Título em Português: BioPrediction: Democratizando a Aprendizagem de Máquina no Estudo de Interações Moleculares

Título em Inglês: bioprediction: democratizing machine learning in the study of molecular interactions

Autor: Bruno Rafael Florentino

Instituição: Universidade de São Paulo

Unidade: Instituto de Física de São Carlos

Orientador: André Carlos Ponce de Leon Ferreira de Carvalho

Área de Pesquisa / SubÁrea: Metodologia e Técnicas da Computação

Agência Financiadora: Outros

BioPrediction: Democratizando Aprendizado de Máquina em Estudos de Interações lncRNA-Proteína

Bruno Rafael Florentino

Robson Parmezan Bonidia, Natan Henrique Sanches

André Carlos Ponce de Leon Ferreira de Carvalho

Universidade de São Paulo

brunorf1204@usp.br

Objetivos

Neste trabalho, é proposto um framework fim-a-fim baseado em Aprendizado de Máquina (AM) automatizado, chamado BioPrediction (uma extensão da ferramenta BioAutoML [1]), capaz de identificar interações implícitas entre sequências, por exemplo, pares de RNA longo não-codificante e proteínas, sem a necessidade de conhecimento especializado em AM de ponta a ponta. Além disso, o framework proposto teve como objetivos secundários, além de alcançar desempenho competitivo em relação à literatura, ser simplificado para tornar-se acessível a pesquisadores no campo das ciências biológicas que não possuem conhecimento profundo em AM. O intuito foi que o modelo de predição fosse construído sem intervenção humana após o usuário inserir os dados. Adicionalmente, pretendeu-se fornecer uma saída útil aos usuários por meio de diversas formas de representação das métricas do modelo e da topologia das interações.

Métodos e Procedimentos

Para construir um modelo abrangente, foi utilizado um pipeline de ponta a ponta que abrange todas as etapas do fluxo de trabalho de AM. Isso inclui desde a extração das características de cada sequência e a seleção das mais eficazes, até a escolha do melhor modelo de classificação baseado em árvore. Além disso, o pipeline lida com dados desbalanceados, otimiza os parâmetros dos algoritmos de AM e produz como saída o melhor modelo, acompanhado de um relatório de interpretabilidade. Todas as características extraídas são baseadas nas bases que as compõem, gerando um vetor que caracteriza as sequências. Em seguida, o par de características do RNA e da proteína é concatenado, criando um vetor único para cada dupla, dando início à tarefa de predição. O framework foi validado em cinco conjuntos de dados disponíveis em [2], seguindo a mesma configuração. Foi utilizada a acurácia balanceada para avaliar o desempenho do modelo, uma vez que as classes, interagente e não interagente, estavam em proporções desbalanceadas. Por fim, o desempenho do BioPrediction foi comparado com diversos modelos mencionados no artigo [2]. Isso incluiu uma comparação com modelos desenvolvidos por especialistas, usando métricas como AUC, precisão, recall, F1 e AUPR para avaliar sua competitividade.

Resultados

Em todos os conjuntos de dados, o desempenho médio do framework automatizado para a acurácia balanceada foi superior a 75%, representando uma primeira evidência da capacidade de distinção entre as classes. Em seguida, temos a comparação de desempenho entre os modelos. O BioPrediction foi o modelo com melhor desempenho em acurácia e AUC

em todos os conjuntos de dados, evidenciando as qualidades do framework nessas medidas. A única medida em que o BioPrediction não superou a média foi em AUPR, sendo uma métrica usada para avaliar o desempenho de modelos treinados com dados desbalanceados, revelando uma limitação nesse aspecto.

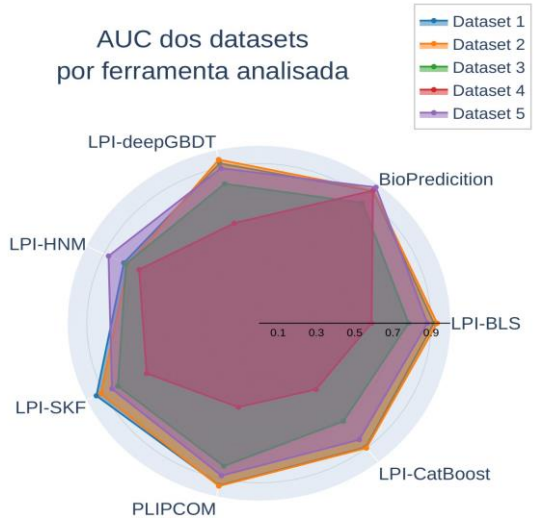

Figura 1: Gráfico de comparação da métrica AUC entre as diversas ferramentas.

