UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

LUIZ FERNANDO DA CUNHA CINTRA

Uma estratégia de pós-processamento para seleção de regras de associação para descoberta de conhecimento

Goiânia
2023

Processo: 23070.041119/2023-17 Documento: 4062042

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Luiz Fernando da Cunha Cintra

3. Título do trabalho

Uma estratégia de pós-processamento para seleção de regras de associação para descoberta de conhecimento

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Rogerio Lopes Salvini**, **Professor do Magistério Superior**, em 21/09/2023, às 14:11, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luiz Fernando Da Cunha Cintra**, **Discente**, em 21/09/2023, às 14:41, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4062042** e o código CRC **B1F6F324**.

Referência: Processo nº 23070.041119/2023-17 SEI nº 4062042

LUIZ FERNANDO DA CUNHA CINTRA

Uma estratégia de pós-processamento para seleção de regras de associação para descoberta de conhecimento

Dissertação apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Programa de Pós-Graduação em Ciência da Computação.

Área de concentração: Ciência da Computação.

Linha de pesquisa: Sistemas Inteligentes e Aplicações.

Orientador: Prof. Rogerio Lopes Salvini

Goiânia
2023

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Cintra, Luiz Fernando da Cunha
Uma estratégia de pós-processamento para seleção de regras de associação para descoberta de conhecimento [manuscrito] / Luiz Fernando da Cunha Cintra. - 2023.
108 f.

Orientador: Prof. Dr. Rogerio Lopes Salvini.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2023.
Bibliografia. Apêndice.
Inclui tabelas, algoritmos, lista de figuras, lista de tabelas.

1. association rules. 2. post-processing. 3. arm. 4. grouping. I. Salvini, Rogerio Lopes, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **16** da sessão de Defesa de Dissertação de **Luiz Fernando da Cunha Cintra**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos vinte e dois dias do mês de agosto de dois mil e vinte e três, a partir das catorze horas, na sala 151 do INF, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Uma estratégia de pós-processamento para seleção de regras de associação para descoberta de conhecimento"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Rogerio Lopes Salvini (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Eduardo José Aguilar Alonso (ICT/UNIFAL), membro titular externo; Professor Doutor Thierson Couto Rosa (INF/UFG), membro titular interno. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Rogerio Lopes Salvini, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos vinte e dois dias do mês de agosto de dois mil e vinte e três.
TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Rogerio Lopes Salvini**, **Professor do Magistério Superior**, em 22/08/2023, às 16:06, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luiz Fernando Da Cunha Cintra**, **Usuário Externo**, em 22/08/2023, às 16:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thierson Couto Rosa**, **Professor do Magistério Superior**, em 22/08/2023, às 16:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo José Aguilar Alonso**, **Usuário Externo**, em 22/08/2023, às 16:12, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3937607** e o código CRC **862E17D9**.

Referência: Processo nº 23070.041119/2023-17 SEI nº 3937607

LUIZ FERNANDO DA CUNHA CINTRA

Uma estratégia de pós-processamento para seleção de regras de associação para descoberta de conhecimento

Dissertação defendida no Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás como requisito parcial para obtenção do título de Mestre em Programa de Pós-Graduação em Ciência da Computação, aprovada em 22 de Agosto de 2023, pela Banca Examinadora constituída pelos professores:

Prof. Rogerio Lopes Salvini
Instituto de Informática – UFG
Presidente da Banca

Prof. Thierson Couto Rosa
Instituto de Informática – UFG

Prof. Eduardo José Aguilar Alonso
Instituto de Ciência e Tecnologia – UNIFAL

Luiz Fernando da Cunha Cintra

Graduou–se em Ciência da Computação na UFG - Universidade Federal de Goiás. Durante sua graduação, foi monitor no Instituto de Informática da UFG. A área de pesquisa durante o Mestrado foi mineração de dados, mais especificamente em Mineração de Regras de Associação. Atualmente trabalha como desenvolvedor de software em programas de automação em faturamento na empresa ZG Soluções.

Resumo

da Cunha Cintra, Luiz Fernando. **Uma estratégia de pós-processamento para seleção de regras de associação para descoberta de conhecimento**. Goiânia, 2023. 108p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

Mineração de regras de associação (ARM, em inglês) é um método tradicional de mineração de dados que fornece informações sobre associações entre itens em bases de dados transacionais. Um problema conhecido da ARM é a grande quantidade de regras geradas, necessitando assim de abordagens para pós-processar essas regras para que um especialista humano seja capaz de analisar as associações encontradas. Além disso, em alguns contextos, o especialista do domínio está interessado em investigar como uma variável de interesse está relacionada às outras variáveis de uma base de dados. Em uma análise exploratória baseada em ARM, isso implica em buscar as associações em que um item de interesse aparece em qualquer parte da regra. Poucos métodos possuem o foco em pós-processar as regras geradas visando um item de interesse. O presente trabalho busca destacar as associações relevantes de um determinado item visando trazer conhecimento sobre o seu papel por meio das suas interações e relações em comum com os demais itens. Para isso, este trabalho propõe uma estratégia de pós-processamento de regras de associação, que seleciona e agrupa regras orientadas a um determinado item de interesse fornecido por um especialista de um domínio de conhecimento. Além do mais, é também apresentado uma forma gráfica para que as associações entre regras e agrupamentos de regras encontrados sejam mais facilmente visualizados e interpretados. Quatro estudos de casos mostram que o método proposto é admissível e consegue reduzir o número de regras relevantes para uma quantidade gerenciável, permitindo a análise do especialista do domínio. Grafos evidenciando as relações entre os agrupamentos foram gerados em todos os estudos de casos e facilitam a análise dos mesmos.

Palavras–chave

regras de associação, pós-processamento, arm, agrupamento.

Abstract

da Cunha Cintra, Luiz Fernando. **A post-processing strategy for selecting association rules for knowledge discovery**. Goiânia, 2023. 108p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

Association rule mining (ARM) is a traditional data mining method that provides information about associations between items in transactional databases. A known problem of ARM is the large amount of rules generated, thus requiring approaches to post-process these rules so that a human expert is able to analyze the associations found. In some contexts the domain expert is interested in investigating only one item of interest, in these cases a search guided by the item of interest can help to mitigate the problem. For an exploratory analysis, this implies looking for associations in which the item of interest appears in any part of the rule. Few methods focus on post-processing the generated rules targeting an item of interest. The present work seeks to highlight the relevant associations of a given item in order to bring knowledge about its role through its interactions and relationships in common with the other items. For this, this work proposes a post-processing strategy of association rules, which selects and groups rules oriented to a certain item of interest provided by an expert of a domain of knowledge. In addition, a graphical form is also presented so that the associations between rules and groupings of rules found are more easily visualized and interpreted. Four case studies show that the proposed method is admissible and manages to reduce the number of relevant rules to a manageable amount, allowing analysis by domain experts. Graphs showing the relationships between the groups were generated in all case studies and facilitate their analysis.

Keywords

association rules, post-processing, ARM, grouping

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

CAPÍTULO 1

Introdução

Este capítulo apresenta os aspectos iniciais sobre o objeto de pesquisa da seguinte maneira. Na Seção 1.1 o assunto de pesquisa é contextualizado; a Seção 1.2 apresenta a justificativa; a Seção 1.3 discorre sobre a motivação do estudo; a Seção 1.4 expõe o problema a ser estudado; a Seção 1.5 especifica os objetivos do estudo. Por fim, a Seção 1.6 apresenta a estrutura do restante do documento.

1.1 Contextualização

Um método bem conhecido para extrair padrões a partir de um conjunto de dados é a mineração de regras de associação (*association rule mining* – ARM) [2]. Uma regra de associação é uma regra do tipo $A \rightarrow B$, onde A é chamado de antecedente e B de consequente da regra. Tais regras representam associações frequentes em uma base de dados. Inicialmente o método foi proposto para descobrir associações em dados de cesta de supermercado, porém durante as últimas décadas foi aplicado para diversos outros domínios, tais como, construção [12], criação de produtos [36], educação [47], esportes [81], manutenção de edifícios [24] [83] [84] [85] [86], medicina [11] [80] e planejamento urbano [6].

O algoritmo *Apriori* [3] foi um dos primeiros, e é o mais conhecido e também um dos mais simples algoritmos de ARM. Sua ideia central é descobrir associações que sejam frequentes e significativas em uma base de dados transacional. O conceito de transação em ARM significa originalmente uma transação comercial em dados de compras, mas outros conceitos podem ser convertidos em transações, como uma ficha de um paciente, eventos gerados por sensores, entre outros. Portanto, é comum que transações sejam mapeadas como linhas em uma tabela. Cada transação é composta de itens, no modelo original um item é um produto que faz parte de uma transação comercial. Ao se aplicar o *Apriori* à outros domínios, os itens correspondem as variáveis de uma base de dados e usualmente são tratados como colunas de uma base de dados tabular. Cada combinação de itens é avaliada utilizando o valor calculado da medida de suporte, que é a frequência em que os itens aparecem juntos em uma transação, mantendo apenas as combinações que possuam

valor acima de um mínimo especificado pelo usuário. A partir das combinações obtidas, o *Apriori* constrói as regras levando em conta um limite mínimo do valor da medida de confiança especificado pelo usuário. A confiança em ARM significa a probabilidade condicional do consequente acontecer dada a ocorrência do antecedente da regra [3]. O processo de minerar regras de associação geralmente envolve dois tipos de usuários: um especialista no domínio de aplicação dos dados (usuário final) e um especialista em mineração de dados (analista) [44].

Outras medidas foram propostas para substituir o suporte e confiança, porém cada uma delas captura determinados aspectos e possui propriedades específicas, sendo assim, não há uma medida que seja boa para qualquer domínio de aplicação [69]. As medidas utilizadas geralmente são dividas em dois grupos, as medidas objetivas e as medidas subjetivas [48]. Uma medida é objetiva (guiada pelos dados) se seu valor pode ser calculado apenas com as informações dos próprios dados. O suporte e a confiança são medidas objetivas. De forma oposta, uma medida é subjetiva (guiada pelo usuário) se seu valor só pode ser calculado utilizando o conhecimento do usuário. Medir a surpresa de uma regra em relação a um conhecimento prévio estabelecido é um exemplo de medida subjetiva [48]. As medidas subjetivas majoritariamente utilizam-se do conhecimento prévio do usuário final.

Diversos algoritmos que mineram regras de associação utilizam algumas medidas objetivas para decidir sobre a qualidade de uma regra. Dessa forma, é possível considerar um conjunto de medidas objetivas como uma função objetivo[1] e, portanto, pode-se tratar a tarefa de minerar regras de associação como um problema de otimização multiobjetivo e usar algoritmos que solucionem essa classe de problemas como, por exemplo, algoritmos evolucionários [53].

A mineração de regras de associação, em geral, é uma atividade de natureza exploratória, ou seja, não existe um item específico (variável de classe) para guiar a construção de um preditor, como nos problemas de classificação. Os algoritmos para minerar as regras de associação buscam qualquer padrão que seja estatisticamente relevante [26]. Esta característica pode levar a um número de regras elevado, tornando difícil a manipulação e análise posterior. Por outro lado, existem estudos cujo objetivo é observar as relações de determinado item de interesse específico dos dados. Portanto, um método capaz de analisar as regras de associação geradas baseando-se em um item de interesse sem perder toda a capacidade de ARM pode ser mais eficaz, pois diminui o espaço de busca em que o especialista irá trabalhar. Por capacidade total de ARM são referidas duas características: 1) diferentemente dos métodos que geram regras de classificação, o item de interesse pode aparecer no antecedente ou no consequente de uma regra de associação;

[1] Uma função onde se deseja maximizar ou minimizar conforme o objetivo do problema

2) regras que não possuam um item de interesse mas que compartilham itens com outras regras que o possuem devem ser mantidas (apenas um filtro simples pelo item de interesse iria ocultar tais relações).

1.2 Justificativa

Um dos principais problemas relativos ao uso de regras de associação é o enorme número de regras que podem ser geradas pelos algoritmos de mineração de regras de associação. Este problema é bem conhecido e tem sido estudado por mais de 20 anos [5]. Porém, uma solução que seja boa para todos os tipos de aplicações é provavelmente impossível, dado que os objetivos podem ser variados, como pode ser visto pelos diferentes tipos de medidas de avaliação [72].

Como geralmente o processo de análise das regras é feito pelo usuário final, isso se torna um problema, levando a um processo de análise custoso e muitas vezes ineficiente, sendo inviável, em alguns casos, analisar todas as regras manualmente. Portanto, métodos para pós-processar regras de associação foram propostos ao longo dos anos, como será visto na revisão sistemática do Capítulo 3.

Poucos estudos foram encontrados em pós-processamento baseado em um item de interesse dos dados. A maioria dos estudos existentes focados nessa abordagem geralmente fixam o item de interesse no consequente [8] [12] [31], apenas um trabalho [80] fixa um item de interesse no antecedente. Pode-se argumentar que esses estudos não utilizam toda a capacidade exploratória do método, pois quando o item de interesse é fixado no consequente observamos apenas o que geralmente leva a ocorrência do item de interesse. O item de interesse posicionado no antecedente permite extrair as consequências da presença do item de interesse. Por exemplo, no trabalho de Wei e Scott [80] o item interesse é uma vacina e o consequente são eventos adversos, nesse caso o consequente possui mais de um item enquanto o antecedente possui apenas um item (o item de interesse). Portanto seria importante um estudo sobre como o item de interesse no antecedente pode contribuir para geração de conhecimento especialista no domínio.

Recentemente alguns estudos focaram-se no relacionamento entre regras, tais como, buscar meta-regras (regras de regras) [8] [23]. Berka [8] argumenta que as meta-regras conseguem gerar um conhecimento complementar as regras de associação sobre um determinado conceito. Grabot [27] relata um interesse de especialistas na existência (ou não) de duplas de regras do tipo $A \rightarrow B$ e $B \rightarrow A$, nomeadas de regras simétricas (ou bidirecionais). Porém um estudo mais aprofundado ainda é necessário, visto que o método proposto por Grabot não foca necessariamente em obter essas duplas. Além disso, não foram encontrados estudos que buscam relação similar a apontada por Grabot para regras com mais de um item no antecedente.

1.3 Motivação

Em estudos anteriores [10], [15] notou-se a necessidade de um método automático capaz de pós-processar regras de associação com foco em um item de interesse e de forma a preservar a natureza exploratória do método de ARM. Ambos trabalhos buscam associações relacionadas a um item de interesse, porém pela falta de métodos existentes para esse tipo de exploração o trabalho de análise se torna custoso.

Um trabalho anterior foi conduzido no Trabalho de Conclusão de Curso do autor, onde um processo automatizado foi proposto para resolver o problema [14]. Algumas limitações deste trabalho foram: a representação das regras que dificultava a visualização da influência do item de interesse, bem como o número ainda elevado de regras de associação, e a necessidade de um estudo mais aprofundado sobre o significado de cada grupo proposto. Portanto esse trabalho visa melhorar os problemas encontrados anteriormente.

Além disso, a disponibilidade da base realizada no trabalho de Castro e colaboradores [10] nos permite realizar os experimentos em uma base de dados real com a qual o grupo de pesquisa já possui familiaridade.

1.4 Problema de pesquisa

O problema de pesquisa consiste em, a partir de um conjunto de regras de associação geradas por um algoritmo de ARM, selecionar as regras que, direta ou indiretamente, estejam relacionadas a um item de interesse. Deseja-se extrair as regras que mostrem como a ocorrência do item de interesse influencia nos demais itens e como os demais itens são influenciados pelo item de interesse, a fim de destacar apenas as mais relevantes ao especialista.

Para esse trabalho é necessário aproveitar o caráter exploratório de ARM, ou seja, o item de interesse pode aparecer tanto no antecedente quanto no consequente. Essa característica é uma das principais diferenças entre ARM e regras de classificação onde, nestas últimas, há um viés indutivo na geração das regras em função do item de interesse (variável de classe) [26].

1.5 Objetivos

O objetivo principal desse trabalho é propor um método de pós-processamento de regras de associação que selecione e organize regras de interesse para um especialista de domínio baseadas em um item alvo específico. O objetivo principal pode ser dividido nos objetivos específicos abaixo:

- Evidenciar regras que mostrem para o especialista a influência do item de interesse nos demais itens
- Evidenciar regras que mostrem para o especialista a influência dos demais itens no item de interesse
- Averiguar diferentes métricas na seleção das regras de interesse
- Averiguar formas de visualização gráfica das regras de interesse a fim de facilitar a análise do especialista

1.6 Estrutura do documento

O restante do trabalho está estruturado da seguinte maneira. O Capítulo 2 fornece um resumo sobre regras de associação, uma explicação sobre o algoritmo *Apriori* e sobre as medidas de interesse para avaliar regras. No Capítulo 3 é reportada uma revisão sistemática da literatura disponível sobre o tema de pós-processamento de regras de associação. A apresentação do método proposto é feita no Capítulo 4. Os resultados são apresentados e discutidos no Capítulo 5. Por fim, no Capítulo 6 uma conclusão é feita, especificando possíveis trabalhos futuros.

CAPÍTULO 2

Fundamentação teórica

Um tipo de tarefa bem estabelecida da mineração de dados é a geração de regras de associação a partir de uma base de dados de transações. Por meio dessa forma de descoberta de conhecimento podemos encontrar não apenas associações que nunca foram cogitadas, como também associações que confirmem ou contradizem o conhecimento prévio [72]. Este capítulo apresenta as definições formais do que são regras de associação, qual a problemática para realizar a mineração delas, alguns métodos de mineração existentes e como as regras são avaliadas.

O capítulo está estruturado da seguinte maneira. Na Seção 2.1 uma definição de regra de associação é apresentada, depois na Seção 2.2 o algoritmo *Apriori* é descrito. Finalmente, na Seção 2.3 algumas medidas de interesse para as regras são apresentadas.

2.1 Regras de associação

Uma regra de associação é uma regra do tipo "se-então", formalizada em 1993 por Agrawal e colaboradores [2], que mostra padrões ou relacionamentos em um conjunto de transações. Uma formalização mais geral também proposta por Agrawal e colaboradores em 1994 [3] define que uma regra de associação tem a forma $A \rightarrow B$, onde A e B são conjuntos de itens (*itemsets*), ou seja, $A = \{a_1, a_2, ..., a_n\}$ e $B = \{b_1, b_2, ..., b_m\}$, sendo a_i e b_j itens de uma base de dados. O tamanho de uma regra é quantidade de itens presentes nela. Dado que I é o conjunto de todos os itens da base de dados, $A \subset I$, $B \subset I$ e $A \cap B = \emptyset$, ou seja, antecedentes e consequente são itens da base de dados e não possuem repetição de itens [3].

Seja D um conjunto de transações, onde cada transação T é um *itemset* tal que $T \subset I$. Uma transação T *contém* um *itemset* A se $A \subset T$. O suporte s de uma regra $A \rightarrow B$ é a porcentagem de transações em D ($s\%$) que contêm os itens de $A \cup B$, ou seja, a porcentagem de transações em que os itens do antecedente e consequente da regra ocorrem juntos. Já a confiança c de uma regra $A \rightarrow B$ é a porcentagem de transações em D ($c\%$) que caso contenha A também contêm B [3]. As Equações 2-1 e 2-2, adaptadas de [17] descrevem o suporte e a confiança de uma regra $A \rightarrow B$.

$$Suporte(A \rightarrow B) = P(A,B) \tag{2-1}$$

$$Confiança(A \rightarrow B) = P(B|A) \tag{2-2}$$

O problema de minerar regras de associação se resume a gerar, a partir de um conjunto de transações D, todas as regras que tenham um suporte e confiança maior que o suporte mínimo e a confiança mínima, ambos especificados pelo usuário [2]. Sendo assim, o problema pode ser decomposto em dois subproblemas [3]:

1. Encontrar todos os *itemsets* que apresentem um suporte maior ou igual ao *suporte mínimo*. Esses conjuntos são chamados de *itemsets frequentes*.
2. A partir dos *itemsets* frequentes, gerar as regras que possuam confiança maior ou igual a confiança mínima estabelecida.

Tabela 2.1: *Exemplo de um conjunto de transações*

Transação	a_1	a_2	a_3
1	✓	✓	✗
2	✓	✗	✓
3	✓	✓	✓
4	✓	✗	✓
5	✓	✓	✗

Através do conjunto de transações da Tabela 2.1, definindo o suporte mínimo em 10% e a confiança mínima em 30%, é possível obter uma regra como a demonstrada a regra abaixo.

$$a_1, a_2 \rightarrow a_3$$

A seguir, o cálculo do suporte e da confiança da regra:

$$Suporte = \tfrac{1}{5} = 0,20$$

$$Confiança = \tfrac{1}{3} = 0,33$$

Existem diversos algoritmos de ARM baseados no modelo Suporte/Confiança para gerações de *itemsets* frequentes, tais como como o *Apriori* [3], *Eclat* [82] e *FP-Growth* [32]. Na Seção 2.2 será detalhado o algoritmo *Apriori*, um dos algoritmos mais comuns e antigos. Além disso, o problema de ARM também pode ser modelado como um problema de otimização multi-objetivo e pode ser resolvido por outras estratégias,

tais como algoritmos evolucionários [53]. Posteriormente, outros algoritmos foram propostos para buscarem padrões diferentes, o livro de Tan e colaboradores [71] destaca alguns deles, tais como, regras de associação difusa (*fuzzy association rules*), regras de associação ponderada (*weighted association rules*), regras de associação espacial (*spatial association rules*), regras de associação sequencial (*sequential association rules*), entre outras.

Regras de associação também podem ser entendidas como um caso mais geral de regras de classificação. Em vez do consequente da regra ser um item alvo pré-definido pelo usuário, ele pode ser qualquer item da base de dados ou ainda um conjunto de itens [53]. Freitas no documento de posicionamento [26] argumenta sobre as diferenças entre regras de classificação e o modelo de Suporte/Confiança de regras de associações, que vão além da posição de um item alvo. Classificação é uma tarefa não determinística e mal estruturada enquanto ARM é uma tarefa determinística e bem estruturada. É importante frisar que dentro do ARM existem tarefas não determinísticas e mal estruturadas, por exemplo, definir se uma regra é relevante. Classificação tem preocupações com *overfitting* e *underfitting* enquanto ARM não, além disso, para classificação é necessário um viés indutivo o que não acontece em ARM. Liu e colaboradores propuseram em 1998 uma interseção entre regras de associação e regras de classificação [46], nomeada de *Classification based Association* (CBA), que minera *Class Association Rules* (CARs) e partir delas constrói um *classificador associativo* (*Associative Classifier*). As principais características que diferem a classificação associativa de outras abordagens tradicionais de classificação é a fácil interpretabilidade dos resultados obtidos e um taxa de erro reduzida [1]. O estudo sobre classificação associativa está fora do escopo desse trabalho, uma extensa revisão sobre o tema pode ser encontrado em [1].

2.2 Algoritmo Apriori

O problema da ARM consiste em descobrir os *itemsets* frequentes, portanto o espaço de busca total equivale a todas as combinações possíveis de itens. Considerando itens como um tipo booleano, que indica presença (ou ausência) do item, temos que $2^{|I|}$ subconjuntos podem ser obtidos de I, o conjunto de todos os itens [2]. A Figura 2.1 mostra todas as combinação possíveis para 5 itens, $2^5 = 32$.

O algoritmo *Apriori* [3] foi proposto para melhorar questões de desempenho do algoritmo anterior o AIS [2]. Para isso o algoritmo se baseia na propriedade de monotonicidade do suporte, ou seja, dados dois *itemsets* A e B, se $A \subset B$ então $Suporte(B) \leq Suporte(A)$. Isso quer dizer que os subconjuntos de um *itemset* frequente é frequente [3]. Portanto, se um subconjunto não é frequente é impossível que uma extensão desse subconjunto seja frequente. Por exemplo, dado que A é subconjunto de B (B é uma extensão

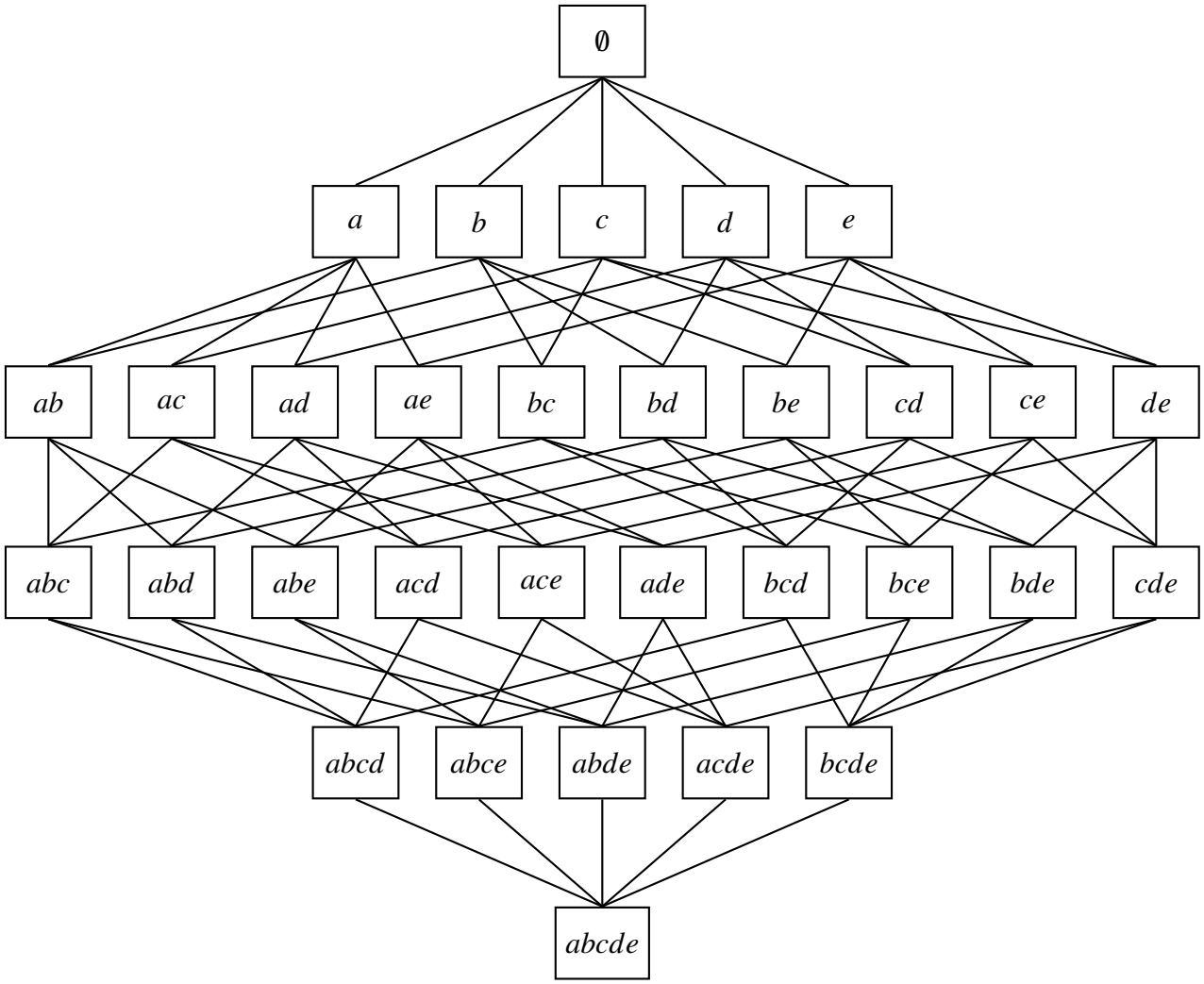

Figura 2.1: *Todas as combinações possíveis para* 5 *itens.*

de A), se A não for frequente então o $Suporte(A) < s$, sendo s o suporte mínimo, portanto pela propriedade do suporte o $Suporte(B) \leq Suporte(A) < s$, logo B também não é frequente. Usando essa propriedade o algoritmo *Apriori* usa uma estratégia de gerar os *itemsets* frequentes apenas combinando os *itemsets* frequentes do passo anterior [3]. A Figura 2.2 ilustra um caso em que um item não é frequente. Os nós em cinza mostram as combinação que não serão analisadas. Após constatar que o item *e* não é frequente 15 combinações deixam de ser feitas. O *Apriori* segue de maneira geral os seguintes passos para gerar regras de associação:

1. Primeiro calcula-se o suporte dos itens individuais e determina se atendem ao suporte mínimo, ou seja, se eles são *itemsets frequentes* com cardinalidade 1.
2. Começando com $k = 2$, repita os passos 3 e 4 até não encontrar mais nenhum novo *itemset* frequente.
3. A partir dos *itemsets* frequentes de cardinalidade $k-1$, gera-se *itemsets potencialmente frequentes* de cardinalidade k, nomeados também de *itemsets candidatos*.

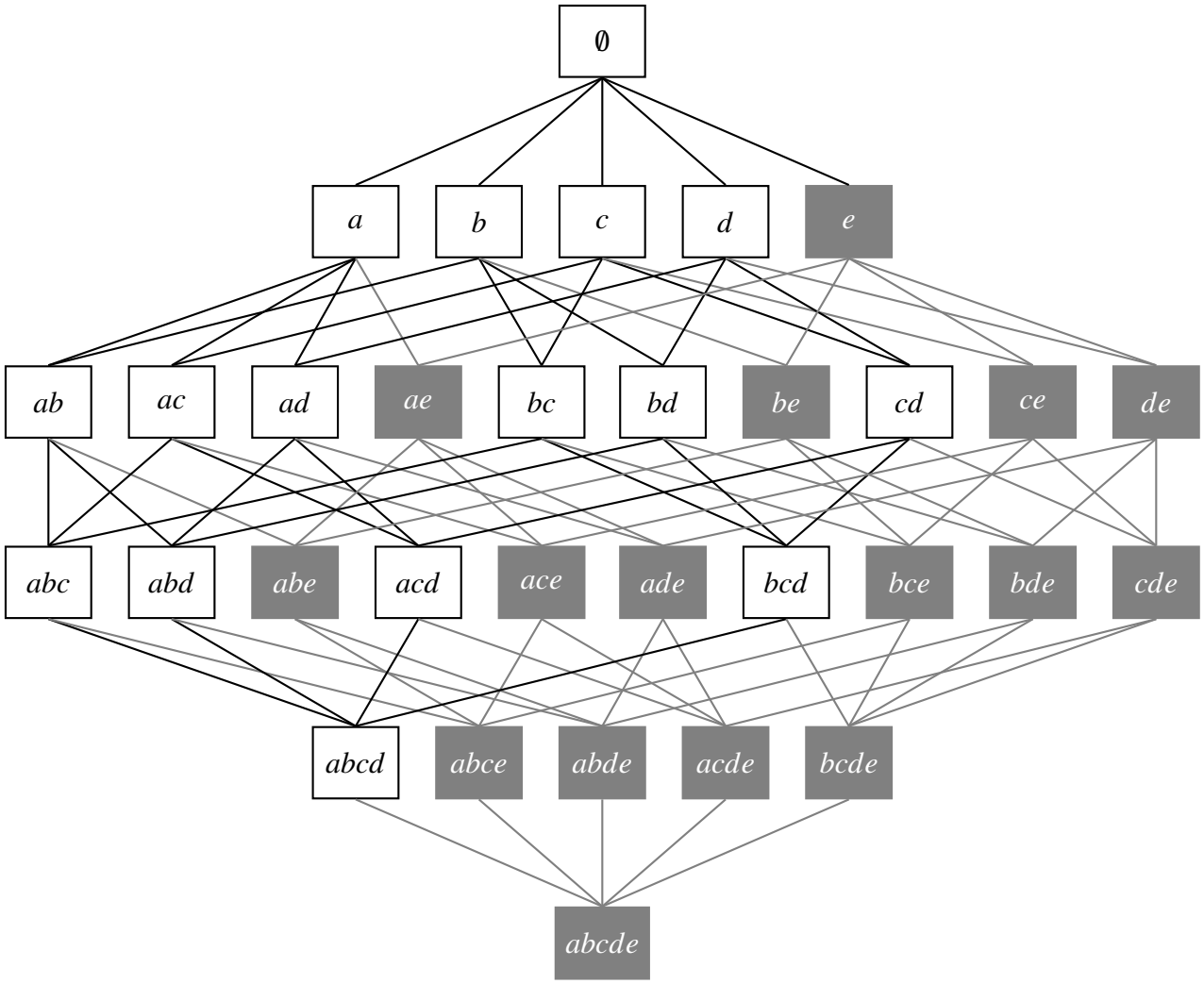

Figura 2.2: *Redução do espaço de busca através do monotonicidade do suporte.*

Ao mesmo tempo em que os *itemsets* candidatos são gerados também é calculado o suporte deles, evitando mais uma passagem sobre os dados.

