Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

半监督聚类和患者相似性分析

数据挖掘组

硕士研究生 谢崇玮

2022 年 9 月 4 日

内容提要

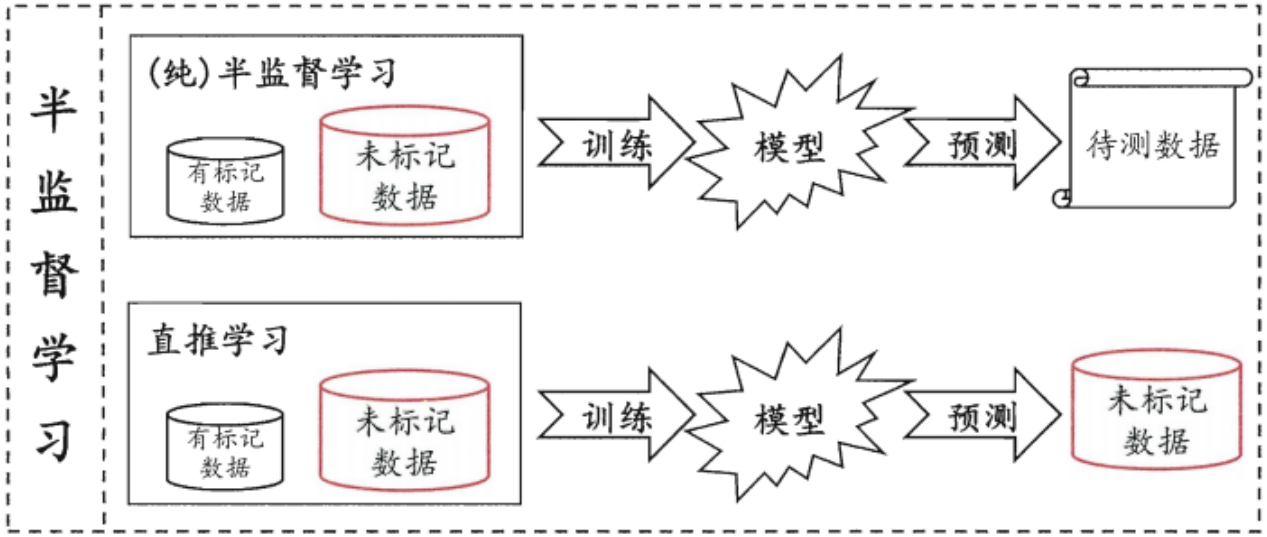
- 背景简介
- 基本概念
- 算法原理
- 总结分析
- 参考文献

内容提要

- 预期收获
  - **1.**了解半监督聚类技术
  - **2.**了解半监督聚类技术在患者相似性分析上的应用

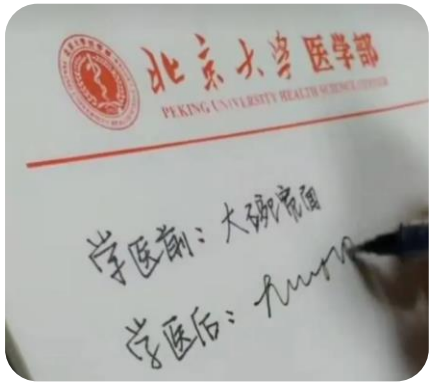
背景介绍

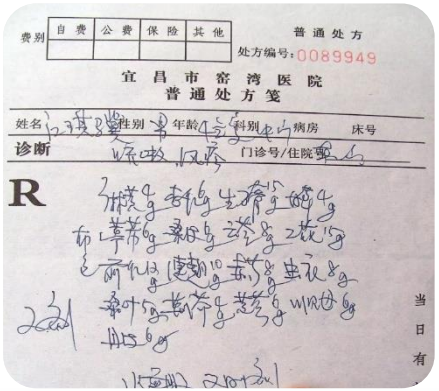
- 半监督学习
  - 大量未标注数据，少量标注数据
  - 仅利用未标注数据聚类分析浪费了宝贵的标注资源
  - 标注数据的获取耗时耗力
- 患者相似性（主要利用电子医疗记录数据）
  - 对患者细粒度区分
  - 构建合适的患者群落划分
  - 为后续医生的诊断和指导提供帮助

