第 40 卷 第 6 期
2017 年 6 月
计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS
Vol. 40 No. 6
June 2017

# 跨媒体语义共享子空间学习研究进展

张 磊[1) ] 赵 耀[2),3)] 朱振峰[2),3)]

[1)](中国科学院信息工程研究所 北京 100093)

[2)](北京交通大学信息科学研究所 北京 100044)

[3)](北京交通大学现代信息科学与网络技术北京市重点实验室 北京 100044)

**摘 要** 随着信息技术的飞速发展,现实世界中涌现出大量的跨媒体数据.所谓跨媒体数据是指那些表达的内容相似,但以不同模态、不同来源、不同背景等形式出现的数据.比如,一张描述花豹的网页通常采用共生的图片和文本等不同的模态刻画花豹的外形和习性.这些跨媒体数据通常呈现出底层特征异构、高层语义相关的特性.传统的单媒体学习方法已无法适应跨媒体数据呈现出的特征异构性.因此,跨媒体学习相关理论与方法的研究是当前数字媒体分析领域的热点研究课题之一.该文主要介绍了跨媒体学习的研究背景和应用价值,概括介绍了各类跨媒体学习相关方法的数学原理和基本特性,并重点介绍了跨媒体共享子空间学习的研究进展,比较了基于投影、矩阵分解、任务和度量等四大类子空间学习方法的优缺点,分析了未来的发展方向.

**关键词** 跨媒体;异构数据;共享子空间;多视角学习;优化;人工智能

**中图法分类号** TP18 **DOI 号** 10.11897/SP.J.1016.2017.01394

# Advances in Semantically Shared Subspace Learning for Cross-Media Data

ZHANG Lei[1)] ZHAO Yao[2),3)] ZHU Zhen-Feng[2),3)]

[1)](*Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093)

[2)](*Institute of Information Science, Beijing Jiaotong University, Beijing* 100044)

[3)](*Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing* 100044)

**Abstract** With the rapid development of information technology, there are many CMD (Cross-Media Data) in the real world. The so-called cross-media data refer to information items with similar underlying contents, which arrive in different modalities, sources or backgrounds, and so on. For example, a webpage describing leopards uses co-occurring text and image of different modalities to represent leopards. These CMD show characteristics of the heterogeneity of low-level features and the correlation of high-level semantics. Traditional Mono-Media Learning (MML) methods have not been able to adapt to the feature heterogeneity of CMD. Thus, the research on related problems in CML (Cross-Media Learning) have been One of the hot research topics in the field of digital media analysis recently. This paper mainly introduces the research background and application value of CML, and provides an overview of mathematical principle and specialties of various related methods in CML. Meanwhile, the advances in cross-media shared subspace learning are presented. Furthermore, the paper compares the advantages and

收稿日期:2015-10-19;在线出版日期:2016-05-05.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316401)、国家自然科学基金(61532005,61572068,61271275,61501457,61601458,61602465)、长江学者和创新研究团队项目(IRT201206)和新世纪优秀人才支持计划(13-0661)资助.**张 磊**,男,1982 年生,博士,助理研究员,中国计算机学会(CCF)会员,主要研究方向为数据挖掘、机器学习.E-mail:zhanglei1@iie.ac.cn.**赵 耀**,男,1967 年生,博士,教授,主要研究领域为图像视频编码、数字水印与取证、视频分析与理解.**朱振峰**,男,1974 年生,博士,教授,主要研究领域为图像视频理解、计算机视觉、机器学习.

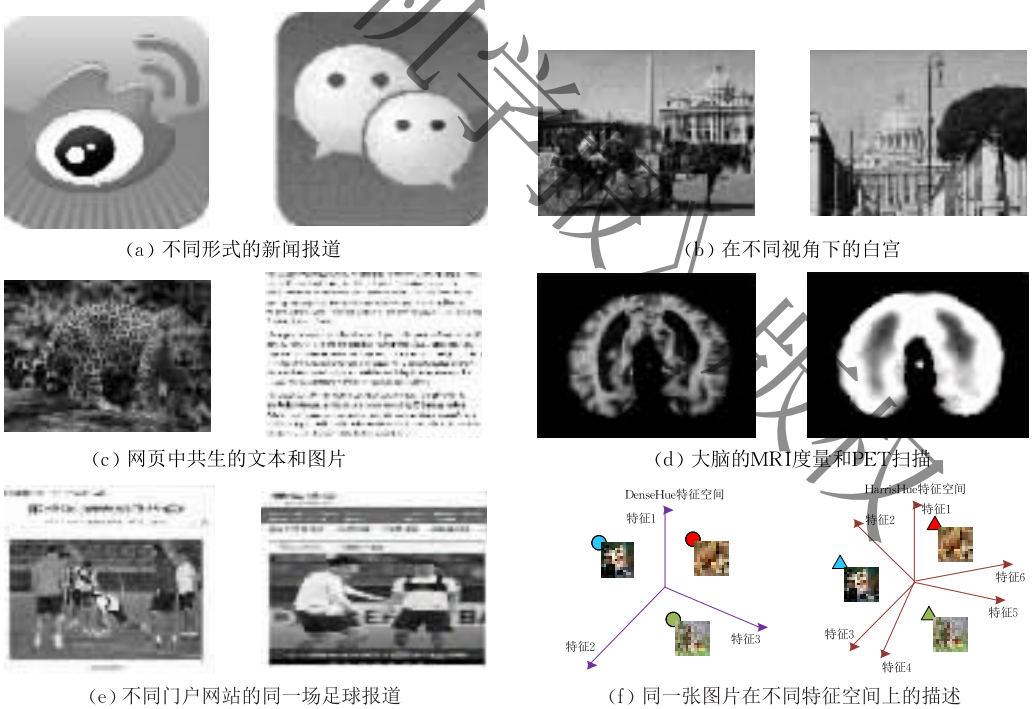
《计算机学报》版权

disadvantages of four kinds of subspace learning methods based on projection, matrix decomposition, task and measurement, respectively. Finally, the future development of CML is analyzed.

**Keywords**　cross-media; heterogeneous data; shared subspace; multi-view learning; optimization; artificiall intelligence

## 1　引　言

近年来，随着大量高技术数码产品的出现，由这些电子设备产生的跨媒体数据(Cross-Media Data, CMD)[1-4]已经遍布到人们现实生活的各个角落. 所谓跨媒体数据是指那些表达的内容语义相似，但以不同模态、不同来源、不同背景等形式出现的数据. 如图 1 所示，新浪微博和腾讯微信关于相同的新闻的不同形式的报道；相同的建筑物白宫可以处在不同的视角之下；同一张描述花豹的网页采用共生的图片和文本等不同的模态来刻画花豹；而人的大脑可以有 MRI 和 PET 两种不同测量的医学成像[4]；同一场足球报道可以来自新浪和搜狐两个不同的门户网站；同一张老虎的图片可以在 DenseHue 和 HarrisHue 特征空间[2]上获得两种异构的描述，等等. 跨媒体数据通常呈现出底层特征异构、高层语义相关的特性. 如何充分利用不同模态、不同来源、不同空间的跨媒体数据，挖掘其一致性和互补信息以充分理解媒体数据的内容是当今媒体数据分析的热点和难点. 传统的单媒体学习(Mono-Media Learning, MML)[5-6]方法已无法适应跨媒体数据的特征异构性. 不同于单媒体学习，跨媒体学习(Cross-Media Learning, CML)[7-10]在学习过程中能够度量不同模态的媒体对象之间的相关性，以实现不同模态间的灵活跨越.




(a) 不同形式的新闻报道　(b) 在不同视角下的白宫

(c) 网页中共生的文本和图片　(d) 大脑的MRI度量和PET扫描

(e) 不同门户网站的同一场足球报道　(f) 同一张图片在不同特征空间上的描述

图 1　典型的跨媒体数据

### 1.1　跨媒体的基本概念

现实世界中广泛存在着大量的跨媒体数据. 目前，跨媒体数据尚没有统一的定义，一般说来，跨媒体数据是指那些表达的内容语义相似，但以不同模态、不同来源、不同背景等形式出现的数据. 跨媒体数据通常具有跨模态、跨源、跨空间等特点.

跨模态(Cross Modality)是指跨媒体数据通过图像、视频、文本等不同的模态共同表达某个概念或事件. 例如，在跨媒体检索中[9-13]，存在于同一网页(也称为多媒体文档(Multimedia Document, MMD)[3-14])中的共生的文本和图片携带了相似的语义信息.

跨源(Cross Source)是指跨媒体数据的来源渠道不同，但表达的语义却相似. 例如，同一事件的新闻报道可以来自新浪和搜狐两个不同的门户网站.

跨空间(Cross Space)指的是跨媒体数据借助交互行为来分析涌现于信息空间和物理世界中的关联数据,揭示在人类信息交互作用下的信息空间、物理世界和人类社会之间耦合映射和相互影响的机理,实现信息空间和物理空间的跨越.例如,网络因与现实社会间的相互融合和映照,使其成为人类生产和生活中不可或缺的一部分.如今互联网数据与现实世界人们行为之间的相互作用和影响愈发明显.如在2010年爆发于突尼斯和埃及等中东国家的一系列动荡中,网络跨媒体数据和现实世界人们行为之间存在着一种增强映射关系:即人们通过在Twitter发布消息、在Facebook张贴相关地图和照片、在Youtube添加相关视频等方式,来表达自己对现实世界中所发生事情的观点,而这些网络媒体数据又进一步影响了现实社会局势,并愈演愈烈,最终导致了国家政权的更迭.

近年来,跨媒体技术已被应用于跨媒体检索(Cross-Media Retrieval)[3,15-17]、跨媒体聚类(Cross-Media Clustering)[18-21]和跨媒体分类(Cross-Media Classification)[22-25]等场合.

跨媒体检索[3,15-17]指的是可以实现不同模态之间灵活跨越的多媒体检索方式,即通过一种模态的实例检索出与之相关的其它模态的样本.如图2所示,在绝大多数情况下,人们通过歌曲的曲调只能搜索到mp3歌曲而无法获得与之相关的歌词.如果可以使用图片检索出与之对应的文本和语音,或者使用文本检索出与之对应的图片和语音,就可以使人们获取更多的媒体内容.

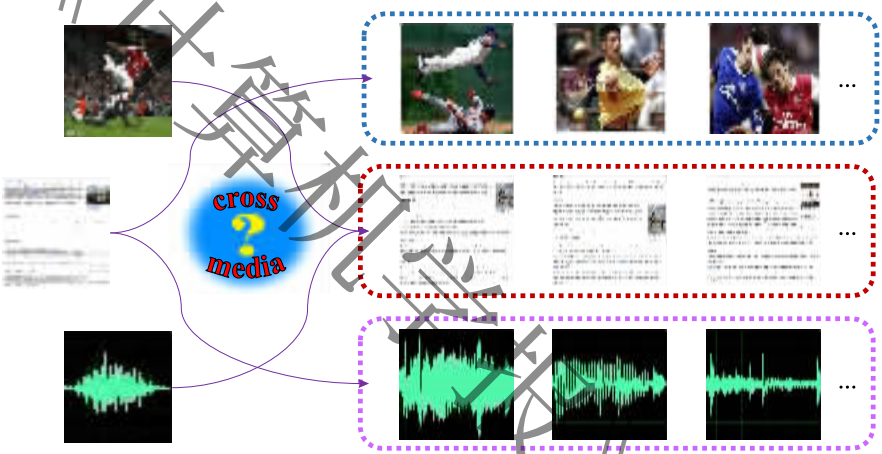



图2 跨媒体检索

近来,Rasiwasia等人[3]研究了一种借助典型相关分析(Canonical Correlation Analysis, CCA)[16-18,26]技术的跨媒体检索方法,以获得不同模态之间的共享描述.在文献[17]中,Hardoon等人呈现了一种使用核CCA的通用方法以学习网页图片和相关文本之间的语义描述.

跨媒体聚类[18-21]利用不同模态间样本的分布相似性,同时聚类多个模态以揭示不同模态间的潜在共享结构.如图3所示,在通常情况下,文本描述携带了更多的信息,这样就可以借助文本模态的分布诱导图片模态的聚类.目前,相关研究人员已经提出了一些卓有成效的跨媒体聚类方法.Chaudhuri等人[18]提出了一种基于CCA的跨媒体聚类方法,将多个模态的数据投影到一个低维空间上.文献[19]利用非负矩阵分解[27]寻找不同模态间兼容的聚类方法.Cai等人[20]提出了一种鲁棒大规模多模态聚类方法以处理大规模数据的异构描述.在文献[21]中,Eaton等人利用交替传播在不同模态间迁移约束,以更新聚类模型.

跨媒体分类[22-25]是利用从多个模态中找出的事物共享特征构建分类器,从而对多媒体数据进行分类.如图4所示,跨媒体分类方法利用已有的有标签跨媒体数据学习出一个性能优越的跨媒体分类器,然后使用这个学习出的分类器预测无标签跨媒体数据的类别.近年来,在跨媒体分类领域已经取得了一些进展.Sharma等人[22]提出了一种基于CCA的跨媒体特征提取通用方法以实现跨媒体分类.在文献[23]中,Su等人提出了一种3D对象类模型,通过姿势的估计与合成识别不可见模态.Zhang等人[24]提出了一种主动迁移学习方法以关联不同场景下的相貌变化.文献[25]基于boosting方法在每个模态上学习一个弱分类器,并通过加权组合产生整合分类器.

不同于单媒体学习,跨媒体学习[7-10]在学习过程中能够度量不同模态的媒体对象之间的相关性,

《计算机学报》

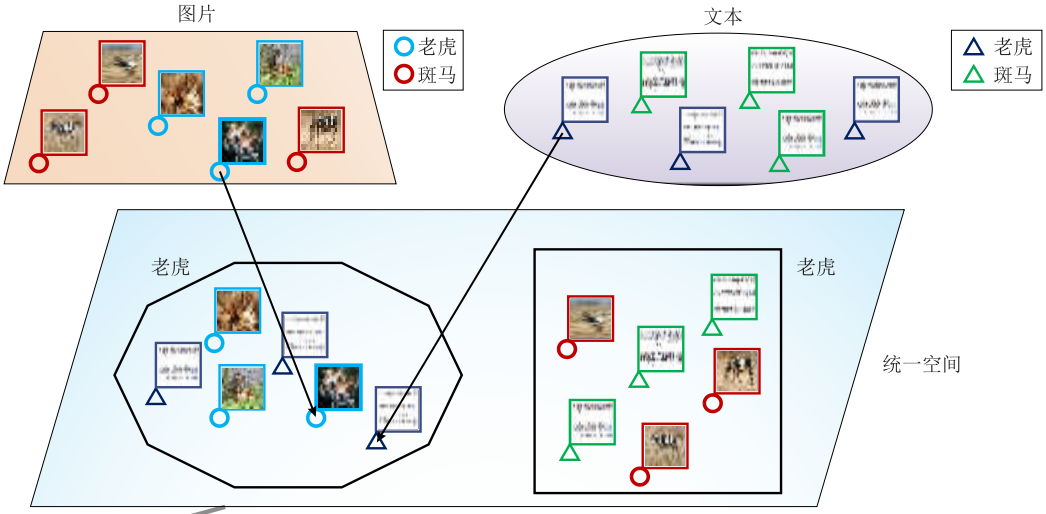



图 3　跨媒体聚类

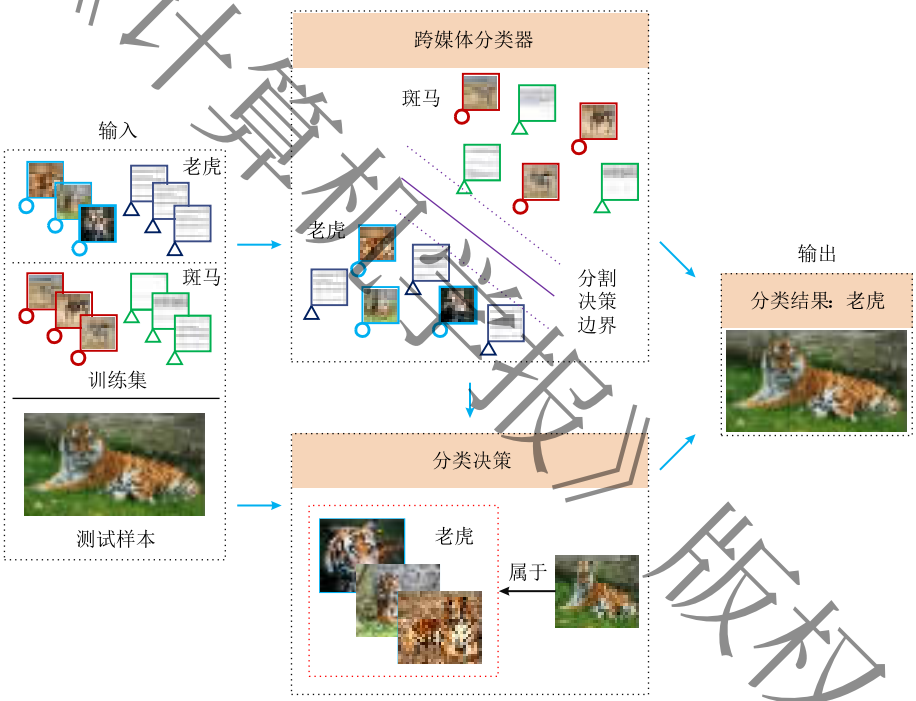



图 4　跨媒体分类

《计算机学报》版权

以实现不同模态间的灵活跨越. 然而，目前跨媒体学习还是一个崭新的研究方向，在相关领域的研究中，虽然国内外的研究人员已经设计出各类跨媒体学习方法，但仍存在许多困难与挑战亟待解决.

(1) 跨媒体数据的一致性描述

由于不同模态的跨媒体数据采用不同维度和不同属性的底层特征表达数据，使得异模间无法根据各自的底层特征直接度量异构描述之间的相关性，不同模态的异构描述间带有特征异构性[28]，从而使得跨媒体数据的相关性度量十分困难. 而在跨媒体学习的研究中，另一个困扰异模间特征分析的难题便是底层内容和高层语义之间的语义鸿沟[29]问题. 因此，挖掘跨媒体数据的一致性描述，以实现不同模态间的灵活跨越是十分必要的.

(2) 跨媒体增量学习

绝大多数跨媒体学习方法都涉及对输入矩阵的奇异值分解(Singular Value Decomposition，SVD)[30]. 因此，一旦新数据插入到原始数据集中，就会花费更多的时间重新计算增广输入数据矩阵的奇异值分解. 所以，为了避免高计算复杂度，研究有效且无损的增量跨媒体学习方法就显得十分重要.

(3) 缺失模态补全

随着信息技术的快速发展，越来越多的跨媒体数据出现在医疗诊断、网页分类和跨媒体分析等应

用中.然而,由于诸如高数据采集代价.不真实性和拒绝反馈等原因,这些应用都面临着某些模态缺失的问题,从而使得这些要求完整模态的应用能够使用的数据量极度有限.所以,当前跨媒体应用领域迫切需要一种跨媒体数据缺失模态补全方法.

(4)不同模态描述间的一一对应

在实际情况下,某些模态中的描述存在噪声,从而会破坏跨媒体描述之间的一一对应关系,所以需要找到一种方法重新建立这种一一对应的关系.

(5)同一模态内的样本冗余

同一模态内的某一对象的描述可能存在多个,如何借助其它模态的描述,消除冗余描述,节省存储空间,也是十分值得研究的问题.

### 1.2 跨媒体学习的相关方法

在一张维基百科描述"老虎"的网页中,存在着共生的图片和语音两种不同的模态共同表达"老虎"的概念.在特征层面上,从网页图片中提取的纹理,形状和色彩等视觉特征与从网页语音中提取的频率与时域等听觉特征共同描述了老虎的语义信息.但是由于这两种模态采用不同维度和不同属性的底层特征描述相同的语义概念——"老虎",所以在它们各自的底层特征空间中很难直接度量异模描述间的相关性.所以,跨媒体学习中首先亟待解决的困难就是不同模态的跨媒体数据在底层空间上的特征异构性问题[28].

尽管传统的基于内容的跨媒体学习研究没有解决跨媒体相关性度量问题,但在数据关系挖掘方面已有许多进展.为了解决跨媒体学习过程中的特征异构性[28]问题,缩减底层内容和高层语义之间的语义鸿沟[29],近年来国内外的研究人员提出了各种各样的跨媒体学习方法来探索跨媒体数据间的潜在关系,以期提高跨媒体学习的效率.这些研究主要包括:协同训练(Co-training)方法、多核学习(Multiple Kernel Learning)方法和共享子空间学习(Subspace Learning)方法.

#### 1.2.1 协同训练(Co-training)方法

Blum 和 Mitchell 提出的协同训练方法是一种充分利用不同模态间的互补性,以此增加其泛化能力的早期方案之一[31-35].该算法基于半监督学习,隐含地利用了模态间的冗余性:(1)所有模态都能够充分描述需要解决的问题,并且每个模态都包含大量的训练样本,从而在各个模态上都能够学习出一个泛化能力较强的学习器;(2)在给定标记时,不同模态之间是条件独立的.

