半监督 AUC 优化的 Boosting 算法及理论

杨智勇[1] 许倩倩[2] 何 源[3] 操晓春[4] 黄庆明[1),2),5),6)]

[1)](中国科学院大学计算机科学与技术学院 北京 101408)

[2)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[3)](阿里安全图灵实验室 北京 100102)

[4)](中国科学院信息工程研究所信息安全国家重点实验室 北京 100093)

[5)](中国科学院大数据挖掘与知识管理重点实验室,中国科学院大学 北京 101408)

[6)](鹏城实验室 广东 深圳 518055)

摘 要 ROC 曲线下面积(Area Under the ROC Curve,AUC)是类不均衡/二分排序等问题中的标准评价指标之一.本文主要聚焦于半监督 AUC 优化方法.现有大多数方法局限于通过单一模型进行半监督 AUC 优化,对如何通过模型集成技术融合多个模型则鲜有涉及.考虑上述局限性,本文主要研究基于模型集成的半监督 AUC 优化方法.具体而言,本文提出一种基于 Boosting 算法的半监督 AUC 优化算法,并提出基于权重解耦的加速策略以降低算法时间/空间复杂度.进一步地,在优化层面,本文通过理论分析证明了所提出的算法相对于弱分类器的增加具有指数收敛速率;在模型泛化能力层面,本文构造了所提出算法的泛化误差上界,并证明增加弱分类器个数在提升训练集性能的同时并不会带来明显的过拟合风险.最后,本文在 16 个基准数据集上对所提出算法的性能进行了验证,实验结果表明所提出算法在多数情况下以 0.05 显著水平优于其他对比方法,并可在平均意义上产生 0.9%～11.28%的性能提升.

关键词 AUC 优化;集成学习;半监督学习;提升法;Rademacher 复杂度

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.01598

Boosting-Based Semi-Supervised AUC Optimization: Theory and Algorithm

YANG Zhi-Yong[1] XU Qian-Qian[2] HE Yuan[3] CAO Xiao-Chun[4] HUANG Qing-Ming[1),2),5),6)]

[1)](*School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 101408)

[2)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3)](*Alibaba Turing Security Lab, Beijing* 100102)

[4)](*State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093)

[5)](*Key Laboratory of Big Data Mining and Knowledge Management (BDKM), University of Chinese Academy of Sciences, Beijing* 101408)

[6)](*Peng Cheng Laboratory, Shenzhen, Guangdong* 518055)

Abstract Area Under the ROC Curve (AUC) is a standard evaluation metric for a wide range of tasks such as class-imbalance classification and bipartite ranking. This paper focuses on the semi-supervised AUC optimization problem. Most existing methods only adopt single-model-based methods, while rarely taking into account the benefit of combine multiple models. To address this issue, this paper studies the problem of how to effectively ensemble a series of semi-supervised

收稿日期:2020-12-04;在线发布日期:2021-10-28.本课题得到科技创新 2030-"新一代人工智能"重大项目(2018AAA0102003)、国家自然科学基金项目(61620106009,61931008,61836002,U2001202,61976202)、中央高校基本科研业务费专项资金资助、中国科学院战略性先导科技专项(XDB28000000)、博士后创新人才支持计划(BX2021298)、中国科学院青年创新促进会、阿里巴巴集团 ARF 项目资助.**杨智勇**,博士,中国计算机学会(CCF)会员,主要研究方向为机器学习理论与方法.E-mail:yangzhiyong21@ucas.ac.cn.**许倩倩**(通讯作者),博士,副研究员,国家自然科学基金优秀青年科学基金入选者,中国计算机学会(CCF)高级会员,主要研究方向为统计机器学习及其在多媒体领域的应用.E-mail:xuqianqian@ict.ac.cn.**何 源**,博士,高级工程师,中国计算机学会(CCF)会员,主要研究方向为计算机视觉、机器学习、AI 安全.**操晓春**,博士,研究员,国家杰出青年科学基金入选者,中国计算机学会(CCF)高级会员,主要研究领域为计算机视觉、多媒体分析.**黄庆明**(通讯作者),博士,讲席教授,IEEE Fellow,国家杰出青年科学基金入选者,中国计算机学会(CCF)会士,主要研究领域为多媒体计算、图像处理、计算机视觉、模式识别.E-mail:qmhuang@ucas.ac.cn.

AUC optimization methods. Specifically, we propose a boosting-based semi-supervised AUC optimization method. On top of this, we provide an acceleration strategy based on a weight decoupling strategy to reduce the time and space complexity. Moreover, we theoretically prove that the proposed algorithm has an exponential convergence rate with respect to the number of weak learners. Meanwhile, we provide a generalization error bound of the proposed method, and further prove that increasing the number of weak learners could improve the performance on the training set without the cost of a significant overfitting effect. Finally, we evaluate our proposed framework on 16 benchmark datasets. Experimental results show that the proposed algorithm outperforms all the competitors with a significance level of 0.05, and achieves a 0.9%—11.28% performance gain in average.

Keywords　AUC optimization; ensemble learning; semi-supervised learning; boosting; Rademacher complexity

1　引　言

AUC(Area Under the ROC Curve),即 ROC 曲线下面积,旨在衡量正样本得分高于负样本得分的概率,被广泛应用于评估得分函数的排序性能.由于对于类别分布、错分代价均不敏感[1-2],AUC 相较于准确率在诸多场景下更适合作为评价指标.一方面,对于类别不均衡任务例如疾病预测[3]和异常事件检测[4],不同类别样本分布十分不平衡,使某些类别样本明显多于其他类别样本.在这种情形下,稀有类别预测精度相比于其他类别往往更为重要.而准确率指标则常常忽略稀有类别的性能,因此不适用于该场景.相反,AUC 的值并不依赖于样本类别分布,自然地成为了样本类别不均衡情形下的常用评价指标.另一方面,在点击率(CTR)预测、推荐系统[5-6]等应用场景中,预测正负类样本之间的相对排序较之预测样本分类更为重要,AUC 由于仅关注正负样本之间的相对排序而成为标准评价指标之一.

在早期的机器学习研究中,往往采用最小化错误率的理念设计模型和优化算法.那么能否在最小化错误率的框架下实现最大化 AUC 的目的呢?Cortes 等人在其工作[7]中指出根据最小化错误率得出的模型可能在 AUC 指标意义下为次优模型,因此有必要直接针对 AUC 指标设计优化方法.在此项工作之后的近二十年内,涌现出了大批 AUC 优化的相关研究[8-23].

绝大多数的 AUC 优化相关研究局限于处理数据标注全部已知的情况,无法适用于数据中存在未标注样本的半监督学习场景.在近期的工作中,已有部分研究聚焦于半监督 AUC 优化问题.Fujino 等人[24]通过对未标注的数据分布函数建模构造了生成模型并由此构造了半监督 AUC 优化方法.Sakai 等人[25]首次基于 PU(Positive Unlabeled)学习框架推导出一个 AUC 的无偏估计,并结合有监督 AUC 和 PU 学习提出一种半监督 AUC 学习框架.该框架无需预生成未标记样本的伪标签,但需预先估计样本正负类的先验概率以对未标记样本进行加权,在已标记样本数量较小的情况下仍然存在局限性.Xie 等人[26]进一步指出在 0-1 损失意义下,无需任何先验分布信息也可由未标记样本估计 AUC 风险.Xie 等人[27]在文献[26]的基础上进一步实现了半监督 AUC 的随机优化方法.

目前,半监督 AUC 优化已取得了初步的成功.但现有半监督 AUC 优化方法仅针对单个线性模型进行设计,对更为复杂的模型则缺乏考虑.鉴于此,本文主要研究基于模型集成的半监督 AUC 优化算法并对其性质进行了系统的理论分析,以期通过融合多个弱学习器突破现有方法的瓶颈.具体而言,本文的主要贡献如下:

(1)提出一种基于 Boosting 的无先验半监督 AUC 优化模型集成方法,并就其效率瓶颈设计了高效的加速算法,大幅降低了更新单个弱分类器的时间/空间复杂度.

(2)在算法收敛速率方面,证明训练集误差随弱分类器的增加以指数速率迅速衰减.

(3)在算法泛化误差性能方面进行了系统的理论分析,首先根据半监督 AUC 优化目标函数自身特点构造了半监督 AUC 优化的 Rademacher 复杂度;其次,针对该复杂度特性提出一种广义最大值不

等式;最终给出了首个在模型集成意义下的半监督AUC优化泛化误差上界,并获得了比现有半监督AUC优化泛化界[25]更为紧致的结果.

此外,本文在16个标准数据集进行了系统的实验分析,实验结果表明本文算法在绝大多数情况下,可在0.05显著水平下优于其他对比方法.

2 相关工作

在介绍本文主要工作前,本节先就所涉及的既往工作及其必要技术细节进行简要回顾,2.1小节主要回顾有监督AUC优化一般化方法,2.2小节进一步延伸至半监督AUC优化方法[25-26],随后在2.3小节中回顾本文所基于的模型集成框架RankBoost[28],最后在2.4小节讨论本文工作与既往工作的主要区别.

2.1 有监督AUC优化

本文聚焦于二分类问题,给定训练样例$(\boldsymbol{x},y)$,其输入特征$\boldsymbol{x}$为d维欧式空间中向量,即$\boldsymbol{x}\in\mathbb{R}^d$,其类标$y$为1或$-1$.令采样过程中样例输入及类标的联合分布为$p(\boldsymbol{x},y)$,则正例样本集$\mathcal{X}_P$及负例样本集$\mathcal{X}_N$服从以下分布:

$$\mathcal{X}_P=\{\boldsymbol{x}_i\}_{i=1}^{n_p}\overset{i.i.d}{\sim}p_P(\boldsymbol{x})=P[\boldsymbol{x}\,|\,y=1],$$

$$\mathcal{X}_N=\{\boldsymbol{x}'_k\}_{k=1}^{n_n}\overset{i.i.d}{\sim}p_N(\boldsymbol{x}')=P[\boldsymbol{x}'\,|\,y=-1],$$

其中n_p,n_n分别为正例及负例样本个数;p_P,p_N分别表示正例及负例样本的条件分布.

给定正负例样本分布$\mathcal{P},\mathcal{N}$,将学习可预测样例得分的得分函数记为$h(\cdot)$.给定得分函数h,AUC指标通过计算ROC曲线下夹面积衡量该得分函数在不同分类阈值下的平均性能.由于本文并不涉及ROC曲线,因此不再详述该定义,读者可参阅文献[1]获得更多细节.下面引入一种更为直观的AUC指标定义.具体来说,Hanley等人[29]指出,AUC指标等价于正例样本根据h获得的得分高于负例样本获得的得分的概率,根据此结论,可将得分函数$h(\cdot)$对应的AUC指标表为

$$AUC(h)=1-\mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{x'\sim\mathcal{N}}[\ell_{0\text{-}1}(f(\boldsymbol{x},\boldsymbol{x}'))]\right],$$

其中$\boldsymbol{x}\sim\mathcal{P}$表示由正例分布采样获得$\boldsymbol{x}$;同理,$\boldsymbol{x}'\sim\mathcal{N}$表示由负例分布采样获得$\boldsymbol{x}'$;$f(\boldsymbol{x},\boldsymbol{x}')$为正负例分差即$f(\boldsymbol{x},\boldsymbol{x}')=h(\boldsymbol{x})-h(\boldsymbol{x}')$;$\ell_{0\text{-}1}$为0-1损失函数,即$\ell_{0\text{-}1}(\boldsymbol{x})=\boldsymbol{I}[\boldsymbol{x}<0]$.由此不难得出AUC数值即为正例负例得分排序的准确率.显然,需要通过最大化该指标获得最优模型.由于机器学习问题中常求解最小化问题,进一步引入AUC的损失形式,即:

$$R_{PN}(h)=1-AUC(h).$$

可将最大化$AUC(h)$目标等效转化为最小化$R_{PN}(h)$.显然,$R_{PN}(h)$的最小化问题为一组合优化问题,时间代价极高,且数据分布往往未知,无法直接计算目标函数.为近似求解该问题,AUC优化方法常以训练样本的平均值代替期望,并以连续可导的替代损失函数(surrogate loss function)ℓ代替0-1损失函数,进而得到替代风险函数(surrogate risk function)$\hat{R}^{\ell}_{PN}$:

$$\hat{R}^{\ell}_{PN}(h)=\frac{1}{n_p n_n}\sum_{\boldsymbol{x}\in\mathcal{X}_P}\sum_{\boldsymbol{x}'\in\mathcal{X}_N}\ell(f(\boldsymbol{x},\boldsymbol{x}')).$$

此时即可获得AUC优化对应的近似优化问题:

$$(\mathrm{OP}_1)\ \min_{h\in\mathcal{H}}\hat{R}^{\ell}_{PN}(h),$$

其中$\mathcal{H}$为所选定的假设空间,由模型类型(决策树、神经网络、线性模型等),以及正则化项决定.由于模型通常由参数确定,设模型h的参数为$\boldsymbol{w}$,并记此时模型为$h_{\boldsymbol{w}}$,并设参数选自集合$\mathcal{W}$,则可将(OP_1)等效转化为(OP_2):

$$(\mathrm{OP}_2)\ \min_{\boldsymbol{w}\in\mathcal{W}}\hat{R}^{\ell}_{PN}(h_{\boldsymbol{w}}).$$

迄今为止,既往工作已基于logistic损失[8-9]$\ell(t)=\log(1+\exp(-t))$、指数损失[10]$\ell(t)=\exp(-t)$、铰链损失[11-12]$\ell(t)=\max(1-t,0)$(hinge loss)及平方损失[13-14]$\ell(t)=(1-t)^2$分别尝试了对该一般框架的实现及加速,并取得了较优性能及效率突破.理论分析方面,文献[15-19]对全监督条件下AUC优化的泛化性能进行了系统研究,文献[22-23]对AUC替代损失相对0-1损失的一致性进行了系统分析.

2.2 半监督AUC优化

相比于全监督条件下的AUC研究,半监督条件下的AUC优化研究尚处于早期阶段.文献[25-26]对离线情况下的半监督AUC优化进行了系统研究,文献[27]对在线条件下的半监督AUC优化进行了系统研究.由于本文主要考虑离线条件下的AUC优化,因此下面针对文献[25-26]进行进一步介绍.

相比于全监督情况,半监督条件下数据集中还存在未标注数据$\mathcal{X}_U$,其生成过程如下:

$$\mathcal{X}_U=\{\widetilde{\boldsymbol{x}}_j\}_{j=1}^{n_u}\overset{i.i.d}{\sim}p(\boldsymbol{x})=\theta_P\cdot p_P(\boldsymbol{x})+\theta_N\cdot p_N(\boldsymbol{x}),$$

其中θ_p,θ_N分别表示$P[y=1],P[y=-1]$即正类和负类的先验概率.显然,引入未标注样集$\mathcal{X}_U$使期望损失$R^{0\text{-}1}_{PN}$无法直接计算.下面讨论如何在半监督条

件下估计真实期望风险 $R_{PN}^{0\text{-}1}$. 首先构建辅助风险损失函数 $R_{PU}^{0\text{-}1}$, $R_{UN}^{0\text{-}1}$ 如下：

$$R_{PU}^{0\text{-}1}=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}}\Big[\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathcal{U}}[\ell_{0\text{-}1}(f(\boldsymbol{x},\tilde{\boldsymbol{x}}))]\Big] \tag{1}$$

$$R_{UN}^{0\text{-}1}=\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathcal{U}}\Big[\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathcal{N}}[\ell_{0\text{-}1}(f(\tilde{\boldsymbol{x}},\boldsymbol{x}'))]\Big] \tag{2}$$

$R_{PU}^{0\text{-}1}$, $R_{UN}^{0\text{-}1}$ 分别表示正例得分低于未标注样例的概率，以及未标注样例得分低于负例的概率. Sakai 等人[25]证明，在可对 θ_P, θ_N 进行较好估计的前提下，可由 $R_{PU}^{0\text{-}1}$, $R_{UN}^{0\text{-}1}$ 估计出全监督损失 $R_{PN}^{0\text{-}1}$（同理替代风险间也满足该关系）. Xie 等人[26]进一步证明，即便在无法估计 θ_P, θ_N 条件下，$R_{PN}^{0\text{-}1}$ 亦可分别由 $R_{PU}^{0\text{-}1}$ 及 $R_{UN}^{0\text{-}1}$ 线性表示，其数学形式可表为

$$R_{PN}^{0\text{-}1}=\frac{1}{\theta_n}R_{PU}^{0\text{-}1}-\frac{1}{2}\cdot\frac{\theta_p}{\theta_n},$$

$$R_{PN}^{0\text{-}1}=\frac{1}{\theta_p}R_{UN}^{0\text{-}1}-\frac{1}{2}\cdot\frac{\theta_n}{\theta_p}.$$

该结论表明，$R_{PN}^{0\text{-}1}$ 数值同时正比于 $R_{PU}^{0\text{-}1}$ 及 $R_{UN}^{0\text{-}1}$，因此可在未知类别分布先验 θ_p, θ_n 的情况下通过优化 $R_{PU}^{0\text{-}1}$、$R_{UN}^{0\text{-}1}$ 优化 $R_{PN}^{0\text{-}1}$. 同理于有监督情况下的 AUC 优化近似问题，Xie 等人[26]采用平方损失 $\ell_{sq}(t)=(1-t)^2$ 作为替代损失构造如下替代风险函数：

$$\hat{R}_{PNU}^{\ell_{sq}}=\gamma\cdot\hat{R}_{PN}^{\ell_{sq}}(h)+(1-\gamma)\cdot\left(\hat{R}_{PU}^{\ell_{sq}}(h)+\hat{R}_{UN}^{\ell_{sq}}(h)-\frac{1}{2}\right),$$

其中：

$$\hat{R}_{PN}^{\ell_{sq}}(h)=\frac{1}{n_pn_n}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\ell_{sq}(f(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})) \tag{3}$$

$$\hat{R}_{PU}^{\ell_{sq}}(h)=\frac{1}{n_pn_u}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\ell_{sq}(f(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)})) \tag{4}$$

$$\hat{R}_{UN}^{\ell_{sq}}(h)=\frac{1}{n_un_n}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\ell_{sq}(f(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})) \tag{5}$$

$$\hat{R}_{PN}^{\ell_{\exp}}(f)=\frac{1}{n_pn_n}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\ell_{\exp}(f(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})) \tag{6}$$

$$\hat{R}_{PU}^{\ell_{\exp}}(f)=\frac{1}{n_pn_u}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\ell_{\exp}(f(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)})) \tag{7}$$

$$\hat{R}_{UN}^{\ell_{\exp}}(f)=\frac{1}{n_un_n}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\ell_{\exp}(f(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})) \tag{8}$$

$$f(\boldsymbol{x},\boldsymbol{x}')=f(\boldsymbol{x})-f(\boldsymbol{x}') \tag{9}$$

$$\ell_{\exp}(t)=\exp(-t) \tag{10}$$

在此基础上，进一步求解以下问题，即可得到 Xie 等人[26]所提出的 SAMULT 算法：

$$(\mathrm{OP}_3)\ \min_{h\in\mathcal{H}}\hat{R}_{PNU}^{\ell_{sq}}(h).$$

由于平方损失的基本性质，该问题存在显式闭式解（closed-form solution），具体细节可见文献[26]，本文不再赘述.

