UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
CAMPUS RECIFE

Tássia Laís Barros Bastos

TRATAMENTO DE KERNELS INCOMPLETOS EM REDES BIPARTIDAS NA PREDIÇÃO DE INTERAÇÕES EM REDES BIOLÓGICAS

RECIFE
2020

TÁSSIA LAÍS BARROS BASTOS

TRATAMENTO DE KERNELS INCOMPLETOS EM REDES BIPARTIDAS NA PREDIÇÃO DE INTERAÇÕES EM REDES BIOLÓGICAS

Trabalho de Conclusão de Curso submetido à Universidade Federal Rural de Pernambuco, como requisito necessário para obtenção do grau de Bacharel em Ciência da Computação

Recife, outubro de 2020

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)

B327t Bastos, Tássia Laís Barros
Tratamento de kernels incompletos em redes bipartidas na predição de interações em redes biológicas / Tássia Laís Barros Bastos. - 2020.
75 f. : il.

Orientador: Andre Camara Alves do Nascimento.
Inclui referências e apêndice(s).

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2020.

1. Aprendizagem de múltiplos kernels. 2. Redes bipartidas. 3. PairwiseMKL. I. Nascimento, Andre Camara Alves do, 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 Tássia Laís Barros Bastos às 15 horas do dia 30 de outubro de 2020, no link https://meet.google.com/cco-siaq-yms, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado **Tratamento de Kernels incompletos em redes bipartidas na predição de interações em redes biológicas**, orientado por André Câmara Alves do Nascimento e aprovado pela seguinte banca examinadora:

André Câmara Alves do Nascimento
DC/UFRPE

Douglas Veras e Silva
DC/UFRPE

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

A decisão. Primeiro eu gostaria de agradecer à Tássia de 2015. Eu costumo brincar que essa foi a melhor decisão da minha vida. E num é que foi?! Agradeço àquela Biomédica recém formada por ter ido conhecer outros desafios, outras perspectivas, outras pessoas. A decisão em si de mudar foi um tanto quanto assustadora, mas essa experiência foi se descobrindo cada vez mais linda. Se eu pensei em desistir? Nunca passou pela minha cabeça!

O apoio. Mudar de carreira não é algo trivial ou simples. Mas dentre os meus tantos privilégios, o maior deles eu chamo carinhosamente de 'Madam'. Não convém ser piegas aqui, ainda mais quando o assunto é ela. Mas nunca será demais agradecer à quem me gerou e me amou sem nunca ter me visto, só me sentido. Obrigada por acreditar nas minhas loucuras e por apoiar as minhas escolhas. Ter você aqui, todos os dias (mesmo quebrando as paredes da casa) é a minha maior gratidão ao universo.

O incentivo. Eu me considero uma pessoa sortuda. Eu já entrei no mundo da computação com assessoria particular. Brincadeiras a parte, teve gente que o olho brilhou quando eu disse que tinha passado em Ciência da Computação. E hoje estamos aqui, me formando! Guigo, gratidão por você ter sempre me feito sentir parte integrante desse mundo e por me incentivar e me ajudar a me enxergar atuante e importante nessa profissão.

A realização. As pessoas que surgem no caminho são fundamentais para o sucesso desse processo. E como eu sou grata por todos os professores e principalmente por todas as professorAs que contribuíram para a minha formação. Gratidão por todo o conhecimento compartilhado, pelas conversas no corredor e pelos cafés (né, Ana Paula e George?). Agradeço também a todos os amigos que ganhei na ruralinda, em especial a Sandra (Rainha de BCC) e a João, Kenedy e Caio, meus fiéis parceiros de projetos.

Ao trio. Agradeço imensamente às minhas meninas, Maria e Ariany, que dividiram muitas histórias comigo. Crescer junto com vocês foi com certeza um dos melhores momentos da faculdade. Tenho um orgulho imenso das minhas meninas! Obrigada por tudo e por reafirmarem que o lugar da mulher é onde ela quiser! Vamos dominar esse mundão aí!

Ao Orie. Agradeço imensamente a Valmir por ter me apresentado o André Orientador. Já admirava André da sala de aula, mas o amigo/professor/orientador/psicólogo que eu ganhei não tem preço. Obrigada por aceitar me orientar e por surtar junto comigo quando os experimentos riam da nossa cara. Obrigada por sua dedicação, paciência, confiança e por nossa amizade, regada a muita cerveja alemã. Gratidão!

"Um passo à frente, e você não está mais no mesmo lugar."
(Chico Science)

Resumo

Na última década, o estudo de redes farmacológicas recebeu bastante atenção dada sua relevância para a produção de novos medicamentos. Os estudos foram propiciados mediante ao grande volume de dados biológicos gerados, possibilitando entender e extrair conhecimento em cima deles. Contudo, apesar de interessante, este é um processo que traz consigo algumas barreiras no quesito viabilidade, particularmente quando os dados aparecem de forma heterogênea e contêm informações ausentes. Muitas abordagens distintas para predição de interações biológicas vêm sendo propostas, com destaque para a área de aprendizagem de múltiplos kernels *Multiple Kernel Learning (MKL)*. O uso de métodos MKL em dados de natureza biológica também são comprometidos pela heterogeneidade das fontes de dados, mas associados aos métodos podem ser utilizadas técnicas de complementação de valores ausentes nas matrizes de *kernel* base. Esse processo de preenchimento geralmente é feito com técnicas simples, como imputação de zeros, média e mediana da matriz. Neste trabalho, técnicas de tratamento de valores faltosos foram avaliadas no contexto de redes bipartidas para solucionar as limitações relativas a heterogeneidade dos dados. Utilizamos três técnicas de imputação de valor único (média, mediana e zero) e uma técnica mais complexa de imputação preditiva (SVD). Todas as técnicas citadas já foram utilizadas para completude de matrizes no contexto de redes unipartidas. Nossas análises demonstraram que a técnica SVD apresentou um desempenho muito superior comparada às demais técnicas nas métricas avaliativas, trazendo resultados expressivos neste domínio para a utilização da técnica em modelos que utilizam redes bipartidas. As técnicas média e mediana apresentaram desempenhos similares, porém inferiores à SVD. E o preenchimento com zero apresentou o pior desempenho em relação a todas as outras técnicas aplicadas.

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 to the production of new drugs. The studies were made possible by the large volume of biological data generated, making it possible to understand and extract knowledge from them. However, although interesting, this is a process that brings with it some barriers in terms of viability, particularly when the data appear heterogeneously and contain missing information. Many different approaches for predicting biological interactions have been proposed, especially in the area of multiple kernel learning (Multiple Kernel Learning (MKL). The use of MKL methods in biological data is also compromised by the heterogeneity of data sources, but associated with the methods, techniques for complementing missing values in the base kernel matrices can be used, this filling process is usually done with simple techniques, such as imputing zeroes, mean and median of the matrix. In this work, techniques for handling false values were evaluated in the context of bipartite networks to solve the limitations related to the heterogeneity of the data. We used three single value imputation techniques (mean, median and zero) and a more complex predictive imputation technique (SVD). All the aforementioned techniques have already been used for matrix completeness in the context of unipartite networks. Our analyzes showed that the SVD technique performed much better than the other techniques in evaluative metrics, bringing encouraging results for the use of the technique in models that use bipartite networks. The average and median techniques showed similar performances, but lower than the SVD. And filling with zero showed the worst performance in relation to all other applied techniques.

Keywords: Multiple kernel Learning, Bipartite Networks, pairwiseMKL.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

MKL - Multiple Kernel Learning

SVM - Support vector machine

MKMC - Mutual kernel Matrix Completion

EM - Expectation Maximization

MKC - Multiple Kernel Clustering

GDSC - Genomics of Drug Sensitivity in Cancer

BLM - Bipartite local models'

HM - Half of the Minimum

RF - Imputation with Random Forest

SVD - Singular Value Decomposition Imputation

kNN - k Nearest Neighbors Imputation

QRILC - Quantile Regression Imputation of Left-Censored data

PKM - Pairwise kernel method

RMSE - Root Mean Square Error

KRR - Kernel Ridge Regression

CG - Conjugate Gradient

GDSC - Genomics of Drug Sensitivity in Cancer

Sumário

1 Introdução

Um dos maiores desafios que as ciências da saúde enfrentam atualmente é o desenvolvimento de novos medicamentos [Csermely et al. 2013]. No entanto, apesar dos crescentes investimentos — financeiros e de pesquisa — na área, o desenvolvimento rápido e de baixo custo de novos fármacos ainda está muito aquém da realidade atual, dado o avanço tecnológico que temos hoje em dia. Em média, são investidos até 2,6 bilhões de dólares e são necessários um pouco mais de 10 anos para lançar um novo medicamento no mercado [Fleming 2018] [ANVISA]. A indústria farmacêutica é uma das maiores investidoras e desenvolvedoras de pesquisa no mundo, e atualmente revelou aproximadamente uma centena de alvos farmacológicos para drogas aprovadas, dentro de um cenário total de mais de 20.000 proteínas não redundantes no proteoma humano [Csermely et al. 2013].

Nos últimos anos, redes droga-proteína receberam bastante notoriedade devido sua relevância para a inovação farmacêutica e produção de novos fármacos [Nascimento, Prudêncio e Costa 2016]. Com a grande produção de dados hoje em dia, extrair o conhecimento deles tornou-se uma tarefa interessante entre os cientistas de dados, mas tal avanço traz consigo algumas barreiras no quesito viabilidade, particularmente quando os dados aparecem de forma heterogênea, ou seja, quando contêm informações ausentes [Rivero, Lemence e Kato 2017].

Diante disso, fazendo uma correlação entre os conceitos biológicos e os conceitos computacionais, o problema de predição droga-proteína pode ser visto como uma abordagem de aprendizagem de máquina supervisionada [Rivero, Lemence e Kato 2017]. Dentro desse contexto, métodos de kernel têm obtido sucesso em vários problemas de classificação, inclusive quando se trata de redes farmacológicas [Shawe-Taylor, Cristianini et al. 2004].

Os métodos de kernel são categorizados como uma família de algoritmos que fazem classificação de padrões e que são capazes de agregar conhecimento prévio baseado na utilização de funções de similaridade, ou simplesmente, um kernel. A aplicação desses métodos tem obtido sucesso em vários problemas de aprendizagem supervisionada [Nascimento, Prudêncio e Costa 2016], como por exemplo, a proposta de [Stražar e Curk 2019] de realizar uma aproximação conjunta de kernel de baixa classificação e de *Multiple Kernel Learning (MKL)* para o problema de regressão. Contudo, a escolha da função de kernel e de seus respectivos parâmetros influenciam diretamente no desempenho obtido pelo classificador construído. Em sua maioria, as funções e os parâmetros são selecionados por meio de um processo de validação cruzada em um conjunto de validação distinto do conjunto de treinamento.

Multiple Kernel Learning (MKL) é uma área com crescente importância dentro

do contexto de aprendizagem de máquina [Gönen e Alpaydın 2011], cuja motivação é equivalente à considerada na combinação de múltiplos classificadores: é mais adequado utilizar um conjunto distinto de kernels e deixar que o algoritmo selecione os melhores ou sua respectiva combinação. As vantagens da aplicação de diferentes kernels para o mesmo problema são que:

- permite diferentes noções de similaridade entre as instâncias do problema, enriquecendo o treinamento do classificador com informações relevantes;
- as noções de similaridade obtidas, mesmo que sejam sobre os mesmos elementos, podem ser extraídas de representações distintas destes elementos.

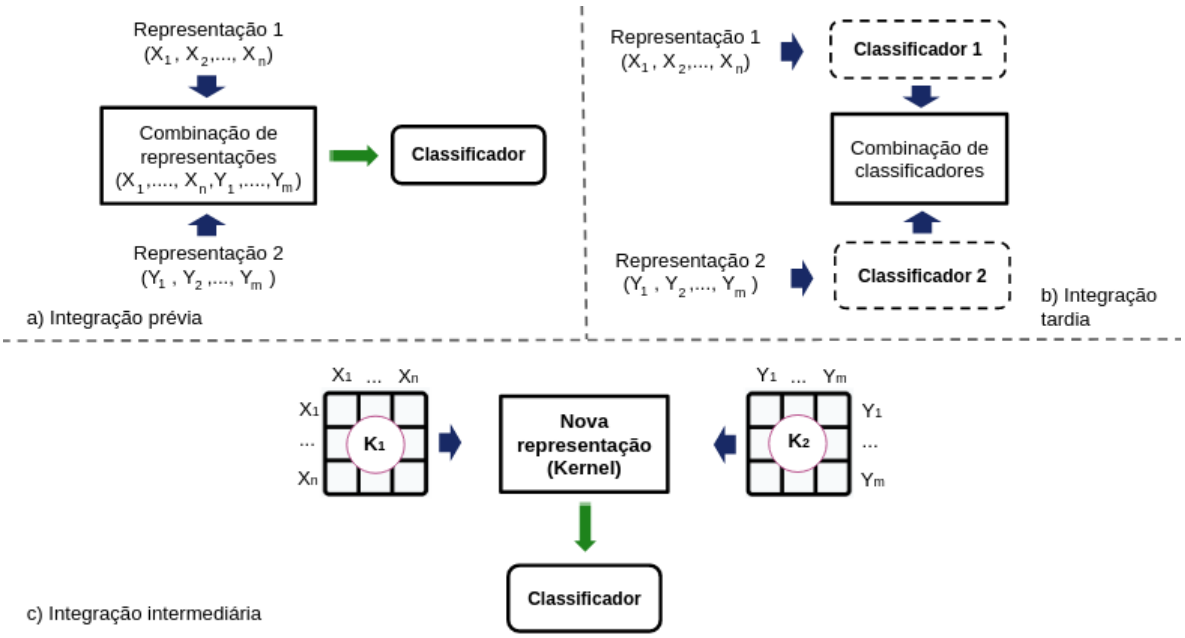


Figura 1 – Métodos de integração para diferentes representações de características: (a) integração prévia, (b) integração tardia, e (c) integração intermediária. Fonte: [Nascimento 2015].

Por conseguinte, a associação de kernels pode ser entendida como uma estratégia de combinação de diferentes bases de informações, ou até mesmo como uma combinação intermediária (Figura 1- c), de acordo com a classificação proposta por [Vert, Tsuda e Schölkopf 2004]. Além da abordagem supracitada, existem outras duas abordagens: a integração prévia (Figura 1- a) e a integração tardia (Figura 1- b).

Por definição, a integração prévia consiste na concatenação de diversos conjuntos de vetores das instâncias *(feature vectors)* em um vetor único para posterior treinamento do classificador. Esta técnica determina um crescimento do vetor resultante obtido, proporcionalmente aos tamanhos das representações base, porém podendo causar uma desproporção no tamanho do conjunto de treinamento, conhecida como a "maldição da dimensionalidade"

(curse of dimensionality) [Jain, Duin e Mao 2000]. A integração tardia, por sua vez, resume-se em treinar diferentes classificadores, sendo cada um referente a uma representação, e ao termino, combiná-los em um classificador final. Tal técnica tem como principal desvantagem o elevado custo de predição associado, visto que é necessário treinar diferentes classificadores para se realizar uma única tarefa [Nascimento 2015].

Como o próprio nome da técnica sugere, a integração intermediária surge como um meio termo entre as abordagens anteriores, e distingue-se pelo fato de que faz uso de novos descritores de representações, diferente das originais, baseados em similaridade entre as matrizes de kernel, que posteriormente são utilizadas como entrada para máquina de kernel (um único classificador) [Nascimento 2015].

1.1 Justificativa

Para utilizar os métodos de kernel para construção de um modelo preditivo é preciso definir uma medida de similaridade (kernel) entre as instâncias do problema, uma vez que o objetivo é predizer a probabilidade de ocorrência de uma ligação entre uma droga e seu alvo na rede. Diante disso, a geração da medida de similaridade é vista como uma limitação computacional dos métodos de kernel por causa da multiplicação explícita dos kernels [Nascimento 2015]. Entretanto, o algoritmo KronRLS-MKL, proposto por [Nascimento, Prudêncio e Costa 2016], apresenta características que contornam as limitações computacionais citadas anteriormente. É um método que modela o problema de interação droga-alvo como uma tarefa de predição de links em redes bipartidas. Para isso, ele intercala a otimização dos parâmetros da função de predição pareada com a otimização de pesos do kernel. Contudo, ele encontra dois conjuntos de pesos de kernel separadamente, um para kernels de drogas e outro para kernels de proteínas. Por não gerar os pesos de kernels emparelhados, o método não explora totalmente as informações contidas no espaço de pares.

De forma análoga ao KronRLS-MKL, o algoritmo pairwiseMKL, proposto por [Cichonska et al. 2018], visa uma melhor performance computacional quando comparado aos métodos mais tradicionais de MKL. Seu funcionamento baseia-se na determinação dos pesos da mistura dos kernels emparelhados de entrada e, em seguida, o algoritmo aprende a função de predição dos pares. Ambas as etapas são feitas de modo eficiente, sem cálculo explícito das matrizes massivas de pares, tornando o método aplicável para a solução de grandes problemas de aprendizagem de pares.

Apesar de algumas novas abordagens resolverem os impasses computacionais, a adoção de métodos MKL em conjuntos de dados de natureza biológica possui outra barreira importante. Atualmente, os métodos de predição de interações em redes farmacológicas assumem que os dados de entrada são composto por matrizes de kernel completas, isto

é, sem valores faltosos. No entanto, na maioria das vezes, os dados obtidos no mundo real são ruidosos e têm informações incompletas [Liu et al. 2019]. A técnica mais simples para lidar com dados perdidos é excluir os pontos de dados com entradas ausentes. No entanto, isso leva a uma redução do conjunto de dados disponíveis, limitando o potencial preditivo dos métodos. Porém, visando a não redução do conjunto de dados, é comum associar aos métodos técnicas simples de complementação de valores ausentes nas matrizes de kernel base. Esse processo de preenchimento geralmente é feito com técnicas simples, como imputação de zeros, média e mediana da matriz [Wei et al. 2018] [Acock 2005] [Zhang 2016] [Tuikkala et al. 2008].

Alguns trabalhos têm abordado o problema de valores ausentes como um problema de aprendizagem supervisionada. Segundo [Kumar et al. 2013], o problema de derivar uma matriz de kernel de um conjunto de matrizes incompletas pode ser contornado fazendo o preenchimento dos valores faltosos com o uso de combinação linear via máquina de vetores de suporte *(Support vector machine - SVM)*, entretanto, seria envolvido um alto custo computacional para realizar tal ação.

