计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(12):375−382 [doi: 10.15888/j.cnki.csa.008871]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于联邦学习和重要性加权的疾病得分预测①

许亚倩[1], 崔文泉[1], 程浩洋[2]

[1](中国科学技术大学 管理学院 统计与金融系, 合肥 230026)
[2](衢州学院 电气与信息工程学院, 衢州 324000)
通信作者: 崔文泉, E-mail: wqcui@ustc.edu.cn

摘　要: 在考虑用户隐私的保护多源域数据背景下预测疾病得分的问题中, 来自不同源域的数据分散存储无法合并, 且可能服从不同的分布, 因此传统的机器学习方法无法合理地利用源域数据的信息. 本文结合联邦学习的思想和基于样本的迁移学习方法, 提出了联邦重要性加权方法, 通过将源域的样本重用于目标域的预测任务, 而且不需要进行源域之间的数据共享, 实现了在保护源域的数据隐私的情况下利用分布不同的多源域的信息提升目标域预测的精度. 并且基于提出的方法, 本文构造了一种加权模型并提供了一个简洁通用的算法用于求解目标域的预测模型. 数值模拟以及实证结果表明, 相对于未考虑分布迁移的传统方法, 联邦重要性加权方法可以有效地利用多源域数据的信息, 在目标域的预测精度上具有优势, 以及在帕金森疾病数据中做出精准的疾病得分预测.

关键词: 联邦学习; 迁移学习; 重要性加权; 加权模型; 疾病得分; 机器学习; 隐私保护

引用格式: 许亚倩,崔文泉,程浩洋.基于联邦学习和重要性加权的疾病得分预测.计算机系统应用,2022,31(12):375–382. http://www.c-s-a.org.cn/1003-3254/8871.html

Disease Scores Predicting Based on Federated Learning and Importance Weighting

XU Ya-Qian[1], CUI Wen-Quan[1], CHENG Hao-Yang[2]

[1](Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China)
[2](College of Electrical and Information Engineering, Quzhou University, Quzhou 324000, China)

Abstract: In the problem of predicting disease scores amid the protection of multi-source domain data considering user privacy, the decentralized data from different source domains cannot be combined and may follow different distributions. Therefore, traditional machine learning methods cannot be applied directly to utilize the information within source domains. In this study, the federated importance weighting method is proposed combining the idea of federated learning and the sample-based transfer learning approach. By re-weighting the samples from the source domains to the prediction task of the target domain, and without data sharing between multiple source domains, it realizes the use of data with different distributions while protecting the data privacy of the source domains. Moreover, this study constructs a weighted model and provides a concise and general algorithm to solve the prediction model for the target domain. Numerical simulation and empirical results show that, compared with the traditional method without considering distribution shift, the federated importance weighting method can effectively utilize the information of the source domain data. It is superior in prediction accuracy of the target domain and can make an accurate prediction of disease scores in the Parkinson's disease data.

Key words: federated learning; transfer learning; importance weighting; weighted model; disease scores; machine learning; privacy protection

① 基金项目: 国家自然科学基金 (71873128, 12171451)
收稿时间: 2022-04-16; 修改时间: 2022-05-22, 2022-05-28, 2022-06-06; 采用时间: 2022-06-13; csa 在线出版时间: 2022-08-12

1 引言

随着数据挖掘方法在各产业中的繁荣发展, 用户数据可以被更加充分地利用以提升生产效率以及用户体验[1]. 然而, 在这些被大规模使用的数据中可能包含用户的隐私信息, 例如个人医疗健康报告[2], 家庭收入与消费记录[3], 以及商业营销数据[4] 等. 因此, 大数据应用中的用户隐私问题受到了越来越多的关注[5]. 为了有效地防止用户隐私数据泄露, 发展关于隐私保护的大数据工具也逐渐成为研究热点[6].

经典的机器学习方法认为训练数据和测试数据来自相同分布, 而且只有一个训练集. 为了训练机器学习算法, 我们通常需要搜集充分多的样本, 来得到算法中参数的准确估计. 然而在实际应用中, 为了保护用户的隐私数据, 我们可能无法确保满足上述要求. 首先, 在现实的隐私保护框架下, 数据经常呈现出数据孤岛的现象[7], 并且单个数据源中拥有的数据量可能不足, 这导致我们仅仅在目标域上进行传统的统计分析或大数据挖掘无法训练得到精准的算法参数, 而且也无法简单地通过将分散的数据汇总来增加目标域的样本量[8]. 除此之外, 由于数据拥有者们可能来自不同的地区或者数据产生于不同的时间, 这些来自不同源域的数据可能因此具有不同的分布[9]. 这导致我们无法在目标域上直接使用其他源域中机器学习算法的训练结果. 一个实际例子是本文关注的医学自动诊断领域中基于多数据源域的帕金森疾病得分预测问题[10]. 某非医学的机构需要预测一家医院 (目标医院) 的早期帕金森患者的疾病得分, 该机构已拥有待预测患者关于帕金森疾病的相关生理指标数据, 而未获得疾病得分. 同时, 该机构可以访问其他一些医院 (源医院) 的帕金森病人数据, 包括相关生理指标数据和疾病得分数据. 出于保护隐私以及预防数据被滥用的目的, 每位病人的疾病数据只在其所属医院本地保存, 而不传输到其他机构中. 通常这些源医院和目标医院来自不同的地区, 各医院的病人数据也发生于不同的时间. 由于人种基因、饮食条件、运动习惯、文化环境影响的认知模式和思维方式等的不同, 不同医院的病人在相关生理指标数据上可能会有很大的不同, 而且呈现出区域性的不同. 然而医学上根据相关生理指标数据计算疾病得分的原理对所有人都是相同的.

