文章编号：1003-0077（2017）00-0000-00

证券市场股东画像的构建

于花蕾　饶 元　汤才芳　任浩然
（西安交通大学 软件学院 社会智能与复杂数据处理实验室，陕西 西安 710049）

摘要：股东画像构建旨在快速了解股东市场行为背后的真实偏好特征，最终指导外部投资者的投资决策。由于十大流通股东频繁的市场行为常会引发股票价格的异常波动，且他们总能近乎精确地把握交易时机，因此股东画像的构建是一项有意义的工作。该文分析了十大流通股东投资行为，从活跃度、偏好特征两个方面构建股东画像，将股东按性质划分为个人、机构、基金，并分别对 3 类股东从全集角度构建画像。此外，还提出了标记股东特征的一些方法，并讨论了处理股东特征时可能遇到问题及应对措施。

关键词：十大流通股东；股东画像；活跃度；偏好特征

中图分类号：TP391　　**文献标识码：**A

Shareholder Profiling in Stock Market

YU Hualei, RAO Yuan, TANG Caifang, REN Haoran
(Lab of Social Intelligence&Complex Data Processing, School of Software,
Xi’an Jiaotong University, Xi’an, Shaanxi 710049, China)

Abstract：The purpose of shareholder profile’s construction is to quickly understand the real preference characteristics behind the shareholders' market behaviors and finally guide the investment decisions of external investors. The construction of shareholder portrait is a meaningful work because of the abnormal fluctuation of stock price caused by the frequent market behavior of the top ten circulating shareholders, which always grasps the opportunity of trading almost accurately. This paper analyzes the investment behavior of the ten circulating shareholders, constructs the shareholder portrait from two aspects of activity degree, preference characteristic. Moreover, the types of shareholder are divided as individual, organization, fund, based on the Nature of shareholder. And we construct the portrait from the complete set angle respectively to 3 kinds of shareholder. In addition, some methods of labeling shareholder characteristics are put forward, and some problems and countermeasures are discussed when dealing with shareholder's characteristics.

Key words: top ten circulation shareholders; shareholder profile; active; preference characteristic

0　引言

随着大数据时代的来临，股票市场已经产生了大量关于股东投资行为的数据信息，并且这些信息公布在上市公司的年报、季报当中，不涉及隐私安全等问题。如何充分利用这些投资行为信息来进行分析，最终指导外围投资者投资决策成为了一个亟待解决的问题。重要股东的市场行为常是各界的重点关注对象，这是因为外围投资者常会跟随重要股东的市场行为做出投资决策。重要股东的市场行为不仅会对股票价格造成直接影响，还与公司的成长性相关联，故其重要性不言而喻。然而，当前针对重要股东的市场行为研究普遍存在以下几点问题：（1）针对性单一，缺乏全局。通常关于股东的市场行为研究都是针对具有某一类市场行为的全体股东展开的，而不是针对市场上全体股东的所有市场行为进行研究。（2）不够精准。当下的粗放式的重要股东持股研究不能为外部投资者、散户等提供更为精准明确的信息，外部投资者不能了解到上市公司大股东的性质（基金、QFII、公司机构、个人股

收稿收稿日期：2017-03-16；**定稿日期：**2017-04-26　六号
基金项目：基金名（基金号）；基金名(基金号)
六号，核实准确完整的基金名称

东等）及重要股东会选择在哪个确切的时刻对哪些股票进行投资。（3）不能对外部投资者提供指导。由于无法详细的了解到感兴趣的重要股东会对哪类股票更具偏好，这些偏好的股票存在哪些共性，应该在何时跟随他们投资等重要信息，最终导致无法对外部投资者提供指导。

针对现有研究存在的以上 3 点不足，本文从一个全新的角度出发，提出一个新的概念—“股东画像”。股东画像的概念可类比于用户画像。相比于用户画像，股东画像的研究对象有所不同，从普通的产品用户变为股东，通过对股东进行标签化，最终生成一个虚拟的可用于表示股东的标签体系。构建股东画像是一种富有成效的股东描述及建模方法。其过程是不具有针对性的，即对市场中的全体股东进行画像；且通过已有数据，可以精确的了解到股东的投资行为及各类信息，最终通过这些信息对外部投资者进行投资指导。

本文的贡献在于提出了一个全新的概念“股东画像”及标记股东特征的一些方法，并讨论了处理股东特征时可能遇到的问题及应对措施，提出了“股东画像”的建模模型。通过对现有数据信息进行各角度的特征构建，最终刻画出当前 A 股上市公司十大流通股东的画像，达到为股票市场中占绝大数量的散户的投资决策提供帮助的目的。

