UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
DEPARTAMENTO DE GENÉTICA E EVOLUÇÃO

LÍVIA UMBERTO BERTONI

Aprendizado de máquina multirrótulo para predição de doenças associadas a RNAs longos não codificantes

SÃO CARLOS
2025

LÍVIA UMBERTO BERTONI

Aprendizado de máquina multirrótulo para predição de doenças associadas a RNAs longos não codificantes

Trabalho de Conclusão de Curso apresentado junto à Universidade Federal de São Carlos como requisito para obtenção do título de Bacharel em Biotecnologia.

Orientador: Ricardo Cerri

Co-orientador: Andrea Fuentes

SÃO CARLOS
2025

RESUMO

RNAs longos não codificantes (lncRNAs) são RNAs com mais de 200 nucleotídeos e que não são traduzidos em proteínas. Esses desempenham funções cruciais em uma ampla gama de processos biológicos, como regulação do ciclo celular, epigenética, diferenciação celular, regulação pós-transcricional e transcricional. Em paralelo, um número crescente de estudos tem identificado associações entre expressões anormais de lncRNAs e diversas doenças humanas, incluindo cardiovasculares, neurológicas e oncológicas. Assim, para identificação e detecção de lncRNAs, têm sido empregadas diferentes abordagens biotecnológicas. No entanto, esses métodos enfrentam desafios consideráveis associados a custo, procedimento operacional e processo experimental. Dessa forma, a construção de modelos baseados em inteligência artificial para predição de doenças a partir de dados de lncRNAs apresenta-se como uma alternativa promissora. Este trabalho tem como objetivo a implementação de algoritmos de aprendizado de máquina para a predição de doenças associadas a RNAs longos não codificantes (lncRNAs). Tendo em vista que um mesmo lncRNA pode estar associado a mais de uma doença, trata-se de um caso de classificação multirrótulo. Neste cenário, foram testados diferentes métodos, de abordagens dependentes e independentes de algoritmo. O desempenho dos classificadores foi avaliado por meio de medidas de avaliação para problemas multirrótulo, sendo estas Precisão, Revocação, MedidaF e Hamming Loss. E, para fins de comparação, testes estatísticos foram performados, sendo estes Teste de Friedman e Teste de Nemenyi. A partir da análise dos resultados, foram formuladas hipóteses a respeito da performance dos algoritmos. No entanto, apesar de diferenças observadas, os métodos apresentaram um desempenho geral insatisfatório, o qual atribuímos à elevada esparsidade do conjunto de dados utilizado. Dessa forma, para trabalhos futuros, foi proposta a investigação de métodos mais robustos para lidar com este problema, de modo a aprimorar as predições e avançar na compreensão das interações biológicas.

Palavras chave: LncRNAs; Doenças; Predição; Aprendizado de Máquina; Classificação Multirrótulo

ABSTRACT

Long non-coding RNAs (lncRNAs) are RNA molecules longer than 200 nucleotides that are not translated into proteins. These RNAs play crucial roles in processes such as cell cycle regulation, epigenetics, cell differentiation, and both post-transcriptional and transcriptional regulation. Abnormal expression of lncRNAs has been increasingly associated with various human diseases, including cardiovascular, neurological, and oncological conditions. While different biotechnological approaches have been employed to identify and detect lncRNAs, these methods face significant challenges related to cost, operational complexity, and experimental procedure. Consequently, the development of artificial intelligence-based models for disease prediction using lncRNA data presents itself as a promising alternative. This study focuses on the implementation of machine learning algorithms for predicting diseases associated with lncRNAs, addressing it as a multi-label classification problem, given that a single lncRNA can be linked to multiple diseases. Various methods, including problem transformation approaches and algorithm adaptation approaches, were tested. The classifiers' performance was assessed using evaluation metrics for multi-label problems, including Precision, Recall, F-measure, and Hamming Loss. Furthermore, for comparative analysis, statistical tests were performed, specifically the Friedman and Nemenyi tests. Based on the analysis of the results, hypotheses regarding the performance of the algorithms were formulated. However, despite the observed differences, the methods showed an overall unsatisfactory performance, which we attribute to the high sparsity presented in the dataset. Thus, for future work, the investigation of more robust methods to address this issue was proposed, aiming to enhance prediction performance and improve the understanding of biological interactions.

Key-words: LncRNAs; Diseases; Prediction; Machine Learning; Multilabel Classification

SUMÁRIO

LISTA DE TABELAS

1 INTRODUÇÃO

RNAs longos não-codificantes (lncRNAs) são RNAs com mais de 200 nucleotídeos que não codificam proteínas (MERCER et al., 2009). Estes foram descobertos recentemente em eucariotos, e têm características estruturais de mRNAs (estrutura CAP e cauda poliA), porém não apresentam quadros de leitura (XU et al., 2021). Os lncRNAs tendem a ser pouco conservados entre as espécies e geralmente apresentam baixos níveis de expressão e alta especificidade tecidual (MERCER et al., 2008; PONTING et al., 2009).

Numerosos estudos demonstraram que os lncRNAs desempenham papel fundamental em diversas atividades vitais. Dentre as funções realizadas por estes RNAs, têm-se a regulação do ciclo celular (LIU et al., 2012), regulação epigenética (HIROTA et al., 2008), controle da diferenciação celular (HU et al., 2012) e regulação pós transcricional (CARRIERI et al., 2012). Outro papel importante consiste na regulação transcricional, promovendo ou prevenindo a ligação de fatores de transcrição e mediadores de transcrição aos promotores (FENG et al., 2006).

Tendo em vista a participação de lncRNAs em um vasto repertório de processos biológicos, mutações e desregulações desses RNAs estão associadas ao desenvolvimento e progressão de diversas doenças humanas complexas (GUPTA et al., 2010). Estas incluem diabetes (Pasmant et al., 2011), AIDS (ZHANG et al., 2013), doenças cardiovasculares (CONGRAINS et al., 2012), doenças neurológicas (JOHNSON, 2012) e diversos tipos de cânceres (SPIZZO, 2012; FANG, 2016).

Sendo assim, a identificação de lncRNAs pode contribuir significativamente para a compreensão dos mecanismos moleculares de doenças associadas. Isso possibilita a descoberta de lncRNAs como potenciais biomarcadores para diagnóstico, tratamento e prognóstico de doenças, bem como alvos terapêuticos em potencial para a descoberta de fármacos (YUAN et al, 2020). Como exemplo, MALAT1 (*Metastasis-Associated Lung Adenocarcinoma Transcript 1*) foi reportado como um biomarcador para detecção de vários cânceres, e sua supressão em células tumorais reduz metástase, proliferação e migração, sugerindo seu potencial terapêutico (AMODIO et al., 2018)

Nos últimos anos, um número crescente de abordagens biotecnológicas têm sido utilizado para identificação e detecção de lncRNAs. Estes métodos incluem microarray, tecnologias de sequenciamento de transcriptoma (RNA-seq), reação em cadeia da polimerase de transcrição reversa em tempo real (qRT-PCR), northern blot, entre outros (ZHU, 2013). No entanto, existem algumas dificuldades envolvidas na realização dessas técnicas, como alto

custo, procedimento operacional complexo e certo dano ao corpo humano durante o processo experimental (XU et al., 2021).

Métodos in silico baseados em aprendizado de máquina oferecem uma abordagem viável para a predição de associações entre lncRNAs e doenças de forma eficiente e rápida. Isso se deve ao fato de que o uso de métodos computacionais contribui para a redução de tempo e custos experimentais em comparação aos métodos tradicionais de pesquisa (WANG, 2024), permitindo análise simultânea e de grande escala. Diante deste cenário, o desenvolvimento de modelos baseados em inteligência artificial para a previsão de doenças a partir de dados de lncRNAs revela-se altamente promissor.

A predição associações lncRNA-doenças permite identificar novos alvos terapêuticos, possibilitando o desenvolvimento de tratamentos mais eficazes. Esta abordagem também viabiliza a detecção de biomarcadores para prognóstico, diagnóstico e monitoramento das condições clínicas. Além disso, a previsão destas interações torna possível a formulação de novas abordagens para prevenção de doenças. Tais contribuições podem representar um avanço significativo da medicina personalizada, proporcionando estratégias mais eficientes e direcionadas.

Neste contexto, a predição de doenças a partir de lncRNAs representa um problema de classificação multirrótulo, dado que um mesmo RNA pode estar simultaneamente associado a mais de uma doença. Desta forma, o presente trabalho tem como objetivo a investigação de métodos de classificação multirrótulo utilizando Aprendizado de Máquina (AM), na predição de doenças associadas a RNAs longos não codificantes (lncRNAs).

