# 股票数据分析汇总报告

朱恬骅

November 16, 2012

## 1 数据集

### 1.1 小规模数据集

在银行、医药、钢铁板块各选择四支股票，构成12支股票数据集（S12）。

1. 银行：
   - 600000 浦发银行
   - 600015 华夏银行
   - 600016 民生银行
   - 600036 招商银行
2. 医药：
   - 600085 同仁堂
   - 600129 太极集团
   - 6000422 昆明制药
   - 000423 东阿阿胶
3. 钢铁：
   - 600019 宝钢股份
   - 000959 首钢股份
   - 000709 河北钢铁
   - 600569 安阳钢铁

### 1.2 中等规模数据集

#### 1.2.1 来自5个不交叉板块的50支股票

选择了来自银行、钢铁、医药、酒、软件这五个板块的50支股票，构成不交叉板块50支股票数据集（S50）。

1. 银行：
   - 000001 深发展A
   - 002142 宁波银行
   - 600000 浦发银行

600015 华夏银行
600016 民生银行
600036 招商银行
601009 南京银行
601166 兴业银行
601169 北京银行
601288 农业银行

2. 钢铁：

000629 攀钢钒钛
000709 河北钢铁
000717 韶钢松山
000898 鞍钢股份
000932 华菱钢铁
600282 南钢股份
600019 宝钢股份
000959 首钢股份
000022 济南钢铁
600569 安阳钢铁

3. 医药：

601607 上海医药
600511 国药股份
600833 第一医药
600713 南京医药
000028 国药一致
600085 同仁堂
600129 太极集团
600422 昆明制药
000423 东阿阿胶
002589 瑞康医药

4. 酒：

000568 泸州老窖
000858 五粮液
600519 贵州茅台
600779 水井坊
000596 古井贡酒
600809 山西汾酒
000799 酒鬼酒
002304 洋河股份
600702 沱牌舍得
600559 老白干酒

5. 软件：

600570 恒生电子
600756 浪潮软件
000948 南天信息
600271 航天信息
002063 远光软件
002065 东华软件
000938 紫光股份
002073 软控股份
002090 金智科技
002230 科大讯飞

### 1.2.2 来自5个交叉板块的50支股票

选择了来自钢铁、煤炭、汽车、航空、电力这5个板块的各10支股票。它们都存在一定的相关性，如钢铁和电力都依赖煤炭，汽车依赖钢铁，航空则与煤炭、电力所代表的能源产业有密切关联。构成交叉板块50支股票数据集（R50）。

1. 钢铁：
   - 000629 攀钢钒钛
   - 000709 河北钢铁
   - 000717 韶钢松山
   - 000898 鞍钢股份
   - 000932 华菱钢铁
   - 600282 南钢股份
   - 600019 宝钢股份
   - 000959 首钢股份
   - 000022 济南钢铁
   - 600569 安阳钢铁

2. 煤炭：
   - 000780 平庄能源
   - 000723 美锦能源
   - 002128 露天煤业
   - 600188 兖州煤业
   - 600348 阳泉煤业
   - 600546 山煤国际
   - 600740 山西焦化
   - 601001 大同煤业
   - 601666 平煤股份
   - 601898 中煤能源

3. 汽车：
   - 000550 江铃汽车
   - 000572 海马汽车
   - 000625 长安汽车
   - 000800 一汽轿车
   - 000868 安凯客车
   - 000927 一汽夏利
   - 000951 中国重汽
   - 000957 中通客车
   - 002594 比亚迪
   - 600006 东风汽车

4. 航空：
   - 600029 南方航空
   - 600115 东方航空
   - 600221 海南航空
   - 600316 洪都航空
   - 000089 深圳机场
   - 600004 白云机场
   - 600009 上海机场
   - 600151 航天机电
   - 600893 航空动力
   - 000901 航天科技

5. 电力：

600011 华能国际
600021 上海电力
600027 华电国际
600644 乐山电力
600101 明星电力
600116 三峡水利
600131 岷江水电
600236 桂冠电力
600292 九龙电力
600310 桂东电力

