Efficient online tree, rule-based and distance-based algorithms

Saulo Martiello Mastelini[1], André Carlos Ponce de Leon Ferreira de Carvalho[1]

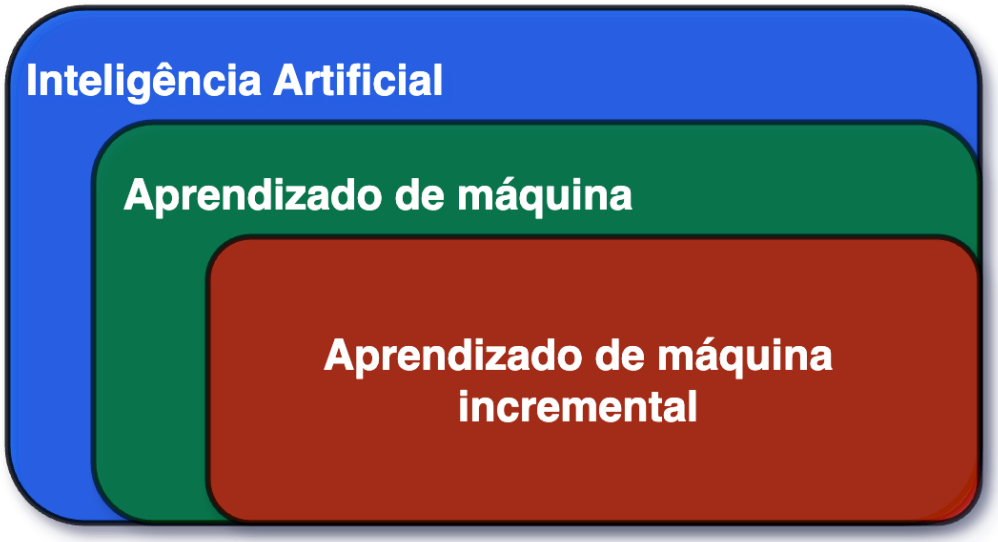
[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.
MASTELINI@ALUMNI.USP.BR

saulomastelini@gmail.com

Julho de 2024

Introdução: qual é o fluxo (de dados)?

- Grande quantidade de dados disponíveis
- Necessidade de processamento incremental
- Aprendizado de máquina incremental
  - Aprendizado de máquina em fluxos de dados
  - *Online Machine Learning* (OML)



A corrente vigente

- Várias soluções propostas no decorrer dos anos
- **Árvores** e **regras de decisão** e **comitês** (*ensembles*) desses modelos estão entre as mais populares
  - Algoritmo base: *Hoeffding Trees*
  - Utilizam a inequalidade de *Hoeffding*[§] como subsídio para decidir quando efetuar divisões
- Algoritmos baseados em proximidade (k-NN – vizinhos mais próximos)
  - Busca exaustiva em janelas deslizantes de dados

[§] Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301), 13-30.

Requisitos de um algoritmo de OML

De acordo com *Bifet e Gavaldà (2009)*[§]:

1. Processar uma instância por vez e inspecioná-la apenas uma vez (no máximo);
2. Estar preparado para realizar previsões a todo momento;
3. Considerar que a natureza dos dados pode evoluir no decorrer do tempo;
4. Assumir que o **fluxo de dados é infinito**, mas processá-lo com **recursos computacionais finitos** (tempo e memória).

[§] Bifet, A., & Gavalda, R. (2009). Adaptive learning from evolving data streams. In *Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31-September 2, 2009. Proceedings 8* (pp. 249-260). Springer Berlin Heidelberg.

Indo contra o fluxo

- Grande enfoque histórico no **desempenho preditivo**
- Aplicações práticas de classificação e regressão necessitam de soluções eficientes
- Regressão recebeu menos atenção do que classificação
  - $f : X \rightarrow y, y \in \mathbb{R}$
- Foco em **reduzir os custos** computacionais de *Hoeffding Trees* e seus algoritmos derivados para regressão
  - Enfoque adicional em busca por vizinhos mais próximos
  - BEPE (Bolsa Estágio de Pesquisa no Exterior) FAPESP – Universidade do Porto

Uma humilde tentativa em aparar algumas arestas

Questões de pesquisa

Três questões de pesquisa guiaram o desenvolvimento da tese:

Q1: Os **custos computacionais** das *Hoeffding Trees* para regressão podem ser reduzidos **sem impactos significativos** no **desempenho preditivo**?

Q2: É possível criar **comitês mais eficientes** baseados em **árvores de decisão** que mantém um **desempenho preditivo comparável** à soluções da mesma categoria já estabelecidas?

Q3: Existe uma **alternativa eficiente** para realizar buscas por **vizinhos mais próximos** em uma **janela deslizante** de dados, ao invés de realizar uma busca exaustiva?

Questões de pesquisa

Três questões de pesquisa guiaram o desenvolvimento da tese:

Q1: Os **custos computacionais** das *Hoeffding Trees* para regressão podem ser reduzidos **sem impactos significativos** no **desempenho preditivo**?

Questões de pesquisa

Três questões de pesquisa guiaram o desenvolvimento da tese:

Q2: É possível criar **comitês mais eficientes** baseados em **árvores de decisão** que mantém um **desempenho preditivo comparável** à soluções da mesma categoria já estabelecidas?

