一种基于单簇核 PCM 的 SVDD 离群点检测方法

杨金鸿[1],邓廷权[1,2]
(1. 哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001;2. 哈尔滨工程大学理学院,黑龙江哈尔滨 150001)

摘 要: 针对支持向量数据描述(Support Vector Data Description,SVDD)的训练集中同时含有正常点和离群点的问题,为降低离群点对 SVDD 训练模型的不利影响,提出了一种基于单簇核可能性 C-均值的 SVDD 离群点检测算法. 本文算法通过单簇核聚类获得每个样本属于正常类的隶属度,将其作为每个样本属于目标类的置信度. 将样本置信度引入到 SVDD 训练模型中,减弱低置信度样本在建立决策边界中的作用. 实验表明,与已有的相关方法相比,本文方法能够显著改善 SVDD 的离群点检测效果.

关键词: 离群点检测; 支持向量数据描述; 可能性 C-均值; 置信度
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)04-0813-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.04.007

A One-Cluster Kernel PCM Based SVDD Method for Outlier Detection

YANG Jin-hong[1], DENG Ting-quan[1,2]
(1. *College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*;
2. *College of Science, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*)

Abstract: In order to reduce the negative influence of outliers on the model of support vector data description (SVDD) when the training dataset contains both normal samples and outliers which are all labeled as target class, a one-cluster kernel possibilistic C-means based SVDD method for outlier detection is proposed. In this paper, each sample of the training dataset is assigned a confidence level based on the membership degree of each sample belonging to the normal class, which is obtained through the one-cluster kernel PCM clustering. The proposed algorithm incorporates the confidence levels into the training model to reduce the importance of the samples which have less confidence levels. The experimental results show that the proposal significantly improves the effect of outlier detection, compared with the existing SVDD-based outlier detection methods.

Key words: outlier detection; support vector data description; possibilistic C-means; confidence level

1 引言

离群点是指数据集中与大部分数据具有显著差异或不一致的少数数据点[1]. 离群点检测旨在挖掘数据背后的罕见模式或有意义的知识. 随着大数据时代的到来,离群点检测备受关注. 近年来,离群点检测成功应用在军事侦察[2]、信用卡欺诈检测[3]、网络入侵检测[4]、故障检测[5]、天气预报[6]、医学辅助诊断[7]等问题中.

在实际应用中,往往可以获得大量的正常样本,而离群数据难以获得和描述. 在此背景下,基于单分类的离群点检测得到了广泛应用. SVDD 是一种经典的单分类方法[8]. SVDD 模型旨在高维特征空间中确定一个最小超球,使得该超球尽可能的包络目标样本(正常点),位于超球面外面的样本点(非目标类)为离群点[8~10]. 其中,目标类为正常点,非目标数据为离群点. 由于 SVDD 离群点检测方法把整个数据集作为目标类进行训练. 当训练集中含有离群数据时,SVDD 的超球面会

收稿日期:2016-08-26;修回日期:2016-10-26;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 11471001)

将离群点包含在训练集超球内部,出现过拟合现象[11,12],进而降低离群点检测精度.

针对上述问题,目前已有改进方法主要分为:基于密度的和基于中心点的方法. 基于密度的方法利用训练集中样本分布的密度信息来度量样本点的重要性,降低离群点对 SVDD 训练过程的影响. Liu 等利用核局部离群因子方法计算每个样本属于目标类的概率[11]. Lee 等提出基于密度的 SVDD 算法(Density-induced SVDD, D-SVDD)[13],通过引入邻域距离和 Parzen 窗口描述每个样本的相对密度. Cha 等提出密度加权的 SVDD 算法(Density Weighted SVDD, DW-SVDD)[14],利用 k-近邻计算每个样本的权值,降低分布稀疏的离群样本在模型训练中的重要性. 基于中心点的方法利用样本与中心点的距离确定样本点的重要性. Liu 等提出不确定数据的 SVDD 离群点检测算法(Uncertain SVDD, UnSVDD)[15],利用核空间中训练集的均值作为中心点. 基于密度的方法仅仅考虑样本局部分布,当离群点形成样本少但密度高的过一个聚簇时,基于密度的方法会错误地赋予其较高置信度. 目前基于中心点的方法在确定中心点这一关键问题时,并没有考虑样本属于正常类的可能性,简单假设每个样本属于正常类的可能性均为 1,忽略了离群点对中心点确定的影响.

针对模糊 C-均值算法(Fuzzy C-means, FCM)[16]对离群点敏感的问题, Krishnapuram 等放松了 FCM 中隶属度之和为 1 的约束,提出了可能性 C-均值算法(Possibilistic C-means, PCM)[17]. 放松了隶属度之和为 1 的约束,更加客观地描述样本属于聚簇的程度. 文献[18]通过单簇 PCM 得到所有训练样本对目标类的隶属度,然后利用设置的隶属度阈值对训练集进行单分类,取得了较好的效果.

