基于多层次集成学习的流特征在线稳定选择算法

王 琦,周 鹏,张燕平
(安徽大学 计算机科学与技术学院,安徽 合肥 230601)

摘 要:特征选择是数据挖掘预处理阶段中的重要组成部分,旨在从原始数据集中选择出最相关的特征子集。传统的特征选择方法假设数据集是静态不变的。然而,在实际应用中,数据可能是动态生成并被处理的。为此,针对特征以流的方式逐个生成的在线流特征选择方法应运而生。目前,大多数研究者所提出的在线流特征选择方法主要关注可扩展性、高准确性和低时间开销,而忽视了算法的稳定性。稳定的特征选择结果才能有效增强用户对算法的可信度,使其具备实用价值。针对在线特征选择算法的稳定性问题,基于多层次集成学习策略,提出了一种新的流特征在线稳定选择算法框架(Multi-level Ensemble Learning Stream Feature Selection, MESFS)。具体来说,在数据集层面采用极限学习机(Extreme Learning Machine, ELM)对样本进行分组和映射来提高算法的准确性;在特征选择层面通过多次迭代和自适应调整阈值的策略对特征进行权重计算和选择,以减少特征选择结果的波动性和随机性。选取 4 种传统静态特征选择算法和 5 种先进的在线流特征选择算法,在 UCI、ARFF 以及 NIPS 等 12 个公开数据集上进行了大量实验对比,结果表明该方法可以在训练数据扰动下取得优秀的预测精度和稳定性平衡。

关键词:特征选择;流特征;稳定性;集成学习;极限学习机
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2025)03-0001-08
doi:10.20165/j.cnki.ISSN1673-629X.2024.0346

Online Stable Streaming Feature Selection Algorithm Using Multi-level Ensemble Learning

WANG Qi, ZHOU Peng, ZHANG Yan-ping
(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: Feature selection is an essential part of the preprocessing phase of data mining, aiming to select the most relevant subset of features from the original data set. Traditional feature selection methods assume that the data set is static and unchanging. However, in real applications, data may be dynamically generated and processed. For this reason, online streaming feature selection methods emerged that generate features one by one in a streaming manner. Currently, most of the online stream feature selection methods proposed by researchers mainly focus on scalability, high accuracy, and low time overhead while ignoring the algorithm´s stability. Stable feature selection results can effectively enhance users´ credibility in the algorithm and make it of practical value. Aiming at the stability problem of the online feature selection algorithm, a new online stable stream feature selection algorithm framework (Multi-level Ensemble Learning Stream Feature Selection, MESFS) is proposed based on the multi-level ensemble learning strategy. Specifically, at the data set level, Extreme Learning Machine (ELM) is used to group and map samples to improve the accuracy of the algorithm. At the feature selection level, multiple iterations and adaptive threshold adjustment strategies are used to calculate the weight of features and selection to reduce the volatility and randomness of feature selection results. Four traditional static feature selection algorithms and five advanced online flow feature selection algorithms were selected, and many experimental comparisons were conducted on public data sets such as UCI, ARFF, and NIPS. The results show that the proposed method can perform excellently under training data disturbance—the balance between prediction accuracy and stability.

Key words: feature selection; streaming feature; stability; ensemble learning; extreme learning machine

0 引 言

特征选择旨在从原始特征空间中选择最少的特征,以提高任务的可解释性、性能和效率[1-2]。特征选择能识别最相关的特征支持决策[3]。根据不同的选择

收稿日期:2024-05-12 **修回日期**:2024-09-12
基金项目:国家自然科学基金面上项目(62376001);安徽省自然科学基金面上项目(2308085MF215)
作者简介:王 琦(1999-),男,硕士研究生,研究方向为机器学习、数据挖掘;周 鹏(1987-),男,博士,副教授,CCF 高级会员(K6292M),研究方向为机器学习、人工智能;张燕平(1962-),女,博士,教授,研究方向为机器学习方法与应用、计算智能与商空间理论。

策略,分为过滤法、包装法和嵌入法[4]。

随着数据增长,传统方法无法存储所有数据[5]。流特征是在流数据中动态生成的特征[6]。在线特征选择方法可针对新特征即时决定[7]。这些方法在大数据分析、物联网、实时机器学习和数据流挖掘等领域有广泛的应用。

在特征选择研究中,除了高预测精度和低时间复杂度外,稳定性也是一个关键问题[8]。不稳定的特征选择方法在训练数据微小变化或扰动时,选定的特征子集会发生显著变化的现象[9]。

稳定的特征选择在处理实时邮件流中特别重要。在图1在线垃圾邮件分类系统中,稳定性对流特征在线选择算法至关重要[10]。该系统通过筛选实时邮件中的特征,如不可信域名、邮件主题关键词、链接可疑性,以区分垃圾邮件和非垃圾邮件。

稳定的算法能在特征到达顺序或分布变化时保持选择相似的特征子集。例如,即使链接可疑性等特征实时变化,稳定算法仍将其视为重要,确保结果一致。反之,缺乏稳定性的算法对特征流变化敏感,导致所选特征子集差异较大,降低用户信任度。

实现稳定的流特征选择面临三大挑战:(1)特征流的随机性。特征生成顺序不固定。(2)实时冗余特征处理,影响特征集的稳定性。(3)样本扰动的影响,削弱特征的信息量,导致选择结果不稳定。

