第 40 卷 第 5 期
2021 年 9 月

应用声学
Journal of Applied Acoustics

Vol. 40, No. 5
September, 2021

◇ 研究报告 ◇

基于改进支持向量机的水声目标-杂波不平衡分类研究*

关 鑫 李然威† 胡 鹏 冯金鹿 何荣钦

(中国船舶第七一五研究所 杭州 310023)

摘要: 针对水声目标-杂波数据集在有限样本下的类不平衡特性导致代价敏感支持向量机难以逼近贝叶斯最优决策的问题,该文提出了一种基于能量统计方法的支持向量机(En-SVM)。该算法通过度量原始数据空间与有限样本空间特征函数之间的加权平方距离,量化少数类样本不完全采样过程中的信息损失,来补偿再生核希尔伯特空间中机器学习算法所需的少数类分类信息,增加少数类样本对决策的影响力。实验结果表明,En-SVM 能够在保持高检测概率的同时获得较低虚警概率,即通过分类可以排除大量的杂波,性能优于标准支持向量机和代价敏感支持向量机,能够有效处理水声不平衡数据的分类问题,实现主动声呐信号处理中的杂波抑制。

关键词: 目标杂波分类;不平衡分类;支持向量机;能量统计

中图法分类号: TB566 **文献标识码:** A **文章编号:** 1000-310X(2021)05-0715-08

DOI: 10.11684/j.issn.1000-310X.2021.05.009

The imbalanced classification of underwater acoustic target-clutter based on improved support vector machine

GUAN Xin LI Ranwei HU Peng FENG Jinlu HE Rongqin

(*China Shipbuilding 715th Research Institute, Hangzhou* 310023, *China*)

Abstract: This paper mainly proposed a novel algorithm En-SVM based on energy statistics method for the imbalance characteristics of acoustic target-clutter data sets resulted in cost sensitive support vector machines(CS-SVM) didn't approach Bayesian optimal decision in limited samples. This algorithm measured the weighted square distance between the original data space and the feature function of the limited sample space, quantifies the information loss in the incomplete sampling process of a few samples, so as to compensate for the minority class classification information required by the machine learning algorithm in the reproducing kernel Hilbert space, and increased the influence of the minority class samples on the decision-making. The experimental results show that the proposed algorithm can obtain a lower false alarm probability while maintaining a high detection probability, which means that a large amount of clutter can be eliminated by classification, and the performance is better than that of standard SVM and CS-SVM, which can effectively deal with the classification problem of underwater acoustic unbalanced data and realize clutter suppression of active sonar signal processing.

Keywords: Target-clutter classification; Imbalance classification; Support vector machine; Energy statistics

2020-10-09 收稿;2021-04-07 定稿

作者简介: 关鑫 (1994–), 男, 陕西丹凤人, 硕士研究生, 研究方向: 水声信号处理。

† 通信作者 E-mail: lirw501@sina.com

0 引言

通常主动声呐较被动声呐具备探测距离优势，但是，在工作过程中经常伴随着大量的杂波虚警，并且随着水下目标隐身降噪技术的发展，探测难度不断加大[1]，尤其是在浅海海域，分布着礁石、海底山脊、山峰和沉船等强散射体，主动发射信号接触这些散射体，会产生和目标强度相近的回波，在探测画面上出现大量类目标杂波亮点。大量杂波的存在对主动声呐探测性能主要有两方面的影响，第一，难以通过调整信噪比门限，在不损失检测概率的同时降低虚警概率；第二，在自动跟踪端生成大量虚假航迹，影响航迹关联，加剧跟踪系统的计算负担，甚至导致跟踪系统瘫痪。因此，杂波抑制是主动声呐信号处理中的重要研究问题，通过对目标和杂波的分类判别，可以有效解决这个问题[2]。

随着大数据时代的到来，从海量数据中挖掘有效信息的需求推动了机器学习的发展，Berg 等[2] 为了解决自主水下潜航器群 (Autonomous underwater vehicles, AUVs) 受制于有限的通信能力而不能共享大量主动声呐探测数据的问题，研究了 k 近邻 (k near neighbor, k-NN)、ID3、朴素贝叶斯 (Naive Bayes) 和神经网络 (Neural network) 等机器学习算法，通过对目标和杂波的分类来缩减探测数据。Stender 等[3−4] 指出在跟踪阶段，由海底地形特征物 (海山、山脊等) 产生的杂波和人造特征物 (无人潜航器 (Underwater unmanned vehicle, UUV)、潜艇等) 产生的回波运动特性不同，建立了包含运动航迹和信噪比特征的数据集，训练机器学习模型，能够准确地从背景中发现人造特征物。可见，机器学习能够利用数据发现一些潜在的变化规律用来预测未知数据，为水声目标和杂波的分类提供了一种新的解决思路。

