文章编号：1673-5196(2018)04-0103-06

基于R语言的网络舆情对股市影响研究

朱昶胜[1]，孙 欣[1]，冯文芳[2]

(1. 兰州理工大学 计算机与通信学院，甘肃 兰州 730050；2. 兰州理工大学 经济管理学院，甘肃 兰州 730050)

摘要：以开源R语言为平台，东方财富网的股评为研究对象，结合中文文本挖掘技术和SVR支持向量回归模型.利用中文挖掘技术，对股评进行去噪声、分词、同义词合并、去停用词、TFIDF、文本向量化将非结构化文本数据转化为结构化的特征向量矩阵，与股票的收益率建立SVR回归模型，通过预测未来的股票收益率来预测股价的涨跌趋势.研究结果表明，预测股价涨跌趋势与实际趋势基本吻合，可以通过分析网络舆情来对股市未来发展趋势进行预测.

关键词：网络舆情；R语言；中文文本挖掘；SVR模型

中图分类号：TP391 **文献标志码**：A

Study on the impact of network public opinion on the stock market based on R-language

ZHU Chang-sheng[1], SUN Xin[1], FENG Wen-fang[2]

(1. School of Computer and Communication, Lanzhou Univ. of Tech, Lanzhou 730050, China; 2. School of Economics and Management, Lanzhou Univ. of Tech, Lanzhou 730050, China)

Abstract: The open source R-language is taken as the platform, the Oriental Fortune Network is taken as the research object, and the Chinese text mining technique and support vector regression(SVR) model are incorporated to examine the stock market. By means of Chinese text mining technique, the denoising, word segmenting, synonyms merging, waste word descarding, TFIDF, and text vectorization are carried out for the stock review text, the unstructuralized text data is transformed into a structuralized eigenvector matrix, the SVR regression model is established along with the return rate of the stock, and the volatility trend of stock price is predicted by means of prediction of the future return rate of the stock. The investigation result shows that the predicted volatility trend of the stock price will basically coincide with the actual one, and the analysis of network public opinion can be adopted to predict the future developing trend of the stock market.

Key words: network public opinion; R-language; Chinese text mining; SVR model

影响股市波动的因素有很多，如市场行情、通货膨胀、净资产收益率等.实际上，与财经相关信息都会影响证券市场股价的波动[1]，这些信息最终可以归结为定量信息和定性信息.定量信息是指可以直接获得的实际观测数据[2]，即科技指标；而定性信息是指不能直接用数据精确描述的因素，如商业环境、文化程度、技术优势、战争、自然灾害、政府经济政策变动等.东方财富股吧中的股评就包含了大量这种定性信息.已有的研究表明，网络舆情对股价走动是有影响的[3-4].

目前，很多领域的学者在进行网络舆情与股市波动之间关系的研究时，新闻源主要是互联网，涉及的新闻是互联网中海量新闻，研究内容主要是互联网新闻信息与股价走向之间的关系.例如，Gunn等[5]运用文本挖掘技术，采用支持向量回归模型来预测新闻对股价的影响；Tang等[6]用移动平均算法计算技术指标对股价的影响，再采用支持向量回归模型预测新闻对股价的影响，最后将两者融合，通过训练集最小化风险函数，得到最终的回归函数，再用得到的模型来预测股价；赵伟等[7]研究了互联网财

收稿日期：2016-12-15

作者简介：朱昶胜(1972-)，男，甘肃秦安人，博士，教授，博导.

经信息发布的数量与股票收益率波动之间的关系，研究表明财经信息量增加时，证券市场股票收益率会出现一定幅度的波动，当财经信息量明显增加数倍时，股票收益率的波动会摆脱随机因素的干扰，显著受到信息量的影响；赵丽丽等[8]提出了融合计算机领域的文本挖掘技术与经济学领域的计量方法，跨学科角度分析新闻如何影响股市波动，即新闻对股市影响第几天最为显著，影响的持续时间多久；马俊伟等[9]采用信息抓取技术获得网络金融舆情文本信息，并根据数据的信息量对金融舆情信息进行分类，建立因子模型和时间序列模型，分析网络金融舆情信息对我国股票市场的影响.

