面向 Kullback-Leibler 散度不确定集的正则化线性判别分析

梁志贞 [1, 2] 张 磊 [1, 2]

摘 要 线性判别分析是一种统计学习方法. 针对线性判别分析的小样本奇异性问题和对污染样本敏感性问题, 目前许多线性判别分析的改进算法已被提出. 本文提出了基于 Kullback-Leibler (KL) 散度不确定集的判别分析方法. 提出的方法不仅利用了 L_s 范数定义类间距离和 L_r 范数定义类内距离, 而且对类内样本和各类中心的信息进行基于 KL 散度不确定集的概率建模. 首先通过优先考虑不利区分的样本提出了一种正则化对抗判别分析模型并利用广义 Dinkelbach 算法求解此模型. 这种算法的一个优点是在适当的条件下优化子问题不需要取得精确解. 投影 (次) 梯度法被用来求解优化子问题. 此外, 也提出了正则化乐观判别分析并采用交替优化技术求解广义 Dinkelbach 算法的优化子问题. 许多数据集上的实验表明了本文的模型优于现有的一些模型, 特别是在污染的数据集上, 正则化乐观判别分析由于优先考虑了类中心附近的样本点, 从而表现出良好的性能.

关键词 判别分析, KL 散度, 不确定集, 正则化, 数据分类

引用格式 梁志贞, 张磊. 面向 Kullback-Leibler 散度不确定集的正则化线性判别分析. 自动化学报, 2022, **48**(4): 1033−1047

DOI 10.16383/j.aas.c210434

Regularized Linear Discriminant Analysis Based on Uncertainty Sets From Kullback-Leibler Divergence

LIANG Zhi-Zhen[1, 2] ZHANG Lei[1, 2]

Abstract Linear discriminant analysis is a statistical learning method. For the singularity problem of small samples and the sensitivity to contaminated samples, now many improved algorithms of linear discriminant analysis have been proposed. In this paper we propose discriminant analysis methods via uncertainty sets from the Kullback-Leibler (KL) divergence. The proposed methods not only employ the L_s norm to define the distance between classes and the L_r norm to define the distance within classes, but also implement the probability modeling for within-class samples and class means based on uncertainty sets from the KL divergence. This paper first proposes a regularized adversarial discriminant analysis model by placing more emphasis on the samples that are difficult to be separated and then the generalized Dinkelbach′s algorithm is used to solve the proposed optimization model. One advantage of this method is that the optimization subproblems do not need to be solved precisely under proper conditions. In addition, this paper also proposes regularized optimistic discriminant analysis and uses the alternative optimization technique to solve optimization subproblems in the generalized Dinkelbach′s algorithm. Experiments on many data sets show that the proposed models are superior to some existing models. Especially on the contaminated data sets regularized optimistic discriminant analysis produces better performance since it places more emphasis on the samples which lie around class means.

Key words Discriminant analysis, KL divergence, uncertainty sets, regularization, data classification

Citation Liang Zhi-Zhen, Zhang Lei. Regularized linear discriminant analysis based on uncertainty sets from Kullback-Leibler divergence. *Acta Automatica Sinica*, 2022, **48**(4): 1033−1047

如今利用现代设备采集高维数据变得方便和容易, 但是获得的高维数据可能包含不相关和冗余的信息. 这不仅增加了学习模型的计算量和存储量, 而且可能导致学习模型的性能下降. 为了解决这些问题, 线性降维[1−4] 通常用于从数据中提取重要和有用的信息. 线性降维的目的是通过优化一些准则函数对原始特征空间进行适当的线性变换. 主成分分析 (Principal component analysis, PCA) 和线性判别分析 (Linear discriminant analysis, LDA) 是两种流行的线性降维方法. 由于 PCA 和 LDA 的简单性和有效性, 它们已经被广泛应用于许多领域, 如

收稿日期 2021-05-19 录用日期 2021-11-02
Manuscript received May 19, 2021; accepted November 2, 2021
国家自然科学基金 (61976216) 资助
Supported by National Natural Science Foundation of China (61976216)
本文责任编委 杨健
Recommended by Associate Editor YANG Jian
1. 中国矿业大学计算机科学与技术学院 徐州 221116 2. 中国矿业大学矿山数字化教育部工程研究中心 徐州 221116
1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116 2. Digitization of Mine, Engineering Research Center of Ministry of Education, Xuzhou 221116

人脸识别[5]、手写体字符识别[6] 和缺陷诊断[7] 等.

当样本的类别信息可用时, 通常情况下 LDA 在提取数据的鉴别特征方面比 PCA 更有效. 线性判别分析的目标是在变换空间中通过最大化类间距离和最小化类内距离来寻找投影矩阵. 从概率的观点来看, 假设每类样本服从高斯分布且具有不同的类中心以及相同的协方差, 则从 Bayes 最优准则可推导出 LDA.

为了改善线性判别分析的特征提取性能, 各种 LDA 的改进算法[8−10] 已经被提出. 使用最优向量替换各类中心[8] 能提高 LDA 的类信息鉴别能力. 分数阶的 LDA[11] 通过在一系列分数阶中引入加权函数来改善 LDA, 但这增加了获得投影向量的代价. 与 Bayes 错误率相关的近似成对精度准则[12] 在原空间计算各类的权重, 从而改善 LDA 的性能. 几何平均[13], 调和平均[14] 以及加权调和平均[15] 被用来定义判别分析的准则函数. 最不利情况下的线性判别分析[16] 考虑了最近的两个类中心和具有最大方差的类来寻找投影方向. 基于最大−最小距离的目标函数[17] 探索了最近的数据对的性质来取得投影方向. Wasserstein 判别分析[18] 利用正则化 Wasserstein 距离获取类之间的全局和局部信息并优化目标函数取得最佳投影方向.

线性判别分析存在小样本的奇异性[19] 以及非线性数据特征提取[20−21] 等问题. 为了克服 LDA 的小样本奇异性问题, 典型的方法包括 PCA+LDA, 正则化 LDA, 伪逆 LDA 以及张量判别分析[22] 等. 为了有效地处理非线性数据, 各种线性判别分析已被拓宽到基于核函数的判别分析[7, 20−21]. 当训练集随着新数据的加入而变化时或处理的数据量大时, 各种增量学习[23−24] 或在线学习方式被用来获得鉴别分析的投影方向. 文献 [24] 提出了两种形式的增量 LDA : 序列增量 LDA 和块增量 LDA, 它们能有效地获取大数据流的特征空间.

数据在采集或传输过程中可能受到污染, 这使得处理的数据包含噪声或离群点. 但经典线性判别分析对噪声数据具有敏感性, 即获得的投影方向偏离真正的投影方向. 为了降低 LDA 对噪声数据的敏感性, 许多工作致力于用鲁棒的目标函数替换 LDA 的原有目标函数[25−26]. 已有的诸多研究发现, 基于 L_1 范数的目标函数比基于 L_2 范数的目标函数在抑制异常点或噪声方面更有效[27−29]. 因此基于 L_1 范数的判别分析方法近年来备受关注. L_1 范数的 LDA[28] 的类内距离和类间距离的定义依赖于 L_1 范数, 这在某种程度上能抑制噪声. L_1 范数的核 LDA 不仅能抑制噪声, 而且能捕捉数据的非线性鉴别特征[28]. L_1 范数的两维 LDA[30] 拓宽了 L_1 范数的 LDA, 这种方法可直接处理图像数据, 而不需要把图像转化为向量形式. 通常 L_1 范数的判别分析通过贪婪算法获取多个投影方向, 而非贪婪迭代算法[31]被用来直接获取 L_1 范数的 LDA 的多个投影向量. 广义弹性网[32] 通过 L_p 范数定义的目标函数来改善判别分析抑制噪声的能力, 而通过优化 Bhattacharyya 的 L_1 范数误差界[33] 可设计出新的鉴别分析模型. 最近提出的基于 L_{21} 范数的 LDA 方法[34] 通过同时优化类中心和投影方向从而在噪声数据方面表现出良好的性能.

在大多数判别分析中, 通常假定类内各个样本以相等的概率 (均匀分布) 取得的, 但是位于类中心附近的样本一般远远多于位于类边界附近的样本. 为了增加类内样本采样的多样性, 可令类内样本的采样概率在均匀分布的概率附近变化, 这种变化有利于区分类中心附近的样本或类边界附近的样本. 不确定优化中的不确定集能描述概率分布的变化范围. 因此本文借助 KL 散度定义的不确定集对类内样本信息进行概率建模. 此外, 为了更好描述各类中心的信息, 本文也利用 KL 散度定义的不确定集对其进行概率建模. 基于此, 本文提出了基于 KL 散度不确定集的线性判别分析方法, 从而进一步改善已有线性判别分析方法. 与以往的方法不同, 本文不仅考虑了一般范数的目标函数, 而且利用不确定集对训练样本信息进行了刻画. 本文采用的不确定集为围绕均匀分布的 KL 散度球且约束中的不确定集被转化为目标函数的正则化项. 本文的主要贡献表现为:

1) 提出了正则化对抗 LDA 和正则化乐观 LDA. 正则化对抗 LDA 优先考虑了难以区分的样本, 而正则化乐观 LDA 优化考虑了易于区分的样本.

2) 采用了广义 Dinkelbach 算法求解正则化对抗 LDA 或正则化乐观 LDA. 对正则化对抗 LDA 运用投影梯度法求解优化子问题, 而对正则化乐观 LDA 运用交替优化求解优化子问题.

3) 在数据集上表明了当数据没有被污染时, 两种判别分析模型取得可竞争的性能, 但在污染数据的情况下, 正则化乐观 LDA 取得更好的性能. 这也从另一方面说明了本文提供两种模型的目的, 即如果在某些验证数据集上正则化乐观 LDA 的最好性能明显优于正则化对抗 LDA 的最好性能, 那说明训练集包含离群点. 因此通过检查正则化对抗 LDA 和正则化乐观 LDA 的性能可判断训练集是否包含离群点.

1 相关工作

假设维数为 d 的 n 个数据点表示为 $\{\boldsymbol{x}_1, \cdots,$

$\boldsymbol{x}_n\}$, 其中 $\boldsymbol{x}_i \in \mathbf{R}^d \ \ (i = 1, \cdots, n)$. 数据集中每个数据点属于 c 类中的一类, 第 i 类的样本数为 n_i, 这样 $n = \sum_{i=1}^{c} n_i$ 为总的样本数. 定义向量 $\boldsymbol{a} = (a_1, \cdots, a_n)^{\mathrm{T}}$ 的 L_s 范数为 $\|\boldsymbol{a}\|_s = (\sum_{i=1}^{n} |a_i|^s)^{\frac{1}{s}}$, 其中 $s > 0$. 当 $0 < s < 1$ 时, $\|\boldsymbol{a}\|_s$ 并不是严格意义上的范数.

线性判别分析的类间离差矩阵和类内离差矩阵分别被定义为:

$$\boldsymbol{S_b} = \sum_{i=1}^{c} \frac{n_i}{n} (\boldsymbol{m}_i - \bar{\boldsymbol{m}})(\boldsymbol{m}_i - \bar{\boldsymbol{m}})^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{S_w} = \frac{1}{n} \sum_{i=1}^{c} \sum_{k=1}^{n_i} (\boldsymbol{x}_k^i - \boldsymbol{m}_i)(\boldsymbol{x}_k^i - \boldsymbol{m}_i)^{\mathrm{T}} \tag{2}$$

其中 $\bar{\boldsymbol{m}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$ 表示所有样本的均值, $\boldsymbol{m}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \boldsymbol{x}_k^i$ 是第 i 类样本的均值, $\boldsymbol{x}_k^i$ 表示第 i 类的第 k 个样本. 线性判别分析在矩阵对 $(\boldsymbol{S_b}, \boldsymbol{S_w})$ 上执行广义特征值分解取得多个投影方向.

