

引用:金鑫,胡滢,李芃,等. 隐私保护下多方高维数据联邦特征选择算法[J]. 聊城大学学报(自然科学版), 2025, 38(4): 497-506.

Citation: JIN Xin, HU Ying, LI Peng, ZHENG Ming. Federated feature selection algorithm for multi-party high-dimensional data under privacy protection[J]. Journal of Liaocheng University (Natural Science Edition), 2025, 38(4): 497-506.

文章编号 1672-6634(2025)04-0497-10

DOI 10.19728/j.issn1672-6634.2024070004

隐私保护下多方高维数据联邦特征选择算法

金鑫^{1,2}, 胡滢², 李芃¹, 郑明²

(1. 铜陵学院 电气工程学院, 安徽 铜陵, 244061; 2. 安徽师范大学 计算机与信息学院, 安徽 芜湖, 241002)

摘要 高维特征选择问题存在“维数灾难”和计算代价高等挑战。受隐私保护的限制,大量高维数据可能分布存储在不同机构(称为参与方)而不能共享,进一步增加了多方高维数据联合特征选择的难度。鉴于此,本文提出一种代理联合辅助的联邦进化特征选择算法,解决隐私保护下多方参与的高维特征选择问题。设计一种代理辅助的联邦进化特征选择算法框架,并基于该框架,给出代理模型联合构建与管理策略、基于代理模型的联合评价策略和个体联合更新策略。最后,将所提算法用于10组测试数据集,并与3种典型的封装式进化特征选择算法进行对比。结果表明,所提算法在充分保护参与方数据隐私的前提下,不仅能够保证算法的分类性能,还能显著提升算法的运行时间。

关键词 特征选择; 进化算法; 代理辅助; 隐私保护

中图分类号 TP309; TP18

文献标志码 A

开放科学(资源服务)标识码(OSID)



Federated feature selection algorithm for multi-party high-dimensional data under privacy protection

JIN Xin^{1,2}, HU Ying², LI Peng¹, ZHENG Ming²

(1. School of Electrical Engineering, Tongling University, Tongling 244061, China; 2. School of Computer and Information, Anhui Normal University, Wuhu 241002, China)

Abstract High-dimensional feature selection faces challenges such as the “curse of dimensionality” and high computational costs. Due to privacy protection constraints, a large amount of high-dimensional data may be distributed and stored across different institutions (referred to as participants) and cannot be shared, further complicating the joint feature selection of multi-party high-dimensional data. In light of this, this paper proposes a surrogate-joint-assisted federated evolutionary feature selection algorithm to address the issue of high-dimensional feature selection involving multiple participants under privacy protection. A framework for a surrogate-assisted federated evolutionary feature selection algorithm is designed, and based on this framework, strategies for joint construction and management of surrogate models, joint evaluation based on surrogate models, and joint updating of individuals are provided. Finally, the proposed algorithm is applied to 10 test datasets and compared with 3 typical wrapper-based evolutionary feature se-

收稿日期: 2024-07-13

基金项目: 国家自然科学基金项目(62306009); 安徽省自然科学基金项目(2308085QF209)资助

通信作者: 胡滢, 女, 汉族, 博士, 讲师, 研究方向: 进化特征选择、联邦进化特征选择, E-mail: hy200712008@126.com.

lection algorithms. The results show that the proposed algorithm not only ensures the classification performance of the algorithm while fully protecting the data privacy of the participants but also significantly improves the algorithm's runtime.

Keywords feature selection; evolutionary algorithm; surrogate-assisted; privacy protection

0 引言

随着大数据时代的到来,人们所获数据的种类、规模及维度正以前所未有的速度增长。为防止重要信息的丢失,决策者所获得的原始数据中常包含许多冗余或者与当前学习任务无关的特征。这些特征不仅会降低学习的精度,还会增加学习算法的时间及空间复杂度。在利用这些高维数据之前,需要对其进行特征降维。特征选择是一类典型的数据降维技术,其目的是从数据集所有特征中选出部分特征构成一个有效特征子集,在减少学习代价的同时使设定的性能指标(如分类精度等)达到最优^[1]。

特征选择过程主要包括子集搜索和子集评价两个关键步骤。子集搜索是指采用某种搜索策略,从原始特征集中搜索最优特征子集的过程,主要包括:全局搜索方法、启发式搜索方法和随机搜索方法^[2],其中,基于进化优化的随机搜索方法已成为解决特征选择问题的热点研究方向。子集评价是指计算所选特征子集的性能指标(如分类精度等)以判定其好坏。按照评价过程与分类器是否相结合,特征选择算法可以分为过滤式特征选择、封装式特征选择和嵌入式特征选择^[3]。研究表明,上述算法已能很好地解决面向数据集中心式存储或面向数据共享环境下分布式存储的常规特征选择问题^[4-5],但应用于隐私保护场景下的联邦特征选择问题时,还需针对问题的特性对上述算法加以改进。

面向高维数据特征选择,学者们结合进化优化技术,提出了很多行之有效的办法,如基于特征空间分组的协同进化特征选择方法^[6]、基于特征空间知识转移的进化多任务特征选择方法^[7],以及基于模糊聚类技术的进化特征选择算法^[8]等。然而,由于分类器训练代价与特征维度呈正相关,这些方法依然存在个体评价耗时较长的问题。代理辅助进化算法(Surrogate-Assisted Evolutionary Algorithm, SAEA)采用代理模型代替耗时的真实目标函数,能够低成本地评价个体,已被广泛应用于求解各类昂贵优化问题^[9]。近年来,一些学者开始尝试将 SAEA 用于特征选择问题^[10],由此,代理辅助进化特征选择方法(Surrogate-Assisted Evolutionary Feature Selection, SAEFS)便应运而生。简而言之,该类方法使用特征子集及其对应的目标函数值来构建目标回归代理模型,并用其代替高计算代价的分类器来预测特征子集的质量,可以解决面向特征空间维度高而带来的高计算代价问题,有效提升进化特征选择方法的收敛速度^[11-12]。

