软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006167]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

可靠多模态学习综述*

杨 杨[1], 詹德川[2], 姜 远[2], 熊 辉[3]

[1](计算机科学与工程学院(南京理工大学), 江苏 南京 210094)

[2](计算机软件新技术国家重点实验室(南京大学), 江苏 南京 210023)

[3](罗格斯商学院, 美国 纽瓦克 07012)

通讯作者: 詹德川, E-mail: zhandc@nju.edu.cn

摘 要: 近年来多模态学习逐步成为机器学习、数据挖掘领域的研究热点之一,并成功应用于诸多现实场景,如跨媒介搜索、多语言处理、辅助信息点击率预估等.传统多模态学习方法通常利用模态间的一致性或互补性设计相应的损失函数或正则化项进行联合训练,进而提升单模态及集成的性能.而开放环境下,受数据缺失及噪声等因素的影响,多模态数据呈现不均衡性.具体表现为单模态信息不充分或缺失,从而导致“模态表示强弱不一致”、“模态对齐关联不一致”两大挑战,而针对不均衡多模态数据直接利用传统的多模态方法甚至会退化单模态和集成的性能.针对这类问题,可靠多模态学习被提出并进行了广泛研究,本文系统地总结和分析了目前国内外学者针对可靠多模态学习取得的进展,并对未来研究可能面临的挑战进行展望.

关键词: 不均衡多模态数据;模态表示强弱不一致;模态对齐关联不一致;可靠多模态学习

中图法分类号:TP311

中文引用格式: 杨杨,詹德川,姜远,熊辉.可靠多模态学习综述.软件学报. http://www.jos.org.cn/1000-9825/ 0000.htm

英文引用格式: Yang Yang, De-Chuan Zhan, Yuan Jiang, Hui Xiong. Reliable Multi-Modal Learning: A Survey. Ruan Jian Xue Bao/Journal of Software, 2019 (in Chinese). http://www.jos.org.cn/1000-9825/0000.htm

Reliable Multi-Modal Learning: A Survey

YANG Yang[1], ZHANG De-Chuan[2], JIANG Yuan[2], XIONG Hui[3]

[1](College of Computer Science and Engineering (Nanjing University of Science and Technology), Nanjing, Jiangsu 210094, China)

[2](State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

[3](Rutgers Business School, Newark 07012, US)

Abstract: Recently, multi-modal learning is one of the important research fields of machine learning and data mining, and has a wide range of practical applications, such as cross-media search, multi-language processing, auxiliary information click-through rate estimation, etc. Traditional multi-modal learning methods usually use the consistency or complementarity among modalities to design corresponding loss functions or regularization terms for joint training, thereby improving the single-modal and ensemble performance. However, in the open environment, affected by factors such as data missing and noise, multi-modal data is imbalanced, specifically manifested as insufficient or incomplete, resulting in “inconsistency modal feature representations” and “inconsistent modal alignment relationships”. Direct use of traditional multi-modal methods will even degrade single-modal and ensemble performance. To solve these problems, reliable multi-modal learning has been proposed and studied. This paper systematically summarizes and analyzes the progress made by domestic and foreign scholars on reliable multi-modal research, and the challenges that future research may face.

* 基金项目: 国家自然科学基金(61673201, 62006118, 61773198, 61632004);CCF-百度松果基金(CCF-BAIDU OF2020011);百度 TIC 项目基金

Foundation item: National Natural Science Foundation of China (61673201, 62006118, 61773198, 61632004); CCF-BAIDU Songguo Foundation(CCF-BAIDU OF2020011); BAIDU TIC Foundation

收稿时间: 2019-06-17; 修改时间: 2020-04-28; 采用时间: 2020-05-20; jos 在线出版时间: 2020-12-02

Key words: Imbalanced multi-modal data, Inconsistent modal feature representations, Inconsistent modal alignment relationships, Reliable multi-modal learning

1 引言

“一本《红楼梦》,经学家看见《易》,道学家看见淫,才子看见缠绵,革命家看见排满,流言家看见宫闱秘事.”——鲁迅.

现实世界中,复杂对象从不同角度分析拥有不同的属性特征.如图 1 所示,网页包含文本、图片和超链接等信息;视频可以分解为图片帧、音频和文本;文章可以通过不同语言表示;手机应用从不同传感器收集信息进行分析等.可见,样本可以通过不同通道的信息加以描述,每一通道信息定义为一种特定的模态.因此,较之于单模态数据,多模态数据可以提供更丰富的信息表示,且基于多模态数据表示也有着极其广泛的应用,如基于图文数据的热点推荐、基于多传感器信号的无人驾驶、基于视频语音的字幕生成等.

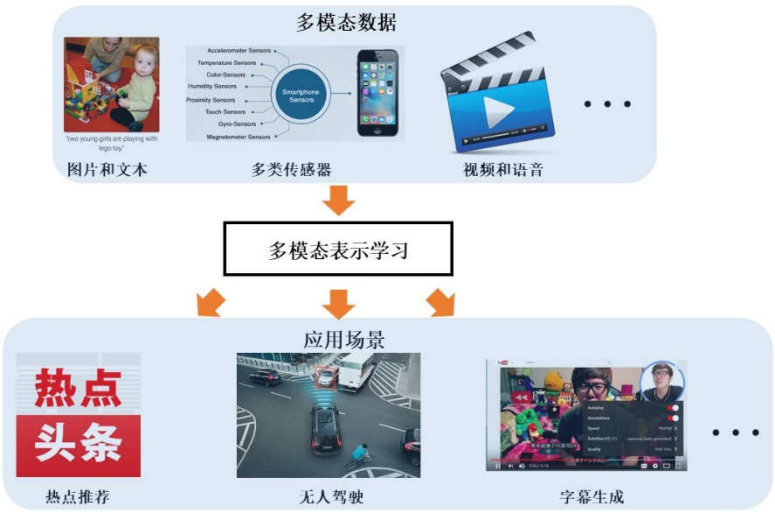


图 1. 多模态数据及应用.现实应用中复杂对象通常可以通过多模态信息加以描述,
多模态学习也有着广泛的应用场景.

较之于单模态学习,多模态学习通常考虑两方面研究内容:(1)单模态学习性能;(2)模态间相关性度量及利用.采用的主要策略是将二者纳入统一框架中进行联合优化,进而为每个模态学习更具有判别性的语义表示,构建模态间的映射关联,提升模型性能.具体地,传统多模态方法大致可分为两类:(1)基于协同训练思想的方法;(2)基于协同正则化思想的方法.协同训练(Co-training)[1]是多模态学习早期学习方法之一,其利用模态间的互补性准则,最大化两个不同模态未标记数据的互一致性(即挑选最置信的未标记样本标记伪标记,提供给其他模态学习)提升单模态的性能.基于此思想设计出众多衍生方法,如 Co-EM[2]、Bayesian co-training[3]、Co-Trade[4]等.作为多模态学习的另一个重要分支,协同正则化(Co-regularization)[5]则是利用模态间的一致性准则,最小化两个不同模态未标记数据的预测差异性来排除不一致的假设.进一步,研究者基于该思路提出其他模型,如 SVM-2K[6]、MSE[7]等.此外,另一方面基于子空间学习方法(如 CCA[8])、基于多核学习方法(如 MKL[9])也可归为利用一致性准则的协同正则化方法.值得注意,早期基于互补性准则的协同训练类型方法通过各模态最置信的未标记样本的伪标记信息进行相互教学,其本质也可看作潜在标记的一致性,因此传统的两类方法都关注利用样本不同模态间的强相关性.相对于早期传统的多模态学习方法,近些年一些研究转而注重学习或度量模态间的互补信息表示,以此增强模态的融合性能[10],本文将在 2.2.3 节具体介绍该类方法.同时,多模态理论研究也有所建树,如协同训练的泛化界[11],基于信息熵的多模态理论框架[12].然而在开放环境下,考虑信息缺失、噪声干扰等问题,模态间的强相关性难以满足,传统多模态学习方法仍面临巨大挑战.同时,多模态学习与机器学习中

的其他研究领域也紧密相关,研究内容丰富,如集成学习[13]、领域适配[14]、主动学习[15],考虑与本文主题关联较低,本文不一一赘述.

