文章编号: 1001-0920(2019)04-0673-16 DOI: 10.13195/j.kzyjc.2018.0865

# 不平衡数据分类方法综述

李艳霞, 柴 毅, 胡友强, 尹宏鹏†

(1. 复杂系统安全与控制教育部重点实验室，重庆 400044；2. 重庆大学 自动化学院，重庆 400044)

**摘 要:** 随着信息技术的快速发展,各领域的数据正以前所未有的速度产生并被广泛收集和存储,如何实现数据的智能化处理从而利用数据中蕴含的有价值信息已成为理论和应用的研究热点. 数据分类作为一种基础的数据处理方法,已广泛应用于数据的智能化处理. 传统分类方法通常假设数据类别分布均衡且错分代价相等,然而,现实中的数据通常具有不平衡特性,即某一类的样本数量要小于其他类的样本数量,且少数类具有更高错分代价. 当利用传统的分类算法处理不平衡数据时,由于多数类和少数类在数量上的倾斜,以总体分类精度最大为目标会使得分类模型偏向于多数类而忽略少数类,造成少数类的分类精度较低. 如何针对不平衡数据分类问题设计分类算法,同时保证不平衡数据中多数类与少数类的分类精度,已成为机器学习领域的研究热点,并相继出现了一系列优秀的不平衡数据分类方法. 鉴于此,对现有的不平衡数据分类方法给出较为全面的梳理,从数据预处理层面、特征层面和分类算法层面总结和比较现有的不平衡数据分类方法,并结合当下机器学习的研究热点,探讨不平衡数据分类方法存在的挑战. 最后展望不平衡数据分类未来的研究方向.

**关键词:** 不平衡数据；机器学习；分类；深度学习

**中图分类号:** TP13 **文献标志码:** A

# Review of imbalanced data classification methods

*LI Yan-xia, CHAI Yi, HU You-qiang, YIN Hong-peng*†

(1. Key Laboratory of Complex System Safety and Control of Ministry of Education，Chongqing 400044，China; 2. College of Automation，Chongqing University，Chongqing 400044，China)

**Abstract:** With the rapid development of information technology, there are so much amount of data produced and collected in different domains. How to efficiently discover knowledge from these data has been a research focus. Classification algorithms, as a basic of data processing methods, are wildly used in data intelligent processing area. Traditional classification algorithms generally assume that the training sets are well-balanced with equal misclassification cost. However, data are usually class-imbalanced in real-world domains, which means one or some of the classes have less number of examples than others. Moreover, the minority class implies heavy cost when it is not well classified. The standard classification algorithms guided by global classification accuracy are often biased towards the majority class due to the imbalanced classification distribution. Therefore, it is required to enhance the accuracy of both the minority classes and the majority classes. The imbalanced data classification problem has received much attention from the machine learning community, and various approaches have been proposed to deal with the problem. This paper reviews the state-of-the-art imbalanced data classification methods in recent years, and analyzes and compares them comprehensively in accordance with essential difference from the data-preprocessing-level, feature-level and algorithm-level respectively. Then, considering the research focuses of machine learning field, the challenge of imbalanced data processing is discussed, followed with the prospects for future work.

**Keywords:** imbalanced data；machine learning；classification；deep learning

## 0 引 言

随着传感器技术、计算机技术、通信技术、数据存储等技术的高速发展,互联网、过程工业等领域产生并存储了大量数据. 如何对这些数据进行智能化分析处理,提取数据中蕴含的有价值信息和知识,已成为目前理论与应用研究的热点. 机器学习作为一种主流的智能数据处理技术,为实现上述目标提供了一种有效途径,其中分类算法作为机器学习的关键

**收稿日期:** 2018-06-25；**修回日期:** 2018-11-27.

**基金项目:** 国家自然科学基金项目(61633005,61773080)；重庆大学科研后备拔尖人才计划项目(cqu2018CDHB1B04).

**责任编委:** 侯忠生.

**作者简介:** 李艳霞 (1991−), 女, 博士生, 从事稀疏表示、数据分类的研究；尹宏鹏 (1981−), 男, 教授, 博士, 从事模式识别与智能系统等研究.

†通讯作者. E-mail: yinhongpeng@gmail.com.

技术之一,能够利用数据集构建一个具有较强泛化能力的分类模型,提取数据中的有用信息而受到广泛关注,成为机器学习领域中一个重要的研究内容.

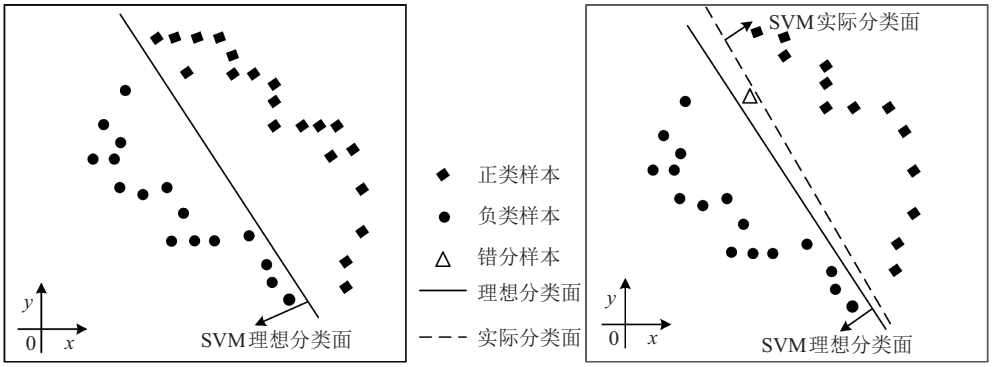
传统的分类方法通常假设数据集中各类别所包含的样本数相同且误分代价相等,然而现实世界中的数据往往具有不平衡特性,即数据集中某一类的样本数量要小于其他类别样本数量,并且具有少数样本的那类数据相比其他类更加重要,错分代价更高. 如在故障诊断中,故障样本通常少于正常运行数据,将“故障”误诊为“正常”使故障的系统继续工作,会导致无法预计的后果和损失. 通常而言,一个数据集称为不平衡数据集需要具备两个条件:类别数量的不平衡和错分代价的不平衡. 以不平衡数据集作为训练样本,构建训练样本与类别之间的关系模型,并对新的样本类别进行判断的问题称为不平衡数据分类问题. 当前,不平衡数据分类问题广泛存在于生物医疗[1-4]、金融[5-8]、信息安全[9-12]、工业[13-16]、计算机视觉[17-19]等诸多领域,表1具体列举出了典型的不平衡数据问题实际应用场景.

**表1 不平衡数据分类相关应用领域**

| 序号 | 应用领域 | 具体应用 | 相关参考文献 |
|---|---|---|---|
| 1 | 生物医学 | 疾病诊断、蛋白质鉴定、基因表达预测、存活率预测 | [1-4] |
| 2 | 金融 | 欺诈检测、贷款违约预测、企业破财预测 | [5-8] |
| 3 | 信息安全 | 软件缺陷检测、网络入侵检测、网络缺陷评估、网络服务质量预测 | [9-12] |
| 4 | 工业系统 | 状态监测、异常检测、故障诊断 | [13-16] |
| 5 | 计算机视觉 | 图像边缘检测、目标检测、异常事件检测 | [17-19] |

不平衡数据分类问题的特性主要体现在不平衡数据集自身的非均衡性以及传统分类算法的局限性两方面. 在数据层面,不平衡数据分类问题主要表现为少数类样本过少,进一步可以分为绝对稀少和相对较少. 绝对过少一般指少数类样本数目本身就稀少,相对过少是指少数类样本本身数量并不少,只是相对多数类而言少数类样本所占的比例较少. 在绝对过少情况下,由于少数类本身含有的信息有限,分类器难以通过训练充分学习到少数类的特征,使得少数类数据难以识别. 而在相对较少的情况下,由于多数类样本会模糊少数类样本的边界,在某些存在类别重叠的区域,很难有效地将少数类样本与多数类区分开来,容易造成分类器对少数类识别度下降.

在分类器层面,传统分类器通常以经验风险或结构风险最小为训练目标,使误分率最低或类间间隔最大. 当数据分布不平衡时,会使得类间重叠部分中属于少数类的样本被大量误分,类间隔面被移动到样本分布比较稀疏的类别那一边,从而无法保证算法对少数类的分类精度. 图1以经典的SVM分类器为例,给出了数据分布不平衡对分类器性能影响的可视化解释. 简便起见,假设该数据集满足线性可分条件. 如图1(a)所示,当数据集正负样本(分别对应图中方形和圆形)分布均衡时,基于支持向量寻找到的SVM分类面位置大致位于两类样本中间. 如果按照一定比例随机抽取出部分正类样本和原负类样本组成新的不平衡数据集,则由于少数类数据的支持向量较少,此时学习得到的SVM分类面会向少数类数据的方向偏移(如图1(b)所示),与理想分类面之间存在一定距离,导致新到达的少类样本被误判为多类样本(三角形),影响少数类样本的正确分类. 因此如何改善现有分类方法在不平衡数据分类问题上的性能,同时保证分类器对多数类和少数类的识别精度,是机器学习领域亟待解决的难点问题之一.




(a) 理想分类面 (b) 分类面发生偏移

**图1 数据的不平衡特性对分类器的影响**

近年来,不平衡数据分类问题已引起国内外众多学者的广泛关注. 2012年国际机器学习大会(International conference on machine learning and applications, ICML)以“Class imbalance: Past, present, future”为题针对不平衡数据学习问题进行了研讨; 2017年国际人工智能发展协会(Association for the advancement of artificial intelligence, AAAI)[20]和国际计算机视觉大会(International conference on computer vision, ICCV)召开的学术研讨会[21]、2016年计算机视觉模式识别会议(Conference on computer vision and pattern, CVPR)[22]和国际自动控制联合会(International federation of automatic control, IFAC)[23]都针对不平衡数据分类问题进行了探讨和研究. 随着对不平衡数据分类问题研究的深入,出现了大量优秀研究成果. 以“Imbalanced data learning”、“Unbalanced data learning”为关键字在Web of Science核心集中搜索到的SCI文章逐年增加,仅在短短的6年时间里,文章数就从2010年的79篇增加到2016年的253篇(如图2所示). 这些国际权威研讨会的召开和研究成果的出现表明了不平衡数据分类及应用的重要性.

