局部保留最大信息差 v–支持向量机

陶剑文 [1, 2] 王士同 [1]

摘 要 针对现有模式分类方法不能较好地保持数据空间的局部流形信息或差异信息等问题, 提出一种基于流形学习的局部保留最大信息差 v–支持向量机 (Locality-preserved maximum information variance v-support vector machine, v-LPMIVSVM). 对于模式分类问题, v-LPMIVSVM 引入局部同类离散度和局部异类离散度概念, 分别体现输入空间局部流形结构和局部差异 (或判别) 信息, 通过最小化局部同类离散度和最大化局部异类离散度, 优化分类器的投影方向. 同时, v-LPMIVSVM 采用适于流形数据的测地线距离来度量数据点对间的相似性, 以更好地反映流形数据的本质结构. 人造和实际数据集实验结果显示所提方法具有良好的泛化性能.

关键词 局部保留投影, v– 支持向量机, 流形学习, 局部同类离散度, 局部异类离散度

DOI 10.3724/SP.J.1004.2012.00097

Locality-preserved Maximum Information Variance v-support Vector Machine

TAO Jian-Wen[1, 2] WANG Shi-Tong[1]

Abstract The state-of-the-art pattern classifiers can not efficiently preserve the local geometrical structure or the diversity (or discriminative) information of data points embedded in high-dimensional data space, which is useful for pattern recognition. A novel so-called locality-preserved maximum information variance v-support vector machine (v-LPMIVSVM) algorithm is presented based on manifold learning to address those problems mentioned above. The v-LPMIVSVM introduces within-locality homogeneous scatter and within-locality heterogeneous scatter, which respectively denote the within-locality manifold information of data points and the within-locality diversity information of data points, thus constructing an optimal classifier with optimal projection weight vector by minimizing the within-locality homogeneous scatter and simultaneously maximizing the within-locality heterogeneous scatter. Meanwhile, the v-LPMIVSVM adopts geodesic distance metric to measure the distance between data in the manifold space, which can reflect the true geometry of the manifold. Experimental results on artificial and real world problems show the outperformed or comparable effectiveness of v-LPMIVSVM.

Key words Locality preserving projections, v-support vector machine (v-SVM), manifold learning, within-locality homogeneous scatter, within-locality heterogeneous scatter

模式分类旨在通过有限的训练样本学习一个分类器, 且该分类器需对未来数据具有良好的泛化能力[1]. 已有多种用于模式分类的方法提出, 其中, 大间隔分类器支持向量机 (Support vector machine, SVM)[2] 及其相关变体是目前实现模式分类的主流方法之一[3], 其通过最大化类间间隔来达到强泛化能力[4]. v–支持向量机 (v-SVM)[5] 是 SVM 的一个扩展变体, 最早由 Schölkopf 等提出, 通过引入一个新的参数 v 来控制支持向量数下界和训练误差上界. 尽管 SVM 及其变体 (v-SVM) 已在机器学习和模式识别领域得到了广泛而成功的应用, 但是 SVMs 只能在有限的样本上进行有监督地学习, 从而导致学习不够充分, 同时该类方法在学习过程中并未充分考虑数据点间的几何结构和数据潜在的判别信息, 因而在一定程度上限制了该类方法在具体模式识别问题上的泛化能力[6].

为了克服 SVM 类方法的上述缺陷, 线性判别分析 (Linear discriminant analysis, LDA)[7] 具备了保持数据的全局数据结构和全局鉴别信息的能力. 但是 LDA 存在小样本问题 (Small sample size, SSS), 且其投影后的结果不能直接用于分类. 为此, Zafeiriou 等[7] 基于 Fisher 线性判别分析 (Fisher LDA, FLDA)[8] 的思想, 提出一种最小类方差 SVM (Minimum class variance SVM, MCVSVM), 其通过类内方差来正则化 SVM, 即在保持类内分布结构最小化的同时最大化类间判别信息. 上述基于数据离散度的方法, 虽然在一定程度上考虑了数据的分布特性和判别信息, 但未能揭示数据的潜在本质几

收稿日期 2010-12-08 录用日期 2011-07-01
Manuscript received December 8, 2010; accepted July 1, 2011
国家自然科学基金 (60975027, 60903100), 宁波市自然科学基金 (2009A610080) 资助
Supported by National Natural Science Foundation of China (60975027, 60903100) and Natural Science Foundation of Ningbo City (2009A610080)
本文责任编委 乔红
Recommended by Associate Editor QIAO Hong
1. 江南大学信息工程学院 无锡 214122 2. 浙江工商职业技术学院信息工程学院 宁波 315012
1. School of Information Engineering, Southern Yangtze University, Wuxi 214122 2. School of Information Engineering, Zhejiang Business Technology Institute, Ningbo 315012

何结构, 尤其是数据的局部几何结构和局部判别信息.

近来所提出的流形学习[9−11] 方法能有效揭示数据点内部所蕴含的局部几何结构, 其中局部保留投影 (Locality preserving projects, LPP)[9] 是一种新颖的线性流形学习方法. LPP 不但可以保持样本间局部几何结构, 而且又可以克服其他流形学习方法难以在新的测试数据上获得低维的投影映射的问题[12], 同时容易被非线性嵌入, 从而发现高维非线性流形结构. 为了充分发挥 LPP 方法长处, Wang 等[13] 将 LPP 和 MCVSVM 相结合, 提出一种最小类内局部保留方差支持向量机 (Minimum class locality preserving variance SVM, MCLPVSVM), 该方法不但继承了传统 SVM 的优点, 还在一定程度上克服了训练不充分和小样本问题, 并且在学习过程中充分考虑了数据的类内几何结构 (或流形信息), 体现了类间的判别信息. 值得一提的是, 文献 [14−16] 通过分析指出, 数据的局部流形信息适于单一流形建模, 对于模式分类问题, 不同类的模式往往处于不同的流形结构, 导致局部保留算法 (如 LPP) 投影后的样本空间出现类信息交叠, 降低了模式分类性能[14−15], LPP 等局部保留方法在保持模式之间的局部结构时, 忽略了局部模式之间的差异 (或判别) 信息, 导致识别性能下降[16]. 从这层含义上来讲, MCLPVSVM 在一定程度上没有充分考虑保持数据局部流形结构和局部差异 (或判别) 信息, 同时, 由文献 [13] 可知, MCLPVSVM 仅考虑了数据类内全局流形结构, 而非局部几何结构, 因此在一定程度上影响了 MCLPVSVM 方法对具体模式进行识别的性能.

针对 MCLPVSVM 方法存在的问题, 本文提出一种新颖的局部保留最大信息差 v-SVM (Locality-preserved maximum information variance v-SVM, v-LPMIVSVM), 对于模式分类问题, v-LPMIVSVM 同时考虑样本空间的局部几何结构和局部差异信息, 通过满足最小局部结构信息 (局部同类离散度) 和最大局部差异信息 (局部异类离散度) 准则来寻求一个最优的模式分割超平面, 从而实现模式最优分割. v-LPMIVSVM 方法的优势在于:

1) 在继承了传统 SVM 方法和 MCLPVSVM 方法特色的同时, 还在一定程度上避免了学习不充分的问题, 并且符合模式分类方法所遵循的类内分布最小和类间间隔最大的准则;

2) 首次将输入空间局部同类离散度和局部异类离散度概念引入到 SVM 中, 在一定程度上不但可以保持数据内在的局部几何机构, 同时还可以在一定程度上保持数据局部的差异信息, 即体现蕴涵于数据间的局部判别信息, 从而克服 LPP 类方法产生的低维嵌入空间中模式交叠现象而导致不能很好地应用于模式分类的问题;

3) 创新性地应用测地线距离而非欧氏距离来度量数据空间的流形距离, 以在一定程度上更好地反映数据流形的本质结构特征, 确保分类模式判别信息最大化;

4) 引入参数 μ, 使之控制分类器的间隔误差的上界和支持向量的下界, 从而在一定程度上增强了分类器的泛化性能.

