**投资者情绪对股价预测影响的行业非对称性**

**——基于随机森林预测效果的比对**

**摘要：**中国股票市场作为一个典型的新兴市场,与国外成熟股票市场相比存在很大差别,其产生的＂金融异常现象＂更加突出和显著。投资者情绪是反映投资者心理的重要因素,系统深入地研究中国股票市场收益、波动性与投资者情绪之间的关系是正确解释中国股票市场＂异常现象＂的有效途径，对于加强中国股票市场的风险管理和控制,改善政府对股票市场的监管效率,保障中国股票市场健康、稳定、持续发展都具有非常重要的理论和现实意义。本文通过对比申万一级行业指数加入投资者情绪和不加入投资者情绪训练随机森林对于股价最终的预测偏差来对比不同行业投资者情绪对股价预测的影响程度。结果发现，随机森林算法对于股指金融序列的预测有良好的表现,并且优于SVR和ANN。对于大部分的股价预测，引入投资者情绪对于确定人们对于某支股票的估值范围帮助最大，但是涨跌方向变化不会非常受投资者情绪的影响。投资者的评论针对日常生活所频繁接触的行业最能充分地反映出投资者内心对该行业股票的内心估值。本文还将申万一级行业指数按照受投资者情绪影响程度分成三类并且对各类行业的金融指标进行了分析，发现受投资者情绪重度影响的行业，存在大量的投资性泡沫，而且这类行业中的企业往往会加大投资或者偿还债务，投资者应当警惕；受投资者情绪轻度影响的行业，股票往往被具有各类投资者情绪的投资者频繁交易，而且其往往是具有良好经营状况的企业且股票在大盘中占有重要地位；不受投资者情绪影响的行业里的企业，往往是具有较强的长期偿债能力，经营状况良好，但是没有太大投资价值的企业。

**关键字：**随机森林 投资者情绪 股价预测 不同行业 非对称性

1. **引言**

精确的股价预测有利于投资者有效地防范风险，管理者制定合理的经营战略和政府更好

地了解经济趋势。股价预测因此总是金融研究的热点话题。然而，这也是在金融研究中普遍被认为是最有挑战性的研究话题，因为它受大量经济的和非经济的因素影响。

在近年来，人工神经网络（ANN）已经成功地被用于金融时间序列的建模（Cheng et al. (1996)， Sharda and Patil (1994) 和David and Suraphan (2005)）。然而，因为研究股市数据是一个多因素分析问题，其中每一个因素可以看作一个维度的指标，这样股市数据就由多个维度的指标体系决定，其中还有大量的噪音，某些研究表明ANN有一些模式学习的局限性。最近，一种叫支持向量机（SVM）的新算法已经被用作金融时间序列的预测（Mukherjee etal.(1997) . Tay and Cao (2001). Kim(2003).）另一种由Breiman（2001）提出的叫随机森林的方法理论则已经被发现在许多研究中表现得更好，例如Creamer and Freund（2004），Lariviere and Poel (2004)。这些机器学习方法例如支持向量机和随机森林成功地实现了在给定其它变量的情况下对某些特定变量进行精确的预测，同时还避免了对一些潜在分布建模时的问题，如神经网络模型可能会倾向于落入局部最优解，而SVM和随机森林的解往往可能是全局最优的。因此，随机森林在回归时往往还可以避免过拟合问题。尽管这些模型已经被科研工作实验过，但目前还没有人尝试过用随机森林的方法来对股市的时间序列预测进行实验。股市同时受经济，政治，金融和社会的因素，噪音和投资者行为影响，股票价格可能在不同的时间范围内有不同的特征（例如分形和混沌的特征）。但是很少研究引入这些价格特征到随机森林中来对股价进行预测。

理论上讲，投资者情绪是投资者信念偏离贝叶斯信念的程度。在一篇较早的系统证明投资者情绪影响资产价格形成的文献中，DeLong et al.（1990）将投资者情绪定义为噪声交易者关于股票未来股价预期偏离理性套利者信念的程度。在行为金融领域，关于投资者情绪及市场波动的经典理论模型较为丰富。Delong et al.（1990）通过建立两类投资者（理性投资者、噪声交易者）的财富效用函数，推导出DSSW资产价格模型，指出噪声交易者创造并承担风险，相比风险厌恶者获得更高的风险回报。Barberis et al.（1998）建立BSV资产价格模型，考察了对不同强度及不同比重的信息，投资者是反应不足还是过度反应。Chen（2011）构造了一个框架来解释投资者情绪的效应并且发现投资者情绪能够影响会计信息的价值相关性。Cornell et al.（2014）研究认为，投资者情绪能够影响预测的收益增长因为投资者通常在投资者情绪高涨时期有对未来积极的态度，并且股票分析者倾向于对那些难于估值的股票发布更高的信用等级。在股市发现一支股票的本质价格是一个决策过程，这可能会影响投资者的心理偏差。这些过于自信的、积极的心理偏差还可能影响金融决定（De Bondit and Thaler（ 1987），Fama and French，（1992）和Thaler（1999））。这些心理偏差会导致股市的牛市或者熊市情绪（Brown and Cliff（2004））。许多作者还证明了这种情绪会控制股市的股票收益（Brown and Cliff（2004）和Fisher and Statman（2000））。然而，很少有人对投资者情绪对股价预测的影响进行分行业的研究。一方面是由于投资者情绪本身数据就很难得，另外一方面是因为股市本身的关联性和复杂性，很难具体研究投资者情绪在不同行业具体的影响。

而本文则利用Kun Guo, Yi Sun and Xin Qian（2016）提出的搜集投资者情绪的方法收集投资者情绪数据，创新性地通过对比申万一级行业指数加入投资者情绪和不加入投资者情绪训练随机森林对于股价最终的预测偏差来对比不同行业投资者情绪对股价预测的影响程度，并且与SVR和ANN的预测效果进行对比。本文第二个部分是投资者情绪数据收集方法和随机森林预测方法的介绍。第三个部分是实证研究和结果。第四个部分是结论。