现实中描述同一个学习任务的数据可能横向(水平)分布存储在不同机构(称为参与方),且各机构所持有数据之间又具有很强的互补性。如不同地区医院收集的同一类典型病例数据、不同地区银行存储的同一类客户数据等。同时,受隐私保护的限制,这些机构往往不愿意或者不能直接分享它们持有的数据。单独对某一个参与方的数据进行特征学习,势必会影响其特征选择结果的分类精度和鲁棒性,带来特征选择偏好问题^[13]。因此,研究面向隐私保护需求下多方参与的特征选择问题具有重要的实际意义^[14]。传统方法采用隐私保护技术进行特征选择,如差分隐私特征选择^[15-17]、基于多方安全计算的隐私保护特征选择^[18-19]等。然而,上述方法在加入噪声或私有特征选择协议的前提下,原始数据仍可能离开本地参与计算,故而仍具有一定的隐私泄露风险,且对于高维特征空间,计算代价较大。近年来,借鉴隐私保护的联邦学习框架^[20],学者们陆续提出了一系列适用于不同问题场景的联邦特征选择算法^[21-22]。

受此启发,本文结合联邦学习思想,融合代理模型高效的评价性能和进化算法优异的全局搜索性能,研究隐私保护下多方参与的高维数据联邦进化特征选择算法,具体贡献如下:(1) 引入代理模型代替耗时的真实评价过程,给出一种面向隐私保护需求的代理联合辅助联邦进化特征选择框架,并基于该框架,设计代理模型的联合管理和更新策略,以提升模型的预测精度;提出基于代理模型的联合评价策略,以提升算法的收敛速度。(2) 给出一种基于头脑风暴优化的联邦进化特征选择方法。设计个体的联合更新策略,即在联邦框架下,使用代理模型联合更新种群中部分较差个体,引导种群朝全局最优方向搜索,提高种群的多样性。

论文组织如下:第1节介绍相关工作,包括头脑风暴优化算法、进化特征选择、代理辅助进化特征选择算

法、隐私保护下的特征选择方法等;第2节提出代理联合辅助的联邦进化特征选择算法;为了验证所提方法的有效性,第3节给出实验对比及结果分析;最后,第4节对全文进行总结并给出研究展望。

1 相关工作

1.1 头脑风暴优化算法

头脑风暴优化算法(Brain Storm Optimization Algorithm, BSO)是2011年由史等人^[23]提出的一种新颖的群体智能优化算法,灵感来源于人类头脑风暴过程。该算法通过模拟人类在头脑风暴过程中产生和交流创意的行为,可以很好地均衡种群的全局探索和局部开发能力,自提出以来一直备受学术界的关注。针对传统BSO的缺陷,学者们已经提出了诸多BSO的改进版本,并将其扩展到多目标优化^[24]、多模态优化^[25]、特征选择^[26]等问题。

BSO算法主要包含3个步骤:个体聚类、新个体生成和个体选择。在个体聚类阶段,当前种群中的所有 n 个个体被分为 m 类,每一类中的最优个体作为类中心;在新个体生成阶段,基于一个或两个类的类中心或个体,通过交叉变异操作生成一个新个体;在个体选择阶段,每个新个体和对应的父代个体进行对比,胜者被保留到下一代。重复以上步骤,直到满足算法终止条件,如最大运行代数。

当采用一个类产生新个体时,选择类中心或者类中的一个个体作为实验个体 X_{try} ,在此基础上加入扰动生成新个体

$$\begin{cases} X'_{\text{new}} = X_{\text{try}} + \epsilon \times r, \\ \epsilon = \text{logsig}(\frac{0.5 \times T - t}{c}), \end{cases} \quad (1)$$

式中 t 为当前迭代次数, T 为算法终止迭代次数, X'_{new} 为基于类中心或个体生成的候选新个体, r 是 $[0,1]$ 之间的随机数, ϵ 为扰动参数, c 为扰动的控制参数。

当采用两个类产生新个体时,首先选择两个类的中心或其任意个体,不妨记为 X_i 和 X_j ,并由式(2)得到实验个体 X_{try} ,然后由式(1)得到新个体

$$X_{\text{try}} = t_{\text{em}} \times X_i + (1 - t_{\text{em}}) \times X_j, \quad (2)$$

式中 $t_{\text{em}} \in [0,1]$ 为生成新个体的控制参数。

1.2 进化特征选择方法

由于不需要问题结构信息且具有较强的全局搜索能力,学者们面向高维数据集提出了大量进化特征选择方法。这些方法包括基于遗传算法(Genetic Algorithm, GA)^[27]、粒子群优化算法(Particle Swarm Optimization Algorithm, PSO)^[7-8]、蚁群优化算法(Ant Colony Optimization algorithms, ACO)^[28]、人工蜂群算法(Artificial Bee Colony Algorithm, ABC)^[29]以及差分进化算法(Differential Evolution Algorithm, DE)^[30]的特征选择方法。Zhou等^[27]针对随机进化的传统GA可能会多次评价劣质解的情况,提出了一种基于相关性引导交叉和变异算子的特征选择方法,提高解的质量,加快算法收敛。Chen等^[7]通过在高维数据的不同搜索空间间进行知识转移,提出了一种基于PSO的有效多任务特征选择方法,并通过执行翻转操作,显著提升了算法的局部搜索能力。Prajapati等^[28]针对具有大量噪声、无关及冗余特征的微阵列基因数据集,采用ACO算法进行特征选择,取得了较好的分类效果。Tong等^[29]将ABC算法与交叉遗传算子相结合,针对高光谱图像中存在波段冗余问题,提出了一种基于人工蜂群融合遗传思想的高光谱波段选择方法。此外,Wang等^[30]将特征选择问题看成多目标优化问题,提出了一种基于小生境的差分进化多目标特征选择方法。作为一种相对较新的进化优化技术,BSO具有相对简单、易于实现,且具有较好多样性保持等优点,因此,本文以BSO为搜索技术,下面重点阐述基于BSO的特征选择方法。

