数据场典型相关分析及其在图像分割中的应用

李文平 [1,2] 杨 静 [1] 印桂生 [1] 张健沛 [1]

摘 要 针对数据场环境下多维数据的低维特征提取问题, 本文将数据之间的相互作用纳入其相关性求解中, 提出一种基于数据场的典型相关分析 (Data field based canonical correlation analysis, DFCCA) 方法. DFCCA 提取的特征具有良好的分布特性, 原空间上相隔较远的数据点对的特征聚集在一个较小区域内, 而相邻数据点对的特征却有规律地分布在其他点所聚集区域的周围. 此特性使得 DFCCA 具有较好的边界辨识能力, 将其应用于图像分割的实验结果表明, DFCCA 提取的复杂图像边界具有较好的保真度.

关键词 典型相关分析, 数据场, 特征提取, 图像分割

引用格式 李文平, 杨静, 印桂生, 张健沛. 数据场典型相关分析及其在图像分割中的应用. 自动化学报, 2015, **41**(4): 772−784

DOI 10.16383/j.aas.2015.c130896

Data Field Based Canonical Correlation Analysis and Its Application to Image Segmentation

LI Wen-Ping[1,2] YANG Jing[1] YIN Gui-Sheng[1] ZHANG Jian-Pei[1]

Abstract In this paper, for extracting low-dimensional features from multi-dimensional data in data field environment, we propose a novel method of canonical correlation analysis (CCA) called DFCCA (data field based CCA) by introducing interactions among data into data correlation solving. The features extracted by DFCCA have better distribution properties, that is the features corresponding to a data point pair that are far apart from each other gather together in a small region, but other features corresponding to the pair of data points that are neighboring each other will scatter regularly around the region. Thanks to these properties, DFCCA has a good capability of frontier identification. Experimental results on image segmentation demonstrate that the frontiers extracted from complex images by DFCCA hold better fidelity.

Key words Canonical correlation analysis (CCA), data field, feature extraction, image segmentation

Citation Li Wen-Ping, Yang Jing, Yin Gui-Sheng, Zhang Jian-Pei. Data field based canonical correlation analysis and its application to image segmentation. *ACTA Automatica Sinica*, 2015, **41**(4): 772−784

典型相关分析 (Canonical correlation analysis, CCA) 是研究两组变量之间线性相关关系的一种著名的多元统计分析方法, 最早由 Hotelling 于 1936 年提出[1]. CCA 将两组随机变量 (向量) 之间的线性相关性转化为少数几对互不相关的随机变量 (向量) 的线性相关性, 目标在于寻找两组基, 使得原向量在此基上投影的 Pearson 相关系数达到最大值. 因此 CCA 不仅是一种检测两组多维数据相关性的基本工具, 而且也是一种常用的数据降维技术. CCA 在聚类[2]、分类[3−4]、特征融合[5−6]、图像处理[7]、模式识别[8−9]、回归分析[10−11] 及气象预测[12] 等领域得到了广泛成功应用.

与应用研究一样, CCA 的理论研究也是学者们持续关注的热点课题. 目前, CCA 研究主要向非线性化、多集化和快速化等方向发展.

经典 CCA 在检测两组变量之间的线性相关关系时具有较好效果, 但在检测数据间的非线性相关关系时能力较弱, 于是非线性 CCA 成为 CCA 发展的一个重要分支. CCA 非线性化技术主要包括核方法、神经网络和流形学习三种.

基于核方法的 CCA, 即 KCCA (Kernel CCA), 将数据映射到特征空间 (往往是高维的), 用特征空间中的线性相关关系近似原空间 (往往是低维的) 中的非线性相关关系[13]. KCCA 能提取蕴含于数据中

收稿日期 2013-09-16 录用日期 2014-11-21
Received September 16, 2013; accepted November 21, 2014
国家自然科学基金 (61370083, 61073043, 61073041, 61402126), 高等学校博士学科点专项科研基金 (20112304110011, 20122304110012) 资助
Supported by National Natural Science Foundation of China (61370083, 61073043, 61073041, 61402126), and Research Fund for the Doctoral Program of Higher Education of China (20112304110011, 20122304110012)
本文责任编委 章毓晋
Recommended by Associate Editor ZHANG Yu-Jin
1. 哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001 2. 嘉兴学院数理与信息工程学院 嘉兴 314001
1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001 2. College of Mathematics Physics and Information Engineering, Jiaxing University, Jiaxing 314001

的非线性低维特征, 它被成功应用于表情识别[14]、fMRI (Functional magnetic resonance imaging) 数据分析[15] 及数据降维[16] 等领域. 然而 KCCA 核参数的选择较困难, 但基于神经网络的方法却不存在此缺陷[17]. 由于神经网络具有较强的学习能力, 在含噪声的数据上, 基于神经网络的 CCA 提取的低维非线性特征较理想[18]. 不过, 基于神经网络的 CCA 需要训练数据集进行学习且速度较慢, 但基于流形学习的方法可避免此不足. 基于流形学习的 CCA 是较新的研究方向, 目前成果主要是基于局部保持思想的扩展模型 LPCCA (Locality-preserving CCA)[19] 及其变种 ALPCCA (A new LPCCA)[20], 该方法在寻找 CCA 的投影基向量时, 考虑了局部邻域结构, 通过局部的线性相关近似全局的非线性关系.

CCA 理论研究的第二个重要分支是多集 CCA, 即 MCCA (Multi-set CCA). 与上述 CCA 仅涉及两组变量不同, MCCA 的目标在于同时分析多组随机变量的相关性, 并提取其低维特征[21]. 由于 MCCA 同时检测多组变量的相关性, 这满足很多现实应用中需同时考虑多组数据的基本条件, 因此 MCCA 得到了广泛应用, 如医疗图像分析[22]、水下目标分类[23]、盲源信号分离[24]、多目标特征融合[25] 以及神经功能链接分析[26] 等.

最近, 随着大数据及数据流研究的兴起, CCA 的快速化问题吸引了部分学者的关注[27], 旨在提高 CCA 计算效率的第三个研究分支已初见端倪. 文献 [28] 将 CCA 转化为一个等价的最小二乘问题, 能快速检测高维大数据间的相关性. 面向数据流的 CCA 扩展模型是目前的一个主攻方向, 文献 [29] 采用不等概抽样技术形成低阶概要矩阵, 在此基础上计算多维数据流之间的典型相关系数; 文献 [30−31] 采用不等概抽样技术, 基于低阶近似理论, 提出适于数据流处理的多变量相关性分析方法; 文献 [32] 研究了一种 CCA 增量式学习算法, 并将其应用于视频跟踪领域; 文献 [33] 基于 GPU, 研究了一种面向数据流的 CCA 并行处理框架; 文献 [34] 基于秩 2 更新理论, 提出了一种面向数据流的 CCA 快速跟踪算法 TCCA.

上述 CCA 方法在各自领域取得较好效果, 但它们在提取低维特征时未考虑数据之间的相互作用. 事实上, 世界是普遍联系的, 数据之间同物理世界的粒子之间一样理应也存在相互作用, 这是数据场理论的基本观点[35]. 数据场理论是我国学者李德毅院士提出的一种新兴的不确定性人工智能方法, 它是受磁场、电场、重力场等物理场概念启发而提出的. 众所周知, 场在物理学中用于刻画物质粒子间一种非接触式相互作用, 反映了粒子间的一种联系. 数据场理论将数据视为具有质量的 “粒子”, 在数据分布的空间中, 不同 “粒子” 间的相互作用形成数据场. 在数据场环境下, 由于数据点受到所处数据场的作用, CCA 提取的低维特征是否会表现出独特的性质还不得而知, 如何将数据之间的相互作用纳入其相关性求解中是一个具有重要意义的研究课题.

本文从数据间相互作用的视角出发, 引入数据场理论, 拟研究一种新的 CCA 方法, 以提取数据场环境下多维数据的低维特征. 本文创新性工作包括: 1) 针对两个数据集相互作用形成的数据场, 提出数据质点、数据场点、互点场、互势值等概念. 互点场刻画了一个数据集对另一个数据集的作用, 而互势值是其作用强度的度量. 2) 为求解不同维数据集形成的互数据场中数据场点的互势值, 提出一种不同维向量间的距离计算公式, 称为向量的广义伪距离. 向量间的 Minkowski 距离以及欧氏距离是向量的广义伪距离的特例. 3) 将数据集间的相互作用纳入其相关性求解中, 提出一种基于数据场的 CCA 方法 DFCCA (Data field CCA). DFCCA 提取的特征具有良好的分布特性, 原空间上相隔较远的数据点对的特征聚集在一个较小区域内, 而原空间上相邻数据点对的特征在特征空间中却有规律地分布在其他点所聚集区域的周围. 此特性使得 DFCCA 具有较好的边界辨识能力, 将其应用于图像分割的实验结果表明, DFCCA 具有较强的图像分割能力, 所提取的复杂图像边界具有较好的保真度. 尽管 CCA 在图像降维和特征提取方面得到了广泛应用, 但就笔者所掌握的资料看, 还未发现将其应用到图像分割的相关报告, 本研究拟在此领域进行初探.

1　基础知识回顾

1.1　CCA

给定 p 维随机向量 $\boldsymbol{X}$ 和 q 维随机向量 $\boldsymbol{Y}$, $p \le q$, CCA 的目标是寻找投影向量 $\boldsymbol{\alpha}_k$ 和 $\boldsymbol{\beta}_k$, 使得在方差 $\mathrm{Var}(\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{X}) = \mathrm{Var}(\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Y}) = 1$ 的约束下, Pearson 相关系数

$$\rho(\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Y}) = \frac{\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{C}_{xy}\boldsymbol{\beta}_k}{\sqrt{(\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{C}_{xx}\boldsymbol{\alpha}_k)(\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{C}_{yy}\boldsymbol{\beta}_k)}} \tag{1}$$

达到最大值. 其中, $\boldsymbol{C}_{xy} = \boldsymbol{C}_{yx}^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ 为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的互协方差矩阵, 而 $\boldsymbol{C}_{xx} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 和 $\boldsymbol{C}_{yy} = \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ 分别为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的自协方差矩阵. 称 $\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{X}$ 和 $\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Y}$ 为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的第 k 对典型相关变量, 其相关系数称为第 k 个典型相关系数.

