文章编号: 1001-0920 (2015) 07-1298-05 DOI: 10.13195/j.kzyjc.2014.0815

聚类分片双支持向量域分类器

梁锦锦[1], 吴 德[2]

(1. 西安石油大学 理学院，西安 710065；2. 西安电子科技大学 计算机学院，西安 710071)

摘 要: 针对支持向量域分类器对大规模样本集的训练时间长且占用内存大的问题, 构造聚类分片双支持向量域分类器. 以均值聚类剖分原始空间, 并选取密度指标大的样本作为初始聚类中心; 对子空间构造双支持向量域分类器, 根据样本与正负类最小包围超球的距离构造分段决策函数; 定义样本的变尺度距离, 以链接规则组合子空间的分类结果. 数值实验表明, 所提出算法的分类精度高且受参数变化的影响不大, 分类时间短且随子空间数的增加而降低.

关键词: 支持向量域分类；分段识别；聚类；密度指标；双支持向量域分类器；变尺度距离

中图分类号: TP301 **文献标志码:** A

Clustering piecewise double support vector domain classifier

LIANG Jin-jin[1], *WU De*[2]

(1. School of Mathematical Sciences，Xi'an Shiyou University，Xi'an 710065，China；2. School of Computer Sciences，Xidian University，Xi'an 710071，China. Correspondent：LIANG Jin-jin，E-mail：myonlyonly@126.com)

Abstract: Support vector domain classifiers have disadvantages like long training time and large memory. The clustering piecewise double support vector domain classifier(CPDSVDC) is proposed. CPDSVDC uses C means algorithm to partition the original space, and selects the initial cluster centers by samples with large density indexes. The dual support vector domain classifier is constructed in each divided subspace, and the corresponding piecewise decision function is also constructed based on the position relationship between the test sample and the two minimum enclosing spheres. The variable distance of the test sample is defined, and linking rule is used to combine classification results in all subspaces. Numerical experiments demonstrate that the CPDSVDC has high classification accuracy that varies slightly with parameters and low training time that decreases with the number of subspaces.

Keywords: support vector domain classifier；piecewise identification；clustering；density indexes；double support vector domain classifier；variable distance

0 引 言

Tax 等[1]提出的支持向量域描述 (SVDD) 已得到广泛应用, 在故障检测[2]、构造集成算法[3]或是进行特征选择[4]中显示出了极大的优越性. 有关 SVDD 的分类算法[5-10]大量涌现: 文献 [5] 根据样本的描述边界构造支持向量域分类器 (SVDC), 降低了分类时间, 但分类精度较低且对参数变化敏感; 文献 [6] 提出了空间支持向量域分类器 (SSVDC), 在一定程度上减弱了分类精度受参数变化的影响; 文献 [7] 根据样本在核空间的位置分布构造了可变的惩罚参数, 提高了分类精度; 文献 [8] 提出了支持向量域多分类器. 这类研究取得了进展[5-8], 却对大规模样本集存在内存占用大和训练时间长的问题. 文献 [9] 寻求 SVDD 特征空间中支持向量的基函数, 提高了测试速度, 却没有克服分类精度依赖参数取值的弊端; 文献 [10] 提出了信赖支持向量域描述 (CSVDD), 依据抽样比例选取信赖度量较大的部分样本训练, 训练样本规模缩减的幅度依赖于抽样比例, 且分类精度依赖于参数取值.

如果能构造出一种分类算法, 既能缩小求解凸二次规划的规模, 又能保证分类精度基本不受参数变化影响, 则可将其应用于大规模训练样本集, 并提高在各个领域的运算效率. 已有学者将分段识别的思想应用于支持向量机 (SVM), 并取得了良好的结果[11-12]. 笔者将分段识别的思想推广应用于 SVDD, 提出聚类分片双支持向量域分类器 (CPDSVDC). CPDSVDC 首先运用 C 均值 (CM) 算法划分原始空间, 定义样本的密度指标, 并选取该值较大的若干样本作为初始聚类中心; 然后构造双支持向量域分类器训练子空间中

收稿日期: 2014-05-22；**修回日期:** 2014-08-25.

