基于隐私保护的粒子群优化联邦特征选择算法

摘要

特征选择是数据挖掘和机器学习领域中的一项重要预处理技术。随着隐私保护意识的提升，在确保所有参与方隐私的前提下选择高质量的特征子集，已成为一个极具实际意义且充满挑战的问题。然而，针对隐私保护下的特征选择问题，目前尚缺乏相关的研究成果。为了解决这一问题，本文首次提出了一个联邦特征选择框架。在该框架中，借鉴联邦学习的思想，引入了一个可信的第三方参与者，用于处理和整合来自多个参与方的最优特征子集。在此框架基础上，本文提出了一种基于粒子群优化的联邦进化特征选择算法，以有效解决多参与方隐私保护下的特征选择问题。该算法设计了两个满足隐私保护需求的新算子，即基于多参与方协作的特征集成策略和基于集成解的群体初始化策略，从而提升了算法的性能。通过对15个数据集的实验分析，与几种典型的集成特征选择算法相比，实验结果表明，本文提出的算法能够显著提高每个参与方选择的特征子集的分类准确率，同时实现数据隐私保护。

关键词：特征选择；联邦学习；进化算法

1 引言

在大数据时代，高维数据的普遍存在使得特征选择（FS）成为数据挖掘和机器学习中的重要预处理技术。通过从原始特征集中挑选关键特征，FS不仅能提高模型的分类性能，还能减少学习成本。然而，高维数据中存在大量冗余或无关特征，这会增加算法复杂性并降低学习准确性。此外，在许多实际应用中，数据分散存储于不同参与方，具有互补性，但因隐私保护需求，直接共享数据的方式难以实现。

现有研究中，针对集中存储数据的FS方法已取得显著进展，包括基于元启发式算法的进化特征选择（EFS）和群体智能优化方法（如粒子群优化，PSO）。这些方法具有全局搜索能力和较好的泛化性能。然而，对于隐私保护下的多参与方场景，现有方法存在局限性：一方面，传统FS算法假设数据可完全共享，不适用于数据分布式存储且敏感信息不可共享的情况；另一方面，现有隐私保护机制多针对集中存储的数据，无法有效处理多参与方间的数据隔离问题。

为弥补这一研究空白，这篇论文首次提出了一个联邦特征选择（Federated Feature Selection, FFS）框架。通过引入可信的第三方参与方，该框架整合了多个普通参与方的特征选择结果，并在隐私保护的前提下构建全局最优的特征子集模型。在此基础上，这篇论文进一步提出了一种基于PSO的联邦进化特征选择算法（FPSO-FS）。该算法设计了两个关键机制：其一，多参与方协作的特征集成策略，通过第三方整合特征索引和分类准确性，避免敏感信息泄露；其二，基于集成解的群体初始化策略，帮助普通参与方找到更优特征子集。

并且最终实验结果表明，与传统方法相比，本文算法显著提高了各参与方特征选择结果的分类性能，同时实现了数据隐私保护。此研究为解决分布式隐私保护下的特征选择问题提供了新思路。

2 相关工作

2.1 传统特征选择问题

假设一个数据集包含 D 个特征和 L 个样本，原始特征集记为 F。特征选择问题可描述为：从 F 中选择 d 个特征（$d \leq D$），以最大化某个指定的性能指标 $H(\cdot)$。使用二进制字符串表示特征子集 X：

$$X = (x_1, x_2, \ldots, x_D), \quad x_j \in \{0, 1\}$$

其中，$x_j = 1$ 表示选择第 j 个特征，否则不选。特征选择问题的公式为：

$$\max\ H(X), \quad \text{s.t. } X = (x_1, x_2, \ldots, x_D), \quad x_j \in \{0, 1\}$$

传统特征选择方法可分为三类：过滤法、包裹法和嵌入法 [23]。

- **过滤法** 根据特征的重要性排序。由于该方法不涉及分类器，其分类准确性相对较低，但计算成本较小。
- **包裹法** 直接使用分类器性能作为特征子集评价标准，典型算法包括顺序前向选择（SFS）[22] 和顺序后向选择（SBS）[14]。此方法分类准确性较高，但需多次训练分类器，计算成本较高。
- **嵌入法** 将特征选择过程与分类器学习相结合，在设计分类器的同时实现特征选择。虽然嵌入法的时间复杂度低于包裹法，但其结果对分类器依赖性强，鲁棒性较差 [27]。

2.2 基于进化优化的特征选择

特征选择本质上是一个离散组合优化问题。Xue 等 [23] 对现有的进化特征选择（EFS）方法进行了详细总结。随后，Hancer 等 [11] 回顾了最新的进化算法在特征选择中的应用。除了粒子群优化（PSO）外，还有许多基于其他进化算法的特征选择方法，例如基于标准误差的人工蜂群算法 [10]、改进的灰狼优化算法 [8]、二值海洋捕食者算法 [1] 和基于生物地理优化算法的特征选择方法 [17]。

本文主要关注基于 PSO 的算法，因其为本文的核心算法。PSO 具有快速收敛、参数少且易于实现的优点 [15]。例如，Ghamisi 等 [9] 结合遗传算法（GA）和 PSO 提出了一种高效的混合特征选择方法；Chen 等 [6] 设计了一种基于 PSO 的混合包裹算法，引入了螺旋机制以挖掘已知的最优区域；Song 等 [21] 提出了协作共进化 PSO 算法，有效提升了 PSO 在高维数据中的表现。然而，这些方法主要针对集中存储数据。尽管它们可以单独应用于一般参与方，但缺乏有效的信息共享机制，难以获得全局最优特征子集。

