面向大规模噪声数据的软性核凸包支持向量机

顾晓清[1,2],倪彤光[2],姜志彬[1],王士同[1]
(1. 江南大学数字媒体学院,江苏无锡 214122; 2. 常州大学信息科学与工程学院,江苏常州 213164)

摘 要: 现有的面向大规模数据分类的支持向量机(support vector machine,SVM)对噪声样本敏感,针对这一问题,通过定义软性核凸包和引入 pinball 损失函数,提出了一种新的软性核凸包支持向量机(soft kernel convex hull support vector machine for large scale noisy datasets,SCH-SVM). SCH-SVM 首先定义了软性核凸包的概念,然后选择出能代表样本在核空间几何轮廓的软性核凸包向量,再将其对应的原始空间样本作为训练样本并基于 pinball 损失函数来寻找两类软性核凸包之间的最大分位数距离. 相关理论和实验结果亦证明了所提分类器在训练时间,抗噪能力和支持向量数上的有效性.

关键词: 大规模数据; 噪声; 软性核凸包; pinball 损失函数; 分类
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2018)02-0347-11
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.02.013

Soft Kernel Convex Hull Support Vector Machine for Large Scale Noisy Datasets

GU Xiao-qing[1,2], NI Tong-guang[2], JIANG Zhi-bin[1], WANG Shi-tong[1]
(1. *School of Digital Media, Jiangnan University, Wuxi, Jiangsu* 214122, *China*;
2. *School of Information Science and Engineering, Changzhou University, Changzhou, Jiangsu* 213164, *China*)

Abstract: Current support vector machines (SVMs) for large-scale datasets classification problems are almost sensitive to noises. To overcome this problem, a new soft kernel convex hull support vector machine called SCH-SVM is proposed based on the soft kernel convex hull and pinball loss function. SCH-SVM extracts the soft convex hull vectors in the kernel space, which can represent geometric profile of data in the kernel space. Then SCH-SVM represents the original samples which projected as the soft convex hull vectors for the training samples, and finds the maximum quantile distance between soft kernel convex hulls belonging to two classes by using pinball loss function. Theoretical analysis and numerical experiments show that SCH-SVM has distinctive ability of training time, noise resistibility, and the number of support vectors.

Key words: large scale datasets; noise; soft kernel convex hull; pinball loss function; classification

1 引言

支持向量机(Support Vector Machine,SVM)使用统计学习理论和最优化方法解决机器学习问题,凭借其优秀的泛化性能成为模式识别领域一个非常重要的分支. 其最优化问题常描述成二次规划问题. 对于样本数为 N 的数据集来说,训练 SVM 分类器的时间复杂度为 $O(N^3)$. 为降低 SVM 的计算复杂度,常用的方法有:逼近替代二次规划问题中的核矩阵计算,如贪婪逼近[1],矩阵分解法[2]和矩阵变换法[3]等;减少训练样本数或限制支持向量数,如核心集向量机[4,5],快速核密度估计[6],近似极值点支持向量机[7]和采样 SVM[8]等.

目前几乎所有的面向大规模数据的 SVM 分类器都基于样本无噪声的前提. 然而,现实世界的真实数据普遍存在噪声或例外点[9]. 当前,抗噪型 SVM 可分为3类,一类是模糊支持向量机,如双权重模糊支持向量机[10]和最小类内散度模糊支持向量机[11];一类是使用噪声不敏感的损失函数,如 pinball 损失函数[12,13];另一类是基于数据相似信息的挖掘,如多任务单类支持向量机[14]. 但这些分类器因计算代价高不适用于大规模数据的分类. 为解决这一问题,本文提出了一种面向大规模噪声数据的软性核凸包支持向量机 (Soft Kernel Convex Hull Support Vector Machine for Large Scale Noisy

收稿日期:2016-10-24;修回日期:2017-10-17;责任编辑:马兰英
基金项目:国家自然科学基金(No. 61572236, No. 61572085);江苏省自然科学基金(No. BK20160187)

Datasets,SCH-SVM),SCH-SVM 首先定义了软性核凸包并选择出代表样本轮廓的软性核凸包向量作为训练样本,并借助 pinball 损失函数寻找垂直平分两类训练样本的最近分位数距离来确定最优分类超平面. 该分类器在保持样本核空间的几何轮廓的同时,忽略样本在核空间几何轮廓内部的噪声,同时对软性核凸包中的噪声不敏感,有效减少了训练时间和支持向量数. 文中定理证明了 SCH-SVM 使用软性核凸包向量作为训练样本对分类精度的影响是极其有限的.

2 pinball 损失函数

传统 SVM 对样本噪声尤其是分类面边界周围的噪声敏感. 受统计学中分位数应用的启发,文献[12,13]将 pinball 损失函数引入到 SVM 中,提出了最大化两类分位数距离的分类策略. 设有离散标量集合 $U=\{u_1,u_2,\cdots,u_m\}$,q 下分位数可以表示为:

$$\min^q\{U\}=\{t:t\in\boldsymbol{R},t \text{ is large than } q \text{ ratio of } u_i\} \tag{1}$$

使用 pinball 损失函数来寻求样本间的最大分位数距离对噪声样本尤其对分类面周围的噪声样本不敏感,可以有效提高分类器的抗噪能力. pinball 损失函数 $L_\tau(u)$[13] 定义为:

$$L_\tau(u)=\begin{cases}u, & \text{if } u\geqslant 0\\ -\tau u, & \text{otherwise}\end{cases} \tag{2}$$

其中,L_τ 表示为 $\tau/(1+\tau)$下分位数. pinball 损失函数参数 τ 与式(1)中的分位数 q 之间存在关系 $\tau=q/(1-q)$.

在二元分类问题中,给定数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 和其类别标签 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_N\}$. 假设序列集Ⅰ和Ⅱ分别表示 $\mathrm{Ⅰ}=\{i\mid y_i=1\}$ 和 $\mathrm{Ⅱ}=\{i\mid y_i=-1\}$. 此时分类器可表示为:

$$\max_{\|w\|=1,b}\{\min^q_{i\in \mathrm{Ⅰ}}y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)+\min^q_{i\in \mathrm{Ⅱ}}y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\} \tag{3}$$

文献[13]提出的 pin-SVM 的优化问题可表示为:

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{C}{N}\sum_{i=1}^{N}\xi_i,\\ &\text{s.t.}\quad y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,i=1,2,\cdots,N,\\ &\qquad y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\leqslant 1+\frac{1}{\tau}\xi_i,i=1,2,\cdots,N\end{aligned} \tag{4}$$

式(4)的求解可转换为二次规划问题,其时间复杂度为 $O(N^3)$. 因此 pin-SVM 只能处理小样本分类问题.

3 面向大规模噪声数据的软性核凸包支持向量机

3.1 SCH-SVM 的基本思想

SVM 的求解问题等价于寻找使得能够最大间隔地分离了两类样本的外围轮廓点的最优分类超平面[14]. 从几何角度看,凸包是一种能够准确描述空间物体轮廓的方法,其定义如下:

定义 1 **(凸包)**[15,16] d 维空间中样本集 $\boldsymbol{X}=\{\boldsymbol{x}_i\in\boldsymbol{R}^d,i=1,2,\cdots,N\}$ 的凸包 $\boldsymbol{X}^*$ 是包含所有样本的最小凸集. 样本集 $\boldsymbol{X}$ 内任何样本 $\boldsymbol{x}_i$ 都可用凸包向量的线性组合来表示,即:

$$\boldsymbol{x}_i=\sum_{\boldsymbol{x}_t\in\boldsymbol{X}^*}\mu_{i,t}\boldsymbol{x}_t \tag{5}$$

其中,$\sum_{\boldsymbol{x}_t\in\boldsymbol{X}^*}\mu_{i,t}=1$ 且 $\mu_{i,t}\geqslant 0$.

根据定义 1,下面给出核凸包和软性核凸包的定义.

定义 2 **(核凸包)** 核映射 ϕ 将样本 $\boldsymbol{x}$ 映射到核空间 $\boldsymbol{F}$,样本集 $\boldsymbol{X}$ 在核空间 F 中的核凸包为 $con(\boldsymbol{X})$,则样本集 $\boldsymbol{X}$ 内任何样本 $\boldsymbol{x}_i$ 在核空间的映射都可用核凸包向量的线性组合来表示,即:

$$\phi(\boldsymbol{x}_i)=\sum_{\phi(\boldsymbol{x}_t)\in con(\boldsymbol{X})}\mu_{i,t}\phi(\boldsymbol{x}_t) \tag{6}$$

其中,$\sum_{\phi(\boldsymbol{x}_t)\in con(\boldsymbol{X})}\mu_{i,t}=1$ 且 $\mu_{i,t}\geqslant 0$.

为了保证核凸包的可解性,在定义 2 的基础上引入误差阈值 ε,得到"软化"了的核凸包.

