

基于多信息源的股价趋势预测

饶东宁¹ 邓福栋¹ 蒋志华²

(广东工业大学计算机学院 广州 510006)¹

(暨南大学信息科学技术学院计算机科学系 广州 510632)²

摘 要 股票价格及趋势预测是金融智能研究的热门话题。一直以来,各种各样的信息源被不断尝试用于股价预测,例如基本经济特征、技术指标、网络舆情、财务公告、财政新闻、金融研报等。然而,此类研究大多数只使用一种或两种信息源,使用3种及以上信息源的极为少见。信息源越多意味着能够提供更加丰富的信息内容和更多不同的信息层面。但是由于各种信息源的本质不同,其对股票市场的影响程度不同,因此将多种信息源融合起来进行股价预测并非易事。此外,多信息源也增加了维度灾难的风险。基于信息融合的目的,尝试同时利用基本经济特征、技术指标、网络舆情3种信息源来进行股价预测。具体做法:先对不同类型的信息源数据进行针对性的处理,使其形成统一的数据集,然后使用SVM分类器建立预测模型。实验结果表明,在选用线性核函数和考虑非交易日数据时,使用这3种信息源组合的预测模型的预测效果要比使用单一信息源或者两两组合的预测效果好。此外,在收集数据时发现,在非交易日(例如周末或停牌期)虽没有买卖但网络舆情剧增。因此,在实验数据中添加了非交易日的舆情情感数据,分类精准度有所提高。研究结果表明,基于多信息源融合的股价预测虽然困难,但是在适当地选择特征和针对性地进行数据预处理后会有较好的预测效果。

关键词 多信息源,股价趋势预测,SVM分类

中图法分类号 TP181

文献标识码 A

DOI 10.11896/j.issn.1002-137X.2017.10.036

Stock Price Movements Prediction Based on Multisources

RAO Dong-ning¹ DENG Fu-dong¹ JIANG Zhi-hua²

(School of Computer, Guangdong University of Technology, Guangzhou 510006, China)¹

(Department of Computer Science, School of Information Science and Technology, Jinan University, Guangzhou 510632, China)²

Abstract Predicting stock price movement is a hot topic in the financial intelligence field. So far, people have continuously attempted to use various data sources in the stock price prediction, such as fundamental economic features, technical indicators, Internet public opinions, financial announcements, financial news, financial research reports and so on. However, most of the previous studies use only one or two distinct data sources to build prediction models. Few of them take advantage of three or more sources simultaneously. Undoubtedly, if more sources are provided, people can extract richer information content and consider more information levels. But, since the natures of various sources are distinct, and they have different effects on the stock market, it is not easy to converge several sources in predicting stock price. In addition, multisources naturally increase the risk of suffering the curse of dimensionality. Based on the idea of information fusion, this paper attempted to use three distinct sources to predict the stock price movement. The three sources are fundamental economic features, technical indicators and Internet public opinions. Our method firstly collects various source data, then implements the specific data preprocessing to form a unified data set, and finally uses the SVM classifier to build prediction models. Experimental results show that the performance of prediction model based on the three sources is better than those which use a single source, or sources in pairs, when the linear core function for the SVM classifier is chosen and the data in the non-trading days are added. Besides, when collecting data, we found that the number of Internet public opinions rose sharply, although there were no transactions in the non-trading days (for example, weekends or the suspension period). Therefore, we added more text sentiment data showing the public opinions in the non-trading days and found that the prediction accuracy is improved. The study in this paper shows that although it is difficult to integrate multisources in the stock prediction, it is possible to produce a good predictor after the appropriate feature selection and the specific data preprocessing.

Keywords Multisources, Stock price movement prediction, SVM classification

到稿日期:2016-09-20 返修日期:2017-03-17 本文受广东省自然科学基金(2016A030313084, 2016A030313700, 2014A030313374), 中央高校基本科研业务费专项资金资助项目(21615438), 广东省科技计划项目(2015B010128007)资助。

饶东宁(1977—),男,博士,副教授,主要研究方向为金融智能、智能规划, E-mail: raodn@gdut.edu.cn; 邓福栋(1991—),男,硕士生,主要研究方向为金融智能; 蒋志华(1978—),女,博士,副教授,主要研究方向为金融智能、智能规划, E-mail: tjiazhzh@jnu.edu.cn(通信作者)。

1 引言

1972 年,尤金·法玛等提出了有效市场假说理论(Efficient Market Hypothesis, EMH)^[1]。根据 EMH,如果市场是有效的,那么对证券价格和技术指标建立预测模型是没有意义的。中国学者对中国股市的大量研究表明^[2-4],中国股市未达到弱式有效,中国证券市场作为新兴的市场,其对信息披露的监管较弱,投资者和上市公司之间存在信息不对称且短时间内并不能被及时消除的问题。这使得中国股票价格波动具有长期记忆性,股市的涨跌具有一定的规律,存在可预测的成分,为我国的股市预测提供了可能。因此预测股价让众多金融专家学者以及计算机领域的学者产生了极大兴趣。

然而,自 20 世纪 80 年代以来,EMH 受到了理论基础和经验检验的挑战。在理论基础方面,3 个假设(即弱式有效市场假说、半强式有效市场假说、强式有效市场假说)都难以实现。在经验检验方面,过度反应现象、价格冲量、小公司效应等与某些假设存在矛盾。针对 EMH 存在的问题,学者们对 EMH 进行了补充,补充分为两类:一类是沿用传统经济学的方法,将 EMH 未能解释的部分描述为金融或者经济特征,把金融市场的变化描述为多个金融和经济特征共同作用的结果^[5-7];另一类吸取了行为金融(Behavioral Finance, BF)理论的观点,提出投资者决策时的心理特征假设来研究投资者的实际投资决策行为^[8-10]。

对 EMH 的补充导致各种风格迥异的信息源涌现。常用的信息源有基本经济特征^[7]、技术指标^[11-12]、网络舆情^[13-14]、财务公告^[15]、财经新闻^[10,16-17]、金融研报^[18-20]等。首先,基本经济特征反映了股票市场的当前价格信息或公司企业经营状况,例如成交量、开盘价、收盘价、市盈率、市净率、收益率等。使用基本经济特征的预测方法称为基本分析法^[7],能反映出股票的内在价值,适合长期预测。其次,技术指标根据证券市场的历史数据,通过数学公式计算得到^[11]。每种技术指标反映着市场某一方面的信息,可分为趋势型指标、超买超卖型指标、能量型指标、停损型指标等多种类型。使用技术指标的预测方法称为技术分析法,其注重市场的行为,适合短期预测。此外,网络舆情、财务公告、财经新闻、金融研报等这些非传统的金融经济特征主要反映了投资者的行为。网络舆情的最鲜明特征是其具有情感倾向,对股票波动具有正效应、负效应或超效应^[14]。财务公告对股票市场的影响取决于公告的用途,例如停牌公告和澄清公告等^[15]。公告中的一些关键词语可能会预示着股价的涨跌,例如,“重组”、“并购”、“重大项目终止”等。财经新闻的内容往往体现在 3 个方面,分别是兼并收购、盈利能力和再融资^[16],这些信息对股票市场具有立竿见影的影响。金融研报体现了投资者的推荐行为^[18],具有强烈推荐倾向的股票价格可能会上涨。综上可知,不同的信息源提供了多元化的数据资源,影响着股票市场的不同层面。

上述研究中的大多数使用某一种信息源进行股价预测,取得了一定的预测效果。但是股票市场错综复杂、灵活多变,考虑到经济、政治和社会等因素,没有一种单独的信息源能完全反映股票市场的信息,因此学者们开始考虑基于多信源的预测,即信源融合(Source Integration),然而这并非易事。其困难在于:首先,需要解决维度灾难。信源增多意味着特征维

