Framework para mineração de subgrupos usando o database METABRIC

Adriano V. M. Zanini[1], Bruno L. Fonseca[1], Gabriel M. M. Fialho[1], José M. da R. Júnior[1], Matheus A. I. de Oliveira[1], Matheus V. L. Lima[1]

[1]Departamento de Computação – Universidade Federal de Minas Gerais (UFMG)

{zanini-adriano, brunolefonseca, gabrielmf2002}@ufmg.br

{josemrjunior, matheusirias, vazleal}@ufmg.br

Abstract. *This study aims to present a framework for subgroup mining using the METABRIC database, a comprehensive collection of genomic data on breast cancer. The framework incorporates SSDP+, CN2-SD, and EsmamDS algorithms, alongside tools like Orange. With gene expression data from approximately 2,000 patients and over 24,000 genes, this framework seeks to identify patient subgroups with specific genetic traits that impact survival and treatment outcomes.*

Resumo. *Este estudo tem como objetivo apresentar um framework para mineração de subgrupos utilizando a base de dados METABRIC, uma coleção abrangente de dados genômicos sobre câncer de mama. O framework incorpora os algoritmos SSDP+, CN2-SD e EsmamDS, além de ferramentas como o Orange. Com dados de expressão gênica de aproximadamente 2.000 pacientes e mais de 24.000 genes, este framework busca identificar subgrupos de pacientes com características genéticas específicas que impactam a sobrevivência e os resultados do tratamento.*

1. Introdução

Os algoritmos de aprendizado descritivo mostram-se como valiosas ferramentas para extração de informações sobre as bases de dados, ajudando a trazer à tona informações intrínsecas às bases de dados, permitindo novas ideias e agregando conhecimento sobre algo antes desconhecido [Wrobel 1997]. Com o crescente aumento de poder de processamento dos computadores e o refinamento dos algoritmos [Boneh et al. 1996], os desafios envolvendo soluções por descrição de dados se tornam cada vez mais atraentes. As pesquisas de descoberta de subgrupos permitem encontrar distribuições locais de um atributo alvo, permitindo representar grupos particulares como um conjunto de seletores sobre os atributos das bases [Lavrac et al. 2004].

A descoberta de subgrupos auxilia na obtenção de resultados facilmente interpretáveis a humanos, algo difícil de ser feito com os algoritmos tradicionais de aprendizado de máquina. Nos algoritmos de regressão e classificação, são ajustados parâmetros do modelo até encontrar o melhor ajuste para uma função objetivo. O número de parâmetros pode variar de dezenas até milhões, o que dificulta a análise humana de como foi possível chegar ao resultado e entender como e o que o modelo aprendeu. Por conta disso, essa abordagem é utilizada em problemas das mais diversas áreas, desde a

otimização de jogadas em esportes para obter a melhor pontuação até na área da saúde, na descoberta de padrões fisiológicos que contribuem para a recuperação de um paciente sobre alguma enfermidade. Em linha com desafios na área de saúde, descoberta de subgrupos é utilizada amplamente no mapeamento genômico e na identificação de características que podem ser usadas para determinar a probabilidade de um paciente suportar uma doença. Essa linha de pesquisa dentro de grandes conjuntos de dados genômicos tem se mostrado uma ferramenta poderosa na medicina personalizada, particularmente no tratamento e prognóstico do câncer de mama. A base de dados METABRIC (*Molecular Taxonomy of Breast Cancer International Consortium*) [Milioli et al. 2015] é uma das maiores e mais detalhadas coleções de dados genômicos sobre câncer de mama. A base de dados abrange informações de expressão gênica de aproximadamente 2.000 pacientes, apresentando a expressão genética de mais de 24.000 genes. Esta riqueza de dados oferece uma oportunidade única para explorar a complexidade molecular do câncer de mama e identificar novos subgrupos de pacientes com características genéticas específicas que possam impactar significativamente a sobrevivência e o tratamento.

