基于简约凸壳的一类模糊支持向量机

周国华[1,2],卢剑炜[1],顾晓清[3],殷新春[2]
(1. 常州工业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;
3. 常州大学信息科学与工程学院,江苏常州 213164)

摘 要: 为解决传统一类支持向量机对噪声数据敏感和不适用于大规模分类等问题,提出了用于大规模噪声环境的基于简约凸壳的一类模糊支持向量机(OC-FSVM-RCH). OC-FSVM-RCH 根据简约凸壳的定义在核空间得到代表正常类数据几何特征的样本,然后基于改进的模糊支持向量域描述算法,使得正常类数据包含在最小超球内,异常数据与超球间隔最大化. OC-FSVM-RCH 剔除正常类数据轮廓边缘处的噪声,同时对数据内部的噪声不敏感. 实验结果表明了所提算法在性能和训练时间上取得了良好的效果.

关键词: 模糊支持向量机; 一类分类; 简约凸壳; 噪声数据
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2019)08-1708-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.08.014

One-Class Fuzzy Support Vector Machine Based on Reduced Convex Hull

ZHOU Guo-hua[1,2], LU Jian-wei[1], GU Xiao-qing[3], YIN Xin-chun[2]
(1. *Department of Information Engineering, Changzhou Institute of Industry Technology, Changzhou, Jiangsu* 213164, *China*;
2. *College of Information Engineering, Yangzhou University, Yangzhou Jiangsu* 225127, *China*;
3. *School of Information Science and Engineering, Changzhou University, Changzhou, Jiangsu* 213164, *China*)

Abstract: The traditional one-class support vector machines are sensitive to noise data and not suitable for large-scale classification. In order to solve the problem, a novel one-class fuzzy support vector machine based on reduced convex hull called OC-FSVM-RCH is proposed for large-scale noise data classification. According to the reduced convex hull, OC-FSVM-RCH obtains the samples representing the geometric characteristics of normal class data in the kernel space. Then OC-FSVM-RCH improves the fuzzy support vector domain description algorithm, in which normal class data is enclosed in the smallest hypersphere, and the margin between abnormal class data and hypersphere is maximized. OC-FSVM-RCH can eliminate the noise at the edge of normal data contour and is insensitive to the noise inside the normal data. Experimental results show that the proposed algorithm achieves good results in terms of performance and training time.

Key words: one-class; fuzzy support vector machine; reduced convex hull; noise data

1 引言

仅使用一类数据建立分类模型的过程称为一类分类,如网络入侵识别、机器故障检测等. 一类分类问题也称为异常检测问题,因为往往正常类数据容易大量获得,异常数据的获取费时费力. 如在机器故障检测中,大多数的数据是在正常工作的条件下获得的,异常工作条件下才能得到的故障数据不多见,且获得故障数据需付出昂贵的代价. 因此,绝大部分的一类分类算法只关注正常类数据,建立的分类器仅对正常类数据进行描述,那些偏离正常数据的样本判定为异常数据. 支持向量机(Support Vector Machine, SVM)因其泛化性能强在一类分类算法中受到了广泛关注[1]. 根据分类面的构建结构,一类 SVM 分为两种:第 1 种是在核空间建立一个超平面,将正常类和异常数据分隔开,如 One-class Support Vector Machine (OCSVM)[2]等;第 2 种是在核空间建立一个封闭紧凑的超球体,使正常类数据尽可能多或全部包括在超球体内部,异常数据没有或尽可能

收稿日期:2018-11-20;修回日期:2019-03-22;责任编辑:马兰英
基金项目:国家自然科学基金(No. 61472343, No. 61806026);江苏省自然科学基金(No. BK20180956);院创新团队项目(No. YB201813101005)

少地包括在超球体内部,如支持向量域描述(Support Vector Data Description,SVDD)[3]等.

随着"大数据"时代的到来,商业经济、工业控制等领域越来越多需要基于数据和分析做出决策,传统的一类 SVM 因时间复杂度为 $O(N^3)$(N 为训练样本数)不适合大规模数据的分类. 为了解决这个难题,SVM 常采用的策略有:(1)使用逼近方法降低二次规划问题的运算量,代表算法有 Nystrøm 算法[4],贪婪逼近[5]和 Sequential Minimal Optimization (SMO)算法[6]等;(2)通过制定筛选策略减少 SVM 的训练数据,代表算法有基于最小包含球的样本筛选[7],凸壳顶点在线分类法(Convex Hull Vertices Selection for Online Classification,CHVS)[8]和凸壳向量机(Convex-Hull Vector Machine,CHVM)[9]. 对于给定集合 $\boldsymbol{X}$,凸壳是包含 $\boldsymbol{X}$ 中所有样本的最小凸集,而 SVM 的分类面与样本分布的几何特性有关,且支持向量大多集中在样本的外部轮廓区域. 因此,基于凸壳的 SVM 能取得了较好的分类性能[9,10]. 由于采集数据环境和手段的不确定性和多样性,现实世界的数据普遍含有噪声,而凸壳的计算受噪声的影响很大. 此外,现有的基于凸壳的 SVM 算法大多研究的是二元分类问题,不能直接应用于一类分类问题.

为解决上述问题,本文提出了一种适用于大规模噪声环境的基于简约凸壳的一类模糊支持向量机(One-class Fuzzy Support Vector Machine based on Reduced Convex Hull,OC-FSVM-RCH). OC-FSVM-RCH 的基本步骤是:首先,使用 BFPRT 算法[11]将正常类数据划分若干个样本子集. 其次,在每个子集中,剔除分布在核空间数据边缘处的噪声样本,得到能表示正常类数据核空间分布的简约凸壳. 然后,在模糊支持向量域描述(Fuzzy Support Vector Data Description,FSVDD)[12]算法上加入极少量的异常样本信息,实现正常数据与异常数据之间的间隔最大化. 本文的贡献有:(1)从大规模数据处理能力来看,OC-FSVM-RCH 采用简约凸壳作为正常类数据的训练样本,有效减少了算法训练时间,实现了快速分类的需要;(2)从抗躁能力来看,OC-FSVM-RCH 使用核空间简约凸壳剔除数据边缘处的噪声样本,同时还继承了模糊支持向量机对噪声数据不敏感的优势;(3)从分类精度来看,OC-FSVM-RCH 得到的简约凸壳能够代表正常类数据的分布特征,并将极少量的异常样本加入到一类分类模型中,实验结果验证了其有效性.

