

基于典型相关分析的多视图学习方法综述

郭陈凤, 伍冬睿

(华中科技大学人工智能与自动化学院, 湖北 武汉 430074)

摘要: 多视图学习是将不同来源的特征子集加以融合的策略。典型相关分析是多视图学习中的重要方法, 旨在最大化不同视图之间的相关性。传统的典型相关分析仅能计算两个视图之间的线性相关性, 无法应用于包含多个视图或包含非线性相关性的数据集。此外, 如果将典型相关分析应用于有监督任务, 其作为一种无监督方法将导致标签信息的浪费。针对上述问题, 提出大量非线性的、针对多个视图的、有监督的基于典型相关分析的多视图学习方法。首先, 概述经典的基于典型相关分析的多视图方法; 然后介绍这些方法在模式识别、跨模态检索和分类, 以及多视图嵌入中的典型应用; 最后, 对基于典型相关分析的多视图学习方法面临的挑战和未来研究方向进行了总结和展望。

关键词: 典型相关分析; 多视图学习; 多模态检索; 多视图嵌入

中图分类号: TP181

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202206

A survey on canonical correlation analysis based multi-view learning

GUO Chenfeng, WU Dongrui

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: Multi-view learning (MVL) is a strategy for fusing data from different sources or subsets. Canonical correlation analysis (CCA) is very important in MVL, whose main idea is to maximize the correlation of different views. The traditional CCA can only calculate the linear correlation between two views. Moreover, it is unsupervised, and the label information is wasted in supervised learning tasks. Many nonlinear, supervised, or generalized extensions have been proposed to accommodate these limitations. Firstly, a comprehensive overview of representative CCA approaches was provided. Then their classical applications in pattern recognition, cross-modal retrieval and classification, and multi-view embedding were described. Finally, the challenges and future research directions of CCA-based MVL approaches were pointed out.

Key words: canonical correlation analysis, multi-view learning, multi-modal retrieval, multi-view embedding

0 引言

通过单一视图对事物进行描述是局部的、不准确的, 如同“盲人摸象”, 无法得知事物全貌。现实场景中包含多个描述视角的多视图数据不胜枚举^[1-3]。多视图数据如图1所示, 被拍摄物体的不同拍摄视角、视频中的音频和图像帧、网页上的文字和对其

他网页的链接、人的脑电图(electroencephalogram, EEG)和心电图(electrocardiogram, ECG)、同一语义对象的不同描述语言等都是多视图数据。使用多视图数据能够帮助被描述对象的信息得到更充分的表达, 从而更好地完成分类、聚类、回归等与机器学习相关的识别任务^[4]。基于此, 多视图学习应运而生。作为机器学习的重要分支, 多视图学习

收稿日期: 2021-04-21; 修回日期: 2021-06-18

通信作者: 伍冬睿, drwu@hust.edu.cn

基金项目: 武汉市应用基础前沿项目(No.2020020601012240); 湖北省技术创新专项(No.2019AEA171)

Foundation Items: Wuhan Science and Technology Bureau (No.2020020601012240), Technology Innovation Project of Hubei Province of China (No.2019AEA171)

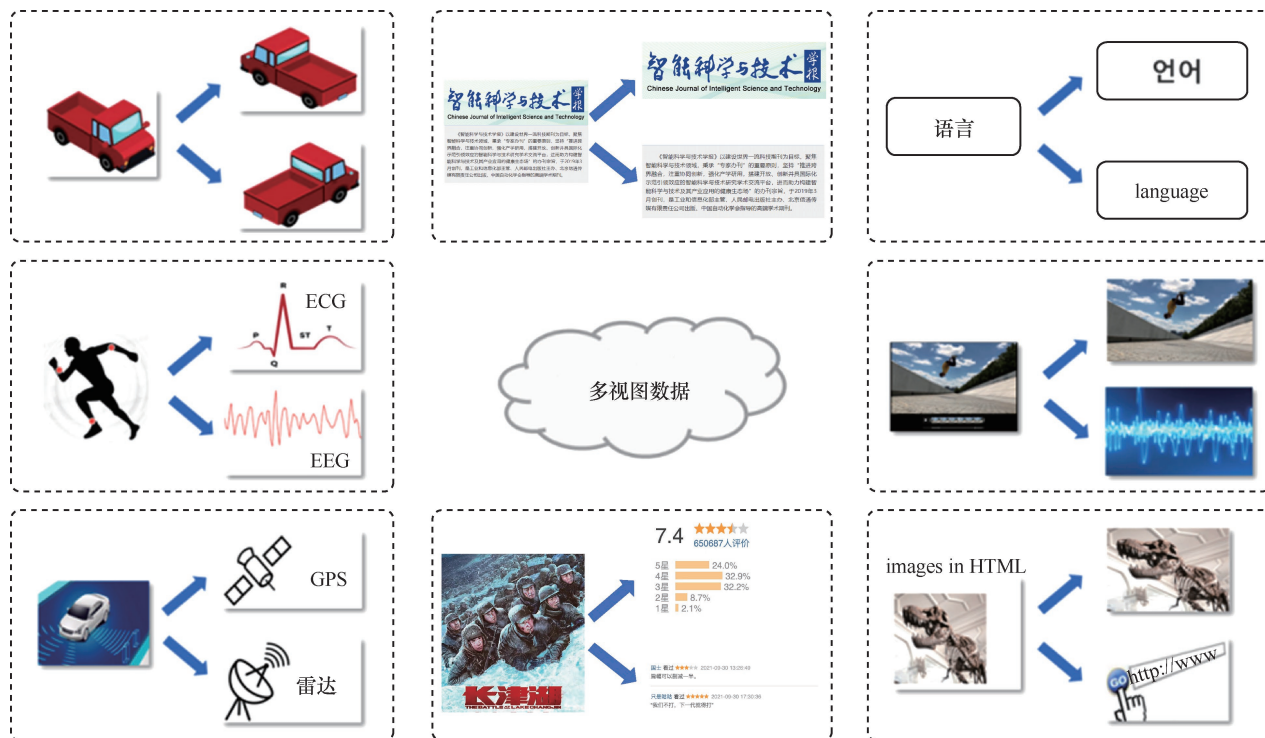


图 1 多视图数据

正在引起越来越广泛的关注和研究。多视图学习的主要目的在于挖掘视图间的一致性和互补性^[5-6],以达到比单视图更好的泛化性能。

多视图学习算法可以分为 3 类^[7-8]: 协同训练、多核学习、子空间学习。协同训练算法^[9-10]是基于分歧的半监督学习方法中相对经典的一类,该算法首先利用已被标记的样本在各视图上分别训练一个学习器;然后利用各学习器对所有未被标记的样本进行预测,交换置信度较高的样本,并重新训练模型;重复上一步骤,直到迭代次数达到预设最大值或者所有未被标记的样本均被标记。多核学习^[11-12]首先利用核函数将原始样本映射到不同的特征空间,然后对各空间中的特征进行线性或非线性组合,以最大限度地挖掘不同空间的特征表达能力,进而提高任务的性能。子空间学习^[13-15]假设每个视图都由一个共同的隐空间生成,各算法通常寻找各视图向共同子空间的映射方向,从而实现从高维向低维的映射。

典型相关分析 (canonical correlation analysis, CCA) 是子空间学习中的重要方法,也是比较经典的统计学方法之一。1936 年, Hotelling H^[13]首次提出了 CCA, 其旨在为各视图寻找合适的映射方向,将视图映射到相关空间,在该空间中不同视图间具有最大的相关性。CCA 的原理如图 2 所示。由于 CCA 考虑了不同特征之间的关系,与多视图学习的思想非常契

合,自然地推广到多视图学习中^[16-18],应用方向包括多视图聚类^[19-20]、多视图回归^[21]、多视图降维^[22]等。

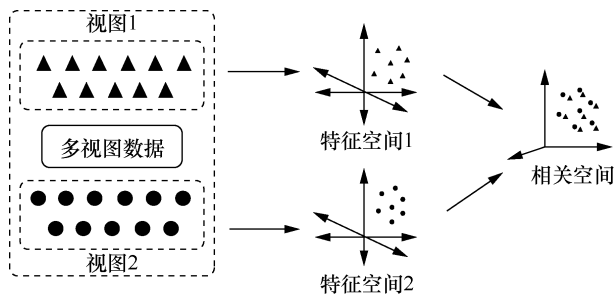


图 2 CCA 的原理

传统的 CCA 只能处理两个视图的线性相关关系,无法应用于包含多个视图或者包含非线性相关性的数据集。此外,如果将 CCA 应用到分类、回归等有监督的任务中,其作为一种无监督的方法将导致标签信息的浪费。针对上述问题,大量非线性的、针对多个视图的、有监督的基于 CCA 的多视图学习方法被提出^[23-36]。

对于包含非线性相关性的数据集而言,通常有 3 种分析方法。第一种是基于核函数的方法,该方法首先使用核技巧将各视图映射到高维特征空间,然后进行线性相关的映射,例如核典型相关分析 (kernel CCA, KCCA)^[24];第二种是基于近邻的方法,这类方法的思想来源于图模型,即通过保留局