Neste contexto, esse trabalho propõe um novo *framework* para descoberta de subgrupos em dados de expressão gênica, tendo como atributos alvo o evento de sobrevivência (paciente se recuperou ou faleceu). O método desenvolvido busca otimizar a qualidade e eficiência dos algoritmos dos algoritmos de descoberta de subgrupos de forma encontrar os subgrupos de atributos genéticos mais relevantes em termos de correlação com a sobrevivência dos pacientes. Em uma validação empírica, este estudo compara três algoritmos de descoberta de subgrupos, sendo eles o SSDP+, CN2-SD e EsmamDS. Foram encontrados agrupamentos excepcionais em linha com a proposta de análise de sobrevivência dos pacientes.

2. Trabalhos Relacionados

Diversos trabalhos na bibliografia utilizam abordagem de descoberta de subgrupos para solucionar os mais diversos problemas. Soluções utilizando o algoritmo CN2-SD como [Cano et al. 2008] visam melhorar a escalabilidade da solução para grandes conjuntos de dados, reduzindo a seleção de características sem perder informações essenciais. Com essa abordagem foi possível identificar padrões genéticos e clínicos que influenciam na sobrevivência dos pacientes utilizando bases de dados como o METABRIC, que foi utilizado neste trabalho. [Mattos et al. 2021] utilizou o algoritmo EsmamDS para explicar as diferenças no comportamento de sobrevivência a partir de uma perspectiva de padrões descritivos, em sentido contrário a modelagem de padrões globais. Seguindo a mesma linha, [Mattos et al. 2020] propôs uma nova abordagem para a descoberta de subgrupos com comportamento de sobrevivência excepcional, utilizando técnicas de otimização por colônia de formigas. Dessa forma, foi possível relevar interações desconhecidas anteriormente entre variáveis oferecendo informações ainda mais profundas sobre os fatores que influenciam na sobrevivência dos pacientes.

Soluções baseadas em algoritmos evolucionários também são utilizadas para a descoberta de subgrupos.

[Pontes et al. 2016] aborda a descoberta de padrões discriminativos em bases de dados de alta dimensionalidade, utilizando o algoritmo SSDP. O SSDP utiliza uma abordagem evolutiva para identificar os top-k padrões discriminativos, simplificando os

parâmetros necessários e focando em problemas de alta dimensionalidade, como bancos de dados de microarranjos Neste trabalho foi aplicado o SSDP com o objetivo de facilitar a extração de conhecimento em bases de dados de alta dimensionalidade, como os microarranjos de genes. Os resultados foram comparados com métodos tradicionais de mineração de dados e demonstraram que o SSDP é capaz de identificar padrões significativos de forma mais eficiente. A relação entre o SSDP e este trabalho está na aplicação de técnicas de mineração de dados para extrair padrões significativos de conjuntos de dados de alta dimensionalidade. Ambos os trabalhos visam superar os desafios impostos pela alta dimensionalidade e complexidade dos dados. O SSDP destaca-se por sua simplicidade e eficácia na busca de padrões discriminativos, características que também buscamos em nossa abordagem para a descoberta de subgrupos no Metabric. [Lucas et al. 2018] propõe o SSDP+ como uma extensão do algoritmo SSDP. O objetivo principal do SSDP+ é melhorar a diversidade e a informatividade dos subgrupos descobertos, abordando as limitações do SSDP original. Enquanto o SSDP foca em maximizar a qualidade dos subgrupos com base em uma única métrica de avaliação, o SSDP+ introduz múltiplas métricas, permitindo uma análise mais abrangente e detalhada. SSDP+ foi comparado com o SSDP e outros algoritmos tradicionais de descoberta de subgrupos, mostrando melhorias significativas em termos de diversidade e qualidade dos subgrupos gerados.

3. Metodologia

Apresenta-se aqui um novo *framework* para descoberta de subgrupos em dados de expressão gênica com foco na otimização da qualidade dos subgrupos descobertos. Essa estrutura consiste em um conjunto de métodos de processamento de dados que são aplicados em sequência de forma a filtrar os dados mais relevantes e assim otimizar os resultados da etapa de descoberta de subgrupos (etapa final).