2　相关知识

2.1　模糊支持向量域描述

为了克服噪声对分类面的影响,模糊一类支持向量机赋予每个样本合适的模糊隶属度,来权衡每个样本对于分类模型的重要性. 对噪声赋予较小的模糊隶属度可以削弱噪声对分类模型的作用. 模糊支持向量域描述 FSVDD[12]旨在寻找一个能够包含所有训练样本的体积最小化的超球体,将正常样本和异常样本分离开,超球内样本为正常样本,超球外样本为异常样本. 假设经模糊化的训练样本集合为 $\boldsymbol{X}=\{(\boldsymbol{x}_1,\mu_1),(\boldsymbol{x}_2,\mu_2),\cdots,(\boldsymbol{x}_N,\mu_N)\}$,其中 $\mu_i(0\leqslant\mu_i\leqslant1)$ 为样本 $\boldsymbol{x}_i$ 的模糊隶属度. 引入非线性映射 ϕ 将 $\boldsymbol{X}$ 投影到核空间,FSVDD 的优化问题可以表示为:

$$\begin{aligned}&\min_{R,\boldsymbol{c},\xi} R^2+C\sum_{i=1}^{N}\mu_i\xi_i,\\ &\text{s.t.}\quad \|\phi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2\leqslant R^2+\xi_i,\\ &\xi_i\geqslant0,i=1,2,\cdots,N\end{aligned}\tag{1}$$

其中,R 和 $\boldsymbol{c}$ 分别是超球的半径和球心,C 是惩罚因子,ξ_i 是松弛变量,样本的惩罚系数为 $\mu_i\xi_i$. 式(1)的求解可转化为一个对偶问题,得出 R 和 $\boldsymbol{c}$ 的最优解后,如果测试样本 $\boldsymbol{x}$ 满足式(2),则被判定为异常数据:

$$\|\phi(\boldsymbol{x})-\boldsymbol{c}\|^2>R^2\tag{2}$$

FSVDD 的时间复杂度为 $O(N^3)$,显然它不适用于大规模噪声数据的分类问题.

2.2　简约凸壳

凸壳能刻画数据样本的全局分布. 数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 的凸壳[9,10]定义为:

$$CH(\boldsymbol{X})=\left\{\sum_{i=1}^{N}\lambda_i\boldsymbol{x}_i\,|\,\boldsymbol{x}_i\in\boldsymbol{X},\ \sum_{i=1}^{N}\lambda_i=1,\lambda_i\geqslant0\right\}\tag{3}$$

但凸壳对噪声非常敏感,如果样本集的边界周围存在噪声,凸壳不能准确表示样本集的几何特征. 为此,Theodoridis 等[13]引入简约因子 γ 提出了简约凸壳 $RCH(\boldsymbol{X},\gamma)$:

$$RCH(\boldsymbol{X},\gamma)=\left\{\sum_{i=1}^{N}\lambda_i\boldsymbol{x}_i\,|\,\boldsymbol{x}_i\in\boldsymbol{X},\ \sum_{i=1}^{N}\lambda_i=1,0\leqslant\lambda_i\leqslant\gamma\right\}\tag{4}$$

其中,$0<\gamma\leqslant1$. 图 1 显示了简约凸壳的示意图,γ 的值分别是 $\gamma=1$,$\gamma=0.85$ 和 $\gamma=0.75$. 由式(4)和图 1 看出,$RCH(\boldsymbol{X},\gamma)\subseteq CH(\boldsymbol{X})$,当 $\gamma=1$ 时,$RCH(\boldsymbol{X},\gamma)=CH(\boldsymbol{X})$.

3　基于简约凸壳一类模糊支持向量机(OC-FSVM-RCH)

噪声数据绝大多数都分布在数据集的边界处[14],为了避免噪声数据对计算凸壳的干扰,本文使用简约凸壳计算噪声环境下的凸壳. OC-FSVM-RCH 将正常类样本划分多个子集,分别在每个子集中计算简约凸壳. 另外,OC-FSVM-RCH 基于 FSVDD 算法加入了极少量的异常样本信息,实现正常数据与异常数据之间的间隔最大化. 整个算法分为 3 个阶段:

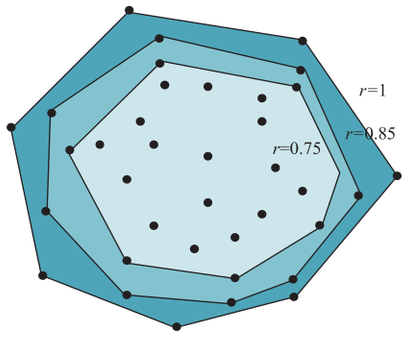

图1 简约凸壳示意图

第1阶段,划分样本子集. OC-FSVM-RCH 按照数据的相似度将正常类数据的训练集分成多个子集,然后分别在每个子集中计算简约凸壳. 划分样本子集常使用聚类法和快速排序法. 聚类法能够根据数据点间的相似性对数据进行有效分组,但聚类法的时间复杂度较高,以层次聚类 Agglomerative[15] 和 Divisive[16] 为例,两者在核空间中的时间复杂度分别是 $O(N^2\log(N))$ 和 $O(2^N)$,显然不适用于大规模数据的场景. 本文使用 BFPRT 算法划分样本子集,单次执行的时间复杂度是 $O(N)$.

假设待划分的每个子集约含 P 个样本,首先随机选择一个样本作为首元素,计算每个样本 $\boldsymbol{x}_i$ 与首元素在核空间中的欧氏距离 d_i,然后根据 d_i 的值搜索样本 $\boldsymbol{x}_k$,将正常类训练集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 分成大致等分的两个子集 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$,其中 $\boldsymbol{X}_1=\{\boldsymbol{x}_i:d_i<d_k,\boldsymbol{x}_i\in\boldsymbol{X}\}$ 和 $\boldsymbol{X}_2=\{\boldsymbol{x}_i:d_i\geq d_k,\boldsymbol{x}_i\in\boldsymbol{X}\}$. 重复执行这一步骤,直至得到 $ceil(N/P)$ ($ceil()$ 表示向上取整)近似等分组.