1. **数据收集和预测方法的介绍**
   1. **投资者情绪数据收集方法的介绍**

我们采用Kun Guo, Yi Sun and Xin Qian（2016）提出的搜集投资者情绪的方法收集投资者情绪数据。原始数据来自雪球网。

1. 第一步：网络爬虫构建：利用Python和Pycharm作为合成的构建环境构建网络爬虫。数据将会利用Webdriver发送请求并且通过内核浏览器收集。
2. 第二步：文字整理：从雪球上收集到的数据是HTML的片段包括大量的HTML标记。为了方便接下来的研究并且得到相对干净的数据，所有的标记都会通过一些特殊的脚本命令去除掉。
3. 第三步：话题分离：话题在这里指的是每支股票的名字。在雪球网里股票名字有一些固定的规则例如“$stock name$”或者“$stock name (stock code)$”。通过这些方法，股票名字的信息可以被从文本中分离出来，所以我们能够从文本中确认出特定的股票。由此，我们便可以统计每支股票的讨论频次。
4. 第四步：情绪分析：情绪分析是获得投资者情绪数据最重要的一步。在这一步，为了从每条评论中获得有用的信息，对中文进行划分是首先需要做的。我们基于中国人民大学的中文情绪汇编构建出情绪分类模型。我们使用被广泛用于机器学习和评估确定事件可能性的Logistic模型来进行分类模型的构建。通过这一步，每一条用户评论都会有确定的情绪，正向的，负向的，中立的。
5. 第五步：情绪指标的构建：因为每一条用户评论都会有一个或两个话题，这意味着一条评论是和一支或者多支股票相关的，并且它有确定的情绪。所以在具体的一天里，一支股票正向的情绪指标是该天对这支股票正向的评论数量。负向和中立的情绪指标由类似的方法构建。同时，我们能够获得由该天这支股票出现的频率所构建的关注指标。因为每一支股票属于特定的行业，我们采用申万一级行业分类来对不同的股票进行划分，行业对应的情绪指标就可能通过包含在这个行业里的所有股票的情绪指标累加而得。由此，整个市场的情绪指标都可以被获得。
6. 第六步：情绪指标的处理：由于雪球网用户群是随着时间的增多而增大的。为了获得更正确的针对某一支股票不同时间的正向，负向和中立的情绪数据。我们选择用每一天正向、负向和中立的评论数量占该天总评论数量的百分比来分别代表该天的正向、负向和中立的投资者情绪数据。
   1. **随机森林**

Bagging（Breiman (1996a)）是并行使集成学习方法最著名的代表，它直接基于自主采样法（Efron and Tibshirani, 1993）。给定包含个样本的数据集，我们先随机取出一个样本放入采样集中，再把该样本放回初始数据集，使得下次采样时该样本仍有可能被选中，这样，经过次随机采样操作，我们得到含个样本的采样集，初始数据集中有的样本在采样集里多次出现，有的则从未出现。照这样，我们可采样出个含个训练样本的采样集，然后基于每个采样集训练出一个基学习器，再将这些基学习器进行结合，这就是Bagging的基本流程。

Breiman（2001）提出的随机森林是Bagging的一个扩展变体。随机森林在以决策树为基学习器构建Bagging集成的基础上，进一步在决策树的训练过程中引入了随机属性选择。具体来说，在随机森林中，一个随机的向量 被生成，独立同分布于原来的随机向量 ；此时有一棵树被用训练集和生成，形成一个分类器 ，其中 是一个输入向量。 随机选择组成大量的独立的从到的随机整数。的性质和维度取决于它在树的构成中的作用。在大量树被生成以后，它们投票选择出最适合的分类。这个过程被称作随机森林。

一个随机森林是一个由多个树型结构基分类器构成的分类器的集成。其中，  是独立同分布的随机向量，并且每棵树为输入的最适合的分类的投出一票。

给定一个分类器的集成 并且随着训练集从随机向量的分布被随机生成，定义边际函数为



其中  是指示函数。这个边际衡量了在 中，正确分类的平均选择数量超出其它分类的程度。这个边际越大，这个分类的置信度越高。总误差由下面的公式给出



其中表明在 空间上的可能性。

在随机森林中， 。 对于大量的树，它服从强大数定律，并且随着树的数量的增加，几乎可以确定出所有的序列 ，…

收敛于



由上面等式的收敛，我们可以知道随机森林不会随着树的增加而过拟合，但会产生一个总误差的极限值。

传统决策树在划分属性选择时是在当前节点的属性集合（假定有个属性）中选择一个最优属性；而在随机森林中，对基决策树的每个节点，先从该节点的属性集合中随机选择一个包含个属性的子集，然后再从这个子集中选择一个最优属性用于划分。这里的参数控制了随机性的引入程度：若令，则基决策树的构建与传统决策树相同；若令，则是随机选择一个属性用于划分；一般情况下，推荐值（Breiman（2001））。

这个策略被采用来达到下面的结果：

1. 在生成的树达到最大的深度的情况下，保证个体误差值足够低。
2. 通过下面两点保证足够低的随机残差相关性：
3. 通过来自训练的数据的样本生成每棵树。
4. 规定  (属性的数量)。在每棵树选择的节点上，选择  个属性并且选择出基于这些属性生成的节点的最佳分割。

随机森林中基学习器的多样性不仅来自样本扰动，还来自属性扰动，这就使得最终集成的泛化性能可通过个体学习器之间差异度的增加而进一步提升。随机森林的起始性能往往相对较差，但随着个体学习器数目的增加，随机森林通常会收敛到更低的泛化误差，而且训练效率常优于Bagging。

总之，对于二进制的输出， 随机森林构造了一个分类树的集成。每棵树是由在每一个分割的导入的样本数据所建立的；用一个随机样本的预测来检验。最终分类的确定是由分类树集成基于每个情况下的最多投票数来确定的。

* 1. **支持向量回归机（SVR）**

支持向量机（SVM）是一种基于统计学习理论的机器学习方法，采用结构风险最小化原则，是处理小样本数据的有效方法。支持向量回归机（SVR）是支持向量机的一种，主要用于解决回归问题。其基本思想是：给定训练集T = {}，其中，输出变量, i = 1...n，通过非线性映射将输入变量映射到高维空间，然后在高维空间中进行线性回归。选取合适的精度和惩罚参数C>0，原始问题可以表示为：

