Acelerando o cálculo do índice Dunn de validação de agrupamento

Eduardo S. Grün[1], Wellington S. Martins[1], Ricardo Franco[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Caixa Postal 74.690-900 – Goiânia – GO – Brazil

eduardo.santos.grun@discente.ufg.br, wsmartins@ufg.br,
ricardofranco@ufg.br

Abstract.

This paper presents a parallel implementation of the Dunn index, utilizing GPUs to accelerate its computation. The Dunn index is a commonly used metric to evaluate the quality of clustering. By exploiting the parallelism of GPUs, we were able to significantly accelerate the calculation of this index, enabling the analysis of larger and more complex datasets. Comparing the parallel implementation with the sequential one, we observed substantial performance gains, demonstrating the effectiveness of the proposed approach.

Resumo.

Este trabalho apresenta uma implementação paralela do índice de Dunn utilizando GPUs para acelerar o cálculo. O índice de Dunn é uma métrica comum para avaliar a qualidade de agrupamentos. Ao explorar o paralelismo das GPUs, conseguimos acelerar significativamente o cálculo desse índice, permitindo a análise de conjuntos de dados maiores e mais complexos. Comparando a implementação paralela com a sequencial, observamos ganhos substanciais de desempenho, demonstrando a eficácia da abordagem proposta.

1. Introdução

Nos últimos anos, a quantidade de dados gerada tem crescido exponencialmente. Essa massificação de dados, conhecida como Big Data, exige novas ferramentas e técnicas para análise e processamento de dados. Neste contexto, o campo do meta-aprendizado, as meta-características são ferramentas poderosas para otimizar processos de aprendizado de máquina. Ao fornecer informações mais ricas sobre os dados, elas permitem tomar decisões mais assertivas sobre a escolha de algoritmos e a configuração de seus parâmetros, resultando em modelos mais precisos e eficientes [Brazdil et al. 2009].

Dunn é um índice de validação cujo objetivo é identificar clusters com alta distância inter-cluster e baixa distância intra-cluster [Silva L. 2023, Luna-Romera et al. 2016, Zerabi et al. 2020]. Neste trabalho foi implementado um cálculo aproximado do índice de Dunn, comparando sua versão sequencial e paralela(GPU).

2. Índice Dunn e trabalhos relacionados

O Índice de Dunn mede a qualidade de um agrupamento. Especificamente o índice de Dunn compara a distância entre os grupos (quanto maior, melhor) com o tamanho dos grupos (quanto menor, melhor). A alta coesão intragrupo e a grande distância intergrupo resultam em um valor elevado do índice, quanto maior o valor do índice, melhor é o

agrupamento. O cálculo destes índices é realizado depois do agrupamento, tendo como entrada o rótulo para cada ponto e os centros para cada um dos grupos encontrados: $C_1, C_2, ..., C_K$, onde K é o número de grupos. [Brazdil et al. 2009].

Estudos anteriores como [Luna-Romera et al., 2016, Zerabi et al., 2020] já abordaram o cálculo do índice de Dunn, propondo aproximações baseadas em centróides e utilizando paralelismo para melhorar o desempenho. Essas pesquisas, no entanto, concentraram-se em ambientes de alta performance, explorando a escalabilidade horizontal por meio da utilização de múltiplas máquinas interconectadas. Em contraste, este trabalho investiga a possibilidade de acelerar o cálculo do índice de Dunn fazendo uso da exploração da escalabilidade vertical, ou seja, utilizando unidades de processamento gráfico (GPUs) em uma única máquina para realizar os cálculos de forma paralela.

3. Ambiente de Testes e Datasets Utilizados

Os experimentos foram realizados em um servidor da UFG, o ambiente de execução utilizado possui 2 CPUs Intel Xeon E5-2620, 16GB de RAM, e uma GPUs RTX 3060 12GB (com 3584 núcleos - 28 SMs c/ 128 núcleos). Para realizar uma análise abrangente, foram utilizados um total de oito conjuntos de dados, incluindo seis conjuntos sintéticos e dois conjuntos reais, destes, 4 são demonstrados posteriormente. Os conjuntos de dados sintéticos, gerados fazendo o uso de um software [Silva L., 2023; Luna-Romera et al., 2016; Zerabi et al., 2020], foram customizados para atender às necessidades específicas deste estudo.

