基于邻域混合抽样和动态集成的不平衡数据分类方法

高 锋 黄海燕
(华东理工大学信息科学与工程学院 上海 200237)

摘 要 不平衡数据严重影响了传统分类算法的性能,导致少数类的识别率降低。提出一种基于邻域特征的混合抽样技术,该技术根据样本邻域中的类别分布特征来确定采样权重,进而采用混合抽样的方法来获得平衡的数据集;然后采用一种基于局部置信度的动态集成方法,通过分类学习生成基分类器,对于每个检验的样本,根据局部分类精度动态地选择最优的基分类器进行组合。通过UCI标准数据集上的实验表明,该方法能够同时提高不平衡数据中少数类和多数类的分类精度。

关键词 数据挖掘,不平衡数据,K-近邻,混合抽样,集成学习
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.038

Imbalanced Data Classification Method Based on Neighborhood Hybrid Sampling and Dynamic Ensemble

GAO Feng HUANG Hai-yan
(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract The class imbalance problems severely affect the performance of the traditional classification algorithm, leading to decrease the recognition rate of the minority. In order to solve this problem, a hybrid sampling technology based on neighborhood characteristic was proposed to enhance the classification accuracy of minority class. This technology changes the sampling weight according to the class distribution in the samples neighborhood, and uses the hybrid sampling to obtain the balanced data subset. Then the base classifiers are generated, for each test sample, a dynamic ensemble method based on local confidence is proposed to select the optimal base classifier sets. The experiments on UCI datasets show that the method has high classification accuracy rate of both minority and majority class for imbalance datasets.

Keywords Data mining, Imbalanced data, K-nearest neighbor, Hybrid sampling, Ensemble learning

1 引言

不平衡数据的分类问题是数据挖掘和机器学习领域中的一个重要研究方向。当数据中某一类数目远多于其他类别的数目时,多数的传统分类算法倾向于将样本划分为多数类,导致少数类的识别率不高。这种现象广泛分布在各个领域中,如医疗诊断[1-2]、人脸年龄估计[3]以及图像标注[4]等。但是事实上我们更加关注的是少数类分类情况,因为其错分的代价更大。因此,如何提高少数类的分类精度,设计适合不平衡数据的分类方法已经成为研究的难点。

现有关于不平衡数据分类方法的研究主要可以分为数据层面和算法层面这两大类。数据层面是通过增加或者删除样本来降低数据的不平衡程度,常用的方法有过抽样和欠抽样技术,如Random-SMOTE[5],SMOTE-RSB[6]等;算法层面则是改进已有的算法或者提出新的算法,Chawla N V等人[7]提出了基于代价敏感的分类方法,该方法强调错分样本的代价,使得分类器重视少数类。这两类算法都能提高少数类的分类精度,各有优势。

目前,将数据采样技术与集成学习算法相结合也是处理不平衡数据分类问题的有效方法之一,受到了学者们的广泛关注。这类方法主要是通过数据采样技术来调整数据类别间的不平衡度,然后采用集成学习算法来分类。Wang等人[8]提出了SmBag(SMOTEBagging)算法,该算法将SMOTE技术与Bagging相结合,获得的效果好于随机过抽样的Bagging算法。Jerzy等人[9]提出了一种新颖的数据不平衡处理方法NBBag(Neighbourhood Balanced Bagging),该算法在bootstrap的过程中增加了少数类的采样权重来提高其分类精度。Chawla等人[10]提出了SMOTEBoost算法,在该算法的每次迭代的过程中都会引入新的合成样本,使得分类器逐步调整分类面,提高了分类性能。李雄飞等人[11]提出了PCBoost算

到稿日期:2016-11-10 返修日期:2017-01-26
高 锋(1990—),男,硕士生,主要研究方向为机器学习、智能计算,E-mail:gf00224@163.com;黄海燕(1972—),女,博士,副教授,主要研究方向为控制与优化、复杂工业过程建模。