2 REVISÃO BIBLIOGRÁFICA

2.1 Classificação multirrótulo

Na literatura de Aprendizado de Máquina (AM), problemas de classificação convencionais são chamados de problemas monorrótulo. Nesses problemas, um classificador é treinado em um conjunto de instâncias que estão associados com uma única classe l de um conjunto de classes disjuntas L, onde $|L| > 1$. Se $|L| = 2$, então o problema é denominado problema de classificação binária, e se $|L| > 2$, o problema é chamado problema de classificação multi-classe. Em uma classificação multirrótulo (CM), as instâncias estão associadas a um conjunto de classes $Y \subseteq L$, sendo $|Y| > 1$ (TSOUMAKAS et al., 2010).

A CM ganhou muita importância nos últimos anos devido à sua ampla gama de domínios de aplicação. As áreas de aplicação incluem categorização de texto (JOACHIMS, 1998), rotulagem de mapas (ZHU;POON, 1999), categorização de imagens e cenas (BOUTELL et al. 2003), detecção de emoções (TROHIDIS, 2008) e bioinformática (CERRI et al., 2016). Além disso, no período de pandemia da Covid-19, a CM possibilitou o diagnóstico da doença juntamente com a pneumonia por meio de imagens de radiografia de tórax (KARAR, 2021), além da identificação de co-infecções associadas ao Covid-19 (BELLO et al, 2021).

A Figura 1 ilustra uma comparação entre um caso de classificação convencional, no qual instâncias podem ser atribuídas a apenas uma classe, e um caso de classificação multirrótulo. A Figura 1 (a) apresenta um problema de classificação em que um filme pode pode pertencer a apenas uma dentre duas classes (“Ação” ou “Ficção Científica”), mas nunca a ambas simultaneamente. A Figura 1 (b) ilustra um problema de classificação em que um filme pode ser atribuído simultaneamente às classes “Ação” e “Ficção Científica”. Assim, na Figura 1 (b), os filmes de dentro da região marcada são classificados tanto como filmes de “Ação” quanto filmes de “Ficção Científica”.

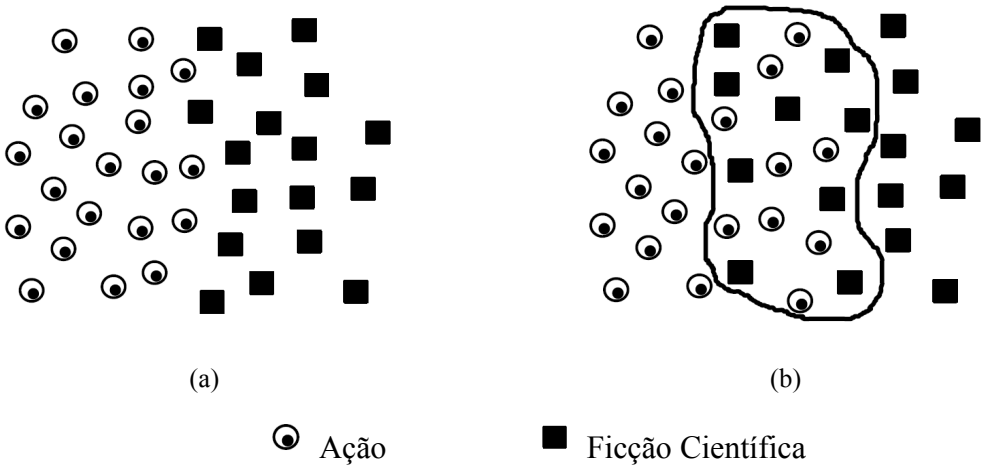

Figura 1: Exemplos de problemas de classificação: (a) classificação convencional (monorrótulo); (b) classificação multirrótulo. Fonte: Adaptado de (SHEN et al., 2003).

Dessa forma, diferentes técnicas têm sido propostas na literatura para tratar problemas de classificação multirrótulo. Estas pertencem a duas abordagens: abordagem independente de algoritmo e abordagem dependente de algoritmo. A abordagem independente de algoritmo utiliza algoritmos tradicionais de classificação para tratar problemas multirrótulo, transformando o problema multirrótulo original em um conjunto de problemas monorrótulo. A abordagem dependente de algoritmo cria algoritmos específicos para tratar o problema multirrótulo. Esses algoritmos podem ser baseados em algoritmos de classificação convencionais, como Máquinas de Vetores de Suporte (VAPNIK, 1999) e Árvores de Decisão (QUINLAN, 1993).

2.2 Abordagem Independente de Algoritmo

O método Binary-Relevance (BR) consiste em um método simples baseado na abordagem independente de algoritmo. Este utiliza L classificadores, sendo L o número de classes que estão envolvidas no problema. Cada classificador é então associado a uma classe e treinado para resolver um problema de classificação binária, na qual é considerada a classe a qual ele está associado contra todos os outros rótulos envolvidos. Esse método é chamado de Binary-Relevance (TSOUMAKAS et al., 2010).

Entretanto, um ponto fraco desse método é que este assume que as classes atribuídas a um exemplo são independentes entre si. Isso nem sempre é verídico, e ignorar possíveis correlações entre as classes pode fazer com que o método tenha pouca capacidade de generalização. Seu processo de transformação é reversível, ou seja, é possível recuperar as classes do problema original a partir do novo problema criado.

Nos trabalhos de Cherman et al. (2012), Read et al. (2009) e Dembczynski et al. (2010) foram propostos métodos baseados na transformação BR e que tentam considerar correlações entre as classes. A ideia é utilizar as classes das instâncias para complementar os vetores de atributos das mesmas, de maneira a incorporar no aprendizado as dependências existentes entre as classes do problema multirrótulo.

Outro método baseado nesta abordagem consiste no Label-Powerset (LP). Neste, para cada instância, todas as classes atribuídas a ela são combinadas em uma nova e única classe, transformando o problema em simples-rótulo. Esta técnica de combinação de rótulos foi utilizada em trabalhos como os de Tsoumakas e Katakis. (2010) e Boutell et al. (2004). Apesar do LP considerar a existência de correlação entre classes, o número de classes envolvidas no problema pode aumentar consideravelmente, e algumas destas podem acabar com poucas instâncias, dificultando a classificação.

No intuito de contornar essa desvantagem, Tsoumakas e Katakis (2010) propuseram o método Random k-Labelsets (RAKEL). Este constroi iterativamente uma combinação de classificadores Label-Powerset, e cada um é treinado utilizando subconjuntos randômicos do conjunto total de classes L, denominados *k-labelsets*. Para classificação de um novo exemplo, cada classificador toma uma decisão binária para cada rótulo do *k-labelset* correspondente. Uma decisão média é calculada para cada rótulo em L, e a decisão final é positiva para dado rótulo se a decisão média for maior que um limiar t.

2.3 Abordagem Dependente de Algoritmo

Um trabalho que utiliza árvores de decisão foi proposto por Clare e King (2001). Neste trabalho, os autores modificaram o algoritmo C4.5 (QUINLAN, 1993) para a classificação de proteínas de acordo com suas funções. O algoritmo C4.5 define os nós da árvore de decisão por meio de uma medida chamada entropia. Os autores modificaram a fórmula dessa medida, originalmente elaborada para problemas monorrótulo, de maneira a permitir seu uso em problemas multirrótulo. Outra modificação feita pelos autores foi a utilização dos nós-folha da árvore para representar conjuntos de rótulos,. Quando um nó-folha, alcançado na classificação de uma instância, contém um conjunto de classes, uma regra separada é produzida para cada classe.

Em Zhang e Zhou (2005) foi proposto um método para classificação multirrótulo baseado no algoritmo KNN, denominado *Multi-Label k-Nearest Neighbors* (MLkNN). Nesse método, para cada instância, as classes associadas com os k exemplos vizinhos mais próximos são recuperadas, e é feita uma contagem dos vizinhos associados a cada classe. Então, o

princípio *maximum a posteriori* é utilizado para definir o conjunto de classes de uma nova instância.

O trabalho de Spyromitros et al. (2008) propõe outro classificador multirrótulo baseado no algoritmo KNN. *Binary Relevance k-Nearest Neighbors* (BRkNN) combina a técnica de transformação Binary Relevance (BR) com kNN, mas de maneira mais eficiente. Isso porque o método estende o algoritmo kNN de forma que previsões independentes são feitas para cada rótulo, após uma única busca pelos *k* vizinhos mais próximos.

No trabalho de Zhang e Zhou (2006) foi proposta uma medida de erro para ser utilizada no treinamento de redes neurais com o algoritmo *Backpropagation*. A medida considera as múltiplas classes dos exemplos no cálculo do erro de classificação.

Em Schapire e Singer (1999) e Schapire e Singer (2000) foram propostas duas extensões para o algoritmo *Adaboost* (FREUND; SCHAPIRE, 1995), de maneira a permitir seu uso em problemas multirrótulo. Na primeira, é feita uma modificação na maneira de se avaliar o desempenho preditivo do modelo induzido, verificando sua capacidade de predizer um conjunto correto de classes para uma dada instância. Na segunda, uma mudança no algoritmo faz com que ele passe a predizer um *ranking* de classes para cada instância de entrada.

