Aplicação de técnicas de Aprendizado de Máquina na classificação de câncer de mama

Maiane Junqueira e Rafael D'Ávila

***Resumo*— O uso de Inteligência Artificial é benéfico em várias aplicações médicas, tais como classificação de tumores de câncer de mama e diagnóstico de doenças como Alzheimer. Neste contexto, o objetivo do nosso projeto é combinar a utilização de técnicas de Aprendizado de Máquina e Computação Evolutiva para auxiliar no diagnóstico precoce de câncer de mama.**

I. Descrição do Problema

Câncer de mama é uma doença que assola mulheres e corpos femininos ao redor do mundo [1]. O diagnóstico precoce é importante para o tratamento adequado e possível cura para a paciente. Usualmente, a análise e identificação da presença de tumor é realizada por profissionais habilitados, no entanto esta abordagem é passível de erro humano. Dessa forma, o uso de Inteligência Artificial (IA) pode auxiliar na identificação de tumores incipientes em estágio inicial [2][4]. Este trabalho tem como objetivo comparar o uso de técnicas de Aprendizado de Máquina (*Multilayer Perceptron* e *K-Nearest Neighbors*) e Computação Evolutiva (Algoritmos Genéticos) para a detecção de tumores malignos de câncer de mama [1][5] [6].

II. Métodos e Implementação

O *dataset* utilizado foi o *Breast Cancer Wisconsin (Diagnostic) Data Set*, obtido através do repositório *UCI Machine Learning Repository*. O *dataset* é composto por 569 instâncias com 32 atributos cada (ID, Diagnóstico e 30 atributos descritivos).

De modo a classificar os tumores entre malignos e benignos, utilizamos os modelos Rede Neural *Multilayer Perceptron* (MLP) e *K-Nearest Neighbors* (kNN). Para o treinamento dos modelos foi utilizado o método de *hold-out*, separando os dados em 80% para o conjunto de treinamento e os outros 20% para o conjunto de teste. A proporção das classes no *dataset* original foi mantida em ambos os conjuntos.

III. Resultados

A. Multilayer Perceptron

Foram construídas 3 arquiteturas diferentes de MLP sendo a primeira delas com uma camada e 128 neurônios, a segunda com duas camadas de 128 e 64 neurônios, respectivamente, e, por último, a terceira com três camadas de 128, 64 e 32 neurônios, respectivamente. Os dados foram padronizados e o treinamento foi realizado por 10 épocas, com taxa de aprendizado 10^{-4} e tamanho de *batch* 16. Foi utilizado o otimizador *Adam* com a função custo de Entropia Cruzada. Ao final do treinamento, comparamos a acurácia das três arquiteturas, cujo resultado pode ser observado na Tabela I.

TABELA I
ACURÁCIA DAS TRÊS ARQUITETURAS DE MLP

MLP	Acurácia
MLP-1	0.94
MLP-2	0.96
MLP-3	0.96

Das três estruturas mencionadas, escolhemos analisar a MLP com uma única camada, dado que o aumento do número de camadas intermediárias não resultou em uma melhora significativa das métricas utilizadas. Sendo assim, na Figura 1 podemos observar o matriz de confusão da MLP com uma camada, cuja acurácia foi de 94%

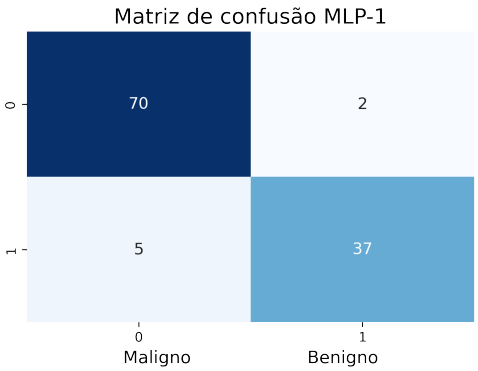

Fig. 1
MATRIZ DE CONFUSÃO DE MLP COM UMA CAMADA INTERMEDIÁRIA.

