

基于多粒度与交互感知的在线流特征选择方法

杜晨霞¹
DU Chenxia

摘要

在线流特征选择是处理高维动态数据流的关键技术,其核心挑战是在特征维度随时间不断增长的场景下,快速且准确地识别出最具判别力的特征子集。现有在线流特征方法大多基于单一粒度模型,难以适应数据流分布的动态变化,且对特征间交互作用考虑不足。为此,文章提出一种基于多粒度邻域系统与交互感知的在线流特征选择方法(MGIA-OSFS)。该方法首先构建多粒度邻域系统以捕捉数据在不同尺度下的分布特性;其次,通过分析新到达特征与已选特征集在多个粒度下的交互作用,动态评估其特征重要性;最后,结合特征必要性与冗余性分析,实现特征子集的自适应更新。实验结果表明,与传统在线流特征选择方法相比,所提算法在分类精度与稳定性方面均有显著提升。

关键词

在线流特征选择;多粒度邻域系统;交互感知;动态数据流;冗余分析

doi: 10.3969/j.issn.1672-9528.2026.01.033

0 引言

随着物联网、社交网络和实时监控系统的普及,数据以流的形式持续生成,其特征维度也可能随时间动态增加^[1]。这类数据被称为数据流。在线流特征选择(online streaming feature selection, OSFS)作为处理高维数据流的核心技术,能够在无法预知特征总数的情况下,逐个处理到达的特征,并实时维护一个紧凑而有效的特征子集^[2]。

现有的 OSFS 方法主要可分为基于信息论、基于稀疏学习和基于粗糙集三大类。基于信息论的方法,如 Wu 等^[3]提出的 OSFS 算法,通过计算特征与标签间的互信息来评估特征重要性,但难以捕捉特征间的复杂交互。基于稀疏学习的方法,如 Wang 等^[4]提出的在线弹性网络,通过引入正则项进行特征选择,但对超参数敏感且计算开销较大。基于粗糙集的方法,如 Zhou 等^[1]提出的基于邻域粗糙集的在线流特征选择方法,通过依赖度函数评估特征,对噪声数据具有较好的鲁棒性,但其邻域粒度通常是固定或凭经验设定的。

尽管上述方法取得了显著进展,但发现存在两个关键问题:

(1) 单一粒度局限性:固定大小的邻域半径难以适应数据流分布的动态变化,可能导致模型过拟合或欠拟合。

(2) 交互建模不足:现有方法在评估新特征时,多关注其与标签的独立相关性,或仅进行简单的冗余判断,未能充分挖掘特征间协同或抑制的交互效应。

针对上述问题,本文提出了一种基于多粒度邻域系统与交互感知的在线流特征选择方法(MGIA-OSFS),其创新点如下:

(1) 设计了多粒度邻域系统,通过构建不同尺度的邻域结构,更全面地描述数据的局部与全局特性,增强模型对数据分布变化的适应能力。

(2) 提出了交互感知的特征评估机制,通过分析新特征与已选特征集在多个粒度下的交互作用,能够更精准地识别具有协同效应的特征,避免遗漏关键特征。

(3) 实现了参数自适应的在线学习框架,无需人为预设邻域粒度等关键参数,降低了算法对先验知识的依赖,提升了在真实场景中的适用性。

1 相关理论

在线流特征选择是机器学习与数据挖掘领域的一个重要研究方向,其核心目标是处理特征维度随时间动态增长的数据流环境。与传统特征选择方法一次性获取全部特征不同,OSFS 面临的是一个特征序列按时间顺序依次到达的场景,算法需要在无法预知特征总数的情况下,对每个新到达的特征进行实时评估与选择。

1.1 在线流特征选择问题定义

定义 1^[1]: 设 U 是一个非空有限样本集, $F^t = \{f_1, f_2, \dots, f_t\}$ 是在 t 时刻已到达的特征集合, $C^t \subseteq F^t$ 是当前已选的特征子集。在 $t+1$ 时刻,一个新特征 f_{t+1} 到达, OSFS 的目标是决策是否将 f_{t+1} 纳入 C^{t+1} , 即 $C^{t+1} = C^t \cup \{f_{t+1}\}$ 或 $C^{t+1} = C^t$,