法，该算法利用数据合成法对少数类进行过采样，并及时删除错分的合成样本。李克文等人[12]提出了一种组合数据采样技术和 Boosting 技术的 RSBoost 算法。这些算法的难点在于如何将这两类方法更好地结合在一起，比如：现有的数据采样技术通过随机的方式或者基于某些启发式的规则来增加或者删除样本，并没有考虑数据邻域的分布特征；此外还有通过选择合适的集成学习方法来提高分类精度。

针对上述算法存在的难点，本文提出了一种基于邻域混合抽样和动态集成的不平衡数据分类方法——DE-NHS(Dynamic Ensemble based on Neighborhood Hybrid Sampling)，该方法首先使用基于邻域特征的混合抽样技术来平衡数据集，主要根据样本邻域中的分布特征来确定其采样权重，进而对数据集进行混合抽样处理，生成多个平衡的训练数据子集；其次使用弱学习算法对采样后的训练数据集进行分类学习以生成基分类器；然后提出一种基于局部置信度的动态集成方法，对于每个检验样本，根据局部置信度动态地选择最优的分类器进行组合；最后通过实验表明该方法能够提高少数类和多数类的分类精度。

2 基于邻域混合抽样和动态集成的不平衡数据分类方法

DE-NHS 算法主要分为 3 个阶段：1)根据数据中邻域的类别分布特征来计算采样权重；2)根据相应的采样权重对少数类和多数类分别进行过抽样和欠抽样，组成多个平衡的训练子集，使用弱学习算法训练得到基分类器；3)对于检验样本，在验证集中确定每个基分类器的局部置信度，然后动态地选择最优的分类器子集进行投票分类。

2.1 基于邻域特征的混合抽样技术

类别数量上的不平衡并不是导致分类器性能下降的唯一因素，数据邻域中的类别分布特征同样是影响分类精度的一个重要原因[13-14]，如少数类内的若干个子区域上的样本分布不平衡[15]。但是现有的方法往往只关注少数类，忽略了多数类的类别分布特征，在提高少数类分类精度的同时对多数类的分类产生了负面影响。本文考虑到不平衡数据邻域中的类别分布特征，提出了一种基于邻域特征的混合抽样技术。该方法增加了数据中“危险样本”的采样权重，然后对少数类和多数类分别进行过抽样和欠抽样，组成平衡的训练数据集。这种做法重视数据邻域中的类别分布特征，通过改变两类样本的“危险程度”来平衡样本分布。

首先使用 KNN(K-Nearest Neighbor)算法来计算每个样本的“危险程度”。令 X_k 为 x 最近 k 邻域中的样本，$x_j(x_j \in X_k)$为其中与 x 类别不相同的样本，其“危险程度”定义如下：

$$p_x=\frac{(\sum_{x_i\in X_k} W_i)^a}{\sum_{i=1}^{k} W_i} \tag{1}$$

其中，$W_i=\frac{1}{d_i}$，a 为尺度参数，d_i 是样本 x 与其邻域中样本 x_j 之间的距离。当 $p_x>0$ 时该样本 x 为“危险样本”，且 x_j 与 x 之间的距离越近、x_j 的个数越多时，其“危险程度”也就越大。然后根据每个样本的“危险程度”分别计算少数类 $X_{\min}$ 和多数类 $X_{\max}$ 的采样权重：

$$r_{\min}=\frac{p_x+1}{\sum_{x\in X_{\min}}(p_x+1)} \tag{2}$$

$$r_{\max}=\frac{p_x+1}{\sum_{x\in X_{\max}}(p_x+1)} \tag{3}$$

其中，p_x+1 可以保证“安全样本”也能以一定的概率被抽样，从而保持了数据的整体性。同时为了提高模型的效率，设置两类样本的抽样数目都为少数类的样本数。由于在少数类中大部分为“危险样本”[16]，采样权重大，因此对于少数类可以将其视为过抽样。而对于多数类，由于抽样样本数目小于样本数，可以将其视为欠抽样。

此外，在 KNN 算法中，计算样本之间的距离时需要同时处理连续属性和离散属性，本文采用 Wilson 等人[17]提出的 HVDM(Heterogeneous Value Difference Metric)距离函数。该函数能反映不同属性对分类结果的影响且能更有效地度量数据之间的差异性。

