计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190400145

基于金融文本情感的股票波动预测

赵 澄 叶耀威 姚明海
浙江工业大学信息工程学院 杭州 310014
(zhaoc@zjut.edu.cn)

摘 要 股票市场的情绪可以在一定程度上反映投资者的行为并影响其投资决策。市场新闻作为一种非结构性数据,能够体现并引导市场的大环境情绪,与股票价格一同成为至关重要的市场参考数据,能够为投资者的投资决策提供有效帮助。文中提出了一种可以准确、快速地建立针对海量新闻数据的多维情绪特征向量化方法,利用支持向量机(Support Victor Machine,SVM)模型来预测金融新闻对股票市场的影响,并通过 bootstrap 来减轻过拟合问题。在沪深股指上进行实验的结果表明,相比于传统模型,所提方法能够将预测准确度提高约 8%,并在 3 个月的回测实验中获得了 6.52%的超额收益,证明了其有效性。

关键词:股票市场预测;金融情感驱动;新闻;文本特征;交易信号;人工智能
中图法分类号 TP391

Stock Volatility Forecast Based on Financial Text Emotion

ZHAO Cheng, YE Yao-wei and YAO Ming-hai
College of Information Engineering, Zhejiang University of Technology, Hangzhou 310014, China

Abstract Emotions in the stock market can reflect investor behavior to a certain extent and influence investors' investment decisions. As a kind of unstructured data, market news can reflect the advantages and disadvantages of the market environment, and become a vital market reference data with stock prices, which can provide effective help for investment decisions effectively. This paper proposes a multidimensional emotional feature vectorization method which can accurately and quickly establish a large amount of news data for massive news data. It uses the support victor machine (SVM) model to predict the impact of financial news on the stock market, and uses bootstrap to mitigate overfitting problems. The results on Shanghai and Shenzhen stock indexes show that compared with the traditional model, the proposed method can improve the prediction accuracy by about 8% and obtain an excess of 6.52% duringthree months, thus proving the effectiveness of the proposed method.

Keywords Stock market prediction, Financial emotion driven, News, Text feature, Trading signal, Artificial intelligence

1 引言

股票市场是一个极为重要的金融市场,投资者作为股票市场的重要参与者,其情绪变化会迅速反映到市场上。随着信息技术的发展,投资者获取信息的渠道和速度发生改变,社交媒体、新闻网站中丰富的信息内容逐渐成为影响投资者投资预期,乃至投资决策的又一重要因素。换言之,股票市场除了受到股价历史数据等结构化数据的影响,还受到一些非结构化数据的间接影响,如新闻事件的发酵对投资者行为的影响[1]。金融新闻诱导放大了投资者对于股票市场的态度倾向[2]。随着新闻、博客、论坛以及社交网络言论的扩散,网络上的文本提供了能够反映投资者心理的信息数据[3]。

然而这些数据虽然重要,却难以直接运用于定价,需要专家参与,并且预测结果严重依赖于主观经验。这样的方法不仅存在滞后和遗漏,同时容易在情感理解和准确度把握上失衡,从而导致结果偏差严重。因此,为确保分析的准确性和及时性,本文针对网络金融新闻以及历史股价数据构建了金融新闻特征对股票市场预测模型。该模型以支撑向量机为基础,基于股票历史数据来对训练集进行构建,并通过 bootstrap 减少过拟合,提高模型泛化能力。

本文主要贡献如下:1)利用多维特征来判定金融新闻的情感极性,根据特征明显程度给予特征词条不同的特征权重,以提高文本数据的分析能力;2)基于股票时序数据对训练集进行构建,规避人为判定的情绪化偏差风险;3)优化模型的重要超参,并且采用 bootstrap 进行随机采样以避免过拟合。实验结果验证了本文模型确实对金融新闻处理具有重要意义。

2 文献综述

以金融社会学、行为经济学和行为金融学[4-6]的视角来说,有效市场理论并不一定是正确的。大量的研究表明,股票

到稿日期:2019-04-26 返修日期:2019-08-28 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61902349)
This work was supported by the National Natural Science Foundation of China (61902349).
通信作者:姚明海(ymh@zjut.edu.cn)