Os trabalhos de [Rivero, Lemence e Kato 2017] e [Liu et al. 2019] apresentam o algoritmo *Mutual kernel Matrix Completion (MKMC)*, que explora o algoritmo *Expectation Maximization (EM)* para minimizar a divergência de *Kullback-Leibler* entre as matrizes de kernels base e o algoritmo *Multiple Kernel Clustering - MKC*, que trata posições faltosas nas matrizes de kernel como variáveis auxiliares a serem otimizadas, respectivamente. Tais métodos foram desenvolvidos buscando solucionar os problemas de *kernel completation*, porém com aplicações apenas no contexto de redes unipartidas.

No entanto, os trabalhos citados acima ainda são limitados. No contexto da aprendizagem de kernels incompletos, não foi observado nenhum trabalho que integre o tratamento de tais kernels ao processo de aprendizagem de kernels em redes bipartidas, mais especificamente, em redes farmacológicas. Além disso, o fato do problema em redes bipartidas envolver informações de conectividade de nós, estas podem ainda ser utilizadas para propor modelos de estimação de valores de kernel baseados em vizinhança [Rivero, Lemence e Kato 2017]. Por isso, neste trabalho, o objetivo é analisar e inferir sobre um ramo específico da Biologia Computacional, a predição de interações em redes biológicas a partir do tratamento de kernels incompletos em redes bipartidas.

1.2 Escopo do trabalho

O escopo do trabalho compreende uma atualização dos estudos descritos em [Cichonska et al. 2018], trazendo resultados de um experimento na abordagem de quatro técnicas para o tratamento de kernels incompletos em redes bipartidas. Adicionalmente, dada a importância do conhecimento de kernels incompletos, de maneira a não afetar

estudos decorrentes, é fundamental que a predição tenha um bom resultado estatístico. Com base nesta necessidade, o preditor descrito em [Cichonska et al. 2018] foi avaliado sobre o impacto de dados faltosos e das diferentes técnicas utilizadas para o tratamento deste problema no processo de integração de dados.

1.3 Formulação de hipóteses

O presente trabalho traz a seguinte questão de pesquisa:

Qual o impacto de diferentes abordagens de tratamento de kernels incompletos em problemas de predição de interações em redes bipartidas?

1.4 Elaboração dos objetivos

1.4.1 Objetivo Geral

Investigar as técnicas de aprendizagem de máquina e tratamento de kernels incompletos para compor a estrutura do experimento de predição de interações em redes bipartidas.

Objetiva-se, neste estudo, entender os métodos de tratamento de kernels incompletos no contexto da aprendizagem de múltiplos kernels (MKL), implementar e adaptar técnicas de tratamento de valores faltosos para o contexto de MKL em redes droga-linha celular (bipartidas) e avaliar o desempenho das técnicas aplicadas.

1.4.2 Objetivos Específicos

Diferentes soluções serão investigadas para reduzir o impacto de valores faltosos na etapa de integração e aprendizado dos kernels a serem usados na predição. Em resumo, é possível elencar como objetivos principais do projeto os seguintes itens:

- Avaliar o impacto de métodos clássicos de tratamento de valores faltosos neste domínio;
- Avaliar métodos baseados em fatoração de matrizes para valores faltosos, especialmente no contexto da aprendizagem de múltiplos kernels.

1.5 Definição da metodologia

Inicialmente, definimos as abordagens utilizadas neste trabalho (desenvolvimento e experimento, Figura 2), destacando as etapas que compõem a metodologia. Tais etapas

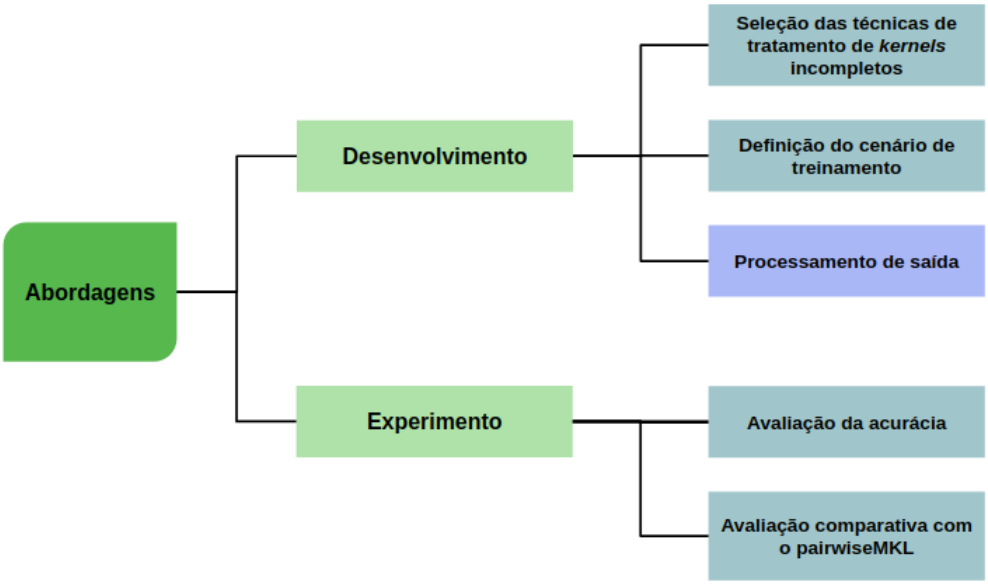


Figura 2 – Metodologia do trabalho. Fonte: elaborado pela autora.

foram construídas baseadas essencialmente no levantamento bibliográfico realizado, e por isso, definimos nossas escolhas de acordo com os melhores resultados alcançados até então.

Para a etapa de desenvolvimento determinamos as técnicas de tratamento de kernels incompletos utilizadas neste estudo, assim como determinamos o cenário de treinamento. A base de dados utilizada (descrita na subseção 3.2.1) foi derivada do projeto *GDSC (Genomics of Drug Sensitivity in Cancer)*, iniciado pelo *Wellcome Trust Sanger Institute* [Yang et al. 2012] e utilizada por [Cichonska et al. 2018], sendo o treinamento realizado com validação cruzada externa, mais especificamente *10-fold cross-validation*. Importante destacar que realizamos uma adaptação no algoritmo original apresentado por [Cichonska et al. 2018], onde utilizamos apenas a validação cruzada externa (*10-fold cross-validation*) e desconsideramos do modelo a validação cruzada interna utilizada.

Para a etapa de experimentação, determinamos que a avaliação de desempenho do modelo será feita de acordo com o resultado do diagrama de dispersão gerado entre os valores reais e os valores preditos da matriz de interação. A avaliação de desempenho do experimento será feita através da comparação entre os resultados obtidos das técnicas aplicadas associadas às porcentagens de valores faltosos e os resultados obtidos no caso base. E por fim, as métricas de avaliação utilizadas no experimento foram as mesmas adotadas por [Cichonska et al. 2018], e.g, RSME, F1-score e o Índice de Correlação de Pearson.

1.6 Estrutura do trabalho

O presente trabalho está organizado da seguinte forma:

Capítulo 2: abordará os conceitos básicos e terminologias utilizadas ao longo do trabalho. Inicialmente são apresentados os conceitos que correspondentes aos temas biológicos, como a análise de redes farmacológicas, a definição e utilização de redes biológicas, e o processo de predição de interações em redes farmacológicas. No capítulo também são abordadas questões de predição de interações envolvendo a heterogeneidade dos dados e os métodos de kernel e *Multiple Kernel Learning*. Além disso, descreve alguns trabalhos que, assim como este, versam sobre o tratamento de kernels incompletos.

Capítulo 3: descreve o desenvolvimento do trabalho, o qual é resultante da metodologia proposta. Está contido, neste capítulo, como foi feito o experimento e treinamento do modelo de aprendizado, a obtenção dos dados e a validação da metodologia proposta do experimento.

Capítulo 4: são discutidos alguns aspectos relativos aos resultados, com as devidas justificativas aos fatos ocorridos no desenvolvimento.

Capítulo 5: contém as conclusões obtidas com os resultados do trabalho bem como algumas sugestões de trabalhos futuros.

2 Conceitos básicos e terminologia

Historicamente, é sabido que os primeiros fármacos foram descobertos ao acaso ou intuitivamente, por meio de observações inesperadas de fatos do dia-a-dia ou em etapas de triagem clínica. Um grande exemplo foi a descoberta da penicilina, por Alexander Fleming em 1928 [Barreiro e Fraga 2005]. Quase 100 anos depois, o desenvolvimento de novos medicamentos ainda é considerado um dos maiores desafios que as ciências da saúde enfrentam atualmente [Csermely et al. 2013]. Apesar dos crescentes investimentos — financeiros e de pesquisa — na área, o desenvolvimento rápido e de baixo custo de novos fármacos ainda está muito aquém da realidade atual, dado o avanço tecnológico que temos hoje em dia. Em média, são investidos até 2,6 bilhões de dólares e são necessários de 12 a 15 anos para lançar um novo medicamento no mercado [Quental e Filho 2006] [Fleming 2018] [ANVISA].

A indústria farmacêutica é uma das maiores investidoras e desenvolvedoras de pesquisa no mundo, e atualmente revelou aproximadamente uma centena de alvos farmacológicos para drogas aprovadas, dentro de um cenário total de mais de 20.000 proteínas não redundantes no proteoma humano [Csermely et al. 2013]. Apesar de todo investimento, a indústria farmacêutica possui alguns gargalos nos deus processos, dificultando o seu desenvolvimento. Um ótimo exemplo são as etapas de uma pesquisa clínica, que usualmente englobam quatro fases: I, II, III e IV [ANVISA].

2.1 Análise de redes farmacológicas

Antes de começar a fase I, testes pré-clínicos são feitos visando a segurança do estudo, realizando, então, testes em animais de experimentação antes da aplicação da droga em seres humanos. Quando a medicação está apta para testes em humanos, fases de investigação clínica são iniciadas a fim de coletar o maior volume possível de informações sobre a ação do medicamento em teste [ANVISA].

A fase I avalia as diferentes vias de administração e diferentes doses do medicamento aplicadas em um grupo de 20 a 100 indivíduos saudáveis que terão contato com o fármaco pela primeira vez. A fase II utiliza um grupo maior indivíduos (100 a 300), porém pessoas que apresentam a doença ou condição para a qual o procedimento está sendo estudado, visando obter mais dados de segurança e também avaliar a eficácia do novo medicamento proposto [ANVISA] [Quental e Filho 2006]. Os testes feitos nessa fase avaliam diferentes dosagens e as diferentes indicações do novo medicamento. As fases I e II compreendem ao estudo piloto da droga em teste. A fase III utiliza um grupo grande de pacientes (5-10 mil), com este número variando a depender da patologia em questão, por um período

maior de tempo, geralmente comparando o novo tratamento à outros já existentes. O objetivo principal desta fase é obter informações sobre a eficácia e interação de drogas. E em posterior análise de todos os dados obtidos na fase III, o medicamento pode ser levado a registro e aprovação para uso comercial pelo órgão sanitário competente [ANVISA].

A fase IV, também conhecida como Farmacovigilância, compreende testes de acompanhamento de uso do medicamento, após ser aprovado e levado ao mercado, e objetiva detectar e definir efeitos colaterais previamente desconhecidos ou incompletamente qualificados, bem como os fatores de risco associados [ANVISA] [Quental e Filho 2006].

Diante do exposto, conclui-se que as quatro fases são demoradas e requerem muito cuidado e planejamento. Então, dentro dessa conjuntura, a comunidade farmacêutica acredita que descobrir novas indicações e alvos para medicamentos já existentes, isto é, 'reposicionamento de drogas' [Vanhaelen 2019] [Nascimento, Prudêncio e Costa 2016] [Ekins et al. 2011], pode suprir a baixa eficiência das abordagens tradicionais na descoberta de novas drogas e o tempo do processo atual de uma pesquisa clínica. Como principais vantagens da abordagem de reutilização de drogas estão: os perfis pré-clínicos, farmaco-dinâmicos, farmacocinéticos e de toxicidade de fármacos já conhecidos, que reduzem o risco do desenvolvimento de compostos que poderiam causar efeitos colaterais e danos aos usuários. Deste modo, os estudos clínico de Fase II e III do medicamento podem desenvolver-se mais rapidamente, resultando em um menor custo de desenvolvimento, um melhor retorno do investimento e menor tempo de desenvolvimento e lançamento no mercado [Vanhaelen et al. 2017] [Ekins et al. 2011].

2.1.1 Redes biológicas

Uma forma de representação e visualização das interações entre dados biológicos é uma rede. Uma rede pode ser definida como uma estrutura composta por um conjunto de nós e um conjunto de relações resultantes da interação entre eles [Nascimento 2015]. Esse tipo de estrutura pode ser utilizada e encontrada em diversas situações do cotidiano, como por exemplo, em aplicativos que escolhem a melhor rota para se chegar a um destino, ou na conexão entre páginas Web. Uma rede considerada simples é composta apenas por um conjunto de nós e das ligações não direcionadas entre eles, e usualmente é representada na matemática como um grafo, ou seja, uma estrutura convencional em que os nós são denominados de vértices e ligações são denominadas de arestas [Moreira, Campos e Jr 2019].

Desta forma, pode-se representar uma rede biológica como um grafo $G = (V, E)$, com $V = \{v_1, ..., v_n\}$ vértices e $E = \{(v_i, v_j)|v_i, v_jV\}$, com $i, j \in \{1, 2, ..., n\}$, o conjunto de interações (arestas). Tal estrutura pode ser representada de várias formas, mas destacam-se as formas gráfica (Figura 3 - a) e de matriz de adjacências (Figura 4), a qual corresponde a uma matriz binária $Y \in \{0, 1\}^{nxn}$, onde $Y_{ij} = 1$ quando $(v_i, v_j) \in E$, e $Y_{ij} = 0$ caso

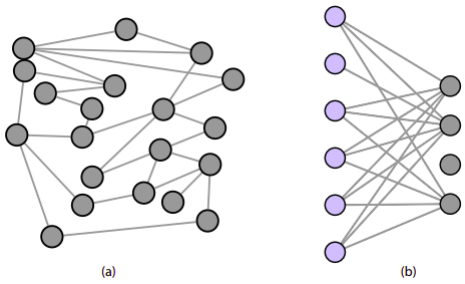
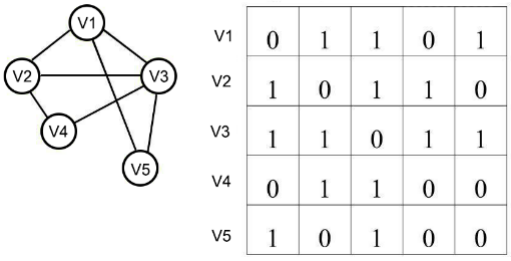


Figura 3 – Representação gráfica de um grafo: (a) Grafo não direcionado unipartido; (b) Grafo bipartido. Fonte: [Nascimento 2015].



	V1	V2	V3	V4	V5
V1	0	1	1	0	1
V2	1	0	1	1	0
V3	1	1	0	1	1
V4	0	1	1	0	0
V5	1	0	1	0	0

Figura 4 – Representação da matriz de adjacência, a qual corresponde a uma matriz binária $Y \in \{0,1\}^{nxn}$, onde $Y_{ij} = 1$ quando $(v_i, v_j) \in E$, e $Y_{ij} = 0$ caso contrário .

contrário [Nascimento 2015].

Além da aplicação prática dos grafos em situações do cotidiano e em problemas matemáticos, grande parte das interações químicas e biológicas podem ser modeladas como uma rede, sob diferentes aspectos [Moreira, Campos e Jr 2019]. Por exemplo, três células de defesa do corpo que atuam no mesmo processo de defesa de uma infecção podem ser representadas como três vértices conectados por arestas; ou um grafo direcionado pode ser modelado para representar as transformações de compostos bioquímicos de uma via metabólica, indicando o fluxo das reações em cadeia.

A utilização de conceitos matemáticos e computacionais como meio para a modelagem dos inúmeros cenários biológicos através de redes de interações, sejam elas entre doenças, fármacos ou alvos celulares, nos permite a observação e descoberta de propriedades e comportamentos que dificilmente seriam identificados quando analisados isoladamente. Atrelado a isso, a modelagem estrutural e estatística das redes de interação biológicas somam mais uma camada ao mecanismo de ação de drogas, auxiliando na explanação de causas do surgimento de doenças. O compartilhamento de conhecimento entre essas áreas tem viabilizado a produção de métodos *in silico* que visam a identificação, validação e predição de interações entre estruturas biológicas, com resultados promissores em ensaios *in vitro* validados [Keiser et al. 2009].

Entretanto, alguns desafios surgem no processo de reconstrução e modelagem de redes biológicas como, por exemplo, o elevado nível de ruido em experimentos dessa natureza; a alta dimensionalidade e complexidade dos dados e principalmente as condições experimentais heterogêneas entre as bases de dados [Nascimento, Prudêncio e Costa 2016].

2.1.2 Predição de interações em redes farmacológicas

A análise de dados biológicos de grande escala possibilitou o entendimento das interações entre os diversos tipos de moléculas de origem biológica. Diante dessa demanda, a Biologia de Sistemas surgiu como uma ferramenta para o entendimento da complexidade dos sistemas biológicos e envolve outras áreas além da biologia, com a teoria dos grafos, a matemática discreta e o processamento computacional [Silva Daniel Luis Notari 2020].

Com o surgimento dessa ferramenta, estudar redes farmacológicas se tornou um paradigma promissor para o desenvolvimento de drogas. Conceitualmente, a farmacologia em rede propõe investigar efeitos sinérgicos e potenciais mecanismos de múltiplos compostos analisando redes complexas e com multicamadas, baseado na farmacologia e farmacodinâmica [Zhao et al. 2019]. Assim, a evolução do estudo em redes farmacológicas pode gerar novas oportunidades para entender as interações entre compostos ativos e alvos relevantes, e consequentemente destacar os mecanismos de ação envolvidos [Wang et al. 2017], [Dong et al. 2019].

A Figura 5 apresenta um exemplo de rede farmacológica, especificamente de interações entre proteínas G, representada por círculos vermelhos, e drogas, representadas por círculos verdes. A aresta que conecta os dois nós indica uma relação entre eles, ou seja, um interação conhecida entre as duas moléculas. Essa representação constitui uma fonte rica de informação, revelando interações sutis que podem auxiliar na predição de novos relacionamentos até então desconhecidos [Nascimento, Prudêncio e Costa 2016] .

