**《金融风险管理》 课 程 作 业**

**(2020-2021学年第2学期)**

**论文题目： 基于行为金融的股市情绪量化模型**

**学生姓名： 邱俊源（201830730406）**

**提交日期： 2021年4月26日 学生签名：邱俊源**

|  |  |  |  |
| --- | --- | --- | --- |
| 学号 | 201830730406 | 座位编号 |  |
| 学院 | 经济与金融学院 | 专业班级 | 金融学（汇丰精英班） |
| 课程名称 | 金融大数据 | 任课教师 | 陈艺云 |
| **教师评语：** | | | |
| **本论文成绩评定：** **分** | | | |

# 基于行为金融的股市情绪量化模型

## 投资者情绪对股票市场的理论性影响

上世纪八十年代左右出现了一系列关于情绪和总股本收益的开创性和著名的研究。它们基本上是理论性地以各种方式判断整个股市是否可能被错误定价。作者寻找：总回报趋向于均值回归的趋势；股指收益率的波动率不能用基本面的波动来说明，这实际上表面了收益率均值回归的另一种方式；或通过一些简单的估值比率，比如股息与股票市场价值之比，来说明总回报率的可预测性。

在这些研究中，情感的作用被隐含了，并且统计上经常不显著。实际上，很难将波动与长期泡沫区分开，尤其是在短时间序列中至多可能包含几个泡沫的情况下。即使统计推断似乎很可靠，但经济上仍解释不清楚。股票收益的可预测性证明了情绪引起的定价错误的纠正，或者，与时间相关的风险或风险规避导致预期股票收益根据时间变化而变化。

最近的研究，例如Baker and Wurgler等[1]，利用行为金融理论，对情绪影响进行了更为精确的检验。并且，在DeLong，Shleifer，Summers和Waldmann的启发下，还得出结论：在许多证券市场行为模型中，投资者分为两种类型：无情感的理性套利者和容易产生外生性情感的非理性交易者[2]。他们在市场上竞争并设定价格和预期收益。但是理性的套利者在各个方面受到限制。这些限制来自套利窗口过短无法套利或者发生交易和做空的成本和风险。结果，价格并不总是等于其基本价值。在这样的模型中，错误定价是由两个因素的组合引起的：非理性交易者的情绪变化，以及理性交易者的套利受到限制。

从近些年的研究中我们可以看出投资者的情绪对市场的影响不可忽视，甚至短期内引起股价巨大的波动，因此建立一个可以量化当前市场投资者整体情绪的模型，可以帮助投资者快速地、有效地把握市场情绪，提前做出反应，以避免情绪波动带来的不必要损失。

## 情绪指标的构建

在东方财富个股股吧中，平均每天都有几十条甚至上百条帖子发布。帖子有看涨的，也有看跌的。如何知晓某只股票股吧一天内整体的股民情绪是看涨还是看跌、市场情绪面是好是坏成为情绪分析的重点。

**首先我们爬取了东方财富华东医药股吧的帖子数据。其次，我们用已经标注情感正负的数据集对8个分类器进行K折训练，找出分类效果最好的模型。之后使用该分类器对每个帖子进行情感分类。最后计算出股吧每天的情绪指数。**

### 源数据获取

我们采取第三方软件——后裔采集器[[1]](#footnote-1)爬取东方财富华东医药（000963.SZ）股吧[[2]](#footnote-2)中2014年1月1日至2021年4月3日的所有评论标题、阅读量、评论数等数据，并利用该软件进行简单的数据清洗（去重，去除空值样本）后将其保存为data.xlsx。数据集具体情况如图1所示。



图 1 **爬虫数据集data.xlsx浏览**

### 函数介绍

下表为模型中使用到的函数名称、参数和说明。

表 1 **各函数名称、参数和说明**

|  |  |  |
| --- | --- | --- |
| **函数名称** | **参数** | **使用说明** |
| load\_dataset\_tokenized |  | 读入已经分词好的并且标注了情绪正负的训练文件 |
| remove\_punctuation | 一段文字 | 删除除了字母，数字和汉字以外的所有符号 |
| stopwordslist |  | 加载中文中的停用词 |
| KFold\_validation | 分类器，特征向量，类别 | 对已标注的数据进行K折验证 |
| eval\_model | 分类器 | 对分类器进行训练并对股吧帖子进行情感分类 |