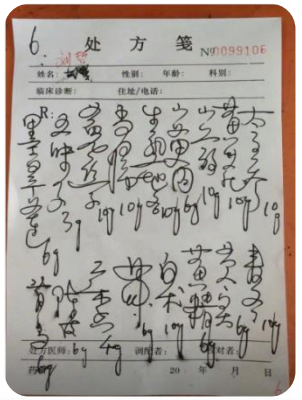


背景介绍

- 电子医疗记录VS纸质记录
  - 结构化、有统一的标准…
  - 非结构化、识别难度大，
  - 重复利用率低…





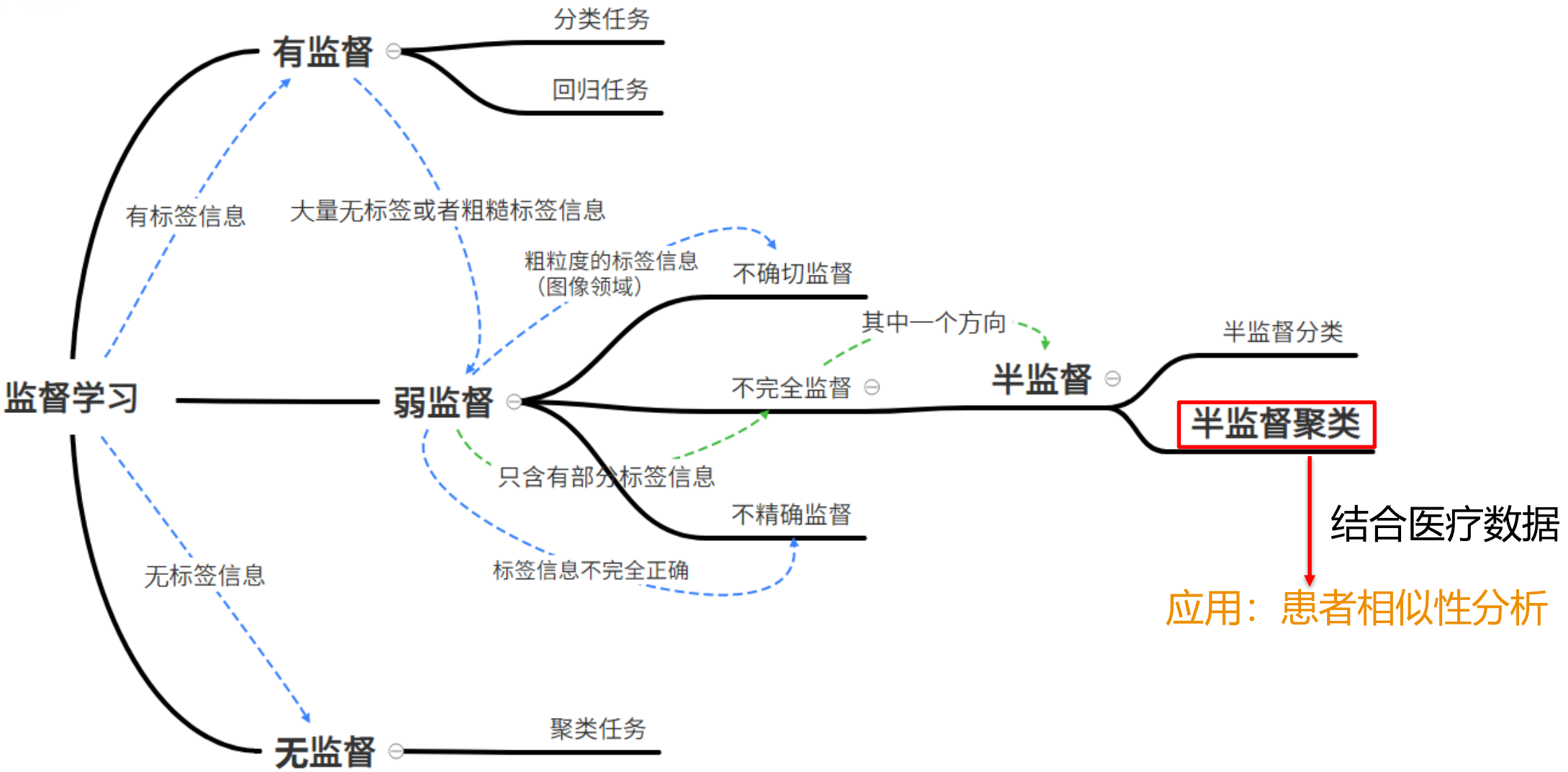


	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息	A_基本信息
2	A6_身份证号码	A1_姓名	A3_性别	A5_出生日期	A24_年龄	健康体检时间	健康评估指导时间	A25_身份证所	A9_家庭类型	A11_月收入	A13_居住地类	A12_职业/前	A2_常住类型	A4_民族	A8_文化程度	A10_1_职业/	A10_2_职业/前	A14_婚姻情况	A15_医疗费用支付	A17_1_有无青	A17_2_有无碘
3	110105196806145509	李宝玲	女	1968-06-14	53	08/24/2021 20:40:32	08/24/2021 20:40:32	北京市	2	9	1	3	1	1	6	1	2	2	1	0	0
4	110107195006171547	李凤梅	女	1950-06-17	71	09/20/2021 19:42:13	09/20/2021 19:42:13	北京市	4	5	1	3	1	1	3	3	2	2	1	0	0
5	110108196408216358	谷仁乔	男	1964-08-21	57	08/23/2021 21:27:42	08/23/2021 21:27:42	北京市	3	8	1	3	1	1	6	1	1	2	1	0	0
6	11010819690625543X	王宝刚	男	1969-06-25	52	08/19/2021 18:31:53	08/19/2021 18:31:53	北京市	2	9	1	3	1	1	6	1	1	2	1	0	0
7	120103195502086411x	罗靖洲	男	1955-02-08	66	08/14/2021 19:34:49	08/14/2021 19:34:49	天津市	2	9	1	3	1	1	5	3	4	2	5	0	0
8	130104195304261318	何平	男	1953-04-26	68	08/13/2021 19:47:41	08/13/2021 19:47:41	河北省	2	9	1	3	1	1	6	3	1	2	7	0	0
9	130404194404170917	李兴余	男	1944-04-17	77	09/13/2021 14:55:31	09/13/2021 14:55:31	河北省	2	4	2	2	1	1	3	3	1	2	1	0	1
10	130426194312181115	吴德全	男	1943-12-18	78	09/03/2021 15:49:55	09/03/2021 15:49:55	河北省	2	4	2	3	2	1	2	3	2	2	1	1	0
11	132336194706040728	孙付爱	女	1947-06-04	74	09/05/2021 17:30:31	09/05/2021 17:30:31	河北省	4	5	1	2	1	1	3	3	4	2	2	0	0
12	132336194810090717	孙书子	男	1948-10-09	73	09/04/2021 16:39:22	09/04/2021 16:39:22	河北省	4	4	1	2	1	1	3	3	4	2	3	0	0
13	132801196412062856	刘尚强	男	1964-12-06	57	08/27/2021 17:08:28	08/27/2021 17:08:28	河北省	2	4	2	3	1	1	5	1	8	2	1	0	0
14	132903196702194740	邓秋斋	女	1967-02-19	54	08/22/2021 18:38:24	08/22/2021 18:38:24	河北省	3	1	4	1	1	1	2	3	5	2	3	0	0
15	13290319670312471x	殷万朝	男	1967-03-12	54	08/22/2021 18:36:04	08/22/2021 18:36:04	河北省	3	2	4	1	1	1	2	3	5	2	3	0	0
16	133030195007010141	王秀芬	女	1950-07-01	71	08/26/2021 18:27:05	08/26/2021 18:27:05	河北省	4	9	3	3	1	1	4	1	4	2	2	0	0
17	133030195010240116	赵云龙	男	1950-10-24	71	08/26/2021 17:16:52	08/26/2021 17:16:52	河北省	4	9	3	2	1	1	3	1	5	2	2	0	0

- 电子病历存在大量未标注数据的问题：
  - 数据标注需要专业人员，且耗时耗力

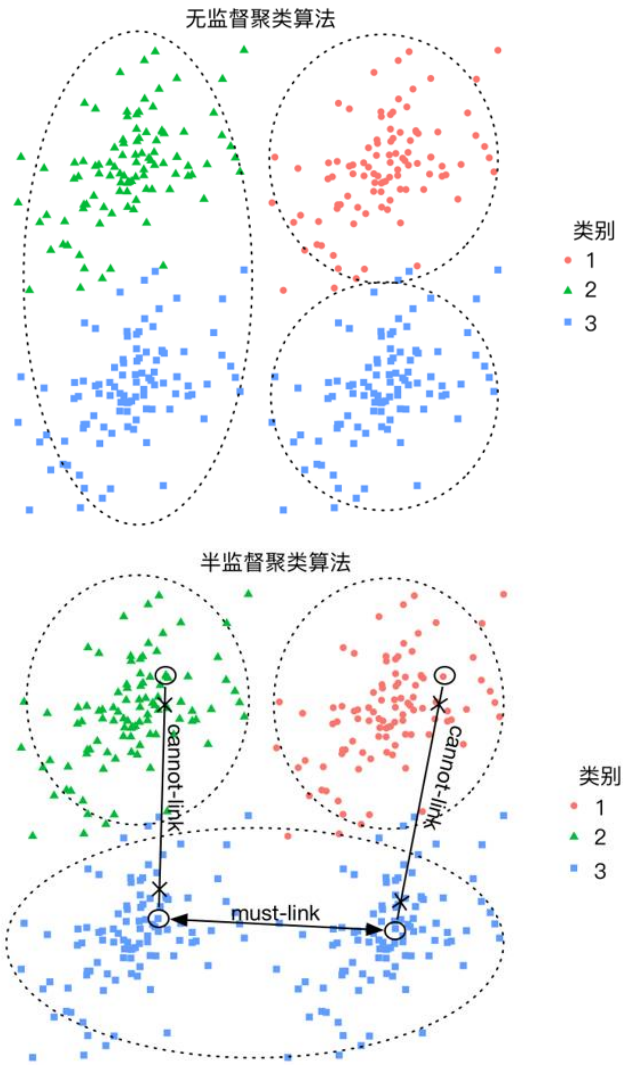
基本概念

基础概念


监督学习
有监督
分类任务
回归任务
有标签信息
弱监督
大量无标签或者粗糙标签信息
粗粒度的标签信息
（图像领域）
不确切监督
其中一个方向
半监督
半监督分类
不完全监督
半监督聚类
只含有部分标签信息
不精确监督
标签信息不完全正确
结合医疗数据
应用：患者相似性分析
无标签信息
无监督
聚类任务

