文章编号: 1001-0920(2018)09-1549-10 DOI: 10.13195/j.kzyjc.2017.0649

基于不平衡数据样本特性的新型过采样SVM分类算法

黄海松†, 魏建安, 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: 针对传统采样方式准确率与鲁棒性不够明显,欠采样容易丢失重要的样本信息,而过采样容易引入冗杂信息等问题,以UCI公共数据集中的不平衡数据集Pima-Indians为例,综合考虑数据集正负类样本的类间距离、类内距离与不平衡度之间的关系,提出一种基于样本特性的新型过采样方式. 首先对原始数据集进行距离带的划分,然后提出一种改进的基于样本特性的自适应变邻域Smote算法,在每个距离带的少数类样本中进行新样本的合成,并将此方式推广到UCI数据集中其他5种不平衡数据集. 最后利用SVM分类器进行实验验证的结果表明:在6类不平衡数据集中,应用新型过采样SVM算法,相比已有的采样方式,少(多)数类样本的分类准确率均有明显提高,且算法具有更强的鲁棒性.

关键词: 数据集不平衡；样本距离；ANBSC-Smote过采样；数据集重构；支持向量机

中图分类号: TP273 **文献标志码:** A

New over-sampling SVM classification algorithm based on unbalanced data sample characteristics

HUANG Hai-song†, WEI Jian-an, KANG Pei-dong

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

Abstract: Aiming at the problem that the accuracy and robustness of the traditional sampling methods are not obvious, under-sampling is easy to lose important sample information, and oversampling is easy to introduce redundant information, the Pima-Indians dataset in the UCI common unbalanced datasets is taken as an example to consider the relationship between the distance within classes, the distance within classes and the imbalance, therefore, a new type oversampling method based on sample characteristics is presented. Firstly, the algorithm divides the original data set into some distance belts. Then an improved adaptive neighborhood neighborhood(Smote) algorithm based on sample characteristics is proposed to synthesize new samples in each class with several samples, and is extended to other five unbalanced data sets of UCI dataset. Finally, experiments are conducted using the traditional SVM classifier, and the results show that, in the six categories of unbalanced data sets, compared with the existing sampling method, the proposed algorithm improves the classification accuracy of the minority or majority class samples, and has stronger robustness.

Keywords: unbalanced datasets； sample distance； ANBSC-Smote oversampling； datasets reconstuction； SVM

0 引 言

大数据时代,随着数据的爆炸式增长,信息量成指数累增,这些数据集包括平衡数据集和不平衡数据集. 而在现实生产生活中,更多且更具研究意义的是那些不平衡数据集[1-2],其分类研究也是近些年的研究热点,例如信用卡欺诈、医疗诊断、机械故障诊断等[3-7].

目前,可以从以下两个方面解决不平衡数据的分类问题: 1)在数据层面上,利用欠(过)采样方式或者内插的方式进行数据的重构,使数据集达到平衡; 2)在算法层面上,通过改进分类器算法以提高对少数类数据的识别,比如引入代价函数、集成学习、单类学

收稿日期: 2017-05-25； **修回日期:** 2017-09-29.

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

责任编委: 阳春华.

作者简介: 黄海松(1977−), 女, 教授, 从事智能制造、制造业信息化等研究； 魏建安(1992−), 男, 硕士生, 从事智能制造、机器学习的研究.

†通讯作者. E-mail: 1046534381@qq.com

习等[8-10]. 然而,在数据层面上将数据进行重构是从SVM分类器的根本原理进行的一种简单有效的处理数据不平衡的方式,且应用广泛,故本文从数据层面上对不平衡数据源进行重构. 常用的数据重构方式主要有以下几种: 1) Smote过采样算法是一种较为常用的算法, 因其在生成新的样本时存在一定的盲目性[11],很多学者提出了改进措施. Han等[12]在Smote的基础上提出了BSmote算法,找到正类(少数类)的边界样本,并只对其进行Smote处理,分类效果有了进一步提升; He等[13]提出了AdaSyn-Somte的改进算法,能够根据数据集的分布情况来控制新样本的分布情况,分类效果又一步提升;此外,还有代价敏感SVM与改进聚类边界Smote过采样方式等[14-15]. 2)在欠采样方面(包括随机欠采样方式[16]),此方式容易删除重要负类(多数类)的样本信息. Batista等[17]提出了借鉴实例简约的DROP与CNN算法,分类效果有了进一步提升;陶新民等[18]提出了基于样本特性的欠取样方式(SPU),首先用原始数据在SVM上找到偏移的超平面,然后从负类找到距离该超平面较近且信息量大的负类样本进行采样,以达到超平面偏移的目的,分类效果又一步提升. 然而,当样本严重失衡时,以上方法并不是很理想. 因此,本文综合考虑传统采样算法的优缺点,以Pima-Indians不平衡数据集为例,剖析了不平衡原始样本集的数据特征,进而综合考虑不平衡数据集的类间距、类内间距以及不平衡度等特性,提出一种基于数据特征的新型过采样方式(New over-sampling based on data feature, NOBDF),并以Pima-Indians为例对数据集进行重构. 将此方式推广到UCI数据集中的其他5类不平衡数据集中,经实验验证, 6种不平衡数据集在NOBDF算法的重构之下,无论是正负类的分类效果,还是综合效果,均具有明显的提升.

1 基于SVM的不平衡数据样本分类性能分析

1.1 支持向量机的分类原理

支持向量机(SVM)是由Cortes等[19]根据统计学习中的VC维理论和结构风险最小化原则,基于有限的样本信息在复杂模型与学习之间求解最佳效果,以期获取最好的泛化能力而提出的一种机器学习方法. 因其在处理数据的小样本、非线性、解决维灾难以及过学习中具有较强的能力,已经成功应用于各种实际生产生活中[20].

