josif 2024
16ª JORNADA CIENTÍFICA E TECNOLÓGICA
13º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
ISSN: 2319-0124
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO CNPq FAPEMIG

MÉTODO PARA SELEÇÃO DE FEATURES E HIPERPARÂMETROS DE DEEP LEARNING COM ALGORITMO GENÉTICO

Luiz R. F. ALVES[1]; Diego SAQUI[2]; Heber R. MOREIRA[3]

RESUMO

Os avanços tecnológicos têm permitido o desenvolvimento de áreas como a Inteligência Artificial, em especial o *Deep Learning* (DL), que possui aplicações na área da saúde, como na predição de diabetes. Um dos desafios enfrentados nessa técnica é a seleção ideal de hiperparâmetros. Neste contexto, os Algoritmos Genéticos (AG) que, normalmente, são utilizados como otimizadores, são propostos em um método para otimizar a escolha de *features* e hiperparâmetros em problemas que utilizam modelos de DL. Para tal, utilizou-se a AG para seleção das melhores *features* de um *dataset* de Diabetes e hiperparâmetros de um modelo de DL. Após aplicação no *dataset* pode-se observar que o modelo com menos *features* pontuou melhor nas métricas de desempenho escolhidas quando comparado ao mesmo modelo que utilizava todas as *features* disponíveis. Portanto, pode-se concluir que o processo se mostrou como uma possível ferramenta na seleção de *features* e hiperparâmetros de modelos de DL.

Palavras-chave:
Inteligência Artificial; Otimização; Aprendizado de Máquina.

1. INTRODUÇÃO

Ao analisar os avanços tecnológicos, percebe-se que estes têm ocorrido de forma mais intensa quando comparados a décadas anteriores. Como consequência, computadores tiveram um aumento de poder computacional, permitindo a implementação de técnicas computacionais que antes não eram possíveis, favorecendo áreas como de Inteligência Artificial (IA). Tal área engloba a *Machine Learning* (ML), que possui subáreas como a *Deep Learning* (DL), a qual, no processo de aprendizagem, utiliza de uma abordagem de representações a partir de dados que enfatizam o aprendizado de camadas sucessivas de representações cada vez mais relevantes (CHOLLET, 2018).

Ainda referente a DL, esta se distingue de outros processos de ML no seu processo de aprendizado de uma determinada função $f(.)$, a qual ocorre por meio da combinação de outras funções $f_i(.)$, com i representando uma camada (PONTI; COSTA, 2017). Além disso, a DL possui utilizações importantes em diversas áreas. Exemplo disso são suas aplicações práticas na área da saúde, como, por exemplo, no campo da diabetes (ZHU; LI; HERRERO; GEORGIOU, 2020).

Entretanto, conforme Goodfellow, Bengio e Courville (2016), um dos desafios enfrentados por modelos de ML e também por modelos de DL, é a definição dos melhores hiperparâmetros para determinado modelo. Tais hiperparâmetros podem ser entendidos como configurações do modelo, os quais podem ser regulados para otimizar a qualidade do algoritmo de aprendizado e desempenho.

Ademais, dentro do contexto de IA, se tem os Algoritmos Genéticos (AG), os quais são

[1] Discente, IFSULDEMINAS – *Campus* Muzambinho. E-mail: luiz.ferlin@alunos.ifsuldeminas.edu.br.
[2] Orientador. Universidade Federal de Lavras - *Campus* São Sebastião do Paraíso. E-mail: diego.saqui@ufla.br.
[3] Coorientador. IFSULDEMINAS – *Campus* Muzambinho. E-mail: heber.moreira@muz.ifsuldeminas.edu.br.

normalmente utilizados em problemas de busca e otimização de parâmetros, baseados no conceito de seleção natural (POSE, 2000). Ele inicia seu processo com uma população de soluções (indivíduos), que são avaliadas de acordo com sua adaptação a um determinado problema (*fitness*). Assim, a partir dessa avaliação, os melhores indivíduos são selecionados para a etapa de reprodução, aplicando operadores como *crossover* e mutação, gerando uma nova população. Esse ciclo se repete até que uma condição de parada seja atingida (DUMITRESCU et al., 2000).

Considerando este contexto, percebe-se a importância da seleção de hiperparâmetros e *features* para a busca de bons resultados em um dado problema. Com base nisso, buscou-se propor e avaliar um método baseado em AG para seleção de *features* e hiperparâmetros de *DL*.

