

市场情绪对股票走势的影响分析及预测

詹冰清, 屈波怡

(上海理工大学 管理学院, 上海 200093)

摘要:为探究投资者情绪对股票走势的影响,利用 R 软件的爬虫技术将人们对股票市场的看法抓取下来,将文本中的隐含情绪分为积极、消极及中性 3 种类别,并依此构建情绪得分作为市场情绪量化的结果。利用单位根检验等方法对上证指数与投资者情绪的因果关系进行探究并建立 VAR 模型。为更好地判断市场情绪对股票走势的影响程度,分别构建加入情绪得分前后的 BP 神经网络模型对上证指数收益率进行预测,比较两个模型的优劣,从而发现当股票预测模型加入市场情绪指标后误差更小,预测更为准确。

关键词:市场情绪;股票预测;VAR 模型;BP 神经网络

中图分类号:F832.5 **文献标志码:**A **文章编号:**1671-1807(2021)04-0051-07

在互联网飞速发展的今天,大数据成为人们分析和关注的重点。股票市场作为金融市场一个重要的分支,它的波动会对金融市场及人们的日常生活产生巨大的影响。自 1990 年在上海成立了新中国第一家证券交易所,中国股市发展 30 年,由较为混乱的波动转变为较为规律性的波动。股份有限公司通过上市发行股票等有偿证券进行资金的融通,而投资者也依据股票进行投资。影响股票波动的因素很多,如金融政策的变动、公司自身的经营状况和投资者自身的行为等。各种股票信息网站也随之增加,新浪、腾讯等网站更是为股票单立了一个版块,为投资者提供股票的相关信息,帮助投资者进行决策。不少学者更是对股票变动的影响因素进行了研究,提出了投资者情绪指数等指标来对股票的波动进行分析,将投资者的情绪与股票的变动联系起来。

现有的针对市场情绪的研究中,部分学者通过选取合适的情绪指标来反映市场情绪,也有学者利用大数据时代文本数据中的隐含信息构建情绪指标。饶兰兰和涂裕荣^[1]、王舒曼^[2]都是利用现有的变量,如好淡指数、市盈率、抄手率、新增开户数等显性或隐性的指标来量化市场情绪;而王婧^[3]则认为文本数据中包含大量的投资者情绪,可以从中构建出情绪指标。

在研究投资者情绪对股票走势的影响时,参考了相关文献中将情绪量化的思想,但并未采用

这些文献中对情绪指数的构建方法。中外学者为了探讨投资者情绪与股票走势的关系,通过构建显性、隐性和好淡指数等情绪指数将投资者的情绪量化,从而对二者的关系进行实证分析。本文基于新浪、腾讯等网站对股票评价的文本数据,利用文本挖掘技术将文本转化为情绪词汇,并利用情绪得分量化情绪从而进入模型,将股票市场的走势与投资者的情绪联系起来,分析市场情绪对上证综指指数的影响^[4],这也符合现今对金融数据分析的要求。

本文主要研究投资者情绪与股票走势的相关关系。通过爬取相关网站文本数据建立投资者情绪指标,并选取大盘指数中的上证综指指数作为研究对象。通过构建 VAR、BP 神经网络模型来对该变量的关系进行探究,从而找出他们之间的因果关系,并进行预测来比较两个模型的优劣。建模过程分为以下 3 个步骤:

1)利用 R 软件对股票与文本数据进行爬取。利用 RCurl、XML 等包对新浪股票版块的文本进行抓取,并利用源代码的路径对抓取的文本进行规整化处理。

2)利用文本挖掘进行情感分析。利用词典提供的情感词汇对所有的词汇赋予情感权重,计算每个文本最终的情感得分,作为反映市场情绪的指标。

3)进行市场情绪与股票走势的关联分析。利用计算得到情绪得分进行与股票相关指标进行分

收稿日期:2020-11-10

作者简介:詹冰清(1997—),女,江西上饶人,上海理工大学管理学院,硕士研究生,研究方向为数量经济学;屈波怡(1996—),女,江苏无锡人,上海理工大学管理学院,硕士研究生,研究方向为统计学。

