引用格式: 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):30-38.

基于边界点的支持向量机分类算法

李福祥*, 王 雪, 张 驰, 周 明
(哈尔滨理工大学 理学院, 黑龙江 哈尔滨 150080)

摘 要: 为了解决标准支持向量机在空间和时间上过度消耗的问题,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式距离,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。该方法只用了少量的边界点对支持向量机进行训练,同时排除了噪声点和孤立点及混杂在异类中的点对决策超平面的影响,提高了分类器的泛化能力。实验结果表明,与传统支持向量机、最近邻支持向量机、K-近邻支持向量机相比,在分类精度相当时,该方法有效地减少了训练样本的数量。

关键词: 支持向量机; 同类近邻; 异类近邻; 边界点; 分类
中图分类号: TP181 **文献标识码:** A **文章编号:** 2096-3998(2022)03-0030-09

支持向量机[1](Support Vector Machine,SVM)于 20 世纪 90 年代末由 Vapnik 等人提出,由于其坚实的数学理论并且具有良好的泛化能力,一经提出就得到了广泛的研究和应用。支持向量机是基于统计学理论的一种新的机器学习方法,它成功地解决了机器学习一直存在的高维度和局部极值问题,目前已经在很多领域取得了巨大的成功[2-4]。尽管传统的支持向量机占据诸多优势,但是仍然存在许多问题。支持向量机的基本动机是找到一个决策超平面,使两类数据之间的间隔最大,根据其间隔最大化,构造目标函数,再转化成其对偶问题进行求解。支持向量机的原问题是一个凸的二次规划问题,在求解其对偶问题时,其运算量取决于样本的规模大小与维度的高低。在实际问题中,如图像处理[5]、雷达一维距离像处理[6]、数据挖掘等领域[7],通常样本集的规模都非常庞大,并且具有很高的维度,如此大规模的数据,占用了大量的内存空间导致支持向量机的训练时间过长。因此,为了减少其运算量,有必要对样本集进行约简。考虑到实际的支持向量机在分类时只用了少部分的支持向量,在求解其对偶问题时,最优分类超平面只与拉格朗日乘子不为零的项有关,如果能在求解二次规划前把包含支持向量的边界点提取出来,并用这些边界样本点来替代所有的训练样本进行支持向量机训练,这无疑会大量减少运算成本,从而可解决支持向量机训练时间过长和内存消耗过大等方面的问题[8]。传统的支持向量机对噪声点和孤立点异常敏感,并且在优化过程中不仅对支持向量进行了优化,同时也对非支持向量进行了优化,这无疑增大了运算成本。近年来,研究者提出大量方法以减少孤立点、噪声点对传统支持向量机的影响,同时针对处理大规模数据集时怎样有效地筛选出支持向量也进行了诸多研究[9-11],大体上有两方面的改进。其一是在大规模数据集上,根据支持向量分布的几何意义,把支持向量提取出来,从而达到

收稿日期: 2021-09-02 修回日期: 2021-11-18
基金项目: 国家自然科学基金项目(11871181);黑龙江省自然科学基金项目(A2018008)
* 通信作者: 李福祥(1972—),男,黑龙江哈尔滨人,博士,教授,主要研究方向为机器学习、非线性数值分析和计算数学。

了缩减样本集规模的目的[12]。其二是根据在分类时每个样本对分类决策超平面起作用的程度不同,给每个样本赋予权重,对远离整体样本集的噪声点和孤立点赋予更小的权重,从而减少了孤立点和噪声点对决策超平面的影响,形成了模糊支持向量[13-14]。此类方法虽然降低了孤立点和噪声点对决策超平面的影响,但是孤立点和噪声点以及大量的非边界点仍然参加了训练,并没有降低运算成本,且分类精度的提高是以计算每个样本的权重为代价的。文献[4]和文献[5]提出了两种针对训练集不同的删减策略,但是它们都需要对训练样本集进行多次的支持向量机模型训练,较为复杂。文献[6]根据每个样本的最近邻是否为同类别来对样本集进行删减,但是这种删减策略有时可能会误删分类时所用到的支持向量,同时也有可能误保留噪声点。文献[7]利用 K-近邻方法来对样本集进行删减,计算每个样本的 K 个最近邻,利用这 K 个最近邻中同类样本的数量是否低于给定的阈值来对样本集进行删减,尽管与文献[6]相比,可以更有效合理的对样本进行删减,但是仍然保留了大量的非边界点。

针对上述支持向量机的缺陷,本文提出了一种改进的支持向量机分类算法——基于边界点的支持向量机分类方法(Support Vector Machine Classification Algorithm Based on Boundary Points,BP-SVM)。

