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

机器学习之贝叶斯文本分类算法的实现：
——大数据研究之指标构建

金融工程研究

重要观点

自然语言处理技术

自然语言处理（NLP）是计算机科学，人工智能，语言学关注计算机和人类（自然）语言之间的相互作用的领域，主要范畴包括切词，词性标注，句法分析，语义分析等。

新闻情绪指数构建概述

新闻情绪指数构建即使用朴素贝叶斯文本分类算法对个股新闻进行正负面分类，正面新闻权重为 1，负面新闻权重为负 W（W 大于零，即负面新闻对个股影响力为正面新闻的-W 倍）。然后根据正负面新闻权重和构建当日新闻情绪指数。

新闻情绪明显偏爱主板

2014 年 1 月 1 日到 2016 年 11 月 30 日，主板日平均情绪指数为 698，而中小板为 357，创业板为 144。主板个股与中小板个股日平均情绪指数为 0.45,而创业板个股日平均情绪指数为 0.28，明显低于主板及中小板。

正面新闻数量占绝对优势

从 2014 年 1 月 1 日到 2016 年 11 月 30 日，正面新闻比例高达 71%，为负面新闻的近 2.5 倍。其中，2014 年正面新闻比例 70.71%，2015 年正面新闻比例 79.67%，2016 年正面新闻比例 62.78%。

新闻情绪指数与大盘走势基本一致

从 2014 年 1 月 1 日到 2016 年 11 月 30 日，大盘经历了疯狂的牛市然后断崖式下跌，最后企稳回升。与此同时，新闻情绪指数走势也是先上升后急剧下跌，最后企稳回升。

未来研究

在本报告中，主要研究如何构建情绪指数，后期我们将研究如何运用这些指数。

丁鲁明

dingluming@csc.com.cn

021-68821623

执业证书编号：S1440515020001

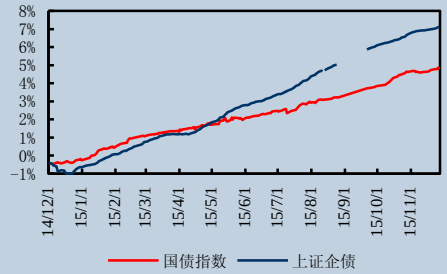
研究助理：喻银尤

yuyinyou@csc.com.cn

021-68821600-808

发布日期： 2017 年 03 月 02 日

市场表现



相关研究报告

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
16.05.31	基本面量化系列之四——量化全球大类资产配置体系
16.05.27	躲不过的中国资本市场宏观对冲时代——非传统型基金产品概述

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。

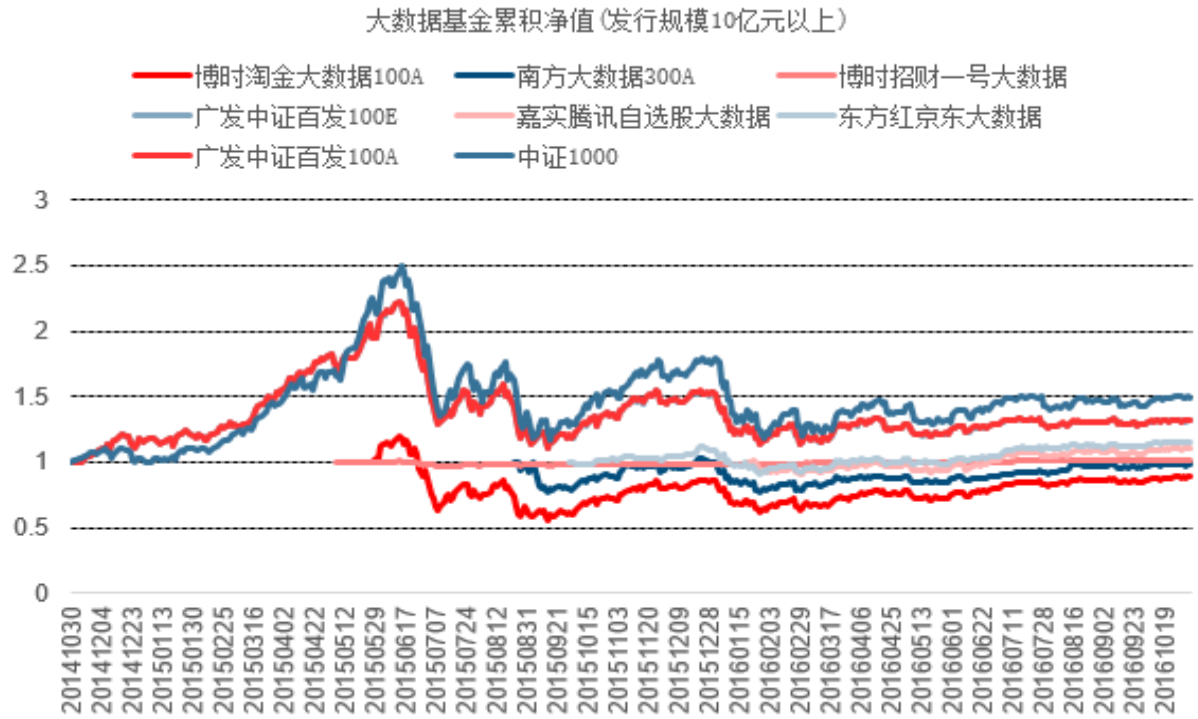
传统量化投资主要包括量化选股、量化择时、股指期货套利、商品期货套利、统计套利、算法交易，资产配置，风险控制等。传统的量化投资研究的数据来源一般是公司的财务指标、交易行情数据、政策宏观方面的投资信息等。而随着量化投资这一领域的快速发展，这些传统数据中所包括的大部分投资信息已经被专业投资者所挖掘，想要从这些信息中获取收益难度将越来越大。大数据将为量化投资这一领域创造前所未有的可量化的新的维度，为量化投资提供了新的研究视野。如何把大数据这一金矿从数据转变为知识则充满挑战和困难，大数据将驱动量化投资的创新。