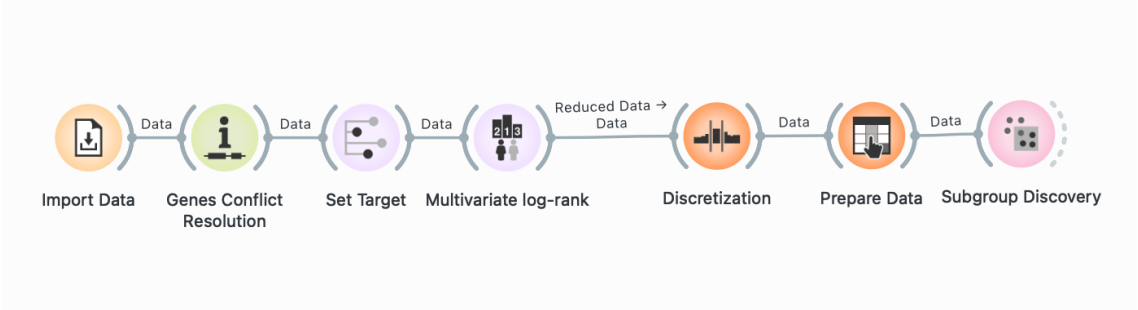


Figura 1. Descoberta de subgrupos em dados de expressão gênica

3.1. Fonte de Dados e Pré-processamento

A base de dados METABRIC (Molecular Taxonomy of Breast Cancer International Consortium)[Curtis et al. 2012] é uma das mais abrangentes e detalhadas coleções de dados sobre câncer de mama disponíveis para pesquisa. Este conjunto de dados foi desenvolvido com o objetivo de melhorar a compreensão das características moleculares do câncer de mama e auxiliar na descoberta de novos biomarcadores e tratamentos.

A base de dados METABRIC contém informações detalhadas de mais de 2.000 amostras de câncer de mama, abrangendo diversos aspectos clínicos, patológicos e

genômicos. Nela, é apresentada a expressão genética de mais de 24.000 genes e os dados clínicos dos pacientes em um banco de dados estruturado. Os dados clínicos incluem informações sobre a idade das pacientes, status menopausal, tamanho do tumor, status dos linfonodos, grau histológico, tipo de câncer de mama (e.g., ductal, lobular), e status de receptor hormonal (ER, PR, HER2). Esses dados são essenciais para correlacionar características genéticas com resultados clínicos e entender como diferentes fatores influenciam a progressão da doença e a resposta ao tratamento.

3.1.1. Seleção de Características

A análise realizada se concentra nas características de expressão gênica disponíveis no conjunto de dados METABRIC. Uma decisão metodológica importante neste estudo foi a exclusão deliberada de dados clínicos convencionais da nossa análise de descoberta de subgrupos. Esta escolha foi motivada por diversos fatores:

1. **Foco na base molecular:** Ao nos concentrarmos exclusivamente nos dados de expressão gênica, buscamos identificar subgrupos baseados puramente em características moleculares. Isso permite uma compreensão mais profunda dos mecanismos biológicos subjacentes que podem influenciar os desfechos de sobrevivência.
2. **Descoberta de novos biomarcadores:** A exclusão de dados clínicos conhecidos aumenta a probabilidade de identificarmos novos biomarcadores moleculares que podem não estar correlacionados com características clínicas já estabelecidas.
3. **Redução de viés:** Dados clínicos podem introduzir vieses baseados em conhecimentos prévios ou práticas clínicas atuais. Ao excluí-los, minimizamos o risco de que nossos subgrupos simplesmente reflitam classificações clínicas já conhecidas. Além disso, é provável que e uma expressão genética que tenha efeito significatívo na sobrêvivência do paciênte, também afete os dados clínicos.
4. **Independência de fatores subjetivos:** Alguns dados clínicos podem ser influenciados por fatores subjetivos ou variações nas práticas de diferentes instituições. A análise baseada em expressão gênica oferece uma abordagem mais objetiva e padronizada.

