计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900168

基于机器学习的股市拐点影响因素研究

袁钰坤[1]　李　刚[1]　赵治翔[1]　徐　力[2]
1 中证数据有限责任公司　北京 100032
2 中国科学院计算技术研究所中国科学院网络数据科学与技术重点实验室　北京 100190
(yuanyk@cmsmc.cn)

摘　要　股票市场的成交情况可以充分反映投资者的行为特征并影响整个股市的走势。股票成交明细数据作为股市最底层的交易数据,能够全面地体现股票交易的情况,成为至关重要的股票市场走势判断的参考数据,能够为资本市场监管者在风险监测领域进行决策提供有效帮助。文中提出了一种可以快速地在海量股票交易明细数据中提取投资者交易特征的方法,然后基于逻辑回归、决策树和随机森林等机器学习算法找到股市大盘较大拐点产生的主要影响因素,并预测交易特征变量对股市较大拐点产生的时间范围。在沪深股指上进行的实验表明,相较于传统的模型,文中提出的方法可以将股市较大拐点预测的准确度提高约10%,并在6个月的回测实验中准确率依旧保持在70%左右的水准,从而证明了模型的有效性。

关键词:股票市场;走势判断;风险监测;股市拐点;机器学习
中图法分类号　TP391

Research on Factors Affecting Stock Inflection Point Based on Machine Learning Algorithms

YUAN Yu-kun[1],LI Gang[1],ZHAO Zhi-xiang[1] and XU Li[2]
1 China Securities Data CO.,LTD,Beijing 100032,China
2 Key Laboratory of Network Data Science & Technology,Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190,China

Abstract　Transaction situation in stock market can fully reflect behavior characteristics ofinvestors and affect the trend of entire stock market. As the bottom-level transaction data of stock market,detailed data of stock transaction can comprehensively reflect the situation of stock transactions and become a vital referencefor judgment of stock market trends. It can also provide regulators in capital market with effective information when making decisions in the field of risk monitoring. In this paper,we propose a method that can quickly extract the characteristics of investor transaction from detailed data of stock transaction,based on machine learning algorithms such as logistic regression,decision tree,and random forest,finding the main influencing factors of large inflection points and predictingtime range over which the larger inflection point occurs. The experimental results on the stock indexes of Shanghai and Shenzhen show that the proposed method can highly improve accuracy of prediction of large inflection point instock market by appoximately 10%,compared with a traditional model,and the accuracy rate in six-month backtesting experiment maintains a level of 70%,which demonstrates validity of the model in this paper.

Keywords　Stock market,Trend judgement,Risk monitoring,Stock inflection point,Machine learning

1　引言

股票投资者作为资本市场中的重要参与者,其交易行为的改变会快速地反映到市场行情上。随着我国A股市场的稳步发展,开户的投资者数量逐渐增加,截至2020年8月,我国A股投资者开户数量达到1.7亿,在部分交易时间段,A股单日成交额以千亿计,庞大的交易量使得股市的成交明细数据成为海量的金融大数据。这些数据中蕴含着丰富的交易信息,合理地利用这些数据可以给股票市场的监管工作提供不小的帮助和支持。

为了找到海量证券交易数据中有价值的信息,采用机器学习算法,按照一定方式学习大数据中隐含的内在规律,从而得到知识和经验。机器学习算法主要是指通过数学及统计方法求解最优化问题的步骤和过程。针对不同的数据和不同模型需求,选择和使用适当的机器学习算法可以更高效地解决一些资本市场监管中的问题。本文基于中国A股市场的底层交易数据构建了用于分析股市拐点预测的模型,以逻辑回归、决策树和随机森林机器学习模型为基础建模,通过历史回测验证模型的有效性。

本文的主要贡献如下:1)通过A股市场全量交易数据来训练模型,减少数据偏差性,并且基于的是实际监管场景,能够助力证券市场科技监管;2)利用多种可解释性较强的机器学习模型,便于找到对股市出现拐点影响较大的因素,有助于监管者追溯风险源头;3)提出一种创新的特征工程方法,有效