Em Thabtah et al. (2004) foi proposto um algoritmo multirrótulo baseado em regras de associação de classes. O algoritmo foi chamado de classificação associativa multi-classe multirrótulo (MMAC). Inicialmente é criado um conjunto de regras, e então são removidas todas as instâncias associadas a esse conjunto. As instâncias restantes são então utilizadas para criar um novo conjunto de regras. Esse procedimento é realizado até que não restem mais instâncias.

Um algoritmo de classificação baseado em entropia foi utilizado por Zhu et al. (2005) para a tarefa de recuperação de informação. Os autores utilizaram o modelo para explorar correlações entre as classes de documentos multirrótulo.

O trabalho de Benites e Sapozhnikova (2015) propõe uma extensão da rede neural *Fuzzy Adaptive Resonance Associative Map* (ARAM), para o caso de CM. O *Multi Label* ARAM introduz um novo nível de organização e agrupamento de protótipos, onde um segundo nível de clustering é incorporado ao processo de aprendizado. O modelo visa lidar com dados de grande volume e de alta dimensionalidade.

Madjarov et al. (2012) publicaram um trabalho no qual vários métodos de classificação, baseados tanto na abordagem dependente e independente de algoritmo, foram comparados. Várias medidas de avaliação também foram utilizadas nos experimentos. Os

melhores desempenhos foram obtidos por métodos que tentam considerar as dependências entre as classes do problema durante a fase de treinamento.

2.4 Trabalhos Relacionados

Recentemente, diversos modelos computacionais foram propostos para prever potenciais associações lncRNA-doenças. No trabalho de Wang et al. (2020), foi proposta a implementação de *deep forest* na CM para predição de associações lncRNA-doenças. Um tópico relevante deste trabalho é que considera correlações entre rótulos como principal informação. Em cada camada, a distribuição de classes estimada é empregada no treinamento de cada floresta. Por fim, resultados de votação de múltiplos classificadores fracos são usados para determinar a qual classe uma amostra de teste deve pertencer. O método foi comparado com outras abordagens de classificação multirrótulo, obtendo o melhor desempenho no conjunto de dados utilizado. Assim, mostrou-se um método eficaz na CM para cenários biológicos.

No estudo de Wei et al (2020) foi proposto um novo preditor chamado iLncRNAdis-FB[1] para identificar novas associações lncRNA-doenças. O método se baseia em Redes Neurais Convolucionais (CNNs) para integrar diferentes fontes de dados usando os blocos de atributos de maneira supervisionada. Uma matriz de semelhança de lncRNA e uma matriz de semelhança de doenças são construídas e, baseando-se nestas, blocos de atributos tridimensionais são gerados. Esses blocos de atributos são então alimentados na CNN para treinar o modelo de modo a prever associações doenças-lncRNA desconhecidas.

O trabalho de Yao et al. (2020) implementou um modelo de previsão baseado em *Random Forest* e seleção de atributos, denominado RFLDA. Primeiramente, RFLDA integra as associações miRNA-doença (MDAs) e associações lncRNA-doença (LDAs), a similaridade semântica de doenças (DSS), a similaridade funcional de lncRNAs (LFS) e as interações lncRNA-miRNA (LMI) como atributos de entrada. Em seguida, o método seleciona os atributos mais relevantes para treinar o modelo. Esta seleção é baseada nos valores de medidas de importância (*feature importance*), que leva em consideração não apenas o efeito de atributos individuais nos resultados de previsão, mas também os efeitos conjuntos de vários atributos. Por fim, um modelo de regressão baseado em *Random Forest* é treinado para pontuar as potenciais associações lncRNA-doença.

[1] http://bliulab.net/iLncRNAdis-FB/

3 OBJETIVOS

3.1 Objetivo Geral

Este trabalho tem como principal objetivo a investigação de métodos de classificação multirrótulo utilizando Aprendizado de Máquina (AM), na predição de doenças associadas a RNAs longos não codificantes (lncRNAs). Desta forma, serão pesquisados e analisados métodos de predição utilizando classificadores.

3.2 Objetivos Específicos

São objetivos específicos deste trabalho:

- Construir conjuntos de dados de predição de doenças associadas a lncRNAs, a partir de bases de dados existentes, para algoritmos de AM;
- Implementar e comparar algoritmos da biblioteca scikit-multilearn neste problema de CM;
- Observar como medidas de densidade e cardinalidade de rótulo influenciam no desempenho dos métodos;
- Utilizar diferentes métricas de avaliação e verificar seu impacto nos resultados da avaliação;
- Aplicar testes estatísticos para análise e comparação das performances dos classificadores.

4 MATERIAIS E MÉTODOS

4.1 Conjunto de Dados

Os conjuntos de dados foram originados a partir de associações lncRNA-doenças experimentalmente documentadas e publicamente disponíveis[2]. O LncRNADisease database apresenta informações como o nome do RNA, doenças associadas, disfunção relativa a doença, descrição da associação, cromossomo localizado, posicionamento genômico, espécie, ALIAS, número de acesso (ID) no Genbank[3] e referências. Este banco apresenta uma ampla cobertura de dados, tratando-se de uma plataforma robusta e atualizada para investigação científica.

O LncRNADisease abrange uma ampla gama de doenças, incluindo câncer, doenças neurodegenerativas, cardiovasculares e metabólicas. Por exemplo, no câncer, o banco de dados inclui doenças como câncer de mama, câncer de pulmão, câncer de ovário, câncer colorretal e leucemia mieloide aguda. Nas doenças neurodegenerativas, são documentadas associações com doenças como Alzheimer, doença de Parkinson e esclerose múltipla. Entre as doenças cardiovasculares, o banco contempla condições como doença cardíaca coronariana e insuficiência cardíaca. No âmbito das doenças metabólicas, exemplos como diabetes são frequentemente investigados.

Este database fornece, também, informações detalhadas a respeito do envolvimento dos lncRNAs na patologia das doenças. A disfunção relativa à doença informa qual anormalidade ou alteração específica no lncRNA está relacionada ao desenvolvimento ou progressão da doença. Isso pode incluir alterações em sua expressão, modificações na sua regulação ou mutações que alteram estrutura ou função. Já a descrição da associação oferece um contexto mais amplo sobre como a interação entre lncRNAs e doenças foi estabelecida, detalhando também os processos celulares específicos envolvidos.

O download do banco de dados foi realizado no formato CSV (Comma-separated values), e o dataset foi manuseado no formato de Pandas DataFrame. Este consiste em uma estrutura bidimensional, de modo que, neste caso, as 2947 linhas originadas correspondem aos lncRNAs, e as 12 colunas às informações.

No processo de filtragem dos dados, colunas foram removidas a fim de manter apenas o *ID* dos RNAs e suas respectivas doenças associadas. Assim, as sequências dos RNAs foram obtidas a partir de seus IDs do Genbank. Para isso, foi utilizado o pacote Bio.Entrez da biblioteca Biopython, o qual fornece código para acessar o NCBI pela internet

[2] https://www.cuilab.cn/lncrnadisease
[3] https://www.ncbi.nlm.nih.gov/genbank/

(WWW). No processo, um desafio encontrado foi a presença de valores “NaN” nos IDs dos RNAs, que consistem em valores “faltantes”. Estes impossibilitaram o acesso às sequências dos RNAs, os quais foram deletados. Como consequência, houve uma redução significativa na quantidade de linhas do Dataframe.

Em seguida, uma matriz binária foi construída para representar as associações entre lncRNAs e doenças, na qual a presença (1) ou ausência (0) de uma doença foi utilizada como classe para cada sequência de lncRNA. Essa abordagem permite a representação de problemas de classificação multirrótulo, em que cada linha da matriz corresponde a um lncRNA e cada coluna corresponde a uma doença. Um valor 1 indica que o lncRNA está associado à respectiva doença, enquanto um valor 0 indica que não há associação conhecida.

Neste cenário, para melhor execução dos algoritmos de classificação, doenças sem qualquer RNA associado e doenças associadas a apenas um único RNA foram excluídas do conjunto de dados. Isso pois esses casos não fornecem um número de exemplos positivos significativamente representativo de cada doença, o que inviabiliza o aprendizado pelos classificadores multirrótulo.

Dessa forma, o conjunto de dados final apresentou 376 lncRNAs e 129 doenças. Para facilitar a consulta, uma lista completa das doenças analisadas neste estudo está disponível no Apêndice deste trabalho. Essa relação detalhada permite uma visão abrangente das condições investigadas, contribuindo para a compreensão do impacto dos lncRNAs em diferentes patologias.

Os procedimentos de construção de conjunto de dados foram efetuados na linguagem de programação *Python*, versão 3.9.15. Neste, foram utilizadas as bibliotecas Biopython[4] (versão 1.78) e Pandas[5] (versão 1.5.1).

