文章编号: 1001-0920 (2015) 02-0203-12 DOI: 10.13195/j.kzyjc.2014.0017

自适应局部图嵌入加权罚支持向量机

廖 剑[1,2], 周绍磊[1], 史贤俊[1]

(1. 海军航空工程学院 控制工程系，山东 烟台 264001；2. 中国人民解放军 91550 部队，辽宁 大连 116000)

摘 要: 针对标准SVM不能有效利用数据流形的局部信息以及对数据中的野值敏感的两点不足, 提出一种基于自适应局部图嵌入加权罚SVM. 算法在保持SVM优化框架不变的情况下, 在目标函数中同时加入了对数据整体类间间隔最大化和数据局部流形分布的要求, 优化了分类决策边界, 简化了核化过程, 同时在软间隔的样本惩罚系数中引入了数据的全局结构信息, 增强了算法的鲁棒性. 在人工、标准和图像数据集上的实验结果表明, 所提出的方法是有效的.

关键词: 支持向量机；流形学习；局部结构信息；局部判别信息；全局结构

中图分类号: TP391 **文献标志码:** A

Adaptive local graph embedding weighted-penalty SVM

LIAO Jian[1,2], *ZHOU Shao-lei*[1], *SHI Xian-jun*[1]

(1. Department of Control Engineering，Naval Aeronautical and Astronautical University，Yantai 264001，China；2. Unit 91550 of PLA，Dalian 116000，China．Correspondent：LIAO Jian，E-mail：251250544@qq.com)

Abstract: As a popular machine learning algorithm, the standard support vector machine(SVM) is faced with two problems: 1) how to effectively use the local information of data manifold; 2) the classification hyperplane sensitive to the outliers in the data. Therefore, a learning algorithm called adaptive local graph embedding weighted-penalty support vector machine(ALGEWP-SVM) is proposed. On the condition of keeping the optimization framework of the standard SVM, the proposed algorithm joins the requirements of maximizing inter-class margin of the entire data and optimizing local distribution of the data manifold in the objective function, which optimizes the hyperplane of classification decision and simplifies the process of kernelization. Meanwhile, the proposed algorithm introduces the global structure information of data to automatically repress the influence of the outliers upon the hyperplane and improve the robustness of the algorithm. The results of the experiment on artificial, standard and image datasets show the effectiveness of the proposed algorithm.

Keywords: support vector machine; manifold learning; local structure information; local discriminative information; global structure

0 引 言

模式分类是指用一定数量的训练样本进行分类器的设计, 然后用该分类器对待识别样本进行分类识别, 使其对未来所有可能样本的预期性能最优[1]. 目前, 已有众多模式识别方法被相继提出, 其中基于统计的机器学习方法得到了广泛研究, 而其中又以支持向量机[2](SVM)及其相关变体更甚[3-7]. SVM通过引入核函数将低维空间中的线性不可分问题映射到高维空间, 然后在该空间中构造最优分类超平面使两类样本以最大间隔分开, 达到强泛化能力[2,8]. 但是, SVM在构造分类超平面时仅仅关注了整体的类间信息, 而忽略了类内样本的先验分布信息, 在一定程度上制约了SVM类方法泛化能力的进一步提高[8-9]. 虽然已有众多学者提出各种SVM方法的变体及参数优选方法[7,10-14], 试图提高SVM类方法的泛化性能, 但其只对某些特定数据集有效.

Lanczos[15]指出, 任何数学技巧都不能补救信息的缺失. 为此, 近年来已有众多学者致力于研究如何将更多先验数据信息融入SVM的优化函数, 并设计出了一系列SVM的改进算法. 最小类方差SVM[14,16](MCVSVM)就是一种融入了全局数据结构和全局鉴别信息的典型SVM改进算法, 它借助于线性判别分

收稿日期: 2014-01-05；**修回日期:** 2014-03-24.

基金项目: 国家青年科学基金项目(61203168).

作者简介: 廖剑(1985−), 男, 博士, 从事模式识别与数据挖掘等研究；周绍磊(1963−), 男, 教授, 博士生导师, 从事模式识别与机器学习等研究.

析 (LDA)[17]的思想, 通过类内方差来正则化 SVM. 上述基于数据全局信息的方法, 在一定程度上弥补了分类器在利用数据分布特征信息上的不足, 但其仍然未能揭示数据的潜在本质几何结构, 尤其是数据的局部结构信息和局部判别信息.

Cover 等[18]曾强调指出, 样本的一半类信息都隐藏在它的邻域里, 而近年来所提出的流形学习方法[1,19-22]能够有效揭示样本点内部所蕴含的局部结构信息和局部判别信息, 因此, 如何将流形学习方法引入 SVM 已迅速成为了研究的焦点. 其中, 最典型的是基于拉普拉斯特征映射 (LE) 的拉普拉斯支持向量机 (LapSVM)[23] , 它首先通过对每个类构建相应的带权无向图, 然后在目标函数中引入相应的流形正则化项来完成对输入空间的内在几何结构确定, 但是流形正则化项的数目等于构造的图个数, 从而造成 LapSVM 需要优化的正则化参数过多, 导致其计算复杂度很高, 性能不稳定. 为了克服 LapSVM 的缺点, 同时引入数据的局部流形信息, Wang 等[24]将 SVM 与局部保持投影 (LPP)[25]相结合, 提出一种最小类局部保持方差 SVM 方法 (MCLPVSVM). 该方法在学习过程中充分考虑了数据的类内局部流形, 但 LPP 本质上属于无监督学习算法, 不能用以揭示数据集的判别信息. 文献 [26] 通过分析指出, 现有局部保持类方法在保持模式之间的局部结构信息时, 忽略了模式之间的判别信息, 同时, 该类方法通过局部惩罚因子来最小化局部离散度, 使得邻域内的样本投影后比较接近, 当邻域内的样本过于接近时, 容易造成邻域内样本之间的判别信息丢失, 因此在一定程度上影响了 MCLPVSVM 方法的模式分类性能[1,24]. 针对上述方法存在的问题, 文献 [1] 通过在 ν-SVM[10]的基础上同时考虑样本空间的局部结构信息和局部差异 (判别) 信息, 提出了一种局部保持 (留) 最大信息差 ν-SVM (ν-LPMIVSVM). 但是, 该方法在引入局部信息时却忽略了数据的全局结构信息, 并没有在目标函数中加入对数据整体类间间隔最大化的要求, 影响了模式识别的性能, 也散失了标准 SVM 解的稀疏性; 同时, ν-LPMIVSVM 方法在构造局部类内图和类间图时忽略了数据集样本的分布特点, 这样容易造成近邻点选择不合适时影响分类器的最优投影方向[27]; 在对错分样本的惩罚上也并没有考虑野值对分类边界的影响, 因此, 在一定程度上可以认为 ν-LPMIVSVM 方法并没有充分考虑数据样本的流形局部信息, 同时也忽略了数据的全局结构信息.

针对 ν-LPMIVSVM 方法存在的问题, 本文提出一种新颖的自适应局部图嵌入加权罚支持向量机 (ALGEWP-SVM). 对于模式分类问题, ALGEWP-SVM 在考虑对数据整体类间间隔最大化的同时引入数据流形的局部信息, 通过满足最小局部结构信息和最大局部判别信息准则来进一步优化模式分割超平面; 同时, 在惩罚系数中引入数据的全局结构信息自动抑制野值对分类边界的影响, 增强算法的鲁棒性. 本文方法的创新之处在于:

1) 全面融入了数据的全局和局部信息, 继承了标准 SVM 和 ν-LPMIVSVM 方法的特色, 在一定程度上不但可以保持数据内在的局部几何结构和判别信息, 还在全局上体现了标准 SVM 的整体类间间隔最大化原则, 避免了学习的不稳定和不充分问题, 同时可以得到比 ν-LPMIVSVM 更稀疏的解;

2) 根据样本分布性状及样本间的相似度自适应计算样本类内和类间近邻点个数, 构造一种对数据分布敏感的局部类内图和类间图拉普拉斯矩阵, 以更准确地刻画样本流形的局部几何结构和局部判别信息, 确保分类性能稳定;

3) 在惩罚系数中引入数据的全局结构信息自动抑制野值对分类超平面的影响, 增强算法的鲁棒性.

1 ALGEWP-SVM

1.1 基本概念

为简单起见, 考虑一个二类分类问题. 对于多类分类问题, 可采用 1vs.1 或 1vs.rest 方法将其转化为多个二分类问题加以解决[2]. 对于一个包含 N 个样本的二分类问题, 设给定训练集

$$T=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\}\in(\boldsymbol{X},\boldsymbol{Y})^l.$$

其中: $\boldsymbol{x}_i\in\boldsymbol{X}=R^n$ 为 n 维输入样本, $y_i\in\boldsymbol{Y}=\{-1,1\}$ 为类标签, $i=1,2,\cdots,l$.