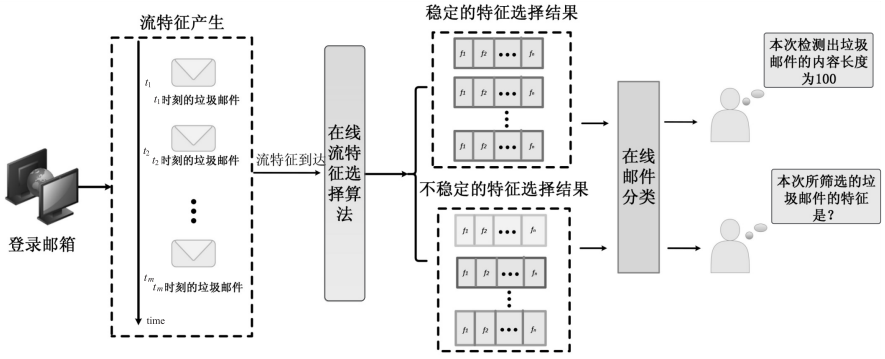

图1　流特征选择算法稳定性应用案例

为了获得稳定的特征选择结果,可采用集成学习的特征选择结果聚合策略。集成学习是通过整合多个模型的选择结果,形成最终稳定的特征子集。该文提出了一种新的稳定流特征在线选择算法(Multi-level Ensemble Learning Stream Feature Selection,MESFS)。该算法首先在数据集层面进行集成,通过极限学习机(Extreme Learning Machine,ELM)[11]对输入样本进行分组和特征映射,生成新的矩阵。然后,在特征选择层面。算法使用自适应阈值为每个特征分配权重,并通过多次迭代整合多个模型信息,以提高特征选择的稳定性。实验结果表明,该算法在多个真实数据集上的表现优于传统方法。该文的贡献主要包括:

(1)首次对流特征在线选择方法的稳定性问题进行定义和研究。

(2)在12个真实数据集进行了充分实验,结果表明,MESFS算法在准确性和稳定性方面优于最先进的流特征在线选择算法。

1　相关工作

迄今为止,已有多种方法来解决传统特征选择算法的不稳定性问题。图2中呈现了解决不同不稳定性来源的各种策略。

(1)特征信息策略:Liu等[12]根据评估标准衡量特征的重要性,以此筛选出稳定特征。Ginsburg[13]提出的非线性嵌入中的特征重要性(FINE)方法通过评估特征对低维空间中分类的贡献进行排序。

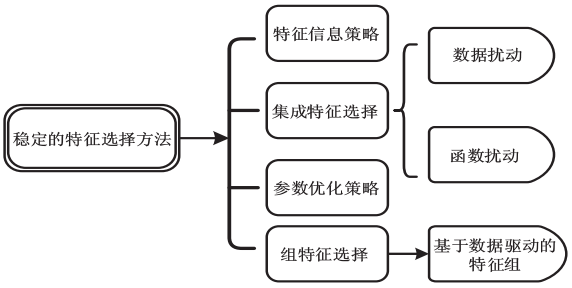

图2　稳定特征选择策略汇总框架

(2)参数优化策略:Isachenko和Strijovti[14]提出基于迭代加权最小二乘法的逻辑回归算法,该算法选择最优的参数集,具有低误差性和高稳定性。

(3)组特征选择策略:Jeitziner等[11,15]提出基于数据驱动的特征分组生成方法,利用聚类分析或密度估计来识别特征组。

(4)集成特征选择策略:集成学习利用"群体智慧"理念,认为集体决策优于个别专家[16]。Pes B[17]提出的传统稳定特征选择算法框架通常分为两阶段:数据扰动和函数扰动。在数据扰动阶段,Ranker $R_n(n \in \{1,2,\cdots,N\})$ 解释数据扰动。每个Ranker利用Bootstrap样本对特征进行排名,生成 M 个排名列表 $L=\{L_n^1,\cdots,L_n^m,\cdots,L_n^M\}$。通过预定义的阈值 t,从 L_n^m 中选择排名靠前的特征,以判断其重要性。

在函数扰动阶段,可以采用多个特征选择方法进行特征筛选,获得多个 Ranker 生成的特征列表,存储重要特征[18]。最后,通过聚合策略将 N 个特征列表组合成一个最优特征子集。

2　MESFS 算法的实现过程

2.1　问题定义

流特征在线稳定选择问题:假设目标动态数据集为 D,样本数量固定为 N,特征以无限流的形式生成。在时间戳 t 处,新流特征为 f_t。流特征在线选择的目标是在每个时间戳 t 筛选 f_t,以获得最佳特征子集。稳定性指的是算法在特征流略有变动时,保持对所选特征子集结果不变或相对一致。

该文采用 Sarah 等[19]提出的特征选择稳定性度量指标作为流特征在线选择算法的稳定性度量标准。该度量标准被定义为:

$$\varphi(M) = 1 - \frac{\frac{1}{J}\sum_{j=1}^{J} S_j^2}{\frac{k}{J}(1 - \frac{k}{J})} \tag{1}$$

