

# 非均衡数据分类经典方法综述与面向医疗领域的实验分析



江昊琛<sup>1</sup> 魏子麒<sup>1</sup> 刘 璘<sup>1</sup> 陈 俊<sup>2</sup>

<sup>1</sup> 清华大学软件学院 北京信息科学与技术国家研究中心 北京 100084

<sup>2</sup> 百度公司 北京 100094

(jhc18@mails.tsinghua.edu.cn)

**摘 要** 近年来,人工智能技术被广泛地应用于多个领域。其中,智慧医疗场景得到了普遍关注,并产生了大量临床辅助诊断和医疗方案推荐的实际应用。然而,由于人工智能技术的本质在于通过从大量真实数据中进行模式抽取,从而预测未知情况,因此真实数据的数据特征和数据质量将直接影响人工智能应用的效果。相比其他智能应用领域,由于罕见病患者在人群中总是占极少数,医疗数据具有天然的非均衡的特点,而高度非均衡的数据在机器学习领域被认为是难于学习的。针对这一应用现状,文中首先围绕“数据非均衡”问题开展了文献调研,尝试通过寻找该问题的通用解决办法来指导在智慧医疗环境下的应用。之后,以数据挖掘领域的会议 SIGKDD(ACM SIGKDD Conference on Knowledge Discovery and Data Mining)近年来涉及非均衡数据集的工作为分析样本,统计针对特定领域的“数据非均衡”问题人们倾向选择的处理方法。最后,通过医学数据分析中的两个典型应用场景,对调研获得的知识和方法进行实验应用,从而验证了调研和统计分析中所得出方法的可用性。

**关键词:** 数据分析;智慧医疗;非均衡数据集;过采样

**中图法分类号** TP311.5

## Imbalanced Data Classification: A Survey and Experiments in Medical Domain

JIANG Hao-chen<sup>1</sup>, WEI Zi-qi<sup>1</sup>, LIU Lin<sup>1</sup> and CHEN Jun<sup>2</sup>

<sup>1</sup> School of Software, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

<sup>2</sup> Baidu Corp., Beijing 100094, China

**Abstract** In recent years, AI technology has been widely adopted in many application domains, amongst which, intelligent medical applications such as clinical decision support systems have attracted much attention. However, since the current wave of AI applications are based on predictive models crystalized from historical data, the feature and quality of data will affect AI applications' performance directly. Medical data are inherently imbalanced as rare disease cases are always the scarce in existing case archives, while considered more important. The “data imbalance problem” is still considered a difficult research problem in machine learning. This paper conducts a literature review on the research efforts targeting at techniques to handle “imbalanced data” in general as well as the ones in intelligent medical area. We then use research publications from the SIGKDD conference dedicated to knowledge discovery and data mining as a sample pool, to find people's preferred approach to address “imbalanced data” problem in a given domain. Finally, based on approaches, we identify from the survey, and conduct experiments on two typical medical predictive model learning scenarios, to validate the know-how we acquired in this study.

**Keywords** Data analysis, Intelligent medical, Imbalanced datasets, Over-sampling

### 1 引言

在数据挖掘和人工智能领域,有许多针对数据实例分布非均衡情况下的分类问题的相关研究<sup>[1-2]</sup>。医疗场景是其中十分典型的一个领域。疾病辅助诊断通过对由病历和治疗记录组成的临床数据集进行分析建模,尝试获取一个高精度和高鲁棒性的分类模型,输出对患者可能患有的疾病类型的预测结果,进而辅助医生进行疾病诊断,或帮助患者根据观察到

的症状进行自诊。

然而,医疗相关的分类模型常常出现对常见疾病的诊断准确率较高但对并发症或罕见病的诊断结果较差的情况。当疾病类型极其罕见时,模型甚至可能在学习过程中主动将其忽略并给出无关结果。而医疗领域的数据集往往都具有非常明显的非均衡特点,如医疗领域最为经典的数据集 MIMIC-III<sup>[3]</sup>,其最大类的疾病诊断有两万余条记录,而同时对于许多罕见病,其疾病诊断记录只有几十条,甚至是几条。这是在处

到稿日期:2021-02-20 返修日期:2021-05-20 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:百度-清华大学 AI 医疗联合科研项目

This work was supported by the Baidu-Tsinghua Collaborated Medical AI Project.

通信作者:刘璘(linliu@tsinghua.edu.cn)

理“非均衡”训练数据集时机器学习模型会遇到的普遍问题。在医学领域,不同疾病的训练样本数量仅取决于该疾病的发生率和累积病例,因此无法从根本上杜绝数据集内各类样本分布不均衡的情况。基于这一状况,本文首先进行了文献调研,以便全面地总结基于非均衡数据集的算法训练策略。

对于经典的机器学习算法,通过训练数据获得的模型能够在测试集获得好的效果的基本假设是不同类数据实例满足“独立同分布”。然而,在真实世界的数据集中,这一假设几乎是不成立的。大类的实例往往占据了数据集的大部分,而小类则仅包含极少数实例的“长尾”现象十分常见。数据分布的“非均衡”问题成为了模型学习的难点之一。采用基于原始的数据分布直接进行模型训练,得到的分类模型将很难对来自较小类别的实例进行准确预测。在许多应用中,较小类的实例数量虽少,但重要性却很高,这种情况被称为“分类器偏差(classifier bias)”问题。如在医学诊断分类应用中,大部分病例均为常见病,而罕见病的病例总是稀缺的。除此之外,在设备故障检测<sup>[4]</sup>、异常检测<sup>[5]</sup>、垃圾邮件过滤<sup>[6]</sup>、图像识别<sup>[7]</sup>等工业应用中也存在类似的问题场景。因此,难以获得足够的罕见类实例导致了这一问题的出现。

