Redes de regras de associação filtradas e multialvo

Dario Brito Calçada

Tese de Doutorado do Programa de Pós-Graduação em Ciências de
Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Dario Brito Calçada

Redes de regras de associação filtradas e multialvo

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
Março de 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C171r Calçada, Dario Brito
Redes de regras de associação filtradas e
multialvo / Dario Brito Calçada; orientadora Solange
Oliveira Rezende. -- São Carlos, 2019.
199 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2019.

1. Regras de Associação. 2. Redes. 3. Medidas
Objetivas. 4. Geração de Hipóteses. 5. Multialvo. I.
Rezende, Solange Oliveira, orient. II. Título.

Dario Brito Calçada

Filtered and multi-target association rules networks

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
March 2019

*Este trabalho é dedicado à minha família
em todos os planos da vida*

AGRADECIMENTOS

Agradecimentos são sempre feitos de forma simples e direta citando nomes e elementos de uma forma geral que contribuíram com este trabalho. Farei um pouco diferente, pois são nas diferenças que percebemos o quanto a nossa vitória realmente compensou.

Dizer que agradeço a Deus é muito pouco diante de tamanha dívida que possuo com o criador. Graças a Ele, tive e tenho a oportunidade de usufruir de uma reencarnação repleta de aprendizados, sendo a conclusão deste trabalho de grande importância em minha vida. Não tendo realmente outras palavras, simplesmente digo OBRIGADO SENHOR!

Muito mais que uma orientadora, uma amiga e quase uma mãe, acima de tudo alguém que confiou em mim, no meu potencial. Ela me motivou desde o início, mesmo antes de saber que seria minha orientadora, e continua me inspirando a seguir firme, em frente, e a sempre “defender meu cinturão” após o doutorado. Prof^a Solange Rezende, obrigado pela veracidade e pelo norteamento de minha vida acadêmica e pessoal, tornando-me um cientista da computação de verdade e um ser humano melhor.

Agradecer a todos que contribuíram diretamente com os experimentos e estudos. Meu amigo Renan pelo tempo compartilhado em várias conversas e debates nos estudos de Mineração de Regras de Associação. Minhas “protetoras” Ivone, Roberta e Camila que me auxiliaram em tantas oportunidades, não só com orientações e sugestões sobre o trabalho, mas também sobre aspectos de minha vida pessoal.

Agradecer aos demais membros da equipe SOL que tive o prazer de conviver. João Antunes, Ricardo Scheicher, Ricardo Marcacini, Vítor e Dildre, além dos outros que contribuíram bastante comigo, mesmo sem saber. Agradecer a todos do LABIC, na pessoa dos professores Alneu e Gustavo, e de todos seus orientandos, por me acolherem tão bem e me ensinarem que para crescer profissionalmente a gente DEVE compartilhar.

Meu muito obrigado aos professores do Programa de Pós-graduação em Ciências da Computação e Matemática Computacional da USP em São Carlos pela paciência a mim direcionada e pelas disciplinas fundamentais em minha formação, destaco o Professor Cláudio Toledo que se tornou um parceiro de pesquisa e um amigo para o resto da vida. A todos os funcionários da USP que eu tanto “perturbei”, muito obrigado!

Emano gratidão a todos que coordenaram esse programa durante a minha passagem, em especial ao professor Ades que muitas vezes pareceu que estava atendendo a um parente próximo bem querido quando eu a ele me dirigia para tirar dúvidas e principalmente pedir ajuda. Ajuda

que nunca foi negada.

Às vezes agradecemos pela realização de sonhos, principalmente àqueles que nem sabíamos que poderiam ser realizados, por isso agradeço à UESPI e à USP-São Carlos por possibilitarem que eu fizesse um doutorado na melhor Universidade da América Latina, e ainda com nota 7 da CAPES, algo inimaginável para mim.

Agradecimento a todos os meus colegas de DINTER, que mesmo sem a possibilidade de convivemos durante todo o processo do doutorado, tornaram as disciplinas extremamente agradáveis. Torço para que todos possam seguir seus rumos com muita sabedoria e amor.

Um agradecimento especial a meu amigo de BRACIS e de Tai Chi, Dimas, obrigado por em tão pouco tempo ter sido um apoio, você nem imagina o quanto me ajudou. Desejo também tudo de melhor ao meu amigo Anandsing pela ajuda nas correções do inglês, pela parceria no artigo da ONU, mas principalmente por me ensinar que constelação não é apenas de estrelas mas também de sentimentos.

Rendo graças a meus familiares pela educação que recebi no ambiente doméstico, pela força e incentivos incondicionais que traziam energias renovadas para superar cada dificuldade que surgia. Mãe, obrigado pelo seu amor e dedicação. Pai, obrigado pelas lições de perseverança e felicidade, nosso reencontro é uma das certezas que me motiva a seguir firme. Minha irmã Rebeca, obrigado por possibilitar a mim o exercício da fraternidade verdadeira.

Neste momento não consigo encontrar uma melhor forma de agradecer àquela que é muito mais que uma esposa, é uma amiga, companheira, irmã, mulher, enfim, alguém que escolheu vir comigo nesta encarnação a fim de que juntos, possamos caminhar no processo evolutivo e alcançarmos níveis espirituais mais elevados de acordo com o nosso merecimento. Jâina, MEU AMOR, Muito obrigado por sempre estar comigo. NEOQETA!

Quando você acha que sabe de alguma coisa, que entende de um assunto ou que consegue compreender determinadas situações, a vida te ensina que o aprendizado é constante e que sábio é aquele que consegue perceber isso. Quando me tornei pai, pude realmente ter a certeza de que não sei de nada e que em todos os momentos aprendemos, principalmente com os pequenos. Lívia, minha princesa, mesmo tão pequena, foi minha companheira em vários momentos deste processo. Papai te ama muito!

Gratidão a todos que fizeram e fazem parte do Centro Espírita e Obras Sociais Luz da Esperança. Sem a certeza de que vocês me apoiariam, nunca teria conseguido realizar esse feito de terminar um doutorado na USP. Paz e bem a todos!

Agradeço por fim ao mestre Allan Kardec, por ter codificado a doutrina espírita, doutrina esta que me dá a base necessária para superar qualquer obstáculo nos ensinando que problemas não existem, são apenas GRANDES OPORTUNIDADES DE CRESCIMENTO. Parafraseando Chico Xavier, um dos grandes exemplos que temos de ser humano:

*“Agradeço todas as dificuldades que enfrentei; não fosse por elas, eu não teria saído do lugar.
As facilidades nos impedem de caminhar. Mesmo as críticas nos auxiliam muito”.*

Devo muito a todas as pessoas mencionadas aqui, intelectual e emocionalmente.

À todos meu muitíssimo obrigado e que Deus abençoe a cada um! Muita LUZ!

“Não te queixes. Trabalha.

Não te irrites. Silencia.

Não pares. Segue adiante.

Não discutas. Demonstra.

Não condenes. Ampara.

Não critiques. Abençoa.

Fala auxiliando para o bem.

Serve sem reclamar.

Não te percas em palavras vazias.

Cabe ao tempo tudo esclarecer em nome de Deus.”

(Emmanuel, no Livro de Respostas - Francisco Cândido Xavier)

AMOR, GRATIDÃO, RESPEITO e ACEITAÇÃO

RESUMO

CALÇADA, D. B. **Redes de regras de associação filtradas e multialvo**. 2019. 199 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

A descoberta de Regras de Associação é uma tarefa de mineração de dados que procura identificar padrões em *datasets*, permitindo, após a sua interpretação, identificar conhecimento específico acerca do problema em análise. A Mineração de Regras de Associação pode ser usada como uma metodologia para descobrir hipóteses ou teorias candidatas em um domínio do conhecimento. No entanto, o processo de Mineração de Regras de Associação gera um grande número de regras superando a capacidade de exploração do usuário. Esse fato pode tornar o processo de análise inviável, além de afetar negativamente o resultado de alguns algoritmos de extração de conhecimento. Diante disso, várias abordagens foram propostas para guiar o usuário na exploração das Regras de Associação descobertas, em especial com a utilização de estruturas de Rede, que permitem analisar as relações existentes entre as regras. Neste contexto, esse trabalho foi motivado pelo potencial uso de Redes na otimização da identificação do conhecimento, em processos de Mineração de Regras de Associação, formulando abordagens explicáveis. Outra motivação surge da lacuna referente ao uso de Redes em tarefas multialvo inerente de várias aplicações do mundo real. O desenvolvimento deste trabalho teve o intento de avançar as pesquisas da área de Mineração de Regras de Associação com o uso de Redes em relação a métodos de geração de hipóteses validáveis com um ou dois itens objetivo, tanto em relação à interpretabilidade como na expressividade das representações construídas. Um Mapeamento Sistemático da literatura da área foi realizado com a finalidade de conhecer o estado da arte sobre como o uso das Redes pode auxiliar nos processos de Mineração de Regras de Associação. Neste trabalho é proposto e desenvolvido um método de seleção e avaliação das medidas de suporte e confiança mínimos referentes a extração de Regras de Associação com o uso de Medidas de Centralidade de Redes, cuja contribuição principal foi a elaboração de um critério objetivo para extração de Regras de Associação. Foram também propostas, desenvolvidas e validadas duas novas Redes, as Redes de Regras de Associação Filtradas (*Filtered-ARNs*) e as Redes de Regras de Associação Multialvo (MTARNs) que promoveram um impacto positivo na identificação do conhecimento por meio da comprovação matemática da influência entre os elementos de uma Regra de Associação e ampliaram a capacidade de extração do conhecimento em estudos de aplicações multialvo.

Palavras-chave: Regras de Associação, Redes, Medidas Objetivas, Geração de Hipóteses, Multialvo.

ABSTRACT

CALÇADA, D. B. **Filtered and multi-target association rules networks**. 2019. 199 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

The discovery of Association Rules is a data mining task that seeks to identify patterns in datasets, allowing, after its interpretation, to determine specific knowledge about the problem under analysis. Association Rules Mining can be used as a methodology for discovering hypotheses or candidate theories in a knowledge domain. However, the Association Rules Mining process generates a large number of rules that exceed the user's ability to exploit. This fact may make the analysis process impracticable, as well as negatively affect the outcome of some knowledge extraction algorithms. Therefore, several approaches were proposed to guide the user in the exploration of the discovered Association Rules, especially with the use of Network structures, which allow to analyze the relations between the rules. In this context, this work was motivated by the potential use of Networks in the optimization of knowledge identification, in Association Rules Mining processes, formulating explainable approaches. Another motivation arises from the gap regarding the use of Networks in multi-target tasks inherent to several real-world applications. The development of this work was intended to advance the research of the Association Rules Mining with the use of Networks with methods of generating validate hypotheses with one or two target items, both about the interpretability and in the expressiveness of representations built. A Systematic Mapping of the literature of the area was carried out with the purpose of knowing the state of the art on how the use of the Networks can help in the Mining processes of Association Rules. In this work, a method of selection and evaluation of the minimum support and trust measures regarding the extraction of Association Rules with the use of Network Centralization Measures was proposed and developed, whose main contribution was the elaboration of an objective criterion for extraction of Association Rules. Two new networks were also introduced, developed and validated, the Filtered Association Rules Networks (*Filtered-ARNs*) and the Multi-Target Association Rules Networks (MTARNs) that promoted a positive impact on the identification of knowledge through mathematical proof of the influence between the elements of an Association Rule and extended the capacity of knowledge extraction in studies of multi-target applications.

Keywords: Association Rules, Networks, Objective Measures, Hypothesis Generation, Multi-target.

LISTA DE ILUSTRAÇÕES

Figura 1 – Algoritmo <i>Apriori</i>	42
Figura 2 – <i>FP-Tree</i>	44
Figura 3 – Papéis das medidas de interesse no processo de mineração de dados	48
Figura 4 – Representação simplificada de uma Rede	53
Figura 5 – Visualização de uma Rede gerada pelo software <i>Gephi</i>	55
Figura 6 – Duas pequenas Redes. (a) Rede Simples (b) Rede Multiarestas com 1 (uma) multiaresta e 2 (duas) auto-arestas	56
Figura 7 – Matrizes de Adjacência	56
Figura 8 – Hipergrafo - em hipergrafos as ligações são simbolizadas por <i>loops</i> que circulam os vértices	57
Figura 9 – Rede Bipartite	57
Figura 10 – Dois esquemas de uma mesma árvore. Na árvore (a) os vértices estão posicionados conforme uma conveniência e na árvore (b) segue-se a estrutura com nó raiz	58
Figura 11 – Exemplo de ARN com hiper-aresta reversa (e_2)	63
Figura 12 – ARN com nó objetivo D. A hiper-aresta e_2 faz parte do ARN. Isso prova a natureza adaptativa da poda local.	63
Figura 13 – ARN de análise de cesta de compras (<i>Market Basket</i>) com pão (<i>bagel</i>) como item objetivo	65
Figura 14 – Distribuição dos trabalhos por base de busca	69
Figura 15 – Distribuição dos trabalhos por ano de publicação	69
Figura 16 – Distribuição dos artigos selecionados pelo local de publicação	72
Figura 17 – Disponibilidade de Base de Dados nos artigos estudados	73
Figura 18 – Execução de validação estatística na pesquisa	74
Figura 19 – Descrição do método proposto aplicado na pesquisa	75
Figura 20 – Realização de validação objetiva	75
Figura 21 – Trabalhos que realizam comparações metodológicas com pesquisas anteriores	76
Figura 22 – Mapa de resultados da combinação das questões de pesquisa	77
Figura 23 – Medidas <i>Added Value</i> (ver Equação 2.3) de arestas conectadas ao item objetivo “[lenses]=no” da ARN	86
Figura 24 – Relação entre Medidas de Rede e Suporte Mínimo com $minconf = 0,3$ (<i>Lenses dataset</i>)	90