基金项目: 国家自然科学基金项目(61373174).

作者简介: 梁锦锦(1983−), 女, 讲师, 博士, 从事最优化理论与算法、数据挖掘与支持向量机的研究；吴德(1979−), 男, 高级工程师, 博士生, 从事网络安全与主机审计的研究.

的样本, 以一个分段函数作为分类决策函数来预测该子空间内待测样本的类别指标; 最后设定链接规则, 合并所有子空间的分类结果. CPDSVDC 每次仅对子空间中的样本进行训练, 具有低的复杂度和短的分类时间; CPDSVDC 根据样本所处的子空间而采取分段分类决策函数, 具有较好的鲁棒性. 不同规模数据集上的数值实验表明了 CPDSVDC 相比于 SVM、SVDC、SSVDC 和 CSVDD 的优越性.

1　聚类分片双支持向量域分类器

记训练集为

$$T=\{(x_i,y_i)\}_{i=1}^{l},\ x_i\in R^n,\ y_i\in\{1,-1\}).$$

其中: $x_i\in R^n$ 为 n 维空间的训练样本, $y_i\in\{1,-1\}$ 为 x_i 的类别指标. CPDSVDC 依次通过 C 均值聚类分片、双支持向量域分类器训练和链接规则设置三步构造分类决策函数. 这里分片意味着将训练空间剖分为若干个子空间, 双意味着对子空间中的正负类样本进行 SVDD 训练以构造两个最小包围超球.

1.1　聚类分片

C 均值聚类作为一种基本的聚类划分方法, 具有简单、快速的优点, 但却依赖于初始聚类中心的选择.

1.1.1　密度指标

去掉 T 的类别指标得到 $X=\{x_1,x_2,\cdots,x_l\}$, 定义

$$P_i^0=\sum_{j=1}^{l}\exp\Big[-\frac{||x_i-x_j||^2}{(r_a/2)^2}\Big] \tag{1}$$

为样本 x_i 的密度指标. 可见, 样本 x_j 与 x_i 的距离越近, 其对样本 x_i 的密度指标 P_i^0 的贡献越大, 即 x_i 周围聚集的样本点越多, 其密度指标值越大.

式 (1) 中邻域半径 r_a 是一个正常数, r_a 之外的数据点对密度指标的影响很小, r_a 的值应与样本的分布特性有关, 可以采用下式计算:

$$r_a=\frac{1}{2}\sqrt{\frac{1}{n(n-1)}\sum_{k=1}^{c}\sum_{i=1}^{l}||x_i-x_k^*||^2}. \tag{2}$$

其中: c 为聚类中心的个数, 由 CM 聚类算法中待划分的子空间数决定; $x_k^*\,(k=1,2,\cdots,c)$ 依次为第 $k\,(k=1,2,\cdots,c)$ 个初始聚类中心. 文献 [13] 中选取初始聚类中心的方法取得了良好的效果, 本文借鉴该文献中的方法选取 $x_k^*\,(k=1,2,\cdots,c)$. 令

$$P_1^*=\max\{P_i^0;i=1,2,\cdots,l\}, \tag{3}$$

同时取对应的 x_1^* 为第 1 个初始聚类中心, 调整后续样本密度指标的关系式为

$$P_i^{(k)}=P_i^{(k-1)}-P_k^*\exp\Big[-\frac{||x_i-x_j||^2}{(r_a/2)^2}\Big],\quad k=1,2,\cdots,c, \tag{4}$$

其中 $P_k^*=\max\{P_i^{(k-1)};i=1,2,\cdots,l\}$, 对应的样本点 x_k^* 取为第 k 个初始聚类中心.

