中信建投证券 CHINA SECURITIES

CHINA SECURITIES RESEARCH

证券研究报告·金融工程深度报告

新闻情绪选股的多空差策略：

--大数据研究之三

金融工程研究

重要观点

数据挖掘概述

数据挖掘基本步骤包括数据采集、数据预处理、数据存储、数据分析、知识发现。数据挖掘常见技术包括监督学习（即分类分析）、无监督学习（即聚类分析）、关联分析、预测分析等。

新闻情绪选股多空差策略概述

新闻情绪选股多空差策略即把 N 日正负面新闻权重和构成的当日新闻情绪指数相加，然后进行排序，选取排名前 topN 的个股作为多头组合或空头组合，选取排名倒数前 topN 的个股作为空头组合或者多头组合的多空差策略。

新闻情绪因子选股不同板块风格各异

主板方面，情绪指数与股票未来表现为正向指标，且持续天数大概 5 个交易日；中小板方面，情绪指数与股票未来表现为反向指标，且持续天数大概 22 个交易日；创业板块方面，情绪指数与股票未来上涨或下跌效果不显著。

主板空头负超额收益十分显著

2014 年 1 月 1 日到 2016 年 12 月 30 日多头组合相对沪深 300 指数最终值为 1.15，多头组合相对沪深 300 年化超额收益 4.6%，而空头组合相对沪深 300 指数最终值为 0.33，空头组合相对沪深 300 年化负超额收益达 30.47%。

多头动量效应明显比空头强劲

新闻情绪选股多空差策略在沪深 300 成份股中，多头组合股票持有 5 天总换手率为 62.4，每次换手率平均值为 0.43；而空头组合中，总换手率为 114.6，每次换手率平均值为 0.8。

最优投资策略

新闻情绪选股多空差模型中，最优投资策略为：以沪深 300 成份股作为候选标的池，负面新闻影响与正面新闻等权，以 5 个交易日为周期，选取 5 个交易日情绪指数和排名前 5 的股票作为多头，选取排名倒数前 5 的作为空头，并持有 5 个交易日。

该策略的组合均剔除掉交易日一字板涨跌停和停牌的股票。该策略多头组合年化收益率为 18.78%，多空收益差年化为 50.44%，多空收益差最大回撤为 27.57%。

丁鲁明

dingluming@csc.com.cn

021-68821623

执业证书编号：S1440515020001

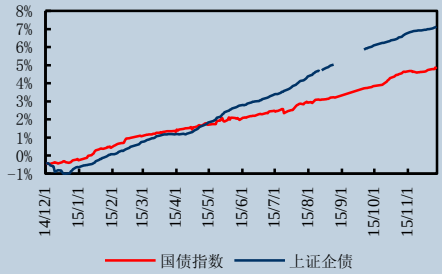
研究助理：喻银尤

yuyinyou@csc.com.cn

021-68821600-808

发布日期： 2017 年 03 月 08 日

市场表现



相关研究报告

17.03.02	大数据研究之指标构建：机器学习之贝叶斯文本分类算法的实现
17.02.27	大数据周报：新闻热度未见突破，大盘继续谨慎；新闻情绪选股空头组合相对沪深 300 指数超额收益为-0.32%
17.02.21	大数据周报：大盘维持谨慎，新闻情绪选股多头组合跑赢沪深 300 达 0.9%
17.02.13	大数据周报：大盘相对谨慎，重点关注沪深 300 新闻情绪选股多空组合
16.10.12	大数据研究之择时：基于新闻热度的多空策略
16.09.14	股票行业配置——基于投资时钟理论
16.09.13	基于残差分析的大类资产轮动策略
16.08.09	基本面量化系列之六——“量化基本面”理论体系及通信行业案例
16.06.24	基本面量化系列之五——投资时钟指路，量化大类资产轮动破局

目录

图形目录

表格目录

一、 大数据与量化投资

IBM 最早定义大数据的 5V 特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）、Veracity（真实性）。当今社会，大数据所带来的信息风暴正在深刻的影响着我们的生活、工作和思维，大数据将开启一次重大的时代转型。

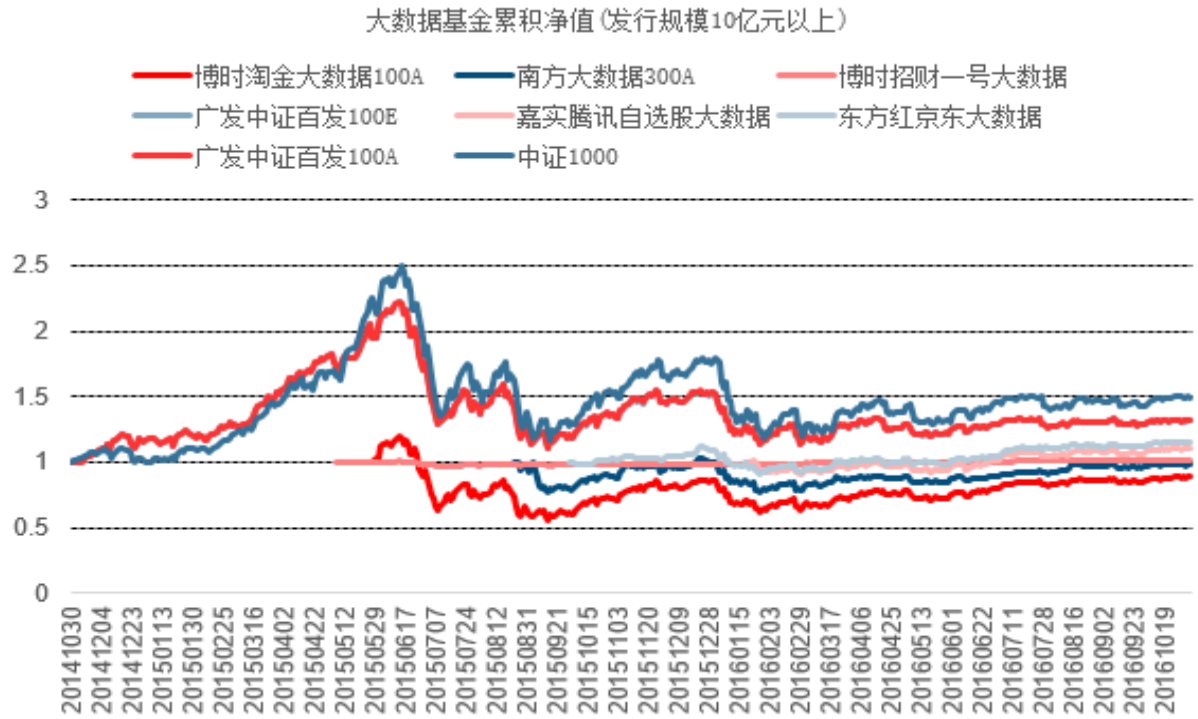
传统量化投资主要包括量化选股、量化择时、股指期货套利、商品期货套利、统计套利、算法交易，资产配置，风险控制等。传统的量化投资研究的数据来源一般是公司的财务指标、交易行情数据、政策宏观方面的投资信息等。而随着量化投资这一领域的快速发展，这些传统数据中所包括的大部分投资信息已经被专业投资者所挖掘，想要从这些信息中获取收益难度将越来越大。大数据将为量化投资这一领域创造前所未有的可量化的新的维度，为量化投资提供了新的研究视野。如何把大数据这一金矿从数据转变为知识则充满挑战和困难，大数据将驱动量化投资的创新。

表 1：市场上大数据基金列表(不完全统计)