1 v-LPMIVSVM

1.1 问题描述

为了简单起见, 本文主要考虑二元分类任务, 对于多类分类问题, 采用一对一方法将其转化为多个二元分类问题解决[6]. 对于一个包含 N 个模式的二元分类问题, 设给定数据对集 $T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$, 其中, $\boldsymbol{x}_i \in X \subset \mathbf{R}^d$ 为 d 维输入数据, $y_i \in \{+1, -1\}$ 为类标签 $(1 \le i \le N)$, 将数据集 X 进一步分割为两个不同的类:

$$X_1 = \{\boldsymbol{x}_i | \boldsymbol{x}_i \in X,\ y_i = 1\}$$

$$X_2 = \{\boldsymbol{x}_j | \boldsymbol{x}_j \in X,\ y_j = -1\}$$

设 $N_k(\boldsymbol{x}_i)$ 为数据点 $\boldsymbol{x}_i$ $(1 \le i \le N)$ 的 k 近邻集, G 代表数据集 X 的加权邻接图[16], 其中第 i 个顶点代表数据点 $\boldsymbol{x}_i$, 如果 $\boldsymbol{x}_i \in N_k(\boldsymbol{x}_j)$ 或 $\boldsymbol{x}_j \in N_k(\boldsymbol{x}_i)$ $(1 \le j \le N)$, 则 G 中顶点 i 与 j 相连.

由文献 [1] 知, MCLPVSVM 的软间隔形式为

$$\min_{\boldsymbol{w},\boldsymbol{\xi},b} f = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}} Z_{\boldsymbol{w}} \boldsymbol{w} + C\sum_{i=1}^{N}\xi_i \tag{1}$$

$$\text{s.t.}\quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \ge 1 - \xi_i,$$

$$\xi_i \ge 0,\ \ i = 1, \cdots, N \tag{2}$$

其中, $Z_{\boldsymbol{w}}$ 为类内局部保留离散度矩阵, 其定义为 $Z_{\boldsymbol{w}} = \sum_{k=1}^{2} Z_k$, Z_k $(k = 1, 2)$ 为第 k 类的局部保留离散度矩阵, 且 $Z_k = X_k(D^k - W^k)X_k^{\mathrm{T}}$, $D^k = \sum_j W_{ij}^k$ 为一对角矩阵, W^k 为第 k 类的邻接图权值矩阵, 其中第 i 行, 第 j 列元素定义为

$$W_{ij}^k = \exp\left(-\frac{\|\boldsymbol{x}_i^k - \boldsymbol{x}_j^k\|^2}{t}\right),\ \ i, j = 1, \cdots, N \tag{3}$$

其中, $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ 代表 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间的欧氏距离, $t > 0$ 是热核参数, 可通过交叉验证确定.

由式 (1) 可看出, MCLPVSVM 通过保持类内局部几何结构来优化分割平面的投影方向, 以确保在选择决策超平面时满足最大间隔原则, 同时, 保持

数据内在的局部流形结构. 但式 (1) 也在一定程度上说明了 MCLPVSVM 方法存在的缺陷, 即 MCLPVSVM 在一定程度上没有充分考虑保持样本内在的局部结构判别信息. 另外, 通过对文献 [13] 分析发现, 为了算法实现简单, MCLPVSVM 方法在一定程度上仅考虑了类内全局几何结构, 即令 k 近邻集简单地等于类集合, 而没有充分考虑类内局部几何信息, 尤其是没能充分考虑局部差异 (或判别) 信息, 从而导致 MCLPVSVM 在一定程度上存在 "过学习" 问题. 为此, 本文通过引入局部同类离散度和局部异类离散度概念, 在传统 v-SVM 方法基础上, 提出一种局部保留最大信息差 v-SVM (v-LPMIVSVM), 在选择决策超平面时充分考虑数据局部流形信息和局部差异信息.

1.2 线性 v-LPMIVSVM

根据著名的谱图理论[8], 加权最近邻接图 G 能有效刻画数据流形的局部几何结构, 但是仅一个全局性的加权邻接图 G 在一定程度上不能充分反映数据间的判别结构, 为此, 按照文献 [8] 的做法, 将上述加权邻接图 G 分割为两个互补的加权邻接图: G_o 和 G_e: $G = G_o \cup G_e$, $G_o \cap G_e = \emptyset$, 分别反映局部同类邻接关系和局部异类邻接关系, 它们的权值矩阵分别为 W^o 和 W^e.

为了保持数据空间局部流形结构, 应该使得局部内距离越小的同类数据点间权值越大, 即在数据点的最近邻集内, 与该点距离越近的点表示与其相似度越大, 彼此属于同一类的可能性越大, 从而连接权值也越大, 故 W^o 的第 i 行, 第 j 列元素定义为

$$W_{ij}^o = \begin{cases} \exp\left(-\dfrac{d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{t}\right), \\ \qquad \text{若 } \boldsymbol{x}_i \in N_k(\boldsymbol{x}_j) \quad \text{或} \quad \boldsymbol{x}_j \in N_k(\boldsymbol{x}_i) \\ \qquad\quad \text{且 } \boldsymbol{x}_i \text{ 与 } \boldsymbol{x}_j \text{ 同类} \\ 0, \quad \text{其他} \end{cases} \tag{4}$$

其中, $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 代表 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 间距离度量, $i, j = 1, \cdots, N$.

而对于局部内那些属于异类的数据点, 其距离越小, 应使其权值越小, 或者说距离越大的点权值越大, 表示其与近邻集的差异性越大, 通过增加其权值, 使得异类数据点间保持最大间隔, 故 W^e 的第 i 行, 第 j 列元素定义为

$$W_{ij}^e = \begin{cases} \exp\left(-\dfrac{h}{d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}\right), \\ \qquad \text{若 } \boldsymbol{x}_i \in N_k(\boldsymbol{x}_j) \quad \text{或} \quad \boldsymbol{x}_j \in N_k(\boldsymbol{x}_i) \\ \qquad\quad \text{且 } \boldsymbol{x}_i \text{ 与 } \boldsymbol{x}_j \text{ 异类} \\ 0, \quad \text{其他} \end{cases} \tag{5}$$

其中, $h > 0$ 是一可调参数.

值得说明的是, 本文方法采用测地线距离来度量数据对间的距离, 即式 (4) 和式 (5) 中的 $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 代表测地线距离, 进一步的说明在第 2 节中讨论.

定义 1 (局部离散度矩阵). 设 L_o, L_e 分别为图 G_o 和 G_e 的 Laplacian 矩阵, 则矩阵 $H_o = XL_oX^{\mathrm{T}} = X(D^o - W^o)X^{\mathrm{T}}$ 称为局部同类离散度矩阵, 矩阵 $H_e = XL_eX^{\mathrm{T}} = X(D^e - W^e)X^{\mathrm{T}}$ 称为局部异类离散度矩阵, D 为一对角矩阵, 其中, $D^{(\cdot)}$ 为对角矩阵, 且 $D_{ii}{}^{(\cdot)} = \sum_j W_{ij}{}^{(\cdot)}$. H_o 和 H_e 统称为局部离散度矩阵.

上述定义中, 局部同类离散度矩阵和局部异类离散度矩阵分别体现了输入空间数据的局部流形结构信息和局部差异 (或判别) 信息.

定义 2 (局部保留最大信息差矩阵). 矩阵 $M = \lambda H_o - (1-\lambda)H_e$ $(0 < \lambda \le 1)$ 称为局部保留最大信息差矩阵, λ 为一非负常量.

上述定义中, 参数 λ 起平衡局部流形结构和局部判别信息的作用, 当 λ 增大时, 偏向于保持局部流形结构信息, 反之, 则加大惩罚局部差异 (或判别) 信息. 从而, 在适当的 λ 值下, 本文方法既能较好地保持局部流形结构, 又能保持较强的模式判别信息.

