DOI: 10.13718/j.cnki.xdzk.2018.12.025

基于凸壳的在线单类学习机①

周国华[1,2]，　申燕萍[1]，　殷新春[2]

1. 常州轻工职业技术学院 信息工程系，江苏 常州 213164；2. 扬州大学 信息工程学院，江苏 扬州 225127

摘要：传统的基于支持向量机的单类分类器因计算复杂度高而无法满足大规模数据实时处理的需求，在线学习方法为解决该问题提供了一种有效途径. 本文在挖掘样本数据在特征空间分布性状的基础上，提出了一种基于凸壳的在线单类学习机(One-class Online Classifier based on Convex Hull，OOCCH). 该方法首先使用凸壳的定义选择能代表特征空间中数据分布的凸壳向量对应的原始样本作为训练样本来缩减训练集的规模；其次在分类器在线更新阶段利用凸壳向量动态地调整分类器的训练样本. 理论分析证明了 OOCCH 的有效性，与现有的在线单类分类器的实验比较，OOCCH 在训练时间和分类性能方面有显著优势.

关　键　词：在线学习；单类；分类；凸壳

中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：1673-9868(2018)12-0163-10

机器学习是人工智能中最具智能特性和最前沿的研究分支之一，它利用经验产生学习模型以提高系统的性能. 机器学习在机器故障诊断、网络流量监测和垃圾邮件分类等领域应用广泛，但这些应用领域的数据常存在类别样本数严重不平衡的情况，如机器故障诊断中正常样本占绝大多数，而故障样本往往稀缺，类型多样且获取代价极高. 机器学习中的单类分类器可以很好地适应这种分类场景，单类分类器通过对一个类别的样本进行学习形成对该类别的数据描述，从而根据设计的相似性度量和阈值来判别新样本的归属[1]. 根据分类原理，单类分类器可以分成 4 类：基于密度估计的算法[2-3]、基于聚类的算法[4-5]、基于神经网络的算法[6-7]和基于支持向量机(SVM)的算法[8-9]. 但这 4 类单类分类器皆存在一定的缺陷：基于密度估计的单类分类器不适用于高维数据场景；基于聚类的单类分类器对聚类的相似性度量不易定义且对噪声敏感；而基于神经网络和 SVM 的算法因其复杂度较高不适用于大规模数据的场景.

随着云计算、传感器等新兴技术的快速发展，产生越来越多的大规模数据，这些大规模数据还同时具有实时性、动态性和突发性等特征[10-11]. 传统单类分类器训练时将数据作为整体进行批量学习，不适应海量数据和流式数据. 在线学习是面对该挑战的有效解决方法之一. 在线学习首先使用部分数据建立一个初始模型，然后按时序一次处理一个或者一小批数据来更新模型，从而有效降低训练模型的复杂度，同时也可以保留数据的最新信息[12]. 根据模型是线性还是非线性，在线学习方法可分为 2 类：第一类是线性在线学习方法，如基于感知器的在线学习算法[13]和稀疏在线学习算法[14]等；第二类是以使用核技术为代表的非线性学习方法，如基于 SVM 的在线学习算法[15]等. SVM 借助统计学习理论和最优化方法解决机器学习的问题，并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. 在处理非线性不可分的数据分类时，SVM 通常采用核化的方法将原始数据映射到高维核空间中，且核化后的最优化问题常描述成二次规划问题. 例如 Wang T. 等人[9]提出了稀疏化的在线最小二乘 SVM 算法；Krell M. M. 等人[15]结合特征

① 收稿日期：2018-06-02
基金项目：国家自然科学基金资助项目(61472343).
作者简介：周国华(1977-)，男，讲师，硕士，主要从事智能学习、模式识别等方面的研究.

选择提出了在线学习的单类学习算法，用于解决不平衡数据的分类问题.

针对上述问题，本文提出了一种适用于大规模数据环境下的基于凸壳的在线单类学习机(OOCCH). OOCCH做了以下4个方面的尝试：

1) OOCCH基于凸壳技术得到特征空间中能代表数据分布信息的凸壳向量，并使用凸壳向量对应的原始样本作为训练数据，在分类精度相当的条件下，比传统的基于SVM在线学习方法训练速度快且高效；

2) OOCCH基于凸壳的定义动态地调整分类器的训练样本，以对后续样本进行在线学习，保证数据的完整性；

3) 从理论上证明凸壳向量对应的原始样本作为训练数据的分类器在精度上是有保证的；

4) 对比4种不同的单类分类在线学习算法，验证OOCCH在保证分类精度的同时能显著提升大规模数据在线分类的实时性，是一种处理大规模数据在线学习的有效工具.

1 单类SVM

单类SVM主要分为基于超球法和基于超平面法这2类. 前者确定一个能够包含所有训练样本的体积最小化的超球，后者构建一个超平面来描述数据集中正常的样本. 这里介绍基于超平面法的单类SVM，假设有N个样本的数据集$\boldsymbol{X}=\{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in \boldsymbol{R}^d, i=1, 2, \cdots, N\}$，特征映射$\varphi$将样本$\boldsymbol{x}$映射到特征空间$\boldsymbol{F}$中，超平面法的单类SVM可以表示为如下的优化问题：

$$\min_{w,\xi,\rho} \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}-\rho+\frac{1}{\nu N}\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } \rho-\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)-\xi_i\leqslant 0,\ \xi_i\geqslant 0 \qquad i=1,2,\cdots,N \tag{1}$$

其中，$\boldsymbol{w}\in\boldsymbol{R}^d$是超平面的法向量，$\rho\in\boldsymbol{R}$是超平面的偏置量，$\nu$为一常数. 其对偶形式可表示为

$$\sum_{i,j=1}^{N}\alpha_i\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

$$\text{s.t. } \sum_{i=1}^{N}\alpha_i=1,\ 0\leqslant\alpha_i\leqslant\frac{1}{\nu l} \qquad i=1,2,\cdots,N \tag{2}$$

单类SVM的决策函数为

$$f(\boldsymbol{x})=\mathrm{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-\rho) \tag{3}$$

对于测试样本，如果$f(\boldsymbol{x})=1$，则判定该样本为正常类；如果$f(\boldsymbol{x})=-1$，则判定该样本为例外点. 由式(2)可以看出单类SVM的计算复杂度为$O(N^3)$.