基金项目:国家自然科学基金(91746301,61902380);北京市科技新星计划(Z201100006820061)
This work was supported by the National Natural Science Foundation of China (91746301, 61902380) and Beijing Nova Program (Z201100006820061).
通信作者:李刚(ligang@cmsmc.cn)

地找到与股市走势拐点相关性较高的模型特征；4）实验结果验证了所构建的模型能在股市拐点的影响因素分析上有较为显著的效果。

2 研究背景和需求

在20世纪90年代，不少国家的金融监管部门逐渐开始尝试使用科技手段支持监管，但直到2008年国际金融危机爆发之后，科技监管才得到真正发展，这背后的推动力主要在以下几方面。

2.1 快速的科技发展

最近十多年以来，数据的存储成本快速下降，计算机的计算能力得到大幅提升，有效推动了机器学习、自然语言处理、强化学习等技术的兴起。与此同时，随着大数据、云计算、区块链等创新技术的快速发展，这些高新技术不仅可用于商业，也可支持资本市场的监管活动。

2.2 严格的监管要求

自从2008年国际金融危机发生后，许多国家的金融监管部门大大加强了对金融机构的监管发展并出台了相应的指引，如美国的《多德－弗兰克法案》[1]、欧洲的《欧盟金融工具市场指导（MiFID）》[2]，更严格的监管需要更先进的监管技术提高监管效率，降低监管成本。

在2018年8月底，证监会正式印发《中国证监督会监管科技总体建设方案》（以下简称《总体建设方案》）[3]，完成了监管科技建设工作的顶层设计，并进入了全面实施阶段。《总体建设方案》详细分析了证监会监管信息化现状、存在的问题以及面临挑战，提出了监管科技建设的意义、原则和目标，明确了监管科技工作需求和工作内容。在2019年发布的证监会深化资本市场改革12条举措中，加快提升科技监管能力位列其中。

2.3 严峻的监管挑战

随着金融科技的迅速发展和应用，监管的方法与金融技术之间的差距越来越大。例如，2010年美国股市发生闪崩事件，股指在几分钟内下跌了700点，美国证监会花了将近半年时间才发现其中根源，这暴露出了目前监管科技的巨大缺陷。面对科技在金融行业的广泛应用，金融监管部门的能力建设亦需要相应跟上。将传统监管方式结合前沿的大数据技术、人工智能等，与当前的金融科技能力相匹配，才能更有效地应对监管挑战。

3 文献综述

机器学习在股市的行情预测中得到应用之前已经发展了几十年，而机器学习本身最早可以追溯到对神经网络的研究。早在19世纪40年代，神经网络的模型被提出[4]，神经网络算法的理论便得到确立，为机器学习的未来奠定了发展基础。1959年，机器学习（Machine Learning）这个词在文献[5]中被首次提出。

随着机器学习算法的发展，近几年许多学术研究人员开始将其应用于股市的走势分析和预测上。Li等[6]基于经验，借助极限学习机来预测股票走势并展示了该方法的有效性；Ebadati等[7]在那之后尝试采用遗传算法和人工神经网络的方式研究了股票类时间序列数据的走势预测，模型的残差平方和达到99%；同年，Xie等[8]基于神经网络集成学习的方法对股票走势进行预测并得到了一定成效；一年后，一种基于深度自编码器的方法被Gao[9]用来对股市走势进行预测；当年，Wang[10]以香港恒生指数为例，采用机器学习模型对股市的大盘指数走势进行预测。以上几项研究虽然均取得了一定成效，但是站在国家监管层面来看，基于深度学习的算法都具有较少的可解释性，难以从监管业务的角度来追根溯源。

为了增加模型的可解释性，决策树[11]和逻辑回归[12]等传统可解释性强的算法也被用来研究股票的走势；2020年，López-Cabarcos等[13]基于GARCH和EGARCH模型分析了投资者情绪和股市行情的关系；之后，Li[14]继续从股市情绪分析的角度，基于机器学习模型实现了以香港为例的股票走势预测，实验结果表示模型的效果在验证集和测试集上的效果均比之前的基准模型更好。在那之后，Zhao等[15]基于金融文本情感分析对股票走势进行预测，模型效果在预测中国沪深股指上比传统方法的准确率高出约8%。以上的研究虽在可解释性上有了进一步提高，但实验都是基于公开的股市行情数据，且股票走势的预测准确性已经逐渐趋近于峰值。