1.1 多模态学习面临的挑战

真实开放环境下,多模态数据通常会受到噪声、自身缺陷及异常点等干扰,使得上述互补性及一致性准则难以满足.究其原因,主要体现在学习过程中出现的未标记样本伪标记噪声、采样偏差,及模态特征表示、模型性能差异等问题,进而导致模态表示强弱以及模态对齐关联的不一致.具体表示为:

1） 模态表示强弱不一致.传统多模态学习方法通常考虑模态间的一致性,即特征或预测的一致性.而在开放环境下,噪声等因素会造成单模态的信息不充分[16],进而导致单模态特征、预测的噪声和模态间的差异性,造成模态存在强弱之分.直接使用传统的互补性或一致性准则会造成模型优化偏差,影响模型联合训练;

2） 模态对齐关联不一致.传统多模态学习方法通常假设同一样本拥有全量的模态信息,且模态间的关联关系也是事先确定的.而开放环境中,考虑到隐私保护、数据收集缺陷等因素,多模态数据存在模态缺失问题[17],即样本可能仅获得部分模态信息,而非全量信息.同时,考虑到人工标注代价等因素,同一任务获得的不同模态间的对应关系也可能不明确[18].

综上所述,模态表示强弱不一致和模态对齐关联不一致是多模态数据在开放环境下凸显的两大新挑战,也是造成传统多模态学习方法在真实数据集上甚至出现性能退化现象的关键因素.针对这些挑战,可靠多模态学习(也称鲁棒多模态学习)开始受到国内外研究的广泛关注.针对模态表示强弱不一致问题,文献[19][20]提出利用强模态作为软监督信息辅助弱模态,文献[21][22]考虑加权等操作排除不一致样本的干扰;针对模态关联不一致问题,文献[17]考虑缺失模态的聚类,文献[23]考虑不对齐多模态的融合.

1.2 多模态学习的主要技术与应用

表 1. 多模态学习的主要技术与应用[24]

方法 / 应用	表示学习	映射学习	对齐学习	融合学习	协同学习
语言合成与辨认					
视听语音识别	✓		✓	✓	✓
(视觉)语音合成	✓	✓			
事件检测					
动作分类	✓			✓	✓
多媒体事件检测	✓			✓	✓
情绪情感					
识别	✓		✓	✓	✓
合成	✓	✓			
媒介描述					
图片描述	✓	✓	✓		✓
视频描述	✓	✓	✓	✓	✓
语音问答	✓	✓		✓	

媒体总结					
媒介搜索 跨模态搜索 跨模态哈希	✓ ✓	✓	✓		✓ ✓

目前已有一些关于多模态学习的综述发表[24][25]错误！未找到引用源。[45],这些综述大多着重于总结传统多模态学习方法及其应用.例如,文献[25]总结了传统多模态子空间学习、多核学习及协同学习,并给出了当前深度多模态学习的进展;文献[24]**错误！未找到引用源。**则从多模态应用层面出发介绍相关的学习方法,包括:(1)模态表示学习;(2)模态映射学习;(3)模态对齐学习;(4)模态融合学习;(5)模态协同学习,并给出其在视觉领域、多媒体领域的诸多应用.表 1 给出了上述五种多模态技术在不同实际场景中的具体应用.

值得注意,大多综述忽略了 1.1 节中描述的多模态学习所面临的挑战,为此本综述将具体分析针对这两个挑战的国内外相关研究现状,并介绍目前可靠多模态学习的研究进展.

1.3 论文的组织

本文首先概述传统多模态学习中基于互补性和一致性准则的方法,其次具体分析开放环境下多模态数据凸显的“模态表示强弱不一致”、“模态对齐关联不一致”两大挑战,并介绍目前针对这两个问题的可靠多模态学习研究进展,内容安排的具体框架如图 2 所示.特别地,随着深度学习的兴起,适应不同领域的深度模型均取得远超传统模型的性能,而目前先进的多模态方法也通常选择相应的神经网络,如卷积神经网络、长短记忆神经网络作为各模态(图片、文本)的基模型,并设计相应的损失函数进行联合训练,为此本文也会着重介绍目前高性能的多模态深度学习模型.

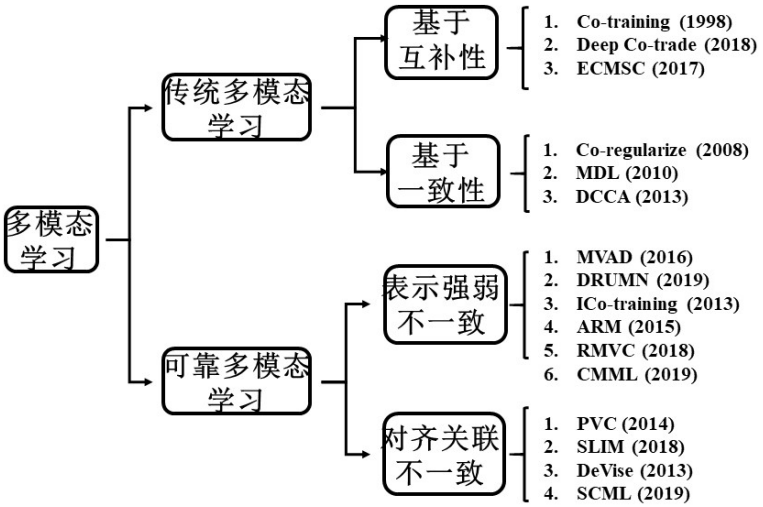


图 2. 论文的整体组织框架.包括传统多模态学习和可靠多模态学习.

2 传统多模态学习

本节先介绍多模态学习的两种基本准则后,具体介绍相应的学习方法.在没有特殊说明情况下,本文所介绍的方法一般以两模态为例,不失一般性,扩展到多模态通常采用两两遍历加和形式.

2.1 两种基本准则

传统多模态学习的精髓在于如何有效地考虑模态间的关联性,通常要求服从两个基本准则:互补性和一致性.互补性准则描述每个模态的数据可能包含其他模态所欠缺的信息,因此综合考虑多模态信息可以更全面地描述数据并提升任务性能.具体地,假设数据集X包含两个模态X_1和X_2,进而单样本可以表示为(x_i^1, x_i^2, y_i),其中y_i是标记信息.数据满足以下三个假设:(1)充分性,即每个模态自身含有充分信息进行分类;(2)兼容性,即两个模态大概率具有共现特征,能够预测相同标签;(3)条件独立,即给定标签情况下模态条件独立.基于上述假设,文献[1]给出如下结论:如两个模态是条件独立的,那么协同训练会提升单模态性能.文献[11]则进一步给出了基于 PAC 理论的协同训练的泛化误差界,证明两个模态的一致性是单模态模型性能的上界.考虑到条件独立假设过强,因此文献[27][28]等工作进一步放松该假设,并给出相应的泛化误差理论证明.

相对于互补性准则,一致性准则旨在最大化两个不同模态的一致性.假设数据集X包含两个模态X_1, X_2,文献[29]证明两个模态的一致性和单模态错误率之间的关联为:

$$P(f^1 \neq f^2) \geq max\{P_{\{err\}}(f^1), P_{\{err\}}(f^2)\}.$$

依据上式可以得出两个独立模态模型不一致的概率是单模态模型最大错误率的上界. 因此,通过最小化两个模态模型的不一致,每个模态模型的错误率将被最小化.殊途同归,可以看出互补性本质上也是一致性的一个变种.