引入核函数K，原始问题可以转化为以下对偶问题：

求解对偶问题，得到最优解

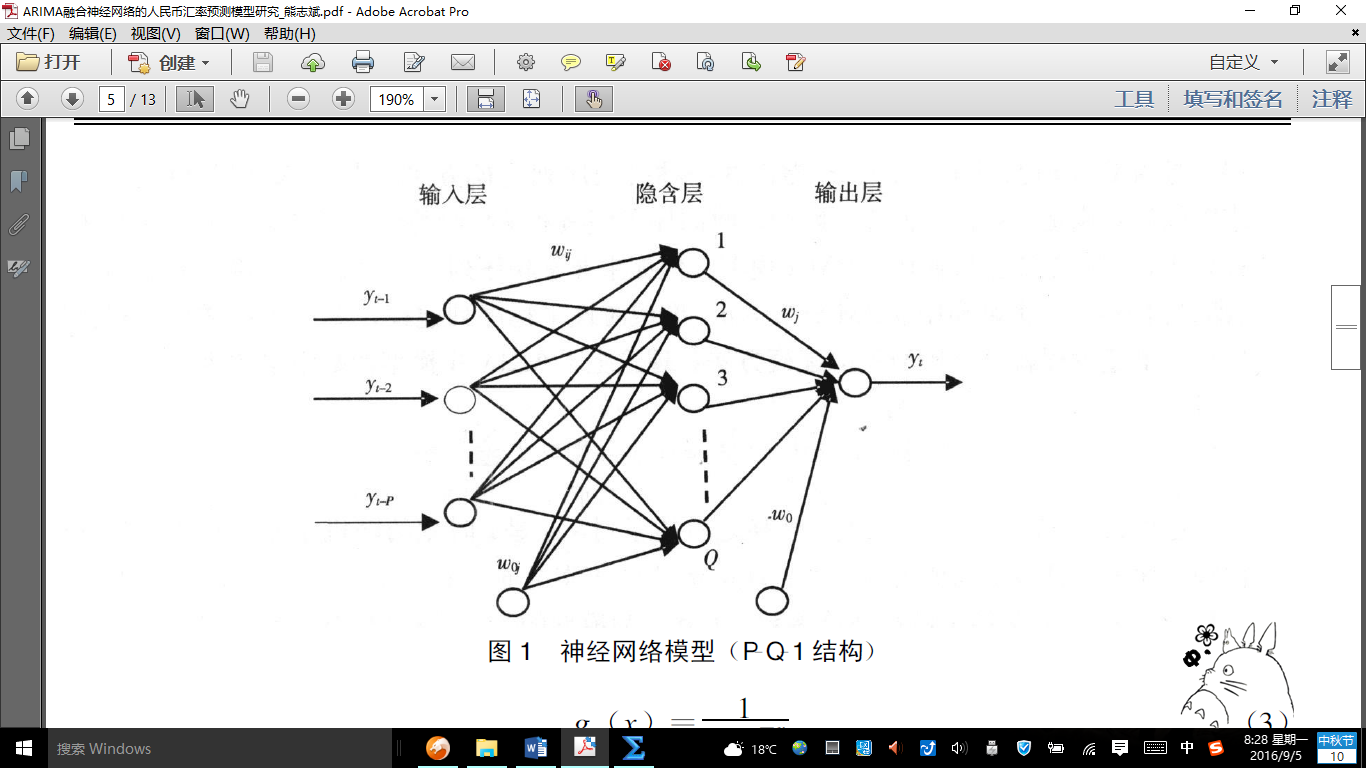
计算：选取位于区间（0，C）中的分量或，如果选到的是，则

如果选到的是，则

构造并求出决策函数：

* 1. **BP神经网络**

迄今为止, 应用于金融时间序列预测的ANN 模型种类很多，ANN是由大量简单的基本元件——神经元相互连接，通过模拟人的大脑神经处理信息的方式，进行信息并行处理和非线性转换的复杂网络系统，前向反馈（BP）和径向基（RBF）网络是目前技术最成熟、应用范围最广泛的两种。其中经典的基于误差反向传播(Back Propagation,BP)算法的多层前馈神经网络(BP 神经网络), 是应用最广的一种ANN 模型。具体结构如下图：



在模型中，输出和输入的关系可用下式表达：



式中，和为模型参数（连接权值向量和阈值向量）；P是输入层节点数；Q是隐含层节点数；g是转换函数，常用的是logistic函数，即：。

BP网络是一种具有三层或者三层以上神经元的神经网络。BP算法的核心是数学中的“负梯度下降”理论，即BP网络的误差调整方向总是沿着误差下降最快的方向进行，常规三层BP网络权值和阈值调整公式如下：





其中，为网络输出和实际输出样本之间的误差平方和，为网络的学习速率即权值调整幅度，为时刻输入层第个神经元与隐含层第个神经元的连接权值；为时刻输入层第个神经元与隐含层第个神经元的连接权值；为时刻输入层第个神经元与隐含层第个神经元的连接权值；为时刻输入层第个神经元与隐含层第个神经元的连接权值；为神经元的阈值，下标的意义与权值的相同，这就是BP神经网络的学习规则。

1. **实证研究与结果**
   1. **实证数据和评估指标**

我们选取2011年1月1日到2016年8月31日的申万一级各行业指数的前收盘价、开盘价、最高价、最低价、收盘价、成交量（股）、成交金额（元）、涨跌、涨跌幅（%）作为未引入投资者情绪之前的特征序列（数据来源于Wind资讯金融终端）。我们利用上述的投资者情绪收集方法获得了2011年1月1日到2015年7月31日的投资者者情绪数据。最后，根据各行业可获得的投资者情绪数据和申万一级各行业指数的前收盘价、开盘价、最高价、最低价、收盘价、成交量（股）、成交金额（元）、涨跌、涨跌幅（%），剔除出现空值的天数后整理出各行业的相关数据进行实证研究。我们将选择前一交易日（t-1）的前收盘价、开盘价、最高价、最低价、收盘价、成交量（股）、成交金额（元）、涨跌、涨跌幅（%）作为未引入投资者情绪时的特征序列来通过随机森林训练预测当天的收盘价（t），然后对比引入前一天的投资者情绪数据作为新的特征序列来预测当天的收盘价（t），通过对比不同行业的预测误差即可探究出投资者情绪对不同行业的股价影响程度。

* + 1. **评估指标**

我们选用均方误差（MSE），均值绝对预测误差（MSPE），方向对称性（DS）来作为评估指标。MSE和MAPE的值越小，则认为预测效果越好；而DS的值越大，则认为预测效果越精确。评估指标的具体式子如下：