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

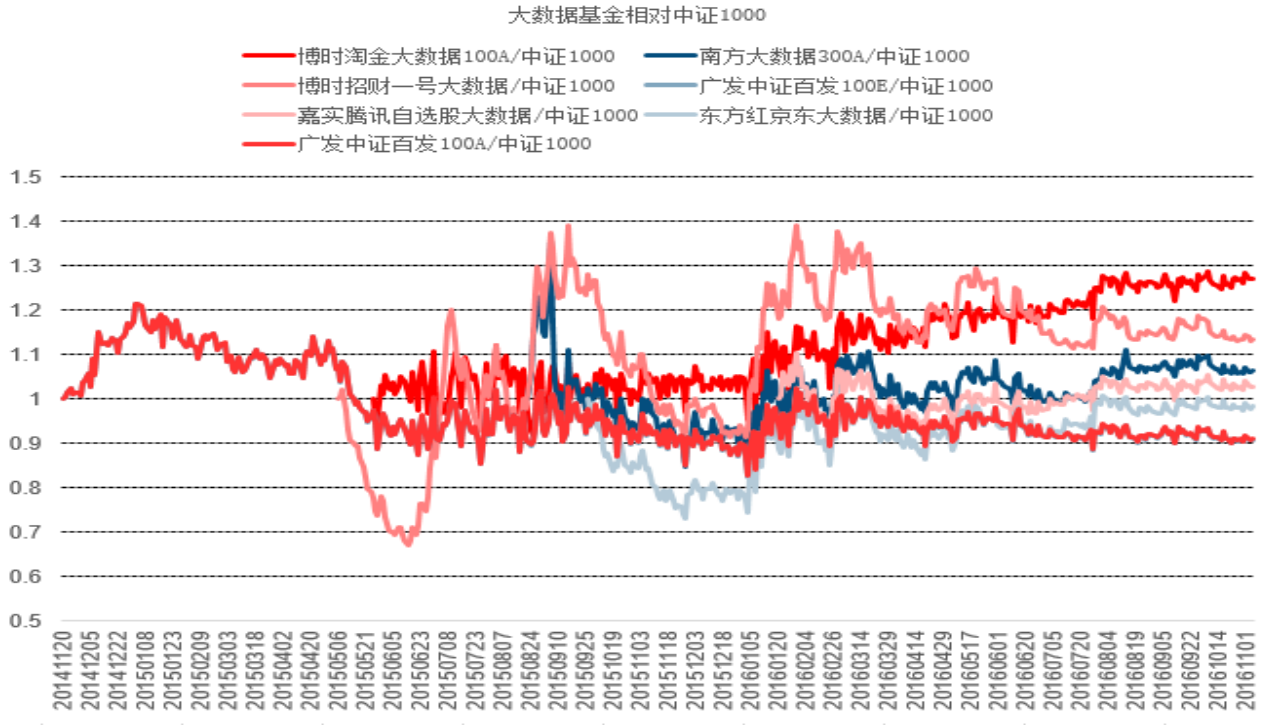
数据来源：wind 资讯，中信建投证券研究发展部

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：

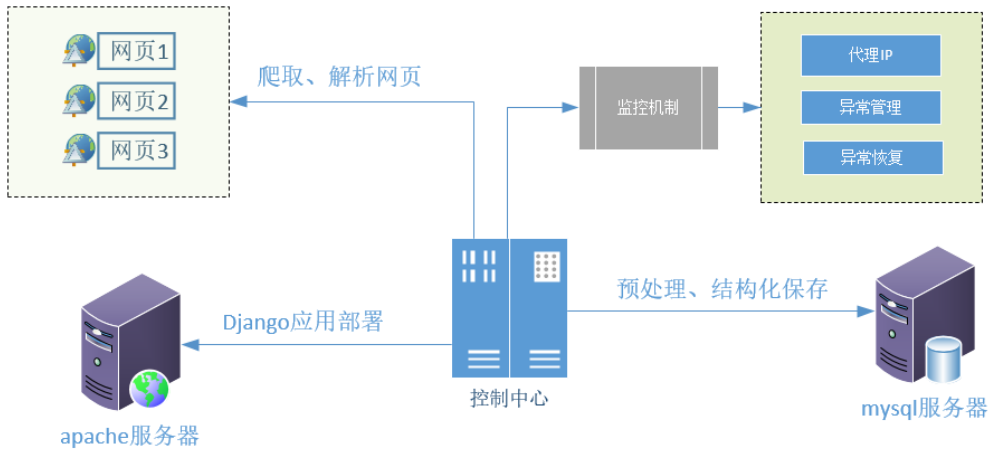
2.1 数据采集与预处理

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、 互联网情绪指标

通过数据挖掘技术，我们可以构建出所有我们需要的指标。目前市场上主要相关指标如下：

3.1 财经新闻指数

个股新闻指数：每支股票每天对应的新闻总数。

个股情绪指数：个股情绪新闻正负面加权汇总。

宏观经济指数：通过财经文本分析，得到新闻正负面，构建加权宏观经济指数等。

图 4：新浪财经个股新闻

数据来源：中信建投证券研究发展部，新浪

3.2 股吧相关指数

个股舆情指数：统计个股相关舆情，判断正负面，加权汇总。

个股每日股吧指数：个股相关每日帖子总数。

个股每周股吧指数：个股相关每周帖子总数。

图 5：股吧

数据来源：中信建投证券研究发展部，东方财富网

3.3 微信公众号指数

个股每日微信推送指数：个股每日被所有相关机构微信公众号推荐总次数。

微信每日推荐个股指数：所有相关机构微信公众号每日推送个股数总数。

图 6：微信每日推送

数据来源：中信建投证券研究发展部，搜狗

3.4 微博相关指数

个股每日微博指数：个股相关每日微博总次数。

个股官方微博指数：个股官方微博文章总数量

个股高官微博指数：公司高管在微博发文总数及情绪加权汇总。

3.5 公司关注度指数

个股搜索指数：个股每天被百度，搜狗等主流搜索引擎被搜索次数总和。

个股访问热度指数：个股行情被用户访问总数。

个股新闻点击率指数：个股新闻被点击总数。

3.6 其它相关指标

电商个股销量指数：淘宝，京东等电商个股相关商品金额总数。

电商行业销量指数：淘宝，京东等电商行业相关商品商量总数。

电商个股综合指数：淘宝，京东等电商个股相关商品的浏览量、点击量、客户评价、客户收藏量等基础数据的综合指标。

