

UM FRAMEWORK DE CÓDIGO ABERTO PARA MINERAÇÃO DE DADOS MULTIOBJETIVO

Rodrigo Lúcio dos Santos Silva – rodrigo.santos@unifei.edu.br
João Luiz Junho Pereira – joaoluizjp@unifei.edu.br
Alexandre Ferreira de Pinho – pinho@unifei.edu.br

Palavras-chave: Otimização Multiobjetivo, Algoritmo de Lichtenberg, Feature Selection, Prototype Selection

1. INTRODUÇÃO

Na mineração de dados, a redução de dados é fundamental para tornar o processo de descoberta de conhecimento mais eficiente e interpretável. A *feature selection* (FS) busca identificar e manter apenas as variáveis mais relevantes, eliminando atributos redundantes ou irrelevantes, o que reduz a dimensionalidade e melhora o desempenho dos algoritmos. Já a *prototype selection* atua sobre as instâncias, escolhendo subconjuntos representativos que preservam as principais características e distribuições dos dados. Juntas, essas técnicas simplificam os dados — a FS reduzindo atributos e a *prototype selection* reduzindo instâncias —, permitindo que os algoritmos de mineração operem de forma mais rápida e eficaz, sem perda significativa de informação (HAN, JIAWEI; KAMBER; PEI, 2011).

Este trabalho adota a abordagem *wrapper* para FS, a qual utiliza um algoritmo de aprendizado de máquina — como o *K-nearest neighbor* (KNN) — para avaliar a qualidade dos subconjuntos de atributos. Embora mais exigente computacionalmente, essa abordagem é amplamente preferida por sua alta acurácia (PEREIRA *et al.*, 2023). Considerando que o problema de FS é de natureza NP-hard, o uso de meta-heurísticas se torna uma escolha natural e eficaz para sua resolução (XIE; SAGE; ZHAO, 2023). Neste contexto, este estudo propõe uma abordagem multiobjetivo baseada no Algoritmo *Lichtenberg* Multiobjetivo (MOLA), uma meta-heurística híbrida inspirada na propagação de raios (PEREIRA *et al.*, 2022). O MOLA é implementado pela primeira vez em *Python* para o problema de seleção de

características em abordagem *wrapper*, buscando equilibrar a acurácia da classificação e a redução de dimensionalidade, e fornecendo um conjunto de soluções *Pareto*-ótimo em vez de uma única solução ideal.

2. DESENVOLVIMENTO

2.1 Referencial teórico

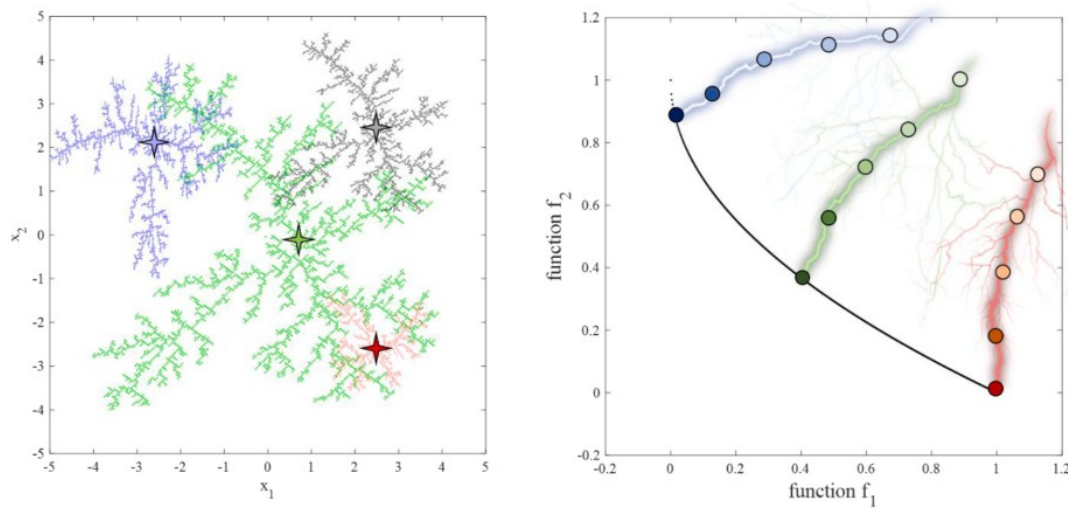
No domínio da seleção de características multiobjetivo (MOFS), grande parte das pesquisas concentra-se na aplicação de meta-heurísticas multiobjetivo (MOMH) para lidar simultaneamente com dois objetivos principais e frequentemente conflitantes: minimizar o tamanho do subconjunto de características selecionadas e reduzir a taxa de erro de classificação. Entre as abordagens mais utilizadas, destacam-se os métodos *wrapper*, que empregam classificadores de aprendizado de máquina — como o KNN — para avaliar a qualidade dos subconjuntos de características ao longo do processo de otimização. A qualidade da frente de *Pareto* obtida por esses métodos é normalmente avaliada por meio de métricas amplamente reconhecidas, como o *Hypervolume* (HV) e o *Inverted Generational Distance* (IGD) (HAN, FEI *et al.*, 2021).