4.2 Métricas de Caracterização de Dados Multirrótulo

Para uma avaliação mais detalhada e precisa das características de um conjunto de dados multirrótulo, é necessário considerar métricas específicas que ajudam a entender a distribuição de rótulos atribuídos às instâncias. Nesse contexto, duas medidas são frequentemente utilizadas: densidade de rótulo e cardinalidade de rótulo, conforme descrito por Tsoumakas et al. (2010). A cardinalidade de um conjunto de dados multirrótulo é a média do número de rótulos das instâncias do conjunto. A densidade de um conjunto de dados

[4] https://biopython.org/
[5] https://pandas.pydata.org/

multirrótulo é a média do número de rótulos das instâncias do conjunto dividida pelo número total de rótulos.

Formalmente, em problemas multirrótulo, a entrada para os algoritmos de classificação é um conjunto de dados *S*, com *N* exemplos T_i , $i = 1, ..., N$, escolhido de um domínio *X* com distribuição fixa, arbitrária e desconhecida, da forma (x_i, Y_i), com $i = 1, ..., N$, para alguma função desconhecida f(x) = *Y*. *L* é o conjunto dos possíveis rótulos e $Y_i \subseteq L$, ou seja, Y_i é o conjunto de rótulos do i-ésimo exemplo. A cardinalidade pode ser definida pela Equação 1, enquanto a densidade pode ser definida pela Equação 2.

$$Card = \frac{1}{N}\sum_{i=1}^{N}|Yi| \quad (1)$$

$$Dens = \frac{1}{N}\sum_{i=1}^{N}\frac{|Yi|}{|L|} \quad (2)$$

O valor de cardinalidade de rótulo deste conjunto de dados é de 2.7. Isto significa que, em média, cada RNA está associado a 2.7 doenças. Tendo em vista um total de 129 doenças no dataset, essa média é relativamente baixa, indicando que cada RNA está associado a uma fração pequena do total de doenças possíveis. Esse valor baixo também pode ser explicado pelo fato de que a classificação não está sendo feita por classes de doenças, e sim de modo individual.

O valor da densidade de rótulo obtido para este conjunto de dados é de 0.02, um valor considerado baixo. Esse resultado pode ser atribuído à alta esparsidade do conjunto de dados. Isto significa que há uma predominância de associações ausentes ou não estabelecidas entre os RNAs e as doenças, o que resulta em um conjunto de dados caracterizado por uma alta proporção de valores nulos na matriz binária.

4.3 Extração de Atributos

A extração de atributos para cada sequência de lncRNA foi realizada utilizando o programa Pse-in-One 2.0[6], um servidor web para gerar modos abrangentes de pseudocomponentes de sequências de DNA, RNA e proteínas. O programa é capaz de gerar diferentes tipos de atributos para sequências de RNA. Assim, diferentes métodos podem ser

[6] http://bliulab.net/Pse-in-One2.0/home/

utilizados para gerar os vetores de atributos com base nos arquivos de sequência de entrada e nos métodos de extração de atributos selecionados (LIU et al., 2015).

O primeiro método é usado para calcular os modos na categoria composição de ácido ribonucleico, apresentando três atributos. K-mer é a abordagem mais simples para representar os RNAs, em que sequências de RNA são representadas como as frequências de ocorrência de *k* ácidos nucleicos vizinhos. Mismatch calcula a ocorrência de ácidos nucleicos vizinhos de comprimento *k* que diferem em, no máximo, *m* ($m<k$). Subsequência é uma abordagem que permite correspondências (*matches*) não contíguos.

O segundo método é usado para calcular os modos na categoria de autocorrelação, apresentando seis atributos. Autocovariância baseada em dinucleotídeos (DAC) mede a correlação do mesmo índice físico-químico entre dois dinucleotídeos separados por uma distância *lag* ao longo da sequência. A covariância cruzada baseada em dinucleotídeos (DCC) mede a correlação de dois índices físico-químicos diferentes entre dois dinucleotídeos separados por *lag* ácidos nucleicos ao longo da sequência. Covariância auto-cruzada baseada em dinucleotídeos (DACC) é uma combinação de DAC, em que o comprimento do vetor é N*N*LAG, onde N é o número de índices físico-químicos e LAG é o máximo de *lag*. A abordagem de autocorrelação de Moreau–Broto normalizada (NMBAC) mede a correlação das mesmas propriedades entre dois resíduos separados por uma distância *lag* ao longo da sequência.

O terceiro método é usado para calcular os modos na categoria composição de pseudo nucleotídeos, apresentando dois atributos. Na abordagem de Composição de pseudo-dinucleotídeos de correlação paralela geral (PC-PseDNC-General), os usuários podem selecionar 22 índices físico-químicos para gerar o vetor de atributos. Composição de pseudo-dinucleotídeos de correlação de série geral (SC-PseDNC-Geral) é uma variante de PC-PseDNC-General, que difere nas equações de cálculo dos fatores de correlação, refletindo na correlação da ordem de sequência entre todos os dinucleotídeos mais contíguos ao longo de uma sequência de RNA.

Ao término da etapa de extração dos atributos das sequências, cada conjunto de dados de atributos foi individualmente concatenado à matriz binária de associação lncRNA-doenças, originando um total de oito conjuntos, os quais combinam atributos e associações. Essa estrutura de conjunto de dados é necessária para a implementação no scikit-multilearn.

4.4 Scikit-multilearn

O scikit-multilearn[7] é uma biblioteca para classificação multirrótulo, licenciada sob BSD, desenvolvida com base na renomada biblioteca de aprendizado de máquina scikit-learn (SZYMANSKI; KAJDANOWICZ, 2017). A biblioteca oferece implementações de métodos amplamente utilizados para classificação multirrótulo, proporcionando soluções avançadas e eficientes para a resolução de problemas nessa área. Totalmente integrada ao ecossistema científico e de aprendizado de máquina em Python, a biblioteca disponibiliza ferramentas eficientes e especializadas, além de suporte para estratificação de dados, acesso e manipulação de conjuntos de dados.

4.5 Métodos de Classificação

Este trabalho tem como foco a comparação de algoritmos do scikit-multilearn. Nesse contexto, foram selecionados classificadores representativos das abordagens dependentes e independentes de algoritmo. A escolha destes foi feita de forma estratégica, com o objetivo de permitir uma análise mais completa e abrangente das diferenças de desempenho entre os métodos na tarefa de predição de associações lncRNA-doenças. Assim, diferentes métodos foram implementados.

Seguindo a abordagem independente de algoritmo, foi utilizada a técnica Classifier Chains (CC) (READ et al., 2009). Baseados na abordagem dependente de algoritmo, foram utilizados os classificadores Multi-label ARAM (MLARAM) (BENITES; SAPOZHANIKOVA, 2015), Multi-label k-Nearest Neighbors (MLkNN) (ZHANG; ZHOU, 2007) e Binary Relevance k-Nearest Neighbors (BRkNN) (SPYROMITROS et al., 2008). Este classificador apresenta duas versões, e ambas foram implementadas neste estudo. A versão A atribui os rótulos associados a pelo menos metade dos vizinhos, enquanto a versão B atribui os *m* rótulos mais frequentes entre os vizinhos, onde *m* é o número médio de rótulos atribuídos aos vizinhos do exemplo.

4.6 Validação cruzada estratificada para problemas multirrótulo

Nesta proposta, foi utilizada uma extensão da validação cruzada específica para problemas multirrótulo. O trabalho de Tsoumakas (2011) propõe um algoritmo iterativo de estratificação multirrótulo que objetiva a manutenção da distribuição de exemplos positivos e negativos de cada classe. A entrada para o algoritmo é um conjunto de dados multirrótulo, *D*, anotado com um conjunto de rótulos $L = \{\lambda 1, ..., \lambda q\}$, um número desejado de subconjuntos *k,*

[7] http://scikit.ml

e uma proporção desejada de exemplos em cada subconjunto, *r1,...rk.* Neste trabalho, os experimentos foram realizados com um valor de *k* igual a 10.

O algoritmo inicia com o cálculo do número de exemplos desejado, c_j, em cada subconjunto. Então, o número desejado de amostras para cada rótulo λ_i em cada subgrupo é calculado. O algoritmo é iterativo e, em cada iteração, o rótulo com o menor número de exemplos remanescentes, designado, é examinado. Então, o algoritmo escolhe um subconjunto aceitável para a distribuição de cada instância *(x, Y)* desse rótulo. O primeiro critério para a seleção do subconjunto é o número atual desejado de exemplos para este rótulo c_j^l . O subconjunto que o maximiza é selecionado. Depois de escolher o subconjunto adequado *m*, adiciona-se o exemplo *(x, Y)* a S_m e o retira de *D*. O algoritmo terminará assim que o conjunto de dados original ficar vazio.

4.7 Medidas de Avaliação

A avaliação de classificadores multirrótulo requer medidas diferentes das utilizadas em problemas de classificação simples. Diferentemente da classificação monorrótulo, na qual uma instância é classificada de maneira correta ou errada, na classificação multirrótulo, uma instância pode ser classificada de modo parcialmente correto ou parcialmente errado. Isto ocorre quando um classificador atribui corretamente a uma instância pelo menos uma das classes a que ela pertence, mas não atribui à instância uma ou mais classes às quais ela pertence. É possível também a atribuição de uma ou mais classes às quais a instância não pertence.