Figura 25 – Relação entre Medidas de Rede e Confiança Mínima com <i>minsup</i> = 0,0 (<i>Lenses dataset</i>)	91
Figura 26 – Relação entre Medidas de Rede e Confiança Mínima com <i>minsup</i> = 0,0 (<i>Hayes-roth dataset</i>)	92
Figura 27 – Relação entre Medidas de Rede e Confiança Mínima com <i>minsup</i> = 0,1 (<i>Soybean Large dataset</i>)	94
Figura 28 – Médias de precipitação pluviométrica em Parnaíba/PI(2015)	96
Figura 29 – Relação entre Medidas de Rede e Confiança Mínima com <i>minsup</i> = 0,1 (<i>Green Manure dataset</i>)	100
Figura 30 – Níveis dos nós de uma <i>Filtered-ARN</i>	107
Figura 31 – <i>Filtered-ARN</i> com “[lenses]=hard” como item alvo	110
Figura 32 – <i>ARN</i> com “[lenses]=hard” como item alvo	111
Figura 33 – Árvore de Decisão para o <i>dataset Lenses</i>	111
Figura 34 – <i>Filtered-ARN</i> com “[class]=3” como item alvo e nós de nível 1	112
Figura 35 – <i>ARN</i> com “[class]=3” como item alvo e nós de nível 1 (vermelho)	113
Figura 36 – Árvore de Decisão para o <i>dataset Hayes-roth</i>	113
Figura 37 – <i>Filtered-ARN</i> com “[class]=rhizoctonia” como item alvo e nós de nível 1 (laranja)	114
Figura 38 – <i>ARN</i> com “[class]=rhizoctonia” como item alvo e nós de nível 1 (laranja)	115
Figura 39 – <i>Filtered-ARN</i> com “[HalfLife]=6” como item objetivo e nós de nível 1 (azul)	117
Figura 40 – <i>ARN</i> com “[HalfLife]=6” como item objetivo e nós de nível 1 (azul)	118
Figura 41 – Árvore de decisão construída com o algoritmo J48 e o <i>dataset</i> de adubos orgânicos	119
Figura 42 – <i>ARN</i> com “Stability = 1.0”(true) como item alvo e nós de nível 1	123
Figura 43 – Nível 1 sem predecessores de <i>ARN</i> com “IsPlayable = 1.0” (true) como item alvo e nós de nível 1	124
Figura 44 – <i>Filtered-ARN</i> com “Stability = 1.0”(true) como item alvo e nós de nível 1	126
Figura 45 – Nível 1 sem predecessores de <i>Filtered-ARN</i> com “IsPlayable = 1.0”(true) como item alvo e nós de nível 1	127
Figura 46 – Acurácias do <i>Angry Birds</i> para a abordagem com Algoritmo Genético	128
Figura 47 – <i>ARN</i> com “[BI-RADS] = 5” como item alvo e nós de nível 1 (vermelho)	131
Figura 48 – <i>Filtered-ARN</i> com “[BI-RADS] = 5” como item alvo e nós de nível 1 (vermelho)	132
Figura 49 – Acurácia obtida em cada etapa dos experimentos com o dataset de mamografias	132
Figura 50 – Exemplo de uma <i>MTARN</i> com “B,C” como itens-alvo	138
Figura 51 – <i>ARN</i> com “[class]=iris-virginica” como item alvo	141
Figura 52 – <i>ARN</i> com “[petal-width]=12” como item alvo	142
Figura 53 – <i>Filtered-ARN</i> com “[class]=iris-virginica” como item alvo	143
Figura 54 – <i>Filtered-ARN</i> com “[petal-width]=12” como item alvo	143
Figura 55 – <i>MTARN</i> com “[petal-width]=12.[class]=iris-virginica” como item alvo	144

Figura 56 – ARN com “[class]=3” como item alvo	145
Figura 57 – ARN com “[marital_status]=3” como item alvo	146
Figura 58 – <i>Filtered</i> -ARN com “[class]=3” como item alvo	146
Figura 59 – <i>Filtered</i> -ARN com “[marital_status]=3” como item alvo	147
Figura 60 – MTARN com “[marital_status]=3.[class]=3” como item alvo	147
Figura 61 – ARN com “[class]=rhizoctonia” como item alvo	148
Figura 62 – ARN com “[fruit-pods]=3” como item alvo	148
Figura 63 – <i>Filtered</i> -ARN com “[class]=rhizoctonia” como item alvo	149
Figura 64 – <i>Filtered</i> -ARN com “[fruit-pods]=3” como item alvo	149
Figura 65 – Nós de nível L = 1 da MTARN com “[class]=rhizoctonia.[fruit-pods]=3” como item alvo	150
Figura 66 – Metodologia aplicada para a realização da Prova de Conceito	151
Figura 67 – Destaque de <i>Filtered</i> -ARN com alvo “[Obito]=Sim” e nós de nível um (Nível = 1) e dois (Nível = 2)	154
Figura 68 – Destaque de <i>Filtered</i> -ARN com alvo “[Complicacoes]=Sim” e nós de nível um (Nível = 1)	154
Figura 69 – Destaque de MTARN com alvo “[Complicacoes]=Sim.[Obito]=Sim” e nós de nível um (Nível = 1)	155

LISTA DE TABELAS

Tabela 1 – Compras diárias	40
Tabela 2 – Construção da <i>String</i> de Pesquisa	68
Tabela 3 – Taxa de Complexidade	88
Tabela 4 – Relevância x Suporte Mínimo (<i>Lenses dataset</i>)	89
Tabela 5 – Relevância x Confiança Mínima (<i>Lenses dataset</i>)	90
Tabela 6 – Relevância x Confiança Mínima (<i>Haye-roth dataset</i>)	92
Tabela 7 – Relevância x Confiança Mínima (<i>Soybean Large dataset</i>)	93
Tabela 8 – Espaçamento e densidade de sementeira de plantas utilizadas como adubos verdes	97
Tabela 9 – Distribuição dos parâmetros em 6 classes	98
Tabela 10 – Categorização dos tempos de meia vida para coleções de 30, 60, 90 e 120 dias.	99
Tabela 11 – Relevância x Confiança Mínima (<i>Green Manure dataset</i>)	100
Tabela 12 – Atributos do <i>dataset</i> “Stability”	121
Tabela 13 – Atributos do <i>dataset</i> “Viability”	121
Tabela 14 – Conhecimento extraído com uso da Árvore de Decisão	122
Tabela 15 – Conhecimento extraído com uso da ARN	125
Tabela 16 – Conhecimento extraído com uso da <i>Filtered-ARN</i>	127
Tabela 17 – Quantificação de forma e margem.	129
Tabela 18 – Variáveis - <i>Coronary dataset</i>	152
Tabela 19 – Atributos - <i>Green Manure dataset</i>	186
Tabela 20 – Distribuição das Classes por Categoria - <i>Green Manure dataset</i>	187
Tabela 21 – Atributos do <i>dataset</i> “Stability”	187
Tabela 22 – Atributos do <i>dataset</i> “Viability”	188
Tabela 23 – Distribuição das Classes por Categoria - <i>Stability dataset</i>	188
Tabela 24 – Distribuição das Classes por Categoria - <i>Stability dataset</i>	189
Tabela 25 – Quantificação de forma e margem.	189
Tabela 26 – Distribuição das Classes por Categoria - <i>Mammography dataset</i>	190
Tabela 27 – Atributos - <i>Coronary dataset</i>	191
Tabela 28 – Distribuição das Classes por Categoria - <i>Coronary dataset</i>	192
Tabela 29 – Taxa de Complexidade	194
Tabela 30 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> IRIS	194
Tabela 31 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> LABOR	195
Tabela 32 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> ABALONE	195

Tabela 33 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> LENSES	196
Tabela 34 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> SOYBEAN	196
Tabela 35 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> LUNG CANCER	197
Tabela 36 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> HAYES-ROTH	197
Tabela 37 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> GLASS IDENTIFI- CATION	197
Tabela 38 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> TEACHING AS- SISTANT	198
Tabela 39 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> STATLOG	199
Tabela 40 – Resultados de Acurácia e Tempo Computacional <i>dataset</i> VOTE	199

SUMÁRIO

1	INTRODUÇÃO	27
1.1	Contextualização	27
1.2	Motivação e Lacunas	30
1.3	Questões de Pesquisa	32
1.4	Principais Resultados	33
1.5	Organização do Texto	35
2	FUNDAMENTOS E MAPEAMENTO SISTEMÁTICO	37
2.1	Considerações Iniciais	37
2.2	Regras de Associação	38
2.2.1	<i>Extração das Regras de Associação</i>	40
2.2.1.1	<i>Apriori</i>	41
2.2.1.2	<i>Apriori-TID</i>	43
2.2.1.3	<i>FP-Growth</i>	43
2.2.1.4	<i>Outros Algoritmos</i>	45
2.2.2	<i>Critérios para Seleção e Classificação das Regras de Associação</i>	46
2.2.3	<i>Pós-processamento das Regras de Associação</i>	52
2.2.3.1	<i>Avaliação por consulta</i>	52
2.2.3.2	<i>Poda de regras</i>	52
2.2.3.3	<i>Técnicas de visualização</i>	53
2.3	Redes	53
2.3.1	<i>Arquiteturas das Redes (representação)</i>	54
2.3.2	<i>Medidas de Centralidade</i>	57
2.4	Redes de Regras de Associação	60
2.5	Mapeamento Sistemático - Uso de Redes Aliado a Regras de Associação	64
2.5.1	<i>Questões de Pesquisa de Mapeamento Sistemático</i>	66
2.5.2	<i>Estratégia e String de Pesquisa</i>	67
2.5.3	<i>Processo de Busca</i>	68
2.5.4	<i>Critérios de Seleção de Estudos</i>	69
2.5.5	<i>Processo de Seleção dos Estudos</i>	70
2.5.6	<i>Processo de Extração dos Dados</i>	70
2.5.7	<i>Resultados</i>	71

2.5.8	<i>Ameaça à Validade do Mapeamento Sistemático</i>	81
2.6	Considerações Finais	81
3	RELAÇÃO ENTRE MEDIDAS DE CENTRALIDADE E MEDIDAS DE REGRAS DE ASSOCIAÇÃO	83
3.1	Considerações Iniciais	83
3.2	Modelo Proposto para Cálculo da Relevância de Redes de Regras de Associação	84
3.3	Estudo das Medidas de Rede Associados a Suporte e Confiança Mínimos - Avaliação Experimental	87
3.3.1	<i>Lenses dataset</i>	89
3.3.2	<i>Hayes-roth dataset</i>	91
3.3.3	<i>Soybean Large dataset</i>	92
3.4	Método Proposto para Avaliação de <i>minsup</i> e <i>minconf</i> com o uso de Medidas de Centralidade	94
3.5	Prova de Conceito	95
3.5.1	<i>Geração de dataset</i>	96
3.5.2	<i>Pré-processamento</i>	97
3.5.3	<i>Avaliação de minsup e minconf ideais</i>	99
3.6	Considerações Finais	101
4	REDES DE REGRAS DE ASSOCIAÇÃO FILTRADAS	103
4.1	Considerações Iniciais	103
4.2	Modelo Proposto para Geração de Hipóteses de Maior Confiabilidade	105
4.3	Avaliação Experimental	108
4.3.1	<i>Lenses dataset</i>	109
4.3.2	<i>Hayes-roth dataset</i>	111
4.3.3	<i>Soybean Large dataset</i>	113
4.4	Provas de Conceito	115
4.4.1	<i>Green Manure dataset</i>	116
4.4.1.1	<i>Filtered-ARN</i>	116
4.4.1.2	<i>ARN</i>	117
4.4.1.3	<i>Árvore de Decisão</i>	119
4.4.2	<i>Angry Birds dataset</i>	119
4.4.2.1	<i>Árvore de Decisão</i>	121
4.4.2.2	<i>ARN</i>	122
4.4.2.3	<i>Filtered-ARN</i>	125
4.4.3	<i>Mammography dataset</i>	128
4.4.3.1	<i>Árvore de Decisão</i>	130
4.4.3.2	<i>ARN</i>	130

4.4.3.3	<i>Filtered-ARN</i>	131
4.5	Considerações Finais	132
5	REDES DE REGRAS DE ASSOCIAÇÃO MULTIALVO	135
5.1	Considerações Iniciais	135
5.2	Modelo Proposto para Geração de Hipóteses em Problemas Multialvo	136
5.3	Avaliação Experimental	138
5.3.1	<i>Iris dataset</i>	140
5.3.2	<i>Hayes-roth dataset</i>	143
5.3.3	<i>Soybean Large dataset</i>	146
5.4	Prova de Conceito	150
5.4.1	<i>Coronary Dataset</i>	151
5.4.2	<i>Resultados e Discussão</i>	153
5.5	Considerações Finais	156
6	CONCLUSÕES	159
6.1	Contribuições Científicas	159
6.2	Publicações	164
6.3	Limitações e Trabalhos Futuros	167
REFERÊNCIAS		169
APÊNDICE A REPOSITÓRIO DE <i>DATASETS</i>		185
A.1	<i>Green Manure dataset</i>	185
A.2	<i>Angry Birds datasets</i>	187
A.2.1	<i>Stability</i>	188
A.2.2	<i>Viability</i>	188
A.3	<i>Mammography dataset</i>	189
A.4	<i>Coronary dataset</i>	190
APÊNDICE B TAXA DE COMPLEXIDADE DE <i>DATASETS</i> DO RE- POSITÓRIO UCI		193

INTRODUÇÃO

1.1 Contextualização

O objetivo da Mineração de Dados é identificar conhecimento para prever comportamentos futuros ou compreender o passado (FOURNIER-VIGER *et al.*, 2017). As técnicas usadas para prever o futuro, como as Redes Neurais Artificiais, geralmente são projetadas para se comportar como caixas-pretas, porque o objetivo é preferencialmente obter um modelo que seja o mais preciso possível, ao invés de obter um modelo que seja explicativo. Além disso, várias técnicas de Mineração de Dados visam descobrir padrões que sejam compreensíveis pelos seres humanos e podem ser classificadas pelos tipos de padrões que são por elas encontrados (AGGARWAL, 2015), tais como: *clusters*, conjuntos de itens, tendências e *outliers*. Um método de identificação de conhecimento, utilizado em pesquisas de exploração de dados, pode ser implementado por meio da elaboração de hipóteses sobre um domínio de aplicação a partir da observação de padrões descobertos em *datasets* com o uso de Regras de Associação (NAMAKI *et al.*, 2017).

Tendo em vista a necessidade de exploração da informação, as técnicas de Mineração de Regras de Associação (ARM - do inglês *Association Rules Mining*) tornam-se essenciais para apoio à extração de conhecimento de conjuntos de dados. Visando à identificação de padrões em *datasets*, o processo de Mineração de Regras de Associação pode ser visto como uma sequência de etapas genéricas, que devem ser instanciadas de acordo com os dados disponíveis e o conhecimento que se espera obter (ALTAF; SHAHBAZ; GUERGACHI, 2017; REZENDE *et al.*, 2003). Por exemplo, uma das aplicações mais comuns de ARM é a análise de cesta de compras. A exploração da cesta de compras ou *market basket* proporciona a descoberta de padrões de compras do consumidor a partir de bases de dados transacionais (VALLE; RUZ; MORRÁS, 2018), nos quais os dados estão distribuídos ao longo de milhares ou milhões de transações. Nesta tarefa, as Regras de Associação desempenham um papel importante para encontrar itens frequentes e relacionamentos de compras entre diferentes produtos nos quais

declarações explícitas do tipo “se o item X é comprado, então o item Y também é comprado” são estabelecidas (AGRAWAL; IMIELINSKI; SWAMI, 1994).

Embora processos de ARM gerem padrões com itens frequentes, o número de regras extraídas geralmente supera a capacidade de exploração do usuário. Várias abordagens foram propostas para guiar o usuário na exploração de regras (HAHSLER; KARPIENKO, 2017; PADUA; CARVALHO; REZENDE, 2016). A extração de um número elevado de regras pode, por exemplo, tornar o processo de análise muito custoso computacionalmente, ou até mesmo inviável, além de afetar negativamente o resultado de alguns algoritmos utilizados para extração de conhecimento. Com o intuito de minimizar esses problemas, várias pesquisas buscam abordagens que realizam a seleção das regras por meio de estratégias de poda, uso de Medidas Objetivas para seleção e estruturas de Rede para extração das regras (CAI *et al.*, 2017; ZHENWEI; LINGYUN; LIZHU, 2016; RAWAT *et al.*, 2016; BELYI *et al.*, 2016; PADUA; REZENDE; CARVALHO, 2014; RODRIGUES; GAMA; FERREIRA, 2012). No entanto, a grande maioria dessas abordagens se concentra na exploração com o uso de todas as regras geradas e não considerando explorações que possam se concentrar em um pequeno conjunto de itens (*itemsets*) ou em um item alvo (variável de estudo).

