ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 4,Jul. 2019,pp. 723－735
DOI:10. 16337/j. 1004-9037. 2019. 04. 018
© 2019 by Journal of Data Acquisition and Processing
http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

针对不平衡数据的PSO-DEC-IFSVM分类算法

魏建安　黄海松　康佩栋

(贵州大学现代制造技术教育部重点实验室,贵阳,550025)

摘　要: 针对不平衡数据集下,传统的模糊支持向量机(Fussy support vector machine,FSVM)算法分类效果不够明显,引入的参数未做优化等缺点,本文提出一种基于粒子群算法(Particle swarm optimization,PSO)优化的改进模糊支持向量机算法,即PSO-DEC-IFSVM算法。该算法首先综合考虑训练样本到其类中心的间距、样本周围的紧密度以及样本的信息量设计模糊隶属度函数,然后将此改进的模糊支持向量机与不同惩罚因子(Different error costs, DEC)算法相结合得到DEC-IFSVM算法,最后利用粒子群算法对DEC-IFSVM算法引入的参数进行优化。实验证明:对于UCI公共数据集中的Pima等6种不平衡数据集,相比已有的FSVM及其改进算法,PSO-DEC-IFSVM算法具有更好的正负类分类效果以及更强的鲁棒性。

关键词: 不平衡数据分类;改进模糊支持向量机;样本信息量;粒子群算法;参数寻优

中图分类号: TP181　　**文献标志码:** A

PSO-DEC-IFSVM Classification Algorithm for Unbalanced Data

Wei Jianan, Huang Haisong, Kang Peidong

(Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang, 550025, China)

Abstract: For the unbalanced datasets, the traditional fuzzy support vector machine (FSVM) algorithm classification effect is not obvious, and the introduced parameters are not optimized. Therefore, this paper proposes an improved fuzzy support vector machine (IFSVM) algorithm based on particle swarm optimization (PSO) algorithm, i.e. PSO-DEC-IFSVM algorithm. First, the algorithm is used to design fuzzy membership function considering the distance from training sample to its center, the tightness around the sample and the amount of information of the sample, and then IFSVM algorithm is combined with different error costs (DEC) algorithm for obtaining the DEC-IFSVM algorithm. Finally the PSO algorithm is used to optimize the introduced parameters in the DEC-IFSVM algorithm. Experiments show that the PSO-DEC-IFSVM algorithm has better positive and negative classification effect and stronger robustness than the existing FSVM algorithm and its improved algorithm for the six unbalanced data sets, such as Pima in UCI public data set.

Key words: unbalanced data classification; improved fuzzy support vector machine; sample information; particle swarm optimization; parameter optimization

---

基金项目: 贵州工业攻关重点(黔科合GZ字[2015]3009)资助项目;贵州省自然科学基金(黔科合J字[2015]2043)资助项目;贵州省重大专项(黔科合JZ字[2014]2001)资助项目;贵州省教育厅(黔教合协同创新字[2015]02)资助项目;贵州大学研究生创新基金(研理工2017037)资助项目。

收稿日期: 2017-07-03;**修订日期:** 2017-09-08

引 言

随着大数据时代的到来,信息量激增,由此产生大量的不平衡数据集,即数据集中某类样本数远小于其他类的样本数,其中样本数较少的类叫做正类,样本数较多的类称为负类。不平衡数据的分类作为数据挖掘与机器学习的重要研究内容,近年来越来越多的国内外学者对其进行了大量的研究[1-3],并将其广泛应用于故障诊断、医疗诊断及信用卡欺诈[4-8]等领域。

在众多机器学习算法中,支持向量机(Support vector machine, SVM)算法是依据统计学习中VC维理论以及结构风险最小化等原则而提出的一种学习方法,能够有效地处理小样本、非线性与高维度等问题,且作为一种有效的分类算法,已经获得广泛的应用。但传统SVM对原始数据的处理是基于样本集是平衡的,即正负类样本的数目相同。显然,对于不平衡数据传统SVM算法的分类效果并不理想,这是因为当数据集不平衡时实际分类超平面会向少数类方向偏移,从而导致少数类样本的识别率变低。目前,对于传统SVM算法可以从以下两个方面进行改进以获得更加理想的分类效果:(1)重构原始数据集,即通过过(欠)采样方式分别对正负类样本集进行重构,常见的方式有:对于过采样有基于SMOTE(Synthetic minority oversampling technique)的过采样方式及其改进算法等[9-10],对于欠采样方式有随机欠采样以及基于样本特性的欠采样等[11-12]。但是实际上以上方法是通过一定的准则通过增加或者减少原始数据集的样本数来调节数据集本身的不平衡性,具有随机性较大、盲目性较高、稳定性较差等缺点,且当数据集严重失衡时,所利用的采样方法可能效果不佳。(2)改进的SVM算法,即针对正负类样本数目上的差异,通对算法本身的改进,以增强算法本身对不平衡数据的适应性。常见的改进算法有:不同惩罚因子(Different error costs, DEC)算法及其改进算法通过正负类样本赋予不同的惩罚因子以提高分类的准确性[13-14];模糊支持向量机(Fuzzy support vector machine, FSVM)及其改进算法通过将模糊数学和支持向量机相结合以克服噪声或野点对支持向量的影响来提高分类的准确性[15-17];此外,还有在赋予不同的惩罚因子的同时,增加新的约束条件的近支持向量机法等[18]。

因模糊支持向量机在处理不平衡数据时有较好的表现,故本文选取FSVM进行不平衡数据的分类。现阶段比较典型的模糊支持向量机的改进方式有:李苗苗等[19]在设计模糊隶属度函数时考虑了每个样本点到类型中心距离的同时还考虑到了该样本点最邻近的K个其他样本点的距离。Batuwita等[20]将模糊支持向量机与DEC算法进行结合提出一种FSVM-CIL算法,用于处理不平衡数据以及噪声样本,该算法在设计模糊隶属函数时与传统FSVM类似,仅考虑样本到类中心的距离;鞠哲等[21]在设计FSVM的模糊隶属度函数时考虑样本到类中心距离的同时还考虑到了样本周围的紧密度,并将FSVM与DEC有机地结合,即DEC-FSVM-Ju算法。但是鞠哲等的算法存在以下缺点:(1)算法复杂程度增加,同时未对增加的参数合理优化;(2)没有考虑到样本特性的影响;(3)优化效果不明显。针对上述算法的缺点,本文在设计模糊隶属度函数时考虑样本到类中心距离以及样本周围紧密度的同时,还考虑到了样本信息量特性的影响并赋予不同样本不同的权值,此外将改进的FSVM算法(Improved fuzzy support vector machine, IFSVM)与DEC算法进行结合,并应用粒子群算法(Particle swarm optimization, PSO)对该改进算法引入的参数进行寻优,得到PSO-DEC-IFSVM算法。最后将PSO-DEC-IFSVM算法应用于UCI机器学习数据库中的6类不同的不平衡数据集中。实验证明:本文所提算法相对于已有算法在处理含有噪声的不平衡数据集分类时具有更好的分类效果。本文结果为不平衡数据的分类提供了一个有效的理论模型。

