# 基于机器学习的场外配资自动识别系统

中泰证券股份有限公司 何波

# 2015年7月股灾，拉开清理场外配资序幕

现公布《关于清理整顿违法从事证券业务活动的意见》，自公布之日起施行。

中国证监会
2015年7月12日

**关于清理整顿违法从事证券业务活动的意见**

一段时期以来，部分机构和个人借助信息系统为客户开立虚拟证券账户，借用他人证券账户、出借本人证券账户等，代理客户买卖证券，违反了《证券法》、《证券公司监督管理条例》关于证券账户实名制、未
益，严重扰乱了股票市场秩序。近日
头，可能再次危及股票市场平稳运行

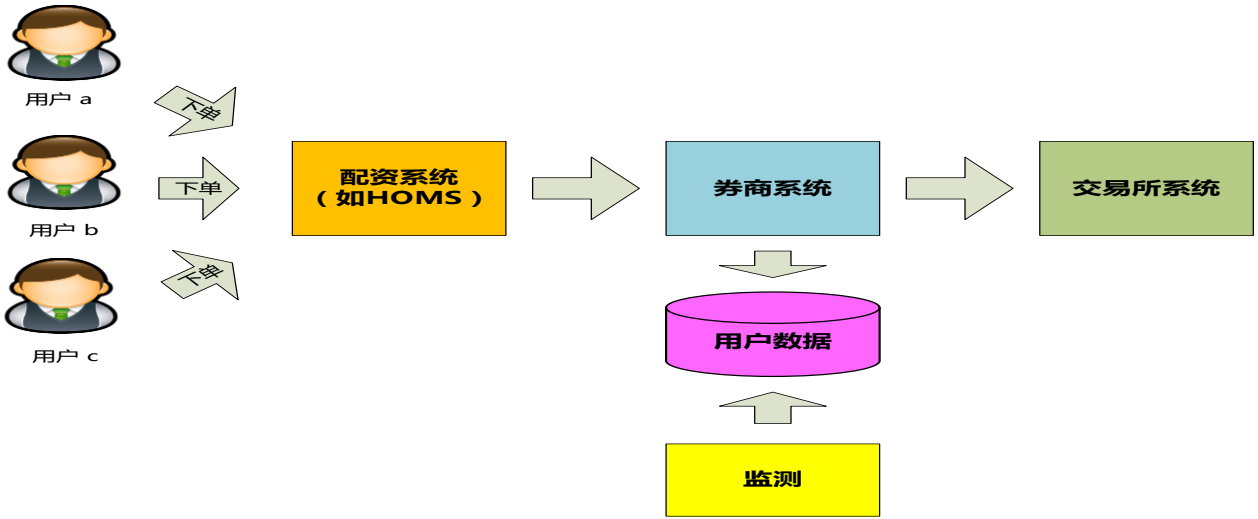
# 总体上配资账户加剧了股市异常波动


## 深交所《关于场外配资账户的研究进展的汇报》：

- 配资账户持股占比越高的股票，股票的波动性越大。
- 配资账户导致股票特质波动率增加。
- 配资账户偏好中小板股票，资金周转率高于深市平均水平，表现出“追高”、“炒小”、“炒热点”等特征。
- 配资账户加大了股票流动性恶化极端事件的出现频率。

# 如何识别一个账户多人交易还是单人交易？

## 配资账户交易模型



# 传统的配资账户查处方法

## 根据分析和经验建立一系列筛选条件

- 账户总资产达到一定规模
- 成交量达到一定规模
- 交易频率满足一定次数

## 传统的配资账户查处方法存在以下几个主要问题

- 筛选规则由监管人员主观决定，缺乏客观依据
- 配资账户行为模式不断发生变化，筛选规则难以及时做出调整
- 筛选规则有限，难以实现多样复杂的筛选规则
- 配资账户样本数据远小于正常账户数目，样本存在严重不平衡问题

# 基于机器学习的配资账户查处方法

## 利用机器学习算法，建立配资账户的自动识别模型，有效提升了配资账户查处的效果：

- 充分利用客户的基础数据，交易数据，资产数据等，多维度判别某一账户是否为配资账户
- 通过设计用以描述配资账户的特征，充分刻画配资账户的独特特性和行为模式，从而区别非配资账户
- 增添查处规则的规模，实现复杂的查处过程
- 利用机器学习算法，通过复杂的模型内部算法提升查处的准确性
- 对于未知的账户，将其相关数据输入模型，模型根据内部算法，实现自动预测该账户是否为配资账户