第2阶段,计算简约凸壳. 首先,OC-FSVM-RCH 使用式(1)在每个正常类样本分组中分别计算核空间内的最小超球体. 依据样本到球心的核空间欧式距离降序排列样本子集内所有样本,选取前20%比例的样本,记为外壳区域 $\tilde{\boldsymbol{X}}$. 然后,将半径为 d 处的样本设为集 $\boldsymbol{Z}^*$,并根据式(4)依次判断降序排列的每个样本 $\boldsymbol{x}_i$($\boldsymbol{x}_i\in\tilde{\boldsymbol{X}}$ and $\boldsymbol{x}_i\notin\boldsymbol{Z}^*$)是否是简约凸壳点:

$$\min_{\mu}\left\|\phi(\boldsymbol{x}_i)-\sum_{t=1}^{|\boldsymbol{Z}^*|}\mu_{i,t}\phi(\boldsymbol{x}_t)\right\|^2,$$
$$\text{s.t.}\quad \phi(\boldsymbol{x}_t)\in\boldsymbol{Z},\ 0\leqslant\mu_{i,t}\leqslant\gamma,\ \sum_{t=1}^{|\boldsymbol{Z}^*|}\mu_{i,t}=1 \tag{5}$$

对式(5)进行整理,舍去常数项,可得如下矩阵形式:

$$\min_{\mu}2\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{Z}^*\boldsymbol{\mu}+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{Z}^{*\mathrm{T}}\boldsymbol{Z}^*\boldsymbol{\mu},$$
$$\text{s.t.}\quad \sum_{t=1}^{|\boldsymbol{Z}^*|}\mu_{i,t}=1,0\leqslant\mu_{i,t}\leqslant\gamma \tag{6}$$

对式(6)求解得到 $\boldsymbol{\mu}$ 的最优解. 如果 $\left\|\phi(\boldsymbol{x}_i)-\sum_{t=1}^{|\boldsymbol{Z}^*|}\mu_{i,t}\phi(\boldsymbol{x}_t)\right\|^2>\varepsilon$,则 $\boldsymbol{x}_i$ 在核空间不能用简约凸壳线性表示,$\phi(\boldsymbol{x}_i)$ 是简约凸壳向量,把 $\boldsymbol{x}_i$ 加入到 $\boldsymbol{Z}^*$ 中;反之,$\phi(\boldsymbol{x}_i)$ 不是简约凸壳向量.

步骤3,构建分类面. 根据文献[17],虽然异常数据的数量有限不足以构成一类,但如果能利用这些异常数据,可以有效提高一类分类模型的性能和泛化能力. 本文在 FSVDD 算法的基础上加入极少量的异常样本,在核空间建立一个封闭紧凑的超球体,使得正常类数据包含在最小超球内,同时异常数据与超球间隔最大化.

假设原始数据集由 N 个正常数据和 m 个极少量的异常样本构成,其中 $m\ll N$. 经过前2个阶段的运行后,正常类数据上得到 N_1 个简约凸壳向量. 改进的 FSVDD 可描述为

$$\min_{R,\rho^2,\xi}R^2-\nu\rho^2+\frac{1}{v_1N_1}\sum_{i=1}^{N_1}\mu_i\xi_i+\frac{1}{v_2m}\sum_{j=N_1+1}^{N_1+m}\mu_j\xi_j,$$
$$\text{s.t.}\quad \|\phi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2\leqslant R^2+\xi_i,\xi_i\geq0,i=1,2,\cdots,N_1$$
$$\|\phi(\boldsymbol{x}_j)-\boldsymbol{c}\|^2\geq R^2+\rho^2-\xi_j,\xi_j\geq0,$$
$$j=N_1+1,N_1+2,\cdots,N_1+m \tag{7}$$

其中,$\boldsymbol{x}_i(i=1,2,\cdots,N_1)$ 是正常数据,$\boldsymbol{x}_j(j=N_1+1,N_1+2,\cdots,N_1+m)$ 是异常数据. $\rho^2\geq0$ 是两数据之间的间隔,v,v_1 和 v_2 均是正常数. 使用拉格朗日函数可得改进的 FSVDD 的最优化问题:

$$\min_{\alpha}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_{i=N_1+1}^{N_1+m}\sum_{j=N_1+1}^{N_1+m}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)$$
$$-\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N_1+m}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=N_1+1}^{N_1+m}\sum_{j=1}^{N_1}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)$$
$$-\sum_{i=1}^{N_1}\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x}_i)+\sum_{j=N_1+1}^{N_1+m}\alpha_jK(\boldsymbol{x}_j,\boldsymbol{x}_j),$$
$$\text{s.t.}\quad 0\leqslant\alpha_i\leqslant\frac{\mu_i}{v_1N_1},i=1,\cdots,N_1,$$
$$0\leqslant\alpha_j\leqslant\frac{\mu_j}{v_2m},j=N_1+1,\cdots,N_1+m,$$
$$\sum_{i=1}^{N_1}\alpha_i=v+1,$$
$$\sum_{j=N_1+1}^{N_1+m}\alpha_j=v \tag{8}$$

其中,α 是拉格朗日因子. 拉格朗日因子大于0的训练样本为支持向量,考虑式(8)中的两个支持向量集合 SV_1 和 SV_2:

$$SV_1=\left\{\boldsymbol{x}_i\,|\,0<\alpha_i\leqslant\frac{\mu_i}{v_1N_1},1\leqslant i\leqslant N_1\right\} \tag{9}$$

$$SV_2=\left\{\boldsymbol{x}_j\,|\,0<\alpha_j\leqslant\frac{\mu_j}{v_2m},N_1+1\leqslant j\leqslant N_1+m\right\} \tag{10}$$

