

【电子与信息科学 / Electronics and Information Science】

基于朴素贝叶斯集成的正类-无标签学习算法

常秀颖¹, 王晓兰², 朱涛^{3,4}, 何芃³, 欧桂良⁴, 何玉林⁴

1) 沧州职业技术学院华为ICT学院, 河北沧州 061001; 2) 沧州职业技术学院信息工程系, 河北沧州 061001; 3) 深圳大学计算机与软件学院, 广东深圳 518060; 4) 人工智能与数字经济广东省实验室(深圳), 广东深圳 518107

摘要: 针对基于集成学习策略处理正类-无标签学习(positive and unlabeled learning, PUL)问题时存在的未能充分考虑正类和无标签样本中噪声对分类器构建影响的缺陷, 提出一种包含类别标注和噪声筛选两个阶段的基于朴素贝叶斯集成的正类-无标签学习(naïve Bayesian ensemble-based positive and unlabeled learning, NBEB-PUL)算法. 为便于对无标签样本标注不确定性进行显式表示, 引入贝叶斯分类器作为基分类器, 并结合AdaBoosting的集成策略, 构建强分类器Ada-NBC计算无标签数据的标注不确定性, 进而基于不确定性较低的可靠正类样本, 形成可靠正类样本集和剩余无标签样本集. 为弱化噪声样本对二类分类器性能的干扰, 使用Ada-NBC分类器对原始数据集进行再分类, 生成基于Ada-NBC预测的正类样本集和负类样本集, 再与类别标注阶段所得的正集和无标签集进行交集筛选, 提取出重合的高置信度样本以形成最终的正负样本集, 并从中筛选出噪声样本. 在23个选自UCI与KEEL的标准数据集上对NBEB-PUL算法的可行性、合理性及有效性进行验证. 结果表明, 随着迭代次数的增加, NBEB-PUL算法在训练过程中呈现出收敛的特性, 且在正类样本比例分别为0.45、0.40、0.35和0.30的情况下, NBEB-PUL算法在分类精度上总体优于S-EM、Biased-SVM、Modified-PUL、PU-LP、LP-PUL和AdaPU等6种先进PUL算法, 取得了最低的精度平均排序值. 研究结果证实了NBEB-PUL是一种高效的能够处理正类-无标签学习问题的优化策略.

关键词: 知识工程; 正类-无标签学习; 标注不确定性; 后验概率; 贝叶斯分类器; 集成学习

中图分类号: TN911.73

文献标志码: A

DOI: 10.3724/SP.J.1249.2026.03338

NBEB-PUL: a naïve Bayesian ensemble-based algorithm for positive and unlabeled learning

CHANG Xiuying¹, WANG Xiaolan², ZHU Tao^{3,4}, HE Peng³, OU Guiliang⁴,
and HE Yulin⁴

1) Huawei ICT Academy, Cangzhou Technical College, Cangzhou 061001, Hebei Province, P. R. China

2) Department of Information Engineering, Cangzhou Technical College, Cangzhou 061001, Hebei Province, P. R. China

3) College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, Guangdong Province, P. R. China

4) Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen), Shenzhen 518107, Guangdong Province, P. R. China

Abstract: Existing ensemble-based positive and unlabeled learning (PUL) methods often fail to adequately account for the impact of noise in both positive and unlabeled samples during classifier construction. To address this issue, this paper proposes a naïve Bayesian ensemble-based PUL algorithm (NBEB-PUL), which consists of two stages:

Received: 2025-07-18; **Accepted:** 2025-12-24; **Online (CNKI):** 2026-04-17

Foundation: Natural Science Foundation of Guangdong Province (2023A1515011667); Guangdong Basic and Applied Basic Research Foundation (2023B1515120020); Science and Technology Major Project of Shenzhen (KJZD20230923114809020)

Corresponding author: Associate professor WANG Xiaolan (187295890@qq.com); Researcher HE Yulin (yulinhe@gmail.ac.cn)

Citation: CHANG Xiuying, WANG Xiaolan, ZHU Tao, et al. NBEB-PUL: a naïve Bayesian ensemble-based algorithm for positive and unlabeled learning [J]. Journal of Shenzhen University Science and Engineering, 2026, 43(3): 338-346. (in Chinese)

Open access: This article is licensed under the CC BY 4.0 License (<https://creativecommons.org/licenses/by/4.0/>).

<http://journal.szu.edu.cn>



label assignment and noise filtering. In the label assignment stage, NBEB-PUL employs naïve Bayesian classifiers as base learners and integrates them using the AdaBoosting ensemble strategy to construct a strong classifier, named Ada-NBC. This strong classifier is then utilized to compute the mean posterior probabilities of a validation set for unlabeled samples, enabling explicit modeling of labeling uncertainty. Based on low-uncertainty predictions a set of reliable positive samples is iteratively identified, along with a residual unlabeled sample set. In the noise filtering stage, NBEB-PUL leverages the ensemble classifier generated in the first stage to reclassify the reliable positive set and the residual unlabeled set, resulting in Ada-NBC-predicted positive and negative sample sets. These sets are then intersected with the first-stage positive and unlabeled sets to extract overlapping high-confidence samples, forming the final positive and negative sample sets. The samples pruned from the dataset during this process are identified as noise. The feasibility, rationality, and effectiveness of NBEB-PUL were validated on 23 benchmark datasets from UCI and KEEL. Experimental results demonstrate that the algorithm exhibits stable convergence during the training process as the number of iterations increases. Moreover, NBEB-PUL outperforms six state-of-the-art PUL algorithms (S-EM, Biased-SVM, Modified-PUL, PU-LP, LP-PUL, and AdaPU) in terms of classification accuracy under varying positive sample proportions of 0.45, 0.40, 0.35, and 0.30. These results confirm that NBEB-PUL provides an effective and robust solution for positive and unlabeled learning in the presence of noise.