为了充分利用分散的数据, 联邦学习通过训练不同源域上的算法进行协作而无需共享自身数据来解决多源域数据处理中的隐私问题[8]. 该方法最近在大数据医疗领域中获得了广泛的关注[11]. 近期研究表明, 联邦学习训练的模型可以达到与在中心训练的绩效水平相当的性能水平. Dowlin 等人[12] 提出加密网以提高数据加密的效率从而实现更好的联邦学习性能. Bonawitz 等人[13] 引入一种聚合机制以在联邦学习框架下更新机器学习模型. Mohassel 等人[14] 提出了联邦学习系统中支持多客户端隐私保护合作培训的安全机器学习. Smith 等人[15] 提出联邦多任务学习, 为每个节点学习一个单独的模型, 其研究成果表明, 多任务学习适合于应对联邦学习中的统计学难题. Liu 等人[16] 展示了一种能够灵活地应用于各种多方安全机器学习任务的半监督联邦迁移学习框架, 允许知识在网络中通过迁移学习进行传输, 且不必损害用户隐私. Cheng 等人[17] 提出了对于纵向联邦学习的安全性提升方法, 这是一种新颖的、无性能损失的、保护隐私的提升树系统架构.

联邦学习可以解决数据分散的问题, 使得孤岛形式的多源域数据可以被充分合理地使用[7]. 而在每个源域本地学习的过程中, 如果源域和目标域的分布不同, 传统的机器学习方法不再适用, 迁移学习是一种有效的模型训练方法, 它旨在从可能与目标域不同的源域中提取知识并将知识应用于目标任务[9]. 迁移学习的研究是为了智能地应用以前学到的知识来更快地解决新问题或提供更好的解决方案. 迁移学习的方法可以分为 4 类: 基于样本的迁移方法[18], 其中主要的技术如样本重新加权和重要性抽样; 基于特征表示的迁移方法[19], 主要想法是通过源域为目标域学习一个特征表示, 其中用于跨域传输的知识被编码到学习的特征表示中; 基于参数的迁移方法[20], 假设源任务和目标任务共享一些参数或者超参数的先验分布, 其中迁移的知识被编码到共享的参数或者先验中; 基于关系的知识迁移[21], 基本假设是源域和目标域中的数据之间的某些关系是相似的, 要传输的知识是数据之间的关系. 通过纠正每个样本数据采样过程中的偏差进行迁移学习的方法, 侧重于数据重要性权重[22] 或类重要性权重[23]. 重要性加权方法赋予每个样本一个密度比形式的权重, Sugiyama 等人[24] 提出在密度比已知情况下的重要性加权交叉验证方法, 将验证风险加权为目标风险的无偏估计. 当密度比未知时, 有一些直接估计密度比的方法, 比如, 核平均匹配方法[25], Kullback-Leiber 重要性估计方法[26], 最小二乘重要性拟合方法[27].

目前, 联邦学习以及迁移学习都分别有丰富的研究以及广泛的应用[28]. 然而, 目前在不同源域和目标域具有相同的特征空间但是允许特征分布不同的数据背景和源域拥有隐私保护的要求下对目标数据进行预测的研究还十分有限. 因此本文针对上述数据特点, 提出了联邦重要性加权方法, 该方法结合了联邦学习和迁移学习中基于样本的重要性加权方法. 联邦重要性加权方法首先在每个源域本地学习一个模型而不用汇总或者传输源域数据以在满足源域数据隐私的要求下充分利用所有源域数据. 在每个源域本地学习中, 我们的方法通过重要性加权方法重用源域样本为目标域的风险函数提出一个合理估计, 通过最小化目标风险的估计获得该源域对目标域贡献的一个模型, 从而解决多源域和目标域之间分布不同的问题. 然后, 本文首次提出, 根据各源域和目标域的分布差异为各源域模型构造权重从而将所有源域模型加权整合成一个模型用于目标域的预测任务.

本文的其余部分组织如下. 在第 2 节中, 我们介绍了本文研究的问题设置. 第 3 节介绍了联邦重要性加权方法的具体内容和执行该方法的具体算法. 第 4 节进行数值模拟, 将目前常用的方法与本文提出的方法进行比较. 第 5 节将联邦重要性加权方法应用于预测帕金森疾病得分. 第 6 节总结了全文.

2 问题设置

考虑上文所述医院的实例, 抽象出研究问题的设置, 本文考虑m个源和一个目标的情况, 源和目标都有一个域和一个任务[29]. 令$\mathcal{X}\subset\mathbb{R}^d$表示$d$维特征空间, $\mathcal{Y}\subset\mathbb{R}$表示实值输出空间. 用$P(x)$表示特征变量$X\in\mathcal{X}$的概率分布, 我们将第$j$个源域和目标域分别表示为: $\{\mathcal{X},P_j(x)\}$和$\{\mathcal{X},P_t(x)\}$.

