基于文本价格融合模型的股票趋势预测*

余传明[1]　龚雨田[1]　王　峰[1]　安　璐[2]
[1](中南财经政法大学信息与安全工程学院　武汉 430073)
[2](武汉大学信息管理学院　武汉 430072)

摘要: 【**目的**】在传统股票预测模型的基础上，提高股票价格预测准确率，降低股票交易风险，研究大数据环境下的股票价格变化趋势。【**方法**】提出一种新的文本价格融合模型。该模型对股票论坛上的评论文本预处理后，通过深度表示学习生成评论文本的特征矩阵，使用 K 均值聚类方法生成文本类别；结合开盘价、收盘价等 15 个原始价格指标，使用多层感知机算法预测股票价格趋势。【**结果**】使用该模型进行预测，所得精度为 65.91%，超出单独使用价格特征的模型 7.76%，超出单独使用文本特征的模型 11.37%，预测性能具有较大提升。【**局限**】只对个股进行预测研究。【**结论**】本文模型从文本和价格结合的角度出发提高股票预测精度，为股价趋势预测相关研究者和从业者提供新的研究方法和研究视角。

关键词: 文本　股票价格　股票价格趋势预测　文本价格融合模型
分类号: TP391.1
DOI: 10.11925/infotech.2096-3467.2018.0420

1　引　言

随着国民经济迅速发展，越来越多的人开始进行金融投资，并将其作为一种重要理财手段。有效市场假说认为，在有效市场中存在着大量追求利益最大化的理性投资者[1]。由于人们在精力、能力、信息等方面的局限，存在过度自信、后悔厌恶、心理账户、处置效应、羊群效应、本土偏差等非理性投资心理和相应的非理性投资情绪，因此在面临投资选择问题时很难达到经济上所说的理性人假设，即所谓“主观上是理性的，但客观上做不到”现象[2]。从英国南海公司投机泡沫破裂[3]到中国股市“铁达尼号”琼民源股海沉船[4]，从荷兰郁金香狂热[3]到中国股市 2015 年暴跌，这些事实表明，在金融市场中仍然存在大量非理性投资者。研究非理性投资者的交易行为和收益对揭示金融市场运行机理、规范金融市场具有重要意义。

在上述背景下，越来越多的学者开始研究非理性投资者对金融市场的影响。例如，王洪良等[5]对上证股市非理性行为进行实证分析，表明上证股票市场存在一定程度的噪音交易风险，指出资本市场只有回归理性才能发挥服务实体经济的作用，市场理性是资本市场与实体经济协同发展的必要条件。Nagy[6]认为无论是 16 世纪的郁金香狂热事件，还是 1987 年的市场泡沫和崩溃，2008 年的网络泡沫和金融危机，羊群心理在创造这些历史性的股市泡沫方面都具有极大的推进作用。股票泡沫的造成是由于股票价格在一段时间内与其内在价值大相径庭，直至达到爆发态势。泡沫可以合理地形成，但泡沫的负面影响通常由非理性投资行为造成。邹辉文[7]指出，大部分股民对股市的非理性心理导致其非理性的投资行为，对市场的新信息和波动做出过激反应，常常造成股票价格异常波动，从而影响股市正常发展、损害股票投资者利益。史青春等[8]研究发现，市场上对公司的负面舆情容易导致股票价格大幅下跌，造成收益率异常。于瑾等[9]研究发现，投

通讯作者：安璐, ORCID: 0000-0002-5408-7135, E-mail: anlu97@163.com。
*本文系国家自然科学基金面上项目“大数据环境下基于领域知识获取与对齐的观点检索研究”(项目编号：71373286)和中南财经政法大学科研项目“证券交易量化投资策略研究”(项目编号：3251612007)的研究成果之一。

资者在非理性加入较高的杠杆交易时，会显著增加股票的暴跌风险。岳衡等[10]研究表明，收盘价格对股票趋势具有一定影响，当收盘价以 0 为尾数时，随后的股票趋势更可能上升；收盘价以 9 为尾数时，则更可能下降。吴璇等[11]证实了网络舆情管理可以缓解信息不对称等问题，降低投资者信息收集成本和偏差，改善股票流动性。林川[12]通过研究上市公司发现，不良的市场情绪导致的非理性投资行为对股市崩盘具有促进作用。上述研究表明，通过研究股票市场，建立股价预测模型，能够在变幻莫测的股票市场中合理投资、控制股票交易风险从而增加股票收益。

在现阶段，对股票进行预测的研究较多地局限于各种价格指标而忽略相应的文本信息；通过文本信息对股票价格进行预测的研究则局限于使用评论文本中的情绪指标，而忽略了其他信息。鉴于此，本文提出一种综合文本信息和价格信息对股票价格趋势进行预测的模型(Text and Price Combined Model, TPCM)，以期为相关研究提供借鉴。

2 相关研究

由于股市内部影响因素众多，特别是股市价格的波动不仅受宏观货币政策[13]影响，还受到宏观经济环境和突发事件影响，因此股价预测行为通常具有较强的时效性。根据股票价格预测机理的不同，本文从以下两个方面进行回顾。

2.1 基于价格的股票预测研究

股票价格预测研究起步较早，很多传统的股票价格预测模型被广泛应用。卢磊[14]选取 5 日均线等 4 个股票技术指标建立多元非线性模型，选取中国农业银行 33 个交易日数据和滨江股份 40 个交易日数据，对这两只企业的股票进行价格预测，预测结果误差小于 2%。陈璐璐[15]对每股收益、每股净资产和上证指数等 7 个股票价格技术指标建立多元线性回归模型，使用 EViews 软件计算回归系数，预测中信银行 2016 年 4 月 1 号的收盘价，尽管预测误差较大，但仍在可接受范围内。这两个研究的模型易于实现，也具有一定的预测能力，但是处理的数据量都相对较小，在面对大规模数据的预测任务时表现相对较弱。