设 $N_k(\boldsymbol{x}_i)$ 为样本点 $\boldsymbol{x}_i$ 的 k 近邻集, G 代表数据集 $\boldsymbol{X}$ 的加权邻接图, 其中第 i 个顶点代表样本点 $\boldsymbol{x}_i$, 称为节点. 如果 $\boldsymbol{x}_i\in N_k(\boldsymbol{x}_j)$ 或 $\boldsymbol{x}_j\in N_k(\boldsymbol{x}_i)$, 则 G 中的顶点相连, 并可定义其连接权重值为 W_{ij}. 其中权值的定义比较常用的方法是采用如下的热核函数[28]:

$$W_{ij}=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/t), \tag{1}$$

其中 $t>0$ 是热核参数. 于是定义图 G 的相应权重矩阵 $\boldsymbol{A}$ 为

$$\boldsymbol{A}_{ij}=\begin{cases}W_{ij},\ \boldsymbol{x}_i\in N_k(\boldsymbol{x}_j)\ \text{or}\ \boldsymbol{x}_i\in N_k(\boldsymbol{x}_j);\\ 0,\text{else}.\end{cases} \tag{2}$$

根据光谱图理论[29], 用一个最近邻图 G 对输入空间的内在几何结构进行建模能有效刻画样本流形的局部结构, 但仅有一个整体图并不足以反映样本间的判别结构, 因此, 可以将加权邻接图 G 进一步分割为两个互补的加权邻接图: 类内图 G^w 和类间图 G^b, 且满足 $G=G^w\bigcup G^b, G^w\bigcap G^b=\varnothing$, 分别用于反映局

部同类邻接关系和局部异类邻接关系, 并分别定义为

$$\boldsymbol{A}_{ij}^{w}=\begin{cases}W_{ij},\ \boldsymbol{x}_i\in N_k^w(\boldsymbol{x}_j)\ \text{or}\ \boldsymbol{x}_i\in N_k^w(\boldsymbol{x}_j);\\ 0,\ \text{else.}\end{cases}\tag{3}$$

$$\boldsymbol{A}_{ij}^{b}=\begin{cases}W_{ij},\ \boldsymbol{x}_i\in N_k^b(\boldsymbol{x}_j)\ \text{or}\ \boldsymbol{x}_i\in N_k^b(\boldsymbol{x}_j);\\ 0,\ \text{else.}\end{cases}\tag{4}$$

其中: $N_k^w(\boldsymbol{x}_i)\subset G^w$ 为样本点 $\boldsymbol{x}_i$ 的 k 同类近邻数据集, $N_k^b(\boldsymbol{x}_i)\subset G^b$ 为样本点 $\boldsymbol{x}_i$ 的 k 异类近邻数据集.

为了更好地描述 ALGEWP-SVM 问题, 现给出如下定义.

定义 1 (局部离散度矩阵)[1]　设 $\boldsymbol{L}^w$ 和 $\boldsymbol{L}^b$ 分别为图 G^w 和 G^b 的拉普拉斯矩阵, 则矩阵

$$\boldsymbol{H}^w=\boldsymbol{X}\boldsymbol{L}^w\boldsymbol{X}^{\mathrm{T}}=\boldsymbol{X}(\boldsymbol{T}^w-\boldsymbol{A}^w)\boldsymbol{X}^{\mathrm{T}}$$

称为局部类内图离散度矩阵, 矩阵

$$\boldsymbol{H}^b=\boldsymbol{X}\boldsymbol{L}^b\boldsymbol{X}^{\mathrm{T}}=\boldsymbol{X}\left(\boldsymbol{T}^b-\boldsymbol{A}^b\right)\boldsymbol{X}^{\mathrm{T}}$$

称为局部类间图离散度矩阵. 其中: $\boldsymbol{A}_{ij}^{(\cdot)}$ 为图 $G^{(\cdot)}$ 的权重矩阵; $\boldsymbol{T}^{(\cdot)}$ 为一对角矩阵, 其对角线上元素定义为 $t_{ii}^{(\cdot)}=\sum\limits_{j=1}^{l}\boldsymbol{A}_{ij}^{(\cdot)}$, 即 $\boldsymbol{T}_i^{(\cdot)}=\sum\limits_{j=1}^{l}\boldsymbol{A}_{ij}^{(\cdot)}$. $\boldsymbol{H}^w$ 和 $\boldsymbol{H}^b$ 统称为局部离散度矩阵.

定义 1 中, 局部类内图离散度矩阵 $\boldsymbol{H}^w$ 描述了输入样本流形的局部结构信息, 局部类间图离散度矩阵 $\boldsymbol{H}^b$ 描述了输入样本流形的局部判别信息.

定义 2 (局部信息度量)　类似于标准 SVM, 假设分类器具有线性形式 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$, 则

$$\begin{aligned}S^w=&\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2\boldsymbol{A}_{ij}^w=\\&\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j)^2\boldsymbol{A}_{ij}^w=\\&\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{T}^w-\boldsymbol{A}^w)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}=\\&\boldsymbol{w}^{\mathrm{T}}\boldsymbol{H}^w\boldsymbol{w}\end{aligned}\tag{5}$$

称为局部类内紧性度量, 而

$$\begin{aligned}S^b=&\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2\boldsymbol{A}_{ij}^b=\\&\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j)^2\boldsymbol{A}_{ij}^b=\\&\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{T}^b-\boldsymbol{A}^b)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}=\\&\boldsymbol{w}^{\mathrm{T}}\boldsymbol{H}^b\boldsymbol{w}\end{aligned}\tag{6}$$

称为局部类间散性度量. 其中: $\boldsymbol{w}$ 为分类器的权重向量, $\boldsymbol{H}^w$ 和 $\boldsymbol{H}^b$ 分别为局部类内图离散度矩阵和局部类间图离散度矩阵. S^w 和 S^b 统称为局部信息度量.

由模式分类的最大间隔准则, 在输出空间中, 期望 G^w 中的近邻样本应尽可能地紧凑, 同时 G^b 中的近邻样本应尽可能地分开. 于是定义局部信息差度量如下.

定义 3 (局部信息差度量)　称

$$\begin{aligned}\Delta S=&\delta S^w-(1-\delta)S^b=\\&\delta\boldsymbol{w}^{\mathrm{T}}\boldsymbol{H}^w\boldsymbol{w}-(1-\delta)\boldsymbol{w}^{\mathrm{T}}\boldsymbol{H}^b\boldsymbol{w}=\\&\boldsymbol{w}^{\mathrm{T}}\Delta\boldsymbol{H}\boldsymbol{w}\end{aligned}\tag{7}$$

为局部信息差度量. 其中: $\delta\in(0,1]$ 为局部信息平衡参数, $\Delta\boldsymbol{H}=\delta\boldsymbol{H}^w-(1-\delta)\boldsymbol{H}^b=\boldsymbol{X}\Delta\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}$ 为局部信息差矩阵, $\Delta\boldsymbol{L}=\delta\boldsymbol{L}^w-(1-\delta)\boldsymbol{L}^b$.

定义 3 中, 参数 δ 用于在输入样本流形的局部几何结构 (局部类内紧性) 与局部判别信息 (局部类间散性) 之间寻求一个满意的平衡. 当 δ 增大时, 倾向于保持局部几何结构, 同时减少对局部判别信息的惩罚; 反之, 则放松对局部几何结构的要求, 加大惩罚局部判别信息. 只要在适当的 δ 值下, ΔS 就能既较好地保持局部几何结构, 又具有较好的模式判别信息[1].

同时可得到整体图 G 下的 $T_i=\sum\limits_{j=1}^{l}(\boldsymbol{A}_{ij}^w+\boldsymbol{A}_{ij}^b)$, 称为节点 i 的度. T_i 度越大, 节点 i 越重要. 在很少的情况下, T_i 的值会为零, 此时节点 i 称为孤立的. 于是得到如下定义.

定义 4 (节点 i 全局权重)　设 T_i 为图 G 中节点 i 的度, 定义

$$\rho_i=T_i/T\tag{8}$$

为节点 i 的全局权重, 其中 $T=\sum\limits_{i=1}^{l}T_i$.

节点 i 的全局权重反映了样本点 $\boldsymbol{x}_i$ 在流形全局分布上的自然度量, 如果样本点 $\boldsymbol{x}_i$ 的全局权重 ρ_i 越大, 则 $\boldsymbol{x}_i$ 的近邻在全局流形结构上分布得越紧凑, 这意味着 $\boldsymbol{x}_i$ 与其近邻之间越相似; 反之, 则说明 $\boldsymbol{x}_i$ 与其近邻越不相似, 越有可能是野值点.