4. Seleciona dentre os *itemsets* candidatos de cardinalidade k os *itemsets* frequentes e incremente k em 1.
5. Após obter todos os *itemsets* frequentes, para cada *itemset* frequente com $k \geq 2$ realiza as combinações sem repetições [1] dos itens para formar regras.
6. Para cada regra calcula a confiança. Se atender a confiança mínima a regra é mantida, caso contrário é descartada.

O *Apriori* usa ambas as restrições, de suporte e de confiança, por que o número de regras possíveis é combinatorial, conforme relatado no livro de Tan e colaboradores

[1] Repetição em um mesmo lado da regra, ou seja, $a_1 \rightarrow a_2 \neq a_2 \rightarrow a_1$ mas $a_1, a_2 \rightarrow a_3 = a_2, a_1 \rightarrow a_3$

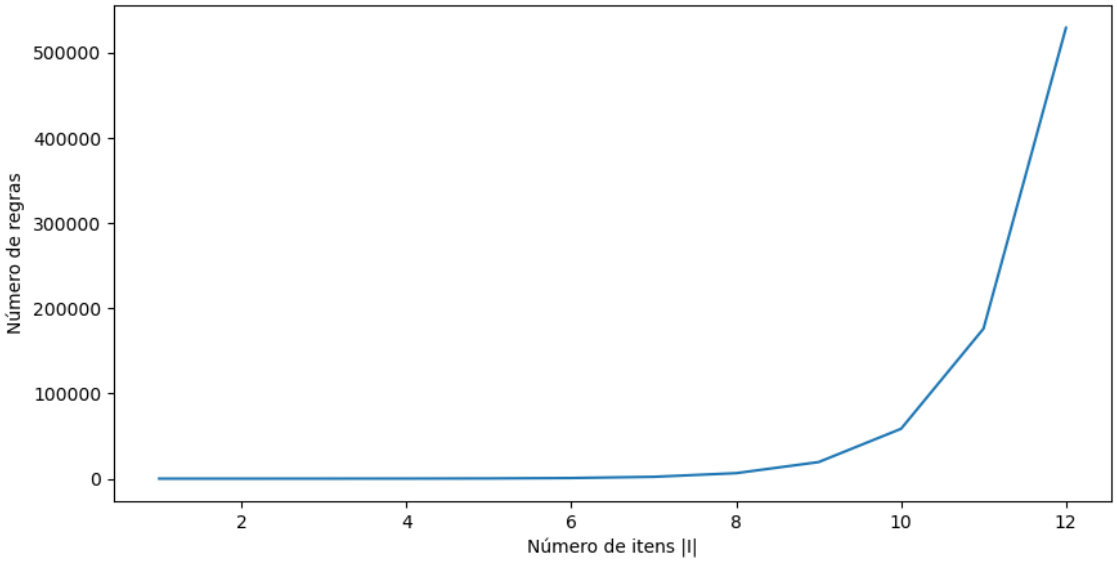

Figura 2.3: *Variação no número de regras conforme o número de itens.*

[71] para um conjunto de itens I o número de regras possíveis é dado pela Equação 2-3. Por exemplo, para 12 itens teremos $3^{12} - 2^{12} + 1 = 527.346$ regras possíveis.

$$3^{|I|} - 2^{|I|} + 1 \tag{2-3}$$

A Figura 2.3 mostra o crescimento do número de regras conforme aumenta o número de itens. Este crescimento exponencial de regras geradas pode ser problemático para uma análise manual posterior.

Como o trabalho se concentra no passo de pós-processamento dessas regras uma explicação sobre o *Apriori* com detalhes de implementação está disponível no Apêndice A.

2.3 Medidas de Avaliação

Desde a proposta de regras de associação, em 1993, diversas medidas de avaliação das regras foram criadas, estudadas ou testadas, visando ajudar o usuário final a obter um conjunto relevante de regras [19], sendo o uso destas medidas a forma mais comum para lidar com o problema do alto número de regras de associação que são geradas. Segundo Tew e colaboradores [72], 61 medidas foram encontradas em uso e cada medida visa aspectos diferentes sobre os itens do banco de dados. Em um dos primeiros trabalhos nesse tópico, Tan e colaboradores [69] relataram que medidas diferentes produzem *rankings* diferentes, portanto, têm objetivos diferentes e não há uma medida que seja boa

para todos os domínios. Porém, também reportaram que algumas medidas produziam um *ranking* parecido, trabalhos posteriores mostraram que mais medidas produzem uma ordenação parecida [67] [72]. O trabalho de Carvalho e colaboradores [19] usa esses grupos de medidas para, utilizando uma medida de cada grupo, gerar um *ranking* mais assertivo.

Na literatura as medidas são divididas em dois tipos, medidas *objetivas* e medidas *subjetivas* [72]. Medidas objetivas dependem apenas da natureza estrutural das regras de associação. Propriedades estatísticas, como suporte e confiança, são exemplos de medidas objetivas. Medidas subjetivas são mais focadas no usuário, por exemplo, dizer que uma regra é mais contraditória em relação ao conhecimento à priori, sendo assim, revelando uma novidade e uma relevância maior [48]. Embora medidas subjetivas pareçam bem vantajosas, elas tem alguns problemas, primeiro se deve de alguma forma colocar especificações do conhecimento do usuário no processo de seleção. Além disso, podem não permitir que padrões inesperados apareçam [48].

Tabela 2.2: *Medidas objetivas usuais para avaliação de regras*

Medida	**Fórmula**
Cobertura($A \rightarrow B$)	$P(A)$
Convicção($A \rightarrow B$)	$\frac{P(A)P(\neg B)}{P(A,\neg B)}$
Coeficiente de correlação ($A \rightarrow B$)	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)P(\neg A)P(\neg B)}}$
Leverage($A \rightarrow B$)	$P(B\|A) - P(A)P(B)$
Lift($A \rightarrow B$)	$\frac{P(A,B)}{P(A)P(B)}$
Odds Ratio($A \rightarrow B$)	$\frac{P(A,B)P(\neg A,\neg B)}{P(A,\neg B)P(\neg A,B)}$

A Tabela 2.2 baseada em [72] mostra algumas medidas objetivas. A *cobertura* mostra a porcentagem de transações que a regra pode ser aplicada. A *convicção* compara a possibilidade de A ocorrer sem B se eles forem dependentes, com a frequência que A ocorre sem B no banco de dados. Um valor igual a 1 significa que A e B são independentes. O intervalo que a convicção pode assumir é de 0 à $+\infty$. O *coeficiente de correlação* mostra o quão forte é o relacionamento entre pares de *itemsets*. Seu valor varia de -1 à 1, sendo -1 uma relação linear perfeitamente negativa, e 1 uma relação linear perfeitamente positiva. O valor 0 indica que não há relação entre os *itemsets*. *Leverage* mede a diferença entre a frequência que A e B aparecem juntos no banco de dados e o valor esperado se A e B fossem dependentes.

Nas seções a seguir, são destacadas as duas medidas (*Lift* e *Odds Ratio*), utilizadas nos estudos de casos apresentados para a validação do método proposto no presente trabalho.

Lift

O *lift* é uma medida bastante conhecida e frequentemente usada em ARM. Esta medida mostra a melhoria trazida por uma regra em relação a uma resposta aleatória [75], ou seja, se a relação aparece porque o antecedente e o consequente são dependentes e não por que um deles, ou até ambos, ocorrem com muita frequência nos dados. A Equação 2-4, fornecida por [17], mostra como o *lift* é calculado. O *lift* divide a probabilidade conjunta de A e B, pela probabilidade de A e de B aparecerem juntos considerando-os independentes. Portanto, o valor do *lift* mostra uma relação de dependência entre o antecedente e o consequente da regra.

$$Lift(A \rightarrow B) = \frac{Suporte(A \cup B)}{Suporte(A) \times Suporte(B)} = \frac{P(A,B)}{P(A)P(B)} \tag{2-4}$$

Para um melhor entendimento desta relação, comparemos com a medida de confiança apresentada na Equação 2-2 e reescrita abaixo:

$$Confiança(A \rightarrow B) = \frac{P(A,B)}{P(A)}$$

A confiança compara o número de ocorrências conjuntas dos dois *itemsets* da regra com o número de ocorrências do *itemset* que aparece como antecedente da regra. No entanto, se o consequente B for muito frequente, pode haver muitas transações tais que $A \rightarrow B$ sem necessariamente A e B serem dependentes, ou seja, regras com alta confiança, mas com o consequente e o antecedente independentes. Por exemplo a Tabela 2.3 mostra alguns *itemsets* com seus respectivos valores de suporte.

Tabela 2.3: *Tabela de itemsets*

Transação	**Suporte**
A_1	60%
A_2	70%
A_3	40%
A_4	30%

Na Tabela 2.4 a regra com maior confiança tem *lift* igual a 1, isso se dá porque o suporte tanto de A_1 quanto A_2 é maior se compararmos com o suporte das outras transações, ou seja, aumenta a chance que essa associação ocorra simplesmente porque os *itemsets* A_1 e A_2 são consideravelmente frequentes. As demais regras possuem uma

confiança pior, porém, o *lift* está distante de 1, isso indica que é mais provável que de fato exista a correlação da associação representada pela a regra.

Tabela 2.4: *Tabela de Regras*

Regra	**Suporte**	**Confiança**	***Lift***
$A_1 \rightarrow A_2$	42%	70%	1,00
$A_1 \rightarrow A_3$	14%	23%	0,58
$A_1 \rightarrow A_4$	36%	60%	2,00

Um valor de *lift* igual a 1 para uma regra $A \rightarrow B$, quer dizer que o consequente da regra (B) é independente do seu antecedente (A), ou seja, não há relação entre o antecedente e o consequente. Se o *lift* for maior que 1, significa que antecedente e consequente estão correlacionados. Porém, se o *lift* for menor que 1, a associação do antecedente com a negação do consequente ($A \rightarrow \neg B$) apresentará correlação, ou seja, embora a regra gerada seja $A \rightarrow B$, o valor de *lift* altera o sentido da associação para $A \rightarrow \neg B$.

Quanto mais distante de 1, maior a dependência entre o consequente e o antecedente, por isso, geralmente procura-se regras com valores de *lift* maiores que 1, chamadas de *regras úteis*. Porém, regras com *lift* menor do 1 também tem importância pois elas mostram que é improvável que os itens contidos nelas ocorram juntos.

Odds ratio

Assim como o *lift* o *odds ratio* também busca medir a independência entre antecedente e consequente. O *odds Ratio*, em regras de associação, é a razão entre: a chance (*odds*) do consequente ocorrer dado que o antecedente ocorre ($\frac{P(A,B)}{P(A,\neg B)}$), e a chance do consequente ocorrer dado que o antecedente não ocorre ($\frac{P(\neg A,B)}{P(\neg A,\neg B)}$) [70]. Portanto se a chance do consequente (B) ocorrer dado o antecedente (A) for igual ou próxima da chance do consequente ocorrer dado que antecedente não ocorra ($\neg A$), então muito provavelmente o consequente (B) ocorre independentemente do antecedente (A) [70]. Isso implica que o valor de *odds ratio* seja 1 ou próximo de 1 para essa situação (Equação 2-5).

$$Odds\ ratio(A \rightarrow B) = \frac{\frac{P(A,B)}{P(A,\neg B)}}{\frac{P(\neg A,B)}{P(\neg A,\neg B)}} = \frac{P(A,B) \times P(\neg A,\neg B)}{P(A,\neg B) \times P(\neg A,B)} \tag{2-5}$$

O *odds ratio* pode gerar valores no intervalo $[0,\infty)$, sendo que quanto mais distante de 1 maior a dependência entre antecedente e consequente. Um valor de *odds ratio* maior que 1 significam que a chance do consequente ocorrer dado que o antecedente ocorreu ($\frac{P(A,B)}{P(A,\neg B)}$) é maior do que a chance do consequente ocorrer dado que o antecedente

não ocorreu ($\frac{P(\neg A,B)}{P(\neg A,\neg B)}$), ou seja, a associação $A \rightarrow B$ de fato acontece. Já um valor de *odds ratio* menor que 1 significa que a chance de um consequente ocorrer dado que o antecedente ocorreu ($\frac{P(A,B)}{P(A,\neg B)}$) é menor do que a chance do consequente ocorrer dado que o antecedente não ocorreu ($\frac{P(\neg A,B)}{P(\neg A,\neg B)}$), ou seja, de forma similar ao *lift* a regra muda seu significado para $\neg A \rightarrow B$.

CAPÍTULO 3

Revisão sistemática

Embora a área de pós-processamento de regras de associação seja estudada por mais de 20 anos [5], não foram encontradas revisões que abordassem todos os tipos de pós-processamentos. Portanto, o estudo atual cataloga alguns métodos para pós-processamento de regras de associação. Além disso, também são apresentados os desafios atuais e possíveis trabalhos futuros. Pela grande quantidade de trabalhos retornados o estudo verificou apenas os artigos de dos últimos 8 anos aproximadamente (2015-2023).

O restante do trabalho está organizado da seguinte maneira. A Seção 3.1 define qual a metodologia seguida. A Seção 3.2 reporta a quantidade de artigos obtidos e os artigos selecionados. A Seção 3.3 mostra uma análise sobre os artigos selecionados. Por fim, a Seção 3.4 resume o que foi discutido nas seções anteriores e envidencia os trabalhos futuros.

3.1 Metodologia da Pesquisa

Para obter os trabalhos recentes mais relevantes, uma revisão sistemática da literatura foi aplicada. Este método foi escolhido por ser uma forma bem estabelecida e validável para obtenção, avaliação e interpretação dos estudos produzidos até o momento. Para guiar essa revisão sistemática da literatura a diretriz de [38] foi utilizada. A diretriz define três estágios principais, o primeiro é definir o protocolo para revisão, depois deve-se conduzir a revisão e por fim documentar o que foi encontrado na análise.

As próximas subseções detalham algumas subtarefas do primeiro estágio. Para a realização dessa revisão sistemática o site Parsifal (https://parsif.al/) foi usado para auxiliar na aplicação da diretriz e a aplicação Mendeley (https://www.mendeley.com/) para leitura, gerenciamento dos documentos dos artigos e de referências bibliográficas.

3.1.1 Objetivo

A primeira tarefa realizada foi definir qual o objetivo da revisão sistemática. O objetivo principal é encontrar os principais métodos existentes na literatura para pós-

processamento de regras de associação, visando entender como eles são avaliados e quais as lacunas de pesquisas existentes.

3.1.2 Questões de pesquisa

A partir do objetivo relatado acima foram formuladas as seguintes questões para alcançar o objetivo da pesquisa.

- (P1) *"Quais os métodos existentes para pós-processamento de regras de associação?"*
- (P2) *"Existem métodos que se concentram em um item de interesse específico?"*
- (P3) *"Os métodos propostos geralmente são avaliados?"*
- (P4) *"Se eles são avaliados, como são avaliados?"*
- (P5) *"Se eles são avaliados, as avaliações podem ser consideradas suficientes?"*
- (P6) *"É possível definir quais são os melhores métodos?"*
- (P7) *"Quais a principais lacunas de pesquisas existentes na área?"*

A questão (P1) visa apenas estabelecer o que existe de relevante na área. (P2) visa verificar se existem métodos que se concentram na busca orientada por um item de interesse e caso existam detalhá-los. A pergunta (P3) busca verificar a qualidade geral dos estudos, isto é, se as propostas passam por uma avaliação mínima. As questões (P4) e (P5) dependem de (P3) ser pelo menos parcialmente verdadeira. (P4) tem como objetivo identificar se há uma avaliação padrão, enquanto (P5) visa uma análise crítica sobre tais avaliações. A pergunta (P6) tem como finalidade verificar se é possível definir uma estratégia de comparação para obter os melhores métodos. Por último, (P7) pretende identificar de modo geral as lacunas deixadas pelas propostas atuais.

3.1.3 Estratégia de busca

Após as questões definidas as seguinte bases foram escolhidas, *ACM Digital Library*, *IEEE Digital Library* e *Scopus*, todas são bases com grande relevância para computação.

Depois das bases escolhidas, foram definidas as palavras-chave e seus sinônimos. A primeira palavra-chave definida foi *association rule* e depois *post-processing*. Para association rule não foram definidos sinônimos, já para *post-processing* os seguintes sinônimos foram definidos: *post-mining*, *clustering*, *pruning*, *ranking*, *summarizing* e *visualization*. Exceto *post-mining* os sinônimos definidos na verdade são tipo de pós-processamento de regras de associação. Essa escolha foi feita porque em uma exploração prévia (sem um protocolo muito bem definido) notou-se que nem sempre o termo *post-processing* era utilizado.

Dessa forma, temos a seguinte *string* base, *"association rules AND (post-processing OR clustering OR post-mining OR pruning OR ranking OR summarizing OR visualization)"*. Para cada base de trabalhos foram necessárias adaptações. Com exceção da *IEEE* as buscas foram feitas procurando no título, resumo e palavras-chave. Para *IEEE*, por padrão, outros metadados também são utilizados. A escolha sobre limitar o período ocorreu já no estágio de condução devido ao grande número de artigos selecionados, por isso não foi aplicado um filtro de data nas *strings*.

3.1.4 Critérios de inclusão e exclusão

Nesta subseção, os critérios de exclusão e inclusão serão apresentados, ao todo 16 critérios de exclusão e 4 critérios de inclusão foram utilizados. Os critérios de exclusão visam registrar o motivo do trabalho ser descartado, enquanto o critério de inclusão visa registrar o motivo porque o trabalho foi aceito. Abaixo os critérios de inclusão:

1. O foco do trabalho é sobre pós-processamento de regras de associação obtida de padrões frequentes básicos.
2. O foco do trabalho é sobre outro tema, porém cria um pós-processamento de regras de associação que pode ser generalizado.
3. O trabalho propõe uma nova medida de interesse.
4. O trabalho faz uma comparação entre medidas de interesse ou as estuda.

A listagem abaixo mostra os critérios de exclusão utilizados:

1. O ano de publicação é anterior a 1993, quando o conceito de regras de associação foi publicado.
2. O trabalho não foi publicado em uma conferência ou periódico.
3. O documento não está acessível.
4. O idioma do documento não é inglês ou português.
5. O trabalho não usa regras de associação.
6. O trabalho usa regras de associação obtidas de padrões não básicos (por exemplo, regras de associação *fuzzy*).
7. Usa regras de associação para classificação (*class-association rules*).
8. O trabalho apenas aplica regras de associação.
9. O trabalho propõe um método de pré-processamento dos dados para *ARM*.
10. O trabalho propõe um novo método de *ARM*, ou seja, não é concentrado na fase de pós-processamento.
11. O método de pós-processamento proposto apenas seleciona as *m* melhores regras de associação, baseados somente em algumas dessas medidas: suporte, confiança, *lift*, ou baseado no conhecimento do especialista no domínio.

12. O pós-processamento proposto/utilizado não visa o problema do alto número de regras de associação.
13. A leitura do resumo foi inconclusiva.
14. O escopo do artigo não é científico.
15. O método utilizado não foi detalhado.
16. O trabalho é sobre o tema, mas foi publicado antes de 2015.

3.1.5 Formulário de avaliação de qualidade

Para avaliar a qualidade do artigos e criar uma ordem de leitura, foram formuladas perguntas com o objetivo de verificar os trabalhos mais alinhados com o tema. As perguntas, as respostas possíveis e seus respectivos valores (em parênteses) estão listadas abaixo.

- (PQ1) *O foco do método proposto é especificamente em pós-processamento de regras de associação?*
  - Sim (2,00)
  - Não, mas cria um método de pós-processamento de regras de associação (1,25)
  - Não (0,50)
- (PQ2) *Onde as palavras-chave aparecem juntas?*
  - Título, resumo e palavras-chave (2,00)
  - Título e resumo (1,75)
  - Título e palavras-chave (1,75)
  - Título (1,50)
  - Resumo e palavras-chave (1,50)
  - Resumo (1,00)
  - Palavras-chave (0,50)
  - Não aparecem juntas em um mesmo metadado (0,25)
- (PQ3) *O resumo do artigo especifica se alguma método de avaliação foi conduzido?*
  - Sim (2,00)
  - Foco na comparação de métodos ou medidas já existentes (1,00)
  - O resumo não detalha se alguma avaliação foi conduzida (0,5)
- (PQ4) *Quais tipos de avaliação foram realizadas?*
  - Comparação entre métodos (2,00)
  - Estudo de caso (1,50)
  - Avaliação pelo especialista (1,00)

- Tipo não especificado (0,75)
- O resumo não detalha se foi realizada (0,5)

Como é possível notar a pergunta (PQ4) está relacionada com a pergunta (PQ3), se a opção marcada para (PQ3) for a última então a opção marcada de (PQ4) também será.

3.2 Condução da revisão

Seguindo a diretriz de Kitchenham e colaborador [38], o próximo estágio é a condução da revisão. Nesta seção será detalhado como a revisão foi realizada. Primeiro, foram executadas as *strings* em suas respectivas bases. A última execução das *strings* em busca de atualizações foi no dia 06 de agosto de 2023. Algumas modificações foram necessárias por conta de especificidades das bases, as *strings* detalhadas com as explicações das adaptações estão disponíveis no Anexo B. Após isso o resultado foi extraído e os artigos foram selecionados por meio da leitura do resumo, por fim a leitura completa dos artigos e a extração dos dados foram feitas.

3.2.1 Seleção dos artigos

Tabela 3.1: *Quantidade de trabalhos obtidos, aceitos e rejeitados para leitura completa dos artigos*

Base	Obtidos	Aceitos	Rejeitados
ACM	261	15	246
IEEE	411	21	390
Scopus	862	68	794
Total	1534	104	1430

Após a execução das *strings* foram extraídas as referências bibliográficas do resultado retornado, essas referências foram importadas no *Parsifal*. Usando o *Parsifal* as duplicatas foram removidas e o processo de aceite ou rejeição com base na leitura dos resumos foi realizado. A Tabela 3.1 mostra a quantidade de artigos obtidos por cada *string*, o número de artigos que foram aceitos e número de artigos que foram rejeitados. Artigos duplicados foram considerados como rejeitados.

3.2.2 Leitura do artigo completo

Depois da fase de aceite e rejeição, os artigos foram importados para o *Mendeley* e foi iniciada a leitura do corpo dos artigos. Durante essa fase, 45 artigos foram considerados como rejeitados, isso aconteceu principalmente por três motivos, por resumos que

não especificavam a tipo de padrão obtido, desconhecimento de alguns padrões pelo autor e indisponibilidade do documento. A Tabela 3.2 mostra os números de artigos após a leitura do corpo dos documentos.

Tabela 3.2: *Quantidade de trabalhos obtidos, aceitos e rejeitados após leitura completa dos artigos*

Base	Obtidos	Aceitos	Rejeitados
ACM	261	6	255
IEEE	411	11	400
Scopus	862	42	820
Total	1534	59	1475

3.3 Resultados obtidos

Esta seção visa compilar as respostas para as questões definidas na Subseção 3.1.2.

(P1) *"Quais os métodos existentes para pós-processamento de regras de associação?"*

Para responder essa pergunta as Tabelas C.1 C.2 foram criadas e estão disponíveis no Apêndice C, nelas temos a referência do trabalho, o tipo de pós-processamento e se o trabalho é focado em um item de interesse. A poda é a abordagem mais comum sendo utilizada em 23 trabalhos, em seguida o ranqueamento que é utilizado em 19 estudos. A visualização é a terceira abordagem mais comum sendo utilizada em 18 trabalhos, seguida de *clustering* que é utilizado em 14 trabalhos. A sumarização é feita em apenas 6 estudos, mostrando uma necessidade de investigação desses métodos. Por fim, 2 estudos focam-se em estudar as medidas de avaliação. Além disso, 20 trabalhos ($33,9\%$) combinam duas ou mais abordagens.

O trabalho de Djenouri e colaboradores [23] tem como objetivo melhorar, em relação ao algoritmo sequencial, o tempo para obter meta-regras a partir de um conjunto de regras de associação. Uma meta-regra é uma regra de regras, ou seja, $mr : R_1 \rightarrow R_2$, onde R_1 e R_2 são conjuntos de regras. Para isso as regras de associação são transformadas em um base de dados transacional da seguinte maneira: as regras que compartilham pelo menos um item no consequente fazem parte de uma mesma transação (o consequente pode ter mais de um item); as regras são representadas como itens na nova base de dados. O algoritmo proposto (GSum-BSO) combina otimização por enxame de abelhas e arquitetura GPU para minerar as meta-regras. O paralelismo é aplicado durante a fase de avaliação das soluções. Após as meta-regras geradas uma poda é aplicada da seguinte

maneira: considere a meta-regra $mr : R_1 \rightarrow R_2$, apenas meta-regras com confiança 1 são selecionadas, e além disso, apenas o consequente da meta-regra é retornado. Os resultados mostraram que o método conseguiu ser mais rápido do que utilizar uma abordagem sequencial. Além disso, gerou uma solução similar a essa abordagem.

Berka propõe em [8] utilizar meta-regras para sumarizar regras de associação, visando diminuir o número de regras para serem analisadas pelo especilista em um domínio. O trabalho utiliza o processo *ASSOC* do método *GUHA* para minerar regras de associações, *GUHA* é um método de análise exploratória de dados. O *GUHA* é capaz de produzir regras em que um item presente em uma regra possa ter mais de uma valor separado por uma conjunção, $a_1 = (v_1 \vee v_2)$, ou ser acompanhado de uma negação. Além disso, cada regra está associada a uma condição que deve ser satisfeita. A configuração usada no trabalho, não especifica nenhuma condição, não permite negação e permite apenas um valor para cada item em uma regra, que faz com que o *ASSOC* gere resultados análogos a regras de associação tradicionais. A forma como o método transforma as regras mineradas para uma base de dados é diferente da feita por Djenouri e colaboradores [23], Berka mapeia cada regra como uma linha e as colunas são todos os itens da base, ou seja, as colunas permanecem iguais a base original. O foco do trabalho é na busca de descrições de conceitos, portanto o consequente é fixado no conceito avaliado. Para cada conceito uma nova base é gerada a partir das regras de associação obtidas, a nova base é submetida ao *ASSOC* novamente ignorando a coluna do conceito e então as meta-regras são geradas. Os resultados apresentados mostram uma redução de pelo menos 50% na quantidade de meta-regras em comparação as regras para a maioria das bases. Além disso, um estudo de caso mostra que as meta-regras obtidas podem demonstrar conhecimentos relevantes e serem úteis ao especialista.

O estudo feito por Grabot em [27], utiliza visualizações e um diagrama *UML* para auxiliar a a busca de regras por um especialista no domínio de manutenção industrial. O método primeiro utiliza um gráfico de colunas para entender o espaço das regras como um todo e ver uma tendência geral com relação ao suporte, confiança e *lift*. Depois utiliza duas variações de *TwoKey plot* para encontrar grupos de regras promissoras com base nas medidas relatas anteriormente. O diagrama *UML* é utilizado para mapear relações entre itens da base de dados que são esperadas, ou seja, as expectativas. A partir dessas visualizações o especialista no domínio pode buscar no arquivo da tabela as regras encontradas na visualização e, junto com as informações mapeadas no *UML*, utilizar filtros para encontrar regras de seu interesse. O que mais aproxima da proposta desse trabalho é que regras simétricas foram encontradas e tiveram um destaque especial, regras simétricas são regras do tipo $r_1 : a_1 \rightarrow a_2$ e $r_2 : a_2 \rightarrow a_1$, ou seja, um conhecimento baseado na relação entre regras. O interesse nessas regras ocorreu principalmente quando era esperado que dois itens fossem intimamente relacionadas mas não foram mineradas

regras simétricas. O trabalho mostra apenas a aplicabilidade do método no contexto de manutenção, portando não é possível afirmar muito sobre sua performance.

(P2) *"Existem métodos que se concentram em um item de interesse específico?"*

A grande maioria dos trabalhos não estão focados em um item de interesse, apenas quatro trabalhos tem esse foco, sendo que três desses [8] [12] [31] utilizam o item de interesse no consequente e apenas um [80] utiliza o item de interesse no antecedente.

O trabalho de Berka [8] foca-se em descrições de conceitos portanto fixa o consequente no conceito escolhido e gera meta-regras conforme descrito anteriormente (**P1**). Cheng e colaboradores [12] propõem uma visualização a partir de um conjunto de regras de associação com mesmo consequente, onde um especialista pode "montar"uma regra pela adição de itens. Essa estratégia permite que o especialista tente construir regras nas quais ele já possui uma suposição prévia. A medida que essa construção é feita ele valida a relevância das regras construídas com base no suporte e confiança disponibilizados pela visualização.

Outra visualização é proposta por Hahsler e Karpienko [31], onde *clusters* dos antecedentes são gerados em regras de mesmo consequente. Primeiro uma matriz é criada onde as linhas representam o consequente e as colunas os antecedentes (ambos *itemsets*). Depois as colunas são agrupadas por meio do algoritmo *k-means* verificando o *lift*. Os autores argumentam que usar o *lift* permite verificar condições de sinônimos ou itens parecidos, por exemplo, manteiga e margarina. A visualização consiste em um *balloon plot*: onde linhas são os consequentes, colunas são os *clusters* de antecedentes, um ponto no gráfico representa uma regra. Os pontos correspondem a balões com propriedades como a cor, o tamanho e a posição do balão sendo utilizadas para destacar os *clusters*. Embora o método não foque em um item de interesse específico, como todos os consequentes estão dispostos nas linhas da visualização isso facilita uma análise de um item específico (no consequente).

Wei e Scott [80] combinam poda, sumarização e visualização para encontrar padrões em reações de efeitos adversos de vacinas nos Estados Unidos da América. No trabalho o item de interesse atua o único antecedente da regra e o consequente é um *itemset*.

(P3) *"Os métodos propostos geralmente são avaliados?"*

A Tabela 3.3 compila a forma de avaliação dos trabalhos selecionados. Não foram considerados os trabalhos que avaliam medidas de interesse, pois não são comparáveis aos trabalhos de pós-processamento. É possível ver que $78,2\%$ são avaliados de

Tabela 3.3: *Quantidade de artigos por método de avaliação*

Método de avaliação	Número de artigos	Porcentagem
Estudo de caso	29	50,9%
Não avalia	*13*	*22,8%*
Comparação	8	14%
Estudo de caso, Comparação	4	7%
Estudo de usuário	2	3,5%
Estudo de caso, Comparação, Estudo de usuário	1	1,8%
Total	**57**	**100** %

alguma maneira [1], portando pode-se concluir que em geral os trabalhos são avaliados, mas que uma quantidade considerável ($22,8\%$) não realizam qualquer avaliação

(P4) *"Se eles são avaliados, como são avaliados?"*

Pela Tabela 3.3 apresentada anteriormente, a maioria ($59,7\%$ [2]) dos métodos avaliam utilizando um estudo de caso. Apenas $22,8\%$ dos trabalhos realizam comparações entre métodos de pós-processamento. Não foram considerados como comparação métodos que comparam os resultados apenas com um método de *ARM*. Estudos de usuário são feitos por $5,3\%$ dos trabalhos e todos propõem uma forma de visualização. Nesse caso os usuários utilizam a visualização durante um período de tempo realizando tarefas pré-determinadas.

(P5) *"Se eles são avaliados, as avaliações podem ser consideradas suficientes?"*

O melhor método para avaliação é a comparação, porém nem sempre é possível realizar essa abordagem. Portanto, realizar um estudo de caso ou mesmo um estudo de usuário são maneiras suficientes para avaliação. Sendo assim, para responder a pergunta (P4) é necessário verificar os aspectos que foram analisados para a avaliação dos métodos.

A Tabela 3.4 mostra quais propriedades foram usadas para avaliar os 44 trabalhos que passaram por um método de avaliação. A lista abaixo apresenta uma explicação de cada forma que foi considerada nesse trabalho.