析,将情绪得分作为一个影响变量加入反映股票走势的模型中。通过格兰杰检验可判断情绪得分是否对股票走势产生影响,并利用 VAR 模型与 BP 神经网络对上证综合指数进行分析和预测^[5]。

1 数据描述及指标建立

1.1 数据来源

主要通过对新浪股票股市观点、博客看市、大盘评述和主力动向等版块进行文本挖掘,从而获得投资者情绪的文本来源,并选择上证综指指数收益率作为研究对象。选取 2016 年 12 月至 2017 年 6 月 180 个日数据进行分析。

1.1.1 数据内容

选取新浪网股票界面的股市观点、大盘评述、主力动向、基金博客等板块,爬取 2017 年 1 月 3 日到 6 月 16 日的文本数据,从中筛选出 4 176 条信息。

1.1.2 爬取的总体框架

利用 R 软件中的 RCurl 包和 XML 包对网站的信息进行爬取^[6]。由于 RCurl 包只能对 url 进行定向爬取,即只能爬取当前网页的信息,因此先对要爬取的模块的主页面中抓取出所需板块的子链接,然后利用循环语句对主页面中各板块的子链接导向的文本进行爬取。

1.1.3 爬取的基础步骤

1)利用 RCurl 包中的 getURL()函数获取网站的源代码,并根据 url 的编码,设置参数.encoding。为了防止网站以为是恶意侵入的,伪装了报头即定义参数 header。

2)利用 XML 包中的 htmlParse()函数对获取的源代码进行解析。需要注意的一点是,XML 包对中文的支持不太理想,并且由于 encoding 的原因,会出现中文乱码的情况。因此在这两步中间加入一个 iconv()函数,对网站进行转码。将网站都转为 utf-8 的编码。

3)通过观察网站源代码中节点,可以利用 XML 包中的 getNodeSet()函数里面的 path 参数,设置路径,然后获取所设路径的源代码。之后利用 xmlValue或者 xmlGetAttr 函数等,xmlValue 是返回所设节点的值,xmlGetAttr 是返回所设节点的属性,这样就可以得到想要的信息。

1.2 软件说明

在对数据处理、分析和建模的过程中,统一使用 R 软件对其进行分析。R 软件是一个开源免费的软件,可以支持进行大部分的分析与操作。利用

R 语言中的多种包对数据进行爬取、分词、去除停用词、构建模型等操作。R 软件包使用说明见表 1。

表 1 R 软件包使用说明

R 软件包	程序包用途
RCurl 包	用于抓取网站的源代码
XML 包	用于解析网站的源代码
Rwordseg 包	对文本进行分词
Plyr 包	将数据框合并
Wordcloud2 包	绘制可视化云图
Urca 包	进行单位根检验
Tsseries 包	构建 VAR 模型
Nnet 包	构建 BP 神经网络
Lmtest 包	进行格兰杰因果检验

2 文本挖掘与情绪分析

2.1 总体框架

在获得大盘评述、股市观点、主力动向、博客看市等板块的文本数据后,利用构建情绪词典的方式来对文本中隐藏的情绪进行挖掘。选取哈尔滨工业大学社会计算与信息中心研发的情绪词典,其中包含正面、反面及中性词汇。通过对爬取的文本进行分词、去除停用词、计算情绪得分等操作,将市场情绪量化,从而有利于分析市场情绪与股价走势的关系。在处理文本的过程中,下载 Rwordseg 包和 rJava 包,利用其中的 SegmentCN()等函数对文本进行分词。利用循环语句对文本情绪进行打分,规定正面词汇的情感得分为+1,负面词汇的情感得分为-1,中性词汇的情感得分为 0,利用每篇文本中正面与负面词汇的评分和作为该篇文档的情绪倾向和得分值^[7]。

2.2 情感分析的具体步骤

2.2.1 数据的清洗

由于在爬取数据时,是根据源代码节点的规律来获取想要的信息,因此其中可能夹杂类似于“新浪财经 App:直播上线博主一对一指导”等无意义文字,且英文单引号、波浪号等标点会在 R 读取时会产生警告信息,从而造成文件读取不完全和乱码。因此在分词之前先对文档中这样的数据进行清洗。利用 gsub()函数以及正则表达式,将不满足要求的文字及英文字符全部剔除。

