计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(9):43−52 [doi: 10.15888/j.cnki.csa.009216]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于超球和 ASSRFOA 的多生支持向量机①

莫源乐, 朱嘉静, 刘勇国, 张　云, 李巧勤

(电子科技大学 信息与软件工程学院 中医知识与数据工程实验室, 成都 610054)
通信作者: 朱嘉静, E-mail: jjzhu@uestc.edu.cn

摘　要: 支持向量机 (support vector machine, SVM) 是一种基于结构风险最小化的机器学习方法, 能够有效解决分类问题. 但随着研究问题的复杂化, 现实的分类问题往往是多分类问题, 而 SVM 仅能用于处理二分类任务. 针对这个问题, 一对多策略的多生支持向量机 (multiple birth support vector machine, MBSVM) 能够以较低的复杂度实现多分类, 但缺点在于分类精度较低. 本文对 MBSVM 进行改进, 提出了一种新的 SVM 多分类算法: 基于超球 (hypersphere) 和自适应缩小步长果蝇优化算法 (fruit fly optimization algorithm with adaptive step size reduction, ASSRFOA) 的 MBSVM, 简称 HA-MBSVM. 通过拟合超球得到的信息, 先进行类别划分再构建分类器, 并引入约束距离调节因子来适当提高分类器的差异性, 同时采用 ASSRFOA 求解二次规划问题, HA-MBSVM 可以更好地解决多分类问题. 我们采用 6 个数据集评估 HA-MBSVM 的性能, 实验结果表明 HA-MBSVM 的整体性能优于各对比算法.

关键词: 超球; 多生支持向量机; 多分类; 自适应缩小步长; 果蝇优化算法

引用格式: 莫源乐,朱嘉静,刘勇国,张云,李巧勤.基于超球和 ASSRFOA 的多生支持向量机.计算机系统应用,2023,32(9):43–52. http://www.c-s-a.org.cn/1003-3254/9216.html

Multiple Birth Support Vector Machine Based on Hypersphere and ASSRFOA

MO Yuan-Le, ZHU Jia-Jing, LIU Yong-Guo, ZHANG Yun, LI Qiao-Qin

(Knowledge and Data Engineering Laboratory of Chinese Medicine, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

Abstract: Support vector machine (SVM) is a machine learning method based on structural risk minimization and can solve classification problems. However, with the complexity of research problems, the real classification problems are often multi-classification ones, whereas SVM can only be adopted to deal with binary classification tasks. To this end, the multiple birth support vector machine (MBSVM) combined with the one-against-all strategy can realize multi-classification with low complexity, but the classification accuracy is low. This study improves MBSVM and proposes a new SVM multi-classification algorithm which is a multiple birth support vector machine based on the hypersphere and fruit fly optimization algorithm with adaptive step size reduction (ASSRFOA). The algorithm is referred to as HA-MBSVM. Through the information obtained from hypersphere fitting, firstly all classes are divided into several blocks and then classifiers are constructed for each class. The constraint distance regulation factor is introduced to properly improve the difference of the classifiers. At the same time, ASSRFOA is employed to solve the quadratic programming problems and HA-MBSVM can better solve the multi-classification problems. Six datasets are utilized to evaluate the performance of HA-MBSVM. The experimental results show that the overall performance of HA-MBSVM is better than that of the comparison algorithms.

Key words: hypersphere; multiple birth support vector machine (MBSVM); multi-classification; adaptive step size reduction; fruit fly optimization algorithm (FOA)

① 基金项目: 国家自然科学基金 (62202084); 国家科技基础资源调查专项 (2022FY102002); 中国博士后科学基金 (2021M690028); 中央高校基本业务费 (ZYGX2021YGLH012, ZYGX2021J020); 四川省自然科学基金 (2022NSFSC0883, 2022NSFSC0958); 四川省重点研发计划 (2022YFS0059, 2023YFS0338)
收稿时间: 2023-02-13; 修改时间: 2023-03-14; 采用时间: 2023-03-30; csa 在线出版时间: 2023-07-14
CNKI 网络首发时间: 2023-07-17

SVM[1] 是一种基于结构风险最小化的机器学习方法, 在解决小样本、高维问题和非线性问题等方面表现出良好的泛化能力和预测性能, 在生物医学[2–5]、故障分析[6–9]、图像识别[10–12] 等领域中被广泛应用, 并具有可观的分类效果[13,14].

随着研究问题的复杂化, 现实的分类问题不单纯是二分类任务, 存在多分类的现象, 而 SVM 仅能用于处理二分类任务. 针对这个问题, 现有的工作已做出重要的贡献[15–19]. 其中一对一 (one-against-one, OAO) 策略、一对一对余 (one-against-one-against-rest, OAOAR) 策略、一对多 (one-against-all, OAA) 策略是 3 种最常用的间接求解多分类问题的分解策略[20]. 对于一个具有 K 个类别的多分类问题, 结合 OAO 策略的一对一支持向量机 (one-against-one support vector machine, OAO SVM) 任选两个类别的样本, 分别作为正反两类, 构建 $K(K-1)/2$ 个 SVM 分类器. 结合 OAOAR 策略的支持向量分类-回归机 (support vector classification-regression machine for K-class classification, K-SVCR)[21] 在正反两类的基础上, 增加了其余类, 减少了样本的误分. 将孪生支持向量机 (twin support vector machine, TWSVM)[22] 与 OAA 策略结合, Yang 等人提出了 MBSVM[23], 通过将“最近”的决策方式转变为“最远”的决策方式, 有效减少了约束条件, 因而具有较低时间复杂度.

MBSVM 虽然在算法效率上具有明显优势, 但其分类精度不够高. 主要原因有两点[24], 一是用超平面进行拟合的类别之间可能相距较远, 很难使得超平面离它们同时都近; 二是约束条件较少, 二次规划问题 (quadratic programming problem, QPP) 的求解容易陷入局部最优. 对于第 1 个问题, 为达到更好的超平面拟合效果, 我们希望根据不同类别之间的相似度将所有类别划分成若干块, 块内类别的样本相似度相对较高, 块间类别的样本相似度相对较低; 对于第 2 个问题, 应该采用全局性更好的求解算法. ASSRFOA 是一种新型的群智能优化算法, 相比于其他算法, 它继承了果蝇优化算法 (fruit fly optimization algorithm, FOA) 优秀的全局寻优能力, 同时通过改进候选解生成机制和搜索步长, 并引入柯西变异, 有效地改善了陷入局部最优的共性问题[25–27]. 本文将采用 ASSRFOA 作为 QPP 的求解算法.

基于以上讨论, 本文从 MBSVM 存在的问题出发, 提出一种新的 SVM 多分类算法: 基于超球 (hypersphere)[28] 和 ASSRFOA 的 MBSVM, 简称 HA-MBSVM. 该算法框架如图 1 所示, 首先通过拟合超球, 获取各类样本对应超球的球心和半径. 利用球心和半径, 计算各类别之间的相似度判定值, 然后将所有类别划分成若干块, 块内类别样本相似度相对较高, 块间类别样本相似度相对较低, 最后再为各类别构建分类器并采用 ASSRFOA 求解 QPP. 我们还引入了约束距离调节因子用以进一步提升 HA-MBSVM 的性能. 本文中的算法仅介绍使用了核方法的情形, 核函数选择径向基核函数 (radial basis function, RBF), 规定矢量加粗表示, 标量不加粗.

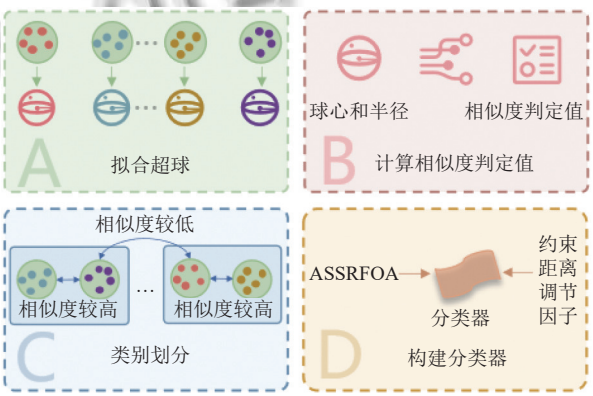


图 1 HA-MBSVM 算法框架

综上所述, 本文的主要贡献如下.

