

基于自监督信息熵学习的深度多模态聚类

王前前¹⁾, 章子豪^{1)*}, 姜洪旭¹⁾, 冯伟²⁾, 高全学¹⁾, 焦李成¹⁾

¹⁾ (西安电子科技大学通信工程学院 西安 710071)

²⁾ (西安交通大学计算机科学与技术学院 西安 710049)

(qqwang@xidian.edu.cn)

摘 要: 为了保持多个模态间聚类空间的一致性, 消除各个模态内的无关信息, 提出一种基于自监督信息熵学习的深度多模态聚类算法. 首先采用多模态卷积自编码器结合重建任务以获取低维的潜层特征; 然后使用深度嵌入技术为多个模态学习一个理想的公共聚类空间, 将其作为标签以自监督的方式来约束各模态的聚类子空间不断接近理想, 保证每个模态的潜层特征具有相似的分布; 最后结合信息熵的理论, 约束标签与各模态潜层特征间的互信息, 保证模态间的相关性, 同时降低模态内数据的冗余性. 此外, 在 Fashion-MNIST, COIL-20, FRGC, YTF, RGB-D 以及 Noisy-MNIST 基准数据集上展开实验. 实验结果表明, 所提算法在 ACC 和 NMI 聚类指标上均优于其他对比算法, 尤其在 Fashion-MNIST 数据集上, ACC 相较于 StSNE 算法提高了 2.2 个百分点; 消融实验和参数分析证明了所提算法的合理性和鲁棒性.

关键词: 多模态聚类; 信息熵; 自监督学习; 自表达学习

中图法分类号: TP391.41

DOI: 10.3724/SP.J.1089.2023-00624

Deep Multimodal Clustering Based on Self-Supervised Information Entropy Learning

Wang Qianqian¹⁾, Zhang Zihao^{1)*}, Jiang Hongxu¹⁾, Feng Wei²⁾, Gao Quanyue¹⁾, and Jiao Licheng¹⁾

¹⁾ (School of Telecommunications Engineering, Xidian University, Xi'an 710071)

²⁾ (School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049)

Abstract: In order to maintain consistency in the clustering space among multiple modalities and eliminate irrelevant information within each modality, a deep multimodal clustering algorithm based on self-supervised information entropy learning is proposed. Firstly, a multimodal convolutional autoencoder is employed in conjunction with a reconstruction task to acquire low-dimensional latent features. Subsequently, deep embedding techniques are utilized to learn an ideal common clustering space for multiple modalities. This common space is used as labels to supervise and guide each modality's clustering subspaces to progressively approach the ideal state, ensuring that the latent features of each modality have similar distributions. Finally, combining the theory of information entropy, the mutual information between labels and latent features of each modality is constrained to guarantee inter-modality correlation while simultaneously reducing redundancy within modality data. Additionally, experiments are conducted on benchmark datasets including Fashion-MNIST, COIL-20, FRGC, YTF, RGB-D, and Noisy-MNIST. Experimental results demonstrate that

收稿日期: 2023-09-20; 修回日期: 2023-12-18. 基金项目: 国家自然科学基金(62176203, 62102306); 陕西省自然科学基金基础研究计划(2023-JC-YB-534); 西安市科技计划(2022JH-JSYF-0009). 王前前(1990—), 女, 博士, 讲师, 硕士生导师, 主要研究方向为机器学习、模式识别、数据挖掘、多模态学习; 章子豪(2000—), 男, 硕士研究生, 论文通信作者, 主要研究方向为机器学习、模式识别; 姜洪旭(2001—), 男, 硕士研究生, 主要研究方向为模式识别; 冯伟(1992—), 男, 博士, 讲师, 硕士生导师, 主要研究机器学习、信息安全; 高全学(1975—), 男, 博士, 教授, 博士生导师, 主要研究方向为机器学习、模式识别、半监督学习、信号稀疏描述和识别; 焦李成(1959—), 男, 博士, 教授, 博士生导师, 主要研究方向为智能感知与图像理解、深度学习与类脑计算、进化优化与遥感解译.

the proposed algorithm outperforms other comparison algorithms in terms of ACC and NMI clustering metrics. Particularly on the Fashion-MNIST dataset, ACC is improved by 2.2 percentage points compared to the advanced StSNE algorithm.

Key words: multimodal clustering; information entropy; self-supervised learning; self-expression learning

随着信息技术的迅猛发展,以不同模态描述同一对象的多模态数据日益积累.不同模态的数据保留着目标事物中的特定信息,这些信息通常分布在不同空间上,具有互补关系和异构性;同时,它们也共享一致或相似的高层语义特征或类分布结构^[1].通过融合不同模态数据中的一致信息和异构互补信息,能够更全面、更精确地描述同一个对象,从而有效地弥补了单一模态数据在描述对象时的局限性和准确性的不足.因此,在当前信息技术快速发展的时代,多模态数据处理具有极其重要的意义.

随着数据量的指数增长,出于成本考虑,难以采用人工方式精确地标注海量数据,而海量数据在质量上参差不齐,需要进行筛选、去噪等预处理后才能被研究者使用.聚类作为一种无监督学习方法,能够在缺乏数据标签的情况下完成对数据的预处理,已经在图像处理、数据挖掘等方面得到广泛应用^[2].聚类能够根据数据潜在相似性和隐含结构对其进行划分,使得相似性高的数据归为一个子类,达到类内相似而类间相异的目的.由于物体的各个模态能够呈现不同的聚类信息,多模态聚类技术能够充分地利用多个模态间的一致性和互补性,深入挖掘模态内/模态间的复杂信息,从而有效地提升聚类性能,因而得到了学术界的广泛青睐.

然而,传统的聚类算法在应用于多模态聚类时面临着一系列挑战.多模态数据通常具备高维特征,而不同维度之间往往呈现低相关性,这使得大部分基于线性嵌入函数的传统聚类算法无法有效地区分数据中的无关维度.因而聚类性能往往随着数据维度的增加而下降,引发了“维度灾难”的问题.一种有效的解决方案是多模态子空间聚类技术,其通过将高维数据投影到低维子空间,进而找到多个模态共享的低维潜在子空间,可在一定程度上避免了上述的“维度灾难”问题.传统的子空间学习技术,如典型相关分析(canonical correlation analysis, CCA)^[3],旨在学习多模态数据的一致表示,通过最大化不同模态数据的子空间间的相关性,增加模态之间的相似性,进而提高后

续聚类的性能.然而,传统的多模态聚类算法通常采用浅层技术(如线性拟合),只能提取出浅层的线性特征,导致聚类性能不佳.因此,基于多核学习的多视图聚类算法应运而生,如核典型相关分析(kernel canonical correlation analysis, KCCA)^[4].这类算法使用多个核函数来学习多模态数据的复杂结构,能够更好地处理非线性数据.

近年来,深度学习凭借其强大的特征提取能力,让深度聚类研究成为学术界的焦点,涌现了众多颇具创新的深度多模态聚类算法^[5-10].深度多模态聚类旨在探索多个模态的一致性信息,并且挖掘适用于聚类的友好表征.例如,为了模拟任意模态间多层次的非线性相关性,Li等^[5]设计了共享权重多模态自编码器网络,用于将原始特征有效地映射到一个共同的低维嵌入空间;考虑单个视图的独有信息能够有益于聚类性能,一些深度多模态聚类算法^[9-10]考虑通过兼容一致性学习和互补性学习,以获得更全面的信息表示.例如,Xu等^[9]提出的深度嵌入多视图聚类算法,通过协同训练多个深度神经网络,能够很好地利用多视图的共同信息和互补信息.这些算法利用深度学习提取并融合多模态特征,最终采用经典聚类算法对整合的潜层特征进行聚类,构成了当前流行的深度多模态聚类框架.