定义 3 **(软性核凸包)** 设原始样本集 $\boldsymbol{X}$ 在核空间的映像 $\boldsymbol{Z}=\{\phi(\boldsymbol{x}_i),\forall\boldsymbol{x}_i\in\boldsymbol{X}\}$,$\boldsymbol{X}$ 在核空间中的软性核凸包表示为 $\boldsymbol{Z}^*$,则 $\boldsymbol{X}$ 内所有样本在核空间的映射与软性核凸包向量的线性关系满足:

$$\max_{\boldsymbol{x}_i\in X}\min\left\|\phi(\boldsymbol{x}_i)-\sum_{\phi(\boldsymbol{x}_t)\in\boldsymbol{Z}^*}\mu_{i,t}\phi(\boldsymbol{x}_t)\right\|^2\leqslant\varepsilon \tag{7}$$

其中,$0\leqslant\mu_{i,t}\leqslant 1$ 且 $\sum_{\phi(\boldsymbol{x}_t)\in\boldsymbol{Z}^*}\mu_{i,t}=1$.

因此核空间映像 $\boldsymbol{Z}$ 中每一个元素 $\phi(\boldsymbol{x}_i)$ 与软性核凸包向量的线性组合可写成:

$$\phi(\boldsymbol{x}_i)=\sum_{\phi(\boldsymbol{x}_t)\in\boldsymbol{Z}^*}\gamma_{i,t}\phi(\boldsymbol{x}_t)+\boldsymbol{\delta}_i \tag{8}$$

其中,$\|\boldsymbol{\delta}_i\|^2\leqslant\varepsilon$ 且

$$\gamma_{i,t}=\begin{cases}\mu_{i,t}, & \text{if } \phi(\boldsymbol{x}_t)\in\boldsymbol{Z}^*,\phi(\boldsymbol{x}_i)\in\boldsymbol{Z} \text{ and } \phi(\boldsymbol{x}_i)\notin\boldsymbol{Z}^*\\ 0, & \text{otherwise}\end{cases}$$

对于噪声样本按几何分布可分成 2 种:第 1 种是非软性核凸包向量;第 2 种是软性核凸包向量,这些噪声样本会作为训练样本参与到分类器构建中. 因此,SCH-SVM 在使用不同类别样本的软性核凸包向量作为训练样本时,同时使用 pin-SVM 作为基础模型,最大化两类分位数距离得到抗噪的分类器模型. SCH-SVM 的无约束原始问题描述为:

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{C}{N}\sum_{t=1}^{M}l(\boldsymbol{w},b,\phi(\boldsymbol{x}_t)) \tag{9}$$

其中,$l(\boldsymbol{w},b,\phi(\boldsymbol{x}_t))$ 为 pinball 损失函数,$l(\boldsymbol{w},b,\phi(\boldsymbol{x}_t))=\max\{-\tau[1-y_t(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_t)+b)],1-y_t(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_t)+b)\}$,$\phi(\boldsymbol{x}_t)$为软性核凸包向量,$M$ 是软性核凸包向量的个数. 引入拉格朗日向量 $\boldsymbol{\alpha}$ 和 $\tilde{\boldsymbol{\alpha}}$,可得到 SCH-SVM 原

始问题的凸二次规划问题:

$$\min_{\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}}} \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j) - \sum_{i=1}^{M}\alpha_i$$

$$\text{s.t.}\quad \sum_{i=1}^{M}\alpha_i y_i = 0, \tag{10}$$

$$\alpha_i + (1+\frac{1}{\tau})\tilde{\alpha}_i = \frac{C}{N}, i=1,2,\cdots,M,$$

$$\alpha_i + \tilde{\alpha}_i \geqslant 0, \tilde{\alpha}_i \geqslant 0, i=1,2,\cdots,M$$

SCH-SVM 的分类决策函数为:

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b \tag{11}$$

3.2　SCH-SVM 分类器的构建

SCH-SVM 分类器的构建包括 3 个阶段:(1)样本在核空间的分组:将训练集划分成若干个样本组;(2)软性核凸包向量的选择:根据定义 3 选择每个分组中的软性核空间凸包向量;(3)分类器的训练:将构建软性核凸包的原始样本作为训练样本代入式(10)得到分类决策函数. 需要指出的是,第 1 和第 2 阶段在两类样本中分别进行. SCH-SVM 分类器的构建原理示意图如图 1 所示.

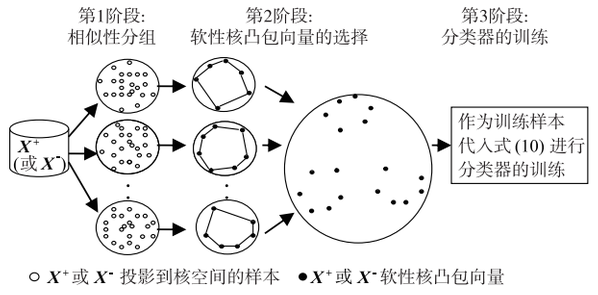


图1　SCH-SVM的构建原理示意图

3.2.1　样本在核空间的分组

从数据的空间划分角度看,属于同一个分组的样本间欧氏距离较小[11]. SCH-SVM 以此作为划分样本组的依据,将 $\boldsymbol{X}$ 在核空间的映像划分成 l 个样本组. 以正类数据集 $\boldsymbol{X}^+$ 为例,步骤如下:

(1)在 $\boldsymbol{X}^+$ 中计算每一个样本 $\boldsymbol{x}_i$ 和第一个元素 $\boldsymbol{x}_1$ 在核空间中的欧氏距离 $d_i(i=1,2,\cdots,|\boldsymbol{X}^+|)$,$d_i$ 值为:

$$d_i = \|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|^2 \tag{12}$$

(2)使用 BFPRT 算法[17,18] 根据 d_i 的值,搜索样本 $\boldsymbol{x}_k$,将 $\boldsymbol{X}^+$ 划分成等分的两个子集 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$,其中 $\boldsymbol{X}_1 = \{\boldsymbol{x}_i : d_i < d_k\}$ 和 $\boldsymbol{X}_2 = \{\boldsymbol{x}_i : d_i \geqslant d_k\}$. 然后,在 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 中分别重复 BFPRT 算法,直至在 $\boldsymbol{X}^+$ 中得到 $ceil(|\boldsymbol{X}^+|/P)$(其中 $ceil$ 函数表示向上舍入最接近的整数操作)个近似等分组,每组约含 P 个样本.

(3)在第 k 个分组 $\boldsymbol{X}_k$ 中,定义 $\boldsymbol{a}_k = \phi(\boldsymbol{x}_t)$,$\boldsymbol{x}_t \in \underset{i}{\operatorname{argmax}}\|\boldsymbol{x}_i\|^2, \boldsymbol{x}_i \in \boldsymbol{X}_k$,计算每个样本与 $\boldsymbol{a}_k$ 的核空间欧氏距离 d_i,然后使用 BFPRT 算法根据 d_i 的值搜索样本 $\boldsymbol{x}_t$,将 $\boldsymbol{X}_k$ 划分为两个子集:规模为 V 的子集 $\boldsymbol{X}' = \{\boldsymbol{x}_i : d_i < d_t\}$ 和子集 $\boldsymbol{X}'' = \{\boldsymbol{x}_i : d_i \geqslant d_t\}$.

(4)在子集 $\boldsymbol{X}''$ 重复 BFPRT 算法,直至将 $\boldsymbol{X}_k$ 划分为规模都不大于 V 的若干个分组.

分组工作结束后,数据集 $\boldsymbol{X}$ 中分组数 l 为:

$$l = (ceil(|\boldsymbol{X}^+|/P) + ceil(|\boldsymbol{X}^-|/P))ceil(P/V) \tag{13}$$

3.2.2　软性核凸包向量的选择

SCH-SVM 第 2 阶段在 l 个分组中并行执行,目标是获得 $\boldsymbol{X}$ 的软性核凸包集. 这一阶段的步骤为:(1)在每个分组中使用核化的支持向量数据域描述(Support Vector Data Description,SVDD)[19] 返回该分组样本的超球球心和半径;(2)依据样本点到球心的核空间欧式距离降序排列分组内所有样本;(3)将超球边界上的向量设为软性核凸包的初始集 $\boldsymbol{Z}^*$,依据排列结果检测每个样本是否是软性核凸包向量:

$$\min_{\mu}\left\|\phi(\boldsymbol{x}_i) - \sum_{t=1}^{|\boldsymbol{Z}^*|}\mu_{i,t}\phi(\boldsymbol{x}_t)\right\|^2$$

$$\text{s.t.}\quad \phi(\boldsymbol{x}_t) \in \boldsymbol{Z}, 0 \leqslant \mu_{i,t} \leqslant 1, \sum_{t=1}^{|\boldsymbol{Z}^*|}\mu_{i,t} = 1 \tag{14}$$

因为 $\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_i)$ 是常数,其值对式(14)的求解没有影响,丢弃该项后上式可表示为矩阵形式:

$$\min_{\mu} 2\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{Z}^*\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{Z}^{*\mathrm{T}}\boldsymbol{Z}^*\boldsymbol{\mu}$$

$$\text{s.t.}\quad \sum_{t=1}^{|\boldsymbol{Z}^*|}\mu_{i,t} = 1, 0 \leqslant \mu_{i,t} \leqslant 1 \tag{15}$$

显然,式(15)是一个标准的凸二次规划问题,对其求解可得向量 $\boldsymbol{\mu}$ 的全局最优解.