2.3 分布式特征选择方法

分布式特征选择的核心思想是将高维或大规模数据集划分为多个低维或小规模子集，然后在这些子集上同步进行特征选择，以减少计算成本并同时提升学习性能 [3]。分布式特征选择方法分为两类：水平分割和垂直分割 [3]。

- **水平分割** 将大规模数据集划分为多个小规模子集并行处理 [7]。
- **垂直分割** 将高维特征空间划分为多个低维特征子空间，并在子空间上进行特征选择 [16]。

例如，Bolón-Canedo 等 [3] 提出了一种分布式过滤特征选择算法，使用传统过滤方法从特征子空间中选择关键特征；Cao 等 [4] 构建了一个多目标特征选择模型，同时考虑分类误差、特征数量和特征冗余。尽管这些方法高效可扩展，但允许参与方共享敏感信息，因此无法满足隐私保护的需求。

2.4 隐私保护下的特征选择

目前，针对隐私保护特征选择的研究较少，尤其是多参与方场景。为防止非法第三方窃取敏感数据，Lu 等 [13] 提出了一种分布式集成特征选择方法，参与方通过加密传输数据。Bhuyan 等 [2] 利用扰动模糊策略处理数据以维护隐私，提出了一种模糊模型选择性能良好的关键特征。Sheikhalishahi 等 [20] 评估了特征的隐私分数和效用分数，提出了一种隐私-效用特征选择方法。Jafer等 [12]通过共享特征分布信息，选择少量隐私泄露但分类准确性高的重要特征。然而，这些方法多基于过滤方法，缺乏全局搜索策略，分类性能相对较低。此外，Qin 等 [18] 结合联邦学习和特征选择，提出了一种基于边缘设备的联合训练方法。然而，其贪婪特征选择算法易陷入局部最优。

综上所述，现有隐私保护特征选择方法仍存在局限性，亟需设计适用于隐私保护场景的高效特征选择算法。

3 本文方法

3.1 本文方法概述

在大数据和隐私保护需求日益增强的背景下，传统特征选择算法面临挑战，因为这些方法通常要求数据集中存储并允许共享。然而，许多实际场景中数据分布于多个参与方，这些参与方由于隐私问题无法直接共享数据。因此本文提出了一种基于隐私保护的联合特征选择框架，基于联邦学习的思想，在数据不共享的前提下实现协作优化特征选择。伪代码可见Algorithm 1。

Algorithm 1 FPSO-FS算法的伪代码

Input: M个B参与者使用的数据集，$\mathrm{Dat}_i(i=1,2,\cdots,M)$

Output: 全局最优特征子集，X^*

while 外层循环的终止条件未满足 **do**
  ——M个B参与者并行执行以下步骤（以B_i为例）——
  设置相关参数并初始化粒子群；
  while 内层循环的终止条件未满足 **do**
    基于数据集Dat_i评估每个粒子的分类准确率；
    执行算法3更新粒子，即生成新的候选解；
  end while
  $X_i \leftarrow Gbest_i$ ▷ 将$Gbest_i$设为B参与者的私有最优特征子集X_i
  将X_i及其分类准确率传输给A参与者；
  等待并接收来自A的$M-1$个私有最优特征子集X_j，$j=1,2,\cdots,i-1,i+1,\cdots,M$；
  基于数据集Dat_i评估这$M-1$个私有最优特征子集，评估结果记作：

$$\mathrm{acc}_{ij}(X_j), j=1,2,\cdots,i-1,i+1,\cdots,M.$$

  将$M-1$个acc值传输给A；
  等待并接收来自A的全局最优特征子集X^*；
  使用算法2初始化粒子群，并回到步骤3；
  ——————————A参与者的步骤——————————
  将M个私有最优特征子集X_i，$i=1,2,\cdots,M$发送给B参与者，并等待其反馈结果；
  接收来自M个B参与者的所有$\mathrm{acc}_{ij}(X_j)$值；
  执行第4.1节提出的特征聚合策略，生成全局最优特征子集X^*；
  将X^*及其分类准确率发送给M个B参与者；
end while
输出X^*。

具体来说：每个参与方（称为B-参与方）独立运行粒子群优化（PSO）。参与方根据其本地数据集生成一组候选特征子集，并计算每个子集在本地数据上的分类准确性。算法以迭代的方式优化特征子集，最终输出当前最优的特征子集，称为“私有特征子集”，并计算其对应的分类准确性。

在每轮迭代结束后，每个 B-参与方将其生成的私有特征子集及分类准确性发送给一个可信的中央节点（称为 A-参与方）。在此过程中，仅共享特征索引和分类准确性，不涉及具体数据，从而确保隐私不被泄露。A-参与方接收来自所有 B-参与方的特征子集及其对应的分类准确性，并利用特定的整合策略生成全局最优特征子集 X^*。并且将全局特征子集 X^* 反馈给所有 B-参与方。每个 B-参与方利用 X^* 指导下一轮特征选择的初始化过程。算法框架如图 1所示：