(1) 提出了一种新的 SVM 多分类算法. 利用拟合超球所得信息完成类别划分, 构建更准确的分类器.

(2) 将 ASSRFOA 应用于 QPP 的求解, ASSRFOA 良好的全局性有利于提高 HA-MBSVM 的性能.

(3) 采用 6 个数据集对 HA-MBSVM 的性能进行评估. 实验结果表明, HA-MBSVM 算法的性能优于常见的 SVM 多分类算法以及 MBSVM.

1 相关工作

本节将介绍与 HA-MBSVM 密切相关的 3 项工作. 其中, 第 1.1 节介绍 MBSVM[23], 这是本文算法的改进出发点; 第 1.2 节介绍超球相关内容, HA-MBSVM 采用文献 [28] 中生成超球的方法为各类别的样本拟合超球, 获取球心和半径用于后续的类别划分; 第 1.3 节介绍的 ASSRFOA 算法[25] 将用于求解 QPP.

1.1 多生支持向量机

为方便讨论, 假定对于多分类问题, 有训练集:

$$T = \{(\boldsymbol{x}_i, y_1), \cdots, (\boldsymbol{x}_l, y_l)\} \tag{1}$$

其中, $(\boldsymbol{x}_i, y_i)$为第 i 个样本, $\boldsymbol{x}_i \in R^n$ 为特征向量, $y_i \in \{1,\cdots,K\}$为类别标签, l 是样本数, K 为类别数.

对于训练集 (1), 假设第 k 类样本构成矩阵$A_k \in R^{l_k \times n}$, 训练集 (1) 中除去第 k 类样本的剩余样本构成矩阵 $B_k = [A_1^{\mathrm{T}},\cdots,A_{k-1}^{\mathrm{T}},A_{k+1}^{\mathrm{T}},\cdots,A_K^{\mathrm{T}}]^{\mathrm{T}}$, $k = 1,\cdots,K$, l_k 为第 k 类样本的数量.

MBSVM 每次选取一个类的样本作为负样本, 其他类的所有样本作为正样本, 寻找一个超平面, 要求该超平面离正样本尽可能近, 离负样本尽可能远, 即获得式 (2)[22,29].

$$K(\boldsymbol{x},E)\boldsymbol{v}_k + b_k = 0, \quad k = 1,\cdots,K \tag{2}$$

其中, $\boldsymbol{v}_k \in R^l$和b_k为模型参数, $E = [A_1^{\mathrm{T}},\cdots,A_K^{\mathrm{T}}]^{\mathrm{T}}$, $K(\boldsymbol{x},E) \in R^l$为由核函数$K(\boldsymbol{x},\boldsymbol{y})$产生的行向量. 为了计算$\boldsymbol{v}_k$和$b_k$, 需要解决如下 QPP 问题, 如式 (3) 所示.

$$\begin{cases} \min\limits_{\boldsymbol{v}_k,b_k,\boldsymbol{\xi}_k} \dfrac{1}{2}\|K(B_k,E)\boldsymbol{v}_k + \boldsymbol{e}_{k1}b_k\|^2 + C_k\boldsymbol{e}_{k2}^{\mathrm{T}}\boldsymbol{\xi}_k \\ \text{s.t.} \quad \begin{cases} K(A_k,E)\boldsymbol{v}_k + \boldsymbol{e}_{k2}b_k \geqslant \boldsymbol{e}_{k2} - \boldsymbol{\xi}_k \\ \boldsymbol{\xi}_k \geqslant 0 \end{cases} \end{cases} \tag{3}$$

其中, $\boldsymbol{\xi}_k \in R^{l_k}$为$A_k$中样本对应松弛变量的列向量, $C_k > 0$ 为惩罚参数, $\boldsymbol{e}_{k1} \in R^{l-l_k}$和$\boldsymbol{e}_{k2} \in R^{l_k}$为元素全是 1 的列向量, $K(A_k,E)$和$K(B_k,E)$分别为由核函数$K(\boldsymbol{x},\boldsymbol{y})$产生的 $l_k \times l$和$(l-l_k)\times l$的矩阵.

由拉格朗日优化方法可将式 (3) 转为对偶问题 (4).

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}_k} \quad \boldsymbol{e}_{k2}^{\mathrm{T}}\boldsymbol{\alpha}_k - \dfrac{1}{2}\boldsymbol{\alpha}_k^{\mathrm{T}}R_k(S_k^{\mathrm{T}}S_k)^{-1}R_k^{\mathrm{T}}\boldsymbol{\alpha}_k \\ \text{s.t.} \quad 0 \leqslant \boldsymbol{\alpha}_k \leqslant C_k \end{cases} \tag{4}$$

其中,$\boldsymbol{\alpha}_k \in R^{l_k}$为拉格朗日乘子的非负列向量, $S_k = [K(B_k,E)\ \boldsymbol{e}_{k1}]$, $R_k = [K(A_k,E)\ \boldsymbol{e}_{k2}]$. 为避免矩阵$S_k^{\mathrm{T}}S_k$的病态化导致不可求逆, 需要为对偶问题 (4) 添加一个正则化项$\varepsilon\boldsymbol{I}$, 其中$\varepsilon > 0$为一固定的小标量, $\boldsymbol{I}$ 为适当大小的单位矩阵, 于是对偶问题 (4) 可写成[22]:

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}_k} \quad \boldsymbol{e}_{k2}^{\mathrm{T}}\boldsymbol{\alpha}_k - \dfrac{1}{2}\boldsymbol{\alpha}_k^{\mathrm{T}}R_k(S_k^{\mathrm{T}}S_k + \varepsilon I)^{-1}R_k^{\mathrm{T}}\boldsymbol{\alpha}_k \\ \text{s.t.} \quad 0 \leqslant \boldsymbol{\alpha}_k \leqslant C_k \end{cases} \tag{5}$$

由各类别对应 QPP 求解得到的超平面, MBSVM 的决策函数计算新样本到各超平面的距离, 相距最远的超平面对应的类别即为新样本的预测类别标签.

1.2 超球

朱美琳等人提出的球结构支持向量机, 通过为每个类别拟合超球来解决多分类问题[28]. 对于训练集 (1), 当为A_k拟合超球时, 假定$\boldsymbol{n}_k$为最小超球的球心, r_k为最小超球的半径, 要求该最小超球尽可能包含A_k中所有样本点, 通过引入松弛变量, 允许存在样本点位于超球外侧. 于是, 可得到如下优化问题 (6).

$$\begin{cases} \min\ (r_k)^2 + C_k\sum\limits_{i=1}^{l_k}\xi_i^k \\ \text{s.t.} \begin{cases} \|\boldsymbol{x}_i^k - \boldsymbol{n}_k\|^2 \leqslant (r_k)^2 + \xi_i^k \\ \xi_i^k \geqslant 0,\ i = 1,\cdots,l_k \end{cases} \end{cases} \tag{6}$$

其中, $\boldsymbol{x}_i^k$为A_k中第 i 个样本点的特征向量, ξ_i^k 为$\boldsymbol{x}_i^k$对应的松弛变量, $C_k > 0$为惩罚参数.

引入核函数并利用拉格朗日优化方法, 可将式 (6) 转为对偶问题 (7).

$$\begin{cases} \max\ \sum\limits_{i=1}^{l_k}\alpha_i^k K(\boldsymbol{x}_i^k,\boldsymbol{x}_i^k) - \sum\limits_{i=1}^{l_k}\sum\limits_{j=1}^{l_k}\alpha_i^k\alpha_j^k K(\boldsymbol{x}_i^k,\boldsymbol{x}_j^k) \\ \text{s.t.} \begin{cases} \sum\limits_{i=1}^{l_k}\alpha_i^k = 1 \\ 0 \leqslant \alpha_i^k \leqslant C_k,\ i = 1,\cdots,l_k \end{cases} \end{cases} \tag{7}$$

其中, α_i^k为式 (6) 中第 i 个约束条件对应的拉格朗日乘子. 通过求解式 (7), 可以得到满足要求的拉格朗日乘子, 进而得到超球的球心和半径.