### 使用8种模型对标注好的数据集进行分类并得到模型评价

我们从github中找到了9214条已经由人工标记好情绪正负（即该条评论是看涨还是看跌）的数据集。之后分别利用线性分类支持向量机、逻辑回归、随机梯度下降、朴素贝叶斯、KNN、决策树、随机森林和AdaBoost等8种分类模型对该数据集进行K折验证（其中K=5），找到分类效果最好的模型。

首先，使用load\_dataset\_tokenized函数读入已经完成分词并且标注了情绪正负的训练文件，类别中1表示正向评论，0表示负向评论。之后对8种分类模型分别进行K折验证：

（1）使用sklearn.feature\_extraction.text中的TfidfVectorizer函数将评论进行文本向量化；

（2）把向量化后的文本放入分类器中进行训练并使用测试集进行验证；

（3）将5次评分均值并保存。

8种分类器在该数据集上的表现如图2所示。根据结果，我们选择线性分类支持向量机（LinearSVC）作为分类器对股吧帖子进行分类。

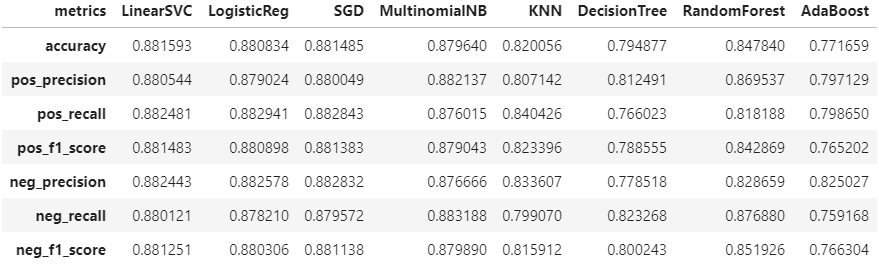


图 2 **8种分类器在标注好的数据集上的分类效果**

### 使用线性分类支持向量机（LinearSVC）对帖子进行情感分类

首先，需要对股吧帖子标题进行分词。我们从github中找到了中文中的停用词数据集，包含“乃至”，“之类”，“大家”等不能体现情感的词语，并用stopwordslist函数读取。配合停用词、remove\_punctuation函数和jieba库，我们对帖子标题进行了处理并分词。

之后使用上述训练集训练线性分类支持向量机，并用训练好的模型对分词好的股吧帖子标题进行分类，得到每条帖子的情感类别（正向—1或负向—0）。

图 3 **对股吧帖子标题进行情感分类**

### 构建股吧每天的情绪指数

我们使用了两种方式来构建股吧每天的情绪指数，分别命名为和，计算方式如下所示。当它们为负时，表示当天的帖子中负面情绪数量更多，反之则正面情绪的帖子更多。



其中，表示正面情绪帖子的数量，表示负面情绪帖子的数量。

根据两种统计方式计算得到2014年1月1日到2021年4月3日东方财富华东医药股吧每天的情绪数值。两个指标的统计数值如下图所示。可以发现相比于，的范围和标准差都更大，换句话说，能更敏感地反映每天的情绪的变化。于是，我们更偏向于选择作为我们的情绪指数。

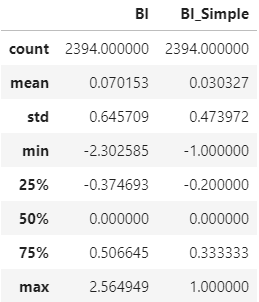


图 4 **两个情绪指标的常见统计量**

图5为2020年1月到5月每天的股吧情绪指标数值。不难发现，股吧短期内情绪变化波动较大，很难发现其中的变化趋势。同时，情绪也具有累计效应[3]。所以，我们使用10天指数加权移动平均来得到今天的情绪数值。具体计算方法如下。