本文提出一种基于单簇核 PCM 的 SVDD 离群点检测算法(One-Cluster kernel PCM based SVDD, OCP-SVDD). 首先,基于文献[18]多子簇单类问题的思想,在核空间中采用单簇核 PCM 聚类计算每个样本属于正常类的隶属度,作为样本点属于目标类的置信度. 然后本文在 SVDD 训练模型中引入样本置信度,利用置信度降低偏离于中心点的样本识别为非目标类的误差,从而降低离群点对 SVDD 训练模型的不利影响,改善离群点检测效果.

2 基于单簇核 PCM 的 SVDD 离群点检测

2.1 单簇核 PCM

若给定的数据集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l\}$,其中 $\boldsymbol{x}_i\in\mathbb{R}^n$ $(1\leqslant i\leqslant l)$,引入核映射 $\varphi(\cdot)$ 的动机在于:若 X 一个非线性可分的复杂数据集,则由 Cover 定理所知,可通过一个非线性映射将数据映射到一个高维的特征空间中,使非线性可分问题转化为线性可分问题. 单簇核 PCM 将整个数据集看作是一个聚簇,其目标函数定义为:

$$\begin{aligned}&\min J(\boldsymbol{U},\boldsymbol{v})=\sum_{i=1}^{l}u_i^m\|\varphi(\boldsymbol{x}_i)-\boldsymbol{v}\|^2+\eta\sum_{i=1}^{l}(1-u_i)^m\\&\text{s.t. } 0\leqslant u_i\leqslant 1,i=1,\cdots,l\end{aligned}\tag{1}$$

其中,$\boldsymbol{U}=[u_1,u_2,\cdots,u_l]$ 为隶属度向量,u_i 表示第 i $(1\leqslant i\leqslant l)$ 个样本属于正常类的隶属度;$\boldsymbol{v}$ 表示正常类的聚簇中心;$\|\cdot\|$ 是欧几里得范数;m 为模糊指数,用于调节隶属度 u_i 的模糊度. 核函数为特征空间中样本的内积,即 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)$. 参数 η 为正则化因子,计算公式如下:

$$\eta=\frac{\sum_{i=1}^{l}u_i^m\|\varphi(\boldsymbol{x}_i)-\boldsymbol{v}\|^2}{\sum_{i=1}^{n}u_i^m}\tag{2}$$

利用拉格朗日方法求解优化问题(1),构建拉格朗日函数[19]如下:

$$L(\boldsymbol{U},\boldsymbol{v})=\sum_{i=1}^{l}u_i^m\|\varphi(\boldsymbol{x}_i)-\boldsymbol{v}\|^2+\eta\sum_{i=1}^{l}(1-u_i)^m$$

求解 $L(\boldsymbol{U},\boldsymbol{v})$ 关于 $\boldsymbol{v}$ 的偏导数,令其为零:

$$\frac{\partial L}{\partial\boldsymbol{v}}=-2\sum_{i=1}^{l}u_i^m(\varphi(\boldsymbol{x}_i)-\boldsymbol{v})=0$$

由此可知:

$$\boldsymbol{v}=\frac{\sum_{i=1}^{l}u_i^m\varphi(\boldsymbol{x}_i)}{\sum_{i=1}^{l}u_i^m}\tag{3}$$

由聚簇中心 $\boldsymbol{v}$ 的计算公式可知,隶属度值 u_i 较高的样本点将发挥较大的作用. 由于离群点偏离数据集中心,由式(3)知离群点属于正常类的可能性较低,对于聚簇中心的确定影响较小. 由此可知,基于核 PCM 的方法能够较精准地定位聚簇中心,一定程度提高样本隶属度确定过程的精确性.

由式(3)知,核空间中样本 $\boldsymbol{x}_i$ 与聚簇中心之间的距离计算如下:

$$\begin{aligned}\|\varphi(\boldsymbol{x}_i)-\boldsymbol{v}\|^2&=\left(\varphi(\boldsymbol{x}_i)-\frac{\sum_{j=1}^{l}u_j^m\varphi(\boldsymbol{x}_j)}{\sum_{j=1}^{l}u_j^m}\right)^2=K(\boldsymbol{x}_i,\boldsymbol{x}_i)\\&-2\frac{\sum_{j=1}^{l}u_j^mK(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sum_{j=1}^{l}u_j^m}+\frac{\sum_{p=1}^{l}\sum_{q=1}^{l}u_p^mu_q^mK(\boldsymbol{x}_p,\boldsymbol{x}_q)}{\sum_{p=1}^{l}\sum_{q=1}^{l}u_p^mu_q^m}\end{aligned}\tag{4}$$

求解 $L(\boldsymbol{U},\boldsymbol{v})$ 关于 u_i 的偏导数,令其为零:

$$\frac{\partial J}{\partial u_i}=mu_i^{m-1}\|\varphi(\boldsymbol{x}_i)-\boldsymbol{v}\|^2-\eta m(1-u_i)^{m-1}=0$$

由此可知:

$$u_i=\frac{1}{1+\left(\frac{\|\varphi(\boldsymbol{x}_i)-\boldsymbol{v}\|^2}{\eta}\right)^{1/(m-1)}} \tag{5}$$

由 u_i 的计算公式可知,在核空间中与聚簇中心具有较大距离的样本属于正常类的可能性较低.