1 相关工作

用户画像(User Profile)，即用户信息的标签化，是建立在一系列数据之上的目标用户模型[1]。意义在于帮助企业猜测用户的真实需求和潜在需求，精细化地定位人群特征等。目前用户画像被运用在许多领域[2]，李恒超[3]等人基于用户的查询词历史记录，提出了一种用于构建用户画像的二级融合算法框架，用于对用户性别、用户年龄及教育程度进行预测。Mueller[4]等针对 Twitter 用户的用户名信息构建了多种词语结构特征，并通过识别用户性别构建了 Twitter 用户画像。黄文彬[5]等在分析移动用户的基站轨迹的基础上，采用频繁模式挖掘、构建概率矩阵等方法，构建了包含地理位置信息的用户画像模型。张慷[6]等通过提取用户的上网行为特征，同时结合相关数据进行数据融合及交叉分析，构建了通信用户的画像模型。费鹏[7]等人基于电网用户进行了分析研究，提出了一种用于构建用户画像的多视角融合框架，该框架能快速、准确地识别电费敏感客户。

重要股东指能对公司的经营运行产生重大影响的股东，他们的存在是公司得以正常经营和运行的关键。本文研究的是重要股东中一类关键群体，即上市公司的十大流通股东。通过了解该类群体的行为偏好、投资偏好，可以帮助外部投资者、散户在投资决策时起到参考作用。倪光耀[8]对于大股东增持的背后考量和公司在被增持后对公司的治理产生的影响进行研究进行了研究，表明股票增持对市场会产生两方面的影响，且为企业对外扩张提供资金的强有力保证，形成宝贵的无形资产。赵红岩[9] 就大股东增减持行为对上市公司股票价格的影响机理进行了分析，并就有关各方如何应对给出相应建议。刘建忠[10]对重要股东持续减持下的公司绩效和成长性进行研究，结果表明重要股东的持续减持行为会对公司的绩效与成长性造成影响。

虽然用户画像和股东行为研究已经成为当前研究的热门话题，但是目前将两者结合考虑，针对股东进行画像的研究还很缺乏。本文的主要工作是从数据集中提取股东特征，并对股东画像进行建模研究。构建股东画像预计有以下用途：其一是了解各个股东基本信息及偏好信息；其二是对外部投资者给予一定程度的指导作用；其三是针对由股东、股票、上市公司生成的三模异构社会网络拓扑图进行链接预测、分类以及演变分析。

2 股东画像的构建步骤

根据网络爬虫爬取到的网易财经、同花顺及金融界等门户网站上股东、股票相关的数据类型及业务需求，可以将股东画像工作进行细化，如图 1 所示。

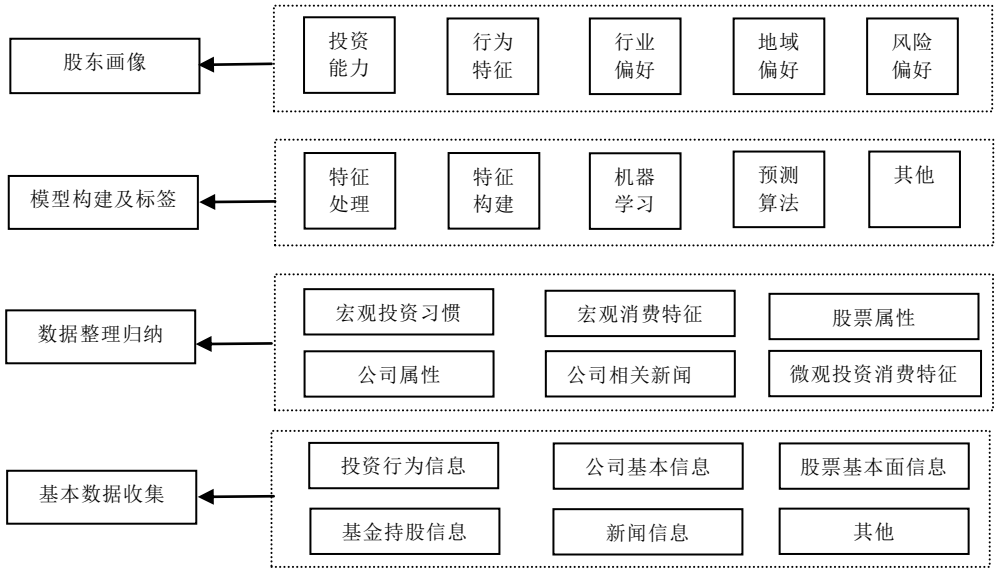


图 1 股东画像构建过程

2.1 基本数据收集

本文数据来自于抓取的网易财经、金融界、同花顺等门户网站及国泰安(CSMAR)数据库，时间为 2004 年 03 月至 2017 年 12 月。本研究重点采用十大流通股东行为数据、上市公司基本数据和 A 股基本面变化 3 类数据，基金持股数据用于数据。预处理阶段，主要是清洗数据，使股东行为表中的数据一致。本研究的数据预处理流程如图 2 所示。

数据集： Π 表示基金持股信息数据表；字段包含 id(唯一标示),jiJinName(基金名称),jiJinNum(基金代号),SumOfStock(持股数量), date(截止日期)等。

Γ 表示原始十大流通股东行为数据表；字段包含 id(唯一标示),stockHolderName(股东名称),nowhaving(现持有数量),date(截止日期),jiJinNum(基金代号)等。