综上所述，很多研究人员对网络舆情与股市之间关系的研究，由于技术的局限性，主要还只是简单地采用新闻数量和标题来进行研究分析网络舆情如何影响股市，然而仅仅分析新闻数量或标题的影响，将忽略新闻文本中包含大量有价值的软信息，从而无法准确获悉网络舆情与股市波动之间的关联.据此，本文以股评文本内容为研究对象，以开源软件 R 为基础，结合文本挖掘思想，将非结构化的股评文本转化为结构化的特征矩阵，提取高频特征矩阵与收益率做 SVR 支持向量回归模型.对原发股评进行未来一个月的收益率预测，实现网络舆情对股票价格影响的研究.

1 R 语言简介

R 是基于统计分析、绘图的语言和操作环境，是属于 GNU 系统的一个自由、免费、源代码开放的软件，是一个用于统计计算和统计制图的优秀工具.R 也是一套完整的数据处理、计算和制图软件系统.其功能包括：数据存储和处理系统，数组运算工具（其向量、矩阵运算方面功能尤其强大），完整连贯的统计分析工具，优秀的统计制图功能，简便而强大的编程语言，可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能.

在数据挖掘方面 R 有天然的优势[10]，R 语言封装了各种基础学科的计算函数，在 R 编程过程中只需调用这些计算函数，就可以构建出面向不同领域、不同行业的复杂数学模型.在 R 上有众多主流机器学习算法包，因此使用 R 结合各种机器学习算法对大数据挖掘非常简洁，其过程如图 1 所示.

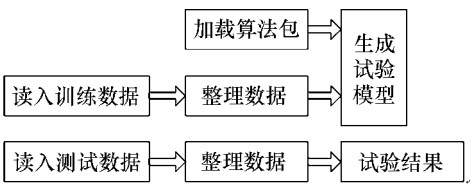

图 1 R 数据挖掘流程

Fig.1 R data mining flow chart

2 中文文本挖掘技术

本文所要研究对象为海量的股评文本数据，如何挖掘文本中包含的有价值的信息，将非结构文本数据转化为结构化数据，这些问题的解决就需要用到中文文本挖掘[11-13]（text mining）技术.主要有以下 7 个步骤.

1）中文分词：将连续的中文语句按照一定规则切分成正确的词串，这些词串含有一个个单独的意义.

2）合并同义词：将意思相似或相同的词语用同一个词语代替，从而实现降维和提高文本处理的准确性.

3）去停用词：去掉文本中没有实际意义，不符合语义单元要求的词语，如语气词、介词等.

4）TFIDF：是一种统计方法，用以评估一个词对于一个文件集或一个语料库中的其中一份文件的重要程度.

5）生成矩阵：把不同文章作为不同样本，以其中分成的词语形成矩阵.

6）特征降维：删除占有比例小的词语.研究表明去掉这些比例小的词语对最终预测结果并不会带来很大影响（维数太多在计算时会浪费大量的时间）.

7）生成模型：运用上面生成的矩阵，选择一种合适的算法，形成最终模型.预测文本基本与训练文本步骤相同，只是在生成矩阵上用训练文本所选择的特征即可.

中文文本挖掘主要流程如图 2 所示.

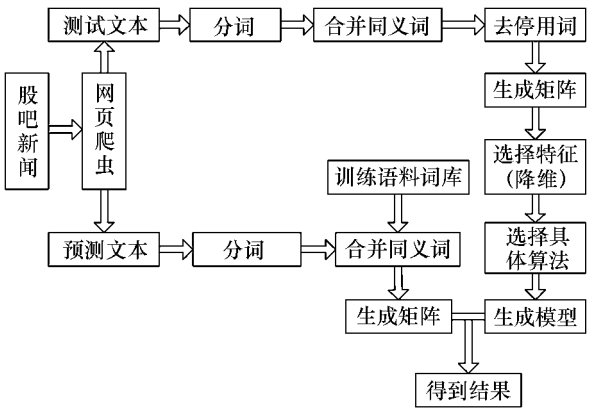

图 2 文本挖掘流程

Fig.2 Flow-chart of text mining

3　支持向量回归机算法

支持向量回归机(SVR)[14-15]寻求的是一个线性回归方程(函数 $y=wx+b$)去拟合固有的样本点,它寻求的最优超平面不是将两类分得最开,而是使样本点离超平面总方差最小.SVR 原理如图 3 所示.