Note que o número de falsos positivos e falsos negativos foram de 2 e 5, respectivamente. Em outras palavras, duas pessoas foram diagnosticadas com câncer sem ter a doença e cinco pessoas com câncer não terão a chance de iniciar o tratamento imediatamente. Uma vez que iniciar o tratamento rapidamente traduz-se em maiores chances de cura, nosso objetivo principal foi diminuir o número de falsos negativos. Para isto, atribuímos diferentes pesos para as classes, sendo peso 10 para a classe positiva e peso 1 para a classe negativa. A matriz de confusão obtida encontra-se na Figura 2. Neste cenário, a acurácia foi de 88% e nenhum caso de falso negativo foi obtido; é interessante notar que o número de falsos positivos aumentou e, consequentemente, a acurácia diminuiu,

Maiane Junqueira e Rafael D'Ávila, Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade Estadual de Campinas (UNICAMP), Campinas-SP. E-mails: m190653@dac.unicamp.br, r149387@dac.unicamp.br .

ou seja, atribuir pesos diferentes para as classes resulta em um *trade-off* entre acurácia e o número de casos falsos negativos.

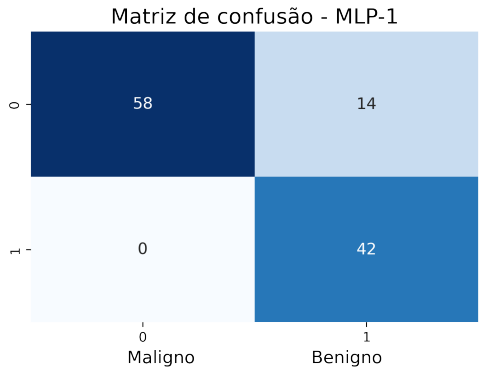

Fig. 2
MATRIZ DE CONFUSÃO MLP-1 COM PESOS.

B. K-Nearest Neighbors

Para o KNN, normalizamos os dados no intervalo [0,1], tomando o cuidado de utilizar os parâmetros calculados no conjunto de treinamento para normalizar o conjunto de teste, de forma a não permitir que o modelo obtenha informações a respeito do conjunto do teste. Para o treinamento, utilizamos os seguintes valores de k: 1, 3, 5, 7, 9 e 11. Computamos a acurácia e número de falsos negativos no conjunto de teste para cada valor de k. Nas figuras 3 e 4, temos os valores de acurácia e número de falsos negativos

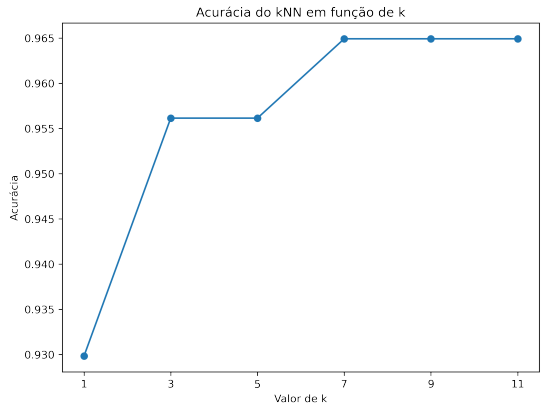

Fig. 3
ACURÁCIA DO KNN PARA DIFERENTES VALORES DE K.

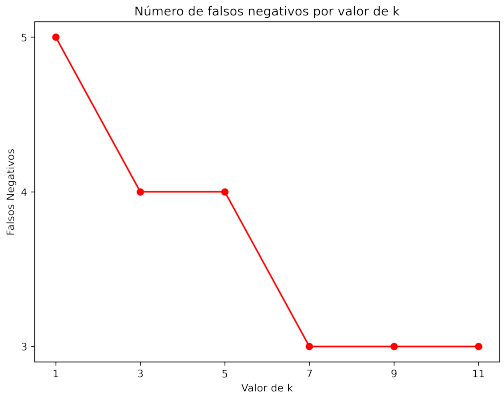

Fig. 4
FALSOS NEGATIVOS PARA DIFERENTES VALORES DE K.