2018年,Papa等^[31]首次提出了一种二元BSO版本,并成功应用于特征选择问题。随后,Pourpanah等^[32]又提出了一种二元BSO的改进版本,应用于特征选择和数据分类的混合模型,取得了较好的分类效果。Zhang等^[33]结合两种个体更新机制,提出了一种基于BSO的新颖特征选择算法,加快了种群的收敛速度。为获得最优的分类模型,Xue等^[26]利用BSO同时搜索由最优特征子集确定的最优分类模型结构以及

不同结构下的最优权向量,实验结果表明,特征选择可以有效简化分类模型的结构。然而,类似于其他进化特征选择算法,基于BSO的特征选择算法同样需要多次调用分类器来评价特征子集或个体,对于维度较高的数据集,其计算代价依然很高。

1.3 代理辅助进化特征选择方法

SAEFS采用代理模型代替昂贵的真实目标函数评价候选解(特征子集),可以大大降低进化算法中特征子集的评价代价。常用的代理模型有径向基函数网络(RBF)^[34]、高斯过程模型(又称Kriging模型)^[35]、支持向量回归模型^[36]等。目前,基于目标函数代理的进化特征选择方法的研究成果相对较少,如Brownlee等^[37]使用马尔可夫网络模型进行适应度评估代理,提出了一种基于GA的特征选择算法。Hu等^[38]将高维特征选择问题划分为2个低维的子优化问题,对应的2个子种群分别采用径向基函数和Kriging代理模型进行个体评估,提出了一种多代理辅助的二元PSO优化方法,并成功用于解决特征选择问题。类似地,Liu等^[10]将高维特征选择问题划分为多个低维的特征选择子问题,并在低维子问题上使用径向基函数代理模型进行评估,提出了一种并行随机分组的代理辅助进化特征选择算法。

上述方法能够大大降低特征子集的评价代价,提升算法的运行效率,然而,它们大都面向数据集中存储的传统特征选择问题,无法有效处理数据分布存储在多个参与方的隐私保护特征选择问题。由于各方数据之间的互补性,对单独一个参与方的数据执行特征选择,势必会带来特征选择偏好问题,各方将很难得到全局最优的特征子集。

1.4 隐私保护特征选择方法

目前,关于隐私保护特征选择方法的研究较少,主要采用传统的隐私保护技术进行特征选择,如差分隐私特征选择^[39]、基于多方安全计算的隐私保护特征选择^[19]等。前者通过添加受控制的随机性噪声,在完成特征选择任务的同时,最大程度保证原始数据的隐私。然而,通过添加噪声进行差分隐私保护,难以准确地度量和处理特征之间的关联关系,并且隐私预算选取不当往往会影响特征选择的效果和隐私保护程度;后者采用私有特征选择协议完成“数据去中心化”架构下的多方联合特征选择,然而,多方安全计算的计算复杂度较高,且对于高维特征空间,会进一步增加其计算代价,带来更大的通信开销,此外,若某一个参与方上存在数据污染情况,无法保证特征选择结果的可靠性。进一步,对于传统的隐私保护技术,原始数据均可出本地,故而仍具有一定的隐私泄露风险。

近年来,联邦学习作为一种分布式学习多方联合建模方法,能够在数据不出域的情况下完成多方联合建模,逐渐成为解决多方参与下隐私保护问题的主流计算框架^[40-41]。Qin等^[42]结合联邦学习思想和传统特征选择技术,提出了一种保证隐私安全的网络入侵检测方法,然而,由于每个边缘设备使用贪婪机制搜索最优特征子集,该方法依然存在局部收敛的不足,一旦数据集的维度较高,将很难得到一个分类性能较好的特征子集。Hu等^[14]结合进化优化技术良好的全局搜索性能,提出了一种基于PSO的联邦特征选择算法。通过设计隐私保护下的集成策略、集成结果引导的粒子群初始化策略等适合联邦学习的新算子,显著提高了PSO处理联邦特征选择问题的能力。面向参与方存在不平衡数据的情形,Hu等^[43]进一步提出了一种面向不平衡数据的联邦进化特征选择方法。通过加入满足隐私保护需求的多层联合样本填充策略,以及不同参与方上最优特征子集的选择和引导策略,提升了参与方处理不平衡数据的能力,保证能够产生高质量的特征子集。

然而,当面向隐私保护需求下多方参与的高维特征选择问题时,已有联邦特征选择方法无法高效对特征子集进行评价,即使在每个参与方上单独采用代理模型辅助评价特征子集,这些构建的局部代理模型也无法兼顾所有参与方上的数据信息,难以使系统收敛到全局最优的特征子集。