2.3　RankBoost 算法的一般化框架

本小节中，将介绍本文所采用的 RankBoost 算法的一般化框架. 考虑一般的逐对排序学习问题，给定逐对比较数据集 $\mathcal{X}_{\text{pair}}$，对所有 $(\boldsymbol{x}_1,\boldsymbol{x}_2)\in\mathcal{X}_{\text{pair}}$，希望得分 $f(\boldsymbol{x}_1)$ 应尽可能大于 $f(\boldsymbol{x}_2)$. RankBoost 近似最小化以下形式的指数排序损失函数实现该目标：

$$\sum_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\in\mathcal{X}_{\text{pair}}}D^0(\boldsymbol{x}_1,\boldsymbol{x}_2)\cdot\exp(-(f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2))).$$

为达成模型集成的目的，RankBoost 以迭代形式逐步学习多个弱学习器 $h^1,\cdots,h^T$，及其权重 $\alpha^1,\cdots,\alpha^T$，并以弱学习器的加权投票形式作为其输出 f，即 $f(\boldsymbol{x})=\sum_{t=1}^{T}\alpha^th^t(\boldsymbol{x})$. RankBoost 的一般框架见过程.

过程 1.　RankBoost 模型一般化过程.

输入：模型输入 $\boldsymbol{X},\boldsymbol{Y}$，弱分类器个数 T

输出：弱分类器权重 $\alpha^1,\cdots,\alpha^T$，弱分类器 $h^1,\cdots,h^T$

初始化样本权重 D^0：

WHILE $t\leq T$ DO

Step 1. 根据权重 D^t，求解弱分类器 h^t 获得其性能 ϵ^t

Step 2. 根据当前弱分类器性能 ϵ^t，更新 α^t

Step 3. 根据 α^t, h^t 计算归一化因子 $\tilde{Z}^t$：

$$\tilde{Z}^t=\sum_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\in\mathcal{X}_{\text{pair}}}D^t(\boldsymbol{x}_1,\boldsymbol{x}_2)\cdot\exp(-\alpha^t\cdot(h^t(\boldsymbol{x}_1)-h^t(\boldsymbol{x}_2))) \tag{11}$$

Step 4. 计算新一轮样本权重 D^t

$$D^{t+1}(\boldsymbol{x}_1,\boldsymbol{x}_2)=\frac{D^t(\boldsymbol{x}_1,\boldsymbol{x}_2)\cdot\exp(-\alpha^t\cdot(h^t(\boldsymbol{x}_1)-h^t(\boldsymbol{x}_2)))}{\tilde{Z}^t} \tag{12}$$

$t=t+1$

END WHILE

生成最终模型 $f(\boldsymbol{x})=\sum_{t=1}^{T}\alpha^t\cdot h^t(\boldsymbol{x})$

与 AdaBoost 算法相同[30]，RankBoost 算法在每次迭代中根据上一轮迭代产生的样本权重 D_t 确定一个新弱学习器 h^t 以及该学习器对应的模型权重 α^t. 随后，RankBoost 算法基于 h^t 及 α^t 为每个排序样例对 $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 确定下一次迭代的样本权重 $D^{t+1}(\boldsymbol{x}_1,\boldsymbol{x}_2)$. $D^{t+1}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 数值越大则该样例对对于下轮迭代模型越为重要. 由 $D^{t+1}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 更新过程可知，$h^t(\boldsymbol{x}_1)-h^t(\boldsymbol{x}_2)>0$ 模型输出与期望序关系一致，此时对应的权重较小；反之，若 $h^t(\boldsymbol{x}_1)-h^t(\boldsymbol{x}_2)<0$ 模型输出与期望序关系不一致，$D^{t+1}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 则有增大趋势. 因此，随着迭代不断进行 RankBoost 模型将逐步聚焦于难以正确排序的样例对，形成由易至难的自适应学习过程.

2.4　与既往工作的比较

首先从方法层面考虑. 相比于 2.1 小节及 2.3 小节中提及的工作，本文主要侧重于对半监督条件

下的 AUC 优化. 相比于 2.2 小节中提及的现有半监督 AUC 优化方法,本文将基于 Boosting 的模型集成技术引入半监督条件下的 AUC 优化问题中,后续理论及实验分析均体现出了本文算法的优势.

其次,从理论层面考虑. 本文首次给出了模型集成条件下半监督 AUC 优化的收敛速率保障及泛化界保障. 相比于文献[25]已有的结果,本文通过构造适用于半监督 AUC 优化的 Rademacher 复杂度(见附录 D.2 中的引理 11)及广义最大值不等式(见附录 D.2 中的引理 9)获得了更紧致的泛化误差上界.

3 半监督 AUC 优化的 Boosting 集成算法

本节针对(OP_4)及 RankBoost 算法设计高效的半监督 AUC 优化模型集成算法.

3.1 总览

沿用 2.2 小节中提及的理论完成半监督条件下的 AUC 替代损失估计. 由于本文采用 RankBoost 进行模型集成,因此将采用指数替代损失函数 $\ell_{\exp}(t)=\exp(-t)$. 记 $\hat{R}_{PN}^{\ell_{\exp}}$, $\hat{R}_{PNU}^{\ell_{\exp}}$, $\hat{R}_{UN}^{\ell_{\exp}}$ 为将式(6)、式(7)、式(8)中的 ℓ_{sq} 替换为 $\ell_{\exp}$ 所得到的经验风险函数,给定模型假设空间 $\mathcal{H}$,本节求解由指数替代损失诱导出的 AUC 优化问题:

$$(\mathrm{OP}_4)\ \min_{f\in\mathcal{H}}\left(\gamma\cdot\hat{R}_{PN}^{\ell_{\exp}}(f)+\frac{(1-\gamma)}{2}(\hat{R}_{PU}^{\ell_{\exp}}(f)+\hat{R}_{UN}^{\ell_{\exp}}(f))\right)\cdot \exp\left(\rho\cdot\sum_{i=1}^{T}\alpha^t\right),$$

其中 $\exp\left(\rho\cdot\sum_{t=1}^{T}\alpha^t\right)$ 为正则项,其作用可见定理 3 中的分析.

对于本文的目标函数而言,直接采用 RankBoost 会造成 $\mathcal{O}(|n_pn_u+n_pn_n+n_un_n|)$ 的时间及空间复杂度,基本正比于训练样本量规模的平方,计算效率较低. 鉴于此,本节将针对(OP_4)设计高效的模型集成方法,其细节见算法 1.

3.2 PNUAUCBoost 算法设计

样本权重解耦. 首先根据(OP_4)的形式,构造过程 1 中 Step 4 中 D^t 的解耦方法,即将样本对的权重 D^t 解耦为两样本权重的乘积,使空间复杂度由 $\mathcal{O}(n_p\cdot n_u+n_u\cdot n_n)$ 降至 $\mathcal{O}(n_p+n_u+n_n)$. 注意到(OP_4)中仅存在三类样例对,分别为 $(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)})$, $(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})$ 以及 $(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})$. 为同时将这三类样例对的权重解耦,构造辅助权重

$$\{\omega_{p,i}^{+,t}\}_{i=1}^{n_p},\ \{\omega_{u,j}^{-,t}\}_{j=1}^{n_u},\ \{\omega_{u,j}^{+,t}\}_{j=1}^{n_u},$$
$$\{\omega_{n,k}^{-,t}\}_{k=1}^{n_n},\ \{\nu_{p,j}^{+,t}\}_{i=1}^{n_p},\ \{\nu_{n,k}^{-,t}\}_{k=1}^{n_n},$$

并将 D^t 表示为

$$D^t(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)})=\omega_{p,i}^{+,t}\cdot\omega_{u,j}^{-,t},\quad i=1,\cdots,n_p,\ j=1,\cdots,n_u \tag{13}$$

$$D^t(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})=\omega_{u,j}^{+,t}\cdot\omega_{n,k}^{-,t},\quad j=1,\cdots,n_u,\ k=1,\cdots,n_n \tag{14}$$

$$D^t(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})=\nu_{p,j}^{+,t}\cdot\nu_{n,k}^{-,t},\quad i=1,\cdots,n_p,\ k=1,\cdots,n_n \tag{15}$$

基于上述解耦方式,进一步设计过程 1 的具体实现细节.

权重初始计算. 将初始值设为

$$\omega_{p,i}^{+,0}=\frac{C_1}{n_p},\ \omega_{u,j}^{+,0}=\frac{C_1}{n_u},\ \omega_{u,j}^{-,0}=\frac{C_1}{n_u},$$
$$\omega_{n,k}^{-,0}=\frac{C_1}{n_n},\ \nu_{p,j}^{+,0}=\frac{C_2}{n_p},\ \nu_{n,k}^{-,0}=\frac{C_2}{n_n} \tag{16}$$

其中 $C_1=\left(\frac{1-\gamma}{2}\right)^{\frac{1}{2}}$, $C_2=(\gamma)^{\frac{1}{2}}$. 此时对应的 $D^0(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})$ 恰为 $\frac{\gamma}{n_pn_n}$, $D^0(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)})$, $D^0(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})$ 恰为 $\frac{1-\gamma}{2n_pn_u}$, $\frac{1-\gamma}{2n_un_n}$. 记 R_{OP_4} 为(OP_4)的目标函数,有:

$$\begin{aligned}R_{\mathrm{OP}_4}=&\left(\sum_{i=1}^{n_p}\sum_{j=1}^{n_p}D^0(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)})\exp(f(\boldsymbol{x}_j^{(u)})-f(\boldsymbol{x}_i^{(p)}))\right)+\\&\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}D^0(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})\exp(f(\boldsymbol{x}_k^{(n)})-f(\boldsymbol{x}_j^{(u)}))+\\&\sum_{j=1}^{n_p}\sum_{k=1}^{n_n}D^0(\boldsymbol{x}_j^{(p)},\boldsymbol{x}_k^{(n)})\exp(f(\boldsymbol{x}_k^{(n)})-f(\boldsymbol{x}_j^{(p)}))\cdot\\&\exp\left(\rho\cdot\sum_{i=1}^{T}\alpha^t\right)\end{aligned} \tag{17}$$

因此上述初始权重设置方法可将算法中的样本权重引入目标函数中.

归一化因子的高效计算. 进一步根据式(13)和式(15)中的解耦规则,实现并加速过程 1 中式(11)的计算. 对于 $\hat{R}_{PN}^{\ell_{\exp}}$ 相关计算,有:

$$\begin{aligned}&\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}D^t(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})\exp(-\alpha^t(h^t(\boldsymbol{x}_i^{(p)})-h^t(\boldsymbol{x}_k^{(n)})))\\&=\left(\sum_{i=1}^{n_p}\nu_{p,j}^{+,t}\cdot\exp(-\alpha^t\cdot h^t(\boldsymbol{x}_i^{(p)}))\right)\cdot\\&\quad\left(\sum_{k=1}^{n_n}\nu_{n,k}^{-,t}\cdot\exp(\alpha^t\cdot h^t(\boldsymbol{x}_k^{(n)}))\right).\end{aligned}$$

对于 $\hat{R}_{PU}^{\ell_{\exp}}$ 及 $\hat{R}_{UN}^{\ell_{\exp}}$,易得类似形式的分解. 鉴于此,可将归一化因子的计算转化为样例归一化因子的乘积之和. 为简化数学表达,于下文中采用以下简写

符号：

$h^t_{p,i}=h^t(\boldsymbol{x}^{(p)}_i)$，$h^t_{u,j}=h^t(\boldsymbol{x}^{(u)}_j)$，$h^t_{n,k}=h^t(\boldsymbol{x}^{(n)}_k)$，则该分解过程可表为

$$Z^t_1=\sum_{i=1}^{n_p}\omega^{+,t-1}_{p,i}\cdot\exp(-\alpha h^t_{p,i}) \tag{18}$$

$$Z^t_2=\sum_{j=1}^{n_u}\omega^{-,t-1}_{u,j}\cdot\exp(\alpha h^t_{u,j}) \tag{19}$$

$$Z^t_3=\sum_{j=1}^{n_u}\omega^{+,t-1}_{u,j}\cdot\exp(-\alpha h^t_{u,j}) \tag{20}$$

$$Z^t_4=\sum_{k=1}^{n_n}\omega^{-,t-1}_{n,k}\cdot\exp(\alpha h^t_{n,k}) \tag{21}$$

$$Z^t_5=\sum_{i=1}^{n_p}\nu^{+,t-1}_{p,j}\cdot\exp(-\alpha h^t_{p,i}) \tag{22}$$

$$Z^t_6=\sum_{k=1}^{n_n}\nu^{-,t-1}_{n,k}\cdot\exp(\alpha h^t_{n,k}) \tag{23}$$

$$\widetilde{Z}^t=Z^t_1\cdot Z^t_2+Z^t_3\cdot Z^t_4+Z^t_5\cdot Z^t_6 \tag{24}$$

样本权重更新. 与归一化因子 $\widetilde{Z}^t$ 的化简相似，可根据式(13)～式(15)中的解耦将过程 1 式(12)中的权重更新过程转化为对解耦权重的更新过程：

$$\omega^{+,t}_{p,i}=\frac{\omega^{+,t-1}_{p,i}\exp(-\alpha h^t_{p,i})}{\sqrt{\widetilde{Z}^t}} \tag{25}$$

$$\omega^{-,t}_{u,j}=\frac{\omega^{-,t-1}_{u,j}\exp(\alpha h^t_{u,j})}{\sqrt{\widetilde{Z}^t}} \tag{26}$$

$$\omega^{+,t}_{u,j}=\frac{\omega^{+,t-1}_{u,j}\exp(-\alpha h^t_{u,j})}{\sqrt{\widetilde{Z}^t}} \tag{27}$$

$$\omega^{-,t}_{n,k}=\frac{\omega^{-,t-1}_{n,k}\exp(\alpha h^t_{n,k})}{\sqrt{\widetilde{Z}^t}} \tag{28}$$

$$\nu^{+,t}_{p,j}=\frac{\nu^{+,t-1}_{p,j}\exp(-\alpha h^t_{p,i})}{\sqrt{\widetilde{Z}^t}} \tag{29}$$

$$\nu^{-,t}_{n,k}=\frac{\nu^{-,t-1}_{n,k}\exp(\alpha h^t_{n,k})}{\sqrt{\widetilde{Z}^t}} \tag{30}$$

模型权重 $\boldsymbol{\alpha^t}$ 的更新. 首先通过以下引理给出损失函数的一个上界：

引理 1.　当算法 T 次迭代结束后，有：

$$R_{\mathrm{OP}_4}=\prod_{t=1}^{T}(\exp(\rho\alpha^t)\cdot\widetilde{Z}^t).$$

证明.　见附录 B.　　证毕.