部线性结构的方式降低数据的维度,例如局部保留典型相关分析(locality preserving CCA, LPCCA)^[25];第三种是基于深度神经网络(deep neural network, DNN)的方法,这类方法首先通过 DNN 对数据进行复杂的非线性映射,然后进行线性相关性分析,代表方法包括深度典型相关分析(deep CCA, DCCA)^[27]、深度变分典型相关分析(deep variational CCA, DVCCA)^[28]。

针对包含两个以上视图的真实场景,常见的思路有通过高维矩阵张量的计算直接最大化所有视图的相关性,例如张量典型相关分析(tensor CCA, TCCA)^[31]。除此之外,还有成对地最大化多个视图间的相关性,例如基于最小二乘的典型相关分析(least squares based CCA, LS-CCA)^[32];也有基于子空间思想的方法,其试图为各视图寻找一个共同的相关子空间,例如广义典型相关分析(generalized CCA, GCCA)^[33]。

样本的标签信息对于分类、回归等任务至关重要。为了充分利用标签中的判别性信息, Sun T K 等人^[34]于 2018 年提出了判别式典型相关分析(discriminative CCA, DisCCA),该方法同时考虑了视图间的类内和类间相似性。在此基础上, Sun S L 等人^[35]结合 CCA 和线性判别分析(linear discriminant analysis, LDA)^[36]提出了多视图线性判别分析(multi-view linear discriminant analysis, MLDA)。

本文对基于 CCA 的多视图学习方法进行了综述。第 1 节介绍了针对两个视图和多个视图(两个以上视图)的具有代表性的扩展方法;第 2 节阐述了这些方法在多视图识别、跨模态检索和分类以及多视图信息嵌入和表示中的典型应用;第 3 节总结全文,并对基于 CCA 的多视图学习面临的挑战和未来研究趋势进行了总结和展望。

1 基于 CCA 的多视图学习方法

基于 CCA 的多视图学习方法按照可处理的视图的数量可以分为针对两个视图和针对多个视图的方法,按照是否使用了标签信息可以分为有监督和无监督的方法,按照是否进行了数据的非线性映射可以分为线性和非线性的方法^[37]。本文主要按照可处理的视图数量以及是否进行非线性映射进行分类,主要分为针对两个视图的线性方法、针对两个视图的非线性方法、针对多个视图的线性方法以及针对多个视图的非线性方法 4 类,同时在各个类

别中分别介绍有监督和无监督的方法。本节介绍的基于 CCA 的多视图学习方法见表 1,本文使用的主要符号的含义见表 2。

表 1 基于 CCA 的多视图学习方法

方法	多个视图	有监督	非线性	基于 DNN
CCA				
sCCA				
KCCA			✓	
RCCA			✓	
LPCCA			✓	
DisCCA		✓		
MLDA		✓		
MULDA		✓		
DCCA			✓	✓
DisDCCA		✓	✓	✓
DCCAE			✓	✓
DisDCCAE		✓	✓	✓
DVCCA			✓	✓
SUMCOR-CCA	✓			
LS-CCA	✓			
GCCA	✓			
TCCA	✓			
DGCCA	✓	✓	✓	✓

表 2 本文使用的主要符号的含义

符号	含义
$\mathbf{X}, \mathbf{Y}$	视图 $\mathbf{X}$ 和视图 $\mathbf{Y}$ 的特征矩阵
$(\mathbf{x}_i, \mathbf{y}_i)$	视图 $\mathbf{X}$ 和视图 $\mathbf{Y}$ 的第 i 对样本
N	各视图的样本数
c	类别数
n_i	类别 i 的样本数
J	视图数
K	典型向量的个数
d_x	视图 $\mathbf{X}$ 的特征维度
$\mathbf{w}_x$	视图 $\mathbf{X}$ 的一个典型向量
$\mathbf{W}_x$	视图 $\mathbf{X}$ 的典型矩阵
$\mathbf{S}_{XY}$	视图 $\mathbf{X}$ 和视图 $\mathbf{Y}$ 的协方差矩阵
$\cdot^\top$	矩阵的转置
$\mathbf{I}$	单位矩阵
$\ \cdot\ _2$	向量的 2 范数
$\ \cdot\ _F$	矩阵的 Frobenius 范数
$\text{tr}(\cdot)$	矩阵的迹

1.1 针对两个视图的线性方法

针对两个视图的线性方法以 CCA 为代表, 包括增加稀疏性约束的稀疏典型相关分析 (sparse CCA, sCCA)^[38], 以及有监督的 DisCCA、MLDA、MULDA。

1.1.1 CCA

CCA 通过寻找几组合适的映射方向以最大化两组随机变量之间的相关性。对于两个去均值后的视图 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{d_x \times N}$ 和 $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N] \in \mathbb{R}^{d_y \times N}$, 其中 N 表示样本个数, d_x 和 d_y 分别表示视图 $\mathbf{X}$ 和视图 $\mathbf{Y}$ 的特征维度, CCA 通过寻找两个投影方向 $\mathbf{w}_x \in \mathbb{R}^{d_x \times 1}$ 和 $\mathbf{w}_y \in \mathbb{R}^{d_y \times 1}$, 最大化 $\mathbf{X}^\top \mathbf{w}_x$ 和 $\mathbf{Y}^\top \mathbf{w}_y$ 的相关系数:

$$\max_{\mathbf{w}_x, \mathbf{w}_y} \rho(\mathbf{X}^\top \mathbf{w}_x, \mathbf{Y}^\top \mathbf{w}_y) = \frac{\mathbf{w}_x^\top \mathbf{S}_{XY} \mathbf{w}_y}{\sqrt{(\mathbf{w}_x^\top \mathbf{S}_{XX} \mathbf{w}_x)(\mathbf{w}_y^\top \mathbf{S}_{YY} \mathbf{w}_y)}} \quad (1)$$

其中, 协方差矩阵的计算方式为:

$$\begin{aligned} \mathbf{S}_{XY} &= \frac{1}{N-1} \mathbf{XY}^\top \\ \mathbf{S}_{XX} &= \frac{1}{N-1} \mathbf{XX}^\top \\ \mathbf{S}_{YY} &= \frac{1}{N-1} \mathbf{YY}^\top \end{aligned} \quad (2)$$

在式 (1) 以及后续所有方法的求解过程中, 为了保证矩阵求逆的稳定性, 可以通过 $\mathbf{S}_{XX} + r_x \mathbf{I}$ 和 $\mathbf{S}_{YY} + r_y \mathbf{I}$ 的形式对协方差矩阵进行正则化, 其中 $\mathbf{I}$ 为单位矩阵, r_x 和 r_y 为正则化系数。正则化系数通常为较小的正数, 例如 0.001。在式 (1) 的函数中, 当 $\mathbf{w}_x$ 和 $\mathbf{w}_y$ 的值呈现倍数变化时, 函数的最大值不会发生变化。为了保证解的唯一性, 通常采用固定分母、优化分子的方式进行优化, 从而可以将式 (1) 转化为如下形式:

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^\top \mathbf{S}_{XY} \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^\top \mathbf{S}_{XX} \mathbf{w}_x = 1, \mathbf{w}_y^\top \mathbf{S}_{YY} \mathbf{w}_y = 1 \end{aligned} \quad (3)$$

式 (3) 可以采用特征值分解和奇异值分解两种方式求解^[27,37]。以特征值分解的方式为例, 除了获得最大特征值对应的映射向量并将其作为典型向量, 也可以同时计算第二大、第三大……的特征值所对应的特征向量, 以构成映射矩阵 $\mathbf{W}_x = [\mathbf{w}_{x,1}, \dots, \mathbf{w}_{x,K}] \in \mathbb{R}^{d_x \times K}$ 和 $\mathbf{W}_y = [\mathbf{w}_{y,1}, \dots, \mathbf{w}_{y,K}] \in \mathbb{R}^{d_y \times K}$, 其中 K 表示典型向量的个数。本文将矩阵

$\mathbf{W}_x$ 和 $\mathbf{W}_y$ 称为典型矩阵, 将 $\mathbf{X}^\top \mathbf{w}_x$ 和 $\mathbf{Y}^\top \mathbf{w}_y$ 称为典型变量。

1.1.2 sCCA

通过 CCA 计算的典型矩阵通常是稠密矩阵, 并且当特征维度高于样本数时, 求解过程中的求逆等运算将变得不稳定。为了解决上述问题, Witten D M 等人^[38]在惩罚性矩阵分解 (penalized matrix decomposition, PMD)^[39]的基础上提出了 sCCA。

PMD 首先使用奇异值分解计算任意矩阵 $\mathbf{M} \in \mathbb{R}^{d \times N}$ 的近似矩阵 $\hat{\mathbf{M}} = \sum_{i=1}^m b_i \mathbf{u}_i \mathbf{v}_i^\top$, 该近似矩阵的秩为 m , 其中 b_i 表示第 i 个最大的奇异值, $\mathbf{u}_i \in \mathbb{R}^{d \times 1}$ 和 $\mathbf{v}_i \in \mathbb{R}^{N \times 1}$ 分别表示对应的左、右奇异向量。当 $m=1$ 时, 最小化 $\|\mathbf{M} - \hat{\mathbf{M}}\|_F^2$ 等价于最大化以下目标函数:

$$\begin{aligned} \max_{\mathbf{u}, \mathbf{v}} \quad & \mathbf{u}^\top \mathbf{M} \mathbf{v} \\ \text{s.t.} \quad & \|\mathbf{u}\|_2^2 = 1, \|\mathbf{v}\|_2^2 = 1, p_1(\mathbf{u}) \leq c_1, p_2(\mathbf{v}) \leq c_2 \end{aligned} \quad (4)$$