Em geral, o conhecimento descoberto por intermédio de processos de Mineração de Regras de Associação, em um paradigma simbólico, é expresso na forma de padrões. A principal função das técnicas de mineração é a descoberta de quais padrões são mais frequentes e interessantes. Existem vários algoritmos para esse fim, tais como: *Apriori* e *FP-Growth*, no entanto, eles são utilizados principalmente em sistemas de computação centralizados e avaliados em pequenas bases de dados sem uma orientação sobre a metodologia de extração do conhecimento. O conhecimento pode ser extraído por meio da elaboração de hipóteses candidatas. Hipóteses são proposições elaboradas a partir da visualização das Regras de Associação e da detecção de padrões nessas regras (SIMARD; ST-PIERRE; BISKRI, 2016). Usman e Usman (2016) revelam o fato de que a maioria das abordagens dependem do conhecimento do domínio e/ou requerem tarefas de descoberta manuais e dependentes do usuário, além de não fornecerem nenhuma assistência gráfica a analistas para exploração das regras descobertas. Para preencher esta lacuna, é proposta nesta tese abordagens visuais que auxiliam na extração do conhecimento sem a necessidade de um domínio da área que está sendo estudada.

As Regras de Associação representam combinações de itens que ocorrem com determinada frequência em uma base de dados. Uma estratégia para a Mineração de Regras de Associação (ARM) em aplicações de transações de compra poderia gerar o seguinte exemplo: $\{\text{feijão}, \text{couve}\} \Rightarrow \{\text{linguiça}\}$. Esta regra é utilizada para gerar a hipótese de que “os clientes que compram os produtos feijão e couve, tendem também a comprar linguiça”. O exemplo ilustra umas das características mais atrativas de uma Regra de Associação, pois ela é expressa em uma forma muito fácil de ser compreendida quando formuladas por *itemsets* de tamanho reduzido (WENG, 2016).

O formato de uma Regra de Associação pode ser representado como uma implicação $LHS \Rightarrow RHS$, em que LHS e RHS são, respectivamente, o lado esquerdo (LHS - do inglês *Left Hand Side*) ou antecedente, e o lado direito (RHS - do inglês *Right Hand Side*) da regra ou consequente, definidos por conjuntos disjuntos de itens.

As regras geradas pela ARM não possuem um atributo alvo como característica, podendo qualquer atributo estar presente na regra. Atributo alvo é uma variável do conjunto de dados selecionada para estudo. Outro aspecto a ser considerado é que as Regras de Associação retratam situações diferentes umas das outras (MELANDA, 2004; REZENDE *et al.*, 2003). Essa condição determina que a avaliação e interpretação do conhecimento extraído por meio do processo de Mineração de Regras de Associação sejam tratados de modo diferenciado.

A quantidade de Regras de Associação extraídas está diretamente conectada ao número de itens que formam a base de dados. Se um conjunto de dados possui uma quantidade elevada de elementos, o conjunto de Regras de Associação geradas torna-se cada vez maior, inviabilizando a observação das regras em sua totalidade. Além de ter um custo computacional alto, a extração das regras não apresenta nenhuma garantia de que conhecimentos úteis serão extraídos. O seguinte exemplo ilustra a geração de uma quantidade elevada de Regras de Associação. Considere um conjunto A formado por apenas 100 elementos ($n(A) = 100$), pode-se então afirmar que:

- A quantidade de *itemsets* unitários, formulados com elementos de A , é igual a 100.
- A quantidade de *itemsets* com dois elementos, que estão contidos em A , é $100 * 99 = 9.900$.
- A quantidade de Regras de Associação possíveis, formuladas apenas com os *itemsets* unitários, é de 9.900.
- A quantidade de Regras de Associação possíveis, formuladas apenas com os *itemsets* com dois elementos é de $9.900 * 9.899 = 98.000.100$.
- Se o número de elementos dos *itemsets* for ampliado, a quantidade de possibilidades de associação aumenta exponencialmente.
- Surge então a necessidade de seleção das Regras de Associação, geralmente utilizando o conhecimento do domínio analisado.

Neste exemplo simples é possível notar duas limitações no processo adotado. 1) Quanto maior o *dataset*, a quantidade de Regras de Associação extraídas torna-se tão grande que fica impossível a visualização dos padrões e, conseqüentemente, a extração do conhecimento, e 2) A escolha dos elementos que podem formar regras interessantes é feita de modo subjetivo, não havendo portanto critérios de escolha para as regras extraídas a fim de que a quantidade de Regras de Associação seja reduzida.

Outra dificuldade proporcionada pela quantidade elevada de Regras de Associação extraídas é a identificação do conhecimento. Nesse sentido o uso de Redes pode facilitar o processo por apresentar um resultado gráfico para extração do conhecimento (PANDEY *et al.*, 2009).

Diversos sistemas no mundo real podem ser representados por meio de Redes, por exemplo, os sistemas comerciais que produzem as relações entre clientes e produtos adquiridos. Desse modo, Redes são uma forma natural de representar matematicamente esses sistemas. A Análise de Redes é a área do conhecimento que investiga a estrutura de uma Rede a fim de obter conhecimento importante sobre seus elementos e suas interações (NEWMAN, 2010).

A área de análise de Redes Complexas está evoluindo de maneira dinâmica e em larga escala. As Redes auxiliam na observação de vários aspectos nos dados minerados, como por exemplo, a presença de estrutura de comunidade, que é representada por grupos de vértices densamente ligados e, ao mesmo tempo, com as conexões esparsas entre os grupos (FERREIRA; ZHAO, 2016).

Muitas pesquisas utilizam estruturas de Rede para realizar o pós-processamento das Regras de Associação a fim de facilitar a extração do conhecimento por intermédio da análise das relações extraídas dos dados (CHERN-TONG; AZIZ, 2018; MARTÍNEZ-BALLESTEROS *et al.*, 2017; LI *et al.*, 2017; MINELGA *et al.*, 2017; MEI; JIANG, 2016; NGUYEN *et al.*, 2015; NGUYEN; NGUYEN, 2015; PAPER *et al.*, 2015; SELMANE; BOUSSAID; BENTAYEB, 2015). Apenas os trabalhos de Pandey *et al.* (2009) e Chawla, Davis e Pandey (2004) apresentam abordagens que possuem um atributo alvo (item objetivo), mas sem nenhum tipo de garantia de que as relações observadas geram hipóteses verdadeiras ou que o conhecimento formulado possa ser utilizado em aplicações de diversas áreas.

Wang e Xu (2018) apresentam o uso de Grafos para auxílio no processo de Mineração de Regras de Associação. As estruturas são utilizadas apenas na etapa de extração das regras e o conhecimento gerado não possui uma avaliação. Para o pós-processamento das Regras de Associação, pode-se utilizar sistemas intuitivos de construção do conhecimento pela observação direta dos Grafos (CHEN *et al.*, 2018). Outro cenário é visto na pesquisa de Makino, Kato e Tanimoto (2017), na qual os autores apresentam a possibilidade de inferência de estruturas imprevisíveis, ampliando a elaboração da geração de hipóteses candidatas no processo de extração do conhecimento. Embora as regras possam ser visualizadas em estruturas de Redes, nenhum dos processos apresenta o direcionamento da mineração para mais de um item objetivo e ainda, não descrevem um método de validação das hipóteses formuladas.

1.2 Motivação e Lacunas

A pesquisa em algoritmos de Mineração de Dados tornou possível filtrar eficientemente grandes quantidades de dados e gerar padrões candidatos que são interessantes, muitas vezes

surpreendentes e em alguns casos mais confiáveis. Liu, Zhai e Pedrycz (2012) afirmam que a etapa da Mineração de Regras de Associação que recebe maior atenção da comunidade é a de pré-processamento. Nesta etapa, pode-se utilizar a modelagem em Rede para remover itens e conexões não interessantes. As Redes também podem ser utilizadas a fim de otimizar a formação dos *itemsets* frequentes, reduzindo a quantidade de itens a serem explorados pelo usuário. Um dos desafios das técnicas de Mineração de Regras de Associação é que elas geralmente geram um grande número de “padrões”, o que torna extremamente difícil para um pesquisador decidir quais regras são confiáveis e valem a pena serem analisadas (ZAKI; MEIRA, 2013). Apesar de diversos trabalhos realizados buscarem alternativas para extração do conhecimento de forma objetiva, a utilização das Redes no processo de ARM, com o uso de um atributo alvo, têm sido pouco explorada na literatura.

Realizando um estudo sobre Mineração de Regras de Associação, Pandey *et al.* (2009) elaboraram uma abordagem que efetua a construção de uma Rede utilizando as Regras de Associação extraídas de uma base de dados. As Redes de Regras de Associação (ARN - do inglês *Association Rules Network*) propiciam uma visualização mais eficiente dos padrões e possuem como objetivo principal encontrar relações entre um item alvo dos dados e os demais itens existentes na base de dados.

Vários dos problemas encontrados no mundo real consistem na relação entre vários itens de modo extremamente complexo, no qual torna-se relevante a análise de mais de um item alvo simultaneamente, a fim de ampliar a extração do conhecimento. Portanto, embora diversos trabalhos busquem a abordagem multialvo, alguns desafios ainda precisam ser superados, como a utilização de Redes para este tipo de abordagem. Embora as ARNs possibilitem um amplo estudo dos dados, elas permitem apenas conjuntos RHS e LHS unitários ($|RHS| = 1$ e $|LHS| = 1$), impossibilitando estabelecer maiores correlações dos itens com mais de um elemento como objetivo.

Nesse contexto, esse trabalho foi motivado pelo potencial uso de Redes na otimização da extração do conhecimento, em processos de Mineração de Regras de Associação, formulando abordagens explicáveis e pela lacuna referente ao uso de Redes em tarefas multialvo inerente de várias aplicações do mundo real. Neste trabalho são propostas técnicas de avaliação de atributos relacionados à extração das Regras de Associação, bem como validações matemáticas objetivas das regras para elaboração de hipóteses com maior probabilidade de serem verdadeiras. O interesse principal é identificar métodos de mineração de padrões, com um ou dois itens objetivos, e avaliar seus impactos tanto na representação do conhecimento gerado quanto em relação à interpretabilidade e expressividade das representações construídas.

1.3 Questões de Pesquisa

Este trabalho de doutorado foi desenvolvido com o intento geral de avançar as pesquisas da área de Mineração de Regras de Associação, em especial na geração de hipóteses para identificação do conhecimento com o uso de Redes. Assim, o desenvolvimento deste trabalho foi guiado por questões de pesquisa referentes ao estado atual das pesquisas de Mineração de Regras de Associação com foco no uso de estruturas de Rede na etapa de pós-processamento das regras e também referentes ao impacto das Redes tanto na representação do conhecimento quanto na tarefa de geração de hipóteses. Essas questões de pesquisa são apresentadas a seguir.

Q1 Como o uso das Redes pode auxiliar nos processos de Mineração de Regras de Associação?

Essa questão de pesquisa foi dividida em quatro questões secundárias:

Q1.1 Quais os tipos de Rede são utilizados nos processos de Mineração de Regras de Associação?

Q1.2 Quais tarefas de mineração utilizam Redes nos processos de Mineração de Regras de Associação?

Q1.3 Quais fases da Mineração de Regras de Associação utilizam Redes?

Q1.4 Como é feita a geração de hipóteses nos processos de Mineração de Regras de Associação que utilizam Redes?

Q2 Como é feita a validação da extração do conhecimento nos processos de Mineração de Regras de Associação?

Q3 Quais as relações que existem entre métricas relacionadas a Regras de Associação e medidas referentes a estruturas de Rede?

Q4 Quais as Medidas Objetivas utilizadas nos processos de Mineração de Regras de Associação que fazem uso de estruturas de Rede?

Essa questão de pesquisa foi dividida em duas questões secundárias:

Q4.1 Como as Medidas Objetivas interferem na geração do conhecimento pela construção de hipóteses?

Q4.2 Como as Medidas Objetivas podem ser utilizadas na validação da extração de conhecimento?

Q5 Em problemas multialvo, como podem ser utilizadas estruturas de Rede nos processos de Mineração de Regras de Associação?

Q6 Qual o impacto nos resultados da geração de hipóteses em processos de Mineração de Regras de Associação com o uso de Redes?

De acordo com as questões de pesquisa estabelecidas, foram definidos os seguintes objetivos para o desenvolvimento deste trabalho.

1. Mapear os trabalhos realizados na comunidade de Mineração de Regras de Associação que se preocupam com o uso de Redes. Tradicionalmente, as Redes são aplicadas em uma etapa específica do processo de mineração, sendo recursos auxiliares para otimização da etapa. No entanto, nos últimos anos, diversas pesquisas surgiram apontando o uso das Redes como protagonistas do processo de Mineração de Regras de Associação, aumentando assim a importância desse tipo de estrutura no resultado final com a geração de hipóteses. Esse objetivo está relacionado à questão de pesquisa **Q1**.
2. Analisar métodos de obtenção do conhecimento por meio da geração de hipóteses em processos de Mineração de Regras de Associação que utilizam Redes, na etapa de pós-processamento, efetuando comparações com outras técnicas já estabelecidas na literatura. Esse objetivo está relacionado à questão de pesquisa **Q2**.
3. Analisar medidas relacionadas a Regras de Associação e métricas de Redes estabelecendo uma relação entre as mesmas para construção de modelos mais confiáveis de Mineração de Regras de Associação. Esse objetivo está relacionado às questões de pesquisa **Q3** e **Q4**.
4. Propor, desenvolver e validar métodos de Mineração de Regras de Associação com o uso de Redes e Medidas Objetivas que possibilitem uma geração de hipóteses com maior probabilidade de serem verdadeiras. Esse objetivo está relacionado às questões de pesquisa **Q4** e **Q6**.
5. Propor, desenvolver e validar métodos de Mineração de Regras de Associação com o uso de Redes em aplicações multialvo. O foco é elaborar abordagens mais amplas para explorar o conhecimento e realizar os testes de validação comparativos com outras técnicas de Mineração de Regras de Associação. Esse objetivo está relacionado às questões de pesquisa **Q5** e **Q6**.

1.4 Principais Resultados

Os principais resultados que foram obtidos com o desenvolvimento deste projeto, visando atingir os objetivos propostos, são listados a seguir.

R1 Mapeamento sistemático da literatura sobre o uso de Redes em processos de Mineração de Regras de Associação. Foi conduzida uma revisão sistemática da literatura desse amplo tema, abrangendo a literatura de Mineração de Regras de Associação que considera o uso de Redes em alguma etapa do processo. Seguindo um protocolo bem definido, revisou-se 2.371 estudos e como resultado obteve-se um mapeamento do tema. A partir da realização

desse mapeamento sistemático respondeu-se a questão de pesquisa **Q1**, por meio das questões secundárias **Q1.1** a **Q1.4**.

