Matrix Factorization Method

朱恬骅
09300240004 计算机科学与技术

September 9, 2012

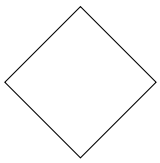
1 Background

Under many circumstances two or more prospectives are possible for observing one series of objects. There might be several situations about the dependence existing amonng different observations. In this paper, we will propose an approach to cope with the following two situations: First, observation data from two perspectives are both available, and we want to utilitize them for possible clusering or classification accuracy improvement. Second, some parts of the data are missing, where just one perspective is available, say X, and we want to estimate the observation for other perspective Y, in order to utilitize an existing and good clustering/classification method on Y.

In this paper, our data is extracted from the Chinese stock market, and the perspectives involved includes (1) price data, (2) news about every stock, and (3) business sphere description.

2 Matrix Factorization Method

2.1 Generative Model



We will discuss the first situation first, that is, exploit observations obtained from two perspectives, to enhance clustering/classification precision. In topic models, we have assumed that a document is generated from several chosen topics, with appropriate proportion. Hierarchical topic models have

assumed farther that some hierarchical structure exists among these topics, and Dirichlet process is introduced in order to solve situations involving infinite or undetermined topics.

In this paper, we combine the idea of hierarchical topic models with normal model. Our assumption is based on the following fact: an object itself is the determinant factor in generating the observations under different perspectives. An individual object, or in our context, stock, is the explicit factor that differs among every possible datum.

To simplify our discussion, we will first consider situations where two perspectives are involved. Denote X for the observations results from the first perspective and Y for the second, and we have correspondingly two sets of topics, named W and Z, who are latent random variables that determined a distribution over the word frequency in every document, i.e. X and Y. The two sets of topics are by iteslf independent, for that we have no physically meaningful measures to apply on the price data and news text. Every generative process, from W to get X, or draw Y from Z, are independent. The de facto dependence between X and Y is caused by the way we draw W and Z and assign proportions to them.

In latent Dirichet allocation, we have known that W and Z each are determined by its Dirichlet parameter, namely αs. Here we introduce a hidden random variable, h, which served as a determinant for choosing appropriate proportion over W and Z.In the reality, h is the abstraction from objects. Thus, the clustering of observation (X, Y) is actually the clustering over different draws from h.

2.2 Matrix Factorization Method

Tranditional topic model may also assume that, two observations, or two sets of feature properties, are results of independent topics. Further more, any two topics are irrelevant. However, in the reality, such constraints seldom hold. Two topics may cover a mutual set of conditions, and two obseravtions, if they are correspondent to one same object, may interfere each other, or display some mutual features. In such situations, we cannot adopt independent identical distribution assumption any longer.

Instead, we introduce a matrix, called topical co-occurence matrix. This matrix, denoted as A in the following sections, describes the posibility of two topics from distinct views occuring in the same context. That is, the relevance between Xs and Ys are assumed to attribute to the relevance between their own generative topics.

Given some observed vectors $X_1, X_2, ..., X_n \in \mathbb{R}^p$ and $Y_1, Y_2, ..., Y_n \in \mathbb{R}^q$, each X_i and Y_i are features for one object, we want to reduce the number

of representation dimensions, by using its topic information. For the reason that topic variables are hidden, thus inaccessible to the observers, we can only estimate the matrix A by factorizing the expectation matrix of the product of the two observation vectors, i.e. $E = \mathbf{E}XY^\top$. In our model, we have two sets of topics, say T_X and T_Y, and note $s = |T_X|$, $t = |T_Y|$, generating results according to the topic-result possiblity matrices P_X and P_Y (and sometimes, generating several results, and the observation is the sum). Then we have the following equation and constraints held:

$$E \simeq PAQ^\top \tag{1}$$

$$\sum_{i,j} A_{ij} = 1 \tag{2}$$

$$\sum_j P_{ij} = 1, \forall i \tag{3}$$

$$\sum_j Q_{ij} = 1, \forall i \tag{4}$$

where $A \in \mathbb{R}^{s\times t}$, and p, q are the length of observation vectors from X and Y, repectively. In order to dispel scale difference, Xs and Ys are normalized to $norm(\mathbf{.}, 1) = 1$.

2.3 Estimating matrix A

We denote words domain in X as $\Sigma = \{\sigma_1, \sigma_2, \sigma_3, ...\}$ and that in Y as $T = \{\tau_1, \tau_2, \tau_3, ...\}$, and the probability of observing σ_i in X and τ_j in the corresponding Y is estimated by $E_{ij} = \frac{1}{N}\sum_{k=1}^{N} X_i^{(k)} Y_j^{(k)}$.