传统研究所使用实验数据仅仅停留在公开的股票数据层面，没能利用到股票市场中交易明细数据的价值。本文实验所用的数据为证券交易明细的数据，且采用可解释性较强的机器学习算法，以Zhao等在2020年发表的模型[15]作为本文的基准模型。基准模型的实验研究是基于中国A股的沪深股指，实验效果属于当前较为前沿的模型，准确率达到68%。本文所构建的模型将以此进行比较。利用股市交易明细数据，来构建对股票市场拐点的预测模型，并找到对股市拐点影响的因素，对于支持证券市场监管具有重要意义。本文的实验流程如图1所示。

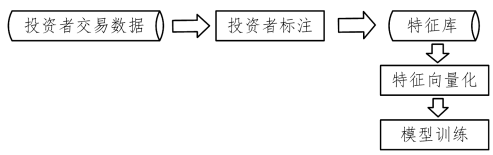


图1 实验流程

Fig. 1 Process of experiment

4 模型构建

4.1 实验目的

本文从资本市场监管领域选择风险监测为研究方向，基于证券市场的交易数据，探索运用机器学习模型，影响股市大盘产生重大拐点的因素，并基于影响因素构建股市大盘重大拐点预警提示模型，以助力股票市场风险的科技监督管业务。

4.2 实验数据和环境

本次研究的数据采用的是证券市场投资者的交易明细数据，使用数据的时间范围是2012年初至2019年中，数据内容主要是投资者在指定时段内的全部交易数据，包括交易时间、交易股票、交易金额、交易方向、交易笔数等明细信息，分析的投资者对象包括了散户投资者、国内机构投资者和外资等在A股市场有进行过交易记录的投资群体。

4.3 特征工程

本次研究中的特征工程采用先假设后验证的方法，首先假设影响股市大盘重大拐点的特征因素，然后基于专家经验，利用基于规则的方法构建模型，并验证假设的有效性，从而得到对股市大盘重大拐点影响较大的因素，并将其作为后续机器

学习训练的数据特征。具体的模型特征构建流程如图 2 所示。

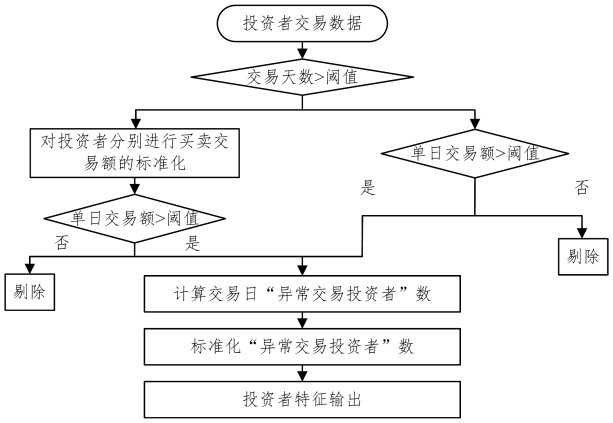


图 2 特征工程流程

Fig. 2 Flowchart of feature engineering

4.3.1 特征有效性假设

本次研究中假设在每个交易日中，当那些较往常交易行为有大波动的投资者（以下简称大波动投资者）数量出现较大变化的时候，股市大盘很有可能出现重大的拐点。

基于以上假设，模型假设具备以下两种交易行为特征的投资者为“大波动投资者”：1）日度交易频率低，且在当日出现较大额度的单向交易行为；2）日度交易频率高，且在当日出现单向交易额度剧增的交易行为。

