Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

多视图聚类技术

硕士研究生 杨景然

2023年12月10日

问题回溯

- 相关内容
  - 2023.04.22 谢崇玮《深度半监督聚类技术》
  - 2022.09.04 谢崇玮《半监督聚类和患者相似性分析》
  - 2021.07.18 董勃《多视角深度学习》

内容提要

- 预期收获
- 内涵解析
- 背景简介
- 知识基础
- 算法原理
  - DMSC
  - IMCCS
- 特点总结与工作展望
- 参考文献

预期收获

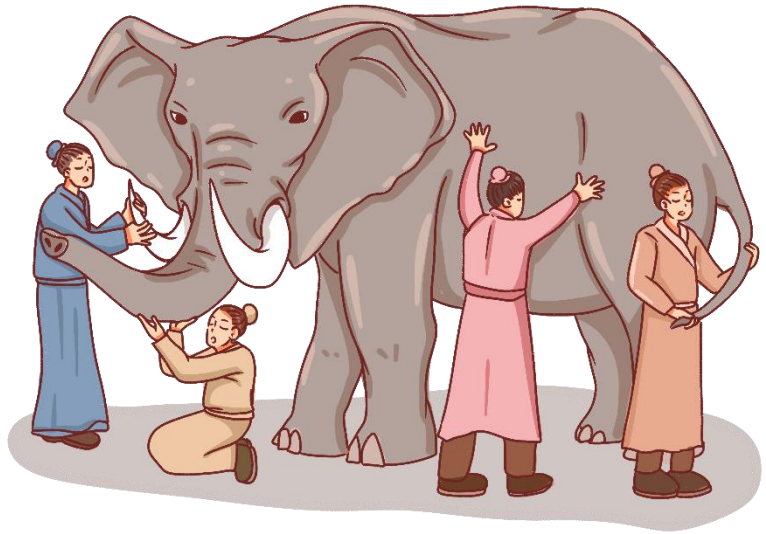
- 预期收获
  - 掌握多视图聚类技术的基本概念
  - 了解完全多视图与不完全视图聚类的方法
  - 明确多视图聚类的应用领域和发展方向

内涵解析与研究目标

- 内涵解析
  - 聚类：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
  - 多视图：在实际应用问题中，对于同一事物可以从多种不同的途径或不同的角度进行描述，这些不同的描述构成了事物的多个视图
    - 使用不同传感器采集一个人的指纹，多种不同的印痕构成了指纹数据的多个视图
    - 同一新闻事件不同媒体的报道构成了新闻文章的多个视图
- 研究目标
  - 面向数据处理领域中从各种特征收集器获得的数据
  - 结合矩阵分解、半监督学习、深度学习、多视图学习等理论
  - 实现对多视图数据对象的准确聚类

研究背景

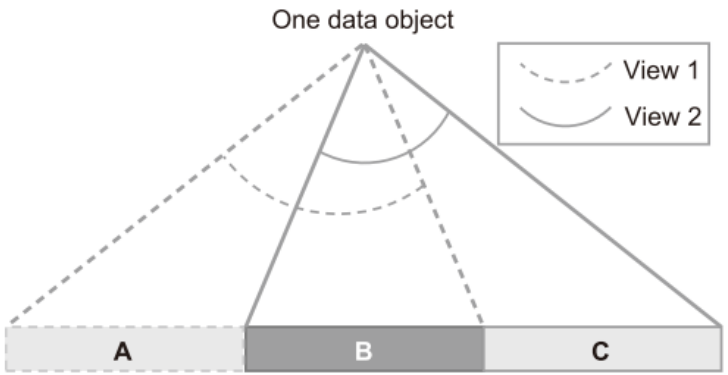
- 传统单视图聚类方法面临困难
  - 信息不完整性：单一的视图可能无法准确地反映数据的全貌
  - 数据噪声影响：单一视图中的噪声或不确定性可能对聚类结果产生负面影响
  - 数据异构性：多领域或多模态的数据通常呈现出异构性，即不同视图之间存在较大的差异
    - 每个单独的视图对于特定的知识发现任务都有其特殊属性
    - 不同的视图也包含着可以利用的互补信息

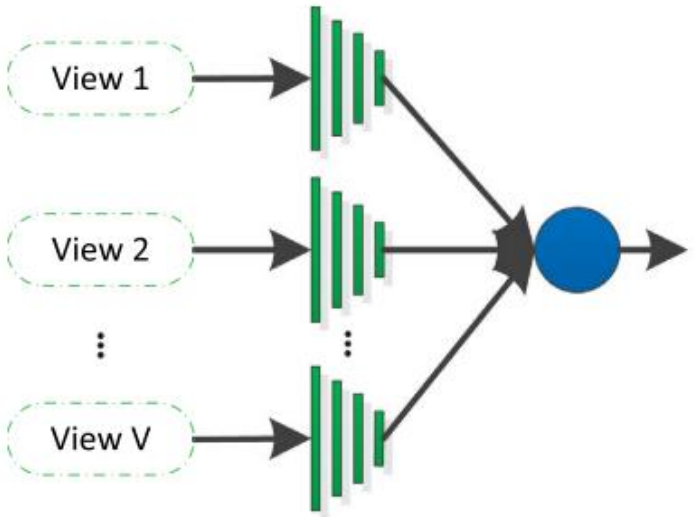


观察事物切忌“盲人摸象”