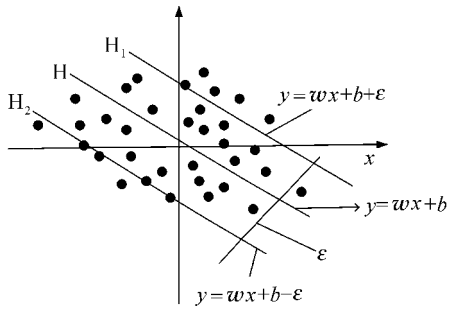

图 3　SVR 原理

Fig.3　Principle diagram of SVR

假设给定训练样本集$\{(x_i,y_i)\mid x_i\in X=R^d,\ y_i\in R,i=1,\cdots,n\}$,其中 x_i 为输入空间 X 的一个数据点,y_i 为对应于 x_i 的输出值.并假设训练集是 $X\times Y$ 上按照某个未知概率分布 $P(x,y)$ 选取的独立同分布的样本点,又设定损失函数 $c(x,y,f)$,回归问题归结为寻找一个函数 $f(x)$,使期望风险 $R[f]=\int c(x,y,f)\mathrm{d}P(x,y)$ 达到最小.假设所有数据都可以在精度 ε 下用下面的回归函数(预测函数)拟合:

$$f(x)=w^{\mathrm{T}}\varphi(x)+b \tag{1}$$

式中:$\varphi(\cdot)$为一个非线性映射,该映射将输入空间时域数据映射到高维特征空间,使得在特征空间中预测函数可以表示为一个线性回归函数;w 为回归函数的系数.如图 3 当样本点位于两条线之间的带内时,则认为该点没有损失.

可通过优化求解式(1)中的 w 和 b:

$$\min\ \frac{1}{2}\ \|w\|^2$$

$$\text{s.t.}\ \begin{cases} y_i-\langle w,\varphi(x_i)\rangle-b\leqslant\varepsilon \\ \langle w,\varphi(x_i)\rangle+b-y_i\leqslant\varepsilon \end{cases} \tag{2}$$

式中:$\langle\cdot,\cdot\rangle$为求内积符号.考虑到允许回归拟合误差的情况,引入松弛因子 $\xi_i\geqslant 0$ 和 $\xi^*\geqslant 0$,回归估计问题转化为

$$\min\ \frac{1}{2}\ \|w\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{3}$$

$$\text{s.t.}\ \begin{cases} y_i-\langle w,\varphi(x_i)\rangle-b\leqslant\varepsilon+\xi_i \\ \langle w,\varphi(x_i)\rangle+b-y_i\leqslant\varepsilon+\xi_i^* \\ \xi_i\geqslant 0 \\ \xi_i^*\geqslant 0 \end{cases} \quad (i=1,\cdots,n)$$

式中:ξ_i 和 ξ_i^* 为松弛变量;C 为惩罚参数;ε 为不敏感损失函数.

为了求解上述优化问题,采用对偶理论将问题转化为二次规划问题,首先通过 Lagrange 乘子法得到式(3)的 Lagrange 方程:

$$\begin{aligned}(w,b,\alpha^{(*)})=&\frac{1}{2}\|w\|^2-\\ &\sum_{i=1}^{n}\alpha_i(\varepsilon+y_i-\langle w,\varphi(x_i)\rangle-b)-\\ &\sum_{i=1}^{n}\alpha^*(\varepsilon-y_i+\langle w,\varphi(x_i)\rangle+b)\end{aligned} \tag{4}$$

其中 $\alpha^{(*)}=(\alpha_1,\alpha_1^*,\cdots,\alpha_n,\alpha_n^*)\geqslant 0$ 为 Lagrange 乘子,根据 Wolf 对偶原理,对 Lagrange 方程关于 w 和 b 求极小,即对 w 和 b 求偏导数并令其为零,则有

