doi: 10.3969/j.issn.1674-8425(z).2019.07.014

本文引用格式: 刘定祥,乔少杰,张永清,等. 不平衡分类的数据采样方法综述[J]. 重庆理工大学学报(自然科学),2019,33(7):102-112.

Citation format: LIU Dingxiang, QIAO Shaojie, ZHANG Yongqing, et al. A Survey on Data Sampling Methods in Imbalance Classification[J]. Journal of Chongqing University of Technology(Natural Science),2019,33(7):102-112.

不平衡分类的数据采样方法综述

刘定祥[1a],乔少杰[1b],张永清[1c],韩 楠[1d],魏军林[1a],张榕珂[2],黄 萍[1d]

(1. 成都信息工程大学 a. 网络空间安全学院; b. 软件工程学院;
c. 计算机学院; d. 管理学院,成都 610225; 2. 西部战区总医院,成都 610083)

摘 要: 如何获得更加精确的分类效果一直是机器学习领域的重要研究内容,现有大多数分类器都是针对平衡的数据集来设计的。虽然平衡的数据训练出来的分类模型能取得较好的正负样本分类正确率,但现实生活中的数据往往是不平衡的,不平衡的数据使得正样本分类正确率急剧下降,不能满足机器学习对分类效果的要求。针对这种情况,综述了当前主流不平衡分类的数据采样方法。首先,阐述了欠采样方法,包括基于聚类和基于整合的欠采样方法; 其次,对过采样方法进行了总结,包括基于 k 近邻、基于聚类、基于半监督、基于深度神经网络和基于进化算法的过采样方法; 再次,对混合采样方法进行了总结; 最后,总结了不平衡分类问题研究的发展趋势。

关 键 词: 机器学习; 不平衡数据; 过采样; 欠采样; 混合采样

中图分类号: TP311 **文献标识码:** A **文章编号:** 1674-8425(2019)07-0102-11

A Survey on Data Sampling Methods in Imbalance Classification

LIU Dingxiang[1a], QIAO Shaojie[1b], ZHANG Yongqing[1c], HAN Nan[1d],
WEI Junlin[1a], ZHANG Rongke[2], HUANG Ping[1d]

(1. a. School of Cybersecurity; b. School of Software Engineering;
c. School of Computer Science; d. School of Management,
Chengdu University of Information Technology, Chengdu 610225, China;
2. Western General Hospital, Chengdu 610083, China)

Abstract: How to achieve highly accurate results on classification is a fundamental research problem in machine learning. Most of classifiers are designed for balanced dataset. The classifiers trained by the balanced dataset can achieve better classification accuracy of positive and negative samples.

收稿日期: 2019-01-27

基金项目: 国家自然科学基金资助项目(61772091,61802035,61702058); 广西自然科学基金资助项目(2018GXNSFDA138005); 四川省科技计划项目(2018JY0448); 四川高校科研创新团队建设计划项目(18TD0027); 成都市软科学研究项目(2017-RK00-00053-ZF); 成都信息工程大学中青年学术带头人科研基金资助项目(J201701); 成都信息工程大学科研基金资助项目(KYTZ201715,KYTZ201750)

作者简介: 刘定祥,男,硕士,主要从事机器学习研究; 通讯作者 乔少杰,男,博士,教授,主要从事移动数据库、人工智能研究,E-mail: sjqiao@cuit.edu.cn。

However, the real data are always imbalanced. The imbalanced data greatly degrade the classification accuracy of positive samples, which fails to satisfy the growing requirement of classification accuracy in machine learning research. This study surveys the state-of-the-art data sample methods in imbalance classification. Firstly the under-sampling methods including clustering based and integration based methods are introduced. Secondly, the over-sampling methods including k nearest neighbor based, clustering based, semi-supervised based, deep neural networks based and evolutionary based methods are presented, and then hybrid-sampling methods are summarized. Lastly, the future development on the problem of imbalance classification is concluded.

Key words: machine learning; imbalance data; over-sampling; under-sampling; hybrid sampling

1 研究背景

针对现实生活中产生的大量数据,人们通过传感器等数据采集设备将其收集、整理,形成了计算机能够批量处理的数据。通过对数据的学习分析,挖掘潜藏在数据背后深层的知识和规律,可提升人们对外界事物的感知和理解能力[1-2]。然而,现实中这些数据大都比例不平衡。例如,癌症基因检测数据[3]中,在几百万个样本基因里可能仅有一个基因是癌症基因; 电信通讯中只有少数通讯是具有欺诈行为的通讯记录[4-5]; 软件检测中也只有不到10%的软件是具有缺陷的[6]。

不平衡数据普遍存在于人类生活的方方面面,不仅数据分布广泛,而且数据比例不均衡。在不平衡数据中数量多的样本称为负样本,数量少的样本称为正样本。正负样本拥有较大的比例差距,例如: 全国1年中雷电天气(正样本)天数占全年天数的比例不到10%; 新生体检中患肺结核疾病的学生人数占比不到1‰。

在数据分类评价指标中,全局分类正确率是指分类正确的正样本与负样本数量之和除以总的正样本与负样本的数量。正样本分类正确率是指分类正确的正样本数量除以总的正样本数量。同理可得负样本分类正确率。通过上述定义可以知道: 不平衡数据中,由于负样本数量远多于正样本,少数正样本被错分并不会大幅度地降低全局分类正确率,但正样本分类正确率会下降。