2.2 基于互补性准则的方法

2.2.1 Co-training

Co-training[1]假设样本有两个条件独立的模态,给定L个有标记样本和U个无标记样本,Co-training 采用如下的迭代训练方式:

Step1. 无放回的从无标记数据集U构造数据池U';

Step2. 分别用两个模态X_1, X_2的有标记数据训练两个朴素贝叶斯学习器(可替换其他弱学习器)h_1, h_2;

Step3. 每个模态用训练好的学习器在U'中为本模态挑选p个最置信正例和n个最置信负例的无标记样本,标上伪标记加到L中重训练.从而X_1可以获得X_2互补的信息,X_2也可以获得X_1互补的信息;

Step4. 从U中重新填充$2p + 2n$个样本到数据池U'.

2.2.2 Deep Co-trade

基于集成学习的思想,文献[4]提出 Co-trade 算法.该算法首先对有标记数据进行可重复取样得到三个训练集并训练三个对应的学习器,且在协同训练的过程中,每个学习器获得的新数据集合都是通过其他两个学习器投票得到.同时,随着深度网络的成功应用,文献[30]基于 Co-trade 的思想提出了 Tri-net.如图 3 所示,Tri-net 首先对训练数据用不同大小的卷积核构造三个不同的训练集,并且采用 Output Smearing 技术(对训练集的真实标记加入随机噪声)来构造差异性更大的无标记数据.随后采用 Tri-training[31]的思想对无标记数据预测标记并带回训练集重新训练.

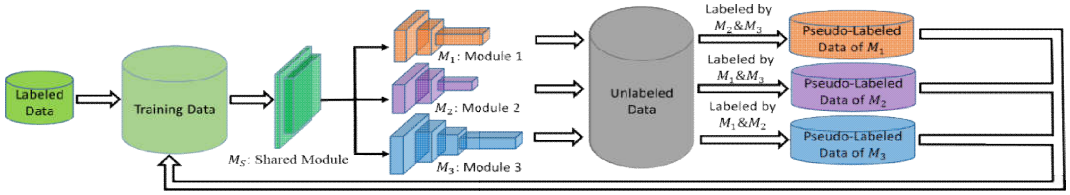


图 3. Tri-net 示意图[30].采用多个学习器集成学习.

而扩展到两模态以上的场景下,Tri-net 也可以衍生很多变种,包括:(1)为每个模态建立学习器,再采用集成思想结合其他模态学习器为当前模态的无标记数据投票得到新标记;(2) 为每个模态基于 Tri-training 思想建立多个学习器,再用两层的堆叠技术(stacking)为无标记数据投票得到新标记.

2.2.3 ECMSC

不难看出,传统协同训练方法局限于运用标记相互教学,仍属于潜在的标记一致,缺乏学习量化模态间的互补信息.因此,文献[10]提出一种新颖的多模态聚类方法 ECMSC (Exclusivity-Consistency Regularized Multi-view Subspace Clustering),ECMSC 兼顾多模态特征表示的差异性和聚类指示矩阵的一致性,其新颖点在于使用了差异化正则凸显模态的互补信息.差异性可通过如下的矩阵 Hadamard 乘积来定义:

定义 1. 两个矩阵$U \in \mathbb{R}^{n\times n}$和$V \in \mathbb{R}^{n\times n}$之间的差异性定义为:$\mathcal{H}(U,V) = \|U\odot V\|_0 = \sum_{i,j}(u_{ij}\cdot v_{ij} \neq 0)$,其中$\odot$表示 Hadamard 乘积(对应位相乘),$\|\cdot\|_0$表示$\ell_0$范数.

ℓ_0范数可以放松到ℓ_1范数,于是两个模态聚类结果的差异性可以表示为$\mathcal{H}(Z_v,Z_w) = \|Z_v\odot Z_w\|_1$.

每个模态聚类指示矩阵和潜在一致的聚类指示矩阵的关联可以延用以往常用约束,具体为:$min\sum_{v=1}^{V}\|Z_v\odot F\|_1\ s.t.\ F^\top F = I$.

将定义 1 中的差异性正则扩展到多模态谱聚类中,新模型表示为:

$$\min_{F,Z_v}\sum_{v=1}^{V}\|E_v\|_1 + \lambda_1\|Z_v\|_1 + \lambda_2\sum_{w\neq v}\|Z_v\odot Z_w\|_1 + \lambda_3\|Z_v\odot F\|_1$$

$$s.t.\ \forall v, X_v = X_vZ_v + E_v, diag(Z_v) = 0, F^\top F = I$$

其中$\|Z_v\|_1$的作用是保证稀疏性,约束项中每个模态的聚类指示矩阵则可以看成字典学习的表示形式,噪声损失项则采用ℓ_1范数来处理稀疏噪声.

该模型的本质思想也是一种对抗学习,一方面希望体现不同模态的差异性(第二项),另一方面则希望单模态的聚类指示函数与潜在真实的聚类指示矩阵一致(第三项).在优化方面,ECMSC 也可以采用 ADMM 进行并行优化.值得注意,第二项的差异正则实质上可以采用很多其它的形式,如 HSIC 等.

2.3 基于一致性准则的方法

基于一致性准则的方法可以分为:(1)约束模态预测一致性;(2)约束模态特征表示的一致性.

2.3.1 Co-regularization

半监督学习方法协同正则化(Co-regularization)[5]考虑预测的一致性约束.具体地,给定少量有标记数据(x_i,y_i)和大量的无标记数据(x_j),协同正则化为每一个模态学习一个最优学习器:

$$\min_{f_1\in H_1, f_2\in H_2}\sum_{i=1}^{N_l}\ell(f_1(x_{i^1}),f_2(x_{i^2}),y_i) + \gamma_1\|f_1\|_{H_1}^2 + \gamma_2\|f_2\|_{H_2}^2 + \sum_{j=N_l+1}^{N_l+N_u}\left\|f_1(x_{j^1}) - f_2(x_{j^2})\right\|_F^2$$

其中$f_1 \in H_1, f_2 \in H_2$分别是两个模态的学习器,H_1, H_2是两个模态的假设空间.$\ell(f_1(x_{i^1}),f_2(x_{i^2}),y_i)$计算两个模态预测集成结果和真实结果的损失.不失一般性,ℓ一般取平方损失,即$\|y_i - (f_1(x_{i^1}) + f_2(x_{i^2}))/2\|_F^2$,$\gamma_1\|f_1\|_{H_1}^2$,$\gamma_2\|f_2\|_H^2$运用 RKHS 范数度量模型 c 复杂度.起关键作用的最后一项$\left\|f_1(x_{j^1}) - f_2(x_{j^2})\right\|_F^2$则是强制不同模态在无监督数据上的一致性,$N_l, N_u$是有标记数据和无标记数据的大小.文献[32]证明通过度量两个函数类的“距离”可以约束无标记数据的一致性,进而降低 Rademacher 复杂度.测试阶段,样本预测结果为:

$$f^*(x) = \frac{1}{2}(f_1^*(x) + f_2^*(x))$$

2.3.2 DCCA

典型性相关分析 CCA(Canonical Correlation Analysis)[8]则是约束模态特征表示的一致性.具体地,对于$X_1 \in \mathbb{R}^{d_1\times N}, X_2 \in \mathbb{R}^{d_2\times N}$两个模态数据,每个模态学习投影向量$\omega_1 \in \mathbb{R}^{d_1}, \omega_2 \in \mathbb{R}^{d_2}$将两个模态投影到相同维度的子空间,并最大化两者投影后特征间的相关系数:

$$\rho = \frac{\omega_1^\top X_1 X_2^\top \omega_2}{\sqrt{(\omega_1^\top X_1 X_1^\top \omega_1)(\omega_2^\top X_2 X_2^\top \omega_2)}}$$

因为ρ对ω_1, ω_2具有伸缩不变性,CCA 等价为:

$$\max_{\omega_1,\omega_2} \omega_1^\top X_1 X_2^\top \omega_2$$

$$s.t.\ \omega_1^\top X_1 X_1^\top \omega_1 = \omega_2^\top X_2 X_2^\top \omega_2 = 1$$

