基于 SFE 的不平衡数据二阶段特征选择算法

沈先倩[1] 杨盛毅[2] 陈 静[3] 何小飞[1] 程俞富[1]
SHEN Xianqian YANG Shengyi CHEN Jing HE Xiaofei CHENG Yufu

摘 要 在数据处理领域，高维特征与类别不平衡问题已成为诸多研究面临的棘手挑战。鉴于此，文章以 SFE 算法作为坚实基石，创新性地提出了一种专门面向不平衡数据情境的二阶段特征选择算法——SFE-TSFS（a two-stage feature selection algorithm for imbalanced data based on SFE）。首先，针对 SFE 算法未能处理类别不平衡的局限，SFE-TSFS 引入了 Borderline-SMOTE 方法生成符合要求的边界样本，以平衡数据分布。其次，为加速算法收敛并降低计算成本，在初始特征筛选阶段结合了模糊互信息方法，有效去除大量冗余特征。实验结果表明，与原 SFE 算法相比，文章所提出的算法在保证分类准确率的同时，运行时间和特征数量上均优于 SFE 算法，验证了 SFE-TSFS 算法的有效性。

关键词 高维数据；不平衡数据；特征选择；Borderline-SOMTE；模糊互信息

doi：10.3969/j.issn.1672-9528.2025.01.039

0 引言

随着计算机和新技术的发展，在医疗保健、电子商务、生物信息学、社交媒体和交通运输等各种应用中产生了大量高维数据集 [1]。机器学习算法被广泛应用于分类问题，然而，维数灾难和类别不平衡问题阻碍了分类结果的实现，因许多算法原本是针对低维数据集所设计，当面对高维数据集时，其中分布的数据特性与低维情形差异较大，往往会致使这些算法无法有效发挥作用，甚至产生诸多不利影响。此外，数据的高维性和类别不平衡增加了机器学习算法在数据分类中时间和空间的复杂度。

在处理高维不平衡数据时，研究人员通常会使用特征选择算法来降低数据的维度。此外，在许多数据集上，特征选择方法还可以减弱数据集中的不平衡因素对结果的影响 [2]。然而，现有研究往往仅聚焦于高维或类别不平衡的单一问题，例如文献 [3-4] 只探讨了高维或不平衡问题。但也有研究同时考虑高维与不平衡数据交叉带来的挑战，甚至探讨其对分类精度的共同影响 [5-6]。

现有的特征选择算法主要分为：过滤式、包裹式和嵌入式 [7]。过滤式通过评估单个特征的贡献来筛选，计算效率高，但忽略了特征间的关联性，导致选择的特征组合未必最佳；包裹式结合模型表现评估特征组合，能够更好地考虑特征间的相互作用，尽管计算成本更高且依赖特定分类器，但通常能得到更优结果；嵌入式算法是在分类器训练过程中同步进行特征选择，相比包裹式计算开销较低，但对分类器依赖较大。混合方法则通过整合不同算法的优势，实现较好性能，但仍面临较高的计算成本。

鉴于此，Ahadzadeh 等人 [8] 提出了 SFE（simple fast and effective high dimensional feature selection）算法，SFE 算法为高维数据集提供了一种高效的特征选择算法，以解决高维数据集中的特征选择问题，能够在减少特征数量和计算成本的同时，实现最高的分类准确率。该算法采用包裹式特征选择方法，通过 KNN 分类器的准确率来衡量特征集的优劣，从而进行特征选择，但在不平衡数据集上，准确率并不能有效衡量分类器的实际表现，如果多数类占主导地位，算法可能会选择对多数类有贡献的特征，而忽略对少数类有利的特征。这会导致最终模型对少数类表现不佳。且在初始筛选特征阶段采用随机选择特征数量，这使得算法易陷入局部最优，计算成本较大。

针对上述问题，本文提出了一种适合高维不平衡数据集的二阶段特征选择算法，主要贡献如下：

（1）为解决高维数据中的类别不平衡问题，本文将 Borderline-SMOTE 方法引入 SFE 算法，有效缓解了不平衡现象。

（2）针对算法容易陷入局部最优以及计算成本较高问题，本文结合 SFE 与模糊互信息方法，有效提升了搜索最优特征子集的效率。

1. 贵州民族大学数据科学与信息工程学院 贵州贵阳 550000
2. 贵州民族大学物理与机电工程学院 贵州贵阳 550000
3. 包头师范学院信息科学与技术学院 内蒙古包头 014030
[基金项目] 贵州民族大学科研基金（GZMUZK〔2023〕CXTD07）

