基于 TA 和 SA 的股价预测系统的实现*

代江波 毛建华 延 丰 刘学锋
(上海大学 通信与信息工程学院 上海 200444)

摘 要：为满足对金融市场的进一步了解以及股价预测的需求，结合投资者的情感倾向提出了一种基于 TA/SA (technical analysis/sentimental analysis)的股票价格预测模型，建立投资者情感与未来股票价格之间的关系的方案。该方案主要包括获取情感指数，建立回归模型以及计算未来股票收盘价。利用该模型预测 200 只股票价格并与 SVM 和 BP 神经网络两种模型预测结果进行比较，结果显示所提出模型的预测正确率分别提高了 10.9%和 7.4%，表明该模型具有更好的预测准确性和实用价值。

关键词：TA/SA；股票价格预测；情感指数；回归模型
中图分类号：TP391；TN91 **文献标识码**：A **国家标准学科分类代码**：520.60

Implementation of stock price forecasting system based on TA and SA

Dai Jiangbo Mao Jianhua Yan Feng Liu Xuefeng
(School of Communication & Information Engineering, Shanghai University, Shanghai 200444, China)

Abstract: In order to meet further understanding of the financial market and the demand for the stock price forecast in advance. This paper proposed a stock price forecasting model based on TA/SA (technical analysis/sentimental analysis), which aims to establish the relationship between sentiment and future stock price. The application mainly includes: get sentiment index, establish regression model and the calculation of future stock closing price. This model predicted 200 stock prices, compared with predicted results of SVM and BP neural network, and the correct rate of the proposed model increased by 10.9% and 7.4%, which shows that the model has better prediction accuracy and practical value.

Keywords: TA/SA; stock price prediction; sentiment index; regression model

0 引 言

金融论坛是投资者交流金融产品信息以及投资策略的平台，投资者通过研究金融产品的历史交易信息以及金融市场环境和政府政策，预测金融产品未来价格变化趋势。随着我国金融市场的不断发展，金融产品投资理论研究领域也得到不断拓展，投资者情绪和投资者行为的关系研究也得到空前重视，并逐渐成为行为金融学的一个重要研究方向[1]。在股票市场领域，有大量研究从不同的视角探讨论坛股评信息与股票价格变化的关系，并应用到股票预测研究中。目前，对于股票价格的预测主要分为 3 种：基于技术分析、基于情感分析以及基于技术分析和情感分析混合 3 种方法。

基于技术分析(technical analysis，TA)方法是从数据和技术角度利用历史股票价格、成交量、财务报表等数据，使用统计或其它一些预测方法预测股票价格的走势[2-3]。TA 仅仅从数据统计分析角度进行训练预测，虽然有很好的分类标准和技术指标，但并未实时考虑投资者情感因素，因此基于 SA 的模型就应运而生。基于情感分析 SA (sentimental analysis)方法是获取金融论坛，网络媒体信息等相关信息，利用情感倾向建立预测模型进行预测股票价格[4-5]；SA 利用投资者以及相关的网络信息的情感极性，进而通过预测模型进行分类预测，相对于 TA 表现出一定的优势。

基于技术和情感分析混合方法是将技术和情感作为股票预测的相关因素，预测股票价格[6-8]。文献[9]基于线性核的 SVM 作为预测模型，对 TA、SA 以及 TA/SA 这 3 种方法进行实验对比，证明 TA/SA 具有更好的效果。综上，TA/SA 混合方法弥补了单方面使用 TA 或 SA 做预测的不足；因此本文使用 TA/SA 组合方法建立股票预测模型。

基于 TA/SA 组合方法建立股票预测模型，利用投资

收稿日期：2017-01
* 基金项目：国家自然科学基金(61271061)项目资助

者情感和多元回归分析技术建立股票预测系统。基于情感指数[10]提出一种情感约束的股票预测方案，通过多元回归分析可以实现利用投资者情感预测未来股票价格，实现人们对金融市场的进一步了解以及股价预测的需求。

1　系统架构

股票预测系统模型包括 3 部分：获取情感指数、构建预测模型和预测股票价格。整体流程为：收集金融论坛股评文本信息，构建情感词典，计算情感指数；Granger 因果关系验证股评情感与股票运行相关关系，建立情感指数和股票价格的初始模型，参数估计与检验，然后建立确定最终回归模型；最后对股票收盘价格进行静态拟合和预测。系统研究框架如图 1 所示。

