第 55 卷 第 1 期
2019 年 1 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol.55, No.1
Jan.,2019

DOI:10.13232/j.cnki.jnju.2019.01.001

面向非平衡多分类问题的二次合成 QSMOTE 方法

韩明鸣[1],郭虎升[1],王文剑[2*]

(1.山西大学计算机与信息技术学院,太原,030006;
2.计算智能与中文信息处理教育部重点实验室,山西大学,太原,030006)

摘 要:近年来非平衡多分类数据的学习问题在机器学习和数据挖掘领域备受关注,上采样技术成为解决数据不平衡问题的主要方法,然而已有的上采样技术仍有很多的不足,例如新合成的少数类样本仍可能分布在对应少数类样本的原始区域内,不能有效改善数据分布的不平衡情况. 此外,若原始样本中不同类别样本分布存在重叠,则新合成的样本会更容易偏离到其他类样本分布中,从而造成过泛化现象,影响少数类样本的分类精度. 为解决上述问题,提出一种二次合成的上采样方法(Quadratic Synthetic Minority Over-sampling Technique,QSMOTE). 首先通过少数类样本的支持度选择包含重要信息的样本来进行第一次合成,然后通过分析指定少数类样本质心的邻域内样本分布情况来调整第二次样本合成范围,并最终进行第二次合成. 在 UCI 和 MNIST 数据集上的实验结果表明,QSMOTE不仅可以改善数据分布的不平衡问题,而且可以尽可能地减少过泛化现象,特别是对少数类样本的分类准确率有大幅提升.

关键词:多类非平衡问题,过泛化,重 叠,合成少数类上采样技术(SMOTE)
中图分类号:TP18 **文献标识码**:A

Quadratic synthetic minority over-sampling technique for classification of multiclass imbalance problems

Han Mingming[1], Guo Husheng[1], Wang Wenjian[2*]

(1.School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, China;
2.Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University, Taiyuan, 030006, China)

Abstract: In recent years, multiclass imbalance data learning has attracted increasing interests on the domain of machine learning and data mining. Over-sampling is the most popular technique to solve the problem of imbalanced classification. However, the existing approaches based on over-sampling have some limits, such as the new synthetic samples located in the area of the initial region of their own class could not improve the class imbalance distribution actively in data space. In addition, the samples belong to different classes in original data space may be overlapping, which will lead the synthetic samples to deviate to the areas of other classes. It will cause serious over generaliza-

基金项目:国家自然科学基金(61673249,61503229),山西省回国留学人员科研基金(2016-004),赛尔网络下一代互联网技术创新项目(NGII20170601)
收稿日期:2018-08-19
* 通讯联系人,E-mail:wjwang@sxu.edu.cn

tion, and then decrease the accuracy of the minority classes. In order to solve these problems, the Quadratic Synthetic Minority Over-sampling Technique, termed QSMOTE, is proposed. Firstly, the samples with more important information are selected from the samples with large support value during the first synthesis. In the second phase, the synthetic sphere is adjusted by the distribution of samples on a certain domain, which is defined by the center of mass of the minority classes. Substantial experiments on UCI and MNIST data sets demonstrat that the proposed QSMOTE algorithm can not only decrease the imbalance of data distribution, but also avoid over generalization as much as possible. Moreover, it can perform well on the classification accuracy of unbalanced data sets, especially for the minority data.

Key words: multiclass imbalance problems, over generation, overlapping, Synthetic Minority Over-sampling Technique(SMOTE)

在很多实际应用中,如网络入侵、文本分类、医疗诊断等,分类处理的对象大多是不平衡数据集,即某些类别的样本要比其他类的样本数目多[1]. 其中,数目较少的类通常被称为少数类,数目较多的类被称为多数类[2]. 传统的分类器在分类判决时总会倾向多数类,而忽略少数类[3],从而导致分类器对于少数类的分类准确率明显下降. 然而在现实中,这些少数类样本往往包含我们更需要的信息,因此,如何有效地提高少数类的分类准确率和分类器的整体性能已成为数据挖掘领域的一个热点[4].

目前对不平衡数据分类问题的处理方法总体上分为两类:一是通过重采样技术对数据进行预处理,包括下采样技术和上采样技术[5]. 其中下采样技术的基本思想是删除部分多数类样本[6],但可能会造成分类信息的丢失;而上采样技术则是增加少数类样本的数量,保留甚至增加少数类的分布信息[7]. 二是对现有的分类算法进行修改或者提出新的分类算法[8],如代价敏感学习[9](Cost-Sensitive Learning)、集成式(Boosting)学习[10]、核方法[11]等,这些方法从算法层面解决非平衡多分类问题,虽然理论上简单易行,且在部分小规模数据的应用中取得了不错的结果,但随着样本规模和样本分布复杂性的增加,算法的运行时间急剧增加而且优化过程也更复杂,此外分类器的重建也将会面临很大的挑战[12]. 上述方法中,重采样技术不依赖数据分布和分类算法[13],还可以和任何分类算法相结合以进一步提升少数类样本分类精度,因此重采样技术是目前解决非平衡多分类问题最常用的方法.

上采样技术是从最根本的层面来解决非平衡问题. 通过向原始的非平衡数据集中引入新的少数类样本使新数据集达到平衡,进而提升少数类样本的分类精度. 此外,由于下采样方法在移除部分多数类样本时可能损失含有重要信息的样本,会对分类造成一定影响,尤其在面对样本数量较小的非平衡数据时,上采样方法相较下采样方法能取得更好的分类效果,因此重采样技术中上采样技术的使用更为流行.

目前针对非平衡多分类问题采用的上采样方法是基于 SMOTE[14](Synthetic Minority Over-sampling Technique)及其衍生的相关方法. SMOTE 利用原始数据中少数类样本及其近邻样本人工合成新的样本,来达到平衡样本的目的. 与较传统的随机上采样技术相比,它能有效地避免分类器的过拟合现象[15],但是 SMOTE 算法也存在着诸如无法精确控制合成样本数量、对少数类样本选择存在盲目性、没有考虑少数类样本与多数类样本的联系、新合成样本容易导致过泛化现象等不足. 因此学者们提出了许多改进型算法. 如基于 Borderline-SMOTE[16]的方法,通过少数类的边界点来合成新样本,对少数类样本进行有区别的选择,可以有效避免冗余样本的生成. 基于 ASMOTE[17-18]的方法考虑了少数类样本的分

布信息,自适应地调整合成样本产生过程中的近邻选择策略,在一定程度上避免了样本重叠现象,但合成少数类新样本分布在其他类的分布范围内,会使样本分类准确率降低. 基于 Random-SMOTE[19-20]的方法通过三个少数类最近邻合成一个新的少数类样本,该方法有效改善了少数类样本内分布的不均匀性. 这些方法在一定程度上提高了 SMOTE 的性能,但是仍然存在如下问题:

(1)它们都没有真正做到对少数类样本进行区别选择,即使通过考虑每个少数类样本的近邻样本信息舍弃了部分可能的孤立样本,但没有充分挖掘和异类样本具有较强联系的样本,使得新合成的样本容易分布在异类样本的分布区域,造成过泛现象.

