

局部化的广义特征值最接近支持向量机

杨绪兵¹⁾ 陈松灿¹⁾ 杨益民²⁾

¹⁾(南京航空航天大学信息科学技术学院 南京 210016)

²⁾(南京财经大学统计系 南京 210003)

摘 要 基于广义特征值的最接近支持向量机(Proximal Support Vector Machine via Generalized Eigenvalues, GEPSVM)是一种新的具有与 SVM 性能相当的两分类方法,通过求解广义特征值来获得两个彼此不平行的拟合两类样本的超平面.其决策是将测试样本归为距其最近的超平面所在的类.然而,该规则在某些情形会导致较差的分类结果.对此,在 GEPSVM 基础上,通过在类拟合超平面上寻找一个包含了所有训练样本投影的局部凸区域,来决定样本的类别.该局部方法不仅具有较 GEPSVM 更优的分类性能,同时还衍生出了求解超平面上凸壳的简单且易于核化的新算法.最后在人工和 UCI 数据集上获得了验证.

关键词 最接近支持向量机;广义特征值问题;凸壳;局部化;分类

中图法分类号 TP181

Localized Proximal Support Vector Machine via Generalized Eigenvalues

YANG Xu-Bing¹⁾ CHEN Song-Can¹⁾ YANG Yi-Min²⁾

¹⁾(College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

²⁾(Department of Statistics, Nanjing University of Finance and Economics, Nanjing 210003)

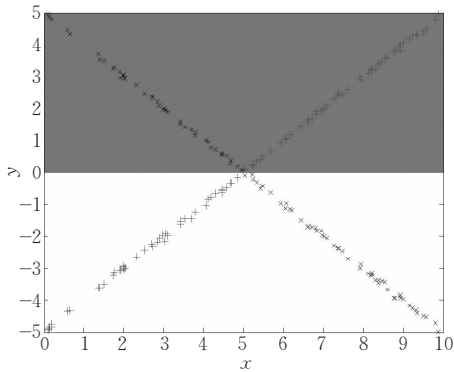
Abstract A binary classifier termed as proximal support vector machine via generalized eigenvalues(GEPSVM), is proposed recently. It aims to obtain two nonparallel planes generated from their corresponding generalized eigenvalue problem and has equivalent test correctness to SVM. In nature, GEPSVM attempts fitting two-class points with two planes. For an unseen sample, according to decision rule of GEPSVM, it will be assigned to the closest planes. In fact, this rule, in most cases, may result in poor test correctness. In this paper, based on GEPSVM, a new classifier named Localized GEPSVM is presented. Instead of two fitting planes, an unknown sample will be classified to the closest localized planes, i. e., convex hull, which are generated from the projections of two-class training points, respectively. Compared to GEPSVM, LGEPSVM outperforms GEPSVM in test correctness. Derivatively, LGEPSVM also develops an algorithm for solving convex hull on the projective hyperplane. Besides simple geometrical interpretation, this algorithm eases up to kernel version. Finally, Test accuracy of LGEPSVM algorithms will be validated on some artificial and real UCI datasets.

Keywords proximal support vector machine; generalized eigenvalues; convex hull; localization; classification

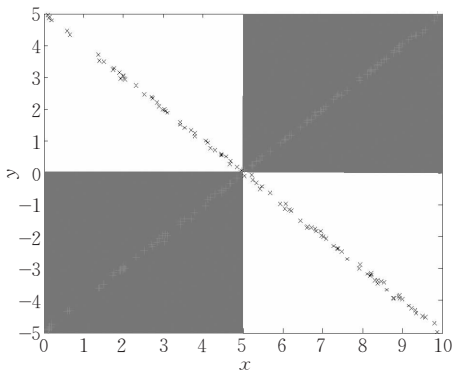
1 引 言

传统的两分类方法均是根据优化准则求解最优分类面,从而将数据空间分成两个不交的子空间(如图 1(a)),如线性判别分析(Linear Discriminant Analysis, LDA)、SVM 等. SVM 是建立在统计学习理论和结构风险最小原理的基础之上,根据有限样本信息在模型的复杂性和推广能力之间寻求最佳折衷,通过求解二次规划获得最优分类面,因此其时间复杂度较高($O(n^{3.5})$ ^[1], n 是训练样本个数). 由 Fung 和 Mangasarian 等提出的最接近支持向量机^[2-3](Proximal SVM, PSVM)采用等式约束替换 SVM 原问题描述中的不等式约束,使问题归结为线性方程组的求解,其时间复杂度为 $O(n^3)$. 实质上,

PSVM 是用两个平行超平面来分别拟合两类样本,同时要求这两个平面间的距离尽可能大. 2006 年, Mangasarian 和 Wild 进一步在 TPAMI 上发表了推广型 PSVM: GEPSVM^[4]. GEPSVM 通过计算两个广义特征方程(时间复杂度为 $O(n^3)$)来求解两个无平行约束的超平面,使得每一个超平面距离本类样本尽可能近,而距离它类样本尽可能远,其本质是用两个超平面来分别拟合两类样本. 与传统二分类方法不同的是,GEPSVM 将数据空间分成四个不相交的子空间,不相邻的两个空间对应同一类别(如图 1(b)),如此,GEPSVM 可以更有效地解决 XOR 等问题. 文献[4]的实验说明,较之 SVM, GEPSVM 具有与 SVM 相当的分类能力,同时,它的计算耗费远小于 SVM.



