SFE：一种适用于高维数据的简单、快速且高效的特征选择算法

Behrouz Ahadzadeh、Moloud Abdar、Fatemeh Safara、Abbas Khosravi，IEEE高级会员、Mohammad Bagher Menhaj、Ponnuthurai Nagaratnam Suganthan，IEEE会士

摘要—本文针对高维数据集提出了一种新的特征选择算法，称为SFE（简单、快速且高效）。SFE算法使用一个搜索智能体和两个算子（非选择算子与选择算子）执行搜索过程。该算法包含两个阶段：探索阶段和开发阶段。在探索阶段，非选择算子在整个问题搜索空间中进行全局搜索，以寻找无关、冗余、无用及噪声特征，并将特征状态从选中模式更改为未选中模式。在开发阶段，选择算子在问题搜索空间中搜索对分类结果影响较大的特征，并将特征状态从未选中模式更改为选中模式。所提出的SFE能够有效地从高维数据集中进行特征选择。然而，在降低数据集维度后，其性能无法显著提升。在这种情况下，可以使用进化计算方法，在新的、缩减后的搜索空间中寻找效率更高的特征子集。为解决这一问题，本文提出了一种混合算法SFE-PSO（粒子群优化），用于寻找最优特征子集。本文在40个高维数据集上比较了SFE和SFE-PSO用于特征选择的效率与有效性，并将其性能与六种近期提出的特征选择算法进行了比较。结果表明，所提出的两种算法显著优于其他算法，可作为高效且有效的算法，用于从高维数据集中选择特征。

索引词— 高维数据集、特征选择、粒子群优化、进化计算方法。

一、引言

计算机系统和新技术的发展，促使各领域产生并存储了大量高维数据集

医疗保健、电子商务、生物信息学、社交媒体和交通运输等应用 [1], [2], [3]。机器学习算法广泛用于分类问题；然而，维数灾难使人们难以获得理想结果，因为高维数据集中的数据分布会对针对低维数据集设计的算法产生不利影响。此外，数据的高维度会增加机器学习算法在数据分类中的时间复杂度和空间复杂度。为了解决这些对机器学习算法效率产生负面影响的问题，通常会提出降维方法 [3]。

一般来说，降维方法通常在整个分类流程的预处理阶段执行，可分为特征提取和特征选择（FS）。特征提取算法的主要目标是根据数据集的主要特征创建新特征，并降低数据集的维度。新构建的特征通常是主要特征的线性或非线性组合 [4]。特征选择算法会选取包含数据重要信息的相关特征子集，从而提升机器学习算法的性能。此外，不重要特征（即无关、冗余、琐碎和噪声特征）会从特征集中移除 [3], [5] ，因为这些特征会导致过拟合并降低算法性能。特征选择算法通常用于预处理阶段，以降低算法的计算和空间复杂度，从而提升机器学习算法的整体性能。此外，算法的泛化能力也会得到改善。

特征提取方法的基础是根据主要特征构建新特征，因此，新特征不具备主要特征的物理意义。因此，这些方法不适用于文本挖掘和遗传分析等应用 [3]。相比之下，特征选择方法会保留特征的物理意义，因为它们直接选择能够为所述应用提供正确解释的相关特征。此外，由于特征提取方法在高维数据集上的计算复杂度很高，因此这些方法在某些应用中可能无法达到预期性能。基于上述原因，近年来特征选择方法得到了研究和改进 [3], [4]。

一般而言，特征选择算法可分为三类：基于过滤器的、基于包装器的和嵌入式 [3]。典型的基于过滤器的特征选择算法包含两个主要步骤。第一步，根据距离、相关性、一致性和信息增益等不同准则，确定每个特征的重要性和排名。第二步，过滤排名较低的特征 [3], [6],[7]。不过，某些基于过滤器的特征选择算法（例如相关性特征选择算法（CFS）[8], ）不一定遵循这两个步骤。

典型的基于包装器的特征选择算法使用机器学习分类器评估选定特征。这些方法通常包含两个主要步骤。第一步，利用搜索方法搜索最优特征子集。第二步，使用 K 最近邻（KNN）和支持向量机（SVM）等各种分类方法评估最优特征子集 [9],

通讯作者：Moloud Abdar（电子邮箱：m.abdar1987@gmail.com（Behrouz Ahadzadeh 任职于伊斯兰阿扎德大学加兹温分校电气、计算机与信息技术工程系，伊朗加兹温（电子邮箱：b.ahadzade@yahoo.com）。Moloud Abdar 和 Abbas Khosravi 任职于迪肯大学智能系统研究与创新研究所（IISRI），澳大利亚（电子邮箱：m.abdar1987@gmail.com 和 abbas.khosravi@deakin.edu.au）。Fatemeh Safara 任职于伊斯兰阿扎德大学伊斯兰城分校计算机工程系，伊朗伊斯兰城（电子邮箱：fsafara@yahoo.com）。Mohammad Bagher Menhaj 任职于阿米尔卡比尔理工大学电气工程系，伊朗德黑兰（电子邮箱：menhaj@aut.ac.ir）。P. N. Suganthan 任职于卡塔尔大学工程学院 KINDI计算研究中心，卡塔尔多哈（电子邮箱：p.n.suganthan@qu.edu.qa）。

[10]。重复这两个步骤，以达到预期的性能 [11], [12]。然而，一些基于包装器的特征选择算法也不一定遵循这两个步骤。嵌入式特征选择方法与基于包装器的方法类似，使用分类算法进行特征选择，不同之处在于特征选择嵌入在分类器结构中 [13]。决策树方法是特征选择方法之一 [14]。选定特征是在训练决策树后最终树中存在的特征 [13], [15]。

一些先前的研究 [3], [16], [17] 将特征选择分为四类，其中包括混合特征选择方法。在混合特征选择方法中，会结合使用基于过滤器的、基于包装器的和嵌入式特征选择算法。总体而言，基于包装器的方法和嵌入式方法比基于过滤器的方法更有效，因为它们会在特征选择过程中与分类器交互。相比之下，基于过滤器的方法的计算成本低于基于包装器的方法和嵌入式方法。嵌入式方法通常比基于包装器的方法成本更低。不过，这些方法大多只能与某些分类方法配合使用，例如决策树、支持向量机 [3], [11], [18]。混合方法也具有令人满意的性能，因为它们利用了不同算法的优势，但仍面临计算成本高的挑战。因此，为了提高特征选择过程的效率，我们应降低基于包装器的方法的计算复杂度；与基于过滤器的方法和嵌入式方法相比，基于包装器的方法效率更高。基于这一原因，本文旨在为高维数据提供一种基于包装器的特征选择方法。

基于包装器的特征选择方法使用搜索算法来寻找最优特征子集。在数据集中搜索最优特征子集的最简单方法，是遍历整个问题搜索空间并评估每个可能子集的性能，这种方法称为穷举搜索方法。在包含 n 个特征的数据集中，可能的子集数量为 $2^n - 1$。因此，对于较大的 n，即使使用当今最强大的计算机，采用完整搜索来寻找最佳特征子集在实际中也几乎不可能 [9],[18]-[20]。因此，研究员没有采用穷举搜索算法，而是使用启发式算法，例如序列前向选择（SFS）[21], 、序列后向选择（SBS） [22], 、加减 r 选择法（LRS）[23], 、序列浮动前向选择（SFFS）和序列浮动后向选择（SFBS）算法 [24]。尽管这些算法能够显著降低完整搜索的时间复杂度，但由于其在搜索过程中的策略一贯不变，可能会陷入局部最优点 [9], [19]。鉴于此，近年来，研究员使用了另一类称为进化计算（EC）方法的搜索算法进行特征选择。这些算法实现简单、具有全局搜索能力、可灵活用于各种应用，能在可接受的空间和时间复杂度内找到近似全局最优解，并且由于具有随机性，能够摆脱局部最优点，因此受到关注并被用于特征选择 [25], [26]。尽管这些方法在低维数据集上进行特征选择的高效率已得到证明，但数据集中的特征数量增加会导致算法陷入局部最优点并发生早熟收敛。其原因是问题搜索空间呈指数级增长，并且问题搜索空间中会增加许多局部最优点 [18]。此外，在高维数据集上评估进化计算方法得到的解，会带来很高的计算成本。

因此，本文旨在为高维数据集提供一种高效的特征选择算法，以解决此类数据集中特征选择面临的问题。我们期望所提出的算法能够以更少的特征和更低的计算成本，实现最高的分类准确率。本文的主要贡献总结如下：

1. 提出一种面向高维数据集的新型特征选择算法，称为 SFE。该算法旨在通过采用两个新算子，以尽可能短的时间、用最少的特征实现最高的分类准确率。

2. 提出一种称为非选择算子的新算子，用于执行所提出算法的探索阶段。该算子在全局范围内探索搜索空间，并以合理的计算成本找出不重要特征。3. 提出一种称为选择算子的新算子，用于执行所提出算法的开发阶段。该算子对非选择算子得到的结果进行局部搜索，并返回在探索阶段被视为不可选择的相关特征。4. 提出一种新的混合算法，称为 SFE-PSO，该算法基于所提出的 SFE 算法和一种进化计算方法——粒子群优化（PSO）算法。该混合算法旨在通过 SFE 算法降低数据集的维度，并利用 PSO 算法的强大潜力，在新的低维数据集中搜索最优子集。5. 提出一种基于 SFE 的通用框架，称为 SFE-ECs，用于考察并展示其他进化计算方法在高维数据特征选择任务中的性能潜力。

II. 相关工作

近年来，研究人员使用进化计算方法从高维数据集中选择特征。大多数基于进化计算方法的特征选择方法，其核心思路是将极大的搜索空间划分为若干较小的空间，并在这些空间中应用进化计算方法。粒子群优化是用于从高维数据集中进行特征选择、最受关注的进化计算方法之一，这是因为它计算速度快、效率高、易于实现且参数较少，同时也已证明该算法在特征选择中的效率 [27]。本节简要回顾此前提出的一些面向高维数据集的特征选择方法。

Tran等人 [28] 提出了一种基于PSO算法、采用变长粒子的特征选择算法。采用变长粒子的目的是通过较短的粒子缩小高维数据集的搜索空间，从而降低计算成本和内存消耗。该算法使用一种称为对称不确定性（SU）方法的基于过滤器的特征选择算法，确定每个特征的排名，并将其用于种群初始化。此外，研究者还提出了一种称为长度变化机制的方法，使PSO算法能够跳出搜索空间中的局部最优点。该方法的效率

高度依赖于SU算法和长度变化机制的效率。

Song 等人 [29] 提出了一种基于变规模协同协进化 PSO 的特征选择算法，用于处理高维数据集。该算法使用对称不确定性算法确定每个特征的重要性，以减少高维数据的搜索空间维度。形成较小的子空间后，使用多群体粒子群优化算法搜索各个子空间。在搜索过程中，采用多种方法移除重复粒子并添加新粒子，以提高种群多样性。最后，通过交叉方法合并各子群获得的解，从而得到最终解。该算法存在一些不足，包括不同群体带来的高计算成本，以及由于不同群体之间缺乏相关特征的交互而陷入局部最优点的可能性。

Tan 等人 [5] 提出了一种名为 DimReM 的新方法，利用进化计算方法降低搜索空间的高维度。该方法旨在从数据集中实际移除一些不重要特征，以缩小问题搜索空间。由于进化计算方法的特性，如果种群中的最佳值陷入局部最优点，种群中的其他成员也可能陷入局部最优解。此外，由于缩小搜索空间规模的速度较慢，要获得预期的解需要付出较高的计算成本。

文献[27]提出了一种用于高维数据集的基于进化多任务的特征选择算法，称为PSO-EMT。该方法研究了两个搜索空间不同的任务。第一个任务使用重要特征，其重要性由ReliefF算法确定[6], ，第二个任务则使用所有特征。该算法利用交叉算子在任务之间共享信息。此外，还采用了另外两种机制来缩小搜索空间并维持种群多样性。ReliefF算法表现不佳，导致算法停滞在局部最优点。此外，由于第二个任务包含所有特征，评估粒子群优化算法得到的解需要付出很高的计算成本。此外，使用多种群优化算法执行搜索过程也需要很高的计算成本。

类似地，文献[30]提出了一种用于高维数据的基于进化多任务的特征选择算法和一种粒子群优化算法，称为MTPSO。在该算法中，采用一种基于ReliefF算法的方法，将较大的搜索空间转换为若干个较小搜索空间的任务。随后，使用多种群粒子群优化算法搜索每个任务的空间。为了提高搜索过程的效率，该方法采用一种称为知识迁移的方法，在各个子集之间交换有关特征的重要信息。

Song 等人 [31] 提出了一种面向高维数据集的特征选择算法，称为 HFS-C-P。该算法包含三个重要阶段。第一阶段使用对称不确定性算法去除不重要特征。第二阶段使用快速相关性引导特征聚类方法对第一阶段保留下来的特征进行聚类。最后，在第三阶段，使用各簇中的特征初始化 PSO算法的粒子，并通过搜索过程获得一个最优子集。第一阶段中对称不确定性算法的性能不佳，以及错误地删除某些相关特征，可能导致该算法在某些数据集上的表现不佳。

Wang 等人 [32] 提出了一种名为 SaWDE 的特征选择算法，该算法基于加权差分进化算法，并采用自适应机制。多种群机制提高了种群多样性，多个子种群在问题搜索空间中搜索，以找到最优特征子集。该算法采用两种自适应机制，以提高差分进化算法在不同数据集上进行特征选择时的效率。这些机制旨在选择合适的变异算子以及差分进化算法参数的合适取值。此外，该算法使用权重模型来确定每个特征的重要性。所提方法的主要缺点是计算成本较高。

Cheng 等人 [33] 提出了一种名为 SM-MOEA 的基于导向矩阵的多目标进化算法，用于高维数据集的特征选择。该算法利用搜索智能体，通过导向矩阵计算特征的重要性，并据此引导种群朝着更优的解搜索。此外，该方法采用降维和个体修复这两种算子，从搜索智能体及问题搜索空间中移除不重要特征。该方法的缺点之一是，对于包含大量特征的数据集，构建导向矩阵会造成较高的内存消耗。此外，降维机制若错误地移除相关且重要的特征，可能导致该算法停滞于局部最优点，从而降低算法性能。

总之，上述方法旨在缩小问题搜索空间，采用进化计算方法在低维数据集上进行搜索，并在高维数据集上达到预期性能。然而，上述某些方法（如 [32] 和 [33], ）在进行特征选择时需要消耗大量内存，且计算成本较高。另一些方法（如 [31], ）的计算成本较低；但由于采用基于过滤器的方法进行降维，在某些数据集上可能会陷入局部最优点。当相关特征未能被正确识别时，就会出现这种情况，导致这些特征被错误地从数据集中筛除。因此，分类效率会较差。总的来说，以往提出的特征选择算法面临的挑战可概括如下：在高维搜索空间中，算法可能陷入局部最优点；计算成本高、内存消耗大；分类效率较差。因此，本文提出一种面向高维数据集的有效且高效的特征选择算法，以克服上述挑战。

III. 所提方法

进化计算方法凭借全局搜索过程中准确的结果以及合理的计算成本和内存消耗，已成功应用于特征选择应用 [10]。然而，在高维数据集上使用这些方法时，大多数方法都会遭遇维数灾难和高昂的计算成本 [31]。在

此外，高维数据集的问题搜索空间呈指数级增长，会产生许多局部极小点。因此，在这些情况下，进化计算方法会陷入局部最优点，并导致早熟收敛 [18]。本文提出了一种特征选择方法，以克服上述挑战。该方法结构简单、计算成本低、内存消耗少，且无需复杂算子。与进化计算方法类似，所提出的方法会对问题搜索空间进行迭代搜索。然而，该方法仅使用一个搜索智能体搜索搜索空间，不同于基于种群的进化算法。

进化计算方法从高维数据集中选择特征时，计算成本高、内存消耗大的原因在于需要将搜索智能体存储在主存中，然后对搜索智能体得到的解进行计算。因此，使用一个搜索智能体可以降低所提出的特征选择算法的计算成本和内存消耗。

特征选择是一个二元问题，其中特征有两种状态：已选择和未选择。在所提出的算法中，搜索智能体 X 是特征选择问题的一个解，用于表示每个特征的状态。因此，智能体 X 的表示方式如图 1 所示。

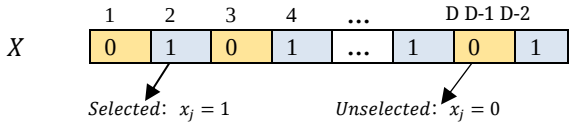

图 1. X 的编码。

其中，$x_j = 0$ 表示未选择第j个特征，$x_j = 1$ 表示选择了该特征。因此，搜索智能体X 采用二进制编码表示。

该方法基于两个简单的算子——选择算子和非选择算子，以及它们在搜索过程中对搜索智能体X 的使用。选择算子将特征的状态从未选中模式变为选中模式，而非选择算子则将特征的状态从选中模式变为未选中模式。在SFE的搜索过程中，选择算子和非选择算子作用于X，以获得算法的更优解。选择算子的使用取决于问题搜索空间的条件以及 X 在问题搜索空间中的位置。换言之，在采用SFE算法的特征选择过程中，在搜索过程的每个阶段，非选择算子都会作用于X。然而，在某些数据集中，选择算子可能从搜索过程一开始就会被使用；而在另一些数据集中，直到搜索过程结束，它都可能从未被使用。本节将详细介绍这两个算子。SFE算法的伪代码见算法1。

A. 非选择算子 有许多不重要的

高维数据集中的特征。另一方面，特征选择旨在以最少的特征实现最高的分类准确率。因此，为了实现这一重要目标，有必要在问题的整个搜索空间中进行全局搜索，以找出不重要特征并将其更改为非选择模式。为了在 SFE算法中进行全局搜索，我们使用非选择算子（算法1第7至12行）。在任何搜索过程中，该算子通过不选择大量不重要特征，使搜索智能体更接近问题的最优点。为了对 X 应用非选择算子，需要计算该算子所作用的特征数量。对于不同的数据集，该算子可以采用不同的值。在 SFE算法中，使用公式 (1) 来确定 UN，即为非选择算子的操作生成的伪随机数数量（算法1第8行）。

$$UN = \lceil UR \times nvar \rceil \quad (1)$$

其中，$nvar$ 是搜索空间中的特征数量，即数据集的特征数量。UR 是非选择算子率，取值范围为0到1。如果为UR 选择较大的值，那么在搜索过程的每个阶段，该算子都会将大量特征的状态更改为未选择模式。如果UR 的值仍然较大，那么经过搜索过程的若干阶段后，除了非选择算子之外，还会使用选择算子。因此，在这种情况下，搜索过程开始时会使用非选择算子进行全局搜索，此时算法处于探索阶段。在接下来的搜索步骤中，算法除了探索阶段外，还会执行开发阶段。要在SFE算法的探索阶段和开发阶段之间建立适当平衡，一个简单的方法是让UR 值线性递减，从较大的值逐渐减小到较小的值。对于SFE算法，UR 值根据以下公式确定（算法1第25行）：

$$UR = (UR_{max} - UR_{min}) \times ((Max_FEs - FEs)/FEs) + UR_{min} \quad (2)$$

其中，UR_{max} 是UR的初始值，UR_{min} 是UR的最终值。$Max\ FEs$ 是函数_评估的最大次数， FEs 是 SFE 当前的函数评估次数。通过公式 (1)确定UN 值后，会生成UN 个介于 1 和解X 中选定特征数量之间的伪随机数（算法1第10 行）。这些伪随机数用于访问索引向量中存储的、处于选定模式的特征。换言之，生成的每个伪随机数都用作索引向量的索引，而该向量中存储着选定特征的编号。

