Victor Vidal dos Santos

Avaliação de Métodos de Imputação de Valores Ausentes para a Predição de Interações Fármaco-Proteína

Recife

2024

Victor Vidal dos Santos

Avaliação de Métodos de Imputação de Valores Ausentes para a Predição de Interações Fármaco-Proteína

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação. Este trabalho tem como objetivo avaliar técnicas de tratamento de valores faltosos em redes bipartidas, com foco na predição de interações fármaco-proteína, visando contribuir para o avanço no campo da descoberta de medicamentos.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: André Camara Alves do Nascimento

Recife

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

S237a Santos, Victor Vidal dos
Avaliação de Métodos de Imputação de Valores Ausentes para a Predição de Interações Fármaco Proteína / Victor Vidal dos Santos. - 2024.
26 f. : il.

Orientadora: Andre Camara Alves do Nascimento Recife.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2024.

1. Aprendizagem de múltiplos kernels. 2. Redes bipartidas. 3. PairwiseMKL.. I. Recife, Andre Camara Alves do Nascimento, orient. II. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Victor Vidal dos Santos às 11 horas do dia 08 de março de 2024, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado **On the impact of missing value imputation methods for multiple kernel learning on bipartite graphs**, orientado por André Câmara Alves do Nascimento e aprovado pela seguinte banca examinadora:

André Câmara Alves do Nascimento
DC/UFRPE

Ruan Vasconcelos Bezerra Carvalho
DC/UFRPE

Dedico este trabalho a todas as pessoas que me inspiraram, apoiaram e incentivaram ao longo da minha jornada acadêmica. Seu amor, apoio e encorajamento foram fundamentais para alcançar este marco. Este trabalho é dedicado a vocês com profunda gratidão e carinho.

Agradecimentos

Gostaria de expressar meus sinceros agradecimentos a todos que contribuíram para a realização deste trabalho. Em primeiro lugar, agradeço à minha família pelo apoio incondicional e pelo incentivo ao longo deste processo. Agradeço aos meus orientadores pelo valioso suporte e orientação ao longo deste projeto. Também gostaria de agradecer aos meus colegas e amigos que me ajudaram e me motivaram durante este percurso acadêmico. Muito obrigado a todos.

“A persistência é o caminho do êxito.”
(Charles Chaplin)

Abstract. Na última década, o estudo de redes farmacológicas tem recebido muita atenção devido à sua relevância no processo de descoberta de medicamentos. Muitas abordagens diferentes para prever interações biológicas têm sido propostas, especialmente na área de aprendizado de múltiplos kernels (MKL). Tais métodos compreendem abordagens integrativas que podem lidar com fontes de dados heterogêneas, mas sofrem com o problema de dados incompletos. Técnicas para lidar com valores faltosos nas matrizes kernel base podem ser utilizadas, geralmente baseadas em técnicas simples, como imputação de zeros, média e mediana da matriz. Neste trabalho, foram avaliadas técnicas de tratamento de valores faltosos no contexto de redes bipartidas. Nossas análises mostraram que, dependendo da quantidade de dados faltantes, a técnica k-NN e SVD teve um desempenho muito melhor do que as outras técnicas, trazendo resultados animadores, enquanto o preenchimento zero apresentou o pior desempenho em relação a todos os outros métodos avaliados.
Palavras-chave: aprendizagem de múltiplos kernels, redes bipartidas, pairwiseMKL.

Abstract. In the last decade, the study of pharmacological networks has received a lot of attention given its relevance drug discovery process. Many different approaches for predicting biological interactions have been proposed, especially in the area of multiple kernel learning (MKL). Such methods comprise integrative approaches that can handle heterogeneous data sources, but suffer from the missing data problem. Techniques to handle missing values in the base kernel matrices can be used, usually based on simple techniques, such as imputing zeroes, mean and median of the matrix. In this work, techniques for handling missing values were evaluated in the context of bipartite networks. Our analyzes showed that the, depending on the amount of missing data, k-NN and SVD technique performed much better than the other techniques, bringing encouraging results, while zero-fill showed the worst performance in relation to all other evaluated methods.

Keywords: Multiple kernel Learning, Bipartite Networks, pairwiseMKL.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

SVD	Decomposição de Valor Único
KNN	K-Vizinhos Mais Próximos
MKL	Aprendizado de Kernel Múltiplo
GDSC	Genômica da Sensibilidade a Drogas no Câncer
iSVD	Decomposição de Valor Único de Baixo Rank Iterativa
RSME	Erro Médio Quadrático da Raiz (Root Mean Squared Error)

Sumário

1 Introdução

Com o crescimento constante e o envelhecimento da população mundial, grandes desafios de saúde, como o combate a vários tipos de câncer e doenças infecciosas, diabetes e doenças neurodegenerativas, estão em grande necessidade de inovações. Apesar desse contexto, o desenvolvimento rápido e econômico de novos medicamentos está longe de atender a essa demanda (CSERMELY TAMáS KORCSMáROS, 2013). O ritmo lento no desenvolvimento de medicamentos é devido à grande quantidade de riscos envolvidos, e esses riscos acabam causando um excesso de cautela na indústria farmacêutica. (CHONG; SULLIVAN, 2007)