由以上引理，若在第 t 次迭代使 $\exp(\rho\alpha^t)\cdot\widetilde{Z}^t$ 尽可能小，则最终将获得较为理想的目标函数值. 因此基于近似最小化 $\exp(\rho\alpha^t)\cdot\widetilde{Z}^t$ 的原则设计 α^t，h^t 的更新方式. 考虑如下放缩，利用 exp 的凸性及 Jensen 不等式，$\forall x\in[-1,1]$ 有：

$$\exp(\alpha\cdot x)\leq\frac{1+x}{2}\cdot\exp(\alpha)+\frac{1-x}{2}\cdot\exp(-\alpha) \tag{31}$$

将弱分类器 h^t 的输出值域限制于区间 $[0,1]$ 内，根据 $\widetilde{Z}^t$ 定义及式(32)，可得：

$$\begin{aligned}\widetilde{Z}^t\leq&\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}D^t(\boldsymbol{x}^{(p)}_i,\boldsymbol{x}^{(u)}_j)\widetilde{\psi}_{i,j}+\\&\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}D^t(\boldsymbol{x}^{(u)}_j,\boldsymbol{x}^{(n)}_k)\widetilde{\psi}_{j,k}+\\&\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}D^t(\boldsymbol{x}^{(p)}_i,\boldsymbol{x}^{(n)}_k)\widetilde{\psi}_{i,k}\\\triangleq&C(\alpha^t)\end{aligned} \tag{32}$$

其中：

$$\begin{aligned}\psi_{i,j}=&\frac{1+h^t_{p,i}-h^t_{u,j}}{2}\cdot\exp(-\alpha)+\\&\frac{1+h^t_{u,j}-h^t_{p,i}}{2}\cdot\exp(\alpha),\\\widetilde{\psi}_{j,k}=&\frac{1+h^t_{u,j}-h^t_{n,k}}{2}\cdot\exp(-\alpha)+\\&\frac{1+h^t_{n,k}-h^t_{u,j}}{2}\cdot\exp(\alpha),\\\widetilde{\psi}_{i,k}=&\frac{1+h^t_{p,i}-h^t_{n,k}}{2}\cdot\exp(-\alpha)+\\&\frac{1+h^t_{n,k}-h^t_{p,i}}{2}\cdot\exp(\alpha).\end{aligned}$$

根据上式对 $\widetilde{Z}^t$ 的放缩，固定弱分类器模型 h^t，并求解 α^t 使 $\exp(\rho\cdot\alpha^t)C(\alpha^t)$ 最小化. 为便于数学上的表达，记：

$$\begin{aligned}\Delta^t=&\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{D^t(\boldsymbol{x}^{(p)}_i,\boldsymbol{x}^{(u)}_j)}{2}\cdot(h^t_{p,i}-h^t_{u,j})+\\&\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{D^t(\boldsymbol{x}^{(u)}_j,\boldsymbol{x}^{(n)}_k)}{2}\cdot(h^t_{u,j}-h^t_{n,k})+\\&\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{D^t(\boldsymbol{x}^{(p)}_i,\boldsymbol{x}^{(n)}_j)}{2}\cdot(h^t_{p,i}-h^t_{n,k})\end{aligned} \tag{33}$$

易证 $\exp(\rho\cdot\alpha^t)\cdot C(\alpha^t)$ 为关于 α^t 的凸函数，因此为最小化该因变量仅需求解：

$$\frac{\mathrm{d}[\exp(\rho\cdot\alpha^t)C(\alpha^t)]}{\mathrm{d}\alpha^t}=0,$$

可解得 α^t 为

$$\alpha^t=\frac{1}{2}\log\left(\frac{1+\Delta^t}{1-\Delta^t}\right)-\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) \tag{34}$$

$\boldsymbol{\Delta^t}$ 的高效计算. 在计算 α^t 过程中需要计算 Δ^t 进而需要遍历所有的样例对权重 D^t. 因此 Δ^t 的计算过程也是本算法的主要计算瓶颈之一. 鉴于上文中提出的权重解耦方法，同样可以给出 Δ^t 的加速计算方式. 由式(13)～式(15)，有：

$$\Delta^t=\sum_{i=1}^{n_p}g^t_{p,i}\cdot h^t_{p,i}+\sum_{j=1}^{n_u}g^t_{u,j}\cdot h^t_{u,j}+\sum_{k=1}^{n_n}g^t_{n,k}\cdot h^t_{n,k},$$

其中：

$$g_{p,i}^{t}=\omega_{p,i}^{+,t}\cdot\Big(\sum_{j=1}^{n_u}\omega_{u,j}^{-,t}\Big)+\nu_{p,j}^{+,t}\cdot\Big(\sum_{k=1}^{n_n}\nu_{n,k}^{-,t}\Big) \tag{35}$$

$$g_{u,j}^{t}=\omega_{u,j}^{+,t}\cdot\Big(\sum_{k=1}^{n_n}\omega_{n,k}^{-,t}\Big)-\omega_{u,j}^{-,t}\cdot\Big(\sum_{i=1}^{n_p}\omega_{p,i}^{+,t}\Big) \tag{36}$$

$$g_{n,k}^{t}=-\omega_{n,k}^{-,t}\cdot\Big(\sum_{j=1}^{n_u}\omega_{u,j}^{+,t}\Big)-\nu_{n,k}^{-,t}\cdot\Big(\sum_{k=1}^{n_n}\nu_{p,j}^{+,t}\Big) \tag{37}$$

注意到上式中所有的权重求和项仅需计算一次，因此仅需$\mathcal{O}(N)$时间复杂度即可完成所有 $g_{p,i}^{t}$，$g_{u,j}^{t}$，$g_{n,k}^{t}$的计算. 为便于后续计算，记 $\boldsymbol{G}_t\in\mathbb{R}^{(n_p+n_u+n_n)\times 1}$为所有 $g_{p,i}^{t}$，$g_{u,j}^{t}$，$g_{n,k}^{t}$拼接而成的列向量，并记 $\boldsymbol{H}_t\in\mathbb{R}^{(n_p+n_u+n_n)\times 1}$为所有 $h_{p,i}^{t}$，$h_{u,j}^{t}$，$h_{n,k}^{t}$拼接而成的列向量. 根据符号 $\boldsymbol{G}^t$，$\boldsymbol{H}^t$，可将 Δ^t进一步简写为

$$\Delta^t=(\boldsymbol{G}^t)^{\top}\boldsymbol{H}^t \tag{38}$$

弱分类器模型 h^t的更新. 给定 α^t，Δ^t，进一步反推出弱分类器 h^t的一个合理更新规则. 为此，构建以下引理.

引理 2. 对于第 t 次算法迭代，若由式(34)更新 α^t，有：

$$\exp(\rho\alpha^t)\cdot\widetilde{Z}^t\leq\exp(\rho\alpha^t)\cdot C(\alpha^t)=\exp\Big(-KL\Big(\frac{1+\rho}{2}\Big\|\frac{1+\Delta^t}{2}\Big)\Big),$$

其中 KL 为二元相对熵，其定义为

$$KL(p\|q)=p\cdot\log\Big(\frac{p}{q}\Big)+(1-p)\cdot\log\Big(\frac{1-p}{1-q}\Big),\quad\forall p\in[0,1],\ q\in[0,1].$$

证明. 见附录 C. 证毕.

由引理 2 可知，在通过式(34)更新 α^t情况下可通过最大化 $KL((1+\rho)/2\|(1+\Delta^t)/2)$近似实现 $\widetilde{Z}^t$的最小化. 与 RankBoost 相同，本文采用决策树桩[10](Decision Stump)学习弱分类器，并将文献[10]中的弱分类器目标函数替换为 $KL((1+\rho)/2\|(1+\Delta^t)/2)$. 给定样本输入 $\boldsymbol{x}=[x^1,x^2,\cdots,x^d]^{\top}$，选定特征维度 e 及阈值 θ，决策树桩函数输出如下：

$$h_{\theta}^{e}(\boldsymbol{x})=\begin{cases}1, & x^e>\theta\\ 0, & \text{其他}\end{cases}.$$

将决策树桩的输出函数代入式(38)，得其对应的 Δ^t可表为

$$\Delta^t=\sum_{x_i^e>\theta}g_i^t,$$

其中 i，θ 为待学习超参数，i 为决策树桩选定的参数决策的输入维度，θ 为其选定阈值. 由上式可知，决策树桩输出为 1 仅当给定样例的第 i 维输入数值大于阈值 θ. 为求解决策树桩的两个参数，为输入特征的每一维度 i 设定阈值备选集 $\boldsymbol{T}_i$，并搜索能够最大化 $KL((1+\rho)/2\|(1+\Delta^t)/2)$的 θ，i，最终生成弱分类器 h^t. 算法细节可见附录 A 中的算法 2.

综合上述所有细节，得到高效的 PNU-AUC 的 Boosting 算法，并将其汇总于算法 1.

算法 1. PNUAUCBoost.

输入：模型输入 $\boldsymbol{X}$，超参数 $\gamma\in[0,1]$，弱分类器个数 T，超参数 ρ

输出：弱分类器权重 $\alpha^1,\cdots,\alpha^T$，弱分类器 $h^1,\cdots,h^T$通过式(16)初始化权重

WHILE $t\leq T$ DO

根据式(35)～式(37)完成 $\boldsymbol{G}^t$计算

根据式(35)～式(37)完成 $\boldsymbol{H}^t$计算

根据算法 2 获得弱分类器 h^t并获得其对应的 Δ^t

更新 α^t：

$$\alpha^t=\frac{1}{2}\log\Big(\frac{1+\Delta^t}{1-\Delta^t}\Big)-\frac{1}{2}\log\Big(\frac{1+\rho}{1-\rho}\Big)$$

由式(18)～式(24)计算归一化因子 $\widetilde{Z}^t$

由式(25)～式(30)更新权重

$t=t+1$

END WHILE

生成最终模型 $f(\boldsymbol{x})=\sum_{i=1}^{T}\alpha^t\cdot h^t(\boldsymbol{x})$

4 理论分析

4.1 总体概述

本节将通过对收敛率及泛化界的联合控制构造最终的泛化性能上界. 本节涉及三个定理，其中定理 1、定理 2 提供收敛及泛化界的中间结果；定理 3 提供最终结果.

在本节证明中需要引入如下辅助函数：

(1) 定义 ρ-错排率($\rho>0$)为

$$r_{\rho}(f)=\frac{\gamma}{n_pn_n}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\boldsymbol{I}\Big[\frac{f(\boldsymbol{x}_k^{(n)})-f(\boldsymbol{x}_i^{(p)})}{\sum_{i=1}^{T}\alpha^t}\leq\rho\Big]+\frac{\gamma}{n_pn_u}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\boldsymbol{I}\Big[\frac{f(\boldsymbol{x}_j^{(u)})-f(\boldsymbol{x}_i^{(p)})}{\sum_{i=1}^{T}\alpha^t}\leq\rho\Big]+\frac{\gamma}{n_un_n}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\boldsymbol{I}\Big[\frac{f(\boldsymbol{x}_k^{(n)})-f(\boldsymbol{x}_j^{(u)})}{\sum_{i=1}^{T}\alpha^t}\leq\rho\Big] \tag{39}$$

(2) 相比普通的错排率，ρ-错排率将错误排序中的分差小于 0 替换为小于 ρ，因此相比一般错误率更为严格. 定理 1 将通过 $r_{\rho}(f)$作为变量给出 $r_{\rho}(f)$的收敛率. 在最终结论定理 3 中将通过 $r_{\rho}(f)$作为桥梁构造误差上界.

(3) 定义 margin 损失函数为

$$\ell_\rho(\boldsymbol{x})=\min\left(1,\max\left(0,1-\frac{\boldsymbol{x}}{\rho}\right)\right).$$

定义 margin $\boldsymbol{x}=y\cdot f(\boldsymbol{x})$，该损失仅在 $\boldsymbol{x}\in(-\infty,0)$ 恒为 1，在 $\boldsymbol{x}\in[0,\rho]$ 区间单调递减，在 $\boldsymbol{x}\in[\rho,+\infty)$ 区间恒定为 0，相比一般的损失函数. 在定理 2 中，可利用 $\ell_\rho(\boldsymbol{x})$ 的性质诱导出 C/ρ 的泛化界，其中 C 取决于模型复杂度.

易得：

$$\ell_{0\text{-}1}(\boldsymbol{x})\leq\ell_\rho(\boldsymbol{x})\leq\boldsymbol{I}[\boldsymbol{x}\leq\rho]\leq\exp(-\boldsymbol{x}+\rho)\quad(40)$$

由此可得以下两个关键不等式，将在本节中对不同变量进行转换：

$$\hat{R}^{PNU}_{\rho,\mathcal{S}}(f)\leq r_\rho\leq R_{\mathrm{OP}_4}\quad(41)$$

以及

$$\mathbb{E}_{\mathcal{S}}[\hat{R}^{PNU}_{0\text{-}1,\mathcal{S}}]\leq\mathbb{E}_{\mathcal{S}}[\hat{R}^{PNU}_{\rho,\mathcal{S}}(f)]\quad(42)$$

4.2　收敛分析

在本小节中，给出算法 1 的收敛速率. 首先，给出相对熵的关键数学性质：

引理 3.　$\forall p\in(0,1),q\in(0,1)$ 有：

$$KL(p\|q)\geq 2\cdot(p-q)^2.$$

证明.　将文献[31]中的定理 1.3 应用于二项分布即可获得本引理.

注意到 $r(f)\leq R_{\mathrm{OP}_4}$，根据引理 1～3，可给出 $R^{0\text{-}1}_{PN}(f)$ 及 $r_\rho(f)$ 的上界：

定理 1.　给定 $\rho>0$，设算法 1 在 T 轮迭代结束后所得到的分类器为 $f(\boldsymbol{x})=\sum_{i=1}^{T}\alpha^t\cdot h^t(\boldsymbol{x})$，$\alpha^t\geq 0$，$\forall\alpha^t\geq 0$，$\forall t$[①]，如下结论成立：

(1) 对于一般情况：

$$R^{0\text{-}1}_{PN}(f)\leq r_\rho\leq R_{\mathrm{OP}_4}\leq\exp\left(-\sum_{i=1}^{T}KL\left(\frac{1+\rho}{2}\middle\|\frac{1+\Delta^t}{2}\right)\right).$$

(2) 若进一步假设 $\min_t\Delta^t\geq\delta>\rho$，有：

$$R^{0\text{-}1}_{PN}(f)\leq r_\rho\leq R_{\mathrm{OP}_4}\leq\exp\left(-\frac{T}{2}\cdot(\delta-\rho)^2\right).$$

证明.　结论(1)可由式(41)及引理 1～2 推知；在结论(1)基础上进一步应用引理 3 即可得结论(2).　　证毕.

上述定理表明，在弱分类器性能较为理想条件下可有 $\min_t\Delta^t\geq\delta>\rho$，此时算法收敛具有指数速率. 换言之，指数收敛速率依赖于结论(2)中的假设，即所有的弱分类器的性能有一致下界，当所有弱分类器性能均较为理想时，该假设较易满足. 若该假设不成立，则算法的收敛与可达成下界的有效弱分类器个数具有指数型关系.

4.3　泛化性能分析

本小结中，继续讨论算法 1 的泛化性能. 首先，将泛化性能定义为样本分布下的期望 AUC 错误率，即 $R^{0\text{-}1}_{PN}$. 在此基础上，希望对于任意由算法 1 产生的学习器，下式以大概率成立：

$$R^{0\text{-}1}_{PN}\leq\epsilon_1+\epsilon_2\quad(43)$$

其中 ϵ_1 与训练集上的模型排序错误率有关，ϵ_2 与算法产生模型的丰富程度以及训练集的样本量有关. 因此，当训练样本充分且模型优化充分的条件下保证上式成立即可保证泛化误差 $R^{0\text{-}1}_{PN}$ 较小. 本文将通过学习论[32]中基于 Rademacher 复杂度的分析方式完成上式的证明. 其证明思路关键在于通过对称化技术(见文献[32]定理 3.3)引入假设空间的 Rademacher 复杂度. 传统对称化技术利用样本损失的独立同分布性，通过 Rademacher 随机变量的期望刻画两个不同数据集上的损失之差. 进一步考察半监督 AUC 损失的定义形式. 不难发现涉及样例对 $(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)})$ 的损失项与所有涉及 $\boldsymbol{x}_i^{(p)}$ 或 $\boldsymbol{x}_j^{(u)}$ 的损失项都无法独立，同理可知，样例对 $(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})$ 及 $(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})$ 同样存在该问题. 鉴于此，传统的对称化技术对 AUC 损失失效.