式中,分子部分是特征选择过程中各个特征选择的样本方差的平均值,$S_j^2 = \frac{r}{r-1}\hat{p}_j(1-\hat{p}_j)$,$k$ 是选中的特征数量,J 是矩阵 $\boldsymbol{M}$ 中的特征数,并且 $\hat{p}_j = \frac{1}{r}\sum_{i=1}^{r} m_{ij}$($m_{ij}$ 是 $\boldsymbol{M}$ 的一个元素)。若某特征在不同的特征集合中的选择结果变化较大,则其 S_j^2 会较大,表示算法的不稳定性。分母部分表示特征选择的平均稳定性与特征的平均选择率之间的关系。将分子除以分母得到稳定性指标 $\Phi(M)$,其取值范围为 0 到 1。若 $\Phi(M)$ 接近 1 时,表示特征选择过程更稳定,结果更一致。若接近 0 时,表示过程不稳定,特征选择结果的变动性较大。

MESFS 算法总体流程:该算法的整体思路是利用 ELMD 函数对数据样本进行集成处理。在特征选择层面,通过多次迭代的自适应特征选择方法,得到特征子集。最后,基于特征排名得分和聚合策略确定最终特征选择结果。具体流程如图 3 所示。

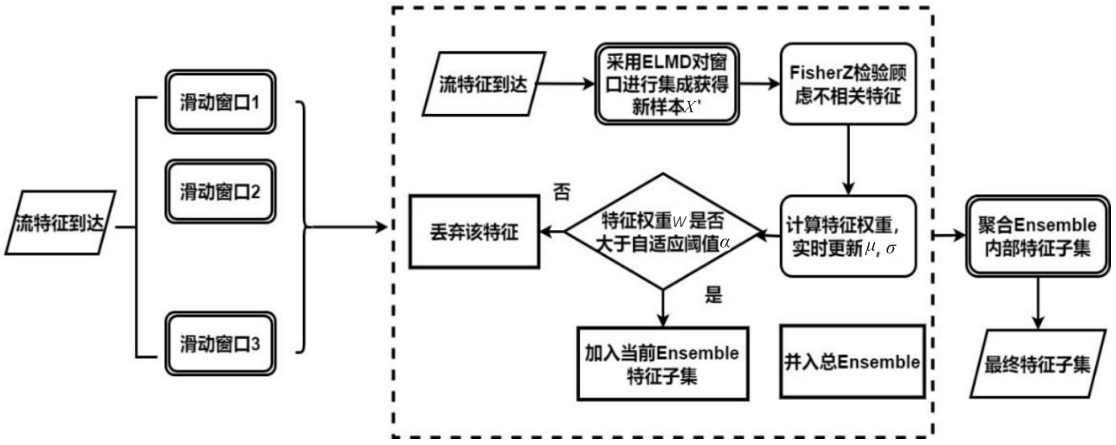

图 3　基于多层次集成学习的稳定流特征选择算法流程

2.2　数据集层面集成策略

算法在样本层面集成策略包括以下步骤:首先,用极限学习机(ELM)进行特征映射。其次,采用 DPC[20]聚类策略来选择每个簇的代表性样本。最后,通过复制与合并标签形成集成后的数据矩阵。

定义 1(特征矩阵):ELM 随机特征映射是指当输出映射 $f(x)$ 根据任意连续采样分布概率随机分配。数据集 X 包含 N 个样本,其中 $x_i \in R^d$。特征从高维空间 X 被映射到 L 维随机特征空间。ELM 利用随机特征映射 $f(x)$ 构建隐含层的权重矩阵 $\boldsymbol{H}$ 和输出矩阵 $\boldsymbol{F}$。矩阵 $\boldsymbol{H}$ 显示输入特征与隐含层的映射关系,反映特征提取结果。权重矩阵 $\boldsymbol{H}$ 为 $N \times (L+1)$ 的矩阵,$W_{(i,j)}$ 表示隐含层中的特征权重,N 为样本数量,$(L+1)$ 包含一个偏置项,矩阵 $\boldsymbol{H}$ 为:

$$\boldsymbol{H} = \begin{pmatrix} W_{(1,1)} & \cdots & W_{(1,L)} & W_{(1,b)} \\ \vdots & \ddots & \vdots & \vdots \\ W_{(N,1)} & \cdots & W_{(N,L)} & W_{(N,b)} \end{pmatrix}_{N\times(L+1)} \tag{2}$$

$\boldsymbol{F}$ 矩阵则反映了隐含层节点的输出,由隐含层通过激活函数进行非线性变换得到的特征值 $f_i(x_j)$ 组成。输出矩阵 $\boldsymbol{F}$ 形式化为:

$$\boldsymbol{F} = \begin{pmatrix} f_1(x_1) & \cdots & f_L(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_N) & \cdots & f_L(x_N) \end{pmatrix}_{N\times L} \tag{3}$$

随机特征映射 $f(x)$ 是由非线性激活函数 $A(S', S, x)$ 所实现的,该函数定义如下:

$$A(S', S, x) = 1/(1 + e^{(-(S'\cdot x + S))}) \tag{4}$$

输入权值 S' 和偏差 S 通过均匀概率分布随机生成。经过 M 次随机特征映射,得到 M 个输出矩阵 $\boldsymbol{F}^1$,

$F^2,\cdots,F^M$,形成投影空间。基于此空间,使用降维技术减少特征维度。因此,采用密度峰值识别的聚类方法(Density Peak Clustering,DPC)[20]将数据样本划分为相似特征的簇,以选择每个簇的代表性样本。因此,提出 ELMD 算法,将 ELM 特征映射和 DPC 聚类策略整合,实现无监督学习。具体流程如算法 1 所示。