3.1.2. Resolução de Conflitos de Identificadores de Genes

Na preparação dos dados para análise foi realizada a resolução de conflitos entre diferentes identificadores de genes. O processo envolveu as seguintes etapas:

1. **Alinhamento com o banco de dados Entrez:** Os identificadores de genes presentes no conjunto de dados Metabric foram alinhados com o banco de dados Entrez Gene [Maglott et al. 2007]. Este passo é necessário para garantir a consistência e atualização dos identificadores de genes.
2. **Remoção de duplicatas:** Genes que apareceram múltiplas vezes devido a diferentes identificadores foram consolidados, mantendo apenas uma entrada por gene.
3. **Resolução de ambiguidades:** Nos casos em que um identificador poderia se referir a múltiplos genes, utilizamos informações adicionais, como a localização cromossômica e a função conhecida do gene, para resolver a ambiguidade.

3.1.3. Análise de Significância

Para se avaliar a significância dos genes em relação aos desfechos de sobrevivência, foi empregado o teste multivariado de log-rank [Harrington and Fleming 1982]. Este método foi escolhido devido à sua capacidade de lidar com múltiplas variáveis simultaneamente, o que é particularmente útil no contexto de dados de expressão gênica de alta dimensionalidade.

O teste multivariado de log-rank compara as curvas de sobrevivência de diferentes grupos, levando em consideração múltiplos fatores (neste caso, níveis de expressão de diferentes genes) simultaneamente. Ele testa a hipótese nula de que não há diferença nas curvas de sobrevivência entre os grupos definidos pelos diferentes níveis de expressão gênica. O emprego dessa análise antes da etapa de descoberta de subgrupos busca:

- Manter um equilíbrio entre a descoberta de genes potencialmente importantes e o controle de falsos positivos.
- Garantir uma base robusta para a subsequente descoberta de subgrupos, focando apenas nos genes mais fortemente associados aos desfechos de sobrevivência.

Foi definido o $p - valor < 0.005$ como limiar de corte. Assim, cerca de 2500 atributos foram considerados relevantes o suficiente para serem candidatos no processo de descoberta de subgrupos.

3.1.4. Discretização de Dados

A discretização dos dados é particularmente útil contexto da análise de dados de expressão gênica a análise de subgrupos, pois:

- Reduz significativamente o número de subgrupos possíveis, o que implica em uma melhora no tempo e capacidade de processamento dos algoritmos de descoberta de subgrupos.
- Facilita a identificação de padrões discretos de expressão gênica que podem ser biologicamente relevantes.
- Reduz o ruído nos dados, potencialmente melhorando o desempenho dos algoritmos de descoberta de subgrupos subsequentes.
- Permite uma interpretação mais intuitiva dos níveis de expressão gênica em comparação com valores contínuos.

O método escolhido para discretização dos dados de expressão gênica nessa implementação foi a heurística de discretização por minimização de entropia introduzida por [Fayyad and Irani 1993]. Nesse método, o atributo é dividido recursivamente em um corte que maximiza o ganho de informação, até que o ganho seja menor que o comprimento mínimo de descrição do corte. Essa discretização pode resultar em um número arbitrário de intervalos. Se o resultado for um único intervalo, a variável é descartada.

3.2. Descoberta de Subgrupos

Na aplicação empírica foram utilizados três algoritmos distintos de descoberta de subgrupos.

3.2.1. CN2-SD

O algoritmo CN2-SD é uma adaptação do algoritmo CN2 para o contexto de descoberta de subgrupos. Neste método, um peso é atribuído a cada instância do conjunto de dados. Iterativamente, o algoritmo, por meio de *beam search*, busca pelo subgrupo que melhor se comporta de acordo com a métrica *WRAcc*, além de ajustar os pesos para a realização da próxima busca.