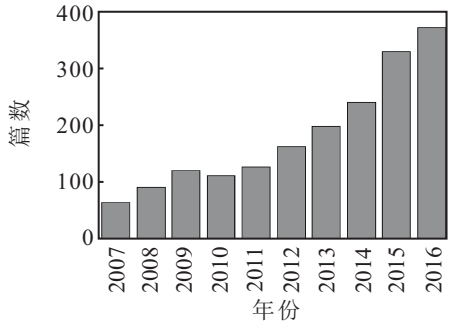



**图2　SCI数据库关于不平衡数据分类论文的发表趋势**

一般而言,典型的数据分类框架主要包含数据预处理、特征和分类器算法3个主要层面,如图3所示. 在数据预处理层面,对数据集进行训练数据集和测试数据集的划分;在特征层面,采用合适的特征模型将数据映射至特征空间,得到训练数据和测试数据的特征表示;在分类器层面,利用训练数据训练分类器,并借助训练好的分类器判断测试数据属于何种类别. 针对不平衡数据分类这一特殊的模式分类问题,相应的研究也主要围绕数据预处理、特征、分类算法3个层面展开. 数据预处理层面上,针对数据稀少问题,通过改变训练集样本分布,降低或消除不平衡性;特征层面上,选择具有较好区分性能的特征,提高少数类的识别率;算法层面上,根据传统分类算法在解决不平衡问题时的缺陷,结合不平衡数据的特点,适当地改进算法以提高对少数类样本的识别率. 数据预处理、特征和分类算法3个层面的平行发展,为不平衡数据分类问题提供了多种解决思路.

国内外的研究学者从数据预处理层面、特征层面和算法层面出发,针对不平衡数据分类问题提出了大量优秀算法,相应的综述性文章[24-33]对这些研究工作进行了总结和梳理. 在已有的综述文献中,部分工作只针对数据层面、特征层面或算法层面的一部分算法进行了归纳和总结,如文献[24]针对不平衡数据处理的各种采样方法展开了综述,文献[25]对基于集成学习的不平衡数据处理方法进行了深入的讨论;部分综述性工作只是侧重讨论了不平衡数据处理分类一个较小的组成部分,如文献[26]研究了数据不平衡程度和分类器对不同采样方法的影响,文献[27]探讨了不平衡数据分类中的一系列开放性问题,文献[28]深入讨论了不平衡数据分类中研究难点和发展方向;当然,也有一些学者对不平衡数据学习进行了比较全面的分析和整理,如文献[29-33];这些工作有效概括了当前的研究热点和发展方向. 然而,随着机器学习、人工智能的升温以及新问题、新技术的不断出现,不平衡数据学习方法也在不断发展,对这些工作进行梳理和总结会有助于把握不平衡数据分类问题的最新研究进展和发展趋势.

本文结合数据分类的一般框架,分别从数据分类的3个层面出发,在介绍不平衡数据分类经典方法的基础上,着重对近年来的研究工作进行了分析和概述,并对该领域面临的挑战和发展趋势进行了展望.

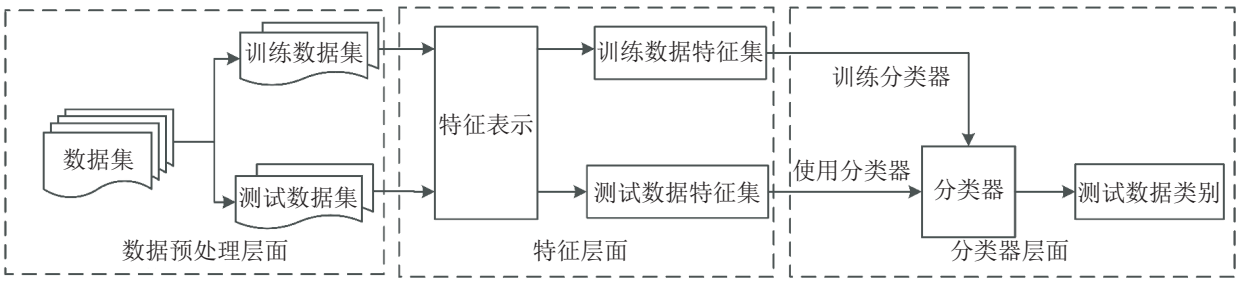



**图3　数据分类典型框架**

## 1 不平衡数据分类方法的研究现状

迄今为止,数据不平衡问题的研究主要在数据预处理层面、特征层面和分类算法层面展开,保证分类器对多数类和少数类的数据都具有较高的分类精度. 具体而言,在数据预处理层面,通过改变训练集样本分布降低或消除不平衡性,代表性方法是一系列重采样方法;在特征层面,考虑样本数量的不平衡分布伴随特征属性分布不平衡这一特性,利用特征选择方法选择具有区分特性的特征,提高少数类的分类精度;在分类算法层面,根据算法在解决不平衡问题时的缺陷,结合不平衡数据的特点,通过适当地改进算法以提高对少数类样本的识别率,典型的方法有代价敏感学习、集成学习、单类学习等. 图4根据分类算法的一般框架,总结了数据预处理层面、特征层面、分类算法层面相应的代表性方法. 本节也将从这3个层面出发,对不平衡数据分类的相关方法进行综述和比较,重点阐述现有不平衡数据分类方法的主要思想和特点.

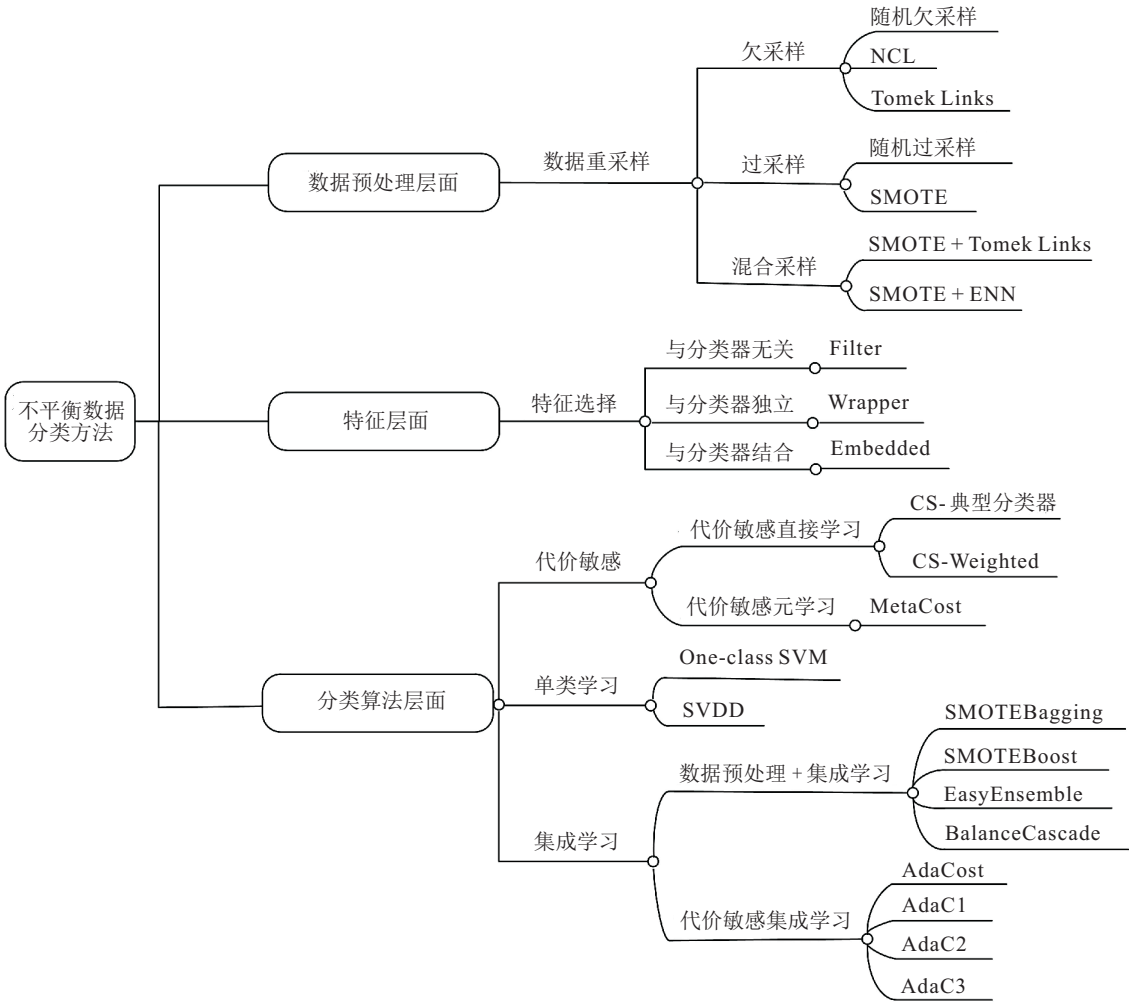



**图 4 不平衡数据分类代表性方法**

### 1.1 数据预处理层面的不平衡数据分类方法

在不平衡数据集中,少数类样本的信息量无法与多数类信息量抗衡,少数类信息被大量多数类样本数据所淹没,导致少数类被大量误分. 因此研究人员希望在数据层面降低少数类样本与多数类样本的不平衡程度,以适应传统的分类方法. 其中,数据重采样是数据层面最具代表性的不平衡数据分类方法. 按照采样方式不同,基于重采样的不平衡数据分类方法主要分为以下3种:欠采样、过采样、欠采样与过采样相结合的混合采样.