1.1.2　均值聚类

CM 聚类的目的是将集合 X 的所有样本划分为互不相交的 c 个子空间 $X_i\,(i=1,2,\cdots,c)$, 满足

$$X=\bigcup_{i=1}^{c}X_i,\ X_i\bigcap X_j=\phi,\ i\neq j, \tag{5}$$

且使各个样本与其所在子空间的中心的误差平方和最小, 即

$$\begin{aligned}&\min\ P(U,Z)=\sum_{i=1}^{c}\sum_{x_k\in X_i}u_{ij}d_{ij}^2;\\&\text{s.t.}\ \sum_{i=1}^{c}u_{ij}=1,\ \forall j=1,2,\cdots,n.\end{aligned} \tag{6}$$

其中: $U=(u_{ij})_{c\times l}$ 为隶属矩阵, $u_{ij}\in\{0,1\}$ 代表第 j 个样本 x_j 隶属第 i 个子空间 X_i 的程度, 若 $u_{ij}=1$, 则 x_j 被分配至 X_i, 若 $u_{ij}=0$, 则 x_j 未被分配至 X_i; $Z=\{z_1,z_2,\cdots,z_c\}$ 为 c 个子空间中心的集合; $d_{ij}=||x_j-z_i||$ 为样本 x_j 与聚类中心 z_i 的距离.

CM 聚类选取密度指标较大的 c 个样本作为初始聚类中心, 根据其余样本与各个子空间中心的距离将其分配到最近的子空间, 求解新形成子空间的中心, 重复此迭代过程, 直到目标函数最小化.

1.2　双支持向量域分类器训练

1.2.1　支持向量域描述

给定样本集 $\{x_t\}_{t=1}^{p}\,(x_t\in R^n)$. 记惩罚因子为 $0<C\in R^1$, 样本 x_t 的松弛为 $0\leqslant\xi_t\in R^1$, 非线性映射为 $\phi:x\to\phi(x)$. SVDD 求解如下规划, 得到半径为 R、球心为 a 的最小包围超球 (R,a):

$$\begin{aligned}&\min_{R,\xi_t}R^2+C\sum_{t=1}^{p}\xi_t;\\&\text{s.t.}\ [\phi(x_t)-a][\phi(x_t)-a]^{\mathrm{T}}\leqslant R^2+\xi_t,\\&\qquad 0\leqslant\xi_t,\ t=1,2,\cdots,p.\end{aligned} \tag{7}$$

通过求解如下对偶规划得到规划 (7) 的最优解:

$$\begin{aligned}&\max_{\alpha}\sum_{t=1}^{p}\alpha_tK(x_t,x_t)-\sum_{t,s=1}^{p}\alpha_t\alpha_sK(x_t,x_s);\\&\text{s.t.}\ \sum_{t=1}^{p}\alpha_t=1,\ 0\leqslant\alpha_t\leqslant C.\end{aligned} \tag{8}$$

其中 $K(x_t,x_s)=\phi(x_t)\cdot\phi(x_s)$ 为核函数. 记规划 (8) 的最优解为 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_p^*)$, 若样本 z 距球心 a 的距离小于半径 R, 则接受该样本. 接受条件为

$$||\phi(z)-a||^2=K(z,z)-2\sum_{t=1}^{p}\alpha_t^*K(z,x_t)+\sum_{t,s=1}^{p}\alpha_t^*\alpha_s^*K(x_t,x_s)\leqslant R^2. \tag{9}$$

1.2.2　双支持向量域分类器

记 $X_i=\{(x_k(i)\}_{k=i_1}^{i_s}\,(i_1<i_2<\cdots<i_s)$ 为第 i 个子空间样本. 根据类别指标 $y_k(i)=1$ 提取对应的正类样本 $X_i^+=\{(x_k(i)\}_{k=i_1}^{i_{s^+}}$, 并根据类别指标 $y_k(i)=$