Questões de pesquisa

Três questões de pesquisa guiaram o desenvolvimento da tese:

Q3: Existe uma **alternativa eficiente** para realizar buscas por **vizinhos mais próximos** em uma **janela deslizante** de dados, ao invés de realizar uma busca exaustiva?

Como naveguei os fluxos de dados

Objetivo principal: obter soluções eficientes (tempo e memória), sem prejuízos notáveis à acurácia

Hipótese: o uso de técnicas de **sumarização** e **amostragem** de dados tem o potencial para gerar soluções **eficientes** e **acuradas**

- A tese está organizada de forma cronológica, de acordo com as principais publicações obtidas no decorrer dos anos

Resumo das principais contribuições

1. Redução de custos para treinamento de árvores e regras de decisão para regressão
2. Criação de floresta de árvores aleatórias eficientes para regressão
3. Proposição de uma estrutura de indexação para busca por vizinhos mais próximos em janelas deslizantes de dados (*First in, first out* – FIFO)

Árvores para regressão

Q1

Os **custos computacionais** das *Hoeffding Trees* para regressão podem ser reduzidos **sem impactos significativos** no **desempenho preditivo**?

Ideia principal: quantização de dados[§]

- Monitoramento de atributos para as divisões:
  - Árvore binária de busca $\rightarrow$ Tabela *hash*

Custo	**Antes**	**Tese**
Inserção	$O(\log n)$	$O(1)$
Memória	$O(n)$	$O(q)$, $q \ll n$
Busca	$O(n)$	$O(q \log q)$

[§] Mastelini, S. M., & de Leon Ferreira, A. C. P. (2021). Using dynamical quantization to perform split attempts in online tree regressors. *Pattern Recognition Letters*, 145, 37-42.

Floresta para regressão

Q2

É possível criar **comitês mais eficientes** baseados em **árvores de decisão** que mantém um **desempenho preditivo comparável** à soluções da mesma categoria já estabelecidas?

Ideia principal: amostragem aleatória[§]

- *Sub-bagging* para treinamento de árvores
- Seleção de limiares aleatórios para cortes
- Escolha aleatória de atributos para monitoramento

Custo	**Antes**	**Tese**
Inserção	$O(\log n)$	$O(1)$
Memória	$O(n)$	$O(1)$
Busca	$O(n)$	$O(1)$

[§] Mastelini, S. M., Nakano, F. K., Vens, C., & de Leon Ferreira, A. C. P. (2022). Online extra trees regressor. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10), 6755-6767.

Busca por vizinhos mais próximos em janela deslizante

Q3

Existe uma **alternativa eficiente** para realizar buscas por **vizinhos mais próximos** em uma **janela deslizante** de dados, ao invés de realizar uma busca exaustiva?

Ideia principal: grafos de proximidade[§]

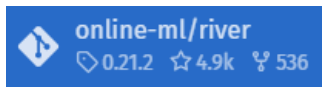
Custo	Antes	Tese
Inserção	$O(1)$	$O(B(c^2+1)+c^h)$
Memória	$O(n)$	$O(n+Bn)$
Busca	$O(n)$	$O(c^h)$

- B: fator de ramificação (depende dos dados)
- h: número de saltos de um vértice até o outro (depende dos dados)
- c: número de vértices candidatos para exploração (definido pelo usuário)

[§] Mastelini, S. M., Veloso, B., Halford, M., de Leon Ferreira, A. C. P., & Gama, J. (2024). SWINN: Efficient nearest neighbor search in sliding windows using graphs. *Information Fusion*, 101, 101979.

Um rio de código aberto

- Criador e mantenedor da Biblioteca River[§] para *Online Machine Learning*



- Biblioteca *open-source* mais popular para *Online Machine Learning*: cerca de 764 mil *downloads*
- Grande comunidade de usuários espalhados por todo o mundo
- Múltiplos projetos e ferramentas são construídos a partir do `River`
- Todas as principais contribuições da tese estão integradas à biblioteca

[§] Montiel, J., Halford, M., Mastelini, S. M., Bolmier, G., Sourty, R., Vaysse, R., ... & Bifet, A. (2021). River: machine learning for streaming data in python. *Journal of Machine Learning Research*, 22(110), 1-8.

Trabalhos futuros

- Florestas totalmente aleatórias: *(Aggregated) Mondrian Forests*
  - Classificação, Regressão e detecção de anomalias
  - Como controlar crescimento?
  - Estratégias eficientes de implementação para aplicações com disponibilidade limitada de recursos computacionais
- Grafos de proximidade
  - Tornar o processo de busca mais eficiente
  - Diminuir falsos positivos em configurações com maior limitação no uso de memória e tempo

Agradecimentos

- Minha família e amigos
- Professores André C. P. L. F. de Carvalho e João Gama
- FAPESP (processos #2018/07319-6 and #2021/10488-7)

Efficient online tree, rule-based and distance-based algorithms

Saulo Martiello Mastelini[1], André Carlos Ponce de Leon Ferreira de Carvalho[1]

[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.
MASTELINI@ALUMNI.USP.BR

saulomastelini@gmail.com

Julho de 2024