景顺长城
Invesco Great Wall

基于机构调研信息的超额收益挖掘

量化及ETF投资部—付亮

2018年9月

目录

01

引言与文献综述

数据介绍

- 2012年底，深交所公告要求上市公司接受机构人员调研后，需在两个工作日内将调研记录发布在深交所互动易（或巨潮资讯网、公司官网上），即“投资者关系活动记录表”。相比于分析师推荐，机构调研时效性高、能反映机构投资者的关注度，但噪音更大。

时效性　关注度

噪音高　数据脏

逻辑解释

- 调研不是简单的关注，机构需要付出时间、经济成本。通常来说，买方的调研是潜在买入，卖方的调研是潜在推荐。从统计意义来说，我们预计 **上市公司被调研越多、调研记录情绪越正面，股价越有可能上涨。**
- 可挖掘信息有：数量特征、内容文本特征。

时间成本　经济成本

价格诉求　潜在买入

现有研究很少

- 海通证券（201805） 构建虚拟变量测试截面溢价，得到显著为正的因子超额收益，月胜率68%，月均溢价0.38%。但未进行策略构建回测。
- 东方证券（201608）根据调研人次构建择时策略，始终买入调研人次最多的四只股票，年化收益75.4%，夏普1.8，回撤34.1%
- 银河证券（201601） 买入调研次数较多的股票，择时分为延迟一个月、两个月买入。

02

数据来源与统计

数据获取渠道

- 深交所互动易、巨潮资讯网（原件PDF、Word）
- 东方财富—特色数据—机构调研
- 第三方数据商，比如通联数据；万得底层数据库

数据展示与统计

- 见网页前端

调研日期：2017年05月22日 星期一

接待基本资料			
公告日期	2017-05-27	上市公司接待人员	副总经理、董事会秘书 黄方红,投资者关系总监 蔡清源,证券事务代表 李义,投资者关系经理 曹静文
接待时间起始	2017-05-22	接待时间截止	-
接待方式	特定对象调研	接待地点	公司总部会议室（包括电话会议）

接待详细对象			
序号	接待对象	接待对象类型	机构相关人员
1	普信资本	-	Eric Moffett
2	国泰基金	基金管理公司	邓时锋,樊利安,王斌,李恒,姚玉涵,孙朝晖
3	国泰君安证券	证券公司	王聪,陈颖,朱颖
4	Tomlinson Capital Management	-	Alvin Yee
5	中国银河国际金融证券	-	Mark Jiang
6	Tiger Pacific Capital	其它	Ryan Yin
7	Cecil Street Capital Management	-	TAY Yi Cheah,Effy Yang,Nicholas Yong
8	Comgest Asset Management	-	Jimmy Chen,前台静
9	Philipse & Co Asset Management	-	Omar Zee,Robertus Van Roozendaal
10	Wilton IS Asset Management	-	Adriaan VAN DER EN
11	Capitael Asset Management	-	Milan Duschka
12	Chenton Asset Management	-	Albertus Van Gaalen
13	Masman Bosman Asset Management	-	Jan Hendrik Wolkers
14	Calamos Advisors	-	Andy Lin
15	Flowering Tree	-	Rajesh Sachdeva,郭梅
16	Columbia Threadneedle Investments	其它	徐进
17	高盛(亚洲)	其它	郑浩云,李嘉平
18	Matthews International	其它	徐伟豪

主要内容资料
Q:关于安防领域人脸识别,现在有很多算法公司也想做智能摄像头,也会选择合作的方式。海康研究院未来的研究方向是什么? A:现在这些算法公司在产业生态中有他们存在的价值,有些公司可能会选择自己做,有些公司也可能选择外部合作。对海康而言,这些算法是放在后台 GPU 上,还是嵌入到前端的摄像机中,这是在产业发展过程中不断演变的,没有固定的模式,我们目

数据清洗流程

爬虫抓取巨潮资讯网原件4万余篇，解析字段存入MongoDB

↓

爬虫抓取东方财富网页6万余篇，解析存入MongoDB，数据格式见网页端

↓

以巨潮原件为基准，匹配东方财富的文档做记录日、内容字段的修改 1597份

↓

其余清洗：调研机构名称、调研类型修改、多个参与者分离为多个调研人次...