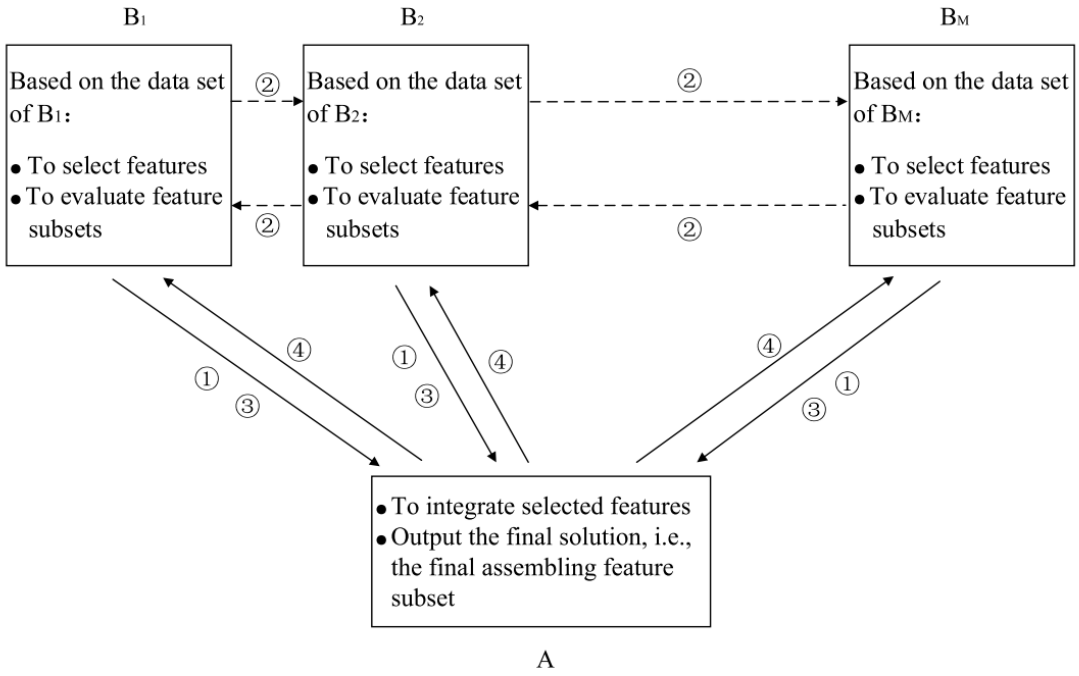


图 1. 方法示意图

3.2 多参与者合作的特征组合策略

特征整合策略的目的是在不共享敏感数据的前提下生成一个全局最优特征子集，使得所有参与方的分类准确性比原始分类结果有所提升。在每一轮FPSO-FS的执行中，A-参与方首先等待来自 M 个 B-参与方的反馈结果。基于 M 个私有特征子集及其对应的分类准确性，A-参与方采用特定的整合方法生成全局最优特征子集。

该论文提出了两种特征整合策略：平均整合策略和最大-最小整合策略，分别用于生成全局最优特征子集。这两种策略的具体内容如下：

- **平均整合策略** 对于第 i 个参与方的私有特征子集，首先计算其在所有参与方上的平均分类准确性，公式如下：

$$\mathrm{acc}_{ij}(X_i, \mathrm{Dat}_j) = \frac{1}{M}\sum_{j=1}^{M}\mathrm{acc}_{ij}(X_i, \mathrm{Dat}_j)$$

其中，X_i 表示第 i 个参与方的私有特征子集，Dat_j 表示第 j 个参与方持有的样本数据，$\mathrm{acc}_{ij}(X_i, \mathrm{Dat}_j)$ 表示 X_i 在 Dat_j 上的分类准确性。接着，选择平均分类准确性最高的特征子集作为全局最优特征子集 X^*，其公式如下：

$$X^* = \arg\max_{X_i}\frac{1}{M}\sum_{j=1}^{M}\mathrm{acc}_{ij}(X_i, \mathrm{Dat}_j)$$

- **最大-最小整合策略** 对于第 i 个参与方的私有特征子集，首先计算其在所有参与方中的最小分类准确性，公式如下：

$$\min_{j=1,\ldots,M}(\mathrm{acc}_{ij}(X_i, \mathrm{Dat}_j))$$

然后，选择最小分类准确性最大的特征子集作为全局最优特征子集 X^*，其公式如下：

$$X^* = \arg\max_{X_i}\min_{j=1,\ldots,M}(\mathrm{acc}_{ij}(X_i, \mathrm{Dat}_j))$$

平均整合策略能够获得在所有参与方上平均分类性能较好的特征子集；最大-最小整合策略能够保证全局最优特征子集在每个参与方上的分类性能不低于某一最低水平。整合策略的选择取决于决策者的实际需求：若希望整体性能最佳，可选用平均整合策略；若需确保所有参与方的最低性能，可以选用最大-最小整合策略。

3.3 群体初始化策略

种群初始化的质量直接影响进化算法的性能。为了提升每个B-参与方特征选择的效果，该论文利用全局最优特征子集 X^* 来协助各参与方生成高质量的初始种群。通过 X^* 的指导，所有参与方可以更快速地收敛到最优解。

在论文中，将每个特征被选中的概率作为粒子的位置编码元素，并用多个元素组成一个粒子。对于一个具有 D 个特征的数据集，第 i 个粒子的位置可以表示为：

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{iD}), \quad x_{ij} \in [0,1], \quad j = 1, 2, \ldots, D$$

其中，x_{ij} 表示第 i 个粒子中第 j 个特征被选中的概率。在评估一个粒子时，如果 $x_{ij} >$ rand（随机数，范围为 $[0,1]$），则第 j 个特征被选入对应的特征子集。

假设当前轮次的全局最优特征子集 X^* 由第 m 个参与方生成，即 $X^* = X_m$，具体的群体初始化策略如下：

- **对于生成 X^* 的参与方（即第 m 个参与方）** 如果某特征包含在 X^* 中，表明该特征在很大程度上是重要的。在初始化过程中，提升该特征被选中的概率；反之，降低未被选中特征的选中概率。具体公式为：

  $$x_{ij} = (1-\alpha) \cdot \text{rand} + \alpha \cdot x_j^*, \quad j = 1, 2, \ldots, D$$