而ω_1, ω_2也可以通过求解广义特征值问题的最大特征值对应的特征向量得到:

$$X_1 X_2^\top (X_2 X_2^\top)^{-1} X_2 X_1^\top \omega_1 = \mu X_1 X_1^\top \omega_1$$

其中,μ是特征向量ω_1的特征值,ω_2也可以类似求得.文献[46]则将 CCA 扩展面向多模态的多重集典型相关分析 MCCA (Multiple CCA),并利用多核稀疏保持投影有效扩展为多模态场景.值得注意,MCCA 采用两两模态关联加和形式.考虑到神经网络强大的非线性表示能力,文献[33]提出了 DCCA (Deep CCA),如图 4 所示,DCCA 为每个模态分别建立单独的神经网络进行特征学习,再将不同模态的特征输出线性投影到共享子空间,最大化模态间的相关性,具体表示为:

$$\max_{\omega_1,\omega_2,\theta_1,\theta_2} \frac{1}{N} tr(\omega_1^\top f_1(X_1,\theta_1) f_2(X_2,\theta_2)^\top \omega_2)$$

$$s.t.\ \omega_1^\top \left(\frac{1}{N} f_1(X_1,\theta_1) f_1(X_1,\theta_1)^\top + r_1 I\right) \omega_1 = I$$

$$\omega_2^\top \left(\frac{1}{N} f_2(X_2,\theta_2) f_2(X_2,\theta_2)^\top + r_2 I\right) \omega_2 = I$$

其中,f_1, f_2表示各模态的神经网络,θ_1, θ_2是其对应网络参数.特别的,文献[33]实验发现全量数据的 L-BFGS 二阶优化效果远好于批量数据的一阶随机优化,说明优化过程中采样数据的大小与相关性计算有着密切联系.

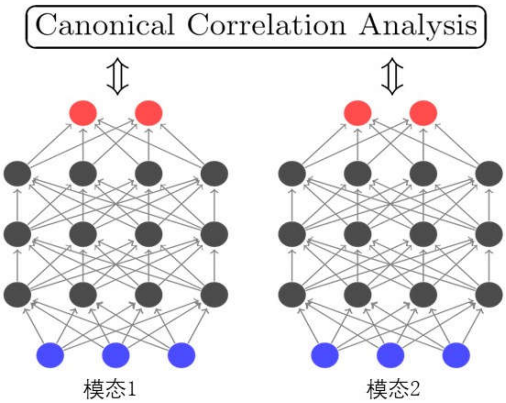


图 4. DCCA 框架[33].该方法结合 CCA 思想和深度模型框架.

进一步,DCCAE(Deep Auto-encoder CCA)[34]综合考虑了自编码网络和 DCCA 思想,相应的模型表示如下:

$$\max_{\omega_1,\omega_2,\theta_1,\theta_2,,\theta_1',\theta_2'} -\frac{1}{N} tr(\omega_1^\top f(X_1,\theta_1) f(X_2,\theta_2)^\top \omega_2) + \frac{\lambda}{N} \sum_{i=1}^{N} \sum_{v=1}^{2} (\|x_{i^v} - p_v(f_v(x_{i^v},\theta_v),\theta_v')\|^2)$$

$$s.t.\ \omega_1^\top \left(\frac{1}{N} f_1(X_1,\theta_1) f_1(X_1,\theta_1)^\top + r_1 I\right) \omega_1 = I$$

$$\omega_2^\top \left(\frac{1}{N} f_2(X_2,\theta_2) f_2(X_2,\theta_2)^\top + r_2 I\right) \omega_2 = I$$

2.3.3 MDL

文献[35]提出了基于模态隐空间表示一致的多模态深度网络 MDL(Multi-modal Deep Learning),如图 5 所示.MDL 在训练阶段利用深度网络学习不同模态在同一子空间共享的隐含表示,再重构不同模态的原始输入.图 5 左图为单模态输入重构多模态,右图为多模态输入重构多模态.值得注意,MDL 共享隐空间表示学习可以自然地扩展为两模态以上的多模态表示学习,无需像子空间表示学习方法两两加和扩展为多模态场景.

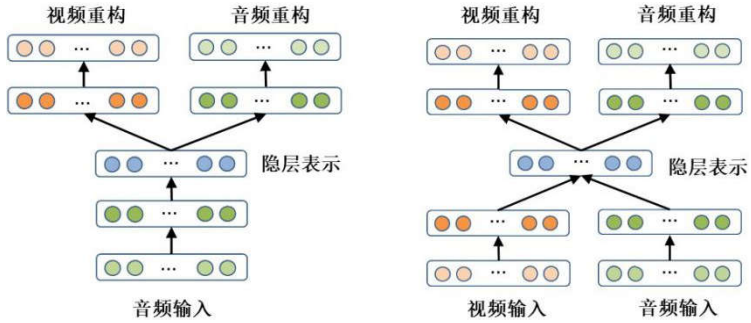


图 5. MDL 框架[35].该方法考虑深度自动编码网路进行模态隐空间表示学习.

2.4 讨论

本节介绍了基于互补性和一致性准则的传统多模态学习方法.万变不离其宗,这两类多模态学习方法都利用了模态间的强相关性:(1)标记预测的强相关性.协同训练类型方法利用潜在一致的伪标记进行互补教学,协同正则化方法利用各模态对齐无标记数据预测的一致性作为正则化项;(2)特征表示的强相关性.子空间特征约束和隐空间特征约束方法均考虑各模态数据相同维度特征表示的相关性度量,其中隐空间特征学习方法可有效扩展为多模态场景,而其他方法则需两两度量.

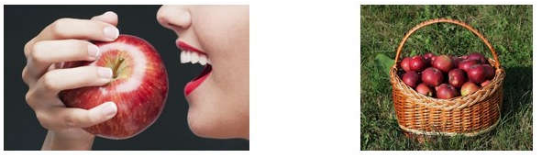
针对传统的聚类、分类等任务,多模态较之于单模态可提供更具判别性的特征表示,其思路可类比于单模态集成学习中的特征抽样、单模态半监督学习中的数据增广,从而在特征层面为样本提供更丰富的表示.基于模态间强相关性有效地利用各模态无标记数据,进而有效提升聚类、分类的集成性能.在聚类、分类任务中,互补性和一致性体现为特征的互补性和标记的一致性,二者相辅相成.另一方面,针对多模态特有的跨模态检索、描述、问答等任务,其需要构建跨模态特征嵌入间的映射关联,这类多模态学习则更注重特征表示的强相关性应用,对互补性考虑较少.

3 可靠多模态学习

在开放环境下,各模态的信息差异性较大,呈现不均衡性,其强相关性很难保证,致使传统多模态学习方法面临巨大挑战.本节将先指出不均衡多模态数据凸显的表示强弱不一致和对齐关联不一致两大挑战,而后具体介绍针对这些挑战目前有关可靠多模态学习方法的最新研究进展.

3.1 不均衡多模态数据

开放环境下,噪音、自身缺陷等因素会导致模态的不充分,进而产生模态间的差异性.如图 6 所示,图文对出现不同程度的不匹配现象.



（1）苹果是蔷薇科苹果亚科苹果属植物　　（2）iPhone是苹果公司研发及销售的智能手机系列

图 6. 表示强弱不一致的数据.图文对呈现不同程度的不匹配问题.

可见,数据的各模态所有拥有的信息呈现差异性,具有强弱之分.又如身份识别中指纹信息更丰富,而受遮挡的人脸信息较难区分;病理检测中核磁共振图像能够提供更有效的病理结构,而 X 光检测提供信息较为局限.因此,针对表示强弱不一致的多模态数据,目前研究主要分为三类:(1)模态表示不一致的异常点检测.较之于单模态异常点检测,多模态异常点检测更为复杂,拥有额外的模态不一致属性的异常点,需设计更鲁棒的多模态不一致度量.3.2.1 节和 3.2.2 节将具体介绍;(2)模态表示不一致的辅助学习.模态信息差异导致强弱之分,而强模态的收集代价通常比弱模态更加昂贵,为有效减少数据收集开销,需利用强模态在训练阶段辅助弱模态建模,进而在测试阶段仅需弱模态即可预测.3.2.3 节和 3.2.4 节将具体介绍;(3)模态表示不一致的加权融合.更一般的场景是不同样本的模态强弱也不尽相同,模态强弱存在自适应性,需自主地学习各样本不同模态的权重,进行加权融合.3.2.5 节和 3.2.6 将具体介绍.

