DOI:10.11918/202407085

局部密度最小不确定性的SVM样本选择算法

周 玉[1],刘虹瑜[1],李京京[2],丁红强[2],白 磊[1]

(1.华北水利水电大学 电气工程学院,郑州 450011; 2.河北省水利工程局集团有限公司,石家庄 050021)

摘 要:为解决支持向量机(SVM)在分类时通常含有大量的冗余样本,从而导致面对较大规模数据集时SVM计算复杂度受到限制的问题,提出一种局部密度最小不确定性的SVM样本选择算法。该方法对决策面影响较大的边界数据进行有效选择,通过提取可能含有支持向量的训练样本,降低计算开销,进而提高SVM性能。首先,计算训练样本的K互近邻个数与高斯核密度估计。其次,将K互近邻个数与高斯核密度估计进行加和得到每个样本点的K局部密度并获取密度矩阵。然后,利用局部密度不确定性平衡优化方法,将密度矩阵进行三值映射后使不确定性改变量达到最小时得到最优阈值,并划分密度矩阵为中心数据与边界数据。最后,提取边界数据并作为SVM的训练样本建立分类模型。结果表明:利用该方法在UCI数据集上与其他6种常用样本选择方法进行实验对比,以准确率、保存率作为性能指标,文中提出的算法可以迅速划分中心数据与边界数据并删除大量冗余的训练样本,有效降低SVM的训练负担的同时提高了分类性能。

关键词:支持向量机(SVM);样本选择;局部密度;不确定性平衡;分类

中图分类号:TP181 **文献标志码**:A **文章编号**:0367-6234(2025)08-0045-12

Sample selection algorithm for SVM with minimum uncertainty in local density

ZHOU Yu[1], LIU Hongyu[1], LI Jingjing[2], DING Hongqiang[2], BAI Lei[1]

(1. School of Electrical Engineering, North China University of Water Resource and Electric Power, Zhengzhou 450011, China;
2. Hebei Water Conservancy Engineering Bureau Group Limited, Shijiazhuang 050021, China)

Abstract: To address the issue that support vector machines (SVM) frequently encompass a considerable number of redundant samples during classification, which restricts the computational complexity of SVM when confronted with large-scale datasets, a SVM sample selection algorithm based on local density minimum uncertainty is put forward. This approach efficiently identifies influential boundary data points that significantly affect the decision boundary, subsequently reducing computational costs by isolating potential support vectors from the training set, thereby bolstering SVM's overall effectiveness. Firstly, the number of K nearest neighbors and Gaussian kernel density estimation of the training samples are computed; Secondly, the sum of the number of K nearest neighbors and Gaussian kernel density estimation is derived for each sample point to acquire the K local density and obtain the density matrix; Subsequently, employing the local density uncertainty balancing optimization method, the density matrix undergoes a triple-mapping process to minimize uncertainty changes, yielding the optimal threshold. This threshold then partitions the density matrix into center data and boundary data. Finally, the boundary data are extracted and utilized as training samples for the SVM, enabling the establishment of an effective classification model. To experimentally evaluate the efficacy of our method, we compared it with six commonly utilized sample selection techniques on UCI datasets, employing accuracy and preservation rate as key performance metrics. The findings indicate that the method introduced in this paper significantly reduces the number of redundant training samples, thereby effectively alleviating the training burden on SVM and enhancing its classification performance.

Keywords: support vector machine(SVM); sample selection; local density; balance of uncertainty; classification

支持向量机(support vector machines, SVM)由Cortes等[1]于1995年提出,是一种基于VC维理论(Vapnik-Chervonenkis dimension)和结构风险最小化(structural risk minimization, SRM)原理的监督学习方法。SVM凭借坚实的统计学习理论在分类问题上具有许多优点,譬如SVM根据结构风险最小化原则同时考虑了经验风险和结构风险,这样避免了因样本量较少时经验风险最小化导致过拟合,提升了泛

收稿日期:2024-07-30;**录用日期**:2024-10-08;**网络首发日期**:2025-07-03
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20250703.1133.008
基金项目:国家自然科学基金(U1504622,31671580);河北省水利科技计划项目(2022-64)
作者简介:周 玉(1979—),男,副教授,硕士生导师
通信作者:周 玉,zhouyu_beijing@126.com

化性能[1-2]；SVM 解决了凸二次规划(quadratic programming, QP)问题,使其不会陷入局部最小值而是转为找到全局最小值[3];此外 SVM 通过最大化两类数据集间的距离得出稀疏解,从而以较低的复杂度提高了其泛化性[4]。由于具有上述优势,SVM 被广泛应用于面部识别[5-6]、疾病诊断[7-8]、文本识别[9-10]、故障诊断[11-12]等领域。目前 SVM 在处理上述实际问题时,往往需要使用核技巧[13-14]将训练数据映射至高维空间,使数据点在高维空间变得线性可分。虽然核技巧提高了支持向量机的分类精度,但映射数据时需要大量的计算量,训练时间也会随之增加,在解决较大规模的数据集时,SVM 分类器的效率会随着训练样本的增加而严重下降[15]。

值得注意的是,仅有一部分被称为支持向量(support vector, SV)的训练样本会影响 SVM 超平面的构建,因此可以去除掉与 SV 不相关的训练样本,保留与 SV 关系较为密切的训练样本,这样既不影响决策超平面的构建,也在一定程度上减少了训练数据。综上所述,减少冗余的非 SV 数据样本后再进行训练是提高 SVM 面对较大规模数据时提高训练效率的办法[16]。

近年来,研究者从基于聚类、几何分析等方面对训练数据进行样本选择,以降低 SVM 的计算开销。基于聚类的样本选择方法是一种被普遍应用的无监督机器学习方法,它将训练样本分为几个簇,每个簇的样本点相比其他簇的样本点更为相似,将聚类后的训练数据进行样本选择,以减少训练时间[15]。如 Barros de Almeida 等[17]提出的 SVM-KM 算法,该算法采用 K 均值聚类将训练样本聚成多个簇,其中包含单类标签的簇和多类标签的簇,然后仅保留单类标签簇的质心以及多类标签的所有数据点,将保留的点对 SVM 进行训练。该方法减少了训练样本的同时也降低了计算复杂度,但容易受训练数据集中特征的影响,并且对高维、稀疏的数据集进行分类时性能会下降。Shen 等[18]提出一种 K 均值聚类后引入 Fisher 判别比的方法来缩减训练样本,该方法采用在簇中寻找边界样本来减少冗余数据的思想,经过两阶段来完成训练样本的筛选。1)使用 K 均值聚类,根据聚类后训练样本的类别标签将簇进一步划分为更小的簇,然后通过使用簇的质心作为训练数据来获得近似超平面,并采用最大最小距离聚类(max-min cluster distance, MMCD)算法去除远离近似超平面的冗余簇;2)通过快速迭代(fast incremental false discovery rate, FIFDR)算法删除每个剩余簇中的数据点,最后将保留点反馈入 SVM 进行分类。该方法相较原始训练数据训练准确率有所下降,但去除了较多冗余点,显著减少了训练时间。此外模糊聚类也被经常应用于样本选择,周玉等[19]提出了一种基于模糊 C 均值聚类(Fuzzy C-means clustering, FCM)诱导出阴影集(shadowed sets)获取核心数据与边界数据的方法,将获取后的边界数据放入分类器训练。结果表明,与传统分类器相比,改进后的分类器不仅节约了训练时间,而且网络的泛化能力和分类识别准确度得到了有效提高。苏小红等[20]利用阴影集对模糊集的分析能力,提出一种基于阴影集的模糊支持向量机样本选择方法,将模糊集划分为 3 个子集后在可信任和不确定子集中进行样本选择,并且采用子空间样本选择和边界向量提取的方法选样。结果表明,该方法在保持 SVM 泛化能力的前提下有效降低了选样率和训练时间。张代俐等[21]通过模糊隶属度函数计算每个样本的隶属度,利用隶属度评估每个样本的重要程度,基于 3 种不同的模糊隶属度函数,分别提出了基于类中心距离、核目标对齐和中心核对齐模糊隶属度函数的 SVM 样本约简算法,相比传统 SVM 在几个分类指标都有提升。上述基于聚类的样本选择算法对聚类性能均有一定要求,当聚类效果在数据集上表现不理想时,会影响后续样本选择的数据点进而导致分类性能下降。