2 基于凸壳的在线单类分类器

2.1 基本思想

单类SVM的分类本质可以表现为寻找一个能最大化与原点之间距离的超平面，并根据训练数据的几何分布，累积求出支持向量. 因为只有在目标类的训练数据中，单类SVM将原点虚设为非目标类唯一的样本. 受这一思想启发，本文认为构建单类SVM分类器仅需使用能代表训练数据轮廓范围的样本而无需使用全部的训练样本，并且删除轮廓内部样本对构建分类器没有影响. 凸壳是计算几何中的概念，对于给定的集合$\boldsymbol{X}$，凸壳是包含所有样本的最小凸集，它由凸壳顶点构成，样本集$\boldsymbol{X}$内所有的样本都可用凸壳向量的线性组合来表示[16]，即

$$\boldsymbol{x}_i=\sum_{\boldsymbol{x}_t\in\boldsymbol{X}^*}\mu_{i,t}\boldsymbol{x}_t \tag{4}$$

其中$\boldsymbol{X}^*$是$\boldsymbol{X}$的凸壳向量集，$\sum_{\boldsymbol{x}_t\in\boldsymbol{X}^*}\mu_{i,t}=1$且$\mu_{i,t}\geqslant 0$.

本文所提的在线单类分类器首先计算得到训练集在特征空间的凸壳向量，在构建初始分类器时将这些凸壳向量对应的原始样本作为训练数据. 在线学习阶段，在接受到新样本后对样本进行分类，同时辨别该

样本是否是特征空间的凸壳向量，如果是则将其对应的原始样本加入到训练数据集中，并更新分类器；如果不是则不更新分类器.

2.2　OOCCH 算法描述

OOCCH 在线分类器的构建可以分为 3 个阶段：1）选择初始训练集在特征空间的凸壳向量；2）训练初始分类器；3）在线分类器的更新. 下面详细介绍 OOCCH 在线分类器的 3 个阶段.

第 1 阶段，选择初始训练集在特征空间的凸壳向量. OOCCH 这一阶段的工作可分为 4 个步骤：1）使用支持向量数据描述(Support Vector Data Description，SVDD)[17] 得到特征空间下能够包含初始训练集样本的体积最小化的超球，并设超球球面上的向量是特征空间下的凸壳集的初始集 $\boldsymbol{M}^*$；2）对得到的特征向量按照其到超球球心的距离进行降序排序，形成特征数据集 $\varphi(\boldsymbol{X}^*)$；3）计算当前凸壳集的向量权重 λ 的值，即：

$$\begin{aligned}&\min_{\lambda} \| \varphi(\boldsymbol{x}_i) - \sum_{t=1}^{|\boldsymbol{M}^*|} \lambda_{i,t}\varphi(\boldsymbol{x}_t) \|^2 \\ &\text{s.t.} \sum_{t=1}^{|\boldsymbol{M}^*|} \lambda_{i,t} = 1,\ 0 \leqslant \lambda_{i,t} \leqslant 1\end{aligned} \tag{5}$$

其中，$|\boldsymbol{M}^*|$ 表示 $\boldsymbol{M}^*$ 中样本的个数. 将式(5)的常数项舍弃后，式(5)的求解可以转化为以下的二次规划形式：

$$\begin{aligned}&\min_{\lambda} 2\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{M}^*\lambda + \lambda^{\mathrm{T}}\boldsymbol{M}^{*\mathrm{T}}\boldsymbol{M}^*\lambda \\ &\text{s.t.} \sum_{t=1}^{|\boldsymbol{M}^*|} \lambda_{i,t} = 1,\ 0 \leqslant \lambda_{i,t} \leqslant 1\end{aligned} \tag{6}$$

4）根据排序结果依次判断非初始集 $\boldsymbol{M}^*$ 中的每个向量 $\varphi(\boldsymbol{x}_i)$ 是否是凸壳向量，其中 $\boldsymbol{x}_i \in \boldsymbol{X}^*$ 且 $\varphi(\boldsymbol{x}_i) \notin \boldsymbol{M}^*$，即

$$\| \varphi(\boldsymbol{x}_i) - \sum_{t=1}^{|\boldsymbol{M}^*|} \lambda_{i,t}\varphi(\boldsymbol{x}_t) \|^2 \leqslant \mu \tag{7}$$

若式(7)的计算结果小于阈值 μ，则表示 $\varphi(\boldsymbol{x}_i)$ 能用当前的凸壳向量线性表示，说明 $\varphi(\boldsymbol{x}_i)$ 不是凸壳向量；反之，若式(7)的计算结果大于阈值 μ，说明 $\varphi(\boldsymbol{x}_i)$ 是凸壳向量，将其并入当前的凸壳向量集 $\boldsymbol{M}^*$ 中，即 $\boldsymbol{M}^* = \boldsymbol{M}^* \cup \varphi(\boldsymbol{x}_i)$. 同时，通过上式也可以得到：

$$\varphi(\boldsymbol{x}_i) = \sum_{\varphi(\boldsymbol{x}_t)\in\boldsymbol{M}^*} \lambda_{i,t}\varphi(\boldsymbol{x}_t) + \tau_i \tag{8}$$

其中，$\| \tau_i \|^2 \leqslant \mu$，$i = |\boldsymbol{M}^*| + 1, |\boldsymbol{M}^*| + 2, \cdots, N$.

