

doi:10.3969/j.issn.1001-4616.2018.04.011

马氏度量下局部化广义特征值最接近支持向量机

周健航¹, 杨绪兵¹, 张福全¹, 业巧林¹, 许等平²

(1. 南京林业大学信息科学技术学院, 江苏 南京 210037)

(2. 国家林业局调查规划设计院, 北京 100714)

[摘要] 局部化广义特征值最接近支持向量机(Localized GEPSVM, LGEPSVM)是从广义特征值最接近支持向量机(GEPSVM; Proximal Support Vector Machine via Generalized Eigenvalues)衍生而来,其原理是在 GEPSVM 通过求解广义特征值获得两个彼此不平行的超平面的基础上,分别求解两个超平面的凸壳,修改 GEPSVM 的分类判据为将测试样本归为距其最近凸壳所属的那一类. 分析和实验表明, LGEPSVM 较之 GEPSVM 具有更高的分类精度. 然而,由于 LGEPSVM 在训练和分类过程中都涉及凸壳计算问题,因而费时较多. 为了缓解这一问题,本文提出的基于马氏度量的最小椭圆凸壳算法 MLGEPSVM(LGEPSVM based on Mahalanobis Metric),即分类时只需要判断样本与对应椭圆凸壳的距离. 较之 LGEPSVM 和 GEPSVM, MLGEPSVM 具有如下几个特点:(1)给出了马氏度量下的椭圆凸壳计算方法,(2)分类速度更快,(3)更低的存储空间,每类样本仅需存储椭圆凸壳(可通过中心和协方差表示),而不是所有的凸壳顶点. 在人工和标准数据集上的实验,验证了 MLGEPSVM 的上述性能.

[关键词] 最接近支持向量机, 广义特征值, 马氏度量, 凸壳

[中图分类号] TP391 **[文献标志码]** A **[文章编号]** 1001-4616(2018)04-0065-07

Localized GEPSVM Based on Mahalanobis Metric

Zhou Jianhang¹, Yang Xubing¹, Zhang Fuquan¹, Ye Qiaolin¹, Xu Dengping²

(1. College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China)

(2. Survey & Planning Institute of State Forestry Administration, Beijing 100714, China)

Abstract: GEPSVM(Proximal Support Vector Machine via Generalized Eigenvalues) have been played more attention in machine learning and pattern recognition. It adopts data fitting to construct classifier, and further leading to two Generalized Eigenvalue problems. One of its variants is Localized GEPSVM, shortly LGEPSVM. Instead of the closest nonparallel planes of GEPSVM, LGEPSVM classifies an unknown sample to the closest convex hulls on the projection plane. Experimental results show that LGEPSVM able to achieve comparable or even better test correctness than GEPSVM. However, due to training convex hull, LGEPSVM would cost much time in training stage. To speed training LGEPSVM, in this paper, we propose a new version LGEPSVM, termed as MLGEPSVM, based on Mahalanobis metric. Concretely, MLGEPSVM aims to find two ellipsoidal convex hulls, and then classify the samples to the class corresponding to its closest ellipsoid. Compared to LGEPSVM and GEPSVM, advantages of MLGEPSVM lie in three aspects: (1) calculation method of ellipsoid convex hull, (2) faster classification speed, and (3) less storage requirement, only ellipsoid convex hull of each class will be stored (the sample center and covariance matrix). Finally, analysis and experiments on artificial and UCI benchmark datasets will validate our foresaid superiorities.

Key words: proximal support vector machine, generalized eigenvalues, Mahalanobis metric, convex hull

支持向量机(Support Vector Machine, SVM)是基于统计学习和结构风险最小理论提出的,它采用间隔来度量模型的泛化能力,在兼顾经验误差最小的同时,最大化两类样本间的间隔. 然而,由于 SVM 的训练时间较长(需要求解二次规划问题),不适合大规模分类任务. 为缓解这一问题,后继学者提出了多种改良版本,例如 LSSVM(Least Square SVM)^[1], Reduced SVM^[2], PSVM(Proximal SVM)^[3]和 GEPSVM^[4]等.

收稿日期:2018-08-16.

基金项目:国家自然科学基金(31670554、50375057)、江苏省自然科学基金(BK20161527、BK20171543).