第j个源任务和目标任务分别表示为: $\{\mathcal{Y},f_j(\cdot)\}$和$\{\mathcal{Y},f_t(\cdot)\}$. 其中, $f(\cdot)$是根据特征变量的观察值预测输出值的函数. 用$p_j(x)$, $p_t(x)$, $p_j(x,y)$和$p_t(x,y)$分别表示第j个源和目标的特征密度函数和联合概率密度函数.

将第j个源的 i.i.d.数据集表示为: $S_j=\{(x_{ji},y_{ji})\}_{i=1}^{n_j}$, 其中, $x_{ji}\in\mathcal{X}$是特征变量的观察值, $y_{ji}\in\mathcal{Y}$是相应的输出值. 类似地, 用: $T=\{(x_{ti})\}_{i=1}^{n_t}$表示目标的 i.i.d.数据集.

本文考虑以下条件.

1) 源支撑目标, 且源和目标的特征分布不同, 即对于每个$j=1,2,\cdots,m$, $p_t(x)>0\Rightarrow p_j(x)>0$, $p_t(x)\neq p_j(x)$.

2) 源和目标的输出变量关于特征变量的条件分布相同, 即: $p_t(y|x)=p_j(y|x)$, $j=1,2,\cdots,m$.

这两个条件确保源数据可用于目标任务[30]. 对应到上述的实际问题, X和$P(x)$是用来计算疾病得分的相关生理指标以及它们的概率分布, Y是疾病得分. 条件 2) 的实际意义是在医学中根据相关生理指标数据计算疾病得分的原理是相同的.

除此之外, 在我们的设置中每个源有隐私要求, 数据只保存在本地. 本文的目的是在这样的问题设置下学习一个预测目标输出值的参数模型.

3 联邦重要性加权方法

3.1 超参数选优

为了获得一个准确的参数模型, 需要找到目标分布上的最优超参数:

$$\theta^*=\arg\min_{\theta\in\Theta}R_t(\theta) \tag{1}$$

其中, $R_t(\theta)$是超参数为$\theta\in\Theta$时定义在目标上的风险函数, $R_t(\theta)=\mathbb{E}_{(X,Y)\sim p_t(x,y)}[l(h(X;\omega,\theta),Y)]$, 我们称之目标风险. Θ是超参数搜索空间. $l:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}^+$是损失函数. $h(x;\omega,\theta):\mathcal{X}\to\mathcal{Y}$是超参数为$\theta\in\Theta$, 模型参数为$\omega\in\Omega$, 根据特征变量的观测值$x$预测输出值的参数模型, Ω是模型参数的取值空间.

如果我们已拥有目标输出值, 令$T^{\text{op}}=\{(x_i,y_i)\}_{i=1}^{n_t^{\text{op}}}$表示目标输出值可获得时的 i.i.d.数据集, 其中, op 表示 output, 用以简记输出值已获得的情形. 我们将其划分为两个不相交的子集$T_{\text{tr}}^{\text{op}}\cup T_{\text{val}}^{\text{op}}$, 样本量分别为$n_t^{\text{tr}}$和$n_t^{\text{val}}$. 传统的超参数优化方法用如下经验均值估计目标风险[31]:

$$\hat{R}^{\text{op}}(\theta)=\frac{1}{n_t^{\text{val}}}\sum_{(x,y)\in T_{\text{val}}^{\text{op}}}l(h(x;\hat{\omega}(\theta),\theta),y) \tag{2}$$

在$T_{\text{tr}}^{\text{op}}$上估计模型参数:

$$\hat{\omega}(\theta)=\arg\min_{\omega\in\Omega}\frac{1}{n_t^{\text{tr}}}\sum_{(x,y)\in T_{\text{tr}}^{\text{op}}}l(h(x;\omega,\theta),y)$$

然后, 把$\hat{R}^{\text{op}}(\theta)$代入式 (1) 获得最优超参数$\theta^*$的估计.

然而, 在第 2 节描述的问题设置中, 目标域的数据的输出值无法获得, 上述过程不再适用. 因此我们需要利用输出值可获得的源数据, 但是源域和目标域的分布不同, 无法直接训练机器学习方法. 迁移学习可以很好地解决分布不同的问题, 将源域数据充分合理地用于目标任务. 由于每个源域的数据只保存在本地, 我们

根据联邦学习的思想让每个源域在本地学习一个参数模型. 在每个源域本地学习的过程中, 本文使用基于样本的迁移学习方法, 将分布不同的源域样本通过重要性加权重用于目标域的预测任务, 以解决分布不同的问题.

3.2 各源本地的重要性加权估计

每个源在本地学习中使用重要性加权方法[27], 这个过程涉及密度函数的比值, 为此我们对$p_j(x,y)>0$, $\forall(x,y)\in(\mathcal{X},\mathcal{Y})$的第$j$个源定义其与目标的密度比函数:

$$r_j(x)=\frac{p_t(x,y)}{p_j(x,y)}=\frac{p_t(x)}{p_j(x)}$$

其中, 第 2 个等式可由第 2 节中的条件 2) 推导得到.