As MOMH são originalmente voltadas a problemas de otimização contínua, enquanto a FS é um problema discreto, em que cada variável indica a seleção (1) ou exclusão (0) de uma característica. Para aplicar algoritmos multiobjetivo nesse contexto binário, é necessário convertê-los por meio de funções de transferência, que transformam as saídas contínuas em valores binários sem alterar a estrutura do algoritmo (ZHANG *et al.*, 2023).

O MOLA é uma meta-heurística híbrida de otimização que utiliza o conceito de figuras de Lichtenberg, padrões de ramificação inspirados em descargas elétricas. O MOLA combina uma busca populacional para explorar amplamente o espaço de soluções com uma busca baseada em trajetória para exploração focada. O algoritmo gera numericamente uma figura de Lichtenberg (LF) que serve como um guia para a exploração do espaço de busca. A busca é balanceada entre a exploração global, com a figura principal, e a exploração local, com uma figura de refinamento menor. A cada iteração, o MOLA avalia um subconjunto de pontos da figura, retém as soluções

não dominadas (a frente de *Pareto*) e utiliza uma dessas soluções superiores como ponto de partida para a próxima busca (Figura 1). Esse mecanismo de auto-orientação permite que o MOLA se concentre progressivamente em regiões que geram soluções de alta qualidade (PEREIRA *et al.*, 2022).

Figura 1 - Estratégia de busca básica de MOLA.



Fonte: (PEREIRA *et al.*, 2022)

2.2 Metodologia

Este trabalho apresenta o MOLA-PY, uma nova implementação em Python do algoritmo MOLA para a seleção de características em aprendizado de máquina. A escolha do Python é justificada por sua natureza de código aberto e pela vasta disponibilidade de bibliotecas científicas, como *NumPy* e *scikit-learn*, que facilitam a integração com tarefas de otimização e treinamento de modelos. Para avaliar seu desempenho, o MOLA-PY foi comparado a uma coleção de outros algoritmos multi-objetivo (MOMHs) já implementados na biblioteca jMetalPy (BENITEZ-HIDALGO *et al.*, 2019). Essa biblioteca foi selecionada por seu design robusto e por sua coleção de algoritmos confiáveis, adequados para pesquisa científica. A seleção de atributos foi modelada como um problema de otimização multi-objetivo, onde os objetivos são minimizar o erro de classificação e o número de atributos selecionados. Para a avaliação de cada solução, foi utilizado um classificador k-NN com validação cruzada.

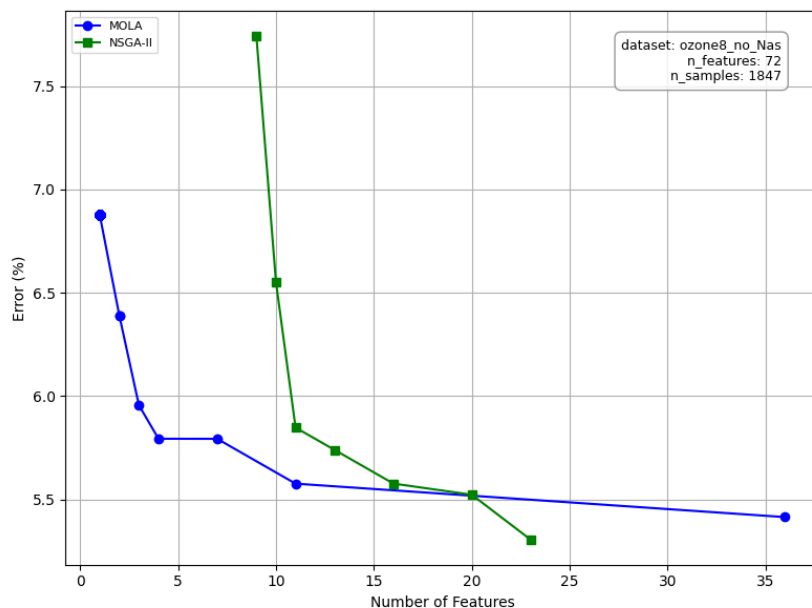
Para garantir uma comparação rigorosa, o estudo utilizou um benchmark de 30 *datasets* com diversas características. Essa seleção foi feita para cobrir uma ampla variedade de cenários, desde dados fáceis até os mais complexos de classificar, permitindo testar a robustez dos algoritmos (PEREIRA *et al.*, 2024). O desempenho foi avaliado usando as métricas HV e IGD. Como a fronteira de *Pareto* ideal é desconhecida, uma fronteira de referência foi construída a partir da união das soluções não dominadas encontradas por todos os algoritmos. As configurações dos parâmetros dos algoritmos foram baseadas em estudos anteriores, o que assegurou a validade das comparações.