1 算法简介

1.1 传统SVM算法

以传统二分类为例,SVM的基本原理为:从样本(或者核)空间内寻求一个最优分类超平面,使得正

负类样本分隔间距达到最大化。假定给定训练集为$(X,Y)=\{(x_i,y_i)\},y_i\in\{-1,1\},i\in 1,2,3,\cdots,n$,其中:$x_i,y_i$分别为训练集的第$i$个样本以及样本的标签。在SVM算法中引入核函数($K$)将训练集引入高维空间,即$K(\boldsymbol{x},\boldsymbol{y})=\varphi(\boldsymbol{x})^{\mathrm{T}}\varphi(\boldsymbol{y})$,其中$\varphi(\boldsymbol{x})$为非线性映射;同时引入松弛变量$\xi_i\geqslant 0,i=1,2,3,\cdots,n$与惩罚因子$C$,综上,给出标准的支持向机一般形式为

$$\min\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{n}\xi_i^2$$
$$\text{s.t. } y_i(\boldsymbol{\omega}^{\mathrm{T}}\varphi(x_i)+b)\geqslant 1-\xi_i \quad \xi_i\geqslant 0,i=1,2,3,\cdots,n \tag{1}$$

对于式(1)的优化求解,可引入Largrandge乘子法转化为对偶形式,即

$$\min\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jK(x_i,x_j)-\sum_{i=1}^{n}\alpha_i$$
$$\text{s.t. } \sum_{i=1}^{n}y_i\alpha_i=0 \quad 0\leqslant\alpha_i\leqslant C,\ \ i=1,2,3,\cdots,n \tag{2}$$

假定对偶问题的最优解为α^*,则可反求出数据集最优分类超平面的法向量$\boldsymbol{\omega}^*$与截距b^*,其解法如式(3,4)所示,最终利用传统SVM方法构造出如式(5)所示的决策函数。

$$\boldsymbol{\omega}^*=\sum_{i=1}^{n}\alpha^*y_i\varphi(x_i) \quad i=1,2,3,\cdots,n \tag{3}$$

$$b^*=y_i-\sum_{i=1}^{n}y_i\alpha_i^*K(x_i,x_j) \quad 0<\alpha_i^*<C,i=1,2,3,\cdots,n \tag{4}$$

$$f(\boldsymbol{x})=\sum_{i=1}^{n}y_i\alpha_iK(x_i,\boldsymbol{x})+b \tag{5}$$

1.2 FSVM算法与DEC算法的结合算法

事实上,实际生产生活中的数据集往往是不平衡的,相比传统SVM算法分配给每一个样本相同的权值,FSVM算法和DEC算法相结合的DEC-FSVM算法根据样本的不平衡性以及重要性分配不同的权值,以提高分类的准确率。同上,对于二分类而言,假定给定训练集为$(X,Y)=\{(x_i,y_i)\},y_i\in\{-1,1\},i\in 1,2,3,\cdots,n$,另假定原始数据集中有$m$个样本为正类样本(即$y_i=1,i=1,2,3,\cdots,m$),则剩余的$n-m$个样本为负类样本(即$y_i=-1,i=m+1,m+2,m+3,\cdots,n$),则用于不平衡数据分类的模糊支持向量机的一般形式如式(6)所示。

$$\min\frac{1}{2}+C^{\mathrm{p}}\sum_{i=1}^{m}S_i^{\mathrm{p}}\xi_i^2+C^{\mathrm{n}}\sum_{i=1+m}^{n}S_i^{\mathrm{n}}\xi_i^2$$
$$\text{s.t. } y_i(\boldsymbol{\omega}^{\mathrm{T}}\varphi(x_i)+b)\geqslant 1-\xi_i \quad \xi_i\geqslant 0,i=1,2,3,\cdots,n \tag{6}$$

式中:$C^{\mathrm{p}},C^{\mathrm{n}}$分别代表正负类样本的惩罚因子,以表示两类间的不平衡性;$S_i^{\mathrm{p}},S_i^{\mathrm{n}}$分别代表正负类样本的隶属度函数,以反映该样本在其所属类别中的重要性。从式(6)可以看出相对于传统SVM算法,DEC-FSVM从惩罚因子与隶属度函数的引入方向作了改进,这将更有利于不平衡数据的分类。

2 改进的模糊支持向量机(IFSVM)工作机理

2.1 模糊隶属度函数的设计

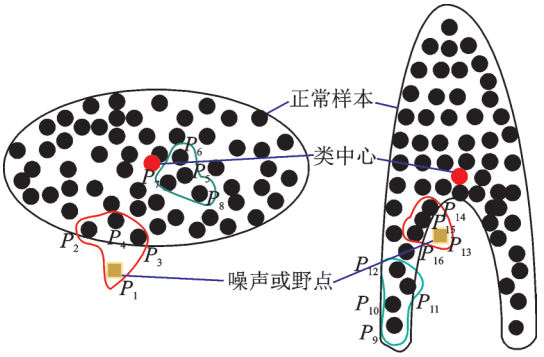
Lin等[15]提出将样本到其类中心的距离作为衡量样本重要性的指标。即将模糊隶属度函数定义为

$$S_i^{\mathrm{p}}=1-\frac{d_i^{\mathrm{cen^p}}}{\max(d_i^{\mathrm{cen^p}})+\delta} \quad i=1,2,3,\cdots,m \tag{7}$$

$$S_i^{\mathrm{n}}=1-\frac{d_i^{\mathrm{cen^n}}}{\max(d_i^{\mathrm{cen^n}})+\delta}\quad i=1+m,2+m,3+m,\cdots,n \tag{8}$$

$$d_i^{\mathrm{cen^p}}=\left\|x_i-\frac{1}{m}\sum_{i=1}^{m}x_i\right\|,d_i^{\mathrm{cen^n}}=\left\|x_i-\frac{1}{n-m}\sum_{i=1+m}^{n}x_i\right\| \tag{9}$$

式中：$d_i^{\mathrm{cen^p}},d_i^{\mathrm{cen^n}}$分别代表正负类的第$i$个样本到其类中心的距离；$\delta$为引入的一个非常小的正数，用来保证隶属度为正。但是当数据集分布不规则时，运用该方式很可能将噪声或野点作为正常的正负类样本进行训练，最终导致算法的整体分类精度降低。如图1(a)数据集1所示，假设P_1为一噪声点，对于正常样本集(以负类为例)来说仅考虑样本到类中心的距离时P_1将被当做正常点进行训练赋予正常隶属度函数值，显然是不合理的。