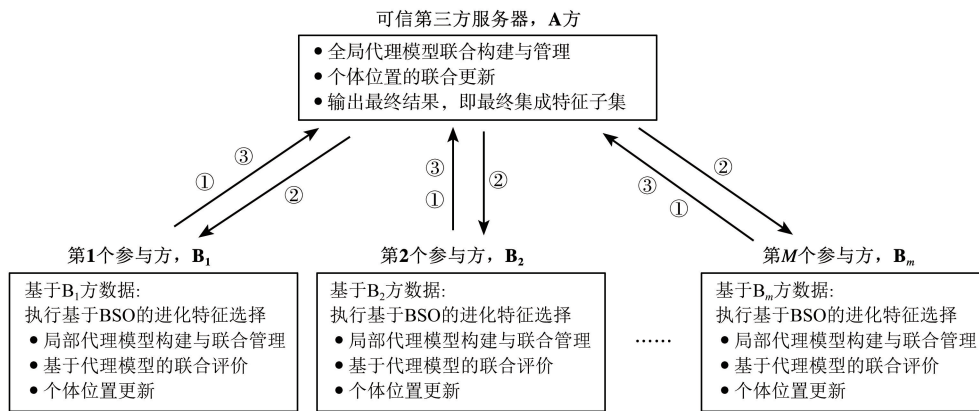
2 隐私保护下多方高维数据联邦特征选择算法

受联邦学习思想的启发,通过引入可信第三方来处理 and 集成各参与方的数据,本文提出一种隐私保护下面向多方高维数据的联邦进化特征选择算法(Federal Evolutionary Feature Selection Algorithm for High-dimensional Dataset, FEFS-HD),以解决有隐私保护需求的多方高维数据特征选择问题。该算法主要包括

代理模型联合构建与管理策略、基于代理模型的联合评价策略,以及个体联合更新策略。首先给出算法的整体框架。

2.1 所提算法框架

将联邦学习思想融合到多方参与的特征选择过程中,本节给出隐私保护下面向多方高维数据的联邦进化特征选择算法框架,如图1所示。假设当前系统包含 m 个参与方(称为 B 方)和一个可信第三方(称为 A 方),结合代理模型的低计算代价,提出一种高效的联邦进化特征选择算法,在约减后的特征空间中搜索系统综合最优的特征子集,同时保证各参与方上的数据隐私安全。该阶段主要包括:代理模型联合构建与管理策略、基于代理模型的联合评价策略,以及个体的联合更新策略。需要注意的是,为确保数据的安全性,B方之间不直接交换任何数据,而是通过 A 方间接交换各参与方的非敏感数据。不失一般性,本文采用基于 BSO 的进化特征选择算法。



注:① 本地代理模型参数,以及各参与方上 SP 采样点的均方误差值; ② 全局代理模型参数; ③ 个体最优位置。

图1 隐私保护下面向多方高维数据的联邦进化特征选择算法框架

Fig. 1 Federated evolutionary feature selection algorithm framework for multi-party high-dimensional data under privacy protection

2.2 代理模型联合构建与管理

本节提出隐私保护下代理模型的联合构建与管理策略,旨在保证数据隐私的前提下,联合多方数据,共同构建与管理各参与方的局部代理模型和可信第三方的全局代理模型。由于径向基函数网络(Radial Basis Function Neural Network, RBFNN)具有结构简单、学习速度快,以及优良的逼近和泛化性能^[34],本文选用RBFNN作为代理模型。

2.2.1 局部代理模型的构建与联合管理。(1) 局部代理模型初始构建。局部代理模型的初始构建过程如下:针对任一参与方,首先,借助拉丁超立方采样产生 N 个初始样本点(初始个体),组成进化算法的初始种群,记为 pop;然后,利用全体训练样本集和分类器评价每个个体的真实分类精度值,这些个体(样本点)及其真实分类精度值将作为采样点,保存到代理样本池 SP 中;最后,使用 SP 中的采样点训练 RBFNN 模型,进而完成参与方上局部代理模型的初始构建。(2) 局部代理模型联合管理。局部代理模型的管理主要体现在两个方面:① 未考虑全局信息的局部代理模型更新。② 考虑了全局信息的局部代理模型自适应更新。第①部分更新过程如下:针对任一参与方,首先,对当前代种群中的所有个体进行局部代理评估,并对评估值由大到小进行排序;随后,选择排序靠前 10% 的个体进一步进行真实评价;接着,将排序靠前 10% 的个体及其真实评价值作为新的采样点,加入到 SP 中,以更新当前代的局部代理模型。需要注意的是,为防止加入较多采样点影响代理模型更新的效率,同时降低质量较差的采样点对模型更新的影响,需要对 SP 的规模进行设置。这里, $|SP|_{\max} = 2 \times \text{popsize}$ 。其中, popsize 为种群规模。当 SP 中采样点的规模超过其上限值时,删除真实评价值最差的采样点,以满足 SP 的规模。

由于第①部分各参与方仅考虑其自身所持有数据的特性,并未考虑其他参与方数据对当前参与方的影响,因此,第②部分利用所有参与方数据信息的全局代理模型进行局部代理模型的自适应更新。具体方法如下:每隔 τ_1 代(τ_1 为代理模型联合更新周期),通过比较全局代理模型与当前局部代理模型在部分个体上的