市场的价格并不遵循随机游走，并且确实在一定程度上能被预测。而情感就是用于反映股票市场状态的一个常用指标。Bollen 等利用道格琼斯指数分析推特的用户情感来进行股票预测，结果显示，该方法在预测日常股票收盘价的升降变化上具有良好的表现，并且将平均绝对误差百分比(Mean Absolute Percent Error，MDPE)减少了 6%[7]。同时，Ann 等也发现新闻的内容和情感表现显著影响市场及市场中的交易量、股价等，甚至是公司未来的利润。其原因在于，新闻或者通告中一些积极性情感的词语(如飙升、增长)和消极性情感的词语(如暴跌、下降)往往会给投资者一个心理暗示，诱使投资者做出相应的投资决策[8]。Makrehchi 利用 Rocchio 方法系统地评估了市场大规模变化可用于预测的情绪类别和其预测能力，并在综合分析后发现情绪具有较高的预测价值，且能够带来超额的交易利润[10]。但显而易见，上述方法均为传统的人为分析方法，并不适用于当前拥有数百万用户、成千上万条实时新闻以及庞大用户情感内容的社交网站分析[9,11]。换言之，人为分析方法已无法满足信息爆炸的局面。由于在金融股票市场中，反映市场情感的非结构性数据(如宏观经济信息、金融新闻以及时间情绪指标)具备非线性、非平稳的特征，随着信息技术的发展，机器学习成为了解决金融股票市场的非结构性情感数据的有效方法。研究表明，SVM 是机器学习领域最优秀、准确且健壮的算法之一，其对维度不敏感，能够同时处理线性可分和线性不可分数据，其分类性能好、稳定性高、算法更新快的特点也尤为显著。对于金融市场情绪指标，基于 SVM 方法模型的结构风险最小化的泛化能力较好，不易陷入局部最优，具备良好的分类性能[12]。Huang 等提出使用 SVM 模型来对股票市场的运动方向进行分类，对 NIKKEI 日经指数周运动方向进行预测，发现 SVM 模型的准确性优于线性判别分析(Linear Discriminant Analysis，LDA)、二次判别分析(Quadratic Discriminant Analysis，QDA)以及 Elman 反向传播神经网络。而后 Rui 等在周日效应的基础上，构建了基于 SVM 的模型，利用更可靠、更现实的情绪指数来对上证 50 进行预测，得到了良好的上证 50 状态预测模型。Chen 等提出了特征加权的 SVM 混合框架，得到了在短中长期对沪深股指预测效果都具有很好表现的预测模型[13-15]。

综上可知，随着数据规模的增长，对金融情感的主观认识分析已经不能满足数据规模快速增长的现状，使用高速计算工具已有其现实必要性，而传统使用信息技术的交易方法仅停留在金融情感数据的层面[16]，忽略了股票市场时序序列(如股价等数据)的重要性。因此，本文基于 SVM 模型，利用最能反映金融情感特征的金融新闻以及股票市场本身所具备的时序数据，来构建对股票市场极性的预测模型，在传统信息技术的基础上，提高了对股票市场时序数据的利用率，扩大了对金融新闻特征的维度，提高了模型预测的准确性。

3 金融新闻情感挖掘模型

3.1 模型结构

传统的新闻情感挖掘模型[17]存在一些缺陷，主要集中在以下几个方面。1)特征词描绘不准确。该模型对金融新闻特征词进行判断，然而只采用了通用词典，缺乏相应的准确性。2)金融新闻的表示情感缺失。金融新闻通常将被表示为术语向量，并且传统模型鲜有使用情感在语义层面分析新闻的行为。3)金融新闻标记不全面。传统模型仅为每条金融新闻做标记并分配一个标签(如积极或消极)，然后在分类模型中新闻片段按照时间戳进行排序，并用标签值进行标记。除此之外，一些模型方法只对金融新闻标记分类标签值。