通讯联系人:杨绪兵,副教授,研究方向:模式识别、神经计算. E-mail:xyang@njfu.edu.cn

PSVM 采用两个彼此平行的超平面来分别拟合两类样本,使得分类超平面的求解问题可通过计算线性方程组完成,简化了传统 SVM 的求解过程,而 GEPSVM 进一步弱化了 PSVM 的拟合平面必须平行的约束条件,即采用两个不平行的平面(nonparallel planes)来拟合样本,导出的问题可通过求解两个广义特征值问题. 实验结果表明,GEPSVM 与 SVM、PSVM 具有相当的分类性能,但它尤其适合解决异或问题. 借鉴 PSVM 和 GEPSVM 的设计思想,后继学者进行了大量的拓展研究,如 DGEPSVM (Distance Difference DGEPSVM)^[6]、TWSVM(Twin SVM)^[5]、PCC(Proximal Classifier with Consistency)^[7]等. 就 GEPSVM 的分类规则(将待测样本归为距其最近的超平面)而言,势必在两个拟合平面的交叉区域产生大量错分样本. 针对这个问题,Yang 和 Chen 提出了 LGEPSVM^[8]:在拟合超平面上先计算两个凸壳(如文献[8]图 3 所示的两个线段),然后将待测样本归为距凸壳最近的那一类. 分析和实验表明,LEGPSVM 的分类性能要优于 GEPSVM. 采用凸壳的局部化思想,与 SVM 或 GEPSVM 结合,亦设计出了很多新算法,如 LCTSVM^[9-10]等,这些算法均表现出了能够降低计算开销、提高分类性能的特点.

由文献[8]知,LEGPSVM 的训练和归类速度均低于 GEPSVM,其本质原因在于:(1)训练过程中,LEGPSVM 是在获得拟合超平面基础上计算凸壳;(2)样本归类过程中,LEGPSVM 需要计算每一个待测样本的投影与其最近的两个凸壳顶点距离,进而判断待测样本与凸壳的距离以实现归类. 因而无法进行批量处理. 本文拟采用马氏度量来设计归类判据,以加快分类速度,期望弥补 LGEPSVM 的上述不足. 由于马氏度量更能融入样本的结构信息,具备尺度无关性^[11]等特点,在机器学习等诸多领域已有许多成功案例,如采用马氏度量的人体身份识别(Person Re-identification)^[12],基于马氏度量的农作物判别和差错检测等^[13-16]. 文献[17]也提供了一种基于马氏度量的凸壳计算方法. 鉴于此,本文拟用马氏度量的椭圆凸壳替换 LGEPSVM 的凸壳,以期提高训练和测试速度,该方法记为 MLGEPSVM,即在 GEPSVM 产生两个拟合超平面上,分别计算两个椭圆凸壳,再据此实现待测样本归类. 与 LGEPSVM 相比,在归类过程中,只需计算样本到椭圆凸壳中心的距离,能够实现批处理. 此外,由于每类样本的椭圆凸壳仅需存储中心和协方差,可大大节省存储空间.

本文后继章节安排如下:第 1 节简单回顾 LGEPSVM 的工作原理及马氏度量性质;第 2 节详细介绍 MLGEPSVM 的设计方法;第 3 节是实验,以 GEPSVM 和 LGEPSVM 为比较对象,验证 MLGEPSVM 的具体性能;第 4 节总结全文,并规划后续工作方向.

1 相关工作

本节简要回顾一下本文的相关工作,并作符号说明.

1.1 GEPSVM

设两类样本 $\{x_i, y_i\}_{i=1}^n, x_i \in R^d, y_i \in \{-1, 1\}$, 在输入空间上寻找两个超平面,

$$P_i: w_i^T x + b_i = 0, \quad i = 1, 2 \tag{1}$$

式中, (w_i, b_i) 分别表示超平面的法向量和阈值.

以 A, B 记两类训练样本,以 P_1 为例,为使其距 A 类样本尽可能近,而距 B 类样本尽可能远,GEPSVM 的优化目标如下,

$$\min_{(w_1, \gamma_1)} \frac{\|Aw_1 - eb_1\|^2 + \delta \left\| \begin{matrix} w_1 \\ b_1 \end{matrix} \right\|^2}{\|Bw_1 - eb_1\|^2}. \tag{2}$$

该优化问题可通过如下广义特征值问题求解,

$$Gz = \lambda Hz. \tag{3}$$

式中, $G = [A \quad -e]^T [A \quad -e] + \delta I, H = [B \quad -e]^T [B \quad -e], z = (w^T, b)^T, e$ 是全为 1 的列向量. $\delta (>0)$ 是正则化因子. 易知是由广义特征值问题(3)式的最小特征值对应的特征向量决定的. 可同理求解第二个平面.