然后使用得到的 $\boldsymbol{\mu}$ 值判断样本是否满足下式:

$$\left\|\phi(\boldsymbol{x}_i) - \sum_{t=1}^{|\boldsymbol{Z}^*|}\mu_{i,t}\phi(\boldsymbol{x}_t)\right\|^2 \leqslant \varepsilon \tag{16}$$

如果其值小于 ε,说明 $\phi(\boldsymbol{x}_i)$ 可以用当前的软性核凸包向量线性表示,则 $\phi(\boldsymbol{x}_i)$ 不是软性核凸包向量;否则 $\phi(\boldsymbol{x}_i)$ 为软性核凸包向量,要将其加入到当前软性核凸包 $\boldsymbol{Z}^*$ 中,即 $\boldsymbol{Z}^* = \boldsymbol{Z}^* \cup \phi(\boldsymbol{x}_i)$.

3.2.3　分类器的训练和 SCH-SVM 的描述

正负类数据集 $\boldsymbol{X}^+$ 和 $\boldsymbol{X}^-$ 经过前 2 个阶段后得到 l 个分组的软性核凸包 $\boldsymbol{Z}_i^*$. SCH-SVM 将 $\boldsymbol{Z}^*$ 对应的原始样本作为训练样本代入式(10)得到决策函数.

根据上述的分析和推导,以下给出 SCH-SVM 在整个数据集上实施的具体步骤,见算法 1.

算法 1　SCH-SVM 描述

输入:正负类集 $\boldsymbol{X}^+$ 和 $\boldsymbol{X}^-$,分组参数 P 和 V,误差阈值 ε,正则化参数 C,核参数 σ,损失函数参数 τ;

输出:决策函数 $f(\boldsymbol{x})$;

令 $\boldsymbol{X} = \boldsymbol{X}^+$;

Step 1　根据式(15)计算样本 $\boldsymbol{x}_i(i=1,2,\cdots,|\boldsymbol{X}|)$ 与 $\boldsymbol{X}$

第一个元素 $\boldsymbol{x}_1$ 在核空间中的欧式距离 d_i；
令 $j=1$
Step 2 使用 BFPRT 算法将 $\boldsymbol{X}$ 划分成等分的两个子集 $\boldsymbol{X}_j=\{\boldsymbol{x}_i:d_i<d_k\}$ 和 $\boldsymbol{X}_j'=\{\boldsymbol{x}_i:d_i\geqslant d_k\}$；
Step 3 若 $j<|\boldsymbol{X}|/2P$，在 $\boldsymbol{X}_j$ 和 $\boldsymbol{X}_j'$ 中分别执行 Step 2，$j=j+1$；否则进入步骤 4；
For $k=1$ to $ceil(|\boldsymbol{X}|/P)$
Step 4 计算 $\boldsymbol{x}_t\in\arg\max_i\|\boldsymbol{x}_i\|^2$，$\boldsymbol{x}_i\in\boldsymbol{X}_k$，令 $\boldsymbol{a}_k=\phi(\boldsymbol{x}_t)$；
Step 5 计算 $\boldsymbol{X}_k$ 中所有点 $\boldsymbol{x}_i\in\boldsymbol{X}_k$ 与 $\boldsymbol{a}_k$ 的距离 d_i；
Step 6 根据 d_i 的值使用 BFPRT 将 $\boldsymbol{X}_k$ 划分为子集 $\boldsymbol{X}'=\{\boldsymbol{x}_i:d_i<d_t\}$ 和 $\boldsymbol{X}''=\{\boldsymbol{x}_i:d_i\geqslant d_t\}$，且 $|\boldsymbol{X}'|\leqslant V$；
Step 7 若 $|\boldsymbol{X}''|\leqslant V$，令 $\boldsymbol{X}_k=\boldsymbol{X}''$，转向步骤 4；否则 $k=k+1$；
End k
Step 8 $\boldsymbol{X}$ 划分 $ceil(|\boldsymbol{X}|/P)ceil(P/V)$ 个分组；
For $t=1$ to $ceil(|\boldsymbol{X}|/P)ceil(P/V)$
Step 9 在 $\boldsymbol{X}_t$ 中用 SVDD 算法计算超球球心和半径；
Step 10 按样本到球心距离值降序排列所有样本；
Step 11 将超球边界上的向量设为初始 $\boldsymbol{Z}^*$，降序排列每个样本 $\boldsymbol{x}_i$ 依次代入式(15)，检测是否满足式(16)，若否，$\boldsymbol{Z}^*=\boldsymbol{Z}^*\cup\boldsymbol{\phi}(\boldsymbol{x}_i)$；
End t
Step 12 令 $\boldsymbol{Z}^{+*}=\boldsymbol{Z}^*$；
Step 13 令 $\boldsymbol{X}=\boldsymbol{X}^-$，并转向 Step 1；
Step 14 令 $\boldsymbol{Z}^{-*}=\boldsymbol{Z}^*$；
Step 15 将 $\boldsymbol{Z}^{+*}$ 和 $\boldsymbol{Z}^{-*}$ 的原始空间样本作为两类训练样本代入式(10)，求解得到参数$(\boldsymbol{w},b)$；
Step 16 将参数$(\boldsymbol{w},b)$代入式(11)，得到决策函数.

4 SCH-SVM 的性质分析

4.1 时间复杂度分析

SCH-SVM 的时间复杂度由 3 部分构成. 第 1 阶段使用 BFPRT 算法将样本分为 l 组，时间复杂度为 $O(N\log_2(N/P))$，随后每个分组中将样本划分 P/V 个分组，时间复杂度为 $O(NP/V)$. 第 2 阶段时间复杂度集中在 SVDD 算法和式(15)的计算上. 本文采用序贯最小优化法(SMO)[20]求解二次规划问题，每个分组的时间复杂度为 $O(VS_i+VA_i^2)$，其中 S_i 为 SVDD 得到的支持向量数；A_i 为第 i 个分组当前核凸包向量数的最大值. 由于 VA_i^2 大于 VS_i，该阶段时间复杂度可写成 $O\left(\sum_{i=1}^{l}VA_i^2\right)$. 第3阶段是分类器的训练，其时间复杂度 $O(|\boldsymbol{Z}^*|^2)$，其中 $\boldsymbol{Z}^*$ 是 $\boldsymbol{X}$ 上获得的软性核凸包集. 因此，SCH-SVM 时间复杂度为 $O(N\log_2(N/P)+(NP/V)+\sum_{i=1}^{l}VA_i^2+|\boldsymbol{Z}^*|^2)$. 需要说明的是，$V$ 为分组中的样本数，其规模一般为 10^3 级别，而 A_i 又小于 V，$\boldsymbol{Z}^*$ 的规模远小于样本规模，因此，SCH-SVM 时间复杂度远小于传统 SVM 的 $O(N^3)$ 的时间复杂度.

4.2 SCH-SVM 分类精度分析

为了从理论上分析使用软性核凸包作为训练集给分类器带来的精度上的影响，将全部训练集代入 SCH-SVM 无约束目标函数，命名为 $F_1(\boldsymbol{w},b)$：

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{C}{N}\sum_{i=1}^{N}l(\boldsymbol{w},b,\phi(\boldsymbol{x}_i)) \tag{17}$$

其中，$l(\boldsymbol{w},b,\phi(\boldsymbol{x}_t))$ 为 pinball 损失函数.

SCH-SVM 无约束目标函数被命名为 $F_2(\boldsymbol{w},b)$. 为了更好地比较 $F_1(\boldsymbol{w},b)$ 和 $F_2(\boldsymbol{w},b)$ 间的关系，根据式(14)、(8)得到的权值 $r_{i,t}(1\leqslant i\leqslant N,1\leqslant t\leqslant M)$，定义新的无约束目标函数 $F_3(\boldsymbol{w},b)$，其 N 个训练样本由软性核凸包向量线性表示，形式如下：

$$\min_{\boldsymbol{w},b}F_3(\boldsymbol{w},b)=\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{C}{N}\sum_{i=1}^{N}l(\boldsymbol{w},b,u_i) \tag{18}$$

其中，$\boldsymbol{u}_i=\sum_{t=1}^{M}r_{i,t}\phi(\boldsymbol{x}_t)$，$\phi(\boldsymbol{x}_t)\in\boldsymbol{Z}^*$.

定理 1 根据无约束目标函数 $F_2(\boldsymbol{w},b)$ 和 $F_3(\boldsymbol{w},b)$ 分别在式(9)和式(18)中的定义，可得 $F_3(\boldsymbol{w},b)\leqslant F_2(\boldsymbol{w},b)$.