为提升股价预测模型在大数据环境下的效果，越来越多的学者尝试使用机器学习算法。张建宽等[16]利用万科 2014 年 235 个交易日的股票数据，选择开盘价、收盘价等 18 个股票技术指标，用支持向量机模型训练股价预测模型，对股票价格波动进行预测，准确率为 62.86%，具有较好的预测性能。黄宏运等[17]采用 2002 年–2016 年沪深 300 股价指数数据，选取开盘价等技术指标，使用前向反馈神经网络算法建立股价预测模型，预测结果误差低于 10%。相对于传统算法，机器学习算法具有处理数据量大、处理数据维度多等特点，并且训练出来的模型预测性能较好，因此越来越多的研究者使用机器学习算法进行股价预测。

在众多机器学习算法中，神经网络算法是使用最多的一类算法。由于其具有很强的泛化能力和较强的非线性映射能力，能够很好地处理非结构化数据，而且有高度并行性，理论上更适合处理股票交易数据。魏文轩[18]通过主成分分析法选取影响股价的主要因素，利用改进的 RBF 神经网络算法、使用中国神华 80 个交易日的数据训练预测模型，并进行股价预测，预测结果误差低于 5%。蔡红等[19]利用主成分分析和 BP 神经网络算法将首创股份 2010 年 3 个月交易数据的收盘价等股票技术指标作为输入变量，进行建模分析和预测，预测误差小于 5%。Göçken 等[20]从股票技术指标与股票市场之间的关系这一角度出发，利用和声搜索 (Harmony Search) 算法和遗传算法 (Genetic Algorithm)，选择最相关的技术指标，将指标应用于人工神经网络中进行股票价格预测。实验结果表明，基于和声搜索和遗传算法的人工神经网络模型的股价预测平均绝对百分比误差分别为 3.38%和 3.36%，优于仅使用人工神经网络算法的模型(3.81%)。此外，在郭建峰等[21]以及 Adebiyi 等[22]使用神经网络算法预测股价的实验中，神经网络算法均有较好的预测性能。值得说明的是，这些研究选取的特征大多局限于收盘价、开盘价和最高价等股票技术指标，结合股票论坛上的股票评论文本信息进行股价预测的研究并不多见。

2.2 基于文本的股票价格预测研究

伴随着智慧金融的兴起和 Web3.0 的高速发展，互联网用户基于 Web 进行的金融活动形成的时空网络越来越庞大。在基于互联网信息的股票交易网络中，各大股票论坛中的用户评论信息对股票交易行为逐渐开始产生影响。Evangelopoulos 等[23]使用潜在语义分析研究与 18 个财富 500 强公司相关的推特(Twitter)发帖，

以主题的形式提取语义和概念内容，使用这些主题因子、Twitter 帖子数量和主题强度拟合一个回归模型，并使用该回归模型预测超出股票市场所能解释的 8.3% 的股价变动率。其实验结果表明推特帖子可以作为股价变动的宏观指标。王健俊等[24]、石勇等[25]、于琴等[26]的研究表明投资者情绪对股票市场具有较大影响，挖掘股票评论文本中的情绪能够提升股票价格变化趋势预测的效果。董理等[27]选取新浪微博作为文本数据源，获取上海证券综合指数的相关股票评论，从这些文本信息中提取情感信息，结合当日的股票技术指标，利用支持向量回归模型(Support Vector Regression, SVR)预测股票价格，试验取得了相对较低的均方误差(0.027)。黄润鹏等[28]使用文本包含的情绪信息进行股票市场预测，取得较好效果。Yan 等[29]利用中文简明心境状态测试(Chinese Profile of Mood States, C-POMS) 分析新浪微博、人人网等社交网站上与股票相关文本数据中的情感因素，使用支持向量机和概率神经网络对 2014 年 6 月–2014 年 7 月的上证综合指数进行预测。模型预测精度为 71.42%，比不使用文本的模型预测精度高出 20%。此外, Nguyen 等[30]、Li 等[31]使用社交媒体上文本中的情感信息研究股票市场，均取得较好的实验效果。

上述研究多利用 TF-IDF 等传统文本特征，不能很好地揭示和获取不同领域文本信息的内在语义表示，从工作量上来看显得较为繁琐。深度学习(Deep Learning)通过构建深层神经网络处理文本、语音和图像等信息，能够更有效地提取不同层次和维度的深层特征，相关研究表明，深度学习应用于股票价格预测能取得较好的效果[32]。目前对如何使用文本深度学习进行股票价格预测的研究仍然相对较少，本文尝试更深层次地挖掘股票评论中隐藏的内在信息，将面向评论文本的深度学习研究与股票价格波动相结合，提出 TPCM 模型并通过多组实证对比研究检验其有效性。

3　研究方法

根据预测结果的不同，现有预测股票价格的研究可以划分为两种，一种是对股票价格即具体的股价数值进行预测，另一种则是对股票价格变动趋势即股价的涨跌进行预测。韩豫峰等[33]研究表明，中国股市存在短期价格趋势，预测股票价格趋势能够显著提升投资收益。本文研究属于第二种，即预测股票价格涨跌。针对上述研究问题，本文提出一种新的文本价格融合模型 TPCM。该模型通过对股票的价格特征和股票论坛上的文本特征进行处理和融合，得到一个更高维度的股票预测特征矩阵；使用多层感知器算法(Multi-Layer Perceptron, MLP)作为最终分类器，利用该矩阵进行股票价格涨跌趋势预测。模型结构如图 1 所示，包括价格特征处理模块、文本特征处理模块以及融合模块三个部分。