2.2 Predição de interações utilizando dados heterogêneos

Como dito anteriormente, muitos problemas biológicos podem ser modelados na estrutura matemática de grafos, e também mais precisamente como problemas de aprendizado em pares [Cichonska et al. 2018]. Nesse contexto, podemos referir uma rede como um grafo bipartido (Figura 3 - b) onde, G = (V, E), V pode ser dividido em V_x e V_y. Deste modo, considerando E(a,b), podemos ter a relação: a $\in V_x$ e b $\in V_y$ ou b $\in V_x$ e a $\in V_y$ [(Org.) 2014].

A modelagem de redes biológicas vem utilizando e aplicando os conceitos de redes bipartidas em vários contextos biológicos, como representação de conexões ecológicas, estudo de reações enzimáticas em vias metabólicas e interações entre drogas-célula alvo [(Org.) 2014].

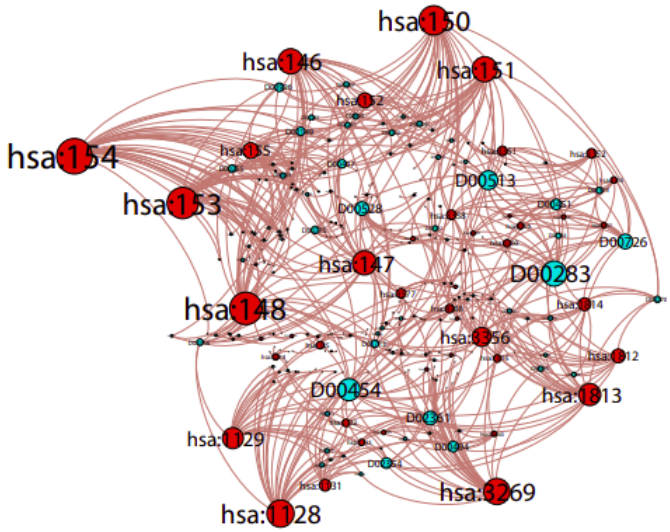


Figura 5 – Rede de interações droga-proteína (G-protein coupled receptors - GPCR's) extraída de bases de dados públicas (interações disponibilizadas por [Yamanishi et al. 2008].

2.2.1 Predição de interações baseada em métodos de kernel e *Multiple Kernel Learning*

Para desenvolver novos fármacos, muitas abordagens diferentes para predição de interações droga-célula alvo vem sendo propostas. Muitas delas são baseadas em uma classe particular de métodos de aprendizagem de máquina, chamada de métodos de kernel. Tal método de aprendizagem possui algoritmos de classificação de padrões que são capazes de incorporar conhecimento prévio na forma de funções de similaridade, obtendo sucesso em vários problemas de aprendizagem supervisionada [Nascimento, Prudêncio e Costa 2016], [Cichonska et al. 2018].

De maneira geral, os métodos de predição atuam sobre um espaço chamado de quimiogenômico, onde interações conhecidas, informações de drogas e alvos celulares estão unificados. Desse espaço são obtidas várias informações. Uma vez definidas as interações conhecidas, é aplicada uma medida de similaridade entre as drogas presentes no espaço quimiogenômico, ou seja, todas as drogas do conjunto serão comparadas entre si. O mesmo é feito para as células alvo/proteínas (Figura 6 - A). No contexto da aprendizagem de múltiplos kernels, são produzidas diversas matrizes com este objetivo (Figura 6 - B). Por fim, todas as informações coletadas são agrupadas para a construção de um modelo preditivo, resultando na inferência de novas interações na rede biológica (Figura 6 - C).

O princípio é que ligantes similares estão mais propensos a interagirem com células alvo/proteínas parecidas. Com isso, a predição do modelo leva em consideração as características topológicas da rede de interações já conhecidas, assim como os descritores de todas as entidades envolvidas. Os descritores podem ser descritos como os efeitos colaterais, o

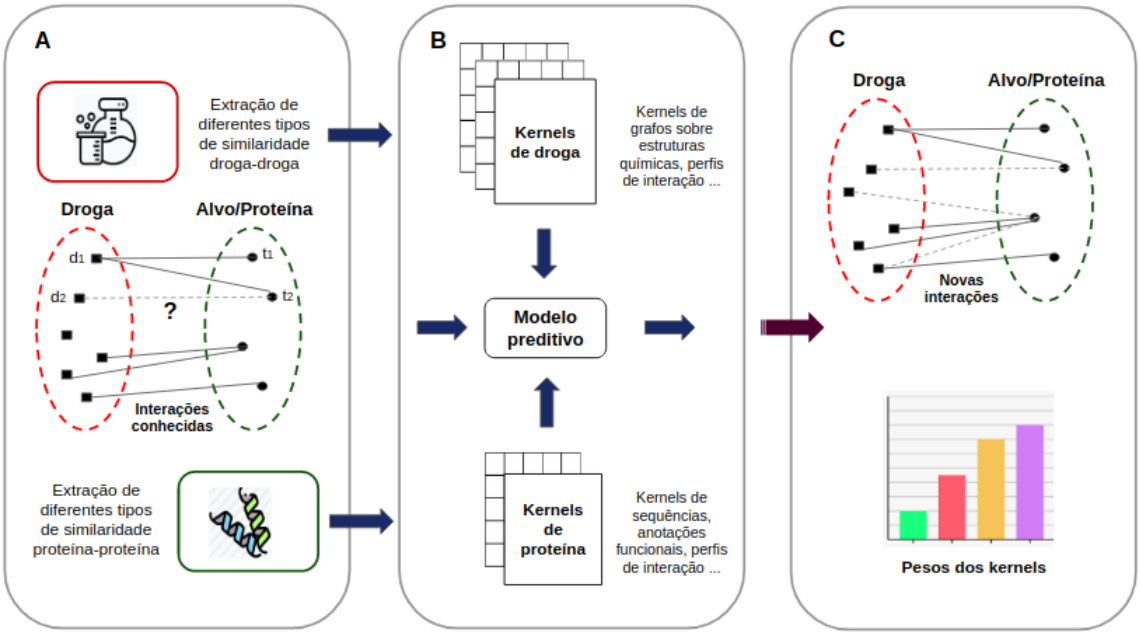


Figura 6 – A predição de interações droga-proteína baseada em similaridade é composta de três etapas principais: a partir de um conjunto de interações conhecidas (A), são extraídas medidas de similaridade entre cada tipo de entidade envolvida (droga-droga e proteína-proteína) (B). Em seguida estas informações são utilizadas para construção de um modelo preditivo, o qual é posteriormente usado para predizer novas interações (C). Figura adaptada de [Nascimento 2015].

perfil de expressão gênica ou qualquer outra informação relevante que ajude a deixar a predição mais rica [Nascimento, Prudêncio e Costa 2016].

Tais características fazem dos algoritmos de aprendizagem baseados em kernel uma alternativa bastante adequada a problemas deste tipo. Dada uma rede de interação conhecida, os métodos baseados no kernel podem ser usados para prever interações entre drogas e alvos celulares desconhecidas [Nascimento, Prudêncio e Costa 2016].

O desempenho e sucesso do classificador é devidamente alcançado com a seleção adequada da função kernel e dos seus respectivos parâmetros. Ultimamente, a aprendizagem de múltiplos kernels vem sendo utilizada para solucionar este problema, permitindo o uso de múltiplos kernels, ao invés de considerar apenas um kernel para uma dada tarefa. Uma das motivações para essa abordagem é similar à considerada na combinação de múltiplos classificadores: ao invés de restringir-se a um único kernel, é preferível utilizar um conjunto de kernels distintos, e deixar que um algoritmo selecione os melhores, ou sua respectiva combinação [Yan, Zhang e He 2019], [Nascimento, Prudêncio e Costa 2016].

Diversas abordagens computacionais vêm se desenvolvendo para realizar as análises e predizer interações entre compostos químicos e proteínas, sendo categorizados geralmente em seis: modelos baseados em vizinhança, que levam em consideração o perfil de interações entre os nós vizinhos mais próximos para realizar as predições; modelos Locais Bipartidos,

do inglês *'Bipartite local models'* (BLMs), que executa a predição em duas etapas separadas, a primeira consistindo da interação droga-proteína e a segunda consiste na agregação das predições de cada domínio para gerar as predições finais; modelos de difusão na rede, onde novas interações são preditas graças às técnicas de difusão de informações em redes complexas; modelos baseados em *docking*, que baseiam-se em simulações de ligação entre compostos, no qual escores de atividade potencial são calculados sobre modelos tridimensionais de estruturas químicas [Ezzat et al. 2019], [Nascimento, Prudêncio e Costa 2016].

Entretanto, uma limitação dos métodos de predição de interações droga-alvo celular é o fato de que a grande maioria deles supõem que todos os kernels base estão completos, isto é, nenhuma das linhas ou colunas de qualquer matriz considerada deve estar ausente. Esta premissa não condiz com a realidade de problemas biológicos, no qual é comum algumas informações estarem presentes em determinadas fontes e não em outras, por exemplo, similaridade 3D, dados de expressão gênica, etc [Nascimento, Prudêncio e Costa 2016]. Soluções comumente adotadas são a remoção das instâncias cujas informações estão incompletas, o que leva a uma redução do conjunto de dados e consequentemente do seu poder preditivo. Ou ainda, a adoção de estratégias simples de preenchimento de tais matrizes, como a média ou o preenchimento da linha/coluna com zeros [Rivero, Lemence e Kato 2017].

2.3 Tratamento de kernels incompletos

O crescente volume de dados de origem biológica acarreta na aparição de dois tipos de problemas. O primeiro deles corresonde ao armazenamento e a gestão dos dados coletados, e o segundo corresponde a extração de informações mediante a manipulação desses dados [Tavares 2015].

O segundo problema é tido atualmente como um dos maiores desafios da biologia computacional, demandando da comunidade científica o desenvolvimento de métodos e ferramentas capazes de extrair o conhecimento biológico de todos esses dados heterogêneos [Rivero, Lemence e Kato 2017] [Tavares 2015].

A heterogeneidade dos dados dificulta a adoção de alguns modelos computacionais já existentes na literatura devido ao fato de que as informações contidas nessas bases de dados muitas vezes vem incompletas ou até mesmo ausentes. Algumas soluções são comumente adotadas para lidar com dados heterogêneos, porém não são muito efetivas. Dentre elas estão a remoção das instâncias cujas informações não estão completas, que leva a uma diminuição do conjunto de dados e consequentemente do poder preditivo daquela amostra. Alguns estudos recentes, que utilizam técnicas de MKL em problemas de natureza unipartida têm investido na complementação de valores ausentes em matrizes de kernel

base. Segundo [Kumar et al. 2013], o problema de derivar uma matriz de kernel de um conjunto de matrizes incompletas pode ser contornado fazendo o preenchimento dos valores faltosos. O preenchimento pode ser feito com a média ou simplesmente colocando zeros nas linhas e colunas da matriz [Nascimento, Prudêncio e Costa 2016] [Rivero, Lemence e Kato 2017]. Diante do exposto, o tratamento de dados faltosos nas matrizes de kernel pode ser amplamente melhorado considerando-se os recentes avanços na pesquisa de métodos de imputação de valores faltosos em problemas MKL [Liu et al. 2019] [Kumar et al. 2013].

No contexto de redes unipartidas, muitos trabalhos na literatura apresentam diversas técnicas que objetivam preencher as lacunas das matrizes. No estudo de [Wei et al. 2018] foi realizada uma comparação, de forma abrangente, entre oito métodos de imputação de valores faltosos no contexto de dados metabolômicos baseados em espectrometria de massas. As técnicas foram: 1) Zero: esta técnica realiza uma imputação simples nos elementos faltosos com o número zero; 2) Meio mínimo (*Half of the Minimum - HM*): executa a substituição dos elementos ausentes pela metade do mínimo de elementos não ausentes na variável correspondente; 3) Média: esta técnica realiza uma imputação simples nos elementos ausentes com valor médio dos elementos não ausentes na variável correspondente; 4) Mediana: de forma similar à técnica anterior, há uma imputação simples nos elementos ausentes utilizando o valor mediano dos elementos não ausentes na variável correspondente; 5) Floresta aleatória (*Imputation with Random Forest - RF*), corresponde a uma técnica de imputação mais sofisticada, que utiliza algoritmos de aprendizagem de máquina para construir um modelo de predição iterativo; 6) Decomposição de valor singular (*Singular Value Decomposition Imputation - SVD*): essa técnica substitui inicialmente todos os elementos ausentes por zero e faz uma estimação com a combinação linear das k variáveis próprias mais significativas iterativamente, até atingir certo limite de convergência; 7) K-vizinhos mais próximos (*k Nearest Neighbors Imputation - kNN*): a utilização do algoritmo de kNN original foi desenvolvida para dados de expressão de genes de microarranjos de alta dimensão (n «p, onde 'n' é o número correspondente de amostras e 'p' é o número correspondente de variáveis). Para cada um dos gene com valores faltosos, o método encontra os k genes mais próximos utilizando a métrica Euclidiana e preenche os espaços vazios com o valor calculado da média dos elementos não ausentes de seus vizinhos; 8) Imputação de regressão quantílica de dados censurados à esquerda (*Quantile Regression Imputation of Left-Censored data - QRILC*): essa técnica foi projetada especificamente para dados censurados à esquerda, ou seja, dados ausentes causados pelos limites de quantificação. A técnica funciona preenchendo os elementos ausentes com o desenho aleatório de uma distribuição truncada estimada por uma regressão de quantis.

[Rivero, Lemence e Kato 2017], apresenta o algoritmo *Mutual kernel Matrix Completion (MKMC)*, que explora o algoritmo *Expectation Maximization (EM)* para minimizar a divergência de *Kullback-Leibler* entre as matrizes de kernels base. Os resultados indicam que a medida que a proporção de dados faltosos aumenta, o algoritmo amplia sua

vantagem sobre estratégias mais simples, como o preenchimento com zeros ou com a média. Abordagens recentes têm buscado a integração da combinação de kernels e o tratamento de kernels incompletos em um único passo, a fim de reduzir o custo computacional de tratar os dois problemas em etapas distintas.

No estudo de [Liu et al. 2019], é proposto um algoritmo de *Multiple Kernel Clustering - MKC* que trata posições faltosas nas matrizes de kernel como variáveis auxiliares a serem otimizadas, também obtendo resultados encorajadores. Este trabalho é estendido no estudo feito por [Zhu et al. 2018], no qual é proposta uma versão localizada do algoritmo, requerendo apenas analisar a vizinhança local (k-vizinhos) de uma amostra para estimar os valores faltosos.

Entretanto, todos os trabalhos citados acima desenvolveram seus experimentos de preenchimento de valores faltosos em redes unipartidas, não sendo observadas evidências para o contexto bipartido.

3 Descrição do Experimento

Para desenvolver novos fármacos, muitas abordagens diferentes para predição de interações droga-linhas celulares vem sendo propostas. Muitas delas são baseadas em uma classe particular de métodos de aprendizagem de máquina chamada de métodos de kernel [Nascimento, Prudêncio e Costa 2016]. Tal método de aprendizagem utiliza algoritmos de classificação de padrões que são capazes de incorporar conhecimento prévio na forma de funções de similaridade, obtendo sucesso em vários problemas de aprendizagem supervisionada [Nascimento, Prudêncio e Costa 2016] [Souza 2010].

Em seu estudo, [Jacob e Vert 2008] formulou o problema de quimiogenômica *in silico* típico como o seguinte problema de aprendizagem: dada uma coleção de n pares alvo/molécula $[(t_1, c_1], ..., (t_n, c_n)]$ conhecidos por formar complexos ou não, estimar uma função $f(t, c)$ poderia prever se qualquer produto químico c se liga a qualquer alvo t.

O algoritmo *Pairwise kernel method (PKM)* [Jacob e Vert 2008] configura esta abordagem, de tal forma que a matriz de kernel de pares é calculada a partir das similaridades dos nós, i.e., $K_E\{(d, p), (d', p') = K_d(d, d') \times K_p(p, p')\}$, representado na Figura 7. Esse método utiliza o algoritmo *SVM (Support Vector Machines)* para treinar o modelo com a matriz de kernel de pares obtida. Contudo, essa abordagem apresenta limitações em relação ao custo computacional de memória para a construção do kernel de pares, inviabilizando,

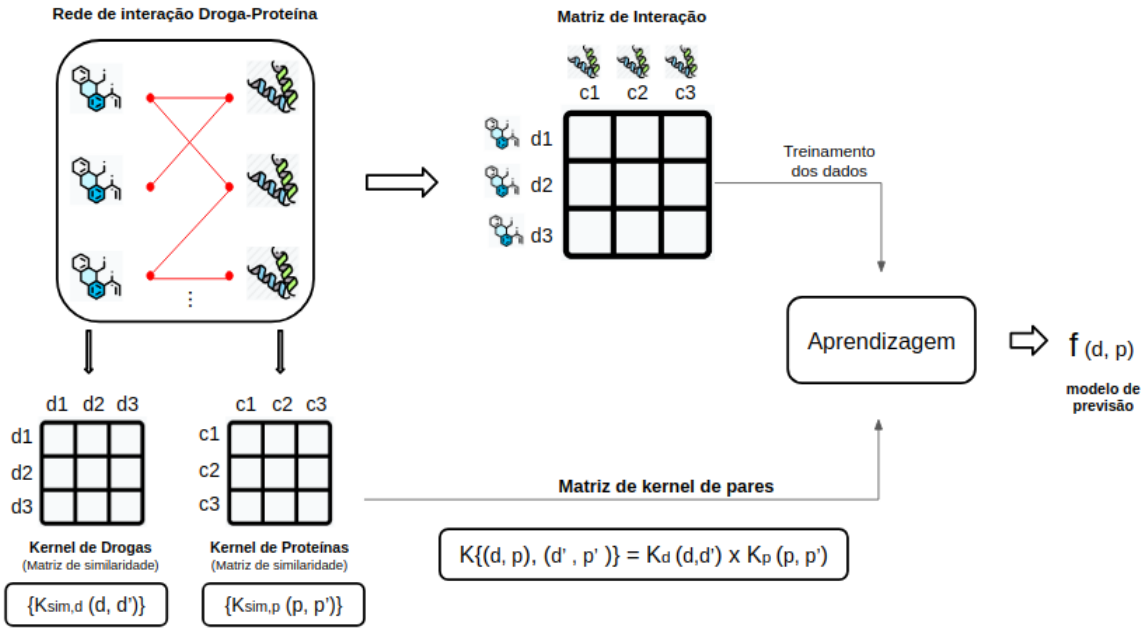


Figura 7 – Figura representativa do Método de pares de kernel. Figura inspirada em: Pairwise kernel method (PKM) (JACOB; VERT, 2008).

portanto, o uso de todo o conjunto de pares possíveis no conjunto de treinamento.