- *Avaliação do especialista*: o especialista avalia as regras. Isso pode ser feito por meio de uma discussão ou o especialista rótula as regras como valiosas. Foram considerados como avaliação do especialista trabalhos que reportam a atividade do especialista ou trabalhos onde os próprios autores eram especialistas no domínio.

[1] $22,8\%$ dos trabalhos não realizaram alguma avaliação, como destacado na linha em itálico da Tabela 3.3.

[2] 7% dos métodos fazem um estudo de caso e comparação, $1,8\%$ fazem um estudo de caso, comparação e estudo de usuário, portando $50,9\% + 7\% + 1,8\%$.

Tabela 3.4: *Quantidade de artigos por forma de avaliação*

Forma de avaliação	Trabalhos	%
Quantitativa	9	20,5%
Discussão	4	9,1%
Avaliação do especialista, Qualidade estatística	4	9,1%
Insuficiente	3	6,8%
Conformidade, Avaliação do especialista, Quantitativa	3	6,8%
Quantitativa, Qualidade estatística	3	6,8%
Qualidade estatística	3	6,8%
Avaliação do especialista	2	4,5%
Avaliação do especialista, Quantitativa	2	4,5%
Tempo de execução	1	2,3%
Quantitativa, Específico do domínio	1	2,3%
Opinião do usuário	1	2,3%
Discussão, Quantitativa	1	2,3%
Específico do domínio, Qualidade estatística	1	2,3%
Discussão, Tempo de execução	1	2,3%
Discussão, Qualidade estatística	1	2,3%
Conformidade, Tempo gasto, Opinião do usuário	1	2,3%
Conformidade, Tempo gasto	1	2,3%
Conformidade, Discussão	1	2,3%
Tempo de execução, Quantitativa, Qualidade estatística	1	2,3%
Total	**44**	**100%**

- *Conformidade*: avalia se o método conseguiu obter regras de um conjunto verdade. Serve para verificar se o método consegue obter um conjunto esperado de regras. Por exemplo, regras que um especialista considera relevantes ou se as regras estão nos *clusters* esperados.
- *Insuficiente*: métodos que simplesmente mostram os resultados sem nenhum discussão ou que fazem afirmações sem alguma referência ou dado que confirme.
- *Discussão*: apresenta as regras obtidas e discute o conhecimento que elas transmitem.
- *Opinião do usuário*: a percepção do usuário final sobre o método de pós-processamento.
- *Qualidade estatística*: verifica a capacidade do método obter regras com altos valores de alguma(s) medida(s) objetiva(s).
- *Quantitativa*: número de regras retornadas ao usuário final. Uma forma de medir a diminuição do esforço do usuário.
- *Tempo de execução*: tempo requerido pelo algoritmo de pós-processamento.
- *Tempo gasto*: tempo gasto pelo usuário para realizar alguma tarefa.

Dos 44 trabalhos 3 avaliações ($6,8\%$) foram consideradas insuficientes. O res-

tante (93,2%) verificaram de forma válida algum aspecto da solução. A maioria dos métodos (45,5%) avaliam as regras de forma quantitativa. Essa forma de avaliação mostra diminuição do esforço do usuário para encontrar regras, pois demonstram a redução do conjunto final de regras a serem avaliadas. Como mostra a Tabela 3.4 a avaliação apenas quantitativa é a mais comum entre os trabalhos (20,5%).

Embora utilizar apenas avaliação quantitativa possa ser considerado satisfatório, isso não permite afirmar que o método produz um conjunto de regras relevantes, portanto em 25% dos estudos a avaliação quantitativa é feita junto com formas de avaliação qualitativas. Dentre esses, destacam-se 3 estudos (6,8%): 2 de Zhang e colaboradores [84] [85] e 1 de Carvalho e colaboradores [19]. Esses estudos avaliaram a conformidade das regras em relação a avaliação prévia do especialista, nesses trabalhos um especialista em um domínio classificou todas as regras geradas pelo algoritmo de *ARM* como valiosas ou não, depois o método de pós-processamento foi aplicado e esses conjuntos foram comparados. Nos trabalhos de Zhang e colaboradores, são verificados tanto a capacidade de obter as regras valiosas corretamente quanto a taxa de erro. Já no trabalho do Carvalho e colaboradores, é verificada a redução do espaço de busca até encontrar uma determinada porcentagem de regras úteis. Ambas abordagens podem ser consideradas a melhor forma de se avaliar um método, pois permite obter métricas claras sobre o conjunto inteiro de regras pós-processadas, tanto quantitativamente quanto qualitativamente.

Gerar um conjunto verdade de regras observando a utilidade demanda um esforço muito grande, portanto, uma alternativa é avaliar, em conjunto com a avaliação quantitativa, a utilidade de apenas algumas regras obtidas. Esse tipo de avaliação mostra que o método consegue obter regras relevantes, porém não fornece uma métrica clara sobre essa capacidade. Dentre os trabalhos analisados 3 métodos (6,8%) utilizam essa estratégia, sendo que 2 desses (4,5%) fazem uma avaliação pelo especialista e o trabalho restante (2,5%) faz uma discussão dos resultados apresentados. Essa forma de avaliação pode ser considerada uma boa estratégia (dentre as possíveis), pois permite ter pelo menos um indício de que regras relevantes podem ser encontradas. Além disso, é preferível que um especialista analise esses resultados.

Outra forma de analisar a qualidade das regras é verificando se possuem bons valores de medidas objetivas, 4 trabalhos (9,1%) realizam esse tipo de avaliação juntamente com a avaliação quantitativa. Esse tipo de estratégia deve escolher cuidadosamente medidas que consigam demonstrar o objetivo do usuário final, por exemplo, *lift* pode ser usado para encontrar sinônimos [31]. Essa pode ser considerada uma boa forma de avaliar, pois pode fornecer bons indícios com base na medida objetiva escolhida.

Quando a avaliação quantitativa não é feita, pelo menos uma avaliação da qualidade do conjunto de regras obtidas ou de parte dele deve ser feita, 38,7% dos métodos são avaliados dessa forma e pode ser considerada satisfatória, sendo preferível

a avaliação do especialista, depois a validação da qualidade estatística e por fim somente uma discussão das regras selecionadas.

Embora os 9% restantes não avaliem nem quantitativamente nem qualitativamente as regras obtidas, nos seus respectivos contextos tais formas de avaliação são boas práticas. Em abordagens que usam visualizações avaliar o tempo gasto e a opinião do usuário pode ser uma boa forma de medir a qualidade. O tempo execução geralmente não é usado, porque o principal problema é a quantidade impeditiva de regras, mas em estudos específicos como em [23] podem ser relevantes. Em dois estudos a avaliação foram feita com algumas informações específicas do domínio aplicado.

Portando, respondendo a pergunta (P4) como mostra a Tabela 3.4 $93,2\%$ dos métodos avaliam de uma maneira suficiente. Porém poucos trabalhos conseguem realizar avaliações com a ajuda de especialistas no domínio o que dificulta a obtenção de resultados mais robustos.

(P6) *"É possível definir quais são os melhores métodos?"*

Uma ideia inicial para comparar esses métodos seria definir um conjunto de regras consideradas boas por alguns trabalhos [19] [84] [85] e verificar a quantidade final de regras, a taxa de erro, a taxa de acerto. Porém, usando como exemplo a proposta de [12], em que o usuário explora o resultado "criando" as regras em vez de ir verificando uma por uma, essa verificação não seria útil. Além disso, como mostrado em [18] e [34] é possível utilizar essas técnicas em conjunto, pois podem possuir objetivos diferentes. Em outros casos, o método propõe apenas uma ranqueamento portanto o que deve ser avaliado é se as regras mais relevantes aparecem primeiro.

Dessa forma podemos concluir que não há uma definição única de melhor método e que essa definição depende mais do objetivo do que os métodos serem necessariamente do mesmo paradigma. Por exemplo, suponha dois trabalhos que tenham como objetivo obter as melhores regras, um trabalho usa *clustering* para agrupar regras e retorna o melhor *cluster* para o usuário, o outro trabalho raqueia as k melhores regras conforme uma nova medida de interesse e retorna as k melhores para usuário, esses trabalhos embora utilizem paradigmas diferentes são passíveis de comparação. Se o método a ser comparado com o primeiro utilizasse *clustering* e tivesse por objetivo agrupar as regras similares com máximo de assertividade a comparação não teria sentido, mesmo utilizando a mesma técnica. Note também que em ambos os casos a tarefa realizada é pós-processamento de regras, mas com finalidades distintas.

(P7) *"Quais as principais lacunas de pesquisas existentes na área?"*

A lacuna mais perceptível é quantidade pequena de artigos que avaliam usando comparações. Dois fatores podem estar relacionados a isso. Primeiro poucos trabalhos

disponibilizam um implementação de sua proposta em código aberto. Um outro ponto é ausência de um *benchmark* de regras, ou seja, não há um conjunto verdade disponível para que métodos possam ser comparados. Sobre este último ponto, seria relevante existir uma base de regras geradas a partir de uma algoritmo de *ARM*, em que as regras ou um subconjunto delas fossem classificadas como valiosas ou não por um ou mais especialistas, similar ao feito em [19] [84] [85], porém aberto e colaborativo. Dessa forma, facilitaria o desenvolvimento de artigos de comparação e a avaliação de novos métodos, pois o processo consistiria em aplicar o novo método sobre esse conjunto de regras e comparar com os resultados já existentes. Além disso, algumas métricas claras poderiam ser usadas como a taxa de erro, a taxa de acerto, a capacidade de descobrir novas regras (caso todas as regras não tenham sido rotuladas). Isso permitiria entender quais são os melhores métodos entre os passíveis de comparação, que é difícil perceber apenas com os estudos existentes na literatura.

Dentre os métodos analisados a maioria realiza um pós-processamento de maneira mais geral, poucos métodos se concentram em estratégias para buscar regras baseado em um item de interesse. Os métodos que fazem essa busca orientada geralmente focam em procurar o item de interesse no consequente, devido a facilidade interpretação [8] [12] [31]. Porém, uma das características de regras de associação é que um item pode aparecer em qualquer lado da regra o que garante ao método uma natureza exploratória, sendo assim, falta estudos em mostrar quais tipos de conhecimento um item de interesse no antecedente pode revelar, tais como, efeitos adversos de vacinas como apresentados por Wei e Scott em [80].

Outro ponto, conforme relatado na pergunta (P1), é que pouquíssimos trabalhos tentam pós-processar regras verificando o relacionamento entre elas. Porém, nesses poucos trabalhos podemos ver que o relacionamento entre regras podem ajudar a reforçar algumas relações [27], bem como trazer conhecimentos novos sobre os itens de uma determinada base [8]. Portanto, estudos com essa abordagem podem encontrar resultados promissores, principalmente aplicado junto a um item de interesse que auxilia no problema do tamanho do espaço de busca.

3.4 Conclusão

Como visto na seção anterior, há um número grande de trabalhos nos últimos 7 anos com as mais variadas abordagens. Vários apresentam resultados promissores, porém não há nenhum estudo tentando definir quais os melhores métodos existentes. Além disso, as avaliações feitas por esses métodos embora sejam satisfatórias são limitadas por alguns desafios, tais como a falta de implementações abertas e *benchmarks* de regras.

Foram encontrados poucos estudos que verificam relacionamento entre regras para gerar conhecimento, embora possam trazer conhecimentos complementares aos produzidos por regras de associação. Outro campo ainda pouco estudado é a seleção de regras com foco em um item de interesse, principalmente com esse item de interesse podendo existir no consequente ou antecedente. Também foram encontrados poucos estudos utilizando abordagens de sumarização.

Um trabalho futuro bastante relevante para área é a criação de um *benchmark* de regras. Uma estratégia para essa criação seria verificar os resultados obtidos em estudos que usaram uma determinada base aberta e rotular essas regras como boas ou ruins (feito por especialistas no domínio). A disponibilização dos métodos em códigos abertos, permitiria maior agilidade para realizar estudos comparações e consequentemente um entendimento melhor sobre as abordagens existentes, portanto seria outro trabalho relevante quanto a aplicabilidade e avaliação dos métodos existentes. Não foram encontrados métodos que façam uma busca por item de interesse e que preservem a natureza exploratória dos algoritmos de *ARM*. Um método desse tipo é relevante para investigações onde existe um interesse bem definido, porém uma investigação exploratória também se faz necessária. Além disso, verificar o relacionamento entre regras orientado por um item de interesse pode mostrar como esse item influência e é influenciado por outras associações, portanto, pode ser uma forma de destacar aspectos relevantes de interesse do especialista.

CAPÍTULO 4

Agrupamento de regras de associação orientado a item de interesse

Neste capítulo será apresentada uma nova abordagem para pós-processamento de regras de associação. A proposta consiste em obter grupos de regras relevantes para um especialista no domínio de aplicação a partir do relacionamento entre regras de associação. Seu principal foco é destacar relacionamentos entre regras de associação que envolvam um item de interesse de estudo a fim de ajudar o especialista no domínio a encontrar regras que atendam aos seus objetivos de estudo. Além disso, diferente dos trabalhos encontrados apresentados no Capítulo 3, o método proposto não fixa o item de interesse em um lado da regra. Desta forma, procura-se utilizar por completo a natureza exploratória do método de ARM.

A abordagem proposta surgiu da observação da busca manual feita por um especialista em Castro e colaboradores [10]. Para o especialista não bastava somente visualizar uma regra como:

$$a_1, a_2 \rightarrow a_3$$

Era necessário saber se os itens do antecedente estavam individualmente relacionados ao consequente, ou seja, se $a_1 \rightarrow a_3$ e $a_2 \rightarrow a_3$ também haviam sido geradas pelo algoritmo de ARM. A partir disso, um método foi proposto no TCC do autor [14], porém, tal método apresentava algumas limitações que dificultavam a análise do especialista. Essas limitações foram sanadas e o método foi estendido na abordagem atual.

Primeiro será apresentado na Seção 4.1 um algoritmo em uma linguagem mais descritiva a fim de definir os conceitos dos relacionamentos entre regras e uma ideia mais geral do processo. Depois, na Seção 4.2 será definido um algoritmo com maiores detalhes de implementação. A Seção 4.3 detalha a representação gráfica proposta no trabalho.

4.1 Descrição do método

Nesta seção, será apresentada a abordagem proposta neste trabalho. Seu principal foco é selecionar subconjuntos de regras de um conjunto de regras de associação geradas por um algoritmo de ARM que utilize o modelo suporte e confiança, concentrando-se em problemas em que haja um item de interesse de estudo, e utilizando uma medida M como medida de importância das regras. Foram consideradas apenas regras cujo consequente possui somente um item. Além disso, considera-se apenas regras com o tamanho máximo igual 3, pois regras maiores tornam mais difícil a sua interpretação e aplicabilidade. O tamanho mínimo das regras foi definido como 2, pois o cálculo da confiança depende da existência de um antecedente.

Algumas implementações de algoritmos de ARM mapeiam uma base de dados tabular de forma que os itens sejam representados pelo par $ATRIBUTO = valor\_atributo$, onde os atributos são as colunas da base de dados. Sendo assim, o método proposto remove regras que possuam itens em que o atributo possua valor faltante, ou seja, $ATRIBUTO =$.

É importante destacar que as regras de associação que o algoritmo recebe como entrada já são regras significativas em termos de suporte, ou seja, os itens ocorrem com uma boa frequência (frequência mínima definida pelo usuário para a geração das regras), e confiança, ou seja, sempre que um antecedente ocorre o consequente costuma ocorrer (acima de um valor mínimo definido pelo usuário antes da geração das regras). Segue abaixo uma descrição em alto nível da abordagem proposta.

Entrada:

- um conjunto R de regras de tamanhos 2 ou 3 geradas por um algoritmo de ARM baseado em suporte e confiança,
- um item de interesse $\bar{a}$, que será usado como base para a seleção das regras,
- uma medida de interesse M capaz de medir a dependência entre antecedente e consequente, por exemplo, *lift*, *conviction* e *odds ratio*,
- um valor de limiar δ (podendo ser zero) representando uma margem de dependência entre o antecedente e consequente da regra baseado na medida M (valor padrão 0),
- um valor α de melhoria mínima da medida M (valor padrão 0),
- um valor c' de confiança mínima para regras de tamanho 3 (valor padrão 0).

Saída:

- conjuntos de agrupamentos de regras categorizados em 8 diferentes tipos.

Procedimento:

1. *Seleção das regras relevantes.* O método seleciona do conjunto R as regras cujo valor M seja maior que $1,0+\delta$, fazendo $R' = \{r_j | r_j \in R \wedge M(r_j) > 1,0+\delta\}$
2. *Seleção das regras de interesse.* O método seleciona do conjunto R' apenas as regras que possuem o item de interesse $\bar{a}$, criando o conjunto $R'' = \{r_j | r_j \in R \wedge \bar{a} \subset r_j\}$
3. *Filtro de confiança mínima.* O método retira do conjunto R'' as regras de tamanho 3 com confiança menor que c', criando o conjunto $R''' = \{r_j | r_j \in R'' \wedge (len(r_j) == 2 \vee Conf(r_j) >= c')\}$
4. *Agrupamento das regras.* O método organiza as regras segundo os tipos a seguir. Vale destacar que, neste ponto, o conjunto de regras R''' consiste apenas de regras que possuem o item de interesse e são relevantes [1].

   (a) Tipo 1: conjuntos de pares de regras de tamanho 2 do tipo:

   $$\{r_1 : a \rightarrow \bar{a}; r_2 : \bar{a} \rightarrow a\}, \text{ onde } r_1, r_2 \in R'''.$$

   (b) Tipo 2: conjuntos de trios de regras do tipo:

   $$\{r_1 : a_1 \rightarrow a_2; r_2 : \bar{a} \rightarrow a_2; r_3 : a_1, \bar{a} \rightarrow a_2\},$$

   onde $r_1 \in R'$, $r_2, r_3 \in R'''$ e $\frac{M(r_3)-M(r_1)}{M(r_1)} \geq \alpha$ e o $\frac{M(r_3)-M(r_2)}{M(r_2)} \geq \alpha$.

   (c) Tipo 3: conjuntos de pares de regras dos tipos:

   $$\{r_1 : a_1 \rightarrow a_2; r_2 : a_1, \bar{a} \rightarrow a_2\},$$

   onde $r_1 \in R'$, $r_2 \in R'''$, $\bar{a} \rightarrow a_2 \notin R'''$ e $\frac{M(r_2)-M(r_1)}{M(r_1)} \geq \alpha$.

   (d) Tipo 4: conjuntos de pares de regras dos tipos:

   $$\{r_1 : \bar{a} \rightarrow a_2; r_2 : a_1, \bar{a} \rightarrow a_2\},$$

   onde $r_1, r_2 \in R'''$, $a_1 \rightarrow a_2 \notin R'$ e $\frac{M(r_2)-M(r_1)}{M(r_1)} \geq \alpha$.

   (e) Tipo 5: conjuntos unitários de regras dos tipos:

   $$\{r : a_1, \bar{a} \rightarrow a_2\},$$

   onde $r \in R'''$ e tanto $\bar{a}$ quanto a_1 não aparecem como antecedentes em regras de tamanho 2 com o consequente a_2.

   (f) Tipo 6: conjuntos de trios de regras com o item de interesse $\bar{a}$ no consequente do tipo:

   $$\{r_1 : a_1 \rightarrow \bar{a}; r_2 : a_2 \rightarrow \bar{a}; r_3 : a_1, a_2 \rightarrow \bar{a}\},$$

   onde $r_1, r_2, r_3 \in R'''$, e a_1 e a_2 aparecem como antecedentes em regras de tamanho 2 em R''' e $\frac{M(r_3)-M(r_1)}{M(r_1)} \geq \alpha$ e o $\frac{M(r_3)-M(r_2)}{M(r_2)} \geq \alpha$.

[1] A medida de interesse M possui um valor que indique dependência entre o antecedente e consequente; as regras de tamanho 3 possuam pelo menos a confiança mínima c'.

(g) Tipo 7: conjuntos de pares de regras com o item de interesse $\bar{a}$ no consequente, do tipo

$$\{r_1 : a_1 \rightarrow \bar{a}; r_2 : a_1, a_2 \rightarrow \bar{a}\},$$

onde $r_1, r_2 \in R'''$, e a_1 (mas não a_2) aparece como antecedente em regras de tamanho 2 em R''' e $\frac{M(r_2)-M(r_1)}{M(r_1)} \geq \alpha$.

(h) Tipo 8: conjuntos unitários de regras com o item de interesse $\bar{a}$ no consequente do tipo:

$$\{r : a_1, a_2 \rightarrow \bar{a}\},$$

onde $r \in R'''$, e tanto a_1 quanto a_2 não aparecem como antecedentes em regras de tamanho 2 em R'''.

Conforme apresentado no algoritmo acima, a seleção de regras proposta se baseia em relacionamentos entre regras, onde cada relacionamento é capaz de expressar um tipo de informação sobre o item de interesse. Esses relacionamentos se baseiam na presença ou ausência de regras, portanto antes de analisar os tipos, é necessário compreender o que significa a presença e a ausência de regras na saída produzida por um algoritmo de ARM.

A ausência de uma regra na saída de um método de ARM baseado em suporte/confiança se dá por dois fatores: não alcançar o suporte mínimo ou não alcançar a confiança mínima. Porém, se existe uma regra de tamanho 3 (r_3), isso implica que qualquer regra de tamanho 2 (r_1 e r_2) contida em r_3 seja frequente (alcança o suporte mínimo). Suponha que exista a regra $\{r_3 : a_1, a_2 \rightarrow a_3\}$ gerada a partir do *itemset* $I_3 = \{a_1, a_2, a_3\}$. O *itemset* I_3 é gerado a partir da combinação dos *itemsets*, $I_1 = \{a_1, a_3\}$ e $I_2 = \{a_2, a_3\}$, caso I_1 ou I_2 não fossem frequentes eles seriam descartados e não poderiam gerar I_3, portanto se I_3 existe, I_1 e I_2 precisam ser frequentes. Por consequência, as regras $\{r_1 : a_1 \rightarrow a_3\}$ gerada a partir de I_1 e $\{r_2 : a_2 \rightarrow a_3\}$ gerada a partir de I_2 são frequentes. Isso ocorre porque os algoritmos de ARM baseados em suporte/confiança usam a propriedade de que apenas um *itemset* frequente pode gerar outro *itemset* frequente. Dada tal propriedade, se existe uma regra de tamanho 3 (r_3) e não existe uma regra de tamanho 2 (r_1 ou r_2) contida em r_3, a única maneira da regra de tamanho 2 (r_1 ou r_2) ter sido descartada é por não ser confiável (não alcança a confiança mínima), pois pela propriedade mostrada anteriormente elas são frequentes (alcançam o suporte mínimo).

Além disso, o algoritmo apresentado também elimina regras que são consideradas irrelevantes para o propósito deste trabalho, ou seja, regras cujo o valor da medida M seja igual a 1 ou, caso o usuário especifique, valores próximos disso, indicando que o antecedente e consequente da regra são independentes. Portanto, durante a geração dos agrupamentos, se existir uma regra de tamanho 3, então as regras de tamanho 2 relacionadas não existirão no conjunto, ou porque não são confiáveis ou porque seus itens são independentes.

O parâmetro c do método não está diretamente relacionado a confiança mínima dos algoritmos de ARM. Em alguns experimentos notou-se que poderia deixar mais evidente a relação entre os itens se fosse permitido valores um pouco mais baixos de confiança para as regras complementares (regras de tamanho 2). Portanto, o valor c' foi criado como um ajuste para não permitir que agrupamentos em torno de regras de tamanho 3 não confiáveis fossem selecionados. No entanto esse filtro é dependente do problema, e o parâmetro c' pode não ser informado, pois possui um valor padrão 0.

Segue abaixo uma discussão sobre cada tipo e o que tais relações podem transmitir a respeito de um item de interesse. Para análise abaixo, estendeu-se o conceito de regras de associação forte (regras com suporte e confiança maiores que os mínimos) [71] para regras que também respeitem a restrição do M maior que $1,00+\delta$. A medida M utilizado nessa discussão foi o *lift*, com margem de dependência $\delta = 0,1$. Nos exemplos dados, o item de interesse é "dengue=sim", ou seja, deseja-se saber quais fatores estão associados com a dengue e as relações onde a dengue é um agravante ou é agravada destas associações. As regras apresentadas não vieram de uma base de dados mas foram informadas por um médico para fins de ilustração dos Tipos definidos na proposta deste trabalho.

Motivação do Tipo 1

$$\{r_1 : a_1 \rightarrow \bar{a}\} \in R'''$$
$$\{r_2 : \bar{a} \rightarrow a_1\} \in R'''$$

O Tipo 1 trata de conjuntos das regras bidirecionais, ou seja, regras onde tanto um item qualquer implica fortemente no item de interesse quanto o item de interesse implica fortemente nesse item qualquer. Esse tipo de informação serve como um reforço em relação as associações, pois é um indicador de que os itens estão intimamente associados.

Exemplo 4.1.1 *Exemplo meramente ilustrativo:*

$$\textit{febre=[>7dias]} \rightarrow \textbf{\textit{dengue=sim}}$$
$$\textbf{\textit{dengue=sim}} \rightarrow \textit{febre=[>7dias]}$$

Motivação do Tipo 2

$$\{r_1 : a_1 \rightarrow a_2,\ lift = 1,50\} \in R'$$
$$\{r_2 : \bar{a} \rightarrow a_2,\ lift = 1,60\} \in R'''$$
$$\{r_3 : a_1, \bar{a} \rightarrow a_2,\ lift = 3,00\} \in R'''$$

Esse conjunto de regras expressa quando o item de interesse e um outro item qualquer estão isoladamente associados ao mesmo consequente e quando aparecem juntos eles têm essa associação reforçada. Nas regras acima, tanto a_1 quanto $\bar{a}$ estão associados a a_2 isoladamente (regras r_1 e r_2). Além disso, a regra r_3 demonstra que quando a_1 e $\bar{a}$ aparecem juntos o *lift* é melhorado, ou seja, os itens se pontencializam em relação ao consequente a_2. Nem sempre o fato de existirem regras como r_1 e r_2 leva a existência de uma regra como r_3.

Exemplo 4.1.2 *Exemplo meramente ilustrativo:*

$$\textit{hemofilia=sim} \rightarrow \textit{hemorragia=sim},\ lift = 1{,}50^2$$
$$\textbf{\textit{dengue=sim}} \rightarrow \textit{hemorragia=sim},\ lift = 1{,}60$$
$$\textit{hemofilia=sim}, \textbf{\textit{dengue=sim}} \rightarrow \textit{hemorragia=sim},\ lift = 3{,}00$$

Motivação do Tipo 3

$$\{r_1 : a_1 \rightarrow a_2,\ lift = 1{,}50\} \in R'$$
$$\{r_2 : a_1, \bar{a} \rightarrow a_2,\ lift = 3{,}00\} \in R'''$$

e uma regra com o corpo $\bar{a} \rightarrow a_2$ não exista após as remoções aplicadas, ou seja, $\{r_3 : \bar{a} \rightarrow a_2\} \notin R'''$.

O relacionamento entre regras expresso pelo Tipo 3 mostra quando o item de interesse contribui para que uma relação já existente se torne mais forte, sem que o item de interesse já estivesse relacionado ao consequente. Levando em conta as regras acima, o item a_1 está fortemente associado a a_2 individualmente (r_1) e quando a_1 aparece junto de $\bar{a}$ essa associação se fortalece (r_2). Além disso, nessa configuração r_3 não pode existir, portanto $\bar{a}$ não aparece individualmente em associação com a_2. Sendo assim, $\bar{a}$ não tem associação forte com a_2, mas potencializa a associação entre a_1 e a_2.

Exemplo 4.1.3 *Exemplo meramente ilustrativo:*

$$\textit{dorAbdominal=intensa} \rightarrow \textit{uti=sim},\ lift = 1{,}50$$
$$\textit{dorAbdominal=intensa}, \textbf{\textit{dengue=sim}} \rightarrow \textit{uti=sim},\ lift = 3{,}00$$

[2]Hemofilia: distúrbio em que o sangue não coagula normalmente.

Motivação do Tipo 4

$$\{r_1 : \bar{a} \rightarrow a_2,\ lift = 1{,}50\} \in R'''$$
$$\{r_2 : a_1, \bar{a} \rightarrow a_2,\ lift = 3{,}00\} \in R'''$$

e uma regra com o corpo $a_1 \rightarrow a_2$ não exista após a remoção de regras irrelevantes, ou seja, $\{r_3 : a_1 \rightarrow a_2\} \notin R'$.

Esse tipo de regras pode expressar quando um item qualquer do banco, sem possuir uma associação forte individual com o consequente, é capaz de potencializar uma associação em que o item de interesse aparece no antecedente. Nas regras acima, a_1 não está fortemente associado a a_2, pois r_3 não existe. Quando a_1 aparece junto do item de interesse $\bar{a}$ a associação de r_2 demonstra ser mais forte que de r_1. Portanto a_1 potencializou a associação presente em r_1.

Exemplo 4.1.4 *Exemplo meramente ilustrativo:*

$$\textbf{\textit{dengue=sim}} \rightarrow \textit{hospitalização=sim},\ lift = 1{,}50$$
$$\textit{gestante=sim}, \textbf{\textit{dengue=sim}} \rightarrow \textit{hospitalização=sim},\ lift = 3{,}00$$

Motivação do Tipo 5

$$\{r_1 : a_1, \bar{a} \rightarrow a_2,\ lift = 3{,}00\} \in R'''$$

e uma regra com o corpo $\bar{a} \rightarrow a_2$ não exista após as remoções aplicadas, bem como, uma regra com o corpo $a_1 \rightarrow a_2$ não exista após a remoção de regras irrelevantes, portanto, $\{r_2 : \bar{a} \rightarrow a_2\} \notin R'''$ e $\{r_3 : a_1 \rightarrow a_2\} \notin R'$.

O Tipo 5 demonstra o caso em que o item de interesse está associado a um consequente apenas quando aparece junto de um outro item qualquer no antecedente, e este por sua vez também não está associado individualmente ao item do consequente. Para melhor visualização considere as regras acima, o item de interesse $\bar{a}$ não está associado ao item a_2 individualmente pois não existe r_2. Além disso, o item a_1 não está associado individualmente a a_2, pois não existe r_3. Como só existe a regra r_1 os itens $\bar{a}$ e a_1 estão associados a a_2 apenas quando aparecem juntos.

Exemplo 4.1.5 *Exemplo meramente ilustrativo:*

$$\textbf{\textit{dengue=sim}}, \textit{dorAbdominal=intensa} \rightarrow \textit{desconfortoRespiratório=sim},\ lift = 3{,}00$$

Motivação do Tipo 6

$$\{r_1 : a_1 \rightarrow \bar{a},\ lift = 1,50\} \in R'''$$
$$\{r_2 : a_2 \rightarrow \bar{a},\ lift = 1,60\} \in R'''$$
$$\{r_3 : a_1, a_2 \rightarrow \bar{a}, lift = 3,00\} \in R'''$$

O conjunto de regras do Tipo 6 mostra quando itens quaisquer que implicam no item de interesse separadamente também implicam no item de interesse quando aparecem juntos, porém de maneira mais forte em relação à medida de avaliação. Na configuração acima, r_1 e r_2 já demonstram a associação entre a_1 e a_2 e o item de interesse $\bar{a}$. A conjunção destes itens na regra r_3 reforçam esta association com o item de interesse.

Exemplo 4.1.6 *Exemplo meramente ilustrativo:*

$$\textit{febre=[>7dias]} \rightarrow \textbf{\textit{dengue=sim}},\ lift = 1,50$$
$$\textit{plaquetopenia=sim} \rightarrow \textbf{\textit{dengue=sim}},\ lift = 1,60^{3}$$
$$\textit{febre=[>7dias], plaquetopenia=sim} \rightarrow \textbf{\textit{dengue=sim}},\ lift = 3,00$$

Motivação do Tipo 7

$$\{r_1 : a_1 \rightarrow \bar{a},\ lift = 1,50\} \in R'''$$
$$\{r_2 : a_1, a_2 \rightarrow \bar{a}, lift = 3,00\} \in R'''$$

e uma regra com o corpo $a_2 \rightarrow \bar{a}$ não exista após as remoções aplicadas, ou seja, $\{r_3 : a_2 \rightarrow \bar{a}\} \notin R'''$.