然而,深度多模态聚类仍然存在一些不足之处.首先,现有的深度多模态子空间聚类算法主要关注提取多模态数据的共享特征,而忽视了模态内部数据和模态之间数据的分布关系.实际上,多模态数据的每个模态都具有其内部的数据分布,并且不同模态的数据之间存在相关性.因此,如何挖掘和建模其模态内数据和模态间数据的分布关系,是进一步优化聚类性能的关键挑战之一.其次,部分深度多模态聚类算法重点考虑了各模态潜层特征间的高度相关性,虽能挖掘到多个模态间的一致性共享信息,但忽略了模态内的冗余信息的干扰,这将不利于提升聚类性能.

针对以上挑战,本文提出了一种基于自监督信息熵学习的深度多模态子空间聚类(deep multimodal subspace clustering based on self-supervised

information entropy learning, DMSC-SI)算法. 采用深度嵌入学习来表征模态数据的分布情况, 并构建了一个理想的分布, 以约束模态之间的分布关系的相似性; 同时, 还利用该分布和信息论原理来减少特征之间的冗余, 以进一步提升聚类性能.

1 相关工作

1.1 传统多模态聚类

传统多模态聚类算法主要包括基于非负矩阵分解、基于子空间聚类的算法、基于 CCA 的算法和基于谱聚类的算法. 例如, Liu 等^[11]提出的联合矩阵分解的算法, 将基于非负矩阵分解方法应用于多模态聚类, 其采用归一化策略对多模态数据直接进行非负矩阵分解, 从而获得多个模态特征的融合表示. Chaudhuri 等^[12]将数据通过 CCA 将多模态数据投影到低维空间, 并在该低维空间进行样本聚类. 此外, Xia 等^[13]提出了鲁棒多模态谱聚类的马尔可夫链算法; Kumar 等^[14]基于规范化聚类假设, 使不同模态的图之间一致, 完成了多模态学习的扩展谱聚类任务. Zhang 等^[15]利用二进制编码技术, 在联合框架中学习了多视图数据的协同二进制编码和聚类结构, 其算法在保持聚类性能的同时, 显著地减少了大型数据集所需的计算时间和内存开销. 尽管这些聚类算法都取得了不错的聚类性能, 但它们所使用的浅层特征提取技术难以处理非线性的高维复杂数据.

1.2 深度多模态子空间聚类

随着深度学习的发展, 基于深度学习的子空间聚类算法得到了国内外学者的广泛关注. Abavisani 等^[16]提出了深度多模态子空间聚类(deep multi-modal subspace clustering, DMSC)算法, 其结合深度学习技巧, 将数据通过非线性投影到潜在子空间中, 以提取和利用模态间共享特征和互补信息. Zhu 等^[17]提出分别对各模态独有的自表达矩阵和模态间共享的自表达矩阵进行学习, 以兼容多模态学习的一致性和互补性. Gao 等^[18]提出基于 CCA 的深度跨模态子空间聚类(cross-modal subspace clustering based on deep CCA, CMSC-DCCA)算法, 在引入了自表达学习的同时, 还考虑了模态内和模态间的相关性. 然而, 上述算法并没有考虑模态的分布差异性, 因此李朝阳^[19]基于生成对抗的思想, 对齐模态间的分布以获取一致性的聚类结构; Wang 等^[20]提出基于自监督 t -SNE 的深度子空间聚类(deep self-supervised t -SNE for multi-modal sub-

space clustering, StSNE)算法, 旨在促使特征分布趋于理想特征分布.

与以往的研究工作不同, 本文采用了一种新的自监督学习算法; 构造了一个理想的公共软标签作为监督标准, 以追求一个具有丰富判别信息的共享子空间表征. 首先, 该监督标准约束模态内和模态间的数据类分布关系, 并将每个模态的分布逼近于理想的聚类分布, 从而保证保持模态间的相关性和一致性; 其次, 它用于指导信息熵理论的学习, 以确保模态间表征的高相关性和最终公共子空间表征的低冗余性. 这种双监督形式能够提高聚类空间的纯净性, 使得样本更加准确地分布在理想的簇中.

2 DMSC-SI 算法

2.1 研究动机

多模态数据呈现多源异质的特点, 多模态聚类的关键挑战在于如何去挖掘模态之间的一些共有信息, 即一致性信息. 目前大多数深度多模态聚类的算法都专注于特征这一层次^[16-17], 通过设计各种借口任务和损失函数来获取良好的特征, 如重构损失. 然而, 现有工作仅关注模态间的一致性, 忽视了数据之间的存在的簇分布关系, 即模态内和模态间的类分布特征^[19-20]. 因此, 为了进一步挖掘这一分布关系, 并将其用于改善聚类性能是本文的动机之一. 此外, 一些基于信息熵的深度跨模态聚类算法, 如跨模态聚类的深度互信息最大最小化(deep mutual information maximin for cross-modal clustering, DMIM)算法^[21], 其关注点也在如何使用互信息来消除模态内的冗余信息. 但是由于缺乏聚类的导向性, DMIM 在面向聚类时很难界定冗余与有效信息的界限, 很可能导致误删除有效的信息, 使得聚类性能无法达到最优. 因此, 本文的另一动机便是将聚类结果作为自监督指导, 构建理想聚类簇分布 $\mathbf{P}$, 使得其能够指导信息熵的学习, 同时进一步消除与聚类无关的冗余信息.

2.2 本文模型设计

2.2.1 模型概述

设模型输入有 N 个样本, 且存在 M 个模态, 则多模态样本可表示为 $\{\mathbf{X}^{(m)}\}_{m=1}^M$. 其中, $\mathbf{X}^{(m)} = \{\mathbf{x}_1^{(m)}, \mathbf{x}_2^{(m)}, \dots, \mathbf{x}_N^{(m)}\} \in \mathbb{R}^{d_m \times N}$ 为第 m 个模态的原始特征矩阵; d_m 为样本的特征维度. 为便于阅读, 表 1 给出了相关的符号说明.

