LPA-SKFST 半监督特征提取方法 *

彭 杰[1]，龚晓峰[1†]，李 剑[2]

(1. 四川大学 电气工程学院, 成都 610065; 2. 浙江农林大学 信息工程学院, 杭州 311300)

摘 要：针对传统 LDA 类半监督特征提取方法的解矢量非正交、解空间不稳定和非线性处理能力不足等问题，提出 LPA-SKFST 方法。该方法的前置级 LPA 通过标签传播提高标记样本容量，后置级 SKFST(半监督核最佳鉴别矢量集)采用双向正则方法对 KFST 引入全局结构保持正则和 Tikhonov 正则，并以成对空间求解方法求取 Fisher 分母矩阵奇异和非奇异时的统一形式解。在 Circle、Iris、Wine 和自有珍珠光谱集的分类实验中，PCA、LDA、SLDA 和 SDG 等组的准确率随样本集、标记样本占比和标签可靠性变化而波动，LPA-SKFST 组则稳定保持在 85%以上。该结果证明，LPA-SKFST 能克服标记样本占比和标记可靠性不足局限，在实际集和线性不可分人工集上取得一致、稳定的优秀表现。

关键词：KPCA；KFST；LDA；双向正则；全局结构保持正则；成对空间求解方法

中图分类号：TP181 doi: 10.19734/j.issn.1001-3695.2020.09.0244

Semi-supervised feature extraction method based on LPA-SKFST

Peng Jie[1], Gong Xiaofeng[1†], Li Jian[2]

(1. *College of Electrical Engineering*, *Sichuan University*, *Chengdu* 610065, *China*; 2. *School of Information Engineering*, *Zhejiang A&F University*, *Hangzhou* 311300, *China*)

Abstract: In order to solve the problems of traditional LDA semi supervised feature extraction methods, such as the solution vector is not orthogonal, the solution space is unstable and there is no linear processing ability, LPA-SKFST is proposed. In this method, the LPA part increases the proportion of labeled samples through label propagation; SKFST (semi supervised kernel optimal discriminant vectors) combines KFST, Tikhonov regularization and Global Preserving regularization by two-way regularization method, and adopts the pair space solution method to ensure the uniform solution form when Fisher's denominator matrix is singular or nonsingular. In the classification experiments of Circle, Iris, Wine and Pearl spectral, the accuracy of PCA, LDA, SLDA and SDG groups fluctuated with the change of sample set, labeled sample proportion and label reliability, while LPA-SKFST group kept stable above 85%. The results show that PA-SKFST can overcome the limitations of low proportion of labeled samples and unreliable labeling, and its performance is stable and excellent in both actual set and linear indivisible artificial set.

Key words: KPCA; KFST; LDA; two-way regularization; global structure regularization; solving method of pairwise space

0 引言

特征提取是机器学习的重要分支，解决高维数据分析问题的重要工具。高维特征数据既显著增加计算资源和存储资源的消耗，又明显加剧小样本问题和过拟合风险，给机器学习任务带来了严峻考验。研究表明，运用特征提取方法，从高维特征空间提取能表征和区分样本的低维表示，可有效提高后续学习任务(过程)的效率和性能[1~8]。

根据是否利用类别信息，特征提取方法通常分为无监督方法和监督方法。典型的，无监督特征提取方法有因子分析(FA)[1]、主成分分析(PCA)[2]、等距特征映射(Isomap)[3]、局部线性嵌入(LLE)[4]等；监督特征提取方法有线性判别分析(LDA)[5]、最佳鉴别平面(ODP)[6]和最佳鉴别矢量集(FSODA 或 FST) [7,8]等。其中，前者主要考虑样本的整体特征，后者则期望最大化同类样本的相似性和异类样本的差异性。工程实践显示，由于无监督方法忽略了不同类别样本间的差异，在分类任务中的表现往往逊色于监督方法。但监督方法要求所有训练样本都具有类别标签，又使样本获取难度和成本增加，可用样本数量减少，过拟合风险加大[9~13]。

为缓解监督方法和无监督方法的矛盾，学界提出了半监督特征提取方法。一种典型做法是在监督方法中加入无监督正则项，降低过拟合风险，以提升算法性能。如 Qiao[9]对 LDA 引入全局结构保持正则，提出全局保持判别分析(GPDA)方法。进一步的，纪明君[10]采用双向正则方法，提出监督与无监督比重可调的 SLDA。Qiao[11]对 LDA 引入拉普拉斯正则和 Tikhonov 正则，并作稀疏表述，提出稀疏保持判别分析(SPDA)。Zhao[12]对 LDA 同时引入全局保持正则和局部保持正则，提出全局约束半监督判别分析(SDG)。孙圣姿[13]则提出全局算法和局部算法相结合的半监督多视角特征提取方法(SS-AMVE)。其中，文献[9,10]和文献[11]方法基于不同理论基础，分别以全局特征和局部特征提升 LDA 性能，文献[12,13]方法则将两者结合。以上方法各有优势，但均基于 LDA 原型，具有特征向量非正交、解空间不稳定和输出维度受样本类别数限制等缺陷。此外，文献[9,10]为线性方法，文献[11~13]的局部正则项只关注近邻特征，均难以有效应对非线性可分数据的分类任务。

为克服上述不足，本文提出 LPA-SKFST 半监督特征提取方法。该方法以核最佳鉴别矢量集 KFST[14]为基础，采用

收稿日期：2020-09-30；**修回日期：**2020-11-19 **基金项目：**浙江省公益技术研究计划资助项目(LGG18F030006)

作者简介：彭杰(1995-)，男，四川成都人，硕士研究生，主要研究方向为检测技术与自动化装置、模式识别、计算机算法；龚晓峰(1965-)，男(通信作者)，四川成都人，教授，硕导，博士，主要研究方向为检测技术与自动化装置(50906650@qq.com)；李剑(1981-)，男，浙江杭州人，副教授，硕导，博士，主要研究方向为检测技术与自动化装置、计算机算法.