用户数据

用户特征

模型训练

生成模型

预测配资账户

# 建立配资账户自动识别系统流程

## 特征设计

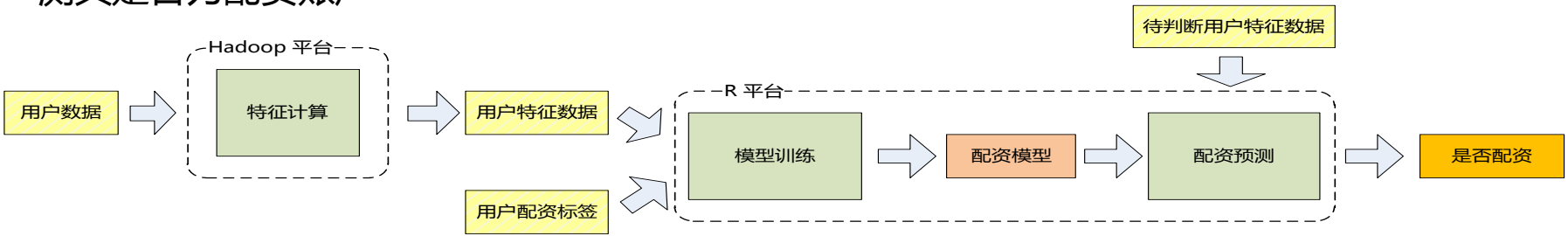
➢通过设计描述配资账户特有行为模式的特征，区别配资和非配资账户

## 模型选择

➢将特征输入到不同的机器学习模型，基于检验结果选择最优模型和参数

## 模型预测

➢对于待识别的账户，基于其原始数据计算特征，将特征输入模型，模型基于内部算法预测其是否为配资账户



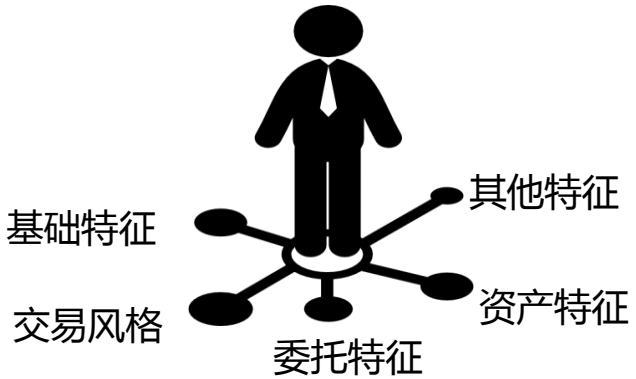
# 特征设计

## 特征设计原则

- 特征用以描述配资账户特点，区别配资和非配资账户

## 经观察配资账户主要有两个特点

- 配资账户内部交易风格差异较大
- 配资账户收益特征不同



# 模型特征选择

## 目前我们设计采用特征包括：

- 账户总资产
- 账户股票资产
- 持仓率
- 持有股票支数
- 平均每支股票持有仓位
- 持有仓位标准差
- 持有股票最大/最小仓位占总持仓比例
- 持有股数少于某一阈值的股票支数
- 每日总成交次数
- 每日单只股票最大成交次数
- 每日成交量少于某一阈值的成交次数
- 成交价格标准差
- 每日成交股票支数
- 总成交量
- 成交量最大值/最小值
- 平均每笔成交量
- 单只股票最大/最小成交量量
- 成交量量标准差
- 单只股票成交量量标准差最大值
- 每日成交量峰度
- 每日成交量偏度
- 股票成交量峰度最大值
- 股票成交量偏度最大值
- 每只股票成交量峰度标准差
- 每只股票成交量偏度标准差
- ……