O Tipo 7 reflete uma informação parecida com a do Tipo 6, porém, neste caso um dos itens não está fortemente associado ao item de interesse de forma individual. Nas regras acima, r_1 demonstra que a_1 está fortemente associado a $\bar{a}$, mas como r_3 não existe o mesmo não vale para a_2. No entanto, r_2 mostra que a_2 é capaz de potencializar a associação entre a_1 e $\bar{a}$ mesmo não estando individualmente relacionado ao item de interesse.

Exemplo 4.1.7 *Exemplo meramente ilustrativo:*

$$\textit{febre=[>7dias]} \rightarrow \textbf{\textit{dengue=sim}},\ lift = 1,50$$
$$\textit{sangramento=sim, febre=[>7dias]} \rightarrow \textbf{\textit{dengue=sim}},\ lift = 3,00$$

[3] Plaquetopenia: nível excepcionalmente baixo de plaquetas no sangue.

Motivação do Tipo 8

$$\{r_1 : a_1, a_2 \rightarrow \bar{a}, lift = 3,00\} \in R'$$

e as regras com os corpos $a_1 \rightarrow \bar{a}$ e $a_2 \rightarrow \bar{a}$ não existam após as remoções aplicadas, ou seja, $\{r_2 : a_1 \rightarrow \bar{a}\} \notin R'''$ e $\{r_3 : a_2 \rightarrow \bar{a}\} \notin R'''$.

O Tipo 8 expressa uma relação diferente do que os Tipos 6 e 7. O Tipo 8 mostra itens que isoladamente não estão fortemente associados ao item de interesse, mas que em conjunto demonstram ter uma relação forte. No caso acima, pode-se notar que não existem r_2 e r_3, mas que existe r_1 onde os itens a_1 e a_2 estão fortemente associados ao item de interesse.

Exemplo 4.1.8 *Exemplo meramente ilustrativo:*

$$petéquias{=}sim, dorAbdominal{=}intensa \rightarrow \textbf{\textit{dengue=sim}},\ lift = 3,00^{4}$$

Em síntese, os agrupamentos dos Tipos 2, 3, 4 e 5, tratam do item de interesse no antecedente da regra destacando como este item participa do reforço dos fatores que acarretam em um outro item. Por outro lado, os agrupamentos dos Tipos 6, 7 e 8 mostram o item de interesse no consequente da regra, ou seja, como as associações de outros itens poderiam reforçar ou potencializar a ocorrência do item de interesse.

A proposta atual estende mas também se difere da abordagem anteriormente proposta no Trabalho de Conclusão de Curso do autor [14] em três aspectos. Primeiro, a abordagem anterior dividia as regras de tamanho 3 em tipos baseado em algumas relações mostradas nos Tipos propostos acima, porém, as regras de tamanho 2 que formavam tais relações não eram atribuídas no mesmo tipo, mas em um tipo com todas as regras de tamanho 2 que possuíam o item de interesse, ou seja, as regras que compunham a relação ficavam separadas e até distantes umas das outras. Isso fazia com que as relações expressas pelos Tipos relatados acima não ficassem evidentes. Na nova abordagem as regras são organizadas em pequenos conjuntos visando corrigir esse problema. Outro fator é que anteriormente todas as regras que não tivessem o item de interesse eram removidas o que ocultava a existência dos Tipos 2 e 3. Por fim, havia-se definido uma hierarquia entre os tipos, porém, no presente estudo foi entendido que não existe tal hierarquia e dependendo do objetivo do estudo um tipo pode ser mais interessante que o outro.

[4] Petéquias: manchas avermelhadas de tamanho pequeno que podem aparecer na pele.

4.2 Implementação

Algoritmo 4.1: *obter-relacionamentos-inter-regras*$(R, \bar{a}, \delta, \alpha, c')$

Entrada:

- Um conjunto de regras R.
- Um item de interesse $\bar{a}$.
- Uma medida de interesse M.
- Um limiar δ para definir a irrelevância (valor padrão 0).
- Uma melhoria mínima α para a medida M (valor padrão 0).
- Um limiar c' de confiança mínima (valor padrão 0).

Saída: Uma lista (G) de agrupamentos regras de associação.

1 $G = \emptyset$
2 $R = \{r_j \mid r_j \in R \wedge M(r_j) > 1,0 + \delta\}$
3 $R_2 = \{r_j | r_j \in R \wedge r_j.tamanho = 2\}$
4 $R_3 = \{r_j | r_j \in R \wedge r_j.tamanho = 3 \wedge Conf(r_j) \geq c' \wedge \bar{a} \subset r_j\}$
5 *hash* = *transforme-em-hash*(R_2)
6 **para todo** $r \in R_3$ **faça**
7 $\quad r_{ant1} = hash[r.antecedente1 + r.consequente]$
8 $\quad r_{ant2} = hash[r.antecedente2 + r.consequente]$
9 $\quad$ **se** *melhoria-percentual*$(r, r_{ant1}, M) \geq \alpha \wedge$ *melhoria-percentual*$(r, r_{ant2}, M) \geq \alpha$ **então**
10 $\quad\quad g =$ *decida-tipo*(r, r_{ant1}, r_{ant2})
11 $\quad\quad G.adicione(g)$
12 $\quad$ **fim**
13 **fim**
14 **para todo** $r \in hash.valores$ **faça**
15 $\quad$ **se** $\bar{a} \subset r$ **então**
16 $\quad\quad r_{inversa} = hash[r.consequente + r.antecedente]$
17 $\quad\quad$ **se** $r_{inversa} \neq nulo$ **então**
18 $\quad\quad\quad g =$ *cria-tipo-bidirecional*$(r, r_{inversa})$
19 $\quad\quad\quad G.adicione(g)$
20 $\quad\quad\quad hash.remova(r.consequente + r.antecedente)$
21 $\quad\quad$ **fim**
22 $\quad$ **fim**
23 **fim**

Nesta seção é apresentado um algoritmo da implementação feita. O Algoritmo 4.1 inicia de maneira similar ao algoritmo conceitual. No passo 2 todas as regras cujo o valor de M sugere independência entre antecedente e consequente são removidas. A partir daí as dois algoritmos se diferenciam. No passo 3 são selecionadas todas as regras de tamanho 2 e são armazenadas no conjunto R_2, nesse passo não é possível descartar regras que não possuam o item de interesse. Já no passo 4 são selecionadas as regras de tamanho 3 e que possuam o item de interesse $\bar{a}$, as regras selecionadas são armazenadas no conjunto R_3.

Após dividir o conjunto inicial R em dois conjunto R_2 e R_3 uma tabela *hash* é criada para evitar iterações sobre todo o conjunto R_2 em busca de regras contidas em R_3. Portanto, no passo 5 o conjunto R_2 é transformado em uma tabela *hash* onde o valor é a própria regra e a chave é o par $antecedente$ e $consequente$, nessa ordem.

No passo 6 é realizada uma iteração sobre o conjunto R_3 (regras de tamanho 3), para cada regra tenta-se obter uma regra menor a partir da combinação de um item do antecedente com o item do consequente, passos 7 e 8, ou seja, r_{ant1} é o primeiro item do antecedente de r implicando no consequente de r e r_{ant2} é o segundo item do antecedente de r implicando no consequente de r. É importante frisar que algum ou ambos, r_{ant1} e r_{ant2}, podem não existir. Após isso, no passo 9 compara, caso existam, se r melhora o valor de M em relação as regras menores em pelo menos $\alpha\%$. Se sim, no passo 10 é decidido pela função *decida-tipo* a qual tipo pertence as regras r, r_{ant1} e r_{ant2}, com base na posição do item de interesse (antecedente ou consequente) e na existência de r_{ant1} e r_{ant2}, ou seja, é somente uma serie de verificações que gasta um tempo constante. Um tipo consiste da regra r e, se existirem, as regras r_{ant1} e r_{ant2}. O passo 11 adiciona o agrupamento obtido na lista de retorno. Caso o valor de M não seja melhorado por r então a regra não entra em nenhum tipo e não será retornada como saída do algoritmo.

Após definir o tipo no qual as regras de R_3 pertencem, o próximo passo é definir os tipos das regras bidirecionais (Tipo 1). O passo 14 inicia uma iteração para percorrer todas as regras da tabela *hash*. Caso o item de interesse exista na regra atual, passo 15, é verificado se existe uma regra para formar um conjunto do Tipo 1. Para isso, tenta obter da tabela *hash* uma regra em que a chave é o antecedente e o consequente da regra atual invertidos, passo 16. Caso exista a regra com tal chave, o agrupamento do Tipo 1 é criado e adicionado a lista de retorno e a regra obtida é removida do *hash* para evitar agrupamentos duplicados.

Ao final do algoritmo um subconjunto de regras do conjunto R estarão organizadas em pequenos conjuntos (de no máximo 3 regras) divididos em 8 tipos, visando mostrar o relacionamento entre regras que possam ajudar o especialista a compreender melhor como o item de interesse está relacionado aos demais itens da base de dados.

4.3 Representação gráfica das relações entre os agrupamentos de regras

Ao se analisar o conjunto de agrupamentos gerados pelo método proposto, notou-se que algumas regras se repetiam em diversos agrupamentos. Assim, verificou-se que seria possível condensar as informações obtidas pelos agrupamentos em uma visualização gráfica. Foi elaborado um grafo contendo todos os agrupamentos que estão conectados

entre si por alguma regra em comum, conforme pode ser visto na Figura 4.1.

Neste grafo, os nós em vermelho representam as regras que os agrupamentos compartilham (nomeada de regra pivô). Os nós azuis, representam os agrupamentos dos Tipos 2, 4, 6 e 7, e os nós amarelos representam os agrupamentos do Tipo 1. Os agrupamentos só se ligam as regras pivôs, essa abordagem foi utilizada para diminuir o número de arestas e facilitar a visualização.

No grafo apresentado não é possível estabelecer ligações entre os agrupamentos dos Tipo 3, 5 e 8. Os Tipo 5 e 8 são de apenas uma regra e portanto não possuem pivô. Já o Tipo 3 também não possui pivô, pois a regra de tamanho 2 só pode aparecer em um único agrupamento dado que não possui o item de interesse.

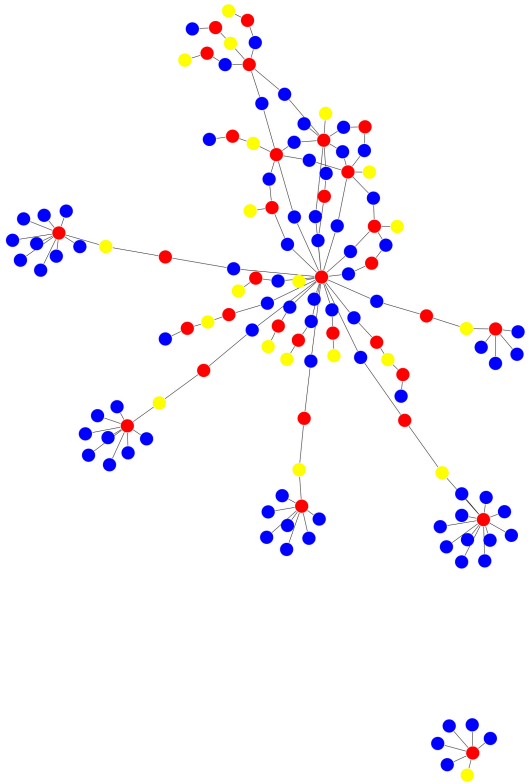

Figura 4.1: *Um grafo gerado para um conjunto de agrupamentos.*

Na Figura 4.2, é destacado um subgrafo mostrando as relações entre os agrupamentos dos Tipos 2 e 4. Este subgrafo mostra as regras relacionadas a associação entre o item de interesse implicando em outro item, em outras palavras, tudo que está relacionado a associação que aparece na regra pivô. O exemplo fictício da Figura 4.2 mostra regras relacionadas ao item de interesse diagnóstico de dengue ($DENGUE = sim$) e ocorrência de hemorragia ($HEMORRAGIA = sim$). Estas regras mostram que os medicamentos, AAS, ibuprofeno e escitalopram não possuem associação individual com hemorragia, porém potencializam a associação entre dengue e hemorragia. Também mostram que doenças hepáticas graves ($HEPATICA = grave$) e doenças hemofílicas ($HEMOFILIA = sim$) es-

tão individualmente a associadas a hemorragia e que quando ocorrem em conjunto com dengue a associação é potencializada. Por fim, a regra bidirecional (em amarelo), mostra que dengue e hemorragia possuem uma relação intrínseca uma com a outra.

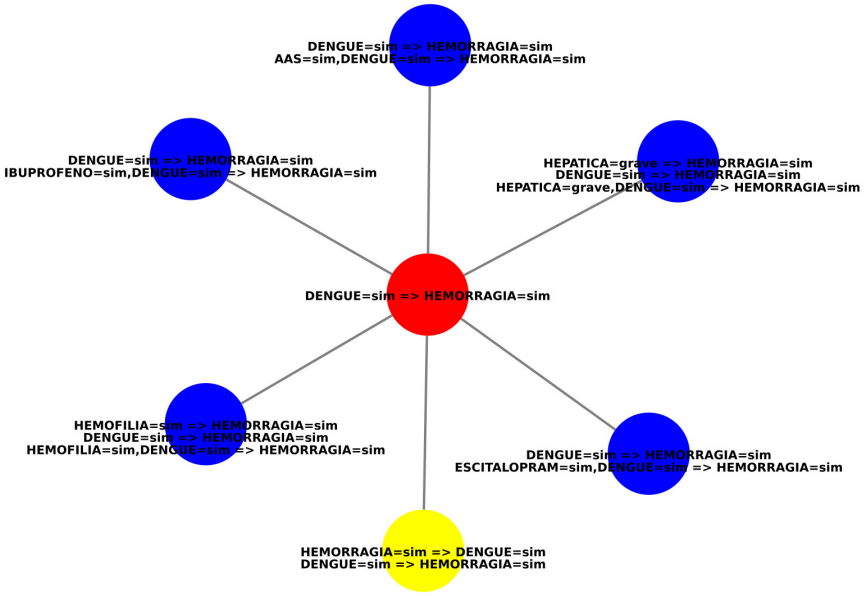

Figura 4.2: *Um grafo que reúne agrupamentos dos tipos 2 e 4.*

A grande quantidade de nós no centro do grafo da Figura 4.1 são dos Tipos 6 e 7, como nesses agrupamentos o item de interesse está no consequente há mais possibilidade de relações entre os agrupamentos, pois para agrupamentos do Tipo 6 existem potencialmente dois pivôs que podem ser ligados ao agrupamento. Por exemplo, note que na Figura 4.3 os nós azuis possuem duas ligações, com dois pivôs distintos (nós vermelhos).

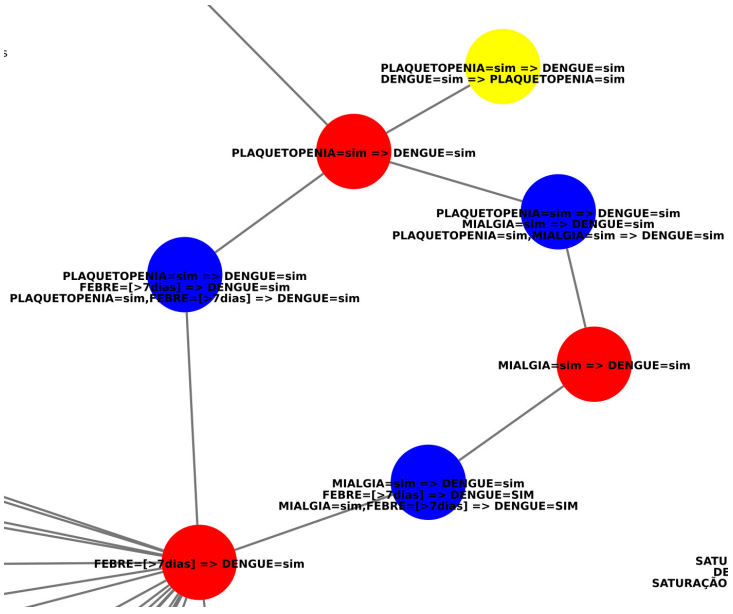

Figura 4.3: *Um grafo que reúne agrupamentos do tipo 6.*

Para as regras no antecedente apenas um pivô pode existir, porque a regra sem o item interesse não pode se repetir. Por exemplo, considere o agrupamento com a associação entre doença hepática grave ($HEPATICA = grave$), ocorrência de hemorragia ($HEMORRAGIA = sim$) e diagnóstico de dengue ($DENGUE = sim$) da Figura 4.2, a única possibilidade de $HEPATICA = grave \rightarrow HEMORRAGIA = sim$ existir no retorno do método é se ela estiver contida em uma regra de tamanho 3 que possua o item de interesse, portanto, apenas uma regra atende essa possibilidade.

O subgrafo apresentado na Figura 4.3 mostra a associação entre os sintomas de plaquetopenia ($PLAQUETOPENIA = sim$), febre prolongada ($FEBRE = [> 7dias]$) e mialgia ($MIALGIA = sim$) com dengue. Ademais, mostra que quando combinados os sintomas potencializam a associação. A regra em amarelo mostra que plaquetopenia está intimamente relacionada a dengue.

Por fim, caso esses agrupamentos estejam relacionados a um agrupamento do Tipo 1 então uma ligação entre as extremidades e o centro do grafo pode ser criada, o agrupamento que faz esse tipo de ligação foi nomeado de agrupamento ponte. Por exemplo, considere a Figura 4.4, existe um agrupamento do Tipo 1 (em amarelo) e ambos os pivôs possuem ligações com outros agrupamentos, os agrupamentos a direita da imagem mostram as associações dos pacientes com dengue ($DENGUE = sim$) que foram internados ($INTERNAÇÃO = sim$), enquanto o agrupamento mais ao centro a internação desempenha um fator para a ocorrência de dengue na base de dados e dessa forma se liga a outros agrupamentos que implicam em dengue (a esquerda na imagem), por exemplo, $FEBRE = [> 7dias]$.

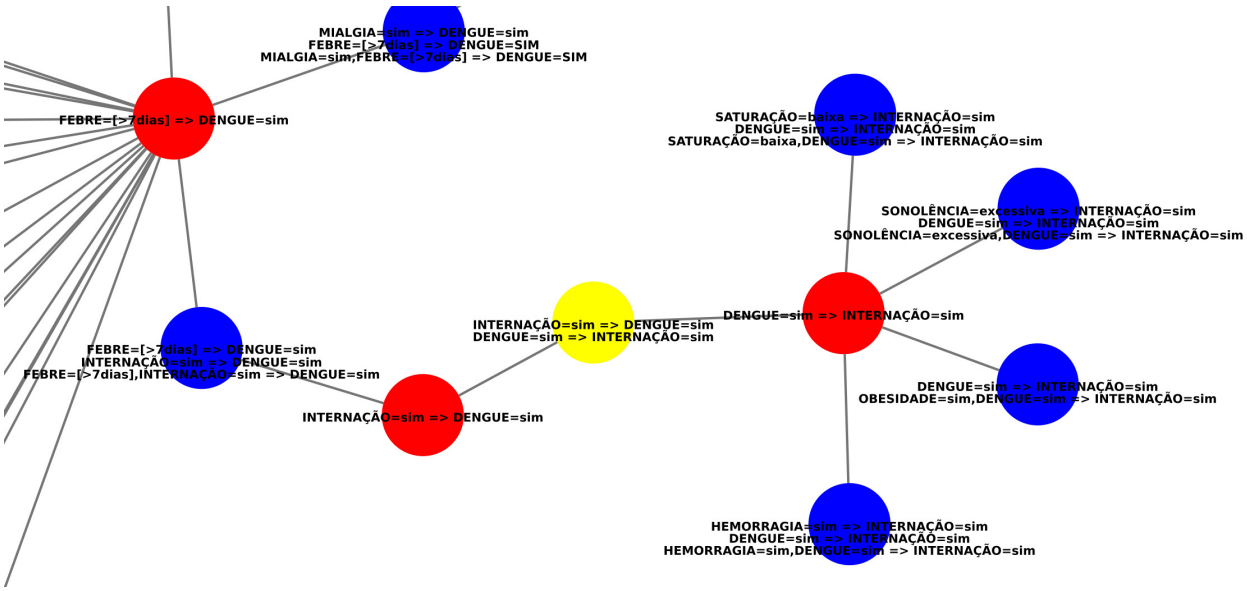

Figura 4.4: *Um exemplo de uma ponte entre o centro e uma extremidade do grafo.*

A parte à direita do subgrafo da Figura 4.4 mostra que baixa saturação

(*SATURAÇÃO* = *baixa*), rebaixamento dos níveis de consciência (*SONOLÊNCIA* = *excessiva*) e hemorragia (*HEMORRAGIA* = *sim*) assim como dengue se associam à internação. Quando estes fatores aparecem juntos a associação com internação é reforçada. No centro do subgrafo o agrupamento em amarelo (agrupamento ponte) mostra que internação e dengue estão intimamente relacionadas. Já a parte à esquerda do subgrafo mostra que a associação entre internação e dengue, bem como, febre prolongada ($FEBRE = [> 7 dias]$ associada à dengue. A partir dessa última associação é possível estabelecer uma ligação com os demais sintomas apresentados no subgrafo da Figura 4.3.

CAPÍTULO 5

Estudos de casos

Neste capítulo, quatro estudos de casos são apresentados com intuito de mostrar a aplicabilidade do método e possíveis interpretações para as saídas que ele produz. Todas as bases utilizadas são bases de dados reais. É importante ressaltar que o autor não é especialista no domínio das aplicações apresentadas, portanto as considerações feitas a seguir não tem como objetivo produzir conhecimento sobre o domínio aplicado, mas mostrar como o método pode ser uma ferramenta de auxílio na interpretação e geração de conhecimento para um especialista por meio de regras de associação.

Para cada estudo de caso, serão apresentados alguns agrupamentos de regras de cada tipo definido pelo método proposto. Os agrupamentos do Tipo 1 mostram itens que estão intrinsecamente associados de forma que a ocorrência de um item implica no outro e vice-versa. Os demais tipos se concentram em estratégias para avaliar e interpretar regras de tamanho 3. Estas estratégias podem ser divididas quando o item de interesse está no antecedente da regra, como possível causa (Tipos 2, 3, 4 e 5), e quando o item de interesse ocorre no consequente da regra, como possível efeito (Tipos 6, 7 e 8). Todos os agrupamentos de regras formados de todos os tipos podem ser acessados em *https://github.com/Luiz-Cintra-Experiments/masters-degree-experiments/tree/main/results*.

O algoritmo de ARM utilizado para a geração das regras de associação foi o Apriori implementado na linguagem R, versão 4.3.1, disponível na biblioteca *arules* versão 1.7-6 [30]. Os parâmetros utilizados foram: suporte mínimo de 1% (exceto no primeiro estudo de caso onde este valor foi de 5%), confiança mínima de 30% e tamanho máximo da regra igual a 3. Este valor de suporte foi escolhido porque alguns atributos possuem uma frequência baixa em determinados valores. Já o valor da confiança foi escolhido com intuito de evidenciar que regras de tamanho 2 com confiança mais baixa podem ter suas associações potencializadas caso o item de seu antecedente ocorra em concomitância com outro item, formando uma regra de tamanho 3 mais confiável.

O método proposto foi implementado na linguagem Python versão 3.8, disponível em *https://github.com/Association-Rules-Post-Processing/ARPPL*. Três medidas foram testadas, a saber: *lift*, *conviction* e *odds ratio*. O *lift* gerou uma quantidade reduzida

de agrupamentos enquanto a *conviction* produziu a maior quantidade de agrupamentos, já o *odds ratio* gerou uma quantidade intermediária de agrupamentos. Portanto, os parâmetros do método proposto foram: medida de interesse $M = Odds\ Ratio$, margem de dependência $\delta = 0,1$, melhoria mínima $\alpha = 10\%$, e confiança mínima para regras de tamanho 3 $c' = 50\%$. Os grafos foram gerados utilizando a biblioteca *Networkx* [28] na versão 2.8.8 [54], com a biblioteca *ForceAtlas2* versão 1.0 [63] sendo utilizada para melhorar a disposição dos nós no grafo.

5.1 Estudo de caso 1: Transtorno Disfórico Pré-menstrual

Os dados do primeiro estudo de caso são provenientes do trabalho publicado por Slyepchenko e colaboradores [66], posteriormente utilizados por Castro e colaboradores [10] como uma tarefa de ARM. O principal objetivo deste estudo foi comparar as características da doença, prevalência de transtornos psiquiátricos comórbidos e problemas de saúde mental específicos do sexo feminino entre mulheres com e sem Transtorno Disfórico Pré-menstrual (*Premenstrual Dysphoric Disorder* – PMDD). Foram estudadas 1.099 mulheres com Transtorno Afetivo Bipolar (TAB) que participaram do Programa de Aprimoramento de Tratamento Sistemático para Transtorno Bipolar (*Systematic Treatment Enhancement Program for Bipolar Disorder* – STEP-BD). O STEP-BD é o maior estudo de TAB financiado pelo Instituto Nacional de Saúde Mental (NIMH) dos Estados Unidos.

O algoritmo *Apriori* gerou ao total 1.659.270 regras. Para a aplicação do método proposto, o item *PMMD=yes* foi definido como item de interesse. Após a aplicação do método foram gerados 1.116 agrupamentos, que contemplam 1.864 regras. A seguir são apresentados alguns destes agrupamentos categorizados em seus tipos.

Tipo 1

O primeiro agrupamento de regras $\{1\text{-}pmdd_1, 1\text{-}pmdd_2\}$, apresentado na Tabela 5.1, mostra que tanto a presença de sintomas de sobrecarga ($OVERWHEL = Yes$) está associado ao PMDD quanto o PMDD está associado aos sintomas de sobrecarga.

Id	Regra	Sup.	Conf.	Odds ratio
$1\text{-}pmdd_1$	$OVERWHEL = Yes \rightarrow PMDD = yes$	34,12%	82,96%	20,96
$1\text{-}pmdd_2$	$PMDD = yes \rightarrow OVERWHEL = Yes$	34,12%	75,45%	20,96
$1\text{-}pmdd_3$	$HYPERSOM = Yes \rightarrow PMDD = yes$	21,66%	81,79%	9,52
$1\text{-}pmdd_4$	$PMDD = yes \rightarrow HYPERSOM = Yes$	21,66%	41,89%	9,52

Tabela 5.1: *PMDD - Regras Tipo 1*

Nesta associação existe um certo grau de simetria entre as duas regras, pois nenhuma possui uma confiança muito maior do a outra (Conf. = 82,96% para *1-pmdd*$_1$ e Conf. = 75,45% para *1-pmdd*$_2$). O que não ocorre para o segundo agrupamento $\{$*1-pmdd*$_3$, *1-pmdd*$_4\}$, que mostram que pacientes com hipersonia ($HYPERSOM = Yes$) tendem a ter PMDD (Conf. = 81,79% para *1-pmdd*$_3$), porém dentre os pacientes com PMDD menos da metade possui hipersonia (Conf. = 41,89% para *1-pmdd*$_4$).

Tipo 2

Os agrupamentos deste tipo visam ajudar a interpretar as regras de tamanho 3. Por exemplo, considere que um especialista obtenha separadamente a seguinte regra: $DRUGDEPP = yes, PMDD = yes \rightarrow IRRITABILITY\_DEP = yes$. Neste caso, a única informação que ele possui é que PMDD e dependência de drogas ($DRUGDEPP$) estão associadas a irritabilidade e depressão ($IRRITABILITY\_DEP$). Porém, sem haver uma busca manual no conjunto de todas as regras geradas, não é possível dizer se, isoladamente, PMDD e $DRUGDEPP$ também estariam associadas a $IRRITABILITY\_DEP$, nem a contribuição de cada item para essa associação. Na Tabela 5.2 o agrupamento de regras $\{$*2-pmdd*$_1$, *2-pmdd*$_2$, *2-pmdd*$_3\}$ mostra diretamente estas associações. A regra *2-pmdd*$_2$ aponta que o PMDD já possui uma associação forte com irritabilidade e depressão (Conf. = 99,8% e Odds Ratio = 454,9). Dependência de drogas também possui associação com irritabilidade e depressão (regra *2-pmdd*$_1$) e a presença deste fator junto com PMDD gerou uma associação em que o paciente sempre apresentará irritabilidade e depressão (Conf. = 100%) (regra *2-pmdd*$_3$).

Id	Regra	Sup.	Conf.	Odds ratio
2-pmdd$_1$	$DRUGDEPP = yes \rightarrow IRRITABILITY\_DEP = yes$	14,56%	82,05%	1,78
2-pmdd$_2$	$PMDD = yes \rightarrow IRRITABILITY\_DEP = yes$	45,13%	99,8%	454,93
2-pmdd$_3$	$DRUGDEPP = yes, PMDD = yes \rightarrow IRRITABILITY\_DEP = yes$	9,28%	100%	∞
2-pmdd$_4$	$PHOBIAPAST = yes \rightarrow ANXIETYPAST = yes$	24,2%	99,63%	292,87
2-pmdd$_5$	$PMDD = yes \rightarrow ANXIETYPAST = yes$	30,94%	68,41%	1,88
2-pmdd$_6$	$PHOBIAPAST = yes, PMDD = yes \rightarrow ANXIETYPAST = yes$	12,01%	100%	∞
2-pmdd$_7$	$POLARITY1ST = manic \rightarrow LENGTH10 = yes$	12,83%	77,9%	1,25
2-pmdd$_8$	$PMDD = yes \rightarrow LENGTH10 = yes$	30,94%	78,47%	1,47
2-pmdd$_9$	$POLARITY1ST = manic, PMDD = yes \rightarrow LENGTH10 = yes$	12,01%	88,89%	2,90

Tabela 5.2: *PMDD - Regras Tipo 2*

No segundo agrupamento de regras $\{$*2-pmdd*$_4$, *2-pmdd*$_5$, *2-pmdd*$_6\}$ da Tabela 5.2, fobia passada ($PHOBIAPAST$) possui uma associação forte com ansiedade passada ($ANXIESTYPAST$) (Conf. = 99,6% e Odds Ratio = 292,8) (regra *2-pmdd*$_4$). PMDD provocou ansiedade em 68,41% dos casos (regra *2-pmdd*$_5$) e, neste caso, a presença de PMDD com fobia completa a associação anterior de forma que todos os pacientes apresentem ansiedade passada (Conf. = 100%) (regra *2-pmdd*$_6$).

Também existem casos onde as associações individuais possuem graus de associação próximos, e a combinação dos itens dos antecedente das regras reforçam a associação com o consequente. Este é o caso do terceiro agrupamento de regras $\{2\text{-}pmdd_7, 2\text{-}pmdd_8, 2\text{-}pmdd_9\}$ da Tabela 5.2. Neste agrupamento, a polaridade do primeiro episódio de TAB ser mania ($POLARITY1ST$) e a presença de PMDD implicam na duração do ciclo menstrual em torno de 10 dias ($LENGTH10$) (regras $2\text{-}pmdd_7$ e $2\text{-}pmdd_8$, respectivamente). E a regra $2\text{-}pmdd_9$ aponta que a conjunção da polaridade com PMDD acarretam em ciclo de 10 dias com confiança e odds ratio maiores que as associações individuais (Conf. = 89,8% e Odds ratio = 2,9 da regra $2\text{-}pmdd_9$ contra Conf. = 77,9% e Odds ratio = 1,2 da regra $2\text{-}pmdd_7$ e Conf. = 78,4% e Odds ratio = 1,47 da regra $2\text{-}pmdd_8$).

