Trabalho Prático de Tópicos Avançados
Predição de Localização Subcelular de Proteínas

Guilherme Sousa Panza[1],**Gustavo Sampaio Lima**[1], **Thaís Ribeiro Lauriano**[1]

[1]Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC - USP)
São Carlos - SP

{guilherme.panza, gustavo.sampaio, thaislauriano}@usp.br

Resumo. *Este trabalho desenvolveu modelos de classificação multirrótulo baseados em redes neurais para prever a localização subcelular de proteínas de plantas e vírus. Técnicas de regularização e ajuste de pesos foram utilizadas para lidar com o desbalanceamento dos dados, melhorando o desempenho em classes minoritárias. Os resultados demonstraram a eficácia das redes neurais para extrair padrões complexos e abordar problemas de classificação em cenários desbalanceados.*

1. Introdução

1.1. Motivação e Contextualização

A biologia molecular tem como um de seus pilares o estudo das proteínas, moléculas essenciais para a regulação e execução de praticamente todos os processos biológicos nos organismos vivos. Entender a localização subcelular das proteínas — os compartimentos celulares onde essas moléculas desempenham suas funções — é crucial para desvendar mecanismos biológicos fundamentais, como os relacionados ao metabolismo, à sinalização celular e à patogênese de infecções virais. Além disso, esse entendimento é frequentemente aplicado ao desenvolvimento de terapias direcionadas, melhorias agrícolas e estudos evolutivos.

Embora técnicas experimentais como fluorescência de proteínas e fracionamento celular sejam amplamente empregadas para determinar a localização subcelular, elas apresentam limitações como alto custo, tempo elevado e necessidade de equipamentos especializados. Nesse contexto, métodos computacionais têm sido desenvolvidos como alternativas eficientes, utilizando dados bioquímicos, genéticos e funcionais para realizar predições. Esses métodos não apenas complementam abordagens laboratoriais, mas também expandem a capacidade de análise em larga escala.

Entre as abordagens computacionais, algoritmos bioinspirados têm se destacado pela sua eficiência na solução de problemas complexos. Modelos baseados em Redes Neurais Artificiais, Algoritmos Genéticos e Otimização por Enxame de Partículas são especialmente adequados para tarefas de classificação, graças à sua habilidade em explorar padrões não lineares e lidar com grandes volumes de dados. No caso específico da predição de localização subcelular, tais algoritmos permitem capturar as características das sequências de proteínas e suas relações com os compartimentos celulares em cenários multirrótulo, onde uma proteína pode estar associada a múltiplas localizações simultaneamente.

Com dois conjuntos de dados fornecidos, um de proteínas de vírus e outro de plantas, categorizados em múltiplos compartimentos celulares, este trabalho visa desenvolver um modelo de classificação multirrótulo utilizando um algoritmo bioinspirado escolhido.

1.2. Objetivos

Os objetivos deste trabalho são:

- Desenvolver um modelo de classificação multirrótulo para a predição da localização subcelular de proteínas, utilizando um algoritmo bioinspirado.
- Escolher um algoritmo bioinspirado adequado, considerando suas características e sua capacidade de resolver o problema proposto.

1.3. Organização

O presente relatório está estruturado em cinco seções principais. A Seção II, intitulada Metodologia, apresenta o algoritmo bioinspirado escolhido, os passos de pré-processamento realizados nos dados e a abordagem utilizada para desenvolver o modelo. A Seção III, Experimentos, descreve detalhadamente os conjuntos de dados, os parâmetros do modelo implementado e as métricas utilizadas para avaliação do desempenho. Na Seção IV, Discussões, são analisados os resultados obtidos, destacando as limitações do modelo e propondo possíveis melhorias. Por fim, a Seção V, Conclusão, sintetiza os principais achados deste trabalho e sugere caminhos para estudos futuros.