分类器偏差问题是数据挖掘领域的经典问题。传统对数据均衡这一需求的应对方法是从不同的类中选择数量规模均衡的实例,通过人工筛选的方式对训练集主动进行均衡化<sup>[8]</sup>;另一种思路则是直接将原始数据作为输入,保留其固有非均衡的分布,寻求其他方法来获得更准确的分类预测结果<sup>[9]</sup>。由于数据驱动的深度神经网络的机器学习算法取得了许多新技术进展,因此需要重新审视存在分类器偏差的问题需求。本文对该问题重新进行系统性分析调研,并从非均衡的数据集中抽取类特征的工程项目,给出系统的分析思路。

本文将首先描述针对“非均衡数据”分类问题的关键特征及常用评价标准;其次通过文献调研结果给出“非均衡数据”常用的处理方案;再次基于数据挖掘会议 SIGKDD 的文献,给出一个针对非均衡问题的统计调查结果和分析;然后基于该统计结果,针对医学研究领域常用的处理方式开展实验分析;最后给出本文的研究结论并对未来工作进行了展望。

2 “非均衡数据”问题定义及评价指标

在对“非均衡数据”处理方式进行调研之前,首先需要对该问题的关键特征及评价标准进行分析,从而做到有的放矢。

2.1 “非均衡数据”问题及特征描述

对于一般的二分类问题,从数据标签的角度进行分析,数据集根据数据总量和非均衡度的不同,可分为 4 种情况,即标准数据集、非均衡数据集、小数据集和稀有数据集<sup>[10]</sup>,如表 1 所列。当可用样本数量足够且非均衡度高时,数据集被归类为非均衡数据集。但被归类为非均衡数据集的数据并不一定存在“非均衡数据”问题。为了更好地描述数据非均衡问题,还需要引入样本重要性这一特征指标。当小类样本的重要性远低于大类样本时,小类样本可被看作噪声,在处理时被直接忽略;只有当小类样本的重要性类似于或大于大类样本时,数据集才被看作存在“非均衡数据”问题,且亟需解决。

表 1 按不同非均衡程度及数据集规模的数据集分类

Table 1 Definition of data sets with different sample quantities and

imbalance level

| 非均衡程度 | 数据集规模  |       |
|-------|--------|-------|
|       | 大      | 小     |
| 低     | 标准数据集  | 小数据集  |
| 高     | 非均衡数据集 | 稀有数据集 |

由此可见,对于“非均衡数据”问题,描述数据集中数量较大类的样本数量与数量较小类的样本数量之比的“非均衡度”是非常重要的数据集特征评价标准。在实际应用中,该变量的取值范围会对分类器的预测结果产生影响。

为了评估非均衡度的实际取值范围及其影响,本文调研了数据挖掘及机器学习领域常用的公开数据集 KEEL<sup>[11]</sup>。KEEL 是一个包含了大量真实数据集及其标准分类结果的数据集仓库,可供研究人员进行分析比较。KEEL 中存在一个专为“非均衡数据集”研究而提供的小类,它包含 6 个小型数据集仓库,分别为 4 个无噪声二分类样例和 2 个多分类样例。包含二分类样例的 4 个数据集仓库分别存储了 22,22,22 和 34 个真实数据集,第一个仓库所包含数据集的非均衡度为 1.5~9,其余仓库中非均衡度均高于 9,在 100 个数据集中,最高的非均衡度高达 129.44。

尽管算法处理非均衡程度较高的数据集的难度更高,但在现实情况中仍需要具体问题具体分析。例如,对于类边界清晰的数据集,即使非均衡度很高,仍可以很容易地通过某些基本分类器(如支持向量机(Support Vector Machine,SVM)等)进行分类。也就是说,单纯使用非均衡率只能描述问题的表现形式,而无法解释分类困难的根本原因。另一方面,较高的非均衡度还会进一步导致“小样本”“类重叠”和“小类互斥”<sup>[12]</sup>等问题。当“小样本”和“类重叠”两种情况同时出现时,分类器将由于训练样本过少和训练样本不够明确导致分类器出现偏差。“小类互斥”描述的是小类为大类的某种子类的情况。在医疗分诊问题中,某些罕见疾病是大类疾病的一个分支类型,但该疾病并不能简单地描述为大类疾病,这一情况即为“小类互斥”。以糖尿病为例,该疾病本身为常见病,但重度脂肪肝糖尿病则相对罕见,从糖尿病人为主类的诊断结果中区分出患“重度脂肪肝糖尿病”的病入的难度较高,会导致分类器偏差。

2.2 “数据均衡”的评估指标体系

由于多分类问题可以被拆解为若干个二分类问题进行处理,为了简化分析,本文将二分类问题作为代表研究非均衡分类问题。当学习模型使用某种方法通过建模来对包含大类样本 A 和小类样本 B 的数据集中的元素进行分类时,常用的评判标准为其分类的“准确率”(Accuracy)。然而,在数据非均衡的情况下,若使用准确率评价分类结果,即使分类器对小数量数据的数据集 B 完全分类错误,只要它能够正确评价大类中的所有实例,准确率的值仍然会处于一个较高的水平。即只要分类器将所有实例均按大类的属性标注,使用准确率作为参照的评价体系便会认为该分类器的效果较好。同时,由于在数据集中大类样本属于主要样本,在实际的训练过程中,可以更广泛地获取数据特征,因此对于大类的识别准确率往往较高,而对于小类的识别精度则较低。如前文所言,此时用传统的评价体系虽可以得到较好的结果,但显然这一训练结果并