Tipo 3

De forma similar ao tipo anterior os agrupamentos deste tipo visam contextualizar melhor uma regra de tamanho 3. A diferença em relação ao Tipo 2, é que neste caso existe apenas uma regra de tamanho 2 que dá suporte à interpretação da regra de tamanho 3. Ou seja, apenas a regra $3\text{-}pmdd_2$ (Tabela 5.3) não fornece uma informação mais completa sobre a associação entre os itens. Uma característica importante do Tipo 3 é que a regra de tamanho 2 não conterá o item de interesse, isso implica que o item de interesse apenas reforça o grau de associação. Por exemplo, no agrupamento da Tabela 5.3 $\{3\text{-}pmdd_1, 3\text{-}pmdd_2\}$ o item de interesse PMDD não está individualmente associado a bipolaridade do tipo 1 ($BDTYPE = 1$), porém o agrupamento mostra que é mais confiável identificar um paciente com $BDTYPE = 1$ quando se verifica PMDD e $AGORAPHOBIA$ do que apenas $AGORAPHOBIA$ (Conf. = 79,28% da regra $3\text{-}pmdd_1$ contra Conf. = 84,62% da regra $3\text{-}pmdd_2$). Além disso, a probabilidade de que a associação $AGORAPHOBIA = Yes, PMDD = yes \rightarrow BDTYPE = 1$ ocorra por acaso é menor do que em $AGORAPHOBIA = Yes \rightarrow BDTYPE = 1$ (Odds ratio = 2,04 para $3\text{-}pmdd_1$ contra Odds ratio = 2,90 para $3\text{-}pmdd_2$).

Id	Regra	Sup.	Conf.	Odds ratio
$3\text{-}pmdd_1$	$AGORAPHOBIA = Yes \rightarrow BDTYPE = 1$	8,01%	79,28%	2,04
$3\text{-}pmdd_2$	$AGORAPHOBIA = Yes, PMDD = yes \rightarrow BDTYPE = 1$	5%	84,62%	2,90

Tabela 5.3: *PMDD - Regras Tipo 3*

Tipo 4

Esse tipo de agrupamento difere do anterior apenas pelo item de interesse estar presente na associação de tamanho 2. Portanto, ao contrário do Tipo 3 o item de interesse necessariamente será o item que mais contribui para a associação, enquanto

o outro item atuará apenas como um reforço. No exemplo da Tabela 5.4, PMDD já está associado a irritabilidade e depressão (*IRRITABILITY_DEP*) como pode ser visto pela regra *4-pmdd*$_1$, porém quando acontece junto de fobia no passado (*PHOBIAPAST*) a associação resultante, regra *4-pmdd*$_2$ é mais forte que a anterior (Conf. = 100% e Odds ratio = ∞ para a *4-pmdd*$_2$ enquanto Conf. = 99,8% e Odds ratio = 454,93% para a regra *4-pmdd*$_1$). Note que *PHOBIAPAST* não está associado individualmente a *IRRITABILITY_DEP*.

Id	Regra	Sup.	Conf.	Odds ratio
4-pmdd$_1$	$PMDD = yes \rightarrow IRRITABILITY\_DEP = yes$	45,13%	99,8%	454,93
4-pmdd$_2$	$PHOBIAPAST = Yes, PMDD = yes \rightarrow IRRITABILITY\_DEP = yes$	12,01%	100%	∞

Tabela 5.4: *PMDD - Regras Tipo 4*

Tipo 5

Enquanto os outros tipos agrupavam regras, essa estratégia armazena as regras de tamanho 3 que não puderam ser agrupadas. Ao contrário dos outros tipos onde era possível verificar quais itens contribuíram mais para uma associação, neste caso, a ausência de associações fortes dos itens individualmente sugere que ambos os itens não possuem uma associação significativa com o consequente [1]. Além disso, estes itens se associam ao consequente apenas quando ocorrem juntos. Como é levado consideração o item de interesse e nesse tipo ele sempre está posicionado no antecedente, isso significa que o item de interesse só está relacionado ao consequente quando ocorre junto de outro item. Por exemplo, a regra *5-pmdd*$_1$ da Tabela 5.5 indica que tanto PMDD quanto dor moderada nos primeiros 5 anos menstruais ($F5Y\_PAIN = moderate$) se associam a depressão como polaridade do 1º episódio ($POLARITY1ST = depressive$) apenas quando aparecem juntas.

Id	Regra	Sup.	Conf.	Odds ratio
5-pmdd$_1$	$F5Y\_PAIN = moderate, PMDD = yes \rightarrow POLARITY1ST = depressive$	12,28%	80,36%	1,57

Tabela 5.5: *PMDD - Regras Tipo 5*

Tipo 6

Similar ao Tipo 2, este tipo de agrupamento também é composto por 3 regras, onde duas delas possuem um item no antecedente com uma associação individual com um mesmo consequente, e o antecedente da terceira regra é a conjunção dos antecedentes

[1] Nesse experimento isso significa que ou possuem suporte abaixo de 5%, ou confiança abaixo de 30%, ou antecedente e consequente são independentes (*odds ratio* igual a 1).

das regras anteriores. Porém, neste tipo, o item de interesse está no consequente da regra. Agrupamentos do Tipo 6 permitem visualizar a associação dos itens que fortalecem a ocorrência do item de interesse. O exemplo apresentado na Tabela 5.6, mostra a associação entre dificuldade de concentração (*DIFFCONC*) e agorafobia (*AGORAPHOBIA*) com PMDD (regra *6-pmdd*$_3$). É possível notar que *DIFFCONC* tem uma contribuição maior, pois possui uma associação individual mais forte com PMDD (Conf. = 83,42% e Odds ratio = 14,33 da *6-pmdd*$_1$, contra Conf. = 58,95% e Odds ratio = 1,95 da regra *6-pmdd*$_2$). Porém *AGORAPHOBIA* causa um grande reforço a associação resultante, pelo aumento tanto da confiança quanto do *odds ratio* (Conf. = 96% e Odds ratio = 33,83 para a regra *6-pmdd*$_3$ em comparação com Conf. = 83,42% e Odds ratio = 14,33 para a *6-pmdd*$_1$.

Id	Regra	Sup.	Conf.	Odds ratio
6-pmdd$_1$	$DIFFCONC = Yes \rightarrow PMDD = yes$	27,93%	83,42%	14,33
6-pmdd$_2$	$AGORAPHOBIA = yes \rightarrow PMDD = yes$	10,19%	58,95%	1,95
6-pmdd$_3$	$DIFFCONC = Yes, AGORAPHOBIA = yes \rightarrow PMDD = yes$	6,55%	96%	33,83

Tabela 5.6: *PMDD - Regras Tipo 6*

Tipo 7

O Tipo 7 segue o mesmo princípio do anterior, porém apenas um item está associado ao item de interesse individualmente. Sendo assim na própria definição fica explícito qual item contribui mais para associação com o item de interesse e qual atua como um reforço. Por exemplo, a regra *7-pmdd*$_2$ da Tabela 5.7 mostra que incontáveis fases de mania ($MANPHASE = uncountable$) e bipolaridade do tipo 1 ($BDTYPE = 1$) estão associadas ao PMDD. Além disso, $BDTYPE = 1$ não está associada individualmente a $PMDD = yes$ (não existe tal regra). Assim, $BDTYPE = 1$ pode estar atuando como um reforço na associação, sendo $MANPHASE = uncountable \rightarrow PMDD = yes$ a associação principal.

Id	Regra	Sup.	Conf.	Odds ratio
7-pmdd$_1$	$MANPHASE = uncountable \rightarrow PMDD = yes$	11,01%	57,62%	1,85
7-pmdd$_2$	$MANPHASE = uncountable, BDTYPE = 1 \rightarrow PMDD = yes$	7,37%	62,79%	2,24

Tabela 5.7: *PMDD - Regras Tipo 7*

Tipo 8

A ideia desse tipo é igual a apresentada no Tipo 5, ou seja, todas as regras que não estão contidas em regras menores são atribuídas a esse tipo. Isso significa, que dois itens só estão associados ao item de interesse quando aparecem juntos.

Id	Regra	Sup.	Conf.	Odds ratio
$8\text{-}pmdd_1$	$F5Y\_CYCLE10 = Yes, CONTRAC = Barrier \rightarrow PMDD = yes$	5,28%	50,43%	1,26

Tabela 5.8: *PMDD - Regras Tipo 8*

No caso apresentado na Tabela 5.8, duração de ciclo em torno de 10 dias nos 5 primeiros anos menstruais ($F5Y\_CYCLE10 = Yes$) e uso de preservativo como método contraceptivo ($CONTRAC = Barrier$) estão associadas ao PMDD apenas quando aparecem juntos (regra $8\text{-}pmdd_1$). Portanto, assim como no Tipo 5, a ausência de regras de tamanho 2 nos permite afirmar apenas que as associações individuais não são relevantes [2]. Logo é razoável supor que ambas contribuem de forma similar nessa associação.

Grafo de relações entre os agrupamentos

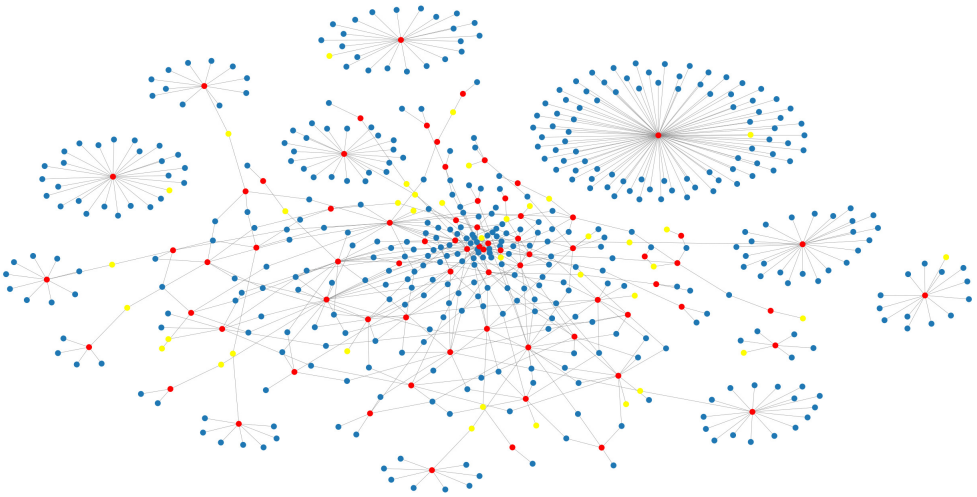

Figura 5.1: *PMDD - Grafo com todos os agrupamentos de regras relacionados*

A Figura 5.1 mostra um grafo de todos os agrupamentos gerados e que possuem relações entre si. As partes mais periféricas do grafo permitem verificar com bastante clareza os agrupamentos relacionados. Entretanto, a parte mais central ainda possui algumas sobreposições que prejudicam na identificação destas relações.

Mesmo assim, é possível destacar várias relações entre os agrupamentos de regras. Por exemplo, na Figura 5.2 aparece todos os conjuntos de regras que possuem a associação entre PMDD e mudanças de humor ($MOODSWNG$) (regra pivô representada como um nó vermelho no grafo), tais como, dependência de drogas no passado ($DRUGDEPP$),

[2] Nesse experimento isso significa que, ou possuem suporte abaixo de 5%, ou confiança abaixo de 30%, ou antecedente e consequente são independentes (*odds ratio* igual a 1).

incontáveis fases de mania ($MANPHASE = uncountable$), agorafobia atual ($PDAGORC$), qualquer ansiedade (ANX), dificuldade de concentração ($DIFFCONC$), além do PMDD se relacionar fortemente com $MOODSWING$ (regra bidirecional representada como um nó amarelo no grafo).

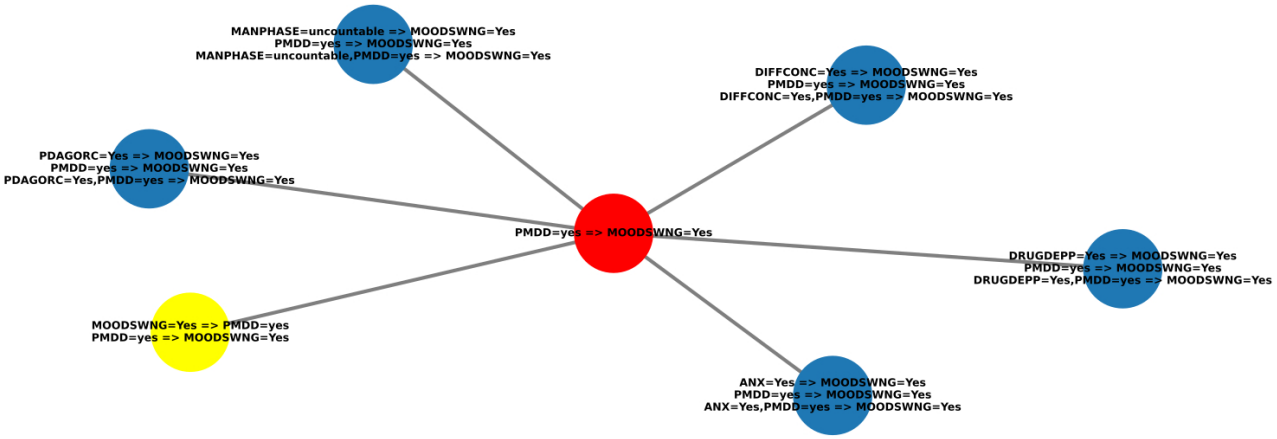

Figura 5.2: *PMDD - Subgrafo com agrupamentos tipos dos 2 e 4.*

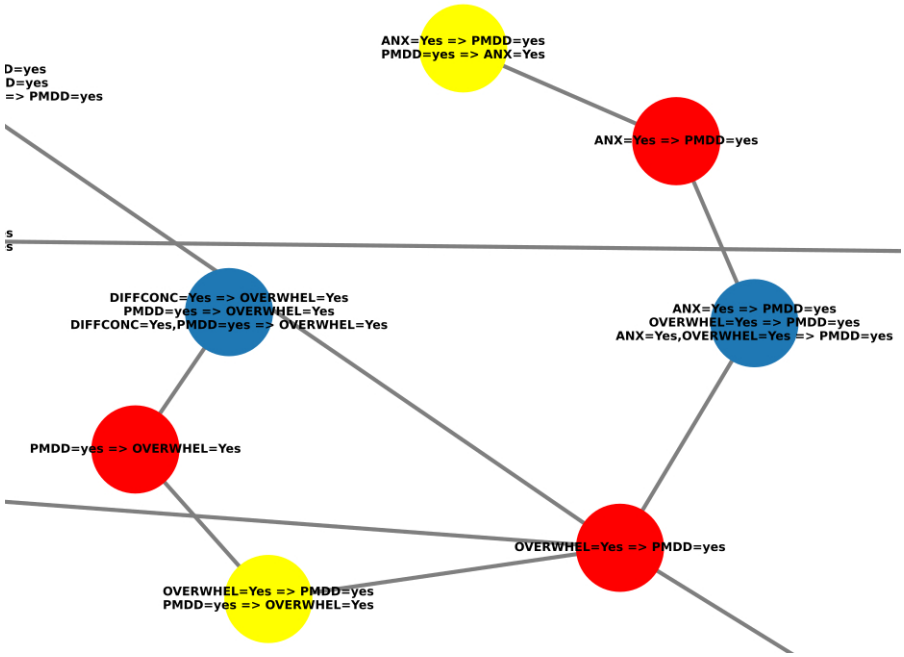

Figura 5.3: *PMDD - Agrupamento ponte. Ligação de agrupamentos com item de interesse no antecedente à agrupamentos com item de interesse no consequente por meio de um agrupamento do Tipo 1 (em amarelo).*

Outra relação que foi facilitada pelos grupos de agrupamentos foi a associação entre $DIFFCONC$ e PMDD com sentimento de sobrecarga ($OVERWHEL$), e que PMDD e $OVERWHEL$ estão relacionados de maneira bidirecional, mostrada na Figura 5.3.

Isso possibilita que agrupamentos com item de interesse no antecedente se ligue a agrupamentos com item de interesse no consequente por meio de uma agrupamento do Tipo 1 (nó amarelo). O agrupamento do Tipo 1 na parte superior do grafo (nó amarelo) também mostra que ansiedade ($ANX = Yes$) está intimamente associada a PMDD, porém não aparece no consequente de regras de tamanho 3 dos tipos 2 e 4, pois o agrupamento ponte dessa associação só se liga a uma regra pivô. Atuando como antecedente, ansiedade e sobrecarga potencializam suas associações com PMDD quando aparecem juntos.

5.2 Estudo de Caso 2: Síndrome Respiratória Aguda Grave

Neste estudo de caso foi utilizada uma base de dados aberta sobre Síndrome Respiratória Aguda Grave (SRAG) do governo brasileiro para investigar fatores ligados à COVID-19 [3]. A base para o estudo de caso foi gerada a partir de quatro bases de dados SRAG, cada uma correspondente aos anos de 2019 até 2022 [56], [57], [58]. Além da integração destas 4 bases, foram executadas operações de processamento dos dados, como: colunas foram renomeadas, valores foram alterados para uma descrição mais legível e seleção de variáveis. Tal processamento resultou em uma base final [4] contendo 3.389.419 registros de pacientes com SRAG e 35 variáveis que correspondem a sintomas, diagnósticos, dados sociodemográficos e resultado de exames. O processo de geração da base de dados está disponível no Apêndice D. A base consiste de casos notificados de SRAG no Brasil entre 2019 e 2022. Na base final 95,48% dos pacientes foram hospitalizados com sintomas de SRAG e 63.92% tiveram o diagnóstico fechado de Covid. Assim, o item de interesse utilizado para o pós-processamento foi $DIAGNOSTICO = COVID$.

A execução do Apriori resultou em 207.787 regras de associação. Após a aplicação do método proposto, 318 agrupamento de regras foram gerados, correspondendo a 424 regras distintas. A seguir são destacados alguns destes agrupamentos.

A regra *1-covid*$_1$ da Tabela 5.9 mostra uma associação de certa forma inesperada, o resutado de tomografia típico de covid ($TOMOGRAFIA = tipico\_covid$) implica em $COVID$ em 92% dos casos com um bom *odds ratio* (Odds ratio = 9,62), se o resultado da tomografia foi típico de covid era de se esperar que a confiança fosse maior, portando obtendo essa regra o especialista pode querer analisar a fundo os casos onde $TOMOGRAFIA = tipico\_covid$ não resultou em um diagnóstico $COVID$. A regra

[3] Url das bases utilizadas: https://opendatasus.saude.gov.br/dataset?tags=SRAG

[4] Base disponível em: https://github.com/Luiz-Cintra-Databases/SRAG-OpenDataSUS-2019-2022/tree/main/srag/database

1-covid$_2$ complementa a regra anterior mostrando os diagnósticos de covid na maioria das vezes não possuem um resultado de tomografia típico de covid (Conf. = 34,65%), isso não necessariamente implica que os paciente tenham sintomas atípicos na maioria dos casos, pois o atributo $TOMOGRAFIA$ pode assumir valores como *não* (não realizado) ou até mesmo não estar preenchido. Além disso, outros métodos de diagnósticos podem ter sido utilizados. Assim como o caso anterior, a regra *1-covid*$_3$ também mostra uma associação inesperada, pois em 79% dos casos em que a causa do óbito foi preenchida como covid o diagnóstico final foi covid (Conf. = 79,43% para a regra *1-covid*$_3$), portanto tal associação pode ser de interesse do especialista durante sua investigação. Além disso, a regra *1-covid*$_4$ mostra a proporção de mortes em paciente com covid que procuraram assistência médica, (Conf. = 31% para a regra *1-covid*$_4$). Por fim, as regras *1-covid*$_5$ e *1-covid*$_6$, mostram que a maioria dos pacientes de UTI estavam com covid (Conf. = 70,18% para *1-covid*$_5$), porém que 32% dos pacientes com covid foram para UTI (Conf. = 32,25% para a regra *1-covid*$_6$).

Id	Regra	Sup.	Conf.	Odds ratio
1-covid$_1$	$TOMOGRAFIA = tipico\_covid \rightarrow DIAGNOSTICO = covid$	22,15%	92,16%	9,62
1-covid$_2$	$DIAGNOSTICO = covid \rightarrow TOMOGRAFIA = tipico\_covid$	22,15%	34,65%	9,62
1-covid$_3$	$EVOLUCAO = obito\_covid \rightarrow DIAGNOSTICO = covid$	19,80%	79,43%	2,70
1-covid$_4$	$DIAGNOSTICO = covid \rightarrow EVOLUCAO = obito\_covid$	19,80%	30,98%	2,70
1-covid$_5$	$UTI = sim \rightarrow DIAGNOSTICO = covid$	20,61%	70,18%	1,49
1-covid$_6$	$DIAGNOSTICO = covid \rightarrow UTI = sim$	20,61%	32,25%	1,49

Tabela 5.9: *Covid - Regras Tipo 1*

Id	Regra	Sup.	Conf.	Odds ratio
2-covid$_1$	$DOR\_ABDOMINAL = sim \rightarrow FEBRE = sim$	3,08%	60,48%	1,33
2-covid$_2$	$DIAGNOSTICO = covid \rightarrow FEBRE = sim$	35,62%	55,73%	1,24
2-covid$_3$	$DOR\_ABDOMINAL = sim, DIAGNOSTICO = covid \rightarrow FEBRE = sim$	2,12%	64,02%	1,55
2-covid$_4$	$DIARREIA = sim \rightarrow TOSSE = sim$	8,45%	74,44%	1,47
2-covid$_5$	$DIAGNOSTICO = covid \rightarrow TOSSE = sim$	43,63%	68,26%	1,11
2-covid$_6$	$DIARREIA = sim, DIAGNOSTICO = covid \rightarrow TOSSE = sim$	6,24%	77,19%	1,70
2-covid$_7$	$PNEUMOPATIA = sim \rightarrow CARDIOPATIA = sim$	2,07%	48,17%	2,41
2-covid$_8$	$DIAGNOSTICO = covid \rightarrow CARDIOPATIA = sim$	19,92%	31,17%	1,40
2-covid$_9$	$PNEUMOPATIA = sim, DIAGNOSTICO = covid \rightarrow CARDIOPATIA = sim$	1,08%	52,02%	2,75

Tabela 5.10: *Covid - Regras Tipo 2*

Em agrupamentos do Tipo 2, Tabela 5.10, foram encontradas regras que mostram o reforço mútuo entre Covid e alguns sintomas em associações com outros sintomas, como dor abnominal e febre $\{2\text{-}covid_1, 2\text{-}covid_2, 2\text{-}covid_3\}$, diarréia e tosse $\{2\text{-}covid_4, 2\text{-}covid_5, 2\text{-}covid_6\}$. O reforço é mútuo pois as confianças de *2-covid*$_1$ e *2-covid*$_2$ (Conf. = 60,48% e Conf. = 55,73% respectivamente), bem como de *2-covid*$_4$ e *2-covid*$_5$ (Conf. = 74,44% e Conf. = 68,26% respectivamente) são próximas. Outro agrupamento destacado mostra que a Covid reforça a associação entre pneumopatia e cardiopatia, as regras $\{2\text{-}covid_7, 2\text{-}covid_8, 2\text{-}covid_9\}$. Neste caso a Covid atua como reforço,

pois a confiança da associação entre Covid e cardiopatia (Conf. = 31,17% para a regra *2-covid*$_8$) é menor que a associações somente entre pneumopatia e cardiopatia (Conf. = 48,17% para a regra *2-covid*$_7$).

Os agrupamentos selecionados como exemplo do Tipo 3 são apresentados na Tabela 5.11. Todos os agrupamentos desse tipo, mostram o item de interesse $DIAGNOSTICO = covid$ atuando como reforço a uma associação individual entre dois itens. Os dois agrupamentos ($\{$*3-covid*$_1$, *3-covid*$_2\}$ e $\{$*3-covid*$_3$, *3-covid*$_4\}$) mostram que dor abdominal está associada a fadiga e desconforto respiratório, além disso, Covid torna tais associações mais fortes (regras *3-covid*$_2$ e *3-covid*$_4$).

Id	Regra	Sup.	Conf.	Odds ratio
3-covid$_1$	$DOR\_ABDOMINAL = sim \rightarrow FADIGA = sim$	2,48%	48,74%	4,60
3-covid$_2$	$DOR\_ABDOMINAL = sim, DIAGNOSTICO = covid \rightarrow FADIGA = sim$	1,83%	55,22%	5,82
3-covid$_3$	$DOR\_ABDOMINAL = sim \rightarrow DESCONFORTO\_RESP = sim$	3,11%	61,06%	1,32
3-covid$_4$	$DOR\_ABDOMINAL = sim, DIAGNOSTICO = covid \rightarrow DESCONFORTO\_RESP = sim$	2,17%	65,36%	1,58

Tabela 5.11: *Covid - Regras Tipo 3*

Um agrupamento do Tipo 4 é apresentado na Tabela 5.12. Ele mostra como Covid e tosse possuem uma associação (*4-covid*$_1$), e que pacientes que também apresentam dor abdominal tem essa associação reforçada (*4-covid*$_2$), embora dor abdominal não esteja diretamente associada com tosse.

Id	Regra	Sup.	Conf.	Odds ratio
4-covid$_1$	$DIAGNOSTICO = covid \rightarrow TOSSE = sim$	43,63%	68,26%	1,11
4-covid$_2$	$DOR\_ABDOMINAL = sim, DIAGNOSTICO = covid \rightarrow TOSSE = sim$	2,44%	73,57%	1,35

Tabela 5.12: *Covid - Regras Tipo 4*

Um agrupamento do Tipo 5 é apresentado na Tabela 5.13. Assim como o exemplo anterior ele mostra a associação entre sintomas e Covid. No caso, Covid e vômito se reforçam na associação com desconforto respiratório (*5-covid*$_1$), ou seja, Covid está associada a desconforto respiratório apenas quando outro sintoma também acontece, no exemplo esse sintoma é vômito.

Id	Regra	Sup.	Conf.	Odds ratio
5-covid$_1$	$VOMITO = sim, DIAGNOSTICO = covid \rightarrow DESCONFORTO\_RESP = sim$	2,98%	58,17%	1,15

Tabela 5.13: *Covid - Regras Tipo 5*

Os agrupamentos do Tipo 6 trazem associações onde Covid é consequência de outros fatores. Os exemplos destacados na Tabela 5.14, mostram um perfil etário dos pacientes com Covid para o primeiro e segundo quadrimestres de 2021. Esses agrupamentos podem forma destacados pelo fato de que não existem outros agrupamentos que associam $DIAGNOSTICO = covid$ à uma outra data e idade. Eles mostram que no primeiro quadrimestre a associação entre idade e o diagnóstico de Covid foi mais ampla

do no segundo quadrimestre, entre 30 e 75 anos no primeiro quadrimestre contra 30 à 60 anos no segundo quadrimestre.

Pode-se dar um destaque maior aos agrupamentos $\{6\text{-}covid_1, 6\text{-}covid_2, 6\text{-}covid_3\}$ e $\{6\text{-}covid_3, 6\text{-}covid_4, 6\text{-}covid_5\}$ que detalham o perfil etário do segundo quadrimestre. Neles as regras mais gerais ($6\text{-}covid_3$ e $6\text{-}covid_6$) aumentam substancialmente tanto o *odds ratio* quanto a confiança. A regra $6\text{-}covid_3$ mais que dobra o *odds ratio* em relação às regras $6\text{-}covid_1$ e $6\text{-}covid_2$ (*Odds ratio* = 4,48 de $6\text{-}covid_3$ contra 1,93 da regra $6\text{-}covid_1$ e 1,98 da regra $6\text{-}covid_2$), bem como a confiança aumenta em 12% (Conf. = 88,36% para $6\text{-}covid_3$ enquanto Conf. = 75,44% para $6\text{-}covid_1$ e Conf. = 76,11% para $6\text{-}covid_2$). Já a regra $6\text{-}covid_6$ tem desempenho ligeiramente menor mas que ainda é relevante, ela não chega a dobrar o *odds ratio* (*Odds ratio* = 4,09 de $6\text{-}covid_6$ contra *Odds ratio* = 2,22 de $6\text{-}covid_5$) mas tem um aumento de confiança um pouco maior que 10% (Conf. = 87,26% para $6\text{-}covid_6$ contra Conf. = 77% para $6\text{-}covid_5$).

Id	Regra	Sup.	Conf.	Odds ratio
$6\text{-}covid_1$	$DT\_SINTOMAS = (5/2021\text{-}8/2021] \rightarrow DIAGNOSTICO = covid$	13,6%	75,44%	1,93
$6\text{-}covid_2$	$IDADE = (30a\text{-}45a] \rightarrow DIAGNOSTICO = covid$	11,69%	76,11%	1,98
$6\text{-}covid_3$	$DT\_SINTOMAS = (5/2021\text{-}8/2021], IDADE = (30a\text{-}45a] \rightarrow DIAGNOSTICO = covid$	3,55%	88,36%	4,48
$6\text{-}covid_4$	$DT\_SINTOMAS = (5/2021\text{-}8/2021] \rightarrow DIAGNOSTICO = covid$	13,6%	75,44%	1,93
$6\text{-}covid_5$	$IDADE = (45a\text{-}60a] \rightarrow DIAGNOSTICO = covid$	17,36%	77%	2,22
$6\text{-}covid_6$	$DT\_SINTOMAS = (5/2021\text{-}8/2021], IDADE = (45a\text{-}60a] \rightarrow DIAGNOSTICO = covid$	4,64%	87,26%	4,09
$6\text{-}covid_7$	$DT\_SINTOMAS = (1/2021\text{-}4/2021] \rightarrow DIAGNOSTICO = covid$	20,23%	81,6%	3,20
$6\text{-}covid_8$	$IDADE = (30a\text{-}45a] \rightarrow DIAGNOSTICO = covid$	11,69%	76,11%	1,98
$6\text{-}covid_9$	$DT\_SINTOMAS = (1/2021\text{-}4/2021], IDADE = (30a\text{-}45a] \rightarrow DIAGNOSTICO = covid$	3,8%	87,89%	4,29
$6\text{-}covid_{10}$	$DT\_SINTOMAS = (1/2021\text{-}4/2021] \rightarrow DIAGNOSTICO = covid$	20,23%	81,6%	3,20
$6\text{-}covid_{11}$	$IDADE = (45a\text{-}60a] \rightarrow DIAGNOSTICO = covid$	17,36%	77%	2,22
$6\text{-}covid_{12}$	$DT\_SINTOMAS = (1/2021\text{-}4/2021], IDADE = (45a\text{-}60a] \rightarrow DIAGNOSTICO = covid$	6%	88,79%	4,83
$6\text{-}covid_{13}$	$DT\_SINTOMAS = (1/2021\text{-}4/2021] \rightarrow DIAGNOSTICO = covid$	20,23%	81,6%	3,20
$6\text{-}covid_{14}$	$IDADE = (60a\text{-}75a] \rightarrow DIAGNOSTICO = covid$	17,64%	70,52%	1,48
$6\text{-}covid_{15}$	$DT\_SINTOMAS = (1/2021\text{-}4/2021], IDADE = (60a\text{-}75a] \rightarrow DIAGNOSTICO = covid$	6,24%	85,95%	3,72

Tabela 5.14: *Covid - Regras Tipo 6*

Os demais agrupamentos da Tabela 5.14 ($\{6\text{-}covid_7, 6\text{-}covid_8, 6\text{-}covid_9\}$, $\{6\text{-}covid_{10}, 6\text{-}covid_{11}, 6\text{-}covid_{12}\}$ e $\{6\text{-}covid_{13}, 6\text{-}covid_{14}, 6\text{-}covid_{15}\}$) mostram o perfil etário dos casos de Covid do primeiro quadrimestre, nesses casos as regras menos específicas ($6\text{-}covid_9$, $6\text{-}covid_{12}$ e $6\text{-}covid_{15}$) tem um incremento menor em relação a suas sub-regras do que as regras do segundo quadrimestre ($6\text{-}covid_3$ e $6\text{-}covid_6$).