(a) 一般的二分类方法所得决策面,它将数据空间一分为二



(b) GEPSVM求得的4个不交子空间

图 1 2 个不交子空间与 4 个不交子空间(两类样本分别用“+”和“×”标示)

GEPSVM 采用的分类规则是:对任一待测样本,它距离哪个拟合超平面最近,就将其判为哪一类. 然而,当两类超平面在训练样本所在区域出现交叉时,即使在线性可分情形下(如图 2),也会导致较差的分类结果. 以图 2 为例来分析 GEPSVM: (1) GEPSVM 的实质是用线性函数来拟合样本,所以当样本呈线性或拟线性分布时 GEPSVM 非常有效,如 XOR 问题; (2) 在图 2 中,样本呈拟线性分布,尽管两类样本不重叠(即线性可分),但两个拟合平面是交叉的(且交叉点位于某一类样本中),如此,造成某些样本错分(特别是在交叉附近的样本)是必然的. 此类问题同样也会出现在高斯型数据分类问题中(如图 3). 解决此类问题,最直接的想法就是寻找两个能够包含所有训练样本的局部区域,通过计算测试样本到这些局部区域的距离来决定其类别,通常情况下,这种局部区域是由文献[5-6]中提到的凸壳来构造,但是,如此将面临两大难题: (1) 目前

尚无有效求解高维凸壳的算法; (2) 需要用二次规划计算点到高维凸壳的距离(其复杂度相当于求解一个支持向量机). 改进方法之一是文献[5-6]中所提出的方法,即对任一测试样本,在 C 类中分别找到它的 k 个近邻,由 k 近邻构造凸集(用凸线性组合),先计算与之对应的 C 类凸集(C 为类别个数),再计算该测试样本到 C 个凸集的距离实现判别. 然而,该方法虽然回避了直接求解凸壳的问题,但对每一个测试样本,仍然需要求解一个线性方程组,而且还必须考虑正则化因子的选择问题. 既然 GEPSVM 的分类问题是基于超平面的,当然更期望这些错分问题在超平面上得以解决. 同时,考虑到超平面自身的特点,此类问题应该能够在超平面上得以解决,如此将大大降低计算开销.

鉴于此,在 GEPSVM 基础上,本文提出了局部化的 GEPSVM (Localized GEPSVM, LGEPSVM),期望能提高 GEPSVM 的分类能力. LGEPSVM 的

基本思想是:在超平面上寻找凸壳,根据样本到凸壳的距离决定其分类(如图 2 和图 3 中所示的线段,其两个凸集顶点标示为“□”).该算法的基本步骤如下:(1)用 GEPSVM 启动 LGEPSVM,产生两个平面;(2)在 GEPSVM 所得的两个拟合超平面(图 2 中所示的两条直线)上找两个凸壳(如图 2 中两个线段,每个线段由投影点的凸线性组合生成,标注为“□”为凸集顶点);(3)分类规则:对任一个测试样本,将其判为距离它最近的凸壳所对应的那一类.此外,本文在求解过程中,提出了一种在超平面上易于核化的凸壳求解方法,该方法对其它算法,诸如子类问题,逐片线性回归等均有借鉴作用.

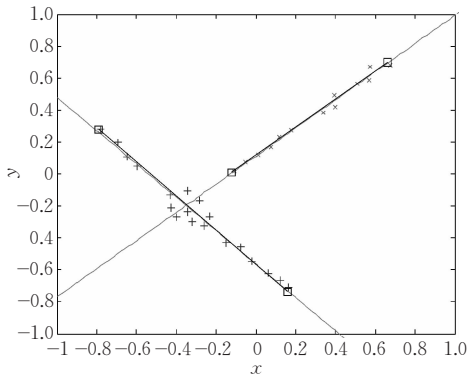


图 2 线性可分的拟线性分布示例

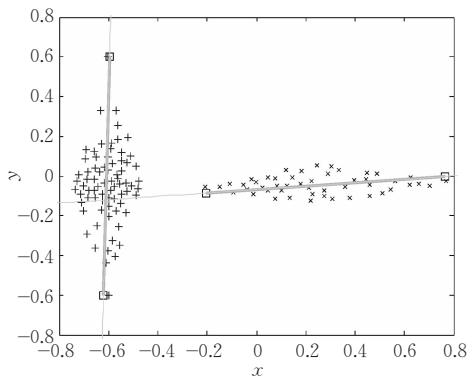


图 3 线性可分的高斯分布示例

本文第 2 节是 GEPSVM 的简单回顾;第 3 节介绍 LGEPSVM 方法及由此发展出来的求超平面上凸壳算法;在人工数据集和标准数据集上的实验结果将放在第 4 节.