  其中，x_j^* 表示 X^* 中第 j 个特征的位置值，$\alpha \in [0,1]$ 是控制 X^* 对粒子初始化影响程度的参数。

- **对于未生成 X^* 的其他参与方**

  – 如果某特征被 X^* 选中（$x_j^* = 1$），则提升该特征的选中概率：

  $$x_{ij} = \text{rand} \cdot [\alpha, 1]$$

  – 如果某特征未被 X^* 和当前参与方的特征子集 X_l 选中（$x_j^* = x_{lj} = 0$），则降低该特征的选中概率：

  $$x_{ij} = \text{rand} \cdot [0, 1-\alpha]$$

  – 如果某特征未被 X^* 选中，但被 X_l 选中（$x_j^* = 0, x_{lj} = 1$），则使用传统随机初始化策略：

  $$x_{ij} = \text{rand}$$

最终，每个参与方的种群部分基于上述策略初始化，其余粒子随机初始化。伪代码可见Algorithm 2所示。

Algorithm 2 提出的种群初始化策略

Input: 全局最优特征子集 X^*；生成 X^* 的 B-参与方的索引 m

Output: 含有 N 个粒子的初始种群 Pop_i

if $i == m$ **then** ▷ X^* 由 B_i 生成

    通过公式初始化 N_l 个粒子的位置；

    随机初始化剩余 $N - N_l$ 个粒子的位置；

else ▷ $X^* \neq X_i$

    通过公式初始化 N_l 个粒子的位置；

    随机初始化剩余 $N - N_l$ 个粒子的位置；

end if

输出 Pop_i。

3.4 粒子更新

本文采用文献[26]中提出的裸骨粒子群优化算法（Bare-bone PSO, BPSO）来更新粒子位置。该算法是一种几乎不需要参数的优化算法，省去了传统粒子群优化算法中的三个敏感参数，同时保留了粒子群优化算法的快速收敛和易于实现的优点。

具体而言，BPSO 使用基于个体最优位置（Pbest）和全局最优位置（Gbest）的高斯采样函数 G 生成新粒子。对于第 i 个粒子 $X_i(t)$，其更新公式如下：

$$X_{i,j}(t+1) = \begin{cases} G\left(\frac{\text{Pbest}_{i,j}(t)+\text{Gbest}_j(t)}{2}, |\text{Pbest}_{i,j}(t) - \text{Gbest}_j(t)|\right), & \text{rand} < 0.5 \\ \text{Pbest}_{i,j}(t), & \text{otherwise} \end{cases}$$

其中： $X_{i,j}(t+1)$ 表示第 i 个粒子在第 j 个维度上的新位置。 rand 是 $[0,1]$ 范围内的随机数。

为了增加种群的多样性，本文设计了一种均匀变异操作。在每次迭代中，先在 $[0,1]$ 范围内生成一个随机数，如果其值小于变异概率 p_c，则当前粒子将发生变异。变异的具体操作如下：首先，随机选择一个个体最优位置（Pbest_k）。然后，从当前粒子中随机选取 U' 元素，并用 Pbest_k 中对应的元素值替换。为了平衡种群的探索能力和开发能力，本文动态调整 p_c 和 U' 的值，其公式如下：

$$p_c = \frac{0.2}{1 + e^{t-5}}, \quad U' = \lceil p_c \cdot N \rceil$$

其中： t 为当前迭代次数,N 为种群大小。

可以看出，在PSO的早期阶段，每个粒子不仅具有较大的变异概率，而且其变异范围也相对较大，从而能够保证种群的多样性。随着迭代次数t 的增加，每个粒子的变异概率和变异范围会同时减小，这可以确保种群在后期阶段具有更强的开发能力。

与现有的进化特征选择（EFS）算法类似[24]，粒子需要通过分类器来评估其对应特征子集的分类准确性。学者们已经提出了多种分类器，例如支持向量机（SVM）、K 最近邻（KNN）和贝叶斯分类器。

尽管分类器的选择会影响最终特征子集的性能，但本文的主要目标是设计一种高效的联合特征选择算法。因此，本文选择了常用的 KNN 作为分类器。此外，本文采用文献[26]中介绍的解码方法将粒子转换为特征子集，并使用文献[26]中提出的增强记忆策略来更新每个粒子的个体最优位置（Pbest）。进一步地，以第i 个 B-参与方为例，伪代码Algorithm 3给出了本文采用的粒子更新策略。

Algorithm 3 粒子更新策略

Input: 通过第 4.3 节方法初始化的粒子群

Output: 私有的最优特征子集

1: **while** 未满足内层循环终止条件 **do**
2: 解码每个粒子的位置 [26] 并评估其适应度；
3: 按照 [26] 中的策略更新每个粒子的 *Pbest*；
4: 更新粒子群的全局最优 *Gbest*；
5: 按公式更新每个粒子的位置；
6: 执行均匀变异操作；
7: **end while**
8: 输出私有的最优特征子集，即 *Gbest*。

4 复现细节

4.1 复现主要内容

这一篇论文在网络上目前没有开源代码，那么我们需要对其框架进行分析，根据伪代码来实现FPSO-FS的主要内容。每一个B-参与方会在自己的本地执行特征选择算法，随后由中央A-参与方进行信息的收集，并收集后传递下去让每方进行评估，从而根据评估结果选出全局最优特征子集帮助指导后续的B-参与方的模型训练过程。

通过这样的方式，我们并不需要共享本地的数据很好的达到了隐私保护的需求，但又能够利用其他方的模型参数、结果帮助训练，吻合现实生活中的实际需求，具有较高的研究前景。

4.2 数据集准备

下载数据集：实验使用 15 个公开的 UCI 数据集（如 Wine、Vehicle 等），可以从 UCI Machine Learning Repository 下载。数据采用两种方式进行划分：在基于分布的划分方式中，首先将每个类别中的样本聚类为 M 组；随后，每组样本会随机分配给 B 个参与方。通过这种方式，每个参与方所持有的数据均具有不同的分布。另一种则是采用基于分布的划分方式进行划分，数据集可见表2。 其中，标注为“*” 的数据集表示使用基于分布的划分方法分配给 M 个参与方。

