DOI:10.11992/tis.201507015
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160314.1431.002.html

动态平衡采样的不平衡数据集成分类方法

胡小生,温菊屏,钟勇
(佛山科学技术学院 电子与信息工程学院,广东 佛山 528000)

摘 要:传统分类算法假定平衡的类分布或相同的误分类代价,处理不平衡数据集时,少数类识别精度过低。提出一种动态平衡数据采样与 Boosting 技术相结合的不平衡数据集成分类算法。在每次迭代初始,综合使用随机欠采样和 SMOTE 过采样获得平衡规模的训练数据,各类别样本数据比例保持随机性以体现训练数据的差异性,为子分类器提供更好的训练平台;子分类器形成后,利用加权投票得到最终强分类器。实验结果表明,该方法具有处理类别不平衡数据分类问题的优势。

关键词:分类;不平衡数据;重采样;集成学习;随机森林
中图分类号: TP181 **文献标志码**:A **文章编号**:1673-4785(2016)02-0257-07

中文引用格式:胡小生,温菊屏,钟勇. 动态平衡采样的不平衡数据集成分类方法[J]. 智能系统学报, 2016, 11(2): 257-263.
英文引用格式:HU Xiaosheng, WEN Juping, ZHONG Yong. Imbalanced data ensemble classification using dynamic balance sampling[J]. CAAI transactions on intelligent systems, 2016, 11(2): 257-263.

Imbalanced data ensemble classification using dynamic balance sampling

HU Xiaosheng, WEN Juping, ZHONG Yong
(College of Electronic and Information Engineering, Foshan University, Foshan 528000, China)

Abstract: Traditional classification algorithms assume balanced class distribution or equal misclassification costs, which result in poor predictive accuracy of minority classes when handling imbalanced data. A novel imbalanced data classification method that combines dynamic balance sampling with ensemble boosting classifiers is proposed. At the beginning of each iteration, each member of the dynamic balance ensemble is trained with under-sampled data from the original training set and is augmented by artificial instances obtained using SMOTE . The distribution proportion of each class sample is randomly chosen to reflect the diversity of the training data and to provide a better training platform for the ensemble sub-classifier. Once the sub-classifiers are trained, a strong classifier is obtained using a weighting vote. Experimental results show that the proposed method provides better classification performance than other approaches.

Keywords: data mining; imbalanced data; re-sampling; ensemble; random forest

分类是机器学习、数据挖掘领域的重要研究内容,通过对输入的训练样本数据进行分析、学习后获得决策模型,随后即可对未知样本进行预测。目前,已经有许多经典的分类算法,例如决策树、支持向量机、人工神经网络,这些算法在类别数据分布均匀的条件下具有良好的分类性能,得到了广泛应用。但是,在许多实际应用领域中,存在着非常明显的类别不平衡数据,例如信用卡欺诈检测、医疗疾病诊断、网络入侵检测等,在这些情况的分类处理过程中,少数类需要受到特别关注,往往具有更大的误分类代价,然而传统分类算法基于平衡的数据分布或者相

收稿日期:2015-07-09. **网络出版日期**:2016-03-14.
基金项目:广东省自然科学基金项目(2015A030313638);佛山科学技术学院校级科研项目.
通信作者:胡小生. E-mail:feihu@fosu.edu.cn.

等的误分类代价之基本假设,为保证算法总体分类准确率,通常将少数类错分至多数类,从而导致少数类识别准确率过低。因此,传统分类算法面对类不平衡数据,分类效果不佳。

当前,不平衡数据分类问题的解决方法主要有 3 个方面:1) 数据层面,移除部分多数类样本或者增加新的合成样例,改变数据分布,降低不平衡度,称之为重采样方法[1-5];2) 算法层面,分析已有算法在面对不平衡数据分类的缺陷,改进算法或者提出新算法来提升少数类的分类准确率,例如代价敏感学习[6]、集成学习[7-8]、单类学习[9]等;3) 评价标准层面,提出新的适合不平衡数据分类的分类器性能评价标准,常见的有基于混淆矩阵基础上的少数类精确度与召回率的调和均值 F_{measure}[10],几何均值 G_{mean}[11] 和 ROC 曲线等。