Id	Regra	Sup.	Conf.	Odds ratio
$7\text{-}covid_1$	$OBESIDADE = sim \rightarrow DIAGNOSTICO = covid$	4,97%	81,1%	2,54
$7\text{-}covid_2$	$OBESIDADE = sim, VACINADO = não \rightarrow DIAGNOSTICO = covid$	1,35%	90,23%	5,30
$7\text{-}covid_3$	$DIABETES = sim \rightarrow DIAGNOSTICO = covid$	14,08%	71,66%	1,55
$7\text{-}covid_4$	$DIABETES = sim, VACINADO = não \rightarrow DIAGNOSTICO = covid$	2,31%	81,78%	2,59
$7\text{-}covid_5$	$CARDIOPATIA = sim \rightarrow DIAGNOSTICO = covid$	19,92%	69,31%	1,40
$7\text{-}covid_6$	$CARDIOPATIA = sim, VACINADO = não \rightarrow DIAGNOSTICO = covid$	3,4%	79,57%	2,26

Tabela 5.15: *Covid - Regras Tipo 7*

Com relação aos agrupamentos do Tipo 7, Tabela 5.15, destaca-se as associações entre as comorbidades: obesidade, diabetes e cardiopatia com diagnóstico de Covid (regras $7\text{-}covid_1$, $7\text{-}covid_3$ e $7\text{-}covid_5$). Embora um paciente não vacinado ($VACINADO = não$) não esteja individualmente associado a diagnóstico de Covid, em pacientes com as comorbidades citadas anteriormente essa chance é aumentada, pois tanto a confiança quanto o *odds ratio* das regras de tamanho 3 são maiores que as regras de tamanho 2 em cada agrupamento ($\{7\text{-}covid_1, 7\text{-}covid_2\}$, $\{7\text{-}covid_3, 7\text{-}covid_4\}$, $\{7\text{-}covid_5, 7\text{-}covid_6\}$).

A Tabela 5.16 apresenta uma regra do Tipo 8. Essa regra mostra que $DOR\_ABDOMINAL$ e $DESCONFORTO\_RESP$ implicam em Covid apenas quando ocorrem juntas.

Id	Regra	Sup.	Conf.	Odds ratio
$8\text{-}covid_1$	$DOR\_ABDOMINAL = sim, DESCONFORTO\_RESP = sim \rightarrow DIAGNOSTICO = covid$	$2,17\%$	$69,72\%$	$1,37$

Tabela 5.16: *Covid - Regras Tipo 8*

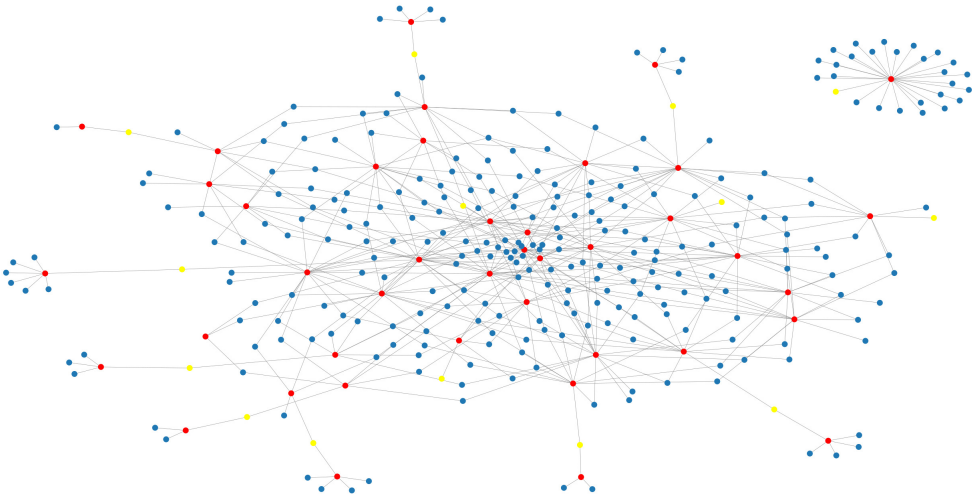

Figura 5.4: *Covid - Grafo de todos os agrupamentos relacionados.*

Um grafo também foi gerado para os agrupamentos obtidos a partir das regras sobre Covid. É possível notar pela Figura 5.4 que o mesmo padrão se repete para esses agrupamentos, o centro do grafo possui menos sobreposições que o grafo do estudo de caso anterior, porém pequenos trechos ainda possuem problema com sobreposições.

A Figura 5.5 destaca uma relação de agrupamentos em torno do associação entre diagnóstico de Covid e fator de risco, onde diagnóstico de Covid implica na ocorrência de fator de risco. Três fatores não associados individualmente a fator de risco aparecem potencializando essa associação. Temos duas datas de primeiro sintomas ($DT\_SINTOMAS = 1/2022\text{-}4/2022$ e $DT\_SINTOMAS = 5/2022\text{-}8/2022$) que mostram que no dois primeiros quadrimestres de 2022 quem tinha Covid a associação entre Covid

e fator de risco foi mais forte. O outro fator mostra o analfabetismo ($ESCOLARIDADE = analfabeto$) como um potencializador da associação entre Covid e fator de risco. Um agrupamento mostra que o resultado do raio-x classificado como outro ($RAIOX\_RES = outro$) possui uma associação individual com fator de risco e quando o resultado do raio-x classificado como outro e diagnóstico de Covid acontecem de forma concomitante a associação é reforçada.

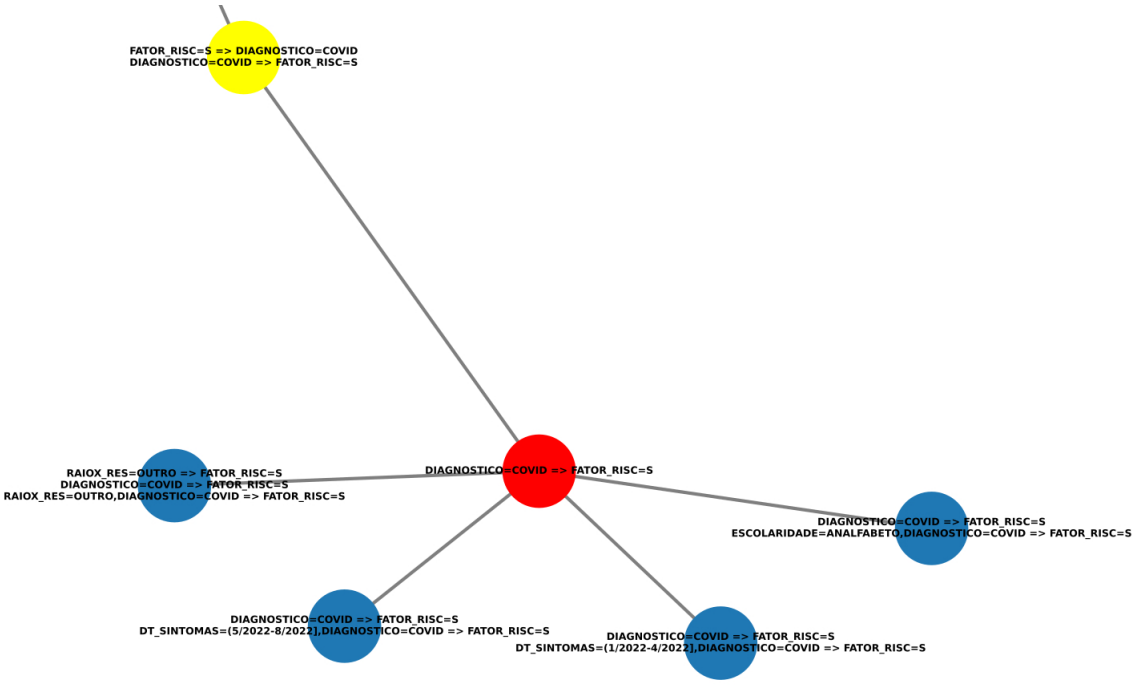

Figura 5.5: *Covid - Relação destacada de agrupamentos.*

5.3 Estudo de Caso 3: Óbito por COVID-19

Para esse estudo a base do estudo de caso 2 foi utilizada para gerar uma nova base, em que apenas os pacientes diagnosticados com Covid foram selecionados. A base resultante consiste em 2.166.443 dados de pacientes e 34 variáveis (a variável diagnóstico ($DIAGNOSTICO$) foi removida) [5].

A execução do Apriori resultou em 187.407 regras de associação. O item de interesse escolhido foi $EVOLUCAO = obito\_covid$. As regras de associação foram submetidas ao método proposto e um total de 116 agrupamentos foram gerados, que correspondeu a 215 regras de associações distintas. Com exceção dos Tipos 7 e 8, os demais tipos geraram agrupamentos, porém apenas os Tipos 1, 2 e 6 foram escolhidos para serem apresentados em detalhes.

[5]Base disponível em: https://github.com/Luiz-Cintra-Databases/SRAG-OpenDataSUS-2019-2022/tree/main/covid/database

A Tabela 5.17 mostra os agrupamentos selecionados do Tipo 1. Os dois primeiros agrupamentos mostram procedimentos médicos que estão associados a mortes por Covid. O primeiro agrupamento ($\{1\text{-}obito_1, 1\text{-}obito_2\}$) mostra que os pacientes que necessitam de suporte respiratório invasivo acabam vindo a óbito em 74% das vezes (Conf. = 74,49% para a $1\text{-}obito_1$), porém menos que a metade dos pacientes que vieram a óbito por Covid chegam a necessitar desse tipo de suporte (Conf. = 41,32% para a regra $1\text{-}obito_2$). Já o segundo agrupamento ($\{1\text{-}obito_3, 1\text{-}obito_4\}$) mostra que as regras possuem uma associação mais simétrica, tanto pouco mais da metade dos pacientes da UTI vieram a óbito pela Covid (Conf. = 53,11% para a regra $1\text{-}obito_3$) quanto pouco mais da métada dos pacientes que morreram de Covid foram internados na UTI (Conf. = 55,28% para a regra $1\text{-}obito_4$).

Id	Regra	Sup.	Conf.	Odds ratio
$1\text{-}obito_1$	$SUPORTE\_VENTILATORIO = invasivo \rightarrow EVOLUCAO = obito\_covid$	12,80%	74,49%	10,38
$1\text{-}obito_2$	$EVOLUCAO = obito\_covid \rightarrow SUPORTE\_VENTILATORIO = invasivo$	12,80%	41,32%	10,38
$1\text{-}obito_3$	$UTI = sim \rightarrow EVOLUCAO = obito\_covid$	17,13%	53,11%	4,41
$1\text{-}obito_4$	$EVOLUCAO = obito\_covid \rightarrow UTI = sim$	17,13%	55,28%	4,41
$1\text{-}obito_5$	$DESCONFORTO\_RESP = sim \rightarrow EVOLUCAO = obito\_covid$	19,31%	35,30%	1,57
$1\text{-}obito_6$	$EVOLUCAO = obito\_covid \rightarrow DESCONFORTO\_RESP = sim$	19,31%	62,33%	1,57
$1\text{-}obito_7$	$BAIXA\_SATURACAO = sim \rightarrow EVOLUCAO = obito\_covid$	21,75%	35,27%	1,72
$1\text{-}obito_8$	$EVOLUCAO = obito\_covid \rightarrow BAIXA\_SATURACAO = sim$	21,75%	70,19%	1,72
$1\text{-}obito_9$	$DISPNEIA = sim \rightarrow EVOLUCAO = obito\_covid$	23,42%	33,87%	1,58
$1\text{-}obito_{10}$	$EVOLUCAO = obito\_covid \rightarrow DISPNEIA = sim$	23,42%	75,58%	1,58
$1\text{-}obito_{11}$	$FATOR\_RISCO = sim \rightarrow EVOLUCAO = obito\_covid$	22,52%	37,28%	2,19
$1\text{-}obito_{12}$	$EVOLUCAO = obito\_covid \rightarrow FATOR\_RISCO = sim$	22,52%	72,68%	2,19

Tabela 5.17: *Óbitos covid - Regras Tipo 1*

Os próximos três agrupamentos mostram a associação de sintomas e óbito por Covid. É interessante notar que em todos os agrupamentos os sintomas, desconforto respiratório, dispneia e baixa saturação, não implicam em óbito na maioria das vezes (baixa confiança das regras $1\text{-}obito_5$, $1\text{-}obito_7$ e $1\text{-}obito_9$), porém, os pacientes que vieram a óbito apresentaram tais sintomas em 70% ou mais dos casos (todos ligados a respiração, regras $1\text{-}obito_6$, $1\text{-}obito_8$ e $1\text{-}obito_{10}$).

Por fim, o último agrupamento ($\{1\text{-}obito_{11}, 1\text{-}obito_{12}\}$) mostra a associação entre algum fator de risco e óbito por Covid, similar aos três agrupamentos anteriores o fator de risco em si não implica em óbito por Covid na maioria dos casos (Conf. = 37,28% para a regra $1\text{-}obito_{11}$), porém a maioria dos paciente que morreram de Covid possuiam algum fator de risco (Conf. = 72,68% para a regra $1\text{-}obito_{12}$).

Dois agrupamentos do Tipo 2 estão ilustrados na Tabela 5.18. O primeiro agrupamento $\{2\text{-}obito_1, 2\text{-}obito_2, 2\text{-}obito_3\}$ mostra as associações individuais entre óbito por Covid ($EVOLUCAO = obito\_covid$) e paciente vacinado ($VACINADO = sim$) à algum fator de risco ($FATOR\_RISCO = sim$), regras $2\text{-}obito_1$ e $2\text{-}obito_2$. Ademais, a associação resultante da combinação de $EVOLUCAO = obito\_covid$ e $VACINADO = sim$ re-

sulta em uma associação mais forte com $FATOR\_RISCO = sim$ (Conf. = 81,11% para $2\text{-}obito_3$ em comparação com Conf. = 72,68% para $2\text{-}obito_1$ e Conf. = 70,52% para $2\text{-}obito_2$). O segundo agrupamento mostra duas associações com um valor de *odds ratio* superior a 10, que representam as associações individuais entre internação na UTI ($UTI = sim$) e óbito por Covid ($EVOLUCAO = obito\_covid$) à necessidade de suporte ventilatório invasivo ($SUPORTE\_VENTILATORIO = invasivo$), regras $2\text{-}obito_4$ e $2\text{-}obito_5$. O agrupamento também mostra que quando $UTI = sim$ e $EVOLUCAO = obito\_covid$ acontecem juntos aumenta a chance que o paciente tenha tido necessidade de $SUPORTE\_VENTILATORIO = invasivo$ (Conf. = 64,73% para $2\text{-}obito_3$ em comparação com Conf. = 44,78% para $2\text{-}obito_1$ e Conf. = 41,32% para $2\text{-}obito_2$).

Id	Regra	Sup.	Conf.	Odds ratio
$2\text{-}obito_1$	$EVOLUCAO = obito\_covid \rightarrow FATOR\_RISCO = sim$	22,52%	72,68%	2,19
$2\text{-}obito_2$	$VACINADO = sim \rightarrow FATOR\_RISCO = sim$	14,14%	70,52%	1,74
$2\text{-}obito_3$	$VACINADO = sim, EVOLUCAO = obito\_covid \rightarrow FATOR\_RISCO = sim$	4,81%	81,11%	2,97
$2\text{-}obito_4$	$UTI = sim \rightarrow SUPORTE\_VENTILATORIO = invasivo$	14,44%	44,78%	19,22
$2\text{-}obito_5$	$EVOLUCAO = obito\_covid \rightarrow SUPORTE\_VENTILATORIO = invasivo$	12,8%	41,32%	10,38
$2\text{-}obito_6$	$UTI = sim, EVOLUCAO = obito\_covid \rightarrow SUPORTE\_VENTILATORIO = invasivo$	11,09%	64,73%	23,10

Tabela 5.18: *Óbitos covid - Regras Tipo 2*

Para os agrupamentos do Tipo 6 os mais destacados estavam relacionados a idade e refletem como a necessidade de alguns procedimentos em idosos apresentam um risco maior de óbito, Tabela 5.19. O agrupamento $\{6\text{-}obito_1, 6\text{-}obito_2, 6\text{-}obito_3\}$ mostra como a necessidade de suporte invasivo aumenta associação com óbitos em idosos acima de 75 anos, note como a idade por si só não é um fator preponderante para óbito (Conf. = 51,48% para a regra $6\text{-}obito_2$), porém atua como um reforço importante quando aparece junto de necessidade de suporte ventilatório (regras $6\text{-}obito_3$), pois tanto a confiança quanto o *odds ratio* tem um aumento significativo em relação as regras $6\text{-}obito_1$ e $6\text{-}obito_2$. Algo similar acontece com pessoas em UTI e com idade acima de 75 anos, porém nesse caso tanto a internação em UTI tanto a idade avançada podem ser consideradas não preponderantes para óbito (Conf. = 53,11% para a regra $6\text{-}obito_7$ e Conf. = 51,48% para a regra $6\text{-}obito_8$), porém a combinação dos dois itens resulta em uma associação forte ($6\text{-}obito_9$).

O sintoma de desconforto respiratório pode ser um pouco mais perigoso em pessoas acima de 75 anos (Conf. = 56,81% para $6\text{-}obito_9$ enquanto Conf. = 51,48% para $6\text{-}obito_8$), como mostrado pelo agrupamento $\{6\text{-}obito_{10}, 6\text{-}obito_{11}, 6\text{-}obito_{12}\}$ da Tabela 5.19.

Já o último agrupamento $6\text{-}obito_{10}, 6\text{-}obito_{11}, 6\text{-}obito_{12}$ da Tabela 5.19 mostra que ser um paciente do segundo quadrimestre de 2020 tem um associação com óbito por Covid, regra $6\text{-}obito_{10}$, embora essa associação não seja tão forte (Conf. = 32,98% e *Odds ratio* = 1,12, ambos muito próximos do limiar mínimo definido) não foram geradas esse tipo de associação para outras datas. Além disso, o item $DT\_SINTOMAS =$

(5/2020-8/2020] reforça a associação de idade acima de 75 anos com morte por Covid, regra $6\text{-}obito_{12}$.

Id	Regra	Sup.	Conf.	Odds ratio
$6\text{-}obito_1$	$SUPORTE\_VENT = invasivo \rightarrow EVOLUCAO = obito\_covid$	12,80%	74,49%	10,38
$6\text{-}obito_2$	$IDADE = (75a-) \rightarrow EVOLUCAO = obito\_covid$	10,03%	51,48%	3,02
$6\text{-}obito_3$	$IDADE = (75a-), SUPORTE\_VENT = invasivo \rightarrow EVOLUCAO = obito\_covid$	3,25%	86,9%	16,40
$6\text{-}obito_4$	$UTI = sim \rightarrow EVOLUCAO = obito\_covid$	17,13%	53,11%	4,40
$6\text{-}obito_5$	$IDADE = (75a-) \rightarrow EVOLUCAO = obito\_covid$	10,03%	51,48%	3,02
$6\text{-}obito_6$	$UTI = sim, IDADE = (75a-) \rightarrow EVOLUCAO = obito\_covid$	4,8%	68,81%	5,63
$6\text{-}obito_7$	$DESCONFORTO\_RESP = sim \rightarrow EVOLUCAO = obito\_covid$	19,31%	35,3%	1,57
$6\text{-}obito_8$	$IDADE = (75a-) \rightarrow EVOLUCAO = obito\_covid$	10,03%	51,48%	3,02
$6\text{-}obito_9$	$DESCONFORTO\_RESP = sim, IDADE = (75a-) \rightarrow EVOLUCAO = obito\_covid$	6,06%	56,81%	3,39
$6\text{-}obito_{10}$	$DT\_SINTOMAS = (5/2020\text{-}8/2020] \rightarrow EVOLUCAO = obito\_covid$	5,53%	32,98%	1,12
$6\text{-}obito_{11}$	$IDADE = (75a-) \rightarrow EVOLUCAO = obito\_covid$	10,03%	51,48%	3,02
$6\text{-}obito_{12}$	$DT\_SINTOMAS = (5/2020\text{-}8/2020], IDADE = (75a-) \rightarrow EVOLUCAO = obito\_covid$	2,09%	58,7%	3,32

Tabela 5.19: *Óbitos covid - Regras Tipo 6 que mostram associações entre idade acima de 75 anos e óbitos por Covid*

O grafo gerado para os agrupamentos obtidos para esse estudo de caso foi parecido com os anteriores no formato, porém seu centro foi menos preenchido o que permite uma melhor visualização dos agrupamentos dos Tipos 6 e 7. O grafo de todos os agrupamentos relacionados pode ser visto na Figura 5.6.

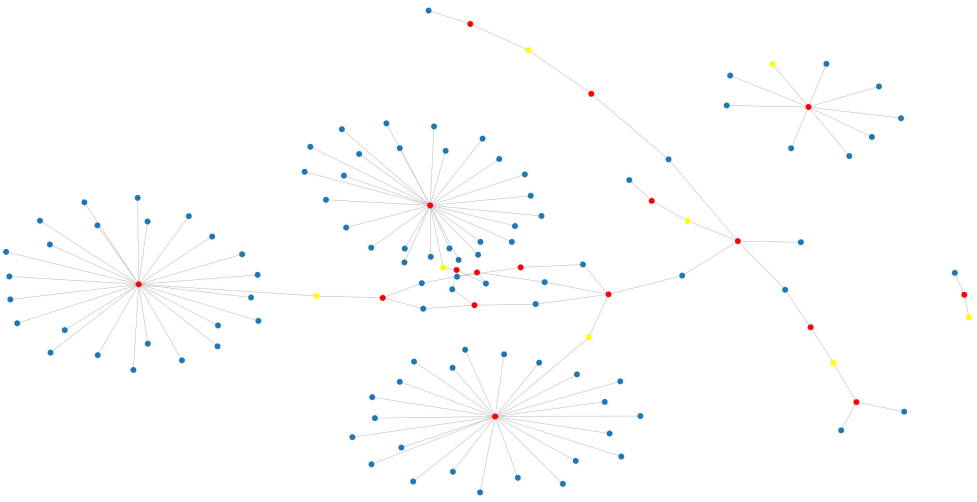

Figura 5.6: *Óbito por Covid - Grafo de todos os agrupamentos relacionados.*

A Figura 5.7 mostra que óbito por Covid está profundamente associado a baixa saturação $SATURACAO = sim$, pois existe uma regra do Tipo 1 (em amarelo). Além da associação com baixa saturação, o subgrafo também mostra diversos sintomas e doenças que se associam a óbito por Covid, tais como dispneia, doenças renais,

doenças neurológicas e pneumopatias. O grafo também mostra que quando esses fatores acontecem em conjunto a associação com óbito é fortalecida.

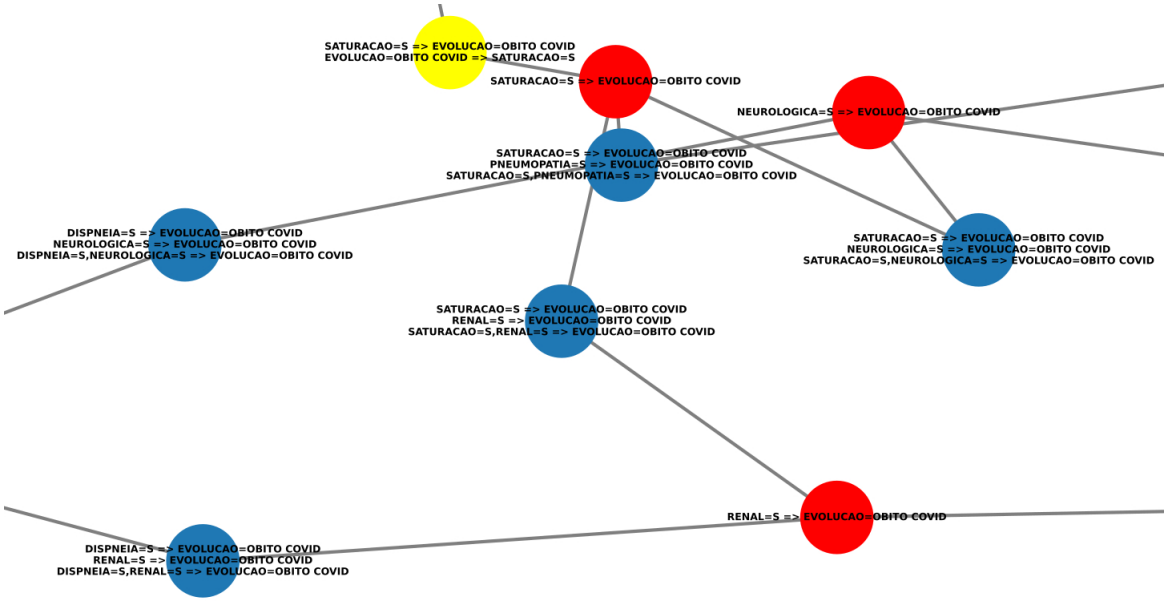

Figura 5.7: *Óbito por Covid - Relação destacada entre agrupamentos.*

5.4 Estudo de Caso 4: Internação em UTI por COVID-19

A mesma base utilizada no estudo de caso 3 foi utilizada para este estudo de caso. Portanto, o mesmo conjunto de regras gerados pelo Apriori foi utilizado como entrada para o método proposto. O método foi aplicado com os mesmo parâmetros, com a única diferença que o item de interesse foi alterado para $UTI = sim$ e gerou 136 agrupamentos que compreende 255 regras distintas. Tais agrupamentos foram analisados pelo autor e alguns foram selecionados para apresentação. Embora agrupamentos dos Tipos 1, 2, 4, 5 e 6 tenham sido gerados apenas agrupamentos do Tipos 1, 2 e 6 foram selecionados para serem apresentados.

O primeiro agrupamento da Tabela 5.20 ($\{1\text{-}uti_1, 1\text{-}uti_2\}$) mostra as mesmas regras que apareceram na seção anterior, a associação entre óbito por covid e uti. O segundo agrupamento ($\{1\text{-}uti_3, 1\text{-}uti_4\}$) mostra que nem todos pacientes que possuem algum fator de risco acabam necessitando de internação na UTI (Conf. = 37,80% para a regra $1\text{-}uti_3$), porém os pacientes em UTI em sua maioria possuem algum fator de risco (Conf. = 70,79% para a regra $1\text{-}uti_4$). Os três últimos agrupamentos ($\{1\text{-}uti_5, 1\text{-}uti_6\}$, $\{1\text{-}uti_7, 1\text{-}uti_8\}$ e $\{1\text{-}uti_9, 1\text{-}uti_{10}\}$) mostram a associação entre sintomas respiratório e internação em UTI. Os sintomas isolados não implicam em internação na UTI pois tem baixa confiança (Conf. $< 40\%$ para as regras $1\text{-}uti_5$), $1\text{-}uti_7$ e $1\text{-}uti_9$, porém os pacientes na

UTI tendem a apresentar os sintomas: desconforto respiratório, baixa saturação e dispnéia (regras *1-uti*$_6$, *1-uti*$_8$ e *1-uti*$_{10}$).

Id	Regra	Sup.	Conf.	Odds ratio
1-uti$_1$	$EVOLUCAO = obito\_covid \rightarrow UTI = sim$	17,13%	55,28%	4,40
1-uti$_2$	$UTI = sim \rightarrow EVOLUCAO = obito\_covid$	17,13%	53,11%	4,40
1-uti$_3$	$FATOR\_RISCO = sim \rightarrow UTI = sim$	22,83%	37,80%	1,95
1-uti$_4$	$UTI = sim \rightarrow FATOR\_RISCO = sim$	22,83%	70,79%	1,95
1-uti$_5$	$DESCONFORTO\_RESP = sim \rightarrow UTI = sim$	19,78%	36,15%	1,49
1-uti$_6$	$UTI = sim \rightarrow DESCONFORTO\_RESP = sim$	19,78%	61,34%	1,49
1-uti$_7$	$BAIXA\_SATURACAO = sim \rightarrow UTI = sim$	22,89%	37,12%	1,83
1-uti$_8$	$UTI = sim \rightarrow BAIXA\_SATURACAO = sim$	22,89%	70,97%	1,83
1-uti$_9$	$DISPNEIA = sim \rightarrow UTI = sim$	24,53%	35,48%	1,65
1-uti$_{10}$	$UTI = sim \rightarrow DISPNEIA = sim$	24,53%	76,06%	1,65

Tabela 5.20: *Covid UTI - Regras Tipo 1*

Para os agrupamentos escolhidos do Tipo 2 os três primeiros agrupamentos ($\{$*2-uti*$_1$, *2-uti*$_2$, *2-uti*$_3\}$, $\{$*2-uti*$_4$, *2-uti*$_5$, *2-uti*$_6\}$ e $\{$*2-uti*$_7$, *2-uti*$_8$, *2-uti*$_9\}$), mostrados na Tabela 5.21, detalham a associação entre UTI e o sintoma de fadiga implicando em outros três sintomas: baixa saturação (*2-uti*$_3$), dispnéia (*2-uti*$_6$ e desconforto respiratório (*2-uti*$_9$). Destes agrupamentos, nos dois primeiros tanto o item de interesse (UTI) quanto o sintomas associado contribuem de forma próxima para a associação (a diferença entre a Conf. de *2-uti*$_1$ e *2-uti*$_2$ é pequena, assim como nas regras *2-uti*$_4$ e *2-uti*$_5$). O terceiro agrupamento o item de interesse UTI atua mais próximo de um reforço (Conf. = 61,34% para *2-uti*$_7$ enquanto Conf. = 72,99% para *2-uti*$_8$). O último agrupamento ($\{$*2-uti*$_{10}$, *2-uti*$_{11}$, *2-uti*$_{10}\}$)mostra que pacientes internados na UTI e com resultado do raio-x infiltrado tem maiores chances de apresentar dispneia.

Id	Regra	Sup.	Conf.	Odds ratio
2-uti$_1$	$UTI = sim \rightarrow BAIXA\_SATURACAO = sim$	22,89%	70,97%	1,83
2-uti$_2$	$FADIGA = sim \rightarrow BAIXA\_SATURACAO = sim$	16,55%	75,77%	2,29
2-uti$_3$	$UTI = sim, FADIGA = sim \rightarrow BAIXA\_SATURACAO = sim$	5,76%	83,19%	3,29
2-uti$_4$	$UTI = sim \rightarrow DISPNEIA = sim$	24,53%	76,06%	1,65
2-uti$_5$	$FADIGA = sim \rightarrow DISPNEIA = sim$	17,61%	80,62%	2,15
2-uti$_6$	$UTI = sim, FADIGA = sim \rightarrow DISPNEIA = sim$	5,96%	86,07%	2,93
2-uti$_7$	$UTI = sim \rightarrow DESCONFORTO\_RESP = sim$	19,78%	61,34%	1,49
2-uti$_8$	$FADIGA = sim \rightarrow DESCONFORTO\_RESP = sim$	15,94%	72,99%	2,75
2-uti$_9$	$UTI = sim, FADIGA = sim \rightarrow DESCONFORTO\_RESP = sim$	5,43%	78,33%	3,21
2-uti$_{10}$	$UTI = sim \rightarrow DISPNEIA = sim$	24,53%	76,06%	1,65
2-uti$_{11}$	$RAIOX = infiltrado \rightarrow DISPNEIA = sim$	9,56%	78,29%	1,71
2-uti$_{12}$	$UTI = sim, RAIOX = infiltrado \rightarrow DISPNEIA = sim$	3,28%	83,35%	2,30

Tabela 5.21: *Covid UTI - Regras Tipo 2*

Um agrupamento do Tipo 6 foi destacado na Tabela 5.22. O agrupamento mostra que obesidade ($OBESIDADE = sim$) e óbito por Covid ($EVOLUCAO = obito\_covid$) estão associados individualmente à internação em UTI ($UTI = sim$), regras *6-uti*$_1$ e *6-uti*$_2$. O agrupamento também mostra que quando os itens ocorrem juntos ($OBESIDADE = sim$

e $EVOLUCAO = obito\_covid$) resulta em uma associação mais forte com $UTI = sim$ (Conf. = 70,84% para $6\text{-}uti_3$ em comparação com Conf. = 47,54% para $6\text{-}uti_1$ e Conf. = 55,28% para $6\text{-}uti_2$).