-1 提取对应的负类样本 $X_i^- = \{(x_k(i)\}_{k=i_{s^+}+1}^{i_s}$. 训练双支持向量域分类器即为求解如下两个凸二次规划:

$$\begin{aligned}&\min_{R^+(i),\xi_k(i)}[R^+(i)]^2 + C\sum_{k=i_1}^{i_{s^+}}\xi_k(i);\\&\text{s.t. } [\phi(x_k(i)) - a^+(i)]^{\mathrm{T}}[\phi(x_k(i)) - a^+(i)] \leqslant\\&\quad [R^+(i)]^2 + \xi_k(i),\\&\quad \xi_k(i) \geqslant 0,\ k = i_1, i_2, \cdots, i_{s^+}.\end{aligned} \tag{10}$$

$$\begin{aligned}&\min_{R^-(i),\xi_k(i)}[R^-(i)]^2 + C\sum_{k=i_{s^+}+1}^{i_s}\xi_k(i);\\&\text{s.t. } [\phi(x_k(i)) - a^-(i)]^{\mathrm{T}}[\phi(x_k(i)) - a^-(i)] \leqslant\\&\quad [R^-(i)]^2 + \xi_k(i),\\&\quad \xi_k(i) \geqslant 0,\ k = i_{s^+}+1, i_{s^+}+2, \cdots, i_s.\end{aligned} \tag{11}$$

其中: $R^+(i)$ 和 $a^+(i)$ 为正类样本最小包围超球 $S_1(i)$ 的半径和球心, $R^-(i)$ 和 $a^-(i)$ 为负类样本最小包围超球 $S_2(i)$ 的半径和球心; $\xi_k(i)$ 为样本 $x_k(i)$ 的松弛. 样本 z 与正类和负类两个超球球心的距离分别通过如下两式计算:

$$\begin{aligned}d_1(i) = D(z, S_1(i)) = ||\phi(z) - a^+(i)||^2 =\\ K(z,z) - 2\sum_{k=i_1}^{i_{s^+}}\alpha_k^*(i)K(z, x_k(i)) +\\ \sum_{p,q=i_1}^{i_{s^+}}\alpha_p^*(i)\alpha_q^*(i)K(x_p(i), x_q(i)) \leqslant\\ R^+(i)^2,\end{aligned} \tag{12}$$

$$\begin{aligned}d_2(i) = D(z, S_2(i)) = ||\phi(z) - a^-(i)||^2 =\\ K(z,z) - 2\sum_{k=i_{s^+}+1}^{i_s}\alpha_k^*(i)K(z, x_k(i)) +\\ \sum_{p,q=i_{s^+}+1}^{i_s}\alpha_p^*(i)\alpha_q^*(i)K(x_p(i), x_q(i)) \leqslant\\ R^-(i)^2.\end{aligned} \tag{13}$$

给定待测样本 z, 双支持向量域分类器以如下分段函数作为分类决策函数来判断 z 的类别指标:

$$f_z(i) = \begin{cases}\mathrm{KNN}(z, x_t),\ z, x_t \in I;\\ \arg\min\limits_{m=1}^{2}(D(z, S_m) - R_m^2),\ z \in M.\end{cases} \tag{14}$$

其中

$$\begin{aligned}&R_1 = R^+(i),\ R_2 = R^-(i),\\&S_1 = (R^+(i), a^+(i)),\ S_2 = (R^-(i), a^-(i)).\end{aligned}$$

当测试样本 z 位于区域 I 时, 以区域 I 中的样本为参考点, 采用 KNN 判断其类别指标; 当测试样本 z 位于区域 M 时, 以其到两个最小包围超球边界的距离最小值来判断类别指标. 区域 I 和 M 中的样本由如下两条准则界定:

1) 如果 $d_1(i) \leqslant R_1$ 且 $R_2/d_2(i) > 1$, 或者 $d_2(i) \leqslant R_2$ 且 $R_1/d_1(i) > 1$, 则 z 处于区域 M;

2) 如果 $d_1(i) \leqslant R_1$ 且 $d_2(i) \leqslant R_2$, 或者 $d_1(i) \geqslant R_1$ 且 $d_2(i) \geqslant R_2$, 则 z 处于区域 I.