## 1.3 大规模数据集

### 1.3.1 100支股票数据集

选择历史数据最多的100支股票构成100支股票数据集（S100）；在所有股票中随机选择500支股票构成500支股票数据集（S500）。具体的股票名单从略。

## 1.4 数据集中股价特征的时间段

对于小规模的数据集，分别选择了2007年7月1日～2009年7月1日、2008年7月1日～2010年7月1日、2009年7月1日～2011年7月1日这三个时间段的数据，分别记为07、08、09，如12支股票数据集在2007～2009年间的数据，记为S12-07。对于两个中等规模的数据集，选择了2009年7月1日～2011年7月1日和2010年7月1日～2012年7月1日这段时间的数据。对于两个大规模数据集，选择2010年1月1日～2012年1月1日这两年的数据。这样我们实际上共有11个不同时间、不同板块的数据集。

# 2 单视图下的股票聚类实验

## 2.1 股价信息

### 2.1.1 表示法

**时序涨跌幅词频表示法** 将每一支股票的走势看作一篇文档。设每支股票取$T+1$天的价格信息，建立一个大小为$2T$的词汇表，包括了“第$i$天涨1%”和“第$i$天跌1%”，$i=2,3,...,T+1$。将股票的走势表现为这$2T$个词上的词频。

**股票涨跌幅词频表示法** 将每一天的走势看作一篇文档，设共有$N$支股票，则建立一个大小为$2N$的词汇表，包括“第$i$支股票涨1%”和“第$i$支股票跌1%”，$i = 1, 2, ..., N$。将每天的股票行情表示为这$2N$个词上的词频。

**正负零收益表示法** 将每一支股票的走势看作一篇文档。设每支股票取$T + 1$天的价格信息，建立一个大小为$3T$的词汇表，包括了“第$i$天涨”、“第$i$天持平”和“第$i$天跌”，$i = 2, 3, ..., T + 1$。将股票的走势表现为这$3T$个词上的词频。

### 2.1.2 聚类方法

**直接K-Means法** 对于选定的股价特征，直接运行K-Means。由于初始中心的随机性，运行多次，选取类与类之间分布最为平均的一次结果。

**LDA法** 对选定的股价特征，运行LDA进行聚类，选取最可能属于的topic作为这支股票的类标记。

**LDA + K-Means法** 对选定的股价特征，先运行LDA进行聚类；将该股票属于这些topic的可能性作为新的特征，运行K-Means进行聚类。

## 2.2 文本信息（经营范围描述）

### 2.2.1 表示法

**全文词频表示法** 即对经营范围描述信息进行分词后，对所有出现的词都计算词频，是最简单的方法。

**构建关键词词典** 为去除一些意义不大的高频词，需要构造一个比较干净的关键词词典。第一种方法是计算一个词的文档间频率DF及其对应的信息熵$H(w)$，进行降序排序，这就构建出了针对特定语料的关键词词典。以这一词典为基础统计的全文词频，将比在所有词或高频词的字典上统计得到的词频更能代表语料的特征。

### 2.2.2 聚类方法

同2.1.2 。

## 2.3 结果

在所有2215支股票的经营范围描述文本信息中，采用构建关键词词典方法，查找出的前100个高频词如下：

| | | | | |
|---|---|---|---|---|
| 经营 | 企业 | 配件 | 代理 | 法律 |
| 技术 | 项目 | 电子 | 机械设备 | 仪表 |
| 销售 | 机械 | 禁止 | 系统 | 法规 |
| 生产 | 工程 | 管理 | 仪器 | 范围 |
| 业务 | 国家 | 公司 | 建筑 | 货物 |
| 出口 | 商品 | 不含 | 营本企业 | 安装 |
| 设备 | 加工 | 进出口业 | 计算机 | 公司经营 |
| 服务 | 咨询 | 进出口业务 | 经营本企业 | 原辅 |
| 开发 | 除外 | 相关 | 金属 | 贸易 |
| 进出口 | 制造 | 许可证 | 规定 | 零配件 |
| 材料 | 本企业 | 化工 | 信息 | 辅材料 |
| 许可 | 投资 | 设计 | 汽车 | 所需的 |