输出： 预处理后的 Π 、Γ 数据

前期处理： 从 Π 中取出所有(jiJinName, jiJinNum)对儿 pair 存入 π 列表中，并对 Γ 中的数据进行数据清洗，去除脏数据。

Step 1: foreach pair p_i **in** π **do**

| **for** length in range (2 , p_i 中的 jiJinName.length) **do**

| | **1.1** 将 jiJinName 截断，取长度为 length 的前几个字符

| | **1.2** 将 1.1 的结果作为条件，在 Γ 中模糊查询，匹配相应的股东名称字

| | 段，并统计总数 sum。若 $sum!=1$, continue；

| | **1.3** 若 $sum=1$，则将对应的 jiJinNum 存入到 Γ 中对应的字段，break；

| **end**

end

Step 2: 把处理后的 Γ 存入 mysql 数据库

图 2 数据预处理算法

2.2 数据整理归纳

收集到海量的十大流通股东及股票的基础数据后，需要对这些数据进行整理归纳，从中获取有用的信息。这一过程需要从真实需求出发，将基础数据划分为宏观统计习惯、宏观消费特征、股票属性、公司属性、公司相关新闻、微观消费投资特征等信息，这些信息结合相关场景将产生巨大的价值。

2.3 模型构建及标签

使用特征工程、机器学习等技术建立行为模型，对股东进行标签化。通过分析股东的宏观投资习惯、宏观消费特征、购买股票的相关属性、

股票对应公司的相关属性、公司相关新闻及微观投资消费特征等信息，可以给股东赋予活跃度标签、行业偏好标签、地域偏好标签等。

2.4 股东画像

针对本文提出的“股东画像”，类比于用户画像，给出如下定义：股东画像是在股东基本信息未知的前提下，根据股东市场投资行为、股票基本面信息、所投资股票所属的上市公司基本信息等抽象出的一个标签化的股东模型。定义的元模型可以形式化表示为一个四元组：

$$< O_j \mid ACT_j, FUND_{j\&k}, Company_{k\&i} > \quad (1)$$

其中，O_j 表示标签化的股东模型；ACT_j 表示股东 j 的市场投资行为；$FUND_{j\&k}$ 表示股东 j 所投资的股票 k_i 的基本面信息集合；$Company_{k\&i}$ 表示股票 k_i 对应的上市公司 i 的基本信息集合。

本文提出的构建股东画像的核心工作是应用上述 4 类信息对股东的活跃度特征、行业偏好、地域偏好等进行标签和确认，从而生成股东画像。

3 证券市场股东画像的构建

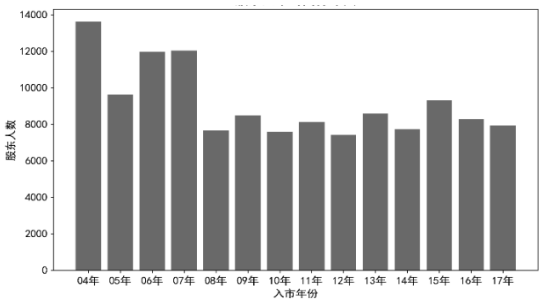


图 3（a）股东进入市场时间与股东数之间的关系

图 4 对假设二做出说明，在数据集的时间间隔内，股东活动区间存在明显区别。活跃区间在一年之内的十大流通股东数占绝大多数，在数据集的时间间隔内，共 65332 名股东，这也说明十大流通股东行列的流通性强。故可以对股东“年龄”赋予适当的权重。

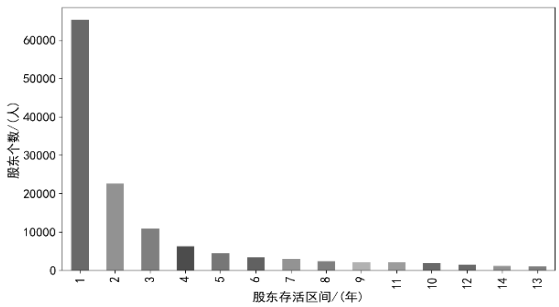


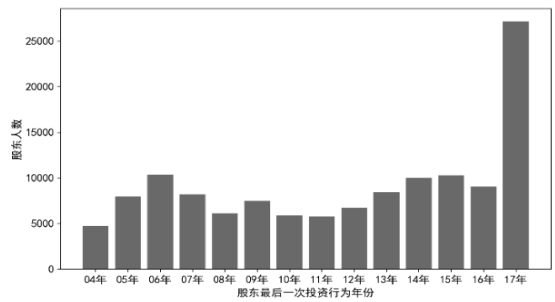
图 4 股东活跃区间长度与股东数量之间的关系

3.1 股东活跃度

股东活跃度可以看作股东在股票市场的交互行为频繁程度及进入十大流通股东的评估。外部投资者迫切想要了解股东的该类信息，从而决定是否跟投。本文定义以下公式用于刻画股东活跃度，即为：