以二分类为例,SVM的基本思想为: 在样本(核)空间内寻找一个使二类样本达到分隔间距最大化的最优超平面,即假设给定的原始训练样本集为$\{(x_i,y_i)\}_{i=1}^n, i=1,2,\cdots,n, x_i$为第$i$个样本,$y_i\in\{-1,+1\}$为第$i$个样本类别标签. 为此,SVM算法引入核函数$K$,将样本集映射到高维空间;引入松弛变量$\xi$,表示训练样本被错分的程度;引入惩罚因子$C$,表示对错分样本的惩罚程度. 另外,引入参数$\omega$和$b$分别代表决策函数$f(x)=w\cdot x+b$的权值向量和阈值,构造如下所示的代价函数,并使其最小,以得到SVM的一般形式:

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

式(1)为一个二次规划求最优解问题,故引入Lagrange乘子法将其转变为如下所示的对偶问题:

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

假定a^*为式(2)的最优Lagrange乘子,进而反向求得最优分类超平面法向量ω^*和阈值b^*,则最终的决策函数如下所示:

$$f(x)=\mathrm{sgn}\Big(\sum_{i=1}^{n}\alpha^*y_iK(x,x_i)+b^*\Big).\tag{3}$$

1.2 不平衡数据集下SVM分类特性的分析

一般传统SVM算法的适用条件为数据集中的各类样本数目是平衡的,然而实际生产生活中的数据大多是不平衡数据. 显然,在此种情况下,利用传统SVM算法分类的效果大多不佳. 其原因如文献[18]中提到的,当数据不平衡时,由于传统SVM算法针对不同类别的错误分类采用了同一的惩罚系数,此时样本密度比较小的少数类,训练后的误差也较小,故在SVM算法本身机制下,为了增大分离间隔的同时降低错分经验风险,正负类的最优分类超平面势必会向小样本的类别移动,从而降低了SVM分类器对小样本类别的识别率. 故寻找合理的采样方式以重构数据集,进而更好地解决分类超平面偏移方向问题,是提高SVM分类效果的关键所在. 另外,为了更加透彻地分析不平衡数据集的特性,以UCI公共数据集中的Pima-Indians数据集为例,进一步剖析不平数据的内部特性. 该数据集的基本属性如下: 正样本数为268,负样本数为500,样本属性为768,不平衡比为1.87,类

别总数为2,正类所属类别标签为1,负类为0. 因为要综合考虑样本的类内间距、类间间距和不平衡度,所以首先要求出正负类的中心

$$D^{+(-)}=\frac{1}{n^{+(-)}}\sum_{1}^{n^{+(-)}}x_i,$$
$$n^{+}=1,2,\cdots,l,n^{-}=1,2,\cdots,m. \tag{4}$$

其中: D^{+}、D^{-}分别代表正负类的中心; n^{+}、n^{-}分别代表正负类样本的个数,其大小分别为l和m. 接下来求正负类之间的间距

$$D=\|D^{+}-D^{-}\|, \tag{5}$$

其中D为正负类之间的间距, Pima-Indians数据集的类间距离为$D=5.877\,2$. 然后,求正负类样本到其所属类的类内间距

$$d_{i(j)}{}^{+(-)}=\|x_{i(j)}-D^{+}(D^{-})\|,$$
$$i=1,2,\cdots,l,j=1,2,\cdots,m, \tag{6}$$

其中$d_i{}^{+}$、$d_j{}^{-}$分别为正类和负类第$i(j)$个样本点到各自类中心的距离. 为了更好地解决分类超平面偏移方向的问题,本文提出将类内距离划分为一系列小距离带的数据处理方法,即正负类样本个体到本类中心的距离带. 距离带的划分原则为: 1) 确定距离带范围(Distance range, Dr),下限L_b的值取为0,上限U_b的值取为$\max\{d_i{}^{+},d_j{}^{-}\},i=1,2,\cdots,l,j=1,2,\cdots,m$,即正负类样本到各自中心的总的最大值; 2) 确定距离带的带宽(Bandwith, Bw),带宽的选择不宜过大或者过小,因为过大会产生较大误差,导致样本特性不易观察,过小同样不易观察样本特性,本文经过多次尝试,暂选取$B_w=1$; 3) 求出在每个距离带中的正类样本(Positve sample, Ps)数和负类样本(Negative samples, Ns)数; 4) 求出每个距离带的密度之差(Density different, Dd). 最终将求解后的结果汇总于表1.

观察并分析表1可知: 1) 每个距离带正负类样本的数目大多具有一定的差距,即数据集在距离带内出现了不均衡性,特别在前17个距离带中更加明显,同时正负类之间的间距在此范围之内; 2) 并非每个距离带内都是负类的数目多于正类,比如12～17组距离带正类的密度大于负类的密度; 3) 在第18个距离带以后,正负类之间的样本密度差距较小,即远离类中心的样本点具有较弱的不均衡性. 考虑到上述Pima-Indians数据集需要在带内进行过采样,故需要将第11～50带进行合并,合并后的样本距离特征如表2所示.

表 1　Pima-Indians数据集的样本距离特征

No.	Dr	Ps	Ns	Dd	No.	Dr	Ps	Ns	Dd
1	(0, 1)	16	18	2	26	(25, 26)	0	4	4
2	(1, 2)	13	22	9	27	(26, 27)	0	4	4
3	(2, 3)	11	26	15	28	(27, 28)	0	1	1
4	(3, 4)	17	30	13	29	(28, 29)	2	3	1
5	(4, 5)	17	30	13	30	(29, 30)	1	1	0
6	(5, 6)	20	38	18	31	(30, 31)	0	2	2
7	(6, 7)	18	40	22	32	(31, 32)	0	4	4
8	(7, 8)	14	47	33	33	(32, 33)	1	1	0
9	(8, 9)	20	38	18	34	(33, 34)	0	3	3
10	(9, 10)	14	70	56	35	(34, 35)	0	2	2
11	(10, 11)	9	69	60	36	(35, 36)	0	2	2
12	(11, 12)	11	2	−9	37	(36, 37)	0	1	1
13	(12, 13)	19	3	−16	38	(37, 38)	0	2	2
14	(13, 14)	13	7	−6	39	(38, 39)	0	0	0
15	(14, 15)	14	6	−8	40	(39, 40)	0	0	0
16	(15, 16)	15	2	−13	41	(40, 41)	0	1	1
17	(16, 17)	9	4	−5	42	(41, 42)	0	0	0
18	(17, 18)	1	2	1	43	(42, 43)	0	0	0
19	(18, 19)	2	3	1	44	(43, 44)	0	0	0
20	(19, 20)	1	3	2	45	(44, 45)	0	0	0
21	(20, 21)	3	1	−2	46	(45, 46)	0	0	0
22	(21, 22)	2	1	−1	47	(46, 47)	0	0	0
23	(22, 23)	2	2	0	48	(47, 48)	0	0	0
24	(13, 24)	1	3	2	49	(48, 49)	0	0	0
25	(24, 25)	2	1	−1	50	(49, 50)	0	1	1