其中, $p_1(\mathbf{u}) = \sum_{i=1}^d |u_i|$ 和 $p_2(\mathbf{v}) = \sum_{i=1}^N |v_i|$ 表示 Lasso 惩罚, 系数 c_1 和 c_2 控制了稀疏的程度。

sCCA 将式 (4) 中的矩阵 $\mathbf{M}$ 替换为视图间的协方差矩阵的形式 $\mathbf{XY}^\top$, 其目标函数如下:

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^\top \mathbf{XY}^\top \mathbf{w}_y \\ \text{s.t.} \quad & \|\mathbf{w}_x\|_2^2 = 1, \|\mathbf{w}_y\|_2^2 = 1, p_1(\mathbf{w}_x) \leq c_1, p_2(\mathbf{w}_y) \leq c_2 \end{aligned} \quad (5)$$

为了保证得到凸目标函数, 首先将式 (5) 中的约束 $\|\mathbf{w}_x\|_2^2 = 1$ 和 $\|\mathbf{w}_y\|_2^2 = 1$ 分别修改为 $\|\mathbf{w}_x\|_2^2 \leq 1$ 和 $\|\mathbf{w}_y\|_2^2 \leq 1$ 。然后, 通过求解多变量 PMD 的方式^[39]计算出 K 对典型向量 $\{(\mathbf{w}_{x,k}, \mathbf{w}_{y,k})\}_{k=1}^K$ 。在 sCCA 的基础上, 有许多改进的方法被提出, 例如考虑结构化信息的结构化稀疏典型相关分析 (group-level sCCA)^[40-42]。总之, sCCA 系列的改进都是为典型向量增加稀疏性约束以提高降维后特征的可解释性。

1.1.3 DisCCA、MLDA、MULDA

DisCCA 是在 CCA 的基础上提出的具有代表性的有监督方法之一。相对于 CCA, DisCCA 在最大化视图间同类样本相关性的同时, 最小化视图间非同类样本的相关性。首先, 将视图 $\mathbf{X}$ 和 $\mathbf{Y}$ 按照类别进行重新排列:

$$\hat{X} = [\mathbf{x}_1^{(1)}, \dots, \mathbf{x}_{n_1}^{(1)}, \dots, \mathbf{x}_1^{(c)}, \dots, \mathbf{x}_{n_c}^{(c)}] \quad (6)$$

$$\hat{Y} = [\mathbf{y}_1^{(1)}, \dots, \mathbf{y}_{n_1}^{(1)}, \dots, \mathbf{y}_1^{(c)}, \dots, \mathbf{y}_{n_c}^{(c)}] \quad (7)$$

式(6)和式(7)中的 $\mathbf{x}_j^{(i)}$ 和 $\mathbf{y}_j^{(i)}$ 分别表示类别 i 的第 j 个样本, n_c 表示第 c 个类别的样本数。DisCCA的目标函数为:

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^\top \hat{X} \hat{A} \hat{Y}^\top \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^\top \hat{X} \hat{X}^\top \mathbf{w}_x = 1, \mathbf{w}_y^\top \hat{Y} \hat{Y}^\top \mathbf{w}_y = 1 \end{aligned} \quad (8)$$

式(8)仍然可以转化为特征值分解的形式进行求解, 其中矩阵 $\hat{A} = \text{diag}(\mathbf{I}_{n_1 \times n_1}, \dots, \mathbf{I}_{n_1 \times n_1}, \dots, \mathbf{I}_{n_c \times n_c})$ 表示关于类别的分块矩阵。DisCCA本质上是将最大化视图间对应样本之间的相关性变成了最大化视图间任意同类样本对之间的相关性。在DisCCA原文(参考文献[34])的相关实验中提到, 两个视图的相关性非常高对分类任务可能没有帮助。这是因为在两个视图完全相关(相关系数为1)的情况下, 对视图进行融合只会导致信息的冗余。因此原文(参考文献[34])在实验中证明了DisCCA的相关性映射能力相较于CCA有所降低, 但是显著提高了分类准确率。

MLDA在CCA的基础上结合了LDA的思想, 在最大化视图间相关性的同时最大化各个视图内的类间方差。定义视图 X 的类间散度矩阵 $S_{b,X} = \frac{1}{N} \hat{X} W \hat{X}^\top$, 其中 $W = \text{diag}(W_1, W_2, \dots, W_c)$ ($W_c \in \mathbb{R}^{n_c \times n_c}$)的元素等于 $\frac{1}{n_c}$, 视图 Y 的类间散度矩阵具有相似定义。MLDA的目标函数为:

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^\top S_{b,X} \mathbf{w}_x + \mathbf{w}_y^\top S_{b,Y} \mathbf{w}_y + \eta \mathbf{w}_x^\top S_{XY} \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^\top S_{XX} \mathbf{w}_x + \sigma \mathbf{w}_y^\top S_{YY} \mathbf{w}_y = 1 \end{aligned} \quad (9)$$

其中, η 为惩罚系数, $\sigma = \frac{\text{tr}(S_{XX})}{\text{tr}(S_{YY})}$ 。采用求和的约束形式, 是为了保证得到解析解。利用拉格朗日乘法, 式(9)最终可以转化为特征值分解的形式进行求解。

MLDA在最大化视图内方差的过程中, 没有考虑不同特征维度之间的冗余。多视图不相关线性判别分析(multi-view uncorrelated linear discriminant analysis, MULDA)^[35]在MLDA的基础上增加了不同映射方向之间的正交约束, 其目标函数为:

$$\begin{aligned} \max_{\mathbf{w}_{x,k}, \mathbf{w}_{y,k}} \quad & \mathbf{w}_{x,k}^\top S_{b,X} \mathbf{w}_{x,k} + \mathbf{w}_{y,k}^\top S_{b,Y} \mathbf{w}_{y,k} + \eta \mathbf{w}_{x,k}^\top S_{XY} \mathbf{w}_{y,k} \\ \text{s.t.} \quad & \mathbf{w}_{x,k}^\top S_{XX} \mathbf{w}_{x,k} + \sigma \mathbf{w}_{y,k}^\top S_{YY} \mathbf{w}_{y,k} = 1 \\ & \mathbf{w}_{x,k}^\top S_{XX} \mathbf{w}_{x,i} = 0, \mathbf{w}_{y,k}^\top S_{YY} \mathbf{w}_{y,i} = 0, i=1, \dots, k-1 \end{aligned} \quad (10)$$

MLDA和MULDA从视图内的类别散度的角度提高所提取特征的判别性, 同时考虑最大化视图内的类间散度与最大化相关性。而DisCCA则是从视图之间的类别相关性来考虑, 由于计算中涉及视图间相同类别样本的任意组合, 因此该方法对异常值更加敏感。Sun S L等人^[35]还进一步提出将MLDA和MULDA目标函数中的CCA替换为DisCCA, 该方法能够同时考虑视图内和视图间的类别信息。

1.2 针对两个视图的非线性方法

第1.1节主要介绍了针对两个视图的线性方法, 本节将介绍4类针对两个视图的非线性方法。

1.2.1 KCCA、RCCA

KCCA首先利用(非线性)核函数将视图映射到高维特征空间后进行相关性映射。设视图中样本的映射为 ϕ_x 和 ϕ_y , 则视图所在的线性空间分别表示为 $\Phi_x = [\phi_x(\mathbf{x}_1), \dots, \phi_x(\mathbf{x}_N)]$ 和 $\Phi_y = [\phi_y(\mathbf{y}_1), \dots, \phi_y(\mathbf{y}_N)]$ 。与CCA的目标一致, 此时KCCA的目标函数为:

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^\top \Phi_x \Phi_y^\top \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^\top \Phi_x \Phi_x^\top \mathbf{w}_x = 1, \mathbf{w}_y^\top \Phi_y \Phi_y^\top \mathbf{w}_y = 1 \end{aligned} \quad (11)$$

其中, $\mathbf{w}_x$ 和 $\mathbf{w}_y$ 可以分别表示为 Φ_x 和 Φ_y 的线性组合:

$$\begin{aligned} \mathbf{w}_x &= \sum_{i=1}^N a^i \phi_x(\mathbf{x}_i) = \Phi_x \mathbf{a} \\ \mathbf{w}_y &= \sum_{i=1}^N b^i \phi_y(\mathbf{y}_i) = \Phi_y \mathbf{b} \end{aligned} \quad (12)$$