算法 1:ELMD
输入:数据样本 x
输出:集成后新数据样本 P'
1. 初始化: $P' = \{\}$
2. 训练 ELM 获取参数对
3. 根据定义 1 获得投影空间 $P = \{F_1,F_2,\cdots,F_M\}$
4. 执行 DPC,获得簇的索引 I 和簇心 C
5. for each $C_k \in C$
6. $C_k' = \{P_k \in P_M \mid I(P_k) = i\}$
7. $R = C'(i)$
8. $P' = P' \cup R$
9. end for
10. return P'

ELMD 算法首先初始化 ELM 模型的参数,并用数据样本来训练模型,同时执行 M 次随机特征映射,获取投影空间(第 1 ~3 行);接着,使用 DPC 算法对投影空间进行聚类,得到簇的索引和中心。然后,从投影空间中选择当前索引为 i 的样本,添加到簇心集合 C'。同时,选择 C' 中索引为 i 的簇心作为代表性样本,放入集合 R,重复此过程,直到没有新的簇索引(第 5 ~9 行)。最后,将代表性样本集合添加至集合 P',完成样本层面的集成(第 10 行)。

2.3 特征选择层面集成策略

特征选择的目标是在每个时间步 t 挖掘稳定的鉴别特征。仅选择固定特征可能导致信息丢失和忽视特征与目标变量的关系。因此,在流特征场景中,需平衡特征的重要性与决策的不确定性。Haug 等人[21]提出了两个属性定理,实现稳定的特征加权。

定理 1(特征权重分配单调性)[21]:特征权重 ω_t 必须在重要性与不确定性方面都严格单调的函数。在 t 时刻,对于任意给定的两个特征,$f_i \neq f_j$,设定 $|\mu_{t_i}|$ 和 $|\mu_{t_j}|$ 为特征的重要性绝对度量,同时设定 $\sigma_{t_i},\sigma_{t_j} \geqslant 0$ 分别为特征的不确定性度量。并且必须具备以下两个条件:

(1)假定 $|\mu_{t_i}| = |\mu_{t_j}|$,等式 $\sigma_{t_i} \geqslant \sigma_{t_j} \Leftrightarrow \omega_{t_i} \leqslant \omega_{t_j}$。否则,如果 $\sigma_{t_i} < \sigma_{t_j}$,表示对特征 f_i 的不确定性低于特征 f_j,那么成立 $\omega_{t_i} > \omega_{t_j}$,反之亦然。

(2)假定 $\sigma_{t_i} = \sigma_{t_j}$,以下成立:$|\mu_{t_i}| \geqslant |\mu_{t_j}| \Leftrightarrow \omega_{t_i} \geqslant \omega_{t_j}$。否则,如果 $|\mu_{t_i}| < |\mu_{t_j}|$,表示特征 f_i 所具有的重要性信息较低于特征 f_j,那么成立 $\omega_{t_i} < \omega_{t_j}$,反之亦然。

定理 2(稳定特征排名)[21]:对于一个稳定的目标分布,特征权重最终必须产生一个一致的排名。设定 $R(\omega_t)$ 为根据特征权重在时间步长 t 时的特征排名。假设 $\exists \bar{t}$,使得 $P(y_t \mid x_t) = P(y_{t+1} \mid x_{t+1})$ 对于任意 $t \geqslant \bar{t}$ 成立。当 $t \geqslant \bar{t} \to \infty$ 时,则认为 $R(\omega_t) = R(\omega_{t+1})$。

基于上述定理,引出该文特征权重计算方案的核心:假设每个输入特征都有一个单一的聚合参数。在时间戳 t 上,用 μ_t 和 σ_t 分别表示特征的预估重要性和不确定性。该算法的目标是在特征具有高重要性时最大化特征权重 ω_t,并且在高不确定性下最小化权重。通过采用这种策略,可以获得既具有良好分类能力又具有稳定性的最优特征子集。该策略所使用的目标函数定义如下:

$$\arg\max_{\omega_r} = \sum_{t=1}^{T} \omega_t(\mu_t^2 - \lambda_m\sigma_t^2 - \lambda_n\omega_t) \tag{5}$$

公式 5 中指定了两个缩放因子,分别为 $\lambda_m \geqslant 0$ 和 $\lambda_n \geqslant 0$,它们分别用于调整不确定性惩罚项和正则化项。在特征选择中,这两种惩罚影响加权方案的敏感性。正则化缩放因子 λ_n 通常设为 0.01,是控制模型复杂度和过拟合的经验性选择。

在流特征场景中,假设特征 f_j 在单个时间步长 t 的特征权重为 ω_{t_j}。为了最大化 ω_{t_j},将选择在零处计算偏导数。ω_{t_j} 的偏导数计算公式定义为:

$$\frac{\partial}{\partial\omega_{tj}} = \mu_{tj}^2 - \lambda_m\sigma_{tj}^2 - 2\lambda_n\omega_{tj} = 0$$

$$\Leftrightarrow \omega_{tj} = \frac{\mu_{tj}^2 - \lambda_m\sigma_{tj}^2}{2\lambda_n} \tag{6}$$