- R2** Identificação de processos de extração de conhecimento por meio da geração de hipóteses com o uso de Redes de Regras de Associação. Com esta análise foram definidos os processos necessários para a elaboração de propostas que propiciem a formação de hipóteses com maior probabilidade de serem verdadeiras. Com esse trabalho foi endereçada a questão de pesquisa **Q2**.
- R3** Análise da complexidade de *datasets* para possibilitar a avaliação dos problemas de Mineração de Regras de Associação com o uso de Redes para uniformização de toda a pesquisa. Esse resultado viabilizou a obtenção dos resultados **R6** e **R10**.
- R4** Construção de *datasets* com aplicações em problemas reais de auxílio a tomada de decisão para possibilitar a validação das hipóteses geradas por meio de classificadores automáticos ou com apoio de especialistas. Esse resultado viabilizou a obtenção dos resultados **R5**, **R6**, **R7** e **R9**.
- R5** Proposta, desenvolvimento e validação de método de avaliação de parâmetros para extração de Regras de Associação com o uso de Medidas de Redes. É proposto o uso de Medidas de Centralidade para avaliação dos parâmetros de suporte e confiança mínimos de extração das Regras de Associação e obtenção de resultados otimizados de acordo com a problemática relacionada. Com o desenvolvimento, análise e avaliação experimental dessas medidas, comparando-as com diferentes métodos da literatura, foram tratadas as questões de pesquisa **Q3** e **Q6**.
- R6** Análise de Medidas Objetivas relacionadas a Regras de Associação para obtenção de estruturas de Redes mais eficientes e eficazes em gerar hipóteses com maior probabilidade de serem verdadeiras por meio de cálculos de influência entre os elementos relacionados nas regras. Com esse trabalho respondeu-se a questão de pesquisa **Q4**, por meio das questões secundárias **Q4.1** e **Q4.2**.
- R7** Proposta, desenvolvimento e validação de Redes de Regras de Associação com o uso de Medidas Objetivas para geração de hipóteses com comprovação de influência entre os itens formadores das regras. É proposta a Rede de Regras de Associação Filtrada (*Filtered-ARN* - do inglês *Filtered-Association Rules Network*) que proporcionam uma extração de conhecimento otimizada. Com o desenvolvimento, análise, avaliação experimental e provas de conceito dessas Redes, comparando-as com diferentes métodos da literatura, foram tratadas as questões de pesquisa **Q2**, **Q4** e **Q6**.
- R8** Análise de Medidas relacionadas a estruturas de Rede para obtenção de métricas que podem ser relacionadas ao processo de seleção de Regras de Associação com maior interesse

ao objetivo do estudo por meio da mineração. Esse resultado viabilizou a obtenção do resultado **R5**.

- R9** Proposta, desenvolvimento e validação de Redes de Regras de Associação para abordagens multialvo. O uso de estruturas de Rede para estudo e geração de hipóteses em problemas relacionados a dois itens objetivos de atributos diferentes. É proposta a Rede de Regras de Associação Multialvo (MTARN - do inglês *Multi-Target Association Rules Network*) que proporcionam uma extração de conhecimento otimizada para dois itens simultâneos de um *dataset* ($RHS = 2$). Com o desenvolvimento, análise, avaliação experimental e prova de conceito dessas Redes, comparando-as com diferentes métodos da literatura, foram tratadas as questões de pesquisa **Q2**, **Q5** e **Q6**.
- R10** Proposta de métrica de Taxa de Complexidade para *datasets* obtidos no repositório da UCI¹. Estudo e análise de taxas de acurácia e tempo de complexidade de 11 *datasets* oriundos do repositório UCI proporciona uma escala de complexidade que poderá ser utilizada como parâmetro de seleção de conjuntos de dados. Com um critério objetivo, pode-se efetuar a comparação de resultados em vários experimentos relacionados a Mineração de Dados e Aprendizado de Máquina. Esse resultado viabilizou a obtenção dos resultados **R5**, **R7** e **R9**.

1.5 Organização do Texto

O restante desta tese está organizado como se segue.

Capítulo 2 - Fundamentos e Mapeamento Sistemático Neste capítulo são apresentados os principais conceitos básicos relacionados a Regras de Associação e como esses elementos são apresentados na literatura. São citados também os processos de obtenção das Regras de Associação por meio dos principais algoritmos encontrados na literatura, além de apresentar métricas utilizadas no processamento de análise dessas regras. O conceito de Redes é apresentado, bem como as principais medidas relacionadas a esse tipo de estrutura. As Redes de Regras de Associação (ARNs) também são apresentadas, bem como suas propriedades e aplicações. Um Mapeamento Sistemático da Literatura sobre o uso de Redes no auxílio à Mineração de Regras de Associação é relatado desde seu planejamento até os resultados encontrados e o mapa gerado do estado da arte das pesquisas nessa área.

Capítulo 3 - Relação entre Medidas de Centralidade e Medidas de Regras de Associação

Nesse capítulo é fornecida uma visão geral da metodologia aplicada para responder a questão de pesquisa **Q3**. É apresentada a proposta de uma nova medida de Relevância aplicada a Redes de Regras de Associação (ARNs) e uma proposta do uso de Medidas

¹ <<http://archive.ics.uci.edu/ml>>

de Rede para avaliação das métricas de suporte e confiança mínimos, bem como uma avaliação experimental com 3 *datasets* oriundos da UCI que auxiliaram na elaboração da abordagem proposta. Por fim, todo o processo da Prova de Conceito da proposta com uma base de dados real é demonstrado.

Capítulo 4 - Redes de Regras de Associação Filtradas Nesse capítulo é apresentada a proposta de uso de filtros de Medidas Objetivas assimétricas para seleção de Regras de Associação, bem como o uso de Redes com filtros dessas medidas, as Redes de Regras de Associação Filtradas (*Filtered-ARNs*) com suas definições e propriedades. A Avaliação Experimental com 3 *datasets* oriundos da UCI são apresentados, além de 3 Provas de Conceito com bases de dados reais. Todos os processos são descritos e seus resultados analisados.

Capítulo 5 - Redes de Regras de Associação Multialvo A proposta de uma abordagem multialvo com o uso de Redes na etapa de pós-processamento da Mineração de Regras de Associação é apresentada nesse capítulo. Descreve-se o processo de Mineração de Regras de Associação com o uso de Redes de Regras de Associação Multialvo (MTARNs) e a metodologia para sua construção. A Avaliação Experimental com 3 *datasets* oriundos da UCI, bem como toda a metodologia realizada nos experimentos de estudo da extração do conhecimento pelo uso da MTARN é detalhada, além de uma Prova de Conceito e os resultados da Mineração de Dados com o *dataset* real de pacientes submetidos a cirurgias de revascularização miocárdica. Todos os resultados são apresentados e avaliados.

Capítulo 6 - Conclusões Nesse capítulo são apresentadas as considerações finais desta tese. As questões de pesquisa apresentadas nesta Introdução são retomadas, sendo sintetizadas as contribuições deste trabalho. Na sequência são apresentadas as publicações originadas deste trabalho, as limitações e direções para trabalhos futuros.

Apêndice A - Repositório de Datasets Nesse apêndice são descritos todos os *datasets* gerados durante a pesquisa e disponibilizados on-line². O título, origem, informações relevantes, número de instâncias e atributos, bem como a descrição de todos os atributos e distribuição das classes que formam cada base de dados são apresentados. São disponibilizados 5 *datasets* com dados reais em diferentes áreas de aplicação.

Apêndice B - Taxa de Complexidade de Datasets do Repositório UCI É apresentada nesse apêndice a proposta para o cálculo da Taxa de Complexidade de 11 *datasets* oriundos da UCI. Todo o processo de coleta das informações e do cálculo da taxa são descritos. Os valores das Taxas de Complexidade são demonstrados e podem ser utilizados como um parâmetro objetivo de seleção de bases de dados do Repositório UCI.

² <<http://sites.labc.icmc.usp.br/dariobcalcada/tese/>>

FUNDAMENTOS E MAPEAMENTO SISTEMÁTICO

2.1 Considerações Iniciais

A descoberta de Regras de Associação é uma tarefa de mineração de dados que procura identificar padrões de dados em *datasets*, permitindo, após a sua interpretação, adquirir conhecimento específico acerca do problema em análise (LE; VO, 2016).

A Mineração de Regras de Associação pode ser usada como uma metodologia para descobrir hipóteses ou teorias candidatas em um domínio do conhecimento. Essa técnica possibilita explorar avanços recentes em processos de pesquisa em conjunto com métodos estatísticos tradicionais, que continuam sendo a base da verificação e validação de teorias (VINAYA; SHAH, 2016).

O ponto de partida do processo de Mineração de Regras de Associação são as observações dos eventos que acionam o pesquisador para acelerar os estudos conceituais e chegar a uma estrutura na qual o processo subjacente (que está gerando os eventos) possa ser eliciado. Uma Regra de Associação (AGRAWAL; SHAFER, 1996; AGRAWAL; SRIKANT., 1994; AGRAWAL; IMIELINSKI; SWAMI, 1994) é um padrão na forma $A \Rightarrow B$, no qual A e B podem ser atributos, itens ou mais geralmente “objetos de dados”.

A Mineração de Regras de Associação corresponde, portanto, a uma vasta gama de áreas de pesquisa. Assim, torna-se necessário limitar o escopo desta tese. Neste trabalho, o interesse está no problema de obtenção de conhecimento com maiores possibilidades de veracidade, por meio do uso de Regras de Associação e estruturas de Rede. O principal foco deste trabalho é avaliar o impacto das Redes no processo de extração de conhecimento e ampliar a técnica de geração automática de hipóteses pela descoberta de padrões de interesse em *datasets*. Conforme já apresentado no Capítulo 1, nesta tese o termo hipótese é toda proposição elaborada a partir da

visualização das Regras de Associação e da detecção de padrões extraídas durante o processo de mineração de dados.

Neste capítulo são apresentados os conceitos fundamentais relacionados aos temas envolvidos nessa tese (Regras de Associação e Redes) e como esses elementos são apresentados na literatura. Na Seção 2.2, são introduzidos os conceitos de Regras de Associação assim como sua definição e como essas regras são decompostas nas tarefas de mineração. São citados também os processos de obtenção das Regras de Associação por meio dos principais algoritmos encontrados na literatura, além de apresentar métricas utilizadas no processamento de análise dessas regras. Na Seção 2.3 são apresentados os conceitos fundamentais relacionados a Redes. O conceito de Redes de Regras de Associação é apresentado na Seção 2.4 bem como suas propriedades e aplicações. Na Seção 2.5 é relatado um Mapeamento Sistemático completo sobre o uso de Redes no auxílio à Mineração de Regras de Associação, desde seu planejamento até os resultados encontrados e o mapa gerado do estado da arte das pesquisas nessa área.

2.2 Regras de Associação

Atualmente, no decorrer das suas atividades cotidianas, empresas e organizações manipulam e armazenam uma grande quantidade de dados. Compreender esses dados, ou seja, conhecer a informação implícita nesses dados assume, cada vez mais, um papel de relevo no apoio à tomada de decisão (ZAKI; MEIRA, 2013).

A mineração de dados é frequentemente descrita como o processo de descobrir padrões interessantes em grandes bases de dados (AGGARWAL, 2015). Dada a grande quantidade de dados digitais que estão sendo constantemente gerados e armazenados, a mineração de dados oferece uma solução para o problema de rapidamente resumir e buscar relacionamentos não óbvios em dados.

Ao se tratar de Mineração de Regras de Associação, pode-se dividir este processo em três etapas (NGUYEN; NGUYEN, 2015; AGGARWAL; PROCOPIUC; YU, 2002; AGRAWAL; IMIELINSKI; SWAMI, 1994):

- **Pré-processamento:** preparação da base para a etapa de extração e preparação do domínio dos dados, podendo ocorrer a remoção de itens não interessantes.
- **Extração de padrões:** são efetuados os cálculos de medidas, construção dos *itemsets* frequentes e formulação das Regras de Associação propriamente ditas.
- **Pós-processamento:** efetua-se a remoção de regras não interessantes, divisão do domínio e redução do número de regras a serem exploradas pelo usuário.

Dada a dimensão das bases de dados atuais, o número de regras descobertas pode ser tão elevado ao ponto de transformar a sua interpretação em um novo problema de mineração. Deste

modo, é importante o entendimento das Regras de Associação e a busca de melhores maneiras para interpretá-las, por meio de técnicas mais fáceis e intuitivas, como o uso de Redes para facilitar as visualizações das regras (PANDEY *et al.*, 2009). Em todas as etapas da Mineração de Regras de Associação, pode-se aplicar algum tipo de abordagem utilizando-se Redes (RASHID; GONDAL; KAMRUZZAMAN, 2015).

Uma Regra de Associação caracteriza o quanto a presença de um conjunto de elementos nos registros de uma base de dados implica na presença de algum outro conjunto distinto de elementos nos mesmos registros (AGRAWAL; SRIKANT., 1994). Desse modo, o objetivo das Regras de Associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados. Por exemplo, observando os dados de vendas de um supermercado, sabe-se que 80% dos clientes que compram o produto Q também adquirem, na mesma ocasião, o produto W . Nessa regra, 80% corresponde a sua confiança (DOMINGUES; REZENDE, 2005).

Definição 1: seja $I = \{i_1, i_2, \dots, i_n\}$ um conjunto de objetos denominados itens que podem assumir valores binários 0 ou 1 (falso ou verdadeiro), que representam a presença ou não de um objeto em particular. Seja T um conjunto de transações, em que cada transação D corresponde a um conjunto de itens tal que $D \subseteq I$. Considera-se ainda que um conjunto de itens A está contido numa transação D , se todos os itens do conjunto tiverem valor “verdadeiro” na transação, ou seja, fizerem parte dessa mesma transação. Uma Regra de Associação R pode ser representada por uma expressão no formato: $A \Rightarrow B$, com $A \subseteq I, B \subseteq I$ e $A \cap B = \emptyset$. É ainda possível tratar as variáveis quantitativas ou qualitativas, criando intervalos de valores e utilizando-as, posteriormente, como variáveis binárias. A é denominado de antecedente (LHS - *Left Hand Side*) da regra e B o consequente (RHS - *Right Hand Side*).

Definição 2: para cada regra ($LHS \Rightarrow RHS$), extraída de um conjunto de transações T , é calculado um valor de suporte (sup), apresentado na Equação 2.1, que verifica a força de associação entre LHS e RHS (probabilidade de ocorrência da transação $LHS \cup RHS$); e um valor de confiança ($conf$), Equação 2.2, que mede a força da implicação lógica da regra (probabilidade condicional de RHS dado LHS) (AGRAWAL; IMIELINSKI; SWAMI, 1994).

$$sup(LHS \Rightarrow RHS) = P(LHS \cup RHS) \quad (2.1)$$

$$conf(LHS \Rightarrow RHS) = P(RHS|LHS) \quad (2.2)$$

O suporte pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto LHS quanto RHS, ao passo que a confiança é a probabilidade de que uma transação satisfaça RHS, dado que ela satisfaz LHS.

2.2.1 Extração das Regras de Associação

De acordo com Vasconcelos e Carvalho (2004), o problema de se descobrir todas as Regras de Associação pode ser decomposto em duas partes:

- Encontrar todos os conjuntos de itens que possuam um suporte de transações acima de um limite mínimo informado. São chamados de conjuntos de itens frequentes aqueles que têm suporte igual ou superior ao mínimo estabelecido.
- Gerar as Regras de Associação a partir dos conjuntos de itens frequentes. Deve-se selecionar apenas as regras que possuam o grau de confiança mínimo, correspondente à confiança mínima estabelecida.