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)x_i=0 \tag{5}$$

$$\frac{\partial L}{\partial b}=\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)x_i=0 \tag{6}$$

把式(5)和式(6)代入可得原问题的对偶问题:

$$\begin{aligned}\min_{\alpha^{(*)}\in R^{2n}}\ &\frac{1}{2}\sum_{i,j}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle\varphi(x_i),\varphi(x_j)\rangle+\\ &\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{n}y_i(\alpha_i^*-\alpha_i)\\ \text{s.t.}\ &\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0\\ &\alpha_i^{(*)}\geqslant 0,\quad i=1,2,\cdots,n\end{aligned} \tag{7}$$

优化上述问题可得最优解 $\bar{\alpha}_i^{(*)}$,并可得到最优的超平面系数向量 $w=\sum_{i=1}^{n}(\bar{\alpha}_i^*-\bar{\alpha}_i)\varphi(x_i)$. 在 SVR 中引入核函数来简化非线性逼近或回归.

核函数满足

$$Q_{ij}=\varphi(x_i)^{\mathrm{T}}\varphi(x_j)=K(x_i,x_j)$$

则式(7)变为

$$\begin{aligned}\min_{\alpha^{(*)}\in R^{2n}}\ &\frac{1}{2}\sum_{i,j}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)+\\ &\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{n}y_i(\alpha_i^*-\alpha_i)\\ \text{s.t.}\ &\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0\\ &\alpha_i^{(*)}\geqslant 0,\quad i=1,2,\cdots,n\end{aligned} \tag{8}$$

核函数的引入,使得函数求解绕过了特征空间,直接在输入空间上求取,从而避免了计算非线性映射 $\varphi(\cdot)$.核函数 $K(x,x')$是对称正实数函数,对于回归函数 $f(x)=w^{\mathrm{T}}\varphi(x)+b$,可以表述为

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i,x)+b \tag{9}$$

常用的核函数定义见表1.

表1 核函数表达式

Tab.1 Expression of kernel function

核函数类型	核函数表达式
线性核函数	$K(x,x_i)=(x,x_i')$
多项式核函数	$K(x,x')=[(x\cdot x_i)+1]^q$
径向基核函数(RBF)	$K(x,x_i)=\exp(-\|x-x_i\|^2)/q^2$
Sigmiod 核函数	$K(x,x_i)=\tanh(\upsilon(x\cdot x_i)+e)$

支持向量回归模型避开了从归纳到演绎的传统过程,实现了高效的从训练样本到预测样本的"转导推理",大大简化了通常的分类和回归等问题,从某种意义上避免了"维数灾难".基于SVR的各种优势,本文选用SVR模型建立股评文本向量与股票收益率之间的回归模型.本文核函数选用RBF函数.

4 股评文本数据对股市收益率走向预测

4.1 数据来源

本文用python对2014年7月1日到12月31日东方财富股吧的上证180的股评文本进行网页爬虫,总共有170万条的文本数据,在数据库中进行初步去重、去空和去噪声,最终得到一个较准确的股评语料库,大约有140万条文本数据,将7~11月的文本数据作为训练文本,12月的数据作为预测文本,文本格式见表2.

表2 训练、预测数据格式

Tab.2 Data format of training and prediction

阅读量	评论数	评论	作者	发表日期
Integer	Integer	String	String	Time

本文所用的股票交易数据来源于同花顺.研究时间窗口为对应的2014年下半年的日收益率.股市分为交易日和非交易日,交易日为工作日;非交易日为节假日和特殊时间段,则丢弃该天对应的股评内容.日收益率计算公式为

$$\eta=\ln(p_t/p_{t-1}) \tag{10}$$

式中:η为日收益率,$\eta>0$时股票价格上涨,$\eta<0$时股票价格下跌;p_t为当日上证180综合指数的收盘价;p_{t-1}为前一天的上证180综合指数的收盘价.