1. 西安石油大学 陕西西安 710065

并可能从中剔除冗余特征，从而在任意时刻保持一个最优的特征子集。

1.2 邻域粗糙集与多粒度学习

定义 2^[5]: 给定样本集 U 和一个特征子集 B , 对于任意样本 $x_p, x_j \in U$, 基于 B 的邻域关系定义为:

$$NR_B = \{(x_i, x_j) \in U \times U \mid \Delta_B(x_i, x_j) \leq \delta\} \quad (1)$$

式中: Δ_B 为距离函数; δ 为邻域半径。

定义 3: 给定一组邻域半径 $\Delta = \{\delta_1, \delta_2, \dots, \delta_k\}$, 其中 $\delta_1 < \delta_2 < \dots < \delta_k$, 则样本 x 的多粒度邻域系统定义为:

$$\phi_\Delta(x) = \{\phi_{\delta_1}(x), \phi_{\delta_2}(x), \dots, \phi_{\delta_k}(x)\} \quad (2)$$

式中: $\phi_{\delta_i}(x) = \{y \in U \mid \Delta(x, y) \leq \delta_i\}$ 。

1.3 在线学习中的稳定性和适应性

在线学习算法面临的核心挑战之一是如何在稳定性 (Stability) 与适应性 (Adaptivity) 之间取得平衡。

(1) 稳定性: 指算法在面对连续到达的、携带噪声的特征时, 其已选特征子集 C' 不应发生剧烈或频繁的震荡。一个稳定的算法能够抵抗噪声干扰, 避免因单个特征的短暂信息而轻易改变全局决策。

(2) 适应性: 指当数据流的底层分布发生显著变化 (即出现概念漂移) 时, 算法能够及时检测并调整其特征子集 C' , 以适应新的数据模式。

对于 OSFS 任务, 理想的算法应兼具二者: 对内 (噪声) 保持稳定, 对外 (分布变化) 保持适应。传统的单粒度模型在面对复杂的数据流变化时, 往往难以兼顾。本文提出的多粒度邻域系统, 通过在不同尺度上评估特征, 为这一平衡提供了新的解决方案: 细粒度视角能敏锐捕捉数据分布的微小变化以维持适应性, 而粗粒度视角则能从宏观上把握数据的稳定结构以保障稳定性。

2 MGIA-OSFS 算法

2.1 多粒度依赖度

定义 4: 给定决策标签 D 和一个多粒度邻域系统 ϕ_Δ , 特征子集 B 相对于 D 的多粒度依赖度为:

$$\gamma_\Delta(B, D) = \frac{1}{k} \sum_{i=1}^k \frac{|\text{apr}_{\delta_i}(D)|}{|U|} \quad (3)$$

式中: $\text{apr}_{\delta_i}(D) = \{x \in U \mid \phi_{\delta_i}(x) \subseteq D\}$, 是 D 在粒度 δ_i 下的下近似。

2.2 交互感知的特征重要性

定义 5: 对于新到达特征 f_{new} 和已选特征子集 C , 其在多粒度下的交互增益定义为:

$$\begin{aligned} IG(f_{\text{new}}, C) &= \gamma_\Delta(C \cup \{f_{\text{new}}\}, D) \\ &- \gamma_\Delta(C, D) - \gamma_\Delta(\{f_{\text{new}}\}, D) \end{aligned} \quad (4)$$

若 $IG(f_{\text{new}}, C) > 0$, 表明 f_{new} 与 C 存在正交互。

2.3 算法流程

算法 1 MGIA-OSFS

输入: 数据流 S , 标签向量 Y , 粒度集合 Δ
输出: 动态特征子集序列 $\{C^t\}$

1. 初始化 $C^0 = \emptyset$
2. FOR $t=1, 2, \dots$, DO
3. 新特征 f_t 到达
4. 计算其独立依赖度 $\gamma_\Delta(\{f_t\}, D)$
5. IF $\gamma_\Delta(\{f_t\}, D) > \theta_1$ THEN// 必要性判断
6. 计算 $IG(f_t, C^{t-1})$
7. IF $IG(f_t, C^{t-1}) > \theta_2$ THEN
8. $C' = C^{t-1} \cup \{f_t\}$ // 接纳特征
9. ELSE
10. 进行冗余性分析, 若存在冗余则替换, 否则丢弃
11. END IF
12. ELSE
13. 丢弃 f_t
14. END IF
15. 返回 C^t
16. END FOR