双向正则方法结合全局结构保持正则项，既保留文献[10]中监督与无监督比重可调优势，又克服输出特征非正交和输出维度受样本类别数限制等缺陷。进一步的，采用成对空间求解方法，给出 Fisher 分母矩阵奇异与非奇异的统一解形式，保证求解过程统一、稳定。此外，受文献[15]启发，前置级联 LPA，引入局部结构信息，提高标记样本容量，以增强算法稳定性。在 Circle、Iris 和 Wine 的分类任务中，LPA-SKFST 表现出比 PCA、LDA、SLDA 和 SDG 等更稳定、一致的优秀性能。进一步的，在基于光谱模式识别的珍珠光泽等级预测任务中展现了其实用价值。

1　LPA-SKFST 特征提取方法

1.1　LPA 半监督学习

给定未标记样本集 $\boldsymbol{X}_u=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_u\}\in\Re^{u\times d}$ 和标记样本集 $(\boldsymbol{X}_l,\boldsymbol{Y}_l)=\{(\boldsymbol{x}_{u+1},\boldsymbol{y}_{u+1}),\ldots,(\boldsymbol{x}_{u+l},\boldsymbol{y}_{u+l})\}\in\Re^{l\times(d+1)}$，通常 $l\ll u$。有总体样本集 $\boldsymbol{X}=\boldsymbol{X}_l+\boldsymbol{X}_u$，样本总数 $n=l+u$，特征维度 d。设样本类别集 $\boldsymbol{C}=\{c_1,\ldots,c_m\}$，类别总数 m，则样本 $\boldsymbol{x}_i\in\boldsymbol{X}_l$ 的类别标签 $y_i\in\boldsymbol{C}$。

以给定样本集 $\boldsymbol{X}$ 的全体样本为节点，绘制全连接图。令 $w_{ij}=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2/2\sigma^2)$ 为连接样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的边权，其中 $\sigma\succ 0$ 为高斯核参数；$\|\|_2$ 为向量 2 范数，有 $\|\boldsymbol{x}_i\|_2^2=\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$。

设置类别预测矩阵 $\boldsymbol{L}=[\boldsymbol{L}_1^{\mathrm{T}},\ldots,\boldsymbol{L}_n^{\mathrm{T}}]^{\mathrm{T}}\in\Re^{n\times m}$。其中，行向量 $\boldsymbol{L}_i=[L_{i1},\ldots,L_{im}]$ 为样本 $\boldsymbol{x}_i$ 的类别预测向量，$0\le L_{ij}\in\boldsymbol{L}_i\le 1$ 对应样本 $\boldsymbol{x}_i$ 属于 c_j 类的后验概率。初始 $\boldsymbol{L}=\boldsymbol{L}(0)$，当 $\boldsymbol{x}_i\in\boldsymbol{X}_l$ 且 $y_i=c_j$，令 $L_{ij}=1$；否则，令 $L_{ij}=0$。

进一步的，构造标记传播矩阵[16] $\boldsymbol{P}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$。有 $\boldsymbol{D}=\mathrm{diag}(d_1,\ldots,d_{l+u})$，$d_i=\sum_{j=1}^{l+u}w_{ij}$，$\boldsymbol{W}=[w_{ij}]_{n\times n}$。标记传播过程中，逐次迭代类别预测矩阵 $\boldsymbol{L}$，

$$\boldsymbol{L}(t)=\begin{bmatrix}\boldsymbol{0}_{ll} & \boldsymbol{0}_{lu}\\ \boldsymbol{0}_{ul} & \boldsymbol{E}_{uu}\end{bmatrix}\boldsymbol{P}\boldsymbol{L}(t-1)+\boldsymbol{L}(0). \tag{1}$$

$t\in N_+$ 为迭代次数，$\boldsymbol{E}_{uu}\in\Re^{u\times u}$ 为单位对角阵。

文献[16]已证明式(1)必收敛，令 $\boldsymbol{L}^*=\lim_{t\to\infty}\boldsymbol{L}(t)$。此时，取样本 $\boldsymbol{x}_i$ 的类别标签

$$y_i=c_j=\arg\max_{1\le j\le m}(L_{ij}^*), \tag{2}$$

使之与后验概率最大的类别相对应，得到完善的总体样本集 $(\boldsymbol{X},\boldsymbol{Y})=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_{l+u},y_{l+u})\}$。

1.2　SKFST 半监督特征提取方法

1.2.1 显式 SKFST 模型

基于完善总体样本集 $\boldsymbol{X}$，采用双向正则方法，提出融合全局结构保持正则项 $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{t}}^{\Phi}\boldsymbol{\omega}$ 和 Tikhonov 正则项 $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{E}^{\Phi}\boldsymbol{\omega}$ 的显式核 Fisher 判别式

$$J(\boldsymbol{\omega})=\frac{\boldsymbol{\omega}^{\mathrm{T}}\varphi\boldsymbol{S}_{\mathrm{b}}^{\Phi}\boldsymbol{\omega}+\boldsymbol{\omega}^{\mathrm{T}}(1-\varphi)\boldsymbol{S}_{\mathrm{t}}^{\Phi}\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}}\varphi\boldsymbol{S}_{\mathrm{w}}^{\Phi}\boldsymbol{\omega}+\boldsymbol{\omega}^{\mathrm{T}}(1-\varphi)\boldsymbol{E}^{\Phi}\boldsymbol{\omega}}. \tag{3}$$

$\varphi\in[0,1]$ 为双向权值，决定监督项 $\boldsymbol{S}_{\mathrm{w}}^{\Phi}$，$\boldsymbol{S}_{\mathrm{b}}^{\Phi}$ 和无监督项 $\boldsymbol{S}_{\mathrm{t}}^{\Phi}$ 在判别式 $J(\boldsymbol{\omega})$ 中的占比。