欠采样方法通过减少多数类样本的数量来提高少数类的分类准确率. 随机欠采样是一种最简单的欠采样方法,通过随机地选择一部分多数类样本,达到平衡样本数量的目的. Oquab等[34]在利用卷积神经网络(Convolutional neural network, CNN)进行目标检测时,在训练过程中随机选取背景图像块的10%进行训练,用以解决背景图像块与目标图像块的数量不平衡问题;方昊等[35]通过多次随机欠采样来减小单次随机欠采样带来的误差,有效解决了软件缺陷检测时的数据不平衡问题. 随机欠采样方法操作简

单,可以有效缩短模型的训练时间. 但随机地舍弃样本可能会移除多数类中潜在的有用信息,从而导致分类器性能降低. 因此,相关学者提出了启发式欠采样方法,经典的欠采样方法是邻域清理(Neighborhood cleaning rule, NCL)和Tome links法. Lin等[36]基于$K$近邻规则,选择聚类中心或聚类中心的邻近样本表示多数类数据,提出了两种基于聚类的欠采样算法;吴园园等[37]根据类重叠度抽取对分类起关键作用的支持向量,提出了一种基于类重叠度不平衡数据欠采样的方法. 这些方法通过启发式地利用欧氏距离和$K$近邻规则去识别可以合理修剪的样本,能有效克服随机欠采样容易丢失重要样本信息的问题. 但欠采样方法通过样本剔除有可能破坏样本集的分布信息,针对此问题,Ng等[38]提出了一种基于灵敏度的多元欠采样方法,通过灵敏度测度合理选择可以修剪的样本,最大限度地保留原数据集的分布信息.

与欠采样方法相反,过采样不对多数类样本进行任何处理,而是增加少数类样本的数量来提高少数类的分类性能. 最简单的过采样方法通过随机复制或者简单的旋转,使两类数据趋于平衡. 此方法实现简单,但是由于反复复制少数类样本而增加了分类算法过拟合的可能性. 为此,相关研究人员引入智能策略来合成新的少数类样本,提出了经典的过采样方法-少类样本合成过采样技术(Synthetic minority oversampling technique, SMOTE)[39]. SMOTE通过随机选择同类近邻的样本进行插值,生成无重复的新的少数类样本,能在一定程度上克服随机过采样方法的过拟合问题,但新样例的产生方法决定了其可能会出现过泛化、样本重叠、噪声等一系列问题. 为了解决以上问题,相继出现了一系列SMOTE的改进算法,如Borderline-SMOTE、N-SMOTE等. 近年来,研究人员也围绕以上问题开展了大量研究工作. Zhu等[40]在选择近邻样本时,引入样本选择权重,能有效解决过泛化问题;黄海松等[41]对不平衡数据集划分距离带,对距离带内的样本采用自适应变邻域SMOTE算法生成样本,但带数的合理划分对数据集分类精度的影响较大;Abdi等[42]提出了一种基于马氏距离的过采样方法,仅在少数类密集的区域中合成样例,能有效克服样本重叠问题,但该方法无法保证少数类的边界样本信息;Douzas等[43]在生成新的样本前,通过$K$均值聚类选择相对安全的数据样本,利用SMOTE算法生成新的数据,能有效克服噪声问题.

过采样方法能够有效扩大其规模使数据达到平衡状态,但上述工作大多基于数据的局部信息进行样本的增加,尽管在样本数量上实现了相对均衡化,但由于未考虑数据的整体分布信息,不能保证过采样后新数据集的数据分布情况. 针对此问题,Das等[44]提出了RACOG(Rapidy converging gibbs)和wRACOG(Wrapperbased rapidy converging gibbs)两种过采样方法,考虑数据的联合概率分布,并结合Gibbs采样器生成新的数据样本;Moreo等[45]在数据生成过程中利用数据的分布信息,提出了分布随机过采样(Distributional random oversampling, DOS)算法,有效提升了文本分类的精度.

此外,近年来相继出现了新的数据生成方法,为不平衡数据分类问题提供了新的解决思路和有效尝试. Liu等[46]将少数类数据的采样问题转化为数据估计问题,通过模糊学习生成新的数据;Razavi等[47]将缺失值填充的思想应用于数据过采样中,通过期望值最大法对缺失值进行估计更新,有效实现了故障数据集的生成;Douzas等[48]利用深度学习模型中的条件生成对抗网络(Conditional generative adversarial nets, CGAN)生成少数类样本,有效改善了数据集的不平衡特性;Ando等[49]提出了一种深度过采样模型(Deep over sampling, DOS),在特征空间中实现少数类的重采样,能有效提升对少数类的识别效果.

基于欠采样的方法和基于过采样的方法都有自身的长处和不足,为取得良好的数据处理效果,可将过采样和欠采样技术相结合进行混合重采样. 其基本思想是增加样本集中少数类样本的个数,减少多数类样本的个数,以此来降低不平衡度. Batista等[50]提出了SMOTE与Tome Link相结合的算法,能很好地克服SMOTE带来的噪声问题;陶新民等[51]在实现多数类样本逐级优化递减降采样后,进一步将该方法与只对边界样本进行升采样的算法相结合,能有效去除训练样本中噪声样本和重复信息,保留有用信息,提高有效数据的利用率;Zhang等[52]在视频图像背景提取过程中,提出一种时空过采样与选择性下采样相结合的数据混合采样方法,用以解决视频图像中前景像素和背景像素的不平衡问题;Li等[53]采用两个独立并行的粒子群优化过程,分别对数据进行欠采样和过采样处理,以相对较短的时间达到平衡数据的目的.

基于重采样的不平衡数据分类方法不依赖于所使用的具体的分类器,具有较好的适应性,但不适用于少数类样本数量太少甚至没有少数类样本的极端不平衡情况. 采用重采样方法处理不平衡数据时,如何合理确定重采样规模仍是基于重采样方法的难点所在. 目前,大多采样到所有类别样本数量相同为止,

然而,采样算法的采样率不仅仅取决于类别之间的不平衡比例. 进一步结合数据的分布信息,合理有效地确定重采样的最佳分类性能的采样率,提升不平衡数据分类性能,是值得关注的方向.

### 1.2 特征层面的不平衡数据分类方法

样本数量分布的不平衡性往往伴随着特征属性分布失衡,导致特征层面上信息的传递、表达不均衡,给少数类识别带来一定困难. 特征选择能根据一定的评价标准,从特征集合中选取具有代表性较优的特征子集,有效区分数据集中的每个数据对象(特征选择具体流程如图5所示). 结合特征选择的思想,研究人员提出一系列特征层面的不平衡数据分类方法,通过特征选择保留不平衡数据集的关键区分特征,增加多数类和少数类的区分度,提高少数类乃至整体的分类正确率. 为便于理解基于特征选择的不平衡数据分类方法,首先对典型特征选择方法进行简单介绍.

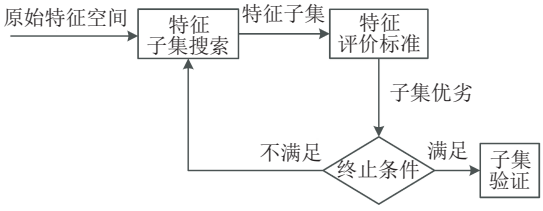



**图 5 特征选择算法流程**

特征选择方法按照所选取的评价指标是否与分类器相关,可分为过滤式(Filter),封装式(Wrapper)和嵌入式(Embedded)3种模型. 过滤式特征选择方法与分类器独立,通过分析原始特征集的内在特性,并结合相应的评价准则来选择出特征子集. 该方法计算复杂度低,通用性强,但由于方法独立于分类器,分类正确率较低. 封装式特征选择方法是一种与分类器相结合的特征选择方法,以分类错误率作为特征子集的评价指标,通过顺序式或启发式搜索策略选择出能够取得较高分类正确率的特征子集. 但封装式特征选择方法每次都要通过分类器的训练和测试来评价子集的优劣,计算代价较大. 为了解决此问题,嵌入式特征选择方法将特征选择与分类模型的学习过程相结合,在分类过程中包含特征选择的功能,具有相对较低计算代价的同时能够保证较好的分类精度,成为目前特征选择的研究热点.

基于特征选择的不平衡数据分类方法利用上述特征选择方法对不平衡数据集进行处理,进而选取有利于分类的区分性特征,提升少数类的识别率. 肖鹰等[54]通过数据集的不同类型属性权衡少数类样本的重要性,通过有效的评估参数筛选出对有效分类更有意义的属性,提高了少数类的分类性能; Hou等[55]通过一系列对比实验验证了Relief、FAST(Feature assessment by sliding thresholds)等过滤式特征选择方法的有效性; Wu等[56]针对不平衡文本分类问题,提出了一种ForesTexter特征选择方法; Han等[57]采用截断梯度法选取相关特征,并有效应用于在线不平衡数据的处理; Yin等[58]基于Hellinger距离对不平衡数据进行特征选择,Hellinger距离能有效度量分布的差异,且不涉及类别先验信息,因此对类别偏置不敏感,能有效克服样本分布不平衡对分类算法的影响; Maldonado等[59]针对高维不平衡数据,采用嵌入式特征选取方法选取出有利于识别目标类别的属性; Viegas等[60]将基因编程思想引入封装式特征选择算法中,指导特征的选取,有效解决了高维不平衡数据的特征分类问题; Zhou等[61]提出了一种基于$K$近邻相关性的高维不平衡数据特征选择方法; Liu等[62]通过优化分类评价指标筛选出有利于预测分类少数类样本的重要特性; Moayedikia等[63]利用对称不确定性衡量特征与分类标签的相关程度,选择与分类标签最为相关的特征,在高维不平衡数据中取得了较好的分类效果.

基于特征选择的不平衡数据分类方法通过选取具有区分能力的特征,提高少数类的分类准确度,主要应用于具有高维特性的不平衡数据处理中. 但特征选择的过程可能存在信息丢失,对后续分类的影响较大. 同时,现有特征层面的不平衡数据分类方法主要以特征选择为主,从特征提取角度解决不平衡数据分类问题的研究较少. 深度学习模型具有强大的特征提取能力,可以在不需要专家手工进行特征设计的条件下,学习表达能力强的抽象特征. 借鉴深度学习模型本身潜在的数据学习能力,提取能表征数据集特性的潜隐特征,可以在特征层面为不平衡数据分类的研究提供新思路和解决途径. 当前已有类似工作出现,如Ng等[64]利用两个激活函数不同的堆叠自编码网络对不平衡数据集进行学习,提取有利于分类的特征,有效解决了不平衡数据的分类问题. 但现有的基于深度学习的不平衡数据分类方法研究仍处于初始阶段,需要进一步探索和推进.