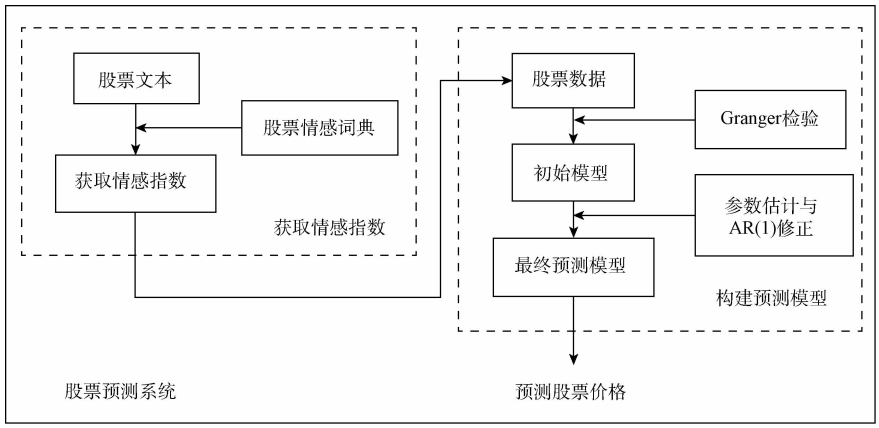

图 1　系统研究框架

1.1　情感指数计算

情感指数计算流程如图 2 所示。

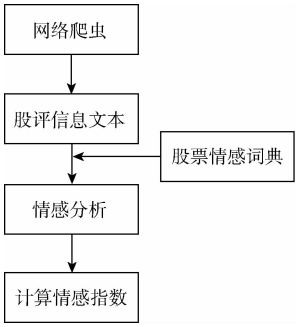

图 2　情感指数计算流程

获取情感指数是该方案的基础，其影响到股票价格预测的准确性。通过网络爬虫获取金融论坛股评信息文本，利用 Java 软件以及人工构建的情感词典[11-12]对文本进行情感分析，将计算得到的情感比例通过式(1)计算情感指数。

$$F_t = \ln\left[\frac{1+S_{正t}}{1+S_{负t}}\right] \tag{1}$$

式中：$S_{正t}$ 为第 t 天的正面情感值，$S_{负t}$ 为第 t 天的负面情感值。当 $F_t \geqslant 0$ 时说明投资者整体情感是看涨的，<0 则整体情况是看跌的。

1.2　构建预测模型

预测股票价格走势不能忽略其外界影响因素，因此考虑到自变量的因子多样化采取多元线性回归模型[13]，使得外界因素和因变量之间的关系更加清晰，有利于对数据的整体分析以及后期预测计算。

使用 Granger 因果关系[14]验证股评情感与股票运行相关关系，证明两者具有相关性后再建立情感指数和股票价格的初始模型，并进行参数估计与检验，然后建立确定最终回归模型；最终模型定义形式如式(2)所示：

$$Y = C + \beta X + AR(p) \tag{2}$$

式中：Y 为被解释变量即预测的股票价格；C 为常数，与股票历史价格有关；X 为情感指数，β 为其系数参数，$AR(p)$ 是一种时序模型，p 自回归阶数，可以优化自相关系数。本系统中经过计算确定 $p=1$。

预测模型的整体研究框架图 3 所示。

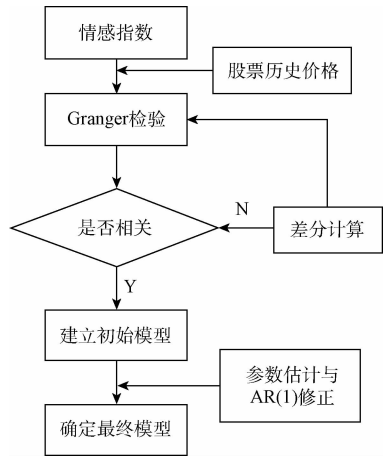

图 3　预测模型的整体研究框架

2 系统实现

2.1 数据获取以及模型构建

本文选取 2016 年 5 月 3 号到 7 月 26 号的 59 天的 200 只股票论坛股评以及股票价格数据,本方案以中国石化为例对数据进行处理并用 Eviews 软件对数据进行分析建立预测模型。