1.2 LGEPSVM

在 GEPSVM 的基础上,LEGPSVM 的主要任务是分别求解两个超平面的局部区域,即计算出这两类超平面上的凸壳.

为避免常规的二次规划求解,文中给出了一种凸壳求解方法,以期提高凸壳的计算速度. 简述其思想如下,图 1 所示. 图中“.”表示投影样本,记为 $\{\varphi_1, \varphi_2, \dots, \varphi_n\}$, x_1 和 x_2 是距离最远两点, \vec{l} 是 $\overrightarrow{x_1x_2}$ 上的法向量. 以 x_2 为定点,寻找 \vec{l} 与 $\overrightarrow{\varphi_i x_2}$ 中的夹角最小时所对应的投影点 φ_i , 可通过 $\max_{\varphi_i \in \Phi} \arccos(\langle \vec{l}, \overrightarrow{\varphi_i x_2} \rangle / \langle \vec{l}, \overrightarrow{\varphi_i x_2} \rangle)$ 来完成,其中 $\langle \cdot, \cdot \rangle$ 表示内积,满足此条件的 φ_i 作为下一个定点. 重复这一操作,直到 x_2 再一次被选为定点. 所有被选作定点均是凸壳顶点.

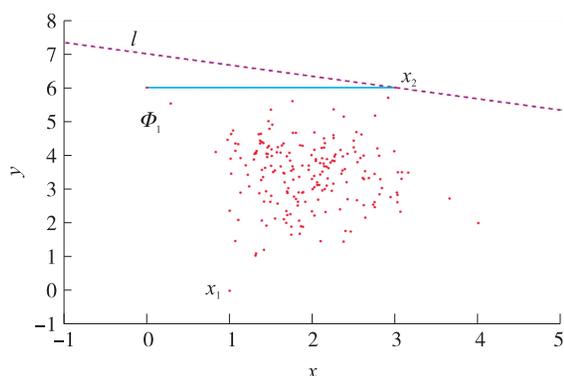


图 1 LGEPSVM 凸壳求解过程

Fig. 1 Illustration of convex hull construction of LGEPSVM

文献[8]的算法迭代是以 x_1 和 x_2 皆为凸壳顶点为前提的,但原文并未证明,本文以定理形式重述此结论,并在附录中给出严格证明.

定理 1 点集中距离最远两点必为凸壳顶点.

1.3 马氏度量特性

如前所述,马氏度量取得了成功应用,为方便下文描述,本节中简介回顾一下相关性质.

定义 1 设 x, y 为空间 R^n 的 d 维向量, Σ 为协方差矩阵,则称

$$\|x - y\|_M = \sqrt{(x - y)^T \Sigma^{-1} (x - y)} \quad (4)$$

为两向量的马氏距离.

由定义知,当 Σ 为单位矩阵时,此时马氏距离则退化为欧式. 文献[17]亦给出了两者的转化关系,以定理 2 记之.

定理 2 同一赋范线性空间中,欧式度量与马氏度量是距离保持的,其转换关系可表述为:

$$\|x - y\|_M = \|\Sigma^{-1/2}(x - y)\|_2. \quad (5)$$

2 MLGEP SVM 算法

以均值为中心,到中心的马氏距离为定长的动点轨迹是椭球面,在二维空间,则退化为椭圆. 三维空间中,椭球面被二维平面所截形成的截痕为一个椭圆. 文献[11]称,经过映射(如核映射)的多数样本更趋向于分布在超椭球体上,并据此认为这是基于马氏度量分类器分类精度较高的原因之一.

2.1 最小椭圆凸壳的计算

如前所述,马氏度量下的凸壳是一个超椭球面,相比于第二节描述 LGEPSVM 的凸壳算法,椭圆凸壳的计算更为简单,只需计算出样本的投影点在超平面的几何中心(即样本均值)和椭圆的半径,如此就可减少训练时间. 具体做法是:以中心为定点,取投影点到中心的最大马氏距离 r 为半径,从而可获得超椭球面. 该思想即为求解最小椭圆凸壳,定理 3 表述如下.