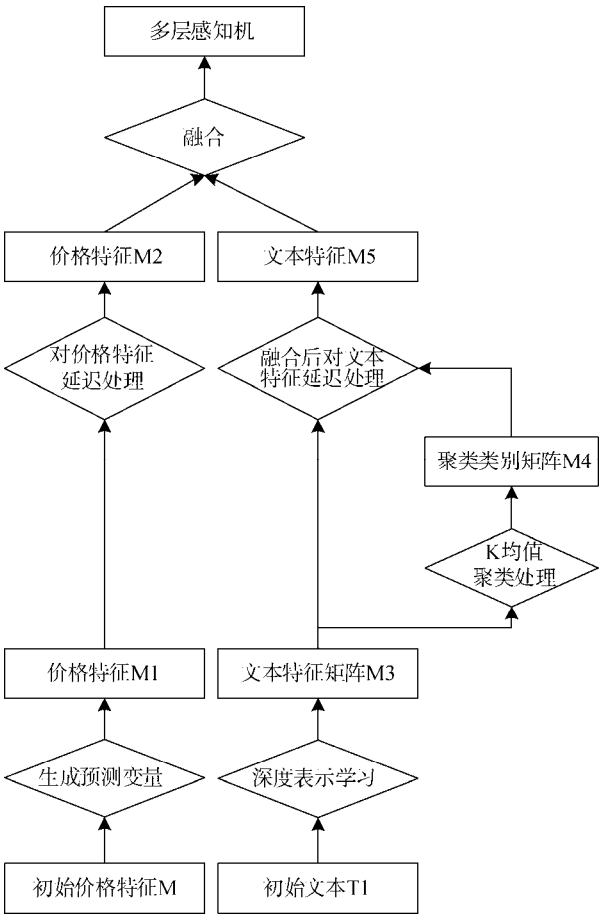

图 1　文本价格融合模型

3.1　价格特征处理模块

价格特征处理模块包括利用价格生成预测变量以及对价格进行延迟处理两部分(如图1左侧所示)。在利用价格生成预测变量这一环节，选取通达信软件中提供的开盘价、收盘价、最高价、最低价、成交量、5 日均线、10 日均线、20 日均线、60 日均线等 15 个技术指标作为价格方面的原始输入矩阵 M。通过计算后一天收盘价和前一天收盘价的差值计算股价的实际涨跌，并将其作为整个模型的预测变量，价格上涨记为

1，价格下跌记为 0。在剔除掉 60 日均线为缺失值的前 59 条记录后，得到初始处理矩阵 M1。

由于股票价格波动具有时序因素，需要考虑股票价格预测的最佳延迟天数，即需要确定当日股票价格信息在预测此后的第几天股价时具有最佳预测效果。将 15 个技术指标与本模型的预测变量(股票价格涨跌)做 1-7 天的延迟处理，并使用神经网络算法进行验证，确定股票价格方面的最佳延迟天数，验证结果见实验部分。在确定延迟天数后，得到价格方面的待融合矩阵 M2。

3.2 文本特征处理模块

文本特征处理模块包括利用深度表示学习生成文本特征矩阵(即深度学习环节)和对文本特征进行延迟处理两部分(如图 1 右侧所示)。

在深度学习环节，首先对股票评论按照交易日进行合并。文本预测模块的最底层数据 T1 是待预测股票相应时间段的股票评论，例如预测某只股票周一到周日的 5 个交易日股票变化趋势，T1 则对应周一到周五 5 个交易日的股票评论。在实际操作中将周六、周日的股票评论并入周五，节假日的股票评论均并入上一个交易日评论中。这种合并方式避免了节假日，即非交易日评论信息丢失导致评论信息不完整的问题。

在对股票评论进行合并后，利用合并后的语料进行深度表示学习。将语料文档中的词投射到 d 维空间，即生成维度为 d 的词向量(Word Vector)；同时，模型将每个交易日的评论视作一个文档，为其生成维度为 d 的文档向量(Document Vector)。在表示学习过程中，模型利用词向量和文档向量(被平均或串联)预测源领域和目标领域中的句子中出现的下一个单词。在模型启动阶段，文档向量和词向量被随机初始化，通过定义深度学习中的损失函数(即量化预测值与实际值之间的差距)和采用一定的优化方法(例如随机梯度下降方法)，最终获得文档向量和词向量作为上述预测任务的间接产物。通过学习，得到文本特征矩阵 M3、M4 和 M5。M3 为深度表示学习生成的 50 维文本特征向量矩阵，行代表每一日所有股票评论文本组成的文档，列代表一个维度特征。M4 为单列矩阵，行表示每一日所有股票评论文本组成的文档，列表示文档在聚类处理(K 均值聚类)中生成的聚类标签。M5 为 M3 和 M4 的组合特征矩阵，行代表每一日所有股票评论文本组成的文档，前 50 列代表维度特征，第 51 列为聚类类别特征。与价格信息的延迟一样，文本方面也存在延迟问题。在延迟处理环节，为检验文本预测模型中时差(文本发布时间与股价波动之间的时间间隔)对预测股票价格趋势的影响，以长短时记忆网络模型(Long-Short Term Memory，LSTM)和双向长短时记忆网络模型(Bi-Long-Short Term Memory，Bi-LSTM)作为工具，验证时差分别为1天-7天的效果。

3.3 文本价格融合模块

文本价格融合的具体步骤如下:

(1) 将价格方面的待融合矩阵 M2(16 维)与文本方面的待融合矩阵 M5(51 维)横向拼接得到最终的融合矩阵 M6(M6 分为训练集和测试集);

(2) 使用神经网络算法对融合矩阵 M6 的训练集进行训练，得到融合文本和价格信息的股票价格趋势预测模型;

(3) 使用M6测试集对模型进行测试，计算评估指标。

4 实验结果与讨论

4.1 数据来源

在价格数据搜集方面，本文选取中金岭南(000060)股票作为预测股票，在通达信软件上下载 2015 年 7 月 2 日至 2017 年 3 月 3 日 420 个交易日的中金岭南股票历史数据，包括收盘价等 15 个技术指标。考虑到前 60 日的 60 日均线存在数据不完整情况，将第 61 日开始的 360 个交易日历史数据作为模型的输入数据，其中前 270 个交易日作为训练数据，后 90 个交易日作为测试数据。