2 GEPSVM 的简单回顾

给定两类 d 维数据集 $\{\mathbf{x}_i^{(j)}\}, i=1,2,\dots,n_j; j=1,2, \mathbf{x}_i^{(j)} \in \mathcal{R}^d$, 样本数分别为 n_1, n_2 . GEPSVM^[4,7] 的目标是要在 d 维空间中寻找两个超平面:

$$\mathbf{x}^T \mathbf{w}^1 - r^1 = 0, \mathbf{x}^T \mathbf{w}^2 - r^2 = 0 \quad (1)$$

要求第一个超平面距第一类(本类)样本尽可能近,距第二类(它类)样本尽可能远.所有公式中的 T 均表示向量(矩阵)转置.

第一类超平面优化准则为

$$\min_{(\mathbf{w}^1, r^1)} \frac{\|\mathbf{A}\mathbf{w}^1 - \mathbf{e}r^1\|^2 + \delta \left\| \begin{bmatrix} \mathbf{w}^1 \\ r^1 \end{bmatrix} \right\|^2}{\|\mathbf{B}\mathbf{w}^1 - \mathbf{e}r^1\|^2} \quad (2)$$

式(2)中 $\|\cdot\|$ 表示 L_2 范数, δ 是正则化因子, \mathbf{A}, \mathbf{B} 是由第一、二类样本构成的样本矩阵,即 $\mathbf{A} = [\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \dots, \mathbf{x}_{n_1}^{(1)}]^T$, $\mathbf{B} = [\mathbf{x}_1^{(2)}, \mathbf{x}_2^{(2)}, \dots, \mathbf{x}_{n_2}^{(2)}]^T$, \mathbf{e} 是分量全为 1 的列向量.式(2)中如果忽略正则项,几何意义十分明确,即在最小化目标要求下,分子要尽可能小,而分母要尽可能大,分子为第一类样本到第一类超平面的距离平方和,分母为第二类样本到第一类超平面的距离平方和;在此目标下,即要求第一类超平面距第一类样本尽可能近,而距第二类样本尽可能远.记 $\mathbf{G} = [\mathbf{A} \quad -\mathbf{e}]^T [\mathbf{A} \quad -\mathbf{e}] + \delta \mathbf{I}$, $\mathbf{H} = [\mathbf{B} \quad -\mathbf{e}]^T [\mathbf{B} \quad -\mathbf{e}]$, $\mathbf{z} = \begin{bmatrix} \mathbf{w}^1 \\ r^1 \end{bmatrix}$, \mathbf{G}, \mathbf{H} 均为 $(d+1) \times (d+1)$ 阶对称矩阵.

式(2)很容易转化为求解下列广义特征值问题:

$$\mathbf{G}\mathbf{z} = \lambda \mathbf{H}\mathbf{z}, \mathbf{z} \neq 0 \quad (3)$$

正则化后, \mathbf{G} 可能由半正定转变为正定,可保证式(3)的广义特征值不会出现奇异性问题.最小特征值对应的特征向量 \mathbf{z} 即为准则(2)的最优解,其前 d 个分量为所求的第一个超平面的 \mathbf{w}^1 , 第 $d+1$ 个分量为 r^1 . 同理可求解第二类超平面.

3 局部化的 GEPSVM(Localized GEPSVM)

从图 2 可直观了解到,按 GEPSVM 的分类规则,很多样本将被错分,但是,如果根据 LGEPSVM,即将测试样本归为距其最近的凸壳所在的那一类,则所有样本均可获得正确分类.自核方法问世以来,原空间中的线性不可分问题变得易于处理,通过非线性隐映射 ϕ ,将原空间中的样本 \mathbf{x} 映射为高维特征空间中的特征样本 $\phi(\mathbf{x})$,并用内积形式刻画问题,如此,原空间中的线性不可分问题,将有望在特征空间中仍能以线性方法获得求解^[8-10].