协同训练[34]算法不断迭代进行,如图 5 所示.此类方法利用不同模态的有标记训练样本,在每个模态上单独训练出一个异模间互为关联的分类器,而所有异模分类器针对同一验证描述的输出结果是相似的.随后,每个异模分类器会预测余下的未标记样本的标记,并从中筛选出若干标记可信度较高的样本加入另一个异模分类器的有标记训练样本集中,以提高对方分类器的泛化能力,如此往复直至某个终止条件.在一致性的原则下,每次迭代的目的就是在验证集上最大化分类器之间的一致性.尽管每个分类器在验证集上的预测结果必然存在差异,但是这些差异会反向传播给训练集,从而训练出更加精准的分类器.进而在下一次迭代中,最小化每个分类器在验证集上预测结果的差异.

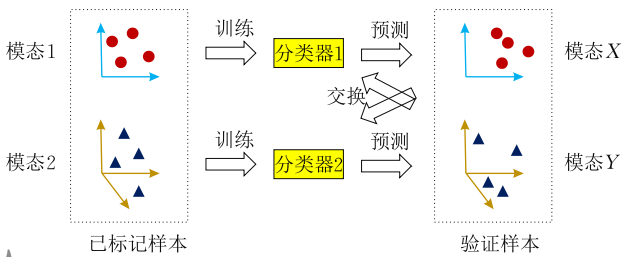



图 5 协同训练的过程

然而,在通常情况下对于绝大多数跨媒体数据集来说,使两个模态充分冗余的条件很难满足.因此,为了进一步增加协同训练算法的应用面,降低其应用的条件,在文献[36]中,Zhou 和 Li 等人设计出一种采用 3 个分类器进行协同训练的 Tri-training 算法.不同于其它协同训练方法,该算法利用集成学习(Ensemble Learning,EL)提高各个异模分类器的泛化能力(异模分类器的种类可以相同),同时对训练样本集也没有过于苛刻的要求.

#### 1.2.2 多核学习(Multiple Kernel Learning)方法

传统的单核方法虽然能够有效解决非线性模式分析问题,但是在面对存在不规则性且在高维特征空间中分布不平坦的海量高维异构数据时,此类方法便显得捉襟见肘.其原因在于,单核方法通过单个简单核映射所有异模描述的方式是不合理的.由此,近年来在非线性模式分析领域提出了不少多核学习方法[37-41].此类方法日渐成为当前核机器学习领域的一个新的热点.

如图 6 所示,多核学习方法是一种非线性的跨媒体模式分析方法.为了提高跨媒体学习的预测精度和分类准确率,该类方法会为每个模态分配一个特定的基本核函数与组合权系数.不同模态的基本

核函数会将对应模态的异构描述映射到一个由多个异构的底层特征空间构建的组合空间中，进而揭示出不同模态之间的关联性. 并且在这个组合空间中，利用各个基本核的非线性映射能力，使异构数据得以更好的表达. 因此，在多核框架下，多核学习的关键难点就在于不同模态的基本核和组合权系数的筛选问题.

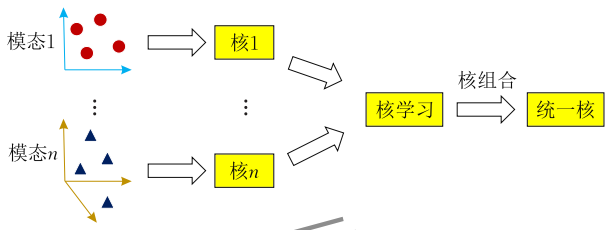



图 6　多核学习的示意图

在学习过程中，不是选择单一的核函数进行多核学习，而是选择一组核函数，组合不同的核. 由于不同的核对应于来自不同模态的各种相似性概念，所以核组合可能是一种更好地整合多个信息源，寻找最优解的方式. 在多核框架下，为了提高跨媒体学习的预测精度和分类准确率，不同模态的基本核函数会将对应模态的异构描述映射到一个由多个异构的底层特征空间构建的组合空间中，使异构数据得以更好的表达. 因此，多核学习的技术难点已经转变为不同模态的基本核和组合权系数的筛选问题. 而重中之重的问题在于如何选择不同基本核之间的组合权系数，进而构建组合空间. 针对这一问题，国内外的研究人员提出了一些卓有成效的多核学习方法，如 Li[39] 研究的超核方法，Xia 等人[42] 设计的基于 Boosting 的算法，Sonnenburg 等人[43] 提出的半无限线性规划方法，基于二次约束型二次规划的学习方法[44]，基于半定规划的多核学习方法[45]，以及近来出现的基于分组 Lasso 思想的多核学习方法[46].

此外，研究人员也提出了各种方法改进多核框架，使组合权系数和基本核函数能够有效结合. 例如，Gönen 等人[47] 设计的局部多核算法，文献[48]中的非平稳多核算法，以及 Kloft 等人[49] 提出的非稀疏算法等等. 另外，Sommer 等人[50] 将具有多尺度表示特性的核函数赋予每个模态，进而推动了多核学习方法在多尺度核方向上的发展. 而 Gehler 等人[51] 又将多核学习方法从有限个基本核函数的组合扩展为无限个核函数的组合.

1.2.3　共享子空间学习(Subspace Learning)方法

共享子空间学习[16-17,52-70] 旨在获得多个模态间共享的潜在子空间，以捕捉不同模态间的互补信息. 如图 7 所示，子空间学习方法利用跨媒体数据间的相关性，将异构数据投影到共享子空间上，从而消除不同模态间的特征异构性.

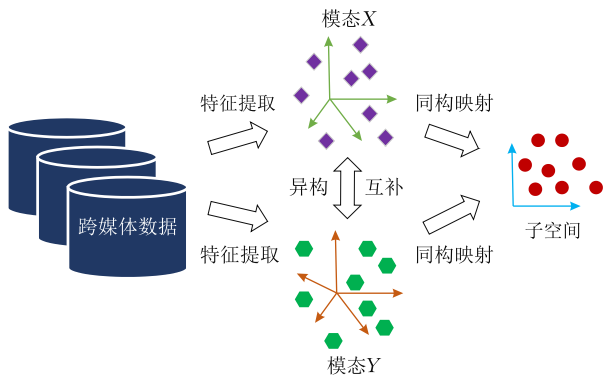



图 7　跨媒体共享子空间学习的示意图

此类方法将不同模态的异构描述投影到一个语义共享的子空间上，从而挖掘出不同模态间的语义互补性. 也就是，挖掘出不同模态描述间共享的语义特征，即语义一致性(Semantic Consistency). 比如，在同一张描述虎的网页中，存在着共生的文本、语音和图片 3 种不同模态的媒体对象，因此就会形成 3 个彼此异构的底层特征空间，人们无法直接度量它们之间的相似性. 但是，由于它们描述的内容都是虎，所以它们之间存在着某些共享的语义特征，只要能挖掘出这些特征，就可以把不同模态的异构描述投影到一个语义共享的子空间上，从而实现不同模态之间的灵活跨越.

当前的子空间学习方法可分为基于投影(Project)[16-17,52-53,70]、矩阵分解(Matrix Decomposition)[54-55]、任务(Task)[56-61] 和度量(Metric)[62-64] 等四大类. 基于投影的共享子空间学习方法利用特征映射提取多个模态间共享的潜在子空间；基于矩阵分解的共享子空间学习方法通过矩阵分解的方式提取不同模态间共享子空间的基向量；基于任务的共享子空间学习方法同时学习多个相关的任务，以提高各个任务的总体泛化性能；基于度量的共享子空间学习方法根据预先设定的邻域边缘，利用学习到的距离度量，将目标近邻拉入邻域边缘，并且将噪声点推出指定边缘，同时将同一数据的多模描述耦合到一起，从而提高描述的分类、聚类和检索性能.

1.2.4　各类跨媒体学习方法的对比与分析

这一小节将总结各类跨媒体学习方法的特性，对比并分析它们的优缺点.

协同训练利用不同模态间的互补信息，采用半

监督学习的方式,提高各个异模分类器的泛化能力.它要求每个模态都有大量的有标记样本,以训练出一个强学习器,然后使用这个学习器对未标记样本进行标记,并从中挑选出若干标记可信度较高的样本,加入对方模态的学习器中,以便对方利用这些新标记的样本进行更新,提高训练模型的泛化能力.这种方法的缺陷在于,它要求每个模态都有大量的有标记样本,学习条件太高.此外,它要求不同的模态间是同构的,所以无法解决不同模态间的特征异构性问题.

而多核学习是一种非线性的跨媒体模式分析方法.为了提高跨媒体学习的预测精度和分类准确率,该类方法会为每个模态分配一个特定的基本核函数与组合权系数.不同模态的基本核函数会将对应模态的异构描述映射到一个由多个异构的底层特征空间构建的组合空间中,进而揭示出不同模态之间的关联性.并且在这个组合空间中,利用各个基本核的非线性映射能力,使异构数据得以更好的表达.因此,在多核框架下,多核学习的关键难点就在于不同模态的基本核和组合权系数的筛选问题.其缺点在于很难针对每个模态选择恰当的基本核函数和组合权系数.

共享子空间学习旨在获得多个模态间共享的潜在子空间.利用跨媒体数据间的相关性和先验知识,将异构数据投影到共享子空间上,从而消除不同模态间的特征异构性,以捕捉不同模态间的互补信息和高层语义.这类方法的优点在于:(1)可以得到异构描述的同构描述,便于直接度量异构描述间的相关性,消除不同模态之间的异构性;(2)在投影过程中,能够充分利用模态间的相关性和相似分布,提取更多的互补信息;(3)能够利用不同模态间的语义互补性,在嵌入空间中,将相关异构描述耦合到一起,并根据不同模态间的分布相似性,利用已有的先验知识(类标签),将同类样本聚集到一起,而异类样本间保持一定距离,从而挖掘出更多的高层语义信息.

表1列出了各类跨媒体学习方法的优缺点.

**表1 各类跨媒体学习方法的优缺点**

| 方法 | 优缺点 |
|---|---|
| 协同训练 | (1)学习条件太高<br>(2)无法处理异构性 |
| 多核学习 | (1)基本核函数选择困难<br>(2)不易选择组合权系数 |
| 共享子空间学习 | (1)可以消除异构性<br>(2)基于模态间的相关性,捕捉互补信息<br>(3)能够利用先验知识,挖掘高层语义 |

从表1中可以看出,共享子空间学习方法和其它方法相比,具有明显的优势.此类方法不仅更加简单易行,而且能够消除不同模态间的特征异构性,捕捉到更多的互补信息和高层语义信息.正因为如此,目前共享子空间学习方法已成为机器学习领域的研究热点,国内外的研究人员已经提出了相当数量的基于共享子空间学习的跨媒体数据处理方法.在国际有关机器学习的主流会议上,如NIPS、ICML、IJCAI、AAAI、CVPR、KDD、ICCV、ICME及ICIP等,以及重要国际杂志上,如TPAMI、TKDE、TIP和JMLR等,每年都有大量的关于共享子空间学习方法的最新工作发表.下面就对有关共享子空间学习方法的国内外研究现状做介绍.

## 2 共享子空间学习

共享子空间学习方法[16-17,52-70]是利用线性或者非线性变换,将多个异构的底层特征空间投影一个潜在的共享子空间,使不同模态的异构描述在这个共享子空间中耦合到一起,从而挖掘出异模间的关联结构,进而得到反映不同模态共性信息的共享表示的机器学习方法.目前,此类方法已经成为跨媒体领域的研究热点之一.国内外的研究人员提出了各种各样的子空间方法以解决跨媒体学习中的相关问题.当前的共享子空间学习方法可分为四大类:基于投影的方法[16-17,52-53,70]、基于矩阵分解的方法[54-55]、基于任务的方法[56-61]和基于度量的方法[62-64].

### 2.1 基于投影的共享子空间学习方法

基于投影的共享子空间学习方法通过特征映射提取多个模态间共享的潜在子空间.此类方法分为线性投影和非线性投影两类.比较经典的线性投影方法分别是:典型相关分析法(Canonical Correlation Analysis,CCA)[16]、偏最小二乘法(Partial Least Squares,PLS)[52]、同构相关冗余变换(Isomorphic Relevant Redundant Transformation,IRRT)[70]和基于相关性的联合特征学习(Correlation-based Joint Feature Learning,CJFL)[70];而主流的非线性投影方法分别为:核典型相关分析法(Kernel Canonical Correlation Analysis,KCCA)[17]和深度典型相关分析法(Deep Canonical Correlation Analysis,DCCA)[53].

#### 2.1.1 典型相关分析法

在跨媒体问题中,CCA[16-18,26]已经成为一种经典的特征提取方法[71].它在计算生物学、金融分析

和信息检索等领域有着广泛的应用.

如图 8 所示，为了研究两组相互关联的随机变量 $\boldsymbol{X}=[x_1,\cdots,x_i,\cdots,x_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_x}$ 和 $\boldsymbol{Y}=[y_1,\cdots,y_i,\cdots,y_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_y}$（其中，$n$ 为样本数量，$d_x$ 为模态 $V_x$ 的维度，$d_y$ 为模态 $V_y$ 的维度）之间的整体线性相关关系，CCA 借助主成分分析的思想，将每一组变量作为一个整体进行研究而不是分析每一组变量内部的各个成分. CCA 分别在每一组变量中提取有代表性的新综合变量，使新综合变量之间的相关关系最大化，利用综合变量之间的相关关系反映原始变量之间的整体相关性.

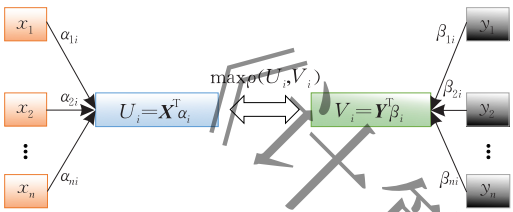



图 8　典型相关分析

CCA 的具体细节如下. 假设两组变量的第 $i$ 对线性组合为 $\boldsymbol{U}_i=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_i$ 和 $\boldsymbol{V}_i=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\beta}_i$，其中，典型变量 $\boldsymbol{\alpha}_i=(\alpha_{1i},\alpha_{2i},\cdots,\alpha_{ni})^{\mathrm{T}}$，$\boldsymbol{\beta}_i=(\beta_{1i},\beta_{2i},\cdots,\beta_{ni})^{\mathrm{T}}$. 在约束条件 $Var(\boldsymbol{U}_i)=\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}_i=1$ 和 $Var(\boldsymbol{V}_i)=\boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}_i=1(i=1,\cdots,p;p\in\{1,\cdots,\min(d_x,d_y)\})$ 下，求 $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\beta}_i$ 使得两者的相关系数 $\rho(\boldsymbol{U}_i,\boldsymbol{V}_i)=\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_i$ 达到最大. CCA 构造相应的优化问题为

$$\begin{aligned}\max_{\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i}\quad & \rho(\boldsymbol{U}_i,\boldsymbol{V}_i)=\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_i\\ \text{s.t.}\quad & Var(\boldsymbol{U}_i)=\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}_i=1\\ & Var(\boldsymbol{V}_i)=\boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}_i=1\end{aligned}\tag{1}$$

其中：$\boldsymbol{\Sigma}_{11}$ 是第 1 组变量的协方差矩阵；$\boldsymbol{\Sigma}_{22}$ 是第 2 组变量的协方差矩阵；$\boldsymbol{\Sigma}_{12}$ 是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差矩阵.

近年来，国内外的研究人员将 CCA 建立在两个不同模态 $V_x$ 和 $V_y$ 对应的矩阵 $X$ 和 $Y$ 上，分析 $X$ 和 $Y$ 的整体线性相关关系，从而将异构描述投影到低维同构空间上，以期提高跨媒体检索[3,17]、跨媒体聚类[18,22]和跨媒体分类[72]的性能. 而当一个模态为由类标签推导而来的预测值时，可以证明 CCA 等同于线性判别分析.

Rasiwasia 等人[3]提出了一种基于 CCA 的跨媒体检索方法以获得不同模态之间的共享描述. Hardoon 等人[17]呈现了一种使用核 CCA 的通用方法以学习网页图片和相关文本之间的语义描述. 此外，Chaudhuri 等人[18]提出了一种基于 CCA 的跨媒体聚类方法，将多个模态的数据投影到一个低维空间上. 进一步，Sharma 等人[72]提出了一种通用的基于 CCA 的跨媒体特征提取方法用于跨媒体分类和检索.

### 2.1.2　偏最小二乘分析法

偏最小二乘法[52]是一种综合了多元线性回归分析、典型相关分析以及主成分分析（Principal CompOnent Analysis，PCA）的算法. 因此，通过 PLS 不仅可以获得一个较为可靠的回归模型，同时还可以得到精简化的数据描述和不同模态间的相关性分析.

PLS 通过最小化误差的平方和，对多因变量向多自变量进行回归建模分析，不仅具有捕捉数据变化信息的能力[73-74]，同时还能够进行类似 CCA 的相关性分析. 在文献[22]中，Abhishek 等人通过提出的广义多视角分析（Generalized Multi-view Analysis，GMA）框架，阐述了主成分分析、典型相关分析以及偏最小二乘法之间的关系.

此外，PLS 回归方法能够在样本点个数少于变量个数且自变量存在严重多重相关性的条件下进行回归建模，并将所有的原有自变量包含在最终模型中. 另外，PLS 在回归建模过程中对非随机性噪声具有一定程度的抗噪能力，能够将数据中的噪声和有用信息区分开.

PLS 的具体细节如下：偏最小二乘回归分析将两组相互关联的随机变量 $\boldsymbol{X}=[x_1,\cdots,x_i,\cdots,x_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_x}$ 和 $\boldsymbol{Y}=[y_1,\cdots,y_i,\cdots,y_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_y}$ 进行中心化处理，得到中心化后的自变量矩阵 $\boldsymbol{E}=(E_1,E_2,\cdots,E_{d_x})\in\mathbb{R}^{n\times d_x}$ 和因变量矩阵 $\boldsymbol{F}=(F_1,F_2,\cdots,F_{d_y})\in\mathbb{R}^{n\times d_y}$. 记 $\boldsymbol{U}_i=\boldsymbol{E}\boldsymbol{\omega}_i$ 是 $\boldsymbol{E}$ 的第 $i$ 个成分，单位向量 $\boldsymbol{\omega}_i$ 是 $\boldsymbol{E}$ 的第 $i$ 个轴，即 $\|\boldsymbol{\omega}_i\|=1$. 记 $\boldsymbol{V}_i=\boldsymbol{F}\boldsymbol{c}_i$ 是 $\boldsymbol{F}$ 的第 $i$ 个成分，$\boldsymbol{c}_i$ 是 $\boldsymbol{F}$ 的第 $i$ 个轴，并且 $\|\boldsymbol{c}_i\|=1$. 基于 PCA 的原理，若要求 $\boldsymbol{U}_i$ 和 $\boldsymbol{V}_i$ 能够最大程度地分别代表 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 中的重要信息，则 $Var(\boldsymbol{U}_i)\to\max$ 和 $Var(\boldsymbol{V}_i)\to\max$. 另一方面，按照 CCA 的思想，$\boldsymbol{U}_i$ 要能够充分解释 $\boldsymbol{V}_i$ 中包含的信息，以满足回归建模的需要，使 $\boldsymbol{U}_i$ 和 $\boldsymbol{V}_i$ 的相关度最大化，即 $\rho(\boldsymbol{U}_i,\boldsymbol{V}_i)\to\max$. 从而需要求解如下优化模型：

$$\begin{aligned}&\max_{\omega_i,c_i}\quad \boldsymbol{E}\boldsymbol{\omega}_i,\boldsymbol{F}\boldsymbol{c}_i=\boldsymbol{U}_i,\boldsymbol{V}_i\\ &\text{s.t.}\quad \begin{cases}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{\omega}_i=1\\ \boldsymbol{c}_i^{\mathrm{T}}\boldsymbol{c}_i=1\end{cases}\end{aligned}\tag{2}$$

另外，尽管 PLS[52]和 CCA[16]都同时计算观测变量的低维嵌入，但它们之间仍然存在一定的区别：CCA 在嵌入空间最大化观测变量之间的相关性，而

计算机学报

PLS 最大化观测变量之间的协方差. 此外,虽然 PLS 与 CCA 都可以用于研究两组互相关联的变量间的彼此依赖关系,但是 CCA 只考虑一组自变量,无预测能力,而 PLS 具有预测功能,能够用一组自变量去预测另一组因变量.

### 2.1.3 核典型相关分析法

核典型相关分析法[17]是 CCA 方法在核空间上的非线性推广. 该方法是目前较为常用的基于核方法的非线性主元提取方法,它通过隐式非线性地将数据投影到一个高维特征空间,为 CCA 提供了一种可替代的解决方法. 当前使用比较频繁的核函数分别为:线性核函数[75],Goldberg 等人[76]提出的 $q$ 阶多项式核函数,Chang 等人[77]设计的高斯径向基(Radial Basis Function,RBF)核函数以及多层感知器(Multi-Layered Perceptron,MLP)核函数[78].