(a) 数据集1 (a) Data set 1　　(b) 数据集2 (b) Data set 2

图1 不同数据集下噪声点与正常样本的位置关系

Fig.1 Relationship between the noise points and the normal samples under different data sets

针对上述问题，文献[21]中提出在设计模糊隶属度函数时需综合考虑样本到类中心的间距及其周围的紧密度，且其紧密度的衡量方式应用K-近邻域准则，即如图1(a)所示：在图中拟取$K=3$，对于负类样本来说对于噪声点P_1的距离最近的3-近邻域点集为$\{P_2,P_3,P_4\}$，负类样本的任一正常样本P_5的距离最近的3-近邻域点集为$\{P_6,P_7,P_8\}$。显然，负类的正常样本点P_5的3-近邻域点集的距离均值大于噪声点P_1的3-近邻域点集的距离均值，故文献[21]引入式(10,11)定义样本周围的紧密度为

$$D_i^{\mathrm{p}}=\frac{1}{K}\sum_{x_j\in N_K^{\mathrm{p}}(x_i)}\left\|x_i-x_j\right\|\quad i=1,2,3,\cdots,m \tag{10}$$

$$D_i^{\mathrm{n}}=\frac{1}{K}\sum_{x_j\in N_K^{\mathrm{n}}(x_i)}\left\|x_i-x_j\right\|\quad i=1+m,2+m,3+m,\cdots,n \tag{11}$$

式中：$N_K^{\mathrm{p(n)}}(x_i)$为正(负)类的第i个样本的K-近邻域的集合，显然如果某样本的$D_i^{\mathrm{p(n)}}$值越小则该样本属于该正(负)类可能性越大。反之如果该样本为噪声或者野点$D_i^{\mathrm{p(n)}}$的值将会较大，故将模糊隶属度函数定义如下

$$S_i^{\mathrm{p}}=(1-\alpha\frac{d_i^{\mathrm{cen^p}}}{\max(d_i^{\mathrm{cen^p}})+\delta}-(1-\alpha)\frac{D_i^{\mathrm{p}}-\min(D_i^{\mathrm{p}})}{\max(D_i^{\mathrm{p}})-\min(D_i^{\mathrm{p}})+\delta})^M\quad i=1,2,3,\cdots,m \tag{12}$$

$$S_i^{\mathrm{n}}=(1-\alpha\frac{d_i^{\mathrm{cen^n}}}{\max(d_i^{\mathrm{cen^n}})+\delta}-(1-\alpha)\frac{D_i^{\mathrm{n}}-\min(D_i^{\mathrm{n}})}{\max(D_i^{\mathrm{n}})-\min(D_i^{\mathrm{n}})+\delta})^M\quad i=1+m,2+m,3+m,\cdots,n \tag{13}$$

式中：α为一个权值，用于均衡样本到类中心与样本的近邻域密度重要性，故对于不同数据集，α($\alpha\in\{0,0.1,0.2,\cdots,1\}$)值合理的选取极为重要；$\delta$的意义同上；$M$($M\in\{0.1,0.2,0.3,\cdots,1\}$)用于调整所有样本模糊隶属度函数的范围，故值的选取亦较为重要；此外，对于样本K-近邻域中的K值，为了简单起见，文献[21]在隶属度函数设计时将所有样本取为同一值，但是由图1(a, b)可以看出，对于1,2两种不同的数据集，如果K值同时取为一定值是不合理的，对于数据集1来说K取为3是合理的，但对于数据集2，假设P_{13}为一噪声点，对于负类样本来说距离噪声点P_{13}最近的3-近邻域点集为$\{P_{14},P_{15},P_{16}\}$，距离负类样本的一正常样本P_9最近的3-近邻域点集为$\{P_{10},P_{11},P_{12}\}$。显然，负类的正常样本点P_5的

3-近邻域点集的距离均值小于噪声点P_{13}的3-近邻域点集的距离均值,在这种情况下,噪声样本P_{13}会被当作正常的负类样本进行处理,这将会在较大程度上影响分类精度。

综上,对于以上所提的α,M,K等参数在利用DEC-IFSVM进行分类时均要进行优化,参数优化将于2.4节进行介绍。

2.2 FSVM算法的改进

当样本分布不规则时,前文提到文献[21]对FSVM算法改进时仅考虑到引入样本的紧密度来设计模糊隶属度函数,而没有考到样本本身的特性。众所周知:在运用传统SVM分类器进行分类时,分类超平面的确定只与支持向量有关,且SVM算法是通过分类间隙的最大化来设计分类超平面,以期获取较好的推广能力。同时文献[12]中提到:样本的信息量,即样本点到决策面的距离是判断该点性质的主要因素,且距离越近对分类超平面的影响越大。故本文在设计模糊隶属度函数时需要对信息量大的样本点赋予较大的隶属度函数值。据此,本文引入如式(14)所示的样本信息量的评价方式。

$$\varphi(x_i^{\mathrm{p(n)}})=-\left\|w^*\cdot x_i^{\mathrm{p(n)}}+b\right\| \quad i\in 1,2,3,\cdots,n \tag{14}$$

式中$\varphi(x_i^{\mathrm{p(n)}})$为第$i$个正(负)类样本信息量。图2为某数据不平衡下的理想超平面与实际超平面的位置示意图。从图2可以看出:对于理想分类超平面,正负类样本中的支持向量都是距离超平面很近的的点,故拥有最大的信息量;而对于偏移过后的分类超平面,正类样本的支持向量为距离分类超平面较远的样本点,负类的支持向量不变仍然为距离超平面较近的点。故运用传统支持向量机进行分类时,由于分类超平面发生严重偏移,正类样本$\varphi(x_i^{\mathrm{p}})$信息量越小,相应的样本信息量越大;反之负类样本$\varphi(x_i^{\mathrm{n}})$信息量越大时相应的样本信息量越大。另$\omega^*$与$b$分别代表传统SVM的分类平类超平面的法向量与阈值,故改进后的FSVM的隶属度函数如式(15,16)所示。