Key words: knowledge engineering; positive and unlabeled learning; labeling uncertainty; posterior probability; Bayesian classifier; ensemble learning

正类-无标签学习 (positive and unlabeled learning, PUL) 是机器学习的一个重要分支^[1]. 传统的半监督二分类模型通过从正类样本和负类样本中学习类别信息从而判断无标签样本集类别, 而 PUL 的目标是仅基于正类样本和无标签样本就能训练出准确的二分类器. 其中, 无标签样本集中的样本数据既有正类也有负类^[2-3]. 对于正类样本数据和大量无标签数据极易获得, 而负类样本数据缺失或者获取负类样本代价非常高^[4]的应用场景, 如传染病大规模筛查和高光谱图像识别^[5]场景, PUL 是解决此类数据分析问题的首选算法.

现有的 PUL 算法主要有两步法、有偏学习法、成本敏感法和基于图的方法. 两步法执行时先在无标签样本集中筛选出正负样本, 再以原始正样本和已确认的可靠负样本作为训练集, 进行分类器学习. 代表性的 PUL 方法包括基于朴素贝叶斯分类器 (naïve Bayesian classifier, NBC) 的间谍样本期望最大化 (spy expectation maximization, S-EM) PUL 算法^[6]、改进的 PUL(modified-PUL)算法^[7], 以及基于聚类的 PUL 算法^[8]等.

有偏学习法^[9]主要是基于支持向量机 (support vector machines, SVM) 构建 PUL 算法, 把未标记样本视作存在标注错误的负类样本, 并在损失函数中引入区别性约束, 为正类和负类样本分别分配不同的惩罚力度. 代表性方法包括偏置 SVM (Biased-SVM) 算法^[10]、循环学习框架有偏 SVM^[11]、加权无

标签 SVM 算法^[12], 以及带偏置的 Bagging-SVM 学习器^[13]等.

成本敏感法^[14]将 PUL 视为一种代价敏感型学习, 通过手动或自动的方式给类别进行加权惩罚, 并将未标记的样本视为负类. 代表性的方法包括普适性风险估计器^[15]、带有统计查询的 PUL 算法^[16], 以及基于无标签样本预测的重加权 PUL 算法^[17]等.

基于图的方法先构造一个表征数据间关联的图模型以获取负样本, 再在此模型上施行标签传递, 借助正负样本的关联特性来增强分类器的决策能力. 代表性方法包括正文档放大正类-无标签 (positive and unlabeled, PU) 分类器^[18]、基于图结构标签传播的正类-无标签学习 (positive and unlabeled learning by label propagation, PU-LP) 算法^[19]、基于标签传播 (label propagation, LP) 的 PUL 算法 LP-PUL^[20]等.

除上述单个分类器的 PUL 算法构建策略外, 采用集成学习的思想解决 PUL 问题也日渐受到研究人员的关注. 例如, 基于 Bagging 技术的 PUL 算法^[21]、基于 AdaBoosting 的迁移学习算法^[22], 以及基于 AdaBoosting 的自适应 PUL 框架 (adaptive PU, AdaPU)^[23]等.

通过分析现有 PUL 算法, 如 S-EM、Biased-SVM、Modified-PUL、PU-LP、LP-PUL 和 AdaPU 等在处理无标签样本时的流程和性能发现, 这些算法的标注过程常伴随着不确定性, 主要表现为分类器

难以有效判断样本的真实类别, 导致预测结果偏离真实结果.

在对无标签样本信息的利用方面, 现有方法多局限于预设规则或依赖于构建分类器进行直接筛选, 未能深入探索和充分利用无标签数据中蕴含的潜在信息. 尽管采用分类器集成机制的 PUL 算法能够通过集成多个基分类器, 在一定程度上提升模型的泛化能力和鲁棒性, 但通常未能充分考虑噪声样本对分类器构建以及对标注准确性的影响.

针对上述 PUL 算法的缺陷, 特别是基于集成学习构建的 PUL 算法存在的未能充分考虑正类和无标签样本中噪声对分类器构建影响的问题, 本研究提出一种包含类别标注和噪声筛选两个阶段的、基于 NBC 集成的正类-无标签学习 (NBC ensemble-based positive and unlabeled learning, NBEB-PUL) 算法, 通过结合集成分类器的类别标注与噪声筛选策略, 提升了无标签样本标注的准确性和鲁棒性.

在 23 个标准数据集上对 NBEB-PUL 算法的性能进行实验, 验证了随着迭代次数的增加, NBEB-PUL 算法的训练具有收敛性. 对比实验结果显示, NBEB-PUL 算法构建的分类器能够获得优于其他算法的预测性能. 在 4 种不同的正类样本比例情况下, NBEB-PUL 算法的测试精度平均排序值均领先于对比算法.

1 NBEB-PUL 算法

NBEB-PUL 算法通过引入集成学习的思想来强化基于后验概率的样本筛选, 且考虑了噪声数据对分类结果的影响, 提升了分类器对无标签样本类别标注的确定性, 增强了学习模型的准确性与泛化性.