1　支持向量机简介

1.1　线性可分支持向量机

给定训练样本集 $F=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,$y_i\in\{-1,1\}$,支持向量机的核心思想是基于训练集 F,在其空间中找到一个最优的分类超平面,使两类样本之间的间隔最大,其超平面方程可描述为

$$\omega^{\mathrm{T}}x+b=0。$$

若超平面(ω,b)能将训练样本正确分类,即对于$(x_i,y_i)\in D$,若 $y_i=1$,则有 $\omega^{\mathrm{T}}x_i+b>0$;若 $y_i=-1$,则有 $\omega^{\mathrm{T}}x_i+b<0$。令

$$\begin{cases}\omega^{\mathrm{T}}x_i+b\geqslant 1, & y_i=1,\\ \omega^{\mathrm{T}}x_i+b\leqslant -1, & y_i=-1,\end{cases} \tag{1}$$

构造支持向量机优化模型

$$\begin{aligned}&\min_{\omega,b}\frac{1}{2}\|\omega\|^2,\\ &\text{s.t.}\quad y_i(\omega^{\mathrm{T}}x_i+b)\geqslant 1,\quad i=1,2,\cdots,n,\end{aligned} \tag{2}$$

这就是支持向量机的基本模型。

注意到式(2)本身是一个凸的二次规划问题,利用拉格朗日乘子法可以转化成其对偶问题进行求解,该凸二次规划问题的拉格朗日函数为

$$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2+\sum_{i=1}^{n}\alpha_i(1-y_i(\omega^{\mathrm{T}}x_i+b)),$$

其中 $\alpha_i\geqslant 0$ 是拉格朗日乘子。

得到式(2)的对偶问题

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j,\\ &\text{s.t.}\quad \sum_{i=1}^{n}\alpha_i y_i=0,\\ &\qquad \alpha_i\geqslant 0,\quad i=1,2,\cdots,n,\end{aligned}$$

决策函数为

$$f(x)=\omega^{\mathrm{T}}x+b=\sum_{i=1}^{n}\alpha_i y_i x_i^{\mathrm{T}}x+b。$$

1.2　线性不可分支持向量机

有时并不是所有的样本都能同时满足式(1)中的约束条件,在这种情况下,向式(2)中引入一个松

弛变量 ξ_i,则目标函数式(2)变为

$$\begin{aligned}&\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i,\\&\text{s.t.}\quad y_i(\omega^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i,\quad i=1,2,\cdots,n,\end{aligned}\tag{3}$$

式中 $\xi=(\xi_1,\xi_2,\cdots,\xi_n)^{\mathrm{T}}$,$\xi_i$ 代表了每个样本的错分程度,所有松弛因子之和 $\sum_{i=1}^{n}\xi_i$ 代表了所有训练样本的错分程度,$\sum_{i=1}^{n}\xi_i$ 随着错分样本的增多而增大,反之亦然。为了控制样本的错分程度,通常会在 $\sum_{i=1}^{n}\xi_i$ 前面加一个错误的惩罚项,也就是惩罚因子 C,当 C 越大时,说明此分类器对错误的惩罚越大,相反当 C 越小时,此分类器对错分样本的惩罚越小。

将式(3)引入拉格朗日乘子 α_i 和 β_i,构造拉格朗日函数得

$$L(\omega,b,\xi,\alpha,\beta)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(\omega x_i+b)-1+\xi_i]-\sum_{i=1}^{n}\beta_i\xi_i,$$

其中 $\alpha_i\geqslant 0$、$\beta_i\geqslant 0$ 是拉格朗日乘子。

得到式(3)的对偶问题为

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j,\\&\text{s.t.}\quad\sum_{i=1}^{n}\alpha_i y_i=0,\\&\qquad 0\leqslant\alpha_i\leqslant C,\quad i=1,2,\cdots,n,\end{aligned}$$

决策函数为

$$f(x)=\omega^{\mathrm{T}}x+b=\sum_{i=1}^{n}\alpha_i y_i x_i^{\mathrm{T}}x+b。$$

1.3 非线性支持向量机

对于非线性问题,首先使用一个变换 $z=\phi(x)$ 将 x 映射到新的特征空间 z,得到支持向量机的对偶问题

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j),\\&\text{s.t.}\quad\sum_{i=1}^{n}\alpha_i y_i=0,\\&\qquad 0\leqslant\alpha_i\leqslant C,\quad i=1,2,\cdots,n,\end{aligned}$$

决策函数为

$$f(x)=\omega^{\mathrm{T}}x+b=\sum_{i=1}^{n}\alpha_i y_i k(x_i,x)+b,$$

其中 $k(\cdot,\cdot)$ 是核函数。常见的核函数有:

线性核函数:$k(x_i,x_j)=x_i^{\mathrm{T}}x_j$;