定理 3 设有 d 维离散点集 $X = \{x_1, x_2, \dots, x_n\}$, 记中心 μ 在拟合超平面 $P_1: w^T x + b = 0$ 上的投影为 μ_p , 则以 μ_p 为定点, μ_p 到 X 最大的马氏距离 r 为半径的超椭球体,被 P_1 截得的超椭球面是 P_1 上投影点 $\Phi = \{\varphi_1, \varphi_2, \dots, \varphi_n\}$ 的最小椭圆凸壳.

证明 设 $r = \max_{x_i \in X} \|\mu_p - x_i\|_M$, $r' = \max_{\varphi_i \in \Phi} \|\mu_p - \varphi_i\|_M$, 有

$$\begin{aligned} \forall x_i \in X, r &\geq \|\mu_p - x_i\|_M \\ \forall \varphi_i \in \Phi, r' &\geq \|\mu_p - \varphi_i\|_M \end{aligned} \quad (6)$$

$$r_d = \max_{X_i \in X} \|\mu_p - X_i\|_2, r'_d = \max_{\varphi_i \in \Phi} \|\mu_p - \varphi_i\|_2.$$

由定理 2, 易证样本中距离 μ_p 最远点 $x_m (\in X)$ (长度为 r_d) 在超平面的投影是所有样本投影点中距离 μ_p 的最远点, 长度即为 r'_d , 有 $r_d = \sqrt{r'^2_d + v^2}$, 其中, $v = \|\mu_p - X_m\|_2$. 故, $r_d > r'_d$ 根据定理 2, 结合(6)式, 有 $r > r' = \max_{\varphi_i \in \Phi} \|\mu_p - \varphi_i\|_M$, 则

$$\forall \varphi_i \in \Phi, r \geq \|\mu_p - \varphi_i\|_M. \tag{7}$$

由(7)可知,以 μ_p 为中心, r 为半径的椭圆凸壳,可包含所有的离散点集,且是最小椭圆凸壳. 证毕.

在构造椭圆凸壳具体算法之前,先回顾一下点到超平面之间的马氏距离,以定理 4 记之.

定理 4 马氏度量下,空间中任一点 v 到超平面 $w^T x + b = 0$ 的距离为:

$$s = \frac{|w^T v + b|}{\sqrt{w^T \Sigma w}}. \tag{8}$$

且其在平面上的投影为:

$$v_p = v - \frac{(w^T v + b) \Sigma w}{w^T \Sigma w}. \tag{9}$$

由上述分析,椭圆凸壳算法描述如表 1 所示.

以下介绍待测样本的分类依据,即需判断待测样本与椭圆凸壳的位置关系.

2.2 测试点到凸壳的距离计算

测试点与凸壳共有 3 种关系:(1)点在壳内,(2)点在壳上,(3)点在壳外. 不同位置关系需要采用不同方法计算测试点与凸壳距离. 同时,由于 LGEPSVM 中凸壳构造的实质是使用凸壳顶点的线性组合构造出多边形凸壳,在某些情况下,会造成顶点共线. 在此情况下,凸壳就从凸多边形退化为线段. 而且在位置关系判断时,先设置一个预检测过程,用来判断使用何种方法决定位置关系,如已判好该测试点的投影位于壳外,还必须进一步计算出该点与凸壳边界上的哪个线段距离近并找出距其最近的点,再通过几何关系计算出点到凸壳距离. 然而,在椭圆凸壳判别准则中,无需这一步骤,直接计算点与投影中心的马氏距离即可. 如前所述, LGEPSVM 时间复杂度为 $O(nd)$,而本文的椭圆凸壳仅为 $O(n)$,即由测试点个数决定.

测试点到凸壳的距离计算分两种情况:(1)测试点在壳内或壳上,(2)测试点在椭圆壳外. 情形 1:点到凸壳的距离等于点到超平面之间的马氏距离;情形 2:直接计算点到凸壳的马氏距离. 以上思想以算法形式描述,表 2 所示.