将式(8)的第1个约束条件变成松弛变量等于0的等式,可得超球半径 R:

$$R=\sqrt{\frac{1}{N}\sum_{\boldsymbol{x}_i\in SV_1}\|\phi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2} \tag{11}$$

同理,将式(8)的第 2 个约束条件变成松弛变量等于 0 的等式,可得超球的间隔和球心:

$$\rho^2 = \frac{1}{m}\sum_{\boldsymbol{x}_i \in SV_2} \|\phi(\boldsymbol{x}_i) - \boldsymbol{c}\|^2 - R^2 \tag{12}$$

$$\boldsymbol{c} = \sum_{i=1}^{N_1} \alpha_i \phi(\boldsymbol{x}_i) - \sum_{j=N_1+1}^{N_1+m} \alpha_j \phi(\boldsymbol{x}_j) \tag{13}$$

最终,改进的 FSVDD 的决策函数是:

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

4　讨论

4.1　模糊隶属度的选择

模糊隶属度函数的设计是模糊支持向量机(Fuzzy Support Vector Machine,FSVM)的关键技术. 目前构造模糊隶属度函数的方法很多,大部分的研究是基于样本到类中心的距离度量隶属度. 本文使用较常用的线性模糊隶属度函数[18]和指数模糊隶属度函数[18],分别如式(15)和式(16)所示. 极少数的异常数据的模糊隶属度函数设置为 1.

$$\mu_i = \begin{cases} 1 - \dfrac{\|\phi(\boldsymbol{x}_i) - \phi(\bar{\boldsymbol{x}})\|}{\max\limits_j(\|\phi(\boldsymbol{x}_j) - \phi(\bar{\boldsymbol{x}})\|) + \delta}, & \boldsymbol{x}_i, \boldsymbol{x}_j \in \text{正常数据} \\ 1, & \boldsymbol{x}_i \in \text{异常数据} \end{cases} \tag{15}$$

$$\mu_i = \begin{cases} \dfrac{2}{1 + \exp(\lambda \|\phi(\boldsymbol{x}_i) - \phi(\bar{\boldsymbol{x}})\|)}, & \boldsymbol{x}_i \in \text{正常数据} \\ 1, & \boldsymbol{x}_i \in \text{异常数据} \end{cases} \tag{16}$$

其中,$\bar{\boldsymbol{x}}$ 是正常数据的平均值.

4.2　时间复杂度分析

OC-FSVM-RCH 算法由 3 个阶段构成,时间复杂度也由 3 部分组成. 第 1 阶段使用 BFPRT 算法将正常类样本划分 $ceil(N/P)$ 个子集,时间复杂度为 $O(P \times ceil(N/P))$,其中 P 是每个划分子集包含的样本数. 第 2 阶段的计算量主要在式(6),本文采用 SMO 算法求解,时间复杂度为 $O\left((N/P)\sum_{i=1}^{N/P} A_i^2\right)$,其中 A_i 为第 i 个正常数据子集的简约凸壳容量. 第 3 阶段是使用式(8)进行分类模型的训练,其时间复杂度 $O(N_1^2)$. 将这 3 个阶段的时间相加,OC-FSVM-RCH 的时间复杂度为 $O(P \times ceil(N/P) + \sum_{i=1}^{N/P}(N/P)A_i^2 + N_1^2)$,其值远小于 FSVDD 高达 $O(N^3)$的时间复杂度.

4.3　理论性质

根据 SVM 的间隔误差界理论,松弛变量 $\xi > 0$ 的训练样本称为间隔误差,改进的 FSVDD 有如下定理:

引理 1[2]　SVM 构建一个的分类超平面满足 VC 维上界 $VC \leqslant \min\{\lceil D^2/\Delta^2 \rceil, d\} + 1$,其中 D 是涵盖全部样本的最小超球的直径,Δ 是两类间的最大间隔,d 是样本的维数.

定理 1　改进的 FSVDD 的 VC 维满足:$VC \leqslant \min\{\lceil D^2/(\rho + R)^2 \rceil, d\} + 1$.

定理 2　设 m^+ 和 s^+ 分别是正常类数据的间隔误差数和支持向量数,m^- 和 s^- 分别是异常数据的间隔误差数和支持向量数,改进的 FSVDD 的参数间存在关系:

$$\overline{\mu_m^+} m^+ \leqslant (v+1) v_1 N_1 \leqslant \overline{\mu_s^+} s^+ \tag{17}$$

$$\overline{\mu_m^-} m^- \leqslant v v_2 m \leqslant \overline{\mu_s^-} s^- \tag{18}$$

其中,$\overline{\mu_m^+}$ 和 $\overline{\mu_m^-}$ 分别是正常类和异常数据间隔误差样本的平均模糊隶属度,$\overline{\mu_s^+}$ 和 $\overline{\mu_s^-}$ 分别是正常类和异常数据支持向量的平均模糊隶属度.

证明　由式(8)得到 $\sum_{i=1}^{N_1} \alpha_i = v + 1$. 当 $\xi_i > 0$ 时 $\alpha_i = 0$,可得 $\alpha_i = \dfrac{\mu_i}{v_1 N_1}$对所有正间隔误差成立,则:

$$\frac{\overline{\mu_m^+} m^+}{v_1 N_1} \leqslant \sum_{i=1}^{N_1} \alpha_i = v + 1 \tag{19}$$

另外,由式(8)的第一个约束项可得正常类数据支持向量对 α_i 至多贡献$\dfrac{\mu_i}{v_1 N_1}$,所以:

$$\sum_{i=1}^{N_1} \alpha_i \leqslant \frac{\overline{\mu_s^+} s^+}{v_1 N_1} \tag{20}$$

综合式(19)、(20)可得证式(17),同理可得证式(18).

5　实验结果与分析

5.1　实验设置

实验的内容为:(1)分组参数 P 和简约参数 γ 的选择;(2)与 FSVDD,FOSVM[19]和 FD-SVM-FC[13]进行性能和训练时间的比较. 实验中 OC-FSVM-RCH$_{\text{lin}}$表示 OC-FSVM-RCH 使用线性模糊隶属度函数,OC-FSVM-RCH$_{\text{exp}}$表示 OC-FSVM-RCH 使用指数模糊隶属度函数. 为了说明简约凸壳对算法性能的影响,将实验中所提算法的第 2 阶段根据定义 1 计算每个分组的简约凸壳,其它阶段不变,该方法命名为 CH-FSVM.