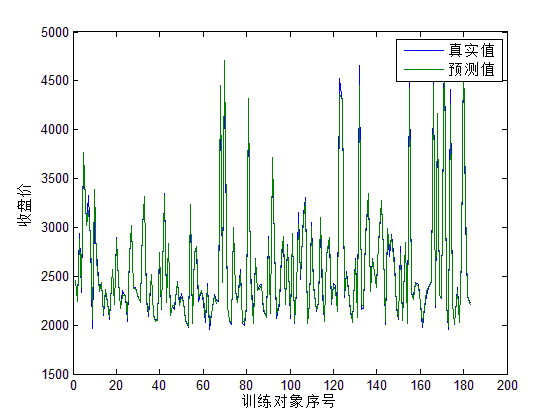


其中，是在时间t处真实的收盘价，是在时间t处预测的收盘价，是真实值或预测值的具体数量。

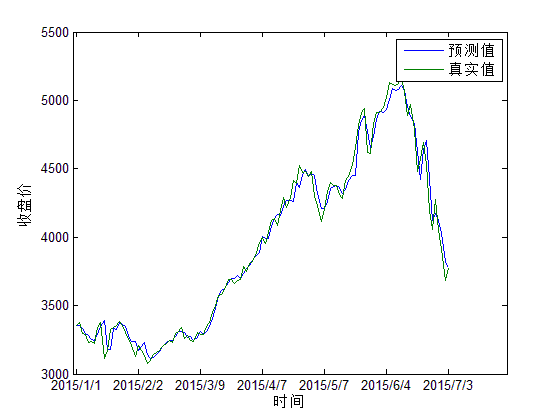
* 1. **投资者情绪对于大盘股价影响程度的实证探究**

为了尽量排除金融系统各种复杂性因素影响而尽可能地控制单一变量只有投资者情绪。我们选择用所能获得的所有数据来构建样本集，随机森林将随机抽取样本集中的进行训练，剩下的进行测试，为了使训练出来的模型尽可能地精确，我们设置的树木个数为100。然后我们人为地选择2015年1月1日所能获得的所有数据作为预测的特征数据，来预测大盘的收盘价，其中大盘是用上证综指的数据来代表（数据来源于Wind资讯金融终端）。

训练结果如下图：



从图中我们可以看出，训练结果的预测值与真实值的图像拟合度极高，说明随机森林算法对于股指金融序列的预测有良好的表现，其所构建的预测模型可被用于我们接下来的股指金融序列的预测实验。我们导入2015年1月5日至2015年7月3日每天前一交易日（t-1）的前收盘价、开盘价、最高价、最低价、收盘价、成交量（股）、成交金额（元）、涨跌、涨跌幅（%）进入随机森林中作为预测集的特征数据。对该交易日（t）的收盘价进行预测。预测结果如下图：



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 操作次数 | MSE | | MAPE | | DS | |
| 未引入投资者情绪 | 引入投资者情绪 | 未引入投资者情绪 | 引入投资者情绪 | 未引入投资者情绪 | 引入投资者情绪 |
| 1 | 7449.838173 | 4526.812085 | 0.013067424 | 0.011613588 | 97 | 96 |
| 2 | 5979.952508 | 4625.360829 | 0.012073096 | 0.012175447 | 95 | 93 |
| 3 | 4348.181411 | 4709.363132 | 0.01167269 | 0.011815204 | 92 | 93 |
| 4 | 5190.705097 | 6267.528252 | 0.012083833 | 0.013147943 | 99 | 88 |
| 5 | 4824.69396 | 4852.847905 | 0.011486361 | 0.011613839 | 96 | 91 |
| 平均 | 5558.67423 | 4996.38244 | 0.012076681 | 0.012073204 | 95.8 | 92.2 |

由上图两条曲线的波动和涨跌幅度趋势的一致性可得，通过引入前一交易日（t-1）的相关股票价格指标到随机森林中进行训练后预测当天交易日（t）的收盘价的效果是非常好的。为了排除随机森林的随机性的影响，我们对每组均进行5次实验后，分别计算出上述的评估指标MSE、MAPE和DS后取其平均值。得到的关于大盘的预测效果比对如下表：

注：MSE和MAPE越小，认为预测的数值范围越精确。DS越大，认为预测的涨跌方向越准确。

由上表可得，引入投资者情绪后的预测效果，MSE和MAPE均小于未引入投资者情绪的预测效果。这说明在随机森林对于大盘股价的数值预测中，引入投资者作为特征序列可以提高预测的数值精度。但是引入投资者情绪后的预测效果DS小于未引入投资者情绪的预测效果。这说明在对于涨跌的方向性预测中，引入投资者情绪后的预测效果反而是会对预测造成干扰的。

这说明投资者情绪的确会影响股价的数值，我们可以通过股市中的投资者情绪来提高对股价范围的预测精度，这意味着投资者情绪可以用来反映股民心中对某支股票价格范围的具体期望。然而，投资者情绪并不能反映投资者实际的操作，例如，正向评论的投资者并不一定会买入该支股票，有可能是持观望态度或者甚至是在高位抛售股票赚取利润。

* + 1. **随机森林、SVR和ANN的预测效果对比**

另外，为了检测随机森林对于引入投资者情绪后对于股价预测的优越性，我们在引入投资者情绪的情况下分别用随机森林、SVR和ANN进行5次预测实验后算出MSE、MAPE和DS的均值来对比三者的预测效果。其中SVR模型我们选择Charles li SVR 1.0[[1]](#footnote-1)，该工具箱使用遗传算法对三参数进行寻优处理，借鉴了 libsvm-[farutoUltimate3.1]的部分内容，根据预测需求删改了部分内容，并对函数句柄加以优化，适合用于金融产品等其他数据的预测问题。ANN模型我们利用MATLAB自带的newff 函数建立BP神经网络，以学习速度为0.05来训练神经网络， 网络误差如果连续6次迭代都没有变化，训练将会自动终止（系统默认）。然后利用训练好的网络进行预测。

最后的结果如下表：

|  |  |  |  |
| --- | --- | --- | --- |
| 预测模型 | MSE | MAPE | DS |
| 随机森林 | 4996.38244 | 0.012073204 | 92.2 |
| SVR | 16517.54286 | 0.024351907 | 58.4 |
| ANN | 8619.069908 | 0.016684611 | 75.8 |

