# ADAPTAÇÃO DO XGBOOST PARA CLASSIFICAÇÃO SEMI-SUPERVISIONADA MULTICLASSE DE FLUXOS DE DADOS COM MUDANÇA DE CONCEITO [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 do Curso de Ciência da Computação – CCT – Bolsista PROBIC
[3] Orientador, Departamento de Ciência da Computação – CCT – fabiano.baldo@udesc.br

Com o surgimento de novas tecnologias como o 5G e o impulsionamento da IoT (Internet das Coisas), o volume de dados está crescendo cada vez mais. Essa geração constante de dados de forma contínua e sem uma previsão de término é conhecida como fluxo de dados. A extração de informações sobre esses dados é feita por meio do aprendizado de máquina. Entretanto, as técnicas tradicionais de classificação supervisionada não são adequadas em cenários reais de fluxos, pois comumente apresentam instâncias não rotuladas e com mudança de conceito. Para resolver esses problemas são utilizadas técnicas de aprendizado semi-supervisionado, que não necessitam que todas as instâncias estejam rotuladas. Isso se torna uma grande vantagem, uma vez que dados rotulados geralmente são difíceis de se obter. Entretanto, na literatura, técnicas de aprendizado semi-supervisionado são escassas e as soluções que existem não são rápidas o suficiente para tratar os fluxos de dados.

Para resolver esse problema são necessários algoritmos rápidos para construção de classificadores de forma incremental, e que suportem mudança de conceitos. Os algoritmos de *boosting* apresentam uma abordagem promissora para o campo de análise de dados. O Gradient Boosting é uma técnica de *boosting* geralmente constituída de árvores de decisão, e requer que cada árvore de decisão seja criada uma por vez, em que o modelo construído depende dos anteriores, o que pode torná-lo lento. Para resolver essa deficiência, foi proposto o XGBoost, algoritmo que tem como objetivo otimizar o tempo de treinamento do Gradient Boosting. Esse algoritmo utiliza a paralelização entre o aprendizado e o treinamento, tornando o processo de aprendizado mais rápido.

Neste trabalho, é proposta uma adaptação ao algoritmo AXGB, proposto por Bonassa et al. (2021), para suportar a classificação semi-supervisionada multiclasse de fluxos de dados com mudança de conceito, chamado SSAFXGB-MC. O trabalho de Bonassa et al. (2021) propor um algoritmo para classificação de fluxos binários usando como base o XGBoost. Para o aprendizado semi-supervisionado, foi utiliza uma estratégia de pseudo-rotulagem dos dados com o algoritmo 1NN. Também foi implementado um mecanismo de reset da janela deslizante sempre que o classificador temporário substitui o principal, para intensificar a atualização dele.

O algoritmo inicia com as instâncias do fluxo sendo adicionadas na janela até atingir sua capacidade máxima. Quando a janela está cheia, começa a classificação dos dados não rotulados por meio do algoritmo 1NN, utilizando no treinamento os dados rotulados da janela. Então, se a distância do vizinho mais próximo for menor que um parâmetro de confiança, a instância é adicionada ao *buffer* que servirá para treinar o classificador supervisionado. Na sequência, o algoritmo checa se existe um classificador principal criado, caso não exista, um classificador XGBoost é treinado e salvo para futuras atualizações. Caso já existir um classificador treinado, um novo modelo é adicionado a ele e é verificado se é necessário iniciar o treinamento do classificador

temporário. Em seguida, é verificado se já ocorreu uma mudança, se sim é feito o reset na janela. Por fim, a janela é atualizada e o contador de ciclos de treinamento é incrementado.

Para avaliação do algoritmo proposto, ele foi comparado com o algoritmo IOE desenvolvido por Parsa et al. (2021), e uma variação do SSAFXGB-MC utilizando o KNN para a pseudo-rotulagem dos dados. O IOE é um algoritmo para classificação semi-supervisionado multiclasse de fluxos de dados com mudança de conceito que utiliza o Hoeffding Tree como classificador base. As métricas avaliadas foram a acurácia, o tempo de treinamento e de teste, e a capacidade de adaptação à mudança de conceito. Foram utilizadas oito datasets para os testes.

A Tabela 1 mostra os resultados obtidos. Nela é possível observar que os algoritmos apresentam comportamentos bastante parecidos referente à acurácia. No entanto, para confirmar se há diferença estatística, foi aplicado o teste de Nemenyi para obter o ranking dos algoritmos, que demonstrou que o SSAFXGB-MC alcançou acurácia semelhante ao IOE. Em relação aos testes para comparação do tempo de execução, na Tabela 2 é mostrado o tempo médio de treinamento, teste e total de cada um dos algoritmos. Nota-se uma grande diferença nos tempos totais, onde o algoritmo proposto foi aproximadamente 15 vezes mais rápido que o IOE. Por fim, para comparar a capacidade de recuperação da mudança de conceito, foi utilizado o dataset SEA com cinco desvios abruptos. Nos testes foi possível observar que antes do terceiro desvio, no ponto 300k, todos os modelos estavam no nível mais alto de acurácia, mas quando acontece o desvio a acurácia dos modelos cai drasticamente. No entanto, enquanto o SSAFXGB-MC se recuperou totalmente após 3.2k instâncias, o IOE ainda tinha acurácia 17% menor a sua acurácia inicial. Dessa forma, é possível dizer que o algoritmo proposto tem uma recuperação mais rápida que o IOE.

Portanto, é possível concluir que o SSAFXGB-MC apresenta acurácia equivalente ao IOE, porém com tempo total de execução 15 vezes menor. Ainda, ele apresenta recuperação mais rápida após mudanças de conceito em relação ao IOE.

**Tabela 1.** *Média das acurácias e ranking para os algoritmos avaliados no estudo em 8 datasets.*

| Dataset | SSAFXGB-MC-KNN | SSAFXGB-MC-1NN | IOE |
|---|---|---|---|
| SEA_A | 0.9205 (3) | **0.9483** (1) | 0.9406 (2) |
| SEA_G | 0.9145 (3) | **0.9389** (1) | 0.9373 (2) |
| RBF_A | 0.6391 (3) | 0.7088 (2) | **0.8056** (1) |
| RBD_G | 0.3078 (1) | 0.3034 (2) | 0.2892 (3) |
| GAS | 0.1254 (3) | 0.5748 (2) | **0.7716** (1) |
| POKER | 0.2525 (3) | 0.6174 (2) | **0.9015** (1) |
| KDDCUP99 | 0.1282 (3) | 0.9439 (2) | **0.9736** (1) |
| COVERTYPE | 0.0348 (3) | 0.3276 (2) | 0.3285 (1) |
| **Avg. acc.** | 0.4153 | 0.6703 | 0.7434 |
| **Avg. rank** | 4.25 | 2.25 | 1.875 |

**Tabela 2.** *Tempos médios de execução dos algoritmos avaliados.*

| | Time (seconds) | | |
|---|---|---|---|
| **Model** | **Avg. training** | **Avg. testing** | **Avg. total** |
| SSAFXGB-MC-1NN | 7540.18 | 4176.16 | 11716.35 |
| SSAFXGB-MC-KNN | 7754.75 | 4173.25 | 11928.01 |
| IOE | 116276.95 | 58224.61 | 174501.57 |

**Palavras-chave:** Classificação rápida. Classificação multiclasse. Aprendizado semi-supervisionado. Fluxos de dados. XGBoost.

BONASSA, Gustavo Miquelluzzi. Adaptação De Classificador Utilizando A Biblioteca XGBoost para Classificação Rápida De Fluxos De Dados parcialmente Classificados Com Mudança De Conceito. 2021.