↓

添加新字段：股票行业、文档问题个数

03

数量因子测试

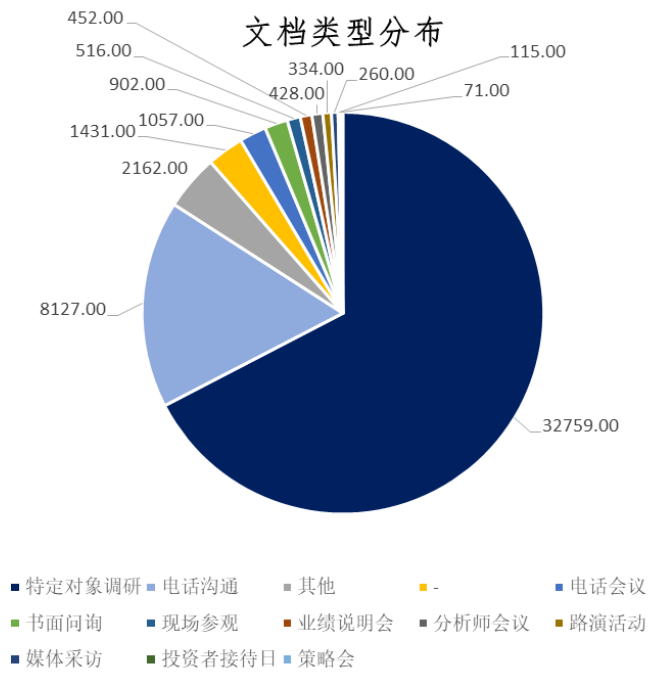
因子提取与处理

股票池选取

取调研频率在一季度一次以上的股票600只

调研文档筛选

保留需要付出调研成本的“特定对象调研”、“分析师会议”、“-”