表 1 主要符号定义参照表

符号	含义	符号	含义
N	样本数	$\mathbf{h}_m$	模态 m 的初层特征
M	模态数	$\mathbf{S}$	共享连接矩阵
k	类别数	$\mathbf{C}$	亲和矩阵
$\{\mu_j\}_{j=1}^k$	簇中心	q_{ij}	实际聚类的簇分配概率
$\mathbf{X}^{(m)}$	模态 m 的原始特征矩阵	$\mathbf{Q}$	实际聚类的簇分布
$\hat{\mathbf{X}}^{(m)}$	模态 m 的重构特征矩阵	p_{ij}	理想聚类的簇分配概率
$\mathbf{Z}_m$	模态 m 的潜层特征	$\mathbf{P}$	理想聚类的簇分布

如图 1 所示, DMSC-SI 算法模型包括 4 个模块: 多模态卷积编码器、自监督信息熵学习网络、自表达层网络和多模态卷积解码器. 将多模态样本 $\{\mathbf{X}^{(m)}\}_{m=1}^M$ 输入多模态卷积编码器, 以获得各个模态下的表征 $\{\mathbf{Z}^{(m)}\}_{m=1}^M$. 在自监督信息熵学习网络模块中, 将各个模态的表征融合成公共表征 $\mathbf{Z}$, 并通过理想聚类簇分布 $\mathbf{P}$ 的不断诱导来获得理想的特征表示. 自表达层网络用于捕捉样本之间成对的亲和力, 获取共享连接矩阵 $\mathbf{S} \in \mathbb{R}^{N \times N}$, 以执行谱聚类. 多模态卷积解码器用于重构表征 $\mathbf{ZS}$, 以确保可靠的潜层表征.

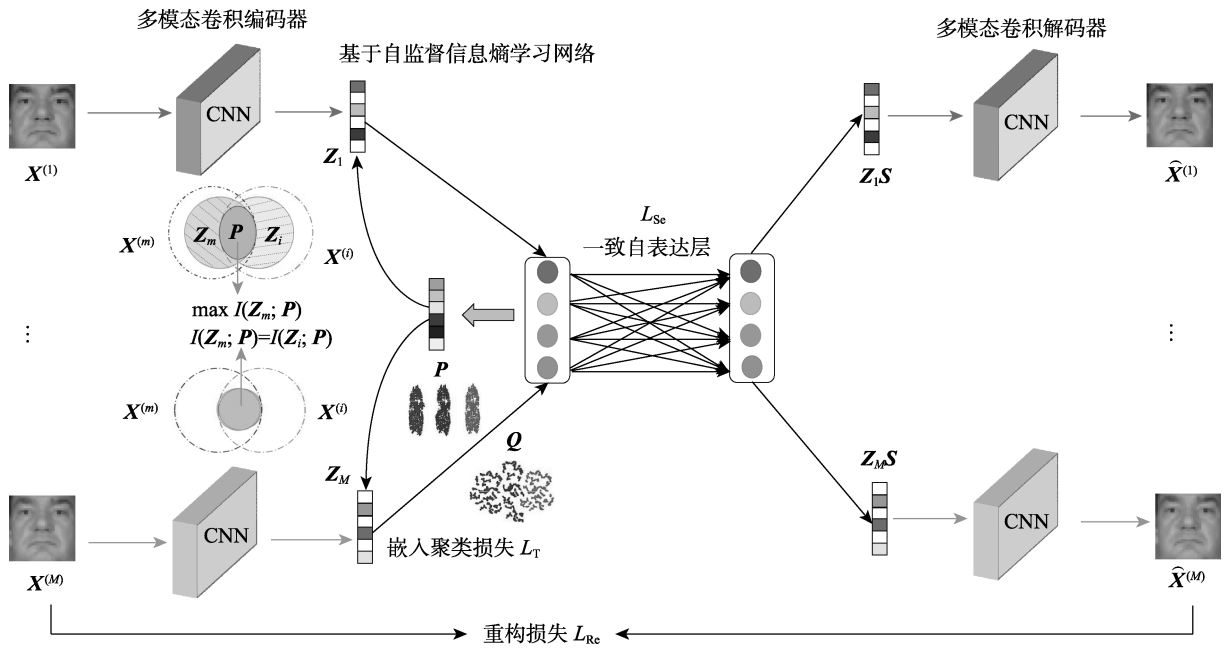


图 1 DMSC-SI 网络框架图

该理想聚类空间将对应的理想聚类的簇分布 $\mathbf{P}$ 作为自监督学习的标签, 能够帮助模型更好地学习区分不同聚类, 并优化公共聚类空间的性能. 因此, 通过引入信息熵约束该聚类空间 $\mathbf{P}$ 与输入数据和子空间表征, 可以进一步共享子空间表征

2.2.2 多模态卷积编码器

对原始高维度数据直接进行处理, 可能会出现维度灾难和特征冗余的问题. 因此, 有必要通过特征转换来降低数据维度, 并提取更具代表性的特征. DMSC-SI 为每个模态设置与其对应的卷积编码器, 其参数记为 $\theta_m (1 \leq m \leq M)$. 即 DMSC-SI 包含 M 个卷积编码器, 它们通过建立非线性映射 $f_m(\mathbf{X}^{(m)}; \theta_m) \rightarrow \mathbf{Z}^{(m)}$ 来转换输入数据. 编码器将该模态下原始数据特征转换为潜层的低维特征 $\mathbf{Z}^{(m)} = \{\mathbf{z}_1^{(m)}, \mathbf{z}_2^{(m)}, \dots, \mathbf{z}_N^{(m)}\} \in \mathbb{R}^{l_m \times N}$; 其中, l_m 表示第 m 个卷积编码器的输出维度.

2.2.3 自监督信息熵学习网络

为充分地考虑各模态之间的信息共享以及所有模态的潜层特征信息, 本文将各个模态的特征进行融合, 以获得公共潜层特征 $\mathbf{Z} = \frac{1}{M} \sum_m \mathbf{Z}_m$.

潜层特征蕴含簇分布结构, 其实际聚类分布缺乏引导, 样本间很难形成其聚类友好的理想簇分布. 因此, 为了解决这个问题, 本文采用 t -Student 分布来衡量公共潜层特征与已知聚类中心之间的概率关系, 并构建一个理想的聚类空间(标签).

学习, 并减少特征的冗余性. 通过这种双重监督形式, DMSC-SI 模型能够不断优化, 进而提高实际聚类空间的质量. 这种优化过程能够提高聚类空间的纯净性, 使得样本更加准确地分布在理想的簇中.

2.2.4 自表达学习层

在同一子空间中, 一个数据点可以用该子空间内其他数据点的线性组合来表示, 这种特性被称为自表达特性. 为了挖掘多模态数据内部的子空间结构, 在 DMSC-SI 中引入了满足自表达特性的矩阵, 即自表达层. 自表达层能够反映样本中点与点的亲和关系, 若将所有模态共享同一个自表达层, 便可以充分地学习多模态数据之间的全局结构. 考虑模态 m 的潜层特征 $\mathbf{Z}_m$ 与共享连接矩阵 $\mathbf{S}$, 自表达层遵循约束条件 $\mathbf{Z}_m = \mathbf{Z}_m \mathbf{S}$. 模型不断地优化能够在 $\mathbf{S}$ 上提取相似度矩阵(邻接矩阵), 并运用谱聚类得到最终的聚类结果.

2.2.5 多模态卷积解码器

为了保证潜层特征能够最大限度地保留原始内在信息, DMSC-SI 在自表达层之后构建了多模态卷积解码器. 该解码器的结构与多模态卷积编码器完全相反、镜像对称. 与卷积编码器相同, 每个卷积解码器也单独重构一个对应模态的重构数据. 重构数据可以通过另一组线性映射来表示, 即 $g_m(\mathbf{Z}_m \mathbf{S}; \delta_m) \rightarrow \hat{\mathbf{X}}^{(m)}$, 其中 δ_m 表示模态 m 的卷积解码器的可学习参数.