第 1 类的特征样本简记为 $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \dots, \boldsymbol{\varphi}_{n_1}$, 它们在第 1 类拟合超平面(即投影平面)上的投影记为 $\boldsymbol{\varphi}_1^p, \boldsymbol{\varphi}_2^p, \dots, \boldsymbol{\varphi}_{n_1}^p$, 记由投影生成的凸集记为 $P = \{\mathbf{p} |$

$p = \sum \alpha_i \phi_i^p, \sum \alpha_i = 1, \alpha_i \geq 0, i = 1, 2, \dots, n_1$. 一般情形, 点 ϕ 到凸集的距离^[5-6] (事实上是点到凸壳的距离) 表示为

$$\min_{\alpha} \left\| \phi - \sum \alpha_i \phi_i^p \right\|^2 \quad (4)$$

s. t. $\sum \alpha_i = 1, \alpha_i \geq 0, i = 1, 2, \dots, n_1$

式(4)是一个典型的二次优化问题, 由此导致分类(测试)阶段过大的计算量. 文献[5-6]将式(4)的约束条件加入目标函数, 将问题(4)转化为一个只需求解线性方程组的无约束问题, 即对每一个测试样本, 均需求解一个线性方程组(时间复杂度 $O(n^3)$), 同时还面临着正则化因子的选择问题, 计算量仍较大.

3.1 最小凸壳顶点集的计算(训练阶段)

本文中, LGPSVM 的设计目标之一, 就是希望能够提高测试速度. 而凸线性组合方法对每个测试样本均要重新计算组合系数, 显然不适合本问题. 其次, 因测试点的投影到凸壳的距离可通过投影到凸壳顶点的距离来计算, 同时, 凸壳的顶点集相对于样本的投影集是稀疏的, 可以有效地节省存储空间, 更为重要的是, 寻找凸壳顶点的工作完全可以在训练阶段完成. 然而, 常用计算凸壳顶点的算法^[11-12]存在如下问题: (1) 局限于二维情况, 无法推广到高维; (2) 算法不能核化, 无法推广到更高维的特征空间. 本文提出的凸壳顶点的求解算法是直接针对超平面而设计的.

算法思想描述如下:

(1) 先计算距离最远的两点 x_1 和 x_2 (可以证明此两点皆为凸壳顶点), 任取一点为定点 (不妨取 x_2 为定点). 过点 x_2 (与下文对应, 记为 v_1) 且以 $x_1 - x_2$ 为法线方向作一超平面 (可以证明除 x_2 外, 所有投影点均在此平面的同侧), 在该平面与投影平面 (两平面垂直) 的交线上任取异于 x_2 的一点 x_0 , 记 $x_0 - x_2$ 为定方向;

(2) 计算除 x_2 外的所有投影点与 x_2 连线与定方向的夹角, 选择夹角最小者对应的投影点 (可通过内积表达), 记为 v_2 (也是凸壳顶点);

(3) 以 $v_2 - v_1$ 为定方向, 计算除 v_1 外的所有投影点与 v_1 连线与定方向的夹角, 选择夹角最小者对应的投影点 (可通过内积表达) 为 v_3 (凸点);

(4) 令 $v_1 = v_2, v_2 = v_3$, 重复步(3), 直到初始定点 x_2 再次被选作凸壳顶点结束.

算法 1. 计算超平面上凸壳顶点集 (以特征样本进行描述).

输入: 投影集 $Pset = \{\phi_1^p, \phi_2^p, \dots, \phi_{n_1}^p\}$, 投影平面:

$\phi(x)^T w_\phi^1 - r^1 = 0; PsetIndex = \{1, 2, \dots, n_1\}$

输出: 最小凸壳的顶点集 $Vset$

1. 初始化, $Vset = \emptyset$, $Vset$ 中元素个数 $Num = 0$;
 $PsetIndex = \{1, 2, \dots, n_1\}$;

2. 计算投影集中距离最远两点, 记为 $\phi_{i_1}^p, \phi_{i_2}^p$ (可以证明这两点是凸壳上的顶点);

3. 任取其中一点, 不妨取 $\phi_{i_2}^p$ 为定点, 令 $FP = \phi_{i_2}^p$ 和 $FIndex = \{i_2\}$, 在以 $\phi_{i_1}^p - \phi_{i_2}^p$ 为法方向且过定点的超平面 (可以证明所有投影均在此平面的同侧) 与投影超平面的交线上, 取一异于 $\phi_{i_2}^p$ 的点 ϕ_0 , 即 ϕ_0 满足

$$\begin{cases} \phi_0^T w_\phi^1 - r^1 = 0 \\ (\phi_{i_1}^p - \phi_{i_2}^p)^T (\phi_0 - \phi_{i_2}^p) = 0 \end{cases} \quad (5)$$

计算 $\phi_0 - \phi_{i_2}^p$ 的单位矢量方向: $FDirect = (\phi_0 - \phi_{i_2}^p) / \|\phi_0 - \phi_{i_2}^p\|$;

$Vset = Vset \cup \{FP\}$; $NewP = \phi_0$;

4. While $NewP \notin Vset$

If $Num \neq 0$ then $Vset = Vset \cup \{NewP\}$; end

$Num = Num + 1$; 差集 $D = PsetIndex - FIndex$;

