Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

深度半监督聚类方法

硕士研究生 谢崇玮

2023年04月22日

内容提要

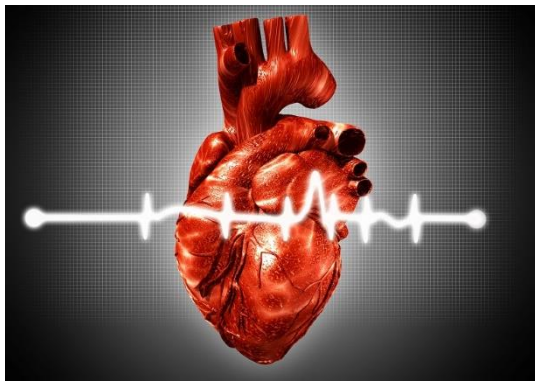
- 案例引入
- 背景简介
  - 聚类
  - 半监督聚类
- 算法原理
  - DEC
  - MCDEC
- 应用总结
- 参考文献

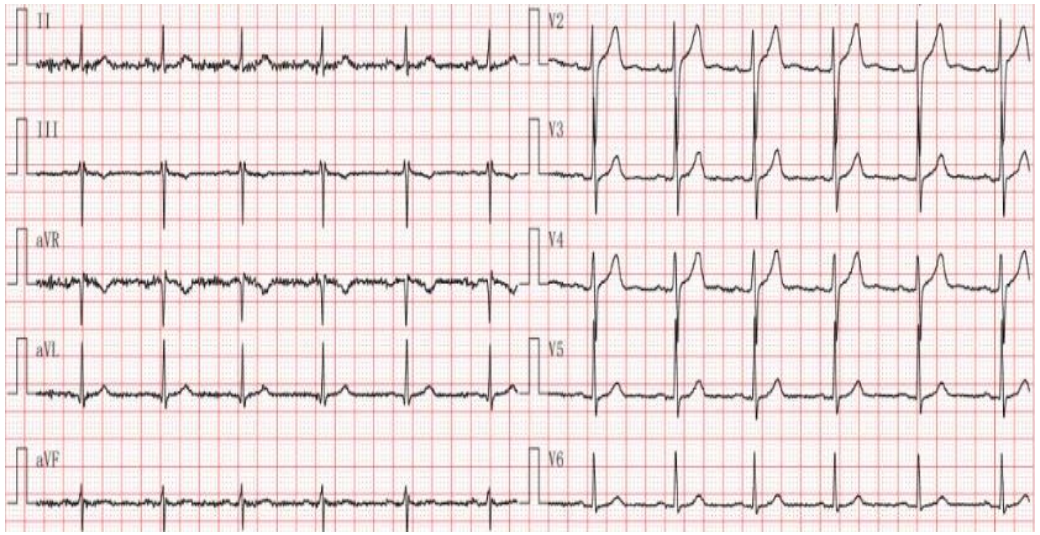
内容提要

- 预期收获
  - 1.理解深度聚类方法原理
  - 2.了解深度半监督聚类的实际应用
  - 3.学会利用不同约束进行聚类优化
  - 4.了解深度半监督聚类算法的优化与发展

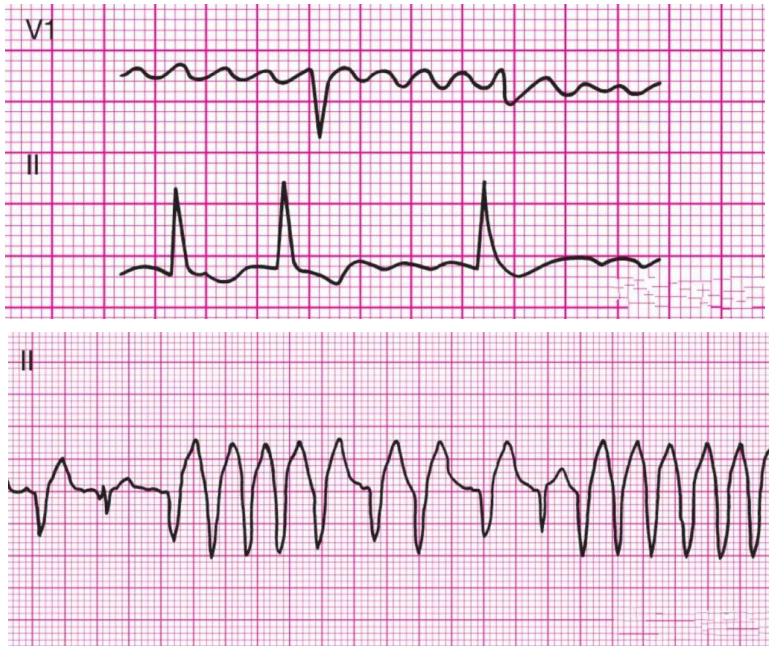
案例引入

- 假设你是一个医学研究人员，想要研究一组病人的心电图数据。这些数据来自于心脏病患者和健康人群，但是大量数据并没有标签说明每个其属于哪个组别。你希望使用半监督聚类相关算法将这些数据点分成不同的组别，以便更好地了解心电图数据的结构和特征，从而为后续的诊断提供帮助。