第 2 阶段，训练初始分类器. 将第 1 阶段得到的凸壳向量集 $\boldsymbol{M}^*$ 中对应的原始数据作为训练样本 $\boldsymbol{X}_{\text{train}}$ 代入式(2)，完成 OOCCH 初始分类器的训练.

第 3 阶段，在线分类器的更新. 传统的在线学习算法中，常采用两类方法更新分类器，一是将新样本直接加入到训练集中重新训练分类器；二是将新样本与上一阶段构成分类器的支持向量合并组成新的训练集，重新训练分类器. 但第一类方法会随着训练集容量的增长而使分类器的训练效率急剧下降，不适用于大数据环境的实时分类要求. 第二类方法保留在线分类器的支持向量作为训练数据，但支持向量仅与分类器构建的分类面相关，不能很好地代表数据在特征空间的轮廓分布. 当新到达的训练数据与历史训练数据的分布存在变化时，这一类方法的分类效果不佳.

OOCCH 在这一阶段使用式(3)判断新到达的一个(一批)样本 $\boldsymbol{x}_{\text{new}}$ 的类别，如果是目标类样本，则将 $\boldsymbol{X}_{\text{train}}$ 代入以下函数判断其是否是候选凸壳向量：

$$f'(\boldsymbol{x}_{\text{new}}) = \boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_{\text{new}}) - \rho \tag{9}$$

如果 $|f'(\boldsymbol{x}_{\text{new}})| \leqslant 1 + \beta$($\beta$ 是设定的很小的常数)，则将新样本加入到训练集 $\boldsymbol{X}_{\text{train}}$ 中，即

$$\boldsymbol{X}_{\text{train}} = \boldsymbol{X}_{\text{train}} \cup \{\boldsymbol{x}_{\text{new}}\}$$

如果$|f'(\boldsymbol{x}_{\text{new}})|>1+\beta$，则不更新训练集$\boldsymbol{X}_{\text{train}}$.

为了减少在线分类器更新阶段的运算量，OOCCH将分布在分类面边界周围的样本纳入到训练集中.但随着在线分类器的不断更新，训练集$\boldsymbol{X}_{\text{train}}$的容量也在不断增长.当$\boldsymbol{X}_{\text{train}}$的容量超过设定的阈值$K$时，OOCCH重新计算训练集的凸壳向量，将新得到的凸壳向量对应的原始样本作为当前的训练数据.

2.3 OOCCH精度和时间复杂度分析

本节首先从理论上证明使用凸壳向量对应的原始样本作为单类分类器的训练样本在精度上是有保证的.这里将式(2)改写成无约束目标函数，命名为$\boldsymbol{F}(\boldsymbol{w},\rho)$，则

$$\min_{\boldsymbol{w},\rho}\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{1}{\nu N}\sum_{i=1}^{N}l(\boldsymbol{w},\rho,\varphi(\boldsymbol{x}_i))-\rho \tag{10}$$

其中，$l(\boldsymbol{w},\rho,\varphi(\boldsymbol{x}_i))=\max\{0,\rho-\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)\}$，$i=1,2,\cdots,N$.

定理1 假设$(\boldsymbol{w}_1^*,\rho_1^*)$是使用全部训练集代入式(2)得到的单类在线分类器的最优解，$(\boldsymbol{w}_2^*,\rho_2^*)$是使用凸壳向量对应的原始样本代入式(2)得到的单类在线分类器的最优解，则

$$\boldsymbol{F}_1(\boldsymbol{w}_2^*,\rho_2^*)-\boldsymbol{F}_2(\boldsymbol{w}_2^*,\rho_2^*)\leqslant\frac{N-|\boldsymbol{M}^*|}{N\cdot\nu}\sqrt{\frac{\mu}{\nu}}$$

证

由

$$\boldsymbol{F}_1(\boldsymbol{w}_2^*,\rho_2^*)=\frac{1}{2}\|\boldsymbol{w}_2^*\|^2+\frac{1}{\nu N}\sum_{i=1}^{N}l(\boldsymbol{w}_2^*,\rho_2^*,\varphi(\boldsymbol{x}_i))-\rho_2^*$$

$$\boldsymbol{F}_2(\boldsymbol{w}_2^*,\rho_2^*)=\frac{1}{2}\|\boldsymbol{w}_2^*\|^2+\frac{1}{\nu N}\sum_{i=1}^{|\boldsymbol{M}^*|}l(\boldsymbol{w}_2^*,\rho_2^*,\varphi(\boldsymbol{x}_i))-\rho_2^*$$

可得

$$\boldsymbol{F}_1(\boldsymbol{w}_2^*,\rho_2^*)-\boldsymbol{F}_2(\boldsymbol{w}_2^*,\rho_2^*)=\frac{1}{N\cdot\nu}\sum_{i=|\boldsymbol{M}^*|+1}^{N}l(\boldsymbol{w}_2^*,\rho_2^*,\varphi(\boldsymbol{x}_i))$$

因为

$$\rho_2^*-\boldsymbol{w}_2^{*\mathrm{T}}\varphi(\boldsymbol{x}_i)\leqslant 0\qquad i=1,2,\cdots,|\boldsymbol{M}^*|$$

根据式(8)可得

$$\rho_2^*-\boldsymbol{w}_2^{*\mathrm{T}}\varphi(\boldsymbol{x}_i)=\rho_2^*-\boldsymbol{w}_2^{*\mathrm{T}}\Big(\sum_{\varphi(\boldsymbol{x}_t)\in\boldsymbol{M}^*}\lambda_{i,t}\varphi(\boldsymbol{x}_t)+\tau_i\Big)=$$

$$\rho_2^*-\boldsymbol{w}_2^{*\mathrm{T}}\sum_{\varphi(\boldsymbol{x}_t)\in\boldsymbol{M}^*}\lambda_{i,t}\varphi(\boldsymbol{x}_t)-\boldsymbol{w}_2^{*\mathrm{T}}\tau_i=$$

$$\sum_{\varphi(\boldsymbol{x}_t)\in\boldsymbol{M}^*}\lambda_{i,t}[\rho_2^*-\boldsymbol{w}_2^{*\mathrm{T}}\varphi(\boldsymbol{x}_t)]-\boldsymbol{w}_2^{*\mathrm{T}}\tau_i\leqslant-\boldsymbol{w}_2^{*\mathrm{T}}\tau_i\leqslant\|\boldsymbol{w}_2^*\|\,\|\tau_i\|$$