由于在我们的超参数优化过程中涉及密度比估计、训练模型参数、估计最优超参数 3 部分, 我们把S_j分为 3 个不相交的子集:

$$S_j^{\mathrm{de}}\cup S_j^{\mathrm{tr}}\cup S_j^{\mathrm{val}}$$

其中,

$$S_j^{\mathrm{de}}=\{(x_{ji}^{\mathrm{de}},y_{ji}^{\mathrm{de}})\}_{i=1}^{n_j^{\mathrm{de}}}$$

$$S_j^{\mathrm{tr}}=\{(x_{ji}^{\mathrm{tr}},y_{ji}^{\mathrm{tr}})\}_{i=1}^{n_j^{\mathrm{tr}}}$$

$$S_j^{\mathrm{val}}=\{(x_{ji}^{\mathrm{val}},y_{ji}^{\mathrm{val}})\}_{i=1}^{n_j^{\mathrm{val}}}$$

令:

$$n_j=n_j^{\mathrm{de}}+n_j^{\mathrm{tr}}+n_j^{\mathrm{val}}$$

$$n^{\mathrm{tr}}=n_1^{\mathrm{tr}}+n_2^{\mathrm{tr}}+\cdots+n_m^{\mathrm{tr}}$$

$$n^{\mathrm{val}}=n_1^{\mathrm{val}}+n_2^{\mathrm{val}}+\cdots+n_m^{\mathrm{val}}$$

对于给定的超参数$\theta\in\Theta$, 在第j个源本地, 目标风险的经验估计式 (2) 被重要性加权方法调整为:

$$\hat{R}_{IW}^j(\theta)=\frac{1}{n_j^{\mathrm{val}}}\sum_{i=1}^{n_j^{\mathrm{val}}}\hat{r}_j(x_{ji}^{\mathrm{val}})l(h(x_{ji}^{\mathrm{val}};\hat{\omega}(\theta),\theta),y_{ji}^{\mathrm{val}}) \tag{3}$$

其中, 密度比估计$\hat{r}_j(\cdot)$通过无约束最小二乘重要性拟合 (unconstrained least-squares importance fitting, uLSIF[27]) 方法在数据集S_j^{de}和T上得到. 模型参数在S_j^{tr}上估计得到:

$$\hat{\omega}(\theta)=\arg\min_{\omega\in\Omega}\frac{1}{n_j^{\mathrm{tr}}}\sum_{i=1}^{n_j^{\mathrm{tr}}}l(h(x_{ji}^{\mathrm{tr}};\omega,\theta),y_{ji}^{\mathrm{tr}}) \tag{4}$$

第j个源本地估计超参数如下:

$$\hat{\theta}^j=\arg\min_{\theta\in\Theta}\hat{R}_{IW}^j(\theta) \tag{5}$$

于是, 相应的第j个源模型为:

$$h(\cdot;\hat{\omega}(\hat{\theta}^j),\hat{\theta}^j)$$

在获得各源在本地学习到的参数模型之后, 我们需要将这m个源模型贡献于目标任务.

3.3 加权模型

我们构造如式 (6) 所示加权模型:

$$\hat{h}_t(\cdot)=\sum_{j=1}^{m}\hat{\beta}(\hat{\theta}^j)h(\cdot;\hat{\omega}(\hat{\theta}^j);\hat{\theta}^j) \tag{6}$$

其中, $\hat{\beta}(\hat{\theta}^j)$是第$j$个源的模型权重, 我们根据 Nomura 等人[32] 提出的方法中将模型权重构造如式 (7):

$$\hat{\beta}(\hat{\theta}^j)=\frac{n_j}{\widehat{d}(\hat{\theta}^j)}\Big/\left(\sum_{k=1}^{m}\frac{n_k}{\widehat{d}(\hat{\theta}^k)}\right) \tag{7}$$

其中,

$$\widehat{d}(\hat{\theta}^j)=\frac{1}{n_j^{\mathrm{val}}}\sum_{i=1}^{n_j^{\mathrm{val}}}[\hat{r}_j(x_{ji}^{\mathrm{val}})l(h(x_{ji}^{\mathrm{val}};\hat{\omega}(\hat{\theta}^j),\hat{\theta}^j),y_{ji}^{\mathrm{val}})]^2-\left[\frac{1}{n_j^{\mathrm{val}}}\sum_{i=1}^{n_j^{\mathrm{val}}}\hat{r}_j(x_{ji}^{\mathrm{val}})l(h(x_{ji}^{\mathrm{val}};\hat{\omega}(\hat{\theta}^j),\hat{\theta}^j),y_{ji}^{\mathrm{val}})\right]^2$$

根据式 (7) 可以验证$\sum_{j=1}^{m}\hat{\beta}(\hat{\theta}^j)=1$.

3.4 算法

为了描述我们提出的联邦重要性加权方法在具体执行时的流程, 我们设计联邦重要性加权算法如算法 1.