2.2.2 对文本进行分词

由于爬下的文本是完整的文档,为了利用词典进行情感分析,采用 segmentCN()函数对文本进行分词处理。在分词之前需要装载自定义词库,本文装载的词库为搜狗输入法下的“财经金融词汇大全

【官方推荐】”词库以及“股票基金词库大全”。可是,由于词典并不十分齐全,导致分词的效果并不是很理想,分出来的词仍然有所欠缺,部分结果如表 2 所示。

表 2 初步分词

方法	分词结果
初步分词	恒、大、安、邦、宝、能、前、海、人、寿、举、牌、特、朗、普等等

显然,R 把“恒”“大”“安”“邦”等能表述一定特定含义的词拆分开了,因此需要人工观察那些不能分开的词,运用 insertWords() 函数,添加不能将其分开的新词。并且 R 还将“特”“朗”“普”之类的人名拆开了,针对这一问题设置 isNameRecognition=TRUE,使 segmentCN() 函数可以辨别人名。接着再重新对文本进行分词,结果如表 3 所示。

表 3 部分分词结果

方法	分词结果
人工添加新词	恒大、安邦、宝能、前海人寿、举牌、特朗普等

可以直观地感受到,通过人为添加词汇可以提高词典对文本的分词效果。通过人工与词典结合的双重分词,最终能够得到的分词结果较为合理,至此,分词工作完成。

2.2.3 删除停用词

通过对在文本中出现次数前 100 名的高频词做可视化云图,发现出现次数最多的是“的”“是”“在”“了”“股”“市场”等对情感分析毫无意义和价值可言的词,这些没有明确的意义,也无明显的情感倾向的词也被称为停用词。由于文章中包含大量的停用词,为了进一步减轻工作量也为了保证后续建模的分析效果,通过下载停用词典,并根据文本特点人工增加一定无意义的停用词,丰富停用词典,利用管道函数将停用词删除。删除前后文本可视化云图如图 1 所示。

由图 1(b)可以看出,删除停用词后的文本出现频率高的为“板块”“增”“跌”“高”“涨”“减”等可以具体体现市场情绪的词,这说明去除停用词是很有必要的,同时也说明了停用词去的效果还不错。

2.2.4 计算情感得分

将文本停用词进行删除之后,下一步的工作是对每篇文章计算情感得分,判断它的情感倾向。采



图 1 删除停用词前后可视化云图

用哈工大情感词典,且将人工判断词典不能识别的重要词加入词典。之后根据此情感词典对处理后的词赋予情感权重,再计算每篇文章的情感得分,导入正面和负面的情感词典,将其与分词后的文本进行匹配,从而对文本进行打分。通过判断评分的正负判断情感倾向。若得分为正,则认为该文档具有正面的情感倾向;若得分为负,则认为该文档具有负面的情感倾向。

对文本分词后的列表按照每一篇文章分词后的词的个数加上标签,然后将文本和词典进行匹配,最后利用 aggregate() 函数,根据标签将情感得分进行汇总。得分结果如图 2 所示。

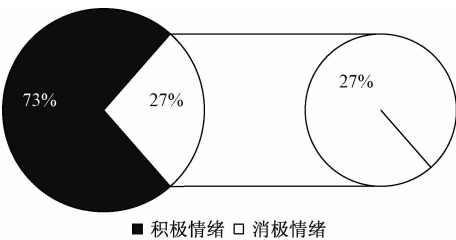


图 2 情绪得分

由图 2 可知,积极情绪的比例大于消极情绪的比例,这说明近半年股票市场的情绪还是以积极为主的,并将市场情绪量化成情绪得分。

3 VAR 模型的建立

3.1 分析总体框架

在对投资者情绪及股票收益关系进行分析时,利用已经爬取的 4 000 多条数据进行分析。由于上证综指指数是大盘指数的重要组成成分之一,因此选取的研究对象为计算出的情绪得分及上证综指的收益率。为了对二者关系进行研究,采用了 VAR