电商行业综合指数：淘宝，京东等电商行业相关商品的浏览量、点击量、客户评价、客户收藏量等基础数据的综合指标。

银联消费指数：银联刷卡相关个股或者行业数据组成的指标。

等等。

四、 新闻情绪指数构建

新闻情绪指数即把新闻分类，根据新闻分类结果，分别给予权重，最后加权求和，得到当天的指标值。对于新闻，可分类为正面，负面。假设正面新闻对相应股票影响程度为 1，则负面对应为负 w(w 为任意数)。他们的加权和构建成当天指标。

常用的文本分类方法有支持向量机、K-近邻算法和朴素贝叶斯方法。其中朴素贝叶斯相对简单，运行速度快且分类准确率较高，应用相对较广。

4.1 朴素贝叶斯文本分类

4.1.1、贝叶斯基本原理

贝叶斯分类是以贝叶斯定理为基础的一类分类算法的总称。

贝叶斯定理：

贝叶斯定理也称贝叶斯推理，早在 18 世纪，英国学者贝叶斯(1702～1763)曾提出计算条件概率的公式用来解决如下一类问题：假设 B[1],B[2]…,B[n]互斥且构成一个完全事件，已知它们的概率 P(B[i]),i=1,2,…,n,现观察到某事件 A 与 B[1],B[2]…,B[n]相伴随机出现，且已知条件概率 P(A/B[i])，求 P(B[i]/A)。

贝叶斯公式：（发表于 1763 年）

P(B[i]/A)=P(B[i])*P(A｜B[i])/{P(B[1])*P(A｜B[1]) +P(B[2])*P(A｜B[2])+…+P(B[n])*P(A｜B[n])}

其中，P(B[i])为先验概率，P(B[i]/A) 是已知 A 发生后 B[i]的条件概率，也由于得到 A 后才取值而被称作 B[i]的后验概率。

朴素贝叶斯文本分类原理：

朴素贝叶斯算法假设前提：在给定目标值时属性值之间相互条件独立。

贝叶斯方法的新实例分类目标是在给定描述实例的属性值（A1,A2…An）下，得到最可能的目标值 V。

V = arg maxP(B[j]/ A1,A2…An)

根据贝叶斯公式:

V = arg maxP(A1,A2…An /B[j]) * P(B[j]) / P(A1,A2…An)

由于 P(A1,A2…An)是不依赖于 B[i]的常量，故简化为：

V = arg maxP(A1,A2…An /B[j]) * P(B[j])