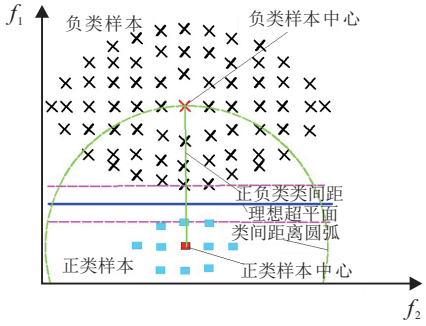
表 2　合并后的Pima-Indians数据集的样本距离特征

No.	Dr	Ps	Ns	Dd
1	(0, 1)	16	18	2
2	(1, 2)	13	22	9
3	(2, 3)	11	26	15
4	(3, 4)	17	30	13
5	(4, 5)	17	30	13
6	(5, 6)	20	38	18
7	(6, 7)	18	40	22
8	(7, 8)	14	47	33
9	(8, 9)	20	38	18
10	(9, 10)	14	70	56
11	(10, 11)	108	141	33

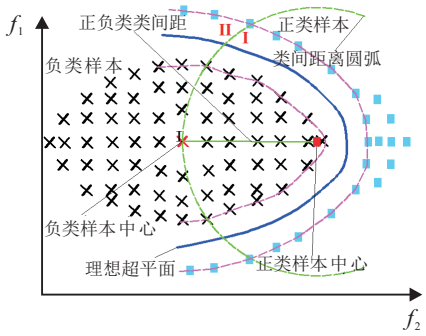
2　基于样本特性的过采样SVM分类算法

前文提到了Pima-Indians不平衡数据集在距离带内出现了数据不平衡,尤其在正负类类间距的附近,现在分4种情况讨论,如图1所示.

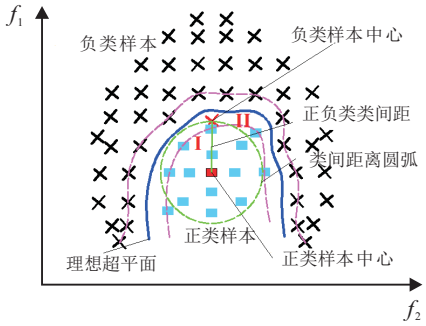
1) 当正负类样本均为如图1(a)所示的凸集时,正负类中心位于最优分类超平面的两侧,且正类的支持向量到其类中心的距离均小于正负类的类间距,故此时需要对在正负类间距的内侧距离带的样本数据进行数据集的重构.



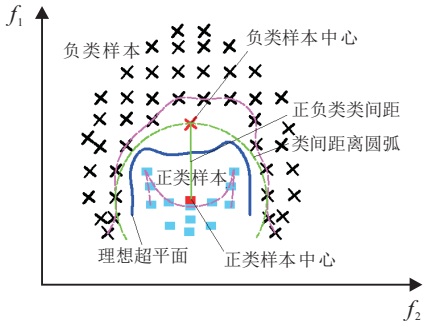
(a) 正负类样本均为凸集



(b) 正类样本为凹集, 负类样本为凸集



(c) 正类样本为凸集, 负类样本为凹集



(d) 正负类样本均为凹集

图 1 不平衡数据分类最优超平面的几种情况

2) 当正类样本为凹集, 负类样本为凸集(如图1(b)所示)时, 由于正类样本为凹集, 可以分为两种情况进行讨论: i) 当正类的支持向量到其类中心的距离小于等于正负类的类间距(如I部分所示)时, 此情况与情况1)的处理方式类似, 需要对正负类间距内侧距离带的样本数据进行数据集的重构; ii) 当正类的支持向量到其类中心的距离大于正负类的类间距(如II部分所示)时, 需要对正负类间距的外侧距离带的样本数据进行数据集的重构. 综合考虑以上两种情况, 对于正类样本为凹集, 负类样本为凸集的数据集, 需要同时对正负类间距的内外两侧的距离带的样本数据进行数据集的重构.

3) 当正类样本为凸集, 负类样本为凹集(如图1(c)所示)时, 由于负类样本为凹集, 亦可以分为两种情况进行讨论: i) 当正类的支持向量到其类中心的距离小于等于正负类的类间距(如I部分所示)时, 此情况亦与情况1)处理方式类似, 需要对正负类间距的内侧距离带的样本数据进行数据集的重构; ii) 当正类的支持向量到其类中心的距离大于正负类的类间距(如II部分所示)时, 需要对正负类间距的外侧距离带的样本数据进行数据集的重构. 综合考虑以上两种情况, 对于正类样本为凸集, 负类样本为凹集的数据集, 亦需要同时对正负类间距的内外两侧距离带的样本数据进行数据集的重构.

4) 当正负类样本均为如图1(d)所示的凹集时, 正负类中心位于最优分类超平面的两侧, 且正类的支持向量到其类中心的距离均小于正负类的类间距, 故此时需要对在正负类间距的内侧距离带的样本数据进行数据集的重构.

综合考虑以上4种情况, 可以得出: 对于一个凹凸情况未知的不平衡数据集, 需要在正负类间距的附近区域内进行数据集的重构.

为此, 本文提出一种基于样本特性的新型过采样算法(New oversampling based on data feature, NOBDF)进行数据集的重构. 该算法首先对数据集进行距离带的划分, 然后在每个距离带少数类样本中提出一种改进的基于样本特性自适应变邻域的Somte算法(An improved Somte algorithm for adaptive neighborhood based on sample characteristics, ANBSC-Smote)进行新样本的合成, 算法原理如图2所示. 以Pima-Indians数据集为例, 详细阐述NOBDF算法, 其步骤如下.

Step 1: 算法开始;

Step 2: 在Pima-Indians数据集完成距离划带;

Step 3: 基于ANBSC-Smote算法进行新样本合成;

Step 4: 重构Pima-Indians原始数据集;

Step 5: 传统支持向量机分类;

Step 6: 算法结束.