机器学习[7]利用训练数据对模型进行训练,使模型能够学习到样本数据特征,实现机器对样本数据的自动分类。精准分类一直是机器学习发展所必需的,但绝大多数流行的分类器是根据平衡数据进行设计,不平衡数据不能够充分训练分类模型,导致分类性能下降[8]。现阶段,机器学习大多通过梯度下降[9]方法训练模型参数,不平衡数据训练分类模型会导致分类模型的参数过多向负样本(majority class)倾斜,从而极大地降低了模型对正样本(minority class)的分类正确率。例如: 对于同一分类方法,利用平衡数据集对分类模型进行训练时,分类模型能够较好地识别正负样本,获得较高的正负样本分类正确率; 利用不平衡数据集对分类模型进行训练时,分类模型对正样本识别能力弱,降低了正样本分类正确率。

为了解决数据不平衡问题,许多数据不平衡处理方法被提出[10]。Anand 等[11]早在1993年就对不平衡数据做了比较深入的研究,发现神经网络反向转播收敛速度慢,其原因是训练集中多数样本均属于同一类。与此同时,Krawczyk 等[12]针对不平衡问题总结了不平衡数据主要应用领域,如表1所示,其中括号内数字表示统计应用的次数。表1充分说明了不平衡数据应用在各个领域,其分布广,使用频率高,是机器学习中普遍存在和亟待解决的问题。

当前,提升分类器在不平衡数据中的学习效果主要采用两种方法:

1) 对不平衡数据分类算法的优化。由于现阶段的分类算法主要是根据平衡数据集进行设计的,故优化不平衡分类算法不仅难度大,且在正样本分类正确率上提升不显著。

2) 对不平衡数据采样算法的优化。采样算

法着手于数据层面，能够有效地解决不平衡数据正负样本分布不平衡的问题，且采样算法的优化设计相对容易，在正样本分类正确率方面性能可得到显著提升。研究不平衡数据采样算法能够有效地提升正样本分类正确率。本文着重从数据采样的角度介绍如何对不平衡数据进行处理。

表 1 不平衡数据的典型应用

应用领域	问题描述
活动识别(19)	罕见活动的检测
行为分析(3)	识别危险的行为
癌症肿瘤分级(30)	分析癌症的严重程度
高光谱数据分析(50)	多维图像中不同区域的分类
工业系统监控(44)	工业机械故障检测
情感分析(65)	文本中的情绪识别
软件缺陷监测(48)	识别代码块中的错误
目标检测(45)	不同频率指定目标出现的分类
文本挖掘(39)	文献关系检测
视频挖掘(20)	识别视频序列中的对象和动作

当前，采样方法主要有以下 3 类：欠采样[13-20]、过采样[21-40]、混合采样[41-51]，这三类方法都有各自的优缺点。

1) 欠采样方法指筛选一些具有代表性的负样本，使负样本和正样本达到比例相当，即所谓的“数据平衡”。其优点是训练集达到了平衡，提升了正样本分类正确率，缺点是丢失了大量的负样本特征，模型不能充分地学习到负样本的样本特征，降低了负样本分类正确率，欠采样过程如图 1(a) 所示。

2) 过采样方法是时下比较流行的方法，其工作原理和欠采样相反，目的是将现有的少数正样本通过模型生成新的正样本，使数据集中正负样本达到平衡。其优点是增加了正样本数量和正样本的多样性，提升了模型对正样本的学习量。缺点是生成的正样本不是真正采集获得的正样本，在增加样本数量和多样性的同时带来了样本噪声(样本不具有的特征)。模型学习样本噪声，降低了模型对正样本的分类正确率。过采样思想如图 1(b) 所示。

3) 混合采样是指将欠采样和过采样结合，正样本通过某种样本生成模型生成一部分新的正样本，负样本通过样本筛选模型筛选一部分具有代表性的负样本，达到正负样本数量平衡。混合采样旨在减少负样本的特征丢失，同时减少正样本的噪声生成，达到正负样本数量平衡，混合采样过程如图 1(c) 所示。

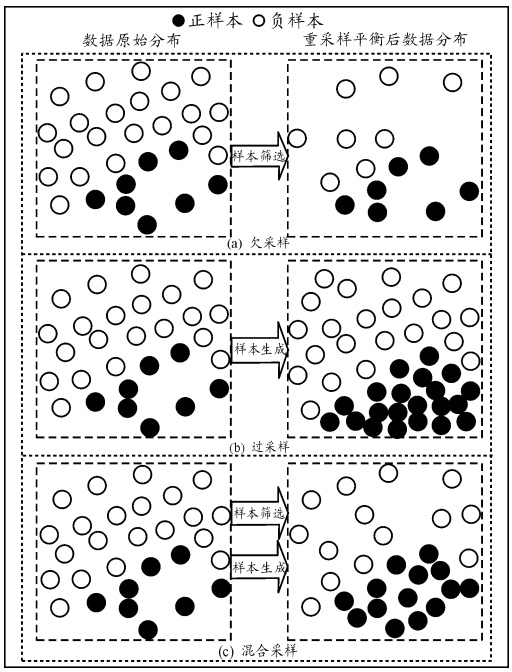

图 1 不同采样算法工作原理

本文通过对当前国内外具有代表性的不平衡分类学习中的采样研究进行统计发现：其 19% 是欠采样的研究，52% 是过采样的研究，29% 是混合采样的研究。单从统计数据上可以发现，不平衡数据采样研究中过采样的研究较多。通过整理，将本文讨论的研究内容汇总如表 2 所示。