此外,传统多模态学习中模态的对齐关联是事先给定的,样本拥有全量的多模态数据.然而考虑到深度学习通常需要大量的数据进行训练,而拥有大规模标注对齐的多模态数据十分困难.现实应用中多模态数据出现对齐关系不一致现象,如图 7 所示:(1)样本模态出现缺失问题,即仅少量样本拥有全量模态;(2)样本仅拥有非平行模态信息,即对齐关联缺失.

苹果是蔷薇科苹果亚科苹果属植物　　iPhone是苹果公司研发及销售的智能手机系列

图 7. 对齐关联不一致的数据.数据出现模态缺失或对齐关系缺失.

针对对齐关系不一致的多模态数据,目前的研究方法主要分为两类:(1)缺失多模态学习.此类方法主要考虑如何利用现有的多模态数据进行跨模态补齐,并进行后续聚类、分类操作.3.3.1 节和 3.3.2 节将具体介绍;(2)非平行多模态学习.此类方法主要考虑如何利用潜在一致的标记信息建立模态间隐含关联,进行辅助学习、跨模态映射.3.3.3 节和 3.3.4 节将具体介绍.

3.2 针对表示强弱不一致的方法

3.2.1 MVAD

文献[21]提出概率隐变量模型 MVAD(Multi-view Anomaly Detection)来检测模态不一致的异常点.MVAD 假设所有一致的样本是由单个隐向量生成,而异常点则由不同隐向量生成.通过狄利克雷过程先验(Dirichlet process priors)可以推断每个样本隐向量的个数,进而获得每个样本异常的概率.如图 9 所示,对于多模态样本X的生成过程如下:

Step1. 刻画参数$\alpha \sim Gamma(a,b)$;

Step2. 对每个样本$n = 1,2,\cdots,N$

(a) 刻画混合权重$\theta_n \sim Stick(\gamma)$;

(b) 对每个隐向量:$j = 1,2,\cdots,\infty$:刻画一个隐向量$z_{nj} \sim N(0,(\alpha r)^{-1}I)$

(c) 对每个视图:$d = 1,2,\cdots,D$

刻画一个隐向量分配$s_{nd} \sim Discrete(\theta_n)$

刻画一个观测向量$x_{nd} \sim N(W_d z_{ns_{nd}}, \alpha^{-1}I)$

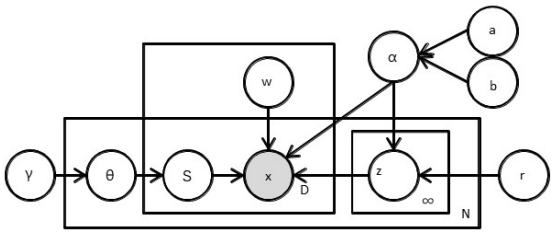


图 8. MVAD 框架[21].该方法利用概率隐变量模型检测模态不一致异常点.

其中Stick(γ)是折棍子过程(stick-breaking)[36],可以利用参数γ为狄利克雷过程生成混合权重,r是对隐向量表示的关联预测.α共享于观测值和隐向量预测.图 8 中阴影部分和未阴影部分分别表示观测值和隐变量.整体的框架可以看成鲁棒概率典型性相关分析对模态不一致异常点检测的扩展,可以运用随机 EM 算法进行贝叶斯推断.

3.2.2 DRUMN

文献[37]基于迭代训练错误率提出一种鲁棒无监督多模态深度网络 DRUMN(Deep robust unsupervised multi-modal network).传统基于模态权重检测多模态异常点的方法存在两个弊端:(1)检测阈值需预先设定且固定不变,不能随学习过程自适应调节;(2)考虑模态两两配对检测,阈值随模态个数的增多而指数增长.为解决上述问题,DRUMN 考虑自适应的为各模态样本及模态对加权.其首先采用能量模型 RBM(Restricted Boltzmann Machine)[38]作为特征学习网络.具体表示为:

$$r\left(f\left(x;\theta_f\right);\theta_r\right) = x - \nabla_x E(x;\theta)$$

$$E_k\left(x_{i^k};\theta\right) = \frac{1}{2}\left\|x_{i^k} - b'\right\|_2^2 - \sum_{j=1}^{N_l} h_{l,j}$$

$$s.t.\ h_l = g\left(f\left(h_{l-1}\right)\right), l \in \{1,\cdots,L\}$$

其中$r\left(f\left(x;\theta_f\right);\theta_r\right)$是自编码网络,$E(x;\theta)$是能量模型,$N_l$是 l 层的维度,$g(\cdot)$是池化层,$f(\cdot)$是对不同网络结构特定的算子,如卷积算子,全连接算子等.$\left\|x_{i^k} - b'\right\|_2^2$相当于先验,使得输入远离$b'$.能量模型也被验证在单模态上能够有效检测异常点[39].DRUMN 利用每个样本在迭代过程中的能量序列计算能量方差为每个模态样本自适应加权:

$$\omega_{i^k} = std_{i^k}^{conf}(H)$$

$$std_{i^k}^{conf}(H) = \sqrt{var\left(E_{H_{i^k}^{t-1}}(x_{i^k};\theta)\right) + \frac{var\left(E_{\left(H_{i^k}^{t-1}\right)}(x_{i^k};\theta)\right)^2}{\left(\left|H_{i^k}^{t-1}\right| - 1\right)}}$$

其中,$H_{i^k}^{t-1}$是第 i 个样本第 k 个模态的能量序列,$var(\cdot)$是估计的能量方差,ω_{i^k}等价于第 i 个样本第 k 个模态的权重,异常样本ω_{i^k}较小.同一个样本两个模态的关系权重可以定义为:

$$\gamma_{i^{m,n}} = \sqrt{var\left(C_{H^{t-1}_{(i^m,i^n)}}(x_{i^m},x_{i^n})\right) + \frac{var\left(C_{H^{t-1}_{(i^m,i^n)}}(x_{i^m},x_{i^n})\right)^2}{\left|H^{t-1}_{(i^m,i^n)}\right|-1}}$$

其中,$C(\cdot)$表示互信息函数,且模态不一致样本$\gamma_{i^{m,n}}$较大.最终的优化函数表示为:

$$\min_{\theta_{f_k},\theta_{r_k},U_k} \sum_{m\neq n}^{K}\sum_{i=1}^{N_C} -\frac{\gamma_{i^{m,n}}}{N_c} tr(U_m^{\top} f_m(x_{i^m}) f_n(x_{i^n})^{\top} U_n) + \sum_{k=1}^{K}\sum_{i=1}^{N_C+N_k} \omega_{i^k} \left\| x_{i^k} - r_k\left(f_k(x_{i^k})\right)\right\|_F^2$$

$$s.t.\ U_m^{\top}\left(\frac{1}{N} f_m(x_{i^m}) f_m(x_{i^m})^{\top} + r_1 I\right) U_m = I$$

总体上,DRUMN 利用各模态的自编码(auto-encoder)网络结构处理模态缺失样本,同时用能量模型自适应地估计样本权重处理模态不一致的样本,进而减小多模态异常点对训练带来的干扰.

