DOI: 10.13382/j.jemi.2016.11.014

一种快速的离群点检测方法

冯 震[1,2] 付敬奇[1] 熊 南[1]

(1. 上海大学 机电工程与自动化学院 上海 200072; 2. 湖北师范大学 机电与控制工程学院 湖北 435002)

摘 要:离群点检测已在许多领域得到了广泛的应用,支持向量数据描述(SVDD)是一种流行的离群点检测方法,但其训练阶段需要二次规划求解,以及决策阶段计算与支持向量数量呈线性关系等导致该方法具有较高时间复杂度。本文提出了一种快速 SVDD 离群点检测方法,首先在训练阶段利用训练集约简和二阶逼近的序列最小优化(SMO)算法降低训练时间,然后在决策阶段通过分析决策函数表达式,利用获取超球球心原像的方式降低决策时间,使得该方法的时间复杂度显著降低。利用标准的公用数据集验证提出的方法,结果表明该方法的时间复杂度明显优于传统的方法。

关键词: 支持向量数据描述;离群点;序列最小优化算法;原像

中图分类号: TP311.13;TN911.23 **文献标识码**: A **国家标准学科分类代码**: 520.60

A novel method for rapidly outlier detection

Feng Zhen[1,2] Fu Jingqi[1] Xiong Nan[1]

(1. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China;
2. College of Mechatronics and Control Engineering, Hubei Normal University, Huangshi 435002, China)

Abstract: Outlier detection is widely used in many fields. Support vector data description (SVDD) method is a popular method for outlier detection. However, SVDD needs to solve quadratic programming problem in the training phase and the time complexity of SVDD is linear in the number of support vectors in the decision-making phase, which lead to the high time complexity of the method. Therefore, a fast SVDD algorithm for outlier detection is proposed in this paper. Firstly, training set reduction strategy and the Sequential Minimal Optimization (SMO) algorithm based on the second order approximation are combined to accelerate the training speed of SVDD. Secondly, through the analysis of the expression decision function, the decision time is reduced by acquiring the pre-image of SVDD hyper-sphere center. Thus the time complexity of the method is significantly reduced. The proposed method is verified by standard public datasets. The experimental results show that the time complexity of the proposed method is obviously better than that of the traditional method.

Keywords: SVDD; outliers; SMO; pre-image

1 引 言

离群点是指某个数据集中出现明显不同于其他对象的数据点,离群点检测是用来确定小部分数据对象与剩余的大部分数据明显不同或者不一致的问题[1]。其广泛的应用于信用卡诈骗、保险诈骗和内幕交易[2]、图像处理[3]、医疗数据的异常检测[4]、网络入侵检测和一些关键安全领域的异常检测[5-6],因此,离群点检测越来越受到国内外学者的关注。

当前,研究者们提出了很多离群点检测方法[7-9]。离群点检测也可认为是一种单类分类问

收稿日期: 2016-07 Received Date: 2016-07

题,也就是通过对正常样本进行学习获取相应的样本模型,然后利用该模型检测任何不同于此模型的离群点。支持向量数据描述(SVDD)是一种有效的单类分类方法[10],并被广泛的应用于离群点检测[11]。对于给定的一组正常样本集,SVDD 是通过寻找一个最小的超球体包围所有或者绝大部分正常样本集,利用获取的超球体边界对未知样本进行检测确定其是否属于离群点。同时,当引入核函数概念后,可以将线性不可分的样本通过核函数映射到高维特征空间,再利用 SVDD 可以获得更灵活的超球体边界适应不同规则形状的目标数据集,从而使得 SVDD 能有效的应用到各种离群点检测的领域[12-14]。但是,在训练阶段,SVDD 在获取超球体边界时需要求解二次规划问题,如果训练样本数为 n,则在此阶段的计算时间复杂度为 $O(n^3)$。在测试阶段,由于引入核函数,SVDD 决策一个未知样本的计算时间复杂度与支持向量数量线性相关。如有大量的样本需要决策时,其时间复杂度将很高。孙提出了一种约减 SVDD 核矩阵尺寸的方法来加快其运算速度,用于飞参数据的新异检测[15]。序列最小优化(SMO)算法是由 platt 提出的求解支持向量机的二次规划问题典型算法。其主要解决 Lagrange 乘子优化和工作集选择策略,其中后者是决定算法收敛速度的关键。文献[16]分析了基于二阶逼近工作集选择策略,该策略能有效的提高 SMO 算法的收敛速度。现有文献主要针对降低训练阶段时间复杂度进行的一些改进工作,很少有文献考虑同时降低训练阶段和决策阶段时间复杂度。

因此,本文从降低训练阶段和决策阶段时间复杂度两个方面出发提出一种快速 SVDD 离群点检测方法。首先利用基于二阶逼近的 SMO 算法降低 SVDD 训练阶段的时间复杂度,同时在此基础上提出了一种通过原像获取的方法,大幅度降低 SVDD 决策阶段的时间复杂度。

2　SVDD 基本原理

对于给定的一组包含 n 个训练数据对象的数据集 $X = \{x_1, x_2, \cdots, x_n\}$, $x_i \in R^d (1 \leqslant i \leqslant n)$, 其中,$d$ 表示数据 x_i 的维度, n 表示训练数据的数量。如果数据是线性不可分的,可通过非线性映射函数 $\phi(\cdot)$ 可将数据映射到高维特征空间得到数据集 $X' = \{\phi(x_1), \phi(x_2), \cdots, \phi(x_n)\}$, 支持向量数据描述算法的主要思想[10]是试图获取一个最小的超球体,使其尽可能的包含训练样本数据集 X' 中所有的正常数据。该问题可表示成式(1)的优化问题。

$$\begin{aligned} &\text{Min} \quad R^2 + C\sum_{i=1}^{n}\varepsilon_i \\ &\text{s.t.} \ \|\phi(x_i) - a\|^2 \leqslant R^2 + \varepsilon_i,\ i = 1,2,\cdots,n \\ &\varepsilon_i \geqslant 0,\ i = 1,2,\cdots,n \end{aligned} \tag{1}$$