(2)它们虽然考虑了多数类样本与少数类样本的联系,但只是利用和异类样本的联系来确定新合成样本在不同区域合成的数量,并不能达到改善样本分布不平衡的目的.

针对上述问题,本文提出了一种二次合成 QSMOTE 算法,通过利用少数类样本中包含重要信息的样本来合成新的样本,从而尽可能减小过泛化影响,并考虑通过少数类的类质心的邻域内样本的分布情况来调节样本合成范围并合成新样本,改善样本分布的不平衡问题,提升少数类样本的分类准确率.

1　SMOTE 方法简介

SMOTE 方法不同于传统的基于样本复制的上采样方法,它通过人工合成新的样本,并控制新合成样本的数量和分布来实现平衡样本集的目的,可以有效解决传统上采样方法由于决策区域过小而引起的分类器过拟合问题. SMOTE 方法本质上是通过在少数类样本与其对应的同类近邻样本间进行线性插值来合成新的少数类样本:首先设置一个上采样倍数 R,并根据 R 从每个少数类样本的 k 个同类最近邻中随机选择 R 个样本,然后利用每个少数类样本与它的 R 个选中的样本分别合成新少数类样本,最后将全部新合成样本加入到原训练样本集中,构成新的训练样本集. 合成一个新样本 x_{syn} 的方法如下:

$$x_{\text{syn}}=x_i+(x_j-x_i)\times\text{rand}(0,1) \quad (1)$$

x_i 表示少数类中参与合成新样本的指定样本;x_j 表示 x_i 对应的第 j 个同类近邻样本,$j=1,\cdots,R$;rand(0,1)将会产生 0 到 1 的随机数.

虽然 SMOTE 方法和传统的上采样技术相比,能对少数类的分类精度有不错的提升,但它合成的新样本容易产生过泛化现象,即由于合成的少数类错误新样本使得少数类样本分布区域被泛化到多数类区域中,降低了少数类样本的可学习性,影响分类精度. 因此当 SMOTE 方法应用到非平衡多分类问题时,不同少数类样本之间的过泛化现象将会更加明显.

2　QSMOTE 方法

SMOTE 方法没有区别对待少数类样本,也没有考虑少数类样本与异类样本的联系,所以导致新合成样本非但没有有效改善原始样本分布的不平衡问题,还造成了过泛化影响,降低了少数类样本的分类精度. 而本文提出的二次合成的上采样方法(Quadratic SMOTE),通过对原始样本中的少数类样本采用二次合成的方式合成新样本,可以改善数据分布的不平衡问题并提升少数类样本的分类准确率. QSMOTE 算法充分利用了异类近邻样本的分布信息,并且对原先少数类样本中包含重要信息的样本进行了选择,使得合成的新样本不仅有效地扩充和强化了少数类的分布区域边界,还尽可能地减少了过泛化现象.

为了便于描述,记训练集为 D,少数类样本集为 P,多数类样本集为 N,则 $D=\{(x_i,y_i)\mid y_i\in\{1,\cdots,n\}\}_{i=1}^{l}$,$P=\{P^1,\cdots,P^{pnum}\}$,$N=\{N^1,\cdots,N^{nnum}\}$,其中 $pnum$ 为少数类样本类别数,$nnum$ 为多数类样本类别数,则 $n=pnum+nnum$. 令 $ave(D)=l/n$,将其作为多数类与少数类的区分标准,若某类样本数目小于 $ave(D)$则被视作少数类,否则视作多数类. 这样可以最大程度找到不平衡的少数类.

2.1 第一次合成 第一次合成的主要目的是从全部少数类样本中选择包含重要信息的样本,并用这些样本来进行第一次新样本合成. 这些含有重要信息的样本应该具备这样的特点:它们位于同类样本分布占据主导的区域内,并且它们和别的类样本之间具有最小距离. 这些样本一般分布在同类样本的边界,因此可以利用它们合成潜在的边界样本.

本文通过定义支持度来获取包含重要信息的样本点参与第一次新样本的合成. 支持度定义为:

$$sop(x_i)=\frac{m}{k} \tag{2}$$

其中 $x^i\in P^i$,m 为 x_i 的 k 近邻中与其同类的样本点的个数,k 为近邻参数. 将不低于支持度阈值的少数类样本的选择权重赋值为 1,否则设置为 0,这样可以排除选择的样本是噪声点或者分布在异类点区域的可能性. 对于选择权重为 1 的全部少数类样本,分别求它们与其他类样本之间的最小类间距,并记录不同最小类间距对应的最佳参照样本. 最小类间距定义为:

$$\mathrm{Min}Dis(A,B)=\min\{Dis(x_i,x_j)\mid x_i\in A,x_j\in B\} \tag{3}$$

其中 $A\in P$,$B\in(D-A)$,$Dis(x_i,x_j)$ 表示两个不同类样本点之间的距离,x_i 称为最小类间距对应的最佳参照样本. 最后用全部的最佳参照样本和它们对应的最小类距,按照下式合成新的同类样本 x_{syn}.

$$x_{\mathrm{syn}}=x_i+\mathrm{rand}(-0.5,1)\times\mathrm{Min}Dis(A,B)\times\frac{x_j-x_i}{\|x_i-x_i\|} \tag{4}$$

其中 x_i 为指定少数类的最佳参照样本,x_j 为同类样本中满足支持度阈值的样本点. 对于满足条件的不同的样本 x_j,需重复上式合成不同的新样本. 最后将所有的少数类均按照上述方法进行第一次合成,并将合成样本与 D 合并组成新的训练集 D'.

第一次合成过程中,因为充分考虑了少数类样本分布与其他类样本的联系,并且利用含有重要信息的少数类样本参与新样本的合成,所以第一次合成的新样本可以有效避免过泛化影响.