Solucionando as limitações computacionais, o algoritmo pairwiseMKL, desenvolvido por [Cichonska et al. 2018], é um método de aprendizagem eficiente em relação ao tempo e memória quando se utiliza múltiplos kernels de pares. Inicialmente, o pairwiseMKL determina os pesos da mistura dos kernels emparelhados de entrada, e em seguida, aprende a função de predição aos pares.

A eficiência do método é tida com a realização de um procedimento de alinhamento de kernel. Ou seja, é feita uma decomposição do *Kronecker* do operador de centralização para o kernel par a par [Cichonska et al. 2018]. Por definição matemática, o produto *Kronecker* é uma operação feita em duas matrizes, de tamanho arbitrário, resultando em uma matriz de bloco [Paleologu, Benesty e Ciochină 2018].

O desempenho e sucesso do pairwiseMKL é devidamente alcançado pois as etapas de determinação dos pesos e aprendizagem da função de predição são realizadas de forma eficiente, tornando o método aplicável para resolver grandes problemas de aprendizagem de pares [Cichonska et al. 2018]. Ultimamente, a aprendizagem de múltiplos *kernels (Multiple Kernel Learning - MKL)* vem sendo utilizada para solucionar este problema, permitindo o uso de múltiplos kernels, ao invés de considerar apenas um kernel para uma dada tarefa.

[Cichonska et al. 2018] demonstra em seu estudo que o pairwiseMKL gera previsões assertivas usando soluções esparsas em termos de kernel selecionados e, consequentemente, identifica automaticamente fontes de dados relevantes para o problema de previsão. As métricas avaliativas utilizadas foram o Erro de Raiz Quadrada Média (*Root Mean Square Error - RMSE*), Correlação de Pearson e Pontuação F1 (*F1-Score*) entre os valores de bioatividade original e previsto.

Portanto, por apresentar resultados promissores e capacidade de desempenho computacional satisfatório, o pairwiseMKL foi escolhido para compor o experimento descrito neste trabalho.

3.1 PairwiseMKL

Nesta seção apresentamos e descrevemos o funcionamento do método pairwiseMKL, desenvolvido por [Cichonska et al. 2018].

Kernels podem ser utilizados no modelo de aprendizagem por pares mediante a construção de uma matriz de kernel de pares $K \in \mathbb{R}^{N \times N}$ relacionando todos os pares de matrizes de drogas e células, formando o par droga-célula. K é calculado, especificamente, como um produto *Kronecker* do kernel de droga $K_d \in \mathbb{R}^{n_d \times n_d}$ e kernel de linha celular $K_c \in \mathbb{R}^{n_c \times n_c}$, formando uma matriz de bloco com o produto dos kernels de entrada K_d e

K_c:

$$K = K_d \otimes K_c = \begin{pmatrix} K_d(X_{d1}, X_{d1})K_c & K_d(X_{d1}, X_{d2})K_c & \cdots & K_d(X_{d1}, X_{dn_d})K_c \\ K_d(X_{d2}, X_{d1})K_c & K_d(X_{d2}, X_{d2})K_c & \cdots & K_d(X_{d2}, X_{dn_d})K_c \\ \vdots & \vdots & \ddots & \vdots \\ K_d(X_{dn_d}, X_{d1})K_c & K_d(X_{dn_d}, X_{d2})K_c & \cdots & K_d(X_{dn_d}, X_{dn_d})K_c \end{pmatrix} \quad (3.1)$$

Logo, a função de previsão para um par de teste (X_d, X_c) é descrita como

$$f\left(X_d, X_c\right) = \sum_{l=1}^{N} \alpha_l k\left(\left(X_{dl}, X_{cl}\right), \left(X_d, X_c\right)\right) = \alpha^T k, \quad (3.2)$$

onde k é um vetor de coluna com os valores de kernel entre cada par de treinamento droga-linha celular (X_{dl}, X_{DL}) e par de teste (X_d, X_c) para o qual a previsão é feita, e $\alpha = (\alpha 1, ..., \alpha N)$ denota um vetor de parâmetros do modelo a serem obtidos pelo algoritmo de aprendizagem por meio da minimização de uma determinada função objetivo.

Na *regressão kernel ridge (KRR)*, [Saunders, Gammerman e Vovk 1998]), a função objetivo é construída baseada na perda quadrática total associada ao regularizador de norma-L2. E a solução para α é encontrada no desenvolvimento da seguinte equação:

$$(K + \lambda I)\alpha = y, \quad (3.3)$$

onde λ corresponde ao hiperparâmetro de regularização (norma-L2) que regula o equilíbrio entre a complexidade do modelo e o erro de treinamento ($\lambda > 0$); e I representa a matriz identidade $N \times N$.

3.1.1 Otimização dos pesos dos pares de kernel

Visando otimizar os pesos dos kernels para o aprendizado par a par, [Cichonska et al. 2018] explorou a identidade conhecida

$$\left\langle K_d \otimes K_c, K'_d \otimes K'_c \right\rangle = \left\langle K_d, K'_d \right\rangle \left\langle K_c, K'_c \right\rangle, \quad (3.4)$$

com o intuito de evitar o cálculo de matrizes massivas. Corrigindo limitações da centralização do kernel de pares, [Cichonska et al. 2018] apresenta uma nova e eficiente decomposição de *Kronecker* do operador de centralização de kernel em pares:

$$C = \sum_{q=1}^{2} {Q_d}^{(q)} \otimes {Q_c}^{(q)}, \quad (3.5)$$

onde ${Q_d}^{(q)} \in \mathbb{R}^{n_d \times n_d}$ e ${Q_c}^{(q)} \in \mathbb{R}^{n_c \times n_c}$ são os fatores de C.

Após sucessivas decomposições das equações, verificou-se que o kernel da resposta linear padrão é adequado para medir semelhanças entre rótulos em tarefas de classificação onde $y \in -1, +1$, porém não é adequado para regressão, onde $y \in \mathbb{R}$. Por isso, [Cichonska et al. 2018] aplica um kernel de resposta gaussiana, considerado um padrão ouro na verificação de semelhanças entre números reais [Shawe-Taylor, Cristianini et al. 2004].

Inicialmente, cada valor de rótulo $y_i, i = 1, ..., N$ é associado com um vetor de característica de comprimento S. O vetor de característica, por sua vez, é representado como um histograma correspondente a uma função de densidade de probabilidade entre todos os rótulos y, centralizado em y_i e armazenado como um vetor de linhas na matriz $\Psi \in \mathbb{R}^{N \times S}$. Desse modo, o kernel da resposta gaussiana faz uma comparação de todos os pares de rótulos do vetor de característica, resultando na soma dos produtos internos de S:

$$K_y = \sum_{s=1}^{S} \psi^{(S)} \psi^{(S)T}, \tag{3.6}$$

onde $\psi^{(S)} \in \mathbb{R}^N$ é um vetor coluna de Ψ.

3.1.2 Treinamento de modelos em pares

Compondo os pesos dos pares de kernel μ na Equação 3.3 de KRR em pares, tem-se a seguinte expressão:

$$\left(\mu_1 K_d^{(1)} \otimes K_c^{(1)} +, ..., +\mu_p K_d^{(P)} \otimes K_c^{(P)} + \lambda I\right)\alpha = y. \tag{3.7}$$

Contudo, valores ausentes na matriz de rótulo $Y \in \mathbb{R}^{n_c \times n_d}, vec(Y) = y$ podem ocorrer devido ao fato de que as bioatividades de todas as combinações de drogas e linhas celulares podem não ser conhecidas, então temos

$$U\alpha = y, \tag{3.8}$$

$$U = B\left(\mu_1 K_d^{(1)} \otimes K_c^{(1)} +, ..., +\mu_p K_d^{(P)} \otimes K_c^{(P)} + \lambda I\right)B^T, \tag{3.9}$$

onde B é a matriz de indexação que representa as associações entre as linhas e as colunas da matriz de kernel e os elementos do vetor $\alpha : B_{il} = 1$ denota que o coeficiente α_i corresponde à l^{th} linha/coluna na matriz de kernel.

De forma análoga a resolução do sistema de equações lineares acima (3.9), é possível encontrar os parâmetros α da função de predição de pares treinando o modelo. [Cichonska et al. 2018] resolvem o sistema utilizando o gradiente conjugado (*CG - Conjugate Gradient*),

fazendo os produtos matriz-vetor entre U e α em um número de iterações proporcionais ao quantitativo de dados, o que melhora iterativamente o resultado.

3.1.3 Descrição do método

A figura 8 apresenta graficamente a descrição do funcionamento do algoritmo. Inicialmente, dois núcleos de droga e três núcleos de proteínas são calculados a partir de bases de dados químicos e genômicos, respectivamente. As matrizes geradas a partir desta interação associam todos os medicamentos a todos os alvos celulares e, portanto, um kernel pode ser considerado como uma medida de similaridade. Uma vez que se está interessado em aprender sobre a bioatividade de pares de objetos de entrada, — no presente estudo, pares de linhagem droga-proteína — são necessários núcleos par a par relacionando todos os kernels de drogas com os kernels de proteínas, sendo eles calculados como produtos *Kronecker* ($\otimes$) de núcleos de drogas e núcleos de proteínas (2 núcleos de drogas 3 núcleos de linha celular = 6 núcleos de pares). No primeiro estágio de aprendizagem, os pesos de mistura de kernel em pares são determinados e, em seguida, uma combinação ponderada de kernels em pares é usada para predição de resposta a drogas anticâncer com um modelo de regressão em pares de mínimos quadrados regularizado. É importante ressaltar que pairwiseMKL executa essas duas etapas de forma eficiente, evitando a construção explícita de quaisquer matrizes massivas pareadas e, portanto, é muito adequado para resolver grandes problemas de aprendizagem pareada.

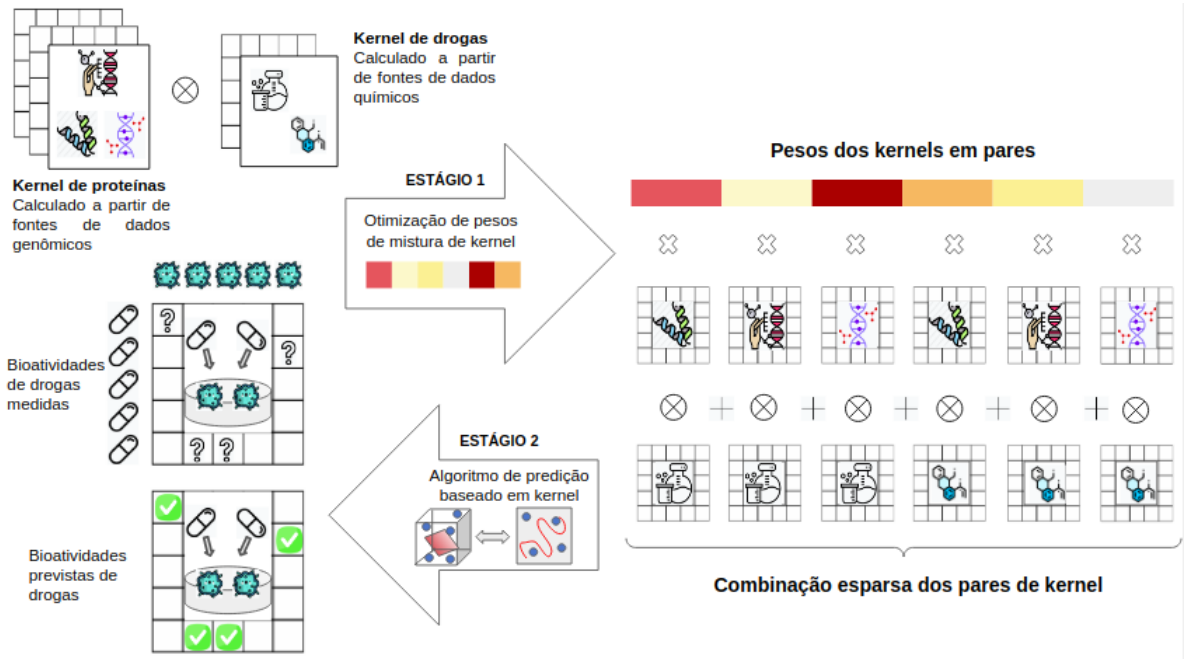


Figura 8 – Figura esquemática que representa a visão geral do método pairwiseMKL usando como exemplo a resposta à droga na predição de proteínas expressas em linhagens de câncer, baseado em múltiplos pares de kernel. Fonte: figura adaptada de [Cichonska et al. 2018]

Em resumo, o pairwiseMKL, no primeiro momento, realiza um procedimento de alinhamento de kernel centralizado com o intuito de evitar o cálculo explícito de várias matrizes de pares — geralmente grandes — na seleção dos pesos de mistura dos kernels de pares de entrada. Para isso, [Cichonska et al. 2018] fez uma nova decomposição do *Kronecker* do operador de centralização para o kernel par a par. Ou seja, o algoritmo gera uma medida de similaridade de matriz entre o kernel final e o kernel ideal, derivado dos valores de rótulo (kernel de reposta gaussiana), a partir de uma combinação convexa de kernels em pares de entrada. Ou seja, essa abordagem torna o método adequado para resolução de problemas no contexto de grandes espaços pareados, que é o caso da previsão de bioatividade de drogas.

3.2 Metodologia experimental

Particularmente neste domínio, o algoritmo do pairwiseMKL realiza 10 validações cruzadas externas e 3 validações cruzadas internas, utilizando 15.376 respostas de drogas em linhas de células de câncer, construindo um total de 120 kernels pareados de droga-linha celular a partir de 10 núcleos de droga (Tabela 2) e 12 núcleos de linhagem celular (Tabela 3). Após gerar a medida de similaridade, a função de previsão aos pares é aprendida pelo algoritmo (Figura 8).

Neste trabalho realizamos uma adaptação simples no algoritmo original do pairwiseMKL, onde retiramos a validação cruzada interna (*3-fold-cross-validation*) e trabalhamos apenas com as 10 validações cruzadas externas do algoritmo. A adaptação foi proposta com o intuito de reduzir o tempo de processamento do experimento.

3.2.1 Base de dados

No experimento foi utilizado um conjunto de base de dados de interação de bioatividade de medicamentos, ou seja, que representa a resposta a medicamentos em linhas celulares de câncer. Esse conjunto de dados já foi utilizado em estudos anteriores [Cichonska et al. 2018], e proposto inicialmente por [Yang et al. 2012]. A base de dados utilizada está disponível no seguinte endereço: *https://github.com/TassiaBastos/Treatment-of-incomplete-kernels*.

3.2.1.1 Dados de bioatividade de drogas

No contexto de respostas a medicamentos em linhas celulares de câncer, o conjunto de dados utilizado no estudo consiste em 124 medicamentos e 124 linhas celulares de câncer humano, para as quais estão disponíveis medidas completas de sensibilidade de $124 \times 124 = 15.376$ na forma $ln(IC_{50})$, em valores nano molares [Ammad-Ud-Din et al. 2016].

A figura 9 apresenta o histograma da distribuição dos valores de bioatividade da base de dados utilizada. Cada coluna da matriz representa uma classe no histograma. Na imagem é possível identificar que a distribuição dos dados segue a distribuição normal, tendo a maior concentração dos dados entre o intervalo de afinidade 0 e 5.

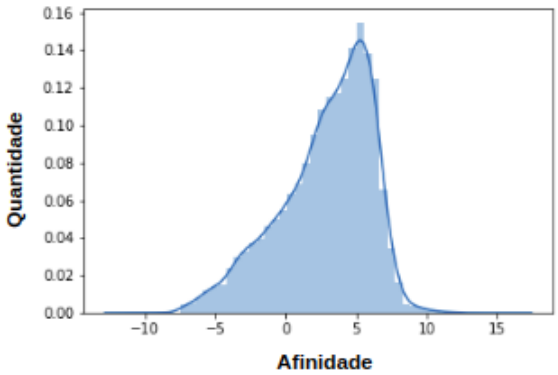


Figura 9 – Histograma de afinidade de interação. A afinidade apresenta uma distribuição normal, centrada no intervalo entre 0 e 5.

Foram utilizados dados de resposta a drogas anticâncer do projeto *GDSC (Genomics of Drug Sensitivity in Cancer)* iniciado pelo *Wellcome Trust Sanger Institute* [Yang et al. 2012]. A tabela 1 apresenta de forma resumida as cinco primeiras linhas da tabela obtida do projeto *GDSC (Genomics of Drug Sensitivity in Cancer)*, onde é possível identificar os nomes, tipo e tecido das 124 linhas de celulares de câncer e os nomes dos 124 medicamentos utilizados no estudo.

Cell line	Type	Tissue	Drug
ES3	Bone	ewings sarcoma	Erlotinib
ES5	Bone	ewings sarcoma	Rapamycin
EW-11	Bone	ewings sarcoma	Sunitinib
NCI-H1395	Lung	lung NSCLC adenocarcinoma	PHA-665752
NCI-H1770	Lung	lung NSCLC not specified	MG-132

Tabela 1 – Representação das cinco primeiras linhas da tabela obtida do projeto *GDSC (Genomics of Drug Sensitivity in Cancer)*. Para visualização completa da tabela, acessar o material suplementar de [Ammad-Ud-Din et al. 2016].

3.2.1.2 Kernels

Nas subseções a seguir serão apresentados os processos de construção dos kernels de drogas e de linhas celulares utilizadas no modelo de previsão pairwiseMKL e também apresentamos duas tabelas com representações resumidas dos cálculos dos kernels de droga e dos kernels de linha celular.

3.2.1.2.1 Kernels de droga

Para o contexto de Kernels de drogas, a sua construção foi baseada no kernel de Tanimoto — uma ferramenta utilizada para descrever a similaridade entre conjuntos de atributos binários [Szedmak e Bach 2020], onde foram calculados usando 10 impressões digitais *(fingerprints)* moleculares diferentes (Tabela 2), ou seja, vetores binários que indicam a presença ou ausência de diferentes subestruturas na molécula, como representado na expressão [Guha et al. 2007]: $k_d(x_d, x'_d) = \frac{H_{x_d}, x'_d}{H_{x_d} + H'_{x_d} - H_{x_d}, x'_d}$, onde H_{x_d} é o número de 1-bit na impressão digital da droga x_d, e H_{x_d}, x'_d corresponde ao número de 1-bit comum a impressões digitais de duas moléculas de drogas x_d e x'_d que estão sendo comparadas entre si [Cichonska et al. 2018].