模型及其相关检验来进行分析^[8]。

1)对投资者情绪与股票收益变动的因果关系进行检验。由于两个变量都为时间序列数据,VAR 模型成立的前提条件为序列平稳,因此对该变量进行 ADF 单位根平稳性检验。当两个变量序列都平稳时,对其构建 VAR 模型(向量自回归模型),并在模型显著的基础上进行格兰杰因果检验。从而得到投资者情绪与上证指数的关系,判断出二者的因果关系。

2)探究模型对股票走势的预测效果。依据已知的数据分别构建 VAR 和 BP 神经网络模型,并利用构建出的投资者情绪与股票走势关系模型对股票的变动进行预测,探究模型的预测效果。当模型预测效果不好时,选择更优的模型对股票走势进行预测。

如图 3 所示,上证综合指数收盘价的波动剧烈,初步看收盘价序列不平稳,但仍应当进行单位根检验判断序列的平稳性。

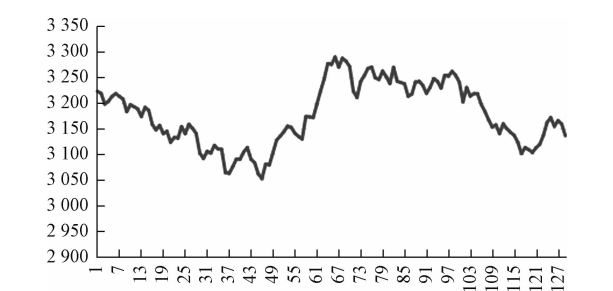


图 3 上证指数收盘价波动图

3.2 对二者因果关系进行检验

由于 VAR 模型的前提是数据变量平稳,因此先对投资者情绪得分(E)和上证综指指数(S)进行单位根检验。由于 fUnitRoots 包无法安装,所以利用 R 中的“urca”包对变量进行单位根检验。对上证综指指数差分定义为 DS,对情绪得分差分定义为 DE 对差分后的变量进行检验,结果如表 4 所示。

表 4 ADF 检验

变量	T 统计量	是否平稳
DE(情绪得分)	-30.096	是
DS(上证综指收益率)	-29.765	是

在保证单位根平稳的情况下,构建 VAR 模型,以 DS 为因变量,DE 为自变量来构建。依据 VAR 模型滞后表判断阶数,如表 5 所示。

表 5 VAR 模型滞后阶数判断

变量	1	2	3	4	5
AIC(n)	3.673	3.582	3.517	3.517	3.507
HQ(n)	3.690	3.610	3.556	3.566	3.568
SC (n)	3.716	3.653	3.617	3.644	3.663
FPE(n)	39.388	35.946	33.684	33.673	33.347

由表 5 可以发现,当阶数为 2 时 AIC 取值较小,此时其他准则取值也较小,因此选择滞后期为 2 的 VAR 模型,如表 6 所示。

表 6 VAR 模型结果

变量	DS(-1)	DS(-2)	DE(-1)	DE(-2)
DE	-1.322	0.219	-0.032	-0.0001
	(0.987)	(0.989)	(0.047)	(0.003)
	[-1.34]	[0.221]	[-0.688]	[0.530]
DS	-0.051	0.011	0.0007	0.0003
	(0.047)	(0.047)	(0.002)	(0.002)
	[-1.086]	[0.233]	[0.32]	[0.138]

所以,上证综指收益率的变动(DS)可以用其自身的滞后期及情绪得分的滞后期解释。可以发现上证指数收益率的变动受过去滞后一期的情绪得分影响最大。随着时间的推移,情绪对股票走势的滞后效果逐渐减弱。

由于 VAR 模型显著,因此对其进行格兰杰因果检验。格兰杰因果检验的实质如下。

$$\begin{cases} H_0: DS \text{ 是 DE 的格兰杰检验} \\ H_1: DS \text{ 不是 DE 的格兰杰检验} \end{cases} \quad (\text{情况 1})$$