2.3 损失函数

2.3.1 嵌入聚类损失函数

对于给定 k 个类别的数据的聚类中心 $\{\mu_j\}_{j=1}^k$, 以 t -Student 分布作为内核来度量公共潜层表示点 $\mathbf{z}_i$ 与中心 μ_j 之间的相似性关系, 即

$$q_{ij} = \frac{(1 + \|\mathbf{z}_i - \mu_j\|^2)^{-1}}{\sum_{j \neq i} (1 + \|\mathbf{z}_i - \mu_j\|^2)^{-1}}.$$

其中, q_{ij} 表示将样本 i 分配给类别 j 的概率(软分配), 它可以反映出当前融合各模态潜层特征后的聚类分布结构. 理想聚类分布应当呈现聚类友好的特点, 即相似数据点更近、差异数据点更远、类内距离较小、类间距离较大. 因此, 对实际分布进行极端化处理, 以追求上述特点, 得到理想聚类分配概率

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_j q_{ij}^2 / f_j}.$$

其中, $f_j = \sum_i q_{ij}$. 由此, 理想聚类的簇分布 $\mathbf{P}$ 可诱导实际聚类的簇分布 $\mathbf{Q}$, 以获得理想的簇表征 $\mathbf{Z}_m$. 本文采用均方误差约束这一关系, 即

$$L_T = \min_{\theta} \sum_{m=1}^M \|\mathbf{Z}_m - \mathbf{P}\|_F^2.$$

其中, $\|\cdot\|_F$ 表示矩阵的 F 范数.

2.3.2 自监督信息熵学习损失

为了挖掘更多模态间共享的特征信息, 并且能够有效地减少潜层特征中与聚类无关的信息. 本文引入自监督信息熵学习损失, 将理想聚类的簇分布作为标签, 以自监督学习的方式指导各模态潜层特征的学习. 以模态 m 为例, 其损失函数可定义为

$$L^m = -I(\mathbf{Z}_m; \mathbf{P}) + \lambda_m I(\mathbf{X}^m; \mathbf{Z}_m | \mathbf{P}) \quad (1)$$

其中, $I(\mathbf{Z}_m; \mathbf{P})$ 项可以度量与标签有关的信息, 即潜层特征 $\mathbf{Z}_m$ 中包含标签中信息量的多少; $I(\mathbf{X}^m; \mathbf{Z}_m | \mathbf{P})$ 用于衡量提取的潜层特征中与标签无关的信息, 即从原始输入 $\mathbf{X}^m$ 中提取出的冗余信息量; λ_m 用于平衡无关信息的舍弃和有关信息的保留.

直接处理式(1)和计算条件概率较为困难, 根据链式法则, 式(1)中的条件概率 $I(\mathbf{X}^m; \mathbf{Z}_m | \mathbf{P})$ 可改写为

$$\begin{aligned} I(\mathbf{X}^m; \mathbf{Z}_m | \mathbf{P}) &= I(\mathbf{X}^m; \mathbf{Z}_m, \mathbf{P}) - I(\mathbf{X}^m; \mathbf{P}) = \\ &= I(\mathbf{X}^m; \mathbf{Z}_m) + I(\mathbf{P}; \mathbf{X}^m | \mathbf{Z}_m) - I(\mathbf{X}^m; \mathbf{P}) = \\ &= I(\mathbf{X}^m; \mathbf{Z}_m) - I(\mathbf{X}^m; \mathbf{P}) \end{aligned} \quad (2)$$

其中, 假设给定特征 $\mathbf{Z}_m$ 时从标签 $\mathbf{P}$ 中捕捉到 $\mathbf{X}^m$ 的信息量为 0, 即 $I(\mathbf{P}; \mathbf{X}^m | \mathbf{Z}_m) = 0$.

由于 $\mathbf{X}^m$ 包含过多与聚类结构无关的冗余信息, 导致训练过程中收敛时间变长、训练参数变多. 因此, 基于 $I(\mathbf{X}^m; \mathbf{P}) \geq I(\mathbf{h}_m; \mathbf{P})$, 将原始输入数据 $\mathbf{X}^m$ 替换为多模态卷积编码器提取的初层特征 $\mathbf{h}_m$.

基于式(1)(2), 模态 m 的损失函数改写为

$$\begin{aligned} L^m &= -I(\mathbf{Z}_m; \mathbf{P}) + \lambda_m (I(\mathbf{X}^m; \mathbf{Z}_m) - I(\mathbf{P}; \mathbf{X}^m)) \leq \\ &= -I(\mathbf{Z}_m; \mathbf{P}) + \lambda_m (I(\mathbf{X}^m; \mathbf{Z}_m) - I(\mathbf{P}; \mathbf{h}_m)) \end{aligned} \quad (3)$$

考虑篇幅和阅读流畅性, 推导流程以及简化思想放置在附录 A 中. 结合模态 m 和 i , 最终基于自监督信息熵学习的损失函数为

$$L_{SI} = (L_{dis}^m + L_{dis}^i) + (L_{cmi}^m + L_{cmi}^i) + (L_{mkl}^m + L_{mkl}^i).$$

2.3.3 自表达学习损失

在理想状态下, 学习到的自表达系数矩阵具有对角块结构, 其每个对角块表示同一个簇. 自表达损失旨在最小化自表达特征和潜层特征之间的全局结构损失, 具体损失函数为

$$\begin{cases} L_S = \min_{\theta, S} \|S\|_1 + \sum_{m=1}^M \|Z_m - Z_m S\|_F^2 \\ \text{s.t. } \text{diag}(S) = 0 \end{cases}$$

其中, $\|S\|_1$ 可避免陷入平凡解 $S = I$ (单位矩阵).

2.3.4 重构损失

以模态 m 为例, 样本 $X^{(m)}$ 输入卷积编码器后, 经过自表达层得到 $Z_m S$, 然后经过解码器输出得到重构样本 $\hat{X}^{(m)}$. 为了提取出更鲁棒的潜层特征, 通过最小化重构样本和原始样本的重构误差, 优化卷积编码器和卷积解码器的参数, 保证提取潜层特征过程的稳定性. 重构损失计算公式为

$$L_{\text{Re}} = \min_{\theta, S, \delta} \sum_{m=1}^M \|X^{(m)} - \hat{X}^{(m)}\|_F^2 \quad (4)$$

2.3.5 总损失

综上, 本节模型的总损失函数可定义为

$$L_{\text{all}} = \min_{\theta, \mu, \phi, S, \delta} L_T + \lambda_S L_S + \lambda_{\text{Re}} L_{\text{Re}} + \lambda_{\text{SI}} L_{\text{SI}} \quad (5)$$

总损失由 t-SNE 损失、自表达层损失、重构损失和自监督信息熵学习损失 4 个部分组成. 其中, θ 表示多模态卷积编码器参数; μ 表示已知的公共聚类中心; ϕ 表示全连接层参数; S 表示共享矩阵; δ 表示多模态卷积解码器参数.

3 实验仿真及其结果分析

本文实验使用的平台为 Ubuntu 16.04 系统, 编程环境为 Python 3.6, 深度学习框架为 Pytorch 1.10.0, 实验均基于 GPU 2080TI 运行.

3.1 数据集介绍

(1) Fashion-MNIST^[22]. 该数据集包含 70 000 幅尺寸为 28×28 像素的图像. 图像为时装产品的 10 个类别, 包括 T 恤、牛仔裤、套衫、裙子等. 本文将原始图像特征和图像边缘特征分别用作第 1 模态和第 2 模态.