基金简称	基金公司	合作方	成立时间	大数据因子	产品类型
银河定投宝	银河基金	腾讯财经	2014.3.14	--	指数型
广发中证百发 100A	广发基金	百度	2014.10.30	百度搜索因子指标	指数型
广发中证百发 100E	广发基金	百度	2014.10.30	百度搜索因子指标	指数型
广发资管互联网+	广发资管	新浪网	2015.4.10	--	集合资产管理计划
南方大数据 100	南方基金	新浪	2015.4.24	个股访问热度及新闻正负面	指数型
博时招财一号大数据	博时基金	蚂蚁金服	2015.4.29	用户行为，行业成长，价格变化等	偏债混合型
博时淘金大数据 100A	博时基金	蚂蚁金服	2015.5.4	用户行为、行业成长、价格变化等因素	指数型
博时淘金大数据 100I	博时基金	蚂蚁金服	2015.5.4	用户行为、行业成长、价格变化等因素	指数型
南方大数据 300A	南方基金	新浪	2015.6.24	个股访问热度及新闻正负面	指数型
南方大数据 300C	南方基金	新浪	2015.6.24	个股访问热度及新闻正负面	指数型
东方红京东大数据	东方资管	京东	2015.7.31	京东电商的销量、浏览量、点击量、客户评价、客户收藏量等基础数据	混合型
广发百发大数据 A	广发基金	百度	2015.9.14	百度搜索因子指标	混合型
广发百发大数据 E	广发基金	百度	2015.9.14	百度搜索因子指标	混合型
广发百发大数据策略成长 A	广发基金	百度	2015.11.18	百度搜索因子指标	混合型
广发百发大数据策略成长 E	广发基金	百度	2015.11.18	百度搜索因子指标	混合型
嘉实腾讯自选股大数据	嘉实基金	腾讯	2015.12.7	用户行为数据	股票型
海富通东财大数据	海富通基金	东方财富	2016.1.29	股票关注度、点击量等投资者行为数据	混合型
大成互联网+大数据	大成基金	360	2016.2.3	360 用户搜索行为	指数型
泰达宏利同顺大数据	泰达宏利基金	同花顺	2016.2.23	网络点击量、新闻发布量、新闻点击量、股吧讨论量等	灵活配置型
银华大数据	银华基金	--	2016.4.7	股票新闻点击率、股票行情浏览量、分析师推荐评级等	指数型
博时银智大数据 100	博时基金	银联	2016.5.20	银联刷卡数据	指数型

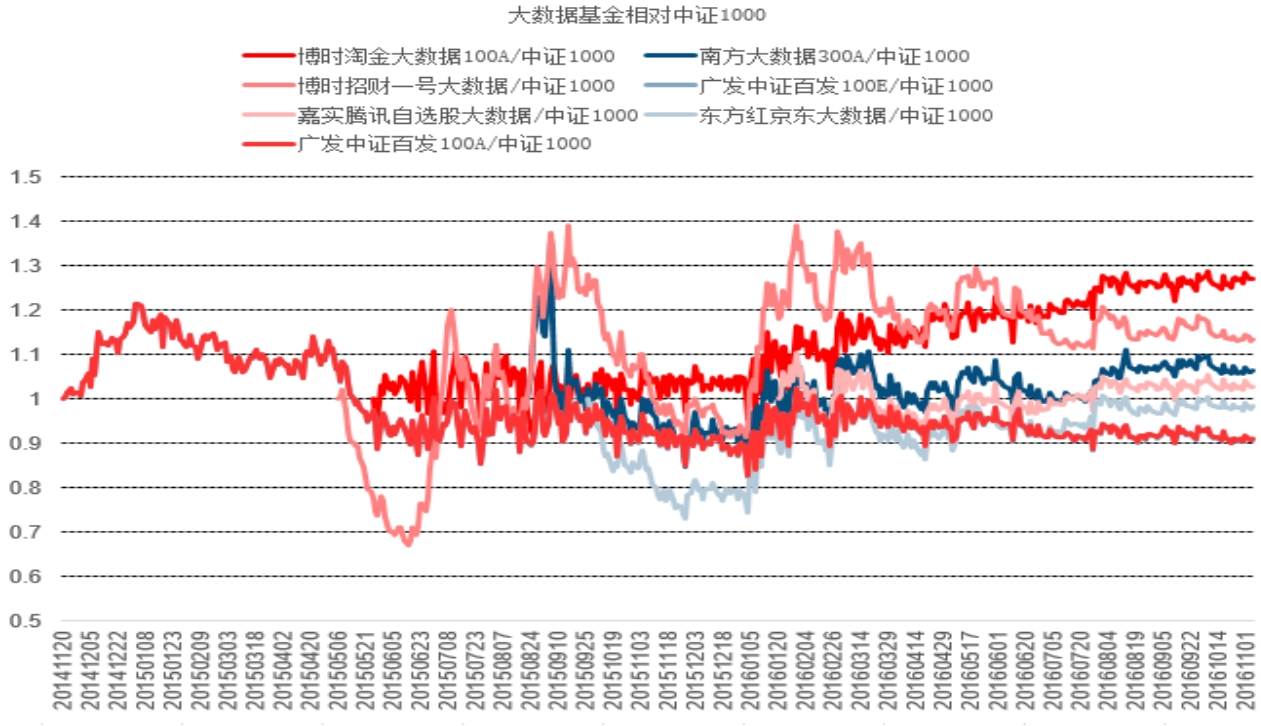
数据来源：wind 资讯，中信建投证券研究发展部

图 1：大数据基金累积净值(发行规模 10 亿元以上)



数据来源：wind 资讯，中信建投证券研究发展部

图 2：大数据基金相对中证 1000(发行规模 10 亿元以上)



数据来源：wind 资讯，中信建投证券研究发展部

二、 大数据体系构建

在大数据时代背景下，完善大数据体系构建是一个长期的、持续的、迭代的过程，其基本过程主要包括以下几个步骤：

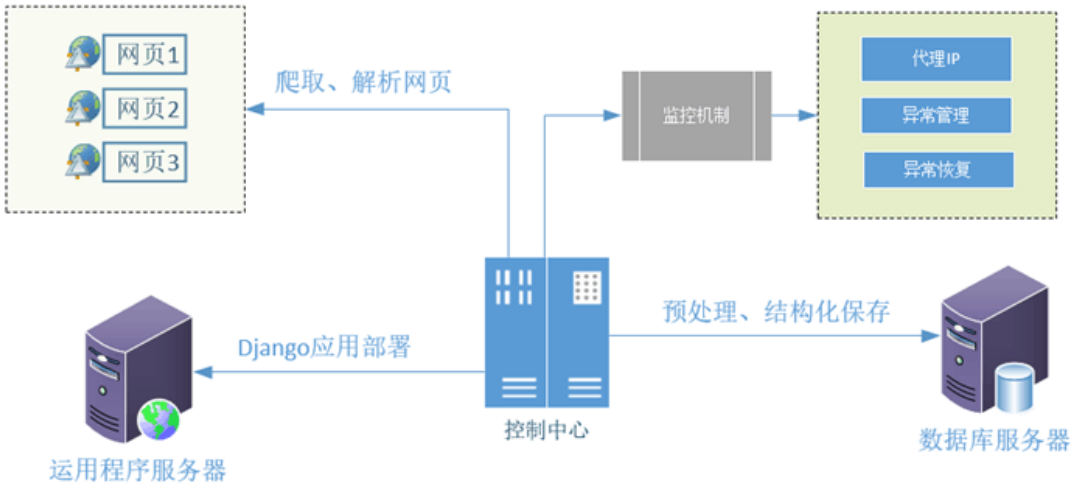
2.1 数据采集与预处理

大数据的源头质量，直接决定我们指标质量，决定着我们的策略优劣性。目前，国内的相关数据来源主要为第一类上交所，深交所等的公告、财报，监管信息等；第二类财经新闻网站，比如新浪财经，第一财经，东方财富网，中国证券网，金融界，雪球财经，腾讯财经，第一财经等的个股新闻，行业新闻，宏观经济等；第三类社交媒体，比如股吧，贴吧，微博等；第四类为关注数据，比如百度，搜狗等个股每天搜索数量及分析师研报提及个股等。我们目前数据主要爬取新浪财经个股相关新闻，包括 200 多家媒体在内的所有个股新闻。