2 欠采样方法

解决数据不平衡问题，最简单的欠采样方法是随机欠采样[13]。它通过随机丢弃一部分负样本使正负样本达到平衡。但这种做法具有很大的缺陷，因为随机丢失了大量的负样本特征，随机欠采样不能大幅度地提升模型的正样本分类正确率。

欠采样主要分为两类: ① 基于聚类的欠采样(clustering based under-sampling): 通过对负样本进行聚类,并在每一个类中选取具有代表性特征的样本作为负样本训练集; ② 通过整合的思想,将负样本分成很多份,利用每一份负样本和唯一一份正样本对多个分类器进行训练,最后对多个结果进行集成。

表 2 典型采样方法及采样策略

种类	重采样策略	文献出处
欠采样	基于聚类的思想	Yen 等(2009),Varassin 等(2013),Ng 等(2017)
	基于整合思想	Liu 等(2009),Tahir 等(2012),Zhang 等(2018)
过采样	基于 K 邻近思想	Chawla 等(2002),Han 等(2005),Bunkhumpornpat 等(2009)
	基于聚类的思想	Sanchez 等(2013),Nekooeimehr 等(2016)
	基于半监督的思想	Dong 等(2016),Ebo 等(2017)
	基于深度神经网络	Konno 等(2018)
	基于进化算法思想	Maleki 等(2017),Lim 等(2016)
	其他思想	Ramentol 等(2016),Pang 等(2013),Barua 等(2014)
混合采样	随机欠采样 + 随机过采样	Seiffert 等(2008)
	SMOTE + Kmeans	戴翔等(2015)
	Borderline + Random	冯宏伟等(2017)
	SMOTE + 聚类 + majority voting	Parchuabsupakij 等(2018),Cao 等(2014)

2.1 基于聚类的欠采样方法

为了解决欠采样的随机性问题,Yen 等[14-15]将负样本进行聚类,选取有代表性的样本作为训练集,以尽可能地提取具有代表性的负样本特征,减少负样本的特征丢失,优化训练效果,在提升对正样本识别率的同时减少负样本的错分率。虽然通过聚类使得训练集包含了更加全面的特征,但依然无法避免样本特征丢失的缺陷。Ng 等[16]认为样本的分布信息有助于代表性样本的选取,通过对负样本进行聚类获取其分布信息,选取每一个类中具有代表性的样本,计算样本的敏感度,再根据敏感度选取 k 个负样本和 k 个正样本,将这 $2k$ 个样本作为训练集。Varassin 等[17]将欠采样的方法运用到 DNA 剪切位点的预测中,通过对负样本进行聚类,选取距聚类中心最近的样本作为代表性的负样本。

2.2 基于整合的欠采样方法

Liu 等[18]针对欠采样提出了一种将负样本划分为多份的思想对模型进行训练,然后对结果进行集成,其基本思想为: 随机将负样本分成和正样本数量相当的若干份,然后对每一份负样本和仅有的一份正样本进行训练,这样可以训练出若干个模型,再将每一个模型的分类结果进行集成得到最终结果。由于考虑了所有负样本的特征,所以应用这一方法可以有效地提升正样本与负样本的分类正确率,其算法流程如下所示:

算法 1 简单集成欠采样算法
输入: 所有正样本 P,所有负样本 N,$|p| < |N|$,数据集个数 T
输出: 分类预测结果
1. 选择合适的模型 X_i,将 N 划分为 T 个子数据集$\{N_1, N_2, N_3, \cdots, N_T\}$;
2. For $i = 1, 2, 3, \cdots, T$
利用 P 和 N_i 对模型 X_i 进行训练,得到结果 R_i;
EndFor
3. 将 T 个结果$\{R_1, R_2, R_3, \cdots, R_T\}$进行集成,得到最终结果;

该方法考虑了所有负样本的特征,能够获得较高的正样本与负样本分类正确率,但是对于一些处于样本边界的数据,并不能有效地提升分类性能。因为在没有强化学习边界样本的情况下,分类器大概率会出现错分的情况。Zhang 等[19]提出了一种反向随机欠采样的方法,其思想是将负样本分成比正样本少的若干份,将每一份负样本和正样本作为训练集,然后对多个结果进行集成得到最终结果。由于欠采样后每一个训练集中负样本比正样本少,所以称为反向欠采样。实验结果表明: 反向欠采样具有一定的有效性与可靠性。

Tahir 等[20]针对上述问题提出了一种寻找正样本和负样本边界的欠采样方法。首先将负样本进行反向欠采样,产生若干个负样本少于正样本的训练集;然后寻找每一个训练集中正负样本的边界,将这些边界进行拟合,得到最终的样本边界,进而通过边界对样本进行分类。实验结果表明:此方法在二分类和多分类问题上均取得了较高的正样本分类正确率。

3 过采样方法

过采样方法主要指通过数学模型或者方法合成正样本。由于合成样本的方法是人为设定的,使得生成的正样本会包含一些原正样本不具有的特征,即噪声数据。该特征被分类器学习进而造成分类正确率下降。如何使生成的正样本具有丰富多样的正样本特征,且使得正样本均匀分布在样本空间是过采样方法研究的关键与核心[21]。已有过采样方法较多,不同方法具有各自的优缺点。最简单的过采样方法是随机过采样[22],其思想是随机复制正样本,单纯地使得正负样本比例达到相对平衡。虽然模型对正样本的分类正确率有一定的提高,但其最大的缺点在于生成的正样本与初始正样本一样,不具有多样性,并不能大幅度地提升正样本的分类正确率。