估计股票收盘价对常数、情感指数的回归方程,并用 AR(1)[15] 修正残差序列相关,拟合得到的模型如图 4 所示。

Dependent Variable: Y
Method: Least Squares
Date: 08/14/16 Time: 16:53
Sample (adjusted): 3 59
Included observations: 57 after adjustments
Convergence achieved after 5 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.739244	0.037919	124.9820	0.0000
X(-1)	0.004851	0.003243	6.495762	0.0000
AR(1)	0.853604	0.072558	11.76446	0.0000

R-squared	0.735437	Mean dependent var	4.750351
Adjusted R-squared	0.725639	S.D. dependent var	0.077367
S.E. of regression	0.040524	Akaike info criterion	-3.522636
Sum squared resid	0.088680	Schwarz criterion	-3.415107
Log likelihood	103.3951	Hannan-Quinn criter.	-3.480847
F-statistic	75.05515	Durbin-Watson stat	1.834922
Prob(F-statistic)	0.000000		

图 4 预测股票回归模型参数估计结果

模型估计结果为:

$$Y_t = 4.739244 + 0.004851 \times X_{t-1} + [AR(1) = 0.853604] \quad (3)$$

对该多元线性回归模型进行拟合优度检验、F 检验、t 检验。由图 4 可知,从拟合优度来看 $R^2 = 0.735\ 4$,调整 $R^2 = 0.725\ 6$,这两个值越高说明其拟合优度越好,自变量可以解释因变量 72.56%的变化。经查阅,在置信度 $\alpha = 0.05$ 情况下,自由度为(1,55)的 F 统计的临界值为 $F_{0.05}(1,55) = 4.016$,模型的 F 统计量的值为 75.05 远高于临界值,即拒绝原假设,认为前一天的情感值对当天的股票收盘价存在显著影响。T 检验在置信度 $\alpha = 0.05$ 情况下,自由度为 55,经查阅临界值 $T_{0.025}(55) = 2.004$,对应得到 T 统计量的绝对值为 6.495 762,其大于临界值,即通过显著性检验。综上,模型成立。

2.2 股票价格预测显示和分析

利用建立的预测模型,股票收盘价进行预测结果如图 5 所示(图 5 中以横坐标 60 表示),预测结果在 YF 文件中,预测结果走势图如图 5 所示。图中 S. E. 为预测标准差,图中的实线即为预测值,虚线为估计区间即预测值加减两个标准差的带状图,这两个标准差带在 95%的置信区间内,表示做预测时因变量的实际值有 95%的可能性落在置信区间内,通过比较其实际值均在给出的 95 的区间预测范围内。

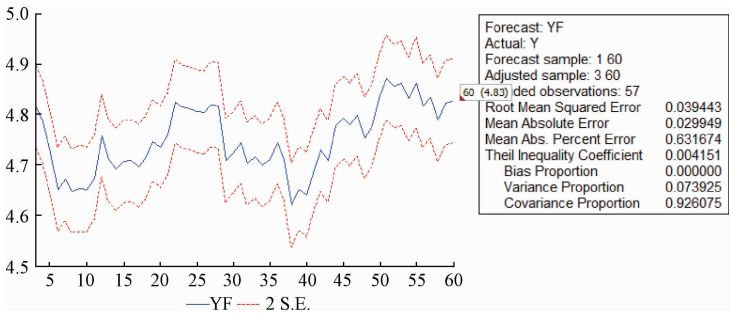

图 5 预测结果分析

图的右侧为预测效果评估,其中均方误差和平均绝对误差表示误差越小,该模型的预测能力越强;其值分别为 0.029 949 和 0.039 443 说明该模型预测能力处于理想状态;泰尔不等系数处于 0 和 1,值越小表示与真实值的拟合效果越好,其值为 0.004 151 说明预测值和实际值拟合效果很好。偏差比、方差比和协方差比总和为 1,偏差比表示预测均值和序列实际值的偏差程度,方差比表示预测方差与序列实际方差的偏离程度,协方差比衡量非系统误差的大小;如果预测结果好,则偏差比、方差比较小,协方差比应该较大;此模型三个指标的值分别为 0、0.073 925 和 0.926 075,很明显其表示最终的预测结果处于理想状态。