综上,单簇核 PCM 算法步骤描述如算法 1:

算法 1　单簇核 PCM 算法

输入:数据集 X,模糊指数 m,循环停止阈值 ε.

输出:隶属度 $\boldsymbol{U}$.

1. 初始化迭代次数 $t=0$,随机初始化隶属度 $\boldsymbol{U}^{(t)}$;
2. 利用公式(2)计算参数 η;
3. 利用公式(3)计算聚簇中心 $\boldsymbol{v}^{(t)}$;
4. 利用公式(5)计算隶属度 $\boldsymbol{U}^{(t)}$;
5. 如果满足条件 $\|\boldsymbol{U}^{(t+1)}-\boldsymbol{U}^{(t)}\|<\varepsilon$,返回 $\boldsymbol{U}$,结束循环;否则,$t=t+1$,返回步骤 3.

2.2　基于单簇核 PCM 的 SVDD

由单簇核 PCM 获得的隶属度 u_i 表示样本 $\boldsymbol{x}_i$ 属于正常类的可能性,将其作为 $\boldsymbol{x}_i$ 属于目标类的置信度. $u_i\xi i$ 表示不同置信度的样本分类为离群点产生的惩罚量. 正常点距离聚簇中心较近,置信度都较高,对训练模型的影响较大. 反之,离群点对训练模型贡献较小. 将样本置信度引入到 SVDD 训练模型中,则基于单簇核 PCM 的 SVDD 的目标函数如下:

$$\begin{aligned}&\min R^2+\gamma\sum_{i=1}^{l}u_i\xi_i,\\&\text{s.t. }\|\varphi(\boldsymbol{x}_i)-\boldsymbol{o}\|^2\leqslant R^2+\xi_i,\\&\xi_i\geqslant 0,i=1,2,\cdots,l\end{aligned} \tag{6}$$

其中 $\boldsymbol{o}$ 表示超球面中心,R 为半径. $\gamma>0$ 是惩罚系数,用于调节超球半径和离群点数目. 目标函数(6)中,置信度 u_i 的值越小,相应的代表惩罚量的松弛变量 ξ_i 发挥的作用越低,样本 $\boldsymbol{x}_i$ 的重要性越小. 通过最小化 $\sum_{i=1}^{l}u_i\xi_i$,使得最优超球面倾向于包含中心附近的正常点区域,而排除偏离于中心的离群点区域. 为了求解优化问题(6),构建拉格朗日函数[19]:

$$\begin{aligned}L(R,\boldsymbol{o},\xi)=R^2+\gamma\sum_{i=1}^{l}u_i\xi_i-\\\sum_{i=1}^{l}\alpha_i(R^2+\xi_i-\|\varphi(\boldsymbol{x}_i)-\boldsymbol{o}\|^2)-\sum_{i=1}^{l}\beta_i\xi_i\end{aligned} \tag{7}$$

其中,$\alpha_i\geqslant 0$ 和 $\beta_i\geqslant 0(i=1,2,\ldots,l)$ 为拉格朗日乘子. 各参数满足如下条件:

$$\frac{\partial L}{\partial R}=0\rightarrow 2R-2R\sum_{i=1}^{l}\alpha_i=0\rightarrow\sum_{i=1}^{l}\alpha_i=1 \tag{8}$$

$$\frac{\partial L}{\partial \boldsymbol{o}}=0\rightarrow 2\boldsymbol{o}-2\sum_{i=1}^{l}\alpha_i\varphi(\boldsymbol{x}_i)=0\rightarrow\boldsymbol{o}=\sum_{i=1}^{l}\alpha_i\varphi(\boldsymbol{x}_i) \tag{9}$$

$$\frac{\partial L}{\partial \xi_i}=0\rightarrow\alpha_i+\beta_i=u_i\gamma \tag{10}$$

则优化问题(6)转化为如下对偶问题:

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{l}\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x}_i)-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j),\\&\text{s.t. }0\leqslant\alpha_i\leqslant\gamma u_i,i=1,2,\cdots,l\\&\sum_{i=1}^{l}\alpha_i=1\end{aligned} \tag{11}$$

