel PCA in stock trading model[J]. Soft Comput, 2015, 19: 1393.

[3] Nair B B, Kumar P K S, Sakthivel N R, *et al*. Clustering stock price time series data to generate stock trading recommendations: An empirical study [J]. Expert Syst Appl, 2017, 70: 20.

[4] Zhang X, Hu Y, Xie K, *et al*. An evolutionary trend reversion model for stock trading rule discovery[J]. Knowledge-Based Systems, 2015, 79: 27.

[5] Chang P C, Liao T W, Lin J J, *et al*. A dynamic threshold decision system for stock trading signal detection [J]. Appl Soft Comput, 2011, 11: 3998.

[6] 赵超, 唐亚勇. 分位点门限自回归时间序列模型的贝叶斯方法[J].四川大学学报:自然科学版, 2016, 53: 748.

[7] Mazeroff G, Gregor J, Thomason M, *et al*. Probabilistic suffix models for API sequence analysis of Windows XP applications[J]. Pattern Recognition, 2008, 41: 90.

[8] 孟海东, 王淑玲, 郝永宽. 基于簇特征的增量聚类算法设计与实现[J]. 计算机工程与应用, 2010, 9: 132.

[9] Begleiter R, El-Yaniv R, Yona G. On prediction using variable order markovmodels[J]. J Artif Intell Res, 2004, 22: 385.

[10] 李丰, 高峰, 寇鹏. 基于分段线性表示和高斯过程分类的股票转折点概率预测[J]. 计算机应用, 2015, 35: 2397.

[11] 张丹辉. 基于概率后缀模型的计算机病毒检测方法研究[D]. 天津: 南开大学, 2011: 108.