基于鲁棒非负矩阵分解的聚类方法研究

答辩人： 杜 亮
导 师： 沈一栋 研究员
专 业： 计算机软件与理论
时 间： 2013年5月27日

计算机科学国家重点实验室
STATE KEY LABORATORY OF COMPUTER SCIENCE
ISCAS

大纲

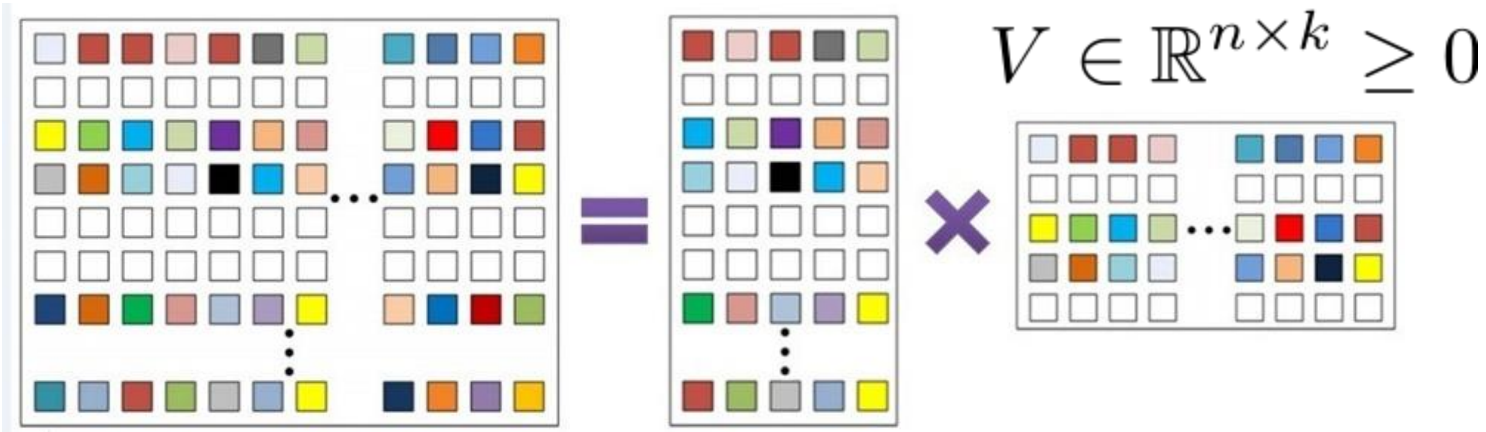
- 研究背景
- 研究内容
  - 鲁棒非负矩阵分解
  - 鲁棒联合聚类
  - 区间矩阵分解
  - 加权图正则非负矩阵分解
- 总结

研究背景

- 非负矩阵分解[NMF, Nature-99]

输入数据　　基因子(basis)　　编码因子

$$X \in \mathbb{R}^{d \times n} \geq 0 \quad U \in \mathbb{R}^{d \times k} \geq 0 \quad V \in \mathbb{R}^{n \times k} \geq 0$$



$$X \approx UV^T$$

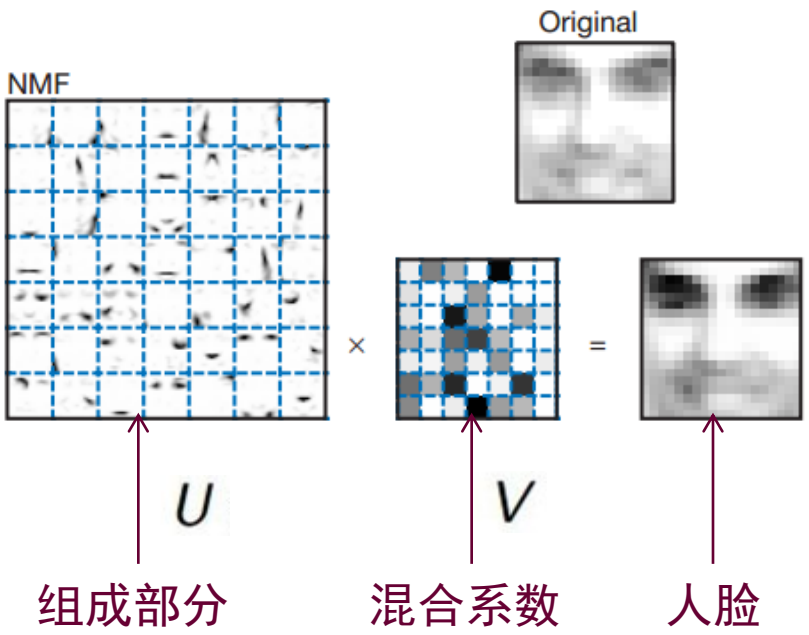
- 非负数据是普遍存在的
  - 文本(TF/IDF)、图像(Pixel)、基因、网络

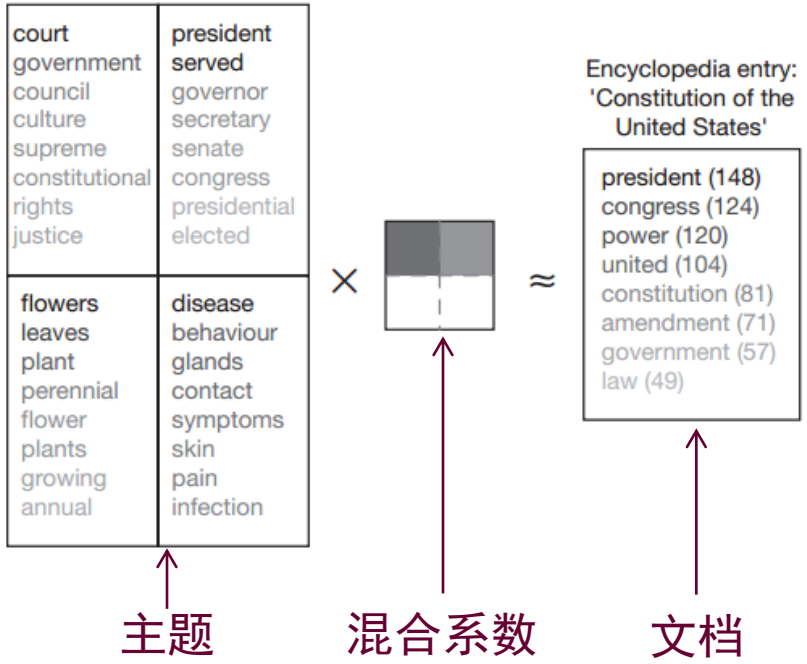
研究背景

非负矩阵分解应用(1)