基础概念

- 半监督聚类
  - 结合半监督学习和聚类的方法（引入一些监督信息来指导聚类过程）
- 基于约束的方法
  - 成对约束（Must-link/Cannot-link）
  - 正负样本约束（A属于Q_1类/B不属于Q_2类）
  - 集群约束（簇大小约束、内部分布约束等）
- 基于距离的方法
  - 首先训练距离度量以满足类别或限制信息，然后使用基于距离度量的聚类算法进行聚类
  - 凸优化的马氏距离、由最短路径算法改进的欧式距离、使用梯度下降算法的KL散度、谱聚类方法



基础概念

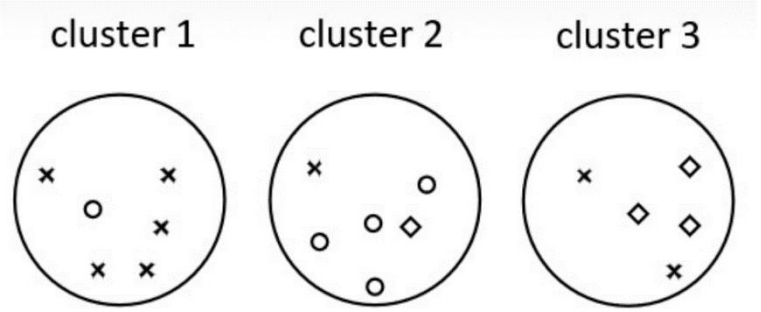
- 聚类效果评判指标
  - 外部指标（监督）

$$RI = \frac{TP + TN}{TP + FP + TF + FN} = \frac{TP + TN}{C_N^2}$$

$$\mathrm{NMI}(\Omega, C) = \frac{I(\Omega; C)}{(H(\Omega) + H(C))/2}$$

  - 内部指标（簇内相似度，簇间分离度）
  - 轮廓系数、SSE

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2, \overline{X}_i = \frac{1}{ni} \sum_{j=1}^{n_i} X_{ij}, i = 1, 2, \ldots, r$$



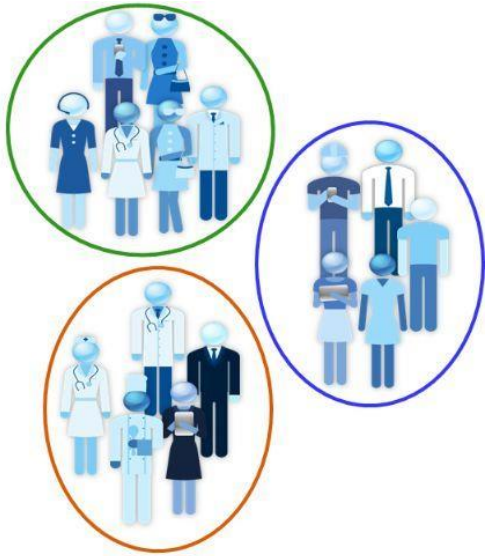
- 聚类中心K的选取标准
  - 随着K值的增大，数据集的划分会更加精确，即SSE会逐渐变小。一般选取SSE随着K值的增大突然大幅下降的位置作为最优K值

	同簇	非同簇
同类	TP	FN
非同类	FP	TN

	同簇	非同簇
同类	TP=20	FN=24
非同类	FP=20	TN=72

基础概念

- 患者相似性分析
  - 选取用户电子病历记录中的特定属性
  - 根据不同类型的特定属性选择不同的相似度计算规则，计算总相似度
  - 根据相似度将患者聚类，找出患者自身与诊断和治疗方案之间的关系



算法原理

算法原理

T	整合多源约束来提高聚类结果的有效性
I	聚类数据集
P	1.通过转换为成对关系矩阵，统一地表示了不同类型的约束 2.整合多源约束指导聚类过程
O	聚类簇

P	算法大多基于单源约束，很少考虑整合多源约束来提高聚类质量
C	多源包含不同类型的约束信息（成对约束、正负标签约束）
D	统一不同类型约束
L	**2021 SCI 1区**

面临问题

- 如何构造统一的表示来保存不同类型的约束?
- 如何整合多源约束来指导聚类过程?