本文通过引理 11(见附录 D.5)给出了适用于半监督 AUC 的对称化技术，其主要思路在于通过二元形式设计 Rachmacher 随机变量与损失之间的复合关系，从而通过定义更为复杂的 Rademacher 随机变量期望刻画不同数据集上的经验损失差. 首先，引入如下适用于 PNU-AUC 问题的 Rademacher 复杂度：

定义 1.　PNU-AUC Rademacher 复杂度. 给定数据集 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m$ 及假设集 $\mathcal{H}$，损失函数 ℓ PNU-AUC Rademacher 经验复杂度由下式给出：

$$\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell\circ\mathcal{H})=\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}Q_{i,k}+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}Q_{j,k}+\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}Q_{i,j}\right],$$

其中

$$Q_{i,k}=\frac{\sigma_i^{(p)}+\sigma_j^{(u)}}{2}\cdot\frac{\gamma}{n_pn_n}\cdot\ell(f,\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)}),$$

$$Q_{j,k}=\frac{\sigma_j^{(u)}+\sigma_k^{(n)}}{2}\cdot\frac{1-\gamma}{2n_un_n}\cdot\ell(f,\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)}),$$

$$Q_{i,j}=\frac{\sigma_i^{(p)}+\sigma_j^{(u)}}{2}\cdot\frac{1-\gamma}{2n_pn_u}\cdot\ell(f,\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)}).$$

变量 $\sigma_i^{(p)},\forall i=1,2,\cdots,n_p$，$\sigma_j^{(u)},\forall j=1,2,\cdots,n_u$，$\sigma_k^{(n)}$，

① $\alpha^t\geq 0$ 可通过设置较小的 ρ 或早停机制实现.

$\forall k=1,2,\cdots,n_n$ 为独立同分布 Rademacher 随机变量①. 在经验复杂度基础上,定义总体水平的 PNU-AUC Rademacher 复杂度为其经验复杂度对 $\mathcal{S}$ 的期望,表示为

$$\mathfrak{R}_{PNU}(\ell\circ\mathcal{H})=\mathbb{E}_{\mathcal{S}}[\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell\circ\mathcal{H})].$$

进一步,给出决策树桩弱分类器的假设空间定义.

定义 2. 决策树桩的假设空间. 将决策树桩的假设空间定义为:$\mathcal{H}_{DS}=\{h_\theta^e:e\in[d],\theta\in\boldsymbol{T}^e\}$,其中 $h_\theta^e(\boldsymbol{x})=\boldsymbol{I}[\boldsymbol{x}^e>\theta]$,$[d]=\{1,2,3,\cdots,d\}$,$\boldsymbol{T}^e$ 为第 e 个维度特征的备选阈值集合,对于每个维度的阈值候选集,固定其阈值备选集 $\boldsymbol{T}^i$,并固定备选阈值个数为 K.

在证明过程中,需要利用假设空间的凸包及其 Rademacher 复杂度的数学性质,其定义如下:

定义 3. 假设空间的凸包. 给定任意假设集 $\mathcal{H}$,定义其凸包为 $co(\mathcal{H})$ 为如下函数集:

$$co(\mathcal{H})=\Big\{h:\exists T\in\mathbb{N},\ \text{s.t.}\ h=\sum_{i=1}^{T}\alpha^t\cdot h^t,h^t\in\mathcal{H},\ \alpha^t\geq0,\forall t\in[T],\sum_{i=1}^{T}\alpha^t=1\Big\}.$$

基于上述定义、不等式以及附录 D 中的引理 4~引理 11,首先给出 $\mathcal{H}_{DS}$ 凸包中函数的泛化界:

定理 2. PNU-AUCBOOST 的泛化界. 给定训练数据集 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$、$\rho>0$,及弱分类器的阈值备选集 $\boldsymbol{T}_e$,设样例均由独立采样生成,则对于任意函数 $f\in co(\mathcal{H}_{DS})$ 以及任意 $\delta\in(0,1)$,下式至少以 $1-\delta$ 概率成立:

$$R_{PN}^{0\text{-}1}(f)\leq\Big(r_\rho(f)+\frac{8\sqrt{2}}{\rho}((\log d+\log K)\cdot\chi(\boldsymbol{Y}))^{\frac12}+\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\chi(\boldsymbol{Y})\Big)^{\frac12}\Big)\cdot\frac{1}{1+\gamma},$$

其中 $\chi(\boldsymbol{Y})=\frac{1}{n_p}+\frac{1}{n_u}+\frac{1}{n_n}$.

证明. 根据引理 12,有下式:

$$\mathbb{E}_{\mathcal{S}}\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)\leq\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)+\frac{8\sqrt{2}}{\rho}((\log d+\log K)\cdot\chi(\boldsymbol{Y}))^{\frac12}+\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\chi(\boldsymbol{Y})\Big)^{\frac12}\tag{44}$$

进一步由式(42),有下式以至少 $1-\delta$ 概率成立:

$$R_{0\text{-}1}^{PNU}(f)\leq r_\rho+\frac{8\sqrt{2}}{\rho}((\log d+\log K)\cdot\chi(\boldsymbol{Y}))^{\frac12}+\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\chi(\boldsymbol{Y})\Big)^{\frac12}\tag{45}$$

综合式(1)及式(2),此时有:

$$\begin{aligned}R_{0\text{-}1}^{PNU}(f)&=\gamma\cdot R_{PN}^{0\text{-}1}+\frac{1-\gamma}{2}(R_{PU}^{0\text{-}1}+R_{UN}^{0\text{-}1})\\&=\gamma\cdot R_{PN}^{0\text{-}1}+\frac{1-\gamma}{2}\Big(R_{PN}^{0\text{-}1}+\frac12\Big)\\&\geq\frac{1+\gamma}{2}R_{PN}^{0\text{-}1}\end{aligned}\tag{46}$$

综合式(45)及式(46),定理得证. 证毕.

下面基于定理 2 考察由算法 1 输出的学习器. 记:

$$\mathcal{A}=\Big\{f:f(\boldsymbol{x})=\sum_{i=1}^{T}\alpha^t\cdot h^t,\alpha^t\geq0,h^t\in\mathcal{H}_{DS},T\in\mathbb{N}_+\Big\}$$

为模型权重非负条件下(见本文第 8 页脚注①),所有可能由算法 1 产生的模型集合. 通过以下定理将 $co(\mathcal{H}_{DS})$ 上的结论推广至假设空间 $\mathcal{A}$.

定理 3. 给定训练数据集 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$、$\rho>0$ 及弱分类器的阈值备选集 $\boldsymbol{T}_e$,设样例均由独立采样生成. 对于算法 1 在 T 轮迭代后的任意输出函数 $f(\boldsymbol{x})\in\mathcal{A}$,以及任意 $\delta\in(0,1)$,下式至少以 $1-\delta$ 概率成立:

$$R_{PN}^{0\text{-}1}(f)\leq\Big(\exp\Big(-\sum_{i=1}^{T}KL\Big(\frac{1+\rho}{2}\Big\|\frac{1+\Delta^t}{2}\Big)\Big)+\frac{8\sqrt{2}}{\rho}((\log d+\log K)\cdot\chi(\boldsymbol{Y}))^{\frac12}+\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\chi(\boldsymbol{Y})\Big)^{\frac12}\Big)\cdot\frac{1}{1+\gamma},$$

另外对于任意满足 $\min_t\Delta^t\geq\delta>\rho$ 的输出函数 f,有:

$$R_{PN}^{0\text{-}1}(f)\leq\Big(\exp\Big(-\frac{T}{2}\cdot(\delta-\rho)^2\Big)+\frac{8\sqrt{2}}{\rho}((\log d+\log K)\cdot\chi(\boldsymbol{Y}))^{\frac12}+\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\chi(\boldsymbol{Y})\Big)^{\frac12}\Big)\cdot\frac{1}{1+\gamma}.$$

证明. 给定 $f=\sum_{i=1}^{T}\alpha^t\cdot h^t\in\mathcal{A}$,构造 $\widetilde{f}=\frac{f}{\sum_{i=1}^{T}\alpha^t}$,显然 $\widetilde{f}\in co(\mathcal{H}_{DS})$ 且 $r_\rho(\widetilde{f})=r_\rho(f)$,$R_{PN}^{0\text{-}1}(f)=R_{PN}^{0\text{-}1}(\widetilde{f})$,因此对 $\mathcal{A}$ 的泛化界即可转化为对 $\mathcal{F}=\{\widetilde{f}:f\in\mathcal{A}\}$ 的泛化界,又因为 $\mathcal{F}\subseteq co(\mathcal{H}_{DS})$,可直接应用定理 2 结论. 综合定理 1 及定理 2 结论,本定理即可得证.

注 1. 对于本定理,有以下结论:

(1)考察定理中两个不等式的基本形式,左端项为有监督情况下的泛化 0-1 误差,右端项中弱分

① Rademacher 随机变量 σ 在 $[-1,1]$ 内随机取值且 $\mathbb{P}[\sigma=1]=\mathbb{P}[\sigma=-1]=\frac12$.

类器个数 T 有关的一项大致随 T 增大以指数形式衰减；右端项中与 T 无关项组成

$$\mathcal{O}\left(\left(\frac{1}{n_p}+\frac{1}{n_u}+\frac{1}{n_n}\right)^{\frac{1}{2}}\right)$$

的残差项，随训练样本个数增大而衰减. 因此本定理即可证明式(43)成立，也即当样本充分时，在半监督条件下运行算法 1 即可有效降低有监督情况下的泛化 0-1 误差. 因此证明了本文方法的可行性.

(2) 相比于目前已有的

$$\mathcal{O}\left(\left(\frac{1}{n_p}\right)^{\frac{1}{2}}+\left(\frac{1}{n_u}\right)^{\frac{1}{2}}+\left(\frac{1}{n_n}\right)^{\frac{1}{2}}\right)$$

半监督 AUC 泛化界[25]，借助文本所提出的广义最大值不等式(引理 9)获得了量级更小的上界.

(3) 本定理中的两个上界均为 T 的单减不增函数. 因此本文算法的另一个优势在于增加弱分类器个数在提升训练集性能的同时并不会带来明显的过拟合风险.

(4) 定理中的两个不等式右端项与 T 无关部分反比于 ρ，该性质表明引入 ρ 可降低泛化误差上界，为所提出算法引入正则项 $\exp\left(\rho\cdot\sum_{i=1}^{T}\alpha^t\right)$ 提供了理论依据.

5 实　验

5.1 数据集

为验证模型的有效性，本文在 16 个常见的二分类数据集上进行实验. 数据集主要来源于 keel①、UCI②、Kaggle③ 和 Libsvm④. 其中 keel 包含的数据集包括 vehicle2、vehicle0、pima、ring、phoneme；UCI 包含的数据集包括 credit-g、glass1、wisconsin、wdbc、shuttle-c0-vs-c4、monk-2、sonar；Kaggle 包含的数据集包括 Surgical-deepnet、insurance 和 numerai；Libsvm 包含的数据集包括 cod-rna. 所有数据集的详细信息如表 1 所示. 为验证本文算法在不同特性数据集上的表现，本文选用的数据集涵盖了较为宽泛的数据集规模(208～381 000)、不平衡比例(0.19～0.97)以及应用场景(金融、保险、交通、医学、生物信息等). 同时，为了测试模型在复杂数据集中的表现情况，本文在 2016 年于 Kaggle 发布的加密股市数据预测竞赛⑤使用的公开 Numerai 数据集上进行实验. Numerai 数据集是基于对冲基金建立的股票数据集，该数据集采用了同态加密，所有的属性均被归一化，特征含义被隐藏，预测难度较大.

表 1　数据集描述

数据集	数据来源	描述	样本数量	特征数量	正负样本比
pima	KEEL	糖尿病人区分数据	768	8	0.5360
ring	KEEL	多元正态分布数据集	7400	20	0.9807
phoneme	KEEL	选手发音数据集	5404	5	0.4154
vehicle0	KEEL	2D 交通工具图像数据集	846	18	0.3075
vehicle2	KEEL	2D 交通工具图像数据集	846	18	0.3471
credit-g	UCI	德国民众信用预测数据集	1000	20	0.4285
glass1	UCI	玻璃杯种类预测数据集	218	9	0.5507
wisconsin	UCI	乳腺癌区分数据集	683	9	0.5382
wdbc	UCI	乳腺肿块区分数据集	569	30	0.5938
shuttle-c0-vs-c4	UCI	航班飞行数据集	58000	9	0.0720
monk-2	UCI	MONK 学习算法竞赛数据集	432	7	0.8947
sonar	UCI	声纳探测岩石数据集	208	60	0.8738
Surgical-deepnet	Kaggle	外科诊疗记录数据集	14600	25	0.3371
insurance	Kaggle	健康护理保险数据集	381000	11	0.1959
numerai	Kaggle	金融股票数据集	96300	21	0.9753
cod-rna	Libsvm	RNA 序列检测数据集	271617	8	0.5000

5.2 对比方法

为验证本文所提出算法的有效性，在实验中采用以下对比方法与本文算法进行对比：

RBAUC(Boosting 集成算法＋AUC 优化). 该对比方法采用 RankBoost[28] 模型进行 AUC 优化，相比于本文提出算法，其训练过程仅利用训练集中所有已标记正负样本，而对未标记数据不加以利用.

PNUAB(半监督学习＋Boosting 集成算法). 该对比方法基于 AdaBoost[30] 模型设计，为了适应半监督实验设定，本文将未标记数据进行复制，一份视作未标记的正样本，一份视作未标记的负样本. 正样

① https://sci2s.ugr.es/keel/datasets.php
② http://archive.ics.uci.edu/ml/datasets
③ https://www.kaggle.com
④ https://www.csie.ntu.edu.tw/～cjlin/libsvmtools/datasets/
⑤ https://www.kaggle.com/numerai/encrypted-stock-market-data-fromnumerai

本、未标记正样本、负样本和未标记负样本的样本权重分别设为$\frac{1-\gamma}{2n_P}$、$\frac{\gamma}{2n_U}$、$\frac{\gamma}{2n_N}$和$\frac{1-\gamma}{2n_U}$，其中n_P、n_U和n_N分别为正样本、未标记样本和负样本的数量，$\gamma\in[0,1]$为实验中的超参数. 相比于本文所提出的算法，相比于本文提出的算法，该算法没有直接利用AUC优化训练模型.

PNU-AUC[25]（单模型半监督AUC优化算法）. 该方法同时利用正负样本以及未标记数据进行AUC优化，由于需要估计未标记样本中正负样本的比例，本文将未标记样本的真实正负样本比例作为输入，采用公开的代码①进行实验. 相比于本文提出算法，该算法仅采用单个模型进行半监督AUC优化，未采用模型集成技术.

Samult[26]（单模型半监督AUC优化算法）. 该对比方法同时利用正负样本以及未标记数据进行AUC优化而无需估计未标记样本中正负样本的比例，本文对Samult模型进行复现从而进行实验. 相比于本文提出算法，该算法仅采用单个模型进行半监督AUC优化，未采用模型集成技术.

LSAUC（基线算法）. LSAUC采用平方损失作为替代损失，替代优化问题的解由闭式解直接得出，同时该算法仅利用训练集中所有已标记正负样本，对未标记数据不加以利用. 由于LSAUC算法既未利用模型集成技术也未利用半监督学习技术，因此为本实验的基线算法.

5.3 实验细节

本文采用分层采样，等比地将正负样本中的70%、15%和15%的样本分别划分为训练集、验证集和测试集. 同时，为了符合半监督实验设定，随机将训练集中85%的数据的标签去掉，作为未标记数据. 最后，为了提高实验的可靠性，独立进行了15次训练集、验证集、测试集的采集，根据模型在验证集上的15次结果的均值来进行模型选择，并将对应的测试集上的15次结果的均值作为最后评估模型的基准.

5.4 实验结果和讨论

各算法在16个数据集上的实验结果如表2所示. 从表2可以看出，本文所提方法在绝大多数情况下都以0.05显著性水平优于其他方法. 在全部16个数据集中，本文所提方法在14个数据集上都表现最优. 从各数据集上的平均性能来看，本文所提方法也达到了最好效果. 此外，给出更加详细的对比分析.