后者则对不同的少数类样本给予不同权重,合成不同个数的新样本。相对而言,欠采样方法实现更加容易,但由于减少了用于学习的实例,其效果也难以保证。过采样方法为分类器的学习提供了更多的样本,使分类器可以对两种分类进行相对平衡的特征学习,从而获取更好的结果;但由于过采样中部分数据为系统生成而非自然存在,也存在使分类器学习到错误样本的可能。综合而言,大量实验表明,好的过采样方法相比欠采样效果更好。

基于算法的处理方法是一种比较特殊的针对数据非均衡问题的处理方法。广义地说,常用的分类器,如贝叶斯分类器、支持向量机、决策树等均为算法层面的分类方法,但这些通用分类器并非专为非均衡数据问题而设计,故并未列入表2。表2中列出的常见的该类方法仅有单边学习一种,它是专为非均衡程度极高或少数类实例难以获取的情况而设计的一种分类算法。例如工程机械的异常检测问题,当机器正常运转时,研究者对故障时的机器情况一无所知。在这种情况下,单边学习是唯一可行的解决方案。常用的单边学习方法有高斯混合模型(Gaussian Mixed Model, GMM)<sup>[18]</sup>和支持向量数据描述(Support Vector Data Description, SVDD)<sup>[19]</sup>。以高斯混合模型为例,由于极少出现故障情况,则只选择正常情况下的数据,使用该方法通过高斯模型尝试完美拟合所有正常情况,若出现的情况在该模型可拟合的实例之外,则可将其划分为故障类。

代价敏感学习方法尝试使用参数从数据和方法层面对问题进行归一化处理。该方法维护了一个代价敏感矩阵,通过人工确定或某些先验学习方法,该矩阵为两类数据分别分配了相对权重值,这些权重值使分类器在学习时可以给予少数类更多的关注,进而强化对少数类数据特征的提取。常见的对基础代价敏感学习方法<sup>[20]</sup>的扩展有CS-SVM<sup>[9]</sup>和FSVM-CIL<sup>[21]</sup>。CS-SVM算法直接修改了SVM的优化方式,对两种不同类的样本误差分别给予不同的惩罚;FSVM-CIL方法使用模糊加权思想,对所有训练样本都分配代价值,希望以此来区分所有样本的权重值差,进而得到更好的SVM分类模型。代价敏感学习一方面受益于算法设计者的理解力,从而建立一个较好的初始代价敏感矩阵;另一方面又可以参考分类结果,自动优化代价敏感矩阵,这就使其兼具了人工经验和机器学习两方面的优势。随着近年来机器学习研究的发展,代价敏感学习的方法也受到了广泛关注。

集成学习方法在21世纪的第一个十年中获得了广泛的关注<sup>[22-23]</sup>。该方法通过对分类器的学习结果进行分析,来修改分类器的数据输入,令其再次学习以提高分类效果。常用的集成学习方法有AdaBoost<sup>[24]</sup>方法和基于知识的欠采样的BalanceCascade<sup>[25]</sup>方法。其中,AdaBoost方法是对基分类器和学习样本均分配不同的权重,之后按照分类器对样本学习的错误率更新分类器和样本权重,再根据这两个权重对基分类器进行重复学习,最终获取较好的结果。相比AdaBoost, BalanceCascade方法则更加复杂,需生成多个平衡的训练子集,其中的每个子集按照AdaBoost的方法进行训练,区别在于前者采用的是并行更新方式,而后的基分类器则以串行

的方式生成。由于对数据集进行了多次学习,集成学习方法有可能导致过拟合。但在大多数情况下,该方法的学习能力较强,学习速度也相对较快。

随着深度学习算法的发展和大规模训练数据集的出现,在方法和数据这两方面均增加了对非均衡数据集问题更好的处理方法。在深度学习领域,出现了若干基于神经网络的用于对非均衡数据集进行分类的分类器<sup>[26-27]</sup>,还出现了针对损失函数定义的相关研究方向和成果<sup>[28]</sup>。一些常用的网络结构经过修改后也被用于非均衡数据处理,如生成对抗网络<sup>[29]</sup>被用于生成小样本数据。

迁移学习的发展给非均衡数据集的分类需求提供了新方法。迁移学习可用于设计比集成学习更强的分类器。其主要思想是将分类器从具有足够样本的数据集所习得的知识迁移到样本数量较小的相似问题中,后者的分类器通过这种方式获取了大量的先验知识,从而使分类效果更好。一种新的迁移学习技术结合了加权多数算法(Weighted Majority Algorithm, WMA)与AdaBoost方法,在扩大训练数据集的同时,均衡了数据集内部分类的数量<sup>[10]</sup>。另一个具有代表性的工作是监督局部权重(Supervised Local Weight, SLW)方法<sup>[30]</sup>,在学习过程中,通过给不同的类分配不同的权重来处理数据非均衡问题。

## 4 SIGKDD 近五年研究文献的统计分析

由第3节可知,针对“非均衡数据”问题已有的处理策略包含基于数据的方法、基于算法的方法、代价敏感方法、集成学习方法以及最新的深度学习模型方法。对于本文所关注的医疗领域相关问题,需要在综合考虑效果及效率两方面的前提下,对上述方法进行取舍。SIGKDD会议作为数据挖掘领域中影响力最大的会议之一,在本领域具有较高的学术地位和权威性。因此,本文针对系列会议SIGKDD中近六年发表的论文,统计分析了其中涉及对非均衡数据集分类问题进行处理的文章,尝试寻找对医疗领域最为适用的非均衡数据处理策略,同时也对其他相关研究领域相同问题的处理方案给出初步的参考意见。