与往常交易日情况相比，当这两类“大波动投资者”的总体数量在当日出现一个较大变化的时候，模型则假设当日附近几天会出现股市大盘的重大拐点。

4.3.2 特征有效性检验

针对日度交易频率低且在当日出现较大额度的单向交易行为的“大波动投资者”，借鉴中心极限定理的思想，模型将单位时间段内总交易天数小于 30 天的投资者筛选出来，在此基础上再筛选出当日净卖出或者净买入在设定的阈值范围内的投资者作为日度交易频率低且在当日出现较大额度的单向交易行为的“大波动投资者”。针对日度交易频率高且在当日出现单向交易额度剧增的交易行为的“大波动投资者”，模型筛选出一年内总交易天数大于 30 且在当天净买入或净卖出超过设定阈值的投资者，然后对每一位投资者的每日卖出和买入分别进行标准化。以每日卖出的标准化为例，定义 $V^i_{交易额}$ 为第 i 个交易日的当日交易额，前$(i-j+1)$个交易日中交易额标准化处理后的结果为 $STD^i_{交易额}$，每个交易日的交易额标准化公式如下所示：

$$\mu^i_{交易额}=\frac{1}{i-j}\sum_{k=j+1}^{i}V^k_{交易额}$$

$$\sigma^i_{交易额}=\sqrt{\frac{1}{i-j}\sum_{k=j+1}^{i}(V^k_{交易额}-\mu^i_{交易额})^2} \tag{1}$$

$$STD^i_{交易额}=\frac{V^i_{交易额}-\mu^i_{交易额}}{\sigma^i_{交易额}}$$

本文借鉴中心极限定理，默认单位时间段内的交易额近似服从正态分布，则当标准化后的数值大于 3 时，有较大的概率认为当日交易额属于异常波动数值。最终，模型将以上两个角度筛选出来的投资者取并集，作为当日总的大波动投资者人数。

基于模型的假设，还需要找出“大波动投资者”数量出现较大变化的时间点。假设定义当日“大波动投资者”数为 $N_{波动}$，前几个交易日的平均“大波动投资者”数为 $\mu_{波动}$，前几个交易日“大波动投资者”数的标准差为 $\sigma_{波动}$，则每个交易日大波动人数标准化处理的计算方法如下所示：

$$\mu^n_{波动}=\frac{1}{n-m}\sum_{k=m+1}^{n}N^k_{波动}$$

$$\sigma^n_{波动}=\sqrt{\frac{1}{n-m}\sum_{k=m+1}^{n}(V^k_{波动}-\mu^n_{波动})^2} \tag{2}$$

$$STD^n_{波动}=\frac{N^n_{波动}-\mu^n_{波动}}{\sigma^n_{波动}}$$

通过模型将每日的大波动投资者数基于前几个交易日进行标准化，如表 1 所列。

表 1 交易日波动人数标准化结果示例

Table 1 Example of fluctuate number of traders after standardization on trading day

交易日期	大波动人数	标准化
Day1	Num1	STD1
Day2	Num2	STD2
Day3	Num3	STD3
…	…	…

标准化后的数值被用于衡量当日异常人数与前几日异常人数相比的突变情况，假设突变跨度越大，则标准化的数值越大，当日越有可能是大盘的较大拐点。根据数据标准化后的数值，结合当日的交易量特点来预警大盘的重大拐点时间。将 2012 年至 2019 年全量的证券交易明细数据作为分析数据，以沪深 300 指数作为回测目标，根据上述的标准化值预警方法得到的效果如表 2 所列。

表 2 基于规则的模型预测效果

Table 2 Performance of the rule-based model

（单位：%）

准确率	召回率	*F-score*
70.4	85.7	77.3

通过表 2 所列结果，可以发现基于本文的假设所得到的特征来进行股市拐点预测，准确率能达到 70% 以上，超过了基准模型。因此，在一定程度上可以验证本文最开始假设的有效性：在每个交易日当中，当那些有异于往常交易行为的投资者数量出现较大变化的时候，股市大盘很有可能出现重大的拐点。同时说明每个交易日的“大波动投资者”数量经过标准化后的值与股市出现较大拐点的情况是相关的，因此本文将其筛选出来作为机器学习算法模型的输入特征。