3 实验结果与对比分析

实验数据采用人工数据集和 UCI 标准数据集,在线性核和非线性核上均进行实验. 实验环境: Windows10 操作系统,中央处理器为 Intel Core(TM) i5-4210M 2.60Hz,内 4.00GB. 计算平台为 MATLAB R2016a.

3.1 椭圆凸壳示例

人工数据集是一个两类的三维样本,样本分布如图 2 所示. 由 GEPSVM 产生出两个拟合超平面后,分别计算两个椭圆凸壳(图中所示的两个椭圆),直观地描述出了凸壳与投影点的位置关系,两

表 1 椭圆凸壳构造算法

Table 1 Ellipse convex hull algorithm

输入:离散点集 $X = \{x_1, x_2, \dots, x_n\}$,超平面 $w^T x + b = 0$
输出:椭圆凸壳半径 r ,投影点中心 u_p

- 1.初始化 $r \leftarrow -\infty$, e 为元素全为 1 的向量
- 2.计算 $u \leftarrow X e / n$, $\Sigma \leftarrow \sum_i (x_i - u)(x_i - u)^T / n$
/* 计算均值和协方差
- 3.由(9)计算投影 $u_p \leftarrow u - (w^T u + b) \sum w / w^T \Sigma w$
- 4.FOR x_i IN X DO,
5. $S \leftarrow \sqrt{(x_i - u)^T \Sigma^{-1} (x_i - u)}$
6. IF $s > r$, THEN $r \leftarrow s$, END IF
7. END FOR
8. RETURN u_p, r

表 2 点与凸壳距离计算

Table 2 Computation of distance between data point and convex hull

输入:点集 $X = \{x_1, x_2, \dots, x_n\}$,超平面 $w^T x + b = 0$
输出:点到凸壳距离 d_0

- 1.初始化,计算协方差矩阵 Σ ,中心点 μ ,椭圆凸壳半径 r , $Flag \leftarrow 0$;
- 2.按(9)式计算投影 $x_p \leftarrow x - (w^T x + b) \sum w / w^T \Sigma w$;
- 3.计算 $s \leftarrow \sqrt{(x_p - u)^T \Sigma^{-1} (x_p - u)}$
/* $Flag = -1$ 壳外; $Flag = 0$ 壳上; $Flag = 1$ 壳内 */
4. IF $s > r$ THEN $Flag \leftarrow -1$, END IF
IF $s < r$ THEN $Flag \leftarrow 1$, END IF
5. IF $Flag = 0$ OR $Flag = 1$ THEN
 $d_0 \leftarrow |w^T x + b| / \sqrt{w^T \Sigma w}$
ELSE
 $\mu_p \leftarrow \mu - (w^T \mu + b) \Sigma w / w^T \Sigma w$
 $d_0 \leftarrow \sqrt{(u_p - x)^T \Sigma^{-1} (u_p - x)} - r$
END IF
6. RETURN d_0

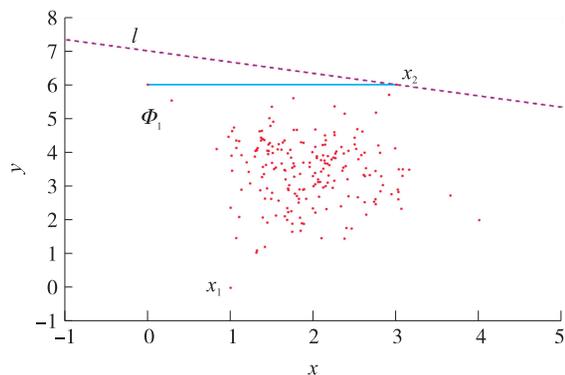


图 2 椭圆凸壳示例

Fig. 2 Illustration of Ellipse convex hull

者均都在同一平面上,验证了定理 3 以及算法 1 的正确性.

3.2 分类精度测试

本节实验数据除 cross2plane 外,其他均采用 UCI 数据,其中 iris12 是 iris 数据集的第一、二两类样本,经 PCA(Principle Component Analysis:主成分分析)降维获得. Cross2plane 数据来自两个相互交叉直线分布^[4],附加了均匀分布的随机噪声,用以验证 GEPSVM 类型分类器解决 XOR 问题的有效性. 实验采用十折交叉验证,通过“分类精度±标准差”报告实验结果,其中两者均是 10 次测试结果的平均值. 为进一步验证 LGEPSVM 和 MLGEPSVM 的差异,对两者的 10 次分类结果进行了 t -检验,显著水平取 0.05,检验统计量采用 p -value 记录,即当 p -value<0.05,表示二者存在显著差异,下表中以“*”标记. 分别采用了线性核和高斯核,分别统计在表 3 和 4 中,其中表 3 是线性核实验结果.