大数据采集则是通过网络爬虫或网站公开 API 等方式从上述相关网站上获取我们所需要的数据信息，将非结构化数据从网页中爬取下来，并解析相关信息，将其存储为统一的本地数据文件，并以结构化的方式存储在我们的数据库中。

我们的数据采集主要包括爬取网页组件、监控组件、控制中心、应用服务器及数据库等。其框架体系图如下：

图 3：中信建投金融工程爬虫系统框架体系图



数据来源：中信建投证券研究发展部

数据预处理指直接从网页爬取的数据并不能直接用于使用，而是需要经过一定的预处理，以保证数据质量和数据安全。因为在大数据应用中，数据来源非常广泛，数据质量良莠不齐，更需要预处理过程。数据预处理主要是去除无法解析的错误网页，删除重复的数据，去除无效的数据等；将不同的数据源爬取到的数据统一存储，建立数据仓库。

2.2 大数据存储技术

我们使用 mysql 存储数据，从 2014 年 1 月 1 号到 2016 年 9 月 26 日，已经有 200 多万条个股新闻数据，共 45g 多，虽然现在不算超级大数据，但随着我们系统的逐渐完善，数据来源的多样化，数据存储一定会成为较大的瓶颈。为了满足大数据访问的效率与要求，大数据处理需要合理地存储与组织各种数据，以减少网络和存储 I/O 开销，提升系统性能； mysql 大数据存储目前我们主要是采用分表和分区技术。

分表技术包括垂直分表：即一个表字段数量控制在一种范围，过多的话应该适当拆分成几个表。在设计阶段就应该考虑好数据库表字段。分表技术还包括水平分表即把数据过多的表拆分成多个表存储。分表后，逻辑上也已经是不同的子表，操作时，要指定子表操作。

分区将表分离在若干不同的表空间上，即把一个大表分割成若干个小表，分区逻辑上还是一个表，实际物理存储成多个数据文件，用来支撑无限膨胀的大表，给大表在物理一级的可管理性。将大表分割成较小的分区可以改善表的维护、备份、恢复、事务及查询性能。目前分区主要包括 1.RANGE 分区：基于属于一个给定连续区间的列值，把多行分配给分区。2.LIST 分区：类似于按 RANGE 分区，区别在于 LIST 分区是基于列值匹配一个离散值集合中的某个值来进行选择。3.HASH 分区：基于用户定义的表达式的返回值来进行选择的分区。4..KEY 分区：类似于按 HASH 分区，区别在于 KEY 分区只支持计算一列或多列，且 MySQL 服务器提供其自身的哈希函数。必须有一列或多列包含>整数值。

以上技术应用于小型大数据还可以完美解决，但是超级大型数据则无能为力。目前有以下几种典型的大数据存储技术解决方案，第一种采用 MPP 架构的新型数据库集群，重点面向行业大数据，采用 Shared Nothing 架构，通过列存储、粗粒度索引等多项大数据处理技术，再结合 MPP 架构高效的分布式计算模式，具有高性能和高扩展性的特点，在企业分析类应用领域获得极其广泛的应用。第二种是基于 Hadoop 的技术扩展和封装，围绕 Hadoop 衍生出相关的大数据技术，应对传统关系型数据库较难处理的数据和场景。第三种是大数据一体机，这是一种专为大数据的分析处理而设计的软、硬件结合的产品，由一组集成的服务器、存储设备、操作系统、数据库管理系统以及为数据查询、处理、分析用途而特别预先安装及优化的软件组成，高性能大数据一体机具有良好的稳定性和纵向扩展性。

2.3 数据分析与指标构建

通过市场情绪分析、财经文本分析、新闻热点捕捉、主题挖掘等从这些大量的新闻中挖掘出有效信息。利用数据挖掘技术，即利用各种方法分析我们需要处理的数据，发现隐藏在海量数据背后的知识和规律。挖掘步骤简单的可以概括为 a.前期数据的准备 b.从这些数据中寻找他们的规律 c.把寻找到的规律表示出来，这 3 个步骤。前期数据的准备是从这些相关的数据源中以一定的规则挑选我们所需的数据，然后整合成我们用于数据挖掘的数据集；寻找这些数据的规律是利用数据挖掘相关的方法将这些数据集所含的规律挖掘出来；把寻找到的规律表示出来是利用比如图表等可视化的技术尽可能以用户可以理解的方式展示出来挖掘出来的规律。

数据挖掘常用的几种方法为：分类分析、聚类分析、关联分析、预测分析、异常分析等等。分类分析是首先从已有的数据中选出已有的分类，且把所有的没有分类的要进行分类的数据按照这些已规定好类别分别进行分类。聚类分类不属于预测性的问题，该算法主要解决的是把一群给定的对象划分成若干个组的问题。划分样本的依据是聚类问题的核心点。聚类分析主要是解决当要分析的数据缺乏描述信息或者是无法组织成任何分类

模式时用于样本的聚类分析。关联分析中主要技术是对象相关度或者他们之间的关系。预测分析主要包括一无线性回归，多元线性回归，Markov 预测模型等。

三、 新闻情绪选股原理

在互联网大数据时代，我们获取信息的途径更加广泛和便捷，我们不再局限于传统的基本面数据和技术面数据。如何利用这些信息则是非常具有挑战性的问题。

资本市场变幻莫测，至今没有任何理论能完全解释并预测股票未来的趋势。互联网大数据的到来，则为我们提供了新思路，每只个股，几乎每天都有相关新闻，我们从新闻正面或者新闻负面对股票的影响进行研究，构建个股情绪因子，来获取超额收益。

3.1 情绪因子构建

新闻情绪指数因子构建即先求出正负面新闻权重和构成的当日新闻情绪指数，然后再把 N 日指数进行相加，得到当期的新闻情绪指数因子。（具体新闻分类方法及当天新闻情绪指数构建可查看深度报告《大数据研究之指标构建：机器学习之贝叶斯文本分类算法的实现》）。

3.2 选股策略原理

该策略类似于多因子选股策略，只是这个策略中，只有 N 日正负面新闻权重和构成的当日新闻情绪指数相加这一因子。

选股策略：

把 N 个交易日正负面新闻权重和构成的当日新闻情绪指数相加，然后进行排序，选取排名前 topN 作为多头组合或空头组合，选取排名倒数前 topN 作为空头组合或者多头组合。其中，多头组合与空头组合都剔除买入当天一字涨跌停和停牌股票，新上市股票一个月内也不能作为候选股，多头与空头组合 N+1 个交易以平均价买入，持有 N 个交易日以平均价卖出，若卖出当天有一字涨跌停和停牌股票，则顺延到下一个交易日以平均价卖出，并买入需要买入的股票，使多头与空头组合始终保持满仓。最后计算多空收益差。

