CONSTRUÇÃO DE UM WRAPPER PARA O ALGORITMO ADAPTIVE EXTREME GRADIENT BOOSTER SUPORTAR A REGRESSÃO SEMI-SUPERVISIONADA DE DADOS[1]

Carlos Eduardo Ritzmann[2], Fabiano Baldo[3]

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

O rápido desenvolvimento de tecnologias e o crescente número de sensores, junto aos avanços da Internet e a adoção de dispositivos inteligentes conectados, proporciona uma geração de dados cada vez maior. Entre as novas formas de geração de dados, uma que se destaca são os fluxos de dados, gerados de forma contínua e não estacionária, ou seja, as propriedades da variável alvo podem mudar com o tempo [2]. Outro detalhe comum em fluxos dados é o surgimento de dados não rotulados, devido ao alto custo de tempo e processamento do processo de rotulagem.

Ao trabalhar com fluxos de dados na área da aprendizagem de máquina é necessário processar uma grande quantidade de dados de forma rápida devido a sua alta frequência de geração. Para isso, o algoritmo utilizado como base neste trabalho será o Adaptive eXtreme Gradient Boosting (AXGB), uma adaptação do eXtreme Gradient Boosting (XGBoost) para trabalhar com fluxos de dados [1]. No entanto, o AXGB não suporta o aprendizado semi-supervisionado. Uma das maneiras de torná-lo apto ao tratamento de dados não rotulados é implementando um método *wrapper* [3]. Dessa forma, o modelo é treinado inicialmente apenas com os dados rotulados, em seguida os dados não rotulados passam pelo *wrapper*, que estima um rótulo para eles (pseudo-rotulagem). Então o modelo pode ser retreinado utilizando tanto dados rotulados quanto os dados pseudo-rotulados fornecidos pelo *wrapper*. A vantagem dessa técnica é que ela possibilita a entrada de dados não rotulados normalmente, sem alterar o funcionamento interno do algoritmo [3]. Os métodos utilizados para a pseudo-rotulagem nesse trabalho foram: *k-Nearest Neighbors* (KNN), um método supervisionado de aprendizado de máquina, onde o valor de k representa a quantidade de vizinhos mais próximos utilizados para prever um novo rótulo, podendo ser utilizado tanto para tarefas de classificação quanto de regressão; 1NN, uma variação do KNN, onde o valor de k é igual a 1; *Gaussian Mixture Model* (GMM), uma técnica de modelagem estatística que assume que o conjunto de dados é formado por uma combinação de diferentes distribuições gaussianas.

Os primeiros testes realizados foram feitos para a classificação multi-classe, comparando a acurácia e o tempo de execução do algoritmo proposto (com os *wrappers* 1NN, KNN e GMM) com o algoritmo IOE (*Improved Online Ensembles*) [4], utilizando 99% dos dados não rotulados. Também foi realizado o teste de Nemenyi para verificar se houve alguma diferença significativa entre os algoritmos. Com base nos resultados obtidos é possível observar na Figura 1 que o algoritmo proposto executou, em média, aproximadamente 10 vezes mais rápido que o IOE, ao mesmo tempo que manteve uma equivalência estatística em relação a acurácia, como mostra a Figura 2.

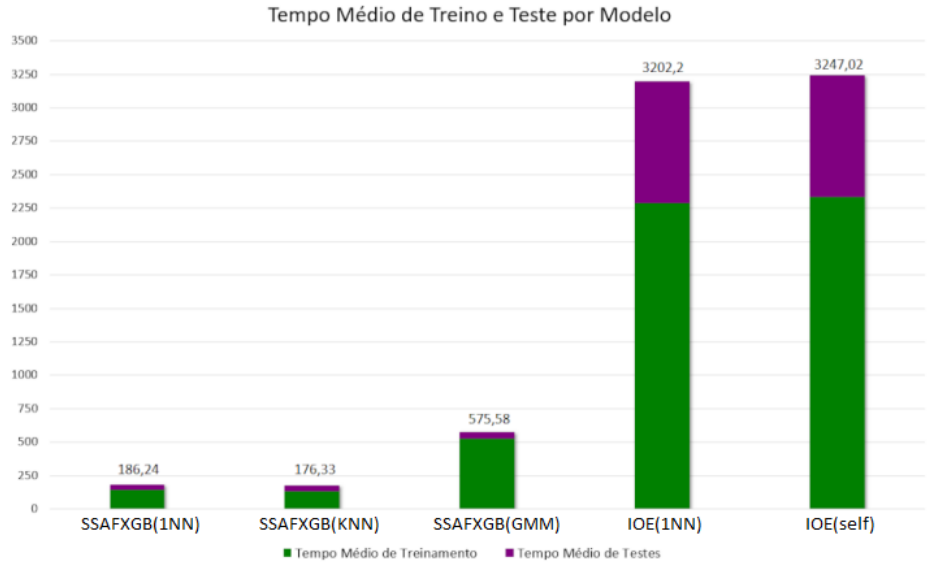

Figura 1. *Tempo médio de treino e teste por modelo.*

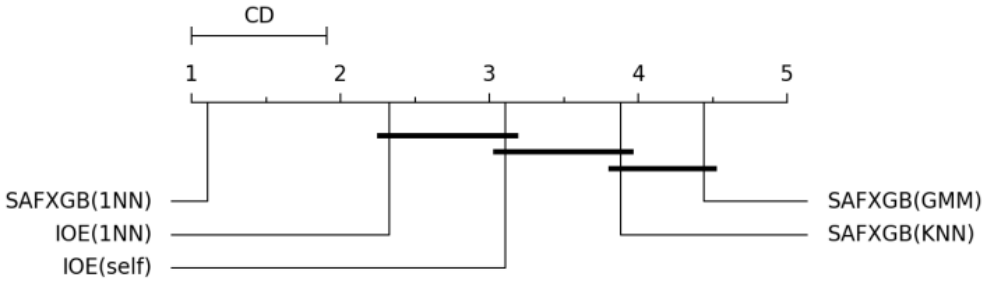

Figura 2. *Teste de Nemenyi dos algoritmos com 99% dos dados não rotulados.*

Palavras-chave: Aprendizado de máquina. Fluxo de dados. Aprendizado semi-supervisionado.

Referências

[1] MONTIEL, J. et al. Adaptive XGBoost for evolving data streams. In: **2020 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2020. p. 1–8.

[2] LU, J. et al. Learning under concept drift: A review. **IEEE Transactions on Knowledge and Data Engineering**, v. 31, n. 12, p. 2346–2363, 2019.

[3] ENGELEN, J. E. V.; HOOS, H. H. A survey on semi-supervised learning. **Machine learning**, Springer, v. 109, n. 2, p. 373–440, 2020.

[4] VAFAIE, P.; VIKTOR, H.; MICHALOWSKI, W. Multi-class imbalanced semi-supervised learning from streams through online ensembles. In: **2020 International Conference on Data Mining Workshops (ICDMW)**. [S.l.: s.n.], 2020. p. 867–874.