因为属性值之间相互条件独立，即联合 A1,A2…An 的概率等于每个单独属性的概率乘积：

Vnb = arg maxP(B[j])* ∏P(Ai | B[j])　　其中 i 为 1 到 n。

Vnb 表示朴素贝叶斯输出的目标值。

4.1.2、朴素贝叶斯分类模型

朴素贝叶斯分类模型为以下几类：

1)　多元分布模型（muiltinomial model）

多元分布模型以单词为粒度，不仅仅计算特征词出现/不出现，还要计算出现的次数。另外多元分布模型实际中分别还可用 tf-idf ,Bool（Bool 表示某个单词是否在某个文档中出现，如果出现则记为 1，否定则记为 0）。BOOL 型特征下的向量空间模型更适合做情绪分类。以下公式以单词出现次数为例。

类条件概率:

P(A1|B[0]) = (Na+1)/(n+N)

其中: Na 为类 B[0]下单词 A1 出现在所有文档中的次数之和

n 为类 B[0]下特征词总数

N 为训练样本中不重复的特征词总数

先验概率:

P(B[0]) = 类 B[0]下单词总数 /训练样本中的特征词总数

2) 伯努利模型

伯努利模型以文件为粒度。

类条件概率:

P(A1 | B[0]) = (类 B[0]下出现 A1 的文件总数 + 1) / (类 B[0]下的文件总数 + 2)

先验概率:

P(B[0]) = (类 B[0]下的文件总数) / (整个训练样本文件总数)

以上两种模型的类条件概率分子中加 1,原因是待分类文本中的属性可能样本中没有,会导致条件概率为 0 的情况. 贝叶斯公式推导前提各个特征不能为 0。因此实现上通常要做一些小的处理，例如把所有计数进行+1（加法平滑(additive smoothing，又叫拉普拉斯平滑(Laplace smothing)）。而如果通过增加一个大于 0 的可调参数 alpha 进行平滑，就叫 Lidstone 平滑。

4.2 新闻情绪分类实现

4.2.1、新闻情绪分类

新闻情绪分类算法主要步奏如下：

1. 新闻数据准备与预处理：

把需要分类的新闻收集，并预处理。此处预处理主要是去掉重复新闻，去掉图片及新闻中穿插的广告等。

2. 训练集收集：

因为分类算法属于有监督学习算法，必须人工先分类好新闻作为训练集，此处我们把新闻分为正

负两个属性。训练集是非常重要的，直接影响到分类效果。我们的训练集共有 1005 条新闻，其中正面新闻 668 条，负面新闻 337 条。这些新闻来自于 2014 年到 2016 年正负面特征比较明显的新闻。在训练测试中，其正确率如下：

表 2：训练集新闻正负面效果检验

	样本总数	训练数	测试数	正确数	正确率
正负面集	1005	804	201	160	0.79602
负面新闻	337	261	76	60	0.789474
正面新闻	668	543	125	100	0.8

数据来源：中信建投证券研究发展部

总样本正确率达到近 80%，这是非常高的。人工也很难把新闻分类达到 90%以上，因为有些新闻本身正负面特征并不明显。

3. 导入自定义词典：

金融行业专业名词及相关重要词语应人工先定义好在自定义词典中，以免分词时，把重要的专业名词分割成多个词语，导致分词效果不好。

4. 切词

即把新闻切割为词语。比如：

“8 月 15 日开盘，万科再出现大涨态势，涨幅一度超 7%，股价最高达 24.56 元，创出多年来的新高，此后有所回落，截至发稿时，成交量接近 30 亿元”

分割为（没有考虑停用词的情况）：

“8 月 15 日 开盘 ， 万科 再 出现 大涨 态势 ， 涨幅 一度 超 7 % ， 股价 最高 达 24.56 元 ， 创出 多年 来 的 新高 ， 此后 有所 回落 ， 截至 发稿 时 ， 成交量 接近 30 亿元”

5. 去除停用词

所谓的停用词，指对新闻特征没有意义的词。比如，“月”，“日”，“我们”，“是”等之类的对新闻特征没有帮助的词。由于我们对新闻分类时，不考虑数字，英文及标点符号。故我们把这几类也归为停用词。在我们系统中，停用词 4200 多个。

上例中，去除停用词之后为：

“开盘 万科 再 出现 大涨 态势 涨幅 一度 超 股价 最高 达元 创出 多年 新高 有所 回落 发稿 时 成交量 接近”

6. 训练集特征提取

特征提取可用 TF-IDF 算法，或者简单的词频统计，或只统计出现与不出现。

7. 待分类新闻利用朴素贝叶斯分类算法进行分类：

后面会具体介绍。

4.2.2、朴素贝叶斯应用

在新闻分类中，目标集由 B[0],B[1]组成。其中，B[0]为负面，B[1]为正面。属性值（A1,A2…An）即为新闻文本经过分词后的结果。比如：

（A1,A2…An）=(开盘 ，万科， 再 ，出现 ，大涨 ，态势， 涨幅， 一度 ，超 ，股价， 最高， 达元，创出， 多年， 新高， 有所， 回落 ，发稿 ，时 ，成交量， 接近)共 21 个属性值。

根据上一节，我们需要得出 Vnb = arg maxP(B[j])* ∏P(Ai | B[j])。需要依次求出 P(Ai | B[j])及 P(B[j])，然后相乘，最大值取为该文本的分类结果。