多项式核函数(POLY):$k(x_i,x_j)=(x_i^{\mathrm{T}}x_j)^d$,$d\geqslant 1$ 为多项式的次数;

径向基核函数(RBF):$k(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)$,$\sigma>0$ 为径向基核函数的带宽。

总体来说支持向量机的核函数主要分为全局核函数和局部核函数。全局核函数的支持向量机外推能力强而学习能力弱,局部核函数的支持向量机学习能力强但泛化能力较弱。在以上列举的核函数中线性核函数和多项式核函数属于全局核函数,径向基核函数属于局部核函数。

2 边界点下的支持向量机分类算法

2.1 K-近邻算法

K-近邻(K-Nearest Neighbor,KNN)算法是机器学习中最基本的一个算法,其思想相对比较容易理

解。K-近邻算法大体思想就是在特征空间中按照某种距离的计算方法,根据距离公式找到每个样本距离最近的 K 个样本,在每个样本的 K 个最近邻中,大多数样本都属于哪个类别,就把这个样本也归到其相应的类别,这就是 K-近邻算法的基本思想。

2.2　提取边界样本点

2.2.1　线性情况下提取边界样本点

设 $F=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,其中 $x_i\in\mathbf{R}^n$,$y_i\in\{-1,1\}$,$i=1,2,\cdots,n$。正类样本 x_i $(i=1,2,\cdots,n^+)$,负类样本 $x_j(j=n^++1,n^++2,\cdots,n^++n^-)$,其中 $n=n^++n^-$。给定正类样本 x_t,提取边界样本的步骤如下:

根据两样本的欧氏距离

$$d(x_t,x_i)=\|x_t-x_i\|=\sqrt{\|x_t-x_i\|^2}=\sqrt{x_t^{\mathrm{T}}x_t+x_i^{\mathrm{T}}x_i-2x_t^{\mathrm{T}}x_i},$$

计算出 x_i 的 K 个同类最近邻 $T_k(x_t)$ 和 x_t 的 K 个异类最近邻 $\overline{T}_k(x_t)$,然后计算 x_t 与 $T_k(x_t)$ 之间的距离以及 x_t 与 $\overline{T}_k(x_t)$ 之间的距离。其中 $T_k(x_t)$ 是与 x_t 同类的距离最近的 K 个样本构成的集合,$\overline{T}_k(x_t)$ 是与 x_t 异类的距离最近的 K 个样本构成的集合。令

$$d(x_t,T(x_t))=\frac{1}{K}\sum_{x_k\in T_k(x_t)}d(x_t,x_k),\qquad \overline{d}(x_t,\overline{T}(x_t))=\frac{1}{K}\sum_{x_k\in\overline{T}_k(x_t)}\overline{d}(x_t,x_k),$$

根据 $d(x_t,T(x_t))$ 和 $\overline{d}(x_t,\overline{T}(x_t))$ 来判断 x_t 是否为边界样本。

定义 1　如果 $d(x_t,T_k(x_t))<\theta_1$,$\overline{d}(x_t,T_k(x_t))<\theta_2$,称 x_t 为边界样本点,否则称 x_t 为非边界样本点,其中 $\theta_1>0$、$\theta_2>0$ 为事先给定的阈值。

通常情况下,如果一个样本点 x_t 到同类样本的 K 个最近邻的距离和到异类样本的 K 个最近邻的距离都相对较小,那么此样本点更可能成为边界点。如图 1 所示,样本点 A 的同类 3-近邻和异类 3-近邻均比较小,所以点 A 为边界样本点,样本点 B 的同类 3-近邻相对较小而异类 3-近邻相对较大,所以点 B 为非边界样本点。

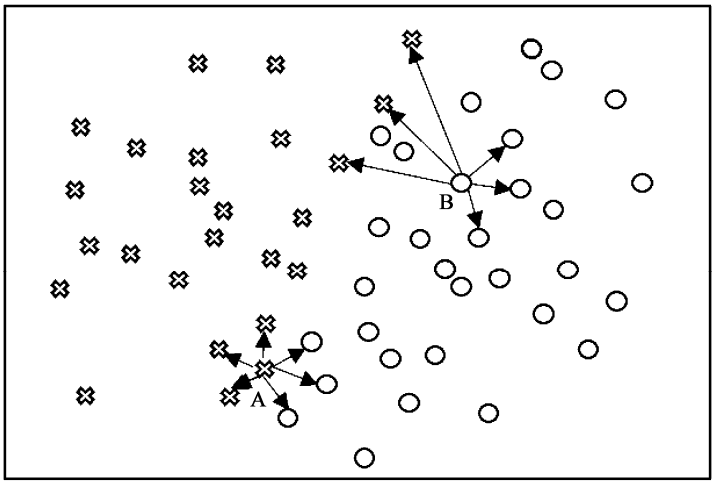


图 1　边界点与非边界点示意图

2.2.2　非线性条件下提取边界样本点

在大多数分类任务中,样本所在的原始空间并不能找到一个能将两类样本正确划分的超平面,通常情况下,通过映射 $\phi(x)$ 将样本映射到一个高维特征空间,使样本在这个高维的特征空间内变得线性可分。令 $\phi(x)$ 表示将 x 映射后的特征向量,两样本在特征空间中的距离为