4. Trabalho realizado

Para o cálculo do índice de Dunn, duas implementações foram desenvolvidas: uma sequencial em C e outra paralela em CUDA. Na versão paralela, os dados são transferidos para a GPU, onde o cálculo das somas parciais das coordenadas dos pontos de cada cluster é distribuído entre múltiplas threads de um kernel CUDA. A memória compartilhada dos blocos é utilizada para acelerar o acesso aos dados e uma redução paralela é empregada para combinar os resultados parciais. O número ideal de threads por bloco, definido em 128, foi determinado experimentalmente. A CPU finaliza o cálculo, combinando as somas parciais e dividindo pelo número de pontos. A combinação da CPU e GPU proporcionou um ganho significativo de desempenho, conforme demonstrado nos resultados.

O algoritmo explora o paralelismo distribuindo o cálculo dos centróides entre múltiplas threads, cada uma responsável por um subconjunto de dados. A redução paralela é empregada para combinar os resultados parciais de cada thread, otimizando o processo através do uso de memória compartilhada e minimizando a latência de acesso à memória.

5. Resultados

A fim de obter resultados confiáveis, adotamos uma metodologia de medição do tempo de execução baseada na média de dez repetições de cada experimento. A cronometragem teve início no momento exato da primeira chamada da função que

calcula o índice de Dunn. Realizamos as comparações entre o algoritmo sequencial em sua versão original e paralelo de Dunn utilizando sua versão aproximada.

Na Tabela 1, observamos os resultados obtidos entre as comparações do algoritmo sequencial e paralelo do índice de Dunn. Na terceira coluna temos os tempos em segundos das execuções das duas versões do algoritmo. A quarta coluna contém o SpeedUp entre a versão sequencial e a paralela.

Tabela 1. Comparação no cálculo de Dunn: sequencial vs paralelo

Dataset	Algoritmo	Tempo(segundos)	SpeedUp
luna_k5_f20_500000	Sequencial	387	86,59x
	Paralelo	4,1	
luna_k5_f20_2500000	Sequencial	1423	87,3x
	Paralelo	16,3	
luna_k9_f20_5000000	Sequencial	2678	50,4x
	Paralelo	54,9	
luna_k9_f20_11000000	Sequencial	6592	19,67x
	Paralelo	335,1	

6. Conclusões

Com base nos resultados apresentados, foi feito um bom uso do paralelismo para o problema do índice de Dunn. É evidente que o algoritmo paralelo é mais rápido que o sequencial, obtendo speedups de até 86.6x. Como trabalho futuro, almejamos testar em outros datasets reais, realizar uma comparação mais abrangente com outros trabalhos e implementar outros índices de validação, como o silhueta, além de outras formas de paralelização como por exemplo OpenMP.

Referências

Luna-Romera, J. M., Martínez Ballesteros, M., Garcıa-Gutierrez, J., and Riquelme, J. (2016). An approach to silhouette and dunn clustering indices applied to big data in spark. Advances in Artificial Intelligence. CAEPIA 2016. Lecture Notes in Computer Science(), vol 9868. Springer, Cham.

Brazdil, P., Giraud-Carrier, C., Soares, C., and Vilalta, R. (2009). Metalearning: Applications to data mining. Springer Publishing Company.

PUMA-VILLANUEVA, W. J.; VON ZUBEN, F. J. Índices de validação de agrupamentos.

Dunn, J.: Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4(1), 95–104 (1974)

Rivolli, A., Garcia, L. P., Soares, C., Vanschoren, J., and de Carvalho, A. C. (2018). Characterizing classification datasets: a study of meta-features for meta-learning. arXiv preprint arXiv:1808.10406.