式中:R 和 a 分别表示超球体的半径和球心, ε_i 表示松弛变量用于排除位于超球外的异常样本数量[17-18], C 表示惩罚因子,控制错分样本比例和超球体体积的平衡。由 Mercer 定理可知,满足 Mercer 条件的函数称为核函数,在此引入核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j)$ 表示特征空间中两个向量的内积。式(1)是一个带约束的优化问题,这里可以通过引入 Lagrange 乘子解决,该优化问题的 Lagrange 函数如式(2)所示。

$$L(R, \boldsymbol{a}_\phi, \varepsilon_i, \alpha_i, \beta_i) = R^2 + C\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\alpha_i(R^2 + \varepsilon_i - \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}_\phi\|^2) - \sum_{i=1}^{n}\beta_i \tag{2}$$

式中:α_i 和 β_i 为 Lagrange 乘子, $\alpha_i = (\alpha_1, \cdots, \alpha_n) \geqslant 0$, $\beta_i = (\beta_1, \cdots, \beta_n) \geqslant 0$。将 L 分别对 R、$\boldsymbol{a}_\phi$、ε_i 求偏导数并令其为 0,得到方程 $\frac{\partial L}{\partial R} = 0 \rightarrow \sum_{i=1}^{n}\alpha_i = 1$, $\frac{\partial L}{\partial \boldsymbol{a}_\phi} = 0 \rightarrow \boldsymbol{a}_\phi = \sum_{i=1}^{n}\alpha_i\phi(\boldsymbol{x}_i)$, $\frac{\partial L}{\partial \varepsilon_i} = 0 \rightarrow a_i + \beta_i = C$。因 $\alpha_i \geqslant 0$, $\beta_i \geqslant 0$, 且 $0 \leqslant \alpha_i \leqslant C$, 可以省略 β_i 的限制。并结合核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, 可得优化问题(1)的对偶问题如式(3)所示。

$$\begin{aligned} &\text{Max}\sum_{i=1}^{n}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &\text{s.t.} \ \sum_{i=1}^{n}\alpha_i = 1 \\ &0 \leqslant \alpha_i \leqslant C,\ i = 1,2,\cdots,n \end{aligned} \tag{3}$$

式(3)对偶问题的优化解可表示为 $\alpha' = (\alpha'_1, \alpha'_2, \cdots, \alpha'_n)^{\mathrm{T}}$。实际上,优化解 $\boldsymbol{\alpha}'$ 中大部分元素的值等于 0,该元素对应的样本点称为非支持向量(NSVs)。然而,SVDD 超球体的区域一般是由优化解中少量元素 $\alpha'_i > 0 \mid \forall i \in \{1,2,\cdots,l\}$ 对应的样本点确定,这些样本点称为支持向量。其中,元素 $C > \alpha'_i > 0$ 对应的样本点称为边界支持

向量(MSVs),元素 $\alpha'_i = C$ 对应的样本点称为非边界支持向量(NMSVs)。因此,SVDD 超球体的半径可以通过计算超球体球心到任意一个边界支持向量的距离来确定,相应的 SVDD 概略图如图 1 所示。

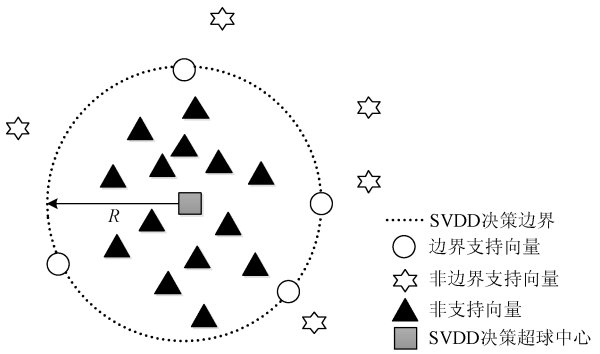

图 1　SVDD 概略图

Fig. 1　Sketch map of SVDD

假设 $\boldsymbol{x}_k$ 是众多支持向量之一,并满足条件 $0 < \alpha'_k < C$,则 SVDD 超球体半径 R 能通过式(4)计算求得。为了判断一个测试样本 x_z 是否属于正常样本,通常采用的方法是通过比较样本 $\boldsymbol{x}_z$ 到球心的距离与超球体半径 R 的大小来确定。当距离小于或等于半径 R 时,则认为样本 x_z 为正常样本,反之,则为异常样本。

$$R^2 = \|\boldsymbol{x}_k - \boldsymbol{a}_\phi\|^2 = K(\boldsymbol{x}_k,\boldsymbol{x}_k) - 2\sum_{i=1}^{n}\alpha_i K(\boldsymbol{x}_k,\boldsymbol{x}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{4}$$

3　改进的 SVDD 算法

SVDD 算法在第 2 节已经提出,它是一种有效的异常检测算法。但是,它在训练和检测阶段具有高的计算复杂度,针对这个问题,提出一种降低 SVDD 计算复杂度的方法。该方法首先降低训练阶段计算复杂度,通过在 SMO 算法中采用二阶逼近方法提高收敛速度,从而提高训练速度,这种方法称为 SMO2-SVDD;其次,在 SMO2-SVDD 基础上,通过在决策阶段提出一种对未知样本的快速决策方法来提高决策速度,这种方法称为 New-SVDD。

3.1　训练集的约简

对于式(3)的对偶问题,其解具有稀疏性的特征。也就是说,最小 SVDD 超球体的范围是由少量支持向量对应的样本点决定,并且不依赖于靠近球体中心的一些样本点。但是,在常规的 SVDD 算法中,超球体的范围确定是通过对整个样本训练集进行训练来获取,这个过程将消耗大量的时间和存储空间,势必延长训练阶段的时间。因此,结合 SVDD 原理和文献[19-20]的相关知识,提出了一种基于欧氏距离的样本约简评估准则。通过对所有训练样本进行评估,去除一定比例位于训练样本中心附近的训练样本。然后将剩余的训练样本进行训练并获取 SVDD 超球体。