健康人群（心电图）

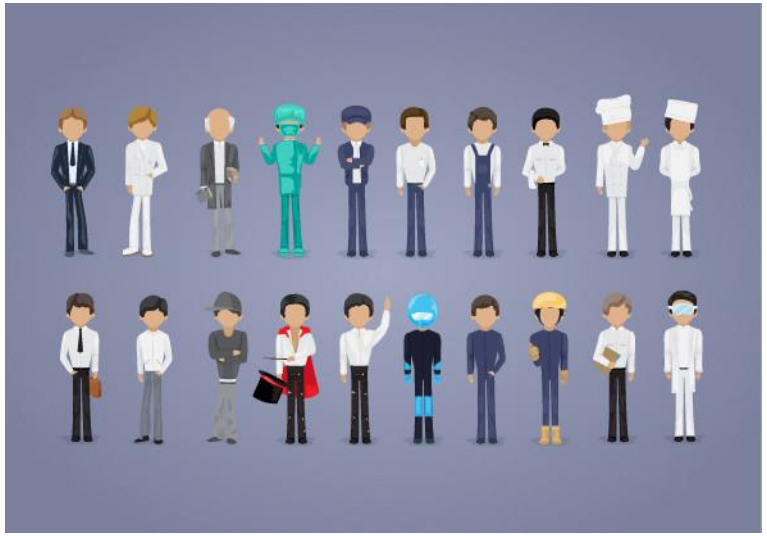


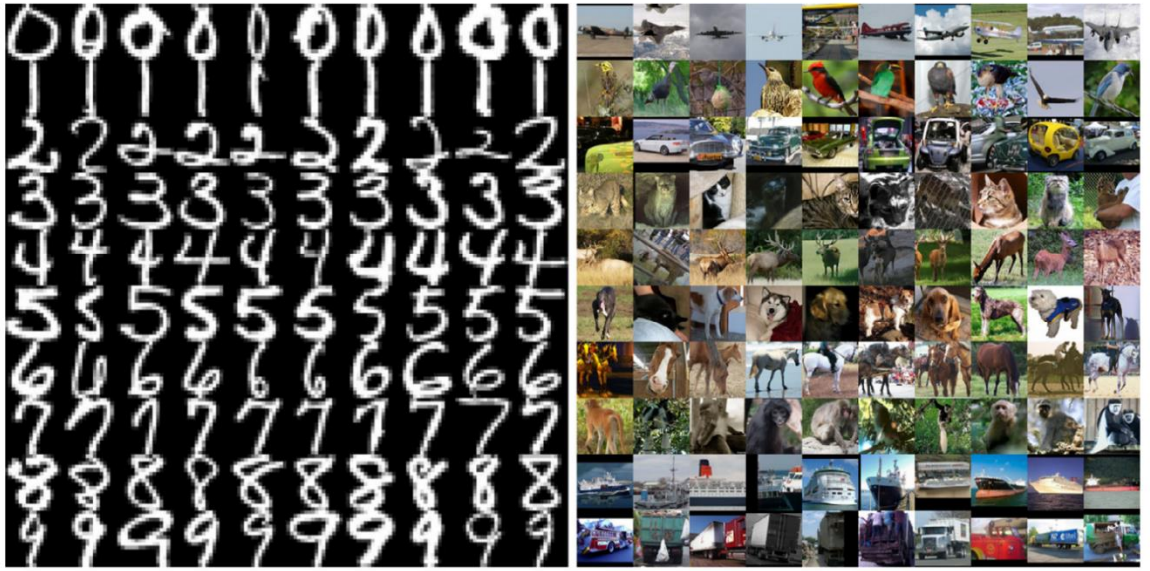
心脏病患者（心电图）

基本概念

基础概念

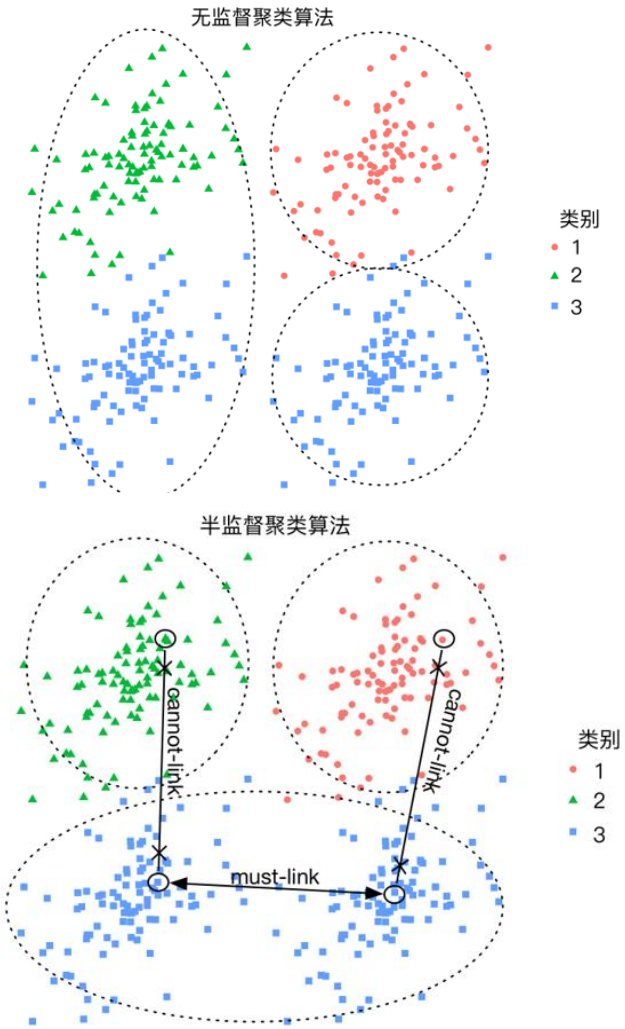
- “物以类聚，人以群分”
- 聚类就是将相似的事物聚集在一起，不相似的事物划分到不同的类别的过程
  - 例如：在图像分析中，人们希望将图像分割成具有类似性质的区域；在文本处理中，希望发现具有相同主题的文本子集；在顾客行为分析中，希望发现消费方式类似的顾客群，以便制订有针对性地客户管理方式和提升营销效率。





基础概念

- 半监督聚类(*附录1)
  - 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



数据维度剧增？

基础概念

- 聚类效果评判指标
  - 外部指标（监督）ACC

$$\mathrm{NMI} = \frac{2 \times I(\mathrm{C}, K)}{H(C) + H(K)}$$

C表示真实类别，K表示聚类结果，I(C,K)表示C和K之间的互信息，H(C)和H(K)代表熵，分别表示真实类别和聚类结果的不确定性

$$\mathrm{I}(C, K) = \sum_{c \in C} \sum_{k \in K} p(c,k) \log \frac{p(c,k)}{p(c) \cdot p(k)}$$

$$\mathrm{H}(C) = -\sum_{c \in C} p(c) \log p(c)$$

  - $\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]}$ (Adjusted Rand Index)