数量因子的生成

orgNum: 当月累积参与调研的机构数量

partNum: 当月累积参与调研的人次数量（一家机构可能派出多人）

plainNum: 当月累积发生的调研次数

questionNum: 当月累积的调研采访问题数量

因子预处理

信息延迟指定天数

半衰期衰减提高覆盖率

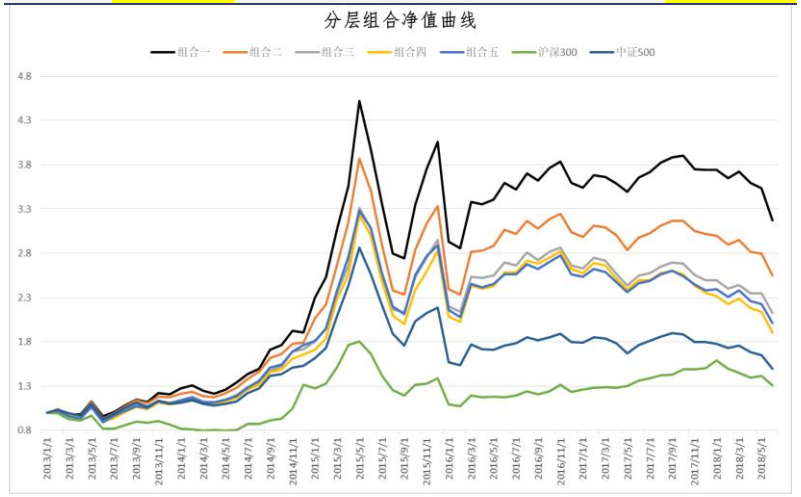
时间序列窗口标准化（剔除部分自有属性带来的差异如市值...）

横截面极值处理（三倍标准差缩极值）

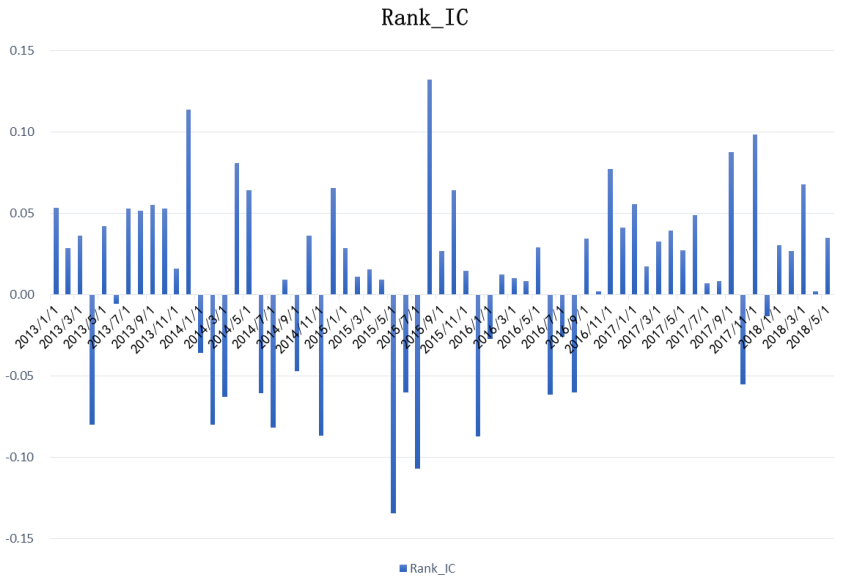
横截面标准化

■ partNum调研人次因子回归测试

T值	|T|均值	|T|>2占比	因子收益率均值	因子收益率序列T值	市值因子|T|均值
	1.1207	16.92%	0.13%	0.1737	3.6405
Rank_IC值	**Rank_IC均值**	**Rank_IC标准差**	**Rank_IC>0占比**	**abs(RankIC)>0.02占比**	**IC_IR**
	0.0106	0.0558	70.77%	75.38%	0.191



■ 调研人次因子IC序列



行业中性分层回测
201301~201806

- 回归结果来看，年化IC_IR=0.67并不高，IC的T值为1.54并不显著；但IC有较稳定的正向收益；
- 分层回测来看，头部分层明显，尾部较弱；组合一回溯期间较稳定跑赢其他组合，稳定跑赢基准
- 回测较粗糙，收益会偏高；更精细的回测参考米筐日度回测结果

买卖双方调研	年化收益率	年化波动率	夏普比率	最大回撤	超额收益年化收益率	超额收益年化波动率	信息比率	超额收益胜率	超额收益最大回撤
组合一	23.34%	32.04%	0.7286	39.22%	15.56%	9.91%	1.5700	66.67%	5.71%
组合二	18.52%	31.13%	0.5948	39.77%	10.98%	7.69%	1.4279	65.15%	7.54%
组合三	14.83%	29.01%	0.5114	35.95%	6.82%	7.38%	0.9248	57.58%	10.25%
组合四	12.33%	30.28%	0.4073	40.96%	4.81%	7.73%	0.6223	54.55%	17.02%
组合五	13.53%	28.83%	0.4694	38.67%	5.62%	6.19%	0.9076	57.58%	10.70%
中证500	7.64%	27.46%	0.2783	47.65%					
沪深300	4.98%	24.55%	0.2030	68.23%					

年化收益率	2013年	2014年	2015年	2016年	2017年	2018年前半年
组合一	21.10%	57.28%	112.78%	-11.41%	4.06%	-27.98%
组合二	17.24%	52.94%	85.76%	-8.79%	-0.58%	-28.96%
组合三	12.22%	54.19%	71.77%	-9.74%	-6.28%	-27.56%
组合四	9.07%	51.11%	70.82%	-7.49%	-10.09%	-34.46%
组合五	9.84%	60.39%	63.94%	-11.24%	-7.14%	-28.72%
中证500	9.24%	35.53%	39.22%	-16.53%	-0.19%	-26.63%

米筐日度回测	年化收益率	年化波动率	夏普比率
组合一	19.30%	30.30%	0.6290
组合二	15.70%	28.60%	0.5410
组合三	12.70%	27.70%	0.4570
组合四	15.90%	28.80%	0.5450
组合五	15.10%	27.50%	0.5350

最优组合在不同行业上的收益

	年化收益率	行业超额收益	股票池数量	夏普比率
有色金属	23.48%	25.20%	19	0.51
计算机	40.59%	21.37%	69	0.87
传媒	28.94%	19.53%	22	0.57
化工	25.19%	17.34%	53	0.75
建筑材料	21.80%	16.00%	9	0.56
国防军工	18.52%	15.35%	7	0.38
银行	17.47%	14.32%	1	0.57
钢铁	18.61%	12.61%	7	0.42
轻工制造	23.18%	12.52%	25	0.63
综合	15.80%	12.11%	1	0.29
商业贸易	13.34%	8.10%	15	0.37
电气设备	14.67%	7.19%	38	0.37
农林牧渔	13.48%	7.04%	15	0.28
电子	19.63%	4.33%	59	0.59
汽车	13.23%	4.19%	19	0.47
建筑装饰	9.62%	4.18%	24	0.27
家用电器	25.18%	3.44%	23	0.81
机械设备	6.86%	2.93%	50	0.22
医药生物	15.23%	1.00%	38	0.45
房地产	6.44%	0.16%	15	0.17
交通运输	6.56%	-1.43%	5	0.16
纺织服装	3.79%	-3.76%	11	0.10
休闲服务	12.64%	-4.19%	9	0.33
公用事业	-1.37%	-5.10%	16	-0.04
通信	5.19%	-6.60%	25	0.14
食品饮料	10.02%	-7.04%	19	0.32
非银金融	-5.32%	-10.21%	3	-0.10
采掘	-25.17%	-16.60%	3	-0.66