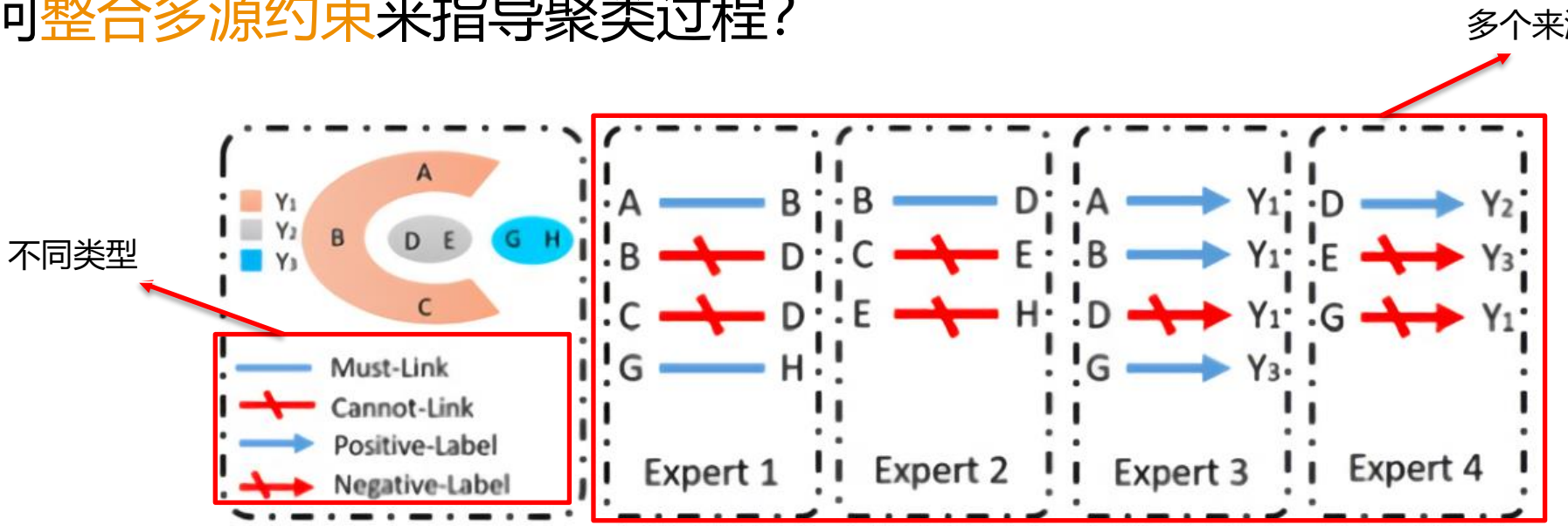


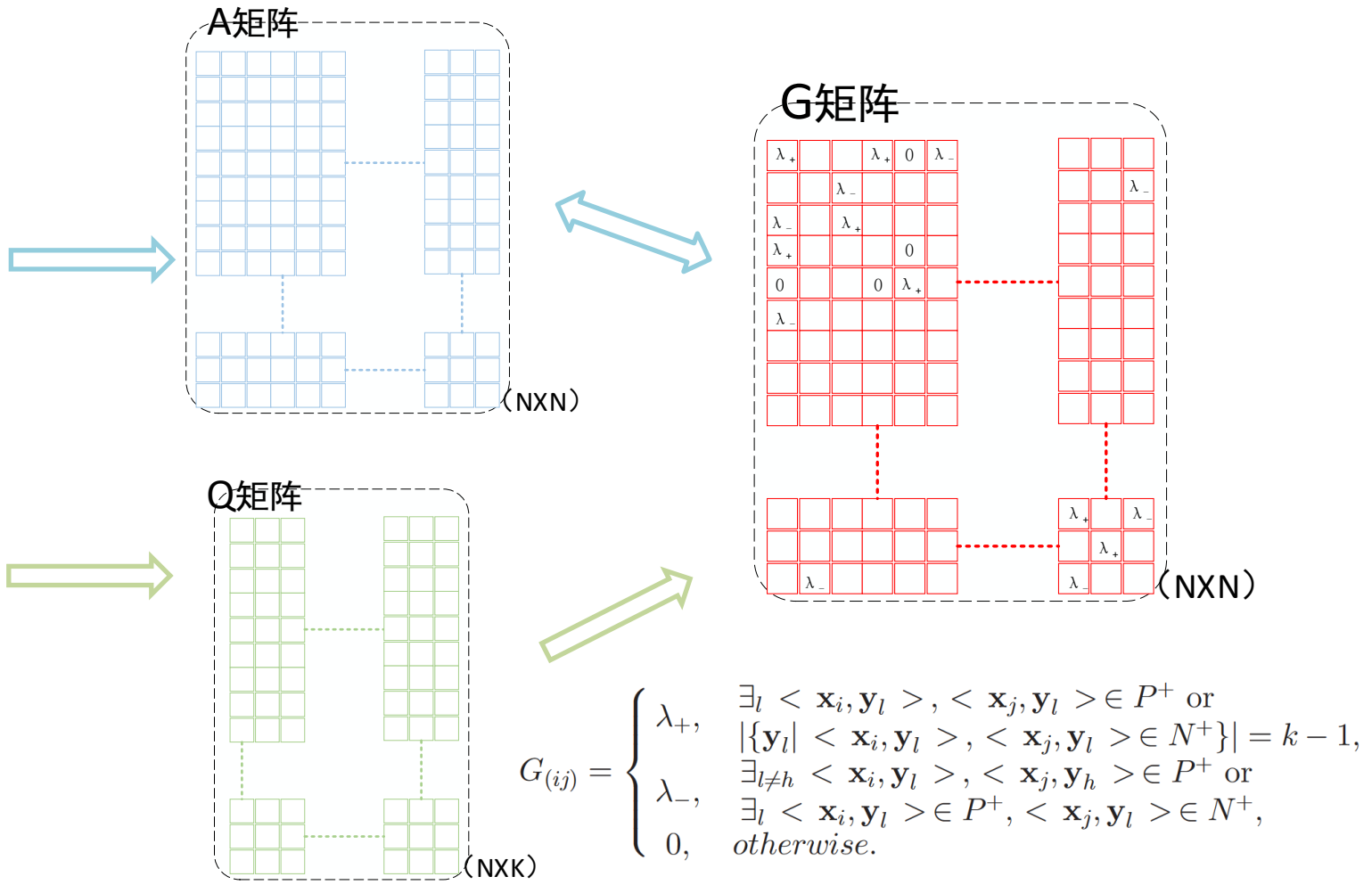
Fig. 1. Example of a data set with 3 clusters and constraints from 4 experts.

算法原理

- 统一表示不同类型的约束
  - 成对约束

    （ML→A_{ij}=正， CL→A_{ij}=负，其它→A_{ij}=0） N*N矩阵
  - 正负例样本

    （正例→ Q_{il}=正，负例→ Q_{il}=负，其它→ Q_{il} =0） N*K矩阵



$$G_{(ij)} = \begin{cases} \lambda_+, & \exists_l < \mathbf{x}_i, \mathbf{y}_l >, < \mathbf{x}_j, \mathbf{y}_l > \in P^+ \text{ or} \\ & |\{\mathbf{y}_l | < \mathbf{x}_i, \mathbf{y}_l >, < \mathbf{x}_j, \mathbf{y}_l > \in N^+\}| = k-1, \\ \lambda_-, & \exists_{l \neq h} < \mathbf{x}_i, \mathbf{y}_l >, < \mathbf{x}_j, \mathbf{y}_h > \in P^+ \text{ or} \\ & \exists_l < \mathbf{x}_i, \mathbf{y}_l > \in P^+, < \mathbf{x}_j, \mathbf{y}_l > \in N^+, \\ 0, & otherwise. \end{cases}$$

算法原理

- 整合多源指导聚类过程
  - 获得m个专家（多源）提供的G矩阵，用U_{n*k}矩阵来示最终聚类结果

$$\min \Theta(U) = F(U) + \alpha E(U)$$

>=0;此处设置=1

$$F(U) = \|\mathbb{X} - UZ\|_F^2$$

$$E(U) = \frac{1}{2}\sum_{t=1}^{m}\|G_t - \hat{U}\hat{U}^T\|_F^2$$ （Z是k*d维中心矩阵，Ū表示U的归一化矩阵）

$$\max_{\hat{U}} Tr\left(\hat{U}^T\left(K + \alpha\sum_{t=1}^{m} G_t\right)\hat{U}\right), \text{ s.t. } \hat{U}^T\hat{U} = I_k.$$

优化目标

$$\lambda_+ = \frac{1}{\gamma_{(ij)}}\left(\max(K) - K_{(ij)}\right) \qquad \lambda_- = \frac{1}{\gamma_{(ij)}}\left(\min(K) - K_{(ij)}\right)$$

（γij是m源中xi和xj之间的约束数量，Kij表示其基于核的特征空间中的相似性）

实验结果

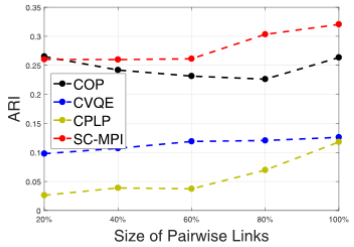
- 实验设置
  - 实验环境：
  - 配备Intel i7-4710MQ CPU@2.5Hz和16GB的RAM的PC上进行
- 实验数据
  - 12个基准数据集
  - 8种对比算法