$$\mathrm{RI} = \frac{a+b}{a+b+c+d}$$

  - 随机聚类结果是通过将样本随机分配到聚类簇中得到的, E[RI]表示随机情况下的兰德指数的期望值
  - 内部指标（簇内相似度，簇间分离度）轮廓系数、SSE

算法原理

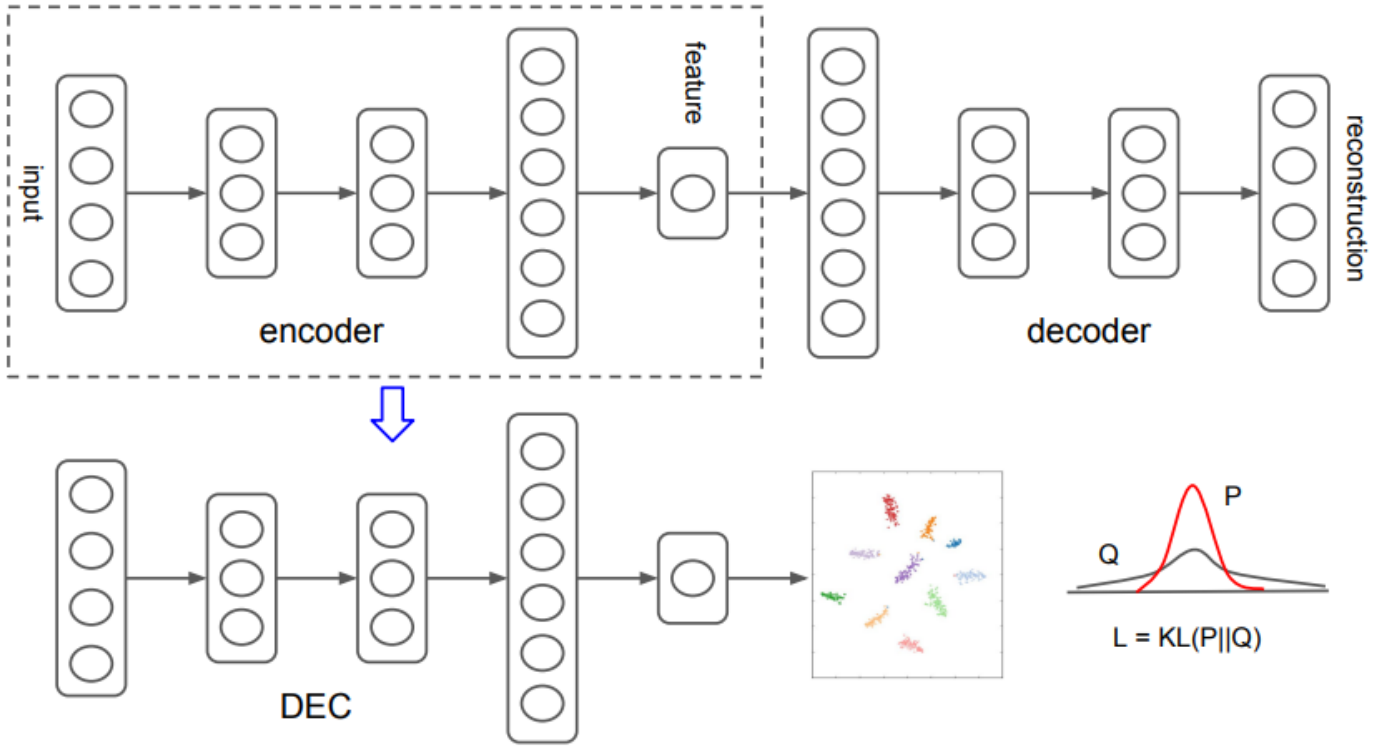
算法原理 DEC

T	目标	对数据进行符合真实分布的聚类
I	输入	UCI数据集(2图像、1文本)
P	处理	1.数据预处理，通过初始化自编码器并进行K-MEANS聚类，获得目标中心 2.根据KL散度最小化（聚类损失）训练自编码器
O	输出	训练好的编码器和最终聚类结果

P	问题	面对高维数据，数据点间距离稀疏，数据相似性和差异性模糊
C	条件	输入数据维度高
D	难点	如何同时学习特征表示和聚类分配
L	水平	CCF B(ICLR)2016

算法原理 DEC

- 预训练自编码器：
  - 使用自编码器对数据进行预训练，将原始数据转换为更具有代表性的特征。
- 初始化聚类中心：
  - 采用K-means对编码器得到的特征进行聚类，得到目标聚类中心。

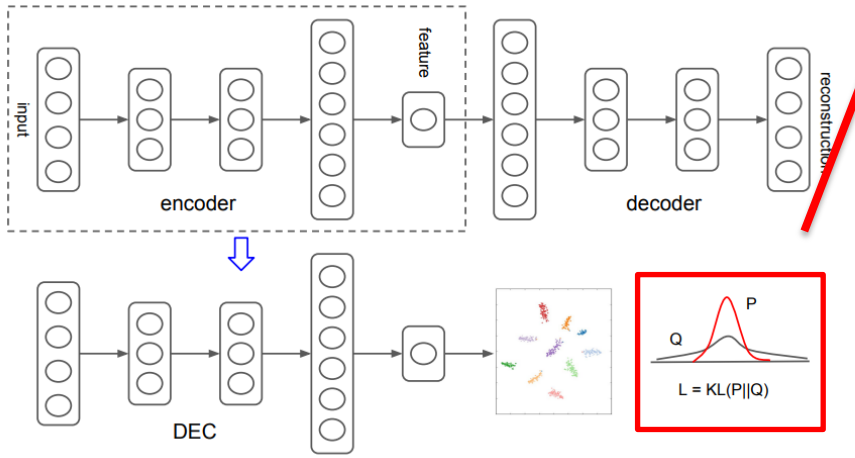


$$l_R = l(g(f(x_i)), x_i)$$

$$l(x, y) = \|x - y\|^2$$

KL散度聚类

- 第一步，计算在编码降维后的数据和聚类质心之间的软分配。 第二步，通过使用辅助目标分布，最小化目标分布和真实分布的KL散度，来更新编码器并定义聚类质心。重复该过程，直到满足收敛标准为止。



$$q_{ij} = \frac{\left(1 + \frac{\|z_i - \mu_j\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'}\left(1 + \frac{\|z_i - \mu_{j'}\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}$$

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_{j'}\left(q_{ij'}^2 / \sum_i q_{ij'}\right)}$$

$$L_C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

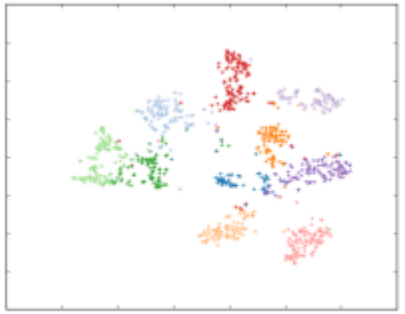
$$\frac{\partial L}{\partial z_i} = \frac{\alpha+1}{\alpha} \sum_j \left(1 + \frac{\|z_i - \mu_j\|^2}{\alpha}\right)^{-1} \times (p_{ij} - q_{ij})(z_i - \mu_j)$$

- 两次迭代之间更改聚类分配（Maxqij）的点小于阈值点或小于评价指标预设的值时停止

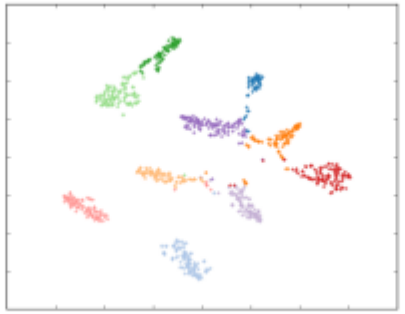
效果与能力

- 论文对比

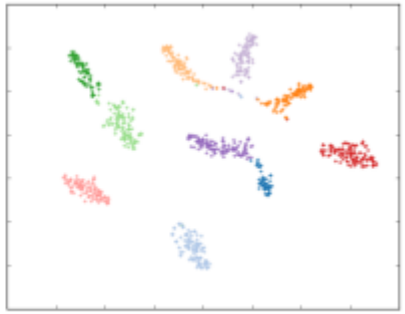
Method	MNIST	STL-HOG	REUTERS-10k	REUTERS
k-means	53.49%	28.39%	52.42%	53.29%
LDMGI	84.09%	33.08%	43.84%	N/A
SEC	80.37%	30.75%	60.08%	N/A
DEC (ours)	**84.30%**	**35.90%**	**72.17%**	**75.63%**