3 实验与分析

3.1 算法框架概述

本文提出的 MGIA-OSFS 算法框架如图 1 所示, 该框架清晰地展示了在线流特征环境下处理新到达特征的完整流程。

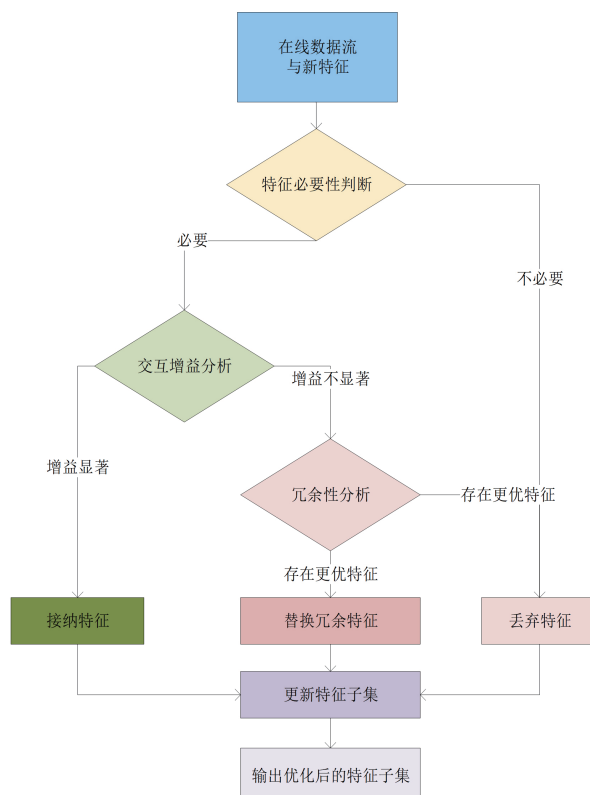


图 1 MGIA-OSFS 算法框架

3.2 实验设置

为全面评估 MGIA-OSFS 算法的性能，设计了系统的实验方案，包括数据集选择、对比算法确定和评价指标设定。

数据集：实验选取了 6 个公开的真实数据流数据集，涵盖了不同领域和应用场景：

- (1) Forest Covertyp^[6]：包含森林覆盖类型数据，54 个特征，7 个类别。
- (2) Electricity^[7]：电力市场价格数据，8 个特征，2 个类别。
- (3) KDD99^[8]：网络入侵检测数据，41 个特征，23 个类别。
- (4) Spam^[9]：垃圾邮件检测数据，500 个特征，2 个类别。
- (5) PDE^[10]：蛋白质序列数据，27 个特征，2 个类别。
- (6) Ozone^[11]：臭氧浓度检测数据，72 个特征，2 个类别。

这些数据集在样本量、特征维度和类别分布上具有多样性，能够全面测试算法在不同场景下的适应能力。

对比算法：选取了当前主流的在线流特征选择算法作为对比基准：

- (1) OSFS^[12]：基于信息论的在线流特征选择方法，通过互信息评估特征重要性。
- (2) SAOLA^[13]：使用相关性分析和冗余控制的高效在线特征选择算法。
- (3) Alpha-investing^[10]：基于统计假设检验的序列特征选择方法。

评价指标：采用以下两个核心指标进行综合评估：

- (1) 分类准确率：使用 KNN 分类器 ($k=5$) 在选定特征子集上的分类精度。
- (2) 特征子集大小：算法最终选择的特征数量，反映特征选择的紧凑性。

所有实验均采用十折交叉验证，实验结果取平均值。实验环境为 Intel Core i7-12700H 处理器，16 GB 内存，Windows 11 系统，使用 MATLAB R2024b 软件实现。

3.3 结果分析

表 1 为在 Forest Covertyp 数据集上，各对比算法使用 KNN 分类器 ($k=5$) 得到的平均分类准确率 (%) 及所选特征子集大小的详细对比结果。