KCCA 的具体思想如下:假设 $\boldsymbol{X}=[x_1,\cdots,x_i,\cdots,x_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_x}$ 和 $\boldsymbol{Y}=[y_1,\cdots,y_i,\cdots,y_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_y}$ 分别表示两组相互关联的随机变量. $\boldsymbol{\Phi}_x$ 表示作用于 $X$ 上的核变换,即 $\boldsymbol{\Phi}_x(\boldsymbol{X})=(\Phi_x(x_1),\cdots,\Phi_x(x_i),\cdots,\Phi_x(x_n))$. 同理,$\boldsymbol{\Psi}_y$ 表示作用于 $Y$ 上的核变换,即 $\boldsymbol{\Psi}_y(\boldsymbol{Y})=(\Psi_y(y_1),\cdots,\Psi_y(y_i),\cdots,\Psi_y(y_n))$. KCCA 借助核理论,首先通过非线性核变换 $\Phi_x$ 和 $\Psi_y$ 将输入样本 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 非线性映射到一个高维(甚至是无穷维)的 Hilbert 空间中,使得原始数据在高维特征空间中成为线性数据;然后在该高维特征空间中利用 CCA 对变换后的数据提取主成分,找到两个向量 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$,使得 $\boldsymbol{U}=\Phi_x(\boldsymbol{X})\boldsymbol{\alpha}$ 与 $\boldsymbol{V}=\Psi_y(\boldsymbol{Y})\boldsymbol{\beta}$ 的相关系数 $\rho(\boldsymbol{U},\boldsymbol{V})$ 最大. 进而将高维空间变换为相对的低维空间,从而间接地实现了原始空间非线性问题的求解.

KCCA 可表示成如下形式的最优化问题:

$$\begin{aligned}&\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\quad \rho(\boldsymbol{U},\boldsymbol{V})=\boldsymbol{\alpha}^{\mathrm{T}}K_xK_y\boldsymbol{\beta}\\&\text{s.t.}\quad \boldsymbol{\alpha}^{\mathrm{T}}K_x\boldsymbol{\alpha}=\boldsymbol{I}\text{ and }\boldsymbol{\beta}^{\mathrm{T}}K_y\boldsymbol{\beta}=\boldsymbol{I}\end{aligned}\tag{3}$$

其中,$K_x=\boldsymbol{\Phi}_x^{\mathrm{T}}(\boldsymbol{X})\Phi_x(\boldsymbol{X})$ 和 $K_y=\boldsymbol{\Phi}_y^{\mathrm{T}}(\boldsymbol{Y})\Phi_y(\boldsymbol{Y})$ 分别对应 Hilbert 空间的核矩阵.

近年来,KCCA 作为 CCA 的一种非线性替代,得到了广泛的应用. Hardoon 等人[17]提出了一种基于高斯分布核函数的 KCCA 方法学习网页图片和相关文本之间的语义描述,以实现基于内容的图像检索. 在文献[79]中,Vinokourov 等人将线性核函数应用于跨语言信息检索,并且取得了显著的效果. 这些研究成果表明,相对于非线性核函数,线性核函数不仅更加快捷而且简单易行,同时也可获得较好的学习效果. 能够克服选择非线性核函数的困难,所以在实际应用中往往采用线性核函数.

### 2.1.4 深度典型相关分析法

尽管 KCCA 能够学习高维非线性变换,但是也存在产生的非线性变换受选择的核函数约束的缺陷. 此外,KCCA 还是一种非参数化的方法,所以它的计算复杂度的伸缩性很差. 而 Andrew 等人[53]提出的深度典型相关分析方法通过深度网络可以灵活地学习两个相关模态之间的非线性变换.

如图 9 所示,DCCA 同时学习两个模态的最大相关深度非线性映射,在学习过程中,每个模态对应一个深度网络,通过多层非线性变换,使输出层最大相关.

不同于 KCCA,DCCA 并不需要内积,从而为 KCCA 提供了一种非线性替代. 此外,DCCA 作为一种参数模型,无需参考训练集就可计算未知数据点的描述. 但是,DCCA 也存在学习过程中需要设置大量参数且学习时间长的缺陷.

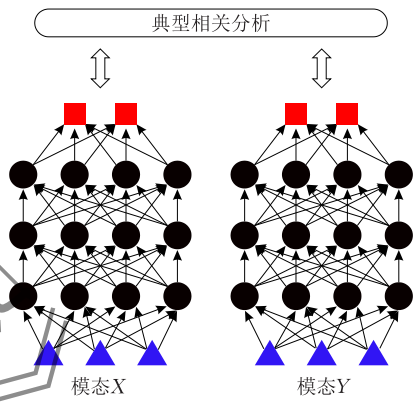



图 9 DCCA 模型(三角形表示输入特征,圆形表示隐藏特征,而输出层由方框表示)

目前,深度网络已经被广泛用于学习描述. 例如,深度波尔兹曼机[80]、深度自动编码机[81]和深度非线性前馈网[82]等. 此外,Ngiam 等人[83]提出的多模态自动编码机与 Srivastava 和 Salakhutdinov[84]提出的多模态有限波尔兹曼机都使用单一的深度网络连接两个模态的网络层进行协同学习. 并且这些方法在缺少一个模态的情况下,可以使用学习出的网络预测另外一个模态.

而 DCCA 和上述方法的主要区别在于,它同时学习两个模态的最大相关深度非线性映射,从而使学习出的描述尽可能相关. 此外,CCA 和 KCCA 可以使用的情况,DCCA 都可以使用. DCCA 也可以看作学习 KCCA 的一个核,但是映射函数并不受限于产生核 Hilbert 空间. 另外,尽管深度网络通常用于学习分类标签或者有监督地到另一个向量空间的映

计算机学报版权

射，但是 DCCA 却使用深度网络学习两个数据集到一个数据最大相关空间的非线性变换.

DCCA 方法的具体细节如下：假设模态 $V_x$ 和 $V_y$ 分别对应一个深度网络. 在模态 $V_x$ 的深度网络中，每个中间层有 $c_x$ 个单元，而输出层有 $o$ 个单元. 设 $x_i \in \mathbb{R}^{d_x}$ 是模态 $V_x$ 的第 $i$ 个实例. 实例 $x_i$ 的第 1 层输出为 $h_1 = s(\boldsymbol{W}_1^x x_i + \boldsymbol{b}_1^x) \in \mathbb{R}^{c_x}$，其中 $\boldsymbol{W}_1^x \in \mathbb{R}^{c_x \times d_x}$ 为权重矩阵，$\boldsymbol{b}_1^x \in \mathbb{R}^{c_x}$ 为基向量，而 $s: \mathbb{R} \mapsto \mathbb{R}$ 是应用于逐个元素的非线性函数. 那么，对于一个 $t_x$ 层网络，输出 $h_1$ 可用来计算下一层的输出 $h_2 = s(\boldsymbol{W}_2^x h_1 + \boldsymbol{b}_2^x) \in \mathbb{R}^{c_x}$，直到最终的描述 $f_x(x_i) = s(\boldsymbol{W}_{t_x}^x h_{t_x - 1} + \boldsymbol{b}_{t_x}^x) \in \mathbb{R}^o$ 被计算出来，其中 $\boldsymbol{W}_{t_x}^x \in \mathbb{R}^{o \times c_x}$，$\boldsymbol{b}_{t_x}^x \in \mathbb{R}^o$. 假设 $y_i$ 是模态 $V_y$ 的第 $i$ 个实例，采用同样的方式计算描述 $f_y(y_i)$，但是使用不同的参数 $\boldsymbol{W}_l^y$，$\boldsymbol{b}_l^y$ $(l = 1, \cdots, t_y)$ 和不同的结构参数 $c_y$ 和 $t_y$.

设 $\boldsymbol{\theta}_x$ 是由模态 $V_x$ 对应的深度网络中的全部参数 $\boldsymbol{W}_l^x$ 和 $\boldsymbol{b}_l^x$ $(l = 1, \cdots, t_x)$ 构成的向量，$\boldsymbol{\theta}_y$ 也由相同的方式构成. DCCA 联合学习两个模态的参数 $\boldsymbol{\theta}_x$ 和 $\boldsymbol{\theta}_y$，使得 $corr(f_x(x_i; \boldsymbol{\theta}_x), f_y(y_i; \boldsymbol{\theta}_y))$ 尽可能地大：

$$(\boldsymbol{\theta}_x^*, \boldsymbol{\theta}_y^*) = \arg\max_{(\boldsymbol{\theta}_x, \boldsymbol{\theta}_y)} corr(f_x(x_i; \boldsymbol{\theta}_x), f_y(y_i; \boldsymbol{\theta}_y)) \quad (4)$$

为了求解出最优解 $(\boldsymbol{\theta}_x^*, \boldsymbol{\theta}_y^*)$，在训练数据上估计相关目标函数的梯度. 假设对于一个具有 $n$ 个样本的数据集，矩阵 $\boldsymbol{H}_x \in \mathbb{R}^{o \times n}$，$\boldsymbol{H}_y \in \mathbb{R}^{o \times n}$ 的列向量是由两个模态的深度模型产生的顶层描述. 设 $\overline{\boldsymbol{H}}_x = \boldsymbol{H}_x - \boldsymbol{H}_x \boldsymbol{I}/n$ 和 $\overline{\boldsymbol{H}}_y = \boldsymbol{H}_y - \boldsymbol{H}_y \boldsymbol{I}/n$ 是中心化的数据矩阵. 定义 $\hat{\boldsymbol{\Sigma}}_{11} = \overline{\boldsymbol{H}}_x \overline{\boldsymbol{H}}_x^{\mathrm{T}}/(n-1)$，$\hat{\boldsymbol{\Sigma}}_{22} = \overline{\boldsymbol{H}}_y \overline{\boldsymbol{H}}_y^{\mathrm{T}}/(n-1)$ 和 $\hat{\boldsymbol{\Sigma}}_{12} = \overline{\boldsymbol{H}}_x \overline{\boldsymbol{H}}_y^{\mathrm{T}}/(n-1)$. $\boldsymbol{H}_x$ 和 $\boldsymbol{H}_y$ 的前 $k$ 个元素的总相关值就是矩阵 $\boldsymbol{P} = \hat{\boldsymbol{\Sigma}}_{11}^{-1/2} \hat{\boldsymbol{\Sigma}}_{12} \hat{\boldsymbol{\Sigma}}_{22}^{-1/2}$ 的前 $k$ 个奇异值的累加和. 如果 $k = o$，那么矩阵 $\boldsymbol{T}$ 的迹范数表示如下：

$$corr(\boldsymbol{H}_x, \boldsymbol{H}_y) = \|\boldsymbol{P}\|_* = tr(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{P})^{1/2} \quad (5)$$

DCCA 通过使用基于梯度的优化方法最优化参数 $\boldsymbol{\theta}_x$ 和 $\boldsymbol{\theta}_y$. 为了关于全部参数 $\boldsymbol{\theta}_x$ 和 $\boldsymbol{\theta}_y$ 计算 $corr(\boldsymbol{H}_x, \boldsymbol{H}_y)$ 的梯度，先计算 $corr(\boldsymbol{H}_x, \boldsymbol{H}_y)$ 关于 $\boldsymbol{H}_x$ 和 $\boldsymbol{H}_y$ 的梯度，然后使用反向传播. 如果矩阵 $\boldsymbol{P}$ 的奇异值分解为 $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$，那么

$$\frac{\partial corr(\boldsymbol{H}_x, \boldsymbol{H}_y)}{\partial \boldsymbol{H}_x} = \frac{1}{n-1}(2 \nabla_{11} \overline{\boldsymbol{H}}_x + \nabla_{12} \overline{\boldsymbol{H}}_y) \quad (6)$$

其中，$\nabla_{12} = \hat{\boldsymbol{\Sigma}}_{11}^{1/2} \boldsymbol{U} \boldsymbol{V}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{22}^{1/2}$，$\nabla_{11} = \frac{1}{2} \hat{\boldsymbol{\Sigma}}_{11}^{1/2} \boldsymbol{U} \boldsymbol{D} \boldsymbol{U}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{11}^{1/2}$. 而 $\partial corr(\boldsymbol{H}_x, \boldsymbol{H}_y)/\partial \boldsymbol{H}_y$ 有一个和式(6)对称的表述.

### 2.1.5　同构相关冗余变换法

作为一种经典的相关性分析方法，CCA[16] 能够获得来自不同模态的异构描述在特征同构空间的低维嵌入. 尽管 CCA 基于良好的数学模型，但是由于其自身固有的缺陷，降低的维度 $p$ 不可能大于 $\min(d_x, d_y)$. 这意味着在低维空间中会丢失一些必要的信息. 实际上，不同模态间的潜在信息往往隐藏在高维空间中[85].

而 Zhang 等人[70] 提出的同构相关冗余变换模型与上述方法的区别在于：它利用带有迹范数约束的同构相关冗余变换，将多个异构底层特征空间投影到一个高维冗余特征同构空间. 从而将来自不同模态的相关描述耦合到一起，以此捕捉更多的互补信息.

IRRT 方法的优化模型如下：

$$\Psi_1: \begin{array}{ll} \min\limits_{A,B} & \|\boldsymbol{X}A - \boldsymbol{Y}B\|_F^2 \\ \text{s.t.} & \|\boldsymbol{X}A_*\| \le \varepsilon \text{ and } \|\boldsymbol{Y}B_*\| \le \gamma \end{array} \quad (7)$$

其中，$\varepsilon$ 和 $\gamma$ 是预先指定的正参数用来控制变换后的数据携带的信息量. 在式(7)中引入迹范数约束的目的是在特征同构空间中捕捉更多不同模态间的潜在互补信息.

需要注意的是，直接求解式(7)中的问题 $\Psi_1$ 不是一个简单的任务，原因有两点. 首先，尽管问题 $\Psi_1$ 单独对于每个变量 $A$ 和 $B$ 都是凸问题，但问题 $\Psi_1$ 整体上是一个非凸问题；第二，迹范数约束是不平滑的，这就使得求解这个模型的最优解变得更加困难. 然而，引理 1 表明在问题 $\Psi_1$ 中对变换数据的迹范数约束可以松弛地转换为问题 $\Psi_2$ 中对投影矩阵的迹范数约束.

**引理 1.**　对于正数 $\delta$ 和任意两个可相乘矩阵 $\boldsymbol{C}$ 和 $\boldsymbol{D}$，如果

$$\|\boldsymbol{C}\|_* \|\boldsymbol{D}\|_* \le \delta,$$

那么

$$\|\boldsymbol{C}\boldsymbol{D}\|_* \le \delta.$$

证明.　由于迹范数是矩阵范数，所以它满足任意两个可相乘矩阵的相容性原理[86]. 由此，可以得到下式：

$$\|\boldsymbol{C}\boldsymbol{D}\|_* \le \|\boldsymbol{C}\|_* \|\boldsymbol{D}\|_*.$$

因此，如果 $\|\boldsymbol{C}\|_* \|\boldsymbol{D}\|_* \le \delta$，那么 $\|\boldsymbol{C}\boldsymbol{D}\|_* \le \delta$.

证毕.

根据引理 1，如果在问题 $\Psi_1$ 中预先指定的正参数 $\varepsilon$ 和 $\gamma$ 满足

$$\|X\|_* \|A\|_* \le \varepsilon \text{ 和 } \|Y\|_* \|B\|_* \le \gamma \quad (8)$$

那么，可以得到 $\|XA\|_* \le \varepsilon$ 和 $\|YB\|_* \le \gamma$. 因此，在问题 $\Psi_1$ 中的迹范数约束可以转化为

$$\|A\|_* \le \varepsilon / \|X\|_* \text{ 和 } \|B\|_* \le \gamma / \|Y\|_* \quad (9)$$

由此，利用式(9)中的松弛约束，可以将问题 $\Psi_1$

《计算机学报》版权

转化为如下问题 $\Psi_2$

$$\Psi_2: \begin{array}{l} \min\limits_{A,B} \ \|XA-YB\|_F^2 \\ \text{s.t.} \ \ \|A\|_* \le \varepsilon / \|X\|_* \text{ and } \|B\|_* \le \gamma / \|Y\|_* \end{array} \tag{10}$$

文献[70]中给出了式(10)的求解方法.

从表面上看,IRRT 模型类似于 CCA 的扩展. 然而,事实上 IRRT 在理论基础上完全不同于 CCA. 其原因有 3 点:(1) IRRT 使用线性高维投影;(2) IRRT 对变换数据施加低秩约束;(3) IRRT 的求解方法不同于 CCA.

(1) IRRT 使用线性高维投影

不同于 CCA 的线性降维投影,即 $p \le \min(d_x, d_y)$,IRRT 模型能够线性地将跨媒体数据映射到高维空间. 也就是说,在 IRRT 算法中 $p$ 可以比 $d_x$ 和 $d_y$ 都大,即 $p \gg \max(d_x, d_y)$. 而且 IRRT 也明显不同于经典的非线性高维投影的核方法(如 KCCA[17] 和 DCCA[53]).

(2) IRRT 对变换数据施加低秩约束

CCA 施加正交约束在典型变量上,从而导致只能将两组模态的数据投影到一个低维空间中,使得这两组变量在低维空间中最大相关. 不同于 CCA 中的正交约束,IRRT 对变换数据施加低秩约束,从而将跨媒体数据投影到一个高维冗余特征同构空间中.

(3) IRRT 的求解方法不同于 CCA

CCA 作为一种经典的统计分析技术,通常将其转化为一个特征值问题才能求得解析解. 然而,IRRT模型能够被转化为一个松弛的凸优化问题,而这个新的问题可以用迭代方法求解,从而收敛于最优解.

因此,在本质上 IRRT 模型和 CCA 没有太多的共同之处. 所以,它不是 CCA 的简单扩展.

### 2.1.6 基于相关性的联合特征学习法

在实际情况下,样本的原始描述中往往包含大量的冗余和噪声,进而严重影响数据的性能. 而基于相关性的联合特征学习[70]模型通过提取不同模态之间共享的低维语义子空间,以此消除原始空间中的冗余和噪声信息,从而得到更加准确、清晰、完整和客观的描述.

不同于其它子空间学习方法,CJFL 模型是一种基于迹率的共享子空间学习算法,它通过利用不同模态的描述之间的相关性,提取一个语义共享子空间. 其具体细节如下:假设 $\boldsymbol{X}=[x_1,\cdots,x_i,\cdots,x_n]^{\mathrm{T}} \in \mathbb{R}^{n\times d_x}$ 和 $\boldsymbol{Y}=[y_1,\cdots,y_i,\cdots,y_n]^{\mathrm{T}} \in \mathbb{R}^{n\times d_y}$ $(d_x=d_y)$ 分别为来自模态 $V_x$ 和 $V_y$ 的两组描述. 设 $\mathcal{C}_{\boldsymbol{X}}^t$ 和 $\mathcal{C}_{\boldsymbol{Y}}^t$ 分别是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中第 $t$ 个类的样本集合. 定义如下各式:

$$\mathcal{S}_{\boldsymbol{X}}^t=\{(x_i,x_j)\,|\,x_i,x_j\in\mathcal{C}_{\boldsymbol{X}}^t, i\ne j\} \tag{11}$$

$$\mathcal{S}_{\boldsymbol{Y}}^t=\{(y_i,y_j)\,|\,y_i,y_j\in\mathcal{C}_{\boldsymbol{Y}}^t, i\ne j\} \tag{12}$$

$$\mathcal{D}_{\boldsymbol{X}}^{tk}=\{(x_i,x_j)\,|\,x_i\in\mathcal{C}_{\boldsymbol{X}}^t \wedge x_j\in\mathcal{C}_{\boldsymbol{X}}^k, i\ne j, t\ne k\} \tag{13}$$

$$\mathcal{D}_{\boldsymbol{Y}}^{tk}=\{(y_i,y_j)\,|\,y_i\in\mathcal{C}_{\boldsymbol{Y}}^t \wedge y_j\in\mathcal{C}_{\boldsymbol{Y}}^k, i\ne j, t\ne k\} \tag{14}$$

设

$$\mathcal{S}_{\boldsymbol{X}}=\sum_t \mathcal{S}_{\boldsymbol{X}}^t \text{ 和 } \mathcal{S}_{\boldsymbol{Y}}=\sum_t \mathcal{S}_{\boldsymbol{Y}}^t \tag{15}$$

$$\mathcal{D}_{\boldsymbol{X}}=\sum_t\sum_k \mathcal{D}_{\boldsymbol{X}}^{tk} \text{ 和 } \mathcal{D}_{\boldsymbol{Y}}=\sum_t\sum_k \mathcal{D}_{\boldsymbol{Y}}^{tk} \tag{16}$$

显然,来自于 $\mathcal{S}_{\boldsymbol{X}}$ 或 $\mathcal{S}_{\boldsymbol{Y}}$ 的每对数据在语义上是彼此相似的,而来自于 $\mathcal{D}_{\boldsymbol{X}}$ 或 $\mathcal{D}_{\boldsymbol{Y}}$ 的每对数据在语义上是彼此不相似的.

