Análise e aplicação de algoritmos na classificação incremental em fluxo de dados parcialmente classificados [1]

Vitor Bernstorff Clemes [1], Fabiano Baldo [2]

[1] Vinculado ao projeto "Stream Mining – Novas Abordagens para Algoritmos de Aprendizagem em Fluxos de Dados Não Estacionários"
[2] Acadêmico do Curso de Ciências da Computação – CCT – Bolsista PROBIC
[3] Fabiano Baldo, Departamento de Ciências da Computação – CCT – fabiano.baldo@udesc.br
[4] Doutor em Engenharia Elétrica – UFSC
[5] Acadêmico do Curso de Ciências da Computação – CCT

Na atualidade, informação é um dos fatores mais importantes para fundamentar a realização de qualquer projeto. Devido a isso, a coleta e análise de dados de forma computacional têm se expandido e desenvolvido exponencialmente. Em muitas situações, o conjunto de dados pode ser muito grande ou até infinito, ou seja, a base de informações cresce ao longo do tempo e em determinada ordem, a isso dá-se o nome fluxo de dados (FEIGENBAUM et al., 2002). Para tratá-los, a área de aprendizagem de máquina tem desenvolvido diversos classificadores, que são programas para mineração de fluxos de dados, permitindo a extração de estruturas de conhecimento a partir de eventos de fluxo (GAMA et al., 2010). Ao serem desenvolvidos, são levados em conta resultados como desempenho computacional, acurácia, generalização, assim como o formato dos dados disponíveis. Este último é profundamente importante, pois dados são coletados de muitas maneiras diferentes, e por mais generalista que os classificadores tendem a ser, os resultados podem variar consideravelmente dependendo da estrutura dos dados. Entre as diversas abordagens, existe o estudo semi-supervisionado, onde as bases de dados de treinamento possuem instâncias parcialmente classificadas.

Um dos principais classificadores existentes são as árvores de classificação. Tais estruturas são montadas como grafos, criadas a partir de um processo que divide os atributos presentes nas instâncias de dados, criando galhos para cada um desses atributos. O processo de separação continua até que cada galho possa ser rotulado com apenas uma classe (BRAMMER et al., 2007). Em uma de suas várias implementações, denominada *Hoeffding Adaptive Tree* e disponível no framework *Massive On-line Analysis* (MOA), busca-se encontrar um equilíbrio entre generalização e desempenho baseando-se na desigualdade de *Hoeffding*. Neste trabalho, buscou-se encontrar uma maneira de introduzir a indução semi-supervisionada baseada no conceito de impureza apresentado por Levatić et al. (2017).

Tal conceito difere dos modelos tradicionais por se propor a utilizar não apenas os rótulos das instâncias, mas também as estatísticas coletadas de instâncias não rotuladas para modelar o classificador. Esse recurso utiliza-se de métricas já documentadas como coeficiente de Gini e variância para calcular o que Levatić denominou como Impureza Semi-supervisionada de um conjunto de dados. O objetivo do algoritmo proposto, denominado *SSL Hoeffding Adaptive Tree*, foi atingir resultados que comprovassem a eficiência do algoritmo em relação aos anteriores.

Para alcançar esse objetivo, foram criadas bases parcialmente classificadas baseadas num valor percentual. Depois, foram introduzidos os cálculos da nova métrica e seus resultados foram utilizados para geração da árvore. Para que isso fosse possível, foi necessária uma adaptação

completa no funcionamento de algumas estruturas internas do classificador. Além disso, foi necessário criar novas rotinas para lidar com dados não classificados dentro do MOA, visto que ele não estava adaptado a dados sem classes.

Os resultados abaixo derivam de uma base de dados sintética gerada por uma ferramenta disponível no próprio MOA denominado *Random Tree Generator*. Essa base contém 1.000.000 de instâncias geradas instantaneamente e introduzidas em formato de fluxo, através do método *Evalute Prequential*. Cada instância possui 5 atributos nominais, 5 atributos numéricos e classificação binária.

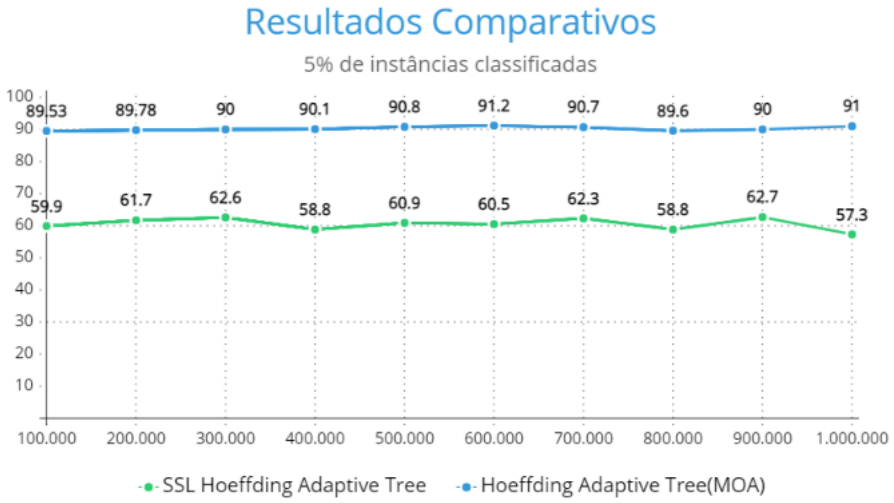

Gráfico 1. *Comparação de acurácia dos algoritmos.*

Conforme pode ser analisado no gráfico, o resultado obtido não foi satisfatório. Suspeita-se que isso se deve à critério(s) conceitual(ais) não explorado(s) acerca da desigualdade de *Hoeffding*, ou então devido às profundas alterações nas estruturas dos componentes internos do classificador. Também foi constatado que, em comparação com outros algoritmos que já adotaram a métrica, a impureza proposta por Levatić não se adapta muito bem a algoritmos incrementais.

Palavras-chave: Classificação de dados, Machine Learning, algorítimos.

Bramer, Max. (2007). Principles of Data Mining. 10.1007/978-1-84628-766-4.

FEIGENBAUM, J.; KANNAN, S.; STRAUSS, M. J.; VISWANATHAN, M. An approximate l 1-difference algorithm for massive data streams. SIAM Journal on Computing, SIAM, v. 32, n. 1, p. 131–151, 2002.

GAMA, J.; RODRIGUES, P. P.; SPINOSA, E. J.; CARVALHO, A. C. P. L. F. de. Knowledge discovery from data streams. [S.l.]: Chapman & Hall/CRC Boca Raton, 2010.

LEVATIĆ, Jurica et al. Semi-supervised classification trees. **Journal of Intelligent Information Systems**, v. 49, n. 3, p. 461-486, 2017.