本文从数据层面和算法层面着手,融合数据采样和 boosting 技术,提出在动态平衡采样基础上集成提升的不平衡数据分类算法,目标旨在提高小类样本的分类精度。为了论述方便,后续部分将少数类称之为正类,多数类称之为负类。

1 采样方法

数据层面的采样技术针对不平衡数据特点,通过过采样、欠采样等方式进行数据处理,以期获得一个相对均衡的数据分布。相关研究表明,平衡的数据分布更加有利于提高传统算法的分类性能[12-13]。

1.1 过采样

最简单的过采样是随机过采样,其随机选择若干正类样本,随后简单复制该样本,添加至训练集。随机过采样仅仅复制正类样本,没有增加任何新的额外合成样例,对于提高正类识别率没有多大帮助;另外,当数据不平衡度非常高时,需要在正类上进行多倍采样才能使最终数据分布趋于平衡,结果使得训练数据规模变大,分类器学习到的决策域变小,容易导致过拟合。

针对随机过采样的不足,Chawla 等[3]提出一种 SMOTE(synthetic minority over-sampling technique)方法,该方法为每个正类样本选择若干(5 或者 7)个近邻,随后在选定样本与近邻样本之间进行线性插值,生成无重复的合成样例。SMOTE 方法能够使正类的决策边界远离负类空间,正类具有更大泛化空间;但是其缺点是没有考虑近邻样本的分布特点,合成样例具有一定的盲目性,容易产生噪声样例,以及出现类间混叠现象,影响后续分类器的分类性能。

为了解决 SMOTE 方法的不足之处,基于 SMOTE 的改进算法相继被提出。Han 等[10]提出仅为靠近类边界的正类样本生成合成样例的 Borderline-SMOTE 方法,更有利于分类器的学习,但是需要依据输入的近邻参数 k 来确定正类边界样本集合,如何合理确定参数 k 以及科学判断边界有待深入研究。He 等[14]提出 ADASYN 算法,将输入数据的密度分布作为确定合成样例数目的依据,自适应方式控制正类样本的分布。Batista 等[15]提出 SMOTE+Tomek 算法,该方法利用 SMOTE 生成合成样例;利用 Tomek 算法对连接样例进行移除,较好地克服了 SMOTE 带来的噪声问题。

1.2 欠采样

随机欠采样是随机性减少负类样本,其方法简单,操作容易,但是存在去除样本的盲目性和去除样本比例参数不确定问题,以及代表性样本的丢失而影响分类精度。

Kubat 等[16]将在负类上的单边采样与 Tomek links 相结合,利用 Tomek link 删除噪声样本,利用压缩最近邻算法删除远离边界区域的样本,将剩下的负类样本与所有正类样本一起构成训练集,用于分类器学习。

文献[17-20]提出利用聚类提取代表性样本的平衡数据分布的方法。算法首先对负类样本进行聚类操作,聚类个数与正类样本数目相同,然后提取各个聚类质心作为聚簇的代表样本,与所有正类样本一起组成平衡训练集。由于用聚类质心代表聚簇内的所有样本,不可避免地损失了数据分布的特征信息,使得抽样后的数据分布与原始数据分布出现一定的差异,从而影响算法的分类性能。

由上述分析可知,过采样和欠采样均存在一定的局限性:

1)过采样不断合成新的正类合成样例使得数据规模变大,增加了算法的学习时间;

2)过采样使得分类器训练得到的决策域变小,容易导致过拟合;

3)欠采样存在富含分类信息样本丢失问题,特别是在高度不平衡数据集中,移除过多负类样本使得信息丢失严重,造成样本代表性差,严重背离初始数据分布;

4)欠采样难以合理确定抽样比例参数。

针对过采样和欠采样方法存在的局限性,本文提出基于动态平衡采样的不平衡数据集成分类方法,在

集成迭代的每次数据采样过程中,无需给定抽样比例参数,而是基于随机生成的样本规模数值,或者对正类进行过采样,或者在负类上进行欠采样,获得类别平衡的训练集,然后参与后续的集成算法训练。