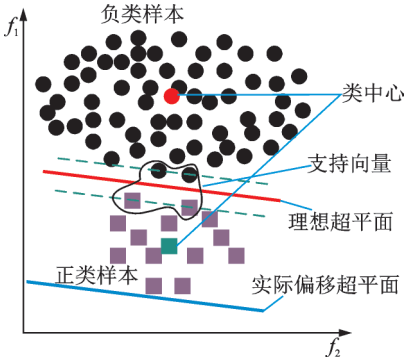


图2 数据不平衡下的理想超平面与实际超平面的位置

Fig.2 Ideal hyperplane and the position of actual hyperplane under data imbanlance

$$S_i^{\mathrm{p}}=\frac{\max(\varphi_i^{\mathrm{p}})-\min(\varphi_i^{\mathrm{p}})}{\dfrac{1}{\varphi_i^{\mathrm{p}}-\min(\varphi_i^{\mathrm{p}})+1}+\delta}*\left(1-\alpha\frac{d_i^{\mathrm{cen^p}}}{\max(d_i^{\mathrm{cen^p}})+\delta}-(1-\alpha)\frac{D_i^{\mathrm{p}}-\min(D_i^{\mathrm{p}})}{\max(D_i^{\mathrm{p}})-\min(D_i^{\mathrm{p}})+\delta}\right)^M$$

$$i=1,2,3,\cdots,m \tag{15}$$

$$S_i^{\mathrm{n}}=Q*\frac{\max(\varphi_i^{\mathrm{n}})-\min(\varphi_i^{\mathrm{n}})}{\varphi_i^{\mathrm{n}}-\min(\varphi_i^{\mathrm{n}})+\delta}*\left(1-\alpha\frac{d_i^{\mathrm{cen^n}}}{\max(d_i^{\mathrm{cen^n}})+\delta}-(1-\alpha)\frac{D_i^{\mathrm{n}}-\min(D_i^{\mathrm{n}})}{\max(D_i^{\mathrm{n}})-\min(D_i^{\mathrm{n}})+\delta}\right)^M$$

$$i=1+m,2+m,3+m,\cdots,n \tag{16}$$

式(15)中:φ_i^{p}为第i个正类样本的信息量,乘号(*)右边部分考虑了样本到类中心的距离及样本紧密度两个因素,而乘号(*)左边为正类样本信息量影响的表达式。上文提到运用传统支持向量机进行不平衡数据分类时,由于分类超平面发生严重偏移,正类样本$\varphi(x_i^{\mathrm{p}})$信息量的值越小相应的样本信息量越大,故引入式(15)用于满足此规律,最终S_i^{p}即为正类样本基于改进的模糊支持向量机的隶属度。同样地,在式(16)中:φ_i^{n}为第i个负类样本的信息量,乘号(*)右边部分亦考虑了样本到类中心的距离及样本紧密度两个因素,乘号(*)左边为负类样本信息量影响的表达式。同样上文提到运用传统支持向

量机进行不平衡数据分类时，由于分类超平面发生严重偏移，负类样本$\varphi(x_i^{\mathrm{p}})$信息量的值越大时相应的样本信息量越大，故引入式(16)用于满足此规律，最终S_i^{n}即为负类样本基于改进的模糊支持向量机的隶属度。

另外，由于利用式(15,16)求正负类样本隶属度时，两式信息量影响的表达式不同，所以需引入平衡因子Q来保证正负类隶属度值范围一致。其算法为：正类所有训练样本的信息量影响值的均值除以负类所有训练样本的信息量影响值的均值所得到，表达式为

$$Q=\left(\sum_{i=1}^{m}\frac{\max(\varphi_i^{\mathrm{p}})-\min(\varphi_i^{\mathrm{p}})}{\dfrac{1}{\varphi_i^{\mathrm{p}}-\min(\varphi_i^{\mathrm{p}})+1}+\delta}*(n-m)\right)\Bigg/\left(\sum_{i=m+1}^{n}\frac{\max(\varphi_i^{\mathrm{n}})-\min(\varphi_i^{\mathrm{n}})}{\varphi_i^{n}-\min(\varphi_i^{\mathrm{n}})+\delta}*m\right) \tag{17}$$

2.3 DEC-IFSVM惩罚因子的设置

众所周知，DEC算法通过赋予正负类样本不同的惩罚因子来提高SVM算法对不平衡数据适应性，对于正类样本赋予较大的惩罚因子，而负类样本赋予较小的惩罚因子。故本文提出DEC协同IFSVM优化算法，既有模糊支持向量机处理噪声(野点)的优势，又可以容易应对不平衡数据。基于样本特性的IFSVM的基本原理与算法上文已作阐述，对于惩罚因子的确定，文献[21-22]采取正负类样本比值的设定方式，且有较好分类效果，故本文亦采取此方式，即正负类的惩罚因子的算法为：$C^{\mathrm{p}}=C(n-m)/m$，$C^{\mathrm{n}}=C$，其中：C^{p}为正类的惩罚因子；C^{n}为负类的惩罚因子；n为训练样本总数；m为训练样本中正类样本的个数；C为惩罚因子的初始参数且$C>0$。

综上，改进的DEC-IFSVM算法的对偶形式为

$$\min\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jK(x_i,x_j)-\sum_{i=1}^{n}\alpha_i;\ \text{s.t.}\ \sum_{i=1}^{n}y_i\alpha_i=0$$

$$0\leqslant\alpha_i\leqslant C_i^{\mathrm{p}}S_i^{\mathrm{p}}\quad i=1,2,3,\cdots,m;\ 0\leqslant\alpha_i\leqslant C_i^{\mathrm{n}}S_i^{\mathrm{n}}\quad i=1+m,2+m,3+m,\cdots,n \tag{18}$$

最终的则决策函数为

$$f(\boldsymbol{x})^{\text{DEC-IFSVM}}=\sum_{i=1}^{n}y_i\alpha_i{}^{\text{DEC-IFSVMK}}(x_i,\boldsymbol{x})+b^{\text{DEC-IFSVM}} \tag{19}$$

2.4 基于PSO算法的DEC-IFSVM参数优化

综合分析上文可知，运用DEC-IFSVM算法进行不平衡数据分类时，在算法复杂度增加的同时，为了得到更加良好的分类效果，需要对引入的α,δ,M,C,K等参数进行优化以及初值赋予。此外本文采用径向基(Radial basis function，RBF)核函数，故核函数中的参数g亦需要进行优化。

在上述需要进行优化的参数中：δ的初值赋予需要多次实验进行择优选取，而K,α,M,C,g五个参数拟利用PSO算法进行优化。

2.4.1 PSO算法简介

PSO算法是受鸟类捕食时搜索附近食物区域行为的启发，将问题的潜在解用不同的粒子来表示，寻找合适的适应度函数来确定各粒子的适应度。另外，PSO算法是一种并行的随机搜索算法，可以实现对解空间的搜索，同时，PSO算法具有控制参数最少、算法简单等优点，一经提出便得到广泛应用[7]。

2.4.2 参数优化

本文以不平衡数据分类效果的评价机制作为目标函数，K,α,M,C,g作为待求粒子，本文实验中采取十折交叉验证，对每一折的参数均进行优化。假定待求解的种群大小为N，迭代代数为G，$P_i(i\in1,2,3,\cdots,N)$表示种群中i个体的位置，$V_i(i\in1,2,3,\cdots,N)$与$\text{fitness}_i(i\in1,2,3,\cdots,N)$分别