Considerando que cada instância de um conjunto multirrótulo é representada por uma tupla (x_i, Y_i) em que i = 1...*m* (sendo *m* o número total de instâncias), $Y_i \subseteq L$ é o conjunto de classes reais e $L = \{\lambda_j : j = 1...q\}$ é o conjunto de todas as classes do problema. Dada uma instância x_i, o conjunto de classes preditas para essa instância é denominado Z_i . Δ representa a diferença simétrica entre dois conjuntos, e corresponde a operação XOR da lógica booleana (TSOUMAKAS; KATAKIS, 2007).

Neste trabalho, serão utilizadas as medidas Hamming Loss, Precisão, Revocação e MedidaF (F1) (SCHAPIRE ; SINGER, 2000; GODBOLE; SARAWAGI, 2004). As métricas estão apresentadas nas Equações 3- 6, respectivamente.

$$Hamming\ Loss = \frac{1}{m} \sum_{i=1}^{m} \frac{Yi\ \Delta\ Zi}{|L|} \tag{3}$$

$$Precisão = \frac{1}{m} \sum_{i=1}^{m} \frac{|Yi \cap Zi|}{|Zi|} \quad (4)$$

$$Revocação = \frac{1}{m} \sum_{i=1}^{m} \frac{|Yi \cap Zi|}{|Yi|} \quad (5)$$

$$MedidaF = \frac{1}{m} \sum_{i=1}^{m} \frac{2|Yi \cap Zi}{|Zi| + |Yi|} \quad (6)$$

4.8 Testes Estatísticos

O teste de Friedman consiste em um procedimento estatístico não paramétrico utilizado para comparar três ou mais amostras pareadas, ou relacionadas (FRIEDMAN, 1937). Baseia-se na atribuição de *rankings* aos dados, ordenando os valores de menor para maior. A aplicação do teste neste trabalho tem como intuito comparar as performances dos algoritmos de classificação multirrótulo, com a finalidade de verificar se existem diferenças significativas entre elas.

Seja r_i^j a posição no ranking do *j*-ésimo algoritmo, de um total de *k* algoritmos, no *i*-ésimo conjunto de dados, de um total de *N* conjuntos de dados. O teste de Friedman compara o ranking médio dos algoritmos, dado por $R_j = \frac{1}{N}\sum_i r_i^j$. Sob a hipótese nula, que considera que todos os algoritmos sendo comparados são equivalentes e têm seus valores de R_j iguais, a estatística de Friedman é dada pela Equação 7 (DEMSAR, 2006).

$$\chi_F^2 = \frac{12N}{k \cdot (k+1)} \left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right] \quad (7)$$

A estatística de Friedman é distribuída de acordo com χ^2_F com *k* - 1 graus de liberdade. Tabelas desta distribuição e seus valores críticos podem ser encontradas em livros de estatística. Assim, o valor crítico obtido na tabela, para determinado valor de α, é comparado com o valor calculado. Se o valor da tabela for menor, a hipótese nula é rejeitada.

Se a hipótese nula for rejeitada, ou seja, se for detectada diferença estatística entre os algoritmos, o teste de Nemenyi (NEMENYI, 1963) é aplicado para uma análise post hoc. O teste foi utilizado para comparar combinações de algoritmos, de modo a identificar quais

pares são diferentes. O desempenho entre dois algoritmos é estatisticamente significativo se seus respectivos *ranking médios* diferem no mínimo de um valor crítico *CD*. O cálculo desse valor é representado na Equação 9. Os valores de q_α são obtidos por meio de tabelas (DEMSAR, 2006).

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{8}$$

5 RESULTADOS E DISCUSSÃO

5.1 Medidas de Avaliação

As tabelas a seguir apresentam os resultados de cada medida de avaliação para os classificadores aplicados a cada conjunto de dados gerado a partir dos atributos extraídos. Nas tabelas, as linhas representam os classificadores multirrótulo utilizados, enquanto as colunas correspondem aos diferentes atributos extraídos das sequências de lncRNAs. Cada tabela é dedicada a uma medida de avaliação específica. As Tabelas 1-4 correspondem, respectivamente, às medidas Precisão, Revocação, MedidaF e Hamming Loss. Os maiores valores para cada atributo (conjunto de dados) estão destacados em negrito.

Tabela 1 - Resultados para a métrica de Precisão

	DAC	**DCC**	**DACC**	**NMBAC**	**Kmer**	**Mismatch**	**Pc**	**Sc**
BRkNNa	0,006	**0,005**	**0,006**	0,004	0,053	0,008	0,006	0,005
BRkNNb	**0,016**	0,001	0,003	0,002	0,011	0,009	**0,028**	0,003
MLkNN	0,006	0,004	0,004	0,004	**0,074**	0,009	0,006	**0,006**
MLARAM	0,009	0,002	0,001	**0,005**	0,003	**0,012**	0,007	0,001
Classifier Chain	0,001	0,001	0,002	0,003	0,023	0,005	0,006	**0,006**

Fonte: Elaborado pelo autor.

Tabela 2 - Resultados para a métrica de Revocação

	DAC	**DCC**	**DACC**	**NMBAC**	**Kmer**	**Mismatch**	**Pc**	**Sc**
BRkNNa	0,004	0,003	0,003	0,001	0,024	0,004	0,003	0,003
BRkNNb	0,011	0,003	0,004	0,002	0,016	0,011	**0,024**	0,007
MLkNN	0,003	0,002	0,002	0,001	**0,061**	0,004	0,003	0,003
MLARAM	**0,017**	**0,011**	**0,008**	**0,013**	0,011	**0,044**	0,021	**0,008**
Classifier Chain	0,001	0,000	0,001	0,001	0,009	0,002	0,003	0,003

Fonte: Elaborado pelo autor.

Tabela 3 - Resultados para a métrica de MedidaF

	DAC	DCC	DACC	NMBAC	Kmer	Mismatch	Pc	Sc
BRkNNa	0,005	**0,003**	**0,004**	0,002	0,031	0,005	0,004	**0,004**
BRkNNb	**0,010**	0,001	0,002	0,001	0,011	0,007	**0,021**	0,003
MLkNN	0,004	0,002	0,003	0,002	**0,061**	0,005	0,004	0,003
MLARAM	**0,010**	**0,003**	0,002	**0,005**	0,004	**0,016**	0,009	0,001
Classifier Chain	0,001	0,000	0,001	0,001	0,012	0,003	0,003	0,003

Fonte: Elaborado pelo autor.

Tabela 4 - Resultados para a métrica de Hamming Loss

	DAC	DCC	DACC	NMBAC	Kmer	Mismatch	Pc	Sc
BRkNNa	0,023	0,025	0,024	0,023	0,020	0,024	0,024	0,024
BRkNNb	**0,050**	**0,043**	**0,042**	**0,045**	**0,043**	0,052	**0,049**	**0,048**
MLkNN	0,023	0,024	0,025	0,023	0,024	0,023	0,024	0,023
MLARAM	0,034	0,028	0,026	0,030	0,030	**0,072**	0,033	0,029
Classifier Chain	0,022	0,021	0,021	0,021	0,020	0,022	0,022	0,022

Fonte: Elaborado pelo autor.

De maneira geral, constata-se que as métricas de avaliação apresentam resultados insatisfatórios, fato que pode ser atribuído à elevada esparsidade dos dados analisados, conforme evidenciado pela cardinalidade e densidade dos rótulos. Observa-se que a maioria dos elementos na matriz binária de associação entre RNAs e doenças possui valor igual a zero, o que indica interações negativas ou desconhecidas. Todavia, os elementos de maior

interesse são aqueles representados pelo valor 1, os quais correspondem às associações efetivas entre RNAs e doenças.

Trata-se de um problema de classificação desafiador, caracterizado por um espaço de classes altamente esparso. Neste cenário, o reduzido número de exemplos positivos para muitas classes compromete o aprendizado de padrões representativos e generalizáveis. Ademais, a capacidade de identificar interações entre diferentes rótulos fica comprometida. Ainda, deve-se destacar que anotações negativas para as classes não indicam, necessariamente, inexistência de associação lncRNA-doença. Em vez disso, esses valores podem refletir falta de evidências experimentais para confirmar uma associação. Entretanto, modelos podem interpretar valores nulos como ausências reais de associações, limitando a capacidade de prever interações futuras.

Assim, a alta esparsidade do conjunto de dados prejudica a performance de predição dos classificadores. Este fator afeta significativamente as métricas de avaliação, como pode ser observado neste trabalho. Em um cenário de escassez de exemplos positivos para aprendizado, os modelos podem frequentemente realizar previsões de associações não existentes, resultando em falsos positivos e, portanto, baixos valores de precisão. Da mesma forma, a medida de Revocação é reduzida, pois a falta de dados positivos dificulta a identificação de todas as instâncias relevantes para uma classe. Como resultado, a MedidaF também fica comprometida, uma vez que esta é a média harmônica entre Precisão e Revocação.