注：MSE和MAPE越小，认为预测的数值范围越精确。DS越大，认为预测的涨跌方向越准确。

由上表可得，随机森林预测的MSE和MAPE均比SVR和ANN的小，而且DS均比SVR和ANN大，从收盘价预测效果的精确度来看， ANN优于SVR，而随机森林则均优于SVR和ANN。这说明随机森林对于股价预测的确比SVR和ANN表现得更好。

* 1. **投资者情绪对不同行业股价影响程度的对比和分析**

下面，我们按照上述操作对申万一级行业分类中各类股指的收盘价分别在引入每天前一交易日（t-1）的前收盘价、开盘价、最高价、最低价、收盘价、成交量（股）、成交金额（元）、涨跌、涨跌幅（%）的基础上，不引入投资者情绪和引入投资者情绪进行5次预测后计算每次的MSE、MAPE、DS后取均值。只有当MSE与MAPE计算出来的情况一致时，我们才认为这个实验结果是有效的，否则我们将对该行业指数的预测再分别重复5次实验后取均值。最后的结

果如下表：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 申万一级行业指数 | MSE | | MAPE | | DS | | 预测结果对比评价 | | |
| 未引入投资者情绪 | 引入投资者情绪 | 未引入投资者情绪 | 引入投资者情绪 | 未引入投资者情绪 | 引入投资者情绪 | MSE | MAPE | DS |
|
| 电气设备 | 22757.00 | 22424.46 | 0.013263 | 0.012517 | 95.6 | 99.6 | 1 | 1 | 1 |
| 电子 | 4960.90 | 4692.94 | 0.013181 | 0.012862 | 97.4 | 97.6 | 1 | 1 | 1 |
| 非银金融 | 2578.96 | 2401.96 | 0.013849 | 0.013514 | 95.0 | 97.0 | 1 | 1 | 1 |
| 国防军工 | 3929.22 | 3897.25 | 0.016443 | 0.015862 | 99.2 | 100.0 | 1 | 1 | 1 |
| 建筑材料 | 12974.67 | 11493.46 | 0.01248 | 0.012156 | 91.8 | 94.4 | 1 | 1 | 1 |
| 休闲服务 | 16921.19 | 13453.10 | 0.011854 | 0.011285 | 96.6 | 99.4 | 1 | 1 | 1 |
| 商业贸易 | 22339.71 | 20037.86 | 0.013951 | 0.01339 | 90.6 | 92.8 | 1 | 1 | 1 |
| 交通运输 | 8175.73 | 7127.16 | 0.014294 | 0.01373 | 91.2 | 91.4 | 1 | 1 | 1 |
| 医药生物 | 29089.22 | 24030.62 | 0.012311 | 0.011727 | 92.4 | 92.8 | 1 | 1 | 1 |
| 纺织服装 | 6612.76 | 6426.77 | 0.013213 | 0.012608 | 89.8 | 94.4 | 1 | 1 | 1 |
| 钢铁 | 7992.01 | 7392.34 | 0.015122 | 0.014885 | 93.4 | 95.4 | 1 | 1 | 1 |
| 有色金属 | 9648.45 | 8109.81 | 0.013836 | 0.012945 | 93.2 | 96.2 | 1 | 1 | 1 |
| 综合 | 6545.80 | 5945.87 | 0.012381 | 0.012109 | 93.0 | 96.2 | 1 | 1 | 1 |
| 银行 | 3427.99 | 3239.47 | 0.011644 | 0.011272 | 98.6 | 98.0 | 1 | 1 | 0 |
| 采掘 | 10475.47 | 8850.25 | 0.014395 | 0.013798 | 95.4 | 95.2 | 1 | 1 | 0 |
| 食品饮料 | 13744.46 | 10908.37 | 0.010028 | 0.009477 | 96.8 | 96.0 | 1 | 1 | 0 |
| 家用电器 | 9628.74 | 8949.47 | 0.011842 | 0.01126 | 98.0 | 98.0 | 1 | 1 | 0 |
| 化工 | 5621.30 | 4568.05 | 0.012179 | 0.011705 | 94.4 | 94.0 | 1 | 1 | 0 |
| 农林牧渔 | 5537.48 | 6180.83 | 0.012809 | 0.01286 | 91.6 | 92.8 | 0 | 0 | 1 |
| 汽车 | 9786.27 | 10101.23 | 0.01115 | 0.011414 | 100.2 | 100.4 | 0 | 0 | 1 |
| 计算机 | 32489.83 | 35503.81 | 0.017441 | 0.017824 | 97.4 | 97.0 | 0 | 0 | 0 |
| 传媒 | 1629.51 | 1725.00 | 0.014147 | 0.014486 | 100.6 | 95.8 | 0 | 0 | 0 |
| 机械设备 | 1783.02 | 2085.08 | 0.012899 | 0.013202 | 95.2 | 94.0 | 0 | 0 | 0 |
| 轻工制造 | 5080.60 | 5244.80 | 0.01128 | 0.011522 | 102.6 | 100.8 | 0 | 0 | 0 |
| 建筑装饰 | 8671.81 | 9886.98 | 0.015589 | 0.016198 | 95.6 | 93.4 | 0 | 0 | 0 |
| 公用事业 | 5493.00 | 6483.20 | 0.012679 | 0.013426 | 94.2 | 93.6 | 0 | 0 | 0 |
| 房地产 | 10530.82 | 10530.82 | 0.012684 | 0.012872 | 94.0 | 92.0 | 0 | 0 | 0 |
| 通信 | 5177.51 | 6050.11 | 0.013436 | 0.013778 | 95.0 | 93.6 | 0 | 0 | 0 |

注：MSE和MAPE越小，认为预测的数值范围越精确。DS越大，认为预测的涨跌方向越准确。预测结果对比评价处，0表示引入投资在这情绪对预测效果的准确度没有提高。1表示引入投资者情绪对预测效果的准确度有提高。