由文献 [1] 可知, 局部类内图和类间图中的最近邻数 k 将严重影响所建立的分类器的投影方向, 最终影响分类性能, 但如何选取最近邻数 k 却没有相应的方法. 为此, 本文提出一种对数据分布敏感的自适应局部图建立方法, 其基本过程如下:

首先, 通过式 (9) 计算样本 $\boldsymbol{x}_i$ 与所有其他样本之间的平均相似度

$$\mathrm{AS}(\boldsymbol{x}_i)=\frac{1}{l}\sum_{j=1}^{l}W_{ij};\tag{9}$$

然后, 按照式 (10) 和 (11) 自适应确定 $\boldsymbol{x}_i$ 的类内近邻点集合 $N_k^w(\boldsymbol{x}_i)$ 及类间近邻点集合 $N_k^b(\boldsymbol{x}_i)$, 即

$$N_k^w(\boldsymbol{x}_i)=\{\boldsymbol{x}_j|l_i=l_j,W_{ij}^w>\mathrm{AS}(\boldsymbol{x}_i)\},\tag{10}$$

$$N_k^b(\boldsymbol{x}_i)=\{\boldsymbol{x}_j|l_i=l_j,W_{ij}^b>\mathrm{AS}(\boldsymbol{x}_i)\}.\tag{11}$$

根据式 (10) 和 (11), $N_k^w(\boldsymbol{x}_i)$ 定义为相似度大于

平均相似度的与$\boldsymbol{x}_i$同类的样本集合, $N_k^b(\boldsymbol{x}_i)$定义为相似度大于平均相似度的与$\boldsymbol{x}_i$异类的样本集合. 于是可以自适应得到数据集$\boldsymbol{X}$的类内图G^w和类间图G^b, 并准确反映样本流形的局部结构信息和局部判别信息.

1.2 线性ALGEWP-SVM

不同于标准SVM和ν-LPMIVSVM方法, 本文方法在考虑数据整体类间间隔最大化的基础上还引入了数据流形的局部信息, 同时在惩罚系数中引入了数据的全局结构信息. 因此, 对于一个二分类问题, ALGEWP-SVM方法的原始优化问题可以描述为

$$\min_{\boldsymbol{w},b,\lambda,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2+\frac{\lambda}{2}\boldsymbol{w}^{\mathrm{T}}\Delta\boldsymbol{H}\boldsymbol{w}+C\sum_{i=1}^{l}\rho_i\xi_i. \tag{12}$$

$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1-\xi_i,\ i=1,2,\cdots,l; \tag{13}$$

$$\xi_i\geqslant 0,\ i=1,2,\cdots,l. \tag{14}$$

其中: $C>0$是惩罚参数; $\lambda\geqslant 0$是正则化参数, 用于调节局部流形信息的相对重要性; $\boldsymbol{\xi}=[\xi_1,\xi_2,\cdots,\xi_l]$为松弛向量; $\rho_i,\Delta\boldsymbol{H}$分别为如上定义的样本点$\boldsymbol{x}_i$的全局权重和局部信息差矩阵.

ALGEWP-SVM方法具有与标准SVM相似的原始优化问题, 当正则化参数$\lambda=0,\rho_i$为常数时, ALGEWP-SVM方法就是标准的SVM, 因此, 可以将SVM看作是ALGEWP-SVM方法的一个特例. 为求解线性ALGEWP-SVM的最优化问题, 类似于标准SVM的推导方法, 将它作为原始最优化问题, 应用拉格朗日对偶性, 通过求解对偶问题得到原始问题的最优解. 于是有如下结论.

定理 1 线性ALGEWP-SVM方法的对偶最优化问题为

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{x}_j-\sum_{i=1}^{l}\alpha_i. \tag{15}$$

$$\text{s.t. } 0\leqslant\alpha_i\leqslant C\rho_i,\ i=1,2,\cdots,l; \tag{16}$$

$$\sum_{i=1}^{l}\alpha_i y_i=0. \tag{17}$$

其中: $\boldsymbol{M}=\boldsymbol{I}+\lambda\Delta\boldsymbol{H}$; α_i为Lagrangian乘子; $\boldsymbol{I}$为单位矩阵; ALGEWP-SVM原始优化问题中权值向量$\boldsymbol{w}^*$和偏置b^*分别为

$$\boldsymbol{w}^*=\sum_{i=1}^{l}\alpha_i^* y_i\boldsymbol{M}^{-1}\boldsymbol{x}_i, \tag{18}$$

$$b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i^*\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{x}_j. \tag{19}$$

为了测试新样本$\boldsymbol{x}\in\boldsymbol{X}$的类别, ALGEWP-SVM的决策函数为

$$f(\boldsymbol{x})=\operatorname{sgn}\Big(\sum_{i=1}^{l}\alpha_i^* y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{x}+b^*\Big). \tag{20}$$

算法 1 线性ALGEWP-SVM算法.

输入: 训练数据集$T=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\}$, 其中$\boldsymbol{x}_i\in\boldsymbol{X}=R^n, y_i\in\boldsymbol{Y}=\{-1,1\}, i=1,2,\cdots,l$;

输出: 分离超平面和分类决策函数.

Step 1: 选择合适的权重函数, 根据式(10)和(11)自适应构造数据集$\boldsymbol{X}$的k近邻加权连接图G^w、G^b;

Step 2: 分别计算图G^w、G^b的拉普拉斯矩阵并根据定义1、定义4计算$\boldsymbol{H}^w$、$\boldsymbol{H}^b$和ρ_i;

Step 3: 根据定义3计算局部信息差矩阵$\Delta\boldsymbol{H}$;

Step 4: 选择适当的惩罚参数C、正则化参数λ、局部信息平衡参数δ, 构造并求解约束最优化问题(15)～(17), 求得最优解$\boldsymbol{\alpha}$, 分别计算式(18)和(19), 得到最优权值向量$\boldsymbol{w}$和偏置b;

Step 5: 求得分离超平面$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$, 分类决策函数如式(20)所示.

需要说明的是, 与ν-LPMIVSVM方法不同, 本文方法通过计算矩阵$\boldsymbol{M}$的逆可避免矩阵的奇异值问题, 即所谓的小样本问题[1]; 只要选择适当的正则化参数λ, 就可以使得矩阵$\boldsymbol{M}$一定可逆, 从而避免通过在ν-LPMIVSVM方法中引入降维算法所引起的降维维数优化、时间消耗等问题.

1.3 非线性扩展

类似于标准SVM, 对于线性不可分问题, 一个自然的想法就是通过引入一个非线性变换ϕ将输入空间对应到一个特征空间(通常为可再生Hilbert空间)中, 使得在输入空间中的超曲面模型对应于特征空间中的超平面模型. 这样, 分类问题的学习任务就可以通过在特征空间中求解线性ALGEWP-SVM来完成, 可以进一步提高算法在处理复杂非线性分类问题时的性能.

定理 2 非线性ALGEWP-SVM的原始优化问题可表示为

$$\min_{\boldsymbol{w},b,\lambda,\boldsymbol{\xi}}\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{\lambda}{2}\boldsymbol{w}^{\mathrm{T}}\Delta\boldsymbol{H}^{\phi}\boldsymbol{w}+C\sum_{i=1}^{l}\rho_i\xi_i. \tag{21}$$

$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,\ i=1,2,\cdots,l; \tag{22}$$

$$\xi_i\geqslant 0,\ i=1,2,\cdots,l. \tag{23}$$

其中 $\Delta\boldsymbol{H}^{\phi}$为特征空间中的相应局部信息差矩阵. 式(21)～(23)相应的对偶问题为

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{M}^{\phi^{-1}}\phi(\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i. \tag{24}$$

$$\text{s.t. } 0\leqslant\alpha_i\leqslant C\rho_i,\ i=1,2,\cdots,l; \tag{25}$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0. \tag{26}$$

其中 $\boldsymbol{M}^{\phi} = \boldsymbol{I} + \lambda\Delta\boldsymbol{H}^{\phi}$.

然而, 特征空间一般是高维的, 甚至是无穷维的, 因此核技巧的想法是, 在学习和预测中只定义核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j)$, 而不显式地定义非线性映射函数 ϕ. 注意到, 在线性 ALGEWP-SVM 方法的对偶问题中, 无论是目标函数还是决策函数都类似于标准 SVM, 都只涉及到实例与实例之间的内积, 因此, 极有可能使用核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j)$ 来代替目标优化函数中可能的内积项. 实际上, 有下面的定理.

定理 3　非线性 ALGEWP-SVM 原始优化问题的对偶问题 (24) ~ (26) 可进一步表示为如下核形式:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(K_{ij} - \lambda\boldsymbol{K}_i^{\mathrm{T}}\Delta\boldsymbol{M}\boldsymbol{K}_j) - \sum_{i=1}^{l}\alpha_i. \tag{27}$$

$$\text{s.t. } 0 \leqslant \alpha_i \leqslant C\rho_i,\ i = 1, 2, \cdots, l; \tag{28}$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0. \tag{29}$$

其中: Mercer 核函数 $\boldsymbol{K} = (K_{ij}) \in R^{l\times l}$, 称为 Gram 矩阵, K_{ij}=$K(\boldsymbol{x}_i, \boldsymbol{x}_j)$; $\boldsymbol{K}_i$ 表示 Gram 矩阵的第 i 列; $\Delta\boldsymbol{M} = \Delta\boldsymbol{L}(\Delta\boldsymbol{L} + \lambda\Delta\boldsymbol{L}\boldsymbol{K}\Delta\boldsymbol{L})^{-1}\Delta\boldsymbol{L}, \Delta\boldsymbol{L} = \delta\boldsymbol{L}^{w} - (1-\delta)\boldsymbol{L}^{b}$.