3.3 回测结果分析

数据说明：

数据区间：2014-01-01 至 2016-12-30。

个股数据：沪深 300 指数成份股每日新闻情绪指数。

选股标的：沪深 300 指数成份股。

多头组合： 把 N 个交易日正负面新闻权重和构成的当日新闻情绪指数相加，然后进行排序，选取排名前 topN 的股票作为多头组合。

空头组合： 把 N 个交易日正负面新闻权重和构成的当日新闻情绪指数相加，然后进行排序，选取排名倒数前 topN 的股票作为空头组合。

策略参数：N，topN，w_neg（即负面新闻对股票影响程度，正面新闻默认为 1）。

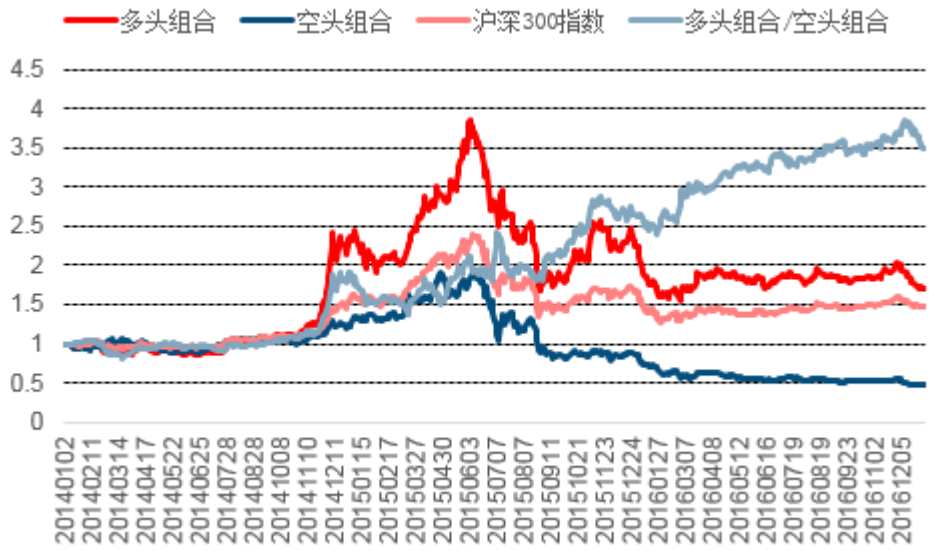
结果分析：

表 2：沪深 300 成份股情绪因子多空差策略结果统计

回测期间	2014-01-01 至 2016-12-30		
初始净值	1	最终净值	3.49
年化多空收益差	50.44%	夏普比	1.55
最大回撤	27.5%	最大回撤区间	2014-12-17 到 2015-03-18
日胜率	52.12%	5 日胜率	59.29%
多头换手率	62.4	空头换手率	114.6

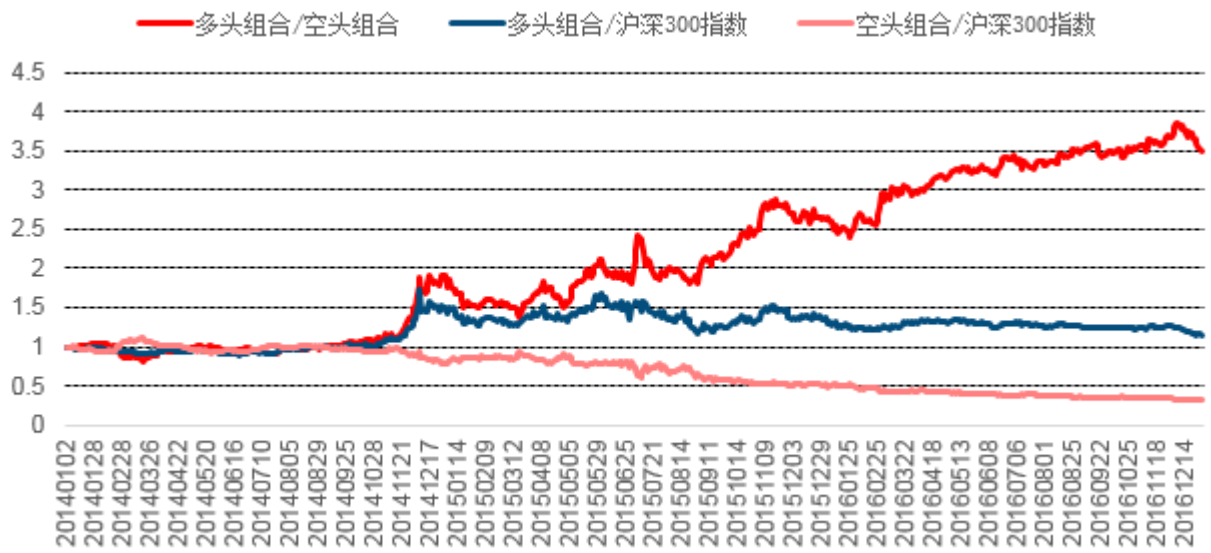
数据来源：中信建投证券研究发展部

图 4：沪深 300 成份股情绪因子多空差策略净值



数据来源：wind 资讯，中信建投证券研究发展部

图 5：沪深 300 成份股情绪因子多空组合与沪深 300 指数比较



数据来源：wind 资讯，中信建投证券研究发展部

从以上结果来看，多头组合相对沪深 300 指数最终值为 1.15，多头组合相对沪深 300 年化超额收益 4.6%，而空头组合相对沪深 300 指数最终值为 0.33，空头组合相对沪深 300 年化负超额收益达 30.47%。

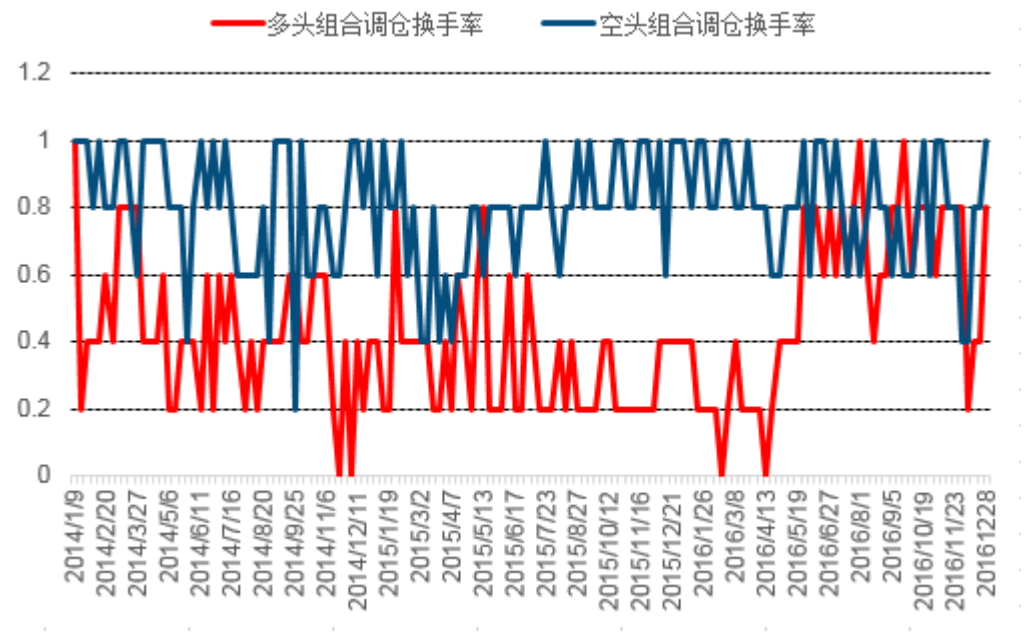
四、 情绪选股在不同板块的表现

4.1 主板因子动量显著

以沪深 300 成份股作为候选标的池，负面新闻影响与正面新闻等权，以 5 个交易日为周期，选取 5 个交易日情绪指数和排名前 5 的股票作为多头，选取排名倒数前 5 的作为空头，并持有 5 个交易日。