Nos testes foram empregados parâmetros conservadores com o objetivo de se otimizar a qualidade dos resultados sem uma penalização excessiva do tempo de execução. Os parâmetros utilizados foram:

- **Rule ordering:** Unordered
- **Evaluation measure:** WRAcc
- **Beam width:** 30
- **Minimum rule coverage:** 1
- **Maximum rule length:** 8

3.2.2. SSDP+

O algoritmo SSDP+ é uma versão aprimorada do algoritmo SSDP. O SSDP é um algoritmo que se utiliza de conceitos presentes em algoritmos genéticos para encontrar os principais subgrupos, especialmente em bases de dados que apresentam alta dimensionalidade, o que é o caso do dataset METABRIC. O SSDP+ estende o SSDP, trazendo mecanismos que visam a diminuição da redundância e a agregação de diversidade aos resultados.

Os parâmetros utilizados nos testes empíricos seguem as recomendações dos autores do SSDP+, sendo empregadas as seguintes configurações:

- **Número de Tops K Subgrupos:** 10
- **Número Máximo de Subgrupos Similares para Cada um dos Tops K Subgrupos:** 5
- **Fator de Similaridade:** 0,10
- **Variável Alvo:** p (evento de sobrevivência). Sendo ‘p’ o paciente não sobreviveu e ‘n’ o paciente sobreviveu.
- **Filtrar Atributos:** “Overall Survival Time [months]”. Este atributo está presente na base de dados para referência, mas não será utilizado na identificação dos subgrupos.

3.2.3. EsmamDS

O algoritmo EsmamDS utiliza a Otimização por Colônia de Formigas (ACO) para descobrir subgrupos excepcionais em dados de sobrevivência. O algoritmo inicializa feromônios, realiza uma busca estocástica e ajusta as trilhas de feromônio iterativamente. Formigas constroem descrições que são podadas localmente, e o conjunto final de subgrupos é atualizado para minimizar redundâncias e maximizar a cobertura, utilizando

operações de generalização. Este processo otimiza a descoberta de subgrupos mais gerais e compactos.

Após testes empíricos, verificamos que os melhores parâmetros utilizados na execução do algoritmo EsmamDS para a base de dados escolhida é:

- **Número de Formigas:** 100
- **Tamanho Mínimo do Subgrupo:** 0.1
- **Regras para Convergência:** 5
- **Iterações para Estagnação:** 40
- **Peso da Pontuação:** 0.9
- **Offset Logístico:** 5

3.3. Ferramentas Utilizadas

A base de dados METABRIC foi obtida pacote de software *Orange: Data Mining* [Demšar et al. 2013] e o pré-processamento e discretização dos dados também foi realizado com os widgets disponíveis nele.

Os testes empíricos para descoberta de subgrupos foram realizados utilizando-se as implementações do SSDP+[1] e do EsmamDS [2], além da implementação do CN2-SD disponibilizada pelo Orange.

Para geração de gráficos a análise exploratória de dados foi utilizada a linguagem Python e as ferramentas JupyterLab, Pandas, Numpy, e Lifelines.

4. Análise dos resultados obtidos

A Tabela 1 apresenta um resumo comparativo dos resultados obtidos pelos métodos CN2-SD e SSDP+, tanto em sua forma original (utilizando todos os dados) quanto com a aplicação do framework proposto. Pelo fato do ESM-AM ser inerentemente mais lento, o tempo de processamento inviabilizou a inclusão do algoritmo nos testes com a base completa. A análise desses resultados revela insights importantes sobre o desempenho e as características de cada abordagem:

- **Número de Regras:** O framework proposto resultou em um aumento no número de regras geradas, tanto para o CN2-SD (de 7 para 13) quanto para o SSDP+ (de 38 para 44). Isso sugere que o framework permite a descoberta de um conjunto mais diversificado de padrões nos dados.
- **Comprimento Médio das Regras:** Observa-se um aumento significativo no comprimento médio das regras com a aplicação do framework. Para o CN2-SD, o comprimento médio quase dobrou (de 4,125 para 8,000), enquanto para o SSDP+ houve um aumento ainda mais expressivo (de 3 para 17). Regras mais longas podem indicar a captura de padrões mais específicos e complexos nos dados.
- **Qualidade Média dos Top 3:** A qualidade média das três melhores regras melhorou consideravelmente com a aplicação do framework. Para o CN2-SD, houve um aumento de 0,0497 para 0,0760 (aproximadamente 53% de melhoria). O SSDP+ também apresentou uma melhoria significativa, passando de 0,0293 para 0,0411 (cerca de 40% de aumento). Isso sugere que o framework proposto está efetivamente identificando subgrupos mais relevantes e informativos.