# 模型特征展示

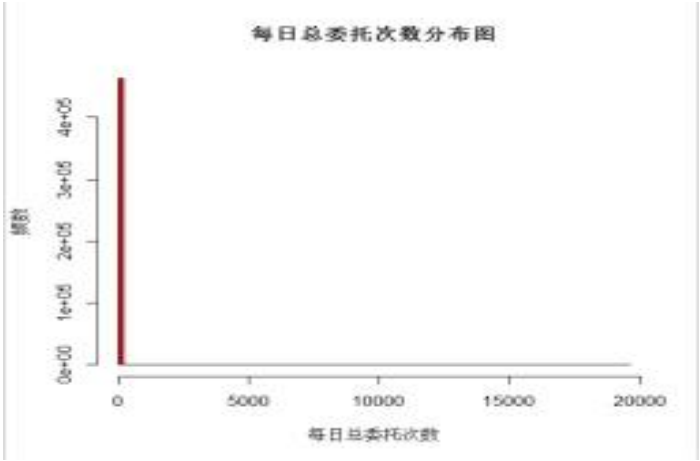
## 用于区别配资账户和非配资账户的特征

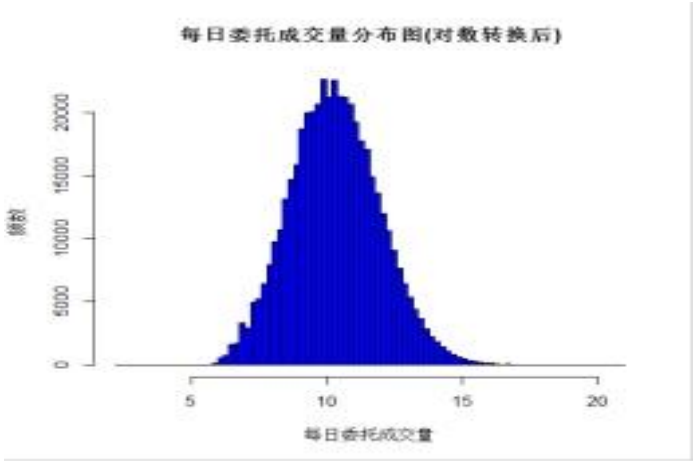
| 特征类型 | 特征 | 配资账户 | 非配资账户 |
|---|---|---|---|
| 规模类型特征 | 成交频率 | 36 | 3 |
| | 成交量 | 2,910,721 | 29,087 |
| | 股票资产 | 5,056,762 | 48,054 |
| | 持有股票数目 | 33 | 3 |
| 差异类型特征 | 成交量标准差 | 1.09 | 0.04 |
| | 股票仓位标准差 | 1.37 | 0.43 |