其中, $\mathbf{a} = [a^1, \dots, a^N]^\top$ 和 $\mathbf{b} = [b^1, \dots, b^N]^\top$ 表示线性系数。式(11)可以改写为:

$$\begin{aligned} \max_{\mathbf{a}, \mathbf{b}} \quad & \mathbf{a}^\top \Phi_x^\top \Phi_x \Phi_y^\top \Phi_y \mathbf{b} \\ \text{s.t.} \quad & \mathbf{a}^\top \Phi_x^\top \Phi_x \Phi_x^\top \Phi_x \mathbf{a} = 1, \mathbf{b}^\top \Phi_y^\top \Phi_y \Phi_y^\top \Phi_y \mathbf{b} = 1 \end{aligned} \quad (13)$$

定义核矩阵 $K_x = \Phi_x^\top \Phi_x$ 和 $K_y = \Phi_y^\top \Phi_y$, 其中 $(K_x)_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ 。 κ 表示核函数, 例如径向基函数。式(13)可以简化为:

$$\begin{aligned} \max_{\mathbf{a}, \mathbf{b}} \quad & \mathbf{a}^\top K_x K_y \mathbf{b} \\ \text{s.t.} \quad & \mathbf{a}^\top K_x K_x \mathbf{a} = 1, \mathbf{b}^\top K_y K_y \mathbf{b} = 1 \end{aligned} \quad (14)$$

与CCA相似, 通常进一步对核协方差矩阵进

行正则化以保证数值的稳定性^[37], 将 $\mathbf{K}_x \mathbf{K}_x$ 和 $\mathbf{K}_y \mathbf{K}_y$ 分别替换为 $\mathbf{K}_x \mathbf{K}_x + r_x \mathbf{K}_x$ 和 $\mathbf{K}_y \mathbf{K}_y + r_y \mathbf{K}_y$, 其中 r_x 和 r_y 为正则化系数。线性系数 $\mathbf{a}$ 和 $\mathbf{b}$ 的求解同样可以转化为特征值分解问题^[37]。

使用核函数将视图映射到新的特征空间, 相当于对视图采取了不同的距离度量方式, 这使得视图在低维空间没有表现出的相关关系, 在高维空间能够更容易地体现。然而, KCCA 的计算复杂度较高。Lopez-Paz D 等人^[43]提出了随机非线性典型相关分析 (randomized nonlinear CCA, RCCA), 即通过对视图进行随机非线性映射, 在损失少许性能的情况下显著降低 KCCA 的计算复杂度。

$$\text{RCCA}(\mathbf{X}, \mathbf{Y}) := \text{CCA}(f_x(\mathbf{X}), f_y(\mathbf{Y})) \approx \text{KCCA}(\mathbf{X}, \mathbf{Y}) \quad (15)$$

其中, $f_x(\cdot)$ 表示视图 $\mathbf{X}$ 的映射函数, 其参数 $\mathbf{Q}_x = (\mathbf{q}_{x,1}, \mathbf{q}_{x,2}, \dots, \mathbf{q}_{x,D}) \in \mathbb{R}^{d_x \times D}$ 是通过在与视图无关的分布上随机采样得到的, 例如高斯分布。D 维随机非线性特征 $\mathbf{Z}_x$ 的计算方式如下:

$$\mathbf{z}_{x,d} := [\cos(\mathbf{q}_{x,d}^\top \mathbf{x}_1 + b_d), \dots, \cos(\mathbf{q}_{x,d}^\top \mathbf{x}_N + b_d)]^\top \in \mathbb{R}^{N \times 1}$$

$$\mathbf{Z}_x = [\mathbf{z}_{x,1}, \dots, \mathbf{z}_{x,D}]^\top \in \mathbb{R}^{D \times N} \quad (16)$$

其中, $\mathbf{b} = [b_1, b_2, \dots, b_D]^\top \in \mathbb{R}^{D \times 1}$ 是随机生成的偏置项, $\cos$ 函数可以替换为其他非线性函数, 如 $\sin$ 函数。针对视图 $\mathbf{Y}$, 仍然采取相似的方式获得随机非线性特征。

1.2.2 LPCCA

除了计算复杂度高, KCCA 使用全局核化的策略还忽略了复杂数据中的局部线性关系。局部流形结构的假设认为在高维空间中互为近邻的样本, 映射到低维空间之后也互为近邻^[44]。Sun TK 等人^[25]基于这一假设提出了 LPCCA, 该方法通过考虑样本邻域内的局部线性关系, 实现全局数据降维。LPCCA 认为不同视图的对应样本在映射到相关空间之后应该距离十分接近, 因此 CCA 的目标函数也可以表达为以下形式:

$$\min_{\mathbf{w}_x, \mathbf{w}_y} \sum_{i=1}^N \|\mathbf{w}_x^\top (\mathbf{x}_i - \bar{\mathbf{x}}) - \mathbf{w}_y^\top (\mathbf{y}_i - \bar{\mathbf{y}})\|_2^2 \quad (17)$$

$$\text{s.t.} \quad \sum_{i=1}^N \|\mathbf{w}_x^\top (\mathbf{x}_i - \bar{\mathbf{x}})\|^2 = 1, \quad \sum_{i=1}^N \|\mathbf{w}_y^\top (\mathbf{y}_i - \bar{\mathbf{y}})\|^2 = 1$$

其中, $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$ 和 $\bar{\mathbf{y}} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i$ 分别表示 $\mathbf{X}$ 和 $\mathbf{Y}$ 的均值向量。经过代数运算, 式 (17) 变换为:

$$\max_{\mathbf{w}_x, \mathbf{w}_y} \mathbf{w}_x^\top \cdot \sum_{i=1}^N \sum_{j=1}^N (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{y}_i - \mathbf{y}_j)^\top \cdot \mathbf{w}_y$$

$$\text{s.t.} \quad \mathbf{w}_x^\top \cdot \sum_{i=1}^N \sum_{j=1}^N (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \cdot \mathbf{w}_x = 1 \quad (18)$$

$$\mathbf{w}_y^\top \cdot \sum_{i=1}^N \sum_{j=1}^N (\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)^\top \cdot \mathbf{w}_y = 1$$

定义相似矩阵 $\mathbf{S}_x$, 矩阵中第 i 行、第 j 列的元素为:

$$S_{x,ij} = \begin{cases} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / t_x\right), & \mathbf{x}_j \in \text{Nei}(\mathbf{x}_i) \\ 0, & \text{其他} \end{cases} \quad (19)$$

其中, $t_x = \sum_{i=1}^N \sum_{j=1}^N \frac{2\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{N(N-1)}$ 表示所有样本的均方距离, $\text{Nei}(\mathbf{x}_i)$ 表示 $\mathbf{x}_i$ 领域内的样本。 $\mathbf{Y}$ 的相似矩阵 $\mathbf{S}_y$ 具有相似定义。

通过将 $\mathbf{S}_x$ 和 $\mathbf{S}_y$ 引入式 (18), 全局相关关系可以被分解为多个局部的、邻域样本之间的线性相关性。LPCCA 的目标函数为:

$$\max_{\mathbf{w}_x, \mathbf{w}_y} \mathbf{w}_x^\top \cdot \sum_{i=1}^N \sum_{j=1}^N S_{x,ij} (\mathbf{x}_i - \mathbf{x}_j) S_{y,ij} (\mathbf{y}_i - \mathbf{y}_j)^\top \cdot \mathbf{w}_y$$

$$\text{s.t.} \quad \mathbf{w}_x^\top \cdot \sum_{i=1}^N \sum_{j=1}^N S_{x,ij} (\mathbf{x}_i - \mathbf{x}_j) S_{x,ij} (\mathbf{x}_i - \mathbf{x}_j)^\top \cdot \mathbf{w}_x = 1$$

$$\mathbf{w}_y^\top \cdot \sum_{i=1}^N \sum_{j=1}^N S_{y,ij} (\mathbf{y}_i - \mathbf{y}_j) S_{y,ij} (\mathbf{y}_i - \mathbf{y}_j)^\top \cdot \mathbf{w}_y = 1 \quad (20)$$

该目标函数同样可以转化为特征值分解的形式进行求解。基于邻域的变形方法强调保留局部近邻关系而忽略非近邻的相关关系^[45]。除此之外, 也有在 CCA 原目标函数的基础上进一步增加局部邻域关系的基于“加法策略”的方法^[46-47]。

1.2.3 DCCA、DisDCCA、DCCAE、DisDCCAE

一类常见的视图非线性相关性分析方法是将 CCA 与 DNN 结合。这类方法首先利用 DNN 对非线性数据进行复杂的非线性变换, 然后进行线性典型相关分析。在 DNN 和 CCA 结合的基础上, 进一步结合自编码器、变分自编码器等, 可以延伸出多种变形。

DCCA 模型结构如图 3 (a) 所示。设 $\mathbf{f}_x$ 和 $\mathbf{f}_y$ 分别表示两个视图的特征映射网络。 $\mathbf{H}_x = \mathbf{f}_x(\mathbf{X}) \in \mathbb{R}^{d'_x \times N}$ 和 $\mathbf{H}_y = \mathbf{f}_y(\mathbf{Y}) \in \mathbb{R}^{d'_y \times N}$ 分别表示两个网络的输出, 其中 d'_x 和 d'_y 分别表示网络输出的特征维度, 通常有 $d'_x = d'_y$ 。根据 CCA 奇异值分解的求解方法,