学习基于局部特征的数据表示

- 人脸由鼻子、嘴巴、眼睛等部分构成
- 文档由主题构成

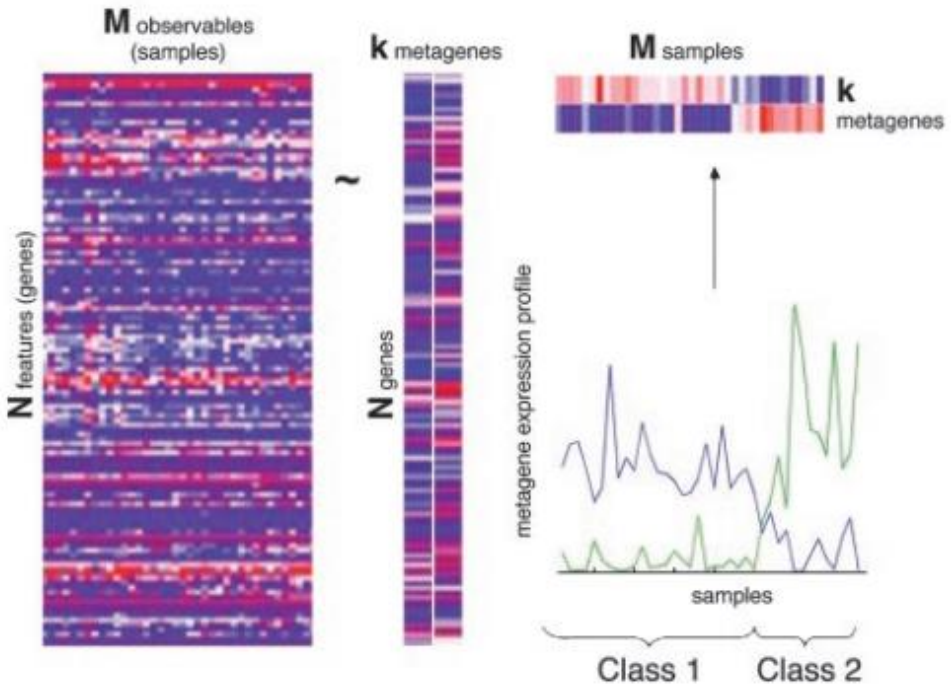




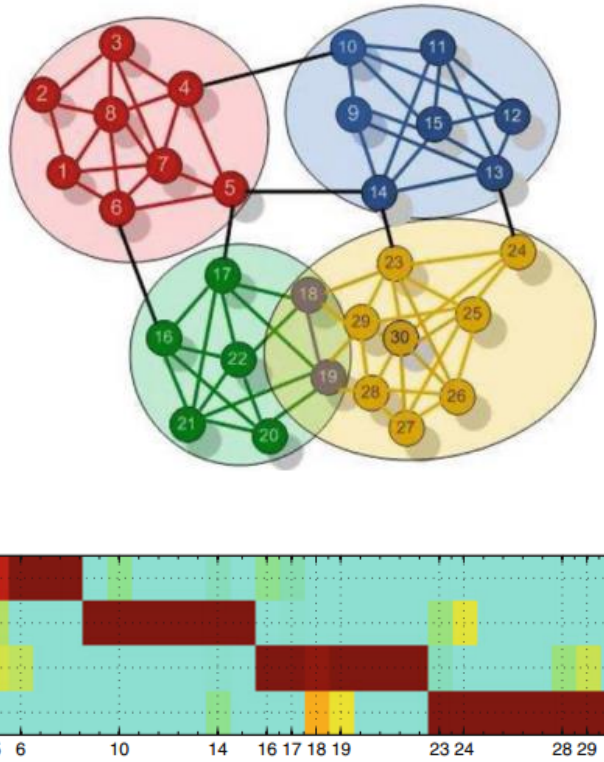
研究背景

非负矩阵分解应用(2)

聚类分析



基因聚类，PNAS-04



网络聚类，DMKD-10

研究背景

非负矩阵分解研究现状

- 损失函数
  - 测度近似的好坏
  - 常见损失：l2(平方损失)，KL失真度等

$$X \approx UV^T$$

- 受限的非负矩阵分解
  - 稀疏性约束
  - 正交、对称性约束
  - 流形约束
  - 判别式约束

$$\min \quad \sum_i \sum_j (X_{ij} - (UV^T)_{ij})^2$$

$$\text{s.t.} \quad U \geq 0, V \geq 0$$

- 优化方法
  - 学习非负因子
  - 如：乘法更新公式

$$U_{ik} = U_{ik} \frac{(XV)_{ik}}{(UV^TV)_{ik}}$$

$$V_{jk} = V_{jk} \frac{(X^TU)_{jk}}{(VU^TU)_{jk}}$$

动机

- **数据质量问题**
  - 数据质量问题是普遍存在的
    - 图像受光照、遮挡等因素干扰
    - 文本中垃圾邮件、网页作弊
    - 基因数据中的测量误差
  - 常见形式
    - 噪声
      - 高斯、泊松、拉普拉斯噪声
      - 偏差期望值较大的噪声(gross error)
      - 结构性：稀疏噪声、异常样本、异常特征
    - 误差
      - 数据误差/图误差

现有非负矩阵分解及其扩展模型缺乏有效的方法处理各类数据质量问题！

目的和意义

- 提高非负矩阵分解的鲁棒性
- 提高聚类分析的效果

大纲

- 研究背景
- 研究内容
  - 鲁棒非负矩阵分解
  - 鲁棒联合聚类
  - 区间矩阵分解
  - 加权图正则非负矩阵分解
- 总结

动机

- **噪声环境下的标准非负矩阵分解**

$$X \approx UV^T$$

$$\begin{aligned} \min \quad & \sum_i \sum_j (X_{ij} - (UV^T)_{ij})^2 \\ \text{s.t.} \quad & U \geq 0, V \geq 0 \end{aligned}$$

观测数据　未观测到的真实数据　实际情况

$$X_{ij} = \bar{X}_{ij} + \epsilon_{ij}$$ ← 噪声

$$X = \bar{X} + \epsilon \approx UV^T$$

$$\bar{X} \approx UV^T$$ ← 理想情况

$$\min \quad \sum_{i=1} \sum_{j=1} (\bar{X}_{ij} + \epsilon_{ij} - (UV^T)_{ij})^2$$

动机

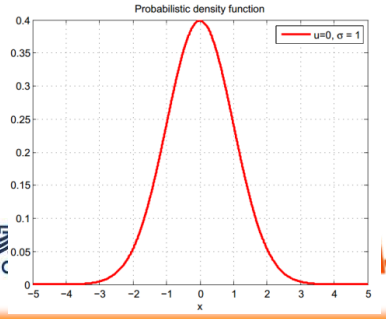
- 标准非负矩阵分解对非高斯噪声敏感
  - 重构误差最小化角度

$$\min_{U,V} \sum_i \sum_j (\bar{X}_{ij} + \epsilon_{ij} - (UV^T)_{ij})^2 = \sum_i \sum_j (E_{ij}^2 - 2E_{ij}\epsilon_{ij} + \epsilon_{ij}^2)$$

    - 噪声较小，甚至接近0时，重构误差支配损失函数
    - 噪声较大时，大噪声支配损失函数！
      - ➔ 单个大噪声都可能任意改变分解结果
  - 概率似然最大化角度
    - 噪声服从0均值高斯分布时

$$\sum_{i=1} \sum_{j=1} (\bar{X}_{ij} + \epsilon_{ij} - (UV^T)_{ij})^2 = \sum_{i=1} \sum_{j=1} (\bar{X}_{ij} - (UV^T)_{ij})^2$$

    - 99.7%的噪声高度集中于均值附近
      - ➔ 当噪声大量远离均值时，噪声不可忽略



鲁棒损失函数

- 标准非负矩阵分解 -> 鲁棒非负矩阵分解
  - 标准二次损失函数(平方损失) -> 鲁棒损失函数

鲁棒损失函数度量矩阵分解质量！

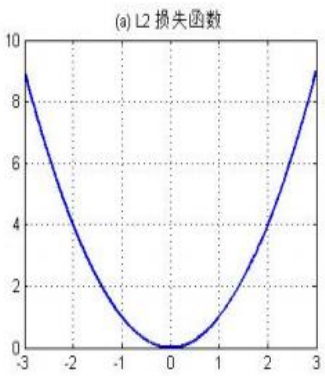
- 基本假设
  - 一个好的非负矩阵分解模型在大噪声部分损失较大

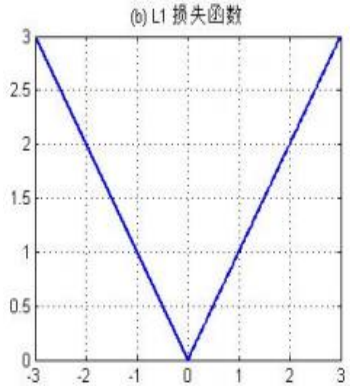
$$\sum_{i=1}\sum_{j=1}(X_{ij}-(UV^T)_{ij})^2=\begin{cases}((\bar{X}_{ij}-(UV^T)_{ij}+\epsilon_{ij})^2 & \text{小噪声，损失小}\\(\bar{X}_{ij}-(UV^T)_{ij}+\epsilon_{ij})^2 & \text{大噪声，损失大}\end{cases}$$