为了消除原始空间中的噪声信息,需要利用先验知识(类信息)学习一个线性变换 $\boldsymbol{\Theta}\in\mathbb{R}^{d_x\times k}$ ($k$ 为子空间的维度),以此参数化语义共享子空间. 在数学上,需要最小化如下类内距离:

$$\begin{aligned} &\sum_{\forall(x_i,x_j)\in\mathcal{S}_{\boldsymbol{X}}} (\boldsymbol{\Theta}^{\mathrm{T}}x_i-\boldsymbol{\Theta}^{\mathrm{T}}x_j)^{\mathrm{T}}(\boldsymbol{\Theta}^{\mathrm{T}}x_i-\boldsymbol{\Theta}^{\mathrm{T}}x_j)+ \\ &\sum_{\forall(y_i,y_j)\in\mathcal{S}_{\boldsymbol{Y}}} (\boldsymbol{\Theta}^{\mathrm{T}}y_i-\boldsymbol{\Theta}^{\mathrm{T}}y_j)^{\mathrm{T}}(\boldsymbol{\Theta}^{\mathrm{T}}y_i-\boldsymbol{\Theta}^{\mathrm{T}}y_j) \\ &=\sum_{\forall(x_i,x_j)\in\mathcal{S}_{\boldsymbol{X}}} tr(\boldsymbol{\Theta}^{\mathrm{T}}(x_i-x_j)(x_i-x_j)^{\mathrm{T}}\boldsymbol{\Theta})+ \\ &\sum_{\forall(y_i,y_j)\in\mathcal{S}_{\boldsymbol{Y}}} tr(\boldsymbol{\Theta}^{\mathrm{T}}(y_i-y_j)(y_i-y_j)^{\mathrm{T}}\boldsymbol{\Theta}) \\ &=tr(\boldsymbol{\Theta}^{\mathrm{T}}J_{\mathcal{S}}\boldsymbol{\Theta})+tr(\boldsymbol{\Theta}^{\mathrm{T}}R_{\mathcal{S}}\boldsymbol{\Theta}) \\ &=tr(\boldsymbol{\Theta}^{\mathrm{T}}(J_{\mathcal{S}}+R_{\mathcal{S}})\boldsymbol{\Theta}) \end{aligned} \tag{17}$$

其中

$$\boldsymbol{J}_{\mathcal{S}}=\sum_{\forall(x_i,x_j)\in\mathcal{S}_{\boldsymbol{X}}} (x_i-x_j)(x_i-x_j)^{\mathrm{T}} \tag{18}$$

$$\boldsymbol{R}_{\mathcal{S}}=\sum_{\forall(y_i,y_j)\in\mathcal{S}_{\boldsymbol{Y}}} (y_i-y_j)(y_i-y_j)^{\mathrm{T}} \tag{19}$$

并且 $\boldsymbol{J}_{\mathcal{S}}+\boldsymbol{R}_{\mathcal{S}}$ 是由数据集 $J$ 和 $R$ 构成的联合类内散度矩阵. 与此同时,还要最大化如下类间距离:

$$\begin{aligned} &\sum_{\forall(x_i,x_j)\in\mathcal{D}_{\boldsymbol{X}}} (\boldsymbol{\Theta}^{\mathrm{T}}x_i-\boldsymbol{\Theta}^{\mathrm{T}}x_j)^{\mathrm{T}}(\boldsymbol{\Theta}^{\mathrm{T}}x_i-\boldsymbol{\Theta}^{\mathrm{T}}x_j)+ \\ &\sum_{\forall(y_i,y_j)\in\mathcal{D}_{\boldsymbol{Y}}} (\boldsymbol{\Theta}^{\mathrm{T}}y_i-\boldsymbol{\Theta}^{\mathrm{T}}y_j)^{\mathrm{T}}(\boldsymbol{\Theta}^{\mathrm{T}}y_i-\boldsymbol{\Theta}^{\mathrm{T}}y_j) \\ &=\sum_{\forall(x_i,x_j)\in\mathcal{D}_{\boldsymbol{X}}} tr(\boldsymbol{\Theta}^{\mathrm{T}}(x_i-x_j)(x_i-x_j)^{\mathrm{T}}\boldsymbol{\Theta})+ \\ &\sum_{\forall(y_i,y_j)\in\mathcal{D}_{\boldsymbol{Y}}} tr(\boldsymbol{\Theta}^{\mathrm{T}}(y_i-y_j)(y_i-y_j)^{\mathrm{T}}\boldsymbol{\Theta}) \end{aligned}$$

计算机学报版权

$$= tr(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{J}_{\mathcal{D}}\boldsymbol{\Theta}) + tr(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{R}_{\mathcal{D}}\boldsymbol{\Theta})$$
$$= tr(\boldsymbol{\Theta}^{\mathrm{T}}(\boldsymbol{J}_{\mathcal{D}} + \boldsymbol{R}_{\mathcal{D}})\boldsymbol{\Theta}) \tag{20}$$

其中

$$\boldsymbol{J}_{\mathcal{D}} = \sum_{\forall (x_i, x_j) \in \mathcal{D}_{\boldsymbol{X}}} (x_i - x_j)(x_i - x_j)^{\mathrm{T}} \tag{21}$$

$$\boldsymbol{R}_{\mathcal{D}} = \sum_{\forall (y_i, y_j) \in \mathcal{D}_{\boldsymbol{Y}}} (y_i - y_j)(y_i - y_j)^{\mathrm{T}} \tag{22}$$

并且 $\boldsymbol{J}_{\mathcal{D}}+\boldsymbol{R}_{\mathcal{D}}$ 是由数据集 $J$ 和 $R$ 构成的联合类间散度矩阵. 为了同时最小化类内距离并且最大化类间距离. 可以直接将上述问题归纳为如下迹率优化问题：

$$\max_{\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}=\boldsymbol{I}} \frac{tr(\boldsymbol{\Theta}^{\mathrm{T}}(\boldsymbol{J}_{\mathcal{D}} + \boldsymbol{R}_{\mathcal{D}})\boldsymbol{\Theta})}{tr(\boldsymbol{\Theta}^{\mathrm{T}}(\boldsymbol{J}_{\mathcal{S}} + \boldsymbol{R}_{\mathcal{S}})\boldsymbol{\Theta}) + \boldsymbol{\alpha}\|\boldsymbol{J}\boldsymbol{\Theta} - \boldsymbol{R}\boldsymbol{\Theta}\|_F^2 + \boldsymbol{\beta}\|\boldsymbol{\Theta}\|_F^2} \tag{23}$$

其中，对 $\boldsymbol{\Theta}$ 的正交约束用来消除原始空间中的冗余信息，联合类内和类间散度矩阵 $\boldsymbol{J}_{\mathcal{S}}+\boldsymbol{R}_{\mathcal{S}}$ 和 $\boldsymbol{J}_{\mathcal{D}}+\boldsymbol{R}_{\mathcal{D}}$ 充分利用了不同模态间样本分布的同一性，$\|\boldsymbol{J}\boldsymbol{\Theta}-\boldsymbol{R}\boldsymbol{\Theta}\|_F^2$ 表示基于相关性的残差，用来避免破坏不同模态描述之间的相关性，而正规化项 $\|\boldsymbol{\Theta}\|_F^2$ 控制模型的复杂度，$\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 为正规化参数.

#### 2.1.7　经典方法的特性对比

虽然 CCA 和 PLS 均为经典的线性低维投影方法，但是在相关分析过程中，只是利用了统计方法，并没有使用任何先验知识或者其它手段捕捉互补信息，从而导致在低维空间中会丢失一些必要的信息. 在文献[85]中，Cheng 等人已经验证不同模态间的潜在信息往往隐藏在高维空间中. KCCA 和 DCCA 作为 CCA 和 PLS 的非线性替代，可以将数据投影到高维空间，但在投影过程中，并没有利用 IRRT 模型中使用的低秩约束，所以能够捕捉到的互补信息有限. 而 CJFL 模型不仅充分利用了不同模态间的语义互补性，而且还考虑了模态间的相似分布，所以能够捕捉到更多的互补信息.

表 2 总结了上述基于投影的共享子空间学习方法的性能.

**表 2　基于投影的共享子空间学习方法的性能对比**

| 方法 | 特点 | 捕捉互补信息的性能 |
|---|---|---|
| CCA | 低维投影 | 较弱 |
| PLS | 低维投影 | 较弱 |
| KCCA | 高维投影 | 强 |
| DCCA | 高维投影 | 强 |
| IRRT | 高维投影 | 较强 |
| CJFL | 低维投影 | 很强 |

### 2.2　基于矩阵分解的共享子空间学习方法

近年来，国内外的研究人员提出了一类基于矩阵分解的共享子空间学习方法，以提取不同模态间的共享子空间的基向量. 如图 10 所示，此类方法将模态 $V_x$ 和 $V_y$ 对应的数据矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分解为两个矩阵的乘积，其中 $W$ 的列向量为不同模态间共享子空间的基向量，而 $P_1$ 和 $P_2$ 的列向量为原始数据在共享子空间中的坐标. 目前该类方法大致可分为两类：一类是基于非负矩阵分解[19]的方法，另一类是基于特征分解[30]的方法.

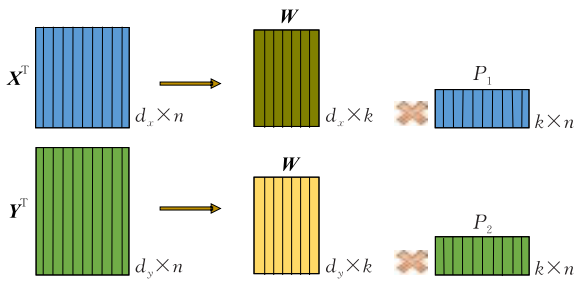



图 10　基于矩阵分解的共享子空间学习方法
（$k$ 为共享子空间的维度）

#### 2.2.1　基于非负矩阵分解的共享子空间学习法

非负矩阵分解（Nonnegative Matrix Factorization，NMF）[19]是一种良好的针对多变量数据的线性变换方法. 但是，不同于其它变换，NMF 基于对矩阵的非负约束.

近年来，Gupta 等人[54]提出了一种无监督的联合共享非负矩阵分解（Joint Shared Nonnegative Matrix Factorization，JSNMF）方法以捕捉两个数据集之间的共享基向量，而两个数据集各自的基向量分别对应于各自的判别空间. JSNMF 方法构建于非负矩阵分解框架之下，并且基于对全部矩阵的非负约束，使用辅助数据源提高主数据源的性能. 然而，JSNMF 算法也存在两个缺陷：(1) 一旦原始数据集发生变化，JSNMF 必须重新计算数据集之间的共享基向量；(2) 这种方法要求两个数据集的维度必须相同.

JSNMF 算法的细节如下：假设非负矩阵 $\boldsymbol{X}=[x_1,\cdots,x_i,\cdots,x_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_x}\geq 0$ 和 $\boldsymbol{Y}=[y_1,\cdots,y_i,\cdots,y_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_y}\geq 0$ 表示两个数据集，其中 $d_x$ 和 $d_y$ 分别为数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 各自特征空间的维度（JSNMF 算法要求两个数据集各自特征空间的维度相同，即 $d_x=d_y$），$n$ 是两个数据集中的样本数量. JSNMF 算法将矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分解为如下形式的非负矩阵的乘积：

$$\boldsymbol{X}^{\mathrm{T}} \approx [W|U]\boldsymbol{H} \text{ 和 } \boldsymbol{Y}^{\mathrm{T}} \approx [W|V]\boldsymbol{L} \tag{24}$$

其中，$W\in\mathbb{R}^{d_x\times k}(k<d_x=d_y)$ 的列向量为共享子空间的基向量，$U\in\mathbb{R}^{d_x\times(r_x-k)}$ 和 $V\in\mathbb{R}^{d_y\times(r_y-k)}$ 代表数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 各自的判别空间，$k$ 是共享基向量的数

计算机学报 JSNMF 版权

量，$r_x$ 和 $r_y$ 是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 各自潜在判别空间的维度，而矩阵 $\boldsymbol{H}\in\mathbb{R}^{r_x\times n}$ 和 $\boldsymbol{L}\in\mathbb{R}^{r_y\times n}$ 的前 $k$ 列向量为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中的样本在共享子空间中的坐标.

JSNMF 方法通过最小化联合分解误差来学习共享子空间，其数学模型如下：

$$\min_{W,U,V,H,L}\ \frac{\|\boldsymbol{X}^{\mathrm{T}}-[W|U]\boldsymbol{H}\|_F^2}{\|\boldsymbol{X}^{\mathrm{T}}\|_F^2}+\frac{\|\boldsymbol{Y}^{\mathrm{T}}-[W|V]\boldsymbol{L}\|_F^2}{\|\boldsymbol{Y}^{\mathrm{T}}\|_F^2}$$
$$\text{s.t.}\quad W,U,V,H,L\geq 0 \tag{25}$$

2.2.2 基于特征分解的共享子空间学习法

多输出正则投影(Multi-Output Regularized Feature Projection，MORFP)[55]算法通过矩阵特征分解将多个模态的特征向量映射到共享子空间，从而捕捉不同模态之间的关联性并保留了各个模态携带的信息.

MORFP 算法的具体细节如下：假设 $\boldsymbol{X}=[x_1,\cdots,x_i,\cdots,x_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_x}$ 和 $\boldsymbol{Y}=[y_1,\cdots,y_i,\cdots,y_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_y}$ 分别为两个不同的模态 $\boldsymbol{V}_x$ 和 $\boldsymbol{V}_y$ 对应的数据矩阵. MORFP 方法通过引入一个潜在变量模型，最小化各个模态的重建误差，并对模型进行矩阵特征分解，从而得到不同模态间的共享子空间，其优化模型如下：

$$\min_{A,B,U}\quad (1-\beta)\|\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{UA}\|_F^2+\beta\|\boldsymbol{Y}^{\mathrm{T}}-\boldsymbol{UB}\|_F^2$$
$$\text{s.t.}\quad \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I} \tag{26}$$

其中，$\boldsymbol{U}\in\mathbb{R}^{d_x\times p}(d_x=d_y)$ 为映射矩阵，$p$ 是共享基向量的数量，正交约束 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}$ 旨在保持 $p$ 个潜在共享基向量间的独立性，$\boldsymbol{A}\in\mathbb{R}^{p\times n}$ 和 $\boldsymbol{B}\in\mathbb{R}^{p\times n}$ 的列向量为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中的样本在共享子空间中的坐标.

然而，MORFP 算法是一种降维方法，在将跨媒体数据映射到共享子空间的同时，会破坏跨媒体数据中的原始结构，易于产生过拟合问题. 此外，如果原始样本集发生变化，多输出正则投影算法也无法解决新样本(Out-of-Sample，OS)问题.

## 2.3 基于任务的共享子空间学习方法

如图 11 所示，基于任务的共享子空间学习方法同时学习多个相关的任务，通过提取不同任务中不同模态间的共享子空间，以提高各个任务的总体泛化性能. 在学习过程中，每个任务通常具有多个模态，拥有不同的训练样本，但是所有的任务共享相同的特征. 到目前为止，基于任务的共享子空间学习方法涉及多任务学习(Multi-Task Learning)[56-57,87-88]、多标签学习(Multi-Label Learning)[58-60]和多类学习(Multi-Class Learning)[61]三大类. 这些方法在现实世界中都已经得到了广泛的应用.

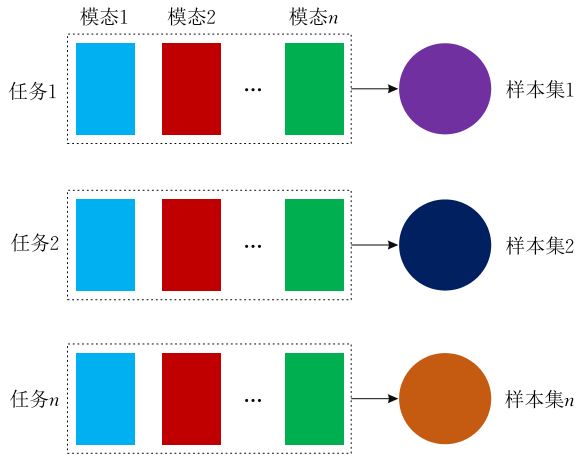



图 11 多任务学习

2.3.1 基于多任务学习的共享子空间学习法

Zhang 和 Huan[87]提出了一种半监督正规化多任务多模态学习(regularized Multi-View Multi-Task Learning，reg-MVMT)方法，利用有标签和无标签样本，提取不同任务间的相似度量矩阵. 该方法假设有 $T$ 个任务，而每个任务具有 $V$ 个模态. 并通过以下优化模型，利用多任务学习，提取任务间的相似度量矩阵 $\boldsymbol{\Omega}_v$：

$$\begin{aligned}\min_{W,\boldsymbol{\Omega}}\ &\sum_t^T\frac{1}{2}\|y_t-\boldsymbol{X}_t w_t\|^2+\\ &\frac{\mu}{2}\sum_{v'>v}^{V}\|\boldsymbol{U}_t^v w_t^v-\boldsymbol{U}_t^{v'}w_t^{v'}\|^2+\\ &\frac{\lambda}{2}\sum_{v=1}^{V}\|w_t^v\|^2+\frac{\gamma}{2}\sum_{v=1}^{V}tr(\boldsymbol{W}^v\boldsymbol{\Omega}_v^{-1}\boldsymbol{W}^{vT})\end{aligned} \tag{27}$$
$$\text{s.t.}\quad \boldsymbol{\Omega}_v\geqslant 0,\ tr(\boldsymbol{\Omega}_v)=1\ \text{and}\ v\in[1:V]$$

其中：$\boldsymbol{y}_t$ 为第 $t$ 个任务的标签矩阵；$\boldsymbol{X}_t$ 为第 $t$ 个任务的样本集；决策矩阵 $\boldsymbol{w}_t$ 是全部 $w_t^v(v\in[1:V])$ 串接的列向量；矩阵 $\boldsymbol{W}^v$ 的第 $t$ 列为 $w_t^v(t\in[1:T])$，$T\times T$ 阶矩阵 $\boldsymbol{\Omega}_v$ 中第 $i$ 行第 $j$ 列的元素表示任务 $i$ 和任务 $j$ 之间的相似度；$U_t^v$ 为第 $t$ 个任务中第 $v$ 个模态里的无标签数据集，$\mu$，$\lambda$ 和 $\gamma$ 为平衡参数. reg-MVMT 框架作为一种协同正规化方法，将多任务学习和跨媒体学习结合起来，可以有效度量不同任务间的相似度. 但是，这种方法只能处理同构的跨媒体数据.

Zuo 等人[88]提出了一种用于图像分类的多模态多任务特征提取(Multi-View Multi-Task Feature Extraction，MVMTFE)框架. MVMTFE 框架同时学习每个模态的特征提取矩阵和模态间的组合系

计算机学报

数，从而不仅可以处理相关和噪声特征，也能利用不同模态间的互补性进一步清除每个模态中的特征冗余. MVMTFE 方法假设有 $T$ 个任务，而每个任务具有 $V$ 个模态，其数学模型如下：

$$\min_{U^v,\theta} \sum_{t=1}^{T}\left\|\frac{1}{V}\sum_{v=1}^{V}\theta_v u_v^t x_v^t - \boldsymbol{y}^t\right\|_2^2 + \gamma_B\sum_{v=1}^{V}\|\boldsymbol{U}^v\|_{2,1} + \gamma_C\|\boldsymbol{\theta}\|_2^2 \tag{28}$$

$$\text{s.t.}\quad \sum_{v=1}^{V}\boldsymbol{\theta}_v = 1 \text{ and } \boldsymbol{\theta}_v > 0$$

其中，$\boldsymbol{y}^t$ 为第 $t$ 个任务的标签矩阵；$x_v^t$ 为第 $t$ 个任务中第 $v$ 个模态里的样本集；$\boldsymbol{U}^v=[u_v^1,u_v^2,\cdots,u_v^{\mathrm{T}}]$ 为第 $v$ 个模态的特征提取矩阵；$\boldsymbol{\theta}=[\theta_1,\theta_2,\cdots,\theta_V]$ 为不同模态间的组合系数；$\gamma_B$ 和 $\gamma_C$ 为平衡系数. MVMTFE 算法利用 $\ell_{2,1}$ 范数筛选特征，并使用交替优化求解 $\boldsymbol{U}^v$ 和 $\boldsymbol{\theta}$. 然而，此方法仍然无法处理异构数据.