2　动态平衡采样不平衡数据分类方法

本文算法包括动态平衡采样的训练数据获取和子分类器学习 2 个步骤,主要包括 4 个阶段:1)对初始数据集的各个样本设置相同的初始权值;2)调用动态平衡采样算法,生成合成样例,组成样本规模一致的训练集,对于新生成的合成样例,需要赋予权值;3)应用 AdaBoost 算法,生成子分类器,之后根据子分类器的分类情况对初始训练集的各个样本进行权值更新,以及权值归一化;2)、3)重复迭代执行 T 次;最后将 T 个子分类器集成。

2.1　动态平衡采样

作为数据预处理的采样技术,需要预先确定数据采样参数,不合理的数据采样参数会导致生成的数据分布严重背离初始数据分布,进而影响算法的分类性能。动态平衡采样依赖随机函数产生的数值确定各类别的采样方式及采样比例,通过重复多次的动态提取初始数据集的样本,获取充分的数据分布特性信息,降低富含分类信息样本点丢失现象。整体算法如算法 1 所示。

算法 1　动态平衡采样算法

输入　初始数据集 $S = \{x_i, y_i\}_{i=1}^{m}, y_i \in Y = \{+1, -1\}$, $+1$ 表示正类样本, -1 表示负类样本;

输出　新数据集 S'。

1) 计算集合 S 中的样本数目,负类样本集合 S_N, 其数量记为 a ,正类样本集合 S_P 的样本数记为 b , $m = a + b$;

2) 利用随机函数,生成一个随机整数 k , $2 < k < m - 2$;

3)　如果 $k < a$,则从数据集 S_N 中进行随机欠采样,采样数目为 k ,将其加入集合 S' ,在集合 S_P 中应用 SMOTE 进行过采样,生成 $m - k - b$ 个新合成样例,连同 S_P 中的 b 个样本,均加入集合 S' ;

4) 如果 $k \geqslant a$,则从数据集 S_P 中进行随机欠采样,采样数目为 $m - k$,将其加入集合 S' ,在集合 S_N 中应用 SMOTE 进行过采样,生成 $k - a$ 个合成样例,连同 S_N 中的 a 个样本,都加入集合 S' ;

5)输出集合 S' 。

算法依据 2)中所产生的随机整数值大小来决定相应的采样操作,如果产生的随机数 k 小于初始数据集的负类样本数量,则在负类样本集进行欠采样,在正类样本集进行过采样,使得最终输出集合 S' 的样本数量与初始数据集数量一致,反之,则进行相反的采样。与传统的采样方法不同的是,在步骤 4 中对正类样本进行欠采样,对负类样本进行过采样,通过随机函数产生的随机数,使得输出集合 S' 在总数量一定的情况下保持对对各类别样本的中立性。

2.2　训练样例权值更新

在第 t 次迭代过程中,需要对两个集合中的样例权重进行更新,分别是动态平衡采样后的输出集合 S' 和子分类器形成之后的初始数据集 S 。

分析动态平衡采样算法过程可知,经过数据采样之后,新数据集 S' 的样例总数与初始数据集 S 一致,均为 m ,其中包括从数据集 S 抽取的部分样例,以及部分由 SMOTE 方法产生的合成样例。S' 中的样本权值按照式(1)更新:

$$D'_t(i) = \begin{cases} \dfrac{1}{m}, & x_i \notin S \\ D_t(i), & x_i \in S \end{cases} \tag{1}$$

式中: $D_t(i)$ 和 $D'_t(i)$ 分别表示第 t 次迭代时,合成样例加入前及加入后的权值。

第 t 次迭代训练结束时,AdaBoost 分类算法在数据集 S' 进行学习后得到子分类器 $h_t: x \rightarrow \{-1, +1\}$, $t = 1, 2, \cdots, T$, $h_t(x)$ 给出数据集 S 中的样例 x 的所属类别,根据子分类器的分类情况,更新样本权值,增加错分样本的权值,减少正确分类样本权值,以便下次迭代时,“错分”样本得到更多关注。

计算子分类器 $h_t(x)$ 的分类错误率 ε_t :

$$\varepsilon_t = \sum_{i=1}^{m} D_t(i) I(h_t(x_i) \neq y) \tag{2}$$

如果 $\varepsilon_t > 0.5$,终止此轮迭代。

计算子分类器投票权重 α_t :

$$\alpha_t = \frac{1}{2} \log\left\{\frac{1 - \varepsilon_t}{\varepsilon_t}\right\} \tag{3}$$