(2) COIL-20^[23]. 该数据集由哥伦比亚大学创建, 包含 20 类物体的不同角度的图像. 该数据集共包含每个类别 72 幅, 共 1 440 幅灰度图像, 其尺寸皆为 128×128 像素. 本文实验中将原始灰度图像特征作为第 1 模态, 图像的边缘特征作为第 2 模态.

(3) YTF(Youtube-Face)^[24]. YTF 数据集是一个面部数据集, 共包含 1 595 个人的面部 RGB 图像. 实验选取前 41 个类别图像, 并采用裁剪和对图像尺寸进行调整的操作, 最终得到 55×55 像素大小的共 10 000 幅图像. 本文实验中将数据集中的原

始 RGB 特征作为第 1 模态, 灰度图像作为第 2 模态, 图像边缘特征将作为第 3 模态.

(4) FRGC^[24]. 此数据集为人脸数据集. 实验中, 随机选择了包含 20 个类别的共 2 462 幅图像, 模态提取和选择方式与 YTF 数据集相同.

(5) Sentences NYU v2 (RGB-D). 此数据集由 13 个类别的共 1 449 幅室内场景的图像和每幅图像的描述组成. 依据文献[25], 本文实验中采用从视觉和文本视图中提取的特征.

(6) Noisy-MNIST^[7]. 此数据集包含着 70 000 幅手写数字图像, 涵盖数字 0~9, 共 10 个类别. 该数据集包含 2 个模态数据, 提供了原始灰度图像和添加了高斯噪声背景的手写图像.

3.2 评价指标

本文实验中采用 2 类常见的聚类评价指标对算法性能进行评价, 其数值越高, 意味着聚类效果越好.

聚类准确度(accuracy, ACC)用于比较获得的分类标签和真实标签, 其计算公式为

$$\text{ACC} = \sum_{i=1}^n \zeta(r_i, \text{map}(c_i)) / N.$$

其中, r_i 表示真实标签; c_i 表示聚类后的标签; N 表示样本总数; $\text{map}(c_i)$ 表示映射聚类标签集合到真实标签集合的映射函数; ζ 表示指示函数, 当两者相等时其值为 1, 否则其值为 0.

归一化互信息(normalized mutual information, NMI)提供了一种对称度量, 以量化 2 个聚类分布之间的共享统计信息, 常用于度量 2 个聚类结果的相近程度, 其计算公式为

$$\text{NMI}(\Omega, C) = \frac{I(\Omega, C)}{\max\{H(C), H(\Omega)\}}.$$

其中, Ω 表示按聚类类别标签划分后的样本集; C 表示按真实类别标签划分的样本集. NMI 表示在给定真实类别簇 C 信息的情况下, 聚类类别簇 Ω 中的信息量.

3.3 实验设置

在本文模型中, 每个卷积模态编码器均包含相同的 3 层卷积神经网络结构. 第 1 层卷积网络的卷积核大小为 4×4×10, 卷积步长为 2; 第 2 层卷积网络的卷积核为 3×3×20, 卷积步长为 1; 第 3 层卷积网络的卷积核大小为 4×4×30, 卷积步长为 2. 多模态卷积解码器结构同样由 3 层卷积神经网络组成, 每层网络的卷积核大小和步长都与对应的卷积编码器成镜像相反结构. 对于 RGB-D 和 Noisy-MNIST

向量数据集, 模型的特征提取骨干网络更换为全连接层, 其维度遵循{数据集维度, 1024, 1024, 40}.

此外, DMSC-SI 使用了自表达层作为正则化方法, 在处理大规模数据时存在一定困难. 为了保证实验的可行性, 本文为 Fashion-MNIST 数据集和 YTF 数据集分别随机挑选了 2000 幅图像进行实验. 从数据集 Noisy-MNIST 中挑选 20000 幅图像, 避免绝大多数对比算法超出内存限制.

3.4 算法流程

由于待优化参数较多, 因此本文将模型优化分为预训练和正式训练 2 个步骤.

在预训练阶段, 模型旨在优化式(4), 以获取可靠的表征 $\mathbf{Z}_m$ 和良好的簇中心 $\{\mu_j\}_{j=1}^k$, 为正式训练阶段打下基础. 模型通过自编码器和重构约束, 不断地更新其多模态卷积编码器和解码器参数, 并获得低维潜层表征 $\mathbf{Z}_m$. 对于大型数据集 Noisy-MNIST, 由于后续的自表达层很难对全体样本进行正则化, 因此本文额外考虑在预训练阶段加入最大化 $I(\mathbf{Z}_m; \mathbf{P})$ 约束以辅助训练. 在此阶段, 将通过 $\mathbf{Z}_m$ 实施 K-means 算法, 以获得簇中心 $\{\mu_j\}_{j=1}^k$ 的初始化.

在正式训练阶段, 联合优化所有损失函数, 即优化式(5), 直至网络收敛. 此外, 监测损失函数值, 一旦 2 个时刻的损失函数值低于特定阈值 ϵ , 将判定算法收敛并停止训练.

DMSC-SI 步骤如下:

输入. 多模态样本 $\{\mathbf{X}^{(m)}\}_{m=1}^M$, 类别数 k .

输出. 聚类结果 ACC, NMI.

Step1. 初始化网络参数 θ , δ , μ , ϕ , 以及 $\mathbf{S}$.

Step2. 预训练网络.

Step3. 正式训练网络.

Step4. 提取邻接矩阵 $\mathbf{C} = 1/2(\mathbf{S} + \mathbf{S}^T)$.

Step5. 对邻接矩阵 $\mathbf{C}$ 执行谱聚类.

Step6. 与真实标签对比, 获取聚类结果.

3.5 对比算法

实验将 DMSC-SI 与当下流行的 14 种算法进行对比, 包括 3 种单模态聚类算法: K-means 聚类^[26]、无监督的深度嵌入式聚类分析(unsupervised deep embedding for clustering analysis, DEC)^[27]、数据增强的自适应自步深度聚类(adaptive self-paced deep clustering with data augmentation, ASPC-DA)算法^[28]; 3 种传统多模态聚类算法: 二元多视图聚类(binary multi-view clustering, BMVC)^[15]、鲁棒的多视图 K-means 聚类(robust multi-view K-means clustering,

RMKMC)^[29]、具有共识锚的高效一次性多视图子空间聚类(efficient one-pass multi-view subspace clustering with consensus anchors, EOMSC-CA)^[30]; 8 种深度多模态聚类算法: 深度典型相关分析(deep canonical correlation analysis, DCCA)^[6]、深度典型相关自编码器(deep canonically correlated auto encoders, DCCAE)^[7]、深度广义典型相关分析(deep generalized canonical correlation analysis, DGCCA)^[8]、联合深度多视图学习的图像聚类(joint deep multi-view learning for image clustering, DMJC)^[31], DMSC^[16], CMSC-DDCA^[18], StSNE^[20]和 DMIM^[21].

由于 K-means^[26], DEC^[27]和 ASPC-DA^[28]算法只能处理单模态数据, 因此选择其中表现最佳的模态结果作为记录. 而对于 DCCA^[6], DCCAE^[7], CMSC-DCCA^[18]和 DMIM^[21]算法, 它们只能处理 2 个模态的数据, 因此针对 FRGC 和 YTF 数据集, 对其所有可能的两两组合进行实验, 并选择取得最佳结果的组合作为记录.