计算 $[\phi_{D(1)}^p - FP, \phi_{D(2)}^p - FP, \dots, \phi_{D(n_1-1)}^p - FP]$ 并按列归一化, 记为 T ; 其中 $D(j)$ 表示集合 D 的第 j 个元素; 记 $R = [FDirect, \dots, FDirect]$;

// R 是 n_1 列全为向量 $FDirect$ 的矩阵;

计算 $R^T T$ 并取其对角线元素, 构成列向量 $M = \text{diag}(R^T T)$;

$[value, FIndex] = \max(M)$;

// 取 M 最大值位置 (亦可为最小值);

// $FP = Vset(Num)$;

$FP = Vset(Num)$; $NewP = Pset(DIndex(FIndex))$;

$FDirect = (NewP - FP) / \|NewP - FP\|$;

End

5. 返回 $Vset$.

几点说明: (1) 如果是二维情形, 式(5)退化为一个方程; (2) 步 4 中 M 值实质是两个直线的夹角余弦, 值越大, 对应的夹角越小; 当最大值相同时, 取距定点距离最大者为 $NewP$; (3) 所得 $Vset$ 集的顶点要求有序 (序可按顶点进入 $Vset$ 的先后次序指定, 此举仅为降低测试阶段的计算量而人为设定, 详见本文 3.2.1 节), 可用链表实现 (因凸壳是封闭的, 故可设计成循环链表).

3.2 测试点在凸壳内外的判断准则

3.2.1 凸壳顶点不共线 (平面上凸壳顶点至少有 3 个) 的情况

按算法 1 中说明(3), 设超平面上的凸壳是由有序顶点集中所有相邻两点的边连接而成. 记凸壳上的有序顶点为 v_1, v_2, \dots, v_l , 凸壳上对应的有序线段 (简称为凸线段) 记为 $\overline{v_1 v_2}, \overline{v_1 v_2}, \dots, \overline{v_l v_1}$, 凸壳中心

$m = \frac{1}{l} \sum_{i=1}^l v_i$ (可证由凸壳顶点的凸线性组合所得的

凸集与全体投影点的凸线性组合所得的凸集相等，从而可得出该中心和全体投影所得的中心重合的结论，限于篇幅，证明略）在各凸线段方向上的投影记为 m_1, m_2, \dots, m_l ，取各凸线段的法线方向 $m-m_1, m-m_2, \dots, m-m_l$ 并记为矩阵 P 。如此构造凸线段的法线方向优势在于：(1) 所有法线方向均指向壳内；(2) 壳内的点 x 到凸线段的代数距离 $(x^T w^1 - r^1)/\|w^1\|$ 均大于零。以上均可在训练阶段完成。

记测试样本 φ 在投影平面上的投影为 φ_p ，构造矩阵 $B=[\varphi_p-m_1, \dots, \varphi_p-m_l]$ ，令

$$Flag=\min(\text{diag}(P^T B)) \tag{6}$$

其中 $\text{diag}(A)$ 表示由矩阵 A 的对角元构成的列向量， $\min(x)$ 表示取向量 x 的最小元素。以上方法用到两个公式^[13-14]：点到超平面距离和点在超平面上的投影公式。因凸线段的法线方向均指向凸壳内，所以壳内的点到该凸线段的代数距离均大于 0。

如此，可根据式(6)判断 φ_p 与凸壳的位置关系：当 $Flag>0$ ，壳内； $Flag=0$ ，壳上； $Flag<0$ ，壳外。

3.2.2 凸壳顶点共线(超平面上的凸壳顶点仅有两个)的情况

3.2.1 节讨论的问题，均是在凸壳顶点不共线条件下得出的结果，下面讨论一个特例，即在投影超平面退化为一条直线，此时，凸壳仅有两个顶点(由算法 1 的第 2 点说明可得此结论)。在此条件下，凸壳的中心也在该直线上，上述方法失效。在本文中，采用如下方法进行判断：

以两个凸壳顶点 v_1, v_2 为端点的线段上的任一点可表示为 $\varphi_p=\lambda v_1+(1-\lambda)v_2, \lambda\in[0,1]$ 。平面上任一点 y 到该线段的最短距离为 $\min f(\lambda)=$

$\|\lambda v_1+(1-\lambda)v_2-\varphi_p\|^2$ ，容易求得在 $\lambda=(\varphi_p-v_2)^T(v_1-v_2)/\|v_1-v_2\|^2$ 时对应的距离为最短(易证 λ 是上述问题极值点且是唯一一个极小值点)。如果 $\lambda\in[0,1]$ ，则 y 在凸壳内；否则在壳外。