# 模型特征展示

## 特征转换

考虑账户特征服从幂率分布的规律（大部分特征值分布在较小的一段区间内，其余特征值分布在较大的区间内），对特征采取对数转化，可以呈现更好的分布效果

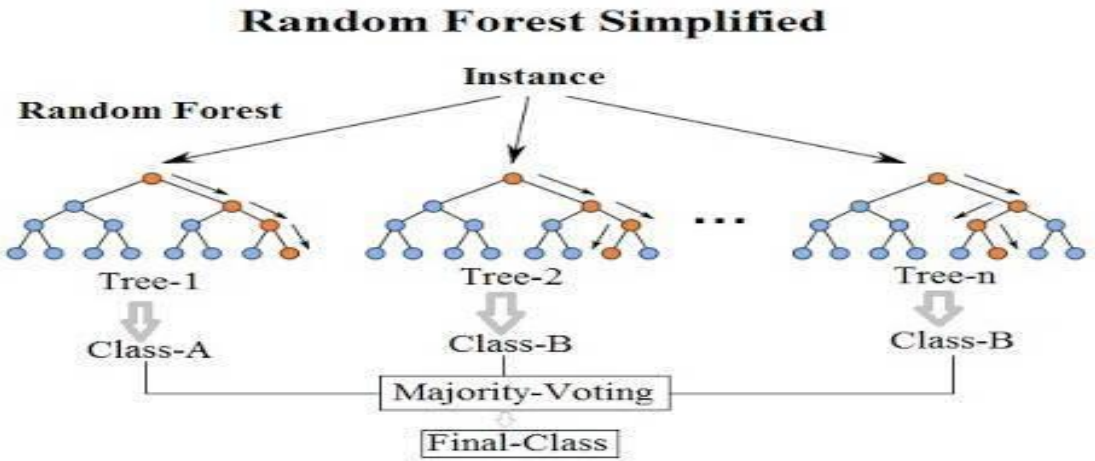


# 模型：随机森林简介

## 经过多种模型对比，随机森林算法具有较好的效果

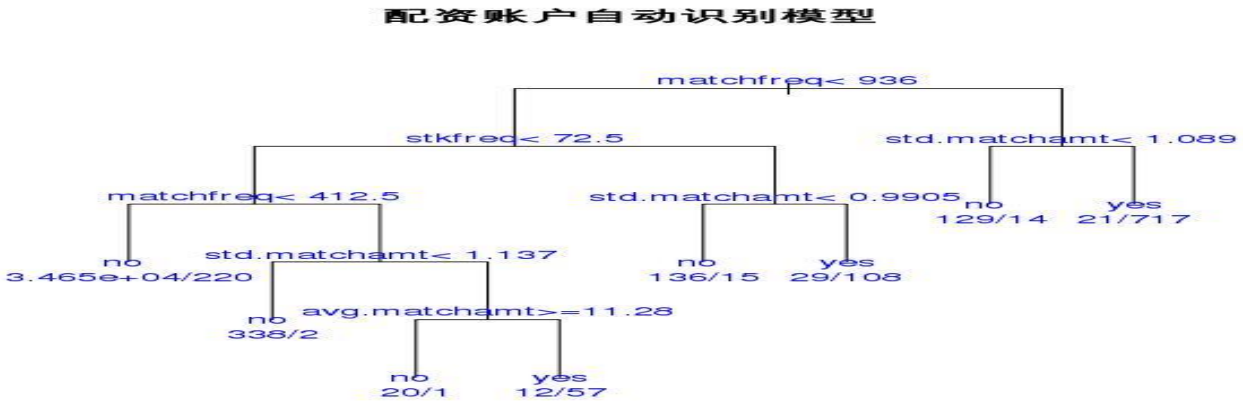
随机森林属于分类算法，建立多个分类树模型，生成一系列的判定筛选条件，根据这些条件判断某一账户是否为配资账户。



# 模型：随机森林简介

## 基于随机森林算法生成的简化配资账户识别模型

- 特征名称：交易频率（matchfreq）、交易股票支数（stkfreq）、成交量标准差（std.matchamt）、平均每笔成交量（avg.matchamt)
- 模型可视化：对于每一个子节点，左边的数字表示该节点中的普通账户数目，右边数字表示该节点中的配资账户数目，可以看出，每一个子节点都很好的实现了配资账户和普通账户的划分，使该节点的大部分账户为同一账户类别。



# 模型：随机森林简介

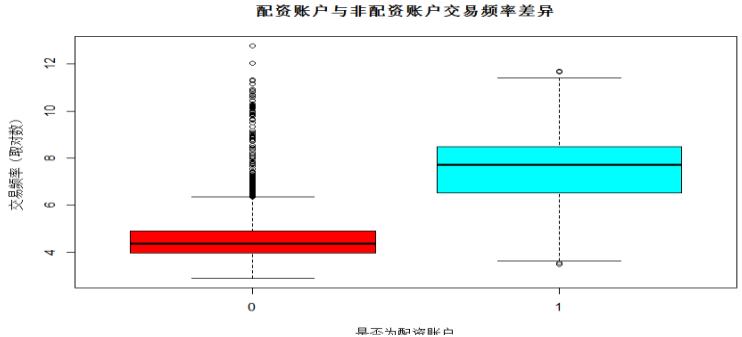
## 基于随机深林模型特征重要性

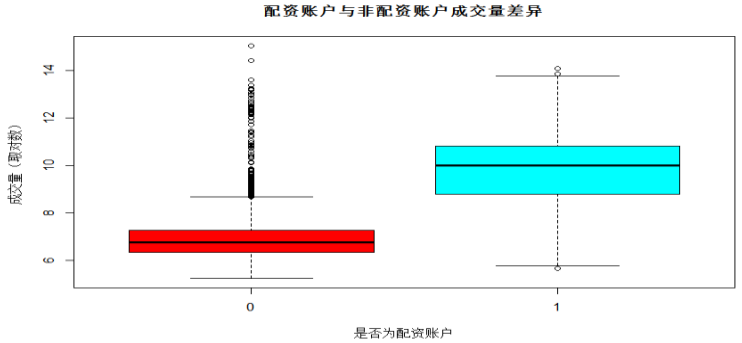
- 随机森林模型根据模型内部算法和模型预测结果，对每一个特征的重要性打分。
- 重要性排名前10特征展示