近来，Ando 和 Zhang[56] 提出了一种交替结构优化（Alternating Structure Optimization，ASO）模型学习一种多个相关任务共享的预测结构. 在 ASO 算法中，每个任务对应一个线性预测函数：

$$f_i(x) = \boldsymbol{u}_i^{\mathrm{T}}x \tag{29}$$

其中，$\boldsymbol{u}_i\in\mathbb{R}^d$ 为第 $i$ 个任务的权重向量，它由该任务的特有成分和不同任务间的共享成分组成，$d$ 为样本空间的特征维度. $\boldsymbol{u}_i$ 的具体形式如下：

$$\boldsymbol{u}_i = w_i + \boldsymbol{\Theta}^{\mathrm{T}}v_i,\ \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}=\boldsymbol{I},\ \boldsymbol{\Theta}\in\mathbb{R}^{h\times d},\ h<d \tag{30}$$

其中，$w_i\in\mathbb{R}^d$ 和 $v_i\in\mathbb{R}^h$ 分别表示高维判别空间和低维共享空间的权重向量；$\boldsymbol{\Theta}$ 表示提取低维共享特征的结构参数. 需要注意的是，ASO 要求 $h<d$（$h$ 为提取的共享结构的维度）. 在数学上，ASO 的优化模型为

$$\min_{\{u_i,v_i\},\Theta} \sum_{i=1}^{m}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathcal{L}(\boldsymbol{u}_i^{\mathrm{T}}x_j^i, l_j^i) + \alpha\|u_i - \boldsymbol{\Theta}^{\mathrm{T}}v_i\|_2^2\right)$$

$$\text{s.t.}\quad \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta} = \boldsymbol{I} \tag{31}$$

其中，$m$ 为任务的数量；$x_j^i$ 为第 $i$ 个任务的第 $j$ 个实例；$l_j^i$ 为第 $i$ 个任务的第 $j$ 个实例的类标签；$n_i$ 是第 $i$ 个任务的样本数量；$\mathcal{L}$ 为损失函数；$\alpha>0$ 为正规化参数. Ando 和 Zhang[56] 也给出了求解式(31)中非凸 ASO 模型的交替优化算法，以获得 ASO 的局部最优解.

此外，Argyriou 等人[57] 提出了一种凸多任务特征学习（Convex Multi-Task Feature Learning，CMTFL）框架以挖掘多个任务共享的特征子空间，从而使所有的任务具有更好的泛化性能. CMTFL 算法没有直接学习共享特征，而是在模型中引入了 $\ell_{2,1}$ 范数间接地筛选共享特征. 其原理在于 $\ell_{2,1}$ 范数可以使矩阵中的某些列变成全零列，从而保证筛选出不同模态间的共享特征. CMTFL 算法的优化模型如下：

$$\min_{\boldsymbol{\Theta},\boldsymbol{U}} \|\boldsymbol{X}\boldsymbol{\Theta}\boldsymbol{U}_1 - \boldsymbol{L}_1\|_F^2 + \|\boldsymbol{Y}\boldsymbol{\Theta}\boldsymbol{U}_2 - \boldsymbol{L}_2\|_F^2 + \lambda\|\boldsymbol{U}\|_{2,1}$$

$$\text{s.t.}\quad \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}} = \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta} = \boldsymbol{I} \tag{32}$$

其中，$\boldsymbol{X}=[x_1,\cdots,x_i,\cdots,x_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_x}$ 和 $\boldsymbol{Y}=[y_1,\cdots,y_i,\cdots,y_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d_y}$ 分别为两个任务各自的样本构成的数据矩阵，$\boldsymbol{\Theta}\in\mathbb{R}^{d_x\times d_x}$（$d_x=d_y$）表示提取共享特征的结构参数，$\boldsymbol{U}_1\in\mathbb{R}^{d_x\times m}$ 和 $\boldsymbol{U}_2\in\mathbb{R}^{d_y\times m}$ 分别为每个任务各自的回归参数，$\boldsymbol{L}_1\in\mathbb{R}^{n\times m}$ 和 $\boldsymbol{L}_2\in\mathbb{R}^{n\times m}$ 分别为每个任务的类标签矩阵，$m$ 为类标签的数量，$\boldsymbol{U}=[U_1,U_2]$，$\lambda$ 为平衡因子，正交约束 $\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}}=\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}=\boldsymbol{I}$ 旨在保持 $d_x$ 个潜在共享基向量间的独立性. 由于式(32)中的模型是非凸的，所以文献[57]将其等价变换为式(33)所示的凸优化模型，以便得到最优解.

$$\min_{D,W} \|XA^*\boldsymbol{W}_1 - \boldsymbol{L}_1\|_F^2 + \|YB^*\boldsymbol{W}_2 - \boldsymbol{L}_2\|_F^2 + \lambda\sum_{t=1}^{2}\langle\boldsymbol{W}_t,\boldsymbol{D}\boldsymbol{W}_t\rangle$$

$$\text{s.t.}\ \boldsymbol{D}\in S_+^{d_x},\ trace(\boldsymbol{D})\le 1,\ \text{and } range(\boldsymbol{W})\subseteq range(\boldsymbol{D}) \tag{33}$$

其中，$\boldsymbol{W}_1=\boldsymbol{\Theta}U_1$，$\boldsymbol{W}_2=\boldsymbol{\Theta}U_2$，$\boldsymbol{W}=[W_1,W_2]$，$S_+^{d_x}$ 表示 $d_x$ 阶对称半正定矩阵的集合；$trace(\cdot)$ 表示矩阵的迹；$range(\cdot)$ 表示矩阵的脊.

### 2.3.2　基于多标签学习的共享子空间学习法

当所有的任务共享相同的训练数据和特征集，多任务学习等价于多标签学习[58-60,89]. 在单标签分类中，每个样本只和一个单一的标签相关，不同的类之间是彼此互斥的. 而在多标签学习中，每个样本和多标签相关. 由于所有的标签共享相同的输入空间，所以不同的标签传达的语义通常是相关的. 而标签之间的相关性会导致类相关，从而为提高分类性能带来了可能. 比如，每部电影都有主题和演员两个类标签. 像成龙这样的动作明星，其电影基本上都是动作片，而像金凯瑞这样的喜剧演员，其电影大多都是以喜剧为题材的. 显然，每一部电影的主题和演员这两个类标签在语义上是相关的. 也就是说，这两个类标签传达的语义是互补，而这里的互补指的就是不同标签之间共享的语义特征. 如果将这些共享的语义特征作为判别信息，就可以有效地区分不同类别的样本. 例如，对于成龙的电影，主题和演员两个类标签共享的语义特征就是动作，而对于金凯瑞的电

计算机学报版权

影，主题和演员两个类标签共享的语义特征就是喜剧. 显然，利用这两个共享的语义特征就可以将成龙和金凯瑞的电影区分开.

近来，Ji 等人[59-60]提出了一种针对多标签分类的共享子空间学习(Shared-Subspace Learning for Multi-label Classification，SSLMC)框架，该框架基于最小二乘方损失，以提取多个标签之间的相关信息. 在这个框架中，针对每个标签构造一个二值分类器，以判别不同的标签. 然而，不同于传统的方法，每个二值分类器不是独立构建的，多个标签之间共享一个低维子空间. SSLMC 算法的具体细节如下. 假设第 $i$ 个标签的预测函数由两个部分组成：一部分是原始数据空间，另一部分为共享子空间. 其形式如下：

$$f_i(x) = \boldsymbol{w}_i^{\mathrm{T}} x + \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{\Theta} x,\ \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}} = \boldsymbol{I} \tag{34}$$

其中：$\boldsymbol{w}_i \in \mathbb{R}^d$ 和 $\boldsymbol{v}_i \in \mathbb{R}^r$ 分别为样本空间和共享子空间的权重向量，$d$ 为样本空间的特征维度，$r$ 为共享子空间的维度，线性变换 $\Theta \in \mathbb{R}^{r\times d}(r<d)$ 为所有标签共享的结构参数，其行向量是彼此正交的. 通过线性变换 $\Theta$，将原始输入数据投影到一个低维共享子空间. 原始数据描述和其在共享子空间上的低维投影共同产生最终的预测结果. 在文献[62-63]中，Ji 等人通过构建如下正规化模型，求得结构化参数 $\{w_i, v_i\}$ 和 $\boldsymbol{\Theta}$：

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Theta}} \|\boldsymbol{XU}-\boldsymbol{L}\|_F^2 + \alpha\|\boldsymbol{U}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{V}\|_F^2 + \beta\|\boldsymbol{U}\|_F^2 \quad \text{s.t. } \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}} = \boldsymbol{I} \tag{35}$$

其中：$\boldsymbol{U} = [u_1, \cdots, u_m]^{\mathrm{T}} \in \mathbb{R}^{d\times m}$，$\boldsymbol{u}_i = \boldsymbol{w}_i + \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}_i$，$\boldsymbol{V} = [v_1, \cdots, v_m]^{\mathrm{T}} \in \mathbb{R}^{r\times m}$，$\alpha>0$ 和 $\beta>0$ 为正规化参数，$\boldsymbol{L} \in \mathbb{R}^{n\times m}$ 为类标签矩阵，$n$ 为样本数量，$m$ 是标签数量. 通过将式(35)中的模型转化为一个广义特征值问题就可求出 $\boldsymbol{\Theta}$ 的最优解[59-60].

#### 2.3.3 基于多类学习的共享子空间学习法

多类学习[61]涉及每个样本只和单一的标签相关的学习情形[90]. 从这个角度来看，多类学习可以看成是多标签学习的特例. 例如，在文本主题分类中，尽管存在大量的类别，但是多个类别之间是彼此相关的，可以基于一些潜在的共享特征识别不同的类别. 比如，老虎和斑马共享的特征为肺呼吸、胎生和四腿，根据这些特征可以将它们识别为哺乳动物. 因此，学习一个潜在的共享子空间，可以显著提高多类学习的性能，从而实现不同类别之间的知识迁移.

为了提取多类数据中的低维共享子空间，Amit 等人[61]提出了一种针对多类分类的共享子空间学习(Shared Structures in Multi-Class Classification，SSMCC)算法，以提取一个低秩共享结构. 在 SSMCC 方法中，每个类对应一个如下形式的线性分类器：

$$h_i(x) = \boldsymbol{w}_i^{\mathrm{T}} x \tag{36}$$

其中，$\boldsymbol{w}_i \in \mathbb{R}^d$ 是第 $i$ 个类的权重向量；$d$ 为原始数据的维度；$\boldsymbol{w}_i$ 由类特定成分和类间共享结构两个部分的乘积组成，其形式如下：

$$\boldsymbol{w}_i = (\boldsymbol{FG})_i \tag{37}$$

其中，$\boldsymbol{F} \in \mathbb{R}^{d\times p}$ 表示 $p$ 个类间共享特征；$\boldsymbol{G} \in \mathbb{R}^{p\times m}$ 表示基于类间共享特征预测 $m$ 个类别. 在数学上，Amit 等人[61]提出构建如下数学模型求解 $\boldsymbol{F}$ 和 $\boldsymbol{G}$：

$$\min_{\boldsymbol{F},\boldsymbol{G}} \frac{1}{2}\|\boldsymbol{F}\|_F^2 + \frac{1}{2}\|\boldsymbol{G}\|_F^2 + \lambda\sum_{i=1}^{m}\mathcal{L}[(\boldsymbol{FG})^{\mathrm{T}}x_i, l_i] \tag{38}$$

其中，$\mathcal{L}$ 为损失函数；$\lambda>0$ 为平衡因子. 由于式(38)中的模型是非凸的，Amit 等人[61]进一步提出将模型等价转化为式(39)中的凸模型，并最终使用梯度下降算法求解 $\boldsymbol{W}$ 的最优解：

$$\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_* + \lambda\sum_{i=1}^{m}\mathcal{L}(\boldsymbol{W}^{\mathrm{T}}x_i, l_i) \tag{39}$$

其中，$\boldsymbol{W} = \boldsymbol{FG}$.

### 2.4 基于度量的共享子空间学习方法

为不同模态的数据学习一个良好的度量，实现模态间的相似度对比，对包括跨媒体检索、聚类和分类这样的现实应用有着十分重要的意义. 因此，研究人员另辟蹊径使用度量学习(Metric Learning，ML)[62]提取不同模态间的共享子空间. 但是，这类方法需要计算各个样本间的欧氏或马氏距离，所以存在计算代价高的缺陷. 另外，此类方法还要求所有模态的描述是同构的，才能计算各个模态间的度量，无法处理异构描述.

如图 12 所示，ML 根据预先设定的邻域边缘，利用学习到的距离度量，将目标近邻拉入邻域边缘，并且将噪声点推出指定边缘，同时将同一数据的多模描述耦合到一起. 从而提高描述的分类、聚类和检索性能.

但是度量学习也存在计算代价高、无法度量异构描述的缺陷. 现有的基于度量的子空间学习方法分为两大类：一类使用欧氏距离[62]度量；另一类基于马氏距离[63]度量.

#### 2.4.1 基于欧氏距离度量的共享子空间学习法

Xie 和 Xing[63]提出了一种基于欧氏距离的多模态距离度量学习(Multi-Modal Distance Metric Learning，MMDML)框架，从而将两个不同模态的

《计算机学报》版权

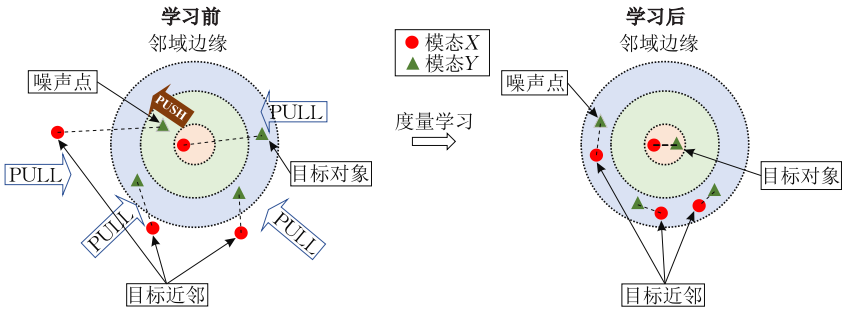



图 12　度量学习

数据嵌入到一个单一的潜在共享子空间，并在这个空间中借助距离监督学习整合不同模态的描述. 但是，MMDML 算法存在只能简单处理同构模态，不能度量异构描述的问题.

MMDML 框架基于多层协调（Multi-Wing Harmonium，MWH）[63]模型，使用欧氏距离度量学习整合不同的模态. 其中 MWH 模型是一个两层随机域，将多个可视特征模态嵌入到一个潜在语义空间. 假设有标签样本分为相似对和不相似对，MMDML 框架在潜在空间中最小化相似对间的欧氏距离，并在不相似对的样本间保持一定欧氏距离的边缘. 通过最大化随机域诱导的数据概率（Data Likelihood），并同时最小化欧氏距离诱导的损失函数，学习多个特征模态到语义空间的嵌入，以合理解释数据，并提供一种有效的距离度量. 其中数据概率作为一种自然形成的正规化项，可以避免在绝大多数度量学习算法中出现的过拟合现象. 不同于其它的度量学习算法，MMDML 框架中不需要进行半定规划（Semi-Definite Programming）或者核嵌入（Kernel Embedding）.

图 13 给出了多层协调模型的示意图. MWH 模型由双模输入 $x=(x_i)$ 和 $z=(z_j)$ 和一组隐藏单元 $h=(h_k)$ 组成. 在这个无向图模型中，输入单元间不存在连接. 每个模态的输入单元和隐藏单元都成了一个全偶图（Complete Bipartite Graph），其中同组单元间无连接，而和其它组单元间存在全连接. 这个拓扑结构具有 3 个条件独立假设：(1) 假设潜在变量 $h$ 和两个模态 $x$ 和 $z$ 是独立的，即 $p(x,z|h)=p(x|h)p(y|h)$；(2) 对于给定的 $x$ 和 $z$，$h$ 中的每个单元是彼此独立的，即 $p(h|x,z)=\prod_k p(h_k|x,z)$；对于给定的 $h$，每个模态中的单元是彼此独立的，即 $p(x|h)=\prod_i p(x_i|h)$，$p(z|h)=\prod_j p(z_j|h)$. 所有的观测变量和隐藏变量都是基于指数簇的，于是可以得到

$$p(x_i)=\exp\{\boldsymbol{\theta}_i^{\mathrm{T}}\phi(x_i)-A(\theta_i)\} \tag{40}$$

$$p(z_j)=\exp\{\boldsymbol{\eta}_j^{\mathrm{T}}\varphi(z_j)-B(\eta_j)\} \tag{41}$$

$$p(h_k)=\exp\{\boldsymbol{\lambda}_k^{\mathrm{T}}\psi(h_k)-C(\lambda_k)\} \tag{42}$$

其中，$\theta_i,\eta_j,\lambda_k$ 是自然参数；$\phi(\cdot),\varphi(\cdot),\psi(\cdot)$ 为充分统计量；$A(\cdot),B(\cdot),C(\cdot)$ 是 log 配分函数.

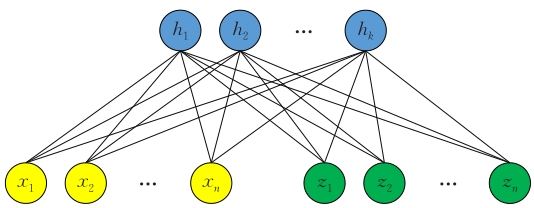



图 13　多层协调模型

MWH 模型通过引入一个额外项，在 log 域内耦合随机变量，从而得到联合分布：

$$p(x,z,h)\propto\exp\Big\{\sum_i\boldsymbol{\theta}_i^{\mathrm{T}}\phi(x_i)+\sum_j\boldsymbol{\eta}_j^{\mathrm{T}}\varphi(z_j)+\sum_k\boldsymbol{\lambda}_k^{\mathrm{T}}\psi(h_k)+\sum_{ik}\boldsymbol{\phi}(x_i)^{\mathrm{T}}\boldsymbol{W}_i^k\psi(h_k)+\sum_{jk}\boldsymbol{\varphi}(z_j)^{\mathrm{T}}U_j^k\psi(h_k)\Big\} \tag{43}$$

利用 $\boldsymbol{\Theta}$ 表示所有的参数 $(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{W},\boldsymbol{U})$，并利用最大似然率法求解 $\boldsymbol{\Theta}$. 通过联合分布 $p(x,z,h)$，可以得出条件分布：

$$p(x_i\mid h)\propto\exp\{\hat{\boldsymbol{\theta}}_i^{\mathrm{T}}\phi(x_i)-A(\hat{\boldsymbol{\theta}}_i)\} \tag{44}$$

$$p(z_j\mid h)=\exp\{\hat{\boldsymbol{\eta}}_j^{\mathrm{T}}\varphi(z_j)-B(\hat{\boldsymbol{\eta}}_j)\} \tag{45}$$

$$p(h_k\mid x,z)=\exp\{\hat{\boldsymbol{\lambda}}_k^{\mathrm{T}}\psi(h_k)-C(\hat{\boldsymbol{\lambda}}_k)\} \tag{46}$$

和变位参数：

$$\hat{\boldsymbol{\theta}}_i=\theta_i+\sum_k W_i^k\psi(h_k) \tag{47}$$

$$\hat{\boldsymbol{\eta}}_j=\eta_j+\sum_k U_j^k\psi(h_k) \tag{48}$$

$$\hat{\boldsymbol{\lambda}}_k=\lambda_k+\sum_i W_i^k\phi(x_i)+\sum_i W_i^k\varphi(z_j) \tag{49}$$

其中移位（Shift）是由观测单元和隐藏单元间的耦合诱导而来.

MMDML 框架的细节如下，假设一个跨媒体数据 $z=(x,y)$ 有两种模态 $V_x$ 和 $V_y$. 设模态 $V_y$ 中的样

《计算机学报》版权

本分为同类集 $\mathcal{S}=\{(y_i,y_j)\}$ 和异类集 $\mathcal{D}=\{(y_i,y_j)\}$ 两个集合. 通过 MWH 模型将模态 $V_y$ 嵌入潜在共享子空间,从而得到它的潜在概率描述 $t$:

$$t = E_{p(h|x,z;\Theta)} \tag{50}$$

其中,$t$ 是 $\Theta$ 的函数;隐藏单元 $h$ 为潜在主题;$\Theta$ 为潜在共享子空间的结构化参数. 由此,MMDML 方法构建如下优化模型:

$$\begin{aligned} &\min_{\Theta} \ \sum \| t_i - t_j \|^2 \\ &\text{s.t.} \ (y_i,y_j) \in \mathcal{S} \ \text{and} \ \| t_i - t_j \|^2 \geq 1 \end{aligned} \tag{51}$$

其中,$t_i$ 是 $y_i$ 在潜在共享子空间的描述. 构建此模型的目的在于最小化同类样本间的欧氏距离,同时使异类样本间的边缘大于等于 1.

2.4.2 基于马氏距离度量的共享子空间学习法

马氏距离[62]度量可以保持欧氏距离度量的所有特性,并且还可对其特征分解,从而转化为欧氏距离度量. 由此,基于马氏距离度量的共享子空间学习也已成为当前的研究热点.