CCA 实质是一个最优化问题. 以第 1 对典型变

量为例 (省略 $\boldsymbol{\alpha}_1$ 和 $\boldsymbol{\beta}_1$ 的下标), 即求:

$$\begin{aligned}&\max_{\boldsymbol{\alpha}\in\mathbf{R}^{p\times1},\boldsymbol{\beta}\in\mathbf{R}^{q\times1}} \quad \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{C}_{xy}\boldsymbol{\beta}\\&\text{s.t.} \quad \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{C}_{xx}\boldsymbol{\alpha}=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{C}_{yy}\boldsymbol{\beta}=1\end{aligned} \tag{2}$$

其中, $\mathbf{R}$ 为实数域. 用拉格朗日乘数法 (Lagrange) 求解式 (2) 有:

$$\begin{aligned}&\boldsymbol{\beta}=\frac{1}{\lambda}\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{\alpha}\\&\boldsymbol{C}_{xy}\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{\alpha}=\lambda^2\boldsymbol{C}_{xx}\boldsymbol{\alpha}\end{aligned} \tag{3}$$

式 (3) 第二式是一个广义特征值问题, 由此解出 $\boldsymbol{\alpha}$ 和 λ, 代入第一式可得 $\boldsymbol{\beta}$. λ 即为所求的典型相关系数.

1.2 数据场

场是物理学的一个基本概念, 用于描述物质粒子间非接触的相互作用. 迄今为止, 物理学家所发现的 4 种相互作用中, 除弱相互作用外, 万有引力、电磁力和强相互作用都产生场. 一般地, 传递物质粒子间相互作用的中间媒质即是场. 如果空间 Ω 中每个点都对应某个物理量或数学函数的一个确定值, 则称在 Ω 上确定了该物理量或数学函数的一个场[35].

如果将 Ω 定义为由某给定的数据集 $\boldsymbol{D}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 所张成的空间, 其中 $\boldsymbol{x}_i=(x_{1i},x_{2i},\cdots,x_{pi})^{\mathrm{T}}$, $i=1,2,\cdots,n$, 并将任意元素 $\boldsymbol{x}_i$ 视作一个具有一定质量的粒子, 且认为 $\boldsymbol{x}_i$ 会对周围其他数据元素产生影响, 那么 $\boldsymbol{D}$ 中数据元素的相互作用便可在 Ω 上形成一个虚拟的场, 即数据场.

势场是受到广泛关注的一类稳定有源场. 势是指把单位质点从场中一点移动到参考点过程中场力所做的功. 势场的分布对应着相互作用的粒子之间由相对位置所确定的势能分布. 空间任意点的势值大小用势函数度量. 拟重力场势函数和拟核力场势函数是数据场理论中两种重要的势函数, 其定义分别如下[35]:

$$\varphi_{\boldsymbol{x}}(\boldsymbol{y})=\frac{m}{1+(\frac{||\boldsymbol{x}-\boldsymbol{y}||}{\sigma})^k} \tag{4}$$

$$\varphi_{\boldsymbol{x}}(\boldsymbol{y})=m\times\exp\left(-\left(\frac{||\boldsymbol{x}-\boldsymbol{y}||}{\sigma}\right)^k\right) \tag{5}$$

其中, $m\geq0$ 代表场源强度, 一般被视为数据对象 $\boldsymbol{x}$ 的质量; $\sigma\in(0,+\infty)$ 称为影响因子, 用于控制数据对象之间的相互作用力程; 而 k 称为距离指数. 势场中任意一点 $\boldsymbol{y}$ 的势值 $\varphi(\boldsymbol{y})$ 是所有数据对象在其上所产生势值的代数和, 即[35]:

$$\varphi(\boldsymbol{y})=\sum_{\boldsymbol{x}\in\boldsymbol{D}}\varphi_{\boldsymbol{x}}(\boldsymbol{y}) \tag{6}$$

2 基于数据场的 CCA

本节从数据间相互作用的视角出发, 基于数据场理论, 提出一种新的典型相关分析方法 DFCCA (Data field CCA) . 下面先定义相关概念, 再阐述 DFCCA 的若干细节.

2.1 相关概念定义

传统数据场理论主要是针对单个数据集而言的, 而 CCA 却需同时考虑两个数据集, 因此有必要扩展传统数据场理论的部分概念. 本节针对两个数据集相互作用形成的数据场, 对数据质点、数据场点、互点场、互势值等作出界定.

定义1 (数据质点). 形成数据场的数据点称为数据质点.

定义 2 (数据场点). 数据场所分布的空间中任意位置对应的点称为数据场点.

定义 3 (点场). 由有限个数据质点在有限个数据场点上形成的数据场称为数据点场, 简称点场.

定义 4 (自点场). 同时以 $\boldsymbol{X}$ 中的点 $\{\boldsymbol{x}_i\}_{i=1}^n$ 为数据质点和数据场点的点场称为 $\boldsymbol{X}$ 上的自点场.

定义 5 (互点场). 以 $\boldsymbol{X}$ 中的点 $\{\boldsymbol{x}_i\}_{i=1}^n$ 为数据场点, 而以 $\boldsymbol{Y}$ 中的点 $\{\boldsymbol{y}_i\}_{i=1}^m$ 为数据质点的点场称为 $\boldsymbol{Y}$ 在 $\boldsymbol{X}$ 上的互点场.

定义 6 (自势值). $\boldsymbol{X}$ 上的自点场中, 数据质点 $\{\boldsymbol{x}_j\}_{j=1}^n$ 在数据场点 $\boldsymbol{x}_i$ 处的势值称为 $\boldsymbol{x}_i$ 在 $\boldsymbol{X}$ 上的自势值, 记为 $\varphi_{\boldsymbol{x}}(\boldsymbol{x}_i)=\sum_{j=1}^n\varphi_{\boldsymbol{x}_j}(\boldsymbol{x}_i)$. 其中, $\varphi_{\boldsymbol{x}_j}(\boldsymbol{x}_i)$ 为势函数.

定义 7 (互势值). $\boldsymbol{Y}$ 在 $\boldsymbol{X}$ 上的互点场中, $\boldsymbol{Y}$ 中的数据质点 $\{\boldsymbol{y}_j\}_{j=1}^n$ 在 $\boldsymbol{X}$ 中的数据场点 $\boldsymbol{x}_i$ 处形成的势值称为互势值, 记为 $\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)=\sum_{j=1}^n\varphi_{\boldsymbol{y}_j}(\boldsymbol{x}_i)$. 其中, $\varphi_{\boldsymbol{y}_j}(\boldsymbol{x}_i)$ 为势函数.

互点场和互势值是 DFCCA 的两个基本概念. 当 $\boldsymbol{Y}=\boldsymbol{X}$ 时, 互点场即为自点场, 而互势值即为自势值. 自势值是同一数据集的数据点所形成的势值, 式 (6) 所示的传统势值即为自势值. 因此互势值将数据场理论的势值概念从一个数据集拓展到两个数据集, 它刻画了一个数据集对另一个数据集中数据点作用的强度.

由于自势值即为传统势值, 因此自势值定义中的势函数 $\varphi_{\boldsymbol{x}_j}(\boldsymbol{x}_i)$ 可以直接取已有势函数, 如式 (4) 所示的拟重力场势函数或式 (5) 所示的拟核力场势函数等. 然而, 将已有势函数用于互势值定义中的势函数 $\varphi_{\boldsymbol{y}_j}(\boldsymbol{x}_i)$ 却存在困难, 其根源在于传统势函数中数据点之间距离的定义. 如果将数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 的坐标视为向量, 传统势函数中数据点之间距离 $\left\|\boldsymbol{x}_i-\boldsymbol{y}_j\right\|$ 要求向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 具有相同维数 (即相同数目的元素个数), 然而此条件在互点场定义中不满足, 求势函数 $\varphi_{\boldsymbol{y}_j}(\boldsymbol{x}_i)$ 时, $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 的维数可能不等.

因此, 为求解不同维数据集形成的互数据场中数据场点的互势值, 有必要研究一种不同维向量之间距离的计算方法, 下面另辟一节阐述此问题.

2.2　向量的广义伪距离

在求定义 7 所述的互势值时, 势函数 $\varphi_{\boldsymbol{y}_j}(\boldsymbol{x}_i)$ 的计算中数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 的维数可能不等, 因为它们分别来自维数可以不同的两个数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$. 这种维数的差异造成势函数 $\varphi_{\boldsymbol{y}_j}(\boldsymbol{x}_i)$ 求解过程中点 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 之间距离计算的困难. 传统距离是针对同维向量而定义的, 如欧氏距离、Minkowski 距离以及统计距离等, 当面对两个维数不等的向量时, 由于其分量无法一一对应, 传统距离定义不再适用. 为采用传统势函数计算互势值, 如何求不同维数据点 (向量) 间的距离是拟解决的关键问题. 本节提出一种不同维向量的距离计算公式, 称为向量的广义伪距离.

定义 8 (向量的广义伪距离). p 维实向量 $\boldsymbol{x}$ 和 q 维实向量 $\boldsymbol{y}$ 之间的广义伪距离定义为

$$d(\boldsymbol{x},\boldsymbol{y})=\begin{cases}\left(\dfrac{1}{r+t}\displaystyle\sum_{i=1}^{p}\sum_{j=(i-1)r+1}^{ir+t}|x_i-y_j|^s\right)^{\frac{1}{s}}, p\le q\\ \left(\dfrac{1}{r+t}\displaystyle\sum_{i=1}^{q}\sum_{j=(i-1)r+1}^{ir+t}|y_i-x_j|^s\right)^{\frac{1}{s}}, p>q\end{cases}\tag{7}$$

其中, $r=\left\lfloor\frac{\max\{p,q\}}{\min\{p,q\}}\right\rfloor$, $t=\max\{p,q\}\bmod\min\{p,q\}$, $\lfloor\cdot\rfloor$ 为向下取整, "mod" 为求余数, $s\in\mathbf{Z}^+$ 为正整数.

由式 (7) 可知, $d(\boldsymbol{x},\boldsymbol{y})$ 满足非负性, 即 $d(\boldsymbol{x},\boldsymbol{y})\ge 0$, 而且 $d(\boldsymbol{x},\boldsymbol{y})=0$ 当且仅当 $p\le q$ 时分量 $x_i=y_j$ 或 $p>q$ 时分量 $y_i=x_j$, $i=1,\cdots,\min\{p,q\}$, $j=(i-1)r+1,\cdots,ir+t$; 此外, $d(\boldsymbol{x},\boldsymbol{y})$ 还满足对称性, 即 $d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x})$.