A análise da evolução dos padrões de relação estrutura-atividade e topologia das redes droga-alvo exibiu uma tendência comportamental em que mais de 80% dos novos medicamentos tendem a se ligar a alvos que também estão conectados a outros medicamentos nas redes biológicas (COKOL IVAN IOSSIFOV, 2005). Assim, uma excelente maneira de mitigar os riscos associados ao desenvolvimento de novos medicamentos é utilizar o conhecimento previamente adquirido. Nesse contexto, é possível compreender a notoriedade que as redes droga-proteína têm recebido nos últimos anos (NASCIMENTO; COSTA, 2016). No entanto, as técnicas que utilizam essas redes sofrem em termos de viabilidade quando há dados ausentes (RIVERO R LEMENCE, 2017). Múltiplos fatores contribuem para os valores ausentes em dados biológicos, incluindo fatores experimentais, limitações de equipamento de laboratório ou o alto custo da aquisição de dados (JIN et al., 2021).

Na ciência de dados, a técnica mais simples usada para tratar dados ausentes em casos em que há muitos dados disponíveis é simplesmente a remoção da instância. No entanto, quando os dados ausentes são excluídos, o tamanho do espaço amostral é reduzido e pode haver uma perda considerável de poder estatístico. Outras técnicas conhecidas na literatura, como imputação com zero e média, também têm suas desvantagens (RIVERO R LEMENCE, 2017). Portanto, escolher a técnica certa é essencial.

Diante desse contexto, este trabalho propõe uma análise sistemática do efeito que as técnicas de imputação de valores ausentes mais simples, como zero, média e mediana, assim como técnicas mais complexas, como SVD (Decomposição de Valor Único) e imputação por KNN (K-Nearest Neighbor), têm no desempenho de métodos de previsão de interações entre drogas e alvos. Mais especificamente, os efeitos da imputação serão analisados em um algoritmo de aprendizado baseado em kernel.

Tais métodos têm sido bem-sucedidos em vários problemas de aprendizado supervisionado e são algoritmos que realizam classificação de padrões e conseguem agregar conhecimento prévio com base em funções de similaridade, ou simplesmente, kernels (NASCIMENTO; COSTA, 2016).

Entre os métodos de kernel existentes, um método conhecido como MKL (Aprendizado de Kernel Múltiplo) foi escolhido. O MKL combina kernels de múltiplas fontes com uma abordagem orientada a dados, o que torna possível usar diferentes noções de similaridade e melhorar a precisão (GONEN, 2011; AIOLLI, 2015). O algoritmo escolhido para este trabalho foi o pairwiseMKL, originalmente proposto por (CICHONSKA T. PAHIKKALA; ROUSU, 2018). Este algoritmo tem um desempenho melhor em comparação com métodos MKL tradicionais, pois sua etapa de aprendizado é realizada sem o cálculo explícito de matrizes em pares.

2 Trabalhos relacionados

Os métodos de kernel na biologia computacional têm um grande potencial para facilitar a integração de dados de uma miríade de fontes heterogêneas. No entanto, as informações contidas nesses bancos de dados biológicos frequentemente são incompletas ou até mesmo ausentes para algumas entidades. Algumas soluções comumente adotadas incluem a remoção de instâncias cuja informação não está completa, o que leva a uma diminuição no conjunto de dados e, consequentemente, no poder preditivo dessa amostra. Alguns estudos recentes que utilizam técnicas de MKL em problemas unipartidos investiram em complementar os valores ausentes em matrizes de kernel.

De acordo com (KUMAR et al., 2013), o problema de derivar uma matriz de kernel a partir de um conjunto de matrizes incompletas pode ser contornado preenchendo os valores ausentes. O preenchimento pode ser feito com a média ou simplesmente colocando zeros nas linhas e colunas da matriz (NASCIMENTO; COSTA, 2016; RIVERO R LEMENCE, 2017). Dado o exposto, o tratamento de dados ausentes em matrizes de kernel pode ser muito melhorado considerando os avanços recentes na pesquisa sobre métodos de imputação de valores ausentes em problemas de MKL (LIU, 2020; KUMAR et al., 2013).

3 Metodologia

Nas subseções seguintes, descrevemos os conjuntos de dados utilizados, bem como a metodologia experimental empregada.

3.1. Afinidades de interação

O experimento utilizou uma base de dados de resposta a drogas anticâncer do projeto GDSC (Genomics of Drug Sensitivity in Cancer) proposto por (YANG, 2012). A base é constituída pelas respostas de 124 linhagens de células cancerosas humanas a 124 drogas, assim, estão disponíveis 15.376 medidas de sensibilidade na forma de $ln(IC50)$, em valores de nanomolar (AMMAD-UD-DIN, 2016).