$$\begin{aligned}D(x_t,x_i)&=\|\phi(x_t)-\phi(x_i)\|=\sqrt{\|\phi(x_t)-\phi(x_i)\|^2}=\\&\sqrt{\phi(x_t)\cdot\phi(x_t)+\phi(x_i)\cdot\phi(x_i)-2\phi(x_t)\cdot\phi(x_i)}=\\&\sqrt{k(x_t,x_t)+k(x_i,x_i)-2k(x_t,x_i)},\end{aligned}\tag{4}$$

其中 $\phi(x_t)\cdot\phi(x_i)$ 是样本 x_t、x_i 映射到特征空间之后的内积,$k(x_t,x_i)=\phi(x_t)\cdot\phi(x_i)$ 是核函数。

计算出 x_t 的 K 个同类最近邻 $T_k(x_t)$ 和 x_t 的 K 个异类最近邻 $\overline{T}_k(x_t)$,然后计算 x_t 与 $T_k(x_t)$ 之间的距离以及 x_t 与 $\overline{T}_k(x_t)$ 之间的距离。其中 $T_k(x_t)$ 是与 x_t 同类的距离最近的 K 个样本构成的集合,$\overline{T}_k(x_t)$ 是与 x_t 异类的距离最近的 K 个样本构成的集合。令

$$D(x_t,T(x_t)) = \frac{1}{K}\sum_{x_k\in T_k(x_t)} D(x_t,x_k),$$

$$\overline{D}(x_t,\overline{T}(x_t)) = \frac{1}{K}\sum_{x_k\in \overline{T}_k(x_t)} \overline{D}(x_t,x_k),$$

根据 $D(x_t,T(x_t))$ 和 $\overline{D}(x_t,\overline{T}(x_t))$ 来判断 x_t 是否为边界样本。

定义 2 如果 $D(x_t,T_k(x_t)) < \theta_1$,$\overline{D}(x_t,T_k(x_t)) < \theta_2$,称 x_t 为边界样本点,否则称 x_t 为非边界样本点,其中 $\theta_1>0$,$\theta_2>0$ 为事先给定的阈值。

无论是线性情况下还是非线性情况下,阈值 θ_1、θ_2 的选取都是依据空间中各个维度的最大值、最小值和平均值来根据经验确定的。如果阈值 θ_1、θ_2 的值选择过大,会导致边界点过多,不能有效地去除非边界点,如果阈值 θ_1、θ_2 的值选择过小,会过滤掉包含支持向量的边界点,造成模型效果不佳。

记 D_1 为所有非边界点构成的集合。

对于非线性情况,在提取边界点时,需要通过核函数将样本映射到特征空间计算其距离来寻找其 K 个最近邻,但是核函数的参数并未能事先取得,为此,给出定理。

定理 当支持向量机核函数为径向基核函数或者指数核函数时,可更简单的采用输入空间的距离公式

$$d(x_i,x_j) = \|x_i - x_j\| = \sqrt{(x_i - x_j)^{\mathrm{T}}(x_i - x_j)}, \tag{5}$$

式(5)是样本在原输入空间下的欧氏距离,当核函数为径向基核函数 $k(x_i,x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$ 或指数核函数 $k(x_i,x_j) = \exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right)$ 时,在原始空间利用公式(5)与特征空间利用公式(4)求得的 K 个最近邻是完全相同的[17]。

证明 对于径向基核函数和指数核函数都有 $k(x_i,x_i) = k(x_j,x_j) = 1$,所以

$$D(x_i,x_j) = \left(2 - 2\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)\right)^{\frac{1}{2}} = \left(2 - 2\exp\left(-\frac{d(x_i,x_j)^2}{2\sigma^2}\right)\right)^{\frac{1}{2}},$$

或者有

$$D(x_i,x_j) = \left(2 - 2\exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right)\right)^{\frac{1}{2}} = \left(2 - 2\exp\left(-\frac{d(x_i,x_j)}{2\sigma^2}\right)\right)^{\frac{1}{2}},$$

而 $d(x_i,x_j)\geqslant 0$,由指数函数和幂函数的性质可知,对于输入空间的两个固定样本 x_i、x_j,$D(x_i,x_j)$ 是关于 $d(x_i,x_j)$ 单调递增的,这表明样本在原始输入空间和特征空间中,样本的相对远近程度并没有发生改变,仅仅只有紧密度发生了变化,从而当核函数为径向基核函数或指数核函数时,在原始输入空间和特征空间所求的 K 个最近邻完全相同。证毕。