鉴于此，本文从以下方面进行改进：首先，在金融新闻的文本处理方面，对情感进行极性判断，构建特征词库，利用多个特征对新闻进行分析，以保证预测准确性，减少判断误差；其次，对于新闻的标注，将其与分钟级别的股票数据结合，并使用程序自动标注，在保证预测准确性的基础上减少了人工标注大量新闻所带来的时间代价，同时也验证了大批量新闻快速标注的可行性；再次，根据 bootstrap 在所有样本中进行随机采样来减少过拟合；此外，将训练集与测试集严格分离，保证结果的可靠性以及准确性。本文提出的处理模型如图 1 所示。

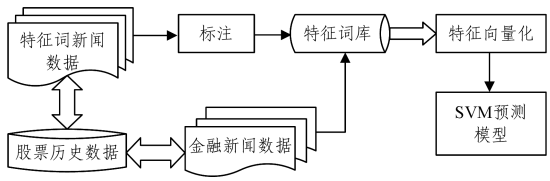

图 1 文本特征挖掘模型的结构

Fig. 1 Structure of text feature mining model

3.2 标签生成

为了排除股票自身波动所带来的影响，本文利用新闻发布前后股票价格异常收益率的变化来量化其对股票市场波动的影响。

1)正常收益率。正常收益指在没有发布金融新闻时股市的期望收益，通常采用固定平均收益模型估计正常收益率。在计算正常收益率之前需要计算收益率，收益率指标采用下式计算：

$$R_{i,t}=(P_{i,t+10}-P_{i,t})/P_{i,t} \tag{1}$$

其中，$R_{i,t}$ 表示金融新闻 i 在发布时间 t 经过 10 min 后沪深股指变化的收益率，$P_{i,t}$ 表示金融新闻 i 在发布时间 t 的指数，$P_{i,t+10}$ 表金融新闻 i 在发布时间 t 经过 10 min 后的指数。

针对金融新闻 i，以 $T_0 \sim T_{10}$ 的股指变化作为滑动窗口，通过回归分析固定平均收益模型[22]，得到 μ 的估计值：

$$\mu_i=R_{i,t}+\xi_{i,t} \tag{2}$$

其中，μ_i 表示金融新闻 i 未发布的正常收益率，$\xi_{i,t}$ 表示 μ 和 $R_{i,t}$ 之间偏差的估计值。

2)异常收益率。异常收益率指金融新闻 i 在发布时间 t 经过 10 min 后的实际收益率与正常收益率之间的差值：

$$AR_{i,t+10}=R_{i,t+10}-\mu_i \tag{3}$$

根据文献[23]，异常收益率可视作金融消息对市场的影响，并分为积极和消极，其公式如下：

$$label=\begin{cases}1, & \text{if } AR_{i,t}\leqslant 0\\ -1, & \text{if } AR_{i,t}>0\end{cases} \tag{4}$$

其中，$label$ 表示对新闻 i 的情感标签。

3.3 特征词挖掘

金融新闻内容存在很多特征词语，无法与结构化数据一样直接用于数据处理，因此需要从新闻本身的特征转化以及理解方面进行改进。首先，将文本数据转化到相应的特征空

间；其次，建立特征词库，用来保存具有代表性的特征词。特征词库用于对训练集的新闻特征进行相似度判断，以提高新闻特征分割的准确度。本文参考 Schumaker 等在分割特征词上采用的特征词分割方法[18]，在上述研究的基础上，选取了更加具有代表性的词。

在实验过程中，对于相关词特征的选取，采用标准卡方统计量来实现，卡方检验能更好地体现一个词向量在不同类别中的相关度，而金融新闻中体现文本特性的词往往并不重复，且整篇文章的极性并不依靠一个词来体现，出现频率高的词与文章极性没有太大的关系[19]，且卡方分布的临界值表可以给出特征词的相关度概率。卡方检验的形式如式(5)所示：