Data Sets	#Ojbects	#Features	#Classes
Yale [43]	165	1,024	15
ORL [43]	400	1,024	40
Banknote [42]	1,372	4	2
COIL20 [43]	1,440	1,024	20
Isolet [43]	1,560	617	26
OpticalDigits [42]	5,620	64	10
Statlog [42]	6,435	36	6
COIL100 [43]	7,200	1,024	100
MNIST [43]	10,000	784	10
PenDigits [43]	10,992	16	10
USPS [43]	11,000	256	10
Letters [42]	20,000	16	26

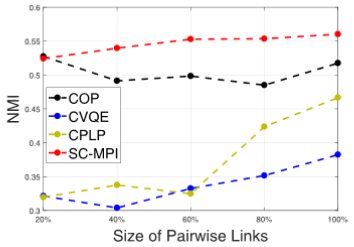
实验结果

- 评估指标
  - ARI、NMI，各自范围内越大聚类效果越好
- 聚类数K设置为给定数据集上的真实类数

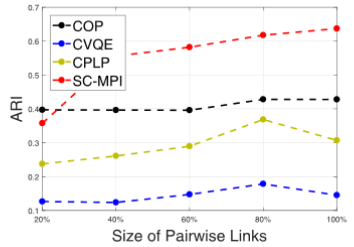
1、与单含成对约束算法对比



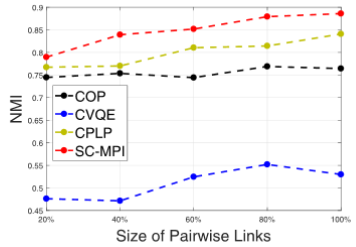
(a) ARI against size of pairwise links on data set Yale



(b) NMI against size of pairwise links on data set Yale

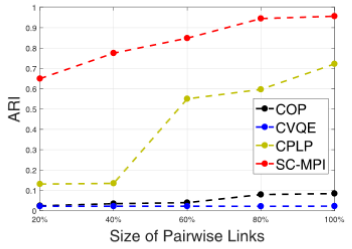


(c) ARI against size of pairwise links on data set ORL

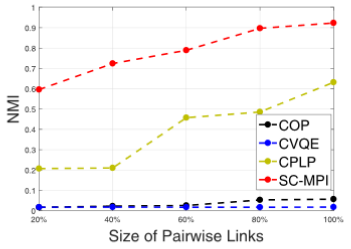


(d) NMI against size of pairwise links on data set ORL

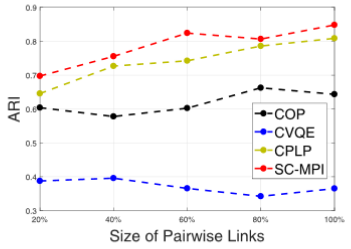
(ML数量=CL数量)



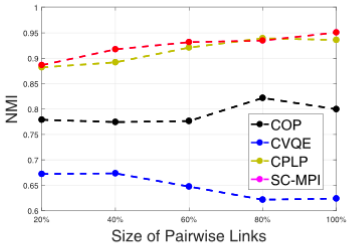
(e) ARI against size of pairwise links on data set Banknote



(f) NMI against size of pairwise links on data set Banknote



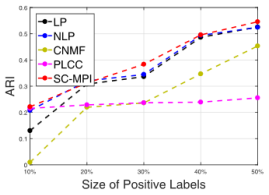
(g) ARI against size of pairwise links on data set COIL20



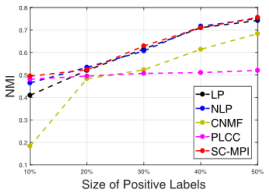
(h) NMI against size of pairwise links on data set COIL20

实验结果

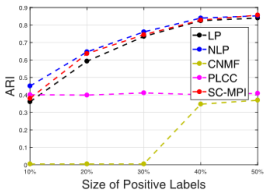
- 2、与给定正标签约束算法对比
  - 对于正标签约束，所提出的表示方法可以克服真实类标签和聚类标签之间的错位问题



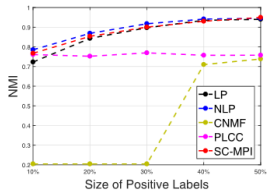
(a) ARI against size of positive labels on data set Yale



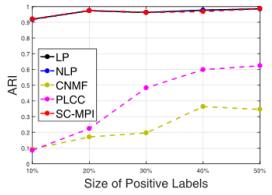
(b) NMI against size of positive labels on data set Yale



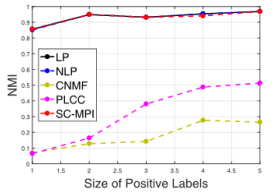
(c) ARI against size of positive labels on data set ORL



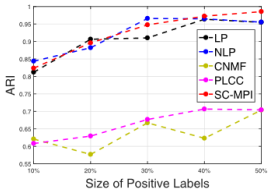
(d) NMI against size of positive labels on data set ORL



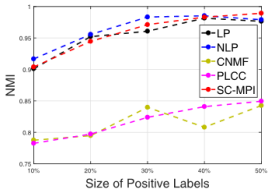
(e) ARI against size of positive labels on data set Banknote



(f) NMI against size of positive labels on data set Banknote



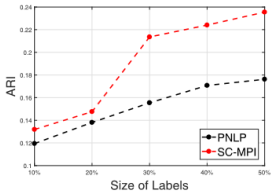
(g) ARI against size of positive labels on data set COIL20



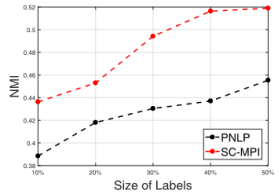
(h) NMI against size of positive labels on data set COIL20

- 3、与同时给定正、负标签约束算法对比

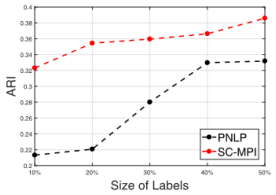
（正标签数量=负标签数量）



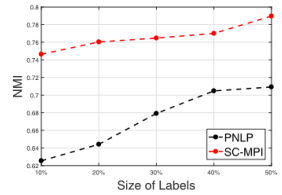
(a) ARI against size of labels on data set Yale



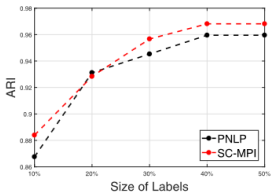
(b) NMI against size of labels on data set Yale



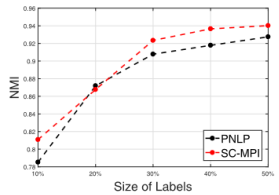
(c) ARI against size of labels on data set ORL



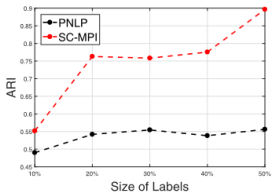
(d) NMI against size of labels on data set ORL



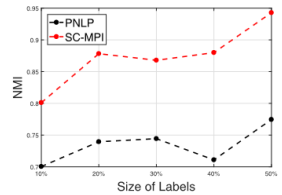
(e) ARI against size of labels on data set Banknote



(f) NMI against size of labels on data set Banknote



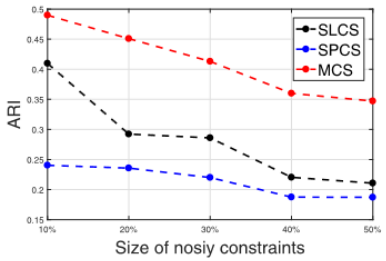
(g) ARI against size of labels on data set COIL20



