北京大学学报(自然科学版) 第 53 卷 第 2 期 2017 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 53, No. 2 (Mar. 2017)
doi: 10.13209/j.0479-8023.2017.037

基于文本信息的股票指数预测

董理 王中卿 熊德意†

苏州大学计算机科学与技术学院, 苏州 215006; † 通信作者, E-mail: dyxiong@suda.edu.cn

摘要 基于情感分析方法, 对股票市场进行预测。将从社交媒体中抽取的文本信息(词信息、情感词信息和情感分类信息)与股票技术指标相结合, 利用支持向量回归构建模型。通过实验与多种预测方法进行比较, 结果表明该方法能够获得较为理想的预测结果。

关键词 股票预测; 情感分析; 支持向量回归(SVR)

中图分类号 TP391

Stock Index Prediction Based on Text Information

DONG Li, WANG Zhongqing, XIONG Deyi †

School of Computer Science and Technology, Soochow University, Suzhou 215006;
† Corresponding author, E-mail: dyxiong@suda.edu.cn

Abstract Sentiment analysis strategy was used to predict stock market index. Support vector machine was applied to construct predict model based on textual information (i.e., lexical information, sentimental words, and sentiment categories) extracted from social media and stock indicators. Experiment results show that the proposed method can obtain the best results, compared with many different predictive model.

Key words stock index prediction; sentiment analysis; support vector regression (SVR)

近年来, 随着金融领域的高速发展, 越来越多的人希望利用股票市场实现财富的保值与增值。因此, 学术界和商业界都希望通过一定的方法来预测股票市场的未来走势。

早期对股票市场的预测是基于股票市场的技术指标, 如成交量、股票价格、指数平滑异同平均线(MACD)和随机指标(KDJ)等。随着 Fama[1−2]有效市场假说(efficient-market hypothesis, EMH)的提出, 围绕股票磐石价值的研究开始流行。然而, 近年来人们发现股票市 场的变化并不像 EMH 描述的那样随机散漫[3~4]。Schumaker 等[5]利用新闻信息对股票市场进行预测, 取得不错的效果。随着行为经济学的发展, 有研究者开始关注大众情感在日常金融领域(如电影票房、图书销量等)中的作用[6−8]。还有研究表明, 大众情感能在一定程度上预测股票市场将来可能的变化[9−10]。

从上述研究中不难看出, 在股票技术的基础上, 结合其他相关信息构建模型对股票市场进行预测, 不失为一种行之有效的方法。因此, 我们以股票技术指标为基础, 利用词信息和情感信息对其加以优化, 用支持向量回归(SVR)方法构建模型来预测股票市场可能的走势。考虑到股票预测所需要的数据量, 我们选择通过新浪微博这类社交媒体来收集数据, 然后通过一定的规则清洗原始数据集。在此基础上, 从规则方法和统计方法的角度出发, 采用情感词典和支持向量机两种方式进行情感分析。最后, 利用 SVR 方法, 结合词信息、情感信息以及股票技术指标构筑模型, 并对其性能加以评判。通过与多种回归预测方法比较, 证明我们提出的方法能够提高模型的预测准确度。

国家自然科学基金(61403269)和江苏省自然科学基金(BK20140355)资助
收稿日期: 2016−07−21; 修回日期: 2016−09−23; 网络出版日期: 2016−11−28

1 相关工作

股票市场预测一直是金融领域研究的重点，常见的预测手段是利用历史股值以及 MACD, KDJ 等技术指标进行分析。然而，股票市场具有高度的复杂性，仅利用技术指标难以获得理想的预测结果。有效市场假说[1−2]指出，由于新闻、历史股价和内部消息的作用，股票市场具有不可预测性。然而，Kavussanos 等[3]对雅典证券交易所进行实例研究，结果与 EMH 假说的股票市场活动规律不相符。Gallagher 等[4]研究需求冲击和供给冲击对股票市场价格的影响，进一步表明 EMH 具有一定的局限性。Qian 等[11]利用神经网络、决策树和 K 近邻算法训练模型来预测股票市场，获得高于 EMH 理论值的预测精度。

鉴于 EMH 假说指出新闻是驱动股票市场价格变化的重要因素之一，学者们从新闻的角度做了大量研究。Fung 等[12]从文本中抽取包含多个新闻的时间序列，研究不同序列之间的关系以及对预测股票市场的作用。Schumaker 等[5]使用词、名词短语和命名实体 3 种不同文本表示方法分析金融文本，抽取其中有价值的词条，利用支持向量机(SVM)训练分类器进行分析，发现利用文本字段和股票价格一起训练能够获得最好的表现。Nikfarjam 等[13]分析对比利用新闻文本和历史价格训练 SVM 分类器进行预测的效果，发现与单纯使用新闻文本相比，将新闻和股票的技术指标相结合能获得更准确的预测结果。