综上, 线性 v-LPMIVSVM 方法的原始优化问题描述为

$$\min_{\boldsymbol{w},\rho,\boldsymbol{\xi},b} f = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}M\boldsymbol{w} - \mu\rho + C\sum_{i=1}^{N}\xi_i \tag{6}$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \ge \rho - \xi_i, \ \ i = 1, \cdots, N \tag{7}$$

$$\xi_i \ge 0, \ \ \mu \ge 0, \ \ \rho \ge 0 \tag{8}$$

其中, ρ 为类间最小间隔, C 为一正则化常量, $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_N]$ 为松弛向量.

v-LPMIVSVM 具有和 v-SVM 相似的原始优化问题形式, 按照文献 [5] 中方法的对偶推导原理, 有如下结论:

定理 1. 线性 v-LPMIVSVM 方法原始优化问题 (6)∼(8) 的对偶问题为

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}H\boldsymbol{\alpha} \tag{9}$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, \ \ i = 1, \cdots, N \tag{10}$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{11}$$

$$\sum_{i=1}^{N}\alpha_i \geq \mu \tag{12}$$

其中, $H=(h_{ij})_{N\times N}$, $h_{ij}=y_iy_j\boldsymbol{x}_i^{\mathrm{T}}M^{-1}\boldsymbol{x}_j$, M^{-1} 为矩阵 M 的逆运算, 且 v-LPMIVSVM 原始问题中投影向量 $\boldsymbol{w}$ 和偏置变量 b 分别为

$$\boldsymbol{w}=\sum_{i=1}^{N}\alpha_i y_i M^{-1}\boldsymbol{x}_i \tag{13}$$

$$b=-\frac{1}{2}\left(\frac{1}{|S_+|}\sum_{\boldsymbol{x}\in S_+}\sum_{j=1}^{N}\alpha_j y_j \boldsymbol{x}_j^{\mathrm{T}}M^{-1}\boldsymbol{x}+\frac{1}{|S_-|}\sum_{\boldsymbol{x}\in S_-}\sum_{j=1}^{N}\alpha_j y_j \boldsymbol{x}_j^{\mathrm{T}}M^{-1}\boldsymbol{x}\right) \tag{14}$$

其中, $S_\pm=\{\boldsymbol{x}_i|0\leq\alpha_i\leq C, y_i=\pm1\}$, $N_{\mathrm{sv}}=|S_+|+|S_-|$, $|\cdot|$ 表示集合基数.

算法 1 (线性 v-LPMIVSVM 算法).

输入. N 个数据集 $X=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$, $\boldsymbol{x}_i\in\mathbf{R}^d$ 代表一个 d 维数据模式.

输出. 分类决策函数 $g(\boldsymbol{x})$.

步骤 1. 构建数据集 X 的 k 近邻加权连接图 G;

步骤 2. 根据 X 的 k 近邻加权连接图 G, 计算测地线距离矩阵 $D=\{d_G(i,j)\}$, $d_G(i,j)$ 代表数据点 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 间测地线距离;

步骤 3. 根据式 (4) 和式 (5), 分别计算连接边权值矩阵 W^o 或 W^e;

步骤 4. 根据定义 1 分别计算 H_o 和 H_e, 再根据定义 2 计算 M;

步骤 5. 选择参数 t, h, C, k, λ, μ. 根据定理 1 求解 Lagrange 乘子向量 α, 再根据式 (13) 和式 (14) 分别计算决策超平面法向量 $\boldsymbol{w}$ 和偏置变量 b;

步骤 6. 输出分类决策函数 $g(\boldsymbol{x})=\mathrm{sgn}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)$.

需要说明的是, 在算法 1 的步骤 4 中, 矩阵 M 可能为奇异矩阵, 即所谓的小样本问题[17], 从而 M 不可逆, 对此, 处理方法较多, 本文按照文献 [18] 中的方法, 利用主成分分析 (Principal component analysis, PCA)[6] 方法将输入空间数据进行降维, 从而使得原始问题转化为一个在低维空间的等价优化问题, 避免了局部保留最大信息矩阵 M 的奇异问题, 从而使矩阵 M 可逆.

从算法 1 可看出, 由于要计算局部保留最大信息差矩阵 M, v-LPMIVSVM 方法与传统的 v-SVM 方法相比具有较高的空间复杂度 ($\mathrm{O}(d^2)$) 和时间复杂度 ($\mathrm{O}(d^3)$), 特别是在处理高维小样本数据时尤为明显, 为了在一定程度上提高本文方法的执行效率, 在训练高维数据时, 首先采用 PCA 方法对数据进行相应的预处理, 以提高所提方法的执行效率.

1.3 非线性扩展

为了处理高维非线性流形数据分类情况, 本文提出非线性 v-LPMIVSVM, 即采用核映射技术 (Kernel trick)[18], 引入一个非线性映射 ϕ, 将输入空间映射到高维甚至无限维特征空间 H 中, 实现模式线性可分, 高维特征空间的线性超平面对应原始输入空间的非线性超平面. 在 H 空间中, 两个向量 $\phi(\boldsymbol{x}_i)$, $\phi(\boldsymbol{x}_j)$ 的内积可利用一个满足 Mercer 条件的核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j)$ 来表示计算.

定理 2. 非线性 v-LPMIVSVM 的原始优化问题为

$$\min_{\boldsymbol{w},\rho,\boldsymbol{\xi},b} f=\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}KLK^{\mathrm{T}}\boldsymbol{\beta}-\mu\rho+C\sum_{i=1}^{N}\xi_i \tag{15}$$

$$\text{s.t.}\quad y_i\left(\sum_{j=1}^{N}\beta_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)+b\right)\geq\rho-\xi_i,\quad i=1,\cdots,N \tag{16}$$

其中, $\xi_i\geq0$, $\mu\geq0$, $\rho\geq0$, Mercer 核函数 $K=(K_{ij})$, $K_{ij}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)$, $L=\lambda L_o-(1-\lambda)L_e$.

证明. 考虑一个非线性映射函数 $\phi:\boldsymbol{x}\rightarrow\phi(\boldsymbol{x})$, 将原始空间中的数据点 $\boldsymbol{x}$ 映射到特征空间 H 中的点 $\phi(\boldsymbol{x})$. 则原始空间数据矩阵在特征空间的表示为 $X^\phi=\{\phi(\boldsymbol{x}_i)\}_{i=1}^N$. 通过对线性 v-LPMIVSVM 方法的决策超平面法向量 $\boldsymbol{w}$ 的分析得知, 法向量 $\boldsymbol{w}$ 与正类样本和负类样本有关, 并结合 Representer Theorems[19], 可以将特征空间中非线性决策平面法向量表示为: $\boldsymbol{w}^\phi=\sum_{i=1}^N\beta_i\phi(\boldsymbol{x}_i)$, 其中, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_N)^{\mathrm{T}}$ 表示权值向量, 则在特征空间中:

$$\boldsymbol{w}^{\phi^{\mathrm{T}}}M^\phi\boldsymbol{w}^\phi=\boldsymbol{\beta}^{\mathrm{T}}X^{\phi^{\mathrm{T}}}X^\phi(\lambda L_o-(1-\lambda)L_e)X^{\phi^{\mathrm{T}}}X^\phi\boldsymbol{\beta}=\boldsymbol{\beta}^{\mathrm{T}}KLK^{\mathrm{T}}\boldsymbol{\beta} \tag{17}$$

其中, M^ϕ 为特征空间局部保留最大信息差矩阵, K 为 $N\times N$ 核矩阵, $L=\lambda L_o-(1-\lambda)L_e$, 结合式 (17) 和式 (6)$\sim$(8), 定理成立. □

定理 3. 非线性 v-LPMIVSVM 方法的原始优化问题 (15) 和 (16) 的对偶问题为

$$\min_{\boldsymbol{\alpha}}\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}H^\phi\boldsymbol{\alpha} \tag{18}$$

$$\text{s.t.}\quad 0 \le \alpha_i \le C,\ \ i = 1, \cdots, N \tag{19}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{20}$$

$$\sum_{i=1}^{N} \alpha_i \ge \mu \tag{21}$$

其中, $H^{\phi} = YK^{\mathrm{T}}(KLK^{\mathrm{T}})^{-1}KY$, $Y = \mathrm{diag}\{y_1, \cdots, y_N\}$.