No histograma da distribuição dos valores de bioatividade da base utilizada, mostrado na Figura 1, é possível observar que os dados seguem uma distribuição normal, onde a maior concentração de dados está na faixa de afinidade de 0 a 5.

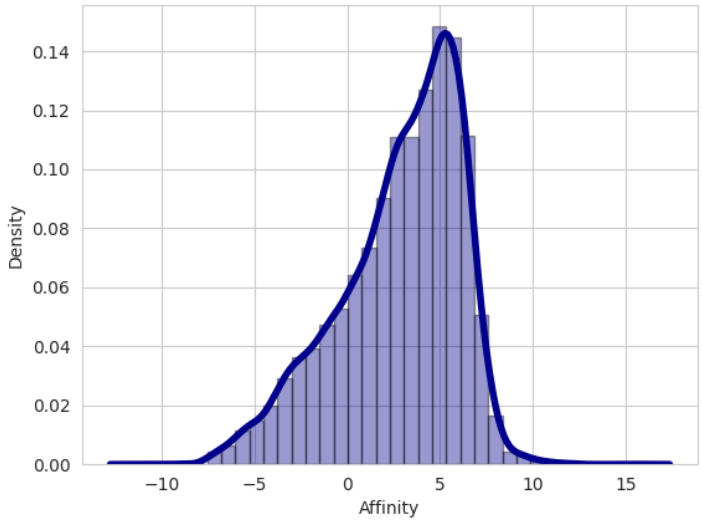

Figura 1 – Histograma dos valores de afinidade de interação.

3.2. Kernels

Nas seções seguintes, os processos de construção dos kernels serão apresentados, bem como duas tabelas com representações resumidas dos cálculos dos kernels de drogas e linhagens celulares usados no modelo pairwiseMKL.

Nome	Descrição da característica e tipo de kernel
Kd-circular	Conectividade Estendida de 1024 bits impressão digital com diâmetro máximo definido como 6 (ECFP6).
Kd-estate	Impressão digital de 79 bits correspondente ao 'Estate' subestruturas descritas por Hall e Kier (1995).
Kd-ext	Impressão digital de bloco de 1024 bits baseada em caminho, levando em consideração sistemas de anel.
Kd-graph	Impressão digital de bloco de 1024 bits baseada em caminho, considerando conectividade.
Kd-hybr	Impressão digital de bloco de 1024 bits baseada em caminho, considerando estados de hibridização.
Kd-kr	Impressão digital de 4860 bits definida por Klekota e Roth (2008).
Kd-maccs	Impressão digital de 166 bits baseada em chaves estruturais MACCS. desenvolvido pela MDL Information Systems.
Kd-PubCh	Impressão digital de 881 bits definida pelo PubChem.
Kd-sp	Impressão digital de 1024 bits baseada nos caminhos mais curtos entre átomos. levando em consideração sistemas de anéis e cargas.
Kd-std	Impressão digital de bloco de 1024 bits baseada em caminho.

Tabela 1 – Breves descrições dos kernels de drogas utilizados no modelo de predição pairwiseMKL. (Fonte: (CICHONSKA T. PAHIKKALA; ROUSU, 2018))

3.2.1. Kernels de droga

A construção dos kernels de drogas foi baseada em uma ferramenta usada para descrever a similaridade entre conjuntos de atributos binários conhecida como kernel de Tanimoto (SZEDMAK, 2020). Foram calculados 10 vetores binários que indicam a presença ou ausência de diferentes subestruturas na molécula. A tabela contendo representações resumidas do cálculo de cada uma das 10 impressões digitais pode ser vista na Tabela 1.

3.2.2. Kernels de linhas celulares

A construção dos kernels de linhagem celular foi baseada no cálculo de kernels gaussianos. (CICHONSKA T. PAHIKKALA; ROUSU, 2018) A tabela contendo representações resumidas do cálculo do kernel gaussiano para cada largura de hiperparâmetro pode ser vista na Tabela 2.

3.3. Técnicas Avaliadas

Neste trabalho, três técnicas de imputação de valor único (zero, média e mediana) e duas técnicas de imputação supervisionada (iSVD e KNN com K=3) foram avaliadas em um conjunto de dados com 10 matrizes de drogas e 12 matrizes de linhagens celulares. Anteriormente, esses conjuntos de dados foram pré-processados, gerando posições ausentes aleatoriamente para posterior imputação com as técni-