相比基于聚类的样本选择算法,基于几何分析的方法较为简单直观,该方法从训练数据的几何形状如凸包(convex hull)或几何距离进行分析,判断处于数据边界的样本点位于超平面附近并将这部分数据点保留作为训练样本。Chau 等[22]提出的凸凹壳 SVM(compressed convex hull SVM, CCHSVM)算法,即网格处理后,利用凸包寻找极值点,然后使用 Jarvis March 方法来确定不可分点的凹(非凸)壳,最后将凸凹包的顶点应用于 SVM 训练。结果表明 CCHSVM 具有良好的分类精度,同时训练速度明显快于传统 SVM 分类器。Xu 等[23]提出一种基于凸包向量和样本距离的 SVM 主动学习算法,通过样本距离和凸包向量可以主动选择对当前 SVM 分类器

最有价值的样本,该算法比随机采样所需的标记样本明显减少,降低了学习中的样本标记成本。Zhang 等[24]借鉴 KNN 分类算法的思想,利用样本的几何特性,提出了一种简单的 SV 预提取方法。只要 K 选择合适,就可以提取出重要 SV 的边界样本,从而减少训练样本,大大加快训练过程。李福祥等[25]提出了一种利用边界点训练支持向量机的新方法。首先计算每两个样本之间的欧氏距离,找出每个样本点的同类近邻集和异类近邻集,根据该样本点到两个集合的距离,判断其是否可能成为边界点。其次根据每个样本近邻集中同类样本数目的多少来删减样本集,该方法在具有较好分类精度的同时有效减少了训练样本的数量。但基于几何分析的样本选择方法也存在不足之处,即选取的数据点大多在簇与簇相邻边界,而在非相邻边界上的一部分点也会被选取作为 SV 构造决策面,且凸包在选取该类点时数量较少,会导致采用高斯核函数时构造的决策面不能很好的区分不同类别数据点,因此在训练样本保存率较低的情况下分类性能不佳。

针对 SVM 在进行分类任务时,训练数据集通常存在大量的冗余数据,这些冗余样本点对决策边界的构造并没有贡献且严重增加了分类器的训练负担的问题,以及根据现有基于聚类、几何分析方法存在的优点与不足,本文提出一种局部密度最小不确定性的 SVM 样本选择算法来搜索并提取出边界数据,首先,计算每个训练样本的 K 互近邻数量与高斯核密度,接着将 K 互近邻数量与高斯核密度进行加和获得 K 局部密度并得出密度矩阵,然后使用局部密度不确定性平衡优化方法得到最优阈值对密度矩阵进行划分,将阈值以上的数据定义为中心数据即在簇中心周围的样本点进行删除,阈值以下的数据定义为边界数据,对其保留放入 SVM 训练。在人工数据集和 UCI 数据集上的实验结果表明,该方法有效地保留了簇与簇相邻与非相邻的边界数据,应用于高斯核或径向基函数核(radial basis function, RBF) SVM 上能构建较好的决策面,并且在保存率较低的情况下还能够获得比不删除训练数据时更高的准确率,提高了 SVM 的分类性能。

1　支持向量机原理简述

给定含有 m 个数据点的二分类样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$,$y_i\in\{-1,+1\}$,其中 $x\in\mathbf{R}^n$,SVM 分类器最基本的思想就是基于训练集 D 在样本空间中找到一个最优的分类超平面。对于线性可分的 SVM,即构造一个线性方程为 $\boldsymbol{w}^{\mathrm{T}}x+b=0$ 的超平面,该超平面能最大化两类间的安全边界[26],其中 $\boldsymbol{w}$ 为权重向量,b 为偏置。最优超平面可以通过以下凸二次规划问题来解决:

$$\min_{\boldsymbol{w},b,\xi_i}\frac{1}{2}\boldsymbol{w}^2+C\sum_i^m\xi_i$$
$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i,(i=1,2,\cdots,m),\xi_i\geqslant 0 \tag{1}$$

式中:ξ_i 为松弛变量用于惩罚错误分类,C 为正则化参数用于控制训练和边界误差之间的权衡。通常上述凸二次规划问题可以通过如下拉格朗日乘子法进行求解,该问题的拉格朗日函数可以写为

$$L(w,b,\alpha,\xi,\mu)=\frac{1}{2}w^2+C\sum_i^m\xi_i+\sum_{i=1}^m\alpha_i(1-\xi_i-y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b))-\sum_i^m\mu_i\xi_i \tag{2}$$

式中 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_3)$。此时求 $L(w,b,\alpha,\xi,\mu)$ 的极大值,即式(1)的极小值,令 $L(w,b,\alpha,\xi,\mu)$ 对 w,b,ξ_i 偏导为零后将 $L(w,b,\alpha,\xi,\mu)$ 中的 w,b 消去转化为 $\boldsymbol{w}^{\mathrm{T}}x+b$ 极值的对偶问题。通常情况下,原始样本空间可能不存在一个能正确划分两类样本的超平面,对于这样的问题,需要将样本从原始空间映射得到一个更高维的特征空间中,使样本在这个特征空间线性可分,此时对于非线性 SVM 来说构造的超平面对应模型表示为

$$f(x)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(x)+b \tag{3}$$

式中:$\boldsymbol{\varphi}(x)$ 为将 x 映射后的特征向量,$\boldsymbol{w}$、b 为模型参数,同时映射到特征空间之后的极值对偶问题的目标函数为

$$\max_{\alpha}\Big(\sum_{i=1}^m\alpha_i-\frac{1}{2}\sum_{i=1}^m\sum_{j=1}^m\alpha_i\alpha_j y_i y_j\boldsymbol{\varphi}(x_i)^{\mathrm{T}}\boldsymbol{\varphi}(x_j)\Big)$$
$$\text{s.t. }\sum_{i=1}^m\alpha_i y_i=0,(i=1,2,\cdots,m),0\leqslant\alpha_i\leqslant C \tag{4}$$

此时求解 $\boldsymbol{\varphi}(x_i)^{\mathrm{T}}\boldsymbol{\varphi}(x_j)$ 非常困难,因此需要引入一个核函数 $K(x_i,x_j)=\boldsymbol{\varphi}(x_i)^{\mathrm{T}}\boldsymbol{\varphi}(x_j)$ 来避免高维特征空间的内积计算,这样只需要通过函数 $K(\cdot,\cdot)$ 就可以得出计算结果。引入核技巧后的非线性

SVM 最优分类超平面的决策函数可以表示为

$$f(x) = \mathrm{sign}\{\sum_{i=1}^{m} \alpha_i y_i K(x_i, x_j) + b\} \tag{5}$$

高斯核也被称为径向基函数核(RBF)泛化性能较好,是目前使用最为广泛的核函数,本文的实验内容均在高斯(或 RBF)核 SVM 上进行。通过式(1) ~ (5)显示出 SVM 的一个重要性质:训练完成后,大部分的训练样本都不需保留,最终模型仅与支持向量 α_i 有关[27]。因此在训练前,尽可能将与支持向量相关的训练样本筛选出来后,再放入 SVM 分类器训练,这样可以减少训练数据量,在保证与未选择样本前的准确率相近的情况下,提高 SVM 的训练速度和分类效率。

2 局部密度的 SVM 样本选择算法

SVM 在实际分类任务中,支持向量 SV 是训练过程用于确定分类超平面的关键点,直接影响到分类决策边界的位置和方向。通常来说,类中心附近的密集样本点不包含 SV,而类外围的稀疏样本点更可能具有 SV[18]。由于支持向量往往位于不同类别的边界附近,这些边界区域的局部密度较低,因此局部密度低的点更有可能成为支持向量,通过计算每个点的局部密度并选择低密度点可以更有效地找到潜在的支持向量,从而提高 SVM 的训练效率和分类精度。为了有效提取样本中低密度的边界数据,提出一种局部密度最小不确定性的方法进行样本选择,由 K 互近邻个数及其高斯核密度加和得到的 K 局部密度,生成密度矩阵后采用不确定性平衡优化方法确定阈值划分密度矩阵,将密集数据视为中心数据删除,不确定数据与稀疏数据视为边界数据选择并应用于 SVM 训练进行分类。