2. MATERIAIS E MÉTODOS

Para atingir o objetivo proposto, utilizou-se o *dataset* Diabetes advindo de um conjunto maior de dados do National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Este *dataset* contém diversos dados (*features*) de pacientes mulheres com pelo menos 21 anos de idade, de herança indígena Pima. Ele possui as *features*: *pregnancies*, sendo o número de vezes que engravidou, *glucose*, sendo a concentração plasmática de glicose em 2 horas em um teste oral de tolerância à glicose, *blood pressure*, sendo a pressão arterial diastólica (mm Hg), *skin thickness*, sendo a espessura da prega cutânea do tríceps (mm), *insulin*, sendo a insulina sérica de 2 horas (mu U/ml), *bmi*, sendo o índice de massa corporal (kg/m²), *diabetes pedigree function*, sendo a função de pedigree do diabetes, *age*, sendo a idade, e do *outcome,* indicando a presença ou não de diabetes.

Outrossim, foi utilizado, para execução deste trabalho, recursos como o Google Collaboratory[4], tal qual é uma ferramenta que permite a execução de código Python em nuvem, não havendo necessidade de instalação. Além disso, foram utilizadas bibliotecas como o Pandas[5], Scikit-Learn[6], seaborn[7], Matplotlib[8] e o PyTorch[9].

Ademais, o processo proposto possui o seguinte fluxo de execução: 1-) ocorre a inicialização da população inicial *P(N)*, os indivíduos gerados nesta etapa são utilizados na criação dos modelos de DL, tal quais são submetidos ao processo de treinamento. 2-) é calculada a pontuação de cada indivíduo. 3-) ocorre o armazenamento do estado atual da geração. 4-) é realizado a seleção dos indivíduos, ocorrem os processos de *crossover* e mutação, estabelecendo a nova população a qual passará pelo mesmo processo até que uma determinada condição de parada seja atingida. Este processo pode ser observado na Figura 1.

---

[4] https://colab.google. Acesso em Setembro de 2024.
[5] https://pandas.pydata.org. Acesso em Setembro de 2024.
[6] https://scikit-learn.org. Acesso em Setembro de 2024.
[7] https://seaborn.pydata.org. Acesso em Setembro de 2024.
[8] https://matplotlib.org. Acesso em Setembro de 2024.
[9] https://pytorch.org. Acesso em Setembro de 2024.

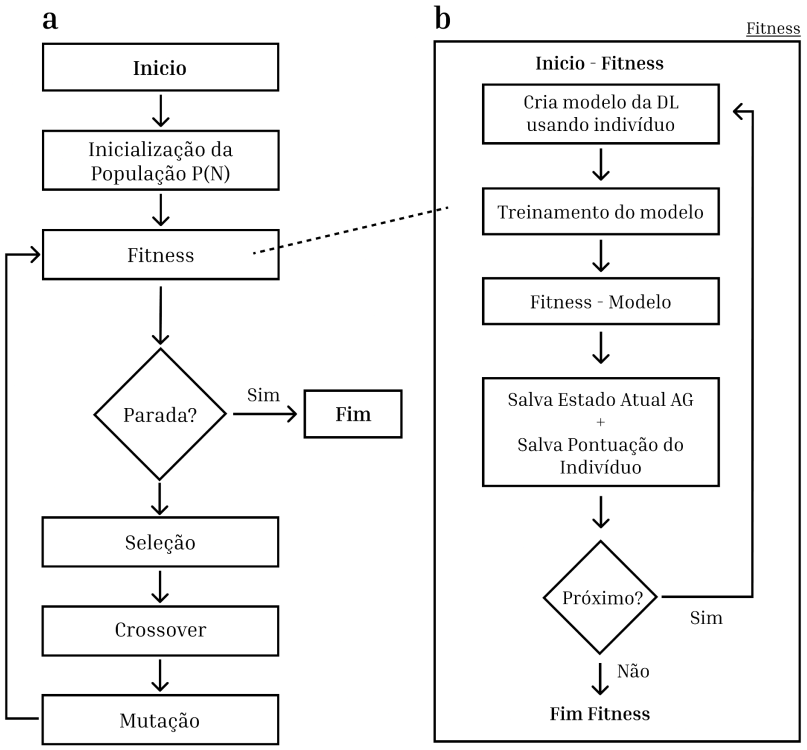

Figura 1: Fluxograma do método proposto

Após se aplicar o método no problema de predição de diabetes, faz-se necessária a avaliação do modelo resultante. Para isso, as métricas de desempenho: acurácia, precisão, revocação e *F1 score*; foram aplicadas, tanto no modelo com as *features* selecionadas, quanto no modelo com todas as *features* disponíveis.