式 (7) 定义的 $d(\boldsymbol{x},\boldsymbol{y})$ 之所以称为向量的广义伪距离, 是鉴于两个方面的考虑: 1) $d(\boldsymbol{x},\boldsymbol{y})$ 是 Minkowski 距离在两个不同维向量间的推广. 当 $p=q$ 时, $r=1$, $t=0$, 有 $d(\boldsymbol{x},\boldsymbol{y})=\left(\sum_{i=1}^{p}\sum_{j=i}^{i}|x_i-y_j|^s\right)^{1/s}=\left(\sum_{i=1}^{p}|x_i-y_j|^s\right)^{1/s}$, 即 $d(\boldsymbol{x},\boldsymbol{y})$ 为 Minkowski 距离; 当 $p=q$ 且 $s=2$ 时, $d(\boldsymbol{x},\boldsymbol{y})$ 为欧氏距离. 从这个意义上说, $d(\boldsymbol{x},\boldsymbol{y})$ 是一种广义距离. 2) 尽管 $d(\boldsymbol{x},\boldsymbol{y})$ 满足非负性和对称性, 但却不满足传统距离测度中的三角不等式性质, 即对任意不同维向量, 不等式 $d(\boldsymbol{x},\boldsymbol{y})\le d(\boldsymbol{x},\boldsymbol{z})+d(\boldsymbol{z},\boldsymbol{y})$ 不一定满足, 一个反例是 $s=2$, $\boldsymbol{x}=[1,1,3]^{\mathrm{T}}$, $\boldsymbol{z}=[4,3]^{\mathrm{T}}$, $\boldsymbol{y}=[4,5,4,1]^{\mathrm{T}}$ 时, $d(\boldsymbol{x},\boldsymbol{y})=5.24$, $d(\boldsymbol{x},\boldsymbol{z})=3.32$, $d(\boldsymbol{z},\boldsymbol{y})=1.73$, 有 $5.24>3.32+1.73=5.05$. 故定义中加上 "伪" 字.

式 (7) 定义的广义伪距离的几何意义是明确的. 图 1 为 $p\le q$ 且 $s=2$ 的条件下, 求向量的广义伪距离时各分量对应关系的三个实例. 图 1 (a) 中 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的维数都为 2, 分量 x_i 与 y_i 一一对应, 此时按式 (7) 计算出的即为欧氏距离; 图 1 (b) 中 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的维数分别为 2 和 3, $\boldsymbol{x}$ 的 1 个分量与 $\boldsymbol{y}$ 的 2 个分量对应, 其中分量 y_2 同时对应 x_1 和 x_2; 图 1 (c) 中 $\boldsymbol{y}$ 的维数恰好是 $\boldsymbol{x}$ 的维数的 2 倍, $\boldsymbol{x}$ 的 1 个分量与 $\boldsymbol{y}$ 的 2 个分量对应, 而 $\boldsymbol{y}$ 的分量无重叠对应. 简言之, 式 (7) 所示定义先将维数大的向量的分量按序均分为分量可重叠的若干段, 段的数目为另一向量的维数 (维数小); 再将维数小的向量的每个分量 x_i 与另一向量中所划分出的段按序对应; 最后求 x_i 与对应段中各分量差的绝对值, 并计算绝对值的 s 次方的均值, 将所有均值和的 s 次方根作为两向量的距离.

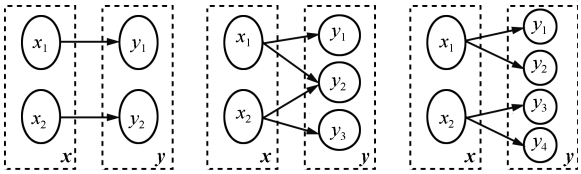


图 1　向量的广义伪距离求解时各分量对应关系实例

Fig. 1　The correspondence of components between two vectors which dimensions may be different from each other in solving their distance

2.3　DFCCA 基本思想及问题描述

对于来自数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的样本点对 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$, CCA 用离均差之积 $(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^{\mathrm{T}}$ 刻画其相关性, 其中 $\bar{\boldsymbol{x}}$ 和 $\bar{\boldsymbol{y}}$ 分别为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的均值向量 (下同). 当数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 给定后, $\bar{\boldsymbol{x}}$ 和 $\bar{\boldsymbol{y}}$ 为常向量, 可见经典 CCA 在刻画点对 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ 的相关性时并未考虑数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_i$ 与其他数据点之间的联系.

数据场理论同物理学中的场论一样, 认为数据点之间像物质粒子之间那样是普遍联系的, 数据场作为一种非接触式的相互作用是这种联系的一种具体形式. 因此, 当从数据点相互作用的视角出发, 在数据场环境中研究数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相关性时, 需要考察每个点对 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ 如何受其他数据点作用, 以及这种作用如何影响其相关性.

DFCCA 的基本思想在于, 将数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相互作用纳入其相关性求解中. 一方面, 以 $\boldsymbol{Y}$ 中的数据点 $\{\boldsymbol{y}_i\}_{i=1}^m$ 为数据质点, 而以 $\boldsymbol{X}$ 中的数据点 $\{\boldsymbol{x}_i\}_{i=1}^n$ 为数据场点, 构造 $\boldsymbol{Y}$ 在 $\boldsymbol{X}$ 上的互点场 DF_{yx}, $\boldsymbol{X}$ 中任意数据点 $\boldsymbol{x}_i$ 均受互点场 DF_{yx} 的影响; 另一方面, 以 $\boldsymbol{X}$ 中的数据点 $\{\boldsymbol{x}_i\}_{i=1}^n$ 为数据质点, 而以 $\boldsymbol{Y}$ 中的数据点 $\{\boldsymbol{y}_i\}_{i=1}^m$ 为数据场点, 构造 $\boldsymbol{X}$ 在 $\boldsymbol{Y}$ 上的互点场 DF_{xy}, $\boldsymbol{Y}$ 中的任意数据点 $\boldsymbol{y}_i$ 均受

互点场 DF_{xy} 的影响.

如何度量互点场 DF_{yx} 对数据点 $\boldsymbol{x}_i$ 的影响, 以及互点场 DF_{xy} 对数据点 $\boldsymbol{y}_i$ 的影响是关键. 本研究用互势值作为这种影响的量化表示. 更确切地说, 互点场 DF_{yx} 对数据点 $\boldsymbol{x}_i$ 的影响表示为 $\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$, 其中 $\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)$ 为互点场 DF_{yx} 在数据场点 $\boldsymbol{x}_i$ 处的互势值; 同理, 互点场 DF_{xy} 对数据点 $\boldsymbol{y}_i$ 的影响表示为 $\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)(\boldsymbol{y}_i-\bar{\boldsymbol{y}})$, 其中 $\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)$ 为互点场 DF_{xy} 在数据场点 $\boldsymbol{y}_i$ 处的互势值.

在互点场影响下, 用加权的离均差之积 $\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\cdot\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^{\mathrm{T}}$ 刻画数据点对 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ 的相关性. 在数据场中, 为获得数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的最大相关性, 点对 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ 的相关表达为: $\boldsymbol{\alpha}^{\mathrm{T}}[\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\cdot\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^{\mathrm{T}}]\boldsymbol{\beta}$. 其中, $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 为式 (1) 需要寻找的投影向量对. 因此, 在数据场环境下, 对来自数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的 n 对样本 $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$, $\boldsymbol{x}_i\in\mathbf{R}^{p\times1}$, $\boldsymbol{y}_i\in\mathbf{R}^{q\times1}$, 式 (2) 所示的 CCA 优化问题变为如下的 DFCCA 优化问题:

$$\begin{aligned}\max_{\substack{\boldsymbol{\alpha}\in\mathbf{R}^{p\times1}\\ \boldsymbol{\beta}\in\mathbf{R}^{q\times1}}}\quad & \boldsymbol{\alpha}^{\mathrm{T}}\sum_{i=1}^{n}\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^{\mathrm{T}}\boldsymbol{\beta}\\ \text{s.t.}\quad & \boldsymbol{\alpha}^{\mathrm{T}}\sum_{i=1}^{n}\varphi_{\boldsymbol{y}}^2(\boldsymbol{x}_i)(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\alpha}=1\\ & \boldsymbol{\beta}^{\mathrm{T}}\sum_{i=1}^{n}\varphi_{\boldsymbol{x}}^2(\boldsymbol{y}_i)(\boldsymbol{y}_i-\bar{\boldsymbol{y}})(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^{\mathrm{T}}\boldsymbol{\beta}=1\end{aligned}\tag{8}$$

2.4 DFCCA 求解及特征表示

记 $\boldsymbol{X}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n]^{\mathrm{T}}$, $\boldsymbol{Y}=[\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n]^{\mathrm{T}}$, 其中 $\boldsymbol{x}_i=[x_{i1},\cdots,x_{ip}]^{\mathrm{T}}$, $\boldsymbol{y}_i=[y_{i1},\cdots,x_{iq}]^{\mathrm{T}}$. 令

$$\boldsymbol{F}_{xy}=(\varphi_{\boldsymbol{x}_i}(\boldsymbol{y}_j))_{n\times n}\tag{9}$$

为 $\boldsymbol{X}$ 的各数据对象在 $\boldsymbol{Y}$ 的所有数据对象处的势函数构成的矩阵, 其中 $\varphi_{\boldsymbol{x}_i}(\boldsymbol{y}_j)$ 为 $\boldsymbol{x}_i$ 在 $\boldsymbol{y}_j$ 处的势函数. 将式 (8) 中目标函数的求和项展开有:

$$\begin{aligned}\boldsymbol{M}_{xy}=&\sum_{i=1}^{n}\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\cdot\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^{\mathrm{T}}=\\ &\boldsymbol{E}_{xy}-\boldsymbol{J}_{xy}-\boldsymbol{G}_{xy}+\boldsymbol{H}_{xy}\end{aligned}$$

其中

$$\begin{aligned}\boldsymbol{E}_{xy}=&\left(\sum_{i=1}^{n}\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)x_{ir}y_{is}\right)_{p\times q}=\\ &\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\boldsymbol{Y}\\ \boldsymbol{J}_{xy}=&\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)x_{ir}y_{js}\right)_{p\times q}=\\ &\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Lambda}_{xy}\mathbf{1}\boldsymbol{Y}\\ \boldsymbol{G}_{xy}=&\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)x_{jr}y_{is}\right)_{p\times q}=\\ &\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\mathbf{1}\boldsymbol{\Lambda}_{xy}\boldsymbol{Y}\\ \boldsymbol{H}_{xy}=&\frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)x_{jr}y_{ks}\right)_{p\times q}=\\ &\frac{\mathrm{tr}(\boldsymbol{\Lambda}_{xy})}{n^2}\boldsymbol{X}^{\mathrm{T}}\mathbf{1}\boldsymbol{Y}\end{aligned}$$