证明. 式 (15) 和式 (16) 对应的 Lagrange 函数为

$$\begin{aligned} L(\boldsymbol{w}, \rho, \boldsymbol{\xi}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \gamma) = \\ \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}} KLK^{\mathrm{T}}\boldsymbol{\beta} - \mu\rho + C\sum_{i=1}^{N}\xi_i - \\ \sum_{i=1}^{N}\alpha_i\left[y_i\left(\sum_{j=1}^{N}\beta_j K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b\right) - \rho + \xi_i\right] - \\ \sum_{i=1}^{N}\eta_i\xi_i - \gamma\rho \end{aligned} \tag{22}$$

其中, $\alpha_i \ge 0$, $\eta_i \ge 0$, $\gamma \ge 0$ 分别为 Lagrange 乘子系数. 根据 KKT 条件:

$$\frac{\partial L}{\partial \rho} = 0 \to \sum_{i=1}^{N}\alpha_i = \mu + \gamma \ge \mu \tag{23}$$

$$\frac{\partial L}{\partial b} = 0 \to \sum_{i=1}^{N}\alpha_i y_i = 0,\ \ i = 1, \cdots, N \tag{24}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \to 0 \le \alpha_i \le C,\ \ i = 1, \cdots, N \tag{25}$$

将式 (23) ∼ (25) 代入式 (22), 得到新的 Lagrange 函数:

$$\begin{aligned} L(\boldsymbol{w}, \rho, \boldsymbol{\xi}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \gamma) = \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}} KLK^{\mathrm{T}}\boldsymbol{\beta} - \\ \sum_{i=1}^{N}\alpha_i\left[y_i\left(\sum_{j=1}^{N}\beta_j K(\boldsymbol{x}_j, \boldsymbol{x}_i)\right)\right] \end{aligned} \tag{26}$$

对式 (26) 中 $\boldsymbol{\beta}$ 求偏导数为

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0 \to \boldsymbol{\beta} = (KLK^{\mathrm{T}})^{-1}KY\boldsymbol{\alpha} \tag{27}$$

其中, $Y = \mathrm{diag}\{y_1, \cdots, y_N\}$, 将式 (27) 代入式 (26), 并结合式 (23) ∼ (25), 定理得证. □

根据定理 2 和定理 3 以及 KKT 条件可得核 v-LPMIVSVM 方法偏置变量 b^{ϕ} 为

$$\begin{aligned} b^{\phi} = -\frac{1}{2}\left(\frac{1}{|S_+|}\sum_{\boldsymbol{x}\in S_+}\sum_{j=1}^{N}\beta_j^K(\boldsymbol{x}_j, \boldsymbol{x}) + \right. \\ \left. \frac{1}{|S_-|}\sum_{\boldsymbol{x}\in S_-}\sum_{j=1}^{N}\beta_j^K(\boldsymbol{x}_j, \boldsymbol{x})\right) \end{aligned} \tag{28}$$

由上可得, 对于某个测试样本 $\boldsymbol{x}$, 非线性 v-LPMIVSVM 的决策函数为

$$g^{\phi}(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N}\beta_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^{\phi}\right)$$

需要说明的是, 模式分类方法一般遵从所谓的大间隔准则, 即: 1) 同类的数据点尽量内聚; 2) 异类的数据点尽量相互远离. 由于本文方法引入了局部同类离散度且使其最小化, 从而使得局部数据点在决策超平面上的投影更加紧致, 这一点显然符合准则 1); 另外, 本文方法还引入了局部异类离散度且使其最大化, 从而确保满足准则 2). 故此, 本文所提方法是一种比较合理的大间隔分类学习方法.

2 讨论

2.1 数据间距离度量

定义 3 (测地线距离 (Geodesic distance)). 数据空间中任意点对 $\boldsymbol{x}_i, \boldsymbol{x}_j$ 在流形 M 上测地线距离 $d_M(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 定义为连接该数据点对的最短曲线 (即最短路径) 的长度.

在 MCLPVSVM, UDP, LPMIP 等方法中, 数据空间邻接图的权值矩阵计算是基于欧氏距离度量, 而文献 [14] 经过分析指出, 在高维数据空间中, 欧氏距离在一定程度上不能很好地反映数据流形的本质几何信息, 而测地线距离能在一定程度上真正反映数据流形的本质特征. 因此, 本文对数据空间的样本距离度量采用测地线距离度量. 根据定义 3, 数据空间内两个样本间测地线距离可通过最短路径算法求得, 数据点对间最短路径的计算算法参见文献 [10], 在此不再赘述.

2.2 参数 μ 属性

按照文献 [4] 中术语, 本文称对应于 Lagrange 乘子系数 $\alpha_i > 0$ 的训练样本为支持向量 (Support vector, SV), 对应于松弛变量 $\xi_i > 0$ 的训练样本称为间隔误差 (Margin error, ME), 则有如下定理:

定理 4. 设在数据集 $X \in \mathbf{R}^d$ 上运行 v-LPMIVSVM 算法, 并获得相应的 Lagrange 乘子向量 $\boldsymbol{\alpha}$ 和优化变量 $\boldsymbol{w}^*$, ρ^*, b^* 的解, 则:

1) 参数 μ 是 v-LPMIVSVM 方法间隔误差的上界;

2) 参数 μ 是 v-LPMIVSVM 方法支持向量的下界.

定理 4 说明 v-LPMIVSVM 中参数 μ 具有和 v-SVM 中参数 v 同样的性质.

2.3 泛化性能分析

Vapnik 等在统计学习理论 (Statistical learning theory, SLT)[2] 中基于 VC 维 (Vapnik Chervonenkis dimension) 理论所提出的结构风险最小化 (Structure risk minimization, SRM) 思想, 为基于数据的机器学习提供了一种优良的归纳推理准则, 基于该准则设计的学习机在一定程度上具有好的泛化能力. VC 维理论同时提供了一个可分析的泛化误差界以估计学习机的泛化误差. 文献 [5] 在一定的分析推导的基础上提出了 v-SVM 学习机的泛化误差界 (文献 [5] 中定理 10), 本文据此提出非线性 v-LPMIVSVM 的泛化误差界模型.

定理 5 (v-LPMIVSVM 泛化误差界). 设 H 为一再生核希尔伯特空间 (Reproducing kernel Hilbert space, RKHS), 核 v-LPMIVSVM 所采用的核形式为: $K(\boldsymbol{x},\boldsymbol{y})=K(\|\boldsymbol{x}-\boldsymbol{y}\|)$, 且满足 $K(0)=1$, 则随机概率 $P(\boldsymbol{x})$ 产生的 N 个数据点 $T=\{(\boldsymbol{x}_i)\}_{i=1}^{N}$ 位于 H 空间的圆心在原点的单位球上, 从而, 核 v-LPMIVSVM 方法的学习函数 $g^{\phi}\in H$ 的泛化误差上界在概率 $1-\eta$ 下满足下式:

$$R(g^{\phi}) \underset{P}{\leq} R_{\text{emp}}^{\rho}(g^{\phi}) + C\sqrt{\frac{4c^2{\boldsymbol{w}^{\phi}}^{\mathrm{T}}M^{\phi}\boldsymbol{w}^{\phi}}{\rho^2}\log_2\left(\frac{2}{C}\right)-1+\ln\left(\frac{2}{\eta}\right)} \leq \mu + C\sqrt{\frac{4c^2{\boldsymbol{w}^{\phi}}^{\mathrm{T}}M^{\phi}\boldsymbol{w}^{\phi}}{\rho^2}\log_2\left(\frac{2}{C}\right)-1+\ln\left(\frac{2}{\eta}\right)} \tag{29}$$

$$M^{\phi}=X^{\phi}(\lambda L_o-(1-\lambda)L_e){X^{\phi}}^{\mathrm{T}} \tag{30}$$

其中, $R_{\text{emp}}^{\rho}(\cdot)$ 为与间隔 ρ 相关的经验风险函数, C 为正则化常量, 用于平衡学习机的结构复杂度和经验风险 R_{emp}, $c<103$ 为一常量[5], 式 (29) 中第 2 个不等式由定理 4 推论可得.

由定理 5 分析可知, 为了学习优化的决策函数, 需使得 $\frac{\boldsymbol{w}^{\mathrm{T}}M^{\phi}\boldsymbol{w}}{\rho^2}$ 最小化, 本文方法的优化形式显然满足该要求. 另外, 根据定理 5, 有两点值得说明: 1) 核 v-LPMIVSVM 方法的泛化误差界基于 VC 理论导出, 且在一定程度上充分考虑了数据空间局部内几何结构和判别信息; 2) 核 v-LPMIVSVM 方法的泛化误差界可以通过参数 t, h, k, C, μ 和 λ 进行调节控制, 从而有望取得相较传统方法更优的泛化性能.