3. RESULTADOS E DISCUSSÃO

Após aplicação e avaliação do modelo resultando do método no problema de predição de diabetes com DL e o *dataset* mencionado anteriormente, pode-se chegar na Tabela 1. Então, com base na Tabela 1, percebe-se que o modelo com o menor número de *features* (Modelo A) obteve uma melhor pontuação em todas as métricas de desempenho utilizadas quando comparado ao Modelo B, que utilizou todas as *features* disponíveis, uma vez que se pode haver *features* que influenciam negativamente no resultado. Porém, deve-se ressaltar que o modelo B, foi o mesmo utilizado no modelo A, advindo do método de seleção.

Métrica de desempenho	Modelo A	Modelo B
Acurácia	0.766 (76,6%)	0.746 (74,6%)
Precisão	0.714 (71,4%)	0.674 (67,4%)
Revocação	0.555 (55,5%)	0.537 (53,7%)
F1 Score	0.625 (62,5%)	0.597 (59,7%)

Tabela 1: Comparação de métricas do Modelo A e Modelo B

Vale salientar que o tempo de processamento da primeira geração foi o maior dentre todos, levando cerca de 5 horas para ser concluído. No entanto, as gerações subsequentes apresentaram tempos de execução menores em relação às anteriores. Isso ocorreu devido ao *fitness* não ser calculado para indivíduos já treinados, o que ocasionou, a partir da nona geração, um tempo de execução inferior a 1 hora. Deve-se ressaltar que isso acontece devido ao número relativamente pequeno de combinações de *features* e hiperparâmetros. Com uma maior quantidade de combinações possíveis, essa redução pode se tornar menos frequente, mas o tempo ainda será inferior ao que seria sem a aplicação desta técnica.

4. CONCLUSÃO

O objetivo deste trabalho foi propor e avaliar um método baseado em algoritmo genético para seleção de *features* e hiperparâmetros de DL. Desta forma, após aplicação da metodologia proposta no *dataset* de Diabetes, pode-se dizer que o método se mostrou como uma possível ferramenta para seleção de *features* e hiperparâmetros de DL sem a necessidade de realizar empiricamente testes para a seleção dos mesmos, embora exija um tempo considerável de processamento para aquisição dos resultados.

REFERÊNCIAS

CHOLLET, François. **Deep Learning with Python**. [S. l.]: Manning Publications Co, 2018. Disponível em: https://www.manning.com/books/deep-learning-with-python.

DUMITRESCU, D. et al. **Evolutionary Computation**. 1. ed. Boca Raton: CRC Press, 2000. Disponível em: https://www.taylorfrancis.com/books/mono/10.1201/9781482273960/evolutionary-computation-dumitrescu-dumitrescu-beatrice-lazzerini-lakhmi-jain.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. Disponível em: https://www.deeplearningbook.org.

PONTI, Moacir Antonelli; COSTA, Gabriel B. Paranhos da. Como funciona o Deep Learning. In: 32ND BRAZILIAN SYMPOSIUM ON DATABASES, 2017, Uberlândia - MG. **Proceedings** [...]. [S. l.: s. n.], 2017. Disponível em: https://sbbd.org.br/2017.

POSE, Marcos Gestal. **Introducción a los algoritmos genéticos**. Universidad de Coruña: Departamento de Tecnologías de la Información y las Comunicaciones, 2000. Disponível em: https://cursa.ihmc.us/rid=1KNKMJ4LN-11XXFSG-1KV5/Algoritmos%20de%20Terminos.pdf.

ZHU, Taiyu; LI, Kezhi; HERRERO, Pau; GEORGIOU, Pantelis. Deep learning for diabetes: a systematic review. **IEEE Journal of Biomedical and Health Informatics**, [s. l.], v. 25, ed. 7, p. 2744-2757, 2020. Disponível em: https://ieeexplore.ieee.org/abstract/document/9268187.