算法 1. 联邦重要性加权算法

输入: 无输出值的目标数据集T, 有输出值的m个源数据集S_j,$j=1,\cdots,m$, 超参数搜索空间Θ, 参数模型 h, 损失函数 l

输出: 目标任务的预测模型, $\hat{h}_t(\cdot)$

1) 目标数据的拥有者将T分别传输给每个源.
2) for j=1 to m
   第j个源的数据拥有者:
   a) 把S_j分成 3 个不相交的子集: $S_j^{\mathrm{de}}\cup S_j^{\mathrm{tr}}\cup S_j^{\mathrm{val}}$.
   b) 在T和S_j^{de}上通过 uLSIF[27] 方法估计密度比函数, $\hat{r}_j(\cdot)$.
   c) 在S_j^{tr}和S_j^{val}上训练模型, 根据式 (3)–式 (5) 获得模型, $h(\cdot;\hat{\omega}(\hat{\theta}^j),\hat{\theta}^j)$.
   d) 在S_j^{val}上根据式 (7) 计算$\hat{\beta}(\hat{\theta}^j)$.
   e) 把$h(\cdot;\hat{\omega}(\hat{\theta}^j),\hat{\theta}^j)$和$\hat{\beta}(\hat{\theta}^j)$传输给目标数据的拥有者.
3) 目标数据的拥有者:
   根据式 (6) 加权各源模型.
4) 返回$\hat{h}_t(\cdot)=\sum_{j=1}^{m}\hat{\beta}(\hat{\theta}^j)h(\cdot;\hat{\omega}(\hat{\theta}^j);\hat{\theta}^j)$.

算法 1 首先让目标将数据分别传输给每个源, 让每个源在本地利用自己的数据和目标的数据学习一个

模型, 将学习得到的参数传输给目标. 因此算法 1 可以很好地保证源域的数据隐私. 为了解决源域和目标域之间分布不同的问题, 算法 1 在每个源域本地学习的过程中使用了基于样本的迁移学习方法, 并且根据每个源域和目标域的分布差异为源域构造了模型权重, 由各源域计算模型权重并传输给目标域. 最终, 目标域获得一个加权模型, 作为自己的预测模型.

4 数值分析

本节我们先通过模拟实验测试我们的方法, 然后在实际的帕金森疾病数据集[10] 上做预测. 后面 FedIW 表示本文提出的方法, 并且每次实验中都与以下 4 个方法对比:

1) Naive: 在每个源本地学习时, 不考虑源和目标之间的协变量移位, 通过最小化源域的损失函数获得各源的超参数.

2) Refer: 假设目标数据的输出值可获得, 并只在目标数据上训练模型, 这在我们的设置中是不可行的, 报告其结果作为参考.

3) LI[33]: 通过最小化元损失函数来找到有希望的超参数. 元损失函数定义为每个源域上的损失函数 (式 (3)) 的总和. 直观地说, 通过最小化元损失函数, LI 估计最优超参数.

4) MS-CS[32]: 根据每个源域和目标域的分布差异为目标风险构造一个估计, 将每个源域上的估计加起来作为目标函数, 通过最小化目标函数获得超参数.

4.1 模拟数据

为了验证我们提出的方法的有效性, 我们用 10 维向量$x=(x_1,x_2,\cdots,x_{10})$模拟疾病的相关生理指标数据, 用$y=\tilde{x}^2+\epsilon$, $\epsilon\sim N(\tilde{x},1)$模拟疾病得分 (输出变量). 其中, $\tilde{x}=(x_1+\cdots+x_{10})/10$. 目标域的特征变量的分布为$N(0_{10},I_{10})$, 0_{10}和I_{10}分别表示元素全为 0 的 10 维向量和 10 维的单位矩阵.

在训练过程中, 我们使用岭回归模型, 超参数是正则化参数λ, 超参数搜索空间$\Theta=[0,1]$, 损失函数为平方损失函数:

$$l(\hat{y},y)=(\hat{y}-y)^2$$

我们用平均绝对误差报告预测误差:

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i|$$

我们进行以下几组模拟实验.

1) 源和目标的样本量都为 100, 第j个源的分布为:

$$c_j\sim U(-5,5),\ x\sim N(c_j\times 1_{10},I_{10})$$

其中, 1_{10}表示元素全为 1 的 10 维向量. 源的个数m依次取 3, 4, 5, 6, 7, 用不同的随机种子进行 100 次实验.

2) 两个源, 特征分布分别为:

$$N(0_{10},I_{10}),\ N(2\times 1_{10},I_{10})$$

样本量分别记为n_1和n_2, 目标的样本量记为n_t, 对于 5 组不同的样本量各根据不同的随机种子进行 100 次实验.

3) 目标样本量为 50, 两个源, 特征分布分别为:

$$N(0_{10},I_{10}),\ N(c\times 1_{10},I_{10})$$

对应的样本量分别为 100 和 200, 对于$c=1,2,3,4$的每个取值, 用不同的随机种子各进行 100 次实验.

4.2 模拟结果

我们将模拟实验 1)–3) 的 100 次实验结果的平均值和标准差分别提供在表 1–表 3 中. 将模拟实验 2) 的实验结果的平均值和标准差关于不同样本量组合的变化情况分别绘制成图 1 和图 2. 将模拟实验 3) 的实验结果的平均值关于c的取值绘制成图 3.