此外，随着行为经济学的发展，越来越多的学者意识到大众情感变化在金融领域的重要作用。Gruhl 等[6]通过文本情感分析的方式，分析网络聊天记录，利用聊天者的情感倾向预测未来图书的销量。Mishne 等[7]从博客评论中提取大众对电影的评价，依靠大众情感预测电影票房。Liu 等[8]构建情感分析模型，用来分析博客文本中的情感信息，预测产品的销售量。

在股票市场方面, Gilbert 等[14]从 LiveJournal 中提取大众焦虑指标，根据指标的变化预测 S&P500 指数，验证了大众情绪对股票市场的作用。Bollen 等[9]在 Twitter 评论的基础上，利用 Google 情感分析工具 GPOMS 和第三方分析工具 OpinionFinder 提取情感倾向，建立模型来预测道琼斯指数，结果表明，将引入大众情绪的分析结果作为参数，能提高原有金融领域预测模型的准确度。Sehgal 等[10]从网页中收集与金融有关的文章，抽取有意义的词条，训练模型对股票市场进行预测。

上述研究中，都是从某个角度(如新闻和情感)切入，忽略了文本最基本的词信息中可能存在与股票价格变化相关的规律。此外，在构建预测模型时，常用的方法是利用简单的线性回归。然而，股票市场的变化在很大程度上难以用简单的线性关系描述。因此，本文采用支持向量回归的方式建立预测模型，希望能获得更好的预测结果。

2 数据收集

本文选择新浪微博作为数据来源，通过访问移动端页面，绕开复杂的反机器人机制，利用关键词搜索功能，收集每日关于上海证券综合指数(简称上证综指)的股票评论。由于新浪微博限制可以返回的搜索结果数量，我们选择采用多关键词的形式补充数据集合，以此满足预测所需的数据量要求。

在此基础上，对原始的数据集合进行一系列的预处理: 1) 解析 html 文件，抽取我们感兴趣的标签的内容; 2) 去重降噪，减少因搜索多个关键词引入的冗余信息; 3) 按照日期，过滤掉非当日(即今日收盘 15:00 到次日开盘 9:00 之间)发布的博文。

在实验过程中，我们利用与股市有关的多个关键词(如股票、股市等)，收集从 2016 年 3 月 5 日至 4 月 29 日，共 40 个交易日的关于上证综指的股票评论微博。每日平均可以收集到 35000 条原始博文。之后，经过上述的预处理步骤，可以提纯出每日平均 13000 条博文。比如，“两市果然小幅翻红，放心大盘会持续上涨。”，“A 股再次进入下跌趋势，将会有巨大跌幅，小心……”，这两句博文都是股民对股票市场的评价，其中蕴含对股票市场的情感。前者代表积极的情感，后者代表消极的情感。利用这些博文构成的数据集，抽取出我们感兴趣的文本信息来构建模型进行预测。数据集相关的信息见表 1。

3 分析方法

我们在清洗后的数据集上，抽取每日的文本信息(词信息和情感信息)，并与当日实际的股票技术指标相结合，利用 SVR 构建模型进行预测。首先利用词信息构建基准模型，然后在词信息模型中逐步添加情感信息和股票技术指标，从而对股票指数

表 1　数据集相关信息
Table 1　Dataset information

项目	内容
时间跨度	2016 年 3 月 5 日至 4 月 29 日(40 个交易日)
数据集对象	上证综指大盘
关键词列表	股票、股市、A 股等共 40 个
原始博文数量(每日)	平均 35000 条
清洗后博文数量(每日)	平均 13000 条
博文示例	1) 两市果然小幅翻红，放心大盘会持续上涨。2) A 股再次进入下跌趋势，将会有巨大跌幅，小心……

进行全面的预测。方法的整体框架如图 1 所示。

3.1　文本信息抽取

文本信息包括词信息和情感信息两部分。我们利用 jieba[①]分词工具，对数据集合中的博文进行分词，提取词信息。情感信息包括情感词和情感分类结果两部分。我们利用情感词典来标识情感词；对于情感分类结果，分别用情感词典和支持向量机两种方法给出。