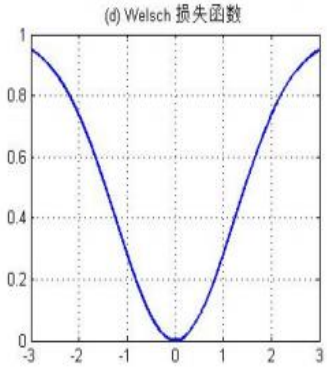
- 思路
  - 选择在大误差部分损失较小的函数
    - 降低大误差部分对整个模型的影响
  - 选择在长尾数据概率较高的分布
    - 提高长尾 (离均值较远)噪声的生成概率

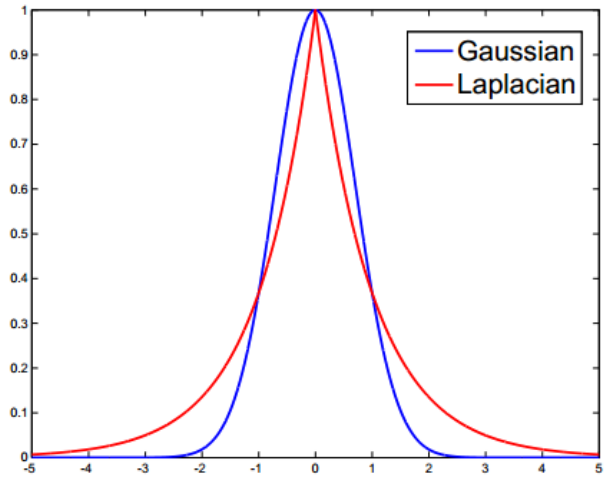
常见的鲁棒损失和分布

- 大误差损失较小的函数
  - L1损失、相关熵失真度、Huber函数、超曲面函数，等
- 长尾概率较高的分布
  - 拉普拉斯分布、学生t-分布、柯西分布、威布尔分布、帕累托分布，等









拉普拉斯分布等价于L1损失

鲁棒损失函数优化

- 问题
  - 如何求解不同鲁棒损失函数引起的非负矩阵分解问题？
    - One by one ？
  - 如何选择合适的鲁棒损失函数?
    - 不同数据集和任务需要不同的损失函数
- 思路
  - 选择一类鲁棒损失函数
    - 半二次损失函数
  - 设计通用优化方法求解
    - 半二次最小化方法

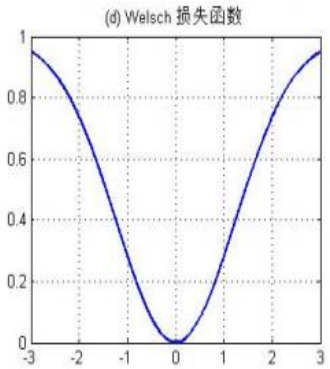
半二次最小化鲁棒非负矩阵分解

- 采用半二次函数测度非负矩阵分解质量
  - 半二次损失函数？
    - 非二次损失函数、对大误差鲁棒性好
    - 可以通过半二次最小化求解
  - 常见的半二次函数包括：l1函数、相关熵失真度、Huber函数、鲁棒M-估计量，等
- 半二次最小化技术求解
  - 通过引入辅助变量，非二次损失函数的优化问题转化为二次优化问题（如乘法形式和加法形式的二次优化问题）
  - 通用性好、速度快

例子：CIM-NMF

- 相关熵失真度(CIM)

$$\mathrm{CIM}(x,\hat{x}) = 1 - \exp(-\frac{(x-\hat{x})^2}{\sigma^2})$$



- 基于CIM的非负矩阵分解(CIM-NMF)

$$\min_{U,V} \quad \sum_{i=1}^{N}\sum_{j=1}^{M}(1 - g(X_{ij} - \sum_{k=1}^{K} U_{ik}V_{jk}, \sigma))$$

$$\text{s.t.} \quad U \geq 0, V \geq 0.$$

CIM-NMF 优化方法(1)

- 基于半二次优化乘法形式的噪声检测算法

引入辅助变量W　　对偶函数

$$\min_{U,V,W} \sum_{i=1}^{N}\sum_{j=1}^{M}[W_{ij}(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk})^2+\phi(W_{ij})]$$

- 计算辅助变量

$$W_{ij}=\begin{cases}\ell''(E_{ij}), \text{if } E_{ij}=0\\ \frac{\ell'(E_{ij})}{E_{ij}}, \text{otherwise}\end{cases} \longrightarrow W_{ij}=\exp(-\frac{(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk})^2}{2\sigma^2})$$

噪声检测：误差越大，权重越小

- 计算非负因子

$$\min_{U,V} \sum_{i=1}^{N}\sum_{j=1}^{M}W_{ij}(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk})^2 \quad \text{s.t.} \quad U\geq 0, V\geq 0,$$

$$U_{ik}=U_{ik}\frac{(W\otimes XV)_{ik}}{(W\otimes(UV^T)V)_{ik}}$$

$$V_{jk}=V_{jk}\frac{((W\otimes X)^TU)_{jk}}{((W\otimes(UV^T))^TU)_{jk}}$$

CIM-NMF 优化方法(2)

- 基于半二次优化加法形式的噪声矫正算法

引入辅助变量S　　对偶函数

$$\min_{U,V,S} \sum_{i=1}^{N}\sum_{j=1}^{M}\left((X_{ij}-(UV)_{ij}-S_{ij})^2+\phi(S_{ij})\right)$$

- 计算辅助变量

$$S_{ij}=E_{ij}-\ell'(E_{ij}) \longrightarrow S_{ij}=E_{ij}-\frac{2E_{ij}}{\sigma^2}\exp(-\frac{E_{ij}^2}{\sigma^2})$$

噪声矫正：误差越大，噪声越大

- 计算非负因子

$$\min \quad ||X-S-UV^T||^2$$
$$\text{s.t.} \quad U\geq 0, V\geq 0,$$

$$U_{ij}=U_{ij}\sqrt{\frac{[(X-S)V]_{ij}^{+}}{[UV^TV+((X-S)V)^{-}]_{ij}}}$$

$$V_{ij}=V_{ij}\sqrt{\frac{[(X-S)^TU]_{ij}^{+}}{[VU^TU+((X-S)^TU)^{-}]_{ij}}}$$

其他例子

- **基于Huber损失函数的鲁棒非负矩阵分解**

$$\ell_{\text{huber}}(e) = \begin{cases} e^2 & \text{if } |e| \le c \\ 2c|e| - c^2 & \text{if } |e| \ge c \end{cases} \qquad \min_{U,V} \sum_{i=1}^{N}\sum_{j=1}^{M} \ell_{\text{huber}}(E_{ij})$$

- **处理异常样本/特征的鲁棒非负矩阵分解方法**
  - 每一行/列看作一个整体

$$\min_{U,V} \sum_{i=1}^{N}(1 - g(||X_{i*} - U_{i*}V^T||, \sigma)) \qquad \text{rCIM-NMF}$$

- **现有鲁棒非负矩阵分解方法是基于半二次最小化鲁棒非负矩阵分解方法的特例**
  - **L1-L2-NMF [TSP-06];L21-NMF [CIKM-11];SR-NMF [FEEE-11]**