4.2 在R上进行预测

4.2.1 加载文本挖掘所需要的包

本文将会使用到以下包:NLP、tm、jiebaR、jiebaRD、e1071.分词用的是jiebaR包,作者没有使用Rwordseg包,之前尝试过用,分词效果不是很理想.要配置Rwordseg包必须加载rJava包,加载比较麻烦,得有JAVA环境和配置好JAVA的各种路径.经过试验决定用jiebaR包.其他的包就没有那么复杂,只要输入install.packages("包名")就可以安装好,用library(包名)把它们加载到R环境中,就能完成本文所需要的平台环境设置.

从以上可以看出,以R平台做文本挖掘设置简单且方便,不像其他平台那样需要手动加载大量函数和包.它简化了操作,使代码更加简洁.

4.2.2 分词、合并同义词并去停用词

使用read.table()函数将训练数据读入R中,运用jiebaR包中的worker()函数对训练文本内容进行分词,分词之前用gsub()函数去掉文本中的数字、英文字符、特殊符号等,在分词时根据不同的模块选用专业术语的词库,如分词词典,一般选用搜狗细胞词库的专有名词库,也可自己手动生成词库.

分词结束后会出现好多同义词,将这些词用对应的同义词替换后,文本就出现了好多相同的词,例如:出现了购买、买、采购、买进、买了、购置等"买"意思的同义词,于是将这些词用"买"替换掉.这样在生成特征矩阵时可以将这些同义词合并为同一维度,在取高频词时避免将许多有意义的词删除,增加了准确性.本文选用哈工大的同义词林进行同义词合并,见表3.

表3 同义词合并示意数据

Tab.3 Schematic data of synonym combination

利好	买入	走势
上涨	购买	趋势
涨	买	趋向
高涨	采购	势头
高潮	买进	走向
飞腾	买了	发展
增长	购置	走势
增升		形势
上升		长期走势
长		

选用好的停用词表,可以去掉不影响预测结果同时又在所有文本中出现较高频率的词,也可以起到最初降维效果,又可节省不必要的运行时间,使运行效率极大提高.本文的停用词表是在网上下载整理的,总共有4 500多个停用词,例如:了、的、啊、吗、呢等.在R中用tm_map()函数对分词后的股评文本去停用词.

4.2.3 特征权重赋予TFIDF

TFIDF是权重计算的重要算法之一,专门用于评估某个词在整个文档集中的重要程度.一个词在特定的文档中出现的频率越高,说明它在区分该文

档内容属性方面的能力越强(TF)；一个词在文档中出现的范围越广，说明它区分文档内容的属性越低(IDF).其经典算法如下：

$$W_{ij}=\mathrm{tf}_{ij}\times\mathrm{idf}_j=\mathrm{tf}_{ij}\times\log\left(\frac{N}{n_j}\right) \qquad (11)$$

其中：tf_{ij} 指特征项 t_j 在文档 d_i 中出现的次数；idf_j 指出现特征项 t_j 的文档的倒数；N 表示总文档数；n_j 指出现特征项 t_j 的文档数.

在 R 语言中用 tm 包实现 TFIDF，代码为：control=list(removePunctuation=T,dictionary=t$V1,wordLengths=c(1,Inf), weighting=weightTfIdf).

4.2.4　生成矩阵和降维

矩阵是做其他计算、建模工作的基础.在 R 语言中矩阵和数据框是其基本数据结构.只有把分好并去停用词的分词文本转化为矩阵，才能进行后面的建模工作.运用 DocumentTermMatrix()函数把去掉停用词的分词转化为矩阵，再选取在所有文本中占有比例大的词做为特征时选用 removeSparseTerms()函数操作，进而降维简化矩阵.

用 removeSparseTerms()函数降维时，第二个数值参数的范围是 0～1.0.目的是选取在所有文本中出现频率大于等于 1 减去这个参数的分词.假如设置为 0.4，就是选择出现频率大于等于 0.6 的词，重新生成一个降维的矩阵.

东方财富网的股评文本经过一系列文本处理，生成的特征矩阵见表 4.