多头组合与空头组合每 5 个交易日调仓一次。多头组合中，换手率平均值为 43%，而空头组合中，换手率平均值为 80%。这说明，多头组合中，当新闻情绪因子排名前 5 时，下一次排名前 5 的概率也非常大，达到 57%，即情绪因子动量十分明显；而在空头组合中，换手率达到 80%，这说明当新闻情绪因子排名倒数前 5 时，下一次排名倒数前 5 时的概率则相对比较小，仅为 20%。

图 6：沪深 300 成份股情绪因子调仓换手率



数据来源：中信建投证券研究发展部

4.2 中小板反向指标

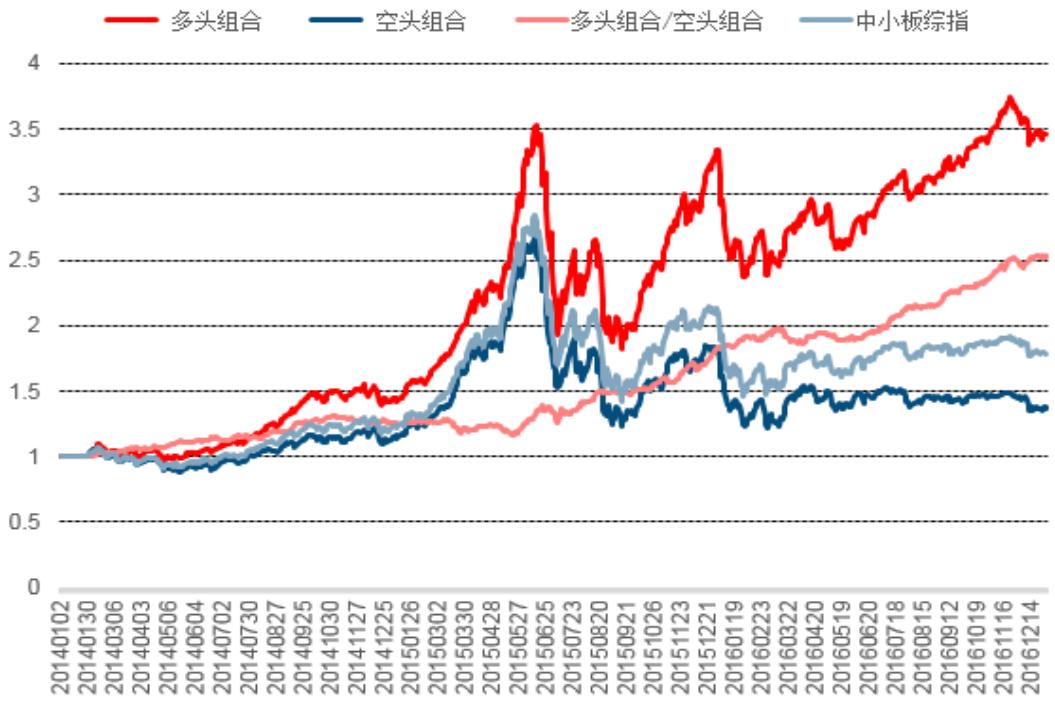
以 2014 年 1 月 1 日到 2016 年 12 月 30 日为回测期间，以中小板成份股作为候选标的池，负面新闻影响为 0.1，以 22 个交易日为周期，选取 22 个交易日情绪指数和排名前 50 的股票作为空头，选取排名倒数前 50 的作为多头，并持有 22 个交易日。

在中小板块中，组合均剔除掉交易日一字板涨跌停和停牌的股票。该策略多头年化收益率为 52.79%，年化多空收益差为 37.33%，多空收益差最大回撤为 11.17%，日胜率为 57.52%，22 日胜率高达 80.65%。

多头组合相对中小板综指最终值为 1.94，多头组合相对中小板综指年化超额收益高达 24.27%，而空头组合相对中小板综指最终值为 0.77，空头相对中小板综指年化负超额收益达 9.40%。这表明，多空收益差的收益主要来自多头组合的超额收益。

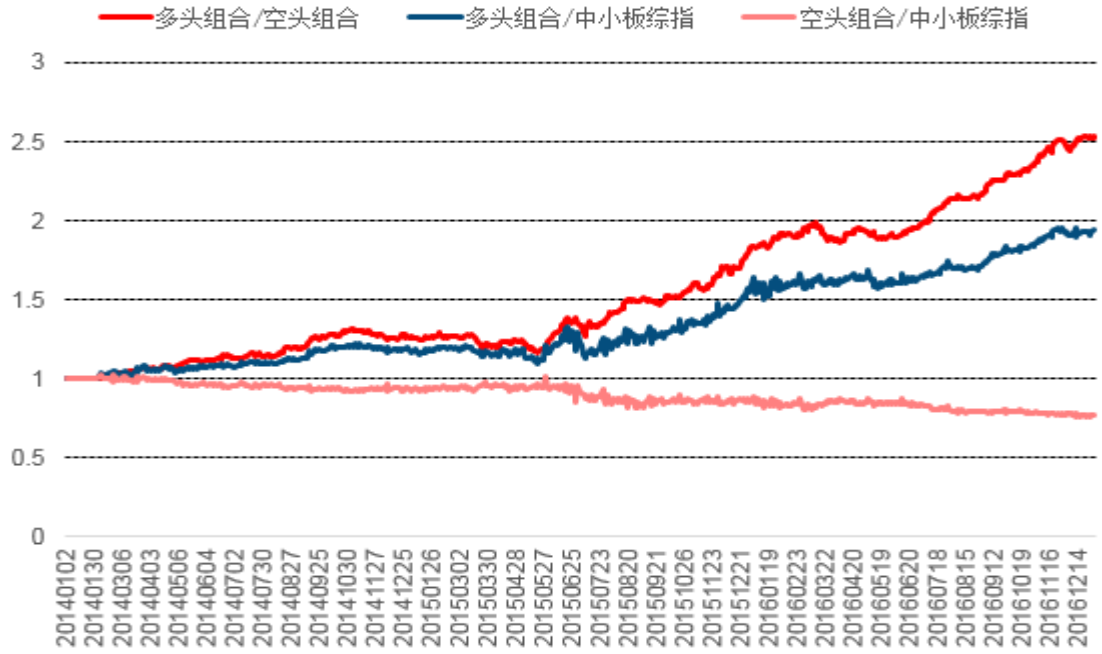
上述结果表明，新闻情绪指数在中小板块为反向指标，即当情绪指数排名靠前，其股票反而未来表现更差，而情绪指数排名倒数的股票，则有明显的超额收益。