3 实验分析

为了说明本文方法的有效性, 将 v-LPMIVSVM 方法分别在人造数据集 (流形结构数据集 two-moons), 真实数据集 (UCI 数据集[4] 和人脸识别[16−17, 20] 数据集) 上进行测试, 并与相关的方法 v-SVM, MCVSVM 和 MCLPVSVM 进行比较, 以显示本文方法的学习泛化能力. 通过测试人造数据来说明本文方法在抉择分类函数过程所依据的基本原理和方法以及参数的影响; 测试真实数据集主要说明本文方法在同时保持数据局部几何结构和判别信息情况下的模式分类性能.

所有实验样本首先归一化为 $[-1,1]$. 对于非线性映射采用径向基函数 (Radial basis function, RBF): $K(\boldsymbol{x},\boldsymbol{y})=\exp(-\frac{1}{\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|^2)$, 其中 γ 值在集合 $\{\sigma^2/8,\ \sigma^2/4,\ \sigma^2/2,\ \sigma^2,\ 2\sigma^2,\ 4\sigma^2\}$ 中选取[4], 其中 σ 为训练样本平均范数的平方根; 方法 v-SVM, MCVSVM, MCLPVSVM 的参数确定将采取文献 [18] 中相同的策略, 即根据最好的交叉验证参数集的平均值来确定优化的实验参数.

3.1 人造数据实验

数据集 two-moons 经常被用于测试一些流形学习方法[11]. 本节将通过与 v-SVM 和 MCLPVSVM 方法进行比较, 测试本文方法在两种不同复杂度 two-moons 数据集上保持非线性局部流形结构和判别信息的性能.

两种 two-moons 数据集大小均为 200, 其中正类的数据数为 101, 负类数据数为 99, 随机抽取训练集和测试集, 重复 10 次, 分别记录实验结果, 且将实验结果的平均值记录于表 1. 三种方法的分类决策边界如图 1 所示.

表 1 三种方法的分类精度比较 (%)

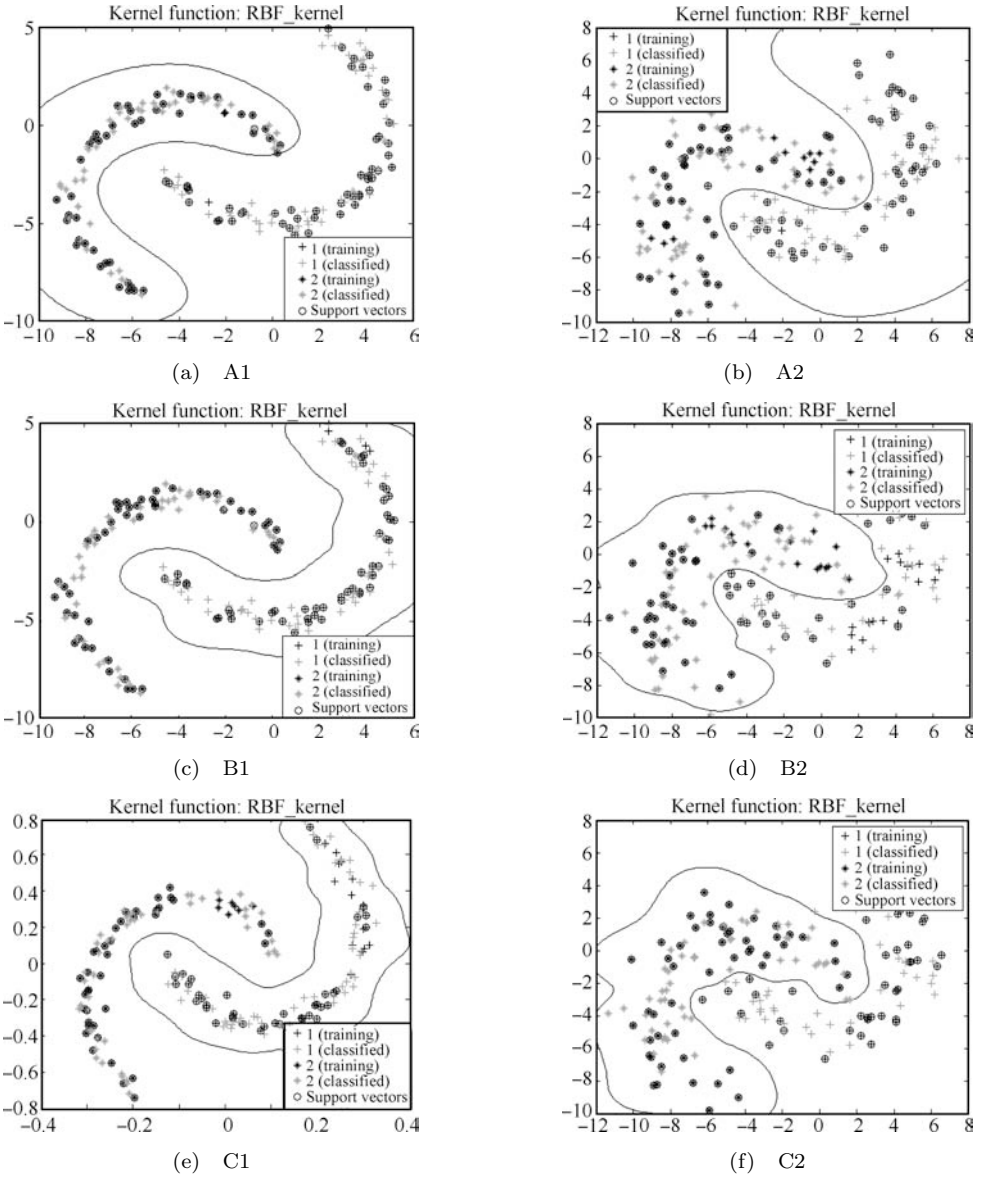
Table 1 Classification accuracy comparison of v-SVM, MCLPVSVM, and v-LPMIVSVM (%)

方法	精度 1 (数据集)	精度 2 (数据集)
v-SVM	100 (A1)	96.7 (A2)
MCLPVSVM	100 (B1)	98.6 (B2)
v-LPMIVSVM	100 (C1)	99.8 (C2)

从图 1 和表 1 结果可以看出: 1) 图 1 (a) 和图 1 (b) 显示, 传统的 v-SVM 方法在两种数据集上的分类边界曲线均较其他两种方法平滑. 但是, 随着数据集的拓扑结构变得更加复杂而不规则 (见图 1 所示的数据结构), v-SVM 方法的分类精度相较其他两种方法下降明显, 特别是在图 1 (b) 所示数据集上的

局部决策边界似乎显示过于平滑, 从而导致 v-SVM 未能充分考虑局部流形结构信息; 2) MCLPVSVM 和本文方法在两种 two-moons 数据集上的决策边界均显示局部不平滑性 (如图 1 (c) ∼ 1 (f), 且随着数据集拓扑结构的复杂度增加, 这种不平滑性更明显, 但是二者分类性能均优于 v-SVM 方法, 这也说明关注数据的局部流形结构信息有利于模式分类性能的增强; 3) 由于在一定程度上仅考虑了局部流形结构信息, 当数据集局部数据交叠严重时 (如图 1 (d)), 导致 MCLPVSVM 方法在一定程度上不能有效识别, 从而分类性能下降. 而本文方法由于充分考虑了局部流形结构和差异信息, 这就导致本文方法在寻找决策超平面时, 不但要考虑保持数据潜在的局部流形结构, 同时还要充分考虑局部差异信息最大化, 即充分保持数据潜在的局部几何结构和局部判别信息, 而这一点有别于现有分类方法.