数据均衡问题的关注点在于数据本身,因此可将机器学习算法根据数据的输入和输出的不同进行分类研究。在输入端,一般认为相同研究领域的数据特征具有更多的相似性,因此可根据研究领域进行分类;在输出端,根据分类器的直接输出,学习任务可归类为下列3种之一<sup>[12]</sup>。

(1)名词分类预测:算法直接输出类标签(Nominal class predictions)。

(2)数值评分预测:算法对实例给出一个数值预测评分(Numerical scoring predictions)。

(3)概率预测:算法输出为某实例属于某一个类别的概率(Probabilistic predictions)。

根据这一分类方式,结合上节提到的问题,本文对文章的标题、摘要和关键字使用关键词“unbalanced/imbalanced/skewed”进行检索,回顾了SIGKDD会议从2015年至2020年发表的所有论文,剔除仅涉及问题讨论、未提出具体解决方案

的论文,共搜索到 24 篇<sup>[31-54]</sup>相关论文。一些论文中提到了数据非均衡性的问题,并采用了某种现有的方法对其进行处理;另一些则关注该问题本身,尝试提出新的处理思路。根据论文的发表年份、目标任务、研究领域和所采用或提出的基本方

法,本文对这 24 篇论文使用表 2 提出的分类法进行了分类,结果如表 3 所列。表 3 第一行中的“Spl”和“Alg”分别表示基于数据和基于算法的方法,“CSL”和“En”分别是代价敏感学习和集成学习的缩写,“DL”则为深度学习模型算法。

表 3 2015—2019 年 SIGKDD 发表论文的情况  
Table 3 Study of KDD publication in 2015—2019

| 年度   | 目标任务   | 研究领域     | 方法                                       | Spl. | Alg. | CSL | En. | DL |
|------|--------|----------|------------------------------------------|------|------|-----|-----|----|
| 2015 | 数值评分预测 | 大数据流     | Kappa/Harmonic                           | —    | ✓    | ✓   | —   | —  |
| 2015 | 数值评分预测 | 社会安全     | SMOTE                                    | ✓    | —    | —   | —   | —  |
| 2016 | 名词类预测  | 社会安全     | Undersampling                            | ✓    | —    | —   | —   | —  |
| 2016 | 名词类预测  | 商品分类     | DeepCN                                   | —    | —    | —   | —   | ✓  |
| 2016 | 名词类预测  | 医疗文本     | Self-Conf/KFF/Balance-EE                 | ✓    | —    | —   | ✓   | ✓  |
| 2016 | 概率预测   | 社会安全     | Normalization                            | ✓    | —    | —   | —   | —  |
| 2016 | 名词类预测  | 通用       | Asymmetric Query Model                   | —    | —    | ✓   | —   | —  |
| 2017 | 名词类预测  | 通用       | Cascaded classification                  | —    | ✓    | —   | —   | —  |
| 2017 | 名词类预测  | 医药相关     | Undersampling                            | ✓    | —    | —   | —   | —  |
| 2018 | 数值评分预测 | 医学诊断     | Random Undersampling                     | ✓    | —    | —   | —   | —  |
| 2018 | 名词类预测  | 青光眼诊断    | PBR/CNNs                                 | —    | —    | —   | —   | ✓  |
| 2018 | 名词类预测  | 医学诊断     | Undersampling                            | ✓    | —    | —   | —   | —  |
| 2018 | 数值评分预测 | 通用       | CIMAP(Oversampling)                      | ✓    | —    | —   | —   | —  |
| 2018 | 概率预测   |          |                                          |      |      |     |     |    |
| 2018 | 名词类预测  | 通用       | OA3                                      | —    | ✓    | ✓   | —   | —  |
| 2018 | 名词类预测  | 通用       | SPARC                                    | —    | —    | —   | —   | ✓  |
| 2019 | 名词类预测  | 通用       | Alternative Loss Function                | —    | —    | ✓   | —   | ✓  |
| 2019 | 概率预测   | 商业广告转化   | Multi-task Learning                      | —    | —    | —   | —   | ✓  |
| 2019 | 名词类预测  | 健康食物记录   | Focal loss                               | —    | —    | —   | —   | ✓  |
| 2019 | 名词类预测  | 天气预报     | Undersampling                            | ✓    | —    | —   | —   | —  |
| 2019 | 概率预测   | 运动水平评估   | Undersampling                            | ✓    | —    | —   | —   | —  |
| 2019 | 名词类预测  | 工作-简历评估  | Randomly Undersampling                   | ✓    | —    | —   | —   | —  |
| 2019 | 数值评分预测 |          |                                          |      |      |     |     |    |
| 2020 | 名词类预测  | 商业金融 NLP | Filtering                                | ✓    | —    | —   | —   | —  |
| 2020 | 名词类预测  | 云服务任务分类  | Partitioning & Undersampling             | ✓    | —    | —   | —   | —  |
| 2020 | 名词类预测  |          |                                          |      |      |     |     |    |
| 2020 | 数值评分预测 | 通用       | TDR(Targeted Data-driven Regularization) | —    | —    | —   | —   | ✓  |
| 2020 | 概率预测   |          |                                          |      |      |     |     |    |
| 总数   |        |          |                                          | 13   | 3    | 4   | 1   | 8  |

在 20 篇单目标任务论文中,有 3 篇目标任务为概率预测的文章分别选用了基于数据和深度学习的方法;目标任务为数值分析预测的 3 篇文章则分别选用了基于数据、基于算法和代价敏感学习这 3 种方法;其余 14 篇目标任务为名词类预测的文章选用的方法则涵盖了这 5 种。可见,由于可选择的方法较多,以训练目标为分类方法难以找出对于同一目标任务更趋向于使用同种处理方法的规律性。