然而，以上研究[1−4] 并未考虑水声数据集的类不平衡特性，即主动声呐使用中海底/海面的不平整性、航船辐射噪声等对水声数据采集带来大量的杂波干扰，一个水下目标回波通常伴随着数百个杂波。因而，相应的机器学习分类问题为不平衡分类问题，即在一个分类问题中某些类的样本数量远多于其他类别的样本数量[5]。一般的机器学习分类算法不适合处理类不平衡数据[6−7]，因为机器学习算法在训练的过程中基于整体分类误差最小构建分类模型，导致多数类样本的分类准确率存在高于少数类样本的趋势[8]，整体分类准确率主要受前者影响而变高，但是少数类样本的分类准确率不能满足实际需求。

支持向量机 (Support vector machine, SVM) 是一种经典的机器学习算法，具有坚实的统计学习理论基础[8−12]，为了探究其在不平衡数据中的分类性能，Lin 等[10] 建立了支持向量机和贝叶斯决策理论之间的关系，在贝叶斯决策理论中，贝叶斯最优决策是最优分类决策[11]，他们从理论上证明了对于错分代价相同的类平衡样本，SVM 可在样本数量趋于无穷时逼近贝叶斯最优决策，但是对于不平衡数据，SVM 无法逼近贝叶斯最优决策，即分类性能差。

代价敏感支持向量机 (Cost sensitive support vector machine, CS-SVM) 由 SVM 结合代价敏感技术发展而来，主要用来解决不平衡分类问题[11−12]。不平衡分类问题与代价敏感学习密切相关，在代价敏感学习中每个类的错分代价不同，不平衡分类问题中，少数类往往具有更高的错分代价[7,13]，对于错分代价不同的类不平衡样本，CS-SVM 理论上在样本数量趋于无穷大时同样可以逼近贝叶斯最优决策[10]。然而实际中的样本数量往往有限，导致 CS-SVM 的分类性能总是次优的。

针对 CS-SVM 在有限不平衡样本中难以逼近贝叶斯最优决策的问题，本文提出了一种基于能量统计法的 En-SVM 算法。利用能量距离量化少数类样本在不完全采样过程中的信息损失，使得少数类样本在再生核希尔伯特空间 (Reproducing kernel Hilbert space, RKHS) 中可以为机器学习算法提供更多的分类信息，提高少数类样本的分类精度。实验结果表明，该算法能够有效地处理不平衡水声数据，同时获得高检测概率及较低的虚警概率，并且随着不平衡比率的增加，仍能保持良好的性能。

1 CS-SVM 的贝叶斯最优决策

1.1 贝叶斯最优决策

水声目标-杂波分类是典型的二分类问题，不失一般性，做如下约定，$(\boldsymbol{X}, Y)$ 代表原始数据空间，$\boldsymbol{X} \in \mathbf{R}^d$，$Y \in \{-1, +1\}$，$(\boldsymbol{X}_s, Y_s)$ 为样本空间，$\boldsymbol{X}_s \in \mathbf{R}^d$，$Y_s \in \{-1, +1\}$，$d$ 表示数据维数，“$Y_s = -1$”代表负样本，“$Y_s = +1$”代表正样本，正

样本为少数类样本，具有更高的错分代价，对应水声目标。则来自 $(\boldsymbol{X}, Y)$ 的某一数据分为正类的贝叶斯后验概率为 $p(\boldsymbol{x}) = \Pr(Y = +1|\boldsymbol{X} = x)$，如式 (1) 所示：

$$p(\boldsymbol{x}) = \frac{k^+ \Pr(\boldsymbol{X} = x|Y = +1)}{k^+ \Pr(\boldsymbol{X} = x|Y = +1) + k^- \Pr(\boldsymbol{X} = x|Y = -1)}, \tag{1}$$

其中，k^+ 和 k^- 分别为原始数据中正负样本的分布概率，$\Pr(\boldsymbol{X} = x|Y = +1)$ 为正样本条件概率，$\Pr(\boldsymbol{X} = x|Y = -1)$ 为负样本条件概率，对于样本空间也有类似的表述。在分类过程中，正类 (正样本) 和负类 (负样本) 具有不同的错分代价，可用代价矩阵表示，如表 1 所示。

表 1　代价矩阵

Table 1 Cost matrix

	负类预测值	正类预测值
负类真实值	0	C^+(FP)
正类真实值	C^- (FN)	0

表 1 中 C^- 为假负例 (False negative instance, FN) 的错分代价，C^+ 为假正例 (False positive instance, FP) 的错分代价。机器学习数据集的建立是对原始数据空间的不完全随机采样过程，正样本和负样本的采样数量并非总是相同的，且正样本和负样本的重要性是不同的，比如具有不同错分代价的不平衡样本。Lin 等 [10] 通过贝叶斯决策理论证明了在有偏采样和错分代价不同的条件下，机器学习算法在原始数据空间和样本空间中的贝叶斯最优决策存在差异。最高的分类准确率在统计意义上对应最小贝叶斯风险：

$$E\{C^+ [1 - p(\boldsymbol{x})] I(\phi(\boldsymbol{x}) = 1) + C^- p(\boldsymbol{x}) I(\phi(\boldsymbol{x}) = -1)\}, \tag{2}$$