1 相关研究

1.1 Borderline-SMOTE 算法

Borderline-SMOTE[9] 是一种过采样算法。在不平衡数据集的研究中，通过增加少数类样本数量来调整数据分布。作为 SMOTE 的变体，Borderline-SMOTE 在不平衡数据研究中应用广泛。其核心思想是将少数类样本分为三类：安全样本、边界样本和噪声样本。在少数类样本 x_i 的 K 近邻样本中，若多数类样本数量 $p=k$，将 x_i 划分为噪声样本；若 $0\le p<k/2$，将 x_i 划分为安全样本；若 $k/2\le p<k$，将 x_i 划分为边界样本。算法通过划分少数类样本，专注于对边界样本进行线性插值生成新样本，不仅提升了合成样本的质量，还强化了对边界样本的关注。

1.2 SFE 高维数据特征选择算法

SFE 算法通过一个搜索代理和两个操作符（选择和非选择操作符）来执行搜索。该算法主要分为探索阶段和开发阶段。在探索阶段，非选择操作符负责全局搜索，遍历整个搜索空间，以识别和标记无关、冗余及噪声特征，并将其状态从选择模式切换为非选择模式。而在开发阶段，选择操作符则会在问题搜索空间中寻找对分类结果至关重要的特征，并将这些特征的状态从非选择模式更改为选择模式。

在 SFE 算法中，搜索代理 X 使用二进制编码来表示。具体来说，设原始数据集为 D，每个搜索代理 $S\{s_1, s_2, \cdots, s_d\}$ 表示一个候选特征子集 [10]，其中 d 为原始数据集特征数量，$s_i=1$ 表示该特征被选中；$s_i=0$ 表示该特征中未被选中。

在探索阶段，在搜索代理 X 上应用非选择操作符在问题的整个搜索空间中进行全局搜索，找到不重要特征，并将特征从选择状态更改为非选择状态。其中，非选择操作符所应用的特征数量 UN 公式为：

$$\mathrm{UN}=\mathrm{UR}\times n_{\mathrm{var}} \tag{1}$$

式中：n_{var} 是搜索空间维数或数据集特征数；UR（$0<\mathrm{UR}<1$）是非选择操作符比率。

在开发阶段，选择操作符在问题搜索空间中搜索对分类结果有影响的重要特征，将 K 近邻（K nearest neighbors，KNN）分类器的准确率作为适应度评估函数，公式为：

$$\mathrm{fit}(X)=\mathrm{Accuracy} \tag{2}$$

利用式（3）自适应线性递减，在 SFE 算法在探索和开发阶段之间使用 UR 值创建适当平衡。

$$\mathrm{UR}=(\mathrm{UR}_{\max}-\mathrm{UR}_{\min})\times(\mathrm{Max}_{\mathrm{EFs}}-\mathrm{EFs})/\mathrm{EFs}+\mathrm{UR}_{\mathrm{end}} \tag{3}$$

式中：$\mathrm{UR}_{\max}$ 为 UR 的最大值；$\mathrm{UR}_{\min}$ 为 UR 的最小值；$\mathrm{Max}_{\mathrm{EFs}}$ 为最大迭代次数；EFs 为当前的迭代次数。

2 本文方法

2.1 提出的 SFE-TSFS 算法

SFE 算法在高维数据集的特征选择中表现出色。但算法并未考虑数据集类别不平衡问题，忽略了对少数类有利的特征，导致最终算法对少数类表现不佳。且算法初始筛选候选特征子集时，特征选择的方式存在随机性过强的问题，未结合特征重要性信息来决定哪些特征应该被取消选择，而是完全依赖随机生成，导致算法在一些情况下不够灵活、造成计算成本过大。针对上述问题，本文提出一种不平衡数据的二阶段特征选择算法（SFE-TSFS）。

第一阶段，为了优化算法并解决数据集的不平衡问题，本文采用了 Borderline-SMOTE 算法，通过增加少数类样本来平衡数据。为提高分类器对少数类样本的学习效果，针对那些被多数类样本包围且难以学习的少数类样本，生成更多合成数据。