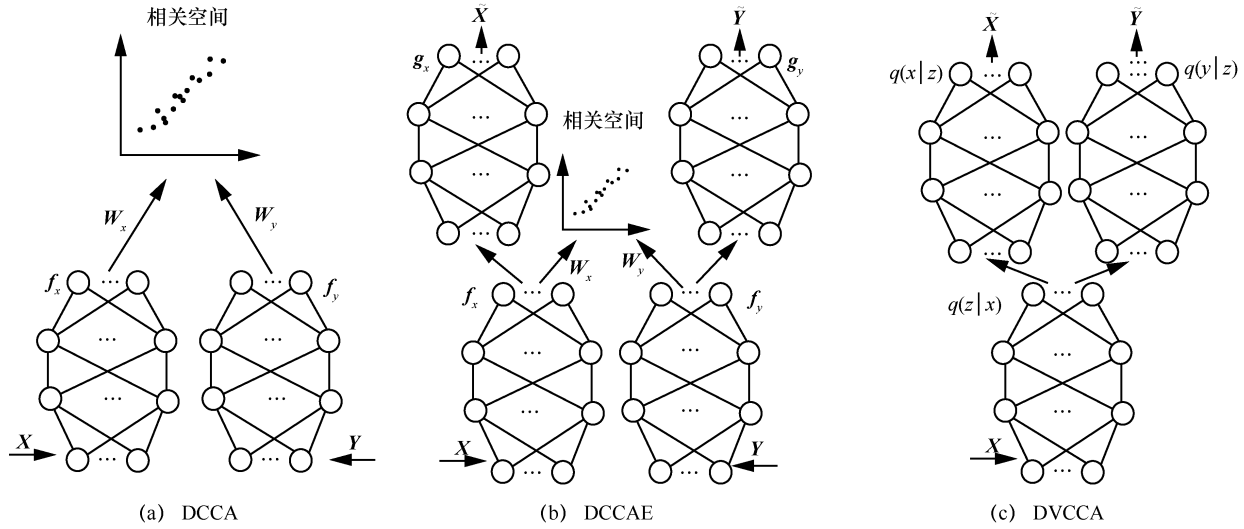


图3 针对两个视图的基于 DNN 的模型

DCCA 的目标为最大化矩阵 T 的前 K 个奇异值之和。矩阵 T 的具体定义为:

$$T = \left(\frac{1}{N-1} H_x H_x^\top \right)^{\frac{1}{2}} \left(\frac{1}{N-1} H_x H_y^\top \right) \left(\frac{1}{N-1} H_y H_y^\top \right)^{\frac{1}{2}} \quad (21)$$

则 DCCA 的目标函数为:

$$\begin{aligned} \max_{f_x, f_y, w_x, w_y} \quad & \sum_{k=1}^K \sigma_k(T) \\ \text{s.t.} \quad & w_x^\top \left(\frac{1}{N-1} H_x H_x^\top \right) w_x = 1 \\ & w_y^\top \left(\frac{1}{N-1} H_y H_y^\top \right) w_y = 1 \end{aligned} \quad (22)$$

其中, $\sigma_k(T)$ 表示矩阵 T 的第 k 个最大的奇异值。

Andrew G 等人^[27]对 DCCA 采用了全批量 (full-batch) 的优化策略, 这是因为在 CCA 的计算过程中涉及对训练集数据协方差矩阵的计算, 如果将数据集划分为多个小的 batch, 协方差矩阵的计算可能变得不准确。然而, 全批量的优化方式内存消耗大、耗时长, 尤其是对于几百万甚至上千万的数据集而言。针对这个问题, Wang WR 等人^[48]提出只要保证划分的批量大小 (batch size) 相对较大, batch 内的数据足以对整体的协方差进行近似估计, 这使得小批量 (mini-batch) 的优化策略可以被使用。

Wang WR 等人^[26]在 DCCA 的基础上结合自编码器提出了深度典型相关自编码器 (deep canoni-

cally correlated auto encoders, DCCAE)。解码器通过对原始视图进行重构, 最大化编码器输出与原始输入之间的互信息。DCCAE 能够使编码的特征在保留自身信息的同时从其他视图中学习有用的信息, 增强了视图的独立性, 有助于处理测试时仅包含单一视图的情况。DCCAE 的模型结构如图 3 (b) 所示, 图中 g_x 和 g_y 分别表示对视图 X 和 Y 进行重构的解码器。设 λ 为惩罚稀疏, DCCAE 的目标函数为:

$$\begin{aligned} \max_{f_x, f_y, g_x, g_y, w_x, w_y} \quad & \sum_{k=1}^K \sigma_k(T) - \\ & \frac{\lambda}{N} \left(\|X - g_x(f_x(X))\|_F^2 + \|Y - g_y(f_y(Y))\|_F^2 \right) \\ \text{s.t.} \quad & w_x^\top \left(\frac{1}{N-1} H_x H_x^\top \right) w_x = 1, \quad w_y^\top \left(\frac{1}{N-1} H_y H_y^\top \right) w_y = 1 \end{aligned} \quad (23)$$

将 CCA 扩展到有监督的 DisCCA 的策略完全可以等价地应用到将 DCCA 扩展到有监督的判别式深度典型相关分析 (discriminative deep CCA, DisDCCA)^[30]中。DCCA 和 DisDCCA 在实现中主要包含两点差异: 一是在计算每个 batch 的损失时, 将式 (21) 中的 $\frac{1}{N-1} H_x H_y^\top$ 替换为 $\frac{1}{N-1} H_x A H_y^\top$, 其中 A 的定义与 DisCCA 中的相同; 二是当模型训练结束后, 对于所有训练集的网络输出, 应用 DisCCA 计算出典型矩阵 W_x 和 W_y 。

同样地, DCCAE 可以扩展为有监督的判别式深度典型相关自编码器 (discriminative deep ca-

nonically correlated autoencoders, DisDCCA), 这与将 DCCA 扩展到 DisDCCA 的方法完全相同。对于损失函数中包含最大化相关性目标的深度学习模型而言, 总是可以通过增加判别式矩阵的方式扩展到有监督。实验表明, 这种扩展方法几乎总是优于无监督的方法。

1.2.4 DVCCA

Bach F R 等人^[49]从隐变量模型的角度解释了 CCA。该模型假设视图中的样本 $\mathbf{x}$ 和 $\mathbf{y}$ 在多元隐变量空间 $\mathbf{z} \in \mathbb{R}^{d_z \times 1}$ 中是条件独立的, 其中 d_z 表示隐变量空间的维度。CCA 旨在最大化变量 $\mathbf{x}$ 和 $\mathbf{y}$ 的联合概率分布:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{z}) &= p(\mathbf{z})p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z}) \\ p(\mathbf{x}, \mathbf{y}) &= \int p(\mathbf{x}, \mathbf{y}, \mathbf{z})d\mathbf{z} \end{aligned} \quad (24)$$

DVCCA^[28]由变分自编码器^[50]扩展而来, 其模型结构如图 3 (c) 所示。首先, 编码器 $\mathbf{f}_x$ 计算视图 $\mathbf{X}$ 中样本 $\mathbf{x}_i$ 的均值向量 $\boldsymbol{\mu}_i$ 和协方差矩阵 $\boldsymbol{\Sigma}_i$; 然后从分布 $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 中采样 L 个样本 $\{\mathbf{z}_i^{(l)}\}_{l=1}^L$; 最后解码器 $\mathbf{g}_x$ 和 $\mathbf{g}_y$ 分别对 $\mathbf{x}$ 和 $\mathbf{y}$ 进行重构。DVCCA 的目标函数为:

$$\begin{aligned} \min_{\mathbf{f}_x, \mathbf{g}_x, \mathbf{g}_y} \quad & \frac{1}{N} \sum_{i=1}^N D_{\text{KL}}(q(\mathbf{z}_i | \mathbf{x}_i) \| p(\mathbf{z}_i)) + \\ & \frac{\lambda}{NL} \sum_{i=1}^N \sum_{l=1}^L (\log p(\mathbf{x}_i | \mathbf{z}_i^{(l)}) + \log p(\mathbf{y}_i | \mathbf{z}_i^{(l)})) \end{aligned} \quad (25)$$

其中, 第一项表示分布 $q(\mathbf{z}_i | \mathbf{x}_i)$ 和 $p(\mathbf{z}_i)$ 之间的 KL 散度 (Kullback-Leibler divergence), 后两项表示近似后验分布下对数似然的期望值, 等价于重构误差。DVCCA 仅仅考虑了两个视图的共同隐变量 $\mathbf{z}$, 不足以对视图 $\mathbf{X}$ 和 $\mathbf{Y}$ 的信息进行充分描述。为了充分挖掘各视图独有的信息, Wang W R 等人^[28]在 DVCCA 的基础上进一步考虑各个视图独有的编码信息, 提出了 DVCCAP (DVCCA-private)。参考文献[28]的实验结果表明, DVCCAP 比 DVCCA 具有更好的表现。