NBEB-PUL 算法主要包括两个关键阶段: ① 基于集成贝叶斯的类别标注; ② 基于 AdaBoost 朴素贝叶斯分类器 (AdaBoosting NBC, Ada-NBC) 的噪声筛选. 算法首先通过集成多个贝叶斯分类器构建的强分类器, 强化样本标注的确定性, 减少单个分类器在类别判定过程中可能产生的偏差. 然后, 通过二次筛选机制对数据集进行再分类, 剔除潜在的噪声样本, 精炼了正负类样本, 以免影响由分类器学习到的决策边界的准确性.

NBEB-PUL 算法的数据处理流程图 (图 1) 直观地呈现了算法执行过程中数据标注的变化情况.

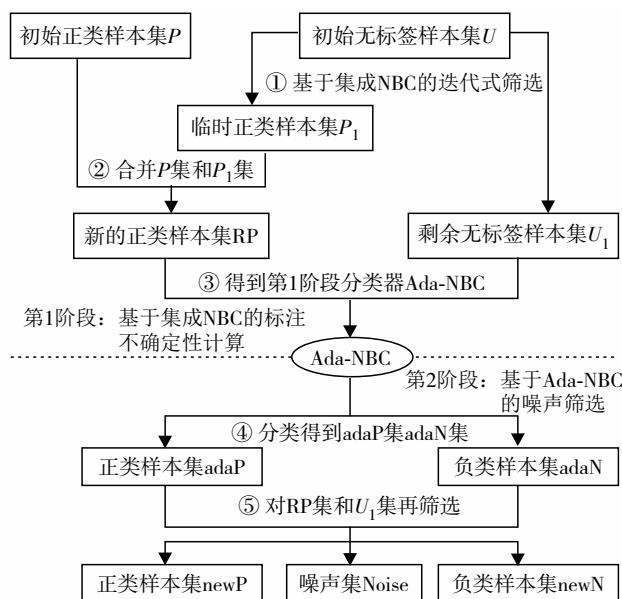


图1 NBEB-PUL算法数据处理流程

Fig. 1 Schematic diagram of the data processing in framework of NBEB-PUL algorithm.

NBEB-PUL算法执行步骤为:

- 1) 通过集成贝叶斯的迭代式类别筛选操作, 从初始的无标签样本集 U 中筛选出临时正类样本集 P_1 和剩余无标签样本集 U_1 ;
- 2) 将 P_1 集和初始正类样本集 P 合并, 得到新的正类样本集 RP ;
- 3) 利用第 1 阶段最后一次迭代得到的分类器 Ada-NBC 对原始数据集进行再分类;
- 4) 使用 Ada-NBC 分类器对原始样本集进行预测, 得到正类样本集 $adaP$ 和负类样本集 $adaN$;
- 5) 从 RP 集和 $adaP$ 集中找出重合样本, 生成最终的正类样本集 $newP$, 对 U_1 集和 $adaN$ 集也进行同样的操作, 得到最终的负类样本集 $newN$, 剩余未被挑选的样本则形成噪声集 $Noise$.

以下将对 NBEB-PUL 算法的关键步骤进行详细介绍与分析, 并讨论算法如何通过两阶段框架提升样本标注过程的确定性, 进而有效识别噪声数据.

1.1 基于贝叶斯集成的类别标注

将原始数据集划分为训练集和测试集, 其中训练集进一步被划分为正类样本集 P 、无标签样本集 U 和验证集 V . 首先, NBEB-PUL 算法利用 P 集和 U 集训练初始贝叶斯集成分类器, 并根据该分类器计算 V 集的初始后验概率均值 a . 然后, 算法进入迭代阶段, 以求精细地判定出每个无标签样本的类别.

在每次迭代中, 算法首先依次将 U 集中的每个无标签样本 u 加入 P 集, 形成新的正类样本集 P' 和新的无标签样本集 U' . 随后, 使用 P' 集和 U' 集训练 1 个新的贝叶斯集成分类器, 并用该分类器对 V 集预测, 计得新的验证集后验概率均值 a' . 最后, 计算 u 与 P 集中心点的距离 D_p , 以及与 U' 的距离 $D_{U'}$.

根据计算结果, NBEB-PUL 算法进行如下判别: 若新的验证集的后验概率均值 $a' > a$, 且 $D_p < D_{U'}$, 则将 u 判定为正类样本, 并将其归入临时正类样本集 P_1 . 持续迭代该过程, 直到 U 集中的样本数量不再变化, 即后验概率均值达到收敛. 最终, 算法保留最后 1 次迭代得到的集成朴素贝叶斯分类器 Ada-NBC, 为后续算法进行基于 Ada-NBC 的噪声筛选做准备. 第 1 阶段的流程图如图 2.