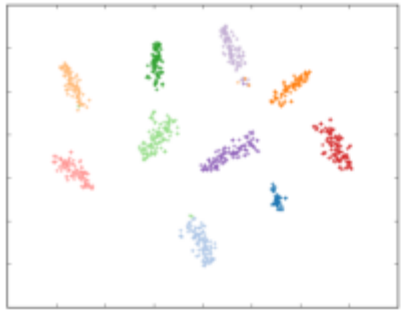
10epochs



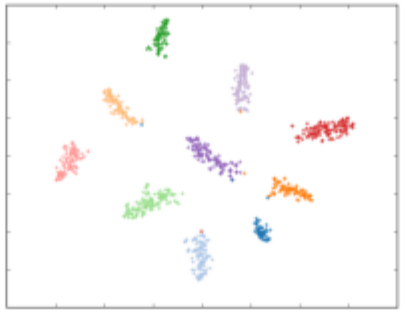
20epochs



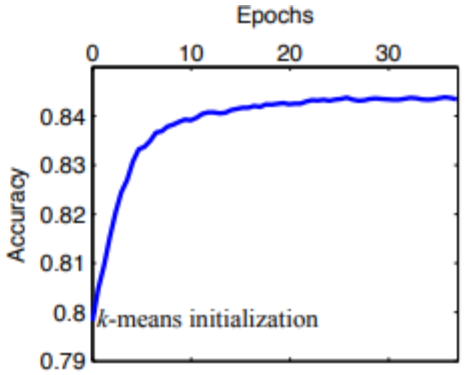
50epochs



100epochs



200epochs



算法原理 MCDEC

T	目标	对数据进行符合真实分布的聚类
I	输入	UCI数据集(7个)
P	处理	1.数据预处理，通过初始化自编码器并进行种子聚类，获得目标中心 2.主动构建成对约束并获得自适应的各类别大小比例 3.利用多种约束信息和聚类损失训练自编码器
O	输出	训练好的编码器和最终聚类结果

P	问题	现有方法随机构建成对约束，容易产生无效和不平衡约束；且K-means随机初始化聚类目标会产生质心偏移问题；同时未引入自适应的类别大小比例约束
C	条件	结合多种约束信息指导聚类
D	难点	成对约束的主动构建
L	水平	聚类效果（ACC、NMI、ARI）优于2021年sci一区论文

框架与原理

- 数据预处理和聚类目标中心选取
  - 数据归一化并且打乱数据顺序
  - 通过挑选种子约束集初始化K-means聚类中心，再通过迭代获得最终聚类目标

Algorithm 1: preprocessing and initialization of target clustering center

Input: Original data $X' \in R^{n \times d}$, number of clusters k.

1: Normalize Original data to (0, 1) range;

2: Disrupt the order of data;

3: E ← an encoder neural network;

4: D ← a feature decoder neural network;

5: Encoded data $Z = E(X)$;

5: Select some samples from each class as the seed sets $S = \cup_{h=1}^{k} S_h$ according to their categories.

6: Initialize cluster centers: $\mu_h^{(0)} \leftarrow \frac{1}{|S_h|} \sum_{z \in S_h} z$, for $h = 1, \ldots, k$; t←0

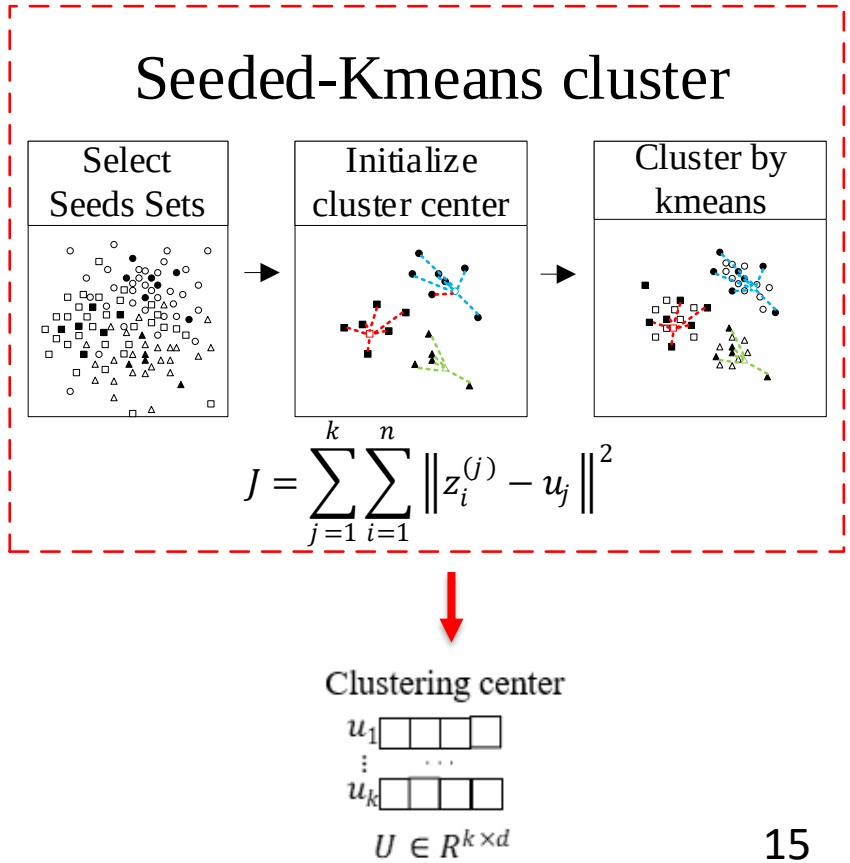
7: Repeat until t=20

7a: Assign cluster: Assign each data point z to the cluster $h^*\left(i.e.\, set\, Z_{h^*}^{(t+1)}\right)$, for $h^* = \arg min \left\| z - \mu_h^{(t)} \right\|^2$;

7b: Estimate means: $\mu_h^{(t+1)} \leftarrow \frac{1}{\left|z_h^{(t+1)}\right|} \sum_{z \in Z_h^{(t+1)}} z$;