其中，$I(\cdot)$ 为指示函数，条件为真，$I(\cdot) = 1$，否则为 0。使得式 (2) 最小的 $\phi_B(\boldsymbol{x})$ 即为贝叶斯最优准则：

$$\phi_B(\boldsymbol{x}) = \begin{cases} +1, & \dfrac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} > \dfrac{C^+}{C^-}, \\ -1, & \text{else.} \end{cases} \tag{3}$$

在原始数据空间中正类与负类满足独立同分布 (Independent and identically distributed, IID) 条件，此时错分代价趋于相等，可得贝叶斯最优决策：

$$\hat{\phi}_B(\boldsymbol{x}) = \text{sign}[p(\boldsymbol{x}) - 1/2], \tag{4}$$

式 (4) 中，sign(·) 为符号函数，然而对具有不同错分代价的不平衡样本 $(\boldsymbol{X}_s, Y_s)$，贝叶斯最优准则为

$$\phi'_B(\boldsymbol{x}) = \begin{cases} +1, & \dfrac{p_s(\boldsymbol{x})}{1 - p_s(\boldsymbol{x})} > \dfrac{C^+ k_s^+ k^-}{C^- k_s^- k^+}, \\ -1, & \text{else.} \end{cases} \tag{5}$$

贝叶斯最优决策变为

$$\hat{\phi}'_B(\boldsymbol{x}) = \text{sign}\left[p_s(\boldsymbol{x}) - \frac{C^+ k_s^+ k^-}{C^+ k_s^+ k^- + C^- k_s^- k^+}\right]. \tag{6}$$

由式 (4) 和式 (6) 可知，在原始数据空间中，后验概率 $p(\boldsymbol{x})$ 只需和 1/2 比较，而在有偏采样和错分代价不同的样本空间中，后验概率 $p_s(\cdot)$ 和 1/2 比较会产生不准确的结果。因此，对于具有不同错分代价的不平衡样本，为了获得良好的分类效果，需要考虑贝叶斯最优决策 $\hat{\phi}'_B(\boldsymbol{x})$。

1.2　代价敏感支持向量机

对于不平衡样本，负类样本主导整体分类准确率，超平面会向正类样本偏移，导致具有更高错分代价的正类样本分类准确率下降，而整体准确率很高。CS-SVM 通过给少数类样本和多数类样本赋予不同的错分代价来处理不平衡样本，它的求解等价于在再生核希尔伯特空间 (RKHS)H_k 中求解关于目标函数的正则问题，决策函数可写为

$$f(\boldsymbol{x}) = h(\boldsymbol{x}) + \gamma,\ h \in \mathrm{H}_k,\ \gamma \in \mathbf{R}. \tag{7}$$

Zhang 证明了 Hinge 损失在 SVM 的求解中具有贝叶斯一致性 (Bayesian consistency)，因此，Hinge 损失常作为 SVM 的目标函数 [14]。在 SVM 的基础上，CS-SVM 引入了调节因子 $L(\cdot)$，如式 (8) 所示：

$$\min_f \frac{1}{\ell}\left\{\sum_{i=1}^{\ell} L(y_i)[1 - y_i f(\boldsymbol{x}_i)]_+\right\} + \lambda\|h\|^2_{\mathrm{H}_k},$$
$$\text{s.t. } y_i f(\boldsymbol{x}_i) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ \forall\, i = 1, \cdots, \ell, \tag{8}$$

其中，$L(-1) = C^+ k_s^+ k^-$，$L(+1) = C^- k_s^- k^+$，$\xi_i = [1 - y_i f(\boldsymbol{x}_i)]_+ = \max\{0,\ 1 - y_i f(\boldsymbol{x}_i)\}$ 为 Hinge 损失。Lin 等 [10] 证明了 CS-SVM 对应最小贝叶斯风险 $E\left[L(Y_s)(1 - Y_s f(\boldsymbol{X}_s))_+\right]$ 的贝叶斯最优

决策为

$$\hat{f} \xrightarrow{\ell\to\infty} \operatorname{sign}\left(p_s - \frac{L(-1)}{L(-1)+L(+1)}\right). \tag{9}$$

需要注意的是，SVM的标准输出为置信度$\hat{f}(\boldsymbol{x})$，经过Sigmoid函数映射得到后验概率p_s[15]。式(9)说明了对于具有不同错分代价的不平衡样本，CS-SVM 是贝叶斯最优的。但是，实际样本总是有限的，在独立同分布的采样过程中，k^+和k^-接近，而对于有限不平衡样本，将k_s^+和k_s^-视为先验概率是不合适的，因为在采样过程中正类样本存在信息损失，比如主动声呐探测过程中，受混响、多径效应等因素影响，目标回波往往会发生畸变并伴有能量损失，导致目标探测数据稀少。因此，正负样本的信息不对称使得式(9)有如下的修正：

$$\hat{f} \xrightarrow{\ell} \operatorname{sign}\left(p_s - \frac{1}{1+\dfrac{C^-}{C^+}\dfrac{k_s^-}{k_s^+}f_H\left(H_{\text{shannon}}\right)}\right), \tag{10}$$