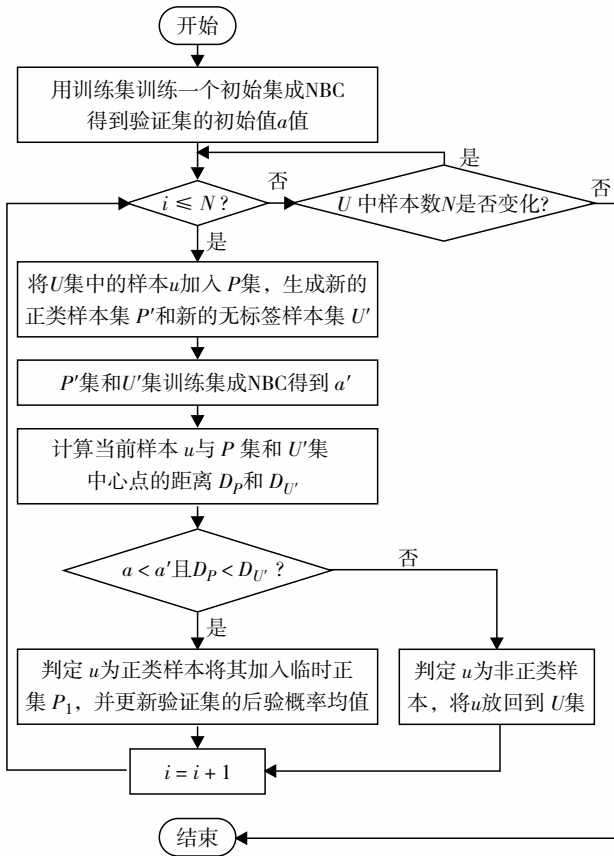


图2 基于贝叶斯分类器集成的类别标注阶段流程图

Fig. 2 Flowchart of category labeling phase based on NBC ensemble.

通过以上迭代机制, 算法能够不断优化样本标注的准确性, 逐步提高正类和无标签样本的判别精度, 最终获得稳定的分类结果.

下面将给出样本后验概率的计算过程. 给定输

入样本 $\mathbf{x}$, 贝叶斯分类器将计算每个类别 C_k 对应的后验概率

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})} \quad (1)$$

其中, $P(\mathbf{x}|C_k)$ 为类别 C_k 下观测到特征 $\mathbf{x}$ 的似然, 表示在假设样本属于 C_k 的情况下, 观察到样本 $\mathbf{x}$ 的概率; $P(C_k)$ 为 C_k 的先验概率; $P(\mathbf{x})$ 为所有类别下样本 $\mathbf{x}$ 出现的总概率.

本研究采用 AdaBoosting 算法对 NBC 进行集成. AdaBoosting 算法不涉及复杂的调参过程, 对噪声和异常值也具有一定的鲁棒性, 且算法通过迭代调整样本权重来训练分类器, 能够更好地既保持较低复杂度又避免过拟合. AdaBoosting 算法中每个弱分类器都会给出 1 个类别的后验概率 $P(C_k|\mathbf{x})$, 并根据分类器的错误率来调整它们对最终分类结果的贡献 (权重 α_t , t 为训练轮数, 亦是基分类器个数, 每轮训练生成 1 个基分类器). 每个分类器的输出概率会乘上它对应的权重 α_t , 再对所有分类器的加权概率进行求和.

假设在 t 轮训练中, 分类器 $h_t(\mathbf{x}_i)$ 为样本 $\mathbf{x}_i$ 关于类别 C_k 的后验概率 $P(C_k|\mathbf{x}_i)$, AdaBoosting 算法需要确定 $h_t(\mathbf{x}_i)$ 的错误率 ϵ_t , 并据此计算该分类器的权重

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) \quad (2)$$

$$\epsilon_t = \frac{\sum_{i=1}^m w_i I(h_t(\mathbf{x}_i) \neq y_i)}{\sum_{i=1}^m w_i} \quad (3)$$

其中, m 为样本个数; w_i 为给定数据集中第 i 个样本 $\mathbf{x}_i$ 的权重, 随着 AdaBoosting 的迭代根据 α_t 进行更新; I 为指示函数, 若样本被错误分类, 则 $I = 1$, 否则 $I = 0$.

对于每个类别 C_k , AdaBoosting 将各贝叶斯分类器的加权后验概率累加后得到 $\mathbf{x}_i$ 的后验概率为

$$P(C_k|\mathbf{x}_i) = \frac{\sum_{t=1}^T \alpha_t I(C_k) P_t(C_k|\mathbf{x}_i)}{\sum_{t=1}^T \alpha_t} \quad (4)$$

其中, $P_t(C_k|\mathbf{x}_i)$ 为第 t 个分类器输出类别 C_k 的后验概率; α_t 为第 t 个分类器的权重; T 为 NBC 的个数.

本阶段 NBEB-PUL 算法利用 V 集后验概率均值

的变化和每个无标签样本与 P 集、 U 集的空间分布特征, 对初始的无标签样本集 U 进行了初步的类别划分. 算法从 U 集中筛选出部分高置信度的正类样本作为临时正类样本 P_1 , 并把剩余的无标签样本集 U_1 视为临时负类样本集. 此阶段的处理为后续对数据集的进一步筛选与优化奠定了基础, 为提升分类器的精度和鲁棒性提供了高质量的数据支持.

1.2 基于 Ada-NBC 的噪声筛选

在本阶段, NBEB-PUL 算法的目标是从数据集中剔除潜在的噪声样本. 考虑到第 1 阶段的分类可能存在误差, 如部分潜在的噪声负类样本可能被误判为可靠正类, 因此, 算法会对第 1 阶段筛选出的正类集和无标签集进行“修剪”, 而非通过迭代增长策略扩大正负类样本集. “修剪”后构建出的正类样本集和负类样本集更可靠, 不仅提升了分类的准确率, 还有效防止了噪声样本对分类器决策边界的干扰.

第 1 阶段最终的 Ada-NBC 分类器是在无标签样本集 U 中样本数量达到稳定、不再变化的条件下构建的, 此时已无法从 U 中提取出更多高置信度的正类样本. 最终分类器利用剩余的无标签样本集 U_1 和所有高置信度的正类样本集 RP 进行训练, 形成第 1 阶段结束时的集成贝叶斯分类器 Ada-NBC.

