计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(2):185−190 [doi: 10.15888/j.cnki.csa.008297]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于遗传理论的改进数据过采样方法①

丁胜夺, 赵　刚, 阎红巧, 刘洪太

(中国石油集团 安全环保技术研究院有限公司, 北京 102206)
通信作者: 丁胜夺, E-mail: dingshengduo@cnpc.com.cn

摘　要: 针对数据分类预测模型的生成中, 高度不平衡的训练数据会大幅降低模型的性能, 本文提出了一种改进的基于遗传思想的不平衡数据集过采样方法, 该方法从生物染色体遗传理论中得到启发, 利用近亲生成相似而又不完全相同的新实例来平衡多数类, 在保证样本分布不变的前提下, 减弱甚至消除不平衡数据对训练结果的偏差影响. 最后, 通过在公共数据集上的对比实验表明, 该方法取得了更高的召回率及 *G-mean* 值, 证明此改进方法行之有效, 所生成模型的综合性能有所提高.

关键词: 过采样; 不平衡数据处理; 分类预测模型; 遗传理论

引用格式: 丁胜夺,赵刚,阎红巧,刘洪太.基于遗传理论的改进数据过采样方法.计算机系统应用,2022,31(2):185–190. http://www.c-s-a.org.cn/1003-3254/8297.html

Improved Data Oversampling Method Based on Genetic Theory

DING Sheng-Duo, ZHAO Gang, YAN Hong-Qiao, LIU Hong-Tai

(Safety and Environmental Technology Research Institute Co. Ltd., China National Petroleum Corporation, Beijing 102206, China)

Abstract: In the generation of data classification prediction models, highly unbalanced training data will significantly degrade the performance of the model. Therefore, this study proposes an improved oversampling method for unbalanced data sets based on genetic ideas. Inspired by the chromosome theory of inheritance in biology, this method uses close relatives to generate similar but not identical new instances to balance the majority of classes. Under the premise of the same sample distribution, the bias influence of unbalanced data on the training results is reduced or even eliminated. Finally, a comparative experiment on a public data set shows that the method has achieved a higher recall rate and *G-mean* value, which proves that the improved method is effective and the comprehensive performance of the generated model has been promoted.

Key words: oversampling; unbalanced data processing; classification prediction model; genetic theory

不平衡数据是指在一个数据集中, 某一类别数据的样本数量远远大于其他类别样本的数量, 其中, 样本数量较少的类别被称为少数类 (或称为正类), 类别数量较多的类被称为多数类 (或称为负类). 类不平衡是一个现实生活中普遍存在的现象, 如故障检测、医学诊断、信用卡欺诈检测、软件缺陷检测、互联网攻击识别等. 传统的分类方法如感知机[1]、决策树[2]、朴素贝叶斯模型[3]、Logistic 回归[4]、支持向量机[5] 等, 众多的方法都是为了提高分类模型的性能而设计提出, 但是不平衡数据的出现会导致训练得到的分类器对多数类的感知能力强于对少数类的感知能力, 这对诸如医学诊断、信用卡欺诈检测等应用场景是十分不利的, 因此, 如何提高不平衡数据集的分类器性能已经成为机器学习领域的热点问题, 众多解决不平衡数据问题的新方法也不断被提出.

目前, 解决非平衡数据集的方式主要有两种: 一种

① 收稿时间: 2021-04-06; 修改时间: 2021-04-29, 2021-05-11; 采用时间: 2021-05-19; csa 在线出版时间: 2022-01-17

是从算法的改进入手, 该方式从生成新的分类策略、分类器集成[6]、代价敏感[7]、特征选择[8] 等方面进行改进; 另一种是从数据集的处理入手, 这也是重要的处理不平衡问题的方式, 主要采用包括欠采样[9]、过采样[10]、训练集划分等降低数据集的不平衡程度. 其中, 欠采样方法容易造成有用信息的丢失, 过采样方法容易造成分类器的过拟合[11]. 本文由此出发, 从生物遗传理论[12]中得到启发, 利用“近亲”(同类临近数据) 生成有相似特性而又和“父类”不完全相同的少数类数据, 在平衡两类数据的同时又极大降低传统方法中过拟合的现象.