{股东活跃度}=（股东“年龄”,宏观投资行为） （2）

1）股东“年龄”。把股东“年龄”作为指标需要考虑：一是在数据集的时间间隔的不同时间阶段里，加入十大流通股东行列的股东构成是否存在明显区别。二是在数据集的时间间隔内，股东存活区间是否存在明显区别。如果上述假设为真，则有必要给不同时间入市的股东赋予不同的权重。图 3（a）和图 3（b）对假设一做出说明，除数据集本身原因造成 2004 年进入市场的股东数显示较多，2017 年为最后投资年份的股东数显示较多外，（a）图说明在数据集时间间期内不断有股东退出和进入十大股东行列中，2006、2007 年股东进入十大股东排行的数量多。（b）图显示 2006、2014、2015 年进行最后一次投资的股东人数也偏多。故不同时间间期内，加入或退出的股东存在明显区别。



（b）股东最后一次投资的时间和股东数之间的关系

这里采用 sigmoid 函数处理股东活动区间得到股东“年龄”指标。设股东进入市场的时间为 t_1，最后一次投资对应的时间为 t_2，$\bar{S}$ 为股东平均投资时间，则有：

$$S_{yearOfHolder} = \frac{1}{1 + \frac{1}{t_2 - t_1 + 1} e^{-(\frac{t_2 - t_1 + 1 - \bar{S}}{\delta})}} \quad (3)$$

2）宏观投资行为。此处主要考虑宏观上十大流通股东的投资行为，需要考虑：一是早期进入市场的股东其投资行为数是否较晚期进入市场的股东更多；二是活动区间长的股东其投资行为数是否较活动区间短的股东更多些。如果上述假设为真，则

有必要给不同时间入市的股东赋予不同的权重。图 5（a）和图 5（b）对上述假设做出说明，早期进入市场的股东的投资行为均值较高，且股东的活动区间长度越长，对应的股东投资行为均值越高。故有必要对不同时间入市的股东赋予不同的权重。

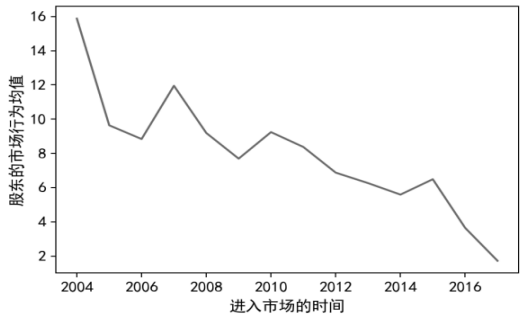


图 5（a）早期进入市场的股东的投资均值

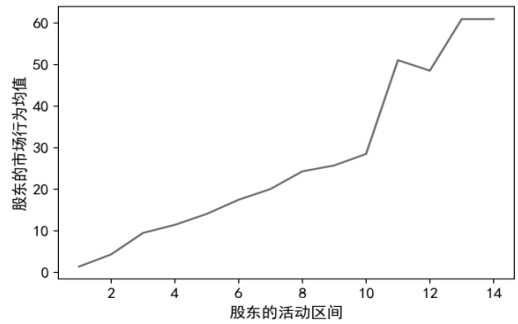


图 5（b）股东活动区间长度与投资均值关系

本文认为应该综合多种特征来刻画股东的投资行为，因此提取以下特征：宏观股东投资行为的总数、宏观股东投资股票的总数。对于上述股东特征引入 sigmoid 函数，设某特征为 x（x 可以为股东投资行为的总数、宏观股东投资股票的总数），则该特征的 x_{index} 为：

$$x_{index} = \frac{1}{1+\frac{1}{x+a}e^{-\frac{x-\bar{x}}{\delta}}} \tag{4}$$

其中，a 的取值极小，如 0.001，是用于平衡可能存在投资行为或投资股票的总数为 0 的可能性。把上述特征值赋予权重 w_x 并求和得：

$$x_u = \sum w_x \cdot x_{index} \tag{5}$$

综上所述，股东活跃度的指标 q_i 为：

$$q_i = w_S \cdot S_{yearOfHolder} + x_u \tag{6}$$

式中，$S_{yearOfHolder}$ 表示公式（1）得到的股东“年龄”指标；x_u 代表股东投资行为的总数、宏观股东投资股票的总数。但是在实际不同阶段，分析人员可根据实际需要调整权重以平衡股东“年龄”和宏观股东行为，从而得到合适的股东活跃度标签。这里取公式（6）中 q_i 指标最高的 20%为高活跃度股东（H），20%~50%为中活跃度股东（M），其他为低活跃度股东（L）。同时可以根据实际需求调整阈值，获得不同的活跃度群体。

3. 2 股东偏好特征

3.2.1 股东行业偏好

股东行业偏好是股东偏好信息中最重要的信息。得知股东行业偏好信息可以帮助外部投资者对自己感兴趣的大股东加深了解。对于股东行业偏好的分析是从两个角度进行的，即行业门类和大类。根据 2012 年修订的《上市公司行业分类指引》，行业门类共 19 种取值，行业大类有 90 种取值。