给定一组样本集 X,通过非线性函数 $\phi(\cdot)$ 映射到特征空间 F 中,其所有数据的中心点表达式为 $\mu_F = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$,样本点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的欧氏距离可表示为:

$$d_F = \|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\| = \sqrt{K(\boldsymbol{x}_i,\boldsymbol{x}_i) - 2K(\boldsymbol{x}_i,\boldsymbol{x}_j) + K(\boldsymbol{x}_j,\boldsymbol{x}_j)} \tag{5}$$

则在特征空间中,样本点 $\phi(x_i)$ 到样本中心 μ_F 之间的距离可表示为:

$$\|\phi(\boldsymbol{x}_i) - \mu_F\|^2 = K(\boldsymbol{x}_i,\boldsymbol{x}_i) - \frac{2}{n}\sum_{j=1}^{n}K(\boldsymbol{x}_i,\boldsymbol{x}_j) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}K(\boldsymbol{x}_j,\boldsymbol{x}_k) \tag{6}$$

由于式(6)中最后一项为常数,因此可定义 $\theta_F(i)$ 为训练样本的约简评估准则并用式(7)表示。

$$\theta_F(i) = K(\boldsymbol{x}_i,\boldsymbol{x}_i) - \frac{2}{n}\sum_{j=1}^{n}K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{7}$$

式(7)表明,越大的 $\theta_F(i)$ 表示样本点 x_i 离样本集中心 μ_F 的距离越远。可以将所有样本点的 $\theta_F(i)$ 按降序排列,并且提取出前面连续的 τn 个样本进行训练来获取 SVDD 超球体。这里的 τ 为比例约简因子,在保证分类性能的基础上根据经验确定。

在训练阶段,传统的 SVDD 方法需要花费大量的时间训练样本集中的非支持向量,而非支持向量占整个样本集的大部分,所以有必要约简一定量的非支持向量来降低训练时间。由于非支持向量一般位于样本集中心附近的位置,因此一种基于欧式距离的约简策略被提出。该策略的实现是通过计算样本点到样本中心的欧氏距离,根据比例约简因

子 τ 去掉一定比例离样本中心较近的样本点,从而降低样本集中样本点使得训练时间大大降低。

3.2 基于二阶逼近的 SMO 算法

根据 SVDD 算法原理可知,在学习过程中需要解一个二次规划问题(QP 问题)。由于求解 QP 问题的计算量较大,导致利用式(3)求解支持向量需要高的计算复杂度。SMO 算法是用于支持向量机(SVM)的一种极端情形的分解算法,它的工作集大小为 2,其主要思想是对 Lagrange 乘子进行两两地优化,直到所有乘子都满足 KKT(Karush-Kuhn-Tucker)条件为止。SMO 算法主要解决两个 Lagrange 乘子的优化问题和工作集的选择问题,其中工作集的选择是决定算法收敛速度的一个关键问题,基于二阶逼近工作集选择法能有效提高 SMO 算法的收敛速度。SVDD 方法的训练阶段也需要求解一个二次规划问题,为了加快 SVDD 的训练速度,将基于二阶逼近的 SMO 算法应用于 SVDD 降低 QP 问题的计算复杂度,使得 SVDD 成为一种快速有效的异常检测方法。

3.2.1 KKT 条件的判别准则

根据最优化理论,当目标函数中的所有 α_i 都满足 KKT 条件时,即为优化问题的一个解。因此,这里给出 α_i 是否满足 KKT 条件的判别准则,作为迭代停止的条件。重写对偶问题(3)为矩阵表达式如下:

$$\begin{aligned}&\text{Min}\quad f(\alpha)=\alpha^{\mathrm{T}}\boldsymbol{Q}\alpha-\boldsymbol{P}^{\mathrm{T}}\alpha\\&\text{s.t.}\quad \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha}=1,\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha}\geqslant 0,Ce-\alpha\geqslant 0\end{aligned}\tag{8}$$

式中,$\boldsymbol{Q}$ 是 $n\times n$ 矩阵,$Q_{ij}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)$,$\alpha$、$\boldsymbol{P}$、$\boldsymbol{e}$ 是 n 维的列向量,$P_i=K(\boldsymbol{x}_i,\boldsymbol{x}_i)$,$e_i=1$。对式(8)取 Lagrange 函数可表示为:

$$\overline{L}(\boldsymbol{\alpha},\lambda,\mu,b)=f(\alpha)-\lambda^{\mathrm{T}}\alpha-\mu^{\mathrm{T}}(Ce-\alpha)+b(\boldsymbol{e}^{\mathrm{T}}\alpha-1)\tag{9}$$

其中,$\lambda_i\geqslant 0$,$\mu_i\geqslant 0$,$b\geqslant 0$,并且他们都是 Lagrange 乘子。令 $\partial\overline{L}/\partial\boldsymbol{\alpha}=2\boldsymbol{Q}\alpha-P-\lambda^{\mathrm{T}}+\mu^{\mathrm{T}}+be^{\mathrm{T}}=0$,又 $\nabla f(\boldsymbol{\alpha})=2\boldsymbol{Q}\boldsymbol{\alpha}-\boldsymbol{P}$,则对于任意的 α_i,如满足问题(9)的 KKT 条件,相当于满足下列条件:

1) $-f(\alpha)_i\leqslant b$,if $\alpha_i<C$;

2) $-f(\alpha)_i\geqslant b$,if $\alpha_i>0$;

3) $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha}=1$。

于是,可定义如下两个下标集:$I_{\mathrm{up}}(\alpha)=\{t\mid\alpha_t<C\}$,$I_{\mathrm{low}}(\alpha)=\{t\mid\alpha_t>0\}$。取 $i\in I_{\mathrm{up}}(\alpha)$,$j\in I_{\mathrm{low}}(\alpha)$,若 $-\nabla f(\alpha)_i\leqslant-\nabla f(\alpha)_j$ 成立,则表明 α_i 和 α_j 满足问题(9)的 KKT 条件;否则称 α_i 和 α_j 是一对违反 KKT 条件的违反对。令 $m(\alpha)\equiv\max_{i\in I_{\mathrm{up}}(\alpha)}-\nabla f(\alpha)_t$,$M(\alpha)\equiv\min_{I_{\mathrm{low}}(\alpha)}-\nabla f(\alpha)_t$,则当 $m(\alpha)\leqslant M(\alpha)+r$ 时,称 α_i 满足 KKT 条件,是一个可行解。其中 $r\geqslant 0$,在实际应用中考虑为一个很小的训练精度。