表2 测试集上AUC（越大越佳）性能对比结果（其中对比方法标记"*"表示本文所提出算法显著优于该算法，且其统计显著性经wilcoxson符号秩和检验p值小于0.05；对比方法标记"-"表示本文所提出算法性能劣于该算法，且其统计显著性经wilcoxson符号秩和检验p值小于0.05；win/tie/loss分别表示本文算法显著优于该算法且$p<0.05$次数/本文算法与该算法差异无显著性次数/本文算法显著劣于该算法且$p<0.05$次数；每个数据最优性能以粗体标出、次优性能以下划线标出）

数据集	Ours	RBAUC	PNUAUC	Samult	PNUAB	LSAUC
pima	**0.8022(0.0481)**	0.7923(0.0476)*	0.6458(0.0546)*	0.6459(0.0480)*	0.6743(0.0370)*	0.6511(0.0447)*
ring	**0.9470(0.0095)**	0.9315(0.0120)*	0.9099(0.0061)*	0.9099(0.0061)*	0.8603(0.0131)*	0.9117(0.0058)*
phoneme	**0.8579(0.0143)**	0.8543(0.0149)*	0.8009(0.0156)*	0.8005(0.0145)*	0.7798(0.0171)*	0.7753(0.0129)*
vehicle0	0.9518(0.0215)	0.9465(0.0207)	0.9780(0.0182)-	**0.9846(0.0171)**-	0.8172(0.0432)*	0.8550(0.0317)*
vehicle2	0.9704(0.0153)	0.9595(0.0228)*	0.9647(0.0195)	**0.9810(0.0134)**-	0.8555(0.0417)*	0.7588(0.1247)*
credit-g	**0.7290(0.0314)**	0.7267(0.0420)	0.6968(0.0553)*	0.7147(0.0523)	0.6296(0.0597)*	0.6785(0.0813)*
glass1	**0.7225(0.0904)**	0.7048(0.0795)	0.5828(0.1199)*	0.6365(0.0898)*	0.6321(0.1086)*	0.5833(0.1304)*
wisconsin	**0.9909(0.0065)**	0.9892(0.0076)	0.9195(0.0282)*	0.9211(0.0275)*	0.9105(0.0257)*	0.9054(0.0384)*
wdbc	**0.9873(0.0100)**	0.9859(0.0100)	0.9704(0.0180)*	0.9786(0.0193)*	0.8367(0.0489)*	0.9725(0.0226)*
shuttle-c0-vs-c4	**1.000(0.0000)**	0.9801(0.0322)*	0.9998(0.0005)	0.9998(0.0008)	0.9526(0.0354)*	0.9922(0.0176)*
monk-2	**0.9844(0.0202)**	0.9839(0.0196)	0.8065(0.0498)*	0.8213(0.0474)*	0.9462(0.0260)*	0.7294(0.0570)*
sonar	**0.7247(0.1128)**	0.7140(0.0768)	0.6664(0.1025)*	0.6719(0.1089)*	0.5991(0.0774)*	0.5673(0.0864)*
Surgical	**0.8332(0.0095)**	0.8266(0.0114)*	0.7797(0.0082)*	0.7910(0.0066)*	0.7820(0.0160)*	0.6291(0.0178)*
insurance	**0.8810(0.0012)**	0.8765(0.0015)*	0.6395(0.0072)*	0.6457(0.0065)*	0.8290(0.0022)*	0.6081(0.0198)*
numerai	**0.5245(0.0049)**	0.5226(0.0045)*	0.5229(0.0040)	0.5229(0.0040)	0.5112(0.0044)*	0.5210(0.0049)*
cod-rna	**0.9681(0.0019)**	0.9362(0.0018)*	0.9313(0.0018)*	0.9363(0.0017)*	0.9218(0.0020)*	0.9314(0.0018)*
总体均值	**0.8672**	0.8582	0.7836	0.8009	0.8101	0.7544
win/tie/lose	/	9/7/0	16/0/0	2012/3/1	2011/3/2	16/0/0

所提方法 vs. 半监督＋AUC优化. PNU-AUC和Samult都是半监督学习和AUC优化结合所产生的方法. 所提方法和PNU-AUC相比，在各个数据集上平均性能提升8.31%. 和Samult相比，平均性能提升7.94%. 其中，在pima数据集上比PNU-AUC提升24.22%，比Samult提升24.20%；在ring数据集上比PNU-AUC提升4.08%，比Samult提升4.08%；在phoneme数据集上比PNU-AUC和Samult分别提升7.12%和7.17%；在vehicle0数

① https://github.com/t-sakai-kure/pywsl

据集上比这两种方法分别降低 2.68%和 3.33%；在 vehicle2 数据集上比 PNU-AUC 提升 0.59%，比 Samult 降低 1.08%；在 credit-g 数据集上比 PNU-AUC 提升 4.62%，比 Samult 提升 2.00%；在 glass1 数据集上比 PNU-AUC 提升 23.97%，比 Samult 提升 13.51%；在 wisconsin 数据集上比这两种方法分别提升 7.77%和提升 7.58%；在 wdbc 数据集上比 PNU-AUC 提升 1.74%，比 Samult 提升 0.89%；在 shuttle-c0-vs-c4 数据集上比两种方法都提升了 0.02%；在 monk-2 数据集上比 PNU-AUC 提升 22.06%，比 Samult 提升 19.86%；在 sonar 数据集上比两种方法分别提升 8.75%和 7.86%；在 Surgical 数据集上比两种方法分别提升 6.86%和 5.34%；在 insurance 数据集上比 PNU-AUC 提升 37.76%，比 Samult 提升 36.44%；在 numerai 数据集上比 PNU-AUC 提升 0.31%，比 Samult 提升 0.31%；在 cod-rna 数据集上比 PNU-AUC 提升 3.95%，比 Samult 提升 3.40%. 相比 PNU-AUC，本文算法有 12 次性能提升是在 0.05 显著性水平之上. 相比 Samult，本文算法有 11 次性能提升的显著性水平超过 0.05. 上述实验结果有效验证了定理 2 中的结论，即将 boosting 方法引入半监督 AUC 优化可在保证泛化能力的同时提升优化收敛速率因此带来提升总体性能提升.

所提方法 vs. Boosting＋AUC 优化. RBAUC 同时利用了 boosting 方法及 AUC 优化方法. 和这一方法相比，本文所提方法带来的平均性能提升为 1.44%. 在各个数据集上的提升分别为：在 pima 数据集上提升 1.25%；在 ring 数据集上提升 1.66%；在 phoneme 数据集上提升 0.42%；在 vehicle0 数据集上提升 0.56%；在 vehicle2 数据集上提升 1.14%；在 credit-g 数据集上提升 0.32%；在 glass1 数据集上提升 2.51%；在 wisconsin 数据集上提升 0.17%；在 wdbc 数据集上提升 0.14%；在 shuttle-c0-vs-c4 数据集上提升 2.03%；在 hepatitis 数据集上提升 10.72%；在 monk-2 数据集上提升 0.05%；在 sonar 数据集上提升 1.50%；在 Surgical 数据集上提升 0.80%；在 insurance 数据集上提升 0.51%；在 numerai 数据集上提升 0.36%；在 cod-rna 数据集上提升 3.41%. 其中，有 9 次性能提升是在 0.05 显著性水平之上. 实验结果表明在 boosting 算法中引入半监督学习能够有效利用未标注数据中的额外信息，从而提升模型在少量已标注数据上的学习效果.

PNUAB vs. 其它半监督方法. PNUAB 是半监督学习和 AdaBoost 的结合. 与其它半监督方法的平均进行对比，PNUAB 在各数据集上的平均性能更低. 以本文所提方法为例，平均性能比 PN 高出 10.79%，其中有 16 次性能提升是在 0.05 显著性水平之上. 该实验观测表明，在半监督条件下进行 AUC 优化可有效地提高 AUC 泛化性能.

LSAUC vs. 其它方法. 其它方法和 LSAUC 相比，性能提升均较为明显. 以本文所提方法为例，平均性能提升 14.76%. 各个数据集上的提升分别为：在 pima 数据集上提升 23.21%；在 ring 数据集上提升 3.87%；在 phoneme 数据集上提升 10.65%；在 vehicle0 数据集上提升 11.32%；在 vehicle2 数据集上提升 27.89%；在 credit-g 数据集上提升 7.44%；在 glass 数据集上提升 23.86%；在 wisconsin 数据集上提升 9.44%；wdbc 数据集上提升 1.52%；在 shuttle-c0-vs-c4 数据集上提升 0.79%；在 hepatitis 数据集上提 10.72%；在 monk-2 数据集上提升 34.96%；在 sonar 数据集上提升 27.75%；在 Surgical 数据集上提升 32.44%；在 insurance 数据集上提升 44.88%；在 numerai 数据集上提升 0.67%；在 cod-rna 数据集上提升 3.94%. 其中，有 16 次性能提升是在 0.05 显著性水平之上. 实验结果表明半监督学习以及 boosting 策略都均能有效克服由未标注数据所带来的性能瓶颈.

算法总体比较. 为比较各个算法总体差异的统计显著性，对各个算法在各个数据集上的性能排序进行了事后假设检验，结果如图 1. 其中本文算法获得了最高排序，其他算法在图中距本文算法距离超

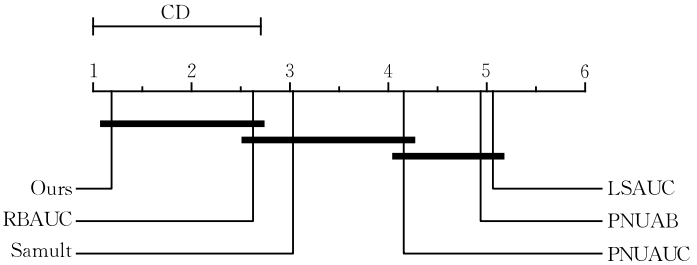

图 1　各算法差异对比图

过 CD(Crtical Difference)长度即可视为具有显著性差异($p<0.05$).可见除 RBAUC 之外,其余算法均于本文算法存在显著差异.另一方面,RBAUC 与本文算法的差距亦接近 CD 距离.由此可见本文所提算法在整体水平也具有显著性优势.

5.5 弱分类器个数的影响

以 vehicle2、insurance、phoneme 三个数据集为例,图 2 中验证了本文算法训练集和测试集平均 AUC 性能(越大越佳)随弱分类器增加的变化趋势,其中曲线阴影部分对应 15 次重复实验上的极差波动.如图 2 所示,随着弱分类器的增加训练集的性能快速提升,并趋于收敛,于此同时,测试集上性能变化趋势与训练集基本一致,未出现过拟合现象.尤其对于 vehicle2 数据集,当弱分类器个数达到 25 左右时,其训练集性能很快到达最佳值 1.00,而测试集在其后的训练过程中未出现下降趋势.综上分析可推知,本文所提出算法可兼顾模型的收敛速度以及泛化能力,再次本文定理 3 的有效性.

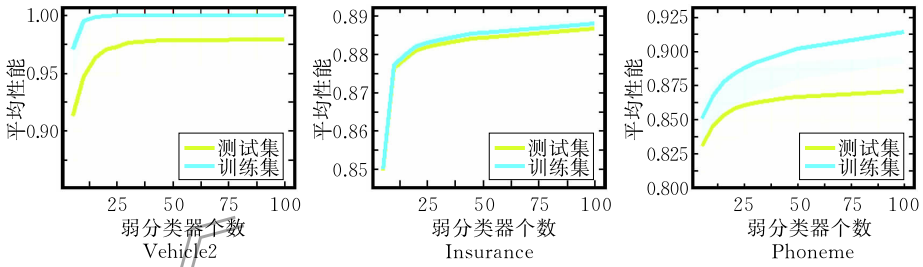

图 2 本文算法性能相对于弱分类器个数的变化趋势

5.6 加速算法验证

为验证本文所提出加速方法的实际加速效果,生成不同规模的仿真数据集对加速前后的算法运行时间进行对比.具体而言,分别生成样本量为 100、150、200、250、300、350、400 的训练数据集.输入特征 $\boldsymbol{X}$ 每维按照正态分布 $\mathcal{N}(0,0.01)$,维度为 10.同时由分布 $\mathcal{N}(0,1)$ 逐维生成线性模型参数 ω,总维度为 10.进一步通过 $s=\boldsymbol{X}\omega+\epsilon$ 生成得分函数,其中 ϵ 为服从 $\mathcal{N}(0,0.0001)$ 的白噪声.最终通过 $y=\boldsymbol{I}[s>0.1]$ 生成样本的标注结果.加速前后运行时间对比如表 3.如表所示,本文加速算法可带来显著的计算效率提升.

表 3 仿真数据集上加速前后运行时间比较 (单位:s)

实现方式	样本量						
	100	150	200	250	300	350	400
加速前	0.2936	0.3818	0.4886	0.6100	0.7890	0.9600	1.1542
加速后	0.0227	0.0258	0.0278	0.0297	0.0356	0.0384	0.0405

6 结 论

本文对半监督条件下基于 Boosting 算法的 AUC 优化模型集成进行了系统性研究.在算法层面,提出一种高效的基于 Boosting 的 AUC 优化模型集成方法,可将单次迭代的空间/时间复杂度由 $O(n_pn_n+n_pn_u+n_un_n)$ 降至 $O(n_p+n_u+n_n)$.进一步的理论分析证明,本文所提出算法可使训练集误差随弱分类器的增加以几何速度衰减且泛化误差随训练样本增加逐渐趋于 0.在泛化误差分析方面,本文通过构造半监督 AUC 的 Rademacher 复杂度以及广义最大值不等式给出了相比于文献[25]更为紧致的上界.实验分析方面,本文在 16 个数据集上对所提方法的性能进行了验证.实验结果证实本文所提出算法可显著提升半监督学习条件下的模型 AUC 性能.

参 考 文 献

[1] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861-874

[2] Hand D J, Till R J. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 2001, 45(2): 171-186

[3] Hao H, Fu H, Xu Y, et al. Open-narrow-synechiae anterior chamber angle classification in AS-OCT sequences. CoRR, abs/2006.05367, 2020

[4] Liu C, Zhong Q, Ao X, et al. Fraud transactions detection via behavior tree with local intention calibration//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3035-3043

[5] Chen Y, Chen B, He X, et al. λOpt: Learn to regularize recommender models in finer levels//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA, 2019: 978-986

[6] Dai L, Yin Y, Qin C, et al. Enterprise cooperation and competition analysis with a sign-oriented preference network//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 774-782

[7] Cortes C, Mohri M. AUC optimization vs. error rate minimization//Advances in Neural Information Processing Systems. Vancouver, Canada, 2003: 313-320

[8] Alan A H, Raskutti B. Optimising area under the ROC curve using gradient descent//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, 2004: 49-56

[9] Calders T, Jaroszewicz S. Efficient AUC optimization for classification//Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery. Warsaw, Poland, 2007: 42-53

[10] Freund Y, Iyer R, Schapire R E, et al. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 2003, 4(Nov): 933-969

[11] Joachims T. A support vector method for multivariate performance measures//Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005: 377-384

[12] Joachims T. Training linear SVMs in linear time//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 217-226

[13] Ying Y, Wen L, Lyu S. Stochastic online AUC maximization//Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 451-459

[14] Natole M, Ying Y, Lyu S. Stochastic proximal algorithms for AUC maximization//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 3707-3716

[15] Agarwal S, Graepel T, Herbrich R, et al. Generalization bounds for the area under the ROC curve. Journal of Machine Learning Research, 2005, 6(4): 393-425

[16] Clémençon S, Lugosi G, Vayatis N, et al. Ranking and empirical minimization of U-statistics. The Annals of Statistics, 2008, 36(2): 844-874

[17] Usunier N, Amini M R, Gallinari P. A data-dependent generalization error bound for the AUC//Proceedings of the ICML Workshop on ROC Analysis in Machine Learning. Bonn, Germany, 2005

[18] Usunier N, Amini M R, Gallinari P. Generalization error bounds for classifiers trained with interdependent data//Advances in Neural Information Processing Systems. Vancouver, Canada, 2005: 1369-1376

[19] Ralaivola L, Szafranski M, Stempfel G. Chromatic PAC-Bayes bounds for non-IID data: Applications to ranking and stationary β-mixing processes. Journal of Machine Learning Research, 2010, 11(Jul): 1927-1956

[20] Lyu S, Ying Y. A univariate bound of area under ROC//Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence. Monterey, USA, 2018: 43-52

[21] Gao W, Wang L, Jin R, et al. One-pass AUC optimization//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 906-914

[22] Agarwal S. Surrogate regret bounds for bipartite ranking via strongly proper losses. Journal of Machine Learning Research, 2014, 15(1): 1653-1674

[23] Gao W, Zhou Z. On the consistency of AUC pairwise optimization//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 939-945

[24] Fujino A, Ueda N. A semi-supervised AUC optimization method with generative models//Proceedings of the IEEE International Conference on Data Mining. Barcelona, Spain, 2016: 883-888

[25] Sakai T, Niu G, Sugiyama M. Semi-supervised AUC optimization based on positive-unlabeled learning. Machine Learning, 2018, 107(4): 767-794

[26] Xie Z, Li M. Semi-supervised AUC optimization without guessing labels of unlabeled data//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 4310-4317

[27] Xie Z, Li M. Cutting the software building efforts in continuous integration by semi-supervised online AUC optimization//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 2875-2881

[28] Vu H T, Gallinari P. Using RankBoost to compare retrieval systems//Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Shanghai, China, 2005: 309-310

[29] Hanley J A, Mcneil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982, 143(1): 29-36

[30] Freund Y, Schapire R E. Experiments with a new boosting algorithm//Proceedings of the 13th International Conference on Machine Learning. Bari, Italy, 1996: 148-156

[31] Popescu P G, Dragomir S S, Slușanschi E I, et al. Bounds for Kullback-Leibler divergence. Electronic Journal of Differential Equations, 2016, 2016(2016): 1-6

[32] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning. Cambridge, MA: MIT Press, 2018

[33] Boucheron S, Lugosi G, Massart P. Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford, UK: Oxford University Press, 2013

[34] Mcdiarmid C. Concentration. Berlin, Germany: Springer, 1998: 195-248

附录 A. 弱分类器的学习.