证明 因为 $\sum_{t=1}^{M}r_{i,t}=1$，可得：

$$\begin{aligned}&L_3(\boldsymbol{w},b,\boldsymbol{Z}^*)\\&=\frac{C}{N}\sum_{i=1}^{N}\max\left\{-\tau\left[1-y_i\left(\boldsymbol{w}^{\mathrm{T}}\sum_{t=1}^{M}r_{i,t}\phi(\boldsymbol{x}_t)+b\right)\right],\right.\\&\left.1-y_i\left(\boldsymbol{w}^{\mathrm{T}}\sum_{t=1}^{M}r_{i,t}\phi(\boldsymbol{x}_t)+b\right)\right\}\leqslant\frac{C}{N}\sum_{i=1}^{N}\sum_{t=1}^{M}\\&\cdot\max\{-\tau r_{i,t}[1-y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_t)+b)],\\&r_{i,t}[1-y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_t)+b)]\}\\&=L_2(\boldsymbol{w},b,\boldsymbol{Z}^*)\end{aligned}$$

由此可得 $F_3(\boldsymbol{w},b)\leqslant F_2(\boldsymbol{w},b)$.

定理 2 根据无约束目标函数 $F_1(\boldsymbol{w},b)$ 和 $F_3(\boldsymbol{w},b)$ 分别在式(17)和式(18)中的定义，可得：

$$\begin{aligned}&-\frac{C}{N}\sum_{i=1}^{N}\max\{y_i\boldsymbol{w}^{\mathrm{T}}\delta_i,-\tau y_i\boldsymbol{w}^{\mathrm{T}}\delta_i\}\\&\leqslant F_1(\boldsymbol{w},b)-F_3(\boldsymbol{w},b)\\&\leqslant\frac{C}{N}\sum_{i=1}^{N}\max\{-y_i\boldsymbol{w}^{\mathrm{T}}\delta_i,\tau y_i\boldsymbol{w}^{\mathrm{T}}\delta_i\}\end{aligned}$$

证明

$$\begin{aligned}L_1(\boldsymbol{w},b,\boldsymbol{X})&=\frac{C}{N}\sum_{i=1}^{N}\max\{-\tau[1-y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)\\&\quad+b)],1-y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\}\\&\leqslant L_3(\boldsymbol{w},b,\boldsymbol{X})\\&\quad+\frac{C}{N}\sum_{i=1}^{N}\max\{-y_i\boldsymbol{w}^{\mathrm{T}}\delta_i,\tau y_i\boldsymbol{w}^{\mathrm{T}}\delta_i\}\end{aligned}$$

即

$$F_1(\boldsymbol{w},b)-F_3(\boldsymbol{w},b)\leqslant\frac{C}{N}\sum_{i=1}^{N}\max\{-y_i\boldsymbol{w}^{\mathrm{T}}\delta_i,\tau y_i\boldsymbol{w}^{\mathrm{T}}\delta_i\}$$

类似

$$F_1(\boldsymbol{w},b)\geqslant F_3(\boldsymbol{w},b)-\frac{C}{N}\sum_{i=1}^{N}\max\{y_i\boldsymbol{w}^{\mathrm{T}}\delta_i,-\tau y_i\boldsymbol{w}^{\mathrm{T}}\delta_i\}$$

定理 3 设$(\boldsymbol{w}_1^*,b_1^*)$是 $F_1(\boldsymbol{w},b)$ 的最优解，$(\boldsymbol{w}_2^*,$

b_2^*)是 $F_2(\boldsymbol{w},b)$的最优解,则 $F_1(\boldsymbol{w}_1^*,b_1^*)-F_2(\boldsymbol{w}_2^*,b_2^*)\leqslant C\sqrt{C\varepsilon}$.

证明 $\frac{1}{2}\|\boldsymbol{w}_2^*\|^2+\frac{C}{N}\|\boldsymbol{\xi}^*\|_1=\|\boldsymbol{\alpha}^*\|_1-\frac{1}{2}\|\boldsymbol{w}_2^*\|^2$,由式(10)得$\|\boldsymbol{\alpha}^*\|_\infty\leqslant C/N$,即$\|\boldsymbol{\alpha}^*\|_1\leqslant C$,因此,$\|\boldsymbol{w}_2^*\|\leqslant\sqrt{C}$. 由定理2可得:

$$\begin{aligned}F_1(\boldsymbol{w}_1^*,b_1^*)-F_3(\boldsymbol{w}_2^*,b_2^*)&\leqslant\frac{C}{N}\sum_{i=1}^{N}\|\boldsymbol{w}_2^*\|\|\boldsymbol{\delta}_i\|\\&\leqslant\frac{C}{N}\sum_{i=1}^{N}\sqrt{C\varepsilon}\\&=C\sqrt{C\varepsilon}\end{aligned}$$

因为 $F_1(\boldsymbol{w}_1^*,b_1^*)\leqslant F_1(\boldsymbol{w}_2^*,b_2^*)$,结合定理1,可得:

$$\begin{aligned}F_1(\boldsymbol{w}_1^*,b_1^*)-F_2(\boldsymbol{w}_2^*,b_2^*)&\leqslant F_1(\boldsymbol{w}_1^*,b_1^*)-F_3(\boldsymbol{w}_2^*,b_2^*)\\&\leqslant F_1(\boldsymbol{w}_2^*,b_2^*)-F_3(\boldsymbol{w}_2^*,b_2^*)\\&=C\sqrt{C\varepsilon}\end{aligned}$$

定理4 设$(\boldsymbol{w}_1^*,b_1^*)$是 $F_1(\boldsymbol{w},b)$的最优解,$(\boldsymbol{w}_2^*,b_2^*)$是 $F_2(\boldsymbol{w},b)$的最优解,则 $F_1(\boldsymbol{w}_2^*,b_2^*)-F_1(\boldsymbol{w}_1^*,b_1^*)\leqslant 2C\sqrt{C\varepsilon}$.

证明 设$(\boldsymbol{w}_3^*,b_3^*)$和$(\boldsymbol{\alpha}_3,\hat{\boldsymbol{\alpha}}_3)$分别是 $F_3(\boldsymbol{w},b)$和其对偶式 $L_3(\boldsymbol{\alpha}_3,\hat{\boldsymbol{\alpha}}_3)$最优解. 设$(\boldsymbol{w}_2^*,b_2^*)$和$(\boldsymbol{\alpha}_2,\hat{\boldsymbol{\alpha}}_2)$分别是 $F_2(\boldsymbol{w},b)$和其对偶式 $L_2(\boldsymbol{\alpha}_2,\hat{\boldsymbol{\alpha}}_2)$最优解. 假设存在映射:

$$h(\tilde{\boldsymbol{\alpha}},\bar{\boldsymbol{\alpha}})=\left\{(a_t,\hat{a}_t):a_t=\sum_{i=1}^{N}r_{i,t}\tilde{a}_i\ \text{and}\ \hat{a}_t=\sum_{i=1}^{N}r_{i,t}\bar{a}_i\right\}$$

则$(\tilde{\boldsymbol{\alpha}}_2,\bar{\boldsymbol{\alpha}}_2)$也是 $L_3(\boldsymbol{\alpha}_3,\hat{\boldsymbol{\alpha}}_3)$最优解:

$$\begin{aligned}L_2(h(\tilde{\boldsymbol{\alpha}}_2,\bar{\boldsymbol{\alpha}}_2))&=\sum_{t=1}^{M}\alpha_t-\frac{1}{2}\sum_{t=1}^{M}\sum_{s=1}^{M}\alpha_t\alpha_s y_t y_s\phi(\boldsymbol{x}_t)^{\mathrm{T}}\phi(\boldsymbol{x}_s)\\&=\sum_{i=1}^{N}\tilde{a}_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{a}_i\tilde{a}_j y_i y_j\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j\\&=L_3(\tilde{\boldsymbol{\alpha}}_2,\bar{\boldsymbol{\alpha}}_2)\end{aligned}$$

因为,$L_3(\boldsymbol{\alpha}_3,\hat{\boldsymbol{\alpha}}_3)\geqslant L_3(\tilde{\boldsymbol{\alpha}}_2,\bar{\boldsymbol{\alpha}}_2)$. 可得,$F_3(\boldsymbol{w}_3^*,b_3^*)\geqslant F_2(\boldsymbol{w}_2^*,b_2^*)$. 从定理1可知,$F_3(\boldsymbol{w}_3^*,b_3^*)\leqslant F_3(\boldsymbol{w}_2^*,b_2^*)\leqslant F_2(\boldsymbol{w}_2^*,b_2^*)$. 因此,$F_3(\boldsymbol{w}_3^*,b_3^*)=F_3(\boldsymbol{w}_2^*,b_2^*)$. 并由定理2可得:

$$-\frac{C}{N}\sum_{i=1}^{N}\max\{y_i\boldsymbol{w}_1^{*\mathrm{T}}\boldsymbol{\delta}_i,-\tau y_i\boldsymbol{w}_1^{*\mathrm{T}}\boldsymbol{\delta}_i\}$$

$$\leqslant F_1(\boldsymbol{w}_1^*,b_1^*)-F_3(\boldsymbol{w}_1^*,b_1^*),F_1(\boldsymbol{w}_2^*,b_2^*)-F_3(\boldsymbol{w}_2^*,b_2^*)$$

$$\leqslant\frac{C}{N}\sum_{i=1}^{N}\max\{-y_i\boldsymbol{w}_2^{*\mathrm{T}}\boldsymbol{\delta}_i,\tau y_i\boldsymbol{w}_2^{*\mathrm{T}}\boldsymbol{\delta}_i\}.\ \text{所以}$$

$$\begin{aligned}F_1(\boldsymbol{w}_2^*,b_2^*)-F_1(\boldsymbol{w}_1^*,b_1^*)&\leqslant\frac{C}{N}\sum_{i=1}^{N}\|\boldsymbol{w}_2^*\|\|\boldsymbol{\delta}_i\|+\|\boldsymbol{w}_1^*\|\|\boldsymbol{\delta}_i\|\\&\leqslant 2C\sqrt{C\varepsilon}\end{aligned}$$

由定理1~4可看出,SCH-SVM 与使用全部训练集的目标函数相比,两者的最优结果非常接近.