Name	Feature description and kernel type
Kc-cn-146	Medições de variação do número de cópias de 43.255 genes, com hiperparâmetro $(\sigma) = 146 delarguradokernelgaussiano.$
Kc-cn-270	Medições de variação do número de cópias de 43.255 genes, com hiperparâmetro $(\sigma) = 270 delarguradokernelgaussiano.$
Kc-cn-417	Medições de variação do número de cópias de 43.255 genes, com hiperparâmetro $(\sigma) = 417 delarguradokernelgaussiano.$
Kc-exp-147	Medições de expressão gênica basal de 13.321 genes, com hiperparâmetro $(\sigma) = 147 delarguradokernelgaussiano.$
Kc-exp-163	Medições de expressão gênica basal de 13.321 genes, com hiperparâmetro $(\sigma) = 163 delarguradokernelgaussiano.$
Kc-exp-177	Medições de expressão gênica basal de 13.321 genes, com hiperparâmetro $(\sigma) = 177 delarguradokernelgaussiano.$
Kc-met-176	Níveis de metilação de 482.892 ilhas de CpG, com hiperparâmetro $(\sigma) = 176 delarguradokernelgaussiano.$
Kc-met-210	Níveis de metilação de 482.892 ilhas de CpG, com hiperparâmetro $(\sigma) = 210 delarguradokernelgaussiano.$
Kc-met-252	Níveis de metilação de 482.892 ilhas de CpG, com hiperparâmetro $(\sigma) = 252 delarguradokernelgaussiano.$
Kc-mut-57	Perfis de valores reais de 12.366 mutações somáticas, com hiperparâmetro $(\sigma) = 57 delarguradokernelgaussiano.$
Kc-mut-71	Perfis de valores reais de 12.366 mutações somáticas, com hiperparâmetro $(\sigma) = 71 delarguradokernelgaussiano.$
Kc-mut-132	Perfis de valores reais de 12.366 mutações somáticas, com hiperparâmetro $(\sigma) = 132 delarguradokernelgaussiano.$

Tabela 2 – Breves descrições do kernel de linhagem celular usado no modelo de predição pairwiseMKL. (Fonte: (CICHONSKA T. PAHIKKALA; ROUSU, 2018))

cas escolhidas. A escolha das técnicas considerou o alto uso em outros estudos e resultados validados na literatura.

A primeira técnica simples de imputação consiste em preencher os valores ausentes com o número zero (TUIKKALA, 2008). Na imputação pela média, a média de cada matriz foi calculada, e cada resultado foi usado para preencher os valores ausentes. Na última técnica de imputação de valor único, foi usada a mediana, que tem um processo semelhante à imputação pela média, mas, neste caso, é usada a mediana da matriz (WEI, 2018).

Para as duas técnicas de imputação supervisionada, foram utilizadas as implementações oferecidas por (RUBINSTEYN; FELDMAN,). A primeira técnica usada foi a iSVD (Decomposição de Valor Único de baixo rank iterativa), cuja implementação foi descrita por (TROYANSKAYA M. CANTOR; ALTMAN1, 2001). A segunda e última técnica de imputação supervisionada utilizada foi o KNN, que consiste em ponderar amostras usando a diferença média ao quadrado nas características para as quais duas linhas têm dados observados.

3.4. Métricas de Avaliação

Para este trabalho, três técnicas de avaliação foram escolhidas. A primeira métrica escolhida foi o F1-score, que corresponde a uma medida de precisão de um modelo em um conjunto de dados, e é comumente usado na avaliação de sistemas de classificação binária. É definido como a média harmônica da sensibilidade (recall) e precisão do modelo (DALIANIS, 2018). A precisão é descrita como:

$$Precision = \frac{TP}{TP + FP},$$

e mede o número de instâncias corretas recuperadas dividido por todas as instâncias recuperadas (DALIANIS, 2018). A sensibilidade (recall) mede o número de instâncias corretas recuperadas dividido por todas as instâncias corretas e é definida matematicamente como:

$$Recall = \frac{TP}{TP + FN}$$

Portanto, o F1-score é definido como:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

A segunda foi o coeficiente de correlação de Pearson (r), que corresponde ao grau de associação linear entre duas variáveis quantitativas (LIU, 2020). A análise de correlação, de forma geral, se inicia com a representação gráfica da relação dos pares de dados através do uso de um diagrama de dispersão. O coeficiente de correlação de Pearson pode ser definido como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

O coeficiente corresponde a um índice adimensional e os valores variam de -1 a +1, refletindo a intensidade de uma relação linear entre dois conjuntos de dados. Ou seja, valores positivos indicam uma tendência de uma variável aumenta ou diminuir em conjunto com outra, e valores negativos indicam uma tendência de que o aumento dos valores de uma variável esteja à redução de outra. Valores próximos de zero indicam baixa associação. (KIRCH, 2008)

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

A última métrica escolhida foi o erro de raiz quadrática médio (RSME - Root Mean Squared Error), que corresponde a raiz quadrada da média do quadrado de todos os erros, onde, Oi determinam as observações, Si os valores previstos de uma variável e n o número de observações disponíveis para análise. O RMSE é bastante utilizado e considerado uma ótima métrica de erro de propósito geral para previsões numéricas. (NEILL, 2018)

3.5. Descrição do experimento

Com o intuito de avaliar o desempenho do algoritmo pairwiseMKL proposto por (CICHONSKA T. PAHIKKALA; ROUSU, 2018) no contexto de redes bipartidas, foi realizado um experimento sistemático visando avaliar a eficácia do método quando se é utilizado um conjunto de dados biológicos heterogêneos incompletos como entrada. Após uma análise dos métodos e técnicas apresentadas na literatura foi desenvolvido um experimento, o mesmo pode ser descrito em 3 fases: geração dos dados ausentes, imputação e treinamento/predição do modelo.