引理 1:公式 10 产生的特征权重满足定理 1 所规定的单调性。

证明:假定两个特征 $f_i \neq f_j$,$|\mu_i| = |\mu_j|$,则等式 $\sigma_i \geqslant \sigma_j \Leftrightarrow \omega_i \leqslant \omega_j$ 成立。

$\sigma_i \geqslant \sigma_j$

$\Leftrightarrow a - b\sigma_i^2 \leqslant a - b\sigma_j^2 (a,b \geqslant 0)$

$a = \frac{\mu_i^2}{2\lambda_n} = \frac{\mu_j^2}{2\lambda_n}, b = \frac{\lambda_m}{2\lambda_n}$

$\Leftrightarrow \omega_i \leqslant \omega_j$

假定 $\sigma_i = \sigma_j$,则等式 $|\mu_i| \geqslant |\mu_j| \Leftrightarrow \omega_i \geqslant \omega_j$ 成立。

$|\mu_i| \geqslant |\mu_j|$

$\Leftrightarrow b\mu_i^2 - c \geqslant b\mu_j^2 - c (b,c \geqslant 0)$

$b = \frac{1}{2\lambda_n}, c = \frac{\lambda_m}{2\lambda_n}\sigma_i^2 = \frac{\lambda_m}{2\lambda_n}\sigma_j^2$

$\Leftrightarrow \frac{1}{2\lambda_n}\mu_i^2 - \frac{1}{2\lambda_n}\lambda_m\sigma_i^2 \geqslant \frac{1}{2\lambda_n}\mu_j^2 - \frac{1}{2\lambda_n}\lambda_m\sigma_j^2$

$\Leftrightarrow \omega_i \geqslant \omega_j$

为了实现完整的特征层面集成策略,需解决以下两个问题:(1)如何选择合适的自适应阈值 α ,以动态确定特征选择或丢弃标准;(2)如何结合不同集成成员的特征子集信息,获得最终的特征选择结果。

针对问题 1,采用自适应调整阈值方案,基于特征权重分布的变化而动态调整阈值,公式为 $\alpha = \mu_\omega - k * \sigma_\omega$ 。其中 μ_ω 是权重均值, σ_ω 是标准差, k 是自定义倍数。通过在线更新新特征来实时更新 μ_ω 和 σ_ω ,灵活调整阈值的灵敏度。

针对问题 2,在特征选择层面,关键是聚合特征子集。算法利用滑动窗口处理新特征,窗口大小 $\boldsymbol{H}$ 根据交叉验证的折数和样本量确定,以提高稳定性。在每个窗口内,计算新特征的权重并存储在数组 W 中。特征排名得分 $G = [g_1, g_2, \cdots, g_p]$,特征子集为 $M = [m_1, m_2, \cdots, m_p]$,其中 m_i 表示第 i 个特征是否被选择。因此,在候选特征子集上的特征排名得分加权平均值 $\bar{G}$ 定义为:

$$\bar{G} = \frac{\sum_{i=1}^{p} g_i * m_i}{\sum_{i=1}^{p} m_i} \tag{7}$$

特征选择层面的集成策略通过加权平均的机制,赋予不同特征不同角色,反映了 Ensemble 成员在特征选择中的相对重要性。为此,基于样本层面和特征选择层面的集成学习方案,提出 MESFS 算法。实现细节如算法 2。

算法 2:MESFS

输入:自适应阈值倍数 k ,不确定性惩罚因子 λ_m 在 t 时刻到达的流特征 f_t

输出:当前 Ensemble 所选特征子集 S_E ;最终聚合结果特征子集 S_A

1. 初始化: $P = \{\}$, $S_E = \{\}$, $S_A = \{\}$
2. for each f_t
3. $P = P \cup f_t$
4. if $|P| > H$
5. 从 P 中移除旧特征,维持窗口大小 $\boldsymbol{H}$
6. else
7. $X' = \mathrm{ELDM}(P)$
8. for each $f_t \in X'$
9. 计算 f_t 在此处的梯度与边际似然
10. 根据梯度与边际似然更新参数 μ_t, σ_t
11. 计算特征权重 ω_{f_t}
12. $\alpha_{f_t} = \mu_{\omega_t} - k * \sigma_{\omega_t}$
13. if $\omega_{f_t} < \alpha_{f_t}$
14. 丢弃 f_t
15. else
16. $S_{E_i} = S_{E_i} \cup \{f_t\}$
17. end for
18. $S_E = S_E \cup S_{E_i}$
19. end for
20. 计算 $\bar{G}$ 调整自适应候选特征数 m

$S_A = \{S_{E_i} \mid S_{E_i} \in S_E, 1 \leqslant i \leqslant m\}$

21. return S_A

MESFS 算法首先初始化候选特征子集和大小为 W 的滑动窗口;动态更新滑动窗口并对其样本进行集成获得新样本(第 4 ~ 8 行);计算新到达特征 f_t 的梯度和边际似然,更新特征 f_t 的重要性和不确定性,根据式 8 计算 f_t 的特征权重,若大于自适应阈值 α 则选择,否则丢弃(第 10 ~ 15 行);最后,使用候选特征子集上的特征排名得分的加权平均值来聚合最终的特征子集。