更新样例权值:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \exp(-\alpha_t h_t(x_i) y_i) \tag{4}$$

式中 Z_t 是归一化常数。

完整算法如算法 2 所示。

算法 2　动态平衡采样的不平衡分类算法

输入　初始数据集 $S = \{x_i, y_i\}_{i=1}^{m}, y_i \in Y =$

$\{+1, -1\}$，其中 +1 表示正类样本，-1 表示负类样本；

输出 $H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} \alpha_t h_t(x)$。

1）初始化数据集 S 中各个样本权重 $D_1(i) = \frac{1}{m}$；

2）for $t = 1, 2, \cdots, T$

① 调用动态平衡采样算法，获得数据集 S'；

② 利用式(1)设置 S' 中的样例权值；

③ 使用数据集 S' 及其中的样例权值，训练基于 AdaBoost 算法的子分类器 $h_t(x)$；

④ 按照式(2)计算分类器 $h_t(x)$ 的误差 ε_t，按照式(3)计算 $h_t(x)$ 的投票权重 α_t；

⑤ 按照式(4)更新数据集 S 中的样本权重；

3）输出模型：$H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} \alpha_t h_t(x)$。

3 实验结果与分析

3.1 评价度量

传统分类器采用分类精度指标衡量分类性能，其追求整体分类准确率，忽略了在不平衡数据分类过程中需要特别关注的正类分类准确率。针对不平衡数据，许多学者提出了在两类混淆矩阵基础上的 F_{measure}[10]、G_{mean}[11]等评价方法。

在混淆矩阵中，TP(true positive)、FN(false negative)、TN(true negative)、FP(false positive)分别代表分类正确的正类样本、假的负类样本、正确的负类样本以及假的正类样本的数目。基于混淆矩阵，F_{measure} 定义如下：

$$F_{\text{measure}} = \frac{(1+\beta^2) \times \text{Recall} \times \text{Precision}}{\beta^2 \times \text{Recall} + \text{Precision}}$$

式中：Recall 为查全率，Precision 为查准率，$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$，$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$，$\beta$ 用于调节 Recall 和 Precision 的相对重要性，通常取为 1。

F_{measure} 定义说明：较大值表示 Recall 和 Precision 都较大，因此，其能够较好评价正类分类性能。

G_{mean} 其定义如下：

$$G_{\text{mean}} = \sqrt{\text{TPR} \times \text{FPR}}$$

式中

$$\text{真正率 TPR} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{真负率 FPR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

G_{mean} 兼顾了正类准确率和负类准确率，比整体分类准确率更适合于不平衡数据分类评价。

本文使用 F_{measure} 准则来衡量正类的分类性能，使用 G_{mean} 准则来衡量数据集整体分类性能。

3.2 UCI 数据

为了检验本文所提方法的有效性，选择 6 组具有实际工程应用背景的 UCI 数据[21]进行测试，对于含有多个类别的数据，取其中某个类为正类，合并其余类为负类，各数据集的基本信息见表 1。

表 1 UCI 数据集信息

Table 1 Information of UCI datasets

数据集	样例数目	少类	大类	不平衡度	属性个数
car	1 728	518	1 210	2.34	6
vehicle	846	199	647	3.25	18
vowel	990	90	900	10	13
sick	3 772	231	3 541	15.33	29
letter	20 000	734	19 266	26.25	16
page-blocks	5 473	115	5 358	46.59	10

3.3 实验结果及分析

实验中对比算法如下：

1）随机森林(random forest, RF)算法，RF 算法作为一种集成算法，在处理不平衡数据时有独特的优势，能够在某种程度上减少不均衡数据带来的影响[22]，因此将其直接应用在初始不平衡数据集进行分类。

2）SMOTEBoost[23]算法，将 SMOTE 方法与 AdaBoost.M2 结合，在每次集成迭代中生成新的合成样例，使得分类器更加关注小类样本。

3）RUSBoost[24]，与 SMOTEBoost 方法相类似，采用随机欠采样从负类样本中随机移除样例，然后应用 AdaBoost 进行多次迭代。

4）文献[4]提出的集成方法 K-means+Bagging，首先在负类样本上应用 K-means 聚类，提取与正类样本数量一致的聚类质心，组成平衡训练集，参与 Bagging 集成。