1.1　最不利情况的线性判别分析

在文献 [16] 中, 类间离差矩阵 (1) 被改写成下面形式:

$$\boldsymbol{S_b} = \sum_{i=1}^{c} \sum_{j=i+1}^{c} \omega_{ij} (\boldsymbol{m}_i - \boldsymbol{m}_j)(\boldsymbol{m}_i - \boldsymbol{m}_j)^{\mathrm{T}} \tag{3}$$

其中 $\omega_{ij} = \omega_i \omega_j$, $i, j = 1, \cdots, c$ 以及 $\omega_i = \dfrac{n_i}{n}$ 表示第 i 类样本的先验概率. 在此基础上, 文献 [16] 定义了如下的线性判别分析:

$$\max_{\boldsymbol{W}} \frac{\min_{ij} \{\omega_{ij} \|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i - \boldsymbol{m}_j)\|_2^2\}}{\max_i \left\{ \dfrac{1}{n_i} \displaystyle\sum_{k=1}^{n_i} \|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i - \boldsymbol{m}_i)\|_2^2 \right\}}$$
$$\text{s.t. } \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \tag{4}$$

其中 $\boldsymbol{W}$ 是一个 $d \times m$ 投影矩阵以及 $\boldsymbol{I}_m$ 表示单位矩阵. 在投影空间中, 模型 (4) 通过优化距离最近的两个类中心和具有最大类内距离的类取得最优投影矩阵. 这种方法实际上寻找不利区分样本的最优投影方向. 这种设计思想来源于分类器的设计. 在分类器设计中, 设计的分类器对难以分类的样本 (边界样本) 进行优先考虑, 即赋予较大的采样概率, 从而使设计的分类器具有更好的泛化性能. 模型 (4) 通过优先考虑难以区分的样本取得最优投影矩阵. Zhang 和 Yeung[16] 提出了两种算法求解模型 (4). 一种方法是将 (4) 转化为度量学习问题, 另一种方法是设计了一种基于约束的凹凸优化算法.

1.2　基于 L_1 范数的线性判别分析

在文献 [28] 中,下面优化模型被用来取得投影矩阵 $\boldsymbol{W}$:

$$\max_{\boldsymbol{W}} \frac{\displaystyle\sum_{i=1}^{c} n_i \|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i - \bar{\boldsymbol{m}})\|_1}{\displaystyle\sum_{i=1}^{c} \sum_{\boldsymbol{x}_j \in l_i} \|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{m}_i)\|_1}$$
$$\text{s.t. } \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \tag{5}$$

其中 $\|\cdot\|_1$ 表示向量的 L_1 范数, l_i 表示第 i 类样本的集合. 优化问题 (5) 的目标函数关于变量 $\boldsymbol{W}$ 是非平滑和非凸的. 文献 [28] 首先利用梯度上升法取得一个投影向量, 然后设计了一个贪婪方法来取得多个投影向量.

1.3　基于 L_{21} 范数的线性判别分析

为了改善线性判别分析模型在投影空间的鉴别能力, 文献 [34] 提出了下面的优化模型:

$$\min_{\boldsymbol{W}, \boldsymbol{m}_i} \frac{\displaystyle\sum_{i=1}^{c} \sum_{\boldsymbol{x}_j \in l_i} \|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{m}_i)\|_2}{\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i\|_2}$$
$$\text{s.t. } \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \tag{6}$$

与以前的一些模型不同, 式 (6) 的类平均 $\boldsymbol{m}_i$ $(i = 1, \cdots, c)$ 是优化变量. 模型 (6) 的目标函数实际上对不同样本采用了 L_1 范数, 而对每个样本的约简特征采用了 L_2 范数, 这通常被称为 L_{21} 范数的线性判别分析. 如果数据包含离群点, 基于算术平均取得的类中心可能偏离样本的真正类中心, 而优化的类中心接近真正的类中心. 模型 (6) 比模型 (5) 包含更多的优化变量. 文献 [34] 设计了一个有效的框架求解模型 (6).

2　基于不确定集的正则化线性判别分析

2.1　正则化对抗 LDA

为了度量两个具有公共支集的离散概率分布之间的差异, 文献 [35] 定义了 KL 散度. 假设给定两个离散概率分布 $\boldsymbol{p} = (p_1, \cdots, p_n)$ 和 $\boldsymbol{q} = (q_1, \cdots, q_n)$, 那么这两个概率分布之间的 KL 散度被定义为:

$$KL(\boldsymbol{p}|\boldsymbol{q}) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \tag{7}$$

如果两个离散概率分布相同, 则 KL 散度的值为零. 由于 KL 散度不满足三角不等式, 它实际上不是一个真正的距离度量. 注意到 KL 散度是非对称的, 即 $KL(\boldsymbol{p}|\boldsymbol{q}) \neq KL(\boldsymbol{q}|\boldsymbol{p})$. KL 散度是非负的,

即 $KL(\boldsymbol{p}|\boldsymbol{q}) \geq 0$. 在鲁棒优化中, 基于 KL 散度的不确定集被定义为[36–37]:

$$K = \{\boldsymbol{p}|KL(\boldsymbol{p}|\boldsymbol{q}) \leq \varepsilon\} \tag{8}$$

式 (8) 的不确定集表明, 离散概率分布 $\boldsymbol{p}$ 在给定的离散概率分布 $\boldsymbol{q}$ 附近变化, 其变化范围由一个非负参数 ε 控制. 参数 ε 越大, 不确定集的变化范围越大. 单个 $\boldsymbol{p}$ 是不确定集内的一个具体的概率分布.

为了探索样本的类间信息, 模型 (4) 试图在低维投影空间中最大化最近的两个类中心, 而且它利用 L_2 范数定义了类间距离. 在 L_s 范数的距离测度下, 与式 (4) 相似的类间信息的优化问题 $\max_{\boldsymbol{W}} \min_{i<j} \omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i - \boldsymbol{m}_j)\|_s^s$ 在 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m$ 约束条件下被写成下面形式:

$$\begin{aligned} &\max_{\boldsymbol{W}} \min_{p_{ij}} \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij}\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i - \boldsymbol{m}_j)\|_s^s \\ &\text{s.t.} \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij} = 1,\ p_{ij} \geq 0,\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \end{aligned} \tag{9}$$

模型 (9) 不仅引入了优化变量 p_{ij}, 而且在投影后的类间距离使用了 L_s 范数. 如果 $s = 1$, 那么使用了 L_1 范数定义了类间距离. 如果 $s = 2$, 那么使用了 L_2 范数定义了类间距离. 模型 (9) 说明了优化变量 p_{ij} 在概率单纯形内变化, 即在这个单纯形中寻找距离最近的类中心对, 并试图取得最优投影矩阵. 因此在提取鉴别特征时, 距离最近的两个类中心所起的作用最大. 显然, 它忽略了其他类中心的信息, 使得这种模型利用类中心的信息不完整. 为了解决这个问题, 本文首先将离散概率分布 p_{ij} 看作类中心 $\boldsymbol{m}_i$ 和 $\boldsymbol{m}_j$ 之间的采样概率, 然后将离散概率分布 p_{ij} 限制在一定的范围内, 使得模型变得更加灵活. 这里使用式 (8) 定义的不确定集, 这个不确定集是由离散概率分布定义的. 因此根据 KL 散度定义的不确定集和各类中心信息, 以下优化问题被提出:

$$\begin{aligned} &\max_{\boldsymbol{W}} \min_{p_{ij}} \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij}\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i - \boldsymbol{m}_j)\|_s^s \\ &\text{s.t.} \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij} = 1,\ p_{ij} \geq 0 \\ &\sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij} \ln \frac{p_{ij}}{q_{ij}} \leq \varepsilon,\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \end{aligned} \tag{10}$$

模型 (10) 引入了满足条件 $\sum_{i<j}^{c} q_{ij} = 1$ 的参数 q_{ij}, 该参数 q_{ij} 表示第 i 类中心和第 j 类中心所提供的先验知识. 如果事先没有提供先验知识, 可令 $q_{ij} = \dfrac{2}{c(c-1)}$, 即它服从均匀概率分布, 这是因为不同类中心构成的数据对的数目为 $\dfrac{c(c-1)}{2}$. 参数 q_{ij} 并不是优化变量. 模型 (10) 允许参数 p_{ij} 在给定 q_{ij} 的邻域内变化. 根据正则化理论, 通过引入一个非负参数 η 可将 (10) 的不确定集转化为目标函数的正则化项. 因此模型 (10) 等价于下面的模型:

$$\begin{aligned} &\max_{\boldsymbol{W}} \min_{p_{ij}} F_1(W, p_{ij}) := \eta KL(\boldsymbol{p}|\boldsymbol{q}) + \\ &\qquad \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij}\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i - \boldsymbol{m}_j)\|_s^s \\ &\text{s.t.} \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij} = 1,\ p_{ij} \geq 0,\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \end{aligned} \tag{11}$$

在模型 (11) 中, KL 散度作为正则化项使用了非负参数 η. 参数 η 越大, 对应的不确定集越小. 参数 η 越小, 对应的不确定集就越大. KL 散度是非负的, 因此 $F_1(\boldsymbol{W}, p_{ij})$ 是非负的. 模型 (11) 仅仅考虑了类间样本的分布信息. 对于类内样本的分布信息, 遵循同样的思想, 这里考虑类边界附近的样本, 在投影矩阵的半正交约束下, $\min_{\boldsymbol{W}} \sum_{i=1}^{c} \max_j \|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_j^i - \boldsymbol{m}_i)\|_r^r$ 被改写成下面的形式:

$$\begin{aligned} &\min_{\boldsymbol{W}} \max_{u_{ik}} \sum_{i=1}^{c}\sum_{k=1}^{n_i} u_{ik}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i - \boldsymbol{m}_i)\|_r^r \\ &\text{s.t.} \sum_{k=1}^{n_i} u_{ik} = 1,\ u_{ik} \geq 0,\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \end{aligned} \tag{12}$$

模型 (12) 采用了 L_r 范数定义类内距离, 参数 r 是正的实数. 模型 (4) 中分式的分母考虑了低维空间中具有最大类内距离的类, 而模型 (12) 考虑了每一类中远离其类中心的那些样本, 那些样本可能位于类之间的交界处. 换言之, 它搜索一些样本, 使它们在类内信息中起主导作用. 这样 u_{ik} 可看成第 i 类的第 k 个样本的采样概率. 通过将 u_{ik} 限制在一个不确定集内, 使样本的采样概率发生变化. 根据 (12) 和不确定集可定义如下的优化问题:

$$\begin{aligned} &\min_{\boldsymbol{W}} \max_{u_{ik}} F_2(\boldsymbol{W}, u_{ik}) := \\ &\qquad \sum_{i=1}^{c}\sum_{k=1}^{n_i} u_{ik}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i - \boldsymbol{m}_i)\|_r^r - \\ &\qquad \sum_{i=1}^{c}\sum_{k=1}^{n_i} \lambda_i u_{ik} \ln \frac{u_{ik}}{v_{ik}} \\ &\text{s.t.}\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m,\ \sum_{k=1}^{n_i} u_{ik} = 1,\ u_{ik} \geq 0 \end{aligned} \tag{13}$$