表 1. 实验数据与结果

No.	Dataset Name	# of Samples	# of Classes	# of Features
1	Wine	178	3	13
2	Vowel	990	11	14
3	Vehicle	846	4	18
4	Segmentation	2310	7	20
5	Ionosphere	351	2	34
6	Satellite*	856	6	36
7	Sonar	208	2	60
8	Hill_valley	606	2	100
9	MUSK1*	476	2	166
10	DNA*	2000	3	180
11	LSVT*	126	2	310
12	CNAE3	200	9	659
13	CNAE1	540	9	856
14	Yale64	165	15	1024
15	ORL32	400	40	1024

4.3 参数设置

参数设置如下：最大迭代轮次 $T_{\max} = 4$，每轮中 PSO 的最大迭代次数设置为 25，种群规模设置为 popsize = 20，用于种群初始化的参数 nummax = 8。对于 SaPSO，根据文献 [24] 的建议，编码策略的阈值设置为 $\theta = 0.6$，每种策略的选择概率为 Prob = 0.2，经验值为 LP = 10。对于 IBSO，根据文献 [25] 的建议，簇的数量设置为 2，两个选择概率值分别设置为 Pcluster = 0.8 和 Pone = 0.4。对于 NOCSA，根据文献 [19] 的建议，飞行长度设置为 $fl = 2$，动态感知概率的阈值设置为 DAPlim = 0.2。为了公平起见，所有基于种群的算法使用相同的种群规模和最大迭代次数。

具体而言，种群规模设置为 20，最大迭代次数设置为 100。对于 KNN 分类器，其参数设置为 $K = 3$ [19]。ReliefF 的最近邻样本数量设置为 5。每个算法独立运行 30 次，并计算其统计结果。

根据文献 [5] 的建议，采用双重 5 折交叉验证以避免特征选择的偏差。在每次验证过程中，使用测试数据评估算法基于训练数据生成的特征选择解的性能。在特征选择方法的迭代过程中，对训练数据进行一个内循环的 5 折交叉验证，以评估粒子或特征子集的分类性能。

在所提出的特征选择算法的搜索过程中，个体 X 的分类准确率基于内循环中的测试数据计算，用 $f(\mathrm{Samp_i})$ 表示。然后，X 在所有 K-折数据子集上的平均分类准确率作为 X 的适应度，定义如下：

$$J(X) = \frac{1}{K}\sum_{i=1}^{K} f(\mathrm{Samp_i})$$

4.4 实验环境

在此次实验中,使用Intel Corei5 处理器,8GB RAM 及充足硬盘空间的计算机，基于 Windows 10 操作系统，使用 Matlab 2024a 版本，从 UCI机器学习库获取了所需的数据集，最终在本地计算机上进行实验。

MATLAB（Matrix Laboratory）是一款功能强大的编程和数值计算软件，广泛应用于工程、科学和经济领域。其核心功能包括矩阵运算、数据可视化、算法开发、模型仿真等。MATLAB提供丰富的工具箱，覆盖信号处理、图像处理、机器学习、控制系统等多个方向，用户可轻松调用内置函数完成复杂任务。用户可通过命令窗口直接运行代码，也可编写.m脚本保存和管理项目。其丰富的文档和社区资源为学习和应用提供了便利，是科研的首选工具之一。

5 实验结果分析

我们依据论文的参数设定进行试验，使用了论文中的多种数据集，包括:Vehicle、Vowel、DNA等等数据集。关键算法参数设置：种群规模 popsize $= 20$，最大迭代次数 $T = 25$，动态概率 $pp = 0.1$ 等。基于此参数设定，我们对九个数据集进行了实验。我们选择了不同类型的数据集，这些数据集覆盖了从小规模到大规模、从低维特征到高维特征的广泛场景。具体来说，这些数据集包括了小规模经典数据集（如 Wine 和 Sonar）、中等规模的数据集（如 Vehicle 和 Ionosphere），以及特征维度较高的大规模数据集（如 DNA 和 Yale64）。数据集的具体统计信息及实验结果展示在表2中。

表 2. 实验数据

No.	Dataset Name	# of Samples	# of Classes	# of Features	# ACC
1	Wine	178	3	13	95.24±2.13
2	Vowel	990	11	14	60.18±4.47
3	Vehicle	846	4	18	67.43±2.37
4	Segmentation	2310	7	20	90.38±1.98
5	Ionosphere	351	2	34	87.45±2.27
6	Sonar	208	2	60	74.76±3.27
7	MUSK1*	476	2	166	74.37±4.17
8	DNA*	2000	3	180	76.53±2.83
9	Yale64	165	15	1024	65.37±4.38

根据表格中的实验结果可以看出，我们的实验结果与原论文中报道的结果基本一致，验证了算法的有效性和稳定性。在绝大多数数据集上，我们的算法都实现了与原论文相近的分类准确度（ACC）。例如，在经典的小规模数据集 Wine 上，我们的算法取得了 95.24 的平均分类准确率，标准差为 2.13，表明算法在低维特征数据集上具有卓越的分类性能。同时，在高维大规模数据集上，算法的表现依然非常稳健，例如在 DNA 数据集上取得了 76.53 的平均准确率，在超高维数据集 Yale64 上则取得了 65.37 的平均准确率。这些结果充分说明了算