通过预测效果可以说明本文建立的预测模型的预测能力和预测结果都处于较好的状态,可以充分利用此模型借助情感指数对股票做出比较理想的预测。以中国石化为例,其实际值和预测值比较如图 6 所示,其残差值如图 7 所示。由图可说明此模型实际值和预测拟合值虽有细微差别,但误差均在允许范围内。

利用本文建立的预测模型对选取的 200 只股票进行分析预测,并对 57 天的股价走势进行统计,统计结果如表 1 所示(只呈现部分结果)。

神经网络是是众多神经网络中应用范围最广使用人数最多一个[16-17];SVM 是在统计学习理论的基础上发展起来的一种新的模式分类方法常用于模式识别、分类以及回归分析领域[19-20];因此分别用本文预测模型、SVM 模型和 BP 神经网络对以上数据集进行实验分别得出对 200 只股票预测的平均正确率,对比结果如表 1 所示。

图 6　实际值和拟合值比较

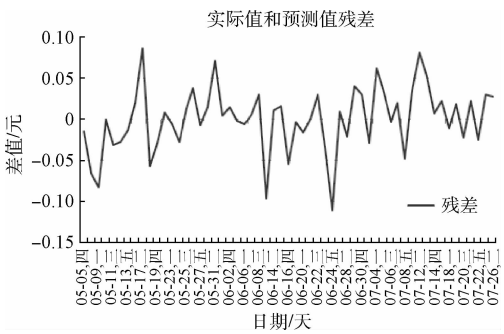

图 7　实际值和预测值的残差

表 1　股票预测正确率部分结果显示

股票名称	正确数	正确率/%	股票名称	正确数	正确率/%
中国石化	35	61.4	中信证券	37	64.9
大秦铁路	36	63.2	建设银行	38	66.7
光大银行	40	70.2	上海机场	40	70.2
国投电力	39	68.4	上汽集团	38	66.7
农业银行	41	71.9	长安汽车	43	75.3
贵州茅台	40	70.2	五粮液	41	71.9
川投能源	42	73.7	长江电力	40	70.2
王府井	43	75.3	江铃汽车	42	73.7
海南海药	39	68.4	同仁堂	42	73.7
甘肃电投	40	70.2	吉电股份	41	71.9

表 2　预测平均正确率

Model	BP 神经网络	SVM	本文模型
正确率/%	62.5	59.0	69.9

从表中可以看出，本文建立的预测模型的平均正确率为 69.9%，相对于 BP 神经网络和 SVM 模型分别有 7.4% 和 10.9%的提升。主要因为本文方法使用基于技术和情感分析混合方法，引入情感指数和多元回归分析建立预测模型并利用 AR(1) 对该模型进行修正残差序列相关。

3　结　　论

TA/SA 的应用对股票价格的预测的准确性有很大的提升。本文提出基于 TA/SA 组合方法建立股票预测模型的方案，实现了通过情感指数对股票价格的预测，实现人们对金融市场的进一步了解以及股价预测的需求。

情感约束的多元回归预测模型的建立为以后问题的深入研究做好了准备。利用金融论坛股评的情感指数分析股票仅仅是其中的一方面，而且利用多元回归模型其考虑的因素越多，模型拟合效果越好，预测也会更准确。所以下一步将引入更多解释变量例如政府政策、股票换手率以及金融门户网站的新闻和专家评论等，以建立拟合度更好的回归模型，提高股票价格预测的正确率。

参考文献

[1] 池丽旭，庄新田. 投资者情绪与股票收益波动溢出效应[J]. 系统管理学报，2009，18(4)：367-372.

[2] 罗海玲，郑根. 基于 R 语言股票市场收益的预测分析[J]. 福建电脑，2015(7)：74-76.

[3] CHEN Y J. Enhancement of stock market forecasting using a technical analysis-based approach[J]. Soft Computing，2016：1-23.

[4] ZHANG K，LI L，LI P，et al. Stock trend forecasting method based on sentiment analysis and system similarity model[C]. 2011 6th International Forum on Strategic Technology (IFOST)，IEEE，2011：890-894.