基于非负矩阵分解的聚类分析

- 数据簇数目==编码因子的个数
  - Max算子
  - 其他聚类算法(如Kmeans)
- 数据簇数目!=编码因子的个数
  - 其他聚类算法(如Kmeans)

$$X \approx UV^T$$

非负矩阵分解结果既可以直接用于聚类;
也可以看成一种新的表示形式,作为其他聚类方法的输入;

实验结果(1)

通用数据集聚类结果对比

Data set	COIL20	JAFFE	CSTR	WebKB
Kmeans	0.631	0.657	0.727	0.520
PCA-Km	0.638	0.780	0.749	0.527
RPCA-Km	0.652	0.753	0.703	0.615
Ncut	0.380	0.795	0.714	0.430
NMF	0.651	0.861	0.758	0.557
SR-NMF	**0.671**	0.839	0.777	0.603
$L_{2,1}$-NMF	0.658	0.893	0.771	0.614
WFS-NMF	0.649	0.879	**0.784**	**0.664**
Huber-NMF	0.661	**0.928**	0.765	0.617
rCIM-NMF	**0.695**	0.882	**0.798**	**0.675**
CIM-NMF	0.670	**0.927**	0.761	0.652

Data set	COIL20	JAFFE	CSTR	WebKB
Kmeans	0.743	0.745	0.651	0.042
PCA-Km	0.743	0.821	0.663	0.056
RPCA-Km	0.755	0.831	0.603	0.022
Ncut	0.578	0.833	0.638	0.151
NMF	0.679	0.859	0.665	0.155
SR-NMF	**0.758**	0.862	0.687	**0.176**
$L_{2,1}$-NMF	0.713	0.908	0.681	0.157
WFS-NMF	0.740	0.872	**0.687**	0.013
Huber-NMF	0.743	**0.932**	0.675	0.172
rCIM-NMF	**0.755**	0.897	**0.691**	**0.177**
CIM-NMF	0.753	**0.942**	0.668	0.153

鲁棒非负矩阵分解方法优于标准NMF

真实数据集存在一定的非高斯噪声！

不同鲁棒方法在不同数据表现并不一致

根据实际数据和任务选择合适的鲁棒模型

实验结果(2)

噪声数据集上数据重构

CIM-NMF可以准确的检测并恢复干净数据

图 2.4: (a) 被破坏的原始图像，(b)、(c)、(d)、(e) 通过NMF、SR-NMF、$L_{2,1}$-NMF和CIM-NMF重构得到的图像，(f)CIM-NMF学习得到的权重矩阵

实验结果(3)

➢ 噪声数据集上聚类结果对比

$r(\%)$	Kmeans	PCA-Km	RPCA-Km	Ncut	NMF	SR-NMF	$L_{2,1}$-NMF	WFS-NMF	Huber-NMF	rCIM-NMF	CIM-NMF
5	51.1±2.9	50.6±1.2	52.1±2.5	44.2±2.9	54.7±3.2	58.2±2.1	56.0±3.4	56.5±3.0	57.1±2.9	56.7±3.3	**61.0±2.1**
10	48.7±2.5	48.4±2.6	48.8±1.9	34.6±1.5	51.5±3.0	54.3±2.7	52.9±2.8	53.6±2.7	55.4±2.2	53.9±2.7	**60.7±3.5**
15	45.5±2.6	45.4±2.0	45.7±2.0	33.8±2.4	49.6±1.7	50.5±3.0	49.9±2.6	50.5±2.5	51.8±3.0	50.6±2.8	**58.8±3.3**
20	43.3±2.4	43.7±2.4	43.8±2.6	35.0±2.2	45.9±2.5	48.0±2.2	47.1±2.5	47.8±2.1	50.2±2.5	47.8±2.3	**57.2±2.6**
25	40.5±2.3	41.5±2.3	41.5±1.7	35.4±2.0	44.2±2.8	46.1±1.8	44.9±2.7	45.5±3.4	47.6±1.8	44.6±2.5	**54.8±3.5**
30	39.1±2.2	39.8±1.7	39.2±1.7	35.1±1.6	42.2±2.8	42.7±2.7	42.5±2.2	42.1±2.8	44.5±2.7	42.1±2.5	**51.6±3.2**
35	37.1±1.9	37.3±1.7	38.1±2.1	34.4±1.8	38.7±1.9	40.6±2.0	39.4±2.1	39.6±2.4	41.3±2.3	40.6±2.0	**47.1±2.4**
40	34.6±1.6	35.0±1.7	35.8±1.7	32.8±1.5	36.9±2.2	38.0±2.0	37.1±1.7	38.7±2.5	38.4±1.6	37.0±2.1	**43.0±2.7**
45	34.1±1.5	34.0±1.4	34.5±1.5	33.3±1.3	36.2±2.1	37.1±1.3	37.9±2.5	36.1±3.1	37.6±2.2	36.8±1.7	**42.0±2.8**
50	32.1±1.6	32.8±1.5	33.2±1.3	32.3±1.7	34.4±1.8	35.4±2.0	35.0±2.1	35.4±2.4	35.9±2.0	35.0±2.1	**39.7±2.1**
Avg.	40.6	40.8	41.3	35.1	43.4	45.1	44.3	44.6	46.0	44.5	**51.6**

CIM-NMF优于其他方法

大纲

- 研究背景
- 研究内容
  - 鲁棒非负矩阵分解
  - 鲁棒联合聚类
  - 区间矩阵分解
  - 加权图正则非负矩阵分解
- 总结

基于非负矩阵分解的联合聚类

- **联合聚类**
  - 同时进行样本聚类和特征聚类
- **图正则非负矩阵三因子分解 [KDD-09;IJCAI-11;PR-12]**
  - 样本-特征关系 => 最小化矩阵分解重构误差
  - 样本-样本关系 => 最小化图正则误差
  - 特征-特征关系 => 最小化图正则误差

重构误差　　特征图正则　　样本图正则

$$\sum_{i=1}^{d}\sum_{j=1}^{n}(X_{ij}-(FHG^T)_{ij})^2+\lambda_F\sum_{i=1}^{d}\sum_{i'=1}^{d}W_{ii'}^F||F_{i\cdot}-F_{i\cdot}||^2+\lambda_G\sum_{j=1}^{n}\sum_{j'=1}^{n}W_{jj'}^G||G_{j\cdot}-G_{j'\cdot}||^2$$

鲁棒联合聚类

- 动机：现有非负矩阵分解联合聚类方法对噪声敏感
  - 平方重构误差对样本-特征关系中非高斯噪声敏感
  - 平方图正则误差对样本-样本(以及特征-特征)噪声关系敏感
- 思路
  - 同时提高矩阵重构和图正则的鲁棒性

鲁棒联合聚类

- **特征-样本矩阵重构**
  - 假设噪声是稀疏的
  - 引入稀疏噪声矩阵，矫正恢复干净数据

$$||X - FHG^T - S||_F^2 + \lambda_S \boxed{||S||_1}$$ ← 噪声矩阵

- **样本-样本(特征-特征)图正则**
  - 假设噪声图正则误差较大
  - 选择大误差损失较小的函数，减轻噪声图正则的影响

$$\sum_{jj'} \begin{cases} W_{jj'}||G_j - G_{j'}||^2 & \text{非噪声正则，误差较小} \\ W_{jj'}||G_j - G_{j'}||^2 & \text{噪声正则，误差较大} \end{cases}$$

$$\Longrightarrow \sum_{jj'} \begin{cases} W_{jj'}||G_j - G_{j'}|| \\ W_{jj'}||G_j - G_{j'}|| \end{cases} = \sum_{jj'} W^G_{jj'}||G_j - G_{j'}||$$