我们回顾 SFE算法在一个包含20个特征的数据集上的搜索过程。图2通过一个示例展示了非选择算子的操作。为了应用非选择算子，使用公式 (1)计算该算子必须处理的特征数量。假设 $UR = 0.3$ 和 $nvar = 20$，那么 UN 将为6。这意味着必须生成6个服从1到9之间均匀分布的伪随机数（9是X中选定特征的数量）。假设生成的6个伪随机数分别为3、7、9、7、3和6。利用这些数可以获得选中模式下特征的编号。如图2所示，这些特征的编号为3、6、8、9、10、13、16、17和20。特征编号存储在索引向量中。为此，我们获取其模式必须改为未选中模式的特征编号（算法1第11行）。

这些特征编号为8、16、20、16、8和13。根据图2，这些特征被改为未选中模式，并由此得到一个新的解，称为 X_{New}（算法1第12行）。如果X_{New} 的适应度值大于X的适应度值，那么X_{New} 将传递到下一代，继续搜索过程。

算法1：SFE特征选择算法

输入：包含原始特征集的训练数据集，$F=(f_1,f_2,f_3,\dots,D)$
输出：包含特征子空间的训练数据集，$S=(s_1,s_2,s_3,\dots,d)$
1 Max_FEs：最大函数评估次数
2 $Nvar$：数据集中的特征数量
3 在搜索空间中初始化一个个体 $X=(x_1,x_2,\dots,x_D)$
4 计算 X 的适应度：$fit(X)$
5 $FEs=0$
6 **While** $FEs \le Max_FEs$ **Do**
7 $X_{New}=X$
8 $UN=\lceil UR\times nvar\rceil$ % 将要切换为未选中模式的特征数量
9 $Index$=在 X 中查找选定特征的索引
10 U=生成一个介于 1 和 X 中选定特征数量之间的随机数 UN
11 $K=index(U)$
12 集合 $X_{New}(K)=0$ % 非选择操作
13 **If** t小时$e\ number\ of\ selected\ features\ in\ X_{New}\ is == 0$
14 $X_{New}=X$
15 $Index$=查找 X 中选定特征的索引
16 $SN=1$ % 要切换为选择模式的特征数量
17 S=在 X 中未选特征的数量范围内，生成一个介于 1 和该数量之间的随机数 SN
18 $K=index(S)$
19 设置 $X_{New}(K)=1$ % 选择操作
20 **End if**
21 **If** $fit(X_{New})\ge fit(X)$
22 $X=X_{New}$
23 **End if**
24 $FEs=FEs+1$
25 $UR=(UR_{max}-UR_{min})\times((Max_FEs-FEs)/FEs)+UR_{min}$
26 **End while**

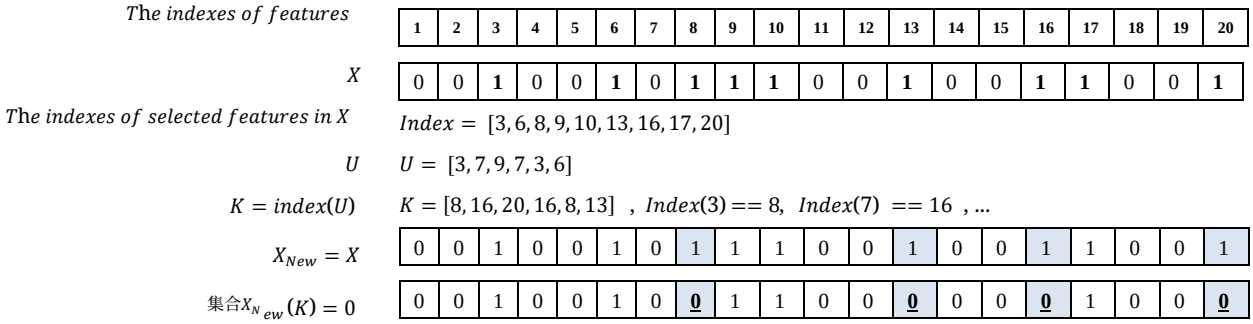

图 2. 非选择操作示意图。

现在假设 X_{New} 适应度值大于 X 适应度值，因此搜索过程继续使用 X_{New}。假设 $UR=0.1$，那么 UN 值将为 2。这意味着必须在 1 到 5 之间生成两个服从均匀分布的伪随机数（5 是 X中选定特征的数量）。假设这两个伪随机数是 3 和 5。在这种情况下，在 X中选定的特征（即特征编号 9、6、3、10 和 17）中，必须由非选择算子将两个特征更改为未选中模式。为此，通过命令 $K=index(U)$ 可得到它们的编号 3 和 17。这样，便使用非选择算子将这两个特征更改为未选中模式。此过程之后，如果 X_{New} 的适应度值大于 X，则 X_{New} 将传递到下一代，以继续搜索过程。但如果 X_{New} 的适应度值不大于 X的适应度值，则搜索过程将继续使用 X。

B. 选择算子

使用非选择算子进行全局搜索后，数据集中的所有特征都可能变为未选中模式。在这种情况下，将应用选择算子（算法1第13至20行）。选择算子使算法能够重新选择状态已变为未选中模式的重要特征。假设我们再使用非选择算子执行一步，且X 变为X_{New}，如图3所示。X_{New} 解中的所有特征都已变为未选中模式，而算法在全局搜索中并未将X解更改为更优的子集。因此，将使用选择算子。为此，首先根据图3将X_{New} 更改为X，然后在1与未选中特征的数量之间生成SN 个伪随机数。假设$SN=1$ 且生成的伪随机数为13，那么特征

编号16的特征将变为选中模式。由于特征选择旨在以尽可能少的特征实现最高准确率，通常应为SN选择较小的值。另一方面，由于经过若干搜索步骤后，算法可能会选择少量特征，因此为SN 使用较大的值可能无法带来预期的改进。因此，较小的值，例如例如，1 可适用于 SN。该算子对解 X 进行微小调整，其行为类似于围绕最优解进行幅度极小的局部搜索；因此，该算子增强了 SFE算法的开发能力。

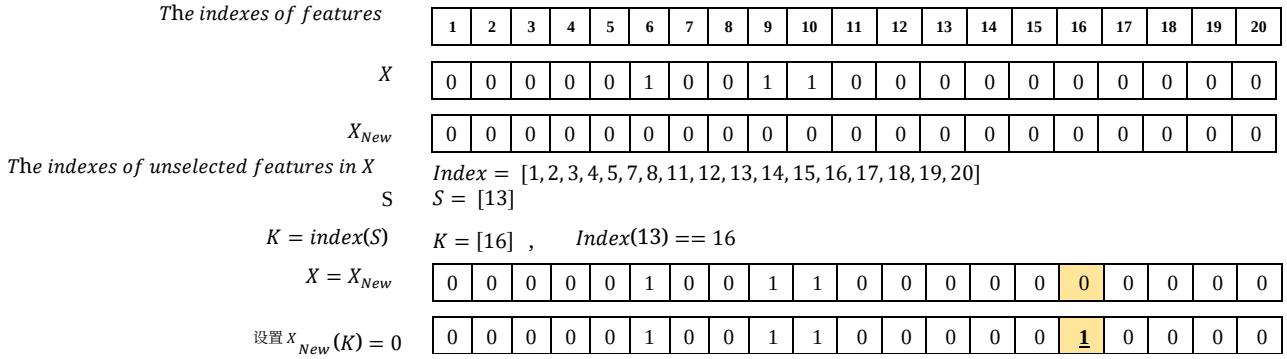

图 3. 选择操作示意图。

C. 适应度函数

分类准确率用于评估搜索过程中每一步获得的新解，如下所示：

$$fit = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \qquad (3)$$

其中， T_P 和 T_N 分别表示分类器正确分类为正类和负类的数据集正例和负例数量。此外， F_P 和 F_N 分别表示分类器错误分类为负类和正类的数据集正例和负例数量。

f_1	f_2	f_3	…	f_{500}	f_{501}	…	f_{1000}	…			f_{1500}	…				f_{2000}
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(a) 特征选择前的训练数据集

f_1	f_2	f_3	…	f_{500}	f_{501}	…	f_{1000}	…			f_{1500}	…				f_{2000}
0	1	0	0	1	1	0	1	0	0	1	1	1	0	0	0	1

(b) 使用 SFE算法进行特征选择后的训练数据集

f_2	f_{500}	f_{501}	…			f_{2000}
s_1	s_2	s_3	…			s_{50}
1	1	1	1	1	1	1

(c) 从数据集中移除未选特征以缩减搜索空间后的训练数据集

图 4. 所提出的混合算法示意图。

D. SFE 与粒子群优化的混合

尽管进化计算方法在高维数据集上进行特征选择的效率并不高，但这些方法在低维数据集上进行特征选择的高效率已得到证实 [34]-[40]。因此，如果能通过从高维数据集中实际移除不重要特征来缩减问题搜索空间，就可以发挥进化计算方法的巨大潜力，从而取得更好的结果。

本文采用这一思路，提出了一种基于 SFE 和 PSO算法的混合算法。图4展示了所提出的混合算法的流程。该混合算法包含以下三个步骤：

1. 在搜索初期使用 SFE 算法，在广阔的搜索空间中进行全局搜索。

2. 完成 SFE 算法的搜索步骤后，物理移除未选中的模式特征，以降低搜索空间的维度。3. 使用 PSO 算法在降维后的搜索空间中进行搜索。

在数据集中，执行特征选择之前，所有特征均处于选定状态（图4(a)）。首先，使用SFE算法从数据集中选出重要特征，并排除不重要特征（图4(b)）。使用SFE算法选出相关特征后，将所有未选定特征从数据集中实际删除，从而创建一个维度远小于原始数据集的新数据集（图4(c)）。因此，与SFE得到的子集相比，PSO算法能够在这一新的搜索空间中获得最优特征子集。算法2给出了SFE-PSO混合算法的伪代码。

算法2：所提出的 SFE-PSO 算法

输入：具有原始特征集的训练数据集，$F = (f_1, f_2, f_3, \dots, D)$
输出：具有特征子空间的训练数据集，$S = (s_1, s_2, s_3, \dots, d)$
1 Max_FEs：最大函数评估次数
3 在搜索空间中初始化一个个体 $X = (x_1, x_2, \dots, x_D)$
4 计算 X 的适应度：$fit(X)$
5 $FEs = 0$
6 **While** $FEs \leq Max_FEs$ **Do**
7 使用算法1 选择和取消选择特征，并将最优解的适应度保存到适应度向量中
8 **If** $FEs > 2000$ *and* (适应度向量(FEs) − 适应度向量$(FEs - 1000) == 0$)
9 S = 从训练数据集中移除未选特征
10 用位置 x 和速度 v 初始化粒子。
11 $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d}),\ v_i = (v_{i,1}, v_{i,2}, \dots, v_{i,d}),\ i = 1,2, \dots, N_p,\ j = 1,2, \dots, d$
12 $Set\ x_{1,j=1\dots d} = 1,$
13 计算每个粒子的适应度 $fit(x_1), \dots, fit(x_{N\ \ p})$
14 更新每个粒子的$pbest$ 和$gbest$
15 **While** $FEs \leq Max_FEs$ **Do**
16 **For** $i = 1$ *to* N_p **Do**
17 **For** $j = 1$ *to* d **Do**
18 $v_{i,j} = w\, v_{i,j} + c_1 r_1 (pbest_{i,j} - x_{i,j}) + c_2 r_2 (gbest_{i,j} - x_{i,j})$
19 $x_{i,j} = \{1, \ if\ rand() \leq Z(v_{i,j}) \ 0, \ 否则$
20 **End for**
21 计算每个粒子的适应度 $fit(x_1), \dots, fit(x_{N\ \ p})$
22 为每个粒子更新 gbest 和 $pbest$
23 **End for**
24 $FEs = FEs + N_p$
25 **End while**
26 **End if**
27 $FEs = FEs + 1$
28 **End while**

从算法2的伪代码可以看出，首先使用SFE算法选出相关特征（算法2第7行）。接下来，假设函数评估计数器大于2000年，且SFE算法在此前的1000次函数评估中未能得到更优解。在这种情况下，搜索过程将使用PSO算法在新的S空间中继续进行（算法2第8行）。

为了在新的子空间中继续使用 PSO算法进行搜索，首先从原始数据集中物理移除处于未选中模式的特征，并创建一个维度较低的新数据集。然后，在初始化粒子速度和位置值后，将第一个粒子的位置更改为 SFE算法获得的最优解。这意味着，第一个粒子中的所有特征都处于选中模式。这样，PSO算法便会在新空间中继续搜索更优的解；与原始搜索空间相比，新空间要小得多。

E. 基于EC的特征选择与 SFE

为了更好地解决特征选择问题，我们可以将 SFE算法与各种进化计算方法相结合。算法3展示了 SFE-EC混合算法的通用框架。SFE-ECs 包含两个阶段。第一阶段使用SFE算法在高维数据的巨大搜索空间中选择特征（算法3第6行）。随后，从原始特征集中物理移除未选定的特征（算法3第8行）。第二阶段在新的搜索空间中继续使用进化计算方法进行搜索；与原始搜索空间相比，该空间要小得多（算法3第9行）。

总体而言，特征选择过程从使用 SFE算法开始，并持续进行，直至满足 SFE算法完成搜索过程的指定条件。随后，搜索过程继续采用针对低维数据集特征选择提出的进化计算方法，例如粒子群优化 [35], [36], 差分进化 [37], 遗传算法 [38],灰狼优化（GWO）[39], 樽海鞘群算法（SSA）[40], 蚁群优化（ACO） [41], 人工蜂群（ABC）算法 [42]。

IV. 实验设计

为评估所提出特征选择方法的有效性和效率，我们在40个真实数据集上进行了实验。我们使用了九种算法与所提出的特征选择方法进行比较。首先，将SFE和SFE-PSO算法的结果与六种算法进行比较：BDE [43],BPSO [44], BGA [5], DimReM-BDE [5], DimReM-BGA [5],和DimReM-PSO [5]。随后，将所提算法的结果与另外三种最先进的基于进化计算的特征选择算法进行比较，包括HFS-C-P [31],SaWDE [32], 和SM-MOEA [33] 算法。选择BDE、BPSO和BGA这三种算法，是为了评估它们作为高维数据集特征选择领域中三种知名进化计算算法的性能。此外，DimReM-BDE、DimReM-BGA和DimReM-PSO这三种算法在降低数据维度方面效率很高，因此也将其与所提方法进行比较。选择HFS-C-P、SaWDE和SM-MOEA这三种算法，也是因为它们是近期针对高维数据集特征选择提出的。

算法3：SFE-EC 框架

输入：具有原始特征集的训练数据集，$F=(f_1,f_2,f_3,\dots,D)$
输出：具有特征子空间的训练数据集，$S=(s_1,s_2,s_3,\dots,l)$
Max_FEs：最大函数评估次数
在搜索空间中初始化一个个体 $X=(x_1,x_2,\dots,x_D)$
计算 X 的适应度：$fit(X)$
While *the sto pping criterion is not met* ***Do***
    使用算法1选择和取消选择特征
    If *the stopping criterion is met*
        S = 通过从训练数据集中移除未选特征
        在新的子空间中调用进化计算方法进行特征选择，$\mathbf{S}=(\mathbf{s_1},\mathbf{s_2},\mathbf{s_3},\dots,l)$
    End if
End while

有关实验的更多信息见补充材料的I节。所提算法的源代码可在 https://github.com/Ahadzadeh2022/SFE.git 公开获取。

五、结果与讨论

A. 分类准确率

分类准确率是评估特征选择算法有效性的重要标准之一。本节根据分类准确率，将所提算法的性能与其他算法进行比较。采用显著性水平为 0.05 的 Wilcoxon 秩和检验，评估 SFE-PSO 算法在 30 次独立运行中的性能。在以下结果中，“+”、“–”和“≈”分别表示 SFE-PSO 显著优于、劣于和与所比较的算法相近。此外，还利用 Friedman 检验的平均秩，将 SFE-PSO 算法与其他算法进行比较，以评估其有效性并检测显著改进。表 I 展示了各算法在 30 次独立运行中的结果。

如表 I 所示，在所考察的大多数数据集上，SFE算法在 30 次独立运行中的平均分类准确率均优于其他六种算法。在某些数据集上，SFE算法相较于其他算法的优势十分显著。例如，在 warpAR10P 和前列腺癌数据集上，SFE算法的准确率分别比其他六种算法中表现最佳的算法高出 25.17% 和 23.40%。然而，在某些数据集上，这种提升幅度很小。例如，在脑肿瘤1数据集上，SFE算法仅比其他六种算法中表现最佳的 BPSO算法高出 0.89%。

考察并比较 SFE算法与 SFE-PSO算法的结果可以发现，在40个数据集中，SFE-PSO算法有22个数据集的表现优于SFE算法。在某些数据集中，这一提升幅度显著。例如，在CML治疗和癌数据集中，SFE-PSO算法分别比 SFE算法高出6.16%和5.60%。而在另一些数据集中，这一提升幅度很小。例如，在 SMK_CAN_187数据集中，SFE-PSO算法比 SFE算法高出0.0981%。图 5展示了各算法在40个数据集上的平均分类准确率结果。BPSO算法的平均分类准确率为83.15%，在六种算法——BDE、BPSO、遗传算法、DimReM-BDE、DimReM-BGA和DimReM-BPSO——中表现最佳。将SFE和SFE-PSO算法的效率与BPSO进行比较可知，这两种算法分别比BPSO算法高出8.71%和10.13%，优势显著。Wilcoxon和Friedman 检验结果表明，SFE-PSO算法表现优异：在所有数据集的分类准确率指标上，其性能均显著优于或与其他算法相当。

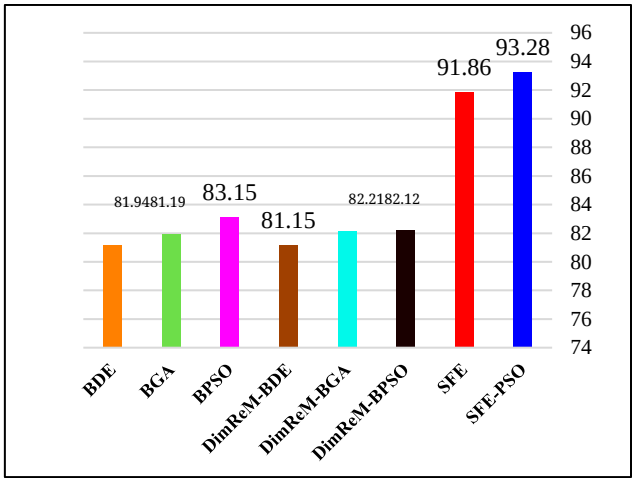

图5. 八种算法的平均分类准确率。

图6展示了SFE算法、SFE-PSO算法以及其他六种算法的收敛图，其中包含有关这些算法在特征选择过程中的行为的重要信息。六种算法BDE、BGA、BPSO、DimReM-BDE、DimReM-BGA和DimReM-BPSO在全部40个数据集上的行为都固定不变。这些算法从问题搜索空间中的一个随机点开始搜索，在搜索早期获得更好的解，最终则停留在包含大量局部最优点的巨大搜索空间中。可以看出，二进制粒子群优化算法相较于其他算法具有相对优势。相比之下，观察SFE和SFE-PSO算法的行为则非常有意义。在乳腺癌、弥漫性大B细胞淋巴瘤和GLI_85等大多数数据集上，SFE表现出预期的行为：从一个随机点开始搜索，并在搜索过程中获得更好的解。补充材料第三节报告了关于图6的更多信息。补充材料中的表III和表IV给出了这些算法在40个数据集上经过30次独立运行后，分类准确率的最佳值、最差值、均值和标准差。