用公式表示为

$$DE = \sum_{i=1}^q \alpha_i DS_{t-i} + \sum_{j=1}^q \beta_j DE_{t-j} + u_{1t} \quad (1)$$

$$DS = \sum_{i=1}^q \alpha_i DE_{t-i} + \sum_{j=1}^q \beta_j DS_{t-j} + u_{2t} \quad (2)$$

当原假设和备择假设为情况 1 时,格兰杰因果检验的公式如式(1)所示;反之,如式(2)所示。格兰杰检验结果如表 7 所示。

表 7 格兰杰检验结果

原假设	格兰杰检验结果
DS 是 DE 的格兰杰检验	是
DE 是 DS 的格兰杰检验	是

由图可以判断出市场情绪和股票收益率互为格兰杰原因。此外对残差进行检验,发现残差此时协整,F 值达到 206.6,P 值远远小于 0.05. 因此,可以认为投资者情绪得分是上证综指变动的格兰

杰原因。

4 模型的选择及预测

4.1 VAR 模型预测

在构建出 VAR 模型后,利用该模型对上证综指的收盘价进行预测。通过对预测结果的分析,能发现 VAR 模型的预测误差较大,不能很好地对上证综指收盘价进行预测,如表 8 所示。

表 8 预测 4 天收益率结果

时间	变动	最低值	最高值	平均值
第 1 天	-0.050	-0.833	0.732	0.783
第 2 天	-0.042	-0.825	0.742	0.784
第 3 天	-0.040	-0.824	0.744	0.784
第 4 天	-0.04	-0.823	0.744	0.784

由表 8 所示,在置信度为 95%的水平下 4 天收益率的预测值相差不大,收益率的置信区间较大,导致精度不够,因此认为 VAR 模型对收益率的预测不够精确,具有较大的误差。为了对股票的走势进行较为准确预测,通过参考文献对众多预测模型的比较找出较为准确的预测模型——BP 神经网络模型。

4.2 BP 神经网络模型预测

4.2.1 BP 神经网络原理^[9]

BP 神经网络属于多层感知器(multilayer perceptrons,MLP)的一种,用于解决预测中的线性不可分问题。神经网络是一种类似于黑匣子的模型,除了输入层和输出层外,还包括若干隐含层。也就是说,一个神经网络模型有三层及三层以上的神经元。其中,全连接层 BP 神经网络相邻的上下层之间的神经元实现全连接,但是同一层神经元之间并无任何联系。其中,输入层与隐含层神经元之间依据网络的权值来进行联系,即两个神经元之间的连接强度^[10]。输入层将数据中包含的信息传入隐藏层中,隐含层则将前一层所有神经元传来的信息进行整合继续向下传递,直到传递到输出层。

4.2.2 模型建立

4.2.2.1 网络层数的确定

尽管增加网络层数可以起到降低误差的效果,但同时也使神经网络复杂化,从而增加了网络权值的训练时间。而误差精度的提高其实也可以通过增加隐藏层中的神经元数目来获得,并且其训练效果比增加网络层数更佳。同时,由 Kosmogorov 定理可知,在神经网络结构合理以及神经节点权值取值恰当的条件下,三层神经网络可以逼近任何连续

函数。因此,基于上述分析,将隐藏层的个数设为 1,也就是构建三层神经网络。

4.2.2.2 输入层神经节点的设计

本文是基于市场情绪指数和上证指数往期数据对上证综合指数进行预测。考虑到中国股票市场的实际情况,一周的交易天数为 5 天,上证综指的滞后期选择四期较为合适。因此,将输入层神经节点数目设置为 6。BP 神经网络结构如表 9 所示。

表 9 BP 神经网络结构

变量	模型 1	模型 2
输入层数据	前 4 期收盘价	前 4 期收盘价及情绪得分
输入层神经元数	4	6
隐含层	1	1
隐含层神经元数	9	9
输出层神经元	1	1

模型 1: $y_t = NN(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$ (3)

模型 2: $y_t = NN(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, x'_{t-1}, x'_{t-2})$ (4)