其中，H_{shannon}代表正类样本采样过程中丢失的信息，用香农熵来表示，$f_H(\cdot)$为其度量准则。基于这一思想，本文提出了改进的CS-SVM。

2 基于能量统计方法的En-SVM

2.1 信息损失度量

根据拉格朗日对偶性，式(8)的对偶问题如下：

$$\min_{\alpha}\frac{1}{4\ell\lambda}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j K\left(\boldsymbol{x}_i,\ \boldsymbol{x}_j\right)-\sum_{i=1}^{\ell}\alpha_i,$$

$$\text{s.t.}\sum_{i=1}^{\ell}\alpha_i y_i=0,\ \ 0\leqslant\alpha_i\leqslant L(y_i),$$

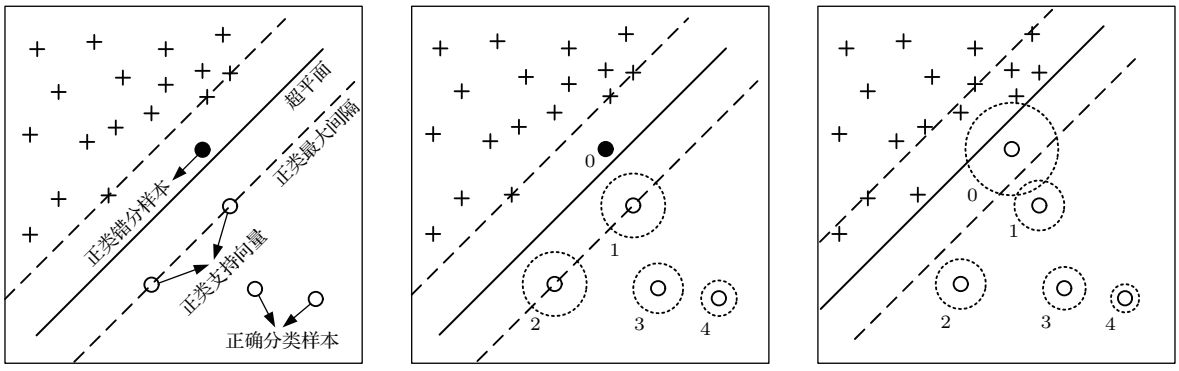
$$\forall\, i=1,\cdots,\ell, \tag{11}$$

式(11)中，$K(\cdot)$为核函数，可将非线性数据映射为希尔伯特空间中的线性数据，因此，在RKHS中认为正负样本线性可分，满足$0<\alpha_i<L(1)$的正样本即为正类支持向量，分类示意图如图1所示。

从图1(b)可知，正类支持向量的后验概率较小，具有较大的自信息(虚线同心圆表示)，含有更多的分类信息，自信息的期望即为香农熵，用来度量样本整体的信息，可以发现多数类样本整体包含的信息大于少数类样本，导致CS-SVM仍有错分的正类样本。En-SVM利用$f_H(H_{\text{shannon}})$，可使分类结果对正类样本更加有利，如图1(c)所示，“0”号错分样本获得了一定的置信度。能量统计方法通过计算特征函数间的加权平方距离来表征不同分布之间的差异[16]，少数类样本经原始数据空间不完全采样得到，存在信息损失H_{shannon}，本质上是其概率分布发生了变化，因此，可以用分布差异来度量信息损失，得到$f_H(H_{\text{shannon}})$近似解。能量距离表示如下：

$$D_E\left(p,\ p'\right)=\int_{\mathbf{R}^d}\left\|\varphi_p(t)-\varphi_{p'}(t)\right\|^2 \times\left[\frac{\pi^{\frac{d+1}{2}}}{\Gamma\left(\dfrac{d+1}{2}\right)}\|t\|^{d+1}\right]^{-1}\mathrm{d}t, \tag{12}$$

式(12)中，p和p'分别表示有限样本和原始数据的概率分布，$\varphi(\cdot)$为其对应的特征函数，对于不同的概率分布，特征函数总是存在且收敛的，$\|\cdot\|$表示欧几里得范数，$\Gamma(\cdot)$为伽马函数，d表示特征向量$\boldsymbol{x}$的维数。能量距离可以等效地表示为