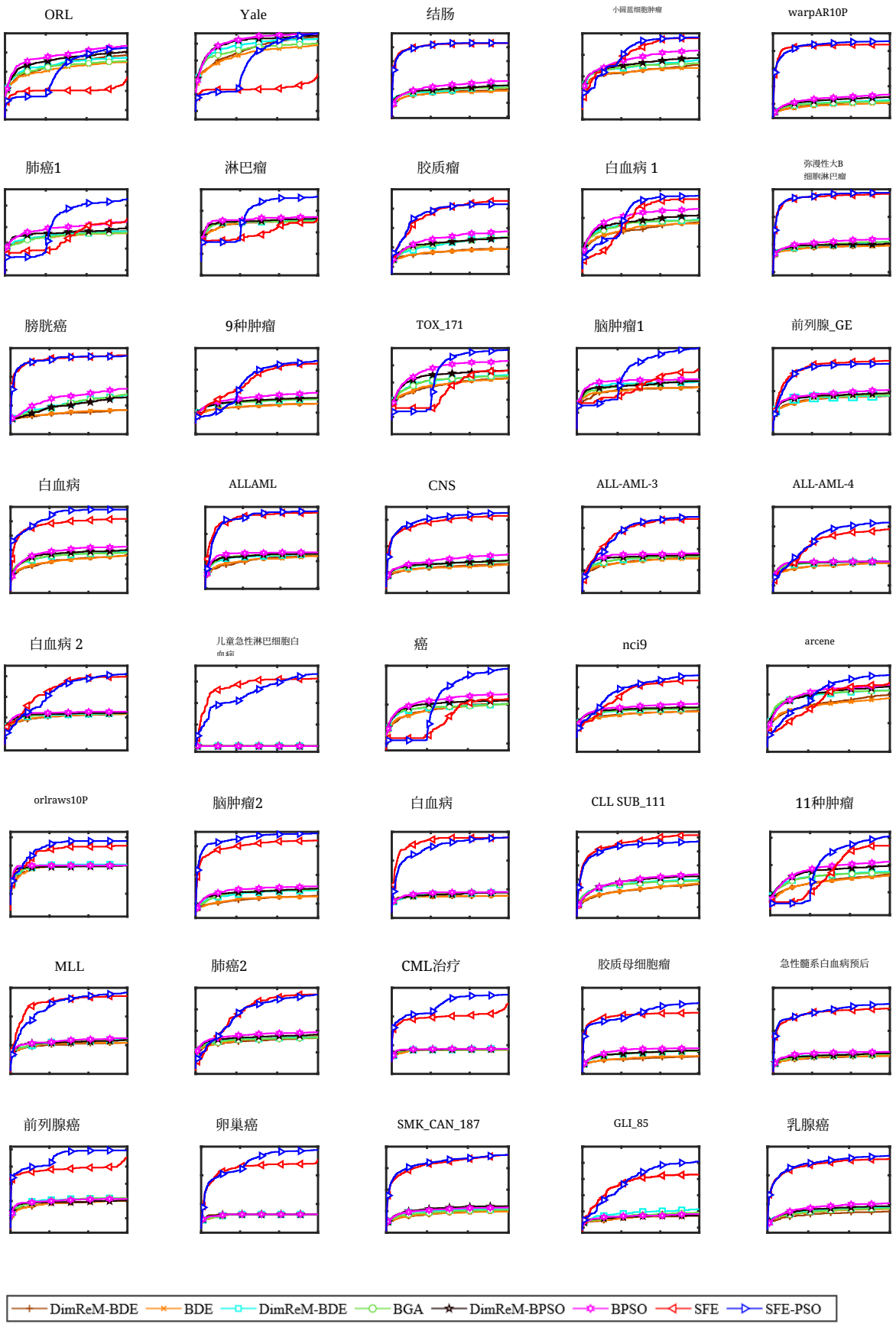

图 6. KNN分类器下，SFE 和 SFE-PSO 与其他算法在所有数据集上的收敛曲线。

表 I 八种算法在 30 次独立运行中获得的分类准确率均值和标准差

	数据集编号	BDE [43]	BGA [5]	BPSO [44]	DimReM-BDE [5]	DimReM-GA [5]	SFE DimReM-BPSO [5]		SFE-PSO
1	ORL	96.49±0.18 (+)	96.55±0.27 (+)	**97.35±0.22 (≈)**	96.86±0.19 (+)	96.73±0.22 (+)	97.05±0.31 (≈)	95.71±0.66 (+)	97.24±0.57
2	Yale	73.82±0.40 (+)	74.02±0.73 (+)	75.15±0.26 (≈)	74.82±0.36 (≈)	74.57±0.51 (≈)	74.95±0.39 (≈)	70.41±1.20 (+)	75.**23±1**.35
3	结肠	79.30±0.91 (+)	80.46±1.19 (+)	82.84±1.20 (+)	79.90±0.99 (+)	81.23±1.59 (+)	81.34±1.55 (+)	96.44±2.44 (≈)	96.**51±2**.46
4	小圆蓝细胞肿瘤	91.94±0.72 (+)	93.04±1.42 (+)	97.64±0.79 (+)	92.66±0.64 (+)	93.84±1.60 (+)	94.27±1.20 (+)	98.89±1.99 (≈)	99.**09±1**.14
5	warpAR10P	60.56±0.72 (+)	63.84±1.10 (+)	68.46±1.25 (+)	61.89±0.69 (+)	65.38±1.18 (+)	63.87±1.28 (+)	93.64±3.39 (+)	95.**53±2**.29
6	肺癌1	96.70±0.22 (+)	96.73±0.29 (+)	97.24±0.33 (+)	96.81±0.24 (+)	96.80±0.38 (+)	96.93±0.36 (+)	97.41±0.64 (≈)	98.**32±0**.37
7	淋巴瘤	97.00±0.31 (+)	97.03±0.36 (+)	97.42±0.60 (+)	97.11±0.42 (+)	97.14±0.45 (+)	97.28±0.51 (+)	97.28±0.95 (+)	99.**35±0**.95
8	胶质瘤	84.53±1.04 (+)	87.73±1.55 (+)	89.13±1.71 (+)	84.66±0.95 (+)	87.66±1.66 (+)	87.53±2.01 (+)	**97.00±2.61 (≈)**	96.13±2.09
9	白血病 1	94.06±0.82 (+)	94.60±1.08 (+)	96.57±0.68 (+)	94.25±0.83 (+)	94.74±0.97 (+)	95.49±0.97 (+)	98.28±1.38 (+)	98.**87±0**.72
10	弥漫性大B细胞淋巴瘤	89.76±0.03 (+)	90.46±0.86 (+)	91.05±0.77 (+)	89.91±0.63 (+)	90.50±0.78 (+)	90.17±1.00 (+)	99.16±1.10 (≈)	99.**36±1**.03
11	膀胱癌	78.40±2.48 (+)	84.00±3.95 (+)	85.90±4.83 (+)	78.40±2.02 (+)	83.60±4.62 (+)	82.80±5.46 (+)	**97.50±2.16 (≈)**	97.30±2.69
12	9种肿瘤	54.22±1.36 (+)	56.22±1.37 (+)	59.22±1.29 (+)	54.16±0.95 (+)	56.27±1.43 (+)	56.88±1.79 (+)	72.77±7.80 (+)	74.**16±4**.92
13	TOX_171	96.44±0.40 (+)	96.73±0.54 (+)	98.54±0.42 (+)	96.54±0.46 (+)	96.87±0.61 (+)	97.37±0.52 (+)	97.44±1.50 (+)	99.**78±0**.32
14	脑肿瘤1	90.37±0.53 (+)	90.96±0.38 (+)	91.11±2.8e-14 (+)	90.37±0.53 (+)	90.85±0.47 (+)	90.92±0.42 (+)	92.00±1.14 (+)	94.**00±0**.95
15	前列腺 GE	88.90±0.50 (+)	89.09±0.61 (+)	90.30±0.63 (+)	88.86±0.51 (+)	88.92±0.66 (+)	89.51±0.78 (+)	**97.05±1.45 (≈)**	96.34±1.68
16	白血病	82.92±0.90 (+)	84.00±1.15 (+)	87.42±0.68 (+)	83.21±0.67 (+)	84.28±1.31 (+)	84.91±1.16 (+)	95.82±2.03 (+)	99.**05±0**.72
17	ALLAML	90.90±0.68 (+)	91.42±0.49 (+)	91.61±4.3e-14 (+)	91.16±0.64 (+)	91.33±0.67 (+)	91.33±0.57 (+)	98.93±0.14 (≈)	99.**22±1**.11
18	CNS	73.00±1.10 (+)	74.22±1.04 (+)	77.05±2.08 (+)	73.27±0.92 (+)	74.22±1.21 (+)	74.61±1.67 (+)	91.50±3.34 (≈)	92.**61±2**.79
19	ALL-AML-3	87.76±0.61 (+)	88.33±0.65 (+)	89.00±0.54 (+)	88.17±0.67 (+)	88.33±0.65 (+)	88.57±0.60 (+)	97.13±2.18 (≈)	97.**68±1**.55
20	ALL-AML-4	85.68±0.57 (+)	86.09±0.36 (+)	86.09±0.36 (+)	85.68±0.57 (+)	86.04±0.26 (+)	86.09±0.36 (+)	94.40±3.10 (+)	96.**00±1**.76
21	白血病 2	85.82±0.46 (+)	85.95±0.24 (+)	86.38±0.64 (+)	85.86±0.54 (+)	86.00±0.00 (+)	86.04±0.26 (+)	94.87±2.34 (≈)	95.**60±1**.70
22	儿童急性淋巴细胞白血病	35.00±0.00 (+)	35.00±0.00 (+)	35.00±0.00 (+)	35.00±0.00 (+)	35.00±0.00 (+)	35.00±0.00 (+)	50.66±17.1 (≈)	51.**80±15**.7
23	癌	90.08±0.33 (+)	90.11±0.59 (+)	91.36±0.45 (+)	90.08±0.42 (+)	90.08±0.60 (+)	90.50±0.58 (+)	90.86±2.25 (+)	96.**47±0**.98
24	nci9	58.72±0.94 (+)	61.66±1.10 (+)	62.38±1.13 (+)	59.11±0.95 (+)	60.11±1.23 (+)	60.61±1.11 (+)	73.11±4.62 (+)	75.**55±3**.31
25	arcene	89.36±0.69 (+)	90.66±1.20 (+)	91.69±0.66 (+)	89.90±0.78 (+)	90.71±0.82 (+)	91.40±0.67 (+)	92.01±3.24 (+)	93.**36±1**.75
26	orlraws10P	98.00±0.00 (+)	98.00±0.00 (+)	98.00±0.00 (+)	98.03±0.18 (+)	98.06±0.25 (+)	98.00±0.37 (+)	99.40±0.77 (≈)	99.**73±0**.44
27	脑肿瘤2	72.46±1.45 (+)	75.00±1.14 (+)	76.00±0.90 (+)	72.73±1.11 (+)	74.66±1.09 (+)	75.20±1.62 (+)	92.00±3.36 (+)	94.**46±2**.90
28	白血病	88.82±0.38 (+)	89.46±0.92 (+)	89.52±0.92 (+)	88.87±0.46 (+)	89.52±0.71 (+)	89.41±1.00 (+)	99.00±1.29 (≈)	99.**06±1**.70
29	CLL SUB_111	73.19±0.78 (+)	74.47±1.27 (+)	76.79±1.01 (+)	73.48±0.77 (+)	77.39±1.47 (+)	76.04±1.06 (+)	**90.81±3.07 (–)**	88.59±2.93
30	11种肿瘤	83.06±0.57 (+)	83.59±0.93 (+)	85.67±0.52 (+)	83.31±0.62 (+)	83.77±0.75 (+)	84.85±0.73 (+)	88.54±2.09 (+)	90.**29±1**.24
31	MLL	90.41±0.50 (+)	90.81±0.77 (+)	91.21±0.76 (+)	90.41±0.61 (+)	91.04±0.75 (+)	90.90±0.68 (+)	98.55±1.54 (≈)	99.**17±1**.27
32	肺癌2	93.48±0.27 (+)	93.36±0.38 (+)	94.09±0.28 (+)	93.32±0.28 (+)	93.40±0.32 (+)	93.64±0.40 (+)	**97.38±0.71 (≈)**	97.37±0.76
33	CML治疗	56.66±1.4e-14 (+)	56.96±1.01 (+)	57.46±1.63 (+)	56.96±1.09 (+)	57.44±1.49 (+)	57.46±1.63 (+)	89.01±5.94 (+)	95.**17±3**.87
34	胶质母细胞瘤	78.08±1.35 (+)	70.32±1.37 (+)	71.68±0.74 (+)	67.84±0.98 (+)	70.64±1.25 (+)	70.64±1.38 (+)	91.50±3.34 (+)	92.**96±3**.06
35	急性髓系白血病预后	68.29±0.82 (+)	69.09±0.90 (+)	70.25±0.82 (+)	68.36±1.04 (+)	69.30±1.30 (+)	69.60±1.01 (+)	88.32±3.85 (+)	92.**58±4**.05
36	前列腺癌	70.00±0.00 (+)	71.80±2.44 (+)	71.40±2.70 (+)	70.20±1.00 (+)	71.60±2.38 (+)	70.20±2.27 (+)	95.20±4.89 (+)	99.**40±1**.65
37	卵巢癌	95.27±1.4e-14 (+)	95.27±1.4e-14 (+)	95.27±1.4e-14 (+)	95.27±1.4e-14 (+)	95.27±1.4e-14 (+)	95.27±1.4e-14 (+)	99.10±0.70 (≈)	99.**76±0**.42
38	69.88±0.44 (+) SMK_CAN_187		70.19±0.52 (+)	70.69±0.46 (+)	70.08±0.43 (+)	70.54±0.57 (+)	71.16±0.53 (+)	83.03±2.36 (≈)	83.**13±3**.18
39	GLI_85	91.52±0.47 (+)	91.85±0.75 (+)	91.67±0.67 (+)	91.45±0.52 (+)	92.27±0.66 (+)	91.48±0.70 (+)	96.54±1.27 (+)	98.**15±1**.23
40	乳腺癌	66.86±0.84 (+)	68.45±0.45 (+)	70.30±0.77 (+)	66.86±0.84 (+)	68.78±0.75 (+)	69.21±1.21 (+)	85.70±4.15 (+)	86.**83±3**.97
Wilcoxon检验 (+ | ≈ | –) 40 | 0 | 0		40 | 0 | 0	38 | 2 | 0	39 | 1 | 0	39 | 1 | 0	38 | 2 | 0	22 | 17 | 1	–
Friedman 检验（平均秩）7.40		5.53	3.29	6.79	5.01	4.58	2.26	**1.15**	

B. 选定特征的数量

选定特征的数量是评估特征选择算法有效性的另一个重要标准。补充材料中的表 V 展示了各算法在 30 次独立运行中选定特征平均数量方面的结果。在 40 个数据集中，有 33 个数据集包含数千个以上的特征，而 SFE算法从中选定的特征少于 100 个；与其他六种算法选定的特征相比，这一结果十分显著。例如，在 12,627 个前列腺癌特征中，SFE算法选定了 24 个特征。其他六种算法从每个数据集中选定的特征约占一半。

此外，SFE-PSO算法的结果与SFE相似。比较SFE算法和SFE-PSO算法时，值得注意的是，在白血病 1、TOX_171、脑肿瘤1和淋巴瘤等一些数据集中，SFE-PSO算法选定的特征数量多于SFE算法。总体而言，当SFE算法在SFE-PSO算法设定的2000年评估次数内，因陷入局部最小值等原因，无法进一步减少特征数量并提高分类准确率时，便会使用PSO算法在新的空间中继续搜索。根据PSO算法的特性，该算法在不选取大量特征方面的能力有限；同时，考虑到问题的搜索空间条件，平均约有一半的特征会转为未选中模式。图6对此进行了清晰展示，其中给出了SFE算法和SFE-PSO算法在白血病 1、TOX_171、脑肿瘤1和淋巴瘤数据集上的收敛图

。观察图6中这些算法在上述数据集上的收敛图可以发现，SFE算法在最初的2000年函数评估中未能取得成功；因此，该算法在此阶段选定的特征数量远多于其在随后4000年函数评估中找到的特征数量。当搜索过程在最初的2000年函数评估后由PSO算法继续时，该算法会将约一半的特征从选中模式切换为未选中模式。因此，在SFE-PSO选定的特征数量多于SFE的数据集中，PSO算法会在新的空间中继续搜索，其搜索空间中的特征数量约为最终解所含特征数量的两倍。此外，表格显示，在卵巢癌和弥漫性大B细胞淋巴瘤等其他一些数据集中，SFE-PSO算法选定的特征数量少于SFE算法。这表明，SFE算法在接下来的4000年评估次数中未能减少特征数量并提高分类准确率；然而，改由PSO继续搜索后，算法以更少的特征取得了更好的结果。

图7展示了八种算法在 40个数据集上选定特征的平均数量。可以看出，SFE和SFE-PSO算法选定的特征数量分别为 92 和 121，显著少于其他算法；其他算法平均选定的特征超过 4400 个。所提算法的运行时间分析见补充材料第二节。