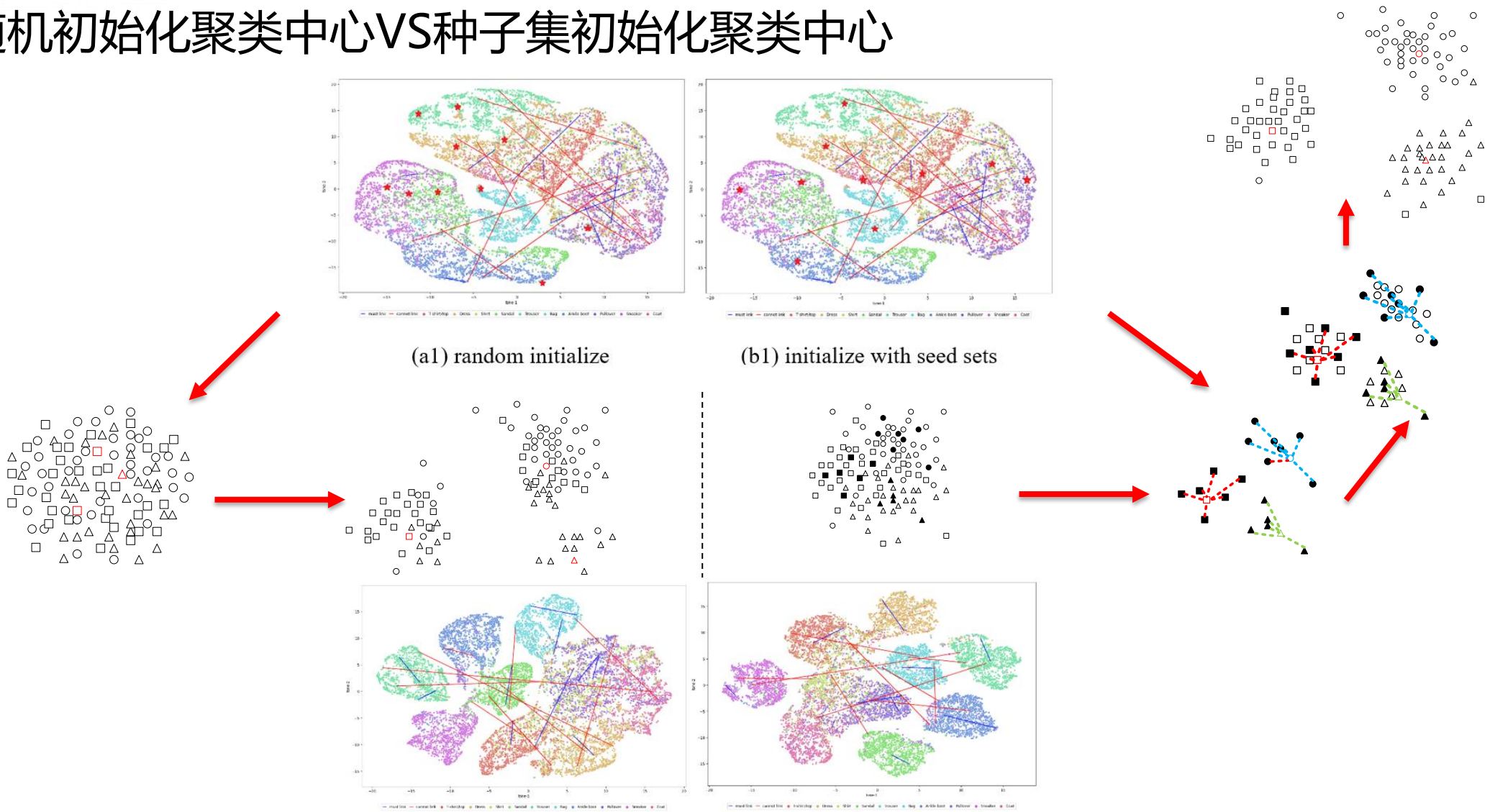
7c: t←(t+1)

Output: The encoder E, the feature decoder D, the disjoint k partitioning $\{Z_h\}_{h=1}^{k}$ of Z such that K-means objective function is optimized.



效果与能力

- 随机初始化聚类中心VS种子集初始化聚类中心



框架与原理

- 主动构建成对约束
  - 随机构建有什么问题？（会产生无效约束和约束不平衡）
  - 主动构建的具体过程

Algorithm 2: Active construction of pairwise constraints

Input: X: data, N: number of data.

1: Train the autoencoder to obtain Z;
2: Initialize membership matrix $U = [u_{ij}]$ with random numbers;
4: **for** $epoch = 1 \rightarrow 20$ **do**
5: **if** $epoch < 10$ do
6: Calculate the centers vectors;
7: Update membership matrix;
8: **else do**
9: Calculate the centers vectors;
10: Update membership matrix;
11: Adjust membership matrix;
12: **end for**
14: Calculate the uncertainty matrix;
15: **for** $epoch = 1 \rightarrow N/2$ **do**
16: Select a pair of samples from top and bottom, respectively;
17: Query their true labels;
18: Put them in pairwise constraints set S (ML/CL);
19: **end for**

Output: pairwise constraints S.

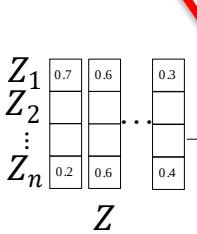
$$u_1 = \left\{\frac{1}{5}, \frac{3}{5}, \frac{0.2}{5}, \frac{0.4}{5}, \frac{0.4}{5}\right\}, u_2 = \left\{\frac{0.4}{5}, \frac{0.1}{5}, \frac{1.2}{5}, \frac{3.2}{5}, \frac{0.1}{5}\right\}$$

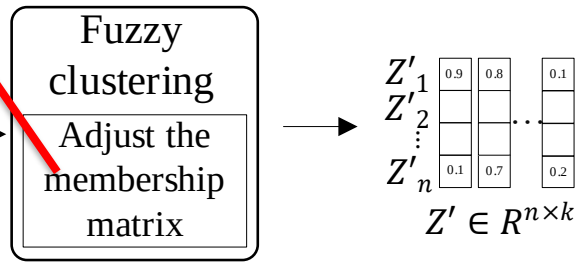
$$u_{1'} = \left\{\frac{0.5}{5}, \frac{4}{5}, \frac{0.1}{5}, \frac{0.2}{5}, \frac{0.2}{5}\right\}, u_{2'} = \left\{\frac{0.2}{5}, \frac{0.05}{5}, \frac{1}{5}, \frac{3.7}{5}, \frac{0.05}{5}\right\}$$

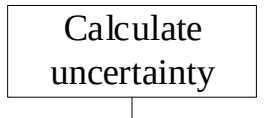
$$J(\boldsymbol{U}, k) = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2$$

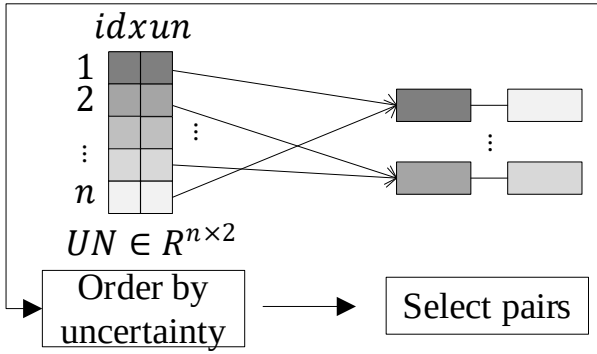
$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

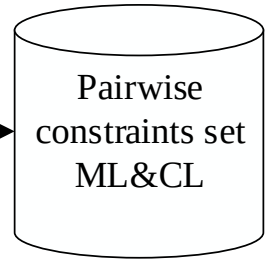
$$u_{pj} = \frac{1}{\sum_{i=1}^{k}\left(\frac{d_{pj}}{d_{ij}}\right)^{\frac{2}{m-1}}}, where\ d_{ij}^2 = \|x_i - v_i\|^2$$