图 7：中小板成份股情绪因子多空差策略净值



数据来源：wind 资讯，中信建投证券研究发展部

图 8：中小板成份股情绪因子多空组合与中小板综指比较



数据来源：wind 资讯，中信建投证券研究发展部

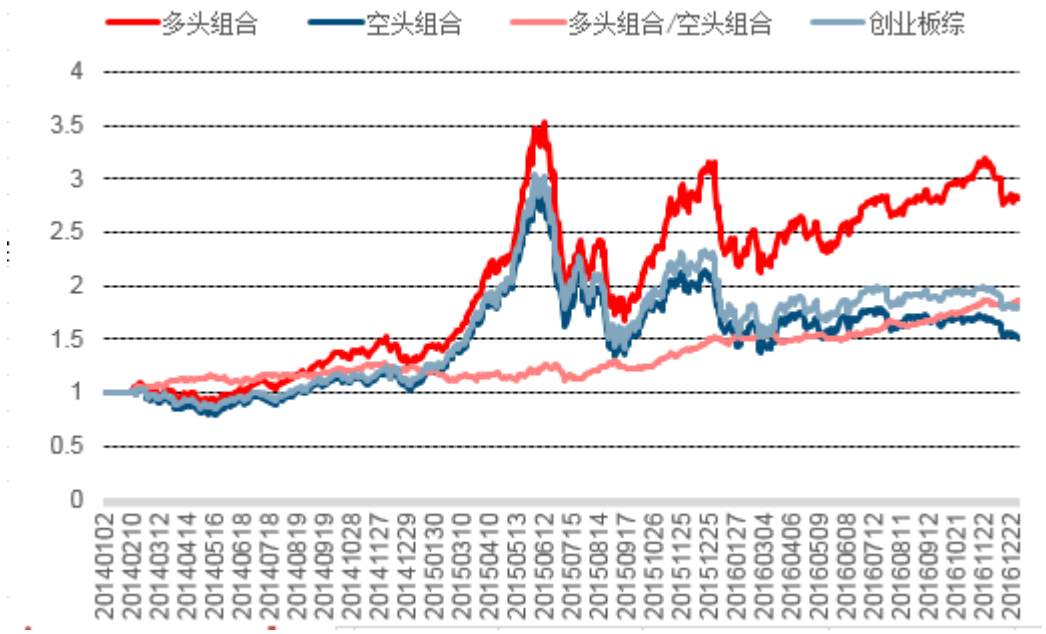
4.3 创业板效果偏弱

以 2014 年 1 月 1 日到 2016 年 12 月 30 日为回测期间，以创业板成份股作为候选标的池，负面新闻影响为 0.1，以 22 个交易日为周期，选取 22 个交易日情绪指数和排名前 60 的股票作为空头，选取排名倒数前 60 的作为多头，并持有 22 个交易日。

在创业板块中，组合均剔除掉交易日一字板涨跌停和停牌的股票。该策略多头组合净值为 2.82，年化收益率为 42.41%，多空收益差年化为 23.64%，多空收益差最大回撤为 12.79%，日胜率为 56.54%，22 日胜率高达 67.74%。

在创业板中，多空收益差为 23.64%，而同期创业板综指年化收益为 22.2%，即新闻情绪因子在创业板中效果并不明显，多空收益差与同期创业板综指差别不大。

图 9：创业板成份股情绪因子多空差策略净值

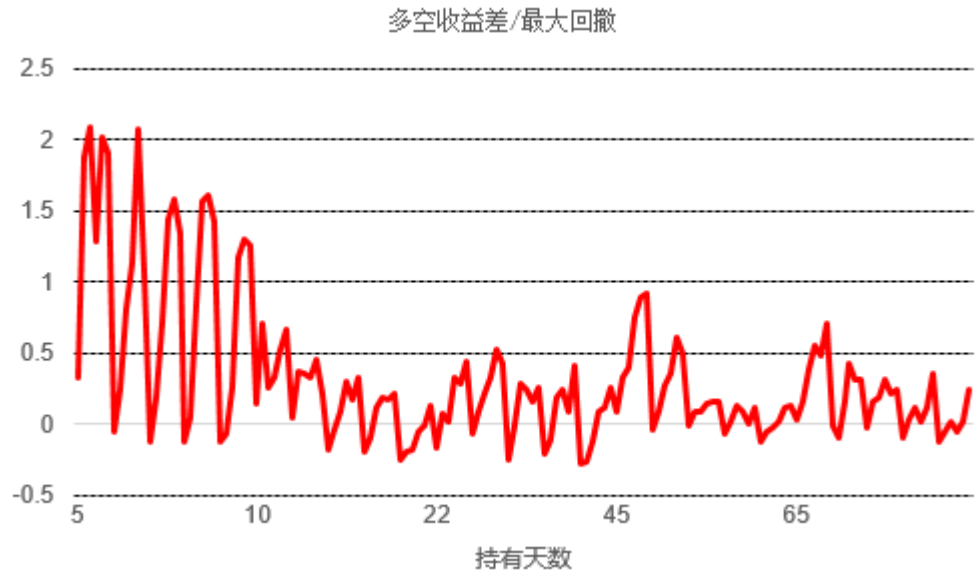


数据来源：wind 资讯，中信建投证券研究发展部

五、 因子敏感性分析

以沪深 300 成份股作为标的池，为了寻找到合适的参数，首先把新闻情绪指数和的天数即买入后持有天数设置为[5,10,22,45,65]，其分别代表持有一周，持有十天，持有一个月，持有二个月，持有一个季度。持有股票数量为[5,10,20,30,50]。负面新闻权重设置为[0.1,0.5,1.0,1.5,2,3]，正面新闻权重为 1。

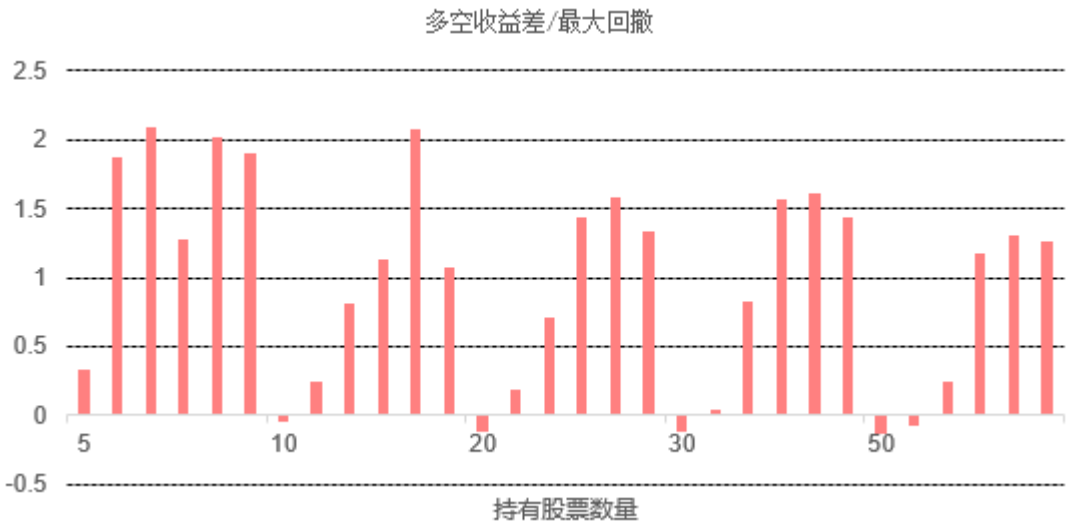
图 10：沪深 300 成份股所有情况收益回撤比



数据来源：中信建投证券研究发展部

其中，纵轴表示年化多空收益差与最大回撤比，横轴表示买入后持有天数。在持有天数从 5 到 10 时，对应有 30 个点，分别表示持有股票数量为[5,10,20,30,50]中 5 种情况时，负面新闻权重设置为[0.1,0.5,1.0,1.5,2,3]的六种情况，共 5*6 得到 30 个数，持有天数从 10 到 22 等依此类推。从上图可知，买入后持有 5 天明显优于其它持有天数。故进一步因子敏感性分析，我们只考虑持有 5 天的情况如下：