(a) CS-SVM分类示意图　(b) 正类样本自信息　(c) En-SVM分类示意图

图1　RKHS中的不平衡分类

Fig. 1　Imbalance classification in RKHS

$$
\begin{aligned}
&D_E(p,p')\\
=&2E_{\boldsymbol{x}\sim p,\boldsymbol{x}'\sim p'}\|\boldsymbol{x}-\boldsymbol{x}'\|-E_{\boldsymbol{x},\bar{\boldsymbol{x}}\sim p}\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|\\
&-E_{\boldsymbol{x}',\bar{\boldsymbol{x}}'\sim p'}\|\boldsymbol{x}-\boldsymbol{x}'\|\\
=&2\boldsymbol{k}^{\mathrm{T}}\boldsymbol{\sigma}-\boldsymbol{k}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{k}+c, \qquad (13)
\end{aligned}
$$

式(13)中，$E_{\boldsymbol{x}\sim p}$表示服从概率密度p的期望，类别数只有两类时，$\boldsymbol{k}=[k^+,k^-]^{\mathrm{T}}$，$c$为与$\boldsymbol{k}$无关的常量，$\boldsymbol{A}$为$2\times 2$阶对称矩阵。对于少数类样本，$D_E(p,p')$可表示为一个相当于常量的$k^+$的函数：

$$J(k^+)=\mu(k^+)^2-2\sigma k^+, \qquad (14)$$

$$\mu=2A_{1,\,-1}-A_{1,\,1}-A_{1,\,-1}, \qquad (15)$$

$$\sigma=A_{1,\,-1}-A_{-1,\,-1}-\sigma_1+\sigma_{-1}, \qquad (16)$$

其中，μ为贝叶斯风险$D_E(p(\boldsymbol{x}|y=1),p(\boldsymbol{x}|y=-1))$，$A_{y,\bar{y}}$和$\sigma_y$可近似给出：

$$\hat{A}_{y,\,\bar{y}}=\frac{1}{n_y n_{\bar{y}}}\sum_{l:y_i=y}\sum_{\bar{l}:y_{\bar{i}}=\bar{y}}\|\boldsymbol{x}_i-\boldsymbol{x}_{\bar{i}}\|, \qquad (17)$$

$$\hat{\sigma}_y=\frac{1}{n' n_y}\sum_{i'=1}^{n'}\sum_{i:y_i=y}\|\boldsymbol{x}'_{i'}-\boldsymbol{x}_i\|. \qquad (18)$$

式(18)中原始数据$\{\boldsymbol{x}_{i'}\}_{i'=1}^{n'}$是未知的，但和有限样本和不平衡率$n_{\bar{y}}/n_y$存在关联，$\hat{\sigma}_y$可近似为

$$\hat{\sigma}_y=\frac{1}{n_{\bar{y}}}\sum_{j=1}^{n_y}\sum_{i:y_i=y}\|\boldsymbol{x}_j-\boldsymbol{x}_i\|. \qquad (19)$$

结合式(14)~(19)可得到信息损失度量：

$$f_H\left(H_{\mathrm{shannon}}\right)\approx J\left(k^+\right). \qquad (20)$$

2.2 En-SVM算法求解

En-SVM算法的核心在于利用少数类样本不完全采样过程的信息损失来补偿分类模型在训练过程中所需的分类信息，使得分类结果对少数类样本更加有利。记$f_H=f_H\left(H_{\mathrm{shannon}}\right)$，由此，可得En-SVM如下：

$$
\begin{aligned}
&\min_f\frac{1}{\ell}\left(\sum_{i=1}^{\ell}\left(f_H I\left(y_i=1\right)+I\left(y_i=-1\right)L\left(y_i\right)\xi_i\right)\right)\\
&\quad+\lambda\|h\|_{H_k}^2,\\
&\text{s.t.}\ y_i f(\boldsymbol{x}_i)\geqslant 1-\xi_i,\ \xi_i\geqslant 0,\ \forall i=1,2,\cdots,\ell. \qquad (21)
\end{aligned}
$$

RKHS理论保证了式(7)有如下的形式：

$$f_{\boldsymbol{\theta},\gamma}(\boldsymbol{x})\sum_{i=1}^{\ell}\theta_i K(\boldsymbol{x}_i,\boldsymbol{x})+\gamma. \qquad (22)$$

为了减少待优化参数的数量，需要利用拉格朗日对偶性得到原始问题式(21)的对偶问题[13]：

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha}}\frac{1}{2\ell\lambda}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{\ell}\alpha_i,\\
&\text{s.t.}0\leqslant\alpha_i\leqslant\left(f_H I(y_i=1)+I\left(y_i=-1\right)\right)L(y_i),\\
&\sum_{i=1}^{\ell}\alpha_i y_i=0,\quad\forall\, i=1,2,\cdots,\ell. \qquad (23)
\end{aligned}
$$

式(23)中，$\boldsymbol{\alpha}$为对偶解，则原始问题的解为

$$\boldsymbol{\theta}=\frac{1}{2\ell\lambda}\left[y_1\alpha_1,\ \cdots,\ y_\ell\alpha_\ell\right]^{\mathrm{T}}. \qquad (24)$$