3.1 基于 K 邻近的过采样方法

为了解决随机过采样的局限性,提升样本的多样性,Chawla 等[23]提出了 SMOTE 方法。SMOTE 算法的主要过程如下所示:

算法 2 基于 K 邻近的 SMOTE 过采样算法

输入:原始样本数据集 N,采样比率 P

输出:新的平衡数据集 N_{new}

1. 对于每一个正样本 X_i,计算 X_i 到正样本集合 N 中所有样本的欧式距离,得到其 K 邻近;

2. 根据采样比率计算生成的正样本数量,从其 K 邻近中随机选择相应数量的邻近配对;

3. 对每个配对的样本(X_i, X_n)按照如下公式生成新的正样本,直到达到采样比率:

$$X_{new} = X_i + \text{rand}(0,1) * |X_i - X_n|;$$

SMOTE 方法是最具代表性的过采样方法,其基本思想是在每一个正样本和其 K 邻近的样本之间随机地生成一个新的样本。由于生成的样本是两个样本之间的随机值,所以该方法解决了样本多样性的问题。Han 等[24]分析了 SMOTE 方法的不足,提出了 Borderline SMOTE 方法,认为在正样本的边界区域样本容易被分类器错分,因而要强化边界区域数据的训练。算法思想:找到正样本的边界区域,对处于边界区域的正样本采用 SMOTE 方法进行样本生成。由于增加了边界样本的数据量,强化了边界样本的学习,正样本分类正确率相比 SMOTE 方法有一定的提升。但 SMOTE 和 Borderline SMOTE 方法均通过寻找 K 邻近生成样本,在选取 K 邻近样本时,均未考虑存在选取到负样本的情况。Bunkhumpornpat 等[25]提出了 Self-Level-SMOTE 方法,通过计算 K 邻近附近正样本的权重来生成正样本,避免了生成样本跨越正样本边界的问题。

3.2 基于聚类的过采样方法

基于聚类的过采样方法思想是:为了将具有相同特征的正样本聚在一起,在每一个类中通过样本生成的方法生成样本。由于对正样本进行了聚类,所以处于每一个类中的样本都会有新样本生成,避免了生成样本过于集中在某一个类中,使得生成的正样本能够均匀地分布在正样本的样本空间,提升了正样本分类正确率。Sanchez 等[26]将正样本进行聚类,根据需要生成的样本数量在每一个类中单独进行过采样。由于聚类后在同一个类中的样本具有相似的属性,新生成的样本不会跨越类边界,同时减少了样本生成的盲目性,提升了采样的效果。Nekooeimehr 等[27]利用层次聚类对正样本进行聚合,在每一个层内部进行过采样,并对边界样本进行识别,对处于边界的样本不进行过采样,避免了生成样本跨越边界的问题。

3.3 基于半监督学习的过采样方法

不平衡数据集中有一些数据集不平衡比例较大,正样本数量非常少,通过简单地样本生成不能有效地生成具有多样性的样本。Dong 等[28]利用半监督的方法解决正样本过于稀少的问题,不断

把新生成的样本合并到原来的样本中进行下一轮迭代,从而达到正负样本平衡。Ebo 等[6] 认为: 无论是 SMOTE 方法,还是通过高斯分布或者基于特征生成的模型都是基于 K 邻近生成的,K 邻近生成的新样本会跨越正样本的边界,如图 2 所示。

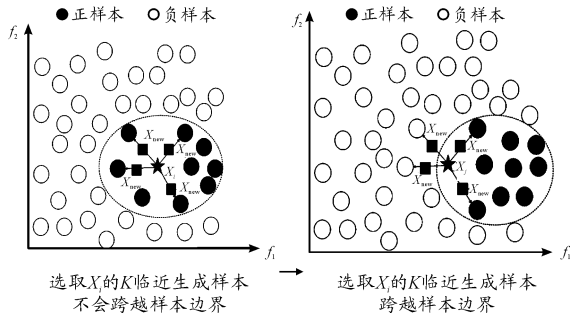

图 2 基于 K 邻近($K=4$) 的样本生成示意图

通过图 2 可以发现: 正样本(用实心五角星表示) 被选取时,基于 K 邻近方法生成的样本有部分跨越了样本边界(虚线以外) 。Ebo 等[6] 提出了一种染色体遗传理论的过采样方法 MAHAKIL,通过计算每一个正样本和正样本中心的马氏距离,按距离大小排序,将距离较大的一半和距离较小的一半分别作为父亲样本和母亲样本进行交配,生成新的样本,然后利用半监督的思想迭代生成需要的样本数量,具体过程如算法 3 所示。