表 1 模拟实验 1) 中各方法的表现 (平均表现±标准差)

m	FedIW	LI	MS-CS	Naive	Refer
3	**0.8482±0.0972**	7.5220±0.2781	7.3113±0.2750	7.9292±0.2825	0.9956±0.1652
4	**2.5108±0.2268**	6.2717±0.2673	6.7423±0.2905	7.0714±0.2798	0.9956±0.1652
5	**1.7180±0.2840**	5.5260±0.2276	5.1914±0.2237	5.1347±0.2191	0.9956±0.1652
6	**2.2419±0.5055**	5.0988±0.2339	4.6867±0.2288	4.9154±0.2252	0.9956±0.1652
7	**1.1400±0.2822**	5.8623±0.1793	5.9753±0.2629	6.2453±0.1833	0.9956±0.1652

表 2 模拟实验 2) 中各方法的表现 (平均表现±标准差)

n_t	n_1	n_2	FedIW	LI	MS-CS	Naive	Refer
50	100	200	**0.9282±0.1356**	2.6687±0.3131	2.5905±0.3087	2.7876±0.3129	1.0095±0.2468
100	200	400	**0.8213±0.1141**	3.1196±0.1906	2.3173±0.1666	2.9483±0.1965	0.8304±0.1231
150	300	600	**0.8467±0.0733**	2.5070±0.1504	1.9032±0.1356	2.2801±0.1434	0.7822±0.0895
200	400	800	**0.7849±0.0670**	2.9953±0.1370	2.2170±0.1232	2.6959±0.1333	0.8774±0.0982
250	500	1 000	**0.7695±0.0594**	2.8088±0.1230	2.0992±0.1106	2.6897±0.1217	0.7544±0.0643

表 3 模拟实验 3) 中各方法的表现 (平均表现±标准差)

c	FedIW	LI	MS-CS	Naive	Refer
1	**1.0652±0.2133**	1.2361±0.1867	1.1307±0.1624	1.3341±0.1964	1.0095±0.2468
2	**1.0225±0.1444**	2.8900±0.3096	2.2697±0.2582	3.1448±0.3292	1.0095±0.2468
3	**0.9679±0.1473**	6.1850±0.3866	4.6967±0.3213	6.6084±0.4083	1.0095±0.2468
4	**1.4993±0.2632**	10.8020±0.4675	8.1594±0.3757	11.4156±0.4913	1.0095±0.2468

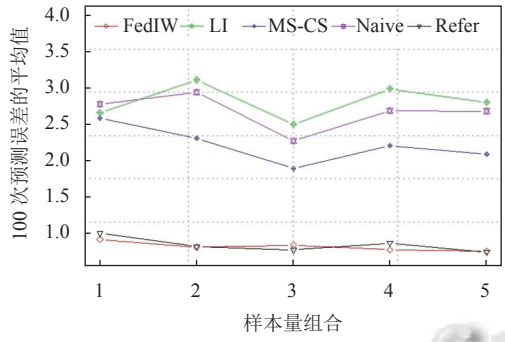

图 1 Case 2 中各方法在不同样本量组合的平均表现

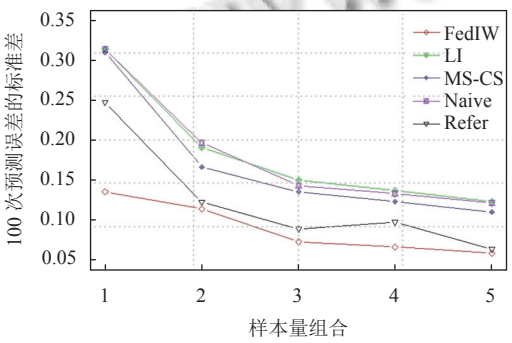

图 2 Case 2 中各方法在不同样本量组合的标准误差

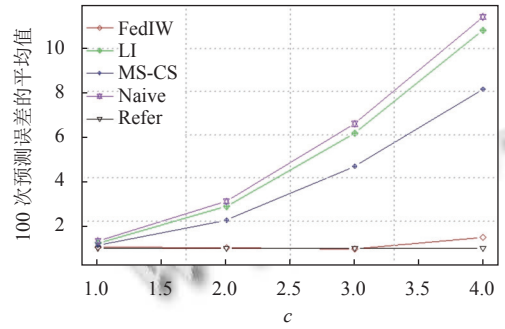

图 3 Case 3 中各方法在c不同取值时的平均表现

4.3 结果分析

表 1 的结果显示在源的个数取不同值时, FedIW 的平均表现优于 Naive, LI 和 MS-CS, 比 Refer 略差一点. 表 2 的结果显示在几组不同的样本量组合下, FedIW 不仅平均表现优于另外 3 种, 而且由图 1 可看出 FedIW 的平均表现和 Refer 很接近. 由图 2 可看出 FedIW 的实验结果的标准差是最小的, 且随着样本量逐渐增大, 各方法的标准差都逐渐减小. 表 3 的结果显示在c的几个取值下, FedIW 的平均表现都优于另外 3 种, 并且和 Refer 相近. 进一步地, 图 3 表明当c的取值越大, 也就是源域和目标域的分布越不同, FedIW 相对 Naive, LI 和 MS-CS 的优势越明显, 说明 FedIW 可以很好地适应源和目标之间的分布迁移.