其中，$\Phi:\boldsymbol{x}\in\Re^d\to\Phi(\boldsymbol{x})\in H$ 表示输入特征空间到高维希尔伯特空间(Hilbert)的非线性映射。$\boldsymbol{S}_{\mathrm{w}}^{\Phi}$，$\boldsymbol{S}_{\mathrm{b}}^{\Phi}$ 和 $\boldsymbol{S}_{\mathrm{t}}^{\Phi}$ 对应样本集 $\boldsymbol{X}$ 在高维 Hilbert 空间的类内、类间和总体散布矩阵，

$$\begin{cases}\boldsymbol{S}_{\mathrm{w}}^{\Phi}=\sum_j^m\sum_{y_i=c_j}(\Phi(\boldsymbol{x}_i)-\boldsymbol{\mu}_j^{\Phi})^{\mathrm{T}}(\Phi(\boldsymbol{x}_i)-\boldsymbol{\mu}_j^{\Phi})\\ \boldsymbol{S}_{\mathrm{b}}^{\Phi}=\sum_j^m n_j(\boldsymbol{\mu}_j^{\Phi}-\boldsymbol{\mu}^{\Phi})^{\mathrm{T}}(\boldsymbol{\mu}_j^{\Phi}-\boldsymbol{\mu}^{\Phi})\\ \boldsymbol{S}_{\mathrm{t}}^{\Phi}=\sum_i^n(\Phi(\boldsymbol{x}_i)-\boldsymbol{\mu}^{\Phi})^{\mathrm{T}}(\Phi(\boldsymbol{x}_i)-\boldsymbol{\mu}^{\Phi})\end{cases}. \tag{4}$$

n_j 是第 j 类样本总数，$\boldsymbol{\mu}_j^{\Phi}=(1/n_j)\sum_{y_i=c_j}\Phi(\boldsymbol{x}_i)$ 是第 j 类样本均值，$\boldsymbol{\mu}^{\Phi}=(1/n)\sum_{\boldsymbol{x}_i\in\boldsymbol{X}}\Phi(\boldsymbol{x}_i)$ 是总体样本均值。$\boldsymbol{E}^{\boldsymbol{\Phi}}\in H$ 是 Hilbert 空间的单位对角阵。

SKFST 特征提取方法旨在寻找正交矢量集 $[\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_h]\in H$，最大化半监督 Fisher 判别式(3)，$h\le n$ 为矢量集的最大可取维度。具体的，取

$$\boldsymbol{\omega}_1=\arg\max_{\boldsymbol{\omega}}(J(\boldsymbol{\omega})),\ \text{s.t. }\|\boldsymbol{\omega}_1\|=1; \tag{5}$$

取第 $i\in\{2,\ldots,h\}$ 维特征矢量

$$\begin{cases}\boldsymbol{\omega}_i=\arg\max_{\boldsymbol{\omega}}(J(\boldsymbol{\omega}))\\ \text{s.t. }\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{\omega}_j=0,\ j=1,2,\ldots,i-1.\\ \text{s.t. }\|\boldsymbol{\omega}_i\|=1\end{cases} \tag{6}$$

1.2.2 隐式 SKFST 模型

显式模型中的 Hilbert 空间通常是未知的，无法直接求解式(5)(6)。常用方法[17,18]是将 Hilbert 空间中的解矢量 $\boldsymbol{\omega}\in H$ 表示为向量集 $\{\Phi(\boldsymbol{x}_1),\ldots,\Phi(\boldsymbol{x}_n)\}$ 的线性组合

$$\boldsymbol{\omega}=\sum_{i=1}^n\zeta_i\Phi(\boldsymbol{x}_i)=\boldsymbol{\Psi}\boldsymbol{\alpha}. \tag{7}$$

其中，向量集矩阵 $\boldsymbol{\Psi}=[\Phi(\boldsymbol{x}_1),\ldots,\Phi(\boldsymbol{x}_n)]$，系数向量 $\boldsymbol{\alpha}=[\zeta_1,\ldots,\zeta_n]^{\mathrm{T}}$。改写显式判别式(3)为等价的隐式判别式

$$J(\boldsymbol{\alpha})=\frac{\boldsymbol{\alpha}^{\mathrm{T}}\varphi\boldsymbol{K}_{\mathrm{b}}\boldsymbol{\alpha}+\boldsymbol{\alpha}^{\mathrm{T}}(1-\varphi)\boldsymbol{K}_{\mathrm{t}}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\varphi\boldsymbol{K}_{\mathrm{w}}\boldsymbol{\alpha}+\boldsymbol{\alpha}^{\mathrm{T}}(1-\varphi)\boldsymbol{K}_{\mathrm{E}}\boldsymbol{\alpha}}. \tag{8}$$

进而，式(5)(6)基于式(3)求解 $\{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_h\}\in H$，可等价替换为基于式(8)求解 $\{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_h\}\in\Re^n$。

隐式判别式(8)中，$\boldsymbol{K}_{\mathrm{b}}$，$\boldsymbol{K}_{\mathrm{w}}$，$\boldsymbol{K}_{\mathrm{t}}$ 和 $\boldsymbol{K}_{\mathrm{E}}$ 分别为 $\boldsymbol{S}_{\mathrm{b}}^{\Phi}$，$\boldsymbol{S}_{\mathrm{w}}^{\Phi}$，$\boldsymbol{S}_{\mathrm{t}}^{\Phi}$ 和 $\boldsymbol{E}^{\Phi}$ 的替换。根据文献[17]，引入核函数 $\mathrm{k}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j)^{\mathrm{T}}$ 和核矩阵 $\boldsymbol{K}=[\mathrm{k}(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{n\times n}$，可得计算式

$$\begin{cases}\boldsymbol{K}_{\mathrm{b}}=\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{b}}^{\Phi}\boldsymbol{\Psi}=\sum_{j=1}^m n_j(\boldsymbol{\eta}_j-\boldsymbol{\eta})^{\mathrm{T}}(\boldsymbol{\eta}_j-\boldsymbol{\eta})\\ \boldsymbol{K}_{\mathrm{w}}=\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}}^{\Phi}\boldsymbol{\Psi}=\boldsymbol{K}\boldsymbol{K}^{\mathrm{T}}-\sum_{j=1}^m n_j\boldsymbol{\eta}_j^{\mathrm{T}}\boldsymbol{\eta}_j\\ \boldsymbol{K}_{\mathrm{t}}=\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{t}}^{\Phi}\boldsymbol{\Psi}=\boldsymbol{K}\boldsymbol{K}\\ \boldsymbol{K}_{\mathrm{E}}=\boldsymbol{K}\end{cases}. \tag{9}$$