具体而言，通过计算少数类样本与其他样本的欧氏距离，确定其 K 近邻。接着，将少数类样本划分为安全样本、边界样本和噪声样本。Borderline-SMOTE 算法强调，位于类别边界区域的少数类样本容易被误分类，因此需要加强分类器对这些样本的训练。通过插值公式（4）对边界样本进行线性插值，从而生成新的少数类合成样本 X_{new}，进而提升分类器的准确性。

$$X_{\mathrm{new}}=X_{\mathrm{new}}+\mathrm{round}(0,1)\times(X_i-X) \tag{4}$$

式中：X 表示原少数类样本；X_i 表示第 i 个少数类样本。

在第二阶段，针对算法在某些情况下缺乏灵活性且计算成本较高的问题，本文引入了模糊互信息 [11] 这一过滤式算法。通过计算特征和标签的相关度，设置阈值，选择具有最高相关性得分的特征，减少了冗余特征的影响，从而降低后续处理的数据量。模糊互信息是基于经典的互信息理论，并结合模糊集理论来处理不确定性和模糊性。一个重要优势在于其能有效应对数据中的模糊性和不确定性。通过对特征重要性进行精准量化，模糊互信息能够提高早期筛选阶段的特征选择效率，进而提升分类器性能，使得后续分类更加准确且高效。

具体地，模糊互信息可以定义为：

$$\mathrm{FMI}(X,Y)=H(X)+H(Y)-H(X,Y) \tag{5}$$

式中：X 和 Y 表示两个模糊变量；$H(X)$、$H(Y)$ 分别为 X 和 Y 的模糊熵；$H(X, Y)$ 为特征和类标签的模糊联合熵。

设有数据集 $X\{x_1, x_2, \cdots, x_n\}$，$A$、$B$ 为定义在 X 上的两个模糊集，第 i 类中第 k 个特征的模糊隶属度计算式为：

$$\mu_{i,k}=\left(\|\overline{x}_i-x_k\|_\delta/\left(r+\varepsilon\right)\right)^{\frac{-2}{m-1}} \tag{6}$$

式中：m 是模糊化系数，在本文中值为 2；ε 是大于 0 的值；δ 是距离计算中涉及的标准差；$\overline{x}_i$ 表示属于类别 i 的数据的均值；数据的半径表示为：

$$r=\max(\|\overline{x}_i-x_k\|_\delta) \tag{7}$$

模糊熵和模糊联合熵在 X 上的表达式分别为：

$$H(A)=-\frac{1}{n}\sum_{x\in X}[\mu_A(x)\log\mu_A(x)+(1-\mu_A(x))\log(1-\mu_A(x))] \quad (8)$$

$$H(B)=-\frac{1}{n}\sum_{x\in X}[\mu_B(x)\log\mu_B(x)+(1-\mu_B(x))\log(1-\mu_B(x))] \quad (9)$$

$$H(A\cup B)=-\frac{1}{n}\sum_{x\in X}[\mu_A(x)\vee\mu_B(x)]\log[\mu_A(x)\vee\mu_B(x)]+[1-\mu_A(x)\vee\mu_B(x)]\log[1-\mu_A(x)\vee\mu_B(x)] \quad (10)$$

按照特征的模糊互信息得分进行特征重要性排序，通过阈值的设置得到初步筛选后的特征子集，经由 SFE 算法进行最优特征子集搜索。

综上所述，通过引入 Borderline-SMOTE 算法解决数据集中的不平衡问题，生成了一个在少数类与多数类边界处保留了边界少数类样本分布特征的平衡数据集。使用结合了优化 SFE 算法与模糊互信息的算法对该平衡数据集进行最优特征子集的搜索。经过两个阶段，不仅有效减少了算法的计算成本，还提升了算法的寻优速度和特征选择的准确性。