近年来,Malave 和 Talbott[64]提出了一种基于马氏距离度量的共享子空间多度量学习(Shared Subspace for Multiple Metric Learning,SSMML)方法,以提取一个共享低秩子空间. SSMML 方法没有直接计算数据之间的距离,而是使用一个线性变换 $L$ 将数据映射到一个共享子空间里. 其距离函数为

$$d(i,j) = (x_i - x_j)^{\mathrm{T}} \boldsymbol{M} (x_i - x_j) \tag{52}$$

其中,$x_i$ 为模态 $V_x$ 中的第 $i$ 个样本;$\boldsymbol{M}=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\geqslant 0$ 为对称半正定度量矩阵. 当 $\boldsymbol{M}$ 为单位矩阵时,距离函数蜕变为欧氏距离. 从而式(52)可进一步转化为

$$d(i,j) = (x_i - x_j)^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} (x_i - x_j) \tag{53}$$

$$d(i,j) = (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} - \boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}})(x_i \boldsymbol{L} - x_j \boldsymbol{L}) \tag{54}$$

$$d(i,j) = (x_i \boldsymbol{L} - x_j \boldsymbol{L})^{\mathrm{T}} (x_i \boldsymbol{L} - x_j \boldsymbol{L}) \tag{55}$$

而对于每一个模态,相应的距离函数为

$$d(i,j) = (x_i - x_j)^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{w}_t \boldsymbol{L} (x_i - x_j) \tag{56}$$

其中,$\boldsymbol{w}_t$ 为第 $t$ 个模态的权重向量. 线性变换 $L$ 将所有模态的样本映射到一个共享子空间上. 设所有模态的样本分为同类集 $\mathcal{S}=\{(x_i,x_j)\}$ 和异类集 $\mathcal{D}=\{(x_i,x_j)\}$ 两个集合. SSMML 算法学习的度量矩阵 $\boldsymbol{M}$ 会使同类样本的马氏距离达到最小,而异类样本间的边缘大于等于 1. 其优化模型为

$$\begin{aligned} &\min_{\boldsymbol{M}} \ \mathbf{1}^{\mathrm{T}} \varepsilon + \sum_{(x_i,x_j)\in\mathcal{S}} d(i,j) \\ &\text{s.t.} \ 1 - \varepsilon \underset{(x_i,x_j)\in\mathcal{S},(x_i,x_k)\in\mathcal{D}}{\leq} d(i,k) - d(i,j) \\ &\varepsilon \geq 0, \ M \geqslant 0 \end{aligned} \tag{57}$$

其中,$\varepsilon$ 为控制边缘的参数. 最终,文献[64]将式(57)中模型转化为特征值问题以求解 $M$ 的最优解.

## 2.5 各类共享子空间学习方法的对比与分析

综上所述,目前共享子空间方法共分为四大类,每类方法都具有不同的功能、处理数据、应用领域和特性.

基于投影的共享子空间学习方法能够处理异构数据,消除不同模态间的特征异构性. 此类方法主要用于异构数据的融合表示,在跨媒体数据的分类、聚类、检索和特征提取中有广泛应用. 但其缺点为,在投影过程中,需要辅助的先验知识或其它辅助手段,否则会丢失重要的互补信息.

基于矩阵分解的共享子空间学习方法被普遍应用于跨媒体数据的分类、检索、特征提取中. 该类方法主要用于子空间基元提取,以获取不同模态间的共享描述. 但目前已有的方法只能处理同构数据,并且一旦原始数据集发生变化,就必须重新计算不同模态间的共享基向量.

基于任务的共享子空间学习方法能够提高各个任务的总体泛化性能. 近年来,国内外的研究人员多将此类方法应用于跨媒体数据的分类、特征学习、特征选择. 虽然该类方法能够处理多个相关任务,但每个任务中涉及的异模数据必须是同构的. 此外,这类方法也存在模型复杂、参数多的缺陷.

基于度量的共享子空间学习方法旨在学习不同模态间的相似性度量矩阵,以度量不同模态间的相似度,从而完成跨媒体数据的聚类、检索、特征学习. 然而,其在学习过程中,需要度量大量样本间的欧氏或者马氏距离,所以计算量大且耗时严重,学习过程漫长. 此外,容易产生过拟合现象,更无法处理异构数据.

表 3 列出了各类共享子空间学习方法的功能、

**表 3 各类共享子空间学习方法的对比**

| 类别 | 功能 | 数据 | 应用 | 特性 |
|---|---|---|---|---|
| 基于投影的共享子空间学习方法 | 异构数据的融合表示 | 异构数据 | 分类、聚类、检索、特征提取 | 需要辅助的先验知识,否则会丢失重要的互补信息 |
| 基于矩阵分解的共享子空间学习方法 | 子空间基元提取,获取不同模态间的共享描述 | 同构数据 | 分类、检索、特征提取 | 一旦原始数据集发生变化,就必须重新计算不同模态间的共享基向量 |
| 基于任务的共享子空间学习方法 | 提高各个任务的总体泛化性能 | 同构数据 | 分类、特征学习、特征选择 | 模型复杂,参数多 |
| 基于度量的共享子空间学习方法 | 学习不同模态间的相似性度量矩阵 | 同构数据 | 聚类、检索、特征学习 | 计算量大,学习过程漫长,容易出现过拟合 |

计算机学报版权

处理数据、应用领域和特性.

## 3　跨媒体学习领域亟待解决的问题

上述这些跨媒体学习方法，都是在假设全部跨媒体数据是完备的条件下，对跨媒体数据进行各种处理. 然而，在现实情况下，由于不确定性和人为因素，在跨媒体数据中必然存在着缺失、噪声和冗余. 这些问题不仅严重干扰跨媒体分析的应用，而且会导致错误的分析结果，浪费大量的计算时间和宝贵的存储空间.

根据解决问题的不同，在以下领域还有待进一步的研究：(1) 跨媒体数据的增量学习；(2) 跨媒体数据的缺失补全；(3) 跨媒体数据的噪声去除；(4) 跨媒体数据的维度缩减；(5) 跨媒体数据的样本选择.

### 3.1　跨媒体数据的增量学习

随着跨媒体数据量的增大，数据会被分散存储在不同的地方，所以通常很难将它们同时收集到一起；此外，许多在线应用会不断地产生诸如流数据这样的新数据. 比如，随着网络技术的快速发展，在万维网上存在着大量的多标签多模态图片. 而每天各大主流网站上又会更新不计其数的图片，因此实时在线分类这些更新的图片就变得越发重要.

然而，上述这些无增量学习方法在学习之前几乎不可能获得所有的输入数据，每次都要重复使用已有数据，从而导致高代价的重复计算. 因此，由于缺少可扩展性，上述方法很难满足低计算复杂度的要求.

特别需要注意的是，相当数量的共享子空间学习算法(比如，基于矩阵分解的共享子空间学习方法[54-55])的计算代价主要取决于输入矩阵的奇异值分解. 因此，一旦新数据插入到原始数据集中，上述算法就会花费更多的时间重新计算增广输入数据矩阵的奇异值分解. 所以，为了避免高计算复杂度，提出一种有效且无损的增量奇异值分解更新算法就显得十分重要.

通常情况下，使用奇异值分解更新(Singular Value Decomposition Updating，SVDU)[91] 算法对增广输入数据矩阵进行奇异值分解. 尽管这种方法是一种经典的更新算法，并且已经成功地应用于增量主成分分析(Incremental Principal CompOnent Analysis，IPCA)，但是 SVDU 方法要求数据矩阵具有所谓的低秩加位移(Low-Rank-Plus-Shift，LRPS)[91] 结构. 然而，在实际中，很难证明数据矩阵满足这种结构. 更糟糕的是，如果数据矩阵没有这种结构，就会导致不正确的分解，必须估算分解误差. 此外，即使当原始数据矩阵拥有这种结构，随着新数据添加到初始数据集中，数据矩阵的结构也会发生变化.

因此，Zhang 等人[92] 提出了一种适用于多标签图像分类的增量共享子空间学习(Incremental Shared Subspace Learning For Multi-Label Classification，ISSLMC)算法. 通过利用提出的增量无损矩阵分解，该算法能够增量地执行而无需使用原始已存在的数据，从而避免了在提取共享子空间过程中的高计算复杂度. 此外，不同于 SVDU 算法，无论输入数据矩阵具有何种结构，ISSLMC 算法都不会引入误差. 然而，ISSLMC 算法存在无法处理异构数据的缺陷.

### 3.2　跨媒体数据的缺失补全

近年来，国内外的研究人员提出了一些基于矩阵补全的跨媒体数据的缺失补全方法. 目前该类方法大致可分为四类：一类是基于欧氏距离的[93] 方法，其次是基于矩阵分解的[30] 方法，再者是基于选择填补[94-95] 的方法，最后是基于特征选择[96] 的方法.

基于样本间的欧氏距离，$K$-最近邻(KNN)[93] 算法通过近邻样本的均值填补缺失值. 尽管这种方法具有简单易行的特点，但若样本中存在噪声，就会大大降低矩阵补全的效果.

奇异值分解(SVD)[30] 是一种基于低秩逼近的补全方法. 该方法首先将缺失值初始为 0，然后对填补矩阵进行奇异值分解，接着利用低秩矩阵中的对应值更新缺失值，直至收敛. 然而，如果存在大量缺失值，SVD 则无法对填补矩阵进行奇异值分解.

Muslea 等人[94] 通过实验已经验证高效而鲁棒的跨媒体学习方法需要将主动(Active)学习和半监督学习相结合. Qian 等人[95] 根据这一点，提出了一种半监督的跨媒体数据缺失标签填补(Semi-Supervised-Dimension Reduction for Multi-label and Multi-view Learning，SSDR-MML)方法，利用异模数据间的相关性，填补大量无标签数据的缺失标签. 假设有 $p$ 个相关任务 $\mathcal{T}=\{t^1,t^2,\cdots,t^p\}$，并且每个任务都有 $n$ 个样本(其中包括有标签和无标签实例)，不同的任务具有不同的标签集，而每个样本具有 $q$ 个模态，$x_i^k$ 表示第 $i$ 个实例的第 $k$ 个模态的特征向量. 对于第 $k$ 个任务 $t^k$ 和对应标签集 $\mathcal{C}^k=\{1,2,\cdots,c^k\}$，都有一个相应的二值分类器 $F^k=[F_l^k;F_u^k]\in\mathbb{B}^{n\times c^k}$ ($F_l^k$ 针对有标签样本，而 $F_u^k$ 预测无标签样本的类别). 每个实例都具有 $p$ 个标签向量 $\{f^1,f^2,\cdots,f^p\}$. 在第 $k$

个任务中，当且仅当第 $i$ 个实例属于第 $j$ 个类时，$f_{ij}^k=0$. 不失一般性，假设第 $k$ 个任务中所有的样本点已排序，前 $l^k$ 个样本是有标签的，后 $u^k$ 个样本是无标签的(其中，$n=l^k+u^k$，$l^k \ll n$). 每个任务 $t^k$ 中有标签样本的标签构成标签矩阵 $\boldsymbol{Y}^k \in \mathbb{B}^{l^k \times c^k}$. SSDR-MML 算法的优化模型如下：

$$\begin{aligned} \min_{W,F} \quad & \sum_{k=1}^{q} a^k \sum_{i=1}^{n} \left\| x_i^k - \sum_{j=1}^{n} w_{ij} x_j^k \right\|_F^2 + \\ & \sum_{k=1}^{p} \beta^k \sum_{i=1}^{n} \left\| f_i^k - \sum_{j=1}^{n} w_{ij} f_j^k \right\|_F^2 + \lambda \| W \|_F^2 \\ \text{s.t.} \quad & \forall i,\ w_i.\mathbf{1} = 1 \text{ and } \boldsymbol{F}_l^k = \boldsymbol{V}^k Y^k \end{aligned} \tag{58}$$

其中，$\boldsymbol{V}^k$ 为第 $k$ 个任务的节点正规化矩阵；$\boldsymbol{F}_l^k$ 表示有标签样本的分类器，$\alpha^k$，$\beta^k$ 和 $\lambda$ 是平衡参数.

Qian 等人在文献[95]中提出了一种交替优化算法，求解式(58)的局部最优解. 在交替求解过程中，为了填补 $\boldsymbol{F}_u^k$ 中缺失的标签，假设权重矩阵 $\boldsymbol{W}$ 不变，式(58)中的问题可以重新简化为如下模型：

$$\begin{aligned} \min_{F^k} \quad & \frac{1}{2} tr\{(\boldsymbol{F}^k)^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{F}^k\} \\ \text{s.t.} \quad & F_l^k = \boldsymbol{V}^k \boldsymbol{Y}^k \end{aligned} \tag{59}$$

其中，$\boldsymbol{I}$ 为单位矩阵；$\boldsymbol{W}=[W_{ll}, W_{lu}; W_{ul}, W_{uu}]$. 文献[103]给出了 $\boldsymbol{F}_u^k$ 的解析解：

$$\begin{aligned} \boldsymbol{F}_u^k = & (\boldsymbol{W}_{lu}^{\mathrm{T}} \boldsymbol{W}_{lu} + (I_u - W_{uu})^{\mathrm{T}} (I_u - W_{uu}))^{-1} \\ & (\boldsymbol{W}_{lu}^{\mathrm{T}} (I_l - W_{ll}) + (I_u - W_{uu})^{\mathrm{T}} \boldsymbol{W}_{ul}) \boldsymbol{V}^k \boldsymbol{Y}^k \end{aligned} \tag{60}$$

SSDR-MML 方法的不足之处在于，它只能填补缺失的标签，对于跨媒体数据的缺失描述却无能为力.

在文献[96]中，Xiang 等人提出了一种能够处理跨媒体逐块(Block-Wise)缺失数据的特征选择(incomplete Source-Feature Selection，iSFS)方法. 其基本思想为：首先，针对每个模态学习一个单独的分类器 $\beta_k$($k=1,2,\cdots,m$，$m$ 为模态的数量)；然后，通过额外的正规化项，将所有模型整合到一起. 为了能够处理跨媒体逐块缺失数据，iSFS 方法根据不同模态的缺失情况，将全部数据分成不同的组. 在每个分组内，都包含完整的和缺失模态的跨媒体数据. 不同分组之间不是完全独立的，同一完整的跨媒体数据可以被包含在不同的分组里.

假设有一组跨媒体数据 $\boldsymbol{X}=[X_1, X_2, \cdots, X_m]$. 针对同一对象，每个跨媒体数据具有 $m$ 个模态. 每个模态都有 $n$ 个描述，$\boldsymbol{X}_i \in \mathbb{B}^{n \times p_i}$($p_i$ 为第 $i$ 个模态的维度). $\boldsymbol{Y}_i$ 是与 $\boldsymbol{X}_i$ 对应的标签矩阵. 对于每个对象，基于相应的跨媒体数据中某一模态是否存在，可以得到一个二值指示向量 $\boldsymbol{I}=[1 \cdots m]$，其中 $\boldsymbol{I}[i]=1$ 表示第 $i$ 模态的描述存在. 这些二值向量可以被完全转化为一个十进制整数(Profile). 所有这些十进制数被存储在一个 $n$ 维向量 $\boldsymbol{pf}$ 中. iSFS 方法通过以下优化模型，学习不同模态间的一致分类器 $\beta$ 和组合劝系数 $\alpha$：

$$\begin{aligned} \min_{\beta,\alpha} \quad & \frac{1}{|\boldsymbol{pf}|} \sum_{k \in \boldsymbol{pf}} f(X_k, \beta, \alpha_k, Y_k) + \lambda R_\beta(\beta) \\ \text{s.t.} \quad & R_\alpha(\alpha_k) \le 1 \text{ and } \forall k \in \boldsymbol{pf} \end{aligned} \tag{61}$$

其中，

$$f(X, \beta, \alpha, Y) = \frac{1}{n} L\left(\sum_{i=1}^{m} \alpha^i \boldsymbol{X}^i \beta^i, Y\right),$$

$R_\beta$ 和 $R_\alpha$ 是关于 $\beta$ 和 $\alpha$ 的正规化项，$k$ 表示受限于包含 $k$ 在自身 profile 里的样本的向量，$\boldsymbol{X}^i$ 表示第 $i$ 个模态的数据矩阵，$\beta^i$ 表示第 $i$ 个模态的分类器，$L$ 为凸损失函数(例如，最小二乘方损失函数或者逻辑损失函数). iSFS 方法的优点在于，能够在跨媒体数据出现逐块缺失的情况下，不必对缺失元素进行插值填补，就可完成多模态分析和特征选择. 然而，它却无法补全缺失的数据.

表 4 总结了上述跨媒体数据缺失补全方法的不足.

**表 4 已有的跨媒体数据缺失补全方法的不足**

| 方法 | 不足 |
|---|---|
| KNN | 若样本中存在噪声，就会大大降低矩阵补全的效果. |
| SVD | 如果存在大量缺失值，则无法对填补矩阵进行奇异值分解. |
| SSDR-MML | 只能填补缺失的标签，对于跨媒体数据的缺失描述却无能为力. |
| iSFS | 只能在存在缺失的情况下，完成多模态分析，却无法补全缺失的数据. |

### 3.3 跨媒体数据的噪声去除

由于人为错误等原因，在现实世界中，某些跨媒体数据受到噪声的腐蚀，并不是现实事物的真实反映. 因此，国内外的研究人员提出了一些多模态去噪方法消除跨媒体数据中的噪声.

在文献[97]中，Zhang 等人通过联合去除不同模态中的噪声，修复所有的针孔成像图片. 他们提出了一种多模态图片去噪(Depth-Guided Multi-View Image Denoising，DGMVID)方法，借助于主成分分析(Principal CompOnent Analysis，PCA)思想，利用多模态图片间的深度估计(Depth Estimation)去除多模态数据中的噪声. Zhang 等人发现将多模态图片用于噪声去除具有独特的优势：一张图片和其它图片的像素间的对应关系取决于其单一深度图. 这一点和视频去噪形成鲜明的对比. 在视频去噪中，帧间运动通常具有更大的自由度. 文献[97]正是利用这一

《计算机学报》录用稿

点，在多模态图片去噪中，寻找相似图片块（Similar Image Patches）. DGMVID 方法包含联合多模态图片匹配（Joint Multi-View Patch Matching，JMVPM）和联合多模态图片去噪（Joint Multi-View Patch Denoising，JMVPD）两个部分. 其中，JMVPM 借助多模态图片间的相关性，使用深度估计度量不同模态的图片中对应像素块的相似度，并根据相似度的大小，将相似图片块分组到一起. 而 JMVPD 为了去除信号相关噪声（Signal-Dependent Noise），利用 PCA 和张量分析（Tensor Analysis），在微光（Low-Light）条件下，消除光子噪声（Photon Noise）.

在音频去噪中，噪声往往夹杂在背景中，即没有其它可见模态的噪声数据. 由于干扰音频模态的噪声多为复合噪声源（例如，背景音乐），往往不稳定且无法直接观测，所以去除单一模态的音频数据的噪声是相当困难的. 针对这一问题，Segev 等人[98]提出了一种基于样本的跨模态去噪（Example-based Cross-Modal Denoising，ECMD）方法，通过不同模态间的关联性，借助无噪的视频模态信息，消除有噪的音频模态中的噪声. ECMD 方法假设“已经看到且听到的事物，如果能再一次见到该事物，那么就能再一次听清楚它”. 根据该假设，ECMD 方法利用先验知识的一致性（Consistency with Prior Knowledge），首先学习跨模态间的相关性. 然后，基于学习到的相关性和来自于无噪模态（视频）的无腐蚀训练样本，协助去除音频模态的噪声. 例如，任何智能手机都具有麦克风和摄像头. 由于无噪的视频数据相对更容易获得，所以若干无噪环境下的视频通话就可以建立一个无腐蚀的样本库. 而某些通话是在嘈杂环境下（例如，火车站、酒吧、商场）完成的. 于是，便可利用无噪的视频样本集去除音频数据的噪声.

文献[99]提出了一种概率深度诱导多模态去噪（Probabilistic Depth-guided Multi-view Denoising，PDMD）算法. 该方法将多模态图片去噪问题转化为有噪环境下深度估计的不确定性问题，利用非局部近邻的几何分布和块相似性，近似深度估计的概率，并基于这些概率平均非局部近邻，完成最小均方差（Minimum Mean Square Error，MMSE）去噪. PDMD 方法假设多模态图片集由 $M$ 个模态组成，并设 $x$ 为某一模态中的噪声块. 从每个模态中，选择 $K$ 个非局部近邻 $u_k(k=1,2,\cdots,K)$，以消除 $x$ 中的噪声成分. 平均这些非局部近邻，每个 $u_k$ 被一个加性噪声腐蚀：

$$u_k = v_k + \boldsymbol{n}_k \tag{62}$$

其中，$v_k$ 表示无噪成分；噪声向量 $\boldsymbol{n}_k$ 服从正态分布 $P(\boldsymbol{n}_k)\sim N(0,\sigma_n^2\boldsymbol{I})$. 而 $x$ 由式(62)中的无噪成分 $v_k$ 和加性噪声 $n_x$ 构成：

$$x = v_k + \boldsymbol{n}_x \tag{63}$$

其中，噪声向量 $\boldsymbol{n}_x$ 服从正态分布 $P(\boldsymbol{n}_x)\sim N(0,\sigma_n^2 I)$. PDMD 方法利用概率深度估计，学习不同模态间的相关性，以估计 $x$ 的无噪近似 $\hat{x}$. 给定噪声块 $x$，通过排斥所有可能的深度候选 $z_m$，第 $m$ 个模态中第 $k$ 个非局部近邻 $v_{m,k}$ 的条件概率为 $P(v_{m,k}|x)$：

$$\begin{aligned} P(v_{m,k}|x) &\cong \sum_{z_m\in\mathcal{Z}_m} P(u_{m,k},z_m|x) \\ &= \sum_{z_m\in\mathcal{Z}_m} P(u_{m,k}|z_m,x)P(z_m|x) \end{aligned} \tag{64}$$

其中，$\mathcal{Z}_m$ 为可能的深度候选集. 对于均值为 0 的加性噪声 $E[u_{m,k}]=E[v_{m,k}]$，利用 $m$ 个模态的数据，无噪块的最小均方差估计为

$$\hat{x}_m = E(v_{m,k} \mid x) = \sum_{k=1}^{K} u_{m,k}P(v_{m,k} \mid x) \tag{65}$$

而无噪估计 $\hat{x}$ 可以通过平均所有模态的 $\hat{x}_m$ 而得

$$\hat{x}=\frac{1}{M}\sum_{m=1}^{1}\sum_{k=1}^{1}\sum_{z_m\in\mathcal{Z}_m} u_{m,k}P(u_{m,k}|z_m)P(z_m|x) \tag{66}$$

其中，$P(z_m|x)$ 表示给定 $x$ 的深度 $z_m$ 的概率，而 $P(u_{m,k}|z_m,x)$ 可以看作给定深度 $z_m$，$u_{m,k}$ 和 $x$ 出自同一无噪块的概率.