2. Metodologia

Este trabalho foi estruturado em três etapas principais: análise e pré-processamento do conjunto de dados, construção dos modelos e avaliação de desempenho. A seguir, detalhamos cada uma dessas etapas:

2.1. Análise e pré-processamento

Dois conjuntos de dados multirrótulo foram utilizados, abrangendo proteínas de plantas e de vírus. Cada instância representa uma sequência de proteína e cada classe corresponde a uma localização subcelular; as proteínas podem estar presentes simultaneamente em dois ou mais compartimentos celulares.

O conjunto de dados do vírus possui seis localizações subcelulares: Proteínas do Capsídeo Viral, Proteínas da Membrana Celular do Hospedeiro, Proteínas do Retículo Endoplasmático do Hospedeiro, Proteínas do Citoplasma do Hospedeiro, Proteínas do Núcleo do Hospedeiro e Proteínas Secretadas. O da planta contém 12 localizações: Proteínas da Membrana Celular, Proteínas da Parede Celular, Proteínas do Cloroplasto, Proteínas do Citoplasma, Proteínas do Retículo Endoplasmático, Proteínas Extracelulares, Proteínas do Complexo de Golgi, Proteínas das Mitocôndrias, Proteínas do Núcleo, Proteínas dos Peroxissomos, Proteínas dos Plastídios e Proteínas do Vacúolo.

Em ambos, a primeira coluna indica o código identificador das proteínas, as últimas colunas indicam a presença ou ausência da proteína em cada uma das localizações subcelulares e as demais representam os códigos de Gene Ontology, com valores que indicam a frequência do código para cada proteína.

Foram aplicadas técnicas de análise exploratória para uma compreensão mais aprofundada dos dados. Entre as análises realizadas, destaca-se o cálculo de estatísticas descritivas, incluindo a identificação dos valores médios de desvio padrão, valores médios dos valores máximos e valores máximos absolutos dos atributos presentes nos conjuntos de dados. Os resultados dessas análises estão sintetizados na Tabela 1.

Tabela 1. Estatísticas Descritivas dos Atributos nos Conjuntos de Dados

Métrica	Conjunto de Plantas	Conjunto de Vírus
Média dos desvios padrão	0.09	0.21
Média dos máximos	1.84	2.05
Máximo dos máximos	122	70

A frequência de rótulos também foi analisada com o intuito de identificar a distribuição no conjunto de dados e a presença de desbalanceamento. Os resultados obtidos a partir dos dados das proteínas das plantas podem ser encontrados na Tabela 2 e aqueles relativos aos dados das proteínas virais estão na Tabela 3.

Tabela 2. Distribuição de Rótulos no Conjunto de Dados de Plantas

Rótulo	Frequência
Proteínas da Membrana Celular	56
Proteínas da Parede Celular	32
Proteínas do Cloroplasto	277
Proteínas do Citoplasma	181
Proteínas do Retículo Endoplasmático	42
Proteínas Extracelulares	22
Proteínas do Complexo de Golgi	21
Proteínas das Mitocôndrias	150
Proteínas do Núcleo	150
Proteínas dos Perixossomos	21
Proteínas dos Plastídios	39
Proteínas do Vacúolo	51

Tabela 3. Distribuição de Rótulos no Conjunto de Dados de Vírus

Rótulo	Frequência
Proteínas do Capsídeo Viral	8
Proteínas da Membrana Celular do Hospedeiro	32
Proteínas do Retículo Endoplasmático do Hospedeiro	20
Proteínas do Citoplasma do Hospedeiro	87
Proteínas do Núcleo do Hospedeiro	84
Proteínas Secretadas	20

O pré-processamento das bases de dados foi realizado em duas etapas principais. Primeiramente, foram eliminadas duplicatas e valores nulos, garantindo a consistência e a qualidade dos dados utilizados nas análises e no treinamento do modelo. Em seguida, os atributos numéricos foram normalizados utilizando o método MinMaxScaler, de forma a restringir os valores ao intervalo [0, 1]. Essa normalização preservou a escala relativa dos dados e facilitou o treinamento do modelo, especialmente considerando a presença de outliers, evidenciados pela discrepância entre a média dos máximos e os valores máximos absolutos observados nos conjuntos de dados.