经过平滑，得到的曲线能更好地反映出股吧情绪变化的趋势。将其与收盘价绘制在同一张图中，如图6，我们可以发现情绪与股票价格有着较高的匹配程度。

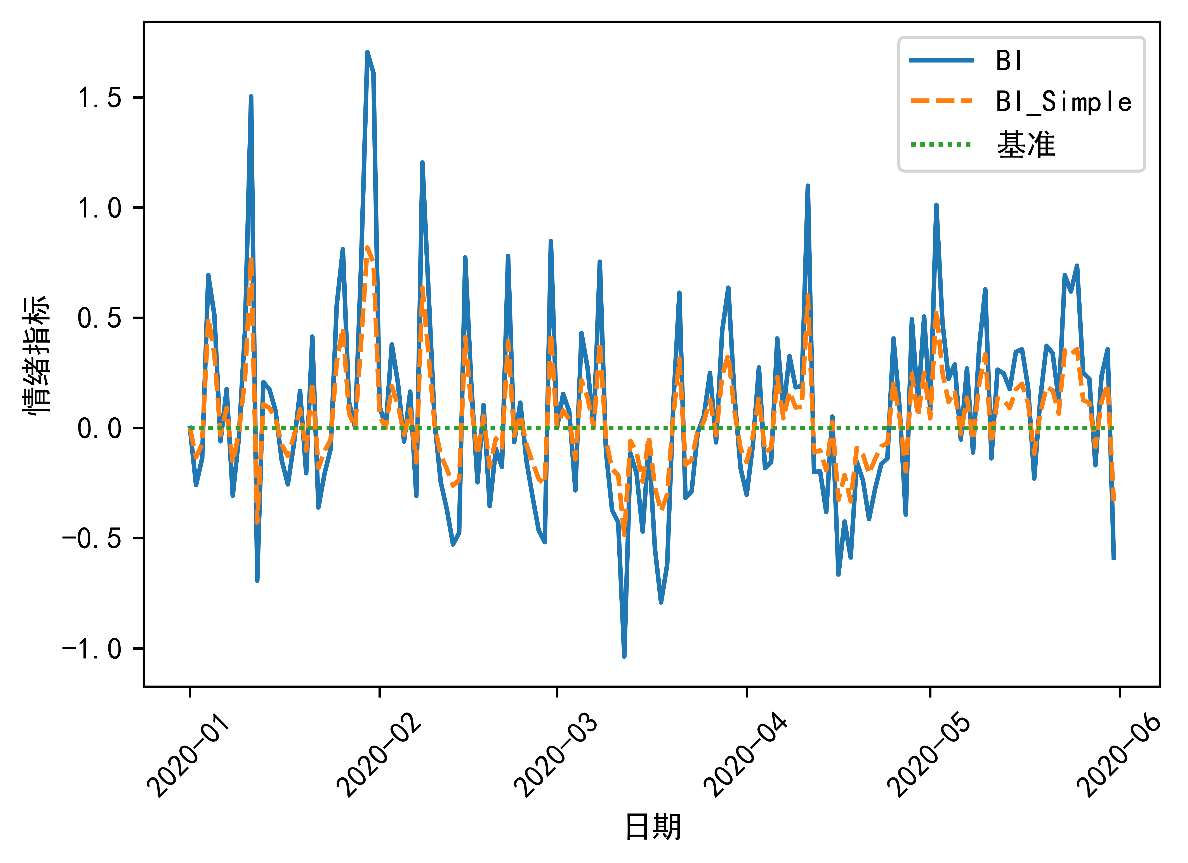


图 5 **股吧每天情绪指数**

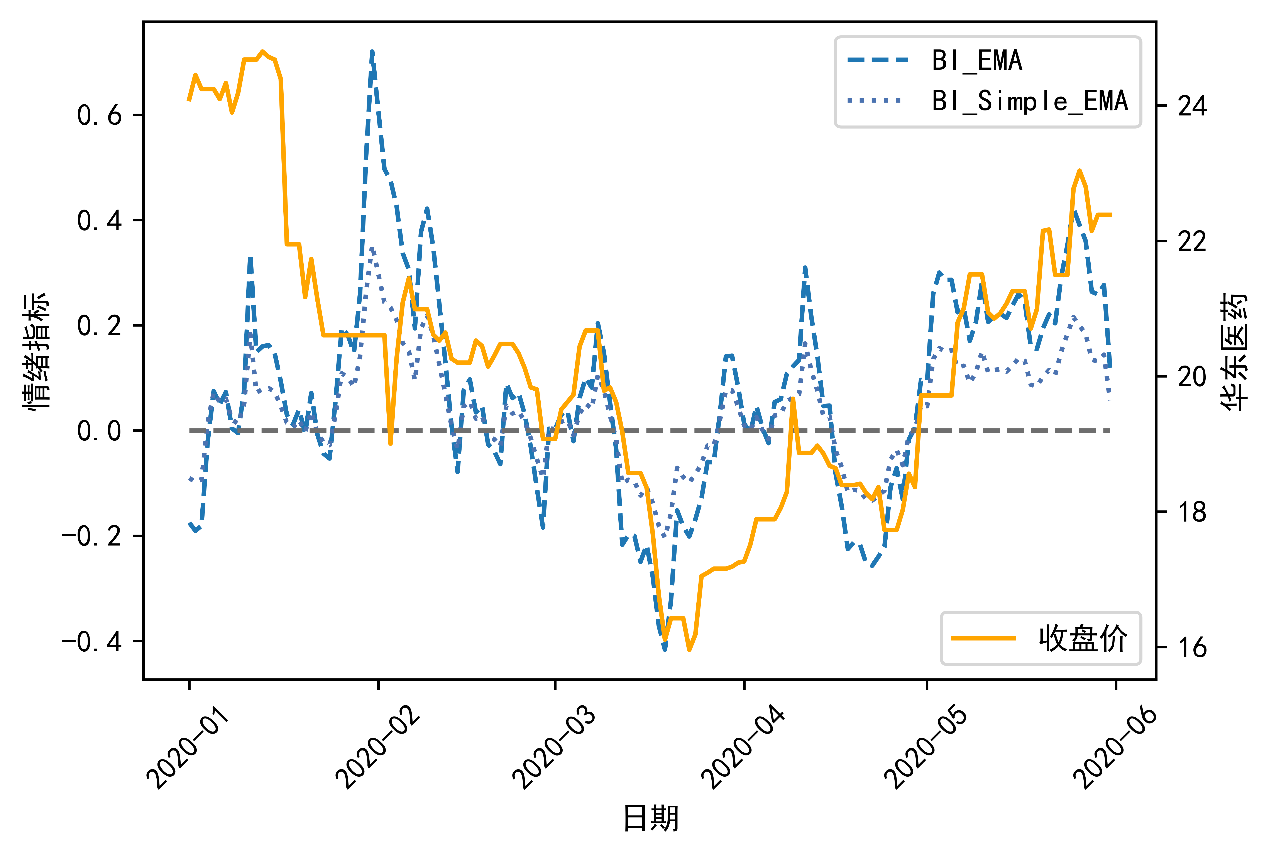


图 6 **经过平滑后的情绪指数与收盘价**

## 股吧避雷

我们希望能够通过分析股吧每天的情绪来预判股票的走势，尤其是能够帮助我们预测未来是否会出现较大的跌幅，以使我们提前退出以保住收益或减少损失。

**首先，我们要构建关于情绪指标的特征，如上文的EWMA，可以反映情绪的趋势。其次，需要选择分类模型，以构建好的特征作为特征向量，未来一段时间的涨跌情况作为标签进行训练，并使用网格搜索（GridSearch）调整模型的参数以使模型达到我们设定的评分标准的最优情况。最后，在测试集上对模型进行检验，以观察模型预测的效果。**

### 特征构建

针对情绪指数，我们构建了6个情绪指标，分别反映之前一段时间内情绪变化的趋势和波动。我们还采用了12个常见的简单技术指标来扩充我们的特征。具体特征如表2所示，特征向量如图7所示。

表 2 **6个情绪指标和12个技术指标**

|  |  |
| --- | --- |
| **情绪指标** | **技术指标** |
| 指标名称 | 指标名称 |
| 前10天情绪指数的均值 | 前10天收益率的均值 |
| 前10天情绪指数的标准差 | 前10天收益率的标准差 |
| 前10天情绪指数的极差 | 前10天收益率的极差 |
| 前10天情绪指数的累计 | 前10天收益率的累计 |
| 前10天情绪指数的加权平均 | 前10天收益率加权平均 |
|  | 前10天每天 （最高价-最低价）/ 最低价的标准差 |
|  | 前10天中 （最高价的最大值-最低价的最小值）/ 最低价的最小值 |
|  | 前10天成交量的均值 |
|  | 前10天成交量的加权平均 |
|  | 前10天成交量的标准差 |
|  | 前10天成交量的（最大值-最小值）/最小值 |