2.2 基于局部置信度的动态集成方法

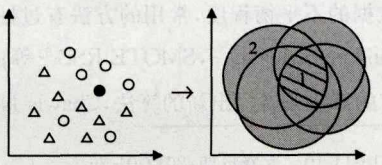
使用基于邻域的混合采样技术生成多个训练子集，然后用弱学习算法进行分类学习，生成基分类器。但由于“危险样本”的权重大而被多次采样，容易导致基分类器间的差异性变小，从而造成模型的过拟合，影响了集成学习的性能，

此外，如何组合基分类器也是集成学习的一个难点[18]。针对这个问题，本文提出了基于局部置信度的动态集成方法，与现有的集成学习算法将所有的基分类进行组合的方式不同，该方法根据每个基分类器在验证集中的局部置信度，动态地选择最优的基分类进行组合，从而提高了分类器的分类精度。其中局部置信度的定义如下。

给定一个检验样本 x_{test}，在验证集中确定其 k 个最近邻点 $xval_j$，$1\leqslant j\leqslant k$，对于基分类器 C，假设能够正确分类其中的 m 个点，则局部置信度 LC 的计算公式为：

$$LC=\frac{m}{k} \tag{4}$$

局部置信度 LC 可以衡量基分类器对于验证样本的分类性能。如果基分类器对其在验证集中邻域点上的分类正确率越高，那么该分类器能够正确分类该样本的概率就越大。因此根据每个基分类器的局部置信度的不同，采用不同的组合方式，即如果存在基分类器，其局部置信度 $LC=1$，则将这些分类器进行组合，否则将局部置信度 $LC>0$ 的基分类器进行组合，其原理如图 1 所示。



(a)原始数据的分布 (b)分类空间

图 1 基于局部置信度的动态集成方法

图 1(a)所示为原始数据的分布，黑色的圆点为验证样本；图 1(b)所示为分类空间，其中区域 1 为分类空间的交集，表示局部置信度 $LC=1$ 的所有基分类器的交集，区域 2 为分类空间的并集，表示局部置信度 $LC>0$ 的基分类器的并集。该组合方法对于每个检验的样本能够选择最优的基分类器集，提高了分类器的精度。

2.3 算法的具体流程

输入：定义训练集 $RS=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中 n 为训练样本的总数，VS 为验证集，x_{test} 为检验样本，$N_{\min}$ 为过抽样样本数量，N_{maj} 为欠抽样样本数量，C 为基分类器，T 为迭代次数，$w(x)$ 为采样权重，设 k_1 为混合采样过程中在训练样本上的最近邻域的样本数，k_2 为基分类器动态组合过程中在验证集上的最近邻域的样本数，$M_{\min}$ 和 M_{maj} 分别为数据集中少数类和多数类的样本数。

Step1　设置过抽样和欠抽样样本数量：

$$N_{\min}=N_{maj}=M_{\min} \tag{5}$$

Step2　for x in RS：

(1)使用 KNN 算法获得 x 最近邻域中的 k_1 个样本；

(2)按照式(1)计算 x“危险程度”的概率 p_{x_i}；

(3)如果 $x_i\in X_{\min}$，按照式(2)计算少数类的采样权重，否则执行步骤(4)；

$$w(x_i)=r_{\min} \tag{6}$$

(4)若 $x_i\in X_{maj}$，按照式(3)计算多数类的采样权重。

$$w(x_i)=r_{\max} \tag{7}$$

Step3　for $i=1$ to T：

(1)根据采样权重 $w(x)$ 对 RS 进行混合抽样并组成平衡的训练子集；

(2)使用弱学习算法训练得到基分类器 C。

Step4　对检验样本 x_{test} 执行如下操作：

(1)如果所有的基分类 C 都指向同一类，则 x_{test} 属于该类，否则执行步骤(2)；

(2)在验证集 VS 上，使用 KNN 算法获得 x_{test} 最近邻域中的 k_2 个样本 $xval_j$，按照式(4)计算基分类器 C 的局部置信度 LC；