2.2. Construção dos Modelos

Dentre os diversos algoritmos bioinspirados, como algoritmos genéticos e otimização por enxame de partículas, as redes neurais artificiais foram selecionadas para este trabalho devido à sua capacidade de extrair padrões complexos e não lineares dos dados, mesmo na presença de ruído e dados incompletos. Essa característica é fundamental para lidar com a variabilidade intrínseca dos dados biológicos.

As redes neurais apresentam alta capacidade de generalização e adaptabilidade, permitindo que o modelo seja atualizado e ajustado facilmente com novas informações e descobertas. Além disso, a escalabilidade e o alto desempenho computacional dessas redes, especialmente quando combinadas com hardware especializado, permitem processar grandes volumes de dados e fornecer resultados em tempo hábil.

Inicialmente, os dados de plantas e vírus foram divididos em conjuntos de treino e teste, de maneira separada para cada dataset, com 30% dos dados reservados para o último conjunto. Em seguida, 15% do conjunto de treino foi alocado para o conjunto de validação.

Além disso, o presente trabalho buscou explorar duas abordagens fazendo uso de redes neurais artifi-

ciais com arquiteturas diferentes. A primeira arquitetura foi desenvolvida com camadas densas, intercaladas com técnicas de regularização, para garantir uma boa generalização e evitar o overfitting. A estrutura do modelo foi aplicada de forma consistente aos dois conjuntos de dados, seguindo estes passos:

- **Camada de entrada:** A primeira camada densa contém 128 neurônios e utiliza a função de ativação *ReLU*, que é eficaz para introduzir não linearidade e facilitar a aprendizagem de padrões complexos.
- **Etapa de Dropout:** 30% dos neurônios foram desativados aleatoriamente para reduzir o overfitting e promover a generalização do modelo.
- **Camadas intermediária:** Uma camada adicional foi inserida, com 64 neurônios e a função de ativação *ReLU*, permitindo uma maior capacidade de aprendizado das características dos dados.
- **2° Etapa de Dropout:** Novamente, 30% dos neurônios foram desativados aleatoriamente, reforçando a regularização do modelo.
- **Camada de saída:** A camada final possui um número de neurônios igual ao número de classes no problema, utilizando a função de ativação *sigmoid*, ideal para problemas de classificação multirrótulo, onde cada classe pode ser tratada independentemente com uma probabilidade associada.

O modelo foi então compilado com o otimizador Adam junto a função de perda de entropia cruzada binária (*binary crossentropy*), adequada para múltiplas classes binárias. A métrica de avaliação escolhida foi a acurácia. Essa arquitetura, mesmo limitada, foi projetada para equilibrar a capacidade de aprendizagem com técnicas de regularização, buscando uma boa performance tanto no tempo de treino quanto na velocidade da inferência do modelo.

A segunda arquitetura foi projetada para lidar com desafios mais complexos, utilizando técnicas mais robustas, especialmente em face da natureza desbalanceada dos rótulos nos conjuntos de dados. Para mitigar o impacto do desbalanceamento, foram calculados pesos para cada classe, ajustando sua contribuição durante o treinamento e penalizando as classes minoritárias. Esses pesos foram calculados separadamente para os conjuntos de plantas e vírus.

- **Camada de entrada:** A primeira camada densa contém 256 neurônios e utiliza a função de ativação *ReLU*. A regularização L2 foi aplicada para ajudar a reduzir o overfitting, promovendo uma aprendizagem mais robusta.
- **Camadas intermediárias:** Foram adicionadas camadas densas adicionais com 128 e 64 neurônios, utilizando também a função de ativação *ReLU*. Para melhorar a generalização e evitar overfitting, foi aplicada regularização L2 em conjunto com *Batch Normalization* e *Dropout* (com taxas de 0.4 e 0.3, respectivamente).
- **Camada de saída:** A camada final possui um número de neurônios igual ao número de classes, utilizando a função de ativação *sigmoid*, que é ideal para problemas de classificação multirrótulo, permitindo que cada classe seja tratada independentemente com uma probabilidade associada.