参照文献[18]，可得

$$\|\boldsymbol{w}_2^*\|\leqslant\sqrt{|\boldsymbol{M}^*|/(\nu\cdot N)}\leqslant\sqrt{1/\nu}$$

所以可得

$$\frac{1}{N\cdot\nu}\sum_{i=|\boldsymbol{M}^*|+1}^{N}l(\boldsymbol{w}_2^*,\rho_2^*,\varphi(\boldsymbol{x}_i))=\frac{1}{N\cdot\nu}\sum_{i=|\boldsymbol{M}^*|+1}^{N}(\rho_2^*-\boldsymbol{w}_2^{*\mathrm{T}}\varphi(\boldsymbol{x}_i))\leqslant$$

$$\frac{1}{N\cdot\nu}\sum_{i=|\boldsymbol{M}^*|+1}^{N}\|\boldsymbol{w}_2^*\|\,\|\tau_i\|\leqslant\frac{1}{N\cdot\nu}\sum_{i=|\boldsymbol{M}^*|+1}^{N}\sqrt{\frac{\mu}{\nu}}=\frac{N-|\boldsymbol{M}^*|}{N\cdot\nu}\sqrt{\frac{\mu}{\nu}}$$

因此得证.

定理2 假设

$$\boldsymbol{F}_1(\boldsymbol{w}_2^*,\rho_2^*)=\frac{1}{2}\|\boldsymbol{w}_2^*\|^2+\frac{1}{\nu N}\sum_{i=1}^{N}l(\boldsymbol{w}_2^*,\rho_2^*,\varphi(\boldsymbol{x}_i))-\rho_2^*$$

则

$$\boldsymbol{F}_1(\boldsymbol{w}_2^*, \rho_2^*) - \boldsymbol{F}_1(\boldsymbol{w}_1^*, \rho_1^*) \leqslant \frac{N - |\boldsymbol{M}^*|}{N \cdot \nu} \sqrt{\frac{\mu}{\nu}}$$

证

由

$$\boldsymbol{F}_2(\boldsymbol{w}_2^*, \rho_2^*) - \boldsymbol{F}_1(\boldsymbol{w}_1^*, \rho_1^*) \leqslant \boldsymbol{F}_2(\boldsymbol{w}_1^*, \rho_1^*) - \boldsymbol{F}_1(\boldsymbol{w}_1^*, \rho_1^*) \leqslant 0$$

由定理 1 可得

$$\boldsymbol{F}_1(\boldsymbol{w}_2^*, \rho_2^*) - \boldsymbol{F}_1(\boldsymbol{w}_1^*, \rho_1^*) =$$
$$\boldsymbol{F}_1(\boldsymbol{w}_2^*, \rho_2^*) - \boldsymbol{F}_2(\boldsymbol{w}_2^*, \rho_2^*) + \boldsymbol{F}_2(\boldsymbol{w}_2^*, \rho_2^*) - \boldsymbol{F}_1(\boldsymbol{w}_1^*, \rho_1^*) \leqslant$$
$$\frac{N - |\boldsymbol{M}^*|}{N \cdot \nu} \sqrt{\frac{\mu}{\nu}}$$

因此得证.

下面讨论 OOCCH 算法的时间复杂度. 在 OOCCH 第 1 阶段，使用序贯最小优化法(SMO) 求解式(6) ～(7)，并以渐进的方式得到凸壳集，其计算复杂度为 $O(\sum_{i=1}^{M} n_i^2)$，其中 $\boldsymbol{M}$ 和 n_i 分别是 SVDD边界向量和当前凸壳向量集的容量. 在 OOCCH 第 2 阶段，仍使用 SMO 方法训练初始分类器，计算复杂度为 $O(|\boldsymbol{M}^*|^2)$，其中 $|\boldsymbol{M}^*|$ 是第 1 阶段最终获得的凸壳向量集的容量. 第 3 阶段是分类器的在线更新阶段，OOCCH 使用式(9) 判断新到达的样本 $\boldsymbol{x}_{\text{new}}$ 是否是候选凸壳向量，其时间复杂度为线性. 因此 OOCCH 总的计算复杂度为 $O(\sum_{i=1}^{M} n_i^2 + |\boldsymbol{M}^*|^2)$，而 $|\boldsymbol{M}^*|$ 值远小于训练样本容量 N，因此该方法能适用于大规模样本的在线分类场景中.

3　实验与分析

3.1　实验设置

本节通过真实数据[19](数据详细信息如表 1 所示)对 OOCCH 分类器进行分析与验证，并与 4 种基于 SVM 的单类分类在线学习算法进行比较：IOCSVM[8]、LS-OC-SVM[9]、WOCSVM [20]和 SO-LS-SVM[21]. 所有的 SVM 分类器均采用高斯核，核参数 σ 范围为$\{10^{-2}, 10^{-1}, \cdots, 10^{2}\}$，正则化参数范围为$\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$(为保持参数的一致性，OOCCH 的正则化参数设为 $1/\nu N$). OOCCH 分类器的误差阈值 μ 范围为$\{10^{-4}, 10^{-3}, 10^{-2}\}$，$K=10^3$ 和 $\beta=10^{-3}$. 另外，各对比算法的参数设置均参照文献中的默认设置. 实验在 2.53-GHz quad-core CPU，8-GB RAM，Windows 7 系统下执行，所有算法均在 Matlab 2016b 环境下实现.