为进一步评估所提算法的有效性和效率，将这些方法的结果进行比较

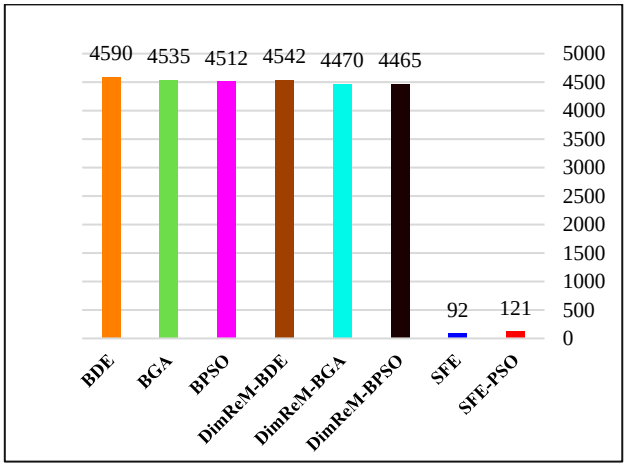

图 7. 八种算法选定特征数量的均值。

与三种最先进的基于进化计算的特征选择算法，即 HFS-C-P [31], SaWDE [32] 和 SM-MOEA [33] 算法。实验分为两部分进行。在第一部分中，使用 SM-MOEA[33] 中的 11 个不同数据集，将所提算法与 SaWDE 和 SM-MOEA 算法进行比较，以考察所提算法的有效性和效率。这些实验中，SaWDE 和 SM-MOEA 算法的结果按照上一节所述的方法获得。也就是说，在特征选择过程中，评价函数使用了 K 值为 =1 的 KNN分类器。此外，采用 5折交叉验证将数据集划分为训练集和测试集。每种算法均独立运行 30 次。最大适应度函数评估次数设为 6000。两种算法的其他参数均采用算法中建议的参数。在第二部分中，将所提算法的结果与 HFS-C-P 算法在两篇论文实验中共同使用的 8 个数据集上的结果进行比较。因此，本节中 SFE 和 SFE-PSO 算法的结果基于 HFS-C-P [31] 中采用的适应度函数（支持向量机和 K 值为 =3 的 KNN分类器，以及 10 折交叉验证）和最大函数评估次数（5000 次）。表 II 展示了所提算法、SaWDE 和 SM-MOEA 算法在 SM-MOEA [33] 中使用的 11 个数据集上的结果，包括分类准确率的最佳值、均值和标准差（Std），选定特征的数量（规模）以及运行时间（30 次独立运行的平均运行时间）。

从表 II 可以看出，在 11 个数据集中，SFE-PSO 的平均分类准确率在 8 个数据集上高于 SM-MOEA算法。在另外 3 个数据集中，两种方法的结果相近。Wilcoxon 检验和 Friedman 检验的结果也证实了 SFE-PSO 的优势，因为它始终显著优于或与 SM-MOEA算法表现相近。此外，表中结果显示，SFE-PSO 的平均分类准确率在 11 个数据集中的 4 个数据集上高于 SaWDE。在 6 个数据集中，两种算法的结果相近；在 1 个数据集中，SaWDE算法优于 SFE-PSO算法。总体而言，Wilcoxon 检验和 Friedman 检验结果表明，SFE-PSO算法优于 SaWDE算法。

表 II SFE、SFE-PSO、SaWDE [32] 和 SM-MOEA [33]的比较结果

数据集编号	时间 规模 算法	Best	均值±标准差
肺癌1 1	97.89±0.4(≈) 463 + **99 02** .( ) 38.267(–) SM-MOEA		
	SaWDE 98.09±0.1(≈) 1343 + 98 53 .( ) 149.23(+)		
	SFE 97.41±0.6(≈) 9853 262 .(–) 127.4 (+)		
	SFE-PSO 77.9 281	**99 02**	**98 32±0 3** . . .
胶质瘤2	SM-MOEA 503 + **100** ( ) 18.76 (–)		94.53±2.1(+)
	SaWDE 1710 + **100** ( ) 87.72 (+)		95.60±2.1(≈)
	SFE 100 235 (–) 26.8 (–)		**97.00±2.6**(≈)
	SFE-PSO 28.633 249	**100**	96.13±2.0
白血病 1 3	SM-MOEA 560 + **100** ( ) 59.36 (–)		98.25±1.0(+)
	SaWDE 2795 + **100** ( ) 212 (+)		98.49±0.5(≈)
	SFE 100 250 (–) 46.43 (–)		98.28±1.3(+)
	SFE-PSO 115.5 261	**100**	**98.78±0.7**
弥漫性大B 细胞淋巴瘤 4	SM-MOEA 592 + **100** ( ) 18.06 (≈)		99.23±0.9(≈)
	SaWDE 3106 + **100** ( ) 188 (+)		97.79±1.4(+)
	SFE 100 244 (–) 29 (+)		99.16±1.1(≈)
	SFE-PSO 18.93 248	**100**	**99.36±1.0**
9种肿瘤 5	SM-MOEA 65.55±5.2(+) 642 + 76 66 .( ) 36.6 (–)		
	SaWDE **74.32±7.8**(≈) 2573 + 82 84 .( ) 301.65(+)		
	SFE 72.77±7.8(+) 8666 250 .(–) 65.83 (–)		
	SFE-PSO 153.53 265	85 00	74 16±4 9 . . .
脑肿瘤 1 6	92.62±1.3 (+) 689 + 95 55 .( ) 67.40 (–) SM-MOEA		
	SaWDE **94.60±0.7** (–) 3597 + 95 49 .( ) 185.38 (–)		
	SFE 92.00±1.1 (+) 9444 251 . (–) 110.23 (–)		
	SFE-PSO 202.8 267	**96 66**	94 00±0 9 . . .
癌 7	90.51±0.8 (+) 882 + 93 10 .( ) 567.13 (+) SM-MOEA		
	SaWDE 93.43±1.2 (+) 9079 + 95 36 .( ) 1088.2(+)		
	SFE 90.86±1.4 (–) 9484 223 . (–) 147.81 (–)		
	SFE-PSO 363.5 235	**99 00**	**96 34±1 6** . . .
脑肿瘤 2 8	87.00±2.3 (+) 1163 + 94 46 .( ) 35.13 (+) SM-MOEA		
	SaWDE 10716 + 96 ( ) 155.2 (+)		88.75±3.4 (+)
	SFE 98 245 (–) 38.23 (+)		92.00±3.3 (+)
	SFE-PSO **23.36** 258	**100**	**94 46±2 9** . .
白血病 3 9	SM-MOEA 93.14±2.1 (+) 1023 + 97 14 . ( ) 36.03 (–)		
	SaWDE 97.86±1.1 (+) 10012 +	**100** ( ) 177 (+)	
	SFE 100 241 (–) 51.44 (–)		99.00±1.2 (≈)
	SFE-PSO 100 255 119 93.		**99.06±1.7**
11种肿瘤 10	85.07±0.8 (+) 1349 + 86 82 .( ) 420.43(≈) SM-MOEA		
	SaWDE 90.12±1.4 (≈) 12965 + 92 45 .( ) 1114.4(+)		
	SFE 88.54±2.0 (+) 9252 252 .(–) 141.03(–)		
	SFE-PSO 416.16 273	**93 09**	**90 29±1 2** . . .
CLL SUB_111 11	SM-MOEA 78.04±2.3 (+) 1144 + 84 66 .( ) 65.46 (+)		
	SaWDE 83.02±3.2 (+) 11692 + 90 00 .( ) 1577.2(+)		
	SFE 90.**81±3.0** (–) 9470 243 .(–) 55.56 (+)		
	SFE-PSO **48.36** 251	**95 57**	88 59±2 9 . . .
Wilcoxon检验(+ | ≈ | –)	3 | 2 | 6 11 | 0 | 0 – SM-MOEA		8 | 3 | 0
	SaWDE 9 | 0 | 1 11 | 0 | 0 –		4 | 6 | 1
	SFE 4 | 0 | 7 0 |0 | 11 –		5 | 4 | 2
	SFE-PSO – – –		–
Friedman 检验 (平均秩)	173 . SM-MOEA 3.00	3.64 3.18	
	SaWDE 3.91 4.00	2.64 2.36	
	SFE 1.91 **1.00**	2.59 2.55	
	SFE-PSO 2.45 2.00	**1.64** **1.41**	

如表 II 所示，在全部 11 个数据集中，SFE-PSO 的运行时间均低于 SM-MOEA 和 SaWDE。Wilcoxon 检验和 Friedman 检验结果也证实了 SFE-PSO算法的优势，因为在所有数据集中，它都显著优于 SM-MOEA 和 SaWDE算法。这些结果表明，SFE-PSO算法相比 SM-MOEA 和 SaWDE算法具有较低的计算成本。值得注意的是，SFE 和 SFE-PSO 算法在所有数据集上的运行时间相近。与此同时，随着数据集特征数量的增加，SM-MOEA 和 SaWDE算法的运行时间急剧增长。此外，SaWDE算法在所有数据集上的运行时间都显著偏高，在 11种肿瘤和 CLL SUB_111 等超高维数据集上尤其如此。例如，在 CLL SUB_111 数据集上，SaWDE算法的运行时间约为 SFE-PSO算法的 47 倍。

SFE-PSO选定的特征数量高于SM-MOEA算法。Wilcoxon检验和Friedman 检验的结果也证明，SM-MOEA算法优于SFE-PSO算法。SM-MOEA算法在选定特征数量方面表现出色的原因是

该算法使用了降维算子。然而，我们注意到，从数据集中错误地移除某些相关特征会导致分类准确率较差。因此，根据统计检验结果，在11个数据集中，有8个数据集上SM-MOEA算法的平均分类准确率低于SFE-PSO算法。在其余3个数据集中，这两种算法的表现相近。

此外，表 III 展示了所提出的算法与 HFS-C-P 算法的结果对比。从表 III 可以看出，在 8 个数据集中，有 5 个数据集上 SFE-PSO 使用 SVM 分类器得到的均值分类准确率高于 HFS-C-P 算法。两种算法在其中两个数据集上的结果相近，只有在 Leukemia 3数据集上，HFS-C-P 算法的表现优于 SFE-PSO 算法。Friedman 检验结果也证明了 SFE-PSO 算法优于 HFS-C-P 算法。相反，结果表明，两种算法使用 KNN 分类器时的性能相近。

表 III SFE、SFE-PSO 和 HFS-C-P 的结果对比 [31]

No.	数据集	算法	KNN 规模	KNN 均值±标准差	KNN 最佳值	SVM 规模	SVM 均值±标准差	SVM 最佳值
1	Yale	HFS-C-P	–	71 34 .	HFS-C-P	357 4	79 52 . .	
		SFE	76	69 2±1 7 . . .	39 396 . SFE	76	75 1±1 0 . . .	28 8521.
		SFE-PSO	64.0			**76.38**	**82 89**	**80 4±1 9** . . .
2	warpAR10P	HFS-C-P	–			**198 40**	– .	**91.1±2.7**
		SFE	**33.4**			100 323 10.		**100±0.0**
		SFE-PSO	34.7			100 470 23.		**100±0.0**
	Colon 3	HFS-C-P	–	**94 77** .	HFS-C-P	155	–	92.47
		SFE	**96**	93 2±2 4 . . .	66 173 . SFE	100 15.93		**95.6±2.0**
		SFE-PSO	**96**	93 8±2 0 . . .	66 251 . SFE-PSO	19 70	**100**	95 3±3 0 . . .
4	小圆蓝细胞肿瘤	HFS-C-P	–	–	HFS-C-P	**100**	99 4	– .
		SFE	100 37 8 .		SFE	98.8±1.0	100 45 32.	**100±0.0**
		SFE-PSO	55 1	**100**	99 1±0 8 . . .	SFE-PSO	**36 19**	**100**
5	弥漫性大B细胞淋巴瘤	HFS-C-P	–	–	HFS-C-P	92.61	149 86	–
		SFE	100 28 6 .		SFE	98.4±1.2	**99**	**97 0±1 4** . . .
		SFE-PSO	38 2	**100**	**99 1±1 0** . . .	SFE-PSO	**23.7**	**99 00**
6	中枢神经系统	HFS-C-P	–	–	82 54 .	HFS-C-P	367 6	–
		SFE	**96**	**87 8±4 2** . . .	66 370 . SFE		98	93 5±4 2 . . .
		SFE-PSO	91	86 7±4 2 . . .	66 578 . SFE-PSO		69 20	**100**
3	白血病 3 7	HFS-C-P	–	–	**100**	152.7	–	**100**
		SFE		100 68.8	98.2±1.4		100 49.66	95.1±3.1
		SFE-PSO	99 36	**100**	98 4±1 3 . . .	SFE-PSO	84 13	**100**
8	卵巢癌症	HFS-C-P	–	–	**99.68**	42.50	–	**100**
		SFE		100 38.66	99.1±0.7		100 51 03.	**100±0.0**
		SFE-PSO	100 39 22 .		SFE-PSO	99.6±0.8	100 41.02	**100±0.0**
	Friedman 检验（平均秩）	HFS-C-P	–	–	HFS-C-P	1.88	2 63	– .
		SFE		–	–	2.38	1 75	– ◦
		SFE-PSO	–	–	SFE-PSO	**1.75**	**1 63**	– ◦

如表 III 所示，使用支持向量机分类器时，SFE-PSO 选定特征的数量少于 HFS-C-P 算法。Friedman 检验结果也证明，SFE-PSO 算法优于 HFS-C-P 算法。值得注意的是，HFS-C-P [31] 未报告使用支持向量机和 KNN 分类器时的最佳分类准确率，也未报告使用 KNN 分类器时选定特征的数量。

VI. 进一步分析

为进一步解释参数 UR 的作用，以及非选择算子和选择算子所选特征的数量，以及在混合 SFE-ECs 算法中使用进化计算方法的时机对所提算法的影响，我们进行了多项实验。本节讨论非选择算子的影响，以及使用进化计算方法继续搜索过程的合适时机。

A. 非选择算子的影响 使用适当的值来确定

UN 非选择算子参数可以显著提高 SFE算法的效率。一般而言，搜索过程开始时，UN 的值应较大；搜索过程结束时，该值应较小，以便算法在特征选择过程的早期进入探索阶段，搜索整个搜索空间，并在特征选择过程的最后阶段进入开发阶段。为实现这一目标，SFE算法使用公式 (1)来确定非选择算子选定特征的数量。

UR 参数从 0.3 线性递减至 0.001，从而实现了预期目标，即在搜索过程的早期阶段进行探索，并在搜索过程的最后阶段进行开发。然而，在所有数据集上使用这种方法，可能无法在探索能力与开发能力之间取得适当的平衡。此外，不同数据集需要不同程度的探索与开发，才能搜索各自的搜索空间。除了UR 值之外，使用其他方法确定UN 值也可能对不同数据集有效。为此，除公式 (1)外，还采用了另一种确定UN 的方法，如公式 (4)所示：

$$UN = \left\lfloor \frac{Rand \times nvar}{k} \right\rfloor \tag{4}$$

其中，$Rand$ 是一个在 0 到 1 之间服从均匀分布的随机数，k 是从范围 $[1, N]$ 中随机选取的整数。本节在五个数据集上进行了实验，以评估公式 (4) 在确定 UN 值方面的效果。图 8 展示了 SFE算法的收敛图，分别采用本节提出的方法和先前实验中使用的方法来确定 UN 值。在本节进行的分析中，N 为 20。因此，每当对 X 应用非选择算子时，都会为 k 生成一个 1 到 20 之间的随机数。如图 8 所示，为确定非选择算子的 UN 而提出的公式提升了 SFE算法的效率。比较 SFE算法在使用公式 1 和公式 4 得到的两种不同 UN 值时的表现可知，SFE算法不会陷入局部最优点，并且能够收敛到更优的解。

表 IV 确定 UN 的两种方法的比较结果

数据集编号		算法	Size	最佳值 最差值	均值±标准差
1	耶鲁	SFE（式 1）	**50.27**	67.87	72 12 . . . 70 41± 1 20. . .
		SFE（式 4）	134.90	**69.09**	**75 75** . . . **73 31± 1 52**. . .
	淋巴瘤 2	SFE（式 1）	**84.63**	95.84	98 94 . . . 97 28±0 95. . .
		SFE（公式 4）	206.83	**97.94**	**100** . . . **98.67±0.77**
	TOX_171 3	SFE（公式 1）	**106.1**	94.13	99 42 . . . 97 44±1 50. . .
		SFE（公式 4）	9650 205 43 . .		**100** . . . **98.55±0.95**
4	儿童期 ALL	SFE（公式 1）	35 1223		80 . . . 50.66±17.1
		SFE（公式 4）	61.66 34.76		**85** . . . **73.83±5.70**
	癌 5	SFE（公式 1）	8616 147 81 . .		**94 84** . . . 90 86 ±2 25. . .
		SFE（公式 4）	8904 422 16 . .		94 26 . . . **91 73± 1 34**. . .

表 IV 展示了采用两种方法时，SFE算法在不同评估标准下的性能。比较结果表明，本节提出的方法对提升

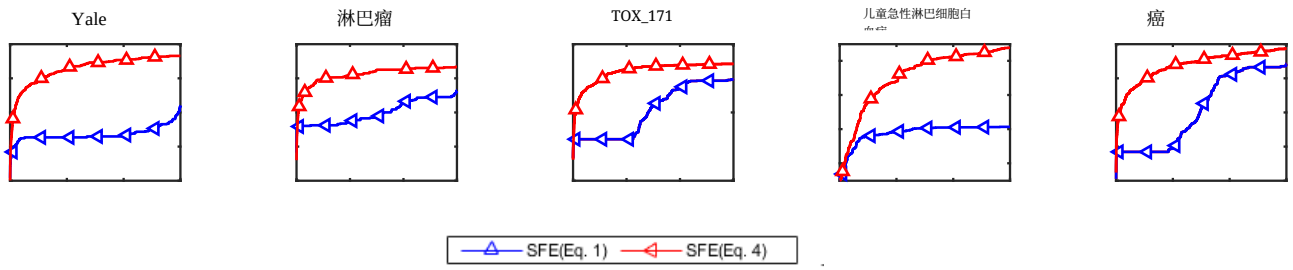

图 8. 两种不同方法确定非选择算子中 UN 对 SFE算法收敛性的影响。

SFE算法效率的作用。例如，在儿童急性淋巴细胞白血病数据集中，SFE算法使该算法的平均分类准确率提高了 23% 以上。因此，这种方法可以显著提升 SFE算法在某些数据集上的性能。

B. 在 SFE-PSO 中使用进化计算方法的时间效应

确定使用进化计算方法在新的子空间中进行搜索的适当条件，并继续搜索过程以获得更优的解，这对于提高算法3所提出框架的效率至关重要。如果 SFE算法未能充分缩小搜索空间的规模，进化计算方法可能会因新搜索空间过大而无法进行搜索。例如，如果某个数据集有 20,000 个特征，而 SFE算法选出了 4,000 个特征，那么搜索空间仍然很大，进化计算方法可能无法得到理想结果。另一方面，SFE算法在搜索并到达这个缩减后的搜索空间时，可能会从数据集中移除一些重要特征，因此进化计算方法在新的子空间中可能无法得到预期结果。例如，对 SFE算法在 Colon 数据集上的结果进行分析表明，该算法只用不到 15 个特征就能达到超过 90% 的分类准确率。因此，如果 SFE算法未能改进已找到的解，使用进化计算方法可能并无帮助。因此，确定一个合适的条件来结束 SFE 搜索过程，随后再使用进化计算方法，对于提高框架3的效率具有重要作用。

C. 计算复杂度

PSO 和二进制蝴蝶优化（蝴蝶优化算法）等进化计算算法方法的计算复杂度为 [45] (it_{max} (D × NP + C × NP))，其中it_{max} 表示最大迭代次数，D 表示数据集维度（即特征数量），C 表示适应度函数的计算成本，NP 表示种群规模。然而，遗传算法和灰狼优化等其他进化计算算法方法的计算复杂度在最佳情况下为 O it_{max} × D × NP +((())，在 $C \times NP + NP\ log\ NP$ O 最差情况下为(it_{max} ×()。这意味着，由于遗传算法和灰狼优化每次迭代都需要对解进行排序，其计算复杂度高于 PSO 和蝴蝶优化算法[45]。 $D \times NP + C \times NP + NP^2$

SFE算法的计算复杂度为O (it_{max} (D + C))。这一时间复杂度优于所有EC方法。此外，在搜索过程的早期阶段不选择大量特征，可降低适应度函数的计算成本，并且算法的计算成本也会显著降低。

