PROPOSTA DE ALGORITMO DE SELF-TRAINING PARA CLASSIFICAÇÃO SEMI-SUPERVISIONADA DE FLUXOS DE DADOS USANDO ENSEMBLE XGBOOST[1]

Igor Schiessl Froehner[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 CNPQ
[3] Orientador, Departamento de Ciência da Computação – CCT – fabiano.baldo@udesc.br

Classificação supervisionada é uma tarefa de aprendizado de máquina que pode ser muito custosa, pois requer uma grande quantidade de dados classificados, que geralmente são gerados por especialistas ou por uma mineração de dados previamente obtidos [2]. Para tratar tal questão existem duas abordagens: aprendizado semi-supervisionado e não supervisionado. O aprendizado semi-supervisionado visa não só extrair informação das instâncias classificadas como também das sem classe, dessa forma há menos custo em classificar as instâncias. Há várias abordagens possíveis para atingir tal objetivo: *co-training*, *self-training*, algoritmos baseados em grafo, *support vector machines*, etc [2]. A maioria dessas abordagens requer algumas premissas sobre o tipo do conjunto de dados para que uma boa acurácia seja atingida, neste trabalho será usado self-training, método que não requer tais premissas. Tal método funciona da seguinte forma: quando uma instância não classificada chega ao classificador ele mesmo a classifica e a usa para o seu treinamento. Porém, como o método se baseia somente na informação que ele mesmo gera, a acurácia pode se deteriorar em algum momento, principalmente quando há poucas instâncias classificadas. Entretanto, há técnicas para tratar a deterioração da eficácia, uma delas é medir a confiança da predição, assim pode-se analisar a probabilidade de ter acertado a classificação e, dessa forma, saber se aquele dado será útil ou não. Ainda sobre a classificação, há cenários em que a natureza dos dados não é estacionária, ou seja, as instâncias chegam de forma contínua e a quantidade pode ser infinita, esse é o caso dos Fluxos de Dados. Em um fluxo pode haver alterações no significado dos dados, ou seja, mudanças de conceito. Os modelos devem identificar essas mudanças e serem treinados para incorporá-las [4].

Neste trabalho, o método de self-training será aplicado da seguinte forma: o próprio classificador que está sendo treinado fará a classificação de uma instância que chegar no fluxo sem classe, dando como saída também a incerteza de que tenha acertado tal classificação. A incerteza foi calculada usando a métrica de *margem*. Então, se a incerteza for baixa, essa instância será usada no treinamento do classificador, usando como classe a pseudo-rotulagem predita previamente, senão, a instância é descartada. O classificador base utilizado foi o AXGB, que é um *ensemble* de árvores de decisão baseado em Extreme Gradient Boost (XGBoost), adaptado para fluxos de dados [3]. O XGBoost conta com várias otimizações quanto à esparsidade dos dados e escalabilidade e é tido como estado da arte em vários cenários de classificação [1].

Os testes foram feitos com objetivo de comparar o desempenho do método proposto com o método usado como base, o AXGB, e também visando testar como o método desempenha diante de alguns cenários diferentes no quesito de mudanças de conceito. Portanto, foram feitos testes nos seguintes cenários: um teste em que o fluxo sofre mudança de conceito repentina

(SEA), outro em que a mudança se dá de forma incremental e outro de forma gradual. O método de avaliação usado nos testes foi o *Evaluate Prequential*, em que o modelo é testado e depois treinado, com 500 mil instâncias e 1% dos dados rotulados, e o mínimo de confiança para que uma instância entrasse no dataset de treinamento foi de 90%. Os parâmetros relevantes usados no treinamento do XGBoost foram os seguintes: taxa de aprendizagem: 0.2; maior profundidade de árvore: 6. Alguns dos testes feitos podem ser vistos na Figura 1.

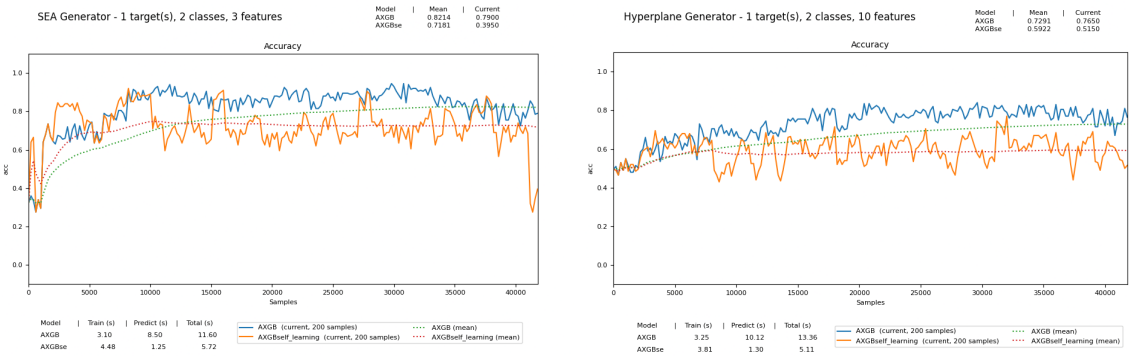

Figura 1. *Gráfico dos Testes*

Pelos gráficos dos resultados apresentados na Figura 1, é possível observar que na maior parte do tempo o método proposto teve relativa equiparação com o método ao qual foi comparado. Em alguns momentos ele teve um desempenho levemente melhor, em outros foi igual e em outros foi pior. Foi possível observar que em momentos de mudança de conceito o AXGB respondeu melhor que o método proposto, portanto uma das considerações que podem ser feitas para melhorar os resultados são no sentido de tornar o método mais adaptável às mudanças de conceito. Outra consideração que pode ser feita para melhorar os resultados é fazer um melhor refinamento nos parâmetros de treinamento do classificador XGBoost. Como conclusão, pode-se inferir que apesar do método proposto não ter se saído melhor do que o AXGB, ele mostrou-se promissor na falta de dados rotulados. Porém, faltou um melhor tratamento para as mudanças de conceito e para a deterioração da acurácia causada pelo self-learning.

Palavras-chave: Aprendizado de Máquina. Classificação semi-supervisionada. Fluxos de dados.

Referências

[1] CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. 2016. p. 785-794.

[2] VAN ENGELEN, Jesper E.; HOOS, Holger H. A survey on semi-supervised learning. **Machine Learning**, v. 109, n. 2, p. 373-440, 2020.

[3] MONTIEL, Jacob et al. Adaptive XGBoost for evolving data streams. In: **2020 International Joint Conference on Neural Networks (IJCNN)**. IEEE, 2020. p. 1-8.

[4] GAMA, Joao. **Knowledge discovery from data streams**. CRC Press, 2010.