2.4　算法时间复杂度分析

本节对 MESFS 算法进行时间复杂度分析,包括数据集成、特征选择和特征子集聚合。样本层面涉及 ELM 模型训练,其时间复杂度为 $O(M^2H + H^2)$ 。

在特征选择和子集聚合中,特征排名得分计算和子集聚合的复杂度分别为 $O(P * \log(P))$ 与 $O(P)$,其中 P 为所到达特征数。外层循环由集成成员数量 K 决定,时间复杂度为 $O(K)$ 。MESFS 算法的总时间复杂度为: $O(M^2H + H^2) + O(P * \log(P)) + O(K)$ 。

3　实验及结果分析

3.1　实验设置

3.1.1　实验数据集

为验证 MESFS 算法的稳定性与预测准确性,实验在 12 种真实世界的数据集上与 9 种对比算法进行比较。实验使用了 5 种来自 UCI 的微型数据集(ACA、LETTER、WPBC、SONA-R 和 ADA)与 4 种传统静态特征选择方法(GPA_FS[22]、ReliefF[23]、MI[24] 和 CFS[25])进行比较。剩余的 7 个高维数据集(DEXTER、MADELON、LUNA、DLBCL、LEUKEMIA_3C、LEUKEMIA_4C 和 CNS)则与 5 种流特征在线选择算法(OSFSMI[26]、SAOLA[27]、Fast_OSFS[28]、OSFS_3WD[29] 和 Alpha_Investing[30])进行比较。有关实验数据集的详细信息见表 1。

表 1　真实世界数据集相关描述

序号	数据集	样本数	特征数	特征类型
1	ACA	690	14	heterogeneous
2	LETTER	20 000	16	Integer
3	SONAR	208	60	real
4	ADA	4 562	48	Integer
5	WPBC	198	33	real
6	DLBCL	77	7 129	Integer
7	COLON	62	2 000	real
8	CNS	60	7 129	Integer

续表1

序号	数据集	样本数	特征数	特征类型
9	LEUKEMIA_4c	203	7 129	Integer
10	LEUKEMIA_3c	181	7 129	Integer
11	LUNA	181	5 000	real
12	MADELON	2 600	500	Integer

3.1.2 实验评价指标

(1)平均预测精度:指模型正确预测的样本数占总样本数的比例,通常以百分比(%)形式表示。

(2)平均算法稳定性:评估特征选择算法在不同数据集和运行条件下表现一致性的关键指标,通常以(%)形式表示。

(3)平均运行时间:指算法在执行特定任务时所需的时间平均值,通常以秒(s)表示。

3.1.3 实验评价方法

统计检验方法:该文采用Frideman[31]检验验证MESFS算法和对比算法是否存在显著差异。该检验基于样本中的等级数据进行计算。其计算方法如下所示:

$$X_F^2 = \frac{12D}{K(K+1)}\left[\sum_{j=1}^{K} \mathrm{AR}_j^2 - \frac{K(K+1)^2}{4}\right] \tag{8}$$

$$\mathrm{AR}_j^2 = \frac{1}{D}\sum_{i=1}^{D} r_i^j \tag{9}$$

Nemenyi[31]检验根据预设的置信水平(通常为95%或99%)确定临界值(Critical Difference,CD)。该值表示在该置信水平下,两个算法平均排名差异的临界点。其计算方法如下:

$$\mathrm{CD} = q_\alpha \sqrt{\frac{K(K+1)}{12D}} \tag{10}$$

3.1.4 实验平台

实验平台配置:操作系统为Windows10,中央处理器为AMD-Ryzen 5800X,运行内存为16 GB,实验所用的IDE以及编程语言均为MATLAB。

3.2 实验结果分析

3.2.1 传统特征选择算法对比结果

为了证明MESFS算法的稳定性,本节将其与4种传统特征选择方法进行实验对比。对比算法包括稳定的演化特征选择算法(GPA_FS)[22]、基于曲率的特征选择算法(CFS)[25]以及两个代表性的算法(ReliefF[23]和MI[24])。实验选取5个低维数据集,所有对比算法均使用默认参数值。

表2展示了MESFS算法与其他算法在特征选择稳定性上的对比。通过Friedman检验,稳定性统计检验值分别为1.701E-3, Nemenyi检验得到临界差值(CD)为2.514 7。分析结果如下:

MESFS相比GPA_FS,后者容易陷入局部最优,特别是在低维数据集中,MESFS的稳定性提升了20~30百分点。与ReliefF相比,MESFS在稳定性上优势明显,因ReliefF依赖局部搜索,无法获得全局最优特征子集。MI算法稳定性差,因其计算依赖有限数据集的联合概率估计,造成偏差。MESFS在稳定性上优于CFS,因为CFS对曲率阈值和邻域参数敏感。

表2 Φ度量标准下的算法稳定性 %

序号	MESFS	GPA_FS	ReliefF	MI	CFS
1	76.67	67.92	58.92	71.67	51.31
2	89.33	86.67	71.33	71.33	74.31
3	85.42	69.30	35.19	66.67	48.15
4	94.86	66.43	71.43	90.81	84.76
5	77.15	36.85	38.10	95.00	14.21
均值	84.69	65.45	54.99	79.10	54.55
平均排名	1.2	3.2	4.1	2.5	4