图 11：沪深 300 成份股买入后持有 5 天收益回撤比



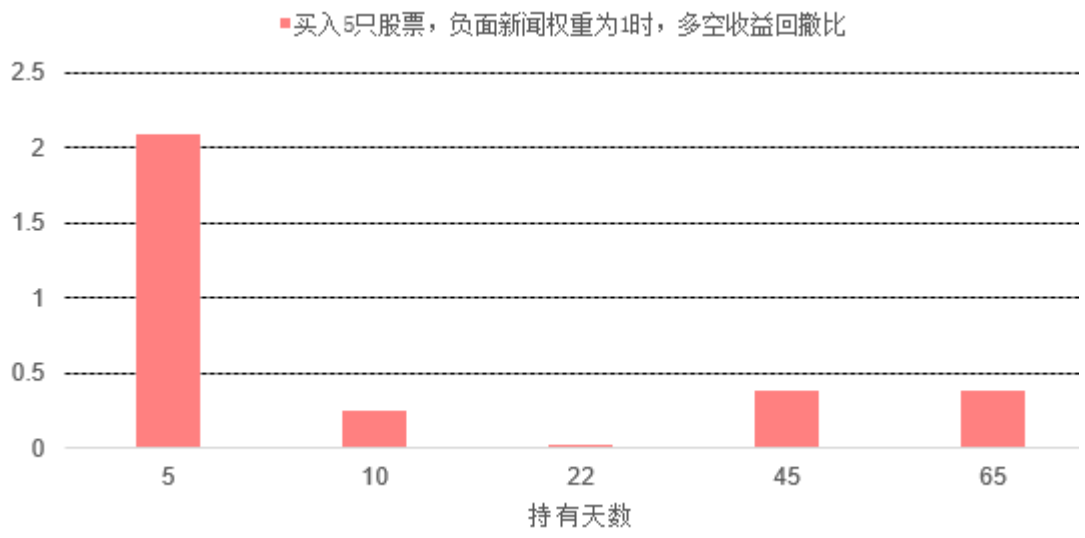
数据来源：中信建投证券研究发展部

其中，纵轴表示年化多空收益差与最大回撤比，横轴表示买入股票数量，股票数量从 5 到 10 时，对应有 6 条柱形图，分别表示持有负面新闻权重设置为[0.1,0.5,1.0,1.5,2,3]的六种情况，股票数量从 10 到 20 等依此类推。

从上图可知，随着持有股票数量越来越多，收益回撤比总体上有递减的趋势，随着负面新闻权重的增大，收益回撤比总体上处于递增的趋势，当持有 5 只股票，负面权重为 1 时，收益回撤比达到最大为 2.09。

综上所述，持有 5 天，买入 5 只股票，负面新闻权重与正面新闻权重等权时，有最优收益回撤比。为了进一步分析最优情况，我们分析在最优时，二个参数不变，所有情况的收益回撤比。

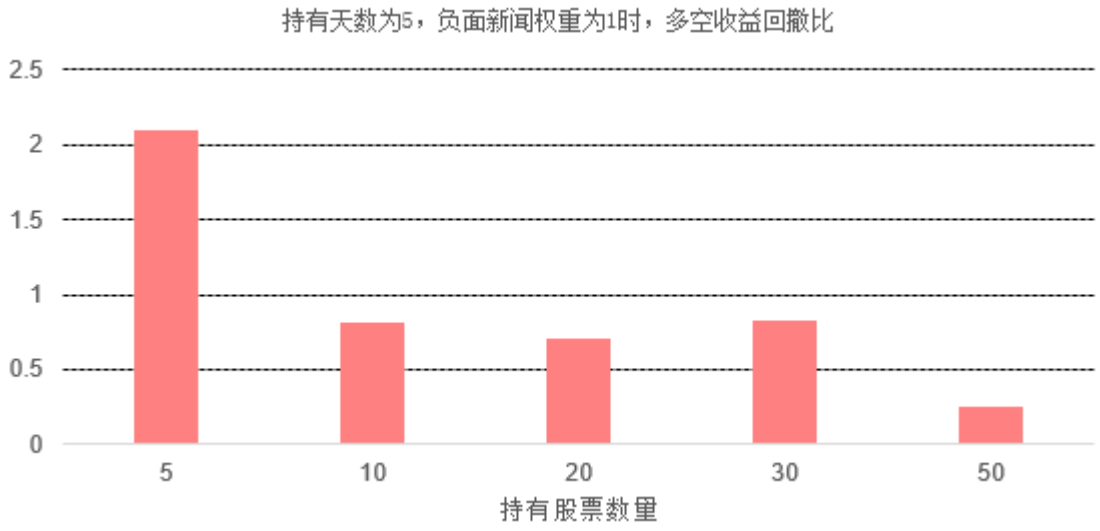
图 12：买入 5 只股票，负面新闻权重为 1 时多空收益回撤比



数据来源：中信建投证券研究发展部

买入 5 只股票，负面新闻权重为 1 时，不同持有天数多空收益回撤比差别很大，持有 5 天远大于其他情况。可理解为新闻情绪对沪深 300 成份股的影响持续 5 个交易日。

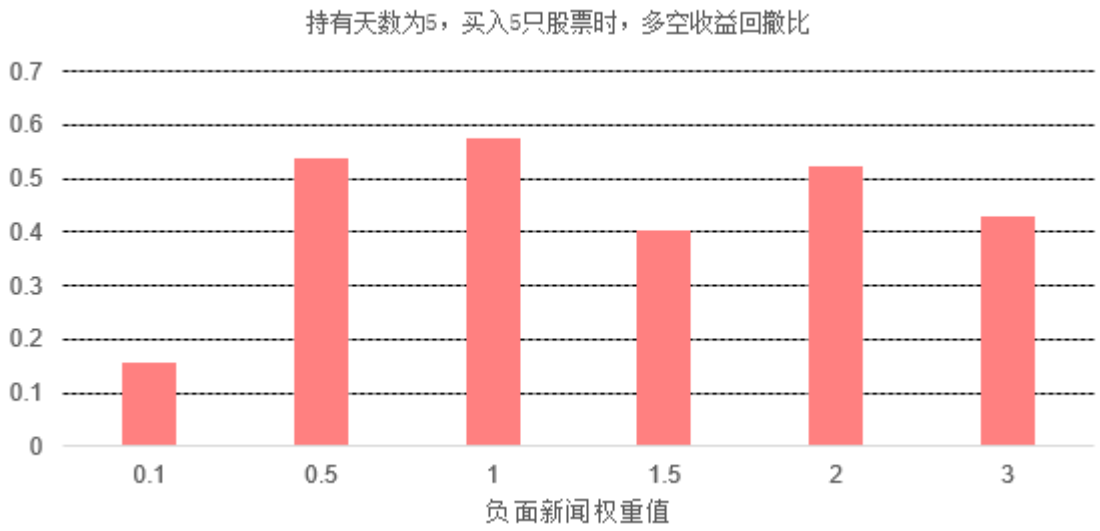
图 13：持有天数为 5，负面新闻权重为 1 时多空收益回撤比



数据来源：中信建投证券研究发展部

持有天数为 5，负面新闻权重为 1 时，持有不同股票数量多空收益回撤比差别很大，且有递减的趋势。这表明新闻情绪指数排名靠前的股票有较明显的超额收益。

图 14：持有天数为 5，买入 5 只股票时，多空收益回撤比



数据来源：中信建投证券研究发展部

持有天数为 5，买入排名前 5 的股票时，不同负面新闻权重多空收益回撤比差别不大。这表明负面新闻权重在最优情况时，基本没有创造超额收益。

六、 总结

传统量化投资主要包括量化选股、量化择时、股指期货套利、商品期货套利、统计套利、算法交易，资产配置，风险控制等。传统的量化投资研究的数据来源一般是公司的财务指标、交易行情数据、政策宏观方面的投资信息等。大数据将为量化投资这一领域创造前所未有的可量化的新的维度，为量化投资提供了新的研究视野。如何把大数据这一金矿从数据转变为知识则充满挑战和困难，大数据将驱动量化投资的创新。

通过市场情绪分析、财经文本分析、新闻热点捕捉、主题挖掘等从这些大量的新闻中挖掘出有效信息。利用数据挖掘技术，即利用各种方法分析我们需要处理的数据，发现隐藏在海量数据背后的知识和规律。挖掘步骤简单的可以概括为 a.前期数据的准备 b.从这些数据中寻找他们的规律 c.把寻找到的规律表示出来，这 3 个步骤。前期数据的准备是从这些相关的数据源中以一定的规则挑选我们所需的数据，然后整合成我们用于数据挖掘的数据集；寻找这些数据的规律是利用数据挖掘相关的方法将这些数据集所含的规律挖掘出来；把寻找到的规律表示出来是利用比如图表等可视化的技术尽可能以用户可以理解的方式展示出来挖掘出来的规律。