证明　证明式 (27) 与 (24) 等价, 实际上只需证明下式成立:

$$\phi(\boldsymbol{x}_i)^{\mathrm{T}}(\boldsymbol{I} + \lambda\Delta\boldsymbol{H}^{\phi})^{-1}\phi(\boldsymbol{x}_j) = K_{ij} - \lambda\boldsymbol{K}_i^{\mathrm{T}}\Delta\boldsymbol{L}(\Delta\boldsymbol{L} + \lambda\Delta\boldsymbol{L}\boldsymbol{K}\Delta\boldsymbol{L})^{-1}\Delta\boldsymbol{L}\boldsymbol{K}_j. \tag{30}$$

由 Sherman-Morrison-Woodbury 公式[30-31]

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\boldsymbol{B}(\boldsymbol{B} + \boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}\boldsymbol{B})^{-1}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}, \tag{31}$$

有

$$(\boldsymbol{I} + \lambda\Delta\boldsymbol{H}^{\phi})^{-1} = (\boldsymbol{I} + \lambda\boldsymbol{X}^{\phi}\Delta\boldsymbol{L}\boldsymbol{X}^{\phi^{\mathrm{T}}})^{-1} = \boldsymbol{I} - \lambda\boldsymbol{X}^{\phi}\Delta\boldsymbol{L}(\Delta\boldsymbol{L} + \lambda\Delta\boldsymbol{L}\boldsymbol{X}^{\phi^{\mathrm{T}}}\boldsymbol{X}^{\phi}\Delta\boldsymbol{L})^{-1}\Delta\boldsymbol{L}\boldsymbol{X}^{\phi^{\mathrm{T}}}. \tag{32}$$

将式 (32) 左右两边各乘 $\phi(\boldsymbol{x}_i^{\mathrm{T}})$ 和 $\phi(\boldsymbol{x}_j)$, 经过化简便可得到式 (30), 由此定理 3 得证. □

选择 $\boldsymbol{\alpha}$ 中的一个分量适合条件 $0 < a_i < C\rho_i$, 计算核 ALGEWP-SVM 方法的偏置 b^{ϕ^*} 为

$$b^{\phi^*} = y_j - \sum_{i=1}^{l} y_i\alpha_i^*(K_{ij} - \lambda\boldsymbol{K}_i^{\mathrm{T}}\Delta\boldsymbol{M}\boldsymbol{K}_j). \tag{33}$$

综上可知, 对于某个测试样本 $\boldsymbol{x} \in \boldsymbol{X}$, 非线性 ALGEWP-SVM 的决策函数为

$$f(x) = \mathrm{sgn}\Big(\sum_{i=1}^{l}\alpha_i^* y_i\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{M}^{\phi^{-1}}\phi(\boldsymbol{x}) + b^{\phi^*}\Big) = \mathrm{sgn}\Big(\sum_{i=1}^{l}\alpha_i^* y_i(K_{i\cdot} - \lambda\boldsymbol{K}_i^{\mathrm{T}}\Delta\boldsymbol{M}K_{i\cdot}) + b^{\phi^*}\Big), \tag{34}$$

其中 $K_{i\cdot} = K(\boldsymbol{x}_i, \boldsymbol{x})$.

需要说明的是, 本文方法在考虑数据整体类间间隔最大化的同时还引入了数据的局部结构信息和局部判别信息, 这既体现了模式分类方法的大间隔准则仍符合流形学习的保持原则, 同时还在惩罚系数中引入了数据的全局结构信息来自动抑制野值对分类面的影响, 增强算法的鲁棒性. 因此, 本文所提出的方法是一种比较合理的大间隔分类学习方法.

2　讨　论

2.1　模型参数选择

机器学习的模型参数选择方法通常采用交叉验证法[11], 即将由 l 个样本组成的样本集分为个数为 l_1 的训练集和个数为 l_2 的测试集, 在不同的给定可调参数下, 用训练集对模型进行训练, 得到相应的模型最优参数, 而后在测试集上进行分类误差检验, 选择具有最低校验误差的可调参数值作为最终的模型参数值. 本文提出的 ALGEWP-SVM 方法也是用这种模式确定可调参数的值.

ALGEWP-SVM 方法主要有 3 个可调参数: 损失函数相对于模型复杂性之间的惩罚参数 C; 正则化参数 λ 和平衡局部结构信息; 局部判别信息的参数 δ. 由于 ALGEWP-SVM 分类算法用于分类的主要性能衡量指标就是预测精度 $\mathrm{acc}(C, \lambda, \delta)$, 它是 C、λ 和 δ 的函数. 分类算法需要根据性能指标选择适当的 C、λ 和 δ 值, 以实现最佳的分类预测效果. 与文献 [1] 不同, 本文并不采用 10 重交叉验证法进行每个参数的选择, 首先是因为这里需要优化的参数较多, 如果对每个参数都进行 10 重交叉验证选择最优, 则对于 ALGEWP-SVM 分类算法, 其组合参数也并不一定是最优的; 其次, 对每个参数都进行 10 重交叉验证, 其计算量太大, 计算效率低下. 因此, 本文最优解路径追踪算法使用一种特殊的网格法, 首先将 3 个参数的取值范围进行人为划分; 然后在固定参数 δ 的情况下, 在 C 与 λ 组成的网格上调用算法 1 进行最优 (C, λ) 组合参数扫描, 求解 $\boldsymbol{w}^*$ 和 b^* 在测试集上预测样本的类标签, 最终选择使得分类精度最大的参数组合; 最后固定最佳的参数组合 (C, λ) 遍历 δ 值, 选取最大的 $\mathrm{acc}(C, \lambda, \delta)$ 进行输出. 具体的参数选择如下: 共取 15 个可能的固定值, 以等比增长分布, 初始值为 0.062 5, 终值为

1 024, 等比为 2. 对于每一个固定的 C 值选取不同的 λ 值, λ 值以等差增长分布, 初始值为 0, 终值为 2, 等差为 0.2. 同样, 参数 δ 也采用等差增长分布, 初始值为 0, 终值为 1, 等差为 0.1. 这样, 全部解路径由 176 个节点组成, 在每个节点上, 进行 10 次重复实验, 将平均值作为该节点的最终分类精度. 获得全部节点数据后绘制解路径, 找出解的变化规律和最优解.

本文之所以采用这种参数选择策略是出于以下考虑: 首先, 类似于标准 SVM, 惩罚参数 C 从一定程度上决定了算法的“过学习”程度, 是用于衡量模型复杂性的主要参数之一, 同时正则化参数 λ 则用于调节局部流形信息的相对重要性, 反映了输入空间的内在几何结构. 这两个参数类似于流形正则化中的外内正则化参数, 是影响 ALGEWP-SVM 分类算法中权值向量的主分量; 其次, 参数 δ 只是强调数据在局部既能保持较好的流形结构, 又具有较强的局部判别信息, 是 ALGEWP-SVM 分类算法中权值向量的微调量. 因此, 本文确定为如上的参数选择策略.

还需要明确指出一点, 对于热核函数参数 t, 本文参照文献 [28] 选择 $t\to\infty$, 即选择矩形核作为近邻图中样本点之间的相似度度量. 核函数参数 (即 RBF 核参数 σ^2) 设置为与随机从训练集中挑出的样本之间的欧氏距离的一半相同[32].

2.2 泛化性能分析

Vapnik 等[2]在传统的基于结构风险最小化为原则的分类器设计中使用 VC 维去估计学习机的推广误差, 在一定程度上为学习机的推广误差分析提供了一个优良的归纳推理准则. 但是, 由于其未考虑训练样本的先验信息, 在很多实际问题中, VC 维其实并不适合用于对分类器的泛化性能进行估计 (实际上 VC 维只适用于线性分类器, SVM 除外[33-34]). 近年来, 针对融合了数据先验分布信息的分类器的泛化性能估计, 不少学者提出了一些新的泛化误差界估计方法来代替传统的 VC 维[8,33-35]. 比如 Rademacher 泛化误差界估计方法, 该方法不但能更有效地用于分类器泛化误差的性能估计, 而且还具有比 VC 维方法更简单的表现形式. 特别地, 在核学习机器中, 对于 Rademacher 泛化误差界的上界有如下定理[8].

定理 4 如果 $K:\boldsymbol{x}\times\boldsymbol{x}\to R$ 是一个核, $S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l\}$ 是数据集 $\boldsymbol{X}$ 上的训练样本, 则分类器 $\mathcal{F}_B$ 的经验 Rademacher 泛化误差界满足

$$R_l(\mathcal{F}_B)\leqslant\frac{2B}{l}\sqrt{\sum_{i=1}^{l}K_{ii}}. \tag{35}$$

其中: $K_{ii}=K(\boldsymbol{x}_i,\boldsymbol{x}_i)$, B 是分类器中权值向量 $\boldsymbol{w}$ 的上界.