本文采用如下特征五元特征组来形式化反映股东行业门类偏好：

$$IC = < SUM_{\sum beh}, SUM_{\sum Ct}, Y_n(Ct_i), Y_n(\sum Ct), m_{Ct_i} > \tag{7}$$

其中，$SUM_{\sum beh}$、$SUM_{\sum Ct}$ 分别表示宏观投资行为总数及投资行业门类总数；$Y_n(Ct_i)$、$Y_n(\sum Ct)$ 分别表示各年股东投资行业门类数量及分布；表示行业门类的平均投资水平。对于上述特征均需 m_{Ct_i} 要计算指标绝对值及其占比。也可根据实际情况赋予权重并制定阈值。

针对某股东投资的每一个行业门类 I_{Ct_i}，其中 Ct_i 表示其中的一个行业门类，计算如下：

$$I_{Ct_i} = w_{mac}\frac{SUM_{\sum beh}}{SUM_{\sum Ct}} + w_{mic}\sum_{n=1}^{j}\mu\frac{Y_n(Ct_i)}{Y_n(\sum Ct)} + w_m m_{Ct_i} \tag{8}$$

式中，w_{mac}、w_{mic}、w_m 分别对应于各个特征的权重，$SUM_{\sum beh}$ 是股东宏观上投资行业门类 Ct_i 的行为总数，$SUM_{\sum Ct}$ 是宏观上投资行业门类的总数。μ 表示距离数据集最后投资年份的跨度，如当处理 2016 年数据门类时，对应的 μ 取 1/2；当处理 2015 年数据门类时，μ 取 1/3,以此类推。m_{Ct_i} 为某股东投资的行业门类 Ct_i 的总次数是否超过该门类平均投资水平 $\overline{C}$，且计算如下：

$$m_{Ct_i} = \{0/1\} \tag{9}$$

当股东对某行业门类的投资数大于该门类的平均投资概率 $\overline{C}$ 时，m_{Ct_i} 取 1；反之，则取 0。对于某行业门类 Ct_i 的平均投资概率 $\overline{C}$,定义如下：

$$\overline{C}=\frac{\text{行业门类}Ct_i\text{的投资总数}}{\text{投资行业门类}Ct_i\text{的股东总人数}} \quad (10)$$

对于求出的 I_{Ct_i} 结果，可按照从高到低的方式进行排序，取前百分 20%最为该股东行业门类偏好，同时可以根据实际需求调整阈值，获得不同的行业门类偏好组群。对于行业大类的分析，与行业门类思路相同，在此不做赘述。

3.2.2 股东地域偏好

在股票证券投资领域，地域因素、地域情结是否是投资者投资决策的重要影响因素，针对该问题，本文对股东的地域偏好进行研究分析。由于数据集自身的条件约束，本文对于股东地域偏好的研究主要考虑以下两方面：（1）对于本国国内地区的偏好（2）对于沪深市场的偏好。而不针对国内外地域偏好进行分析，但从分析十大流通股东中 QFII 所占的比例可以看出投资者更倾向于投资本国内的证券市场。

股东对于国内各地区的投资偏好，本文采用如下四元特征组来形式化表示：

$$DAP=<SUM_{\mathrm{ar}_i},SUM_{\sum ar},Y_n(ar_i),Y_n(\sum ar)> \quad (11)$$

其中，SUM_{ar_i}、$SUM_{\sum ar}$ 分别表示宏观投资地域数量及分布；$Y_n(ar_i)$、$Y_n(\sum ar)$ 分别表示各年投资地域数量及分布。计算公式如下：

$$V=\mathrm{w}_{mac_i}\frac{SUM_{ar_i}}{SUM_{\sum ar}}+w_{mic_i}\sum_{n=1}^{j}\mu\frac{Y_n(ar_i)}{Y_n(\sum ar)} \quad (12)$$

其中，w_{mac_i} 表示宏观投资地区的数量的权重；w_{mic_i} 表示各年投资地区的数量的权重。SUM_{ar_i} 表示宏观投资地区 air_i 的总次数，$SUM_{\sum ar}$ 表示宏观投资所有地区的总次数，μ 表示距离数据集最后投资年份的跨度。对于股东投资的各个地区，均采用上述公式进行计算，结果中的前 20%作为股东国内各地区偏好的组群。

股东对于沪深市场的偏好，本文采用各年投资沪市股票的总数量、各年投资深市股票的总数量作为特征，对于上述股东特征引入 sign 函数，计算方式如下：

$$\mathrm{per}=sign(\sum_{n=1}^{j}\mu\frac{Y_n(SUM_{hu})-Y_n(SUM_{shen})}{Y_n(SUM_{hu\&shen})}) \quad (13)$$

式中，$Y_n(SUM_{hu})$、$Y_n(SUM_{\mathrm{shen}})$ 分别表示在第 n 年，股东对于沪市股票及深市股票的投资总数，μ 表示距离数据集最后投资年份的跨度。当 $\mathrm{per}>0$ 时，即股东更倾向于投资沪市，反之，股东更倾向于投资深市。