比如“负面”先验概率 P(B[0]) = 类“负面”下特征词总数 /训练样本中的特征词总数

比如，“开盘”在为类“负面”的条件概率为：

类条件概率 P(A1 | B[0])=p(开盘|负面) = (类“负面”下单词“开盘”出现在所有文档中的次数之和+1)/(类“负面”下特征词总数 +训练样本中不重复的特征词总数)。

同样可以求出剩下的属性特征词，(万科， 再 ，出现 ，大涨 ，态势， 涨幅， 一度 ，超 ，股价， 最高， 达元， 创出， 多年， 新高， 有所， 回落 ，发稿 ，时 ，成交量， 接近)

得出由这些属性值组成的新闻文本归类为“负面”的概率为：

V0 = P(B[0])* P(A1 | B[0])* P(A2 | B[0])*…* P(A21 | B[0])

同样的方法，求出由这些属性值组成的新闻文本归类为“正面”的概率为：

V1 = P(B[1])* P(A1 | B[1])* P(A2 | B[1])*…* P(A21 | B[1])

最后，比较 V0 与 V1 的大小,若 V0>V1,则该新闻文本应该归类为负面。否则为正面。

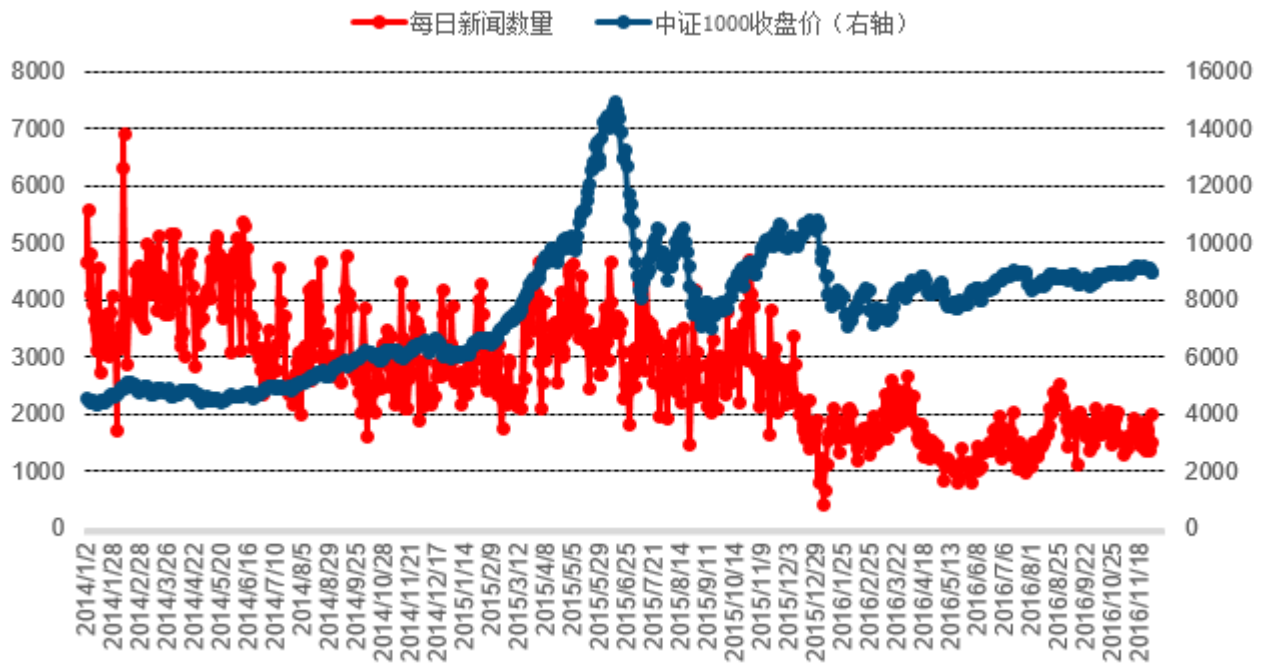
4.2.3、情绪指数构建

新闻情绪指数即把新闻分类，根据新闻分类结果，分别给予权重，最后加权求和，得到当天的指标值。对于新闻，可分类为正面，负面，在传统的新闻分类研究中，大家分类之后，一般赋予正面新闻与负面新闻等权

重，即负面新闻与正面新闻效果等同。而现实中，正面新闻与负面新闻的影响并不是相同的。假设正面新闻对相应股票影响程度为 1，则负面对应为负 w(w 为任意数)。他们的加权和构建成当天指标。N 天指标和相加可构成 N 天指标。