研究背景

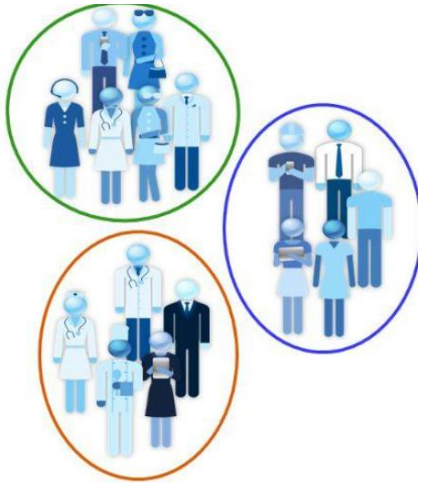
- 多视图学习的普及
  - 信息收集和提取技术的发展带动了多视图数据的普及，这使得在数据处理过程中能够获得多个角度下数据的特征
  - 核心思想：通过融合来自不同数据源或不同特征收集器的信息来提高模型的性能
  - 使用准则：共识准则、互补准则
    - 共识准则：旨在最大限度地保持多个不同观点的一致性
    - 互补准则：每个视图都有着自己独特的属性，但不同的视图之间有着可以相互补充、利用的信息

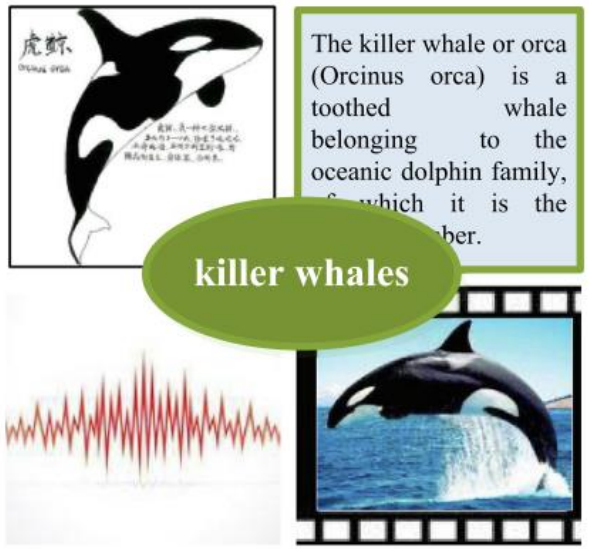




研究意义

- 研究意义
  - **提升聚类准确率**：多视图聚类通过综合不同视图的信息，提高聚类的准确性
  - **增强聚类鲁棒性**：通过结合多个视图，可以减轻某个视图受噪声影响的问题，从而提高聚类的鲁棒性
  - **跨模态关联**： 对于涉及多模态数据的任务，多视图聚类能够更好地学习和理解不同模态之间的关联，从而提供更一致和综合的聚类结果





研究历史

1996 J.B.MacQueen 提出了 k-means聚类算法，该算法是数据聚类领域内最经典的算法，该算法复杂度低、容易理解

2004 Bickel 和 Scheffer 将经典的k-means和期望最大化算法扩展到两个视图。提出了一个适用于文本数据的协同式最大期望多视图聚类算法

2009 Cleuziou 等人提出了基于协同的多核模糊聚类算法，将局部的核构造和全局模糊聚类形成一个统一的学习框架

2012 Jiang等人在概率潜在语义模型中引入协同正则化的思想，使同一对象在不同视图下的语义主题空间保持最大的一致性

2015 Zhang 等人提出了基于代表点一致性约束的多视图模糊聚类算法，利用代表点一致性约束进行多视图间全协同学习

2020 Liu 等人提出了基于簇加权的多视图核 k-means 聚类算法，该算法为各个视图内的每个簇分配了一个权重，使得拥有较高簇内相似度的簇拥有较高的权重值

2021 Li 等人提出了自适应一致性传播的图聚类方法，该算法通过从近到远传播数据点之间的拓扑连接，充分利用输入数据的流形结构去学习初始图结构

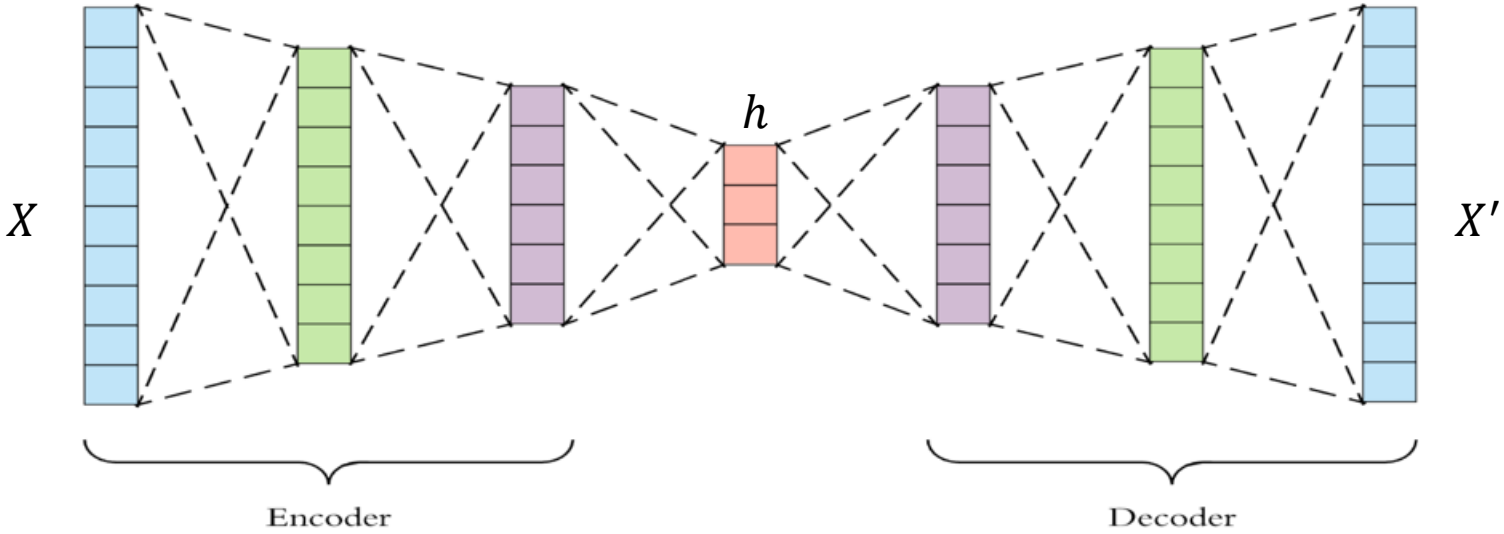
2022 Fu等人介绍了两种深度多视图联合聚类模型，其中可以同时学习多个潜在嵌入，加权多视图学习机制和聚类预测

