Simpósio de Integração Acadêmica
"Das Montanhas de Minas ao Oceano: Os Caminhos da Ciência para um Futuro Sustentável"
SIA UFV 2025

UFV Universidade Federal de Viçosa

Análise de Desempenho de Algoritmos de Aprendizado de Máquina com GPU e Estratégias de Seleção de Atributos

Isabela de Castro Freitas, Ricardo dos Santos Ferreira, Olavo Alves Barros Silva, Pedro Henrique Coura Pereira

ODS9: Construir infraestruturas resilientes, promover a industrialização inclusiva e sustentável e fomentar a inovação

Categoria: Pesquisa

Introdução

Este trabalho avalia o desempenho do algoritmo K-means em duas abordagens: execução sequencial em CPU (scikit-learn) e paralela em GPU (CUDA C++ com T4). Além da análise de tempo de execução, investigou-se a redução de dimensionalidade por meio da seleção de subconjuntos de atributos, utilizando o índice de Gini como métrica de pureza e validando os melhores casos com o classificador XGBoost. Essa combinação busca equilibrar eficiência computacional e desempenho preditivo.

Objetivos

O trabalho teve como objetivo investigar se é possível reduzir o número de atributos em conjuntos de dados de alta dimensionalidade sem comprometer significativamente a acurácia. Para isso, foram avaliadas combinações de três atributos usando K-means e índice de Gini, com validação via XGBoost. Além disso, buscou-se otimizar o processo por meio de uma implementação paralela em CUDA C++, tornando viável a análise em casos de maior volume de dados.

Material e Métodos ou Metodologia

Foram testadas todas as combinações de três atributos em cada dataset, aplicando K-means, com 8 clusters, e avaliando a pureza com índice de Gini. As representações reduzidas foram classificadas com XGBoost e comparadas com classificações que utilizaram todas as features.

Para acelerar o processo, o K-means foi implementado em CUDA C++, permitindo execução paralela de múltiplas combinações e maior eficiência na busca por subconjuntos representativos.

Apoio Financeiro

CAPES FAPEMIG CNPq

Resultados e/ou Ações Desenvolvidas

A redução de dimensionalidade apresentou resultados melhores para datasets binários, nos quais mostrou que os clusters mais puros (0,5%, pontos vermelhos na Figura 1) geram maiores acurácias no XGBoost, chegando a a 85%, em alguns casos ultrapassando os resultados do dataset original.

A paralelização do K-means em GPU T4 manteve tempos estáveis mesmo em datasets maiores, sendo muito mais rápida que a versão sequencial.

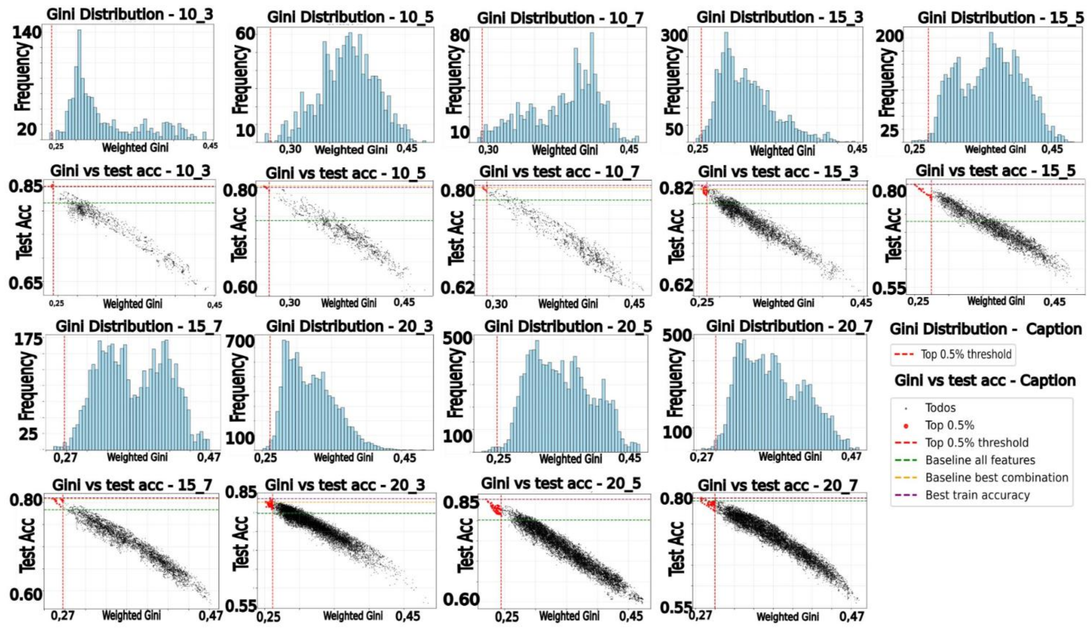

Figura 1: Resultados dos experimentos de redução de dimensionalidade para algoritmos binários

Conclusões

A redução de dimensionalidade, transformando clusters em novos atributos, destacou características relevantes e aumentou a acurácia em datasets binários, mas não obteve bom desempenho para datasets com mais classes.

A paralelização do K-means em GPU apresentou tempos estáveis e escaláveis para casos com mais combinações. Assim, essas abordagens podem melhorar a representatividade dos dados e garantir processamento rápido, equilibrando desempenho e eficiência em aprendizado de máquina em algumas aplicações.

Bibliografia

ARORA, P.; VARSHNEY, S. et al. Analysis of k-means and k-medoids algorithm for big data.
Procedia Computer Science, Elsevier, v. 78, p. 507–512, 2016.
LABER, E.; MURTINHO, L. Minimization of gini impurity: Np-completeness and
approximation algorithm via connections with the k-means problem. Electronic Notes in
Theoretical Computer Science, Elsevier, v. 346, p. 567–576, 2019.
PENHA, J. C. et al. A gpu/fpga-based k-means clustering using a parameterized code generator.
In: IEEE. High Performance Computing Systems (WSCAD). [S.l.], 2018.