Abreviação dos kernels de droga	**Descrição do recurso e tipo de kernel**
Kd-circular	Conectividade estendida Impressão digital de 1 024 bits com diâmetro máximo definido para 6 (ECFP6).
Kd-estate	Impressão digital de 79 bits correspondente às subestruturas 'Estate' descritas por Hall e Kier (1995).
Kd-ext	Impressão digital com blocos de 1.024 bits baseada em caminho, levando em consideração os sistemas de anel.
Kd-graph	Impressão digital com blocos de 1.024 bits baseada em caminho, considerando a conectividade.
Kd-hybr	Impressão digital com blocos de 1.024 bits baseada em caminho, considerando os estados de hibridização.
Kd-kr	Impressão digital de 4.860 bits definida por Klekota e Roth (2008).
Kd-maccs	Impressão digital de 166 bits baseada em chaves estruturais MACCS desenvolvidas por MDL Information Systems.
Kd-PubCh	Impressão digital de 881 bits definida por PubChem.
Kd-sp	Impressão digital de 1.024 bits com base nos caminhos mais curtos entre os átomos, levando em consideração sistemas de anéis e cargas.
Kd-std	Impressão digital com blocos de 1024 bits baseada em caminho.

Tabela 2 – Representação resumida dos cálculos dos núcleos de droga e dos núcleos da linha celular utilizados no modelo de previsão pairwiseMKL. Fonte: [Cichonska et al. 2018].

3.2.1.2.2 Kernels de linha celular

Para o contexto de Kernels de linhas de células, foram utilizados e calculados kernels gaussianos (Tabela 3) $k_c(x_c, x'_c) = exp\left(\frac{-\left\|x_c - x'_c\right\|^2}{2\rho_c^2}\right)$, onde x_c e x'_c correspondem

a representação de característica de duas linhas celulares que podem ser: a medida de expressão gênica; o padrão de metilação; a variação do número de cópias de genes ou o perfil de mutação somática; e ρ_c indica um hiperparâmetro de largura do kernel [Cichonska et al. 2018].

[Cichonska et al. 2018] derivou vetores de características de perfil de mutação de valor real, ao invés de empregar indicadores de mutação binários, que são comumente utilizados. Nesse estudo, cada elemento x_{c_i}, $i = 1, ..., M$, corresponde a uma das M mutações. Ou seja, se uma linha celular representada por x_c tiver um status de mutação i negativo, então $x_{c_i} = 0$; caso contrário x_{c_i} indica um logaritmo negativo de proporção de todas as linhas de células com status de mutação positiva. Dessa forma, x_{c_i} apresenta valor elevado para uma mutação específica em uma linha celular representada por x_c, dando mais importância a tal variante genética.

Abreviação dos kernels de células	**Descrição do recurso e tipo de kernel**
Kc-cn-146	Medições da variação do número de cópias de 43.255 genes, com hiperparâmetro de largura do kernel gaussiano (σ) = 146.
Kc-cn-270	Medições da variação do número de cópias de 43.255 genes, com hiperparâmetro de largura do kernel gaussiano (σ) = 270.
Kc-cn-417	Medições da variação do número de cópias de 43.255 genes, com hiperparâmetro de largura do kernel gaussiano (σ) = 417.
Kc-exp-147	Medições de expressão gênica basal de 13.321 genes, com hiperparâmetro de largura do kernel gaussiano (σ) = 147.
Kc-exp-163	Medições de expressão gênica basal de 13.321 genes, com hiperparâmetro de largura do kernel gaussiano (σ) = 163.
Kc-exp-177	Medições de expressão gênica basal de 13.321 genes, com hiperparâmetro de largura do kernel gaussiano (σ) = 177.
Kc-met-176	Níveis de metilação de 482.892 ilhas de CpG, com hiperparâmetro de largura do kernel gaussiano (σ) = 176.
Kc-met-210	Níveis de metilação de 482 892 ilhas de CpG, com hiperparâmetro de largura do kernel gaussiano (σ) = 210.
Kc-met-252	Níveis de metilação de 482.892 ilhas de CpG, com hiperparâmetro de largura do kernel gaussiano (σ) = 252.
Kc-mut-57	Perfis de valor real de 12 366 mutações somáticas, com hiperparâmetro de largura do kernel gaussiano (σ) = 57.
Kc-mut-71	Perfis de valor real de 12 366 mutações somáticas, com hiperparâmetro de largura do kernel gaussiano (σ) = 71.
Kc-mut-132	Perfis de valor real de 12 366 mutações somáticas, com hiperparâmetro de largura do kernel gaussiano (σ) = 132.

Tabela 3 – Representação resumida dos cálculos dos núcleos de droga e dos núcleos da linha celular utilizados no modelo de previsão pairwiseMKL. Fonte: [Cichonska et al. 2018].

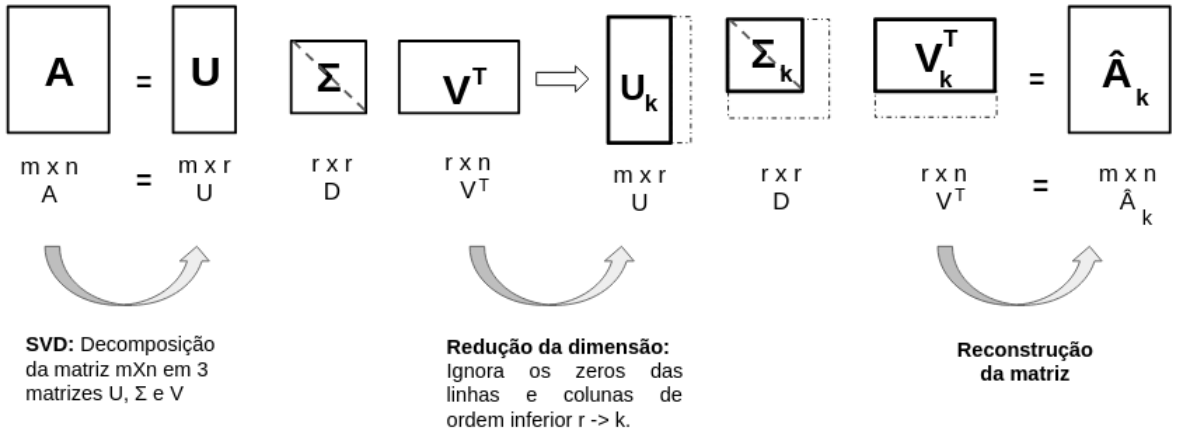


Figura 10 – Decomposição em Valores Singulares. Adaptado de [Bougiatiotis e Giannakopoulos 2018]

3.2.2 Técnicas

Neste trabalho foram avaliadas três técnicas de imputação de valor único (média, mediana e zero) e uma técnica de imputação supervisionada (*Single-value Decomposition, SVD*) em um conjunto de dados com 10 matrizes de droga e 12 matrizes de linhas celulares. Esses conjuntos de dados foram primeiro pré-processados gerando posições ausentes nas matrizes para posterior completude com as técnicas escolhidas. A escolha das técnicas foram determinadas pelo alto índice de utilização em outros trabalhos e por terem resultados validados na literatura.

As primeiras três técnicas utilizadas correspondem a imputação de valores simples — média, mediana e zero. Na técnica média foi calculada a média de cada matriz e cada resultado foi utilizado para preencher as lacunas das matrizes respectivamente. A técnica mediana possui um processo de desenvolvimento semelhante à técnica média, mas nesse caso utilizando o valor mediano da matriz para preenchimento dos espaços faltosos [Wei et al. 2018]. A terceira e última técnica de imputação simples, compreende preencher as lacunas diretamente com o número zero [Tuikkala et al. 2008].

A quarta técnica — Decomposição em Valores Singulares (SVD) — é tida como a base dos métodos mais precisos quando o objetivo é a resolução de problemas de mínimos quadrados, e especialmente para determinação do espaço nulo de matrizes. O SVD é o método de decomposição/fatoração de matrizes mais confiável, porém sua utilização demanda um maior tempo de execução [Yuan et al. 2019]. Visando melhorar o desempenho do SVD, [Kurucz András A. Benczúr 2007] traz uma modificação na implementação tradicional do código Lanczos, que permite a imputação de dados ausentes assim como o manuseio de bases de entradas muito grandes. Neste trabalho, utilizamos a técnica de SVD melhorada proposta por [Kurucz András A. Benczúr 2007].

O funcionamento da técnica consiste em decompor a matriz no produto dos fatores de outras matrizes (Figura 10). Por exemplo, toda matriz A, $m \times n$ pode ser fatorada da seguinte forma:

$$A = U\Sigma V^T \tag{3.10}$$

Sendo essa etapa chamada de Decomposição em valores singulares da matriz A, onde: V denota uma matriz ortogonal $n \times n$, construída a partir de um conjunto ortogonal de auto-vetores da matriz $A^T A$, $v_1, v_2, ..., v_n$ e U corresponde a uma matriz ortogonal $m \times m$, onde os elementos são determinados por:

$$u_i = \frac{1}{\rho_i} A v_i \tag{3.11}$$

Se A tem r valores singulares não nulos, Σ é uma matriz da forma:

$$\Sigma = \begin{pmatrix} & & & 0 & \cdots & 0 \\ & D & & \vdots & \ddots & \vdots \\ & & & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}, \text{ em que D} = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{pmatrix}$$

A matriz A, $m \times n$, com r valores singulares não nulos, pode ser escrita como:

$$A = \rho_1 u_1 {v_1}^T + \cdots + \rho_r u_r {v_r}^T = \sum_{i=1}^{r} \rho_i u_i {v_i}^T \tag{3.12}$$

em que u_i são as colunas de U, ditos vetores singulares à esquerda, e v_i são colunas de V, ditos vetores singulares à direita.

3.2.3 Métricas avaliativas

Medir a qualidade de um modelo de aprendizagem de máquina de acordo com o objetivo proposto é fundamental durante o seu processo de desenvolvimento. Saber escolher as métricas corretas é muito importante para uma avaliação correta e precisa [Sun, Lai e Pei 2018]. Alguns fatores devem ser levados em consideração ao escolher as métricas como o objetivo da previsão e a proporção de dados de cada classe na base de dados, por exemplo. Para esse trabalho, escolhemos 3 métricas descritas a seguir.

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 utilizado na avaliação de sistemas de classificação binários, categorizando os exemplos em dois grupos: "positivos"ou "negativos". Ele é definido como a média harmônica do *recall* e da precisão do modelo [Dalianis 2018]. A precisão é descrita como:

$$P = \frac{tp}{tp + fp} \tag{3.13}$$

e mede o número de instâncias corretas recuperadas dividido por todas as instâncias recuperadas [Dalianis 2018]. O Recall mede o número de instâncias corretas recuperadas dividido por todas as instâncias corretas e é definida matematicamente como:

$$R = \frac{tp}{tp + fn} \tag{3.14}$$

. Logo, o F1-Score é descrito como:

$$F - score : F1 = F = 2 \times \frac{P \times R}{P + R} \tag{3.15}$$

A segunda foi o coeficiente de correlação de Pearson (r), que corresponde ao grau de associação linear entre duas variáveis quantitativas (Equação 3.16) [Liu et al. 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.

$$r = \frac{\sum_i (x_i - x)(y_i - y)}{\sqrt{\sum_i (x_i - x^2} \sum_i (y_i - y^2} \tag{3.16}$$

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 aumentar ou diminuir junto a outra variável, e valores negativos indicam uma tendência de que o aumento dos valores de uma variável esteja associado à redução dos valores da outra variável e vice-versa. Valores próximos de zero indicam baixa associação entre os dois conjuntos de dados. [Kirch 2008]

E por fim, escolhemos o erro de raiz quadrática médio (*RSME - Root Mean Squared Error*), que corresponde a raiz quadrada da média do quadrado de todos os erros (Equação 3.17),

$$RSME = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (S_i - O_i)^2} \tag{3.17}$$

onde, O_i determinam as observações, S_i 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 e Hashemi 2018]

3.2.4 Estrutura do método desenvolvido

A fim de avaliar o desempenho do algoritmo pairwiseMKL proposto por [Cichonska et al. 2018] no contexto de redes bipartidas, foi realizado um experimento sistemático

com o intuito de avaliar a eficácia do método quando se é utilizado um conjunto de dados biológicos heterogêneos incompletos como entrada. Após uma ampla análise dos métodos e técnicas apresentadas na literatura, a metodologia desenvolvida no presente estudo contém três fases: pré-experimental, experimental - fase I e experimental - fase II.

Um ponto fundamental que deve ser levado em consideração para o planejamento e desenvolvimento do experimento é que, nesse estudo, estamos tratando de bases de dados biológicos incompletos, ou seja, dados que possuem um alto grau de heterogeneidade. Por isso, simulamos esse ambiente criando lacunas nas matrizes, ou seja, apagando os dados contidos em posições conhecidas e preenchendo os espaços aplicando as técnicas escolhidas.

A primeira fase corresponde a etapa pré-experimental (Figura 11), onde destacamos:

- Escolha/identificação das matrizes de droga e de células utilizadas no trabalho, onde a base de entrada de dados do algoritmo pairwiseMKL contempla 12 kernels de células e 10 kernels de drogas. A descrição completa e detalhada da base de dados utilizada está na seção 3.4 deste trabalho;
- Escolha das técnicas para preenchimento de valores faltosos, das quais destacamos as técnicas escolhidas e aplicadas no presente estudo: média, mediana, zero e SVD.
- Escolha do percentual para simulação de valores faltosos.

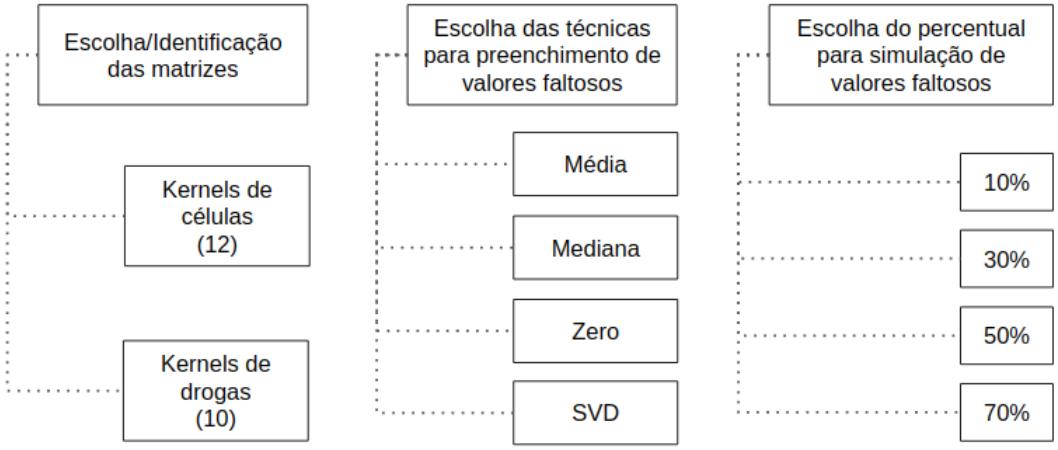


Figura 11 – Esquema figurativo da etapa pré-experimental. Descrição de escolha das matrizes, das técnicas de preenchimento e dos percentuais aplicados.

A segunda etapa corresponde ao experimento - fase I (Figura 12), onde foi desenvolvido um *script* em *Python*, *Matrix Cleaner*, que é responsável por todo o processo de verificação, limpeza e aplicação das técnicas associadas às porcentagens nas matrizes de entrada. O *Matrix Cleaner* funciona da seguinte forma: o algoritmo recebe uma matriz de entrada, de droga ou de célula, verificando se ela é uma matriz quadrada e se tem

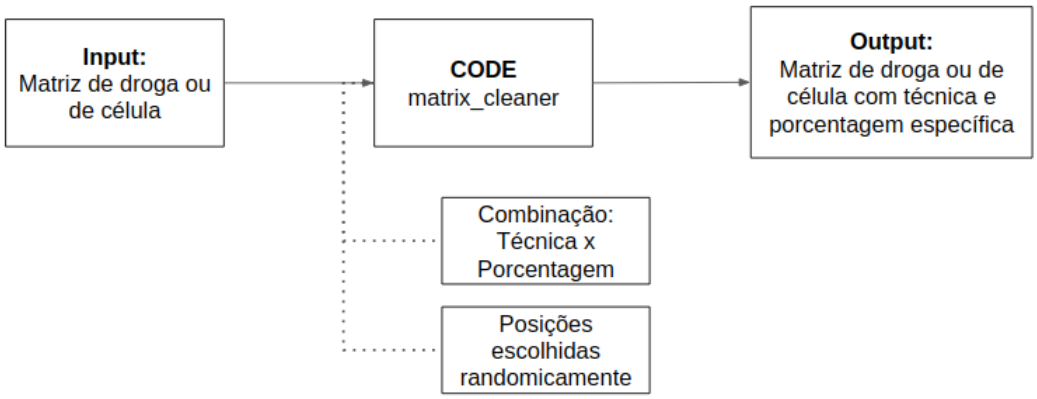


Figura 12 – Figura representativa da etapa de experimento - fase I, que descreve o funcionamento do *script Matrix Cleaner*. A partir do conjunto de dados (de droga ou de célula), o *Matrix Cleaner* recebe uma entidade (kernel de droga ou de células) como entrada e é passado como parâmetro um valor percentual (10%, 30%, 50% ou 70%) para que posições aleatórias escolhidas proporcionalmente ao percentual fornecido sejam apagadas e preenchidas com uma das técnicas (média, mediana ou zero), gerando uma matriz modificada como saída.

cabeçalho ou não. Caso a tenha, o algoritmo retira essas informações da matriz de entrada, restando apenas os dados a serem analisados. Importante destacar que todas as 22 matrizes (12 de células e 10 de drogas) da base de dados utilizadas nesse estudo são processadas individualmente como dados de entrada para o *Matrix Cleaner*.