SFE-PSO算法的计算复杂度为O (SFE_{it} (D + C)+ PSO_{it} (D × NP + C × NP))，其中SFE_{it} 是SFE算法执行的迭代次数，PSO_{it} 是PSO算法执行的迭代次数。我们还应注意，$it_{max} = SFE_{it} + PSO_{it}$。这一复杂度意味着，在为SFE-PSO算法搜索过程设定的最大适应度函数评估次数（迭代次数）中，SFE_{it} 次迭代由SFE算法执行，PSO_{it} 次迭代由PSO算法执行。关于SFE-PSO算法复杂度，一个值得注意之处是，SFE算法会在搜索过程的第一阶段从解中取消选择无关特征。因此，PSO算法继续搜索时，由于搜索空间较小，与在原始搜索空间中搜索相比，搜索的计算成本显著降低。因此，SFE-PSO算法的计算成本通常显著低于所有进化计算算法。补充材料中的表 II 和表 VI 展示了实验中的运行时间，清楚地体现了这一点。

VII. 结论

本文提出了一种新颖的[illegible]算法，用于处理高维数据。SFE算法旨在以最少的特征实现最高的分类准确率，同时将高维数据的计算成本和内存消耗降至最低。实验结果表明，与其他针对相同目的提出的算法相比，所提出的 SFE 算法具有明显优势。尽管 SFE 算法在实现高维数据的特征选择目标方面效率很高，但在一些数据集中，SFE 算法在降低数据维度后可能会停留在局部最优点。为解决这一问题，本文提出了一种结合 SFE 算法和 PSO 算法的混合算法。SFE-PSO 混合算法的结果表明，该方法比 SFE 算法更有效。因此，可以将 SFE 算法与其他进化计算方法结合，以获得更高的性能。

尽管所提出的方法具有有效性和较高的效率，但这两种算法都存在可在未来改进的不足。通过算法的行为分析（见图6）以及 ORL、Yala、Carcinom 和 TOX_171 等一些数据集发现，SFE算法的一个不足是其UN 和 SN 参数不适用于这些数据集。为克服这一不足，可以采用一些机制来设计具有自组织参数的 SFE

算法。此外，本文提出的非选择算子和选择算子也可以进一步改进。虽然结果表明 SFE-PSO算法性能优异，但错误地移除相关特征（在物理移除阶段）可能导致性能不佳，并使算法陷入局部最优点。此外，我们可以不实际移除特征，而是在 SFE算法搜索过程的不同阶段，使用 PSO算法在选定特征中执行局部搜索。此外，为了提高 SFE-PSO算法的有效性和效率，我们可以不使用 PSO算法，而是采用该算法的改进版本 [36] 以及其他进化计算方法，包括[37]-[42] 中提出的方法，这些方法能显著提高低维数据集中特征选择的效率。

参考文献

[1] J. Cai, J. Luo, S. Wang, and S. Yang, “机器学习中的特征选择：一种新视角”，《神经计算》，第300卷，第70–79页，2018年。[2] J. Li and H. Liu, “大数据分析中的特征选择挑战”，IEEE Intell. Syst.，第32卷，第2期，第9–15页，2017年。[3] J. Li 等，“特征选择：数据视角”，ACM Comput. Surv.，第50卷，第6期，第1–45页，2018年。[4] H. Huang and H. Liu, “利用标签的联合语义信息和结构信息进行层次分类的特征选择”，Knowl. Based Syst.，第195卷，第105655期，第105655页，2020年。[5] P. Tan, X. Wang, and Y. Wang, “面向运动想象脑机接口的进化算法特征选择中的降维”，Swarm Evol. Comput.，第52卷，第100597期，第100597页，2020年。[6] M. Robnik- ˇ Sikonja and I. Kononenko, “ReliefF 和 RReliefF 的理论与实证分析”，Machine learning，第53卷，第1-2期，第23–69页，2003年。[7] Wang, X. Li, and R. Ruiz, “最大相关性与最小监督冗余的特征选择”，IEEE 控制论汇刊，第1–11页，2022年。[8] M. A. Hall, “面向离散类与数值类机器学习的基于相关性的特征选择”，researchcommons.waikato.ac.nz，2000年5月1日。[9] B. H. Nguyen, B. Xue, and M. Zhang, “数据挖掘中特征选择的群体智能方法综述”，Swarm Evol. Comput.，第54卷，第100663期，第100663页，2020年。[10] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “进化计算特征选择方法综述”，IEEE 进化计算汇刊，第20卷，第4期，第606–626页，2016年。[11] M. Abdel-Basset, D. El-Shahat, I. El-henawy, V. H. C. de Albuquerque, and S. Mirjalili, “灰狼优化算法与两阶段变异相结合的新型特征选择方法”，《专家系统及其应用》，第139卷，第112824期，第112824页，2020年。[12] D. P. Muni, N. R. Pal, and J. Das, “用于同时进行特征选择和分类器设计的遗传编程”，IEEE Trans. Syst. Man Cybern. B Cybern.，第36卷，第1期，第106–117页，2006年。[13] J. Tang, S. Alelyani, and H. Liu, “分类的特征选择：综述”，《数据分类：算法与应用》。编辑：Charu Aggarwal，CRC Press，2013年。[14] J. R. Quinlan, “决策树的归纳”，Machine Learning，第1卷，第1期，第81–106页，1986年3月。[15] S. Wang, J. Tang, and H. Liu, “嵌入式无监督特征选择”，AAAI人工智能会议论文集，第29卷，第1期，2015年。[16] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, “有监督、无监督和半监督特征选择：基因选择综述”，IEEE/ACM Trans. Comput. Biol. Bioinform.，第13卷，第5期，第971–989页，2016年。[17] Z. Zhu, Y.-S. Ong, and M. Dash, “基于模因框架的封装器-过滤器特征选择算法”，IEEE Trans. Syst. Man Cybern. B Cybern.，第37卷，第1期，第70–76页，2007年。[18] Y. Xue, B. Xue, and M. Zhang, “用于分类大规模特征选择的自适应粒子群优化”，ACM Trans. Knowl. Discov. Data，第13卷，第5期，第1–27页，2019年。[19] L. Wang, Y. Wang, and Q. Chang, “大数据生物信息学的特征选择方法：搜索视角综述”，Methods，第111卷，第21–31页，2016年。[20] Y. Xue, T. Tang, W. Pang, and A. X. Liu, “用于多分类器大规模特征选择问题的自适应参数和策略粒子群优化”，《应用软计算》，第88卷，第106031页，2020年3月。[21] A. W. Whitney, “一种非参数测量选择的直接方法”，IEEE Trans. Comput.，第C–20卷，第9期，第1100–1103页，1971年。

[22] T. Marill 和 D. Green，“论受体在识别系统中的有效性”，IEEE 信息论汇刊，第 9 卷，第 1 期，第 11–17 页，1963年。[23] S. D. Stearns，“关于为模式分类器选择特征”，载于《模式识别国际会议论文集》，美国加利福尼亚州科罗纳多，第 71–75 页，1976年。[24] P. Pudil、J. Novovičová 和 J. Kittler，“特征选择中的浮动搜索方法”，《模式识别快报》，第 15 卷，第 11 期，第 1119–1125 页，1994年。[25] J. Hu 等，“针对灰狼优化器缺陷的正交学习协方差矩阵：见解、平衡、多样性与特征选择”，Knowl. Based Syst.，第 213 卷，第 106684 号，第 106684 页，2021年。[26] M. Mafarja 等，“用于特征选择问题的进化种群动力学与蚱蜢优化方法”，Knowl. Based Syst.，第 145 卷，第 25–45 页，2018 年。[27] K. Chen 、 B. Xue 、 M. Zhang 和 F. Zhou，“ 一种基于进化多任务的高维分类特征选择方法”，IEEE 控制论汇刊，第 PP 卷，第 1–15 页，2021年。[28] B. Tran、B. Xue 和 M. Zhang，“用于高维分类特征选择的变长粒子群优化”，IEEE 进化计算汇刊，第 23 卷，第 3 期，第 473–487 页，2019年。[29] X.-F. Song、Y. Zhang、Y.-N. Guo、X.-Y. Sun 和 Y.-L. Wang，“用于高维数据特征选择的可变规模协同协同进化粒子群优化”，IEEE 进化计算汇刊，第 24 卷，895，2020 5，882第。页 – 。[30] K. Chen、B. Xue、M. Zhang 和 F. Zhou，“通过粒子群优化实现高维分类特征选择的进化多任务方法”，IEEE Transactions on Evolutionary Computation，第 1–1 页，2021年。[31] X.-F. Song、Y. Zhang、D.-W. Gong 和 X.-Z. Gao，“一种基于相关性引导聚类和粒子群优化的高维数据快速混合特征选择方法”，IEEE 控制论汇刊，第 PP 卷，第 1–14 页，2021年。[32] X. Wang、Y. Wang、K.-C. Wong 和 X. Li，“一种用于大规模特征选择的自适应加权差分进化方法”，Knowledge-Based Systems，第 235 卷，第 107633 页，2022年1月。[33] F. Cheng、F. Chu、Y. Xu 和 L. Zhang，“一种基于导向矩阵的高维特征选择多目标进化算法”，IEEE Transactions on Cybernetics，第 52 卷，第 9 期，第 9695–9708 页，2022年9月。[34] J. Hu 等，“分散觅食黏菌算法：用于全局优化和基于封装器的特征选择的连续与二进制变体”，Knowl. Based Syst.，第 237 卷，第 107761 号，第 107761 页，2022 年。[35] F. Kılıç 、 Y. Kaya 和 S. Yildirim，“一种新颖的基于多种群的粒子群优化特征选择方法”，Knowledge-Based Systems，第 219 卷，第 106894 页，2021年5月。[36] X. Song、Y. Zhang、D. Gong 和 X. Sun，“利用结合互信息的裸骨粒子群优化进行特征选择”，《模式识别》，第 112 卷，第 107804 页，2021年4月。[37] Y. Zhang、D.-W. Gong、X.-Z. Gao、T. Tian 和 X.-Y. Sun，“用于多目标特征选择的自学习二进制差分进化”，《信息科学》，第 507 卷，第 67–85 页，2020年。[38] A. K. Das、S. Das 和 A. Ghosh，“使用双目标遗传算法的集成特征选择”，Knowl. Based Syst.，第 123 卷，第 116–127 页，2017年。[39] P. Hu、J.-S. Pan 和 S.-C. Chu，“改进的二进制灰狼优化器及其在特征选择中的应用”，Knowl. Based Syst.，第 195 卷，第 105746 号，第 105746 页，2020年。[40] M. Tubishat、N. Idris、L. Shuib、M. A. M. Abushariah 和 S. Mirjalili，“基于反向学习和新型局部搜索算法的改进樽海鞘群算法及其在特征选择中的应用”，《专家系统及其应用》，第 145 卷，第 113122 号，第 113122 页，2020年。[41] Z. Wang、S. Gao、M. Zhou、S. Sato、J. Cheng 和 J. Wang，“用于分类中多目标特征选择的基于信息论的非支配排序蚁群优化”，IEEE Transactions on Cybernetics，第 1–14 页，2022年。[42] X.-H. Wang、Y. Zhang、X.-Y. Sun、Y.-L. Wang 和 C.-H. Du，“基于人工蜂群的多目标特征选择：一种采用可变样本规模的加速方法”，《应用软计算》，第 88 卷，第 106041 号，第 106041 页，2020年。[43] L. Wang、X. Fu、Y. Mao、M. Ilyas Menhas 和 M. Fei，“一种新型改进的二进制差分进化算法及其应用”，《神经计算》，第 98 卷，第 55–75 页，2012年。[44] J. Kennedy 和 R. C. Eberhart，“粒子群算法的离散二进制版本”，载于《1997年IEEE系统、人类与控制论国际会议：计算控制论与仿真》，2002年。[45] S. Arora 和 P. Anand，“用于特征选择的二进制蝴蝶优化方法”，《专家系统及其应用》，第 116 卷，第 147–160 页，2019年。



Behrouz Ahadzadeh 于2014年获得伊朗加兹温伊斯兰阿扎德大学加兹温分校电气、计算机与信息技术工程系人工智能与机器人工程理学硕士学位。他的研究兴趣包括计算智能、机器学习、计算机视觉、知识发现与数据挖掘。



Moloud Abdar 于2015年获得伊朗达姆甘大学计算机工程学士学位，2018年获得日本会津会津大学计算机科学硕士学位，并于2022年获得澳大利亚迪肯大学智能系统研究与创新研究所（IISRI）工程学（计算机、机器学习）博士学位。他目前是澳大利亚维多利亚州迪肯大学 IISRI 的副研究员。其研究兴趣包括机器学习、深度学习、计算机视觉、医学图像分析、情感分析、可解释性和不确定性量化。



Fatemeh Safara 获得伊斯兰阿扎德大学应用数学理学学士（荣誉）学位，以及德黑兰塔比阿特莫达勒斯大学软件工程理学硕士学位。她于2014年获得马来西亚博特拉大学（UPM）人工智能博士学位。她2000年至2010年，他在伊朗电信研究中心信息技术部担任研究员，从事图像挖掘和生物特征信号研究。2010年，他加入伊斯兰阿扎德大学伊斯兰城分校，目前任计算机工程学院助理教授。他目前的研究兴趣包括机器学习、数据挖掘、深度学习、信号处理，尤其是生物信号处理、医学图像分析和医疗物联网。



Abbas Khosravi（IEEE高级会员）于2010年获得澳大利亚维多利亚州沃恩庞兹迪肯大学的机器学习博士学位。目前，他是迪肯大学智能系统研究与创新研究所的副教授。他的广泛研究兴趣包括人工智能、深度学习和不确定性量化。他目前正在研究并应用概率深度学习理念以及医疗保健和工程领域中的不确定性感知解决方案。



Mohammad Bagher Menhaj于1992年获得俄克拉荷马州立大学（OSU）斯蒂尔沃特校区的博士学位。随后，他成为OSU的博士后研究员。1993年，他加入伊朗德黑兰的阿米尔卡比尔理工大学。2000年12月至2003年8月期间，他任职于OSU电气与计算机工程学院计算机科学系，担任访问教员兼研究学者。他发表了500余篇论文。其研究兴趣包括计算智能、学习自动机、电力系统、图像处理、模式识别、通信、认知科学和知识发现。



Ponnuthurai Nagaratnam Suganthan于1990年、1992年和1994年分别获得英国剑桥大学电气与信息工程专业的学士学位和硕士学位。他于2020年获得斯洛文尼亚马里博尔大学授予的荣誉博士（即Doctor Honoris Causa）学位。在此之后1995年完成博士研究后，他于1995—1996年在悉尼大学电气工程系担任博士前研究助理，并于1996—1999年在昆士兰大学计算机科学与电气工程系担任讲师。自2022年8月起，他在卡塔尔大学KINDI计算研究中心担任研究教授。他曾任麻省理工学院出版社《进化计算》期刊编委（2013—2018年）。他曾任或现任《应用软计算》（爱思唯尔，2018年至今）、《神经计算》（爱思唯尔，2018年至今）、《IEEE控制论汇刊》（2012—2018年）、《IEEE进化计算汇刊》（2005—2021年）、《信息科学》（爱思唯尔，2009年至今）、《模式识别》（爱思唯尔，2001年至今）和《IEEE系统、人及控制论汇刊：系统》（2020年至今）的副编辑。他是《群体与进化计算》（2010年至今）的创刊联合主编，该刊是SCI收录的爱思唯尔期刊。他与他人合著的SaDE论文（发表于2009年4月）于2012年获得“IEEE进化计算汇刊杰出论文奖”。他的研究兴趣包括基于随机化的学习方法、群体与进化算法、模式识别、深度学习，以及群体算法、进化算法和机器学习算法的应用。2015年至2022年，他每年都入选汤森路透科学引文评选的计算机科学领域高被引研究人员。他曾担任IEEE SSCI 2013大会主席。自1991年以来，他一直是IEEE会员（S'91、M'92、SM'00，2015年当选会士），并于2014—2016年当选IEEE计算智能学会（CIS）行政委员会成员。他于2018—2021年担任IEEE CIS杰出演讲人（DLP）。

《SFE：一种用于高维数据的简单、快速且高效的特征选择算法》补充材料

Behrouz Ahadzadeh、Moloud Abdar、Fatemeh Safara、Abbas Khosravi（IEEE高级会员）、Mohammad Bagher Menhaj、Ponnuthurai Nagaratnam Suganthan（IEEE会士）

I. 实验设计

A. 数据集 为了

评估所提出的特征选择方法的有效性和效率，我们在表 II 所示的 40 个真实数据集上进行了实验。这些数据集可从以下网址下载：https://data.mendeley.com/datasets/fhx5zgx 2zj/1、https://ckzixf.github.io/dataset.html、http://csse.szu.edu.cn/staff/zhuzx/datasets.html、https://jundongl.github.io/scikit-feature/datasets.html、https://figshare.com/articles/dataset/Microarray_data_rar/7345880/2、https://file.biolab.si/biolab/supp/bicancer/projections/info/gastricGSE2685.html。

B. 比较算法和参数设置 我们采用了九种算法进行比较，

所提出的特征选择方法。首先，将 SFE 和 SFE-PSO 算法的结果与六种算法 BDE [43], BPSO [44],BGA [5], 以及三种最先进的基于进化计算的特征选择算法（即 DimReM-BDE [5], DimReM-BGA [5],和 DimReM-PSO [5]）进行比较。表 I 列出了这些算法的参数设置详情。所有实验均使用表 II 中相同的 40个数据集。随后，将所提算法的结果与三种最先进的基于进化计算的特征选择算法（即 HFS-C-P [31], SaWDE[32], 和 SM-MOEA [33] 算法）进行比较。

表 I 参数设置

算法	**参数值**
BDE [43] 和 DimReM-BDE [5]	F = 0.8; CR = 0.2; b = 20
BPSO [44] 和 DimReM-BPSO [5]	w = 1; c1 = 2; c2 = 1.5
遗传算法 [5] 和 DimReM-GA [5]	pc = 0.6; pm = 0.03
SFE 和 SFE-PSO	UR_{max} = 0.3; UR_{mIN} = 0.001 , SN = 1

选择 BDE [43],BPSO [44], 和 BGA [5], 这三种算法，是为了评估这些作为特征选择领域知名进化计算算法的算法在高维数据上的性能。此外，鉴于 DimReM-BDE [5], DimReM-BGA [5]，和 DimReM-PSO [5], 在缩减高维数据规模方面效率很高，且其源代码可用，因此也将这些算法与所提出的方法进行比较。选择 HFS-C-P [31], SaWDE [32], 和 SM-MOEA [33] 这三种算法，是因为它们是近期针对高维数据集中的特征选择提出的。在实验的第一部分，进行了 6000 次适应度评估。此外，六种算法的种群规模均为 20。在特征选择过程中，使用 K=1 近邻（KNN）作为分类器，对算法找到的解进行评估。随后，采用 5折交叉验证方法将数据集划分为训练集和测试集。每种算法均独立运行 30 次。然后，计算并比较各算法的平均分类准确率、平均选定特征数和运行时间。本文所有实验均在配备英特尔 (R) core (TM) i5-5200 U 中央处理器、主频 2.2 GHz、12.00 GB 随机存取存储器的设备上进行，并使用 MATLAB R2019a 实现。本节评估 SFE 和 SFE-PSO算法的性能，并将其与 BDE、BPSO、BGA、DimReM-BDE、DimReM-BGA 和 DimReM-BPSO 的性能进行比较。