通过求解线性约束二次优化问题(11)可以获得 α_i. 仅当 $\alpha_i>0$ 时,样本点 $\boldsymbol{x}_i$ 对超球面的中心产生影响,相应的样本点 $\boldsymbol{x}_i$ 称为支持向量. 另外,KKT 理论满足:

$$\xi_i\beta_i=0,i=1,2,\cdots,l, \tag{12}$$

$$(R^2+\xi_i-\|\varphi(\boldsymbol{x}_i)-\boldsymbol{o}\|^2)\alpha_i=0,i=1,2,\cdots,l. \tag{13}$$

如果 $0<\alpha_i<\gamma u_i$,则由式(10)知 $\beta_i\neq 0$,且由式(12)知 $\xi_i=0$. 又由式(13)知,若支持向量满足如下条件:

$$R^2-\|\varphi(\boldsymbol{x}_i)-\boldsymbol{o}\|^2=0,i=1,2,\cdots,l. \tag{14}$$

则其分布于超球面上. 如果 $\boldsymbol{x}_k$ 是分布于超球面上的支持向量,即 $0<\alpha_k<\gamma u_k$,则半径 R 的计算公式如下:

$$\begin{aligned}R^2=\|\varphi(\boldsymbol{x}_k)-\boldsymbol{o}\|^2=K(\boldsymbol{x}_k,\boldsymbol{x}_k)\\-2\sum_{i=1}^{l}\alpha_iK(\boldsymbol{x}_k,\boldsymbol{x}_i)+\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\end{aligned} \tag{15}$$

对于给定测试样本 $\boldsymbol{x}\in\mathbb{R}^n$,可根据如下决策函数对其进行分类:

$$f(\boldsymbol{x})=R^2-\|\varphi(\boldsymbol{x})-\boldsymbol{o}\|^2 \tag{16}$$

若 $f(\boldsymbol{x})\geqslant 0$,即 $\boldsymbol{x}$ 到超球中心的距离小于等于 R,则 $\boldsymbol{x}$ 判定为正常点;否则,$\boldsymbol{x}$ 判定为离群点.

综上,基于单簇核 PCM 的 SVDD 算法描述如算法 2.

算法 2　OCP-SVDD 算法

输入:样本及其置信度值$(\boldsymbol{x}_i,u_i)$ $(1\leqslant i\leqslant l)$,参数 γ.

输出:离群点.

训练:

1. 求解优化问题(11);
2. 获得每个样本的 $\alpha_i(1\leqslant i\leqslant l)$ 值;
3. 通过式(15) 计算半径 R;

测试:

1. 对于样本 $\boldsymbol{x}$,由决策函数式(16)判断 $\boldsymbol{x}$ 是否为离群点,将离群点返回.

3　实验分析

为验证本文方法的有效性,分别在模拟数据集和

UCI 数据集上,将 OCP-SVDD 算法与经典的 SVDD[8]、基于中心点的 UnSVDD[15]算法、基于密度的 D-SVDD[13]和 DW-SVDD[14]算法进行对比分析. 本文实验环境为: Windows XP 操作系统, Pentium 双核 CPU, 主频为 3GHz,2GB 内存,运行平台为 MATLAB R2009a. 实验中所有算法均统一采用高斯核函数.

3.1 评价方法

检测率和误警率是评价离群点检测方法性能的常用准则[12]. 检测率定义为 TPR = TP/(TP + FN),误警率为 FPR = FP/(TN + FP),见表 1 中的混淆矩阵. ROC (Receiver Operating Characteristic) 曲线主要考察离群点检测率和误警率之间的平衡关系,其横坐标是误警率,纵坐标是检测率. 一般地,AUC (Area Under the ROC Curve)指 ROC 曲线下的面积, AUC 值越大,离群点检测的效果越佳. 由于算法具有一定的随机性,本文每组实验重复运行 10 次,并将各算法 AUC 值的平均统计量用作实验对比.

表 1 检测结果的混淆矩阵

	检测为离群点	检测为正常点
实际为离群点	True Positive (TP)	False Negative (FN)
实际为正常点	False Positive (FP)	True Negative (TN)

3.2 模拟数据集实验

实验采用的模拟数据集如图 1(a)所示. 在实验中,各算法设置的惩罚系数 γ 和高斯核函数带宽 σ 分别见图 1(b) ~ (f)标题. 此外,D-SVDD 算法设置的权值因子 ω 值,以及 OCP-SVDD 算法设置的模糊因子 m 值,具体见图 1(c)和 2(f)标题. 图 2 所示为各算法计算出的模拟数据集的样本置信度,其中 1 ~ 115 样本点为正常点,116 - 136 样本点为离群点,即竖直虚线左侧为正常点,竖直虚线右侧为离群点. 图 1(b) ~ (f)分别是各算法在模拟数据集上的离群点检测结果,其中“.”表示目标类样本,“×”表示检测出的离群点,圆圈“○”表示支持向量对应的样本点,虚线表示决策边界.