Ainda referente ao experimento - fase I, a etapa de combinação entre a técnica e a porcentagem aplicada a uma matriz de entrada ocorre de forma similar, independentemente se a matriz for de droga ou de célula (Figura 13), e é descrita a seguir. A partir do conjunto de dados, o algoritmo *Matrix Cleaner* recebe uma entidade (kernel de droga ou proteína) como entrada e é passado como parâmetro um valor percentual (10%, 30%, 50% ou 70%) para que posições aleatórias escolhidas proporcionalmente ao percentual fornecido sejam apagadas e preenchidas com uma das técnicas (média, mediana ou zero). Por exemplo, escolhemos a entidade K_{c1}, referente ao primeiro kernel de células, como entrada e passamos como parâmetro a técnica média e o percentual de 10%. O *script* apresentará como saída a matriz de entrada K_{c1} modificada, agora K_{c1}_media_10, com 10% das posições da matriz preenchidos com a da média da matriz. Esse procedimento é similar à base de dados com as matrizes de drogas, sendo o processo (Figura 12) executado 352 vezes, devido a combinação entre as 4 técnicas escolhidas e as 4 porcentagens de deleção, para cada uma das 22 matrizes do conjunto total de kernels. Tal processo descrito representa uma iteração completa, sendo que, no total, foram realizadas 3 iterações/repetições com o intuito de obter resultados mais confiáveis estatisticamente. Importante salientar que todas as execuções do *Matrix Cleaner* são feitas de forma independente, ou seja, cada matriz de

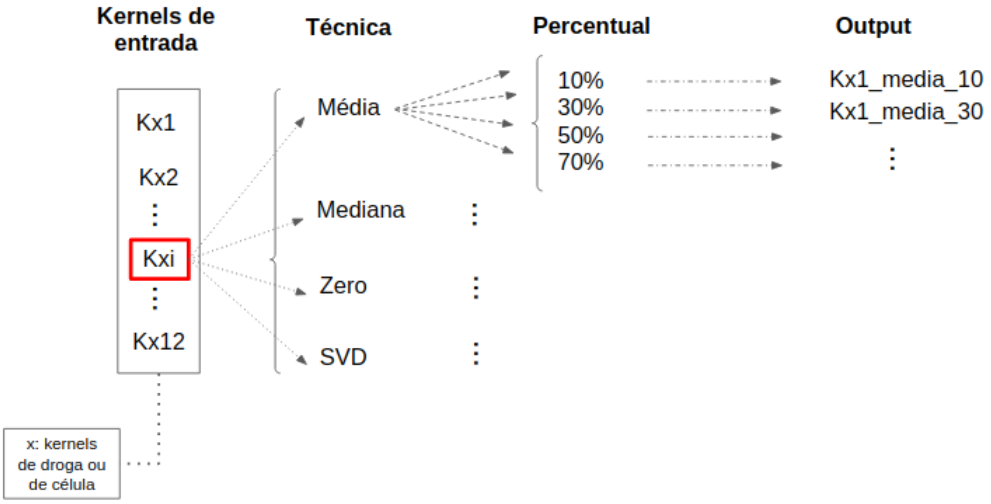


Figura 13 – Descrição da etapa de combinação: técnica X porcentagem para as matrizes de droga ou célula. Por exemplo, se escolhermos a entidade K_{c1}, como entrada e passarmos como parâmetro a técnica média e o percentual de 10%, o *script* apresentará como saída a matriz de entrada K_{c1} modificada, agora K_{c1}_media_10, com 10% das posições da matriz preenchidos com a da média da matriz.

kernel processada pelo *script* terá a escolha das posições feita de forma aleatória.

A terceira e última etapa compreende ao experimento - fase II (Figura 14), onde os kernels de droga e de proteínas gerados na fase anterior serão a base e entrada para o algoritmo pairwiseMKL. Em sua estrutura de arquivos, o pairwiseMKL apresenta uma pasta onde armazena a base de dados utilizada na execução do algoritmo. Essa pasta contém duas subpastas, a primeira delas contendo todos os 12 kernels de células e a segunda todos os 10 kernels de drogas. Para executar o algoritmo, todos os 22 arquivos de entrada devem estar presentes nas pastas. Nesse estudo, a pasta de base de dados do pairwiseMKL foi adaptada da seguinte forma: os arquivos originais foram substituídos pelos arquivos modificados, gerados a partir da etapa de experimento - fase I, colocando as matrizes modificadas com mesma técnica e porcentagem juntas, como demonstrado na Figura 14.

O procedimento para cada associação técnica-porcentagem é feito três vezes, de modo a gerar mais confiança estatística nos resultados, uma vez que as posições são escolhidas aleatoriamente. Por fim é calculada uma média sobre dos resultados obtidos em cada uma das métricas aplicadas.

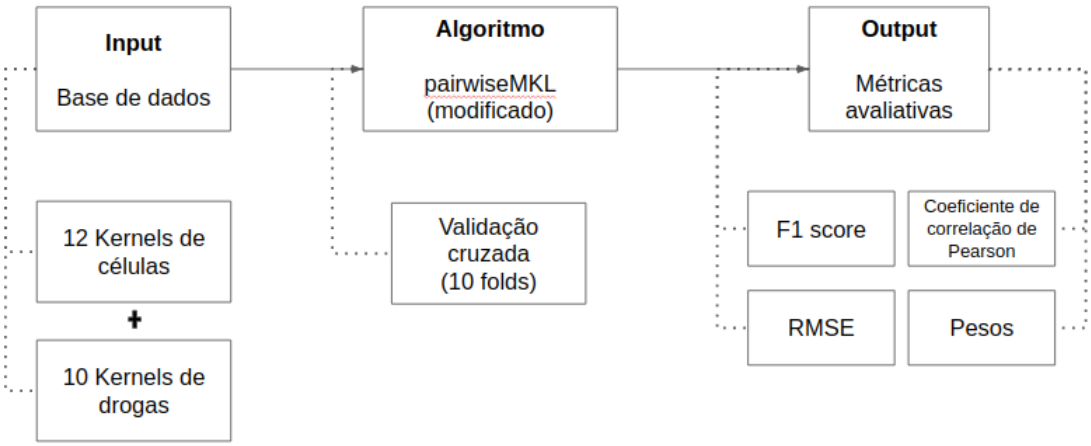


Figura 14 – Figura representativa da etapa de experimento - fase II, que descreve o funcionamento do pairwiseMKL-modificado. A partir do conjunto de matrizes com técnicas e porcentagens específicas, o algoritmo processa os dados e avalia seu desempenho com base em métricas avaliativas.

4 Análise dos resultados

Os resultados obtidos com as diferentes estratégias de preenchimento de valores faltosos em kernels de droga e de linha celular são apresentados nesta capítulo. A seção 4.1 aborda uma análise comparativa dos resultados baseada em métricas de avaliação. E a título de avaliação do padrão de comportamento dos pesos empregados pelo pairwiseMKL, a sub-seção 4.1.1 apresenta uma análise dos pesos baseada em *heatmaps*.

4.1 Análise comparativa

A figura 15 contém uma representação gráfica que analisa a relação entre a matriz real de interação e a matriz de interação predita, ou seja, duas variáveis quantitativas distintas. Na imagem, podemos observar que o modelo conseguiu aprender bem o problema, acertando razoavelmente bem em todos os *ranges* de valores reais. A relação entre as variáveis está fortemente associada à linha de tendência e apresentando poucos *outliers*, o que configura um correlação forte e positiva.

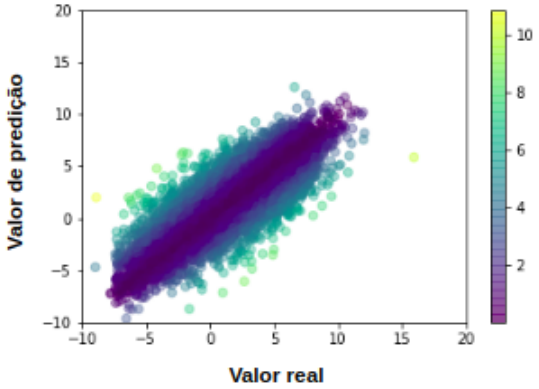


Figura 15 – Gráfico de dispersão entre os valores reais e os valores de predição.

4.1.0.1 Cenário original - caso base

Inicialmente partimos do caso base, onde utilizamos os kernels originais disponibilizados por [Cichonska et al. 2018] e executamos o algoritmo do pairwiseMKL modificado (apenas com a validação cruzada externa). Os resultados obtidos estão descritos na tabela 4.

Os resultados obtidos no cenário original foram utilizados como modelo comparativo para a avaliação das técnicas empregadas nas matrizes. Quanto mais próximo das métricas originais, melhor é classificada a técnica utilizada no pairwiseMKL.

Previsão de resposta a drogas anticâncer	F1 Score	Correlação de Pearson	RSME
PairwiseMKL	0.636589	0.837781	1.799855

Tabela 4 – Desempenho de previsão na tarefa de resposta a drogas na previsão de linha celular de câncer. As medidas de desempenho foram calculadas realizando-se a média dos 10 folds de validação cruzada externa.

4.1.0.2 Cenário modificado

Geramos valores ausentes aleatórios para cada um dos kernels de droga e linha celular nas proporções de 10%, 30%, 50% e 70%. Quatro métodos de imputação distintos foram realizados em todos os conjuntos de dados ausentes, sendo três dos métodos com imputação simples de valor (média, mediana e zero) e o quarto com um método preditivo (SVD).

A figura 17 apresenta três gráficos de linhas com o comparativo de desempenho médio entre as métricas avaliativas aplicadas e as porcentagens de valores faltosos nos conjuntos de kernels. Na figura 17-A, o desempenho do método SVD em relação ao F1-score foi superior em todas as proporções percentuais aplicadas em comparação com as demais técnicas. As técnicas média e mediana obtiveram um desempenho um pouco inferior comparado com o SVD, porém bastante próximos entre si. A média apresentou melhor resultado avaliativo nos percentuais de 30% e 50% em comparação com a mediana. A mediana apresentou melhor resultado avaliativo nos percentuais de 10% e 70% em comparação com os valores obtidos com a média. A técnica zero (figura 16-A) apresentou o pior desempenho de *F1-score* de forma gradativa, do percentual 10% até 70%.

Na figura 17-B, o desempenho do método SVD em relação ao Índice de Correlação de Pearson também foi superior em todas as proporções percentuais aplicadas em comparação com as demais técnicas. As técnicas mediana e média obtivera desempenhos semelhantes entre sim e inferior quando com parado com o SVD. A média apresentou melhor resultado avaliativo no percentual de 50% em comparação com a mediana, entretanto a mediana apresentou resultados ligeiramente superiores nas demais porcentagens (10%, 30% e 70%) em relação com a média. A técnica zero (figura 16-B) também apresentou o pior desempenho nessa métrica avaliativa, apresentando uma queda pela metade de desempenho entre as porcentagens de 10% e 70%.

A tabela 5 abaixo descreve os resultados médios das três interações relativos às técnicas utilizadas. Cada técnica apresenta uma comparação direta entre os quatro valores de porcentagens aplicadas.

Na figura 17-C, o desempenho do método SVD em relação a métrica RSME foi melhor e mais estável em todas as proporções percentuais aplicadas em comparação com as demais técnicas. Inclusive, os resultados apresentados por essa técnica foram mais bem avaliados em comparação com o resultado apresentado no cenário original nas

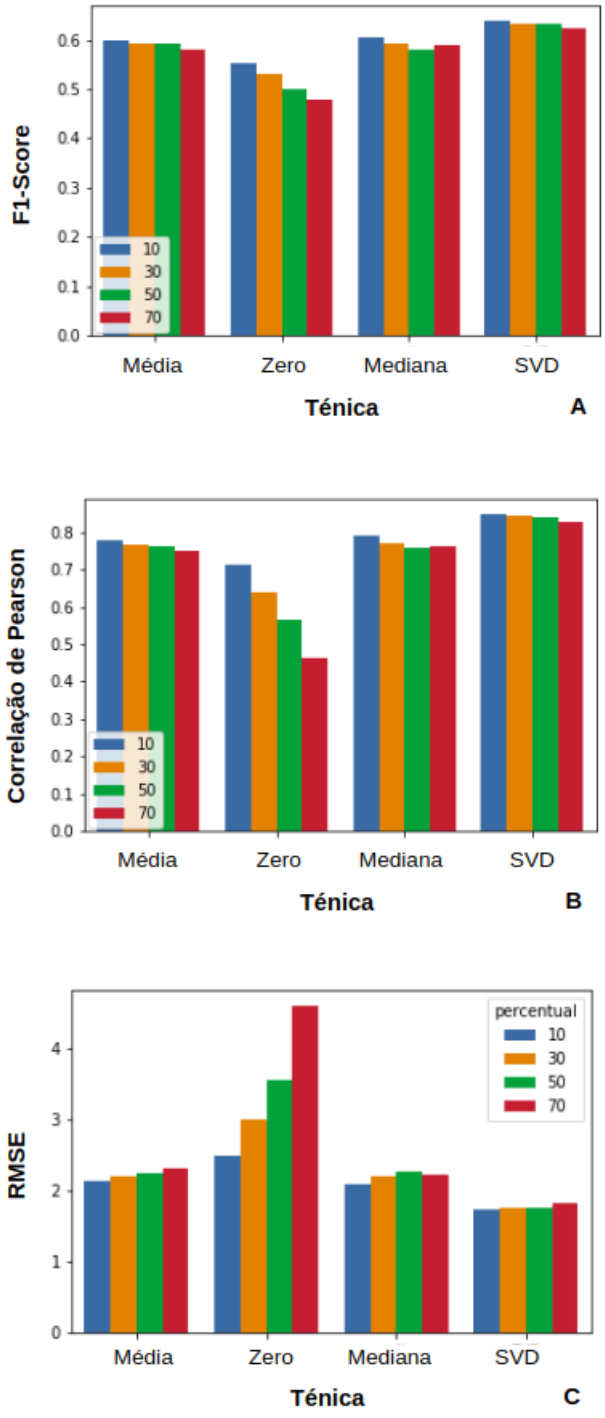


Figura 16 – Avaliação comparativa das métricas em relação aos tipos de técnicas aplicadas.

porcentagens de 10%, 30% e 50%. As técnicas média e mediana obtivera desempenhos semelhantes entre sim e inferior quando com parado com o SVD. A mediana apresentou melhor resultado avaliativo no percentual de 50% em comparação com a média, entretanto a média apresentou resultados ligeiramente superiores nas demais porcentagens (10%, 30% e 70%) em relação com a mediana. A técnica zero (figura 16-C) também apresentou o pior desempenho nessa métrica avaliativa, dobrando o valor do erro entre as porcentagens de

Técnica	**Percentual**	**F1 Score**	**Índice de correlação de Person**	**RSME**
Média	10%	0.5994 $(\pm 2.22E-16)$	0.7811 $(\pm 1.05E-16)$	2.1349 $(\pm 5.61E-16)$
Mediana		0.6038 $(\pm 2.45E-16)$	0.7904 $(\pm 1.28E-16)$	2.0856 $(\pm 1.26E-15)$
Zero		0.5509 $(\pm 1.28E-16)$	0.7150 $(\pm 8.19E-17)$	2.5032 $(\pm 5.61E-16)$
SVD		**0.6388** $(\pm 2.22E-16)$	**0.8495** $(\pm 2.80E-16)$	**1.7281** $(\pm 6.55E-16)$
Média	30%	0.5937 $(\pm 1.98E-16)$	0.7691 $(\pm 1.98E-16)$	2.2072 $(\pm 8.89E-16)$
Mediana		0.5933 $(\pm 1.98E-16)$	0.7698 $(\pm 3.39E-16)$	2.1996 $(\pm 6.08E-16)$
Zero		0.5301 $(\pm 1.40E-16)$	0.6399 $(\pm 8.19E-17)$	3.0089 $(\pm 5.14E-16)$
SVD		**0.6326** $(\pm 8.19E-17)$	**0.8435** $(\pm 3.86E-16)$	**1.7610** $(\pm 4.68E-16)$
Média	50%	0.5921 $(\pm 5.85E-17)$	0.7610 $(\pm 1.05E-16)$	2.2501 $(\pm 8.89E-16)$
Mediana		0.5813 $(\pm 2.10E-16)$	0.7571 $(\pm 4.68E-17)$	2.2704 $(\pm 7.02E-16)$
Zero		0.4997 $(\pm 1.57E-16)$	0.5655 $(\pm 1.52E-16)$	3.5581 $(\pm 1.45E-15)$
SVD		**0.6318** $(\pm 1.28E-16)$	**0.8421** $(\pm 2.69E-16)$	**1.7693** $(\pm 2.80E-16)$
Média	70%	0.5798 $(\pm 2.10E-16)$	0.7492 $(\pm 3.39E-16)$	2.3055 $(\pm 1.87E-16)$
Mediana		0.5886 $(\pm 2.10E-16)$	0.7648 $(\pm 1.98E-16)$	2.2213 $(\pm 7.95E-16)$
Zero		0.4783 $(\pm 1.34E-16)$	0.4632 $(\pm 1.40E-16)$	4.6098 $(\pm 1.59E-15)$
SVD		**0.6226** $(\pm 1.87E-16)$	**0.8292** $(\pm 1.63E-16)$	**1.8352** $(\pm 2.34E-16)$

Tabela 5 – Análise comparativa das métricas baseada na média das 3 interações para cada técnica com todos os percentuais.

10% e 70%.

Agrupando todos os percentuais relativos a cada técnica, a figura 16 apresenta três gráficos de barras com uma análise comparativa de desempenho médio entre as métricas avaliativas e as técnicas utilizadas. De forma geral, nas três métricas avaliativas, o desempenho de todas as técnicas piora de acordo com o aumento do percentual de valores faltosos das matrizes, com raras exceções, por exemplo o desempenho da técnica mediana-70% comparada com a técnica mediana-50% é um pouco melhor. De acordo com a figura, a técnica zero apresenta um padrão piora contínuo e expressivo de acordo com o aumento da porcentagem envolvida na técnica.

4.1.1 Pesos dos kernels

A análises dos pesos dos kernels atribuídos pelo algoritmo pairwiseMKL podem ser utilizados para verificar a capacidade do método em identificar corretamente as fontes de informação mais relevantes. A abordagem aqui adotada foi uma análise simples e individual dos pesos de cada combinação de kernel droga-linhas celulares, expressa em *heatmaps*.

Em relação a análise dos pesos, dado que a escolha das posições apagadas em todos os kernels foi feita de forma randômica, ou seja, nenhuma matriz possui exatamente as mesmas posições deletadas, há um impedimento na comparação direta entre os pesos das técnicas aplicadas. Entretanto, é possível observar que a distribuição dos pesos no cenário original foi bastante similar com a distribuição apresentada pela técnica SDV-10% (Figura 18). Tal distribuição semelhante pode ter influenciado no desempenho superior da técnica SVD-10% quando comparado com o cenário original (Figura 17 - A e B).