实验通过 8 个真实大规模数据集[20](基本信息见表 1)来验证和比较 OC-FSVM-RCH$_{\text{lin}}$和 OC-FSVM-RCH$_{\text{exp}}$的性能. 数据集分别随机选取 90% 正类样本和部分负类样本,使得产生的数据集中 98% 的训练样本属于正常类,2% 的训练样本属于异常数据. 实验中设计了两种不同的加噪模式:(1)参照文献[21]的噪声模式 1,数据集的边界和内部各添加数量是正常类样本 5% 的额外的噪声向量,噪声向量带有均值为 0 且方差为样本特征值的 5% 的高斯白噪声,其中数据的边界根据 SVDD 算法获得;(2)参照文献[22]的噪声

模式 2,随机选择 20% 的样本加入均值为 0 的高斯白噪声,方差为样本特征值的 15%. 噪声模式 2 的强度大于噪声模式 1;噪声模式 1 的噪声数大于噪声模式 2.

参数设置如下: OC-FSVM-RCH$_{lin}$ 和 OC-FSVM-RCH$_{exp}$ 中 P 和 γ 的取值范围分别为 $\{2\times10^3,4\times10^3,6\times10^3\}$ 和 $\{0.85,0.9,0.95\}$, $\varepsilon=10^{-3}$, v 的取值范围为 $\{1,10,\cdots,80\}$, v_1 和 v_2 的取值范围均为 $\{0.001,0.01\}$. 所有分类模型都使用高斯核,核参取值范围为 $\{2^{-4},2^{-3},\cdots,2^2\}$, SVM 分类器中的正则化参数取值为 $\{10^{-3},10^{-2},\cdots,10^3\}$,线性模糊隶属度函数参数 $\delta=0.001$ 和指数模糊隶属度函数参数取值为 $\lambda=\{0.2,0.4,0.6,0.8,1\}$. 参数采取 10 重交叉验证法. 实验采用 G-mean 作为分类效果的评价准则. 实验环境为 2.53-GHz quad-core CPU, 8-GB RAM, Windows 7 系统,所有算法均在 Matlab2016b 下执行.

表 1 数据集的基本信息

数据集	特征数	样本数
Seizure detection (DET)	3	10000
Forest cover type(FOR)	53	20000
Kdd99(KDD)	41	310000
Localization(LOC)	7	164860
Moore-WWW(MOO)	24	245000
Porker(POR)	10	10810
ncRNA(RNA)	8	486201
Seismic(SEI)	60	18686

5.2 算法参数选择

根据文献[23]分析,正则化参数和高斯核参数适合在一定的范围内采用交叉验证法得到,因此本小节对分组参数 P 和简约参数 γ 进行分析. 实验给出了选取不同的 P 和 γ 时各数据集简约凸壳的规模和运行时间,噪声模式 1 和噪声模式 2 情况下的实验结果如表 2 所示. 从表 2 的结果可以看出:

(1)噪声模式 1 额外添加新的噪声向量,噪声模式 2 在原始数据上添加噪声,且两种模式的噪声数和噪声强度均不同. 由表中数据可知,两种噪声模式取得了接近的简约凸壳的规模. 这一结果可以理解为相比噪声模式,简约凸壳的规模与参数 P 和 γ 更密切相关. 这是因为数据集的简约凸壳总规模为各分组简约凸壳数之和,当 P 值较大时,分组数就少,简约凸壳的规模也小. 当 γ 较大时,简约凸壳覆盖的区域较大,简约凸壳的规模自然就大.

(2)尽管两种噪声模式构造方法不同,但两种噪声模式取得了接近的运行时间. 计算简约凸壳的时间随着 P 值的增加而增加. 因为 P 值较大时,每个分组中的样本数较多,每个分组计算简约凸壳的运行时间就较多,全部数据集上计算简约凸壳的运行时间也较多. 同时,计算简约凸壳的时间随着 γ 值的增加而增加. 因为 γ 值较大时,计算简约凸壳的样本数相对较多,而 γ 值较少时,计算简约凸壳的样本数相对较少.

(3)为了平衡简约凸壳规模和运行时间,后续的实验中 P 值固定为 4×10^3. 另外,从表中数据可以看出,简约参数 γ 的值在取值范围 $\{0.85,0.9,0.95\}$ 里变化时,算法的运行时间是相当的. 同时考虑 γ 和噪声强度、简约凸壳规模和运行时间之间的关系,后续的实验中噪声模式 1 的 γ 值固定为 0.95,噪声模式 2 的 γ 值固定为 0.9.

表 2 不同参数 P 和 V 对应的第 1 -2 阶段的运行时间(单位:s)和简约凸壳容量(括号内表示)

数据集:DET	噪声模式 1			噪声模式 2		
P \ γ	0.85	0.90	0.95	0.85	0.90	0.95
2×10^3	0.68(145)	0.70(148)	0.72(155)	0.73(144)	0.77(152)	0.80(160)
4×10^3	0.71(142)	0.77(146)	0.78(151)	0.76(141)	0.80(150)	0.85(163)
6×10^3	0.75(140)	0.85(144)	0.82(150)	0.80(138)	0.90(148)	0.90(155)
数据集:FOR	噪声模式 1			噪声模式 2		
P \ γ	0.85	0.90	0.95	0.85	0.90	0.95
2×10^3	3.36(2768)	3.40(2850)	3.46(2980)	3.39(2846)	3.43(2875)	3.46(2908)
4×10^3	3.58(2476)	3.60(2525)	3.64(2588)	3.53(2675)	3.58(2765)	3.60(2810)
6×10^3	3.76(2137)	3.80(2168)	3.86(2297)	3.70(2557)	3.74(2254)	3.82(2289)
数据集:KDD	噪声模式 1			噪声模式 2		
P \ γ	0.85	0.90	0.95	0.85	0.90	0.95
2×10^3	139.61(6001)	148.33(6045)	159.46(6100)	138.45(6078)	146.67(6100)	150.99(6187)
4×10^3	168.20(5832)	185.43(5928)	197.47(5977)	162.36(5813)	178.49(5903)	194.13(5932)
6×10^3	184.63(5575)	203.68(5618)	209.90(5703)	180.53(5641)	210.50(5622)	218.43(5900)