度会大量增加,需适当进行降维处理。其次,不同信源的数据格式的差别可能较大,需进行归一化处理以统一到同一数据集。再次,信源并非越多越好,有些信源组合后反而会降低预测效果。最后,不同信源对股票市场的影响程度不同,需平衡其中的权重。已有的基于多个信息源的股票预测研究工作往往选用两个信息源进行组合,其中以财经新闻加上传统经济特征的组合居多^[10,21-24];同时使用 3 种及以上的信息源进行预测的研究工作相当少^[25],原因是信源融合一方面提供了信息的多元化,但另一方面也增加了处理的复杂性。

本文工作是基于 3 种信源组合来进行股价预测的一种尝试。选取了基本经济特征、技术指标、网络舆情 3 种不同的信息源,基于 SVM 分类器建立股价预测模型。其中,通过网络舆情产生的情感文本是文本型数据源,而技术指标和传统经济特征是数值型数据。首先,经过实验测试,选择适当的特征维度,避免维度灾难。其次,对网络舆情建立针对金融领域的情感词典,计算其情感值。然后,对各种数据进行归一化处理以统一到同一数据集。再次,对 3 种信源的所有组合分别进行测试,实验结果表明,在某些情况下同时使用 3 种信源的预测模型的预测效果比使用单一信源或者两两组合的预测效果更好。最后,考虑到交易日和非交易日对不同信源的影响(比如在周末或者停牌期间,网络舆情的情感倾向会比较强烈),增加了非交易日的舆情情感倾向数据,实验结果表明,增加非交易日的数据后提高了预测精度。

本文第 2 节介绍基于各种信息源的股价预测的研究现状;第 3 节陈述基于多信源的股价预测问题,并介绍 SVM 的分类原理;第 4 节介绍针对不同信息源特征的数据处理方法,以及多信源的股价预测流程;第 5 节是实验设计及分析;最后总结全文并展望未来工作。

2 研究背景

针对本文所选取的信息源,本节将逐一介绍这些信息源的特征、基于其的股价预测成果以及基于多信源的股价预测的研究现状。

2.1 基本经济特征

基本经济特征是指直接反映股票价格的数据或者反映公司企业经营状况的信息。最常用的反映股票价格的数据包括成交量、开盘价、收盘价、最高价、最低价等,这些也是计算技术指标的基础数据;最常用的反映公司企业经营状况的指标包括市盈率、市净率、收益率、利润总额等。使用基本经济指标的股价预测方法称为基本分析法,它通过研究影响证券在市场上的供求关系的基本因素,依靠经济学、金融学、财务管理学和投资学等基本原理,对决定证券价值及价格的基本要素进行分析,评估证券的内在价值,判断证券的合理价位并提出相应投资决策的意见。这种方法具有较强的经济学含义,但是由于目前金融市场上炒作事件频发,股票价格往往与其基本价值背离,使得基本分析方法在股票预测的应用中并未得到广泛认可。

在传统的股票预测模型中,最早的成果是 Fama 的三因子模型^[26]。该模型认为,一个投资组合(包括单个股票)的超额回报率可由它对 3 个因子的暴露来解释。这 3 个因子分别为市场资产组合($R_m - R_f$)、市值因子(SMB)、账面市值比因子

(HML)。由于三因子模型中还存在着许多未解释的部分,因此许多学者对其进行了改进。Carhart 在三因子模型的基础上引入了动量因子,构造了四因子模型^[5],它对于基金绩效的解释能力较前者有了很大的提高。四因子模型把基金收益描述为在市场因素(MKT)、规模因素(SMB)、价值因素(HML)与动量因素(UMD)共同作用下的结果。Novy-Marx 通过大量的实验验证发现,相对于 HML 盈利能力在描述股票横截面收益上具有更好的解释性^[6]。受此启发,Fama 等把盈利能力和投资因素添加到三因子模型上,得到一个新的五因子模型^[7]。基本分析法能够很好地反映股票的内在价值,因而在较长的周期中,能够对股票的走势起到很好的预测作用;但基本分析法的缺点是价格变动时反应较为迟钝,不适合用于短期预测。

2.2 技术指标

技术指标反映市场某一方面的信息,它基于基本经济数据通过数学公式计算得来。使用技术指标的预测方法称为技术分析法,该方法利用证券市场的历史数据,通过图表、公式等寻求股票价格变化的规律来进行预测。技术分析法忽略了市场的因素和经济定律,只注重市场的行为,其包含 3 大假设:1)市场行为涵盖一切信息,直接分析市场行为更为便捷和可行;2)证券价格沿趋势运动,其运动方向由供求关系决定;3)价格的运行方式往往会重复历史。技术分析包括多种方法,大致分为 K 线理论、波浪理论、技术指标、循环周期、切线理论、形态理论等。目前市场上的技术指标数不胜数,最常用的是相对强弱指标(RSI)、简单移动平均(MA)、指数平滑异同平均(MACD)、随机指标(KDJ)等。

Lee^[12]在 NASDAQ 指数上进行实验,构建了一个 SVM 模型来预测股市变化。他提出了一种新的特征选择方法 F-SSFS,先通过 F 得分进行特征筛选,然后基于 SVM 的前向搜索特征选择方法(SSFS),对输入变量进行筛选。Patel 等^[11]也使用技术指标对印度的股票价格进行预测。他们选取了 10 个技术指标,同时结合了 4 个股票价格特征,分别在两个股票和两个股指上进行实验;使用 4 种不同类型的机器学习方法进行全面的对比,包括人工神经网络(Artificial Neural Network, ANN)、支持向量机(Support Vector Machine, SVM)、随机森林(Random Forest, RF)、朴素贝叶斯分类(Naive Bayesian Classification)。实验结果表明,RF 具有最好的分类性能,其次是以高斯核作为核函数的 SVM。此外,Zhai 等^[21]结合公司新闻和技术指标,利用 SVM 分类器对公司股票价格进行预测;其使用 7 个技术指标将新闻分为公告新闻和市场新闻两类,实验结果显示组合的预测效果较好。

2.3 情感文本

情感分析(也称观点挖掘)的任务是根据输入文本的不同主题,发现、提取和分类观点或态度。其预处理过程包括:标记、移除停止词、词性标注以及特征的提取和表示。由于投资者的行为表现在他们接收到的信息和对待这些信息的态度上,因此许多学者把目光转向于网络舆情、财务公告、财政新闻、金融研报等,试图探索出它们对股票市场的影响。

情感分析技术已广泛用于股票预测,特别是在博文方面。Bollen 等^[13]通过 twitter 对金融数据进行情感分析来研究市场预测,使用两个心情追踪的工具,通过 6 个心情维度来计算

情感值。徐琳^[14]研究了网络舆情(主要是微博)对我国股票市场的影响,提出了网络舆情对股票波动可能具有正效应、负效应以及超效应,并通过实验验证了这 3 种效应。

股票公司的公告也会影响投资者的情绪,进而影响股价。例如,停牌期间的公告对股价的影响可能会反映在复牌后的股价上。公告中的一些关键词,例如“重组”、“并购”等有可能造成股价的上涨,而“维持股价稳定”、“重大项目终止”等有可能造成股价的下跌。王莉莉^[15]研究了澄清公告与股价波动之间的关系,提出了 4 个假设:1)市场传闻会对股票价格产生影响,且会因传闻性质的好坏产生不同方向的影响;2)针对市场传闻的澄清公告会对股票价格产生影响,且与澄清公告的回应方式有关;3)发布澄清公告后,若投资者对其表现为反应不足,则澄清公告不能在短期内使股价恢复传闻前的水平;4)澄清公告发布的及时性将影响投资者的市场反应,澄清公告发布得越及时,澄清效果越好。王莉莉通过实验验证了前 3 个假设。