参照文献[16]的实验设置，本文实验分为 4 个步骤：第 1 步，产生实验数据. 首先随机选取 90%正类样本和部分负类样本，使得产生的数据集中 95%的训练样本属于正常类，而 5%的训练样本属于异常类；然后将数据集随机分成 10 份，训练集、扩展集和测试集所占比为 3∶4∶3，其中训练集用于初始分类器的训练，扩展集用于分类器的更新，测试集用于分类器的分类测试. 实验中这一操作重复 10 次，产生 10 组不同的训练集、扩展集和测试集. 第 2 步，初始分类器的训练. 各算法完成给定训练数据集的初始分类器的训练，各调节参数的设置通过 5 重交叉验证法来选取最优值. 在这一步骤中，OOCCH 基于凸壳的定义选择能代表样本在特征空间轮廓分布的凸壳向量，使用凸壳向量对应的原始样本作为训练数据来完成初始分类器的训练. 第 3 步，分类器的在线更新，各算法使用扩展集对分类器进行在线更新. 实验中扩展集分成 10 份，每次使用 1 份扩展集更新分类器. 第 4 步：测试评估. 当所有在线分类器完成更新后，我们使用测试集评估分类器的性能.

为了更好地评价单类分类器的性能，实验采用 *G-mean*[22] 和 *F-measure*[23] 评价准则，

$$G\text{-}mean=\sqrt{Positive\ Accuracy\times Negative\ Accuracy} \tag{11}$$

$$F\text{-}measure=\frac{2\times Recall\times Precision}{Recall+Precision} \tag{12}$$

其中，*G-mean* 能有效评价数据整体的分类性能，*F-measure* 侧重评价异常类样本的分类准确率. *Positive Accuracy* 是异常类样本的分类正确率，*Negative Accuracy* 是正常类样本的分类正确率，$Precision=TP/(TP+FP)$为查准率，其中 TP 是被正确分类的异常类样本数，FP 是被错误分类的正常类样本数，$Recall=Positive\ Accuracy$.

表1 数据集的基本信息

数 据 集	特征数	样本数
Forest cover type(COV)	53	20 000
ncRNA (RNA)	8	147 000
Moore-MAIL(MAIL)	24	30 070
Moore-WWW(WWW)	24	245 000
KDD intrusion detection (KDD)	41	100 000
Seizure detection (SEI)	3	10 000

3.2 OOCCH 参数选择

OOCCH 分类器中有3个参数需要从给定的范围内选取最优值，这3个参数分别是：误差阈值 μ、正则化参数和高斯核参数 σ. 根据文献[24]分析，正则化参数和高斯核参数 σ 宜在一定范围内采用交叉验证法得到，而 OOCCH 中误差阈值 μ 与特征空间中凸壳的选择和算法运行时间密切相关. 因此，本小节从凸壳向量数量和凸壳选择运行时间2方面讨论误差阈值 μ 的选择. 考虑到 σ 与特征空间的选择有关，表2和表3分别列出了在 COV、KDD 和 SEI 数据集上不同 μ 和 σ 时 OOCCH 得到的凸壳向量数量(标准差)和运行时间(标准差). 表4列出了在 COV、KDD 和 SEI 数据集上不同 μ 和 σ 时 OOCCH 在测试集上的 *G-mean* 和 *F-measure* 值.

通过表2～4的结果可以发现：

1) 随着高斯核参数 σ 的增加，凸壳向量的数目也随着增加. 这是因为：σ 值与特征空间的选择有关，σ 值较小时，数据在特征空间内的间距较小而分布较集中，因此获得较少的凸壳向量；反之，σ 值较大时，数据在特征空间内的间距较大而分布较分散，因此获得较多的凸壳向量.

2) 当误差阈值 μ 较小时，满足 $\|\varphi(\boldsymbol{x}_i)-\sum_{t=1}^{|M^*|}\lambda_{i,t}\varphi(\boldsymbol{x}_t)\|^2\leqslant\mu$ 的样本较少，因此获得的凸壳向量较多，同时运行时间较长；反之，当误差阈值 μ 较大时，获得的凸壳向量较少，同时运行时间较小.

3) μ 值与凸壳向量数量、凸壳选择运行时间和测试集上的 *G-mean* 和 *F-measure* 密切相关，凸壳向量数量较多则凸壳选择运行时间较长，但获得的 *G-mean* 和 *F-measure* 值较大，这是因为凸壳向量数量越多时，越能较好地表示数据在特征空间的轮廓分布. 因此在实际应用中需要从运行时间和凸壳向量数量2个方面权衡选择 μ 值，在下面的实验中 μ 值固定为 $\mu=10^{-3}$.

表2 OOCCH 在 COV、KDD 和 SEI 数据集上凸壳向量的数量(标准差)

		$\sigma=10^{-2}$	$\sigma=10^{-1}$	$\sigma=10^{0}$	$\sigma=10^{1}$	$\sigma=10^{2}$
COV	$\mu=10^{-2}$	70.6±1.96	76.1±1.65	78.9±1.37	81.8±1.61	86.3±1.71
	$\mu=10^{-3}$	74.0±1.84	77.8±1.48	80.6±1.42	83.7±1.62	87.8±1.64
	$\mu=10^{-4}$	75.2±1.87	78.9±1.77	82.5±1.55	86.4±1.59	90.6±1.66
KDD	$\mu=10^{-2}$	106.6±2.90	127.5±2.35	148.7±2.98	173.0±2.34	198.4±2.36
	$\mu=10^{-3}$	126.8±2.53	144.8±2.41	167.9±2.82	182.1±2.51	203.9±2.68
	$\mu=10^{-4}$	132.5±2.00	159.7±2.04	188.0±2.35	195.8±2.65	210.7±2.01
SEI	$\mu=10^{-2}$	61.7±1.35	62.7±1.44	64.4±1.06	65.3±1.34	66.9±1.54
	$\mu=10^{-3}$	62.4±1.17	63.4±1.37	65.1±1.14	66.8±1.22	68.1±1.21
	$\mu=10^{-4}$	63.8±1.19	64.9±1.50	65.8±1.18	67.2±1.30	69.3±1.17