续表

数据集:LOC	噪声模式 1			噪声模式 2		
P \ γ	0.85	0.90	0.95	0.85	0.90	0.95
2×10^3	15.14(3167)	15.19(3228)	15.26(3401)	15.06(3172)	15.18(3256)	15.23(3375)
4×10^3	15.70(3105)	15.89(3211)	15.97(3269)	15.80(3134)	16.00(3211)	15.85(3304)
6×10^3	16.88(3013)	17.00(3133)	17.28(3178)	16.67(3075)	16.79(3104)	17.21(3198)
数据集:MOO	噪声模式 1			噪声模式 2		
P \ γ	0.85	0.90	0.95	0.85	0.90	0.95
2×10^3	176.36(12235)	177.01(12383)	178.00(12402)	178.04(12278)	179.97(12351)	179.99(12438)
4×10^3	178.41(11564)	178.76(11843)	178.99(11907)	178.79(11698)	180.31(11854)	182.65(11896)
6×10^3	180.91(11069)	183.45(11000)	187.03(11000)	179.82(11103)	184.76(11007)	189.23(11000)
数据集:POR	噪声模式 1			噪声模式 2		
P \ γ	0.85	0.90	0.95	0.85	0.90	0.95
2×10^3	0.84(245)	0.87(256)	0.90(260)	0.80(244)	0.83(257)	0.86(268)
4×10^3	1.00(227)	1.12(249)	1.17(256)	0.96(236)	1.07(242)	1.08(249)
6×10^3	1.16(223)	1.19(230)	1.26(248)	1.04(219)	1.08(237)	1.16(244)
数据集:RNA	噪声模式 1			噪声模式 2		
P \ γ	0.85	0.90	0.95	0.85	0.90	0.95
2×10^3	139.55(3024)	142.26(3149)	146.67(3212)	137.56(3028)	138.96(3123)	141.45(3205)
4×10^3	143.65(2878)	150.92(2900)	152.50(3044)	141.38(2954)	148.56(2964)	164.34(3021)
6×10^3	145.29(2809)	152.64(2866)	156.58(2906)	154.37(2802)	158.59(2877)	165.70(2899)
数据集:SEI	噪声模式 1			噪声模式 2		
P \ γ	0.85	0.90	0.95	0.85	0.90	0.95
2×10^3	1.25(356)	1.30(378)	1.35(388)	1.31(371)	1.33(390)	1.38(400)
4×10^3	1.50(332)	1.54(356)	1.59(368)	1.48(347)	1.62(358)	1.66(382)
6×10^3	1.80(315)	1.89(320)	2.00(327)	1.70(312)	1.76(320)	1.89(328)

5.3　对比实验

本小节进行了 OC-FSVM-RCH$_{lin}$、OC-FSVM-RCH$_{exp}$ 与 FSVDD,FOSVM,FD-SVM-FC 和 CH-FSVM 的性能比较,实验首先对比了算法在噪声模式 1 情况下的 G-mean 值和各算法的训练时间,实验结果分别图 2 和表 3 所示.

从图 2 实验结果可以看出,OC-FSVM-RCH$_{lin}$ 和 OC-FSVM-RCH$_{exp}$ 在 8 个大规模噪声数据集上取得了令人满意的 G-mean 结果. OC-FSVM-RCH$_{lin}$ 在 SEI 数据集上最优,OC-FSVM-RCH$_{exp}$ 在其它 7 个数据集上取得了最佳值. 这是因为指数模糊隶属度可以通过 λ 参数调节每个样本的模糊隶属度来权衡对分类模型的贡献. FOSVM 是快速一类 SVM 分类算法,没有处理噪声数据的能力,分类效果相对较差. FSVDD,FD-SVM-FC 和 CH-FSVM 均属于模糊 SVM,使用模糊隶属度赋予样本不同的权重,能削弱了噪声样本对分类器的作用. 但这 3 个算法的 G-mean 结果均低于 OC-FSVM-RCH$_{lin}$ 和 OC-FSVM-RCH$_{exp}$. 这是因为:(1)简约凸壳能够表现噪声环境下数据在核空间的分布;(2)使用简约凸壳作为训练样本对分类器的精度有提升的作用. 噪声按几何分布

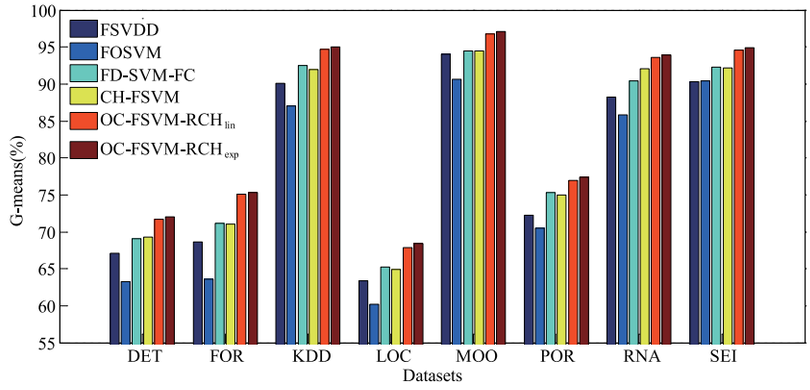

图2　各算法在噪声模式1情况下的G-mean值比较(%)

可分成2种:①分布在样本的边缘;②分布在样本的内部区域. OC-FSVM-RCH$_{lin}$和OC-FSVM-RCH$_{exp}$算法通过简约凸壳剔除样本边缘的噪声数据,同时继承了FSVM的优势,能够对噪声赋予较小的模糊隶属度而削弱其分类作用;(3) OC-FSVM-RCH$_{lin}$和OC-FSVM-RCH$_{exp}$算法改进FSVDD模型,将正常数据的简约凸壳和极少量的异常样本共同参与到模糊超球分类模型的构建中,实现正常数据与异常数据之间的间隔最大化. 因此,OC-FSVM-RCH$_{lin}$和OC-FSVM-RCH$_{exp}$算法比4种对比算法的分类精度高.