Id	Regra	Sup.	Conf.	Odds ratio
$6\text{-}uti_1$	$OBESIDADE = sim \rightarrow UTI = sim$	3,7%	47,54%	2,02
$6\text{-}uti_2$	$EVOLUCAO = obito\_covid \rightarrow UTI = sim$	17,13%	55,28%	4,41
$6\text{-}uti_3$	$OBESIDADE = sim, EVOLUCAO = obito\_covid \rightarrow UTI = sim$	2,07%	70,84%	5,39

Tabela 5.22: *Covid UTI - Regras Tipo 6*

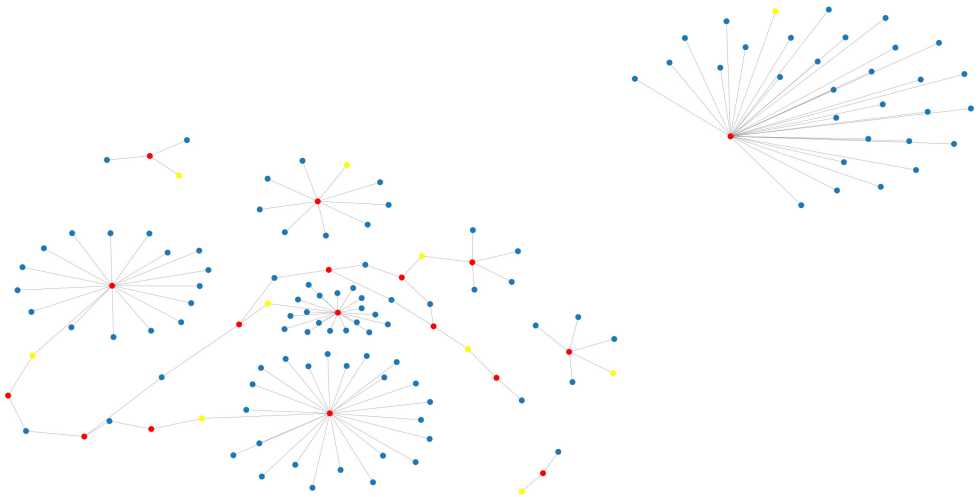

Figura 5.8: *Covid UTI - Grafo de todos os agrupamentos relacionados.*

O grafo gerado para esse estudo de caso foi parecido ao grafo do estudo de caso sobre óbito por Covid. O centro do grafo foi mais esparso possibilitando um interpretação mais facilitada do que nos dois primeiros estudos de caso. A Figura 5.8 mostra o grafo de todos os agrupamentos relacionados.

Para ilustrar isso a Figura 5.9 mostra o subgrafo do centro do grafo da Figura 5.8, é possível notar que no centro do grafo, onde há a maior concentração de nós, não há intersecção entre os rótulos dos nós. A relação dos agrupamentos da Figura 5.9 mostra associações relacionadas à uti e saturação baixa ($SATURACAO = sim$), a regra pivô mais a direita (nó vermelho mais à direita), que incluem diversos itens da base dados. Também mostra que essa associação é bidirecional através do agrupamento ponte (nó em amarelo) e isso proporciona uma ligação com as regras em que o item de interesse ocorre no antecedente, como o pivô mais a esquerda (nó vermelho mais a esquerda) que mostra associações relacionadas a baixa saturação implicando na ocorrência de internação em UTI.

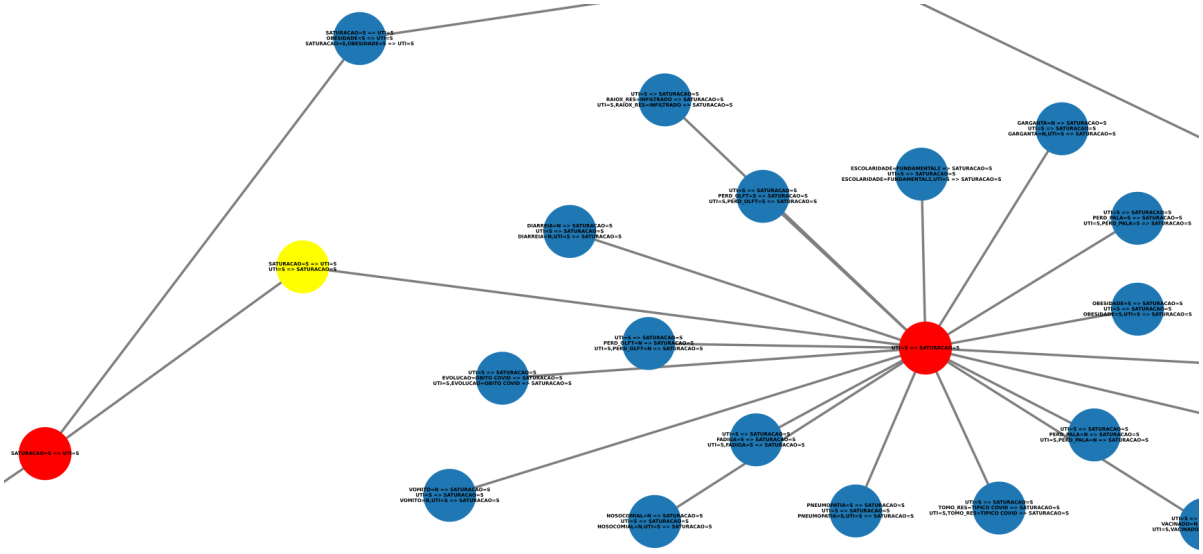

Figura 5.9: *Covid UTI - Grafo de todos os agrupamentos relacionados.*

5.5 Considerações finais

A principal contribuição do método é permitir ao especialista uma visão mais aprofundada das associações por meio do relacionamento entre regras, permitindo entender como o item de interesse se associa com os demais itens da base e se ele atua como reforço a uma associação ou se é reforçado por outro item. Isso pode auxiliar o especialista do domínio a interpretar e tirar conclusões sobre o seu objeto de estudo (item de interesse), podendo agregar novos conhecimentos para a área específica.

Os resultados apresentados nas seções anteriores mostram a aplicabilidade do método em bases de dados reais. Também fornecemos uma nova visão sobre a aplicação de regras de associação ao utilizar o Apriori com valores baixos de confiança para regras de tamanho 2, a maioria dos trabalhos encontrados na literatura priorizam altas confianças. Isso permite ver de maneira clara o reforço em itens cuja a associação direta é menor, mas a associação em conjunto com outro item é confiável.

Uma visualização gráfica foi fornecida a fim de verificar a relação entre agrupamentos. O grafo proposto utiliza regras pivôs, dessa forma, agrupamentos que compartilham uma regra podem ser agrupados formando assim uma visão geral sobre uma determinada associação (a associação da regra pivô). Também foi possível notar a relação entre os agrupamentos dos Tipos 1, 2, 4, 6 e 7 que podem formam um grafo complexo com os agrupamentos do Tipos 1 servindo como ponte entre dois grupos de agrupamentos. Porém, a visualização apresentada não fornece informações sobre as métricas das regras e isso oculta qual item é responsável por apenas reforçar a associação. Além disso, a visualização não contempla os agrupamentos dos Tipos 3, 5 e 8.

Performance quantitativa do método

Em todos os estudos de casos houve um diminuição substancial no número de regras comparando com não realizar nenhum pós-processamento. Embora o método filtre por um item de interesse os agrupamentos dos Tipos 2, 3, 4 e 5 envolvem buscar a presença/ausência de regras de tamanho 2 que não contenham o item de interesse, portanto, uma busca manual resultaria em uma procura sobre todas as regras.

A Figura 5.10 exemplifica como o método reduz o número de regras para análise de maneira significativa. O gráfico está usando um escala logarítmica devido ao intervalo grande do eixo y, portanto, não representa com precisão o tamanho da redução, porém permite verificar que houve um queda substancial no número de regras. Como mostrado pela Figura 5.10 o total de regras geradas (em azul) é diminuído a poucas centenas (em verde) nos estudos de casos 2, 3 e 4; e um pouco mais de 1800 para o estudo de caso 1. Além disso, essas regras são condensadas em uma quantidade menor ainda de agrupamentos (em amarelo). Um hipótese aberta pelo trabalho, é que um agrupamento pode ter um esforço próximo a avaliar um regra individual após uma curva de aprendizado sobre seus significados.

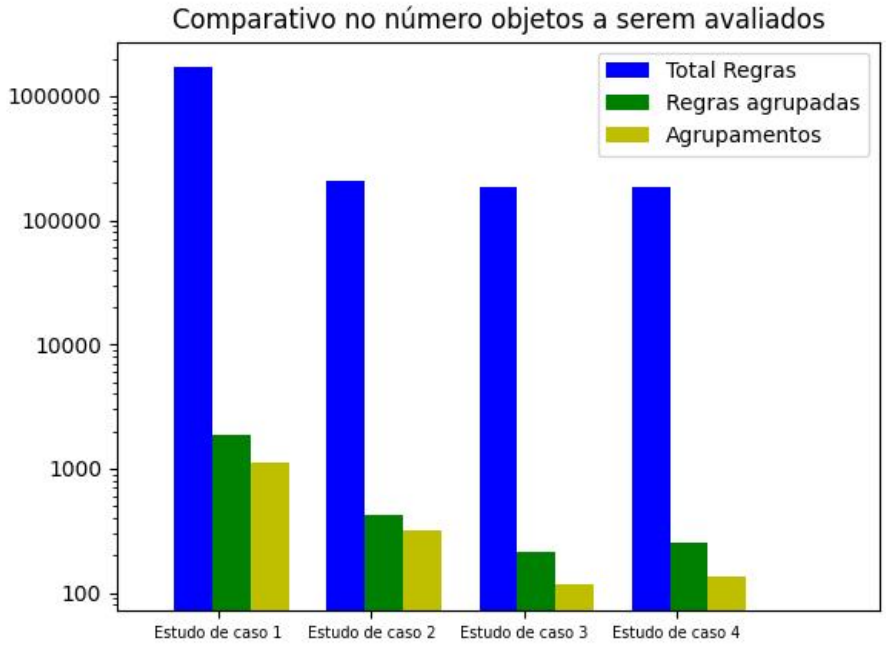

Figura 5.10: *Quantidade de objetos a serem avaliados. O eixo y representa a quantidade de objetos (regras ou agrupamentos) que o especialista precisaria verificar. Em azul o total de regras geradas pelo Apriori sem nenhum tratamento. Já em verde temos o total de regras que existem em algum agrupamento. Por fim, em amarelo a quantidade de agrupamentos gerados.*

CAPÍTULO 6

Conclusões

Dado o problema intrínseco do excesso de regras geradas pelos algoritmos padrões de ARM e a dificuldade de se extrair um conhecimento específico a partir destas regras, este trabalho apresenta um método para selecionar regras que destaquem associações relacionadas a um item de interesse. A revisão sistemática da literatura que foi realizada mostrou que poucos artigos tinham foco em um item de interesse específico e não foram encontrados trabalhos que verificassem a influência do item de interesse nos demais itens (e vice-versa) por meio de regras relacionadas.

O método apresentado mostrou-se ser capaz de organizar as regras de forma a evidenciar a influência de um item de interesse em associações presentes. Isso permite ao especialista, além de ter uma visão mais aprofundada acerca de como o item de interesse interage com os demais itens nas próprias regras onde ele aparece, mas também entender qual a importância do item de interesse nas associações existentes com outras regras. Uma análise gráfica permite condensar a informação presentes nesses agrupamentos de uma maneira mais centralizada, tornando a exploração mais dinâmica.

Os resultados apresentados estão restritos pelo modelo Suporte/Confiança dos métodos de ARM. Como relatado por Freitas [26], regras de associação não se preocupam com problemas de *overfitting* e *underfitting*, portanto seus resultados não podem ser interpretados como uma predição e sim como um descrição sobre a disposição dos itens em uma base de dados. Algoritmos de classificação são enviesados pelo atributo alvo (atributo de classe) a fim de construir um modelo de predição. Além disso, é comum que nos algoritmos de classificação seja esperado que os atributos de entrada sejam independentes ou tenham baixa correlação. Por outro lado, em ARM, o objetivo é encontrar associações/dependências entre os atributos na base de dados, o que diferencia bastante da tarefa de classificação. Por exemplo, experimentos realizados durante este trabalho utilizando bases de dados de problemas de classificação gerou poucos e insignificantes agrupamen-

tos de regras. As bases utilizadas foram *adult*[1], *mushroom*[2] e *zoo*[3], todas disponibilizadas *UC Irvine Machine Learning Repository* (UCI).

6.1 Limitações do estudo

Apesar das bases de dados utilizadas nos estudos de casos serem bases reais, por um questão de disponibilidade, não foi possível avaliar a utilidade das regras com um especialista no domínio. Desta forma, os resultados apresentados estão limitados à compreensão do autor sobre os respectivos temas.

A visualização proposta apresentada possui a limitação de não fornecer informações sobre as métricas das regras encontradas e não contemplar todos os tipos de agrupamentos que o método produz.

Durante a análise dos resultados foram constatadas algumas regras com itens contraditórios, tais como as regras abaixo. Será necessário um estudo mais aprofundado para entender o que leva ao aparecimento dessas regras bem como qual a melhor forma de tratar esta situação.

$$EVOLUCAO = obito\_covid, \textbf{DOR\_ABDOMINAL=sim} => DESCONFORTO\_RESP = sim$$
$$EVOLUCAO = obito\_covid, \textbf{DOR\_ABDOMINAL=nao} => DESCONFORTO\_RESP = sim$$

6.2 Trabalhos futuros

Neste trabalho, três métricas foram testadas, porém existem muitas outras que podem ser utilizadas. Portanto um estudo mais aprofundado aplicando o método utilizando outras métricas pode ser realizado, bem como uma comparação entre os agrupamentos gerados por cada métrica.

Uma melhoria na visualização de forma que o centro do grafo não fique tão sobreposto, além de contemplar os agrupamentos dos tipos 2, 5 e 8 também pode ser feito. A nova visualização também deve mostrar as informações completas que um agrupamento possui, como qual item atua como reforço e qual é reforçado, bem como o valor das métricas utilizadas.

Uma extensão que do método seria permitir uma busca orientada por atributo interesse em vez de item de interesse, por exemplo, seria usado *PMDD* podendo ser *sim* ou *não* no lugar de $PMDD = sim$.

[1] https://archive.ics.uci.edu/dataset/2/adult
[2] https://archive.ics.uci.edu/dataset/73/mushroom
[3] https://archive.ics.uci.edu/dataset/111/zoo

O método proposto manteve algumas regras opostas que não conseguem trazer um conhecimento claro ao especialista. É necessário um estudo mais aprofundado sobre a natureza dessas regras e qual a melhor abordagem para tratar esses casos.

Um estudo comparativo sobre o esforço em interpretar uma regra com o esforço em interpretar um agrupamento pode ajudar a estabelecer um parâmetro para comparação com outros métodos de pós-processamento. A avaliação do resultados produzidos pelo método feito por um especialista pode ajudar na compreensão de como os resultados podem auxiliar os especialistas do domínio na exploração das regras e geração de conhecimento. Ademais, tal avaliação é uma prática melhor para a análise do método, como levantado na análise da literatura.

Somente bases de dados de saúde foram utilizadas como estudos de caso neste trabalho, portanto, estudos em bases de dados em outros domínios podem ser realizados para se verificar a eficiência do método proposto.

Referências Bibliográficas

[1] ABDELHAMID, N.; THABTAH, F. **Associative Classification Approaches: Review and Comparison**. *Journal of Information & Knowledge Management*, 13(03):1450027, sep 2014.

[2] AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. **Mining association rules between sets of items in large databases**. *ACM SIGMOD Record*, 22(2):207–216, jun 1993.

[3] AGRAWAL, R.; SRIKANT, R. **Fast Algorithms for Mining Association Rules in Large Databases**. In: *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, p. 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[4] AIT-MLOUK, A.; GHARNATI, F.; AGOUTI, T. **Multi-criteria decisional approach for extracting relevant association rules**. *International Journal of Computational Science and Engineering*, 15(3/4):188, 2017.

[5] BAESENS, B.; VIAENE, S.; VANTHIENEN, J. **Post-processing of association rules**. *Katholieke Universiteit Leuven, Open Access publications from Katholieke Universiteit Leuven*, 01 2000.

[6] BALASUBRAMANI, B. S.; SHIVAPRABHU, V. R.; KRISHNAMURTHY, S.; CRUZ, I. F.; MALIK, T. **Ontology-based urban data exploration**. In: *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics - UrbanGIS '16*, p. 1–8, New York, New York, USA, oct 2016. ACM Press.

[7] BENHACINE, F. Z.; ATMANI, B.; ABDELOUHAB, F. Z. **Contribution to the association rules visualization for decision support: A combined use between boolean modeling and the colored 2d matrix**. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5:38, 2019.

[8] BERKA, P. **Comprehensive concept description based on association rules: A meta-learning approach**. *Intelligent Data Analysis*, 22(2):325–344, mar 2018.

[9] BOUZIRI, A.; LATIRI, C.; GAUSSIER, E. **LTR-expand: query expansion model based on learning to rank association rules**. *Journal of Intelligent Information Systems*, 55(2):261–286, oct 2020.

[10] CASTRO, G.; SALVINI, R.; SOARES, F. A.; NIERENBERG, A. A.; SACHS, G. S.; LAFER, B.; DIAS, R. S. **Applying Association Rules to Study Bipolar Disorder and Premenstrual Dysphoric Disorder Comorbidity**. In: *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*, p. 1–4. IEEE, may 2018.

[11] CAZER, C. L.; AL-MAMUN, M. A.; KANIYAMATTAM, K.; LOVE, W. J.; BOOTH, J. G.; LANZAS, C.; GRÖHN, Y. T. **Shared Multidrug Resistance Patterns in Chicken-Associated Escherichia coli Identified by Association Rule Mining**. *Frontiers in Microbiology*, 10(APR), apr 2019.

[12] CHENG, C.-W.; SHA, Y.; WANG, M. D. **InterVisAR: An Interactive Visualization for Association Rule Search**. In: *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '16, p. 175–184, New York, NY, USA, oct 2016. ACM.

[13] CHENG, Y.; YU, W.-D.; LI, Q. **GA-based multi-level association rule mining approach for defect analysis in the construction industry**. *Automation in Construction*, 51(C):78–91, mar 2015.

[14] CINTRA, L. F. D. C. **Seleção e agrupamento de regras de associação baseados em fator de interesse**. Undergraduate Thesis, Instituto de Informática, Universidade Federal de Goiás, Goiânia, Goiás. p. 45, 2019.

[15] DA COSTA, N. R.; MANCINE, L.; SALVINI, R.; TEIXEIRA, J. D. M.; RODRIGUEZ, R. D.; LEITE, R. E. P.; NASCIMENTO, C.; PASQUALUCCI, C. A.; NITRINI, R.; JACOB-FILHO, W.; LAFER, B.; GRINBERG, L. T.; SUEMOTO, C. K.; NUNES, P. V. **Microcephaly measurement in adults and its association with clinical variables**. *Revista de Saúde Pública*, 56:38, may 2022.

[16] DAHBI, A.; JABRI, S.; BALOUKI, Y.; GADI, T. **A new method for ranking association rules with multiple criteria based on dominance relation**. In: *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, p. 1–7. IEEE, nov 2016.

[17] DAHBI, A.; JABRI, S.; BALOUKI, Y.; GADI, T. **A New Method to Select the Interesting Association Rules with Multiple Criteria**. *International Journal of Intelligent Engineering and Systems*, 10(5):191–200, oct 2017.

[18] DAHBI, A.; JABRI, S.; BALOUKI, Y.; GADI, T. **Selecting, sorting and ranking association rules with multiple criteria using dominance relation**. *Advances in Mathematics: Scientific Journal*, 9(11):9489–9508, 2020.

[19] DE CARVALHO, V. O.; DE PAULA, D. D.; PACHECO, M. V.; DOS SANTOS, W. E.; DE PADUA, R.; REZENDE, S. O. **Ranking Association Rules by Clustering Through Interestingness**. In: Castro, F.; Miranda-Jiménez, S.; González-Mendoza, M., editors, *Advances in Soft Computing*, p. 336–351, Cham, 2018. Springer International Publishing.

[20] DE PADUA, R.; CARMO, L. P. D.; REZENDE, S. O.; DE CARVALHO, V. O. **An Analysis on Community Detection and Clustering Algorithms on the Post-Processing of Association Rules**. In: *Proceedings of the International Joint Conference on Neural Networks*, volume 2018-July. Institute of Electrical and Electronics Engineers Inc., oct 2018.

[21] DE PADUA, R.; DE CARVALHO, V. O.; REZENDE, S. O. **Post-processing Association Rules: A Network Based Label Propagation Approach**. In: Freivalds, R. M.; Engels, G.; Catania, B., editors, *SOFSEM 2016: Theory and Practice of Computer Science*, p. 580–591, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

[22] DE PAUW, J.; MOENS, S.; GOETHALS, B. **SubSect—An Interactive Itemset Visualization**. *Communications in Computer and Information Science*, 1196:165–181, 2020.

[23] DJENOURI, Y.; BENDJOUDI, A.; DJENOURI, D.; BELHADI, A.; NOUALI-TABOUDJEMAT, N. **New GPU-based swarm intelligence approach for reducing big association rules space**. In: *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, p. 1–6. IEEE, aug 2017.

[24] FAN, C.; XIAO, F.; YAN, C. **A framework for knowledge discovery in massive building automation data and its application in building diagnostics**. *Automation in Construction*, 50(C):81–90, feb 2015.

[25] FISTER, I.; FISTER, D.; IGLESIAS, A.; GALVEZ, A.; OSABA, E.; DEL SER, J.; FISTER, I. **Visualization of Numerical Association Rules by Hill Slopes**. In: *21th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL*, p. 101–111, 2020.

[26] FREITAS, A. A. **Understanding the crucial differences between classification and discovery of association rules**. *ACM SIGKDD Explorations Newsletter*, 2(1):65–69, jun 2000.

[27] GRABOT, B. **Rule mining in maintenance: Analysing large knowledge bases**. *Computers & Industrial Engineering*, 139:105501, jan 2020.

[28] HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. **Exploring network structure, dynamics, and function using networkx**. In: Varoquaux, G.; Vaught, T.; Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, p. 11 – 15, Pasadena, CA USA, 2008.

[29] HAHN, J. **Evaluating systematic transactional data enrichment and reuse**. In: *PervasiveHealth: Pervasive Computing Technologies for Healthcare*. ICST, may 2019.

[30] HAHSLER, M. rdocumentation.org, 2023. **R package arules - mining association rules and frequent itemsets**. Disponível em: <https://www.rdocumentation.org/packages/arules/versions/1.7-6>. Acessado em: 5 de agosto de 2023.

[31] HAHSLER, M.; KARPIENKO, R. **Visualizing association rules in hierarchical groups**. *Journal of Business Economics*, 87(3):317–335, apr 2017.

[32] HAN, J.; PEI, J.; YIN, Y.; MAO, R. **Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach**. *Data Mining and Knowledge Discovery*, 8(1):53–87, jan 2004.

[33] HASHIMOTO, K.; OTAKE, K.; NAMATAME, T. **Proposal of the visualization system of simultaneous purchasing relation using POS data with the ID of the supermarket**. In: *2016 Future Technologies Conference (FTC)*, p. 610–615. IEEE, dec 2016.

[34] IDOUDI, R.; ETTABAA, K. S.; SOLAIMAN, B.; HAMROUNI, K. **Ontology Knowledge Mining Based Association Rules Ranking**. In: *Procedia Computer Science*, volume 96, p. 345–354. Elsevier, jan 2016.

[35] JILONG, H. **Research on association rules data mining based on improved k-means algorithm**. In: *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, p. 610–613. IEEE, 8 2022.

[36] KARIMI-MAJD, A.-M.; MAHOOTCHI, M. **A new data mining methodology for generating new service ideas**. *Information Systems and e-Business Management*, 13(3):421–443, aug 2015.

[37] KIRCHGESSNER, M.; LEROY, V.; AMER-YAHIA, S.; MISHRA, S. **Testing Interestingness Measures in Practice: A Large-Scale Analysis of Buying Patterns**. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, p. 547–556. IEEE, oct 2016.

[38] KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing systematic literature reviews in software engineering**. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 2007.

[39] KUMARA SWAMY, M.; KRISHNA REDDY, P. **Improving Diversity Performance of Association Rule Based Recommender Systems**. In: *Database and Expert Systems Applications*, p. 499–508. Springer, Cham, 2015.

[40] KUMARA SWAMY, M.; KRISHNA REDDY, P.; BHALLA, S. **Association Rule Based Approach to Improve Diversity of Query Recommendations**. In: Benslimane, D.; Damiani, E.; Grosky, W. I.; Hameurlain, A.; Sheth, A.; Wagner, R. R., editors, *Database and Expert Systems Applications*, p. 340–350, Cham, 2017. Springer International Publishing.

[41] KUMARI, D.; RAJNISH, K. **A new approach to find predictor of software fault using association rule mining**. *International Journal of Engineering and Technology*, 7(5):1671–1684, 2015.

[42] KWON, J.-H.; KIM, E.-J. **Accident Prediction Model Using Environmental Sensors for Industrial Internet of Things**. *Sensors and Materials*, 31(2):579, feb 2019.

[43] LEESUTTHIPORNCHAI, P.; PRADUBSUWUN, D. **Association Extraction from Functional Testing Scenarios**. In: *Proceedings of the 2019 International Electronics Communication Conference*, p. 27–31, New York, NY, USA, jul 2019. ACM.

[44] LENCA, P.; MEYER, P.; VAILLANT, B.; LALLICH, S. **On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid**. *European Journal of Operational Research*, 184(2):610–626, jan 2008.

[45] LIAN, S.; GAO, J.; LI, H. **A Method of Mining Association Rules for Geographical Points of Interest**. *ISPRS International Journal of Geo-Information*, 7(4):146, apr 2018.

[46] LIU, B.; HSU, W.; MA, Y. **Integrating Classification and Association Rule Mining**. In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, p. 80–86. AAAI Press, 1998.

[47] MATETIC, M.; BAKARIC, M. B.; SISOVIC, S. **Association rule mining and visualization of introductory programming course activities**. In: *Proceedings of the 16th International Conference on Computer Systems and Technologies - CompSysTech '15*, volume 1008, p. 374–381, New York, New York, USA, jun 2015. ACM Press.

[48] MCGARRY, K. **A survey of interestingness measures for knowledge discovery**. *The Knowledge Engineering Review*, 20(1):39–61, mar 2005.

[49] MIANI, R. G. L.; HRUSCHKA, E. R. **Eliminating redundant and irrelevant association rules in large knowledge bases**. In: *ICEIS 2018 - Proceedings of the 20th International Conference on Enterprise Information Systems*, volume 1, p. 17–28. SciTePress, 2018.

[50] MOAHMMED, S. A.; A., M.; M., E.-S. **Clustering of Association Rules for Big Datasets using Hadoop MapReduce**. *International Journal of Advanced Computer Science and Applications*, 12(3):536–545, 2021.

[51] MOUHIR, M.; BALOUKI, Y.; GADI, T. **Selecting and Filtering Association Rules within a Semantic Technique**. *International Review on Computers and Software (IRECOS)*, 11(6):530–538, jun 2016.

[52] MUKHERJI, A.; LIN, X.; TOTO, E.; BOTAISH, C. R.; WHITEHOUSE, J.; RUNDENSTEINER, E. A.; WARD, M. O. **FIRE: a two-level interactive visualization for deep exploration of association rules**. *International Journal of Data Science and Analytics*, 7(3):201–226, apr 2019.

[53] MUKHOPADHYAY, A.; MAULIK, U.; BANDYOPADHYAY, S.; COELLO, C. A. C. **Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part II**. *IEEE Transactions on Evolutionary Computation*, 18(1):20–35, feb 2014.

[54] NETWORKX DEVELOPERS. networkx.org, 2022. **Networkx - network analysis in python**. Disponível em: <https://networkx.org/documentation/networkx-2.8.8/>. Acessado em: 5 de agosto de 2023.

[55] NOGO, A.; ZUNIC, E.; DONKO, D. **Identification of association rules in orders of distribution companies' clients**. In: *IEEE EUROCON 2019 -18th International Conference on Smart Technologies*, p. 1–4. IEEE, jul 2019.

[56] OPENDATASUS. openDataSUS - Ministério da Saúde do Brasil, 2019. **Srag 2019 - banco de dados de síndrome respiratória aguda grave**. Disponível em: <https://opendatasus.saude.gov.br/dataset/srag-2019>. Acessado em: 11 de maio de 2023.

[57] OPENDATASUS. openDataSUS - Ministério da Saúde do Brasil, 2020. **Srag 2020 - banco de dados de síndrome respiratória aguda grave - incluindo dados da covid-19**. Disponível em: <https://opendatasus.saude.gov.br/dataset/srag-2020>. Acessado em: 27 de abril de 2023.

[58] OPENDATASUS. openDataSUS - Ministério da Saúde do Brasil, 2023. **Srag 2021 a 2023 - banco de dados de síndrome respiratória aguda grave - incluindo dados da covid-19**. Disponível em: <https://opendatasus.saude.gov.br/dataset/srag-2021-a-2023>. Acessado em: 11 de maio de 2023.

[59] OUNIFI, M. S.; AMDOUNI, H.; ELHOUSSINE, R. B.; SLIMANE, H. **New 3D Visualization and Validation Tool for Displaying Association Rules and Their Associated Classifiers**. In: *2016 20th International Conference Information Visualisation (IV)*, volume 2016-Augus, p. 152–158. IEEE, jul 2016.

[60] OZAKI, T. **Evaluation Measures for Frequent Itemsets Based on Distributed Representations**. In: *2018 Sixth International Symposium on Computing and Networking (CANDAR)*, p. 153–159. IEEE, nov 2018.

[61] PEJIC, A.; STANIC MOLCER, P. **Relationship Mining in PISA CBA 2012 Problem Solving Dataset Using Association Rules**. In: *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, p. 000549–000554. IEEE, may 2018.

[62] SHEN, X.; BAO, L.; ZHANG, L. **Research on Visualization Algorithm of Association Rules Based on Concept Lattice**. In: *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies*, p. 22–27, New York, NY, USA, dec 2020. ACM.

[63] SHINN, M. networkx.org, 2016. **Forceatlas2 for python**. Disponível em: <https://github.com/mwshinn/forceatlas2-python/>. Acessado em: 5 de agosto de 2023.

[64] SHUKLA, S.; MOHANTY, B.; KUMAR, A. **A Multi Attribute Value Theory approach to rank association rules for leveraging better business decision making**. *Procedia Computer Science*, 122:1031–1038, jan 2017.

[65] SHUKLA, S.; MOHANTY, B.; KUMAR, A. **A fuzzy approach to prioritise DEA ranked association rules**. *International Journal of Business Intelligence and Data Mining*, 14(1/2):155, 2019.

[66] SLYEPCHENKO, A.; FREY, B. N.; LAFER, B.; NIERENBERG, A. A.; SACHS, G. S.; DIAS, R. S. **Increased illness burden in women with comorbid bipolar and**

premenstrual dysphoric disorder: data from 1 099 women from STEP-BD study. *Acta Psychiatrica Scandinavica*, 136(5):473–482, nov 2017.

[67] SUDARSANAM, N.; KUMAR, N.; SHARMA, A.; RAVINDRAN, B. **Rate of change analysis for interestingness measures**. *Knowledge and Information Systems*, 62(1):239–258, jan 2020.

[68] SUN, C.; YUAN, L.; CAO, S.; XIA, G.; LIU, Y.; WU, X. **Identifying supply-demand mismatches in district heating system based on association rule mining**. *Energy*, 280:128124, 10 2023.

[69] TAN, P.-N.; KUMAR, V.; SRIVASTAVA, J. **Selecting the right interestingness measure for association patterns**. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, KDD '02, p. 32, New York, New York, USA, 2002. ACM Press.

[70] TAN, P.-N.; KUMAR, V.; SRIVASTAVA, J. **Selecting the right objective measure for association analysis**. *Information Systems*, 29:293–313, 6 2004.

[71] TAN, P.-N.; STEINBACH, M.; KARPATNE, A.; KUMAR, V. **Introduction to Data Mining (2nd Edition)**. Pearson, 2nd edition, 2018.

[72] TEW, C.; GIRAUD-CARRIER, C.; TANNER, K.; BURTON, S. **Behavior-based clustering and analysis of interestingness measures for association rule mining**. *Data Mining and Knowledge Discovery*, 28(4):1004–1045, jul 2014.

[73] THAKAR, S.; KALBANDE, D. **A pipeline for business intelligence and data-driven root cause analysis on categorical data**. In: Shakya, S.; Balas, V. E.; Haoxiang, W., editors, *3rd International Conference on Sustainable Expert Systems, ICSES 2022*, p. 389–398. Springer Singapore, 9 2022.

[74] THAKUR, R. S. **Intelligent decision making in medical data using association rules mining and fuzzy analytic hierarchy process**. *International Journal of Recent Technology and Engineering*, 7(6):1813–1819, 2019.

[75] TUFFÉRY, S. **Association analysis**. In: *Data Mining and Statistics for Decision Making*, chapter 10, p. 287–299. John Wiley & Sons, Ltd, 1 edition, 2011.