Além disso, a predominância de valores nulos pode impactar, também, a métrica Hamming Loss, que calcula a fração de rótulos incorretamente previstos. Primeiramente, pode-se ter um aumento da probabilidade de erros em predições negativas, ou seja, não associações erroneamente preditas como positivas. Ademais, os modelos podem apresentar dificuldades em capturar corretamente as associações positivas reais, havendo falha na predição de associações positivas corretas.

5.2 Teste de Friedman

A tabela a seguir apresenta os resultados do teste estatístico de Friedman. O teste verifica se existem diferenças significativas entre os algoritmos para cada métrica de avaliação analisada, considerando os valores do p-value obtidos.

Para esta análise, adotamos um nível de significância de 5% (α=0,05). Dessa forma, se o p-value for menor que 0,05, rejeitamos a hipótese nula de que as performances dos algoritmos são equivalentes, concluindo que há diferenças estatisticamente significativas.

A Tabela 5 apresenta os valores de p-value obtidos pelo teste de Friedman, para cada métrica de avaliação. Os casos em que se observou significância estatística estão destacados em negrito.

Tabela 5 - Valores de p-value resultantes do teste de Friedman

Métrica de Avaliação	**p-value**
Precisão	0.3485
MedidaF	**0.0434**
Revocação	**0.0002**
H. Loss	**5.198e-06**

Fonte: Elaborado pelo autor.

Inicialmente, no que tange à medida de Precisão, observa-se que o p-value é superior a 0,05. Dessa forma, as diferenças nas performances dos algoritmos para essa métrica podem ser atribuídas ao acaso. Conclui-se, portanto, que não existem diferenças estatisticamente significativas entre os algoritmos em relação a esta métrica.

A respeito das medidas F1 e Revocação, observa-se diferenças estatisticamente significativas entre os algoritmos. Em relação à medida F1, o valor de p-value obtido foi inferior a 0,05, sugerindo que existem diferenças significativas nas performances dos algoritmos. Da mesma forma, o p-value para a métrica de Revocação foi substancialmente menor que 0,05, apontando para a conclusão de que as diferenças observadas são altamente significativas.

Por fim, para a métrica Hamming Loss, o p-value obtido foi extremamente baixo e muito inferior ao nível de significância adotado. Isto evidencia que as diferenças observadas nas performances dos algoritmos em relação a essa métrica são estatisticamente significativas. Assim, a magnitude do p-value ressalta que as discrepâncias entre os algoritmos não podem ser atribuídas ao acaso, mas sim a diferenças consistentes em seus desempenhos.

5.3 Teste de Nemenyi

Para aprofundar a análise das diferenças identificadas pelo teste de Friedman, foi realizado o teste de Nemenyi para comparações pareadas entre os algoritmos de classificação multirrótulo. Este teste permite identificar especificamente quais algoritmos diferem significativamente entre si, avaliando os pares de métodos em cada métrica de avaliação.

O teste de Nemenyi calcula o valor de p-value para cada pareamento, indicando a probabilidade de que as diferenças observadas sejam atribuídas ao acaso. Assim como no teste de Friedman, adotou-se um nível de significância de 5% (α=0,05). Assim, se p-value for menor que 0,05, rejeitamos a hipótese nula para o par de algoritmos analisados, concluindo que suas performances diferem de forma estatisticamente significativa.

As tabelas a seguir apresentam os resultados do teste de Nemenyi para cada métrica de avaliação considerada. As Tabelas 6-9 correspondem, respectivamente, às medidas Precisão, Revocação, Medida F e Hamming Loss. Cada tabela exibe as comparações par a par entre os algoritmos, juntamente com os valores de p-value associados. Os casos em que se observou significância estatística estão destacados em negrito.

Tabela 6 - Resultados do teste de Nemenyi para a métrica de Precisão

	BRkNNa	**BRkNNb**	**MLkNN**	**MLARAM**
BRkNNb	0.97	-	-	-
MLkNN	1.00	0.88	-	-
MLARAM	0.99	1.00	0.93	-
Classifier Chain	0.51	0.88	0.32	0.80

Fonte: Elaborado pelo autor.

Tabela 7 - Resultados do teste de Nemenyi para a métrica de Revocação

	BRkNNa	**BRkNNb**	**MLkNN**	**MLARAM**
BRkNNb	0.56109	-	-	-
MLkNN	0.99486	0.31864	-	-
MLARAM	0.17454	0.95395	0.06872	-
Classifier Chain	0.40974	**0.01046**	0.66359	**0.00074**

Fonte: Elaborado pelo autor.

Tabela 8 - Resultados do teste de Nemenyi para a métrica MedidaF

	BRkNNa	**BRkNNb**	**MLkNN**	**MLARAM**
BRkNNb	0.908	-	-	-
MLkNN	0.954	1.000	-	-
MLARAM	0.995	0.990	0.998	-
Classifier Chain	**0.045**	0.319	0.240	0.123

Fonte: Elaborado pelo autor.

Tabela 9 - Resultados do teste de Nemenyi para a métrica Hamming Loss

	BRkNNa	**BRkNNb**	**MLkNN**	**MLARAM**
BRkNNb	**0.022**	-	-	-
MLkNN	1.000	**0.017**	-	-
MLARAM	0.240	0.878	0.205	-
Classifier Chain	0.363	**1.4e-05**	0.410	**0.001**

Fonte: Elaborado pelo autor

Na métrica de Precisão, portanto, não foi observada diferença estatisticamente significativa entre os métodos, conforme já evidenciado pelo teste de Friedman. Na métrica de Revocação, observamos diferenças entre Classifier Chains e BRkNNb, e Classifier Chains e MLARAM. A respeito da MedidaF, existe diferença entre os métodos Classifier Chains e BRkNNa.

Quanto à métrica Hamming Loss, quatro pares de algoritmos diferem entre si. São estes BRkNNa e BRkNNb, MLkNN e BRkNNb, Classifier Chains e MLARAM e, por fim, Classifier Chains e BRkNNb, onde foi observado o menor valor de p-value, evidenciando maior discrepância entre os dois métodos.

5.4 Análise dos Algoritmos

Com base nos resultados obtidos, algumas hipóteses podem ser formuladas. Primeiramente, apesar de não ter sido detectada significância estatística na métrica de precisão, o algoritmo Classifier Chains apresenta os menores valores nesta métrica. Além disso, para a

métrica de Revocação, apresenta desempenho significativamente menor que algoritmos como BRkNNb e MLARAM. Da mesma forma, para a MedidaF, o método CC apresenta menores valores que BRkNNa.

Pode-se sugerir que esta baixa performance de CC, comparado aos outros classificadores, se deve a dois principais fatores. Primeiramente, ao alto número de classes no dataset, que torna mais difícil a captação correta de correlações entre os rótulos. Segundo, devido à ausência da técnica de Ensemble. O método de *Ensemble of Classifier Chains* (ECC) faz o treinamento de várias cadeias em ordens diferentes. Tendo em vista que CC é altamente influenciado pela ordem dos classificadores na cadeia, ECC reduz esse efeito ao agregar predições de múltiplas cadeias.

É viável, também, fazermos suposições a respeito da maior robustez dos algoritmos BRkNN e MLARAM, quando comparados a CC. No algoritmo BRkNN, cada rótulo é tratado de forma independente. Isso implica que, se um erro for cometido em uma previsão de rótulo, este não afeta as previsões dos outros rótulos. Já MLARAM, com sua capacidade de aprender de forma robusta com dados de alta dimensionalidade e seu mapeamento de protótipos auto-organizáveis, é eficaz em lidar com a complexidade dos dados esparsos.

Já no caso da métrica Hamming Loss, Classifier Chains apresentou melhor desempenho quando comparado a outros algoritmos. Isso porque obteve valores menores, o que indica que o modelo tem menos erros de predição. Podemos atribuir este fato a sua capacidade de acertar mais verdadeiros negativos em problemas de CM. Esta se deve ao efeito sequencial de classificadores no CC, somado ao aprendizado de interações entre os rótulos.

6 CONCLUSÃO E PERSPECTIVAS FUTURAS

Neste trabalho, foi proposta a implementação de algoritmos de classificação multirrótulo para predição de doenças associadas a lncRNAs. Os conjuntos de dados foram construídos a partir de associações experimentalmente validadas, e estes foram utilizados para treinar diferentes algoritmos de Classificação Multirrótulo. O desempenho dos classificadores foi computado por meio de medidas de avaliação específicas e, para fins de comparação, testes estatísticos foram performados.

A análise de resultados dos testes permitiu inferir em que casos observaram-se diferenças no desempenho dos métodos. Nestes, pudemos formular hipóteses a respeito da performance do algoritmos. Comparações envolvendo Classifier Chains mostraram-se interessantes pois, embora o método tenha apresentado resultados inferiores em métricas como Precisão, Revocação e MedidaF, destacou-se positivamente na métrica de Hamming Loss. Entretanto, no problema abordado neste estudo, a identificação de associações corretas (valor 1) é de maior interesse, visto que o propósito é predizer as doenças associadas aos lncRNAs.