代表i个体的速度与适应度值,故本文所采用的粒子群算法的求解步骤如下:(1)算法开始;(2)种群的初始化:包括粒子的位置P_i与速度V_i的随机初始化;(3)个体适应度值:根据目标函数来计算粒子的适应度值fitness_i;(4)循环迭代:在循环迭代过程中,寻找个体的极值P_{best}以及整个群的极值G_{best};(5)算法终止:在满足最优解的条件下,终止循环。

2.4.3 优化结果

由上文可知,DEC-IFSVM引入的参数值需要进行优化,本文选取UCI数据集中的Pima等6种数据集,每个数据集进行十折交叉验证,由于每一折正负类样本数目不同,故需要对每一折的参数进行优化。最终,经粒子群算法优化后的K,α,M,C,g五个参数在不同数据集的最优参数如表1所示。

表1 PSO优化后DEC-IFSVM的最优参数

Tab. 1 Optimized parameters of DEC-IFSVM after PSO optimization

Dataset	NO.	α,M,C,g,K	Dataset	NO.	α,M,C,g,K
Pima	1	0.5,0.9,1.26,0.01,3	Wpbc	1	0.4,0.5,0.01,69.0,6
	2	0.6,0.9,1.14,0.01,3		2	0.4,0.9,0.01,6.075 9,6
	3	0.7,0.9,1.90,0.01,3		3	0.6,0.9,0.01,72.75,6
	4	0.7,0.9,1.12,0.01,3		4	0.6,0.8,0.01,50.046 8,6
	5	0.7,0.9,1.31,0.01,3		5	0.95,0.8,0.01,100,6
	6	0.9,0.9,1.33,0.01,3		6	0.95,0.9,0.01,79.93,6
	7	0.9,0.9,1.12,0.01,3		7	0.95,0.9,0.01,2.534 3,6
	8	0.5,0.9,0.98,0.01,3		8	0.98,0.9,0.01,33.03,6
	9	0.4,0.9,1.21,0.01,3		9	0.9,0.5,0.01,87.699 5,6
	10	0.3,0.9,1.97,0.01,3		10	0.4,0.95,0.01,100,6
Haberman	1	0.1,0.7,15.441 3,0.01,3	Yeast	1	0.5,0.6,65.29,33.67,4
	2	0.5,0.5,0.638 0,0.01,3		2	0.5,0.6,18.96,49.74,4
	3	0.5,0.7,2.819,0.01,3		3	0.4,0.5,100,0.11,4
	4	0.5,0.7,2.815 1,0.01,3		4	0.4,0.5,9.13,0.26,4
	5	0.4,0.5,0.01,55.73,3		5	0.4,0.5,64.14,1.20,4
	6	0.1,0.6,4.964 8,0.01,3		6	0.6,0.5,7.95,1.70,4
	7	0.2,0.6,90.51,0.003 9,3		7	0.5,0.5,9.25,0.70,4
	8	0.1,0.5,1,0.031 25,3		8	0.5,0.5,1.80,2.95,4
	9	0.3,0.6,0.71,0.007 8,3		9	0.5,0.5,61.31,0.01,4
	10	0.4,0.6,1,0.015 6,3		10	0.5,0.5,100,0.01,4
German	1	0.1,0.9,59.20,92.48,5	Abalone	1	0.9,0.9,65.29,33.67,5
	2	0.9,0.5,2.05,0.01,5		2	0.9,0.9,64,0.03125,5
	3	0.9,0.6,1.95,0.01,5		3	0.1,0.5,45.25,0.003 9,5
	4	0.9,0.7,2.79,0.01,5		4	0.1,0.5,58.33,0.005 5,5
	5	0.95,0.8,0.01,100,5		5	0.4,0.5,25.6,0.50,5
	6	0.4,0.95,1.53,0.01,5		6	0.1,0.6,72.24,1.14.99,5
	7	0.95,0.9,3.14,0.01,5		7	0.1,0.6,47.32,10.87,5
	8	0.5,0.95,28.44,9.63,5		8	0.1,0.5,52.70,1.82,5
	9	0.5,0.9,0.01,61.42,5		9	0.1,0.6,9.7584,32.44,5
	10	0.4,0.95,0.01,90.11,5		10	0.1,0.5,114.33,18.23,5

3 实验与结果分析

3.1 不平衡数据分类评价机制的引入

在数据集平衡的条件下,一般用数据集分类的总准确率对其分类效果进行评判,即:分类的总准确率越高,则分类器的分类效果越好;但是当数据集不平衡时,特别是不平衡比较大时,存在即使正类样本具有很低的辨识率的情况下,整体的分类准确率很高的情况,故该方式对于不平衡数据的分类准确率的评判是不准确的。为了克服单一分类准确率评价方式不令人信服的弊端,一些学者又提出了一些更加合理的评价机制:灵敏度(Sensitivity, SEN),即正类样本的分类准确率的评价机制;特异性(Specificity, SPE),即负类样本的分类准确率的评价机制;几何平均值(*G*-mean),即分类器的综合评价机制。各评价机制的算法表达式为

$$\mathrm{SEN}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{20}$$

$$\mathrm{SPE}=\mathrm{TN}/(\mathrm{TN}+\mathrm{FP}) \tag{21}$$

$$G\text{-mean}=\sqrt{\mathrm{SEN}\times\mathrm{SPE}} \tag{22}$$

式中:TP(++)为分类正确的正类样本的数目;FN(+−)为分类错误的正类样本的数目;FP(−+)为分类错误的负类样本的数目,TN(−−)为分类正确的负类样本的数目,构成的混淆矩阵如表2所示。

表2 混淆矩阵

Tab. 2 Confusion matrix

Attribute	+	−
+	TP	FN
−	FP	TN

分析上述3种评价机制可知:SEN的值越大正类样本的辨识率就越高;同样SPE的值越大负类样本的辨识率就越高;当SEN与SPE都较大时*G*-mean值就越大,反之*G*-mean值就越小。故对于不平衡数据选取*G*-mean值进行分类器的评价更加合理。

3.2 实验数据以及实验环境

为了突出本文所提算法在不平衡数据下分类的优越性,将所提算法(PSO-DEC-IFSVM)与现有算法进行对比,即:支持向量机(SVM)算法、模糊支持向量机(FSVM)算法、DEC算法、DEC结合FSVM的算法(DEC-FSVM)、DEC-FSVM-Ju算法以及利用PSO算法参数寻优前的DEC-IFSVM算法。同时,为了使实验结果更加具有说服力,本文在UCI机器学习数据中选取6种不同空间结构以及不同维度的不平衡数据进行实验验证,且这些不平衡数据必定会含一些噪声或野点个体。此外,为了减少训练的时间,每种不平衡数据集均随机选择部分作为实验,选取的6种不平衡数据集的基本特征如表3所示。