问题形式化

$$\mathcal{J}_1 = ||X - FHG^T - S||_F^2 + \lambda_S||S||_1$$
$$+ \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2 + \lambda_G \sum_{jj'} W_{jj'}^G ||G_{j\cdot} - G_{j'\cdot}||_2$$
$$\text{s.t.} \quad F, H, G \geq 0, F\mathbf{1}_{k1} = \mathbf{1}_d, G\mathbf{1}_{k2} = \mathbf{1}_n,$$

- **变量**
  - 稀疏噪声 **S**
  - 特征簇 **F**
  - 样本簇 **G**
  - 特征簇-样本簇关联因子 **H**

优化方法：分块坐标下降法

优化方法(1)

- 计算稀疏噪声(变量S)子问题

$$\mathcal{J}_2 = ||X - FHG^T - S||_F^2 + \lambda_S||S||_1$$

  - 分解为单变量的优化问题
    - 软门限算子计算最优解

$$S_{ij} = \begin{cases} 0 & \text{if } |E_{ij}| \leq \lambda_S/2, \\ E_{ij} - \frac{\lambda_S}{2}\text{sign}(E_{ij}) & \text{otherwise}, \end{cases}$$

小误差，噪声为0

优化方法(2-1)

- 计算特征簇(变量F)子问题

$$\mathcal{J}_3 = ||X - FHG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2$$

$$\text{s.t.} \quad F \geq 0, F\mathbf{1}_{k_1} = \mathbf{1}_d.$$

  - 难点1：l1图正则不可微
    - 将l1重写为加权的l2图正则

引入辅助变量

$$\text{引入}\widetilde{W}_{ii'}^F = \frac{W_{ii'}^F}{2||F_{i\cdot} - F_{i'\cdot}||_2}, \widetilde{D}_{ii}^F = \sum_{i'} \widetilde{W}_{ii'}^F, \quad P_{ij}^- = (|P_{ij}| - P_{ij})/2$$

$$P = (X - S)GH^T, \quad Q = HG^TGH^T \quad P_{ij}^+ = (|P_{ij}| + P_{ij})/2$$

优化方法(2-2)

- 计算特征簇(变量F)子问题

$$\mathcal{J}_4 = \mathrm{tr}(-2F^T P^+ + 2F^T P^- + FQF^T + \lambda_F F^T \widetilde{D}^F F - \lambda_F F^T \widetilde{W}^F F)$$

$$\text{s.t.} \quad F \geq 0, F\mathbf{1}_{k_1} = \mathbf{1}_d,$$

  - 难点2：归一化约束

引入辅助变量 $\widetilde{F} \in \mathbb{R}^{d \times k_1}$

满足归一化约束的F可由辅助变量计算 $F_{ik} = \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}$

优化方法(2-3)

- 计算特征簇(变量F)子问题

$$\mathcal{J}_5(\widetilde{F}) = -2\sum_{ik} P_{ik}^{+}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} + 2\sum_{ik} P_{ik}^{-}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} + \sum_{ikk'} Q_{kk'}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}} +$$
$$\lambda_F \sum_{ii'k} (\widetilde{L}_F)_{ii'}^{+}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}\frac{\widetilde{F}_{i'k}}{\sum_s \widetilde{F}_{is}} - \lambda_F \sum_{ii'k} (\widetilde{L}_F)_{ii'}^{-}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}\frac{\widetilde{F}_{i'k}}{\sum_s \widetilde{F}_{is}}, \qquad (3.12)$$

  - 非负矩阵分解中常用的辅助函数法求解

    构造辅助函数 $Z(\widetilde{F}, \widetilde{F}^t)$ 见Eq.(3.13), p53

    根据一阶最优性条件：辅助变量导数为零 $\frac{\partial Z(\widetilde{F}, \widetilde{F}^t)}{\partial \widetilde{F}_{ik}} = \sum_{k'} \frac{\partial Z(\widetilde{F}, \widetilde{F}^t)}{\partial (\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})} \frac{\partial (\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})}{\partial \widetilde{F}_{ik}} = 0$

    结合F与辅助变量的关系 $F_{ik} = \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}$

  - 难点3：没有闭式解

优化方法(2-4)

- 计算特征簇(变量F)子问题

$$A_{ik}F_{ik}^2+\sum_s[C_{is}-A_{is}F_{is}^2]F_{ik}-C_{ik}=0$$

$$A=(P^-+F^tQ+\lambda_F\widetilde{D}^FF^t)\oslash F^t$$

$$C=(P^++\lambda_F\widetilde{W}^FF^t)\odot F^t$$

  - 计算F的不动点
    - 采用文献[94]的方法

优化方法(3)

- 计算簇关联因子(变量H)子问题

$$\mathcal{J}_6 = ||X - FHG^T - S||_F^2, \quad \text{s.t.} \quad H \geq 0$$

  - 更新公式

$$H_{ij} = H_{ij}\sqrt{\frac{[F^T(X-S)G]_{ij}^+}{[F^TFHG^TG]_{ij} + [F^T(X-S)G]_{ij}^-}}$$

- 计算样本簇 (变量G)子问题

$$\mathcal{J}_7 = ||X^T - GH^TF^T - S^T||_F^2 + \lambda_G \sum_{jj'} W_{jj'}^G ||G_{j\cdot} - G_{j'\cdot}||_2$$

$$\text{s.t.} \quad G \geq 0, G\mathbf{1}_{k_2} = \mathbf{1}_n.$$

  - 变量G与变量F对偶，采用变量F的计算方法更新

实验结果(1)

聚类准确性

	JAFFE	MFEA	OPTDIGIT	LUNG	GLIOMA
Kmeans	0.7239	0.6862	0.7280	0.6911	0.5520
NCut	0.8392	0.7296	0.7855	0.7667	0.5680
ONMTF	0.7777	0.6511	0.6733	0.6931	0.5640
GNMF	0.9371	0.9158	0.7908	0.7697	0.5800
DRCC	0.9305	0.8759	0.7854	0.7719	0.5800
DNMTF	0.9305	0.9255	0.8055	0.8266	0.5800
IGNMTF	0.9256	0.8358	0.7910	0.8162	0.5640
LDCC	0.8249	0.7974	0.7700	0.7574	0.6520
RCC	**0.9765**	**0.9438**	**0.8558**	**0.8808**	**0.6840**

RCC优于当前主流联合聚类算法

实验结果(2)

- 归一化互信息

	JAFFE	MFEA	OPTDIGIT	LUNG	GLIOMA
Kmeans	0.7942	0.7016	0.7314	0.5069	0.4710
NCut	0.7974	0.8320	0.7203	0.5982	0.4630
ONMTF	0.8188	0.6189	0.6269	0.5137	0.4405
GNMF	0.9384	0.9051	0.8220	0.6670	0.3523
DRCC	0.9480	0.8257	0.8306	0.5552	0.4841
DNMTF	0.9507	0.9088	0.8361	0.5977	0.5047
IGNMTF	0.9175	0.8528	0.8301	0.5786	0.4880
LDCC	0.8798	0.8400	0.7929	0.5595	0.4979
RCC	**0.9756**	**0.9117**	**0.8611**	**0.6909**	**0.5350**

RCC优于当前主流联合聚类算法

鲁棒联合聚类方法的半二次扩展

- 动机：
  - 改善鲁棒联合聚类算法稀疏噪声假设的局限性
    - 得到应用面更广的鲁棒联合聚类方法
- 思路
  - 证明鲁棒联合聚类算法等价于采用特定的半二次损失函数
  - 采用一般的半二次损失函数测度非负矩阵分解质量，设计通用的优化方法