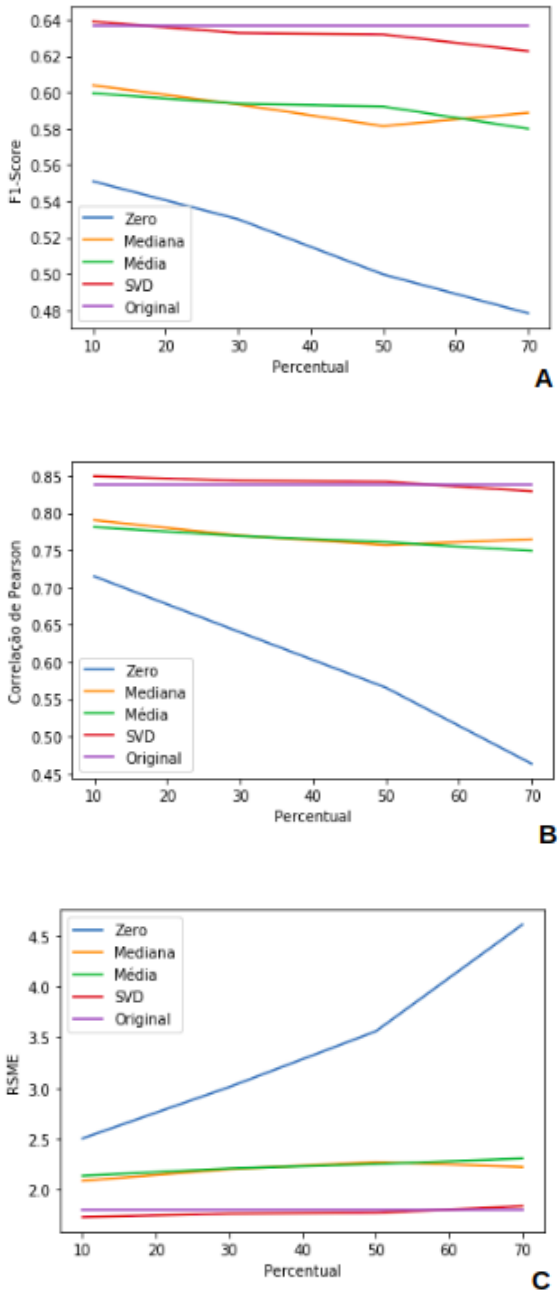


Figura 17 – Avaliação comparativa das métricas em relação ao percentual aplicado.

Além disso, é possível observar a distribuição completa dos pesos em cada par de kernel nos anexos deste documento, sendo uma comparação válida a disposição entre as porcentagens aplicadas dentro da mesma técnica.

4.2 Discussão

Para determinar o parâmetro ideal definido para a imputação com a técnica SVD, o método foi avaliado usando 10, 30, 50 e 70% de valores faltosos para estimativa (Figura

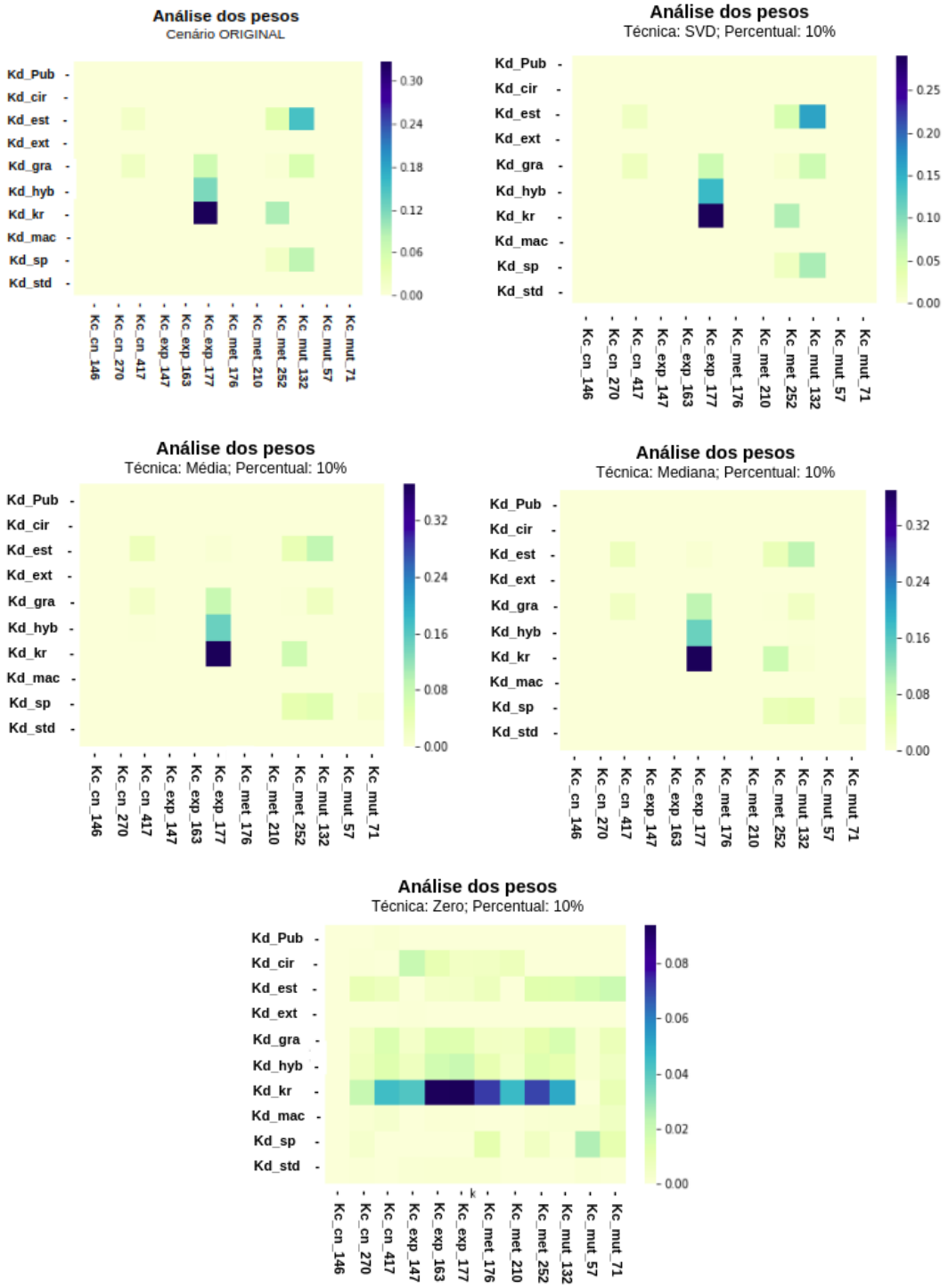


Figura 18 – Comparação entre a distribuição aleatória dos pesos em todas as técnicas, com o percentual de 10%.

17 - C). A estimativa mais bem avaliada, em comparação com o modelo base, é tida quando apenas 10% da matriz possui valores faltosos. Além disso, a técnica SVD apresenta resultados muito similares em todas as porcentagens aplicadas, sendo considera para esse conjunto de dados a técnica mais eficiente dentre as demais. Entretanto, seus resultados pioram de acordo com o aumento de posições faltosas no kernel. Como a estimativa baseada em SVD é essencialmente um método de regressão linear num espaço dimensional inferior, essa piora no desempenho não é surpreendente para dados de séries não temporais, onde um padrão de expressão claro muitas vezes não está presente.

De acordo com a literatura, a técnica média é uma das mais usadas para imputação de valores faltosos [Wei et al. 2018], [Cichonska et al. 2018], [Zhang 2016]. Embora tenha apresentado resultados melhores em relação a substituição dos valores ausentes por zero, a técnica média rendeu uma precisão bem menor do que a apresentada pela técnica SVD (Figura 17 - C). O mesmo pode ser inferido quando a técnica aplicada é a mediana. O fato é que, técnicas de imputação de valores simples apresentam desempenho inferior quando comparadas a outras técnicas mais sofisticadas, como é o caso do SVD.

Portanto, métodos mais sofisticados como o SVD, associado ao pairwiseMKL, fornecem maneiras mais precisas de estimar os valores ausentes para o conjunto de dados de interação de bioatividade de medicamentos. A técnica SVD, aplicada no contexto de redes bipartidas, apresentou desempenho muito superior em relação as soluções mais simples, tirando vantagem da estrutura de correlação dos dados para estimar os valores de expressão ausentes.

5 Considerações finais

No campo da bioinformática, muitos problemas, incluindo a previsão de bioatividade de drogas, podem ser formulados como problemas de aprendizagem em pares, onde o interesse é fazer previsões para pares de objetos, por exemplo, drogas e seus alvos [Cichonska et al. 2018]. Na literatura, algumas abordagens baseadas em kernel surgiram com o intuito de resolver esses problemas, e especialmente o aprendizado de multiplos kernels (MKL) apresenta benefícios promissores, como a integração de vários tipos de fontes de informações biológicas complexas na forma de kernels, juntamente com o aprendizado de sua importância para a tarefa de previsão.

No entanto, os métodos de MKL focam principalmente no problema de aprendizagem de kernels constituídos com o mesmo tipo de dado, por exemplo kernels droga-droga ou proteína-proteína, o que não corresponde ao problema da predição de interações em redes bipartidas. Além disso, boa parte dos algoritmos de MKL existentes são computacionalmente inviáveis, pois o imenso tamanho dos kernels emparelhados geram um gargalo no processamento do algoritmo, mesmo quando o número de pares de entrada é pequeno.

O algoritmo pairwiseMKL, utilizado nesse estudo, preenche todas as lacunas deixadas pelas abordagens anteriores, uma vez que consiste em um método para aprendizagem eficiente em termos de tempo e memória com vários kernels emparelhados, implementando otimização de pesos de kernel em pares de forma eficiente e treinamento de modelo em pares [Cichonska et al. 2018]. Os resultados obtidos no experimento realizado indicaram que o método pairwiseMKL, associado às técnicas de preenchimento de valores faltosos, é capaz de identificar e selecionar as fontes de informação mais relevantes para a interação droga-linha celular.

O desempenho das técnicas associada ao pairwiseMKL modificado foi considerado satisfatório. Principalmente quando a técnica utilizada foi a SVD, onde obteve os melhores resultados nas métricas avaliativas quando comparado com o cenário original (modelo base). Embora os resultados apresentados sejam superiores em relação a outras técnicas mais simples, análises futuras deverão ser feitas com o intuito de analisar se o desempenho da técnica é sensível ao tipo de dados que está sendo analisado.

Como discutido anteriormente, ainda não existem, na literatura, trabalhos que analisem técnicas de preenchimento de matrizes em kernels de natureza bipartida, sendo os resultados obtidos animadores para o desenvolvimento da área. Entretanto, várias outras técnicas podem ser aplicadas, além da utilização de outras bases de dados para uma avaliação mais completa do modelo.

5.1 Recomendações e trabalhos futuros

- Uso de outras técnicas mais complexas de preenchimento de matrizes, como a Multiplo Kernel Clustering - MKC, Mutual Kernel Matrix Completation - MKMC e a Random Forest - RF por exemplo, realizando uma posterior comparação de desempenho entre elas;
- Uso de outras bases de dados biológicas, como a descrita em [Yamanishi et al. 2008] e também avaliar bases maiores, como o bindingDB. O intuito é verificar o desempenho do algoritmo e das técnicas aplicadas em outras bases de dados;
- Avaliação em outros domínios, como sistemas de recomendação hibrido baseados em filtragem colaborativa e em conteúdo, ou sistemas de recomendação sensível ao contexto, por exemplo;
- Análise de custo de memória e de tempo empregado no experimento.

Referências

ACOCK, A. C. Working with missing values. *Journal of Marriage and family*, Wiley Online Library, v. 67, n. 4, p. 1012–1028, 2005. Citado na página 24.

AMMAD-UD-DIN, M. et al. Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics*, Oxford University Press, v. 32, n. 17, p. i455–i463, 2016. Citado 3 vezes nas páginas 15, 44 e 45.

ANVISA, A. N. de V. S. Regularização de empresas - medicamentos: Pesquisa clínica. Citado 3 vezes nas páginas 21, 29 e 30.

BARREIRO, E. J.; FRAGA, C. A. M. A questão da inovação em fármacos no brasil: Proposta de criação do programa nacional de fármacos (pronfar). *Química Nova*, SciELO Brasil, v. 28, p. S56–S63, 2005. Citado na página 29.

BOUGIATIOTIS, K.; GIANNAKOPOULOS, T. Enhanced movie content similarity based on textual, auditory and visual information. *Expert Systems with Applications*, Elsevier, v. 96, p. 86–102, 2018. Citado 2 vezes nas páginas 13 e 48.

CICHONSKA, A. et al. Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics*, Oxford University Press, v. 34, n. 13, p. i509–i518, 2018. Citado 19 vezes nas páginas 13, 15, 23, 24, 25, 26, 32, 33, 40, 41, 42, 43, 44, 46, 47, 50, 55, 61 e 63.

CSERMELY, P. et al. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology Therapeutics*, v. 138, n. 3, p. 333 – 408, 2013. ISSN 0163-7258. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0163725813000284>. Citado 2 vezes nas páginas 21 e 29.

DALIANIS, H. Evaluation metrics and evaluation. In: *Clinical Text Mining*. [S.l.]: Springer, 2018. p. 45–53. Citado 2 vezes nas páginas 49 e 50.

DONG, Y. et al. Network pharmacology-based prediction and verification of the targets and mechanism for panax notoginseng saponins against coronary heart disease. *Evidence-Based Complementary and Alternative Medicine*, Hindawi, v. 2019, 2019. Citado na página 32.

EKINS, S. et al. In silico repositioning of approved drugs for rare and neglected diseases. *Drug discovery today*, Elsevier, v. 16, n. 7-8, p. 298–310, 2011. Citado na página 30.

EZZAT, A. et al. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in bioinformatics*, Oxford University Press, v. 20, n. 4, p. 1337–1357, 2019. Citado na página 35.

FLEMING, N. How artificial intelligence is changing drug discovery. *Nature*, Nature Publishing Group, v. 557, n. 7706, p. S55–S55, 2018. Citado 2 vezes nas páginas 21 e 29.

GÖNEN, M.; ALPAYDIN, E. Multiple kernel learning algorithms. *Journal of machine learning research*, v. 12, n. Jul, p. 2211–2268, 2011. Citado na página 22.

GUHA, R. et al. Chemical informatics functionality in r. *J Stat Softw*, v. 18, n. 5, p. 1–16, 2007. Citado na página 46.

JACOB, L.; VERT, J.-P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, Oxford University Press, v. 24, n. 19, p. 2149–2156, 2008. Citado na página 39.

JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, Ieee, v. 22, n. 1, p. 4–37, 2000. Citado na página 23.

KEISER, M. J. et al. Predicting new molecular targets for known drugs. *Nature*, Nature Publishing Group, v. 462, n. 7270, p. 175–181, 2009. Citado na página 31.

Pearson's correlation coefficient. In: KIRCH, W. (Ed.). *Encyclopedia of Public Health*. Dordrecht: Springer Netherlands, 2008. p. 1090–1091. ISBN 978-1-4020-5614-7. Disponível em: <https://doi.org/10.1007/978-1-4020-5614-7_2569>. Citado na página 50.

KUMAR, R. et al. Multiple kernel completion and its application to cardiac disease discrimination. In: IEEE. *2013 IEEE 10th International Symposium on Biomedical Imaging*. [S.l.], 2013. p. 764–767. Citado 2 vezes nas páginas 24 e 36.

KURUCZ ANDRáS A. BENCZúR, K. C. M. Methods for large scale svd with missing values. *KDDCup.07*, ACM Digital Library, 2007. Citado na página 48.

LIU, X. et al. Multiple kernel k-means with incomplete kernels. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, 2019. Citado 3 vezes nas páginas 24, 36 e 37.

LIU, Y. et al. Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Processing Letters*, Springer, p. 1–17, 2020. Citado na página 50.

MOREIRA, E.; CAMPOS, G. O.; JR, W. M. Dense hierarchy decomposition for bipartite graphs. In: SBC. *Anais do VII Symposium on Knowledge Discovery, Mining and Learning*. [S.l.], 2019. p. 105–112. Citado 2 vezes nas páginas 30 e 31.

NASCIMENTO, A. C.; PRUDÊNCIO, R. B.; COSTA, I. G. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics*, Springer, v. 17, n. 1, p. 46, 2016. Citado 9 vezes nas páginas 21, 23, 30, 32, 33, 34, 35, 36 e 39.

NASCIMENTO, A. C. A. do. *Combinação de Kernels para predição de interações em redes biológicas*. Tese (Doutorado), 2015. Citado 6 vezes nas páginas 13, 22, 23, 30, 31 e 34.

NEILL, S. P.; HASHEMI, M. R. *Fundamentals of ocean renewable energy: generating electricity from the sea*. [S.l.]: Academic Press, 2018. Citado na página 50.

(ORG.), H. V. *Bioinformática: da Biologia à Flexibilidade Molecular*. [S.l.]: Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, 2014. 282 p. ISBN 978-85-69288-00-8. Citado na página 32.

PALEOLOGU, C.; BENESTY, J.; CIOCHINĂ, S. Linear system identification based on a kronecker product decomposition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, IEEE, v. 26, n. 10, p. 1793–1808, 2018. Citado na página 40.

QUENTAL, C.; FILHO, S. S. Ensaios clínicos: capacitação nacional para avaliação de medicamentos e vacinas. *Revista brasileira de Epidemiologia*, SciELO Public Health, v. 9, p. 408–424, 2006. Citado 2 vezes nas páginas 29 e 30.

RIVERO, R.; LEMENCE, R.; KATO, T. Mutual kernel matrix completion. *IEICE TRANSACTIONS on Information and Systems*, The Institute of Electronics, Information and Communication Engineers, v. 100, n. 8, p. 1844–1851, 2017. Citado 4 vezes nas páginas 21, 24, 35 e 36.

SAUNDERS, C.; GAMMERMAN, A.; VOVK, V. Ridge regression learning algorithm in dual variables. 1998. Citado na página 41.

SHAWE-TAYLOR, J.; CRISTIANINI, N. et al. *Kernel methods for pattern analysis*. [S.l.]: Cambridge university press, 2004. Citado 2 vezes nas páginas 21 e 42.

SILVA DANIEL LUIS NOTARI, G. D. Scheila de Avila e. *Bioinformática [recurso eletrônico]: contexto computacional e aplicações*. [S.l.]: Educs, UNIVERSIDADE DE CAXIAS DO SUL, 2020. 297 p. ISBN 978-65-5807-001-6. Citado na página 32.

SOUZA, B. F. d. *Meta-aprendizagem aplicada à classificação de dados de expressão gênica*. Tese (Doutorado) — Universidade de São Paulo, 2010. Citado na página 39.

STRAŽAR, M.; CURK, T. Approximate multiple kernel learning with least-angle regression. *Neurocomputing*, Elsevier, v. 340, p. 245–258, 2019. Citado na página 21.

SUN, T.; LAI, L.; PEI, J. Analysis of protein features and machine learning algorithms for prediction of druggable proteins. *Quantitative Biology*, Springer, v. 6, n. 4, p. 334–343, 2018. Citado na página 49.