由上表可以看出，电气设备、电子、非银金融、国防军工、建筑材料、休闲服务、商业贸易、交通运输、医药生物、纺织服装、钢铁、有色金属、综合的股指在引入投资者情绪后，无论是对数据预测范围的精确度还是对于涨跌方向的准确度都有所提高；而银行、采掘、食品饮料、家用电器、化工的股指则在引入投资者情绪后，只对数据预测范围的精确度有所提高而对涨跌方向的准确度没有提高；与之相反的是，农林牧渔、汽车的股指在引入投资者情绪后，对数据预测范围的精确度没有提高却对涨跌方向的准确度有所提高。至于计算机、传媒、机械设备、轻工制造、建筑装饰、公用事业、房地产、通信的股指的预测，引入投资者情绪反而会降低对数据和涨跌方向的预测效果。

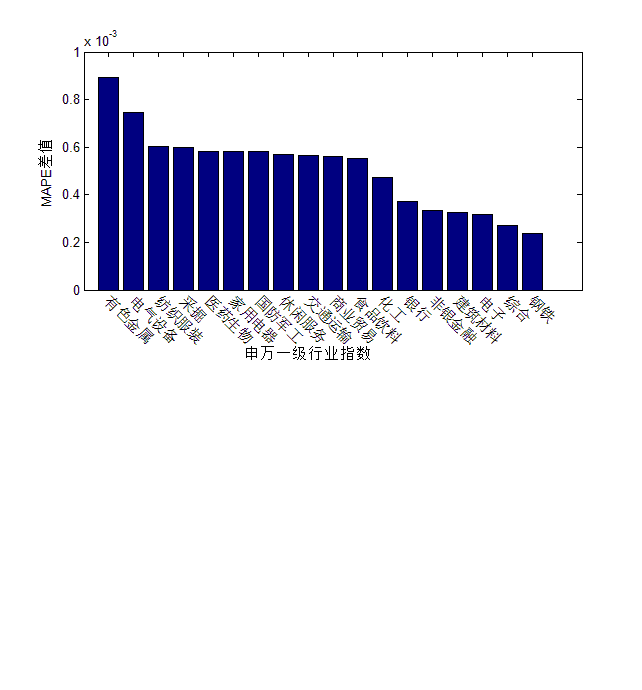
在28个行业分类中，有20个行业分类，即71.43%的行业在引入投资者情绪后提高了对股指收盘价的预测效果，其中有64.29%的行业在引入投资者情绪后有利于提高对数值的预测效果，有53.57%的行业在引入投资者情绪后有利于提高对涨跌方向的预测效果，有46.43%的行业在引入投资者情绪后有利于提高对数值和涨跌幅的预测效果。

由此说明，对于大部分的股价预测，引入投资者情绪作为指标来考虑是非常有必要性的，对于确定人们对于某支股票的估值范围帮助最大，同时，对于大部分行业尽管某支股票可能会被热烈讨论，甚至投资者情绪高涨，但是只要股指在投资者期望的价格范围内，它的涨跌方向变化就不会非常受投资者情绪的影响。由此也可以说明，羊群效应在中国股市的投资者中已经开始减弱，人们开始更关注股票的业绩，而且对于股票的买卖逐渐回归对股票业绩的考察而非概念的炒作。

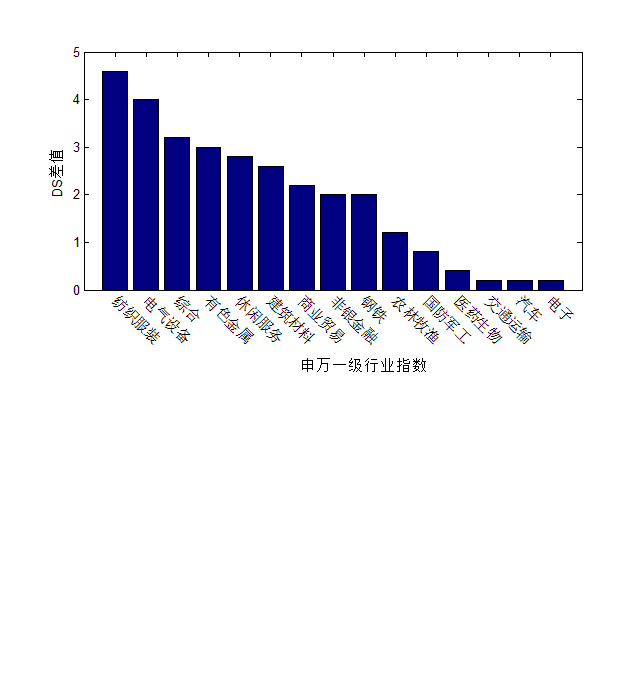
2016年，上海和深圳交易所明显加大了交易的监管力度，对交易中影响价格波动的异常申报也进行了严格监管，随着这样的监管强化，必然会降低市场股票价格的波动幅度，对凭借资金大的优势，通过打高股价出货或打低股价低吸获取交易性博弈投资收益，甚至做庄拉高股价套现获取投资收益的盈利模式的投资者来说，这样的监管力度无疑对他们是巨大的打击。证券监管部门通过对交易异常申报的强化监管，以规范对市场的短炒和博弈性资金差价操作。目前管理层全面加大了监管力度。从重组新规、信息披露、交易异常申报、交易特停、欣泰直接退市、处罚力度等都反映了前面三个监管正在推进，随着这样的监管力度加大，投机成本必然加大。管理层是有引导A股投资理念向价值为基础的合理健康的市场转变的思路。价值投资与公司成长发展相结合是投资的基石，长期看无疑对资本市场发展和投资获得持续赚钱效应都是有好处的，对过去长时间靠炒作、打涨停板获取博弈收益的人来说，是一个打击，短期市场会波动减少。从中长期看，市场回到按照价值＋成长的思路挖掘行业和个股机会。

* + 1. **不同行业股指引入投资者情绪后对数据预测提高效果的对比分析**

为了深入对比不同行业引入投资者情绪的预测具体效果，我们将采用消除了量纲的各股指的MAPE和DS未引入投资者情绪和引入了投资者情绪的差额来进行对比分析。得到的结果如下图：