É possível notar que, para $k = 7$, temos a acurácia máxima e o menor número de casos falsos negativos, de modo que não ocorre melhora nos valores subsequentes de k. Portanto, é possível concluir que o valor ótimo, neste caso, é dado por $k = 7$. Na figura 5 abaixo, encontra-se a matriz de confusão do modelo para $k = 7$

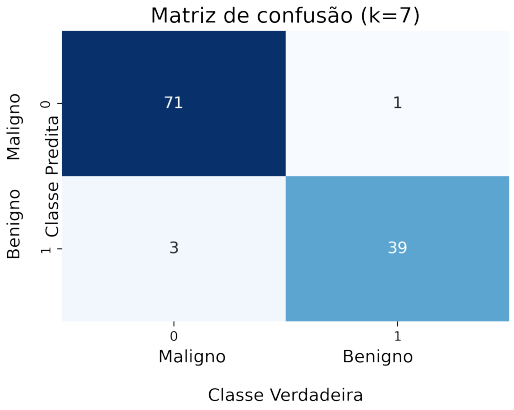

Fig. 5
MATRIZ DE CONFUSÃO KNN PARA NÚMERO DE VIZINHOS IGUAL A 7.

Para o cenário acima, o número de falsos negativos e falsos positivos foram 3 e 1, respectivamente. Além disso, a acurácia do kNN foi 97%, enquanto a acurácia da MLP com 1 camada intermediária foi 94%, de acordo com a Tabela I. O fato de termos um *dataset* pequeno e relativamente simples, aliado com a propriedade do kNN de explorar a similaridade local dos dados, provavelmente explica o desempenho superior do mesmo em relação à rede MLP.

C. Algoritmos genéticos

Dado que avaliamos o desempenho da MLP e KNN para classificar adequadamente os tumores em malignos e benignos, utilizaremos um algoritmo genético (AG) para a seleção de *features*.

Para isto, definimos um indivíduo como um vetor binário com 30 genes, em que cada posição corresponde a um atributo

do *dataset*. Se o valor daquela posição for 1, significa que a respectiva *feature* será usada e 0 caso contrário. Para que o modelo utilizasse um menor número de atributos, estabelecemos a função de *fitness* como

$$\text{acurácia} - \lambda \cdot \left(\frac{\text{número de atributos utilizados}}{\text{número total de atributos}} \right) \quad (1)$$

com $\lambda = 0.5$ escolhido arbitrariamente.

Utilizamos o torneio com alta pressão seletiva para selecionar aleatoriamente dois indivíduos da população (pais + filhos) e mantivemos o indivíduo de maior *fitness*. O operador de recombinação foi *crossover* simples, que seleciona aleatoriamente uma posição do vetor para o cruzamento e mutação em que todos os bits do indivíduo apresentavam 10% de chance de mutação. Foi criada uma população de 50 indivíduos, a qual se reproduziu por 10 gerações.

No caso da rede MLP, foi utilizada a arquitetura de uma camada intermediária com 128 neurônios para avaliar o *fitness* dos indivíduos. Na figura 6 encontra-se a evolução do *fitness* do melhor indivíduo ao longo das gerações. É possível notar que o desempenho da MLP oscila ao longo das gerações, isto se deve por conta de uma população inicial de soluções inadequadas ou mutações desfavoráveis.De um modo geral, o algoritmo foi capaz de convergir e apresentar uma solução final adequada com uma acurácia de 97% na décima geração.

