Aplicação de redes neurais multilayer perceptron em bioinformática para a classificação de espécies do conjunto de dados íris

Vitor Amadeu Souza[1]; 0009-0002-1857-6799

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
vitor.amadeu@foa.org.br

Resumo: Este trabalho apresenta a implementação e análise de uma rede neural artificial do tipo Multilayer Perceptron (MLP) para classificação de espécies de flores do conjunto de dados Iris. O estudo utilizou a biblioteca scikit-learn do Python para desenvolver um modelo classificador capaz de distinguir entre três espécies distintas de íris (Setosa, Versicolor e Virginica) com base em características morfológicas das flores. A metodologia empregou uma divisão dos dados em 80% para treinamento e 20% para teste, utilizando uma arquitetura de rede com três camadas: entrada com 4 neurônios, camada oculta com 100 neurônios e saída com 3 neurônios. Os resultados demonstraram uma acurácia de 93,33%, evidenciando a eficácia das redes neurais artificiais na resolução de problemas de classificação multiclasse. A análise dos parâmetros da rede revelou o uso da função de ativação ReLU e convergência após 200 iterações, confirmando a robustez do modelo proposto.

Palavras-chave: Redes neurais artificiais. Multilayer Perceptron. Classificação. Conjunto de dados Iris. Aprendizado de máquina.

INTRODUÇÃO

As redes neurais artificiais (RNA) representam uma das principais ferramentas computacionais inspiradas no funcionamento do sistema nervoso biológico, tendo revolucionado o campo do aprendizado de máquina e da inteligência artificial nas últimas décadas. O Multilayer Perceptron (MLP), proposto inicialmente por Rosenblatt (1958) e posteriormente desenvolvido por Rumelhart, Hinton e Williams (1986), constitui uma das arquiteturas mais fundamentais e amplamente utilizadas em problemas de classificação e reconhecimento de padrões.

O conjunto de dados Iris, introduzido por Fisher (1936), tornou-se um benchmark clássico na área de aprendizado de máquina, sendo frequentemente utilizado para avaliar algoritmos de classificação devido às suas características bem definidas e à natureza multiclasse do problema. Este conjunto contém 150 amostras de flores de íris, distribuídas igualmente entre três espécies: Iris setosa, Iris versicolor e Iris virginica, cada uma caracterizada por quatro atributos morfológicos: comprimento e largura das sépalas, e comprimento e largura das pétalas (Anderson, 1935).

A aplicação de redes neurais artificiais em problemas de classificação tem demonstrado resultados promissores em diversas áreas do conhecimento. Haykin (2001) destaca que as MLPs são aproximadores universais de funções, capazes de mapear relações não-lineares complexas entre variáveis de entrada e saída. Esta capacidade torna-se particularmente relevante em problemas de classificação biológica, onde as relações entre características morfológicas e taxonomia podem apresentar complexidades não capturadas por métodos lineares tradicionais.

Diversos estudos têm explorado a eficácia das redes neurais na classificação de espécies. Glorot e Bengio (2010) demonstraram a importância da inicialização adequada dos pesos em redes profundas, enquanto LeCun, Bengio e Hinton (2015) consolidaram os fundamentos teóricos do aprendizado profundo.

O presente trabalho objetiva investigar a aplicabilidade de redes neurais artificiais do tipo MLP na classificação automática de espécies de íris, avaliando a performance do modelo

em termos de acurácia e analisando os parâmetros arquiteturais que contribuem para o desempenho obtido. A pesquisa justifica-se pela necessidade de compreender melhor os mecanismos de funcionamento das RNA em problemas de classificação biológica, contribuindo para o avanço do conhecimento na área de bioinformática e taxonomia computacional.

MÉTODOS

Os dados foram carregados por meio da função load_iris() da biblioteca sklearn.datasets e, em seguida, divididos em conjuntos de treinamento e teste. A divisão adotou a estratégia convencional de utilizar 80% das amostras (120 instâncias) para treinamento e 20% (30 instâncias) para teste, implementada através da função train_test_split com os parâmetros test_size=0.2 e random_state=10, de forma a garantir a reprodutibilidade dos experimentos. Essa abordagem permitiu que o modelo fosse treinado com uma quantidade substancial de dados, preservando ao mesmo tempo um conjunto independente para avaliar sua capacidade de generalização.