3.6 实验结果分析

在基准数据集上, 测试各个算法的聚类性能, 结果如表 2 所示. 结合各对比算法进行分析, 可以得出以下结论.

(1) 本文提出的 DMSC-SI 模型在 6 个数据集上都拥有最佳性能, 充分地验证了本文通过引入自监督标签和信息熵来学习最优化自表达的聚类空间的设计. 该设计增加了聚类子空间的判别性, 同时消除了数据内的冗余信息, 学习到了更准确的模态间一致的聚类结果.

(2) 对于单模态聚类算法, DMSC-SI 的聚类性能更为优越. 在只考虑单一模态信息的情况下, 多模态数据内部的聚类结构特征难以被挖掘, 同时模态间的相关性和互补性被忽视, 造成聚类性能的降低. 例如, 对于 YTF 数据集, DEC 算法^[27]的 ACC 指标仅为 37.10%, NMI 指标仅为 44.60%. 此外, BMVC^[15], RMKMC^[29]以及 EOMSC-CA^[30]都属于浅层的聚类算法; 但鉴于数据集的复杂性, 与深度多模态算法相比, 它们的聚类精度稍显不足.

(3) 由于 DCCA 算法^[6]没有对潜层特征进行数据重构, 无法确保提取特征的可靠性, 可能丢失了大量的有效信息, 因此其聚类性能较差. 而 DCCAE 算法^[7]考虑使用解码器进行重构, 保留了全局的结构分布, 自然使得聚类表现优于 DCCA 算法^[6]; 但与 DMSC-SI 相比, 因为其未关注模态内数据的固有分布结构, 且未考虑模态内数据的分布关系, 使得其聚类性能也较差.

表 2 不同算法在 6 个数据集上的聚类表现对比

%

算法	Fashion-MNIST		COIL-20		FRGC		YTF		RGB-D		Noisy-MNIST	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
<i>K</i> -means ^[26]	51.27	49.99	57.49	73.22	23.62	27.12	56.01	75.23	47.14	40.09	54.64	48.62
DEC ^[27]	51.80	54.60	68.00	80.25	37.80	50.50	37.10	44.60	55.14	41.95	64.33	60.47
ASPC-DA ^[28]	58.45	58.70	61.04	67.90	33.51	32.27	59.95	76.61	45.41	39.35	69.46	62.43
BMVC ^[15]	45.36	38.05	34.31	40.33	41.51	45.92	28.13	38.28	14.28	10.25	81.27	76.12
RMKMC ^[29]	53.32	52.87	60.97	74.93	23.52	25.85	57.21	74.56	45.76	38.38	41.53	40.25
EOMSC-CA ^[30]	52.70	56.72	57.43	74.69	41.31	49.57	59.65	83.73	32.23	18.45	59.96	50.03
DCCA ^[6]	52.74	53.82	55.76	64.91	22.91	24.75	45.19	60.35	39.68	25.78	85.53	89.44
DCCAE ^[7]	55.95	53.01	61.60	71.56	32.33	31.22	45.57	60.15	41.25	29.03	81.60	84.69
DGCCA ^[8]	56.28	57.04	54.01	62.40	23.76	24.53	47.26	61.38	40.22	28.71	82.64	83.47
DMJC ^[31]	61.41	63.41	72.99	81.58	44.07	59.79	61.15	77.40	46.15	40.17	57.16	62.44
DMSC ^[16]	63.20	65.07	74.10	86.82	72.83	80.96	62.80	80.16	57.14	43.35	69.43	61.72
CMSC-DCCA ^[18]	62.95	68.33	82.64	91.45	70.80	78.55	66.15	82.67	57.83	43.26	87.18	82.10
StSNE ^[20]	65.65	68.99	82.43	91.72	74.33	81.44	67.20	83.44	58.30	42.71	73.29	69.95
DMIM ^[21]	56.10	58.06	OOM	OOM	34.12	41.01	57.10	74.80	42.10	26.72	50.79	48.17
DMSC-SI	67.85	69.22	84.09	92.33	75.46	81.86	68.52	84.47	59.14	43.82	88.67	84.86

注. 粗体表示最优值; OOM 表示超出内存限制.

(4) 作为子空间聚类算法在多模态领域的应用, DMSC 算法^[16]存在无法保证模态间数据相关性的问题. 其与 StSNE 算法^[20]相比, 因未考虑冗余信息, 导致聚类性能也不佳. DMSC-SI 引入自监督标签, 通过互信息的约束在保证模态间相关性的同时, 使得各模态潜层特征接近理想, 因而能够得到一致的理想聚类空间.

(5) DMIM 算法^[21]与 DMSC-SI 的方向类似, 旨在最大化跨模态的潜层特征之间的互信息, 并且压缩冗余信息. 但由于其缺乏聚类导向, 很难去界

定关于聚类结构的冗余信息与有效信息, 导致聚类性能无法达到最优. 在 COIL-20 数据集上涉及到一个很大的全连接层, 由于采取过度聚类的方法, 因此导致其运行内存不够.

3.7 可视化结果

在 COIL-20 数据集上, 本文对 6 种对比算法学习到的聚类表示进行了 *t*-SNE 可视化实验, 结果如图 2 所示. 具体而言, 不同算法学习到的公共聚类表示均被应用 *t*-SNE 降维算法, 并生成了用于聚类的嵌入特征. DCCA 算法^[6]和 DCCAE 算法^[7]的聚类

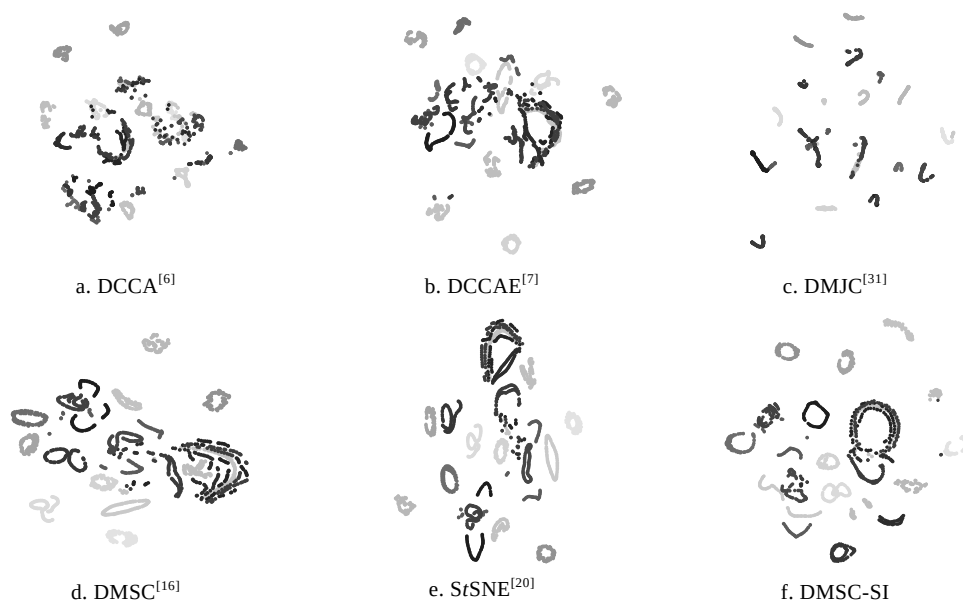


图 2 不同算法在 COIL-20 数据集上的 *t*-SNE 可视化结果

簇存在没有明显边界、聚类结构松散的问题. DMJC 算法^[31]的聚类结构虽然比较紧凑,但聚类精度有待提高,存在不同类别数据混杂的现象. DMSC 算法^[16]在一定程度上改善了聚类结构,但中部与右侧的聚类结构并不明显,类间界限不清晰. StSNE 算法^[20]在 DMSC 算法的基础上优化了聚类结构,显示出了更明显的集簇结构,但聚类的簇不够清晰,同时仍存在较多错误聚类的数据. 因此,与这些算法相比, DMSC-SI 提取出的嵌入特征呈现出更清晰的簇结构,聚类结构更加紧凑;同时,噪声点的数量明显减少,聚类效果显著提高.