表 3　OOCCH 在 COV、KDD 和 SEI 数据集上计算凸壳向量的运行时间(标准差)

		$\sigma=10^{-2}$	$\sigma=10^{-1}$	$\sigma=10^{0}$	$\sigma=10^{1}$	$\sigma=10^{2}$
COV	$\mu=10^{-2}$	1.11±0.004	1.25±0.006	1.29±0.006	1.31±0.004	1.36±0.005
	$\mu=10^{-3}$	1.34±0.008	1.37±0.007	1.38±0.007	1.40±0.005	1.43±0.006
	$\mu=10^{-4}$	1.68±0.009	1.80±0.007	1.91±0.006	2.23±0.008	2.47±0.009
KDD	$\mu=10^{-2}$	3.50±0.010	3.67±0.009	3.87±0.008	3.99±0.009	4.02±0.009
	$\mu=10^{-3}$	3.68±0.009	3.85±0.008	3.96±0.009	4.25±0.010	4.66±0.010
	$\mu=10^{-4}$	3.92±0.010	4.62±0.010	4.91±0.009	6.18±0.011	6.23±0.008
SEI	$\mu=10^{-2}$	1.06±0.007	1.08±0.005	1.14±0.004	1.15±0.007	1.22±0.004
	$\mu=10^{-3}$	1.08±0.006	1.10±0.005	1.23±0.006	1.24±0.006	1.25±0.005
	$\mu=10^{-4}$	1.29±0.005	1.62±0.006	1.96±0.008	2.28±0.009	2.80±0.006

表 4　OOCCH 在 COV、KDD 和 SEI 数据集上的 *G-mean*(标准差)和 *F-measure*(标准差)

			$\sigma=10^{-2}$	$\sigma=10^{-1}$	$\sigma=10^{0}$	$\sigma=10^{1}$	$\sigma=10^{2}$
COV	$\mu=10^{-2}$	*G-mean*/%	68.90±0.68	70.01±0.57	72.06±0.45	72.00±0.50	72.00±0.61
		F-measure/%	34.29±0.23	35.88±0.26	38.55±0.30	38.40±0.27	38.39±0.28
	$\mu=10^{-3}$	*G-mean*/%	70.42±0.74	71.75±0.50	73.95±0.51	73.70±0.61	73.07±0.68
		F-measure/%	35.48±0.38	38.02±0.31	43.02±0.39	41.86±0.33	41.06±0.33
	$\mu=10^{-4}$	*G-mean*/%	70.49±0.77	71.80±0.54	73.95±0.60	73.74±0.55	73.09±0.67
		F-measure/%	35.48±0.30	35.50±0.30	43.03±0.31	41.90±0.29	41.09±0.29
KDD	$\mu=10^{-2}$	*G-mean*/%	72.09±0.65	74.86±0.60	74.45±0.62	76.71±0.56	76.38±0.63
		F-measure/%	58.73±0.54	62.09±0.34	62.00±0.43	66.59±0.40	66.10±0.44
	$\mu=10^{-3}$	*G-mean*/%	72.86±0.61	75.40±0.58	76.08±0.60	77.50±0.53	77.07±0.58
		F-measure/%	60.21±0.35	63.11±0.36	65.70±0.34	67.27±0.35	66.84±0.30
	$\mu=10^{-4}$	*G-mean*/%	72.89±0.58	75.52±0.44	76.26±0.52	77.61±0.48	77.25±0.52
		F-measure/%	60.27±0.46	63.19±0.42	65.88±0.49	67.34±0.47	67.10±0.45
SEI	$\mu=10^{-2}$	*G-mean*/%	97.21±0.69	97.61±0.57	97.74±0.60	97.10±0.52	97.00±0.61
		F-measure/%	95.29±0.54	95.48±0.48	95.51±0.50	94.93±0.51	94.67±0.53
	$\mu=10^{-3}$	*G-mean*/%	97.42±0.68	97.91±0.60	97.89±0.62	97.13±0.58	97.04±0.59
		F-measure/%	95.53±0.49	95.73±0.49	95.70±0.48	95.36±0.47	95.17±0.48
	$\mu=10^{-4}$	*G-mean*/%	97.57±0.47	97.93±0.52	97.90±0.50	97.15±0.49	97.06±0.52
		F-measure/%	95.66±0.39	95.73±0.41	95.72±0.40	95.36±0.41	95.18±0.40

3.3　性能比较

本小节我们在 6 个真实数据集上进行 OOCCH 与 4 种单类分类在线学习算法 IOCSVM、LS-OC-SVM、WOCSVM 和 SO-LS-SVM 的性能比较. 图 1 和图 2 分别显示了 OOCCH 和另外 4 种在线学习算法训练初始分类器的运行时间和在线更新分类器的运行时间(单位：s). 表 5 显示了 OOCCH 和对比算法在测试集上的 *G-mean*(标准差)和 *F-measure*(标准差)比较. 根据图 1,2 和表 5 的结果可以看出：

1) OOCCH 在训练初始分类器和在线更新分类器上花费时间最少，因为 OOCCH 能够根据数据在特征空间的分布得到代表其轮廓分布的凸壳向量，并以凸壳向量对应的原始样本作为训练数据，这样在保证分类精度的情况下能有效缩减训练集的容量；

2) OOCCH 在线学习中基于凸壳的定义调整分类器的训练数据，既可以动态调整分类器，使之适应新的数据分布，同时又不增加分类器的训练负担；

3) 前文证明使用凸壳向量对应的原始样本作为训练数据不会降低分类的性能，表 5 的结果显示从

G-mean 和 *F-measure* 2个性能指标上OOCCH取得了令人满意的分类效果. 在6个真实数据集上，除了在WWW数据集上的性能略逊于LS-OC-SVM算法外，在其他5个数据集上均取得了最优的 *G-mean* 和 *F-measure* 结果.