财经新闻反映股票市场的最新事件,对其进行情感分析是比较适宜的。新闻数据多数情况下是结合其他金融数据进行预测^[10,21-23],而且一般都进行分类。杨娟^[16]将收集到的新闻分为三大类,分别是兼并收购、盈利能力、再融资。Shynkevich 等^[17]也对新闻进行分类,按照其与股票市场的相关性设定权值,并且对每种类别指定一个核函数;实验结果表明,针对不同类别的设置增加了预测精准度。

金融研报往往反映投资者的推荐行为。当某个研报“强烈推荐”某只股票时,该股票的价格很可能上涨;当某个研报“减持”某只股票时,该股票的价格很可能下跌。Duan 等^[18]提出基于分析者的推荐行为的后验概率模型来预测股票的回报;进一步地,他们结合股票交易信息和投资者评价进行股票收益预测^[20];特别地,使用规则提取技术建立预测系统,解决了大多数预测模型不具有解释性的问题。此外,Newman 等^[19]也研究了投资者推荐行为,与大多数相关研究不同,他们使用买方数据进行分析。

2.4 基于多信源的股价预测

多信息源的股票预测指的是同时利用多种信息源进行预测。由于不存在能反映股票市场的所有信息的单一信息源,因此信息源的综合利用可能会提高精准度。但是,信源并非越多越好,信源数量的增多会使得信源之间的相互影响变得复杂,有时甚至会出现相互干扰的现象,得不偿失。目前的研究多数是利用两种信息源来进行股价预测,搭配的模式往往为基本的股票价格数据加上新闻事件或研报推荐等文本信息。少量研究采用 3 种信息源进行预测^[25]。实验结果表明多源结合比单源的预测效果更好。

在采用两种以上信息源的为数不多的研究工作中,Li 等^[25]使用张量(tensor)来构建多源的信息空间,用高价的张量回归算法来产生预测模型。张量表示有别于传统表示方法,它不是把多源数据简单地汇集成一个超大矢量,而是构建一个具有多个数据层面的整体,通过张量分解和重构来区分不同源的影响。Li 等采用 3 种信息源(公司股票价格、新闻事件和情感文本)构建了一个性能良好的交易决策平台。另外,在较早期的研究工作中,Wu 等^[24]声称使用了 4 种信息源:市场状态、当前价格、历史价格、新闻。其中,当前价格和

历史价格属于基本经济特征,而市场状态指的是股票价格涨跌的趋势。他们根据新闻和股票价格来识别隐藏的市场状态,所使用的信源实质上包括两种:基本经济数据和金融新闻文本。

采用两种信息源来进行预测的研究工作较多。如前所述,金融新闻往往是一种被结合的信息源。Zhai 等^[21]结合公司新闻和技术指标,利用 SVM 分类器对公司股票价格进行预测;其使用 7 个技术指标将新闻分为公告新闻和市场新闻两类。Li 等^[10,22]使用市场新闻和股票价格两种信源来构建预测模型,他们使用 ELM (Extreme Learning Machine) 方法来挖掘隐藏数据,预测精准度与 SVM 类似,但学习速度更快。Gunduz 等^[23]也使用新闻和股票价格来进行股价预测,但与众不同的是,他们收集的是土耳其新闻,这需要开发土耳其文本挖掘技术;同时他们还提出了基于互斥信息的特征选择方法。其他信源结合方式也都各具特色。Duan 等^[20]结合股票交易信息和投资者评价来进行股票收益预测,其使用规则提取技术建立预测系统,大大增加了预测模型的可解释性。Tsai 等^[27]选取了 19 个金融特征和 11 个经济指标作为输入变量,构造了一个集成分类预测模型。Lam^[28]使用 16 个金融变量和 11 个宏观经济变量,构建了一个基于后向传播神经网络的预测模型。

3 问题陈述

3.1 多信源的股价预测问题

目前对股票预测的研究存在不足,尤其是信源融合的问题。首先,传统经济方法只考虑了传统的经济金融特征对股价的影响,这些特征一般都是数值型。其次,基于机器学习技术的各种预测方法虽然扩展了影响股价的特征,并且考虑了一些文本型数据对股价的影响,但是并没有考虑多个文本型特征共同作用时股价的变化。基于这些不足,我们在考虑影响股价的因素时,既考虑了传统的经济金融数值型的变量,也考虑了各种类型的文本型数据变量。

本文选取基本经济特征、技术指标、网络舆情等 3 种不同的信息源,基于 SVM 分类器来构造一个多信源的股价预测模型。该预测模型使用 3 种信源融合的数据,其经过处理后形成统一的数据集,然后输出表示股价涨跌振幅的判断。它可以使投资者考虑更全面的信息,进行更好的投资组合选择。已有不少相关工作^[11-12,21-22]表明 SVM 分类是性能较好的分类方法,因此本文选择 SVM 分类器来构建预测模型。

3.2 SVM 分类原理

SVM 作为一类新型机器学习方法,是对神经网络的发展,能较好地解决小样本、非线性、高维数等神经网络不能解决的问题,克服了传统方法的诸多缺点,并且具有更高的精确度。

SVM 是从线性可分情况下的最优分类发展而来,其本质在于寻找一个把训练空间 R^d 分成两部分的最优线性分类面 $w \cdot x + b = 0$,使得两类不仅能够被分开,而且分类间隔最大,最终得到一个决策函数。对于线性可训练空间 $X_i \in R^d$,在 d 维特征空间上通过最大化几何间隔得到 SVM 优化模型^[11]:

$$\begin{aligned} \min_{w_i, b_i, \xi_i} & \frac{1}{2} \|w\|_2^2 + c \sum_{i=1}^n \xi_i \\ \text{s. t. } & y_i (w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \quad (1)$$

其中, c 为代价参数, ξ_i 为松弛因子。求解式(1),对于线性不可分的情况,将原特征向量映射到高维空间,得到决策函数:

$$f(x) = \text{sign}(\sum a_i y_i \langle \varphi(x_i), \varphi(x_j) \rangle + b) \quad (2)$$

引入核函数,决策函数转化为:

$$f(x) = \text{sign}(\sum a_i y_i k(x_i, x) + b) \quad (3)$$

其中, $k(x_i, x) = \langle \varphi(x_i), \varphi(x_j) \rangle$ 称为核函数。本文选择径向基核函数(Radial Basis Function, RBF):

$$k(x_i, x_j) = \exp\left[-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right] \quad (4)$$

和线性核函数(Linear Function):

$$k(x_i, x_j) = x_i \times x_j \quad (5)$$

作为实验的备选核函数,以找出合适的分类预测模型。其中,参数 σ^2 为高斯函数的方差; σ 控制了函数的径向作用范围, σ 过小则容易出现“过拟合”, σ 过大则容易出现“欠拟合”。

3.3 精准度

我们可以通过计算各项指标来评判模型预测结果。

(1) 平均绝对误差

$$MAE = \frac{\sum_{i=1}^n |Y_i - y_i|}{n} \quad (6)$$

(2) 均方误差

$$MSE = \frac{\sum_{i=1}^n (Y_i - y_i)^2}{n} \quad (7)$$

(3) 平均绝对百分比误差

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{Y_i - y_i}{Y_i} \right| \quad (8)$$