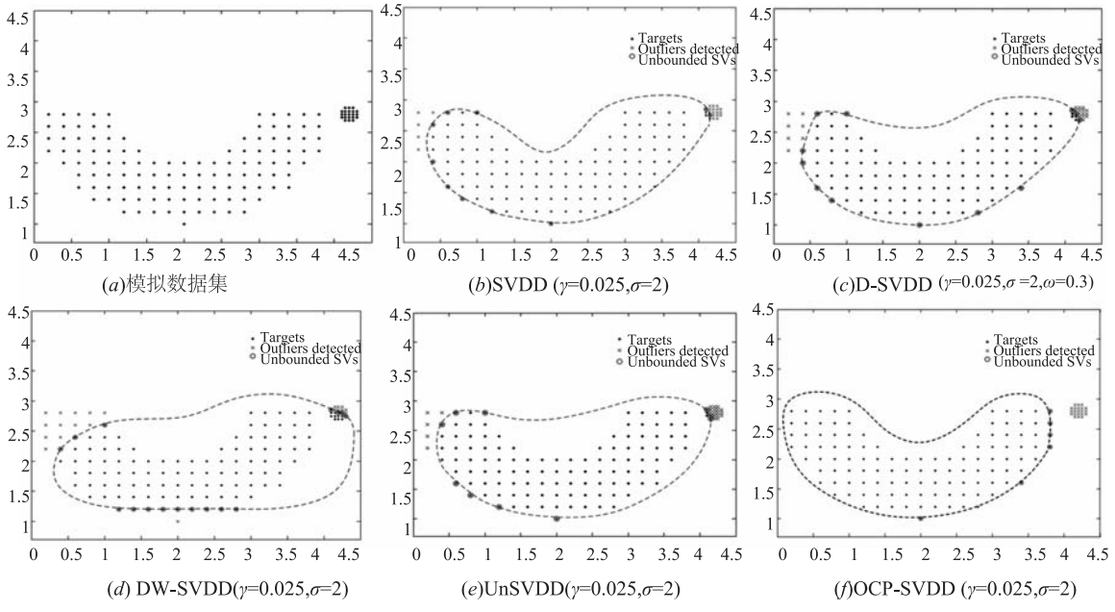

图1 不同算法在模拟数据集上的离群点检测结果

图 1 表明,OCP-SVDD 算法边界面探测结果明显优于其他算法. 原因在于,SVDD 算法假设每个样本点的置信度均为 1,如图 2 所示. 离群点的存在影响边界面的判定,图 1(b)中的边界面向右偏离. D-SVDD 和 DW-SVDD 算法均是基于密度的样本置信度计算方法,密度较高的样本点赋予较高的置信度. 模拟数据集中离群点分布较密集,造成离群点的置信度远高于正常点,如图 2 所示. 因而造成 D-SVDD 和 DW-SVDD 算法的超球面严重向右偏离,如图 1(c)-(d)所示. UnSVDD 算法采用基于中心点的置信度计算方法,但由于右侧离群点的存在,UnSVDD 算法中心点偏右,从而造成边界面向右偏离,如图 1(e). 本文算法采用单簇核 PCM 计算置信度,偏离聚类中心的样本具有较小的置信度,如图 2 所示,偏离聚类中心的离群点具有较低的置信度,使得边界面能够准确地描述目标类,如图 1(f).

图 3 为各算法在模拟数据集上的 ROC 曲线及 AUC 值,显然 OCP-SVDD 算法离群点检测结果的 AUC 值明显高于其他对比算法. 原因在于:与其他算法相比,OCP-SVDD 算法能够精确地计算样本置信度,减弱离群点对训练模型的影响,获得更准确的超球面来描述目标类数据.