2.2 第二次合成 第二次合成首先对 D' 中不同类别样本数量进行统计,求得各个少数类的质心和最大类内距离. 最大类内距离的定义为:

$$\mathrm{Max}Dis(A)=\max\{Dis(x_i,\bar{x})\mid x_i\in A\} \tag{5}$$

其中 $A\in P$,$Dis(x_i,\bar{x})$ 为少数类 A 中任意样本 x_i 与其质心 $\bar{x}$ 的距离. 然后以质心 $\bar{x}$ 为圆心,以最大类内距离 $\mathrm{Max}Dis(A)$ 为半径的超球空间作为第二次样本合成的合成范围,这个合成范围又称为质心的邻域. 计算每个少数类样本质心的邻域内异类点占比值 com,对异类点占比值不满足条件(不低于给定阈值)的少数类适当调整最大类内距离,直到重新确定的合成范围内的异类点占比值满足条件. com 的定义为:

$$com=\frac{others}{|A'|} \tag{6}$$

其中 $others$ 为指定少数类样本质心的邻域内异类点的个数,$|A'|$ 为指定少数类的质心的邻域内全部样本点的个数. 若最大类内距离需要调整,则其值大小按照质心和其他样本的距离远近依次变化. 最后用下式进行第二次新样本 x_{syn} 的合成.

$$x_{\mathrm{syn}}=\bar{x}+\mathrm{Max}Dis(A)\times\frac{x_j-\bar{x}}{\|x_j-\bar{x}\|} \tag{7}$$

其中 x_j 为指定少数类的质心的邻域内同类样本.

当样本数有多个时,用式(7)重复生成新样本,并将合成样本与 D' 结合形成最终训练集 D^*. 第二次合成在第一次合成新样本的基础上,在满足条件的质心邻域内合成新样本,保证了新合成样本在尽可能减少过泛化现象的同时,最大程度扩充了少数类样本的分布区域.

图 1 是本文算法二次合成样本过程的示意图. 图中共包括四类样本点,其中有一类是多数类(用圆形表示),其余三类是少数类,且三角形代表用于合成新样本的指定少数类;近邻参数 $k=3$,支持度阈值 $sop=0.3$.

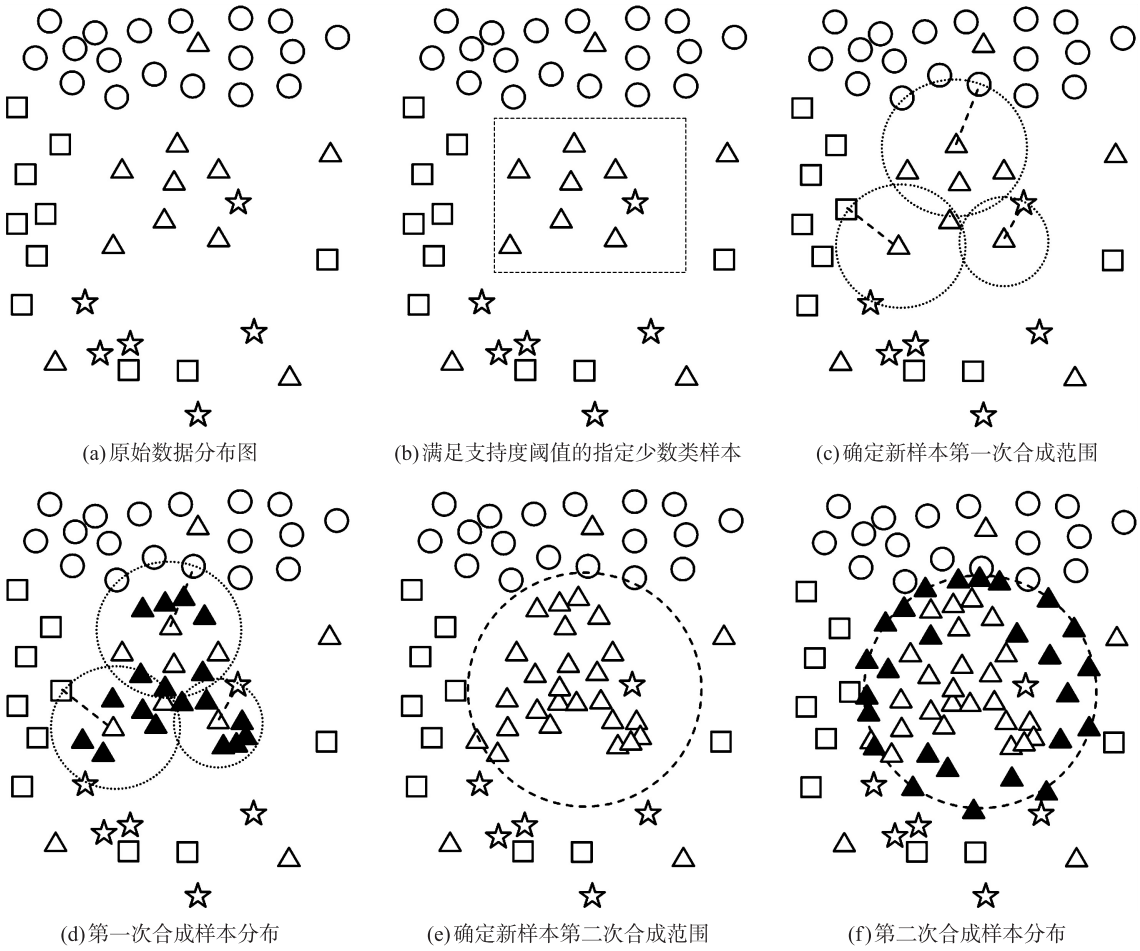

图 1　QSMOTE 算法二次合成样本示意图

Fig. 1　The diagrammatic sketch of QSMOTE algorithm synthesis samples

图 1a 为样本的初始分布情况,图 1b 中虚线矩形框选中的样本为满足少数类样本支持度阈值的指定少数类样本,图 1c 中虚线圆表示新合成样本的合成区域,且不同的虚线圆区域代表以不同最佳参照样本为中心的新样本合成范围,图 1d 为第一次样本合成后全部样本的分布情况(黑色三角形代表全部新生成样本,共 18 个),图 1e 中虚线圆为第二次合成新样本的最大范围,图 1f 为第二次合成后的样本分布(黑色三角形表示新生成的样本,共 26 个).

从图中可以看出,指定少数类样本经过两次合成之后,不仅数量关系上和多数类样本基本达到平衡,而且合成之后样本的分布区域在没有入侵到其他类样本所属区域的同时,相较原始分布有了一定程度的扩展.