4 方法

4.1 整体流程

多信源的股价预测的整体流程如图 1 所示。该过程分为 4 个阶段:数据收集、数据预处理、核函数选择、性能评估。其中,数据预处理是本文工作的关键。

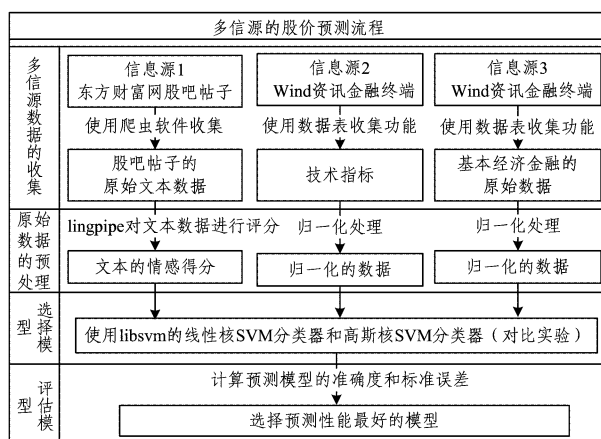


图 1 多信源的股价预测的整体流程

首先,进行多信源的数据收集。使用爬虫软件从东方财富股吧收集创业板股票所在的股吧在某一时期内发表的帖子,形成网络舆情文本。使用 wind 金融数据终端¹⁾收集创业板股票同期的基本经济金融数据和各项技术指标。

然后,进行数据的预处理。先去所有空白数据。对于

¹⁾ <http://www.wind.com.cn>

数值型数据,为了消除不同特征的取值范围对股价预测的影响差异,需进行归一化处理。对于文本型数据,由于其来自网络舆情,有明显的情感倾向,因此需计算其情感值。采用基于词典的文本评分方法把文本型数据转换为适合预测模型的数据。可借助开源的多语言处理工具 lingpipe¹⁾ 的中文分词和情感分析功能,针对不同类型的文本建立不同的词典,并根据词典完成文本的评分。

接着,确定预测模型。已有的实验结果^[11-12,21-22]表明,SVM在股价预测问题上具有较好的分类性能。本文采用 libsvm²⁾ 来构造预测模型,libsvm是台湾大学林智仁教授等开发设计的一个简单、易用且快速有效的 SVM 模式识别与回归的软件包,它提供了很多默认参数,利用这些默认参数可以解决很多问题。本文使用线性核函数和高斯核函数来分别建立预测模型,进行实验对比。

最后,评价预测模型的性能。计算两种分类器的准确度,根据不同的股票类型选择适合的预测模型。

4.2 针对各种信源的数据预处理

多信源的股价预测需要针对各种信源的特点对数据进行处理才能达到融合的目的。为此,本节介绍实验所涉及的各种信源数据的预处理过程。总体来说,对于数值型数据,由于各种特征的数据取值差异很大,为了消除取值范围对影响程度的差异,需进行归一化处理。对于文本型数据,由于其来自网络舆情,有明显的情感倾向,需计算其情感值。

对于数值型数据,归一化的好处在于:一方面,能够提升 SVM 对数据的学习速度;另一方面,能避免某些特征由于取值范围过大,对股价的影响程度覆盖了其他特征。尤其对于各式各样的技术指标,它们是通过不同的数学公式计算得出,在数值上的变化范围不相同,还可能处在不同的数量级上。因此,使用原始数据进行输入可能会使得模型在进行学习和分析时产生较大的偏差。故需要对原始数据进行归一化处理。

常用的归一化方法有 3 种:最小最大值法、对数法、统计法。本文采用最小最大值法,它的计算公式为:

$$X_i^* = (y_{\max} - y_{\min}) \times \frac{X_i - x_{\min}}{x_{\max} - x_{\min}} + y_{\min} \quad (9)$$

其中, $x_{\min}$ 与 $x_{\max}$ 为原始数据集的最小值和最大值, $y_{\min}$ 与 $y_{\max}$ 为归一后数据集的最小值和最大值。考虑到某些特征有负值的情况,把原始数据的归一化区间设置为 $[-1, 1]$ 。

对于网络舆情(帖子),我们主要考虑了帖子长度和情感值两个特征。为了表示这两个特征如何影响股价预测,定义情感倾向比值为帖子情感值与帖子长度的比值,因此,情感值越强烈且长度越短的帖子被认为是越有价值的帖子。而某只股票在某天的舆情情感倾向值则为当天所有关于它的帖子的情感倾向的比值之和。当然,计算得到所有的情感倾向值后也需对其进行归一化处理。

情感值的计算涉及两个关键技术:中文分词以及情感词典(或极性词典)的构造。前者可以使用通用工具,后者需要

对已有词典进行金融情感词的扩充,这也是本文的主要贡献之一。对于分词问题,我们采用 n-gram 算法^[29] 进行中文分词和匹配,同时采用统计语言模型。n-gram 算法通过构建字串的等价类来减少概率计算量,即当两个字串的最近 $N-1$ 个词(或字)相同时,映射两个字串到同一个等价类。对于情感值问题,流行的解决方法是 TF-IDF(Term Frequency-Inverse Document Frequency)算法^[30]。TF-IDF 是一种统计方法,用以评估一个字词对于一个文件集的重要程度。字词的重要性随着它在文件中的出现次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。因此,如果某个词或短语在一篇文章中出现的频率 TF 高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来进行分类。TF-IDF 倾向于过滤掉常见的词语,保留重要的词语。

由于已有的情感词典 NTUSD 并没有针对金融领域的网络舆情构建情感词,因此我们需要扩展 NTUSD 词典作为股吧帖子的情感词典。具体扩展内容如下:1)由于股吧帖子多数是由网友发表的,言语一般比较随意,一些词在帖子中具有明显的感情倾向,但在正规的文本中不具备感情倾向(例如“顶”,在帖子中具有明显的支持和赞同情感,但在正式的文章中并不具备任何感情倾向),因此要把这些特殊词加入 NT-USD 中。2)一些股票术语也带有作者对股价变化的明显感情倾向,如“飘红”表明作者对股票的上涨充满信心,“飘绿”表明作者对股票的上涨不抱信心,这一部分词也要扩展到 NT-USD 中。然后,根据词典中每个词对股价变化方向的影响程度,给每个词赋予相应权值。

NTUSD 的部分扩展内容如表 1 所列,一部分是看跌的情感词以及权值,另一部分是看涨的情感词以及权值。金融情感词的权值在理论上也可以通过 TF-IDF 算法来确定,但是由于目前还没有公认的金融情感词语料库,因此本文通过实验及经验来设置金融情感的权值。

表 1 补充到 NTUSD 的金融情感词(部分)

词	权值	词	权值	词	权值	词	权值
超买	-10	超卖	10	补跌	-10	涨跌	10
超涨	-10	超跌	10	大阴	-15	大阳	15
逃顶	-10	抄底	10	中阴	-10	中阳	10
空头	-12	多头	12	小阴	-5	小阳	5
亏损	-8	盈利	8	利空	-10	利好	10

5 实验

为了验证多信源对股价预测的影响,我们选取了基本经济特征、技术指标和网络舆情等 3 个信源来构建数据集。实验假设如下:

假设 1 以基本面、技术面和舆情情感倾向特征作为输入,构造出来的分类器的分类效果最好;

假设 2 增加非交易日的舆情情感倾向特征数据可以提高分类器的分类效果。

如果假设 1 成立,则表明在本文设计的实验中使用 3 种

¹⁾ <http://alias-i.com/lingpipe>

²⁾ <https://www.csie.ntu.edu.tw/~cjlin/libsvm>

信源组合的预测模型的预测效果比使用单一信源或者两两组合的预测效果更好,由此可见多信源融合对于股价预测问题的积极影响。假设 2 的设置是由于在实验中我们发现在非交易日(例如停牌期间或者周末)基本面数据和技术面数据没有变化,但是网络舆情的情感倾向特别强烈,因此将其作为额外数据补充来考查其对股价预测效果的影响。如果假设 2 成立,则表明周末发出的帖子对股价的影响会在下一周的股价上得到体现,或者停牌期间发出的帖子对股价的影响会在复牌后的股价上得到体现。