从表3实验结果可以看出,OC-FSVM-RCH$_{lin}$和OC-FSVM-RCH$_{exp}$的训练时间明显优于其它算法. 因为OC-FSVM-RCH$_{lin}$和OC-FSVM-RCH$_{exp}$算法仅模糊隶属度函数不同,因此两者训练时间相近. FSVDD算法在所有数据集上的训练时间均最长,而FOSVM在大规模一类分类问题中也不具有优势. CH-FSVM算法的第1阶段和第3阶段与本文所提算法相同,因此这两个算法的训练时间是相似的. FD-SVM-FC使用聚类加采样的方法对正常类数据的训练集进行筛选,最终训练分类器的数据小于原始数据的规模. 因此CH-FSVM、FD-SVM-FC、OC-FSVM-RCH$_{lin}$和OC-FSVM-RCH$_{exp}$均能运用于大规模噪声环境的一类数据分类问题.

其次,实验首先对比了各算法在噪声模式2情况下的G-mean值和训练时间,实验结果分别图3和表4所示. 噪声模式2的强度大于噪声模式1,但从图3实验结果可以看出,随着噪声强度的提高,OC-FSVM-RCH$_{lin}$和OC-FSVM-RCH$_{exp}$的G-mean值下降得最少,OC-FSVM-RCH$_{exp}$的G-mean值略高于OC-FSVM-RCH$_{lin}$. FOSVM的G-mean值下降得最多,因为FOSVM对噪声很敏感,特别是数据边界点处的噪声对分类模型的影响较大. FSVDD、FD-SVM-FC和CH-FSVM通过模糊隶属度赋予样本不同的权重,对噪声样本较为不敏感.

表3 各算法在噪声模式1情况下的训练时间和标准差的比较(单位:s)

	FSVDD	FD-SVM-FC	FOSVM	CH-FSVM	OC-FSVM-RCH$_{lin}$	OC-FSVM-RCH$_{exp}$
DET	12.80 ±0.19	2.27 ±0.16	1.96 ±0.05	1.08 ±0.05	**1.04 ±0.03**	1.05 ±0.02
FOR	50.15 ±0.64	18.36 ±0.17	10.52 ±0.07	6.64 ±0.07	**5.72 ±0.05**	5.75 ±0.04
KDD	2078.36 ±2.36	276.32 ±1.97	294.17 ±1.13	236.54 ±0.70	232.17 ±0.80	**229.32 ±0.72**
LOC	1007.27 ±5.69	118.30 ±1.35	89.05 ±0.67	**20.32 ±0.49**	20.32 ±0.47	20.39 ±0.78
MOO	1270.34 ±4.45	540.15 ±2.09	570.71 ±2.11	254.54 ±1.89	**238.25 ±1.38**	**236.72 ±1.37**
POR	36.97 ±0.43	2.67 ±0.08	3.54 ±0.06	2.14 ±0.08	**1.89 ±0.05**	1.93 ±0.06
RNA	2008.18 ±5.44	210.58 ±2.05	288.09 ±1.16	204.11 ±0.69	192.35 ±0.46	**194.18 ±0.49**
SEI	27.71 ±0.39	9.69 ±0.09	8.59 ±0.09	2.22 ±0.07	**2.02 ±0.06**	**2.04 ±0.05**

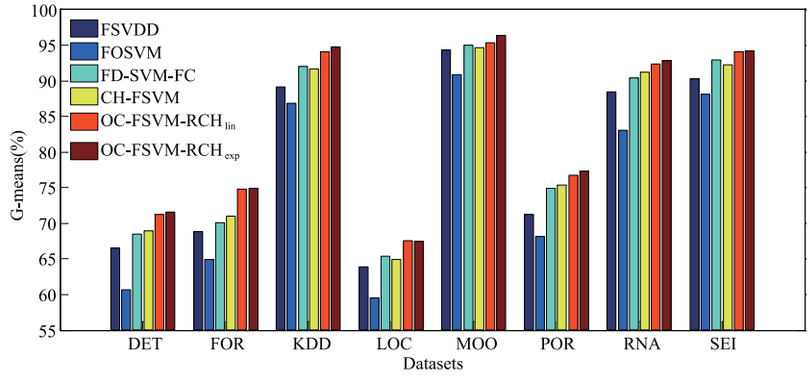

图3 各算法在噪声模式2情况下的G-mean值比较(%)

表4实验结果显示各分类算法的训练时间与表3的结果相似,说明不同噪声模式下算法的训练时间是接近的,也就是说噪声强度的变化与训练时间关联不大. 6个比较算法中OC-FSVM-RCH$_{lin}$和OC-FSVM-RCH$_{exp}$的训练时间接近,都较另4种对比算法少. FSVDD因为时间复杂度高训练时间最长.

表 4　各算法在噪声模式 2 情况下的训练时间和标准差的比较(单位:s)

	FSVDD	FD-SVM-FC	FOSVM	CH-FSVM	OC-FSVM-RCH$_{lin}$	OC-FSVM-RCH$_{exp}$
DET	13.28 ±0.22	2.32 ±0.15	1.85 ±0.06	1.09 ±0.04	**1.05 ±0.04**	1.07 ±0.05
FOR	52.34 ±0.30	18.46 ±0.27	10.90 ±0.07	5.90 ±0.08	**5.81 ±0.07**	5.80 ±0.08
KDD	1950.79 ±2.45	279.62 ±1.60	285.41 ±1.48	245.24 ±0.25	**232.25 ±0.54**	233.01 ±0.24
LOC	1018.65 ±4.67	117.98 ±1.51	87.01 ±0.66	**20.39 ±0.53**	20.71 ±0.48	20.68 ±0.49
MOO	1259.29 ±4.51	546.46 ±2.45	550.23 ±2.37	245.71 ±1.69	240.22 ±1.37	**239.12 ±1.32**
POR	38.25 ±0.76	3.01 ±0.17	3.64 ±0.07	1.74 ±0.05	1.82 ±0.04	**1.80 ±0.03**
RNA	2077.49 ±6.04	204.28 ±0.80	303.11 ±1.40	206.27 ±0.75	**198.99 ±0.72**	201.43 ±0.70
SEI	28.89 ±0.51	9.91 ±0.08	8.28 ±0.07	2.32 ±0.09	**2.10 ±0.08**	2.11 ±0.07

