DecaLearn: Desafio de Decatlo em Aprendizado de Máquina

Prof. Dr. Artur Jordão Lima Correia

Departamento de Engenharia de Computação e Sistemas Digitais

Escola Politécnica da Universidade de São Paulo

E-mail: arturjordao@usp.br

6 de junho de 2023

Resumo

Aprendizado de máquina tem impulsionado avanços sem precedentes em direção à automatização de diversas tarefas cognitivas. Modelos modernos de aprendizado de máquina vêm adotando redes neurais e aprendizado profundo com o paradigma central de aprendizado. Sob este paradigma, em alguns cenários, modelos que compartilham propriedades de aprendizado similares são desconsiderados na avaliação. Esse problema se estende quando a avaliação considera, por exemplo, modelos não-paramétricos ou aprendizado profundo, porém não ambos. Tais fatores impossibilitam estabelecer o estado da arte e demonstrar se a representação do conhecimento de um modelo se transfere ao longo de diferentes domínios. Para contornar essas questões, este projeto propõe DecaLearn: uma avaliação unificada que desafiará diversos modelos de aprendizado de máquina em 10 (Decatlo) domínios distintos. Os objetivos deste projeto compreendem o desenvolvimento de modelos representativos de aprendizado de máquina, a avaliação desse modelos nas tarefas de geologia, indústria 4.0, cidades inteligentes e reconhecimento de atividades.

Palavras-chave: Aprendizado de Máquina, Aprendizado Profundo, *Benchmarking*.

1 Introdução e Justificativa

Inteligência artificial (IA) constitui uma importante linha de pesquisa na automação de tarefas cognitivas abrangendo tópicos como processamento de linguagem natural, visão computacional, cidades inteligentes e indústria 4.0. Nessas, e em muitas outras, tarefas o campo de aprendizado de máquina em IA vem obtendo resultados sem precedentes [1, 2, 3]. A ideia geral por trás do aprendizado de máquina reside na elaboração de modelos capazes de produzir hipóteses a partir de observações (dados) representando estados do mundo. Essa ideia corresponde à fase de aprendizado, onde os modelos aprendem a distinguir padrões a partir das observações

prévias. Após essa etapa, os modelos (denominados modelos preditivos) realizam predições a partir de dados novos.

Três principais fatores impulsionam grande parte do progresso no aprendizado de máquina: algoritmo, dados e hardware [4]. Os dois primeiros fatores referem-se, respectivamente, à inovação algorítmica dos modelos preditivos e à quantidade e qualidade dos dados disponíveis; ambos fatores essenciais para que os modelos representem adequadamente as hipóteses sobre o mundo e tomem decisões (processo de predição) precisas ao lidar com novos dados. O terceiro fator diz respeito aos avanços da tecnologia de hardware que possibilitam a utilização de modelos mais complexos e promovem alta performance ao manipular grandes volumes de dados.

No contexto de aprendizado de máquina, uma família de modelos em particular tem alavancado o estado da arte em várias aplicações: as redes neurais. Estudos recentes demonstraram que existem cenários onde redes neurais atingem resultados comparáveis, até mesmo superiores, aos humanos [5, 6]. Consequentemente, redes neurais e aprendizado profundo tornaram-se o paradigma central de aprendizado. Adotar precipitadamente esse paradigma, entretanto, pode criar uma percepção equivocada de que outros modelos não são igualmente efetivos. Importantemente, essa questão persiste mesmo em modelos pertencentes à categoria de redes neurais. Colocando em perspectiva, na área de geologia, estudos existentes na predição da produção de óleo predição e classificação de litologia, frequentemente, adotam modelos recorrentes (ex. RNNs e LSTMs) [7, 8]; entretanto, a avaliação nesses trabalhos desconsidera modelos que compartilham propriedades de aprendizagem similares e até mesmo superiores [9]. Em algumas aplicações, portanto, a avaliação não é abrangente suficiente para estabelecer o estado da arte. Além disso, o desempenho preditivo de um modelo em particular pode não se traduzir adequadamente quando avaliado em outro domínio. Para demonstrar e contornar esse último ponto, Rebuffi et al. [10] propôs o *Visual Domain Decathlon* que avaliou redes neurais em 10 aplicações distintas. Apesar de promover contribuições importantes, o trabalho de Rebuffi et al. [10] limitou-se ao domínio visual, em particular, à classificação de imagens.

Motivado pela avaliação limitada (em termos da faixa de modelos preditivos considerados) em algumas áreas de aplicação e por estudos prévios [10], este projeto de pesquisa propõe o DecaLearn, uma avaliação unificada que desafiará diversos modelos de aprendizado de máquina em diferentes aplicações. Especificamente, o projeto planeja quantificar a habilidade preditiva de modelos representativos de aprendizado de máquina em resolver simultaneamente 10 problemas cognitivos de diferentes aplicações, abrangendo temas como geologia, indústria 4.0, cidades inteligentes e reconhecimento de atividades.