选取一个满足$0<\alpha_i<(f_H I(y_i=1)+I(y_i=-1))L(y_i)$的样本，则根据KKT条件(Karush-Kuhn-Tucker condition)可得

$$
\gamma=\frac{\displaystyle\sum_{i=1}^{\ell}\alpha_i\left(L\left(y_i\right)f_H-\alpha_i\right)\left[y_i-\sum_{j=1}^{\ell}\boldsymbol{\theta}_j K\left(\boldsymbol{x},\boldsymbol{x}_j\right)\right]}{\displaystyle\sum_{i=1}^{\ell}\alpha_i\left(L\left(y_i\right)f_H-\alpha_i\right)}. \qquad (25)
$$

代入式(24)和式(25)到式(22)，得到最终决策$\hat{f}_{\boldsymbol{\theta},\gamma}(\boldsymbol{x})$。

3 海试数据处理结果及分析

为验证本文算法，使用某海域的水下目标历史探测数据来构建目标-杂波数据集，由于数据集的样本量较小，为了能够得到有效的机器学习模型，采用“交叉验证(Cross validation)”方法来处理数据。

3.1 评价指标

对于类别不平衡数据，ROC曲线(Receiver operating characteristic curve)不易受到数据分布影响，是一种评价机器学习模型性能的常用方法[13]。ROC曲线以真正率(true positive rate)为横坐标，以假正率(False positive rate)为纵坐标，反映了检测概率和虚警概率之间的制约关系。ROC曲线下的面积被称为Auc(Area of under curve)值，值越大表明分类效果越好。

3.2 水声目标杂波数据集

不平衡样本中，多数类样本与少数类样本的数量之比称为不平衡率(Imbalanced rate, IR)，本文所采用数据集$(\boldsymbol{X}_s,Y_s)$的IR $\approx$ 245.3，数据维数为

11 (对应11类特征)，即 $\boldsymbol{X}_s \in \mathbf{R}^{11}$，$Y_s \in \{-1, +1\}$，“−1”代表杂波，“+1”代表目标，为少数类样本。在该数据集上做10 次3折交叉验证[13]，即每一次交叉验证前分别将杂波和目标样本随机等分为3份(每一份称为一折)，即Data1、Data2和Data3，如表2所示，并形成3组训练集和测试集：(1) 训练集Data1 + Data2，测试集Data3；(2) 训练集Data1 + Data3，测试集Data2；(3) 训练集Data2 + Data3，测试集Data1。

分别在(1)、(2)和(3)上训练并测试，重复进行10次，以减小实验过程中的随机性。

表2　水声目标-杂波样本

Table 2　Underwater acoustic target-clutter sample

	Data1	Data2	Data3	总计
杂波数	8586	8586	8586	25758
目标数	35	35	35	105
总计	8621	8621	8621	25863

3.3　实验结果及分析

为便于比较，标准SVM、CS-SVM和本文算法En-SVM均采用径向基核函数，核自由参数δ取1，采用序列最小最优化(Sequential minimal optimization, SMO)算法，由于涉及样本间距离的计算，为防止受到具有过高特征值或过低特征值样本的影响，输入数据均做标准化处理。CS-SVM和En-SVM中的假负例FN与假正例FP的代价之比C^-/C^+取和IR相同的值，算法在表2所示的数据集上做10次3折交叉验证。

(1) 算法性能比较

为了有效比较标准SVM、CS-SVM和En-SVM在贝叶斯最优准则(式(5))下的性能，始终以0.5 作为概率决策门限，即算法输出的后验概率大于0.5时，该样本$(\boldsymbol{x}, y)$被判断为目标，否则为杂波。为了保证实验结果的可靠性，按照10次3折交叉验证的方式进行，统计每次每折的分类后验概率预测值绘制ROC曲线并通过梯度法计算Auc值，有效地消除了ROC 曲线中的“锯齿”，使得固定门限下的数值更加准确。

依照图例顺序，图2中所示曲线分别表示SVM、CS-SVM和En-SVM算法的ROC性能，其中，每条曲线上会标记一个与曲线同色的实心点，该点表示决策门限值为0.5时，算法能够达到的检测概率和虚警概率。为了使得算法输出结果具有一定的统计意义和可信度，本文将机器学习算法的输出通过Sigmoid函数统一映射为正样本(目标)的后验概率值，即未知样本数据是目标的可能性，后验概率值越大，是目标的可能性越大。对于一条ROC性能曲线，当取不同的后验概率值作为决策门限时，该门限将对应一组不同的检测概率和虚警概率，为了防止人为的先验知识对结果产生干扰，同时，为了使得不同算法具有相同的衡量标准，本文选取了概率值为0.5处作为决策门限，大于0.5，则该未知样本数据就是目标，否则是杂波，实现了从统计意义上的可能性向确定性决策的转变。Auc值说明了ROC性能曲线接近左上角的程度，而实心点处对应的检测概率和虚警概率则进一步说明了算法在统计意义上的优劣。