3.2.2 流特征在线选择算法对比结果

在本小节中,将采用不同度量指标对比该文提出的流特征在线选择算法(MESFS)与最新的流特征在线选择算法,包括OSFSMI[26]、SAOLA[27]、Fast_OSFS[28]、OSFS_3WD[29]和Alpha_Investing[30]。实验将在多个不同类型和维度的数据集上进行,统计结果包括算法的稳定性、预测精度、运行时间。

图4展示了MESFS算法与其他对比算法在特征选择稳定性和KNN分类器平均预测精度上的统计检验结果。根据表3~6,可以得出以下观察:

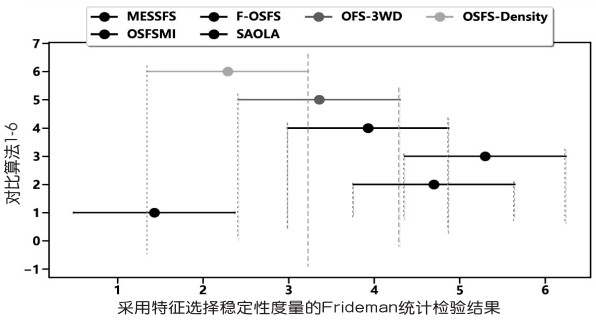

图4 Φ度量标准下算法稳定性的统计检验结果

OSFSMI算法在稳定性和分类准确性上显著低于MESFS,因其未能有效应对数据集的不平衡性和动态变化。MESFS与Fast_OSFS的特征选择稳定性差异明显,后者对特征分布变化适应性差。尽管MESFS与SAOLA在预测精度上无明显差异,但SAOLA仅分析两两特征相关性,导致稳定性较低。OSFS_3WD对微小数据变化敏感,因此稳定性差于MESFS。Alpha_Investing在高维数据集处理效率较低,进而算法稳定性差。

综上所述,MESFS算法在8个高维数据集上展现了最高的稳定性,得益于采用多层次集成学习的方法,提供了更可靠的特征选择能力。

表 3　Φ 度量标准下的算法稳定性　%

序号	MESFS	OSFSMI	FOSFS	SAOLA	OFS3WD	AINV
6	61.01	24.94	2.32	31.55	40.27	20.78
7	88.30	43.83	7.57	52.86	18.10	96.00
8	56.20	25.94	10.68	17.4	9.53	27.17
9	97.30	18.68	26.04	40.02	56.76	41.23
10	79.29	32.36	23.86	35.26	60.09	37.70
11	71.42	25.61	38.11	58.91	58.91	59.94
12	77.52	42.54	66.34	31.81	71.52	79.80
均值	75.86	30.56	24.98	38.26	45.03	51.80
平均排名	1.43	4.7	5.3	3.93	3.36	2.29

表 4　KNN 分类器上的平均预测精度　%

序号	MESFS	OSFSMI	FOSFS	SAOLA	OFS3WD	AINV
6	88.13	70.13	88.13	88.67	81.33	79.33
7	71.67	61.67	75.00	70.00	78.33	45.00
8	75.00	56.67	55.00	56.67	58.33	58.33
9	71.43	31.43	87.15	85.71	88.57	68.57
10	68.57	48.57	77.14	80.00	84.29	81.43
11	96.67	77.22	93.89	88.67	95.66	80.56
12	57.46	50.62	54.77	52.65	70.73	51.92
均值	76.97	56.62	75.85	74.62	79.61	66.45
平均排名	2.5	5.64	3.36	3.36	1.79	4.36

表 5　算法在不同数据集上的运行时间　s

序号	MESFS	OSFSMI	FOSFS	SAOLA	OFS3WD	AINV
6	0.396 7	1.368 8	0.384 7	0.526 1	0.269 1	1.719
7	0.453 2	1.045 2	0.101 3	0.107 1	0.084 6	0.025
8	0.234	0.654 6	0.307 4	0.333 5	0.274 8	1.367
9	0.694	0.588 1	0.443 1	0.643 9	0.318 6	1.786
10	0.368 8	0.524 6	0.437 4	0.591 7	0.290 8	1.727
11	3.078 8	2.079 3	1.027 9	1.944 9	0.206 5	0.303
12	0.147	0.039 9	0.072 9	0.043 8	0.043 1	0.037
均值	0.767 5	0.900 7	0.396 4	0.598 7	0.212 5	0.995
平均排名	3	2.71	4	2.86	5.43	3

4　结束语

当前的流特征在线选择算法通常关注时间效率、可扩展性和高精度,但忽视了结果的稳定性。为改善这一问题,该文提出了一种流特征在线选择算法(MESFS),在数据集和特征选择层面引入多层次集成策略。

通过在 4 个低维数据集和 8 个高维数据集上的对比实验表明,MESFS 算法在流特征场景中能获得稳定的特征选择结果,并在 SVM 和 CART 分类器上实现良好的分类精度。

参考文献:

[1] SARWAR T,SEIFOLLAHI S,CHAN J,et al. The secondary use of electronic health records for data mining: data characteristics and challenges[J]. ACM Computing Surveys,2022,55(2):1-40.

[2] DOGAN A,BIRANT D. Machine learning and data mining in manufacturing[J]. Expert Systems with Applications,2021,166:114060.