在文本数据搜集方面，以东方财富网的股票论坛作为主要信息来源。此前有大量研究者对股票论坛进行分析和研究，这些研究表明股票论坛上包含大量股价相关信息[34-37]，这些信息有助于分析和预测股票价格。选择东方财富网的原因在于其作为国内最大的股票论坛之一，具有权威性、全面性、专业性和即时性等优势，是股票交易者获取股市信息、进行股票信息交流的最重要平台之一。对中国股票网络论坛最流行的前三名(东方财富网股吧、和讯、金融界)进行统计，发现东方财富网股吧的用户数和浏览量最大[38]。笔者抓取东方财富网上 2015 年 7 月 2 日至 2017 年 3 月 3 日股民对中金岭南股票发表的评论作为文本数据，原始评论实例如图 2 所示。

阅读	评论	标题	作者	发表日	最后更新
14454	13	赣锋锂业，2016年营收28亿，净利润4.64亿，2017	岭南野人大师兄	07-26	07-26 07:02
19694	19	关于新南威尔士州发现世界级的大型铜钴矿和中金	岭南野人大师兄	07-20	07-26 07:00
926	0	中金岭南中期业绩增长1000%，看大涨！	股友Cnz	07-26	07-26 05:11
5295	1	新闻 中金岭南(000060)融资融券信息(07-24)	中金岭南资讯	07-26	07-26 03:27
5149	12	净入400万，还是绿的。。。你妹	友缘利丰	07-25	07-26 01:13
2314	1	看样子锌价还要持续涨涨涨 供应短缺库存减少	利好支持信心	07-25	07-26 00:36
3075	1	世界金属统计局：2017年1-5月全球锌市供应短缺	日日进万金	07-25	07-25 23:39
3538	4	现价沪锌期货每吨升380元，升幅大2%↑	日日进万金	07-25	07-25 23:36

图 2　文本评论示例

4.2　基线方法

选取集成算法(AdaBoosting)[39]、决策树(Decision Tree)[40]、最近邻(K-Nearest Neighbor, KNN)[41]、朴素贝叶斯(Naive Bayes, NB)[42]和支持向量机(Support Vector Machine, SVM)[43]这 5 种算法作为基线方法，与 TPCM 模型进行效果比较。

4.3　评价指标

选取数据挖掘与信息检索等领域应用较为广泛的准确度(Precision, P)、召回率(Recall, R)、F1 值、精度(Accuracy, ACC)和曲线下面积(Areas Under the Curve, AUC)作为评价指标。

4.4　实验结果

股票技术指标对股票价格的预测存在延迟效应，需要确定当天的股票技术指标更适合预测此后第几日的股票价格。针对神经网络算法，计算当延迟时间从 1 到 7 天变化时，使用 15 个维度的价格实验数据进行预测，得到预测结果如图 3 所示。

当选择 2 天的延迟，即当日股票技术指标预测第三个交易日的收盘价运动趋势时，股票技术指标能发挥最好的预测性能。因此本研究将价格方面的延迟时间设置为 2 天。

同样，文本信息对股票价格的预测也存在延迟效应，使用文本特征处理模块中确定文本延迟天数的方法并进行验证，得到结果如表 1 所示。时间间隔为 3 天时，两种模型均取得最好的精度。因此确定文本信息延迟天数为 3 天。

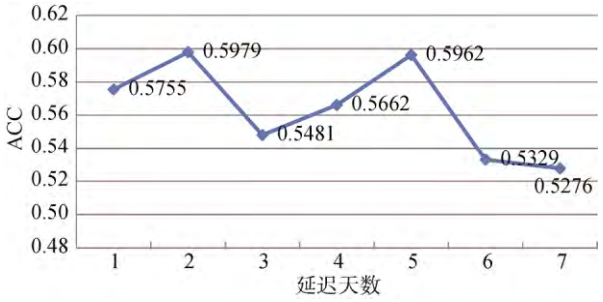

图 3　模型预测精度随股票价格技术指标延迟天数变化

表 1　文本预测模型精度随时间延迟变化

延迟时间	LSTM	Bi-LSTM
1 daylag	46.67%	52.22%
2 daylags	50.56%	44.94%
3 daylags	**53.41%**	**54.54%**
4 daylags	45.98%	49.43%
5 daylags	45.34%	43.02%
6 daylags	45.88%	45.88%
7 daylags	45.23%	39.29%

将价格和文本延迟分别设置为 2 天和 3 天，使用价格和文本实验数据对 TPCM 模型进行训练和预测。本文使用 AdaBoosting、DT、KNN、NB、SVM 和 MLP 共 6 种算法进行预测，单独使用价格因素进行预测的实验结果和使用融合模型预测的结果如表 2 所示。

表 2　各算法预测实验结果

算法		P	R	F	ACC	AUC
Price	AdaBoosting	45.12%	55.69%	50.89%	56.18%	**56.36%**
	DT	**59.86%**	58.12%	58.49%	57.90%	53.99%
	KNN	56.32%	56.42%	56.36%	54.68%	50.00%
	NB	40.58%	47.76%	44.63%	47.73%	52.00%
	SVM	54.68%	55.83%	55.46%	54.62%	53.31%
	MLP	57.67%	**58.22%**	**58.09%**	**58.15%**	50.00%
Price+Text	AdaBoosting	50.65%	51.53%	49.58%	51.21%	57.95%
	DT	41.94%	42.05%	41.86%	42.05%	42.06%
	KNN	51.17%	51.14%	50.83%	51.14%	51.12%
	NB	59.17%	59.09%	59.09%	59.09%	56.03%
	SVM	57.37%	56.82%	56.00%	56.82%	55.68%
	TPCM(MLP)	**66.78%**	**65.91%**	**65.46%**	**65.91%**	**62.66%**