5　实验结果与分析

5.1　实验设置

为验证 SCH-SVM 的有效性,本文实验分成两个部分:(1)SCH-SVM 中参数的选择;(2)SCH-SVM 与基线分类器 pin-SVM 以及4种大规模数据分类器的比较,包括基于最小包含球技术的 CVM[4] 和 BVM[21],基于代表点技术的 AESVM[7] 及基于核密度估计采样技术的 FastKDE[6]. 实验采用3个评价指标:分类器的训练时间,分类精度(方差)和分类器得到的支持向量数.

实验使用了8个真实 UCI 数据集[22](基本信息见表1),参照文献[13,14,23]的方法,随机选择50%和80%的样本并加入均值为0的高斯白噪声,为了检验分类器对噪声强度的敏感性,所加噪声强度为4种:方差分别为样本特征值的10%,20%,50%和100%. 另外,实验参数设置如下:SCH-SVM 中分组参数 P 和 V 的范围分别为$\{10^4,5\times10^4,10^5\}$和$\{10^3,2\times10^3,5\times10^3\}$,pinball 损失函数参数 τ 范围为$\{0.1,0.2,0.5,1\}$. 所有分类器中的核函数采用高斯核,核参数范围为$\{10^{-3},10^{-2},\cdots,10^3\}$,正则化参数范围为$\{10^{-3},10^{-2},\cdots,10^3\}$(为保持参数的一致性,SCH-SVM 的正则化参数为 C/N). SCH-SVM,CVM,BVM 和 AESVM 的误差阈值 $\varepsilon=10^{-4}$. 另外,根据各分类器的默认参数设置,AESVM 中分组参数为 10^5 和 10^3;FastKDE 的采样策略为随机抽取训练集2%的样本. 实验采取10重交叉验证法来选取参数最优值,且每个数据集重复实验5次. 本文实验在2.53-GHz quad-core CPU,8-GB RAM,Windows 7 系统下执行,所有分类器均在 VC 环境下实现.

表1　数据集的基本信息

ID	数据集	规模	维数	ID	数据集	规模	维数
ijc	ijcnn1	49990	22	kdd	kdd99	310000	41
shu	shuttle	58000	9	cod	cod-ncRNA	486201	8
5sn	5s-ncRNA	95172	8	che	checkerboard	510000	2
loc	localization	164860	7	cov	covtype	590012	54

5.2 SCH-SVM 中参数的选择

SCH-SVM 中分组参数{P,V}以及高斯核参数 σ 与软性核凸包规模及算法运行时间密切相关,另外 σ 值对分类精度的影响也较大,而根据文献[24]分析,σ 值宜在一定范围内搜索得到. 因此下面列出了参数 P 和 V 在 *ijc* 和 *kdd* 集上不同噪声比和噪声强度下对软性核凸包规模和第 1 至第 2 阶段运行时间的影响(固定 σ = 1),实验结果如表 2 ~ 3 所示. 表中数据集描述为:(噪声比,噪声强度). 可得出:(1)随着 P 值的增加,SCH-SVM 第 1 ~ 2 阶段的运行时间有所增加,同时软性核凸包的规模有所下降. 这是因为在固定参数 V 的情况下,P 的取值与分组个数成反比关系,P 值越小,分组数就越多,则软性核凸包的规模就大. 同时,分组数又影响着 SCH-SVM 第 2 阶段的运行时间,分组数越多,每个分组中包含的样本数就越少,运行时间就越短. 为兼顾软性核凸包规模和运行时间,P 设置为 5×10^4. (2)软性核凸包的规模随着每个分组中样本数 V 值增大而降低. 因为 V 值越大,分组数越少,虽然在单个分组中软性核凸包数可能会增加,但总的软性核凸包数却随着分组数减少而减少. 从表 2 ~ 3 可以看出,随着每个分组中样本数 V 值增加,SCH-SVM 第 2 阶段的运行时间亦增加. 考虑到软性核凸包规模与运行时间的平衡,V 设置为 2×10^3.

表2 不同 P 和 V 的 SCH-SVM 在 ijc 集的运行时间(秒)和软性核凸包向量数 (括号内为软性核凸包向量数)

	(50, 10)			(50, 20)			(50, 50)			(50, 100)		
P \ V	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3
10^4	3.5 (4027)	3.8 (2984)	6.8 (2792)	3.4 (4082)	3.8 (2875)	6.7 (2750)	3.5 (4024)	3.9 (2835)	6.7 (2770)	3.3 (3938)	4.3 (2902)	6.8 (2774)
5×10^4	3.7 (3610)	4.0 (2746)	7.6 (2685)	3.8 (3587)	4.6 (2738)	7.7 (2639)	3.7 (3557)	4.2 (2755)	7.5 (2688)	3.9 (3472)	4.6 (2743)	7.7 (2655)
10^5	3.8 (3481)	4.5 (2657)	8.4 (2653)	3.9 (3416)	4.9 (2679)	8.5 (2626)	3.8 (3549)	4.5 (2618)	8.4 (2669)	3.9 (3452)	4.9 (2698)	8.5 (2643)
	(80, 10)			(80, 20)			(80, 50)			(80, 100)		
P \ V	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3
10^4	3.5 (3998)	4.5 (3028)	6.8 (2786)	3.5 (4073)	4.5 (3063)	7.0 (2836)	3.4 (4012)	4.5 (3110)	6.7 (2875)	3.6 (3970)	4.5 (3218)	6.5 (2557)
5×10^4	3.9 (3565)	4.7 (2762)	7.5 (2724)	3.9 (3561)	4.7 (2741)	7.6 (2683)	3.7 (3590)	4.9 (2827)	7.4 (2666)	3.9 (3603)	4.8 (2928)	6.7 (2488)
10^5	3.9 (3525)	5.1 (2719)	8.4 (2682)	4.0 (3498)	5.2 (2722)	7.9 (2669)	4.0 (3535)	5.3 (2813)	8.0 (2619)	4.0 (3521)	4.9 (2803)	6.7 (2472)

表3 不同 P 和 V 的 SCH-SVM 在 kdd 集的运行时间(秒)和软性核凸包向量数 (括号内为软性核凸包向量数)

	(50, 10)			(50, 20)			(50, 50)			(50, 100)		
P \ V	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3
10^4	200.4 (6019)	257.2 (5587)	427.3 (5274)	210.2 (6032)	264.8 (5518)	432.9 (5293)	207.0 (6026)	228.7 (5660)	430.2 (5604)	210.7 (6123)	264.8 (5481)	455.2 (5262)
5×10^4	255.2 (5801)	290.9 (4922)	496.5 (4864)	248.0 (5805)	280.9 (4805)	478.0 (4790)	260.1 (5739)	309.5 (4729)	502.1 (4770)	260.4 (5736)	281.5 (4869)	480.3 (4803)
10^5	277.1 (5497)	379.7 (4754)	532.2 (4480)	286.5 (5568)	388.3 (4694)	519.7 (4404)	289.3 (5399)	357.2 (4703)	552.7 (4401)	280.5 (5507)	364.6 (4401)	527.6 (4504)
	(80, 10)			(80, 20)			(80, 50)			(80, 100)		
P \ V	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3	10^3	2×10^3	5×10^3
10^4	197.4 (6128)	254.8 (5613)	419.4 (5329)	204.6 (6083)	260.3 (5669)	443.0 (5258)	250.3 (5997)	259.3 (5490)	442.0 (5307)	207.6 (6187)	244.6 (5506)	468.7 (5210)
5×10^4	269.5 (5758)	316.1 (4922)	500.7 (4812)	255.6 (5892)	300.6 (4805)	492.5 (4800)	258.3 (5791)	289.8 (4729)	473.7 (4701)	270.9 (5758)	286.0 (4869)	496.5 (4784)
10^5	288.2 (5406)	365.7 (4728)	538.0 (4420)	283.2 (5590)	360.8 (4782)	549.3 (4399)	271.1 (5500)	370.4 (4717)	520.4 (4618)	296.6 (5381)	358.7 (4580)	528.4 (4508)