2.1 K 互近邻及其高斯核密度估计

K 局部密度由 K 互近邻及其高斯核密度估计组成,因此需要先引入有关 K 近邻(K-nearest neighbors, KNN)的相关概念。假设数据集 $D = \{x_1, x_2, \cdots, x_n\}$,对于任意数据点 x 来说,K 近邻个数表示为

$$K_{\mathrm{KNN}(x)} = \{x_i \in D \mid d(x, x_i) \leqslant d(x, x_{(k)})\} \tag{6}$$

式中:$d(x, x_i)$ 为点 y 和 x_i 之间的欧氏距离,$d(x, x_{(k)})$ 为点 x 到其第 K 近邻的欧氏距离。由 K 近邻个数可以得出 K 互近邻个数的定义,对于集合 D 中的两个任意两个数据点 x_i, y,如果 x_i 是 y 的 K 近邻,同时 y 也是 x_i 的 K 近邻,那么该两点就是 K 互近邻(K-mutual nearest neighbors, KMNN),KMNN 的个数可以用集合表示为

$$K_{\mathrm{KMNN}(x)} = \{x_i \in D \mid x_i \in \mathrm{KNN}(x), x \in \mathrm{KNN}(x_i)\} \tag{7}$$

式中 $K_{\mathrm{KNN}(x)}$ 为点 x 的 K 个最近邻点。因为 $K_{\mathrm{KMNN}(x)}$ 有效反映了样本点在同一簇中的紧密联系程度,所以在聚类中可以识别密集的簇,同时 $K_{\mathrm{KMNN}(x)}$ 能够直观地将数据中的稀疏区域和密集区域等密度分布信息展现出来。因此,采用 $K_{\mathrm{KMNN}(x)}$ 的思想在 SVM 样本选择中有助于将含有少量 SV 存在的高密度样本点进行删除,这类点往往对分类超平面的构造作用较小。

此外,核密度估计(kernel density estimation, KDE)是一种非参数技术,用于估计样本数据上概率密度函数,它是数据分析的重要工具[28]。本文将利用核密度估计来分析 SVM 训练样本中样本分布的疏密情况,其中核函数选择高斯核,高斯核具有理想的数学特性[29],并且采用高斯核计算局部核密度可以保留数据点本身的分布特点。在数据点邻域有相同的 $K_{\mathrm{KMNN}(x)}$ 时能够突出局部核密度最高的数据点[30]。这进一步提高了对密集样本点的检索,帮助本文将其认为是对构造 SVM 决策超平面不重要的点从而进行删除。高斯核函数以及高斯核密度估计(Gaussian kernel density estimation, GKDE)的数学表达式如下:

$$\mathrm{Gaussian}(x) = \frac{\mathrm{e}^{(-\|x\|^2/2)}}{2\pi^d} \tag{8}$$

$$K_{\mathrm{GKDE}(x)} = \sum_{x_i \in K_{\mathrm{KNN}(x)}} \frac{\mathrm{e}^{(-d(x,x_i)^2/2)}}{2\pi^d} \tag{9}$$

式中:$\|x\|$为 x 的范数,d 为数据样本的维度,$d(x, x_i)$为点 x 与第 K 个最近邻点 x_i 的欧氏距离。最后将 KMNN 与 GKDE 进行相加得到新的 K 局部密度 $\boldsymbol{\rho}_k(x)$ 为

$$\boldsymbol{\rho}_k(x) = K_{\mathrm{KMNN}(x)} + K_{\mathrm{GKDE}(x)} \tag{10}$$

2.2 局部密度不确定性平衡优化方法

局部密度不确定性平衡方法参考了阴影集的思想。阴影集由模糊集诱导而来,目的是解决模糊集中使用具有精确数值的隶属度来描述模糊逻辑的缺陷问题,用于观察和解释不确定现象。在一个模糊

集中,找到一个阈值 α,将大于 $1-\alpha$ 部分的隶属度升高至 1,并将小于 α 的隶属度降低至 0,保持整体不确定性平衡便可以将传统隶属度函数转变为具有三值逻辑的阴影集[19]。其中求解最优阈值 α 是构造阴影集的关键,Pedrycz 等[31]提出一种基于模糊平衡的优化方法,使隶属度的增减量达到整体平衡(即不确定性平衡),如在离散情况下使下式中 V 应达到最小值,此时的阈值视为最优阈值 α_{opt}。其中 $A(x)$ 为在论域 X 中的模糊集诱导出的阴影集,映射关系为 $A: X \rightarrow \{0,1,[0,1]\}$。

$$V(\alpha) = \left| \sum_{i:A(x_i)<\alpha} A(x_i) + \sum_{i:A(x_i)>(1-\alpha)} [1-A(x_i)] - \text{card}\{x_i \in X \mid \alpha < A(x) < (1-\alpha)\} \right| \tag{11}$$

$$\alpha_{\text{opt}} = \arg\min_{\alpha} V(\alpha) \tag{12}$$

在聚类分析中,局部密度可以用来描述一个点在某个区域内的密集程度,若样本点局部密度越高,则说明该样本点更有可能属于某一类,通过式(10)可以得出每个点的 K 局部密度 ρ_k,将数据集 $D=\{x_1,x_2,\cdots,x_n\}$ 所有点的 ρ_k 输出后得出密度矩阵有 $\rho_k(x)=[\rho_k(x_1),\rho_k(x_2),\cdots,\rho_k(x_n)]$,本文将密度矩阵归一化至 $[0,1]$ 后类比于模糊集中的隶属度,进行构造阴影集的三值映射即 $\varphi: \rho_k \xrightarrow{\alpha} \{0,1,[0,1]\}$,其中 ρ_k 为密度矩阵的集合,φ 为密度矩阵的三值映射关系。为了确定三值划分的最优阈值,提出局部密度不确定性平衡优化方法如下式来选择阈值对密度矩阵进行划分:

$$V(\alpha) = \left| \sum_{i:\rho_k(x_i)<\alpha} \rho_k(x_i) + \sum_{i:\rho_k(x_i)>(1-\alpha)} [1-\rho_k(x_i)] - \text{card}\{x_i \in D \mid \alpha < \rho_k(x_i) < (1-\alpha)\} \right| \tag{13}$$

定义 1(密集数据)　对于 $\forall(x,\rho_k(x))$,$\rho_k(x)$ 为样本点 x 的 K 局部密度,若 $\varphi(x,\rho_k(x))=1$,即 $\rho_k(x)>1-\alpha$,则称样本点 x 为密集数据。

定义 2(稀疏数据)　对于 $\forall(x,\rho_k(x))$,$\rho_k(x)$ 为样本点 x 的 K 局部密度,若 $\varphi(x,\rho_k(x))=0$,即 $\rho_k(x)<\alpha$,则称样本点 x 为稀疏数据。

定义 3(不确定数据)　对于 $\forall(x,\rho_k(x))$,$\rho_k(x)$ 为样本点 x 的 K 局部密度,若 $\varphi(x,\rho_k(x))=(0,1)$,即 $\alpha \leqslant \rho_k(x) \leqslant 1-\alpha$,则称样本点 x 为不确定数据。

上述定义和公式解释为,当求出最优阈值 α 后,大于 $1-\alpha$ 的部分定义为 x 完全处于某类的密集核心区,即映射 φ 将局部密度升至 1 的部分;处于 $[\alpha,1-\alpha]$ 的部分定义为 x 不确定是否完全属于某类的区域,即映射 φ 处于 $[0,1]$ 局部密度不变的部分;而小于 α 的部分定义为稀疏边界区,即映射 φ 将局部密度降至 0 的部分。式(13)表示的是经过映射 φ 后,尽可能使局部密度的改变量达到整体不确定平衡,即使映射在 $[0,1]$ 的不确定性改变量 $V(\alpha)$ 达到最小时,取得最优阈值 α_{opt}。本文将密集数据视为对决策面影响较小的中心数据进行删除,将不确定数据以及稀疏数据视为边界数据予以保留。

2.3　算法步骤

局部密度最小不确定性的 SVM 样本选择算法步骤描述如下:

输入:算法参数 K、T,训练集。

输出:边界数据训练后的 SVM 分类器模型。

Step1　将数据划分为训练集与测试集。

Step2　通过式(7)、(9)计算训练样本的 $K_{\text{KMNN}(x)}$ 与 $K_{\text{GKDE}(x)}$ 并做归一化处理。

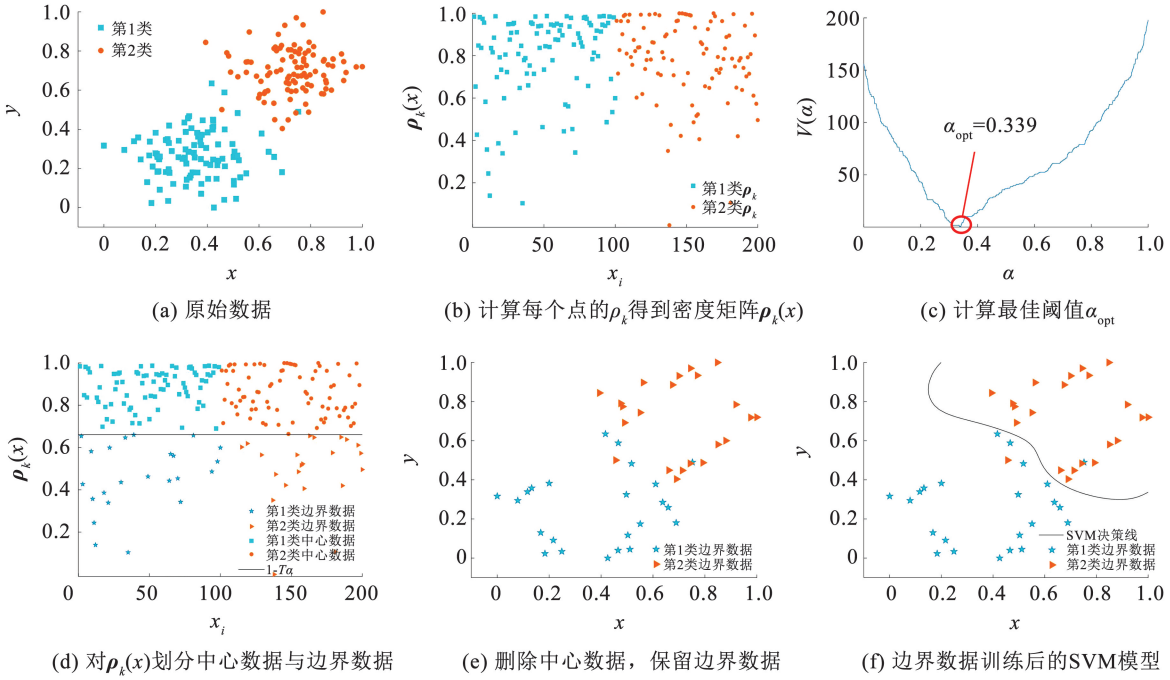
Step3　根据式(10)对 $K_{\text{KMNN}(x)}$ 和 $K_{\text{GKDE}(x)}$ 加和并进行归一化后得到每个样本点 ρ_k 以及密度矩阵 $\boldsymbol{\rho}_k(x)$。

Step4　对 $\boldsymbol{\rho}_k(x)$ 进行不确定性平衡优化,根据式(12)、(13)得出最优阈值 α_{opt}。

Step5　由定义 1 ~ 定义 3 将 $\boldsymbol{\rho}_k(x)$ 划分为 3 种数据后保留小于 $1-T\alpha_{\text{opt}}$ 的边界数据,其余样本点删除,其中控制参数 T 用来控制所选样本的数量。

Step6　将边界数据应用于 SVM 训练后得出分类模型在测试集上验证模型性能。

算法对边界数据进行样本选择的过程如图 1 所示,生成每类数据量 100 的二维随机数据,其中 x, y 轴表示人工数据集的两项特征做 $[0,1]$ 归一化处理。图 1(a)为原始数据;图 1(b)为计算出每个样本点 K 局部密度后归一化的密度矩阵 $\boldsymbol{\rho}_k(x)$,即每个样本点 $\{x_1,x_2,\cdots,x_{200}\}$ 对应的 K 局部密度 ρ_k;图 1(c)为求出最优阈值的过程,即式(13)中 $V(\alpha)$ 为最小值 0 时对应的 α_{opt};图 1(e)将 $\boldsymbol{\rho}_k(x)$ 中低于 $1-T\alpha_{\text{opt}}$ 视为边界数据予以保留,高于 $1-T\alpha_{\text{opt}}$ 视为中心数据进行删除;图 1(f)为保留的边界数据放入高斯核函数 SVM 训练后得出的决策线。



(a) 原始数据 (b) 计算每个点的ρ_k得到密度矩阵$\boldsymbol{\rho}_k(x)$ (c) 计算最佳阈值α_{opt}

(d) 对$\boldsymbol{\rho}_k(x)$划分中心数据与边界数据 (e) 删除中心数据，保留边界数据 (f) 边界数据训练后的SVM模型

图 1 算法可视化过程

Fig. 1 Visualization process of the algorithm

可以看出,使用本文提出的方法选出边界数据应用于高斯核 SVM 训练后的模型,与不做任何筛选样本的全部数据训练后得出的 SVM 模型(见图 2)相似。原因是本方法提取簇边界数据包含了足够多对高斯核决策线有重要影响的 SV,仅用了约原数据量 1/4 的训练样本即可生成与全数据训练相似的决策线。

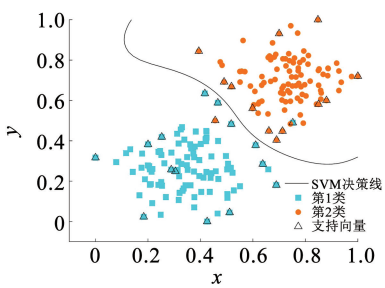


图 2 全部数据训练的 SVM 模型

Fig. 2 SVM model trained on the full dataset

2.4 计算复杂度分析

对于 n 个训练样本,局部密度最小不确定性的 SVM 样本选择算法计算复杂度分析为:算法中计算每个训练样本欧氏距离的时间复杂度为 $O(n^2)$;计算训练样本 $K_{\mathrm{KMNN}(x)}$、$K_{\mathrm{GKDE}(x)}$、计算 $\boldsymbol{\rho}_k(x)$ 以及局部密度不确定性平衡优化的时间复杂度均为 $O(n)$,因此本文样本选择算法的时间复杂度近似为 $O(n^2)$。相较于传统 SVM 的时间复杂度为 $O(n^3)$,本文提出的算法降低了时间复杂度,提高了训练效率。

3 实验与分析

为了验证局部密度最小不确定性的 SVM 样本选择算法的可靠性,本文介绍提出算法与 4 种样本选择算法即 KNN[24]、SS[19]、DR[18]、BPLSH[32] 以及两种常用对比方法即不进行样本选择(ALL)与同等样本量进行随机选择[23](Rand)的 LIBSVM,各方法详见表 1,在 12 个 UCI 数据集[33](见表 2)对比,测试本文方法与其他算法在 UCI 数据集的分类性能,以及介绍本文算法不同的 K、T 参数在 UCI 数据集对性能的影响。为了直观展示提出算法与其他对比方法提取的边界数据的区别,在构造的 Normal 数据集上进行实验,展示各个样本选择算法保存对决策面有关键影响样本的能力,在选择样本后的对原始数据进行测试后的实验结果见图 3、表 3。其中性能指标设置为准确率(accuracy rate, AR)、保存率(preservation rate, PR)可以分别表示为:

$$A=\frac{N_{\mathrm{CTE}}}{N_{\mathrm{TE}}}\times 100 \tag{14}$$

$$P=\frac{N_{\mathrm{STR}}}{N_{\mathrm{TR}}}\times 100 \tag{15}$$

式中:A 为准确率,P 为保存率,N_{TE}为测试集中的样本数目,N_{CTE}为 SVM 正确分类测试集的个数,N_{TR}为训练集的样本个数,N_{STR}为经过样本选择算法后的训练样本数[32]。实验环境为: Windows11 操作系统,Intel(R) Core(TM) i7 - 12650H 2.30 GHz,16 GB 内存,采用 MATLAB R2022b 编写。

表 1　对比方法介绍

Tab. 1　Introduction to comparison methods

方法	基本原理	年份
ALL	LIBSVM	1995
Rand[23]	LIBSVM	1995
KNN[24]	距离	2008
SS[19]	聚类	2013
DR[18]	聚类、密度	2016
BPLSH[32]	相似函数	2021