算法 3 基于多样性的 MAHAKIL 过采样算法
输入: 原始数据集 N,生成比例 P
输出: 和生成比例 P 相当的数据集
1. 将数据集 N 分成正样本 $N_{\min}$ 和负样本 N_{maj},获取正样本的个数 $NN_{\min}$ 和负样本的个数 NN_{maj};
2. 根据生成比例 P 得到样本生成后总的正样本数量 $N_G = NN_{\mathrm{maj}} * P$;
3. while $NN_{\min} < N_G$
4. 计算每一个正样本和正样本均值中心的马氏距离;
5. 对所有正样本按照上述距离大小进行排序,距离较大的一半为父亲样本 SF,距离较小的一半为母亲样本 SM;
6. For $i = 1,2,3,\cdots,NN_{\min}/2$
$N_{\mathrm{new}} = \mathrm{average}(SF_i, SM_i)$;
EndFor
得到合成的新样本 N_{new} 及其数量 NN_{new};
7. 将新样本合并到原始样本中: $N_{\min} = N_{\min} + N_{\mathrm{new}}$;
更新其个数: $NN_{\min} = NN_{\min} + NN_{\mathrm{new}}$;

MAHAKIL 方法按照半监督的原理生成样本,在增加样本多样性的同时不会降低样本的有效性。算法实现简单,生成的新样本在多数数据集上不会跨越样本边界,且生成的样本分布均匀,在计算成本、时间成本、采样效果方面均取得较好的效果。由于 MAHAKIL 方法按照距离样本中心的马氏距离进行聚类,导致该方法在数据小析取项[29-31] 的问题上存在缺陷。小析取项又称为类内不平衡问题,是指正样本分布并不都在一个连续的样本空间,可能分布在两个或者多个不连续的样本空间,其结构如图 3 所示。

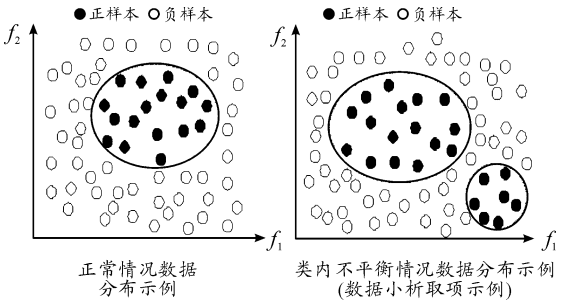

图 3 正样本类内分布不平衡的情况(数据小析取项)

3.4 基于深度神经网络的过采样方法

Konno 等[32] 将深度神经网络技术应用于过采样中,其思想是: 通过深度神经网络(DNN) 提取正样本特征作为样本基本特征,在基本特征上加入一部分伪特征(pseudo feature) 产生新的样本。该方法的特点在于伪特征的加入能增加样本的多样性。通过深度神经网络能够有效地提取样本特征,具有较好的普适性,但存在一些不足,例如: 伪特征中仍会产生许多噪声,特征提取过程中会有部分样本特征丢失。算法思想如图 4 所示。

3.5 基于进化算法的过采样方法

样本生成过程中减少噪声是影响过采样性能的关键因素之一。为了减少噪声,同时提升生成样本的多样性,学者们提出了一些基于进化算法的过采样方法[33-35] 。进化算法的基本原理是通过选择、交叉、变异等操作在问题空间寻找最优解,其主要步骤包括: 首先,选择合适的正样本分别作为父亲类和母亲类; 然后,父亲类样本和母亲

类样本进行交叉生成新的样本；最后，在新样本上进行变异操作，增加样本的多样性。当前较新的进化算法是 Lim 等[36]提出的基于进化理论的过采样算法 ECO-Ensemble。该方法通过优化正样本中的类内和类间的样本生成比例，使得生成的样本具有多样性和均匀分布的特性。