1.3 ASSRFOA 具体步骤

步骤 1. 初始化最大迭代次数 *Maxgen*、果蝇种群规模 *Sizepop*、味道浓度方差阈值δ和果蝇群体位置 $\boldsymbol{X}_{\text{axis}} \in R^D$、$\boldsymbol{Y}_{\text{axis}} \in R^D$, D 为待求解未知量的个数.

步骤 2. 赋予果蝇个体随机的搜索方向和距离.

$$\begin{cases} \boldsymbol{X}_i = \boldsymbol{X}_{\text{axis}} + \boldsymbol{Rand} \\ \boldsymbol{Y}_i = \boldsymbol{Y}_{\text{axis}} + \boldsymbol{Rand} \end{cases} \tag{8}$$

其中, $\boldsymbol{Rand} \in R^D$, 其元素均为 [−1, 1] 之间的随机数, $\boldsymbol{X}_i$和$\boldsymbol{Y}_i$为各果蝇个体基于群体位置随机分散后所处的位置, $i = 1,\cdots,Sizepop$.

步骤 3. 设$\boldsymbol{X}_i = (p_i,\cdots,p_D)$, $\boldsymbol{Y}_i = (q_1,\cdots,q_D)$, 利用式 (9) 计算果蝇个体到原点的间距$\boldsymbol{Dist}_i \in R^D$, 再利用式 (10) 计算味道浓度判定值$\boldsymbol{SD}_i \in R^D$.

$$\boldsymbol{Dist}_i = \left(\sqrt{p_1^2+q_1^2},\cdots,\sqrt{p_D^2+q_D^2}\right) \tag{9}$$

$$\boldsymbol{SD}_i = \left(\frac{sign(Rand)}{\sqrt{p_1^2+q_1^2}},\cdots,\frac{sign(Rand)}{\sqrt{p_D^2+q_D^2}}\right) \tag{10}$$

其中, 随机数$Rand \in [-1, 1]$, *sign* 函数如式 (11) 所示[30].

$$sign(x)=\begin{cases}1, & x>0\\ 0, & x=0\\ -1, & x<0\end{cases} \tag{11}$$

步骤 4. 把$\boldsymbol{SD}_i$代入到味道浓度判定函数 $Fitness$ (即目标函数), 计算出味道浓度值$smell_i$.

$$smell_i=Fitness(\boldsymbol{SD}_i) \tag{12}$$

步骤 5. 找出果蝇种群中味道浓度值最优的果蝇, 并记录此果蝇的位置信息和相应的味道浓度值.

$$[Currentbest, bestindex]=\max(\boldsymbol{Smell}) \tag{13}$$

其中, $Currentbest$ 为最优味道浓度值, $bestindex\in\{1,\cdots,Sizepop\}$为最优味道浓度值对应下标, $\boldsymbol{Smell}=(smell_i,\cdots,smell_{Sizepop})$.

步骤 6. 计算味道浓度方差值σ^2, 且当迭代次数不为 0 时, 转至步骤 7, 否则转至步骤 8.

$$\sigma^2=\frac{1}{Sizepop}\sum_{i=1}^{Sizepop}\left(smell_i-\sum_{i=1}^{Sizepop}\frac{smell_i}{Sizepop}\right)^2 \tag{14}$$

步骤 7. 判断 $Currentbest$ 是否优于历史最优味道浓度值 $Globalbest$, 是则通过式 (15) 和式 (16) 保留 $Currentbest$ 并更新果蝇群体位置, 然后转至步骤 8, 否则直接转至步骤 8.

$$Globalbest=Currentbest \tag{15}$$

$$\begin{cases}\boldsymbol{X}_{\text{axis}}=\boldsymbol{X}_{bestindex}\\ \boldsymbol{Y}_{\text{axis}}=\boldsymbol{Y}_{bestindex}\end{cases} \tag{16}$$

步骤 8. 判断是否陷入局部最优, 若$\sigma^2<\delta$, 则转至步骤 9, 否则转至步骤 10.

步骤 9. 利用式 (17) 对果蝇个体位置进行柯西变异, 然后转至步骤 11.

$$\begin{cases}\boldsymbol{X}_i=\boldsymbol{X}_i+\boldsymbol{X}_i\times C(0,1)\\ \boldsymbol{Y}_i=\boldsymbol{Y}_i+\boldsymbol{Y}_i\times C(0,1)\end{cases} \tag{17}$$

其中, $C(0, 1)$ 为标准柯西分布.

步骤 10. 根据式 (18) 更新果蝇个体位置, 然后转至步骤 11.

$$\begin{cases}\boldsymbol{X}_i=\boldsymbol{X}_{\text{axis}}+\tau e^{\frac{-(\beta\times g)}{Maxgen}}\times\boldsymbol{Rand}\\ \boldsymbol{Y}_i=\boldsymbol{Y}_{\text{axis}}+\tau e^{\frac{-(\beta\times g)}{Maxgen}}\times\boldsymbol{Rand}\end{cases} \tag{18}$$

其中, $\tau e^{\frac{-(\beta\times g)}{Maxgen}}$为搜索步长, τ为步长调控因子, β为指数调节因子, g 为当前迭代次数.

步骤 11. 当前迭代次数不大于 $Maxgen$, 执行步骤 3 至步骤 10, 否则执行步骤 12.

步骤 12. 输出全局最优解.

2 HA-MBSVM

2.1 类别划分

在 MBSVM 中, 被作为正类的样本间可能相距较远, 很难构建一个超平面离它们同时都近. 于是我们尝试先将所有类别根据彼此间的相似度和给定阈值划分成若干块, 块内类别样本相似度相对较高, 块间类别样本相似度相对较低, 之后再分别以各块的样本作为约束条件来构建多个超平面, 从而避免因单个超平面难以同时有效地满足多个差异较大的类别样本的约束而导致分类器性能下降. 相似度判定值既应考虑超球球心, 也应考虑超球半径[31], 由此我们得到如下相似度判定值的计算式 (19).

$$d_{ab}=\frac{\|\boldsymbol{n}_a-\boldsymbol{n}_b\|}{r_a+r_b} \tag{19}$$

其中, $\boldsymbol{n}_a$和$\boldsymbol{n}_b$, r_a和r_b分别为第 a 类和第 b 类样本对应超球的球心和半径. d_{ab}值越小则两类样本相似度越高, 反之越低. 我们将所有类别从 0 开始标号, 得到类别标签集$S_0=\{0,\cdots,K-1\}$. 由式 (19) 可以计算出各类别两两之间的相似度判定值, 其中的全局最大值$d_{\max}$如式 (20) 所示.

$$d_{\max}=\max_{a,b\in S_0,a<b}d_{ab} \tag{20}$$

对S_0进行一次类别划分, 将其分为差异性较大的两块, 块中的类别相似度较高, 得到此时类别划分集$S=\{S_1,S_2\}$, 其中S_1和S_2为类别标签的集合. 然后再对各块分别进行一次类别划分, 重复上述过程, 块停止划分的标准为块中类别之间最大的相似度判定值与$d_{\max}$的比值不大于给定阈值或块中只剩下一个类别. 由此得到最终的类别划分集$S=\{S_1,\cdots,S_i,\cdots,S_m\}$, $S_i=\{s_{i1},\cdots,s_{ij},\cdots,s_{in_i}\}$, 其中$s_{ij}$为第 i 块中第 j 个类别标签, m 为块总数, n_i为第 i 块中的标签数, $i=1,\cdots,m$. 完整的类别划分算法见算法 1, 一次类别划分算法的细节见算法 2, 用 $len(x)$ 表示 x 中的类别标签数, 用 a 和 b 暂存类别标签, 用$Algorithm2(d_{ab},S_i)$表示对算法 2 的调用.