Assim, dado um conjunto de transações, o problema de mineração por Regras de Associação está em gerar todas as regras que contenham o suporte e confiança iguais ou maiores do que os valores mínimos determinados pelo usuário, referenciados como *suporte_mínimo* (*minsup*) e *confiança_mínima* (*minconf*), respectivamente.

Por exemplo, considere a base de dados “Compra” apresentada na Tabela 1, com os dados referentes a compras diárias de um indivíduo por um período de 10 (dez) dias. Para essa questão define-se que *minsup* é 0,3 (30%) e que *minconf* é de 0,8 (80%).

Tabela 1 – Compras diárias

TID	Lista de Itens
1	café, pão, manteiga
2	leite, cerveja, pão, manteiga
3	café, pão, manteiga
4	leite, café, pão, manteiga
5	cerveja
6	manteiga
7	pão
8	feijão
9	arroz, feijão
10	feijão

Fonte: Elaborada pelo autor.

Considerando-se que LHS = {CAFÉ} e RHS = {PÃO}, pode-se calcular o suporte e a confiança para a regra (CAFÉ $\Rightarrow$ PÃO) tendo como resultados $sup(CAFÉ \cup PÃO) = 0,3$ ou 30% e $conf(CAFÉ \Rightarrow PÃO) = 1$ ou 100%. Este resultado implica duas afirmações i) “em

30% das compras (em 10 dias) desse indivíduo ele comprou café e pão” e ii) “sempre que esse indivíduo comprou café, ele comprou pão”. Essas afirmações podem auxiliar na elaboração de hipóteses que podem conduzir estudos futuros sobre o comportamento padrão de compras do sujeito analisado.

2.2.1.1 *Apriori*

O algoritmo mais conhecido para obtenção de Regras de Associação é o *Apriori* (AGRAWAL; IMIELINSKI; SWAMI, 1994). Ele encontra todos os conjuntos de itens frequentes, denominados *itemsets* frequentes (L_k).

O algoritmo emprega busca em profundidade e gera conjuntos de itens candidatos (padrões) de k elementos a partir de conjuntos de itens de $k - 1$ elementos. Os padrões não frequentes são eliminados. Toda a base de dados é rastreada e os conjuntos de itens frequentes são obtidos a partir dos conjuntos de itens candidatos.

O algoritmo principal (Figura 1) faz uso de duas funções: a função *Apriori_gen*, para gerar os candidatos e eliminar aqueles que não são frequentes, e a função *Gen_rules*, utilizada para extrair as Regras de Associação.

Outro algoritmo foi formulado na mesma época por Houtsma e Swami (1993). Chamado de SETM, o sistema executava o algoritmo e promovia uma extração orientada pelos *itemsets* frequentes, porém o algoritmo *Apriori*, (AGRAWAL; SRIKANT., 1994) em sua segunda versão desenvolvido em 1994, tornou-se o “algoritmo padrão”.

Em resumo o algoritmo *Apriori* divide a descoberta de Regras de Associação em dois subproblemas:

- Descoberta de todos os conjuntos de itens frequentes e;
- Geração das Regras de Associação (utilizando os conjuntos de itens frequentes descobertos).

Dentre estes passos, o primeiro é mais exigente sob o ponto de vista computacional, e tem merecido especial atenção por parte da comunidade de mineração de dados (ZHENJIANG; CHIN; TAKEICHI, 2001), levando ao surgimento de diferentes algoritmos cujo principal objetivo reside na melhoria da eficiência computacional do *Apriori*.

O algoritmo *Apriori* possui algumas desvantagens, como o fato de não suportar uma grande quantidade de dados brutos (AGGARWAL; PROCOPIUC; YU, 2002). Esse fato constitui um problema, já que atualmente a quantidade de dados está aumentando exponencialmente. Desse modo, daqui a alguns anos será impossível processar os dados que circulam em *gigabytes* a *terabytes* muito rapidamente. Portanto, é necessário encontrar técnicas eficientes que analisem

Figura 1 – Algoritmo *Apriori*

```

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on
candidate generation
Input:
  • D, a database of transactions:
  • min_sup, the minimum support count threshold.
Output: L, frequent itemsets in D.
Method:
1. L1 = find_frequent_1-itemsets(D);
2. for(k=2; Lk-1 ≠ ∅; k++){
3.   Ck = apriori_gen(Lk-1);
4.   for each transaction t ∈ D { // scan D for counts
5.     Ct = subset(Ck, t); // get the subsets of t that are candidates
6.     for each candidate c ∈ Ct
7.       c.count++;
8.   }
9.   Lk = {c ∈ Ck | c.count ≥ min_sup }
10. }
11. return L = ∪k Lk;
procedure apriori_gen(Lk-1; frequent (k-1)-itemsets)
1. for each itemset l1 ∈ Lk-1
2.   for each itemset l2 ∈ Lk-1
3.     if(l1[1] - l2[1]) ^ (l1[2] - l2[2]) ^ ... ^ (l1[k-2] - l2[k-2]) ^ (l1[k-1] - l2[k-1]) then {
4.       C = l1 JOIN l2;
5.       if has_infrequent_subset(c, Lk-1) then
6.         delete c; // prune step: remove unfruitful candidate
7.       else
8.         add c to Ck;
9.     }
10. return Ck;
procedure has_infrequent_subset(c: candidate k-itemset; Lk-1: frequent (k-1) = itemsets); //
use prior knowledge
1. for each (k-1) = subset s of c
2.   if s ∈ Lk-1 then
3.     return TRUE;
4.   return FALSE;

```

Fonte: Han e Kamber (2006).

e gerenciem dados escaláveis (KAKKAD; ALUVALU, 2013). Todavia, o algoritmo *Apriori* ainda continua sendo o mais adequado para a Mineração de Regras de Associação.

Existem outros algoritmos que buscaram corrigir a necessidade de percorrer todo o conjunto de elementos mais de uma vez. Esses algoritmos diferem-se do *Apriori* pela existência de estruturas que auxiliam no processamento das regras com apenas uma passagem pelo elemento da base de dados, consequentemente melhorando o desempenho computacional em relação ao algoritmo *Apriori*, mas com o mesmo resultado. Pode-se citar como exemplos o *Apriori-TID* e o *FP-Growth*.

2.2.1.2 Apriori-TID

Uma das melhorias que ocorreram ao algoritmo *Apriori* levaram à elaboração do algoritmo *Apriori-TID* (*Transaction ID*) (AGRAWAL; SHAFER, 1996), que buscou corrigir a necessidade de percorrer todo o conjunto de elementos mais de uma vez. Esse algoritmo difere-se do primeiro pela existência de uma estrutura que auxilia no processamento das regras com apenas uma passagem pelo elementos da base de dados, consequentemente melhorando o desempenho computacional em relação ao primeiro, mas com o mesmo resultado.

2.2.1.3 FP-Growth

Algoritmos tradicionais de associação adotam uma abordagem igual ou semelhante àquela adotada pelo algoritmo *Apriori*. No entanto, a geração de conjunto candidato é ainda um processo de grande custo computacional, especialmente quando há um grande número de padrões e/ou estes são longos, ou seja, padrões formados por um número expressivo de itens. Os algoritmos tradicionais necessitam percorrer repetidas vezes a base de dados para verificar e testar todos os conjuntos de candidatos e seus padrões correspondentes (HAN; KAMBER, 2006).

O algoritmo *FP-Growth* utiliza uma abordagem diferente do *Apriori*, pois não efetua a geração do conjunto de candidatos como também não realiza o paradigma gerar e testar do *Apriori*. O algoritmo *FP-Growth* codifica o conjunto de dados em uma estrutura compacta em forma de árvore chamada *Frequent Pattern Tree* (*FP-tree*) e extrai os conjuntos de itens frequentes diretamente desta estrutura. Isso possibilita uma melhor eficiência na geração das Regras de Associação, pois evita constantes acessos na base de dados (GRAHNE; ZHU, 2005).

O algoritmo necessita de no mínimo dois parâmetros de entrada (NANDI *et al.*, 2014) suporte (*sup*) e confiança (*conf*), métricas básicas em Regras de Associação. Definidos os parâmetros de entrada e uma base de dados a ser analisada, o primeiro passo do *FP-Growth* consiste em montar a *FP-tree* que irá armazenar os primeiros *itemsets* frequentes. Para isso, é necessário ler a base de dados duas vezes para que a *FP-tree* seja criada. Durante a primeira leitura é identificado o conjunto de *itemsets* frequentes F , de tamanho um, e seus respectivos suportes.

Logo após a primeira leitura da base de dados, os primeiros *itemsets* infrequentes, que possuem valor de suporte menor que o *minsup*, são eliminados e os demais são ordenados em ordem decrescente e armazenados em uma lista L . Inicia-se a segunda leitura da base de dados, na qual cada transação é lida novamente, os *itemsets* infrequentes de tal transação são desconsiderados e os frequentes são armazenados em uma lista T . A lista T é ordenada de acordo com o valor de suporte dos *itemsets* armazenados na lista L . Em seguida a lista T é passada como parâmetro para o método que insere a transação na *FP-tree*. Na Figura 2 é mostrado um exemplo de *FP-tree*.

Figura 2 – *FP-Tree*

Fonte: [Nandi et al. \(2014\)](#).

Na construção da *FP-tree*, o algoritmo *FP-Growth* cria uma tendência. Isso ocorre devido ao modo como as informações, que estão contidas nas transações que estão sendo lidas, são organizadas. Essa tendência, Padrão Frequente de Crescimento (do inglês - *Frequent Pattern Growth*), dá nome ao algoritmo e contribui para um melhor aproveitamento dos ramos inseridos na estrutura e a torna compacta.

Ao término da segunda leitura da base de dados, todas as transações estarão processadas e inseridas na *FP-tree* que estará corretamente criada. Ela irá representar a base de dados original e será passada como parâmetro para outro método do algoritmo *FP-Growth* responsável pela descoberta dos *N itemsets* frequentes. O método consiste no desenvolvimento de uma estratégia baseada na técnica de dividir para conquistar, na qual o problema é fracionado em subproblemas. O algoritmo faz isso considerando cada item da tabela de cabeçalhos. O primeiro item a ser processado é o último, o segundo é o penúltimo, e assim sucessivamente até o término do processamento de todos os itens.

Para cada *itemset* presente na tabela de cabeçalhos, bases de padrões condicionais são geradas, as quais são utilizadas para a construção das *FP-trees* condicionais, que relacionam os caminhos frequentes que se conectam aos nós correspondentes ao *itemset* em questão. Uma vez criadas, as *FP-trees* condicionais são utilizadas para encontrar os padrões frequentes que apresentam o *itemset* como consequente ([HAN et al., 2004](#)).

Tendo como base o conjunto de *N itemsets* frequentes encontrados por meio da *FP-tree*, o algoritmo *FP-Growth* gera as Regras de Associação e pode realizar o cálculo de algumas métricas como: *lift*, *leverage* e *convicção*, que serão descritos na Seção 2.2.2 ([NANDI et al., 2014](#)).

2.2.1.4 Outros Algoritmos

A Mineração de Regras de Associação tem sido amplamente utilizada em sistemas para suporte à decisão e para personalização do atendimento a clientes. Muitos destes sistemas são caracterizados por uma adição constante de dados às bases, o que inevitavelmente determina reavaliações constantes dos resultados das minerações. Nestes casos, utilizar algoritmos tradicionais, que mineram as bases inteiras a cada execução pode ser dispendioso e até impraticável (THOMAS *et al.*, 1997).

Com esta finalidade, Siqueira *et al.* (2002) desenvolveram o algoritmo *iFP-Growth* que, a partir das Regras de Associação calculadas para a base de dados original e dos dados adicionados posteriormente, calcula as Regras de Associação do conjunto de dados resultante. Este algoritmo é uma extensão do algoritmo *FP-Growth* e se distingue não apenas pelo seu caráter incremental, mas também pela otimização do uso dos recursos computacionais.

Outras abordagens baseadas em grade (LBA, do inglês - *Lattice-based approaches*) (LE; VO, 2016) foram elaboradas. Os métodos tradicionais de obtenção de Regras de Associação incluem duas tarefas: determinar *itemsets* frequentes e gerar as Regras de Associação. LBAs para obtenção de Regras de Associação implementam duas etapas diferentes: determinar a grade de *itemsets* frequente e gerar as Regras de Associação por meio dessa grade. O tempo total de mineração pela LBA supera os métodos tradicionais de obtenção de Regras de Associação. Além disso, a vantagem mais importante das LBAs para a mineração é que os algoritmos só precisam construir a estrutura uma única vez, tendo um grande ganho computacional em velocidade de processamento.

O algoritmo *Apriori* e suas muitas variantes melhoradas usam uma primeira abordagem que percorre a base de dados inúmeras vezes para a exploração do espaço de busca, o que os torna ineficiente ao lidar com conjuntos de dados densos, cargas de padrões comuns e suas formas longas, ou limiares de apoio muito baixos. Para resolver estes problemas, é necessário o uso de algoritmos de mineração que utilizem abordagens mais inteligentes. *TreeProjection* (AGARWAL; AGGARWAL; PRASAD, 2001) é um dos algoritmos que utilizam uma estrutura bem organizada de dados (árvore lexicográfica) para representar padrões frequentes gerando uma pesquisa mais eficiente.

No entanto, esse algoritmo ainda encontrou os mesmos problemas que *Apriori* para lidar com os conjuntos de dados de padrões densos e longos. Agarwal, Aggarwal e Prasad (2000) já haviam proposto uma abordagem mais profunda (*Depth*) que a primeira citada, mas com base na mesma estrutura de dados: *DepthProjection*. Em ambos os casos houveram ganho de velocidade de processamento, mas os algoritmos mostraram limitações ao lidar com grandes bases de dados esparsos.

A Mineração de Regras de Associação baseada em agrupamentos é mais complexa do que a Mineração de Regras de Associação tradicional, devido principalmente à necessidade de

identificação dos agrupamentos para a contagem correta do suporte de um *itemset*. O algoritmo *Apriori-Group* estende o algoritmo *Apriori* e permite obter Regras de Associação confiáveis em bases com agrupamentos de transações (RIBEIRO; VIEIRA; TRAINA, 2005). Os princípios são os mesmos do *Apriori*, porém esse algoritmo consegue melhores resultados por obter Regras de Associação em bases com conjuntos de transações relacionadas.

Uma abordagem que pode auxiliar a análise de Regras de Associação é o uso de taxonomias durante o pós-processamento (DOMINGUES; REZENDE, 2005). O uso de taxonomia ajuda a solucionar o problema da grande quantidade de padrões extraídos pela técnica de Regras de Associação (ADAMO, 2001). As taxonomias refletem uma visão coletiva ou individual de como os itens podem ser hierarquicamente classificados, podendo ser utilizadas para eliminar regras não interessantes e/ou redundantes.

Independente de qual algoritmo é utilizado para extração das regras, os critérios de seleção e classificação das mesmas deve ser claro e conciso.