表 II 40个数据集的基本信息

数据集 编号。	#实例数	#特征数	领域	#类别
ORL 1	400	1024	40	人脸图像
耶鲁 2	165	1024	15	人脸图像
Colon 3	62	2000	2	生物学
小圆蓝细胞肿瘤 4	83	2308	4	生物学
warpAR10P 5	130	2400	10	人脸图像
肺癌1 6	203	3312	5	生物学
淋巴瘤 7	96	4026	9	生物学
胶质瘤 8	50	4434	4	生物学
白血病 1 9	72	5327	3	生物学
弥漫性大B细胞淋巴瘤 10	77	5469	2	生物学
膀胱癌 11	40	5724	3	生物学
9种肿瘤 12	60	5726	9	生物学
TOX_171 13	171	5748	4	生物学
脑肿瘤1 14	90	5920	5	生物学
前列腺_GE 15	102	5966	2	生物学
白血病 16	72	7070	2	生物学
ALLAML 17	72	7129	2	生物学
中枢神经系统 18	60	7129	2	生物学
ALL-AML-3 19	72	7129	3	生物学
ALL-AML-4 20	72	7129	4	生物学
白血病 2 21	72	7129	4	生物学
儿童急性淋巴细胞白血病 22	60	8280	4	生物学
癌 23	174	9182	11	生物学
nci9 24	60	9712	9	生物学
arcene 25	200	10000	2	生物学
orlraws10P 26	100	10304	10	人脸图像
脑肿瘤2 27	50	10367	4	生物学
白血病 28	72	11225	3	生物学
慢性淋巴细胞白血病 SUB_111 29	111	11340	3	生物学
11种肿瘤 30	174	12533	11	生物学
混合谱系白血病 31	72	12582	3	生物学
肺癌2 32	203	12600	5	生物学
CML治疗 33	28	12625	2	生物学
胶质母细胞瘤 34	50	12625	4	生物学
急性髓系白血病预后 35	54	12625	2	生物学
前列腺癌 36	20	12627	2	生物学
卵巢癌 37	253	15154	2	生物学
SMK_CAN_187 38	187	19993	2	生物学
GLI_85 39	85	22283	2	生物学
乳腺癌 40	97	24481	2	生物学

表 III 八种算法在 30 次独立运行中获得的分类准确率的最差值、最佳值、均值和标准差

数据集编号	指标	BDE [43]	遗传算法 [5]	BPSO [44]	DimReM-BDE [5]	DimReM-BGA [5]	DimReM-BPSO [5]	SFE	SFE-PSO
ORL 1	最差值	96.2500	96.0000	97.0000	96.5000	96.2500	96.5000	94.5000	96.2500
	Best	96.7500	97.0000	97.7500	97.2500	97.0000	97.5000	97.0000	98.2500
	Mean	96.8600(+)	97.3500(≈)	96.5500(+)	96.4900(+)	96.7396(+)	97.0500(≈)	95 7167 +	97 2417. . 0
	Std	0.18370	0.27000	0.2282	0.1920	0.2270	0.3146	0.6612	0.5704
耶鲁 2	最差值	73.3333	72.1212	74.5455	73.9394	73.9394	73.9394	67.8788	71.5152
	Best	75.1515	75.1515	75.7576	75.1515	75.7576	75.7576	72.1212	77.5758
	Mean	74.9587(≈)	74.5730(≈)	74.8209(≈)	75.1515(≈)	74.0220(+)	73.8292(+)	70 4132 +	**75 2323**. . 0
	Std	0.40270	0.73150	0.2645	0.3611	0.5114	0.3917	1.2069	1.3574
Colon 3	最差值	77.3077	78.8462	80.5128	78.8462	78.8462	78.8462	90.5128	91.9231
	Best	82.1795	82.1795	85.5128	82.1795	85.5128	85.5128	100.000	100.000
	Mean	79.9060(+)	82.8462(+)	80.4615(+)	79.3034(+)	81.2393(+)	81.3462(+)	96 4402 ≈	**96 5128**. . 0
	Std	0.91200	1.19160	1.2066	0.9913	1.5904	1.5538	2.4443	2.4659
小圆蓝细胞肿瘤 4	最差值	90.5147	90.5147	94.0441	91.6912	90.5147	91.6912	92.7941	96.4706
	Best	94.1176	96.4706	96.0319	94.1176	96.4706	96.4706	100.000	100.000
	Mean	92.6618(+)	97.6471(+)	93.0490(+)	91.9461(+)	93.8480(+)	94.2721(+)	98 8971 ≈	**99 0931**. . 0
	Std	0.7286	1.42760	0.7979	0.6463	1.6069	1.2075	1.9921	1.1494
warpAR10P 5	最差值	59.2308	59.2308	63.0769	60.7692	60.0000	61.5385	83.0729	92.3077
	Best	62.3077	61.4359	65.3077	63.0769	62.0769	66.9231	99.2308	99.2308
	Mean	61.8974(+)	68.4618(+)	63.8462(+)	60.5641(+)	65.3846(+)	63.8718(+)	93 6410 +	**95 5385**. . 0
	Std	0.72650	1.10150	1.2510	0.6920	1.1805	1.2853	3.3919	2.2979
肺癌1 6	最差值	96.5732	96.0854	96.5732	96.5732	96.0854	96.0854	96.0854	97.5610
	Best	97.0732	97.5610	97.5610	97.0732	97.5610	97.5610	98.5366	99.0244
	Mean	96.8199(+)	97.2472(+)	96.7374(+)	96.7069(+)	96.8049(+)	96.9358(+)	97 4122 ≈	**98 3252**. . 0
	Std	0.2219	0.29770	0.3309	0.2485	0.3826	0.3649	0.6435	0.3775
淋巴瘤 7	最差值	96.8947	96.8947	96.8947	96.8947	96.8947	96.8947	95.8421	96.8947
	Best	97.9474	97.9474	99.0000	97.9474	97.9474	97.9474	98.9474	100.000
	Mean	97.1175(+)	97.4211(+)	97.0351(+)	97.0088(+)	97.1421(+)	97.2825(+)	97 2860 +	**99 3526**. . 0
	Std	0.31880	0.36390	0.6025	0.4226	0.4519	0.5147	0.9554	0.9532
胶质瘤 8	最差值	84.0000	84.0000	84.0000	84.0000	84.0000	82.0000	90.0000	92.0000
	Best	88.0000	90.0000	92.0000	86.0000	90.0000	90.0000	100.000	100.000
	Mean	84.6667(+)	89.1333(+)	87.7333(+)	84.5333(+)	87.6667(+)	87.5333(+)	**97 0000** ≈	96 1333. . 0
	Std	1.04170	1.55220	1.7167	0.9589	1.6678	2.0126	2.6130	2.0965
白血病 1 9	最差值	93.1429	93.0476	95.9048	93.1429	93.0476	94.4762	94.5714	97.2381
	Best	95.9048	95.9048	97.3333	95.9048	97.3333	97.3333	100.000	100.000
	Mean	94.2508(+)	96.5746(+)	94.6000(+)	94.0667(+)	94.7429(+)	95.4952(+)	98 2825 ≈	**98 8794**. . 0
	Std	0.82960	1.08990	0.6815	0.8394	0.9781	0.9748	1.3848	0.7234
弥漫性大B细胞淋巴瘤 10	最差值	89.7500	88.5000	89.7500	88.5000	89.7500	88.5000	95.0000	96.2500
	Best	89.8333	92.3333	92.3333	91.0000	92.3333	92.2500	100.000	100.000
	Mean	89.9194(+)	91.0500(+)	90.4667(+)	89.7639(+)	90.5083(+)	90.1722(+)	99 1611 ≈	**99 3639**. . 0
	Std	0.0316	0.8629	0.7733	0.6337	0.7828	1.0079	1.1075	1.0393
膀胱癌 11	最差值	75.0000	75.0000	75.0000	75.0000	75.0000	75.0000	92.5000	90.0000
	Best	82.5000	90.0000	90.0000	82.5000	90.0000	90.0000	100.000	100.000
	Mean	78.4000(+)	85.9000(+)	84.0000(+)	78.4000(+)	83.6000(+)	82.8000(+)	**97 5000** ≈	97 3000. . 0
	Std	2.4875	3.9528	4.8348	2.0259	4.6256	5.4639	2.1651	2.6926
9种肿瘤 12	最差值	51.6667	55.0000	56.6667	53.3333	53.3333	53.3333	56.6667	65.0000
	Best	56.6667	58.3333	61.6667	56.6667	58.3333	61.6667	86.6667	85.0000
	Mean	54.1667(+)	59.2222(+)	56.2222(+)	54.2222(+)	56.2778(+)	56.8889(+)	72 7778 +	**74 1667**. . 0
	Std	1.3655	1.3795	1.2935	0.9539	1.4306	1.7904	7.8092	4.9276
TOX_171 13	最差值	95.8992	95.3109	97.6471	95.8992	95.8992	96.4874	94.1345	98.8235
	Best	97.0588	97.6639	98.8235	97.6471	98.8235	98.8235	99.4286	100.000
	Mean	96 7356 +	98 5490. . 0	96.4443(+)	96.5445(+)	96.8717(+)	97.3776(+)	97 4420 +	**99 7849**. . 0
	Std	0.40050	0.5460	0.4296	0.4669	0.6196	0.5248	1.5030	0.3264
脑肿瘤1 14	最差值	90.0000	90.0000	91.1111	90.0000	90.0000	90.0000	90.0000	92.2222
	Best	91.1111	91.1111	91.1111	91.1111	91.1111	91.1111	94.4444	96.6667
	Mean	90.3704(+)	91.1111(+)	90.9630(+)	90.3704(+)	90.8519(+)	90.9259(+)	92 0000 +	**94 0000**. . 0
	Std	0.53270	0.38420	2.89e-14	0.5327	0.4780	0.4212	1.1451	0.9500
前列腺 GE 15	最差值	88.1905	88.1905	89.1429	88.1905	88.1429	88.1905	93.1429	92.0952
	Best	90.0952	90.1429	92.0952	90.1429	90.1429	91.1429	99.0000	99.0000
	Mean	88.8667(+)	90.3079(+)	89.0905(+)	88.9016(+)	88.9270(+)	89.5159(+)	**97 0524** ≈	96 3429. . 0
	Std	0.5000	0.6184	0.6384	0.5184	0.6636	0.7897	1.4538	1.6871
白血病 16	最差值	81.7143	81.7143	84.5714	81.8095	81.7143	81.7143	90.2857	97.2381
	Best	84.5714	86.0000	86.1429	84.5714	87.4286	86.0952	100.000	100.000
	Mean	83.2159(+)	87.4286(+)	84.0063(+)	82.9270(+)	84.2889(+)	84.9111(+)	95 8254 +	**99 0571**. . 0
	Std	0.9066	1.1578	0.6867	0.6793	1.3183	1.1698	2.0384	0.7255
ALLAML 17	最差值	90.1905	90.1905	91.6190	90.1905	88.8571	90.1905	95.7143	95.9048
	Best	91.6190	91.6190	91.6190	91.6190	91.6190	91.6190	100.000	100.000
	Mean	91.1683(+)	91.6190(+)	91.4286(+)	90.9016(+)	91.3397(+)	91.3365(+)	98 9302 ≈	**99 2222**. . 0
	Std	0.6830	0.4939	4.33e-14	0.6488	0.6743	0.5750	1.1447	1.1183
中枢神经系统 18	最差值	70.0000	71.6667	73.3333	71.6667	71.6667	71.6667	85.0000	85.0000
	Best	75.0000	75.0000	80.0000	75.0000	76.6667	78.3333	98.3333	96.6667
	Mean	73.2778(+)	77.0556(+)	74.2222(+)	73.0000(+)	74.2222(+)	74.6111(+)	91 5000 ≈	**92 6111**. . 0
	Std	1.1073	1.0480	2.0846	0.9268	1.2172	1.6772	3.3434	2.7917
ALL-AML-3 19	最差值	87.3333	87.3333	88.7619	87.3333	87.3333	87.3333	93.0476	94.4762
	Best	88.7619	88.7619	90.1905	88.7619	88.7619	90.1905	100	100
	Mean	88.1778(+)	89.0000(+)	88.3397(+)	87.7619(+)	88.3397(+)	88.5778(+)	97 1397 ≈	**97 6857**. . 0
	Std	0.6145	0.6564	0.5415	0.6798	0.6564	0.6075	2.1855	1.5575
ALL-AML-4 20	最差值	84.6667	86.0000	86.0000	84.6667	86.0000	86.0000	88.7619	92.8571
	Best	86.0000	87.4286	87.4286	86.0000	87.4286	87.4286	98.6667	100.000
	Mean	85.6889(+)	86.0952(+)	86.0952(+)	85.6889(+)	86.0476(+)	86.0952(+)	94 4063 +	**96 0095**. . 0
	Std	0.5736	0.3624	0.3624	0.5736	0.2608	0.3624	3.1007	1.7693

表 IV 八种算法在 30 次独立运行中获得的分类准确率的最差值、最佳值、均值和标准差

数据集编号		指标	BDE [43]	BGA [5]	BPSO [44]	DimReM-BDE [5]	DimReM-BGA [5]	DimReM-BPSO [5]	SFE	SFE-PSO
21	白血病 2	最差值	84.6667	84.6667	86.0000	84.6667	86.0000	86.0000	88.8571	91.6190
		Best	86.0952	86.000	87.4286	87.4286	86.0000	87.4286	98.5714	100.000
		Mean	85.8698(+)	86.3810(+)	85.9556(+)	85.8286(+)	86.0000(+)	86.0479(+)	94 8794 ≈	**95 6063**. . 0
		Std	0.4641	0.2434	0.6425	0.5459	0.00000	0.2608	2.3460	1.7084
22	儿童急性淋巴细胞白血病	最差值	35.0000	35.0000	35.0000	35.0000	35.0000	35.0000	35.0000	35.0000
		Best	35.0000	35.0000	35.0000	35.0000	35.0000	35.0000	80.0000	75.0000
		Mean	35.0000(+)	35.0000(+)	35.0000(+)	35.0000(+)	35.0000(+)	35.0000(+)	50 6667 ≈	**51 8000**. . 0
		Std	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	17.1931	15.7301
23	癌	最差值	89.0420	89.0420	90.7563	89.0420	89.0420	89.6134	86.1681	94.6179
		Best	90.7563	91.3445	92.4874	90.7731	91.3277	91.9160	94.8403	100.000
		Mean	90.0863(+)	91.3636(+)	90.1100(+)	90.0817(+)	90.0817(+)	90.5019(+)	90 8625 +	**96 4706**. . 0
		Std	0.3366	0.5962	0.4524	0.4234	0.6017	0.5825	2.2512	0.9897
24	nci9	最差值	56.6667	58.3333	60.0000	56.6667	58.3333	58.3333	63.3333	68.3333
		Best	60.0000	60.1667	65.0000	60.0000	61.6667	61.6667	81.6667	85.0000
		Mean	59.1111(+)	62.3889(+)	61.6667(+)	58.7222(+)	60.1111(+)	60.6111(+)	73 1111 +	**75 5556**. . 0
		Std	0.9472	1.1029	1.1315	0.9522	1.2328	1.1145	4.6265	3.3141
25	arcene	最差值	88.0000	88.5000	90.0000	89.0000	88.5000	90.0000	86.0000	90.0000
		Best	90.5000	93.0000	93.0000	91.5000	92.5000	93.0000	97.5000	97.0000
		Mean	89.9000(+)	91.6957(+)	90.6667(+)	89.3667(+)	90.7167(+)	91.4000(+)	92 0167 +	**93 3667**. . 0
		Std	0.6940	1.2058	0.6698	0.7812	0.8272	0.6747	3.2417	1.7564
26	orlraws10P	最差值	98.0000	98.0000	98.0000	98.0000	98.0000	97.0000	97.0000	99.0000
		Best	98.0000	98.0000	98.0000	99.0000	99.0000	99.0000	100.000	100.000
		Mean	98.0333(+)	98.0000(+)	98.0000(+)	98.0000(+)	98.0667(+)	98.0000(+)	99 4000 ≈	**99 7300**. . 0
		Std	0.00000	0.00000	0.00000	0.1826	0.2537	0.3718	0.7701	0.4498
27	脑肿瘤2	最差值	70.0000	72.0000	74.0000	70.0000	72.0000	70.0000	86.0000	88.0000
		Best	76.0000	76.0000	78.0000	74.0000	76.0000	78.0000	98.0000	100.000
		Mean	72.7333(+)	76.0000(+)	75.0000(+)	72.4667(+)	74.6667(+)	75.2000(+)	92 0000 +	**94 4667**. . 0
		Std	1.4559	1.1447	0.9097	1.1121	1.0933	1.6274	3.3631	2.9094
28	白血病	最差值	88.6667	88.6667	88.6667	88.6667	88.6667	87.3333	95.8095	91.5238
		Best	90.0952	91.5238	91.5238	90.0952	90.0952	91.5238	100.000	100.000
		Mean	88.8724(+)	89.5221(+)	89.4667(+)	88.8229(+)	89.5238(+)	89.4171(+)	99 0095 ≈	**99 0635**. . 0
		Std	0.3858	0.9295	0.9221	0.4631	0.7143	1.0091	1.2965	1.7068
29	CLL SUB_111	最差值	72.0158	71.9367	74.6245	72.0158	71.1067	72.9249	82.8063	83.7945
		Best	74.6640	76.5217	78.3399	74.6640	74.7905	77.4308	94.7036	95.5731
		Mean	73.4862(+)	76.7905(+)	74.4730(+)	73.1963(+)	77.3913(+)	76.0422(+)	885942 908116 . . (–)	
		Std	0.7801	1.2780	1.0127	0.7793	1.4743	1.0635	3.0744	2.9313
30	11种肿瘤	最差值	81.5798	81.5462	85.0084	82.1513	82.7059	82.7227	85.0420	88.4706
		Best	83.8655	85.0252	86.7395	84.4370	85.0084	86.1681	92.5210	93.0924
		Mean	83.3115(+)	85.6749(+)	83.5950(+)	83.0627(+)	83.7703(+)	84.8588(+)	88 5417 +	**90 2919**. . 0
		Std	0.5734	0.9367	0.5281	0.6265	0.7512	0.7343	2.0926	1.2420
31	MLL	最差值	90.1905	90.1905	88.8571	88.8571	90.1905	90.1905	94.2857	95.9048
		Best	91.5238	92.9524	92.9524	91.5238	92.9524	91.6190	100.000	100.000
		Mean	90.4127(+)	91.2159(+)	90.8159(+)	90.4127(+)	91.0413(+)	90.9079(+)	98 5587 ≈	**99 1714**. . 0
		Std	0.5054	0.7708	0.7651	0.6148	0.7520	0.6830	1.5405	1.2771
32	肺癌2	最差值	93.1463	92.6463	93.6463	92.6585	92.6463	93.1463	95.5976	96.0976
		Best	94.1463	94.1341	94.6341	93.6463	94.1463	94.6341	98.5366	99.0244
		Mean	93.3283(+)	94.0951(+)	93.3644(+)	93.4873(+)	93.4049(+)	93.6454(+)	**97 3839** ≈	97 3797. . 0
		Std	0.277	0.3820	0.2812	0.2822	0.3255	0.4070	0.7177	0.7622
33	CML治疗	最差值	56.6667	56.6667	56.6667	56.6667	56.6667	56.6667	78.0000	88.6667
		Best	56.6667	60.6667	60.6667	60.6667	60.6667	60.6667	100.000	100.000
		Mean	56.9600(+)	57.4667(+)	56.9600(+)	56.6667(+)	57.4400(+)	57.4667(+)	89 0133 +	**95 1733**. . 0
		Std	1.45e-14	1.0198	1.6330	1.0198	1.4967	1.6330	5.9446	3.8791
34	胶质母细胞瘤	最差值	66.0000	68.0000	70.0000	66.0000	68.0000	68.0000	85.0000	86.0000
		Best	72.0000	72.0000	72.0000	70.0000	72.0000	72.0000	98.3333	98.0000
		Mean	67.8400(+)	71.6800(+)	70.3200(+)	78.0800(+)	70.6400(+)	70.6400(+)	91 5000 +	**92 9600**. . 0
		Std	1.3515	1.3760	0.7483	0.9866	1.2543	1.3808	3.3434	3.0616
35	急性髓系白血病预后	最差值	66.5455	68.3636	68.3636	66.5455	66.5455	68.3636	82.0000	84.9091
		Best	70.1818	70.1818	72.0000	70.1818	72.0000	72.0000	94.0000	98.1818
		Mean	68.3636(+)	70.2545(+)	69.0909(+)	68.2909(+)	69.3018(+)	69.6000(+)	88 3200 +	**92 5818**. . 0
		Std	0.8266	0.9091	0.8266	1.0497	1.3045	1.0123	3.8592	4.0588
36	前列腺癌	最差值	70.0000	70.0000	65.0000	70.0000	70.0000	65.0000	80.0000	95.0000
		Best	70.0000	75.0000	75.0000	75.0000	75.0000	75.0000	100.000	100.000
		Mean	70.2000(+)	71.4000(+)	71.8000(+)	70.0000(+)	71.6000年(+)	70.2000年(+)	95 2000 +	**99 4000**. . 0
		Std	0.0000	2.4495	2.7080	1.0000	2.3805	2.2730	4.8905	1.6583
37	卵巢癌	最差值	95.2784	95.2784	95.2784	95.2784	95.2784	95.2784	97.6157	98.0314
		Best	95.2784	95.2784	95.2784	95.2784	95.2784	95.2784	100.000	100.000
		Mean	95.2784(+)	95.2784(+)	95.2784(+)	95.2784(+)	95.2784(+)	95.2784(+)	99 1072 ≈	**99 7642**. . 0
		Std	1.46e-14	1.46e-14	1.46e-14	1.46e-14	1.46e-14	1.46e-14	0.7018	0.4208
38	SMK_CAN_187	最差值	68.9758	69.2532	70.0427	69.5021	69.5021	70.0427	79.1750	76.9986
		Best	71.1095	71.1238	71.8625	70.5832	71.1238	72.1764	87.7098	89.3314
		Mean	70.0825(+)	70.6952(+)	70.1952(+)	69.8846(+)	70.5405(+)	71.1651(+)	83 0337 ≈	**83 1318**. . 0
		Std	0.4431	0.5214	0.4695	0.4395	0.5703	0.5333	2.3694	3.1881
39	GLI_85	最差值	90.5882	90.5882	90.5852	90.5882	90.5882	90.5882	94.1176	92.3077
		Best	91.7647	92.9412	92.9412	91.7647	92.9412	92.9412	98.8235	99.2308
		Mean	91.4510(+)	91.6706(+)	91.8588(+)	91.5294(+)	92.2745(+)	91.4824(+)	96 5490 +	**98 1569**. . 0
		Std	0.4786	0.7533	0.6724	0.5291	0.6686	0.7026	1.2714	1.2235
40	乳腺癌	最差值	65.9474	66.2514	69.1053	65.9474	68.0526	67.0000	79.3684	80.5263
		Best	69.1053	69.8564	71.2105	69.1053	70.1579	72.2105	93.8947	94.9474
		Mean	66.8658(+)	70.3053(+)	68.4532(+)	66.8658(+)	68.7816(+)	69.2132(+)	85 7018 +	**86 8351**. . 0
		Std	0.8434	0.4565	0.7757	0.8434	0.7531	1.2170	4.1574	3.9705