2.2 分类器构建

MBSVM 在为各类别构建分类器时, 每次选取一个类的样本作为负样本, 其他类的所有样本作为正样本, 要求超平面离正样本尽可能近, 离负样本尽可能远. 这种做法虽然能够有效减少 QPP 的约束条件, 提高训练效率, 但在这种做法下, 决策函数将新样本归入相距最远的超平面对应的类别. 在样本空间中, 与某一超平

面对应类别不同的样本也可能离该超平面较远, 因此相比于“最近”的决策方式, “最远”的决策方式错误分类的可能性更高.

我们利用 HA-MBSVM 为各类别构建分类器时, 每次选取一个类的样本作为正样本, 假设该类别标签s_{ij}位于类别划分集S的块S_i中,其余块为$S_j \in S(1 \leqslant j \leqslant m, j \neq i)$, $S_i' = S_i - \{s_{ij}\}$, 依次以S_j和S_i'中类别的样本为负样本, 构建 m 个子分类器 (当S_i中仅有s_{ij}一个类别时, 构建$m-1$ 个子分类器), 即寻找 m 个超平面, 如式 (21) 所示, 要求超平面离正样本尽可能近, 离负样本尽可能远, 构建过程如图 2 所示.

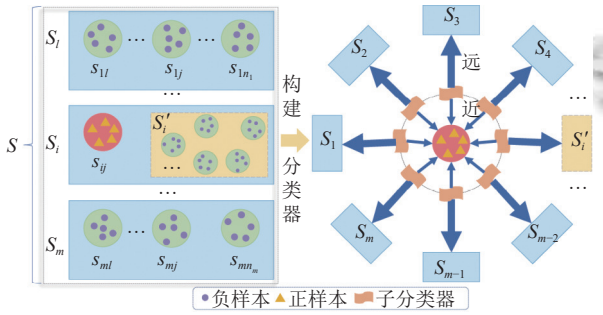


图 2 HA-MBSVM 分类器构建示意图

算法 1. 完整的类别划分

输入: 各类样本对应超球的球心$\boldsymbol{n}_k$和半径r_k, $k=0,\cdots,K-1$; 类别标签集S_0; 划分停止阈值ξ.
输出: 类别划分集 S.

1. 由式 (19) 计算出各类别两两之间的相似度判定值, 由式 (20) 得到其中的全局最大值$d_{\max}$;
2. $flag \leftarrow true$;
3. $S \leftarrow S_0$;
4. WHILE $flag$ /*不断进行划分, 直到所有块均满足停止条件*/
5. $flag \leftarrow false$;
6. FOR S_i IN S /*依次处理当前S中所有的块*/
7. $d'_{\max} \leftarrow \max\limits_{a,b\in S_i, a<b} d_{ab}$;
8. IF $len(S_i)>1$ AND $\frac{d'_{\max}}{d_{\max}} > \xi$ THEN /*若不满足停止条件, 则进行一次类别划分*/
9. $S_{i1}, S_{i2} \leftarrow Algorithm2(d_{ab}, S_i)$; /*调用算法 2*/
10. $S_i \leftarrow S_{i1}$, $S \leftarrow S \cup \{S_{i2}\}$;
11. $flag \leftarrow true$;
12. END IF
13. END FOR
14. END WHILE
15. RETURN S;

假设类别s_{ij}对应正样本构成矩阵$P_{ij} \in R^{l_{ij} \times n}$, 各子分类器对应负样本构成矩阵$Q_{ij}^k \in R^{l_{ij}^k \times n}, k=1,\cdots,m$, l_{ij}和l_{ij}^k分别为正样本和负样本的数量.

$$K(\boldsymbol{x}, E_{ij}^k)\boldsymbol{v}_{ij}^k + b_{ij}^k = 0, \quad k = 1,\cdots,m \tag{21}$$

其中, $\boldsymbol{v}_{ij}^k$和b_{ij}^k为模型参数, $E_{ij}^k = [P_{ij}^{\mathrm{T}}, (Q_{ij}^k)^{\mathrm{T}}]^{\mathrm{T}}$, $K(\boldsymbol{x}, E_{ij}^k) \in R^{l_{ij}+l_{ij}^k}$为由核函数$K(\boldsymbol{x},\boldsymbol{y})$产生的行向量. 类似于 MBSVM, 我们有如下 QPP 式 (22).

$$\begin{cases} \min\limits_{\boldsymbol{v}_{ij}^k, b_{ij}^k, \boldsymbol{\xi}_{ij}^k} \dfrac{1}{2}\left\| K(P_{ij}, E_{ij}^k)\boldsymbol{v}_{ij}^k + \boldsymbol{e}_{ij}^{k1} b_{ij}^k \right\|^2 + C_{ij}^k (\boldsymbol{e}_{ij}^{k2})^{\mathrm{T}} \boldsymbol{\xi}_{ij}^k \\ \text{s.t.} \quad \begin{cases} K(Q_{ij}^k, E_{ij}^k)\boldsymbol{v}_{ij}^k + \boldsymbol{e}_{ij}^{k2} b_{ij}^k \geqslant \boldsymbol{e}_{ij}^{k2} - \boldsymbol{\xi}_{ij}^k \\ \boldsymbol{\xi}_{ij}^k \geqslant 0 \end{cases} \end{cases} \tag{22}$$

其中, $\boldsymbol{\xi}_{ij}^k \in R^{l_{ij}^k}$为$Q_{ij}^k$中样本对应松弛变量的列向量, $C_{ij}^k > 0$为惩罚参数, $\boldsymbol{e}_{ij}^{k1} \in R^{l_{ij}}$和$\boldsymbol{e}_{ij}^{k2} \in R^{l_{ij}^k}$为元素全是 1 的列向量.

算法 2. 一次类别划分

输入: 各类别间的相似度判定值; 待划分块S_i.
输出: 类别划分结果S_{i1}和S_{i2}.

1. IF $len(S_i)$=2 THEN /*若S_i中仅有两个类别标签, 则可直接输出结果*/
2. 将S_i中的两个类别标签分别作为S_{i1}和S_{i2};
3. RETURN S_{i1}和S_{i2};
4. END IF
5. $a,b \leftarrow \arg\min\limits_{a,b\in S_i, a<b} d_{ab}$, $S_{i1} \leftarrow \{a,b\}$, $S_i \leftarrow S_i - S_{i1}$;
6. $b \leftarrow \arg\max\limits_{b\in S_i} \sum\limits_{a\in S_{i1}} d_{ab}$, $S_{i2} \leftarrow \{b\}$, $S_i \leftarrow S_i - S_{i2}$;
7. WHILE *true* /*不断将S_i中的类别标签分配到S_{i1}和S_{i2}中, 当S_i为空时输出结果*/
8. IF $len(S_i)$=0 THEN
9. RETURN S_{i1}和S_{i2};
10. ELSE
11. $b \leftarrow \arg\min\limits_{b\in S_i} \sum\limits_{a\in S_{i2}} d_{ab}$;
12. $S_{i2} \leftarrow S_{i2} \cup \{b\}$, $S_i \leftarrow S_i - S_{i2}$;
13. IF $len(S_i)$=0 THEN
14. RETURN S_{i1}和S_{i2};
15. ELSE IF $len(S_i)$=1 THEN
16. IF $\sum\limits_{a\in S_i, b\in S_{i1}} d_{ab} \leqslant \sum\limits_{a\in S_i, b\in S_{i2}} d_{ab}$ THEN
17. $S_{i1} \leftarrow S_{i1} \cup S_i$;
18. ELSE
19. $S_{i2} \leftarrow S_{i2} \cup S_i$;
20. END IF
21. RETURN S_{i1}和S_{i2};
22. ELSE
23. $a \leftarrow \arg\min\limits_{a\in S_i} \sum\limits_{b\in S_{i1}} d_{ab}$, $S_{i1} \leftarrow S_{i1} \cup \{a\}$;
24. $S_i \leftarrow S_i - S_{i1}$;
25. END IF
26. END IF
27. END WHILE