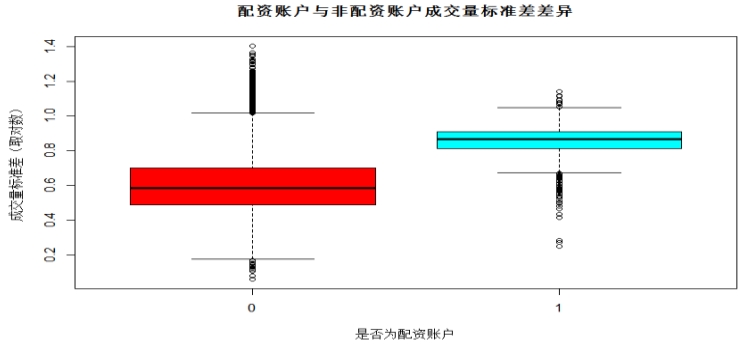
| 特征排名 | 特征名称 | 特征重要性 | 特征类型 | 特征排名 | 特征名称 | 特征重要性 | 特征类型 |
|---|---|---|---|---|---|---|---|
| 1 | 交易频率 | 16.0% | 规模类特征 | 6 | 个股成交量标准差波动 | 5.4% | 差异类特征 |
| 2 | 总成交量 | 14.3% | 规模类特征 | 7 | 个股成交量偏度 | 3.1% | 差异类特征 |
| 3 | 交易股票支数 | 12.66% | 规模类特征 | 8 | 个股成交量峰度 | 2.7% | 差异类特征 |
| 4 | 成交量标准差 | 10.22% | 差异类特征 | 9 | 交易频率偏度 | 2.7% | 差异类特征 |
| 5 | 个股成交量标准差极值 | 7.0% | 差异类特征 | 10 | 交易频率峰度 | 2.6% | 差异类特征 |

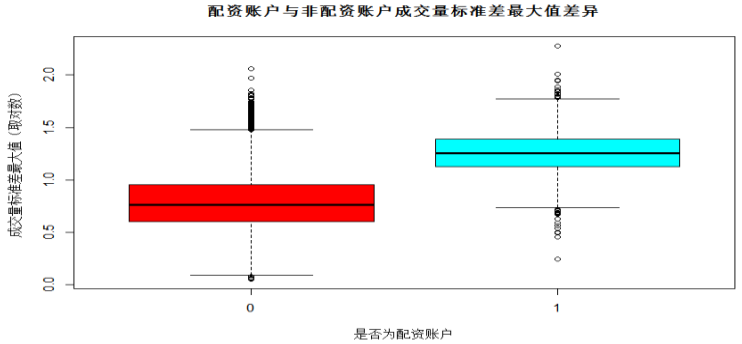
# 模型：随机森林简介

## 重要特征在配资账户与非配资账户间分布差异



# 模型检验效果

- 使用交叉验证的方法，按8:2的比例把原始数据集分成模型生成样本和模型检验样本
- 模型生成样本用于生成模型，模型检验样本用以检验模型的效果，以避免模型的过度拟合。
- 模型检验样本总共含有7295个账户，其中配资账户227个，非配资账户7068个，配资账户占总样本比例为3.1%。
- 经多次抽样迭代，模型平均能够识别配资账户182个，召回率80%左右，误识别的非配资账户10个，准确率95%左右。

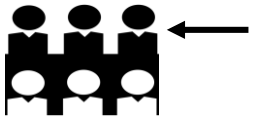
# 配资账户检测

- 配资监测模型每日运行，对当天交易和资产数据进行监测
- 定期报送可疑名单到运营管理部门以及CRM系统
- 对可疑名单进行电话回访，发现明确提供配资服务。

数据仓库

算法集群

客服系统



疑似配资账户

正常账户

# 模型优化

## 模型优化的主要方向

- 增加特征，配资账户画像更加完备
- 对于模型中的特征进行分析，识别强特征和弱特征，挖掘强特征背后的逻辑和原因
- 对模型的参数进行调试和优化
- 通过增加模型的复杂程度来优化模型，提升模型准确率