3.2.3 ICo-training

针对强弱模态辅助学习,文献[16]证明模态不充分条件下,Co-training 适用的理论分析:两个模态预测置信度的差异性较大,Co-training 在模态信息不充分的条件下仍然能够通过利用无标记数据提升学习器性能,并提出一种基于大间隔算法 ICo-training:

Step1. 无放回的从无标记数据U构造大小为u的数据池U';

Step2. 分别运用两个模态X_1, X_2的有标记数据训练两个学习器h_1, h_2;

Step3. 每个模态用训练好的学习器在U'中本模态无标记样本中挑选p个最置信的正例和n个最置信的负例,挑选最置信的样本需要预测概率大于设定的阈值;

Step4. 标上伪标记加到L中重训练.

不难发现,随着学习器性能的变化,设定的阈值也应变化.为此,文献[16]进一步提出了基于迭代间隔的 ICo-training 算法,迭代的阈值表示为:

$$\gamma_v^{i+1} = y_v^0 - C_v^{\mathcal{L}}(R_v - \eta_v)\left(1 - \frac{n\sqrt{m_0}}{(m_i + |T_i|)^{3/2}}\right)$$

其中,y_v^0为初始值,$C_v^{\mathcal{L}}, \eta_v$为常数,$R_v = max_{h_v} R(h_v)$,$n$ 为有标记和无标记样本的总数,m_0为初始有标记的数量,m_i为当前迭代有标记样本数量,$|T_i|$为挑选的样本数量.

3.2.4 ARM

但上述方法仍需手动设定阈值参数挑选样本.为此,文献[20]提出了 ARM(Auxiliary regularized machine)方法,旨在训练阶段利用强模态学习器辅助弱模态进行有效的特征抽取.ARM 利用先验知识,将模态分为强模态和弱模态两个模态并分别建立学习器,同时利用强模态的预测和弱模态的邻接矩阵构造流形正则项,起到强模态辅助弱模态的作用.ARM 模型表示如下:

$$\min_{F_{v_2},\omega} \|F_{v_2}\|_2^2 + \lambda_1 tr(F_{v_2} L_{v_1} F_{v_2}) + \lambda_2 \left\| {X_{v_1}}^{\top}\omega - (YY^{\top})^{-\frac{1}{2}}Y\right\|_F^2$$

$$s.t.\ \ y_i f_{v_2}(x_{i^{v_2}}) \geq 1, \qquad \forall i \in \{1,2,\cdots,N_l\}$$

其中,f_{v_2}是强模态学习器,X_{v_1}是弱模态特征矩阵,ω是弱模态的特征映射矩阵.在约束项中对强模态进行硬间隔约束,强制要求强模态分类正确.弱模态只采用了平方损失函数(可以替换成其他凸函数).同时,第二项是强模态和弱模态构成的流形正则化项,L_{v_1}是弱模态特征映射后构造的拉普拉斯矩阵,该项目的在于使得弱模态尽可能与强模态一致,起到强模态辅助弱模态的作用.值得注意,ICo-training 和 ARM 忽略了交叉迭代训练过程中弱模态可能给强模态带来的噪声问题,影响强模态的模型训练.

3.2.5 RMVC

模态不充分场景下,传统多模态聚类会产生性能退化现象.为此,文献[40]提出了可靠多模态聚类方法RMVC(Reliable Multi-view Clustering),自适应地为不同候选聚类结果学习相应的权重,并最大化最优单模态在最坏聚类设定下的信息增益,以此提高多模态集成聚类的性能.该方法先提出χ^2距离,度量不同聚类指示矩阵$Y_1 \in \mathbb{R}^{n\times K_1}, Y_2 \in \mathbb{R}^{n\times K_2}(K_1, K_2$可不相等)的差异:

$$d^2_{\chi^2}(Y_1, Y_2) = \|Y_1 Y_1^\top - Y_2 Y_2^\top\|_F^2$$

$d^2_{\chi^2}$是二次函数,且可证等价于错分距离(Misclassification Error distance $\mathrm{d_{ME}}$),d_{ME}对应聚类准确度.可见 $d^2_{\chi^2}$和聚类效果直接关联.基于$d^2_{\chi^2}$,RMVC 表示为:

$$\max_{Y\in\mathcal{Y}} \min_{\alpha\in M} \sum_{i=1}^{m} \alpha_i (d^2_{\chi^2}(Y_0, Y_i) - d^2_{\chi^2}(Y, Y_i))$$

其中,α服从单纯型$M = \{\alpha | 1^\top \alpha = 1; \alpha \geq 0\}$,$Y$为待优化的潜在聚类结果.$d^2_{\chi^2}(Y_0, Y_i) = min_{1\leq v\leq V} d^2_{\chi^2}(Y_0^v, Y_i)$,$Y_0^v$是预先获得的单模态聚类结果,$Y_i$是运行 m 个多模态聚类算法获得的 m 个聚类结果.Y_0等价于所有单模态聚类结果中最优的聚类结果.分开看,$d^2_{\chi^2}(Y_0, Y_i)$这一项可确定每种多模态聚类效果的权重α_i.而最大化$-d^2_{\chi^2}(Y, Y_i)$相当于对 m 个多模态聚类的集成学习,可看出最终的聚类结果和Y_i密切相关,文献[40]证明了如下结论:如最优聚类结果属于Y_i,那么优化得到的聚类结果肯定优于单模态的聚类结果.

3.2.6 CMML

针对分类任务,文献[41]提出了半监督多模态学习方法 CMML(Comprehensive Multi-modal learning),其利用注意力机制自适应地为每个样本的不同模态学习相应的权重,并提出差异性度量和鲁棒一致性度量来体现模态间的互补性并进行自适应加权融合.充分性度量表示为:

$$L_s = \sum_{i=1}^{N_l} \sum_{j=1}^{M} \ell(\alpha_{i,j} f_j(x_i), y_i)$$

其中,$f_j(\cdot)$是每个模态的学习器,这里表示为深度网络,$\alpha_{i,j} = \frac{h\left(f_j\left(x_i^j\right)\right)}{\sum_{m=1}^{M} h\left(f_m\left(x_i^m\right)\right)}$表示第 i 个样本的第 j 个模态的权重,$h(\cdot)$是额外的注意力神经网络,如两层浅层全连接网络.

差异性度量可以表示为:

$$Com(F) = \frac{1}{\sum_{1\leq i\neq j\leq M} 1} \sum_{1\leq i\neq j\leq M} sim(f_i, f_j)$$

$$sim(f_i, f_j) = \frac{1}{N} \sum_{k=1}^{N} cos(f_i(x_k^i), f_j(x_k^j))$$

$F = \{f_m\}_{m=1}^M$是所有模态学习器,cos是 cosine 函数,$Com(F)$可以凸显模态的互补性.鲁棒一致性度量可以表示为:

$$R_\delta(F) = \frac{1}{\sum_{1\leq i\neq j\leq M} 1} \sum_{1\leq i\neq j\leq M} H_\delta(f_i, f_j)$$

$$H_\delta(f_i, f_j) = \begin{cases} \frac{1}{2}\left(2 - cos(f_i, f_j)\right)^2, & |2 - cos(f_i, f_j)| \leq \delta \\ \delta|2 - cos(f_i, f_j)| - \frac{1}{2}\delta^2, & otherwise \end{cases}$$

$H_\delta(f_i, f_j)$使用改进的 Huber loss[44]替换原始的平方损失函数,从而缓解模态不一致样本所带来的影响.模型的优化函数构建如下:

$$\min_{f_j} \sum_{i=1}^{N_l} \sum_{j=1}^{M} \ell(\alpha_{i,j} f_j(x_i), y_i) + \|f_j\|_F^2 + Com(F) + R_\delta(F)$$

该方法借用图像、文本领域常用的注意力机制,自适应地为每个模态学习相应的权重进行加权融合,从而有效缓解模态不均衡带来的弱相关问题.