鲁棒联合聚类方法的半二次扩展

- **鲁棒联合聚类等价于特定的半二次损失函数**
  - 稀疏噪声假设等价于 **Huber**损失的加法形式

基于L1的稀疏噪声

$$||X - FHG^T - S||_F^2 + \lambda_S||S||_1$$

$$S_{ij} = \begin{cases} 0 & \text{if } |E_{ij}| \le \lambda_S/2, \\ E_{ij} - \frac{\lambda_S}{2}\text{sign}(E_{ij}) & \text{otherwise,} \end{cases}$$

Huber损失

$$\mathcal{J} = \begin{cases} E_{ij}^2 & \text{if } |E_{ij}| \le \lambda_S/2, \\ \lambda_S|E_{ij}| - (\frac{\lambda_S}{2})^2 & \text{otherwise.} \end{cases}$$

$$\ell_{Huber}(E_{ij}) = \sum_{ij} \left\{ (E_{ij} - S_{ij})^2 + g_A(S_{ij}) \right\},$$

半二次加法

$$S_{ij} = \begin{cases} 0 & |E_{ij}| \le c \\ E_{ij} - c\text{sign}(E_{ij}) & |E_{ij}| > c \end{cases}$$

  - **L1图正则是一种半二次正则**

鲁棒联合聚类方法的半二次扩展

- 基于半二次损失函数的鲁棒联合聚类方法
  - 半二次损失函数测度重构误差
  - 半二次损失函数测度图正则误差
- 基于半二次最小化的优化方法
  - 基于半二次乘法形式
  - 基于半二次加法形式

大纲

- 研究背景
- 研究内容
  - 鲁棒非负矩阵分解
  - 鲁棒联合聚类
  - 区间矩阵分解
  - 加权图正则非负矩阵分解
- 总结

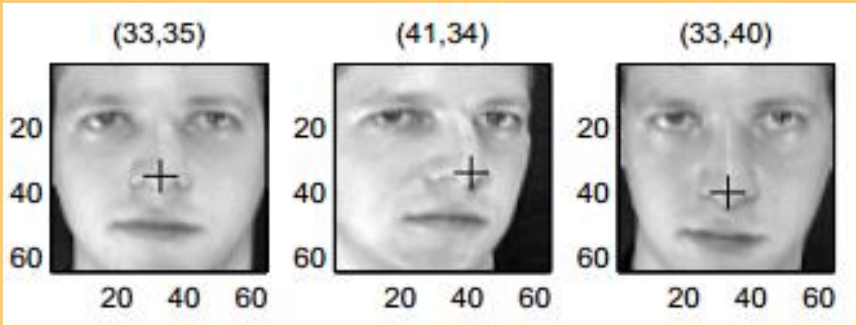
区间矩阵分解

动机

观测数据的不确定性

- 单值观测数据中不可避免的误差

应用：人脸分析



理想数据：同一个像素对应相同位置

应用：协同过滤

	m_1	m_2	m_3	m_4	m_5
u_1		1	4		5
u_2	3		1	2	
u_3		1		1.8 ? 2.2	4
u_4	5				
u_5		1	4	2	
u_6		3		2	5

理想数据：评分更准确刻画偏好

单值表示无法刻画理想数据

区间矩阵分解

- 思路
  - 利用某种概率分布刻画观测数据的不确定性
- 方法
  - 假设观测数据的每个元素来自某个均匀分布
  - 构造每个元素的均匀分布对应的上下界
  - 构造整个观测数据基于均匀分布上下界的区间表示

$$X_{ij} \xrightarrow{X_{ij} \sim \text{uniform}(X_{ij}^{\text{low}}, X_{ij}^{\text{up}})} [X_{ij}^{low}, X_{ij}^{up}]$$

均值：X_{ij}

半径：δ_{ij}

$$X_{ij}^{low} = X_{ij} - \delta_{ij}$$

$$X_{ij}^{up} = X_{ij} + \delta_{ij}$$

通过周围的观测数据估计

区间矩阵分解

- **区间矩阵联合分解**
  - 数据由其上下界构造 $E(X_{ij}) = \frac{1}{2}(X_{ij}^{\text{low}} + X_{ij}^{\text{up}})$
  - 上下界矩阵联合分解，共享编码矩阵**U**

$$\hat{X}^{\text{low}} \leftarrow \boxed{U}V^{\text{low}} \quad \hat{X}^{\text{up}} \leftarrow \boxed{U}V^{\text{up}}$$

- **区间非负矩阵分解**

$$\mathcal{L}_{\text{I-NMF}} = ||X^{\text{low}} - UV^{\text{low}}||^2 + ||X^{\text{up}} - UV^{\text{up}}||_F^2$$

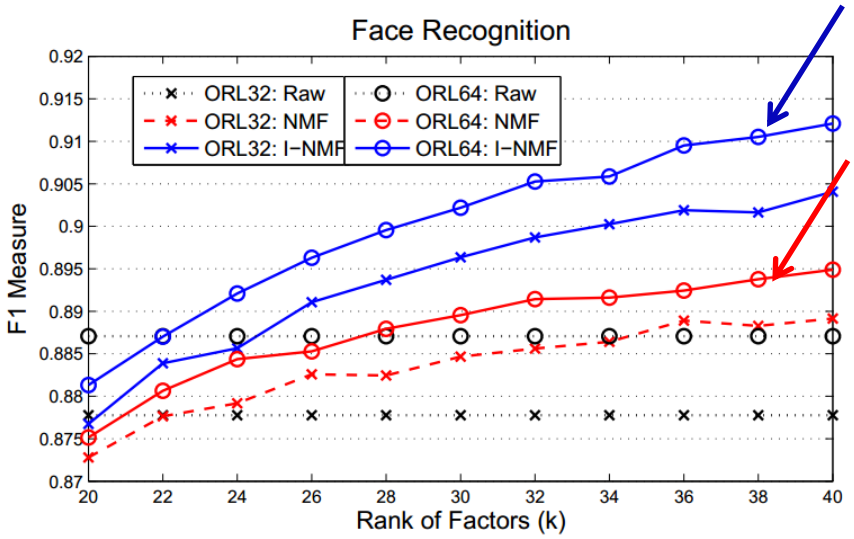
$$\text{s.t.} \quad U \geq 0, V^{\text{low}} \geq 0, V^{\text{up}} \geq 0$$

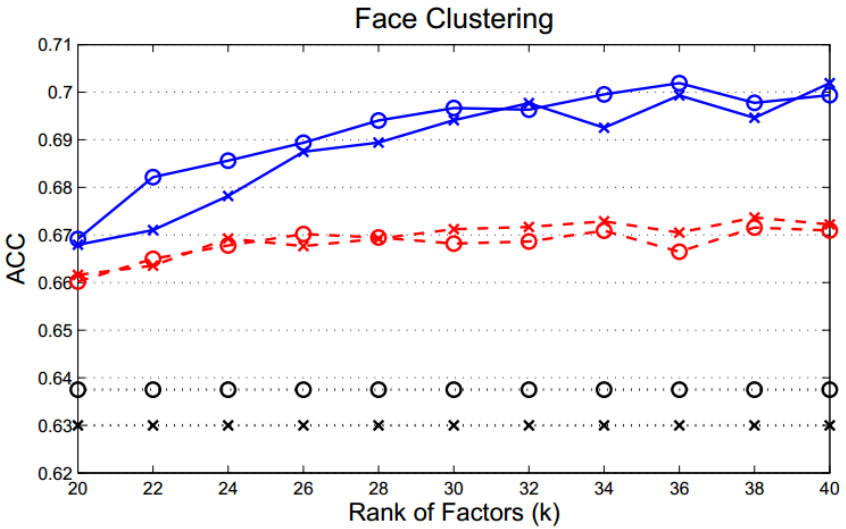
- **区间概率矩阵分解**

$$\mathcal{L}_{\text{I-PMF}} = ||X^{\text{low}} - UV^{\text{low}}||^2 + ||X^{\text{up}} - UV^{\text{up}}||_F^2 + \lambda(||U||^2 + ||V^{\text{low}}||^2 + ||V^{\text{up}}||^2)$$