使用 Ada-NBC 对原始数据集中的样本进行分类, 得到正类样本集 $adaP$ 和负类样本集 $adaN$. 随后对 $adaP$ 集和 RP 集取交集, 筛选出两集重合的样本, 形成最终的正类样本集 $newP$. 类似地, 对 $adaN$ 集和 U_1 集取交集, 筛选出最终的负类样本集 $newN$. 未被划入 $newP$ 集和 $newN$ 集的剩余样本则被视为潜在的噪声样本, 并划入噪声集 $Noise$. NBEB-PUL 算法的具体执行流程可扫描论文末页右下角二维码查看补充材料图 S1. 其中, 后验概率均值为验证集样本加权后验概率的算术平均值.

2 实验验证与结果分析

2.1 实验设置

实验数据选自 UCI 与 KEEL 的 23 个基准数据集, 见表 1. 首先, 对所有数据集进行预处理, 剔除离散属性, 以适配贝叶斯分类器对连续型特征概率的估计. 然后, 针对多类数据集, 筛选出样本量最大的两个类别并进行均衡化调整, 防止类别比例失衡. 最后, 通过随机指派, 将这 2 个类别中的一

个定为正类, 另一个定为负类.

表 1 实验所用基准数据集信息

Table 1 The information of data sets 个

数据集	属性数	样本数	数据集	属性数	样本数
abalone	7	1 268	seeds	7	140
ALFD	11	212	spambase	57	3 624
balance	4	576	thyroid	6	735
customer Churn	8	990	titanic	3	1 422
forest	27	318	vertebral	6	398
housevotes	16	336	breast	10	480
HTRU_2	8	3 278	bd	30	422
optdigits	56	1 142	credit	6	612
page-blocks	10	658	band	23	456
phoneme	5	3 172	pima	8	536
saheart	8	192	vehicle	8	432
satimage	36	3 016			

针对每个数据集, 先将其随机划分为测试集 (30%) 与训练基础 (70%), 再构建无标签集 U : 从正类中随机抽取占比为 d 的样本, 并与全部负类样本合并. 从剩余的正类样本中, 再留出 30% 作为验证集 V , 其余样本则构成正类样本集 P . 实验分别测试了 d 为 10%、20%、30% 及 40% 时的情况, 并重点展示了对应正类占比 R 为 45%、40%、35% 和 30% 时的实验结果.

对比算法选择 2002—2025 年报道的 6 种具有代表性的 PUL 算法, 包括 S-EM^[6]、Modified-PUL^[7]、Biased-SVM^[10]、PU-LP^[19]、LP-PUL^[20] 和 AdaPU^[23]. 所有算法均使用 Python 语言实现, 参数设置遵从原始文献中的设定, 其中部分关键参数设置如下: S-EM 算法中正类样本比例 $s = 10\%$, 噪声样本比例 $l = 15\%$; PU-LP 算法中路径权重控制参数 $\alpha = 0.02$, 正类样本规模控制参数 $\lambda = 0.8$, 迭代次数 $m = 2$, 近邻数 $k = 4$; LP-PUL 算法中构建图的算法采用 k 近邻算法, 其中近邻数 $k = 7$; 负面文档提取中数量 $\text{num_neg} = 1 \times |P|$; 标签传播部分采用高斯场谐波函数 (Gaussian field harmonic function, GFHF) 算法.

实验评估指标设为分类准确率 (测试集中被正确分类的样本占比). 对比算法在产出最终的样本类别标记后, 均使用 1 个标准的 NBC 进行性能测试. 为保障结果稳健, 每种算法均记录独立运行 5 次后的精度均值. 所有实验统一在 Intel(R) Core

(TM) i5-9400 处理器(2.90 GHz), 16.0 Gbyte 内存的计算机上执行。

2.2 NBEB-PUL 算法的可行性验证

可行性实验主要检查参数选取对算法性能的影响, 以及验证算法的收敛性。学习率是 AdaBoost 算法的重要参数, 用于控制每个弱分类器在最终模型中的权重调整幅度。学习率最常见的取值范围为 $(0, 1]$: 较低的学习率(如 0.1 或 0.5)会使模型参数更新幅度较小, 通常可避免过拟合, 并提供更稳定的学习过程; 较高的学习率(如 1.0)会使模型参数更新幅度较大, 加快收敛速度, 但易出现过拟合, 产生发散或振荡, 难以收敛到全局或局部最优解, 尤其是在数据较复杂或噪声较大的情况下。

算法还通过验证集后验概率的变化来指导样本筛选。在循环迭代过程中, 随着 U 中样本 u 不断被判定为高置信度的正类样本并加入正类样本集, 验证集的后验概率均值会逐步上升。此趋势反映了分类器对样本标注的不确定性逐渐降低, 模型的分类能力不断增强的过程。随着 U 中剩余样本的减少, 验证集后验概率的提升速率逐渐放缓, 直至达到收敛状态, 算法完成对无标签样本的筛选。

选取 housevotes、page-blocks、phoneme 和 vertebral 数据集分析学习率的选取对验证集后验概率收敛情况的影响。设 $R = 40\%$, 分类器个数为 50, 则 4 个数据集上 NBEB-PUL 算法验证集后验概率的收敛情况可扫描论文末页右下角二维码查看补充材料图 S2。由图 S2 可见, 当学习率为 0.1 时, 算法表现出较高的稳定性, 不易出现发散或振荡的现象, 避免了收敛困难的问题。此外, 在学习率为 0.1 时, 算法收敛后对验证集后验概率的预测均值显著提升, 表明预测的确定性较高, 优于学习率取值较大的情形。同时, 该学习率设定下算法在一定的迭代次数内即可达到收敛状态, 展现出较好的收敛效率。实验结果表明, NBEB-PUL 算法具备从无标签样本中甄别出可信正类样本的能力, 并能够通过迭代优化不断提升贝叶斯集成分类器的性能, 从而证实了算法的可行性。