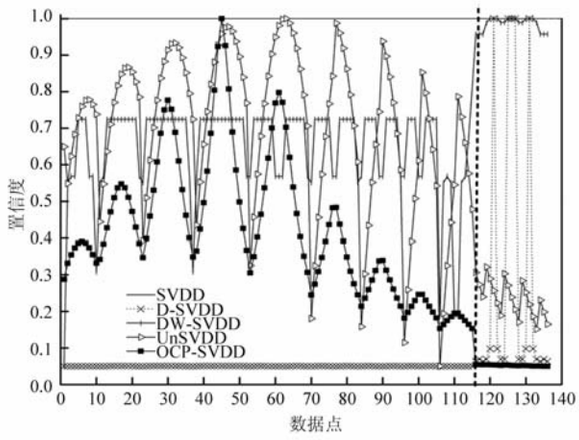

图2　模拟数据集的样本置信度

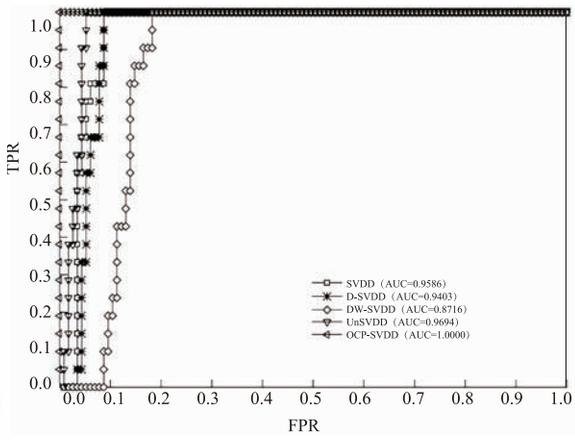

图3　各算法在模拟数据集上的ROC曲线及AUC值

3.3　真实数据集实验

实验数据来自 8 个 UCI 数据集,数据集的具体介绍参见表 2. 实验将每个数据集中样本点最多的一类作为目标类,其余类作为非目标类. 从非目标类中随机选择少量样本点作为离群点,利用离群点检测效果评估算法对非目标类的识别能力.

表 2　真实数据集介绍

数据集简称	数据源	目标类	目标样本数	离群点数	特征数
Liv	Liver Disorders	Class2	200	10	7
Aba	Abalone	Class9	689	30	8
Ion	Ionosphere	Class1	225	30	34
Win	Wine	Class2	72	10	4
Ima(1)	Image Segmentation	Grass	330	10	19
Ima(2)	Image Segmentation	Window	330	10	19
Bre	Breast Cancer Wisconsin	Benign	458	30	9
KDDCUP	KDDCUP1999	Normal	56237	4177	41
Ima(3)	Image Segmentation	Window, Grass, Sky	690	25	19
Iri	Iris	Setosa, Versicolor	100	10	4

3.3.1　精度分析

图 4 为各算法在 10 个 UCI 数据集上离群点检测结果的平均 AUC 值. 在实验中所有算法的惩罚系数 γ 值和高斯核函数带宽 σ 值设置已分别标注在图 4(*a*) ~