3.2.2 基于二阶逼近的工作集选择策略

依据 Zoutendijk 可行方向法,确定可行方向 $\boldsymbol{d}^{\mathrm{T}}\equiv[\boldsymbol{d}_B^{\mathrm{T}},\boldsymbol{0}_N^{\mathrm{T}}]$,选择工作集 α_i 和 α_j 时,为加快收敛速度,需使目标函数 $f(\alpha^q)$ 在可行方向上目标值下降最大。这里 q 为迭代次数,在 $q+1$ 次迭代时,用 $\boldsymbol{\alpha}^q+\boldsymbol{d}$ 代替 α^q,对 $f(\boldsymbol{\alpha}^q+\boldsymbol{d})$ 在 α^q 处进行二阶泰勒展开得:

$$f(\boldsymbol{\alpha}^q+\boldsymbol{d})-f(\boldsymbol{\alpha}^q)=\nabla f(\boldsymbol{\alpha}^q)\boldsymbol{d}+\frac{1}{2}\boldsymbol{d}^{\mathrm{T}}\nabla^2 f(\boldsymbol{\alpha}^q)\boldsymbol{d}=\nabla f(\boldsymbol{\alpha}^q)_B\boldsymbol{d}_B+\frac{1}{2}\boldsymbol{d}_B^{\mathrm{T}}\nabla^2 f(\boldsymbol{\alpha}^q)_{BB}\boldsymbol{d}_B\tag{10}$$

$B(i,j)$ 为工作集,$N=(1,2,\cdots,l)\backslash B$ 为非工作集。为使 $f(\boldsymbol{\alpha}^q)$ 在 $\boldsymbol{d}$ 上下降最大,相当于解最优化问题:

$$\begin{aligned}&\text{Min}Sub(B)=\nabla f(\boldsymbol{\alpha}^q)_B\boldsymbol{d}_B+\frac{1}{2}\boldsymbol{d}_B^{\mathrm{T}}\nabla^2 f(\boldsymbol{\alpha}^q)_{BB}\boldsymbol{d}_B\\&\text{s.t.}\ \boldsymbol{e}^{\mathrm{T}}\boldsymbol{d}_B=0\\&d_t\geqslant 0,\text{if}\ \alpha_t^q=0,t\in B\\&d_t\leqslant 0,\text{if}\ \alpha_t^q=C,t\in B\end{aligned}\tag{11}$$

由 $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{d}_B=0$ 知 $d_i=-d_j$,将其带入目标函数 $sub(B)$ 并整理化简,得:

$$Sub(B)=p_{ij}d_j+\frac{1}{2}\eta_{ij}d_j^2\tag{12}$$

式中:$p_{ij}=-\nabla f(\alpha^q)_i+\nabla f(\alpha^q)_j$,$\eta_{ij}=K_{ii}+K_{jj}-2K_{ij}$,$K_{ij}$ 表示核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$。当 $i\neq j$,且 (i,j) 是一对违反对时,$\eta_{ij}>0$,$p_{ij}>0$,此时 $sub(B)$ 在 $\hat{d}_i=-\hat{d}_j=p_{ij}/\eta_{ij}$ 处取得最小值 $-p_{ij}^2/2\eta_{ij}$。因此,基于二阶逼近的工作集选择策略如下:

1)取 $i\in\text{argmax}_t\{-\nabla f(\alpha^q)_t\mid t\in I_{up}(\alpha^q)\}$;

2)取 $j\in\text{argmin}_t\{-p_{it}^2/\eta_{it}\mid t\in I_{low}(\alpha^q),-\nabla f(\alpha^q)_t<-\nabla f(\alpha^q)_i\}$;

3)返回工作集 $B(i,j)$,如若无违反对,则迭代

结束。

其中，q 表示迭代次数，$p_{it} = -\nabla f(\alpha^q)_i + \nabla f(\alpha^q)_t$，$\eta_{it} = K_{ii} + K_{tt} - 2K_{it}$，$K_{it}$ 表示核函数 $K(x_i, x_t)$。

3.2.3　Lagrange 乘子的优化

首先获取违反 KKT 条件的两个 Lagrange 乘子 α_i^q 和 α_j^q，并将其他 Lagrange 乘子视作常数，对 α_i^q 和 α_j^q 进行优化得到优化值 α_i^{q+1} 和 α_j^{q+1}。根据线性约束条件 $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha} = 1$，可得 $\alpha_i^{q+1} + \alpha_j^{q+1} = \alpha_i^q + \alpha_j^q = \delta$，其中 $\delta = 1 - \sum_{t=1, t\neq i,j}^{n} \alpha_t^q$ 为常数。一般情况下，先计算 α_j^{q+1} 再利用该值计算 α_i^{q+1}。

由于 α_j^{q+1} 的取值范围是 $L < \alpha_j^{q+1} < C$，其中 $L = \max(0, \alpha_i^q + \alpha_j^q - C)$，$H = \min(C, \alpha_i^q + \alpha_j^q)$。因此，优化后的 α_i^q 和 α_j^q 取值为：

$$\alpha_i^{q+1} = \alpha_i^q - (\alpha_j^{q+1} - \alpha_j^q)$$

$$\alpha_j^{q+1} = \begin{cases} H, & \alpha_j^q - p_{ij}/2\eta_{ij} > H \\ \alpha_j^q - p_{ij}/2\eta_{ij}, & L \leqslant \alpha_j^q - p_{ij}/2\eta_{ij} \leqslant H \\ L, & \alpha_j^q - p_{ij}/2\eta_{ij} < L \end{cases}$$

式中：$p_{ij} = -\nabla f(\alpha^q)_i + \nabla f(\alpha^q)_j$，$\eta_{ij} = K_{ii} + K_{jj} - 2K_{ij}$，$K_{ij}$ 表示核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_t)$。