研究现状

- 完全多视图聚类
  - 获得的多视图数据没有缺失，利用多视图信息提高聚类模型的鲁棒性和聚类准确率
- 不完全多视图聚类
  - 由于数据采集技术的限制或人为因素等原因常导致视图或样本缺失的问题
    - 基于多核学习
    - 基于矩阵分解
    - 基于深度学习
    - 基于图学习
  - 解决方式
    - 学习跨视图数据的一致信息
    - 从已有的数据中恢复缺失的数据

知识基础 自动编码器

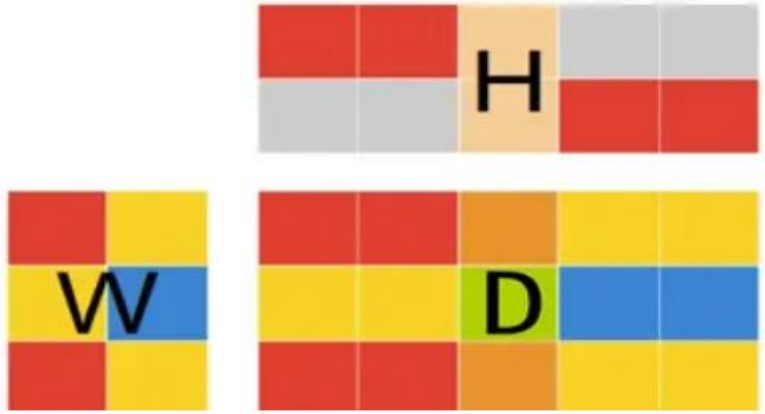
- 自动编码器结构
  - 包含两个主要的部分：Encoder（编码器）和Decoder（解码器）
    - Encoder：把高维输入X编码成低维的隐变量h，让神经网络可以学习最有信息量的特征
    - Decoder：把隐变量h尽可能地还原为原始数据
- 自动编码器分类
    - 堆叠自编码器：将多个编码器连接在一起构建一个深层的神经网络
    - 卷积自编码器：采用卷积层代替全连接层



知识基础 非负矩阵分解

- 非负矩阵分解（NMF）
  - 基本原理：将大的非负矩阵可以分解为两个小的非负矩阵，解决数据庞大而带来的复杂超额的计算量，所以常用于数据降维
  - 公式概况：设非负矩阵$D = m * n$， m是数据的个数， n 是数据维度，则D可以表示为$D_{(m*n)} \approx D'_{(m*n)} = W_{(m*k)} * H_{(k*n)}$， k为从矩阵D中抽取的特征个数
    - W：包含k个特征的基矩阵，也就是降维的目标矩阵
    - H：为对应的系数矩阵
  - 常见的损失函数
    - $min\|D - D'\|^2$
    - 基于KL散度