[1] `https://github.com/tarcisiodpl/ssdp_plus`
[2] `https://github.com/jbmattos/EsmamDS`

Em resumo, a aplicação do framework proposto demonstrou melhorias consistentes em ambos os métodos, CN2-SD e SSDP+, em termos de qualidade das regras geradas. O aumento no número e no comprimento das regras sugere que o framework está permitindo uma exploração mais profunda e detalhada dos padrões nos dados. Isso é particularmente evidenciado pelo aumento significativo na qualidade média das três melhores regras.

Tabela 1. Resumo dos resultados obtidos

Method	# Rules	Avg. Length	Top 3 Avg Quality
CN2-SD	7	4.125	0.0497
Framework + CN2-SD	13	8.000	0.0760
SSDP+	38	3	0.0293
SSDP+ + Framework	44	17	0,0411

Estes resultados indicam que o framework pode ser uma ferramenta valiosa para aprimorar a descoberta de subgrupos em dados genéticos, oferecendo insights mais ricos e potencialmente mais relevantes para a compreensão dos fatores de expressão gênica que influenciam a sobrevivência em pacientes.

5. Análise das Curvas de Kaplan-Meier

As Figuras (a), (b) e (c) apresentam as curvas de sobrevivência de Kaplan-Meier para os dois principais subgrupos identificados pelos métodos Esmam-DS, SSDP+ e CN2-SD, respectivamente. Essas curvas fornecem uma comparação visual das probabilidades de sobrevivência entre os subgrupos descobertos por cada método.

5.1. Observações Gerais

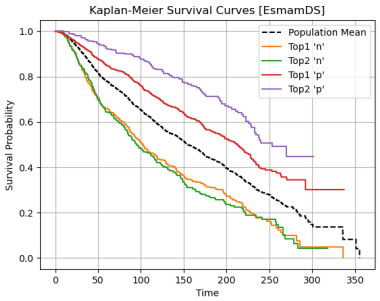
- Todos os três métodos (Esmam-DS, SSDP+ e CN2-SD) identificaram com sucesso subgrupos com padrões de sobrevivência distintos.
- Em todos os casos, observa-se uma clara separação entre as curvas de sobrevivência dos dois principais subgrupos (rotulados como 'n' e 'p'), indicando que estes métodos encontraram distinções significativas nos desfechos de sobrevivência.
- A média populacional (linha tracejada preta) geralmente se situa entre as curvas dos subgrupos, o que é esperado se os subgrupos representarem desfechos divergentes dentro da população geral.
- Para todos os três métodos, os subgrupos 'p' (linhas roxas) consistentemente mostram melhores desfechos de sobrevivência em comparação com os subgrupos 'n' e a média populacional.

5.2. Comparação entre os Métodos

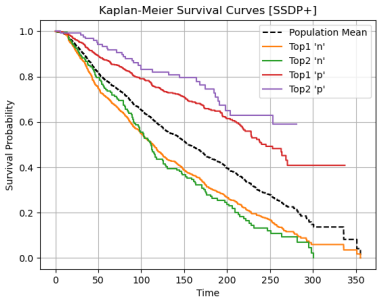
- O método CN2-SD (Figura c) parece ter identificado subgrupos com a diferença mais pronunciada nos desfechos de sobrevivência, particularmente para o subgrupo Top1 'n', que mostra uma sobrevida notavelmente pior do que nos outros métodos.
- O método SSDP+ (Figura b) aparenta ter encontrado subgrupos com diferenças mais moderadas na sobrevivência, especialmente nos períodos iniciais de tempo.