SZEDMAK, S.; BACH, E. On the generalization of tanimoto-type kernels to real valued functions. *arXiv preprint arXiv:2007.05943*, 2020. Citado na página 46.

TAVARES, C. L. S. Cluster: um software para auxílio em estudos de dados biológicos. Universidade Federal de Minas Gerais, 2015. Citado na página 35.

TUIKKALA, J. et al. Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC bioinformatics*, BioMed Central, v. 9, n. 1, p. 1–14, 2008. Citado 2 vezes nas páginas 24 e 48.

VANHAELEN, Q. *Computational Methods for Drug Repurposing*. [S.l.]: Springer, 2019. Citado na página 30.

VANHAELEN, Q. et al. Design of efficient computational workflows for in silico drug repurposing. *Drug Discovery Today*, Elsevier, v. 22, n. 2, p. 210–222, 2017. Citado na página 30.

VERT, J.-P.; TSUDA, K.; SCHÖLKOPF, B. A primer on kernel methods. *Kernel methods in computational biology*, MIT press Cambridge, MA, v. 47, p. 35–70, 2004. Citado na página 22.

WANG, N. et al. A network pharmacology approach to determine the active components and potential targets of curculigo orchioides in the treatment of osteoporosis. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, International Scientific Information, Inc., v. 23, p. 5113, 2017. Citado na página 32.

WEI, R. et al. Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific reports*, Nature Publishing Group, v. 8, n. 1, p. 1–10, 2018. Citado 4 vezes nas páginas 24, 36, 48 e 61.

YAMANISHI, Y. et al. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, Oxford University Press, v. 24, n. 13, p. i232–i240, 2008. Citado 3 vezes nas páginas 13, 33 e 64.

YAN, X.-Y.; ZHANG, S.-W.; HE, C.-R. Prediction of drug-target interaction by integrating diverse heterogeneous information source with multiple kernel learning and clustering methods. *Computational biology and chemistry*, Elsevier, v. 78, p. 460–467, 2019. Citado na página 34.

YANG, W. et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, Oxford University Press, v. 41, n. D1, p. D955–D961, 2012. Citado 3 vezes nas páginas 26, 44 e 45.

YUAN, X. et al. Singular value decomposition based recommendation using imputed data. *Knowledge-Based Systems*, Elsevier, v. 163, p. 485–494, 2019. Citado na página 48.

ZHANG, Z. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, AME Publications, v. 4, n. 1, 2016. Citado 2 vezes nas páginas 24 e 61.

ZHAO, H. et al. A network pharmacology approach to explore active compounds and pharmacological mechanisms of epimedium for treatment of premature ovarian insufficiency. *Drug Design, Development and Therapy*, Dove Press, v. 13, p. 2997, 2019. Citado na página 32.

ZHU, X. et al. Localized incomplete multiple kernel k-means. In: *IJCAI*. [S.l.: s.n.], 2018. p. 3271–3277. Citado na página 37.

Apêndices

APÊNDICE A – Heatmaps dos pesos

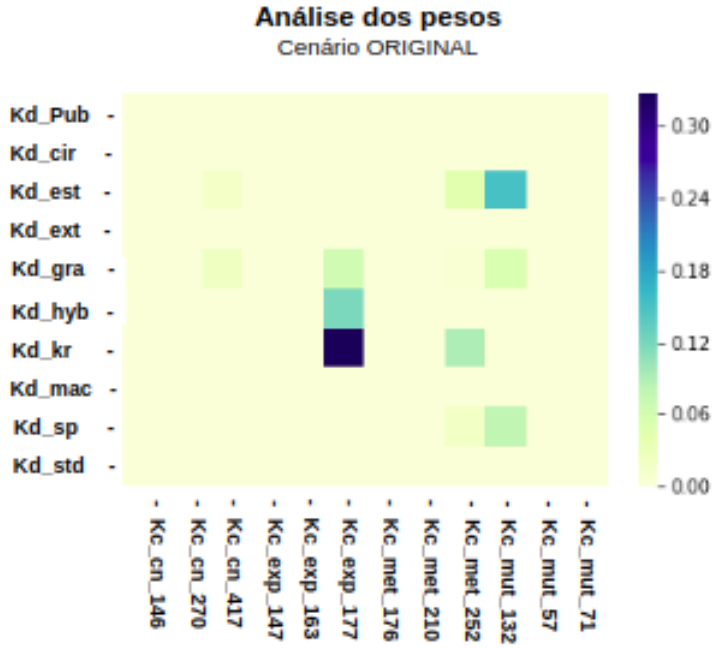


Figura 19 – Heatmap dos pesos do cenário original.

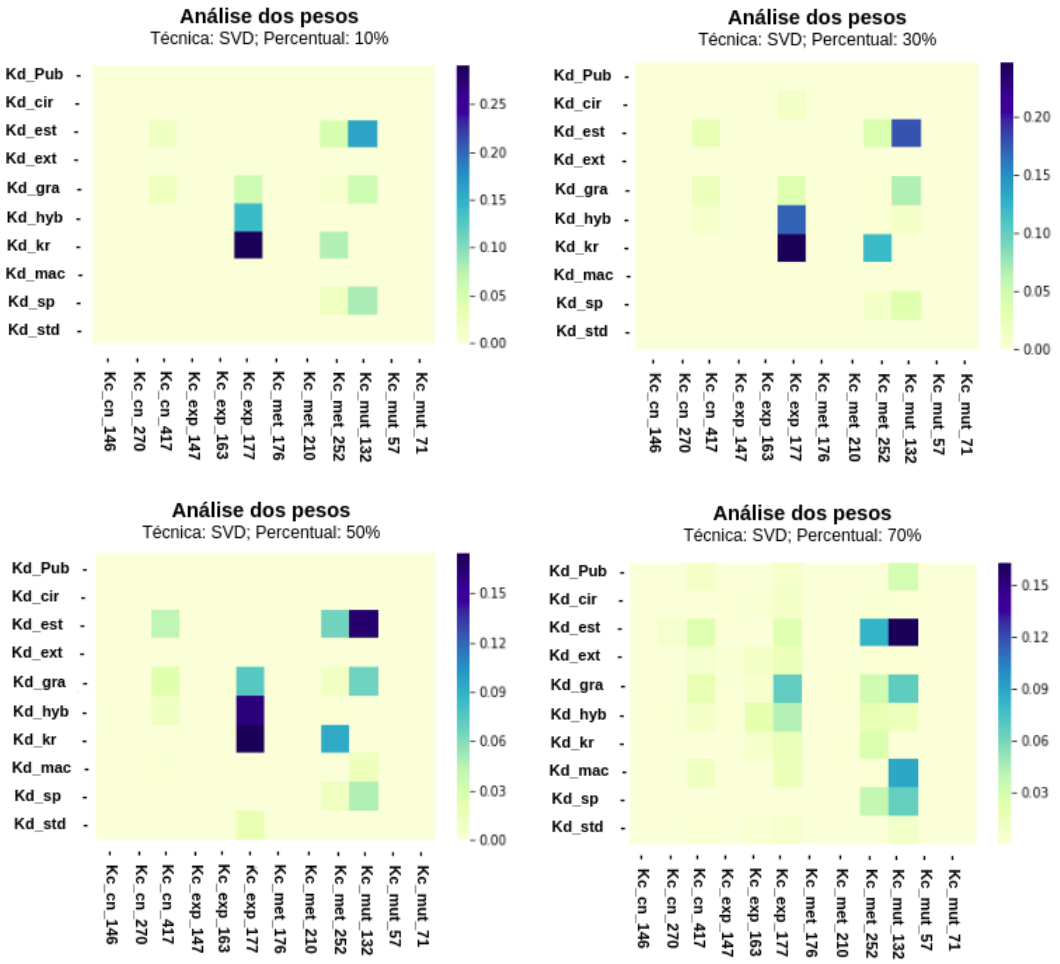


Figura 20 – Heatmap dos pesos: análise dos pesos de acordo com a combinação dos arquivos de drogas e células. Resultado da avaliação com a técnica 'SVD' e com as porcentagens de 10%, 30%, 50% e 70%, respectivamente.

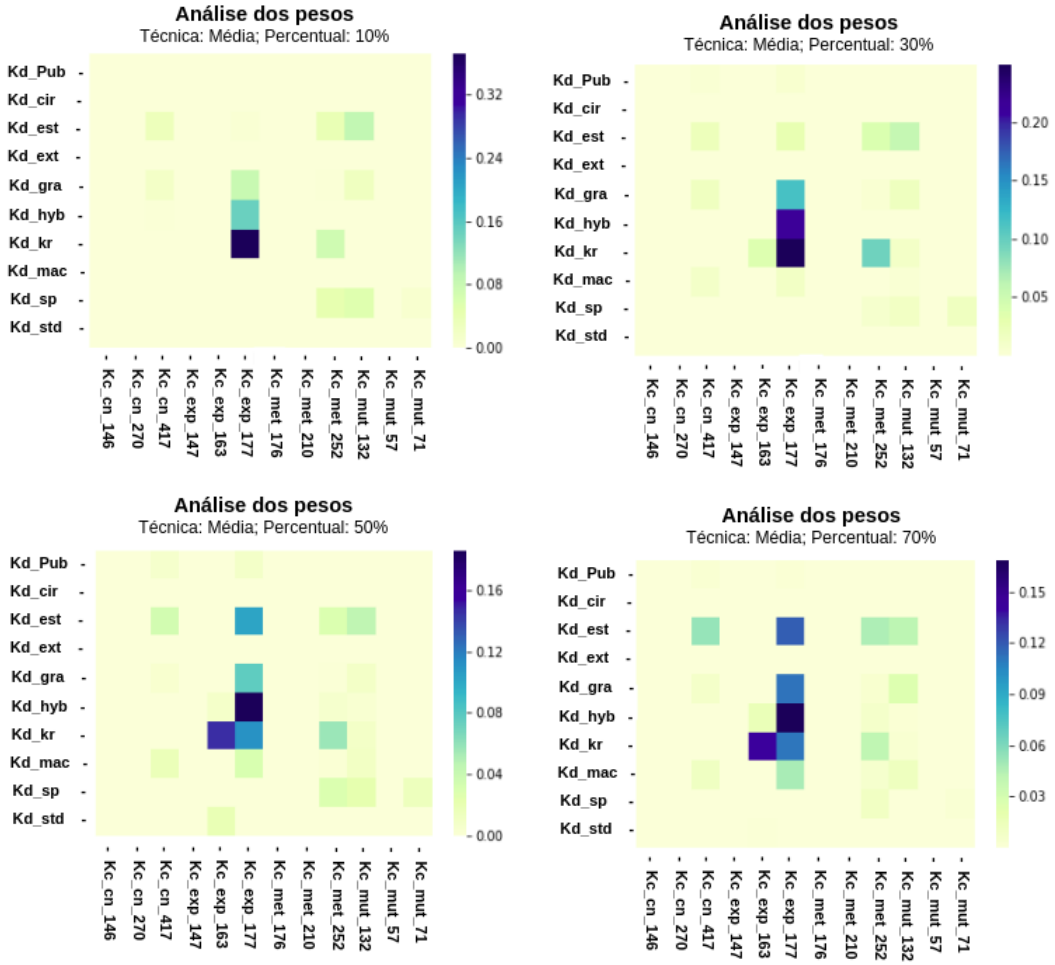


Figura 21 – Heatmap dos pesos: análise dos pesos de acordo com a combinação dos arquivos de drogas e células. Resultado da avaliação com a técnica 'Média' e com as porcentagens de 10%, 30%, 50% e 70%, respectivamente.

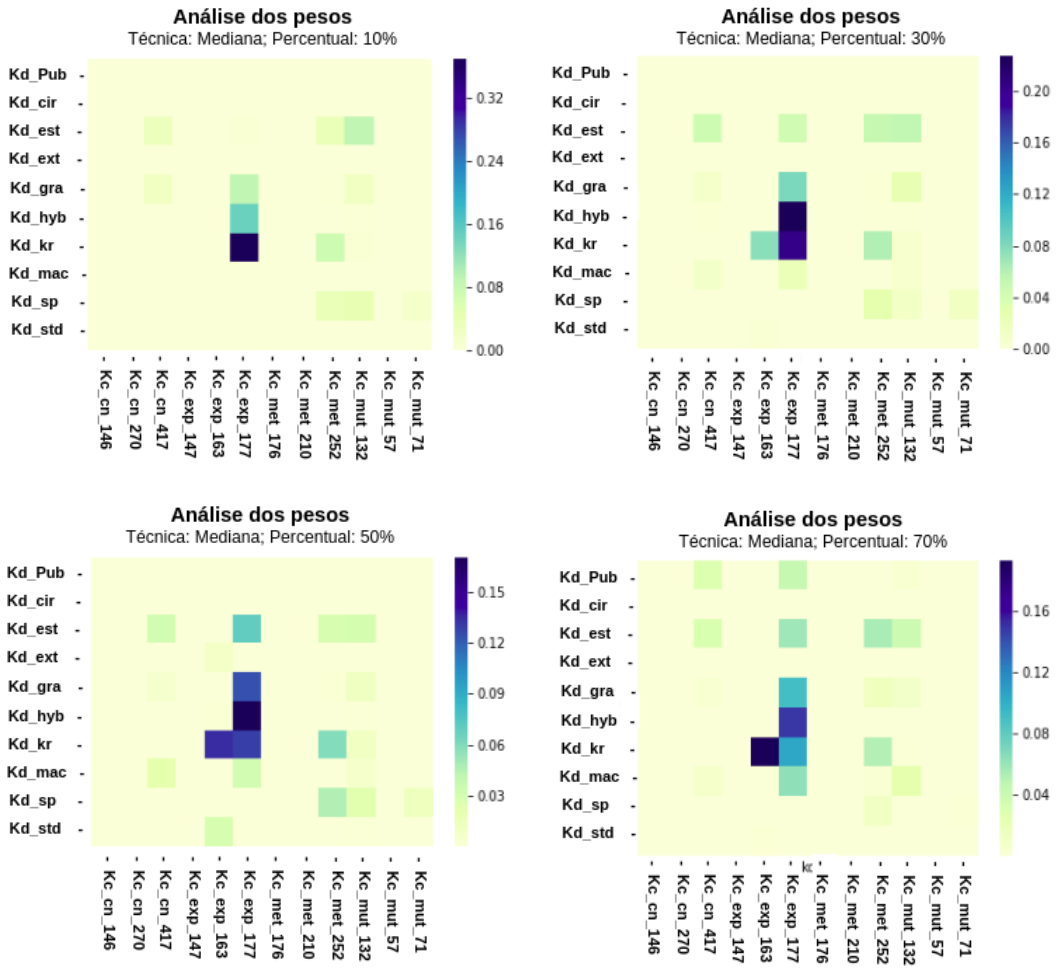


Figura 22 – Heatmap dos pesos: análise dos pesos de acordo com a combinação dos arquivos de drogas e células. Resultado da avaliação com a técnica 'Mediana' e com as porcentagens de 10%, 30%, 50% e 70%, respectivamente.

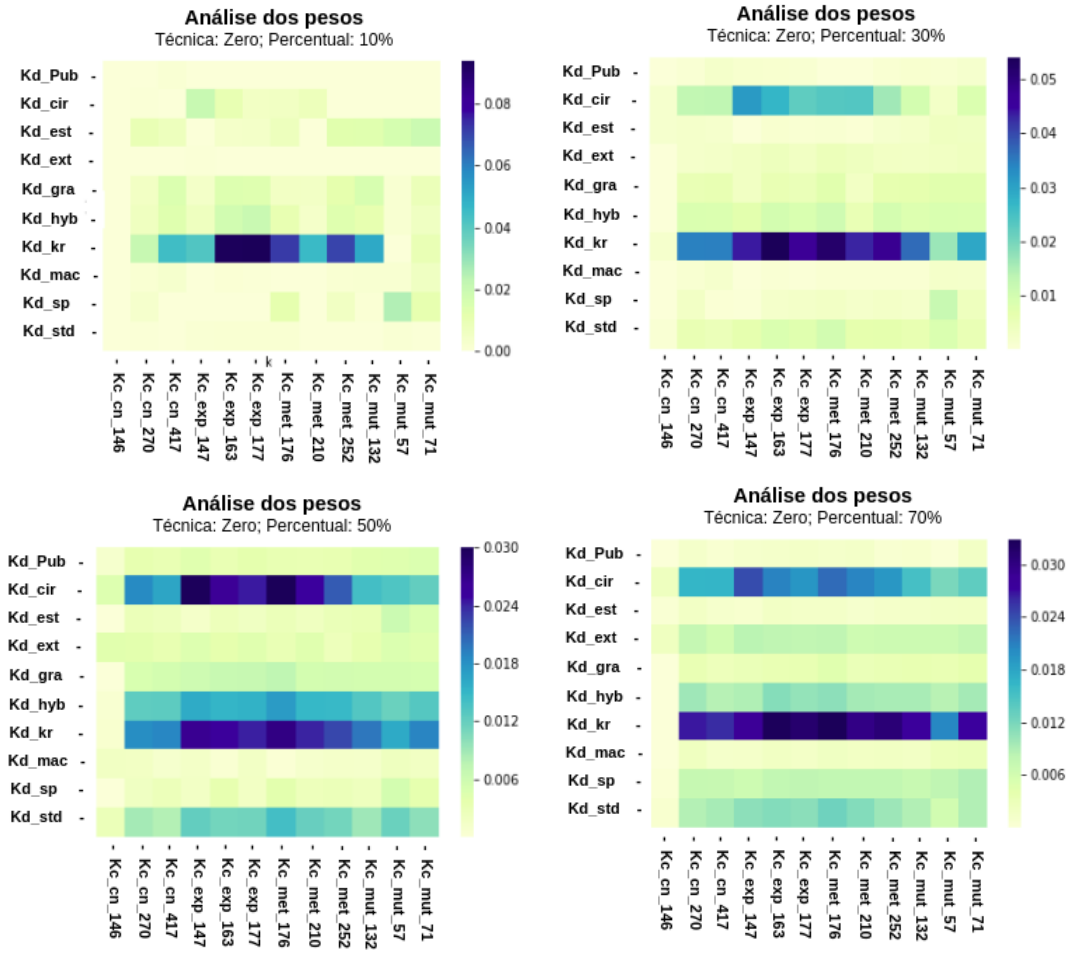


Figura 23 – Heatmap dos pesos: análise dos pesos de acordo com a combinação dos arquivos de drogas e células. Resultado da avaliação com a técnica 'Zero' e com as porcentagens de 10%, 30%, 50% e 70%, respectivamente.