其中，$\boldsymbol{\eta}=(1/n)\boldsymbol{Q}_{\mathrm{A}}\boldsymbol{K}^{\mathrm{T}}$，$\boldsymbol{\eta}_j=(1/n_j)\boldsymbol{Q}_j\boldsymbol{K}^{\mathrm{T}}$。行向量 $\boldsymbol{Q}_j\in\Re^n$ 是第 j 类样本的指示向量，当 $\boldsymbol{x}_i\in\boldsymbol{X}$ 且 $y_i=c_j$ 时，$\boldsymbol{Q}_j$ 的第 i 个元素 $q_{ji}=1$，否则 $q_{ji}=0$。特别的，$\forall q_i\in\boldsymbol{Q}_{\mathrm{A}}$ 有 $q_i=1$。

1.2.3 成对空间求解方法

解空间不稳定是 Fisher 类方法的共同困扰，学界已提出多种求解方法[19~22]。在此基础上，本文针对性地给出 SKFST 分母矩阵 $[\varphi\boldsymbol{K}_{\mathrm{w}}+(1-\varphi)\boldsymbol{I}_{\mathrm{k}}]$ 奇异和非奇异的统一解形式。

令分母矩阵 $\boldsymbol{M}=\varphi\boldsymbol{K}_{\mathrm{w}}+(1-\varphi)\boldsymbol{K}_{\mathrm{E}}$，分子矩阵 $\boldsymbol{N}=\varphi\boldsymbol{K}_{\mathrm{b}}+(1-\varphi)\boldsymbol{K}_{\mathrm{t}}$，改写式(8)为常规 Fisher 准则形式 $J(\boldsymbol{\alpha})=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\alpha}/\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\alpha}$。对分母矩阵 $\boldsymbol{M}$ 做特征分解 $\boldsymbol{M}\boldsymbol{v}=\lambda\boldsymbol{v}$，有非零特征向量集 $\boldsymbol{U}=\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_r\}$ 和零特征向量集 $\boldsymbol{V}=\{\boldsymbol{v}_{r+1},\ldots,\boldsymbol{v}_{r+z}\}$，且 $\forall\boldsymbol{v}\in\{\boldsymbol{U},\boldsymbol{V}\}$ 满足 $\boldsymbol{v}_i^{\boldsymbol{T}}\boldsymbol{v}_j=0$，$\boldsymbol{v}_i^{\boldsymbol{T}}\boldsymbol{v}_i=1$。称生成空间 $\overline{\boldsymbol{M}(0)}=\mathrm{span}(\boldsymbol{U})$ 为非零特征空间，$\boldsymbol{M}$ 在该空间非奇异，有 $|\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{M}\boldsymbol{U}|\ne 0$；称生成空间 $\boldsymbol{M}(0)=\mathrm{span}(\boldsymbol{V})$ 为零特征空间，$\boldsymbol{M}$ 在该空间必奇异，有 $|\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{M}\boldsymbol{V}|=0$。

文献[19,22]指出，通常意义的最佳鉴别矢量 α 存在于交空间 $\boldsymbol{M}(0)\cap\overline{\boldsymbol{N}(0)}=\mathrm{span}(\boldsymbol{R})$，张成向量集 $\boldsymbol{R}=\{\boldsymbol{v}|\boldsymbol{v}\in\boldsymbol{V},\boldsymbol{v}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{v}\ne 0\}$。将线性组合

$$\boldsymbol{\alpha}=\sum_{\boldsymbol{v}_i\in\boldsymbol{R}}\zeta_i\boldsymbol{v}_i=\boldsymbol{R}\boldsymbol{\theta} \tag{10}$$

带入判别式(8)，并改写判别式为 $J(\boldsymbol{\theta})=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{R}\boldsymbol{\theta}$。可求解矢量 $\boldsymbol{\alpha}\in\mathrm{span}(\boldsymbol{R})$ 的系数向量 $\boldsymbol{\theta}=[\zeta_1,\ldots,\zeta_s]^{\mathrm{T}}$，$s=\mathrm{rank}(\boldsymbol{R})$，再由 $[\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_s]=\boldsymbol{R}[\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_s]$ 取得解矢量集。但上述解易发生特征简并过度风险[14]，且 $\mathrm{span}(\boldsymbol{R})$ 并不总是存在，在成对的 $\overline{\boldsymbol{M}(0)}$ 空间增补求取矢量集 $[\boldsymbol{\alpha}_{s+1},\ldots,\boldsymbol{\alpha}_{s+r}]\in\mathrm{span}(\boldsymbol{U})$ 则可有效规避该问题。类似的，将线性组合

$$\boldsymbol{\alpha}=\sum_{\boldsymbol{v}_i\in\boldsymbol{U}}\zeta_i\boldsymbol{v}_i=\boldsymbol{U}\boldsymbol{\theta}. \tag{11}$$

带入判别式(8)，并改写判别式为

$$J(\boldsymbol{\theta})=\frac{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{U}\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{U}\boldsymbol{\theta}}, \tag{12}$$

便可解得增补矢量 $\boldsymbol{\alpha}_i\in[\boldsymbol{\alpha}_{s+1},\ldots,\boldsymbol{\alpha}_{s+r}]$ 的系数向量 $\boldsymbol{\theta}_i=[\zeta_1,\ldots,\zeta_r]^{\mathrm{T}}$。

1.3 LPA-SKFST 算法实现步骤

算法 1　LPA-SKFAT 算法实现步骤

输入：已标记样本集 $(\boldsymbol{X}_l,\boldsymbol{Y}_l)$，未标记样本 X_u，类别集 C，高斯核参数 σ，核函数 $k(x_i,x_j)$，双向权值 φ；