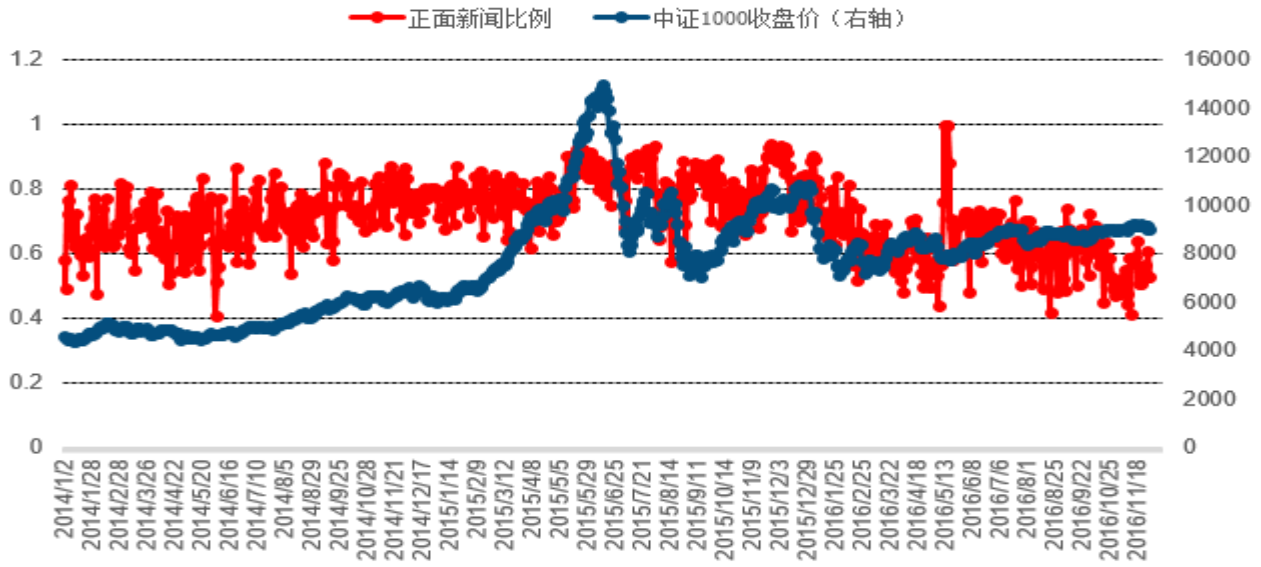
某个股当天新闻情绪指数=当天所有正面新闻数 n – w*(当天所有负面新闻数量 m)