表 2　UCI 数据集

Tab. 2　UCI datasets

UCI 数据集	数量	特征
Diabetes	768	8
Heartstatlog	270	13
WBC	683	9
Bupa	345	6
Cancer	683	9
Blood	748	4
Haberman	306	3
Mammographic	961	5
WDBC	569	31
HTRU2	17 898	8
Magic	19 020	8
Credit Card	30 000	24

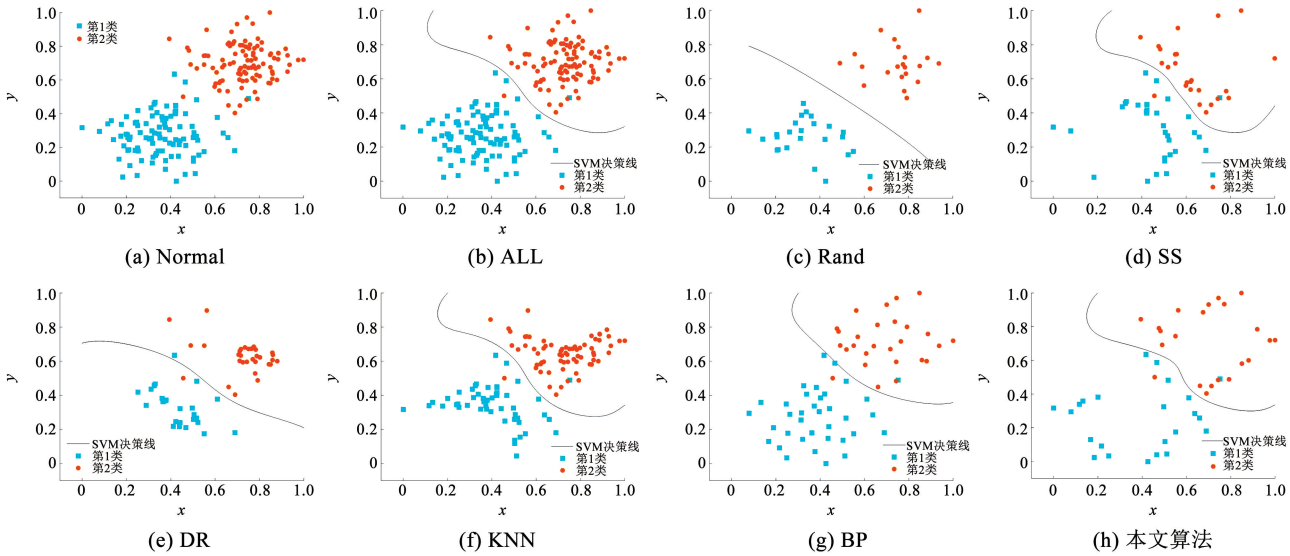


图 3　人工数据集上各样本选择算法得出的 SVM 模型

Fig. 3　SVM models obtained by various sample selection algorithms on the artificial dataset

表 3　人工数据集实验结果

Tab. 3　Experimental results on the artificial dataset

人工数据集	指标	Rand	KNN	SS	DR	ALL	BPLSH	本文算法
Normal	A	97.500 0	99.000 0	99.000 0	98.000 0	99.000 0	98.500 0	**99.000 0**
	P	22.500 0	60.500 0	28.500 0	29.500 0	100.000 0	31.500 0	**22.500 0**

3.1　参数对算法性能的影响

本文介绍局部密度最小不确定性的 SVM 样本选择算法中参数 K、T 在 UCI 数据集中对准确率、保存率的影响。图 4 展示了提出算法对全数据进行样本选择后与未进行样本选择(ALL)在十折交叉验证的准确率对比,图 5 为参数对保存率的影响变化。其中 LIBSVM 为 RBF 核函数,HTRU2 数据集参数 $c=100$,其余数据集 $c=10$,参数 g 为默认。K 选取范围与训练样本数量有关,在 Cancer 等 9 个数据集上,K 取 2 ~ 50,Credit Card、HTRU2、Magic 3 个数据集上 K 取训练样本数量 10% 左右的 20 个值,本文在 UCI 数据集中取 $T=0.25$,$T=1.00$,$T=1.15$ 这 3 个值进行测试。

由图 4、5 可以看出,参数 T 对训练样本的保存率的影响较为关键,$T=1.00$ 时保存率大都处于 50% 以下,参数 T 增大保存率下降,反之保存率上

升。并且在确定 T 后,K 的增大对样本保存率的影响会逐渐变小。经实验得出 T 取较小时,准确率随 K 的变化较为稳定,在规模较小的数据集如 Haberman,Bupa 上,$T=0.25$ 的准确率变化明显比其他值更稳定。此外,K 的变化对准确率也有很大影响,K 不宜过大也不宜过小。其中有以下原因,在保存率一定的情况下,当样本量较小时 K 增大会使簇间相邻边界数据的局部密度上升,此类样本点可能被划分为中心数据删除,导致其中包含的支持向量减少进而构建的决策模型分类性能不佳;同理,面对样本量较大的数据集时 K 过小也会使簇间相邻边界数据的选取率下降。如在 Haberman 上,K 过大会导致准确率出现大幅度下降;在 Credit Card 等 3 个大型数据集上,K 过小准确率也处于整体较低水平,因此要尽量避免上述现象的发生。总的来说,无论在全数据进行十折交叉训练测试或者 UCI 数据集分为 2∶1 训练测试的情况下,准确率达到峰值时对应的 K 与训练样本量密切相关,根据经验判断最优 K 通常处于训练样本量的 15% 以内。