为了进一步评价参数 μ, k, λ 对所提方法的性能影响, 抽取上述 two-moons 实验中针对图 1 (f) 数据集进行 10 次实验的 μ, k, λ 参数值, 并在其他参数固定的情况下, 记录参数 μ, k, λ 值变化分别对本文方法的性能影响如表 2 所示. 从表 2 可看出, 参数 k, μ, λ 的选取对本文方法的性能影响较大, 对于本实验, 最优的参数值对为: $k = 5$, $\mu = 0.5$, $\lambda = 0.8$. 可见, 通过调节参数 μ, k, λ 的不同取值, 可以明显改



(a) A1　(b) A2

(c) B1　(d) B2

(e) C1　(f) C2

图 1　在两种 two-moons 数据集上的决策边界: v-SVM ((a) ∼ (b)), MCLPVSVM ((c) ∼ (d)), v-LPMIVSVM ((e) ∼ (f))

Fig. 1　The discriminant boundaries in the two two-moon datasets: v-SVM ((a) ∼ (b)), MCLPVSVM ((c) ∼ (d)), v-LPMIVSVM ((e) ∼ (f))

表 2 v-LPMIVSVM 在不同参数值下的分类结果 (%)

Table 2 The classification accurancies of v-LPMIVSVM based on different parameters (%)

参数 μ 变化的分类精度影响 ($\lambda = 0.8$, $k = 5$), 其中 "−" 代表无优化解								
μ	0.01	0.05	0.1	0.3	0.5	0.7	0.8	1
分类精度	23.43	46.8	61.4	93	99.8	87.68	−	−

参数 λ 变化的分类精度影响 ($\mu = 0.5$, $k = 5$)								
λ	0.1	0.3	0.4	0.5	0.6	0.8	0.9	1
分类精度	97.6	87	98	62	87	99.8	88	82.4

参数 k 变化的分类精度影响 ($\mu = 0.5$, $\lambda = 0.8$)								
k	3	5	7	9	10	11	13	15
分类精度	93.18	99.8	97.67	74.54	87	93.87	67.7	82.6

进本文方法的分类性能, 这也进一步验证了定理 5 的结论.

3.2 实际数据集

为了更全面地说明本文方法的分类性能, 在本节分别测试 UCI 数据集, 人脸数据集 (ORL 数据集, Yale 数据集), 同时与三种现有方法 (v-SVM, MCVSVM, MCLPVSVM) 进行比较, 以进一步说明本文算法的有效性.

3.2.1 UCI 数据集实验

UCI 数据集经常用来测试模式分类方法的性能[4, 12−13]. 本文综合文献 [4, 13, 16] 中所采用的 UCI 数据集, 选取 14 个典型数据集进行实验, 分别测试线性, 非线性情况下的四种方法 (v-SVM, MCVSVM, MCLPVSVM, v-LPMIVSVM) 的分类性能.

数据集详细信息如表 3 所示. 数据集中同时包括二类和多类, 对于多类分类采取 one-against-one 策略进行, 实验前所有数据集进行了归一泛化处理. 评价所有分类方法的泛化性能的标准是基于数据集总体样本的 5 重交叉验证精度, 在 5 重交叉验证测试中, 数据集被随机划分为 5 个子集, 每次验证取这 5 个子集中的 1 份作为测试集, 其他作为训练集, 该过程重复 5 次, 取其平均值作为最后的实验结果.

本文记录 5 重交叉验证的平均精度, 表 4 显示在线性和非线性情况下所提方法 v-LPMIVSVM 与 v-SVM, MCVSVM 及 MCLPVSVM 方法在给定数据集上的模式分类精度. 需指出的是, 这里仅记录了在最优参数下的实验结果值.

从表 4 结果可得出如下结论:

1) 采用核方法能明显增强四种分类方法在几乎所有数据集上的分类性能. 2) 结果显示, 在线性或非线性情况下, 传统的 v-SVM 方法在所有数据集上的分类效果在一定程度上均逊于其他三种方法. 这也验证了仅仅考虑数据整体类间的最大间隔, 而忽视数据空间的结构信息尤其是保持局部流形信息, 一定程度上不能取得较优的分类性能. 而在考虑了数据空间的本质流形结构的情况下, MCLPVSVM 和 v-LPMIVSVM 均表现出明显的泛化能力优势. 3) 与 v-SVM, MCVSVM, MCLPVSVM 等方法相比, 所提方法对于所有数据集均具有优于或可比较的模式分类性能, 这进一步说明, 在充分考虑数据空间的局部几何结构信息和局部判别信息的情况下, v-LPMIVSVM 方法能取得较好的泛化性能.

表 3 模式分类 UCI 数据集

Table 3 The UCI datasets for pattern recognition

Dataset	模式数	特征数	类数
Iris	150	4	3
Breast	699	9	2
Heart	303	13	2
Vehicle	846	8	4
Glass	214	9	6
Ionosphere	351	34	2
Wine	178	13	4
Waveform	900	21	3
Balance-scale	625	4	3
Sonar mines	208	60	2
Hepatitis	155	19	2
Biomed	194	5	2
Diabetes	768	8	2
Liver	345	6	2

3.2.2 人脸识别实验

人脸图像数据集呈现出明显的非线性流形结构[10, 12], 因此, 被许多流形学习方法用于测试数据集[10, 12, 17−20], 以反映流形学习方法的有效性. 文献 [10] 分析指出, 大量研究成果证实, 图像模式是嵌入

表 4 四种方法在 UCI 数据集上的分类结果 (%)

Table 4 The classification accurancies of the four approaches based on UCI datasets (%)

Dataset	v-SVM		MCVSVM		MCLPVSVM		v-LPMIVSVM	
	线性	非线性	线性	非线性	线性	非线性	线性	非线性
Iris	97.3	**100**	98.7	**100**	98.6	**100**	**100**	**100**
Breast	97.12	99.67	96.94	99.54	**97.2**	**99.88**	96.71	99.87
Heart	79.9	84.8	83.4	88.94	84.3	91.13	**89.12**	**94.28**
Vehicle	76.7	78.76	80.5	87.42	82.1	**88.3**	**85.7**	**88.3**
Glass	63.2	68.4	60.5	70.3	63.57	70.41	**70.85**	**72.29**
Ionosphere	87.67	84.28	83.6	85.30	89.8	90. 33	**90.71**	**91.42**
Wine	93.2	98.4	93.48	98.2	95.51	98.61	**97.89**	**100**
Wave form	88.7	90.46	86.44	89.74	85.17	90.87	**89.3**	**90.9**
Balance scale	**94.2**	96.51	93.9	94.89	93.67	95.27	**94.2**	**97.2**
Sonar mines	69.9	69.9	75.45	75.45	76.26	76.26	**80.21**	79.6
Hepatitis	84.41	**88.81**	87.01	86.2	85.46	88.86	**87.66**	88.04
Biomed	86.9	86.9	87.5	87.5	90.82	90.82	**91.67**	**93.37**
Diabetes	**76.2**	**83.4**	75.26	86.1	75.67	79.3	75.8	81.4
Liver	65.7	67.18	71.5	74.68	71.84	77.92	**74.27**	**79.87**

在某个高维空间的低维流形, 对于图像模式分类问题, 考虑数据空间的本质流形结构, 尤其是同时考虑数据的局部内本质流形结构和局部判别信息, 有望能在一定程度上提升图像模式识别性能. 为此, 本部分主要测试所提方法在处理非线性流形结构的人脸数据集上所具有的有效性, 分别采用两种距离度量方式 (欧氏距离和测地线距离) 对所提方法进行测试 (为区别起见, 本文称欧氏距离度量的 v-LPMIVSVM 为 v-eLPMIVSVM), 以进一步显示本文方法所采用的度量方式的有效性.

实验采用 ORL 数据库和 Yale 数据库[20] 作为测试数据. 其中 ORL 人脸数据集包含 40 个对象的人脸数据, 每类对象由不同表情的 10 张图片构成. Yale 人脸数据集包括 15 个类别的 165 幅灰度级图像, 同一类由不同光照条件和脸部表情的 11 张人脸数据组成. 实验前, 对上述图像集进行预处理, 使其缩放到 32 像素 × 32 像素大小, 且每个像素为 256 灰度级, 则在图像空间, 每张图像由一个 1 024 维向量表示, 更多信息可参考文献 [20].

由于人脸图像是一种典型的小样本高维数据集, 为了降低算法运行过程中由于矩阵奇异性引起的计算误差, 本文按照文献 [13] 中做法, 首先利用 PCA 对上述 2 个数据集进行特征降维, 为了对各方法进行有效比较, 实验前将数据降维到同样维度. 文献 [12] 经测试分析指出, 对于一个 1 024 维的图像数据, 降维到 9 维时能保持原始图像的 90 %以上的信息量. 据此, 本文在实验前先利用 PCA 将实验数据集进行特征降维到 9 维, 且每个数据集所选取的训练数据数大于数据维数, 这样能降低在矩阵求逆过程中所产生的计算误差. 同时, 由于本段各实验数据子集的训练样本数均相对较少, 故测试中采取留一法交叉验证.