上述 3 种集成方法以及本文算法均使用 C4.5 决策树算法作为基分类器算法。

为客观对比上述不平衡数据分类方法，实验数据采用 10 折交叉验证方式，重复 10 次，以平均值作为最终的分类结果。

表 2 和表 3 分别列出 5 种方法在 6 个 UCI 数据集上的正类 F_{measure} 值和数据集整体的 G_{mean} 值，最后一行列出每种方法在所有数据集上的平均结果。

表 2　5 种方法的 $F_{measure}$ 值比较

Table 2　Comparison of $F_{measure}$ between five methods

数据集	RF	SMOTEBoost	RUSBoost	K-means+Bagging	本文算法
car	0.951	0.954	0.982	0.925	0.992
vehicle	0.932	0.955	0.973	0.738	0.987
vowel	0.845	0.992	0.896	0.705	0.998
sick	0.828	0.986	0.961	0.816	0.983
letter	0.964	0.959	0.884	0.863	0.994
page-blocks	0.68	0.904	0.744	0.622	0.988
平均值	0.867	0.958	0.907	0.778	0.990

表 3　5 种方法的 G_{mean} 值比较

Table 3　Comparison of G_{mean} between five methods

数据集	RF	SMOTEBoost	RUSBoost	K-means+ Bagging	本文算法
car	0.967	0.972	0.974	0.952	0.993
vehicle	0.963	0.977	0.979	0.826	0.987
vowel	0.868	0.974	0.892	0.742	0.995
sick	0.862	0.984	0.921	0.789	0.983
letter	0.965	0.988	0.981	0.825	0.999
page-blocks	0.767	0.952	0.905	0.713	0.992
平均值	0.899	0.975	0.942	0.808	0.992

从表 2 的 $F_{measure}$ 值可以看出,本文方法除了在 sick 数据集稍微低于 SMOTEBoost 算法之外,在其他 5 个数据集上均有最佳表现,比较各种算法在 6 组 UCI 数据上的平均值,本文方法比随机森林 RF 算法有 14.2%的提升,与基于聚类欠采样的集成算法相比有 27.3%的提升,说明本文所提方法在少数类分类性能方面有巨大的提升。

比较各个算法的整体分类性能 G_{mean}, 从表 3 可以看出,本文方法也仅在 sick 数据集上稍逊于最优算法 SMOTEBoost,二者精度相差不超过 1‰;在 6 个数据集上的平均分类性能上,本文方法获得最优精度。

结合表 1~3 可以看出,随着数据不平衡度的提高,无论是随机欠采样还是基于聚类的欠采样,由于都会对原始数据集造成样本丢失,分类性能都有所下降,特别是在 letter 和 page-blocks 数据集上,差距比较明显。与之对比,本文方法在数据采样过程中也需要对某类样本进行欠采样,通过多次动态、随机性采样调和,使得抽样数据能够较好地保持对原始数据的分布;与此同时,对另外一类样本进行 SMOTE 过采样,在没有增加数据规模条件下,保持对各类样本的中立性,或者对正类过采样,或者对负类过采样。从最终分类结果来看,本文方法在不降低数据集整体 G_{mean} 值的基础上,提高了正类的 $F_{measure}$ 值,对正类和负类都具有较高的识别率。

本文算法中经过动态平衡采样后参与基分类器训练的数据集样本规模与初始数据集一致,即集合数据大小比例为 100%,为考察参与训练的不同数据规模比例对算法分类性能的影响,选取本文算法、随机森林和 SMOTEBoost 3 种算法,同时选择以 letter 数据集为例,在 20%~100%范围内每次增加 20%比例的数据,参与集成学习,迭代 10 次,相关算法的 $F_{measure}$、G_{mean} 均值如图 1 所示。