输出：矢量集 $[\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_{s+1},\ldots,\boldsymbol{\alpha}_{s+r}]$，特征值集 $[\lambda_1,\ldots,\lambda_s,\lambda_{s+1},\ldots,\lambda_{s+r}]$；

a) 由 σ 和 $\forall x_i,x_j\in X$ 计算边权 w_{ij}，生成矩阵 W；

b) 基于边权矩阵 W 构造标记传播矩阵 P；

c) 初始化类别预测矩阵 L；

d) 基于迭代式(1)得到收敛 L^*；

e) 基于 L^* 和式(2)为 $\forall X_i\in X_u$ 分配伪标签 y_i；

f) 由 $\forall x_i,x_j\in X$ 和 $k(x_i,x_j)$ 计算核矩阵 K，再根据式(9)计算 K_b，K_w，K_t 和 K_E；

g) if $0\le\varphi\le1$

h)　特征分解矩阵 $\varphi K_b+(1-\varphi)K_t$，取得非零特征值向量集 U 和零特征值向量集 V；

i)　抽取子集 $R=\{\upsilon\,|\,\upsilon\in V,\upsilon^{\mathrm{T}}N\upsilon\ne0\}$；

j) end if

k) if $rank(R)\ne0$

l)　特征分解矩阵 $R^{\mathrm{T}}NR$，取得特征值 $[\lambda_1,\cdots,\lambda_s]$ 和特征向量集 $[\upsilon_1,\cdots,\upsilon_s]$；(按特征值降序)

m)　求解 $[\alpha_1,\cdots,\alpha_s]=R[\theta_1,\cdots,\theta_s]$；

n) end if

o) 初始化增补矢量的系数向量集 $[\boldsymbol{\theta}_{s+1},\ldots,\boldsymbol{\theta}_{s+r}]=[\boldsymbol{0}]_{r\times r}$；

p) 令 $[\boldsymbol{\theta}_{s+1},\ldots,\boldsymbol{\theta}_{s+r}]$ 为矩阵 $\boldsymbol{H}^{\perp}$，即 $\boldsymbol{H}^{\perp}=[\boldsymbol{\theta}_{s+1},\ldots,\boldsymbol{\theta}_{s+r}]$；

q) for $i=1:\mathrm{rank}(\boldsymbol{U})$

u)　令矩阵 $\boldsymbol{H}$ 为 $\boldsymbol{H}^{\perp}$ 的正交补，即 $\boldsymbol{H}$ 为方程组 $\boldsymbol{H}^{\perp}\boldsymbol{\chi}=\boldsymbol{0}$ 的单位正交解集，其中 $\boldsymbol{\chi}\in\Re^{r\times1}$ 为未知数向量，$\boldsymbol{0}\in\Re^{r\times1}$ 为零向量；

v)　特征分解 $(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{U}\boldsymbol{H})^{-1}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{U}\boldsymbol{H})$，取得最大特征值 $\lambda_{\max}$ 和对应特征向量 $\varepsilon_{\max(\lambda)}$；

w)　令 $\boldsymbol{\alpha}_{s+i}=\boldsymbol{U}\boldsymbol{H}\boldsymbol{\varepsilon}_{\max(\lambda)}$，$\lambda_{s+i}=\lambda_{\max}$；

x)　令 $\boldsymbol{\theta}_{s+i}=\boldsymbol{H}\boldsymbol{\varepsilon}_{\max(\lambda)}$，更新 $\boldsymbol{H}^{\perp}$；

y) end for

2　实验与分析

2.1　实验数据

为验证 LPA-SKFST 特征提取方法的实际效果，本文分别在人工集 Circle、UCI 公开集 Iris 和 Wine 的分类任务中，同 PCA、LDA、SLDA[10]和 SDG[12]等进行了实验对比。各数据集所含类别及样本分布如表 1，Circle 含 2 个样本类别，Iris 和 Wine 各含 3 个样本类别。其中，Circle 有 300 例样本和 2 个特征维度，成同心环状分布，如图 1。Circle 的非线性特征突出，异类样本间分界明显、界线规整，常用于验证算法的非线性处理能力。鸢尾花数据集 Iris 有 150 例样本和 4 个特征维度，葡萄酒数据集 Wine 有 178 例样本和 13 个特征维度，获取自 http://archive.ics.uci.edu。Iris 和 Wine 是机器学习算法检验的经典实测数据集，用于检验算法在现实数据集上的实际表现。

表 1　各数据集的构成情况

Tab. 1　Composition of each data set

Data set	Class_1	Class_2	Class_3
Circle	150	150	—
Iris	50	50	50
Wine	59	71	48

2.2　实验设计

不失一般性的，分别将 Circle、Iris 和 Wine 中样本，按 4:1 随机划分为训练集(Train)和测试集(Test)，且各样本类别均尽可能满足该划分比例。进一步的，对各训练集，随机删除 $(1-Per)$% 样本的类别标签，模拟不同的未标记样本占比情况。本实验依次取 $Per=\{0.1,0.2,\ldots,1\}$。

对各训练集，分别以 3.1 中方法训练特征提取模型，提取对应训练集和测试集的低维特征。其间，PCA 与标签无关，无论 *Per* 取值，均使用全体训练样本。LDA 、SLDA 和 SDG 均只使用有标记样本，且以 PCA 作预降维，使预降维度 $d\le n-m$，并限制输出特征为 $m-1$ 维[10,12]。SKFST 使用 LPA 处理所得训练集，公平起见，也取 $m-1$ 维输出特征。

特别的，SKFST 的 $k(x_i,x_j)$ 采用高斯核函数，其核参数 σ_2 同双向权值 φ 和 LPA 的核参数 σ 一起通过网格法确定。设置 φ 的搜索序列为 $\{0,0.1,\ldots,1\}$，σ 和 σ_2 的搜索范围为 $[2^{-10},2^{10}]$，取被检验指标最佳的最小 $(\sigma,\sigma_2,\varphi)$。类似的，SLDA 的双向权值 φ_{SLDA} 和 SDG 的正则权值 λ_{SDG}、γ_{SDG} 与 η_{SDG} 通过相同方法搜索取得。其中，根据文献[12]设置 $\lambda_{SDG},\gamma_{SDG}$ 和 η_{SDG} 的搜索序列为 $\{10^{-6},10^{-4},10^{-2},10^{0},10^{2},10^{4},10^{6}\}$。