3.3 针对对齐关联不一致的方法

3.3.1 PVC

在模态缺失情况下如直接应用现有的多模态方法,须丢弃模态缺失的样本或先补全缺失模态特征,这会丢失有效信息或引入额外噪声.为此,文献[17]提出了 PVC(Partial view clustering)方法对模态缺失样本进行聚类.不同于传统多模态方法优化投影矩阵将不同模态投影到同维度子空间表示,PVC 基于字典学习将子空间表示也作为优化变量投影回各模态的原始表示空间,再利用优化得到的子空间表示进行聚类:

$$\min_{\{U^v, \bar{P}^v\}} \sum_{v=1}^{2} \left\| \begin{bmatrix} X_c^v \\ \hat{X}^v \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^v \end{bmatrix} U^v \right\| + \lambda \|\bar{P}^v\|_1$$

$$s.t.\ U^v \geq 0,\ \bar{P}^v \geq 0$$

其中,X_c^v表示无模态缺失的样本表示,$\hat{X}^v$表示模态缺失的样本表示,$\bar{P}^v = [P_c; \hat{P}^v]$是两个模态子空间上的表示形式.值得注意,样本可表示成非负矩阵分解形式,$\bar{P}^v$是子空间表示(聚类指示矩阵),U^v是样本在该模态子空间基向量(字典模型),$\|\bar{P}^v\|_1$表示其稀疏性.因此所有数据可以表示为:$P = [P_c; \hat{P}^1; \hat{P}^2]$,利用 P 可以直接聚类.在 PVC 中,不需要$\hat{P}^1; \hat{P}^2$一致,仅共享无模态缺失的样本表示P_c.值得注意,此方法本质上对模态缺失样本并无实际操作,仅约束无缺失的模态子空间表示投影后与原空间表示的一致性.

3.3.2 SLIM

考虑利用对齐的无缺失模态样本信息辅助缺失模态进行学习,文献[42]提出半监督多模态学习方法 SLIM(Semi-supervised learning with incomplete modalities).SLIM 有效地利用数据预测的潜在一致性,利用预测概率补全各模态的相似性矩阵,从而在统一的框架中为每个模态学习单独的学习器和所有未标记样本的聚类学习器,进而可以同时进行分类和聚类任务:

$$\min_{W_k, b_k, F} \sum_{k=1}^{K} \frac{1}{2\eta_k} ||\hat{X}_k W_k + 1 b_k^\top \odot P_k - F \odot P_k||_F^2 + \frac{\lambda_1}{2} ||W_k||_F^2 + \frac{\lambda_2}{2\eta_k} \left|\left| R_\Omega(M_k) - R_\Omega(YY^\top) \right|\right|_F$$

其中,$W_k \in R^{d_k \times C}$是线性学习器,$b_k \in R^C$是当前预测的偏差,1 是一个全 1 向量,$\odot$表示对应元素的点乘算子,$P_k \in R^{N \times C}$ 是指示矩阵,其中$[P_k]_{I,\cdot} = 1$ 表示第 i 个示例的第 k 个模态上完整,否则$[P_k]_{i,\cdot} = 0$.在多类情况下,x_i的标签y_i扩展为一个 C 维的向量,其中 $y_{i,j} = 1$表示第 i 个示例为第 j 个标签,否则,$y_{i,j} = 0$;类似地,$F \in R^{N \times C}$表示所有示例的预测标记,η_k是第 k 个模态的完整样本的个数.$M_k \in R^{N \times N}$ 是第 k 个模态的相似度矩阵.$[R_\Omega(M_k)]_{i,j} = [M_k]_{i,j}$表示第 i 个样本和第 j 个样本的第 k 个模态完整,否则为 0.其中第三项进一步采用平方根损失函数代替方程中的最小二乘函数,减少噪音数据的影响.换言之,此项等价于一个加权正则化的最小二乘形式,其中每个模态的权重为:$\frac{1}{\eta_k ||R_\Omega(M_k) - R_\Omega(YY^\top)||_F}$,进而可以通过考虑所有模态的不同噪声水平来校准每个模态.最终,SLIM 利用模态的一致性来补全各模态缺失的相似性矩阵,从而获得潜在一致的预测矩阵 F.

3.3.3 DeVise

针对模态对齐关联缺失问题,文献[18]提出一种启发式辅助学习方法 DeVise(deep visual-semantic embedding model).具体地,DeVise 在训练图片模型时随机抽样文本模态的异类样本构造三元组损失函数辅助图片深度网络训练,利用文本基模型获得的特征嵌入辅助图片缩小类内距离,扩大类间距离.最终可利用文本模态样本增广训练数据,从而减少图片训练样本的数量.具体公式如下:

$$\ell(image, label) = \sum_{j \neq label} max\,[0, margin - e_{label} M e_{image} + e_{text} M e_{image}]$$

其中,margin 是人为定义的距离参数,e_{label}是标记的语义表示,e_{image}是图片的特征嵌入表示,M 是映射矩阵,e_{text}是文本模态的特征表示.值得注意,该方法无需模态间的对齐关联,仅利用标记一致性进行样本挑选,适用于分类等任务,而针对面向模态样本对齐的跨模态检索等任务则效果甚微.

3.3.4 SCML

针对模态对齐关联缺失下的跨模态检索问题,文献[43]提出 SCML(Sequential cross-modal learning),该方法基于共享预测模型的序列化训练方式进行多模态模型联合训练,进而利用共享模型挖掘跨模态潜在一致的特征表示.

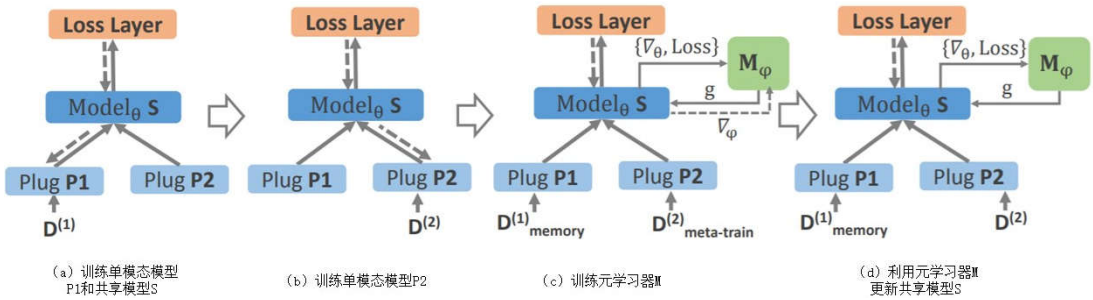


图 9. SCML 框架[43].该方法基于共享预测模型进行序列化训练,通过保证共享模型性能不下降获得模态间潜在一致的特征嵌入.

如图 9 所示,SCML 首先训练单模态模型 P1(S)和共享模型 S,再固定共享模型 S 训练单模态模型 P2,此步固定 S 旨在防止 S 对 P1 学到知识的遗忘.而后仅利用少量的 P1 和 P2 数据训练元学习器 M,此步是为了利用元学习器更新共享 S,进一步获得潜在一致的语义表示.值得注意,SCML 训练共享模型使得各模态预测性能不下降这一思路,以此获得跨模态潜在一致的映射关联,但这并不是样本级别的映射关联,因此该方法在 NDCG 指标中性能较好,而 Rank 指标中性能较差.

3.4 讨论

本节主要介绍了针对不均衡多模态数据所提出的可靠多模态学习方法.考虑模态表示强弱不一致的方法主要思考如何有效度量模态的不一致性,并考虑利用性能优异的模态进行辅助学习.而考虑模态对齐关联不一致的方法主要考虑如何缓解模态缺失的影响,补齐模态缺失数据.而面向关联缺失的方法主要思考如何学习并利用模态间潜在一致的关联性,如标记关联.但目前仍有诸多挑战有待解决:(1)模态不充分性度量.目前强弱模态是靠训练数据的性能或者先验知识决定,且绝大多数方法局限于两模态.如何更有效地界定模态的不充分性,并度量更细粒度的样本级别的模态不充分性有待研究;(2)模态缺失数据处理.目前对于模态缺失问题实质上对样本缺失模态仅作为单模态处理,如何利用样本无缺失的模态对缺失的模态进行有效操作有待研究;(3)非平行多模态学习.目前针对模态关联缺失的方法大多为启发式方法,如何有效扩展为仅利用少量对齐数据进行对齐标签传播有待研究.