评价误差,然后由参与方自适应选择是否将全局代理模型替换为当前的局部代理模型。

2.2.2 全局代理模型的联合构建与管理。(1) 全局代理模型联合构建。联合不同参与方上的局部代理模型,在 A 方构建系统的全局代理模型。具体过程如下:针对任一参与方,首先将其初始构建的局部代理模型参数上传给 A 方;随后,借鉴联邦学习思想,对来自不同参与方的模型参数进行加权聚合,聚合后的结果即为 A 方构建的初始全局代理模型。由于无法判定不同局部代理模型对全局代理模型的影响,因此本文采用平均聚合方式构建全局代理模型。(2) 全局代理模型联合管理。考虑到通信代价,全局代理模型每隔 τ_1 代进行一次更新。对于任一参与方,当达到 τ_1 时,向 A 方传输当前代局部代理模型参数,以及部分个体的局部代理模型评估值与真实分类器评价值的均方误差。一方面,传输局部代理模型参数是为构建当前代的全局代理模型。另一方面,传输均方误差的目的是判断当前局部代理模型是否需要替换为全局代理模型。均方误差由 SP 中采样点计算。记 SP 中当前代采样点集合为 Z , $Z = \{z_1, z_2, \dots, z_{|SP|}\}$, 其中, $z_i = [x_i, f(x_i)]$, $i=1, 2, \dots, |SP|$, x_i 表示第 i 个采样点对应的个体, $f(x_i)$ 表示 x_i 的真实评价值。针对第 j 个参与方 B_j , 其均方误差计算公式如下:

$$M_{SE}(j) = \frac{1}{|SP|} \sum_{i=1}^{|SP|} |f_j(x_i) - \hat{f}_j(x_i)|^2, j=1, 2, \dots, m, \quad (3)$$

式中 $\hat{f}_j(x_i)$ 表示参与方 B_j 方上个体 x_i 用局部代理模型评估的预测适应度值。

A 方接收到所有参与方的均方误差值,随后计算均方误差均值,记为 $M_{MSE} = \sum_{k=1}^m M_{SE}(k)/m$ 。判断 $M_{SE}(j)$ 与 M_{MSE} 的大小。当 $M_{SE}(j) > M_{MSE}$, 表示 B_j 方的局部代理模型误差相对较大,需要对其进行更新,即使用当前全局代理模型进行替换;反之,表示 B_j 方的局部代理模型与真实分类器拟合程度较高,不需要对其进行更新或替换。

2.3 基于代理模型的联合评价策略

本节采用代理模型联合辅助的评价方法,对参与方上个体的分类精度值进行评估,提出一种融合多方代理模型的联合评价策略。

针对当前待评价的子代种群,首先采用当前局部代理模型预测种群中所有个体的分类精度值,并按照分类精度值由大到小对个体进行排序;接着选择排序靠前 10% 的个体,由真实分类器评价它们的分类精度值。对于每个个体来说,最新的分类精度值就是其在本次迭代中的最终分类精度值。尽管分类器的选择会影响最终所得特征子集的性能,考虑到本文主要目的是设计一种高效的联邦进化特征选择算法,因此本文选择常用的 KNN 作为真实分类器,用来评价每个个体的分类精度值。由于评价当前个体的局部代理模型可能为联合所有参与方构建的全局代理模型,因此,该过程体现了局部代理模型和全局代理模型的“联合”评价过程。

2.4 个体联合更新策略

每个参与方均采用文献^[44]提出的二进制 BSO 算法来产生新的个体,即特征子集。然而,由于需要兼顾所有参与方的数据信息,共同搜索系统综合最优的特征子集,因此需要周期性地对不同参与方上的最优特征子集进行联合更新。记个体的联合更新周期为 τ_2 , 其更新策略如下:对于任一参与方,首先当到达联合更新周期 τ_2 时,在前 10% 真实评价的较优个体中,选择真实评价最高值所对应的最优个体,并传输给 A 方。接着以 A 方为中转,各参与方从 A 方接收其他 $m-1$ 个参与方的最优特征子集,来替换当前种群中较差的 $m-1$ 个个体。

最后,当所有参与方运行的 BSO 算法均达到最大迭代次数时,终止特征选择算法,并将得到的最优特征子集传输给 A 方,由 A 方对所有特征子集进行集成,输出系统综合最优的特征子集。特征子集的集成过程参考作者前期提出的平均集成策略^[14]。

2.5 所提算法的计算代价

不妨设参与方数目为 m , 种群规模为 popsize , BSO 算法的聚类数目为 K_0 , 迭代次数为 T 。对于每个参与方上的计算代价,首先任意一个参与方执行 BSO 算法的计算代价为 $O(\text{popsize} \times K_0)$ 。其次,代理模型的

计算复杂度主要来自于 RBFNN 模型的训练复杂度,为 $O(|SP| \times K^2)$,其中 $|SP|$ 为代理样本池的规模, K 为隐藏层神经元数目。最后,粒子的评价过程包括粒子的真实评价和基于代理模型的联合评价,该过程的计算复杂度主要由粒子的真实评价代价决定。在算法每次迭代中,需要使用整个数据集真实评估 $0.1 \times \text{pop-size}$ 个粒子的真实适应度值。对于 A 方的计算代价,每隔代理模型联合更新周期,需要进行全局代理模型的构建与管理,其计算复杂度为 $O(m)$,特征子集集成的计算代价亦为 $O(m)$ 。综上,类似其它基于进化优化的特征选择算法一样,在处理高维数据集时,其真实适应度评价所需的计算代价远远大于其他所有算子。

3 实验

为验证本文所提 FEFS-HD 算法的性能,本节将其用于 10 组常用的 UCI 数据集。实验分为如下三个部分:第 1 部分介绍所用对比算法及参数设置;第 2 部分分析 SaFEFS 中关键参数对算法性能的影响;第 3 部分给出 FEFS-HD 与所有对比算法的实验结果并对其进行分析。每个算法在每个数据集上分别独立运行 30 次,统计它们的平均优化结果。表 1 给出了所用数据集的基本信息。