O segundo modelo foi compilado de maneira similar ao primeiro, com a adição da definição da taxa de aprendizado em 0.001. Além disso, para prevenir o overfitting e otimizar o tempo de treinamento, foi aplicada a técnica de parada antecipada, interrompendo o treinamento caso o desempenho na validação não melhorasse após 10 épocas consecutivas. Por fim, foi implementada uma técnica que ajusta automaticamente a taxa de aprendizado caso o modelo não apresente melhorias, com o intuito de evitar que o modelo fique preso em um mínimo local. O treinamento foi realizado por até 100 épocas, com um tamanho de batch de 32. Essa abordagem visou otimizar a performance do modelo e garantir sua generalização para novos dados.

2.3. Avaliação de Desempenho

A avaliação de desempenho dos modelos foi realizada utilizando várias métricas, com ênfase na *Hamming Loss* para avaliar o erro de classificação binária multirrótulo, além de outras métricas para obter uma visão abrangente do desempenho do modelo.

A *Hamming Loss* é uma métrica importante para problemas de classificação multirrótulo, que calcula a proporção de rótulos incorretos, ou seja, quantas vezes as previsões do modelo divergem dos rótulos reais, considerando as instâncias e as classes individualmente. Quanto menor o valor de *Hamming Loss*, melhor o modelo.

Para o cálculo do *Hamming Loss*, as previsões dos modelos são comparadas com os valores reais dos rótulos das classes, onde um valor de 1 representa uma previsão errada e um valor de 0 indica que a previsão está correta. Essa métrica é útil quando o modelo deve prever múltiplos rótulos para uma mesma instância e é necessário avaliar a precisão de cada classe individualmente.

O Relatório de Classificação fornece uma visão detalhada do desempenho dos modelos em termos de precisão, recall, F1-score e suporte para cada classe. Estas métricas são particularmente úteis em problemas de classificação desbalanceada, pois fornecem insights sobre como o modelo está se comportando com relação a cada classe individual.

3. Experimentos

Para avaliar a classificação multirrótulo, foram realizados experimentos nos dois conjuntos de dados.

3.1. Conjunto de Dados de Proteínas de Planta

Neste conjunto, dois experimentos foram feitos para classificação em 12 rótulos diferentes: Proteínas da Membrana Celular, Proteínas da Parede Celular, Proteínas do Cloroplasto, Proteínas do Citoplasma, Proteínas do Retículo Endoplasmático, Proteínas Extracelulares, Proteínas do Complexo de Golgi, Proteínas das Mitocôndrias, Proteínas do Núcleo, Proteínas dos Peroxissomos, Proteínas dos Plastídios e Proteínas do Vacúolo.

3.1.1. 1° Experimento: Rede Neural Base

O modelo foi treinado por 50 épocas com o otimizador Adam, usando a função de perda de entropia cruzada binária e sem considerar pesos específicos para as classes. Os resultados obtidos estão na Tabela 4.

Tabela 4. Relatório de classificação (1º experimento)

Rótulo	Precisão	Revocação	F1-Score	Suporte
Proteínas da Membrana Celular	0.73	0.53	0.62	15
Proteínas da Parede Celular	0.83	0.67	0.74	15
Proteínas do Cloroplasto	0.90	0.94	0.92	84
Proteínas do Citoplasma	0.89	0.82	0.85	50
Proteínas do Retículo Endoplasmático	0.57	0.36	0.44	11
Proteínas Extracelulares	1.00	0.20	0.33	10
Proteínas do Complexo de Golgi	1.00	0.33	0.50	6
Proteínas das Mitocôndrias	0.86	0.84	0.85	37
Proteínas do Núcleo	0.84	0.93	0.88	45
Proteínas dos Peroxissomos	1.00	0.14	0.25	7
Proteínas dos Plastídios	1.00	0.53	0.69	17
Proteínas do Vacúolo	0.50	0.29	0.37	17
Média Micro	0.85	0.75	0.80	314
Média Macro	0.84	0.55	0.62	314
Média Ponderada	0.85	0.75	0.77	314
Média por Amostra	0.77	0.75	0.75	314

3.1.2. 2° Experimento: Rede Neural com Regularização e Pesos de Classes

No segundo modelo, os pesos das classes desbalanceadas foram ajustados e variaram de acordo com a proporção de amostras, como mostrado abaixo na Tabela 5.

