

Data Structure

Tuesday, August 24, 2021 3:58 PM

1. Stock ID

a. General

- i. 名字
- ii. 所属行业
- iii. 类型 (所属规模)

b. Monthly Performance (past 2 years - present)

- i. 收盘价
- ii. 成交均价
- iii. 换手率
- iv. 振幅
- v. 融资融券差额

c. Quarterly report

- i. 最新市值
- ii. 基金持股比例总和
- iii. 持股基金个数总和
- iv. 持股排名前十基金名字 (需通过基金信息创建)
- v. 持股排名前十基金持股比例 (需通过基金信息创建)

2. Fund ID

a. General

- i. 名字
- ii. 所属基金公司
- iii. 类型 (如股票开放型, 混合开放型)
- iv. 基金经理
- v. 该基金经理管理的基金数目 (需推算)
- vi. 该基金经理管理的基金名称 (包含本基金, 需推算)

b. Quarterly performance (past 2 years - present)

- i. 最新资金规模
- ii. 选证能力评级
- iii. 择时能力评级
- iv. 基金资产总值
- v. 净值
- vi. 净值增长率
- vii. 收益
- viii. 前十大重仓股名字
- ix. 前十大重仓股合计占基金净资产值比 (最好有breakdown, 需推算)
- x. 不同风格占比 (需推算)

通过股票信息可以汇总出行业信息

通过股票月度表现，与股票季报中的基金信息关联起来

通过股票季报中的基金信息，结合月度表现，寻找建仓点

通过股票季报中的基金信息，查看基金公司的行动一致性

通过基金的投资风格变化趋势，来绘制基金总体的倾向

通过基金在重仓股票的足迹来绘制基金投资去向及趋势

通过基金在重仓股的投资力度，与股票的表现关联

通过基金经理的评级，规模等，构建不同基金的影响力权重

Feature:

1. 股票所处行业的静态信息（行业信息是指行业对应的个股的某个指标的总和，以个股的第一行业为准）
 - a. 行业个股在沪深指数300中的占比的平均值（单位：%）
 - 如一个行业中有10个个股在300指数中，则占比为10/300
 - b. 行业个股的大中小盘比例（单位：%）
 - 大盘股比例
 - 中盘股比例
 - 小盘股比例
 - c. 行业个股的股票风格分类统计（单位：%）
 - 混合型比例
 - 价值型比例
 - 成长型比例
2. 股票所处行业的整体趋势描述（行业信息是指行业对应的个股的某个指标的总和，以个股的第一行业为准）
 - a. 过往N个季度的市值平均变化率和标准差（单位：%）
 - 举例：从季度1到季度N，变化率的计算方式为：以季度1为baseline，其他季度的值为sample，
变化率=(sample - baseline)/baseline，平均变化率各个sample的平均值，标准差为各个sample的标准差。
 - 如果是NAN，mean=0，std=0
 - b. 过往N个季度的基金持股总值的平均变化率和标准差（单位：%）
 - 同上
 - 如果是NAN，mean=0，std=0
 - c. 过往N个季度的持股的基金公司数目的变化率和标准差（单位：%）
 - 同上
 - 如果是NAN，mean=0，std=0
 - d. 过往N个季度的持股基金的能力评级的变化率和标准差（单位：%）

- 同上
 - 如果是NAN, mean=0, std=0
3. 个股本身的趋势描述（个股信息是指单个股票的对应指标）：
- a. 过往N个季度的持股基金的能力评级的变化率（单位：%）
 - 同上
 - 如果是NAN, mean=0, std=0
 - b. 过往N个季度的收盘价平均变化率和标准差（单位：%）
 - 同上
 - 如果是NAN, mean=0, std=0
 - c. 过往N个季度的平均成交价变化率和标准差（单位：%）
 - 同上
 - 如果是NAN, mean=0, std=0
 - d. 过往N个季度的平均换手率和标准差（单位：%）
 - 同上
 - 如果是NAN, mean=0, std=0
 - e. 过往N个季度的平均振幅和标准差（单位：%）
 - 同上
 - 如果是NAN, mean=0, std=0
 - f. 过往N个季度的平均融资融券差额和标准差（单位：%）
 - 同上
 - 如果是NAN, mean=0, std=0
 - g. 过往N个季度的基金持股比例的平均值和标准差，最大值和最小值的差（单位：%）
 - 平均值和标准差同上
 - 最大值和最小值的差(下面2个数值相减):
 - 最大值为过去N个季度的最大值
 - 最小值为过去N个季度的最小值
 - 如果是NAN, mean=0, std=0, max-min=0
 - h. 过往N个季度的持股基金个数的平均值和标准差，最大值和最小值的差（单位：%）
 - 同上
 - 如果是NAN, drop this stock.
 - i. 过往N个季度的持股基金的基金公司集中度的平均值和标准差（Top10持股基金，单位：%）
 - 每个季度取Top 10的持股基金，获取他们的基金公司。
 - 然后去同，比如AABBBCDEFG，则为ABCDEFGFG，共7个基金公司
 - 则当季的集中度为： $1-7/10=0.3$
 - 获得每个季度的集中度后，再计算平均值和标准差
 - 如果是NAN, mean=0, std=0
 - j. 过往N个季度的持股基金的回访率（或多次持有次数的平均值）（单位：%）

- 每个季度取Top10个持股基金，N个季度就有 $N*10$ 个sample
- 在 $N*10$ 个sample中，遍历查同，获得所有出现超过1次的基金的名单
- 该名单长度/ $(N*10)$ 就为回访率
- 如果是NAN，mean=0，std=0

k. 过往N个季度的所属行业相对排名的平均值和标准差 (单位：%)

- 计算每个季度的相对排名：(行业总个数 - 个股所在行业排名) / 行业总个数
- 计算平均值和标准差

l. 是否为周期性行业 (0或1)

- xiaoying给出计算方式
- 不预期是NAN

m. 过往N个季度的大基金持有比例 (单位：%)

- xiaoying给出计算方式