■ 所有机构调研人次

买卖双方调研	年化收益率	年化波动率	夏普比率	最大回撤	额收益年化收益	额收益年化波动	信息比率
组合一	23.34%	32.04%	0.7286	39.22%	15.56%	9.91%	1.5700
组合二	18.52%	31.13%	0.5948	39.77%	10.98%	7.69%	1.4279
组合三	14.83%	29.01%	0.5114	35.95%	6.82%	7.38%	0.9248
组合四	12.33%	30.28%	0.4073	40.96%	4.81%	7.73%	0.6223
组合五	13.53%	28.83%	0.4694	38.67%	5.62%	6.19%	0.9076

■ 仅考察买方调研人次

买方	年化收益率	年化波动率	夏普比率	最大回撤	额收益年化收益	额收益年化波动	信息比率
组合一	21.91%	31.70%	0.6914	39.28%	14.13%	9.71%	1.4558
组合二	20.93%	32.08%	0.6526	40.13%	13.36%	9.34%	1.4304
组合三	19.81%	30.10%	0.6579	36.54%	11.75%	7.98%	1.4727
组合四	11.61%	29.92%	0.3882	42.70%	4.14%	6.69%	0.6191
组合五	14.05%	28.79%	0.4878	36.39%	6.00%	7.02%	0.8548

■ 仅考察卖方调研人次

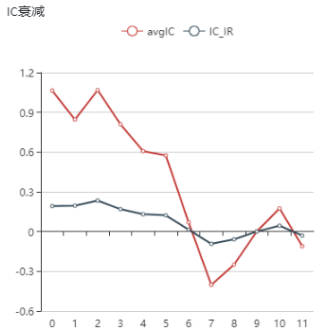
卖方调研	年化收益率	年化波动率	夏普比率	最大回撤	额收益年化收益	额收益年化波动	信息比率
组合一	21.56%	31.39%	0.6870	39.60%	13.83%	8.23%	1.6807
组合二	17.24%	30.22%	0.5707	38.87%	9.53%	7.02%	1.3582
组合三	18.41%	30.90%	0.5959	38.14%	10.65%	8.75%	1.2178
组合四	14.21%	29.92%	0.4748	42.36%	6.53%	6.87%	0.9501
组合五	11.17%	28.55%	0.3912	41.22%	3.29%	6.86%	0.4796

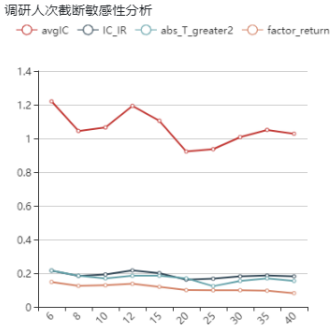
结论：

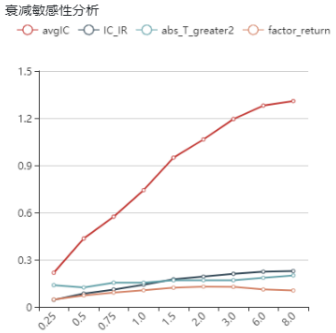
- 在有色金属、计算机、传媒、化工行业众多股票中，调研人次因子有稳健超额收益；电子与机械设备超额收益相对较少；
- 买方调研人次整体收益更高，卖方调研人次分层效果更明显，但并无显著提升，推荐使用更稳健的所有机构调研人次因子