2.3　QSMOTE 算法　QSMOTE 算法的主要步骤如下:

QSMOTE **算法**

输入:原始训练数据集 D,近邻参数 k,支持度阈值 sop,$com=0.1$

输出:将新样本合并后新的训练集 D^*

Step1:将 D 中全部样本根据每类样本数量与 $ave(D)$ 比较,划分到多数类 N 与少数类 P;

Step2:对每个少数类 P^i 进行第一次合成;

Step2.1:求得 P^i 中全部样本的支持度 $sop(P^i)$,并将满足 $sop(P^i)\geqslant sop$ 的样本选择权重赋值为 1,不满足的赋值为 0,将 P^i 中全部选择权重为 1 的样本构成集合 $Select(P^i)$;

Step2.2:分别求得 $Select(P^i)$ 与其他类的最小类

间距 MinDis($Select$(P^i),B),其中 $B\in(D-P^i)$,并记录其各自对应的最佳参照样本;

Step2.3:用式(4)将 Step2.2 得到的最佳参照样本和对应的最小类间距合成属于 P^i 类的新样本,且合成数目为各少数类的 $Select(P^i)$求和;

Step3:将全部合成样本与 D 合并构成 D',重新划分多数类 N 和少数类 P;

Step4:对 D'中的每个少数类 P^i 进行第二次合成;

Step4.1:求每个少数类 P^i 的质心 $\bar{x}$ 和最大类内距离 MaxDis(P^i);

Step4.2:计算每个少数类 P^i 的质心的邻域内 com 值,若 $com\geqslant 0.1$ 则适当调整最大类内距离并重复 Step4.2,否则执行 Step4.3;

Step4.3:在满足条件的最大合成范围内按式(7)进行第二次合成,且合成样本数目为各少数类最大类内距离内同类样本数的求和;

Step5:将合成样本与 D'合并构成 D^*;

Step6:算法结束.

本文提出的二次合成算法其时间复杂度主要由计算少数类样本支持度、计算最小类间距、计算最大类内距离三部分组成. 假设少数类样本总数为 t,每个少数类样本均需计算它和剩余 $l-1$ 个样本的距离并最终获得其对应的 k 近邻样本,则少数类样本支持度的时间复杂度为 $O(t\times(l-1)\times k)<O(l^2)$;最小类间距计算过程中,每个指定少数类均和其他类样本进行距离计算,其时间复杂度不超过 $O(tl)$;最大类内距离只考虑每个少数类样本内部的计算,所以其时间复杂度不超过 $O(l)$. 综上所述,该算法的时间复杂度不超过 $O(l^2)$.

3 实验与结果

3.1 实验数据集和参数设置 为了验证本文提出的 QSMOTE 算法的有效性,选取 UCI 数据集和数字手写集 MNIST 进行实验,数据集说明如表 1 所示. 实验数据按照训练集∶测试集为 4∶1 的比例进行划分,按照相同比例在原数据集中重复抽取 10 次,并对每次划分结果采用 10-折交叉验证进行实验,得到对应的最优参数. 将 QSMOTE 方法与两种经典的采样方法 SMOTE 和 Borderline-SMOTE 进行比较,分类采用 LibSVM 工具箱中的 SVM 分类器进行分类预测,选择 RBF 核为核函数,通过实验获得不同数据集下的最优参数 C 和 σ. 数据合成算法中近邻参数 k 依次为 5,10,15,20;支持度阈值 sop 依次为 0.3,0.5,0.8.

表 1 实验数据集

Table 1 Summary description of datasets

	Data set	# Attributes	# Examples	# Classes	# The minority classes	# The minority/# The majority
UCI	Glass	9	214	6	4	9/76
	Page_Block	10	5473	5	4	28/4913
	Yeast	8	600	3	2	50/450
	Abalone	8	2200	3	2	100/1500
	Ecoli	7	272	3	1	52/143
MNIST		784	4450	10	9	50/4000

3.2 评价指标 本文采用 g-$means$,$precision$,$recall$,F 值等指标来衡量分类器的分类性能. g-$means$ 用来衡量分类器对全部样本点的分类性能,由少数类样本和多数类样本分类召回率共同决定,其表达式如下:

$$g\text{-}means=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \quad (8)$$

其中 TP 表示正类(少数类)样本判为正类的个数,TN 表示负类(多数类)样本判为负类的个数,FN 和 FP 分别表示判决错误的实际正类和负类样本数目. 在非平衡问题中,g-$means$ 值越大说明分类器性能越好,少数类样本的召回率越高. $recall$ 表示样本的召回率,$precision$ 表示样本的分类预测准确率,具体表示为:

$$recall=\frac{TP}{TP+FN} \tag{9}$$

$$precision=\frac{TP}{TP+FP} \tag{10}$$

F 值充分考虑了 $recall$ 和 $precision$ 结果,只有当两者都足够高,F 值才会越高,β 在通常情况下为 1. 但在非平衡的分类问题中,在保持整体分类准确率的基础上,能识别到更多的少数类样本更值得关注. F 值的表达式为:

$$F=\frac{2\times recall\times precision}{recall+precision} \tag{11}$$

3.3　实验结果及其分析

3.3.1　参数对算法的影响　本文算法主要有 k 和 sop 两个参数,它们通过改变不同样本的近邻样本个数和近邻样本分布情况来影响最佳参照样本的获取,并最终影响合成样本的分布. 表 2 至表 5 为在不同 k 值下,不同算法的 $g\text{-}means$ 和 F 值结果比较,其中黑色字体表示不同参数下的最优结果. 所有实验结果均为循环 100 次所取得的平均值.

