山西大学学报(自然科学版)41(1):1－14,2018
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.cnki.shanxi.univ(nat.sci.).2018.01.001

* 基于 BPNN 和 SVR 的股票价格预测研究

冉杨帆,蒋洪迅*
(中国人民大学 信息学院,北京 100872)

摘　要:将情感分析和机器学习方法相结合,以股票新闻数据为基础,分别采用 BP 神经网络(BPNN)和支持向量机回归(SVR)两种方法,对股票价格进行预测分析。首先选取交易量较大的 20 只股票作为研究对象,抓取相关的新闻数据。然后邀请专家对高频词进行人工情感打分,得到一个针对性更强、粒度更细[－5,＋5]的情感词典,同时考虑否定词、程度副词、假设疑问词和情感词间的相互作用,归纳出 9 种常见的语义规则,给不同的语义规则下的情感词赋予不同的权重,对情感值进行修正。最后分别采用 BPNN 和 SVR 两种方法构造股价预测模型,并对模型的预测效果进行对比分析。结果表明,文章提出的人工情感词典和语义规则在股价预测领域表现良好,情感得分正负方向与股价涨跌方向的一致程度显著提升,另外,SVR 股价预测模型的均方误差更小,且股价走势方向正确率更高。

关键词:股票价格;BP 神经网络;支持向量机回归;情感分析
中图分类号:TP183　　**文献标志码**:A　　**文章编号**:0253-2395(2018)01-0001-14

Stock Prices Prediction based on Back Propagation Neural Network and Support Vector Regression

RAN Yangfan, JIANG Hongxun*
(*School of Information, Renmin University of China, Beijing* 100872, *China*)

Abstract: Prediction models combining with sentiment analysis and machine learning method are proposed to predict the price of the stock based on the stock news and the machine learning algorithm of Back Propagation Neural Network (BPNN) and Supportive Vector Regression (SVR). 20 stocks with high transaction amount are chosen and the related news data are crawled. The words with high frequency are scaled and ranked to produce a preciser sentiment dictionary with granularity [－5, ＋5]. Moreover, considering the 9 different combination among the negative words, adverb of degree, hypothesis words and the sentiment words, semantic rules are generated with different weight to the sentiment words to modify the emotion scale. Finally, the prediction model is built on the BPNN and SVR algorithm and a comparison is conducted. The result shows that the sentiment dictionary and semantic rule performs well in predicting the price and a more accurate match between the emotion scale and the price trend is achieved. Moreover, the SVR algorithm can produce a higher accuracy and less mean squared error.

Key words: stock price; Back Propagation Neural Network; Support Vector Regression; sentiment analysis

* **收稿日期**:2017-06-24;**接受日期**:2017-09-14
基金项目:国家自然科学基金(71571183);教育部人文社会科学基金(12YJA630046)
作者简介:冉杨帆(1995－),男,硕士研究生,主要研究方向为机器学习、社交网络数据挖掘。E-mail:yf_ran@163.com
* **通信作者**:蒋洪迅(JIANG Hongxun),E-mail:jianghx@ruc.edu.cn
引文格式:冉扬帆,蒋洪讯. 基于 BPNN 和 SVR 的股票价格预测研究[J]. 山西大学学报(自然科学版),2018,**41**(1):1-14.

0 引言

股票市场发展至今已有400多年的历史。股票市场反映了国民经济的发展情况,被称为金融市场的"晴雨表"和"预警器"。在我国,股票市场在资本融通、财富再分配、优化资源配置、金融资产价格发现等方面有着重要的作用[1]。20世纪90年代,沪、深两家交易所陆续成立,我国的股票市场进入高速发展阶段,但是由于起步较晚,制度尚不健全,在飞速发展的同时也暴露出很多问题,如股市容易大幅度波动,产生暴涨或暴跌现象,存在大量人为恶意操纵等。这些现象对广大投资者,乃至我国国民经济的发展有着极大的危害,因此,掌握股票市场的变化规律,是股票市场健康发展的重要前提。

随着人们生活水平不断提高,许多人开始将闲钱用于投资理财,进入股票市场。与发达国家相比,我国的投资者多为个人,机构投资者占比较少,大多数股民风险承担能力较弱,股票市场的异常波动对广大投资者有着巨大的损害。一直以来,股票市场都需要一种比过去更高效的投资理论,为广大投资者提供更好的投资方法,以提高投资收益、规避投资风险[2]。

应用现有的技术在一定程度上实现对股票价格的预测对股票市场、投资者都具有十分重大的意义。然而,股票市场是一种极其复杂的系统,由于股票市场影响因素的多样性和不确定性,以及股票市场特有的高噪声和非线性的特性,想要得到高精确度的预测结果十分艰难。

随着大数据时代的到来和海量数据库的发展,对股票市场每天产生的海量数据加以利用变为可能,越来越多的学者开始考虑使用更多的数据来源,如舆论、股票新闻等,各种文本挖掘和机器学习的方法也被广泛地应用到了股票研究中,主流的机器学习方法,如支持向量机、神经网络在对复杂信息的综合处理上表现良好,可以克服传统预测方法中的许多局限。本次研究将新闻文本作为输入之一,通过人工构建专门针对股票市场的情感词典和语义规则,计算每篇新闻的情感值,并分别使用BP神经网络和支持向量机回归的方法建立股价预测模型。本次研究是一次很有意义的尝试和探索,可为情感分析、机器学习等技术应用于股票市场价格预测提供一定的理论价值和实践价值。

股票市场是一个极其复杂的系统,股票的价格受政治、经济、行业、公司、人为因素和心理因素等诸多因素影响[3]。股票系统的复杂性、多变性导致股票价格的变化通常具有较高的非线性特征,而传统的预测方法大多是基于线性模型,不能很好地反映这些因素对股价的影响[4]。

20世纪60年代,许多学者将股市中的历史数据按时间先后顺序列出,建立起时间序列预测模型,该模型可以较为准确的预测股票市场未来几天的变化。这些时间序列模型都着眼于历史市场数据,通过观察历史的变化规律来推测股价未来的走势,较为出名的有移动平均模型(MA)[5],指数平滑和ARIMA模型[6]。

一些研究者认为,股票市场是一种极其复杂的系统,股票价格预测的输入不应只包含股票市场历史的数据,还应该包含其他类型的数据。随着大数据时代的到来和海量数据库的发展,使得股票市场每天产生的海量数据的使用变为可能,越来越多的学者开始考虑更多的数据源,把股票舆论、股票新闻信息作为输入加入预测模型,各种文本挖掘和机器学习的方法也被广泛地应用到了股票研究中。另外,许多学者开始将情感分析和机器学习相结合,来预测股票价格,都取得了不错的效果。