A primeira fase do experimento consistiu na geração de dados ausentes nas matrizes de kernels descritas anteriormente. Para isso foi implementado um algoritmo que recebeu como entrada uma matriz de kernel completa e uma porcentagem de valores faltosos que deveria ser gerada, em seguida o algoritmo substituiu valores selecionados randomicamente das matrizes por NaN até que a proporção desejada fosse atingida. Vale salientar que as propriedades inerentes as matrizes de kernel foram mantidas após a execução do algoritmo. O algoritmo descrito foi executado para cada uma das 22 matrizes de kernel existentes na base de dados e cada uma das porcentagens de valores ausentes escolhidas (10%, 30%, 50% e 70%) totalizando um total de 88 execuções. As matrizes incompletas geradas foram então armazenadas para uso posterior na próxima fase do experimento.

Após a geração das matrizes de kernel incompletas foi realizada a etapa de imputação. Para esta etapa foi implementado um algoritmo que recebeu como entrada uma matriz de kernel com valores ausentes e uma técnica de imputação, em seguida o algoritmo realizou a imputação dos dados ausenta na matriz obtida de acordo com a técnica recebida. O algoritmo em questão foi executado para cada uma das 88 matrizes geradas na fase anterior e então as matrizes de kernel imputadas foram armazenadas para uso. O produto resultante do processo de imputação da etapa anterior foram 22 matrizes (10 kernels de drogas e 12 de linhas celulares) imputadas para cada uma das 4 escolhas de porcentagem de dados ausentes (10%, 30%, 50% e 70%), utilizando cada um dos 5 métodos de imputação abordados (zero, média, mediana, iSVD e KNN), totalizando 22 x 4 x 5 = 440 kernels imputados.

Em seguida foi realizado o processo de treinamento e predição utilizando o algoritmo original proposto por (CICHONSKA T. PAHIKKALA; ROUSU, 2018), com uma pequena modificação. A quantidade de inner folds utilizadas no processo de validação cruzada foi reduzida de 3 para 1. Esta modificação foi realizada visando diminuir o tempo de execução do algoritmo. O processo de treinamento e validação cruzada foi executado separadamente em cada conjunto de 22 kernels imputados utilizando cada uma das técnicas escolhidas e porcentagem de dados ausentes.

O resultado do processo descrito previamente foram 3 arquivos de texto, um para cada métrica avaliativa escolhida (F1-score, Pearson e RSME), nos quais cada linha representa o valor da métrica em questão para cada outer fold da validação cruzada executada na combinação técnica-porcentagem em questão. Sendo assim, essa etapa resultou em 3x4x5=60 arquivos de métrica.

4 Resultados e discussão

Nesta seção serão discutidos os resultados presentes nos arquivos de métrica resultantes da última etapa do experimento descrito na seção anterior.

Inicialmente, devemos partir do caso base, ou seja, as métricas provenientes da execução do algoritmo modificado com os kernels completos. Os resultados obtidos no cenário original podem ser utilizados como modelo comparativo para a avaliação da perfomance das técnicas empregadas. Quanto mais próximo das métricas originais, melhor é classificada a técnica utilizada no pairwiseMLK. A tabela 3 contendo as métricas originais pode ser visualizada a seguir.

Cenário original - PairwiseMKL		
F1-score	Pearson	RSME
0.630376	0.857668	1.6816

Tabela 3 – Métricas de desempenho do cenário original

Na tabela 4, podemos ver os valores das métricas resultantes de cada combinação de técnica e porcentagem de valores ausentes. A primeira técnica de imputação de valor único, imputação por zero, apresentou o menor F1-score e o valor do coeficiente de Pearson em todos os cenários de porcentagem de valores ausentes. A imputação por zero também obteve o maior RMSE em todas as iterações. Também é possível observar que as técnicas de imputação pela média e mediana obtiveram resultados semelhantes em todas as iterações, mas é interessante observar que essas duas técnicas apresentaram valores ligeiramente superiores às técnicas de imputação supervisionada quando os valores ausentes corresponderam a 70% dos valores da matriz de kernel. Considerando o fato de que em todos os outros cenários de dados ausentes (10%, 30% e 50%), métodos de imputação mais complexos alcançaram melhores resultados, os resultados ligeiramente melhores obtidos pelo método de imputação mediana no cenário de 70% podem ser devido a efeitos aleatórios, ou mesmo pela maior esparsidade nos dados de treinamento. Mais discussões sobre como esses efeitos podem ser mitigados são discutidas na Seção 5.