Assume that the distribution of word σ_i on topic $w \in W$ is given by $p_X(i|w)$, and $p_Y(j|z)$ depicted word τ_j on topic $z \in Z$, we have $E_{ij} = p_X(i|w)p(w, z)p_Y(j|z)$, where $p(w, z)$ denotes the topical co-occurence probability of w and z, which is obtainable from A.

It is evident that any distribution of $p_X(i|w)$ or $p_Y(j|z)$ is possible, amongst which we consider multinomial as the most appropriate one. The distribution may degenerate to a skew distribution when the dictionary, or say σs and τs are not appropriately chosen, and to a uniform distribution when the dictionary is built on some keywords.

The distribution of $p_X(i|w)$ and $p_Y(j|z)$ can be estimated by using EM method, which can form matricies P and Q. Then we can get A by calculatng

$$A = \arg\min_A S(E - PAQ^\top) \tag{5}$$

where S the square error function.

2.4 Describing h by using A

For the situations where multiple observations for each object is possible, we can get several (X, Y) from one same object. Denote its mean observation as $(\bar{X}, \bar{Y})$. To exploit them, we can use A to depict h, whereby at the same time elimate differences in observation samples. In this case, we use the globally trained P and Q, which is obtained from $\bar{X}$s and $\bar{Y}$s. We calculates E using the observations of the same object, and A for this object can be also estimated by using equation 5. Hence, we have a matrix to indicate the object, and such matricies resembles when their owner objects are similar. We may then adopt clustering or classifciation over these A matricies, by using 3D k-means or other tensor methods. A is the distribution by which we draw w and z from the topic sets W and Z.

2.5 Knowledge transfering

In this subsection, we will discuss the second problem, that is, to estimate reasonable values for the missing perspective Y from available data X.

Since we have estimated the probability of σ_i co-occurring with τ_j as in matrix E, and P and Q for distribution over words given some topic, it is easy to estimate $p(w, z)$ by adopting equation 5. Hence, for the incoming observation $X = (p(\sigma_1), p(\sigma_2), p(\sigma_3), ...)$, our estimation for the probability of the unseen observation Y is generate by topic z, i.e., $p_Y(z|X)$ is given by the following equation:

$$p_Y(z|X) = \sum_w \frac{p(w|X)p(w, z)}{p_Y(z)} \tag{6}$$

And we may further imply the possible Y is that $\{\sum_z p_Y(z|X)p_Y(j|z)\}_{j=1}^q$. In many cases where topics are used to reduce dimension, using $p_Y(z|X)$ is more suitable.

3 Dataset

3.1 Small-scale dataset

Select four stocks from each of banking, pharmaceutical and steel industries, to compose the small-scale dataset, named S12.

1. Banking 600015 华夏银行Huaxia Bank

   600000 浦发银行Pufa Bank

600016 民生银行Minsheng Bank

600036 招商银行Zhaoshang Bank

2. Pharmaceutical

600085 同仁堂Tongrentang

600129 太极集团Taiji Group

6000422 昆明制药Kunming Pharm

000423 东阿阿胶Dong'e E-jiao

3. Steel

600019 宝钢股份Baosteel

000959 首钢股份Shougang Steel

000709 河北钢铁Hebei Steel

600569 安阳钢铁Anyang Steel

3.2 Medium-scale dataset

3.2.1 50 stocks from 5 irrelevant industries

Noted as S50.

1. Banking

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. Steel

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. Pharmaceutical

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. Wine and Alcoholic Drinks
   - 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. Software Development
   - 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

3.2.2 50 stocks from 5 relevant industries

Noted as R50.

1. Steel
   - 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. Charcoal
   - 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. Vehicle manufacturing
   - 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. Airlines and Aircraft Industry

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. Power

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

3.3 Large-scale dataset

3.3.1 S100 and S500

Select randomly 100 and 500 stocks, to compose the S100 and S500 dataset.

3.4 Notes on time periods

We denote the time period by adding a suffix composed of a hyphen followed by two digits indicating the starting year. The time period is always chosen from July 1 of that year to June 30 two years later. For example, S12-09 indicates the dataset contains information generated in the time period starting from July 1, 2009 to June 30, 2011, relating to the chosen 12 stocks.

4 Experiments on single-view dataset

4.1 股价信息

4.1.1 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。

4.1.2 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。

4.2 文本信息（经营范围描述）

4.2.1 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。

4.2.2 聚类方法

同4.1.2 。

4.3 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。