2.3 NBEB-PUL 算法的合理性验证

合理性实验重点检验了 NBEB-PUL 算法的两阶段框架能否够实现以下目标: ① 在第 1 阶段从无标签样本集中准确筛选出可靠的正类样本; ② 在第 2 阶段能有效识别并剔除易混淆的噪声样本。因此, 本研究基于 bd 数据集, 比较第 1 阶段中单个贝叶斯

分类器与集成贝叶斯分类器的分类效果; 而在第 2 阶段, 对比 NBEB-PUL、PU-LP 和 LP-PUL 算法的分类能力。

bd 数据集共有 422 条样本, 将 30%(127 条)分配为测试集, 剩余 70%(295 条)用于定义训练范畴。当 $R = 30\%$ 时, 训练范畴被明确界定为: 验证集 26 条、正类集 59 条、无标签集 210 条。该无标签集由 153 个真实负例和 57 个真实正例混合而成。NBEB-PUL 算法在此设定下的分类效果可扫描论文末页右下角二维码查看补充材料图 S3。图 S3(a)呈现了训练集正负样本的分布情况。从中可见, 在正负样本交界处的样本较多, 增加了分类的难度。这些样本可能包含错误标注或异常样本, 原因是 bd 数据集是一个真实的医疗数据集, 记录了献血者在未来某段时间内是否会再次献血的情况。由于献血行为可能受到随机因素的影响(例如个人健康状况或临时意愿), 标签可能无法完全准确反映献血者的真实意图。此外, 在数据采集过程中可能存在记录错误或标注偏差, 进一步增加了数据的不确定性。这些特点为 NBEB-PUL 算法在处理复杂和噪声数据方面的能力提供了可用的测试场景。

图 S2(b)展示了训练集初始阶段的 P 集与 U 集的散点分布, 其中部分随机挑选的正类样本被划入 U 集。图 S3(c)为第 1 阶段基于贝叶斯集成的类别标注结果, P_1 集中的样本在此阶段里选出的可靠正类样本。图 S3(d)为第 2 阶段基于 Ada-NBC 的噪声筛选结果。从中可见, 大多数噪声样本分布在正负样本的交界区域, 符合预期。图 S3(e)为算法对训练集分类完成后的样本分布结果。从图中可观察到, 绝大多数样本已被正确分类。

对于第 1 阶段的分类效果, 本研究对比了单个贝叶斯分类器与集成贝叶斯分类器在从无标签样本集中挑选正类样本时的性能, 结果请扫描论文末页右下角二维码查看补充材料图 S4。由图 S4 可见, 单个贝叶斯分类器从无标签样本集中挑选出 29 个正类样本, 全部为正类样本, 占 U 集中正类样本总数的 50.88%; 集成贝叶斯分类器从无标签样本集中挑选出了 54 个正类样本, 其中 52 个为正类样本, 2 个为负类样本, 占 U 集中正类样本总数的 91.23%。这表明, 集成贝叶斯分类器的分类效果较单个贝叶斯分类器有显著提升, 它不仅能挑选出更多的正类样本, 还保持了较高的分类正确率, 有效减少了错误标注的风险。

通过与 LP-PUL 算法和 PU-LP 算法的对比, 评估 NBEB-PUL 算法在第 2 阶段的分类性能, 结果请扫描论文末页右下角二维码查看补充材料图 S5. 由图 S5 可见, NBEB-PUL 算法在最终的分类结果中表现出色: 最终的正类样本集 newP 包含 107 个样本, 其中正类样本 106 个, 负类样本 1 个; 最终的负类样本集 newN 包含了 153 个样本, 其中正类样本 2 个, 负类样本 151 个; 噪声集 Noise 包含了 9 个样本, 其中正类样本 8 个, 负类样本 1 个. 值得注意的是, 噪声集中的正类样本主要分布在负类样本区域内, 而负类样本则分布在正类样本区域内. NBEB-PUL 算法在正类样本上的分类正确率为 82.46%, 在负类样本上的分类正确率为 98.69%, 总分类精度达到 93.70%.

相比之下, LP-PUL 算法的最终正类样本集 fin_P 包含 114 个样本, 其中正类样本 73 个, 负类样本 41 个; 最终的负类样本集 N 包含了 155 个样本, 其中正类样本 57 个, 负类样本 112 个. 算法中正类样本的分类正确率仅为 24.56%, 负类样本的分类正确率为 72.26%, 总分类精度为 85.83%.

PU-LP 算法的最终正类样本集 fin_P 包含了 105 个样本, 其中正类样本 104 个, 负类样本 1 个; 负类样本集 N 包含了 105 个样本, 全部为负类样本; 噪声集 Noise 包含了 59 个样本, 其中正类样本 12 个, 负类样本 47 个. 噪声样本主要集中在正负样本边界和负类样本内部, 筛选结果稍显杂乱, 有很多负类样本区域内部的负集被误分类为噪声样本. PU-LP 算法在正类样本上的分类正确率为 78.95%, 负类样本的分类正确率为 68.63%, 总分类精度为 90.55%.