实验结果(1)

- 人脸识别/聚类

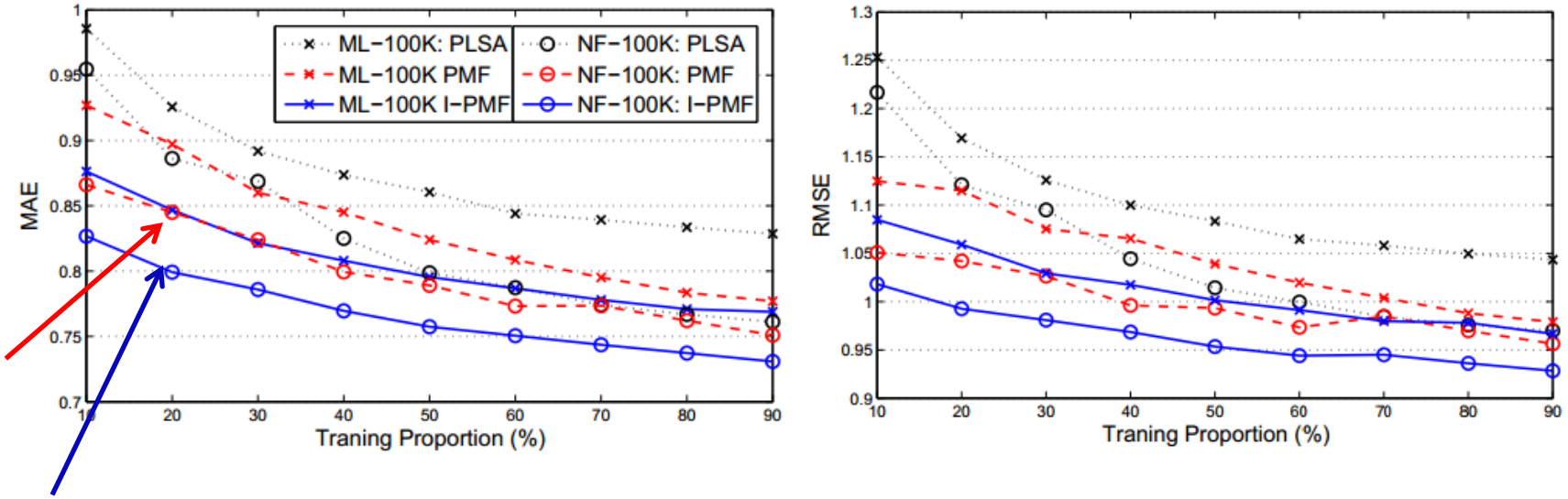




区间非负矩阵分解优于标准非负矩阵分解

实验结果(2)

- 协同过滤



区间概率矩阵分解优于标准概率矩阵分解

大纲

- 研究背景
- 研究内容
  - 鲁棒非负矩阵分解
  - 鲁棒联合聚类
  - 区间矩阵分解
  - 加权图正则非负矩阵分解
- 总结

加权图正则非负矩阵分解

- 背景：数据间的相似性对聚类结果非常重要
- 动机：数据相似关系中的不确定性
  - 单个关系图无法准确刻画数据相似性
- 聚类分析中的多关系问题
  - 可以构造多种关系图刻画数据相似性
    - 图构造：K-近邻准则、epsilon-球邻域准则
    - 边权配置：0-1权重、高斯核、逆欧式距离，等
- 聚类集成中的多关系问题
  - 将多个聚类结果合并成一个聚类结果
  - 每个聚类结果都可以用一个关系图表示

加权图正则非负矩阵分解

- **思路**
  - 利用多关系的线性组合近似数据的相似性

$$\hat{W} = \sum_{c=1}^{m} \boldsymbol{\alpha}_c W^c, \quad \sum_{c=1}^{m} \boldsymbol{\alpha}_c = 1, \boldsymbol{\alpha}_c \geq 0$$

集成关系　　单个关系矩阵

- **方法**
  - 基于欧式距离和KL失真度的多关系图正则

$$\mathcal{R}_{\mathrm{Euc}} = \frac{1}{2} \sum_{c=1}^{m} \sum_{i=1}^{n} \sum_{i'=1}^{n} \boldsymbol{\alpha}_c W_{ii'}^c ||U_{i\cdot} - U_{i'\cdot}||^2 = \mathrm{tr}(U^T (\sum_{c=1}^{m} \boldsymbol{\alpha}_c L^c) U)$$

$$\mathcal{R}_{\mathrm{KL}} = \frac{1}{2} \sum_{c=1}^{m} \sum_{i=1}^{n} \sum_{i'=1}^{n} \boldsymbol{\alpha}_c W_{ii}^c (\mathrm{kl}(U_{i\cdot}||U_{i'\cdot}) + \mathrm{kl}(U_{i'\cdot}||U_{i\cdot}))$$

加权图正则非负矩阵分解

- **基于欧式距离**

$$\mathcal{L}_{\text{WGNMF-Euc}} = ||X - UV^T||^2 + \lambda \text{tr}(U^T(\sum_{c=1}^{m} \boldsymbol{\alpha}_c L^c)U)$$

$$\text{s.t.} \quad U \geq 0, V \geq 0, \sum_{c=1}^{m} \boldsymbol{\alpha}_c = 1, \boldsymbol{\alpha}_c \geq 0$$

- **基于KL失真度**

$$\mathcal{L}_{\text{WGNMF-KL}} = \sum_{i=1}^{n}\sum_{j=1}^{d}(X_{ij}\log\frac{X_{ij}}{\sum_{l=1}^{k}U_{il}V_{jl}} - X_{ij} + \sum_{l=1}^{k}U_{il}V_{jl})$$

$$+ \frac{\lambda}{2}\sum_{c=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{l=1}^{k}\boldsymbol{\alpha}_c W_{ij}^c(U_{il}\log\frac{U_{il}}{U_{jl}} + U_{jl}\log\frac{U_{jl}}{U_{il}})$$

$$\text{s.t.} \quad U \geq 0, V \geq 0, \sum_{c=1}^{m} \boldsymbol{\alpha}_c = 1, \boldsymbol{\alpha}_c \geq 0$$

实验结果(1)

聚类集成

- 集成结果：准确性，归一化互信息
- 对比方法：图划分方法，概率生成方法，矩阵分解方法

	Iris	Glass	Ecoli	Soybean	Zoo
Kmeans	0.81	0.42	0.65	0.72	0.69
KC	0.85	0.49	0.64	0.65	0.67
CSPA	0.87	0.43	0.56	0.69	0.58
HGPA	0.62	0.40	0.51	0.72	0.55
MCLA	0.89	0.46	0.61	0.73	0.74
BCE	0.89	0.49	0.66	0.73	0.74
GWCA	0.89	0.53	0.64	0.73	0.74
WGNMF-Euc	0.89	0.54	0.67	0.75	0.77
WGNMF-KL	0.89	0.51	0.65	0.73	0.73

	Iris	Glass	Ecoli	Soybean	Zoo
Kmeans	0.69	0.31	0.58	0.72	0.69
KC	0.72	0.33	0.59	0.67	0.70
CSPA	0.71	0.29	0.51	0.63	0.59
HGPA	0.39	0.26	0.40	0.69	0.60
MCLA	0.74	0.32	0.56	0.71	0.74
BCE	0.74	0.35	0.59	0.69	0.70
GWCA	0.75	0.37	0.57	0.71	0.73
WGNMF-Euc	0.74	0.38	0.59	0.73	0.74
WGNMF-KL	0.74	0.37	0.58	0.71	0.73