(3)如果存在 $LC=1$，则用相应的基分类器对 x_{test} 进行投票分类，否则执行步骤(4)；

(4)用 $LC>0$ 的基分类器对 x_{test} 进行投票分类，否则执行步骤(5)；

(5)使用所有的基分类器进行投票分类。

输出：检验样本 x_{test} 的标签。

该算法通过基于邻域的混合抽样技术来调整数据集的不平衡度，相比于其他抽样技术，该抽样技术更加关注“危险样本”，通过采用混合抽样的方式缩小了训练子集的规模，提高了模型的分类效率。同时，使用基于局部置信度的动态集成方式提高了算法的分类精度。

3 实验与分析

为了便于讨论，本文主要关注二类非平衡数据分类问题，通常情况以多数类为负类，以少数类为正类，相应的类别标签取值分别为{−1，+1}。

3.1 数据集

为了验证算法的有效性，选取 8 组 UCI 数据集，每组数据集的样本总数、特征数、不平衡度如表 1 所列。对于含有多个类别的数据集，将其中一类作为正类，其余统一作为负类。

表 1　UCI 数据集

数据集	样本总数	特征数	正类	不平衡度
Pima	768	8	1	1.87
Breat-w	699	9	Malignant	1.90
Credit-g	1000	20	bad	2.33
Haberman	306	3	2	2.78
Vehicle	846	18	Van	3.25
Cmc	1473	9	2	3.42
Abalone	4177	8	6	12.13
Yeast	1484	8	ME2	28.10

3.2 不平衡数据分类的评价方法

目前都是采用混淆矩阵作为不平衡数据分类的评价指标，如表 2 所列，其中 TP 和 TN 分别表示预测正确的正类和负类的样本数目，FP 和 FN 分别表示预测错误的正类和负类的样本数目。

表 2　二分类混淆矩阵

	预测为正类	预测为负类
实际为正类	TP	FN
实际为负类	FP	TN

相关评价指标分别如下。

准确率：

$$Precision=\frac{TP}{TP+FP} \tag{8}$$

准确率表示所有预测为正类的样本中实际为正类样本所占的比值。

召回率：

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

召回率表示所有实际为正类的样本中预测为正类样本所占的比值。

$$F\text{-}value=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision} \tag{10}$$

其中，参数 β 表示 $Recall$ 与 $Precision$ 的相对重要性，通常取 $\beta=1$。

上面 3 个指标都是正类的评价标准，对于不平衡数据，同样需要考虑分类的整体性能。

$$G\text{-}mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \tag{11}$$

G-mean 以正类和负类的分类正确率为基础，其通常作为衡量不平衡数据集整体分类性能的评价指标。

为了全面评估算法性能，本文使用 G-mean 值，Recall 值和 F-value 值 3 个指标对分类结果进行评估。

3.3 实验结果

本文实验基于 python2.7 环境实现，使用 UCI 的部分公

共数据集，对比 SmBag[8]，NBBag[9]，SMOTEBoost[10]，PCBoost[11]，RSBoost[12]和本文提出的 DE-NHS 等方法的分类性能。其中 DE-NHS 的参数设置详细见 3.4 节，SMOTE 的过采样率为 200%，NBBag 的参数设置为 0.5，迭代次数 T 都设为 100。实验中使用的基分类器均为 C4.5 决策树，采用十折交叉验证，即将数据分为 10 份，其中 8 份用来训练，1 份用来验证，1 份用来测试，最后将 10 次测试结果的平均值作为该算法的最终结果。图 2—图 4 分别示出了 6 种方法在 8 个数据集上的 G-mean 值、Recall 值和 F-value 值的对比柱形图。