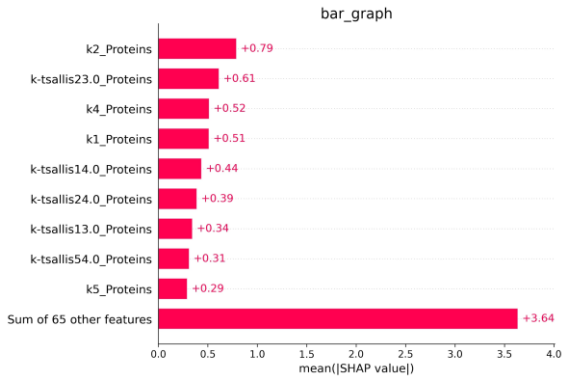

Figura 2: Exemplo de gráfico de interpretabilidade que mostra as características mais relevantes.

Complementarmente, foi implementado o módulo de interpretabilidade que analisa as entradas e saídas do modelo em busca de padrões. Isso foi realizado a partir da teoria dos SHAP Values. Com base nisso, foram criados vários gráficos que mostram, por exemplo, quais são as características mais importantes e o quanto o valor que cada característica assume em uma dada amostra influencia na classificação final. Finalmente, o presente trabalho foi aceito para publicação no 20th Encontro Nacional de Inteligência Artificial e Computacional (2023).

Conclusões

Com base nos resultados, constatou-se que o framework demonstrou ter um desempenho competitivo em relação a alguns modelos presentes na literatura. A criação do framework BioPrediction representa um possível avanço no campo da predição de interações, ao democratizar o acesso ao tratamento de dados biológicos e treinamento automatizado de modelos de AM, de forma que esses possuam desempenho competitivo quando comparados aos criados por especialistas.

Agradecimentos

Todo agradecimento ao CNPQ pelo financiamento (número do processo: 117295/2022-1), ICMC pelo apoio financeiro e pela orientação e ao Canada's International Development Research Centre (IDRC) (Grant No. 109981).

Referências

Bonidia, et al. BioAutoML: automated feature engineering and metalearning to predict noncoding RNAs in bacteria. Briefings in Bioinformatics, 23(4), 2022.

Zhou et al. (2021). LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncrna–protein interaction identification. BMC Bioinformatics, 22:479

BioPrediction: democratizing machine learning in lncRNA-protein interaction studies

Bruno Rafael Florentino

Robson Parmezan Bonidia and Natan Henrique Sanches

André Carlos Ponce de Leon Ferreira de Carvalho

Universidade de São Paulo

brunorf1204@usp.br

Objectives

In this work, we propose an end-to-end framework based on automated Machine Learning (ML), named BioPrediction (an extension of the BioAutoML tool [1]), capable of identifying implicit interactions between sequences, such as pairs of long non-coding RNA and proteins, without the need for specialized knowledge in end-to-end ML. Furthermore, the proposed framework had secondary objectives, in addition to achieving competitive performance compared to the literature, to be simplified to become accessible to researchers in the field of biological sciences who do not possess deep knowledge in ML. The aim was for the prediction model to be constructed without human intervention after the user inputs the data. Additionally, we aimed to provide users with a useful output through various forms of representation of the model's metrics and interaction topology.