É importante salientar, também, que, de maneira geral, todos os algoritmos executados apresentaram resultados insatisfatórios, como evidenciado pelos valores das métricas. Isso pode ser justificado, sobretudo, pela predominância de valores nulos na matriz binária, o que dificulta substancialmente a aprendizagem pelos modelos. Ressalta-se, assim, a necessidade de investigar métodos mais robustos e eficazes para lidar com esses contextos, de modo a mitigar os efeitos da alta esparsidade dos dados.

Nesse cenário, alguns classificadores empregam técnicas como a fatoração matricial para preencher a matriz binária, visando mitigar os efeitos provocados pela elevada esparsidade dos dados. A técnica decompõe a matriz original binária em duas ou mais matrizes de menor dimensão, que representam os fatores latentes das amostras e das classes. Estes capturam características subjacentes às associações, permitindo identificar padrões não diretamente evidentes na matriz esparsa.

A vantagem do uso da fatoração matricial, neste caso, é o preenchimento de valores nulos, reconstruindo valores desconhecidos ou ausentes. Esses valores preenchidos representam predições probabilísticas sobre a existência de associações entre amostras (lncRNAs) e classes (doenças). Desta forma, valores originalmente nulos na matriz esparsa são transformados em estimativas contínuas que variam entre 0 e 1. Como resultado, obtém-se uma estrutura mais abrangente e completa em informações, impactando no aprimoramento do desempenho dos algoritmos de classificação.

No método proposto por Li et al (2019), a fatoração matricial foi usada para modelar a probabilidade de iteração de cada par lncRNA-doença. Neste trabalho, medidas de avaliação indicam melhor desempenho comparado a outros métodos estado-da-arte. Em (Ban et al. 2019), o novo modelo de previsão de interação droga-alvo utiliza também da fatoração matricial, havendo melhora da previsão para pares medicamento-alvo com menos informações de interação. Da mesma forma, em (Pliakos et al. 2020), a aplicação da fatoração matricial teve efeitos significativos no desempenho da previsão para pares medicamento-alvo. O trabalho forneceu ótimo desempenho preditivo enquanto foi computacionalmente eficiente e escalável.

Em síntese, a adaptação e implementação de tais técnicas, no contexto de lncRNAs e doenças, têm o potencial de melhorar a qualidade das predições, possibilitando avanços significativos na compreensão das interações biológicas. Assim, futuras investigações poderão focar na aplicação de novas abordagens que combinem eficiência computacional e capacidade preditiva robusta, contribuindo para o progresso em pesquisas bioinformáticas e para expansão do entendimento das bases moleculares de doenças.

REFERÊNCIAS

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. Machine Learning, v. 6, n. 1, p. 37–66, 1991.

AMODIO, N. MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. Journal of Hematology & Oncology, n. 11, p. 1–19, 2018.

BAN, T.; OHUE, M.; AKIYAMA, Y. NRLMFβ: Beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction. Biochemistry and biophysics reports, v. 18, n. 100615, p. 100615, 2019.

BAYES, T. An essay towards solving a problem in doctrine of chances. Philosophical Transactions of the Royal Society of London, v. 53, p. 293–315, 1763.

BASGALUPP, M. P.; CERRI, R.; SCHIETGAT, L.; TRUGUERO, I.; VENS, C. Beyond global and local multi-target learning. Information Sciences, v. 579, p. 508-524, 2021.

BELLO, M.; AGUILERA, Y.; NÁPOLES, G.; et al. Layer-Wise Relevance Propagation in Multi-label Neural Networks to Identify COVID-19 Associated Coinfections. Progress in Artificial Intelligence and Pattern Recognition. Lecture Notes in Computer Science, vol. 13055, Springer, 2021.

BENITES, F.; SAPOZHNIKOVA, E. HARAM: A hierarchical ARAM neural network for large-scale text classification. 2015 IEEE International Conference on Data Mining Workshop (ICDMW). Anais...IEEE, 2015.

BISWAS, A. K.; ZHANG, B.; WU, X.; et al. A multi-label classification framework to predict disease associations of long non-coding RNAs (lncRNAs). In: Lecture Notes in Electrical Engineering, vol. 322. Cham: Springer Verlag, p. 821–830, 2015.

BOUTELL, M. R.; LUO, J.; SHEN, X.; BROWN, C. M. Learning multi-label scene classification. Pattern Recognition, v. 37, n. 9, p. 1757–1771, 2004.

CARRIERI, C.; et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature, v. 491, p. 454–457, 2012.

CARVALHO, A.; FREITAS, A. A tutorial on multi-label classification techniques. Studies in Computational Intelligence, v. 205, p. 177–195, 2009.

CERRI, R.; SILVA, R.; CARVALHO, A. Comparing methods for multilabel classification of proteins using machine learning techniques. In: IV Brazilian Symposium on Bioinformatics. Lecture Notes in Bioinformatics, vol. 5676, p. 109–120, 2009.

CHERMAN, E. A.; METZ, J.; MONARD, M. C. Incorporating label dependency into the binary relevance framework for multi-label classification. Expert Systems with Applications, v. 39, n. 2, p. 1647–1655, 2012.

CLARE, A.; KING, R. D. Knowledge discovery in multi-label phenotype data. In: 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2001). LNAI, v. 2168, p. 42–53, 2001.

COHEN, W. W. Fast effective rule induction. In: Twelfth International Conference on Machine Learning, p. 115–123, 1995.

CONGRAINS, A.; KAMIDE, K.; OGURA, R.; et al. Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. Atherosclerosis, v. 220, p. 449–455, 2012.

DEMBCZYNSKI, K.; CHENG, W.; HÜLLERMEIER, E. Bayes optimal multilabel classification via probabilistic classifier chains. In: Proceedings of the 27th International Conference on Machine Learning, p. 279–286, 2010.

DEMSAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research, n. 7, p. 1–30, 2006.

FANG, Y.; FULLWOOD, M. J. Roles, functions, and mechanisms of long non-coding RNAs in cancer. Genomics, Proteomics and Bioinformatics, v. 14, n. 1, p. 42-54, 2016.

FENG, J.; BI, C.; CLARK, B. S. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev, v. 20, 2006.

FREUND, Y.; SCHAPIRE, R.; et al. A decision-theoretic generalization of on-line learning and an application to boosting. In: European Conference on Computational Learning Theory, p. 23–37, 1995.

GODBOLE, S.; SARAWAGI, S. Discriminative methods for multi-labeled classification. In: 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, p. 22–30, 2004.

GOYAL, B.; et al. Diagnostic, prognostic, and therapeutic significance of long non-coding RNA MALAT1 in cancer. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, v. 1875, n. 2, p. 188502, 2021.

GUPTA, R. A.; SHAH, N.; WANG, K. C.; et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature, v. 464, p. 1071–1076, 2010.

HAYKIN, S. Neural Networks: A Comprehensive Foundation. 2. ed. Upper Saddle River, NJ: Prentice Hall PTR, 1999.

HIROTA, K.; et al. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature, v. 456, p. 130–134, 2008.

HU, W.; ALVAREZ-DOMINGUEZ, J. R.; LODISH, H. F. Regulation of mammalian cell differentiation by long non-coding RNAs. EMBO Reports, v. 13, p. 971–983, 2012.

JOACHIMS, T. Text categorization with support vector machines: learning with many relevant features. In: Machine Learning: ECML-98. Lecture Notes in Computer Science, v. 1398, p. 137–142, 1998.

JOHNSON, R. Long non-coding RNAs in Huntington's disease neurodegeneration. Neurobiology of Disease, v. 46, p. 245–254, 2012.

KARAR, M. E.; HEMDAM, E. E. D.; SHOUMAN, M. A. Cascaded deep learning classifiers for computer-aided diagnosis of COVID-19 and pneumonia diseases in X-ray scans. Complex Intelligent Systems, v. 7, p. 235–247, 2021.

LI, Y.; LI, J.; BIAN, N. DNILMF-LDA: Prediction of lncRNA-disease associations by dual-network integrated logistic matrix factorization and Bayesian optimization. Genes, v. 10, n. 8, p. 608, 2019.

LIU, B.; LIU, F.; WANG, X.; CHEN, J.; FANG, L.; CHOU, K.-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Research, v. 43, p. W65–W71, 2015.

LIU, X.; LI, D.; ZHANG, W.; GUO, M.; ZHAN, Q. Long noncoding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. The EMBO Journal, v. 31, p. 4415–4427, 2012.

MADJAROV, G.; KOCEV, D.; GJORGJEVIKJ, D.; DZEROSKI, S. An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, v. 45, n. 9, p. 3084–3104, 2012.

MADHAVAN, M.; GOPAKUMAR, G. DBNLDA: Deep Belief Network based representation learning for lncRNA-disease association prediction. Applied Intelligence, v. 52, p. 5342–5352, 2022.