2.3 约束距离调整

QPP 式 (22) 中的约束条件用$\boldsymbol{e}_{ij}^{k2}$限制负样本尽可

能离超平面有 1 的约束距离[23]. 但对于同一块中相似度较高的两个类别, 这一约束距离可能是不足够的, 由于两个类别的子分类器可能较为相似, 分类器对较小的距离差异不敏感, 此时样本容易被错误分类. 因此, 我们在 QPP 式 (22) 中添加参数d_{ij}^k, 称为约束距离调节因子, 用于调整超平面与负样本之间的约束距离, 提高分类器的差异性, 由此得到 QPP 式 (23).

$$\begin{cases} \min\limits_{\boldsymbol{v}_{ij}^k,b_{ij}^k,\boldsymbol{\xi}_{ij}^k} \dfrac{1}{2}\left\|K(P_{ij},E_{ij}^k)\boldsymbol{v}_{ij}^k+\boldsymbol{e}_{ij}^{k1}b_{ij}^k\right\|^2+C_{ij}^k(\boldsymbol{e}_{ij}^{k2})^{\mathrm{T}}\boldsymbol{\xi}_{ij}^k \\ \text{s.t.} \quad \begin{cases} K(Q_{ij}^k,E_{ij}^k)\boldsymbol{v}_{ij}^k+\boldsymbol{e}_{ij}^{k2}b_{ij}^k\geqslant \boldsymbol{e}_{ij}^{k2}d_{ij}^k-\boldsymbol{\xi}_{ij}^k \\ \boldsymbol{\xi}_{ij}^k\geqslant 0 \end{cases} \end{cases} \tag{23}$$

由拉格朗日优化方法可得式 (23) 的对偶问题 (24).

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}_{ij}^k} -\dfrac{1}{2}(\boldsymbol{\alpha}_{ij}^k)^{\mathrm{T}}R_{ij}^k[(S_{ij}^k)^{\mathrm{T}}S_{ij}^k+\varepsilon I]^{-1}(R_{ij}^k)^{\mathrm{T}}\boldsymbol{\alpha}_{ij}^k+d_{ij}^k(\boldsymbol{e}_{ij}^{k2})^{\mathrm{T}}\boldsymbol{\alpha}_{ij}^k \\ \text{s.t. } 0\leqslant\boldsymbol{\alpha}_{ij}^k\leqslant C_{ij}^k \end{cases} \tag{24}$$

其中, $\boldsymbol{\alpha}_{ij}^k\in R^{l_{ij}^k}$为拉格朗日乘子的非负列向量, $S_{ij}^k=[K(P_{ij},E_{ij}^k)\boldsymbol{e}_{ij}^{k1}]$, $R_{ij}^k=[K(Q_{ij}^k,E_{ij}^k)\boldsymbol{e}_{ij}^{k2}]$.

通过求解 m 个对偶问题 (24), 我们得到类别s_{ij}的分类器, 该分类器由 m 个子分类器组成. 类似地, 我们为每一个类别都构建相应的分类器, 由此得到 HA-MBSVM 的决策函数 (25).

$$f(\boldsymbol{x})=\arg\min_{i,j}\frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}}\frac{\left|K(\boldsymbol{x},E_{ij}^k)\boldsymbol{v}_{ij}^k+b_{ij}^k\right|}{\sqrt{(\boldsymbol{v}_{ij}^k)^{\mathrm{T}}K(E_{ij}^k,E)\boldsymbol{v}_{ij}^k}} \tag{25}$$

其中, $i=1,\cdots,m$, $j=1,\cdots,n_i$, 指向类别划分集中第 i 块第 j 个类别标签, N_{ij}为类别s_{ij}的子分类器数量. 决策函数 (25) 计算新样本到各类别子分类器的平均距离, 将其归入平均距离最小的类别.

2.4 求解 QPP

Yang 等人[23] 采用逐次超松弛迭代法 (successive over relaxation, SOR) 求解 MBSVM 中的 QPP, 该方法求解速度快但求解质量不高. 为了得到性能更好的模型, 本文采用全局优化能力较好的 ASSRFOA 来求解 QPP. HA-MBSVM 关键需要求解对偶问题 (24), 注意到其中有约束条件$0\leqslant\boldsymbol{\alpha}_{ij}^k\leqslant C_{ij}^k$, 因此在利用 ASSRFOA 求解式 (24) 时, 应对$\boldsymbol{\alpha}_{ij}^k$的候选解进行越界处理: 当候选解小于 0 时, 变为 0; 当候选解大于C_{ij}^k时, 变为C_{ij}^k, 当候选解介于 0 与C_{ij}^k之间, 保留原值.

3 实验分析

在本节中, 我们从 3 方面评估所提出的 HA-MBSVM 算法. 第 3.1 节是 HA-MBSVM 与其他 3 种算法在 4 项评估指标下的性能对比, 第 3.2 节对 HA-MBSVM 中的部分参数进行讨论, 第 3.3 节是消融实验结果.

实验选用了以下 6 个用于多分类的数据集 (https://github.com/renatopp/arff-datasets): iris、 wine、 zoo、 tae、 lymph 和 dermatology. 所有实验均在 PC 机 (16 GB RAM, AMD Ryzen 5 5600U 处理器, Windows 10 操作系统) 上采用 Python 环境实现. 数据集的具体信息见表 1. 规定 HA-MBSVM 中生成的最小超球至少包含 90% 的样本点. 所有算法中的核函数选择 RBF, 即式 (26).

$$K(\boldsymbol{x},\boldsymbol{y})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right) \tag{26}$$

其中, $\boldsymbol{x}$与$\boldsymbol{y}$为维度相同的向量, $\sigma>0$为 RBF 的带宽.

表 1 数据集具体信息

数据集	样本个数	样本维数	类别数
iris	150	4	3
wine	178	13	3
zoo	101	17	7
tae	151	7	3
lymph	148	18	4
dermatology	366	34	6

3.1 分类性能对比

我们将 HA-MBSVM 与 OAO SVM、一对多支持向量机 (one-against-all support vector machine, OAA SVM)、MBSVM 进行性能对比, 4 项评估指标分别为准确率, 查准率, 查全率和 $F1$ 值, 采用 5 折交叉验证寻找最优参数. 具体实验设置见如下.

对于 OAO SVM、OAA SVM、MBSVM, RBF 带宽σ的候选值集合为$[2^{-10},2^{-9},2^{-8},\cdots,2^4]$, 惩罚参数$C$的候选值集合为$[2^{-2},2^{-3},2^{-4},\cdots,2^{12}]$. 对于 HA-MBSVM, 在拟合超球时, RBF 带宽σ_1和惩罚参数C_1的候选值集合均为$[2^{-2},2^4,2^{10}]$, 类别划分停止阈值ξ的候选值集合为$[0.2,0.4,0.6,0.8,1]$; 在构建分类器时, RBF 带宽σ_2的候选值集合为$[2^{-3},2^{-1.75},2^{-0.5},\cdots,2^7]$, 惩罚参数$C_2$的候选值集合为$[2^3,2^9,2^{15}]$, 约束距离调节因子 d 的候选值集合为$[2^0,2^1,2^2,2^3,2^4,2^5,2^6]$. 各类别样本在拟合超球和构建分类器时使用统一的初始参数, 两个环节中的 QPP 使用 ASSRFOA 完成求解. 对于 ASSRFOA 中的初始化参数, 在拟合超球时的最大迭代次数 $Maxgen_1$

默认为 100, 在构建分类器时的最大迭代次数 $Maxgen_2$ 候选值集合为[90,150,210], 其余 ASSRFOA 中的参数采用文献 [25] 中实验的取值作为默认参数值, 即种群规模 $Sizepop$=30, 步长调控因子τ=0.2, 指数调节因子β=2, 味道浓度方差阈值δ=10^{-6}.