表 1 数据集
Tab. 1 Datasets

No.	Datasets	Features	Samples	Classes
1	Musk1	166	476	2
2	GFE	301	1 062	2
3	madelon	500	2 600	2
4	Isolet	617	1 560	26
5	MFD	649	1 000	10
6	CNAE-9	856	1 080	9
7	Coil20	1 024	1 440	20
8	PCMAC	3 289	1 943	2
9	RELATHE	4 322	1 427	2
10	BASEHOCK	4 862	1 993	2

3.1 对比算法及参数设置

为验证所提 FEFS-HD 算法的准确性和高效性,选择两类经典的基于代理模型的高维特征选择方法, MDEFS 算法^[45]和 CUS-SPSO 算法^[11],并分别将其嵌入到封装式集成特征选择算法框架^[14]中作为对比算法,分别记为 WAFS-1 和 WAFS-2。这样做的原因是,其一,目前没有面向高维数据集的联邦进化特征选择方法;其二,封装式集成框架利用可信第三方对各参与方所得结果进行集成,可以直接扩展到隐私保护特征选择问题。此外,选择已有联邦进化特征选择算法 FPSO-FS^[14]作为第三个对比算法。

鉴于本文目的是验证隐私保护场景下,面向高维数据集的联邦进化特征选择方法的有效性,不失一般性,将数据集分成三个参与方进行讨论。首先,随机均匀划分数据集,并将其随机分配给三个参与方。

为保证比较的公平性,所有基于种群的对比算法迭代次数均设为 100,种群规模均设为 50。在代理模型联合管理中,代理模型联合更新周期 τ_1 直接影响全局代理模型对局部代理模型的更新效果以及后续代理评价的损失差异。若 τ_1 较大,会导致特征选择模型更新不及时,缺乏多样性;若 τ_1 较小,会导致代理模型训练不充分,进而使模型评价不准确。综合上述原因, τ_1 取值设为 20。在个体联合更新过程中,其联合更新周期 τ_2 决定了参与方之间的通信次数。若 τ_2 较大,可能导致每个参与方不能及时学习到其他参与方提供的最优信息,使联合构建的特征选择模型更新不及时;若 τ_2 较小,可能使参与方之间交互一些局部最优特征子集。综合上述原因, τ_2 取值设为 30。算法的参数设置如表 2 所示。在所有封装特征选择算法或策略中,使用 10 折交叉验证计算每个候选解的分类精度。采用简单高效的 KNN 分类器进行真实评价,参数 k 设置为 3^[14]。所有结果采用分类准确率(Classification accuracy, acc)作为评价指标。

表 2 参数设置

Tab. 2 Parameter setting

Algorithms	Parameter setting
WAFS-1	$c1=rand, c2=1-c1, Th=0.7$
WAFS-2	$c1=c2=1.49445, \omega=0.9-0.5 \times (iter/T_{max}), n_c=2, A=0.15, B=0.05$
FPSO-FS	$\alpha=0.1, T_r=4$
FEFS-HD	$c1=rand, c2=1-c1, \tau_1=20, \tau_2=30$

3.2 实验结果对比及分析

本节将对对比算法 WAFS-1、WAFS-2 和 FPSO-FS 与所提 FEFS-HD 算法进行对比,从分类性能和运行时间两方面验证 FEFS-HD 的性能。表 3 给出 FEFS-HD 与对比算法所得 acc 指标值和运行时间。其中,‘+’表示 FEFS-HD 显著优于所选对比算法,‘-’表示 FEFS-HD 显著劣于所选对比算法,‘ $\approx$ ’表示 FEFS-HD 和所选对比算法无显著性差异。

从表 3 可以看出,(1)从分类性能上看,FEFS-HD 算法在 7 个数据集上取得了最优的 acc 指标值。这可能是因为,WAFS-1 中的 MDEFS 算法在集成过程中采用并集、交集和投票法选择最优个体,只考虑了候选解的统计信息,并未考虑特征之间的冗余性。此外,对于具有离散变量的特征选择问题,WAFS-2 中的 CUS-SPSO 算法采用欧式距离来预测每个个体的适应度值,这会显著影响预测值的准确性。此外,FPSO-FS 通过直接交互每个参与方上的最优特征子集及其真实分类精度值引导各参与方上种群的进化,一定程度上减少了种群的多样性。(2)从运行时间上看,除 Musk1、GFE 和 Madelon 外,FEFS-HD 算法均取得了最短的运行时间。这可能是因为,当特征维度较低时,FEFS-HD 中代理联合更新的时间相对较长,因此总体耗时较大。当特征维度较高时,FEFS-HD 大部分评价过程均采用代理模型进行评价,与 WAFS-1、WAFS-2 和 FPSO-FS 中所有粒子都进行真实评价相比,显著降低了算法的运行时间。显然,面向高维数据集,本文算法在运行时间上取得了很大的提升。

表 3 FEFS-HD 与对比算法所得 acc 指标值和运行时间

Tab. 3 acc values and running time of FEFS-HD and comparative algorithms

	FEFS-HD		WAFS-1		WAFS-2		FPSO-FS	
	acc/%	T/s	acc/%	T/s	acc/%	T/s	acc/%	T/s
Musk1	85.302 6	123.11	76.912 6(+)	18.93	77.737 5(+)	26.39	83.427 6(+)	70.69
GFE	94.082 7	128.78	85.837 6(+)	79.63	85.711 1(+)	139.01	92.531 7(+)	96.41
Madelon	65.270 0	236.35	53.956 1(+)	683.80	56.015 6(+)	1 038.75	65.498 4($\approx$)	207.25
Isolet	79.739 3	142.11	80.300 2($\approx$)	450.06	82.735 0(-)	533.73	82.544 0(-)	269.36
MFD	96.719 8	146.69	95.551 9($\approx$)	150.58	96.363 9($\approx$)	272.74	96.471 9($\approx$)	172.78
CNAE-9	75.821 0	159.04	67.476 9(+)	180.64	73.916 7(+)	335.04	74.853 4(+)	232.29
Coil20	91.561 5	226.18	88.706 8($\approx$)	554.76	90.166 7(+)	646.23	91.209 6(+)	379.83
PCMAC	67.720 6	531.41	65.467 7(+)	2648.55	65.877 2(+)	5 752.23	67.532 2($\approx$)	1 885.40
RELATHE	73.784 9	354.37	69.334 9(+)	1904.84	73.466 1($\approx$)	2 882.34	74.659 8(-)	1 404.39
BASEHOCK	77.268 4	653.26	68.077 9(+)	4085.97	72.848 3(+)	6 198.93	75.443 9(+)	2 343.57