上述这些方法的缺陷在于，它们都只能用来处理特定模态的噪声（例如，图片或者语音），而无法解决其它形式的跨媒体数据中的噪声问题.

近来，在机器学习和数据挖掘领域，知识迁移（Knowledge Transfer）技术越来越受到国内外研究人员的青睐. 在文献[100]中，Quanz 等人便利用知识迁移技术，提出了一种基于稀疏编码（Sparse Coding）的特征提取（Knowledge Transfer with Low-Quality Data，KTLQD）方法，借助高质量的辅助有标签数据源的信息，预测主源无标签数据的类别，并提高主源低质数据的性能. 目前，稀疏编码已广泛应用在高维数据处理上，以清除高维数据中的噪声. 假设有 $n$ 个 $p$ 维数据点 $\{x_1,x_2,\cdots,x_n\}\subset\mathbb{R}^p$，从而构成 $p\times n$ 阶的数据矩阵 $\boldsymbol{X}$. 稀疏编码的目的就是学习 $r$ 个 $p$ 维基向量 $\{b_1,b_2,\cdots,b_r\}\subset\mathbb{R}^p$ 形成 $p\times r$ 阶的基向量矩阵 $\boldsymbol{B}$ 和 $n$ 个 $r$ 维稀疏权重向量 $\{w_1,w_2,\cdots,w_n\}\subset\mathbb{R}^r$ 形成 $r\times n$ 阶的权重矩阵 $\boldsymbol{W}$. 利用 $\boldsymbol{B}$ 和 $\boldsymbol{W}$，便可近似原始数据 $X$. KTLQD 方法将主源数据嵌入到低维空间中，并利用辅助源高质量数

《计算机学报》

据的数据分布，诱导主源数据的分布. 为了达到这一目的，KTLQD 方法在优化模型中添加了针对主源数据集 $U$ 和辅助源数据集 $V$ 的嵌入权重的可调正则化项，调整数据的分布距离. KTLQD 方法的优化模型如下：

$$\min_{\boldsymbol{B},\boldsymbol{W}} \quad \| \boldsymbol{X}-\boldsymbol{B}\boldsymbol{W} \|_F^2 + \alpha \sum_{i=1}^{n} \| w_i \|_1 + d(p_U, q_V)$$

$$\text{s.t.} \quad \| b_i \|_2^2 \le c, \ \forall i = 1,2,\cdots,n \tag{67}$$

其中，$p$ 和 $q$ 为预先指定的概率密度函数；$d(\cdot)$ 是预先指定的分布距离函数. 在文献[100]中，Quanz 等人利用交替优化求解式(67)中的问题. 然而，KTLQD 方法的缺陷在于，它无法处理异构数据.

表 5 总结了上述跨媒体数据噪声去除方法的不足.

**表 5 已有的跨媒体数据噪声去除方法的不足**

| 方法 | 不足 |
|---|---|
| DGMVID | 只能用来处理图片模态的噪声，而无法解决其它形式的跨媒体数据中的噪声问题. |
| ECMD | 只能用来处理音频模态的噪声，而无法解决其它形式的跨媒体数据中的噪声问题. |
| PDMD | 只能用来处理图片模态的噪声，而无法解决其它形式的跨媒体数据中的噪声问题. |
| KTLQD | 无法处理异构数据. |

### 3.4 跨媒体数据的维度缩减

近年来，随着跨媒体数据的不断增加和现代存储技术的快速发展，跨媒体数据中不可避免地出现大量冗余. 于是，国内外的研究人员设计了各种各样的机器学习方法降低跨媒体数据的维度，去除跨媒体数据中的冗余特征，节省存储空间和计算时间.

Christoudias 等人[101]提出了一种基于分布编码的无监督多模态图片特征选择(Joint Feature Histogram Model，JFHM)算法. 该方法通过统计不同模态信息的高斯处理模型过滤不同模态数据中的冗余信息，并在接收端获取多模态数据的联合编码，以降低数据维度，提高对象识别准确率. JFHM 方法假设 $Z^v(v=1,2,\cdots,V)$ 为每个模态的编码特征集. 为了便于特征流的联合解码，解码器首先利用全局特征词典(Global Feature Vocabulary)，计算特征直方图(Feature Histogram) $h^v = h(Z^v)$. 在 JFHM 算法中，函数 $h(\cdot)$ 的形式为平型(Flat)直方图，或分层(Hierarchical)直方图. 而 Christoudias 等人提出的分布编码方法适合于任何词包(Bag-of-Words)技术. 在编码器中，联合流统计由以下特征直方图表示：

$$p(h^1, h^2, \cdots, h^V) = p(h^1) \prod_{v=2}^{V} p(h^v \mid h^{v-1}, \cdots, h^1) \tag{68}$$

其中，条件概率由训练数据学习得来. 假设直方图的不同箱(Bin)间独立，而直方图间也逐对独立，式(68)中的条件概率便可定义为

$$p(h^v \mid h^{v-1}, \cdots, h^1) = \prod_{k=1}^{v-1} \prod_{b=1}^{B} p(h^{v,b} \mid h^k),$$

其中，$h^{v,b}$ 是直方图 $h^v$ 中的第 $b$ 个箱，而 $B$ 为箱的数量. 式(68)中的模型可以判别给定模态中哪些特征是冗余的，也就是说，冗余特征是那些和当前模态中冗余箱相关的特征. JFHM 算法正是基于这些特征直方图，消除冗余特征，提取判别特征. 然而，JFHM 算法只能用来处理多模态图片冗余.

在文献[102]中，Zhu 等人提出了一种多模态典型特征选择(Multi-modality Canonical Feature Selection，MCFS)方法，利用不同模态之间的相关性，将不同模态的特征投影到由 CCA 诱导的典型空间中，从而实现多模态典型特征选择. MCFS 方法将不同模态间的相关信息，整合到稀疏多任务学习(SparseMulti-Task Learning)中. MCFS 方法首先利用 CCA 获取典型特征空间的典型基向量，再将不同模态的异构描述嵌入到这个空间中，并利用带有典型正规化项的稀疏多任务学习筛选典型特征. 假设 $\boldsymbol{X}^{(1)} \in \mathbb{R}^{d\times n}$ 和 $\boldsymbol{X}^{(2)} \in \mathbb{R}^{d\times n}$ 是来自两个不同模态的样本构成的数据矩阵($d$ 为样本维度，$n$ 为样本数量). $\boldsymbol{Z}^{(1)} \in \mathbb{R}^{d\times n}$ 和 $\boldsymbol{Z}^{(2)} \in \mathbb{R}^{d\times n}$ 是 $\boldsymbol{X}^{(1)}$ 和 $\boldsymbol{X}^{(2)}$ 由 CCA 产生的典型描述. 设 $\boldsymbol{Y} \in \mathbb{R}^{c\times n}$ 为标签矩阵，$c$ 为类的数量. 为了选择最相关的典型特征，MCFS 方法将稀疏多类线性回归整合到多任务学习框架中，利用 $\ell_{2,1}$ 范数筛选特征，从而得到如下优化模型：

$$\min_{\boldsymbol{W}} \ \frac{1}{2} \| \boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}} \boldsymbol{Z} \|_F^2 + \beta \| \boldsymbol{W} \|_{2,1} + \gamma \| \boldsymbol{W} \|_{\mathrm{CCA}}^2 \tag{69}$$

其中，$\boldsymbol{Z} = [Z^{(1)}; Z^{(2)}] \in \mathbb{R}^{2d\times n}$；$\boldsymbol{W} \in \mathbb{R}^{2d\times c}$ 为回归系数矩阵；参数 $\beta$ 控制 $\boldsymbol{W}$ 的行稀疏度；$\gamma$ 为平衡参数；$\| \boldsymbol{W} \|_{\mathrm{CCA}}$ 为权重系数矩阵 $\boldsymbol{W}$ 的典型正规化项，具体形式如下：

$$\begin{aligned} \| \boldsymbol{W} \|_{\mathrm{CCA}}^2 &= \sum_{i=1}^{2d} \| w^i \|_{\mathrm{CCA}}^2 \\ &= \sum_{i=1}^{d} \frac{1-\lambda_i}{\lambda_i} \sum_{j=1}^{c} \{ (w_{i,j})^2 + (w_{i+d,j})^2 \}, \end{aligned}$$

其中，$w^i$ 为 $\boldsymbol{W}$ 的第 $i$ 行，$\{\lambda_i\}_{i=1:d}$ 为一组典型相关系数. 在式(69)中添加典型正规化项的目的是保持不同模态描述间的相关性. 然而，MCFS 方法本身无法处理异构数据，必需借助 CCA 方法，才能进行关联

《计算机学报》版权

分析，消除冗余.

而文献[103]提出了一种降低无标签样本的复杂度的半监督多模态学习(Reducing the Unlabeled Sample Complexity of Semi-Supervised Multi-view Learning，RUSCSSML)方法. 在半监督学习中，同时利用有标签样本和无标签样本训练分类器. 而样本复杂度(Sample Complexity)是训练样本有效性的通用度量标准. Lan 和 Huan 将这两类样本的复杂度定义为有标签样本复杂度(Labeled Sample Complexity，LSC)和无标签样本复杂度(Unlabeled Sample Complexity，USC)两类. 在松弛条件下，RUSCSSML 方法将 USC 从 $O(1/\varepsilon)$ 降低到 $O(\log(1/\varepsilon))$($\varepsilon$ 为错误率). Lan 和 Huan 在理论上证明了分类器的泛化错误率和不兼容性之间的联系，并证明了在半监督多模态学习中，具备大量无标签样本，就可以学习到一个不兼容性较低的分类器. 通过结合这两项证明结果，Lan 和 Huan 进一步证明了半监督多模态学习的可能近似正确(Probably Approximately Correct，PAC)的学习范围. 然而，上述方法却无法得到不同模态间的共享描述.

Guo 等人[104]提出了一种用于跨语言文本分类的低维子空间协同正规化多模态学习(Subspace Co-regularized Multi-view Learning，SCML)方法. 该方法基于机器翻译的平行语料库，联合最小化每个模态上的分类器的训练误差，并同时最小化低维子空间中描述间的距离. Guo 等人假设一篇文档(源语言(Source Language))和其翻译文本(目标语言(Target Language))以两种不同的模态描述了相同的对象. 于是，对于相同的分类任务，同一对象的不同模态描述在潜在共享子空间中应该是相似的. 因此，SCML 方法针对不同的语言(模态)，同时学习不同的分类器，并通过半监督优化模型，最小化所有模态的有标签数据的训练损失，惩罚所有对象异构描述在子空间中的距离. 假设源语言具有 $n_s$ 个样本(其中，$n_s=l_s+u_s$，$l_s$ 个有标签样本，$u_s$ 个无标签样本.)，而目标语言具有 $n_t$ 个样本(其中，$n_t=l_t+u_t$，$l_t$ 个有标签样本，$u_t$ 个无标签样本.). 利用机器翻译，可以将源语言的文档翻译为目标语言的平行文档，反之亦然. 从而可以得到两个平行数据矩阵 $\boldsymbol{X}_1\in\mathbb{R}^{n\times d_1}$(源模态)和 $\boldsymbol{X}_2\in\mathbb{R}^{n\times d_2}$(目标模态)，其中 $n=n_s+n_t$，$d_1$ 为源模态的维度，$d_2$ 为目标模态的维度. $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 中的前 $l=l_s+l_t$ 行分别构成有标签子矩阵 $\boldsymbol{X}_1^l$ 和 $\boldsymbol{X}_2^l$，而 $y\in\{-1,+1\}^l$ 为相应的标签向量. SCML 方法针对每个模态学习一个线性预测函数：

$$f_i(\boldsymbol{X}_i^l)=\boldsymbol{X}_i^l\boldsymbol{\Theta}_i\boldsymbol{w}_i+\boldsymbol{b}_i \tag{70}$$

其中，$\boldsymbol{w}_i\in\mathbb{R}^m$ 为线性权重向量；$\boldsymbol{b}_i\in\mathbb{R}$ 为偏置参数；$\boldsymbol{\Theta}_i\in\mathbb{R}^{d_i\times m}$ 为将数据投影到低维子空间的线性变换矩阵，$m$ 为子空间的维度，并对 $\boldsymbol{\Theta}_i$ 加以正交约束 $\boldsymbol{\Theta}_i^{\mathrm{T}}\boldsymbol{\Theta}_i=\boldsymbol{I}$($\boldsymbol{I}$ 为单位矩阵). 由于不同的模态共享相同的分类任务，不同模态的平行文档的各自预测子空间描述应该是相似的. 根据这一原理，SCML 方法将跨语言文本分类问题塑造为一个半监督多模态优化问题，以最小化不同模态的有标签样本的训练损失，并惩罚不同模态的样本在子空间中的距离. SCML 方法的优化模型如下：

$$\begin{aligned}\min_{\{\theta_i,w_i,b_i\}}\quad & \sum_{i=1}^{2}\|X_i^l\theta_i w_i+b_i-y\|^2+a_i\|w_i\|^2+\\ & \gamma\|X_1\theta_1-X_2\theta_2\|_F^2\\ \text{s.t.}\quad & \boldsymbol{\theta}_i^{\mathrm{T}}\theta_i=\boldsymbol{I}\ \text{and}\ \boldsymbol{\theta}_2^{\mathrm{T}}\theta_2=\boldsymbol{I}\end{aligned} \tag{71}$$

但是，SCML 方法在降维过程中，只考虑了不同模态间的相关性，而没有充分利用不同模态间的分布相似性，所以必然会丢失一些重要的信息.

表 6 总结了上述跨媒体数据冗余缩减方法的不足.

**表 6　已有的跨媒体数据冗余缩减方法的不足**

| 方法 | 不足 |
|---|---|
| JFHM | 只能用来处理多模态图片冗余. |
| MCFS | 无法处理异构数据，必需借助 CCA 方法，才能进行关联分析，消除冗余. |
| RUSCSSML | 无法得到不同模态间的共享描述. |
| SCML | 只考虑了不同模态间的相关性，而没有充分利用不同模态间的分布相似性，所以必然会丢失一些重要的信息. |

### 3.5　跨媒体数据的样本选择

为了进一步节省存储空间，提高学习效率和性能，国内外的研究人员相继提出了一些跨媒体数据的样本选择方法，挑选精简样本子集.

多模态图片是一组由多个摄像机在同一场景下拍摄的图片. 不同模态的图片之间必然存在互补信息. 在多模态立体声(Multi-View Stereo，MVS)的重建中，并不是所有的图片都能提高重建模型输出结果的质量，而且大量的多模态图片还会耗费过多的处理时间. 针对这一问题，Hornung 等人[105]设计出一种多模态图片选择(Image Selection for Improved Multi-View Stereo，ISIMVS)方案，利用预先定义的质量标准，根据多模态图片中的互补信息，筛选相关性最大的图片合成重建结果. ISIMVS 方法通过以下 3 个预先定义的质量标准(Criteria)完成 MVS 重建：(1) 初始表面精简子集(Initial Surface Proxy)，

《计算机学报》版权

即选择一组输入图片的精简子集，该子集不仅能够充分代表原始数据集，并能充分逼近未知对象的表面；(2)表面可见度(Surface Visibility)，即在设定的最小可视距离内，挑选的精简子集中的多模态样本是清晰可见的；(3)自适应性(Adaptivity)，即针对多模态图片中的不一致区域，通过选择一些其它的相关图片，以提高在这些区域中的重建性能，增加精选子集的可靠性. 根据上述 3 个质量标准，ISIMVS 方法的样本选择过程可分为 3 步：第 1 步，从若干模态中，挑选一些有利于快速收敛的模态；第 2 步，为每个对象挑选至少两个模态的图片，以达到足够的覆盖率；第 3 步，挑选一些突出多模态图片中不一致区域的额外的相关图片，进而达到良好的 MVS 重建效果. 但是，ISIMVS 方法存在只能精简多模态图片的缺陷.

文献[106]提出的多模态视频编码(Multi-view Video Coding using View Interpolation and Reference Picture Selection，MVCVIRPS)方法. MVCVIRPS 方法基于 H. 264/AVC 的动作/不一致性补偿完成多模态视频编码. 该方法利用不同模态图片之间的关联性，通过选择相关的异模图片作为参考图片，插值补偿视差. MVCVIRPS 方法假设索引 $c=1,2,\cdots,C$ 对应 $C$ 台不同的摄像机(基于 H. 264/AVC 编码)，而另一索引 $C_c\subset\{1,2,\cdots,C\}$ 表示摄像机 $c$ 用于不一致性补偿的参考摄像机的索引. 在视频编码过程中，MVCVIRPS 方法同时利用不同模态间的时间(temporal)、空间(spacial)和模间(inter-view)相关性，从摄像机 $c$ 的参考摄像机的索引中选择相关索引，完成异模间动作/不一致性补偿.

在图像提取(Image Matting)中，利用基于颜色采样的抠图方法获取的颜色信息，可以挑选出针对未知像素的前景(Foreground(F))和背景(Background(B))颜色的最匹配样本. 然而，如果前景和背景区域的颜色分布有重叠，那么利用颜色信息就很难区分这些区域，挑选的样本便无法估计遮片(Matte). Shahrian 等人[107]提出了一种基于内容的样本选择(Weighted Color and Texture Sample Selection for Image Matting，WCTSSIM)方法，该方法借助图片的纹理(模态)信息，使用基于样本的颜色(模态)匹配方法，通过最佳匹配图片填补未知锁的前景和背景颜色. 在 WCTSSIM 方法中，利用图片的颜色(Color)和纹理(Texture)特征挑选一组 $(F,B)$ 候选集. 并根据两种特征，基于图片的内容，自动确定 $F$ 和 $B$ 间的权重，进而从候选集中挑选出最佳样本.

然而，虽然上述方法在精简多模态图片数据集问题上，都取得了不错的效果，但是对于其它形式的多模态数据(例如，文本)的冗余却捉襟见肘.

表 7 总结了上述跨媒体数据样本选择方法的不足.

**表 7 已有的跨媒体数据样本选择方法的不足**

| 方法 | 不足 |
|---|---|
| ISIMVS | 只能用来精简多模态图片数据集. |
| MVCVIRPS | 只能用来精简多模态图片数据集. |
| WCTSSIM | 只能用来精简多模态图片数据集. |

### 3.6 小 结

综上所述，虽然目前国内外的研究人员已经提出了一些针对跨媒体数据中增量学习、缺失补全、噪声去除以及冗余缩减等问题的解决方案，但绝大多数方法都必须和已有的经典算法(如 CCA 和 IRRT 等)相结合，才能处理跨媒体数据的异构性. 此外，已有的方法大多只能用来处理固定形式的多模态数据(例如，多模态图片)，对于其它形式的跨媒体数据则束手无策.

因此，除跨媒体数据的一致性描述外，针对跨媒体数据的研究还需要进一步解决增量学习、缺失补全、噪声去除以及冗余缩减等问题. 然而，目前相关领域的研究仍然较为初步，其研究潜力巨大，很多问题还有待进一步的解决.

## 4 结论及未来展望

本文主要介绍了跨媒体学习的研究背景和应用价值，概括介绍了各类跨媒体学习相关方法的数学原理和基本特性. 并重点介绍了跨媒体共享子空间学习的研究进展，比较了基于投影、矩阵分解、任务和度量等四大类子空间学习方法的优缺点，分析了未来的发展方向.

不同于单媒体学习，跨媒体学习还是一个崭新的研究方向. 目前相关领域的研究仍然较为初步. 跨媒体学习的研究潜力巨大，很多方面还有待进一步的研究. 传统方法虽然在相关性提取上做出了一定的贡献，但在现实情况下，跨媒体数据中必然存在冗余、噪声和缺失. 如何在这些情况下挖掘跨媒体数据的相关性，仍然是目前跨媒体学习领域亟待解决的问题.