表 V 八种算法在 40个数据集上 30 次独立运行所选特征数量的均值

No.	数据集	粒子群优化 [43]	[44] BGA [5] BDE		DimReM-BDE [5]	DimReM -BGA [5]	DimReM -BPSO [5]	SFE	SFE-PSO
1	ORL	497.88(+)	487.52(+)	471.68(+)	427.16(+)	528.37(+)	394.64(+)	94.5 (+)	**57.9**
2	Yale	500.13(+)	491.09(+)	474.18(+)	441.81(+)	424.22(+)	411 (+)	50.27 (–)	57.0667
3	结肠	988.13(+)	945.66(+)	934.53(+)	936.5 (+)	862.93(+)	865.06(+)	17.4 (+)	**13.0333**
4	小圆蓝细胞肿瘤	1138 (+)	1109 (+)	1084 (+)	1083 (+)	1032 (+)	1020 (+)	32.23(–)	43.1
5	warpAR10P	1180 (+)	1154 (+)	1146 (+)	1122 (+)	1080 (+)	1077 (+)	33.23(+)	**26.8667**
6	肺癌1	1625 (+)	1593 (+)	1568 (+)	1581 (+)	1510 (+)	1502 (+)	77.9 (–)	127.4
7	淋巴瘤	1974 (+)	1849 (+)	1909 (+)	1927 (+)	1846 (+)	1854 (+)	84.63(–)	149.9
8	胶质瘤	2188 (+)	2137 (+)	2116 (+)	2139 (+)	2050 (+)	2056 (+)	26.8 (–)	28.6333
9	白血病 1	2647 (+)	2593 (+)	2560 (+)	2587 (+)	2531 (+)	2507 (+)	46.43(–)	115.5
10	弥漫性大B细胞淋巴瘤	2695 (+)	2650 (+)	2625 (+)	2653 (+)	2561 (+)	2565 (+)	29 (+)	**18.9333**
11	膀胱癌	2820 (+)	2767 (+)	2734 (+)	2487 (+)	2669 (+)	2695 (+)	31.80(+)	**18.6**
12	9种肿瘤	2830 (+)	2785 (+)	2766 (+)	2791 (+)	2716 (+)	2715 (+)	65.83(–)	153.5333
13	TOX_171	2846 (+)	2806 (+)	2762 (+)	2800 (+)	2746 (+)	2727 (+)	106.1(–)	221.1333
14	脑肿瘤1	2933 (+)	2864 (+)	2837 (+)	2862 (+)	2788 (+)	2781 (+)	110.2(–)	202.8
15	前列腺_GE	2958 (+)	2912 (+)	2897 (+)	2908 (+)	2863 (+)	2835 (+)	29.8 (+)	**23.7**
16	白血病	3542 (+)	3487 (+)	3458 (+)	3499 (+)	3416 (+)	3422 (+)	37.26(–)	70.8667
17	ALLAML	3541 (+)	3472 (+)	3425 (+)	3493 (+)	3384 (+)	3387 (+)	32.66(+)	**27.8333**
18	CNS	3531 (+)	3484 (+)	3467 (+)	3486 (+)	3403 (+)	3403 (+)	36.4 (+)	**30.4667**
19	ALL-AML-3	3536 (+)	3473 (+)	3434 (+)	3491 (+)	3396 (+)	3396 (+)	53.96(–)	60.6667
20	ALL-AML-4	3543 (+)	3427 (+)	3465 (+)	3492 (+)	3396 (+)	3382 (+)	62 (–)	120.8
21	白血病 2	3537 (+)	3459 (+)	3431 (+)	3487 (+)	3401 (+)	3394 (+)	59.43(–)	100.0333
22	儿童急性淋巴细胞白血病	12232 (+)	12232 (+)	12232 (+)	12232 (+)	12232 (+)	12232 (+)	1223 (+)	**1204**
23	癌	4563 (+)	4523 (+)	4490 (+)	4523 (+)	4470 (+)	4490 (+)	147.8(–)	363.5667
24	nci9	4815 (+)	4759 (+)	4730 (+)	4785 (+)	4694 (+)	4675 (+)	60.2 (+)	**57.5667**
25	arcene	4992 (+)	4970 (+)	4978 (+)	4958 (+)	4933 (+)	4926 (+)	208.6(–)	276.5333
26	orlraws10P	5081 (+)	5014 (+)	4986 (+)	5046 (+)	4941 (+)	4962 (+)	33.26(+)	**28.8**
27	脑肿瘤2	5148 (+)	5070 (+)	5036 (+)	5109 (+)	4993 (+)	4995 (+)	38.23(+)	**23.3667**
28	白血病	5573 (+)	5455 (+)	5440 (+)	5536 (+)	5398 (+)	5400 (+)	51.44(–)	119.9333
29	慢性淋巴细胞白血病 SUB_111	5657 (+)	5629 (+)	5646 (+)	5614 (+)	5600 (+)	5573 (+)	55.56(+)	**48.3667**
30	11种肿瘤	6239 (+)	6221 (+)	6223 (+)	6223 (+)	6180 (+)	6147 (+)	141.0(–)	416.1667
31	MLL	6205 (+)	6149 (+)	6117 (+)	6205 (+)	6079 (+)	6089 (+)	60.3 (+)	**41.7**
32	肺癌2	6272 (+)	6220 (+)	6123 (+)	6223 (+)	6132 (+)	6117 (+)	74.52(–)	109.9
33	CML治疗	6246 (+)	6143 (+)	6092 (+)	6195 (+)	6073 (+)	6105 (+)	30.24(+)	**21.68**
34	胶质母细胞瘤	6263 (+)	6194 (+)	6137 (+)	6245 (+)	6132 (+)	6114 (+)	55.64(+)	**35.36**
35	急性髓系白血病预后	6287 (+)	6186 (+)	6141 (+)	6220 (+)	6109 (+)	6109 (+)	61.72(+)	**43.96**
36	前列腺癌	6260 (+)	6159 (+)	6127 (+)	6235 (+)	6076 (+)	6102 (+)	24 (+)	**18.88**
37	卵巢癌	7494 (+)	7469 (+)	7452 (+)	7435 (+)	7332 (+)	7324 (+)	43.63(+)	**23.9**
38	SMK_CAN_187	9958 (+)	9915 (+)	9965 (+)	9943 (+)	9935 (+)	9954 (+)	81.03(+)	**68.5667**
39	GLI_85	11088 (+)	10919 (+)	10928 (+)	11029 (+)	10867 (+)	10889 (+)	115.3(–)	225.1
40	乳腺癌	12177 (+)	12232 (+)	12092 (+)	12232 (+)	12042 (+)	12023 (+)	84 (+)	**59.6667**
Wilcoxon检验(+ | ≈ | –)		40 | 0 | 0	40 | 0 | 0	40 | 0 | 0	40 | 0 | 0	40 | 0 | 0	40 | 0 | 0	21 | 0 | 19	–
Friedman 检验（平均秩）		7.83	6.11	5.36	6.29	3.78	3.64	1.53	**1.48**

II. 结果与讨论

A. 运行时间

运行时间是评估特征选择算法效率时最重要的标准之一。在采用进化计算方法进行特征选择时，特征选择算法的运行时间取决于多种因素，例如特征选择算法所用方法的运行时间，以及适应度函数评估的运行时间。例如，在某些特征选择方法中，可能会使用基于过滤器的算法来识别重要特征并缩减搜索空间的维度。使用这些算法会增加特征选择算法的运行时间。值得注意的是，适应度函数评估的运行时间取决于数据集中的特征数量。评估特征数量较多的解所需的运行时间，比评估特征数量较少的解更长。因此，如果特征选择算法能够以最少的特征实现最高的分类准确率，就能显著降低计算成本。为确保公平比较，所有算法的评估次数均为 6000。

表 VI 展示了各算法在每个数据集上进行 30 次独立运行的平均运行时间（秒）。结果表明，与其他算法相比，SFE 的速度更快。这是因为 SFE 能够利用针对不重要特征的非选择算子探索搜索空间，并将这些特征从主特征集中移除。例如，乳腺癌数据集包含 24,481 个特征；其余六种算法在继续搜索过程时仍处理约 12,000 个特征，而 SFE算法在大多数搜索阶段继续搜索时处理的特征不到 100 个。结果表明，SFE-PSO算法在所有数据集上的运行时间都略长于 SFE算法。图 1 展示了不同算法在所有数据集上的平均运行时间；与其他算法相比，SFE 和 SFE-PSO 的计算成本最低。

SFE算法的运行时间表明，与其他算法相比，该算法速度很快。SFE算法在全部40个数据集上运行一次的平均计算成本为159分钟。相比之下，BDE、BGA、BPSO、DimReM-BDE、DimReM-BGA和DimReM-BPSO算法的计算成本约为420分钟。SFE-PSO算法的计算时间比SFE算法长9分钟。因此，由于在SFE和SFE-PSO

算法中，搜索过程从非选择算子开始，大量不重要特征会转变为非选择模式，这些算法的运行时间该算法的运行时间显著短于其他六种算法。在特征数量较多的数据集上，这一优势尤为明显，如表 VI 所示。

表 VI 八种算法在40个数据集上的运行时间（单位：秒）

No.	数据集	BDE [43]	BPSO [5]	[44] BGA	DimReM-BDE [5]	DimReM-BGA [5]	DimReM-BPSO [5]	SFE	SFE-PSO
1	ORL	541.2461(+)	583.4642(+)	544.4080(+)	533.0367(+)	528.3768(+)	539.2996(+)	274.7719(–)	307.4716
2	Yale	500.1364(+)	491.0909(+)	474.1818(+)	441.8182(+)	424.2273(+)	411.0000(+)	241.0767(–)	246.2473
3	结肠	283.9458(+)	246.6358(+)	248.5017(+)	263.4985(+)	283.9458(+)	302.9982(+)	203.5401(–)	210.2160
4	小圆蓝细胞肿瘤	259.0920(+)	279.9288(+)	268.5830(+)	272.3974(+)	274.6476(+)	264.8922(+)	206.2531(–)	215.5863
5	warpAR10P	315.0313(+)	310.9073(+)	302.5929(+)	311.6762(+)	293.6838(+)	297.8050(+)	204.4825(–)	212.6028
6	肺癌1	553.6767(+)	543.3894(+)	624.7057(+)	566.1042(+)	535.3923(+)	539.2301(+)	262.2813(–)	281.0564
7	淋巴瘤	377.8539(+)	350.6657(+)	349.7745(+)	368.9600(+)	346.6688(+)	350.7426(+)	239.0331(–)	262.2968
8	胶质瘤	351.5166(+)	442.2780(+)	277.6318(+)	258.5156(+)	288.3924(+)	257.9036(+)	235.8901(–)	249.7273
9	白血病 1	354.0600(+)	330.2223(+)	332.2564(+)	374.9271(+)	324.3148(+)	326.7226(+)	250.8541(–)	261.6521
10	弥漫性大B细胞淋巴瘤	389.2433(+)	380.3789(+)	375.1956(+)	384.2798(+)	365.1615(+)	364.0092(+)	244.6797(–)	248.8541
11	膀胱癌	290.9187(+)	289.7900(+)	286.8089(+)	283.8560(+)	272.7860(+)	287.1854(+)	221.7138(–)	235.1802
12	9种肿瘤	320.5015(+)	302.1543(+)	304.5801(+)	317.9044(+)	298.9058(+)	302.3159(+)	250.8441(–)	265.2514
13	TOX_171	603.0840(+)	590.4988(+)	590.3512(+)	593.5654(+)	581.6480(+)	590.3741(+)	277.8532(–)	295.6514
14	脑肿瘤1	394.2107(+)	371.7299(+)	427.3627(+)	416.1046(+)	369.0812(+)	399.1046(+)	251.6583(–)	267.2031
15	前列腺 GE	424.1751(+)	450.2496(+)	456.9916(+)	427.5325(+)	439.9723(+)	441.6824(+)	274.4060(–)	290.2154
16	白血病	377.8539(+)	350.6657(+)	349.7749(+)	368.9600(+)	346.6688(+)	350.7428(+)	251.7532(–)	267.5201
17	ALLAML	368.2364(+)	344.0183(+)	344.7949(+)	358.6048(+)	343.5811(+)	343.6836(+)	217.2698(–)	242.1219
18	CNS	380.4685(+)	347.4508(+)	375.7153(+)	477.1344(+)	348.2764(+)	351.6070(+)	223.4291(–)	246.1087
19	ALL-AML-3	440.2320(+)	422.6172(+)	381.1674(+)	489.5207(+)	376.4393(+)	431.2530(+)	251.5603(–)	247.1452
20	ALL-AML-4	469.0674(+)	441.9734(+)	442.2845(+)	390.1674(+)	376.7197(+)	365.1251(+)	245.5339(–)	266.5186
21	白血病 2	426.1322(+)	406.2893(+)	431.5671(+)	422.5816(+)	395.4409(+)	418.8220(+)	251.0294(–)	270.3005
22	儿童急性淋巴细胞白血病	467.9675(+)	460.8471(+)	451.0124(+)	482.6325(+)	464.3200(+)	460.9524(+)	271.4392(–)	283.3181
23	癌	479.2894(+)	471.0015(+)	465.1265(+)	479.2029(+)	470.2564(+)	472.2954(+)	223.8957(–)	235.5214
24	nci9	395.6600(+)	499.8097(+)	359.5930(+)	392.7161(+)	370.1463(+)	414.9878(+)	209.2154(–)	218.9003
25	arcene	1386.102(+)	1340.037(+)	1401.403(+)	1392.215(+)	1368.321(+)	1375.573(+)	255.1887(–)	269.2514
26	orlraws10P	561.4201(+)	690.8217(+)	521.7263(+)	547.4365(+)	527.7035(+)	580.4538(+)	261.0213(–)	275.8608
27	脑肿瘤2	402.2516(+)	398.4953(+)	399.8541(+)	405.0210(+)	401.1234(+)	408.0094(+)	245.8012(–)	258.6450
28	白血病	540.9726(+)	462.8705(+)	534.8302(+)	530.9527(+)	469.2992(+)	486.2307(+)	241.1309(–)	255.8521
29	CLL SUB_111	838.0785(+)	803.5214(+)	801.0213(+)	784.8991(+)	758.3350(+)	789.2531(+)	243.9980(–)	251.5293
30	11种肿瘤	1372.096(+)	1223.859(+)	1222.165(+)	1572.976(+)	1565.088(+)	1362.643(+)	252.1077(–)	273.8871
31	MLL	554.3502(+)	480.3871(+)	568.1775(+)	526.9200(+)	488.3199(+)	515.9554(+)	229.1854(–)	239.2937
32	肺癌2	1593.254(+)	1513.923(+)	1532.202(+)	1734.090(+)	1562.636(+)	1723.717(+)	265.1183(–)	299.6739
33	CML治疗	362.4804(+)	346.4675(+)	327.8606(+)	352.1110(+)	335.6387(+)	352.6826(+)	202.9308(–)	205.9978
34	胶质母细胞瘤	442.6616(+)	418.4480(+)	453.7888(+)	431.3714(+)	380.8300(+)	458.6237(+)	204.8273(–)	209.7009
35	急性髓系白血病预后	363.2683(+)	332.4697(+)	328.2568(+)	346.6846(+)	303.0070(+)	332.1513(+)	201.6439(–)	215.4942
36	前列腺癌	396.5907(+)	303.8168(+)	353.2002(+)	364.8926(+)	342.3289(+)	325.8817(+)	215.6388(–)	223.3617
37	卵巢癌	3210.561(+)	3111.514(+)	3118.526(+)	3110.967(+)	3105.965(+)	3109.001(+)	276.3198(–)	291.6594
38	SMK_CAN_187	2328.012(+)	2319.251(+)	2306.294(+)	2311.351(+)	2318.351(+)	2316.951(+)	236.9516(–)	238.6916
39	GLI_85	917.1863(+)	890.7940(+)	895.1259(+)	895.9701(+)	949.3433(+)	882.7051(+)	224.4827(–)	228.5900
40	乳腺癌	1410.124(+)	1386.253(+)	1330.001(+)	1382.025(+)	1389.201(+)	1342.012(+)	244.6513(–)	251.3931
Wilcoxon检验(+ | ≈ | –)		40 | 0 | 0	40 | 0 | 0	40 | 0 | 0	40 | 0 | 0	40 | 0 | 0	40 | 0 | 0	0 | 0 | 40	–
Friedman 检验（平均秩）		6.99	5.23	5.28	6.15	4.21	5.15	**1.02**	1.98