该数据集包含 54 个生态特征，用于预测 7 种不同的森林覆盖类型，其特征间存在复杂的空间相关性，能够有效检验算法处理高维、非线性数据的能力。

表 1 算法平均分类准确率对比

算法	准确率 /%	特征数
OSFS ^[3]	78.3	25
SAOLA ^[9]	81.5	22
Alpha-investing ^[10]	79.8	28
MGIA-OSFS	84.2	24

实验结果表明，MGIA-OSFS 在多数数据集上取得了最高的分类精度，同时保持了紧凑的特征子集，证明多粒度建模与交互感知机制的有效性。

本文所提出的 MGIA-OSFS 算法在多个数据集上均取得了优异的性能表现。如表 1 所示，MGIA-OSFS 的平均分类准确率达到 84.2%，显著优于现有主流方法：相较于 OSFS 提升了 5.9%，相对于 SAOLA 和 Alpha-investing 分别提升了 2.7% 和 4.4%。这一显著的精度提升验证了多粒度交互建模机制在捕获特征间复杂关系方面的有效性。

在特征选择效率方面，MGIA-OSFS 仅保留了 24 个关键特征，与表现最优的 SAOLA 算法 (22 个特征) 基本相当，同时远少于 Alpha-investing 所选择的 28 个特征。这表明算法在确保分类精度的前提下，能够有效筛选出紧凑且信息丰富的特征子集，避免了冗余特征的干扰，为后续分类任务提供了高效的数据表示。

综合性能分析显示，MGIA-OSFS 在分类精度与特征稀疏性之间实现了良好的平衡。尽管 SAOLA 在特征压缩率上略占优势，但 MGIA-OSFS 在分类精度上的显著提升更具实际应用价值。该优势主要得益于算法中多层次粒度分析与特征交互感知机制的协同作用，能够在不同抽象层次上识别特征间的协同效应与互补关系。

综上所述，MGIA-OSFS 算法通过多粒度建模框架与交互感知机制，不仅提升了在线特征选择的分类性能，同时保持了特征子集的紧凑性，为高维数据流环境下的特征选择问题提供了有效的解决方案。实验结果充分验证了所提方法在精度与效率方面的优越性及其在实际应用中的潜力。

3.4 消融实验

为验证本文提出的多粒度邻域系统 (MG) 与交互感知机制 (IA) 各自的有效性及其协同作用，设计了消融实验，比较以下算法变体在相同数据集上的性能：Base：基础模型，使用固定粒度 (取多粒度集合的中位数) 且不进行交互感知分析，仅基于特征独立依赖度进行选择；Base+MG：在基础模型上引入多粒度邻域系统；Base+IA：在基础模型上引入交互感知机制；MGIA-OSFS (Ours)：完整模型，同时包含多粒度邻域系统与交互感知机制。

实验在 Forest Covertyp 和 Electricity 两个具有代表性的数据集上进行，结果如表 2 所示 (评价指标为平均分类准确率 %)。

表 2 消融实验中两个典型数据集的评估指标

模型变体	Forest Covertyp	Electricity
Base	79.5	82.1
Base+MG	81.8	84.3
Base+IA	82.6	85.0
MGIA-OSFS	84.2	86.5

结果分析:

(1) 多粒度邻域系统 (MG) 的有效性: 与 Base 模型相比, 引入多粒度邻域系统的 Base+MG 模型在两个数据集上的准确率分别提升了 2.3% 和 2.2%。这表明, 通过在不同尺度下评估特征, 模型能够更全面地捕捉数据的分布特性, 从而做出更鲁棒的特征选择决策, 有效克服了单一粒度模型的局限性。

(2) 交互感知机制 (IA) 的有效性: Base+IA 模型相比 Base 模型, 准确率分别提升了 3.1% 和 2.9%, 提升幅度更为显著。这证明了在在线流特征选择中, 考虑特征间的协同效应至关重要。该机制能够识别并保留那些单独与标签相关性不强, 但与已选特征组合后能显著提升判别力的特征, 避免了有用特征的遗漏。