将神经网络方法运用到股票和指数价格预测领域,最早可追溯到20世纪90年代[7],1996年Gencay等[8]建立了向前人工神经网络模型,先将道琼斯工业评价指数的历史数据进行移动平均,得到其平均价格作为输入,然后用该模型对1967年到1988年的指数进行预测,研究结果显示,神经网络模型的预测效果优于简单的移动平均法,不过,当时的预测效果还达不到实际应用的要求。2003年,Zhang等[9]把神经网络模型和时间序列模型ARIMA进行了对比实验,得出在非线性数据的处理方面,神经网络模型的预测精度好于ARIMA模型。Murat等[10]使用神经网络模型对TKC证券进行预测,将趋势、波动性和动力等不同类型的数据作为输入,并使用2006年的数据进行测试,研究结果显示神经网络模型的预测效果优于其他模型,但该研究没有解释模型输入向量的甄选理由,存在较强的主观性。另外,一些学者将小波包方法和神经网络方法进行结合,例如张坤、郁湧等[11]把股价进行小波分解,生成多个不同尺度的分层数据,然后先预测得到各层的预测结果,再把各层的预测结果作为输入,利用BP神经网络得出最终的预测结果,最终该模型能取得了较好的预测效果。

支持向量机模型(Support Vector Machine,简记为 SVM)是另外一种经常被用于股票价格预测研究的机器学习模型,SVM 方法使用结构风险最小化原理,具有较好的泛化能力[12]。1995 年,Vapnik 出版了《The Nature of Statistical Learning》一书,标志着支持向量机理论的成熟,经过二十多年的发展,SVM 相关的理论和方法得以完善和丰富。2003 年,Kim[13]把支持向量机回归模型(SVR)用于股票价格预测中,并和 BP 神经网络模型进行对比,发现 SVR 的精确度好于 BP 神经网络。2005 年,Huang 等[14]将支持向量机用于日经 225 指数的预测研究中,得出其预测精确度比神经网络方法更高。2006 年 Xu 等提出了改进的最小二乘支持向量机(简称 LS-SVM),优化参数后用于对纳斯达克指数进行预测,得到了令人满意的预测结果。2013 年,施剑[15]把支持向量机算法应用到新股 IPO 首日价格变动预测上,解决了其他预测方法对历史数据具有较大依赖性的问题,为今后的研究提供了有力参考。2014 年张世军[16]将网络舆情加入 SVM 模型,获得了更佳的预测效果。2014 年龙真真等[17]在 SVM 模型的基础上,加入模糊核超球快速分类算法建立预测模型,研究表明该模型能较为准确地预测沪市上市公司的回报率。从一些学者的研究结果中可以看出,在他们的研究所适用的范围内,SVM 模型的预测效果优于神经网络模型。

综上所述,国内外多数研究的焦点,集中在机器学习模型的优化,或是将原有模型与其他算法相结合以求得到更好的预测效果,把网络新闻文本数据作为输入加入股价预测模型的研究尚不多见。

1 数据采集与预处理

1.1 新闻数据的采集

股票市场是一个高收益与高风险并存的场所,投资者愿意花大量的时间去了解它们所关注的公司的动态。各大财经新闻网、证券新闻网是大多数投资者的重要信息来源,它们所发布的新闻、评论也深刻地影响着投资者的决策。

根据影响力、重要性、原创性对各大新闻网址进行对比分析,划定新闻文本的采集范围,选取具有代表性的股票新闻门户网站作为数据源。同时,尽可能获取较长时间段的新闻数据用于研究。

最后选定了中证网、中华财会网、华夏时报、证券时报网、中国财经新闻网五家权威网站作为数据来源,抓取的时间跨度从 2008 年 1 月 1 日到 2016 年 12 月 31 日。通过编写 Java 爬虫进行抓取,抓取的字段有:标题、网址、发布时间、内容。其中,对中证网、中华财会网、证券时报网、中国财经新闻网采用搜索抓取方式,即先按关键词进行搜索,抓取其搜索结果,华夏时报由于新闻量不大,采取直接搜索方式。最后爬取到新闻共 1 447 440 篇,共 7 GB 左右。

1.2 股价数据的采集

本次实验选取了 20 只股票作为研究对象。在选择时,交易量是最重要的因素,高交易量的股票一般发行量也大,也间接说明了发行公司实力雄厚,这样的公司一般更易受到投资者的关注。同时,高交易量也反映了股票的不稳定性,这意味着有更多的新闻报道它们,当股票价格不稳定时,新闻报道也更容易对投资者的决策产生影响。

通过 Wind 客户端导出成交量最大的前 20 只股票,然后再分别导出每只股票从 2008 年 1 月 1 日到 2016 年 12 月 31 日每天的价格。最后,通过关键词匹配将新闻数据和各只股票进行简单的匹配,删去重复项、空值后,得到的数据集如表 1 所示。

1.3 文本分词

中文分词的方法主要分为三类[18]。第一类是基于词典的分词方法,该方法是依据现有词典进行正向或逆向的匹配,其特点是分词速度快,同时能保持较高的准确率,但对词典中没有的新词识别能力差;第二类是基于统计的分词方法,该方法是通过统计字与字相邻一起出现的频次,两个字相邻出现的次数多则可能组成一个词语;第三类是基于理解的分词方法,该方法是用计算机模拟人脑的功能,这种方法需要的信息较多,实现起来相对复杂。

表 1 数据抓取结果

Table 1 Results of data crawling

股票代码	名称	关键词	数量
601988	中国银行	中国银行中行 601988	70 028
000725	京东方 A	京东方 A 000725	4 099
601288	农业银行	农业银行 农行 601288	39 793
601668	中国建筑	中国建筑中建 601668	18 826
600795	国电电力	国电电力 600795	5 964
601989	中国重工	中国重工 601989	6 501
601398	工商银行	工商银行 工行 601398	72 875
600028	中国石化	中国石化中石化 600028	59 639
601818	光大银行	光大银行 601818	16 929
600050	中国联通	中国联通联通 600050	36 528
601328	交通银行	交通银行交行 601328	40 304
000100	TCL 集团	TCL 集团 000100	7 871
600010	包钢股份	包钢股份 包钢 600010	12 830
600016	民生银行	民生银行 600016	33 741
600030	中信证券	中信证券 600030	52 130
601766	中国中车	中国中车中车 601766	8 007
601390	中国中铁	中国中铁中铁 601390	16 971
601899	紫金矿业	紫金矿业 601899	7 014
601166	兴业银行	兴业银行 601166	31 708
600221	海南航空	海南航空 海航 600221	11 065

常见的分词工具有 IKAnalyzer、结巴分词、最大熵分词、ICTCLAS 等。本次采用中科院张华平博士团队研发的 ICTCLAS2016 工具(http://ictclas.nlpir.org/),对新闻文本的标题和内容进行中文分词。ICTCLAS2016 工具包括中文分词、用户自定义词典、词性标注、命名实体识别等功能,并且提供 Java、C++、C# 等多种版本,在分词方面具有出色的表现。

1.4 去停用词

为了提高处理效率和节省存储空间,在文本处理之前通常会对一些无信息量的词进行过滤,这些词就是停用词。停用词通常有以下两个特点:一是停用词通常为功能词,不具有实际含义,如“是”、“在”;二是停用词通常使用十分广泛,如“我”、“他们”。在本阶段去除停用词,能大大加快之后数据处理效率。