表 4 不同 τ 的 SCH-SVM 在 ijc 集上的分类精度(%)和标准差

τ	(50, 10)	(50, 20)	(50, 50)	(50, 100)	(80, 10)	(80, 20)	(80, 50)	(80, 100)
0.1	96.38±1.14	95.87±1.08	94.41±1.25	94.15±1.17	95.60±1.21	95.45±1.22	95.00±1.32	93.64±1.29
0.2	**96.56**±1.08	96.14±1.21	94.60±1.32	94.19±1.20	95.69±1.18	**95.62**±1.23	95.00±1.28	93.78±1.24
0.5	96.47±1.02	**96.18**±1.10	**94.70**±1.25	**94.28**±1.19	**96.00**±1.20	95.61±1.19	95.05±1.13	**93.99**±1.30
1	96.32±1.13	96.12±0.97	94.53±1.19	94.20±1.23	95.87±1.54	95.42±1.20	**95.09**±1.30	93.75±1.31

SCH-SVM 第 3 阶段涉及 3 个参数:正则化参数,高斯核核宽 σ 和 pinball 损失函数参数 τ. 其中正则化参

数直接影响其泛化性能,宜在设定的范围内搜索得到[24],因此本文实验采用网格搜索方式来寻找正则化参数和高斯核核宽的最优值. 受篇幅限制,以下列出参数 τ 的敏感性实验,结果如表 4 ~ 5 所示. 可以看出,在不同的噪声比和噪声强度下,所有数据集在 τ 取值范围{0.1,0.2,0.5,1}的时候均取得了较高的分类准确率和较小的方差. 为减少寻优参数的个数,在后面的实验中 τ 值固定为 0.5.

表5　不同 τ 的 SCH-SVM 在 kdd 集上的分类精度(%)和标准差

τ	(50,10)	(50,20)	(50,50)	(50,100)	(80,10)	(80,20)	(80,50)	(80,100)
0.1	93.37±0.85	93.19±0.78	92.45±0.98	92.05±0.98	92.65±1.00	91.82±0.66	91.57±0.75	90.46±0.69
0.2	**93.82**±0.86	93.14±0.86	92.67±0.78	91.98±1.02	**92.98**±0.85	**92.40**±0.75	91.80±0.78	90.77±0.84
0.5	93.77±0.69	**93.27**±0.83	**92.89**±0.85	**92.30**±0.87	92.96±0.90	92.29±0.72	**91.86**±0.73	**91.00**±0.71
1	93.64±0.80	93.07±0.88	**92.89**±0.91	91.98±0.75	92.64±1.02	92.29±0.59	91.76±0.70	**91.00**±0.72

5.3　对比实验

本节给出各分类器在 8 个大规模噪声数据集上的性能比较,实验首先比较了在噪声比 50% 情况下各分类器的训练时间,分类精度和包含的支持向量数,实验结果如图 2 ~ 4 所示.