4.4 机器学习模型的构建

考虑到基于规则的模型中按照单日净交易额大于单一阈值来筛选“大波动投资者”结果可能有偏，在机器学习模型中将筛选各个阈值范围的“大波动投资者”并进行标准化，作为机器学习模型训练的多个特征输入值。所有维度的特征可以描述为$\{D_1,D_2,\cdots,D_n\}$，将特征的总维度空间设置为 D，D 由各个交易额阈值筛选出的标准化后的大波动投资者数值 D_i 组成，如下所示：

$$D=\overset{n}{\underset{i=1}{\times}}D_i \tag{3}$$

其中，D 在实验中的机器学习模型训练过程中将作为模型的数据特征值输入。

通过构建交易频率、成交金额、成交量、异常交易时间点等参数，同时调整相关参数的阈值范围，最终得到 10 个相关的特征，并将计算得到的特征数值标准化。

通过特征工程步骤得到包含成交额和成交量信息的高阶特征，将特征值以及是否为大盘拐点的标签输入到机器学习模型中。考虑到模型可解释性，本次研究主要采用了决策树和逻辑回归这两种可解释性较强的机器学习算法模型，最后再采用随机森林算法来提高模型的效果。具体模型的训练流程如图 3 所示。

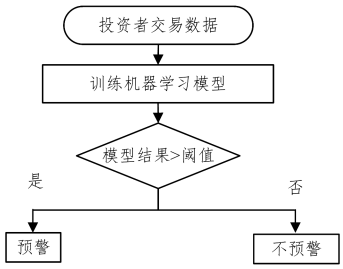


图 3 机器学习模型训练流程

Fig. 3 Training process of the machine learning-based model

5 实验与结果分析

将 2012 年 1 月至 2019 年 6 月的证券市场交易明细数据进行特征工程处理，将数据按照大约 80% 训练集合、20% 验证集进行划分，将特征值输入机器学习算法中。

本次实验回测选择沪深 300 股指为基准，分别采用逻辑回归、决策树和集成学习的方法进行模型训练，在沪深 300 股指上进行历史回测，回测的效果如图 4 所示。

图 4 逻辑回归模型回测效果

Fig. 4 Performance of the model based on logistic regression

图 5 决策树模型回测效果

Fig. 5 Performance of the model based on decision tree

图 6 随机森林模型回测效果

Fig. 6 Performance of the model based on random forest

图 4—图 6 中，沪深 300 指数为图中上方的折线，算法回测结果为图中下方柱状图。左侧主坐标是沪深 300 股指的点位，右侧次坐标轴是机器学习算法的预测情况，预测数值分布在 0 到 1 之间。模型回测效果如表 3 所列。

表 3 机器学习模型回测效果

Table 3 Performance of the machine learning-based models

（单位：%）

	准确率	召回率	*F-score*
逻辑回归	71.2	87.5	78.5
决策树	75.7	90.7	82.5
随机森林	77.8	95.1	85.8

从表 3 中可以看出，3 个机器学习模型的准确率均超过了基准模型的 68%，其中随机森林模型的效果最好，准确率为 77.8%，比基准模型的预测效果高出约 10%。决策树模型和逻辑回归模型表现较弱，但是由于其模型可解释性强，可以通过这两个模型找出股市产生拐点的显著影响因素。通过模型训练结果可以看出，股市产生拐点的影响因素主要是较大波动投资者的交易额和交易量。

为了进一步验证模型的效果，本文基于表现最佳的随机森林算法模型，在 2019 年下半年的交易数据上进行回测，测试的准确率约为 74%，虽然模型准确率有所下降，但是仍然保持在 70% 的水准之上，效果超过基准模型，在一定程度上说明本文所构建的模型是有效的。

结束语 股市的走势预测一直是许多股市研究的学者和专家较为关注的问题，但由于股票交易的明细数据无法获取，很难在传统的研究上有较大的提升空间。针对这一问题，本文采用逻辑回归、决策树和随机森林等机器学习算法，基于真实的证券市场交易明细数据，对股市大盘重大拐点的影响因素进行了分析和研究。

由于中国股市尚处于发展阶段，历史数据相较于成熟的欧美股市还是有一定的差距，即使实验中有区分数据的训练集和测试集，但在有限的数据中构建的模型仍然容易产生过拟合的现象。所以，尽管模型在实验结果上展示出一定效果，但是本模型的实际效果还需要在未来的新数据上进行测试。

参 考 文 献

[1] SONG L Z, HU H B. An Interpretation of the "Dodd-Frank Act" of the United States, with a Discussion on the Reference and Enlightenment to Country's Financial Supervision[J]. Research on Macro Economy, 2011(1): 67-72.