2.2 SFE-TSFS 算法时间复杂度分析

设原始数据集有 n 个样本和 d 个特征，数据集被分为 c 个类别。则本文算法时间复杂度取决于以下部分：在计算特征重要性阶段需要 $O(nd)$，后续 SFE 算法阶段并未改变算法执行时间，所以 SFE 与模糊互信息结合的算法时间复杂度为 $O(\mathrm{it}_{\max}\times(d'+f))$，其中，$\mathrm{it}_{\max}$ 表示最大迭代次数；d' 是经过模糊互信息算法过滤后的数据集维度；f 是适应度函数计算成本。综上，SFE-TSFS 算法的总时间复杂度为 $O(nd+\mathrm{it}_{\max}\times(d'+f))$。并且，在初步筛选特征阶段，算法大幅度减少特征数量。因此，SFE-TSFS 算法中数据维度远小于传统的 SFE 算法。

3 实验与结果分析

为了验证 SFE-TSFS 算法的有效性，实验选择了 5 个数据集来评估所提出算法的性能，数据集均来源于 UCI 数据集。同时，将其结果与传统的 SFE 算法和 BGA 算法进行比较，这两者在特征选择领域具有较高的知名度，以检验所提出算法在高维不平衡数据集中的表现。此外，实验环境为 Windows11 64 位操作系统 , Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz 1.80 GHz，8 GB 内存，用 MATLABR2023a 进行实验。

相关数据集的详细信息见表 1，各数据集的最佳结果则在表 2 中以粗体标出。

表 1 各真实数据集的数据信息

数据集	样本数	特征数	类别	不平衡率 IR
Colon	62	2000	2	1.82
CNS	60	7129	2	1.73
ALL_AML_4	72	1024	4	9.25
CML treatment	28	12 625	2	1.33
Leukemia	72	7130	2	1.88

表 2 各算法的分类准确率和运行时间对比

数据集	评价指标	BGA	SFE	SFE-TSFS
Colon	Accuracy	80.461	**96.140**	95.154
	Time	246.630	35.940	22.939
CNS	Accuracy	73.220	90.280	**95.490**
	Time	147.450	30.890	25.083
ALL_AML_4	Accuracy	84.670	94.400	**97.510**
	Time	441.970	25.530	22.120
CML treatment	Accuracy	56.960	90.040	**96.694**
	Time	346.460	20.93	25.620
Leukemia	Accuracy	84.006	89.440	**95.947**
	Time	140.670	44.120	23.230

表 2 结果显示，相较于 BGA 算法，SFE 和 SFE-TSFS 算法在高维数据集上表现更为出色，具体地，本文算法在 5 个数据集上的平均分类准确率为 96.2%，相较于两个对比算法而言，总的平均分类准确率分别提高了 20.3% 和 4.1%，其中，在 ALL_AML_4 数据集上表现最为出色，相较于对比算法，分类准确率分别提高了 12.8% 和 3.11%。这一显著提升主要得益于 SFE-TSFS 算法结合过滤式和包裹式方法的优势，通过模糊互信息得出重要特征将进行最优子集搜索，有效避免了陷入局部最优，从而提升了分类准确率。

在算法的运行时间对比上，SFE 算法最大迭代次数为 900，本文算法的最大迭代次数为 700 次，并且在初步筛选特征阶段选择 FMI 值最大的前 500 个特征，这有助于本文算法保留重要特征，减少后续特征选择所需时间。因此，本文在 4 个数据集上的运行时间是最短的。特征子集数量方面，SFE 算法在 5 个数据集的最终特征数分为 38、31.7、62、30.2 和 41.7，相较于 SFE 算法，本文算法的在这 5 个数据集上的最终特征数量分别为 21.7、21.3、25.4、25.6 和 23。这是由于在特征初步筛选阶段，本文算法利用过滤式方法去除了大量的冗余特征，从而大大缩短了算法运行时间。

综上所述，SFE-TSFS 算法在数据集上的出色表现充分验证了其搜索最优特征子集方面的有效性。与传统的 SFE 算法相比，既保证了分类器精度，又减少了运行时间和特征数量。

4 总结

为了应对高维不平衡数据集的问题，本文基于 SFE 算法提出了一种新的二阶段特征选择算法，称为 SFE-TSFS，

（下转第 174 页）

[8] MOISES A V, PAULO AARON A A, ALFREDO P M, et al. Flexible convolver for convolutional neural networks deployment onto hardware-oriented applications[J].Applied sciences, 2023,13(1):93.