表 3 线性核下 3 种算法的分类精度比较

Table 3 Comparison on accuracy of three linear classifiers

Data	GEPSVM	LGEPSVM	MLGEPSVM	P-value
Cross2plane 200×2	99.50±1.12	99.50±1.12	99.50±1.58	0.088 1
Monk2 432×6	56.28±6.23	69.17±6.41	80.00±16.25	0.073 5
Monk3 432×6	80.38±3.97	80.73±4.41	81.00±18.23	0.964 9
Muskclean 6 598×166	38.79±1.52	51.09±1.20	81.34±20.33	0.008 8*
Waveform 5 000×21	60.62±0.56	59.66±0.52	91.00±1.29	6.074 0E-17*
Pimadata 768×8	70.71±3.44	68.37±0.21	71.54±6.14	0.183 4
Iris12 100×2	100.00±0.00	100.00±0.00	98.00±4.22	0.167 9

在表 3 中,数据集大小为 $n \times m$,其中 n 为样本数, m 为样本维度,粗体数据为 3 种分类器中的最高分类精度. MLGEPSVM 不仅能够保持与 LGEPSVM 相当分类精度,甚至在一些样本上分类性能要优于 LGEPSVM 和 GEPSVM. 在 waveform 数据集上,MLGEPSVM 的分类准确率相比较于其他两者,有相当大幅度的提升. 说明在线性核中,MLGEPSVM 相较于 LGEPSVM,有着更加优越的分类性能. 大多数数据,3 个分类器的分类精度相当,说明 MLGEPSVM 在提升 LGEPSVM 性能同时,继承了 GEPSVM 原有的优点. 然而,表 3 同时也反映出 MLGEPSVM 的标准差较大,表现出算法存在不稳定性,可能是由于此时协方差阵加了扰动以达到矩阵正定所致. 例如在 monk3 数据集中,三者性能相似,但 MLGEPSVM 标准差远高于其他两者. 图 3 直观地展示了 3 种分类器的分类效果.

表 4 是 3 种方法在 UCI 数据上的实验结果. 采用经验核映射,选择高斯核,核参数在 $\{10^i \mid i = -1, -2, -3\}$ 范围内选取,按文献[4,8],在训练集中随机取 5%作为验证集,用于调整参数,最终保存验证集上分类精度的最高时对应的参数为本轮验证参数,重新训练分类器,并据此获得分类精度. 如表 4 所示,MLGEPSVM 较之 GEPSVM 和 LGEPSVM 表现不俗,尤其是数据集 Checkdata 中较为明显. 值得注意的是,在高斯核上,MLGEPSVM 在 Waterdata 的准确率虽低于 GEPSVM,但在线性核上却表现优异,这也再次验证了核选择能够直接影响到分类精度.

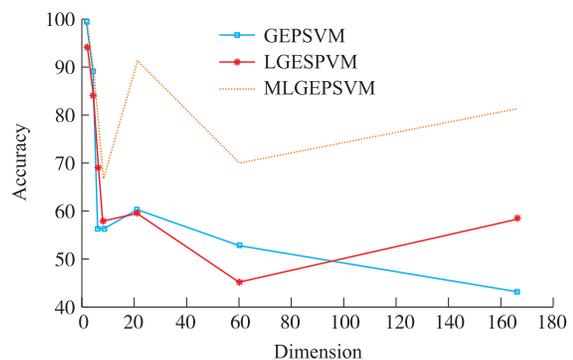


图 3 3 种分类器在不同维度数据集的比较

Fig. 3 Accuracy as dimension increasing in classifiers

表 4 高斯核下 3 种算法的分类精度比较

Table 4 Comparison three classifiers in Gaussian Kernel

Data	GEPSVM	LGEPSVM	MLGEPSVM
BUPA 345×6	63.41±3.29	64.39±2.19	67.85±12.59
Water 116×38	68.22±8.15	66.49±10.20	59.64±11.30
Checkdata1000×2	55.61±5.33	55.88±4.76	76.77±14.52
Pima 768×8	45.08±4.23	52.96±2.97	54.04±10.39
WPBC 194×32	64.32±5.83	74.21±3.90	52.50±5.27