式中: y_t 为 t 时期上证指数的收盘价; x_t 为 t 时期市场情绪指数; y_{t-i} 为从 t 时期起滞后 i 期的收盘价, $i=1,2,3,4$,即 $t-i$ 时期的收盘价; x_{t-i} 为从 t 时期起滞后 i 期的市场绪指数, $i=1,2$,即 $t-i$ 时期的情绪指数。

4.2.3 模型拟合

4.2.3.1 只基于上证指数的 BP 神经网络模型

运用 R 软件对 2017 年 1 月 3 日到 2017 年 6 月 11 日的 104 个数据进行处理,作为驯良样本,再用接下来的 4 个数据作为模型的预测集,用前 4 期的数据对下一期做预测,本文主要对收盘价进行了预测,图 4 为实际数据和预测数据之间的拟合,其中红色表示实际的收盘价,蓝色表示预测的收盘价。模型的拟合优度为 96.62%,拟合结果较为理想,同时,均方误差为 166.662 2,可见整体的偏差也不是很大,说明了神经网络预测模型的有效性。同时,可以看出整个 2017 年上半年中,上证指数呈先缓慢上升再下降的趋势。

4.2.3.2 加入情感得分的 BP 神经网络模型

基于前文的分析研究,考虑将情感得分加入模型,作为一个新的输入变量,依旧基于收盘价进行预测,将最后得出的结果先与实际值做对比,看模型是否可行,再与未加入情感得分的 BP 神将网络模型做对比,看情感得分是否在能够提高预测的精度。模型拟合的结果如图 5 所示。MSE 值约为

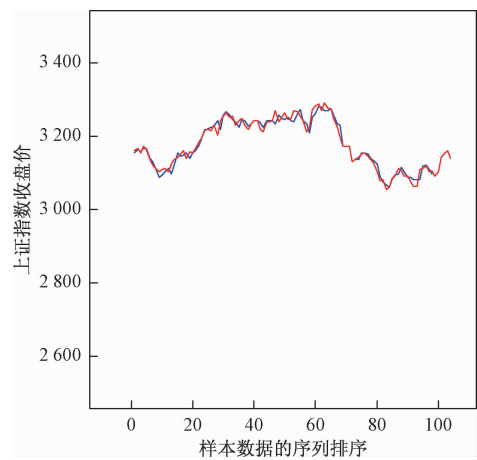


图 4 没加入情感得分的模型拟合

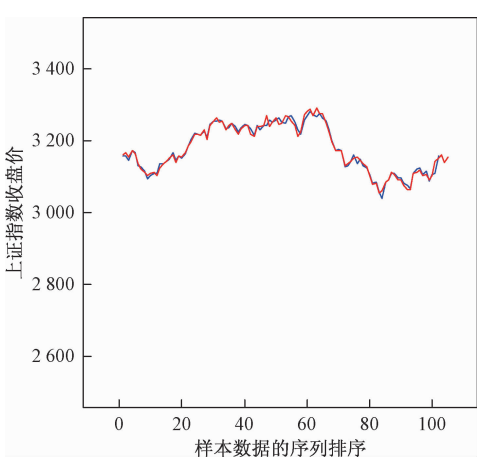


图 5 加入情感得分后的模型拟合

80,模型拟合良好。
基于本文所建立的 BP 神经网络模型,对接下
来的 4 期数据做预测,并与实际结果进行对比,预测
结果如表 10 所示。

表 10 BP 神经网络预测效果

日期	实际值	预测值			
		模型 1		模型 2	
		预测值	偏差	预测值	偏差
2017 年 6 月 13 日	3 153.742 9	3 109.457	44.285 9	3 115.01	38.733
2017 年 6 月 14 日	3 130.674 0	3 166.222	35.548 0	3 162.59	31.916
2017 年 6 月 15 日	3 132.486 3	3 150.766	18.279 7	3 161.14	28.655
2017 年 6 月 16 日	3 123.166 2	3 161.213	38.046 8	3 143.35	20.186