根据定理 4, 可以给出 ALGEWP-SVM 的泛化误差界.

定理 5 ALGEWP-SVM 的泛化误差界的上界 $R_{\mathrm{ALGEWP-SVM}}(f)$ 至多为 SVM 的上界 $R_{\mathrm{SVM}}(f)$.

证明 重写式 (24) 如下:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{M}^{\phi^{-1}}\phi(\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i. \tag{36}$$

对每个特征空间中的样本 $\phi(\boldsymbol{x})$ 作如下变换:

$$\widetilde{\phi}(\boldsymbol{x})=(\boldsymbol{I}+\lambda\Delta\boldsymbol{H}^{\phi})^{-1/2}\phi(\boldsymbol{x}). \tag{37}$$

于是, 式 (36) 可重新表示为

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j\widetilde{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i. \tag{38}$$

式 (38) 非常类似于标准 SVM 中的对偶问题, 因此有

$$\begin{aligned}&\sum_{i=1}^{l}K_{ii}=\sum_{i=1}^{l}\widetilde{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)=\\&\sum_{i=1}^{l}\phi(\boldsymbol{x}_i)(\boldsymbol{I}+\lambda\Delta\boldsymbol{H}^{\phi})^{-1}\phi(\boldsymbol{x}_i)=\\&\sum_{i=1}^{l}(K_{ii}-\lambda\boldsymbol{K}_i^{\mathrm{T}}\Delta\boldsymbol{M}\boldsymbol{K}_i),\end{aligned} \tag{39}$$

且 $\Delta\boldsymbol{M}$ 是一个对称半正定矩阵, $\lambda\geqslant 0$, 所以

$$\begin{aligned}&R_{\mathrm{ALGEWP-SVM}}(f)\leqslant\\&\frac{B}{2l}\sqrt{\sum_{i=1}^{l}(K_{ii}-\lambda\boldsymbol{K}_i^{\mathrm{T}}\Delta\boldsymbol{M}\mathbf{K}_i)}\leqslant\\&\frac{B}{2l}\sqrt{\sum_{i=1}^{l}K_{ii}}=R_{\mathrm{SVM}}(f),\end{aligned} \tag{40}$$

当且仅当 $\lambda=0$ 时等式成立, 由此定理得证. □

由定理 5 分析可知, 考虑数据整体类间间隔最大化的同时将数据的局部信息融入到分类器的设计中, 将进一步降低分类器的泛化误差, 这意味着 ALGEWP-SVM 具有比 SVM 更好的推广性能.

2.3 算法复杂度分析

ALGEWP-SVM方法的时间复杂度主要由 3 方面决定: 1) 自适应近邻图近邻点 k 的计算; 2) 矩阵逆的求解; 3) 凸二次规划问题的求解.

ALGEWP-SVM 方法中, 关于近邻点 k 的计算问题只涉及到图 G^w、G^b 近邻点 k 的计算, 对于图 G 的近邻点可通过图 G^w、G^b 近似相加得到, 因此其时间复杂度为 $O(k_w l^2+k_b l^2+3nl^2)$. 其中: $O(3nl^2)$ 代表图 G、G^w、G^b 计算任意两个样本的欧氏距离的时间复杂度, n 为样本的维度, l 为样本的个数; $O(k_w l^2+k_b l^2)$ 代表寻找图 G^w、G^b 近邻点个数的时间复杂度, k_w、k_b 分别代表图 G^w、G^b 近邻点个数. 又有矩阵 $\Delta\boldsymbol{L}$

$+\lambda\Delta\boldsymbol{LK}\Delta\boldsymbol{L}$ 为 $l\times l$ 矩阵, 求解其逆矩阵的时间复杂度为 $O(l^3)$. 一般的凸二次规划问题的求解时间复杂度也是 $O(l^3)$. 因此, ALGEWP-SVM 算法的时间复杂度为 $O(2l^3+(k_w+k_b+3n)l^2)$. 由于图近邻个数 k 远小于样本维数 n, ALGEWP-SVM 算法的时间复杂度近似为 $O(l^3+nl^2)$. 相较于标准 SVM 的时间复杂度 $O(l^3)$ 而言, 本文算法还与样本的维度有关, 因此, 在处理高维样本问题时, 为了在一定程度上提高本文方法的执行效率, 可以首先采用降维方法 (如 PCA、LDP 等) 对数据进行预处理, 以提高所提出方法的执行效率.

3　实验分析

为了验证本文方法的有效性, 分别将其用于人工双月 two-moons 数据集[32]、UCI 数据集[1,27]和实际的图像识别数据集[36]进行测试, 并与相关方法进行对比, 以显示本文方法的分类效果. 通过测试人工数据集直观地演示 ALGEWP-SVM 方法在二维平面上选择分类决策边界的过程、原理及优势; 测试 UCI 数据集和实际的图像识别数据集, 用以说明本文方法在实际的模式识别问题中的分类性能. 表 1 列出了实验中的比较方法.

表 1　实验中的比较算法

缩写	全称	引用
SVM	支持向量机	文献 [2]
LapSVM	拉普拉斯支持向量机	文献 [23]
MCLPVSVM	最小类内局部保留方差支持向量机	文献 [24]
ν-LPMIVSVM	局部保留最大信息差 ν-支持向量机	文献 [1]
ALGEWP-SVM	自适应局部图嵌入加权罚支持向量机	本文方法

需要说明的是, 非线性 ALGEWP-SVM 方法的核函数采用径向基函数 (RBF)

$$K(\boldsymbol{x},\boldsymbol{y})=\exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/\sigma^2),$$

其中 σ^2 的取值如 2.1 节所述, 即设置为随机从训练集中挑出的样本之间的欧氏距离的一半. SVM, MCLPVSVM, ν-LPMIVSVM 方法的参数确定与文献 [1] 相同, LapSVM 方法的内外正则化参数选择采用文献 [32] 中的策略, 而方法 ALGEWP-SVM 的参数选取策略如 2.1 节所述.

3.1　人工数据实验

人工双月 two-moons 数据集经常被用于测试一些流形学习方法[1,32]. 本节将通过与 SVM 和 ν-LPMIVSVM 方法进行比较, 分析本文方法在 3 种不同复杂度数据集上得到的决策边界的分类性能.

3 种 two-moons 数据集中, 每类均有 100 个样本, 随机抽取训练集和测试集各一半, 重复 10 次, 分别记录实验结果. 表 2 和表 3 分别列出了算法相应的训练和测试精度, 表 4 列出了算法中 Lagrangian 乘子中非零值的个数, 即与标准 SVM 中支持向量的概念相同. 3 种方法的分类决策边界如图 1 所示.

表 2　SVM, ν-LPMIVSVM 和 ALGEWP-SVM 方法的训练精度　%

方法	情形 1	情形 2	情形 3
SVM	100	97.4	91.6
ν-LPMIVSVM	100	100	98
ALGEWP-SVM	100	100	99.5

表 3　SVM, ν-LPMIVSVM 和 ALGEWP-SVM 方法的测试精度　%

方法	情形 1	情形 2	情形 3
SVM	100	97.4	88.4
ν-LPMIVSVM	100	98	96
ALGEWP-SVM	100	100	99

表 4　SVM, ν-LPMIVSVM 和 ALGEWP-SVM 方法的支持向量个数

方法	情形 1	情形 2	情形 3
SVM	2.4	5.7	15.6
ν-LPMIVSVM	49.3	49.7	50
ALGEWP-SVM	12.9	18.6	31.2

从表 2 ~ 表 4 以及图 1 结果可以看出:

1) 当两类数据相距较远, 相对容易分类时, 标准 SVM、ν-LPMIVSVM 和本文 3 种方法的分类决策边界都保持了其平滑性, 如图 1(a)、图 1(d) 和图 1(g)所示. 但是, 当数据集的拓扑结构变得更加复杂, 分类复杂度逐步增加时, 标准 SVM 方法的分类决策边界依然保持了平滑性, 导致其分类精度急剧下降, 如图 1(c) 所示的决策边界似乎明显过于平滑, 而充分考虑局部流形结构的 ν-LPMIVSVM 和本文方法其分类决策边界更加符合样本的几何分布, 这验证了仅仅关注整体类间间隔最大化对于复杂的分类问题是远远不够的.

2) 数据集拓扑结构变得复杂时, ν-LPMIVSVM 和本文方法的决策边界均显示出局部不平滑性, 见图 1(e)、图 1(f)、图 1(h) 和图 1(i). 但是, ν-LPMIVSVM 仅关注数据的局部流形结构和判别信息, 造成其分类决策边界更加靠近数据分布急剧变化区 (见图 1(e)); 相反的, 本文方法从全局的观点出发, 在强调数据整体分布结构的同时考虑了数据的局部信息, 既充分保持了数据整体类间间隔, 又充分挖掘了数据的局部流形结构和判别信息, 因此, 本文算法得到了很高的训练和测试精度 (见图 1(i)).