W与H不唯一



Deep multi-view semi-supervised clustering with sample pairwise constraints

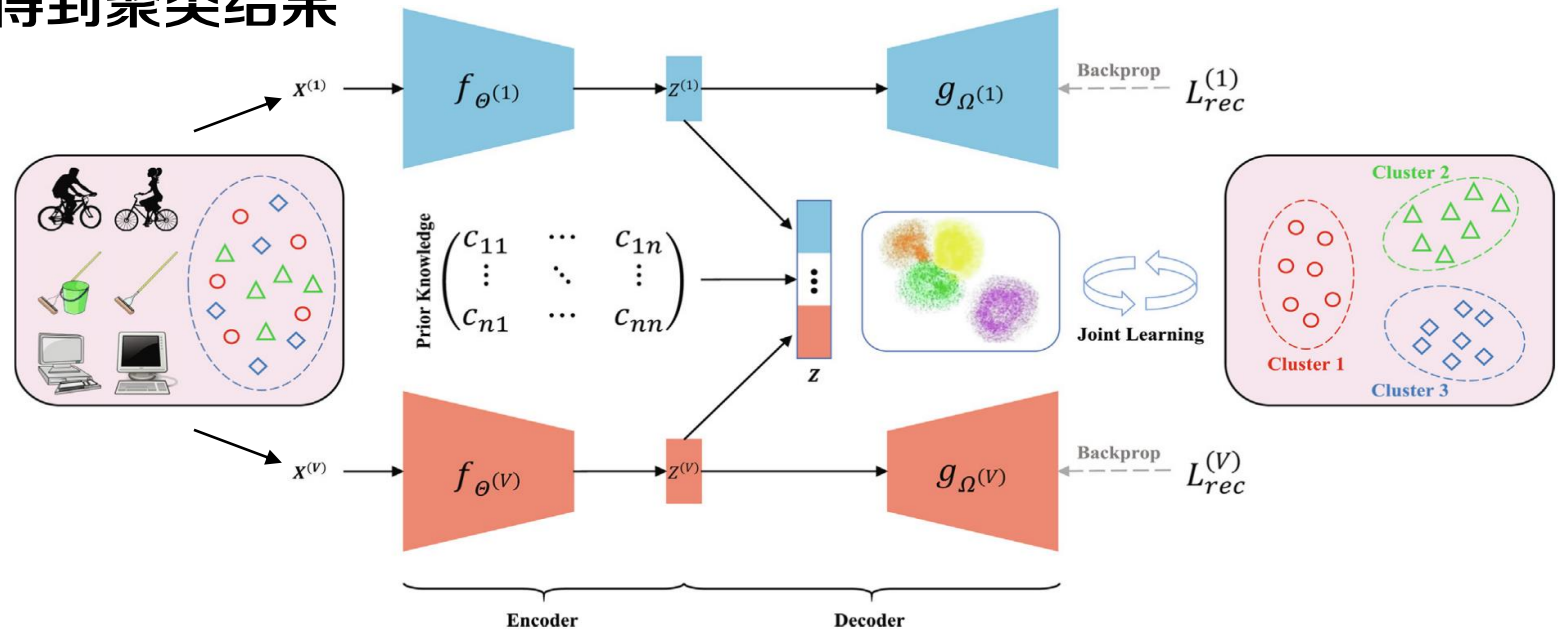
DMSC TIPO

T	目标	提取数据多视图信息，利用互补性和一致性提高聚类准确率
I	输入	图像数据集*n、成对约束信息*n
P	处理	1.预处理：使用神经网络提取原始数据的不同特征 2.构建多视图特征空间并初始化其聚类中心 3.使用初始聚类中心、成对约束矩阵、自编码器重建损失组成自编码器的损失函数，迭代更新聚类中心并得到聚类结果
O	输出	聚类分配矩阵*1

P	问题	现有单视图深度聚类未考虑数据特征的多样性
C	条件	数据具有多种类型的特征
D	难点	如何挖掘并提取出数据能够利用的不同特征并用于提升聚类准确率
L	水平	Neurocomputing2021（SCI 二区）

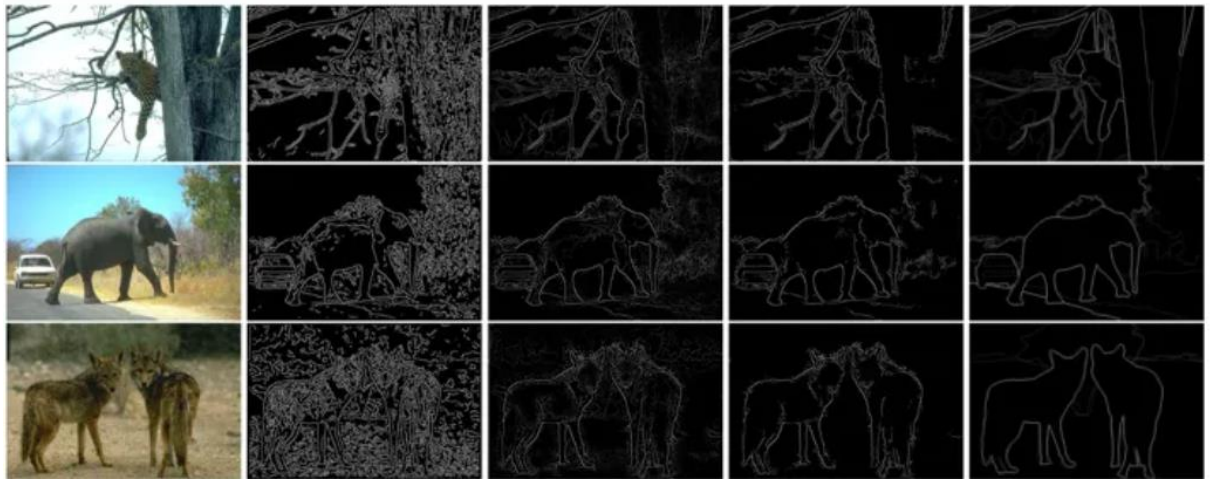
DMSC 算法原理图

- 算法原理图
  - 预处理：使用卷积神经网络提取原始数据的不同特征
  - 构建特征空间：将特征视图通过自编码器网络构建多视图特征空间
  - 聚类中心初始化：使用k-means进行多视图特征的聚类中心初始化
  - 聚类：使用初始聚类中心、成对约束矩阵、自编码器重建损失组成损失函数，迭代更新聚类中心并得到聚类结果



创新分析 多视图特征提取

- 多视图特征提取
  - 使用各类神经网络对数据的不同特征进行提取
  - 常用的神经网络
    - VGG16、VGG19、ResNet50、InceptionV3、AlexNet、DenseNet121等
    - 所用的神经网络都使用ImagNet数据集进行训练
  - 要点：特征的提取选用的神经网络视数据集的不同而不同
    - 对于彩色物体数据集，可以选用VGG16、ResNet50、InceptionV3等神经网络



创新分析 约束损失与编码器重构损失

- **损失函数**
  - $L = \gamma \cdot L_{clu} + \lambda \cdot L_{con} + L_{rec}$
    - L_{clu}为聚类损失，根据KL散度构建： $L_{clu} = KL(P||Q) = \sum_{i=1}^{n}\sum_{j=1}^{K} p_{ij} \log \frac{p_{ij}}{q_{ij}}$
    - L_{con}为约束损失， $L_{con} = \sum_{i=1}^{n}\sum_{k=1}^{n} c_{ik}\|Z_i - Z_K\|_2^2$
    - L_{rec}为自编码器重构损失，$L_{rec}^{(v)} = \sum_{i=1}^{n} \left\| X_i^{(v)} - \hat{X}_i^{(v)} \right\|_2^2$
    - p_{ij}和q_{ij}为软分配和辅助目标分配，Z为多视图特征空间， c_{ik} 为成对约束矩阵的元素值； $\hat{X}_i^{(v)}$表示视图V中的数据经过编码器重建后的值
  - 引入约束信息用于引导聚类快速收敛（见附录A）
  - 引入重建损失，既保护了局部特定视图的独特性，又保证了全局共享视图完整性

实验设计

- 数据资源
  - 数据集如右图所示
- 对比方法
  - 传统单视图聚类方法
    - DEC、IDEC、SDEC
  - 现有的多视图聚类方法
    - RMKMC、MSPL、DCCAE等
- 实验设置
  - 本实验的对比实验中使用两个视图，如下表所示
- 评价指标
  - 准确率（ACC）
  - 归一化互信息（NMI）
  - 调整兰德指数（ARI）

数据集	数量	类别数	大小	通道数
STL10	9298	10	96×96	3
COIL100	10000	100	128×128	3
CALTECH101	8677	101	-	3
CIFAR10	60000	10	32×32	3