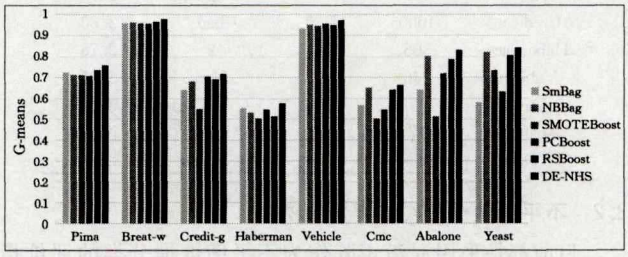


图 2 各种方法的 G-mean 值对比柱形图

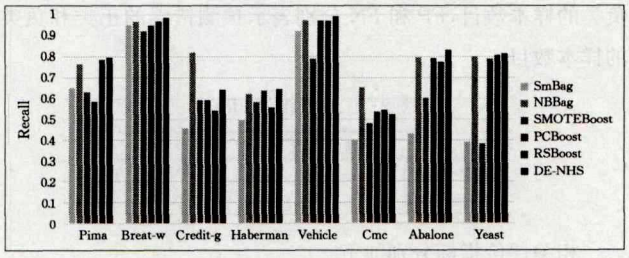


图 3 各种方法的 Recall 值对比柱形图

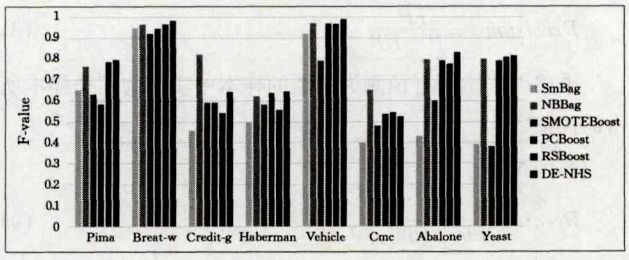


图 4 各种方法的 F-value 值对比柱形图

从图中可以看出，DE-NHS 在多数数据集上的结果均优于其他算法，并且随着不平衡度的增加，其优势越明显。由于 DE-NHS 算法是基于每类样本邻域中的类别分布特征进行混合抽样的，因此能够改变样本的“危险程度”。对比 Pima 和 Breat-w 这两个数据集，它们在样本的数量、特征数和不平衡度上都相差不多，但在分类的精度上却相差很大。根据数据集中两类样本邻域中的类别分布情况（表 3 列出了两类样本中“危险样本”的占比情况），Pima 数据集的正类中 86.19%的样本和负类中 63.8%的样本都属于“危险样本”，而在 Breat-w 数据集中只有 24.48%的正类样本与 4.58%的负类样本属于“危险样本”，可以看出两个明显不同邻域的类分布特征导致了这两个数据集的分类精度不同。相对于其他算法，DE-NHS 在“危险样本”占比较大的 Pima 上，在 3 个度量指标上都具有一定的优势，在 Breat-w 上的表现也是良好的。

表 3 两类样本中“危险样本”的占比情况/%

数据集	负类	正类
Pima	63.8	86.19
Breat-w	4.58	24.48
Credit-g	68.42	98.33
Haberman	61.33	98.76
Vehicle	16.07	48.24
Cmc	61.31	96.99
Abalone	14.70	99.01
Yeast	8.51	98.04

同时，从实验结果也可以看出，SmBag，SMOTEBoost，PCBoost 和 RSBoost 在处理多个数据集时体现出不错的分类性能。但是这些方法在抽样的过程中只从全局的角度处理数据，忽略了样本邻域中的类分布特征，影响了分类效果。NBBag 虽然考虑到了样本邻域中的类分布特征，但由于只重视正类邻域的类分布特征，而且没有对基分类器进行选择，因此在提高正类分类精度的同时对负类产生了负面影响，如在 Pima 和 Credit-g 数据集上，NBBag 虽然取得了较高的 Recall 值，但是 G-mean 值却相对不高。

为了进一步验证 DE-NHS 算法的性能，本文使用 Wilcoxon 秩和检验[19]来分别比较 DE-NHS 与其他算法在实验结果中的差异性，具体的 p-value 如表 4 所列。