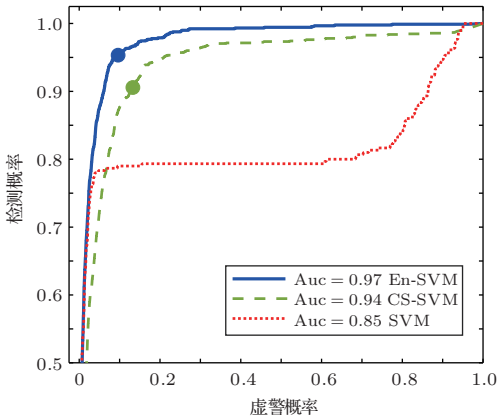


图2　算法性能比较

Fig. 2　Algorithm performance comparison

可以看到图2中SVM的ROC性能曲线上没有出现实心点，原因在于其实心点对应的检测概率小于50%，一般更加关注检测概率大于90%时对应的虚警概率，为了便于观察不同算法性能曲线的差异，图2中仅绘制出了检测概率大于50%的部分。SVM算法的Auc值低于CS-SVM和En-SVM，且检测概率低于50%，分类性能差。相较于CS-SVM算法，本文算法En-SVM的Auc值高出0.03，并且固定决策门限下的性能更靠近左上角，虚警概率降低了3.4个百分点，检测概率提高了5个百分点，分别达到了9.9%和95.6%，分类性能优于CS-SVM，即En-SVM算法在获得高检测概率时，可以排除约90.1%的杂

波。实验结果表明，对于不平衡数据的分类问题，本文算法 En-SVM 因为考虑了少数类样本不完全采样过程中的信息损失，而具有更好的分类性能，更加接近贝叶斯最优决策 (式 (9))。

(2) 数据集不平衡率对算法的影响

本文算法 En-SVM 的核心思想在于度量原始数据空间 $(\boldsymbol{X}, Y)$ 和样本空间 $(\boldsymbol{X}_s, Y_s)$ 正类样本分布的能量距离来量化正类样本不完全采样过程中的信息损失，来补偿 CS-SVM 在 RKHS 中正类样本的香农熵，使得正类样本能在分类过程中为算法提供更强的分类信息，从而使 En-SVM 能够在有限样本中逼近贝叶斯最优决策，获得更好的分类性能。为了进一步验证算法效果，将数据集 (表 2 所示) 中的目标数量 (+1 表示) 依次从 105 随机下采样为 90、60、30，对应的不平衡率 IR 从 245.3 变为 286.2、429.3 和 858.6，统计 10 次 3 折交叉的 Auc 值，以“均值 ± 标准差”的形式给出，并得到对应的 ROC 曲线。

由表 3 可以看出，随着 IR 的增大，标准 SVM 的性能明显下降，CS-SVM 性能也有所下降，而 En-SVM 的性能保持稳定，Auc 值高于其他两者。

表 3　不平衡率对 Auc 值的影响

Table 3　Effect of unbalance rate on Auc value

IR	SVM	CS-SVM	En-SVM
245.3	0.85 ± 0.16	0.94 ± 0.08	$\mathbf{0.97} \pm 0.07$
286.2	0.85 ± 0.14	0.95 ± 0.20	$\mathbf{0.97} \pm 0.06$
429.3	0.78 ± 0.20	0.90 ± 0.04	$\mathbf{0.97} \pm 0.15$
858.6	0.56 ± 0.11	0.86 ± 0.15	$\mathbf{0.95} \pm 0.06$

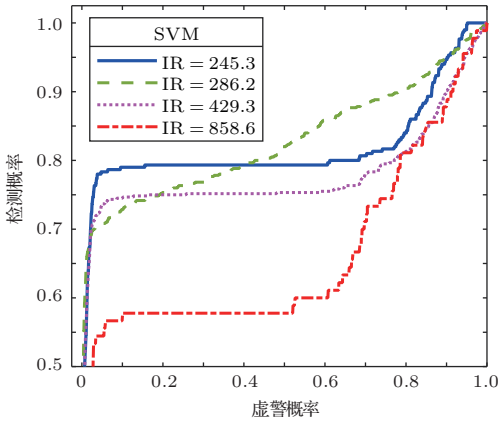


图 3　标准 SVM 的 ROC 曲线

Fig. 3　ROC of Standard SVM

图 3~5 分别为 SVM、CS-SVM 和 En-SVM 在不同 IR 数据下得到的 ROC 曲线，可以看出，随着数据 IR 的增大，En-SVM 能够保持良好的性能，且 0.5 决策门限下的性能波动程度比 SVM 和 CS-SVM 小。实验结果表明，En-SVM 能够充分利用少数类样本不完全采样过程中的信息损失，提升算法性能，并具有一定的稳定性。