2.2.2 Critérios para Seleção e Classificação das Regras de Associação

Medir o nível de interesse dos padrões descobertos é uma área ativa e importante da pesquisa de mineração de dados (USMAN; USMAN, 2016). Com base na diversidade das definições apresentadas, o interessante é tratado como um conceito amplo que enfatiza concisão, cobertura, confiabilidade, peculiaridade, diversidade, novidade, surpreendimento, utilidade e capacidade de ação. Esses nove critérios específicos são usados para determinar se um padrão é ou não interessante. Eles são descritos da seguinte forma:

- **Concisão:** um padrão (*itemset* frequente) é conciso se contiver relativamente poucos pares atributo-valor.
- **Generalidade ou Cobertura:** um padrão é geral se abrange um subconjunto relativamente grande de um conjunto de dados. A generalidade (ou cobertura) mede a abrangência de um padrão, ou seja, a fração de todos os registros no conjunto de dados que corresponde ao padrão.
- **Confiabilidade:** um padrão é confiável se o relacionamento descrito pelo padrão ocorrer em uma porcentagem elevada de casos aplicáveis. Por exemplo, uma Regra de Associação é confiável se tiver alta confiança. Muitas medidas de probabilidade, estatística e recuperação de informação foram propostas para medir a confiabilidade das Regras de Associação.
- **Peculiaridade:** um padrão é peculiar se estiver longe de outros padrões descobertos de acordo com alguma medida de distância. Padrões peculiares são gerados a partir de dados peculiares (*outliers*), que são relativamente poucos em número e significativamente diferentes do resto dos dados. Padrões peculiares podem ser desconhecidos para o usuário, portanto, interessantes.

- **Diversidade:** um padrão (*itemset* frequente) é diversificado se os seus elementos diferirem significativamente uns dos outros. A diversidade é um fator comum para descoberta de regras de interesse.
- **Novidade:** um padrão é novo para um usuário se ele não souber antes e não é capaz de deduzi-lo de outros padrões conhecidos. Nenhum processo de mineração de dados representa tudo o que um usuário conhece e, portanto, a novidade não pode ser medida explicitamente com referência ao conhecimento do usuário.
- **Surpreendimento:** um padrão é surpreendente (ou inesperado) se contradizer o conhecimento ou expectativas existentes de uma pessoa.
- **Utilidade:** um padrão é de utilidade se o uso por parte de uma pessoa contribui para alcançar um objetivo. Diferentes pessoas podem ter objetivos divergentes em relação ao conhecimento que pode ser extraído de um conjunto de dados.
- **Capacidade de ação/Aplicabilidade:** um padrão é acionável (ou aplicável) em algum domínio se ele possibilita a tomada de decisão sobre ações futuras neste domínio.

Esses nove critérios podem ser incluídos em três classificações: medidas objetivas, medidas subjetivas e medidas baseadas em semântica. Uma medida objetiva baseia-se apenas nos dados originais. Nenhum conhecimento sobre o usuário ou aplicativo é necessário. As medidas de interesse subjetivas consideram principalmente a opinião de um analista para determinar a força da regra (GONÇALVES, 2005) e as medidas baseadas em semântica utilizam a estruturas semânticas das regras para estabelecimento de sua importância (HAN; KAMBER; PEI, 2012).

A maioria das medidas objetivas baseiam-se em probabilidade, estatística ou teoria da informação. As medidas de interesse objetivas são índices estatísticos utilizados para selecionar regras dentre as muitas que podem ser descobertas por um algoritmo de Mineração de Regras de Associação. Algumas das medidas objetivas consideram que uma Regra de Associação é interessante apenas quando o valor de seu suporte real é maior do que o valor de seu suporte esperado ($supEsp$). O suporte esperado é computado baseado no suporte da união dos itens que compõem a regra $supEsp(A \cup B) = sup(A) \times sup(B)$ (GONÇALVES, 2005).

Tendo considerado nove critérios para determinar se um padrão é interessante, pode-se utilizar três métodos para realizar a determinação de interesse. Primeiro, pode-se classificar cada padrão como interessante ou desinteressante. Em segundo lugar, pode-se determinar uma relação de preferência para representar que um padrão é mais interessante do que outro. Em terceiro lugar, pode-se elencar os tipos de padrões. Assim, o uso de medidas de interesse facilita uma abordagem geral e prática para identificar automaticamente padrões interessantes em Regras de Associação.

Durante o processo de mineração de dados, as medidas de interesse podem ser usadas em três etapas. A Figura 3 mostra essas três etapas. Uma das funcionalidades demonstram que as

medidas podem ser usadas para podar padrões desinteressantes durante o processo de mineração, de modo a restringir o espaço de busca e, assim, melhorar a eficiência da mineração (AGRAWAL; SRIKANT., 1994).

Apesar da reconhecida importância do suporte (*sup*) e confiança (*conf*), como forma de caracterizar e avaliar o interesse das Regras de Associação, várias outras medidas podem ajudar nessa caracterização (NEVES, 2002).

Figura 3 – Papéis das medidas de interesse no processo de mineração de dados

Fonte: Adaptada de Geng e Hamilton (2006).

Uma das funções das medidas objetivas de interesse é demonstrar como os itens influenciam uns aos outros. Esta influência pode ser de modo direto, quando os itens variam de modo proporcional, ou inverso, quando a elevação da incidência de um item leva à diminuição de outro. Por exemplo, ao se verificar o número de animais em uma casa e a quantidade de ração consumida, diz-se que são grandezas diretamente proporcionais, nesse caso uma dependência direta, ao contrário, seria uma dependência inversa, como a relação entre a velocidade média de um trajeto e o tempo de duração do deslocamento.

Definição 3: medidas de interesse objetivas podem ser classificadas como simétricas ou assimétricas, isto é, uma medida M é simétrica se $M(A \Rightarrow B) = M(B \Rightarrow A)$ (TAN; STEINBACH; KUMAR, 2005). As medidas assimétricas apresentam resultados distintos para as regras $A \Rightarrow B$ e $B \Rightarrow A$.

Além das medidas usuais de suporte (*sup*) e confiança (*conf*), outras medidas para a regra ($LHS \Rightarrow RHS$) podem ser calculadas. Para a elaboração dos experimentos deste trabalho, foram utilizadas apenas medidas assimétricas devido à natureza direcionada da Redes de Regras de Associação, entre elas *Added Value*, *Cartainty Factor*, *Convicção*, *Gini Index*, *J-measure* e *Laplace*. Porém, algumas medidas simétricas são muito utilizadas como *Lift* (BRIN *et al.*, 1997; BERZAL *et al.*, 2002), *Rule Interest* (PIATETSKY-SHAPIO, 1991) e Teste do χ^2 (qui-quadrado) (LIU; HSU; MA, 1999).

Added Value [-1..0..1]: a medida *Added Value* (AV) indica o quanto a frequência do consequente aumenta na presença do antecedente, ou seja, mede o ganho de RHS na presença de LHS (SAHAR, 2003). Se AV for positivo, então a frequência de RHS aumenta na presença de LHS. Sendo AV negativo, a frequência de RHS diminui na presença de LHS. Se AV possuir valor nulo (zero), tem-se uma coincidência aleatória, ou seja, a frequência de LHS não altera em nada a frequência de RHS.

$$AV = Conf(LHS \Rightarrow RHS) - sup(RHS) \quad (2.3)$$

Certainty Factor [-1..0..1]: a medida *Certainty Factor* (CF) mede o aumento da crença em RHS em consequência da observação de LHS (SHORTLIFFE; BUCHANAN, 1975), ou seja, a relação proporcional entre a descrença de RHS como resultado da observação de LHS. Se CF = 1 então LHS e RHS são diretamente proporcionais (dependência direta). Se CF = -1 então os itens são inversamente proporcionais (dependência inversa). Se CF = 0 significa que LHS não confirma nem contradiz RHS, ou seja, LHS e RHS são independentes. A medida CF pode ser vista como sendo o valor adicionado (AV, Equação 2.4) em relação ao máximo ganho que LHS pode proporcionar a RHS.

$$CF = \frac{sup(RHS|LHS) - sup(RHS)}{1 - sup(RHS)} = \frac{AV}{1 - sup(RHS)} = \frac{AV}{sup(\overline{RHS})} \quad (2.4)$$

Convicção [0..1..∞]: a medida de interesse convicção (Conv, em inglês *Conviction*) (Equação 2.5) (BRIN *et al.*, 1997) é proposta com o objetivo de avaliar uma Regra de Associação como uma verdadeira implicação (medida assimétrica). Foi desenvolvida como uma alternativa à confiança. A intuição por trás dessa medida recai sobre a lógica proposicional. Se “ $\Rightarrow$ ” for interpretado como uma implicação lógica, tem-se que $LHS \Rightarrow RHS \equiv (\neg LHS) \vee RHS \equiv \neg(LHS \wedge \neg RHS)$. Portanto, a força de $LHS \Rightarrow RHS$ é a mesma de $\neg(LHS \wedge \neg RHS)$ e pode ser medida em termos de quanto a ocorrência de $\neg(LHS \wedge \neg RHS)$ se distancia da ocorrência conjunta dos eventos LHS e $\neg RHS$, considerando os mesmos independentes (ADAMO, 2001). Se Conv = 1, LHS e RHS são estatisticamente independentes. Se Conv < 1, existe uma dependência inversa entre LHS e RHS. Se Conv > 1 a correlação é direta. No caso de Conv = ∞, quando o antecedente nunca aparece sem o consequente (conf = 100%), tem-se uma implicação lógica total (BRIN *et al.*, 1997). A principal desvantagem dessa medida é o seu intervalo sem limite superior, o que dificulta a comparação da medida entre todas as demais regras, pois seus valores não possuem diferenças significativas, sendo também difícil a definição de um *threshold* para a mesma.

$$Conv = \frac{sup(LHS).sup(\overline{RHS})}{sup(LHS \Rightarrow \overline{RHS})} \quad (2.5)$$

Gini Index [0..0,5]: a medida Gini Index (GI) (Equação 2.6) é frequentemente utilizada como critério de seleção de atributos na indução de árvores de decisão. Ela é utilizada para

medir o decréscimo da incerteza de uma determinada classe (meta) em observação de uma determinada classe preditora (FISHER, 1996). Desta forma, se duas variáveis estiverem altamente associadas, o grau de redução da incerteza entre as mesmas será grande, ou seja, quanto maior for o valor da medida, maior será a associação entre as variáveis.

$$GI = \sup(LHS) \cdot [conf(LHS \Rightarrow RHS)^2 + conf(LHS \Rightarrow \overline{RHS})^2] + \sup(\overline{LHS}) \cdot [conf(\overline{LHS} \Rightarrow RHS)^2 + conf(\overline{LHS} \Rightarrow \overline{RHS})^2] - \sup(RHS)^2 - \sup(\overline{RHS})^2 \quad (2.6)$$

J-Measure [0..1]: a medida J representa o produto de dois termos (Equação 2.7). O primeiro termo, $\sup(LHS)$, pode ser visto como uma preferência para a generalidade, ou seja, o antecedente da regra deve ocorrer frequentemente para que uma regra seja considerada interessante (GOODMAN; SMYTH, 1991). O segundo termo, representado entre colchetes, mede a diferença entre a confiança da regra $conf(LHS \Rightarrow RHS)$ e o suporte do anterior $\sup(LHS)$, ou seja, o poder discriminativo de LHS em RHS (CARVALHO, 2007). Portanto, maximizar o produto entre esses dois termos é equivalente a maximizar simultaneamente tanto a generalidade da hipótese LHS, quanto realizar um ajuste entre LHS e RHS (GOODMAN; SMYTH, 1991). Quanto maior for o valor de J, maior será a influência de LHS em RHS. O valor máximo de J é atingido quando a $conf(LHS \Rightarrow RHS) = 1$ e $conf(LHS \Rightarrow \overline{RHS}) = 0$, demonstrando total influência de LHS em RHS. Quando $conf(LHS \Rightarrow RHS) = \sup(LHS)$ e, por consequência, $conf(LHS \Rightarrow \overline{RHS}) = \sup(\overline{RHS})$, o valor mínimo de J é gerado, provando a falta total de influência entre os elementos da regra.

$$J = \sup(LHS) \cdot [conf(LHS \Rightarrow RHS) \cdot \log\left(\frac{conf(LHS \Rightarrow RHS)}{\sup(RHS)}\right) + conf(LHS \Rightarrow \overline{RHS}) \cdot \log\left(\frac{conf(LHS \Rightarrow \overline{RHS})}{\sup(RHS)}\right)] \quad (2.7)$$

$$J = \sup(LHS \cap RHS) \cdot \log\left(\frac{conf(LHS \Rightarrow RHS)}{\sup(RHS)}\right) + \sup(LHS \cap \overline{RHS}) \cdot \log\left(\frac{conf(LHS \Rightarrow \overline{RHS})}{\sup(RHS)}\right) \quad (2.8)$$

Observando a Equação 2.7 e realizando a operação distributiva, aliado com as devidas simplificações, obteve-se a Equação 2.8.

Laplace [0..1]: uma medida muito utilizada nos algoritmos de indução de regras que foi desenvolvida com o objetivo de penalizar regras muito específicas (regras que cobrem poucos exemplos (transações)) a fim de evitar *overfitting* (SMALDON; FREITAS, 2006). Para atingir a este objetivo, foram adicionados à medida de confiança elementos que agregam penalidades, como por exemplo, se uma determinada regra ocorre apenas em uma única transação. Sem a correção proposta pela medida Laplace, a Confiança dessa regra, considerando um valor de $N = 100$, seria de 100% (0.01/0.01), um valor muito otimista para uma regra extremamente específica. Sendo assim, levando-se em consideração o número de classes contidas no domínio pela medida Laplace, obtém-se um valor mais real, que no

caso do exemplo anterior seria de 66.7% $((100 * 0.01 + 1)/(100 * 0.01 + 2))$ (Equação 2.9) (SMALDON; FREITAS, 2006). A variável N indica o número de transações e o número 2 no denominador, no caso das Regras de Associação, indica o número de classes, além de representar a presença ou ausência de um item em uma transação.

$$L = \frac{N.\text{sup}(LHS \cap RHS) + 1}{N.\text{sup}(LHS) + 2} \quad (2.9)$$

Gain [0..1]: é uma medida proposta por Fukuda *et al.* (1996) que dá um *trade-off* entre suporte e confiança, auxiliando na seleção das regras de acordo com a frequências da mesma em relação à confiança mínima.

$$\text{Gain} = \text{sup}(LHS \cap RHS) - \text{minconf}.\text{sup}(LHS) \quad (2.10)$$

Se considerarmos a Equação 2.2, podemos inferir que:

$$\text{sup}(LHS \cap RHS) = \text{sup}(LHS).\text{conf}(LHS \Rightarrow RHS)$$

portanto, fazendo a substituição na Equação 2.10, e colocando-se em evidência o fator $\text{sup}(LHS)$ obtém-se a Equação 2.11.

$$\text{Gain} = [\text{conf}(LHS \Rightarrow RHS) - \text{minconf}].\text{sup}(LHS) \quad (2.11)$$

Por meio da Equação 2.11, percebe-se que a medida objetiva *Gain* funciona como uma normalização da medida de *Confiança*. Quando o valor de $\text{Gain} = 0$ a confiança da regra é igual a confiança mínima ($\text{conf}(LHS \Rightarrow RHS) = \text{minconf}$), e a partir daí todos os outros valores são calculados, sendo que $\text{Gain} = 1$ evidencia um valor de dependência estatística total, pois seria possível apenas quando o valor de *minconf* é insignificante ($\text{minconf} = 0$, i.e. sem limitação de confiança), a $\text{conf}(LHS \Rightarrow RHS) = 1$ e o $\text{sup}(LHS) = 1$, ou seja, em todos os momentos que ocorresse LHS, também ocorreria RHS, com LHS em todas as instâncias da base de dados analisada.