通过从活跃度、偏好特征两方面构建股东画像，外部投资者可以了解到想要跟投的股东的市场行为频繁程度及各类偏好信息，从而加深度了解，并从中选取真正符合自己投资意愿的股东进行跟投。最终实现指导外部投资者进行投资决策的目的。

4 证券市场股东画像构建实例

本文的数据集主要涉及到爬虫得到的财经门户网站上 3107 只上市 A 股对应的十大流通股东行为数据、国泰安数据库中上市公司基本信息及基金持股数据。其中涉及到 2004 年至 2017 年期间共 1060700 条投资行为数据及 128045 位股东数据。结合图 2 的算法预处理过程得到干净数据集后，首先在全集上考察股东群体的整体活跃度、偏好等趋势，其次将全集股东分为 3 类分别从个人、基金、机构说明这 3 类性质的股东群体在股东活跃度、股东投资偏好方面存在的相似性与差异性。

图 6 展示的是数据集全时间段内各年份对应的 3 类不同性质的股东数量及变化情况。由图可知，三类股东中，个人股东占比始终最高，机构次之。从 2007 年开始，机构股东的数目基本保持稳定不变。且随着时间的变化，基金类型的股东数量不断增加，总体上看，全集股东人数呈日益扩张趋势。

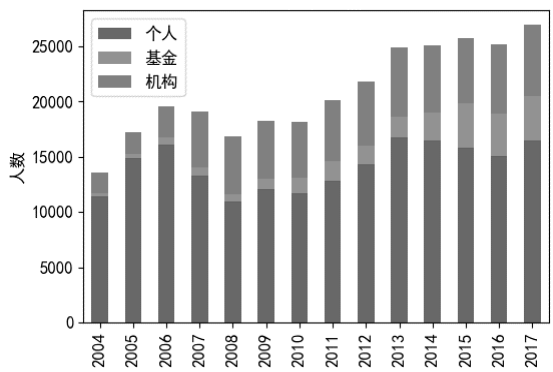


图 6 3 类性质股东数量分布图

通过在数据全集上考察上述定义的各特征变量是否影响股东活跃度，给出图 7 作为解释，图中显示了具有高活跃度的股东在各类特征上的表现。图 7 横轴表示活跃度的取值，纵轴分别表示最早最晚投资时间、活动区间、股东“年龄”、投资股票总数和行为总数。

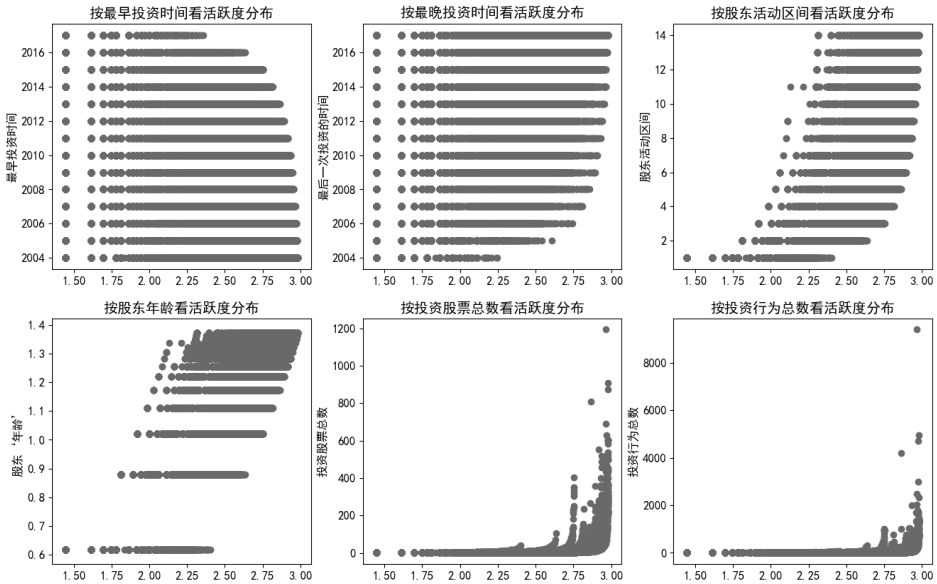


图 7 在全集上考察各特征与活跃度之间的相关性

图7可以看出具有高活跃度的股东应该具有的特点是，最早投资时间更偏向早期、最晚的投资时间接近于数据集的截止时间、股东的活动区间长度越长，该股东活跃度越高的可能性越大、股东“年龄”更多的处于 1.1 到 1.4 之间、投资股票总数及行为总数与活跃度存在指数关系。

利用本文式(2)定义的股东活跃度模型，分别对个人、机构、基金三类股东进行建模，得出如图 8 所示的结果。分析表明：（1）图中显示的大多数股东的活跃度偏低，且 3 类股东的活跃度较多的集中于 1.5 左右，是由于股东活动区间为 1 的全集股东占绝大多数所造成的。（2）3 类不同性质股东的活跃度呈现出的分布变化趋势是大致相同的，这也从另一方面说明了，用于描述股东的“活跃度”是均衡的。