本实验中，被检验指标即测试集样本的分类准确率，其同一于分类任务的最终目标。实验采用线性支持向量积分类方法，通过台湾大学林智仁团队开发的 LIBSVM 工具包实现。

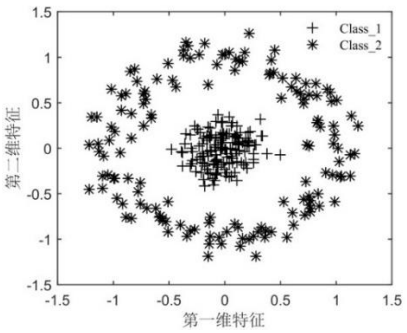


图 1　Circle 数据集

Fig. 1　Circle dataset

2.3　实验结果与分析

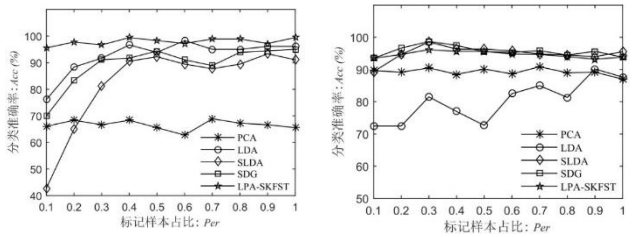
2.3.1 实验结果分析

根据实验设计，在各数据集、各 *Per* 设定下，分别对各实验组做 5 折交叉验证，统计平均准确率如图 2。为确保实验结果可靠，每折交叉验证均做重复实验，取最佳准确率。

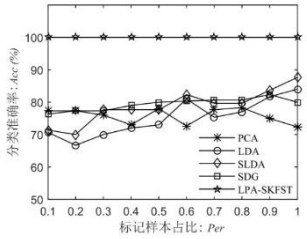
各实验组在 Circle 集的表现如图 2(A)。其中，PCA 组的准确率在 72.33%至 78.33%区间波动。LDA、SLDA 和 SDG 组的准确率随 *Per* 值上升而缓慢上升。$Per=0.2$ 时，LDA 和 SLDA 组分别取得最小准确率 66.67%和 70%；$Per=1$ 时，分别取得最大准确率 84%和 87.67%。SDG 组准确率的最大值和最小值分别为 76.33%和 82.33%，在 $Per=0.1$ 和 $Per=0.9$ 时取得。LPA-SKFST 组则在各 *Per* 设定下均取得 100%的分类准确率，明显优于其余各组。这主要得益于 LPA-SKFST 优秀的非线性处理能力，在此基础上，Circle 规整、明显的异类样本间界线是 LPA-SKFST 组取得 100%分类准确率的另一重要原因。

各实验组在 Iris 和 Wine 集的表现如图 2(b)(c)。在 Iris 集，PCA 组准确率在 87%至 90.83%区间波动，SLDA 和 SDG 组准确率保持在 89.33%至 98.5%和 93.5%至 98.67%区间，LDA 组准确率随 *Per* 上升而由 72.5%波动上升至 90.17%。特别的，实验中 LDA 组频繁发生过拟合现象，即使增加重复次数，也未取得相对平滑的准确率变化曲线。Wine 集中，PCA 组准确率波动区间为 66.11%至 68.89%。LDA、SDG 和 SLDA 组准确率在 $Per\in[0.1,0.4]$ 时，分别由 76.11%、70%和 42.78%上升至 96.67%、91.67%和 90.56%，在 $Per\ge0.4$ 时趋于稳定。以上结果显示，当 PCA(无监督方法)表现优秀时，SLDA 和 SDG 确能有效改善 LDA 过拟合现象，提升算法稳定性和分类准确率指标，如图 2(b)。反之，PCA 表现明显逊色于 LDA 时，SLDA 和 SDG 的实验表现也并不比 LDA 优秀，如图 2(c)。

LPA-SKFST 组在 Iris 和 Wine 集的准确率分别保持在 93.5%至 96.17%和 95.56%至 99.44%区间。纵向而言，该准确率较 Circle 集有所下降。原因在于现实数据集 Iris 和 Wine 维度更高、样本分布和类间界线更复杂，数据处理难度更大。SDG 等组的准确率有不同程度的提高，则表明 Iris 和 Wine 的线性不可分程度较 Circle 弱，这便削弱了非线性方法的相对优势。横向上，在 Iris，SDG、SLDA 与 LPA-SKFST 的表现趋于一致，并在 $Per=0.4$ 等个别条件下具有微弱优势。然而，Wine 实验显示 SDG 和 SLDA 易因数据集和 *Per* 变化而产生较大波动。与之不同，LPA-SKFST 在各数据集和 *Per* 设定的优秀表现是一致稳定的。一方面，这得益于 LPA 提高标记样本容量，使监督项能够获得充足的类别信息。另一方面，双向正则方法使 SKFST 的理论表现应不弱于 KPCA 和 KFST，进一步增强了算法稳定性。



(a)wine　　(b)iris



(c)circle

图 2　准确率统计

Fig. 2　Statistical chart of accuracy

2.3.2 LPA 增益分析

为验证 LPA 的实际作用，本文新增 SKFST 实验组，除只使用标记样本训练模型外，其余设定同 LPA-SKFST 组一致。统计两组准确率的均值(Avg)、最小值(Min)和最大值(Max)如表 2。

结果显示，LPA 确有助于提高准确率指标和实验表现的稳定性。具体的，LPA-SKFST 组在 Circle、Iris 和 Wine 的平均准确率分别为 100%、97.94%和 94.61%，一致高于 SKFST 组的 99.73%、97%和 94.08，即拥有更高平均准确率。此外，两组在各集的最大准确率相近，LPA-SKFST 组的最小准确率却平均高出 SKFST 组 2%，即实验表现更稳定。