式中, $r=1,\cdots,p$; $s=1,\cdots,q$; $\mathbf{1}$ 是元素全为 1 的 n 阶方阵; $\boldsymbol{\Lambda}_{xy}$ 为如下所示的对角阵:

$$\boldsymbol{\Lambda}_{xy}=\mathrm{diag}\left\{\boldsymbol{F}_{xy}^{\mathrm{T}}\mathbf{1}_n\circ\boldsymbol{F}_{yx}^{\mathrm{T}}\mathbf{1}_n\right\}\tag{10}$$

其中, "$\circ$" 为 Hadamard 积, $\mathbf{1}_n$ 是元素全为 1 的 n 维列向量. 因为 Hadamard 积满足交换率, 故有 $\boldsymbol{\Lambda}_{xy}=\boldsymbol{\Lambda}_{yx}$. 对任意 n 阶方阵 $\boldsymbol{D}$, 若记:

$$f(\boldsymbol{D})=\boldsymbol{D}-\frac{1}{n}\boldsymbol{D}\mathbf{1}-\frac{1}{n}\mathbf{1}\boldsymbol{D}+\frac{\mathrm{tr}(\boldsymbol{D})}{n^2}\mathbf{1}\tag{11}$$

则有: $\boldsymbol{M}_{xy}=\boldsymbol{X}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{xy})\boldsymbol{Y}$. 同理可得: $\boldsymbol{M}_{xx}=\boldsymbol{X}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{xx})\boldsymbol{X}$, $\boldsymbol{M}_{yy}=\boldsymbol{Y}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{yy})\boldsymbol{Y}$. 因此, 式 (8) 所示的优化问题改写为

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathbf{R}^{p\times1},\boldsymbol{\beta}\in\mathbf{R}^{q\times1}}\quad & \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{xy})\boldsymbol{Y}\boldsymbol{\beta}\\ \text{s.t.}\quad & \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{xx})\boldsymbol{X}\boldsymbol{\alpha}=1\\ & \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{yy})\boldsymbol{Y}\boldsymbol{\beta}=1\end{aligned}\tag{12}$$

利用拉格朗日乘数法, 式 (12) 变为如下所示的广义特征值问题:

$$\begin{aligned}&\begin{pmatrix}\mathbf{0}&\boldsymbol{X}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{xy})\boldsymbol{Y}\\ \boldsymbol{Y}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{yx})\boldsymbol{X}&\mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}=\\ &\lambda\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{xx})\boldsymbol{X}&\mathbf{0}\\ \mathbf{0}&\boldsymbol{Y}^{\mathrm{T}}f(\boldsymbol{\Lambda}_{yy})\boldsymbol{Y}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}\end{aligned}\tag{13}$$

求式 (13) 所得向量 $[\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}}]^{\mathrm{T}}$ 即为式 (8) 的解.

由式 (13) 可解出 d 个正特征值 $\lambda_1\geq\cdots\geq\lambda_d>0$ 及对应的 d 个特征向量 $[\boldsymbol{\alpha}_1^{\mathrm{T}},\boldsymbol{\beta}_1^{\mathrm{T}}]^{\mathrm{T}},\cdots,[\boldsymbol{\alpha}_d^{\mathrm{T}},\boldsymbol{\beta}_d^{\mathrm{T}}]^{\mathrm{T}}$, $d\leq\min\{p,q\}$, 不妨记:

$$\boldsymbol{A}=[\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_d],\quad \boldsymbol{B}=[\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_d]$$

对来自数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的数据点对 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$, $\boldsymbol{x}_i\in\mathbf{R}^{p\times1}$, $\boldsymbol{y}_i\in\mathbf{R}^{q\times1}$, 本文称 $(\boldsymbol{u}_i,\boldsymbol{v}_i)$ 为 DFCCA 从数据点对 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$提取的特征, 其中:

$$\boldsymbol{u}_i=\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$$

$$\boldsymbol{v}_i = \varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}) \tag{14}$$

式中, $\varphi_{\boldsymbol{y}}(\boldsymbol{x}_i)$ 为 $\boldsymbol{Y}$ 在 $\boldsymbol{X}$ 上的互点场 DF_{yx} 于数据场点 $\boldsymbol{x}_i$ 处的互势值, 而 $\varphi_{\boldsymbol{x}}(\boldsymbol{y}_i)$ 为 $\boldsymbol{X}$ 在 $\boldsymbol{Y}$ 上的互点场 DF_{xy} 于数据场点 $\boldsymbol{y}_i$ 处的互势值. 可见, 特征 $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ 是原数据点对 $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ 的低维表示 (因为 $d \le \min\{p, q\}$). 因此, DFCCA 是一种基于数据场的数据降维技术.

对于容量为 n 的数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$, 可以获得 n 对特征 $(\boldsymbol{u}_i, \boldsymbol{v}_i)$. 不妨记:

$$\boldsymbol{U} = [\boldsymbol{U}_1, \cdots, \boldsymbol{U}_d], \quad \boldsymbol{U}_j = [u_{1j}, \cdots, u_{nj}]^{\mathrm{T}}$$
$$\boldsymbol{V} = [\boldsymbol{V}_1, \cdots, \boldsymbol{V}_d], \quad \boldsymbol{V}_j = [v_{1j}, \cdots, v_{nj}]^{\mathrm{T}}$$

其中, u_{ij} 和 v_{ij} 分别为特征 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 的第 j 个元素, $i = 1, \cdots, n$; $j = 1, \cdots, d$. 本研究分别称 $\boldsymbol{U}_j$ 和 $\boldsymbol{V}_j$ 为 DFCCA 从 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 提取的第 j 特征.

本研究发现, 由于 DFCCA 在求解相关性时考虑了数据点之间的相互作用, 在数据场环境下 DFCCA 提取的特征具有良好的分布特性, 原空间上相隔较远的数据点对的特征聚集在一个较小的区域内, 而原空间上相邻的数据点对的特征在特征空间中却有规律地分布在其他点所聚集区域的周围. 这一良好的分布特性使得数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相邻边界点可被直观地提取出来, 因此 DFCCA 有望用于图像分割等领域, 下一节通过实验考察 DFCCA 所提取特征的这一分布特性及其在图像分割中的应用.

3　实验

本节通过实验考察 DFCCA 所提取特征的分布. 实验首先在仿真数据集上考察 DFCCA 所提取特征的分布特点, 再将其应用于图像分割中. 实验过程中, 数据场距离指数 $k = 2$; 势函数选取式 (5) 所示的拟核力场势函数, 其中数据点之间距离通过式 (7) 定义的向量的广义伪距离计算, 取 $s = 2$.

3.1　DFCCA 特征分析

3.1.1　DFCCA 特征分布规律探析

先考察 DFCCA 所提取特征的分布规律, 实验在容量为 30 的两个 2 维非线性数据集 $\boldsymbol{X} = (x_{ij})_{30\times 2}$ 和 $\boldsymbol{Y} = (y_{ij})_{30\times 2}$ 上进行, 数据集中数据点对按下式产生:

$$\begin{cases} x_{ij} = 2 + \theta\sin(3\theta) + \varepsilon_1 \\ y_{ij} = \mathrm{e}^{\frac{\theta}{4}}\cos(2\theta)\sin(2\theta) + \varepsilon_2 \end{cases} \tag{15}$$

其中, $\theta \sim \mathrm{U}(-\pi, \pi)$ 为 $[-\pi, \pi]$ 上均匀分布的随机变量, 而 $\varepsilon_1 \sim \mathrm{N}(0, 0.01^2)$, $\varepsilon_2 \sim \mathrm{N}(0, 0.5^2)$ 为随机噪声, $i = 1, \cdots, 30$; $j = 1, 2$.

图 2 为数据场影响因子 $\sigma = 0.85$, 数据质量 $m = 1$ 时的实验结果, 图 2 中每个点侧面的数字为数据点或特征点编号. 图 2 (a) 为原始数据分布及其数据场等势线图, 封闭曲线为数据场等势线, 星号标示的点对应数据集 $\boldsymbol{X}$, 而实心圆标示的点对应数据集 $\boldsymbol{Y}$. 由图 2 (a) 可知, 数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别分布在右上角和左下角区域, 在两数据集相邻边界点附近, 数据场等势线发生了交叠. 本实验的目的在于考察 DFCCA 提取的相邻边界点的特征分布有何特点.

图 2 (b) 为经典 CCA 提取的第 1 特征的分布情况. 此图表明, 经典 CCA 提取的第 1 特征分布比较分散, 且两数据集相邻边界点的特征分布并无显著的几何直观. 可见, 经典 CCA 在提取两数据集相邻边界点方面的能力较弱. 实验也观察了经典 CCA 提取的第 2 特征的分布, 结果与第 1 特征的分布一样不存在几何直观.

图 2 (c) 和 (d) 分别为 DFCCA 所提取的第 1 特征和第 2 特征的分布情况. 由此两图可以看出, DFCCA 所提取的第 1 特征和第 2 特征的分布均具有显著的几何直观, 概括为:

1) 原空间上相隔较远的数据点对的特征聚集在一个较小的区域内, 不妨称此区域为特征 Body, 简称 B 区, 如图中矩形框所示. 可见, B 区的空间跨度很小, 且聚集了大量特征点, 这些特征点之间相距非常近, 几乎重叠在一起;

2) 原空间上相邻的数据点对的特征在特征空间中有规律地分布在 B 区周围. 分别沿 B 区的 4 条边界画直线可以将 B 区周围分为 L、R、D、T 4 个区域, 其中 L 区为 B 区左边界至 $-\infty$ 远处、R 区为 B 区右边界至 $+\infty$ 远处、D 区为 B 区下边界至 $-\infty$ 远处、T 区为 B 区上边界至 $+\infty$ 远处. 特征的分布规律为:

a) 分散在 L 区和 R 区的特征与 $\boldsymbol{X}$ 数据集中靠近 $\boldsymbol{Y}$ 数据集的数据点对应. 例如, 图 2 (c) 所示第 1 特征中, 细线圆标注的处于 R 区的 11 和 18 号特征以及 L 区的 28 号特征与 $\boldsymbol{X}$ 数据集的同编号数据点对应. 显然, 图 2 (a) 中 $\boldsymbol{X}$ 数据集的 11、18 和 28 号数据点距离 $\boldsymbol{Y}$ 非常近, 它们落入了 $\boldsymbol{Y}$ 在 $\boldsymbol{X}$ 上的互点场的第 7 级等势线内, 且 18 号数据点已接近第 4 级等势线, 而图 2 (c) 对应的 18 号特征距离 B 区也最远. 图 2 (d) 中第 2 特征也呈现出与第 1 特征类似的分布规律, L 区细线圆标注的 8、10、11、18、22、28 和 30 号特征对应 $\boldsymbol{X}$ 中的同编号数据点, 显然它们与 $\boldsymbol{Y}$ 临近.