Branch	Input
View 1 (SAE)	DenseNet121 feature
View 2 (SAE)	InceptionV3 feature

实验设计 对比实验

- 实验结果
  - 与其他方法相比，DMSC方法在三种评价指标下均优于其他算法
  - 多视图概念的引入有助于提高聚类算法的准确率

Type	Method	STL10			COIL100			CALTECH101			CIFAR10		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SvC	AE-View1	0.7521	0.7218	0.6367	0.7546	0.9278	0.7473	0.5088	0.7555	0.4259	0.5300	0.4384	0.3335
	AE-View2	0.8716	0.8401	0.8023	0.7079	0.9152	0.7018	0.5877	0.8019	0.4636	0.6586	0.5697	0.4696
	AE-View1,2	0.9098	0.8706	0.8478	0.7676	0.9393	0.7706	0.6096	0.8278	0.4924	0.6658	0.5883	0.4965
	DEC [2]	0.9574	0.9106	0.9091	0.7794	0.9459	0.7779	0.6282	0.8364	0.5261	0.6744	0.5930	0.5137
	IDEC [41]	0.9605	0.9150	0.9155	0.7921	0.9481	0.7955	0.6373	0.8393	0.5398	0.6866	0.6072	0.5298
	DCN [42]	0.9318	0.8965	0.8781	0.7771	0.9399	0.7726	0.6626	0.8418	0.6022	0.6828	0.6326	0.5308
	ASPC [43]	0.9381	0.9061	0.8908	0.7854	0.9497	0.7869	0.6729	0.8495	0.6087	0.6692	0.6162	0.5153
	SDEC [44]	0.9585	0.9120	0.9115	0.7942	0.9523	0.8009	0.6433	0.8450	0.5471	0.6953	0.6141	0.5353
MvC	RMKMC [51]	0.8344	0.8273	0.7635	-	-	-	-			0.5714	0.4688	0.3679
	MSPL [52]	0.7414	0.7174	0.6370	-	-	-	-			0.7156	0.5948	0.5174
	DCCA [21]	0.8411	0.7477	0.6917	-	-	-	-			0.4242	0.3385	0.2181
	DCCAE [22]	0.8235	0.7273	0.6632	-	-	-	-			0.3960	0.3226	0.2034
	DGCCA [23]	0.8960	0.8218	0.7970	-	-	-	-	-	-	0.4703	0.3577	0.2634
	DMJCS [24]	0.9374	0.9063	0.8989	0.7841	0.9532	0.7991	0.6998	0.8578	0.7054	0.7184	0.6188	0.5527
	DEMVC [25]	0.9582	0.9121	0.9132	0.7563	0.9382	0.7626	0.6719	0.8419	0.6991	0.6998	0.6351	0.5457
	DMSC (ours)	**0.9679**	**0.9268**	**0.9305**	**0.8077**	**0.9569**	**0.8159**	**0.7161**	**0.8593**	**0.7230**	**0.7337**	**0.6442**	**0.5712**

实验设计 消融实验

- 实验设置
  - 在USPS数据集上进行鲁棒性测试，使用三个视图（添加了原始像素视图）
- 实验结果
  - 随着微调迭代进行直到模型收敛，使用三个特征视图实现了比两个视图更出色的聚类性能
  - 合理地增加视图数量能够有效提升聚类性能

Stage	Method	ACC	NMI	ARI
Initialization	View1	0.7112	0.6926	0.5940
	View2	0.7317	0.7169	0.6348
	View3	0.7252	0.7114	0.6268
	View1,2,3	0.7492	0.7458	0.6678
Finetuning	View1,2	0.7866	0.8163	0.7380
	View1,3	0.7782	0.8131	0.7362
	View2,3	0.7804	0.8119	0.7317
	View1,2,3	**0.7873**	**0.8263**	**0.7452**