5.1 数据收集

本文选取 39 只创业板的股票作为实验的样本,收集 2015 年 7 月到 2015 年 12 月的数据作为实验数据。根据万得金融数据终端,最终筛选出 35 个特征(如表 2 所列),包括反映公司企业经营状况的基本面特征,从 K 线图得出影响股价变动的技术面特征,以及反映股民买卖股票倾向的网络舆情特征。具体来说,经济数据特征和技术指标特征都直接来源于万得金融数据终端。首先,尽可能从万得金融数据终端获取更多的经济数据特征。其次,在这些特征中,某些以季度作为周期计算的经济特征没有有效地提高分类效果,因此被忽略掉。

表 2 实验中各信息源的特征集(部分)

基本面特征	技术面特征	情感面特征
市盈率 PE(TTM)	RSI 相对强弱指标	帖子的长度(length)
市净率 PB(LF)	WR 威廉指标	情感值(score)
市销率 PS(TTM)	MA 简单移动平均	
净资产收益率 ROE(平均)	换手率	
总资产净利率 ROA	MACD 指数平滑	
	异同平均(DIFF)	
销售毛利率	MACD 指数平滑	
	异同平均(DEA)	
销售净利率	MACD 指数平滑	
	异同平均(MACD)	
财务费用/营业总收入	KDJ 随机指标(K)	
利润总额(同比增长率)	KDJ 随机指标(D)	
速动比率	KDJ 随机指标(J)	
资产负债率	BOLL 布林带(UPPER)	
流动负债/负债合计	BOLL 布林带(MID)	
净利润(同比增长率)	BOLL 布林带(LOWER)	
营业利润(同比增长率)	DMA 平均线差(DDD)	
	DMA 平均线差(AMA)	
存货周转率	CCI 顺势指标	
	BIAS 乖离率	
	VR 成交量比率	

如表 2 所列,最后选取的特征量为 35 个,即情感文本特征 2 个,技术指标特征 18 个,基本经济数据特征 15 个。其中,情感文本只考虑帖子长度和帖子情感值两个特征。由于

SVM 存在维度灾难的问题,在预测问题中设计了含有 100 多个特征及 10000 多条数据的问题实例,用 SVM 进行测试,发现能够在可接受的时间间隔内得出成果。因此,对于本文的实验,学习问题的维度灾难不严重。

5.2 实验设计

根据股票的涨跌幅度,我们将股票的涨跌分为 4 类,这是一种较为精细的分类方法。为了后文比较,我们将股票分为两类(即只有涨跌)。首先,四分类的定义如下:

$$class_{four} = \begin{cases} 2, & ratio \geq 5\% \\ 1, & 0 \leq ratio < 5\% \\ -1, & -5\% \leq ratio < 0 \\ -2, & ratio < -5\% \end{cases} \quad (10)$$

其中, $ratio$ 表示某只股票的涨跌率, $class_{four}$ 表示四分类的类别。当下跌超过 5% 则为 -2 类,若下跌在 0~5% 之间则是 -1 类,若上涨在 0~5% 之间则是 1 类,若上涨超过 5% 则是 2 类。

由于预测股票价格具有时序性,将收集到的 39 只股票在有效交易日内的数据的前 80% 的交易日数据作为训练集,后 20% 的交易日数据作为测试集。以 300001 这只股票为例,有效交易日为 126 天,因此前 101 天的交易日数据作为训练集,后 25 天的数据作为测试集。在实验中还采用了 n -fold 交叉验证法,本文取 $n=10$ 。具体做法:将训练集平均分为 10 等份,依次取 1 份作为测试子集,其余 9 份作为训练子集,来构造分类器并计算分类的准确率,最后计算出 10 个分类器的准确率的平均值作为最后的准确率。

为了验证以上两个假设,我们设计了对比实验。对于假设 1,用 SVM 分类器测试 7 种特征组合的输入模式,分别是:只用基本经济特征作为输入,只用技术指标特征作为输入,只用舆情情感倾向特征作为输入,同时使用基本经济特征和技术指标特征作为输入,同时使用基本经济特征和舆情情感倾向特征作为输入,同时使用技术指标特征和舆情情感倾向特征作为输入,同时使用基本经济特征、技术指标以及舆情情感倾向特征作为输入。对于假设 2,在原有数据集中增加了非交易日的舆情情感数据,同样测试 7 种特征组合的输入模式。此外,关于 SVM 核函数的选取,采用线性核函数和高斯核函数分别建立预测模型来进行对比实验,从而分析不同股票适宜选取的核函数以及特征组合。为了方便表示,后文的图表中特征组合均用简写的符号名称来表示。符号名称的说明如表 3 所列,简单来说,fun 表示基本面特征,tech 表示技术指标特征,post 表示舆情情感倾向特征,Union 表示特征组合,add 表示增加了非交易日数据的特征组合。

表 3 7 类特征组合输入在图 2—图 9 中的符号表示

不考虑非交易日		考虑非交易日	
图示名称	特征组合	图示名称	特征组合
Union-fun	经济特征	Union-fun-add	经济特征
Union-tech	技术指标	Union-tech-add	技术指标
Union-post	舆情情感	Union-post-add	舆情情感
Union-fun-tech	经济特征+技术指标	Union-fun-tech-add	经济特征+技术指标
Union-fun-post	经济特征+舆情情感	Union-fun-post-add	经济特征+舆情情感
Union-tech-post	技术指标+舆情情感	Union-tech-post-add	技术指标+舆情情感
Union-full	经济特征+技术指标+舆情情感	Union-full-add	经济特征+技术指标+舆情情感

5.3 实验结果与分析

由于可以选择核函数(线性核或高斯核)和训练集(是否考虑非交易日),因此实验中产生了不同类型的预测模型。为了比较它们的分类效果,我们定义以下评价函数:

$$evaluation_{type, flag} = (verify_set_{type, flag} + train_set_{type, flag}) / 2 \quad (11)$$

$$verify_set_{type, flag} = verify_set_average_accuracy_{type, flag} - verify_set_stdev_accuracy_{type, flag} \quad (12)$$

$$train_set_{type, flag} = train_set_average_accuracy_{type, flag} - train_set_stdev_accuracy_{type, flag} \quad (13)$$

其中, $type$ 表示核函数的类型,若取值为 linear,则表示线性核函数;若取值为 rbf,则表示高斯核函数。 $flag$ 表示是否考

虑非交易日数据的影响,若取值为 add,则表示考虑非交易日数据的影响;若取值为 noadd,则表示不考虑非交易日数据的影响。评价函数 $evaluation_{type, flag}$ 由两部分组成,一部分是在测试集上的分类准确率 $verify_set_{type, flag}$ (测试集由后 20% 的交易日数据构成),另一部分是在训练集上由交叉验证法($n=10$)得到的准确率 $train_set_{type, flag}$,每种准确率由平均分类准确率 $average_accuracy$ 和标准差 $stdev_accuracy$ 共同决定。 $evaluation$ 越大,表明对应输入的分类准确率越高,波动越小,分类效果越好。实验结果见表 4—表 5 以及图 2—图 9。表 4 和表 5 列出了不同配置下的平均分类准确率和标准差。图 2—图 9 展示了根据式(12)和式(13)得到的各种准确率。

表 4 在测试集上的平均准确率和标准差

	不考虑非交易日				考虑非交易日			
	线性核		高斯核		线性核		高斯核	
	准确率	标准差	准确率	标准差	准确率	标准差	准确率	标准差
经济特征	0.419341	0.170898	0.399774	0.149397	0.419341	0.170898	0.399773	0.149397
技术指标	0.477050	0.170376	0.405734	0.120959	0.477050	0.170376	0.405734	0.120959
舆情情感	0.400288	0.137962	0.402087	0.156546	0.403760	0.144763	0.403249	0.156823
经济特征+技术指标	0.507980	0.142991	0.408040	0.125641	0.507980	0.142991	0.408040	0.125641
经济特征+舆情情感	0.412438	0.142108	0.417595	0.152790	0.417773	0.159869	0.420658	0.153404
技术指标+舆情情感	0.481260	0.165020	0.410559	0.121036	0.485249	0.172122	0.408114	0.120446
经济特征+技术指标+舆情情感	0.491259	0.144329	0.407861	0.131376	0.514136	0.155823	0.414454	0.126611