目前使用较为广泛的停用词表有:哈工大停用词词库、百度停用词表、四川大学机器学习智能实验室停用词库等。通过比较分析,最后选用了哈工大的停用词词库。通过去除停用词,所有新闻总词数减少了 15%左右。

2 股价预测模型的实现

2.1 新闻情感分析

2.1.1 构造情感词典

情感分析的方法主要可分为两种[19],一是基于情感词典的分析方法,其主要有人工构造情感词典或直接引用权威的情感词典两种方式,得出文中每个词的情感倾向(积极、消极、不相关)。二是基于机器学习的分析方法,即运用机器学习的方法进行情感分析。

目前,学术研究中使用较为广泛的中文情感词典有以下三种:中国知网 HowNet 中包括正面情感词语、正面评价词语、负面情感词语、负面评价词语 4 个词语表,词语较为全面,但其仅是把情感词语分入相应的词表,未对词语的情感强度进行描述。台湾大学的 NTUSD 情感词典由 2 810 个正向情感词和 8 276 个负向情感词构成,但其也未对词语的情感强度进行描述。以上两个情感词典都仅是把各情感词进行正负向分类,不能很好地区分同类情感词之间的程度差异。大连理工大学信息检索研究室的情感词汇本体库(以下简称 DUT 情感词典)包含正向、负向、中性三类词语共 27 466 个,并且对每个词的词性、情感极性和强度都进行

了详细地描述,它把情感强度分为 0－9 十个程度,适合用于新闻情感值的计算。

在使用 DUT 情感词典进行情感值计算的过程中,笔者发现,虽然该词典情感词数量较多,却不适合用于股票新闻。以"中国银行"相关的 70 028 条新闻数据为例,经统计,经过分词、去停用词后,新闻平均每篇 760 词,使用大连理工大学情感词典进行匹配,平均每篇新闻匹配到的正向和负向情感词的次数之和不足 20 次。另外,DUT 情感词典缺少很多股票市场常见情感词,如"暴跌"、"下降"、"上涨"、"萎缩"等。

因此,本次研究考虑通过人工打分的方式构造一个专门针对股票市场的情感词典。首先,对所有新闻文本进行词频统计,按词语出现的总次数从高到低进行排序,选出出现次数最多的前 3 000 个词(出现的总次数占文本总词数的比重超过 90%)进行情感值打分。分值区间为[－5,＋5],分值大于 0 说明该词语为积极正向的情感词,且分值越高,正向情感强度越大,反之分值小于 0 则说明该词语为消极负向的情感词,分值越低,负向情感强度越大,分值为 0 表示该词语为中性情感词。人工打分邀请了 3 位该领域的专家(证券从业人员、股民)分别进行,最后计算平均值即为该词最终的情感得分。

2.1.2 构造语义规则

相同的情感词在不同的语句环境中所表达出的情感正负、强度可能完全不同,例如"股票价格上涨"、"股票价格稍稍上涨"两句中,由于副词不同,两个句子中"上涨"一词所表达的情感强度发生变化。又如"股票价格如果上涨"、"股票价格上涨",前者由于表假设语气的词"如果"的出现,使得股票价格是否上涨变为不确定。又如"股票价格不会上涨"、"股票价格上涨",前者由于否定词"不会"的出现,表达的情感完全不同。

因此,本次研究希望通过构造相关的语义规则来更深层次地挖掘情感词在各句子环境中的真实情感。杨希[20]在对微博文本进行情感分析的研究中,总结出 6 种常见的语义规则,结合实际情况,对这 6 种语义规则进行扩展,得到如下规则:

(1)只有情感词在句子中起情感表达作用。如"股市状况良好",只有"良好"表达出积极正向的情感,"股市"、"状况"为中性情感。

(2)程度副词加情感词。程度副词对情感的表达有加深或减弱两种作用,如"今天很开心","开心"表达出积极正向的情感,"很"作为程度副词,使积极的情感加深。又如"经济稍稍好转","稍稍"作为程度副词,反而使"好转"表达出的正向情感减弱。

(3)否定词加情感词。如"没有好转","没有"是否定词,在句子中对情感词"好转"所表达的正向情感起了相反作用,最后句子的情感变为负向情感。

(4)否定词加程度副词加情感词。如"没有一点改善","改善"表达积极正向的情感,"一点"对情感词的积极强度起减弱作用,"没有"对"一点改善"的组合起相反作用。

(5)程度副词加否定词加情感词。如"很不乐观","乐观"表达积极正向的情感,"不"对情感词起相反作用,"很"对"不乐观"的组合起增强作用。

(6)多个否定词加情感词。如"并非不乐观","乐观"表达积极正向的情感,"不"对情感词起相反的作用,"并非"对"不乐观"的组合起再次相反的作用,最后该组合表达的情感仍为正向情感。

(7)假设疑问词加情感词。如"一旦上涨","上涨"一词表达出正向情感,"一旦"一词表示假设推断,表达出不确定性,对情感词进行修正后,不能明确得知是否会上涨,因此,表达的情感为中性。

(8)假设疑问词加否定词加情感词。如"万一不乐观","不乐观"组合表达出消极负面的情感,"万一"表达假设、不确定,因此单单根据这一句话,不能推断最后情感为积极或是消极。

(9)假设疑问词加程度副词加情感词。如"如果很差",同理,"很差"组合经过"如果"修正后,表达的情感也不能确定。

以上仅列出了情感词和程度副词、否定词、疑问词间的常见组合,还有部分复杂的组合未全部列出。不难发现,程度副词对情感词表达的情感起加强或减弱的作用,否定词起相反作用,假设疑问词则是使表达的情感变为不确定。

根据知网 HowNet 中的程度级别词语(中文),结合新闻数据的词频统计情况,人工进行一定的简化,留下日常使用较多的副词,按 HowNet 里的划分分为 6 个级别,并给其赋予了不同的权重,具体如表 2 所示。同时,结合新闻文本的词频统计情况,选取常用的 20 个否定词,构建否定词表,如表 3 所示。

表 2 程度副词词表

Table 2 List of adverb of degree

程度	程度副词	权重
极其(extreme)、最(most)	倍加、充分、非常、极、极其、极为、尽、截然、绝、绝对、满、十分、十足、完全、万般、异常、最为、最	2
很(very)	不少、不胜、多么、很、很是、颇、颇为、太、特别、特、尤、尤其、尤为	1.5
较(more)	更、更加、更为、较、较为、那么、那样、远、远远、越来越、这样、这般	1.3
稍微(－ish)	略、略微、略为、稍、稍稍、稍微、相当、一些、一点、有些、些微	1.1
欠(insufficiently)	不大、不甚、不怎么、没怎么、轻度、微、丝毫、相对	0.8
超(over)	超、超额、过度、过分、过于、偏	1.7