DVCCA 是较早从概率模型的角度出发对典型相关进行建模的方法, 属于概率 CCA。在这类方法中, 还有多种基于贝叶斯理论的概率模型^[51-53]。

1.3 针对多个视图的线性方法

实际场景中可能包含多个视图, 本节分别介绍针对多个视图进行典型相关分析的线性方法和非线性方法。

1.3.1 SUMCOR-CCA、LS-CCA

求和形式的典型相关分析 (sum of correlation-CCA, SUMCOR-CCA)^[54]是针对多个视图提出的一种自然的扩展形式, 该方法成对地最大化多个视图的相关性。设 $\{\mathbf{X}_j \in \mathbb{R}^{d_j \times N}\}_{j=1}^J$ 表示包含 J 个去均值后的视图, d_j 表示视图 j 的特征维度。

SUMCOR-CCA 的目标函数为:

$$\begin{aligned} \max_{\{\mathbf{w}_j\}_{j=1}^J} \quad & \sum_{i=1}^J \sum_{j=1}^J \mathbf{w}_i^\top \mathbf{X}_i \mathbf{X}_j^\top \mathbf{w}_j \\ \text{s.t.} \quad & \mathbf{w}_j^\top \mathbf{X}_j \mathbf{X}_j^\top \mathbf{w}_j = 1, \quad j=1, \dots, J \end{aligned} \quad (26)$$

由于式 (26) 没有关于 $\mathbf{w}_j$ 的分析解, 通常采用 Manopt 工具包^[55]进行迭代求解^[56]。

与 SUMCOR-CCA 相似, LS-CCA 同样成对地最大化多个视图的相关性。该方法假设不同视图的对应样本在相关空间中距离很近, 甚至相互重叠。LS-CCA 旨在为每个视图寻找一组典型向量, 使得映射后不同视图的对应样本之间距离最近。其目标函数为:

$$\begin{aligned} \min_{\{\mathbf{w}_j\}_{j=1}^J} \quad & \frac{1}{2J(J-1)} \sum_{i=1}^J \sum_{j=1}^J \|\mathbf{X}_i^\top \mathbf{w}_i - \mathbf{X}_j^\top \mathbf{w}_j\|_2^2 \\ \text{s.t.} \quad & \frac{1}{J} \sum_{j=1}^J \mathbf{w}_j^\top \mathbf{X} \mathbf{X}^\top \mathbf{w}_j = 1 \end{aligned} \quad (27)$$

式 (27) 可以通过特征值分解对 $\mathbf{w}^{(k)} = [\mathbf{w}_1^{(k)}, \dots, \mathbf{w}_J^{(k)}]$ 进行求解。由于矩阵较大, 计算复杂度高, 可以使用偏最小二乘^[32]进行迭代求解。Via J 等人^[32]证明了当 $J=2$ 时, SUMCOR-CCA 和 LS-CCA 都退化为 CCA。

1.3.2 GCCA

GCCA 假设每个视图都由一个共同的多变量隐空间 $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_N]^\top \in \mathbb{R}^{N \times K}$ 生成, 其目标函数为:

$$\min_{\mathbf{G}, \{\mathbf{w}_j\}_{j=1}^J} \sum_{j=1}^J \|\mathbf{G} - \mathbf{X}_j^\top \mathbf{w}_j\|_F^2 \quad \text{s.t.} \quad \mathbf{G}^\top \mathbf{G} = \mathbf{I} \quad (28)$$

其中, $\mathbf{W}_j \in \mathbb{R}^{d_j \times K}$ 表示视图 j 的典型矩阵, $\mathbf{I}$ 表示单位矩阵, 约束项保证了 $\mathbf{G}$ 为一个单位正交矩阵。式 (28) 可以通过特征值分解的方式进行求解, 实际计算中为了减少计算代价, 会首先对每个视图进行奇异值分解。在 GCCA 的基础上, Benton A 等人^[29]提出了考虑不同视图重要性差异的加权 GCCA。

1.3.3 TCCA

前文介绍的方法几乎是成对地优化视图的相关性,忽略了多个视图的高阶统计特征,例如结构信息等。TCCA 则通过计算多个视图的高阶协方差矩阵直接最大化所有视图的相关性,其目标函数表示为:

$$\begin{aligned} \max_{\{w_j\}_{j=1}^J} \quad & \rho(w_1^\top X_1, w_2^\top X_2, \dots, w_J^\top X_J) \\ \text{s.t.} \quad & w_j^\top X_j X_j^\top w_j = 1, j = 1, \dots, J \end{aligned} \quad (29)$$

其中, ρ 表示 J 个视图的相关性,其计算方式为:

$$\rho(w_1^\top X_1, \dots, w_J^\top X_J) = C_{1,2,\dots,J} \times_1 w_1^\top \cdots \times_J w_J^\top \quad (30)$$

其中, $\times_j$ 表示 j 模态积, $C_{1,2,\dots,J} \in \mathbb{R}^{d_1 \times d_2 \times \dots \times d_J}$ 表示协方差张量,其计算方式为:

$$C_{1,2,\dots,J} = \frac{1}{N} \sum_{i=1}^N x_{1i} \circ x_{2i} \circ \dots \circ x_{Ji} \quad (31)$$

其中, x_{ji} 表示视图 j 的第 i 个样本, $\circ$ 表示张量乘积。

式(29)可以通过交替最小二乘的方式进行求解,以获得各视图对应的典型矩阵。TCCA 利用了张量计算,同时考虑多个视图的相关关系,因此能够更有效地挖掘多视图中包含的高阶相关性,从而获得更可靠的公共子空间。

1.4 针对多个视图的非线性方法

深度广义典型相关分析 (deep generalized CCA, DGCCA) [57] 是针对多个视图的非线性方法,该方法将 GCCA 和 DNN 进行了结合。就原理而言,该非线性的扩展方式同样适用于其他针对多个视图的线性方法。

DGCCA 由 GCCA 扩展而来,被用于处理多个视图的非线性关系,其扩展方法与将 CCA 扩展到 DCCA 的方式相同。DGCCA 的模型结构如图4所示。首先将每个视图通过一个多层神经网络,以获得视图的非线性映射特征 $H_j = f_j(X_j)$, $j = 1, \dots, J$, 然后计算 $[H_j]_{j=1}^J$ 对共同相关空间 G 的重构误差,并将其作为 DGCCA 的损失函数:

$$\min_{\{f_j\}_{j=1}^J} \sum_{j=1}^J \|G - f_j(X_j)^\top W_j\|_F^2 \quad \text{s.t. } G^\top G = I \quad (32)$$

为了将 DGCCA 扩展为有监督的方法, Benton A 等人 [57] 将视图的标签信息作为一个额外的视图加入模型的优化中,用于处理有监督的分类任务。

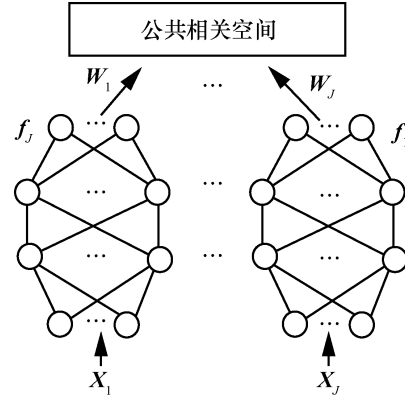


图4 DGCCA 的模型结构

2 基于 CCA 的多视图学习方法的应用场景

作为一种经典的统计学习方法,CCA 及其扩展方法被广泛地应用到多个领域的机器学习任务中。根据任务类型的不同,本文将其分为3类:多视图识别、跨模态检索和分类,以及多视图信息嵌入和表示。

2.1 多视图识别

识别任务通常是指有监督或者无监督的分类或回归。在计算机视觉领域中常常使用基于 CCA 的多视图学习方法,参考文献[28]将 Daubechies 小波变换的低频图像和原始图像分别作为两个视图,完成人脸识别的任务;Zheng W M 等人 [58] 将心理学家的语义评分作为第二视图,对人脸的图像的表情进行识别;在性别识别任务中,Shan C F 等人 [59] 通过 CCA 对步态和面部特征进行了融合。除此之外,另一项具有挑战性的任务是三维空间中的物体姿态估计,解决此问题的一种代表性的方法 [25] 是首先在训练集上使用基于 CCA 的多视图学习方法建立不同姿态之间的联系,然后将测试集映射到相关空间,通过计算近邻的方式估计测试集图像的姿态参数。Xing X L 等人 [60] 采用类似的思路完成了步态识别的任务。

除了计算机视觉任务中的应用之外,其他领域也常常使用基于 CCA 的多视图学习方法,以语音识别为例,参考文献[34,48,61-62]中证明了通过 CCA 对发音和声学特征进行融合能够提高语音识别的准确率;在神经科学领域中,结合大脑生理成像特征和临床症状等非成像特征为疾病的理解和诊断提供了新视角 [63]。其他相似的多视图学习任务包括文本分类 [28]、手写数字识别 [28,30,35]、生物特征结构预测 [31]、动作识别 [64] 等,不同任务的差别主要