模型 (13) 为每类引入了非负参数 λ_i $(i = 1, \cdots, c)$, 这些非负参数在类内信息和不确定集之间作出

适当的折衷. 参数 v_{ik} 表示不确定集的中心, 并满足 $\sum_{k=1}^{n_i} v_{ik}=1,\ v_{ik}\geq 0$. 通常每类的 v_{ik} 为均匀分布, 即 $v_{ik}=\dfrac{1}{n_i},\ k=1,\cdots,n_i$. 模型 (11) 通过对各类中心进行概率建模探索了类间信息, 而模型 (13) 通过对类内样本的采样概率进行建模探索了类内信息. 为了同时利用类内信息和类间信息, 从式 (11) 和 (13) 可建立如下单目标优化模型:

$$\begin{aligned}&\min_{\boldsymbol{W}}\frac{\max_{u_{ik}}F_2(\boldsymbol{W},u_{ik})}{\min_{p_{ij}}F_1(\boldsymbol{W},p_{ij})}\\&\text{s.t. }\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}_m,\sum_{i=1}^{c}\sum_{j=i+1}^{c}p_{ij}=1\\&p_{ij}\geq 0,\ u_{ik}\geq 0,\sum_{k=1}^{n_i}u_{ik}=1\end{aligned}\tag{14}$$

式 (14) 的内层优化是关于变量 u_{ik} 和 p_{ij} 的优化问题以及外层优化是关于变量 $\boldsymbol{W}$ 的优化问题. 模型的内层优化优先考虑了可分离性不好的样本 (边界附近的样本有大的采样概率), 而外层优化使得它们在投影空间的可分离性变好. 从 (12), (13) 以及 (14) 知, 内层优化和外层优化的目标具有不一致 (对抗) 的性质. 本文将模型 (14) 称为正则化对抗 LDA(Regularized adversarial LDA, RALDA). 式 (14) 的内层优化取得的概率分布被称为不确定集的对抗概率分布. 模型 (14) 是一个非凸优化问题. 模型 (14) 包含三组优化变量, 即 $\boldsymbol{W}$, u_{ik} 和 p_{ij}. 如果同时考虑这三组优化变量, 这实际上是一个鞍点优化问题. 如果令 $u_{ik}^*(\boldsymbol{W})=\arg\max_{u_{ik}}F_2(\boldsymbol{W},u_{ik})$, 那么从 $u_{ik}^*(W)$ 可得到关于 $\boldsymbol{W}$ 的梯度信息[38], 这是因为 $F_2(\boldsymbol{W},u_{ik})$ 是变量 u_{ik} 的严格凸函数. 同样地令 $p_{ij}^*(W)=\arg\max_{p_{ij}}F_1(\boldsymbol{W},p_{ij})$, 从 $p_{ij}^*(\boldsymbol{W})$ 可得到关于 $\boldsymbol{W}$ 的梯度信息. 换句话说, 此情况下可计算 $u_{ik}^*(\boldsymbol{W})$ 和 $p_{ij}^*(W)$ 关于 $\boldsymbol{W}$ 的 (次) 梯度. 注意非可微函数采用了次梯度. 通过下面两个优化问题可取得 $u_{ik}^*(\boldsymbol{W})$ 和 $p_{ij}^*(\boldsymbol{W})$:

$$\begin{aligned}&\max_{u_{ik}}F_2(\boldsymbol{W},u_{ik})\\&\text{s.t. }\sum_{k=1}^{n_i}u_{ik}=1,\ u_{ik}\geq 0\end{aligned}\tag{15}$$

$$\begin{aligned}&\min_{p_{ij}}F_1(\boldsymbol{W},p_{ij})\\&\text{s.t. }\sum_{i<j}^{c}p_{ij}=1,\ p_{ij}\geq 0\end{aligned}\tag{16}$$

优化问题 (15) 对应模型 (14) 中分式的分子, 优化问题 (16) 对应模型 (14) 中分式的分母. 注意到函数 $F_1(\boldsymbol{W},p_{ij})$ 是变量 p_{ij} 的强凸函数以及 $F_2(\boldsymbol{W},u_{ik})$ 是变量 u_{ik} 的强凹函数, 这说明优化问题 (15) 和 (16) 分别存在唯一解. 式 (15) 和 (16) 的唯一解分别表示为:

$$u_{ik}^*=\frac{v_{ik}\exp(\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i-\boldsymbol{m}_i)\|_r^r/\lambda_i)}{\sum\limits_{k=1}^{n_i}v_{ik}\exp(\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i-\boldsymbol{m}_i)\|_r^r/\lambda_i)}\tag{17}$$

$$p_{ij}^*=\frac{q_{ij}\exp(-\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i-\boldsymbol{m}_j)\|_s^s/\eta)}{\sum\limits_{i=1}^{c}\sum\limits_{j=i+1}^{c}q_{ij}\exp(-\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i-\boldsymbol{m}_j)\|_s^s/\eta)}\tag{18}$$

把式 (17) 代入 $\max_{u_{ik}}F_2(\boldsymbol{W},u_{ik})$ 可取得:

$$\tilde{F}_2(\boldsymbol{W})=\max_{u_{ik}}F_2(\boldsymbol{W},u_{ik})=\sum_{i=1}^{c}\lambda_i\ln\sum_{k=1}^{n_i}v_{ik}\exp(\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i-\boldsymbol{m}_i)\|_r^r/\lambda_i)\tag{19}$$

把式 (18) 代入 $\min_{p_{ij}}F_1(\boldsymbol{W},p_{ij})$ 可取得:

$$\tilde{F}_1(\boldsymbol{W})=\min_{p_{ij}}F_1(\boldsymbol{W},p_{ij})=-\eta\ln\sum_{i=1}^{c}\sum_{j=i+1}^{c}q_{ij}\exp(-\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i-\boldsymbol{m}_j)\|_s^s/\eta)\tag{20}$$

式 (19) 和 (20) 实际上求解模型 (14) 的内层优化问题. 利用式 (19) 和 (20), 优化模型 (14) 被改写成下面形式:

$$\begin{aligned}&\min_{W}F(\boldsymbol{W}):=\frac{\tilde{F}_2(\boldsymbol{W})}{\tilde{F}_1(\boldsymbol{W})}\\&\text{s.t. }\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}_m\end{aligned}\tag{21}$$

模型 (21) 仅包含优化变量 $\boldsymbol{W}$. 它是一个非常复杂的非线性优化问题. 从模型 (21) 的约束来看, 它属于 Stiefel 流形上的优化问题[41].

2.2　求解优化模型 (21)

式 (21) 是一个比率优化问题. 从 $F_1(\boldsymbol{W},p_{ij})$ 的非负性和式 (20) 可得出 $\tilde{F}_1(\boldsymbol{W})$ 是非负的. 从 (19) 可推导出 $\tilde{F}_2(\boldsymbol{W})$ 是非负的. 这样式 (21) 中的函数 $F(\boldsymbol{W})$ 是非负的. 尽管存在求解模型 (21) 的诸多优化算法, 但本文采用了广义 Dinkelbach 算法[39] 求解模型 (21). 算法 1 描述了求解式 (21) 的主要步骤.

算法 1. 求解模型 (21) 的广义 Dinkelbach 算法

1) 令 $\boldsymbol{W}^1$ 是满足 $(\boldsymbol{W}^1)^{\mathrm{T}}\boldsymbol{W}^1=\boldsymbol{I}_m$ 的初始可行解, 设定参数 $\lambda_i,\ \eta,\ q_{ij}=2/(c(c-1)),\ u_{ik}=1/n_i,\ \gamma_1=F(\boldsymbol{W}^1)$,

2) 对于 $t=1$ to T 执行

a) 求解 $\bar{F}(\gamma_t)=\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}_m}\{\tilde{F}_2(\boldsymbol{W})-\gamma_t\tilde{F}_1(\boldsymbol{W})\}$, 其解表示为 $\boldsymbol{W}^{t+1}$,

b) 如果 $\bar{F}(\gamma_t)=0$, 停止且 $\boldsymbol{W}^{t+1}$ 是最优解, 否则取得 $\gamma_{t+1}=F(\boldsymbol{W}^{t+1})$,

3) 输出投影矩阵 $\boldsymbol{W}$.

算法 1 把比率优化问题 (21) 归结为两个函数差的优化问题. 文献 [27] 已采用这种思想求解 L1 范数的 LDA, 从而避免矩阵逆的计算, 这也克服小样本的奇异性问题. 文献 [27] 的方法实际上是广义 Dinkelbach 算法的一种形式. 这样本文的算法也不会出现矩阵逆计算问题. 如果 KL 散度的值不为 0, $F_1(\boldsymbol{W}, p_{ij})$ 作为式 (14) 的分母也不为零. 这些因素促使了提出的算法克服了小样本的奇异性问题.

算法 1 的停止条件是 $\bar{F}(\gamma_t)=0$. 在实际应用中, 由于受浮点数计算精度的影响, 这通常是不可行的. 通常采用最大迭代次数或目标函数的相对变化小于一个数作为停止条件. 如文献 [39] 所述, 如果约束集是闭有界集且优化子问题取得精确解, 则算法 1 将在有限次迭代中结束, 或者它将产生一个无穷序列, 其极限点是式 (21) 的平稳点. 注意到投影矩阵 W 位于一个闭有界集中, 并且模型 (21) 的目标函数是连续的. 因此式 (21) 的目标函数是有上下界的. 在实际求解中可能无法取得优化子问题的精确解. 如果采用闭映射的下降算法求解优化子问题, 那么算法 1 在这种情况下也收敛[39]. 因此本文利用梯度下降算法解决优化子问题以确保算法 1 的收敛性. 从算法 1 可看出, 其关键问题是如何求解优化子问题. 算法 1 中的子问题被表示如下:

$$\bar{F}(\gamma_t)=\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}_m}\{\tilde{F}_2(\boldsymbol{W})-\gamma_t\tilde{F}_1(\boldsymbol{W})\} \tag{22}$$

从 $F(\boldsymbol{W}^1)$的非负性可知 γ_t是非负的. 如文献 [39] 所述, 函数 $\bar{F}(\gamma_t)$ 是变量 γ_t 的凹函数和递减函数, 并且 $\bar{F}(\gamma_t)=0$ 有唯一解. 算法 1 的优点在于不需要取得优化子问题的精确解, 只要 $\boldsymbol{W}$ 满足 $\{\tilde{F}_2(\boldsymbol{W})-\gamma_t\tilde{F}_1(\boldsymbol{W})\}<0$, 那么 $\boldsymbol{W}$ 比前一次的解更接近目标函数 $F(\boldsymbol{W})$ 的最优解. 这样可采用简单的梯度投影法求解优化问题 (22). 本文采用了基于 Armijo 线搜索的投影次梯度法求解模型 (22). 为了方便描述算法, 这里定义了梯度 $\dfrac{\partial\tilde{F}(\boldsymbol{W})}{\partial\boldsymbol{W}}=\dfrac{\partial\tilde{F}_2(\boldsymbol{W})-\lambda_t\tilde{F}_1(\boldsymbol{W})}{\partial\boldsymbol{W}}$. 算法 2 描述了求解模型 (22) 的主要步骤.