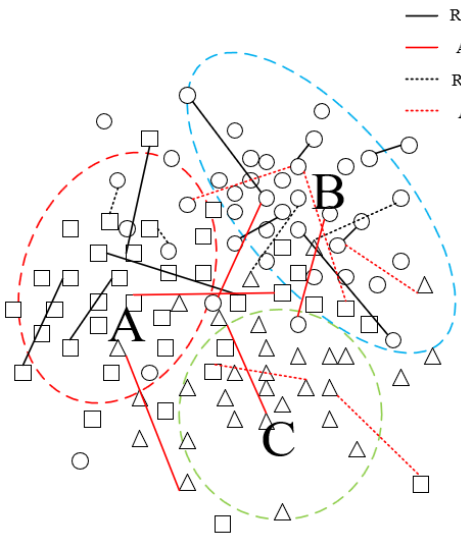






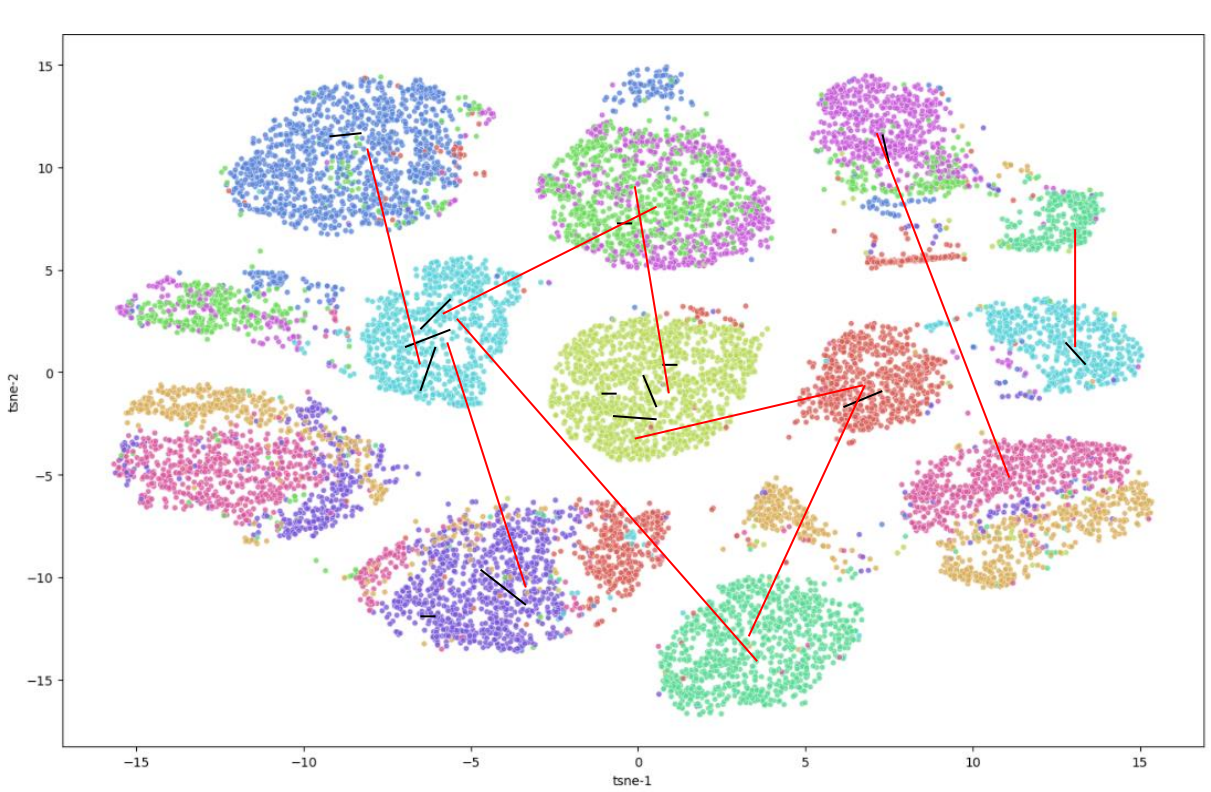


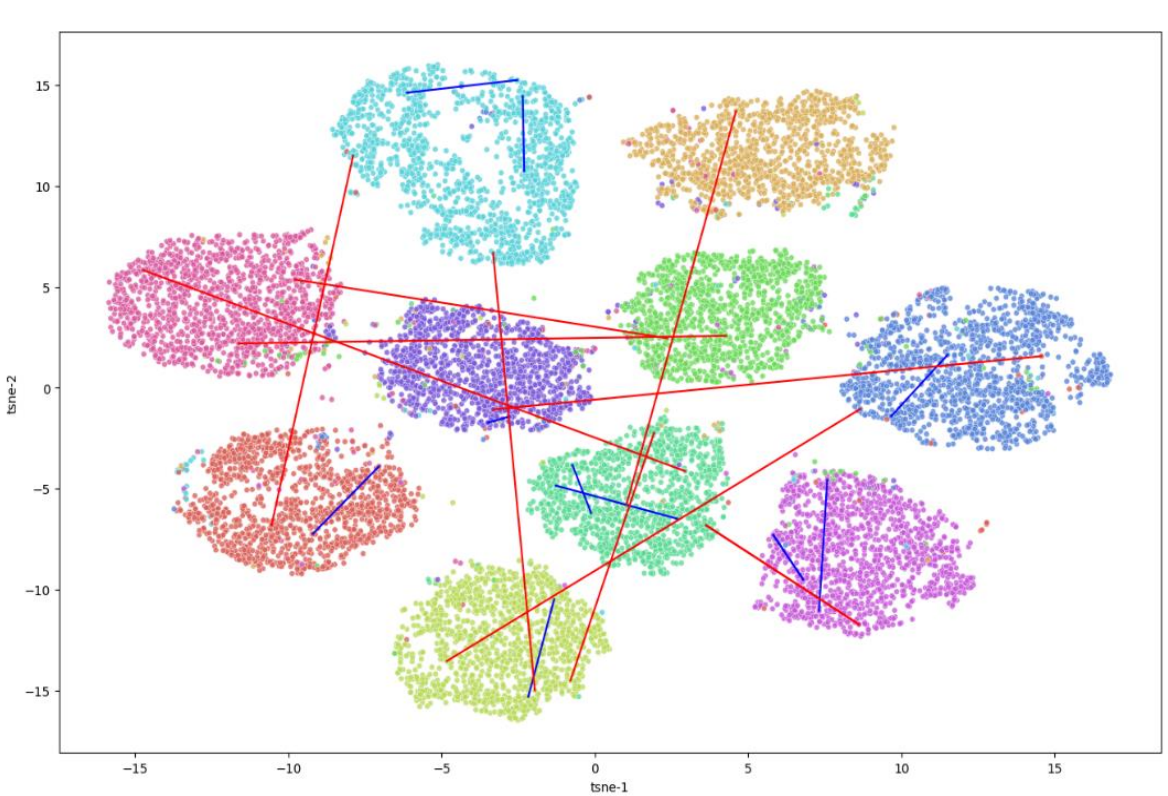




效果与能力

- 随机构建成对约束VS主动构建成对约束





框架与原理

- 训练自编码器，输出最终聚类结果

Algorithm 2: Data clustering

Input: X: data, m: maximum epochs, k: number of clusters, N: total number of batches and N_C: total number of pairwise constraints batches.