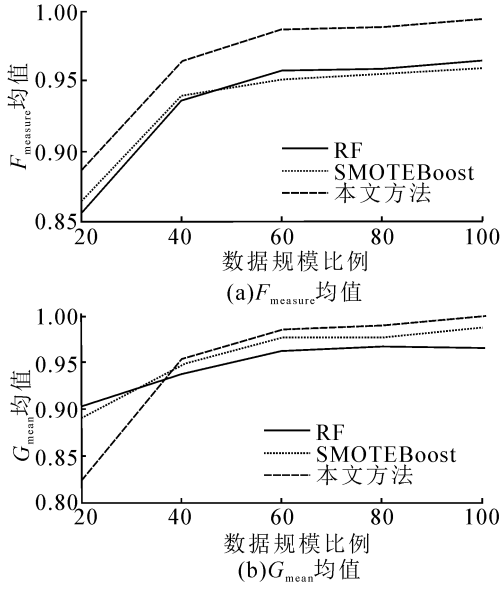


图 1　不同数据规模对分类性能影响

Fig.1　Performance measures of different ensemble size

从图 1 可看出,随着参与训练数据集比例的增大,无论是正类分类性能还是整体分类精度,都有所上升,但是随着数据比例的增大,相应的分类性能提升幅度有限。另外,在数据比例为 20%、40%时,3 种算法相对应的 $F_{measure}$ 和 G_{mean} 值几乎是线性提升,这说明过低比例的抽样数据由于损失太大的原始数据分布信息,会严重影响算法的分类性能。

4 结束语

针对类别不平衡数据分类问题,本文提出了一种混合数据采样与 Boosting 技术相结合的集成分类方法。该方法统筹运用欠采样和过采样,在保持训练集数据规模一致条件下,灵活调整各类别样本数量比例,较好地保持原始数据分布,然后采用 Boosting 技术进行多次迭代学习,获得更强性能分类器。实验结果表明,该方法能够有效提高正类样本的分类性能。

由于数据集本身的多样性和复杂性,诸如类重叠分布、噪声样本等均会影响不平衡数据性能,如果进行有针对性的数据预处理工作,将会使得动态平衡采样的数据分布更加合理,对正类的分类性能将会进一步提高。此外,将本文方法应用于多类别不平衡数据分类,也是今后需要进一步研究的方向。

参考文献:

[1] CATENI S, COLLA V, VANNUCCI M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems[J]. Neurocomputing, 2014, 135: 32-41.

[2] ZHANG Huaxiang, LI Mingfang. RWO-Sampling: a random walk over-sampling approach to imbalanced data classification[J]. Information fusion, 2014, 20: 99-116.

[3] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16(1): 321-357.

[4] 郭丽娟, 倪子伟, 江弋, 等. 集成降采样不平衡数据分类方法研究[J]. 计算机科学与探索, 2013, 7(7): 630-638.
GUO Lijuan, NI Ziwei, JIANG Yi, et al. Research on imbalanced data classification based on ensemble and undersampling[J]. Journal of frontiers of computer and technology, 2013, 7(7): 630-638.

[5] 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.
LI Xiongfei, LI Jun, DONG Yuanfang, et al. A new learning algorithm for imbalanced data-PCBoost[J]. Chinese journal of computers, 2012, 35(2): 202-209.

[6] CHEN Xiaolin, SONG Enming, MA Guangzhi. An adaptive cost-sensitive classifier[C]//Proceedings of the 2nd International Conference on Computer and Automation Engineering. Singapore: IEEE, 2010, 1: 699-701.

[7] 李倩倩, 刘胥影. 多类类别不平衡学习算法: EasyEnsemble. M[J]. 模式识别与人工智能, 2014, 27(2): 187-192.
LI Qianqian, LIU Xuying. EasyEnsemble. M for multiclass imbalance problem[J]. Pattern recognition and artificial intelligence, 2014, 27(2): 187-192.

[8] 韩敏, 朱新荣. 不平衡数据分类的混合算法[J]. 控制理论与应用, 2011, 28(10): 1485-1489.
HAN Min, ZHU Xinrong. Hybrid algorithm for classification of unbalanced datasets[J]. Control theory & applications, 2012, 28(10): 1485-1489.

[9] WANG Shijin, XI Lifeng. Condition monitoring system design with one-class and imbalanced-data classifier[C]//Proceedings of the 16th International Conference on Industrial Engineering and Engineering Management. Beijing, China: IEEE, 2009: 779-783.

[10] 叶志飞, 文益民, 吕宝粮. 不平衡分类问题研究综述[J]. 智能系统学报, 2009, 4(2): 148-156.
YE Zhifei, WEN Yimin, LV Baoliang. A survey of imbalanced pattern classification problems[J]. CAAI transactions on intelligent systems, 2009, 4(2): 148-156.