# 智查配资账户识别系统开发

## 智查配资账户自动识别系统开发

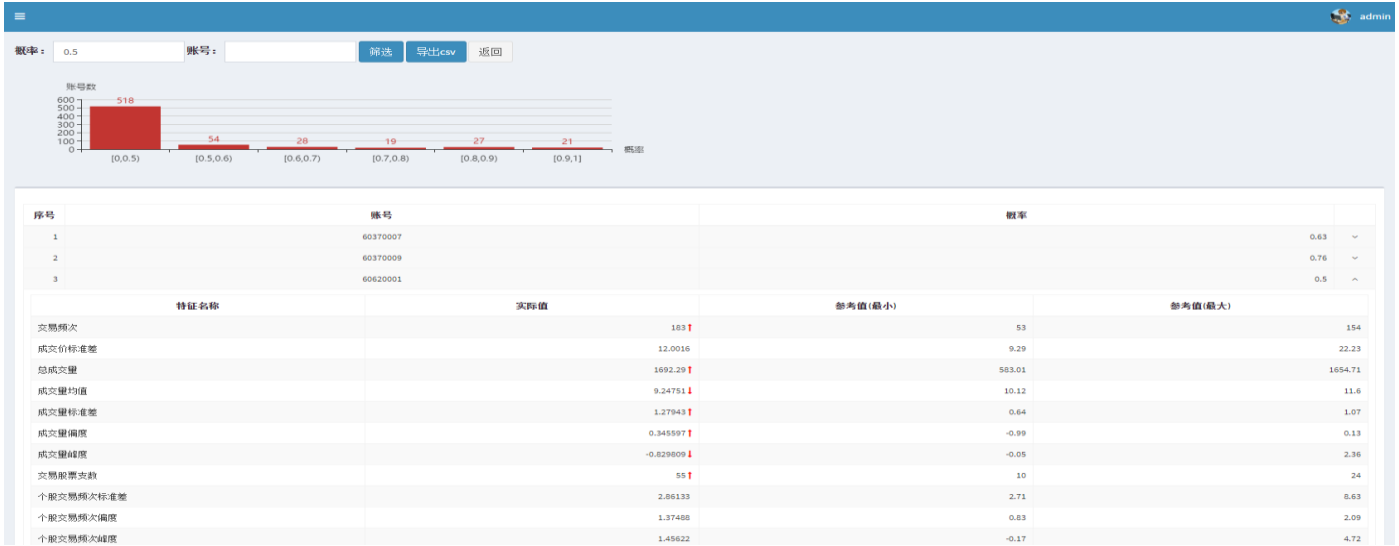
- 在星环TDH上完成整个配资系统产品化开发
- 使整个配资账户识别流程一体化

## 第一版本与2017.06开发完成

# 智查配资账户识别系统开发

## 查询结果展示：

- 显示每个未知账户是配资账户的可能性
- 显示每个账户的各个特征值和该特征值的正常区间




| 序号 | 账号 | 概率 |
|---|---|---|
| 1 | 60370007 | 0.63 |
| 2 | 60370009 | 0.76 |
| 3 | 60620001 | 0.5 |

| 特征名称 | 实际值 | 参考值(最小) | 参考值(最大) |
|---|---|---|---|
| 交易频次 | 183↑ | 53 | 154 |
| 成交价标准差 | 12.0016 | 9.29 | 22.23 |
| 总成交量 | 1692.29↑ | 583.01 | 1654.71 |
| 成交量均值 | 9.24751↓ | 10.12 | 11.6 |
| 成交量标准差 | 1.27943↑ | 0.64 | 1.07 |
| 成交量偏度 | 0.345597↑ | -0.99 | 0.13 |
| 成交量峰度 | -0.829809↓ | -0.05 | 2.36 |
| 交易股票支数 | 55↑ | 10 | 24 |
| 个股交易频次标准差 | 2.86133 | 2.71 | 8.63 |
| 个股交易频次偏度 | 1.37488 | 0.83 | 2.09 |
| 个股交易频次峰度 | 1.45622 | -0.17 | 4.72 |