表 4 Wilcoxon 秩和检验的 p-value

方法	G-mean	Recall	F-value
SmBag	2.35e−09	3.54e−12	0.0003
NBBag	0.0466	0.1593	0.0002
SMOTEBoost	9.14e−09	3.54e−12	0.0017
PCBoost	5.32e−06	0.0001	0.0302
RSBoost	0.0402	0.0391	0.0126

从表 4 可以发现，在显著性水平设为 0.01 的情况下，DE-NHS 分别与 SmBag 和 SMOTEBoost 在 3 个指标上都显示出了明显的差异性。而在显著性水平设为 0.05 的情况下，DE-NHS 除了与 NBBag 在 Recall 这个指标上的显著性不明显之外，与其他方法在 3 个指标上都显示出了明显的差异性。主要原因是这两类方法都重视正类邻域的类分布特征，没有引入噪声，从而都具有较高的 Recall 值。

3.4 参数的选择对分类性能的影响

DE-NHS 算法中尺度参数 α、混合采样过程中在训练样本上的最近邻域的样本数 k_1 以及基分类器动态组合过程中在验证集上的最近邻域的样本数 k_2 都影响着分类的准确性，但是最佳的参数组合很难通过经验获得。为了便于分析，本文只研究参数 α 和 k_2 的变化对分类结果的影响，以 Yeast 数据集进行演示，设 $k_1=5$，通过改变 α 和 k_2 来观察其对分类性能的影响，如图 5 所示。

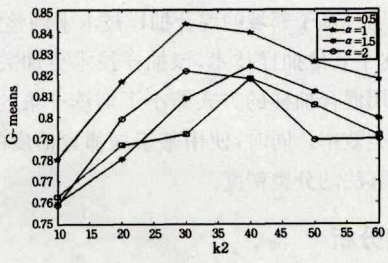


图 5 改变 α 和 k_2 对 Yeast 数据集的影响

通过图 5 的曲线可以看出，参数 $\alpha=1, k_2=30$ 时取得最好的 G-mean 曲线，α 值较小时“危险样本”的采样权重太小，分类面偏向负类；当 α 过大时，“危险样本”采样权重太大，分类面偏向正类。参数 k_2 也是影响分类精度的一个重要因素，k_2 过大或者过小都无法显示出分类器动态组合的优势。

最佳的参数组合依赖于具体的数据分布，为了提高 DE-NHS 算法的分类性能，本文通过对每个数据集中的训练集进行交叉验证来获得最佳的参数组合。参数 a 的取值范围设置为[0.5,2]，步长为 0.1，$k_1=5$，k_2 的取值范围设置为[10,60]，步长为 5，使用 G-mean 作为优化参数组合的评价指标。最后，将获得的最佳参数组合作为该数据集的最佳参数组合，分类结果（即前面对比实验的结果）和参数组合如表 5 所列。

表 5　DE-NHS 分类结果及其参数组合

数据集	G-mean	Recall	F-value	a	k_1	k_2
Pima	0.7559	0.7963	0.7361	1.1	5	10
Breat-w	0.9776	0.9849	0.9613	1.0	5	15
Credit-g	0.7161	0.6429	0.6158	2.0	5	35
Haberman	0.5779	0.6459	0.6012	1.2	5	10
Vehicle	0.9732	0.992	0.9221	1.0	5	20
Cmc	0.6623	0.5242	0.4634	1.5	5	25
Abalone	0.8304	0.8313	0.5102	0.9	5	40
Yeast	0.8417	0.8141	0.3045	1.0	5	30

结束语　本文提出了一种将数据采样技术与集成学习相结合的不平衡数据分类算法，算法将邻域混合抽样和动态集成结合起来，重视少数类和多数类邻域中的类别分布特征，通过改变采样权重，使用混合抽样技术将不平衡数据转化为多个平衡的训练子集，然后进行分类学习并生成基分类器；同时提出基于局部置信度的动态集成方法，对于每个检验的样本，动态地选择局部置信度高的基分类器子集，进一步提高了分类精度。实验结果表明，与其他不平衡分类算法相比，该算法对于少数类和整体都具有更高的分类精度，但是如何进一步优化参数是今后需要研究的内容。

参 考 文 献

[1] KRAWCZYK B, WOŹniak M. Hypertension Type Classification Using Hierarchical Ensemble of One-Class Classifiers for Imbalanced Data[M]// ICT Innovations 2014. Springer International Publishing, 2015: 341-349.