法在处理不同规模、不同特征维度数据集时的适应性与鲁棒性，验证了其在广泛应用场景下的实用价值。

我们还对于参与者的数量进行了研究。在实验中，我们分别设置参与者数量（M）为2、3、4和5，讨论其对分类性能的影响。实验中采用了平均集成策略，图2中展示了不同参与者数量下分类准确率的变化情况。从图中可以看出：

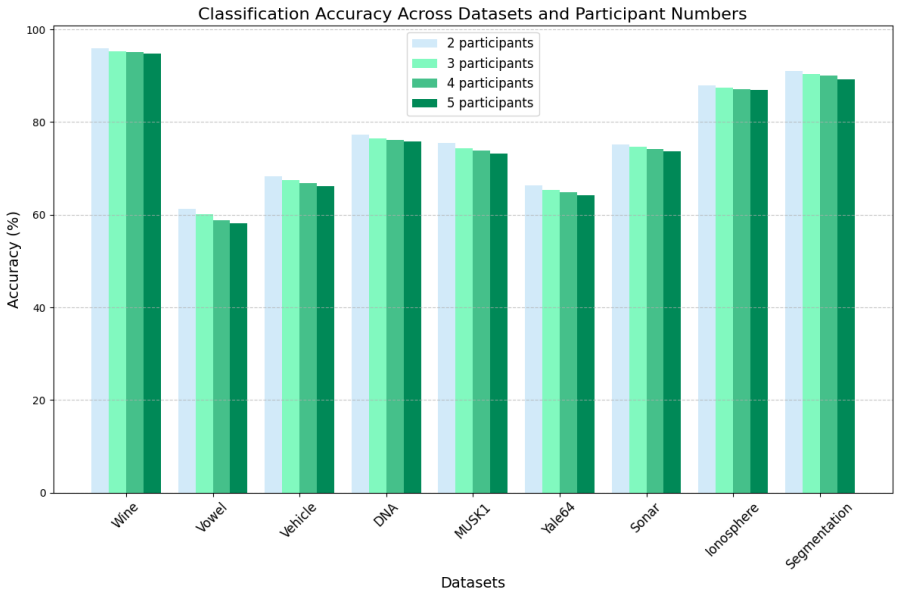


图 2. 参与者数量与分类准确率

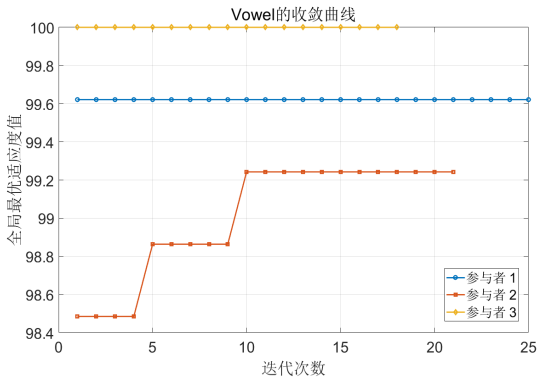
- 对于大多数数据集，当 $M = 2$ 时，分类准确率（Acc）最高，因为两个参与者拥有最多的样本数量。当参与者拥有足够的样本时，构建的分类模型具有较高的准确性；反之，当样本数量减少时，模型的分类准确率会随之下降。

- 随着参与者数量的增加，在大多数数据集上，分类性能呈现下降趋势。例如，在 Yale64 数据集上，当 $M = 2$ 时，Mean-Acc 值为 66.29%；而当 M 增加至 3 和 5 时，Mean-Acc 值分别降至 65.37% 和 64.26%。这主要是由于样本分配到每个参与者后，单个参与者的样本量不足以支撑其构建更准确的模型。

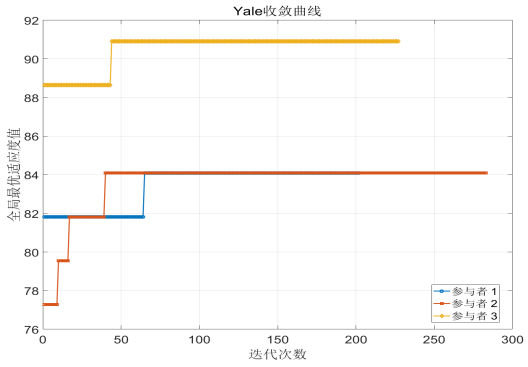
- 对于其他部分数据集，参与者数量的变化对算法性能影响较小。这是因为在 M 从 2 增加到 5 的过程中，每个参与者仍然能够获得足够的样本以构建有效的模型。

我们还对于收敛情况进行了研究，以三个参与者的情况为例，绘制了每方的收敛情况图如图3所示，这里以四个数据集为例进行详细展开。

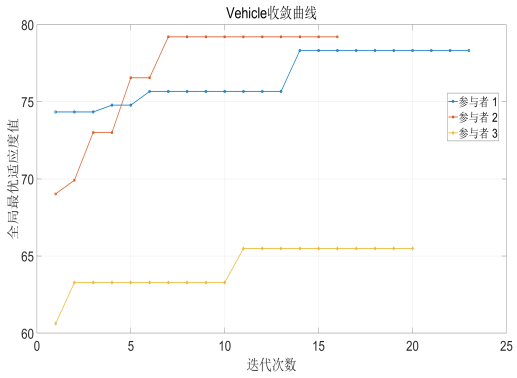
这张图展示了四个不同数据集（Vowel、Yale、Vehicle、Sonar）上的收敛曲线，分别比较了三位参与者的模型表现。在 Vowel 数据集中，所有参与者的准确率表现较为平稳，尤其是参与者 2，准确率几乎保持在 100%，显示出较强的收敛性。相比之下，Yale 数据集的初始准确率差异较大，参与者 2 的表现明显优于其他参与者，最终达到 92%，而其他两位参与者的准确率也趋于稳定，但稍逊一筹。在 Vehicle 数据集中，参与者 1 和 2 的表现相对较好，最终准确率稳定在 75%-80% 左右，而参与者 3 的准确率始终偏低，仅维持在 60%-65%，表明模型在该数据集上的适应性有较大差异。Sonar 数据集中，所有参与者的准确率均随着