1 相关工作

解决数据不平衡问题时, 在数据处理方面, 采样的方式分为非启发式采样和启发式采样, 本文提出的合成少数类样本是典型的启发式方法. 此外, 启发式过采样方法还结合了 K 近邻准则 (K-nearest neighbor, KNN)[13]、邻域清理准则 (neighborhood cleaning rule, NCL)[14]、OSS(one-sided selection)[15] 和 IRUS (inverse random under sampling)[16] 等.

Chawla 等在 2002 年提出了少数类样本合成技术, 即 SMOTE (synthetic minority over-sampling technique) [17]. 此方法与随机过采样方法不同, 它是通过在少数类样本与其 k 个最近邻居连线上合成新样本, 合成公式如下:

$$x_{\text{new}} = x + rand(0,1) \times (\bar{x} - x) \tag{1}$$

使用传统的 SMOTE 算法会增加子集群的大小, 生成的每个数据实例都属于指定的集群, 加剧了类内的数据不平衡. 该方法中对实例点 A 和 B 进行过采样, 生成的点会落到各自的簇中, 估计的决策边界会向实例密集的群靠近, 而不是向实际边界靠近. 正如 Barua 的结论: 这些问题的出现是因为这些方法选择了所有 k 个最近的邻居, 而忽略了数据点与这些邻居间的位置关系和距离关系.

随后, Freund 等提出了 AdaBoost 算法[18], 是一种经典的集成算法, 该算法相对传统算法具有更好的泛化能力, 更高的分类精度. Chawla 等将 SMOTE 方法和 AdaBoost.M2 算法相结合, 在每次迭代中引入合成样本, 提出 SMOTEBoost 分类算法[17]. Kinoshita 等提出了联合随机降采样与 Boosting 的 RUSBoost 算法[19], 是 SMOTEBoost 的变形. 李克文等[20] 提出基于 RSBoost 的不平衡数据分类方法, 该方法采取 SMOTE 过采样和随机降采样, 再将其与 Boosting 算法相结合对数据进行分类: 李雄飞等提出一种新的不平衡数据学习算法 PCBoost[21], 每次迭代初始, 考虑属性特征, 合成新的少数类样本, 平衡训练信息.

上述研究中, 各种改进过采样的方法改善了类间的数据不平衡, 一定程度上提高了分类器的性能, 虽然较原生 AdaBoost 方法取得了较大的进步, 但仍然需要继续关注和改进, 进一步提高不平衡数据的分类精度. 本文提出的改进方法充分考虑了类间和类内的不平衡, 使少数类边界最大化, 更好地模拟数据的分布, 提高样本的总体质量.

2 基于遗传理论的不平衡数据过采样方法

2.1 染色体遗传理论

根据基础生物学和遗传学, 位于染色体上的基因是遗传的基本单位, 受精卵形成过程中, 有父母双方各一半染色体等量组合, 这就是染色体遗传理论.

引入集合 $M=(m\_1, m\_2, m\_3, \cdots, m\_q)$ 和 $F=(f\_1, f\_2, f\_3, \cdots, f\_q)$ 分别代表来自父母双方的一对染色体, A、B 为控制个体性状的等位基因, 新个体产生过程中同源染色体上的等位基因彼此分离, 非同源染色体上的非等位基因自由组合, 如图 1 所示. 生物遗传理论的发现对生物学、农业等领域都掀起了巨大的轰动, 对该理论的应用取得了重大的成果. 生物学家通过基因工程得到了更高产, 抵抗灾害能力更强的作物极大促进了人类社会生产力的发展. 从中得到启发, 我们可以通过相似性度量来分离数据样本, 合理的对有缺陷的类进行过采样. 与遗传理论不同的是: 遗传理论强调基因对个体性状的影响, 而用于采样方法中的遗传理论强调个体的多样性; 遗传理论中, 生成新的个体后父类个体逝去, 而该采样方法中, 父类数据个体和子数据个体会同时存在于集合中.