为明确区分类别标注和噪声筛选的独立贡献, 添加类别标注和噪声筛选的消融实验. 选择 bd、band、forest、spambase 和 seeds 数据集, 在同等条件下, 以 45% 的正类样本比例重复 5 次独立实验并且取平均值作为最终结果. 由于噪声筛选阶段是对类别标注阶段的优化, 因此前后训练集和测试集的筛选都是相同的. 基于此前提进行消融实验, 得到分类别标注和噪声筛选的测试精度结果请扫描论文末页右下角二维码查看补充材料图 S6. 由图 S6 可见, 在 5 个数据集中, 噪声筛选阶段较类别标注阶段的测试精度都有显著提升, 尤其是在 bd 数据集, 升幅接近 12%. 消融实验结果进一步证明基于 Ada-NBC 的噪声筛选能够消除类别标注阶段分类存在的

误差, 通过此修剪过程, 可构建出更准确的类样本集和负类样本集, 进一步提升分类器的泛化性能.

综上, NBEB-PUL 算法在分类精度上显著优于 LP-PUL 和 PU-LP 算法. 特别是在处理复杂数据分布和噪声样本时, NBEB-PUL 算法展现出良好的鲁棒性和高效性. 其通过两阶段框架实现了对正类和负类样本的高效筛选, 以及对噪声样本的有效识别, 为提高分类器的分类性能提供了有力的支持.

2.4 NBEB-PUL 算法的有效性验证

基于两种不同的评估标准, 在 23 个标准数据集上通过与 S-EM、Biased-SVM、Modified-PUL、PU-LP、LP-PUL 和 AdaPU 算法进行对比, 以验证 NBEB-PUL 算法的有效性. 为考察不同样本分布对分类任务的影响, 分别设置正类样本比例(R)为 45%、40%、35% 和 30%, 结果请扫描论文末页右下角二维码查看补充材料表 S1 至表 S4. 所有结果都是基于 5 次独立运行的均值, 并引入测试精度与临界差(critical difference, CD)值作为衡量标准.

在整个实验评估中, 若算法单次运行时长超过 30 min, 其测试结果将被视为无效. LP-PUL 算法在 HTRU_2、phoneme、satimage 和 spambase 数据集上均因耗时超出上限而未被采用. 在不同训练设置下, NBEB-PUL 算法的平均分类准确率均显著优于基准算法, 较对比算法中的最优结果分别提高了 1.64%、1.52%、2.70% 和 2.23%, 充分证明 NBEB-PUL 算法是一种具备实用性的偏标记学习解决方案.

从表 S1 至 S4 可见, NBEB-PUL 算法在不同正类样本比例条件下均表现出显著的性能优势. 当 R 为 45% 时, NBEB-PUL 算法的平均精度较对比算法的最优结果提升了约 1.64 个百分点; 当 R 为 40% 时, 其平均精度提升了约 1.52 个百分点; 当 R 为 35% 时, 平均精度提升幅度达到 2.70 个百分点, 该比例下算法改进效果最为显著; 而当 R 为 30% 时, 算法的平均精度仍提升了约 2.23 个百分点. 综合来看, NBEB-PUL 算法在 23 个数据集上的分类精度总体优于其他 6 种 PUL 算法, 尤其在正类样本比例较低时表现出更强的鲁棒性和优越性, 进一步验证了该算法在处理 PUL 任务中的有效性.

采用临界差值(critical difference, CD)图对 NBEB-PUL 算法的分类性能进行统计显著性检验. 由于 LP-PUL 算法在 HTRU_2、phoneme、satimage 和 spambase 数据集上的运行结果不符合时效要求,

因此排除这4个数据集的实验结果。

基于5次独立重复实验的平均结果,补充材料在正类样本比例 R 为45%、40%、35%和30%时,19个数据集的测试精度CD图可扫描论文末页右下角二维码查看补充材料图S7。其中,纵轴为各个对比算法,横轴为算法的精度平均排序值。对于给定的0.05显著性水平,临界差值为

$$D_c = q_{0.05} \sqrt{\frac{k(k+1)}{6N}} \quad (5)$$

其中, k 为对比算法数量; N 为数据集个数。本研究取 $k=7$, $N=19$,则有 $D_c \approx 1.997$,因此,最终采用Bonferroni分布的临界值 $q_{0.05} = 2.849$ 。

分析图S7可知,NBEB-PUL算法的平均序位被一个临界差区间所环绕;凡序位超出此区间的算法,均与NBEB-PUL存在显著性能差异。在 R 分别为45%、40%、35%和30%时,NBEB-PUL算法均取得了7种方法中的最佳(最低)平均排序,验证了其优越性,可见该算法是一种能够高效解决PUL问题的机器学习方案。

结 语

针对现有集成式正类-无标签学习算法对样本噪声敏感性关注不足的问题,提出基于朴素贝叶斯集成的新型PUL方法——NBEB-PUL。该方法通过集成学习增强对无标签样本的判别一致性,并引入再分类机制滤除潜在噪声样本,从而提升了分类边界的鲁棒性。在多个公开数据集上的实验表明,NBEB-PUL算法在可行性、合理性和分类性能方面均表现良好,展现出较强的无标签样本识别能力,为PUL任务提供了一种稳定可行的解决方案。

为进一步优化算法性能,未来研究工作将围绕以下方向展开:①拓展基分类器类型。当前NBEB-PUL以朴素贝叶斯为基础分类器,后续可引入如决策树、神经网络、支持向量机等更多类型的分类器,探索其对算法整体性能的影响;②面向大规模数据场景。拟将NBEB-PUL算法拓展至分布式计算框架^[24],以适应跨域存储与处理需求,提升其在大规模PUL任务中的适用性与效率。

基金项目: 广东省自然科学基金资助项目(2023A1515011667);广东省基础与应用基础研究基金粤深联合基金重点资助项目(2023B1515120020);深圳市科技重大专项资助项目(KJZD20230923114809020)

作者简介: 常秀颖(775427738@qq.com), 沧州职业技术学院副教授。研究方向:移动通信网络数据分析。

引 文: 常秀颖, 王晓兰, 朱涛, 等. 基于朴素贝叶斯集成的正类-无标签学习算法[J]. 深圳大学学报理工版, 2026, 43(3): 338-346.