4 结论

针对隐私保护下面向多方高维数据的联邦特征选择问题,本文将代理模型用于辅助模型评价,给出了基于 BSO 的代理辅助联邦进化特征选择算法。本文设计的代理模型联合构建与管理策略、代理模型联合评价策略等适合在高维特征空间进行联邦学习的新算子,显著提升了算法处理高维特征选择问题的能力。将所提算法应用于 10 组典型测试数据集,并与 3 种封装式集成特征选择算法进行对比,结果表明,所提算法不仅

能够从数据层面保证原始数据的隐私安全,而且在保证分类性能的同时,显著缩短了算法的运行时间。

然而,在诸多实际场景中,受人为或客观等因素的影响,参与方上采集的数据特征维度非常高且样本量非常小,例如医疗诊断中的基因数据,对这类高维小样本数据进行联合建模时,如何进行联邦进化特征选择是一个亟待解决的问题。同时,不同参与方可能采集的特征目录不完全相同,面向这类异构数据的联邦进化特征选择问题也将是今后需要重点研究的一个方向。进一步,我们考虑将联邦特征选择算法与某些实际应用场景相结合,例如医疗诊断、金融风险控制等,验证算法的实际应用价值。

参 考 文 献

- [1] GARCÍA-TORRES M, RUIZ R, DIVINA F. Evolutionary feature selection on high dimensional data using a search space reduction approach[J]. *Engineering Applications of Artificial Intelligence*, 2023, 117: 105556.
- [2] YANG J Q, YANG Q T, DU K J, et al. Bi-directional feature fixation-based particle swarm optimization for large-scale feature selection [J]. *IEEE Transactions on Big Data*, 2023, 9(3): 1004-1017.
- [3] NGUYEN B H, XUE B, ZHANG M J. A survey on swarm intelligence approaches to feature selection in data mining[J]. *Swarm and Evolutionary Computation*, 2020, 54: 100663.
- [4] MORAN-FERNANDEZ L, BOLON-CANEDO V, ALONSO-BETANZOS A. Centralized vs. distributed feature selection methods based on data complexity measures[J]. *Knowledge-Based Systems*, 2017, 117: 27-45.
- [5] LUO C, WANG S Z, LI T R, et al. RHDOFS: a distributed online algorithm towards scalable streaming feature selection[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2023, 34(6): 1830-1847.
- [6] RASHID A N M B, AHMED M, SIKOS L F, et al. Cooperative co-evolution for feature selection in big data with random feature grouping [J]. *Journal of Big Data*, 2020, 7(1): 107.
- [7] CHEN K, XUE B, ZHANG M J, et al. An evolutionary multitasking-based feature selection method for high-dimensional classification [J]. *IEEE Transactions on Cybernetics*, 2022, 52(7): 7172-7186.
- [8] ZHANG Y, WANG Y H, GONG D W, et al. Clustering-guided particle swarm feature selection algorithm for high-dimensional imbalanced data with missing values[J]. *IEEE Transactions on Evolutionary Computation*, 2022, 26(4): 616-630.
- [9] JI J Y, TAN Z S, ZENG S Y, et al. A surrogate-assisted evolutionary algorithm for seeking multiple solutions of expensive multimodal optimization problems[J]. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024, 8(1): 377-388.
- [10] LIU S L, WANG H D, PENG W, et al. A surrogate-assisted evolutionary feature selection algorithm with parallel random grouping for high-dimensional classification[J]. *IEEE Transactions on Evolutionary Computation*, 2022, 26(5): 1087-1101.
- [11] CHEN K, XUE B, ZHANG M J, et al. Correlation-guided updating strategy for feature selection in classification with surrogate-assisted particle swarm optimization[J]. *IEEE Transactions on Evolutionary Computation*, 2022, 26(5): 1015-1029.
- [12] HU P, PAN J S, CHU S C, et al. Multi-surrogate assisted binary particle swarm optimization algorithm and its application for feature selection[J]. *Applied Soft Computing*, 2022, 121: 108736.
- [13] DASH M, LIU H. Feature selection for classification[J]. *Intelligent Data Analysis*, 1997, 1(1-4): 131-156.
- [14] HU Y, ZHANG Y, GAO X Z, et al. A federated feature selection algorithm based on particle swarm optimization under privacy protection [J]. *Knowledge-Based Systems*, 2023, 260: 110122.
- [15] ALISHAHI M, MOGHATAIE V, NAVIDAN H. Add noise to remove noise: local differential privacy for feature selection[J]. *Computers & Security*, 2022, 123: 102934.
- [16] 靳珂, 荣存庆, 常锦才. 基于集成特征选择的个性化差分隐私 LightGBM 算法[J]. *华北理工大学学报(自然科学版)*, 2024, 46(2): 145-155.
- [17] 刘中锋. 基于局部学习的差分隐私集成特征选择算法[J]. *计算机技术与发展*, 2018, 28(10): 79-82.
- [18] ZHANG Q, XIN C S, WU H Y. Privacy-preserving deep learning based on multiparty secure computation: a survey[J]. *IEEE Internet of Things Journal*, 2021, 8(13): 10412-10429.
- [19] 武宁宁. 基于安全多方计算的隐私保护特征选择算法研究[D]. 成都: 电子科技大学, 2024.
- [20] LI Y N, YANG S S, REN X B, et al. Multi-stage asynchronous federated learning with adaptive differential privacy[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(2): 1243-1256.
- [21] QIN Y, KONDO M. Federated learning-based network intrusion detection with a feature selection approach[C]//The 3rd International Conference on Electrical, Communication and Computer Engineering (ICECCE), Kuala Lumpur, Malaysia, 2021: 1-6.
- [22] ZHANG X Z, MAVROMATIS A, VAFEAS A, et al. Federated feature selection for horizontal federated learning in IoT networks[J].