此文研究中，我们利用个股新闻情绪指数作为唯一因子来选股，即把 N 日正负面新闻权重和构成的当日新闻情绪指数相加，然后进行排序，选取排名前 topN 的股票作为多头或空头，选取排名倒数前 topN 的股票作为空头或者多头,最后计算多空收益差。

根据研究得出，新闻情绪因子选股在不同板块效果相关很大，甚至是决然相反的结论。

主板方面，尤其是沪深 300 成份股，情绪指数与股票未来上涨为正向指标，且持续天数大概 5 个交易日。组合均剔除掉交易日一字板涨跌停和停牌的股票，新闻情绪指数排名靠前的股票表现明显优于排名靠后的股票，我们以 5 个交易日正负面新闻权重和构成的当日新闻情绪指数相加，然后进行排序，把选取排名前 5 的股票作为多头组合，选取排名倒数前 5 作为空头组合，多空组合的年化收益差高达 50.44%，夏普比为 1.55，最大回撤 27.5%，5 日胜率 59.29%；

主板方面,多头组合中新闻情绪指数动量效应明显。每 5 个交易日调仓一次，换手率平均值为 43%，而空头组合中，换手率平均值为 80%。这说明，多头组合中，当新闻情绪因子排名前 5 时，下一次排名前 5 的概率也非常大，达到 57%，即情绪因子动量十分明显；而在空头组合中，换手率达到 80%，这说明当新闻情绪因子排名倒数前 5 时，下一次排名倒数前 5 时的概率则比较小，仅为 20%。

中小板方面，情绪指数与股票未来上涨为反向指标，且持续天数大概 22 个交易日。以 22 个交易日为周期，选取 22 个交易日情绪指数和排名前 50 的股票作为空头，选取排名倒数前 50 的作为多头，并持有 22 个交易日。组合均剔除掉交易日一字板涨跌停和停牌的股票。该策略多头年化收益率为 52.79%，年化多空收益差为 37.33%，多空收益差最大回撤为 11.17%，日胜率为 57.52%，22 日胜率高达 80.65%。

创业板块方面，情绪指数与股票未来上涨或下跌效果不显著。我们选取较优的参数，以 22 个交易日为周期，选取 22 个交易日情绪指数和排名前 60 的股票作为空头，选取排名倒数前 60 的作为多头，并持有 22 个交易日。组合均剔除掉交易日一字板涨跌停和停牌的股票。该策略多头组合净值为 2.82，年化收益率为 42.41%，多空收益差年化为 23.64%，多空收益差最大回撤为 12.79%，日胜率为 56.54%，22 日胜率高达 67.74%。而同期创业板综指年化收益为 22.2%，即新闻情绪因子在创业板中效果并不明显，多空收益差与同期创业板综指差别不大。

存在的不足：

当然，该策略也有不足的地方，首先新闻来源比较单一，目前主要来源于新浪财经网站，来自 200 多家媒体的新闻。其次，指标比较单一，该策略仅对沪深 300 成份股的新闻数量进行了统计，虽然具有一定的代表性，但并不全面。

以上相关结论仅来自 2014 年 1 月 1 日到 2016 年 12 月 30 日期间的统计分析得出，时间相对并不太长，不能完全代表未来及 14 年以前的过去。

七、 风险提示

以上结论来自历史统计，并不代表未来一定有这种规律。

大数据预测的前提是数据大而全，并且数据质量可靠。由于数据来源有限，目前主要用新浪财经的个股新闻来做研究，虽然具有代表性，但并不能完全代表市场

注：以上相关计算中，交易手续费为双向千分之三，无风险利率为 2.5%，剔除一字涨跌停及停牌股票，剔除上市不足一个月的新股。

分析师介绍

丁鲁明：同济大学金融数学硕士，中国准精算师，现任中信建投证券研究发展部金融工程方向负责人，首席分析师。9 年证券从业，历任海通证券研究所金融工程研究员、量化资产配置方向负责人；先后从事转债、选股、高频交易、行业配置、大类资产配置等领域的量化策略研究，对国内证券市场的量化策略构建具备资深经验。曾多次荣获：新财富最佳分析师上榜，包括 2009 第 4、2012 第 4、2013 第 1、2014 第 3 等；水晶球奖：2009 第 1、2013 第 1 等。

研究助理 喻银尤：021-68821600-808 yuyinyou@csc.com.cn
复旦大学硕士，两年上交所相关部门工作经验。专注于大数据处理，数据挖掘，文本分析，舆情分析等相关策略研究。

研究服务

社保基金销售经理

彭砚苹 010-85130892 pengyanping@csc.com.cn
姜东亚 010-85156405 jiangdongya@csc.com.cn

机构销售负责人

赵海兰 010-85130909 zhaohailan@csc.com.cn

北京地区销售经理

张博 010-85130905 zhangbo@csc.com.cn
黄玮 010-85130318 huangwei@csc.com.cn
李祉瑶 010-85130464 lizhiyao@csc.com.cn
朱燕 010-85156403 zhuyan@csc.com.cn
李静 010-85130595 lijing@csc.com.cn
赵倩 010-85159313 zhaoqian@csc.com.cn
黄杉 010-85156350 huangshan@csc.com.cn
任师蕙 010-85159274 renshihui@csc.com.cn
王健 010-65608249 wangjianyf@csc.com.cn
周瑞 18611606170 zhourui@csc.com.cn
刘凯 010-86451013 liukaizgs@csc.com.cn

上海地区销售经理

陈诗泓 021-68821600 chenshihong@csc.com.cn
邓欣 021-68821600 dengxin@csc.com.cn
黄方禅 021-68821615 huangfangchan@csc.com.cn
戴悦放 021-68821617 daiyuefang@csc.com.cn
李岚 021-68821618 lilan@csc.com.cn
潘振亚 021-68821619 panzhenya@csc.com.cn
肖垚 021-68821631 xiaoyao@csc.com.cn
吉佳 021-68821600 jijia@csc.com.cn
朱丽 021-68821600 zhuli@csc.com.cn
杨晶 021-68821600 yangjingzgs@csc.com.cn

深广地区销售经理

胡倩 0755-23953859 huqian@csc.com.cn
芦冠宇 0755-23953859 luguanyu@csc.com.cn
张苗苗 020-38381071 zhangmiaomiao@csc.com.cn
许舒枫 0755-23953843 xushufeng@csc.com.cn
王留阳 0755-22663051 wangliuyang@csc.com.cn
廖成涛 0755-22663051 liaochengtao@csc.com.cn

券商私募销售经理

任威 010-85130923 renwei@csc.com.cn

评级说明

以上证指数或者深证综指的涨跌幅为基准。

买入：未来 6 个月内相对超出市场表现 15%以上；

增持：未来 6 个月内相对超出市场表现 5—15%；

中性：未来 6 个月内相对市场表现在-5—5%之间；

减持：未来 6 个月内相对弱于市场表现 5—15%；

卖出：未来 6 个月内相对弱于市场表现 15%以上。

重要声明

地址

北京中信建投证券研究发展部

中国北京 100010

东城区朝内大街 2 号凯恒中心 B 座 12 层

电话：(8610) 8513-0588

传真：(8610) 6518-0322

上海中信建投证券研究发展部

中国上海 200120

浦东新区浦东南路 528 号上海证券大厦北塔 22 楼 2201 室

电话：(8621) 6882-1612

传真：(8621) 6882-1622