[3] 许华杰,刘冠霆,张　品,等. 采用动态相关度权重的特征选择算法[J]. 计算机工程与应用,2024,60(4):89-98.

[4] 许召召,申德荣,聂铁铮,等. 融合信息增益比和遗传算法

的混合式特征选择算法[J]. 软件学报,2022,33(3):1128-1140.
[5] BENSAID F, ALIMI A M. Online feature selection system for big data classification based on multi-objective automated negotiation [J]. Pattern Recognition, 2021, 110:107629.
[6] ALNUAIMI N, MASUD M M, SERHANI M A, et al. Streaming feature selection algorithms for big data: a survey [J]. Applied Computing and Informatics, 2020, 18(1/2): 113-135.
[7] 张小清,王晨曦,吕 彦,等. 基于 ReliefF 的层次分类在线流特征选择算法[J]. 计算机应用,2022,42(3):688-694.
[8] 刘 艺,曹建军,刁兴春,等. 特征选择稳定性研究综述[J]. 软件学报,2017,29(9):2559-2579.
[9] KHAIRE U M, DHANALAKSHMI R. Stability of feature selection algorithm: a review [J]. Journal of King Saud University - Computer and Information Sciences, 2022, 34(4):1060-1073.
[10] MANSOOR R, JAYASINGHE N D, MUSLAM M M A. A comprehensive review on email spam classification using machine learning algorithms [C]//2021 international conference on information networking (ICOIN). Jeju Island: IEEE, 2021:327-332.
[11] WANG J, LU S, WANG S H, et al. A review on extreme learning machine[J]. Multimedia Tools and Applications, 2022, 81(29):41611-41660.
[12] LIU Y, DIAO X, CAO J, et al. Evolutionary algorithms' feature selection stability improvement system[C]//Bio-inspired computing: theories and applications: 12th international conference, BIC-TA 2017. Harbin: Springer, 2017:68-81.
[13] GINSBURG S B, LEE G, ALI S, et al. Feature importance in nonlinear embeddings (FINE): applications in digital pathology[J]. IEEE Transactions on Medical Imaging, 2015, 35(1):76-88.
[14] ISACHENKO R V, STRIJOV V V. Quadratic programming optimization with feature selection for nonlinear models[J]. Lobachevskii Journal of Mathematics, 2018, 39:1179-1187.
[15] JEITZINER R, CARRIERE M, ROUGEMONT J, et al. Two-tier mapper, an unbiased topology-based clustering method for enhanced global gene expression analysis [J]. Bioinformatics, 2019, 35(18):3339-3347.
[16] 刘洪涛,李 航,王 进,等. 基于标签特定特征的多目标回归稀疏集成方法[J]. 电子学报,2020,48(5):906-913.
[17] PES B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains[J]. Neural Computing and Applications, 2020, 32(10):5951-5973.
[18] SALMAN R, ALZAATREH A, SULIEMAN H. The stability of different aggregation techniques in ensemble feature selection[J]. Journal of Big Data, 2022, 9(1):1-23.
[19] NOGUEIRA S, SECHIDIS K, BROWN G. On the stability of feature selection algorithms[J]. Journal of Machine Learning Research, 2018, 18(174):1-54.
[20] CHEN Y, HU X, FAN W, et al. Fast density peak clustering for large scale data based on KNN[J]. Knowledge-Based Systems, 2020, 187:104824.
[21] HAUG J, PAWELCZYK M, BROELEMANN K, et al. Leveraging model inherent variable importance for stable online feature selection[C]//Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. CA: ACM, 2020:1478-1502.
[22] SALESI S, COSMA G. Generalisation power analysis for finding a stable set of features using evolutionary algorithms for feature selection[J]. Knowledge-Based Systems, 2021, 231:107450.
[23] CUI X, LI Y, FAN J, et al. A novel filter feature selection algorithm based on relief[J]. Applied Intelligence, 2022, 52(5):5063-5081.
[24] VERGARA J R, ESTÉVEZ P A. A review of feature selection methods based on mutual information[J]. Neural Computing and Applications, 2014, 24:175-186.
[25] ZUO Z, LI J, XU H, et al. Curvature-based feature selection with application in classifying electronic health records[J]. Technological Forecasting and Social Change, 2021, 173:121127.
[26] RAHMANINIA M, MORADI P. OSFSMI: online stream feature selection method based on mutual information[J]. Applied Soft Computing, 2018, 68:733-746.
[27] YU K, WU X, DING W, et al. Scalable and accurate online feature selection for big data[J]. ACM Transactions on Knowledge Discovery from Data, 2016, 11(2):1-39.
[28] WU X, YU K, DING W, et al. Online feature selection with streaming features [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(5): 1178-1192.
[29] ZHOU P, ZHAO S, YAN Y, et al. Online scalable streaming feature selection via dynamic decision [J]. ACM Transactions on Knowledge Discovery from Data, 2022, 16(5):1-20.
[30] JI T, GUO X, LI Y, et al. Multi-label online streaming feature selection algorithms via extending alpha-investing strategy[C]//International conference on big data analytics and knowledge discovery. [s. l.]: Springer International Publishing, 2022:112-124.
[31] LÜDECKE D, BEN-SHACHAR M S, PATIL I, et al. Performance: an R package for assessment, comparison and testing of statistical models [J]. Journal of Open Source Software, 2021, 6(60):31-39.