**致 谢** 在此，对各位评审专家表示衷心的感谢！

《计算机学报》版权

## 参　考　文　献


[1] Van Breukelen M, Duin Robert P W, Tax David M J, Den Hartog J E. Handwritten Digit Recognition by Combined Classifiers, Kybernetika, 1998, 34(4): 381-386

[2] Guillaumin M, Verbeek J, Schmid C. Multiple instance metric learning from automatically labeled bags of faces//Proceedings of the European Conference on Computer Vision. Crete, Greece, 2010: 634-647

[3] Rasiwasia N, Costa Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval//Proceedings of the ACM International Conference on Multimedia. Firenze, Italy, 2010: 251-260

[4] Liu M H, Zhang D Q, Shen D G. Ensemble sparse classification of Alzheimer's disease. NeuroImage, 2012, 60(2): 1106-1116

[5] Fukunaga K. Introduction to Statistical Pattern Recognition. San Diego, USA: Academic Press Professional, Inc., 1991

[6] Candès E J, Li X D, Ma Y, Wright J. Robust principal component analysis. Journal of the ACM, 2011, 58(3): 1-78

[7] Lu Z K, Chi Z R, Siu W C. Extraction and optimization of B-spline PBD templates for recognition of connected handwritten digit strings. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(1): 132-139

[8] Boutell M R, Luo J B, Shen X P, Brown C M. Learning multi-label scene classification. Pattern recognition, 2004, 37(9): 1757-1771

[9] Yang Y, Nie F P, Xu D, et al. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 723-742

[10] Tian Chun-Na. Multi-View Face Detection and Recognition in Complex Background [Ph. D. dissertation]. Xidian University, Xi'an, 2008(in Chinese)
(田春娜. 复杂背景下多视角人脸检测与识别[博士学位论文]. 西安电子科技大学，西安，2008)

[11] Zhang Hong. Correlation Mining Based Cross-Media Retrieval [Ph. D. dissertation]. Zhejiang University, Hangzhou, 2007 (in Chinese)
(张鸿. 基于相关性挖掘的跨媒体检索研究[博士学位论文]. 浙江大学，杭州，2007)

[12] Liu Hong. Study on Profitability of Commercial Bank of China Based on the Perspectives[Ph. D. dissertation]. Jilin University, Changchun, 2010(in Chinese)
(刘宏. 基于多视角的我国商业银行盈利能力研究[博士学位论文]. 吉林大学，长春，2010)

[13] Feng Wei. Study on Distributed Multi-View Target Tracking Algorithm [Ph. D. dissertation]. Fudan University, Shanghai, 2011(in Chinese)
(冯巍. 分布式多视角目标跟踪算法研究[博士学位论文]. 复旦大学，上海，2011)

[14] Yang Y, Nie F P, Xiang S M, et al. Manifold learning based cross-media retrieval: A solution to media object complementary nature. The Journal of VLSI Signal Processing, 2007, 46(2): 153-164

[15] He Ning. Cross-Modal Semantic Information Acquisition for Image Retrieval [Ph. D. dissertation]. Wuhan University, Wuhan, 2013(in Chinese)
(何宁. 图像检索中跨模语义信息获取方法研究[博士学位论文]. 武汉大学，武汉，2013)

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

[17] Hardoon D R, Szedmák S, Taylor J S. Canonical correlation analysis: An overview with application to learning methods. Neural Computation, 2004, 16(12): 2639-2664

[18] Chaudhuri K, Kakade S M, Livescu K, et al. Multi-view clustering via canonical correlation analysis//Proceedings of the ACM International Conference on Machine Learning. Montréal, Canada, 2009: 129-136

[19] Gao J, Han J W, Liu J L, Wang C. Multi-view clustering via joint nonnegative matrix factorization//Proceedings of the SIAM International Conference on Data Mining. Austin, USA, 2013: 252-260

[20] Cai X, Nie F P, Huang H. Multi-view *K*-means clustering on big data//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2598-2604

[21] Eaton E, Desjardins M, Jacob S. Multi-view constrained clustering with an incomplete mapping between views. Knowledge and Information Systems, 2014, 38(1): 231-257

[22] Sharma A, Kumar A, Daumé III H, Jacobs D W. Generalized multiview analysis: A discriminative latent space//Proceedings of the International Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 2160-2167

[23] Su H, Sun M, Li F F, Savarese S. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories//Proceedings of the International Conference Computer Vision. Kyoto, Japan, 2009: 213-220

[24] Zhang Z X, Zhao Y H, Wang Y H, Liu J Y, Yao Z J, Tang J. Transferring training instances for convenient cross-view object classification in surveillance. IEEE Transactions on Information Forensics and Security, 2013, 8(10): 1632-1641

[25] Koço S, Capponi C. A boosting approach to multiview classification with cooperation//Proceedings of the European Conference on Machine Learning and European Conference on Principles and Practice of Knowledge Discovery. Athens, Greece, 2011: 209-228

[26] Sun L, Ji S W, Ye J P. Canonical correlation analysis for multi-label classification: A least-squares formulation, extensions, and analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 194-200


计算机学报版权

[27] Lee D D, Seung H S. Algorithms for non-negative matrix factorization//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2001, 32(6): 556-562

[28] Dorai C, Venkatesh S. Computational media aestheties: Finding meaning beautiful. IEEE Multimedia, 2001, 8(4): 10-12

[29] Zhao R, Grosky W I. Negotiating the semantic gap: From feature maps to semantic landscapes. Pattern Recognition, 2002, 35(3): 593-600

[30] Golub G H, Van Loan C F. Matrix Computations. Baltimore, USA: The Johns Hopkins University Press, 1996

[31] Blum A, Mitchell T M. Combining labeled and unlabeled data with co-training//Proceedings of the ACM International Conference on Computational Learning Theory. Madison, USA, 1998: 92-100

[32] Gu Ping. Research on Text Classification with Bayesian Model and Relational Technologies[Ph. D. dissertation]. Chongqing University, Chongqing, 2006(in Chinese)
(古平. 基于贝叶斯模型的文档分类及相关技术研究[博士学位论文]. 重庆大学, 重庆, 2006)

[33] Huang Dong-Shan. Research on Feature Selection and Semi-Supervised Classification [Ph. D. dissertation]. Huazhong University of Science and Technology, Wuhan, 2011 (in Chinese)
(黄东山. 特征选择及半监督分类方法研究[博士学位论文]. 华中科技大学, 武汉, 2011)

[34] Zhang Bo-Feng. Research on Text Categorization Method Oriented to Content Security [Ph. D. dissertation]. National University of Defense Technology, Changsha, 2007 (in Chinese)
(张博锋. 面向内容安全的文本分类研究[博士学位论文]. 国防科学技术大学, 长沙, 2007)

[35] Zhu Jian-Zhang. Studies on Real-Time Visual Object Tracking in Complex Environment [Ph. D. dissertation]. Wuhan University, Wuhan, 2014(in Chinese)
(朱建章. 复杂场景下实时视觉目标跟踪的若干研究[博士学位论文]. 武汉大学, 武汉, 2014)

[36] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541

[37] Gönen M, Alpaydin E. Multiple kernel learning algorithms. Journal of Machine Learning Research, 2011, 12: 2211-2268

[38] Liu Lei. Doppler Ultrasound—Based Pulse Signal Analysis and Classification Research [Ph. D. Dissertation]. Harbin Institute of Technology, Harbin, 2013(in Chinese)
(刘磊. 基于多普勒超声信号的脉象分析与分类研究[博士学位论文]. 哈尔滨工业大学, 哈尔滨, 2013)

[39] Li Xiao-Dong. Kernel ELM Theory and Algorithms and Application in Image Processing [Ph. D. dissertation]. Zhejiang University, Hangzhou:, 2014(in Chinese)
(李小冬. 核极限学习机的理论与算法及其在图像处理中的应用[博士学位论文]. 浙江大学, 杭州, 2014)

[40] Zeng Pu. Research on Key Techniques of Image Classification for Semantic Extraction [Ph. D. dissertation]. National University of Defense Technology, Changsha, 2009 (in Chinese)
(曾璞. 面向语义提取的图像分类关键技术研究[博士学位论文]. 国防科学技术大学, 长沙, 2009)

[41] Shao Xi-Gao. The Research and Application of Multiple Kernel Prediction Model Based on Statistical Learning Theory [Ph. D. dissertation]. Central South University, Changsha, 2013(in Chinese)
(邵喜高. 基于统计学习理论的多核预测模型研究及应用[博士学位论文]. 中南大学, 长沙, 2013)

[42] Xia H, Wu P C, Hoi S C H, Jin R. Boosting multi-kernel locality- sensitive hashing for scalable image retrieval//Proceedings of the International Conference on Research and Development in Information Retrieval. Portland, USA, 2012: 55-64

[43] Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B. Large scale multiple kernel learning. Journal of Machine Learning Research, 2006, 7: 1531-1565

[44] Yuan Gan-Zhao. Low-Rank Matrix Approximation Algorithm and Its Application [Ph. D. dissertation]. South China University of Technology, Guangzhou, 2013(in Chinese)
(袁淦钊. 矩阵的低秩近似算法及其应用[博士学位论文]. 华南理工大学, 广州, 2013)

[45] Jian Ling. Kernel Based Learning Algorithm and Application [Ph. D. dissertation]. Dalian University of Technology, Dalian, 2012(in Chinese)
(渐令. 基于核的学习算法与应用[博士学位论文]. 大连理工大学, 大连, 2012)

[46] Xu Z L, Jin R, Yang H Q, et al. Simple and efficient multiple kernel learning by group lasso//Proceedings of the International Conference on Machine Learning. Haifa, Israel, 2010: 1175-1182

[47] Gönen M, Alpaydin E. Localized multiple kernel learning//Proceedings of the International Conference on Machine Learning. Helsinki, Finland, 2008: 352-359

[48] Lewis D P, Jebara T, Noble W S. Nonstationary kernel combination//Proceedings of the International Conference on Machine Learning. Pittsburgh, USA, 2006: 553-560

[49] Kloft M, Brefeld U, Sonnenburg S, Zien A. ℓp-norm multiple kernel learning. Journal of Machine Learning Research, 2011, 12: 953-997

[50] Sommer S, Lauze F, Nielsen M, Pennec X. Sparse multiscale diffeomorphic registration: The kernel bundle framework. Journal of Mathematical Imaging and Vision, 2013, 46(3):292-308

[51] Gehler P V, Nowozin S. Infinite Kernel Learning. Neural Information Processing Systems (NIPS) Workshop on Auto-

《计算机学报》版权

matic Selection of Kernel Parameters, Vancouver, British Columbia, Canada, MPI Technical Report 178, 2008

[52] Wold H. Partial Least Squares. Encyclopedia of Statistical Sciences, Hoboken, USA: Wiley Online Library, 2006

[53] Andrew G, Arora R, Bilmes J, Livescu K. Deep canonical correlation analysis//Proceedings of the ACM International Conference on Machine Learning. Atlanta, USA, 2013: 1247-1255

[54] Gupta S K, Phung D Q, Adams B, et al. Nonnegative Shared Subspace Learning and its application to social media retrieval//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 1169-1178

[55] Yu S P, Yu K, Tresp V, Kriegel H P. Multi-output regularized feature projection. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(12): 1600-1613

[56] Ando R K, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 2005, 6: 1817-1853

[57] Argyriou A, Evgeniou T, Ponti M. Convex multitask feature learning. Machine Learning, 2008, 73(3): 243-272

[58] Kong X, Ng M K, Zhou Z H. Transductive multi-label learning via label set propagation. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(3): 704-719

[59] Ji S W, Tang L, Yu S P, Ye J P. Extracting shared subspace for multi-label classification//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 381-389

[60] Ji S W, Tang L, Yu S P, Ye J P. A shared-subspace learning framework for multi-label classification. ACM Transactions on Knowledge Discovery from Data, 2010, 4(2): 890-895

[61] Amit Y, Fink M, Srebro N, Ullman S. Uncovering shared structures in multiclass classification//Proceedings of the International Conference on Machine Learning. Corvalis, USA, 2007: 17-24

[62] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 2009, 10(1): 207-244

[63] Xie P T, Xing E P. Multi-modal distance metric learning//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 1806-1812

[64] Malave V L, Talbott W. A Shared Subspace for Multiple Metric Learning. San Diego, CA, USA: Charlotte. ucsd. edu, 2005

[65] Xu Jie. Feature Extraction and Classification Based on Subspace Analysis and Its Applications[Ph. D. dissertation]. Nanjing University of Science and Technology, Nanjing, 2012 (in Chinese)
(徐洁. 基于子空间分析的特征抽取及分类方法研究[博士学位论文]. 南京理工大学，南京，2012)

[66] Sun Tong-Feng. Image Fusion and Recognition Based on Multi-scale Analysis and Subspace Learning [Ph. D. dissertation]. China University of Mining and Technology, Xuzhou, 2012(in Chinese)
(孙统风. 基于多尺度与子空间的图像融合和识别研究[博士学位论文]. 中国矿业大学，徐州，2012)

[67] Li Da-Rui. Large-Scale Visual Pattern Learning for High-Performance Image Representation [Ph. D. dissertation]. University of Science and Technology of China, Hefei, 2014 (in Chinese)
(李大瑞. 基于大规模视觉模式学习的高性能图像表示[博士学位论文]. 中国科学技术大学，合肥，2014)

[68] Liu Bo. Large Margin of Group Sparse Subspace for Feature Selection [Ph. D. dissertation]. Chongqing University, Chongqing, 2013(in Chinese)
(刘波. 组稀疏子空间的大间隔特征选择[博士学位论文]. 重庆大学，重庆，2013)

[69] Wen Jing. Visual Tracking Based on Tensor Subspace Learning [Ph. D. dissertation]. Xidian University, Xi'an, 2010(in Chinese)
(温静. 基于张量子空间学习的视觉跟踪方法研究[博士学位论文]. 西安电子科技大学，西安，2010)

[70] Zhang L, Zhao Y, Zhu Z F, et al. Mining semantically consistent pattern for cross-view data. IEEE Transactions on Knowledge and Data Engineering, 2014, 11(26): 2745-2758

[71] Yang F L, Li B M. Unsupervised learning of spatial structures shared among images. The Visual Computer, 2012, 28(2): 175-180

[72] Kusakunniran W, Wu Q, Zhang J, Li H D. Support vector regression for multi-view gait recognition based on local motion feature selection//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 974-981

[73] Wang Hui-Wen. Partial Least Squares Regression—Linear and Nonlinear Methods. Beijing: National Defence Industry Press, 1999(in Chinese)
(王惠文. 偏最小二乘回归的线性与非线性方法. 北京：国防工业出版社，1999)

[74] Wang Yu, Zhou Zhi-Hua, Zhou Ao-Ying. Machine Learning and Its Applications. Beijing: Tsinghua University Press, 2006(in Chinese)
(王珏，周志华，周傲英. 机器学习及其应用. 北京：清华大学出版社，2006)

[75] Lay D C. Linear Algebra and Its Applications. Hoboken, NJ, USA: Wiley-Interscience, 2007

[76] Goldberg Y, Elhadad M. SplitSVM: Fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Columbus, USA, 2008: 237-240

[77] Chang Y W, Hsieh C J, Chang K W, et al. Training and testing low-degree polynomial data mappings via linear SVM.

《计算机学报》版权

Journal of Machine Learning Research, 2010, 11(11): 1471-1490

[78] Rauber T W, Berns K. Kernel multilayer perceptron//Proceedings of the SIBGRAPI Conference on Graphics, Patterns and Images. Alagoas, Brazil, 2011: 337-343

[79] Vinokourov A, Taylor J S, Cristianini N. Inferring a semantic representation of text via cross-language correlation analysis//Proceedings of the Advances of Neural Information Processing Systems. Vancouver, Canada, 2002: 91-98

[80] Salakhutdinov R, Hinton G. Deep boltzmann machines//Proceedings of the International Conference on Artificial Intelligence and Statistics. Clearwater Beach, USA, 2009: 448-455

[81] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507

[82] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527-1554

[83] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning//Proceedings of the International Conference on Machine Learning. Bellevue, USA, 2011: 689-696

[84] Srivastava N, Salakhutdinov R. Multimodal learning with deep boltzmann machines//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 689-696

[85] Cheng Q, Zhou H B, Cheng J. The fisher-markov selector: Fast selecting maximally separable feature subset for multi-class classification with applications to high-dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(6): 1217-1233

[86] Meyer C D. Matrix Analysis and Applied Linear Algebra. Philadelphia, USA: SIAM Publishers, 2000

[87] Zhang J T, Huan J. Inductive multi-task learning with multiple view data//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 543-551

[88] Zuo Z Q, Luo Y, Tao D C, Xu C. Multi-view multi-task feature extraction for Web image classification//Proceedings of the ACM International Conference on Multimedia. Orlando, USA, 2014: 1137-1140

[89] Tsoumakas G, Katakis I. Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13

[90] Aly M. Survey on multiclass classification methods. Neural Networks, 2005: 1-9

[91] Zha H Y, Simon H D. On updating problems in latent semantic indexing. SIAM Journal on Scientific Computing, 1999, 21(2): 782-791

[92] Zhang L, Zhao Y, Zhu Z F. Extracting shared subspace incrementally for multi-label image classification. The Visual Computer, 2014, 12(30): 1359-1371

[93] Cover T M, Hart P E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1967, 13(1): 21-27

[94] Muslea I, Minton S, Knoblock C A. Active+Semi-Supervised learning=Robust Multi-view learning//Proceedings of the ACM International Conference on Machine Learning. Sydney, Australia, 2002: 435-442

[95] Qian B Y, Wang X, Ye J P, Davidson I. A reconstruction error based framework for multi-label and multi-view learning. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3): 594-607

[96] Xiang S, Yuan L, Fan W, et al. Multi-source learning with block-wise missing data for Alzheimer's disease prediction//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 185-193

[97] Zhang L, Vaddadi S, Jin H L, Nayar S K. Multiple View image denoising//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1542-1549

[98] Segev D, Schechner Y Y, Elad M. Example-based cross-modal denoising//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 486-493

[99] Lee C, Kim C S, Lee S U. Probabilistic depth-guided multi-view image denoising//Proceedings of the IEEE International Conference on Image Processing. Melbourne, Australia, 2013: 905-908

[100] Quanz B, Huan J, Mishra M. Knowledge transfer with low-quality data: A feature extraction issue. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(10): 1789-1802

[101] Christoudias C M, Urtasun R, Darrell T. Unsupervised feature selection via distributed coding for multi-view object recognition//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Ancherage, AK, USA, 2008: 1-8

[102] Zhu X F, Suk H II, Shen D G. Multi-modality canonical feature selection for alzheimer's disease diagnosis. Springer Medical Image Computing and Computer-Assisted Intervention, 2014, 17(2): 162-169

[103] Lan C, Huan J. Reducing the unlabeled sample complexity of semi-supervised multi-view learning//Proceedings of the ACM SIGKDD International Conference Knowledge Discovery Data Mining. Sydney, Australia, 2015: 627-634

[104] Guo Y H, Xiao M. Cross language text classification via subspace co-regularized multi-view learning//Proceedings of the ACM International Conference on Machine Learning. Edinburgh, UK, 2012: 1615-1622

[105] Hornung A, Zeng B Y, Kobbelt L. Image selection for improved multi-view stereo//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8

《计算机学报》版权

[106] Kitahara M, Kimata H, Shimizu S, et al. Multi-view video coding using view interpolation and reference picture selection //Proceedings of the IEEE International Conference on Multimedia and Expo. Toronto, Canada, 2006: 97-100

[107] Shahrian E, Rajan D. Weighted color and texture sample selection for image matting. IEEE Transactions on Image Processing, 2013, 22(11): 4260-4270

**ZHANG Lei**, born in 1982, Ph. D., assistant researcher. His research interests include data mining and machine learning.

**ZHAO Yao**, born in 1967, Ph. D., professor, Ph. D. supervisor. His research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding.

**ZHU Zhen-Feng**, born in 1974, Ph.D., professor, Ph. D. supervisor. His research interests include image and video understanding, computer vision and machine learning.

**Background**

With the rapid development of information technology, there are many CMD (Cross-Media Data) in the real world. These CMD show characteristics of the heterogeneity of low-level features and the correlation of high-level semantics. Traditional Mono-Media Learning (MML) methods have not been able to adapt to the feature heterogeneity of CMD. Thus, the research on related problems in CML (Cross-Media Learning) have been One of the hot research topics in the field of digital media analysis recently. This paper mainly introduces the research background and application value of CML, and provides an overview of mathematical principle and specialties of various related methods in CML. Meanwhile, the advances in cross-media shared subspace learning are presented. Furthermore, the paper compares the advantages and disadvantages of four kinds of subspace learning methods based on projection, matrix decomposition, task and measurement, respectively. Finally, the future development of CML is analyzed.

This work was supported in part by the National Basic Research Program of China (No. 2012CB316401), the National Natural Science Foundation of China (Nos. 61532005, 61572068, 61271275, 61501457, 61601458, and 61602465), the Program for Changjiang Scholars and Innovative Research Team in University (No. IRT201206), and the Program for New Century Excellent Talents in University (No. 13-0661).

《计算机学报》版权