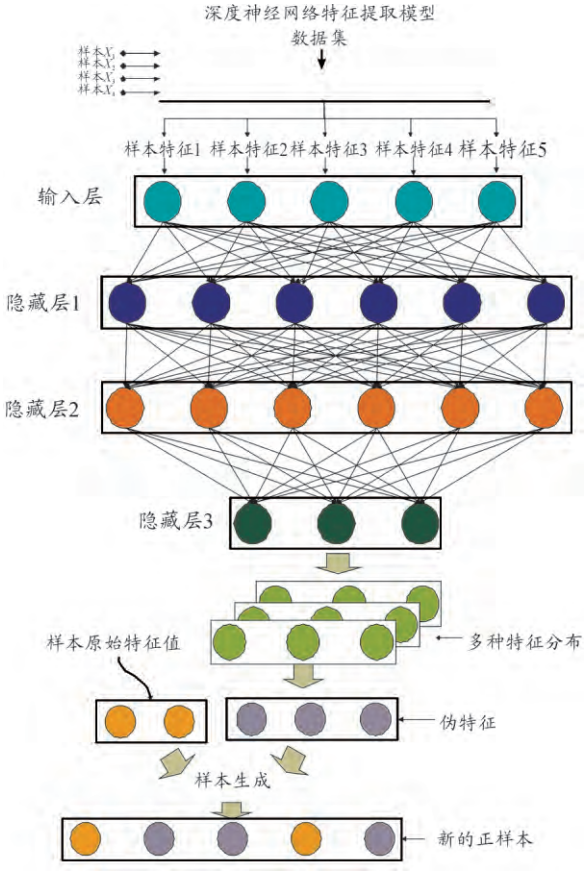

图 4 基于深度神经网络的过采样模型

3.6 其他过采样方法

目前，将过采样方法应用于机器学习技术的研究日益普及。Ramentol 等[37]将模糊粗糙集的编辑技术应用于过采样中，取得了较好的正样本分类正确率。Pang 等[38]利用不平衡时间序列和稀疏混合的高斯模型对正样本进行过采样，降低了过采样的随机性。Moreo 等[39]认为要提取样本的分布特征，根据样本分布特征生成新的正样本，使生成的正样本具有合理的分布，其缺点在于：不同的数据具有不同的分布特征，基于特征分布来生成新样本的方法不具有普适性。Barua 等[40]提出了一种根据样本权重生成新样本的过采样方法 MWMOTE。首先，算法识别一些分类器比较难识别的正样本；然后计算这些正样本和最近负样本的欧式距离，根据距离大小赋予正样本相应的权重，依照权重对正样本进行聚类；最后在每一个类中应用 SMOTE 方法对样本进行过采样。该方法提升了比较难识别（权重较大）的样本的学习效果。为了解决数据不平衡的问题，已有研究虽然取得了一些进展，提升了不平衡数据分类正确率，但仍存在诸多不足，典型过采样研究方法简介如表 3 所示。

表 3 典型过采样研究方法简介

过采样方法	算法特点
随机过采样	平衡了正负样本训练集，产生的样本不具有多样性
SMOTE 过采样方法	生成样本具有多样性、但生成样本有可能跨越边界
Borderline SMOTE	强化了边界样本的学习
基于聚类的过采样方法	生成样本更符合样本类内的分布
基于深度神经网络特征提取的过采样方法	具有很好的普适性，样本特征容易丢失，产生噪声较多
基于半监督思想的过采样方法	解决样本稀少问题，但新样本噪声较多
基于样本分布特征的过采样方法	样本生成效果好，不具有普适性
基于进化理论的过采样方法	样本生成效果好，实现困难，代价大

4 混合采样方法

混合采样是将过采样方法与欠采样方法结合以达到平衡正负样本的采样方法，其主要从以下两个方面提升正样本与负样本分类正确率[41]：① 不会造成大量的负样本特征丢失，模型能学习到足够多的负样本特征；② 不会产生过多的噪声，模型学习到的噪声少。

为了验证混合采样的性能，Seiffert 等[42]将随机过采样和随机欠采样技术结合，通过实验验证了混合采样技术能够显著提升决策树的正样本分

类正确率。戴翔等[43]综合过采样与欠采样的优点,将 SMOTE 算法运用于少数类样本的生成,利用 K-means 聚类对负样本进行欠采样,提升了正样本与负样本的分类正确率。Li 等[44]将混合采样技术运用于支持向量机 SVM 中,并利用 K 邻近方法对混合采样的结果做进一步约减,解决了数据混淆的问题,提高了支持向量机的泛化性能。Cervantes 等[45]利用欠采样和支持向量机得到初始 SVs 和超平面,将这些实例作为遗传算法的初始种群。原始数据集包含生成和演化的数据,通过学习达到最小化不平衡数据的目的。该方法提高了支持向量机在不平衡数据集上的泛化能力。高锋等[46]提出一种基于邻域特征的混合采样技术,根据样本领域分布特征赋予采样权重,利用局部置信度的动态集成方法对不同的数据选择不同的分类器,并将不同分类器结果集成。实验结果表明,在查全率和查准率上该混合采样技术都有较大的提升。冯宏伟等[47]认为,位于边界区域的样本是最容易错分的样本,于是针对边界样本进行 SMOTE 过采样以强化边界样本的学习,然后针对负样本进行随机欠采样。该方法的正样本与负样本分类正确率较经典的采样方法有较大的提升。Gazzah 等[48]提出的方法不是单纯地进行过采样和欠采样,进行过采样时重点考虑具有代表性的正样本,进行欠采样时丢弃相关性较小的负样本。Cao 等[49]认为正负样本比例大时,单纯应用混合采样的效果不理想,将混合采样和集成的思想结合能够有效地提升模型正样本与负样本的分类正确率。基于集成的混合采样方法工作原理如图 5 所示。

算法基本思想: 首先,将正样本进行一次过采样,并随机地将负样本欠采样成与正样本相当的若干份; 然后,将每一份负样本和正样本进行混合,形成多个训练集,得到多个训练好的分类器; 最后,将不同分类器的结果进行集成,得到最终结果。模型通过该方法能够充分学习负样本特征,是目前比较流行的混合采样方法[50-51]。

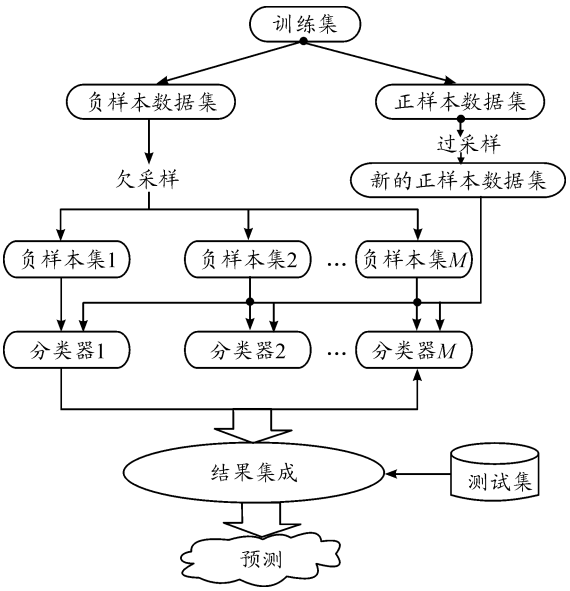

图 5 基于集成的混合采样方法工作原理

5 结束语

不平衡数据分类问题是当前机器学习领域比较热门的研究内容,已经吸引越来越多学术界和工业界专家对其进行广泛和深入的研究。本文详述了不平衡分类问题中采样方法的研究现状和发展趋势,介绍了欠采样、过采样和混合采样 3 大类采样方法原理和典型算法。应用这些方法可以提高不平衡数据分类的正确率。