A rede neural artificial foi implementada utilizando a classe MLPClassifier da biblioteca scikit-learn. A arquitetura configurada corresponde a um Multilayer Perceptron (MLP) de três camadas: a camada de entrada com quatro neurônios associados às variáveis do conjunto de dados, uma camada oculta composta por 100 neurônios e a camada de saída com três neurônios representando as três classes de espécies. A função de ativação empregada foi a Rectified Linear Unit (ReLU), bastante utilizada em redes neurais por suas propriedades de não saturação e eficiência computacional (Nair e Hinton, 2010).

O treinamento do modelo foi realizado por meio do método fit(), estabelecendo-se um número máximo de 200 iterações. A convergência foi monitorada a partir da tolerância padrão de $1e^{-4}$ aplicada à melhoria da função de perda. Após a etapa de treinamento, foram realizadas predições sobre o conjunto de teste por meio do método predict(). O desempenho do classificador foi avaliado com base na métrica de acurácia, calculada com a função accuracy_score da biblioteca sklearn.metrics. Essa métrica, que expressa a proporção de classificações corretas em relação ao total de predições, mostrou-se apropriada para o

problema, visto que o conjunto Iris apresenta classes balanceadas (Sokolova e Lapalme, 2009).

Por fim, foi realizada uma análise dos parâmetros aprendidos pela rede, contemplando o número de camadas (n_layers_), a distribuição de neurônios por camada, a função de ativação empregada, o número de iterações efetivamente realizadas (n_iter_), os pesos das conexões (coefs_) e os termos de bias (intercepts_) da primeira camada. Essa análise possibilitou compreender a dinâmica do aprendizado do modelo e os aspectos estruturais da rede neural implementada.

O código-fonte está disponível para download através do link: https://github.com/vitor-souza-ime/mlp.

RESULTADOS E DISCUSSÃO

A rede neural artificial implementada demonstrou performance na classificação das espécies de íris, alcançando uma acurácia de 93,33% no conjunto de teste. Este resultado evidencia a robustez do modelo MLP, que utilizou 80% dos dados (120 amostras) para treinamento e foi validado em 20% das amostras (30 amostras), seguindo as boas práticas da área de aprendizado de máquina. A análise comparativa das predições com os valores reais revelou apenas duas classificações incorretas em 30 amostras de teste. Especificamente, observaram-se erros nas posições 10 e 11, onde a rede predisse a classe Iris virginica quando o valor real era Iris versicolor, sugerindo maior similaridade entre estas duas espécies em comparação com a Iris setosa.

A análise dos parâmetros arquiteturais confirmou a estrutura de três camadas da rede neural: camada de entrada com quatro neurônios correspondentes às características morfológicas, camada oculta com 100 neurônios e camada de saída com três neurônios representando as espécies. Esta configuração mostrou-se adequada para o problema, oferecendo capacidade suficiente para capturar as relações não-lineares entre as variáveis de entrada e as respectivas classificações. A função de ativação utilizada, ReLU, revelou-se apropriada para este tipo de tarefa, uma vez que apresenta vantagens significativas em relação às funções sigmoidais, como maior velocidade de convergência e mitigação do problema de desaparecimento do gradiente (Glorot, Bordes e Bengio, 2011). O número

máximo de iterações foi alcançado (200), indicando que o algoritmo de otimização L-BFGS explorou completamente o espaço de parâmetros dentro dos critérios de convergência definidos (Liu e Nocedal, 1989).

A análise dos pesos da primeira camada e dos termos de bias forneceu insights relevantes sobre o aprendizado da rede. A matriz de pesos, de dimensão 4×100, representa as conexões entre os atributos de entrada e os neurônios da camada oculta, enquanto o vetor de bias ajusta o limiar de ativação de cada neurônio. A distribuição dos valores aprendidos reflete a importância relativa de cada característica morfológica na diferenciação das espécies. Conexões com valores absolutos mais elevados sugerem maior relevância desses atributos para a classificação, o que está em consonância com estudos clássicos que destacam medidas específicas, como comprimento e largura das pétalas, como discriminativas entre espécies (Fisher, 1936).