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{5}$$

其中，N 表示训练数据集文档总数；A 表示包含词条 t，同时属于类别 c 的文档数量；B 表示包含词条 t，但不属于类别 c 的文档数量；C 表示属于类别 c，但不包含词条 t 的文档数量；D 表示不属于类别 c，同时也不包含词条 t 的文档数量；χ 表示最终计算的卡方值。

对于常用的 tf-idf 方法，词频越高，文档的 $tf\text{-}idf$ 值越高，这种特性对提取金融新闻的特征词条的帮助不大。显然，在文档中并不需要提取高频词，而需要提取特征比较明显的词。在卡方检验中，由于 A,B,C,D 只标记词在某一文档中出现与否，不标记词在文档中的出现频率，因此其更适合用于提取金融新闻的特征词条，故本文用卡方检验提取文档的分类特征词条。特别地，卡方检验能计算不同特征词条的卡方值，并且卡方值越高，该特征词条和相关类别的关联度就越高，其具备的代表性就更强。因此，利用卡方值可以获取特征区分度从大到小排列的相关的特征词。为了使得这些词对于新闻的判断更加准确，在后续实验中为不同的特征词赋予相应的特征权重，有利于提高模型的性能。

3.4　模型训练

完成特征词库的构建后，需要对训练文本进行预处理，除去无用的停词、标点符号，并将其与词库中的特征词进行相似性比较，将最后得到的特征向量进行标记，再使用 SVM 进行模型分类训练。实验对训练过程进行了迭代，每次迭代都进行交叉验证，并且使用 bootstrap 进行随机采样，随机分配训练集(80%)以及测试集(20%)，最后得到一个具备合适超参数的分类模型。训练过程的伪代码如下。

算法 1

输入：训练集 $S=\{(x_{11},x_{12},\cdots,x_{1m},y_1),(x_{21},x_{22},\cdots,x_{2m},y_2),\cdots,(x_{n1},x_{n2},\cdots,x_{nm},y_n)\}$

输出：具备合适超参数的 SVM 模型

1. while 准确率$<$0.75 do
2. $S_{train},S_{test}\leftarrow S$

   其中，S_{train} 为抽样出来的训练集，S_{test} 为剩下的测试集。该步骤主要利用 bootstrap 方法，从数据集中有放回地进行随机采样。
3. 进行交叉验证获得准确率
4. end while

4　实验结果以及评估

4.1　实验准备

本文抓取新浪财经、同花顺以及中证网在 2016－2018 年间共 15 968 篇有关于沪深宏观股票市场的金融新闻[20]，金融新闻均为简体中文，实验以相同时间段内的上证 500 指数作为市场参照。实验环境采用 Python3.7，编译平台为 Pycharm。

4.2　超参数优化

在实验模型设置中，需要调整的参数主要是惩罚因子以及相应的判别特征词的数量。不同惩罚因子 C 有不同的分类效果，实验结果如图 2、图 3 所示。

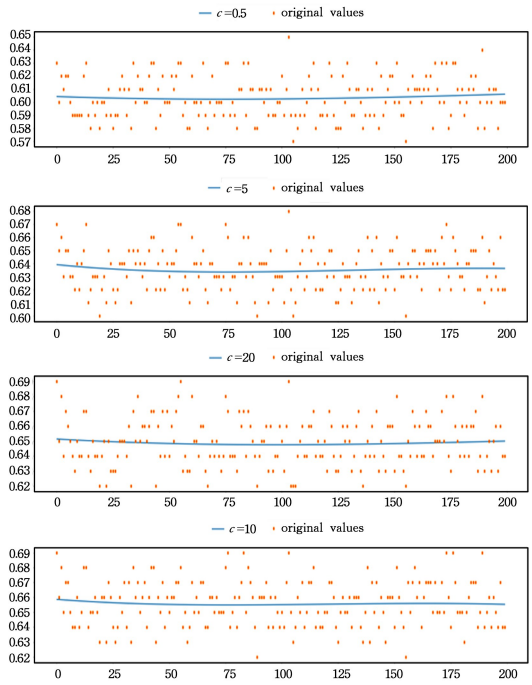

图 2　不同惩罚因子 C 对预测准确率的影响(积极预测)