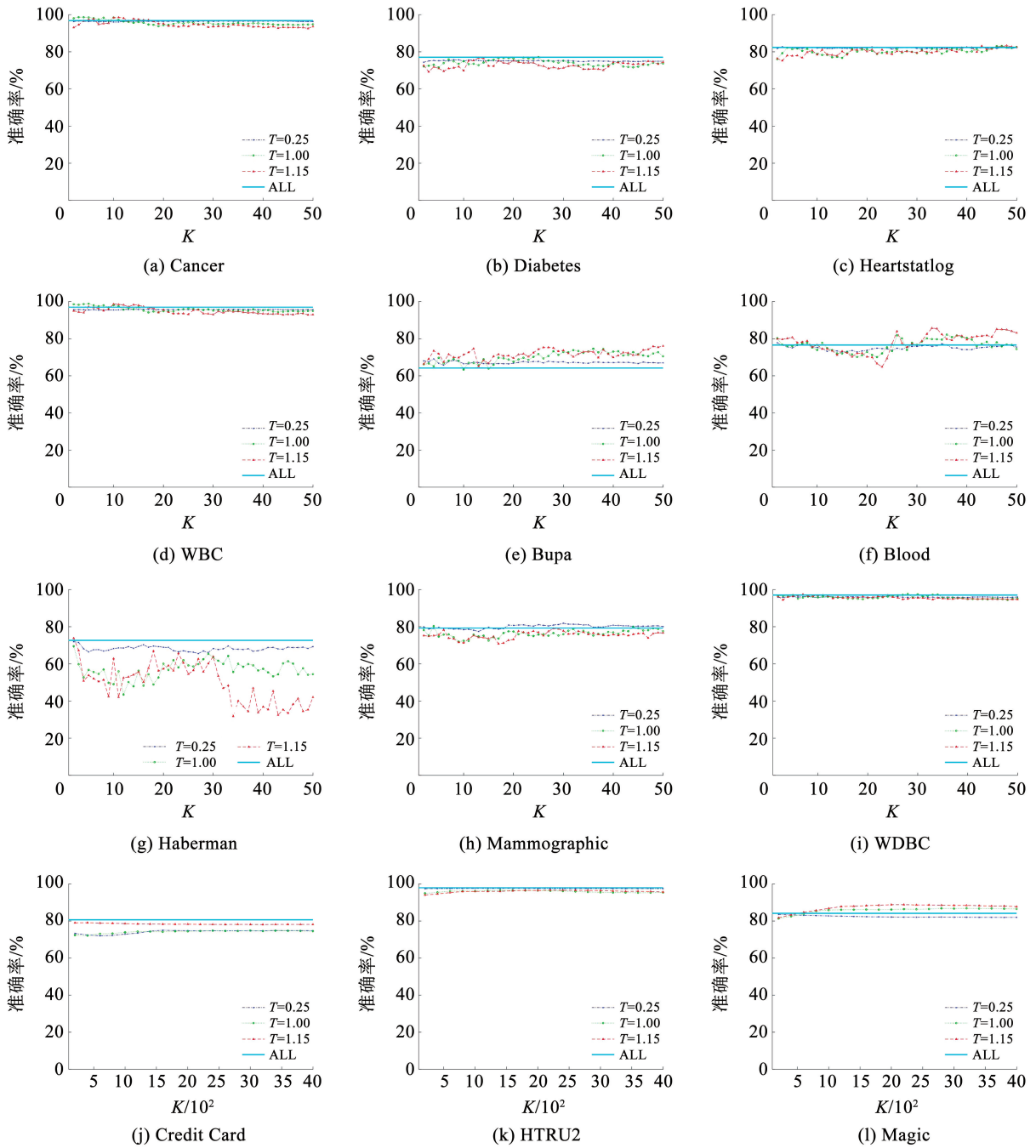


图 4 UCI 数据集上不同参数对准确率的影响

Fig. 4 Impact of different parameters on accuracy rate on UCI datasets

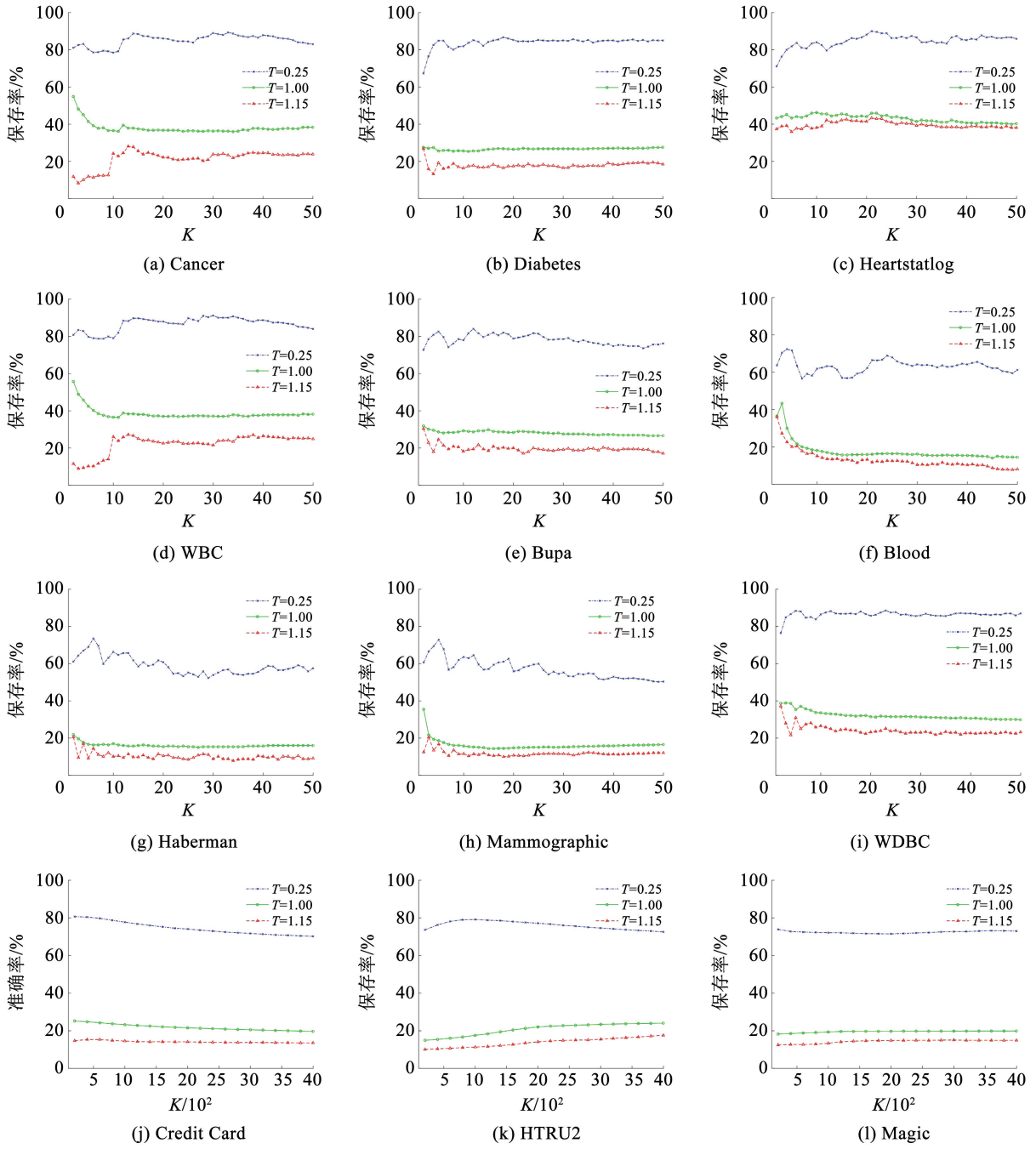


图 5　UCI 数据集上不同参数对保存率的影响

Fig. 5　Impact of different parameters on preservation rate on UCI datasets

3.2　UCI 数据集

本文在 12 个 UCI 数据集上与 6 种方法(见表 1)进行对比实验,以准确率(accuracy rate, AR)、保存率(preservation rate, PR)作为性能指标。实验过程将数据集分为 2/3 训练集与 1/3 测试集,每个数据集的各个算法运行 10 次后取均值进行对比测试。为减轻参数对结果的影响,在一定保存率的范围内,每次测试采用循环选取性能表现最好的参数作为结果进行记录见表 4。此外,为进一步测试算法泛化性能,对全部数据样本选择后再进行十折交叉验证并循环 10 次取平均值,准确率与保存率平均值见表 5。上述实验为避免不同特征维度之间的量纲可能会出现差异影响分类性能,实验中所有数据集的各项特征都进行了归一化处理,并对缺失值使用平均值代替。

表 4 UCI 数据集实验结果

Tab. 4 Experimental results on UCI datasets

UCI 数据集	指标	Rand	KNN	SS	DR	ALL	BPLSH	本文算法
Cancer	*A*	95.784 8	97.174 9	97.309 4	96.636 8	97.130 1	97.085 2	**97.623 3**
	P	24.217 4	27.739 1	30.913 0	**16.152 2**	100.000 0	38.043 5	24.217 4
Diabetes	*A*	74.244 6	76.223 0	75.143 9	72.474 8	77.122 3	77.482 0	**78.093 5**
	P	28.102 0	63.857 1	34.755 1	29.326 6	100.000 0	57.367 4	**28.142 9**
Heartstatlog	*A*	79.888 9	82.222 2	82.333 2	75.777 8	83.000 1	83.444 5	**83.888 9**
	P	40.611 1	51.111 1	49.166 7	**27.055 6**	100.000 0	68.555 6	40.611 1
WBC	*A*	96.053 8	96.771 3	96.591 9	95.156 9	96.681 6	96.816 1	**97.040 4**
	P	22.108 7	40.587 0	30.913 4	**12.347 8**	100.000 0	38.587 0	22.108 7
Bupa	*A*	56.381 0	60.285 7	57.999 9	57.904 8	60.000 0	61.809 5	**65.714 2**
	P	29.166 7	63.208 3	67.958 3	31.875 0	100.000 0	63.750 0	**29.166 7**
Blood	*A*	75.480 0	75.560 0	76.840 0	76.600 0	76.526 1	76.080 0	**78.440 0**
	P	12.228 9	59.116 5	21.285 1	27.088 3	100.000 0	48.393 6	**12.148 6**
Haberman	*A*	73.431 4	72.058 8	73.137 3	74.117 7	74.656 9	71.470 6	**74.804 0**
	P	20.392 1	60.245 1	24.068 6	30.490 2	100.000 0	45.098 0	**20.392 2**
Mammographic	*A*	76.465 9	81.059 2	81.059 2	70.841 1	80.078 1	77.352 0	**81.682 2**
	P	17.859 3	80.453 1	69.296 8	24.078 1	100.000 0	60.875 0	**17.859 3**
WDBC	*A*	97.142 8	97.542 9	97.460 5	96.171 4	97.614 2	97.428 6	**98.171 4**
	P	42.538 1	64.644 7	47.563 5	**36.091 4**	100.000 0	59.010 1	42.538 1
Credit Card	*A*	78.766 0	78.305 0	81.096 0	80.594 0	80.558 0	80.525 0	**81.359 0**
	P	20.681 5	40.736 0	22.087 0	**17.229 0**	100.000 0	91.691 5	20.681 5
HTRU2	*A*	97.789 1	97.926 6	97.772 4	94.421 8	97.918 2	97.968 5	**97.975 1**
	P	20.448 4	19.338 8	15.052 8	**12.581 3**	100.000 0	29.656 4	20.448 4
Magic	*A*	83.505 7	83.841 1	83.595 6	81.905 4	83.862 8	84.042 6	**83.979 5**
	P	72.007 1	76.134 9	76.791 8	**31.958 2**	100.000 0	81.070 2	72.007 1