表3 实验中的6种不平衡数据集的特征

Tab. 3 Characteristics of the six unbalanced data sets in the experiment

Dataset	Attribute	Unbalanced-ratio	Experimental data	Label
Pima	9	268:500	200/(67:133)	1:0
German	25	300:700	200/(68:132)	2:1
Wpbc	35	46:148	190/(46:144)	R:N
Haberman	4	126:225	200/(52:148)	2:1
Yeast	9	429:463	200/(89:111)	NUC:CYT
Abalone	9	634:689	200/(99:101)	10:9

本文所涉及的所有算法均采取十折交叉验证,且为了减少随机影响,每折运行十次,即对于一个不平衡数据将产生100组数据,最终将所得的100组数据的均值作为每种评价机制的最终值。本文所有算法的初始参数均为:$\delta=10^{-13}$,$\alpha=0.5$,$m=0.5$,$C=2$,$g=0.01$以及$K=3$。此外,本文所有结果均是在3.20 GHz/4.0 GB的PC机上利用MATLAB2012a软件编程实现。

3.3 结果与分析

对于6种不同不平衡数据集的3种评价机制的实验对比效果如表4所示。分析表4可知:(1)在不平衡数据集下,传统的SVM算法效果最差,甚至有的数据集中G-mean的值为0,特别是样本集严重失衡时,这是因为分类超平面向正类样本方向发生了严重的偏移,其他算法作为SVM算法的改进形式,使分类超平面偏回负类样本方向,使得分类效果获得提升。(2)传统的DEC算法仅考虑到了样本平衡性的影响,没有考虑样本中噪声或野点影响;相反传统的FSVM算法仅考虑到了样本噪声或野点影响,而忽略了样本平衡性的影响。故在不平衡数据集中传统的DEC与FSVM算法的分类效果提升不是很明显,特别是SEN与G-mean两个评价机制较低,即这两种算法对于分类超平面的向负类偏移影响较小。(3)DEC-FSVM算法将传统的DEC与FSVM方式相结合,融合了两种算法的优点,分类效果得到进一步提升,尤其是SEN或G-mean。(4)DEC-FSVM-Ju算法是在DEC-FSVM算法基础上进行改进,相比DEC-FSVM算法,其分类效果亦有提升,这是因为在设置模糊隶属度函数时DEC-FSVM算法仅考虑了样本到达类中心的距离,而DEC-FSVM-Ju算法考虑样本到类中心距离的同时还考虑了样本的K-近邻域的密度。(5)同样地,DEC-IFSVM作为DEC-FSVM-Ju的改进算法,分类效果亦有提升,这是因为DEC-IFSVM算法除了考虑样本到类中心的距离以及样本的K-近邻域密度外,还考虑到了样本的信息量,在设计模糊隶属度函数时给予样本不同的权值,这样可以赋予支持向量较大的权值,故分类效果进一步提升。(6)对比PSO优化前后的DEC-IFSVM算法可知,经过PSO参数优化后的DEC-IFSVM算法,相比优化前的算法对6种不平衡数据集在分类器的分类效果均有较大提升。

综上,本文所提的算法在综合考虑样本到类中心距离、K-近邻域密度以及样本的信息量设计模糊隶属度函数,并将其与DEC算法相结合,最终引入的参数经过PSO算法优化,与现有的算法相比在不同空间结构以及不同维度的不平衡数据集中具有更好的分类性能。

4 分类器鲁棒性的对比

为了进一步说明本文所提算法的优越性,对本文所有算法的鲁棒性进行比较。本文采用文献[23]中所提算法鲁棒性的评价方式,即算法m在某一特定数据集上的鲁棒性为用该算法求解目标问题时的相对性能,文中选取G-mean值作为不平衡数据分类效果鲁棒性的比较值,求解文中所有算法G-mean值的相对性能,此相对性能的求解算法为

$$b_m=R_m/\max_k(R_k)\quad i\in 1,2,3,\cdots,k \tag{23}$$

式中:R_m为算法m在某一数据集的Adjusted rand index值;b_m为算法m鲁棒性的相对性能。由式(23)可知,当某一算法在特定数据集上表现最好时b_m的值即为1,而其他算法$b_m\leqslant 1$,且b_m的值越大,算法的相对性能就越好。故算法m在不同数据集的鲁棒性可以利用$\sum_{j=1}^{l}b_m^l$(Sum of ARI,S-ARI)表示,其中l为算法的总数,且本文的算法总数为7。同样$\sum_{j=1}^{l}b_m^l$的值越大代表该算法的综合鲁棒性越强。利用上述方法求解本文7种算法在6种平衡数据集上G-mean值的鲁棒性,其结果如图3所示。

分析图3可知:(1)传统的SVM算法S-ARI的值远小于其余算法,证明SVM算法的鲁棒性最差;

表4 6种不平衡数据集下运用各类算法分类的效果

Tab. 4 Classification effect of different algorithms in the six kinds of unbalanced data sets