算法 2. 求解模型 (22) 的投影次梯度法

1) 初始的 W, Armijo 参数 $0<\beta<1$,

2) 重复直到收敛

a) 计算 (次) 梯度 $\dfrac{\partial\tilde{F}(\boldsymbol{W})}{\partial\boldsymbol{W}}$ 和 $g=1$,

b) 计算 $\boldsymbol{W}_n=P(\boldsymbol{W}-\beta^g\dfrac{\partial\tilde{F}(\boldsymbol{W})}{\partial\boldsymbol{W}})$,

c) 计算 $\tilde{F}(\boldsymbol{W}_n)$; 如果 $\tilde{F}(\boldsymbol{W}_n)<\tilde{F}(\boldsymbol{W})$, 那么 $\boldsymbol{W}=\boldsymbol{W}_n$; 否则 $g=g+1$, 转到 b).

算法 2 定义了保证投影矩阵半正交性的投影算子 $P(\boldsymbol{W})=\boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-\frac{1}{2}}$, 即 $(P(\boldsymbol{W}))^{\mathrm{T}}P(\boldsymbol{W})=\boldsymbol{I}_m$. 从算法 1 可知, 算法 1 涉及函数 $F(\boldsymbol{W})$ 的计算和算法 2 的运行. 计算 $F(\boldsymbol{W})$ 的复杂度为 $\mathrm{O}(c^2dm+ndm)$ 以及算法 2 的复杂度在于梯度的计算. 梯度计算的复杂度为 $\mathrm{O}(c^2dm+ndm)$. 这样算法 1 的计算复杂度为 $\mathrm{O}(T_1((c^2dm+ndm)+T_2(c^2dm+ndm)))$, 其中 T_2 为算法 2 的迭代次数以及 T_1 为算法 1 的迭代次数.

2.3 正则化乐观判别分析

正则化对抗判别分析探索了不确定集中的对抗概率分布, 这实际上优化考虑了训练集中位于类边界附近的样本 (远离类中心的样本). 然而当数据集包含离群点时, 这些离群点可能远离各类中心. 在这种情况下, 优先考虑类中心附近的样本是有益的, 即对接近类中心附近的样本分配大的采样概率. 不同于第 2.1 节中的模型, 本小节试图在不确定集中寻找另外一种概率分布, 这对应于优先考虑类中心附近的样本点. 因此借助不确定集以及类内和类间信息可定义以下优化问题:

$$\begin{aligned}&\max_{\boldsymbol{W}}\max_{p_{ij}}H_1(\boldsymbol{W},p_{ij}):=-\eta KL(\boldsymbol{p}|\boldsymbol{q})+\\&\qquad\sum_{i=1}^{c}\sum_{j=i+1}^{c}p_{ij}\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i-\boldsymbol{m}_j)\|_s^s\\&\text{s.t.}\sum_{i=1}^{c}\sum_{j=i+1}^{c}p_{ij}=1,\ p_{ij}\geq 0,\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}_m\end{aligned} \tag{23}$$

$$\begin{aligned}&\min_{\boldsymbol{W},u_{ik}}H_2(\boldsymbol{W},u_{ik}):=\sum_{i=1}^{c}\sum_{k=1}^{n_i}\lambda_iu_{ik}\ln\frac{u_{ik}}{v_{ik}}+\\&\qquad\sum_{i=1}^{c}\sum_{k=1}^{n_i}u_{ik}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i-\boldsymbol{m}_i)\|_r^r\\&\text{s.t.}\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}_m,\sum_{k=1}^{n_i}u_{ik}=1,\ u_{ik}\geq 0\end{aligned} \tag{24}$$

模型 (23) 优先考虑了距离大的类中心对, 这实际上优先考虑区分性比较好的类. 因为 $KL(p|q)$ 取非负值, 这不能从 $H_1(\boldsymbol{W},p_{ij})$ 的定义得出其非负性. 对于模型 (24), 它优先考虑区分性比较好的样本, 即对类中心附近的样本赋予大的采样概率和类边界附近的样本赋予小的采样概率. 当处理的数据包含异常点或噪声时, 它优先考虑了那些可能不是异常点的样本, 即对异常点赋予小的采样概率. 因为 KL 散度是非负的, 从式 (24) 知, $H_2(\boldsymbol{W},u_{ik})$ 不

小于零. 根据 (23) 和 (24) 可建立以下单目标函数:

$$\begin{aligned}&\min_{\boldsymbol{W}} H(\boldsymbol{W}) := \frac{\min_{u_{ik}} H_2(\boldsymbol{W}, u_{ik})}{\max_{p_{ij}} H_1(\boldsymbol{W}, p_{ij})}\\ &\text{s.t. } \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m, \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij} = 1\\ &p_{ij} \geq 0,\, u_{ik} \geq 0,\, \sum_{k=1}^{n_i} u_{ik} = 1 \end{aligned} \tag{25}$$

从 (23), (24) 和 (25) 知, 外层优化 (优化 $\boldsymbol{W}$) 和内层优化 (优化 u_{ik}, p_{ij}) 的目标有一致的性质. 根据不确定规划中的乐观模型的概念, 本文将模型 (25) 称为正则化乐观 LDA(Regularized optimistic LDA, ROLDA). 内层优化取得的概率分布被称为不确定集的乐观概率分布. 模型 (25) 利用不确定集优先考虑类中心附近的样本点. 对训练集而言, 模型 (25) 优先考虑了区分性好的类和样本. 如果数据集包含异常点或噪声, 那么这些异常点或噪声可能远离类中心且具有较低的采样概率. 因此模型 (25) 在某种程度上能抑制异常点或噪声. 模型 (25) 是非凸优化问题, 同样地也采用了广义 Dinkelbach 算法求解模型 (25). 求解式 (25) 涉及下面优化子问题:

$$\begin{aligned}&\bar{H}(\gamma_t) = \min_{\boldsymbol{W}}\{\min_{u_{ik}} H_2(\boldsymbol{W}, u_{ik}) -\\ &\qquad \gamma_t \max_{p_{ij}} H_1(\boldsymbol{W}, p_{ij})\}\\ &\text{s.t. } \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij} = 1, \sum_{k=1}^{n_i} u_{ik} = 1\\ &p_{ij} \geq 0,\, u_{ik} \geq 0,\, \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \end{aligned} \tag{26}$$

其中 γ_t 是来源于 $H(\boldsymbol{W})$ 的值. 函数 $H_1(\boldsymbol{W}, p_{ij})$ 关于变量 p_{ij} 是强凹函数, 函数 $H_2(\boldsymbol{W}, u_{ik})$ 关于变量 u_{ik}是强凸函数. 这样 p_{ij}和 u_{ik}存在唯一解, 分别表示为:

$$p_{ij} = \frac{q_{ij}\exp(\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i - \boldsymbol{m}_j)\|_s^s/\eta)}{\sum\limits_{i<j} q_{ij}\exp(\omega_{ij}\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_i - \boldsymbol{m}_j)\|_s^s/\eta)} \tag{27}$$

$$u_{ik} = \frac{v_{ik}\exp(-\left\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i - \boldsymbol{m}_i)\right\|_r^r/\lambda_i)}{\sum\limits_{k=1}^{n_i} v_{ik}\exp(-\|\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_k^i - \boldsymbol{m}_i)\|_r^r/\lambda_i)} \tag{28}$$

因为简化 $H_1(\boldsymbol{W}, p_{ij})$ 和 $H_2(\boldsymbol{W}, u_{ik})$ 会导致模型是复杂的非线性问题, 所以没有借助式 (27) 和 (28) 简化 $H_1(\boldsymbol{W}, p_{ij})$ 和 $H_2(\boldsymbol{W}, u_{ik})$. 注意到如果把式 (27) 代入 $H_1(\boldsymbol{W}, p_{ij})$, 那么可得出 $\max_{p_{ij}} H_1(\boldsymbol{W}, p_{ij})$ 是非负的. 如果把式 (28) 代入 $H_2(\boldsymbol{W}, u_{ik})$, 那么可得出 $min_{u_{ik}} H_2(\boldsymbol{W}, u_{ik})$ 是非负的. 因此 γ_t 是非负的. 考虑到 γ_t 的非负性, 模型 (26) 可转化为下面的优化问题:

$$\begin{aligned}&\min_{\boldsymbol{W}} \min_{u_{ik}, p_{ij}} \{H_2(\boldsymbol{W}, u_{ik}) - \gamma_t H_1(\boldsymbol{W}, p_{ij})\}\\ &\text{s.t. } \sum_{i=1}^{c}\sum_{j=i+1}^{c} p_{ij} = 1, \sum_{k=1}^{n_i} u_{ik} = 1\\ &p_{ij} \geq 0,\, u_{ik} \geq 0,\, \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_m \end{aligned} \tag{29}$$

模型 (29) 属于约束的最小化问题. 因为约束集是闭有界集且目标函数是连续的, 模型 (29) 存在解. 注意到优化变量 u_{ik}, p_{ij} 以及 $\boldsymbol{W}$ 的约束是独立的, 这样交替优化算法能有效地求解模型 (29). 模型 (22) 仅涉及优化变量 $\boldsymbol{W}$ 且目标函数是复杂的非线性优化问题, 但模型 (29) 涉及三组优化变量, 每一组优化变量对应一个优化子问题. 算法 3 概括了求解模型 (29) 的过程.

算法 3. 求解式 (29) 的交替优化算法

1) 给定初始 p_{ij} 和 u_{ik},

2) 重复直到收敛

a) 固定 p_{ij} 和 u_{ik}, 更新 $\boldsymbol{W}$,

b) 固定 u_{ik} 和 $\boldsymbol{W}$, 通过 (27) 更新 p_{ij},

c) 固定 p_{ij} 和 $\boldsymbol{W}$, 通过 (28) 更新 u_{ik}.

如果 $s = r = 2$, 算法 1 中的步骤 a) 可通过特征值分解取得投影矩阵 $\boldsymbol{W}$. 在其他情况下, 可通过在 Stiefel 流形上的梯度下降法取得投影矩阵 $\boldsymbol{W}$. 不同于正则化对抗 LDA 的求解, 如果采用 Stiefel 流形上的梯度下降法取得 $\boldsymbol{W}$ [41], 此时函数的梯度并不包含指数函数, 这实际上把指数函数放在了 p_{ij} 和 u_{ik} 的更新上. 在 $s = r = 2$ 情况下, 取得 W 的计算复杂度为 $\mathrm{O}(d^3)$. 当样本的维数较大时, 执行特征值分解可能比较耗时, 那么可采用流形上的梯度下降法取得 $\boldsymbol{W}$. 采用 Stiefel 流形上的梯度下降法取得 $\boldsymbol{W}$ 的复杂度为 $\mathrm{O}(T_g(c^2dm + ndm))$, 其中 T_g 为梯度下降法的迭代次数. 更新 p_{ij} 和 u_{ik} 的计算复杂度分别为 $\mathrm{O}(c^2dm)$ 和 $\mathrm{O}(ndm)$. 这样如果 $s = r = 2$ 且采用特征值分解取得投影矩阵 $\boldsymbol{W}$, 那么算法的计算复杂度为 $\mathrm{O}(T_3(c^2dm + ndm + d^3))$, 其中 T_3 为算法 3 的迭代次数. 如果采用 Stiefel 流形上的梯度下降法取得投影矩阵 $\boldsymbol{W}$, 那么算法 3 的计算复杂度为 $\mathrm{O}(T_3(c^2dm + ndm + T_g(c^2dm + ndm)))$.