给定已排序的输入特征集合 $\boldsymbol{X}=\{\boldsymbol{X}^1,\cdots,\boldsymbol{X}^d\}$($\boldsymbol{X}^i$ 为所有训练样本第 i 维特征构成的集合),及已排序的阈值候选集 $\boldsymbol{T}=\{\boldsymbol{T}^1,\cdots,\boldsymbol{T}^d\}$($\boldsymbol{T}^i$ 第 i 维特征候选阈值构成的集合),通过如下搜索方式完成最佳参数的选择.

算法 2. 基于决策树桩的弱分类器.

输入:模型输入特征 $\boldsymbol{X}=\{\boldsymbol{X}^1,\cdots,\boldsymbol{X}^d\}$,其中 $\boldsymbol{X}^i$ 为所有训练样本第 i 维特征数值的集合,令 X_j^i 表示第 j 个样本的第 i 维特征数值. 进入算法前,预先对每个 $\boldsymbol{X}^i$ 元素进行降序排序,使得:$X_1^i\geq X_2^i\geq\cdots\geq X_N^i$;

阈值备选集 $\boldsymbol{T}=\{\boldsymbol{T}^1,\cdots,\boldsymbol{T}^d\}$. 其中 $\boldsymbol{T}^i$ 为第 i 维特的阈值备选集集合,令 T_j^i 表示第 i 个特征的第 j 个阈值备选值,同样预先将其排序并使 $T_1^i\geq T_2^i\geq\cdots\geq T_N^i$

输出:最佳特征维度 i^*,最佳阈值 $\theta_{i^*}^*$,Δ^t

$RE_0\leftarrow-\infty$

FOR $i\leftarrow1:d$ DO

  $last\leftarrow-1$

  $L\leftarrow0$

  FOR $j\leftarrow1:|T_i|$ DO

    $offset\leftarrow last+1$

    FOR $k\leftarrow(last+1):N$ DO

      $t\leftarrow T_j^i$

      IF $X_k^i>t$ THEN

        $L+=g_{t_i}$

        $last\leftarrow k$

      ELSE

        break

      END IF

    END FOR

    $RE_1\leftarrow KL\left(\frac{1+\rho}{2}\middle\|\frac{1+L}{2}\right)$

    IF $RE_1>RE_0$ THEN

      $\Delta_t\leftarrow L$

      $i^*\leftarrow i$

      $\theta_{i^*}^*\leftarrow t$

      $RE_0\leftarrow RE_1$

    END IF

  END FOR

END FOR

附录 B. 引理 1 的证明.

证明. 展开过程 1 中 Step 4 的迭代规则,有

$$D^{T+1}(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_j)=\frac{D^0(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_j)\exp(f(\tilde{\boldsymbol{x}}_j)-f(\boldsymbol{x}_i))}{\prod_{t=1}^{T}\tilde{Z}^t}\tag{47}$$

$$D^{T+1}(\boldsymbol{x}_u,\boldsymbol{x}_n)=\frac{D^0(\boldsymbol{x}_u,\boldsymbol{x}_n)\exp(f(\boldsymbol{x}_u)-f(\boldsymbol{x}_n))}{\prod_{t=1}^{T}\tilde{Z}^t}\tag{48}$$

$$D^{T+1}(\boldsymbol{x}_p,\boldsymbol{x}_n)=\frac{D^0(\boldsymbol{x}_p,\boldsymbol{x}_n)\exp(f(\boldsymbol{x}_p)-f(\boldsymbol{x}_n))}{\prod_{t=1}^{T}\tilde{Z}^t}\tag{49}$$

将式(47)、式(49)代入式(17)得

$$R_{\mathrm{OP}_4}=\prod_{t=1}^{T}\exp(-\rho\cdot\alpha^t)\cdot\tilde{Z}^t\cdot\Big(\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}D^{T+1}(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_j)+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}D^{T+1}(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})+\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}D^{T+1}(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})\Big).$$

由式(13)~式(15)、式(18)~式(24)、式(25)~式(30),易得样例对权重 D^{T+1} 在算法中已被归一化,即:

$$\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}D^{T+1}(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_j)+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}D^{T+1}(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})+\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}D^{T+1}(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})=1.$$

由式(50),此引理得证. 证毕.

附录 C. 引理 2 证明.

证明. 首先对式(32)中的 $C(\alpha^t)$ 进行化简,有:

$$\exp(\rho\alpha^t)\cdot C(\alpha^t)=\frac{1+\Delta^t}{2}\exp((\rho-1)\cdot\alpha^t)+\frac{1-\Delta^t}{2}\exp((\rho+1)\cdot\alpha^t)\tag{50}$$

将式(33)代入式(50),将 $\exp(\rho\alpha^t)\cdot C(\alpha^t)$ 重写为

$$\underbrace{\frac{1-\Delta^t}{2}\exp\left[\frac{\rho+1}{2}\log\left(\frac{(1+\Delta^t)\cdot(1-\rho)}{(1-\Delta^t)\cdot(1+\rho)}\right)\right]}_{\mathrm{(I)}}+\underbrace{\frac{1+\Delta^t}{2}\exp\left[\frac{\rho-1}{2}\log\left(\frac{(1+\Delta^t)\cdot(1-\rho)}{(1-\Delta^t)\cdot(1+\rho)}\right)\right]}_{\mathrm{(II)}}\tag{51}$$

对于(I),有:

$$\begin{aligned}\mathrm{(I)}&=\frac{1}{2}\exp\left[-KL\left(\frac{1+\rho}{2}\middle\|\frac{1+\Delta^t}{2}\right)+\log\left(\frac{1-\rho}{1-\Delta^t}\right)+\log(1-\Delta^t)\right]\\&=\frac{1-\rho}{2}\cdot\exp\left[-KL\left(\frac{1+\rho}{2}\middle\|\frac{1+\Delta^t}{2}\right)\right]\end{aligned}\tag{52}$$

对于(II),有:

$$\begin{aligned}\mathrm{(II)}&=\frac{1}{2}\exp\left[-KL\left(\frac{1+\rho}{2}\middle\|\frac{1+\Delta^t}{2}\right)+\log\left(\frac{1+\rho}{1+\Delta^t}\right)+\log(1+\Delta^t)\right]\\&=\frac{1+\rho}{2}\cdot\exp\left[-KL\left(\frac{1+\rho}{2}\middle\|\frac{1+\Delta^t}{2}\right)\right]\end{aligned}\tag{53}$$

联立式(51)、式(52)及式(53),有:

$$\begin{aligned}\exp(\rho\alpha^t)C(\alpha^t)&=\left(\frac{1+\rho}{2}+\frac{1-\rho}{2}\right)\cdot\exp\left[-KL\left(\frac{1+\rho}{2}\middle\|\frac{1+\Delta^t}{2}\right)\right]\\&=\exp\left[-KL\left(\frac{1+\rho}{2}\middle\|\frac{1+\Delta^t}{2}\right)\right].\end{aligned}$$

由此引理得证. 证毕.

附录 D. 定理 2 的证明.

D.1. 预备引理.

定义 1. 有限差分性质(Bounded Difference Property). 给定一组独立随机变量 $X_1,\cdots,X_n$ 及其定义域 $\mathbb{X}$,对于函数 $f(X_1,X_2,\cdots,X_n)$ 若存在非负常量 $c_1,c_2,\cdots,c_n$ 使:

$$\sup_{x_1,\cdots,x_n,x_i'} |f(x_1,\cdots,x_n)-f(x_1,\cdots,x_{i-1},x_i',\cdots,x_n)|\le c_i,$$

$$\forall\, 1\le i\le n \tag{54}$$

则称函数 f 满足有限差分性质.

对于作用于随机变量上且满足该性质的函数,有以下不等式:

引理 4. 有限差分不等式(Bounded Difference Inequality)(见文献[34]命题 6.1 及定理 6.2). 设 $X_1,\cdots,X_n,X_i\in\mathbb{X}$ 为一组独立随机变量,令 $Z=f(X_1,\cdots,X_n)$,若 f 满足有限差分性质,且对应常数为 $c_1,c_2,\cdots,c_n$,则有:

$$\log\mathbb{E}[\exp(\lambda(Z-\mathrm{E}[Z]))]\le\frac{\lambda^2 v}{2} \tag{55}$$

对于任意 $\lambda>0$ 成立,其中:

$$v=\frac{1}{4}\sum_{i=1}^{n}c_i^2 \tag{56}$$

引理 5. 最大值不等式(Maximal Inequality)(见文献[33]2.5 节). 令 $Z_1,\cdots,Z_n$ 为一组实数值随机变量且存在 $v>0$ 使任意 $i=1,2,\cdots,n$, 有 $\log(\mathbb{E}[\exp(\lambda Z_i)])\le\frac{\lambda^2 v}{2}$,则有:

$$\mathbb{E}[\max_{i=1,2,\cdots,n}Z_i]\le\sqrt{2v\log n}.$$

引理 6. Mcdiarmid 不等式(见文献[34]). 令 $X_1,\cdots,X_m$ 为一组取值于集合 $\mathcal{X}$ 内的独立随机变量,令 $f:\mathcal{X}\to\mathbb{R}$ 满足:

$$\sup_{x,x'}|f(x_1,\cdots,x_i,\cdots,x_m)-f(x_1,\cdots,x_i',\cdots,x_m)|\le c_i,$$

其中 $\boldsymbol{x}\ne\boldsymbol{x}'$,则对于任意 $\epsilon>0$ 有:

$$P[\mathbb{E}(f)-f\ge\epsilon]\le\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}c_i^2}\right),$$

$$P[f-\mathbb{E}(f)\ge\epsilon]\le\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}c_i^2}\right).$$

引理 7. Talagrand 压缩引理. 令 $\ell_1,\cdots,\ell_l$ 为一组 ϕ-Lipshiptz 连续函数,$\sigma_1,\cdots,\sigma_m$ 为相互独立的一组 Rademacher 随机变量,则:

$$\frac{1}{m}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\cdot(\ell_i\circ f)(\boldsymbol{x})\Big]\le\frac{\phi}{m}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\cdot f(\boldsymbol{x})\Big].$$

引理 8. 见文献[32]引理 7.4. 给定 $\mathcal{H}$ 为由函数 $f:\mathcal{X}\to\mathbb{R}$ 构成的函数集,记 $co(\mathcal{H})$ 为

$$co(\mathcal{H})=\Big\{f:\sum_{i=1}^{T}\alpha^t h^t,\sum_{i=1}^{T}\alpha^t=1,\alpha^t\ge0,h^t\in\mathcal{H}\Big\},$$

$$\frac{1}{m}\mathbb{E}_\sigma\Big[\sup_{f\in co(\mathcal{H})}\sum_{i=1}^{m}\sigma_i\cdot f(\boldsymbol{x})\Big]=\frac{1}{m}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\cdot f(\boldsymbol{x})\Big].$$

D.2. 本文提出的引理.

引理 9. 广义最大值不等式. 给定实数值随机变量 $\{M_i^{(k)}\}_{1\le i\le N,1\le k\le K}$,若满足以下条件:

(1) 对于任意 $k_1\ne k_2$,$M_{i_1}^{(k_1)}$ 与 $M_{i_2}^{(k_2)}$ 相互独立.

(2) $\forall i,k,\mathbb{E}[M_i^{(k)}]=0$,且 $\log(\mathbb{E}[\exp(\lambda M_i^{(k)})])\le\frac{\lambda^2 v_k}{2}$ 有:

$$\mathbb{E}\Big(\sum_{k=1}^{K}\max_i M_i^k\Big)\le\Big(2\log N\cdot\sum_k v_k\Big)^{\frac{1}{2}}.$$

证明. 见附录 D.3. 证毕.

引理 10. PNU-AUC 的对称化技术. 将数据集 $\mathcal{S}$ 及 $\mathcal{S}'$ 的类标集固定为 $\boldsymbol{Y},\boldsymbol{Y}'$,则对任意假设集 $\mathcal{H}$ 及损失函数 ℓ,以下结论成立:

$$\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\sup_{f\in\mathcal{H}}(\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f))]\le4\mathfrak{R}_{PNU}(\ell\circ\mathcal{H}).$$

证明. 见附录 D.5. 证毕.

引理 11. $\hat{\mathfrak{R}}_{PNU}(\ell_\rho\circ co(\mathcal{H}_{DS}))\le2\cdot(2(\log d+\log K)\cdot\rho_\gamma(\boldsymbol{Y}))^{\frac{1}{2}}$,其中

$$\rho_\gamma(\boldsymbol{Y})=\Big(\frac{1}{n_p}+\frac{1}{n_u}+\frac{1}{n_n}\Big).$$

证明. 见附录 D.5. 证毕.

引理 12. PNU-AUCBOOST 的泛化界. 给定训练数据集 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$,$\rho>0$,及弱分类器的阈值备选集 $\boldsymbol{T}_e$,设样例均由独立采样生成,则对于任意函数 $f\in co(\mathcal{H}_{DS})$ 以及任意 $\delta\in(0,1)$,下式至少以 $1-\delta$ 概率成立:

$$\mathbb{E}_{\mathcal{S}}\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)\le\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)+4\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell_\rho\circ\mathcal{H}_{DS})+\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\Big(\frac{1}{2n_p}+\frac{1}{2n_u}+\frac{1}{2n_n}\Big)\Big)^{\frac{1}{2}} \tag{57}$$

其中 $\chi(\boldsymbol{Y})=\frac{1}{n_p}+\frac{1}{n_u}+\frac{1}{n_n}$.

D.3. 引理 9 的证明.

证明:

$$\begin{aligned}
&\exp\Big(\mathbb{E}\Big(\lambda\sum_{k=1}^{K}\max_i M_i^k\Big)\Big)\\
&\overset{(1)}{\le}\mathbb{E}\Big(\prod_{k=1}^{K}\exp(\lambda\max_i M_i^k)\Big)\\
&\overset{(2)}{=}\prod_{k=1}^{K}\mathbb{E}(\exp(\lambda\max_i M_i^k))\\
&\overset{(3)}{=}\prod_{k=1}^{K}\mathbb{E}(\max_i\exp(\lambda M_i^k))\\
&\le\prod_{k=1}^{K}\mathbb{E}\Big(\sum_{i=1}^{N}\exp(\lambda M_i^k)\Big)\\
&\overset{(4)}{\le}N\exp\left(\frac{\lambda^2\sum_{k=1}^{K}v_k}{2}\right)
\end{aligned}\tag{58}$$

其中(1)由 Jensen 不等式获得;(2)由本引理假设条件(1)获得;(3)由指数函数 $\exp(x)$ 的严格单增性获得;(4)由本引理假设条件(2)获得. 由上式进一步获得:

$$\mathbb{E}\Big(\sum_{k=1}^{K}\max_i M_i^k\Big)\le\frac{\log\left(N\exp\left(\frac{\lambda^2\sum_{k=1}^{K}v_k}{2}\right)\right)}{\lambda} \tag{59}$$

对上式求最大值,引理得证. 证毕.

D.4. 引理 10 的证明.