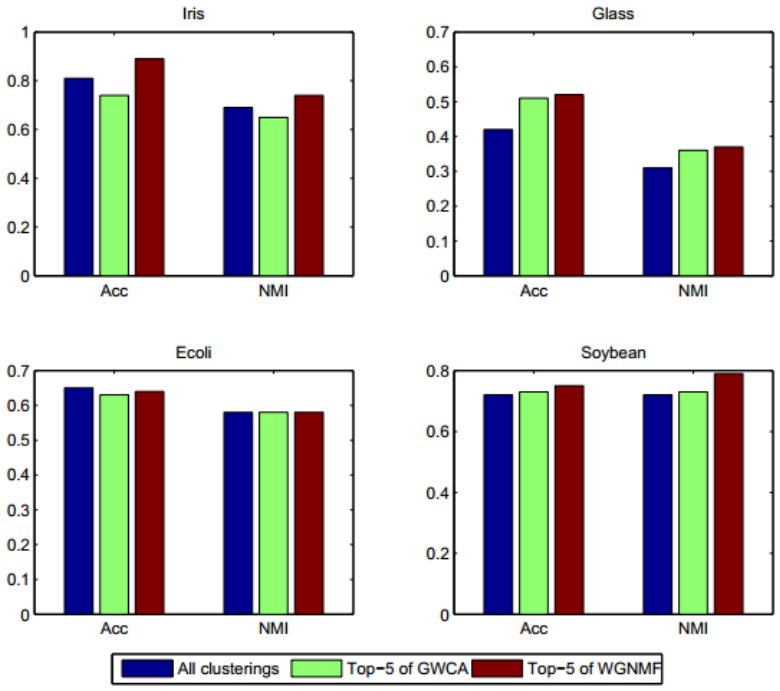
矩阵分解方法由于其他方法

实验结果(2)

单个聚类质量评估和选择

$$\hat{W} = \sum_{c=1}^{m} \alpha_c W^c, \quad \sum_{c=1}^{m} \alpha_c = 1, \alpha_c \geq 0$$

集成关系　　单个关系矩阵



权重越高，聚类质量越好

聚类效果对比：全部聚类vs. Top-5聚类

大纲

- 研究背景
- 研究内容
  - 鲁棒非负矩阵分解
  - 鲁棒联合聚类
  - 区间矩阵分解
  - 加权图正则非负矩阵分解
- 总结

研究成果

- 提出基于半二次最小化的鲁棒非负矩阵分解方法

鲁棒非负矩阵分解

- **ICDM 2012 部分评审意见**
  - **“The authors make *a fundamental contribution* to the discipline.” (本文的工作是对该领域的一个根本性贡献。)**
  - **"The methods and formulations in this paper can have *broader impacts* on various other disciplines.“ (本文提出的方法会对其他相关领域产生广泛的影响。)**

Liang Du, Xuan Li and Yi-Dong Shen. Robust nonnegative matrix Factorization via half-quadratic minimization. ICDM 2012. (Full paper, acceptance rate 81/756 = 10.7%.)

ICDM Student Travel Award

研究成果

- 提出基于半二次最小化的鲁棒非负矩阵分解方法
- 提出基于稀疏噪声的鲁棒联合聚类方法

鲁棒联合聚类

- **IJCAI 2013 部分评审意见**
  - **“The proposed approach can be considered *one of the first attempts* to systematically deal with noisy data.” (本文提出的方法可被认为是最早系统性处理噪声数据非负矩阵分解问题的方法之一。)**
  - **“*It advances the state of art*, at least from the point of view of co-clustering methods based on non-negative matrix tri-factorization.”(它领先并推动了基于非负矩阵分解的联合聚类方面的研究)**

Liang Du and Yi-Dong Shen. Towards Robust Co-Clustering, IJCAI 2013. (Oral paper, acceptance rate 195/1473=13.2%.)

研究成果

- 提出基于半二次最小化的鲁棒非负矩阵分解方法
- 提出基于稀疏噪声的鲁棒联合聚类方法
- 提出区间非负矩阵分解方法
- 提出加权图正则非负矩阵分解方法

区间矩阵分解和加权图正则非负矩阵分解

- **区间矩阵分解方法**
  - 区间非负矩阵分解、区间概率矩阵分解
  - 人脸识别/聚类、协同过滤

Zhiyong Shen, Liang Du, Xukun Shen and Yi-Dong Shen. Interval-valued matrix factorization with applications. ICDM 2010.

- **加权图正则非负矩阵分解**
  - 聚类集成/聚类分析

Liang Du, Xuan Li and Yi-Dong Shen. Cluster ensembles via weighted graph regularized nonnegative matrix factorization. ADMA 2011.

研究成果

- 提出基于半二次最小化的鲁棒非负矩阵分解方法
- 提出基于稀疏噪声的鲁棒联合聚类方法
- 提出区间非负矩阵分解方法
- 提出加权图正则非负矩阵分解方法

已发表论文(1)

1. Liang Du and Yi-Dong Shen. Towards robust co-clustering. The 23rd International Joint Conference on Artificial Intelligence (IJCAI), 2013, (Oral paper, accepted rate 195/1473 = 13.2%).
2. Liang Du, Xuan Li and Yi-Dong Shen. Robust nonnegative matrix factorization via half-quadratic minimization. In Proceedings of IEEE 12th International Conference on Data Mining (ICDM), 2012, pages 201-210. (Full paper, accepted rate 81/756 = 10.7%).
3. Xuan Li, Liang Du and Yi-Dong Shen. Update summarization via graph-based sentence ranking. IEEE Transactions on Knowledge and Data Engineering (TKDE), May 2013, vol.25, no.5, pp.1162-1174.
4. Zhiyong Shen, Liang Du, Xukun Shen and Yi-Dong Shen. Interval-valued matrix factorization with applications. In Proceedings of the IEEE 10th International Conference on Data Mining (ICDM), 2010, pages 1037-1042.

已发表论文(2)

5. Liang Du and Yi-Dong Shen. Joint clustering and feature selection. The 14th International Conference on Web-Age Information Management (WAIM), 2013, (Full paper, Accepted).
6. Liang Du, Yi-Dong Shen, Zhiyong Shen, Jianying Wang and Zhiwu Xu. A self-supervised framework for clustering ensemble. The 14th International Conference on Web-Age Information Management (WAIM), 2013, (Full paper, ccepted).
7. Liang Du, Xuan Li and Yi-Dong Shen. Cluster ensembles via weighted graph regularized nonnegative matrix factorization. Advanced Data Mining and Applications (ADMA), 2011, pages 215-228.
8. Liang Du, Xuan Li and Yi-Dong Shen. User graph regularized pairwise matrix factorization for item recommendation. Advanced Data Mining and Applications (ADMA), 2011, pages 372-385.
9. Xuan Li, Liang Du and Yi-Dong Shen. Graph-based marginal ranking for update summarization. In Proceedings of the Eleventh SIAM International Conference on Data Mining (SDM), 2011, pages 486-497.

已发表论文(3)

10. Xuan Li, Yi-Dong Shen, Liang Du and Chen-Yan Xiong. Exploiting novelty, coverage and balance for topic-focused multi-document summarization. In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM), 2010, pages 1765-1768.
11. Liang Wu, Alvin Chin, Guandong Xu, Liang Du, Xia Wang, Kangjian Meng, Yonggang Guo and Yuanchun Zhou. Who Will Follow Your Shop? Exploiting Multiple Information Sources in Finding Followers. The 18th International Conference on Database Systems for Advanced Applications (DASFAA), 2013, pages 401-415.

致谢

谢谢各位评委老师！

QA