Para elaborar o Decatlo, as seguintes categorias de modelos preditivos serão consideradas: *ensembles*, modelos paramétricos e não-paramétricos, redes neurais e aprendizado profundo. Além disso, o projeto incluirá mecanismos de *Automated Machine Learning*, tema que vem se popularizando cada vez mais no campo da IA [11]. A ideia por trás do Decatlo é responder a seguinte questão: Qual o desempenho das diversas moda-

lidades de modelos preditivos ao resolver simultaneamente[1] 10 problemas cognitivos de diferentes aplicações? Naturalmente, ao responder essa questão surgem outras questões que serão também respondidas durante o desenvolvimento do projeto:

- A alta habilidade preditiva de um modelo se transfere a múltiplos domínios?
- Qual o estado da arte nas tarefas consideradas?
- Qual modelo obtém o melhor desempenho levando em conta todas as aplicações? Em termos computacionais, esse modelo é adequado a todas essas tarefas?
- A aquisição de dados constituiu um processo árduo, custoso e suscetível a erros; desta forma, todo os modelos são apropriados a todas as tarefas?
- Algumas tarefas impõem desafios maiores de aprendizagem? Em caso positivo, quais modelos foram capazes de contornar tais desafios.

Acreditamos que responder as questões acima por meio do Decatlo impulsionará avanços ainda mais significativos nas tarefas consideradas.

2 Objetivos

Os objetivos deste projeto consiste em organizar um Decatlo (denominado DecaLearn) para modelos representativos de aprendizado de máquina, permitindo analisar o desempenho preditivo e computacional dos modelos (simultaneamente) em várias aplicações. Mais concretamente, este projeto possui os seguintes os objetivos. (i) Desenvolvimento de diversos modelos de aprendizado de máquina. (ii) Organizar bases de dados publicamente disponíveis para as aplicações consideradas no Decatlo. (iii) Avaliação dos modelos implementados nas bases de dados. (iv) Comparação e análise dos resultados obtidos.

Os objetivos e questões de pesquisa deste projeto limitam-se aos modelos supervisionados de aprendizado de máquina.

3 Detalhamento das Atividades

Durante o desenvolvimento deste projeto de iniciação científica as seguintes atividades serão desempenhas pelo aluno:

- Estudar conceitos teóricos e práticos de aprendizado de máquina incluindo aprendizado supervisionado, redes neurais, aprendizado profundo e *Automated Machine Learning*;

[1] O termo simultaneamente, refere-se ao desempenho preditivo geral (ex. médio) de um modelo em todas as aplicações.

- Ler artigos científicos relacionados ao tema do projeto;
- Implementar modelos representativos de aprendizado de máquina supervisionado;
- Organizar bases de dados disponíveis publicamente de aplicações de geologia, indústria 4.0, cidades inteligentes e reconhecimento de atividades;
- Organizar e documentar os códigos produzidos seguindo boas práticas de programação;
- Escrever relatórios técnicos e artigos científicos;
- Submeter artigos científicos a eventos de iniciação científica como, por exemplo, SIICUSP;
- Participar de eventos científicos.

4 Método

Para alcançar os objetivos propostos neste projeto, será adotado o seguinte planejamento metodológico. (1) Estudar conceitos teóricos e práticos sobre modelos preditivos que contemplem os seguintes temas de aprendizado de máquina: aprendizado supervisionado, redes neurais e aprendizado profundo, e *Automated Machine Learning*. (2) Revisar a literatura sobre as aplicações contempladas pelo Decatlo. (3) Estudar o desenvolvimento de modelos preditivos utilizando os *frameworks* scikit-learn, TensorFlow e/ou PyTorch. (4) Selecionar e organizar bases de dados das aplicações consideradas no Decatlo. (6) Selecionar modelos preditivos de alta performance e que sejam representativos das diferentes técnicas de aprendizado de máquina, incluindo mecanismos de *Automated Machine Learning*. (5) Conduzir experimentos (treinamento e teste) dos modelos selecionados nas bases de dados produzidas após a etapa de planejamento (4).

Devido à alta complexidade envolvendo o tema *Automated Machine Learning*, o aluno utilizará ferramentas prontas que necessitam somente designar os dados e definir poucos parâmetros de configuração. Exemplos representativos dessas ferramentas são: AutoSklearn[2], AutoKeras[3] e Fedot[4].

Para calcular o desempenho dos modelos serão utilizadas métricas de avaliação bem definidas na literatura das aplicações. Tipicamente essas métricas correspondem à acurácia e RMSE (*Root Mean Squared Error*) para problemas de classificação e regressão, respectivamente [12, 13, 9, 14]. Em virtude da diversidade e magnitude dessas métricas, para mensurar o desempenho geral dos modelos no Decatlo será contabilizado o número de ocorrências em que o modelo obteve o primeiro (Top1), segundo (Top2) e terceiro (Top3) melhor resultado. Seguindo tendências recentes em relação a questões ambientais do aprendizado de máquina [15, 16, 17, 18], o custo computacional dos modelos também será reportado; mais especificamente, o tempo de treinamento

[2]https://www.automl.org/automl-for-x/tabular-data/auto-sklearn/
[3]https://autokeras.com/
[4]https://fedot.readthedocs.io/en/latest/

que possibilita estimar a emissão de gás poluente (decorrente do consumo energético). Em particular, para os modelos de aprendizado profundo as métricas de número de parâmetros e quantidade operações de ponto flutuante também serão consideradas.