若从研究领域进行分析,则可找到比较明显的问题处理规律。在 24 篇文章中,除研究领域为通用技术的 16 篇外,有关医疗/健康数据分析研究的有 6 篇,商业领域有 4 篇文章以及社会公共领域有 3 篇。剩余的 3 篇文章分别研究了天气、竞技运动和云服务任务分类的相关课题。

在医疗健康领域的 6 篇论文中,4 篇选择了数据采样处理,另外 2 篇选择了深度学习框架;而社会公共领域的 3 篇文章则均采用了数据采样的非均衡数据处理方式;在商业领域,采用数据采样方式进行处理后,通过深度学习进行分类和选择直接使用深度学习框架的文章各有 2 篇。其余 3 篇文章则均选取了基于数据采样的处理方式。由此可见,针对特定的研究领域,人们选择的处理方法存在一定的规律。

除了不同领域的处理方法存在规律性外,在明确应用领域的 16 篇文章中,有 12 篇采用了数据采样的处理方式进行数据平衡预处理,占总数的 75%。这也意味着,虽然针对数据非均衡问题有许多不同的处理方法,但最直观且最容易实现的是基于数据的方法,因此其也是最常使用的方法。另外,在标明研究领域的文章中,医疗数据处理相关的文章数量最多,且大部分选择了基于数据的处理方法。基于这些发现可以得到以下结论:

- (1)若面对的是经典机器学习的应用场景,例如,在社会安全领域,可选择基于数据采样的方式对数据集的内部分布进行均衡化,从而提升评估目标指标。
- (2)如果基于数据采样方法的效果不够理想,则可考虑选择基于深度学习网络的分类模型,从而提高分类器在最终评价指标上的表现。
- (3)若上述两种方法均无法达到让人满意的效果,则可考虑使用数据采样方法来平衡数据分布,同时使用深度学习模型进行分类,从而提升分类器的效果。
- (4)若研究内容不限于常见的机器学习应用需求,且侧重于理论与技术研究,则需要广泛比较多种已有的方法,从而针

对不同数据集和评价指标选择更有针对性的处理方法。

基于这些发现,对于本文所关注的医疗相关任务,首选方法应是采用基于采样的方法对其进行处理。第 5 节将采用医疗数据对这一结论进行实验论证。

5 基于数据层面过采样方法的非均衡数据集处理效果的实验验证

针对第 4 节得到的结论,本节选用治疗方案预测与分诊两个实际应用对其进行验证,以考察数据采样类方法在医疗领域对分类器结果的影响。方案预测与分诊均是医疗数据挖掘领域的经典问题,前者的目标是帮助院方优化治疗方案,从而降低预测在院死亡率。后者则是在患者进入医院后的首要需求。本节选取 3 个基于 SMOTE 的方法用于对比,分别为 SMOTE,Borderline SMOTE 以及 SMOTE-Tomek。3 种算法的区别在于,前两者只进行了单轮过采样,而 SMOTE-Tomek 则是在过采样后按照 Tomek 原则剔除了冗余项,随后进行迭代式过采样,直到达到预期的非均衡程度。

5.1 治疗方案与患者的在院死亡率预测

对于治疗方案预测与患者在院死亡率预测的分类测试任务,本文选取 MIMIC-III<sup>[3]</sup>数据集作为分析对象,该数据集包含了参与数据收集的医院从 2001 年至 2012 年中 4 万多例 ICU 病人的就诊记录。由于其中的部分疾病为罕见病,同时不同的患者的情况差别较大,因此该数据集本身就具有数据分布高度非均衡的特点。对于用于比较的基础方法,本文采用了发表于 2017 年和 2018 年的未对非均衡数据进行均衡化预处理的两篇 SIGKDD 会议文章中的算法<sup>[55-56]</sup>作为基础方法,并分别对其添加了针对非均衡的数据处理技术,用于评估均衡数据集能否对结果产生影响。

5.1.1 特征选择和数据预处理

本文选择的两种基础算法采用了患者的全部基本信息以及就诊时所作检查产生的所有时序信息作为训练特征。然而,由于需要对数据进行过采样,本文最终选择了以下特征对学习系统进行训练:年龄、性别、血压、血氧浓度、心跳、呼吸速率、吸入氧气浓度、血 pH、血糖、体温、排泄情况等。

在训练前,需要进行数据清洗。首先剔除了年龄小于 18 岁的未成年患者。未成年人的治疗方案与成年人的治疗方案差别较大,可能会对模型的判别过程造成扰动。根据数据完整性的需求,本文在实验中遵循原论文的方式,剔除了数据缺项数 10 个及以上的实例。完成数据清洗后,采用 K 近邻(K-Nearest Neighbor,KNN)算法对实例中的缺失数据项进行人工补全。尽管实验目标是评估采样算法对非均衡数据集的处理结果的影响,但过于罕见的病例仍然可以被视作噪音。针对这一问题,本文选择了诊断数量前 2 000 的疾病来进行实验,此时数据集仍然保留了较高的非均衡度。对经过以上处理的数据集进行随机分割,将其 80%作为训练集,10%作为测试集,剩余的 10%作为验证集。实验只对训练集的数据进行均衡化处理。根据初始非均衡度所在的区间,还对采样后的非均衡度进行了调整,具体如下表 4 所列。

表 4 不同非均衡度数据集在抽样后的目标值

Table 4 Target value of data sets with different imbalanced

| ratio after sampling |            |
|----------------------|------------|
| 初始非均衡度               | 人工采样后的非均衡度 |
| $\leq 100$           | 1          |
| 100~1 000            | 0.2        |
| $\geq 1\,000$        | 0.04       |