[11] 翟云, 杨炳儒, 曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学, 2010, 37(10): 27-32.
ZHAI Yun, YANG Bingyu, QU Wu. Survey of mining imbalanced datasets[J]. Computer science, 2010, 37(10): 27-32.

[12] HAN Hui, WANG Wenyuan, MAO Binghuan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//International Conference on Intelligent Computing. Berlin Heidelberg, Germany: Springer, 2005: 878-887.

[13] HE Haibo, BAI Yang, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[C]//Proceedings of IEEE International Joint Conference on Neural Networks. Hong Kong, China: IEEE, 2008: 1322-1328.

[14] BATISTA G, 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.

[15] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C]//Proceedings of the 14th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 1997: 179-186.

[16] 蒋盛益, 苗邦, 余雯. 基于一趟聚类的不平衡数据下抽

样算法[J]. 小型微型计算机系统, 2012, 33(2): 232-236.
JIANG Shengyi, MIAO Bang, YU Wen. Under-sampling method based on one-pass clustering for imbalanced data distribution [J]. Journal of Chinese computer systems, 2012, 32(2): 232-236.

[17]胡小生, 钟勇. 基于加权聚类质心的 SVM 不平衡分类方法[J]. 智能系统学报, 2013, 8(3): 261-265.
HU Xiaosheng, ZHONG Yong. Support vector machine imbalanced data classification based on weighted clustering centroid [J]. CAAI transactions on intelligent systems, 2013, 8(3): 261-265.

[18]胡小生, 张润晶, 钟勇. 两层聚类的类别不平衡数据挖掘算法[J]. 计算机科学, 2013, 40(11): 271-275.
HU Xiaosheng, ZHANG Runjing, ZHONG Yong. Two-tier clustering for mining imbalanced datasets [J]. Computer science, 2013, 40(11): 271-275.

[19]陈思, 郭躬德, 陈黎飞. 基于聚类融合的不平衡数据分类方法[J]. 模式识别与人工智能, 2010, 23(6): 772-780.
CHEN Si, GUO Gongde, CHEN Lifei. Clustering ensembles based classification method for imbalanced data sets [J]. Pattern recognition and artificial intelligence, 2010, 23(6): 772-780.

[20]UCI machine learning repository[EB/OL]. (2009-10-16) [2015-3-20]. http://archive.ics.uci.edu/ml.

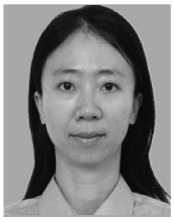
[21]李建更, 高志坤. 随机森林针对小样本数据类权重设置[J]. 计算机工程与应用, 2009, 45(26): 131-134.
LI Jiangeng, GAO Zhikun. Setting of class weights in random forest for small-sample data[J]. Computer engineering and applications, 2009, 45(26): 131-134.

[22] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTBoost: improving prediction of the minority class in boosting[C]//Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin Heidelberg: Springer, 2003, 2838: 107-119.

[23]SEIFFERT C, KHOSHGOFTAAR T M, VAN HULSE J, et al. RUSBoost: a hybrid approach to alleviating class imbalance[J]. IEEE transactions on system, man and cybernetics-part a: systems and humans, 2010, 40(1): 185-197.

作者简介:

胡小生,男,1978 年生,讲师/高级工程师,主要研究方向为机器学习、数据挖掘、人工智能。主持广东省教育厅育苗工程项目 1 项,参与省级、市厅级科研项目 6 项,发表学术论文 12 篇,其中被 EI、ISTP 检索 4 篇。



温菊屏, 女,1979 年生,讲师,主要研究方向为虚拟现实、数据挖掘。主持广东省教育厅科研项目 1 项,参与省级、厅级科研和教改项目 4 项,发表学术论文 9 篇。

钟勇, 男,1970 年生,教授,博士,主要研究方向为访问控制、隐私保护、信息检索、云计算。主持和参与国家自然科学基金、国家星火科技计划、省自然科学基金等国家级、省级科研项目 10 余项,发表学术论文 30 多篇,其中被 SCI、EI 检索 10 篇。