在单独使用股票技术指标进行股价预测的实验中，MLP 算法表现出最好的预测效果，其 ACC 值达到 58.15%；其次为 DT 模型，其 ACC 值和 AUC 值分别达到 57.90%和 53.99%；KNN 算法和 AdaBoosting 算法分别排在第三位和第四位；SVM 和 NB 算法效果最弱。由于 MLP 算法具有很强的泛化能力和较强的非线性映射能力，能够很好地处理非结构化数据，而且具有高度并行性，因而理论上 MLP 算法更适合处理股票交易数据。

将 MLP 作为最终分类器的 TPCM 模型在整个股价预测实验中的 ACC 值(65.91%)和 AUC 值(62.66%)均高于 AdaBoosting、DT、KNN、NB 和 SVM 等基线算法。这表明 TPCM 模型在股票价格预测中优于传统基线算法。此外，实验效果的显著提升表明，通过深度学习引入的文本特征能够有效提高股价预测效果。

对比表 2 上下两部分来看，在价格特征的基础上加入文本特征后，NB、MLP 和 SVM 算法的 ACC 值和 AUC 值均有所提升(多数提升近 1%左右)，AdaBoosting 的 ACC 值降低但 AUC 值有所提升，DT 算法的 ACC 值和 AUC 值则有所降低。由此可以推论，在基于价格信息进行股价预测的基础上，加入股票评论文本信息总体而言(即在大多数算法)有助于提高股价预测准确率，但这种提升并不能推广到所有的算法。

4.5 实验结果讨论

(1) 延迟天数对股价预测的影响

从股票交易实际操作来看，交易者往往在获取足够的股票相关信息后才会做出交易决策。对于价格信息和文本信息，从产生到传播，再到转化为指导交易决策信号，通常需要一定的时间。

对比图 3 和表 1 可看出，当价格指标选取延迟天数为 2 天时，各种指标预测性能达到最优；当文本评论指标选取延迟天数为 3 天时，各种指标预测性能达到最优。这表明股票交易者对价格和文本的反应延迟并不完全同步；从延迟天数对比来看，股票交易者对价格变动比文本内容变动有更敏捷的反应速度。

(2) 文本特征对股价预测的影响

由 4.4 节实验结果可以看出，针对部分机器学习算法，在模型中加入文本信息有助于提高对股票价格预测的 ACC 值和 AUC 值。通过观察财富网上的用户发现，在股票市场上，无论是在线股票交易还是线下交易，交易者通过使用“股吧”或者类似的股票交流论坛作为获取股票信息的平台，形成一个规模庞大的社交网络。在这个社交网络中，用户通过关注他人、添加好友、转发评论等行为相互影响。其中，以文本形式进行信息传递和转换是其他行为的重要基础。例如，如果用户对某个评论内容产生共鸣，则更倾向于关注该评论作者。在社交网络中，用户会受到彼此发表的文本信息中情绪、逻辑和可信度等因素不同程度的影响，这些影响在某种意义上会反应到交易者的股票交易行为上，从而影响股票价格。综上，文本信息从一定程度上影响到用户的社交网络行为，继而反馈到股票价格上，理论上可作为股票价格的预测变量。

在上述实验中，引入 K 均值聚类后得到的特征矩阵存在数据冗余。为检验 K 均值聚类特征是否会降低整个方法的效果，在使用 M5(M4+M3)进行预测实验的基础上，增加单独使用 M3 进行预测的实验，结果如图 4 所示。

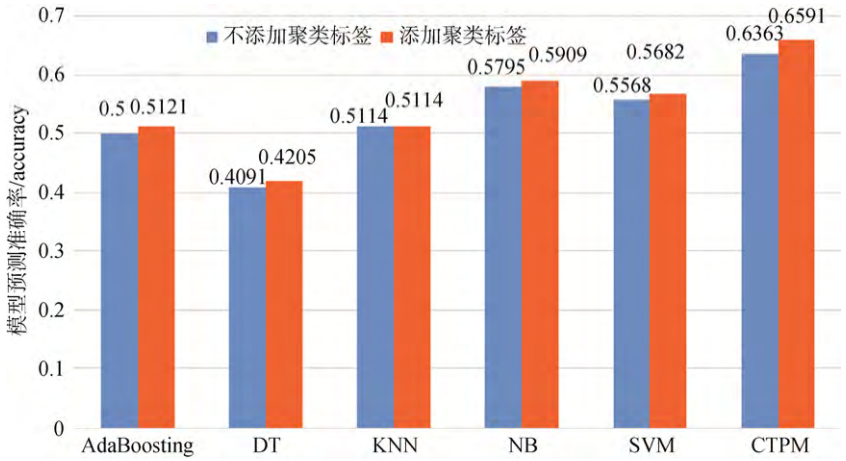

图 4 添加文本聚类标签与不添加文本聚类标签预测结果对比

在 6 种算法的对比实验中，除了 KNN 算法外，其他算法使用矩阵 M5 的 ACC 值高于单独使用 M3 的 ACC 值。实验结果表明，在文本特征方面，结合聚类标签能够有效提升股票价格趋势预测效果。

(3) 深度表示学习对股价预测的影响

对比 TPCM 模型与其他基线方法的实验结果可以看出，通过深度学习抽取文本特征能够有效提高股价预测效果。在深度学习过程中，模型能够在小的语境中考虑到单词顺序，这点与 n-gram 模型方式相同。n-gram 模型保留段落的大量信息，包括单词顺序。传统的 n-gram 模型往往需要创建一个非常高维的表示，而表示学习模型能够创建一个相对低维的表示，因此表示学习模型比传统 n-gram 模型具有更好的性能。此外，在深度表示学习环节，词向量和文档向量同时从未标记的用户评论中学习，特征的获取以利用上下文预测单词作为切入点，能够脱离繁重的人工标注过程，使模型具有更好的推广性。