3.3 测试点到凸壳的距离计算方法

凸壳顶点共线时的距离较简单，以下仅对不共线问题简单描述。对一未知类别的样本 x ，按式(6)，判断它的投影(投影超平面由 GEPSVM 求得)和超平面上凸壳的位置关系，如果在壳内(包括在壳上)，该样本到凸壳的距离就等于它到超平面的距离；如果在壳外，它到凸壳的距离可由以下两个距离(三者构成直角三角形)算出：(1) 该点投影到凸壳的距离；(2) 它到投影超平面的距离。比较测试点到两个凸壳的距离后，距谁最近就将其归为那一类。

下面将通过实验，比较 GEPSVM 及 LGEPSVM 的分类性能。

4 实验结果

4.1 凸壳算法示例

为验证本文提出的凸壳算法的有效性，下面先给出两个凸壳的示例，图 4(a)是一个高斯分布的二维数据(用“·”表示)，实线段围成的一个凸多边形即为凸壳(凸壳顶点是有序的，已在图中用数字标注)。图 4(b)是一个三维样本(用“×”表示)，服从高斯分布，它们在给定平面上的投影(“·”标示)、包含所有投影的凸壳(三维上的凸多边形)和凸壳顶点与样本的对应关系均显示在图 4 中。

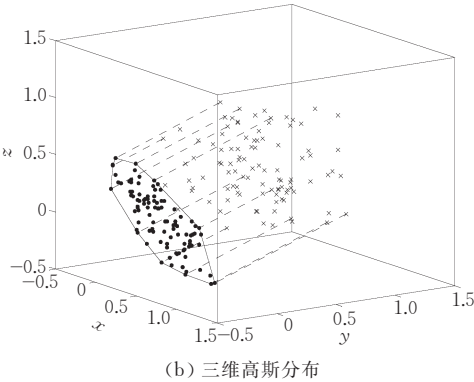
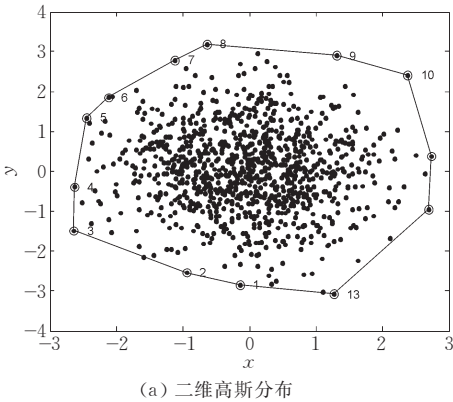


图 4 凸壳求解示例

凸壳算法有效性的进一步验证将通过分类精度来反映。下节将测试 LGEPSVM 的分类精度。

4.2 分类能力测试

数据采用人工数据集和 UCI^① 标准数据集。采用

10 重(10-fold)交叉验证测试 GEPSVM 和 LGEPSVM 的分类性能。在每一重的训练集中随机选择 10% 用

① Datasets available at <http://www.ics.uci.edu/~mllearn/MLRepository.html>

于参数(包括正则化因子和核参数)选择,正则化因子 δ 的选择集为 $\{10^i | i = -7, \dots, 7\}$,高斯核参数的选择集为 $\{10^i | i = -4, \dots, 4\}$. 两种算法的分类精度均是 10 重交叉验证的平均结果. 为克服随机性,本文还对两者的分类精度进行了 t -检验,显著性水平取 0.05.

4.2.1 线性核

本节中,GEPSVM 和 LGEPSVM 均采用线性核,实验结果如表 1. 两种算法的分类精度与标准差如表 1 所示,其中 p 值表示根据现有样本计算出来的统计量,粗体表示具有较好的分类能力,“*”表示两者的平均分类精度存在显著差异.

表 1 采用线性核时,两种算法的分类性能比较

数据集大小	分类精度±标准差/%		
	GEPSVM	LGEPSVM	p 值
Sep2Sphere 350×2	86.30±6.23	100±0.00*	0.002
Cross2Plane 200×2	99.50±1.12	99.50±1.12	—
Cmc 1473×8	72.60±3.70	81.37±4.39*	0.001
Monk2 432×6	62.97±3.02	71.80±5.34*	0.010
Monk3 432×6	80.86±4.83	81.57±5.46	0.070
Liver 345×6	58.26±5.46	59.42±3.97	0.596
Pima 768×8	75.41±4.29	73.33±3.86	0.076

表 1 中的数据集 Sep2Sphere 由图 3 所示的两类高斯分布的样本中加入 10% 的噪声扩充而来,Cross2Plane 是文献[4]中图 1 所示的两类相互交叉样本. 从表 1 可以得知,LGEPSVM 弥补了 GEPSVM 分类规则在线性可分时的不足(线性可分时,Sep2Sphere 的分类精度达到了 100%). 对于 Cross2Plane 数据集,因为测试点的投影均在对应的凸壳内,此时,GEPSVM 与 LGEPSVM 等价,正是因为两个算法在该数据集上的分类精度完全一样,无法进行 t -检验.