3.8 消融实验

为了更好地验证 DMSC-SI 的网络模块,本文在 Fashion-MNIST 数据集上依次去除 t -SNE 损失、自表达层损失、重构损失以及自监督信息熵学习损失进行上述实验,结果如表 3 所示.

表 3 Fashion-MNIST 上的消融实验 %

模块设计	ACC	NMI
无嵌入聚类损失	65.11	68.59
无自表达损失	64.30	67.53
无重构损失	53.42	58.94
无自监督信息熵损失	65.79	66.16
DMSC-SI	67.85	69.22

注: 粗体表示最优值.

与缺失某组成部分的 DMSC-SI 模型相比,完整的 DMSC-SI 模型都具有更好的性能. 具体表现为: (1) 自监督信息熵学习网络对聚类性能有较大帮助,通过 t -SNE 生成的理想标签进行自监督学习,使得各模态聚类结构更接近理想聚类结构;通过互信息约束各模态特征之间的关系,保证了模态间的相关性. (2) 自表达层对模型性能有重大影响,通过构建模态内数据的共享表示学习到模态内的数据分布,构建了模态间一致的聚类结构. (3) 卷积编解码器结构在网络结构中最为重要,保证了各模态特征提取的有效性和可靠性,在很大程度上决定了聚类性能的好坏.

3.9 参数分析

为了探究参数变化对模型性能的影响,本节在 Fashion-MNIST 数据集上进行参数分析实验. 3 个正则化参数 λ_S , λ_{Re} 和 λ_{SI} 在 {0.01, 0.1, 1, 10, 100} 的区间中依次改变,并固定其他参数. 根据实验结果可绘出如图 3 所示的柱状图,分析图 3 可以看出: (1) 当 $\lambda_S=100$, $\lambda_{Re}=0.1$, $\lambda_{SI}=1$ 时, DMSC-SI 取得的结果最好. (2) 尽管 $\lambda_{Re}=0.01$ 时的性能

略逊一筹,但仍能证明算法是较为稳定的,不同参数对聚类性能精度的波动不明显,但微调参数能取得更好的聚类结果.

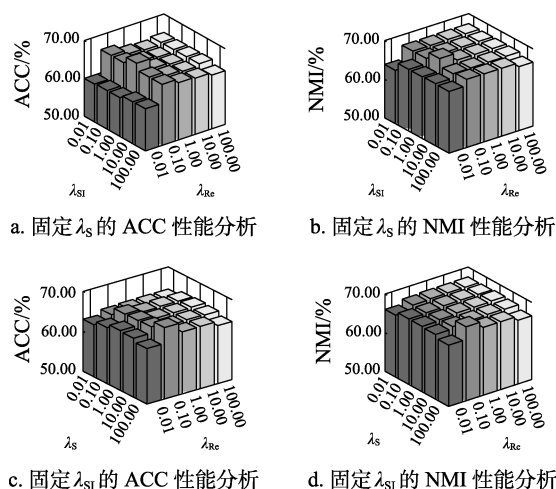


图 3 Fashion-MNIST 数据集上的参数分析实验

4 结 语

本文针对多模态聚类算法的研究现状进行了调研,分析了现有深度多模态聚类算法的局限性和不足,并提出了一种基于自监督信息熵学习的深度多模态子空间聚类算法. 利用深度嵌入学习一个理想的公共聚类空间,将其作为软标签,以自监督的方式约束各个模态的聚类子空间逐渐接近理想聚类空间,深度多模态子空间聚类算法. 利用深度嵌入学习一个理想的公共聚类空间,将其作为软标签,以自监督的方式约束各个模态的聚类子空间逐渐接近理想聚类空间,以确保每个簇的分布更加紧凑;引入信息熵理论,通过最大化软标签信息与各个模态潜层特征之间的互信息,同时最小化原始数据与潜层特征之间的互信息,从而在保证模态间相关性的同时,降低了特征提取的冗余性. 由于理想公共空间具有自监督特性,因此模型能够有效地提取与聚类任务最相关的潜层特征;最后,通过在主流数据集上的测试和对比,有效地验证了 DMSC-SI 具有更好的聚类性能. 在未来的工作中,将重点研究信息熵在缺失多模态聚类中的应用,以应对缺失场景对聚类性能的影响.

参考文献(References):