本研究对股票评论的文本数据进行训练，得到的词向量维度为 50 维。为检验词向量维度对实验结果的影响，在原实验 50 维词向量的基础上，增加了 10 维、100 维、150 维、200 维、250 维的词向量，分别使用上述维度的数据进行实验，得到预测精度如图 5 所示。

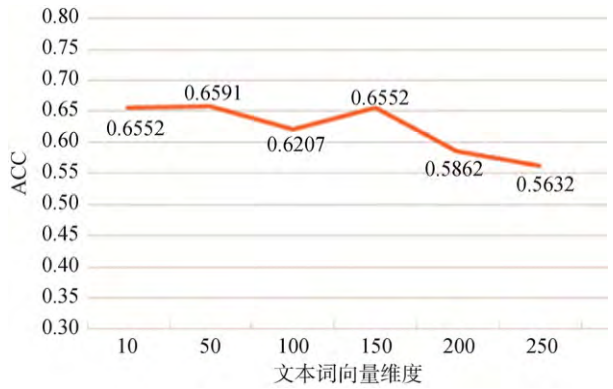

图 5　不同文本向量维度对应预测实验结果对比

当词向量维度为 50 维时，模型预测准确率最高。这表明，在利用文本评论进行深度学习时，选择词向量维度对股价预测结果具有一定影响，随着向量维度的增加，模型整体预测性能先趋于平稳，后呈现下降趋势。

(4) 分类算法选择对股价预测的影响

在基于价格信息进行股价预测的基础上，加入股票评论信息总体而言(即在大多数算法)有助于提高股价预测的准确率，但这种提升并不能推广到所有分类算法。例如，DT 和 KNN 算法的 ACC 值和 AUC 值在加入文本特征后反而降低。这表明选择或者建立适当的预测模型在股价预测中同样具有重要作用。如果模型不能对文本信息有效建模，则预测模型的效果在加入文本信息后有可能降低。

从上述实验结果可以看出，神经网络在加入文本信息和未加入文本信息的两种模型中均具有较好的表现，在股票价格预测方面具有较好的预测能力[15-19]。

5　结　语

本文提出一种新的文本价格融合模型。该模型在对股票论坛上的评论文本预处理后，使用文本深度表示学习生成评论文本的特征矩阵，使用 K 均值聚类方法生成文本矩阵类别；结合开盘价、收盘价等 15 个原始技术指标，使用多层感知机算法对股票价格进行预测，取得了优于基线方法的预测效果。对比实验对从事文本深度学习以及股票价格预测方面的研究者有一定的借鉴意义。

本文的局限性在于仅对个股进行实证研究。在后续研究中将对相关指数和板块等进行更多测试，以进一步验证该模型的预测能力。

参考文献:

[1] Fama E F. The Behavior of Stock-Market Prices[J]. Journal of Business, 1965, 38(1): 34-105.

[2] 庄树田. 浅谈投资心理和投资行为[J]. 东南大学学报：哲学社会科学版，2015, 17(S2): 41, 46. (Zhuang Shutian. A Preliminary Analysis of Investment Psychology and Investment Behavior[J]. Journal of Southeast University: Philosophy and Social Science, 2015, 17(S2): 41, 46.)

[3] 张健. 近代西欧历史上的泡沫事件及其经济影响[J]. 世界经济与政治论坛，2010(4): 99-109. (Zhang Jian. Economic Bubbles in the History of Modern Western Europe and Influences[J]. Forum of World Economics & Politics, 2010(4): 99-109.)

[4] 师萍，李丽青，杨洵. 上市公司与审计机构信息披露的博弈模型与实证分析[J]. 管理工程学报，2004, 18(1): 44-47. (Shi Ping, Li Liqing, Yang Xun. A Game Theory Analysis Between Public Company and Audit Office in Securities Market[J]. Journal of Industrial and Engineering

Management, 2004, 18(1): 44-47.)

[5] 王洪良，詹奕椿. 上证股市非理性行为的实证分析[J]. 长春大学学报，2015, 25(7):24-29. (Wang Hongliang, Zhan Yichun. An Empirical Analysis on the Irrational Behavior of Shanghai Stock Market[J]. Journal of Changchun University, 2015, 25(7): 24-29.)

[6] Nagy J L. Behavioral Economics and the Effects of Psychology on the Stock Market [EB/OL]. [2017-08-30]. http://digitalcommons.buffalostate.edu/economics_theses/24/.

[7] 邹辉文. 投资者非理性心理行为的综合效应与股价波动[J]. 福州大学学报: 哲学社会科学版, 2008, 22(1): 25-29. (Zou Huiwen. Combined Effects of Non-rational Trade Behavior of Investors and Fluctuation of Stock Prices[J]. Journal of Fuzhou University: Philosophy and Social Sciences, 2008, 22(1): 25-29.)

[8] 史青春，徐露莹. 负面舆情对上市公司股价波动影响的实证研究[J]. 中央财经大学学报，2014(10): 54-62. (Shi Qingchun, Xu Luying. Empirical Research on the Listed Companies' Stock Prices Affected by Negative Public Opinion[J]. Journal of Central University of Finance & Economics, 2014(10): 54-62.)

[9] 于瑾，侯伟相. 杠杆交易、机构投资者行为与资产价格暴跌风险——来自股票市场的证据[J]. 金融监管研究，2017(12): 17-34. (Yu Jin, Hou Weixiang. Leveraged Transactions, the Behavior of Institutional Investor and the Risk of Asset Price Crash: Evidences from the Stock Market[J]. Financial Regulation Research, 2017(12): 17-34.)