b) 而分散在 T 区和 D 区的特征与 $\boldsymbol{Y}$ 数据集中靠近 $\boldsymbol{X}$ 数据集的数据点对应. 例如, 图 2 (c) 所示第 1 特征中, 粗线圆标注的处于 D 区的 4、22、23、25 号特征以及 T 区的 1、9、18、19 和 26 号特征与 $\boldsymbol{Y}$

数据集的同编号数据点对应. 显然, 图 2 (a) 中 $\boldsymbol{Y}$ 数据集的 1、4、9、18、19、22、23、25 和 26 号数据点距离 $\boldsymbol{X}$ 最近, 其中 19、23 和 25 号数据点落入了 $\boldsymbol{X}$ 在 $\boldsymbol{Y}$ 上的互点场的第 11 级等势线内. 图 2 (d) 所示的第 2 特征也呈现出与第 1 特征类似的分布规律, 粗线圆标注的处于 D 区的 10、19、25 号特征以及 T 区的 22 号特征对应 $\boldsymbol{Y}$ 中的同编号数据点, 显然它们与 $\boldsymbol{X}$ 临近.

c) 特征点距离 L、R、D、T 线越远, 其对应的数据点距离另一数据集越近, 所谓 "近" 包括嵌入另一数据集内部. 以图 2 (c) 中 R 区和 T 区的 18 号特征点为例, 该点处于 R 区最右侧, 距离 R 线最远, 观察图 2 (a) 中 $\boldsymbol{X}$ 数据集 (星号) 的同编号数据点发现, 该点已嵌入 $\boldsymbol{Y}$ 中, 即它距离 $\boldsymbol{Y}$ 最 "近"; 同为 18 号特征点, 它与 T 线的距离不如处于 D 区的 25 号特征点与 D 线的距离远, 因此图 2 (a) 中 $\boldsymbol{Y}$ 数据集 (实心圆) 的 18 号数据点没有 25 号数据点距离 $\boldsymbol{X}$ 近.

可见, 增大 B 区的空间跨度, 所提取的相邻边界点变少, 但这些点的相邻边界特性更加明晰, 即距离另一数据集更近; 反之, 缩小 B 区的空间跨度, 则所提取的相邻边界点增多, 但所增加点的边界特性逐渐减弱. 因此, B 区范围的选择应根据所需提取的相邻边界点数目的多寡而定.

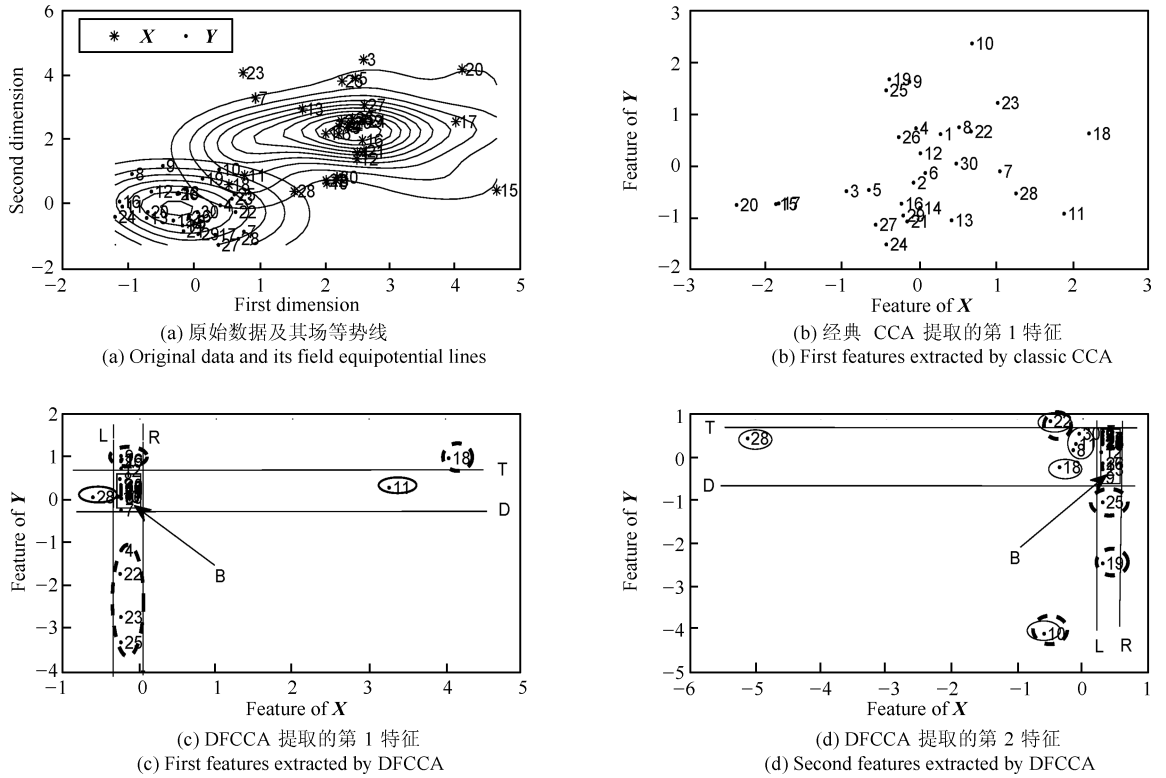
为直观起见, 本文实验仅仅考虑二维数据情况, 故后文将 B 区表示为形如 $(l, r; d, t)$ 的形式, 其中 l 表示 L 区右边界对应的横轴坐标值, r 表示 R 区左边界对应的横轴坐标值, d 表示 D 区上边界对应的纵轴坐标值, 而 t 表示 T 区下边界对应的纵轴坐标值.

3.1.2 DFCCA 在复杂数据上的边界辨识研究

本实验在球形带噪数据集上考察 DFCCA 所提取特征的分布特点, 以探讨其在复杂分布情况下的边界辨识能力, 即探究 DFCCA 能否提取边界点. 实验产生了两个含噪声的球形数据集 $\boldsymbol{X} = (x_{ij})_{50\times 2}$ 和 $\boldsymbol{Y} = (y_{ij})_{50\times 2}$, 含 2 维样本 50 对, 数据集中数据点对按下式生成:

$$\begin{cases} x_{i1} = \cos(\theta_i) + \varepsilon_1, & x_{i2} = \sin(\theta_i) + \varepsilon_1 \\ y_{i1} = 2\cos(\theta_i) + \varepsilon_2, & y_{i2} = 2\sin(\theta_i) + \varepsilon_2 \end{cases} \tag{16}$$

其中, $\theta_i = 2(i-1)\pi/50$, $i = 1, \cdots, 50$, $\varepsilon_1 \sim \mathrm{N}(0, 0.8^2)$, $\varepsilon_2 \sim \mathrm{N}(0, 0.5^2)$ 为随机噪声. 实验过程中, 数据场影响因子 $\sigma = 0.65$, 数据质量 $m = 1$.



(a) 原始数据及其场等势线
(a) Original data and its field equipotential lines

(b) 经典 CCA 提取的第 1 特征
(b) First features extracted by classic CCA

(c) DFCCA 提取的第 1 特征
(c) First features extracted by DFCCA

(d) DFCCA 提取的第 2 特征
(d) Second features extracted by DFCCA

图 2 DFCCA 所提取特征的分布规律

Fig. 2 Feature distributions extracted by DFCCA

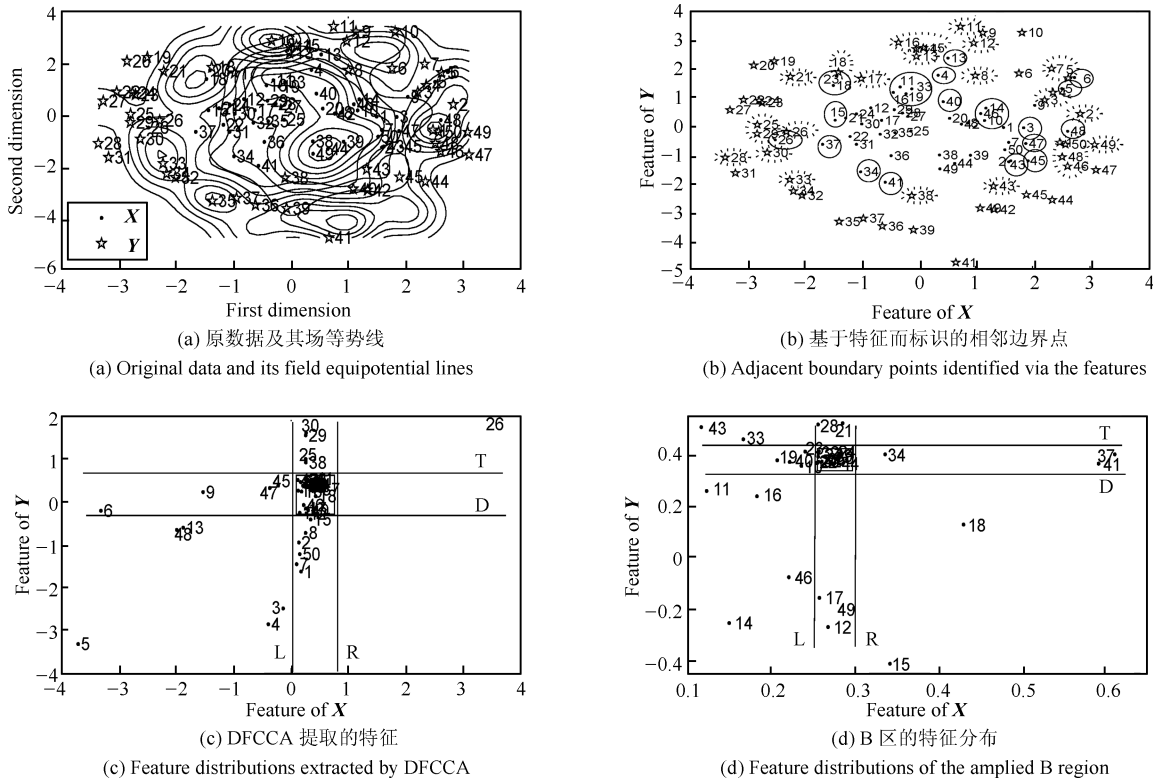
实验结果如图 3 所示, 图中每个点侧面的数字为数据点或特征点编号. 图 3 (a) 为原始数据分布及其数据场等势线图, 封闭曲线为数据场等势线, 实心圆标示的点对应数据集 $\boldsymbol{X}$, 而五角星标示的点对应数据集 $\boldsymbol{Y}$. 图 3 (b) 为图 3 (a) 去掉等势线后的原始数据分布. 由图 3 (a) 可知, 数据集 $\boldsymbol{X}$ 嵌套在数据集 $\boldsymbol{Y}$ 内, 两者的数据场交叠复杂, 部分相邻边界点潜入对方内部. 本实验的目的在于考察 DFCCA 能否提取出这些复杂交叠的相邻边界点.