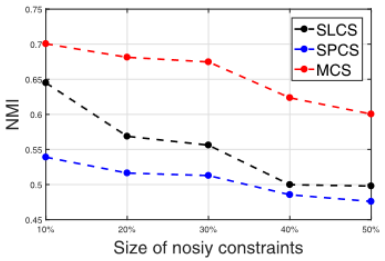
(h) NMI against size of labels on data set COIL20

实验结果

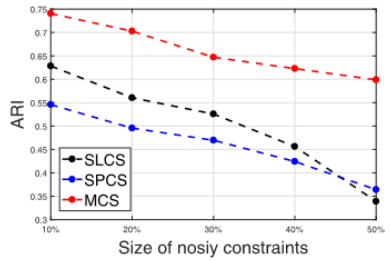
- 4、引入不正确约束信息的情况



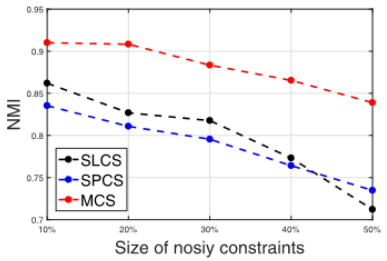
(a) ARI against size of noisy constraints on data set Yale



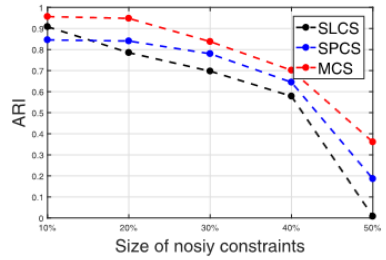
(b) NMI against size of noisy constraints on data set Yale



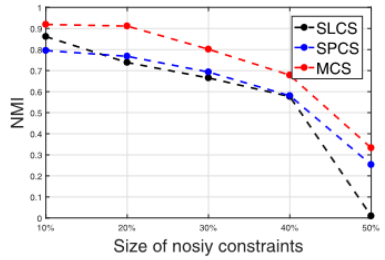
(c) ARI against size of noisy constraints on data set ORL



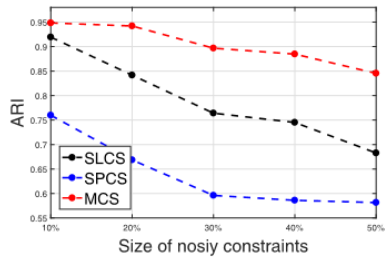
(d) NMI against size of noisy constraints on data set ORL



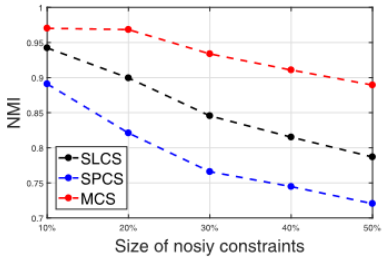
(e) ARI against size of noisy constraints on data set Banknote



(f) NMI against size of noisy constraints on data set Banknote



(g) ARI against size of noisy constraints on data set COIL20



(h) NMI against size of noisy constraints on data set COIL20

（整合多源约束，降低不正确信息的影响）

算法原理

算法原理

T	结合电子医疗记录，构建患者配对约束，找到与目标患者最相似的患者集
I	电子病历的入院记录
P	**1.**从电子病历中获取患者自身的多种特征，并构建患者属性体系 **2.**结合各类属性特征计算患者总相似度，获取成对监督信息 **3.**利用成对信息指导聚类，实现患者相似组划分
O	各患者相似组

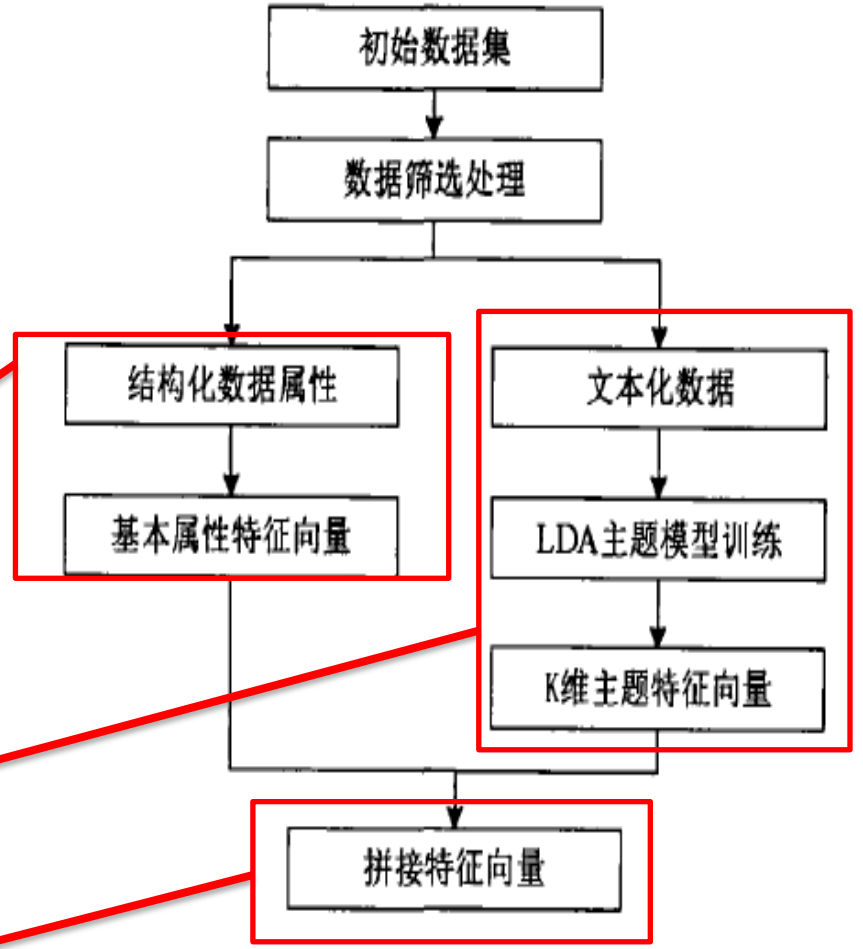
P	如何构建患者间约束信息指导患者相似集聚类
C	根据电子医疗记录可提取计算患者相似度
D	不同类别属性的相似度统一
L	**2019 SCI二区**