证明. 此处记 $\ell(f,\boldsymbol{x}_1,\boldsymbol{x}_2)=\ell(f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2))$,定义

$$\begin{aligned}
Q_\sigma^{p,n,i,k}=&\frac{\sigma_i^{(p)}+\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_i^{(p)},\tilde{\boldsymbol{x}}_k^{(n)})+\\
&\frac{\sigma_i^{(p)}-\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_i^{(p)},\tilde{\boldsymbol{x}}_k^{(n)})-\\
&\frac{\sigma_i^{(p)}-\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_i^{(p)},\tilde{\boldsymbol{x}}_k^{(n)})-\\
&\frac{\sigma_i^{(p)}+\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_i^{(p)},\tilde{\boldsymbol{x}}_k^{(n)}),\\
Q_\sigma^{p,n,i,j}=&\frac{\sigma_i^{(p)}+\sigma_j^{(u)}}{2}\ell(f,\tilde{\boldsymbol{x}}_i^{(p)},\tilde{\boldsymbol{x}}_j^{(u)})+\\
&\frac{\sigma_i^{(p)}-\sigma_j^{(u)}}{2}\ell(f,\tilde{\boldsymbol{x}}_i^{(p)},\tilde{\boldsymbol{x}}_j^{(u)})-\\
&\frac{\sigma_i^{(p)}-\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_i^{(p)},\tilde{\boldsymbol{x}}_j^{(u)})-\\
&\frac{\sigma_i^{(p)}+\sigma_j^{(u)}}{2}\ell(f,\tilde{\boldsymbol{x}}_i^{(p)},\tilde{\boldsymbol{x}}_j^{(u)}),\\
Q_\sigma^{u,n,j,k}=&\frac{\sigma_j^{(u)}+\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_j^{(u)},\tilde{\boldsymbol{x}}_k^{(n)})+\\
&\frac{\sigma_i^{(p)}-\sigma_j^{(u)}}{2}\ell(f,\tilde{\boldsymbol{x}}_j^{(u)},\tilde{\boldsymbol{x}}_k^{(n)})-\\
&\frac{\sigma_j^{(u)}-\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_j^{(u)},\tilde{\boldsymbol{x}}_k^{(n)})-\\
&\frac{\sigma_j^{(u)}+\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_j^{(u)},\tilde{\boldsymbol{x}}_k^{(n)})
\end{aligned}\tag{60}$$

首先证明:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S},\mathcal{S}'}\Big[\sup_{f\in\mathcal{H}_{DS}}(\hat{R}_{\mathcal{S}'}^{PNU}(f)-\hat{R}_{\mathcal{S}}^{PNU}(f))\Big]=\\
&\mathbb{E}_{\mathcal{S},\mathcal{S}'}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\gamma}{n_pn_n}\cdot Q_\sigma^{p,n,i,k}+\\
&\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{1-\gamma}{2n_pn_u}\cdot Q_\sigma^{p,u,i,j}+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{1-\gamma}{2n_un_n}\cdot Q_\sigma^{u,n,j,k}\Big]
\end{aligned}\tag{61}$$

给定样本集

$$\mathcal{S}=\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_p}\cup\{\boldsymbol{x}_j^{(u)}\}_{j=1}^{n_u}\cup\{\boldsymbol{x}_k^{(n)}\}_{k=1}^{n_n},$$

$$\mathcal{S}'=\{\tilde{\boldsymbol{x}}_i^{(p)}\}_{i=1}^{n_p}\cup\{\tilde{\boldsymbol{x}}_j^{(u)}\}_{j=1}^{n_u}\cup\{\tilde{\boldsymbol{x}}_k^{(n)}\}_{k=1}^{n_n}.$$

考虑 $\mathcal{S},\mathcal{S}'$ 中的样例由独立采样获得,因此交换数据集中的相同类标的样例不影响数据分布.考虑该数据交换过程,记交换数据集 $\mathcal{S}$ 及 $\mathcal{S}'$ 任意一个或多个对应位置的样例,也即交换任意一个或多个 $\boldsymbol{x}_i^{(p)}$ 与 $\tilde{\boldsymbol{x}}_i^{(p)}$,$\boldsymbol{x}_j^{(u)}$ 与 $\tilde{\boldsymbol{x}}_j^{(u)}$,或 $\boldsymbol{x}_k^{(n)}$ 与 $\tilde{\boldsymbol{x}}_k^{(n)}$ 所得的两个新数据集分别为 $\tilde{\mathcal{S}}$ 及 $\tilde{\mathcal{S}}'$,易得如下等式关系:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S},\mathcal{S}'}\Big[\sup_{f\in\mathcal{H}}(\hat{R}_{\mathcal{S}'}^{PNU}(f)-\hat{R}_{\mathcal{S}}^{PNU}(f))\Big]=\\
&\mathbb{E}_{\mathcal{S},\mathcal{S}'}\Big[\sup_{f\in\mathcal{H}}(\hat{R}_{\tilde{\mathcal{S}}'}^{PNU}(f)-\hat{R}_{\tilde{\mathcal{S}}'}^{PNU}(f))\Big]
\end{aligned}\tag{62}$$

受上式启发,为证明式(61),首先证明对于任意独立同分布 Rademacher 随机变量序列 $\sigma=\{\sigma_i^{(p)}\}_i\cup\{\sigma_j^{(u)}\}_j\cup\{\sigma_k^{(n)}\}_k$ 以及任意 $\mathcal{S}$、$\mathcal{S}'$,存在交换后产生的数据集 $\tilde{\mathcal{S}}^\sigma$ 及 $\tilde{\mathcal{S}}'^\sigma$ 使得:

$$\begin{aligned}
&\sup_{f\in\mathcal{H}}\Big[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\gamma}{n_pn_n}\cdot Q_\sigma^{p,n,i,k}+\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{1-\gamma}{2n_pn_u}\cdot Q_\sigma^{p,u,i,j}+\\
&\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{1-\gamma}{2n_un_n}\cdot Q_\sigma^{u,n,j,k}\Big]=\sup_{f\in\mathcal{H}}[\hat{R}_{\tilde{\mathcal{S}}^{\sigma\prime}}(f)-\hat{R}_{\tilde{\mathcal{S}}^\sigma}(f)]
\end{aligned}\tag{63}$$

下面通过数学归纳法证明式(63).

步骤 1. 考虑平凡情况:$n_p=1,n_u=1,n_n=1$,$\mathcal{S}=\{\boldsymbol{x}_1^{(p)},\boldsymbol{x}_1^{(n)},\boldsymbol{x}_1^{(u)}\}$,$\mathcal{S}'=\{\tilde{\boldsymbol{x}}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(u)},\tilde{\boldsymbol{x}}_1^{(n)}\}$,$\sigma=(\sigma_1^{(p)},\sigma_1^{(u)},\sigma_1^{(n)})$.下面通过分类讨论证明式(63)成立:

情况 1-1. $\sigma_1^{(p)}=1,\sigma_1^{(u)}=1,\sigma_1^{(n)}=1$:

$$\begin{aligned}
Q_\sigma^{p,n,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(n)})-\ell(f,\boldsymbol{x}_1^{(p)},\boldsymbol{x}_1^{(n)}),\\
Q_\sigma^{p,u,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(u)})-\ell(f,\boldsymbol{x}_1^{(p)},\boldsymbol{x}_1^{(u)}),\\
Q_\sigma^{u,n,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(u)},\tilde{\boldsymbol{x}}_1^{(n)})-\ell(f,\boldsymbol{x}_1^{(u)},\boldsymbol{x}_1^{(n)})
\end{aligned}\tag{64}$$

比较 Rademacher 的基本定义,若令 $\tilde{\mathcal{S}}_\sigma=\mathcal{S},\tilde{\mathcal{S}}'_\sigma=\mathcal{S}'$,则式(63)成立.

情况 1-2. $\sigma_1^{(p)}=-1,\sigma_1^{(u)}=-1,\sigma_1^{(n)}=-1$.通过与情况 1-1 类似的推导方法,可知,若令 $\tilde{\mathcal{S}}_\sigma=\mathcal{S}',\tilde{\mathcal{S}}'_\sigma=\mathcal{S}$,则式(63)成立.

情况 1-3. $\sigma_1^{(p)}=1,\sigma_1^{(u)}=1,\sigma_1^{(n)}=-1$,此时有:

$$\begin{aligned}
Q_\sigma^{p,n,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(p)},\boldsymbol{x}_1^{(n)})-\ell(f,\boldsymbol{x}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(n)}),\\
Q_\sigma^{p,u,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(u)})-\ell(f,\boldsymbol{x}_1^{(p)},\boldsymbol{x}_1^{(u)}),\\
Q_\sigma^{u,n,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(u)},\boldsymbol{x}_1^{(n)})-\ell(f,\boldsymbol{x}_1^{(u)},\tilde{\boldsymbol{x}}_1^{(n)})
\end{aligned}\tag{65}$$

令 $\mathcal{S},\mathcal{S}'$ 交换 $\boldsymbol{x}_1^{(n)}$ 与 $\tilde{\boldsymbol{x}}_1^{(n)}$,分别得到 $\tilde{\mathcal{S}}_\sigma,\tilde{\mathcal{S}}'_\sigma$,则此时式(63)成立.

情况 1-4. 包括其余 5 种情况.这些情况与情况 3 类似,可证明 $\mathcal{S},\mathcal{S}'$ 交换 $\sigma_i=-1$ 位置对应的样例则可得到 $\tilde{\mathcal{S}}_\sigma,\tilde{\mathcal{S}}'_\sigma$,细节此处从略.

综上讨论,基条件得证,下面证明递推关系成立.

步骤 2. 考察递推关系,给定 $1<n_1<n_p,1<n_2<n_u,1<n_3<n_n$,样本

$$\mathcal{S}^0=\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_1}\cup\{\boldsymbol{x}_j^{(u)}\}_{j=1}^{n_2}\cup\{\boldsymbol{x}_k^{(n)}\}_{k=1}^{n_3},$$

$$\mathcal{S}^{0\prime}=\{\tilde{\boldsymbol{x}}_i^{(p)}\}_{i=1}^{n_1}\cup\{\tilde{\boldsymbol{x}}_j^{(u)}\}_{j=1}^{n_2}\cup\{\tilde{\boldsymbol{x}}_k^{(n)}\}_{k=1}^{n_3}$$

以及 Rademacher 随机变量序列:$\sigma^0=\{\sigma_i^{(p)}\}_{i=1}^{n_1}\cup\{\sigma_j^{(u)}\}_{j=1}^{n_2}\cup\{\sigma_k^{(n)}\}_{k=1}^{n_3}$,假设存在 $\tilde{\mathcal{S}}_{\sigma^0}^0,\tilde{\mathcal{S}}_{\sigma^0}^{0\prime}$ 使式(63)对 $\mathcal{S}^0,\mathcal{S}^{0\prime},\sigma^0$ 成立.下面证明,任意给出新样本 $\boldsymbol{x}_b^{(p)},\tilde{\boldsymbol{x}}_b^{(p)},\sigma_b^{(p)},\boldsymbol{x}_b^{(u)},\tilde{\boldsymbol{x}}_b^{(u)},\sigma_b^{(u)}$ 或 $\boldsymbol{x}_b^{(n)},\tilde{\boldsymbol{x}}_b^{(n)},\sigma_b^{(n)}$,式(63)仍然成立.

显然仅 $Q_\sigma^{p,n,b,k},k=1,2,\cdots,n_n$,$Q_\sigma^{p,u,b,j},j=1,2,\cdots,n_u$ 的计算与新样本有关.首先考虑任意给定的 $Q_\sigma^{p,n,b,k}$.先考虑 $\sigma_b^{(p)}=1$ 的情况:

情况 2-1. $\sigma_k^{(n)}=1$,此时有:

$$Q_\sigma^{p,n,b,k}=\ell(f,\tilde{\boldsymbol{x}}_b^{(p)},\tilde{\boldsymbol{x}}_k^{(n)})-\ell(f,\boldsymbol{x}_b^{(p)},\boldsymbol{x}_k^{(n)}).$$

情况 2-2. $\sigma_k^{(n)}=-1$,有:

$$Q_\sigma^{p,n,b,k}=\ell(f^{(i)},\tilde{\boldsymbol{x}}_b^{(p)},\boldsymbol{x}_k^{(n)})-\ell(f,\boldsymbol{x}_b^{(p)},\tilde{\boldsymbol{x}}_k^{(n)}).$$

同理,可以证明若 $\sigma_b^{(p)}=-1$,则有:

情况 3-1. $\sigma_k^{(n)}=1$,此时有:

$$Q_\sigma^{p,n,b,k}=\ell(f,\boldsymbol{x}_b^{(p)},\tilde{\boldsymbol{x}}_k^{(n)})-\ell(f,\tilde{\boldsymbol{x}}_b^{(p)},\boldsymbol{x}_k^{(n)}).$$

情况 3-2. $\sigma_k^{(n)}=-1$,有:

$$Q_\sigma^{p,n,b,k}=\ell(f^{(i)},\boldsymbol{x}_b^{(p)},\boldsymbol{x}_k^{(n)})-\ell(f,\tilde{\boldsymbol{x}}_b^{(p)},\tilde{\boldsymbol{x}}_k^{(n)}).$$

对于 $Q_\sigma^{p,u,b,j}$,若 $\sigma_b^{(p)}=1$,有:

情况 4-1. $\sigma_j^{(u)}=1$,此时有:$\sigma_m^{(j)}=1$,

$$Q_\sigma^{p,n,b,k}=\ell(f,\tilde{\boldsymbol{x}}_b^{(p)},\tilde{\boldsymbol{x}}_j^{(u)})-\ell(f,\boldsymbol{x}_b^{(p)},\boldsymbol{x}_j^{(u)}).$$

情况 4-2. $\sigma_j^{(u)}=-1$，有：

$$Q_\sigma^{p,n,b,k}=\ell(f^{(i)},\widetilde{\boldsymbol{x}}_b^{(p)},\boldsymbol{x}_j^{(u)})-\ell(f,\boldsymbol{x}_b^{(p)},\widetilde{\boldsymbol{x}}_j^{(u)}).$$

若 $\sigma_b^{(p)}=-1$，有：

情况 5-1. $\sigma_j^{(u)}=1$，此时有：

$$Q_\sigma^{p,n,b,k}=\ell(f,\boldsymbol{x}_b^{(p)},\widetilde{\boldsymbol{x}}_j^{(u)})-\ell(f,\widetilde{\boldsymbol{x}}_b^{(p)},\boldsymbol{x}_j^{(u)}).$$

情况 5-2. $\sigma_j^{(u)}=-1$，有：

$$Q_\sigma^{p,n,b,k}=\ell(f^{(i)},\boldsymbol{x}_b^{(p)},\boldsymbol{x}_j^{(u)})-\ell(f,\widetilde{\boldsymbol{x}}_b^{(p)},\widetilde{\boldsymbol{x}}_j^{(u)}).$$

综上讨论，若 $\sigma_b^{(p)}=1$，令：

$$\widetilde{\mathcal{S}}_\sigma=\widetilde{\mathcal{S}}_{\sigma^0}\cup\{\widetilde{\boldsymbol{x}}_b^{(p)}\},\ \widetilde{\mathcal{S}}'_\sigma=\widetilde{\mathcal{S}}'_{\sigma^0}\cup\{\boldsymbol{x}_b^{(p)}\},$$

若 $\sigma_b^{(p)}=-1$，令：

$$\widetilde{\mathcal{S}}_\sigma=\widetilde{\mathcal{S}}_{\sigma^0}\cup\{\boldsymbol{x}_b^{(p)}\},\ \widetilde{\mathcal{S}}'_\sigma=\widetilde{\mathcal{S}}'_{\sigma^0}\cup\{\widetilde{\boldsymbol{x}}_b^{(p)}\},$$

可保证式(63)仍然成立.

对于新样本为 $\boldsymbol{x}_b^{(u)},\widetilde{\boldsymbol{x}}_b^{(u)},\sigma_b^{(u)}$ 或 $\boldsymbol{x}_b^{(n)},\widetilde{\boldsymbol{x}}_b^{(n)},\sigma_b^{(n)}$ 的情况，仅需将上述 $\widetilde{\mathcal{S}}_\sigma,\widetilde{\mathcal{S}}'_\sigma$ 构造方式中的 $\sigma_b^{(p)}$ 替换为 $\sigma_b^{(u)}$ 或 $\sigma_b^{(n)}$，将 $\widetilde{\boldsymbol{x}}_b^{(p)},\boldsymbol{x}_b^{(p)}$ 替换为 $\widetilde{\boldsymbol{x}}_b^{(u)},\boldsymbol{x}_b^{(u)}$ 或 $\widetilde{\boldsymbol{x}}_b^{(n)},\boldsymbol{x}_b^{(n)}$ 即可保证式(63)成立，由证明过程与上述内容类似，此处从略.

综上讨论，式(63)得证. 基于式(63)，可进一步得出：

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S},\mathcal{S}'}\mathbb{E}_\sigma\Big[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\gamma}{n_pn_n}\cdot Q_\sigma^{p,n,i,k}+\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{1-\gamma}{2n_pn_u}\cdot Q_\sigma^{p,u,i,j}+\\
&\qquad\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{1-\gamma}{2n_un_n}\cdot Q_\sigma^{u,n,j,k}\Big]\\
&=\mathbb{E}_\sigma\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\sup_{f\in\mathcal{H}}(\hat{R}_{\widetilde{\mathcal{S}}^{\sigma'}}(f)-\hat{R}_{\widetilde{\mathcal{S}}^{\sigma}}(f))]\\
&=\frac{1}{2^N}\cdot\sum_\sigma\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\sup_{f\in\mathcal{H}}(\hat{R}_{\widetilde{\mathcal{S}}^{\sigma'}}(f)-\hat{R}_{\widetilde{\mathcal{S}}^{\sigma}}(f))]\\
&=\frac{1}{2^N}\cdot\sum_\sigma\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\sup_{f\in\mathcal{H}}(\hat{R}_{\mathcal{S}'}^{PNU}(f)-\hat{R}_{\mathcal{S}}^{PNU}(f))]\\
&=\frac{2^N}{2^N}\cdot\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\sup_{f\in\mathcal{H}}(\hat{R}_{\mathcal{S}'}^{PNU}(f)-\hat{R}_{\mathcal{S}}^{PNU}(f))]\\
&=\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\sup_{f\in\mathcal{H}}(\hat{R}_{\mathcal{S}'}^{PNU}(f)-\hat{R}_{\mathcal{S}}^{PNU}(f))]
\end{aligned}\tag{66}$$

由于 Rademacher 随机变量的符号不影响其期望及样例水平的独立性，有

$$\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\sup_{f\in\mathcal{H}}(\hat{R}_{\mathcal{S}'}^{PNU}(f)-\hat{R}_{\mathcal{S}}^{PNU}(f))]\le 4\mathfrak{R}_{PNU}(\ell\circ\mathcal{H}).$$

由此引理得证. 证毕.