A construção do DecaLearn será na linguagem de programação Python. Finalmente, a realização dos experimentos será, preferencialmente, em ambientes de computação gratuitos na nuvem (ex. Google Colab[5] Gradient Papers Code[6] e DeepNote[7]) que fornecem suporte para diversas ferramentas de aprendizado de máquina e aprendizado profundo.

5 Resultados Esperados

Os resultados do presente projeto incluem a contribuição ao desenvolvimento da ciência e da engenharia. Mais concretamente, ao longo do desenvolvimento deste projeto espera-se que os resultados abaixo sejam obtidos.

- Desenvolvimento de uma variedade de modelos de aprendizado de máquina, contemplando desde redes neurais até modelos modernos gerados por *Automated Machine Learning*;
- Disponibilização dos códigos gerados em repositórios públicos utilizando a plataforma github;
- Definição de *Benchmarkings* e, consequentemente, o estado da arte para diversas aplicações cognitivas;
- Estabelecer boas práticas e estratégias de aprendizado de máquina;
- Facilitar o uso das bases de dados das aplicações consideradas a pesquisas futuras;

6 Cronograma de Execução

A Tabela 1 detalha o cronograma, em bimestres, das atividades envolvidas no projeto.

Tabela 1: Cronograma das atividades em bimestres.

Atividade	1	2	3	4	5	6
Leitura de artigos científicos	X	X	X	X	X	X
Seleção das aplicações e base de dados	X					
Organização das bases de dados das aplicações	X	X				
Implementação dos modelos preditivos	X	X	X	X		
Execução dos experimentos (Decatlo)		X	X	X	X	
Documentação dos códigos			X			X
Escrita de relatório parcial			X			
Escrita de artigo científico			X	X	X	X
Relatório Final						X

[5]https://research.google.com/colaboratory/
[6]https://www.paperspace.com/
[7]https://deepnote.com/

7 Outras Informações

Os dados que serão utilizados neste projeto estão disponíveis publicamente e as considerações éticas sobre os dados são de responsabilidade dos autores/pesquisadores que coletaram as bases de dados. Desta forma, este projeto não envolve questões éticas e está em conformidade legal. Finalmente, a utilização desses dados disponíveis publicamente permite transparência e reprodutibilidade na pesquisa.

Referências

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems (NeurIPS)*, 2017.

[2] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[3] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[4] OpenAI, "Ai and compute," https://openai.com/blog/ai-and-compute/, accessed: 2023-05-20.

[5] A. P. Badia, B. P. S. K. P. S. A. V. Z. Guoand, and C. Blundell, "Agent57: Outperforming the atari human benchmark," in *International Conference on International Conference on Machine Learning (ICML)*, 2020.

[6] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel, "Partial success in closing the gap between human and machine vision," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[7] D. T. d. Santos, M. Roisenberg, and M. d. S. Nascimento, "Deep recurrent neural networks approach to sedimentary facies classification using well logs," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[8] J. Lin, H. Li, N. Liu, J. Gao, and Z. Li, "Automatic lithology identification by applying LSTM to logging data: A case study in X tight rock reservoirs," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 8, pp. 1361–1365, 2021.

[9] A. Jordao, J. P. da Ponte Souza, M. C. Kuroda, M. F. de Rezende, H. Pedrini, and A. C. Vidal, "Towards automatic and accurate core-log processing," *Journal of Applied Geophysics*, p. 104990, 2023.

[10] S. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Neural Information Processing Systems (2017)*, 2017.

[11] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning - Methods, Systems, Challenges*, 2019.

[12] J. Sena, J. B. Santos, C. Caetano, G. Cramer, and W. R. Schwartz, "Human activity recognition based on smartphone and wearable sensors using multiscale DCNN ensemble," *Neurocomputing*, 2021.

[13] L. Jiang, X. Wang, W. Li, L. Wang, X. Yin, and L. Jia, "Hybrid multitask multi-information fusion deep learning for household short-term load forecasting," *IEEE Trans. Smart Grid*, 2021.

[14] S. Yin, Y. Huang, T. Chang, S. Chang, and V. S. Tseng, "Continual learning with attentive recurrent neural networks for temporal data classification," *Neural Networks*, 2023.

[15] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Conference of the Association for Computational Linguistics (ACL)*, 2019.

[16] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," in *Neural Information Processing Systems (NeurIPS)*, 2019.

[17] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, 2020.

[18] D. A. Patterson, J. Gonzalez, U. Hölzle, Q. V. Le, C. Liang, L. Munguia, D. Rothchild, D. R. So, M. Texier, and J. Dean, "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, 2022.