A vantagem desta medida sobre a confiança é que pode-se calcular com mais exatidão a influência do elemento antecedente sobre o consequente, provocando assim uma maior credibilidade à medida, podendo também, ser utilizada para selecionar regras.

Todas as medidas objetivas podem ser utilizadas na seleção de regras interessantes, e consequentemente ser aplicadas para elaboração de metodologias de extração de conhecimento de modo mais eficaz, como aqueles que utilizam Redes.

2.2.3 Pós-processamento das Regras de Associação

O Pós-Processamento é uma etapa importante do processo de Mineração de Regras de Associação na qual o conhecimento extraído pode ser simplificado, avaliado, visualizado ou simplesmente documentado (PIRI *et al.*, 2018). O processo de extração de Regras de Associação pode ser considerado uma tarefa simples, porém a compreensão de um conjunto elevado de regras torna a etapa de pós-processamento um grande desafio. Três etapas são definidas no pós-processamento de Regras de Associação (NAMAKI *et al.*, 2017; SIMARD; ST-PIERRE; BISKRI, 2016; HENDRICKX *et al.*, 2015; ZHAO; ZHANG; CAO, 2009):

- **Avaliação:** consiste em avaliar o conhecimento extraído do conjunto de dados por meio de critérios, tais como confiabilidade, aplicabilidade.
- **Interpretação e explicação:** consiste em documentar, visualizar, modificar e/ou comparar (ao conhecimento pré-existente) o conhecimento extraído do conjunto de dados de forma a torná-lo compreensível ao usuário.
- **Filtragem:** consiste em filtrar o conhecimento que foi extraído do conjunto de dados, podendo ser realizada por vários mecanismos que variam de acordo com a técnica utilizada.

Alguns estudos foram realizados por diversos pesquisadores, cujo objetivo principal era o de facilitar a interpretação de um grande número de Regras de Associação. Foram propostas diferentes abordagens, como avaliação por consulta, eliminação de regras redundantes e utilização de técnicas de visualização (FOURNIER-VIGER *et al.*, 2017).

2.2.3.1 Avaliação por consulta

A avaliação por consulta permite que o usuário explore o conjunto de regras por meio do uso de uma linguagem de consulta (tipicamente inspirada no SQL - Structured Query Language). Usando esse tipo de facilidade é possível obter subconjuntos de regras contendo determinados elementos e/ou satisfazendo alguma restrição definida pelo usuário, que envolvam, por exemplo, o suporte e a confiança (MEO; PSAILA; CERI, 1996).

2.2.3.2 Poda de regras

A poda (*pruning*) é uma das abordagens utilizadas para reduzir o número de regras geradas. Tem o objetivo de eliminar (excluir) regras redundantes ou que não são interessantes para o usuário. A poda de regras pode ser formulada pelo uso de diferentes métodos, obtendo também resultados diferenciados (TOIVONEN, 1996).

2.2.3.3 Técnicas de visualização

Uma área que também tem sido desenvolvida nos processos de mineração de dados e, mais concretamente, na interpretação de Regras de Associação, é a visual ou gráfica (HAHSLER; KARPIENKO, 2017; SANTOS; CARVALHO, 2017).

A visualização utiliza de recursos gráficos, como Redes, para uma estruturação visual do conhecimento, possibilitando o uso de técnicas gráficas, bem como a observação direta das relações obtidas no processo de mineração de dados (KLEMETTINEN *et al.*, 1994).

2.3 Redes

Uma Rede é descrita como um conjunto de elementos e suas relações, e com isso, podem ser representadas por um Grafo (NEWMAN, 2010). Segundo Caldarelli (2007), o termo “Rede” é usado para fazer referência para todo e qualquer sistema real que possa ser representado por um Grafo, e o termo “Grafo” é usado para representar um conjunto de regras e ferramentas conceituais específicas para descrever a Rede.

Uma Rede é uma representação simplificada (Figura 4) que reduz um sistema para uma estrutura abstrata que captura apenas o básico dos padrões de conexão. Uma Rede, em sua forma mais simples, é um conjunto de pontos unidos em pares por linhas. De maneira mais informal, os pontos são referidos como “vértices” ou “nós” e as linhas são referidas como “arestas”. Muitos objetos de interesse nas ciências físicas, biológicas e sociais podem ser modelados como Redes (NEWMAN, 2010).

Figura 4 – Representação simplificada de uma Rede

Fonte: Newman (2010).

Em geral, uma Rede R é representada como $R = (V, E)$, na qual V é um conjunto de vértices (ou nós) e E é um conjunto de arestas (ou links), que ligam alguns pares de vértices em V . Estatisticamente, um Grafo pode ser caracterizado por valores derivados, tais como o grau médio dos nós e o comprimento médio (caminho) entre os nós. Características adicionais como:

diâmetro da rede, número de triângulos, número de isomorfismos e o coeficiente de agrupamento, também podem ser analisados (NETTLETON, 2013).

Dada uma Rede $R = (V, E)$, vários links e auto-conexões não são permitidas dependendo do tipo de Rede que está sendo implementada. Se G é uma Rede Dirigida, considera-se o conjunto universal, denotado por U , contendo todas as $|V| * (|V| - 1)$ potenciais ligações dirigidas entre um par de nós em V , o qual $|V|$ denota o número de elementos em V . Se R é uma Rede Sem Direção, o conjunto universal U contém $|V| * (|V| - 1)$ links. Deste modo, a representação da Rede está relacionada diretamente ao tipo de dado que ela representa (VALVERDE-REBAZA; De Andrade Lopes, 2014).

Newman (2010) afirma que uma variedade de sistemas pode ser representada como Redes em que os dados podem ser reunidos seguindo algum critério. A função do sistema que a Rede representa pode indicar qual a forma ideal da Rede. O primeiro passo na análise da estrutura de uma Rede é muitas vezes fazer uma imagem da mesma (Figura 5). Cada Rede pode ser visualizada por uma imagem gerada por um programa de computador especializado projetado para construção gráfica da Rede. Atualmente, existem muitos desses programas disponíveis, dos quais, pode-se citar o *Gephi* (BASTIAN; HEYMANN; JACOMY, 2009) como exemplo.

A visualização pode ser uma ferramenta extremamente útil na análise de dados de Rede, permitindo que se possa ver imediatamente características estruturais importantes que de outra forma seria difícil detectar com os dados brutos. O olho humano é extremamente talentoso em escolher padrões, e visualizações permitem ao ser humano colocar esta cognição para trabalhar em problemas de Rede (NEWMAN, 2010).

2.3.1 Arquiteturas das Redes (representação)

Na última década, a análise de um grande número de Redes do mundo real revelou que algumas estruturas ou arquiteturas, são presentes em muitos sistemas naturais e artificiais. Essas estruturas são de relevância, uma vez que constituem a base da classificação dos sistemas reais. Uma Rede, também chamada de Grafo em literatura matemática, é uma coleção de vértices unidos por aros. Vértices e aros são também chamados de nós e arestas em ciência da computação, *sites* e títulos em física e atores e os laços em sociologia (NEWMAN, 2010).

A maior parte das Redes possui, no máximo, uma única aresta entre qualquer par de vértices. Nos casos raros, em que podem haver mais do que uma aresta entre o mesmo par de vértices, refere-se a estas arestas como uma multiaresta. Na maioria das Redes, também não há arestas que conectam vértices para si mesmos, embora tais arestas podem ocorrer em casos particulares. Tais arestas são chamadas de auto-arestas ou *auto-loops*. Uma Rede que não tem nem auto-arestas nem multiarestas é chamada de Rede Simples ou Grafo Simples. Uma Rede com multiarestas e/ou auto-arestas é chamada um Multigrafo (Figura 6).

Das inúmeras maneiras de representar matematicamente uma Rede, uma que se apresenta

Figura 5 – Visualização de uma Rede gerada pelo software *Gephi*

Fonte: [Bastian, Heymann e Jacomy \(2009\)](#).

com bons resultados é a Matriz de Adjacências que representa as ligações existentes entre os vértices de uma Rede. As matrizes de adjacências das Redes na Figuras 7(a) e 7(b) indicam a quantidade de arestas entre cada um dos vértices que são identificados pela linha i e coluna j da matriz, por exemplo: na matriz A da primeira Rede na segunda linha tem-se apenas uma ligação nas colunas 1, 3 e 4, e nenhuma nas demais, estando zeradas, para a primeira Rede. Já na matriz B para a segunda Rede tem-se uma ligação nas colunas 1, 2 e 4, duas na coluna 3, conforme descrito a seguir.

As matrizes também podem receber outros valores. Conforme [Newman \(2010\)](#), esses valores podem representar o peso, ou seja, o grau de aproximação entre os vértices de uma Rede formando uma Rede Ponderada. As matrizes também podem indicar algum tipo de direcionamento da Rede formando, nesse caso, uma Rede Direcionada que pode seguir a seguinte regra, por exemplo:

Figura 6 – Duas pequenas Redes. (a) Rede Simples (b) Rede Multiarestas com 1 (uma) multiaresta e 2 (duas) auto-arestas

Fonte: Adaptada de Newman (2010).

Figura 7 – Matrizes de Adjacência

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 0 & 1 & 0 & 0 & 3 & 0 \\ 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{pmatrix}$$

Fonte: Elaborada pelo autor.

$$A_{ij} = \begin{cases} 1, & \text{se for de } i \text{ para } j \\ 0, & \text{se for de } j \text{ para } i \end{cases}$$

Em alguns tipos de ligações, é necessário juntar mais de dois vértices de uma só vez. Por exemplo, pode-se querer criar uma Rede Social que representa famílias de uma comunidade maior de pessoas. As famílias podem ter mais de duas pessoas, sendo assim, a melhor maneira de representar os laços familiares em tais situações é usar um tipo generalizado de aresta que junta mais de dois vértices. Tal aresta é chamada hiper-aresta e uma Rede com hiper-arestas é chamado de Hipergrafo (Figura 8).

A associação de vértices em grupos representados por hiper-arestas em um Hipergrafo pode igualmente, e de modo mais convenientemente, ser representado como uma Rede Bipartida. Em tal Rede, há dois tipos de vértices, um representando os vértices originais e o outro representando os grupos a que pertencem (Figura 9).

Outra forma de representação de Redes é por meio de árvores. Uma árvore é uma Rede Conectada, não direcionada e que não contém circuitos fechados. Cada vértice na Rede é

Figura 8 – Hipergrafo - em hipergrafos as ligações são simbolizadas por *loops* que circulam os vértices

Fonte: Newman (2010).

Figura 9 – Rede Bipartite

Fonte: Newman (2010).

acessível a partir de todos os outros por meio de algum caminho através da Rede. Uma Rede pode também consistir em duas ou mais partes desconexas uma da outra, e se um nó individual não tem conexões também é chamado árvore. Se todas as partes da Rede são árvores, a Rede completa é chamada floresta. As árvores são frequentemente representadas graficamente de forma enraizada, ou seja, com um nó raiz no topo e uma estrutura de ramificação para baixo, na qual os vértices na parte inferior que estão ligados a apenas um outro vértice são chamados de folhas (Figura 10).

2.3.2 Medidas de Centralidade

Se a estrutura de uma Rede é conhecida, pode-se calcular, a partir de uma variedade de medidas quantitativas quais são aquelas características que conseguem se adaptar melhor a cada tipo de Rede. Algumas métricas são oportunas e selecionadas de acordo com o conhecimento que se deseja extrair (NEWMAN, 2010).

As medidas de centralidade são métricas importantes aplicadas em diversos estudos, como Redes Sociais, Redes de transportes e mercado financeiro. Elas podem ser utilizadas para medir o grau de relevância dos vértices de uma Rede R , ou seja, o quanto um vértice v_k é mais ou menos importante em relação aos demais vértices v_i de acordo com determinada medida, com

Figura 10 – Dois esquemas de uma mesma árvore. Na árvore (a) os vértices estão posicionados conforme uma conveniência e na árvore (b) segue-se a estrutura com nó raiz

Fonte: Newman (2010).

$1 \leq i \leq n, i \neq k$, sendo n o número total de vértices da Rede (SILVA, 2010).

Definição 4 Centralidade de grau ou informação (*Degree*): a concepção mais simples e intuitiva no que diz respeito à centralidade de um vértice é o número de contatos diretos que ele possui. A centralidade de grau é a contagem do número de adjacências de um vértice (Equação 2.12) (FREITAS, 2010).

$$Degree_i = \sum_{j=1}^n A_{ij} \quad (2.12)$$

O grau de um vértice também pode ser ponderado quando é atribuído algum valor às arestas conectadas ao mesmo. Pode-se calcular o Grau Médio de uma Rede por meio da média do grau de todos os seus nós, bem como o Grau Ponderado Médio, caso as arestas possuam pesos.

Definição 5 Centralidade de Auto-vetor (*Eigenvector*): a centralidade de auto-vetor, definida em (BONACICH, 2001), atribui relevância para um vértice em função de sua relação com os demais vértices da Rede.

Caso um vértice esteja ligado a outros que se encontram em uma posição central na Rede, o referido vértice terá uma alta Centralidade de Auto-vetor. Se um vértice estiver ligado a apenas um, e somente um outro vértice, que por sua vez esteja influenciando diversos outros vértices, que também atuam sobre vários outros, certamente terá Centralidade de Auto-vetor alta, mesmo influenciando diretamente apenas um vértice.

Definição 6 Centralidade de Intermediação (*Betweenness*): a centralidade de intermediação atribui importância a um vértice em função do número de vezes que existe um menor caminho que passa por ele. Pode ser definida como a porcentagem de vezes que um vértice necessita de outro vértice, cuja centralidade está sendo medida, para atingir o primeiro

(BRANDES, 2001). O vértice com maior centralidade de intermediação é aquele que participa de maneira mais ativa em um processo de interação, no qual os caminhos mais curtos são percorridos (Equação 2.13) (SILVA, 2010).

$$Betweenness_i = \sum_{j=1}^n \frac{n_{st}^i}{g_{st}} \quad (2.13)$$

em que n_{st}^i é o número de menores caminhos entre os vértices s e t que passam pelo vértice i e g_{st} é o número total de menores caminhos entre os vértices s e t .

Da mesma forma pode-se definir a Centralidade de Intermediação para uma aresta, bastando para isso ter como referência o número de menores caminhos que utilizam a aresta e (n_{st}^e), ocasionando na Equação 2.14.

$$ShortPath_e = \sum_{j=1}^n \frac{n_{st}^e}{g_{st}} \quad (2.14)$$

De fato, a Centralidade de Intermediação é dependente também da estrutura topológica. Para fins de análise preliminar de topologias alternativas em Redes de comunicação, por exemplo, projetistas podem optar por avaliar com mais detalhes as Redes com menor Centralidade de Intermediação, já que este resultado indica que os menores caminhos estarão mais distribuídos na topologia. Uma topologia com *Betweenness* alto pode sofrer maior impacto (perda de tráfego) no caso de falhas.