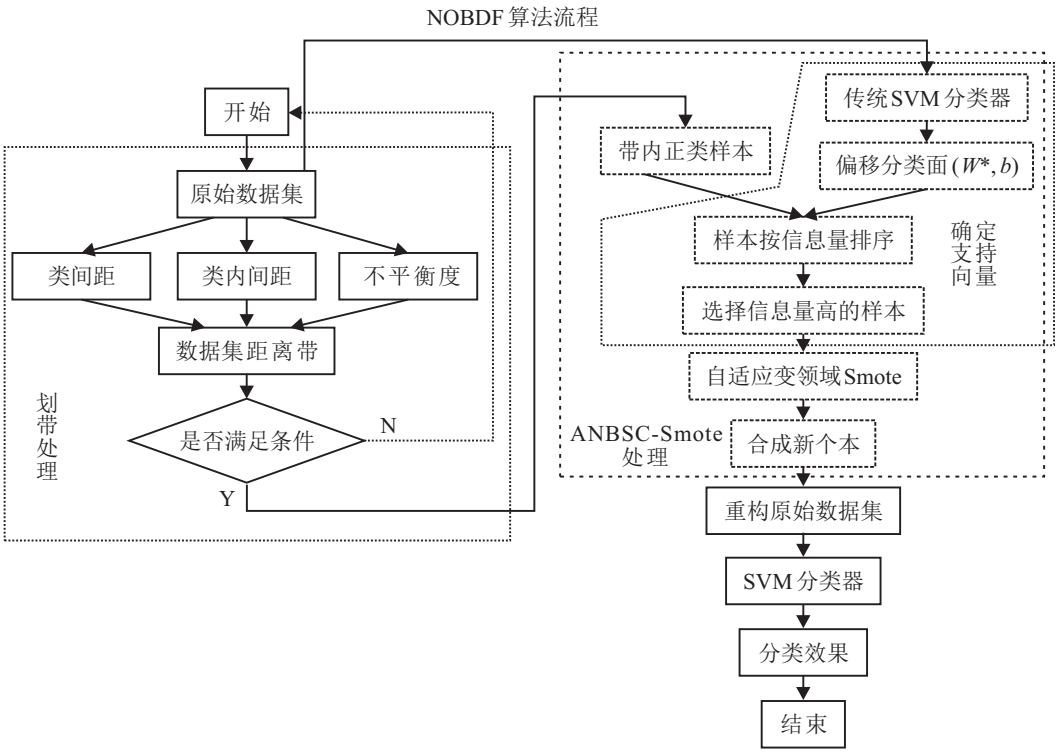


图 2　不平衡数据下, NOBDF算法的分类原理

上面的算法有两个创新点, 一个是划带, 前文已作阐述; 另一个是基于ANBSC-Smote算法进行新样本的合成. ANBSC-Smote算法描述如下: 以表2的第2带为例, 正类有13个样本, 负类有22个样本, 因此在此带中正类需要重新合成9个不同的个体. 为了克服传统Smote算法合成样本时的盲目性, 在样本合成之前需要对此带内的正类样本作相应处理进行排序, 如文献[18]中所提到的“信息量大的样本对分类界面影响较大, 且距离越近信息量越大”, 故本文按照信息量对每个带内的正类样本进行排序, 其公式如下:

$$\varphi(x_i{}^+) = -\|w^* \cdot x_i{}^+ + b\|,\ i = 1, 2, \cdots, n. \quad (7)$$

其中: $\varphi(x_i{}^+)$ 为带内第 i 个正类样本信息量, ω^* 和 b 分别为传统SVM的分类平类超平面的法向量和阈值. 另外, 在正类样本支持向量的确定上: 因为对于某个不平衡数据集而言, 选取的正负类样本集的凹凸情况可能不可知, 所以正负类边界可能包含前文所讨论的情况的1种, 或者2～3种, 则此时理想状态下所有支持向量可能同时落于正负类间距附近的几个距离带中; 同时, 为了使正负类样本集的数目一致, 不妨在划定的每个距离带中选取信息量较大的样本点进行新样本的合成. 其中, 在本文首次运用SVM算法对所选的不平衡数据集进行分类时, 实际上正类样本中绝大部分可能为支持向量的边界点, 由于实际分类超平面的偏移, 这些点很可能被分为负类, 则此时 $\varphi(x_i{}^+)$ 的值越小信息量越大, 越容易成为支持向量; 反之, 也有极少部分(或者没有)正类样本被正确分类, 则此时 $\varphi(x_i{}^+)$ 的值越大信息量越大, 该点越容易成为支持向量, 以上即为正类支持向量的确定过程.

对带内正类样本排序之后, 将信息量最大的样本个体拟作为域中心, 邻域值 k 拟选为 $k = \mathrm{Ps} - 1$, 而Dd作为 k 邻域合成个体的总数, 取样过程如下: 假设正类样本在带中信息量排序为 $a > b > c > d > e$, 则以 a 作为邻域中心, 此时邻域值 k 为4, 如果Dd为36, 则需要合成36个新个体, 故在 $a-b, a-c, a-d, a-d$ 所组成的直线上各合成9个新个体. 以 $a(x_1) - b(x_2)$ 为例, 如图3所示, 首先取二者所在直线的中点 (x_3) 作

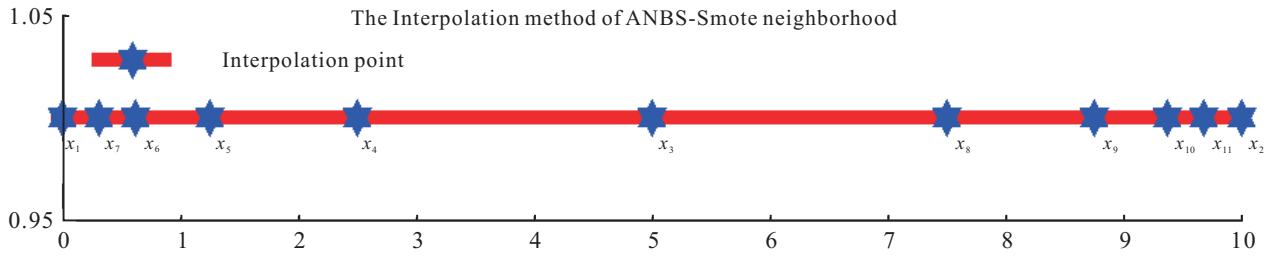


图 3　不平衡数据下, ANBSC-Smote算法新样本合成方式

为分界点,再以x_3为边界分别向两端取中点,即x_4为x_1和x_3的中点,x_8为x_2和x_3的中点,以此类推得到$x_5 \sim x_{11}$. 同时,本算法应该遵循以下原则:若合成个体数不能均分,则在靠近邻域中心的个体数目应大于远离邻域中心的数目. 综上,ANBS-Smote不但消除了传统Smote算法在合成新样本时的盲目性,而且弥补了合成新个体时发生混叠现象的弊端.