2.4　与其他线性判别分析的关系

从式 (17) 和 (18) 可得到如下推论.

推论 1. 假定 p_{ij}^* 和 u_{ik}^* 如式 (17) 和 (18) 那样定义, 那么可得出

$$\lim_{\eta\to\infty} p_{ij}^* = q_{ij}, \qquad \lim_{\lambda_i\to\infty} u_{ik}^* = v_{ik}$$

推论 1 说明了当参数 η 和 λ_i 趋向正无穷大时,

p_{ij} 和事先定义的 q_{ij} 一致, 以及 u_{ik} 和事先定义的 v_{ik} 一致. 在这种情况下, 根据 q_{ij} 和 v_{ik} 定义可得出 $p_{ij}=2/(c(c-1))$ 和 $u_{ik}=1/n_i$. 此时正则化对抗 LDA 退化为不同范数的 LDA. 即当 $s=r=2$ 时, 模型 (14) 成为 L_2 范数的 LDA; 当 $s=r=1$ 时, 模型 (14) 成为 L_1 范数的 LDA. 当参数 η 和 λ_i 趋向 0 时, 模型 (14) 变成了最强对抗概率分布下的线性判别分析. 这样参数 η 和 λ_i 的变化为模型 (14) 和各种范数的 LDA 建立了联系. 因此 RALDA 推广了以前的模型.

从式 (27) 和 (28) 可知: 当参数 η 和 λ_i 趋向正无穷大时, $p_{ij}=2/(c(c-1))$ 和 $u_{ik}=1/n_i$. 此时正则化乐观 LDA 退化为不同范数的 LDA. 即 $s=r=2$ 时, 模型 (25) 成为 L_2 范数的 LDA; 当 $s=r=1$ 时, 模型 (25) 成为 L_1 范数的 LDA. 当参数 η 和 λ_i 趋向 0 时, 模型 (25) 变成了最乐观概率分布下的线性判别分析. 从式 (17) 和 (18), 以及式 (27) 和 (28) 可知, RALDA 和 ROLDA 中的 p_{ij} 和 u_{ik} 是不同的. 从 (17) 可知类边界附近的样本具有大的采样概率; 从 (28) 可知类中心附近的样本具有大的采样概率. 这样模型 (14) 优先考虑那些不利区分的样本, 而模型 (25) 优先考虑了那些有利区分的样本. 尽管这两种模型的机制是不同的, 但是当参数趋向无穷时, 它们是等价的.

3 实验结果

本节通过在数据集上的一系列实验来评估所提出模型的有效性. 当涉及到分类问题时, 本文采用了最近邻分类器且度量准则采用了欧氏范数. 为了进行比较, 本文编程实现了几种鲁棒的特征提取方法, 包括 L1-LDA[6]、最不利 LDA (Worst-case LDA, WLDA[16])、LDA-L1[28] 和 L21-LDA[34]. L1-LDA 和 LDA-L1 的参数设置与文献 [34] 中的相同. 注意到提出的模型涉及多个参数. 为了简单起见, 令参数 $\lambda=\lambda_1=\cdots=\lambda_c$, 即所有的类都被赋予了相同的参数 λ, 这样模型的参数被约简为两个参数 λ 和 η, 两个参数取自集合 $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. 对于参数 r 和 s, 本文只考虑参数取特殊值的情况, 即把 $s=r=1$ 和 $s=r=2$ 的 RALDA 分别简记为 L1RALDA 和 L2RALDA, 以及把 $s=r=1$ 和 $s=r=2$ 的 ROLDA 分别简记为 L1ROLDA 和 L2ROLDA. 迭代方法的初始解采用了 LDA 的正交化投影矩阵. 实验环境是一台内存为 8 GB 的奔腾 1.6 GHz 计算机和 Matlab (7.0.0) 编程语言.

3.1 人脸和物体数据集的实验

本小节在四个人脸数据库 (Yale、ORL、UMIST 和 AR) 上和一个物体数据库 (COIL) 上测试了提出的模型. ORL 人脸数据库包含 40 个人, 每个人都有 10 幅不同的图像, 所有的图像都是在一个均匀背景下拍摄的. UMIST 人脸数据库包含 20 个人的 564 幅人脸图像, 每个人都有不同的种族、性别和外貌, 例如不同的表情、照明、戴眼镜/不戴眼镜、留胡须/不留胡须、不同的发型. 耶鲁大学的人脸数据库包含 15 个人的 165 幅灰度图像, 其人脸图像涉及光照条件和面部表情的变化. AR 人脸数据库包含 4 000 多幅彩色人脸图像, 每个人有不同的面部表情、光照条件和遮挡. AR 人脸图像在两个不同的时间段拍摄且时间间隔是两周, 每阶段获取每个人的 13 幅图像. COIL 数据库包含 1 440 幅灰度图像和 20 个物体的黑色背景, 每个物体有 72 幅不同的图像. 为了提高计算效率, 将所有图像归一化为 32×32 大小的灰度图像.

考虑到提出的算法是一种迭代算法, 第一组实验测试了算法的收敛性. 实验数据来自 Yale 数据集. 我们随机选取每个人的四幅图像组成训练集, 其它图像用于测试. 假设约简维数为 14. 训练集中有一半的样本被矩形噪声污染. 矩形噪声包含等概率的白点和黑点, 它们在图像中的位置是随机的且块的大小是 32×32 像素. 在实施算法 2 时, 令 $\beta=0.5$. 图 1 列出了 L2RALDA, L1RALDA, L2ROLDA 和 L1ROLDA 四种方法的收敛性. 为了简单性, 在这组实验中令 $\lambda=\eta$.

从图 1 可知, 四种算法的目标函数值随着迭代次数的增加而递减. 当迭代次数超过某一数值时, 目标函数值几乎保持稳定. 文献 [27] 指出: 当出现小样本的奇异性问题时, 采用梯度上升法求解 L1 范数的判别分析 (目标函数为最大化问题) 可能使得目标函数值发生振荡现象. 在这组实验中, 图像的维数远远大于训练样本的个数, 这存在奇异性问题, 但图 1 显示了提出的算法的目标函数值并没有出现振荡的情况, 这意味着提出的算法在某种程度上避免了小样本的奇异性问题. 从图 1 可观察到基于 L_1 范数的模型比基于 L_2 范数的模型一般需要更多的迭代次数, 这是因为基于 L_1 范数的模型有不可微的目标函数. 注意到不同的参数也影响算法的收敛速度. 因此在后面的实验中, 本文设定算法的停止条件为最大迭代次数为 50 或目标函数值的相对改变不超过 10^{-4}.

在约简的参数下提出的模型有两个重要参数 λ 和 η. 这组实验探索了参数 λ 和 η 对识别性能的影响. 对于图像识别问题, 为了降低计算代价, 采用主成分分析降维但保持图像的百分之九十九的能量. 对于耶鲁数据集, 随机选取每个人的四幅图像组成

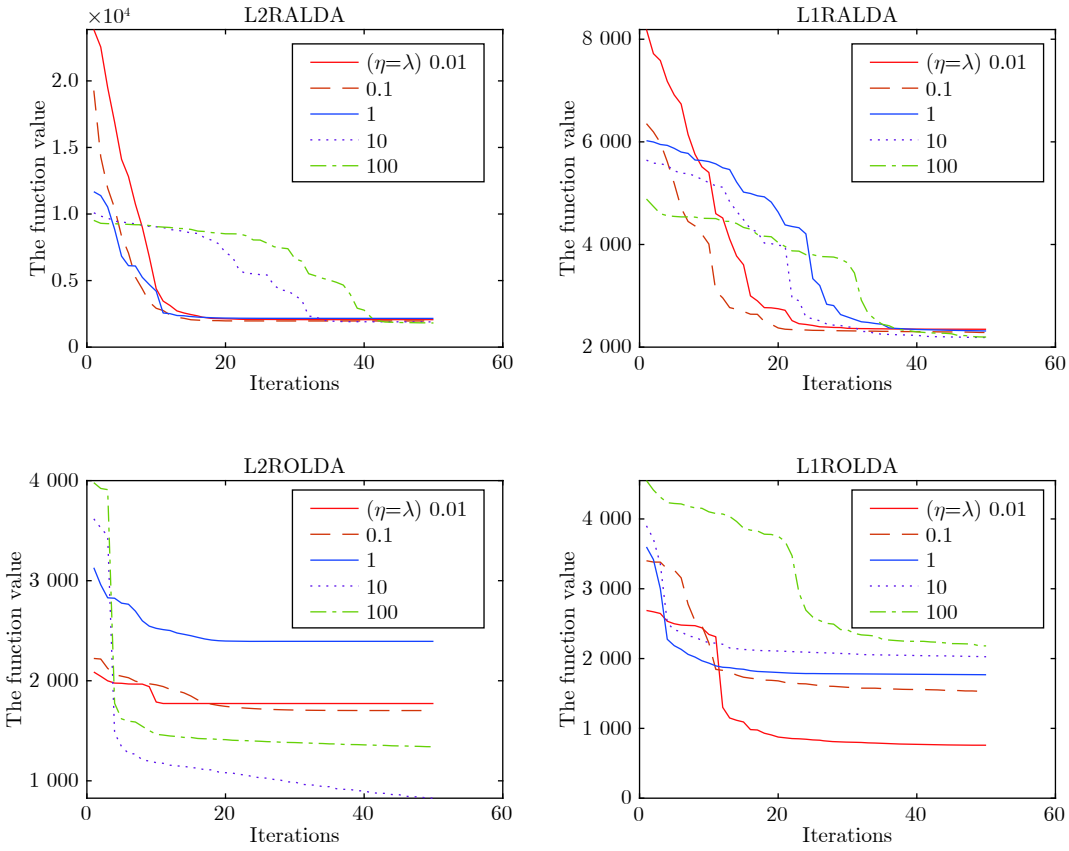


图 1　L2RALDA, L1RALDA, L2ROLDA 和 L1ROLDA 的收敛性分析

Fig. 1　Convergence analysis of L2RALDA, L1RALDA, L2ROLDA and L1ROLDA

训练集, 其他样本用于测试. 假设约简维数为 14, 固定 $r = s = 1$ 或 $r = s = 2$, 参数 λ 和 η 从集合 $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ 中取值. 因此每个参数取七个值. 实验结果来自十次的平均实验. 图 2 显示了每种算法随参数变化的错误率. 在图 2 的每个子图中, 其中 x 轴表示对数尺度的参数 λ, y 轴表示对数尺度的参数 η, 以及 z 轴表示算法的错误率.

从图 2 可看出, 模型的错误率随着参数的变化而变化. 当参数 λ 和 η 取较小值时, L1ROLDA 的错误率较低. 对于 L2ROLDA, 如果参数 η 取较大值且参数 λ 取较小值, 则分类性能较好. 对于 L1RALDA, 如果参数 η 取较小值且参数 λ 取较大值, 则在很大范围内获得较低的错误率. 对于 L2RALDA, 其参数也影响算法的性能. 这组实验表明了 L1ROLDA 和 L1RALDA 的作用机理是不同的. 因此在实际应用中需要选择合适的参数才能获得最佳的分类性能, 这可通过交叉验证等方法来获得最佳参数.