[2] XIAO F. From Efficiency to Transparency-Small Discussion on EU "Financial Instruments Market Directive II" and other regulatory measures [J]. Bond, 2018(7): 87-93.

[3] CSRC. The Securities Regulatory Commission officially released the overall construction plan for the implementation of regulatory technology[J]. Information Technology and Informatization, 2018(9): 10.

[4] MCCULLOCH W, PITTS W. A logical calculus of the ideas immanentin nervous activity[J]. The Bulletin of Mathematical Biophysics, 1943, 5(4): 115- 113.

[5] SAMUEL A L. Some Studies in Machine Learning Using the Game of Checkers[J]. IBM Journal of Research and Development, 1959, 3(3): 210-229.

（下转第 177 页）

SARIMA models forecasting of weather parameters for Thrissur district [J]. International Journal of Statistic and Applied Mathematics, 2018; 3(1): 360-367.

[16] LEI L. Wavelet Neural Network Prediction Method of Stock Price Trend Based on Rough Set Attribute Reduction [J]. Applied Soft Computing, 2018, 62: 923-932.

[17] GÜRSOY O, ENGIN S N. A wavelet neural network approach to predict daily river discharge using meteorological data [J]. Measurement and Control, 2019, 52(5/6): 599-607.

[18] JEWSON S, CABALLERO R. The use of weather forecasts in the pricing of weather derivatives [J]. Meteorological Applications, 2003, 10: 377-389.

ZHANG Xue, born in 1989, Ph. D, lecturer. Her main interests include risk management and so on.

LUO Zhi-hong, born in 1976, Ph.D, professor. Her main interests include risk management and so on.

（上接第 168 页）

[6] LI X, XIE H, WANG R, et al. Empirical analysis: stock market prediction via extreme learning machine [J]. Neural Computing and Applications, 2016, 27(1): 67-78.

[7] EBADATI O M E, MORTAZAVI M T. An Efficient Hybrid Machine Learning Method For Time Series Stock Market Forecasting [J]. Neural Network World, 2018, 28(1).

[8] XIE Q. Research Based on Stock Predicting Model of Neural Networks Ensemble Learning [C] // Shanghai University Of Engineering Science. Proceedings of 2018 2nd International Conference on Electronic Information Technology and Computer Engineering (EITCE 2018). 2018.

[9] GAO Z. Research on Financial Data Prediction Based on Deep Autoencoder [C] // International Informatization and Engineering Associations. Proceedings of 2019 2nd International Conference on Financial Management, Education and Social Science (FMESS 2019). International Informatization and Engineering Associations: Computer Science and Electronic Technology International Society, 2019: 394-400.

[10] WANG W. Prediction of Hang Seng Index Based on Machine Learning [C] // Institute of Management Science and Industrial Engineering. Proceedings of 2019 3rd International Conference on Artificial intelligence, Systems, and Computing Technology (AISCT 2019). Institute of Management Science and Industrial Engineering: Computer Science and Electronic Technology International Society, 2019: 252-256.

[11] DENG S, WANG C, WANG M, et al. A gradient boosting decision tree approach for insider trading identification: An empirical model evaluation of China stock market [J]. Applied Soft Computing Journal, 2019, 83.

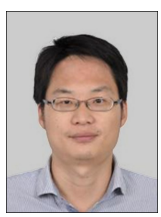
[12] ZHOU F, ZHANG Q, SORNETTE D, et al. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices [J]. Applied Soft Computing Journal, 2019, 84.

[13] LÓPEZ-CABARCOS M A, PÉREZ-PICO A M, PIÑEIRO-CHOUSA J, et al. Bitcoin volatility, stock market and investor sentiment. Are they connected? [J]. Finance Research Letters, 2019.

[14] LI X, WU P, WANG W. Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong [J]. Information Processing and Management, 2020.

[15] ZHAO C, YAO W. Stock Volatility Forecast Based on Financial Text Emotion [J]. Computer Science, 2020, 47(5): 79-83.

YUAN YU-kun, born in 1994, postgraduate. His main research interests include machine learning and natural language processing.



LI Gang, born in 1980, postgraduate. His main research interests include compute science and quantitative finance.