4.2.2 高斯核

采用高斯核,两种算法的分类性能比较如表 2 所示.

表 2 采用高斯核时,两种算法的分类性能比较

数据集大小	分类精度±标准差/%		
	GEPSVM	LGEPSVM	p 值
Sep2Sphere 350×2	86.42±5.74	98.00±4.47*	0.014
WDBC 569×30	73.66±4.31	78.81±9.60*	0.041
Pima 768×8	45.08±4.23	52.96±2.97	0.349
Water 116×38	68.22±8.15	66.49±10.20	0.688
WPBC 194×32	64.32±5.83	74.21±3.90*	0.036
BUPA 345×6	63.41±3.29	64.39±2.19	0.594
Checkdata 1000×2	55.61±5.33	55.88±4.76	0.920

综合两表,多数情况下,LGEPSVM 的分类精度均高于 GEPSVM,而且在某些数据集上的表现更为显著. 对于 Pima 数据集,在线性核时略低于

GEPSVM,但在高斯核时,LGEPSVM 反而高于 GEPSVM,这说明:(1)分类精度与所选择的核有关;(2)因高维数据本身的复杂性(如非线性性,包含多个子类等问题),因 LGEPSVM 以 GEPSVM 的拟合平面为基础,难以克服 GEPSVM 算法在训练阶段的缺陷,不可避免地会影响到凸壳的求解.

4.3 测试时间比较

在设计 LGEPSVM 算法时,目标之一就是尽可能提高分类速度. 下面将以 UCI 数据集 MuskClean 为例,比较两者的测试时间. 该两类数据集有 6598 个样本,166 维. 选择线性核,比较两种算法的分类时间(10 次平均). 实验环境:Windows XP,CPU 是 Pentium IV 3.0GHz,内存 512MB,计算平台 Matlab 7.0. 用 Matlab 函数 cputime 记录分类时间,单位为 s. 结果如表 3 所示.

表 3 两种算法在数据集 MuskClean(6598×166)上的平均测试时间与分类精度

算法	平均 CPU 时间/s	测试精度±标准差/%
GEPSVM	0.3197	38.79±1.52
LGEPSVM	1.1677	51.09±1.20

由表 3 可知,GEPSVM 的分类速度优于 LGEPSVM,但其中涉及到的两种因素必须考虑:(1)相对其它工具,Matlab 的优势之一在于做矩阵运算,而不是循环操作. 本实验中,对于 GEPSVM,测试集可以按矩阵形式进行运算;而 LGEPSVM,必须先判断每一测试样本的投影与凸壳的位置关系,然后才能做分类,无法进行批量运算;(2)尽管如此,对此庞大的数据集,LGEPSVM 的分类速度只需 1.2s,在很多应用问题中影响不大,更何况它具有较之 GEPSVM 更优的分类结果.

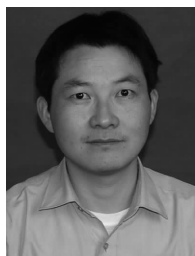
5 结束语

本文提出了一个局部化的 GEPSVM 算法,该算法修正了 GEPSVM 的分类规则,实验证明,LGEPSVM 具有比 GEPSVM 更优的分类性能. 此外,在求解过程中,发展出一个求解超平面上的凸壳算法,该算法不仅求解简单,易于核化,对诸如平面局部化、子类问题,逐片线性回归等均有借鉴作用.

参 考 文 献

[1] Kojima M, Mizuno S, Noma T, Yoshise A. A unified approach to interior point algorithms for linear complementarity