图 7：中信建投金融工程每日个股新闻数量



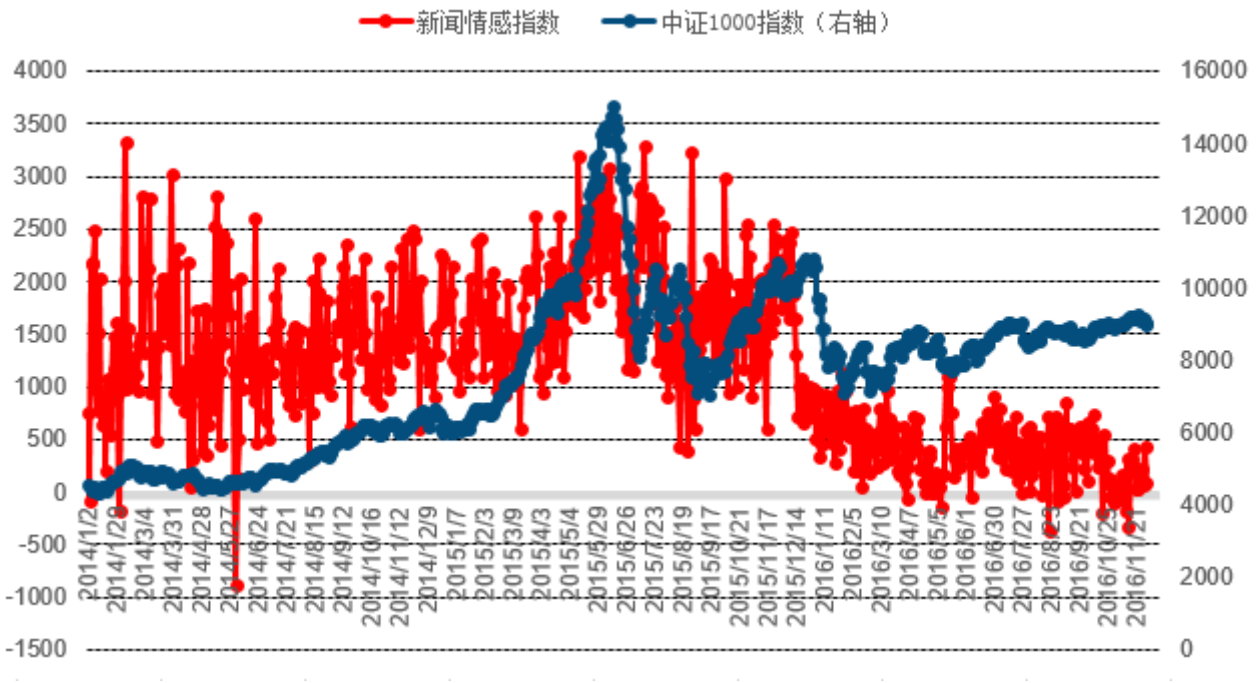
数据来源：wind 资讯，中信建投证券研究发展部

图 8：中信建投金融工程每日个股新闻正面比例



数据来源：wind 资讯，中信建投证券研究发展部

图 9：中信建投金融工程新闻情绪指数



数据来源：wind 资讯，中信建投证券研究发展部

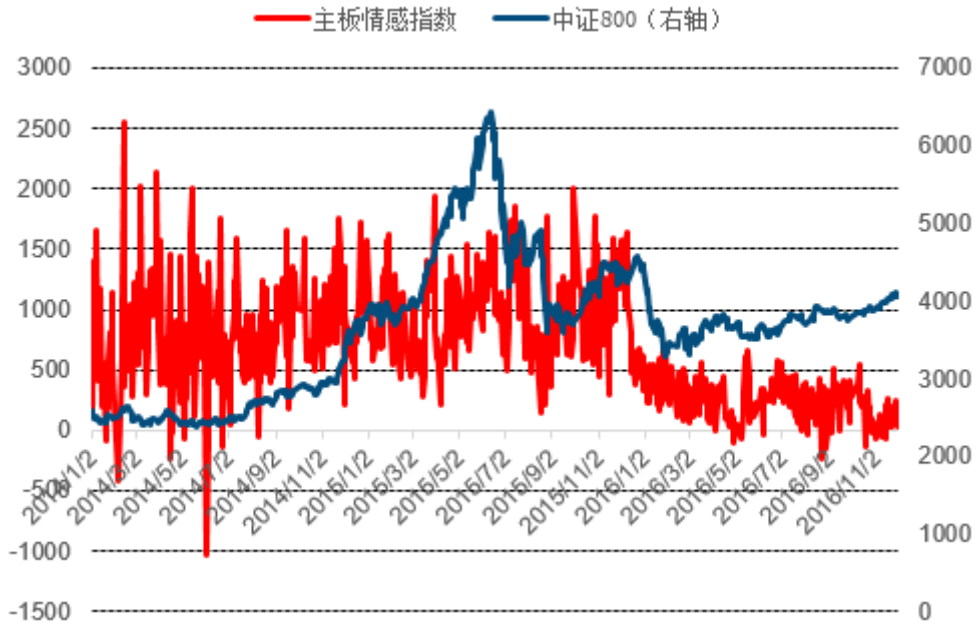
五、 情绪指数与板块

在选个股时，不同投资者偏好各异，比如财务偏好等，比如喜欢现金流充沛、高 ROE、低估值的个股；比如行业偏好，有的偏向金融、地产和稳定成长类公司；也有股权偏好，比如喜欢高分红、股权分散、中等市值等；情绪指数也一样，在不同板块，不同行业其有效性差异可能很大。以下主要展示主板，中小市值及创业板与对应的情绪指数的关系。