[10] 岳衡，赵龙凯. 股票价格中的数字与行为金融[J]. 金融研究, 2007(5): 98-107. (Yue Heng, Zhao Longkai. Figures and Behavioral Finance in Stock Prices[J]. Journal of Financial Research, 2007(5): 98-107.)

[11] 吴璇，田高良，司毅，等. 网络舆情管理与股票流动性[J]. 管理科学, 2017, 30(6): 51-64. (Wu Xuan, Tian Gaoliang, Si Yi, et al. Internet Media Management and Stock Liquidity[J]. Journal of Management Science, 2017, 30(6): 51-64.)

[12] 林川. 过度投资、市场情绪与股价崩盘——来自创业板上市公司的经验证据[J]. 中央财经大学学报，2016(12): 53-64. (Lin Chuan. Excessive Investment, Market Sentiment and Share Prices Crash: Empirical Evidence from GEM Listed Companies[J]. Journal of Central University of Finance & Economics, 2016(12): 53-64.)

[13] 郭红玉，许争，佟捷然. 日本量化宽松政策的特征及对股票市场短期影响研究——基于事件分析法[J]. 国际金融研究, 2016(5): 38-47. (Guo Hongyu, Xu Zheng, Tong Jieran. The Characteristics of Japan's Quantitative Easing Policy and Its Short-Term Impact on Stock Market——Based on Event Analysis[J]. Studies of International Finance, 2016(5): 38-47.)

[14] 卢磊. 基于多元回归与技术分析的组合股票价格预测[J]. 上海应用技术学院学报: 自然科学版，2014, 14(3): 274-276. (Lu Lei. Combinational Stock Price Forecasting Based on Multiple Regression and Technical Analysis[J]. Journal of Shanghai Institute of Technology: Natural Science, 2014, 14(3): 274-276.)

[15] 陈璐璐. 基于多元线性回归分析的股价预测——以中信银行为例[J]. 经济研究导刊，2016(19): 75-76. (Chen Lulu. Based on Multivariate Linear Regression Analysis——Forecasting Stock Prices in China Citic Bank[J]. Economic Research Guide, 2016(19): 75-76.)

[16] 张建宽，盛炎平. 支持向量机对股票价格涨跌的预测[J]. 北京信息科技大学学报: 自然科学版，2017, 32(3): 41-44. (Zhang Jiankuan, Sheng Yanping. Prediction of Stock Price Fluctuation with Support Vector Machine[J]. Journal of Beijing Information Science & Technology University: Natural Science, 2017, 32(3): 41-44.)

[17] 黄宏运，吴礼斌，李诗争. BP 神经网络在股票指数预测中的应用[J]. 通化师范学院学报, 2016, 37(5): 32-34. (Huang Hongyun, Wu Libin, Li Shizheng. Application of Neural Network in Prediction of Stock Index[J]. Journal of Tonghua Normal University, 2016, 37(5): 32-34.)

[18] 魏文轩. 改进型 RBF 神经网络在股票市场预测中的应用[J]. 统计与决策，2013(15): 70-72. (Wei Wenxuan. Application of Improved RBF Neural Network in Stock Market Forecasting[J]. Statistics & Decision, 2013(15): 70-72.)

[19] 蔡红，陈荣耀. 基于 PCA-BP 神经网络的股票价格预测研究[J]. 计算机仿真, 2011, 28(3):365-368. (Cai Hong, Chen Rongyao. Stock Price Prediction Based on PCA and BP Neural Network[J]. Computer Simulation, 2011, 28(3): 365-368.)

[20] Göçken M, Özçalıcı M, Boru A, et al. Integrating Metaheuristics and Artificial Neural Networks for Improved Stock Price Prediction[J]. Expert Systems with Applications, 2016, 44: 320-331.

[21] 郭建峰，李玉，安东. 基于LM遗传神经网络的短期股价预测[J]. 计算机技术与发展，2017, 27(1): 152-155. (Guo Jianfeng, Li Yu, An Dong. Prediction for Short-term Stock Price Based on LM-GA-BP Neural Network[J]. Computer Technology and Development, 2017, 27(1): 152-155.)

[22] Adebiyi A A, Adewumi A, Ayo C. Comparison of ARIMA and

Artificial Neural Networks Models for Stock Price Prediction[J]. Journal of Applied Mathematics, 2014(1): 1-7.

[23] Evangelopoulos N, Magro M, Sidorova A. The Dual Micro/Macro Informing Role of Social Network Sites: Can Twitter Macro Messages Help Predict Stock Prices?[J]. Informing Science: The International Journal of an Emerging Transdiscipline, 2012, 15: 247-269.

[24] 王健俊, 殷林森, 叶文靖. 投资者情绪、杠杆资金与股票价格——兼论 2015-2016 年股灾成因[J]. 金融经济学研究, 2017, 32(1): 85-98. (Wang Jianjun, Yin Linsen, Ye Wenjing. Investor Sentiment, Leveraged Fund and Stock Price: Reflection on the Cause of Stock Crash in 2015-2016[J]. Financial Economics Research, 2017, 32(1): 85-98.)

[25] 石勇, 唐静, 郭琨. 社交媒体投资者关注、投资者情绪对中国股票市场的影响[J]. 中央财经大学学报, 2017(7): 45-53. (Shi Yong, Tang Jing, Guo Kun. The Study of Social Media Investor Attention and Sentiment's Influence on Chinese Stock Market[J]. Journal of Central University of Finance & Economics, 2017(7): 45-53.)

[26] 于琴, 张兵, 虞文微. 新闻情绪是股票收益的幕后推手吗[J]. 金融经济学研究, 2017, 32(6): 95-103. (Yu Qin, Zhang Bing, Yu Wenwei. Are the Emotions of News a Wire-puller of Stock Returns?[J]. Financial Economics Research, 2017, 32(6): 95-103.)