3 实验与分析

3.1 不平衡数据分类效果的评价机制

对于不平衡数据而言,以单纯的准确率为评价机制是毫无意义的,因为存在少数类识别率很低而总体准确率很高的情况. 为了克服这种评价机制的弊端,不少学者引入了一些更加合理的评价机制,包括特异性(Specificity)、敏感性(Sensitivity)、几何平均值(G-mean)、少数类(正类)的查准率(Precision)以及少数类的(F-measure)值等[18],并引入包括TP(样本为正类,预测亦为正类的个数)、FN(样本为正类,被错分为负类的个数)、FP(样本为负类,被错分为正类的个数)以及TN(样本为负类,预测为负类的个数)等变率.

各种评价机制的算法如下.

1) 特异性

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}} + \text{FN}; \tag{8}$$

2) 敏感性

$$\text{Specificity} = \frac{\text{TN}}{\text{FP}} + \text{TN}; \tag{9}$$

3) 几何平均值

$$G\text{-mean} = \sqrt{\text{Sensitivity}} \cdot \text{Specificity}; \tag{10}$$

4) 正类的查准率

$$\text{Precision} = \frac{\text{TP}}{\text{FP}} + \text{TP}; \tag{11}$$

5) 正类的F-measure值

$$F\text{-measure} = \frac{2 \cdot \text{Sensitivity} \cdot \text{Precision}}{\text{Sensitivity}} + \text{Precision}. \tag{12}$$

分析式(8)~(12)可知:Sensitivity和Specificity分别代表正确预测正负类样本的比率,显然二值越大越好,但考虑到二者之间的相互制约性,特别是在数据不平衡时,很难同时用这两个指标进行分类效果评价;评价机制G-mean综合考虑了二者之间的相互制约性,当分类面向某类样本偏移时,G-mean值会减小,故此评价机制具有更好的参考性;F-measure将正类的查全率与查准率相结合,且二者均可影响F-measure值,因此它更能体现分类器对正类的识别效果.

3.2 将NOBDF算法推广至其余不平衡数据集的准则

将NOBDF算法推广到UCI数据集的其他5类不平衡数据集中,应遵循如下准则:

1) 所有距离带内的正类样本数目应该少于负类样本数目,反之则重新确定带宽;

2) 如果有任意一个带内正或负类样本数为0,则需要重新划定带宽;

3) 带宽的划定不宜太窄或者太宽,且每个距离带内正类样本的个数须大于等于2.

3.3 实验数据

为了进行实验结果的对比,本文实验数据来源为UCI-machine learning repository机器学习数据库的6种不同的数据集,且引用文献[18]中实验基本参数,如表3所示.

表3 6种不平衡数据集的特征

Datasets	Attributes	Unbalanced-ratio	Labels
pima	9	268/500	1: 0
german	25	300/700	B: A
wpbc	35	46/148	R: N
haberman	4	126/225	2: 1
yeast	9	429/463	NUC: CYT
abalone	9	634/689	10: 9

3.4 不同采样方式的分类性能比较

为了突出本文所提算法的优越性,将本文的NOBDF-SVM算法与上文提到的SPU-SVM、RU-SVM、Smote-SVM、BSmote-SVM、Weight-SVM、RU-Smote-SVM和AdaSyn-SVM等算法对表3的6种不平衡数据集的处理结果进行对比.

3.4.1 实验条件以及分类器的设置

实验采用十折交叉验证,为了去除随机影响,每折运行10次,待实验结束后,计算评价机制G-mean和F-measure的统计均值;同时,考虑到不平衡数据下的特性,实验中正负类样本的选取按1:10的比例随机选取;另外,对比的各类算法都选取各自的最优值,为了便于比较,其设置同文献[18].

SVM分类器的参数设置为:核函数为高斯径向基,核宽度数δ为10,惩罚因子C为1 000;SPU-SVM算法中的$L = 5 \times \text{MI}, u = 7, \lambda = 2, \delta = 10$;Smote-SVM和BSMOTE-SVM算法中的最近邻域参数$k = 6$;Weigh-SVM算法中的正类与负类的代价比为$C_{\text{MA}}/C_{\text{MI}} = 0.1$,其余的欠采样算法负类保留与正类相同的样本数;本文算法中6种数据集的距离带的个数均为2. 6种数据集各个评价机制汇总于表4.