表 5 在训练集上交叉验证的平均准确率和标准差

	不考虑非交易日				考虑非交易日			
	线性核		高斯核		线性核		高斯核	
	准确率	标准差	准确率	标准差	准确率	标准差	准确率	标准差
经济特征	0.336406	0.071442	0.346334	0.065003	0.340596	0.070706	0.339510	0.064342
技术指标	0.521066	0.072551	0.430569	0.060368	0.509606	0.068360	0.424825	0.071422
舆情情感	0.333260	0.044279	0.330057	0.043524	0.341232	0.041537	0.328704	0.044345
经济特征+技术指标	0.508651	0.064266	0.331272	0.046805	0.503141	0.070488	0.405872	0.052203
经济特征+舆情情感	0.344894	0.057285	0.409418	0.058148	0.357707	0.066345	0.334309	0.045967
技术指标+舆情情感	0.504822	0.071031	0.427784	0.063909	0.512972	0.061119	0.426997	0.064690
经济特征+技术指标+舆情情感	0.514235	0.072083	0.399252	0.063024	0.508165	0.068652	0.405774	0.063720

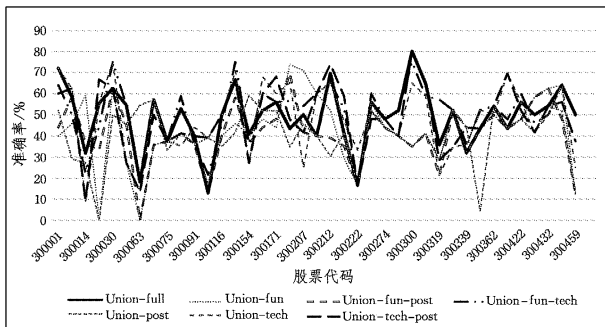


图 2 不考虑非交易日数据时线性核 SVM 的准确率

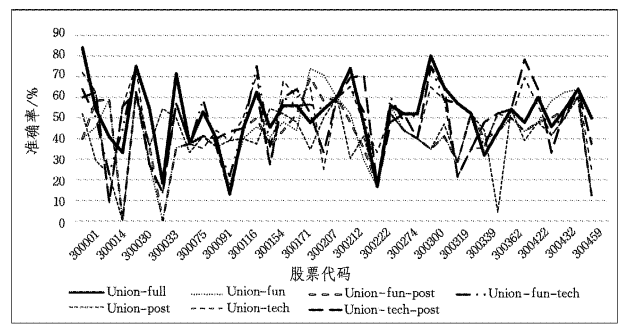


图 4 考虑非交易日数据时线性核 SVM 的准确率

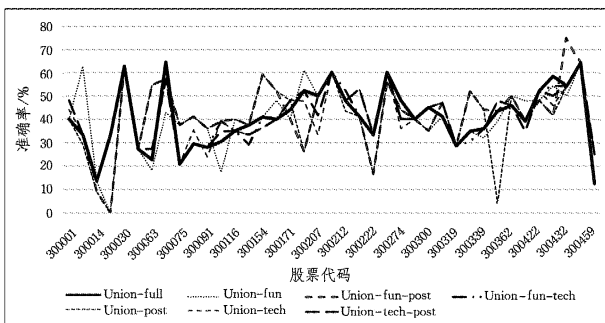


图 3 不考虑非交易日数据时高斯核 SVM 的准确率

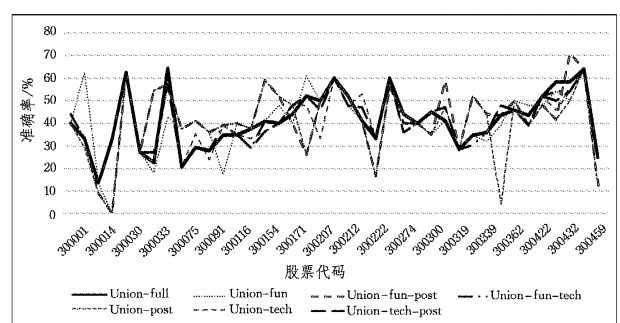


图 5 考虑非交易日数据时高斯核 SVM 的准确率

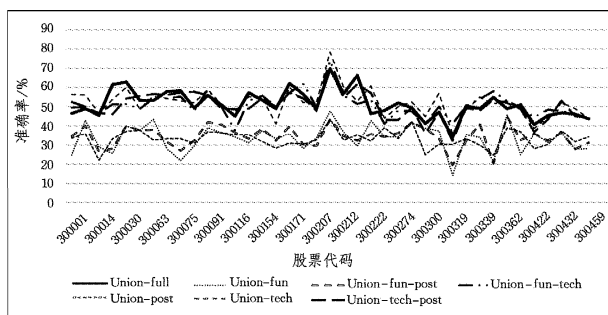


图6 不考虑非交易日数据时线性核 SVM 在交叉验证法下的准确率

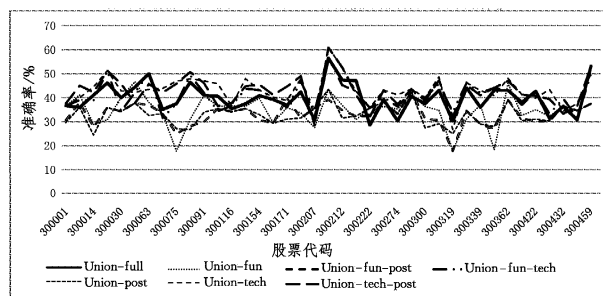


图7 不考虑非交易日数据时高斯核 SVM 在交叉验证法下的准确率

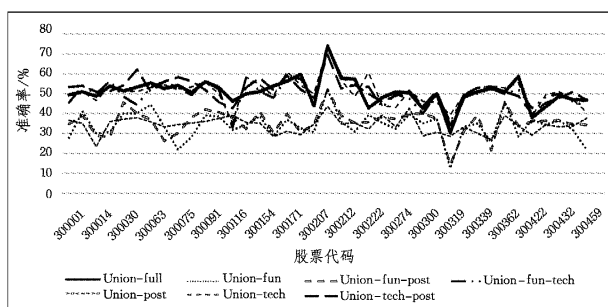


图8 考虑非交易日数据时线性核 SVM 在交叉验证法下的准确率

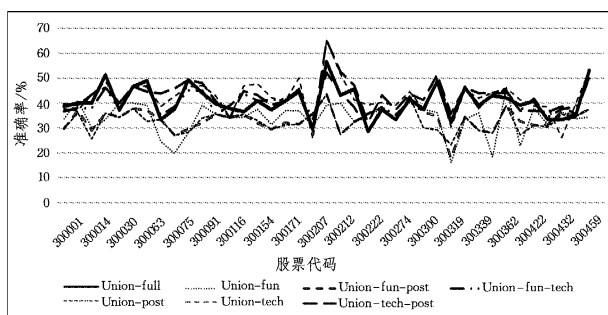


图9 考虑非交易日数据时高斯核 SVM 在交叉验证法下的准确率

观察图 2—图 9 可得如下结果。一方面,观察每条曲线的走势(从横向看)可以发现,不同股票用同样的输入特征组合和同样的核函数构造出来的分类器的(交叉验证)准确率差异比较大,原因可能源于 3 个方面:1)不同股票的训练样本数目不同或停牌日期不完全相同,因此它们的有效交易日不完全相同。2)这 39 只股票虽然同为创业板股票,但是它们所属的行业是不同的,因此行业差异导致的内部隐藏信息会影响准确率。3)在舆情情感倾向特征方面,不同股票的评论帖子数相差较大,一些其他属性(例如帖子热度、评论量或者最新回复等)并没有作为特征量输入,因此情感文本的特征种类及