表 2 为 6 个数据集上, HA-MBSVM 与其他 3 种算法关于 4 项评估指标 5 折交叉验证的最佳 (以准确率为基准) 平均结果. 表 3 为 3 个对比算法在最佳平均结果下的最优参数, 表 4 为 HA-MBSVM 在最佳平均结果下的最优参数.

表 2 HA-MBSVM 与其他 3 种算法的分类性能对比

数据集	算法	准确率	查准率	查全率	$F1$值
iris	OAO SVM	0.9800	0.9819	0.9821	0.9812
	OAA SVM	0.9733	0.9756	0.9756	0.9756
	MBSVM	0.9667	0.9656	0.9656	0.9647
	HA-MBSVM	**0.9867**	**0.9859**	**0.9888**	**0.9867**
wine	OAO SVM	0.9943	0.9867	0.9952	0.9901
	OAA SVM	0.9943	0.9944	**0.9956**	**0.9948**
	MBSVM	0.9890	0.9875	0.9893	0.9878
	HA-MBSVM	**0.9947**	**0.9949**	0.9944	0.9944
zoo	OAO SVM	0.9514	0.9060	0.8976	0.8901
	OAA SVM	0.9500	0.9052	0.8802	0.8830
	MBSVM	0.9605	**0.9381**	0.9155	**0.9156**
	HA-MBSVM	**0.9610**	0.9305	**0.9295**	0.9151
tae	OAO SVM	0.5972	0.5963	0.5979	0.5901
	OAA SVM	0.6226	0.6249	0.6402	0.6182
	MBSVM	0.6361	0.6545	0.6375	0.6254
	HA-MBSVM	**0.7019**	**0.7132**	**0.7082**	**0.6995**
lymph	OAO SVM	0.8466	0.7421	0.7311	0.7286
	OAA SVM	0.8571	0.7836	**0.7771**	**0.7761**
	MBSVM	0.8522	0.7353	0.7410	0.7356
	HA-MBSVM	**0.8791**	**0.7936**	0.7667	0.7735
dermatology	OAO SVM	0.9691	0.9664	0.9559	0.9587
	OAA SVM	0.9664	0.9663	0.9584	0.9605
	MBSVM	0.9690	0.9671	**0.9680**	**0.9650**
	HA-MBSVM	**0.9721**	**0.9723**	0.9518	0.9565

表 3 对比算法的最优参数 (评估指标为准确率)

数据集	OAO SVM		OAA SVM		MBSVM	
	σ	C	σ	C	σ	C
iris	2^{-2}	2^{4}	2^{0}	2^{2}	2^{0}	2^{-2}
wine	2^{0}	2^{1}	2^{0}	2^{-1}	2^{-2}	2^{-2}
zoo	2^{-1}	2^{3}	2^{0}	2^{2}	2^{0}	2^{0}
tae	2^{2}	2^{10}	2^{3}	2^{10}	2^{-3}	2^{5}
lymph	2^{-2}	2^{4}	2^{-3}	2^{2}	2^{4}	2^{0}
dermatology	2^{-2}	2^{1}	2^{-2}	2^{0}	2^{1}	2^{2}

由表 2 可知, 除了 dermatology 数据集, HA-MBSVM 在所有数据集上的所有指标均为最优或仅稍次于最优; 在 iris 和 tae 数据集上, HA-MBSVM 的 4 项指标全部优于其他算法; 在 tae 数据集上, HA-MBSVM 的提升幅度尤为显著, 相比于表现最差的算法, 4 项指标的提升幅度可达 17.53%–19.60%, 这证明了 HA-MBSVM 的有效性和较好的通用性. 从整体上看, HA-MBSVM 在所有数据集上对 MBSVM 的分类性能都有所提升, 但在 zoo 和 dermatology 数据集上, 提升幅度有限, HA-MBSVM 与 MBSVM 表现相似, 这一方面是因为这两个数据集的样本个数较少而类别数较多, 所以各类别数据分布的差异性不突出, 不易生成合适的类别划分结果, 另一方面是因为数据集本身分布相对均衡, 在不进行类别划分等操作的情况下就能达到较好的分类性能.

表 4 HA-MBSVM 的最优参数 (评估指标为准确率)

数据集	σ_1	C_1	ξ	σ_2	C_2	$Maxgen_2$	d
iris	2^{-2}	2^{-2}	0.6	$2^{0.75}$	2^{3}	150	2^{0}
wine	2^{-2}	2^{-2}	0.4	$2^{0.75}$	2^{3}	210	2^{3}
zoo	2^{-2}	2^{-2}	0.6	2^{2}	2^{15}	90	2^{0}
tae	2^{4}	2^{-2}	0.4	2^{-3}	2^{9}	90	2^{2}
lymph	2^{-2}	2^{-2}	0.4	$2^{0.75}$	2^{3}	90	2^{3}
dermatology	2^{-2}	2^{-2}	0.6	2^{2}	2^{15}	90	2^{6}

3.2 参数讨论

在本节中, 我们讨论类别划分停止阈值ξ和构建分类器时 RBF 的带宽σ_2对算法性能的影响.

对于ξ, 我们将 HA-MBSVM 中的其他参数固定为表 4 中的最优参数, ξ的候选值集合为 [0.2, 0.4, 0.6, 0.8, 1], 在各数据集上进行 5 次 5 折交叉验证实验, 取 5 次实验的平均准确率为实验结果, 具体数值如表 5 所示. ξ用于控制类别划分的过程, ξ取值越小, 对类别标签集进行划分的次数越多, 产生的标签块越多. 由表 5 可见, 不同数据集的最优ξ取值不完全相同, 但较多地集中在 0.4 与 0.6 之中. 这是因为 5 折交叉验证中每一折的实验都会采用相同的ξ进行类别划分, 0.4 与 0.6 的取值能够较好地依据每一折训练集的数据分布产生合适的类别划分结果. 除了 tae 和 wine 数据集, HA-MBSVM 在其余数据集上, ξ不为 1 时的性能均显著优于ξ为 1 时的性能, 这说明了进行类别划分再构建分类器的必要性与有效性. 此外, 同一数据集上不同ξ取值的实验结果可能较为接近, 如 iris 数据集上, ξ取值为 0.4 与 0.6 时的实验结果仅相差 0.0013, 这是因为相近的ξ取值可能会产生相同的类别划分结果.

对于σ_2, 我们将 HA-MBSVM 中其他参数固定为表 4

的最优参数, σ_2的候选值集合为$[2^{-3},2^{-1.75},2^{-0.5},\cdots,2^{7}]$, 具体实验结果如表 6 所示. σ_2为 RBF 的带宽, 利用核函数可以将样本从原始空间映射到一个更高维的特征空间, 使得样本在这个特征空间内线性可分. σ_2的改变实质上是 HA-MBSVM 向高维度映射的特征空间复杂度的改变. 当σ_2增大时, 高维特征空间的复杂度降低, 线性可分程度也将降低; 而当σ_2趋向于 0 时, 高维特征空间的复杂度趋向于无穷, 此时虽能将任意数据映射为线性可分, 但往往会造成严重的过拟合问题[32]. 由表 6 可见, 当σ_2取值较小时, 除 iris 数据集外, HA-MBSVM 在其余数据集上的性能都出现严重下降, 而当σ_2取较大值时, HA-MBSVM 在所有数据集上的表现都很差, 这是高维特征空间复杂度的极端变化带来的影响. 此外, 不同数据集的最优σ_2取值不完全相同, 但主要集中在$2^{0.75}$和2^{2}之中, 这是因为这两个取值相对适中, 映射的高维特征空间的复杂度能够较好地权衡样本的线性可分程度与模型的泛化能力.