为了比较各种模型的特征提取性能, 在 Yale、ORL、UMIST 和 COIL 数据集上随机选取每个人的 4 幅图像构成训练集, 其余图像用作测试目的. 为了评价算法的鲁棒性, 在 Yale、ORL、UMIST 和 COIL 数据集上人工模拟了遮挡图像, 那就是训练集中有一半的样本被矩形噪声污染. 矩形噪声包含等概率的白点和黑点, 它们在图像中的位置是随机的, 块的大小是 20×20 像素. 由于 AR 数据集包含实际遮挡的情况, 本文考虑了 120 个人的两种遮挡情况: 其一是太阳镜遮挡, 其二是围巾遮挡. 在 AR 数据集上, 对于太阳镜遮挡, 从第一时间段的 7 幅无遮挡图像中随机选择 3 幅图像, 然后将这些图像与随机选择的 3 幅太阳镜遮挡的图像构成训练集. 第二时间段得到的 7 幅无遮挡的图像被用做测试集. 因此每个人的训练样本数为 6. 对于围巾遮挡, 从第一时间段的 7 幅无遮挡图像中随机选择 3 幅图像, 然后将这些图像与随机选择的 3 幅围巾遮挡图像构成训练集, 第二时间段得到的 7 幅无遮挡的图像被用做测试集.

图 3 显示了在 Yale 数据库上几种算法的错误率随约简特征变化的情况. 表 1 列出了各种方法的最好性能. 另外表 1 中给出了各个人脸数据集上和 COIL 数据集上的最好性能. 每种方法的参数在额外的五次运行上得到. 表 1 的实验结果来自 20 次随机运行的平均, 其中 C- 表示污染的数据集.

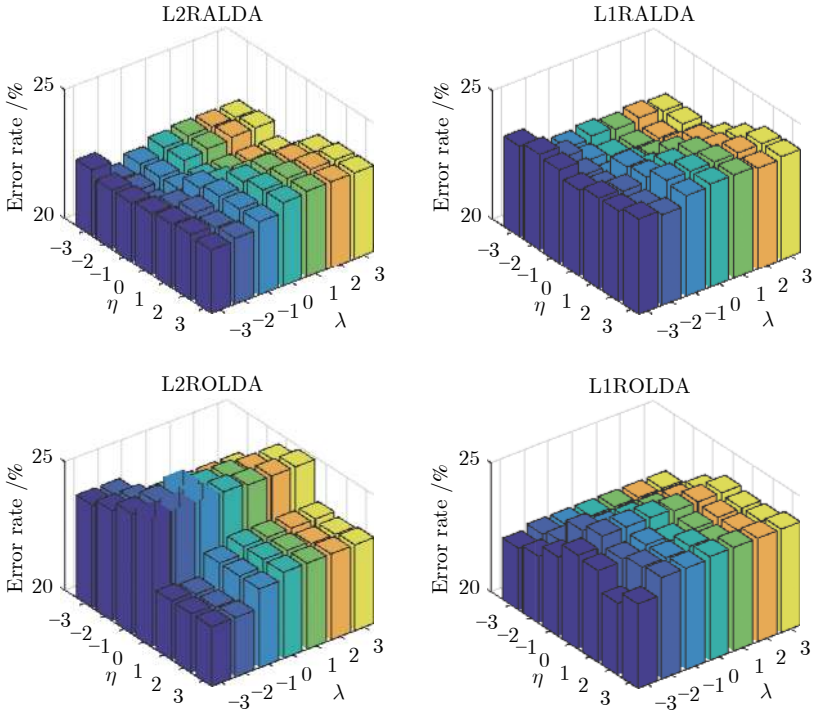
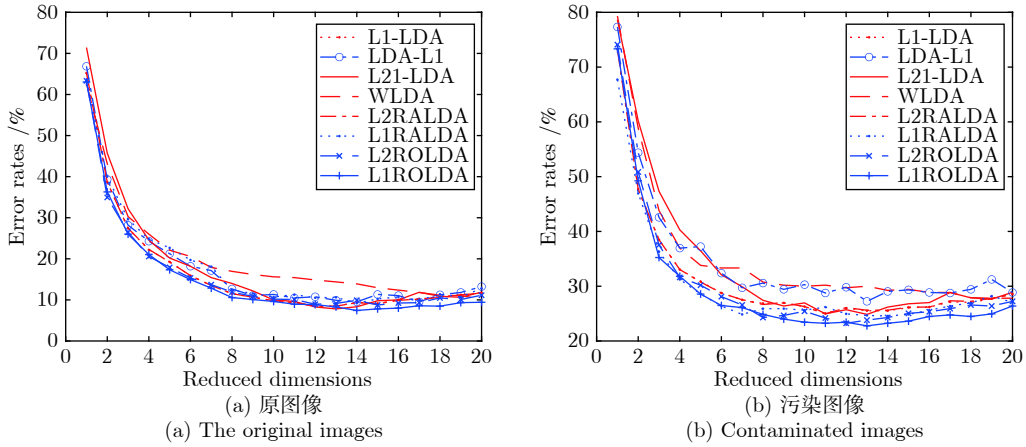


图 2 L2RALDA, L1RALDA, L2ROLDA 和 L1ROLDA 的错误率与参数的关系

Fig. 2 Error rates of L2RALDA, L1RALDA, L2ROLDA and L1ROLDA versus the parameters



(a) 原图像 (a) The original images

(b) 污染图像 (b) Contaminated images

图 3 数据集上不同方法随维数变化的错误率

Fig. 3 Error rates of various methods with varying dimensions on the Yale database

从图 3 可知, 随着约简维数的增加各种方法的错误率下降, 但过多的特征反而使得错误率上升. 这表明特征的维数是一个重要的参数. 从表 1 可看出, 对于原图像集, L1RALDA 在 ORL 和 UMIST 数据集上可获得很好的分类效果, 这表明在适当的条件下, 优先考虑边界点的判别分析算法是合理的. 当部分图像被污染后, L1ROLDA 在大多数情况下优于其他方法. 在实际遮挡的 AR 数据上, 围巾遮挡比太阳镜遮挡导致更大的错误率.

同其他方法比较, L1ROLDA 不仅采用了 L_1 范数的损失函数, 而且考虑了样本的采样概率, 其采样概率在不确定集中自适应地变化, 从而得到较好的效果. 从表 1 可知, 当部分图像被污染后, 各种方法错误率的标准偏差一般大于原始图像上各种方法错误率的标准偏差. 这表明污染的数据使得各种方法的性能具有更大的不确定性.

这组实验比较了各种算法在五个图像数据集上的时间消耗. 提出算法的停止条件如前面所述. 在

表 1　各种方法在原始数据集和污染数据集上的平均错误率 (%) 和标准偏差

Table 1　Average error rates (%) of various methods and their standard deviations on the original and contaminated data sets

Data sets	L1-LDA	LDA-L1	L21-LDA	WLDA	L2RALDA	L1RALDA	L2ROLDA	L1ROLDA
Yale	8.48 (3.42)	9.52 (3.47)	7.81 (4.21)	10.19 (3.24)	8.48 (4.25)	9.19 (3.96)	8.76 (3.84)	**7.46 (3.10)**
C-Yale	24.95 (4.76)	25.05 (5.05)	24.86 (4.98)	27.81 (4.87)	27.24 (4.92)	24.57 (4.58)	23.24 (4.49)	**22.76 (4.12)**
ORL	9.89 (2.13)	0.21 (2.06)	8.86 (2.45)	10.33 (2.02)	8.98 (2.15)	**8.34** (2.12)	9.66 (2.18)	9.19 (1.92)
C-ORL	14.62 (2.41)	15.27 (2.32)	13.98 (2.73)	15.92 (2.85)	15.82 (2.67)	13.13 (2.63)	13.45 (2.49)	**12.58** (2.52)
UMIST	8.99 (2.09)	9.23 (2.07)	8.87 (2.75)	10.15 (2.02)	9.42 (2.15)	**8.83** (2.12)	9.07 (2.18)	8.98 (1.99)
C-UMIST	24.52 (3.89)	26.33 (3.93)	22.98 (3.85)	29.23 (3.84)	23.39 (4.04)	23.52 (3.92)	23.22 (3.88)	**21.90** (3.72)
COIL	18.45 (2.02)	19.46 (1.64)	18.21 (1.65)	19.97 (1.79)	19.05 (1.64)	17.98 (1.46)	18.31 (2.12)	**17.42** (2.14)
C-COIL	28.34 (3.41)	29.66 (3.49)	27.35 (3.55)	29.01 (3.15)	28.46 (3.43)	27.65 (2.43)	28.32 (3.01)	**26.22** (3.32)
AR-sunglasses	9.26 (1.73)	9.38 (1.46)	8.05 (1.57)	10.02 (1.70)	9.21 (1.53)	9.01 (1.25)	8.25 (1.23)	**7.33** (1.79)
AR-scarf	21.29 (1.10)	20.81 (1.25)	19.03 (1.28)	28.02 (0.92)	26.35 (0.89)	20.34 (1.34)	19.38 (1.24)	**17.24** (1.34)

ORL, UMIST, Yale 和 COIL 数据集上假定约简维数为 40. 由于 AR 数据集包含更多的类数, 其约简维数设定为 80. 图 4 显示了各种算法在五个数据集上的运行时间 (单位: 秒).

从图 4 可看出, WLDA 的运行时间远远高于其它算法的运行时间, 这是因为这种方法采用了二阶锥规划. L1-LDA 和 LDA-L1 都采用了贪婪算法取得多个投影向量, 它们的运行时间相差不大. 在 Yale 和 ORL 数据集上, L21LDA 的运行时间是最少的, 但在 UMIST 和 COIL 数据集上, L2ROLDA 需要的时间最少. 由于在 AR 数据集上的约简维数为 80 且类数较多, 每种方法的训练时间明显大于在其他数据集上的训练时间. 一般来说, L1RALDA 和 L1ROLDA 训练时间分别大于 L2RALDA 和 L2ROLDA 的训练时间, 这是因为 L1RALDA 和 L1ROLDA 采用了 L_1 范数导致其目标函数是不可

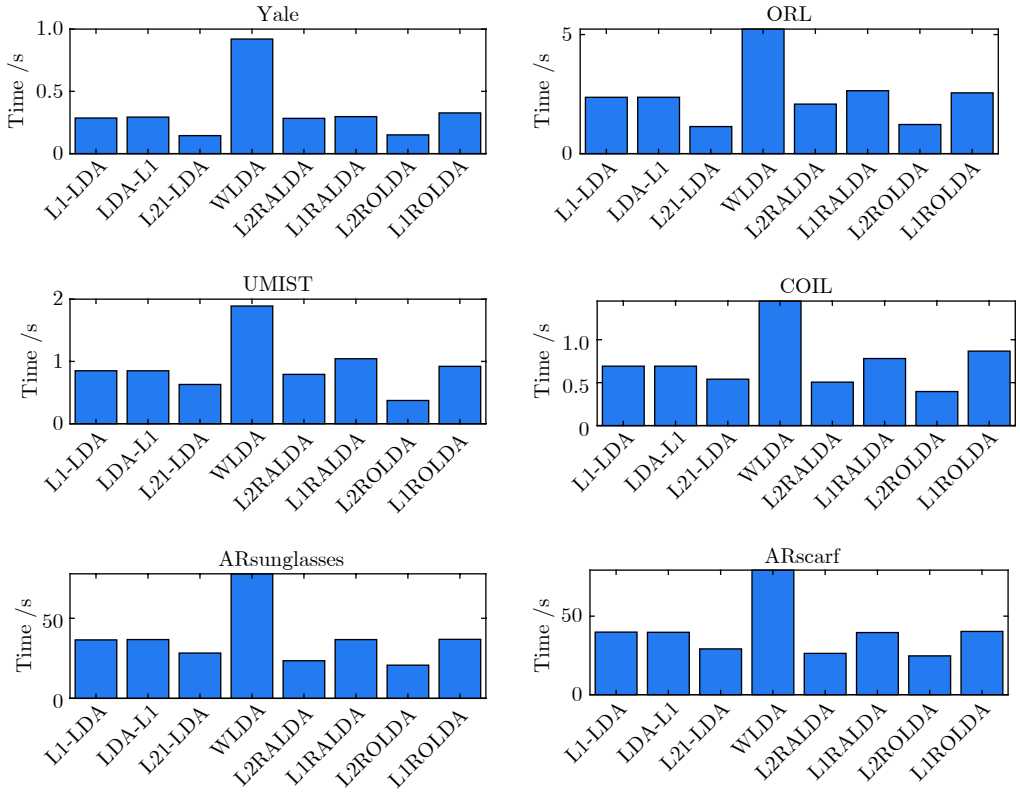


图 4　五个图像数据集上各种算法的运行时间

Fig. 4　Running time of various methods on five image data sets

微的, 而 L2RALDA 和 L2ROLDA 的目标函数是可微函数.