Fig. 2　Influence of different penalty factors C on predictive accuracyrate (positive prediction)

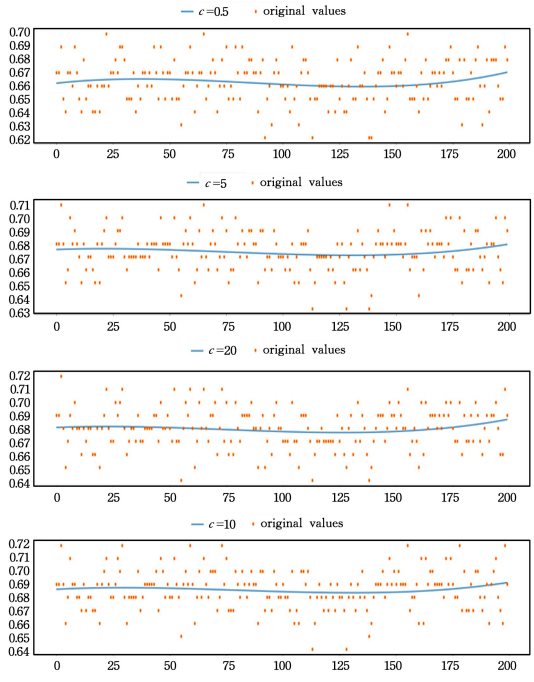

图 3　不同惩罚因子 C 对预测准确率的影响(消极预测)

Fig. 3　Influence of different penalty factors C on predictive accuracy rate (Negative prediction)

同样地，特征词数量会影响预测的准确率以及模型训练速度，过多的特征词会影响模型训练以及分类的速度，并增加数据的噪声；过少的特征词会降低分类准确度。图 4、图 5 显示了不同数量的特征词对预测结果的影响。

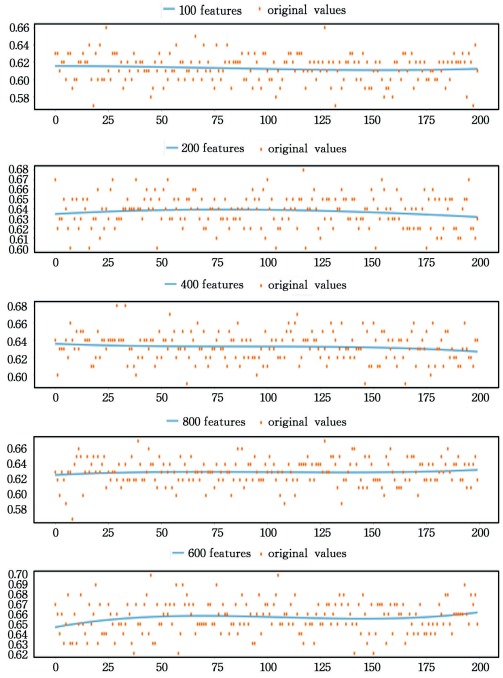

图 4 不同特征词数量下预测精度的比较(积极)

Fig. 4 Comparison of prediction accuracy under different feature words (positive)