为了进一步研究更高效稳定的不平衡学习方法,未来可以从以下几个方面展开研究:

1) 在样本信息获取中,样本信息获取不完善是导致不平衡分类学习性能下降的最根本原因。对样本的物理、化学属性进行分析,以便多角度、多方位获取更多的样本属性,提升正负样本的区分度,达到提升不平衡分类学习性能的目的。

2) 在样本聚类中,通过单一的距离指标进行聚类不能全面地衡量样本间的距离,应结合多个距离指标进行聚类,并引入普适高效的聚类方法提升聚类效果。

3) 在过采样特征提取中,需要研究多层次的样本特征提取模型,强化样本特征的提取,生成噪声量少、多样性丰富的新样本,提升过采样的有

效性。

4) 在强调边界的采样方法中,需要研究有效的边界寻找方法,结合多个评价指标对样本边界进行拟合,并对样本进行降噪处理,提升样本边界的有效性。

5) 基于进化算法的过采样中引入更多参数,优化生成样本的分布,提升生成样本的多样性与有效性。

参考文献:

[1] COTE D. Using machine learning in communication networks [J]. IEEE Journal of Optical Communications and Networking,2018,10(10):D100 - D109.

[2] JORDAN M I, MITCHELL T M. Machine learning: Trends, perspectives, and prospects [J]. Science, 2015, 349(6245):255 - 260.

[3] YU H,NI J,DAN Y,et al. Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets [J]. Tsinghua Science and Technology,2012,17(6):666 - 673.

[4] OLSZEWSKI D. A probabilistic approach to fraud detection in telecommunications [J]. Knowledge Based Systems,2012,26:246 - 258.

[5] LIMA R F,PEREIRA A C M. A fraud detection model based on feature selection and under-sampling applied to web payment systems [C]//Proceedings of the 2015 IEEE International Conference on Web Intelligence and Intelligent Agent Technology. Piscataway, NJ: IEEE, 2016:219 - 222.

[6] EBO B K,KEUNG J,PHANNACHITTA P,et al. MAHAKIL: diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction [J]. IEEE Transactions on Software Engineering,2017,44(6):534 - 550.

[7] MURPHY K P. Machine learning: a probabilistic perspective [J]. Chance,2012,27(2):62 - 63.

[8] STEFANOWSKI J. Dealing with data difficulty factors while learning from imbalanced data [M]. Berlin German: Springer,2016:333 - 363.

[9] LI J,ZHOU T. On gradient descent algorithm for generalized phase retrieval problem [C]//Proceedings of the 2017 IEEE International Conference on Signal Processing. Piscataway,NJ: IEEE,2017:320 - 325.

[10] JAPKOWICZ N,STEPHEN S. The class imbalance problem: a systematic study [J]. Intelligent Data Analysis, 2002,6(5):429 - 449.

[11] ANAND R,MEHROTRA K G,MOHAN C K,et al. An improved algorithm for neural network classification of imbalanced training sets [J]. IEEE Transactions on Neural Networks,1993,4(6):962 - 969.

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

[13] PRUSA J,KHOSHGOFTAAR T M,DITTMAN D J,et al. Using random under-sampling to alleviate class imbalance on tweet sentiment data [C]//Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration. Piscataway,NJ: IEEE,2015:197 - 202.

[14] YEN S,LEE Y. Cluster based under-sampling approaches for imbalanced data distributions [J]. Expert Systems with Applications,2009,36(3):5718 - 5727.

[15] YEN S J,LEE Y S. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset [J]. Lecture Notes in Control and Information Sciences,2006,344(2):731 - 740.

[16] Ng W W,Hu J,Yeung D S,et al. Diversified sensitivity based under-sampling for imbalance classification problems [J]. IEEE Transactions on Cybernetics,2017,45(11):2402 - 2412.

[17] VARASSIN C G,PLASTINO A,LEITAO H C D G,et al. Under-sampling strategy based on clustering to improve the performance of splice site classification in human genes [C]//Proceedings of the 24th International Workshop on Database and Expert Systems Applications. Piscataway,NJ: IEEE,2013:85 - 89.

[18] LIU X Y,WU J,ZHOU Z H. Exploratory under-sampling for class-imbalance learning [J]. IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics), 2009,39(2):539 - 550.

[19] ZHANG Y,LIU G,LUAN W,et al. An approach to class imbalance problem based on stacking and inverse random

under sampling methods [C] //Proceedings of the 2018 IEEE 15th International Conference on Networking, Sensing and Control. Piscataway, NJ: IEEE, 2018.

[20] TAHIR M A, KITTLER J, YAN F. Inverse random under sampling for class imbalance problem and its application to multi-label classification [J]. Pattern Recognition, 2012, 45(10): 3738 - 3750.

[21] CAO H, LI X L, WOON Y K, et al. Integrated oversampling for imbalanced time series classification [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(12): 2809 - 2822.

[22] ESTABROOKS A, JO T, JAPKOWICZ N. A multiple resampling method for learning from imbalanced data sets [J]. Computational Intelligence, 2010, 20(1): 18 - 36.