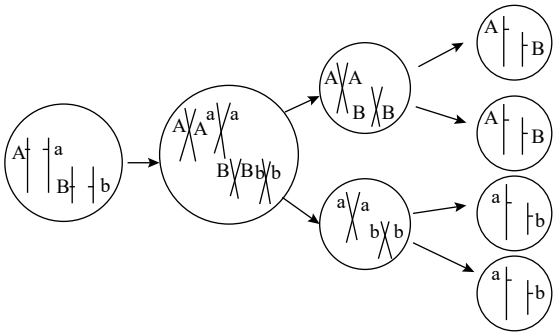

图 1 生物遗传理论染色体结合图解

2.2 基于遗传理论的过采样原理

利用染色体遗传理论, 将缺陷模块的特征指标作为染色体. 改进的过采样方法分为 3 个阶段: 首先, 分离出少数类与多数类样本, 并按照少数类样本相对本类样本的马氏距离进行降序排列; 将已排序少数类样本从中心点分割为两份数据集, 并依次为相应数据集中的每个实例分配唯一标签; 最后, 从两个分区中选择有相同唯一标签的实例求均值生成新的实例. 算法 1 列出了整个过程.

算法 1. 基于遗传理论的过采样算法

(1) 将数据集按照少数类与多数类进行划分获得集合 N_{min}、N_{maj};
(2) 计算可使数据集达到平衡的所需合成的少数类样本数 T, 并记 k 为少数类样本集样本数;
(3) 建立容纳合成数据的集合X_{new}, 初始化为空;
(4) 建立记录合成数据数量的数据集X_{newc}, 初始化为空;
(5) 计算 N_{min} 中样本的马氏距离 D^2;
(6) 创建马氏距离矩阵 $N_{mindist}$, 将数据按照递减顺序进行存储;
(7) 确定中间实例 N_{mid};
(8) 将 $N_{mindist}$ 以 N_{mid} 为界分为两个子集, 分别记为 $N_{bin1}=\{x_1, x_2, x_3, \cdots, x_{mid}\}$和 $N_{bin2}=\{x_{min+1}, x_{min+2}, x_{min+3}, \cdots, x_k\}$, 其中$x_i \in N_{midist}$;
(9) 为$x_i \in N_{bin1}$和$x_i \in N_{bin2}$中的样本按次序分配标签l_i, i=1, 2, 3,⋯, mid;
(10) **for** i=1, 2, 3,⋯, mid
(11)
  1) 选择N_{bin1}和N_{bin2}中标签相同的样本l_i, 如$x_a(l_i)==x_b(l_i)$;
  2) 通过取x_a, x_b均值生成少数类样本 x
  3) 将 x 添加到X_{new}中, 增加X_{newc}
(12) **end for**
(13) 如果$X_{newc}<T$, 将X_{new}中的实例与每个父N_{bin}中的实例配对, 并重复步骤 (10)–(12) 来创建两个不同的集合$X_{new[j]}(j$=1,2,⋯). 如果仍然$X_{newc}<T$, 将当前的每一个X_{new}集合与它的直接父集合进行配对后, 然后将其与祖先集合进行配对, 并为后续迭代重复步骤 (10)–(12);
(14) 如果$X_{newc} \geqslant T$, 将X_{new}中的所有样本赋值为少数类 (defective) 的标签, 并添加到数据集 N 中.

图解 (图 2) 过程如下: 第 1 步, 根据数据点的马氏距离找到起始双亲即 S1 和 S2, 合成新样本 C01, 为了防止传统 SMOTE 方法中的渗透现象, 设定初始父节点作为边界, 将后续生成的子节点限定在父类的范围内, 如果需要更多的实例, 在第 2 代中, 利用新生成的实例分别与父节点 (S1、S2) 继续生成新的节点. 在第 3 代中, 如果子节点和直接父节点配对生成的样本小于所需的样本数量, 则利用祖父节点与当前节点继续配对生成新的节点, 当前层级的所有配对情况均已完成后若仍不满足样本需求, 则继续按照此规则进行下一层级新样本的生成直至满足所需样本数量. 从第 1 代开始, 调用的是算法 1 的步骤 (10)–(14).