通过二阶逼近的方法训练 SVDD，求得最优解 $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_n^*)^{\mathrm{T}}$ 和相应的支持向量集。为了提高计算精度，在计算超球体半径 R 时采用所有 M 个边界支持向量的平均值来计算，表示如下：

$$R^2 = \frac{1}{M}\sum_{x_k \in MSVs} \|\boldsymbol{x}_k - \boldsymbol{a}_\phi\|^2 = K(\boldsymbol{x}_k, \boldsymbol{x}_k) - \frac{2}{M}\sum_{x_k \in MSVs}\sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}_k, \boldsymbol{x}_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{13}$$

对于给定的未知样本 $\boldsymbol{x} \in R^d$，可以依据下面的决策函数表达式判断样本是否异常：

$$f(\boldsymbol{x}) = R^2 - \|\phi(\boldsymbol{x}) - \boldsymbol{a}_\phi\|^2 \tag{14}$$

本文采用的核函数为高斯函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2h^2)$（$h$ 为高斯核的带宽参数），则式(14)可简化为：

$$f(\boldsymbol{x}) = 2\sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) - v \tag{15}$$

其中，

$$v = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + 1 - R^2 \tag{16}$$

v 是一个可以计算的常量。当 $f(\boldsymbol{x}) \geqslant 0$ 时，该样本为目标样本，否则为异常样本。根据式(15)可知，对于一个未知样本进行决策的计算复杂度为 $O(n)$。但是 $\alpha_i = 0$ 的 Lagrange 乘子是不参与式(15)的计算，只有 $\alpha_i \geqslant 0$ 的 Lagrange 乘子对应的支持向量集(SVs)参与计算，所以决策一个未知样本的计算复杂度 $O(|SVs|)$。如果给定 N 个未知样本，则总的决策计算复杂度为 $O(N|SVs|)$。通常决定超球半径 R 的 SVs 不能太小，否则影响超球体的 R 的精度使得目标样本被错分的可能性加大。故而，当 SVs 和 N 很大时，则决策函数的计算复杂度将增加。因此，本文在前文加快训练速度的基础上，进一步提出一种新方法加快 SVDD 决策复杂度，从而使得 SVDD 的异常检测的性能得到提高。

3.3　SVDD 的决策方法

根据式(14)的决策函数知道，如果 $\boldsymbol{a}_\phi$ 的原像 $\boldsymbol{a}$ 存在，那么 $a_\phi = \phi(\boldsymbol{a})$。则决策函数为：

$$f(x) = R^2 - \|\phi(\boldsymbol{x}) - \phi(\boldsymbol{a})\|^2 = 2K(x, \boldsymbol{a}) - v' \tag{17}$$

式中：$v' = K(\boldsymbol{x}, \boldsymbol{x}) + K(\boldsymbol{x}, \boldsymbol{a}) - R^2$ 是某一常量。从式(17)可以看出，计算 $K(\boldsymbol{x}, \boldsymbol{a})$ 的复杂度为 $O(1)$，也就是决策一个未知样本的复杂度为 $O(1)$，而由式(15)知决策一个未知样本的复杂度为 $O(|SVs|)$，若能在原空间 R 中找到 $\boldsymbol{a}_\phi$ 的原像 $\boldsymbol{a}$，这将使得决策复杂度显著降低。为了获取原像，利用局部线性嵌入的思想[21-22]获取一个众所周知的假设：空间中的点可通过其领域内点的线性组合近似表示。因此给定 $\boldsymbol{a}$ 的某一邻域 δ，则 $\boldsymbol{a} \approx \sum_i \gamma_i\delta_i$，其中 $\delta_i \in \delta$，$\boldsymbol{\gamma} = (\gamma_i, \cdots, \gamma_{|\delta|})$ 为权向量，且 $\gamma_i > 0$ 和 $\sum_i \gamma_i = 1$。又因为 $\boldsymbol{a}$ 在所有样本点中，因此我们可以取相应的领域 δ 为边界支持向量构成的邻域，即 $\delta_i \in MSVs$。则：

$$\hat{\boldsymbol{a}} = \sum_{\boldsymbol{x}_i \in MSVs} \gamma_i \boldsymbol{x}_i \tag{18}$$

现在如何选定权向量 $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{|\delta|})$，使得损失函数 $\|\hat{\boldsymbol{a}} - \boldsymbol{a}\|$ 达到最小值。根据中值定理，

可知：

$$\phi(\hat{\boldsymbol{a}}) \approx \phi(\boldsymbol{a}) + \varphi'(\zeta)(\hat{\boldsymbol{a}} - \boldsymbol{a})$$
$$\Leftrightarrow \phi(\hat{\boldsymbol{a}}) - \phi(\boldsymbol{a}) \approx \phi'(\zeta)(\hat{\boldsymbol{a}} - \boldsymbol{a})$$
$$\Rightarrow \|\phi(\hat{\boldsymbol{a}}) - \phi(\boldsymbol{a})\| \geqslant \|\hat{\boldsymbol{a}} - \boldsymbol{a}\| \min(\phi'(\zeta)) \quad (19)$$

从式(19)可知，为使 $\|\hat{\boldsymbol{a}} - \boldsymbol{a}\|$ 达到最小下界，可以通过使 $\|\phi(\hat{\boldsymbol{a}}) - \phi(\boldsymbol{a})\|$ 达到下界来近似求解。所以，只需构建 $\boldsymbol{\gamma}$ 的累计平方误差(integrated squared error, ISE)，即 $ISE(\boldsymbol{\gamma}) = \|\phi(\hat{\boldsymbol{a}}) - \phi(\boldsymbol{a})\|^2$，并使其尽可能小，则：

$$\hat{\gamma} = \min_{\gamma} ISE(\boldsymbol{\gamma}) = \min\Big(\sum_{\boldsymbol{x}_i \in MSVs}\sum_{\boldsymbol{x}_j \in MSVs} \gamma_i\gamma_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) - 2\sum_{\boldsymbol{x}_i \in MSVs}\gamma_i \sum_{\boldsymbol{x}_j \in SVs}\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) + \sum_{\boldsymbol{x}_i \in SVs}\sum_{\boldsymbol{x}_j \in SVs}\alpha_i\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)\Big) \quad (20)$$