[23] 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.

[24] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [C] //Proceedings of the 2005 International Conference on Advances in Intelligent Computing. Berlin, Germany: Springer, 2005: 878 - 887.

[25] BUNKHUMPORNPAT C, SINAPIROMSARAN K, LURSINSAP C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem [C] //Proceedings of the 2009 Pacific-Asia conference on knowledge discovery and data mining. Berlin, Germany: Springer, 2009: 475 - 482.

[26] SANCHEZ, ATLANTIDA I, Morales E F, et al. Synthetic oversampling of insistences using clustering [J]. International Journal on Artificial Intelligence Tools, 2013, 22(2): 475 - 482.

[27] NEKOOEIMEHR I, LAIYUEN S K. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets [J]. Expert Systems with Applications, 2016, 46: 405 - 416.

[28] DONG A, CHUNG F, WANG S. Semi-supervised classification method through oversampling and common hidden space [J]. Information Sciences, 2016, 349: 216 - 228.

[29] 石凤兴. 针对类内不平衡样本分类方法的研究 [D]. 哈尔滨: 哈尔滨工业大学, 2016.

[30] 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.

[31] JAPKOWICZ N, STEPHEN S. The class imbalance problem: a systematic study [J]. Intelligent data analysis, 2002, 6(5): 429 - 449.

[32] KONNO T, IWAZUME M. Pseudo-Feature generation for imbalanced data analysis in deep learning [P]. arXiv: 1807.06538, 2018.

[33] PREZGODOY M D, FERNNDEZ A, RIVERA A J, et al. Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets [J]. Pattern Recognition Letters, 2010, 31(15): 2375 - 2388.

[34] CAO P, LI B, LI W, et al. Imbalanced data learning based on particle swarm optimization [J]. Journal of Computer Applications, 2013, 33(3): 789 - 792.

[35] GAO M, HONG X, CHEN S, et al. On combination of SMOTE and particle swarm optimization based radial basis function classifier for imbalanced problems [C] //Proceedings of the 2011 International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2011: 1146 - 1153.

[36] LIM P, GOH C K, TAN K C. Evolutionary cluster based synthetic over-sampling ensemble (ECO-Ensemble) for imbalance learning [J]. IEEE Transactions on Cybernetics, 2016, 47(9): 2850 - 2861.

[37] RAMENTOL E, GONDRES I, LAJES S, et al. Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance: the SMOTE-FRST-2T algorithm [J]. Engineering Applications of Artificial Intelligence, 2016, 48: 134 - 139.

[38] PANG J Z F, CAO H, TAN V Y F. MOGT: oversampling with a parsimonious mixture of gaussian trees model for imbalanced time-series classification [C] //Proceedings of the 2013 IEEE International Workshop on Machine Learning for Signal Processing. Piscataway, NJ: IEEE, 2013.

[39] MOREO A, ESULI A, SEBASTIANI F. Distributional random oversampling for imbalanced text classification [C] //Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information

Retrieval. New York: ACM,2016: 805 - 808.

[40] BARUA S,ISLAM M M,YAO X,et al. MWMOTE: majority weighted minority oversampling technique for imbalanced data set learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26 (2): 405 - 425.

[41] 欧阳源遊. 基于混合采样的非平衡数据集分类研究 [D]. 重庆: 重庆大学,2014.

[42] SEIFFERT C,KHOSHGOFTAAR T M,VAN H J. Hybrid sampling for imbalanced data [J]. Integrated Computer-Aided Engineering,2009,16(3): 193 - 210.

[43] 戴翔,毛宇光. 基于集成混合采样的软件缺陷预测研究 [J]. 计算机工程与科学,2015,37(5): 930 - 936.

[44] LI P,QIAO P L,LIU Y C. A hybrid re-sampling method for SVM learning from imbalanced data sets [C] //Proceedings of the 2018 International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway,NJ: IEEE, 2008: 65 - 69.

[45] CERVANTES J,HUANG D S,FARID G L,et al. A hybrid algorithm to improve the accuracy of support vector machines on skewed data sets [M]. Berlin German: Springer,2014: 782 - 788.

[46] 高锋,黄海燕. 基于邻域混合抽样和动态集成的不平衡数据分类方法 [J]. 计算机科学,2017,44(8): 225 - 229.

[47] 冯宏伟,姚博,高原,等. 基于边界混合采样的非均衡数据处理算法 [J]. 控制与决策,2017,32(10): 1831 - 1836.

[48] GAZZAH S,HECHKEL A,AMARA N E B. A hybrid sampling method for imbalanced data [C] //Proceedings of the 2015 International Multi-conference on Systems. Piscataway,NJ: IEEE,2015.

[49] CAO P,ZHAO D,ZAIANE O. Hybrid probabilistic sampling with random subspace for imbalanced data learning [J]. Intelligent Data Analysis, 2014, 18 (6): 1089 - 1108.

[50] PRACHUABSUPAKIJ W. A new hybrid sampling classification for imbalanced data [C] //Proceedings of the 2015 International Joint Conference on Computer Science and Software Engineering. Piscataway, NJ: IEEE, 2015: 281 - 286.

[51] CAO P, YANG J, LI W, et al. Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD [J]. Computerized Medical Imaging and Graphics the Official Journal of the Computerized Medical Imaging Society,2014,38(3): 137 - 150.

(责任编辑 杨黎丽)