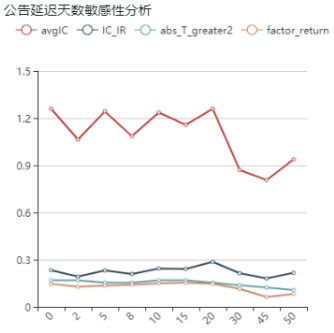
敏感性分析

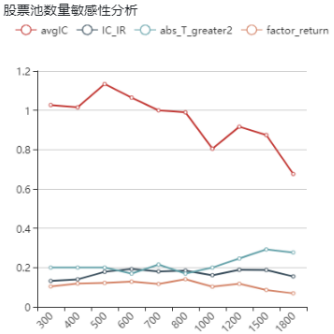
■ 回归测试各参数敏感性分析



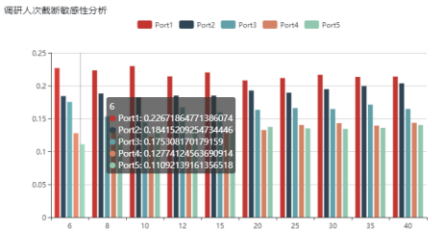


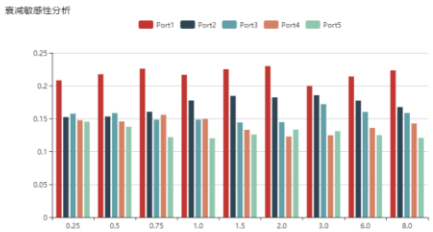


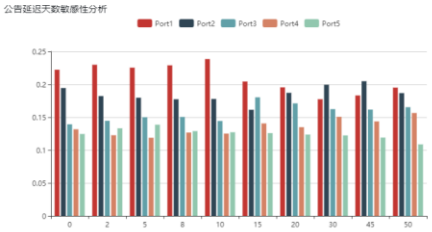


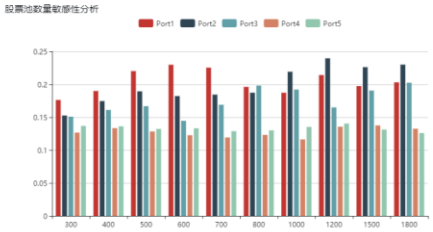


■ 分层回测参数敏感性分析









结论：

- IC在延期两个月后，开始快速衰减
- 调研人次截断在10左右，信息保留度最佳
- 尽管预处理中因子衰减半衰期越长，IC越高（在6左右开始平缓），但对分层组合的收益影响不是特别大。考虑到现实意义，取半衰期=2个月
- 调研信息发布延迟天数并不敏感，但在延迟20天后发布，IC显著下降。
- 股票池选取对因子分层效果有显著影响，越多调研频率少的股票纳入股票池，因子有效性越低。可能因为这些股票调研数据稀缺，调研人次因子解释力低，其它风险因子干扰了分层效果。就如一致预期评级上调因子只对分析师覆盖较多的股票有意义。
- 股票池选取800只以上后，分层变弱；但次优组合及前三组合均稳健优于尾部组合。

其余因子回测

调研人次数量：partNum

Rank_IC值	Rank_IC均值	Rank_IC标准差	Rank_IC>0占比;(RankIC)>0.02		IC_IR
	0.0106	0.0558	70.77%	75.38%	0.191

买卖双方调研	年化收益率	年化波动率	夏普比率	最大回撤	额收益年化收益	额收益年化波动
组合一	23.34%	32.04%	0.7286	39.22%	15.56%	9.91%
组合二	18.52%	31.13%	0.5948	39.77%	10.98%	7.69%
组合三	14.83%	29.01%	0.5114	35.95%	6.82%	7.38%
组合四	12.33%	30.28%	0.4073	40.96%	4.81%	7.73%
组合五	13.53%	28.83%	0.4694	38.67%	5.62%	6.19%

调研次数数量：plainNum

Rank_IC值	Rank_IC均值	Rank_IC标准差	Rank_IC>0占比;(RankIC)>0.02		IC_IR
	0.0146	0.0502	69.23%	73.85%	0.291

买卖双方调研	年化收益率	年化波动率	夏普比率	最大回撤	额收益年化收益	额收益年化波动
组合一	21.99%	31.66%	0.6946	39.08%	14.21%	9.39%
组合二	19.22%	31.26%	0.6151	40.87%	11.69%	7.61%
组合三	15.84%	29.57%	0.5355	36.43%	8.00%	6.55%
组合四	12.11%	28.93%	0.4186	39.31%	4.32%	6.23%
组合五	12.56%	28.52%	0.4405	38.54%	4.63%	6.82%

调研机构数量：orgNum

Rank_IC值	Rank_IC均值	Rank_IC标准差	Rank_IC>0占比;(RankIC)>0.02		IC_IR
	0.0126	0.0546	66.15%	80.00%	0.231