2.3 K-近邻删除样本点

传统支持向量机对于一些交错严重的样本集,因为在寻找最优超平面时需要照顾到每一个样本点,会导致分类过程较为复杂,使分类过程变得复杂的原因正是因为混杂在异类中的样本点在分类时,并不能对决策超平面起到积极的作用,相反会对支持向量机在寻找最优决策超平面时造成干扰,从而影响其泛化能力。针对传统支持向量机的这种弊端,提出了一种利用 K-近邻的方法对样本集进行删减,计算出每个样本的 K 个最近邻。其中同类的样本数量小于给定阈值 δ(一般取 $0<\delta\leqslant\frac{K}{2}$)的样本构成的集合记为 D_2,在 2.2 中提取边界点时已经得到了非边界点集合 D_1,将两次删减的样本集合 D_1、D_2 分别从总样本集 F 中删去,将最终保留下来的样本作为新的训练样本集,应用传统支持向量机进行训练得到最终的分类器。

2.4　边界点下的支持向量机的步骤

步骤一：根据公式(5)计算两两样本之间的欧式距离；

步骤二：计算每个样本点的同类 K-近邻 $T_k(x_t)$ 和异类 K-近邻 $\overline{T}_k(x_t)$；

步骤三：根据 2.2 提取出边界样本点，删除非边界点；

步骤四：利用 K-近邻的方法对样本集进行删减，计算出样本 x_t 的 K 个最近邻，判断其同类的样本数量是否小于给定阈值 δ（一般取 $0<\delta\leqslant\frac{K}{2}$），如果 x_t 的 K 个近邻集中，同类的样本数量小于给定阈值 δ，则将 x_t 从样本集中删去，否则保留；

步骤五：对删减过后的样本集利用传统的支持向量机进行训练。

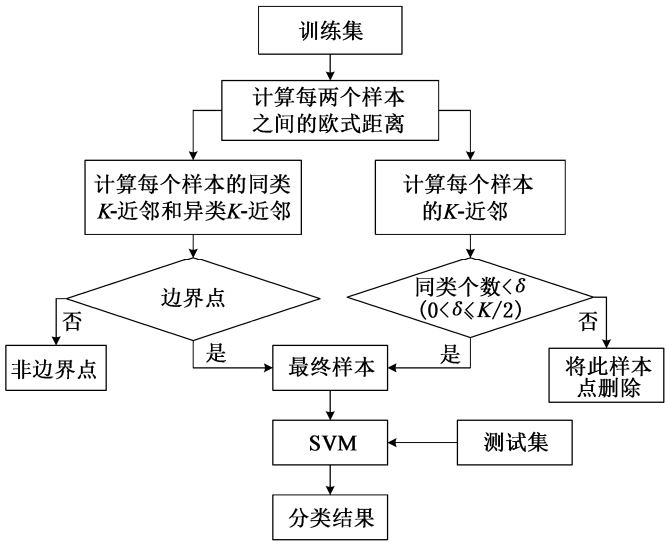


图 2　边界点下的支持向量机步骤流程图

边界点下的支持向量机步骤流程如图 2 所示。

2.5　算法的复杂度分析

假设训练数据集共有 N 个样本，每个类别中的样本数为 $\frac{N}{2}$，对于第 $c_i(i=1,2)$ 类中的每一个样本，从另一类中选取 K 个近邻的复杂度为 $\frac{N}{2}\log\frac{N}{2}$，而传统支持向量机的算法复杂度为 $O(NL)^2$，其中 N 为样本数量，L 为特征维数。综上所述，BP-SVM 算法的复杂度远远小于传统支持向量机算法的复杂度，从而提高了时间效率。

3　实验结果及分析

3.1　实验设计

为了验证 BP-SVM 算法的有效性，对算法进行测试，实验环境如下：

1) 硬件环境：1.80 GHz CPU，内存 8 GB，硬盘 100 GB 以上的个人计算机；

2) 软件环境：Windows10 Professional 操作系统，用 Python3.8 编程实现。

线性条件下的实验数据集采用随机产生的 400 个服从高斯分布的线性不可分数据，其中正样本的个数为 200，负样本的个数为 200，正样本集服从均值为 1、方差为 2 的高斯分布，负样本集服从均值为 4，方差为 1 的高斯分布。支持向量机采用线性核函数。非线性条件下的实验数据集采用 Banana 数据集，支持向量机采用径向基核函数。除此之外，本文从 UCI 机器学习库中共选取 6 个数据集，分别是 Banana、Breast、Thyroid、Heart、Titanic、Diabetis，除 Diabetis 数据集采用了线性核函数，其他数据集均采用径向基核函数。