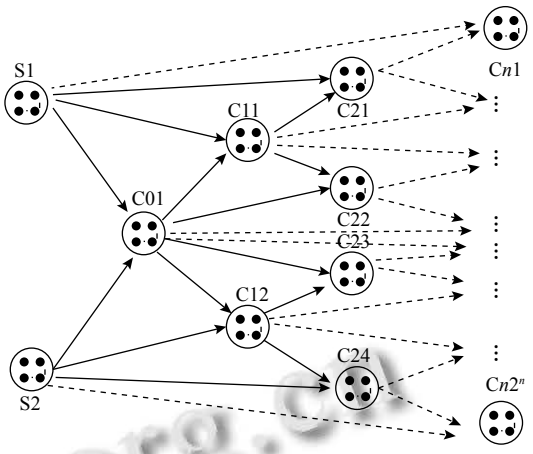

图 2 图解少数类样本生成过程

实例点 $x=(x_1, x_2, x_3, \cdots, x_n)^T$ 与实例点 $y=(y_1, y_2, y_3, \cdots, y_n)^T$ 之间的马氏距离可表示为: $d_m(\vec{x}, \vec{y}) = \sqrt{(\vec{x}-\vec{y})^T M^{-1}(\vec{x}-\vec{y})}$(其中, M 是多维随机变量的协方差矩阵, 它的幂为–1 表示求其逆矩阵), 我们使用这个度量将数据点进行降序排列, 以便于我们区分出数据点离中心点的距离. 将本文所提出的过采样方法记为 GOS (genetic over-sampling) 算法, 其执行过程如下:

GOS 算法的流程图如图 3所示. 基于 P_{fp} 的值, 算法计算要生成的合成数据实例的数量, 对已生成但不需要的数据进行剔除. 最后一层以上的合成数据点全部存储至最终数据集中. 冗余数据的删除从所生成的最后一层数据开始, 方法是将完成所需最终集的剩余数据样本量除以该层上的样本总数得到选择概率 P_c. 然后以概率值 P_c 在最后一层中选择保留样本, 这意味着上一级别的每个样本可为所需新生成数据作出相同贡献.

基于所设定的两个分区, 新生成的实例与其父实例是密切相关的, 两个分区之间的顺序限定保证了子实例的遵循父实例的趋势, 新的实例就被限制在了少数类样本的边界之内, 同时避免了样本的重叠现象. 相对于 KNN, 改进算法所生成的样本分布更加均匀, 携带的信息量更大.

3 实验分析

3.1 实验设置

为验证本文遗传过采样算法的表现, 探究分类预测模型能否借助本文采样方法提升预测精度, 实验部分将本文方法 (GOS) 同 SMOTE、Borderline-SMOTE、

随机过采样 (ROS), 和非采样方法 (None) 进行比较. 实验数据集采用 UCI 公共数据集中的数据, 数据详情如表 1 所示, 其中 Yeast 数据集中将 EM1、EXC、VAC 作为正类, CYT、NUC、MIT 作为负类; Ecoli 数据集中将第 2、4、5、6 类作为正类, 其余作为负类.