3.1.1　情感词典构建

由于泛用的情感词典没有收录有关股票市场的情感词，所以需要针对股票市场的特点自行构建情感词典。针对 2016 年 3 月 5—6 日的样本，利用 jieba 分词工具分词，统计每个词出现的频率，人工标注热门词语的情感极性。据此，我们获得一个有 932 词的词典，其中含积极倾向的词 442 个，消极倾向的词 490 个(表 2)。

3.1.2　情感信息抽取

1) 情感词典。

首先，利用 jieba 分词工具对博文进行分词，检索情感词典，获取每个词对应的情感值。然后，将句子中所有词的情感值之和作为句子的情感值，将句子的情感值作为判断博文情感倾向的依据。最后，将博文情感分为积极、消极和中性三类。

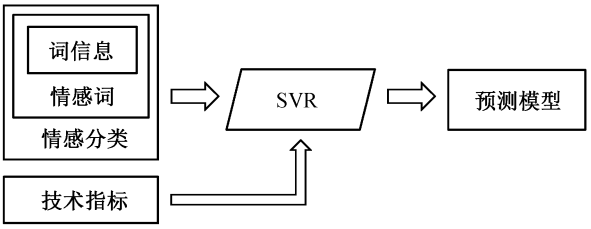

图 1　预测模型
Fig. 1　Predict model

表 2　情感词典示例
Table 2　Example of sentiment lexicon

情感极性	数量	示例
积极	442	飘红、走高、看好、涨停、利好 等
消极	490	萎缩、减持、回落、探底、跳水 等

将由情感词典方法得出的分析结果记为 Lex〈posLex, negLex, calmLex〉。其中，posLex 代表情感词典方法中积极情感倾向所占的比例，negLex 代表情感词典方法中消极情感倾向所占的比例，calmLex 代表情感词典方法中中性情感倾向所占的比例。

2) 基于 SVM 的机器学习方法。

从 2016 年 3 月 5—6 日的样本中人工标注正负(分别对应积极和消极情感倾向)800 个样本，随机打乱，选择其中正负 700 个作为训练样本，剩余的正负 100 个作为测试样本。利用训练所得分类器，将文本情感分为积极和消极两类。

将 SVM 分类方法得出的情感结果记为 SVM〈posSVM, negSVM〉，其中 posSVM 代表 SVM 结果中积极情感所占的比例，negSVM 代表 SVM 结果中消极情感所占的比例。

在实验过程中，我们将情感词典与 SVM 分类结果相结合，一并作为当日的情感分类结果，希望能结合两者的优势，获得更全面的分析结果。

3.2　股票市场技术指标

在选取股票技术指标的过程中，我们参考以往金融学研究和日常股票投资经验，选择日常生活中股民经常使用的能够充分反映股票市场变化的信

① https://github.com/fxsjy/jieba

息，包括大盘的股票指数、成交量、涨跌程度、指数平滑异同平均线(MACD)和随机指标(KDJ)。其中，股票指数反映股票价格的变化，是股民最关心的指标之一；成交量客观地反映股票市场的交易情况，结合股票价格，可以推测目前和将来的行情走势；涨跌程度显示当前市场的稳定程度；MACD 暗示着中长期的买卖信号，KDJ 象征着短期的买卖信号，在实际分析过程中经常将 MACD 与 KDJ 结合起来判断。

我们将技术指标记为 tech_indicator〈stockIndex, volume, tendency, MACD, *K*, *D*, *J*〉, stockIndex 代表股票指数，volume 代表成交量，tendency 代表涨跌趋势，MACD 代表指数平滑异同平均线，*K*，*D* 和 *J* 代表随机指标 KDJ 对应的 3 个分量。

3.3 基于 SVR 预测

支持向量回归(SVR)[15]是支持向量机(SVM)在回归领域的应用。与 SVM 不同，SVR 寻求的最优超平面不是使两类样本点分得“最开”的平面，而是所有样本点离超平面的“总偏差”最小的平面。这时，样本点都在两条边界之间，求最优回归超平面等价于求最大间隔。

支持向量回归的出发点是结构风险最小化，既考虑样本的拟合性，又考虑样本的复杂性，具有良好的外推预测能力。SVR 方法的核心是模型参数的选择，包括损失函数参数 ε、惩罚因子 C 以及核参数等。

我们在多特征信息的基础上，利用 SVR 构建预测模型。首先利用词信息构建原始模型，然后逐步添加情感词信息和情感分类信息，最后加入股票技术指标信息。通过分析对比，研究不同特征信息对预测结果的影响。