O resultado obtido posiciona o modelo MLP entre os classificadores eficazes para o conjunto Iris, apresentando desempenho comparável a métodos tradicionais como Support Vector Machines, k-Nearest Neighbors e árvores de decisão. Esse desempenho está alinhado a estudos anteriores que demonstraram a efetividade das redes neurais em problemas de classificação não-linear (Bishop, 2006). A capacidade da rede de alcançar alta acurácia com uma divisão convencional dos dados reforça sua eficiência em extrair padrões relevantes durante o treinamento e generalizar para novos dados, característica fundamental em aplicações práticas de classificação biológica.

Os resultados obtidos demonstram o potencial das redes neurais artificiais na automatização de processos de classificação taxonômica, com implicações diretas para áreas como botânica sistemática, ecologia e conservação. A alta acurácia alcançada sugere que modelos MLP podem ser utilizados em sistemas de identificação automática de espécies, contribuindo para estudos de biodiversidade e monitoramento ambiental.

CONCLUSÕES

Este estudo demonstrou a eficácia das redes neurais artificiais do tipo Multilayer Perceptron na classificação automática de espécies do conjunto de dados Iris, alcançando uma acurácia de 93,33% utilizando uma divisão convencional de 80% dos dados para treinamento e 20%

para teste. A arquitetura de três camadas com 100 neurônios na camada oculta e função de ativação ReLU mostrou-se adequada para capturar as relações não-lineares entre as características morfológicas e a taxonomia das espécies.

Os resultados confirmam o potencial das RNA como ferramenta robusta para problemas de classificação biológica, oferecendo vantagens em termos de capacidade de generalização e adaptabilidade a padrões complexos. A análise dos parâmetros aprendidos forneceu insights sobre o funcionamento interno do modelo, contribuindo para o entendimento dos mecanismos de decisão da rede neural.

As implicações práticas deste trabalho estendem-se além do conjunto Iris, sugerindo aplicabilidade em sistemas de identificação automática de espécies e monitoramento de biodiversidade. Trabalhos futuros podem explorar a aplicação de arquiteturas mais complexas, como redes neurais convolucionais, em conjuntos de dados biológicos mais desafiadores, além de investigar métodos de interpretabilidade que possam elucidar os critérios utilizados pelos modelos na classificação de espécies.

A integração de técnicas de aprendizado profundo com conhecimento biológico tradicional representa uma fronteira promissora para o avanço da taxonomia computacional e da bioinformática, potencializando descobertas científicas e aplicações práticas em conservação e gestão da biodiversidade.

REFERÊNCIAS

ANDERSON, E. The irises of the Gaspé Peninsula. Bulletin of the American Iris Society, v. 59, p. 2-5, 1935.

BISHOP, C. M. Pattern Recognition and Machine Learning. New York: Springer, 2006. 738 p.

FISHER, R. A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, v. 7, n. 2, p. 179-188, 1936.

GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. In: PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 13., 2010, Sardinia. Proceedings... Sardinia: JMLR, 2010. p. 249-256.

GLOROT, X.; BORDES, A.; BENGIO, Y. Deep sparse rectifier neural networks. In: PROCEEDINGS OF THE FOURTEENTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 14., 2011, Fort Lauderdale. Proceedings... Fort Lauderdale: JMLR, 2011. p. 315-323.

HAYKIN, S. Redes Neurais: Princípios e Prática. 2. ed. Porto Alegre: Bookman, 2001. 900 p.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. Nature, v. 521, n. 7553, p. 436-444, 2015.

LIU, D. C.; NOCEDAL, J. On the limited memory BFGS method for large scale optimization. Mathematical Programming, v. 45, n. 1-3, p. 503-528, 1989.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 27., 2010, Haifa. Proceedings... Haifa: ICML, 2010. p. 807-814.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, v. 65, n. 6, p. 386-408, 1958.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. Nature, v. 323, n. 6088, p. 533-536, 1986.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. Information Processing & Management, v. 45, n. 4, p. 427-437, 2009.