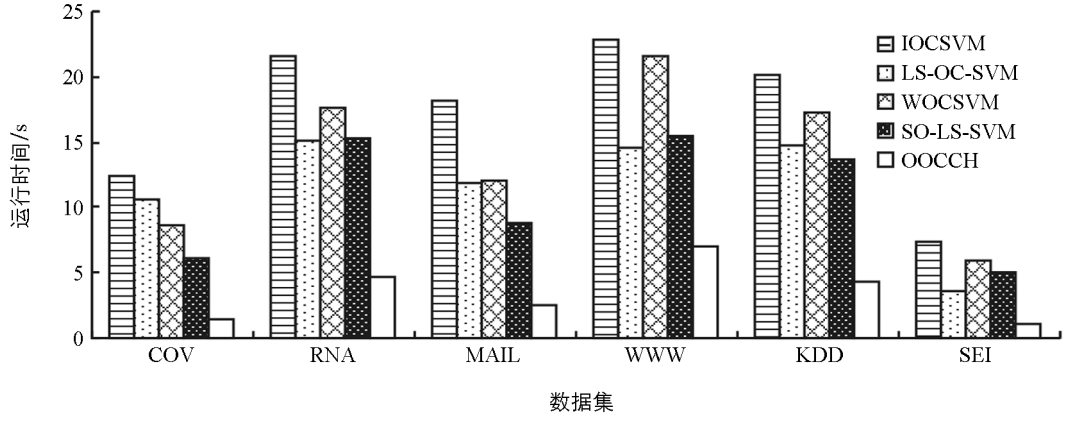

图1 OOCCH和对比算法训练初始分类器的时间比较(单位：s)

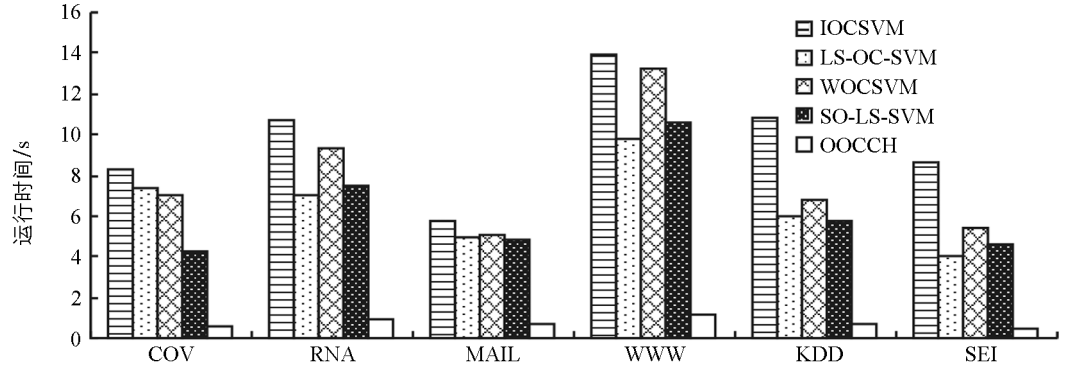

图2 OOCCH和对比算法在线更新分类器的时间比较(单位：s)

表5 OOCCH和对比算法在测试集上的 *G-mean*(标准差)和 *F-measure*(标准差)比较

		IOCSVM	LS-OC-SVM	WOCSVM	SO-LS-SVM	OOCCH
COV	*G-mean*/%	71.06±0.57	71.89±0.60	73.08±0.58	73.07±0.62	73.95±0.51
	F-measure/%	42.03±0.30	42.76±0.39	42.97±0.36	42.81±0.38	43.02±0.39
RNA	*G-mean*/%	93.27±0.51	93.84±0.57	95.41±0.58	95.19±0.57	96.43±0.50
	F-measure/%	92.18±0.43	92.65±0.60	94.75±0.56	94.60±0.60	95.99±0.40
MAIL	*G-mean*/%	95.56±0.52	96.07±0.62	97.70±0.59	97.35±0.67	98.21±0.50
	F-measure/%	93.61±0.52	94.85±0.58	95.82±0.54	95.33±0.60	96.43±0.51
WWW	*G-mean*/%	96.09±0.61	97.02±0.67	98.56±0.60	98.24±0.64	98.45±0.55
	F-measure/%	94.77±0.57	95.68±0.60	97.93±0.64	97.20±0.61	97.82±0.53
KDD	*G-mean*/%	77.08±0.46	76.85±0.55	76.44±0.54	77.04±0.58	77.50±0.53
	F-measure/%	66.59±0.48	65.30±0.47	65.01±0.50	66.51±0.50	67.27±0.35
SEI	*G-mean*/%	94.87±0.54	94.07±0.53	96.67±0.69	97.14±0.57	97.91±0.60
	F-measure/%	92.40±0.49	90.95±0.51	93.99±0.67	95.36±0.60	95.73±0.49

4 总　结

本文基于凸壳的定义提出了一种新的在线单类学习机 OOCCH. OOCCH 在分类器初始训练阶段使用凸壳技术对训练集进行选择，使所选样本能够在最大程度上表示数据集在特征空间的轮廓分布. 在线更新阶段 OOCCH 再次使用凸壳技术对新到达的样本进行筛选，仅保留能改变数据集分布形状的凸壳样本作为训练集的补充. 理论分析和真实数据集的仿真实验证明了本文算法具有优良的分类性能和较少的运行时间.