****

注：MAPE差值越大，说明投资者情绪对于该行业指数的收盘价预测的数值范围的精确度的提高效果越显著。



注：DS差值越大，说明投资者情绪对于该行业指数的收盘价预测的涨跌方向的准确度的提高效果越显著。

为了消除实验随机性的影响，我们取各行业分类的预测提高效果的平均排名进行排序，确定出投资者情绪对股指预测效果提高的行业排序如下：

|  |  |  |
| --- | --- | --- |
| 申万一级行业指数 | 平均排名 | 排序 |
| 电气设备 | 2 | 1 |
| 纺织服装 | 2 | 1 |
| 有色金属 | 2.5 | 3 |
| 休闲服务 | 6.5 | 4 |
| 医药生物 | 8.5 | 5 |
| 商业贸易 | 8.5 | 6 |
| 国防军工 | 9 | 7 |
| 综合 | 10 | 8 |
| 建筑材料 | 10.5 | 9 |
| 交通运输 | 11 | 10 |
| 非银金融 | 11 | 11 |
| 钢铁 | 13.5 | 12 |
| 电子 | 15.5 | 13 |

注：平均排名越小，说明投资者情绪对于该行业指数的收盘价预测的数值范围和涨跌方向提高效果越显著，排序越前。

由上表可得，投资者情绪对于电气设备、纺织服装、有色金属、休闲服务、医药生物、商业贸易的预测的数值范围和涨跌方向提高效果是最显著的。这些大部分都是属于普通大众熟悉的和最频繁接触的东西。由此说明，投资者的评论针对日常生活所频繁接触的行业最能充分地反映出投资者内心对该行业股票的内心估值。而对于其它行业，虽然投资者情绪也能在一定程度上提高预测的效果，但是仅仅关注网络评论并不能真实地反映股民内心对该行业股票的估值。对于这些其它行业，投资者在网络上的评论很多时候只是评论而已，并不会真实地反映他们的下一步的投资操作。而对于电气设备、纺织服装、有色金属、休闲服务、医药生物、商业贸易这些与普通大众日常生活息息相关的行业，网络评论很大程度上可以反映出投资者未来的投资策略，因此引入投资者情绪作为这些行业股指预测的特征序列是最有利于股指预测的。由此说明，中国投资者很大程度上仍然是理性的，舆论与实际操作往往是分开的，对于投资者了解得越全面的行业，投资者的舆论与他们的实际操作才会越密切。

* + 1. **不同行业金融指标对比分析**

为了具体研究和分析不同行业受投资者情绪影响的本质原因。我们将上述的申万一级行业指数按照受投资者情绪影响程度分成三类，分别是股指受投资者情绪深度影响的行业，包括电气设备、纺织服装、有色金属、休闲服务、医药生物、商业贸易、国防军工、综合、建筑材料、交通运输、非银金融、钢铁、电子；股指受投资者情绪轻度影响的行业，包括银行、采掘、食品饮料、家用电器、化工、农林牧渔、汽车；股指不受投资者情绪影响的行业，包括计算机、传媒、机械设备、轻工制造、建筑装饰、公用事业、房地产、通信。我们选取申万一级各行业指数的所有样本股的交易日期为2014年12月31日的营业总收入、总市值、市盈率、市销率、市现率、自由流通市值；2014年的年成交额、年均价、年平均换手率；2014年报的流动比率、速动比率、已获利息倍数、现金比率（数据来源为Wind资讯金融终端），按照上述受投资者情绪影响程度的分的三类归类后取平均值，得到的结果如下表：

|  |  |  |  |
| --- | --- | --- | --- |
| 行业受投资者情绪影响程度 | 重度 | 轻度 | 不受 |
| 营业总收入(TTM) （万元） | 658179.3281 | 3203471.6258 | 402126.4207 |
| 年成交额 （万元） | 3809946.1705 | 5194507.9731 | 2759843.6130 |
| 年均价 （元） | 13.99357493 | 12.79272087 | 15.8301434 |
| 年平均换手率 （%） | 2.55055586 | 2.417176597 | 3.123560402 |
| 总市值 （万元） | 1843049.1186 | 8902605.8673 | 1148704.5143 |
| 市盈率PE(TTM) （倍） | 855.4200893 | 24.76601249 | 23.75366147 |
| 市销率PS(TTM) | 16.33839599 | 7.425893121 | 21.78212509 |
| 市现率PCF(经营现金流TTM) | -27.0878712 | 31.30260784 | 39.38104909 |
| 自由流通市值 （万元） | 761180.6219 | 1851427.1396 | 461325.0663 |
| 流动比率 | 2.257185726 | 2.390845686 | 2.470903859 |
| 速动比率 | 1.793371288 | 1.865829087 | 1.894920808 |
| 已获利息倍数(EBIT/利息费用) | 11.10045087 | 20.19785455 | 52.55652171 |
| 现金比率 | 1.038255244 | 1.173123879 | 1.050498322 |

由上图可以看出，受投资者情绪重度影响的行业，市盈率PE（TTM）为855.4201倍，分别为受投资者情绪轻度影响的行业的34.54倍和不受投资者情绪影响的行业的36.01倍；股价取决于市场需求，即变相取决于投资者对企业的表现和未来发展前景、推出的产品或服务以及该行业的前景的期望，其余影响股价的因素还包括市场气氛、[新兴行业](http://baike.baidu.com/view/1857483.htm)热潮等。市盈率把股价和利润联系起来，但利润没有变化，甚至下降，则市盈率将会上升。这反映出在受投资者情绪重度影响的行业里，投机性[泡沫](http://baike.baidu.com/subview/73683/5119258.htm)严重。另外，其市现率为-27.0879，为三者之中唯一的负值；市现率是股票价格与每股现金流量的比率，净现金流包括投资现金流、融资现金流和经营现金流三个部分，一般来说，经营现金流为正值（表明企业是营利的），但投资现金流和融资现金流的正负用于判断企业的运营状况比较复杂。市现率为负，说明企业加大投资，用现金购买资产导致投资现金流为负或者偿还了债务导致融资现金流为负最后导致三者加总以后的净现金流为负。由此说明，受投资者情绪重度影响的行业，投资者的情绪的确对于股价具有非常大的决定性作用，投资者情绪高涨，导致股价被推高，存在大量的投资性泡沫，而且这类行业中的企业往往会加大投资或者偿还债务，投资者应当警惕。