3.2 UCI 数据集上的实验

这组实验使用的数据集来自 UCI 机器学习库. 这些数据集已被广泛用于评估一些学习算法的性能. 本文在 8 个数据集上测试了提出的模型. 这 8 个数据集分别是 Australia (690 样本/14 特征/2 类), Diabetes (768/8/2), German (1 000/24/2), Heart (270/13/2), Liver(345/6/2), Sonar (208/60/2), WPBC (198/33/2) 和 Waveform (5000/21/3). 不同于图像数据集, 对于这些数据集, 训练样本的维数远远小于训练样本的个数, 即小样本的奇异性问题不会出现. 每个数据集样本的属性被转化为 [−1, 1] 区间. 对于每个数据集, 随机选取 70% 的样本构成训练集, 剩下的样本作为测试集. 除了原始数据集, 我们也模拟了训练集中样本的某些特征被污染的情况. 具体来说, 训练集中一半样本的 50% 特征被替换为由−1 和 1 组成的随机噪声.

每种方法的参数在训练集上通过五叠交叉验证方法学习得到. 实验性能取自十次随机实验的平均结果. 为了比较两种算法在同一个数据集上的性能, 在显著性水平 0.05 的情况下计算 L1ROLDA 和其他算法的配对 t 检验的 p-值. 如果 p- 值大于 0.05 时, 这表明 L1ROLDA 和其他算法没有显著的差别. 当 p-值小于 0.05 时, 这表明 L1ROLDA 和其他算法存在显著差别. 表 2 和表 3 列出在原始数据集和污染数据集上的实验结果, 其实验结果包括平均正确率 (Average correct rates, ACR) 和标准偏差 (standard deviations,SD) 以及 L1ROLDA 和其他方法的配对 t 检验的 p-值.

从表 2 可知, L21-LDA 在 Diabetes 数据集上取得最好的性能, L1RALDA 在 German 和 WPBC 数据集上取得最好的性能, L1ROLDA 在 Australian, Heart 和 Waveform 数据集上取得最好的性能, L1-LDA 在 Liver 数据集上取得最好的性能, 这表明了没有一种方法在这些数据集上取得一致好的性能. 从表 3 可知, 当特征被污染后, 每种方法的性能都会存在某种程度的下降. 一般来说基于 L_2 范数方法不比基于 L_1 范数方法好. 从表 2 和表 3 的 p-值可知, L1ROLDA 和其它方法是否存在明显的差别. 例如: 从 p-值可知在特征没有被污染的情况下, L1ROLDA 和 L21LDA 在 German 和 Heart 数据集上存在统计意义上的差别, 但当特征被污染后, L1ROLDA 和 L21LDA 在 Australian, German, Heart, Liver 和 Waveform 数据集上存在统计意义上的差别.

当特征被污染后, L1ROLDA 方法在多数情况

表 2 各种方法在原始数据集上的平均正确率 (ACR(%)), 标准偏差 (SD) 和 p-值

Table 2 Average correct rates (ACR(%)), standard deviations (SD), and p-values of various methods on the original data sets

Data sets	L1-LDA ACR (SD) p-值	LDA-L1 ACR (SD) p-值	L21-LDA ACR (SD) p-值	WLDA ACR (SD) p-值	L2RALDA ACR (SD) p-值	L1RALDA ACR (SD) p-值	L2ROLDA ACR (SD) p-值	L1ROLDA ACR (SD) p- 值
Australian	82.22 (3.44)	80.15 (3.36)	83.44 (3.57)	80.12 (3.29)	79.15 (3.61)	82.99 (3.46)	79.77 (3.27)	**84.12** (3.63)
	7.98×10^{-3}	3.13×10^{-5}	0.43	3.00×10^{-5}	1.45×10^{-5}	0.028	1.78×10^{-5}	—
Diabetes	72.55 (4.51)	71.68 (4.62)	**73.28** (4.33)	70.19 (4.40)	71.18 (4.71)	72.87 (4.26)	71.99 (4.27)	72.68 (4.39)
	0.22	0.15	0.84	0.0084	0.10	0.46	0.19	—
German	74.45 (3.66)	72.02 (3.69)	74.68 (3.88)	69.34 (3.77)	72.06 (3.54)	**74.99** (3.49)	72.67 (3.66)	73.74 (3.48)
	0.03	1.30×10^{-3}	0.04	6.39×10^{-4}	7.84×10^{-3}	0.04	0.01	—
Heart	75.89 (5.11)	74.36 (5.13)	77.32 (5.16)	73.53 (5.17)	74.22 (5.22)	77.52 (5.34)	75.67 (5.54)	**78.98** (5.19)
	8.94×10^{-4}	5.99×10^{-5}	0.012	3.72×10^{-6}	4.81×10^{-5}	0.043	1.14×10^{-4}	—
Liver	**65.25** (4.33)	63.27 (4.78)	64.34 (4.99)	62.87 (4.60)	62.99 (4.71)	64.12 (4.37)	63.01 (4.48)	64.54 (4.29)
	0.52	0.89	0.28	0.074	0.08	0.68	0.51	—
Sonar	72.11 (4.98)	70.99 (4.96)	73.16 (5.52)	70.21 (5.43)	70.68 (5.06)	72.45 (5.21)	70.99 (5.29)	**73.22** (5.16)
	0.06	6.07×10^{-4}	0.72	3.80×10^{-4}	4.45×10^{-4}	0.074	6.69×10^{-4}	—
Waveform	83.27 (1.99)	82.18 (2.12)	85.12 (1.88)	81.23 (1.94)	81.53 (2.15)	83.69 (2.22)	81.49 (2.10)	**86.28** (1.98)
	1.93×10^{-5}	8.97×10^{-6}	0.08	1.42×10^{-6}	1.83×10^{-6}	7.10×10^{-5}	1.57×10^{-6}	—
WPBC	77.89 (5.19)	75.32 (5.23)	78.23 (5.44)	72.12 (5.37)	73.14 (5.21)	**79.33** (5.36)	72.99 (5.28)	77.89 (5.29)
	0.47	1.67×10^{-4}	0.17	1.19×10^{-5}	1.58×10^{-5}	4.76×10^{-3}	6.08×10^{-6}	—

表 3　各种方法在污染数据集上的平均正确率 (ACR(%)), 标准偏差 (SD) 和 p -值

Table 3　Average correct rates (ACR(%)), standard deviations (SD), and p -values of various methods on the contaminated data sets

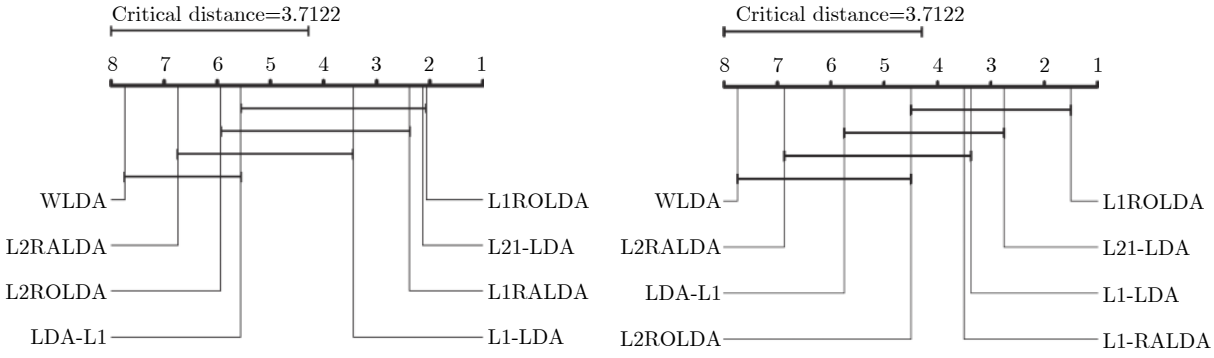
Data sets	L1-LDA	LDA-L1	L21-LDA	WLDA	L2RALDA	L1RALDA	L2ROLDA	L1ROLDA
	ACR (SD)	ACR (SD)	ACR (SD)	ACR (SD)	ACR (SD)	ACR (SD)	ACR (SD)	ACR (SD)
	p -值	p -值	p -值	p - 值	p -值	p -值	p -值	
Australian	80.45 (3.56)	78.34 (3.77)	81.65 (3.46)	75.22 (3.89)	77.26 (3.45)	81.78 (3.66)	79.62 (3.78)	**82.51 (3.52)**
	1.99×10^{-4}	7.89×10^{-6}	0.025	6.02×10^{-7}	5.17×10^{-6}	0.04	2.10×10^{-5}	—
Diabetes	70.63 (4.22)	69.44 (4.29)	70.32 (4.35)	65.26 (4.65)	69.37 (4.60)	70.65 (4.05)	70.38 (4.30)	**70.37 (4.41)**
	0.41	0.055	0.29	7.55×10^{-5}	0.037	0.49	0.39	—
German	71.34 (3.48)	70.08 (3.55)	71.76 (3.22)	64.45 (3.79)	70.05 (3.86)	71.09 (3.94)	71.39 (3.68)	**72.36 (3.77)**
	5.08×10^{-3}	0.92×10^{-3}	1.41×10^{-2}	1.30×10^{-7}	1.80×10^{-3}	0.027	0.099	—
Heart	72.05 (5.26)	72.24 (5.45)	72.44 (5.13)	66.53 (4.98)	70.22 (5.26)	70.35 (5.39)	71.51 (4.99)	**74.88 (5.10)**
	3.06×10^{-3}	1.58×10^{-3}	8.12×10^{-3}	2.03×10^{-6}	0.35×10^{-4}	1.06×10^{-4}	0.67×10^{-3}	—
Liver	62.67 (4.33)	60.67 (4.59)	62.53 (4.25)	59.36 (4.78)	60.08 (4.32)	62.04 (4.64)	61.01 (4.13)	**63.98 (4.31)**
	0.047	9.79×10^{-4}	0.039	1.51×10^{-4}	2.75×10^{-4}	0.032	7.22×10^{-3}	—
Sonar	70.56 (5.71)	68.89 (5.96)	**71.37 (5.34)**	66.19 (5.39)	68.34 (5.30)	70.32 (5.41)	69.82 (5.27)	71.02 (5.19)
	0.17	9.16×10^{-4}	0.55	6.18×10^{-5}	2.34×10^{-4}	0.12	6.99×10^{-2}	—
Waveform	80.46 (1.89)	79.04 (2.03)	81.08 (1.96)	79.28 (1.89)	80.56 (1.95)	81.75 (2.02)	80.73 (2.11)	**82.28 (2.03)**
	4.47×10^{-3}	1.54×10^{-5}	0.03	4.69×10^{-5}	5.27×10^{-3}	0.18	6.33×10^{-3}	—
WPBC	73.44 (5.10)	71.35 (5.15)	73.31 (5.27)	70.25 (5.29)	70.21 (5.33)	72.21 (5.39)	70.82 (5.42)	**74.76 (5.22)**
	0.49	2.30×10^{-2}	0.34	3.17×10^{-3}	2.96×10^{-3}	3.59×10^{-2}	8.34×10^{-3}	—

下优于其他方法, 这是因为在不确定集下, 优先考虑了那些有利区分的样本, 这些样本可能不是污染的数据点, 而对离群点赋予较低的采样概率. 因此当数据被污染后, 应当优先选择 L1ROLDA 方法. 实际上, 当测试 RALDA 和 ROLDA 时, 如果在验证数据集上 ROLDA 的性能远远好于 RALDA 的性能, 那么说明训练集包含离群点. 在这种情况下, 我们可采用一些去离群点的方法去除训练集中的离群点, 然后再执行特征提取算法.