在于视图特征的差异。

总而言之, 基于 CCA 的多视图学习方法首先为不同的视图学习新的特征表示, 即典型变量, 然后使用新的特征完成接下来的识别任务。在获得多个视图的典型变量后, 一方面可以将各个视图的特征拼接或者求和之后用于训练学习器^[26,28,34], 另一方面也可以将第一个视图的典型变量用于训练学习器, 对任一视图进行预测^[31,35,64]。这一做法主要考虑到两种实际场景: 一种是测试时可能仅包含一个视图的数据; 第二种是训练时直接将标签作为第二个视图进行特征学习, 因此针对测试样本利用视图 1 的特征计算近邻, 从而实现预测。

2.2 跨模态检索和分类

随着互联网多媒体数据的快速增长, 跨模态检索逐渐成为人工智能领域的热点。跨媒体检索的主要挑战在于“异构型差距”^[65], 即来自不同媒体类型的数据结构空间不一致, 这种差距使得不同媒体数据之间的相似性难以度量, 从而导致跨模态检索变得困难。例如文档通常包含图像和文本, 跨模态检索的目的是当用户输入查询的图像时, 系统自动返回关联的文本, 反之亦然。CCA 是跨模态检索中一种常用的解决方法^[66-68], 它首先将图像和文本映射到 K 维的相关空间中。考虑到不同视图的对应样本在相关空间中位置十分接近, 当输入查询图像时, 直接计算映射到相关空间后的查询图像与所有文本之间的距离, 将距离近的文本作为检索的结果。

跨模态分类^[69]和跨模态检索具有相同的数据分布假设, 即对应的样本在映射到共同相关空间之后距离仍然十分接近, 因此在一个视图上训练的分类器可以用于完成对另一个视图的分类。

2.3 多视图信息嵌入和表示

基于 CCA 的多视图学习方法的又一个重要应用是自然语言处理中的词嵌入^[70-72]。Faruqui M 等人^[71]提出了使用 DCCA 学习多种语言的词嵌入向量, 并证明了这种方式能够提升使用词相似度度量任务的性能。参考文献[29,57]通过 GCCA 学习用户的多视图词嵌入表示, 以完成好友推荐任务。

总而言之, 多视图词嵌入方法具有以下两点优势: 第一点是使用多个视图特征学习到的特征表示能够捕捉更充分的信息, 从而更好地进行相似性的度量; 第二点是学习的嵌入特征能够降低原始特征的维度, 有效降低算法的计算复杂度。

除以上任务外, 还有语音转录^[57]、超分辨率的图像重建^[73-74]、工业过程中的故障检测^[72,75-76]等应用。

3 结束语

随着人工智能和机器学习关注度的增加, 多视图学习也在快速发展。CCA 是多视图学习中的经典方法, 该方法以最大化视图的相关性为目标寻找原始特征空间到公共子空间的映射。本文首先按照各方法所适用的视图数量、是否进行非线性映射的分类方式介绍了基于 CCA 的多种改进算法, 从理论上对比了各个方法的相似性和差异性; 与此同时, 介绍了这类多视图学习方法在多视图识别、跨模态检索和分类以及多视图信息嵌入和表示等场景中的应用。虽然基于 CCA 的多视图学习方法已经有一些成功的应用, 但是仍然存在很多挑战。基于本文的研究内容, 对未来研究方向的展望如下。

(1) CCA 的核心在于最大化视图间的相关性, 然而最大化相关性真的有利于任务相关信息的提取吗? 在有监督的分类任务中, 最大化相关性也可能导致噪声信息的增加和有用信息的损失。因此将 CCA 应用到其他具体任务中时, 需要根据任务和数据本身进行改进和优化。进一步地, 对于多视图学习的众多场景而言, 如何利用 CCA 的优势设计更通用有效的学习框架, 是值得深入探讨的话题。

(2) 绝大部分基于 CCA 的多视图学习方法涉及矩阵分解和矩阵求逆的计算, 该过程往往是导致计算耗时的主要原因。基于矩阵分解的求解方式的计算量通常与特征维度直接相关。研究计算量更小的、更优化的求解方式能够直接提高基于 CCA 的多视图学习方法的运算效率。

(3) 从数据的角度来看, 将多个视图拼接为单个视图的方式和直接使用多个视图的方式相比, 缺失的信息只在于不同视图间的区分。因此, 如何将单视图数据视图划分为多视图数据, 对于有效地利用视图间的差异性和互补信息, 降低视图中与任务无关的噪声而言至关重要。

(4) 根据实验, 在某些情形下使用多视图的方法可能不如直接使用单视图的方法, 甚至不如直接使用单个视图的方法。因此, 需要根据视图的信息量, 设计合理的方法和指标用于快速判断数据集更适用于多视图还是单视图学习方法, 此外, 研究如何剔除无用的视图, 对于多视图方法的应用实践而言同样具有重要的意义。

参考文献:

- [1] 王建民. 工业大数据技术综述[J]. 大数据, 2017, 3(6): 3-14.
WANG J M. Survey on industrial big data[J]. Big Data Research, 2017, 3(6): 3-14.
- [2] 郭斌, 翟书颖, 於志文, 等. 群智大数据: 感知、优选与理解[J]. 大数据, 2017, 3(5): 57-69.
GUO B, ZHAI S Y, YU Z W, et al. Crowdsensing big data: sensing, data selection, and understanding[J]. Big Data Research, 2017, 3(5): 57-69.
- [3] 李娜, 李永超, 邹颖, 等. 基于多模态影像下的抑郁症大脑异常[J]. 智能科学与技术学报, 2020, 2(2): 116-125.
LI S, LI Y C, ZOU Y, et al. Brain abnormalities in depression based on multimodal imaging[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(2): 116-125.
- [4] 梁吉业, 钱宇华, 李德玉, 等. 面向大数据的粒计算理论与方法研究进展[J]. 大数据, 2016, 2(4): 13-23.
LIANG J Y, QIAN Y H, LI D Y, et al. Research development on granular computing theory and method for big data[J]. Big Data Research, 2016, 2(4): 13-23.
- [5] TANG J J, TIAN Y J, ZHANG P, et al. Multiview privileged support vector machines[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(8): 3463-3477.
- [6] XU C, TAO D C, XU C. A survey on multi-view learning[J]. arXiv preprint, 2013, arXiv:1304.5634.
- [7] XUE X, NIE F, WANG S, et al. Multi-view correlated feature learning by uncovering shared component[C]//Proceedings of the 31th AAAI Conference on Artificial Intelligence. CA: AAAI, 2017: 2810-2816.
- [8] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.
- [9] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory. New York: ACM Press, 1998: 92-100.
- [10] ABNEY S. Bootstrapping[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. PA: ACL, 2002: 360-367.
- [11] GEHLER P, NOWOZIN S. On feature combination for multiclass object classification[C]//Proceedings of 2009 IEEE 12th International Conference on Computer Vision. Piscataway: IEEE Press, 2009: 221-228.
- [12] GONEN M, ALPAYDIN E. Multiple kernel learning algorithms[J]. Journal of Machine Learning Research, 2011, 12: 2211-2268.
- [13] HOTELLING H. Relations between two sets of variates[J]. Biometrika, 1936, 28(3/4): 321-377.
- [14] DIETHE T, HARDOON D R, SHAW-TAYLOR J. Multiview fisher discriminant analysis[C]//Proceedings of the Neural Information Processing Systems Workshop on Learning from Multiple Sources. Cambridge: MIT Press, 2008: 1-8.
- [15] XIA T, TAO D C, MEI T, et al. Multiview spectral embedding[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2010, 40(6): 1438-1446.
- [16] ZHAO J, XIE X J, XU X, et al. Multi-view learning overview: recent progress and new challenges[J]. Information Fusion, 2017, 38: 43-54.
- [17] SUN S L. A survey of multi-view machine learning[J]. Neural Computing and Applications, 2013, 23(7): 2031-2038.
- [18] ZHANG T, DENG Z H, WU D R, et al. Multiview fuzzy logic system with the cooperation between visible and hidden views[J]. IEEE Transactions on Fuzzy Systems, 2019, 27(6): 1162-1173.
- [19] CHAUDHURI K, KAKADE S M, LIVESCU K, et al. Multi-view clustering via canonical correlation analysis[C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 129-136.
- [20] BLASCHKO M B, LAMPERT C H. Correlational spectral clustering[C]//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2008: 1-8.
- [21] KAKADE S M, FOSTER D P. Multi-view regression via canonical correlation analysis[M]//Learning theory. Heidelberg: Springer, 2007: 82-96.
- [22] ZHANG Y Y, ZHANG J C, PAN Z S, et al. Multi-view dimensionality reduction via canonical random correlation analysis[J]. Frontiers of Computer Science, 2016, 10(5): 856-869.
- [23] WU G, LI F. A randomized exponential canonical correlation analysis method for data analysis and dimensionality reduction[J]. Applied Numerical Mathematics, 2021, 164: 101-124.
- [24] LAI P L, FYFE C. Kernel and nonlinear canonical correlation analysis[J]. International Journal of Neural Systems, 2000, 10(5): 365-377.
- [25] SUN T K, CHEN S C. Locality preserving CCA with applications to data visualization and pose estimation[J]. Image and Vision Computing, 2007, 25(5): 531-543.
- [26] WANG W R, ARORA R, LIVESCU K, et al. On deep multi-view representation learning[C]//Proceedings of the 32th International Conference on Machine Learning. Cambridge: JMLR, 2015: 1083-1092.
- [27] ANDREW G, ARORA R, BILMES J, et al. Deep canonical correlation analysis[C]//Proceedings of the 30th International Conference on Machine Learning. Cambridge: JMLR, 2013: 1247-1255.
- [28] WANG W R, YAN X C, LEE H, et al. Deep variational canonical correlation analysis[J]. arXiv preprint, 2016, arXiv:1610.03454.
- [29] BENTON A, ARORA R, DREDZE M. Learning multiview embeddings of twitter users[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 14-19.
- [30] EL DIN ELMADANY N, HE Y F, GUAN L. Multiview learning via deep discriminative canonical correlation analysis[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2016: 2409-2413.
- [31] LUO Y, TAO D C, RAMAMOHANARAO K, et al. Tensor canonical correlation analysis for multi-view dimension reduction[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(11): 3111-3124.
- [32] VÍA J, SANTAMARÍA I, PÉREZ J. A learning algorithm for adaptive canonical correlation analysis of several data sets[J]. Neural Networks, 2007, 20(1): 139-152.