Técnica	Porcentagem	F1-score	Pearson	RMSE
Zero	10%	0.553766	0.726793	2.45866
Mean		0.589447	0.773677	2.16993
Median		0.593042	0.779026	2.1413
iSVD		0.617436	0.787961	2.09885
KNN		0.625912	0.847979	1.73262
Zero	30%	0.533408	0.647054	2.9584
Mean		0.596227	0.773153	2.17983
Median		0.598501	0.774271	2.17475
iSVD		0.613883	0.819591	1.90926
KNN		0.608417	0.812906	1.9292
Zero	50%	0.507404	0.575169	3.52906
Mean		0.582878	0.749684	2.31245
Median		0.578097	0.749635	2.31263
iSVD		0.597902	0.781713	2.11929
KNN		0.555198	0.658958	2.85667
Zero	70%	0.490236	0.517869	4.08531
Mean		0.586777	0.759897	2.25142
Median		0.593031	0.76949	2.19618
iSVD		0.577555	0.759954	2.24384
KNN		0.565799	0.721587	2.47796

Tabela 4 – Análise comparativa das métricas para cada combinação técnica-percentual

Em seguida, é possível notar que a técnica de imputação supervisionada iSVD (iterative low-rank Single-Value Decomposition) apresentou valores melhores que a outra técnica supervisiona (KNN) quando o percentual de valores faltosos eram 30% e 50%. No entanto, é curioso notar que a técnica KNN (K=3) apresentou resultados consideravelmente melhores que o iSVD na primeira iteração, porém houve uma degradação gradativa em sua performance conforme o percentual de números faltosos aumentava. Tal comportamento pode se dar devido ao número baixo do parâmetro K escolhido. A evolução de cada métrica de avaliação em diferentes cenários de valores ausentes são apresentadas nas Figuras 2, 3 e 4. A seguir é possível observar os gráficos que demonstram a evolução dos valores das métricas de acordo com o aumento da porcentagem de dados ausentes. É possível constantar que a segunda maior degradação pertence a técnica supervisionada KNN (K=3)

Para avaliar a relevância estatística dos resultados, os resultados para cada métrica de avaliação em todas as dobras foram analisados. Para simplificar, iremos focar nossa análise da métrica F1 no cenário de 70%. Dado que os resultados seguiram uma distribuição normal e demonstraram ser homocedásticos (de acordo com o teste de Bartlett), utilizamos a ANOVA de medidas repetidas como teste omnibus para determinar se há diferenças significativas entre os valores médios das popula-

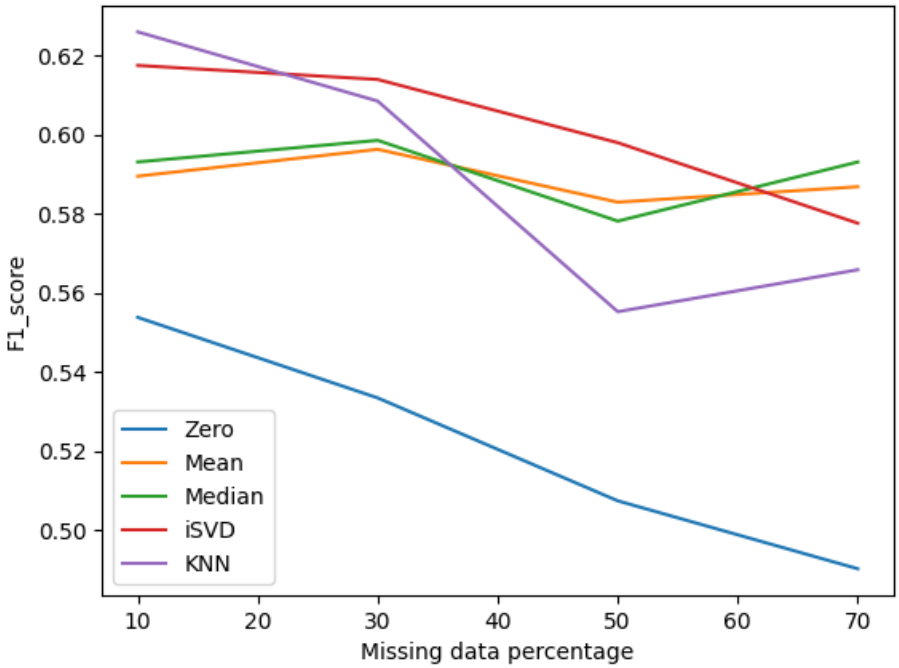

Figura 2 – F1-Score em diferentes proporções de dados ausentes.