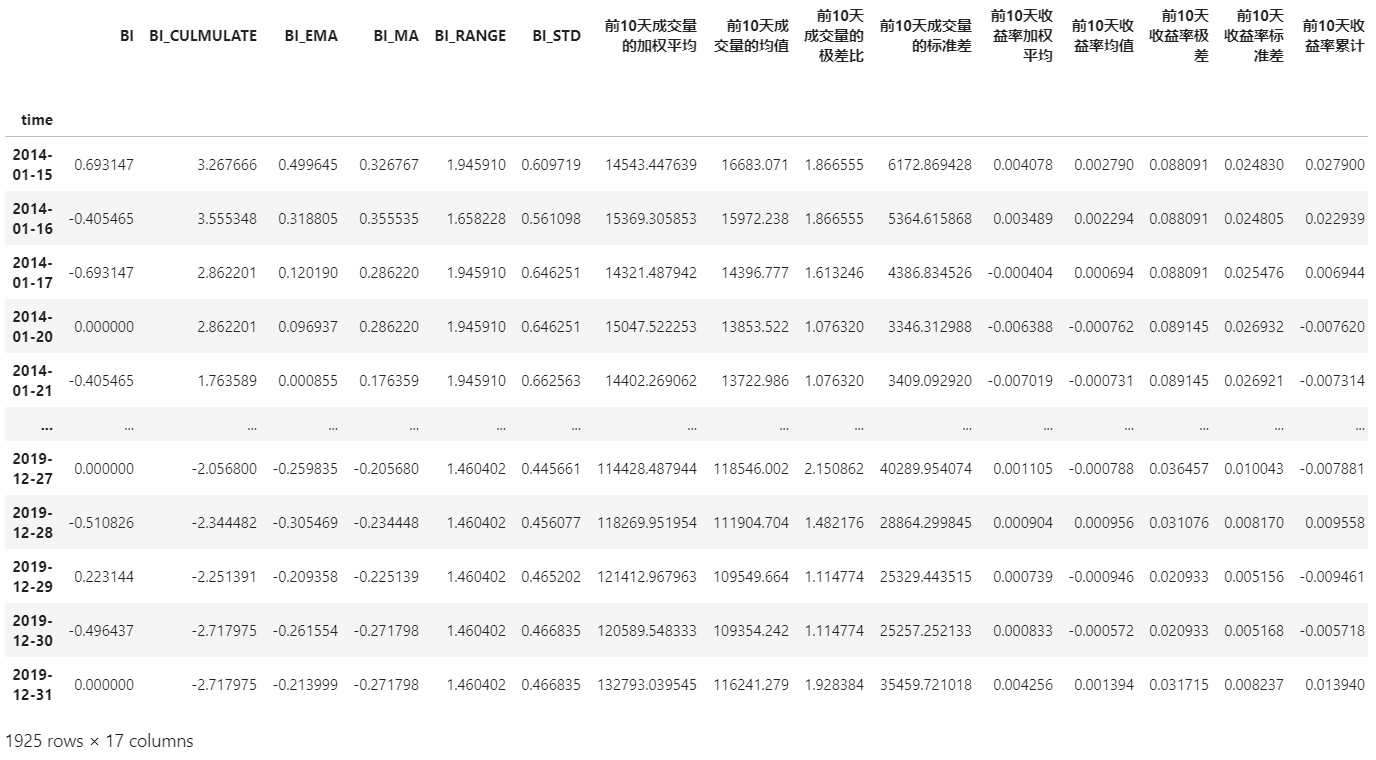


图 7 **特征向量矩阵**

以未来5天的涨跌幅为标签，若未来5天收益率为正，则为类别1，反之则为类别0。在训练集中，类别1占51.39%，类别0占48.61%，不存在类别不均衡的情况。

### 模型选择

因为只有1925条训练样本（2014年1月15日至2019年12月31日），数据量较少，所以我们选择了在小样本数据集上表现较好的分类器：SVM、RandomForest和AdaBoosting。