4 结束语

多模态学习近些年受到广泛关注并拥有诸多实际应用.传统多模态学习方法面向真实不均衡多模态数据会出现性能退化甚至低于单模态性能,这通常归结于模态表示强弱的不一致和模态对齐关联的不一致问题.为此可靠多模态学习被提出,针对上述两个挑战的可靠多模态学习体现较之于传统多模态学习具有更优异的性能.在未来研究者,本文认为还存在如下几方面挑战:(1)针对表示不一致的可解释性研究.目前的方法大多局限

于基于各模态最终的特征嵌入进行不一致的度量及后续处理,缺乏考虑导致模态间不一致的因素,如局部区域信息的不一致性.如何利用多示例学习细粒度刻画各模态样本,并结合诸如图模型解释模态不一致具有巨大的研究前景和广阔的应用价值;(2)针对关联不一致的隐关联学习.目前的方法大多还是启发式方法,在模态对齐映射学习过程中可能引入额外的噪声,如何利用少量的对齐模态数据初始化模态间的映射函数,并利用非平行数据结合对偶学习或循环生成网络进一步训练值得研究;(3)动态环境下的多模态学习.当前多模态学习大多是静态的,即给定训练集训练模型并在测试集验证,而现实环境是动态变化的,流式数据具有分布变化、特征增广、新类检测等问题,如何将现有的多模态学习扩展到动态环境下值得研究.

5 参考文献:

[1]A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In COLT, pages 92–100, 1998.

[2]K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In CIKM, pages 86–93, 2000.

[3]S. Yu, B. Krishnapuram, R. Rosales, and R.B. Rao. Bayesian co-training. In NeurIPS, pages 1–7, 2007.

[4]M.L. Zhang, Z.H. Zhou: CoTrade: Confident Co-Training With Data Editing. IEEE TMC, Part B41(6): 1612-1626, 2011.

[5]V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In ICML Workshop, 2005.

[6]J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: Svm-2k, theory and practice. In NeurIPS, pages 355-362, 2005.

[7]T. Xia, D. Tao, T. Mei, and Y. Zhang. Multiview spectral embedding. IEEE TSMC, 40(6):1438–1446, 2010.

[8]H. Hotelling. Relations between two sets of variates. Biometrika, 28(3/4):321–377, 1936.

[9]M. Gonen and E. Alpaydın. Multiple kernel learning algorithms. JMLR, 12:2211–2268, 2011.

[10]X. Wang, X. Guo, Z. Lei, C. Zhang, S. Z. Li. Exclusivity-Consistency Regularized Multi-view Subspace Clustering. In CVPR, pages 1-9, 2017.

[11]W. Wang, Z.H. Zhou. A New Analysis of Co-Training. In ICML, pages 1135-1142, 2010.

[12]K. Sridharan, S.M. Kakade. An Information Theoretic Framework for Multi-view Learning. In COLT, pages 403-414, 2008.

[13]T.G. Dietterichl. Ensemble learning. 2002.

[14]B. Wei and C. Pal. Cross lingual adaptation: An experiment on sentiment classifications. In ACL, pages 258–262, 2010.

[15]I. Muslea, S. Minton, and C.A. Knoblock. Selective sampling with naive co-testing:Preliminary Results. In CRM Workshop, 2000.

[16]W. Wang and Z.H. Zhou: Co-Training with Insufficient Views. In ACML, pages 467-482, 2013.

[17]S. Li, Y. Jiang, and Z. Zhou, Partial multi-view clustering. In AAAI, pages 1968–1974, 2014.

[18]A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In NeurIPS, pages 2121–2129, 2013.

[19]I. Mulsea, S. Minton and C.A. Knoblock. Active Learning with Strong and Weak Views: A Case Study on Wrapper Induction. In IJCAI, pages 415-420, 2003.

[20]Y. Yang, H.J. Ye, D.C. Zhan, and Y. Jiang. Auxiliary Information Regularized Machine for Multiple Modality Feature Learning. In IJCAI, pages 1033–1039, 2015.

[21]T. Iwata and M. Yamada. Multi-view Anomaly Detection via Robust Probabilistic Latent Variable Models. In NeurIPS, pages 1136-1144, 2016.

[22]H.D. Zhao, Y. Fu: Dual-Regularized Multi-View Outlier Detection. In IJCAI, pages 4077-4083, 2015.

[23]Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Zhi-Bin Liu, Yuan Jiang. Complex Object Classification: A Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport. In KDD, page 2594–2603, 2018.

[24]T. Baltrusaitis, C. Ahuja, L.P. Morency, Multimodal Machine Learning: A Survey and Taxonomy. PAMI, 41(2): 423-443, 2019.

[25]D. Ramachandram, G.W. Taylor. Deep Multimodal Learning: A Survey on Recent Advances and Trends. In SPM, pages 96-108, 2017.

[26]Shiliang Sun, A survey of multi-view machine learning. Neural Computing and Applications, 23(7-8): 2031-2038, 2013.

[27]M.F. Balcan, A. Blum, and Y. Ke. Co-training and expansion: Towards bridging theory and practice. In NeurIPS, pages 89-96, 2004.

[28]W. Wang and Z.H. Zhou. Analyzing co-training style algorithms. In ECML, pages 454–465, 2007.

[29]S. Dasgupta, M.L. Littman, and D. McAllester. Pac generalization bounds for co-training. In NeurIPS, pages 375–382, 2002.

[30]D.D. Chen, W. Wang, W. Gao, and Z.H. Zhou. Tri-net for semi-supervised deep learning. In IJCAI, pages 2014–2020, 2018.

[31]Z.H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. IEEE TKDE, 17(11):1529–1541, 2005.

[32]V. Sindhwani and D.S. Rosenberg. An rkhs for multi-view learning and manifold coregularization. In ICML, pages 976–983, 2008.

[33]G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, Deep canonical correlation analysis, In ICML, pages 1247–1255, 2013.

[34]W. Wang, R. Arora, K. Livescu, Jeff A. Bilmes: On Deep Multi-View Representation Learning. In ICML, pages 1083-1092, 2015.

[35]J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A.Y. Ng. Multimodal deep learning. In ICML, pages 689–696, 2010.

[36]J. Sethuraman. A constructive definition of dirichlet priors. Statistica sinica, pages 639–650, 1994.

[37]Y. Yang, Y.F. Wu, D.C. Zhan, and Y. Jiang. Deep robust unsupervised multi-modal network. In AAAI, pages 5652-5659, 2019.

[38]P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. JMLR, 11:3371–3408, 2010.

[39]S. Zhai, Y. Cheng, W. Lu, and Z. Zhang. Deep structured energy based models for anomaly detection. In ICML, pages 1100–1109, 2016.

[40]H. Tao, C. Hou, X. Liu, T. Liu, D. Yi and J. Zhu: Reliable Multi-View Clustering. In AAAI, pages 4123-4130, 2018.

[41]Y. Yang, K.T Wang, D.C zhan, H. Xiong and Y. Jiang. Comprehensive Semi-Supervised Multi-Modal Learning. In IJCAI, pages 4092-4098, 2019.

[42]Y. Yang, D. Zhan, X. Sheng, and Y. Jiang, Semi-supervised multi-modal learning with incomplete modalities. In IJCAI, pages 2998–3004, 2018.

[43]G. Song, X.Y Tan, Sequential Learning for Cross-Modal Retrieval, In CVPR Workshop, pages 4531-4539, 2019.

[44]P.J. Huber. Robust estimation of a location parameter. AMS, 35(1):73–101, 1964.

附中文参考文献:

[45]王魏, 周志华. 多视图在利用未标记数据学习中的效用.见:张长水,杨强, 机器学习及其应用 2013,北京: 清华大学出版社, 2013, 27-45.

[46]张荣. 基于稀疏表示的多重集典型相关分析算法研究.南京理工大学, 2015.