在分类任务中，通常用正确率和错误率来衡量分类器性能的好坏，如果样本总数为 n，其中有 m 个样本被误分，则错误率可以表示为

$$error=\frac{m}{n},$$

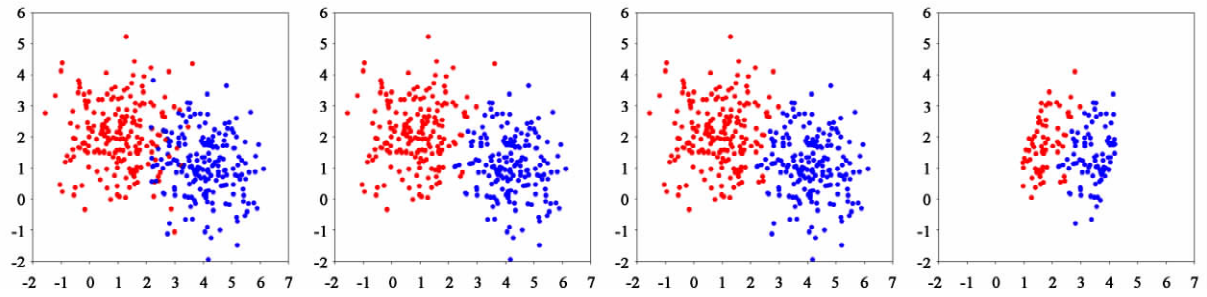
分类正确率可表示为

$$accuary=1-error=\frac{n-m}{n}。$$

本文用正确率来衡量分类器的性能。

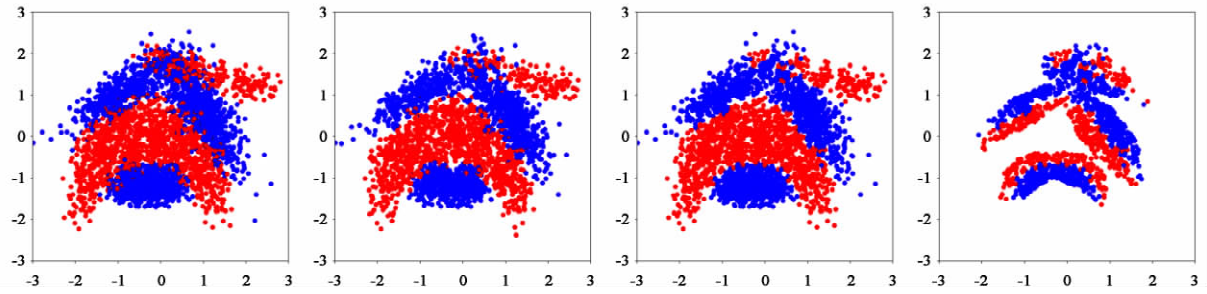
3.2　各类算法的训练点数量对比图

图 3 和图 4 分别是在线性条件下和非线性条件下,传统支持向量机(SVM)、最近邻支持向量机(NN-SVM)、*K*-近邻支持向量机(KNN-SVM)和 BP-SVM 算法训练点的数量对比图。从图中可以看出,最近邻支持向量机有效地删除了混杂在异类中的点,但是同时也误删了支持向量机在分类时所用的边界点。*K*-近邻支持向量机删除了混杂在异类中的点,同时克服了最近邻支持向量机的缺陷,保留了支持向量机在分类时所用的边界点,但是仍有许多孤立点和大量的非边界点存在。本文算法既有效地保留了支持向量分类时起作用的边界点,同时也删除了混杂在异类中的点和一些远离整体样本的孤立点、噪声点,大大缩减了训练样本的数量。