5.1.2 评估标准与结果分析

本实验直接采用原论文中的模型来预测死亡率,选择了预测患者在院死亡率以及 Jaccard 相似度这两个指标进行评测。预测在院死亡率的降低是评估治疗方案是否为最优的关键指标,而 Jaccard 相似度则体现了模型给出的治疗方案与医生给出的治疗方案的相似程度<sup>[55]</sup>。如表 5 所列,采用 SMOTE 的算法在两方面的评价结果均有小幅提升。其中,预测在院死亡率降低了 2%,Jaccard 相似度小幅提高。

表 5 基于 MIMIC-III 数据的不同方法效果的比较

Table 5 Performance comparsion on MIMIC-III for different

| methods                  | 预计死亡率 | Jaccard 相似度 |
|--------------------------|-------|-------------|
| LEAP                     | 0.246 | 0.351       |
| LEAP+SMOTE               | 0.229 | 0.356       |
| LEAP+Borderline SMOTE    | 0.226 | 0.363       |
| LEAP+SMOTE-Tomek         | 0.231 | 0.358       |
| SRL-RNN                  | 0.180 | 0.411       |
| SRL-RNN+SMOTE            | 0.169 | 0.417       |
| SRL-RNN+Borderline SMOTE | 0.164 | 0.424       |
| SRL-RNN+SMOTE-Tomek      | 0.167 | 0.422       |

对于非均衡度大于 1 000 的样本,经过过采样处理后得到了最佳的预测结果。但由于只选择诊断数量前 2 000 的疾病的限制,非均衡度达到如此高值的样本总量较少。大量的样本处于原始非均衡度为 100 到 1 000 的区间中,经过处理后,这些样本的评估结果获得了提高。小类样本预测效果的提高是评估结果小幅提升的主要原因。然而,对于大类样本,如高血压,采样算法引发了扰动,这也是 Jaccard 相似度的提升相对较小的原因。

在 3 种 SMOTE 算法中,Borderline SMOTE 在本任务中的表现比其他两种算法有微弱优势。这是由于在本实验中,许多小类样本聚集于边界附近,而 Borderline SMOTE 增加了边界附近的样本数量,使得学习系统有更多的样本可学习,从而提高了预测效果。

5.2 分诊任务

分诊指通过患者的症状,医护人员根据经验辅助患者选择科室的过程。该选择基于医护人员的专业知识,普通患者往往并不具备。该过程是一个典型的分类问题,许多研究者尝试通过设计机器学习模型来将其自动化。本实验数据来源于一家三级甲等口腔专科医院的诊疗数据集。该数据集包含了 2 万余条患者的就诊记录以及就诊科室,这些数据都已进行了脱敏的处理。就诊科室包括急诊科、儿科、牙体牙髓科、正畸科、修复科、种植科、关节科、牙周科、颌面外科、黏膜科。

5.2.1 数据预处理和特征抽取

在数据预处理阶段,首先需要删除不适合用于机器学习的科室数据。由于分类器的训练是基于患者主诉的文本内

容,而儿科患者的科室选择过程是基于患者年龄,急诊科的选择是基于患者的就诊时间,因此在本实验中删除这两个科室的数据。

在本问题中,对于需要分诊的 8 个科室,根据样本数量可将其分为 3 类:颌面外科和牙体牙髓科拥有最多的样本数量,修复科、牙周科、黏膜科以及种植科的样本数量中等,而正畸科与关节科的样本数最少。上述 3 类科室的样本数量比约为 8:4:1。

本实验选取通用预训练模型 BERT<sup>[57]</sup> 对每一段主诉生成其对应的文本特征向量。由于 BERT 本身的预训练模型的能力较为强大,为防止过拟合,实验中选择了分类能力较差的通用基础分类器——支持向量机对问题进行预测,输入为文本特征向量及其对应的科室标签。该分类器并未对非均衡数据集分类问题进行特别优化,因此可以更好地用于评价 SMOTE 方法的处理结果。与 5.1 节相同,实验采用了 8:1:1 的数据分割比例来进行验证。

5.2.2 结果分析

实验最终的评估标准为各科室的分诊准确率,由图 2 可知,经过过采样处理后,除 Borderline SMOTE 方法外,整体准确率均得到了提升。

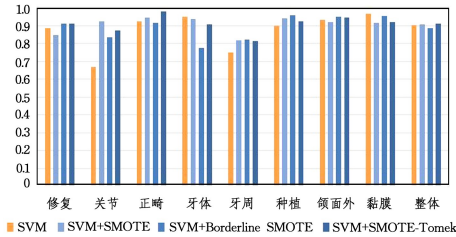


图 2 不同方法下分诊问题的效果比较

Fig. 2 Performance comparison on different departments for different methods

对于两个小类样本,最终结果却表现出了不同的特点。对正畸科来说,不论是否进行过采样,其准确率一直较高。这是因为其特征词与其他科室区别明显,因此采样类算法对其没有特别大的影响。而对于关节科来说,由于其特征词与其他部分科室较为相近,因此在基本算法中,其准确率较低。在均衡化之后,被错分类的样本明显减少。

另一个值得关注的现象是,Borderline SMOTE 实际上降低了最终的整体准确率,原因是该方法导致牙体科的准确率降低。由于该科室的数据量较大,因此拉低了整体准确率。牙体科准确率下降是由于其特征词与其他科室较相似,而采用 Borderline SMOTE 算法后,其边界与其他科室混淆,最终导致准确率下降。

**结束语** 本文从医疗领域的非均衡数据的分类训练问题出发,对非均衡数据集的处理方法进行了系统性的调研,内容包括其定义、常用评价标准及已有的处理方案。之后,通过对数据挖掘领域主流会议 SIGKDD 中发表的论文进行分析,从统计的角度分析了常用处理方案与数据领域间的关系。基于这些分析,得出以下结论:

(1)对于经典机器学习应用,如在社会安全领域,可选择

基于数据采样的方式对数据集的内部分布进行均衡化,从而提升评估目标指标。

(2)如果基于数据采样方法的效果不够理想,则可考虑选择基于深度学习网络的分类模型,从而提升分类器在最终评价指标上的表现。

(3)若上述的两种方法均无法达到让人满意的效果,则可考虑使用数据采样方法平衡数据分布,同时使用深度学习模型进行分类,从而提升分类器的效果。

(4)针对通用理论与技术研究,需要广泛比较多种已有的方法,从而针对不同数据集和评价指标选择更有针对性的处理方法。

最终基于医疗数据的分类预测任务,验证了过采样方法的综合效果和可用性。未来工作包括继续针对医疗大数据应用中的非均衡分类问题进行进一步的研究,并验证是否存在合适的算法层面的处理方案。

参 考 文 献

[1] MAIMON O,ROKACH L. Introduction to Knowledge Discovery and Data Mining[J/OL]. Springer US. [https://link.springer.com/chapter/10.1007/978-0-387-09823-4\\_1](https://link.springer.com/chapter/10.1007/978-0-387-09823-4_1).

[2] SHEARERC. The CRISP-DM model:the new blueprint for data mining[J]. Journal of Data Warehousing,2000,5(4):13-22.

[3] JOHNSON A E W,POLLARD T J,SHEN L,et al. MIMIC-III, a freely accessible critical care database [J]. Scientific Data, 2016,3(1):1-9.

[4] HAO W,LIU F. Imbalanced Data Fault Diagnosis Based on an Evolutionary Online Sequential Extreme Learning Machine[J]. Symmetry,2020,12(8):1204.

[5] WANG G,YANG J,LI R. Imbalanced SVM-Based Anomaly Detection Algorithm for Imbalanced Training Datasets [J]. ETRI Journal,2017,39(5):621-631.

[6] ZHAO C,XIN Y,LI X,et al. A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data[J]. Applied Sciences,2020,10(3):936.

[7] SUN Y,WONG A K,KAMEL M S. Classification of imbalanced data:A review[J]. International Journal of Pattern Recognition and Artificial Intelligence,2009,23(4):687-719.

[8] TOMKEK I. Two modifications of CNN[J]. IEEE Transactions on Systems,Man,and Cybernetics,1976,6:769-772.

[9] VEROPOULOS K,CAMPBELL C,CRISTIANINI N. Controlling the sensitivity of support vector machines [C]// Proceedings of the International Joint Conference on AI. 1999.

[10] AI-STOUHI S,REDDY C K. Transfer learning for class imbalance problems with inadequate data[J]. Knowledge and Information Systems,2016,48(1):201-228.

[11] ALCALA-FDEZ J,FERNANDEZ A,LUENGO J,et al. KEEL Data-Mining Software Tool:Data Set Repository[J]. Journal of Multiple-Valued Logic and Soft Computing,2011,17(2/3):255-287.

[12] FERNÁNDEZ A,GARCÍA S,GALAR M,et al. Learning from imbalanced data sets [M]. Berlin:Springer,2018.