在所有可能的词（组）中，信息熵最大的词（组）如下：

| | | | | |
|---|---|---|---|---|
| 咨询 | 业务 | 相关 | 建筑 | 及其 |
| 国家 | 企业 | 电子 | 零配件 | 各类 |
| 材料 | 开发 | 自产 | 制品 | 法律 |
| 机械 | 设备 | 设计 | 范围 | 批发 |
| 除外 | 公司 | 仪器 | 规定 | 法规 |
| 商品 | 投资 | 化工 | 进口 | 配件 |
| 禁止 | 管理 | 产品 | 货物 | 自营 |
| 项目 | 加工 | 生产 | 租赁 | 零售 |
| 进出口 | 不含 | 代理 | 安装 | 汽车 |
| 进出口业 | 许可 | 仪表 | 信息 | 限制 |
| 服务 | 许可证 | 限定 | 房地产 | 系统 |
| 制造 | 工程 | 技术 | 计算机 | |

然后使用贝叶斯平均方法提取了所有2215支股票经营范围描述的关键词。得到的部分结果如下：

000001 监管人民币汇款借款放款非贸易有价证券汇兑信托业外币见证资信承兑各项存款贴现调查票据结算外汇代理业人民保险境内买卖允许发行有关境外办理

002142 十一十三十二金融债中国银公众银行卡信用证发放中期款项长期收

付兑付短期吸收债券监督保险业政府承兑银行拆借存款担保中国贴现保管同业票据
600000 外汇托管保险箱全国离岸保障外币借款汇款兑换委员会社会银行业见证资信中国银拆借结汇股票存款担保公众贴现同业信用证发放中期款项长期收付
600015 金融债委员会中国银结汇公众银行卡债券信用证发放款项中期长期收付兑付短期政府吸收监督承兑拆借存款担保贴现保管同业买卖票据结算贷款代理业
600016 本行十四十一十三十二可以银行业结汇金融债公众银行卡信用证发放中期款项长期收付兑付短期吸收监督保险业承兑银行拆借债券存款担保中国政府
这几个都是银行股。

000028 医用区域性救护车口腔科化验缝合一次性灭菌诊断同化第一器具激素手术室急救室精神抗生素射线麻醉药超声敷料附属临床诊疗蛋白毒性疫苗分析消毒合剂
000423 膏剂合剂糖浆口服液保健颗粒剂胶囊药品批准食品范围许可证进出口业商品生产销售
002589 保存常温毒液罂粟助听器隐形眼镜同化激素体外麻醉药蛋白毒性疫苗健身器护理诊断抗生素精神配送三类化学药饮片日用品生化试剂中药材制毒生物制品中成药化妆品
600085 营养液老年病乌鸡作用妇产科儿科梅花鹿乌骨鸡外科冷食品中医科内科马鹿涂膜剂同仁皮肤科供暖定型皮肤北京诊疗其中股份动植物西药饲养有限公司图书保健饮片
600129 执业中草药旅馆水产西药作业二级首饰前不副食品保健金银土地中成药养殖以下工艺美术维护种植经济印刷不得百货医疗旅游器械出租自有化学包装
这几个都是医药方面的。

但也有不好的例子，比如东华软件的：
002065 决定国务院机关注册选择行政自主工商登记不得活动开展批准后方法规法律审批自营规定各类限定许可代理禁止进出口业公司管理商品除外进出口
完全没有体现出它经营范围是什么。

# 3 双视图下的股票聚类实验

## 3.1 视图间聚类的相关性

对于同一对象的描述，因其所选择的视图有所不同，会导致观察到的结果反映了同一对象的不同特征。然而，由于这些特征反映的是同一对象，我们有理由认为这些特征之间可能存在一定的相关性，而不是如之前的模型所假设的那样是独立的。因此，利用这些特征之间的相关性，以组合不同视图下观察到的数据，并以此进行聚类，可能会得到比简单地将两组特征合并进行聚类更好的结果。

矩阵拆分是建立在词频模型上的。将观察到的数据用词频的形式表示。

## 3.2 不同视图下聚类结果的合并

## 3.3 结果