(3) 组件的协同效应: 完整的 MGIA-OSFS 模型取得了最佳性能, 其效果并非 Base+MG 与 Base+IA 效果的简单叠加。在 Forest Covertype 数据集上, 完整模型比 Base+MG 提升了 2.4%, 比 Base+IA 提升了 1.6%。这证明了多粒度系统为交互感知提供了更丰富、更稳健的评估上下文, 而交互感知则充分利用了多粒度系统所提供的信息, 两者协同工作, 共同促成了最终性能的显著提升。

消融实验结论明确: 本文提出的多粒度邻域系统与交互感知机制均为有效的创新点, 且二者之间存在积极的协同作用, 共同构成了 MGIA-OSFS 算法卓越性能的核心。

4 结语

本文针对在线流特征选择中存在的单一粒度适应性差和特征交互挖掘不足问题, 提出了一种基于多粒度邻域系统与交互感知的在线流特征选择方法 (MGIA-OSFS)。通过构建多粒度邻域系统, 算法能够在不同尺度下全面感知数据分布特性, 增强对数据流动态变化的适应能力; 通过设计交互感知机制, 能够精准识别特征间的协同效应, 避免有用特征的遗漏。两个创新组件相互配合, 共同提升了特征选择的判别力与鲁棒性。在多个真实数据集上的实验结果表明, MGIA-OSFS 在分类准确率和特征子集紧凑性方面均显著优于现有主流方法。消融研究进一步证实了多粒度系统和交互感知机制各自的有效性及其协同作用。未来工作将从三个方向展开: 一是研究自适应粒度选择策略, 实现粒度参数的动态优化; 二是将算法与概念漂移检测机制结合, 增强在非平稳数据流中的鲁棒性; 三是探索算法在多标记学习等复杂场景下的扩展应用, 进一步提升其适用价值。

参考文献:

- [1] ZHOU P, HU X G, LI P P, et al. Online streaming feature selection using adapted neighborhood rough set[J]. Information sciences, 2019, 481: 258-279.
- [2] HU X G, ZHOU P, LI P P. A survey on online feature selection

with streaming features[J]. Frontiers of computer science: selected publications from Chinese universities, 2018, 12(3): 479-493.

- [3] WU X D, 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.
- [4] WANG J, WANG M, LI P P, et al. Online feature selection with group structure analysis[J]. IEEE transactions on knowledge and data engineering, 2015, 27(11): 3029-3041.
- [5] HU Q H, YU D R, LIU J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information sciences, 2008, 178(18): 3577-3594.
- [6] BLACKARD J A, DEAN D J. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables[J]. Computers and electronics in agriculture, 1999, 24(3): 131-151.
- [7] HARRIES M. SPLICE-2 comparative evaluation: electricity pricing[EB/OL]. 1999[2025-12-22]. <https://www.semanticscholar.org/paper/SPLICE-2-Comparative-Evaluation%3A-Electricity-Harries/31c493f93edf4a2440add1db8797cbd1143612b8>.
- [8] KAYACIK G, ZINCIR-HEYWOOD A N, HEYWOOD M I. Selecting features for intrusion detection: a feature relevance analysis on kdd 99 intrusion[C]//Proceedings of the Third Annual Conference on Privacy, Security and Trust. Piscataway: IEEE, 2005: 1722-1723.
- [9] 陈诗佳. 面向垃圾邮件检测的参数高效混合专家模型的研究[D]. 合肥: 中国科学技术大学, 2025.
- [10] 王伟申. 结合动态特征的 Python 软件缺陷预测研究[D]. 镇江: 江苏科技大学, 2025.
- [11] 王健. 高效臭氧溶气氧化技术研究[D]. 扬州: 扬州大学, 2023.
- [12] YOU D L, WANG Y, XIAO J W, et al. Online multi-label streaming feature selection with label correlation[J]. IEEE transactions on knowledge and data engineering, 2021, 35(3): 2901-2915.
- [13] ZHOU J, FOSTER D, STINE R, et al. Streaming feature selection using alpha-investing[C]//Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. New York: ACM, 2005: 384-393.

【作者简介】

杜晨霞 (1998—), 女, 山西吕梁人, 硕士研究生, 研究方向: 机器学习、粒计算、分层分类等。

(收稿日期: 2025-11-24 修回日期: 2026-01-06)