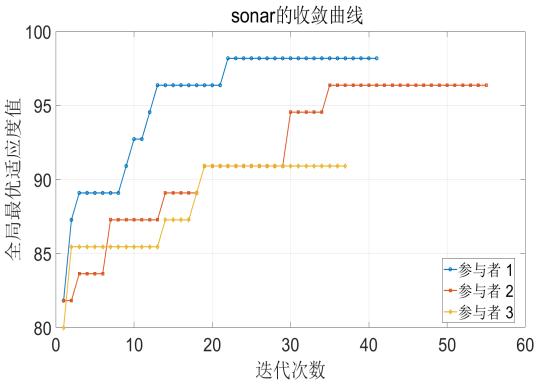
(a) Vowel的收敛曲线



(b) Yale的收敛曲线



(c) Vehicle的收敛曲线



(d) Sonar的收敛曲线

图 3. 收敛曲线

迭代逐步提升，参与者 1 的表现最佳，最终接近 100%，而参与者 2 和 3 分别稳定在 95% 和 90%。总体来看，不同数据集的收敛性和参与者间的表现差异显著，这可能与数据集的特性、模型的初始化条件及优化策略有关。改进方向可以集中在优化低准确率的数据集模型表现，并分析参与者间的差异以提升整体性能。

为了进一步分析和评估联邦学习策略的必要性，我们同作者一样通过比较 FPSO-FS（结合联邦学习的特征选择算法）与传统非联邦特征选择算法在分类准确率和特征选择效率上的表现，验证 FPSO-FS 是否在数据隐私保护场景中同时具备高效性和准确性。FPSO-FS/FL的具体策略如下：每个参与者独立运行PSO算法，直接输出其私有最优特征子集的Mean-Acc，具体结果可见表3所示。

从表格中可以看出，FPSO-FS 相较于 FPSO-FS/FL 在不同数据集上的分类准确率（Acc）表现出了一致的提升，这表明引入联邦学习策略能够有效提高特征选择的质量，而对不同数据集的改进幅度存在差异。FPSO-FS通过FL策略在参与者之间有效共享有用信息，使得特征选择更加全面，缓解了单个参与者独立运行时因数据量不足或分布不均导致的过拟合问题。同时，某些数据集（如Sonar）的改进幅度较小，可能与其特征分布或数据特性有关，需进一步优化算法以适应更复杂的数据集场景。

表 3. FPSO-FS 和 FPSO-FS/FL 在3个B-参与者中获得的平均Acc值。

Data	**Measure**	**FPSO-FS/FL**				**FPSO-FS**			
		B1	B2	B3	Avg.	B1	B2	B3	Avg.
Wine	Acc	89.05	93.14	87.32	90.07	94.93	93.32	96.85	95.39
Vowel	Acc	53.64	53.91	52.71	52.71	56.27	59.22	62.27	60.04
Vehicle	Acc	58.72	60.46	57.62	60.41	68.23	69.48	61.63	66.83
DNA	Acc	68.36	71.37	67.62	71.94	73.46	71.78	79.36	76.37
Sonar	Acc	71.84	69.73	73.10	72.05	71.84	75.19	73.28	73.71

6 总结与展望

论文提出了一种新型的联邦粒子群优化特征选择算法，旨在解决分布式环境下的特征选择问题。在方法设计中，通过引入改进的粒子更新策略、增强的记忆机制以及统一变异操作，有效平衡了算法的全局探索能力与局部开发能力。同时，为了适应联邦学习的特性，本文提出了特征集装策略，以在多参与方之间整合最佳特征子集。实验结果表明，该算法在分类准确性和特征选择效率方面均显著优于现有方法。

尽管论文提出的联邦粒子群优化特征选择算法（FPSO-FS）在隐私保护下的特征选择问题上展现了良好的性能，但仍存在一些不足之处和需要进一步研究的方向。

与其他联邦特征选择算法类似，FPSO-FS 的运行时间相对较高。未来可以探索利用新的技术手段（如代理模型）来降低算法的计算成本，从而提高其效率。并且FPSO-FS 在某些特殊场景下的表现仍有待改进，例如部分参与方数据中包含空类别的情况，这可能会对算法的性能造成影响。联邦学习中数据分布的非独立同分布特性（Non-IID）和隐私保护的要求可能对算法的性能产生一定影响。未来可以结合差分隐私或安全多方计算等技术，进一步增强算法的鲁棒性和隐私保护能力，以应对更复杂的数据分布和更严格的隐私保护需求。自适应特性选择也是未来研究的重要方向。针对不同任务和数据集的多样性，可以设计自适应的特征选择机制，根据实际需求动态调整参数和策略，从而提高算法的灵活性和适用性。此外，将提出的算法应用于更多实际问题也是未来研究的重点之一。例如，该算法可以被用于解决一些潜在的实际问题，如罕见疾病的跨医院联合学习、跨银行的消费者行为分析等。随着隐私保护意识的提升和大数据技术的广泛应用，此类问题将持续增加，对算法的实用性和扩展性提出更高要求。

综上所述，论文提出的 FPSO-FS 算法为隐私保护下的特征选择问题提供了一种有效的解决方案，也为联邦学习领域的进一步发展奠定了基础。未来的研究将着重于优化算法性能、增强鲁棒性与隐私保护能力、设计自适应特性选择机制，以及验证其在实际问题中的效果，以推动该领域的持续发展。

参考文献

[1] Zahra Beheshti. Bmpa-tvsinv: A binary marine predators algorithm using time-varying sine and v-shaped transfer functions for wrapper-based feature selection. *Knowledge-Based Systems*, 252:109446, 2022.