式(20)中的第 3 项独立于 $\boldsymbol{\gamma}$，故式(20)的优化模型也可以表示为：

$$\hat{\gamma} = \max_{\gamma} \sum_{\boldsymbol{x}_i \in MSVs}\gamma_i \sum_{\boldsymbol{x}_j \in SVs}\alpha_j 2K(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_{\boldsymbol{x}_i \in MSVs}\sum_{\boldsymbol{x}_j \in MSVs}\gamma_i\gamma_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)$$
$$\text{s.t. } \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{e} = 1, \gamma_i \geqslant 0, 1 \leqslant i \leqslant |MSVs| \quad (21)$$

式(21)是一个 QP 问题，为了求解上式，采用直接求解方法将式(21)对 γ_k 求偏导数等于 0，可得

$$\frac{\partial ISE(\boldsymbol{\gamma})}{\partial \gamma_k} = 2\sum_{\boldsymbol{x}_j \in SVs}\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) - 2\sum_{\boldsymbol{x}_j \in MSVs}\gamma_k K(\boldsymbol{x}_i,\boldsymbol{x}_j) = 0$$
$$\Rightarrow \gamma_k = \frac{\sum_{\boldsymbol{x}_j \in SVs}\alpha_j K(\boldsymbol{x}_j,\boldsymbol{x}_k)}{\sum_{\boldsymbol{x}_j \in MSVs} K(\boldsymbol{x}_j,\boldsymbol{x}_k)} \quad (22)$$

显而易见，选定的权向量 $\boldsymbol{\gamma}$ 可以通过式(22)求得，从而有效获取原像 $\boldsymbol{a}$。最后，利用式(17)实现决策一个未知样本的决策时间复杂度从 $O(|SVs|)$ 降到 $O(1)$。

New-SVDD 样本决策的实现步骤如下：

1)初始化 SVDD 的核参数和惩罚因子 C；

2)利用前面提出的基于二阶逼近的 SMO 算法解 SVDD 中的 QP 问题；

3)利用式(13)计算超球体半径 R；

4)利用式(22)计算加权因子 $\boldsymbol{\gamma}$；

5)根据式(18)的原像获取方法估计出球心 $\phi(\boldsymbol{a})$ 的原像 $\hat{\boldsymbol{a}}$；

6)根据式(17)实现样本的异常检测。

4　仿真实验

通过仿真实验，对本文提出的离群点检测方法 New-SVDD 与 SMO2-SVDD 和 SVDD 方法进行了比较分析。SMO2-SVDD 方法是通过提高 SVDD 方法的训练阶段速度来提高算法的速度。New-SVDD 方法是在 SMO2-SVDD 方法的基础上通过提高决策阶段的速度来提高算法的速度。仿真实验环境是 Win7 操作系统计算机上的 MATLAB 2013a。本文利用表 1 所示的 UCI 数据集验证 3 种算法的性能，并对实验结果进行了比较分析。UCI 数据集表 1 中的第 1 列是名称，第 2 列维数，第 3 列是目标类，第 4 列目标类对应的样本数。

表 1　UCI 实验数据集

Table 1　Experimental datasets of UCI

数据集	维数	目标类	样本数
Balance Scale(B. S)	4	1	288
		2	49
		3	288
Breast Cancer(B. C.)	30	1	357
		2	212
Wine(W.)	13	1	71
		2	59
		3	48
Iris(I.)	4	1	50
		2	50
		3	50
Liver(L.)	6	1	200
		2	145
Connectionist Bench(C. B.)	60	1	111
		2	97
Spambase(S. B.)	57	1	2 788
		2	1 813
Waveform(W. F.)	21	1	1 696
		2	1 647
		3	1 657
Landsat Satellite(L. S.)	36	1	1 533
		2	1 508
Blood Transfusion Service center(B. S. C.)	4	1	570
		2	178

在所有仿真实验中的核函数采用高斯核函数，高斯核宽度参数 h 和惩罚因子 C 分别从 $\left\{\frac{s^2}{128},\frac{s^2}{64},\frac{s^2}{32},\frac{s^2}{16},\frac{s^2}{8},\frac{s^2}{4},\frac{s^2}{2},s^2,2s^2,4s^2,8s^2\right\}$ 和 $\{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$ 中选择获取，s 表示所有训练样本 2 范数的平均值，这里采用交叉验证

法确定这两个参数。为了对几种算法进行有效的比较,仿真实验采用相同的训练样本和测试样本以及相同的参数。同时,实验过程中使用样本集的目标类 1 的 70% 的样本做为训练样本集,剩余的 30% 的样本和其他类的样本做为测试样本集,也就是把所有其他类的样本认为是离群点。

为了实现合理的对比分析,UCI 数据集被应用到各种算法上时,评估算法的性能指标主要考虑 3 个方面的参数:几何平均数、训练速度、测试速度。几何平均数指标用来评估几种算法的分类精确度,由于该指标同时考虑了目标类和非目标类的分类精度结果,因此在处理不平衡数据集时被广泛的应用。其表达式如下:

$$g-means=\sqrt{Acc^{+}\cdot Acc^{-}} \tag{23}$$

式中:Acc^{+} 和 Acc^{-} 分别表示目标样本和非目标样本的分类精度,表达式为 $Acc^{+}=\frac{\text{被正确分类的目标样本数量}}{\text{所有样本数量}}\times 100\%$,$Acc^{-}=\frac{\text{被正确分类的非目标样本数量}}{\text{所有样本数量}}\times 100\%$。