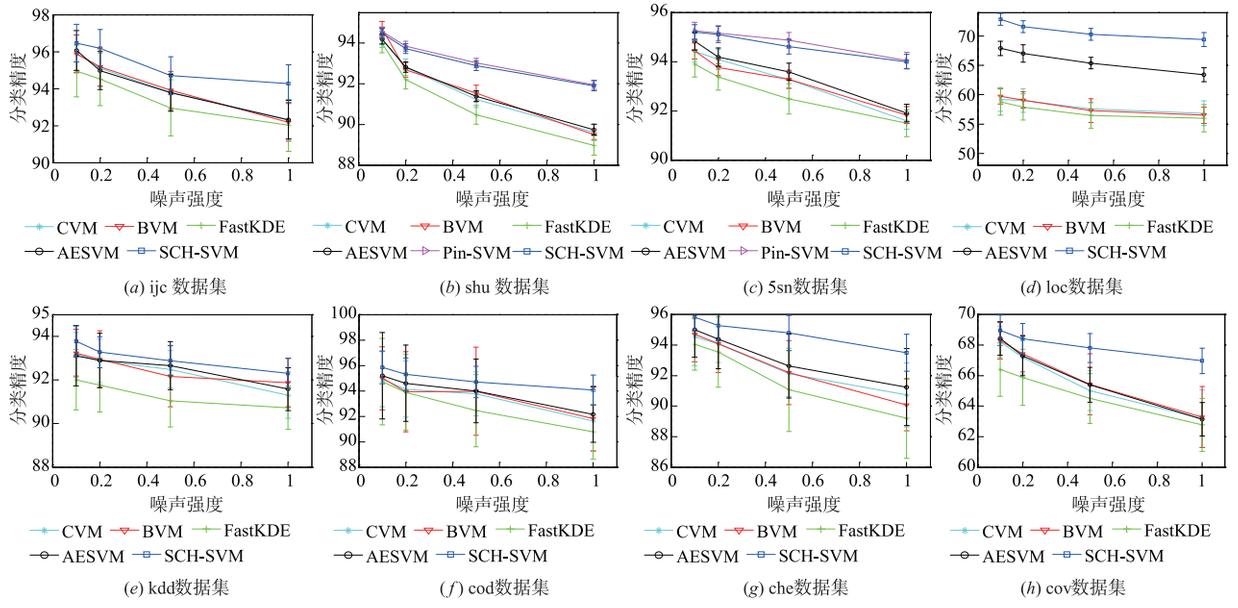


图2　所有方法在噪声比50%情况下的分类精度(%)比较

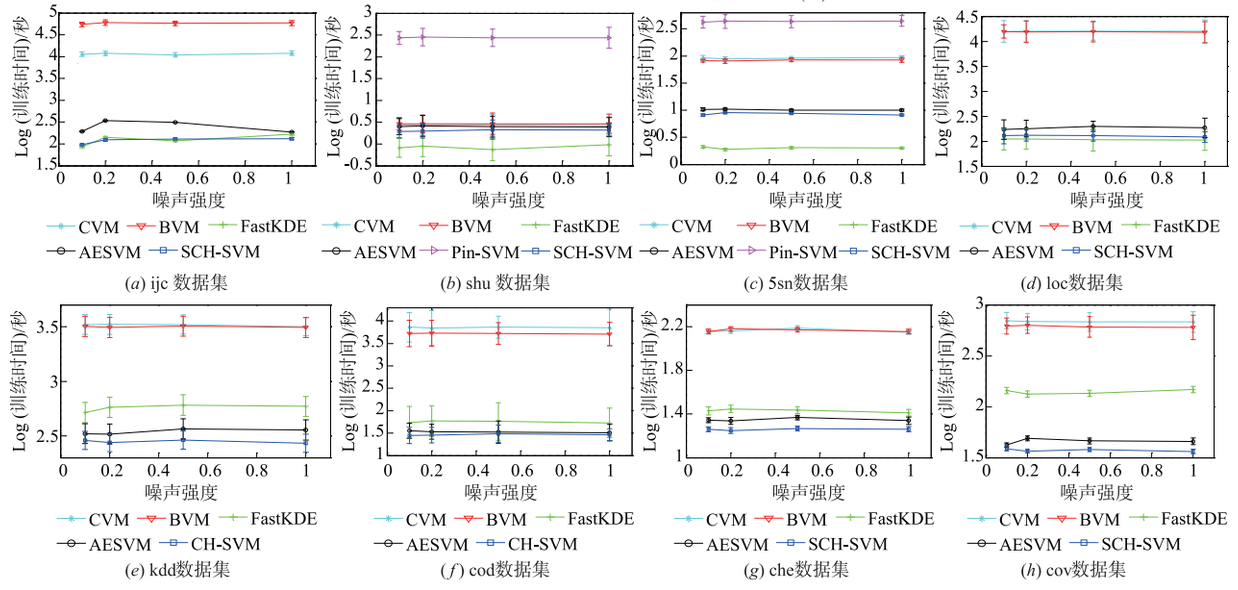


图3　所有方法在噪声比50%情况下的训练时间(秒)比较

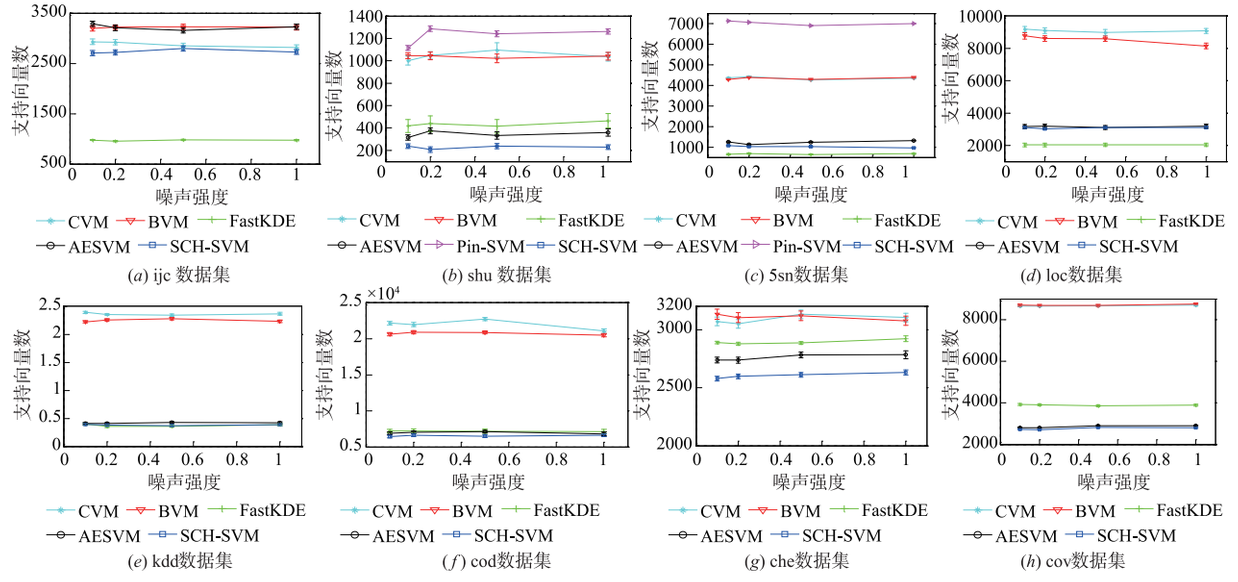


图4 所有方法在噪声比50%情况下的支持向量数比较

(1)从图2可见,随着噪声强度的增加各分类器的精度都呈现出不同程度地下降. 特别是CVM、BVM、AESVM和FastKDE,随着噪声强度增加,其分类精度迅速降低. 而软性核凸包能体现样本在核空间的几何轮廓,加之使用最大化两类分位数距离得到抗噪的分类器模型,SCH-SVM在分类精度上取得令人满意的结果. pin-SVM在有训练结果的2个数据集上也获得良好的分类结果,说明pinball损失函数是对噪声不敏感的损失函数.

(2)对于训练时间而言,SCH-SVM与pin-SVM、CVM和BVM相比优势明显. 特别是pin-SVM,仅在2个规模较小的集上有训练结果,而在其它6个集上无法求解(设定训练时间上限为3小时). CVM和BVM计算量主要集中在找到相应的近似最小包含球,随着最小包含球的规模变大,其计算量相当耗时. 此外,FastKDE使用简单采样得到的训练集规模很大时,在训练时间上就不具有优势.

(3)图4显示了各分类器在训练后获得的支持向量数以及标准差,其中SCH-SVM获得的支持向量数较少. 由于FastKDE使用的训练样本仅占全部训练集的2%,其获得的支持向量数也较少. 再者是AESVM,其通过选择代表点来缩减训练集的规模,得到的支持向量数也较少. 根据SVM的基本性质,支持向量数的多少直接决定了其执行分类操作的时间,因此,从图4结果可以看出SCH-SVM不仅在训练时间上具有优势,在分类

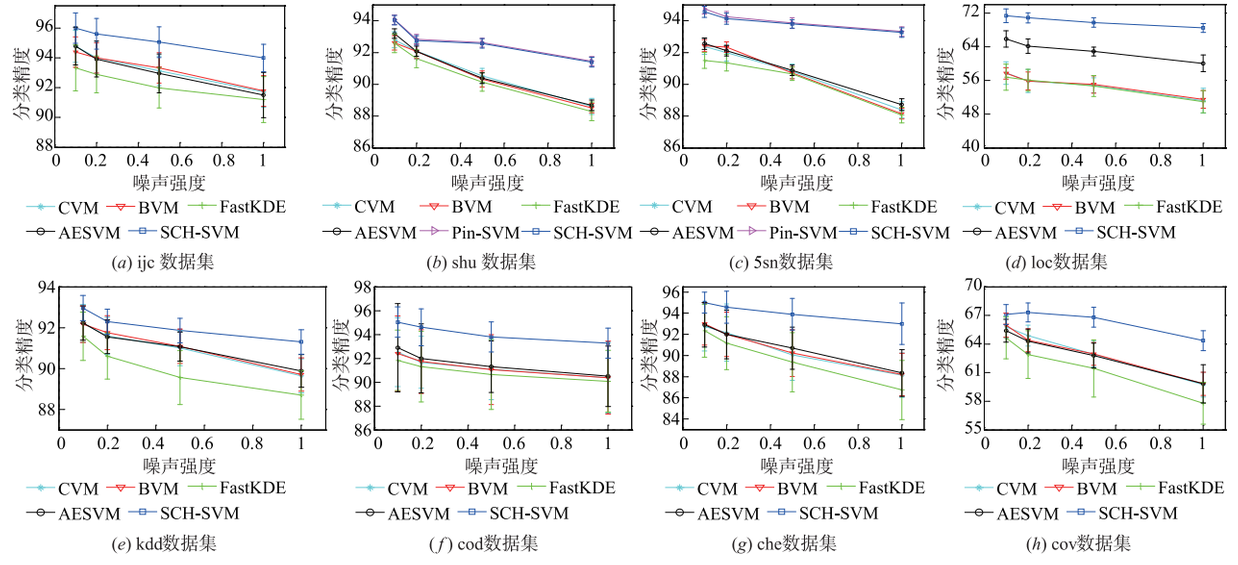
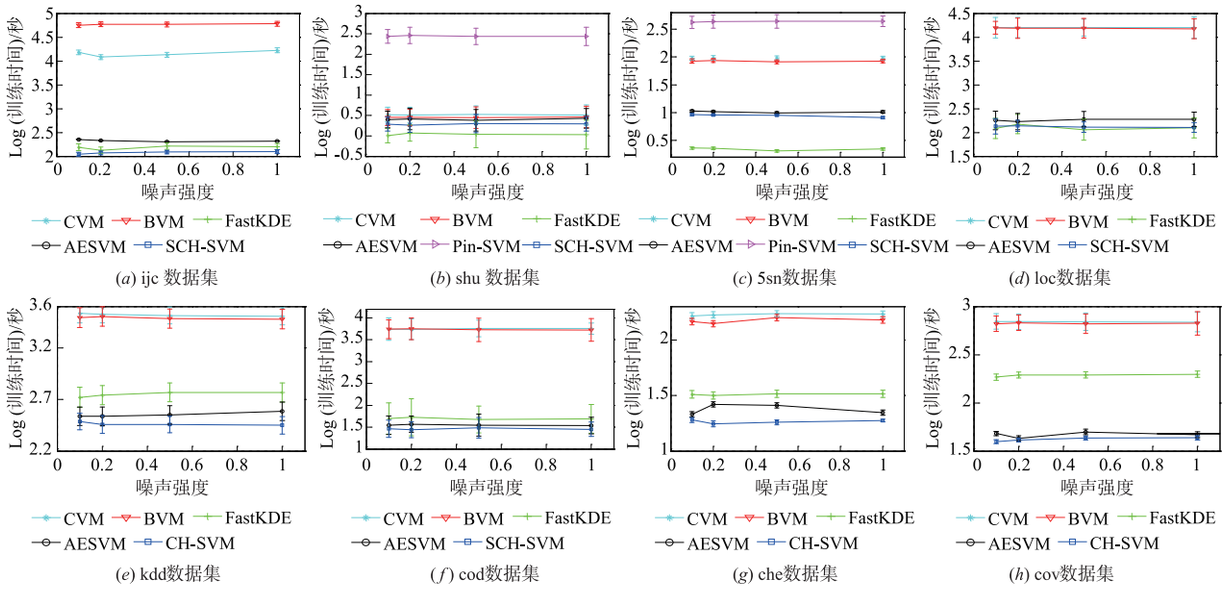


图5 所有方法在噪声比80%情况下的分类精度(%)比较

时间上也同样具有优势.