从表 2 中可以看出,当 $k=5$ 时,直接使用 SVM 进行分类的方法的 $g\text{-}means$ 和 F 值有很多为 0,且在不为 0 的结果中 $g\text{-}means$ 和 F 值明显低于别的方法,说明未经采样方法处理的数据在进行分类实验时,少数类样本的召回率和准确率很低,甚至存在少数类样本未被识别的现象. 采用 SMOTE,Borderline-SMOTE 和 QSMOTE 处理过的数据,再用 SVM 进行分类实验,则分类结果整体好于前者,这表明使用上采样技术确实可以提升非平衡数据分类的性能,尤其是提升对少数类样本分类的性能. 采用 SMOTE 和 Borderline-SMOTE 的结果显示,在数据集 Glass 和 Yeast 中,sop 虽然不同,但 SMOTE 的分类结果均低于 Borderline-SMOTE;在 Page_Block 数据集中,当 sop 为 0.3

表 2　不同支持度下各算法的实验结果($k=5$)

Table 2　The experimental results of different algorithms with different sop ($k=5$)

sop	Dataset	SVM		SVM+SMOTE		SVM+Borderline-SMOTE		SVM+QSMOTE	
		$g\text{-}means$	F	$g\text{-}means$	F	$g\text{-}means$	F	$g\text{-}means$	F
0.3	Glass	0.7581	0.8000	0.7581	0.8000	0.8476	0.9091	**0.8526**	**0.9251**
	Page_Block	0.3991	0.2553	**0.4100**	**0.2676**	0.3991	0.2553	**0.4100**	**0.2676**
	Yeast	0	0	0.5235	0.3509	0.6164	0.4318	**0.6495**	**0.4789**
	Abalone	0	0	0.2994	0.1451	0.1420	0.3396	**0.4890**	**0.4268**
	Ecoli	0	0	**0.2821**	**0.3896**	0	0	0.1784	0.3038
	MNIST	0.9006	0.8439	**0.9178**	**0.8492**	0.9006	0.8439	**0.9178**	**0.8492**
0.5	Glass	0.7581	0.8000	0.8476	0.9091	0.8476	0.9091	**0.8476**	**0.9091**
	Page_Block	0.3991	0.2553	**0.4206**	**0.2797**	0.4100	0.2676	0.4098	0.2657
	Yeast	0	0	0.5164	0.3390	0.5888	0.4045	**0.6495**	**0.4789**
	Abalone	0	0	0.2896	0.1376	0.1205	0.3333	**0.4832**	**0.4281**
	Ecoli	0	0	**0.2821**	**0.3896**	0	0	0.2523	0.3117
	MNIST	0.9006	0.8439	0.8995	0.8249	0.9006	0.8439	**0.9129**	**0.8671**
0.8	Glass	0.7581	0.8000	0.7581	0.8000	0.8305	0.9091	**0.8476**	**0.9091**
	Page_Block	0.3991	0.2553	0.3991	0.2553	**0.4100**	**0.2676**	0.3991	0.2553
	Yeast	0	0	0.5029	0.3279	0.6495	0.4789	**0.6498**	**0.4834**
	Abalone	0	0	0.2294	0.1451	0.1320	0.3365	**0.4871**	**0.4288**
	Ecoli	0	0	**0.2821**	**0.3896**	0	0	0.1784	0.3038
	MNIST	0.9006	0.8439	0.9001	0.8202	0.9006	0.8439	**0.9129**	**0.8523**

和 0.5 时,SMOTE 的分类结果高于 Borderline-SMOTE,而当 *sop* 为 0.8 时,SMOTE 的分类结果却低于 Borderline-SMOTE,且 Borderline-SMOTE 取得最优的分类结果低于 SMOTE 的最优分类结果;在其余数据集中,不同 *sop* 值下,SMOTE 的分类结果均高于 Borderline-SMOTE,特别在 Ecoli 数据集中的表现尤为突出. 可以看出这两种方法受参数影响较大,而 QSMOTE 在不同 *sop* 值下不同数据集的结果统计中 13 次取得最佳结果,表现出了很强的稳定性,表明 QSMOTE 有效提升了少数类样本的分类准确率.

从表 3 至表 5 可以看出,在 $k=10,15,20$ 的情况下各算法的实验结果也表现出和 $k=5$ 大致相同的趋势. 为了更加直观地比较不同算法的优劣,表 6 列出了 QSMOTE 和其他方法的 *g-means* 和 *F* 值的比较结果统计.

从表 6 可以看出,QSMOTE 方法在所有实验中两个指标均为最优的次数明显多于别的方法,且在 44 次最优结果中仅有 5 次最优结果与其他方法的最优结果相同,这表明 QSMOTE 方法不仅对不同数据集具有更好的适用性,而且算法性能相较别的方法更优. 在全部方法仅一项指标为最优的统计次数相对持平情况下,QSMOTE 方法两项指标均不为最优的次数大大低于别的方法的次数,这可以说明 QSMOTE 方法在解决非平衡问题时具有相对稳定的优点,在对于改善非平衡分类问题时效果明显优于别的方法.

为了区别不同参数对算法性能的影响,本文将全部数据集的实验结果在对应的参数下求均值进行比较,表 7 和表 8 分别为不同算法在参数 *k* 和 *sop* 下的均值实验结果,表中的黑色字体表示相同参数下不同算法结果的最大值,斜体表示同一个算法在不同参数下结果的最大值.

表 3 不同支持度下各算法的实验结果($k=10$)

Table 3 The experimental results of different algorithms with different *sop* ($k=10$)

sop	Dataset	SVM		SVM+SMOTE		SVM+Borderline-SMOTE		SVM+QSMOTE	
		g-means	*F*	*g-means*	*F*	*g-means*	*F*	*g-means*	*F*
0.3	Glass	0.7638	**0.9091**	0.7638	0.8333	0.7638	0.8333	**0.7817**	**0.9091**
	Page_Block	0.3991	0.2553	0.4206	0.2797	0.3986	0.2517	**0.4409**	**0.3014**
	Yeast	0	0	0.5497	0.3692	0.5501	0.3918	**0.6107**	**0.4270**
	Abalone	0	0	**0.4132**	**0.2490**	0.1420	0.4310	0.3880	0.2175
	Ecoli	0	0	**0.2821**	**0.3896**	0	0	0.1708	0.2785
	MNIST	0.9006	0.8439	0.9117	0.8427	0.9006	0.8439	**0.9238**	**0.8603**
0.5	Glass	0.7638	**0.9091**	0.7638	**0.9091**	0.7638	**0.9091**	**0.7993**	**0.9091**
	Page_Block	0.3991	0.2553	0.4098	0.2657	0.3988	0.2553	**0.4206**	**0.2797**
	Yeast	0	0	0.5538	0.3692	**0.5900**	**0.4255**	0.5872	0.4086
	Abalone	0	0	0.3059	0.1500	0.1205	0.2791	**0.4392**	**0.4291**
	Ecoli	0	0	**0.2821**	**0.3896**	0	0	0.1708	0.2785
	MNIST	0.9006	0.8439	0.9001	0.8295	0.9006	0.8439	**0.9184**	**0.8539**
0.8	Glass	0.7638	**0.9091**	0.7638	**0.9091**	0.7817	**0.9091**	**0.7993**	**0.9091**
	Page_Block	0.3991	0.2553	**0.4310**	**0.2917**	0.3991	0.2553	0.4206	0.2797
	Yeast	0	0	0.5538	0.3750	0.5578	0.4118	**0.6272**	**0.4500**
	Abalone	0	0	0.3153	0.1576	0.4318	0.3352	**0.4378**	**0.3396**
	Ecoli	0	0	**0.2821**	**0.3896**	0	0	0.1708	0.2785
	MNIST	0.9006	0.8439	**0.9238**	**0.8603**	0.9006	0.8439	0.9006	0.8439