FedIW 明显优于 Naive, 因为我们提出的联邦重要性加权估计可以很好地适应源和目标之间的分布迁移, 而 Naive 则不能. FedIW 表现优于 LI 和 MS-CS, 说明了我们让每个源域本地学习而不必考虑其他源域的做法以及构造的模型权重有助于提升目标域的预测精度. 有时候, FedIW 的表现很接近 Refer 的表现, 甚至略优于 Refer, 说明我们提出的方法很有效.

5 帕金森疾病得分预测

我们使用帕金森远程监控数据集[10], 该数据集收录了 42 名早期帕金森患者的一系列生物医学语音测量数据, 这些患者被招募参加为期 6 个月的远程监测症状进展测试. 这些记录是在患者家中自动记录的. 在这里, 我们将每个患者的家视为一家小医院, 患者的所有记录都是这家医院拥有的数据. 数据集的列包括受试者编号、受试者年龄、受试者性别、从基线招募日期开始的时间间隔、UPDRS (帕金森病评分量表) 总评分和 16 个生物医学语音测量值. 每行对应于这些人 5 875 条语音记录中的一条. 该数据集的主要目的是预测 16 次语音测量的 UPDRS 总分.

我们选择一个医院作为目标, 把其他的医院都作为源. 各源医院的患者数据不离开本地不允许泄露, 不同的源医院的病人疾病数据不可以混合. 我们依然用岭回归模型训练数据, 超参数为正则化参数. 损失函数是平方损失.

实验过程: 使用不包含输出值的目标数据和包含输出值的源数据; 将目标数据按 7:3 划分为训练集和测试集; 用目标数据的测试集和源数据训练模型; 预测目标数据的测试集的输出值; 计算预测的平均绝对误差; 用不同的随机种子重复 100 次以上步骤.

我们将这 100 次实验结果的平均表现、标准差和最坏情况表现展示在表 4 中, 结果表明 FedIW 在平均和最坏情形下都拥有比其他方法更好的表现, 并且拥有较小的标准差. 特别地, FedIW 的表现比 Refer 更好, 因为在帕金森数据集上的实验中, 源的样本量总和远大于目标样本量, 而 FedIW 可以充分有效地将源样本信息迁移到目标任务.

表 4 各方法在实际数据集上的实验结果

表现	方法				Refer
	FedIW	LI	MS-CS	Naive	
平均表现	**0.8667**	0.8803	0.9030	0.8736	0.9283
标准差	0.0620	**0.0594**	0.0705	0.0628	0.1220
最差情形	**0.9959**	1.0061	1.0330	1.0024	1.3343

6 结论与展望

本文首次在疾病得分预测上提出了一种新的基于联邦学习和迁移学习的联邦重要性加权方法, 为了保证用户隐私保护, 联邦重要性加权方法在各源域中分别训练模型, 基于重要性加权将多个源域中的样本重用于目标域的预测任务, 以适应源和目标分布不同的情形并且有效地利用了各源域提供的信息, 提升了在目标域上的预测精度. 为了将分散的多源域的模型用于目标域的预测任务, 联邦重要性加权方法为每个源域根据它和目标域的分布差异构造了一个模型权重, 若源域的分布与目标域的差异较大, 则会赋予较小的模型权重, 反之赋予较大的模型权重, 从而获得一个加权模型用于目标域的预测任务, 因而联邦重要性加权方法具有突出的预测效果.

在实际数据中, 本文研究了帕金森疾病的得分预测问题. 研究结果显示, 本文的联邦重要性加权方法相比于没有考虑源域和目标域之间存在分布差异的方法、将各源域的模型用于目标域的预测任务时不考虑根据分布差异区别化采用源域模型的方法, 表现出了更加突出且稳定的预测效果.

在我国大数据医疗飞速发展的背景下, 本文的研究结果为某些疾病的自动化诊断提供了一个有效的参考方法, 更对我国医疗的发展有着深远意义.

首先, 联邦重要性加权方法得出的疾病得分预测为相关部门决定有关疾病的保险预算以及相关政策的制定提供了可靠的依据. 其次, 医院可以依据联邦重要性加权方法的预测结果提前对潜在患者进行医疗干预, 进而延缓或阻止疾病的到来, 从而提高患者的长期生活质量. 总的来说, 我们的实证研究清楚地验证了联邦重要性方法在预测疾病得分时的优异性能, 对于医院的治疗手段、 治疗流程以及疾病形式研判等方面具有重要的参考意义.

参考文献

1 Sagiroglu S, Sinanc D. Big data: A review. 2013 International Conference on Collaboration Technologies and Systems (CTS). San Diego: IEEE, 2013. 42–47.

2 Malin BA. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. Journal of the American Medical Informatics Association, 2005, 12(1): 28–34.

3 Madden M, Gilman M, Levy K, *et al*. Privacy, poverty, and big data: A matrix of vulnerabilities for poor Americans. Washington University Law Review, 2017, 95(1): 53–125.