USPS		
ACC	NMI	ARI
0.7727	0.7941	0.7207
0.7825	0.8087	0.7326
0.7780	0.8142	0.7353
0.7866	**0.8163**	**0.7380**

IMCCS

Incomplete multi-view clustering with cosine similarity

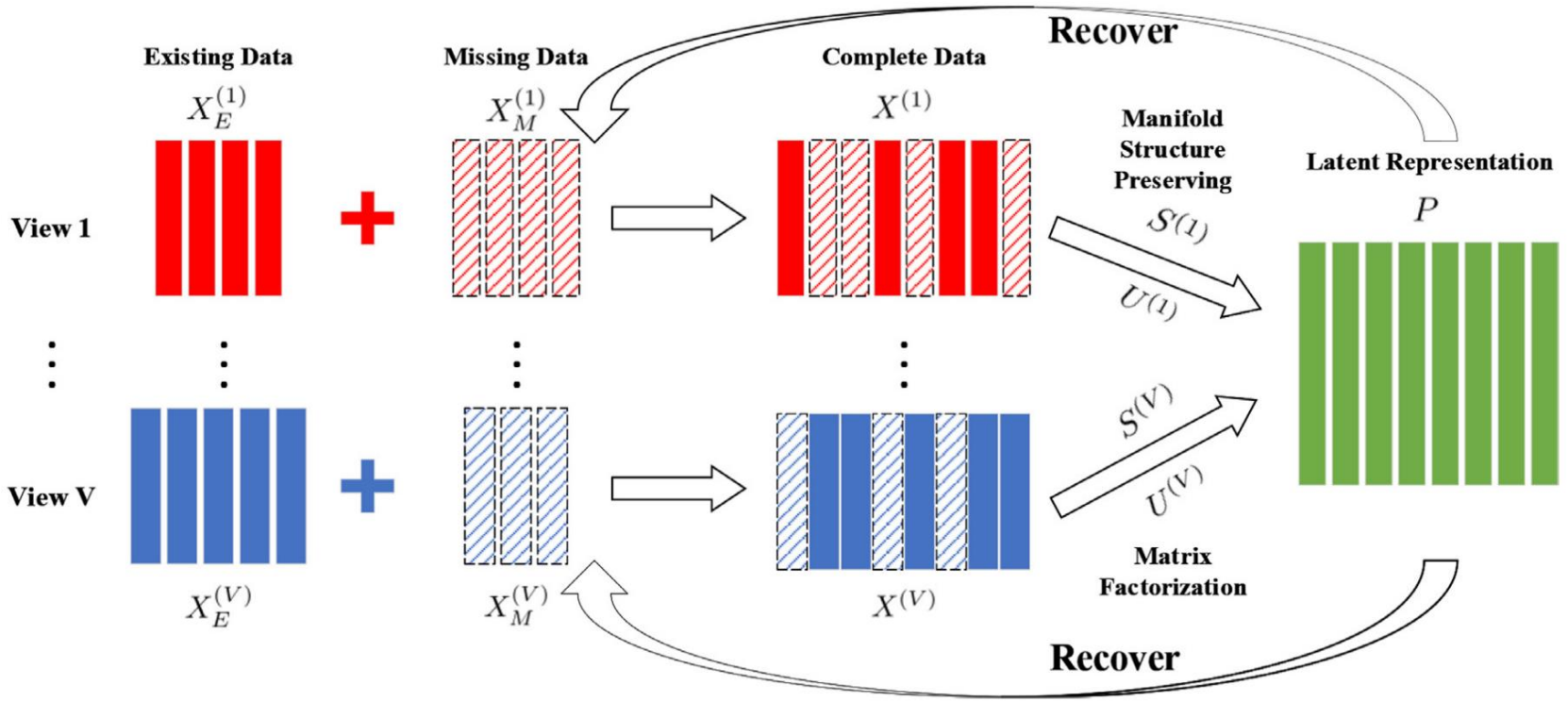
T	目标	使用矩阵分解方法重建缺失的视图信息，提升聚类准确率
I	输入	多视图数据集（图像、文本）*n
P	处理	1.计算每个视图数据的基矩阵和公共系数矩阵 2. 交替更新视图数据、系数矩阵、基矩阵 i)根据基矩阵和系数矩阵重建视图数据 ii)根据重建后的视图数据和系数矩阵更新基矩阵 iii)根据视图数据和新的系数矩阵更新基矩阵 3.输出聚类结果
O	输出	重建好的数据*1份、聚类分配矩阵*1

P	问题	实际应用中会出现视图信息缺失的情况
C	条件	视图数据在处理后可以进行非负矩阵分解
D	难点	在交替更新中同时完成视图数据的重建和聚类
L	水平	Pattern Recognition2022（SCI 二区/模式识别领域顶刊）

IMCCS 算法原理图

- 算法原理
  - 根据已有的多视图信息计算每个视图数据的基矩阵和公共系数矩阵
  - 交替更新视图数据、系数矩阵（公共潜在表示）、基矩阵
  - 达到终止条件后输出聚类结果

$$X^{(V)} \approx U^{(V)} P$$



X_E：现有数据

X_M：缺失数据

S：余弦相似度

U：是视图v的潜在子空间的基矩阵

P：公共潜在表示

IMSCC 算法原理

- 多视图思想
  - 利用NMF重建丢失的视图，同时重建的视图也用于进行NMF
  - 重建的损失函数为：$min_{X^{(V)},U^{(V)},P}\sum_{v=1}^{V}(\left\|X^{(V)}-U^{(V)}P\right\|_F^2+\lambda\left\|U^{(V)}\right\|_F^2)$
    - 不同的重建视图都是基于公共的系数矩阵P导出的，所以每个视图中缺失的数据会在重建过程中被部分地恢复
    - 该过程加强了不同视图之间的联系，使所有视图都能得到统一的利用
- 聚类实现
  - 聚类的损失函数为：$min_P\sum_{v=1}^{V}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|P_i-P_j\right\|_2^2 S_{ij}^{(V)}$，$S_{ij}^{(V)}$为余弦相似度
  - 如果两个数据在原始多视图空间中具有高度的相似性，则使它们在公共潜在空间中彼此接近，以此来实现在重建过程中完成聚类

$$S_{ij}^{(v)}=\frac{X_i^{(v)T}X_j^{(v)}}{\left\|X_i^{(v)}\right\|_2\left\|X_j^{(v)}\right\|_2}$$

IMSCC 算法原理

- 算法整体损失函数
  - 结合多视图与聚类，得到整体的损失函数为：
  - $min_{X^{(V)},U^{(V)},P} \sum_{v=1}^{V}(\left\|X^{(V)} - U^{(V)}P\right\|_F^2 + \lambda\left\|U^{(V)}\right\|_F^2 + \beta \sum_{i=1}^{N}\sum_{j=1}^{N}\left\|P_i - P_j\right\|_2^2 S_{ij}^{(V)})$

- 交替更新
  - 第一步：使用$U^{(V)}$和P重建$X^{(V)}$
  - 第二步：使用重建的$X^{(V)}$和潜在表示P计算新的基矩阵$U^{(V)}$
  - 第三步：使用重建的$X^{(V)}$和新的基矩阵计算新的潜在表示P

Algorithm 1 IMCCS algorithm.

Input:
Incomplete multi-view data $X_E^{(v)}|_{v=1}^{V}$; parameters λ and β.

Output:
Complete multi-view data $X^{(v)}|_{v=1}^{V}$; basis matrices $U^{(v)}|_{v=1}^{V}$; common latent representation P.

1: Initialize $X^{(v)}|_{v=1}^{V}$, $U^{(v)}|_{v=1}^{V}$ and P.
2: **repeat**
3: With fixed $U^{(v)}|_{v=1}^{V}$ and P, update $X^{(v)}|_{v=1}^{V}$ using Eq. (9), Eq. (10), Eq. (11) and Eq. (12).
4: With fixed $X^{(v)}|_{v=1}^{V}$ and P, update $U^{(v)}|_{v=1}^{V}$ using Eq. (15).
5: With fixed $X^{(v)}|_{v=1}^{V}$ and $U^{(v)}|_{v=1}^{V}$, update P using Eq. (19).
6: **until** Eq. (3) converges.