表 4　不同支持度下各算法的实验结果($k=15$)

Table 4　The experimental results of different algorithms with different *sop* ($k=15$)

sop	Dataset	SVM		SVM+SMOTE		SVM+Borderline-SMOTE		SVM+QSMOTE	
		g-means	*F*	*g-means*	*F*	*g-means*	*F*	*g-means*	*F*
0.3	Glass	0.7112	0.8000	**0.7951**	**0.9091**	**0.7951**	**0.9091**	**0.7951**	**0.9091**
	Page_Block	0.3991	0.2553	0.4310	0.2917	0.4096	0.2639	**0.4412**	**0.3034**
	Yeast	0	0	0.5302	0.3492	0.1963	0.3562	**0.6443**	**0.4634**
	Abalone	0	0	0.4222	0.2581	0.1420	0.2767	**0.4368**	**0.4325**
	Ecoli	0	0	**0.2821**	**0.3896**	0.1995	0.3797	0.1708	0.2785
	MNIST	0.9006	0.8439	**0.9123**	0.8427	0.9006	0.8439	0.9056	**0.8515**
0.5	Glass	0.7112	0.8000	**0.7951**	**0.9091**	0.7112	0.8000	0.6948	0.8000
	Page_Block	0.3991	0.2553	**0.4310**	**0.2917**	0.3988	0.2553	**0.4310**	**0.2917**
	Yeast	0	0	0.5938	0.4235	0.1414	0.3649	**0.6110**	**0.4242**
	Abalone	0	0	0.4352	**0.2742**	0.1205	0.2253	**0.4392**	0.2387
	Ecoli	0	0	0.2821	0.3896	0.1995	0.3797	**0.2913**	**0.4156**
	MNIST	0.9006	0.8439	0.9129	**0.8674**	0.9006	0.8439	**0.9178**	0.8429
0.8	Glass	0.7112	0.8000	0.7951	0.9091	0.7581	0.9091	**0.8710**	**1.0000**
	Page_Block	0.3991	0.2553	0.4206	0.2797	0.4098	0.2657	**0.4308**	**0.2897**
	Yeast	0	0	0.6146	0.4478	0.1414	0.3649	**0.6278**	**0.4491**
	Abalone	0	0	0.4204	0.2570	**0.4871**	0.2177	0.4387	**0.2811**
	Ecoli	0	0	0.2821	0.3896	0.1995	0.3797	**0.2913**	**0.4156**
	MNIST	0.9006	0.8439	**0.9238**	**0.8603**	0.9006	0.8439	0.9129	0.8571

表 5　不同支持度下各算法的实验结果($k=20$)

Table 5　The experimental results of different algorithms with different *sop* ($k=20$)

sop	Dataset	SVM		SVM+SMOTE		SVM+Borderline-SMOTE		SVM+QSMOTE	
		g-means	*F*	*g-means*	*F*	*g-means*	*F*	*g-means*	*F*
0.3	Glass	0.8710	1.0000	0.8710	0.1000	0.8710	1.0000	**0.8710**	**1.0000**
	Page_Block	0.3991	0.2553	**0.4310**	**0.2917**	0.4102	0.2695	0.4206	**0.2797**
	Yeast	0	0	0.6254	0.4507	0.5292	0.3711	**0.6498**	**0.4706**
	Abalone	0	0	0.4570	0.3062	0.1420	0.2397	**0.4638**	**0.3767**
	Ecoli	0	0	**0.2443**	**0.3864**	0.1995	0.3797	0.1784	0.3038
	MNIST	0.9006	0.8439	0.9117	0.8380	0.9006	0.8439	**0.9250**	**0.8701**
0.5	Glass	0.8710	1.0000	0.8710	0.1000	0.8710	1.0000	**0.8906**	**1.0000**
	Page_Block	0.3991	0.2553	0.4310	**0.2917**	0.4204	0.2778	**0.4333**	0.2596
	Yeast	0	0	**0.6301**	**0.4571**	0.5780	0.4314	0.6218	0.4390
	Abalone	0	0	**0.4659**	0.3202	0.1205	0.2076	0.4638	**0.3767**
	Ecoli	0	0	**0.1995**	**0.3797**	**0.1995**	**0.3797**	0.1708	0.2785
	MNIST	0.9006	0.8439	0.9056	0.8315	0.9006	0.8439	**0.9184**	**0.8588**
0.8	Glass	0.8710	1.0000	0.8710	0.1000	0.8710	1.0000	**0.8906**	**1.0000**
	Page_Block	0.3991	0.2553	**0.4310**	**0.2917**	0.3991	0.2553	**0.4310**	**0.2917**
	Yeast	0	0	**0.6495**	0.4789	0.5371	**0.7416**	0.6272	0.4500
	Abalone	0	0	0.4570	**0.3062**	**0.4871**	0.2424	0.4387	0.2811
	Ecoli	0	0	0.1961	0.0741	**0.1995**	**0.3797**	0.1708	0.2785
	MNIST	0.9006	0.8439	0.8995	0.8202	0.9006	0.8439	**0.9184**	**0.8588**

表 6 QSMOTE 方法与其他方法实验结果比较

Table 6 The experimental results of QSMOTE and other algorithms

k	＃指标均最优		＃一项指标最优		＃指标均不为最优	
	QSMOTE	其他	QSMOTE	其他	QSMOTE	其他
5	13	7	0	0	5	11
10	11	7	0	3	7	8
15	11	5	4	4	3	9
20	9	6	3	4	6	8
合计	44	25	7	11	21	36

表 7 不同近邻参数下算法均值实验结果

Table 7 The experimental results of QSMOTE and other SMOTE algorithms with different k

k	SMOTE		Borderline-SMOTE		QSMOTE	
	g-means	*F*	*g-means*	*F*	*g-means*	*F*
5	0.5288	0.4678	0.4831	0.4652	***0.5847***	***0.5414***
10	0.5459	0.4922	0.4778	0.4567	**0.5560**	**0.5141**
15	0.5711	*0.5189*	0.4451	0.4933	**0.5751**	**0.5302**
20	**0.5860**	0.3791	*0.5298*	***0.5393***	0.5824	0.5374