参考文献 / References:

- [1] DE COMITÉ F, DENIS F, GILLERON R, et al. Positive and unlabeled examples help learning [C]// Proceedings of the Tenth International Conference on Algorithmic Learning Theory. Berlin: Springer Berlin Heidelberg, 1999: 219-230.
- [2] DENIS F Ç. PAC learning from positive statistical queries [C]// Proceedings of the Ninth International Conference on Algorithmic Learning Theory. Berlin: Springer Berlin Heidelberg, 1998: 112-126.
- [3] LETOUZEY F, DENIS F, GILLERON R. Learning from positive and unlabeled examples [C]// Proceedings of the 11th International Conference on Algorithmic Learning Theory. Berlin: Springer-Verlag, 2000: 71-85.
- [4] FUNG G P C, YU J X, LU Hongjun, et al. Text classification without negative examples revisit [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 6-20.
- [5] LI Wenkai, GUO Qinghua, ELKAN C. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data [J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49(2): 717-725.
- [6] LIU Bing, LEE W S, YU P S. Partially supervised classification of text documents [C]// Proceedings of the 19th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers Inc., 2002: 387-394.
- [7] HERNÁNDEZ FUSILIER D, MONTES-Y-GÓMEZ M, ROSSO P, et al. Detecting positive and negative deceptive opinions using PU-learning [J]. Information Processing & Management, 2015, 51(4): 433-443.
- [8] CHAUDHARI S, SHEVADE S. Learning from positive and unlabelled examples using maximum margin clustering [C]// Proceedings of the Nineteenth International Conference Neural Information Processing. Berlin: Springer Berlin Heidelberg, 2012: 465-473.
- [9] ELKAN C, NOTO K. Learning classifiers from only positive and unlabeled data [C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008: 213-220.
- [10] LIU Bing, DAI Yang, LI Xiaoli, et al. Building text

- classifiers using positive and unlabeled examples [C]// Proceedings of the 3rd IEEE International Conference on Data Mining. Piscataway, USA: IEEE, 2003: 179-186.
- [11] KE Ting, YANG Bing, ZHEN Ling, et al. Building high-performance classifiers using positive and unlabeled examples for text classification [C]// Advances in Neural Networks-ISBN. Berlin: Springer Berlin Heidelberg, 2012: 187-195.
- [12] LIU Zhigang, SHI Wenzhong, LI Deren, et al. Partially supervised classification-based on weighted unlabeled samples support vector machine [C]// The 1st International Conference on Advanced Data Mining and Applications. Wuhan, China: Springer, 2005, 3584: 118-129.
- [13] MORDELET F, VERT J P. A bagging SVM to learn from positive and unlabeled examples [J]. Pattern Recognition Letters, 2014, 37: 201-209.
- [14] DU PLESSIS M C, NIU Gang, SUGIYAMA M. Analysis of learning from positive and unlabeled data [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2014: 703-711.
- [15] DU PLESSIS M, NIU Gang, SUGIYAMA M. Convex formulation for learning from positive and unlabeled data [C]// Proceedings of the 32nd International Conference on Machine Learning. Sardinia, Italy: PMLR, 2015: 1386-1394.
- [16] DENIS F, GILLERON R, LETOUZEY F. Learning from positive and unlabeled examples [J]. Theoretical Computer Science, 2005, 348(1): 70-83.
- [17] STERCKX L, CARAGEA C, DEMEESTER T, et al. Supervised keyphrase extraction as positive unlabeled learning [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2016: 1924-1929.
- [18] YU Shuang, LI Chunping. PE-PUC: a graph based PU-learning approach for text classification [C]// Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition. Berlin: Springer Berlin Heidelberg, 2007: 574-584.
- [19] MA Shuangxun, ZHANG Ruisheng. PU-LP: A novel approach for positive and unlabeled learning by label propagation [C]// Proceeding of the 2017 IEEE International Conference on Multimedia & Expo Workshops, Hongkong, China, IEEE, 2017: 537-542.
- [20] CARNEVALI J C, GERALDELI ROSSI R, MILIOS E, et al. A graph-based approach for positive and unlabeled learning [J]. Information Sciences, 2021, 580: 655-672.
- [21] WU Bangyu, QIU Weirong, JIA Junxiong, et al. Landslide susceptibility modeling using bagging-based positive-unlabeled learning [J]. IEEE Geoscience and Remote Sensing Letters, 2021, 18(5): 766-770.
- [22] LIU Bo, LIU Changdong, XIAO Yanshan, et al. AdaBoost-based transfer learning method for positive and unlabelled learning problem [J]. Knowledge-Based Systems, 2022, 241: 108162.
- [23] ZHAO Yawen, ZHANG Mingzhe, ZHANG Chenhao, et al. A boosting framework for positive-unlabeled learning [J]. Statistics and Computing, 2024, 35(1): 2.
- [24] SUN Xudong, HE Yulin, WU Dingming, et al. Survey of distributed computing frameworks for supporting big data analysis [J]. Big Data Mining and Analytics, 2023, 6(2): 154-169.

【中文责编：英子；英文责编：木柯】



补充材料