买卖双方调研	年化收益率	年化波动率	夏普比率	最大回撤	额收益年化收益	额收益年化波动
组合一	22.67%	32.30%	0.7021	39.30%	15.01%	10.00%
组合二	18.26%	30.92%	0.5907	40.91%	10.68%	7.61%
组合三	17.81%	29.13%	0.6115	33.95%	9.67%	7.10%
组合四	13.01%	30.39%	0.4280	40.68%	5.43%	8.13%
组合五	11.11%	28.33%	0.3923	41.71%	3.24%	5.95%

问题数量：questionNum

Rank_IC值	Rank_IC均值	Rank_IC标准差	Rank_IC>0占比;(RankIC)>0.02		IC_IR
	0.0067	0.0648	60.00%	73.85%	0.103

买卖双方调研	年化收益率	年化波动率	夏普比率	最大回撤	额收益年化收益	额收益年化波动
组合一	18.74%	31.38%	0.5973	37.69%	11.08%	9.37%
组合二	16.77%	30.18%	0.5558	36.87%	9.01%	7.71%
组合三	18.77%	29.63%	0.6334	35.68%	10.70%	7.36%
组合四	16.22%	29.84%	0.5437	39.46%	8.43%	6.98%
组合五	10.85%	29.11%	0.3727	45.14%	3.20%	6.47%

- 数量因子间相关性分析

	partNum	orgNum	uestionNur	plainNum
partNum	1.000	0.975	0.704	0.856
OrgNum	0.975	1.000	0.726	0.896
QuestionNum	0.704	0.726	1.000	0.770
plainNum	0.856	0.896	0.770	1.000

- 与传统单因子测试对比

	orgNum	BP	反转	低换手率
组合一	22.67%	22.07%	23.15%	20.42%
组合二	18.26%	21.09%	17.04%	18.93%
组合三	17.81%	16.55%	15.84%	14.24%
组合四	13.01%	12.44%	11.16%	16.08%
组合五	11.11%	8.87%	7.67%	7.72%