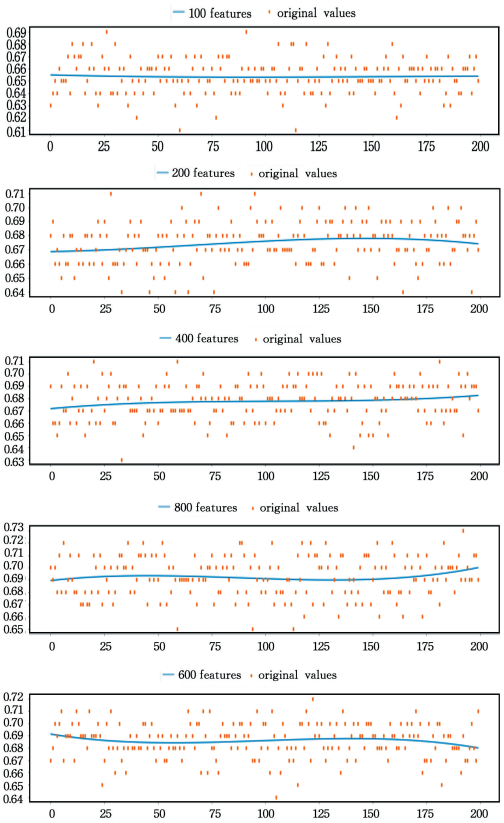

图 5 不同特征词数量下预测精度的比较(消极)

Fig. 5 Comparison of prediction accuracy under different feature words (negative)

可以看到，特征词数量在 100～400 时预测精度不断增加，在 400～800 时预测精度并没有显著增长，并且随着特征词数量的增加，模型的训练时间以及对验证文本的处理时间显著增加。因此，结合训练结果的准确率和模型训练速度，本文将特征词数量定为 600，超参 C 的值为 10。

4.3 实验结果及比较

本文将文献[21]提出的 BP 神经网络(BP-NN)，被广泛运用于文本分类的伯努利朴素贝叶斯模型(B-NB)，以及本文提出的情感判别模型进行对比实验。每个模型的准确率如表 1、表 2 所列。

表 1 各个模型在测试集中的准确率(积极)

Table 1 Accuracy of each model in test set (positive)

积极预测	准确率	召回率
BP-NN	0.53	0.54
SVM	0.65	0.68
B-NB	0.62	0.55

表 2 各个模型在测试集中的准确率(消极)

Table 2 Accuracy of each model in test set (negative)

消极预测	准确率	召回率
BP-NN	0.54	0.59
SVM	0.62	0.62
B-NB	0.68	0.55

可以看出，在积极(Positive)新闻的判定上，SVM 相比于另外两个模型都有较好的准确率。但是在消极(Negative)新闻的分类上，伯努利朴素贝叶斯相比于 SVM 有一定的优势。这可能是因为市场力求营造向上氛围，积极新闻占据主导，因此消极新闻一般可信度更高，与股价波动的线性相关更强，更适合伯努利朴素贝叶斯算法。根据前面的分析，市场总体的氛围是向上的，因此在保证消极新闻预测的基础上，显然对积极新闻预测较好的 SVM 更有优势。实验，BP-NN 的效果并非很出色，这是因为语料总量有限，而神经网络对于较小数据集的处理效果并不是很好。

图 6 给出本文模型应用于上证 500 进行为期 3 个月的收益回测的结果，Strategy Return 代表本文模型的收益，Base Return 代表基准收益。实验中排除了手续费和冲击成本。可以看出，超额收益一直稳定增加，在模拟回测中达到了 6.52%，超额年化利率达到了 26.58%，证明了所提方法的有效性。

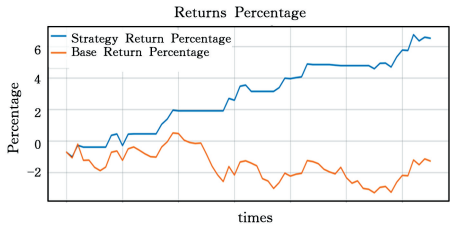

图 6 收益对比

Fig. 6 Income comparison

结束语 本文提出了一个通过金融新闻的情感极性变化来预测市场变化的 SVM 模型，并通过实验进行验证。下一步的工作目标是进一步扩展金融文本的来源，不局限于新闻和公告，并将其极性和影响力因子结合，以进一步探索金融文本挖掘的广度和深度。

参 考 文 献

[1] OLIVEIRA N，CORTEZ P，AREAL N. Stock market sentiment lexicon acquisition using microblogging data and statistical measures[J]. Decision Support Systems，2016，85：62-73.