### 1.3 分类算法层面的不平衡数据分类方法

经典分类算法在对不平衡数据分类时,结果会向多数类倾斜,使得少数类容易被忽视. 以支持向量机为例,由于少数类样例较少,基于支持向量寻找到的分割超平面会向少数类偏移,这样少数类便会被错分

为多数类. 为了克服已有算法在解决不平衡问题时的缺陷,研究人员从分类算法本身出发,结合不平衡数据分布的特点,对现有的分类算法进行适当改进,以提高对少数类数据的识别. 典型的方法有:代价敏感法、单类学习法、集成学习法等.

### 1.3.1 代价敏感学习

经典分类方法通常假设各类误分代价相等,以全局误分率最低为目标,而在实际的分类问题中,不同的分类错误往往会带来不同的损失. 基于代价敏感学习的不平衡数据分类方法以代价敏感理论为基础,关注错误代价较高类别的样本,以分类错误总代价最低为诊断算法的优化目标. 目前对代价敏感学习的研究主要集中在代价敏感直接学习和代价敏感元学习两个方面.

代价敏感直接学习的基本思想是在传统学习算法的基础上引入代价敏感因子,通过改进分类器模型的内部构造,使基于最小错误率的分类器转化为基于最小代价的代价敏感分类器. 目前,主流的分类算法——人工神经网络、SVM和决策树等都有着相应的代价敏感扩展算法. 如代价敏感神经网络[65]直接嵌入各个类别样本的误分类代价以减少平均误分类代价;代价敏感支持向量机[66]分别给不同类别赋予不同的代价因子,获得最佳代价最小的分类结果;代价敏感决策树[67]在分裂标准和剪枝阶段注入代价敏感的思想,提高少数类的权重. 此外,一些工作在算法内部增加少数类的权重(CS-weighted),提高少数类的识别精度来解决不均衡数据分类问题,本质上也与代价敏感的思想类似[68-69]. 近年来,一些分类领域最新的研究成果,如稀疏表示、深度学习等,也引入了代价敏感学习的思想,以提升对少数类的识别精度. 代价敏感字典学习(Cost sensitive dictionary learning, CSDL)[70]结合稀疏表示字典学习以及代价敏感因子,在提升分类性能的同时解决了类不平衡的问题;史作婷等[71]在稀疏重构度量学习的基础上,为度量矩阵学习阶段引入代价敏感因子来减小样本错分代价;Chung等[72]将代价因子嵌入CNN的损失函数中,提升经典CNN对少数类样本的识别精度;Shen等[17]对CNN模型中的softmax分类器损失函数引入代价敏感因子,实现图像边缘检测;Raj等[73]对少数类和多数类的分类误差分别赋予不同的权值,能有效地处理极端不平衡数据集;Khan等[74]提出一种基于代价敏感学习的特征学习方法,对CNN模型中特征学习和分类器学习过程进行联合优化,使学习到的特征具有鲁棒性和判别性. 但将代价敏感学习方法应用于深度学习中,用于解决不平衡数据分类问题目前仍处于初始探索阶段,需要进一步分析研究和探索.

代价敏感直接学习通过对分类器进行改造,使已有的分类算法适合于不平衡数据的分类. 但仍存在一些分类器不能直接应用代价敏感学习机制,此时只能间接地进行代价敏感学习,代表性的方法是代价敏感元学习. 代价敏感元学习将样本相关的错分代价转换成样本权重,按权重对原始样本集进行重构. MetaCost[75]是一种典型的代价敏感元学习方法,通过估计训练样本的后验概率密度,结合代价矩阵计算每个训练样本的理想类别,然后根据理想类别修改原训练样本的类别,得到新的训练集,最后使用基于错误率的分类器学习这个新的训练集. 另外,Chew等[76]通过训练集先验信息的分析,利用支持向量机为不同类的样本设置惩罚系数,给不同的训练样本赋予不同的权值也起到代价敏感学习的作用;Zhou等[77]改变训练数据集中各个类别的实例数据比例,使得分类器偏向于少数类,从而具有代价敏感的性质,其缺点是重构的过程中丢失了一些有用的样本信息,并改变了样本的分布情况.

基于代价敏感学习的不平衡数据分类方法通过引入代价来指导学习过程,是处理不平衡数据分类问题的有效方法,但当少数类样本数量太少甚至没有少数类样本时,代价敏感的方法不再适用. 同时,基于代价敏感学习的不平衡数据分类方法的核心是错分代价的确定,在大多数情况下,很难对真实的错分代价做出准确的估计. 当前,误分类代价的确定与各类别的样本个数密切相关,通常以各类在数据集中的样本数量或各类样本数的比例作为错误分类代价,不能较好地反映数据真实的类分布特征,无法保证代价敏感学习的效果. 可行的解决思路是以多数类和少数类的分类性能最佳为标准,结合数据的分布特性构造代价分布空间,利用搜索和优化技术,寻找一个最适应数据集特性的误分类代价.

### 1.3.2 单类学习

在很多情况下,比较容易得到大量的多数类数据样本,而少数类样本难以获得,表现为极端不平衡问题. 为了解决该问题,利用只含有单一类样本进行训练成为有效的解决方法. 单类学习作为一种可以利用少量的标注样本和大量无标注样本进行学习的方法,为解决不平衡数据学习问题提供了一种新思路.

基于单类学习的不平衡数据分类方法的主要思想是只对多数类样本进行训练,形成一个对该类别的数据模型. 其目标是从测试样本中识别出多数类

样本,而不是对少数类和多数类进行区分. 对于新的样本,通过设计相似度度量并设定阈值来判断新样本的归属. 代表性工作主要有单类支持向量机(One-class SVM)及各种优化算法[78-79]、支持向量数据描述(Support vector data description, SVDD)[80]等.

单类支持向量机通过在高维特征空间求解一个最优超平面实现多数类样本与坐标原点的最大分离. 支持向量数据描述的基本思想是,通过在映射到高维的特征空间中找出一个尽可能小的超球体描述多数类样本的边界分布情况. 二者均具有非常强的单类数据处理能力. Luca等[81]针对异常检测中正常数据与异常数据的极端不平衡问题,采用单类支持向量机有效实现了异常检测; Yin等[82]针对单类支持向量机存在易受噪声样本干扰的问题,通过自适应调节样本的权值,有效提高了算法的鲁棒性;韩志艳等[83]将SVDD中的经验误差替换为对样本类别分布鲁棒的曲线下面积,提出了不平衡支持向量数据描述算法,能有效地描述多数类和少数类样本之间的边界,实现了故障诊断;姚宇等[84]提出加权超椭球体支持向描述方法,充分考虑样本分布信息,并对不同类样本赋予不同权重,具有较强的异常检测能力.

基于单类学习的不平衡数据分类方法只需要一类数据集作为训练样本,能有效减少时间开销,适用于少数类样本非常少或类间不平衡度很高的极端情况,但容易陷入对训练集中少数类样本的过拟合而导致泛化能力下降. 在单类学习中,样本类别的确定依赖于阈值,如何选取合理的阈值是需要解决的关键问题之一. 同时,现有的单类学习方法大多基于核函数来寻找非线性决策边界,性能在很大程度上还依赖于核函数的选取,针对此问题,可通过采用一定的优化算法对参数进行优化选择,提高算法的收敛速度和泛化能力.

### 1.3.3 集成学习

集成学习方法通过将多个基分类器的分类结果按一定方式集成来提升分类器的泛化性能,进而获得较高的分类结果. 典型的集成学习方法有Bagging、Boosting等. 基于集成学习的不平衡数据分类方法的基本思想是将标准的集成学习算法与现有的不平衡数据分类方法相结合,以适应不平衡数据处理问题的需求,在一定程度上属于一种混合的不平衡数据分类方法. 文献[25]对基于集成学习的不平衡数据分类方法进行了较为完整的梳理和总结,根据结合方法的不同,将目前基于集成学习的不平衡数据处理方法分为数据预处理与集成学习相结合的方法和代价敏感集成学习方法两大类,如图6所示.

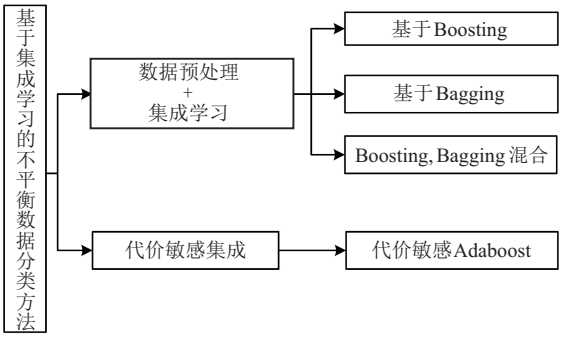



**图6 基于集成学习的不平衡数据分类方法[25]**

数据预处理与集成方法相结合的不平衡数据处理方法主要是将重采样方法嵌入集成学习过程中,根据集成方式不同,进一步分为基于Bagging、基于Boosting和混合集成的不平衡数据处理方法. 基于Bagging的不平衡数据处理方法的关键在于处理不平衡数据获得有效分类器的同时保证基分类器的多样性,代表性的算法有SMOTEBagging、UnderOverBagging[85]. 基于Boosting的不平衡数据处理方法将采样技术嵌入到Boosting算法中,朝着小类的方向改变用于训练下一个分类器的数据分布. 典型的算法有SMOTEBoost[86]、RUSBoost[87]等. 混合方法采用双重集成,将Bagging与Boosting组合起来. EasyEnsemble和BalanceCascade[88]是典型的双重集成算法,采用Bagging作为基本的集成学习方法,并在训练每个Boostrap数据时,使用AdaBoost作为分类算法. 相关研究表明,集成间隔是影响集成学习分类性能的关键因素,因此Feng等[89]提出了一种基于集成间隔的不平衡数据分类方法,在采样过程中考虑样本的间隔分布信息,选择间隔较小的样本,与SMOTEBagging和UnderBagging相比,能取得更好的分类效果; Sun等[90]认为采样会影响数据集的分布信息,因此仅采用聚类的方式构造多个相对平衡的子集,并集成多个子集的分类结果,能够获得更好的分类性能;陈圣灵等[91]结合SMOTE、Bagging、Boosting等方法,提出了一种通过数据采样间接改变样本权重的集成学习方法,能够提升少数类样本的分类准确率; Yuan等[92]提出了一种正则化的集成深度学习框架,在采样处理的基础上,通过正则化因子修正基分类器的分类误差,有效提升分类性能; Zhao等[93]考虑多数类样本的内部结构信息,通过无监督方式将多数类样本分解为多个子集,再集成各子集的分类结果,也取得了较好的分类结果.