表 2　LPA-SKFST 和 SKFS 组准确率

Tab. 2　Accuracy of LPA-SKFST and SKFST groups　　%

Statistical	LPA-SKFST			SKFST		
items	Circle	Wine	Iris	Circle	Wine	Iris
Avg	100	97.94	94.61	99.73	97	94.08
Min	100	95.56	93.50	97.33	92.78	92.83
Max	100	99.44	95.67	100	99.44	95.17

3　LPA-SKFST 应用实例

3.1.1 案例描述

进一步的，将 LPA-SKFST 作为珍珠光谱模式识别任务的特征提取方法，以验证其实用性。该任务以测量自珍珠表面随机区域的可见光光谱为基础，旨在通过模式识别技术，训练珍珠光泽等级预测模型。光谱数据维度大，冗余度、耦合度高，极大地考验着特征提取方法的实际性能。

此外，前期研究[23]发现，珍珠表面光泽“预期均匀，实际非均匀”，使部分光谱样本表征的光泽等级同所赋标签(采样珍珠的整体性光泽等级)不一致，降低了带监督特征提取方法和预测模型的准确性。为提高样本可靠性，常需大量地重复采样和筛选光谱，操作繁琐，样本利用率低。应用 LPA-SKFST 方法则可有效缓解上述矛盾。

3.1.2 过程设计

具体的，自光谱集 $\boldsymbol{X}$ 每个样本类别中各抽取 n 组标记可靠的样本，组成“标记样本集” $\boldsymbol{X}_l$，其余样本组成“未标记样本集” $\boldsymbol{X}_u$。利用前置级 LPA 为 $\boldsymbol{x}_i \in \boldsymbol{X}_u$ 分配伪标签 y_i，得到完善光谱样本集 $(\boldsymbol{X},\boldsymbol{Y})$。在此基础上，以后置级 SKFST 提取低维特征向量，并训练 SVM 分类预测模型。其中，SKFST 和 SVM 均采用高斯核函数，核参数取得方法与 2.2 节一致。特别的，结合前期研究经验，取低维特征向量为 8 维。

上述过程，通过 LPA 的标记传播功能，基于后验概率最大准则，为所有标记非可靠样本重新分配了类别标签。该操作既提高了样本标记可靠性，又避免了繁琐的重复采样和光谱筛选，简化了工作任务。进一步的，由于避免了大量地筛除光谱，便也提高了样本利用率。

3.1.3 数据介绍

训练集有 600 例光谱样本，分属高光泽、中光泽和低光泽三个样本类别。其中，每个类别均含 30 组标记可靠的光谱样本，用做 LPA 的“标记样本” $\boldsymbol{x}_i \in \boldsymbol{X}_l$。其余 510 例样本中，300 例样本未赋标签，210 例样本的标签可靠性未知。另取 100 例标记可靠的光谱样本组成测试集，用以检验模型效果。上述样本均为实测珍珠可见光光谱集，随机展示其中 10 例如图 3。

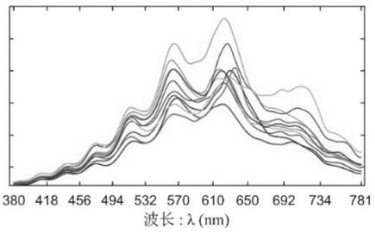


图 3　10 例光谱样本(随机抽取自样本集)

Fig. 3　Ten spectral (randomly obtained from the sample set)

3.1.4 结果分析

根据前述设计，统计测试集光谱样本的分类准确率如图 4。LPA-SKFST 组(L-S)的准确率为 86%，略低于 O-S 组的 89%，高于 PCA、LDA、SDG、SLDA 和 SKFST 组的 72%、59%、63%、65%和 74%。其中，PCA 等组的实验原则与 2.2 节一致，LDA、SLDA 和 SDG 组取 2 维低维特征向量，PCA 和 SKFST 组取 8 维低维特征向量。特别的，O-S 组的训练集由“重复采样+光谱筛选”方法[23]取得，并以 SKFST 提取低维特征向量。

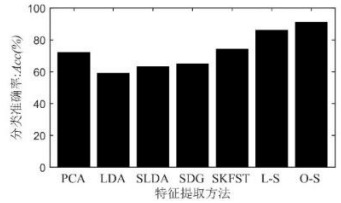


图 4　光谱分类准确率统计图

Fig. 4　Statistical chart of spectral classification accuracy

结果显示，LDA、SLDA 和 SDG 组的准确率均未超过 65%，较 PCA 组低 7%至 13%。该现象是由于 LDA 等方法存在输出维度限制，对维度高、类别少的数据集，具有特征约简过度问题。SKFST 则克服了该缺陷，其准确率也较 PCA 组提高了 2%。遗憾的是，上述各组均未能改善样本可靠性，使

准确率仍落后于 L-S 和 O-S 组 10%以上。

实验中，L-S 组的准确率较 O-S 组低 3%，表明经 LPA 取得的样本可靠性较文献[23]仍有差距。L-S 组和 SKFST 组的结果对比则显示，前置级 LPA 的准确率增益达 12%，展现了 LPA-SKFST 方法应对标签不可靠样本集的优越性。进一步的，L-S 组的表现虽略低于 O-S 组，但有效简化了工作任务，提高了样本利用率，仍具有突出的实用价值。

4　结束语

本文基于 LPA、KFST、Tikhonov 正则和全局保持正则，提出 LPA-SKFST 特征提取方法，并在分类任务中检验了其实际表现。结果显示，该方法能有效克服标记样本占比和标记可靠性不足局限，在实际集和线性不可分人工集上取得一致、稳定的优秀表现。分析发现，LPA 提高标记样本容量，增加可用类别信息是取得上述表现的原因之一；双向正则方法主动调节监督和无监督比重，也保证其理论表现不弱于 KPCA 和 KFST。此外，该方法输出特征空间正交，输出特征维度不受样本类别数限制，求解过程统一、稳定。但该方法对核参数 σ 取值敏感，不易匹配，故增强其鲁棒性是笔者后续研究的主要工作之一。

参考文献：

[1] Argelaguet R, Velten B, Arnol D, *et al.* Multi-Omics factor analysis-a framework for unsupervised integration of multi-omic data sets [J]. Molecular Systems Biology, 2018, 14 (6): e8124. (2018-06-20) [2020-06-01]. https://pubmed. ncbi. nlm. nih. gov/29925568/

[2] Ledoit O, Wolf M. Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions [J]. Journal of Multivariate Analysis, 2015, 139: 360-384.