- O método Esmam-DS (Figura a) mostra uma clara separação entre os subgrupos, com padrões de certa forma intermediários entre SSDP+ e CN2-SD.

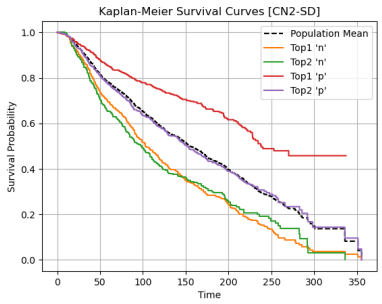
A **Figura 2** contém a plotagem usando Kaplan-Meier, onde comparamos o resultado da plotagem entre os três métodos, Esmam-DS, SSDP+ e CN2-SD.



(a) Kaplan-Meier dos Top 2 subgrupos do Esmam-DS



(b) Kaplan-Meier dos Top 2 subgrupos do SSDP+



(c) Kaplan-Meier dos Top 2 subgrupos do CN2-SD

Figura 2. Comparação de diferentes métodos

5.3. Sintetização dos Resultados

A análise comparativa dos métodos CN2-SD e SSDP+ revelou que o framework proposto resultou em um aumento no número de regras geradas e no comprimento médio das regras, sugerindo uma exploração mais profunda dos padrões nos dados. A qualidade média das três melhores regras também melhorou consideravelmente, indicando que o framework está identificando subgrupos mais relevantes e informativos.

As curvas de sobrevivência de Kaplan-Meier para os subgrupos identificados pelos métodos EsmamDS, SSDP+ e CN2-SD mostram uma clara separação entre as curvas dos principais subgrupos, indicando distinções significativas nos desfechos de sobrevivência. O método CN2-SD identificou subgrupos com diferenças mais pronunciadas nos desfechos, enquanto o SSDP+ encontrou diferenças mais moderadas, especialmente nos períodos iniciais. O EsmamDS mostrou padrões intermediários.

Os resultados do SSDP+ indicam uma boa cobertura de grupos locais da variável alvo, com subgrupos apresentando uma confiança considerável. No entanto, os valores de WRAcc foram mais elevados apenas para o subgrupo principal. A análise dos

subgrupos revelou que a expressão combinada de genes pode afetar significativamente a sobrevivência dos pacientes.

O EsmamDS, aplicado para minerar subgrupos com comportamentos de sobrevivência excepcionais, destacou variações específicas nos genes que podem influenciar a sobrevivência. A análise desses subgrupos pode revelar interações genéticas desconhecidas e potenciais alvos para tratamentos personalizados, fornecendo dados valiosos para a pesquisa clínica e biomédica.

6. Conclusão

O desenvolvimento e a implementação do framework para a mineração de subgrupos utilizando a base de dados METABRIC apresentaram resultados significativos em suas métricas e, principalmente, na abordagem inovadora que facilita a aplicação em outras áreas, como a biomedicina. Os três algoritmos apresentados - CN2-SD, EsmamDS e SSDP+ - demonstraram ser eficazes na identificação de subgrupos de pacientes com características genéticas específicas que impactam na sobrevida e nos resultados do tratamento do câncer de mama. Esta abordagem não só aprimora a análise de dados genéticos, mas também oferece uma ferramenta poderosa para a pesquisa clínica, permitindo o desenvolvimento de tratamentos personalizados e a melhoria das estratégias terapêuticas existentes.

6.1. Principais Contribuições

- **Integração de Algoritmos Avançados**:
  - O *framework* incorporou algoritmos de descoberta de subgrupos como o SSDP+, CN2-SD e EsmamDS, que se mostraram eficientes na identificação de padrões genéticos complexos e relevantes.
  - A combinação desses algoritmos permitiu a exploração de diferentes abordagens e métricas, aumentando a diversidade e a informatividade dos subgrupos identificados.
- **Processamento de Dados de Alta Dimensionalidade**:
  - O uso de técnicas robustas de pré-processamento, como a resolução de conflitos de identificadores de genes e a discretização dos dados de expressão gênica, foi essencial para a preparação adequada dos dados e a maximização da eficiência dos algoritmos de descoberta de subgrupos.
  - A análise de significância com o teste multivariado de *log-rank* garantiu que apenas genes fortemente associados aos desfechos de sobrevida fossem considerados, reduzindo falsos positivos e focando nos atributos mais relevantes.