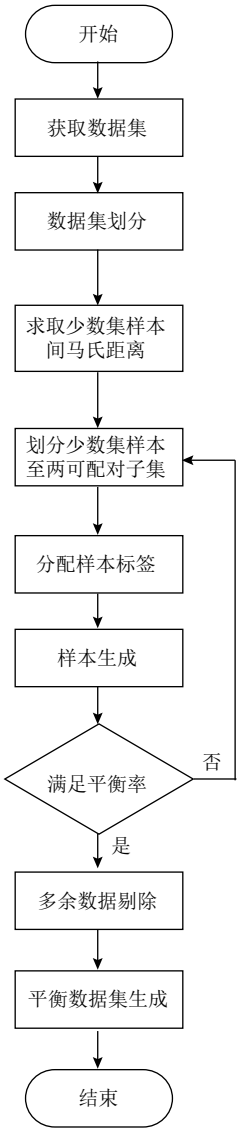

图 3 GOS 算法流程图

实验结果的评价指标采用召回率及几何平均值 *G-mean*, 其表达式如式 (2)、式 (3) 所示, 式中变量含义如表 2 所示. 其中, 召回率越高, 则说明分类器对少数类样本的识别性能越好, 可以反映出分类器对少数类样本的识别敏感度; *G-mean* 值弥补了召回率作为评价指标的片面性, 该评价公式将少数类的识别准确率和多数类的识别准确率同时考虑在内, 可更加综合的反映出算法的总体预测分类性能.

$$Recall = \frac{TP}{TP+FP} \tag{2}$$

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{3}$$

表 1 实验数据

数据集编号	数据集名称	属性数量	样本总数	平衡率(P_{fp})
1	Breast Cancer	9	286	0.42
2	Spambase	57	4 601	0.07
3	Hepatitis	19	155	0.26
4	Yeast	8	947	0.03
5	Steel Pastry	27	1 941	0.09
6	Ecoli	7	281	0.03

表 2 混淆矩阵

	预测正类 (P)	预测负类 (N)
实际正类 (P)	TP	FN
实际负类 (N)	FP	TN

对实验数据的准备主要包括预处理、备份、数据划分. 具体为首先经过对实验数据执行集成、规约、变换等预处理后将 6 个数据集进行复制备份至 5 份, 分别采用 5 种采样方式进行采样得到目标实验数据集, 对每个数据集按照 4:1 的比例划分训练集和测试集, 而后, 分别在 3 种分类模型 (BP、SVM、决策树) 的作用下进行测试, 以对比各采样方式在不同分类模型中对最终结果的影响, 其中每项最终实验数据均采用 10 折交叉验证的方式产生. 其中, 采用 3 种对比算法的目的是减小实验结果的偶然性, 提升实验结果的说服力.

3.2 实验结果分析

经过对比试验, 各分类算法在不同采样方式的作用下产生的分类结果如表 3、表 4 所示, 表 3 为各算法在召回率 (*Recall*) 评价中的表现. 由表中数据可以看出, 由于数据集 Breast Cancer 和 Hepatitis 平衡率较高, 采样算法甚至无需执行, GOS 采样方式对其召回率的提升并不明显, 除了 SMOTE 采样方式下的 BP 实验结果和 Borderline-SMOTE 采样方式下的 SVM 实验结果外, 本文采样方式下的分类结果和相对应的分类算法所获取的次优结果相比, 仍然以最低 1%, 最高 4% 的优势取得最优的召回率值; 对于数据集 Spambase 和 Steel Pastry, 其不平衡率相对加剧, GOS 采样方式对其召回率的提升相对明显, 和次优结果相比平均提升了 4.1%; Yeast 和 Ecoli 数据集的平衡率最低, GOS 采样

方法对各算法的召回率性能提升也最为明显, 达到了4.8%. 表 3 数据表明采样方法可以提升算法对正类样本的错分概率, 有效缓解不平衡数据对算法性能的影响.

表 4 为各分类算法在不同采样方式的作用下取得的 *G-mean* 值, 从表中数据可得, Breast Cancer 和 Hepatitis 两个较为平衡的数据集在 SMOTE 和 Borderline-SMOTE 采样方法中以 BP 分类算法和 SVM 分类算法取得了 1% 优势, 这是由于采样方法对数据集的改变较弱, 实验结果主要取决于原始数据, 采样对算法性能提升不明显; 除此之外, 本文采样法下的算法均取得了最优的 *G-mean* 值. 表 4 的实验结果进一步印证了本文算法的有效性.