将 New-SVDD, SMO2-SVDD 和 SVDD 三种算法在相同的训练样本、测试样本和参数条件下分别运行 30 次,通过计算评估性能指标的均值和方差结果对算法性能进行比较。3 种算法在 UCI 数据集上的精度比较体现于表 2,表中列出了 UCI 各个数据集分别运行 3 种算法的几何平均数的均值和方差。从该表可以明显看出,提出的 New-SVDD 方法几何平均数的均值略低于传统 SVDD 方法。例如,在 B. S. 数据集中,改进的 New-SVDD 分类精度为 80.65%,SMO2-SVDD 分类精度为 80.68%,而传统的 SVDD 分类精度为 80.69%;在 B. C. 数据集中,改进的 New-SVDD 分类精度为 81.55%,SMO2-SVD 分类精度为 81.54%,而传统的 SVDD 分类精度为 81.56%。可以看出在大部分数据集上它们的分类精度几乎相同,因此 New-SVDD 方法在提高算法的训练速度后,其分类精度与其他方法具有一定的可比性。

表 2　几何平均数和标准差

Table 2　Average g-means and the standard derivation　　(%)

数据集	New-SVDD		SMO2-SVDD		SVDD	
	几何平均数	标准差	几何平均数	标准差	几何平均数	标准差
B. S.	80.65	0.57	80.68	0.48	80.69	0.45
B. C.	81.55	3.85	81.54	3.94	81.56	3.62
W.	89.49	4.86	89.46	5.14	89.39	4.35
I.	91.75	3.51	91.76	3.64	91.74	3.57
L.	65.52	5.21	65.51	4.86	65.25	5.14
C. B.	58.42	3.45	58.43	3.63	57.97	3.52
S. B.	78.35	1.18	78.32	1.25	78.05	1.02
W. F.	87.41	0.45	87.42	0.62	87.35	0.41
L. S.	90.54	0.65	90.52	0.75	90.12	0.52
B. S. C.	79.28	2.38	79.22	2.82	79.18	2.64

同时为了比较改进的 SVDD 方法在训练阶段和决策阶段的时间复杂度,将 3 种算法分别运行在各个数据集上,得到的训练时间和决策时间展现在表 3 中,从该表可以看出,SVDD 方法的训练时间明显高于 New-SVDD 和 SMO2-SVDD 方法,例如,在B. C. 数据集中,New-SVDD 和 SMO2-SVDD 的训练时间平均值是4.958 7 s,而传统 SVDD 的训练时间的平均值是 14.021 5 s,这是因为前两种方法采用了本文提出的训练集约减和二阶逼近的 SMO 算法训练目标样本集使得训练时间复杂度降低的结果。同时,New-SVDD 方法的决策时间非常短,明显优于 SVDD 和 SMO2-SVDD 方法,例如,在 B. S. 数据集中,New-SVDD 方法的决策时间的平均值是 0.004 5 s,而 SVDD 和 SMO2-SVDD 方法的决策时间的平均值是 1.652 4 s,这是因为前者采用了本文提出的快速决策方法使得算法的决策时间复杂度从 $O(|SVs|)$ 降低到 $O(1)$,使得改进的方法决策时间明显比传统方法短。特别地,在大样本数据

集 S. B. 中,SVDD 方法的训练时间和决策时间的平均值分别是 992.648 和 68.253 6 s;New-SVDD 的训练时间和决策时间的平均值分别是 415.582 7 和 0.045 1 s。结果表明本文提出的 New-SVDD 方法在训练阶段和决策阶段的时间复杂度都有显著提高。

表 3　平均训练时间和决策时间

Table 3　Average training and testing time on the datasets　　(s)

数据集	New-SVDD				SMO2-SVDD				SVDD			
	训练时间		决策时间		训练时间		决策时间		训练时间		决策时间	
	平均值	标准差	平均值	标准差	平均值	标准差	平均值	标准差	平均值	标准差	平均值	标准差
B. S.	0.814 5	0.275 2	0.004 5	0.000 1	0.814 5	0.275 2	1.652 4	0.004 2	1.784 2	0.412 5	1.652 4	0.004 2
B. C.	4.958 7	0.304 2	0.008 5	0.000 1	4.958 7	0.304 2	2.521 0	0.008 8	14.021 5	0.262 5	2.521 0	0.008 8
W.	1.352 4	0.328 5	0.001 5	0.000 0	1.352 4	0.328 5	0.210 0	0.001 7	2.812 4	0.401 6	0.210 0	0.001 7
I.	0.495 2	0.195 4	0.001 2	0.000 0	0.495 2	0.195 4	0.160 0	0.001 2	1.085 1	0.105 2	0.160 0	0.001 2
L.	2.125 2	0.585 4	0.003 5	0.000 1	2.125 2	0.585 4	1.350 0	0.005 7	5.015 1	0.746 2	1.350 0	0.005 7
C. B.	5.563 7	0.528 4	0.002 7	0.000 1	5.563 7	0.528 4	0.842 6	0.002 5	15.418 3	0.621 5	0.842 6	0.002 5
S. B.	415.582 7	49.256 4	0.045 1	0.000 5	415.582 7	49.256 4	68.253 6	0.123 6	992.648 0	78.582 9	68.253 6	0.123 6
W. F.	295.681 7	34.538 2	0.028 9	0.000 3	295.681 7	34.538 2	48.215 4	0.070 5	685.325 1	54.287 6	48.215 4	0.070 5
L. S.	190.523 6	29.634 2	0.019 5	0.000 2	190.523 6	29.634 2	27.254 8	0.185 6	452.368 4	49.528 3	27.254 8	0.185 6
B. S. C.	3.285 3	0.628 5	0.009 2	0.000 1	3.285 3	0.628 5	3.436 2	0.019 5	8.053 4	0.925 1	3.436 2	0.019 5

5　结　论

传统的 SVDD 方法在许多离群点检测问题上的应用取得了好的结果。但是,SVDD 方法也存在两方面的问题导致该方法具有较高的时间复杂度,首先,训练阶段需要求解二次规划问题,本文提出一种 SMO2-SVDD 方法,通过训练集约减和二阶逼近 SMO 算法训练目标样本集,从而降低 SVDD 的训练时间。同时在此基础上,提出一种 New-SVDD 方法,通过获取原像的方法使得单个样本的决策时间由 $O(|SVs|)$ 降到 $O(1)$,从而降低测试阶段的时间。最后,通过 UCI 数据集验证了 SVDD 和 SMO2-SVDD 和 New-SVDD 三种方法。例如在 B. S. 数据集中,改进的 New-SVDD 方法的训练时间平均值为 0.814 5 s,明显低于传统 SVDD 的 1.784 2 s;并且决策时间平均值为 0.004 5 s,明显低于传统 SVDD 的 1.652 4 s。实验结果验证了提出方法的训练阶段和决策阶段时间复杂度明显优于传统 SVDD 方法。

参考文献

[1] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: A survey [J]. ACM Computing Surveys, 2009, 41(3):75-79.