1.3 链接规则

双支持向量域分类器共得到 c 个分类决策函数 $f_z(i)\ (i = 1, 2, \cdots, c)$, 如何链接这些结果对待测样本进行识别是取得好的分类性能的一个重要环节.

由于训练是分子空间进行的, 一种自然的想法就是在测试时亦如此: 根据待测样本所处子空间的不同, 选取相应的分类准则, 即令决策函数为

$$f_z(j) = \sum_{i=1}^{c} I_{X_i}(z) f_z(i). \tag{15}$$

其中: $f_z(i)$ 为式 (14) 的决策函数, $I_{X_i}(z)$ 为示性函数

$$I_{X_i}(z) = \begin{cases}1,\ z \in X_i;\\ 0,\ z \notin X_i.\end{cases} \tag{16}$$

对待测样本 z, 定义其与 c 个子空间的聚类中心 $v_i\ (i = 1, 2, \cdots, c)$ 的变尺度距离为

$$d_i = ||z - v_i||/r_i,\ i = 1, 2, \cdots, c. \tag{17}$$

这里, 聚类半径 r_i 的引入是为了平衡不同大小的聚类之间的影响, 由第 i 个子空间 X_i 中的样本与该子空间的聚类中心 v_i 的最大距离确定, 即

$$r_i = \max_{z_n \in X_i}||z_n - v_i||,\ i = 1, 2, \cdots, c. \tag{18}$$

比较得到最小的 d_i 所对应的脚标值 $j = \arg(\min d_i)$, 并将 z 归类于第 j 个子空间 X_j.

1.4 复杂度分析

给定含有 l 个样本的二分类训练集 T. 不妨假设正负类样本数目同为 $l/2$, 且经 CM 聚类为 c 个子空间后, 各个子空间的样本数亦同为 l/c.

SVM 对整个训练集训练, 空间和时间复杂度为 $O(l^2)$ 和 $O(l^3)$; SVDC 仅对正或负类样本训练, 空间和时间复杂度为 $O(l^2/4)$ 和 $O(l^3/8)$; CPDSVDC 先对子空间解耦求解, 子空间的空间和时间复杂度为 $O((l/c)^2)$ 和 $O((l/c)^3)$, 再链接分类结果, 总的空间和时间复杂度为 $O(l^2/c)$ 和 $O(l^3/c^2)$.

显然, 当 $c \geqslant 4$ 时, 3 种分类器同时满足 $O(l^2/c) \leqslant O(l^2/4) \leqslant O(l^2)$ 和 $O(l^3/c^2) \leqslant O(l^3/8) \leqslant O(l^3)$, 即 CPDSVDC 的复杂度最低, SVDC 次之, SVM 的复杂度最高.

2 数值实验

为了验证 CPDSVDC 的性能, 选取不同规模的二分类基准数据集和正态分布数据集进行实验. 所有实验均在 CPU 为 P4, 3.06 GHz, 内存为 0.99 GB 的 PC 机上进行; 所有实验均采取高斯径向基核函数 $K(x, y) = \exp(-||x - y||^2/\sigma^2)$; 所有程序均采用 Matlab 7.01 编写. 首先阐明 CPDSVDC 的分类表现受子空间数 c 和最近邻个数 k 的影响; 然后对比 CPDSVDC 与已有算法在大规模数据集上的性能优劣.

例 1 小规模训练和测试数据集 Diabetics.

Diabetics 为含有 768 个样本的 8 维数据集. 随机选取 468 个数据参与训练, 其余 300 个参与测试. 设定惩罚参数 $C = 1$, 径向基核参数 $\sigma = 0.5$, 最近邻个数 $k = 1$. 在 10 次随机抽取实验下, CPDSVDC 随子空间数 c 的变化所得到的分类表现如表 1 所示.