- problems//Lecture Notes in Computer Sciences. Berlin, Germany: Springer Verlag, 1991, 538(10): 247-254
- [2] Fung G, Mangasarian O. Proximal support vector machine classifiers//Proceeding of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA, 2001: 77-86
- [3] Agarwal D K. Shrinkage estimator generalizations of proximal support vector machines//Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining. Edmonton, Edmonton Canada, 2002: 173-183
- [4] Mangasarian O, Wild E. Multisurface proximal support vector machine classification via generalized eigenvalues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(1): 69-74
- [5] Vincent P, Bengio Y. K-local hyperplane and convex distance nearest neighbor algorithms//Dietterich T G, Becker S, Ghahramani S eds. Advances of Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002: 985-992
- [6] Bennett K P, Bredensteiner E J. Duality and geometry in SVM classifiers//Langley P ed. Proceedings of the 17th International Conference on Machine Learning (ICML2000). San Francisco: Morgan Kaufmann, 2000: 57-64
- [7] Yang Xu-Bing, Chen Song-Can. Proximal support vector machine based on prototypal multiclassification hyperplanes. Journal of Computer Research and Development, 2006, 10: 1700-1705(in Chinese)
(杨绪兵, 陈松灿. 基于原型超平面的多类最接近支持向量机. 计算机研究与发展, 2006, 10: 1700-1705)
- [8] Mika S, Ratsch G, Weston J. Fisher discriminant analysis with kernels//Hu Y H, J Larsen J, Wilson E, Douglas S eds. Neural Networks for Signal Processing 9. New York: IEEE Press, 1999: 41-48
- [9] Mika S, Schölkopf B, Smola A, K Müller K R. Kernel PCA and de-noising in feature spaces//Kearns M S, Solla S A, Cohn D A eds. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1999, 11: 536-542
- [10] Akaho S. A kernel method for canonical correlation analysis//Proceedings of the International Meeting on Psychometric Society (IMPS2001). Osaka, Japan, 2001. Berlin, Germany: Springer-Verlag, 2002: 101-105
- [11] Zhou Pei-De. Computational geometry: Algorithm Analysis and Design. 2nd Edition. Beijing: Tsinghua University Press, 2005(in Chinese)
(周培德. 计算几何: 算法分析与设计. 第二版. 北京: 清华大学出版社, 2005)
- [12] Mukhopadhyay A, Chatterjee S, Lafreniere B. On the all-farthest-segments problem for a planar set of points. Information Processing Letters, 2006, 100(3): 120-123
- [13] Bian Zhao-Qi, Zhang Xue-Gong. Pattern Recognition. 2nd Edition. Beijing: Tsinghua University Press, 2000(in Chinese)
(边肇祺, 张学工. 模式识别. 第二版. 北京: 清华大学出版社, 2000)
- [14] Duda R O, Hart P E, Stock D G. Pattern Classification. 2nd Edition. New York: John Wiley & Sons, Inc, 2001



YANG Xu-Bing, born in 1973, Ph.D. candidate. His current research interests include neural calculation and pattern recognition.

CHEN Song-Can, born in 1962, Ph.D., professor and Ph.D. supervisor. His main research interests include pattern recognition, medical image processing and neural networks.

YANG Yi-Min, born in 1955, M. S., professor. His current research interests include statistical algorithms and their estimation rule.

Background

This work is supported by two National Natural Science Foundation projects of China (grant Nos. 60473035 and 70671052). The former project is "Enhanced Linear Discriminant Analysis and its Generalization". Due to its simplicity and efficiency, linear discriminant analysis (LDA) has shown its outstanding classification performance and effective dimension-reduction in many applications such as handwritten digit recognition, face detection in image, text categorization and target tracking. However, there are also many embarrassments in LDA, such as singularity of scatter matrix, sin-

gle training sample each class, limitation problem of rank and so on. To breakthrough the foresaid limits, the authors have directly defined new optimization criterion to widen application range of LDA. Nowadays, another popular linear classifier, support vector machine (SVM) is based on the structural risk minimization (SRM) principle and aims at maximizing the margin between the points of two-class data classification. However, SVM requires a solution of quadratic programming (QP) problem. Recently, Fung and Mangasarian introduced a proximal SVM (PSVM), which replaces ine-

quality with equality constraints of the SVM framework. In doing so, the authors claim that the computational complexity can be greatly decreased without resulting in discernible loss of classification accuracy. Furthermore, PSVM classifies two-class points to the closest of two parallel planes that are pushed apart as far as possible. With similar starting point in defining objective function of LDA, Proximal support vector via generalized eigenvalue (GEPSVM) can be interpreted as replacing class centers of LDA with proximal planes. In a viewpoint of proximal planes, GEPSVM is another version of PSVM, in which only a set of linear algebra problem needs to be solved instead of a QP problem of traditional SVM. Its proximal planes are generated by a generalized eigenvalue problem such that GEPSVM is superior to SVM in computational time but has still comparable test correctness. The latter project is "Topology Synthesis of Evaluating Statistical Index Architecture and Generating Projection Set to be Evaluated", which aims to construct a new mechanism to

evaluate classifier' performance in statistical comparison.

For an unseen test point, according to decision rule of GEPSVM, it is assigned to that class of its closest plane. However, this rule, even in linear-separable cases, may result in poor test correctness, i. e. , generalization. In this paper, based on GEPSVM, a new classifier termed as Localized GEPSVM (LGEPSVM) is present. Instead of two proximal planes, an unknown sample is classified to the closest localized planes, i. e. , convex hulls, which are generated from the projection of two-class training points, respectively. In addition to reporting the average accuracies, the authors performed paired *t*-tests comparing LGEPSVM to GEPSVM. In most real datasets here, LGEPSVM outperforms GEPSVM in test accuracy. Derivatively, LGEPSVM is also used to develop an algorithm for solving minimal convex hull on the projection hyperplane which can solve subclass classification and piecewise linear regression problems etc.