首先我们对每一个特征用下面的公式进行标准化。



之后，我们设置了一个新的评分指标以满足“避雷”的要求，该指标计算方式如下所示。对类别0，我们使用作为评价指标以使召回率的权重是精准率的 两倍；对类别1，则使用作为评价指标。因为类别0（未来发生亏损）对我们来说更重要，在总体评分时，给予类别0的评价指标更高的权重使其重要形式类别1的5倍。



最后，使用sklearn. model\_selection中的GridSearchCV函数，设置其中的cross-validation函数为sklearn.model\_selection中的StratifiedShuffleSplit以使每次抽样的正负类样本比例和总体比例相同，并使用训练集对三个模型的主要参数进行调参。在具体实践过程中，我们发现SVM和RandomForest十分容易出现高度过拟合的情况，而AdaBoosting在设置弱分类器为单层决策树（decision stump）时不容易出现过拟合的情况。于是，我们以该情况为基准，对弱分类器的ccp\_alpha参数进行调参以使模型达到最优的分数。

网格搜索得到的最优AdaBoost模型参数设置如图8所示。

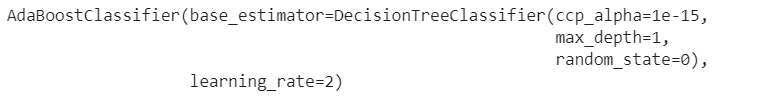
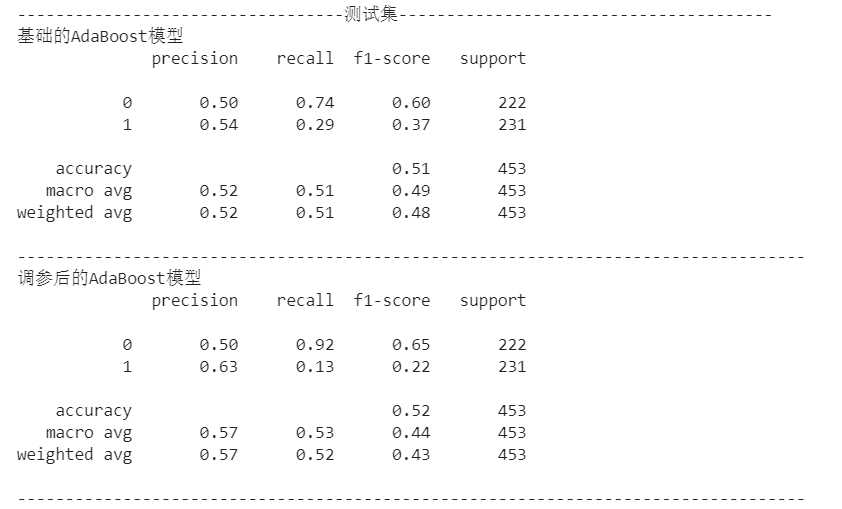


图 8 **最优AdaBoost模型参数设置**

### 模型评估

使用上述的模型对测试集453个样本（2020年1月1日至2021年4月3日）进行预测，基础AdaBoost模型和调参后的AdaBoost模型的预测结果如图9、10所示。可以发现，对类别0的预测中，调参后的模型在精度不变的情况下，拥有更高的召回率。对类别1的预测中，调参后的模型拥有更高的精度，但其召回率却下降了。观察发现，有近9成的数据被预测成了类别0，仅有1成左右的数据被分类为类别1。

之后再经过多次调参和实验，均得不到一个较为满意的结果：类别0分类结果同时具有较高的精度和召回率。推测可能的原因为股票市场影响因素较多，同时存在大量噪音，仅靠这些特征无法十分准确地预测未来股票走势。