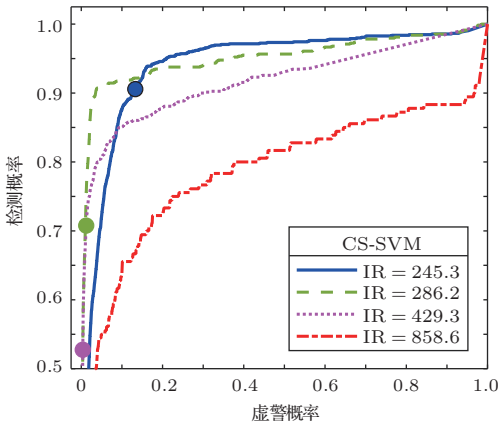


图 4　CS-SVM 的 ROC 曲线

Fig. 4　ROC of CS-SVM

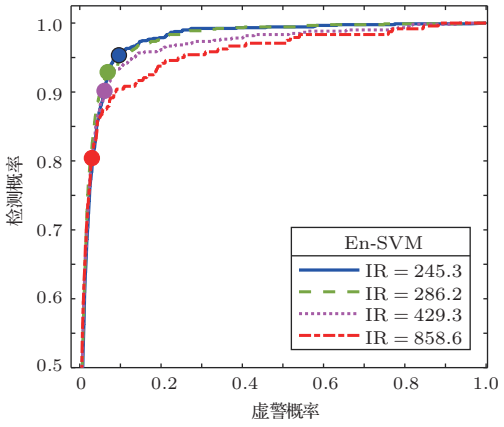


图 5　En-SVM 的 ROC 曲线

Fig. 5　ROC of En-SVM

4　结论

本文针对少数类样本在不完全采样过程中存在信息损失，结合能量统计法提出了 En-SVM 算法，在处理水声目标-杂波不平衡数据中有着良好的分类效果。实际海试数据的处理结果表明，En-SVM 算法能够在有限样本中更加逼近贝叶斯最优决策，并且对样本的不平衡率变化不敏感，验证了算法的有效性和稳定性。本文采用的水声数据集建立在高

于最小可检测阈6 dB的数据上，未来将进一步研究该算法在更低可检测信噪比数据集上的不平衡分类效果。

参 考 文 献

[1] Berg H, Hjelmervik K T. Classification of anti-submarine warfare sonar targets using a deep neural network[C]. OCEANS Marine Technology Society. IEEE Charleston, 2018: 1–5.

[2] Berg H, Hjelmervik K T. A comparison of different machine learning algorithms for automatic classification of sonar targets[C]. OCEANS Marine Technology Society. IEEE Monterey, 2016: 1–8.

[3] Stender D H, Hjelmervik K T, Berg H, et al. Sensitivity to target behavior in automatic classification on kinematic track features[C]. OCEANS Marine Technology Society. IEEE Kobe, 2018: 1–5.

[4] Stender D H, Hjelmervik K T, Berg H, et al. The classification performance of signal-to-noise ratio and kinematic features in varying environments[C]. OCEANS. IEEE Aberdeen, 2017: 1–5.

[5] 赵楠, 张小芳, 张利军. 不平衡数据分类研究综述 [J]. 计算机科学, 2018, 45(6A): 22–27.

[6] Liu Z, Cao W, Gao Z, et al. Self-paced ensemble for Highly imbalanced massive data classification[C]. IEEE 36th International Conference on Data Engineering. IEEE Computer Society, 2020: 841–851.

[7] Han J, Kamber M, Pei J. 数据挖掘: 概念与技术 [M]. 范明, 孟小峰, 译. 第三版. 北京: 机械工业出版社, 2019: 250–251.

[8] Núez H, Gonzalez-Abril L, Angulo C. Improving SVM classification on imbalanced datasets by introducing a new bias[J]. Journal of Classification, 2017, 34(3): 427–443.

[9] Vanhoeyveld J, Martens D. Imbalanced classification in sparse and large behaviour datasets[J]. Data Mining & Knowledge Discovery, 2018, 32(1): 1–58.

[10] Lin Y, Yoonkyung L, Grace W. Support vector machines for classification in nonstandard situations[J]. Machine Learning, 2002, 46(1/2/3): 191–202.

[11] Zheng E, Li P, Song Z. Cost sensitive support vector machines[J]. Control & Decision, 2006, 21(4): 473–476.

[12] Liu N, Qi E, Xu M, et al. A novel intelligent classification model for breast cancer diagnosis[J]. Information Processing & Management, 2019, 56(3): 609–623.

[13] 于化龙. 类别不平衡学习: 理论与算法 [M]. 北京: 清华大学出版社, 2017: 4–5.

[14] Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization[J]. The Annals of Statistics, 2004, 32(1): 56–85.

[15] Lin H T, Lin C J, Weng R C. A note on Platt's probabilistic outputs for support vector machines[J]. Machine Learning, 2007, 68(3): 267–276.

[16] Székely G J, Rizzo M L. Energy statistics: a class of statistics based on distances[J]. Journal of Statistical Planning and Inference, 2013, 143(8): 1249–1272.