4 实验

4.1 实验设置

通过 40 个词的关键词列表，收集新浪微博中关于上证综指的评论，再用一定的方法进行数据清洗，获得每日 1.2~1.5 万条的原始数据集。该数据集的时间跨度为 2016 年 3 月 5 日至 4 月 29 日，共 40 个交易日。在此基础上，选择 2016 年 3 月 7 日至 4 月 12 日(共 25 天)的数据来建立模型进行预测，利用 2016 年 4 月 13—29 日(共 13 天)的数据进行测试，将股票指数归一化，采用均方误差(MSE)衡量预测结果。对 SVR 方法的预测效果与普通线性拟合的预测效果进行比较，同时，比较不同特征值对预测效果的影响。

实验过程中，利用最小二乘法实现普通线性拟合，利用 libSVM①实现 SVR 算法。抽取文本信息过程中，利用 SVM-light②实现支持向量机的情感分类方法。这些工具的参数设置均为默认值。

我们构建 4 个模型：Model_words，Model_sentiment_words，Model_sentiment_analysis 和 Model_text_tech，用于比较特征信息对预测结果的影响。其中，Model_words 为词信息的模型，Model_sentiment_words 为增加情感词的模型，Model_sentiment_analysis 为增加情感分类结果的模型，Model_text_tech 为增加技术指标的模型。此外，在对模型输入特征的时候，我们用一定程度的缩放倍数，对特征值进行简单的归一化处理。

选取最佳的特征组合构建模型 Model_SVR，与基准模型 Model_baseline 及 Model_WE 和 Model_LR 进行比较，对比预测能力。其中，Model_baseline 是我们设计的简单预测模型：利用前 7 天股票指数的平均值作为预测值；Model_WE 是使用深度学习方法构建的模型：利用 gensim③中实现的 Word2Vec 方法，求取词向量和文档向量，将文档向量中最大的维度作为特征值，与词信息一起，利用 SVR 构建预测模型；Model_LR 是将股票技术指标进行线性拟合构建的预测模型。用 MSE 作为评判标准，衡量归一化后的预测值和真实值。

4.2 实验结果与分析

根据上述实验框架，对比各个特征值对股票预测结果的影响。选择利用文本信息和技术指标构建的预测模型与基准模型进行对比，实验结果如表 3 所示。

我们发现，Model_LR 的结果并没有基准系统 Model_baseline 精确。这可能是由于股票市场价格虽然变化程度大，但是相对比较集中，不能真正地反映股票市场的变化。线性拟合方法中，技术指标参数较多，拟合情况相对复杂，容易出现过拟合的

① http://www.csie.ntu.edu.tw/~cjlin/libsvm/

② http://www.cs.cornell.edu/People/tj/svm_light

③ http://radimrehurek.com/gensim/

表 3　实验结果对比
Table 3　Results of different methods

系统模型	方法	MSE
Model_baseline	简单假设	0.035163
Model_LR	线性拟合	0.103503
Model_WE	深度学习	0.035491
Model_SVR	SVR	0.027454

情况。

此外，Model_WE 的预测结果比 Model_LR 更理想，充分说明利用文本信息预测股票市场变化具有合理性。利用 SVR 方法构建的模型 Model_SVR 具有相对最小的 MSE 值。在趋势变化上，Model_SVR 比其他模型更接近真实股指的变化，预测值和真实值的离散程度相对较小。这些结果充分说明，利用技术指标和文本信息进行 SVR 分析可以获得更理想的预测结果(图 2)。

我们之所以选择利用文本信息和技术指标的组合构建模型，是因为在各特征模型的对比实验中发现该组合具有更好的预测性能。不同特征组合的预测结果如表 4 所示。

不难发现，仅用技术指标构建的模型比仅用词性构建的模型预测效果好。这是因为技术指标是在长期的研究与实践中总结得到的，与股票市场的相关性比较高；而词信息相对零散，与股票价格变化之间的相关性相对较弱。

通过对词信息构建的基准模型增加特征值，我们发现在引入情感词和情感分类结果之后，预测结果有显著的提升，甚至超过技术指标的预测结果。这表明，股票市场变化与大众情感之间存在紧密的联系。我们也发现，添加情感分类结果对预测模型的提升更明显。这说明股票市场与大众情感的相关性并不是体现在某一个或某一些词上，而是更多地表现在整个句子的情感变化中。

最后，我们发现将文本信息(词信息、情感信息和情感分类结果)与技术指标相结合，能够获得相对最高的预测准确度。这与 Bollen 等[9]的研究结果一致，也就是说，为金融模型添加情感参数，的确能提升预测的准确度。