表 1　不同算法的分类表现

算法	子空间数	分类精度/%	分类时间/s
SVM	$c = 1$	73.67	54.35
SVDC	$c = 1$	66.67	17.26
CPDSVDC	$c = 1$	73.35	31.87
	$c = 2$	75.84	24.32
	$c = 3$	76.19	12.93
	$c = 4$	74.07	7.19
	$c = 5$	69.93	5.05
	$c = 6$	63.56	2.41

表 1 中, 分类精度是训练和测试精度的平均值, 分类时间是训练和测试时间的总和. 由表 1 可知: 1) CPDSVDC 具有与 SVM 相当, 且优于 SVDC 的分类精度; 当 $c = 3$ 时, CPDSVDC 比 SVM 和 SVDC 的分类精度分别高出 2.66 % 和 9.58 %. 2) CPDSVDC 的分类精度先随 c 的增加而增加, 后随其增加而减少; 子空间数 c 较大时, 分类时间缩短的优势更为明显, 如 $c = 6$ 时, CPDSVDC 的分类时间为 2.41 s, 仅是 SVM 和 SVDC 分类时间的 4.43 % 和 13.96 %.

子空间数 c 需要根据实际情况取值, 以便在高的分类精度和短的分类时间之间取得好的折衷. 在本例中, 当子空间数 $c = 3$ 时, CPDSVDC 的性能最优. 其对不同核参数及最近邻个数的性能如表 2 所示.

表 2　CPDSVDC 的性能表现

核参数	$k = 1$		$k = 3$		$k = 5$	
	精度/%	时间/s	精度/%	时间/s	精度/%	时间/s
$\sigma = 0.1$	74.98	25.87	75.67	25.91	73.53	25.93
$\sigma = 0.2$	74.98	19.93	75.67	19.68	75.33	19.73
$\sigma = 0.3$	75.37	14.03	76.33	14.37	75.81	14.51
$\sigma = 0.8$	75.18	11.06	76.29	11.17	76.05	11.29
$\sigma = 1$	74.97	9.93	75.72	10.07	75.72	10.28

核参数	$k = 7$		$k = 9$	
	精度/%	时间/s	精度/%	时间/s
$\sigma = 0.1$	75.33	26.17	75.33	26.56
$\sigma = 0.2$	75.33	19.95	73.53	19.98
$\sigma = 0.3$	75.46	14.67	73.53	14.65
$\sigma = 0.8$	75.33	11.33	73.46	11.37
$\sigma = 1$	75.05	10.28	74.44	10.31

由表 2 可知: 1) CPDSVDC 的分类时间随最近邻个数 k 的增加而产生的变化很微小, 这说明分类时间主要由子空间数 c 决定; CPDSVDC 的分类时间随径向基核参数 σ 的增加而逐步减小, 这是由于核参数取值过小会导致过拟合, 从而增加分类时间. 2) CPDSVDC 的分类精度随径向基核参数 σ 的增加而变化的范围很微小, 以 $k = 5$ 为例, 两个过程分类精度的变化范围不超过 1 %, 这说明 CPDSVDC 具有好的鲁棒性, 其分类精度基本不受参数变化的影响.

对 Banana 和 Pima Indians 数据分别进行实验, CPDSVDC 的分类精度和时间随子空间数 c 和最近邻个数 k 的变化亦有类似结论. 一般而言, 根据经验取 $c \in [2, 5]$ 和 $k = 3$ 可得到好的分类表现.

例 2　大规模训练和测试数据集 Image.

Image 数据集共含有 2 310 个 18 维数据, 随机选取 1 300 个作为训练集, 其余 1 000 个作为测试集. 取 CPDSVDC 的子空间数为 $c = 3$, 取 CPDSVDC 和 SSVDC 的最近邻个数为 $k = 3$, 取 CSVDD 中抽样比例参数为 $\varepsilon = 0.3$; 对 SVM、SVDC、SSVDC、CSVDD 和 CPDSVDC 设定相同的惩罚参数 $C = 1$. 分类精度和分类时间的变化趋势分别如图 1 和图 2 所示.