5.1 主板情绪指数

主板情绪指数指主板成份股当日新闻情绪指数之和构建的当日指数。

图 10：主板情绪指数指与中证 800

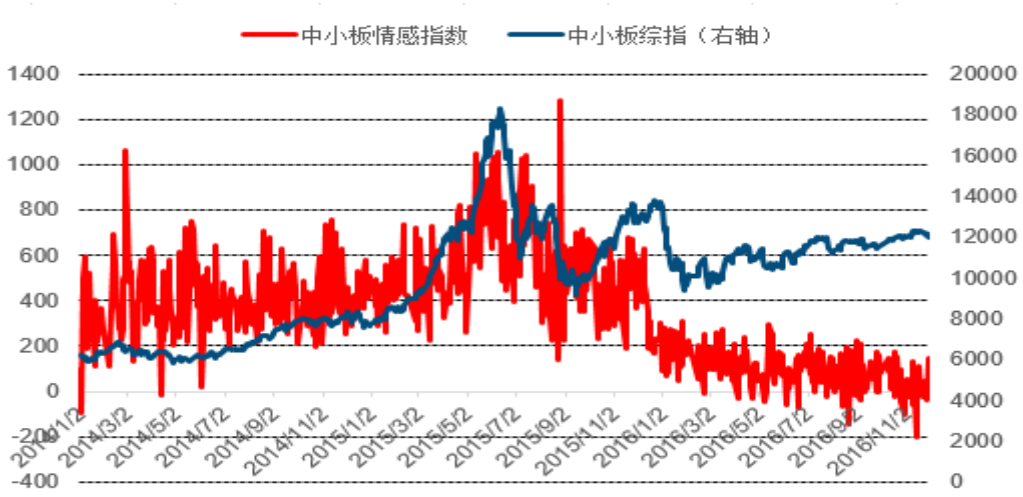


数据来源：wind 资讯，中信建投证券研究发展部

5.2 中小板情绪指数

中小板情绪指数指中小板成份股当日新闻情绪指数之和构建的当日指数。

图 11：中小市值情绪指数与中小板综指

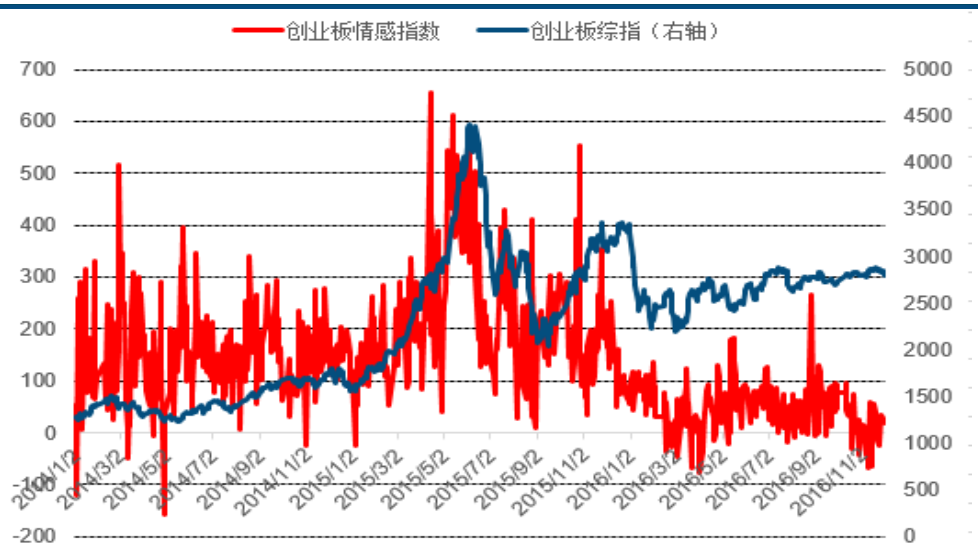


数据来源：wind 资讯，中信建投证券研究发展部

5.3 创业版情绪指数

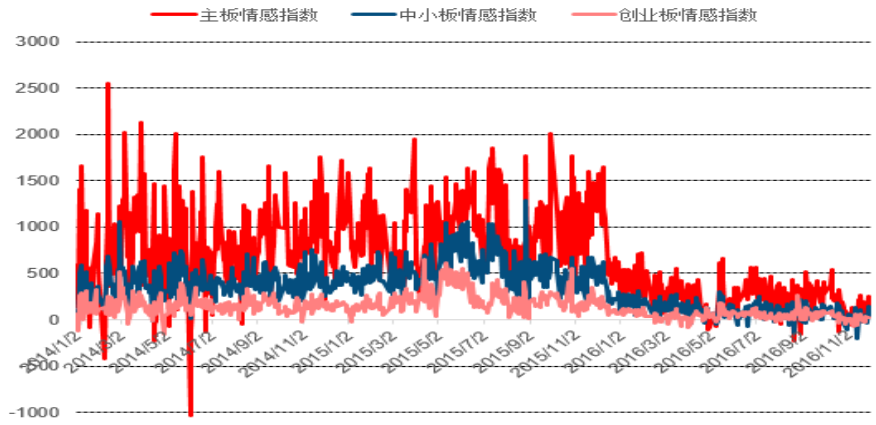
创业板情绪指数指创业板成份股当日新闻情绪指数之和构建的当日指数。

图 12：创业板情绪指数与创业板综指



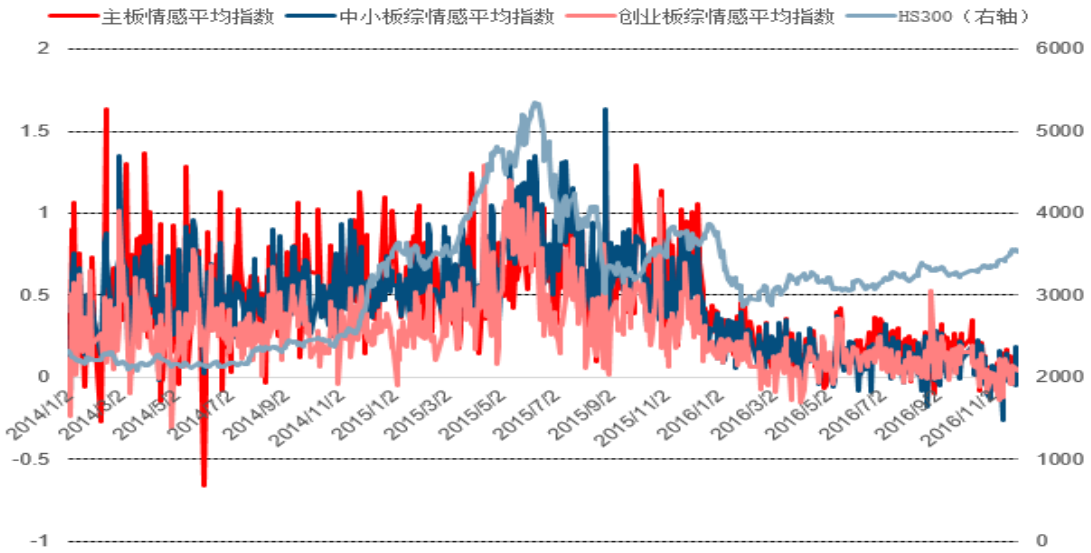
数据来源：wind 资讯，中信建投证券研究发展部

图 13：主板，中小板，创业板情绪指数



数据来源：中信建投证券研究发展部

图 14：主板，中小板，创业板各情绪平均指数与沪深 300 指数



数据来源：wind 资讯，中信建投证券研究发展部

六、 风险提示

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：以上相关情绪指数中，正负面新闻都是赋予同等权重。

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