3. RESULTADOS E DISCUSSÃO

Os resultados experimentais mostraram que o algoritmo MOLA apresentou desempenho superior em relação a outros métodos de otimização multiobjetivo, especialmente em métricas que avaliam a qualidade e a diversidade das soluções. Ele se destacou em bases de dados com alta dimensionalidade, demonstrando boa capacidade de equilibrar a redução de características com a manutenção da precisão de classificação. Testes estatísticos (*Friedman e Nemenyi*) confirmaram diferenças significativas entre os algoritmos analisados, e o MOLA obteve as melhores posições médias em comparação com seus concorrentes, mantendo desempenho consistente nas métricas avaliadas e exibindo frentes de *Pareto* bem distribuídas e dominantes. A Figura 2 mostra a frente de *Pareto* comparando o MOLA com o NSGA-II.

Além disso, a análise de tempo de processamento mostrou que o MOLA foi o algoritmo mais eficiente e equilibrado, unindo rapidez, estabilidade e capacidade de resolver todos os problemas testados. Embora outros métodos, como MOEA/D e NSGA-II, também tenham apresentado bom desempenho, o MOLA se destacou por sua robustez e eficiência, oferecendo um excelente compromisso entre velocidade, qualidade das soluções e interpretabilidade. Esses resultados indicam que o MOLA é uma opção altamente promissora para seleção de características em cenários complexos e de alta dimensionalidade.

Figura 2 – Frente de *Pareto* MOLA vs NSGA-II.



Fonte: Elaboração própria.

4. CONSIDERAÇÕES FINAIS

O estudo apresentou e avaliou o algoritmo MOLA para o problema de FS multiobjetivo, demonstrando sua eficácia em equilibrar a minimização do erro de classificação e a redução do número de atributos. Testado em 30 *datasets* e comparado com nove algoritmos, o MOLA obteve desempenho superior segundo os indicadores IGD e HV, além de se destacar estatisticamente nos testes de *Friedman* e *Nemenyi*. Também se mostrou o mais eficiente em termos computacionais, sendo o mais rápido na maioria dos casos. Assim, o MOLA se consolida como uma solução de última geração, combinando precisão, eficiência e interpretabilidade para aplicações reais em aprendizado de máquina.

O próximo passo da pesquisa será avaliar o MOLA para o problema de seleção de protótipos visando reduzir o número de exemplos de treinamento, removendo instâncias redundantes, ruidosas ou pouco informativas, ou escolhendo exemplos representativos do conjunto original. Enquanto a seleção de características atua no eixo das colunas (variáveis), a seleção de instâncias atua no eixo das linhas (amostras).

AGRADECIMENTOS

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo contínuo apoio à pesquisa e à divulgação científica.

REFERÊNCIAS

BENITEZ-HIDALGO, Antonio *et al.* jMetalPy: a Python Framework for Multi-Objective Optimization with Metaheuristics. 17 abr. 2019. Disponível em: <<http://arxiv.org/abs/1903.02915>>.

HAN, Fei *et al.* Multi-objective particle swarm optimization with adaptive strategies for feature selection. *Swarm and Evolutionary Computation*, v. 62, 1 abr. 2021.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*. . [S.l: s.n.], 2011.

PEREIRA, João Luiz Junho *et al.* A comparison between chaos theory and Lévy flights in sunflower optimization for feature selection. *Expert Systems*, v. 40, n. 8, 1 set. 2023.

PEREIRA, João Luiz Junho *et al.* Multi-objective lichtenberg algorithm: A hybrid physics-based meta-heuristic for solving engineering problems. *Expert Systems with Applications*, v. 187, 1 jan. 2022.

PEREIRA, João Luiz Junho *et al.* Optimal selection of benchmarking datasets for unbiased machine learning algorithm evaluation. *Data Mining and Knowledge Discovery*, v. 38, n. 2, p. 461–500, 1 mar. 2024.

XIE, Jiarui; SAGE, Manuel; ZHAO, Yaoyao Fiona. *Feature selection and feature learning in machine learning applications for gas turbines: A review. Engineering Applications of Artificial Intelligence*. [S.l.]: Elsevier Ltd. , 1 jan. 2023

ZHANG, Min *et al.* Multi-objective optimization algorithm based on clustering guided binary equilibrium optimizer and NSGA-III to solve high-dimensional feature selection problem. *Information Sciences*, v. 648, 1 nov. 2023.