[27] 董理, 王中卿, 熊德意. 基于文本信息的股票指数预测[J]. 北京大学学报: 自然科学版, 2017, 53(2): 273-278. (Dong Li, Wang Zhongqing, Xiong Deyi. Stock Index Prediction Based on Text Information[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2017, 53(2): 273-278.)

[28] 黄润鹏, 左文明, 毕凌燕. 基于微博情绪信息的股票市场预测[J]. 管理工程学报, 2015, 29(1): 47-52. (Huang Runpeng, Zuo Wenming, Bi Lingyan. Predicting the Stock Market Based on Microblog Mood[J]. Journal of Industrial Engineering and Engineering Management, 2015, 29(1): 47-52.)

[29] Yan D F, Zhou J, Zhao X, et al. Predicting Stock Using Microblog Moods[J]. China Communications, 2016, 13(8): 244-257.

[30] 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.

[31] 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.

[32] 苏治, 卢曼, 李德轩. 深度学习的金融实证应用: 动态、贡献与展望[J]. 金融研究, 2017(5): 111-126. (Su Zhi, Lu Man, Li Dexuan. Deep Learning in Financial Empirical Application: Dynamics, Contributions and Prospects[J]. Journal of Financial Research, 2017(5): 111-126.)

[33] 韩豫峰, 汪雄剑, 周国富, 等. 中国股票市场是否存在趋势?[J]. 金融研究, 2014(3): 152-163. (Han Yufeng, Wang Xiongjian, Zhou Guofu, et al. Are There Trends in Chinese Stock Market?[J]. Journal of Financial Research, 2014(3): 152-163.)

[34] 金德环, 李岩. 投资者互动与股票收益——来自社交媒体的经验证据[J]. 金融论坛, 2017(5): 72-80. (Jin Dehuan, Li Yan. Investor Interaction and Stock Returns——Empirical Evidences of Social Media[J]. Finance Forum, 2017(5): 72-80.)

[35] 刘向强, 李沁洋, 孙健. 互联网媒体关注度与股票收益:认知效应还是过度关注[J]. 中央财经大学学报, 2017(7): 54-62. (Liu Xiangqiang, Li Qinyang, Sun Jian. Internet Media Coverage and Stock Returns: Investor Recognition or Over Attention[J]. Journal of Central University of Finance & Economics, 2017(7): 54-62.)

[36] 段江娇, 刘红忠, 曾剑平. 中国股票网络论坛的信息含量分析[J]. 金融研究, 2017(10): 178-192. (Duan Jiangjiao, Liu Hongzhong, Zeng Jianping. Analysis on the Information Content of China's Internet Stock Message Boards[J]. Journal of Financial Research, 2017(10): 178-192.)

[37] 杨晓兰, 沈翰彬, 祝宇. 本地偏好、投资者情绪与股票收益率: 来自网络论坛的经验证据[J]. 金融研究, 2016(12): 143-158. (Yang Xiaolan, Shen Hanbin, Zhu Yu. The Effect of Local Bias in Investor Attention and Investor Sentiment on Stock Markets: Evidence from Online Forum[J]. Journal of Financial Research, 2016(12): 143-158.)

[38] Huang Y, Qiu H, Wu Z. Local Bias in Investor Attention: Evidence from China's Internet Stock Message Boards[J]. Journal of Empirical Finance, 2016, 38: 338-354.

[39] Rätsch G, Onoda T, Müller K R. Soft Margins for AdaBoost[J]. Machine Learning, 2001, 42(3): 287-320.

[40] Safavian S R, Landgrebe D. A Survey of Decision Tree Classifier Methodology[J]. IEEE Transactions on Systems, Man and Cybernetics, 2002, 21(3): 660-674.

[41] Guo G, Wang H, Bell D, et al. KNN Model-Based Approach in Classification[J]. Lecture Notes in Computer Science, 2003, 2888: 986-996.

[42] Rish I. An Empirical Study of The Naive Bayes Classifier[C]// Proceedings of the 2001 Workshop on Empirical Methods in Artificial Intelligence. 2001, 3(22): 41-46.

[43] Hearst M A, Dumais S T, Osuna E, et al. Support Vector Machines[J]. IEEE Intelligent Systems & Their Applications, 1998, 13(4): 18-28.

作者贡献声明:

余传明：提出研究思路，实施及优化技术方案与路线，数据收集，论文修改；
龚雨田：数据收集和处理、实验操作，论文初稿撰写；
王峰：实验操作，论文初稿撰写；
安璐：提出研究思路，数据分析，论文修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: yucm@zuel.edu.cn。
[1] 余传明. data.zip. 股票价格和股票评论文本对应的训练集和测试集数据.

收稿日期: 2018-04-16
收修改稿日期: 2018-06-13

Predicting Stock Prices with Text and Price Combined Model

Yu Chuanming[1] Gong Yutian[1] Wang Feng[1] An Lu[2]
[1](School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China)
[2](School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: [Objective] This paper tries to predict stock price fluctuation with the help of big data, aiming to improve the accuracy of the forecasting and reduce the trading risks. **[Methods]** We proposed a new Text and Price Combined Model (TPCM) to process comments retrieved from a stock forum. Then, we employed deep representation learning algorithm to generate text feature matrix and utilized the K-means clustering method to generate text category. Finally, we used the Multi-Layer Perceptron (MLP) to predict stock price fluctuation based on the opening price, closing price and other 15 original price indicators. **[Results]** The accuracy of TPCM was 65.91%, which was 7.76% higher than that of the model (58.15%) employing price features only, and 11.37% higher than that of the model (54.54%) employing text features only. **[Limitations]** The study only used one stock to examine the proposed model. **[Conclusions]** Stock price forecasting could be improved through the combination of text and price, which creates novel perspectives for future studies.

Keywords: Text Stock Price Stock Price Fluctuation Prediction Text and Price Combined Model