表 3 否定词词表

Table 3 List of negative words

否定词	权重
不、不能、不要、不比、不必、不可、不会、不得、别、莫、没有、没、未、未必、并未、非、并非、无、无法、难以	－1

最后,结合词频统计情况,选取常用的 22 个表假设疑问的词,构建假设疑问词表,如表 4 所示。

至此,已经构造好了情感词典、程度副词表、否定词表、假设疑问词表和语义规则,接下来即可进行新闻文本的情感值计算。

表 4 假设疑问词词表

Table 4 List of hypothesis words

表假设疑问的词	权重
如果、如若、假如、假若、假使、假设、若、若是、倘若、倘使、猜测、猜想、或许、也许、万一、一旦、只要、是否、能否、以免、避免、否认	0

2.1.3 计算新闻情感值

新闻文本的情感值计算分为两步:首先计算每个句子的情感值,然后进行加和即可得到整篇文章的情感值。

假设 e 代表句子的情感值,p 代表情感词的情感强度,w 代表程度副词的权重,根据上节中定义的语义规则,可以得到情感值计算规则,如表 5 所示。

表 5 情感值计算规则

Table 5 Rules of sentiment scale calculation

序号	语句结构	情感值计算公式
1	只有情感词	e＝p
2	程度副词＋情感词	e＝w＊p
3	否定词＋情感词	e＝(－1)＊p
4	否定词＋程度副词＋情感词	e＝(－1)＊w＊p
5	程度副词＋否定词＋情感词	e＝w＊(－1)＊p
6	否定词＋否定词＋情感词	e＝(－1)＊(－1)＊p
7	假设疑问词＋情感词	e＝0＊p
8	假设疑问词＋否定词＋情感词	e＝0＊(－1)＊p
9	假设疑问词＋程度副词＋情感词	e＝0＊w＊p

考虑到程度副词在语句中可能位于情感词前,也可能位于情感词后(如“价格没怎么上涨”和“价格上涨程度不大”),所以在具体实现时,先定位到正向情感词和负向情感词所在的位置,然后从情感词所在位置开始向前和向后扫描,直到出现“,”“。”“!”“?”“;”等符号为止,如果出现程度副词、否定词或假设疑问词,则按上面的计算规则对情感词的情感值进行修正,即可得到该句子的情感值。

对于一条新闻文本来说,先得到包含正负向情感词的子句的情感值 Ei(其中 E_0 表示标题的情感值),然后根据下述公式,即可计算出整条新闻的情感值。

$$E = 5 * E_0 + \sum_{i=1}^{n} E_i \tag{1}$$

最后,再把同一天的所有新闻的情感值进行加总,即可得到每天新闻数据的总情感得分,另外,对于当天无新闻的,情感值置为零。

2.2 模型搭建

2.2.1 实验平台的选取

为了实现 BP 神经网络和支持向量机回归两种模型,本次实验采用 MatlabR2014a 为实验平台,操作系统环境为 WIN7。

在建立 BP 神经网络股价预测模型时,直接运用 Matlab 自带的工具包,而在建立 SVR 股价预测模型时,运用台湾大学的林智仁教授编写的 LibSVM 工具包,该工具包使用 Matlab 和 C++混合编程,比 Matlab 自带的 SVM 工具包运行速度更快,功能更齐全,模型中参数的修改也更加容易。

2.2.2 数据归一化

经过数据爬取、预处理、情感分析,最终得到的用于实验的数据共包括两部分:一部分是股价历史数据{P},通过股票价格序列可以得到该股票 T-1 日的价格序列、T-2 日的价格序列、T-3 日的价格序列等,另一部分数据是通过对新闻文本进行情感挖掘得到的情感值序列{E}。由此,可以得到模型的两种输入向量,其中一种为混合数据,包括滞后期为 n 的股票价格数据和 T-1 日的新闻文本的情感值,可表示为 $\boldsymbol{X}=\{p_{t-1},p_{t-2},\cdots,p_{t-n},e_{t-1}\}$,另一种输入向量仅包含股票数据,可表示为 $\boldsymbol{X}=\{p_{t-1},p_{t-2},\cdots,p_{t-n}\}$。关于两种输入向量的优劣和滞后期 n 的选择将在 2.2.3 节中进行说明。

在进行 BP 神经网络模型训练之前,通常需要先将数据进行归一化处理[21],其主要原因如下:(1)部分数据的范围可能特别大,可能会导致网络收敛速度减慢、学习时间过长;(2)数据范围和量级较大的数据项在模型中的作用可能会偏大,而数据范围较小的数据作用则可能会偏小;(3)在输出层中,激活函数的值域是有限制的,所以需要把训练的目标数据都映射到激活函数的值域中去。本次实验采用 Log-Sigmoid 函数,因此将数据归一化到[0,1]区间里去。

在建立支持向量机回归模型时,归一化同样有助于加快训练速率、降低因数量级的差异所造成的影响[22],因此同样在实验前对数据进行归一化处理。

2.2.3 输入向量的选择

股票市场瞬息万变,新闻文档的时间滞后效应是很短的,我们需要关注的是股票数据(即股票价格)的滞后天数。设置的时间滞后期越大,所包含的信息也就越多,模型可能表现更好,但是,过大的滞后期也会增加模型的复杂程度,影响计算速率。因此,寻找最优滞后期对模型至关重要。

均方误差(Mean Squared Error,简称 MSE)是数据预测中常用的评价指标,MSE 的值越小,误差的离散程度越小,预测效果也就越好。MSE 计算公式如下:

$$\mathrm{MSE} = \frac{1}{n}\sqrt{\sum_{t=1}^{n}(y_t - y'_t)^2} \tag{2}$$

分别将混合数据和股票数据作为输入向量,比较两种输入的优劣。在每次实验中,将数据分成训练集和测试集两部分,比例为 3∶1,将滞后天数从 1 逐渐增到 20,以 MSE 为评价指标,每次试验均重复进行 10 次,取其平均值为最终结果。

本次对比实验统一采用三层网络模型,隐藏层节点个数统一设置为 6。通过图 1 可以看出,当输入数据为混合数据时,随着滞后天数的增大,MSE 先呈下降趋势,当滞后天数增长到 3 日后,MSE 停止下降,随后在 0.004 1 到 0.006 4 间上下波动。当输入数据为股票数据时,MSE 也先随着滞后天数的增大呈现下降趋势,当滞后天数增加到 6 时停止下降,随后在 0.003 3 到 0.005 5 间上下波动。可以看出,BP 神经网络模型的 MSE 不稳定,波动较大,即使在完全相同条件下,得到的 MSE 也差别较大,迭代次数也不尽相同。在 BP 神经网络模型中,混合数据的表现略差于股票数据,说明新闻数据在神经网络模型中没能为预测带来有效的信息输入,所以在建立 BP 神经网络预测模型时,均采取股票数据作为输入。从股票数据的曲线可以看出,当滞后天数增加到 6 天后,继续增加滞后天数未能使模型效果得到较为明显的改善,较高的维度不仅会降低