(a) SVM　(b) NN-SVM　(c) KNN-SVM　(d) BP-SVM

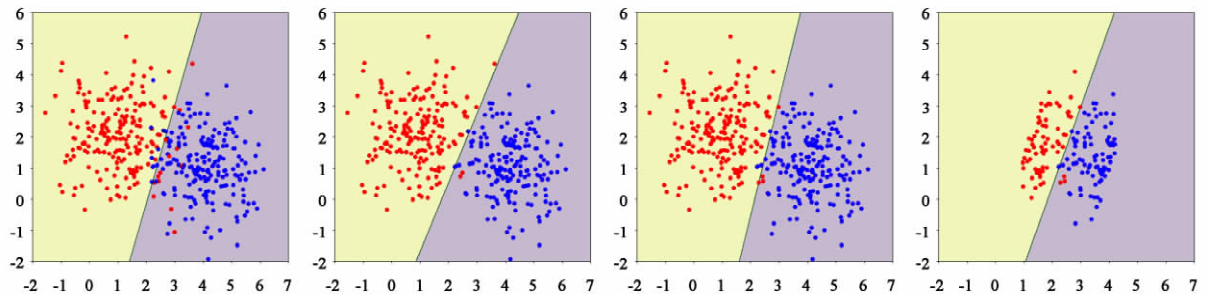
图 3　线性条件下训练点的数量对比图



(a) SVM　(b) NN-SVM　(c) KNN-SVM　(d) BP-SVM

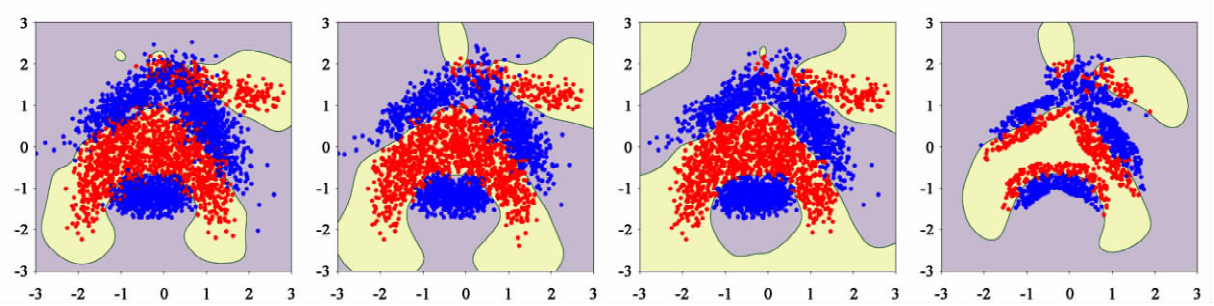
图 4　非线性条件下训练点的数量对比图

图 5 和图 6 分别是在线性条件下和非线性条件下,传统支持向量机、最近邻支持向量机、*K*-近邻支持向量机和 BP-SVM 算法分类效果对比图。从图中可以看出,线性条件下,BP-SVM 算法只用了少量的边界点对支持向量机进行训练,同时排除了噪声点和孤立点对决策超平面的影响,使分类器有更好的泛化能力;非线性条件下,与传统支持向量机、最近邻支持向量机和 *K*-近邻支持向量机相比,BP-SVM 算法删去了大量的非边界点、和混杂在异类中的样本点,只利用了少量的边界点对支持向量机进行训练,从而大大缩减了在优化时所带来的计算损失。



(a) SVM　(b) NN-SVM　(c) KNN-SVM　(d) BP-SVM

图 5　线性情况下分类效果对比图



(a) SVM　　(b) NN-SVM　　(c) KNN-SVM　　(d) BP-SVM

图 6　非线性情况下分类效果对比图

3.3　实验结果

本文从 UCI 机器学习库中共选取 6 个数据集，分别是 Banana、Breast、Thyroid、Heart、Titanic、Diabetis，数据集描述见表 1。各种算法的训练点数量和分类正确率见表 2 和表 3。从表 2 和表 3 可以看出，与传统支持向量机、最近邻支持向量机和 K-近邻支持向量机相比，BP-SVM 算法大幅度裁减了冗余的、无用的训练样本的数量，只利用了少量的边界点进行训练。从表 3 可以看出，本文算法在 Breast、Thyroid 数据集中，正确率等于或者大于传统支持向量机，其原因是本文算法在提取边界点时，删除了非边界点，这些非边界点包括噪声点和孤立点，从而提高了正确率。

表 1　数据集描述

数据集	实例数	特征数	类数
Banana	4000	2	2
Breast	2000	9	2
Thyroid	2800	5	2
Heart	1500	13	2
Titanic	750	3	2
Diabetis	1000	8	2

表 2　各种算法的训练点数量

数据集	SVM	NN-SVM	KNN-SVM	BP-SVM
Banana	4000	3664	3661	1798
Breast	2000	2000	1953	920
Thyroid	2800	2742	2742	1948
Heart	1500	1500	1488	975
Titanic	750	750	744	568
Diabetis	1000	905	848	384

表 3　各种算法的分类正确率　%

数据集	SVM	NN-SVM	KNN-SVM	BP-SVM
Banana	91.0	90.1	91.1	90.9
Breast	76.7	76.7	76.7	76.7
Thyroid	95.7	97.2	97.3	96.1
Heart	73.2	73.2	73.5	71.6
Titanic	70.5	70.5	70.5	66.6
Diabetis	78.6	78.6	79.0	78.5

4　结论

本文针对传统支持向量机在训练时应用了所有的样本点进行训练，提出了一种利用边界点进行训练的支持向量机(BP-SVM)。与传统支持向量机、最近邻支持向量机、K-近邻支持向量机相比，只利用了少量的边界点训练支持向量机，同时排除了噪声点和孤立点对决策超平面的影响。在算法复杂度方面，BP-SVM 算法的复杂度明显低于传统支持向量机的复杂度。在正确率方面，在 Breast、Thyroid 数据集中，BP-SVM 算法的正确率大于或等于传统支持向量机。最后得出结论，与传统支持向量机、最近邻支持向量机、K-近邻支持向量机相比，在正确率相当的情况下，BP-SVM 方法有效地缩减了训练样本数目。

[参　考　文　献]

[1] VAPNIK V. The Natural of Statistical Learning Theory [J]. Technometrics, 1995, 38(4): 409.