[5] MAKREHCHI M，SHAH S，LIAO W. Stock prediction using event-based sentiment analysis[C]. IEEE/WIC/ACM International Joint Conferences on Web Intelligence. IEEE Computer Society，2013：337-342.

[6] WU J L，SU C C，YU L C，et al. Stock price predication using combinational features from sentimental analysis of stock news and technical analysis of trading information[J]. International Proceedings of Economics Development & Research，2013，3(3)：239-253.

[7] LI X，XIE H，CHEN L，et al. News impact on stock price return via sentiment analysis[J]. Knowledge-Based Systems，2014，69(1)：14-23.

[8] HO K Y，WANG W. Predicting Stock Price Movements with News Sentiment：An Artificial Neural Network Approach[M]. Springer International Publishing：Artificial Neural Network Modelling，2016.

[9] NGUYEN T H，SHIRAI K，VELCIN J. Sentiment analysis on social media for stock movement prediction[J]. Expert Systems with Applications，2015，42(24)：9603-9611.

[10] 张对. 网络股评影响股市走势吗——基于股票情感

分析的视角[J]. 现代经济信息, 2015(1): 355-357.
[11] 张建华, 梁正友. 基于情感词抽取与 LDA 特征表示的情感分析方法[J]. 计算机与现代化, 2014(5): 79-83.
[12] 阳爱民, 林江豪, 周咏梅. 中文文本情感词典构建方法[J]. 计算机科学与探索, 2013, 7(11):1033-1039.
[13] 刘明, 李明莉. 线性回归模型统计检验方法体系构建[J]. 统计与决策, 2014(2):8-11.
[14] 李永立, 吴冲. 基于多变量的 Granger 因果检验方法[J]. 数理统计与管理, 2014, 33(1):50-58.
[15] 陈庆堂, 宋一然, 黄宜坚. 基于 AR 模型-分形维的磁流变减振器性能研究[J]. 仪器仪表学报, 2016, 37(12):2774-2781.
[16] 李小珉, 尹明. 基于遗传算法的 BP 神经网络电子系统状态预测方法研究[J]. 电子测量技术, 2016, 39(9):182-186.
[17] 丁硕, 巫庆辉, 常晓恒, 等. 基于灰色 BP 神经网络的实验材料供应预测[J]. 国外电子测量技术, 2016, 35(12):78-82.
[18] 黄秋萍, 周霞, 甘宇健, 等. SVM 与神经网络模型在股票预测中的应用研究[J]. 微型机与应用, 2015, 34(5):88-90.
[19] 汪济洲, 鲁昌华, 蒋薇薇. 一种新的基于混合粒子的粒化支持向量机算法[J]. 电子测量与仪器学报, 2015(4):591-597.
[20] 宋晓琳, 郑亚奇, 曹昊天, 等. 基于 HMM-SVM 的驾驶员换道意图辨识研究[J]. 电子测量与仪器学报, 2016, 30(1):58-65.

作者简介

代江波(通讯作者),硕士研究生,主要研究方向为文本事件抽取与分析。
E-mail:djbo2016@163.com

毛建华,博士后,副教授,主要研究方向为文本事件抽取与分析。
E-mail:mjh@shu.edu.cn

延丰,硕士研究生,主要研究方向为 web 文本数据挖掘。
E-mail:314627745@qq.com

刘学锋,博士后,教授,主要研究方向为遥感与空间信息处理。
E-mail:lxf02@shu.edu.cn

(上接第 47 页)

[8] 高云泽, 叶盛波, 张晓娟, 等。基于电磁感应和超宽带雷达的新型探测系统[J]. 电子测量技术, 2015, 38(9):134-140.
[9] 杨俊. 基于声学波动方程的偏移速度误差分析[D]. 成都:成都理工大学, 2006.
[10] 陈正宁,张金全. 一种雷达脉内特征仿真验证系统[J]. 国外电子测量技术,2016, 35(7):96-100.
[11] JIANG SH Y, GAO T CH, LIU X CH. The simulation and error analysis of raindrop size distribution obtained by micro rain radar[J]. Instrumentation, 2015(3):43-54.

作者简介

沃远,1991 年出生,硕士研究生,主要研究方向为探地雷达信号处理。
E-mail:woyuan14@mails.ucas.ac.cn