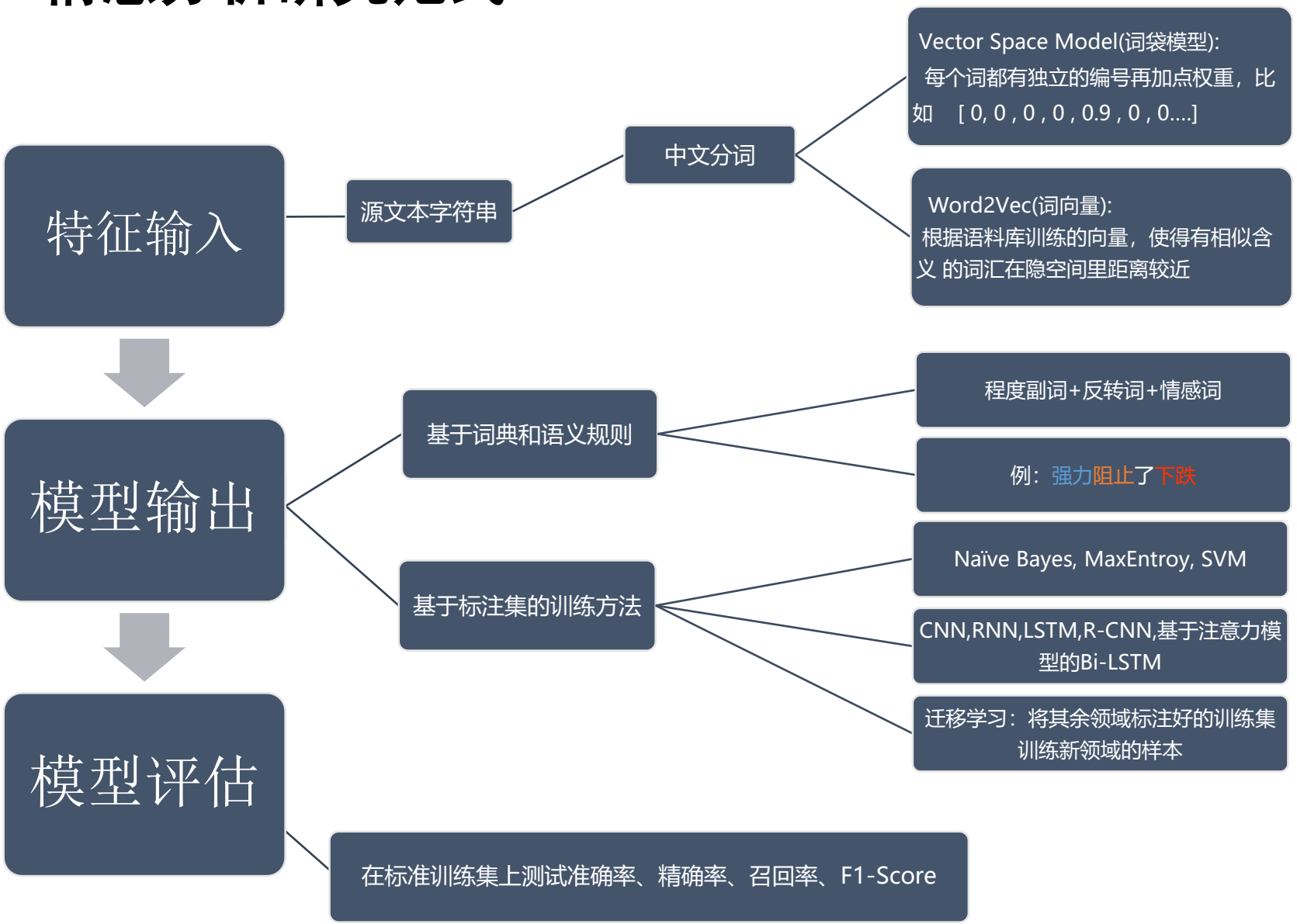
结论：

- 数量因子间相关度都很高，表现类似；但问题数量因子由于解析噪音较多，表现最差
- 调研人次因子最优组合的收益最高，调研次数因子的IC最稳定IC_IR最高，调研机构数量因子单调性最好
- 数量因子中推荐使用调研机构数量因子
- 调研机构因子并不弱于传统因子！BP因子换手低，表现最优

04

文本情感信息提取

情感分析研究范式


特征输入
源文本字符串
中文分词
Vector Space Model(词袋模型):
每个词都有独立的编号再加点权重，比如 [0, 0 , 0 , 0 , 0.9 , 0 , 0....]
Word2Vec(词向量):
根据语料库训练的向量，使得有相似含义 的词汇在隐空间里距离较近
模型输出
基于词典和语义规则
程度副词+反转词+情感词
例：强力阻止了下跌
基于标注集的训练方法
Naïve Bayes, MaxEntroy, SVM
CNN,RNN,LSTM,R-CNN,基于注意力模型的Bi-LSTM
迁移学习：将其余领域标注好的训练集训练新领域的样本
模型评估
在标准训练集上测试准确率、精确率、召回率、F1-Score

研究现状

- 银河证券(201412)，根据分析师研报摘要，依据情感热词提取情绪
- 兴业证券(201710)，使用SVM构建雪球网帖子情感打分系统
- 国外关于earnings conference calls情绪挖掘，多数基于词典；关于Twitter类社交媒体文本，多数也用的SVM