3) 表 4 显示, 标准 SVM 的解具有高度的稀疏性, 而 ν-LPMIVSVM 方法的解完全丧失了稀疏性; 相对于上述两种方法, 本文方法取得的解尚在可接受的范围内, 有利于提高大数据集的测试速度.

4) 标准 SVM 和 ν-LPMIVSVM 方法对数据集中的所有样本点均采用相同的惩罚系数, 导致其决策边

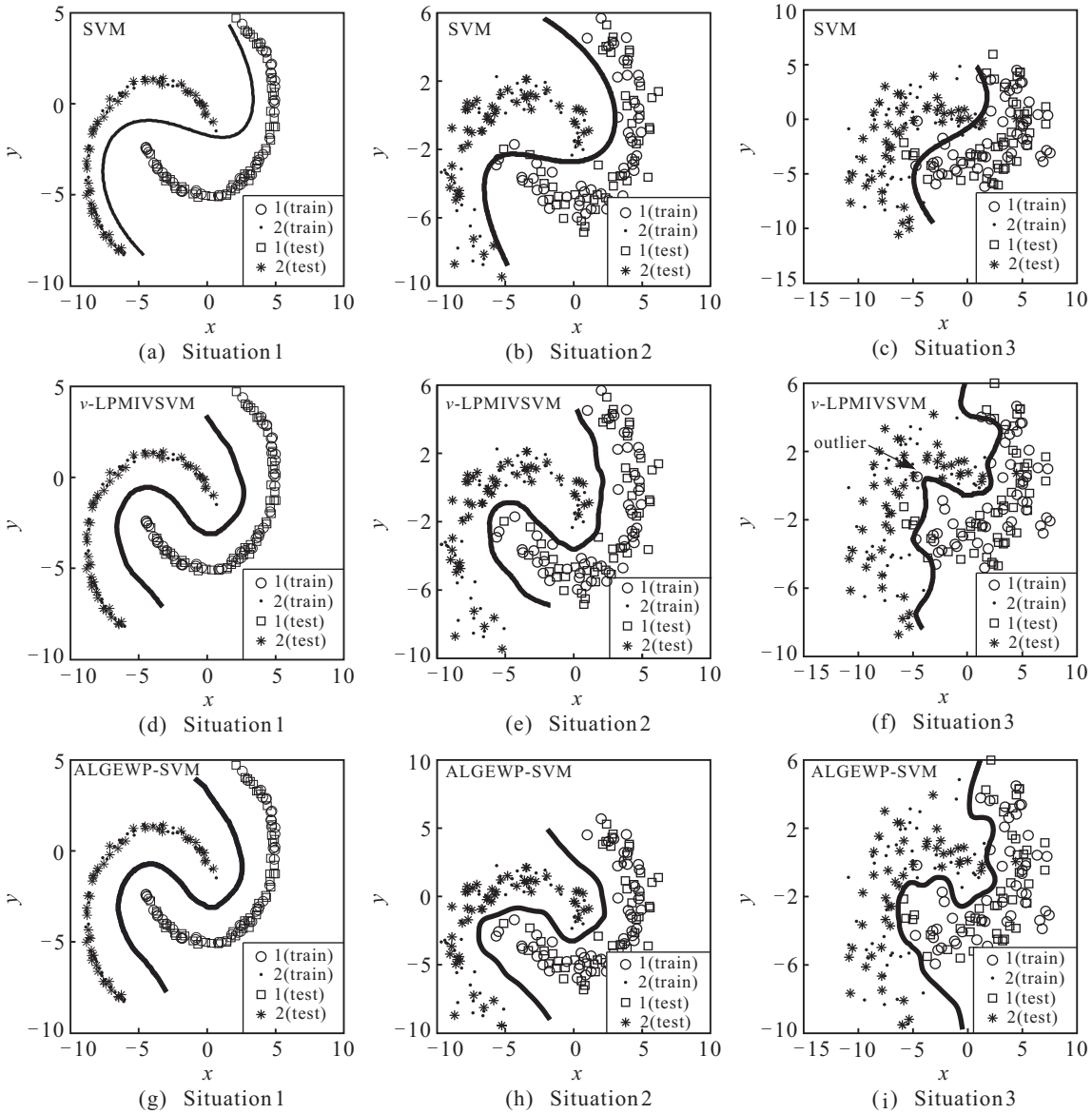

图 1 SVM, ν-LPMIVSVM 和 LGEWP-SVM 方法在不同数据集上的决策边界

界对数据集中的野值点敏感, 如图 1(f) 中的标示点, 可以看出其决策边界明显偏向于数据集中的野值点; 而本文方法兼顾了全局权重, 在相应的惩罚参数前乘以样本的相应全局权重, 能增强算法的鲁棒性, 取得更加合理的决策边界.

同时, 为了考察参数 C、λ 和 δ 对本文方法的性能影响, 选取情形 3 所示的数据集分别进行 10 次实验, 分别记录参数 C、λ 和 δ 单独变化时对本文方法性能的影响, 如表 5 所示.

表 5 本文方法在不同参数值下的分类精度 %

	参数 C 单独变化对分类精度的影响 ($\lambda = 1.2, \delta = 0.9$)									
C	2	4	8	16	32	64	128	256	512	1024
分类精度	44	61	79	89	94	96	98	98	99	96
	参数 λ 单独变化对分类精度的影响 ($C = 512, \delta = 0.9$)									
λ	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
分类精度	89	90	92	94	96	99	96	95	92	92
	参数 δ 单独变化对分类精度的影响 ($C = 512, \lambda = 1.2$)									
δ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
分类精度	91	91	91	93	94	96	97	98	99	96

从表 5 中可以看出, C、λ 和 δ 的选取对本文方法都有不同程度的影响, 但相较于参数 δ, 参数 C、λ 的选取对算法分类性能的影响更大, 这与 2.1 节的分析是一致的. 可见, 要想获得较理想的分类性能, 需要通过调节不同的 C、λ 和 δ 的参数组合.

3.2 标准数据集实验

为了更全面地评价本文方法, 在 UCI 标准数据集上对算法进行测试, 同时系统地比较了本文方法与表 1 中所列出的各种方法的分类性能, 以进一步说明本文算法的优越性.

UCI 数据集经常被用来测试模式分类方法的性能[39]. 下面实验将从 UCI 数据库选取 5 个两类数据集和 5 个多类数据集作为测试数据库 (该数据库的下载地址为 http: // www.ics.uci.edu/~mlearn/MLRepository.html), 数据集的详细信息如表 6 所示.

表 6　实验中所用 UCI 数据集基本情况

数据集	维数	类数
Biomed	5	2
Bupa	6	2
Hepatitis	19	2
Pid	8	2
Water	38	2
Iris	4	3
Lenses	4	3
Tae	5	3
Waveform	21	3
Ecoli	6	6

对于每个 UCI 数据集, 随机选择数据集的一半作为训练集, 剩下的为测试集, 对每个数据集都重复 10 次这样的实验, 最后将 10 次测试精度的平均作为实验结果进行记录.

表 7 和表 8 分别记录了采用线性 ALGEWP-SVM 方法和非线性 ALGEWP-SVM 方法情况下的实验结果, 非线性 ALGEWP-SVM 采用 RBF 核函数. 需要指出的是, 各种方法的参数选取方法与 3.1 节相同, 同时对于多分类问题使用 1vs.rest 策略进行.

表 7　5 种方法 (线性核) 在 5 个 UCI 数据集上的分类结果　%

数据集	分类精度					win/tie/loss
	SVM	LapSVM	MCLPVSVM	ν-LPMIVSVM	ALGEWP-SVM	
Biomed	86.9	89.92	90.82	91.67	93.59	4/0/0
Bupa	66.99	68.9	67.3	70.3	72.14	4/0/0
Hepatitis	84.41	85.3	85.46	87.66	86.97	3/0/1
Pid	73.96	70.19	71.04	76.1	79.34	4/0/0
Water	98.64	96.89	95.69	96.72	99.1	3/1/0
Iris	98.56	96.57	97.3	97.15	98.89	3/1/0
Lenses	75.31	81.58	82.01	84.26	83.79	3/0/1
Tae	50.39	52.31	52.1	50.32	51.64	2/0/2
Waveform	88.7	86.72	86.46	88.98	93.21	4/0/0
Ecoli	84.95	84.02	83.59	86.9	87.5	4/0/0
win/tie/loss	9/1/0	9/0/1	8/1/1	6/2/2		

表 8　5 种方法 (非线性核) 在 5 个 UCI 数据集上的分类结果　%

数据集	分类精度					win/tie/loss
	SVM	LapSVM	MCLPVSVM	ν-LPMIVSVM	ALGEWP-SVM	
Biomed	89.2	89.92	91.43	93.7	94.73	4/0/0
Bupa	72.43	78.1	77.31	75.8	81.57	4/0/0
Hepatitis	88.72	89.16	89.07	88.21	88.83	0/3/1
Pid	76.47	73.64	74.1	79.52	79.93	3/1/0
Water	98.64	97.03	96.15	98.9	99.32	4/0/0
Iris	100	99.6	99.3	100	100	0/4/0
Lenses	76.78	86.38	86.69	87.69	85.32	1/0/3
Tae	52.61	58.44	57.83	60.1	61.74	4/0/0
Waveform	91.12	89.63	90.83	90.19	93.26	4/0/0
Ecoli	88.87	89.61	86.59	88.52	89.12	3/1/0
win/tie/loss	8/2/0	7/2/1	8/1/1	6/3/1		

从表 7 和表 8 的结果可以得到如下结论:

1) 核技巧在绝大部分数据集上都能够增强 5 种方法的分类性能. 例如, 在 Iris 数据集上, SVM、ν-LPMIVSVM 和本文方法在采用 RBF 核时都取得了 100% 的正确率.