- [1] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443

- [2] Xu R, Wunsch D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678
- [3] Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639-2664
- [4] Lai P L, Fyfe C. Kernel and nonlinear canonical correlation analysis[J]. International Journal of Neural Systems, 2000, 10(5): 365-377
- [5] Li Z Y, Wang Q Q, Tao Z Q, *et al.* Deep adversarial multi-view clustering network[C] //Proceedings of the 28th International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 2952-2958
- [6] Andrew G, Arora R, Bilmes J, *et al.* Deep canonical correlation analysis[C] //Proceedings of the 30th International Conference on Machine Learning. New York: ACM Press, 2013: III-1247-III-1255
- [7] Wang W R, Arora R, Livescu K, *et al.* On deep multi-view representation learning[C] //Proceedings of the 32nd International Conference on Machine Learning. New York: ACM Press, 2015: 1083-1092
- [8] Benton A, Khayrallah H, Gujral B, *et al.* Deep generalized canonical correlation analysis[C] //Proceedings of the 4th Workshop on Representation Learning for NLP. Stroudsburg: ACL Press, 2019: 1-6
- [9] Xu J, Ren Y Z, Li G F, *et al.* Deep embedded multi-view clustering with collaborative training[J]. Information Sciences, 2021, 573: 279-290
- [10] Zhang C Q, Liu Y Q, Fu H Z. AE²-Nets: autoencoder in autoencoder networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 2572-2580
- [11] Liu J L, Wang C, Gao J, *et al.* Multi-view clustering via joint nonnegative matrix factorization[C] //Proceedings of the SIAM International Conference on Data Mining. Philadelphia: SIAM Press, 2013: 252-260
- [12] Chaudhuri K, Kakade S M, Livescu K, *et al.* Multi-view clustering via canonical correlation analysis[C] //Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 129-136
- [13] Xia R K, Pan Y, Du L, *et al.* Robust multi-view spectral clustering via low-rank and sparse decomposition[C] //Proceedings of the 28th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2014: 2149-2155
- [14] Kumar A, Rai P, Daumé H. Co-regularized multi-view spectral clustering[C] //Proceedings of the 24th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2011: 1413-1421
- [15] Zhang Z, Liu L, Shen F M, *et al.* Binary multi-view clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1774-1782
- [16] Abavisani M, Patel V M. Deep multimodal subspace clustering networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(6): 1601-1614
- [17] Zhu P F, Hui B Y, Zhang C Q, *et al.* Multi-view deep subspace clustering networks[OL]. [2023-09-20]. <https://arxiv.org/abs/1908.01978>
- [18] Gao Q X, Lian H H, Wang Q Q, *et al.* Cross-modal subspace clustering via deep canonical correlation analysis[C] //Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 3938-3945
- [19] Li Zhaoyang. Deep learning-based multimodal clustering analysis[D]. Xi'an: Xidian University, 2021(in Chinese)
(李朝阳. 基于深度学习的多模态聚类分析[D]. 西安: 西安电子科技大学, 2021)
- [20] Wang Q Q, Xia W, Tao Z Q, *et al.* Deep self-supervised t-SNE for multi-modal subspace clustering[C] //Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 1748-1755
- [21] Mao Y Q, Yan X Q, Guo Q, *et al.* Deep mutual information maximin for cross-modal clustering[C] //Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 8893-8901
- [22] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[OL]. [2023-09-20]. <https://arxiv.org/pdf/1708.07747>
- [23] Li F F, Qiao H, Zhang B. Discriminatively boosted image clustering with fully convolutional auto-encoders[J]. Pattern Recognition, 2018, 83: 161-173
- [24] Yang J W, Parikh D, Batra D. Joint unsupervised learning of deep representations and image clusters[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 5147-5156
- [25] Trosten D J, Løkse S, Jenssen R, *et al.* Reconsidering representation alignment for multi-view clustering[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 1255-1265
- [26] Hartigan J A, Wong M A. A K-means clustering algorithm[J]. Journal of the Royal Statistical Society Series C: Applied Statistics, 1979, 28(1): 100-108
- [27] Xie J Y, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis[C] //Proceedings of the 33rd International Conference on Machine Learning. New York: ACM Press, 2016: 478-487
- [28] Guo X F, Liu X W, Zhu E, *et al.* Adaptive self-paced deep clustering with data augmentation[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(9): 1680-1693
- [29] Cai X, Nie F P, Huang H. Multi-view K-means clustering on big data[C] //Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2013: 2598-2604
- [30] Liu S Y, Wang S W, Zhang P, *et al.* Efficient one-pass multi-view subspace clustering with consensus anchors[C] //Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 7576-7584
- [31] Xie Y, Lin B Q, Qu Y Y, *et al.* Joint deep multi-view learning for image clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(11): 3594-3606

附录 A 互信息推导优化

由于式(3)存在多个损失项, 对去掉其模态下标, 进行分类优化, 即

$$\begin{cases} L_{\text{cmi}} = -I(\mathbf{P}; \mathbf{h}) \\ L_{\text{dis}} = -I(\mathbf{Z}; \mathbf{P}) \\ L_{\text{mkl}} = \lambda I(\mathbf{X}; \mathbf{Z}) \end{cases}.$$

(1) 初始特征与标签间的互信息 L_{cmi}

在计算互信息时, 常使用 KL 散度对互信息进行表示, 即 $I(\mathbf{P}; \mathbf{h}) = D_{\text{KL}}[p(\mathbf{P}; \mathbf{h}) \| p(\mathbf{P})p(\mathbf{h})]$. 然而两者分布相距较远时的 KL 散度会难以收敛, 因此使用 JS 散度代替 KL 散度来衡量分布之间的差异^[19], 以追求最大化互信息这一目标, 而不是其具体数值, 即

$$\max_{E(\mathbf{P}; \mathbf{h})} [E_{\mathbf{P}}[D(\mathbf{P}; \mathbf{h})] + E_{\mathbf{P}}[\ln(1 - D(\mathbf{P}; \hat{\mathbf{h}}))].$$

其中, 同一样本生成的 $\mathbf{P}$ 与 $\mathbf{h}$ 定义为正样本对, 模态内随机采样的 $\hat{\mathbf{h}}$ 与 $\mathbf{P}$ 定义为负样本对. D 表示判别器, 用于估计输入样本对的概率. E 表示分布期望.

(2) 潜层特征与标签间的互信息 L_{dis}

类似地, $I(\mathbf{Z}; \mathbf{P})$ 可由

$$\max_{E(\mathbf{P}; \mathbf{Z})} [E_{\mathbf{P}}[D(\mathbf{P}; \mathbf{Z})] + E_{\mathbf{P}}[\ln(1 - D(\mathbf{P}; \hat{\mathbf{Z}}))]$$

优化.

(3) 潜层特征与原始输入间的互信息 L_{mkl}

采用变分近似的方法计算互信息 $I(\mathbf{X}; \mathbf{Z})$, 假定分布 $q(\mathbf{Z})$ 为分布 $p(\mathbf{Z})$ 的变分近似, 将“求两类远近”问题转化为求“两分布距离”的优化问题, 即

$$\begin{aligned} I(\mathbf{X}; \mathbf{Z}) &= \int_{\mathbf{Z}} \int_{\mathbf{X}} p(\mathbf{X}; \mathbf{Z}) \ln \frac{p(\mathbf{Z}, \mathbf{X})}{p(\mathbf{Z})p(\mathbf{X})} d\mathbf{X} d\mathbf{Z} = \\ &= \int_{\mathbf{Z}} \int_{\mathbf{X}} p(\mathbf{X}; \mathbf{Z}) \ln \frac{p(\mathbf{Z} | \mathbf{X})}{p(\mathbf{Z})} d\mathbf{X} d\mathbf{Z} = \\ &= \int_{\mathbf{Z}_n} \int_{\mathbf{X}} p(\mathbf{X}; \mathbf{Z}) \ln \frac{p(\mathbf{Z} | \mathbf{X}) q(\mathbf{Z})}{p(\mathbf{Z}) q(\mathbf{Z})} d\mathbf{X} d\mathbf{Z} = \\ &= \int_{\mathbf{Z}} \int_{\mathbf{X}} p(\mathbf{X}; \mathbf{Z}) \ln \frac{p(\mathbf{Z} | \mathbf{X})}{q(\mathbf{Z})} d\mathbf{X} d\mathbf{Z} - \\ &= \int_{\mathbf{Z}_n} p(\mathbf{Z}) \ln \frac{p(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} = \\ &= E_{p(\mathbf{X})} [D_{\text{KL}}(p(\mathbf{Z} | \mathbf{X}) | q(\mathbf{Z}))] - \\ &= E_{p(\mathbf{X})} [D_{\text{KL}}(p(\mathbf{Z}) | q(\mathbf{Z}))] \leq \\ &= E_{p(\mathbf{X})} [D_{\text{KL}}(p(\mathbf{Z} | \mathbf{X}) | q(\mathbf{Z}))]. \end{aligned}$$

由于 $D_{\text{KL}}(p(\mathbf{Z}) | q(\mathbf{Z})) \geq 0$, $I(\mathbf{X}; \mathbf{Z})$ 具有最小上界. 通过最小化其最小上界, 可以达到最小化 $I(\mathbf{X}; \mathbf{Z})$.