规模也影响着准确率。另一方面,观察不同曲线的截面(从纵向看)可以发现,对同一只股票用不同的输入特征组合构造出来的分类器的准确率也有所不同,这表明同一特征组合并非对所有股票的预测都具有绝对优势。此外,交叉验证的准确率普遍比测试集上的分类准确率高。

为了从统计意义上比较不同特征组合的准确率,基于图 2—图 9,在同一特征组合、同一核函数、同一数据集上对所有股票的准确率求平均值(即每条曲线的平均值)。根据式(11)计算出综合准确率 $evaluation$,并从高到低排序,排名前 3 位的输入特征组合如表 6 所列。

表6 分类效果最好的前 3 位特征组合(四分类)

evaluation 排名的前三名			
	1	2	3
不考虑非交易日			
	Union-fun-tech	Union-full	Union-tech
$verify_set_{linear, noadd}$	0.364989	0.346930	0.306674
$train_set_{linear, noadd}$	0.444384	0.442152	0.448514
$evaluation_{linear, noadd}$	0.404687	0.394541	0.377595
	Union-tech	Union-tech-post	Union-fun-post
$verify_set_{rbf, noadd}$	0.284776	0.289522	0.264804
$train_set_{rbf, noadd}$	0.370202	0.363874	0.312692
$evaluation_{rbf, noadd}$	0.327489	0.326699	0.308037
考虑非交易日			
	Union-full	Union-fun-tech	Union-tech-post
$verify_set_{linear, add}$	0.358312	0.364989	0.313127
$train_set_{linear, add}$	0.439513	0.432654	0.451852
$evaluation_{linear, add}$	0.398913	0.398822	0.382490
	Union-tech-post	Union-tech	Union-fun-tech
$verify_set_{rbf, add}$	0.287668	0.284776	0.282399
$train_set_{rbf, add}$	0.362307	0.353403	0.353669
$evaluation_{rbf, add}$	0.324988	0.319090	0.318035

由表 6 可知:1)线性核 SVM 比高斯核 SVM 的综合准确率更高,这表明对于本文的数据,线性核比高斯核更适合作为核函数。2)排名前 3 的多数为 2 种或者 3 种信源的组合,这说明多种信源组合比单一信源的预测效果更好。3)由于涨跌类型分为 4 类,随机预测的准确度为 0.25,而加入了任何一种信源后的预测准确度均大于 0.25,这说明了信源信息对预测模型的作用。4)没有一种信源组合的准确率远超其他组合,这说明信源之间如何组合最有效需分情况讨论。相对来说,技术指标在前 3 名排序中出现的次数较多,可见其反映的信息更全面,对提高预测模型的精准度的作用较大;而包含舆情与不包含舆情的特征组合的准确率差别不大,这说明了舆情信息对散户影响大,对机构的影响不大,对预测模型的整体影响不大。5)在训练集上交叉验证的准确率 $train_set$ 普遍比在测试集上的分类准确率 $verify_set$ 更高,这表明学习出现了轻微的过拟合性。

从表 6 中还可以看到多源组合对本文两个假设的支持。首先,对于 $evaluation_{linear, add}$, Union-full 排名第一,这表明在选择线性核和考虑非交易日数据的情况下假设 1 成立。其次,对于 $evaluation_{linear, noadd}$,前 3 名分别是 Union-fun-tech, Union-full, Union-tech;而对于 $evaluation_{linear, add}$,前 3 名变为

Union-full, Union-fun-tech, Union-tech-post, 由此可见包含网络舆情的信源组合排名更靠前, 其支持假设 2。类似地, 对于 $evaluation_{rf, noadd}$, Union-tech-post 排名第二; 而对于 $evaluation_{rf, add}$, Union-tech-post 排名第一, 这也支持了假设 2。

但是, 表 6 中的实验结果存在一个问题, 即准确率普遍过低 (小于 0.5), 这主要是由于将股票涨跌分为 4 种类型, 即股票价格趋势不仅有涨跌之分, 而且还有振幅超过 5% 的区别。这样的考虑能够使交易更加慎重和有信心, 也能顾及交易成本以及 $t+1$ 交易制度的影响。关于准确率的问题, 对于同样的数据, 我们做了股票价格趋势二分类的实验 (即只考虑涨跌两种类型)。将样本标注类型修改为:

$$class_{new} = \begin{cases} 1, & 0 < ratio \\ -1, & ratio \leq 0 \end{cases} \quad (14)$$

其中, 1 表示涨, -1 表示跌。

实验结果如表 7 所列。可以看到, 分类准确率得到提高, 同样支持假设, 稳定性也得以提升。

表 7 分类效果最好的前 3 位特征组合 (二分类)

evaluation 排名的前 3 名			
	1	2	3
不考虑非交易日			
	Union-fun-tech	Union-full	Union-tech-post
$verify_set_{linear, noadd}$	0.587255	0.578457	0.551014
$train_set_{linear, noadd}$	0.723491	0.727497	0.725385
$evaluation_{linear, noadd}$	0.655373	0.652977	0.6381995
	Union-tech-post	Union-full	Union-tech
$verify_set_{rf, noadd}$	0.490563	0.491770	0.488590
$train_set_{rf, noadd}$	0.616971	0.613175	0.610388
$evaluation_{rf, noadd}$	0.553767	0.5524725	0.549489
考虑非交易日			
	Union-fun-tech	Union-full	Union-tech-post
$verify_set_{linear, add}$	0.587255	0.590967	0.551014
$train_set_{linear, add}$	0.724999	0.713086	0.703772
$evaluation_{linear, add}$	0.656127	0.6520265	0.627393
	Union-tech-post	Union-full	Union-tech
$verify_set_{rf, add}$	0.497563	0.489818	0.488590
$train_set_{rf, add}$	0.610285	0.611173	0.611885
$evaluation_{rf, add}$	0.553924	0.550496	0.550238

结束语 本文建立了基于多信源的股价预测模型。为了实现更广泛的信源融合, 选取基本经济特征、技术指标、网络舆情 3 种不同的信息源, 基于 SVM 分类器建立股价预测模型。首先, 对网络舆情扩展情感词典并补充金融情感词及权值。然后, 对各种数据进行归一化处理。最后, 对 3 种信源的所有组合进行测试。实验结果表明, 在选用线性核函数和考虑非交易日数据时, 以 3 种信源全面组合构造出来的分类器的分类效果最好; 进一步, 考虑了非交易日数据的影响, 其可以提高分类器的分类效果。

未来工作主要体现在 5 个方面: 考虑不同信源的权重贡献; 具体信源对具体类型股票的作用; 考虑更多信源组合; 探讨各信源之间的相关性; 考虑非结构化的大规模数据。

1) 融合多源信息时, 需要注意不同数据源的权重贡献。在目前的实验中将权值均设置为相等是因为 3 种数据源的数

据归一化之后聚合在一起, 形成一个统一的输入向量, 而该输入向量的权值不能通过 libsvm 调整, 因此权值都相等。在未来的工作中, 可以通过增加预处理阶段对特征设置不同的权值, 再将其作为 libsvm 的输入向量。关于权值设置的趋势, 目前的观察是: 技术指标的特征较多, 影响较大, 应减少权值; 而情感文本的特征较少, 应增大权值, 这样可以平衡不同数据源对预测精准度的影响。