Tabela 5. Pesos para cada classe (2º experimento)

Rótulo	Peso (Classe 0)	Peso (Classe 1)
Proteínas da Membrana Celular	0.5276	9.5667
Proteínas da Parede Celular	0.5143	17.9375
Proteínas do Cloroplasto	0.7034	1.7289
Proteínas do Citoplasma	0.6226	2.5398
Proteínas do Retículo Endoplasmático	0.5218	11.9583
Proteínas Extracelulares	0.5080	31.8889
Proteínas do Complexo de Golgi	0.5116	22.0769
Proteínas das Mitocôndrias	0.5979	3.0532
Proteínas do Núcleo	0.5942	3.1538
Proteínas dos Peroxissomos	0.5116	22.0769
Proteínas dos Plastídios	0.5181	14.3500
Proteínas do Vacúolo	0.5266	9.8966

Os resultados obtidos estão na Tabela 6.

Tabela 6. Relatório de classificação (2º experimento)

Rótulo	Precisão	Revocação	F1-Score	Suporte
Proteínas da Membrana Celular	0.85	0.73	0.79	15
Proteínas da Parede Celular	0.85	0.73	0.79	15
Proteínas do Cloroplasto	0.98	1.00	0.99	84
Proteínas do Citoplasma	0.94	0.94	0.94	50
Proteínas do Retículo Endoplasmático	0.86	0.55	0.67	11
Proteínas Extracelulares	1.00	0.40	0.57	10
Proteínas do Complexo de Golgi	1.00	0.33	0.50	6
Proteínas das Mitocôndrias	0.97	0.89	0.93	37
Proteínas do Núcleo	0.89	0.93	0.91	45
Proteínas dos Peroxissomos	1.00	0.57	0.73	7
Proteínas dos Plastídios	1.00	0.82	0.90	17
Proteínas do Vacúolo	0.75	0.53	0.62	17
Média Micro	0.93	0.85	0.89	314
Média Macro	0.92	0.70	0.78	314
Média Ponderada	0.93	0.85	0.88	314
Média por Amostra	0.87	0.86	0.86	314

3.2. Conjunto de Dados de Proteínas de Vírus

Neste conjunto, dois experimentos foram feitos para classificação em 6 rótulos diferentes: Proteínas do Capsídeo Viral, Proteínas da Membrana Celular do Hospedeiro, Proteínas do Retículo Endoplasmático do Hospedeiro, Proteínas do Citoplasma do Hospedeiro, Proteínas do Núcleo do Hospedeiro e Proteínas Secretadas.

3.2.1. 1° Experimento: Rede Neural Base

O modelo foi retreinado por 50 épocas com o otimizador Adam, usando a função de perda de entropia cruzada binária e sem considerar pesos específicos para as classes. Os resultados obtidos estão na Tabela 7.

Tabela 7. Relatório de classificação (1º experimento)

Rótulo	Precisão	Revocação	F1-Score	Suporte
Proteínas do Capsídeo Viral	0.00	0.00	0.00	3
Proteínas da Membrana Celular do Hospedeiro	1.00	0.67	0.80	12
Proteínas do Retículo Endoplasmático do Hospedeiro	1.00	0.50	0.67	4
Proteínas do Citoplasma do Hospedeiro	0.89	0.96	0.92	25
Proteínas do Núcleo do Hospedeiro	0.93	0.84	0.89	32
Proteínas Secretadas	0.00	0.00	0.00	2
Média Micro	0.92	0.78	0.85	78
Média Macro	0.64	0.50	0.55	78
Média Ponderada	0.87	0.78	0.82	78
Média por Amostra	0.81	0.78	0.78	78

Para esse modelo, o valor da *Hamming Loss* obtido foi de 0.0082 para o conjunto de treino e 0.0538 para o conjunto de teste.

3.2.2. 2° Experimento: Rede Neural com Regularização e Pesos de Classes

Novamente, os pesos das classes desbalanceadas foram ajustados e variaram de acordo com a proporção de amostras, como mostrado abaixo na Tabela 8.