代价敏感的集成学习方法不再调整单个分类器来体现不同类的误分代价,而是通过集成学习算法来引导代价的最小化过程. 通过这种方法,避免

了对基分类器的不断调整. 目前解决不平衡分类问题的代价敏感集成方法主要采用不同的方法更新Adaboost 的权重, 使得算法对于不同类别区别对待. 在权重更新方法中引入不同的代价则形成了各种代价敏感Boosting算法, 代表性方法有: AdaCost、AdaC1、AdaC2和AdaC3等[94]. Krawczyk等[95]通过集成多个代价敏感分类器, 有效提升了疾病监测的准确率; 付忠良[96]以误检标签代价和漏检标签代价之和为误分代价, 提出了一种多标签代价敏感分类集成学习算法, 可以自动学习多个弱分类器来组合成强分类器, 实现平均错分代价的最小化. 近年来, 出现了一些工作将栈式集成的方法应用于不平衡数据处理中. Cao等[97]分别在数据层和特征层面采样代价敏感方法学习, 通过堆叠集成的方式获得最终的分类结果, 具有更好的泛化性能; Yan等[98]提出采用栈式集成的方法对不平衡数据进行处理, 在采样集成的基础上进一步进行代价敏感集成, 能够取得比单个集成模型更好的分类效果.

基于集成学习的不平衡数据处理方法取得了一定的成功, 但集成算法对基分类器的训练过程较为复杂, 时间花费较高, 并在处理高维数据上存在一定的局限性. 同时, 现有的基于集成学习的不平衡数据处理方法大多存在基分类器类型和数量的选择难题. 针对此问题, 可尝试结合数据集的特性, 选择不同类型的基分类器进行组合, 并以不平衡数据分类性能为目标, 对各基分类器的数量进行寻优, 从而获得较好的分类性能.

## 2 不平衡数据分类方法性能评价

构建完成一个分类器, 需要选择评价标准对分类器的性能进行评估. 常规的分类方法一般以准确率(Accuracy)作为评价指标. 当样本呈现不平衡分布时, 由于少数类样本对总体准确率的影响较小, 即使分类算法将全部样本视为多数类, 仍然可以获得较高的准确率, 可见单独使用准确率作为评价指标, 难以准确反映出分类器在不平衡数据集上的分类性能. 因此, 对于不平衡数据处理方法的性能评估问题, 还需要适应于不平衡数据集特点的评价指标. 本节主要介绍面向不平衡数据分类的性能评价指标, 并给出了常用的公开数据集和典型的不平衡数据处理方法性能评价结果.

### 2.1 不平衡数据分类方法性能评价指标

针对不平衡数据分类性能评估问题, 相关学者提出一系列新的评价方法和指标. 典型的评价指标有$F$-measure、几何平均准则($G$-means metric)和接收者操作特征曲线(Receiver operating characteristic, ROC). 以上评价方法大多采用混淆矩阵进行表示, 为了表述方便, 先对混淆矩阵进行简单的说明. 在两类分类问题中, 将分类学习任务关注的少数类定义为正类, 多数类定义为负类, 混淆矩阵如表2所示. 其中: TP表示正类样本预测仍为正类, FN表示正类样本预测结果为负类, FP表示负类样本预测为正类, TN表示负类预测仍为负类. 在混淆矩阵的基础上, 下式给出了准确率(Accuracy)这一基本评价指标:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}. \tag{1}$$

**表 2 混淆矩阵**

| 类别 | 正类样本 | 负类样本 |
|---|---|---|
| 正类样本 | True positive(TP) | False negatives(FN) |
| 负类样本 | False positive(FP) | True negatives(TN) |

$G$-means是一种简单有效的衡量不平衡数据分类方法的指标, 定义为

$$G\text{-means} = \sqrt{A^+ \times A^-}. \tag{2}$$

可见, $G$-means综合考虑了两类样本的分类准确率, 相对于准确率而言, 能更好地衡量分类方法在不平衡数据集上的分类效果, 而且简便有效易于理解, 已成为不平衡数据处理领域常用的方法之一. 其中

$$A^+ = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{3}$$

$$A^- = \frac{\text{TN}}{\text{TN}+\text{FP}}. \tag{4}$$

$F$-measure也是一种常用的不平衡数据集分类问题评价指标, 定义为

$$F\text{-measure} = \text{Recall} \times \text{Precision}. \tag{5}$$

其中

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}. \tag{7}$$

可见, $F$-measure值的大小与查全率(Recall)和查准率(Precision)直接相关, 只有当查全率和查准率都大时, $F$-measure的值才会相应增大. 该方法主要是在查全率和查准率平衡的前提下尽可能将其最大化, 因此可以正确评价分类器对于多类和少数类样本的分类性能.

ROC曲线是以对多数类样本的分类错误率为横坐标, 对少数类样本分类准确率为纵坐标绘制的曲线如图7所示, 反映了当分类器参数变化时TPR与FPR之间的相对变化情况, 曲线越靠近左上角表示分类器性能越好. ROC曲线能够比较全面地描述分类器的性能, 是目前评价不平衡数据集分类器性能的常用方法之一, 主要通过判断ROC曲线的斜率度来判定

性能. 但不同ROC可能存在交叉情况,同时,ROC曲线不能定量评价分类器的性能. 针对以上问题,研究人员提出了采用ROC曲线下的面积AUC(area under the curve, AUC)来定量评估分类器的性能,AUC的值越大说明分类器的性能越好.

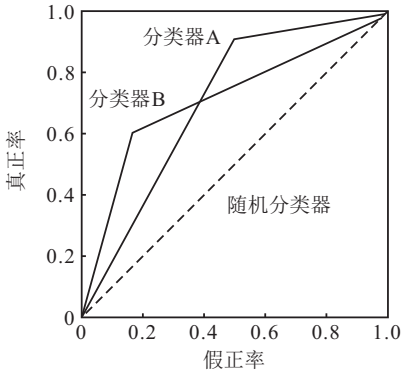



**图 7 ROC曲线示意图**

以上典型的衡量指标可实现不平衡数据集下分类方法的性能评价,当应用这些指标进行算法性能评估时,指标的选取要结合具体的应用场景和对象进行合理有效的选择. 同时,现有的不平衡数据处理方法评价指标大多针对两类分类问题,尽管多分类问题通常可以简化为两分类问题来解决,但仍需要针对多类类别不平衡分类问题的指标评价标准,这方面的工作需要进一步的探索和研究.

### 2.2 典型不平衡数据分类方法性能评价

目前,在不平衡数据处理领域已经公开发布了许多可供算法评测的数据集,表3对典型的数据集及其特点进行了简要的归纳,并给出了相关数据集的下载链接. 本文选取KEEL数据集中的5个代表性不平衡数据集,并以SVM作为基本分类器,以$F$-measure作为评价指标,对典型的不平衡数据分类方法进行比较. 表4列举了所选取数据集的相关信息,评估的算法主要包括随机欠采样(Random undersampling, RUS)、 随机过采样(Random oversampling, ROS)、SMOTE、代价敏感学习方法CS-SVM、集成学习方法SMOTEBagging(简写为SBAG)、RUSBoost(简写为RUSB)等. 评估结果如表5所示.

**表 3 典型的公共不平衡数据集**

| 序号 | 数据集 | 特点 | 主页链接 | 参考文献 |
|---|---|---|---|---|
| 1 | UCI | 机器学习领域常用数据集,包含大量不平衡数据集 | http://www.ics.uci.edu/mlearn/MLRepository | [99-103] |
| 2 | KEEL | 数据集包含不平衡程度1.5～9以及9以上两部分数据集 | http://www.keel.es/dataset.php | [104-107] |
| 3 | KDD-99 | 典型的用于入侵检测的数据集 | http://kdd.ics.uci.edu/ | [108,109] |
| 4 | Microarray | 数据集数据具有高维特性 | http://www.cs.binghamton.edu/lyu/KDD08/data/ | [60,110] |
| 5 | LIBSVM | 数据集数据量较大 | https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/ | [111-113] |
| 6 | Library | 数据集数据量较大 | http://www.vision.uji.es/sanchez/Databases/ | [114] |

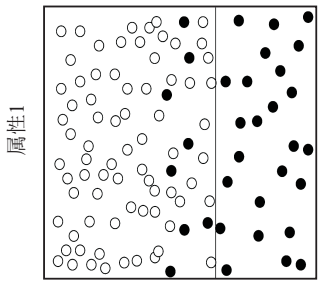
**表 4 不平衡数据集**

| 数据集名称 | 样本数 | 属性数 | 不平衡程度 |
|---|---|---|---|
| Vehicle1 | 846 | 18 | 2.9 |
| Ecoli034vs5 | 200 | 7 | 9 |
| Ecoli01vs5 | 240 | 6 | 11 |
| Abalone9-18 | 731 | 8 | 16.4 |
| Shuttle6vs23 | 230 | 9 | 22 |

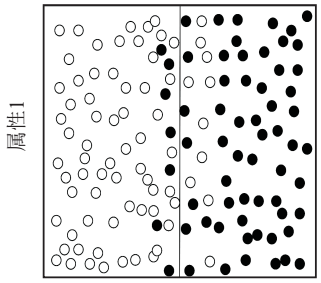
**表 5 以$F$-measure为指标的典型算法性能对比**

| 数据集 | SVM | RUS | ROS | SMOTE | CS-SVM | RUSB | SBAG |
|---|---|---|---|---|---|---|---|
| Vehicle1 | 0.746 | **0.787** | 0.739 | 0.742 | 0.751 | 0.647 | 0.642 |
| Ecoli034vs5 | 0.835 | 0.647 | **0.842** | 0.767 | 0.829 | 0.641 | 0.834 |
| Ecoli01vs5 | 0.812 | 0.693 | 0.754 | 0.736 | **0.820** | 0.775 | 0.792 |
| Abalone9-18 | 0.403 | 0.282 | 0.337 | 0.353 | 0.362 | 0.331 | **0.421** |
| Shuttle6vs23 | 0.794 | 0.526 | 0.717 | **0.854** | 0.818 | 0.713 | 0.812 |

从评估结果可以看出,数据的不平衡特性会对分类算法的分类性能产生影响,不平衡数据分类方法能够有效提升分类器的性能. 其中,过采样算法在不平衡程度较高的数据集上的性能较好,欠采样算法在不平衡程度较低的数据集上的性能更优. 在基于集成学习的不平衡数据分类方法中,与过采样相结合的集成方法能在不平衡程度较高的数据集上取得较好的效果,与欠采样相结合的集成方法更利于处理不平衡程度较低的数据集. 同时,分类器性能不仅与数据的不平衡程度有关,数据集的其他特性如样本数量、数据分布等因素也会影响分类器的性能,因此并不是每种不平衡分类方法都能得到较为显著的提升效果. 以过采样为例,当生成的样本使类间的重叠度变高时,反而会使得误分的样本增多,导致分类性能下降,如图(8)所示. 在处理不平衡数据分类问题时,无法判断具体何种不平衡数据分类方法更优,应根据具体的数据特性和需求选取合适的方法.