5　总结

本文在文本信息和股票技术指标的基础上，首次利用 SVR 技术构建模型。以 MSE 为依据，对比了各个特征值对预测结果的作用。同时，从 MSE

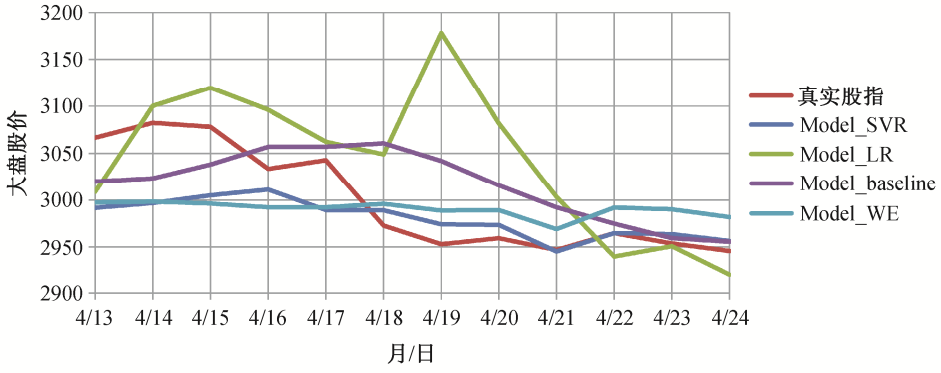

图 2　实验结果对比
Fig. 2　Results of different methods

表 4　不同特征组合的预测结果
Table 4　MSE of predict model based on different feature

模型名称	特征值	MSE
Model_tech	技术指标	0.030442
Model_words	词信息	0.035536
Model_sentiment_words	词信息、情感词信息	0.035469
Model_sentiment_analysis	词信息、情感词信息、情感分类结果	0.029486
Model_text_tech	词信息、情感词信息、情感分类结果、技术指标	0.027454

和股指变化趋势两方面，比较了该模型与其他模型的预测结果。我们发现，结合文本信息和技术指标构建的模型能够获得较高的预测准确度，其中情感分类结果和股票技术指标起着相对主要的作用。下一步的工作是选择更合适的股票技术指标，结合深度学习等方法，抽取更有代表性的特征值，增加样本的时间跨度，提升现有模型的效果。

参考文献

[1] Fama E F. Market efficiency, long-term returns, and behavioral finance. Journal of Financial Economics, 1998, 49: 283−306

[2] Fama E F. Efficient capital markets: a review of theory and empirical work. The Journal of Finance, 1970, 25: 383−417

[3] Kavussanos M G, Dockery E. A multivariate test for stock market efficiency: the case of ASE. Applied Financial Economics, 2001, 11(5): 573–579

[4] Gallagher L A, Taylor M P. Permanent and temporary components of stock prices: evidence from assessing macroeconomic shocks. Southern Economic Journal, 2002, 69: 345–362

[5] Schumaker R P, Chen H. Textual analysis of stock market prediction using breaking financial news: the AZFinText system. ACM Transactions on Information System, 2009, 27(2): 1139–1141

[6] Gruhl D, Guha R, Kumar R, et al. The predictive power of online chatter // KDD’05 Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York: ACM, 2005: 78–87

[7] Mishne G, Glance N. Predicting movie sales from blogger sentiment // AAAI 2006 Spring Symposium on Computational Approachesto Analysing Weblogs. Palo Alto, 2006: 155−158

[8] Liu Y, Huang X, An A, et al. ARSA: a sentiment-aware model for predicting sales performance using blogs // SIGIR’07 Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2007: 607–614

[9] Bollen J, Mao H, Zeng X J. Twitter mood predicts the stock market. Journal of Computational Science, 2010, 2(1): 1–8

[10] Sehgal V, Song C. SOPS: stock prediction using web sentiment // Workshops IEEE International Conference on Data Mining. Omaha, 2007: 21–26

[11] Qian B, Khaled R. Stock market prediction with multiple classifiers. Applied Intelligence, 2007, 26: 25–33

[12] Fung G P C, Yu J X, Lam W. Stock prediction: integrating text mining approach using real-time news // IEEE International Conference on Computational Intelligence for Financial Engineering. Hong Kong, 2003: 395–402

[13] Nikfarjam A, Emadzadeh E, Muthaiyah S. Text mining approaches for stock market prediction // International Conference on Computer & Automation Engineer. Singapore, 2010: 256–260

[14] Gilbert E, Karahalios K. Widespread worry and the stock market // Fourth International AAAI Conference on Weblogs and Social Media. Washington, DC: ICWSM, 2010: 58−65

[15] Somla A J, Schölkopf B. A tutorial on support vector regression. Statistics & Computing, 2004, 14(3): 199–200