受投资者情绪轻度影响的行业，营业总收入（TTM）为3203471.6258万元，分别为受投资者情绪重度影响的行业的4.87倍和不受投资者情绪影响的行业的7.97倍；营业收入是指企业在从事销售商品，提供劳务和让渡资产使用权等日常经营业务过程中所形成的经济利益的总流入，说明这类行业往往在实体经济中盈利状况良好。其年成交额为5194507.9731万元，分别为受投资者情绪重度影响的行业的1.36倍和不受投资者情绪影响的行业的1.88倍，说明该类行业交易比另外两类行业更加频繁。其总市值为8902605.8673万元，分别为受投资者情绪重度影响的行业的4.83倍和不受投资者情绪影响的行业的7.75倍；自由流通市值为1851427.1396万元，分别为受投资者情绪重度影响的行业的2.43倍和不受投资者情绪影响的行业的4.01倍；说明该类行业的股价波动对大盘的影响往往更为显著。由此说明，受投资者情绪轻度影响的行业，股票往往被具有各类投资者情绪的投资者频繁交易，而且其往往是具有良好经营状况的企业且股票在大盘中占有重要地位。

不受投资者情绪影响的行业，市销率PS（TTM）为21.78213，分别为受投资者情绪重度影响的行业的1.33倍和受投资者情绪轻度影响的行业的2.93倍，市销率为股票总市值除以主营业务收入或者股价除以每股销售额，没有销售就没有收入，市销率越大，说明企业的投资价值越小。另外，其已获利息倍数为52.55652，分别为受投资者情绪重度影响的行业的4.73倍和受投资者情绪轻度影响的行业的2.60倍；指上市公司[息税前利润](http://baike.baidu.com/view/694608.htm)相对于所需支付债务利息的倍数，可用来分析公司在一定盈利水平下支付债务利息的能力，数值越高，企业长期偿债能力越强。由此说明，不受投资者情绪影响的行业里的企业，往往是具有较强的长期偿债能力，经营状况良好，但是没有太大投资价值的企业。

1. **结论**

本文利用Kun Guo, Yi Sun and Xin Qian（2016）提出的搜集投资者情绪的方法收集投资者情绪数据，创新性地通过对比申万一级行业指数加入投资者情绪和不加入投资者情绪训练随机森林对于股价最终的预测偏差来对比不同行业投资者情绪对股价预测的影响程度。结果发现，随机森林算法对于股指金融序列的预测有良好的表现，并且进一步发现其优于SVR和ANN。对于大部分的股价预测，引入投资者情绪作为指标来考虑是非常有必要性的，对于确定人们对于某支股票的估值范围帮助最大，由此也可以说明，尽管某支股票可能会被热烈讨论，甚至投资者情绪高涨，但是只要股指在投资者期望的价格范围内，它的涨跌方向变化就不会非常受投资者情绪的影响。由此也可以说明，羊群效应在中国股市的投资者中已经开始减弱，人们开始更关注股票的业绩，而且对于股票的买卖逐渐回归对股票业绩的考察而非概念的炒作。投资者的评论针对日常生活所频繁接触的行业最能充分地反映出投资者内心对该行业股票的内心估值。而对于其它行业，虽然投资者情绪也能在一定程度上提高预测的效果，但是仅仅关注网络评论并不能真实地反映股民内心对该行业股票的估值。而对于普通大众日常生活息息相关的行业，网络评论很大程度上可以反映出投资者未来的投资策略，因此引入投资者情绪作为这些行业股指预测的特征序列是最有利于股指预测的。由此说明，中国投资者很大程度上仍然是理性的，舆论与实际操作往往是分开的，对于投资者了解得越全面的行业，投资者的舆论与他们的实际操作才会越密切。

另外，本文还将上述的申万一级行业指数按照受投资者情绪影响程度分成三类，分别是股指受投资者情绪深度影响的行业、股指受投资者情绪轻度影响的行业和股指不受投资者情绪影响的行业。对各类行业的金融指标进行了分析发现，受投资者情绪重度影响的行业，投资者的情绪的确对于股价具有非常大的决定性作用，投资者情绪高涨，导致股价被推高，存在大量的投资性泡沫，而且这类行业中的企业往往会加大投资或者偿还债务，投资者应当警惕；受投资者情绪轻度影响的行业，股票往往被具有各类投资者情绪的投资者频繁交易，而且其往往是具有良好经营状况的企业且股票在大盘中占有重要地位；不受投资者情绪影响的行业里的企业，往往是具有较强的长期偿债能力，经营状况良好，但是没有太大投资价值的企业。

投资者情绪是个难以度量的概念，它反映了[市场](http://wiki.mbalib.com/wiki/%E5%B8%82%E5%9C%BA)参与者的[投资意愿](http://wiki.mbalib.com/wiki/%E6%8A%95%E8%B5%84%E6%84%8F%E6%84%BF)或者预期。投资者能感觉到它的客观存在，但是要问它到底有多高、近期发生了何种变化，每个[个体](http://wiki.mbalib.com/wiki/%E4%B8%AA%E4%BD%93)的投资者有因为有[持仓](http://wiki.mbalib.com/wiki/%E6%8C%81%E4%BB%93)、风格、[财富](http://wiki.mbalib.com/wiki/%E8%B4%A2%E5%AF%8C)、地位等因素的不同，给出不同的答案。不过所有人都难以否认的是，投资者情绪有是个非常重要的概念，在[经济活动](http://wiki.mbalib.com/wiki/%E7%BB%8F%E6%B5%8E%E6%B4%BB%E5%8A%A8)中，情绪是个不确定因素，它影响到[投资者](http://wiki.mbalib.com/wiki/%E6%8A%95%E8%B5%84%E8%80%85)对未来收益的主观判断，进而影响到其[投资行为](http://wiki.mbalib.com/wiki/%E6%8A%95%E8%B5%84%E8%A1%8C%E4%B8%BA)，形成合力后，对市场会形成很大的影响。投资者情绪对未来市场波动的影响逻辑在于对正面消息、负面消息的逐级正反馈放大。但是过去国内外的对于投资者情绪对于市场波动大部分都是针对市场整体，希望本文对于不同行业的投资者情绪对股价预测的影响的研究结果可以对行为金融研究者和投资者在未来考虑投资者情绪和股价关联时起到分行业角度考虑的启发，并且对于未来股市政策制定在考虑投资者情绪时可以起到分行业考虑的建议。

**参考文献**

1. Charles li SVR 1.0工具箱下载地址：http://bbs.pinggu.org/forum.php?mod=viewthread&tid=2398764&pid=39243335&page=1&extra=#pid39243335 [↑](#footnote-ref-1)