表 5 HA-MBSVM 在不同ξ取值下的准确率

数据集	0.2	0.4	0.6	0.8	1
iris	0.9600	**0.9640**	0.9627	0.9587	0.9227
wine	0.9866	0.9876	0.9845	**0.9898**	0.9865
zoo	0.9170	0.9327	**0.9370**	0.9305	0.8846
tae	0.5919	**0.6397**	0.6003	0.6051	0.6141
lymph	0.8147	0.8133	**0.8238**	0.8204	0.7873
dermatology	0.9542	**0.9552**	0.9481	0.9429	0.8744

表 6 HA-MBSVM 在不同σ_2取值下的准确率

数据集	2^{-3}	$2^{-1.75}$	$2^{-0.5}$	$2^{0.75}$	2^{2}	$2^{3.25}$	$2^{4.5}$	$2^{5.75}$	2^{7}
iris	0.9467	0.9400	0.9547	0.9600	**0.9640**	0.9307	0.3587	0.3333	0.3333
wine	0.4694	0.9823	0.9649	**0.9910**	0.9193	0.9773	0.9482	0.4450	0.3566
zoo	0.5307	0.6218	0.8733	**0.9408**	0.9307	0.9289	0.7759	0.4062	0.4061
tae	0.5931	**0.6183**	0.4883	0.4367	0.4156	0.3669	0.3934	0.3342	0.3379
lymph	0.5479	0.6028	0.7887	0.8095	**0.8110**	0.7736	0.7953	0.5796	0.4937
dermatology	0.3103	0.3215	0.9005	0.9513	**0.9537**	0.9508	0.8978	0.5106	0.3193

3.3 消融实验

为了更好理解 HA-MBSVM 中不同部分对算法整体性能提升的影响, 我们设计了 HA-MBSVM 的 5 种变体, 进行全面的消融实验.首先, 我们将 MBSVM 设置为 base 模型, 以此为基础逐步添加 HA-MBSVM 中不同的设计组件, 得到对应的变体如下所示.

(I) base+超球与类别划分: base 中的 RBF 带宽与惩罚参数取表 3 中 MBSVM 的最优参数, 拟合超球时 RBF 带宽与惩罚参数候选值集合为$[2^{-2},2^{4},2^{10}]$, 划分停止阈值的候选值集合为$[0.2,0.4,0.6,0.8]$.

(II) base+ASSRFOA: base 中的 RBF 带宽与惩罚参数取表 3 中 MBSVM 的最优参数, ASSRFOA 最大迭代次数的候选值集合为$[90,150,210]$.

(III) base+“最近”的决策方式: base 中的 RBF 带宽候选值集合为$[2^{-10},2^{-9},2^{-8},\cdots,2^{4}]$, 惩罚参数候选值集合为$[2^{-2},2^{-3},2^{-4},\cdots,2^{12}]$.

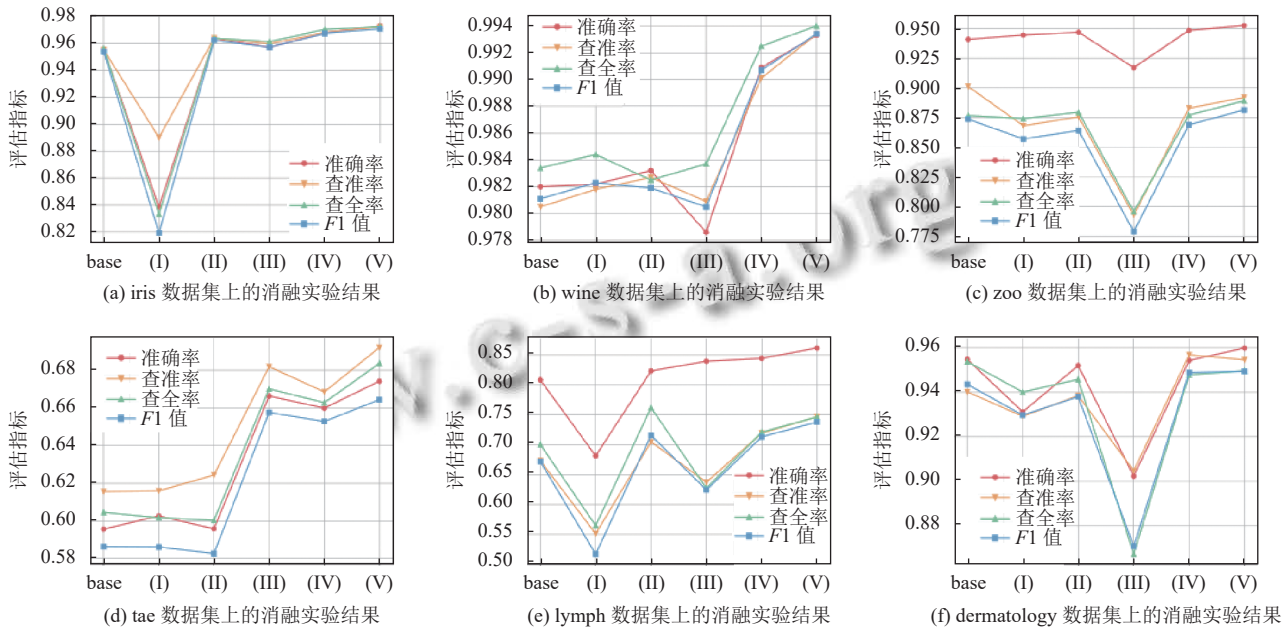
(IV) base+超球与类别划分+ASSRFOA+“最近”的决策方式: 我们通过将 HA-MBSVM 中的约束距离调节因子固定为 1, 其余参数取表 4 中 HA-MBSVM 的最优参数来实现该变体.

(V) base+超球与类别划分+ASSRFOA+“最近”的决策方式+约束距离调节因子: 该变体即为 HA-MBSVM. 我们取约束距离调节因子候选值集合为$[2^{1},2^{2},2^{3},2^{4},2^{5},2^{6}]$, 其余参数取表 4 中的最优参数.

依据如上参数设定, 我们在各数据集上进行 5 次 5 折交叉验证实验, 最优结果作为单次实验结果, 5 次实验的平均结果作为最终实验结果, 如图 3 所示.

由图 3 可知, 变体 (I)、(II)、(III) 在各数据集上表现不一, 从整体上看, 这 3 种变体无法有效提升 base 模型性能. 变体 (I) 添加了超球与类别划分模块, 但仍采用“最远”的决策方式, 在各数据集上表现均较差, 这说明“最远”的决策方式不适用于进行了类别划分的情况. 变体 (II) 仅将 QPP 求解算法换为 ASSRFOA, 除了 dermatology, 在其他数据集上均有两项以上的指标优于 base 模型, 尤其在 iris 和 lymph 上, 各指标均优于 base 模型, 这说明使用 ASSRFOA 求解 QPP 的有效性. 但由于变体 (II) 缺少超球与类别划分模块, 且未采用“最近”的决策方式和添加约束距离调节因子, 所以性能提升幅度有限. 变体 (III) 采用“最近”的决策方式, 但缺少超球与类别划分模块, 在各数据集上表现很不稳定, 在 iris 和 tae 上各指标均优于 base 模型, 但在 zoo 和 dermatology 上所有指标均不如 base 模型, 这说明“最近”的决策方式和超球与类别划分模块结合使用的

重要性. 对于变体 (IV), 仅在 zoo 上的查准率和 $F1$ 值以及 dermatology 上的准确率和查全率略低于 base 模型, 其余数据集上 4 项指标均优于 base 模型, 这说明组合使用 HA-MBSVM 的各组件能有效提升 base 模型性能. 变体 (V) 在变体 (IV) 基础上加入约束距离调节因子, 仅在 dermatology 上的查准率略低于变体 (IV), 而在其余数据集上 4 项评估指标均优于变体 (IV), 这说明引入约束距离调节因子的有效性. 图 3(b) 与 (d) 可见在 wine 和 tae 上, 变体 (IV) 和变体 (V) 对 base 模型性能有显著提升.