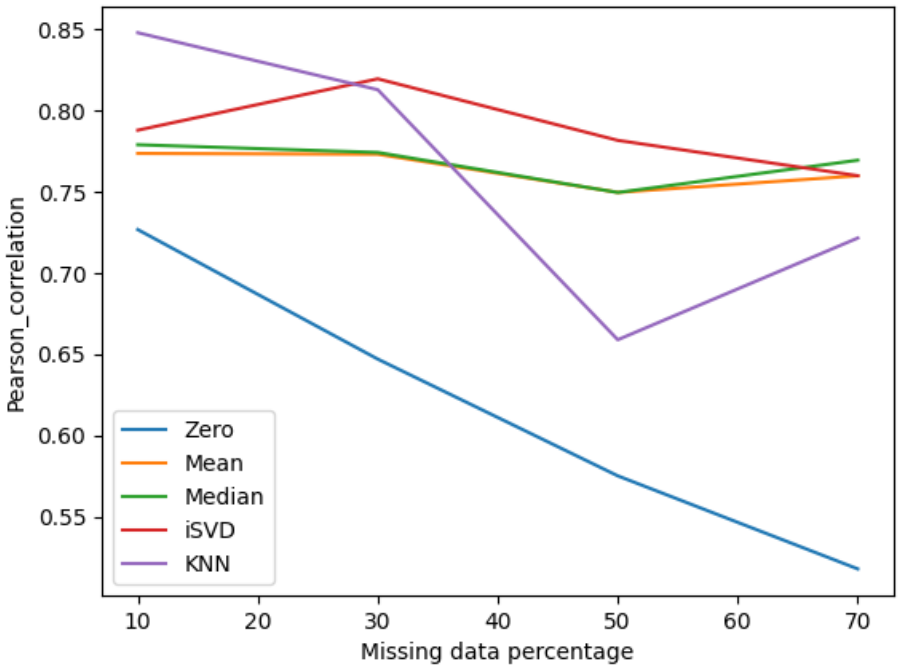

Figura 3 – Correlação de Pearson em diferentes proporções de dados ausentes.

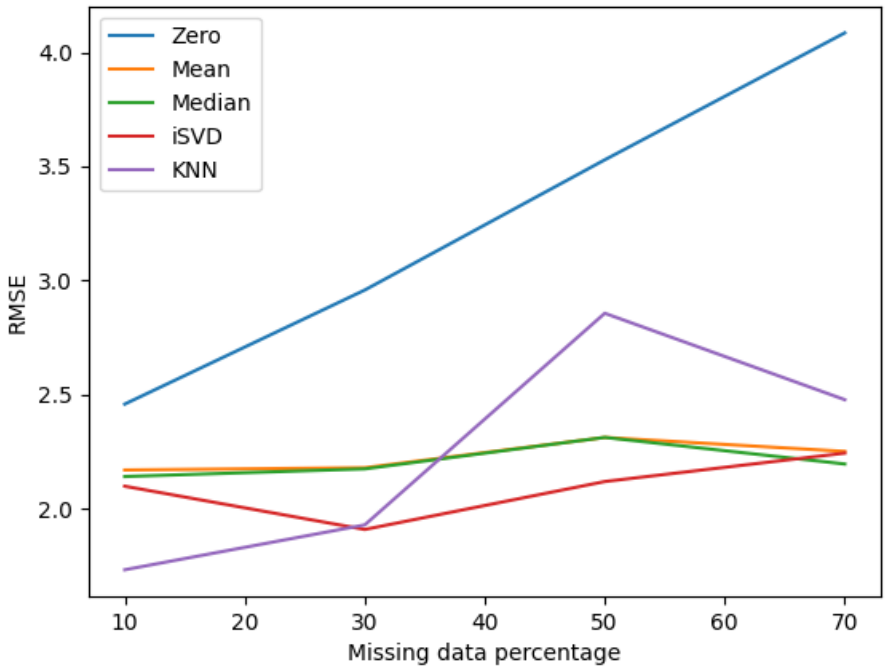

Figura 4 – RMSE em diferentes proporções de dados ausentes.

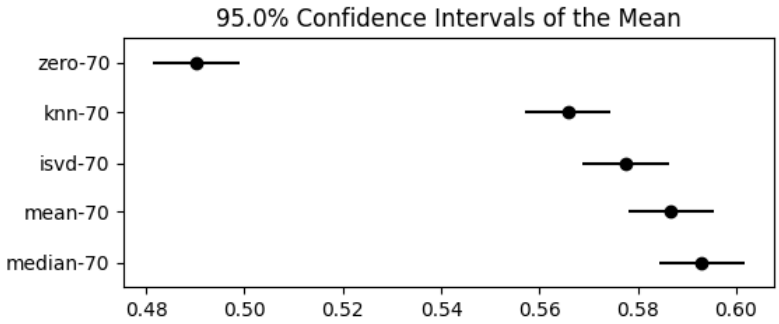

Figura 5 – Intervalos de confiança dos resultados médios para a métrica F1-score no cenário de 70% de valores ausentes.

ções. Com base no teste post-hoc de Tukey HSD, assumimos que não há diferenças significativas dentro dos seguintes grupos: média-70 e mediana-70. Todas as outras diferenças são significativas. A Figura 5 apresenta os intervalos de confiança dos resultados médios para cada método. Os resultados são significativamente diferentes se os intervalos de confiança não se sobrepuserem.