Methods and Procedures

To construct a comprehensive model, we employed an end-to-end pipeline encompassing all stages of the workflow. This pipeline includes the extraction of features from each sequence and the selection of the most effective ones, as well as the choice of the best tree-based classification model. Furthermore, the pipeline handles imbalanced data, optimizes model parameters, and yields the best model as output, accompanied by an interpretability report. All extracted features are based on the constituent bases, generating a vector that characterizes the sequences. Subsequently, the pair of RNA and protein features are concatenated, creating a unique vector for each pair, initiating the prediction task. The framework underwent validation on five datasets available in [2], following the same configuration. Balanced accuracy was used to evaluate the model's performance, given the imbalanced proportions of interacting and non-interacting classes. Lastly, the performance of BioPrediction was compared to various models mentioned in the article [2]. This comparison involved assessing an AutoML model against models developed by experts, utilizing metrics such as AUC, precision, recall, F1, and AUPR to evaluate its competitiveness.

Results

In all datasets, the average performance of the automated framework for balanced accuracy exceeded 75%, representing an initial indication of its ability to distinguish between classes. Next, we have the performance comparison among the models. BioPrediction outperformed other models in terms of accuracy and AUC in all datasets, highlighting the framework's strengths in these metrics. One metric where BioPrediction did not consistently outperform was AUPR, which is used to evaluate the performance of models trained with imbalanced

data, revealing its limitations in this aspect.

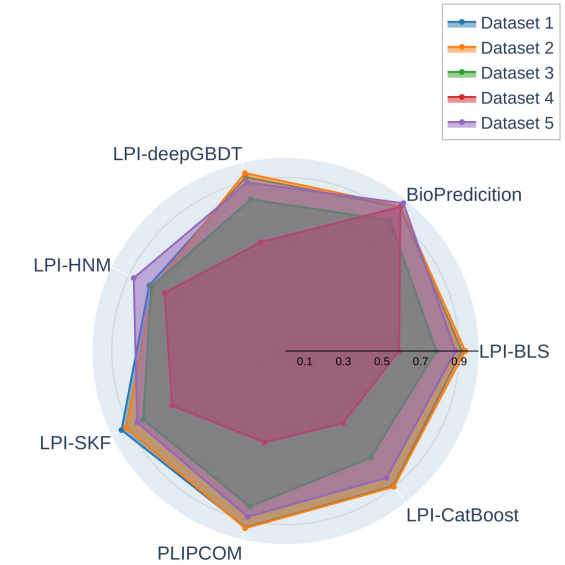

Figure 1: A comparison chart of the AUC metric among the various tools.

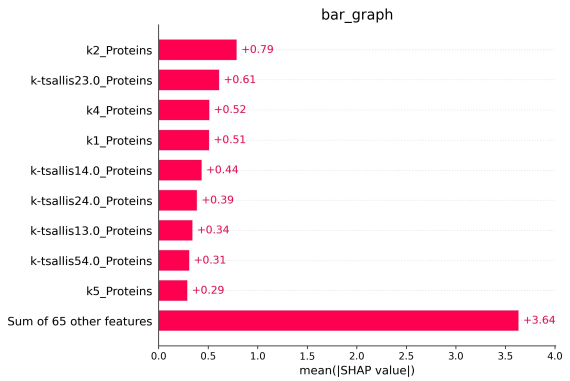

Figure 2: Example of an interpretability chart displaying the most relevant features.

Additionally, an interpretability module was implemented to analyze the model's inputs and outputs for patterns. This was accomplished using the theory of SHAP Values. Based on this, several graphs were created to show, for example, the most important features and how much the value each feature takes on in a given sample influences the final classification. Finally, this work has been accepted for publication at the 20th National Meeting on Artificial Intelligence and Computational Intelligence (2023).

Conclusions

Based on the results, it was found that the framework demonstrated competitive performance compared to some models in the literature. The creation of the BioPrediction framework represents a potential advancement in the field of interaction prediction by democratizing access to biological data processing and automated ML model training. This ensures that these models achieve competitive performance when compared to those created by experts.

Acknowledgments

I would like to extend my gratitude to CNPQ for their funding support (grant number: 117295/2022-1), ICMC for their financial assistance and guidance, and Canada's International Development Research Centre (IDRC) (Grant No. 109981). Your support has been instrumental in the success of this research.

References

Bonidia, et al. BioAutoML: automated feature engineering and metalearning to predict noncoding RNAs in bacteria. Briefings in Bioinformatics, 23(4), 2022.

Zhou et al. (2021). LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncrna–protein interaction identification. BMC Bioinformatics, 22:479.