表 4 6种不平衡数据集的评价机制数值表

Dataset	Methods	Specificity	Sensitivity	G-mean	F-measure
	SVM	1.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	RU	0.898 ± 0.075	0.521 ± 0.046	0.682 ± 0.038	0.673 ± 0.038
	Smote	0.768 ± 0.073	0.687 ± 0.045	0.725 ± 0.034	0.789 ± 0.029
	BSmote	0.772 ± 0.076	0.685 ± 0.041	0.726 ± 0.035	0.788 ± 0.027
pima	Weight	0.778 ± 0.064	0.679 ± 0.033	0.736 ± 0.036	0.797 ± 0.024
	RUS	0.778 ± 0.058	0.611 ± 0.064	0.687 ± 0.025	0.733 ± 0.045
	AdaSyn	0.760 ± 0.052	0.704 ± 0.039	0.730 ± 0.030	0.800 ± 0.027
	SPU	0.718 ± 0.134	0.767 ± 0.114	0.733 ± 0.043	0.833 ± 0.068
	NOBDF	0.969 ± 0.015	0.883 ± 0.010	0.925 ± 0.009	0.834 ± 0.006
	SVM	0.998 ± 0.004	0.035 ± 0.005	0.039 ± 0.046	0.007 ± 0.011
	RU	0.856 ± 0.037	0.482 ± 0.053	0.641 ± 0.034	0.630 ± 0.048
	Smote	0.824 ± 0.029	0.481 ± 0.032	0.629 ± 0.024	0.626 ± 0.029
	BSmote	0.831 ± 0.036	0.472 ± 0.031	0.626 ± 0.025	0.619 ± 0.028
german	Weight	0.794 ± 0.058	0.555 ± 0.054	0.662 ± 0.030	0.569 ± 0.042
	RUS	0.807 ± 0.038	0.384 ± 0.045	0.555 ± 0.031	0.531 ± 0.046
	AdaSyn	0.811 ± 0.029	0.486 ± 0.032	0.627 ± 0.025	0.629 ± 0.029
	SPU	0.776 ± 0.098	0.604 ± 0.092	0.679 ± 0.026	0.719 ± 0.024
	NOBDF	0.899 ± 0.021	0.981 ± 0.003	0.938 ± 0.011	0.958 ± 0.005
	SVM	1.0 ± 0.0	0.019 ± 0.027	0.086 ± 0.114	0.037 ± 0.051
	RU	0.824 ± 0.058	0.371 ± 0.138	0.543 ± 0.109	0.498 ± 0.144
	Smote	0.811 ± 0.080	0.426 ± 0.072	0.584 ± 0.044	0.559 ± 0.063
	BSmote	0.825 ± 0.077	0.422 ± 0.089	0.586 ± 0.057	0.556 ± 0.081
wpbc	Weight	0.831 ± 0.079	0.439 ± 0.076	0.600 ± 0.049	0.574 ± 0.068
	RUS	0.763 ± 0.073	0.451 ± 0.095	0.585 ± 0.072	0.573 ± 0.084
	AdaSyn	0.804 ± 0.073	0.435 ± 0.079	0.589 ± 0.055	0.566 ± 0.072
	SPU	0.717 ± 0.119	0.532 ± 0.117	0.612 ± 0.076	0.633 ± 0.091
	NOBDF	1.0 ± 0.0	0.764 ± 0.018	0.874 ± 0.010	0.780 ± 0.012
	SVM	1.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	RU	0.961 ± 0.063	0.128 ± 0.106	0.281 ± 0.211	0.207 ± 0.169
	Smote	0.769 ± 0.161	0.469 ± 0.082	0.588 ± 0.057	0.597 ± 0.061
	BSmote	0.769 ± 0.171	0.469 ± 0.082	0.590 ± 0.062	0.597 ± 0.071
haberman	Weight	0.777 ± 0.158	0460 ± 0.109	0.586 ± 0.053	0.589 ± 0.082
	RUS	0.645 ± 0.142	0.562 ± 0.052	0.599 ± 0.083	0.661 ± 0.051
	AdaSyn	0.750 ± 0.138	0.491 ± 0.093	0.599 ± 0.053	0.614 ± 0.071
	SPU	0.823 ± 0.106	0.460 ± 0.073	0.612 ± 0.061	0.599 ± 0.069
	NOBDF	0.851 ± 0.050	0.972 ± 0.017	0.909 ± 0.031	0.841 ± 0.031
	SVM	1.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	RU	0.963 ± 0.042	0.121 ± 0.088	0.297 ± 0.170	0.204 ± 0.104
	Smote	0.719 ± 0.095	0.546 ± 0.065	0.623 ± 0.033	0.689 ± 0.052
	BSmote	0.715 ± 0.094	0.547 ± 0.064	0.622 ± 0.029	0.690 ± 0.050
yeast	Weight	0.750 ± 0.102	0.504 ± 0.057	0.611 ± 0.028	0.656 ± 0.046
	RUS	0.671 ± 0.082	0.520 ± 0.033	0.590 ± 0.037	0.667 ± 0.028
	AdaSyn	0.704 ± 0.095	0.560 ± 0.064	0.625 ± 0.036	0.700 ± 0.052
	SPU	0.749 ± 0.093	0.520 ± 0.089	0.618 ± 0.033	0.667 ± 0.075
	NOBDF	0.912 ± 0.020	0.868 ± 0.028	0.889 ± 0.016	0.884 ± 0.021
	SVM	1.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	RU	0.966 ± 0.047	0.112 ± 0.105	0.268 ± 0.189	0.186 ± 0.165
	Smote	0.630 ± 0.112	0.628 ± 0.082	0.623 ± 0.035	0.753 ± 0.059
	BSmote	0.635 ± 0.118	0.629 ± 0.077	0.626 ± 0.046	0.754 ± 0.058
abalone	Weight	0.639 ± 0.110	0.620 ± 0.089	0.623 ± 0.029	0.747 ± 0.064
	RUS	0.616 ± 0.063	0.580 ± 0.077	0.595 ± 0.041	0.716 ± 0.060
	AdaSyn	0.593 ± 0.089	0.657 ± 0.074	0.620 ± 0.037	0.773 ± 0.055
	SPU	0.658 ± 0.047	0.606 ± 0.105	0.630 ± 0.189	0.740 ± 0.165
	NOBDF	0.977 ± 0.018	0.751 ± 0.019	0.857 ± 0.010	0.764 ± 0.011

分析表 4 可见: 传统 SVM 分类器在不平衡数据集下进行分类, Specificity 评价机制全部为 1, 而 Sensitivity、G-mean、F-measure 等评价机制几乎为 0, 这表明分类超平面向正类方向发生了严重的偏移, 以至于分类效果严重不良; 而其余改进算法形式的 Sensitivity、G-mean、F-measure 评价机制具有较明显的提升, 其中本文所提出的 NOBDF 算法在 4 种评价机制上均明显优越于其余算法, 且具有大幅度提升, 这说明本文所提算法在不同空间结构以及不同维度的不平衡数据集下拥有更强的正负类识别率, 构成的 NOBDF-SVM 分类器拥有更好的综合性能; 此外, 观察每种算法的均方差, 发现除传统 SVM 算法外, 相对于其余改进算法, 本文所提出的 NOBDF 算法具有最小的均方差值, 这表明本文算法的随机性影响远小于其他算法.

3.4.2 新型过采样方式与 SVM 分类器改进算法对比

模糊支持向量机是对传统 SVM 的一种改良算法, 且在不平衡数据下有较好的分类效果[21], 故为了突出本文所提出算法的优越性, 将所提出的分类方式与现有模糊支持向量机及其改进算法的分类效果进行对比, 并将结果汇总于表 5.