如前文所述, MLGEPsvm 的一大特点或是对 LGEPsvm 改进的方面就是在分类时间上. LGEPsvm 分类性能上的提升需要以时间为代价^[8], 这也是 MLGEPsvm 需要突破的地方. 本实验使用 UCI 数据集 muskclean, 该数据集有两类样本点, 6598 个样本, 166 维. 在交叉验证过程中, 先对每折的 10 次验证计算平均时间, 再对十折的总时间再取一次平均, 得出最终的分类时间. 实验使用 matlab 命令 tic 和 toc 记录训练时间和分类时间, tic 为开始计时, 并在 toc 处结束计时.

由表 5 知, LGEPsvm 的分类时间约为 GEPsvm 的 3.6 倍, 时间成本较高, 而 MLGEPsvm 的分类时间仅为 0.624 s, 在运算速度上相比于 LGEPsvm 提升了 1.87 倍. 由表 3 知, MLGEPsvm 在 muskclean 上的分类精度远高于 GEPsvm 和 LGEPsvm. 由于 MLGEPsvm 与 LGEPsvm 相似, 在分类中仍然需要针对每一个测试点在判别过程中计算一次位置关系, 导致无法进行批量运算, 所以分类速度上仍比 GEPsvm 慢. 但是在空间上, MLGEPsvm 相较于 LGEPsvm 和 GEPsvm 的开销更小.

4 结论

本文基于马氏度量提出一种新的最小凸壳算法, 简化 LGEPsvm 的分类规则, 并给予相关证明. 实验得出 MLGEPsvm 分类性能不仅与原 LGEPsvm 相当, 而且在某些数据集上表现更加优越. 不仅如此, 较之 LGEPsvm, MLGEPsvm 有着更快的分类速度. 同时, 由于椭圆凸壳只需存储样本均值和协方差, 有利于降低存储开销. 由实验结果知, MLGEPsvm 分类过程中存在较大标准差问题, 我们需要在后继工作中加以改进, 根据文献[18]提到的一种凸壳收缩方法, 希望能够据此在已有凸壳算法基础上, 进一步完善凸壳的设计方法及拓宽到更多的应用领域.

表 5 GEPsvm, LGEPsvm, 以及 MLGEPsvm 时间比较
Table 5 Time comparison on GEPsvm, LGEPsvm and MLGEPsvm on muskclean data

算法	平均分类时间	分类精度±标准差
GEPsvm	0.3197	38.79±1.52
LGEPsvm	1.1677	51.09±1.20
MLGEPsvm	0.6240	81.34±20.33

[参考文献]

[1] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.

[2] LEE Y J, MANGASARIAN O L. RSVM: Reduced support vector machines[C]//Proceedings of the 2001 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. Chicago: Wellesey-Cambridge Press, 2001.

[3] MANGASARIAN O L, WILD E W. Proximal support vector machine classifiers[C]//Proceedings KDD-2001: Knowledge Discovery and Data Mining. San Francisco: ACM Press, 2001.

[4] MANGASARIAN O L, WILD E W. Multisurface proximal support vector machine classification via generalized eigenvalues[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(1): 69-74.

[5] KHEMCHANDANI R, CHANDRA S. Twin support vector machines for pattern classification[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(5): 905-910.

[6] YE Q, ZHAO C, ZHANG H, et al. Distance difference and linear programming nonparallel plane classifier[J]. Expert systems with applications, 2011, 38(8): 9425-9433.

[7] SHAO Y H, DENG N Y, CHEN W J. A proximal classifier with consistency[J]. Knowledge-based systems, 2013, 49(49): 171-178.

[8] 杨绪兵, 陈松灿, 杨益民. 局部化的广义特征值最接近支持向量机[J]. 计算机学报, 2007, 30(8): 1227-1234.

[9] YE Q, ZHAO C, YE N, et al. Localized twin SVM via convex minimization[J]. Neurocomputing, 2011, 74(4): 580-587.