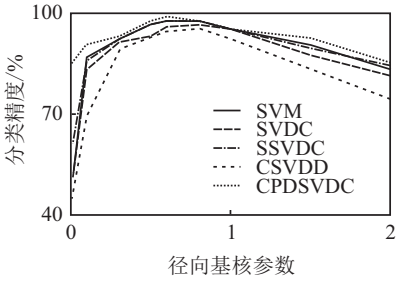

图 1　分类精度随径向基核参数的变化曲线

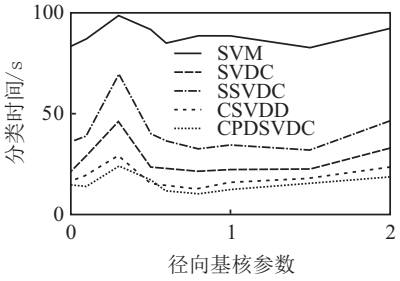

图 2　分类时间随径向基核参数的变化曲线

从图 1 可以看出: CPDSVDC 的分类精度最高, SVM 和 SSVDC 的分类精度次之, SVDC 的分类精度略低, CSVDD 的分类精度最低; CPDSVDC 的鲁棒性最好, 其分类精度几乎不随径向基核参数变化, SSVDC 的鲁棒性次之, SVM、SSVDC 和 CSVDD 的鲁棒性差, 分类精度依赖径向基核参数的取值.

从图 2 可以看出: CPDSVDC 的分类时间最低, CSVDD 的分类时间次之, SVDC 的分类时间略低, SSVDC 的分类时间较高, SVM 的分类时间最高.

例 3　正态分布数据集.

分别产生正负类样本数目均等的正态分布数据 2 000、4 000、10 000 和 20 000 个, 随机交换 5 % 样本的类别指标, 并选取 50 % 的样本参与训练. 选取标准 SVDC 和具有较短训练时间的 CSVDD (以正类样本作为目标类), 以具有较好鲁棒性的 SSVDC 作为参照, 对比给出 CPDSVDC 的分类表现. 对所有算法设置相同的惩罚参数 $C = 1$ 和核参数 $\sigma = 1$, 取 CPDSVDC 的子空间数为 $c = 3$, 取 CPDSVDC 和

SSVDC的最近邻个数为$k = 3$, 取CSVDD的抽样比例参数为$\varepsilon = 0.3$. 不同算法的分类性能如表3所示, 其中"/ "表示由于内存溢出无法运算.

表 3 大规模样本集上的分类性能

规模	SVDC		SSVDC	
	精度/%	时间/s	精度/%	时间/s
2 000	91.23	54.92	93.44	239.84
4 000	92.29	221.57	94.55	779.12
10 000	/	/	/	/
20 000	/	/	/	/
规模	CSVDD		CPDSVDC	
	精度/%	时间/s	精度/%	时间/s
2 000	88.67	32.35	95.25	21.16
4 000	92.12	122.16	96.83	59.07
10 000	93.57	610.28	96.83	257.93
20 000	/	/	95.01	799.30

由表3可知: 1) CPDSVDC具有最高的分类精度. 对不同规模的训练样本集, CPDSVDC将SVDC的分类精度提高了4.02%～4.54%, 将SSVDC的分类精度提高了1.81%～2.28%, 将CSVDD的分类精度提高了3.26%～6.58%. 2) CPDSVDC具有最短的分类时间, 且缩短的幅度随样本规模的增加更为明显. 当样本规模为2 000时, CPDSVDC的分类时间是21.16 s, 依次为SVDC、SSVDC、CSVDD分类时间的38.53%、8.82%、65.40%; 当样本规模为4 000时, CPDSVDC的分类时间是59.07 s, 依次为SVDC、SSVDC、CSVDD训练时间的26.76%、7.61%、48.54%; 当样本规模为10 000时, SVDC和SSVDC无法计算, CPDSVDC的分类时间是257.93 s, 为CSVDD分类时间的42.26%; 当样本规模增至20 000时, 仅CPDSVDC可以计算, 分类时间是799.30 s.