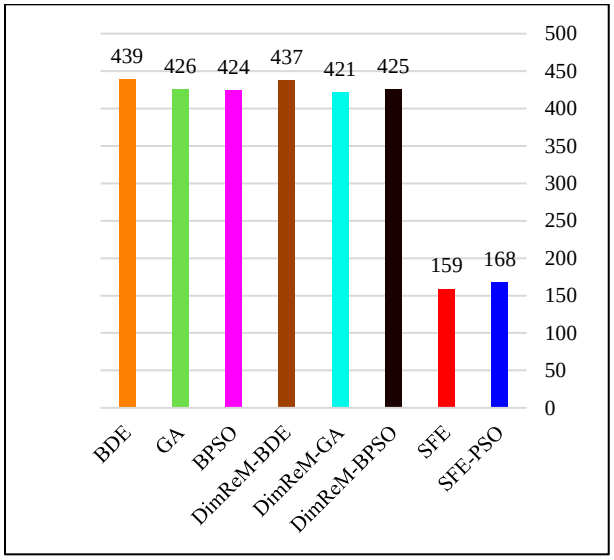

图 1. 八种算法在40个数据集上的平均训练时间（单位：分钟）。

III. 实验结果的进一步分析

本节将分析根据分类准确率和选定特征数量评估的实验结果，以及 SFE 和 SFE-PSO 算法在搜索过程中的表现。

C. 分类准确率

SFE和SFE-PSO算法在平均准确率方面优于其他算法，一个重要原因是使用非选择算子将无关特征的状态更改为非选择模式，并使用选择算子将相关特征添加到最优特征子集中。事实上，在高维数据集中，尤其是微阵列数据中，许多特征与分类任务无关。因此，如果我们能通过某种机制识别出这些特征，并将其设为非选择模式，就能达到预期的效率。SFE算法中的非选择算子正是这种机制。

此外，由于算子具有随机性，该算法能够跳出局部最小值并获得最优解。例如，图2展示了SFE算法在Colon数据集上的一次运行中的搜索过程。

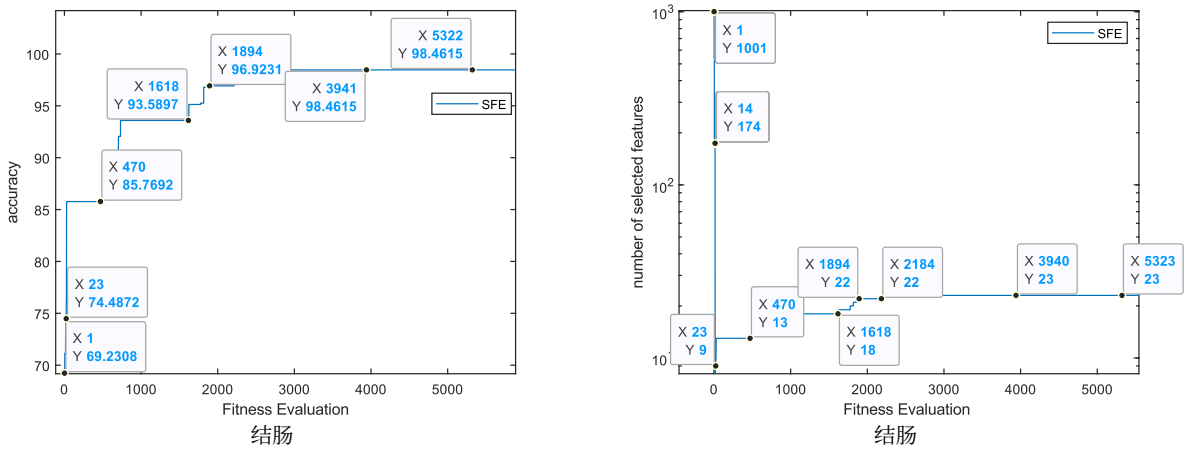

图2. SFE算法在Colon 数据集上的一次运行中的搜索过程。

右图显示了运行过程中特征的数量。算法以 1001 个特征开始搜索过程；在第 23 次迭代时，选定特征的数量达到 9 个（使用无选择算子）。在后续搜索过程中，这使算法的计算成本降低，而重要特征（通过选择算子）则被添加到特征集中。左图显示了该算法本次执行的收敛准则。算法的分类准确率达到 96.92，经过 1894 次迭代；搜索从第 1 次迭代时 69.23 的准确率开始。

相反，进化计算方法在搜索过程中没有这类机制，因此，启动搜索并经过若干次迭代后，它们会停留在问题的某个局部最优解上，无法达到预期的效率。例如，图3展示了BPSO算法在 Colon 数据集上一次运行中的搜索过程。

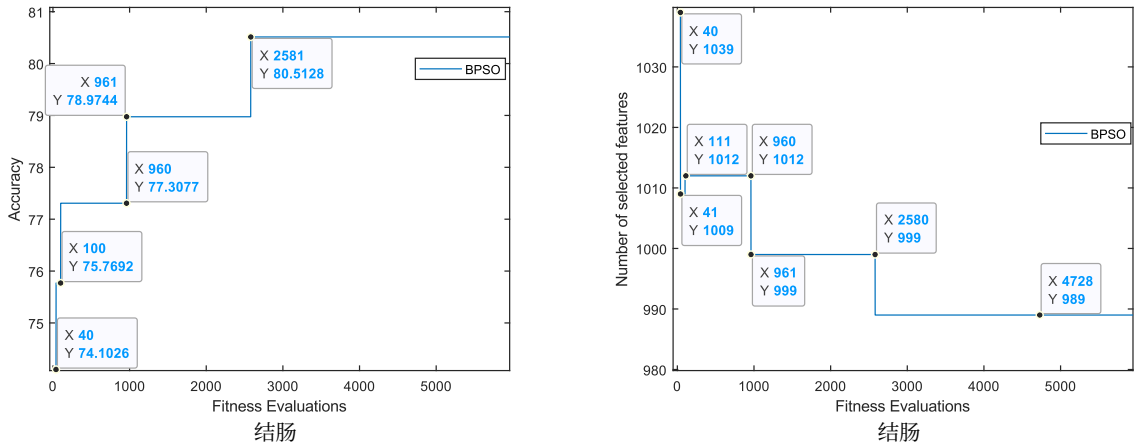

图 3. BPSO算法在 Colon 数据集上的一次运行中的搜索过程。

右图显示了一次运行中的特征数量。在第40次迭代中，BPSO算法选定了1039个特征；在第41次迭代中，选定特征的数量达到1009个。随后，该算法以989个特征结束搜索过程。事实上，该算法未能选出重要特征。左图显示了该算法在此次执行中的收敛准则。

该算法在迭代 40 次后达到了 74.10 的分类准确率。在后续搜索过程中，分类准确率指标有所提高。然而，最终算法在迭代 2581 次后停滞于局部最优。SFE 和 BPSO 算法在 SMK_CAN_187、CNS 和 CLL SUB_111 数据集上的结果也见图4和图 5。

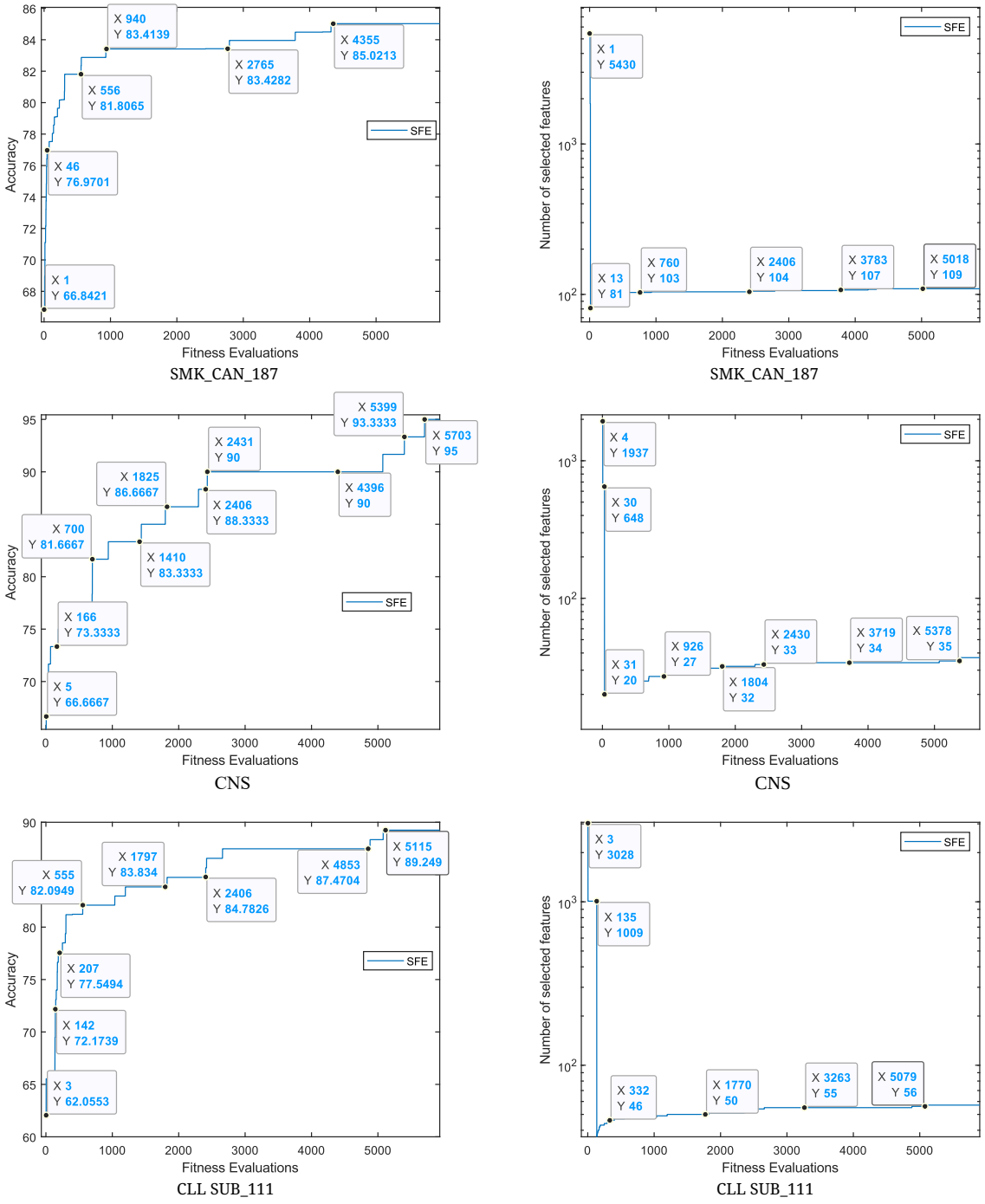

图4. th 的搜索过程　　在 SMK_CAN_187、CNS 上一次运行 SFE算法　　，以及 CLL SUB_111 数据ets.

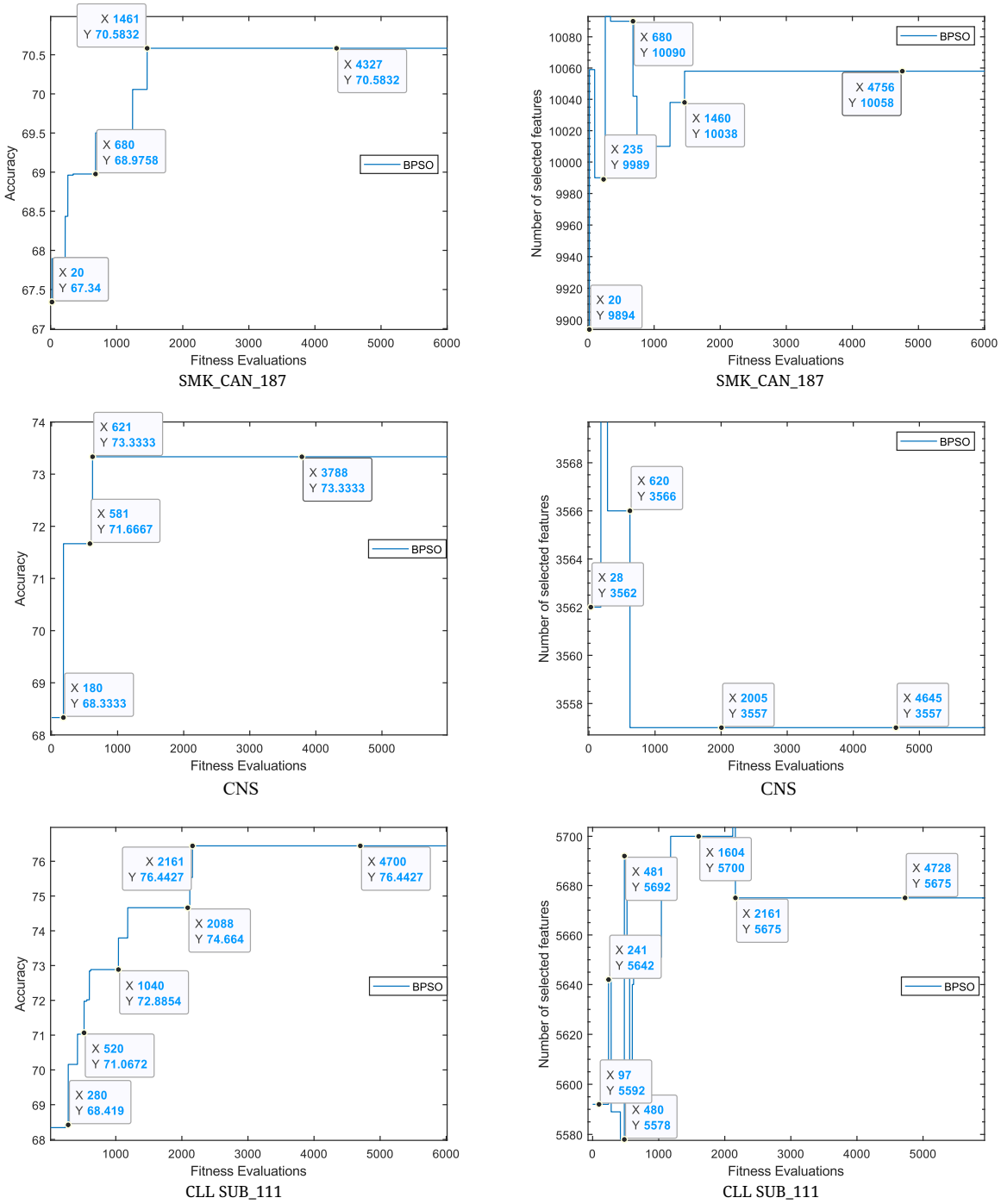

图 5. BPSO算法在 SMK_CAN_187、CNS 和 CLL SUB_111 数据集上的一次运行中的搜索过程。

此外，在 SFE-PSO算法中，PSO 会继续搜索过程，从 SFE算法确定的特征集中寻找最佳特征子集。相反，BDE、BGA 和 BPSO 这三种算法会停留在高维数据的某个局部最优点，并且由于其搜索策略的限制，在搜索过程中无法获得更高的分类准确率。此外，DimReM-DBE、DimeRem-BGA 和 DimeReM- BPSO 这三种算法由于移除无关特征的过程缓慢，无法实现高准确率分类。其他最先进的基于进化计算的特征选择算法，即 HFS-C-P、SaWDE 和 SM-MOEA，取得了与所提算法相当的结果。然而，Wilcoxon 检验和 Friedman 检验的结果表明，SFE-PSO算法优于这些算法。

D. 选定特征的数量

与 BDE、BGA、BPSO、DimReM-BDE、DimReM-BGA 和 DimReM-BPSO 算法相比，SFE 和 SFE-PSO 算法在选定特征数量这一指标上具有显著优势，其原因在于使用了非选择算子。事实上，SFE算法利用非选择算子，将不重要特征的状态从选择模式切换为非选择模式。因此，该算法能够找到特征数量最少的解。此外，在搜索过程的不同阶段，算法会利用选择算子将相关且重要的特征加入最优特征子集。相反，BDE、BGA 和 BPSO 这三种算法在整个搜索过程中，其解中几乎都包含数据中一半的特征，其中大多数是不重要特征。因此，由于这些方法没有识别无关特征的机制，实际上无法找到特征数量更少的解。此外，对于 DimReM-BDE、DimReM-BGA 和 DimReM-BPSO 这三种算法，识别并移除无关特征的过程较为缓慢，因此最终无法获得特征数量更少的解。

E. SFE和SFE-PSO算法在搜索过程中的行为 观察图6，即……的收敛图

不同数据集上的算法表现，为我们提供了有关算法在搜索过程中的行为的重要信息。观察 SFE 和 SFE-PSO 算法在 ORL、Yala、Carcinom、TOX_171 和 Carcinom 等数据集上的表现可以发现，在这些数据集中，由于 UN 和 SN 参数值不合适等原因，SFE算法的表现较差。因此，这些算法在搜索过程的早期阶段无法获得更高的准确率。与此同时，在搜索过程继续进行时，SFE算法能够在 TOX_171 等一些数据集上获得更好的解。另一方面，在 Yale 等一些数据集中，即使到了最后一次迭代，该算法也未能显著提高分类准确率。观察 SFE-PSO 算法的表现可以发现，经过 2000年次迭代，并通过 PSO算法继续搜索过程后，该算法的分类准确率得到了显著提升。这是因为PSO算法能够在低维数据中选择特征。事实上，PSO算法在 SFE算法缩减后的搜索空间中继续搜索后，能够获得更好的解。这一点可以从该算法在 Yale、淋巴瘤和 TOX_171 等数据集上的结果中看出。