[2] GAO Z, FANG S C, LUO J, et al. A Kernel-Free Double Well Potential Support Vector Machine with Applications [J]. European Journal of Operational Research, 2020, 290(1): 248-262.

[3] DING S, ZHANG N, ZHANG X, et al. Twin Support Vector Machine: Theory, Algorithm and Applications [J]. Neural Computing and Applications, 2017, 28(11): 3119-3130.

[4] LIU Y,DING H,HUANG Z,et al. Distributed and Robust Support Vector Machine [J]. International Journal of Computational Geometry and Applications,2021,30(3):213-233.
[5] TONG Y,SUN W. The Role of Film and Television Big Data in Real-Time Image Detection and Processing in the Internet of Things Era [J]. Journal of Real-Time Image Processing,2021,18(4):1115-1127.
[6] GUO X K,JIAN T,DONG Y L. Radar One-Dimensional Range Profile Recognition Based on Improved Wavelet Denoising [J]. Radar Science and Technology,2019,17(4):360-364.
[7] HAUBEN M,PATADIA V,GERRITS C,et al. Data Mining in Pharmacovigilance [J]. Drug Safety,2005,28(10):835-842.
[8] UTAMI N A,MAHARANI W,ATASTINA I. Personality Classification of Facebook Users According to Big Five Personality Using SVM (Support Vector Machine) Method [J]. Procedia Computer Science,2021,179(1):177-184.
[9] GUO S M,CHEN L C,TSAI J. A Boundary Method for Outlier Detection Based on Support Vector Domain Description [J]. Pattern Recognition,2009,42(1):77-83.
[10] ARUMUGAM P,JOSE P. Efficient Decision Tree Based Data Selection and Support Vector Machine Classification [J]. Materials Today: Proceedings,2018,5(1):1679-1685.
[11] FAYED H A,ATIYA A F. Speed Up Grid-Search for Parameter Selection of Support Vector Machines [J]. Applied Soft Computing,2019,80:202-210.
[12] ARUMUGAM P,JOSE P. Efficient Decision Tree Based Data Selection and Support Vector Machine Classification [J]. Materials Today: Proceedings,2018,5(1):1679-1685.
[13] SARIMVEIS H,ALEXANDRIDIS A,BA F G. A Fast Training Algorithm for RBF Networks Based on Subtractive Clustering [J]. Neurocomputing,2003,51(Apr):501-505.
[14] YAO J,DASH M,TAN S T,et al. Entropy-Based Fuzzy Clustering and Fuzzy Modeling [J]. Fuzzy Sets and Systems,2000,113(3):381-388.
[15] KE H,ZHANG X. Editing Support Vector Machines [C]. International Joint Conference on Neural Networks. IEEE,2001:1464-1467.
[16] 李红莲,王春花,袁保宗. 一种改进的支持向量机 NN-SVM [J]. 计算机学报,2003,26(8):1015-1020.
[17] 和文全,薛惠峰,解丹蕊,等. 基于 K 近邻的支持向量机分类方法 [J]. 计算机仿真,2008,25(11):161-163.

[责任编辑:李 莉]

Support vector machine classification algorithm based on boundary points

LI Fu-xiang, WANG Xue, ZHANG Chi, ZHOU Ming
(College of Science, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: In order to solve the excessive consumption of space and time in standard support vector machines, a new method of training support vector machines using boundary points is proposed. Firstly, the Euclidean distance between each two samples is calculated to find out the homogeneous nearest neighbor set and heterogeneous nearest neighbor set of each sample point. According to the distance from the sample point to the two sets, it is judged whether it may become a boundary point. Secondly, the sample set is deleted according to the number of similar samples in each sample nearest neighbor set. This method only uses a small number of boundary points to train the support vector machine, eliminates the influence of noise points, isolated points and points mixed in different classes on the decision hyperplane, and improves the generalization ability of the classifier. Experimental results show that compared with traditional support vector machines, nearest neighbor support vector machines and *K*-nearest neighbor support vector machines, the proposed method can greatly reduce the number of training samples when the classification accuracy is similar.

Key words: support vector machine; homogeneous nearest neighbor; heterogeneous nearest neighbor; boundary point; classification