表 5 NOBDF-SVM 与模糊支持向量机及其改进形式的对比

Dataset	Methods	Specificity	Sensitivity	G-mean
Pima	FSVM	0.808 ± 0.008	0.560 ± 0.016	0.623 ± 0.009
	IFSVM	0.741 ± 0.005	0.739 ± 0.013	0.740 ± 0.007
	NOBDF-SVM	0.969 ± 0.015	0.883 ± 0.010	0.925 ± 0.009
haberman	FSVM	0.768 ± 0.014	0.364 ± 0.031	0.529 ± 0.021
	IFSVM	0.791 ± 0.010	0.538 ± 0.019	0.652 ± 0.012
	NOBDF-SVM	0.851 ± 0.050	0.972 ± 0.017	0.909 ± 0.031

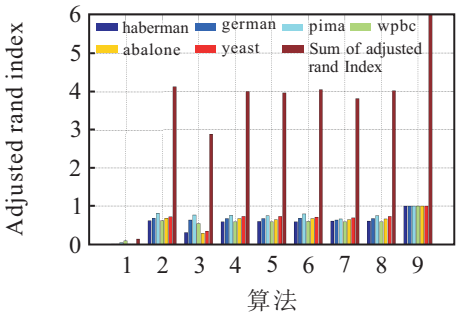
分析表 5 可知: 相对于在算法上改进 SVM 分类器, NOBDF-SVM 在 Specificity、Sensitivity、G-mean 上也有较大的提升, 虽然抗随机性影响并不一定比 FSVM 等算法强, 这表明在原始数据集重构数据, 必定会加入一些随机影响的干扰. 综上可知, 无论是在原始数据集上进行重构, 还是对分类器进行改进, NOBDF-SVM 算法对不平衡数据的分类效果明显优于已有的 SVM 改进算法.

3.5 NOBDF 采样算法鲁棒性的研究

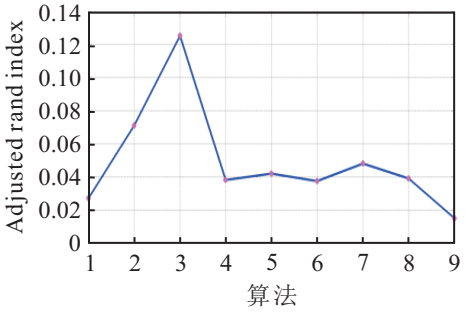
为了验证 NOBDF 算法的优越性, 本文对所提出的算法与其他 8 种算法的鲁棒性进行对比研究, 采用参考文献 [18] 的算法鲁棒性的评价机制, 即算法 m 在某一特定数据集上的相对性能以该算法求解该问题时得到的 Adjusted rand index 的值与最大的 Adjusted rand index 值的比值进行衡量, 计算公式如下:

$$b_m = R_m / \max(R_k),\ m = 1, 2, \cdots, k. \tag{13}$$

则在某个数据集上效果最好的算法 m^* 的相对性能即为 $b_{m^*} = 1$, 而其他算法的相对性能 $b_m \leqslant 1$, 且 b_m 值越大, 相应算法 m 在所有算法中的相对性能越好. 因此, 算法 m 的鲁棒性可以用该算法在所有数据集上 b_m 值的总和来评价, 且总和越大的算法鲁棒性越强. 为了验证本文所提算法的鲁棒性以及在不平衡比进一步加大时的算法性能, 在选用前文所提的 6 种数据集的同时, 正负类取样比变为 1: 20, 同样采用十折交叉验证, 各类算法的设置亦同上, 最后将 9 种算法的 G-mean 评价机制的鲁棒性及其统计平均误差进行比较, 如图 4 所示. 图 4 中算法 1 ~ 算法 9 分别代表 SVM、SPU、RU、Smote、BSmote、Weight、RUS、AdaSyn、NOBDF.



(a) 鲁棒性的比较



(b) 鲁棒性统计平均误差的比较

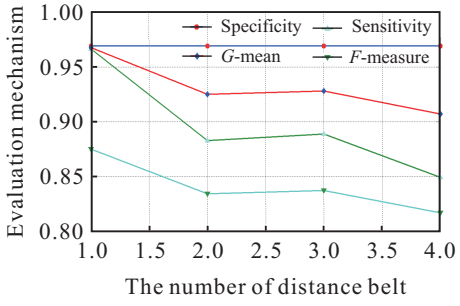
图 4 不平衡数据集下, 9 种算法 G-mean 的鲁棒性比较

分析图4可知: 本文所提NOBDF算法的b_m均为1, 获得最高的总和值, 这表明对于不同空间结构以及不同维度的不平衡数据集的分类, NOBDF算法均表现出良好的性能. 其次文献[18]中的SPU算法也表现不错, 但此算法是基于样本特性的欠取样选择负类样本的, 可能会丢失重要的样本信息. 而本文的算法不但考虑到样本的特性, 还对样本划分距离带, 此新型过采样方式不会丢失原有的数据集信息, 且克服了一般过采样方式的盲目性以及易发生混叠样本的弊端, 故NOBDF能够更好地适用于SVM算法, 更好地解决了分类超平面偏移问题. 此外, 还可以看出: 在数据集的不平衡比发生变化的情况下, NOBDF算法仍具有较好的效果.

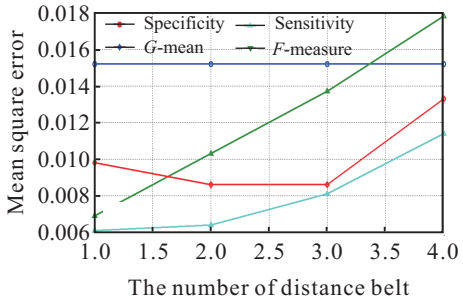
3.6　距离带的数目对NOBDF算法的影响

NOBDF算法的核心之一就是初始带的划定, 因为带数必定会影响带内的正负类个体的数目, 以影响ANBSC-Smote合成新个体的数目、邻域值等重要因素, 故在此探究一下带宽的划定对分类效果的影响. 选用Pima和Abalone两个数据集, 为了简化带宽探究过程, 正负类取样比为1:10, 同样采用十折交叉验证, SVM分类器同样采用以上的设置, 其中两个数据集的带数均分别为1、 2、 3、 4. 将不同带数的Specificity、Sensitivity、G-mean以及F-measure的值与统计平均误差汇总于图5和图6.