表 8 不同支持度下算法均值实验结果

Table 8 The experimental results of QSMOTE and other SMOTE algorithms with different *sop*

sop	SMOTE		Borderline-SMOTE		QSMOTE	
	g-means	*F*	*g-means*	*F*	*g-means*	*F*
0.3	*0.5602*	0.4658	0.4732	0.4860	**0.5715**	**0.5327**
0.5	0.5594	*0.4720*	0.4668	**0.4784**	**0.5739**	**0.5269**
0.8	0.5543	0.4557	*0.5119*	*0.5017*	***0.5784***	***0.5328***

从表 7 和表 8 中可以看出 SMOTE, Borderline-SMOTE和 QSMOTE 三种方法均受参数 k 的影响,且参数 k 对于 SMOTE 方法的性能影响最大,Borderline-SMOTE 次之,QSMOTE 受参数 k 的影响最小,且不同参数下的实验结果相对稳定. 表 8 中 $k=20$ 的结果分析表明 k 取值在一定范围内,SMOTE 方法实验结果与 k 值成正比,但随着 k 不断增大,新样本偏离到其他样本分布区域,降低了分类效果. *sop* 通过比较样本近邻中同类样本的比例来确定参照合成样本,当 k 较小时,*sop* 影响可以忽略不计;随着 k 在合理范围内的增大,*sop* 越大表明参照合成样本不是噪声点和容易产生 overlapping 的可能性越低,因此实验结果有提升;但当 *sop* 和 k 增长同时超过合理范围时,一方面导致合成样本的分布密度增加造成过拟合,另一方面更多的异类点被考虑容易造成过泛化,导致结果下降. Borderline-SMOTE 和 QSMOTE 均受到 *sop* 影响,但结果表明 *sop* 对不同方法实验结果的影响明显低于 k 的影响.

从表 7 和表 8 的实验结果可以看出,QSMOTE取得了最好的实验结果,说明相较于其他方法需要参照更多近邻样本来合成新样

本,QSMOTE 仅需少量的近邻样本就可以生成有效提升分类效果的新样本,同时还可以减小算法运行时间,证明 QSMOTE 相较别的方法对少数类样本采样具有明显优势.

3.3.2　算法性能比较　对于不同的数据集,分别选取最佳近邻参数 k 和最佳 sop 参数,然后对各算法的分类结果进行比较.

图 2 给出了不同算法的 $g\text{-}means$ 和 F 值结果. 在 Glass,Page_Block和 MNIST 数据集中,包括直接用 SVM 分类在内的四种算法均取得不同的$g\text{-}means$和 F 值,且 QSMOTE 的结果最好. 在其他数据集中,直接使用 SVM 分类实验中$g\text{-}means$和 F 值均为 0,表明对少数类分类准确率和召回率均为 0;采用上采样技术的其他三种方法 $g\text{-}means$ 和 F 值均大于前者,且在 Abalone 和 Ecoli 中,QSMOTE 的实验结果高于其他方法,在 Yeast 中 Borderline-SMOTE 取得了最好的分类结果.

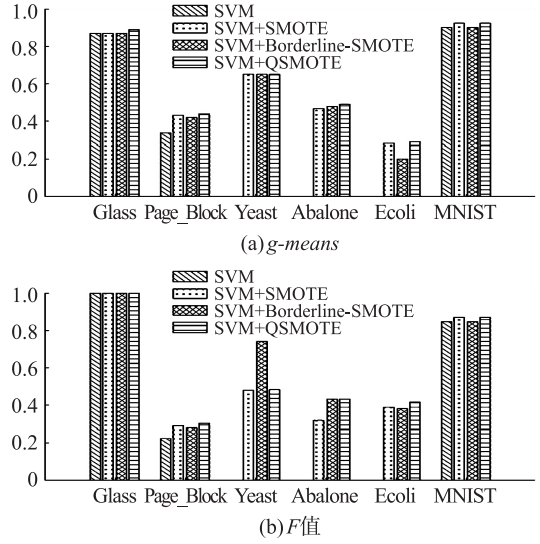

图 2　最优参数下各算法的分类指标结果

Fig. 2　The experimental results of different algorithms with optimal parameters

在保证整体分类准确率不下降的前提下,QSMOTE 算法更关注少数类的分类结果. 对各自数据集取最优参数,计算各数据集少数类样本分类准确率的算术均值进行比较,结果如图 3 所示.

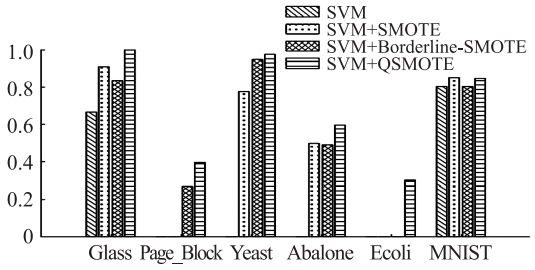

图 3　最优参数下各算法在全部少数类上的分类结果

Fig. 3　The experimental results of different algorithms under the minority with optimal parameters

从图 3 可以看出,QSMOTE 算法在全部数据集中均取得最好的分类准确率,在 Ecoli 数据集上的表现尤其突出. Ecoli 数据集上只有 QSMOTE 准确识别了少数类样本,提升了少数类样本的分类准确率,而别的方法少数类的分类准确率均为 0,对少数类样本的分类效果很差. 在 Glass 和 MNIST 数据集中,全部方法均取得了不同的分类准确率,且 QSMOTE 的效果最好. 在 Yeast 和 Abalone 数据集中,除了直接使用 SVM 分类的少数类分类准确率为 0 外,其他方法均取得不同的少数类分类准确率,其中 QSMOTE 的结果最佳.