6　总结

本文提出了适用于大规模噪声环境的基于简约凸壳的一类模糊支持向量机 OC-FSVM-RCH. 算法分为 3 个阶段:核空间内数据分组,基于简约凸壳的样本选择和改进 FSVDD 的训练. 简约凸壳能够准确地表示数据在噪声环境下的几何分布,且能有效减少训练样本数提高算法的时间效率. 但是,本研究现阶段只能处理大规模静态数据的单分类问题,还不能应用到云计算等新技术场景下的动态数据. 如何处理大规模面向动态的带噪声的流数据是本研究下一阶段的研究重点. 另外,对于模糊隶属度的选择没有深入探讨,如何根据算法使用的实际场景选择合适的模糊隶属度也是值得研究的课题.

参考文献

[1] 杜栋栋,任星彰,陈坤,等. 一种基于 One-Class SVM 和 GP 安全事件关联规则生成方法研究[J]. 电子学报,2018,46(8):1793 – 1803.
Du Dong-dong, Ren Xing-zhang, Chen Kun, et al. A security event correlation rule generation method research based on one-class SVM and genetic programming [J]. Acta Electronica Sinica, 2018, 46(8): 1793 – 1803. (in Chinese)

[2] Burges C J C. Atutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery,1998,2(2):955 – 974.

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

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

[5] Drineas P, Mahoney M W. On the Nystrøm method for approximating a gram matrix for improved kernel-based learning[J]. Journal of Machine Learning Research,2005,6:2153 – 2175.

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

[7] 张景祥,王士同. 基于共同决策方向矢量的多源迁移及其快速学习方法[J]. 电子学报,2015,43(7):1349 – 1355.
Zhang Jing-xiang, Wang Shi-tong. Common-decision-vector based multiple source transfer learning classification and its fast learning method[J]. Acta Electronica Sinica,2015,43(7):1349 – 1355. (in Chinese)

[8] Ding S, Nie X, Qiao H, et al. A fast algorithm of convex hull vertices selection for online classification [J]. IEEE Trans. on Neural Networks and Learning Systems,2018,29(4):792 – 806.

[9] Gu X, Chung F L, Wang S. Fastconvex-hull vector machine for training on large-scale ncRNA data classification tasks [J]. Knowledge-Based Systems,2018,151(6):149 – 164.

[10] 顾晓清,倪彤光,姜志彬,等. 面向大规模噪声数据的软性核凸包支持向量机[J]. 电子学报,2018,46(2):347 – 357.
Gu Xiao-qing, Ni Tong-guang, Jiang Zhi-bin, et al. Soft kernel convex hull support vector machine for large scale noisy datasets[J]. Acta Electronica Sinica,2018,46(2):347 – 357. (in Chinese)

[11] 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.

[12] Forghani Y, Yazdi H S, Effati S. An extension to fuzzy support vector data description (FSVDD)[J]. Pattern Analysis & Applications,2012,15(3):237 – 247.

[13] Theodoridis S, Mavroforakis M. Reducedconvex hulls: a geometric approach to support vector machines[J]. Signal Processing Magazine IEEE,2007,24(3):119 – 122.

[14] Almasi O N, Rouhani M. Fast and de-noise support vector machine training method based on fuzzy clustering method for large real world datasets[J]. Turkish Journal of Electrical Engineering & Computer Sciences,2016,24(1):219

－233.
[15] Bouguettaya A, Yu Q, Liu X. Efficient agglomerative hierarchical clustering[J]. Expert Systems with Applications, 2015, 42(5): 2785－2797.
[16] Székely G J, Rizzo M L. Hierarchical clustering via joint between-within distances: extending ward's minimum variance method[J]. Journal of Classification. 2005, 22(2): 151－183.
[17] Wu M, Ye J. Asmall sphere and large margin approach for novelty detection using training data with outliers[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088－2092.
[18] Batuwita R, Palade V. FSVM-CIL: fuzzy support vector machines for class imbalance learning[J]. IEEE Trans. Fuzzy Systems, 2010, 18(3): 558－571.
[19] Le T, Phung D, Nguyen K. Fastone-class support vector machine for novelty detection[A]. Pacific-Asia Conference on Knowledge Discovery and Data Mining[C]. Vietnam: Springer International Publishing, 2015. 189－200.
[20] Bache K, Lichman M. UCI database[J/OL]. https://archive. ics. uci. edu/ml/datasets. html, 2018-02-28.
[21] 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.
[22] Luukka P, Lampinen J. Differential evolution classifier in noisy settings and with interacting variables[J]. Applied Soft Computing Journal, 2011, 11(1): 891－899.
[23] Ni T G, Gu X Q, Wang J, et al. Scalable transfer support vector machine with group probabilities[J]. Neurocomputing, 2018, 273(17): 570－582.

作者简介

周国华(通信作者) 男,1977 年出生,江苏东台人,硕士,现常州工业职业技术学院信息工程系讲师,主要研究领域为智能学习,模式识别.
E-mail: tiddyddd@ sina. com. cn

卢剑炜 男,1982 年出生,江苏靖江人,硕士,现常州工业职业技术学院信息工程系副教授,主要研究领域为智能信息处理.
E-mail: ljw@ czili. edu. cn

顾晓清 女,1981 年出生,江苏常州人,博士,现常州大学信息科学与工程学院硕士生导师,研究方向为模式识别,模糊系统.

殷新春 男,1962 年出生,江苏姜堰人,博士,教授,现扬州大学博士生导师. 研究方向为信息安全,软件质量保障、高性能计算.