2) SVM 方法在几乎所有数据集上的分类性能都相对较差, 这清楚地说明仅仅强调数据整体类间间隔的最大化对分类问题是远远不够的, 而在考虑了数据空间流形分布结构的情况下, LapSVM、MCLPVSVM、ν-LPMIVSVM 和本文方法均表现出明显的优势. 因此, 在分类器的设计过程中应该融入更多的数据结构分布信息, 以指导更有效的分类.

3) 特别地, 本文方法和 LapSVM、MCLPVSVM、ν-LPMIVSVM 都在分类器的构造中引入了数据的流形结构, 但 LapSVM 和 MCLPVSVM 方法仅侧重于保持数据的流形结构, ν-LPMIVSVM 方法强调利用数据的判别信息去保持数据的流形结构, 而本文方法则在保持数据整体类间间隔与利用数据的流形结构和判别信息之间寻求一个平衡. 实验结果表明, 在大多数情况下, 本文方法都要优于其他方法.

4) 为了验证本文方法是否从统计意义上显著优于其他方法, 对每个数据集的 10 轮分类过程分别进行纠正重复取样测试. 表 7、表 8 分别列出了在显著性水平为 0.05 的情况下, 本文方法与其他 4 种方法比较的双侧成对测试统计结果. 表中: “win”代表本文方法优于其他方法, “tie”代表无显著差异, “loss”代表本文方法逊于其他方法, 最后一行统计了本文方法与其他方法在所有数据集上比较的 win /tie /loss 个数, 最

后一列统计了在某一特定数据集上, 本文方法与其他方法分别比较的 win /tie /loss 个数. 综合表 7、表 8 可知, 在大部分数据集上本文方法都具有显著优于其他方法的分类性能, 这与上面的分析结果是一致的.

3.3 图像识别实验

图像数据集呈现明显的非线性流形结构, 常常可以嵌入在一个低维内在流形上[15,29], 已被许多基于流形的学习方法用于测试数据集, 以反映流形学习方法的有效性[28], 且由于图像识别数据集一般都是高维数据, 很适合用于对算法复杂度的分析. 因此, 本节继续将 ALGEWP-SVM 方法应用于图像识别问题中, 以考察其对非线性流形数据的分类识别性能和计算效率. 实验中采用了 3 个典型的图像分类数据库, 分别为人脸图像 (ORL)[38]、物体图像 (COIL-20)[1]和美国邮政手写数字库 (USPS)[33]. 其中: ORL 人脸数据集包含 40 个人的 400 张灰度图, 同一类由不同面部表情和面部细节的 10 张图片构成, 实验前将图像分辨率统一裁剪为 40×40 大小; COIL-20 物体数据库含有 20 个物体的 1 440 张灰度图, 每个物体具有在同一位置的 72 张不同视角的图像, 实验前将其裁剪为 32×32 大小; USPS 手写数字数据集包含从 0 到 9 的 10 个数字, 每个数字由 1 100 张图像构成, 图像大小为 16×16.

在整个实验过程中, 每个图像库都被分为不同的训练集和测试集, 参照文献 [39], 用 Gm /Pn 代表从每个数据子集中随机抽取每类对象的 m 个图像用于训练, n 个图像用于测试. 每组实验执行 10 次这个过程, 并计算其平均值作为最终的分类精度. 特别地, 对于 ORL、COIL-20 数据库, 实验中有可能出现样本数小于数据维度的问题, 且在 ν-LPMIVSVM 方法中由于涉及到矩阵求逆的奇异值问题, 需要先采用 PCA 方法对实验数据进行降维以满足矩阵的求逆要求, 然后再采用降维后数据训练 ν-LPMIVSVM 分类器. 需要指出的是, 为了处理高维数据的非线性问题, 实验中所有方法采用的核函数均为 RBF 核函数, 且实验中的结果均是在最优参数下取得的.

图 2 显示了参数 λ 和 δ 对本文方法在非线性流形数据子集 ORL.1-2 上的识别性能影响曲线. 表 9 ～表 12 分别记录了 5 种方法在上述 3 种不同数据集上不同训练集和测试集采样精度下的实验结果. 另外, 表 12 还列出了在 USPS 数据集中的不同数字对中掺杂部分相应最难分的相似数字的分类结果.

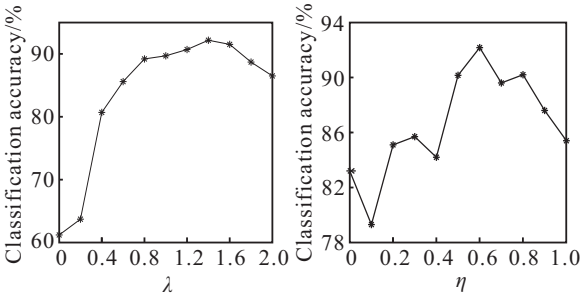

图 2 参数 λ 和 δ 对分类精度的影响

表 9 ORL 实验结果 %

方法	分类精度			
	G3/P7	G5/P5	G7/P3	G9/P1
SVM	50.50	56.82	63.34	61.12
LapSVM	56.32	62.33	70.67	73.36
MCLPVSVM	61.63	70.78	71.46	73.44
ν-LPMIVSVM	66.87	77.43	82.72	86.33
ALGEWP-SVM	70.49	85.56	90.42	92.17
DR-ALGEWP-SVM	68.76	81.57	87.32	90.26

表 10 COIL-20 实验结果 %

方法	分类精度			
	G12/P60	G24/P48	G36/P36	G48/P24
SVM	93.62	95.37	96.33	96.89
LapSVM	94.15	96.48	97.16	97.02
MCLPVSVM	95.56	96.24	97.33	97.33
ν-LPMIVSVM	97.10	98.43	98.78	99.25
ALGEWP-SVM	98.52	99.05	99.23	99.31

表 11 USPS 实验结果 %

方法	分类精度			
	G10/P1090	G100/P1000	G250/P850	G550/P550
SVM	93.96	98.84	99.38	99.6
LapSVM	94.48	98.16	99.25	99.56
MCLPVSVM	94.12	99.32	99.54	99.70
ν-LPMIVSVM	95.60	99.72	99.73	99.74
ALGEWP-SVM	96.30	99.70	99.81	99.92

从表 9 ～表 12 以及图 2 的结果可得到如下结论:

1) 显而易见, 随着训练样本数据的增加, 所有分类算法的识别率都相应增加, 但 SVM 方法在几乎所有图像数据集上的分类性能始终都是最差的. 这说明, 当数据集分布很复杂时, 训练样本数太少, 分类器学习不够充分时, 任何学习方法都很难取得理想的分类效果, 这正应验了引言中的一句话 —— 任何数学技巧都不能补救信息的缺失. 然而, 随着训练样本数目的增加, 考虑数据分布结构的 LapSVM、MCLPVSVM、ν-LPMIVSVM 和 ALGEWP-SVM 方法的分类性能都取得了明显的增加, 尤其是考虑了数据流形局部结构信息和判别信息的 ν-LPMIVSVM 和 ALGEWP-SVM 方法, 在所有数据集上均取得了最佳的识别性能.

2) 由于 ORL 数据集是典型的小样本高维数据, 在 ν-LPMIVSVM 方法的实验中对其进行了 PCA 降维处理, 降维过程有可能会造成信息缺失, 可能降低算法最后的分类性能. 为了公平对比 ν-LPMIVSVM 和 ALGEWP-SVM 方法的分类性能, 表 9 中最后一列列出了两种方法在数据降到相同维度后, ALGEWP-SVM (表中标示为 DR-ALGEWP-SVM) 方法的分类精度. 从中可以看出, 虽然 PCA 降维后的信息损失确实造成了分类性能降低, 但 ALGEWP-SVM 方法仍然取得了更优的分类精度.

表 12　USPS 部分子集掺杂相应最难分的相似数字的分类结果　　%

USPS 数据集	分类精度				
	SVM	LapSVM	MCLPVSVM	ν-LPMIVSVM	ALGEWP-SVM
1-7 (4, 9)	96.83	97.26	97.34	97.38	99.81
2-3 (5, 8)	97.12	97.80	98.10	98.73	99.26
3-5 (8, 6)	96.90	97.01	97.24	97.93	98.68
3-8 (9, 6)	96.47	96.83	96.43	96.89	98.50

注: 表中 1-7 (4, 9) 表示在数字 1 和 7 的分类中掺杂部分数字 4 和 9, 其他类似.