(a) 少数类数据过采样前的分类面




(b) 少数类数据过采样后的分类面

**图 8 类间重叠对不平衡数据分类的影响**

## 3 面临的挑战与研究展望

不平衡数据分类方法经过长时间的发展,已经得到了广泛的研究并取得了一系列研究成果,但仍存在一些挑战性的问题有待进一步研究,具体体现在以下几个方面:

1) 大规模不平衡数据分类问题.

近年来,随着信息技术的飞速发展和日益成熟,数据量呈现爆发式增长,产生和收集了大量数据. 大规模数据集在带来丰富数据量的同时,也给不平衡数据分类方法本身带来了一定挑战,当利用采样方法、代价敏感方法处理大规模数据时,其计算复杂度和内存消耗非常大,算法执行时间会随着样本数量和不平衡度的增加而增加. 如何克服数据采样、特征选择、集成学习等方法在处理大规模数据集时存在的计算量过大、计算效率低等问题,是当前不平衡数据分类亟待解决的难点所在. 目前的研究主要集中于大规模数据分类问题,对如何提高大规模不平衡数据分类性能讨论较少.针 对大规模数据分类问题,一般是将传统分类算法进行改进,利用MapReduce思想将任务分成若干个小模块并行处理[28].

2) 高维不平衡数据分类问题.

现实中的数据往往呈现出高维和不平衡双重特性,即数据属性较多且类别分布不均匀. 在高维特征空间下,数据分布更加稀疏,含有更多的冗余特征和不相关特征,从中获取有效信息却具有一定的困难性,从而导致少数类更加难以识别,给不平衡数据分类带来一定挑战,如何有效分类高维不平衡数据,是不均衡数据分类研究亟需解决的问题. 当前,高维不平衡数据分类问题的主要解决方式是通过降维、特征选择等预处理过程减少属性数量,使得预处理后的高维不平衡数据可以直接适用于已有不平衡数据方法,如代价敏感子空间学习[115]、以$F$-measure为优化目标的特征选择方法[62]等. 这些方法在进行高维数据处理时,往往忽略了数据内部的结构信息,对后续分类性能的影响较大.

3) 不平衡数据流分类问题.

现实中的数据通常以流的形式出现,呈现出高度动态变化的特性. 新到达的样例类别分布不确定,使得数据分布会随着时间发生变化,由此可能导致数据的多数类和少数类的变化. 然而,目前大多不平衡数据分类模型都是静态模型,即一旦被训练确定,则不会动态调整其参数和结构,使其无法实时地学习新增数据的特征. 因此,如何在线实时地处理连续动态变化的不平衡流数据,使得分类器随着数据分布的变化而动态更新,是不平衡数据分类的难点之一. 已有的数据流分类方法大多只关注数据流本身的动态变化特性,而缺少对其中类别分布不平衡性的考虑. 不平衡数据流分类方面的研究仅将数据流分类、不平衡数据分类两方面的研究成果进行简单组合[57,116-117],在算法设计时综合考虑数据流和类别不平衡问题的研究较少.

4) 少量标签数据的不平衡数据分类问题.

现有的不平衡数据分类方法主要是在有监督学习的框架下,通过充足的带标签数据确定分类边界. 而在现实分类问题中,由于标签数据获取的代价较大,存在大量只有少量标签且类别分布不平衡的数据. 数据的标记率低使得能参与训练的少数类样本数量进一步减少,难以有足够的数据建立分类模型,导致少数类的识别精度进一步降低. 如何提高不平衡数据分类方法在少量标签数据集上的分类表现,是不平衡数据分类中的又一瓶颈问题. 利用大量未标注样本信息改善现有不平衡数据分类学习算法的性能,是当前少量标签数据不平衡数据分类问题的主要解决思路. 基于数据层面的不平衡数据分类方法大多需要根据分类边界来确定人工样本的生成位置,因为有标签数据量较少,难以确定分类边界,所以在基于特征层面和分类算法层面的不平衡数据分类算法中引入未标注的数据信息[118-120],具有较高的时间复杂度.

针对上述挑战,不平衡数据分类问题的进一步研究可从以下4个方面展开:

1) 基于分布式计算的不平衡数据分类方法.

当处理大规模不平衡数据时,传统的集中式单机环境下的分类方法需要消耗大量的时间和空间物理资源,难以满足计算需求. 分布式计算技术利用多台

计算机协同工作,能够提供高性能的计算能力. 将集中式计算模式转变为分布式,实现计算模型的并行化处理,可以有效解决大规模数据处理问题. 近年来,随着各类分布式计算技术的发展与推广,特别是以MapReduce为代表的大数据分布式计算与编程平台的出现和成熟,为大规模数据的高效分析处理提供了强大的技术支撑. 结合现有的成熟的分布式计算框架,设计高效的分布式并行不平衡数据分类算法,是解决大规模不平衡数据分类问题的重要思路,值得进一步研究和探索. 其主要难点是如何设计以高效能为目标的大规模处理系统的系统架构、计算框架和处理方法.

2) 基于稀疏表示的不平衡数据分类方法.

稀疏表示能在稀疏约束条件下利用过完备字典对数据进行抽象,是表示和压缩数据的有效方法. 高维不平衡数据中存在大量无关、冗余的信息,借鉴稀疏表示潜在的数据表示能力,深度挖掘潜藏在高维数据内部的稀疏本征结构,能为高维不平衡数据分类问题的研究提供新思路和解决途径. 对基于稀疏表示的高维不平衡数据分类方法的进一步研究可能会产生突破性成果,最终将促进该领域的发展. 主要难点是如何设计与优化适用于高维不平衡数据分类问题的低维判别性字典.

3) 基于增量学习的不平衡数据分类方法.

在增量学习中,由历史数据训练得到的模型可通过新增训练样本的更新来得到新的模型,以符合当前的数据分布. 针对不平衡数据流特点,设计基于增量学习的不平衡数据分类方法,将新增的训练样本作为增量,通过增量式地学习新数据、更新已有模型,能有效学习到新增数据的特征,实现动态实时数据的处理,进而提升分类器的分类精度,是对现有不平衡数据分类方法的补充和扩展,值得进一步研究和关注. 主要难点是如何动态感知不平衡数据流中多数类和少数类的分布变化.

4) 基于半监督学习的不平衡数据分类方法.

半监督学习可根据已知的标记样本信息,利用自学习机制从未标记样本中学习新的标记样本,实现对无监督学习和有监督学习的扩展. 将半监督学习的思想引入不平衡数据分类方法中,在半监督学习的框架下,利用未标记样本对只从已标记样本得到的假设进行调整,弥补标记样本的不足,有效地将大量的未标记数据用于不平衡数据的分类,进一步提高不平衡数据分类方法在少量标签数据集上的分类表现,是不平衡数据分类中值得研究的又一重要方向. 主要难点是在考虑数据的不平衡特性的情况下,如何选取最有价值的无标签数据,有效提升不平衡数据分类问题的性能.

## 4 结 论

不平衡数据分类问题在机器学习领域具有重要的研究意义和应用价值. 随着稀疏表示、深度学习等机器学习相关技术的迅速发展,不平衡数据分类问题也出现了一系列新的研究成果. 本文从数据预处理层面、特征层面以及分类算法层面对现有的不平衡数据分类方法进行了梳理和分析,同时结合机器学习的研究热点,讨论了不平衡数据分类问题面临的挑战并给出了相应的解决思路.

## 参考文献(References)


[1] Bhattacharya S, Rajan V, Shrivastava H. ICU mortality prediction: A classification algorithm for imbalanced datasets[C]. Proc of the 31st AAAI Conf on Artificial Intelligence. San Francisco: AAAI, 2017: 1288-1294.

[2] Herndon N, Caragea D. A study of domain adaptation classifiers derived from logistic regression for the task of splice Site prediction[J]. IEEE Trans on Nanobioscience, 2016, 15(2): 75-83.

[3] Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data[J]. Bmc Bioinformatics, 2013, 14(1): 106-122.

[4] Li J, Fong S, Mohammed S, et al. Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms[J]. J of Supercomputing, 2016, 72(10): 3708-3728.

[5] Zakaryazad A, Duman E. A profit-driven artificial neural network (ANN) with applications to fraud detection and direct marketing[J]. Neurocomputing, 2016, 175: 121-131.

[6] Li H, Wong M L. Financial fraud detection by using Grammar-based multi-objective genetic programming with ensemble learning[C]. IEEE Conf on Evolutionary Computation. Sendai: IEEE, 2015: 1113-1120.

[7] Sanz J A, Bernardo D, Herrera F, et al. A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data[J]. Chemical Geology, 2015, 90(4): 973-990.

[8] Lin S J, Chang C, Hsu M F. Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction[J]. Knowledge-Based Systems, 2013, 39: 214–223.

[9] Wang S, Yao X. Using class imbalance learning for software defect prediction[J]. IEEE Trans on Reliability, 2013, 62(2): 434-443.

[10] Zhong W, Raahemi B, Liu J. Classifying peer-to-peer applications using imbalanced concept-adapting very fast decision tree on IP data stream[J]. Peer-to-Peer Networking and Applications, 2013, 6(3): 233-246.

[11] Xiong W, Li B, He L, et al. Collaborative web service QoS prediction on unbalanced data distribution[C]. IEEE Int Conf on Web Services. Anchorage: IEEE, 2014:

377-384.