在 ORL 人脸数据集中随机抽取 5 个 2 分类的数据子集, 即 ORL.1-2 (指 ORL 库中第 1 类和第 2 类, 以下类同), ORL.3-5, ORL.6-10, ORL.20-32, ORL.18-40, 对于每个数据子集, 随机选取 10 个数据作为训练集, 其余作为测试数据集. 实验结果取自 10 重交叉验证实验的平均值. 在 Yale 人脸数据集中随机抽取 3 个 2 分类的数据子集, 即 Yale.3-5, Yale.6-10, Yale.8-12, 在每个数据子集中分别对每类对象随机抽取 $p = \{2, 3, 5\}$ 个图像数据作为训练子集, 其余数据作为测试子集, 实验中以符号 P_m/T_n 表示每个数据子集中每类对象抽取 m 个图像用于训练, n 个图像用于测试. 所有数据子集的实验结果分别取自 10 重交叉验证的平均值.

图 2 和图 3 分别显示了参数 λ 和 k 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. 另外, 表 5 记录了 ORL 数据集在不同二分类数据集下的实验结果, 表 6 ∼ 8 分别记录了 Yale 数据库中 3 个随机 2 分类数据集在抽取类对象不同训练图像数据下的实验结果.

从表 5 ∼ 8, 图 2 和图 3 的结果可得如下几点结论:

1) v-SVM 方法在大多数二分类图像数据集上的识别性能均低于其他几种方法, 这说明, 对于图形识别这类具有非线性流形结构的数据集, 仅考虑数据类间的最大间隔在一定程度上已不能很好地满足模式识别性能的要求. 而考虑了数据类内分布结构的

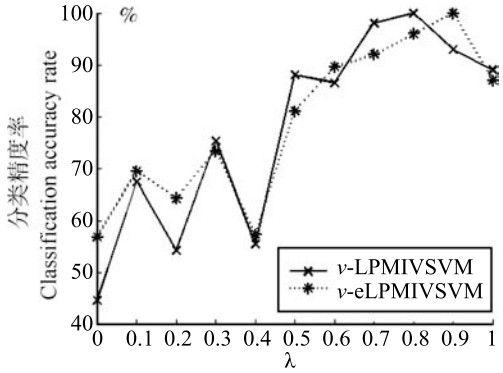


图 2 参数 λ 影响 (ORL.1-2)

Fig. 2 The performance influence of v-LPMIVSVM based on different λ

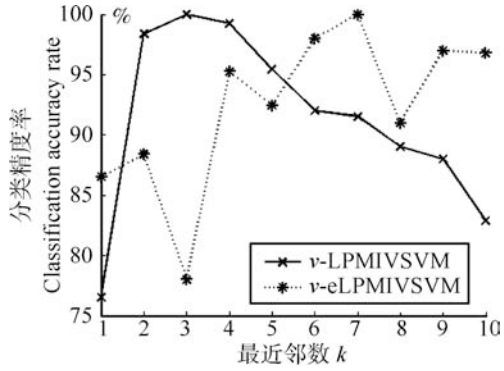


图 3 参数 k 影响 (ORL.1-2)

Fig. 3 The performance influence of v-LPMIVSVM based on different k

表 5 ORL 数据库实验结果 (%)

Table 5 The experimental results for ORL datasets (%)

Dataset		v-SVM	MCVSVM	MCLPVSVM	v-eLPMIVSVM	v-LPMIVSVM
ORL	ORL.1-2	85.29	96.22	99.54	99.46	100
	ORL.3-5	88.82	85.76	**90.34**	87.81	90.17
	ORL.6-10	90	87.47	88.33	88.76	**91.24**
	ORL.20-32	**100**	98.93	**100**	**100**	**100**
	ORL.18-40	96.5	95.10	**98.72**	98.0	98.70

表 6 Yale 数据库 P_2/T_9 实验结果 (%)

Table 6 The experimental results for Yale datasets P_2/T_9 (%)

Dataset		v-SVM	MCVSVM	MCLPVSVM	v-eLPMIVSVM	v-LPMIVSVM
Yale	Yale.3-5	60.0	64.30	**79.10**	79.02	78.68
	Yale.6-10	90.0	88.61	91.11	87.01	**91.57**
	Yale.8-12	80.0	77.24	85.51	85.76	**88.14**

表 7 Yale 数据库 P_3/T_8 实验结果 (%)

Table 7 The experimental results for Yale datasets P_3/T_8 (%)

Dataset		v-SVM	MCVSVM	MCLPVSVM	v-eLPMIVSVM	v-LPMIVSVM
Yale	Yale.3-5	62.74	71.04	79.42	79.27	**80.18**
	Yale.6-10	88.33	87.28	**91.76**	88.22	**91.76**
	Yale.8-12	78.50	80.53	82.01	81.76	**82.93**

表 8 Yale 数据库 P_5/T_6 实验结果 (%)

Table 8 The experimental results for Yale datasets P_5/T_6 (%)

Dataset		v-SVM	MCVSVM	MCLPVSVM	v-eLPMIVSVM	v-LPMIVSVM
Yale	Yale.3-5	58.89	78.29	79.57	79.48	**80.35**
	Yale.6-10	76.58	87.88	92.23	91.94	**92.46**
	Yale.8-12	79.00	81.00	87.21	86.80	**88.03**

MCVSVM, MCLPVSVM, v-LPMIVSVM (v-eLPMIVSVM) 方法均取得了相对较好的识别性能, 尤其是在考虑了数据局部流形结构的 MCLPVSVM, v-LPMIVSVM (v-eLPMIVSVM) 方法的识别性能明显优于仅考虑大间隔的 v-SVM 和考虑类内离散度的 MCVSVM 方法. 尤其值得说明的是, 本文方法在充分考虑数据空间局部流形结构和局部判别信息的情况下, 在所有数据集上均取得了最优或相当的人脸识别性能.

2) 在绝大多数数据集实验上, 采取测地线距离度量的 v-LPMIVSVM 方法在一定程度上均优于或等同于采用欧氏距离度量的 v-eLPMIVSVM 方法, 从而说明测地线距离度量在一定程度上有利于图像数据模式识别性能的提升.

3) 从表 6 ∼ 8 可看出, 在各类对象训练样本较少时, 由于不能较好地刻画数据类内的潜在本质流形结构, 从而导致 5 种方法均不能取得较好的模式识别率, 但是, 相较之下, 本文方法在一定程度上依然取得了较好的识别性能. 随着各类对象训练数据集大小的增加, 数据空间的潜在流形结构呈现复杂, 传统的 v-SVM 方法的识别性能有所下降, 其他几种方法的识别性能虽在一定程度上有所增强, 但本文方法性能上升更明显, 从而说明, 不管是在简单的或是复杂的非线性流形结构的数据下, 由于充分考虑了数据的局部几何结构和局部差异 (或判别) 信息, 导致本文方法均具备较强的模式识别能力.

4) 由图 2 可看出, 当 $\lambda = 0$ 时, 即为忽略局部流形结构信息时, v-LPMIVSVM 和 v-eLPMIVSVM 方法的识别率最低, 随着 λ 逐渐增加, 两种方法识别率均上升, 但当 $\lambda = 1$ 时, 即为忽略数据流形局部差异信息时, v-LPMIVSVM 和 v-eLPMIVSVM 的识别率均下降, 且 v-eLPMIVSVM 方法下降更明显, 由此可看出, 在具有非线性流形结果的图像识别问题上, 确保局部流形结构的保持要强于局部判别信息的保持, 而同时考虑局部流形结构和局部差异信息确实能在一定程度上增强图像模式识别能力. 另外, 从图 3 可以看出, 在 k 值相对较小时, 本文方法能达到最优识别率, 而 v-eLPMIVSVM 方法在 k 值较大时才达到优化识别率, 随着 k 值进一步增大, v-LPMIVSVM 的人脸识别率呈下降趋势, 这是因为随着最近邻数 k 值增大, 最短路径已不能很好地近似计算测地线距离, 从而导致对数据的本质几何流形结构的估计误差增大. 由上分析可知, 参数 λ 和 k 值的优化选取确实对本文方法的识别性能具有较大的影响.