[76] VERA, J. C. D.; ORTIZ, G. M. N.; MOLINA.; VILA, M. A. C. **Knowledge redundancy approach to reduce size in association rules**. *Informatica*, 44(2):167–181, jun 2020.

[77] WANG, B.; ZHANG, T.; CHANG, Z.; RISTANIEMI, T.; LIU, G. **3D Matrix-Based Visualization System of Association Rules**. In: *2017 IEEE International Conference on Computer and Information Technology (CIT)*, p. 357–362. IEEE, aug 2017.

[78] WANG, X.; WANG, H. **Design and implementation of digital book recommendation platform based on data mining visualization technology**. In: *2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, p. 1–8. IEEE, 4 2023.

[79] WEI, H.; NIU, C.; XIA, B.; DOU, Y.; HU, X. **A refined selection method for project portfolio optimization considering project interactions**. *Expert Systems with Applications*, 142:112952, mar 2020.

[80] WEI, L.; SCOTT, J. **Association rule mining in the US Vaccine Adverse Event Reporting System (VAERS)**. *Pharmacoepidemiology and Drug Safety*, 24(9):922–933, sep 2015.

[81] WEIDNER, D.; ATZMUELLER, M.; SEIPEL, D. **Finding Maximal Non-redundant Association Rules in Tennis Data**. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12057 LNAI, p. 59–78. Springer, sep 2020.

[82] ZAKI, M. **Scalable algorithms for association mining**. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, 2000.

[83] ZHANG, C.; XUE, X.; ZHAO, Y.; ZHANG, X.; LI, T. **An improved association rule mining-based method for revealing operational problems of building heating, ventilation and air conditioning (HVAC) systems**. *Applied Energy*, 253:113492, nov 2019.

[84] ZHANG, C.; ZHAO, Y.; LI, T.; ZHANG, X. **A post mining method for extracting value from massive amounts of building operation data**. *Energy and Buildings*, 223:110096, sep 2020.

[85] ZHANG, C.; ZHAO, Y.; LU, J.; LI, T.; ZHANG, X. **Analytic hierarchy process-based fuzzy post mining method for operation anomaly detection of building energy systems**. *Energy and Buildings*, 252:111426, dec 2021.

[86] ZHANG, C.; ZHAO, Y.; ZHANG, X. **An improved association rule mining-based method for discovering abnormal operation patterns of HVAC systems**. *Energy Procedia*, 158:2701–2706, feb 2019.

APÊNDICE A

Algoritmo Apriori

Este apêndice detalha e explica a implementação do algoritmo *Apriori* [3] para a geração de regras de associação a partir de uma base de dados de transações (formadas por *itemsets*). A seguinte notação será usada na apresentação do algoritmo:

- Tamanho: significa o número de itens em um *itemset*.
- k-*itemset*: é um *itemset* de tamanho k. Os itens são mantidos em ordem lexicográfica no *itemset*. Seja um k-*itemset* $c = \{c[1], c[2], ..., c[k]\}$, onde $c[1] < c[2] < ... < c[k]$, ele pode ser representado por $c[1] \cdot c[2] \cdot, ..., \cdot c[k]$, que significa a concatenação dos itens $c[i]$, $i = 1, ..., k$.
- L_k: é o conjunto de k-*itemsets* frequentes.
- C_k: é o conjunto de k-*itemsets* candidatos. A geração de *itemsets* candidatos é explicada na Subseção A.1.

Algoritmo A.1: $Apriori(D)$

Entrada: Um conjunto de transações D.

Saída: Um conjunto com todos os *itemsets* frequentes.

```
1  L_1 = {1-itemsets frequentes};
2  para k = 2; L_{k-1} ≠ ∅; k++ faça
3      C_k = apriori-gen(L_{k-1});
4      para todo transação t ∈ D faça
5          C_t = subset(C_k, t);
6          para todo candidato cnd ∈ C_t faça
7              cnd.contador++;
8          fim
9      fim
10     L_k = {cnd ∈ C_k | cnd.contador ≥ suporte mínimo}
11 fim
12 Retorne ⋃_k L_k;
```

O Algoritmo A.1 [3] descreve os passos do Apriori. Ele começa gerando um conjunto de *itemsets* frequentes de tamanho 1 (1-*itemsets*) a partir de D. Seguindo, no passo 2, verifica se foi possível gerar o conjunto de $(k-1)$-*itemsets* frequentes, para $k > 2$. O passo 3, é onde gera-se os *itemsets* candidatos, C_k, a partir do conjunto de (k-1)-*itemsets* gerado anteriormente. Depois o algoritmo percorre todas as transações, passo 4, e no passo 5, 6 e 7, ele seleciona os candidatos que estão contidos na transação t, definindo C_t, para cada candidato contido em C_t incrementa-se o contador desse candidato. Após percorrer todas as transações ele verifica quais *itemsets* candidatos são frequentes, passo 10. Por fim, o algoritmo repete o processo a partir do passo 2. A resposta retornada é a união dos L_k.

Na Subseção A.1 é mostrado com detalhe a função usada no passo 3, pois ela evidencia como o *Apriori* constrói os candidatos a partir de combinação dos *itemsets* frequentes e não das transações em si. Depois é dado um exemplo do algoritmo funcionando por completo.

A.1 Geração de Itemsets Candidatos

Algoritmo A.2: $apriori\text{-}gen(L_{k-1})$

Entrada: Um conjunto de (k-1)-*itemsets* frequentes.

Saída: Um conjunto com todos os k-*itemsets* candidatos.

1 **para todo** $p \in L_{k-1}$ **faça**
2 **para todo** $q \in L_{k-1}$ **faça**
3 **se** $p.item_i = q.item_i,\ i = 1,2,...,k-2 \ \wedge\ p.item_{k-1} < q.item_{k-1}$ **então**
4 $C_k = C_k \cup \{p.item_1, ..., p.item_{k-1}, q.item_{k-1}\}$;
5 **fim**
6 **fim**
7 **fim**
8 **para todo** $a \in C_k$ **faça**
9 **para todo** $(k-1) - subset\ b\ de\ a$ **faça**
10 **se** $b \notin L_{k-1}$ **então**
11 $C_k = C_k - a$;
12 **fim**
13 **fim**
14 **fim**
15 Retorne C_k;

Como detalhado no Algoritmo A.2, pode-se dividir o processo em duas partes, cada uma consistindo de um para-todo mais externo. No primeiro laço na linha 2, o algoritmo percorre todos os $(k-1)$-*itemsets* frequentes e adiciona um $(k-1)$-*itemset* frequente, p, a C_k quando encontra outro $(k-1)$-*itemset* frequente, q, onde q se difere de p apenas pelo último item e o último item de q é lexicograficamente maior que o último item de p, esse passo é chamado de passo de junção. Na segunda parte, o passo de poda, ele retira de C_k todos os *itemsets* frequentes que possuem um subconjunto que não esteja em L_{k-1}. Por fim, retorna C_k.

Exemplificando, se possuímos um conjunto $L_2 = \{\{a_1a_2\}, \{a_1a_3\}, \{a_1a_4\}, \{a_2a_4\}, \{a_3a_4\}\}$, ao final do passo de junção o valor de C_k seria $\{\{a_1a_2a_3\}, \{a_1a_2a_4\}, \{a_1a_3a_4\}\}$. Após o passo de poda, $C_k = \{\{a_1a_2a_4\}, \{a_1a_3a_4\}\}$, pois o subconjunto $\{a_2a_3\}$ de $\{a_1a_2a_3\}$ não pertence a L_2.

A.2 Exemplo do *Apriori*

Tabela A.1: *Banco de dados de transação*

Transação	a_1	a_2	a_3	a_4	a_5
1	✓	✓	✗	✗	✓
2	✓	✗	✗	✓	✗
3	✗	✓	✓	✗	✗
4	✓	✗	✗	✓	✗
5	✓	✓	✓	✗	✗

Considere a Tabela A.1 onde a partir dela será construído todos os *itemset* frequentes, com mínimo suporte de $40,00\%$.

Tabela A.2: *1-itemsets*

Itemset	Suporte
a_1	80%
a_2	60%
a_3	40%
a_4	40%
a_5	20%

Tabela A.3: *1-itemsets frequentes*

Itemset	Suporte
a_1	80%
a_2	60%
a_3	40%
a_4	40%

Primeiramente é mostrado o conjunto de 1-*itemsets* e o conjunto de 1-*itemsets* frequentes, nas Tabelas A.2 e A.3 respectivamente. Note que o item a_5 não atende o suporte mínimo, portanto para o próximo passo a_5 não será usado para gerar os *itemsets* de tamanho 2.

As Tabelas A.4 e A.5 mostram nessa ordem o conjunto de 2-*itemsets* e o conjunto de 2-*itemsets* frequentes. Os *itemsets* $\{a_2, a_4\}$ e $\{a_3, a_4\}$ possuem suporte $0,00$, ou seja, nenhuma transação no banco de dados possui esses itens juntos. O *itemset* $\{a_1 a_3\}$ não possui suporte mínimo portanto não é usado no próximo passo.

Tabela A.4: *2-itemsets*

Itemset	Suporte
a_1,a_2	40%
a_1,a_3	20%
a_1,a_4	40%
a_2,a_3	40%
a_2,a_4	0%
a_3,a_4	0%

Tabela A.5: *2-itemsets frequentes*

Itemsets	Suporte
a_1,a_2	40%
a_1,a_4	40%
a_2,a_3	40%

Por fim, a Tabela A.6 mostra o conjunto de 3-*itemsets*. Como o conjunto de 3-*itemsets* frequentes é vazio o programa para. O retorno do algoritmo é o conjunto de 1-*itemsets* frequentes e o conjunto de 2-*itemsets* frequentes.

Tabela A.6: *3-(itemset)*

Itemsets	Suporte
a_1,a_2,a_4	0%
a_2,a_3,a_4	0%

A.3 Geração de Regras

A partir dos *itemsets* frequentes obtidos é possível gerar as regras de associação, para isso tome como base a união dos *itemsets* frequentes cujo $k \geq 2$, no exemplo anterior seria apenas os 2-*itemsets* frequentes mostrado na Tabela A.5.

O processo consiste em permutar itens dentro de cada *itemset* e formar regras cuja confiança seja maior que a confiança mínima estabelecida, denominadas *regras confiáveis*. As regras que atendem suporte e confiança também são denominadas como regras fortes [71].

Para este exemplo, considere a confiança mínima de 60%. A Tabela A.7 mostra todas as regras que poderiam ser geradas, no entanto, algumas dessas regras não atendem o requisito da confiança mínima estabelecida, logo essas regras, chamadas de *regras não confiáveis*, serão descartadas. As regras confiáveis são mostradas na Tabela A.8.

Tabela A.7: *Regras Geradas*

Itemset	Confiança
$a_1 \rightarrow a_2$	50%
$a_2 \rightarrow a_1$	67%
$a_1 \rightarrow a_4$	50%
$a_4 \rightarrow a_1$	100%
$a_2 \rightarrow a_3$	67%
$a_3 \rightarrow a_2$	100%

Tabela A.8: *Regras Confiáveis*

Itemsets	Confiança
$a_2 \rightarrow a_1$	67%
$a_4 \rightarrow a_1$	100%
$a_2 \rightarrow a_3$	67%
$a_3 \rightarrow a_2$	100%

O *tamanho* de uma regra é número de itens que a compõe. É possível gerar regras com o antecedente vazio, ou seja, regras de tamanho 1. Porém, essas regras mostram apenas que um item é frequente, logo não há como atestar a sua confiança.

Além disso, também é possível gerar regras com mais de um item no consequente, optou-se por considerar, no exemplo, apenas regras com um item no consequente, porque a implementação do *Apriori* utilizada no trabalho considera apenas regras com essa configuração.

APÊNDICE B

Strings revisão sistemática

Scopus

```
TITLE-ABS-KEY ( "association rule")
AND (TITLE-ABS-KEY ("post-mining")
      OR TITLE-ABS-KEY ("post-process*")
      OR ((TITLE-ABS-KEY (visuali*)
            OR TITLE-ABS-KEY (cluster*)
            OR TITLE-ABS-KEY (prun*)
            OR TITLE-ABS-KEY (summari*)
            OR TITLE-ABS-KEY (rank*))
              W/1 TITLE-ABS-KEY (rule))
)
AND NOT TITLE("fuzzy association rule")
AND NOT TITLE("class association rule")
AND NOT TITLE("associative classifier")
AND NOT TITLE("associative classification")
```

Alguns operadores foram utilizados para melhorar a busca. O operador *wildcard* (*) foi usado para buscar variações nas palavras-chave. O operador de proximidade (*W/x*) foi utilizado para ajudar a filtrar os resultado, esse operador define que a palavra *rule* deve aparecer junto com as outras palavras-chave e permite haver até 1 palavra entre elas. O limitador foi utilizado devido a grande quantidade de trabalhos retornados, sua escolha foi feita porque durante a leitura notou-se que os trabalhos relacionados ao tema só usavam os tipos de pós-processamento para se referir a tarefa realizada (não usavam *post-processing* ou *post-mining*) geralmente essses termos estavam acompanhados ou acompanhavam as palavras *rule(s)* ou *association rule(s)*. Além disso, visando eliminar estudos que utilizavam tipos de regras de associação que não estavam no escopo desse trabalho foram adicionadas as cláusulas *NOT* no título.

IEEE Digital Library

```
"association rule?"
AND ("post-mining"
      OR "post mining"
      OR "postmining"
      OR "post-process*"
      OR "post process*"
      OR "postprocess*"
            OR ((visuali*
            OR cluster*
            OR prun*
           OR summari*
           OR ranking)
                 NEAR/2 rule?)
) AND NOT "Document Title":"fuzzy association rule?"
AND NOT "Document Title":"class-association rule?"
AND NOT "Document Title":"class association rule?"
AND NOT "Document Title":“associative classifier?”
AND NOT "Document Title":“associative classification”
```

A *string* usada foi quase a mesma da *Scopus*. O *wildcard* também foi utilizado aqui, assim como o operador de proximidade (*NEAR/x*). Um outro operador (a saber, ?) foi utilizado para lidar com o plural, já que o motor de busca não lida automaticamente. Além disso, foi necessário repetir as palavras-chave onde existia hífen (-), pois o motor de busca também não lida automaticamente com uma possível correspondência entre hífen e espaço. Na base da *IEEE* além do título, resumo e palavras-chave, a busca também ocorreu em outros metadados, como o resultado foi gerenciável não foi limitada aos três campos.

ACM Digital Library

```
(Abstract:("association rule"OR "association rules")
      OR Title:("association rule"OR "association rules")
      OR Keyword:("association rule"OR "association rules"))
AND (
      Abstract:(("post-mining"OR "post mining"OR "postmining"
            OR "post-process*"OR "post process*"OR "postprocess*"
```

OR visuali* OR cluster* OR prun* OR summari* OR “ranking”))
OR Title:(("post-mining"OR "post mining"OR "postmining"
OR "post-process*"OR "post process*"OR "postprocess*"
OR visuali* OR cluster* OR prun* OR summari* OR “ranking”))
OR Keyword:(("post-mining"OR "post mining"OR "postmining"
OR "post-process*"OR "post process*"OR "postprocess*"
OR visuali* OR cluster* OR prun* OR summari* OR “ranking”))
) AND Title:(NOT "fuzzy association rule"
AND NOT "fuzzy association rules"
AND NOT "class-association rule"
AND NOT "class-association rules"
AND NOT “associative classifier”
AND NOT “associative classifiers”
AND NOT “associative classification”)

A principal diferença da base da *ACM* em relação as outras é que não há operador de proximidade, porém como isso não levou a uma quantidade ingerenciável de trabalhos não foi um problema. Pela falta do operador de proximidade (ajuda a limitar em bases muito grandes) e pelas outras bases terem retornado uma quantidade significativa de trabalhos, a base da *ACM Digital Library* escolhida foi a *The ACM Full-Text Collection* que é restrita a artigos patrocinados ou publicados pela *ACM*.

APÊNDICE C

Tabela trabalhos selecionados

C = *Clustering*
M = Medida de interesse
P = Poda
R = Ranqueamento
S = Sumarização
V = Visualização

Tabela C.1: *Artigos selecionados*

Trabalho	**Abordagem**						**Atributo de interesse**			
	C	M	P	R	S	V	Ambos	Antecedente	Consequente	Nenhum
[4]				✓						✓
[6]			✓							✓
[7]					✓					✓
[8]					✓				✓	
[9]				✓						✓
[11]			✓			✓				✓
[12]						✓			✓	
[13]			✓							✓
[16]				✓						✓
[17]			✓							✓
[18]	✓		✓	✓						✓
[19]	✓			✓						✓
[20]	✓									✓
[21]			✓	✓						✓
[22]						✓				✓
[23]					✓					✓
[24]			✓							✓
[25]						✓				✓
[27]						✓				✓
[29]						✓				✓
[31]	✓					✓			✓	
[33]	✓					✓				✓

Tabela C.2: *Artigos selecionados (continuação)*

Trabalho	Abordagem						Atributo de interesse			
	C	M	P	R	S	V	Ambos	Antecedente	Consequente	Nenhum
[34]	✓			✓						✓
[35]	✓		✓							✓
[36]	✓					✓				✓
[37]		✓								✓
[39]			✓	✓						✓
[40]				✓						✓
[41]			✓	✓						✓
[42]						✓				✓
[43]				✓						✓
[45]				✓						✓
[47]	✓		✓			✓				✓
[49]			✓							✓
[50]	✓		✓							✓
[51]			✓							✓
[52]	✓			✓		✓				✓
[55]						✓				✓
[59]						✓				✓
[60]				✓						✓
[61]			✓							✓
[62]						✓				✓
[64]				✓						✓
[65]				✓						✓
[67]	✓		✓							✓
[68]		✓								✓
[73]					✓					✓
[74]				✓						✓
[76]			✓							✓
[77]						✓				✓
[78]						✓				✓
[79]				✓						✓
[80]			✓		✓	✓		✓		
[81]			✓							✓
[83]	✓		✓							✓
[84]			✓		✓					✓
[85]			✓	✓						✓
[86]	✓		✓							✓

APÊNDICE D

Base de dados sobre Síndrome Respiratória Aguda Grave

Quatro bases de dados foram utilizadas para gerar uma base sobre a ocorrência de Síndrome Respiratória Aguda Grave (SRAG) entre 2019 e 2022[1] [56] [57] [58]. As bases consistem de sintomas, comorbidades, resultados de exames, dados sociodemográficos, bem como a classificação final do paciente (qual a doença que levou a procura por atendimento) disponibilizadas pelo Ministério da Saúde do Brasil via o site OpenDataSUS. Ao longo dos anos novos atributos foram adicionados as bases de dados, na versão de 2022 a base contava com 173 colunas, onde 40 foram utilizadas no processamento e estão listadas abaixo.

SURTO_SG	*ASMA*	*TOMO_RES*
NOSOCOMIAL	*DIABETES*	*CS_ESCOL_N*
FEBRE	*NEUROLOGIC*	*CLASSI_FIN*
TOSSE	*PNEUMOPATI*	*RAIOX_RES*
GARGANTA	*IMUNODEPRE*	*SUPORT_VEN*
DISPNEIA	*RENAL*	*TP_IDADE*
DIARREIA	*OBESIDADE*	*NU_IDADE_N*
DESC_RESP	*HOSPITAL*	*DT_SIN_PRI*
VOMITO	*UTI*	*CS_SEXO*
SATURACAO	*DOR_ABD*	*EVOLUCAO*
PUERPERA	*FADIGA*	
CARDIOPATI	*PERD_OLFT*	
SIND_DOWN	*PERD_PALA*	
HEMATOLOGI	*FATOR_RISC*	
HEPATICA	*VACINA_COV*	

Os valores de alguns atributos eram numéricos e para facilitar a leitura das regras de associação esses valores foram alterados para campos textuais, utilizando o dicionário de dados disponibilizado nas respectivas páginas de acessos. O mapeamento dos valores foram feitos conforme mostrado a seguir.

[1] fonte: https://opendatasus.saude.gov.br/dataset?tags=SRAG

SURTO_SG
NOSOCOMIAL
FEBRE
TOSSE
GARGANTA
DISPNEIA
DIARREIA
DESC_RESP
VOMITO
SATURACAO
PUERPERA
CARDIOPATI
SIND_DOWN
HEMATOLOGI
HEPATICA
ASMA
DIABETES
NEUROLOGIC
PNEUMOPATI
IMUNODEPRE
RENAL
OBESIDADE
HOSPITAL
UTI
DOR_ABD
FADIGA
PERD_OLFT
PERD_PALA
FATOR_RISC
VACINA_COV

- 1.0 ⇒ *S* (sim)
- 2.0 ⇒ *N* (não)
- 9.0 ⇒ *I* (ignorado)

CS_ESCOL_N:

- 0 ⇒ ANALFABETO
- 1 ⇒ FUNDAMENTAL1
- 2 ⇒ FUNDAMENTAL2
- 3 ⇒ MEDIO
- 4 ⇒ SUPERIOR
- 5 ⇒ NSA (Não se aplica)
- 9 ⇒ IGNORADO

RAIOX_RES:

- 1 ⇒ NORMAL
- 2 ⇒ INFILTRADO
- 3 ⇒ CONSOLIDACAO
- 4 ⇒ MISTO
- 5 ⇒ OUTRO
- 6 ⇒ NAO (Não realizado)
- 9 ⇒ IGNORADO

CLASSI_FIN:

- 1 ⇒ INFLUENZA
- 2 ⇒ OUTRO VIRUS
- 3 ⇒ OUTRO AGENTE
- 4 ⇒ NAO ESPECIFICADO
- 5 ⇒ COVID

TOMO_RES:

- 1.0 ⇒ TIPICO COVID
- 2.0 ⇒ COVID INDETERMINADO
- 3.0 ⇒ ATIPICO COVID
- 4.0 ⇒ SEM PNEUMONIA
- 5.0 ⇒ OUTRO
- 6.0 ⇒ NAO
- 9.0 ⇒ IGNORADO

SUPORT_VEN:

- 1.0 ⇒ INVASIVO
- 2.0 ⇒ NAO INVASIVO
- 3.0 ⇒ NAO (Não realizado)
- 9.0 ⇒ IGNORADO

O campo *VACINA_COV* foi criado nas bases de dados de 2019 e 2020 [56] [57] com valor *NE* (Não existente) para diferenciar os casos onde a vacina não existia dos casos de não preenchimento (*N/A*). Além disso, para nas bases de 2021 e 2022 [58] os valores *N/A* foram substituídos pelo valor I (Ignorado).

Os atributos *TOMO_RES*, *DOR_ABD*, *FADIGA*, *PERD_OLFT* e *PERD_PALA*, não existiam na base de dados de 2019 e foram criados com valores *N/A*. Além disso, a base de 2019 também não possuía o campo *FATOR_RISC*. Para a criação desse campo foram adotados dois critérios de preenchimentos: se pelo menos um atributo que é um fator de risco estivesse como S (Sim) então era atribuído S a *FATOR_RISC*; se todos os fatores de risco estivessem com N (Não) ou *N/A* era atribuído N a *FATOR_RISC*. Nas demais bases (2021 e 2022), para cada fator de risco específico, se *FATOR_RISC* for N e o fator de risco específico estive com I (Ignorado) ou *N/A* foi atribuído N ao fator de risco específico. Por exemplo, para uma linha onde *FATOR_RISC* é N e *DIABETES* é I ou *N/A*, então é atribuído N a *DIABETES*

Os atributos *TP_IDADE* e *NU_IDADE_N* foram utilizados para criar um novo atributo *IDADE*. O atributo *TP_IDADE* representa a medida da idade: 1 para dias, 2

para meses e 3 para anos; enquanto *NU_IDADE_N* é quantidade usando a medida em *TP_IDADE*. A partir de ambos os atributos foi calculado a idade em anos e discretizado nos seguintes intervalos abaixo. Após isso, *TP_IDADE* e *NU_IDADE_N* foram removidos.

- *invalido* se $IDADE < 0$
- $[0,1)$
- $[1,5]$
- $(5,10]$
- $(10,18]$
- $(18,30]$
- $(30,45]$
- $(45,60]$
- $(60,75]$
- $(75,\infty)$

O atributo *DT_SIN_PRI* corresponde a data dos primeiros sintomas e foi discretizado em quadrimestres.

Após as mudanças nos atributos as bases foram unidas em uma única base. Na nova base gerada foram removidos todas as linha cujo atributo *CLASSI_FIN* era *N/A*. Além disso, os atributos *PUERPERA*, *HEMATOLOGI*, *SIND_DOWN*, *HEPATICA* possuíam menos de 1% dos seus valores iguais a S e como o suporte mínimo usados no estudos de casos foi de 1%, esses atributos foram removidos.

Alguns atributos foram renomeados por um questão de legibilidade, o mapeamento segue como mostrado abaixo.

$CLASSI\_FIN \Rightarrow DIAGNOSTICO$
$CS\_SEXO \Rightarrow SEXO$
$CS\_ESCOL\_N \Rightarrow ESCOLARIDADE$
$PNEUMOPATI \Rightarrow PNEUMOPATIA$
$VACINA\_COV \Rightarrow VACINADO$
$DT\_SIN\_PRI \Rightarrow DT\_SINTOMAS$
$CARDIOPATI \Rightarrow CARDIOPATIA$
$NEUROLOGIC \Rightarrow NEUROLOGICA$
$IMUNODEPRE \Rightarrow IMUNODEPRESSAO$

O resultado final foi uma base com 3.389.419 linhas e 35 atributos. As Tabelas D.1 D.2 D.3 D.4 mostram informações sobre os atributos da base final.

Atributo	Descrição	Distribuição
DT_SINTOMAS	Data de 1º sintomas do caso.	(1/2021-4/2021] = 24,79% (5/2021-8/2021] = 18,03% (5/2020-8/2020] = 16,9% (9/2020-12/2020] = 12,66% (1/2022-4/2022] = 7,57% (9/2021-12/2021] = 6,13% (5/2022-8/2022] = 5,03% (1/2020-4/2020] = 4,45% (9/2022-12/2022] = 3,03% (5/2019-8/2019] = 0,72% (1/2019-4/2019] = 0,41% (9/2019-12/2019] = 0,26% (9/2018-12/2018] ≈ 0%
SEXO	Sexo do paciente.	M = 53,75% F = 46,23% I = 0,02%

Tabela D.1: *Atributos da base de dados*

Atributo	Descrição	Distribuição
ESCOLARIDADE	Nível de escolaridade do paciente. Para os níveis fundamental e médio deve ser considerada a última série ou ano concluído.	N/A = 32,68% IGNORADO = 30,42% MEDIO = 9,55% FUNDAMENTAL1 = 9,46% FUNDAMENTAL2 = 6,01% NSA = 4,27% SUPERIOR = 4,19% ANALFABETO = 3,42%
SURTO_SG	Caso é proveniente de surto de Síndrome Gripal?	N/A = 71,66% N = 18,26% S = 6,74% I = 3,34%
NOSOCOMIAL	Caso de SRAG com infecção adquirida após internação.	N = 73,82% N/A = 16,99% I = 7,21% S = 1,98%
FEBRE	Paciente apresentou febre?	S = 53,79% N = 29,09% N/A = 15,84% I = 1,28%
TOSSE	Paciente apresentou tosse?	S = 67,4% N = 19,32% N/A = 12,24% I = 1,05%
GARGANTA	Paciente apresentou dor de garganta?	N = 53,97% N/A = 27,94% S = 15,84% I = 2,25%
DISPNEIA	Paciente apresentou dispneia?	S = 67,61% N = 19,01% N/A = 12,43% I = 0,95%
DESC_RESP	Paciente apresentou desconforto respiratório?	S = 54,74% N = 25,64% N/A = 18,41% I = 1,21%
SATURACAO	Paciente apresentou saturação $O_2 < 95\%$?	S = 58,69% N = 23,55% N/A = 16,39% I = 1,37%
DIARREIA	Paciente apresentou diarreia?	N = 57,47% N/A = 29,18% S = 11,36% I = 2%
VOMITO	Paciente apresentou vômito?	N = 59,07% N/A = 29,9% S = 8,97% I = 2,06%
FATOR_RISC	Paciente apresenta algum fator de risco.	S = 58,53% N = 41,47% N/A $\approx$ 0%
CARDIOPATIA	Paciente possui Doença Cardiovascular Crônica?	N = 58,44% S = 28,74% N/A = 12,31% I = 0,5%
ASMA	Paciente possui Asma?	N = 73,7% N/A = 21,91% S = 3,56% I = 0,83%

Tabela D.2: *Atributos da base de dados (Continuação)*

Atributo	Descrição	Distribuição
DIABETES	Paciente possui Diabetes mellitus?	N = 64,17% S = 19,65% N/A = 15,57% I = 0,6%
NEUROLOGICA	Paciente possui Doença Neurológica?	N = 73,21% N/A = 21,65% S = 4,29% I = 0,84%
PNEUMOPATIA	Paciente possui outra pneumopatia crônica?	N = 73,19% N/A = 21,66% S = 4,29% I = 0,86%
IMUNODEPRESSAO	Paciente possui Imunodeficiência ou Imunodepressão (diminuição da função do sistema imunológico)?	N = 74,11% N/A = 22,26% S = 2,74% I = 0,88%
RENAL	Paciente possui Doença Renal Crônica?	N = 73,73% N/A = 22,01% S = 3,41% I = 0,84%
OBESIDADE	Paciente possui obesidade?	N = 71,57% N/A = 21,19% S = 6,13% I = 1,11%
HOSPITAL	O paciente foi internado?	S = 95,48% N = 2,24% N/A = 2,06% I = 0,22%
UTI	O paciente foi internado em UTI?	N = 56,11% S = 29,37% N/A = 12,59% I = 1,93%
SUPORT_VEN	O paciente fez uso de suporte ventilatório?	NAO INVASIVO = 47,01% NAO = 21,98% INVASIVO = 15,15% N/A = 12,58% IGNORADO = 3,28%
RAIOX_RES	Resultado de Raio X de Tórax.	N/A = 37,73% NAO = 25,65% INFILTRADO = 13,47% IGNORADO = 9,57% OUTRO = 6,14% NORMAL = 3,41% CONSOLIDACAO = 2,02% MISTO = 2%
DIAGNOSTICO	Diagnóstico final do caso. Se tiver resultados divergentes entre as metodologias laboratoriais, priorizar o resultado do RT-PCR.	COVID = 63,92% NAO ESPECIFICADO = 32,66% OUTRO VIRUS = 1,98% INFLUENZA = 1,05% OUTRO AGENTE = 0,39%
EVOLUCAO	Evolução do caso	CURA = 66,43% OBITO COVID = 24,93% N/A = 4,59% IGNORADO = 2,47% OBITO OUTRO = 1,58%
DOR_ABD	Paciente apresentou dor abdominal?	N = 50,95% N/A = 41,87% S = 5,1% I = 2,08%

Tabela D.3: *Atributos da base de dados (Continuação)*

Atributo	Descrição	Distribuição
FADIGA	Paciente apresentou fadiga?	N = 40,35% N/A = 39% S = 18,73% I = 1,92%
PERD_OLFT	Paciente apresentou perda do olfato?	N = 50,15% N/A = 41,45% S = 5,92% I = 2,48%
PERD_PALA	Paciente apresentou perda do paladar?	N = 49,96% N/A = 41,52% S = 6% I = 2,52%
TOMO_RES	Resultado da tomografia.	N/A = 44,95% TIPICO COVID = 24,03% NAO = 18,31% IGNORADO = 5,16% OUTRO = 2,94% COVID INDETERMINADO = 2,41% ATIPICO COVID = 1,76% SEM PNEUMONIA = 0,44%
IDADE	Idade do paciente.	(60a-75a] = 25,01% (45a-60a] = 22,54% (75a-) = 19,9% (30a-45a] = 15,35% (18a-30a] = 5,1% [1a-5a] = 5,01% [0d-1a) = 4,18% (5a-10a] = 1,58% (10a-18a] = 1,32% N/A $\approx$ 0% INVALIDO $\approx$ 0%
VACINADO	Informar se o paciente recebeu vacina COVID-19, após verificar a documentação/caderneta.	NE = 35,78% N = 22,74% S = 20,93% I = 20,56%

Tabela D.4: *Atributos da base de dados (Continuação)*