- [33] CARROLL J D. Generalization of canonical correlation analysis to three or more sets of variables[C]//Proceedings of the 76th Annual Convention of the American Psychological Association. [S.l.:s.n.], 1968: 227-228.
- [34] SUN T K, CHEN S C, YANG J Y, et al. A novel method of combined feature extraction for recognition[C]//Proceedings of 2008 8th IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2008: 1043-1048.
- [35] SUN S L, XIE X J, YANG M. Multiview uncorrelated discriminant analysis[J]. IEEE Transactions on Cybernetics, 2016, 46(12): 3272-3284.
- [36] FISHER R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics, 1936, 7(2): 179-188.
- [37] HARDOON D R, SZEDMAK S, SHAW-TAYLOR J. Canonical correlation analysis: an overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639-2664.
- [38] WITTEN D M, TIBSHIRANI R J. Extensions of sparse canonical correlation analysis with applications to genomic data[J]. Statistical Applications in Genetics and Molecular Biology, 2009, 8(1): 1-27.
- [39] WITTEN D M, TIBSHIRANI R, HASTIE T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis[J]. Biostatistics, 2009, 10(3): 515-534.
- [40] CHEN X. Structured sparse canonical correlation analysis[C]//Proceedings of the 15th International Conference Artificial Intelligence and Statistics. New York: ACM Press, 2012: 199-207.
- [41] LIN D, ZHANG J, LI J, et al. Group sparse canonical correlation analysis for genomic data integration[J]. BMC Bioinformatics, 2013, 14: 245.
- [42] LIN D D, CALHOUN V D, WANG Y P. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis[J]. Medical Image Analysis, 2014, 18(6): 891-902.
- [43] LOPEZ-PAZ D, SRA S, SMOLA A, et al. Randomized nonlinear component analysis[C]//Proceedings of the 31st International Conference on Machine Learning. Cambridge: JMLR, 2014: 1359-1367.
- [44] HOEGAERTS L, SUYKENS J A K, VANDEWALLE J, et al. Subset based least squares subspace regression in RKHS[J]. Neurocomputing, 2005, 63: 293-323.
- [45] YANG J F, ZHANG X. Feature-level fusion of fingerprint and finger-vein for personal identification[J]. Pattern Recognition Letters, 2012, 33(5): 623-628.
- [46] WANG F S, ZHANG D Q. A new locality-preserving canonical correlation analysis algorithm for multi-view dimensionality reduction[J]. Neural Processing Letters, 2013, 37(2): 135-146.
- [47] TIAN Y, SIGAL L, DE LA TORRE F, et al. Canonical locality preserving latent variable model for discriminative pose inference[J]. Image and Vision Computing, 2013, 31(3): 223-230.
- [48] WANG W R, ARORA R, LIVESCU K, et al. Unsupervised learning of acoustic features via deep canonical correlation analysis[C]//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2015: 4590-4594.
- [49] BACH F R, JORDAN M I. A probabilistic interpretation of canonical correlation analysis[R]. 2005.
- [50] KINGMA D P, WELING M. Auto-encoding variational bayes[J]. arXiv preprint, 2013, arXiv:1312.6114.
- [51] KLAMI A, VIRTANEN S, KASKI S. Bayesian canonical correlation analysis[J]. Journal of Machine Learning Research, 2013, 14(1): 965-1003.
- [52] WANG C. Variational Bayesian approach to canonical correlation analysis[J]. IEEE Transactions on Neural Networks, 2007, 18(3): 905-910.
- [53] VIRTANEN S, KLAMI A, KASKI S. Bayesian CCA via group sparsity[C]//Proceedings of the 28th International Conference on Machine Learning. Madison: Omnipress, 2011: 457-464.
- [54] HORST P. Generalized canonical correlations and their applications to experimental data[J]. Journal of Clinical Psychology, 1961, 17(4): 331-347.
- [55] BOUMAL N, MISHRA B, ABSIL P A, et al. Manopt, a Matlab toolbox for optimization on manifolds[J]. Journal of Machine Learning Research, 2014, 15(1):1455-1459.
- [56] ASENDORF N A. Informative data fusion: beyond canonical correlation analysis[D]. Ann Arbor: The University of Michigan, 2015.
- [57] BENTON A, KHAYRALLAH H, GUJRAL B, et al. Deep generalized canonical correlation analysis[C]//Proceedings of the 4th Workshop on Representation Learning for NLP. Stroudsburg: Association for Computational Linguistics, 2019: 1-6.
- [58] ZHENG W M, ZHOU X Y, ZOU C R, et al. Facial expression recognition using kernel canonical correlation analysis (KCCA)[J]. IEEE Transactions on Neural Networks, 2006, 17(1): 233-238.
- [59] SHAN C F, GONG S G, MCOWAN P W. Fusing gait and face cues for human gender recognition[J]. Neurocomputing, 2008, 71(10/11/12): 1931-1938.
- [60] XING X L, WANG K J, YAN T, et al. Complete canonical correlation analysis with application to multi-view gait recognition[J]. Pattern Recognition, 2016, 50: 107-117.
- [61] ARORA R, LIVESCU K. Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2013: 7135-7139.
- [62] ZHUANG X W, YANG Z S, CORDES D. A technical review of canonical correlation analysis for neuroscience applications[J]. Human Brain Mapping, 2020, 41(13): 3807-3833.
- [63] YONG S, CHEONG P. Analysis of correlation based dimension reduction methods[J]. International Journal of Applied Mathematics and Computer Science, 2011, 21(3): 549-558.
- [64] ELMADANY N E D, HE Y F, GUAN L. Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis[J]. IEEE Transactions on Image Processing, 2018, 27(11): 5275-5287.
- [65] CHI J, HUANG X, PENG Y. Zero-shot cross-media retrieval with external knowledge[C]//Proceedings of the 9th International Conference on Internet Multimedia Computing and Service. Singapore: Springer, 2017: 200-211.
- [66] RASIWASIA N, PEREIRA J C, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]//Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press, 2010: 251-260.
- [67] PENG Y X, HUANG X, ZHAO Y Z. An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges[J].

- IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(9): 2372-2385.
- [68] WU F, ZHANG H, ZHUANG Y T. Learning semantic correlations for cross-media retrieval[C]//Proceedings of 2006 International Conference on Image Processing. Piscataway: IEEE Press, 2006: 1465-1468.
- [69] CHANDAR S, KHAPRA M M, LAROCHELLE H, et al. Correlation-al neural networks[J]. Neural Computation, 2016, 28(2): 257-285.
- [70] DHILLON P, FOSTER D, UNGAR L. Multi-view learning of word embeddings via CCA[C]//Proceedings of the 25th Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2011: 199-207.
- [71] FARUQUI M, DYER C. Improving vector space word representations using multilingual correlation[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2014: 462-471.
- [72] CHEN Z W, YANG C H, PENG T, et al. A cumulative canonical correlation analysis-based sensor precision degradation detection method[J]. IEEE Transactions on Industrial Electronics, 2019, 66(8): 6321-6330.
- [73] HUANG H, HE H T, FAN X, et al. Super-resolution of human face image using canonical correlation analysis[J]. Pattern Recognition, 2010, 43(7): 2532-2543.
- [74] AN L, BHANU B. Face image super-resolution using 2D CCA[J]. Signal Processing, 2014, 103: 184-194.
- [75] JIANG Q C, YAN X F. Multimode process monitoring using variational

Bayesian inference and canonical correlation analysis[J]. IEEE Transactions on Automation Science and Engineering, 2019, 16(4): 1814-1824.

- [76] SONG B, SHI H, TAN S, et al. Multi-subspace orthogonal canonical correlation analysis for quality related plant wide process monitoring[J]. IEEE Transactions on Industrial Informatics, 2020.

[作者简介]



郭陈凤（1995—），女，华中科技大学人工智能与自动化学院硕士生，主要研究方向为脑机接口、机器学习。



伍冬睿（1982—），男，博士，华中科技大学人工智能与自动化学院教授，主要研究方向为机器学习、脑机接口、计算智能、情感计算。