4 Martin KD, Murphy PE. The role of data privacy in marketing. Journal of the Academy of Marketing Science, 2017, 45(2): 135–155. [doi: 10.1007/s11747-016-0495-4]

5 Shokri R, Shmatikov V. Privacy-preserving deep learning. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. Denver: Association for Computing Machinery, 2015. 1310–1321.

6 Jain P, Gyanchandani M, Khare N. Big data privacy: A technological perspective and review. Journal of Big Data, 2016, 3(1): 25. [doi: 10.1186/s40537-016-0059-y]

7 Yang Q, Liu Y, Cheng Y, *et al*. Federated learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019, 13(3): 1–207. [doi: 10.2200/S00960ED2V01Y201910AIM043]

8 Yang Q, Liu Y, Chen TJ, *et al*. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 12.

9 Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359. [doi: 10.1109/TKDE.2009.191]

10 Tsanas A, Little MA, McSharry PE, *et al*. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Transactions on Biomedical Engineering, 2010, 57(4): 884–893. [doi: 10.1109/TBME.2009.2036000]

11 Lee J, Sun JM, Wang F, *et al*. Privacy-preserving patient similarity learning in a federated environment: Development and analysis. JMIR Medical Informatics, 2018, 6(2): e20. [doi: 10.2196/medinform.7744]

12 Dowlin N, Gilad-Bachrach R, Laine K, *et al*. CryptoNets: Applying neural networks to encrypted data with high

throughput and accuracy. Proceedings of the 33rd International Conference on Machine Learning. New York: ACM, 2016. 201–210.
13 Bonawitz K, Ivanov V, Kreuter B, *et al*. Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017. 1175–1191.
14 Mohassel P, Zhang YP. SecureML: A system for scalable privacy-preserving machine learning. 2017 IEEE Symposium on Security and Privacy (SP). San Jose: IEEE, 2017. 19–38.
15 Smith V, Chiang CK, Sanjabi M, *et al*. Federated multi-task learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 4427–4437.
16 Liu Y, Kang Y, Xing CP, *et al*. A secure federated transfer learning framework. IEEE Intelligent Systems, 2020, 35(4): 70–82. [doi: 10.1109/MIS.2020.2988525]
17 Cheng KW, Fan T, Jin YL, *et al*. SecureBoost: A lossless federated learning framework. IEEE Intelligent Systems, 2021, 36(6): 87–98. [doi: 10.1109/MIS.2021.3082561]
18 Yao Y, Doretto G. Boosting for transfer learning with multiple sources. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 1855–1862.
19 Raina R, Battle A, Lee H, *et al*. Self-taught learning: Transfer learning from unlabeled data. Proceedings of the 24th International Conference on Machine Learning. Corvalis: Association for Computing Machinery, 2007. 759–766.
20 Lawrence ND, Platt JC. Learning to learn with the informative vector machine. Proceedings of the 21st International Conference on Machine Learning. Banff: Association for Computing Machinery, 2004. 65.
21 Mihalkova L, Huynh T, Mooney RJ. Mapping and revising markov logic networks for transfer learning. Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver: AAAI Press, 2007. 608–614.
22 Zadrozny B. Learning and evaluating classifiers under sample selection bias. Proceedings of the 21st International Conference on Machine Learning. Banff: Association for Computing Machinery, 2004. 114.
23 Lipton Z, Wang YX, Smola A. Detecting and correcting for label shift with black box predictors. Proceedings of the 35th International Conference on Machine Learning. Stockholm: Proceedings of Machine Learning Research, 2018. 3122–3130.
24 Sugiyama M, Krauledat M, Müller KR. Covariate shift adaptation by importance weighted cross validation. The Journal of Machine Learning Research, 2007, 8: 985–1005.
25 Huang JY, Smola AJ, Gretton A, *et al*. Correcting sample selection bias by unlabeled data. Proceedings of the 19th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2006. 601–608.
26 Sugiyama M, Suzuki T, Nakajima S, *et al*. Direct importance estimation for covariate shift adaptation. Annals of the Institute of Statistical Mathematics, 2008, 60(4): 699–746. [doi: 10.1007/s10463-008-0197-x]
27 Kanamori T, Hido S, Sugiyama M. A least-squares approach to direct importance estimation. The Journal of Machine Learning Research, 2009, 10: 1391–1445.
28 Nguyen XL, Wainwright MJ, Jordan MI. Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 2010, 56(11): 5847–5861. [doi: 10.1109/TIT.2010.2068870]
29 Kouw WM, Loog M. A review of domain adaptation without target labels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(3): 766–785. [doi: 10.1109/TPAMI.2019.2945942]
30 Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference, 2000, 90(2): 227–244. [doi: 10.1016/S0378-3758(00)00115-4]
31 Bergstra J, Bardenet R, Bengio Y, *et al*. Algorithms for hyper-parameter optimization. Proceedings of the 24th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2011. 2546–2554.
32 Nomura M, Saito Y. Efficient hyperparameter optimization under multi-source covariate shift. Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: Association for Computing Machinery, 2021. 1376–1385.
33 Wistuba M, Schilling N, Schmidt-Thieme L. Learning hyperparameter optimization initializations. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris: IEEE, 2015. 1–10.

(校对责编: 牛欣悦)