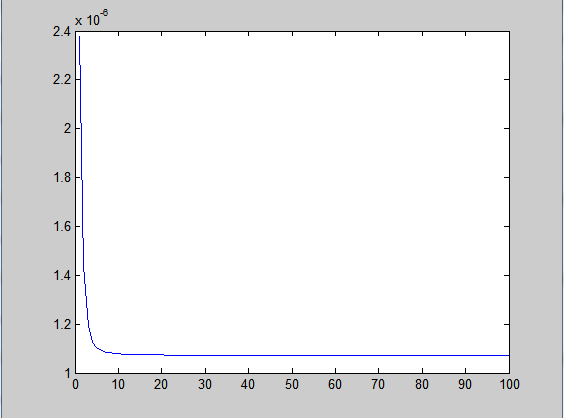
矩阵分解最近研究结果

算法：

1. 明确三个矩阵的含义。

将两个view的内容都视作文本词频。矩阵E：期望（平均）意义上两个view中两个词的共现概率（频率）。矩阵A：两个view上的topic之间的共现概率。P、Q每个view上，每个topic对应的词频概率。

1. 数值算法
   1. 初始化：用k-means聚类出的各类中心作为初始的每个topic上对应的词频概率，初始化的P、Q矩阵。
   2. 固定P、Q不动，用最小二乘法计算A。
   3. 固定A、Q不动，计算P。
   4. 固定A、P不动，计算Q。
   5. 计算残差 ，如果没有小到预设的值且没有超过最大迭代次数，则跳转到b。

残差

1. 用上述办法进行聚类。

聚类先分两个view分别进行。由于P、Q有topic对应词频的概率，我们可以计算出产生一个输入数据的概率。假定所有输入数据的原始文档等长，我们可以计算出对数下的生成输入数据的概率。对于P、Q中为零的项，则全部拉到1e-5。对于概率最大的那个topic，我们就认为该文档属于它。从而我们得到在两个view下各自的聚类结果。将两者组成二元组，每个二元组就对应一个原object所在的类。

1. 用矩阵分解法测试聚类的数量。

令输入数据的两个view相同，且令P=Q，进行矩阵分解之后A矩阵代表的就是该view下每个topic的自相关。对于那些和为0的行，我们可以认为这个topic是不必要的。

1. 用矩阵分解法进行feature selection

在第4点的约束下，令P=Q为一对角阵，对角线上的每个元素代表一个词在文本中的总体概率，以此计算出的A矩阵中，和为0的行（列）对应的词可以认为是不必要的，可以从特征中去除。

实验结果：

1. 聚类结果
   1. 检验。用两种办法：一是与人工分类的结果比较。二是对原始输入数据（股价）二等分，将两次聚类的结果进行比较。
   2. 两分法。一种是将股价按照前后等分成两份，另一种是用奇数日和偶数日。显然，用前后等分的先决条件是，股票所在的聚类在给定的时间长度上（两年）不会发生变化。而用奇偶法则没有那么严格的要求。
   3. 聚类比较。对半M和对半K分别是 矩阵分解 和 K-MEANS 用前一半的数据和后一半数据聚类结果的差异程度。

人工1M和人工1K分别是 矩阵分解 和 K-MEANS 前一半数据和人工分类结果的差异程度。2M和2K分别是后一半。S100没有人工分类的信息。

差异程度的度量：还是先计数，两个聚类结果有一个对应情况的矩阵，比如说对于S12-09，矩阵分解的两次聚类结果是：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 5 | 5 | 5 | 5 | 15 | 15 | 15 | 10 | 4 | 4 | 4 | 4 |
| 16 | 16 | 16 | 16 | 6 | 6 | 6 | 6 | 9 | 3 | 9 | 9 |

然后我们可以得到一个计数矩阵：

[5, 16] = 4

[15, 6] = 3

[4, 9] = 3

[10, 6] = 1

[4, 3] = 1

排序后，我们把5和16进行对应，15和6对应，4和9对应，[10,6]是异常，因为第二次的6已经和第一次的15对应掉了；[4,3]是一个异常，因为第一次的4已经和第二次的9对应掉了

所以一共有2个不同的分类结果，除以12得到0.1667。

* 1. 结果。
     1. 前后分（最多迭代1000次）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 对半M | 对半K | 人工1M | 人工1K | 人工2M | 人工2K |
| S12-07 | 0.5000 | **0.2500** | **0.1667** | 0.2500 | **0.3333** | 0.4167 |
| S12-08 | 0.3333 | **0.1667** | 0.2500 | 0.2500 | **0.0833** | 0.3333 |
| S12-09 | **0.1667** | 0.3333 | **0.0833** | 0.3333 | **0.0833** | 0.4167 |
| S50-09 | **0.1800** | 0.3600 | **0.1000** | 0.3200 | **0.2200** | 0.3800 |
| S50-10 | **0.2200** | 0.5000 | **0.2400** | 0.3000 | **0.1200** | 0.4800 |
| R50-09 | **0.3200** | 0.5200 | **0.2200** | 0.6400 | **0.3000** | 0.5800 |
| R50-10 | 0.4800 | **0.4600** | **0.3000** | 0.6800 | **0.4000** | 0.5800 |
| S100-10 | 0.6200 | **0.4500** | / | / | / | / |

* + 1. 奇偶分（最多迭代1000次）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 奇偶M | 奇偶K | 人工1M | 人工1K | 人工2M | 人工2K |
| S12-07 | **0** | 0.2500 | **0** | 0.1667 | **0** | 0.2500 |
| S12-08 | **0** | 0.3333 | **0** | 0.2500 | **0** | 0.2500 |
| S12-09 | **0** | 0.1667 | **0** | 0.2500 | **0** | 0.2500 |
| S50-09 | **0.1200** | 0.2800 | **0.1800** | 0.4200 | **0.2200** | 0.3400 |
| S50-10 | **0.0800** | 0.2600 | **0.0800** | 0.4200 | **0.0400** | 0.4200 |
| R50-09 | **0.3400** | 0.3800 | **0.3200** | 0.5400 | **0.3000** | 0.6400 |
| R50-10 | **0.2800** | 0.5000 | **0.3000** | 0.5200 | **0.3200** | 0.6200 |
| S100-10 | 0.5700 | **0.5000** | / | / | / | / |

1. Feature selection结果

在文本上进行feature selection。在100个词上，选择出：信监管 人民币 放款 非贸易 汇兑 托业 外币 见证 存款 贴现 000004 000007 000008 基地 计算机网 禽畜 生产 000017 运动 000001 汇款 借款 结算 技术 经营 销售 资信 票据 产品