(a) iris 数据集上的消融实验结果 (b) wine 数据集上的消融实验结果 (c) zoo 数据集上的消融实验结果

(d) tae 数据集上的消融实验结果 (e) lymph 数据集上的消融实验结果 (f) dermatology 数据集上的消融实验结果

图 3 HA-MBSVM 的消融实验结果

4 结论

本文对 MBSVM 进行改进, 提出了基于超球和 ASSRFOA 的多生支持向量机: HA-MBSVM. 该算法利用拟合超球得到的球心和半径先进行类别划分再构建分类器, 同时运用全局优化能力较好的 ASSRFOA 来求解算法中的 QPP, 为了调整超平面与负样本之间的约束距离, 适当提高分类器的差异性, 我们还添加了一个约束距离调节因子. 在 6 个数据集上的实验结果表明 HA-MBSVM 的整体性能优于常见的 SVM 多分类算法以及 MBSVM. 但 HA-MBSVM 中涉及参数较多, 需要求解的 QPP 较多, 如何更高效地选取最优参数, 减少训练时间是未来的研究方向.

参考文献

1 Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, *et al*. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 2020, 408: 189–215. [doi: 10.1016/j.neucom.2019.10.118]

2 张侠. 基于 SVM 和逻辑回归的糖尿病数据分析与研究. 沧州师范学院学报, 2023, 39(1): 19–23, 84. [doi: 10.13834/j.cnki.czsfxyxb.2023.01.018]

3 李邦凤, 付玉苹, 龚良庚, 等. CT 影像组学结合支持向量机对偶发急性及陈旧性椎体压缩性骨折的鉴别诊断价值. 中国 CT 和 MRI 杂志, 2023, 21(2): 149–150, 174. [doi: 10.3969/j.issn.1672-5131.2023.02.050]

4 滕凯迪, 赵倩, 谭浩然, 等. 基于 SVM-KNN 算法的情绪脑电识别. 计算机系统应用, 2022, 31(2): 298–304. [doi: 10.15888/j.cnki.csa.008332]

5 韩伟, 韩士举, 魏延, 等. 支持向量机在消化系统疾病诊疗中的应用. 胃肠病学和肝病学杂志, 2022, 31(4): 454–458. [doi: 10.3969/j.issn.1006-5709.2022.04.021]

6 肖永茂, 鄢威, 龚青山. 模糊熵特征选择与 SVM 在三相异步电机故障诊断中的应用. 机械设计与制造, 2023, (3): 207–211. [doi: 10.19356/j.cnki.1001-3997.2023.03.010]

7 李燕飞, 李春光, 卜笛祺. SVM 在机械液压传动系统故障预测中的应用研究. 自动化与仪器仪表, 2023, (2): 42–45. [doi: 10.14016/j.cnki.1001-9227.2023.02.042]

8 江勋林. 多目标支持向量机及其在少样本故障诊断中的应用. 计算机系统应用, 2022, 31(9): 287–293. [doi: 10.15888/j.cnki.csa.008716]

9 郝万亮, 边英杰, 申献芳, 等. 基于支持向量机回归的航空

装备故障预测. 直升机技术, 2022, (4): 1–4, 9. [doi: 10.3969/j.issn.1673-1220.2022.04.001]
10 胡牡华. 支持向量机的舰船图像识别与分类技术. 舰船科学技术, 2022, 44(11): 156–159. [doi: 10.3404/j.issn.1672-7649.2022.11.032]
11 潘惠苹, 任艳, 徐春. 基于核典型相关分析和支持向量机的图像识别技术. 南京理工大学学报, 2022, 46(3): 284–290. [doi: 10.14177/j.cnki.32-1397n.2022.46.03.005]
12 吴晔, 李成辉, 姚骏. 基于支持向量机的眼底图像视盘定位算法. 工业控制计算机, 2023, 36(2): 98–99, 101. [doi: 10.3969/j.issn.1001-182X.2023.02.039]
13 张松兰. 支持向量机的算法及应用综述. 江苏理工学院学报, 2016, 22(2): 14–17, 21. [doi: 10.3969/j.issn.1674-8522.2016.02.004]
14 刘方园, 王水花, 张煜东. 支持向量机模型与应用综述. 计算机系统应用, 2018, 27(4): 1–9. [doi: 10.15888/j.cnki.csa.006273]
15 Sheng WJ, Liu YT, Söffker D. A novel adaptive boosting algorithm with distance-based weighted least square support vector machine and filter factor for carbon fiber reinforced polymer multi-damage classification. Structural Health Monitoring, 2023, 22(2): 1273–1289. [doi: 10.1177/14759217221098173]
16 Han SJ, Wang HR, Hu XY, *et al*. Research on tower mechanical fault classification method based on multiclass central segmentation hyperplane support vector machine improvement algorithm. Applied Sciences, 2023, 13(3): 1331. [doi: 10.3390/APP13031331]
17 Clement D, Agu E, Suleiman MA, *et al*. Multi-class breast cancer histopathological image classification using multi-scale pooled image feature representation (MPIFR) and one-versus-one support vector machines. Applied Sciences, 2022, 13(1): 156. [doi: 10.3390/APP13010156]
18 Li Q, Liu C, Guo QX. Support vector machine with robust low-rank learning for multi-label classification problems in the steelmaking process. Mathematics, 2022, 10(15): 2659. [doi: 10.3390/MATH10152659]
19 Barman U, Choudhury RD. Soil texture classification using multi class support vector machine. Information Processing in Agriculture, 2020, 7(2): 318–332. [doi: 10.1016/j.inpa.2019.08.001]
20 王乃芯. 多分类支持向量机的研究 [硕士学位论文]. 上海: 华东师范大学, 2020. [doi: 10.27149/d.cnki.ghdsu.2020.001477]
21 Angulo C, Parra X, Català A. *K*-SVCR. A support vector machine for multi-class classification. Neurocomputing, 2003, 55(1–2): 57–77. [doi: 10.1016/S0925-2312(03)00435-1]
22 Jayadeva, Khemchandani R, Chandra S. Twin support vector machines for pattern classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905–910. [doi: 10.1109/TPAMI.2007.1068]
23 Yang ZX, Shao YH, Zhang XS. Multiple birth support vector machine for multi-class classification. Neural Computing and Applications, 2013, 22(1): 153–161. [doi: 10.1007/s00521-012-1108-x]
24 丁世飞, 张健, 张谢锴, 等. 多分类孪生支持向量机研究进展. 软件学报, 2018, 29(1): 89–108. [doi: 10.13328/j.cnki.jos.005319]
25 王丽娜. 果蝇优化算法的改进研究 [硕士学位论文]. 赣州: 江西理工大学, 2021. [doi: 10.27176/d.cnki.gnfyc.2021.000431]
26 高栋. 群智能优化算法的改进研究 [硕士学位论文]. 赣州: 江西理工大学, 2020. [doi: 10.27176/d.cnki.gnfyc.2020.000149]
27 张水平, 王丽娜. 果蝇优化算法的进展研究分析. 计算机工程与应用, 2021, 57(6): 22–29. [doi: 10.3778/j.issn.1002-8331.2011-0174]
28 朱美琳, 刘向东, 陈世福. 用球结构的支持向量机解决多分类问题. 南京大学学报 (自然科学), 2003, 39(2): 153–158. [doi: 10.3321/j.issn:0469-5097.2003.02.002]
29 Mangasarian OL, Wild EW. Multisurface proximal support vector machine classification via generalized eigenvalues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(1): 69–74. [doi: 10.1109/TPAMI.2006.17]
30 王念, 张靖, 李博文, 等. 基于加权果蝇优化算法的多区域频率协同控制. 电力系统保护与控制, 2020, 48(11): 102–109. [doi: 10.19783/j.cnki.pspc.190864]
31 谢志强, 高丽, 杨静. 基于球结构的完全二叉树 SVM 多类分类算法. 计算机应用研究, 2008, 25(11): 3268–3270, 3274. [doi: 10.3969/j.issn.1001-3695.2008.11.019]
32 周广悦, 李克文, 刘文英, 等. 灰狼优化的混合参数多分类孪生支持向量机. 计算机科学与探索, 2020, 14(4): 628–636. [doi: 10.3778/j.issn.1673-9418.1905024]

(校对责编: 孙君艳)