DFCCA 所提取的第 1 特征如图 3 (c) 所示. 此图表明, 在非线性可分的复杂分布情况下, DFCCA 所提取的特征与图 2 (a) 所示的线性可分情况下一样具有良好的分布特性, 其分布呈现出明显的 B 区, 原空间上相隔较远的数据点对的特征聚集在 B 区内, 而原空间上相邻的数据点对的特征在特征空间中却有规律地分布在 B 区周围. 可见, DFCCA 也能有效提取非线性可分的两数据集边界点.

观察图 3 (b) 中实线圆和虚线圆圈注的点发现, 分散在图 3 (c) 的 L、R、D、T 区的特征已能提取出数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的边界轮廓, 但是此轮廓还比较粗糙, 有必要再提取部分相邻边界点, 这可以通过缩小 B 区实现. 图 3 (d) 为图 3 (c) 的 B 区在视觉效果上的放大图 (注: 特征点之间原有的空间位置未变). 在原 B 区 (大 B 区) 放大后发现, 其间的特征分布仍然呈现出显著的 B 区 (小 B 区) 及 L、R、D、T 区, 这表明 DFCCA 提取的特征分布具有层级或嵌套性. 作者将图 3 (d) 中 B 区再次在视觉效果上放大, 并将图 2 (c) 和图 2 (d) 中 B 区在视觉效果上放大, 都发现了类似的嵌套性, 这实属研究预期之外的一个发现.

图 3 (d) 中处于 L 区的 10、11、14、16、19、23、33、40、43 和 46 号特征点, 以及处于 R 区的 15、18、34、37 和 41 号特征点, 它们在图 3 (b) 中对应的 $\boldsymbol{X}$ 数据集的同编号数据点用实线圆圈注; 而处于 D 区的 11、12、14、15、16、17、18、46 和 49 号特征点, 以及处于 T 区的 21、28、33 和 43 号特征点, 它们在图 3 (b) 中对应的 $\boldsymbol{Y}$ 数据集的同编号数据点用虚线圆圈注. 可见, 缩小 B 区后, 相邻边界点增多, 边界轮廓逐渐清晰, 但所增加点的边界特性逐渐减弱, 即离另一个数据集的边界越远.

本实验所生成数据集的复杂性不仅在于两数据集的非线性可分性, 还在于它们边界点的叠加性, 即边界点相互嵌入. 如图 3 (b) 或图 3 (a) 中 $\boldsymbol{X}$ 数据集的 5 号和 6 号数据点已嵌入 $\boldsymbol{Y}$ 数据集内部, 它们在图 3 (c) 中对应的同编号特征点距离 L 线较远; 同样, $\boldsymbol{Y}$ 数据集的 3、4、5 号特征点和数据点也呈现出类似特性. 可见, $\boldsymbol{X}$ 数据集的 5、6 号数据点和 $\boldsymbol{Y}$ 数据集的 3、4、5 号数据点形成了一个叠加区, 所有



(a) 原数据及其场等势线
(a) Original data and its field equipotential lines

(b) 基于特征而标识的相邻边界点
(b) Adjacent boundary points identified via the features

(c) DFCCA 提取的特征
(c) Feature distributions extracted by DFCCA

(d) B 区的特征分布
(d) Feature distributions of the amplied B region

图 3　含噪声球形数据集下 DFCCA 所提取特征的分布及边界辨识

Fig. 3　Feature distributions extracted from spherical noisy data sets by DFCCA and class frontier identification

的叠加区共同构成了一条边界带. 因此, DFCCA 提取的是一条边界带, 其中的数据点在空间上相互交织.

3.1.3 数据场影响因子对 DFCCA 特征的影响分析

为考察数据场影响因子对 DFCCA 所提取特征分布的影响, 本实验基于图 2 (a) 所示数据集考察了 100 种情况, 数据场影响因子 σ 以 0.05 为步长, 从 0.05 增至 5. 图 4 为选取的特征分布较有代表性的部分实验结果, 其中 σ 依次为: 0.05、0.35、0.65、1.05、1.55 和 3.65.

由图 4 可见, σ 较小时, DFCCA 所提取的特征聚集在少数几个区域内, 如 $\sigma = 0.05$ 时, 其特征聚集在三个区域, 且同一区域内特征点之间的距离如此之近而交叠在一起, 以至于看上去像聚集在三个点上一样; 随着 σ 的增大, DFCCA 所提取特征的聚集区域数目逐渐增多, 如 $\sigma = 0.65$ 时, 特征聚集到 7 个区域, 右下角密集区呈现出 B 区特点, 其相邻边界点对应特征主要分布在 L 区和 T 区; 当 σ 增大到 3.65 时, DFCCA 所提取的特征已比较分散, 其分布未呈现出显著的聚集特性, 与图 2 (b) 所示的经典 CCA 所提取特征的分布特点类似.

因此, σ 的取值应根据不同数据集而定. 作者认为, 应以能呈现出显著的 B 区为宜, 如本实验数据集下, σ 取值介于 [0.65, 1.05] 区间能获得显著的 B 区.

需要补充的是, 尽管势函数的另一参数, 即式 (5) 所示的距离指数 k, 也是影响势函数值的一个因素, 但已有研究[35] 指出, 数据场的空间分布主要取决于影响因子 σ, 而与距离指数 k 的选取关系不大, 故本文未考察 k 对 DFCCA 所提取特征的影响.

3.2 DFCCA 在图像分割中的应用

前述实验结果表明, DFCCA 提取的特征具有良好的分布特性, 基于分散在 L、R、D、T 区的特征, 能辨识两数据集的边界, 即可提取出相邻边界点, 本实验将此特性应用于图像分割领域.

3.2.1 DFCCA 分割图像的实证研究

本实验对 DFCCA 的图像分割效果进行直观考察, 包括双色纹理图像和含复杂线条无背景的花瓶图像. 实验将图像转换为灰度图, 数据质量 m 用灰度值表示, 坐标用像素点位置刻画. 将图的灰度值大于其均值加 1 倍方差的像素作为 $\boldsymbol{X}$ 数据集, 其余像素点作为 $\boldsymbol{Y}$ 数据集. 由于 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的容量只有巧合才一致, 实验对容量大的数据集进行均匀随机抽样, 抽取的样本容量与小数据集一致. 在对大数据集抽样前, 分别建立 $\boldsymbol{X}$ 对 $\boldsymbol{Y}$ 的互点场以及 $\boldsymbol{Y}$ 对 $\boldsymbol{X}$ 的互点场.

DFCCA 在纹理图像分割的结果如图 5 所示. 图 5 (a) 为黑白相间、粗细渐变的波浪形斑马纹原图, 实验的目的在于分别提取出黑色纹理和白色纹理的边界点. 图 5 (b) 为数据场影响因子 $\sigma = 0.05$ 情况下的特征分布, 其分布呈现出显著的 B 区, 相邻边界点对应特征在 L、R、D、T 区皆有分布, 可见数据场影响因子的选取是合理的. 图 5 (c) 和图 5 (d) 分别为黑色纹理和白色纹理边界点, 其中 B 区为 $(-0.308, -0.304; -0.395, -0.393)$. 由图 5 (c) 和图 5 (d) 可见, DFCCA 提取的纹理边界具有较好的保真性, 其轮廓不仅保持了原纹理的波浪形斑马纹状, 而且粗细渐变的纹理宽度在提取出的边界点中也得以体现.

DFCCA 在含复杂线条无背景花瓶图像上的分割结果如图 6 所示, 其中图 6 (a) 为黑白双色的原花瓶图, 可见线条形状复杂, 有平行直线、波浪线、弧形线及其他不规则线条, 线粗细不等, 且填充区与空白区复杂交错. 图 6 (b) 为数据场影响因子 $\sigma = 2.3$ 情况下的特征分布, 可见其特征呈现出显著的 B 区, 相邻边界点的特征主要分布在 L、R 和 T 区, 且大部分分散在 L 和 T 区的交叉区域, 可见数据场影响因子的选取是合理的. 图 6 (c) 和图 6 (d) 分别为 B 区取 $(0.05, 0.55; -0.7, -0.5)$ 时所提取的黑色边界和白色边界, 而图 6 (e) 为两色边界的叠加图. 可见, DFCCA 所分割出的花瓶轮廓与原图相当吻合, 其中一个细节显示, 线条具有大角度弯曲的花瓶两耳在图 6 (c) 中得以完美呈现, 且右耳根部向左下开口的狭长空白区也被区别开了. 但是, 在瓶颈下部的拱形条纹处提取的边界并不理想, 拱形线之间被相邻边界点填充了, 这是由于对这部分选取的相邻边界点过多而造成的, 因为它们线细而间隔小, 可对这部分单独重新选取 B 区进一步完善.