Dataset	Algorithm	SEN	SPE	*G*-mean
Pima	SVM	0.0 ± 0.0	1.0 ± 0.0	0.0 ± 0.0
	FSVM	0.400 9 ± 0.103 6	0.685 6 ± 0.027 8	0.507 0 ± 0.058 0
	DEC	0.680 9 ± 0.156 5	0.607 4 ± 0.032 0	0.620 7 ± 0.070 1
	DEC-FSVM	0.646 0 ± 0.088 2	0.789 0 ± 0.008 7	0.697 4 ± 0.032 9
	DEC-FSVM-Ju	0.723 0 ± 0.121 5	0.753 6 ± 0.029 2	0.690 8 ± 0.041 2
	DEC-IFSVM	0.756 0 ± 0.061 7	0.781 2 ± 0.017 4	0.750 2 ± 0.020 4
	PSO-DEC-IFSVM	0.811 4 ± 0.025 0	0.817 4 ± 0.011 4	0.807 3 ± 0.007 2
Haberman	SVM	0.119 4 ± 0.020 2	0.865 2 ± 0.094 4	0.232 9 ± 0.065 4
	FSVM	0.581 4 ± 0.102 2	0.703 4 ± 0.137 7	0.584 0 ± 0.012 9
	DEC	0.531 6 ± 0.063 5	0.768 5 ± 0.047 4	0.553 9 ± 0.061 6
	DEC-FSVM	0.586 9 ± 0.047 0	0.775 9 ± 0.032 1	0.654 4 ± 0.015 4
	DEC-FSVM-Ju	0.583 2 ± 0.042 3	0.796 7 ± 0.031 0	0.657 1 ± 0.023 6
	DEC-IFSVM	0.605 1 ± 0.041 0	0.768 2 ± 0.029 9	0.668 5 ± 0.018 0
	PSO-DEC-IFSVM	0.694 2 ± 0.017 7	0.788 1 ± 0.021 4	0.735 8 ± 0.013 3
German	SVM	0.213 1 ± 0.063 1	0.834 8 ± 0.046 2	0.255 2 ± 0.074 8
	FSVM	0.566 2 ± 0.029 5	0.776 0 ± 0.031 8	0.647 5 ± 0.013 7
	DEC	0.575 7 ± 0.093 0	0.702 3 ± 0.042 0	0.647 1 ± 0.012 9
	DEC-FSVM	0.648 9 ± 0.054 4	0.691 3 ± 0.032 6	0.690 0 ± 0.008 4
	DEC-FSVM-Ju	0.755 8 ± 0.047 5	0.709 9 ± 0.014 0	0.714 3 ± 0.005 6
	DEC-IFSVM	0.815 2 ± 0.041 4	0.651 9 ± 0.038 3	0.708 0 ± 0.009 3
	PSO-DEC-IFSVM	0.743 3 ± 0.034 1	0.750 7 ± 0.022 0	0.734 4 ± 0.007 8
Wpbc	SVM	0.0 ± 0.0	1.0 ± 0.0	0.0 ± 0.0
	FSVM	0.694 5 ± 0.075 2	0.699 6 ± 0.051 2	0.655 9 ± 0.011 3
	DEC	0.540 1 ± 0.068 6	0.781 0 ± 0.044 6	0.563 1 ± 0.050 3
	DEC-FSVM	0.702 7 ± 0.043 9	0.657 3 ± 0.037 2	0.662 6 ± 0.010 1
	DEC-FSVM-Ju	0.721 9 ± 0.028 2	0.653 8 ± 0.048 2	0.667 2 ± 0.010 0
	DEC-IFSVM	0.647 3 ± 0.048 0	0.756 5 ± 0.037 4	0.678 9 ± 0.007 0
	PSO-DEC-IFSVM	0.757 9 ± 0.028 5	0.731 7 ± 0.026 0	0.733 2 ± 0.009 7
Yeast	SVM	0.107 0 ± 0.048 5	0.962 5 ± 0.006 3	0.168 0 ± 0.068 3
	FSVM	0.412 4 ± 0.077 3	0.831 9 ± 0.029 7	0.538 5 ± 0.023 8
	DEC	0.469 0 ± 0.028 5	0.834 4 ± 0.016 7	0.609 0 ± 0.002 8
	DEC-FSVM	0.517 1 ± 0.030 9	0.720 1 ± 0.033 7	0.588 3 ± 0.013 0
	DEC-FSVM-Ju	0.563 9 ± 0.033 0	0.807 3 ± 0.012 9	0.661 1 ± 0.003 8
	DEC-IFSVM	0.615 4 ± 0.023 7	0.733 5 ± 0.043 7	0.673 8 ± 0.011 6
	PSO-DEC-IFSVM	0.564 1 ± 0.008 9	0.840 6 ± 0.021 2	0.683 5 ± 0.001 4
Abalone	SVM	0.217 3 ± 0.044 8	0.792 5 ± 0.051 5	0.264 0 ± 0.045 7
	FSVM	0.532 0 ± 0.022 1	0.770 0 ± 0.039 3	0.627 2 ± 0.012 8
	DEC	0.581 3 ± 0.038 9	0.680 2 ± 0.074 8	0.597 6 ± 0.015 7
	DEC-FSVM	0.568 6 ± 0.031 7	0.743 5 ± 0.053 2	0.626 5 ± 0.011 0
	DEC-FSVM-Ju	0.542 6 ± 0.040 1	0.867 8 ± 0.008 8	0.669 0 ± 0.010 4
	DEC-IFSVM	0.615 3 ± 0.032 9	0.749 8 ± 0.040 5	0.656 7 ± 0.010 7
	PSO-DEC-IFSVM	0.629 2 ± 0.034 7	0.764 7 ± 0.019 8	0.676 2 ± 0.010 1

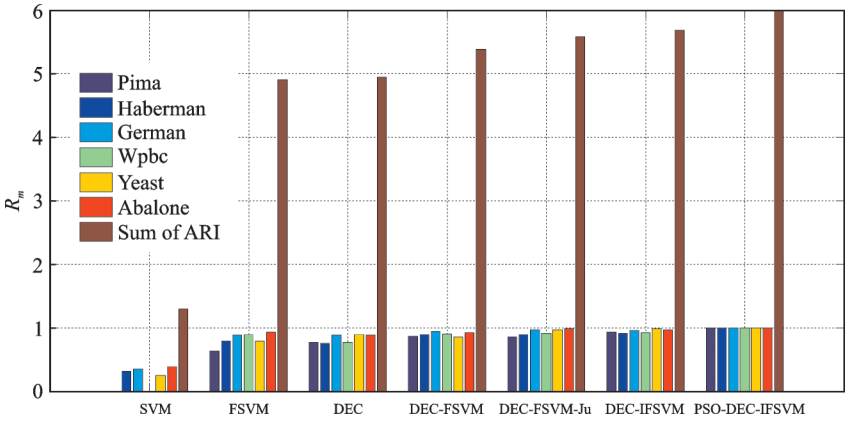


图3 不平衡数据集下7种算法 G-mean值的鲁棒性比较

Fig.3 Robustness comparison of G-mean value of seven algorithms under Unbalanced data sets

(2)分析FSVM与DEC算法的S-ARI值可知:FSVM与DEC算法分类器的总体效果不是很理想,DEC-FSVM算法相比FSVM与DEC算法鲁棒性进一步增强,显示了综合考虑样本距离以及不平衡度的优势;(3)DEC-IFSVM作为DEC- FSVM-Ju的改进算法,其S-ARI值有所增加,证明鲁棒性增强不够明显,这是由于算法引入参数增加时,算法复杂度增加且初始参数不是最优值,导致结果不明显;(4)本文所提的PSO-DEC-IFSVM算法对DEC-IFSVM算法引入的参数进行优化,其S-ARI值最大为6,明显大于DEC-IFSVM算法以及其他算法,故在不同的不平衡数据集上均有最好的鲁棒性。

5 结束语

针对传统的模糊支持向量机在不平衡数据集下分类效果不够明显、引入的参数未做优化等缺点,本文提出一种新型的基于粒子群优化的改进支持向量机算法(PSO-DEC-IFSVM)。该算法在设计模糊隶属函数时,综合考虑训练样本到期类中心的间距与样本周围的紧密度以及样本的信息量,并将其与DEC算法相结合,最后利用粒子群算法对DEC-IFSVM算法引入的K,α,M,C以及g五个参数进行优化。实验证明:本文算法相比已有的FSVM算法,正负类的分类精度进一步增加,且此算法拥有更好的鲁棒性。结果证明:本文算法可以更好地降低样本集中含有噪声或野点影响,同时,可以更好地应对数据集不平衡问题。故此算法为不平衡数据的分类问题提供了一个重要的理论模型,该模型可以应用于机械故障诊断、医疗诊断等异常诊断领域,因为在这些领域中故障数据收集相对困难,极易形成不平衡数据集,且数据集中很可能含有噪声或者野点。

本文在利用粒子群算法对DEC-IFSVM分类器进行参数寻优时,仅将分类器的综合评价机制(G-mean)作为优化目标,这可能会导致正负类分类准确率(SEN,SPE)不一定同时比优化前效果理想,所以将SEN,SPE,G-mean同时作为优化目标进行协同优化,即:寻求一种适用于多目标寻优的智能算法,将是课题组下一步的研究重点。

参考文献:

[1] 李勇,刘战东,张海军.不平衡数据的集成分类算法综述[J].计算机应用研究,2014,31(5):1287-1291.
Li Yong, Liu Zhandong, Zhang Haijun. Review on ensemble algorithms for imbalanced data classification[J]. Application Research of Computers, 2014,31 (5): 1287-1291.