3.3 参数选择讨论

所提方法中, 参数 t, h, k 和 λ 直接影响局部保留最大信息差矩阵矩阵 M 的计算, 从而由定理 5 可知, 所提方法中参数 t, h, k, C, μ 和 λ 的取值在一定程度上直接 (C 和 μ) 或间接 (t, h, k 和 λ) 影响所提方法的泛化性能的提升. 因此, 上述参数的取值策略以及取值优化与否均在一定程度上决定了所提方法的泛化性能. 本节将对上述实验参数的选取策略展开讨论.

文献 [5] 分析指出, 在 SVM 中引入 μ 参数后, 正则参数 C 可以为一个固定常数 $1/N$, 其中 N 为总体样本数, 据此, 本文实验中令参数 $C = 1/N$; 依据文献 [14] 策略, 热核参数 t 选取为 $t = 2^m\sigma_0$, 其中, σ_0 为训练样本范数平方的标准差, $m \in \{-10, -9, \cdots, 0, \cdots, 9, 10\}$, 参数 m 采取 10 重交叉验证法选取; 参数 h 的选取策略与参数 t 相同; 参数 k 采取 10 重交叉验证法选取, 由于所提方法采用测地线距离来度量样本间的相似性, 因此样本近邻集大小 k 不能取值太大, 否则导致测量精度下降 (如图 3 所示), 故此, 确定实验参数 k 的选取范围为 {1, 2, $\cdots$, 15}; 依据文献 [5] 策略, 参数 μ 的取值范围为 $\{0.01\,n, 0.1\,n\}$, 其中 n 为 1 ∼ 10 间整数.

当参数 $\lambda = 1$ 时, 本文方法仅考虑局部流形结构信息, 等同于文献 [13] 中方法, 当 λ 在 $(0, 1)$ 间取值时, 本文方法既能较好地保持局部流形结构, 又能保持较强的局部判别信息. 参数 λ 在一定程度上起到平衡局部流形结构信息和局部判别信息的能量变化的作用, 为了简单起见, 实验参数 λ 定义为局部同类离散度矩阵能量与局部总体离散度矩阵 (局部同类离散度矩阵与局部异类离散度矩阵之和) 能量的比率[14], 即 $\lambda = 2^{\frac{a}{4.5}}\frac{\lambda_{\max}(H_o)}{\lambda_{\max}(H_o+H_e)}$, 其中, $\lambda_{\max}(A)$ 指矩阵 A 的最大特征值, 常量 a 依据 10 重交叉验证选取法的取值范围为 $\{-20, -19, \cdots, 0, \cdots, 19, 20\}$.

从第 3.1 节和第 3.2 节的实验分析结果可知, 尽管采取上述参数选择策略在一定程度上能确保所提方法取得可比较的实验优化性能, 但是如何更有效地协调上述实验参数并使之分别达到最优值仍然是一个值得关注并进一步研究的公开问题.

4 结论

本文方法与传统的大间隔分类机 v-SVM, MCVSVM 和 MCLPVSVM 等方法具有某些相同的优化特性: 基于结构风险最小化思想来构建原始优化问题形式, 通过求解一个凸优化问题来取得全局最优解, 采用现有软件包 (如 LibSVM 等) 使得算法实现简单等, 不同的是, 所提方法在考虑数据空间局部本质流形结构的同时还考虑了局部差异 (或判别) 信息, 另外, 在流形数据空间的距离度量方式上, 所提方法采用了适于流形数据距离测量的测地线距

离度量, 其更能反映流形数据的本质距离, 从而使得本文方法在一定程度上具备较强的学习泛化能力. 在人造和实际数据集上测试证实, 所提方法具有优于或等同于相关方法的优良性能. 对待小样本问题上, 尽管 v-LPMIVSVM 采用 PCA 对原始输入数据进行降维处理来达到解决问题的目的, 但是仍然还有许多随之而产生的问题 (如降维数的优化问题, 降维后信息缺失问题等) 有待进一步解决; 另外, 如何更有效地协调选取参数 t, h, k, μ, k, λ 的优化取值也是一个有待进一步研究的问题.

References

1 Tao J W, Wang S T, Hu W J, Ying W H. ρ-margin kernel learning machine with magnetic field effect for both binary classification and novelty detection. *International Journal of Software and Informatics*, 2010, **4**(3): 305−324

2 Vapnik V N. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995. 69−83

3 Wen Chuan-Jun, Zhan Yong-Zhao, Chen Chang-Jun. Maximal-margin minimal-volume hypersphere support vector machine. *Control and Decision*, 2010, **25**(1): 79−83
(文传军, 詹永照, 陈长军. 最大间隔最小体积球形支持向量机. 控制与决策, 2010, **25**(1): 79−83)

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

5 Scholkopf B, Smola A J, Williamson R C, Bartlett P L. New support vector algorithms. *Neural Computation*, 2000, **12**(5): 1207−1245

6 Shivaswamy P K, Jebara T. Maximum relative margin and data-dependent regularization. *Journal of Machine Learning Research*, 2010, **11**: 747−788

7 Zafeiriou S, Tefas A, Pitas I. Minimum class variance support vector machines. *IEEE Transactions on Image Processing*, 2007, **16**(10): 2551−2564

8 Cai D, He X F, Zhou K, Han J W, Bao H J. Locality sensitive discriminant analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India: IJCAI Press, 2007. 708−713

9 He X F, Yan S C, Hu Y X, Niyogi P, Zhang H J. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(3): 328−340

10 Tenenbaum J B, Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, **290**(5500): 2319−2323

11 Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled unlabeled examples. *Journal of Machine Learning Research*, 2006, **7**: 2399−2434

12 Gao Jun, Wang Shi-Tong, Deng Zhao-Hong. Global and local preserving based semi-supervised support vector machine. *Acta Electronica Sinica*, 2010, **38**(7): 1626−1634
(皋军, 王士同, 邓赵红. 基于全局和局部保持的半监督支持向量机. 电子学报, 2010, **38**(7): 1626−1634)

13 Wang X M, Chung F L, Wang S T. On minimum class locality preserving variance support vector machine. *Pattern Recognition*, 2010, **43**(8): 2753−2762

14 Wang H X, Chen S, Hu Z L, Zheng W M. Locality-preserved maximum information projection. *IEEE Transactions on Neural Networks*, 2008, **19**(4): 571−585

15 Li H F, Jiang T, Zhang K S. Efficient and robust feature extraction by maximum margin criterion. *IEEE Transactions on Neural Networks*, 2006, **17**(1): 157−165

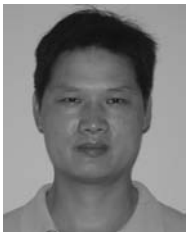
16 Gao Quan-Xue, Xie De-Yan, Xu Hui, Li Yuan-Zheng, Gao Xi-Quan. Supervised feature extraction based on information fusion of local structure and diversity information. *Acta Automatica Sinica*, 2010, **36**(8): 1107−1114
(高全学, 谢德燕, 徐辉, 李远征, 高西全. 融合局部结构和差异信息的监督特征提取算法. 自动化学报, 2010, **36**(8): 1107−1114)

17 Jun G, Chung F L, Wang S T. Matrix pattern based minimum within-class scatter support vector machines. *Applied Soft Computing*, 2011, **11**(8): 5602−5610

18 Muller K R, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 2001, **12**(2): 181−201

19 Scholkopf B, Herbrich R, Smola A J. A generalized representer theorem. In: Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory. Amsterdam, Netherlands: Springer, 2001. 416−426

20 Cai D, He X F, Han J W, Zhang H J. Orthogonal Laplacianfaces for face recognition. *IEEE Transactions on Image Processing*, 2006, **15**(11): 3608−3614



陶剑文 江南大学信息工程学院博士研究生. 主要研究方向为模式识别与数据挖掘技术.
E-mail: jianwen_tao@yahoo.com.cn
(**TAO Jian-Wen** Ph. D. candidate at Southern Yangtze University. His research interest covers pattern recognition and data mining.)

王士同 江南大学信息工程学院教授. 主要研究方向为人工智能, 机器学习. 本文通信作者.
E-mail: wxwangst@yahoo.com.cn
(**WANG Shi-Tong** Professor at Southern Yangtze University. His research interest covers artificial intelligence and machine learning. Corresponding author of this paper.)