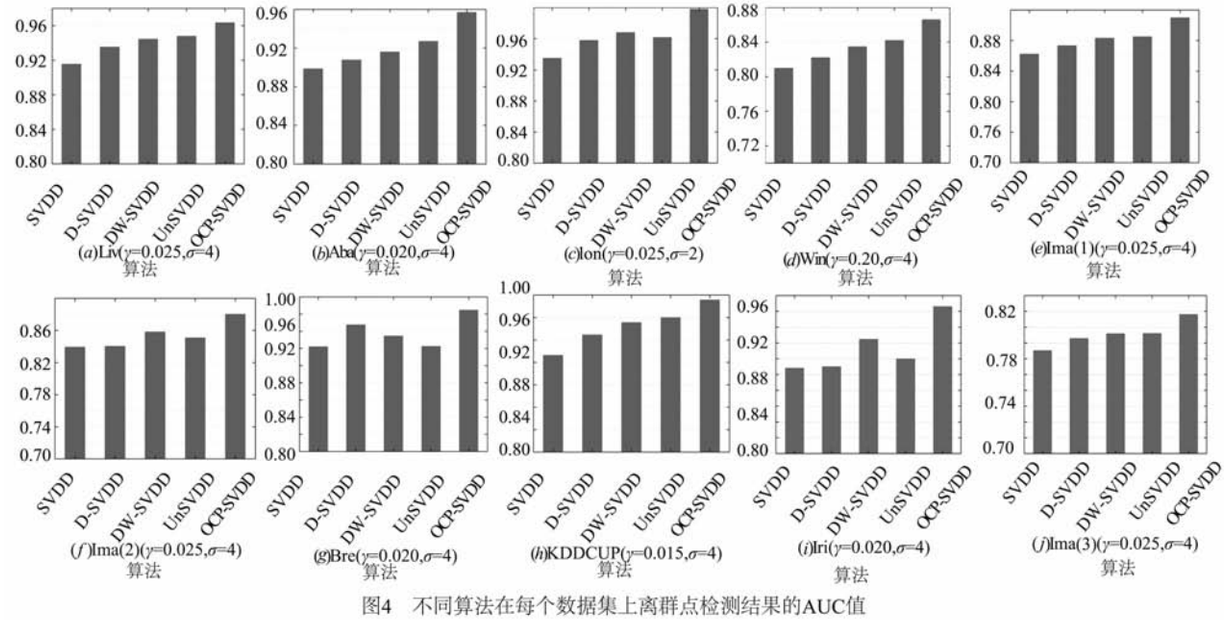

图4　不同算法在每个数据集上离群点检测结果的AUC值

(j)标题中. 同时设置 D-SVDD 算法的权值因子 $\omega = 0.3$,设置 OCP-SVDD 算法的模糊因子 $m = 2.6$. 图 4 表明,OCP-SVDD 算法在 10 个 UCI 数据集的离群点检测结果的 AUC 值均高于其他 4 种对比算法. AUC 值是一种同时评价目标类和非目标类识别能力的指标. OCP-SVDD 算法在识别目标类数据和非目标类数据的能力均高于其他算法. 数据集 Iri 和 Ima(3)的目标类中含有多个子簇,通过 OCP-SVDD 算法可获得较为合理的样本置信度,与其他 4 种算法相比,仍能获得较好的离群点检测效果,说明 OCP-SVDD 算法同样适用于目标类中含有多个子簇的数据集.

3.3.2 参数分析

与经典的 SVDD 算法相比,OCP-SVDD 算法引入了参数 m. 参数 m 为单簇核 PCM 算法的模糊指数,用于调节隶属度的模糊度. 不同的 m 值对样本置信度有一定的影响,进而影响 OCP-SVDD 算法的离群点检测结果. 图 5 所示为不同 m 值下 OCP-SVDD 算法离群点检测结果的平均 AUC 值,可见 m 取值在 2.5 至 4.0 时较为适当.

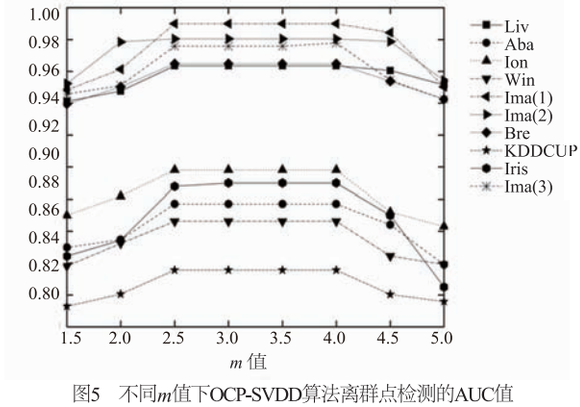

图5 不同m值下OCP-SVDD算法离群点检测的AUC值

3.3.3 运行时间分析

经典的 SVDD 算法运行时间主要消耗在训练阶段中求解线性约束二次优化问题,其时间复杂度为 $O(l^3)$. 与 SVDD 算法相比,D-SVDD 算法和 DW-SVDD 算法需要计算每个样本的邻域,进而计算每个样本的权重,需要多消耗时间 $O(l^2)$. OCP-SVDD 算法首先需要通过计算单簇核 PCM 算法获得样本置信度,较 SVDD 算法需要多消耗的时间复杂度为 $O(l^2)$. 不同算法在 UCI 数据集上的运行时间分析见图 6.

4 结束语

本文提出一种基于单簇核 PCM 的 SVDD 离群点检测算法 OCP-SVDD,该算法利用单簇 PCM 聚类确定训练集的聚类中心,利用隶属度获得每个样本的置信度,将置信度引入训练模型,减弱离群点对决策边界的不

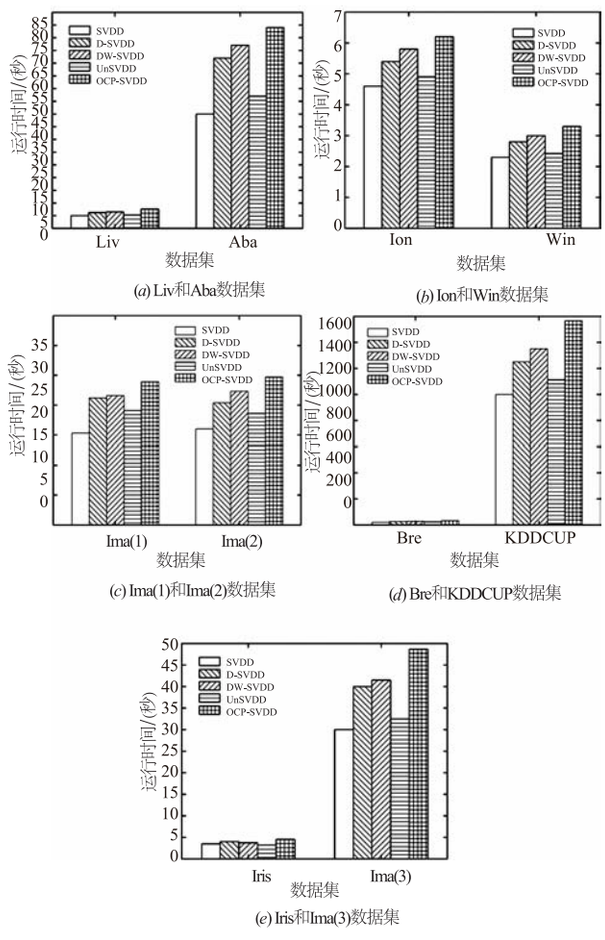

图6 不同算法在UCI数据集上的运行时间

利影响. 实验结果表明,OCP-SVDD 算法对目标类和非目标类的识别能力均高于经典 SVDD 和目前改进的 SVDD 算法,具体包括基于中心点的 UnSVDD 算法、基于密度的 D-SVDD 和 DW-SVDD 算法,说明 OCP-SVDD 能够有效提高 SVDD 的离群点检测能力.