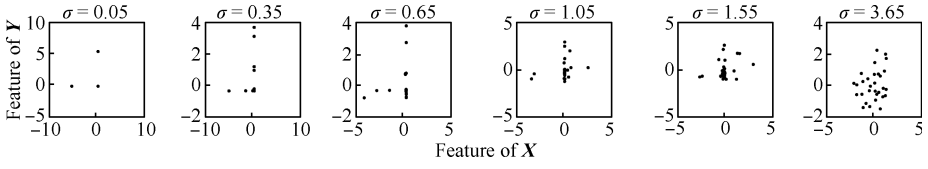


图 4 数据场影响因子对 DFCCA 所提取特征分布的影响

Fig. 4 Impacts on the feature distributions extracted by DFCCA of the impact factors of data field

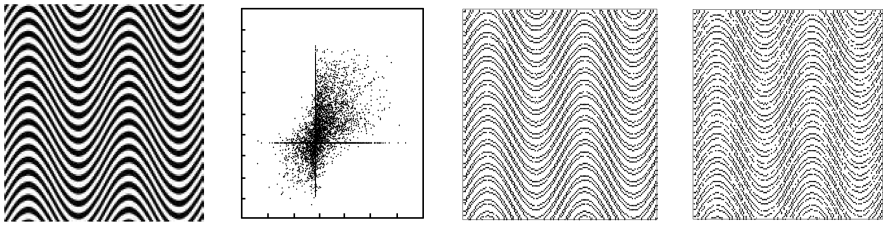


图 5　DFCCA 分割双色纹理

Fig. 5　Bicolor texture image segmentation by DFCCA

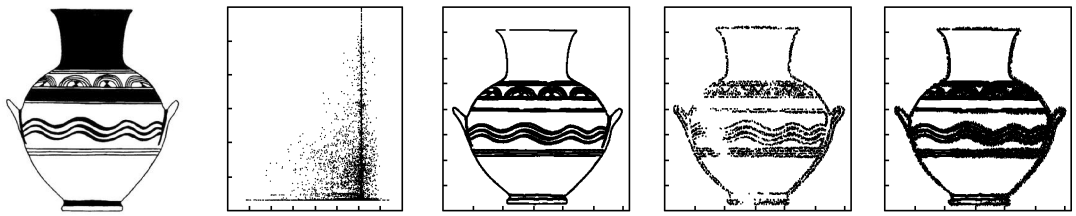


图 6　DFCCA 提取的花瓶轮廓

Fig. 6　Profile of vase segmented by DFCCA

3.2.2　DFCCA 在复杂图像上的分割能力评估

为了客观评价 DFCCA 在图像分割中的应用效果, 本实验将其与 Laplacian[36]、Sobel[37]、Roberts[38] 等图像分割的经典算法进行比较, 重点考察它们在复杂图像上分割能力和计算效率的差异, 其中分割能力用块相似系数 (Dice similanity coefficient, DSC)、整体一致误差 (Global consistency error, GCE)、分割错误率 (Ratio of segmentation error, RSE) 和信息变异量 (Variation of information, VI) 等 4 个指标进行评价[39−40], 其中 DSC 越大, 而 GCE、RSE、VI 越小, 说明分割保真度越好. 对于 DFCCA, 实验将每幅图灰度大于其均值加 1 倍方差的像素作为 $\boldsymbol{X}$ 数据, 其余作为 $\boldsymbol{Y}$ 数据集, 分别选取两数据集第一特征四分位数内的值构成 B 区.

图 7 为 4 种算法在复杂图像上的分割示例, 其中第 1 列为原始图像, 后 4 列依次对应 DFCCA、Laplacian、Sobel 和 Roberts 的分割结果. 直观考察的结果表明, 4 种算法都能有效提取图像中的主要对象, 但细节上存在差异, 总体而言, DFCCA 提取曲线型边界的能力较其他 3 个算法略好, 如第 1 幅图机头的驾驶舱和第 3 幅图熊前的冰面凹槽, DFCCA 分割的保真度较好.

为比较 DFCCA 与另外 3 个算法的分割能力, 实验采用得到广泛应用的 Berkeley 图像库, 将分割结果与 Berkeley 的标准分割结果对比. 实验重复了 60 次, 每次随机挑选一幅图像, 分别考察 4 种算法在 4 个分割指标上的值, 最后用指标的均值刻画各算法的分割能力.

图 8 为不同算法在 4 个分割指标上对比的箱盒图 (Box plot)、考察缺口 (Notches) 和中心标志 (Central marks) 发现, 在指标 DSC 上, DFCCA 的缺口跨度 (Intervals) 与另外三个经典分割方法均未重叠, 且中心标志最大, 这说明 DFCCA 的 DSC 分割指标较好; 在指标 GCE 和 VI 上, DFCCA 的缺口跨度也未与其他算法重叠, 且中心标志最小, 这说明 DFCCA 的 GCE 和 VI 分割指标较好; 在指标 RSE 上, DFCCA 的分割效果与 Laplacian 差异不显著, 但明显好于另外两个算法.

本研究还对算法的计算效率进行了比较, 共进行 60 次实验, 每次从 Berkeley 图像库中随机挑选一幅图, 然后对图像进行压缩, 考察 12 种不同压缩率. 图 9 为 4 种算法在不同图像大小下的平均运行时间比较图, 横坐标表示压缩率, 其计算公式为 $Si = 1 - (12 - i) \times 40)/\max(r, c)$, 其中 r 和 c 分别表示图像高和宽. 由图 9 可知, 当图像较小时, DFCCA 的平均运行时间比其他三个算法略低, 但当图像逐渐增大时, DFCCA 的计算效率下降较快, 其原因可能与数据场势值求解效率随图像增大而快速降低有关.

总之, 上述实验结果表明, DFCCA 提取的特征具有良好的分布特性, 根据分散在 L、R、D、T 区的特征能直观地辨识相邻边界点, DFCCA 具有较好的图像分割能力, 其分割的图像保真度较高.

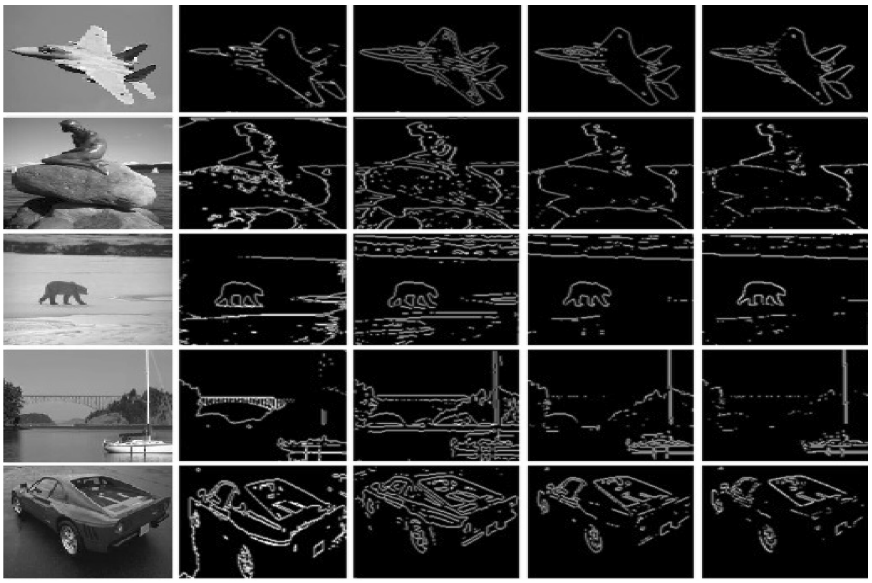


图 7 不同算法在 4 个分割指标上的对比

Fig. 7 Comparasions of different segmentation methods on the four segmentation evaluations

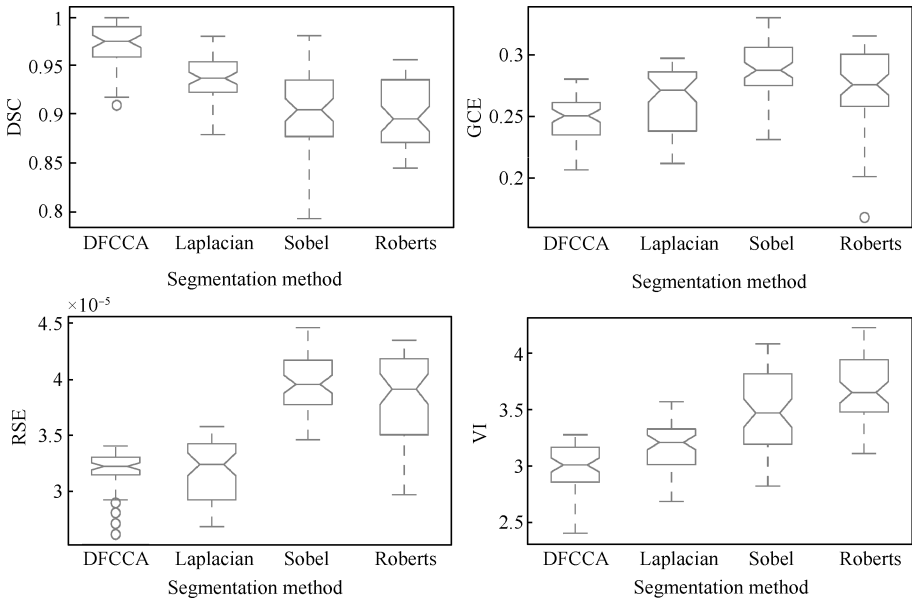


图 8 平均运行时间比较图

Fig. 8 Comparasions of different segmentation methods on average running time

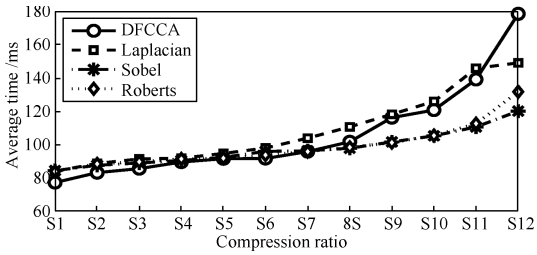


图 9 平均运行时间比较图

Fig. 9 Comparisons of different segmentation methods on average running time

4 结论

本文从数据间相互作用的视角出发, 引入数据场理论, 研究了一种新的典型相关分析方法 DFCCA, 以提取数据场环境下数据的低维特征.

针对两个数据集相互作用形成的数据场, 提出数据质点、数据场点、互点场、互势值等概念; 为求解不同维数据集形成的互数据场中数据场点的互势值, 提出一种不同维向量之间的距离计算公式, 称为向量的广义伪距离; 重点研究了 DFCCA 的数学描

述及解的推导.

由于将数据之间的相互作用纳入其相关性求解中, 在数据场环境下 DFCCA 提取的特征呈现出经典 CCA 所不具有的崭新的良好分布特性: 1) 原空间上相隔较远的数据点对的特征聚集在一个较小的区域内, 而相邻的数据点对的特征却有规律地分布在其他点所聚集区域的周围; 2) 特征分布具有嵌套性, 将原 B 区放大后, 其间的特征分布仍然呈现出显著的 B 区及 L、R、D、T 区. 前述特性使得 DFCCA 具有较好的边界辨识能力, 实验部分将其应用到图像分割领域, 结果表明, DFCCA 提取的复杂图像边界具有较好的保真度.

DFCCA 有望作为一种崭新的数据降维及特征提取的基本工具应用于图像分割等领域. 但 DFCCA 所分割的图像边界点数目的优选还有待完善, 作者目前正在进行的另一项工作是将主曲线法应用于 DFCCA 提取的图像边界带以实现图像的精细分割, 下一步将研究数据场影响因子及 B 区选择的优化策略以实现图像边界点数目的优选.