参考文献：

[1] 潘志松，陈　斌，缪志敏，等. One-Class 分类器研究 [J]. 电子学报，2009，37 (11)：2498－2503.

[2] LIU J C，MIAO Q G，SUN Y N，et al. Modular Ensembles for One-Class Classification Based on Density Analysis [J]. Neurocomputing，2016，171(C)：262－276.

[3] LIU B，XIAO Y，YU P S，et al. Uncertain One-Class Learning and Concept Summarization Learning on Uncertain Data Streams [J]. IEEE Transactions on Knowledge and Data Engineering，2013，26 (2)：468－484.

[4] KRAWCZYK B，WOZNIAK M，CYGANEK B. Clustering-Based Ensembles for One-Class Classification [J]. Information Sciences，2014，264(6)：182－195.

[5] 张辉荣，唐　雁，何　茨，等. 面向分类数据的重叠子空间聚类算法 SCCAT [J]. 西南大学学报(自然科学版)，2016，38(3)：171－176.

[6] MANEVITZ L，YOUSEF M. One-Class Document Classification via Neural Networks [J]. Neurocomputing，2007，70 (7/9)：1466－1481.

[7] ZHANG M，WU J，LIN H，et al. The Application of One-Class Classifier Based on CNN in Image Defect Detection [J]. Procedia Computer Science，2017，114：341－348.

[8] 梁修荣，杨正益. 基于聚类和 SVM 的数据分类方法与实验研究 [J]. 西南师范大学学报(自然科学版)，2018，43(3)：91－96.

[9] WANG T，CHEN J，ZHOU Y，et al. Online Least Squares One-Class Support Vector Machines-Based Abnormal Visual Event Detection [J]. Sensors，2013，13(12)：17130－17155.

[10] KIVINEN J，SMOLA A J，WILLIAMSON R C. Online Learning with Kernels [J]. IEEE Transactions on Signal Processing，2004，52(8)：2165－2176.

[11] YANG H Q，LYU M R，KING I. Efficient Online Learning for Multitask Feature Selection [J]. ACM Transactions on Knowledge Discovery from Data，2013，7(2)：1－27.

[12] ZHENG J，SHEN F，FAN H，et al. An Online Incremental Learning Support Vector Machine for Large-Scale Data [J]. Neural Computing and Applications，2013，22(5)：1023－1035.

[13] SAITOH D，HARA K. Mutual Learning Using Nonlinear Perceptron [J]. Journal of Artificial Intelligence and Soft Computing Research，2015，5(1)：71－77.

[14] LANGFORD J，LI L H，ZHANG T. Sparse Online Learning via Truncated Gradient [J]. Journal of Machine Learning Research，2009，10(2)：777－801.

[15] KRELL M M，WILSHUSEN N，SEELAND A，et al. Classifier Transfer with Data Selection Strategies for Online Support Vector Machine Classification with Class Imbalance [J]. Journal of Neural Engineering，2017，14 (2)：025003.

[16] WANG D，QIAO H，ZHANG B. Online Support Vector Machine Based on Convex Hull Vertices Selection [J]. IEEE Transaction on Neural Networks and Learning Systems，2013，24(4)：593－609.

[17] TAX D M J，DUIN R P W. Support Vector Data Description [J]. Machine Learning，2004，54(1)：45－66.

[18] SHALEV-SHWARTZ S，SINGER Y，SREBRO N，et al. Pegasos：Primal Estimated Sub-Gradient Solver for SVM [C]// The 24th International Conference on Machine Learning. Corvalis，USA：ACM，2007：807－814.

[19] UC Irvine Machine Learning Repository. UCI database [EB/OL]. [2013－1－12]. https：//archive. ics. uci. edu/ml/datasets. html.

[20] KRAWCZYK B, WOZNIAK M. One-Class Classifiers with Incremental Learning and Forgetting for Data Streams with Concept Drift [J]. Soft Computing, 2015, 19(12): 3387－3400.

[21] UDDIN M S, KUH A. Online Least-squares One-Class Support Vector Machine for Outlier Detection in Power Grid Data [C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016: 2628－2632.

[22] BATUWITA R, PALADE V. FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning [J]. IEEE Transactions on Fuzzy Systems, 2010, 18(3): 558－571.

[23] GU X Q, CHUNG F L, ISHIBUCHI H S, et al. Imbalanced TSK Fuzzy Classifier by Cross-Class Bayesian Fuzzy Clustering and Imbalance Learning [J]. IEEE Transactions on Systems, Man, and Cybernetics Systems, 2017, 47(8): 2005－2020.

[24] CHANG C C, LIN C J. LIBSVM: A Library for Support Vector Machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1－27.

A One-Class Online Classifier Based on Convex Hull

ZHOU Guo-hua[1,2], SHEN Yan-ping[1], YIN Xin-chun[2]

1. *Department of Information Engineering, Changzhou Institute of Light Industry Technology, Changzhou Jiangsu 213164, China;*
2. *College of Information Engineering, Yangzhou University, Yangzhou Jiangsu 225127, China*

Abstract: Facing the challenge of large-scale data processing, the traditional SVM(support vector machine) based one-class classifier suffers from its high computational complexity. The online learning technique is an effective way to solve this problem. In this paper, a one-class online classifier based on convex hull (OOCCH) is proposed by considering the distribution characteristics of the data in the feature space. In order to reduce the number of training sets, OOCCH selects the samples corresponding to the convex hull vectors in the feature space as training samples. In the online update stage of the classifier, OOCCH dynamically adjusts the training samples based on the definition of convex hull. Theoretical analysis proves the effectiveness of OOCCH. Compared with the existing online one-class classifiers in experiments, OOCCH has significant advantages in training time and classification performance.

Key words: online learning; one-class; classification; convex hull

责任编辑 崔玉洁