表 4　生成的特征矩阵示意数据

Tab.4　Schematic data of generated eigenmatrix

发表日期	垃圾	拉升	利好	利空	买入	卖出	…
7 月 1 日	36.062 25	39.560 46	95.574 53	78.556 63	102.512 60	68.939 52	…
7 月 2 日	43.760 21	25.362 90	117.790 19	62.122 09	102.767 26	52.382 44	…
7 月 3 日	53.160 80	38.080 48	119.827 01	60.336 67	122.524 89	85.416 59	…
7 月 4 日	49.004 83	34.922 50	100.457 75	90.604 30	124.254 64	70.321 07	…
7 月 7 日	1.741 33	0	21.255 14	24.205 45	22.755 23	8.451 50	…

4.2.5　生成模型

把得到简化的矩阵按日期从小到大进行排序并进行合并，用支持向量回归模型 e1071 包的 svm()函数进行建模，使用 svm()函数时，可以选择是回归还是分类.当训练文本中已知道的结果值保存为因子时，则自动选择为作分类；如为数值类型，则自动选择为作回归.

将股评特征矩阵作为 SVR 模型的输入，对应股票收益率作为 SVR 的输出，用 scale()函数将数据标准化，再用 svm()函数进行回归模型建立.

4.2.6　得到结果

预测数据经过分词，提取训练数据最终的特征形成最后的矩阵，再用 predict()函数对生成的模型与预测矩阵进行预测.代码 predict(model,test.data)得到最终的预测结果见表 5.

从表 5 可以看出，有 4 天的收益率的实际值和预测值符号相反，说明股票的涨跌趋势预测错误；有 18 天的收益率的实际值和预测值符号相同，说明股票的涨跌趋势预测正确.为了更直观地观察，用＋1 代表该天的收益率为正值，股价上涨；用－1 代表该天的收益率为负值，股价下跌.实际股价涨跌情况如图 4 所示，预测股价涨跌情况如图 5 所示，实际与预测股价涨跌的差值如图 6 所示.

表 5　实际值与预测值

Tab.5　Actual value and forecast value

发表日期	实际值	预测值
12 月 1 日	0.040 882 198	0.026 521 421
12 月 2 日	0.011 269 971	0.018 631 949
12 月 3 日	0.049 342 743	0.099 173 501
12 月 4 日	0.012 127 328	0.018 669 049
12 月 5 日	0.041 871 433	0.041 219 388
12 月 8 日	－0.048 046 219	－0.167 541 368
12 月 9 日	0.033 845 737	0.094 220 531
12 月 10 日	□－0.018 024 703	□0.056 208 762
12 月 11 日	0.003 265 042	0.096 882 422
12 月 12 日	0.006 932 338	0.007 805 614
12 月 15 日	0.033 308 582	0.007 604 775
12 月 16 日	□0.023 290 152	□－0.014 539 240
12 月 17 日	□－0.004 014 561	□0.057 399 837
12 月 18 日	0.013 700 506	0.031 047 660
12 月 19 日	0.007 276 521	0.038 809 125
12 月 22 日	－0.022 263 033	－0.005 209 584
12 月 23 日	－0.033 250 694	－0.064 202 351
12 月 24 日	0.036 321 564	0.057 648 301
12 月 25 日	0.036 788 662	0.056 654 598
12 月 26 日	□0.004 668 342	□－0.027 212 071
12 月 29 日	0.004 668 342	0.024 908 796
12 月 30 日	0.021 801 081	0.081 870 302

注：表中加□数据为预测错误数据.