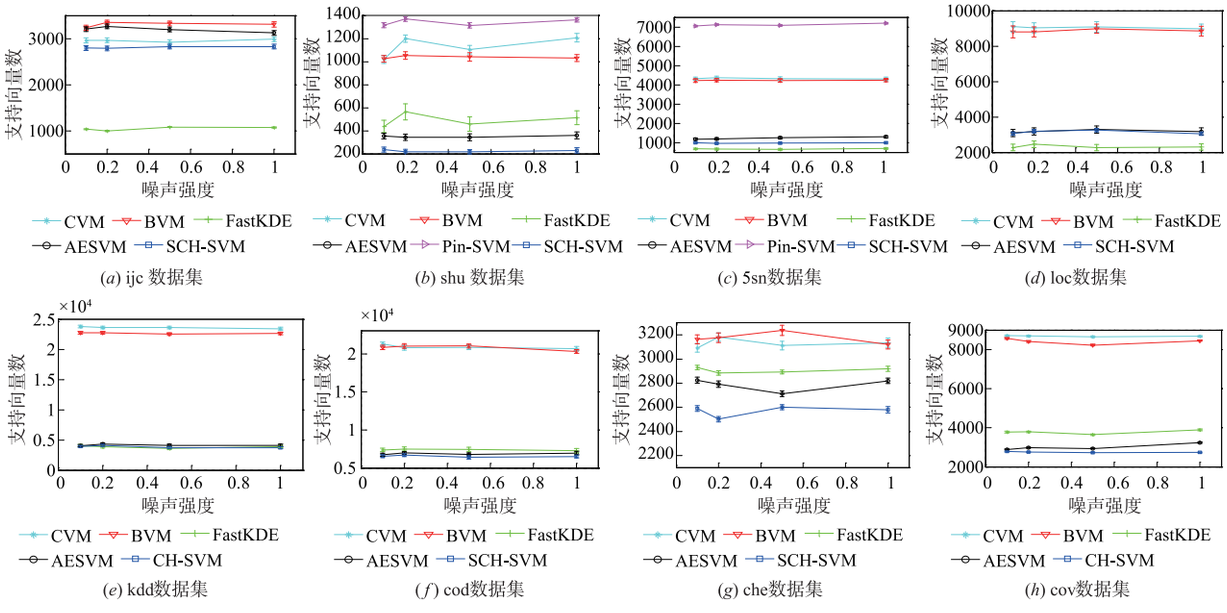
实验还比较了噪声比 80% 时各分类器的分类精度,训练时间和支持向量数,各分类器对应的 3 个评价指标的变化曲线如图 5 ~ 7 所示. 从实验结果可以看出,噪声比 80% 时获得的实验结果与噪声比 50% 时的实验结果是一致的. 具体实验结论如下:



(*a*) ijc 数据集　(*b*) shu 数据集　(*c*) 5sn数据集　(*d*) loc数据集

(*e*) kdd数据集　(*f*) cod数据集　(*g*) che数据集　(*h*) cov数据集

图6　所有方法在噪声比80%情况下的训练时间(秒)比较



(*a*) ijc 数据集　(*b*) shu 数据集　(*c*) 5sn数据集　(*d*) loc数据集

(*e*) kdd数据集　(*f*) cod数据集　(*g*) che数据集　(*h*) cov数据集

图7　所有方法在噪声比80%情况下的支持向量数比较

(1)从图 5 可以看出,SCH-SVM 在噪声比增加的情况下更能体现较强的噪声不敏感性,与图 2 噪声比 50% 时的结果相比,分类精度仅略有下降. 而 CVM、BVM、AESVM 和 FastKDE 在噪声比增加的情况下,分类精度迅速下降,其中 FastKDE 简单采样的策略更易受到噪声的干扰,分类精度最低.

(2)从图 6 可以看出,噪声比的增加对训练时间影响不明显,各分类器的训练时间与图 2 中数据差别不大. SCH-SVM 在 5 个数据集的训练时间最少. 另外,AESVM 较 CVM 和 BVM 也花费较少的时间. 而 pin-SVM 在处理大规模样本的分类问题时,计算量太大仅在 shu 和 5sn 集上训练出了分类器模型.

(3)从图 7 可以看出,噪声比的增加对各分类器得到的支持向量数略有影响,因为噪声样本改变了样本

在核空间的几何轮廓. 与图 4 结果对比发现,两种结果的偏差在 10% 以内. 另外,SCH-SVM 和 FastKDE 获得的支持向量数最少,由于 pin-SVM 使用全部训练集进行分类器的训练,在有结果的 2 个数据集上得到的支持向量数最多.

6 结论

本文从样本投影到核空间的几何轮廓出发给出了软性核凸包向量的定义,并依据这一定义对训练集样本分组选择出全部训练集上的软性核凸包向量,然后将所选软性核凸包对应的原始空间样本作为训练集进行使用了 pinball 损失函数的 SVM 分类器的训练. 此外,文中理论证明了本文方法在有效约减训练样本的同时能有效地保持分类器的分类精度. 实验结果亦表明,在大规模含噪声数据的分类问题中本文方法与 CVM,BVM,FastKDE 和 AESVM 等分类器相比,在分类精度,支持向量数和训练时间上具有明显的优势. 但应当指出,本文方法依然面临一些进一步需要探讨的问题:如何将本文方法扩展到的大规模噪声数据的多分类问题等.

参考文献

[1] Bo L F, Wang L L, Jiao C. Training hard-margin support vector machines using greedy stagewise algorithm [J]. IEEE Trans. Neural Networks, 2008, 19(8): 1446 – 1455.

[2] Fine S, Scheinberg K. Efficient SVM training using low-rank kernel representations [J]. The Journal of Machine Learning Research, 2001, 2(3): 243 – 264.

[3] Ni T G, Chung F L, Wang S T. Support vector machine with manifold regularization and partially labeling privacy protection [J]. Information Sciences, 2015, 294(10): 390 – 407.

[4] Tsang I W, Kwok J T, Cheung P M. Core vector machines: fast SVM training on very large data sets [J]. The Journal of Machine Learning Research, 2005, 6(12): 363 – 392.

[5] 胡文军,王士同. 隐私保护的 SVM 快速分类方法[J]. 电子学报,2012,40(2):280 – 286.
Hu Wenjun, Wang Shitong. Fast classification approach support vector machine with privacy preservation [J]. Acta Electronica Sinica, 2012, 40(2): 280 – 286. (in Chinese)

[6] Wang S T, Wang J, Chung F L. Kernel density estimation, kernel methods, and fast learning in large data sets [J]. IEEE Trans. Cybernetics, 2014, 44(1): 1 – 20.

[7] Nandan M, Khargonekar P P, Talathi S S. Fast SVM training using approximate extreme points [J]. Journal of Machine Learning Research, 2014, 15: 59 – 98.

[8] Kim W, Stankovc M S, Johansson K H, et al. A distributed support vector machine learning over wireless sensor networks [J]. IEEE Trans Cybernetics, 2015, 45(11): 2599 – 2611.

[9] Batuwita R, Palade V. FSVM-CIL: fuzzy support vector machines for class imbalance learning [J]. IEEE Trans Fuzzy Systems, 2010, 18(3): 558 – 571.

[10] Wang Y, Wang S, Lai K. A new fuzzy support vector machine to evaluate credit risk [J]. IEEE Trans Fuzzy System, 2005, 13(6): 820 – 831.

[11] An W J, Liang M G. Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises [J]. Neurocomputing, 2013, 110(6): 101 – 110.

[12] Huang X L, Shi L, Pelckmansb K, Suykens J A K. Asymmetric ν-tube support vector regression [J]. Computational Statistics and Data Analysis, 2014, 77: 371 – 382.

[13] Huang X L, Shi L, Suykens J A K. Support vector machine classifier with pinball loss [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2014, 36(5): 984 – 997.

[14] He X Y, Mourot G, Maquin D, et al. Multi-task learning with one-class SVM [J]. Neurocomputing, 2014, 133: 416 – 426.

[15] Wang D, Qiao H, Zhang B, et al. Online support vector machine based on convex hull vertices selection [J]. IEEE Trans. Neural Networks and Learning Systems, 2013, 24(4): 593 – 609.

[16] Dattorro J. Convex Optimization and Euclidean Distance Geometry [M]. Mεβoo Publishing USA, 2015.

[17] Blum M, Floyd R W, Pratt V R, et al. Time bounds for selection [J]. Journal of Computer and System Sciences, 1973, 7(4): 448 – 461.

[18] Xua J, Jiang Y X, Zeng C Q, et al. Node anomaly detection for homogeneous distributed environments [J]. Expert Systems with Applications, 2015, 42(20): 7012 – 7025.

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

[20] Takahashi N, Nishi T. Rigorous proof of termination of SMO algorithm for support vector machines [J]. IEEE Trans Neural Network, 2005, 16(3): 774 – 776.

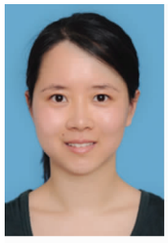
[21] Tsang I, Kwok A, Kwok J. Simpler core vector machines with enclosing balls [A]. International conference on Machine learning [C]. Corvallis, USA, 2007. 911 – 918.

[22] Bache K, Lichman M. UCI database [DB/OL]. http://www. ics. uci. edu/%20mlearn/MLRepository. html.

[23] Luukka P, Lampinen J. Differential evolution classifier in noisy settings and with interacting variables [J]. Applied Soft Computing, 2011, 11: 891 – 899.

[24] Chang C, Lin C. LIBSVM: A library for support vector machines [J]. ACM Trans Intelligence and System Technology, 2011, 2(3): 1 – 27.

作者简介



顾晓清　女,1981 年出生,江苏常州人,2017 年获江南大学博士学位,现任常州大学讲师,研究方向为模式识别,模糊系统.
E-mail:czxqgu@163.com

倪彤光(通信作者)　男,1978 年出生,河北邢台人,2015 年获江南大学博士学位,现任常州大学讲师,研究方向为模式识别与人工智能.
E-mail:hbxtntg-12@163.com

姜志彬　男,1991 年出生,山东烟台人,江南大学博士研究生,主要研究方向为模式识别.

王士同　男,1964 年出生,江苏扬州人,江南大学教授、博士生导师,主要从事人工智能、模式识别、模糊系统、医学图像处理和生物信息学等方面的研究工作.