实验设计

- 数据资源
  - 数据集：3Sources、Wikipedia、BBC
- 对比方法
  - BSV、Spec-Pair、Spec-Cent等
- 评价指标
  - 准确率（ACC）
  - 归一化互信息（NMI）
- 实验设置
  - BSV、Multi-NMF、Spec-Pair和Spec-Cent不能直接处理不完整的多视图数据
    - 对于这些方法，使用现有数据的平均值填充缺失数据
  - 在数据集上，随机选取10%到90%的样本作为不完全视图信息，随机选择不完整的例子运行10次，计算平均值

数据集	描述
3Sources	它包含六个主题标签的416个新闻故事，这些故事来自三个新闻来源
Wikipedia	它包含属于10个类别的2866份文件，这些文档是文本-图像对，它们形成双视图数据集
BBC	来自BBC新闻网站的2225份文件组成，这些文件分为五类，通过对文档的分割获得文档的四个视图

实验设计 对比、消融实验

- 实验结果
  - 在3Sources数据集上，IMCCS在所有IER设置下都表现最好
  - 在Wikipedia数据集上同样有着良好的性能
  - 当IER很小时，仍然存在很多具有完整视图数据，所以在IER为10%时BSV取得了更佳的性能

NMI under different IER settings on 3Sources dataset.

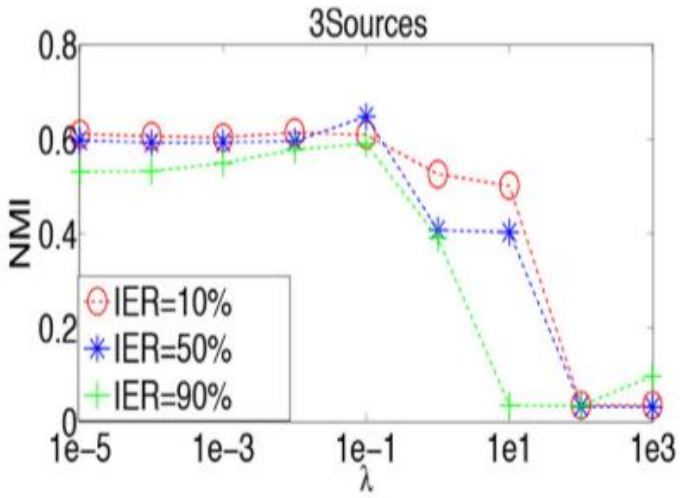
Method\IER	10%	30%	50%	70%	90%
BSV	0.5562	0.5116	0.4438	0.4034	0.3597
MultiNMF	0.4614	0.4541	0.4353	0.4106	0.3922
Spec-Pair	0.4476	0.4233	0.4041	0.3947	0.3633
Spec-Cent	0.4937	0.4772	0.4594	0.4583	0.4172
PVC	0.5497	0.5660	0.5382	0.5303	0.4928
IMVDG	0.5550	0.5368	0.5440	0.5245	0.4952
DAIMC	0.5307	0.4197	0.3813	0.3490	0.3123
SRLC	0.5162	0.4822	0.4688	0.4302	0.4226
UEAF	0.5481	0.5290	0.4917	0.4785	0.4262
IMCCS	**0.6533**	**0.6188**	**0.6217**	**0.6138**	**0.6101**

NMI under different IER settings on Wikipedia dataset.

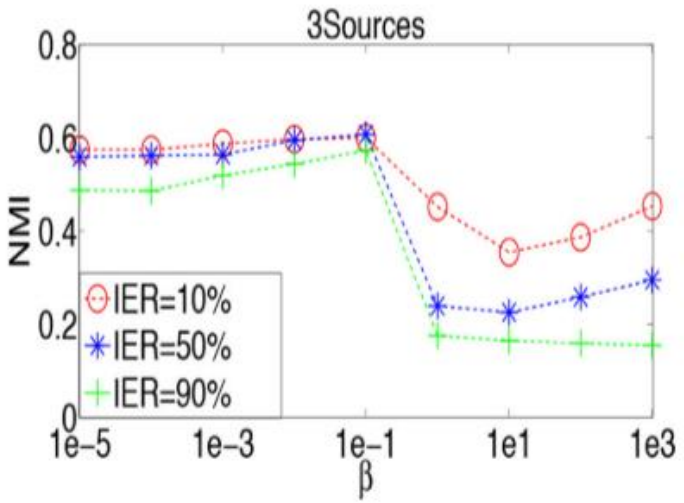
Method\IER	10%	30%	50%	70%	90%
BSV	**0.5201**	0.4502	0.3865	0.3335	0.2731
MultiNMF	0.3621	0.3262	0.2845	0.2502	0.2042
Spec-Pair	0.3926	0.3534	0.3064	0.2427	0.1652
Spec-Cent	0.4435	0.3837	0.3381	0.3022	0.2690
PVC	0.5069	0.4089	0.3411	0.2773	0.2276
IMVDG	0.4993	0.4313	0.3683	0.3125	0.2555
DAIMC	0.4588	0.3176	0.1821	0.0808	0.0254
SRLC	0.3630	0.3472	0.3021	0.2723	0.2319
UEAF	0.4877	0.4212	0.3608	0.3123	0.2616
IMCCS	0.5034	**0.4575**	**0.3994**	**0.3655**	**0.3203**