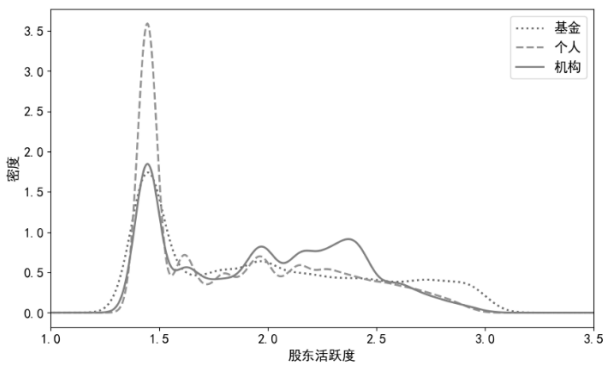


图 8 3 类股东活跃度对比图

从 3 类不同性质的股东出发，在全集上考察股东的行业偏好。全集包括 22921 条机构类型股东，个人股东 93170 名，基金股东 11954 名。具体的偏好是经过式（8）计算得到的。取前 20%为各个股东的偏好行业，再汇总综合考虑 3 类不同的股东群体，从整体考察股东偏好。最终得出图 9 所示的股东群体偏好。

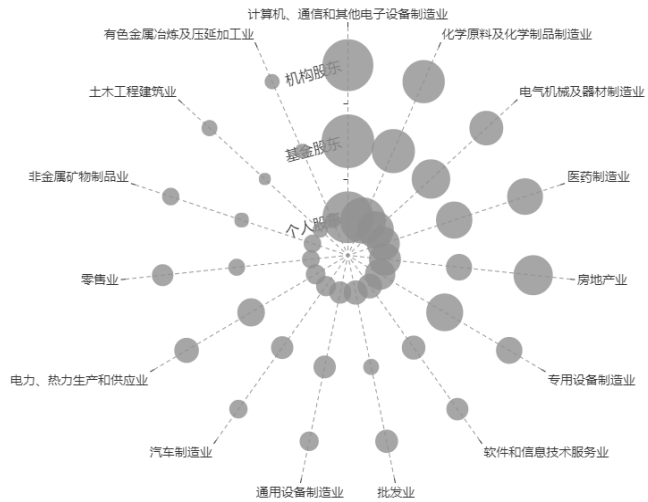


图 9 3 类股东群体偏好

图 9 从内到外由个人股东、基金股东和机构股东 3 层组成，显示了 15 类排行靠前的股东喜好行业。圆圈的大小表示偏好值，某位置的圆圈越大意味着该类股东对于该行业的偏好强度越高。不难发现，3 类股东的行业偏好相似度极高，基本完全一致。这也从侧面反应了股东群体所偏好的股票所属的行业均是发展前景较好、且被多数重要股东认可的，具有一定的参考价值。

图 10 展示了沪深市场的投资变化情况。我们可以看出深市的投资数量较沪市随时间呈激增变化趋势；3 类股东中，机构股东对于沪深市场的投资数量基本持平；个人股东对于深市的投资整体呈上升趋势，而对于沪市的投资波动较大；基金股东在 2015 年之前对于沪深的投资数量呈上升趋势，之后存在明显的骤降现象。且经过式（13）计算可得出，个人股东 $per=-0.82$，故更偏好于深市；基金股东 $per=-0.53$，更偏好于深市；机构股东

$per=-0.05$偏好于深市。三类股东中个人、基金股东的沪深偏好较为明显，而机构的沪深偏好差异不大。与图 10 所示的投资变化一致。

图 10　沪深市场投资数量变化图

图(11)展示了 3 类股东投资国内各省份的占比情况。由图可知，(1)3 类股东他们的投资倾向具有相似性,且相似性极好；(2)在这些区域中，股东投资比例超过 6%的区域有，广东省、北京市、上海市、江苏省及浙江省，可以看出一线或经济相当高度发达区域的企业的收益可能会更好，原因在于这些区域经济可能发展的更活跃，机会可能会更多，所以相应的企业发展就会更好些。