D.5. 引理 11 证明.

证明. 根据 $\hat{\mathfrak{R}}_{PNU}(\ell_\rho\circ co(\mathcal{H}_{DS}))$ 定义，有：

$$\begin{aligned}
\hat{\mathfrak{R}}_{PNU}(\ell_\rho\circ co(\mathcal{H}_{DS}))\le&\underbrace{\mathbb{E}_\sigma\Big[\sup_{f\in co(\mathcal{H}_{DS})}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}Q_{i,k}\Big]}_{\text{(I)}}+\\
&\underbrace{\mathbb{E}_\sigma\Big[\sup_{f\in co(\mathcal{H}_{DS})}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}Q_{i,j}\Big]}_{\text{(II)}}+\\
&\underbrace{\mathbb{E}_\sigma\Big[\sup_{\sigma}\sum_{f\in co(\mathcal{H}_{DS})}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}Q_{j,k}\Big]}_{\text{(III)}}.
\end{aligned}$$

由引理 7、引理 8、ℓ_ρ 函数的 $\frac{1}{\rho}$-Lipschiptz 连续性及上确界的次可加性，有以下结论：

$$\begin{aligned}
\text{(I)}\le&\frac{\gamma}{2n_pn_n\rho}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}_{DS}}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_i^{(p)}}{2}(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_k^{(n)}))\Big]+\\
&\frac{\gamma}{n_pn_n\rho}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}_{DS}}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_k^{(n)}))\Big],\\
\text{(II)}\le&\frac{1-\gamma}{2n_pn_u\rho}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}_{DS}}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_i^{(p)}}{2}(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_j^{(u)}))\Big]+\\
&\frac{1-\gamma}{2n_pn_u\rho}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}_{DS}}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_j^{(u)}}{2}(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_j^{(u)}))\Big],\\
\text{(III)}\le&\frac{1-\gamma}{2n_un_n\rho}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}_{DS}}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_j^{(u)}}{2}(f(\boldsymbol{x}_j^{(u)})-f(\boldsymbol{x}_k^{(n)}))\Big]+\\
&\frac{1-\gamma}{2n_pn_n\rho}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}_{DS}}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}(f(\boldsymbol{x}_j^{(u)})-f(\boldsymbol{x}_k^{(n)}))\Big],
\end{aligned}$$

定义

$$\begin{aligned}
\text{(IV)}=&\mathbb{E}_\sigma\Big[\max_{e\in[d]}\frac{\gamma}{n\in\boldsymbol{T}_e}\Big[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_i^{(p)}}{2}(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)}))\Big]+\\
&\max_{\substack{e\in[d]\\\theta\in\boldsymbol{T}_e}}\Big[\frac{1-\gamma}{2n_pn_u}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_j^{(u)}}{2}(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)}))\Big]+\\
&\max_{\substack{e\in[d]\\\theta\in\boldsymbol{T}_e}}\Big[\frac{1-\gamma}{2n_un_n}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}(h_\theta^e(\boldsymbol{x}_j^{(u)})-h_\theta^e(\boldsymbol{x}_k^{(n)}))\Big]\Big],\\
\text{(IV)}=&\mathbb{E}_\sigma\Big[\max_{e\in[d]}\frac{\gamma}{n\in\boldsymbol{T}_e}\cdot\Big[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)}))\Big]+\\
&\max_{\substack{e\in[d]\\\theta\in\boldsymbol{T}_e}}\Big[\frac{1-\gamma}{2n_pn_u}\cdot\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_i^{(p)}}{2}(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_j^{(u)}))\Big]+\\
&\max_{\substack{e\in[d]\\\theta\in\boldsymbol{T}_e}}\Big[\frac{1-\gamma}{2n_un_n}\cdot\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_j^{(u)}}{2}(h_\theta^e(\boldsymbol{x}_j^{(u)})-h_\theta^e(\boldsymbol{x}_k^{(n)}))\Big]\Big]
\end{aligned}$$

有

$$\text{(I)}+\text{(II)}+\text{(III)}\le\text{(IV)}+\text{(V)}$$

固定训练样本，仅将随机数 σ 视作随机变量，(IV)、(V)均满足引理 9 条件，下面由该引理分别给出其上界. 首先关注(IV)，可证明 $\frac{\gamma}{n_pn_n}\cdot\Big[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)}))\Big]$ 相对于随机变量序列 $\{\sigma_i^{(p)}\}$ 满足有限差分性质，且对应系数为

$$c_1=c_2=\cdots=c_{n_p}=\frac{\gamma}{n_p}.$$

同理可知 $\frac{1-\gamma}{2n_pn_u}\cdot\Big[\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_j^{(u)}}{2}(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_j^{(u)}))\Big]$ 满足有限差分性质，且对应系数为

$$c_1=c_2=\cdots=c_{n_u}=\frac{1-\gamma}{2n_u}$$

$\frac{1-\gamma}{2n_un_n}\cdot\Big[\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\sigma_k^{(n)}(h_\theta^e(\boldsymbol{x}_j^{(u)})-h_\theta^e(\boldsymbol{x}_k^{(n)}))\Big]$ 满足有限差分性质，且对应系数为

$$c_1=c_2=\cdots=c_{n_n}=\frac{1-\gamma}{2n_n}.$$

综合引理 4 及引理 9，令

$$v_1=\sum_{i=1}^{n_p}\frac{\gamma^2}{n_p^2},$$

$$v_2=\sum_{j=1}^{n_u}\frac{(1-\gamma)^2}{4n_u^2},$$

$$v_3=\sum_{k=1}^{n_n}\frac{(1-\gamma)^2}{4n_u^2}.$$

由于假设集$\{h_\theta^e:e\in[d],\theta\in\boldsymbol{T}_e\}$包含$d\cdot K$个函数,有:

$$\begin{aligned}(\mathrm{IV})&\leq(2(\log d+\log K)\cdot(v_1+v_2+v_3))^{\frac{1}{2}}\\&\leq(2(\log d+\log K)\cdot\rho_\gamma(\boldsymbol{Y}))^{\frac{1}{2}}.\end{aligned}$$

同理可证

$$(\mathrm{V})\leq(2(\log d+\log K)\cdot\rho_\gamma(\boldsymbol{Y}))^{\frac{1}{2}}.$$

综合(IV)及(V)上界,引理得证. 证毕.

D.6. 引理 12 证明.

证明.

步骤 1. 固定所有训练样本类别$\boldsymbol{Y}$,由引理 6 中的第二式构造$\sup_{f\in co(\mathcal{H}_{DS})}[\mathbb{E}_{\mathcal{S}'}\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f))]$的大概率上界,且在上界中引入 PNU-Rademacher 复杂度.由引理 10,有:

$$\mathbb{E}_{\mathcal{S},\mathcal{S}'}\Big[\sup_{f\in co(\mathcal{H}_{DS})}(\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f))\Big]\leq4\mathfrak{R}_{PNU}(\ell_\rho\circ co(\mathcal{H}_{DS})) \tag{67}$$

给定训练样本$\mathcal{S}$,定义$\mathcal{S}_o=(\mathcal{S}/\{(\boldsymbol{x}_o,y_o)\})\cup\{(\tilde{\boldsymbol{x}}_o,y_o)\}$.考察$\sup_{f\in co(\mathcal{H}_{DS})}[\mathbb{E}_{\mathcal{S}'}\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)]$相对于训练样本的有限差分性质系数.

若当前$y_o=1$,有:

$$\begin{aligned}c_o=&\sup_{\boldsymbol{x}_o^{(p)},\tilde{\boldsymbol{x}}_o}\Big|\sup_{f\in co(\mathcal{H}_{DS})}(\mathbb{E}_{\rho,\mathcal{S}}\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f))-\\&\sup_{f\in co(\mathcal{H}_{DS})}(\mathbb{E}_{\mathcal{S}_i}\hat{R}_{\rho,\mathcal{S}_i}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}_i}^{PNU}(f))\Big|\\\leq&\sup_{\boldsymbol{x}_o^{(p)},\tilde{\boldsymbol{x}}_o^{(p)}}\sup_{f\in co(\mathcal{H}_{DS})}\Big|\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}_i}^{PNU}(f)\Big|\\=&\sup_{\boldsymbol{x}_o^{(p)},\tilde{\boldsymbol{x}}_o^{(p)}}\sup_{f\in co(\mathcal{H}_{DS})}\Bigg[\frac{\gamma}{n_pn_n}\sum_{k=1}^{n_n}|\ell_\rho(f,\boldsymbol{x}_o^{(p)},\boldsymbol{x}_k^{(n)})-\\&\ell_\rho(f^{(i)},\tilde{\boldsymbol{x}}_o^{(p)},\boldsymbol{x}_k^{(n)})|+\\&\frac{1-\gamma}{2n_pn_u}\sum_{j=1}^{n_u}|\ell_\rho(f,\boldsymbol{x}_o^{(p)},\boldsymbol{x}_j^{(u)})-\ell_\rho(f,\tilde{\boldsymbol{x}}_o^{(p)},\boldsymbol{x}_j^{(u)})|\Bigg].\end{aligned}$$

由此可得

$$c_o\leq\Big(\gamma+\frac{1-\gamma}{2}\Big)\frac{1}{\rho\cdot n_p}\leq\frac{1}{\rho\cdot n_p}.$$

同理,当$y_o=-1$有$c_o\leq\frac{1}{\rho\cdot n_n}$;当$y_o=0$有$c_o\leq\frac{1}{\rho\cdot n_u}$.此时令

$$\begin{aligned}v=&\frac{1}{\rho}\sum_{o=1}^{N}c_o^2=\frac{1}{\rho}\cdot\max\Big\{\gamma^2,\frac{(1-\gamma)^2}{4}\Big\}\Big(\frac{1}{n_p}+\frac{1}{n_u}+\frac{1}{n_n}\Big)\\\leq&\frac{1}{\rho}\cdot\Big(\frac{1}{n_p}+\frac{1}{n_u}+\frac{1}{n_n}\Big).\end{aligned}$$

对$\sup_{f\in co(\mathcal{H}_{DS})}[\mathbb{E}_{\mathcal{S}'}\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)]$运用引理 6 第二式,可知,下式至少以$1-\frac{\delta}{2}$概率成立:

$$\begin{aligned}&\sup_{f\in co(\mathcal{H}_{DS})}[\mathbb{E}_{\mathcal{S}'}\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)]\\&\leq\mathbb{E}_{\mathcal{S}}\sup_{f\in co(\mathcal{H}_{DS})}[\mathbb{E}_{\mathcal{S}'}\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)]+\\&\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\Big(\frac{1}{2n_p}+\frac{1}{2n_u}+\frac{1}{2n_n}\Big)\Big)^{\frac{1}{2}}\end{aligned} \tag{68}$$

为引入 PNU-Rademacher 复杂度,构造$\mathbb{E}_{\mathcal{S}}[\sup_{f\in co(\mathcal{H}_{DS})}(\mathbb{E}_{\mathcal{S}'}\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f))]$的上界,根据 Jensen 不等式:

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\Big[\sup_{f\in co(\mathcal{H}_{DS})}(\mathbb{E}_{\mathcal{S}'}\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f))\Big]\\&=\mathbb{E}_{\mathcal{S}}\sup_{f\in co(\mathcal{H}_{DS})}\mathbb{E}_{\mathcal{S}'}[\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)]\\&=\mathbb{E}_{\mathcal{S},\mathcal{S}'}\Big[\sup_{f\in co(\mathcal{H}_{DS})}(\hat{R}_{\rho,\mathcal{S}'}^{PNU}(f)-\hat{R}_{\rho,\mathcal{S}}^{PNU}(f))\Big].\end{aligned}$$

因此有对于任意$f\in co(\mathcal{H}_{DS})$,下式至少以概率$1-\frac{\delta}{2}$成立

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)\leq&\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)+4R_{PNU}(\ell_\rho\circ co(\mathcal{H}_{DS}))+\\&\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\Big(\frac{1}{2n_p}+\frac{1}{2n_u}+\frac{1}{2n_n}\Big)\Big)^{\frac{1}{2}}\end{aligned} \tag{69}$$

至此步骤 1 证毕.

步骤 2. 固定样例类别,对$\mathfrak{R}_{PNU}(\ell_\rho\circ co(\mathcal{H}_{DS}))$运用引理 6 第一式,在大概率上界中由$\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell_\rho\circ\mathcal{H}_{DS})$替换$\mathfrak{R}_{PNU}(\ell_\rho\circ co(\mathcal{H}_{DS}))$,并由引理 11 将$\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell_\rho\circ\mathcal{H}_{DS})$化简.

通过再次考察$\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell_\rho\circ\mathcal{H}_{DS})$相对于训练样本的有限差分性质,可得下式至少以$1-\frac{\delta}{2}$概率成立:

$$\begin{aligned}&\hat{\mathfrak{R}}_{PNU}(\ell_\rho\circ co(\mathcal{H}_{DS}))\leq\\&\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell_\rho\circ\mathcal{H}_{DS})+\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\Big(\frac{1}{2n_p}+\frac{1}{2n_u}+\frac{1}{2n_n}\Big)\Big)^{\frac{1}{2}}\end{aligned} \tag{70}$$

联立式(69)及式(70)以及概率的次可加性,下式以$1-\delta$概率成立:

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)\leq&\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)+4\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell_\rho\circ\mathcal{H}_{DS})+\\&\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\Big(\frac{1}{2n_p}+\frac{1}{2n_u}+\frac{1}{2n_n}\Big)\Big)^{\frac{1}{2}}\end{aligned} \tag{71}$$

代入引理 D.5,有:

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)\leq&\hat{R}_{\rho,\mathcal{S}}^{PNU}(f)+\frac{8\sqrt{2}}{\rho}((\log d+\log K)\cdot\chi(\boldsymbol{Y}))^{\frac{1}{2}}+\\&\frac{1}{\rho}\Big(\log\Big(\frac{2}{\delta}\Big)\cdot\chi(\boldsymbol{Y})\Big)^{\frac{1}{2}}\end{aligned} \tag{72}$$

至此步骤 2 证毕.

步骤 3. 解除对$\boldsymbol{Y}$的固定.

推导过程同工作[15]中定理 8,可证明式(72)可以至少$1-\delta$的概率成立. 证毕.

YANG Zhi-Yong, Ph. D. His research interests include theoretical and algorithmic aspects of machine learning.

XU Qian-Qian, Ph. D., associated professor. Her research interests include statistical machine learning, with applications in multimedia and computer vision.

HE Yuan, Ph. D., senior staff engineer. His research interests include computer vision, machine learning and AI security.

CAO Xiao-Chun, Ph. D., professor. His main research interests include computer vision and multimedia analysis.

HUANG Qing-Ming, Ph. D., chair professor. His research areas include multimedia computing, image processing, computer vision and pattern recognition.

Background

AUC (Area Under the ROC Curve) is a well-known metric targeting at imbalance learning problems. By definition, AUC is the area under the Receiver Operating Characteristic curve plotted by sensitivity against specificity. More specifically, it works by separating the instance distributions conditioned on the positive/negative labels, which naturally avoids the dependency on the label distribution priors. Besides this merit, AUC is also known to have other good properties such as insensitivity toward costs, thresholds, and magnitude of the classification scores.

Over the past two decades, the importance of AUC has raised an increasing favor in the machine learning community to explore direct AUC optimization methods. Nonetheless, the vast majority of studies along this line merely focus on the fully supervised setting. In the recent few years, there is a new wave to study semi-supervised AUC optimization problems which has achieved partial success. The existing studies only focus on single-model-based semi-supervised AUC optimization methods, while leaving how to effectively integrate multiple models as an open problem.

To address this issue, this paper studies the problem of how to ensemble multiple semi-supervised AUC optimization methods under the context of the boosting strategy. Specifically, we propose a boosting-based semi-supervised AUC optimization method. On top of this, we provide an acceleration strategy based on weight decoupling to reduce the time and space complexity. Moreover, we theoretically prove that the proposed algorithm has an exponential convergence rate with respect to the increase of weak learners. Meanwhile, we provide a generalization error bound of the proposed method, and further prove that increasing the number of weak learners could improve the performance on the training set without the cost of a significant overfitting effect. Finally, we evaluate our proposed framework on 16 benchmark datasets. Experimental results show that the proposed algorithm outperforms all the competitors with a significance level of 0.05, and achieves a 0.9%—11.28% performance gain in average.

This work was supported in part by the National Key R&D Program of China under Grant No. 2018AAA0102003, in part by the National Natural Science Foundation of China (Nos. 61620106009, 61931008, 61836002, U2001202, 61976202), in part by the Fundamental Research Funds for the Central Universities, in part by the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB28000000), in part by the National Postdoctoral Program for Innovative Talents under Grant No. BX2021298, in part by the Youth Innovation Promotion Association CAS, in part by the Alibaba Group through Alibaba Research Fellowship Program.