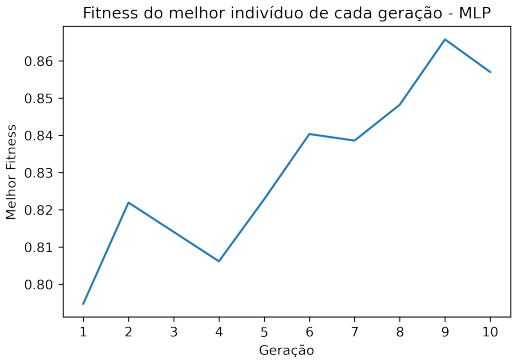

Fig. 6

FITNESS DO MELHOR INDIVÍDUO POR GERAÇÃO - MLP.

Para o kNN, foi utilizado o valor ótimo $k = 7$. Na figura 7 observamos a evolução do *fitness* do melhor indivíduo ao longo das gerações. Nela, vemos que o kNN apresentou uma evolução praticamente constante ao longo das gerações e estabilizou próximo à nona geração com uma acurácia de 96%. Sendo assim, o kNN mostrou-se menos sensível às inicializações e/ou mutações do AG.

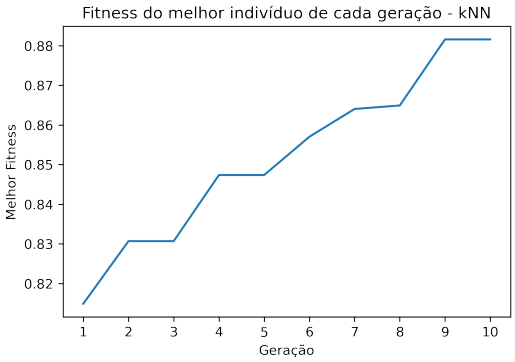

Fig. 7

FITNESS DO MELHOR INDIVÍDUO POR GERAÇÃO - KNN.

IV. DISCUSSÃO

Podemos notar que o kNN não melhorou seu desempenho com a extração de *features* dado que esta caiu de 97% para 96%. É uma leve queda que não afeta o algoritmo, mas faz com que tenhamos que analisar se é necessário utilizar extração de atributos através de um AG para o kNN. Neste caso, é melhor optar pela simplicidade do algoritmo. Por outro lado, a MLP se beneficiou com a extração de *features* pelo AG, pois a acurácia desta foi de 94% para 97%. Provavelmente, a rede MLP conseguiu extrair uma relação mais complexa entre os atributos selecionados. Embora o uso do AG tenha sido benéfico, não podemos nos esquecer que a MLP é um modelo com maior complexidade e custo computacional quando comparado ao kNN.

Por último, queremos salientar que o avanço do uso de inteligência artificial no cotidiano deve se estender para possibilitar o diagnóstico precoce de câncer e de outras doenças, de modo que um tratamento gratuito e de boa qualidade seja acessível para toda a sociedade.

REFERÊNCIAS

[1] Mohammed, Siham A., et al. "Analysis of breast cancer detection using different machine learning techniques." Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5. Springer Singapore, 2020.

[2] Amrane, Meriem, et al. "Breast cancer classification using machine learning." 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT). IEEE, 2018.

[3] Mohammed, Siham A., et al. "Analysis of breast cancer detection using different machine learning techniques." Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5. Springer Singapore, 2020.

[4] Epimack Michael, He Ma, Hong Li, Shouliang Qi, "An Optimized Framework for Breast Cancer Classification Using Machine Learning", BioMed Research International, vol. 2022, Article ID 8482022, 18 pages, 2022. https://doi.org/10.1155/2022/8482022

[5] Aalaei, Shokoufeh, et al. "Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets." Iranian journal of basic medical sciences 19.5 (2016): 476.

[6] Talatian Azad, Saeed, Gholamreza Ahmadi, and Amin Rezaeipanah. "An intelligent ensemble classification method based on multi-layer perceptron neural network and evolutionary algorithms for breast cancer diagnosis." Journal of Experimental & Theoretical Artificial Intelligence 34.6 (2022): 949-969.