ADAPTAÇÃO DO XGBOOST PARA CLASSIFICAÇÃO BINÁRIA DE FLUXOS DE DADOS COM MUDANÇA DE CONCEITO[1]

Kawan Moreira Weege[2], Fabiano Baldo[3].

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

Com a geração cada vez mais abundante e rápida de dados na forma de fluxos, alavancada pelas redes de alta velocidade, a aplicação de métodos tradicionais de Aprendizado de Máquina para sua análise se torna inviável. Isso se deve ao fato, principalmente, dos fluxos da dados serem infinitos e apresentarem mudança de conceito. Essas características fazem com que os algoritmos de Aprendizado de Máquina precisem ser capazes de se atualizarem ao longo do tempo, de forma rápida para que não haja perda de informação, e precisem detectar adequadamente as mudanças de conceito, de tal forma que o modelo não perca sua precisão ao longo do processamento dos dados.

Dentre os algoritmos para classificação de dados o XGBoost (eXtreme Gradient Boosting) é um dos mais recentes e que apresenta os melhores resultados, tanto de precisão quanto de tempo de processamento. Ele implementa uma estratégia de conjunto de classificadores na forma de *bagging*. Originalmente, o XGBoost não foi projetado para analisar fluxos de dados e, consequentemente, para lidar com a mudança de conceito. Montiel et al. (2020) propôs um algoritmo chamado AXGB, baseado no XGBoost, para análise de fluxos de dados. Já Bonassa (2021) desenvolveu um algoritmo baseado no AXGB para classificação rápida denominado AFXGB (Adaptive Fast XGBoost). Apesar do novo algoritmo trazer melhorias na diminuição do tempo de execução, os testes realizados estavam incompletos, e esse novo modelo trazia alguns erros de implementação.

O objetivo desse trabalho foi realizar as correções do algoritmo AFXGB, executar testes com outras bases de dados e comparar os resultados com outros classificadores propostos na literatura. Além disso, foi implementado um mecanismo de detecção de mudança de conceito ativo, e outro que realiza o reset da janela deslizante de instâncias, ambos podendo ser ativados ou desativados. Feitas as alterações, foram realizados experimentos comparando o AFXGB com os seguintes classificadores: AXGB, AFXGB, ARF e HAT. Também foram comparadas as seguintes combinações de configurações do AFXGB: com detecção ativa de mudança de conceito desativado e reset da janela desativado, com o primeiro ativado e o segundo desativado, com o primeiro desativado e o segundo ativado, e ambos ativados. Foram feitas 5 execuções de cada algoritmo para cada um dos 7 datasets escolhidos, 5 destas sendo sintéticas e 2 reais, cada um tendo em média 500.000 instâncias. As métricas analisadas para determinar o desempenho de cada modelo são acurácia, estatística KAPPA, tempo de treinamento, tempo de previsão, e tempo total (soma entre tempo de treinamento e previsão). Para cada dataset foi feito o ranqueamento dos modelos de acordo com a suas acurácias. Foram feitos também os testes estatísticos de Friedman e o Nemenyie, que compararam se havia diferença estatística entre os modelos. A Figura 1 mostra a recuperação mais rápida da mudança de conceito realizada pelo algoritmo AFXGB, na figura denominado como incremental_r.

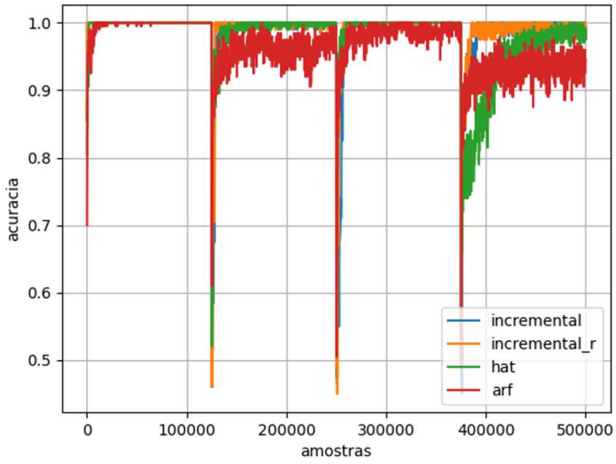

Figura 1. *Resultados das acurácias dos algoritmos citados no dataset sintético AGRAWAL.*

Os testes de Friedman e Nemenyie mostraram que não houve uma diferença estatisticamente significativa em relação às acurácias entre os quatro classificadores. A Tabela 1 mostra os tempos de execução média dos algoritmos.

Tabela 1. *Tempos de execução média (em minutos) dos algoritmos AFXGB, AXGB, HAT e ARF.*

Modelo	Média Treinamento	Média Teste	Média Total
ARF	71.05	6.55	77.60
HAT	2.90	0.47	3.37
AXGB	0.21	378.54	378.76
AFXGB	0.31	16.26	16.57

O AFXGB teve tempos menores em relação ao AXGB. O ARF e o HAT tiveram os maiores tempo de treinamento e os menores tempo de teste. Além disso, o HAT teve o menor tempo total em comparação aos outros classificadores testados. No entanto, foi possível perceber que o AFXGB tem um tempo total de processamento muito menor que o AXGB, conforme esperado, pois o algoritmo AXGB utilizam um conjunto de XGBoost em sua arquitetura, enquanto o AFXGB tem apenas dois XGBoost em sua arquitetura, um principal e outro auxiliar, que substitui o principal quando o primeiro está se tornando superespecializado.

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

MONTIEL, Jacob et al. Adaptive xgboost for evolving data streams. 2020.
BONASSA, Gustavo Miquelluzzi. Adaptação De Classificador Utilizando A Biblioteca XGBoost para Classificação Rápida De Fluxos De Dados parcialmente Classificados Com Mudança De Conceito. 2021.