Tabela 8. Pesos para cada classe (2º experimento)

Rótulo	Peso (Classe 0)	Peso (Classe 1)
Proteínas do Capsídeo Viral	0.5169	15.2500
Proteínas da Membrana Celular do Hospedeiro	0.5865	3.3888
Proteínas do Retículo Endoplasmático do Hospedeiro	0.5596	4.6923
Proteínas do Citoplasma do Hospedeiro	0.8591	1.1960
Proteínas do Núcleo do Hospedeiro	0.7625	1.4523
Proteínas Secretadas	0.5809	3.5882

Os resultados obtidos estão na Tabela 9 a seguir.

Tabela 9. Relatório de classificação (2º experimento)

Rótulo	Precisão	Revocação	F1-Score	Suporte
Proteínas do Capsídeo Viral	1.00	1.00	1.00	3
Proteínas da Membrana Celular do Hospedeiro	1.00	0.25	0.40	12
Proteínas do Retículo Endoplasmático do Hospedeiro	1.00	0.25	0.40	4
Proteínas do Citoplasma do Hospedeiro	0.96	1.00	0.98	25
Proteínas do Núcleo do Hospedeiro	0.97	0.88	0.92	32
Proteínas Secretadas	0.00	0.00	0.00	2
Média Micro	0.97	0.77	0.86	78
Média Macro	0.82	0.56	0.62	78
Média Ponderada	0.95	0.77	0.81	78
Média por Amostra	0.81	0.77	0.78	78

Para esse modelo, o valor da *Hamming Loss* obtido foi de 0.0055 para o conjunto de treino e 0.0591 para o conjunto de teste.

4. Discussões

Os experimentos realizados revelaram importantes insights sobre o desempenho das redes neurais em cenários multirrótulo. No caso dos dados de plantas, a introdução de regularização e ajuste de pesos melhorou significativamente métricas como precisão e F1-score para classes menos representadas, como ”Proteínas Extracelulares”e ”Proteínas do Complexo de Golgi”. No entanto, classes com muito poucas instâncias ainda apresentaram desafios, como observado nos dados de vírus para ”Proteínas Secretadas”.

Embora as redes neurais tenham demonstrado capacidade para lidar com variabilidade nos dados biológicos, a Hamming Loss indicou que erros em classes minoritárias contribuíram mais para a perda total. Além disso, o uso de estratégias de balanceamento, embora eficaz, aumentou o tempo de treinamento devido à necessidade de ajustes iterativos nos pesos das classes.

A análise comparativa entre os dois conjuntos de dados também destacou a influência da diversidade de rótulos e do desbalanceamento no desempenho geral. O conjunto de dados de plantas, com mais classes e instâncias, permitiu uma avaliação mais robusta das estratégias aplicadas. Já os dados de vírus apresentaram limitações devido ao pequeno número de instâncias, impactando a capacidade de generalização do modelo.

Por fim, a escolha de redes neurais como abordagem principal mostrou-se efetiva. No entanto, a integração de métodos híbridos, como redes neurais combinadas com algoritmos evolutivos, pode ser explorada para mitigar limitações observadas.

5. Conclusão

Este trabalho demonstrou que redes neurais são ferramentas poderosas para a predição de localização subcelular de proteínas, especialmente em cenários multirrótulo e desbalanceados. As estratégias de regularização e ajuste de pesos foram essenciais para melhorar o desempenho em classes minoritárias, ampliando a aplicabilidade do modelo em cenários reais.

Apesar dos resultados promissores, o trabalho identificou limitações relacionadas ao desbalanceamento extremo e à escassez de dados em certas classes. Estudos futuros podem explorar técnicas complementares, como aumento de dados ou integração de métodos híbridos, para superar essas barreiras. Adicionalmente, a aplicação dos modelos desenvolvidos em novos conjuntos de dados pode validar sua robustez e escalabilidade. A continuidade desse trabalho tem o potencial de contribuir significativamente para avanços em biologia molecular e áreas correlatas.