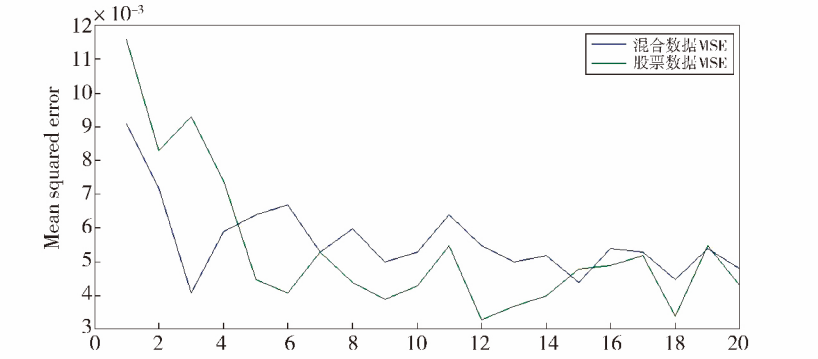

Fig. 1 MSE diagram under different time lag(BPNN)

图 1 不同滞后期下的 MSE 图(BPNN)

BP 神经网络模型的学习效率,而且会影响训练后的预测效果。因此,在接下来的实验中,选择 6 天作为滞后天数。

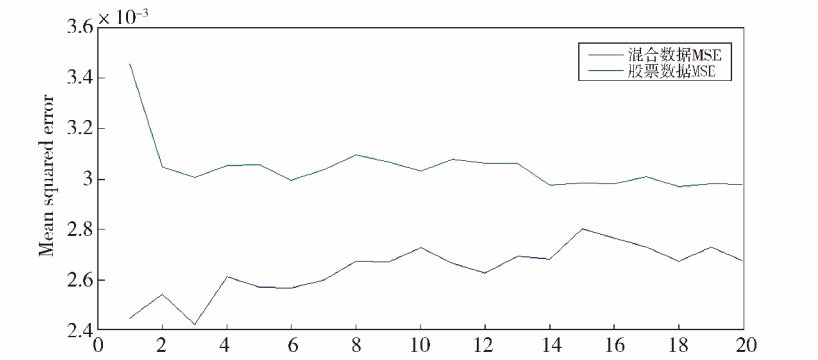

Fig. 2 MSE diagram under different time lag(SVR)

图 2 不同滞后期下的 MSE 图(SVR)

在 SVR 模型的对比实验中,参数均采用默认参数。通过图 2 可以看出,从滞后天数增加到 4 天开始,混合数据和股票数据的 MSE 均较平稳,且混合数据的 MSE 均低于股票数据的 MSE,所以在使用 SVR 模型建立股价预测模型时,采用混合数据作为模型的输入,由于支持向量机算法能避免"维数灾难",为包含更多的信息,SVR 模型中滞后天数选为 20 天。

2.2.4 模型参数寻优

在 BP 神经网络模型中,选择合适数量的隐含层节点非常重要,对网络性能影响非常大。隐含层节点数目设定得过少,容易使训练过程不收敛,节点数目设定过多,又容易训练过度,使学习时间过长[23]。隐含层节点数目的设定是一个复杂的问题,尚没有权威的计算方法,如今多采用多次训练取最优的方法,开始时先选择一个较小的估计数,保持其他条件不变,逐渐地增加节点数目,反复地进行训练和测试,最后误差最小时所对应的隐含层节点数目即为最佳节点数。现有的计算公式计算出来的结果都只是经验上的估计量,并不一定是最佳的节点数,常用的计算公式主要有以下三种:

$$m = \sqrt{nl} \tag{3}$$

$$m = \log 2n \tag{4}$$

$$m = \sqrt{n+l} + a \tag{5}$$

其中,m 代表隐含层节点数目,n 是输入层的节点数目,l 是输出层的节点数目,α 是 1 到 10 之间的常数。本文首先选用公式(5)来确定一个大致范围,保持其他条件不变,仅改变隐含层节点数反复进行实验,通过比较 MSE 的大小来确定最优的隐含层节点数目。

由于 n 为 6(BPNN 模型输入向量的选取已在 2.2.3 节中说明),l 为 1,所以隐含层节点数目的范围可初步确定为[3,13]。实验采用三层网络结构,训练目标误差设为 e^{-5},学习速率设为 0.01,每次实验 10 次取平均值,最终结果如表 6 所示。在相同训练条件下,隐含层节点数目设为 4 时,MSE 最小。

在应用 SVR 方法时,需要设置很多参数,参数设置是否合理会对模型的预测效果产生巨大的影响。本次实验选取了两个最普遍接受的参数 c 和 g(gamma)进行参数寻优,诸多研究表明,这两个参数的设定对最终的训练效果有最显著的影响。c 是模型的惩罚因子,表示模型对误差的宽容程度,c 的值越高,说明模型对误差的容忍程度越低。gamma 是核函数的一个参数,它隐性地决定了原数据映射到高维特征空间后的分布情况。

表 6 不同隐含层节点数下的 MSE 值(BPNN)

Table 6 MSE value under different number of hidden layer nodes(BPNN)

隐含层节点个数	MSE	隐含层节点个数	MSE
3	0.003 9	9	0.008 5
4	0.003 6	10	0.007 8
5	0.004 1	11	0.006 7
6	0.004 1	12	0.004 5
7	0.006 5	13	0.008 5
8	0.004 3		

关于 SVR 的参数寻优,国际上并没有公认的最好方法,常见的寻优方法有交叉验证—网格搜索方法(Grid Search)、遗传算法(GA)、启发式算法(PSO)。本文使用 Grid Search 方法寻找最优的 c 和 g,该方法虽然简单,但却有以下两个优点:可以得到全局最优;c 和 g 相互独立,便于并行化进行。通过 Grid Search 方法,得到每只股票最优的 c 和 g,如表 7 所示,在后续的实验中,参数均采用该只股票最优的 c 和 g 进行实验。

表 7 20 只股票的最优参数(SVR)

Table 7 Optimal parameters of 20 stocks(SVR)

序号	股票名称	Best c	Best g	MSE
1	中国银行	256	0.003 9	0.000 718
2	京东方 A	8	0.003 9	0.001 473
3	农业银行	0.5	1	0.001 552
4	中国建筑	32	0.015 6	0.000 649
5	国电电力	128	0.003 9	0.000 700
6	中国重工	64	0.007 8	0.000 751
7	工商银行	8	0.25	0.000 911
8	中国石化	16	0.015 6	0.000 828
9	光大银行	256	0.003 9	0.001 730
10	中国联通	128	0.003 9	0.001 495
11	交通银行	4	0.5	0.000 922
12	TCL 集团	64	0.015 6	0.000 558
13	包钢股份	128	0.003 9	0.000 635
14	民生银行	0.5	0.25	0.001 218
15	中信证券	0.5	2	0.001 161
16	中国中车	0.5	4	0.002 163
17	中国中铁	0.5	2	0.002 420
18	紫金矿业	32	0.031 3	0.000 732
19	兴业银行	1	1	0.000 920
20	海南航空	1	1	0.000 956