算法原理

- 医嘱辅助决策模型框架
  - 数据获取与处理
  - 患者相似组划分
  - 诊疗医嘱推荐

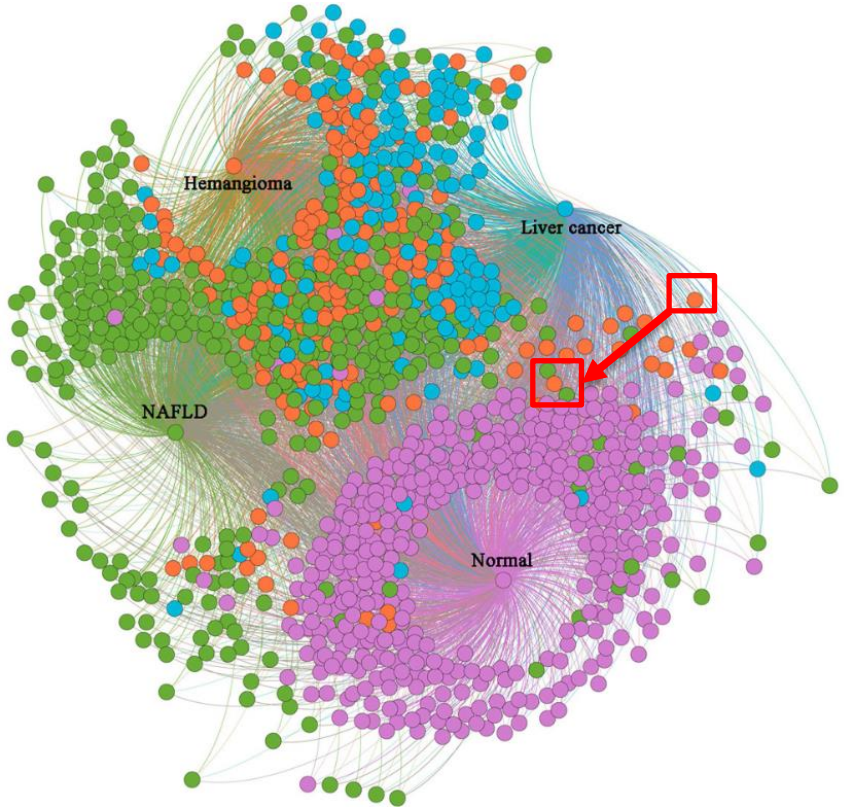
算法原理

- 数据获取与处理
  - 电子病历的入院记录
    - 患者基本信息、患者自述信息、
    - 医生检查信息、辅助检查信息、初步诊断信息
  - 特征提取
    - 结构化字段：医生建议保留可能影响患者相似性的字段、同时为保证数据隐私，将病人ID、姓名等身份识别字段筛除。
    - 非结构化文本字段：对N个字段中每个部分的关键词提取数或主题数定为K，共可产生N*K维属性值
    - 属性特征与文本主题特征进行拼接，形成最终属性体系



算法原理

- 患者相似组划分
  - 成对约束指导相似组构建
  - 聚类评估指标：轮廓系数



算法 4-2. 基于成对约束的患者相似组划分 K-means 聚类算法

输入： 文档数据集 $D=\{x_1,x_2,...,x_n\}$，must-link 集合 ML 和 cannot-link 集合 CL，初始聚类中心数 K → 人工选取的影响

输出： K 个聚簇的集合

1) 随机选择样本数据集 D 中的 K 个数据样本 $C=\{C_1,C_2,...,C_k\}$ 作为初始聚类中心；

2) 重复执行下面的步骤。直到收敛为止

3) **for each** x_i **in** D **do**

4) 计算各样本 x_i 与均值向量间的距离，找出与样本 x_i 距离最近的簇 C_r，即使 x_i 满足 $Distance = \arg\min_k \|x_i - C_i\|^2$；

5) 检测将 x_i 划入当前 C_r 簇中是否违背 ML 与 CL 集合中的约束；

6) **if ! is_voilated then**: $C_r = C_r \cup \{x_i\}$

1) **else:**将当前目标簇舍去，重新计算 $Distance$，即找到两个聚类中心 C_r 和 C_l，使 $\min(\|x_i - C_i\|^2 + \|x_j - C_l\|^2)$

2) 循环 5）至 7）步，直到归入某类簇中

7) 重新计算 K 个类别中各自的聚类中心，判断是否产生变化，若变化，则更新其均值向量，使聚类中心 $C_i = (\sum_{j=i}^{nj} x_j^i)/n$ 进行迭代计算，直至聚类中心不产生变化，或变化幅度小于所设阈值时停止迭代