[12] Drown D J, Khoshgoftaar T M, Seliya N. Evolutionary sampling and software quality modeling of high-assurance systems[J]. IEEE Trans on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2009, 39(5): 1097-1107.

[13] 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.

[14] Dufrenois F. A one-class kernel fisher criterion for outlier detection[J]. IEEE Trans on Neural Networks and Learning Systems, 2015, 26(5): 982-994.

[15] Martin-Diaz I, Morinigo-Sotelo D, Duque-Perez O, et al. Early fault detection in induction motors using adaboost with imbalanced small data and optimized sampling[J]. IEEE Trans on Industry Applications, 2017, 53(3): 3066-3075.

[16] Wang F, Xu T, Tang T, et al. Bilevel feature extraction-based text mining for fault diagnosis of railway systems[J]. IEEE Trans on Intelligent Transportation Systems, 2016, 18(1): 49-58.

[17] Shen W, Wang X, Wang Y, et al. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection[C]. Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3982-3991.

[18] Chen K, Gong S, Xiang T, et al. Cumulative attribute space for age and crowd density estimation[C]. Proc of the 2013 IEEE Conf on Computer Vision and Pattern Recognition. Portland: IEEE, 2013: 2467-2474.

[19] Pouyanfar S, Chen S C. Automatic video event detection for imbalance data using enhanced ensemble deep learning[J]. Int J of Semantic Computing, 2017, 11(1): 85-109.

[20] Daniels Z A, Metaxas D N. Addressing imbalance in multi-Label classification using structured hellinger forests[C]. Proc of the 31st AAAI Conf on Artificial Intelligence. San Francisco: AAAI, 2017: 1826-1832.

[21] Buda M, Maki A, Mazurowski M A. A systematic study of the class imbalance problem in convolutional neural networks[J]. Neural Networks, 2018, 106: 249-259.

[22] Huang C, Li Y, Chen C L, et al. Learning deep representation for imbalanced classification[C]. Proc of IEEE Conf on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 5375-5384.

[23] Wagner C, Saalmann P, Hellingrath B. Machine condition monitoring and fault diagnostics with imbalanced data sets based on the KDD process[J]. IFAC-Papers OnLine, 2016, 49(30): 296-301.

[24] García V, Sanchez J S, Mollineda R A. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance[J]. Knowledge-Based Systems, 2011, 25(1): 13-21.

[25] Galar M, Fernandez A, Barrenechea E, et al. A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and Hybrid-based approaches[J]. IEEE Trans on Systems, Man, and Cybernetics-Part C, 2012, 42(4): 463-484.

[26] Prati R C, Batista G E, Silva D F. Class imbalance revisited: A new experimental setup to assess the performance of treatment methods[J]. Knowledge and Information Systems, 2015, 45(1): 247-270.

[27] Krawczyk B. Learning from imbalanced data: Open challenges and future directions[J]. Progress in Artificial Intelligence, 2016, 5(4): 221-232.

[28] Fernández A, del Río S, Chawla N V, et al. An insight into imbalanced big data classification: Outcomes and challenges[J]. Complex and Intelligent Systems, 2017, 3(2): 105-120.

[29] Haixiang G, Yijing L, Shang J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017, 73: 220-239.

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

[31] 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 Y, Yang B R, Qu W. Survey of mining imbalced datasets[J]. Computer Science, 2010, 37(10): 27-32.)

[32] 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 Z F, Wen Y M, Lv B L. A survey of imbalanced pattern classification problems[J]. Caai Trans on Intelligent Systems, 2009, 4(2): 148-156.)

[33] 林智勇, 郝志峰, 杨晓伟. 不平衡数据分类的研究现状[J]. 计算机应用研究, 2008, 25(2): 332-336.
(Lin Z Y, Han Z F, Yang X W. Current state of research on imbalanced datasets classification learning[J]. Application on Research of Conputers, 2008, 25(2): 332-336.)

[34] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]. Proc of the 2014 IEEE Conf on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 1717-1724.

[35] 方昊, 李云. 基于多次随机欠采样和POSS方法的软件缺陷检测 [J]. 山东大学学报: 工学版, 2017, 47(1): 15-21.
(Fang H, Li Y. Random undersampling and POSS method for software defect prediction[J]. J of Shandong University: Engineering Science, 2017, 47(1): 15-21.)

[36] Lin W C, Tsai C F, Hu Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017, 409: 17-26.

[37] 吴园园, 申立勇. 基于类重叠度欠采样的不平衡模糊多类支持向量机[J]. 中国科学院大学学报, 2018, 35(4): 536-543.
(Wu Y Y, Shen L Y. Imbalanced fuzzy multiclass support

vector machine algorithm based on class-overlap degree undersampling[J]. J of University of Chinese Academy of Sciences, 2018, 35(4): 536-543.)

[38] Ng W W Y, Hu J, Yeung D S, et al. Diversified sensitivity-based undersampling for imbalance classification problems[J]. IEEE Trans on Cybernetics, 2015, 45(11): 2402-2412.

[39] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. J of Artificial Intelligence Research, 2002, 16(1): 321-357.

[40] Zhu T, Lin Y, Liu Y. Synthetic minority oversampling technique for multiclass imbalance problems[J]. Pattern Recognition, 2017, 72: 327-340.

[41] 黄海松, 魏建安, 康佩栋. 基于不平衡数据样本特性的新型过采样SVM分类算法[J]. 控制与决策, 2018, 33(9): 1549-1558.
(Huang H S, Wei J A, Kang P D. A new over-sampling SVM classification algorithm based on unbalanced data sample characteristics[J]. Control and Decision, 2018, 33(9): 1549-1558.)

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

[43] Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on $k$-means and SMOTE[J]. Information Sciences, 2018, 465: 1-20.

[44] Das B, Krishnan N C, Cook D J. RACOG and wRACOG: Two probabilistic oversampling techniques[J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27(1): 222-234.

[45] Moreo A, Esuli A, Sebastiani F. Distributional random oversampling for imbalanced text classification[C]. Proc of the 39th Int ACM SIGIR Conf on Research and Development in Information Retrieval. Pisa: ACM, 2016: 805-808.

[46] Liu S, Zhang J, Xiang Y, et al. Fuzzy-based information decomposition for incomplete and imbalanced data learning[J]. IEEE Trans on Fuzzy Systems, 2017, 25(6): 1476-1490.

[47] Razavi F R, Farajzadeh-Zanjani M, Saif M. An integrated class-imbalanced learning scheme for diagnosing bearing defects in induction Motors[J]. IEEE Trans on Industrial Informatics, 2017, 13(6): 2758-2769.

[48] Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks[J]. Expert Systems with Applications, 2018, 91: 464-471.

[49] Ando S, Huang C Y. Deep over-sampling framework for classifying imbalanced data[C]. Proc of the 2017 Joint European Conf on Machine Learning and Knowledge Discovery in Databases. Skopje: Springer, 2017: 770-785.

[50] 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.

[51] 陶新民, 童智靖, 刘玉. 基于ODR和BSMOTE结合的不均衡数据SVM分类算法[J]. 控制与决策, 2011, 26(10): 1535-1541.
(Tao X M, Tong Z J, Liu Y. SVM classifier for unbalanced data based on combination of ODR and BSMOTE[J]. Control and Decision, 2011, 26(10): 1535-1541.)

[52] Zhang X, Zhu C, Wu H, et al. An imbalance compensation framework for background subtraction [J]. IEEE Trans on Multimedia, 2017, 19(11): 2425-2438.

[53] Li J, Fong S, Wong R K, et al. Adaptive multi-objective swarm fusion for imbalanced data classification[J]. Information Fusion, 2018, 39: 1-24.

[54] 肖鹰, 吴哲夫, 张彤, 等. 一种基于特征选择的不平衡数据分类算法[J]. 集成技术, 2016, 5(1): 68-74.
(Xiao Y, Wu Z F, Zhang T, et al. Feature selection based classifi cation algorithm with imbalanced data[J]. J of Integration Technology, 2016, 5(1): 68-74.)

[55] Hou X, Zhang T, Ji L, et al. Combating highly imbalanced steganalysis with small training samples using feature selection[J]. J of Visual Communication and Image Representation, 2017, 49: 243-256.

[56] Wu Q, Ye Y, Zhang H, et al. ForesTexter: An efficient random forest algorithm for imbalanced text categorization[J]. Knowledge-Based Systems, 2014, 67: 105-116.

[57] Han C, Tan Y K, Zhu J H, et al. Online feature selection of class imbalance via pa algorithm[J]. J of Computer Science and Technology, 2016, 31(4): 673-682.

[58] Yin L, Ge Y, Xiao K, et al. Feature selection for high-dimensional imbalanced data [J]. Neurocomputing, 2013, 105(3): 3-11.

[59] Maldonado S, Weber R, Famili F. Feature selection for high-dimensional class-imbalanced data sets using support vector machines[J]. Information Sciences, 2014, 286: 228-246.

[60] Viegas F, Rocha L, Gonalves M, et al. A genetic programming approach for feature selection in highly dimensional skewed data[J]. Neurocomputing, 2018, 273: 554-569.

[61] Zhou P, Hu X, Li P, et al. Online feature selection for high-dimensional class-imbalanced data[J]. Knowledge-Based Systems, 2017, 136: 187-199.

[62] Liu M, Xu C, Luo Y, et al. Cost-sensitive feature selection by optimizing $F$-measures[J]. IEEE Trans on Image Processing, 2018, 27(3): 1323-1335.

[63] Moayedikia A, Ong K L, Boo Y L, et al. Feature selection for high dimensional imbalanced class data using harmony search[J]. Engineering Applications of Artificial Intelligence, 2017, 57: 38-49.

[64] Ng W W Y, Zeng G, Zhang J, et al. Dual autoencoders features for imbalance classification problem[J]. Pattern Recognition, 2016, 60: 875-889.

[65] Zhang Z L, Luo X G, Garca S. Cost-sensitive back-propagation neural networks with binarization techniques in addressing multi-class problems and non-competent classifiers[J]. Applied Soft Computing, 2017, 56(C): 357-367.

[66] Dhar S, Cherkassky V. Development and evaluation of cost-sensitive universum-SVM[J]. IEEE Trans on Cybernetics, 2017, 45(4): 806-818.