Train the autoencoder to obtain Z;
Initialize centroids μ via Seeded-K-means on embedding Z;
for $epoch = 1 \rightarrow m$ **do**
  for $batch = 1 \rightarrow N$ **do**
    Calculate L_C, L_R;
    Calculate L_G;
    Calculate total loss as $L_C + L_R + L_G$;
    Update network parameters based on total loss;
  end for
  for $batch = 1 \rightarrow N_C$ **do**
    Calculate L_P;
    Update network parameters based on L_P;
  end for
  Forward pass to compute Z and $S_i = argmax_j q_{ij}$.
end for

Output: latent embeddings Z, cluster assignment S.

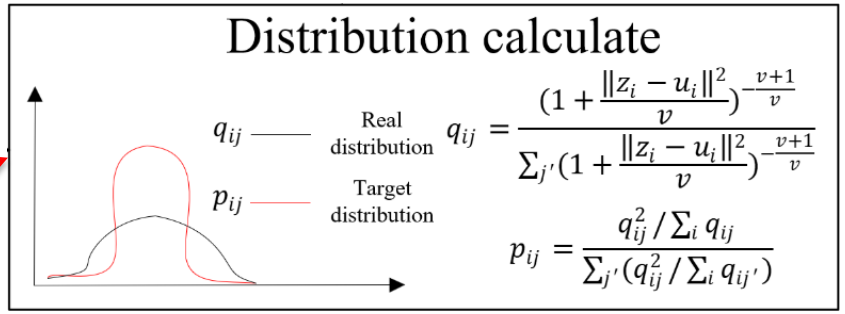
$$L_C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

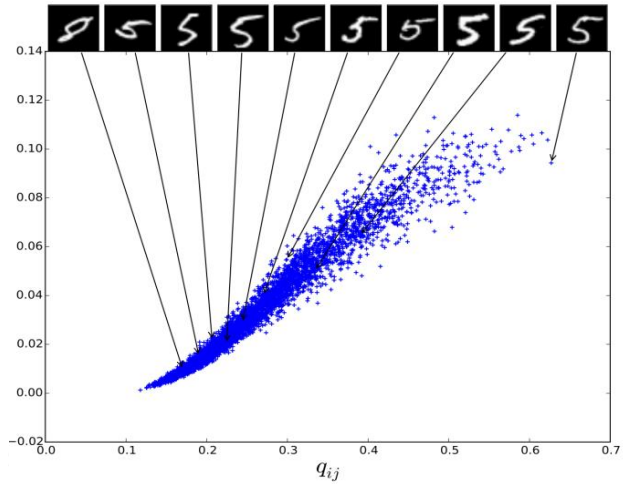
$$l_R = l(g(f(x_i)), x_i)$$

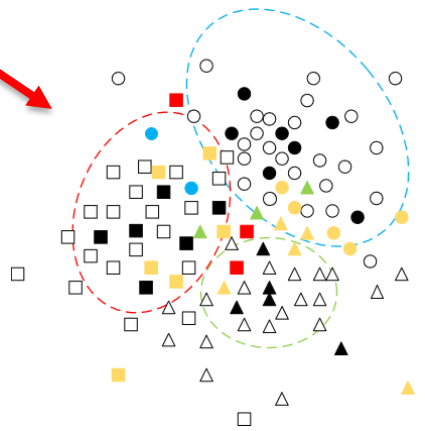
$$L_G = \sum_{c=C_1}^{C_n} \left(\sum_{i=1}^{n} \frac{q_{ic}}{n} - \frac{c}{C_1 + \cdots + C_n} \right)$$