3) **end**

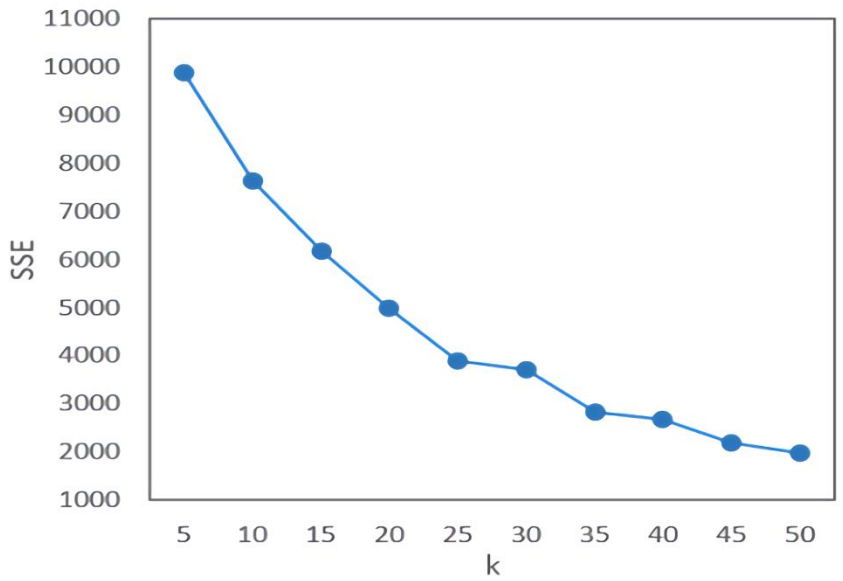
4) **Return** K 个聚簇的集合

上一步所得约束集

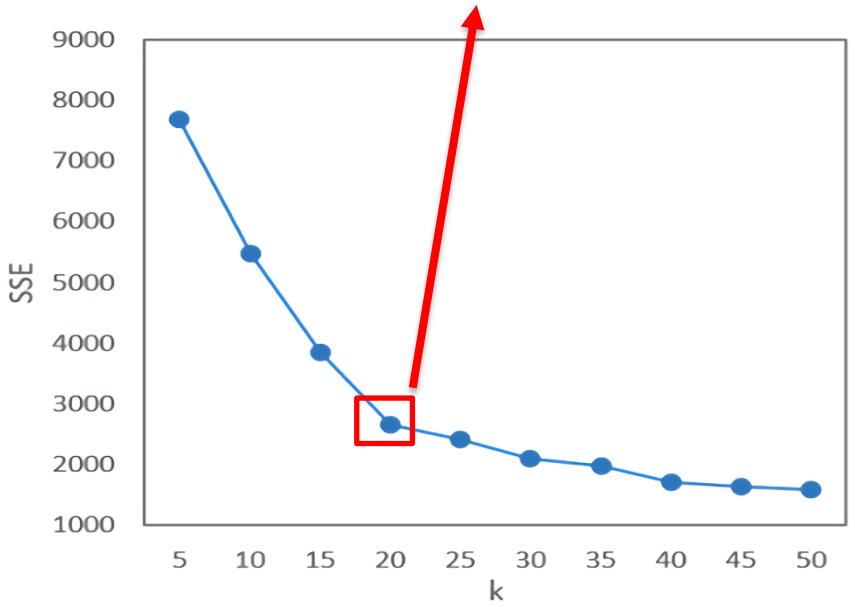
实验结果

- 聚类中心K的选取
  - 与K-Means对比
  - 成对约束信息更有针对性，因此聚类结果更清晰，样本数据分类更清晰。

K值的增大突然大幅下降的位置作为最优K值



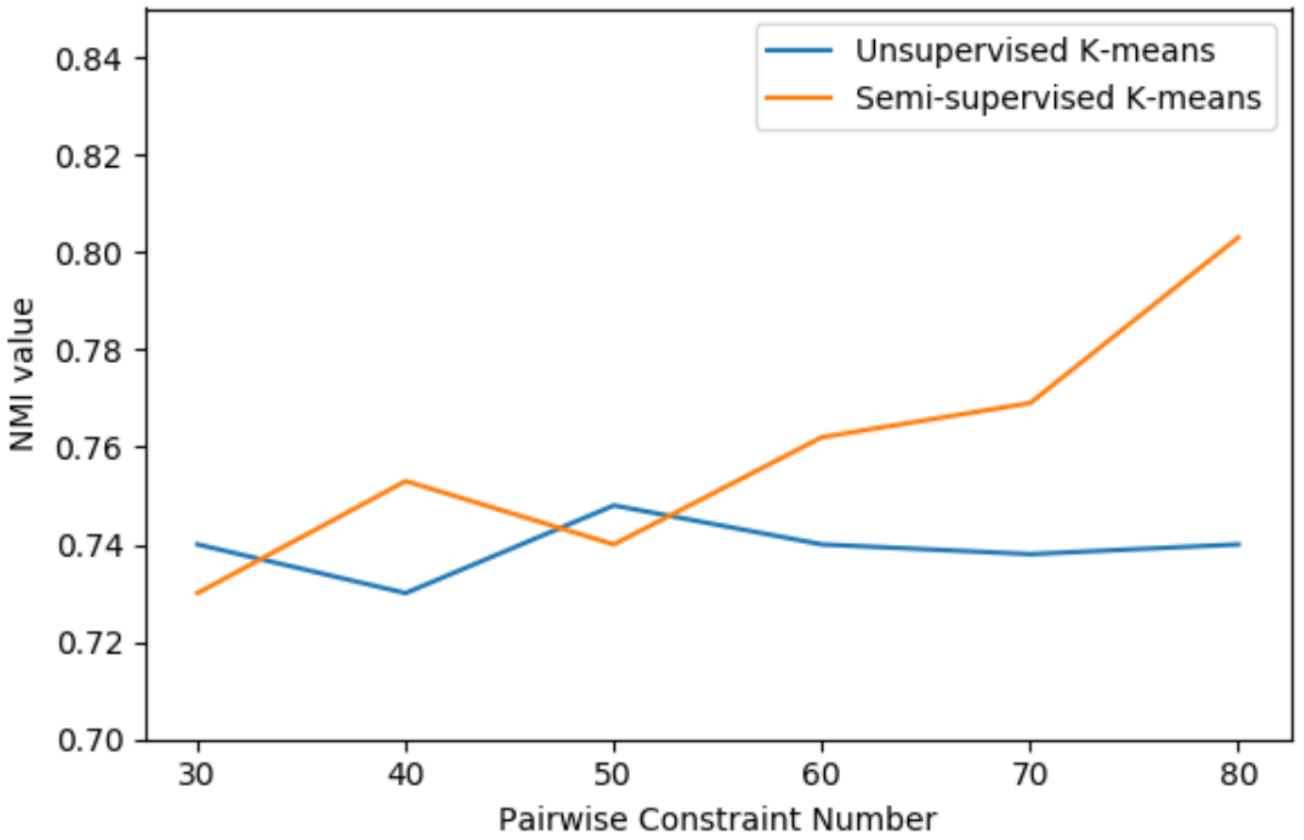
无监督



半监督

实验结果

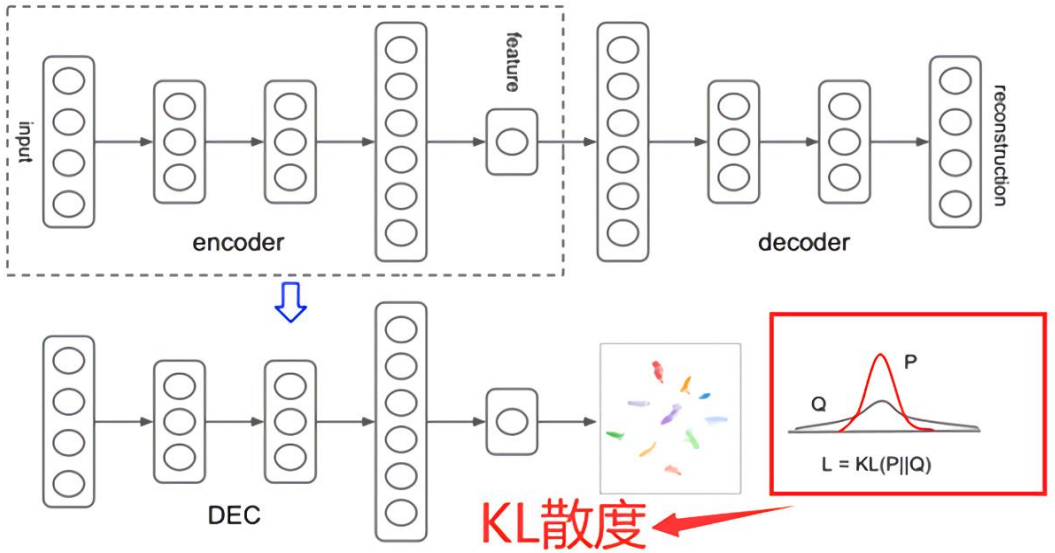
- 聚类效果
  - **NMI**
  - 半监督算法随着患者成对约束信息的增加，聚类效果不断优化。
  - 由于无监督K-means算法不受成对约束信息的影响，并且初始聚类中心是随机选择的，因此聚类效果略有不同，但总体波动不大。



总结分析

总结

- 深度聚类网络结合
  - 这两种算法都是基于传统的划分聚类方法
  - 深度聚类网络主要贡献是面向高维数据，处理转换为低维特征后再进行聚类
- 老年人运动功能特征挖掘
  - 对老年人运动相关属性提取
  - 依据属性构建约束信息，对老年人群进行划分
  - 探究其簇中隐含的，能反映簇内相似性，簇间相离性的特性



参考文献

- [1] Zhang J, Chang D. Semi-supervised patient similarity clustering algorithm based on electronic medical records[J]. IEEE Access, 2019, 7: 90705-90714.
- [2] Bai L, Liang J, Cao F. Semi-supervised clustering with constraints of different types from mult iple information sources[J]. IEEE Transactions on pattern analysis and machine intelligence, 2021, 43(9): 3247-3258.
- [3] Huang D, Hu J, Li T, et al. Consistency regularization for deep semi-supervised clustering with pairwise constraints[J]. International Journal of Machine Learning and Cybernetics, 2022: 1-14.

道德经

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒胜热。清静为天下正。

谢谢！