3) 从图 2 中可以看出, 当 $\lambda = 0$ 时, 即忽略数据流形的局部结构信息和判别信息时, ALGEWP-SVM 方法等价于标准 SVM 方法, 相应的分类精度最低, 随着 λ 逐渐增加, ALGEWP-SVM 方法的识别率上升, 并在 $\lambda = 1.4$ 附近取得最大值, 说明同时兼顾整体类间间隔最大化和数据局部信息确实有利于提高分类器的识别性能. 同时, 通过 δ 值调整局部流形几何结构和判别信息所占的比例也能影响分类器的识别性能, 在这点上与文献 [1] 所得出的结论是一致的.

4) 从表 12 中可以看出, 在掺杂了其他数据子集中的样本时, LapSVM、MCLPVSVM、ν-LPMIVSVM 方法由于只关注数据的局部流形结构, 使得数据中的野值点影响了其分类决策边界的分布, 造成其识别率相应降低. 而本文方法在相应的惩罚系数前引入了数据的全局分布权重, 因此, ALGEWP-SVM 方法受野值的影响最小, 依然保持较好的识别率, 算法具有较强的鲁棒性.

最后, 本节利用 USPS 数据集的高维数据量大的特点, 测试了本文方法的计算效率, 并与标准 SVM 方法进行了比较. 标准 SVM 和 ALGEWP-SVM 方法都采用 Matlab 2010 b 编写, 在配置主频为 3.00 GHz 的 CPU (Pentium Dual-Core E5700 CPU) 以及 2 G 内存的计算机上运行. 表 13 为标准 SVM 和 ALGEWP-SVM 方法进行训练集学习所花费的时间. 由表 13 可以看出, ALGEWP-SVM 方法的训练时间增加不少, 但随着训练样本数的增加, 其与标准 SVM 方法的训练时间相差越来越少, 这与 2.3 节的分析基本一致.

表 13　训练时间　　s

方法	G10/P1090	G100/P1000	G250/P850	G550/P550
标准 SVM	0.3	4.7	1 029.4	3 890.3
ALGEWP-SVM	1.6	19.4	2 616.1	7 646.2

4　结　论

标准 SVM 是当前的主流机器学习算法, 但它在构造分类超平面时仅仅关注了整体的类间信息, 而忽略了类内样本的先验分布信息, 从而在一定程度上制约了 SVM 类方法泛化能力的进一步提高. 为此, 本文在保持数据整体类间间隔最大化的同时还充分考虑了数据流形的局部结构信息和判别信息, 而且在软间隔的样本惩罚系数中引入了数据的全局结构信息, 提出了一种新的自适应局部图嵌入加权罚支持向量机 ALGEWP-SVM, 使得本文方法在一定程度上取得了较强的学习泛化能力. 在人工、标准和图像数据集上进行的实验测试都验证了本文方法具有优于相关方法的分类性能. 虽然本文在分类性能上取得了一定的提高, 但随之而造成的参数选择问题却是一个值得考虑的问题, 同时如何提高其计算效率也是一个有待进一步研究的问题.

参考文献(References)

[1] 陶剑文, 王士同. 局部保留最大信息差 ν-支持向量机[J]. 自动化学报, 2012, 38 (1): 97-108.
(Tao J W, Wang S T. Locality-preserved maximum information variance ν-support vector machine[J]. Acta Automatic Sinica, 2012, 38(1): 97-108.)

[2] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995: 59-87.

[3] Bradford J R, West D R. Improved prediction of protein binding sites using a support vector machines approach[J]. Bioinformatics, 2005, 21(8): 1487-1494.

[4] Tao Q, Chu D J, Wang J. Recursive support vector machines for dimensionality reduction[J]. IEEE Trans on Neural Networks, 2008, 19(1): 189-193.

[5] Wu M R, Ye J P. A small sphere and large margin approach for novelty detection using training data with outliers[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088-2092.

[6] Song H, Lee I, Zhao L, et al. Adaptive virtual support vector machine for reliability analysis of high-dimensional problems[J]. Structural and Multidisciplinary Optimization, 2013, 47(4): 479-491.

[7] Hwang J P, Choi B, Kim E. Multiclass lagrangian support vector machine[J]. Neural Computing and Applications, 2013, 22(3): 703-710.

[8] Xue H, Chen S C, Yang Q. Structural support vector machine[J]. Lecture Notes in Computer Science, 2008, 5263(1): 501-511.

[9] Shivaswamy P K, Jebara T. Maximum relative margin and data-dependent regularization[J]. J of Machine Learning Research, 2010, 32(11): 747-788.

[10] Scholkopf B, Smola A J, Williamson R C, et al. New support vector algorithms[J]. Neural Computation, 2000, 12(5): 1207-1245.

[11] Chapelle O, Vapnik V N, Bousquet O, et al. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46(1): 131-159.

[12] Suyken J A K, Brabanter J, De Lukas L, et al. Weighted least square support vector machines: Robustness and sparse approximation[J]. Neurocomputing, 2002, 48(4): 85-105.

[13] Cao P, Zhao D Z, Zaiane O. An optimized cost-sensitive SVM for imbalanced data learning[J]. Lecture Notes in Computer Science, 2013, 7819(1): 280-292.

[14] Peng X J, Xu D. Robust minimum class variance twin support vector machine classifier[J]. Neural Computing and Applications, 2013, 22(5): 999-1011.

[15] Lanczos C. Linear differential operators[M]. London: Van Nostrand, 1964: 31-33.

[16] Zafeiriou S, Tefas A, Pitas I. Minimum class variance support vector machines[J]. IEEE Trans on Image Processing, 2007, 16(10): 2551-2564.

[17] Fisher R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics, 1936, 7(2): 179-188.

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

[19] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

[20] Tenenbaum J B, Silva V, Langford J. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2322.

[21] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering[J]. Advances in Neural Information Processing System, 2002, 14(9): 585-591.

[22] Tao Y T, Yang J. The maximized discriminative subspace for manifold learning problem[J]. Lecture Notes in Computer Science, 2013, 7751(1): 784-792.

[23] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples[J]. J of Machine Learning Research, 2006, 7(12): 2399-2434.

[24] Wang X M, Chung F L, Wang S T. On minimum class locality preserving variance support vector machine[J]. Pattern Recognition, 2010, 43(8): 2753-2762.

[25] Gui J, Wang C, Zhu L. Locality preserving discriminant[J]. Lecture Notes in Computer Science, 2009, 5755(1): 566-572.

[26] 高全学, 谢德燕, 徐辉, 等. 融合局部结构和差异信息的监督特征提取算法[J]. 自动化学报, 2010, 36(8): 1101-1114.
(Gao Q X, Xie D Y, Xu H, et al. Supervised feature extraction based on information fusion of local structure and diversity information[J]. Acta Automatic Sinica, 2010, 36(8): 1101-1114.)

[27] 谢钧, 刘剑. 一种新的局部判别投影方法[J]. 计算机学报, 2011, 34(11): 2243-2250.
(Xie J, Liu J. A new local discriminant projection method[J]. Chinese J of Computers, 2011, 34(11): 2243-2250.)

[28] Wang F. A general learning framework using local and global regularization[J]. Pattern Recognition, 2010, 43(9): 3120-3129.

[29] Yan S, Xu D, Zhang H, et al. Graph embedding and extensions: A general framework for dimensionality reduction[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(1): 40-51.

[30] Woodbury M A. Inverting modified matrices[R]. Princeton: Statistical Research Group, Institute for Advanced Study, 1950: 78-80.

[31] Chun Y D. A generalization of the Sherman-Morrison-Woodbury formula[J]. Applied Mathematics Letters, 2011, 24(9): 1561-1564.

[32] Simon H K. Neural networks and learning machines[M]. 3rd ed. Englewood Cliffs: Prentice Hall, 2009: 198-228.

[33] Bartlett P L, Mendelson S. Rademacher and gaussian complexities: Risk bounds and structural results[J]. J of Machine Learning Research, 2002, 3(3): 463-482.

[34] Koltchinskii V. Rademacher penalties and structural risk minimization[J]. IEEE Trans on Information Theory, 2001, 47(5): 1902-1914.

[35] Sun S L. Multi-view Laplacian support vector machines[J]. Lecture Notes in Computer Science, 2011, 7121(1): 209-222.

[36] Zhang T, Tao D, Li X, et al. Patch alignment for dimensionality reduction[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(9): 1299-1313.

[37] Saul L K, Roweis S T. Think globally, fit locally: Unsupervised learning of low dimensional manifolds[J]. J of Machine Learning Research, 2003, 4(12): 119-155.

[38] Belhumeour P, Hespanha J, Kriegman D. Eigenfaces vs fisherfaces: Recognition using class specific linear projection[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720.

[39] Xue H, Chen S C, Yang Q. Discriminatively regularized least-squares classification[J]. Pattern Recognition, 2009, 42(1): 93-104.

(责任编辑: 曹洪武)