对各自数据集选取最优参数,比较不同算法对于每个数据集样本数量最小的少数类的召回率和准确率结果,如图 4 所示. 可以看出,除了 Glass 和 MNIST 数据集外,直接使用 SVM 分类的实验中,最小类样本的准确率和召回率均为 0,别的数据集中只有 QSMOTE 能够全部得到最小样本的少数类召回率和准确率结果,表明只有 QSMOTE 能够在全部的数据集中识别到全部种类的少数类样本,在 Ecoli 数据集中的表现更为突出. 在 Page_Block 中只有 QSMOTE和 Borderline-SMOTE 可以得到最小类样本的召回率和分类准确率,且 QSMOTE 在比 Borderline-SMOTE 的召回率更大时,还拥有相同的准确率,表明 QSMOTE 对于最小类样本的分类效果更好,而 SMOTE 对该数据集中最小类样本的分类效果很差. 其他数据集中,不同的上采样方法均得到不同准确率和召回率,且 QSMOTE 都能在得到较高

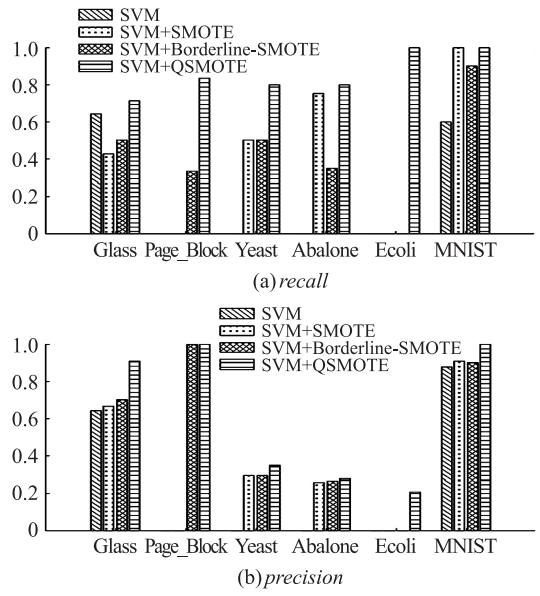

图 4 最优参数下各算法在最小类的实验结果

Fig. 4 The experimental results of different algorithms under the smallest class with optimal parameters

召回率的同时得到最高的准确率,因此,对于少数类样本的识别和准确分类,QSMOTE 比其他方法表现出更强的效率.

结合图 3 和图 4 的实验结果,可以发现,除 QSMOTE 外,其余采样方法虽然可以提升少数类的整体分类准确率,但在面对非平衡度较高,特别是少数类类别多的分类问题时,都具有一定的局限性,而 QSMOTE 则可以在一定程度上减少这些局限造成的少数类样本分类准确率较低的问题,因此 QSMOTE 算法能够更好地解决像疾病诊断这样具有较大错分代价的非平衡分类问题.

4 结束语

非平衡数据的处理在机器学习领域受到越来越多的重视,然而现有方法的效果并不理想.本文提出的 QSMOTE 算法考虑了少数类样本中包含重要信息且与异类样本之间具有最小距离的样本点,并通过这些样本点来合成新的样本.这些新样本在一定程度上避免了对其他类样本可能造成的过泛化影响,并且在一定程度上丰富和扩展了少数类样本点分布区域.因此,和目前常用的其他采样方法相比,本文算法无论对所有样本的分类效果,还是对所有少数类样本及最小类样本的分类效果都有不同程度的提升.然而合成样本数目的多少对实验有一定影响,如何选择合理的样本数目,使其既不会因为样本过多造成过泛化以及计算代价提升,也不会因为合成样本数量过少而使分类结果没有提升,是本文未来进一步研究的工作.

参考文献

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

[2] Abdi L, Hashemi S. To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(1): 238－251.

[3] Lim P, Goh C K, Tan K C. Evolutionary cluster-based synthetic oversampling ensemble (ECO-Ensemble) for imbalance learning. IEEE Transactions on Cybernetics, 2017, 47(9): 2850－2861.

[4] Lin M L, Tang K, Yao X. Dynamic sampling approach to training neural networks for multiclass imbalance classification. IEEE Transactions on Neural Networks & Learning Systems, 2013, 24(4): 647－660.

[5] Wang S, Yao X. Multiclass imbalance problems: Analysis and potential solutions. IEEE Transactions on Systems, Man, & Cybernetics, Part B (Cybernetics), 2012, 42(4): 1119－1130.

[6] Maciejewski T, Stefanowski J. Local neighbourhood extension of SMOTE for mining imbalanced data // 2011 IEEE Symposium on Computational Intelligence and Data Mining. Paris, France: IEEE, 2011: 104－111.

[7] Wang B X, Japkowicz N. Imbalanced data set learning with synthetic examples // Proceedings of IRIS Machine Learning Workshop. Piscataway, NJ, USA: IEEE, 2004, 153－162.

[8] Cateni S, Colla V, Vannucci M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing, 2014, 135: 32－41.

[9] Liu Z, Tang D Y, Li J C, *et al*. Objective cost-sensitive-boosting-WELM for handling multi class imbalance problem // 2017 International Joint Conference on Neural Networks. Anchorage, AK, USA: IEEE, 2017: 1975－1982.

[10] Dong Y J, Wang X H. A new over-sampling approach: Random-SMOTE for learning from imbalanced data sets // International Conference on Knowledge Science, Engineering and Management (KSEM 2011). Springer Berlin Heidelberg, 2011: 343－352.

[11] Zhao Y G, Li M M, Chung R, *et al*. Multi-class kernel margin maximization for kernel learning. Neurocomputing, 2016, 207: 843－847.

[12] Napierala K, Stefanowski J. Identification of different types of minority class examples in imbalanced data // International Conference on Hybrid Artificial Intelligent Systems (HAIS 2012). Springer Berlin Heidelberg, 2012: 139－150.

[13] Sáez J A, Krawczyk B, Woźniak M. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognition, 2016, 57: 164－178.

[14] Chawla N V, Bowyer K W, Hall L O, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321－357.

[15] Jian C X, Gao J, Ao Y H. A new sampling method for classifying imbalanced data based on support vector machine ensemble. Neurocomputing, 2016, 193: 115－122.

[16] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning // Advances in Intelligent Computing (ICIC 2005). Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2005: 878－887.

[17] He H B, Bai Y, Garcia E A, *et al*. ADASYN: Adaptive synthetic sampling approach for imbalanced learning // 2008 IEEE International Joint Conference on Neural Networks. Hong Kong, China: IEEE, 2008: 1322－1328.

[18] Blagus R, Lusa L. Evaluation of SMOTE for high-dimensional class-imbalanced microarray data // 2012 11th International Conference on Machine Learning and Applications. Boca Raton, FL, USA: IEEE, 2013: 89－94.

[19] Wu Y C, Lee Y S, Yang J C. Robust and efficient multiclass SVM models for phrase pattern recognition. Pattern Recognition, 2008, 41(9): 2874－2889.

[20] Wu D H, Wang Z L, Chen Y, *et al*. Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset. Neurocomputing, 2016, 190: 35－49.

（责任编辑　杨可盛）