表 3 各算法在不同采样方式下的 *Recall* 值

采样方法	分类算法	Breast Cancer	Spambase	Hepatitis	Yeast	Steel Pastry	Ecoli
SMOTE	BP	**0.94**	0.79	0.81	0.77	0.83	0.81
	SVM	0.93	0.85	0.86	0.81	0.77	0.74
	决策树	0.93	0.89	0.83	0.88	0.84	0.83
Borderline-SMOTE	BP	0.94	0.81	0.85	0.81	0.84	0.82
	SVM	0.94	0.85	**0.86**	0.90	0.77	0.80
	决策树	0.93	0.83	0.85	0.91	0.91	0.85
ROS	BP	0.92	0.78	0.69	0.65	0.69	0.64
	SVM	0.90	0.81	0.75	0.69	0.72	0.71
	决策树	0.92	0.69	0.72	0.70	0.75	0.72
None	BP	0.90	0.60	0.67	0.57	0.66	0.59
	SVM	0.89	0.63	0.57	0.61	0.71	0.58
	决策树	0.91	0.59	0.69	0.68	0.75	0.71
GOS	BP	0.93	**0.88**	**0.89**	**0.84**	**0.86**	**0.89**
	SVM	**0.94**	**0.86**	0.85	**0.96**	**0.83**	**0.86**
	决策树	**0.96**	**0.90**	**0.87**	**0.96**	**0.93**	**0.87**

表 4 各算法在不同采样方式下的 *G-mean* 值

采样方法	分类算法	Breast Cancer	Spambase	Hepatitis	Yeast	Steel Pastry	Ecoli
SMOTE	BP	**0.89**	0.78	0.82	0.76	0.81	0.77
	SVM	0.91	0.82	0.85	0.77	0.78	0.75
	决策树	0.88	**0.87**	0.80	0.83	0.79	0.82
Borderline-SMOTE	BP	0.93	0.81	0.85	0.75	0.83	0.75
	SVM	0.92	**0.82**	0.82	0.81	0.78	0.79
	决策树	0.85	0.76	0.79	0.88	0.89	0.80
ROS	BP	0.92	0.75	0.66	0.64	0.67	0.63
	SVM	0.93	0.82	0.75	0.69	0.78	0.71
	决策树	0.89	0.69	0.71	0.69	0.71	0.67
None	BP	0.91	0.55	0.63	0.44	0.65	0.57
	SVM	0.93	0.62	0.56	0.54	0.63	0.54
	决策树	0.85	0.56	0.55	0.59	0.72	0.67
GOS	BP	0.88	**0.82**	**0.87**	**0.86**	**0.85**	**0.84**
	SVM	**0.96**	0.81	**0.88**	**0.85**	**0.82**	**0.87**
	决策树	**0.93**	0.86	**0.84**	**0.91**	**0.90**	**0.85**

对本文所提出方法的表现更为优异的原因进行简要分析, 这可以归因于该方法所生成的样本在少数类边界内简洁且同原样本的分布更为相似, 扩散更加均匀. 从样本多样性角度考虑, 通过计算重采样数据样本中少数类中每个度量的多样性, 每个指标的多样性计算使用香农多样性指数来评价, 通过该评价指标我们发现 GOS 对样本多样性的增加是最为明显的.

4 总结与展望

本文针对现行分类算法对不平衡数据集的正类数据预测性能偏低的情况提出一种基于遗传思想的过采样方法, 该方法在不改变数据分布的前提下, 通过合成少数类数据实例来平衡数据集中的正负样本成分. 该采样方式避免了常见合成方式会产生错误或重复的数据导致高负类率的情形, 马氏距离的使用确保了合成数据不会跨越分类算法的决策边界. 通过在 6 个公共数据集上使用 3 种分类模型, 将本文方法和其他 4 种采样方法进行比较, 经 90 个实验组合结果验证, 本文采样方式在召回率和 *G-mean* 两个评价指标上均取得了综合最优的结果, 证明了本文采样方式的有效性. 在未来的研究中, 对 GOS 在多分类模型中的引入应用可进一步扩展该采样法方法的应用价值.

参考文献

1 丁雪松, 黄立群, 张步忠, 等. 基于多层感知机的蛋白质变性温度预测. 计算机应用研究, 2019, 36(8): 2421–2423.