表 10 说明 BP 神经网络模型预测效果极佳。
由于时间有限,获得情感得分数据较少,只有
6 个月的时间,导致 BP 神经网络的训练集数据较
少,神经网络的结果还不是很理想。所以,针对这
些问题提出了进一步优化,主要有以下 3 点:

- 1)需要增大样本量,获得更长时间且精度更高的情感得分数据,只对新浪财经上的股评进行了爬取,在接下来的工作中,可以根据相关指标选取更多的股票评论网站,综合多方面的意见爬取数据,同时,在分词时,可以结合句法及朴素贝叶斯对文本数据进行分词以及计算情感得分的数据清洗工作,从数据来源上进行改进,
- 2)本文在构建 BP 神经网络时选取的是上证综合指数得到收盘价,收盘价只是股票价格变动的一个侧面,可以考虑将上证综合指数的成交量、交易金额、市盈率等加入模型中,从多方面来探究股评的情感趋势对大盘变动的影响。
- 3)本文在构建 BP 神经网络模型的过程中,采取的算法比较简单,同时神经网络模型本身也有限制,可以考虑在今后的工作中采取更为有效的算法

来提高预测的精度,并将 BP 神经网络和其他算法如支持向量机等结合起来构建预测精度更高的模型。

5 结论

利用多种分析与处理方法,构建了 VAR 和 BP 神经网络模型,通过这两个模型来探究市场投资者情绪与上证综指指数的关系,并比较了两种模型的预测效果。在对模型的比较中,发现 BP 神经网络在预测方面具有其他模型无法比拟的优势,不用建立复杂的数学模型,具有很强的自适应、自学习能力,所以 BP 神经网络模型在股票预测方面更为准确合理。

参考文献

[1] 饶兰兰,涂裕荣.牛熊市下投资者情绪对股票收益影响的非对称性分析[J].商业经济,2016(8):141—143.
[2] 王舒曼.投资者情绪变动与股市收益率关系的实证研究[J].商丘师范学院学报,2016,32(2):98—101.
[3] 王婧.投资者情绪对股票市场收益的影响——基于上证综合指数的实证分析[J].知识经济,2016(6):34—35.
[4] 王镇.投资者情绪对中国股市收益影响的实证研究[D].大

连;东北财经大学,2015.

[5] 王美今,孙建军. 中国股市收益、收益波动与投资者情绪[J]. 经济研究,2004(10):75—83.

[6] MUNZERT S,RUBBA C,NYHUIS D. 基于 R 语言的自动数据搜集[M]. 吴今朝,译. 北京:机械工业出版社,2016.

[7] 韩磊. 利用 BP 神经网络系统对股票市场进行预测与分析的研究[D]. 天津:天津大学,2013.

[8] 梁希喆. 投资者情绪与股票收益关联性研究[D]. 北京:北京交通大学,2017.

[9] 田鑫. 基于情绪指数和神经网络的上证指数预测研究[D]. 哈尔滨:哈尔滨工业大学,2013.

[10] 郭晓菲,李清军,潘子颖,等. 基于投资者情绪指数的上证综指预测研究[C]//2015 年(第四届)全国大学生统计建模大赛论文. 北京:中国统计教育学会,2015:24.

Analysis and Forecast of the Influence of Market Sentiment on Stock Trend

ZHAN Bing-qing, QU Bo-yi

(Business School,University of Shanghai for Science and Technology,Shanghai 200093,China)

Abstract: In order to explore the impact of investor sentiment on the stock trend, the crawler technology of R software was used to capture people's views on the stock market, and the implied emotions in the text were divided into three categories, namely positive, negative and neutral, based on which the sentiment score was constructed as the quantitative result of market sentiment. Then, the unit root test and other methods were used to explore the causal relationship between The Shanghai Composite Index and investor sentiment and establish the VAR model. In order to judge the degree of influence of market sentiment on stock trend, BP neural network models were built respectively before and after the addition of sentiment score to predict the return rate of Shanghai Stock Index. The advantages and disadvantages of the two models were compared, and it was found that the error of stock prediction model after the addition of market sentiment index become smaller and more accurate.

Key words: market sentiment;stock prediction;VAR model;BP neural network