ADAPTAÇÃO DE UM ALGORITMO BASEADO EM GRADIENT TREE BOOSTING PARA CLASSIFICAÇÃO MULTICLASSE DE FLUXOS DE DADOS UTILIZANDO AUTOAPRENDIZADO[1]

Yuji Yamada Correa[2], Fabiano Baldo[3]

1 Vinculado ao projeto "StreamMining - Novas abordagens para Algoritmos de Aprendizadem em Fluxos de Dados Não Estacionários"
[2]Acadêmico (a) do Curso de Ciência da Computação – CCT – Bolsista PROBIC
3 Orientador, Departamento de Ciência da Computação – CCT – fabiano.baldo@udesc.br

A popularização das tecnologias de sensoriamento e conectividade, como 5G e IoT, está impulsionando a geração de fluxos de dados. Esses tipos de dados são sequências sem previsões de término de objetos de dados não estacionários que são continuamente gerados em altas taxas. O conceito não estacionário, também conhecido como mudança de conceito, significa que o comportamento do fluxo pode mudar de vez em quando, estando em constante evolução. A literatura apresenta alguns algoritmos para classificação de fluxo de dados que lidam com mudança de conceito. No entanto, apesar de apresentarem boa acurácia, eles consomem um tempo considerável de processamento, uma vez que utilizam técnicas de ensemble. O XGBoost é um algoritmo de boosting que cria árvores de decisão em paralelo, utilizando vários processadores a fim de minimizar o tempo de treinamento e melhorar o desempenho do processo.

Neste trabalho propomos uma adaptação do algoritmo de Baldo et al. (2022), para classificar fluxos de dados não estacionários com múltiplas classes, chamado AFXGB-MC. O AFXGB-MC começa treinando um modelo que será utilizado para classificar os dados do fluxo. Esse modelo é treinado e usado para classificar dados enquanto não atingir um limite máximo definido por parâmetro. Em paralelo, um novo classificador começa a ser treinado assim que o classificador mais antigo atinge o limite de dados treinados. O classificador principal é substituído quando o novo modelo atinge um nível de treinamento adequado. Além disso, o AFXGB-MC inclui um mecanismo que atualiza todas as árvores de decisão já treinadas e pertencentes ao conjunto XGBoost. Isso garante que não apenas a última árvore treinada esteja atualizada com as características do fluxo, mas também as árvores mais antigas incorporem parcialmente as características dos dados mais recentes. Além disso, ele aplica um algoritmo de detecção de mudança de conceito que utiliza um mecanismo de *reset* da janela de dados para intensificar a atualização do classificador. O algoritmo proposto foi comparado com outros da literatura para avaliar sua acurácia e desempenho em tempo de processamento.

Para avaliação do AFXGB-MC, ele foi comparado com outros seis modelos propostos na literatura, que são: ARF, HAT, LbHT, ObHT, SAM-kNN e IOE. As métricas avaliadas foram a acurácia, o tempo de treinamento e de teste, e a capacidade de adaptação à mudança de conceito. Foram utilizadas onze *datasets* para os testes, sendo seis sintéticas com mudanças de conceito gradual ou abrupta. Para cada *dataset*, os modelos foram ranqueados por acurácia. A Tabela 1 mostra os tempos de execução média dos algoritmos.

Para comparar os modelos e determinar se havia diferenças significativas entre eles, foram realizados os testes estatísticos de Friedman e Nemenyi. Os testes indicaram que os classificadores LbHT, ARF, AFXGB-MC e SAMkNN obtiveram os melhores resultados em termos de acurácia,

sem diferença estatística significativa entre eles. Além disso o AFXGB-MC e o SAMkNN obteram melhor acurácia em *datasets* reais. Isso significa que são mais adequados para processar fluxos de dados reais, o que é uma diferença competitiva, pois o processamento de dados reais é mais útil na prática.

A Figura 1 apresenta o tempo médio dos 7 modelos em um gráfico de barras cumulativo. O AFXGB-MC teve o menor tempo total médio de todos os algoritmos. Quando comparado ao SAMkNN, que teve o segundo menor tempo, o AFXGB-MC é 2.3 vezes mais rápido. Em comparação ao LbHT, o algoritmo com o melhor ranking de acurácia, o AFXGB-MC é 51.5 vezes mais rápido. Quando comparado ao ARF, o segundo melhor algoritmo de acurácia, o AFXGB-MC é 22 vezes mais rápido.

Tabela 1. *Acurácia média dos algoritmos avaliados no estudo.*

Dataset	LBHT	ARF	AFXGBMC	OBHT	HAT	IOE	SAMKNN
RBF_A	0.8957	0.8556	0.8883	0.8475	0.7675	0.8812	0.9289
RBF_G	0.8934	0.8521	0.8797	0.8436	0.7809	0.8790	0.9266
SEA_A	0.9951	0.9941	0.9904	0.9545	0.9867	0.9657	0.9819
SEA_G	0.9903	0.9889	0.9859	0.9519	0.9832	0.9609	0.9771
LED_A	0.3364	0.3480	0.3592	0.3410	0.3638	0.2973	0.1964
LED_G	0.3318	0.3443	0.3581	0.3402	0.3581	0.2975	0.1939
GAS	0.8793	0.9697	0.7260	0.5580	0.5555	0.7785	0.9710
POKER	0.5854	0.5373	0.6142	0.5308	0.5195	0.3474	0.4928
KDDCUP	0.9989	0.9992	0.9872	0.9980	0.9964	0.9939	0.9990
COVTYPE	0.9290	0.9406	0.8246	0.8740	0.8276	0.9027	0.9513
WINNIPEG	0.9813	0.9813	0.9867	0.9611	0.9572	0.9765	0.5991
Acc. média	**0.8015**	**0.8010**	**0.7818**	**0.7455**	**0.7360**	**0.7528**	**0.7471**

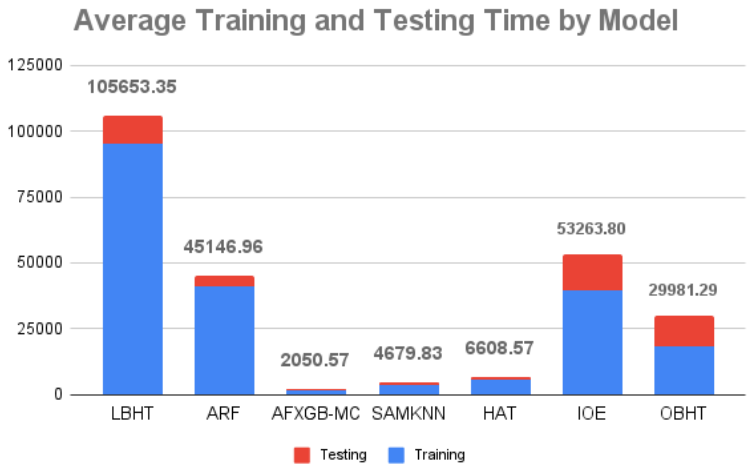

Figura 1. *Tempo médio de treinamento e teste de todos os algoritmos (em segundos).*

Palavras-chave: Aprendizado de Máquina. Fluxos de dados. Mudança de conceito. Classificação de dados.

BALDO, Fabiano et al. Adaptive Fast XGBoost for Binary Classification. 2022.