# 单账户配资

## ➢ 后续工作：

配资账户识别系统主要针对伞型配资账户，后续准备展开针对单账户配资的研究以及相关模型的开发工作。

## ➢ 背景工作：

2015年股市大牛市中出现的配资账户主要为伞型配资账户，在证监会及相关机构的大力监管下，伞型配资账户逐渐减少，目前流行的配资形式为单账户配资。

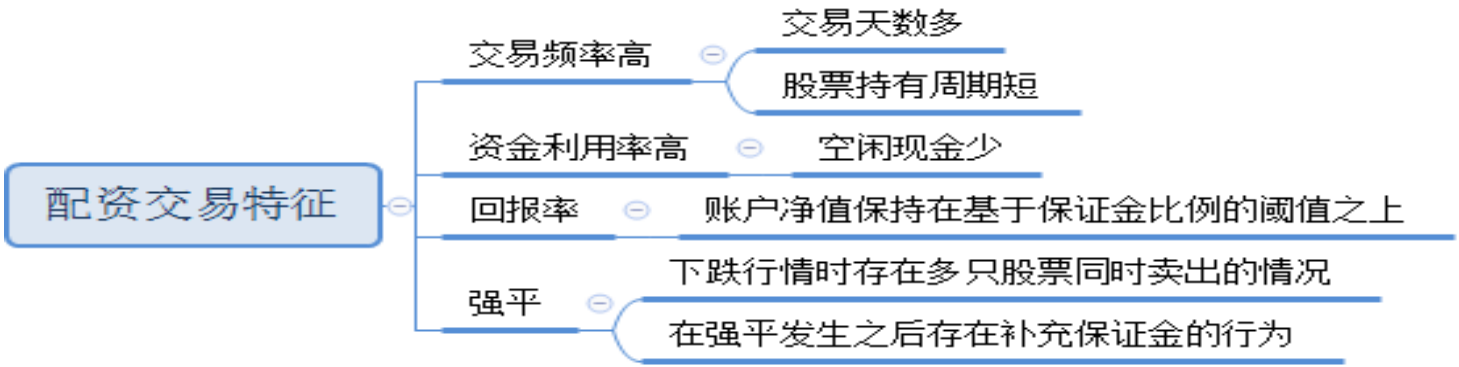
## ➢ 单账户配资特点：

一个人操作一个账户

一个账户在不同的时间点可能被不同的人使用

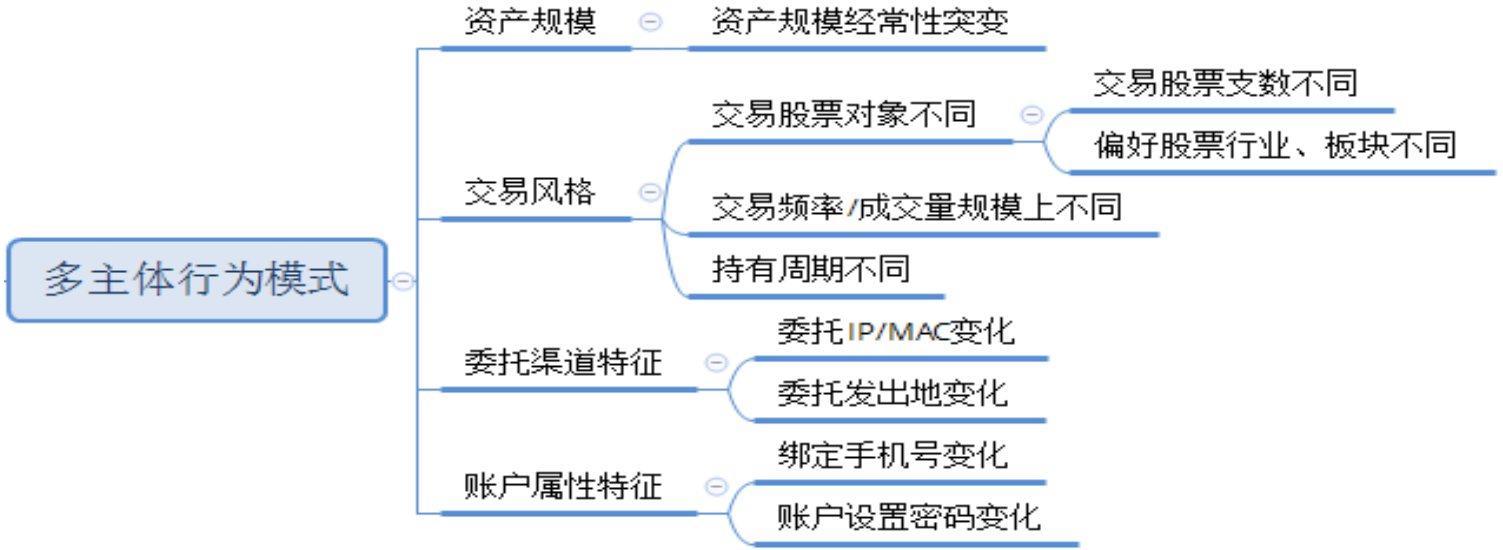
# 单账户配资交易特征

单账户配资具有配资交易的一些共性交易特征



# 单账户配资时间维度特征

由于单账户配资会在不同时间分配给不同的交易客户，在时间维度上存在一些显著特征可以区别单账户配资和普通账户



谢谢！