- IEEE Internet of Things Journal, 2023, 10(11): 10095-10112.
- [23] SHI Y H. Brain storm optimization algorithm[C]//The 2th International Conference in Swarm Intelligence (ICSI), Berlin, Heidelberg, 2011: 303-309.
- [24] MA X M, FU Y P, GAO K Z, et al. A multi-objective scheduling and routing problem for home health care services via brain storm optimization[J]. Complex System Modeling and Simulation, 2023, 3(1): 32-46.
- [25] DAI Z Y, FANG W, TANG K, et al. An optima-identified framework with brain storm optimization for multimodal optimization problems[J]. Swarm and Evolutionary Computation, 2021, 62: 100827.
- [26] XUE Y, ZHAO Y. Structure and weights search for classification with feature selection based on brain storm optimization algorithm[J]. Applied Intelligence, 2022, 52(5): 5857-5866.
- [27] ZHOU J, HUA Z S. A correlation guided genetic algorithm and its application to feature selection[J]. Applied Soft Computing, 2022, 123: 108964.
- [28] PRAJAPATI S, DAS H, GOURISARIA M K. Feature selection using ant colony optimization for microarray data classification[C]//The 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023: 1-6.
- [29] TONG X Y, ZHOU X C. Hyperspectral band selection algorithm based on artificial bee colony fusion genetic idea[C]//The 3rd International Symposium on Computer Technology and Information Science (ISCTIS), Chengdu, China, 2023: 323-329.
- [30] WANG P, XUE B, LIANG J, et al. Differential evolution-based feature selection: a niching-based multiobjective approach[J]. IEEE Transactions on Evolutionary Computation, 2023, 27(2): 296-310.
- [31] PAPA J P, ROSA G H, DESOUSA A N, et al. Feature selection through binary brain storm optimization[J]. Computers & Electrical Engineering, 2018, 72: 468-481.
- [32] POURPANAH F, SHI Y H, LIM C P, et al. Feature selection based on brain storm optimization for data classification[J]. Applied Soft Computing, 2019, 80: 761-775.
- [33] ZHANG W Q, ZHANG Y, PENG C. Brain storm optimization for feature selection using new individual clustering and updating mechanism[J]. Applied intelligence, 2019, 49(12): 4294-4302.
- [34] LI G H, ZHANG Q F, LIN Q Z, et al. A three-level radial basis function method for expensive optimization[J]. IEEE Transactions on Cybernetics, 2022, 52(7): 5720-5731.
- [35] WEI F F, CHEN W N, MAO W T, et al. An efficient two-stage surrogate-assisted differential evolution for expensive inequality constrained optimization[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2023, 53(12): 7769-7782.
- [36] WANG Y G, WU J R, HU Z H, et al. A new algorithm for support vector regression with automatic selection of hyperparameters[J]. Pattern Recognition, 2023, 13: 108989.
- [37] BROWNLEE A E I, REGNIER-COUDERT O, MCCALL J A W, et al. An application of a GA with Markov network surrogate to feature selection[J]. International Journal of Systems Science, 2013, 44(11): 2039-2056.
- [38] HU P, PAN J S, CHU S C, et al. Multi-surrogate assisted binary particle swarm optimization algorithm and its application for feature selection[J]. Applied Soft Computing, 2022, 121: 108736.
- [39] ALISHAHI M, MOGHATAIE V, NAVIDAN H. Add noise to remove noise: local differential privacy for feature selection[J]. Computers & Security, 2022, 123: 102934.
- [40] LI Y N, YANG S S, REN X B, et al. Multi-stage asynchronous federated learning with adaptive differential privacy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(2): 1243-1256.
- [41] WU J H, ZHANG W Z, LUO F C. On the security of "LSFL: a lightweight and secure federated learning scheme for edge computing" [J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 3481-3482.
- [42] QIN Y, KONDO M. Federated learning-based network intrusion detection with a feature selection approach[C]//The 3rd International Conference on Electrical, Communication and Computer Engineering (ICECCE), Kuala Lumpur, Malaysia, 2021: 1-6.
- [43] HU Y, ZHANG Y, GONG D W, et al. Multiparticipant federated feature selection algorithm with particle swarm optimization for imbalanced data under privacy protection[J]. IEEE Transactions on Artificial Intelligence, 2023, 4(5): 1002-1016.
- [44] ZHANG W Q, ZHANG Y, PENG C. Brain storm optimization for feature selection using new individual clustering and updating mechanism[J]. Applied Intelligence, 2019, 49(12): 4294-4302.
- [45] JIANG Z, ZHANG Y, J. WANG J. A multi-surrogate-assisted dual-layer ensemble feature selection algorithm[J]. Applied Soft Computing, 2021, 110(10): 107625.

(责任编辑:赵海军)