MERCER, T. R.; DINGER, M. E.; et al. Specific expression of long noncoding RNAs in the mouse brain. Proceedings of the National Academy of Sciences of the United States of America, v. 105, p. 716–721, 2008.

PASMANT, E.; SABBAGH, A.; VIDAUD, M.; et al. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB Journal, v. 25, p. 444–448, 2011.

PLIAKOS, K.; VENS, C. Drug-target interaction prediction with tree-ensemble learning and output space reconstruction. BMC Bioinformatics, v. 21, n. 1, art. no. 49, 2020. DOI: 10.1186/s12859-020-3379-z.

PONTING, C. P.; OLIVER, P. L.; REIK, W. Evolution and functions of long noncoding RNAs. Cell, v. 136, p. 629–641, 2009.

QUINLAN, J. R. C4.5: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 1993.

READ, J.; PFAHRINGER, B.; HOLMES, G.; FRANK, E. Classifier chains for multi-label classification. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II, ECML PKDD '09, p. 254–269, Berlin, Heidelberg: Springer-Verlag, 2009.

SCHAPIRE, R. E.; SINGER, Y. Improved Boosting Algorithms Using Confidence-rated Predictions. Machine Learning, v. 37, p. 297–336, 1999.

SCHAPIRE, R. E.; SINGER, Y. BoosTexter: A boosting-based system for text categorization. Machine Learning, v. 39, p. 135–168, 2000.

SHEN, X.; BOUTELL, M.; LUO, J.; BROWN, C. Multilabel machine learning and its application to semantic scene classification. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, v. 5307, p. 188–199, 2003.

SHENGCHANG; JIAQING; FENG, S. Prediction of lncRNA and disease associations based on residual graph convolutional networks with attention mechanism. Scientific Reports, 2024.

SPIZZO, R.; et al. Long non-coding RNAs and cancer: a new frontier of translational research? Oncogene, v. 31, p. 4577–4587, 2012.

SPYROMITROS, E.; TSOUMAKAS, G.; VLAHAVAS, I. An empirical study of lazy multilabel classification algorithms. Em: Artificial Intelligence: Theories, Models and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 401–406.

SZYMAŃSKI, P.; KAJDANOWICZ, T. A scikit-based Python environment for performing multi-label classification. arXiv preprint, arXiv:1702.01460, 2017.

THABTAH, F. A.; COWLING, P.; PENG, Y.; et al. MMAC: A new multi-class, multi-label associative classification approach. In: Fourth IEEE International Conference on Data Mining, p. 217–224, 2004.

TROHIDIS, K.; TSOUMAKAS, G.; KALLIRIS, G.; VLAHAVAS, I. Multilabel classification of music into emotions. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, USA, 2008.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. In: Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications, p. 64–74, 2008.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. P. Mining Multi-label Data. In: MAIMON, O.; ROKACH, L. (Ed.). Data Mining and Knowledge Discovery Handbook. 2. ed. p. 667–685, 2010.

VAPNIK, V. N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

VAPNIK, V. N. The Nature of Statistical Learning Theory. Information Science and Statistics. New York: Springer-Verlag, 1999.

VENS, C.; STRUYF, J.; SCHIETGAT, L.; DZEROSKI, S.; BLOCKEEL, H. Decision trees for hierarchical multi-label classification. Machine Learning, v. 73, n. 2, p. 185–214, 2008. DOI: 10.1007/s10994-008-5077-3.

WANG, W.; DAI, Q.; LI, F.; et al. MLCDForest: multi-label classification with deep forest in disease prediction for long non-coding RNAs. Briefings in Bioinformatics, v. 22, 2021.

WEI, H.; LIAO, Q.; LIU, B. iLncRNAdis-FB: identify lncRNAdisease associations by fusing biological feature blocks through deep neural network. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

XU, L.; JIAO, S.; et al. Identification of long noncoding RNAs with machine learning methods: a review. Briefings in Functional Genomics, v. 20, p. 174–180, 2021.

YAO, D. et al. A random forest based computational model for predicting novel lncRNA-disease associations. BMC bioinformatics, v. 21, n. 1, p. 126, 2020.

YUAN, L. Long non-coding RNAs towards precision medicine in gastric cancer: early diagnosis, treatment, and drug resistance. Molecular Cancer, p. 1–22, 2020.

ZHANG, M.-L.; ZHOU, Z.-H. A k-Nearest Neighbor Based Algorithm for Multi-label Classification. IEEE Computational Intelligence Society, v. 2, p. 718–721, 2005.

ZHANG, M.-L.; ZHOU, Z.-H. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. IEEE Transactions on Knowledge and Data Engineering, v. 18, p. 1338–1351, 2006.

ZHANG, M.-L.; ZHOU, Z.-H. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, v. 40, p. 2038–2048, 2007.

ZHANG, Q.; CHEN, C.-Y.; YEDAVALLI, V. S. R. K.; et al. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. Journal of Virology, v. 87, p. 596–612, 2013.

ZHU, B.; POON, C. K. Efficient approximation algorithms for multi-label map labeling. In: International Symposium on Algorithms and Computation. Berlin, Heidelberg: Springer, p. 143–152, 1999.

ZHU, S.; JI, X.; XU, W.; GONG, Y. Multi-labelled classification using maximum entropy method. In: International Conference on Research and Development in Information Retrieval. p. 274–281, 2005.

APÊNDICE

Lista de doenças presentes no conjunto de dados utilizado neste estudo

AIDS
Doença de Alzheimer
Neoplasias de células B
Síndrome de Beckwith-Wiedemann
Linfoma de Burkitt
Distrofia muscular de Duchenne
HIV
Doença de Huntington
Doença de Parkinson
Síndrome de Prader-Willi
Síndrome de Silver-Russell
Estados patológicos associados à TDP-43
Tumor de Wilms
Leucemia linfoblástica aguda
Leucemia mieloide aguda
Infarto agudo do miocárdio
Leucemia promielocítica aguda
Escoliose idiopática do adolescente
Envelhecimento
Esclerose lateral amiotrófica
Aneurisma da aorta
Astrocitoma
Aterosclerose
Doença autoimune
Carcinoma basocelular
Transtorno bipolar
Câncer de bexiga
Síndrome de blefarofimose
Câncer de mama
Câncer
Hipertrofia cardíaca
Doença cardiovascular
Câncer do colo do útero
Coriocarcinoma
Leucemia linfocítica crônica
Leucemia mieloide crônica
Câncer de cólon
Câncer colorretal
Doença arterial coronariana
Depressão
Diabetes mellitus
Cardiomiopatia diabética
Retinopatia diabética
Linfoma difuso de grandes células B
Cardiomiopatia dilatada

Abuso de drogas
Carcinoma embrionário
Câncer endometrial
Endometriose
Câncer epitelial de ovário
Adenocarcinoma esofágico
Câncer de esôfago
Carcinoma espinocelular do esôfago
Câncer da vesícula biliar
Adenocarcinoma gástrico
Câncer gástrico
Adenocarcinoma da cárdia gástrica
Glioblastoma
Glioma
Carcinoma espinocelular de cabeça e pescoço
Insuficiência cardíaca
Neoplasias hematológicas
Carcinoma hepatocelular
Telangiectasia hemorrágica hereditária
Hipertensão
Infertilidade
Degeneração do disco intervertebral
Aneurisma intracraniano
Lesão por isquemia-reperfusão
Insuficiência cardíaca isquêmica
Acidente vascular cerebral isquêmico
Câncer renal
Carcinoma espinocelular de laringe
Leucemia
Câncer de fígado
Fibrose hepática
Adenocarcinoma pulmonar
Câncer de pulmão
Carcinoma espinocelular de pulmão
Linfoma
Melanoma
Meningioma
Mieloma múltiplo
Esclerose múltipla
Síndrome mielodisplásica
Leucemia mieloide
Infarto do miocárdio
Carcinoma nasofaríngeo
Neuroblastoma
Câncer de pulmão de células não pequenas
Obesidade
Carcinoma espinocelular do esôfago
Carcinoma espinocelular oral

Osteoartrite
Osteossarcoma
Câncer de ovário
Câncer de pâncreas
Adenocarcinoma ductal pancreático
Carcinoma papilífero da tireoide
Adenoma hipofisário
Mesotelioma pleural
Síndrome do ovário policístico
Pré-eclâmpsia
Síndrome de Sjögren primária
Mielofibrose primária
Leucemia promielocítica
Câncer de próstata
Doença psiquiátrica
Carcinoma renal
Carcinoma de células renais
Carcinoma de células claras renais
Neurodegeneração da retina
Retinoblastoma
Artrite reumatoide
Esquizofrenia
Melanoma cutâneo
Câncer de pulmão de pequenas células
Carcinoma espinocelular
Adenocarcinoma gástrico
Acidente vascular cerebral (AVC)
Lúpus eritematoso sistêmico
Câncer testicular
Câncer de tireoide
Carcinoma espinocelular da língua
Diabetes mellitus tipo 2
Câncer urotelial
Melanoma uveal
Defeitos do septo ventricular
Carcinoma espinocelular da vulva