[2] Hemanta Kumar Bhuyan and Narendra Kumar Kamila. Privacy preserving sub-feature selection in distributed data mining. *Applied soft computing*, 36:552–569, 2015.

[3] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. Distributed feature selection: An application to microarray data classification. *Applied soft computing*, 30:136–150, 2015.

[4] Bin Cao, Jianwei Zhao, Po Yang, Peng Yang, Xin Liu, Jun Qi, Andrew Simpson, Mohamed Elhoseny, Irfan Mehmood, and Khan Muhammad. Multiobjective feature selection for microarray data via distributed parallel algorithms. *Future Generation Computer Systems*, 100:952–981, 2019.

[5] Ke Chen, Bing Xue, Mengjie Zhang, and Fengyu Zhou. Evolutionary multitasking for feature selection in high-dimensional classification via particle swarm optimization. *IEEE Transactions on Evolutionary Computation*, 26(3):446–460, 2021.

[6] Ke Chen, Feng-Yu Zhou, and Xian-Feng Yuan. Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. *Expert Systems with Applications*, 128:140–156, 2019.

[7] Kamalika Das, Kanishka Bhaduri, and Hillol Kargupta. A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to-peer networks. *Knowledge and information systems*, 24:341–367, 2010.

[8] Kusum Deep et al. A random walk grey wolf optimizer based on dispersion factor for feature selection on chronic disease prediction. *Expert Systems with Applications*, 206:117864, 2022.

[9] Pedram Ghamisi and Jon Atli Benediktsson. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. *IEEE Geoscience and remote sensing letters*, 12(2):309–313, 2014.

[10] Kazım Hanbay. A new standard error based artificial bee colony algorithm and its applications in feature selection. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4554–4567, 2022.

[11] Emrah Hancer, Bing Xue, and Mengjie Zhang. A survey on feature selection approaches for clustering. *Artificial Intelligence Review*, 53(6):4519–4545, 2020.

[12] Yasser Jafer, Stan Matwin, and Marina Sokolova. A framework for a privacy-aware feature selection evaluation measure. In *2015 13th Annual Conference on Privacy, Security and Trust (PST)*, pages 62–69. IEEE, 2015.

[13] Yunmei Lu, Mingyuan Yan, Meng Han, Qingliang Yang, and Yanqing Zhang. Privacy preserving feature selection and multiclass classification for horizontally distributed data. *Mathematical Foundations of Computing*, 1(4):331–348, 2018.

[14] Thomas Marill and D Green. On the effectiveness of receptors in recognition systems. *IEEE transactions on Information Theory*, 9(1):11–17, 1963.

[15] Kamlesh Mistry, Li Zhang, Siew Chin Neoh, Chee Peng Lim, and Ben Fielding. A micro-ga embedded pso feature selection approach to intelligent facial emotion recognition. *IEEE transactions on cybernetics*, 47(6):1496–1509, 2016.

[16] Laura Morán-Fernández, Verónica Bolón-Canedo, and Amparo Alonso-Betanzos. Centralized vs. distributed feature selection methods based on data complexity measures. *Knowledge-Based Systems*, 117:27–45, 2017.

[17] Xiaodong Na, Min Han, Weijie Ren, and Kai Zhong. Modified bbo-based multivariate time-series prediction system with feature subset selection and model parameter optimization. *IEEE Transactions on Cybernetics*, 52(4):2163–2173, 2020.

[18] Yang Qin and Masaaki Kondo. Federated learning-based network intrusion detection with a feature selection approach. In *2021 International conference on electrical, communication, and computer engineering (ICECCE)*, pages 1–6. IEEE, 2021.

[19] Behrouz Samieiyan, Poorya MohammadiNasab, Mostafa Abbas Mollaei, Fahimeh Hajizadeh, and Mohammadreza Kangavari. Novel optimized crow search algorithm for feature selection. *Expert Systems with Applications*, 204:117486, 2022.

[20] Mina Sheikhalishahi and Fabio Martinelli. Privacy-utility feature selection as a privacy mechanism in collaborative data classification. In *2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 244–249. IEEE, 2017.

[21] Xian-Fang Song, Yong Zhang, Yi-Nan Guo, Xiao-Yan Sun, and Yong-Li Wang. Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data. *IEEE Transactions on Evolutionary Computation*, 24(5):882–895, 2020.

[22] A Wayne Whitney. A direct method of nonparametric measurement selection. *IEEE transactions on computers*, 100(9):1100–1103, 1971.

[23] Bing Xue, Mengjie Zhang, Will N Browne, and Xin Yao. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on evolutionary computation*, 20(4):606–626, 2015.

[24] Yu Xue, Bing Xue, and Mengjie Zhang. Self-adaptive particle swarm optimization for large-scale feature selection in classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(5):1–27, 2019.

[25] Yu Xue, Qi Zhang, and Yan Zhao. An improved brain storm optimization algorithm with new solution generation strategies for classification. *Engineering Applications of Artificial Intelligence*, 110:104677, 2022.

[26] Yong Zhang, Dunwei Gong, Ying Hu, and Wanqiu Zhang. Feature selection algorithm based on bare bones particle swarm optimization. *Neurocomputing*, 148:150–157, 2015.

[27] Jun Zhao, Long Chen, Witold Pedrycz, and Wei Wang. Variational inference-based automatic relevance determination kernel for embedded feature selection of noisy industrial data. *IEEE Transactions on Industrial Electronics*, 66(1):416–428, 2018.