打分流程

中文分词，去常用停词；对每个问题进行打分

↓

遍历句子词汇，遇到情感词则暂停

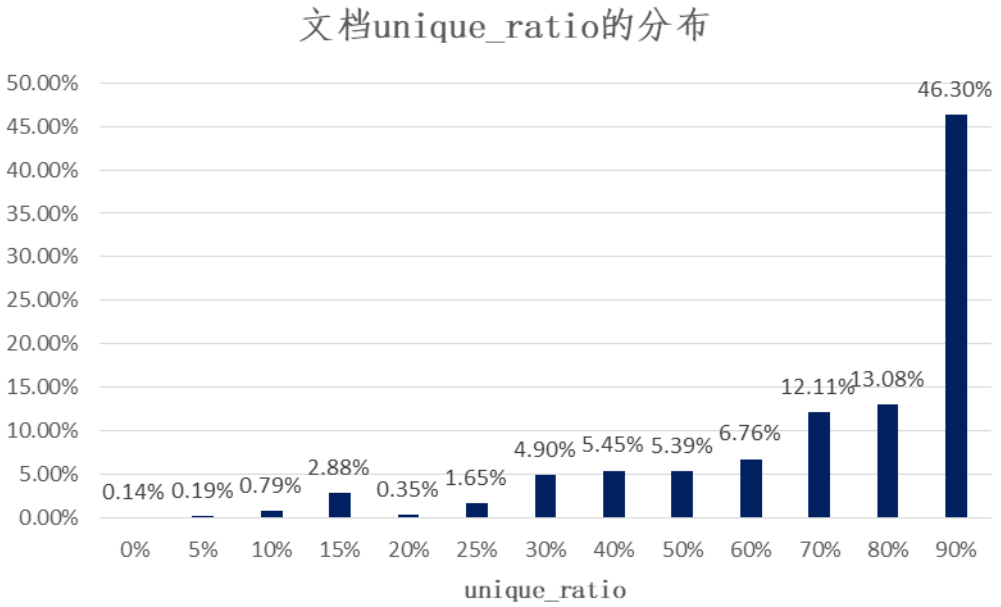
↓

前向搜索：如遇否定词进行极性反转；如遇程度词，进行加权

↓

每个问题打分后，得到调研记录平均分数

文档重合度



- 缺乏好的情感词典，情感打分性能较差
- 调研记录的重合度比较高，文本上的信息增益有待商榷

05

总结与后续展望

研究结论

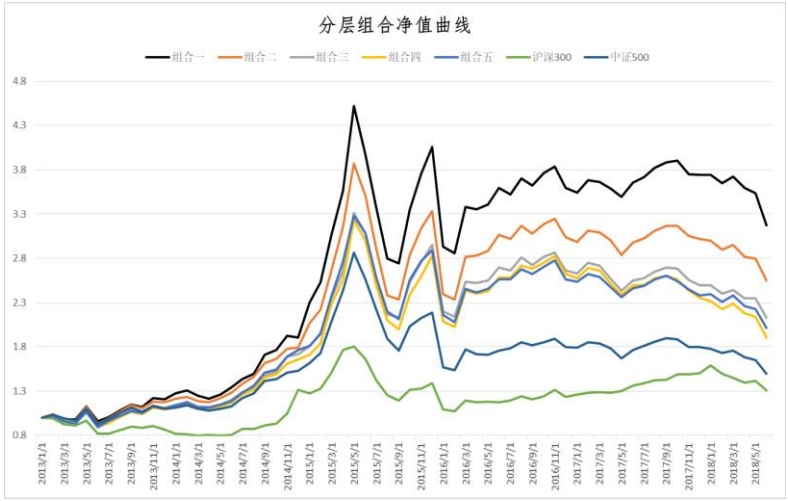
■ 机构调研的数量信息有超额收益，回归测试有较为稳定正的IC，分层回测中分层效果明显。

Rank_IC值	Rank_IC均值	Rank_IC标准差	Rank_IC>0占比	(RankIC)>0.02占比	IC_IR
	0.0126	0.0546	66.15%	80.00%	0.231

买卖双方调研	年化收益率	年化波动率	夏普比率	最大回撤	额收益年化收益	额收益年化波动
组合一	22.67%	32.30%	0.7021	39.30%	15.01%	10.00%
组合二	18.26%	30.92%	0.5907	40.91%	10.68%	7.61%
组合三	17.81%	29.13%	0.6115	33.95%	9.67%	7.10%
组合四	13.01%	30.39%	0.4280	40.68%	5.43%	8.13%
组合五	11.11%	28.33%	0.3923	41.71%	3.24%	5.95%

对比传统因子

	orgNum	BP	反转	低换手率
组合一	22.67%	22.07%	23.15%	20.42%
组合二	18.26%	21.09%	17.04%	18.93%
组合三	17.81%	16.55%	15.84%	14.24%
组合四	13.01%	12.44%	11.16%	16.08%
组合五	11.11%	8.87%	7.67%	7.72%



■ 因子覆盖度比较低，股票池越大其他风险因子噪音越高，单因子测试结果更差。但头部组合均高于尾部组合。

■ 数量因子中，调研机构累积数量效果最好；在自定义的600股票池中，选股能力逼近传统的BP、反转因子。

■ 调研内容有一定重合度，加上因子覆盖度本来就不高，文本信息噪音更大。

■ 情感分析需要高质量的情感词典或者训练语料库，需要一定的精力与资源。

■ 后续可考虑对数量信息进行细分加权（加入调研机构的权威度、资金量、内容长度等信息）

■ 高质量的文本挖掘可考虑引入数据商合作

感谢量化及ETF投资部的所有领导、同事！

谢谢聆听