2) 不同类型的信源对不同类型的股票的影响可能不同。由图 2—图 9 可知, 虽然在实验中选取的都是创业板股票, 但是在所有配置相同的情况下, 不同股票的预测准确率差异较大。即使同为创业型, 由于公司规模或者所在行业的不同, 也会使得某种信源对它们的影响不同。信源对股票类型的针对性影响更值得深入研究。

3) 应拓展更多信源, 如停牌公告或金融研报。停牌期间的公告对股价的影响会反映在复牌后的股价上, 或者股价会随着研报的推荐行为发生变化。对于公告或研报, 与处对网络舆情扩展情感词典类似, 找出其中关联股价的情感词。

4) 信源的相关性是更深层次的信息。同一信源内部特征之间存在相关性, 不同信源之间也存在着相关性, 后者更值得研究。信源并非越多越好, 有时会相辅相成, 有时会背道而驰。比如, 股吧帖子有可能是根据上市公司发布的公告而进行讨论, 也有可能是针对某个研报进行讨论, 它们之间存在着相关性。可以参考 Li 等的工作^[25]用张量来表示多源信息空间, 从而便于描述信源之间的相关性。

5) 非结构化数据应与结构化数据联合使用。已有研究甚至本文几乎均使用结构化数据, 其实互联网空间中有大量的非结构化数据, 例如视频、照片、音像等。这些数据具有极大的价值, 但难以分析和解释, 并且存在噪音, 如果能够将它们与结构化数据一并使用, 则可能揭示更深层次的相关性, 以预测有关的经济指数。可参考 Radzimski 等的处理非结构数据的方法^[31], 使用大数据分析技术, 根据上市公司之间的共生性来估算股票收益率的时间相关性。

参 考 文 献

- [1] FAMA E F. Efficient Capital Markets: A Review of Theory and Empirical Work[J]. Journal of Finance, 1970, 25(2): 383-417.
- [2] CAO Q, PARRY M E, LEGGIO K B. The three-factor model and artificial neural networks: predicting stock price movement in China[J]. Annals of Operations Research, 2011, 185(185): 25-44.
- [3] GAO S J, XU D M, WANG H Q, et al. Knowledge-based anti-money laundering: A software agent bank application[J]. Journal of Knowledge Management, 2009, 13(2): 63-75.
- [4] CUI B G, WANG H Q, YE K, et al. Intelligent agent-assisted adaptive order simulation system in the artificial stock market [J]. Expert Systems with Applications, 2012, 39(10): 8890-8898.
- [5] CARHART M M. On Persistence in Mutual Fund Performance

- [J]. Journal of Finance, 1997, 52(1): 57-82.
- [6] NOVY-MARX R. The other side of value; The gross profitability premium ☆ [J]. Journal of Financial Economics, 2013, 108(1): 1-28.
- [7] FAMA E F, FRENCH K R. A five-factor asset pricing model [J]. Journal of Financial Economics, 2014, 116(1): 1-22.
- [8] FAMA E F. Market efficiency, long-term returns, and behavioral finance [J]. Journal of Financial Economics, 1998, 49(3): 283-306.
- [9] SHILLER R J. From Efficient Market Theory to Behavioral Finance [J]. Social Science Electronic Publishing, 2003, 17(1): 83-104.
- [10] LI X, XIE H, WANG R, et al. Empirical analysis: stock market prediction via extreme learning machine [J]. Neural Computing & Applications, 2014, 27(1): 67-78.
- [11] PATEL J, SHAH S, THAKKAR P, et al. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques [J]. Expert Systems with Applications, 2015, 42(1): 259-268.
- [12] LEE M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction [J]. Expert Systems with Applications, 2009, 36(8): 10896-10904.
- [13] BOLLEN J, MAO H, ZENG X. Twitter mood predicts the stock market [J]. Computer Science, 2010, 2(1): 1-8.
- [14] XU L. Empirical research on the impact of network public opinion on the stock price volatility [D]. Chengdu: Southwestern University of Finance and Economics, 2013. (in Chinese)
徐琳. 网络舆情对股价波动影响的实证研究 [D]. 成都: 西南财经大学, 2013.
- [15] WANG L L. Clarification Announcement of listed companies and share price volatility-based on the research of investor behavior [D]. Nanjing: Nanjing University of Science & Technology, 2014. (in Chinese)
王莉莉. 上市公司“澄清公告”与股价波动——基于投资者行为的研究 [D]. 南京: 南京理工大学, 2014.
- [16] YANG J. Empirical analysis of the impact of Internet financial news on stock-based on perspective of semantic analysis of company news [D]. Chengdu: Southwestern University of Finance and Economics, 2012. (in Chinese)
杨娟. 互联网财经新闻对股票影响的实证分析——基于公司新闻语义分析的视角 [D]. 成都: 西南财经大学, 2012.
- [17] SHYNKEVICH Y, MCGINNITY T M, COLEMAN S A, et al. Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning [J]. Decision Support Systems, 2016, 85(C): 74-83.
- [18] DUAN J, LIN H, ZENG J. Posterior probability model for stock return prediction based on analyst's recommendation behavior [J]. Knowledge-Based Systems, 2013, 50: 151-158.
- [19] NEWMAN M R, GAMBLE G O, CHIN W W, et al. An Investigation of the Impact Publicly Available Accounting Data, Other Publicly Available Information and Management Guidance on Analysts' Forecasts [M] // New Perspectives in Partial Least Squares and Related Methods. New York: Springer, 2013: 315-339.
- [20] DUAN J, ZENG J. Forecasting stock return using multiple information sources based on rules extraction [C] // 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 15). Piscataway, New Jersey: IEEE, 2015: 1183-1188.
- [21] ZHAI Y, HSU A, HALGAMUGE S K. Combining News and Technical Indicators in Daily Stock Price Trends Prediction [C] // International Symposium on Neural Networks: Advances in Neural Networks. Springer-Verlag, 2007: 1087-1096.
- [22] LI X, HUANG X, DENG X, et al. Enhancing quantitative intraday stock return prediction by integrating both market news and stock prices information [J]. Neurocomputing, 2014, 142(1): 228-238.
- [23] GUNDUZ H, CATALTEPE Z. Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection [J]. Expert Systems with Applications, 2015, 42(22): 9001-9011.
- [24] WU D, FUNG G P C, YU J X, et al. Integrating Multiple Data Sources for Stock Prediction [J]. Web Information Systems Engineering, 2008, 5175: 77-89.
- [25] LI Q, CHEN Y, JIANG L L, et al. A tensor-based information framework for predicting the stock market [J]. ACM Transactions on Information Systems, 2016, 34(2): 11.
- [26] FAMA E F, FRENCH K R. The Cross-Section of Expected Stock Returns [J]. Journal of Finance, 1992, 47(2): 427-465.
- [27] TSAI C, LIN Y, YEN D C, et al. Predicting stock returns by classifier ensembles [J]. Applied Soft Computing, 2011, 11(2): 2452-2459.
- [28] LAM M. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis [J]. Decision Support Systems, 2004, 37(4): 567-581.
- [29] SIDOROV G, VELASQUEZ F, STAMATATOS E, et al. Syntactic N-grams as machine learning features for natural language processing [J]. Expert Systems with Applications, 2014, 41(3): 853-860.
- [30] WU H C, LUK R W P, WONG K F, et al. Interpreting TF-IDF term weights as making relevance decisions [J]. ACM Transactions on Information Systems, 2008, 26(3): 55-59.
- [31] RADZIMSKI M, SÁNCHEZ-CERVANTES J L, CUADRADO J L L, et al. Predicting stocks returns correlations based on unstructured data sources [C] // Joint Proceedings of the Second International Workshop on Semantic Web Enterprise Adoption and Best Practice and Second International Workshop on Finance and Economics on the Semantic Web Co-Located with European Semantic Web Conference. Anissaras, Greece, May, 2014.