[2] CAO P, LI B, LI W, et al. Hybrid Sampling Algorithm Based on Probability Distribution Estimation[J]. Control and Decision, 2014(5): 815-520. (in Chinese)
曹鹏，李博，栗伟，等. 基于概率分布估计的混合采样算法[J]. 控制与决策，2014(5)：815-520.

[3] CHAO W L, LIU J Z, DING J J. Facial age estimation based on label-sensitive learning and age-oriented regression[J]. Pattern Recognition, 2013, 46(3): 628-641.

[4] ZHANG D, ISLAM M M, LU G. A review on automatic image annotation techniques[J]. Pattern Recognition, 2012, 45(1): 346-362.

[5] LI J, LI H, YU J L. Application of Random-SMOTE on Imbalanced Data Mining[C]// 2011 Fourth International Conference on Business Intelligence and Financial Engineering (BIFE). 2011: 130-133.

[6] RAMENTOL E, CABALLERO Y, BELLO R, et al. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory[J]. Knowledge & Information Systems, 2011, 33(2): 245-265.

[7] CHAWLA N V, CIESLAK D A, HALL L O, et al. Automatically countering imbalance and its empirical relationship to cost [J]. Data Mining and Knowledge Discovery, 2008, 17(2): 225-252.

[8] WANG S, YAO X. Diversity analysis on imbalanced data sets by using ensemble models[C]// IEEE Symposium on Computational Intelligence and Data Mining, 2009(CIDM'09). IEEE, 2009: 535-548.

[9] BŁASZCZY ŃS J, STEFANOWSKI J. Neighbourhood sampling in bagging for imbalanced data[J]. Neurocomputing, 2015, 150: 529-542.

[10] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: Improving Prediction of the Minority class in Boosting [J]. Lecture Notes in Computer Science, 2003, 2838: 107-119.

[11] LI X F, LI J, DONG Y F, et al. A New Learning Algorithm for Imbalanced Data-PCBoost[J]. Chinese Journal of Computers, 2012, 35(2): 202-209. (in Chinese).
李雄飞，李军，董元方，等. 一种新的不平衡数据学习算法 PCBoost[J]. 计算机学报，2012，35(2)：202-209.

[12] LI K W, YANG L, LIU W Y, et al. Classification Method of Imbalanced Data Based on RSBoost[J]. Computer Science, 2015, 42(9): 249-252. (in Chinese)
李克文，杨磊，刘文英，等. 基于 RSBoost 算法的不平衡数据分类方法[J]. 计算机科学，2015，42(9)：249-252.

[13] NAPIERA, KRYSTYNA A, STEFANOWSKI J, et al. Learning from imbalanced data in presence of noisy and borderline examples[C]// International Conference on Rough Sets and Current Trends in Computing. Springer-Verlag, 2010: 158-167.

[14] NAPIERALA K, STEFANOWSKI J. Identification of different types of minority class examples in imbalanced data[C]// International Conference on Hybrid Artificial Intelligent Systems. Springer-Verlag, 2012: 139-150.

[15] WEISS G M. The impact of small disjuncts on classifier learning [M]// Data Mining. Springer US, 2010: 193-226.

[16] NAPIERALA K. Improving rule classifiers for imbalanced data [D]. Poznan University of Technology, 2013.

[17] WILSON D R, MARTINEZ T R. Improved heterogeneous distance functions[J]. Journal of Artificial Intelligence Research, 2000, 6(1): 1-34.

[18] LI L, ZOU B, HU Q, et al. Dynamic classifier ensemble using classification confidence[J]. Neurocomputing, 2013, 99(99): 581-591.

[19] JAPKOWICZ N, SHAH M. Evaluating Learning Algorithms: A Classification Perspective[OL]. http://www.openisbn.com/download/0521196000.pdf.