3 实验与结果分析

3.1 情感词典效果对比分析

为对比大连理工大学(DUT)情感词典和人工情感词典在股票市场价格预测领域的适用性,做了以下对比实验,实验以“中国银行 601988.SH”为例,时间从 2008 年 1 月 1 日到 2016 年 12 月 31 日,除去周末和法定节假日,共计 2 190 天。

分别运用大连理工大学情感词典和人工情感词典对新闻文本进行打分,得到每篇新闻的情感值,加总同一天的所有新闻得到当天的总情感值,然后与股票价格的涨跌方向进行对比。

表 8 大连理工大学情感词典测试效果

Table 8 Performance on DUT sentiment dictionary

DUT 情感词典+语义规则预测结果	情感值为正	情感值为负	情感值为零	正确率
股价上涨	919	1	0	99.8%
股价下跌	973	0	0	0%
股价不变	297	0	0	0%

表 9 人工情感词典测试效果

Table 9 Performance on artificial sentiment dictionary

人工情感词典+语义规则预测结果	情感值为正	情感值为负	情感值为零	正确率
股价上涨	587	333	0	63.8%
股价下跌	504	469	0	48.2%
股价不变	182	114	1	0%

如表 8 和表 9 所示,运用大连理工大学情感词典进行打分得到的情感值几乎都为正,当股价下跌和股票不变时,不能很好地进行预测。而针对股票市场进行人工打分得到的情感词典,在股票上涨、股票下跌时,分值与涨跌方向保持了较好的一致性。其中,股票下跌时预测正确率相对较低,其可能原因是在人工进行打分时,人们更容易偏乐观。另外,在股票不变时,无论是运用大连理工大学情感词典,还是运用人工情感词典,新闻文本的情感得分几乎都不为零,这是由于在每一篇新闻文本中,或多或少都会包含正向或负向的情感词,所以最终新闻文本的得分很难为零。经过初步对比每天的股票涨跌和新闻情感值的正负,人工情感词典表现明显优于大连理工大学情感词典。

选取 2008 年 1 月 1 日至 2008 年 6 月 31 日的股票价格涨跌数据和情感值数据,测试结果见图 3、图 4。

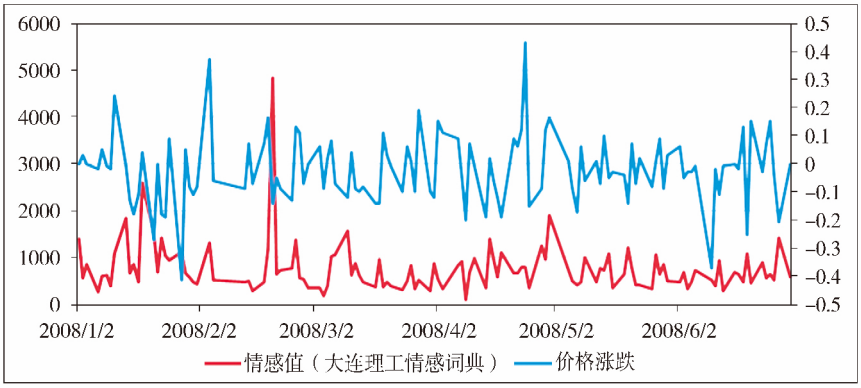

Fig. 3 Performance diagram on DUT sentiment dictionary

图 3 大连理工大学情感词典测试效果图

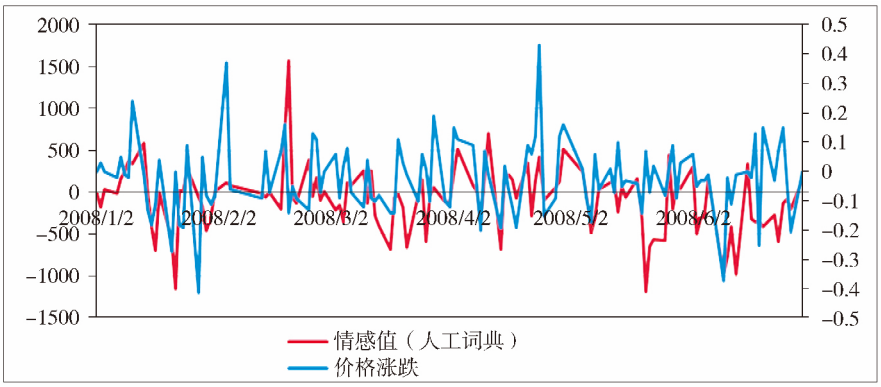

Fig. 4 Performance diagram on artificial sentiment dictionary

图 4 人工情感词典测试效果图

由图可得,人工情感词典打分结果与股票价格保持了较好的一致性,主要体现在:当情感值为正时,股票

价格多为涨,或是跌幅相对之前减小;当情感值为负时,股票价格多为跌,或是涨幅相对之前减小。而大连理工大学情感词典打分结果效果明显不如人工情感词典。

综上,可以得出以下两个结论:

(1)无论是 DUT 情感词典,还是人工情感词典,得分都偏乐观,在价格下跌时的预测效果较差。

(2)相比于人工情感词典,DUT 情感词典存在以下劣势:一、运用 DUT 情感词典得到的情感值几乎都为正,说明该词典在股票市场的适用性较差;二、运用 DUT 情感词典得到的情感值与股票价格涨跌方向的一致程度较差。

3.2 语义规则效果分析

以股票"中国银行 601988.SH"为例,应用人工词典加语义规则进行情感值计算,同时,统计"正负情感词"出现的总次数和"正负情感词+语义规则"组合出现的总次数。

经统计,在七万条新闻文本数据中,人工词典中的正向和负向情感词出现的总次数为 3 360 319 次,而这些情感词和程度副词、否定词、假设疑问词等一同出现的次数为 837 072 次,其占比情况如图 5 所示。

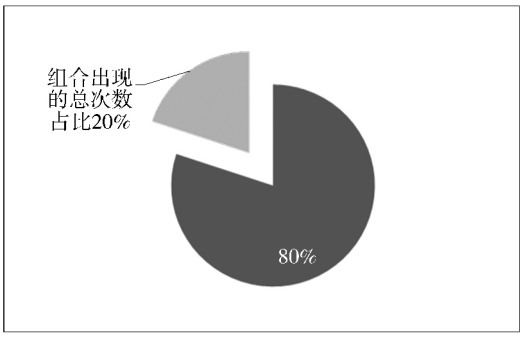

Fig. 5 Pie chart of 'positive and negative sentiment words+semantic rules'

图 5 "正负向情感词+语义规则"组合占比图

在所有匹配到正向和负向情感词的情况中,情感词和程度副词、否定词、假设疑问词一同出现的总次数占了情感词出现总次数的五分之一,说明语义规则所列出的 9 种组合在句子中是较为常见的,也说明是否应用语义规则将会对最终的情感值产生显著的影响。