References

1 Hotelling H. Relations between two sets of variates. *Biometrika*, 1936, **28**(3): 321−377

2 Chaudhuri K, Kakade S M, Livescu K, Sridharan K. Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th International Conference on Machine Learning. Montreal, Canada: ICML, 2009. 129−136

3 Olcay K, Ethem A, Oleg V F. Canonical correlation analysis using within-class coupling. *Pattern Recognition Letters*, 2011, **32**(2): 134−144

4 Peng Yan, Zhang Dao-Qiang. Semi-supervised canonical correlation analysis algorithm. *Journal of Software*, 2008, **19**(11): 2822−2832
(彭岩, 张道强. 半监督典型相关分析算法. 软件学报, 2008, **19**(11): 2822−2832)

5 Sun Quan-Sen, Zeng Sheng-Gen, Heng Pheng-Ann, Xia De-Shen. The theory of canonical correlation analysis and its application to feature fusion. *Chinese Journal of Computers*, 2005, **28**(9): 1524−1533
(孙权森, 曾生根, 王平安, 夏德深. 典型相关分析的理论及其在特征融合中的应用. 计算机学报, 2005, **28**(9): 1524−1533)

6 Hou Shu-Dong, Sun Quan-Sen. Sparsity preserving canonical correlation analysis with application in feature fusion. *Acta Automatica Sinica*, 2012, **38**(4):659−665
(侯书东, 孙权森. 稀疏保持典型相关分析及在特征融合中的应用. 自动化学报, 2012, **38**(4): 659−665)

7 Huang H, He H T, Fan X, Zhang J P. Super-resolution of human face image using canonical correlation analysis. *Pattern Recognition*, 2010, **43**(7): 2532−2543

8 Jia C C, Wang S J, Peng X J, Pang W, Zhang C Y, Zhou C G, Yu Z Z. Incremental multi-linear discriminant analysis using canonical correlations for action recognition. *Neurocomputing*, 2012, **83**: 56−63

9 Hong Quan, Chen Song-Can, Ni Xue-Lei. Sub-pattern canonical correlation analysis with application in face recognition. *Acta Automatica Sinica*, 2008, **34**(1): 21−30
(洪泉, 陈松灿, 倪雪蕾. 子模式典型相关分析及其在人脸识别中的应用. 自动化学报, 2008, **34**(1): 21−30)

10 Yuan Y H, Sun Q S, Ge H W. Fractional-order embedding canonical correlation analysis and its applications to multi-view dimensionality reduction and recognition. *Pattern Recognition*, 2014, **47**(3): 1411−1424

11 An B G, Guo J H, Wang H S. Multivariate regression shrinkage and selection by canonical correlation analysis. *Computational Statistics & Data Analysis*, 2013, **62**(6): 93−107

12 Singh A, Kulkarni M A, Mohanty U C, Kar S C, Robertson A W, Mishra G. Prediction of Indian summer monsoon rainfall (ISMR) using canonical correlation analysis of global circulation model products. *Meteorological Applications*, 2012, **19**(2): 179−188

13 Fukumizu K, Bach F R, Gretton A. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 2007, **8**(2): 361−383

14 Zheng W M, Zhou X Y, Zou C R, Zhao L. Facial expression recognition using kernel canonical correlation analysis (KCCA). *IEEE Transactions on Neural Networks*, 2006, **17**(1): 233−238

15 Hardoon D R, Mourão Miranda J, Brammer M, Taylor J S. Unsupervised analysis of fMRI data using kernel canonical correlation. *Neuroimage*, 2007, **37**(4): 1250−1259

16 Zhu X F, Huang Z, Shen H T, Cheng J, Xu C S. Dimensionality reduction by mixed kernel canonical correlation analysis. *Pattern Recognition*, 2012, **45**(8): 3003−3016

17 Shu C, Ouarda T B M J. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. *Water Resources Research*, 2007, **43**(7), doi: 10.1029/2006WR005142

18 Karageorgiou E, Lewis S M, Mccarten J R, Leuthold A C, Hemmy L S, Mcpherson S E, Rottunda S J, Rubins D M, Georgopoulos A P. Canonical correlation analysis of synchronous neural interactions and cognitive deficits in Alzheimer′s dementia. *Journal of Neural Engineering*, 2012, **9**(5): 056003

19 Gu Jing-Jing, Chen Song-Can, Zhuang Yi. Localization in wireless sensor network using locality preserving canonical correlation analysis. *Journal of Software*, 2010, **21**(11): 2883−2891
(顾晶晶, 陈松灿, 庄毅. 用局部保持典型相关分析定位无线传感器网络节点. 软件学报, 2010, **21**(11): 2883−2891)

20 Wang F S, Zhang D Q. A new locality-preserving canonical correlation analysis algorithm for multi-view dimensionality reduction. *Neural Processing Letters*, 2013, **37**(2): 135−146

21 Yuan Y H, Sun Q S. Graph regularized multiset canonical correlations with applications to joint feature extraction. *Pattern Recognition*, 2014, **47**(12): 3907−3919

22 Gomez D D, Maletti G, Nielsen A A, Ersboll B. Multiset multitemporal canonical analysis of psoriasis images. In: Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging. Washington D. C., USA: IEEE, 2004. 1151−1154

23 Thompson B, Cartmill J, Azimi S M R, Schock S G. A multichannel canonical correlation analysis feature extraction with application to buried underwater target classification. In: Proceedings of the 2006 IEEE International Joint Conference on Neural Network. Vancouver, Canada: IEEE, 2006. 4413−4420

24 Li Y O, Adali T, Wang W, Calhoun V D. Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing*, 2009, **57**(10): 3918−3929

25 Yuan Y H, Sun Q S. Fractional-order embedding multiset canonical correlations with applications to multi-feature fusion and recognition. *Neurocomputing*. 2013, **122**(12): 229−238

26 Deleus F, Van Hulle M M. Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis. *Journal of Neuroscience Methods*, 2011, **197**(1): 143−157

27 Yang Jing, Li Wen-Ping, Zhang Jian-Pei. Canonical correlation analysis of big data based on cloud model. *Journal of Communications*, 2013, **34**(10): 121−134
(杨静, 李文平, 张健沛. 大数据典型相关分析的云模型方法. 通信学报, 2013, **34**(10): 121−134)

28 Sun L A, Ji S W, Ye J P. Canonical correlation analysis for multilabel classification: a least-squares formulation, extensions, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(1): 194−204

29 Yang Xue-Mei, Dong Yi-Sheng, Xu Hong-Bing, Liu Xue-Jun, Qian Jiang-Bo, Wang Yong-Li. Online correlation analysis for multiple dimensions data streams. *Journal of Computer Research and Development*, 2006, **43**(10): 1744−1750
(杨雪梅, 董逸生, 徐宏炳, 刘学军, 钱江波, 王永利. 高维数据流的在线相关性分析. 计算机研究与发展, 2006, **43**(10): 1744−1750)

30 Wang Yong-Li, Xu Hong-Bing, Dong Yi-Sheng, Qian Jiang-Bo, Liu Xue-Jun. A correlation analysis algorithm based on low-rank approximation for multiple dimension data streams. *Acta Electronica Sinica*, 2006, **34**(2): 293−300
(王永利, 徐宏炳, 董逸生, 钱江波, 刘学军. 基于低阶近似的多维数据流相关性分析. 电子学报, 2006, **34**(2): 293−300)

31 Wang Y L, Zhang G X, Qian J B. ApproxCCA: an approximate correlation analysis algorithm for multidimensional data streams. *Knowledge-Based Systems*, 2011, **24**(7): 952−962

32 Kim M. Correlation-based incremental visual tracking. *Pattern Recognition*, 2012, **45**(3): 1050−1060

33 Zhou Yong, Lu Xiao-Wei, Cheng Chun-Tian. Parallel computing method of canonical correlation analysis for high-dimensional data streams in irregular streams. *Journal of Software*, 2012, **23**(5): 1053−1072
(周勇, 卢晓伟, 程春田. 非规则流中高维数据流典型相关性分析并行计算方法. 软件学报, 2012, **23**(5): 1053−1072)

34 Yang Jing, Li Wen-Ping, Zhang Jian-Pei. A tracking algorithm based on rank two modifications for canonical correlation analysis of multidimensional data streams. *Acta Electronica Sinica*, 2012, **40**(9): 1765−1774
(杨静, 李文平, 张健沛. 基于秩 2 更新的多维数据流典型相关跟踪算法. 电子学报, 2012, **40**(9): 1765−1774)

35 Li De-Yi, Du Yi. *Artificial Intelligence with Uncertainty*. Beijing: National Defence Industry Press, 2005. 224−227
(李德毅, 杜鹢. 不确定性人工智能. 北京: 国防工业出版社, 2005. 224−227)

36 Milyaev S, Barinova O. Learning graph Laplacian for image segmentation. *Transactions on Computational Science XIX*, 2013, **7870**: 92−106

37 Sun J. Image edge detection based on relative degree of grey incidence and Sobel operator. *Artificial Intelligence and Computational Intelligence*, 2012, **7530**: 762−768

38 Klette R. *Image Segmentation*. London: Springer Press, 2014. 167−214

39 He C J, Wang Y, Chen Q. Active contours driven by weighted region-scalable fitting energy based on local entropy. *Signal Processing*, 2012, **92**(2): 587−600

40 Cho M, Mulee M K. Authority-shift clustering: hierarchical clustering by authority seeking on graphs. In: Proceedings of the 33th IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 3193−3200

李文平 哈尔滨工程大学国家大学科技园博士后, 嘉兴学院讲师. 主要研究方向为数据挖掘, 隐私保护, 膜计算.
E-mail: liwenping@hrbeu.edu.cn
(**LI Wen-Ping** Postdoctor at National Science Park of Harbin Engineering University, and lecturer at Jiaxing University. His research interest covers data ming, privacy preservation, and membrane computing.)

杨 静 哈尔滨工程大学教授. 主要研究方向为数据库理论, 数据挖掘, 隐私保护. 本文通信作者.
E-mail: yangjing@hrbeu.edu.cn
(**YANG Jing** Professor at Harbin Engineering University. Her research interest covers database, data mining, and privacy preservation. Corresponding author of this paper.)

印桂生 哈尔滨工程大学教授. 主要研究方向为可信软件, 数据库理论, 虚拟现实和信息安全.
E-mail: yinguisheng@hrbeu.edu.cn
(**YIN Gui-Sheng** Professor at Harbin Engineering University. His research interest covers trusted software, database, virtual reality, and information security.)

张健沛 哈尔滨工程大学教授. 主要研究方向为数据库理论, 数据挖掘, 数据流, 社会网络.
E-mail: zhangjianpei@hrbeu.edu.cn
(**ZHANG Jian-Pei** Professor at Harbin Engineering University. His research interest covers database, data mining, data stream, and social network.)