[9] 高强 , 邵春霖 , 李京润 , 等 . 面向图卷积神经网络的 FPGA 部署及加速研究 [J]. 现代电子技术 ,2024,47(10):39-46.

[10] 魏秀参 . 解析深度学习 [M]. 北京 : 电子工业出版社 ,2018.

[11] 刘凡平 . 神经网络与深度学习应用实战 [M]. 北京 : 电子工业出版社 :2018.

【作者简介】

钟戴元（2000—），男，湖南怀化人，硕士研究生，研究方向：SoC FPGA 系统、数字信号处理。

曾庆立（1975—），通信作者（email:2878627@qq.com），男，湖南永州人，博士在读，硕导、高级实验师，研究方向：FPGA 综合应用、RISC-V SoC 系统生态。

周佳凯（2000—），男，湖南永州人，硕士研究生，研究方向：FPGA IP 核、AXI 总线。

薛浪（1999—），女，贵州六盘水人，硕士研究生，研究方向：FPGA、数字信号处理。

唐瑞东（2000—），男，湖南娄底人，硕士研究生，研究方向：FPGA、图像信号处理。

（收稿日期：2024-10-11）

（上接第 169 页）

用于处理不平衡数据。在数据集预处理阶段加入 Borderline-SMOTE 克服不平衡问题，并结合模糊互信息在初步筛选特征的阶段去除大量的冗余特征，有效的加快了算法收敛速度和降低计算成本。在 5 个数据集上都有较高的分类准确率，较低的运行时间，这些实验很好地展示了 SFE-TSFS 的优势。

但本文的算法中也存在一定的局限性。具体表现为，算法在初步筛选特征阶段阈值的设定在很大程度上影响了算法的运行时间和分类准确率。未来的研究方向将围绕这点进行研究探索。

参考文献：

[1] KOLECK T A, DREISBACH C, BOURNE P E, et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review[J]. Journal of the American medical informatics association, 2019, 26(4): 364-379.

[2] WASIKOWSKI M, CHEN X W. Combating the small sample class imbalance problem using feature selection[J]. IEEE transactions on knowledge and data engineering, 2009, 22(10): 1388-1400.

[3] PRASETIYOWATI M L, MAULIDEVI N U, SURENDRO K. The accuracy of Random Forest performance can be improved by conducting a feature selection with a balancing strategy[J/OL]. PeerJ computer science, 2022[2024-02-13]. https://pubmed.ncbi.nlm.nih.gov/35875646/.

[4] GU S K, CHENG R, JIN Y C. Feature selection for high-dimensional classification using a competitive swarm optimizer[J]. Soft computing, 2016, 22(10): 811-822.

[5] YIN L Z, GE Y, XIAO K L, et al. Feature selection for high-dimensional imbalanced data[J]. Neurocomputing, 2013, 105(4): 3-11.

[6] 陈祥焰 , 林耀进 , 王晨曦 . 基于邻域粗糙集的高维类不平衡数据在线流特征选择 [J]. 模式识别与人工智能 , 2019, 32(8): 726-735.

[7] 苏逸 , 李晓军 , 姚俊萍 , 等 . 不平衡数据分类数据层面方法 : 现状及研究进展 [J]. 计算机应用研究 , 2023, 40(1): 11-19.

[8] AHADZADEH B, ABDAR M, SAFARA F, et al. SFE: a simple, fast, and efficient feature selection algorithm for high-dimensional data[J]. IEEE transactions on evolutionary computation, 2023, 27(6): 1896-1911.

[9] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[J/OL]. Advances in Intelligent Computing.878-887[2024-01-19]. https://link.springer.com/chapter/10.1007/11538059_91.

[10] 刘佳璇 , 李代伟 , 任李娟 , 等 . 面向不平衡医疗数据的多阶段混合特征选择算法 [J/OL]. 计算机工程与应用 , 1-13[2024-03-11].http://kns.cnki.net/kcms/detail/11.2127.tp.20240712.1803.026.html.

[11] HOQUE N, AHMED H A, BHATTACHARYYA D K, et al. A fuzzy mutual information-based feature selection method for classification[J]. Fuzzy information and engineering, 2016, 8(3): 355-384.

【作者简介】

沈先倩（1997—），女，贵州毕节人，硕士，研究方向：统计建模与模式识别。

（收稿日期：2024-10-10）