[10] HUANG H, WEI X, ZHOU Y. Twin support vector machines: a survey[J]. Neurocomputing, 2018, 300: 34-43.

[11] 张凯军, 梁循. 马氏距离多核支持向量机学习模型[J]. 计算机工程, 2014, 40(6): 219-225.

[12] ROTH P M, HIRZER M, KÖSTINGER M, et al. Mahalanobis distance learning for person re-identification[C]//Person re-identification. London: Springer Press, 2014.

[13] BEHERA S K, DOGRA D P, ROY P P. Fast recognition and verification of 3D air signatures using convex hulls[J]. Expert systems with applications, 2018, 100: 106-119.

- [14] GARCÍA S I D, PAJARES G. On-line crop/weed discrimination through the Mahalanobis distance from images in maize fields[J]. *Biosystems engineering*, 2018, 166:28–43.
- [15] DE LA HERMOSA GONZÁLEZ, R R. Wind farm monitoring using Mahalanobis distance and fuzzy clustering[J]. *Renewable energy*, 2018, 123:526–540.
- [16] SUO M, ZHU B, ZHANG Y, et al. Fuzzy Bayes risk based on Mahalanobis distance and Gaussian kernel for weight assignment in labeled multiple attribute decision making[J]. *Knowledge-based systems*, 2018, 152:26–39.
- [17] 杨绪兵, 王一雄, 陈斌. 马氏度量学习中的几个关键问题研究及几何解释[J]. *南京大学学报(自然科学版)*, 2013, 49(2):133–141.
- [18] BENNETT K P, BREDENSTEINER E J. Duality and geometry in SVM classifiers[C]//Seventeenth International Conference on Machine Learning. Stanford: Morgan Kaufmann Publishers Inc. 2000.
- [19] CHAU A L, LI X, YU W. Convex and concave hulls for classification with support vector machine[J]. *Neurocomputing*, 2013, 122:198–209.
- [20] SHANG J, CHEN M, ZHANG H. Fault detection based on augmented kernel Mahalanobis distance for nonlinear dynamic processes[J]. *Computers & chemical engineering*, 2018, 109:311–321.
- [21] VERMA N, BRANSON K. Sample complexity of learning mahalanobis distance metrics[J]. *Computer science*, 2015:2584–2592.
- [22] MAHALANOBIS P C. On the generalized distance in statistics[J]. *National institute of science of India*, 1936, 2:49–55.
- [23] DAI H. *Theory of matrices*[M]. Beijing: Science Press, 2001.
- [24] BIAN Z Q, ZHANG X G. *Pattern recognition*[M]. Beijing: Tsinghua Press, 2005.
- [25] XU H L, LONG G Z, BIE X F, et al. Active learning algorithm of SVM combining tri-training semi-supervised learning and convex-hull vector[J]. *Pattern recognition and artificial intelligence*, 2016, 1:39–46.
- [26] REN D W, HU Z P. Construction of classifiers of approximative convex hull with reduced dimension[J]. *Mathematics in practice and theory*, 2014(18):166–174.
- [27] WANG W B. *The research of SVM classifier method based on Gilbert algorithm and scaled convex hulls*[D]. Fuzhou: Fuzhou University, 2014.
- [28] ZHANG X K. *Research on high-dimensional data visualization methods and visualization classification techniques* (Doctoral dissertation)[M]. Harbin: Harbin Institution of Technology, 2013.
- [29] CAI G X. *Research and application of convex hull support vector machine based on two-phase method*[J]. *Coal technology*, 2013, 32(5):200–202.
- [30] 杨铭, 赵文雨. 基于内点法的凸壳算法在支持向量中的应用[J]. *信息与电脑(理论版)*, 2013(4):144–145.
- [31] YAN J, BI S B, WANG D, et al. Parallel algorithm for computing convex hulls in multi-processor architecture[J]. *Computer science*, 2013, 40(2):16–19.
- [32] HU Z P, LU L, FENG C S. Research on classification algorithm based on high-dimensional data description[J]. *Journal of Yanshan university*, 2011, 35(4):370–376.
- [33] LIU Z B, CHEN Z, LIU J G. A novel geometric nearest point algorithm for constructing SVM classifiers[J]. *Acta automatica sinica*, 2010, 36(6):791–797.

[责任编辑:黄敏]