表 5 UCI 数据集交叉验证结果

Tab. 5 Cross validation experiment results on UCI datasets

UCI 数据集	指标	Rand	KNN	SS	DR	ALL	BPLSH	本文算法		
								$T=0.25$	$T=1.00$	$T=1.15$
Cancer	*A*	96.742 5	94.842 9	94.628 2	91.087 5	96.853 9	92.464 3	96.651 0	**98.797 6**	98.308 3
	P	44.948 8	59.004 0	57.833 1	**14.055 6**	100.000 0	36.456 8	90.483 2	44.948 8	23.279 6
Diabetes	*A*	72.812 9	59.997 8	69.800 9	64.344 1	77.112 3	63.287 6	77.098 4	**77.456 7**	76.298 5
	P	26.822 9	55.989 6	35.156 2	29.036 5	100.000 0	53.125 0	81.640 6	26.822 9	**16.927 1**
Heartstatlog	*A*	79.921 3	67.546 2	75.793 0	68.194 8	82.370 4	78.287 4	82.403 6	83.134 8	**81.387 9**
	P	40.370 4	51.851 9	51.481 5	**31.111 1**	100.000 0	81.111 1	62.963 0	40.370 4	38.148 1
WBC	*A*	94.714 1	91.204 7	91.251 6	89.436 8	96.926 1	91.231 6	95.730 2	**98.753 6**	98.414 9
	P	52.173 9	40.434 8	31.304 3	**15.217 4**	100.000 0	37.826 1	88.579 8	42.459 7	23.865 3
Bupa	*A*	60.582 1	54.117 5	63.497 7	61.085 9	64.157 9	54.008 4	69.104 6	**74.578 7**	74.267 6
	P	27.246 4	59.710 1	70.724 6	26.956 5	100.000 0	62.318 8	86.087 0	27.246 4	**14.203 0**
Blood	*A*	74.123 3	64.220 0	72.719 8	74.285 7	76.620 0	59.908 8	78.362 9	80.331 8	**86.178 6**
	P	15.508 0	56.417 1	17.780 7	25.267 4	100.000 0	49.465 2	58.689 8	15.508 0	**10.561 5**
Haberman	*A*	70.826 4	54.349 7	63.083 3	63.182 5	72.682 0	43.420 0	71.937 4	69.495 2	**73.823 8**
	P	21.895 4	57.843 1	23.856 2	26.470 6	100.000 0	47.058 8	61.111 1	21.895 4	**20.582 0**

表 5(续)

UCI 数据集	指标	Rand	KNN	SS	DR	ALL	BPLSH	本文算法		
								$T=0.25$	$T=1.00$	$T=1.15$
Mammographic	*A*	77.889 9	74.695 2	80.194 0	70.530 5	79.295 8	69.339 4	**81.883 1**	80.347 4	79.057 6
	P	19.354 8	77.835 6	69.406 9	21.852 2	100.000 0	60.353 8	76.845 1	19.354 8	**16.208 4**
WDBC	*A*	96.044 9	94.588 5	96.297 7	95.745 2	97.099 7	95.527 4	96.274 2	**97.606 8**	97.398 5
	P	31.458 7	54.481 5	47.100 2	35.149 4	100.000 0	50.087 9	83.724 0	31.458 7	**27.416 5**
Credit Card	*A*	78.282 7	56.514 3	74.553 5	63.478 9	**80.672 0**	76.352 4	78.181 6	74.655 2	74.775 8
	P	21.583 3	40.613 3	22.033 3	24.150 0	100.000 0	87.666 7	71.765 3	21.583 3	**14.146 7**
HTRU2	*A*	97.490 8	88.400 1	94.922 8	89.042 6	**97.935 0**	91.530 3	97.352 3	96.595 2	96.850 4
	P	20.840 3	19.102 7	**12.336 6**	12.398 0	100.000 0	23.930 0	77.824 3	20.840 3	14.141 2
Magic	*A*	82.488 1	79.160 3	82.500 1	82.749 7	84.157 2	80.750 8	82.434 7	86.374 2	**88.982 3**
	P	19.763 4	75.326 0	77.681 4	**13.959 0**	100.000 0	78.916 9	71.629 9	19.763 4	14.826 5

从表 4、5 可以看出,本文提出的样本选择算法在训练数据保存率较低的同时仍然有着高于未进行样本选择的准确率,而其他样本选择算法在多种数据集上删除大部分训练数据会无可避免地使准确率下降。在交叉验证的实验中,提出算法的准确率在多个数据集中都有较好的表现,表明局部密度最小不确定性方法保留的边界数据相比其他算法有着更高的泛化性能,应用于 SVM 训练后准确率得到了提高,其中有以下原因,阴影集方法对边界数据进行筛选时易受聚类准确率影响,在聚类准确率较低时,必须扩大边界数据的数量才能够有较高准确率。KNN 方法筛选后的边界数据往往在簇与簇相邻的边界,而高斯核(或 RBF 核)SVM 的支持向量在相邻边界与非相邻边界上都可能存在,这可能会出现过拟合而无法形成良好的决策面导致准确率下降。DR 方法能保证较低的保存率,Fisher 比有效的删除了每个聚类簇的核心数据,但处于 K 均值聚类后边界簇的核心数据也可能包含 SV 的训练样本,删除这类样本会导致准确率下降。BPLSH 法保留了边界数据的同时也对核心数据进行保留,这在一定程度上提高了准确率并防止了过拟合现象,但也因此增加了许多非 SV 的冗余样本。而本文提出局部密度最小不确定性的 SVM 样本选择算法不受聚类的影响,能搜索并保留簇周围稀疏的边界样本点,其中的相邻边界数据是距离决策边界点较近的点,非相邻边界数据含有的部分 SV 防止出现过拟合现象,本文方法尽可能地将大部分对构建分类决策面有帮助的边界样本点保存下来,同时删除了大量处于类中心周围的冗余样本,提高了 SVM 的分类性能。

4　结　论

1)通过将 K 互近邻个数与高斯核密度估计进行加和定义了一种 K 局部密度,该局部密度能很好地突出簇中的高密度区域,在三值映射后,利用不确定性平衡优化方法可以将数据高效划分为边界数据与中心数据。

2)对边界数据进行样本选择并应用于 SVM 训练后,以准确率和保存率作为性能指标,在 UCI 数据集上的实验表现优秀,大幅减少了冗余训练样本,在有效降低训练负担情况下提高了分类准确率。

3)相较于其他 SVM 样本选择方法,局部密度最小不确定性方法在对相邻边界数据选择的同时还保留了足够多的非相邻边界数据,这样在面对多种形状的簇时仍然可以保持边界特征,可以建立分类性能较好的 SVM 模型。

4)尽管局部密度最小不确定性方法容易取得高准确率的 K、T,但分类性能达到最优时的算法参数仍然较难确定,因此未来的研究将结合不同数据集的分布特性,分析如何自适应求出最优参数的方法,进一步减少因主观因素导致算法性能下降的影响。

参考文献

[1] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273. DOI: 10.1007/BF00994018

[2] CRISTIANINI N, SHAWE-TAYLOR J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge: Cambridge University Press, 2000. DOI: 10.1017/CBO9780511801389

[3] ALMASI O N, ROUHANI M. A geometric-based data reduction approach for large low dimensional datasets: Delaunay triangulation in SVM algorithms[J]. Machine Learning with Applications, 2021, 4: 100025. DOI: 10.1016/j.mlwa.2021.100025