实验设计 超参实验

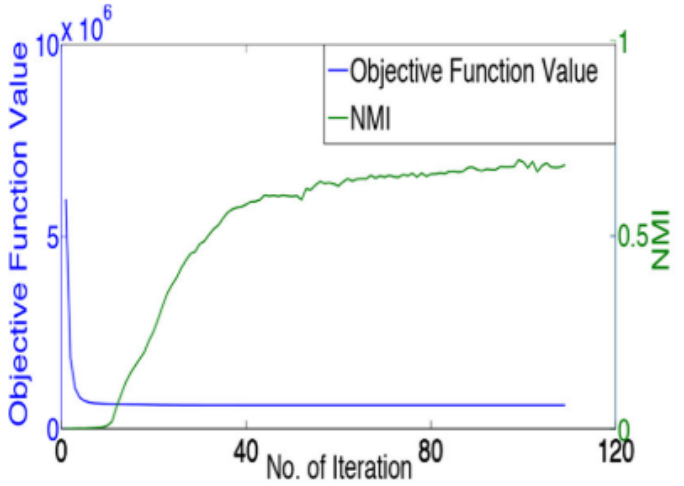
- 实验目的
  - 对参数λ和β以及算法收敛性进行探讨研究
- 实验结果
  - 参数λ和β的最佳取值约为0.1
  - 在目标函数值不变的情况下算法收敛，几乎可以获得最大的NMI



(a) $\beta = 0.1$



(b) $\lambda = 0.01$



(a) BBC

特点总结与未来展望

特点总结与未来展望

- **DMSC**
  - 利用神经网络作为图像的特征提取器，提取不同视角下的特征视图，利用多视图的互补性和一致性增强聚类模型对数据的属性识别，提高聚类准确率
  - 目前只在图像数据集上进行了实验，没有深入研究文本、视频等数
- **IMSCC**
  - 基于非负矩阵的分解和重建将不同视图下的信息进行关联和统一利用，重建了缺失数据，增强聚类模型的鲁棒性
  - λ和β的具体设置需要根据实际应用进行调整
- 未来发展
  - 扩展视图的类型，提升模型的通用性
  - 在大规模数据上实现良好的多视图聚类性能

参考文献

[1] Chen R, Tang Y, Zhang W, et al. Deep multi-view semi-supervised clustering with sample pairwise constraints[J]. Neurocomputing, 2022, 500: 832-845.

[2] Yin J, Sun S. Incomplete multi-view clustering with cosine similarity[J]. Pattern Recognition, 2022, 123: 108371.

[3] Xie Y, Lin B, Qu Y, et al. Joint deep multi-view learning for image clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 33(11): 3594-3606.

[4] Yang Y, Wang H. Multi-view clustering: A survey[J]. Big Data Mining and Analytics, 2018, 1(2): 83-107.

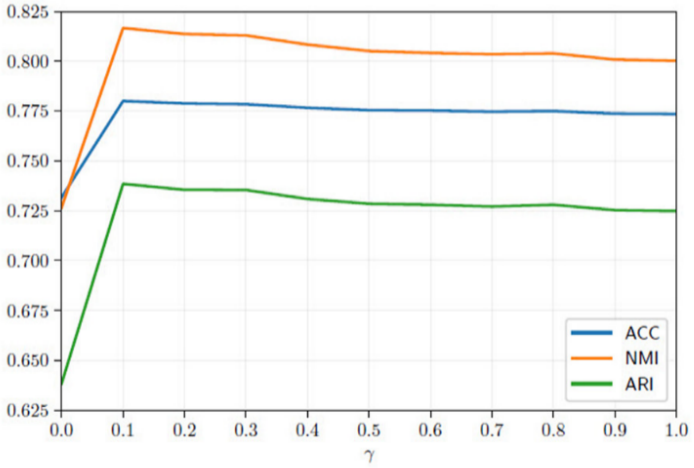
道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

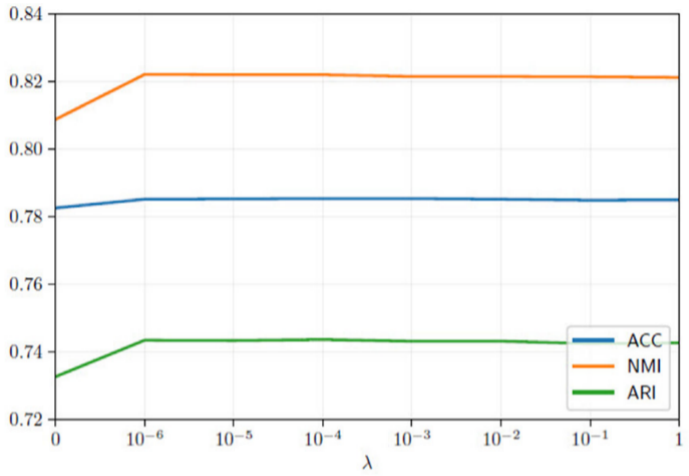
谢谢！

附录A 部分实验结果（DMSC）

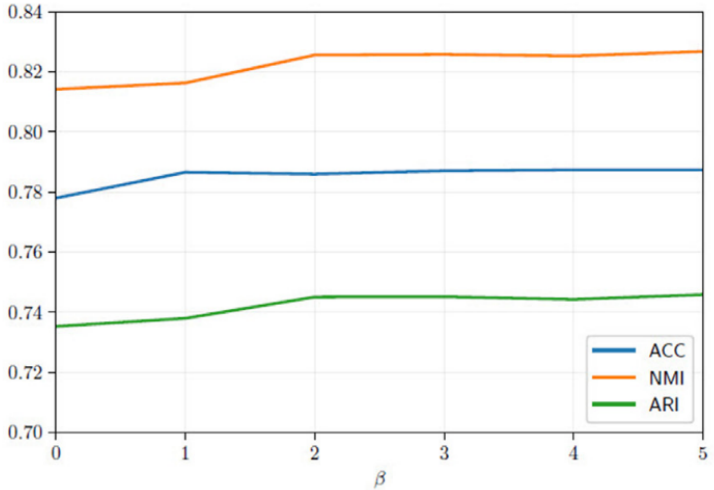
- **超参实验**
  - γ：聚类损失系数
  - λ：约束损失系数
  - β：先验知识比例，即引入成对约束数量与数据总量的比值



(a) USPS



(a) USPS



(a) USPS