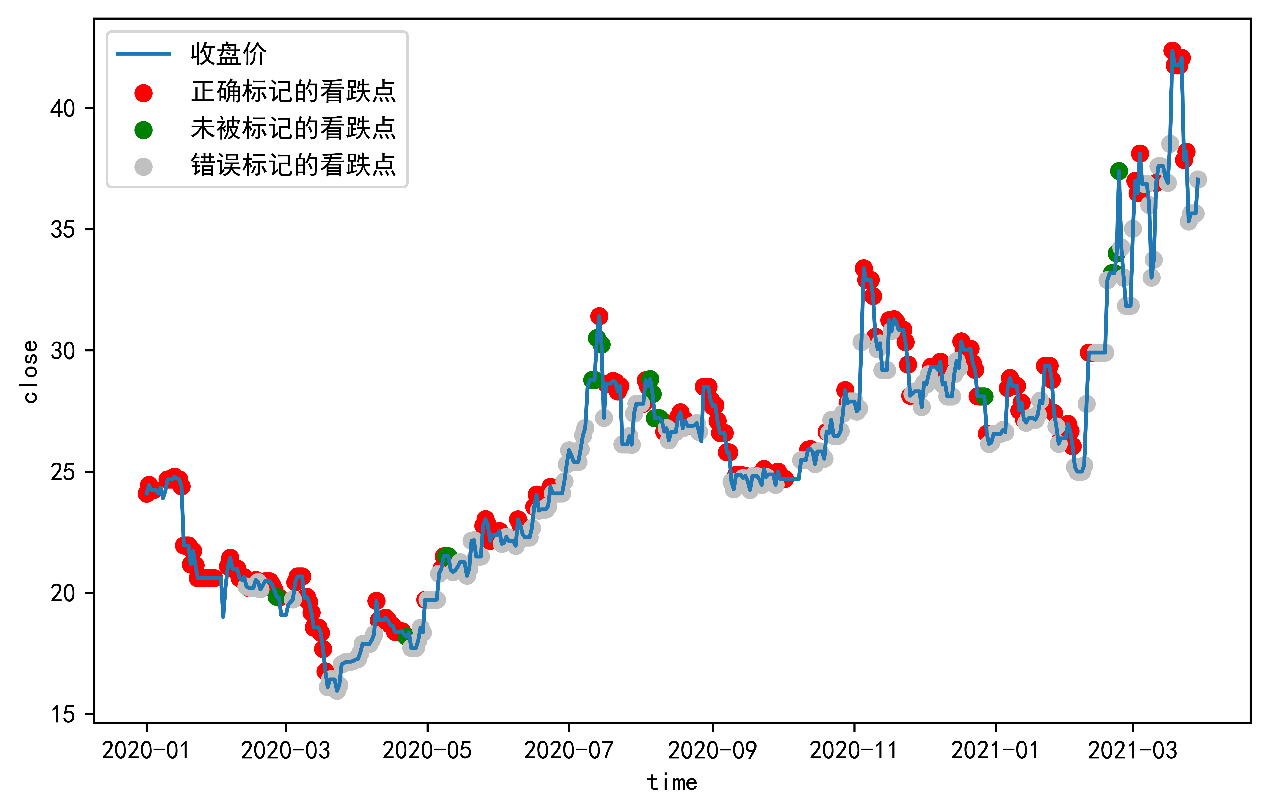
图 9 **测试集测试结果**

图 10 **测试集结果可视化**

## 总结

首先，我们爬取了东方财富股吧的帖子时间和标题。其次通过情感分析划分每条帖子是正向感情还是负向感情。之后构建情绪指数以反映每天股吧整体的情绪。最后使用技术指标和情绪指标预测未来一段时间内股票的涨跌以达到“避险”的目的

**研究发现：**

**1、经过平滑后的情绪指数与股票价格有着较高的匹配程度。**

**2、对未来股票价格的预测难以起到有较高精度的“避险”，推测可能的原因为股票市场影响因素较多，同时存在大量噪音，仅靠这些特征无法十分准确地预测未来股票走势。**

**参考文献**

[1] Baker M,Wurgler J. Investor Sentiment and the Cross‐section of Stock Returns[J]. The Journal of Finance, 2006, 61(4): 1645-1680.

[2] De long JB,Shleifer A,Summers LH, et al. Noise Trader Risk in Financial Markets[J]. Journal of Political Economy, 1990, 98(4): 703-738.

[3] 洪杨闻. 基于投资者情绪的累积效应与资产定价模型研究[D]. 华南理工大学, 2016.

1. 后裔采集器官网：http://www.houyicaiji.com/ [↑](#footnote-ref-1)
2. 东方财富华东医药吧：http://guba.eastmoney.com/list,000963.html [↑](#footnote-ref-2)