2 李卫疆, 常伟, 余正涛. 加快排序文档的剪枝决策树和分块方法. 计算机应用研究, 2020, 37(1): 193–197.

3 Liu SY, Xiao J, Xu XK. Sign prediction by motif naive bayes model in social networks. Information Sciences, 2020, 541: 316–331. [doi: 10.1016/j.ins.2020.05.128]

4 赖永凯, 陈向宇, 刘海. 基于贝叶斯 Logistic 回归的软件缺陷预测研究. 计算机工程与应用, 2019, 55(11): 204–208, 220. [doi: 10.3778/j.issn.1002-8331.1812-0254]

5 Xu LX, Wang XF, Bai L, *et al*. Probabilistic SVM classifier ensemble selection based on GMDH-type neural network. Pattern Recognition, 2020, 106: 107373. [doi: 10.1016/j.patcog.2020.107373]

6 Shahraki A, Abbasi M, Haugen Y. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. Engineering Applications of Artificial Intelligence, 2020, 94: 103770. [doi: 10.1016/j.engappai.2020.103770]

7 Greiner R, Grove AJ, Roth D. Learning cost-sensitive active classifiers. Artificial Intelligence, 2002, 139(2): 137–174. [doi: 10.1016/S0004-3702(02)00209-6]
8 赵婧, 邵雄凯, 刘建舟, 等. 文本分类中一种特征选择方法研究. 计算机应用研究, 2019, 36(8): 2261–2265.
9 李春雪, 谢林森, 卢诚波. 面向不平衡数据集的一种基于聚类的欠采样方法. 数学的实践与认识, 2019, 49(1): 203–209.
10 Bellinger C, Drummond C, Japkowicz N. Manifold-based synthetic oversampling with manifold conformance estimation. Machine Learning, 2018, 107(3): 605–637. [doi: 10.1007/s10994-017-5670-4]
11 黄贤英, 熊李媛, 刘英涛, 等. 基于类别特征改进的 KNN 短文本分类算法. 计算机工程与科学, 2018, 40(1): 148–154. [doi: 10.3969/j.issn.1007-130X.2018.01.022]
12 吴凯, 罗朝晖. 昆虫学案例在遗传学教学中的应用. 遗传, 2019, 41(4): 349–358.
13 Zhao PN, Lai LF. Analysis of KNN information estimators for smooth distributions. IEEE Transactions on Information Theory, 2020, 66(6): 3798–3826. [doi: 10.1109/TIT.2019.2945041]
14 刘云, 肖雪. 基于邻域维护准则的特征选择算法优化研究. 重庆大学学报, 2019, 42(3): 58–64.
15 Martin B. Persistent bias on wikipedia: Methods and responses. Social Science Computer Review, 2018, 36(3): 379–388. [doi: 10.1177/0894439317715434]
16 Zhang CX, Wang GW, Zhang JS, *et al*. IRUSRT: A novel imbalanced learning technique by combining inverse random under sampling and random tree. Communications in Statistics-Simulation and Computation, 2014, 43(10): 2714–2731. [doi: 10.1080/03610918.2013.765467]
17 Chawla NV, Lazarevic A, Hall LO, *et al*. SMOTEBoost: Improving prediction of the minority class in boosting. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin Herdelberg: Springer, 2003. 107–119.
18 Freund Y, Schapire RE. A decision theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 1997, 55(1):119–139.
19 Kinoshita T, Fujiwara K, Kano M, *et al*. Sleep spindle detection using RUSBoost and synchrosqueezed wavelet transform. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2020, 28(2): 390–398. [doi: 10.1109/TNSRE.2020.2964597]
20 李克文, 林亚林, 杨耀忠. 一种改进的基于欧氏距离的 SDRSMOTE 算法. 计算机工程与科学, 2019, 41(11): 2063–2070.
21 李雄飞, 李军, 董元方, 等. 一种新的不平衡数据学习算法 PCBoost. 计算机学报, 2012, 35(2): 202–209.