[13] GU Q,YUAN L,NING B,et al. A Novel Classification Algo-

- rithm for Imbalanced Datasets Based on Hybrid Resampling Strategy[J]. *Computer Engineering & Science*, 2012, 34(10): 128-134.
- [14] WILSON D. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data[J]. *IEEE Transactions on Systems, Man, and Cybernetics*, 1972, 2(3): 408-421.
- [15] 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: 321-357.
- [16] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [C]// *International Conference on Intelligent Computing*. 2005: 878-887.
- [17] HE H, BAI Y, GARCIA E, et al. Adaptive synthetic sampling approach for imbalanced learning [C]// *IEEE International Joint Conference on Neural Networks*. 2008.
- [18] CHRISTOPHER M. *Neural Networks for Pattern Recognition* [M]. New York: Oxford University Press, 1995.
- [19] TAX D M, DUIN R P. Support vector domain description[J]. *Pattern Recognition Letters*, 1999, 20(11/12/13): 1191-1199.
- [20] ELKAN C. The foundations of cost-sensitive learning [C]// *International Joint Conference on Artificial Intelligence*. 2001: 973-978.
- [21] BATUWITA R, PALADE V. FSVM-CIL: fuzzy support vector machines for class imbalance learning[J]. *IEEE Transactions on Fuzzy Systems*, 2010, 18(3): 558-571.
- [22] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTE-Boost: Improving prediction of the minority class in boosting [C]// *European Conference on Principles of Data Mining and Knowledge Discovery*. 2003: 107-119.
- [23] SAGI O, ROKACH L. Ensemble learning: A survey[J]. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018, 8(5): e1249.
- [24] SUN Y, KAMEL M S, WONG A K, et al. Cost-sensitive boosting for classification of imbalanced data[J]. *Pattern Recognition*, 2007, 40(12): 3358-3378.
- [25] LIU X Y, WU J, ZHOU Z H. Exploratory Undersampling for Class-Imbalance Learning[J]. *IEEE Transactions on Systems Man & Cybernetics Part B*, 2009, 39(2): 539-550.
- [26] HANSON S J, BRUNSWICK S N, KULIKOWSKI C, et al. Concept-Learning in the Absence of Counter-Examples: an Autoassociation-Based Approach to Classification [J/OL]. New Brunswick Rutgers the State of New Jersey, 1999. <http://dl.acm.org/citation.cfm?id=929980>.
- [27] YPMA E, DUIN R. Novelty detection using Self-Organizing Maps[J/OL]. *Proc. of Iconip*, 1998. [https://www.researchgate.net/publication/2722672\\_Novelty\\_detection\\_using\\_Self-Organizing\\_Maps](https://www.researchgate.net/publication/2722672_Novelty_detection_using_Self-Organizing_Maps).
- [28] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection[C]// *IEEE Transactions on Pattern Analysis & Machine Intelligence*. IEEE, 2017: 2999-3007.
- [29] DOUZAS G, BACAO F. Effective data generation for imbalanced learning using conditional generative adversarial networks [J]. *Expert Systems with Application*. 2018, 91(Jan. ): 464-471.
- [30] LIANG G, JING G, NGO H, et al. On handling negative transfer and imbalanced distributions in multiple source transfer learning [J]. *Statistical Analysis & Data Mining*, 2014, 7(4): 254-271.
- [31] ZHANG X, YANG T, SRINIVASAN P. Online Asymmetric Active Learning with Imbalanced Data [C]// *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016: 2055-2064.
- [32] BIFET A, DE FRANCISCI MORALES G, READ J, et al. Efficient online evaluation of big data stream classifiers [C]// *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015: 59-68.
- [33] SHABANI E, ALEALI A, SHAKARIAN P, et al. Early identification of violent criminal gang members [C]// *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015: 2079-2088.
- [34] DU B, LIU C, ZHOU W, et al. Catch me if you can: Detecting pickpocket suspects from large-scale transit records [C]// *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016: 87-96.
- [35] HA J W, PYO H, KIM J. Large-scale item categorization in ecommerce using multiple recurrent neural networks[C]// *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016: 107-115.
- [36] NGUYEN H, PATRICK J. Text Mining in Clinical Domain: Dealing with Noise [C]// *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016: 549-558.
- [37] WANG H, KIFER D, GRAIF C, et al. Crime rate inference with big data [C]// *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016: 635-644.
- [38] DADKHAHI H, MARLIN B M. Learning Tree-Structured Detection Cascades for Heterogeneous Networks of Embedded Devices [C]// *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017: 1773-1781.
- [39] KOPELOV M, ZIMMERMANN A, BONNET P, et al. PrePeP: A Tool for the Identification and Characterization of Pan Assay Interference Compounds [C]// *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2018: 462-471.
- [40] SATO I, NOMURA Y, HANAOKA S, et al. Managing Computer-Assisted Detection System Based on Transfer Learning with Negative Transfer Inhibition [C]// *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2018: 695-704.
- [41] SUGIURA H, KIWAKI T, YOUSEFI S, et al. Estimating Glaucomatous Visual Sensitivity from Retinal Thickness with Pattern-Based Regularization and Visualization [C]// *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2018: 783-792.
- [42] SUN M, TANG F, YI J, et al. Identify susceptible locations in medical records via adversarial attacks on deep predictive models [C]// *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2018: 793-801.

[43] WANG J,ZHANG M L. Towards mitigating the class-imbalance problem for partial label learning [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018;2427-2436.

[44] ZHANG Y,ZHAO P,CAO J,et al. Online adaptive asymmetric active learning for budgeted imbalanced data [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018;2768-2777.

[45] ZHOU D,HE J,YANG H,et al. Sparc:Self-paced network representation for few-shot rare category characterization [C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018;2807-2816.

[46] DING D,ZHANG M,PAN X,et al. Modeling Extreme Events in Time Series Prediction. knowledge discovery and data mining [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019;1114-1122.

[47] KITADA S,IYATOMI H,SEKI Y. Conversion Prediction Using Multitask Conditional Attention Networks to Support the Creation of Effective Ad Creative [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019;2069-2077.

[48] SAHOO D,HAO W,KE S,et al. FoodAI:Food ImageRecognition via Deep Learning for Smart Food Logging [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019;2260-2268.

[49] SCHON C,DITTRICH J,MULLER R. The Error is the Feature;How to Forecast Lightning using a Model Prediction Error [C] // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019; 2979-2988.

[50] SICILIA A,PELECHRINIS K,GOLDSBERRY K. DeepHoops: Evaluating Micro-Actions in Basketball Using Deep Feature Representations of Spatio-Temporal Data [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019;2979-2988.

[51] YAN R,LE R,SONG Y,et al. Interview Choice Reveals Your Preference on the Market; To Improve Job-Resume Matching through Profiling Memories [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019;914-922.

[52] LI H,YANG Q,CAO Y,et a l. Cracking Tabular Presentation

Diversity for Automatic Cross-Checking over Numerical Facts [C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020;2599-2607.

[53] PHAM P,JAIN V,DAUTERMAN L,et al. DeepTriage: Automated Transfer Assistance for Incidents in Cloud Services [C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020;3281-3289.

[54] KAMANI M M,FARHANG S,MAHDAVI M,et al. Targeted Data-driven Regularization for Out-of-Distribution Generalization [C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020;882-891.

[55] WANG L,ZHANG W,HE X,et al. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018;2447-2456.

[56] ZHANG Y,CHEN R,TANG J,et al. LEAP: learning to prescribe effective and safe treatment combinations for multimorbidity [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017; 1315-1324.

[57] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding [OL]. <https://arxiv.org/abs/1810.04805>.



**JIANG Hao-chen**, born in 1996, post-graduate,is a student member of China Computer Federation. His main research interests include big data techniques in health care and imbalanced data analysis.



**LIU Lin**, born in 1973, Ph.D, is a member of China Computer Federation. Her main research interests include software requirements engineering and health data analytics.

(责任编辑:李亚辉)