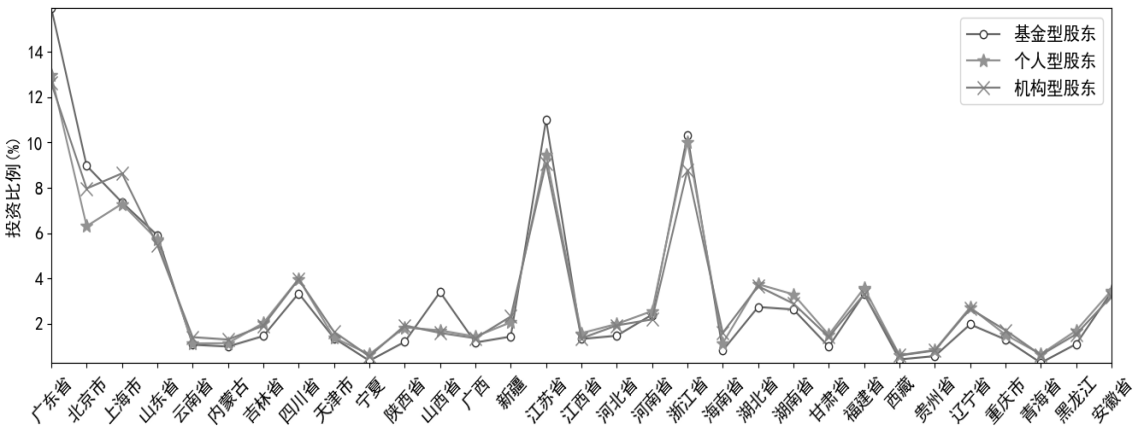


图 11　3 类股东投资各个地区的占比图

之后对投资占比超过 6%的地区开展进一步研究，去发现在这些区域中 3 类股东重点投资的行业都有哪些，存在什么样的差异性。首先假定投资超过 6%的地区为较为发达的地区，结果表明在 90 个行业大类中，无论是个人、基金还是机构，他们最为重点投资的行业大类都是计算机通信和其他电子设备制造业。且 3 类股东投资的排名前 10 的行业大类差异不大，仅在排名先后顺序上略有不同。考虑到投资者投资时会选择发展潜力大的企业，由此我们也可以推断出在较为发达的地区，龙头企业所属的行业主要是,计算机通信和其他电子设备制造业、房地产业、电气机械及器材制造业、化学原料及化学制品制造业、医药制造业等。

5 结论

本文将用户画像的概念迁移至证券市场，针对我国上市A股的十大流通股东提出“股东画像”的概念，以便于深刻了解股东偏好，力图从更详细、更多角度对外部投资者的投资提供指导作用，本研究可以作为分析具有相似投资行为的股东群体、股东风险偏好的基础。其次利用爬取和国泰安数据库（CSMAR）获取到的股东、股票相关数据从活跃度、偏好特征 2 个方面构建了证券市场股东画像，将股东按性质划分为个人、机构、基金，并分别对 3 类股东从全集角度构建画像。除此之外，本文还提出了一些处理股东特征的思路和方式，特别是针对如何刻画股东“年龄”特征及股东行业门类偏好。

由于数据集的限制，本文没有对于国内外地域偏好进行针对性研究。对于股东的投资行为并没有细化为增持、减持、不变、新进 4 类，从更加细致的角度分析股东画像。未来，笔者会针对细化后的股东投资行为、相似投资行为的股东群体做出分析，得到更有意义的证券市场股东画像。并针对由股东、股票、上市公司生成的三模社会网络拓扑图进行链接预测、分类以及演变分析。

致谢

感谢国家自然科学基金（61741208），教育部“云数融合科教创新”基金项目（2017B00030），中央高校基本科研业务费（ZDYF2017006），陕西省科技厅协同创新项目（2015XT-21）以及陕西省软科学重点项目（2013KRZ10）的联合支持。

参考文献

[1] 曾鸿，吴苏倪．基于微博的大数据用户画像与精准营销[J]．现代经济信息，2016(16):306-308
[2] LEE W J，OH K J，LIM C G，et al. User profile extraction from Twitter for personalized news recommendation [C]//International Conference on Advanced Communication Technology，2014: 779-78
[3] 李恒超，林鸿飞，杨亮,等．一种用于构建用户画像的二级融合算法框架[J]．计算机科学，2018，45(1):157-161.
[4] Mueller J，Stumme G. Gender Inference using Statistical Name Characteristics in Twitter[C]//The，Multidisciplinary International Social Networks Conference on Socialinformatics 2016，Data Science. ACM，2016:47.
[5] 黄文彬，徐山川，吴家辉,等．移动用户画像构建研究[J]．现代情报，2016，36(10):54-61..
[6] 张慷．手机用户画像在大数据平台的实现方案[J]．信息通信，2014(2):266-267.
[7] 费鹏，林鸿飞，杨亮,等．一种用于构建用户画像的多视角融合框架[J]．计算机科学，2018，45(1):179-182.
[8] 倪光耀．大股东增持股票的目的和影响的研究[J]．现代经济信息，2017(2):298-298.
[9] 赵红岩，仲达飞．大股东增减持行为对我国上市公司股票价格的影响分析[J]．时代金融，2017(5).
[10] 刘建忠．重要股东持续坚持下的公司绩效和成长性研究[J]．纳税，2008(4)

于花蕾（出生于 1994 年），在读研究生，硕士，主要研究领域为自然语言处理、复杂网络分析。
E-mail：13227805056@163.com

饶元（出生于 1973 年），通讯作者，博士学位，副教授，这要研究领域为文本数据挖掘、自然语言处理、机器学习以及社会网络分析等相关领域的知识基础。
E-mail：yuanrao@163.com

汤才芳（出生于 1992 年），在读研究生，硕士，主要研究领域为自然语言处理、知识图谱。
E-mail：tang_c_fang@163.com