接着,仅运用人工词典对同样的文本数据再次进行情感值计算。将两次计算分别得到的情感值与股票价格涨跌数据进行对比,结果如表 10、表 11 所示。

表 10 仅使用人工情感词典的测试效果

Table 10 Performance when merely artificial sentiment dictionary is applied

人工情感词典预测结果	情感值为正	情感值为负	情感值为零	正确率
股价上涨	525	392	3	57.1%
股价下跌	542	428	3	44.0%
股价不变	171	126	0	0%

表 11 使用人工情感词典+语义规则的测试效果

Table 11 Performance when both artificial sentiment dictionary and semantic rules are applied

人工情感词典+语义规则预测结果	情感值为正	情感值为负	情感值为零	正确率
股价上涨	587	333	0	63.8%
股价下跌	504	469	0	48.2%
股价不变	182	114	1	0%

由表 10、表 11 和图 6 可以看出,无论是在股票价格上涨,还是股票价格下跌时,加入语义规则进行修正后,情感值的正负向与当天股价的上涨下降方向的一致程度均有所上涨。

综上可知:

(1)本文提出的 9 条语义规则在新闻文本中出现频繁,"正负向情感词+语义规则"的组合出现的总次数占正负向情感词出现总次数的五分之一,语义规则的使用将对最终情感得分有显著影响。

(2)无论是在股价上涨还是股价下跌时,经过语义规则修正后的情感得分,比单独使用人工词典得到的

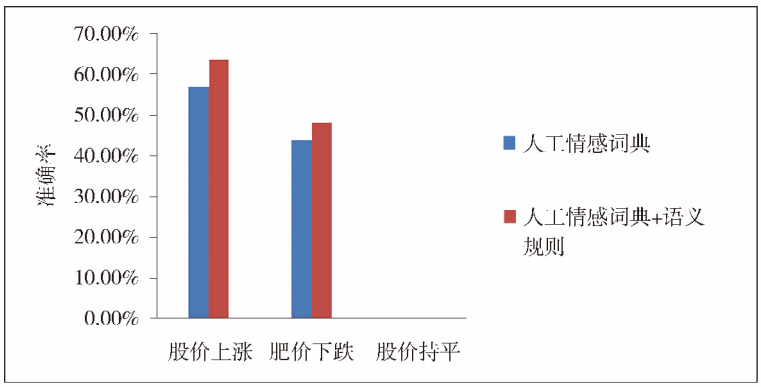

Fig. 6 Performance comparison with and without the semantic rules deployed

图 6 加入语义规则前后的效果对比

情感得分表现更好,情感值正负方向与股价涨跌方向一致程度更高。

3.3 模型预测效果分析

本次分析选用 2008 年 1 月 1 日到 2016 年 12 月 31 日共两千多个交易日的数据和与之日期相匹配的新闻文本情感值作为数据源,分析和比较 BP 神经网络和支持向量机回归模型。用于训练和测试的数据比例为 3∶1,部分股票在某些时间段内没有交易记录,所以在训练集和测试集的量上会有所减少。

用 20 只股票分别进行实验,统计每次实验的均方误差 MSE、走势方向正确率如表 12 所示。

表 12 BPNN 模型与 SVR 模型效果对比

Table 12 Performance comparison between the BPNN and SVR

序号	股票名称	MSE		走势方向正确率	
		BP	SVR	BP	SVR
1	中国银行	0.004 623	0.000 809	44.3%	51.7%
2	京东方 A	0.000 210	0.000 118	47.0%	45.2%
3	农业银行	0.001 954	0.001 309	55.2%	57.8%
4	中国建筑	0.010 766	0.001 333	62.3%	59.8%
5	国电电力	0.069 419	0.000 786	46.0%	46.6%
6	中国重工	0.010 524	0.000 954	50.5%	76.8%
7	工商银行	0.002 986	0.002 109	48.3%	50.0%
8	中国石化	0.001 121	0.000 153	52.1%	51.4%
9	光大银行	0.003 245	0.000 568	54.5%	60.2%
10	中国联通	0.001 436	0.000 567	56.7%	59.3%
11	交通银行	0.002 030	0.000 675	47.8%	55.9%
12	TCL 集团	0.050 411	0.002 030	42.6%	43.0%
13	包钢股份	0.014 994	0.000 701	47.0%	46.6%
14	民生银行	0.005 345	0.009 773	47.2%	51.8%
15	中信证券	0.011 632	0.079 480	50.7%	51.8%
16	中国中车	0.083 147	0.082 633	54.9%	53.9%
17	中国中铁	0.013 623	0.057 014	52.1%	54.9%
18	紫金矿业	0.000 569	0.000 442	49.8%	48.9%
19	兴业银行	0.010 921	0.070 723	53.4%	54.3%
20	海南航空	0.002 155	0.003 983	45.1%	47.3%

除“民生银行”、“中信证券”、“中国中铁”、“兴业银行”、“海南航空”5 只股票外,其他股票使用 SVR 模型得到的 MSE 均小于使用 BP 神经网络模型得到的 MSE。

从走势方向正确率角度看,有 6 只股票使用 BP 神经网络模型得到的正确率更高,另外 14 只股票使用 SVR 模型时效果正确率更高,通过计算平均值,使用 BP 神经网络模型时,20 只股票的平均走势方向正确率为 50.3%,而使用 SVR 时,平均走势方向正确率为 53.3%。

从以上两方面对比可得，在本次实验中，基于 SVR 的股价预测模型的效果优于基于 BP 神经网络的股价预测模型。

以“中国银行”为例，对 2015 年和 2016 年的股价进行预测，与实际价格进行对比，结果如图 7 和图 8 所示，从图形直接可以看出，对于“中国银行”这只股票，SVR 的预测效果远好于 BPNN 模型。