$$L_{ML} = - \sum_{(a,b) \in ML} log \sum_j q_{aj} * q_{bj}$$

$$L_{CL} = - \sum_{(a,b) \in CL} \left(1 - \sum_j q_{aj} * q_{bj}\right)$$

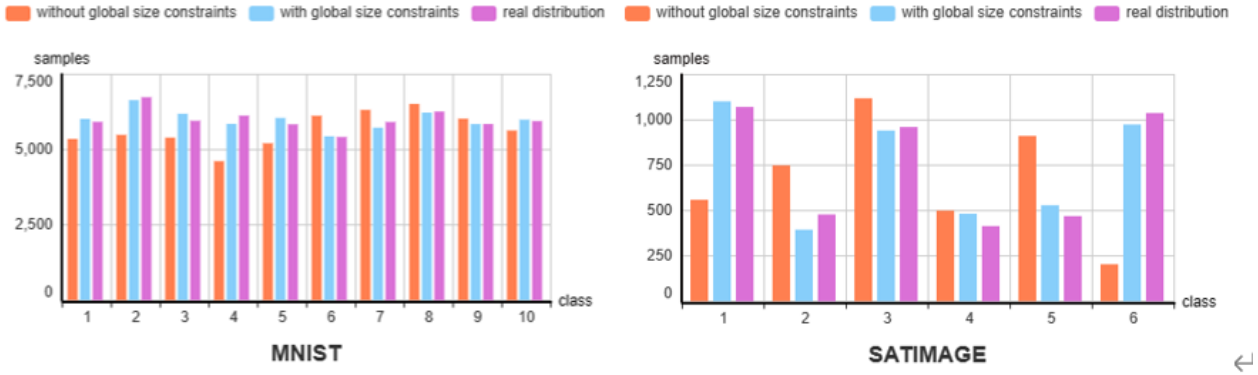






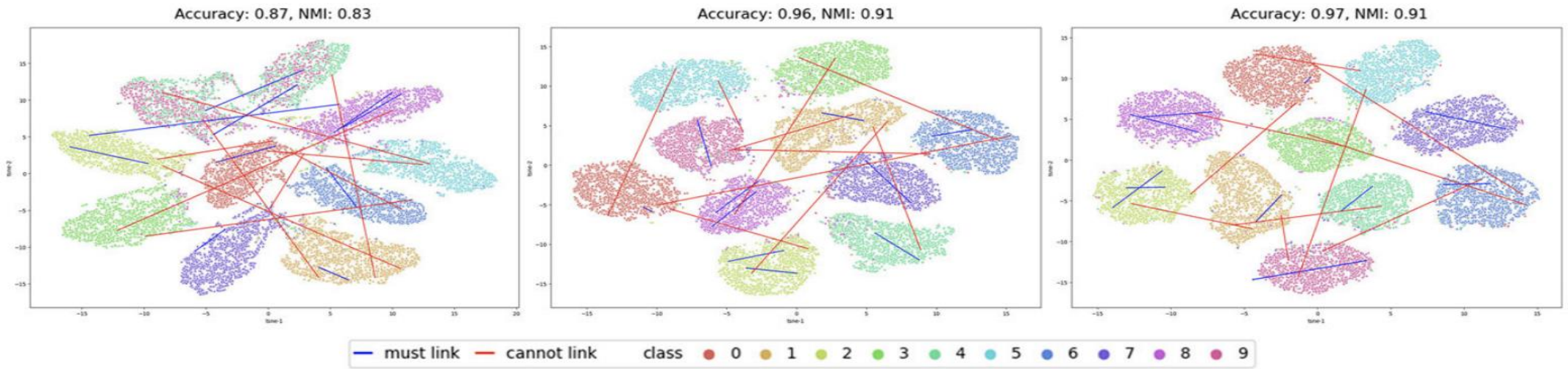
效果与能力

- 添加簇大小约束



- 与近几年高水平论文对比

Data		DEC	IDEC	SDEC	SGAE	DCC	Ours
MNIST	ACC	82.67	88.67	96.67	47.07	96.33	**96.90**
	NMI	86.67	86.82	83.79	66	90.38	**91.99**
	ARI	79.21	83.26	79.86	53.24	91.63	**93.31**
FMNIST	ACC	57.49	52.03	59.87	58.46	79.64	**81.14**
	NMI	63.41	56.33	63.22	61.23	71.33	**73.84**
	ARI	45.85	39.48	46.95	45.77	65.46	**68.69**



100epochs 200epochs 300epochs

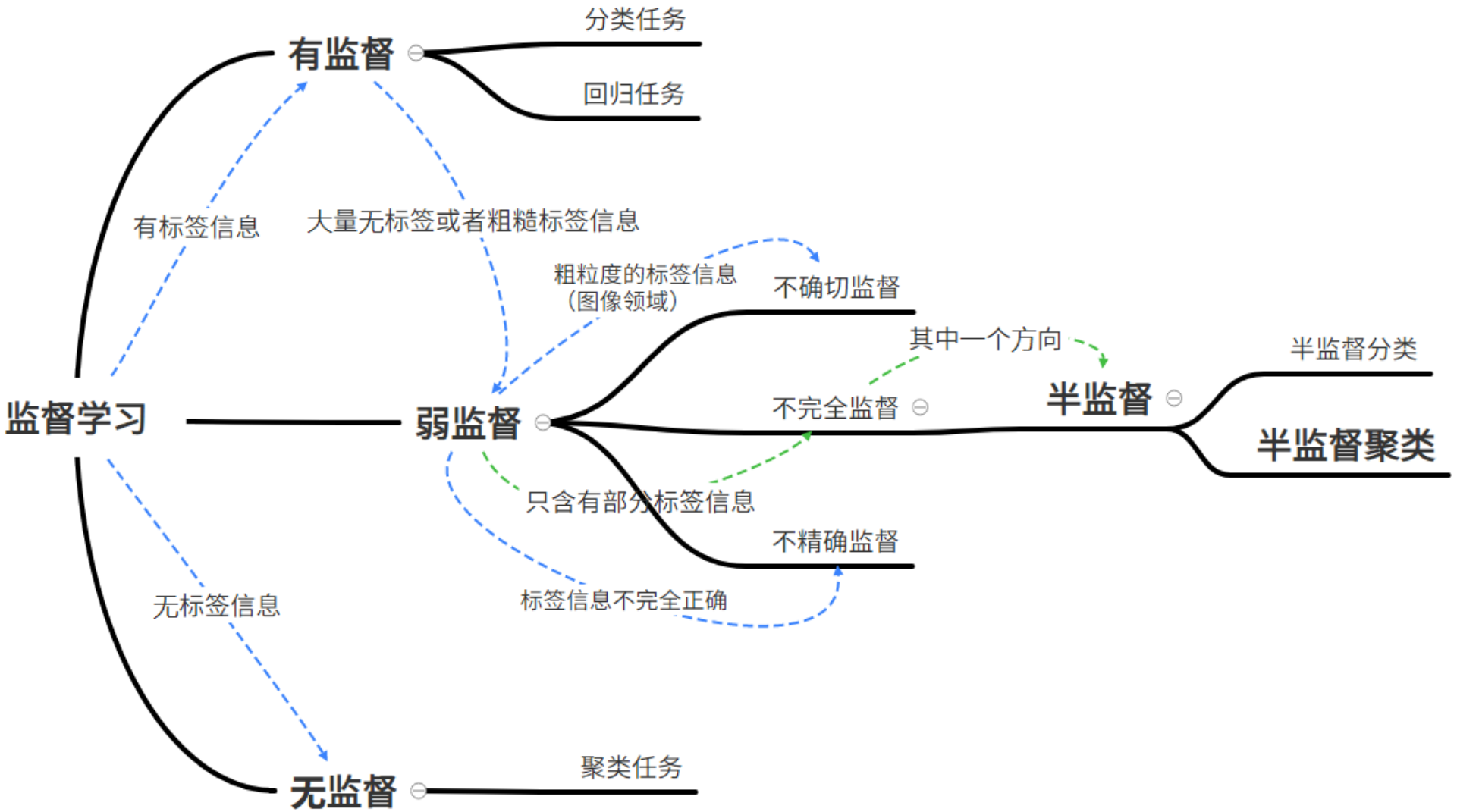
应用总结

- 应用领域
  - 老年人运动功能人群划分
  - 图像中物体分类
  - 患者相似性分析
- 未来改进方向
  - 实现对标签含噪数据集划分
  - 对聚类簇数未知的数据集进行划分
  - 使用卷积网络替代线性网络，提升对图像数据集的聚类效果
  - 采用不同的其他类别的聚类方式（密度聚类等）替换K-means聚类获取聚类目标

参考文献

- [1] Zhang H, Zhan T, Basu S, et al. A framework for deep constrained clustering [J]. Data Mining and Knowledge Di scovery, 2021, 35: 593-620.
- [2] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. International conference on machine learning [C]. PMLR, 2016: 478-487.
- [3] Ren Y, Hu K, Dai X, et al. Semi-supervised deep embedded clustering [J]. Neurocomputing, 2019, 325: 121-130.
- [4] Basu S. Semi-supervised clustering by seeding. Proceedings of the 19th International Conference on Machine Lea rning [C]. Sydney: PMLR, 2002: 19-26.
- [5] Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge. Proceedings of the 18th International Conference on Machine Learning [C]. Williams College: PMLR, 2001: 577-584.
- [6] Yang J, Parikh D, Batra D. Joint unsupervised learning of deep representations and image clusters. Proceedings o f the IEEE conference on computer vision and pattern recognition [C]. Las Vegas: IEEE, 2016: 5147-5156.
- [7] Huang P, Huang Y, Wang W, et al. Deep embedding network for clustering. 22nd International conference on patt ern recognition [C]. Beijing: IEEE, 2014: 1532-1537.
- [8] Ji P, Zhang T, Li H, et al. Deep subspace clustering networks [J]. Neural information processing systems, 2017, 30.

附录1


监督学习
有监督
分类任务
回归任务
有标签信息
大量无标签或者粗糙标签信息
弱监督
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
半监督聚类
不完全监督
只含有部分标签信息
不精确监督
标签信息不完全正确
无标签信息
无监督
聚类任务

谢谢！

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。