[67] Sahin Y, Bulkan S, Duman E. A cost-sensitive decision tree approach for fraud detection[J]. Expert Systems with Applications, 2013, 40(15): 5916-5923.

[68] Gu X, Chung F L, Ishibuchi H, et al. Imbalanced TSK fuzzy classifier by cross-class Bayesian fuzzy clustering and imbalance learning[J]. IEEE Trans on Systems, Man, and Cybernetics: Systems, 2017, 47(8): 2005-2020.

[69] 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 Vibrationand Shock, 2016, 35(20): 178-182.)

[70] Zhang G, Sun H, Ji Z, et al. Cost-sensitive dictionary learning for face recognition[J]. Pattern Recognition, 2016, 60: 613-629.

[71] 史作婷, 吴迪, 荆晓远, 等. 类不平衡稀疏重构度量学习软件缺陷预测[J]. 计算机技术与发展, 2018, 28(6): 125-128.
(Shi Z T, Wu D, Jin X Y, et al. Prediction of defect of class-imbalance sparse reconstruction metric learning software[J]. Computer Technology and Development, 2018, 28(6): 125-128.)

[72] Chung Y A, Lin H T, Yang S W. Cost-aware pre-training for multiclass cost-sensitive deep learning[C]. Proc of the 25th Int Joint Conf on Artificial Intelligence(IJCAI). New York: IJCAI, 2016: 1411-1417.

[73] Raj V, Magg S, Wermter S. Towards effective classification of imbalanced data with convolutional neural networks[C]. Proc of the 7th IAPR Workshop on Artificial Neural Networks in Pattern Recognition. Ulm: Springer, 2016: 150-162.

[74] Khan S H, Hayat M, Bennamoun M, et al. Cost-sensitive learning of deep Feature representations from imbalanced data[J]. IEEE Trans on Neural Networks and Learning Systems, 2018, 29(8): 3573-3587.

[75] Domingos P. MetaCost: A general method for making classifiers cost-sensitive[C]. Proc of the 5th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. San Diego: ACM, 1999: 155-164.

[76] Chew H, Bogner R, Lim C C. Dual $v$-support vector machine with error rate and training size biasing[C]. Proc of the 2001 IEEE Int Conf on Acoustics, Speech, and Signal. Salt Lake City: IEEE, 2001: 1269-1272.

[77] Zhou Z H, Liu X Y. On multi-class cost-sensitive learning [C]. Proc of the 21st National Conf on Artificial Intelligence. Boston: AAAI, 2006: 567-572.

[78] Maldonado S, Montecinos C. Robust classification of imbalanced data using one-class and two-class SVM-based multiclassifiers[J]. Intelligent Data Analysis, 2014, 18(1): 95-112.

[79] Dufrenois F. A one-class kernel fisher criterion for outlier detection[J]. IEEE Trans on Neural Networks and Learning Systems, 2014, 26(5): 982-994.

[80] Chaki S, Verma A K, Routray A, et al. A one class classifier based framework using SVDD: Application to an imbalanced geological dataset[C]. Proc of the 3rd IEEE Students' Technology Symposium. Kharagpur: IEEE, 2014: 76-81.

[81] Luca S, Clifton D A, Vanrumste B. One-class classification of point patterns of extremes[J]. The J of Machine Learning Research, 2016, 17(1): 6581-6601.

[82] Yin S, Zhu X, Jing C. Fault detection based on a robust one class support vector machine[J]. Neurocomputing, 2014, 145: 263-268.

[83] 韩志艳, 王健. 基于不平衡支持向量数据描述的故障诊断算法[J]. 计算机工程, 2017, 34(5): 156-162.
(Han Z Y, Wang J. Fault diagnosis algorithm based on imbalanced support vector data description[J]. Computer Engineering, 2017, 34(5): 156-162.)

[84] 姚宇, 冯健, 张化光, 等. 一种基于椭球体支持向量描述的异常检测方法[J]. 山东大学学报: 工学版, 2017, 47(5): 195-202.
(Yao Y, Feng J, Zhang H G, et al. Weighted hyper-ellipsoidal support vector data description with negative samples for outlier detection[J]. J of Shandong University: Engineering Science, 2017, 47(5): 195-202.)

[85] Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models[C]. IEEE Symposium on Computational Intelligence and Data Mining. Nashville: IEEE, 2009: 324-331.

[86] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]. Proc of the 7th European Conf on Principles and Practice of Knowledge Discovery in Databases. Cavtat-Dubrovnik: Springer, 2003: 107-119.

[87] Seiffert C, Khoshgoftaar T M, Hulse J V, et al. RUSBoost: A hybrid approach to alleviating class imbalance[J]. IEEE Trans on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2010, 40(1): 185-197.

[88] Liu X Y, Wu J, Zhou Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Trans on Systems, Man, and Cybernetics-Part B, 2009, 39(2): 539-550.

[89] Feng W, Huang W, Ren J. Class imbalance ensemble learning based on the margin theory[J]. Applied Sciences, 2018, 8(5): 815-843.

[90] Sun Z, Song Q, Zhu X, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5): 1623-1637.

[91] 陈圣灵, 沈思淇, 李东升. 基于样本权重更新的不平衡数据集成学习方法[J]. 计算机科学, 2018, 45(7): 31-37.

(Chen S L, Shen S Q, Li D S. Emsemble learning method for imbalanced data based on sample weight updating[J]. Computer Science, 2018, 45(7): 31-37.)

[92] Yuan X, Xie L, Abouelenien M. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data[J]. Pattern Recognition, 2018, 77: 160-172.

[93] Zhao Y, Shrivastava A K, Tsui K L. Imbalanced classification by learning hidden data structure[J]. IIE Trans, 2016, 48(7): 614-628.

[94] Sun Y, Kamel M S, Wong A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12): 3358-3378.

[95] Krawczyk B, Schaefer G, Woniak M. A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification[J]. Artificial Intelligence Medicine, 2015, 65(3): 219-227.

[96] 付忠良. 多标签代价敏感分类集成学习算法[J]. 自动化学报, 2014, 40(6): 1075-1085.
(Fu Z L. Cost-sensitive ensemble learning algorithm for multi-label classication problems[J]. Acta Automatic Sinica, 2014, 40(6): 1075-1085.)

[97] Cao C, Wang Z. IMCStacking: Cost-sensitive stacking learning with feature inverse mapping for imbalanced problems[J]. Knowledge-Based Systems, 2018, 150: 27-37.

[98] Yan J, Han S. Classifying imbalanced data sets by a novel RE-sample and cost-sensitive stacked generalization method[J]. Mathematical Problems in Engineering, 2018, DOI: 10.1155/2018/5036710.

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

[100] Triguero I, Garcia S, Herrera F. SEG-SSC: A framework based on synthetic examples generation for self-labeled semi-supervised classification[J]. IEEE Trans on Cybernetics, 2015, 45(4): 622-634.

[101] Zhang X, Hu B G. A new strategy of cost-free learning in the class imbalance problem[J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(12): 2872-2885.

[102] Yang Z, Shrivastava A K, Tsui K L. Imbalanced classification by learning hidden data structure[J]. Iise Trans, 2016, 48(7): 614-628.

[103] Tan S C, Wang S, Watada J. A self-adaptive class-imbalance TSK neural network with applications to semiconductor defects detection[J]. Information Sciences, 2018, 427: 1-17.

[104] Triguero I, Garcia S, Herrera F. SEG-SSC: A framework based on synthetic examples generation for self-labeled semi-supervised classification[J]. IEEE Trans on Cybernetics, 2015, 45(4): 622-634.

[105] Huang C, Loy C C, Tang X. Discriminative sparse neighbor approximation for imbalanced learning[J]. IEEE Trans on Neural Networks and Learning Systems, 2018, 29(5): 1503-1513.

[106] Rivera W A. Noise reduction a priori synthetic over-sampling for class imbalanced data sets[J]. Information Sciences, 2017, 408: 146-161.

[107] Li J, Fong S, Wong R K, et al. Adaptive multi-objective swarm fusion for imbalanced data classification[J]. Information Fusion, 2018, 39: 1-24.

[108] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(1): 63-77.

[109] Tang M, Yang C, Zhang K, et al. Cost-sensitive support vector machine using randomized dual coordinate descent method for big class-imbalanced data classification[J]. Abstract and Applied Analysis, 2014: 1-9.

[110] Chuang L Y, Yang C H, Wu K C, et al. A hybrid feature selection method for DNA microarray data[J]. Computers in Biology and Medicine, 2011, 41(4): 228-237.

[111] Du J, Vong C M, Pun C M, et al. Post-boosting of classification boundary for imbalanced data using geometric mean[J]. Neural Networks, 2017, 96: 101-114.

[112] Maurya C, Toshniwal D, Venkoparao G. Distributed sparse class-imbalance learning and its applications[J]. IEEE Trans on Big Data, 2018, DOI: 10.1109/TBDATA.2017.268837.

[113] Maurya C K, Toshniwal D, Venkoparao G V. Online sparse class imbalance learning on big data[J]. Neurocomputing, 2016, 216: 250-260.

[114] García V, Sánchez J S, Mollineda R A. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance[J]. Knowledge-Based Systems, 2012, 25(1): 13-21.

[115] Lu J, Tan Y P. Cost-sensitive subspace analysis and extensions for face recognition[J]. IEEE Trans on Information Forensics and Security, 2013, 8(3): 510-519.

[116] Wang S, Minku L L, Yao X. Resampling-based ensemble methods for online class imbalance learning[J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27(5): 1356-1368.

[117] Wang B, Pineau J. Online bagging and boosting for imbalanced data streams[J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(12): 3353-3366.

[118] Huang G, Song S, Gupta J N D, et al. Semi-supervised and unsupervised extreme learning machines[J]. IEEE Trans on Cybernetics, 2014, 44(12): 2405-2417.

[119] Frasca M, Bertoni A, Re M, et al. A neural network algorithm for semi-supervised node label learning from unbalanced data[J]. Neural Networks, 2013, 43: 84-98.

[120] Zhang Q, Sun J, Zhong G, et al. Random multi-graphs: A semi-supervised learning framework for classification of high dimensional data[J]. Image and Vision Computing, 2017, 60: 30-37.

(责任编辑：郑晓蕾)