3 结 论

本文利用分段识别思想提出了CPDSVDC, 运用C均值聚类将原始空间划分为若干子空间; 采用双支持向量域分类器, 对子空间中较小规模的样本进行训练, 并运用链接规则综合所有子空间中的分类结果. 不同规模数据集上的数值实验验证了CPDSVDC具有高的分类精度、好的鲁棒性和短的分类时间, 从而为大规模样本集提供了一种可行的手段. 由于分类精度和时间随子空间数c和最近邻个数k变化, CPDSVDC的一个重要步骤是确定这些参数的取值, 文中通过经验为c和k赋值. 下一步将研究参数的有效取值方法, 并将CPDSVDC推广应用于多分类算法.

参考文献(References)

[1] Tax D M J, Duin R P W. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.

[2] Zhao Yang, Wang Shengwei, Xiao Fu. Pattern recognition-based chillers fault detection method using support vector data description(SVDD)[J]. Applied Energy, 2013, 112(1): 1041-1048.

[3] Niazmardi Saeid, Homayouni Saeid, Safari Abdolreza. An improved FCM algorithm based on the svdd for unsupervised hyperspectral data classification[J]. IEEE J of Selected Topics in Applied Earth Observations and Remote Sensing, 2013, 6(2): 831-839.

[4] Lan Jingchuan. Research on the fast ICA and SVDD based fault feature extraction algorithm for analog circuit[J]. Int J of Digital Content Technology and Its Applications, 2012, 6(6): 107-115.

[5] 陆从德, 张太镒, 胡金燕. 基于乘性规则的支持向量域分类器[J]. 计算机学报, 2004, 27(5): 690-694.
(Lu C D, Zhang T Y, Hu J Y. Support vector domain classifier based on multiplicative updates[J]. Chinese J of Computers, 2004, 27(5): 690-694.)

[6] 梁锦锦, 刘三阳, 吴德. 空间支持向量域分类器[J]. 西安电子科技大学学报, 2008, 35(6): 1080-1088.
(Liang J J, Liu S Y, Wu D. Space support vector domain classifier[J]. J of Xidian University, 2008, 35(6): 1080-1088.)

[7] 刘富, 侯涛, 刘云, 等. 可变惩罚因子的支持向量数据描述算法[J]. 吉林大学学报: 工学版, 2014, 44(2): 440-445.
(Liu F, Hou T, Liu Y, et al. A variable trade-off parameter support vector domain description[J]. J of Jilin University: Engineering and Technology Edition, 2014, 44(2): 440-445.)

[8] 吴德, 刘三阳. 多类支持向量域分类器[J]. 西安交通大学学报, 2012, 46(6): 87-91.
(Wu D, Liu S Y. Multiple support vector domain classifier[J]. J of Xi'an Jiaotong University, 2012, 46(6): 87-91.)

[9] Zhao Feng, Yan Liu, Zhen Hua, et al. Simplified solution for support vector domain description[J]. Int J of Digital Content Technology and Its Applications, 2011, 5(2): 292-299.

[10] Liu Sanyang, Liang Jinjin, Wu De, et al. Confidence support vector domain description[J]. J of Systems Engineering and Electronics, 2009, 20(4): 852-857.

[11] 任双桥, 杨德贵, 黎湘, 等. 分片支撑矢量机[J]. 计算机学报, 2009, 32(1): 77-85.
(Ren S Q, Yang D G, Li X, et al. Piecewise support vector machines[J]. Chinese J of Computers, 2009, 32(1): 77-85.)

[12] Ye Qixiang, Han Zhenjun, Jiao Jianbin. Human detection in images via piecewise linear support vector machines[J]. IEEE Trans on Image Processing, 2013, 22(2): 778-789.

[13] 毛韶阳, 李肯立. 优化K-means初始聚类中心的研究[J]. 计算机工程与应用, 2007, 43(22): 179-181.
(Mao S Y, Li K L. Research of optimal K-means initial clustering center[J]. Computer Engineering and Applications, 2007, 43(22): 179-181.)

(责任编辑: 齐 霁)