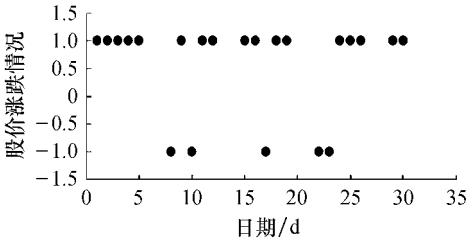

图 4 实际股价涨跌

Fig.4 Volatility diagram of actual stock price

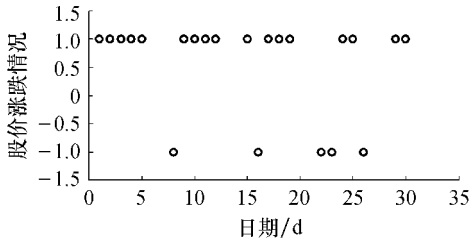

图 5 预测股价涨跌

Fig.5 Volatility diagram of predictive stock price

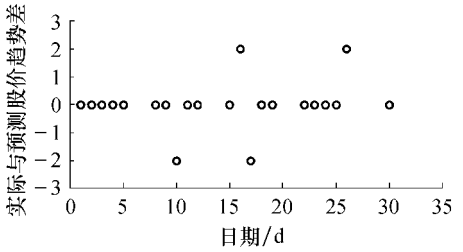

图 6 实际与预测股价趋势差值

Fig.6 Difference graph of actual and predicted stock price trend

从图 6 可以看出，实际股票涨跌趋势与预测股票涨跌趋势一致时，两者的差值为 0；不一致时，两者的差值为±2.只有个别天数预测错误，可得到股票涨跌趋势预测正确率为

$$\beta=\frac{\text{股价涨跌趋势预测正确天数}}{\text{总天数}}=\frac{18}{22}=81.81\%$$

5 结语

1）利用中文文本挖掘技术将非结构化股评文本转化为结构化特征矩阵，并将结构化的特征矩阵与对应股票收益率建立模型，用建立的模型预测未来一个月股价涨跌趋势.

2）将预测日收益率与实际日收益率进行对比，验证网络舆情是否对股市有影响、是否可以通过网络舆情预测股票收益率来预测股价涨跌趋势.

3）股价涨跌趋势预测是可行的，它可以用过去的网络舆论来预测当前以及未来的股票趋势，给股民进行股票投资提供了良好的参考价值，并做好应对的准备.

致谢：本文受到兰州理工大学红柳杰出人才基金项目(J201304)的资助，在此表示感谢.

参考文献：

[1] FAMA E F.The behavior of stock-market prices [J].The Journal of Business，1965，38(1)：34-105.

[2] 郑 冲.中国股市波动性的影响因素研究 [M].北京：北京交通大学出版社，2012.

[3] TAKEDA F，YAMAZAKI H.Stock price reactions to public TV programs on listed Japanese companies [J]. Economics Bulletin，2006，13(7)：1-7.

[4] ANTWEILER W，FRANK M Z.Do US stock markets typically overreact to corporate news stories [J/OL]. (2005-10-24) [2016-10-15].http://www.docin.com/p-93781965.html.

[5] GUNN S R.Support vector machines for classification and regression [R/OL]. (1998-05-14) [2016-10-15].https://wenku.baidu.com/view/a17ce9f90242a8956bece457.html.

[6] TANG X，YANG C，ZHOU J.Stock price forecasting by combining news mining and time series analysis [J].IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies，2009，1：279-282.

[7] 赵 伟，梁 循.互联网金融信息量与收益率波动关联研究 [J].计算机技术与发展，2009，19(12)：1-4.

[8] 赵丽丽，赵茜倩，杨 娟，等.财经新闻对中国股市影响的定量分析 [J].山东大学学报(理学版)，2012，47(7)：70-75.

[9] 马俊伟，王铁军，李 庆，等.基于网络信息挖掘的股市影响因素分析 [J].吉林大学学报(信息科学版)，2014，32(2)：195-200.

[10] 朱昶胜，王莎莎，王永贤.基于 R+Hadoop 的中药材大数据的分析及预测 [J].兰州理工大学学报，2017，43(1)：98-103.

[11] 康 东.中文文本挖掘基本理论与应用 [D].苏州：苏州大学，2014.

[12] 鹿小明.文本挖掘及其在信息检索中的应用 [J].情报资料工作，2004(6)：26-28.

[13] 谌志群，张国煊.文本挖掘研究进展 [J].模式识别与人工智能，2005，18(1)：65-74.

[14] 李 悦.基于支持向量机的股市预测 [M].上海：复旦大学出版社，2014.

[15] 郭明玮，赵宇宙，项俊平，等.基于支持向量机的目标检测算法综述 [J].控制与决策，2014，29(2)：193-199.