[3] Kashniyal J, Verma S, Singh K P. Wireless sensor networks localization using progressive Isomap [J]. Wireless Personal Communications, 2017, 92: 1281-1302.

[4] 张潞瑶, 季伟东, 程昊. 基于 LLE 降维思想的自然计算方法 [J]. 系统仿真学报, 2020, 32 (10): 1943-1955. (Zhang Luyao, Ji Weidong, Cheng Hao. Natural computing method base on LLE dimension reduction [J]. Journal of System Simulation, 2020, 32 (10): 1943-1955.)

[5] Dellacasa Bellingegni A, Gruppioni E, Colazzo G, *et al.* NLR, MLP, SVM, and LDA: a comparative analysis on EMG data from people with trans-radial amputation [J]. Journal of Neuroengineering & Rehabilitation, 2017, 14 (1): 82-108.

[6] Cao Suqun, Wang Shitong, Zhu Quanyin, *et al.* Unsupervised optimal discriminant plane based feature extraction method [C]// Fifth International Conference on Fuzzy Systems and Knowledge Discovery. 2008 (2): 315-319.

[7] Huang Xinyu, Xu Jiaolong, Guo Gang. Incremental kernel null Foley-Sammon transform for person re-identification [C]/ Proc of the 24th International Conference on Pattern Recognition (ICPR) . (2018-08-01) [2020-06-01] https://doi. org/10. 1109/ICPR. 2018. 8546301

[8] Shi Yanjiao, Yi Yugen, Zhang Qing, *et al.* Kernel null-space-based abnormal event detection using hybrid motion information [J]. Journal of Electronic Imaging, 2019, 28 (2): 021011. 1-021011. 12.

[9] Qiao Lishan, Zhang Limei, Chen Songcan. An empirical study of two typical locality preserving linear discriminant analysis methods [J]. Neurocomputing, 2010, 73 (10-12): 1587-1594.

[10] 纪明君, 刘漫丹, 才乐千. 基于半监督 LDA 特征子空间优化的人脸识别算法 [J]. 计算机工程与科学, 2018, 40 (10): 1851-1857. (Ji Mingjun, Liu Mandan, Cai Leqian. A face recognition algorithm base on semi-supervised LDA feature subspace optimization [J]. Computer Engineering & Science, 2018, 40 (10): 1851-1857.)

[11] Qiao Lishan, Chen Songcan, Tan Xiaoyang. Sparsity preserving discriminant analysis for single training image face recognition [J]. Pattern Recognition Letters, 2010, 31 (5): 422-429.

[12] Zhao Xin, Li Xue, Pang Chaoyi, *et al.* Human action recognition based on semi-supervised discriminant analysis with global constraint [J]. Neurocomputing, 2013, 105 (apr. 1): 45-50.

[13] 孙圣姿, 万源, 曾成. 自适应嵌入的半监督多视角特征降维方法 [J]. 计算机应用, 2018, 38 (12): 3391-3398. (Sun Shengzi, Wan Yuan, Zeng Cheng. Semi-supervised adaptive multi-view embedding method for feature dimension [J]. Journal of Computer Applications, 2018, 38 (12): 3391-3398.)

[14] Zheng Wenming, Zhao Li, Zou Cairong. Foley-Sammon optimal discriminant vectors using kernel approach [J]. IEEE Trans on Neural Networks, 2005, 16 (1): 1-9.

[15] Nie Feiping, Xiang Shiming, Jia Yangqing, *et al.* Semi-supervised orthogonal discriminant analysis via label propagation [J]. Pattern Recognition, 2009, 42 (11): 2615-2627.

[16] Zhou Dengyong, Bousquet O, Lal T N, *et al.* Learning with local and global consistency [C]// Advances in Neural Information Processing Systems. 2004, 16 (3): 321-328.

[17] Scholkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation, 1998, 10 (5): 1299-1319.

[18] Shang Liangliang, Yan Ze, Qiu Aibing, *et al.* Efficient recursive kernel principal component analysis for nonlinear time-varying processes monitoring [C]/ 2019 Chinese Control And Decision Conference (CCDC) . (2019-06-03) [2020-06-01]. https://doi. org/10. 1109/CCDC. 2019. 8832617.

[19] 黄新宇, 郭纲, 许娇龙, 等. 基于增量零空间 Foley-Sammon 变换的行人重识别 [J]. 光电子•激光, 2018, 29 (04): 405-410. (Huang Xinyu, Guo Gang, Xu Jiaolong, *et al.* Incremental null Foley-Sammon transform for pedestrian re-identification [J]. Journal of Optoelectronics • Laser, 2018, 29 (04): 405-410.)

[20] Liu Shenglan, Feng Lin, Qiao Hong. Scatter balance: an angle-based supervised dimensionality reduction [J]. IEEE Trans on Neural Networks and Learning Systems, 2015, 26 (2): 277-289.

[21] Zhang Di, Li Xueqiang, He Jiazhong, *et al.* A new linear discriminant analysis algorithm based on L1-norm maximization and locality preserving projection [J]. Pattern Analysis & Applications, 2018, 21 (3): 685-701.

[22] 于雪莲, 刘本永. 最优的核判别分析用于雷达目标识别 [J]. 电子科技大学学报, 2008, 37 (06): 883-885, 937. (Yu Xuelian, Liu Benyong. Optimal kernel discriminant analysis for radar target recognition [J]. Journal of University of Electronic Science and Technology of China, 2008, 37 (06): 883-885, 937.)

[23] 彭杰, 龚晓峰, 雒瑞森, 等. 基于可见光光谱的珍珠光泽检测方法: 中国, CN201910842953. 5 [P]. 2020. 04. 07. (Peng Jie, Gong Xiaofeng, Luo Ruisen et. Pearl luster detection method based on visible light spectrum: China, CN201910842953. 5 [P]. 2020. 04. 07.)