[2] LONG W，TANG Y，TIAN Y. Investor sentiment identification based on the universum SVM[J]. Neural Computing and Applications，2018，30(2)：661-670.

[3] PERIKOS I，HATZILYGEROUDIS I. Recognizing emotions in text using ensemble of classifiers[J]. Engineering Applications of Artificial Intelligence，2016，51：191-201.

[4] WU B，ZHOU X，JIN Q，et al. Analyzing Social Roles Based on a Hierarchical Model and Data Mining for Collective Decision-Making Support[J]. IEEE Systems Journal，2015：1-10.

[5] JIANG F，LEE J，MARTIN X，et al. Manager sentiment and stock returns [J]. Journal of Financial Economics，2019，132(1)：126-149.

[6] MIWA K. Investor sentiment，stock mispricing，and long-term growth expectations[J]. Research in International Business and Finance，2016，36：414-423.

[7] BOLLEN J，MAO H，ZENG X. Twitter mood predicts the stock market[J]. Journal of Computational Science，2011，2(1)：1-8.

[8] SUL H K，DENNIS A R，YUAN L. Trading on twitter：Using social media sentiment to predict stock returns[J]. Decision Sciences，2017，48(3)：454-488.

[9] OLIVEIRA N，CORTEZ P，AREAL N. On the predictability of stock market behavior using stocktwits sentiment and posting volume[C]// Portuguese Conference on Artificial Intelligence. Berlin，Heidelberg：Springer，2013：355-365.

[10] MAKREHCHI M，SHAH S，LIAO W. Stock prediction using event-based sentiment analysis[C]// Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01. IEEE Computer Society，2013：337-342.

[11] CHECKLEY M S，HIGÓN D A，ALLES H. The hasty wisdom of the mob：How market sentiment predicts stock market behavior[J]. Expert Systems with Applications，2017，77：256-263.

[12] NIKKINEN J，SAHLSTRÖM P. Impact of Scheduled US Macroeconomic News on Stock Market Uncertainty：A Multinational Perspecive[J]. Multinational Finance Journal，2011，5(2)：129-148.

[13] REN R，WU D D，LIU T. Forecasting stock market movement direction using sentiment analysis and support vector machine [J]. IEEE Systems Journal，2018，13(1)：760-770.

[14] CHEN Y，HAO Y. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction[J]. Expert Systems with Applications，2017，80：340-355.

[15] HUANG W，NAKAMORI Y，WANG S Y. Forecasting stock market movement direction with support vector machine[J]. Computers & Operations Research，2005，32(10)：2513-2522.

[16] CHEN W，ZHANG Y，YEO C K，et al. Stock market prediction using neural network through news on online social networks [C]// 2017 International Smart Cities Conference (ISC2). IEEE，2017：1-6.

[17] HÁJEK P. Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns[J]. Neural Computing and Applications，2018，29(7)：343-358.

[18] SCHUMAKER R P，CHEN H. A discrete stock price prediction engine based on financial news[J]. COMPUTER，2010，43(1)：51-56.

[19] CI Y X，ZHAO S L，LUO Y，et al. Text data preprocessing method based on word frequency statistics[J]. Computer Science，2017，44(10)：276-282，288.

[20] LI L，ZHANG G Y，LI Z W，et al. Research on topic crawler technology based on SVM[J]. Computer Science，2015，42(2)：118-122.

[21] LI X，XIE H，WANG R，et al. Empirical analysis：stock market prediction via extreme learning machine[J]. Neural Computing and Applications，2016，27(1)：67-78.

[22] YAO W D，WANG R J. An Empirical Study of the Relationship between Stock Market Volatility and Policy Events from the Perspective of Structural Decomposition—Based on EEMD Algorithm [J]. Shanghai Economic Research，2016(1)：71-80.

ZHAO Cheng，born in 1985，Ph.D，senior engineer. His main research interests include quantitative financial and artificial intelligence.

YAO Ming-hai，born in 1963，professor，Ph.D，doctoral tutor. His main research interests include pattern recognition and intelligent control，control theory and control engineering，and computer application.