[2] QUAH J T S, SRIGANESH M. Real-time credit card fraud detection using computational intelligence [J]. Expert Systems with Applications, 2008, 35(4): 1721-1732.

[3] 刘松松,张辉,毛征,等. 基于 HRM 特征提取和 SVM 的目标检测方法[J]. 国外电子测量技术, 2014, 33(10): 38-41.
LIU S S, ZHANG H, MAO ZH, et al. Target detection method based on HRM feature extracting and SVM [J]. Foreign Electronics Measurement Technology, 2014, 33(10): 38-41.

[4] 马莉,蒋建波,张磊邦,等. 基于智能终端的心血管疾病诊断设备前端设计[J]. 电子测量技术, 2015, 38(2): 87-90.
MA L, JIANG J B, ZHANG L B, et al. Design of front-end device for CVD diagnosis based on cortex-Ms [J]. Electronic Measurement Technology, 2015, 38(2): 87-90.

[5] 吴新宇,郭会文,李楠楠,等. 基于视频的人群异常事件检测综述[J]. 电子测量与仪器学报, 2014, 28(6): 575-584.
WU X Y, GUO H W, LI N N, et al. Survey on the video-based abnormal event detection in crowd scenes [J]. Journal of Electronic Measurement and Instrumentation, 2014, 28(6): 575-584.

[6] PETERSON G L, MCBRIDE B T. The importance of

generalizability for anomaly detection [J]. Knowledge and Information Systems, 2008, 14(3): 377-392.

[7] ZAMANI M, MOVAHEDI M. Machine learning techniques for intrusion detection [J]. arXiv preprint arXiv:1312.2177, 2013.

[8] DUA S, DU X. Data Mining and Machine Learning in Cybersecurity[M]. BocaRaton: CRC press, 2016.

[9] BUTUN I, MORGERA S D, SANKAR R. A survey of intrusion detection systems in wireless sensor networks[J]. IEEE Communications Surveys and Tutorials, 2014, 16(1): 266-282.

[10] LEE K Y, 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.

[11] WU M R, YE J P. A small sphere and large margin approach for novelty detection using training data with outliers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088-2092.

[12] LIU Y H, LIN S H, HSUEH Y L, et al. Automatic target defect identification for TFT-LCD array process inspection using kernel FCM-based fuzzy SVDD ensemble [J]. Expert Systems with Applications, 2009, 36(2): 1978-1998.

[13] PARK J, KANG D, KIM J, et al. SVDD-based pattern denoising [J]. Neural Computation, 2007, 19(7): 1919-1938.

[14] BANERJEE A, BURLINA P, DIEHL C. A support vector method for anomaly detection in hyperspectral imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(8): 2282-2291.

[15] 孙文柱,曲建岭,袁涛,等. 基于改进SVDD的飞参数据新异检测方法[J]. 仪器仪表学报, 2014, 35(4): 932-939.
SUN W ZH, QU J L, YUAN T, et al. Flight data novelty detection method based on improved SVDD [J]. Chinese Journal of Scientific Instrument, 2014, 35(4): 932-939.

[16] 曾志强,吴群,廖备水,等. 改进工作集选择策略的序贯最小优化算法[J]. 计算机研究与发展, 2009(11): 1925-1933.
ZENG ZH Q, WU Q, LIAO B SH, et al.. An Improved Working Set Selection Strategy for Sequential Minimal Optimization Algorithm [J]. Journal of Computer Research and Development, 2009(11): 1925-1933.

[17] KANG W S, CHOI J Y. Domain density description for multiclass pattern classification with reduced computational load [J]. Pattern Recognition, 2008, 41(6): 1997-2009.

[18] LEE S W, PARK J, LEE S W. Low resolution face recognition based on support vector data description[J]. Pattern Recognition, 2006, 39(9): 1809-1812.

[19] RICO-JUAN J R, INESTA J M. Adaptive training set reduction for nearest neighbor classification [J]. Neurocomputing, 2014, 138(11): 316-324.

[20] ZHU F, WEI J F. A new SVM reduction strategy of large-scale training sample sets [C]. Advanced Materials Research. Trans Tech Publications Ltd, 2013: 512-515.

[21] SIRIPANADORN S, HATTAGAM W, TEAUMROONG N. Anomaly detection in wireless sensor networks using self-organizing map and wavelets [C]. International Conference on Applied Computer Science-Proceedings. World Scientific and Engineering Academy and Society, 2010:381-387.

[22] BRANCH J W, GIANNELLA C, SZYMANSKI B, et al. In-network outlier detection in wireless sensor networks [J]. Knowledge and Information Systems, 2013, 34(1): 23-54.

作者简介

付敬奇,1962年出生,1995年于南京理工大学获得博士学位,现为上海大学教授博士生导师,主要研究方向为无线传感网络仪表智能化网络化等。

E-mail: jqfu@ staff.shu.edu.cn

Fu Jingqi was born in 1962, received Ph. D. from Nanjing University of Science and Technology in 1995. Now, he is a professor and Ph. D. supervisor in Shanghai University. His main research interests include wireless sensor networks, intelligentized and networked instrument.

冯震,1982年出生,在读博士,讲师,主要研究方向为异常检测,机器学习等。

E-mail: fengzhen2003@ shu.edu.cn

Feng Zhen was born in 1982. And now he is Ph. D. candidate in Shanghai University and lecturer in Hubei Normal University. His research interest is outlier detection, and machine learning and so on.

熊南,1987年出生,在读博士,主要研究方向为工业网络化系统分析与控制。

E-mail: nanxiong.shu@gmail.com

Xiong Nan was born in 1987. And now he is Ph. D. candidate in Shanghai University. His main research interest is industrial networked systems analysis and control.