[4] BENNETT K P, BREDENSTEINER E J. Duality and geometry in SVM classifiers[C]//Proceedings of the Seventeenth International Conference on Machine Learning. New York: ACM, 2000: 57

[5] ALI ALHUSSAN A, TALAAT F M, EL-KENAWY E M, et al. Facial expression recognition model depending on optimized support

vector machine[J]. Computers, Materials & Continua, 2023, 76(1): 499. DOI: 10.32604/cmc.2023.039368

[6] Rangayya, Virupakshappa, PATIL N. Improved face recognition method using SVM-MRF with KTBD based KCM segmentation approach[J]. International Journal of System Assurance Engineering and Management, 2024, 15(1): 1. DOI:10.1007/s13198-021-01483-3

[7] HOURIA L, BELKHAMSA N, CHERFA A, et al. Multimodal magnetic resonance imaging for Alzheimer's disease diagnosis using hybrid features extraction and ensemble support vector machines[J]. International Journal of Imaging Systems and Technology, 2023, 33(2): 610. DOI: 10.1002/ima.22824

[8] ARUKONDA S, CHERUKU R. A novel stacking framework with PSO optimized SVM for effective disease classification[J]. Journal of Intelligent & Fuzzy Systems, 2023, 45(3): 4105. DOI: 10.3233/jifs-232268

[9] HAMAD R M, ABUSHAALA A M. Medical named entity recognition in Arabic text using SVM[C]//2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA). Benghazi: IEEE, 2023: 200. DOI: 10.1109/MI-STA57575.2023.10169454

[10] HAO Shule, ZHANG Peng, LIU Sen, et al. Sentiment recognition and analysis method of official document text based on BERT-SVM model[J]. Neural Computing and Applications, 2023, 35(35): 24621. DOI: 10.1007/s00521-023-08226-4

[11] SARITA K, KUMAR S, SAKET R K. OC fault diagnosis of multilevel inverter using SVM technique and detection algorithm [J]. Computers & Electrical Engineering, 2021, 96: 107481. DOI: 10.1016/j.compeleceng.2021.107481

[12] HAN Tian, ZHANG Longwen, YIN Zhongjun, et al. Rolling bearing fault diagnosis with combined convolutional neural networks and support vector machine [J]. Measurement, 2021, 177: 109022. DOI: 10.1016/j.measurement.2021.109022

[13] WANG Qiyu. Support vector machine algorithm in machine learning [C]//2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). Dalian: IEEE, 2022: 750. DOI: 10.1109/ICAICA54878.2022.9844516

[14] ELEN A, BAŞ S, KÖZKURT C. An adaptive Gaussian kernel for support vector machine [J]. Arabian Journal for Science and Engineering, 2022, 47(8): 10579. DOI: 10.1007/s13369-022-06654-3

[15] BIRZHANDI P, KIM K T, YOUN H Y. Reduction of training data for support vector machine: a survey[J]. Soft Computing, 2022, 26(8): 3729. DOI: 10.1007/s00500-022-06787-5

[16] BIRZHANDI P, KIM K T, LEE B, et al. Reduction of training data using parallel hyperplane for support vector machine [J]. Applied Artificial Intelligence, 2019, 33(6): 497. DOI: 10.1080/08839514.2019.1583449

[17] BARROS DE ALMEIDA M, DE PADUA BRAGA A, BRAGA J P. SVM-KM: speeding SVMs learning with a priori cluster selection and k-means[C]//Proceedings of Vol.1. Sixth Brazilian Symposium on Neural Networks. Rio de Janeiro: IEEE, 2002: 162. DOI: 10.1109/SBRN.2000.889732

[18] SHEN Xiangjun, MU Lei, LI Zhen, et al. Large-scale support vector machine classification with redundant data reduction [J]. Neurocomputing, 2016, 172: 189. DOI: 10.1016/j.neucom.2014.10.102

[19] 周玉, 钱旭, 王自强. 基于阴影集数据选择的可拓神经网络性能改进[J]. 北京工业大学学报, 2013, 39(3): 430
ZHOU Yu, QIAN Xu, WANG Ziqiang. Performance improvement of extension neural network using data selection method based on shadowed sets[J]. Journal of Beijing University of Technology, 2013, 39(3): 430

[20] 苏小红, 赵玲玲, 谢琳, 等. 阴影集的模糊支持向量机样本选择方法[J]. 哈尔滨工业大学学报, 2012, 44(9): 78
SU Xiaohong, ZHAO Lingling, XIE Lin, et al. Shadowed sets-based sample selection method for fuzzy support vector machine[J]. Journal of Harbin Institute of Technology, 2012, 44(9):78. DOI: 10.11918/j.issn.0367-6234.2012.09.014

[21] 张代俐, 汪廷华, 朱兴淋. 基于模糊隶属度函数的 SVM 样本约简算法[J]. 山西大学学报(自然科学版), 2024, 47(1): 18
ZHANG Daili, WANG Tinghua, ZHU Xinglin. SVM sample reduction algorithm based on fuzzy membership functions [J]. Journal of Shanxi University (Natural Science Edition), 2024, 47(1): 18. DOI: 10.13451/j.sxu.ns.2023138

[22] CHAU A L, LI Xiaoou, YU Wen. Large data sets classification using convex-concave hull and support vector machine [J]. Soft Computing, 2013, 17(5): 793. DOI: 10.1007/s00500-012-0954-x

[23] XU Hailong, LI Longyue, GUO Pengsong, et al. Uncertainty SVM active learning algorithm based on convex hull and sample distance [C]//2021 33rd Chinese Control and Decision Conference (CCDC). Kunming: IEEE, 2021: 6815. DOI: 10.1109/ccdc52312.2021.9602182

[24] ZHANG Li, YE Ning, ZHOU Weida, et al. Support vectors pre-extracting for support vector machine based on K nearest neighbour method [C]//2008 International Conference on Information and Automation. Changsha: IEEE, 2008: 1353. DOI: 10.1109/ICINFA.2008.4608212

[25] 李福祥, 王雪, 张驰, 等. 基于边界点的支持向量机分类算法[J]. 陕西理工大学学报(自然科学版), 2022, 38(3): 30
LI Fuxiang, WANG Xue, ZHANG Chi, et al. et al. Support vector machine classification algorithm based on boundary points [J]. Journal of Shaanxi University of Technology (Natural Science Edition), 2022, 38(3): 30

[26] BIRZHANDI P, YOUN H Y. CBCH (clustering-based convex hull) for reducing training time of support vector machine[J]. The Journal of Supercomputing, 2019, 75(8): 5261. DOI: 10.1007/s11227-019-02795-9

[27] 周志华. 机器学习[M]. 北京:清华大学出版社,2016
ZHOU Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2016

[28] GILLALA R, REDDY V K, TYAGI A K. KDOS-Kernel density based over sampling-A solution to skewed class distribution[J]. Journal of Information Assurance and Security (JIAS), 2020, 15(2): 44

[29] LIN Mingwei, XU Wenshu, LIN Zhanpeng, et al. Determine OWA operator weights using kernel density estimation [J]. Economic Research-Ekonomska Istraživanja, 2020, 33(1): 1441. DOI: 10.1080/1331677X.2020.1748509

[30] 周玉,夏浩,刘虹瑜,等. 基于 K 互近邻与核密度估计的 DPC 算法[J]. 北京航空航天大学学报,2025, 51(6): 1978
ZHOU Yu, XIA Hao, LIU Hongyu, et al. DPC algorithm based on K-reciprocal neighbors and kernel density estimation[J]. Journal of Beijing University of Aeronautics and Astronaucics, 2025, 51(6): 1978. DOI:10.13700/j.bh.1001-5965.2023.0342

[31] PEDRYCZ W. From fuzzy sets to shadowed sets: interpretation and computing[J]. International Journal of Intelligent Systems, 2009, 24(1): 48. DOI: 10.1002/int.20323

[32] ASLANI M, SEIPEL S. Efficient and decision boundary aware instance selection for support vector machines [J]. Information Sciences, 2021, 577: 579. DOI: 10.1016/j.ins.2021.07.015

[33] Markelle Kelly, Rachel Longjohn, Kolby Nottingham. The UCI Machine Learning Repository [EB/OL]. 1987. https://archive.ics.uci.edu

(编辑 张 红)