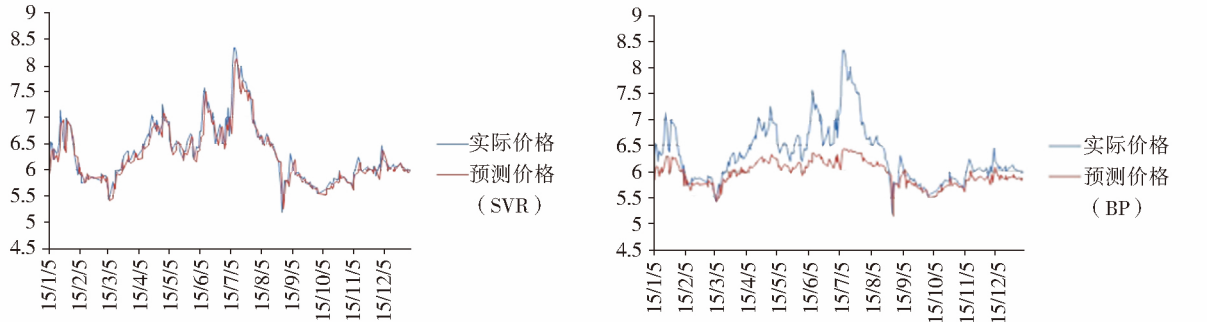

Fig. 7 Prediction diagram of the price of stock ‘601988. SH’ in 2015

图 7 2015 年“601988. SH”股价预测图

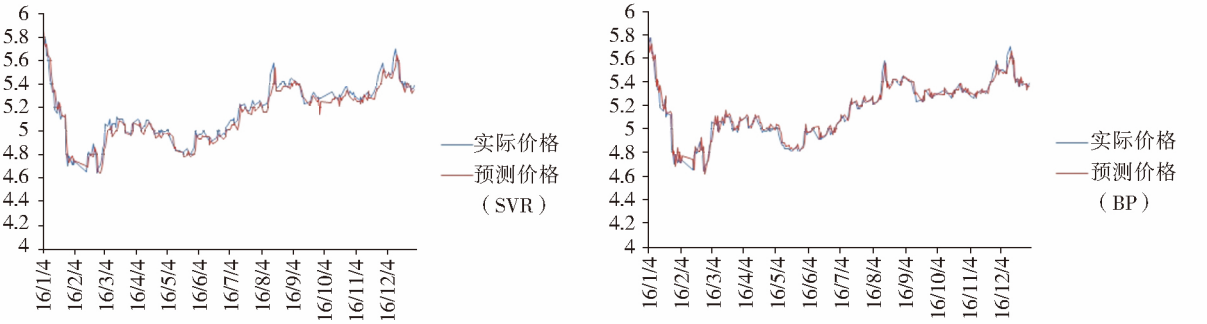

Fig. 8 Prediction diagram of the price of stock ‘601988. SH’ in 2016

图 8 2016 年“601988. SH”股价预测图

4 结论

本文以交易量较大的 20 只股票为研究对象，抓取相关的新闻数据，通过建立有针对性的情感词典和相应的语义规则对新闻文本进行打分，对传统的情感分析方法做出了改进，再加入股票价格的历史数据，分别采用 BP 神经网络和支持向量机回归两种方法建立股价预测模型。本文通过邀请领域内专家对高频词汇进行人工情感打分，得到了一个更具针对性、粒度更细[－5，＋5]的情感词典，与大连理工大学情感词典的打分效果进行对比，应用人工词典得到的情感得分的正负与当天股票价格的涨跌方向一致程度更高；同时，通过给不同语义规则下的情感词赋予不同的权重，对情感值进行修正，在同一样本上分别使用情感词典加语义规则的打分方法和仅使用情感词典的打分方法，发现加入语义规则后，情感得分正负方向与股价涨跌方向的一致程度显著提升。在 BP 神经网络股价预测模型中，新闻情感数据和股票历史价格数据组成的混合数据表现略差于仅使用股票历史价格数据的表现，当股价的滞后天数设定为 6，BPNN 隐含层节点数目设定为 4 时，模型预测效果最佳。而在支持向量机模型回归中，混合数据表现更佳，最佳滞后天数为 20，通过 Grid Search 进行参数寻优，找出了每只股票的最优参数 c 和 g。最终，将两种方法的预测效果进行对比，发现 SVR 模型的均方误差更小，且股价走势方向正确率略高于 BPNN 模型。

参考文献：

[1] 胡照跃. 人工神经网络在股票预测中的应用[D]. 太原：中北大学，2016.

[2] 殷光伟. 中国股票市场预测方法的研究[D]. 天津：天津大学，2003.

[3] 尹璐. 基于 GA_BP 神经网络的股票预测理论及应用[D]. 北京：华北电力大学，2010.

[4] 郑睿颖，伍应环. 神经网络在股票价格预测中的研究[J]. 计算机仿真，2011，**28**(10)：393-396. DOI：10. 3969/j. issn. 1006-9348. 2011. 10. 095.

[5] Bollerslev T. Generalized Autoregressive Conditional Heteroskedasticity[J]. *Journal of Econometrics*, 1986, **31**: 307-327. DOI: 10.1016/0304-4076(86)90063-1.

[6] Rao T S, Gabr M M. An Introduction to Bispectral Analysis and Bilinear Time Series Models[J]. *Lecture Notes in Statistics*, 1984, **150**(150). DOI: 10.1007/978-1-4684-6318-7.

[7] White H. Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns[J]. *IEEE International Conference on*, 1988, **2**(6): 451-458. DOI: 10.1109/ICNN.1988.23959.

[8] Gencay R. Non-linear Prediction of Security Returns with Moving Average Rules[J]. *Journal of Forecasting*, 1996, **15**(3): 43-46.

[9] Zhang G P. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model[J]. *Neurocomputing*, 2003, **50**: 159-175. DOI: 10.1016/S0925-2312(01)00702-0.

[10] Ozbayoglu A M. Neural Based Technical Analysis in Stock Market Forecasting[J]. *Intelligent Engineering Systems through Artificial Neural Networks*, 2008, **18**: 261-265. DOI: 10.1115/1.802655.paper40.

[11] **张坤, 郁湧, 李彤. 基于小波和神经网络相结合的股票价格模型**[J]. **计算机工程与设计**, 2009, **30**(23): 5497-5498.

[12] **秦玉平. 基于支持向量机的文本分类算法研究**[D]. **大连: 大连理工大学**, 2008.

[13] Kim K J. Financial Time Series Forecasting Using Support Vector Machines[J]. *Neurocomputing*, 2003, **55**(1-2): 307-319. DOI: 10.1016/S0925-2312(03)00372-2.

[14] 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. DOI: 10.1016/j.cor.2004.03.016.

[15] **施剑. 基于** SVM **的** IPO **首日投资策略分析**[J]. **计算机系统应用**, 2013, **22**(10): 206-209. DOI: 10.3969/j.issn.1003-3254.2013.10.042.

[16] **张世军. 基于网络舆情的** SVM **股票价格预测研究**[D]. **南京: 南京信息工程大学**, 2011.

[17] **龙真真, 张正文. 基于模糊核超球的快速分类算法在股票预测中的应用**[J]. **计算机系统应用**, 2014, **23**(1): 197-201. DOI: 10.3969/j.issn.1003-3254.2014.01.040.

[18] **熊泉浩. 中文分词现状及未来发展**[J]. **科技广场**, 2009(11): 222-225. DOI: 10.3969/j.issn.1671-4792.2009.11.067.

[19] **马力, 宫玉龙. 文本情感分析研究综述**[J]. **电子科技**, 2014, **27**(11): 180-184. DOI: 10.3969/j.issn.1007-7820.2014.11.052.

[20] **杨希. 基于情感词典与规则结合的微博情感分析模型研究**[D]. **合肥: 安徽大学**, 2011.

[21] **李克文, 王秋宝, 于明晓. 基于改进** ACO **优化** BPNN **的软件缺陷预测模型**[J]. **计算机工程与设计**, 2017(8): 2137-2141.

[22] **高雷阜, 佟盼. 融合改进遗传和人工蜂群的** SVM **参数优化算法**[J]. **计算机工程与应用**, 2016, **52**(18): 36-39. DOI: 10.3778/j.issn.1002-8331.1411-0111.

[23] **白淼. 基于** BP **和** SOM **神经网络的股票价格预测的研究**[D]. **阜新: 辽宁工程技术大学**, 2009.