参考文献

[1] Hawkins D M. Identication of Outliers[M]. London:Chapman and Hall,1980.

[2] Hodge V J,Austin J. A survey of outlier detection methodologies[J]. Artificial Intelligence Review,2004,22(2):85 – 126.

[3] Aral K D,Güvenir H A,et al. A prescription fraud detection model[J]. Computer Methods & Programs in Biomedicine,2012,106(1):37 – 46.

[4] 江峰,杜军威,葛艳,等. 基于粗糙集理论的序列离群点检测[J]. 电子学报,2011,39(2):345 – 350.
Jiang F,Du J W,Ge Y,et al. Sequence outlier detection based on rough set theory[J]. Acta Electronica Sinica,

2011,39(2):345 – 350. (in Chinese)
[5] Yang Z, Wang S, Fu X. Pattern recognition-based chillers fault detection method using support vector data description (SVDD) [J]. Applied Energy, 2013, 112 (4): 1041 – 1048.
[6] Shepherd J M, Burian S J. Detection of urban-induced rainfall anomalies in a major coastal city[J]. Earth Interactions, 2002, 7(4): 1 – 17.
[7] Prastawa M, Bullitt E, Ho S, et al. A brain tumor segmentation framework based on outlier detection[J]. Medical Image Analysis, 2004, 8(3): 275 – 283.
[8] Tax D M J, Duin R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11 – 13): 1191 – 1199.
[9] 方景龙,王万良,王兴起,等. 求解多示例问题的支持向量数据描述方法[J]. 电子学报,2013,41(4):763 – 767.
Fang J L, Wang W L, Wang X Q, et al. Support vector data description method for solving multiple instance problems [J]. Acta Electronica Sinica, 2013, 41(4): 763 – 767. (in Chinese)
[10] 胡正平,冯凯. 高维空间多分辨率最小生成树模型的自适应一类分类算法[J]. 自动化学报,2012,38(5):769 – 775.
Hu Z P, Feng K. An adaptive one-class classification algorithm based on multi-resolution minimum spanning tree model in high-dimensional space [J]. Acta Automatica Sinica, 2012, 38(5): 769 – 775. (in Chinese)
[11] Liu B, Xiao Y, Yu P S, et al. An efficient approach for outlier detection with imperfect data labels [J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(7): 1602 – 1616.
[12] Chen G, Zhang X, Wang Z J, et al. Robust support vector data description for outlier detection with noise or uncertain data[J]. Knowledge-Based Systems, 2015, 90(C): 129 – 137.
[13] Lee K, Kim D W, Lee K H, et al. Density-induced support vector data description[J]. IEEE Transactions on Neural Networks, 2007, 18(1): 284 – 289.
[14] Cha M, Kim J S, Baek J G. Density weighted support vector data description [J]. Expert Systems with Applications, 2014, 41(7): 3343 – 3350.
[15] Liu B, Xiao Y, Cao L, et al. SVDD-based outlier detection on uncertain data[J]. Knowledge & Information Systems, 2013, 34(3): 597 – 618.
[16] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. Plenum Press, 1981.
[17] Krishnapuram R, Keller J M. A possibilistic approach to clustering [J]. IEEE Transactions on Fuzzy Systems, 1993, 1(2): 98 – 110.
[18] 陈斌,冯爱民,陈松灿,等. 基于单簇聚类的数据描述[J]. 计算机学报,2007,30(8):1325 – 1332.
Chen B, Feng A M, Chen S C, et al. One-cluster clustering based data description[J]. Chinese Journal of Computers, 2007, 30(8): 1325 – 1332. (in Chinese)
[19] Vapnik V N. The Nature of Statistical Learning Theory [M]. Springer, 2000. 988 – 999.

作者简介

杨金鸿　女,1987 年生于黑龙江哈尔滨. 哈尔滨工程大学计算机科学与技术学院博士研究生. 研究方向为数据挖掘、机器学习以及不确定性理论等.
E-mail: yangjinhong. 66@163. com

邓廷权　男,1965 年生于四川三台. 哈尔滨工程大学理学院教授、博士生导师,研究方向为不确定性理论、数据挖掘和数字图像处理等.