[2] 张晶,冯林.针对动态非平衡数据集鲁棒的在线极端学习机[J].计算机研究与发展,2015,52(7): 1487-1498.
Zhang Jing, Feng Lin. An algorithm of robust online extreme learning machine for dynamic imbalanced datasets[J]. Journal of Computer Research and Development, 2015, 52(7): 1487-1498.

[3] Shao Y H, Chen W J, Zhang J J, et al. An efficient weighted Lagrangian twin support vector machine for imbalanced data classification[J]. Pattern Recognition, 2014, 47(9): 3158-3167.

[4] He Haibo, Garcia E A. Learning from imbanlanced data[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[5] 汪友生, 胡百乐, 张丽杰,等. 基于支持向量机的动脉硬化斑块识别[J]. 数据采集与处理, 2012, 27(3): 283-286.
Wang Yousheng, Hu Baile, Zhang Lijie, et al. Recognition of atherosclerotic plaque based on support vector machine[J]. Journal of Data Acquisition and Processing, 2012, 27(3): 283-286.

[6] Dey S, Sarkar R, Chatterjee K, et al. Pre-cancer risk assessment in habitual smokers from DIC imag es of oral exfoliative cells using active contour and SVM analysis[J]. Tissue & Cell, 2017, 49(2): 296-306.

[7] 段礼祥,郭晗,王金江.数据集不均衡下的设备故障程度识别方法研究[J].振动与冲击,2016,35(20): 178-182.
Duan Lixiang, Guo Han, Wang Jinjiang. A mechanical fault severity identification method under unbalanced datasets[J]. Journal of Vibration and Shock, 2016, 35(20): 178-182.

[8] Duan L, Xie M, Bai T, et al. A new support vector data description method for machinery fault diagnosis with unbalanced datasets[J]. Expert Systems with Applications, 2016, 64: 239-246.

[9] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[J]. Lecture Notes in Computer Science, 2005, 3644(5): 878-887.

[10] 楼晓俊,孙雨轩,刘海涛.聚类边界过采样不平衡数据分类方法[J].浙江大学学报(工学版),2013,47(6): 944-950.
Lou Xiaojun, Sun Yuxuan, Liu Haitao. Clustering boundary over-sampling classification method for unbalanced data sets[J]. Journal of Zhejiang University (Engineering Science), 2013, 47(6): 944-950.

[11] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets[J]. Lecture Notes in Computer Science, 2004, 3201: 39-50.

[12] 陶新民, 郝思媛, 张冬雪,等. 基于样本特性欠取样的不均衡支持向量机[J]. 控制与决策, 2013(7): 978-984.
Tao Xinmin, Hao Siyuan, Zhang Dongxue, et al. Support vector machine for unbalanced data based under-sampling approaches on sample properties[J]. Control and Decision, 2013(7): 978-984.

[13] Ao S, Wu J, Cai Z. Kernel function pre-processed SVM and its application in imbalanced data sets[J]. Energy Procedia, 2011, 13: 3316-3324.

[14] 刘东启,陈志坚,徐银,等. 面向不平衡数据分类的复合SVM算法研究[J]. 计算机应用研究,2018,35(4): 1023-1027.
Liu Dongqi, Chen Zhijian, Xu Yin, et al. Hybrid SVM algorithm oriented to classifying unbalanced datasets[J]. Application Research of Computers, 2018, 35(4): 1023-1027.

[15] Lin C F, Wang S D. Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471.

[16] Guo W, Jia S, Xu T, et al. A new motion control method for omnidirectional intelligent wheelchair based on improved fuzzy support vector machine[C]// IEEE International Conference on Mechatronics and Automation. [S.l.]: IEEE, 2015: 1567-1572.

[17] 李忠国, 侯杰, 王凯,等. 模糊支持向量机在路面识别中的应用[J]. 数据采集与处理, 2014, 29(1): 146-151.
Li Zhongguo, Hou Jie, Wang Kai, et al. Application of fuzzy support vector machine on road type recognition[J]. Journal of Data Acquisition and Processing, 2014, 29(1): 146-151.

[18] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(1): 2-10.
Ding Shifei, Qi Bingjuan, Tan Hongyan. An overview on theory and algorithm of support vector machines[J]. Journal of University of Electronic Science and Technology of China, 2011, 40(1): 2-10.

[19] 李苗苗, 向凤红, 刘新旺. 一种新颖隶属度函数的模糊支持向量机[J]. 计算机工程与科学, 2009, 31(9): 92-94.
Li Miaomiao, Xiang Fenghong , Liu Xinwang. A novel membership function for fuzzy support vector machines[J]. Computer

Engineering & Science, 2009, 31(9): 92-94.

[20] 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.

[21] 鞠哲, 曹隽喆, 顾宏. 用于不平衡数据分类的模糊支持向量机算法[J]. 大连理工大学学报, 2016, 56(5): 525-531.
Ju Zhe, Cao Junzhe, Gu Hong. A fuzzy support vector machine algorithm for unbalanced data classification[J]. Journal of Dalian University of Technology, 2016, 56(5): 525-531.

[22] He H, Ma Y. Class imbalance learning methods for support vector machines[M]. [S.l.]: Wiley-IEEE Press, 2013: 83-99.

[23] 公茂果, 焦李成, 马文萍, 等. 基于流形距离的人工免疫无监督分类与识别算法[J]. 自动化学报, 2008, 34(3): 367-375.
Gong Maoguo, Jiao Licheng, Ma Wenping, et al. Unsupervised classification and recognition using an artificial immune system based on manifold distance[J]. Acta Automatica Sinica, 2008, 34(3): 367-375.

作者简介:

魏建安(1992-),男,博士研究生,研究方向:智能制造与数据挖掘,E-mail:jiananwei0811@foxmail.com。

黄海松(1977-),女,**通信作者**,博士,教授,博士生导师,研究方向:智能制造、制造业信息化等,E-mail:1046534381@qq.com。

康佩栋(1992-),男,硕士研究生,研究方向:智能制造与数据获取。

(编辑:张黄群)