尽管配对 t 检验能比较两种算法在同一个数据集上的性能差别, 但是它不能取得多种算法在多个数据集上的性能排序问题. 由于在多个数据集上测试了多种方法, 本文采用 Friedman 检验[40] 评估多种算法的性能. 在 Friedman 检验中, 如果零假设成立, 即假设所有的算法在性能上是相等的, Friedman 统计量可用概率分布来计算. 本文使用关键差别 (Critical difference, CD) 图[40] 比较不同算法的性能. 图 5 表示了在原始数据和污染数据上的性能分析的 CD 图. CD 图的平均秩提供了算法的性能的排序. 平均秩越小, 算法的性能越好. 从图 5(a) 可知, L1ROLDA 给出了最小的平均秩, 但是 L21-LDA 的平均秩和 L1ROLDA 的平均秩相差不大. WLDA 有最大的平均秩. 从图 5(b) 可知, 当数据被污染后, L1ROLDA 仍然取得了最小的平均秩, 但 L21-LDA 的平均秩明显大于 L1ROLDA 的平均秩. 总的来说, 由于采用了 L_1 范数和优先考虑了类中心附近的样本, 当数据被污染后, 正则化乐观线性判别分析取得最好的性能.

4　结论

本文提出了基于不确定集和混合范数的线性判别分析方法. 不同于以前的方法, 本文借助 Kullback-Leibler 不确定集描述样本的变化信息, 这样可探索不确定集中的概率分布. 由于样本或类中心的采样概率在不确定集中灵活变化, 从而使得模型适应数据的内在特性. 模型是比率优化和非凸优化问题. 广义 Dinkelbach 算法被用来求解优化模型. 本文利用投影梯度法或交替优化技术求解算法的优化子问题. 这样算法的收敛性直接来自广义 Dinkelbach 算法的收敛性. 在图像数据集和 UCI 数据集上做了一系列实验. 实验结果表明: 在没有污染的数据集上, 由于 RALDA 考虑了类边界附近的样本, RALDA 应当被优先考虑, 但在污染数据集上, ROLDA 应当被优先考虑. 由于本文简化了模型中一些参数, 这些参数对模型的性能会产生一定的影响. 今后我们将重点研究如何自动学习多个参数, 并将本文的基本思想推广到其它基于核函数的非线性判别分析和张量判别分析.



(a) 在原始数据上的 CD 图
(a) CD diagram on the original data

(b) 在污染数据上的 CD 图
(b) CD diagram on the contaminated data

图 5 不同方法性能的显著性分析

Fig. 5 Performance significance analysis of various methods

References

1 Kan M N, Shan S G, Zhang H H, Lao S H, Chen X L. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(1): 188−194

2 Kwak N. Principal component analysis based on L1-norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(9): 1672−1680

3 Kwak N. Principal component analysis by L_p-norm maximization. *IEEE Transactions on Cybernetics*, 2014, **44**(5): 594−609

4 Gao Yun-Long, Luo Si-Zhe, Pan Jin-Yan, Chen Bai-Hua, Zhang Yi-Song. Robust PCA using adaptive probability weighting. *Acta Automatica Sinica*, 2021, **47**(4): 825−838
(高云龙, 罗斯哲, 潘金艳, 陈柏华, 张逸松. 鲁棒自适应概率加权主成分分析. 自动化学报, 2021, **47**(4): 825−838)

5 He Jin-Rong, Bi Ying-Zhou, Ding Li-Xin, Liu Bin. Local variation regularized margin discriminant projection. *Chinese Journal of Computers*, 2018, **41**(4): 780−795
(何进荣, 闭应洲, 丁立新, 刘斌. 局部差异正则化的边界判别投影. 计算机学报, 2018, **41**(4): 780−795)

6 Zheng W M, Lin Z C, Wang H X. L1-norm kernel discriminant analysis via Bayes error bound optimization for robust feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**(4): 793−805

7 Zhong K, Han M, Qiu T, Han B. Fault diagnosis of complex processes using sparse kernel local fisher discriminant analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(5): 1581−1591

8 Iosifidis A, Tefas A, Pitas I. On the optimal class representation in linear discriminant analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, **24**(9): 1491−1497

9 Zheng S, Ding C, Nie F P, Huang H. Harmonic mean linear discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2019, **31**(8): 1520−1531

10 Nie F P, Wang Z, Wang R, Li X L. Submanifold-preserving discriminant analysis with an auto-optimized graph. *IEEE Transactions on Cybernetics*, 2020, **50**(8): 3682−3695

11 Lotlikar R, Kothari R. Fractional-step dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(6): 623−627

12 Loog M, Duin R P W, Haeb-Umbach R. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, **23**(7): 762−766

13 Tao D C, Li X L, Wu X D, Maybank S J. Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(2): 260−274

14 Bian W, Tao D C. Harmonic mean for subspace selection. In: Proceedings of the 19th International Conference on Pattern Recognition. Tampa, USA: IEEE, 2008. 1−4

15 Li Z H, Nie F P, Chang X J, Yang Y. Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2017, **29**(10): 2100−2110

16 Zhang Y, Yeung D Y. Worst-case linear discriminant analysis. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2010. 2568−2576

17 Bian W, Tao D C. Max-min distance analysis by using sequential SDP relaxation for dimension reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(5): 1037−1050

18 Flamary R, Cuturi M, Courty N, Rakotomamonjy A. Wasserstein discriminant analysis. *Machine Learning*, 2018, **107**(12): 1923−1945

19 Zhao X W, Guo J, Nie F P, Chen L, Li Z H, Zhang H X. Joint principal component and discriminant analysis for dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(2): 433−444

20 Cao M, Chen C, Hu X Y, Peng S L. Towards fast and kernelized orthogonal discriminant analysis on person re-identification. *Pattern Recognition*, 2019, **94**: 218−229

21 Shawe-Taylor J, Cristianini N. *Kernel methods for Pattern Analysis*. New York: Cambridge University Press, 2004.

22 Tao D P, Guo Y N, Li Y T, Gao X B. Tensor rank preserving discriminant analysis for facial recognition. *IEEE Transactions on Image Processing*, 2018, **27**(1): 325−334

23 Liu J C, Lian Z H, Wang Y, Xiao J G. Incremental kernel null space discriminant analysis for novelty detection. In: Proceedings of the 2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 4123−4131

24 Pang S N, Ozawa S, Kasabov N. Incremental linear discriminant analysis for classification of data streams. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2005, **35**(5): 905−914

25 Zhao H F, Wang Z, Nie F P. A new formulation of linear discriminant analysis for robust dimensionality reduction. *IEEE Transactions on Knowledge and Data Engineering*, 2019, **31**(4): 629−640

26 Zheng W M, Lu C, Lin Z C, Zhang T, Cui Z, Yang W K. ℓ_1-norm heteroscedastic discriminant analysis under mixture of Gaussian distributions. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(10): 2898−2915

27 Ye Q L, Yang J, Liu F, Zhao C X, Ye N, Yin T M. L1-norm

distance linear discriminant analysis based on an effective iterative algorithm. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, **28**(1): 114−129

28 Zhong F J, Zhang J S. Linear discriminant analysis based on L1-norm maximization. *IEEE Transactions on Image Processing*, 2013, **22**(8): 3018−3027

29 Wang H X, Lu X S, Hu Z L, Zheng W M. Fisher discriminant analysis with L1-norm. *IEEE Transactions on Cybernetics*, 2014, **44**(6): 828−842

30 Li X L, Pang Y W, Yuan Y. L1-norm-based 2DPCA. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2010, **40**(4): 1170−1175

31 Liu Y, Gao Q X, Miao S, Gao X B, Nie F P, Li Y S. A non-greedy algorithm for L1-norm LDA. *IEEE Transactions on Image Processing*, 2017, **26**(2): 684−695

32 Li C N, Shang M Q, Shao Y H, Xu Y, Liu L M, Wang Z. Sparse L1-norm two dimensional linear discriminant analysis via the generalized elastic net regularization. *Neurocomputing*, 2019, **337**: 80−96

33 Li C N, Shao Y H, Wang Z, Deng N Y, Yang Z M. Robust Bhattacharyya bound linear discriminant analysis through an adaptive algorithm. *Knowledge-Based Systems*, 2019, **183**: Article No. 104858

34 Nie F P, Wang Z, Wang R, Wang Z, Li X L. Towards robust discriminative projections learning via non-greedy $\ell_{2,1}\ell 2$,1-norm MinMax. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(6): 2086−2100

35 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

36 Liang Z Z, Chen X W, Zhang L, Liu J, Zhou Y. Correlation classifiers based on data perturbation: New formulations and algorithms. *Pattern Recognition*, 2020, **100**: Article No. 107106

37 Ben-Tal A, den Hertog D, De Waegenaere A, Melenberg B, Rennen G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 2012, **59**(2): 341−357

38 Bonnans J F, Shapiro A. Optimization problems with perturbations: A guided tour. *SIAM Review*, 1998, **40**(2): 228−264

39 Ródenas R G, López M L, Verastegui D. Extensions of Dinkelbach's algorithm for solving non-linear fractional programming problem. *Top*, 1999, **7**(1): 33−70

40 Demšar J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 2006, **7**: 1−30

41 Absil P A, Mahony R, Sepulchre R. *Optimization Algorithms on Matrix Manifolds*. Princeton: Princeton University Press, 2008.

梁志贞 中国矿业大学副教授. 2005 年获得上海交通大学模式识别与智能系统专业博士学位. 主要研究方向为模式识别, 生物特征识别. 本文通信作者.
E-mail: liang@cumt.edu.cn
(**LIANG Zhi-Zhen** Associate professor at China University of Mining and Technology. He received his Ph.D. degree in pattern recognition and intelligence system from Shanghai Jiaotong University in 2005. His research interest covers pattern recognition and biometric recognition. Corresponding author of this paper.)

张 磊 中国矿业大学副教授. 主要研究方向为最优化方法和数据挖掘.
E-mail: zhanglei@cumt.edu.cn
(**ZHANG Lei** Associate professor at China University of Mining and Technology. His research interest covers optimization methods and data mining.)