分析图5(a)和图6(a)可知: 以G-mean值为例, pima数据集的效果最优出现在带数为1的情况下, 而abalone数据集的最优效果出现在带数为2的情况下,

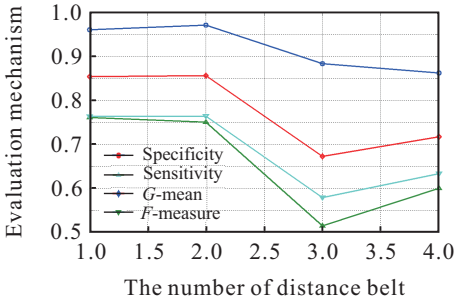


(a) 分类效果

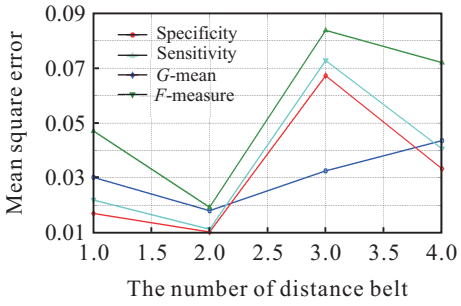


(b) 统计平均误差

图5　不同距离带数对Pima数据集分类效果的影响



(a) 分类效果



(b) 统计平均误差

图6　不同距离带数对Abalone数据集分类效果的影响

且两个数据集在随带数的增加出现上下波动, 分析Specificity、Sensitivity以及F-measure值也有类似的情况. 分析图5(b)和图6(b)可知: pima数据集在带数为1的情况下总平均误差最低(即同一带数下4种评价机制平均误差的加和), 且不难发现abalone数据集在带数为2的情况下总平均误差最低. 同时分析图5和图6还可以看出: 分类精度高的带数必然会有较低的统计平均误差. 综上可知, 距离带的划分还是很有必要的, 因为合理的划分距离带不但会提高分类精度, 而且产生的误差也会减少, 但并非带数越多分类的效果就越好, 如果想要寻找更加理想的结果, 须寻找更加合理的带数.

4　结　论

针对已有欠(过)采样方式在不平衡数据集下分类效果不明显、鲁棒性不强等缺点, 本文综合考虑数据样本的类内间距、 类间间距以及不平衡度等因素, 提出了一种基于样本特性的新型过采样算法(NOBDF). 该算法在不平衡数据集划分距离带的基础上, 利用一种改进的基于样本特性自适应变邻域的Smote算法(ANBSC-Smote)对带内正类进行新样本的合成, 以达到带内数据平衡, 最终使整个数据集达到平衡. 实验结果表明: 相比已有的欠(过)采样方式, 本文所提出的算法无论在正负类分类精度上, 还是在总体分类性能上, 均有明显提升, 且具有更强的鲁棒性, 这为不平衡数据的分类提供了一个有效的理论模型. 此外, 本文还进行了原始数据集距离带数目对NOBDF算法的影响探究. 研究结果表明, 带数的划分

将会影响样本合成及数据集分类结果,如何对带数进行合理的划分,是原始数据集分类的重要影响因素,研究团队将在后续研究中进行深入探讨.

参考文献(References)

[1] 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 J, Feng L. An algorithm of robust online extreme learning machine for dynamic imbalanced datasets[J]. J of Computer Research and Development, 2015, 52(7): 1487-1498.)

[2] 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.

[3] 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 Y, Liu Z J, Zhang H J. Review on ensemble algorithms for unbalanced data classification[J]. Application Research of Computers, 2014, 31(5): 1287-1291.)

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

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

[6] 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 L X, Guo H, Wang J J. A mechanical fault severity identification method under unbalanced datasets[J]. J of Vibration and Shock, 2016, 35(20): 178-182.)

[7] 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.

[8] 陶新民, 刘福荣, 童智靖, 等. 不均衡数据下基于SVM的故障检测新算法[J]. 振动与冲击, 2010, 29(12): 8-12.
(Tao X M, Liu F R, Tong Z J, et al. A new fault detection method of unbalanced data based on SVM[J]. Expert Systems with Applications, 2010, 29(12): 8-12.)

[9] 付忠良. 通用集成学习算法的构造[J]. 计算机研究与发展, 2013, 50(4): 861-872.
(Fu Z L. A univer-sal ensemble learning algorithm[J]. J of Computer Research and Development, 2013, 50(4): 861-872.)

[10] Wang S, Xi L. Condition monitoring system design with one-class and imbalanced-data classifier[C]. Int Conf on Industrial Engineering and Engineering Management. Beijing: IEEE, 2009: 779-783.

[11] 杨智明, 乔立岩, 彭喜元. 基于改进SMOTE的不平衡数据挖掘方法研究[J]. 电子学报, 2007, 35(b12): 22-26.
(Yang Z M, Qiao L Y, Peng X Y. Research on datamining method for imbalanced dataset based on improved Smote[J]. Acta Electronica Sinica, 2007, 35(b12): 22-26.)

[12] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]. Int Conf on Intelligent Computing. Berlin: Springer, 2005: 878-887.

[13] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. IEEE Int Joint Conf on Neural Networks. Hong Kong: IEEE Xplore, 2008: 1322-1328.

[14] Masnadishirazi H, Vasconcelos N, Iranmehr A. cost-sensitive support vector machines[J]. 2015, 1(1): 1-26.

[15] 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 X J, Sun Y X, Liu H T. Clustering boundary over-sampling classification method for unbalanced data sets[J]. J of Zhejiang University: Engineering Science, 2013, 47(6): 944-950.)

[16] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets[C]. The 15th European Conf on Machine Learning. Pisa: Springer, 2004: 39-50.

[17] Batista G E A P A, Prati R C, Monard M C. A study of the behavior of several methods for balancing machine learning training data[J]. Acm Sigkdd Explorations Newsletter, 2004, 6(1): 20-29.

[18] 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, 28(7): 978-984.
(Tao X M, Hao S Y, Zhang D X, et al. Support vector machine for unbalanced data based under-sampling approaches on sample properties[J]. Control and Decision, 2013, 28(7): 978-984.)

[19] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[20] 曹愈远, 张建, 李艳军, 等. 基于模糊粗糙集和SVM的航空发动机故障诊断[J]. 振动测试与诊断, 2017, 37(1): 169-173.
(Cao Y Y, Zhang J, Li Y J, et al. Aero-engine fault diagnosis based on fuzzy rough set and SVM[J]. J of Vibrations, Measurement & Diagnosis, 2017, 37(1): 169-173.)

[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 Z, Cao J Z, Gu H. A fuzzy support vector machine algorithm for unbalanced data classifications[J]. J of Dalian University of Technology, 2016, 56(5): 525-531.)

(责任编辑: 齐 霁)