6.2. Limitações e Futuras Direções

Apesar dos resultados promissores, algumas limitações foram identificadas:

- **Dimensionalidade dos Dados**: A alta dimensionalidade dos dados de expressão gênica continua a ser um desafio, exigindo a contínua melhoria das técnicas de pré-processamento e dos algoritmos de descoberta de subgrupos.
- **Validação Externa**: A validação externa dos subgrupos descobertos, utilizando outros conjuntos de dados de câncer de mama, é essencial para confirmar a generalizabilidade e a robustez dos resultados.

Para trabalhos futuros, sugere-se:

- **Incorporação de Dados Clínicos**: A inclusão controlada de dados clínicos pode fornecer uma visão mais holística e integrada dos fatores que influenciam a sobrevida dos pacientes.
- **Desenvolvimento de Interfaces de Visualização**: Ferramentas de visualização interativas podem facilitar a interpretação e a exploração dos subgrupos descobertos, tornando as descobertas mais acessíveis para profissionais de saúde.
- **Expansão para Outros Tipos de Câncer**: A aplicação do *framework* a outros tipos de câncer pode ajudar a identificar padrões comuns e específicos, contribuindo para a medicina personalizada em uma ampla gama de contextos.

Em conclusão, o *framework* mostrou-se uma ferramenta versátil e poderosa, oferecendo novas perspectivas para a compreensão do câncer de mama. As melhorias contínuas e a validação adicional são passos necessários para maximizar o impacto clínico e científico.

Referências

Boneh, D., Dunworth, C., Lipton, R. J., and Sgall, J. (1996). On the computational power of dna. *Discrete Applied Mathematics*, 71(1-3):79–94.

Cano, J.-R., Herrera, F., Lozano, M., and García, S. (2008). Making cn2-sd subgroup discovery algorithm scalable to large size data sets using instance selection. *Expert Systems with Applications*, 35(4):1949–1965.

Curtis, C., Shah, S., Chin, S.-F., Turashvili, G., Rueda, O., Dunning, M., Speed, D., Lynch, A., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Caldas, C., Aparicio, S., Brenton, J., and Børresen-Dale, A.-L. (2012). The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups. *Nature*, 486:–.

Demšar, J., Curk, T., Erjavec, A., Gorup, v., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2013). Orange: data mining toolbox in python. *J. Mach. Learn. Res.*, 14(1):2349–2353.

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *International Joint Conference on Artificial Intelligence*.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566.

Lavrac, N., Kavsek, B., Flach, P., and Todorovski, L. (2004). Subgroup discovery with cn2-sd. *J. Mach. Learn. Res.*, 5(2):153–188.

Lucas, T., Vimieiro, R., and Ludermir, T. (2018). Ssdp+: A diverse and more informative subgroup discovery approach for high dimensional data. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE.

Maglott, D., Ostell, J., Pruitt, K., and Tatusova, T. (2007). Entrez gene: Gene-centered information at ncbi. *Nucleic Acids Res.*, 39:D52–57.

Mattos, J. B., Neto, P. S., and Vimieiro, R. (2021). Esmamds: A more diverse exceptional survival model mining approach. *arXiv preprint arXiv:2109.02610*.

Mattos, J. B., Silva, E. G., de Mattos Neto, P. S., and Vimieiro, R. (2020). Exceptional survival model mining. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part II 9*, pages 307–321. Springer.

Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., and Moscato, P. (2015). The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the metabric data set. *PLoS One*, 10(7):e0129711.

Pontes, T., Vimieiro, R., and Ludermir, T. B. (2016). Ssdp: a simple evolutionary approach for top-k discriminative patterns in high dimensional databases. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 361–366. IEEE.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *European symposium on principles of data mining and knowledge discovery*, pages 78–87. Springer.