5 Conclusão

A pesquisa apresentada neste artigo tem como objetivo estudar o uso de diferentes técnicas para a imputação de valores ausentes no contexto em que métodos de kernel são usados para prever interações entre drogas e proteínas. Através do experimento realizado, foi possível obter um conhecimento mais profundo sobre os efeitos que o aumento na porcentagem de valores ausentes tem sobre cada técnica aplicada. Os resultados mostraram que as técnicas de imputação supervisionada têm melhor desempenho do que as técnicas de imputação de valor único quando a porcentagem de dados ausentes é baixa, mas também foi possível observar que técnicas mais simples, como imputação pela média e mediana, podem ser desejáveis em casos em que a porcentagem é alta.

Essas observações podem levar a considerações relevantes sobre o método de imputação mais apropriado para um determinado cenário. No entanto, como limitações dos experimentos presentes, é importante destacar que mais repetições para cada experimento são necessárias para controlar os efeitos aleatórios dos métodos de imputação de valores ausentes. Além disso, a otimização dos hiperparâmetros por meio de procedimentos de CV interno também é importante, especialmente para o parâmetro λ do algoritmo pairwiseMKL, bem como para o k no método de imputação KNN.

Agradecimentos: Os autores agradecem aos autores dos estudos de (CICHONSKA T. PAHIKKALA; ROUSU, 2018) por disponibilizarem seus dados e código publicamente.

Referências

AIOLLI, M. D. F. Easymkl: a scalable multiple kernel learning algorithm. Neurocomputing, 2015. Citado na página 12.

AMMAD-UD-DIN. Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. Bioinformatics, Oxford University Press, 2016. Citado na página 14.

CHONG, C.; SULLIVAN, D. New uses for old drugs. Nature, 2007. Citado na página 11.

CICHONSKA T. PAHIKKALA, S. S. H. J. A. A. M. H. T. A. A.; ROUSU, J. Learning with multiple pairwise kernels for drug bioactivity prediction. Bioinformatics, Oxford University Press, 2018. Citado 7 vezes nas páginas 8, 12, 15, 16, 18, 19 e 24.

COKOL IVAN IOSSIFOV, C. W. A. R. M. Emergent behavior of growing knowledge about molecular interactions. Nat Biotechnol, 2005. Citado na página 11.

CSERMELY TAMáS KORCSMáROS, H. J. K. G. L. R. N. P. Structure and dynamics of molecular networks: A novel paradigm of drug discovery a comprehensive review. Pharmacol Ther, 2013. Citado na página 11.

DALIANIS, H. Evaluation metrics and evaluation. Clinical text mining, Springer, 2018. Citado na página 17.

GONEN, E. A. M. Multiple kernel learning algorithms. The Journal of Machine Learning Research, 2011. Citado na página 12.

JIN, L. et al. A comparative study of evaluating missing value imputation methods in label-free proteomics. Scientific reports, Nature Publishing Group, v. 11, n. 1, p. 1–11, 2021. Citado na página 11.

KIRCH, W. Pearson's correlation coefficient. Encyclopedia of Public Health, Dordrecht: Springer Netherlands, 2008. Citado na página 17.

KUMAR, R. et al. Multiple Kernel Completion and its application to cardiac disease discrimination. Proceedings - International Symposium on Biomedical Imaging, p. 764–767, 2013. ISSN 19457928. Citado na página 13.

LIU, Y. Daily activity feature selection in smart homes based on pearson correlation coefficient. Neural Processing Letters, Springer, 2020. Citado 2 vezes nas páginas 13 e 17.

NASCIMENTO, R. B. C. P. A. C. A.; COSTA, I. G. A multiple kernel learning algorithm for drug-target interaction prediction. BMC Bioinformatics, 2016. Citado 3 vezes nas páginas 11, 12 e 13.

NEILL, R. M. H. S. P. Fundamentals of ocean renewable energy: generating electricity from the sea. Academic Press, 2018. Citado na página 18.

RIVERO R LEMENCE, T. K. R. Mutual kernel matrix completion. IEICE, 2017. Citado 2 vezes nas páginas 11 e 13.

RUBINSTEYN, A.; FELDMAN, S. fancyimpute: An Imputation Library for Python. Disponível em: <https://github.com/iskandr/fancyimpute>. Citado na página 16.

SZEDMAK, E. B. S. On the generalization of tanimoto-type kernels to real valued functions. arXiv:2007.05943, 2020. Citado na página 15.

TROYANSKAYA M. CANTOR, G. S. P. B. T. H. R. T. D. B. O.; ALTMAN1, R. B. Missing value estimation methods for dna microarrays. Bioinformatics (2001). DOI - PubMed, 2001. Citado na página 16.

TUIKKALA, J. Missing value imputation improves clustering and interpretation of gene expression microarray data. BMC bioinformatics, BioMed Central, 2008. Citado na página 16.

WEI, R. Missing value imputation approach for mass spectrometry-based metabolomics data. Scientific reports, Nature Publishing Group, 2018. Citado na página 16.

YANG, W. Genomics of drug sensitivity in cancer: a resource for therapeutic biomarker discovery in cancer cells. Nucleic acids research, Oxford University Press, 2012. Citado na página 14.