Definição 7 Page Rank: na internet, os sites, ou páginas, são interligados por meio de links (hiperlinks) formando uma Rede de Informações (BRIN; PAGE, 1998). O *Page Rank* (Equação 2.15) avalia a importância de um site, considerando a quantidade de links que cada página possui.

$$p(i) = \frac{q}{n} + (1 - q) \sum_{j:j \rightarrow i} \frac{p(j)}{k_{out}j}, i = 1, 2, \dots, n \quad (2.15)$$

em que $p(i)$ é o *PageRank* do vértice v_i , k_{out} é o grau do vértice v_i considerando apenas as arestas de saída, e a soma é executada nos vértices v_j nos quais as arestas se conectam ao v_i . O parâmetro q é o peso da união entre o caminho aleatório e a salto de página.

Pode-se citar mais uma série de outras métricas importantes no estudo das Redes. Com os valores da centralidade de todos os vértices, ou de todas as arestas, define-se o Comprimento Médio do Caminho (BRANDES, 2001) e o Diâmetro da Rede (COLEMAN; MORÉ, 1983). Além dessas, outras medidas que caracterizam a Rede de forma geral podem ser formuladas, como Densidade (COLEMAN; MORÉ, 1983).

Definição 8 Densidade: A Densidade D de um Grafo mede o quão perto a Rede está de ser completa. Uma Rede Completa possui todas as arestas possíveis ($D = 1$). A densidade é definida de acordo com o tipo de Rede, sendo calculada para Grafos Não Direcionados (Equação 2.16) e Grafos Direcionados (Equação 2.17), sendo E o número de arestas e V o número de vértices da Rede.

$$D = \frac{2 * E}{V(V - 1)} \quad (2.16)$$

$$D = \frac{E}{V(V - 1)} \quad (2.17)$$

Unindo os conceitos de utilização de Redes no auxílio da Mineração de Regras de Associação, principalmente em sua etapa de pós-processamento, bem como a poda das regras para otimização da extração do conhecimento, Pandey *et al.* (2009) elaboraram as Redes de Regras de Associação (ARN - do inglês *Association Rules Networks*).

2.4 Redes de Regras de Associação

Com o objetivo de facilitar a obtenção do conhecimento por meio das Regras de Associação, foram propostas algumas abordagens que aliam o uso de Redes à Mineração de Regras de Associação (ARM - do inglês *Association Rules Mining*) nas três principais fases de mineração (Pré-processamento, Extração e Pós-processamento) (RASHID; GONDAL; KAMRUZZAMAN, 2015) que serão mais detalhadas na Seção 2.5. Liu, Zhai e Pedrycz (2012) afirmam que a etapa da Mineração de Regras de Associação que recebe maior atenção da comunidade é a de pré-processamento que visa utilizar a modelagem em Rede para remover itens e conexões não interessantes, ou até mesmo para otimizar a formação dos *itemsets* frequentes, reduzindo a quantidade de itens a serem explorados pelo usuário.

Para a extração das regras, pode-se utilizar as Redes com o intuito de facilitar o processo e alcançar uma estrutura capaz de obter as regras automaticamente. Como exemplo, o algoritmo proposto por Nguyen, Vo e Hong (2015) utiliza uma estrutura de árvore para manter a informação relacionada de *itemsets* nos nós, e assim, acelerar o processo de geração de regras.

Abordagens de pós-processamento também são propostas com o uso de Redes, das quais pode-se citar, a abordagem com aprendizado transdutivo, no qual são selecionadas algumas regras a serem classificadas pelo usuário, direcionando-se esforços do mesmo com as regras consideradas de maior impacto na Rede, de acordo com alguma Medida de Rede (PADUA; REZENDE; CARVALHO, 2014). Uma operação importante no pós-processamento das regras é a poda, ou seja, a eliminação de regras de não-interesse, com isso o uso de Hipergrafos Direcionados é uma abordagem eficiente no auxílio deste processo (CHAWLA; DAVIS; PANDEY, 2004).

Nesta tese, foi feito o estudo das Redes de Regras de Associação com o intuito de expandir as suas aplicações, bem como otimizar o processo de extração do conhecimento pela elaboração de hipóteses com a observação dos padrões observados na Rede.

Redes de Regras de Associação (ARNs) possuem uma estrutura para sintetizar, podar, e analisar um conjunto de Regras de Associação para a construção de hipóteses candidatas. De uma perspectiva de descoberta de conhecimento, a ARN possibilita uma análise orientada para o contexto, o que facilita o entendimento dos relacionamentos e a utilidade das informações dos dados. Do ponto de vista matemático, ARNs são instâncias de Hipergrafos Direcionados (PANDEY *et al.*, 2009).

A ideia central da ARN é que as Regras de Associação descobertas pelo algoritmo de mineração podem ser sintetizadas, podadas, e integradas no contexto de objetivos específicos da pesquisa. Em particular, se houver uma variável de interesse (“alvo” ou “objetivo”), pode-se formar uma Rede com as variáveis mais relevantes e relacionadas ao objetivo, e, em seguida, elaborar uma estrutura que pode ser testada usando métodos estatísticos ou seja, acoplar uma tarefa de mineração de dados com análise estatística. Na prática, a Mineração de Regras de Associação envolve centenas de variáveis, portanto, a estratégia de poda que faz parte das ARNs pode ser usada para remover inconsistências locais entre as variáveis, como ciclos. Resumindo, ARNs oferecem as seguintes características:

- Poda no contexto: usa-se ARN para podar regras no contexto de um objetivo específico. Alterando o objetivo resultará na poda de regras diferentes.
- Estrutura de rede: ARNs fornecem um mecanismo para determinar a relação entre as variáveis relevantes e o objetivo utilizando a construção de uma Rede. Isso pode ajudar na análise dos efeitos de mudanças ocorridas de modo direto e indireto na mineração das Regras de Associação.
- Geração e avaliação de hipóteses: ARN pode servir como uma ponte entre as saídas geradas pela Mineração de Regras de Associação e sua avaliação.

Conforme Chawla (2010) descreve, ARNs utilizam como representação um Hipergrafo Inversamente-direcionado (B-graph - do inglês *Backward-directed hypergraph*), que após os processos de poda, podem ir transformando a ARN conforme o objetivo. Para a criação da ARN, são realizadas quatro etapas:

- (1) Passo A: dado um banco de dados D, o suporte mínimo e a confiança mínima, deve-se primeiro extrair todas as Regras de Associação utilizando um algoritmo padrão como *Apriori* (AGRAWAL; SRIKANT., 1994), *Apriori-TID* (AGRAWAL; SHAFER, 1996) e *FP-Growth* (GRAHNE; ZHU, 2005).

- (2) Passo B: escolher um item frequente z , que será representado no conjunto de regras como o nó objetivo, e construir um B -graph que flui de forma recursiva para z .
- (3) Passo C: realizar a poda do B -graph gerado no Passo B retirando hiperciclos e hiper-arestas reversas. O B -graph resultante é chamado de ARN.
- (4) Passo D: encontrar caminhos mais curtos entre o nó objetivo e os demais nós no nível mais superior (uma variante da distância entre extremidades) na ARN. O conjunto destes caminhos representa a Rede exploratória para o nó objetivo.

Definição 9: o nível de um determinado vértice $v \in \text{ARN}$, é o número de arestas necessárias para acessar o item objetivo z .

Por exemplo, o item z têm nível zero (Nível = 0), pois ele não precisa passar por nenhuma aresta para alcançar o próprio item z . Itens na parte LHS de regras que possuem $\text{RHS} \subset z$ terá o nível um (Nível = 1). Eles estão uma aresta longe do item em z .

Na Figura 11 é representado um exemplo de ARN, na qual foi selecionado o item “A” como objetivo. Seleciona-se então todas as regras que possuem “A” como consequente, neste caso apenas as regras $(B \Rightarrow A)$ e $(D \Rightarrow A)$. Assim, os itens “B” e “D” são modelados no nível 1 da ARN e passam a ser os objetivos, passando a executar o algoritmo de forma recursiva até que não restem mais itens como objetivos nos níveis mais altos da abordagem. Neste caso, foram modeladas as regras que possuem “B” como objetivo, depois as regras que possuem “D” e então “E”, “C” e “F”, respectivamente. Nesse exemplo, não existem regras que possuem “F” como subsequente. A hiper-aresta “ e_2 ” em destaque será umas das eliminadas no processo de poda, pois mesmo possuindo o item “C” como consequente, o item “B” já estava inserindo na ARN em um nível abaixo, inviabilizando esta regra.

Ampliando a utilidade das Regras de Associação, além da análise exploratória simples, ARNs são ferramentas que fornecem um contexto para a compreensão do relacionamento entre as regras durante a sua obtenção. Mais especificamente, ARNs oferecem os seguintes benefícios (PANDEY *et al.*, 2009):

- Organização de regras em um contexto: a principal utilidade da ARN está na organização de um potencialmente grande conjunto de Regras de Associação, de tal forma que um raciocínio objetivo pode ser explicado usando as regras mais relevantes no conjunto. Esta explicação pode se estender além das regras imediatas, que tem o item objetivo como um consequente e, portanto, fornece uma estrutura para organizar estas regras em um contexto, além de gerar hipóteses estatisticamente viáveis a partir de um conjunto de Regras de Associação derivados de dados brutos.
- Poda local: ARNs fornecem um método gráfico para podar regras pela associação de regras redundantes com hiperciclos e hiper-arestas reversas. Além disso, a poda ocorre

Figura 11 – Exemplo de ARN com hiper-aresta reversa (e_2)

Fonte: Adaptada de Pandey *et al.* (2009).

no contexto de um nó objetivo. Assim, uma regra que é redundante para um nó objetivo particular pode tornar-se relevante para o outro alvo, desta forma, essa abordagem é mais flexível do que a poda com base em medidas estatísticas de interesse.

Considere o ARN apresentado na Figura 11. Quando o nó objetivo é “A”, a hiper-aresta “ e_2 ” é uma regra redundante, e será eliminada. Por outro lado, quando o nó objetivo é “D” (Figura 12), a mesma hiper-aresta é relevante. Assim, a poda de uma regra torna-se dependente do contexto do nó alvo. Refere-se a este tipo de poda como uma poda local.

Figura 12 – ARN com nó objetivo D. A hiper-aresta e_2 faz parte do ARN. Isso prova a natureza adaptativa da poda local.

Fonte: Adaptada de Pandey *et al.* (2009).

ARNs implementam uma metodologia para integrar sistematicamente o processo de pesquisa de mineração de dados com métodos estatísticos tradicionais. Aplica-se mineração de dados com o objetivo de aperfeiçoar a abordagem teórica existente, enquanto que os métodos

estatísticos são utilizados para validar as conjecturas em relação aos dados. Esta abordagem híbrida possibilita explorar técnicas de pesquisa computacionalmente eficientes para grandes bases de dados em conjunto com métodos estatísticos tradicionais. A ARN fornece um mecanismo para sintetizar Regras de Associação de uma forma estruturada.

Também pode-se utilizar ARNs para seleção sistemática de atributos (CHAWLA, 2010) adaptando os processos de obtenção das Regras de Associação e relacionando-as ao atributo objetivo (alvo), para, em seguida, aplicar um algoritmo de agrupamento em Hipergrafo (HAN *et al.*, 1997) na ARN para extrair as características relevantes (no âmbito do domínio de destino). Deduz-se que os primeiros nós (nível = 1) tem um efeito imediato sobre o nó alvo, enquanto os nós de nível superior têm uma influência indireta.

Raeder e Chawla (2011) ilustram um exemplo (Figura 13) de análise de cesta de compras (*Market Basket*) com o uso da ARN, efetuando uma comparação de ARNs com outras técnicas de mineração, demonstrando assim o potencial desta técnica e como ela pode complementar algoritmos de detecção de comunidades, e outros algoritmos de mineração de dados.

O uso de ARNs é amplo e estende-se a várias áreas, porém ainda existe um vasto campo de pesquisa com relação à sua implementação em Mineração de Dados. Com intuito de mapear o estado da arte do uso de Redes aliado à Mineração de Regras de Associação foi realizado um mapeamento sistemático que será explanado na Seção 2.5.

2.5 Mapeamento Sistemático - Uso de Redes Aliado a Regras de Associação

A revisão relatada nesta seção é o resultado de um mapeamento sistemático de estudos (MSE), que foi realizado a fim de se obter uma visão geral do uso de Redes em processos de Mineração de Regras de Associação em publicações de pesquisa. Um mapeamento sistemático é um tipo particular de revisão sistemática da literatura (KITCHENHAM; CHARTERS, 2007).

Revisão sistemática da literatura é uma revisão sistematizada para identificar, avaliar e sintetizar evidências de resultados empíricos, a fim de responder a uma pergunta de pesquisa. Pode-se encontrar relatórios importantes sobre o uso de revisões sistemáticas especialmente na comunidade de engenharia de software (FELIZARDO *et al.*, 2014; KITCHENHAM *et al.*, 2010). Outras iniciativas esparsas também podem ser encontradas em outras áreas da ciência da computação, como ambientes baseados em nuvem (MOGHADDAM; LAGO; GROSSO, 2015), reconhecimento de padrões em imagens (PEDRO; NUNES; MACHADO-LIMA, 2013) e inteligência computacional (BRAGA *et al.*, 2015).

Uma revisão sistemática é realizada a fim de responder a uma pergunta de pesquisa e deve seguir um protocolo definido. O protocolo é desenvolvido ao planejar a revisão sistemática e é composto principalmente pelas questões de pesquisa, as estratégias e critérios para a

Figura 13 – ARN de análise de cesta de compras (*Market Basket*) com pão (*bagel*) como item objetivo

Fonte: Raeder e Chawla (2011).

busca de estudos primários, seleção do estudo e extração de dados (KITCHENHAM; DYBÅ; JØRGENSEN, 2004).

Para uma revisão ser considerada sistemática deve-se apresentar uma orientação muito útil para o planejamento e realização, sendo assim, deve-se seguir um protocolo bem definido e documentado, evitando subjetividades e tornando o processo mais reprodutível do que uma revisão da literatura regular.

Quando o campo de interesse é amplo e o objetivo é ter uma visão geral do que está sendo desenvolvido no domínio da investigação, recomenda-se a aplicação de um tipo particular de revisão sistemática chamado mapeamento sistemático (KITCHENHAM; CHARTERS, 2007). Mapeamentos sistemáticos de estudos seguem um protocolo bem definido como em qualquer revisão sistemática. As principais diferenças entre uma revisão sistemática tradicional e um MSE são a sua amplitude e profundidade. Enquanto uma revisão sistemática detalha estudos primários, em um MSE um maior número de estudos são analisados, em geral de modo menos

detalhado. Assim, os termos de pesquisa de um mapeamento sistemático são mais amplos e os resultados são normalmente apresentados por meio de gráficos. O MSE pode ser usado para obter um mapa das publicações sobre algum assunto ou campo. Além disso, pode identificar áreas que requerem o desenvolvimento de mais estudos primários e áreas em que uma mais refinada revisão sistemática da literatura seria de grande ajuda para a comunidade de pesquisa.

Nesta Seção será apresentado o mapeamento sistemático que visou verificar o estado da arte sobre o uso de Redes aliados à Regras de Associação e quais os *gaps* existentes nessa área do conhecimento. Os resultados do MSE apresentados nesse trabalho seguem um conjunto de diretrizes com o objetivo de identificar e sumarizar estudos relacionados ao uso Redes em processos de Mineração de Regras de Associação, principalmente sobre a implementação das Redes de Regras de Associação (ARN - do inglês *Association Rules Network*) (PANDEY *et al.*, 2009) e como as mesmas são aplicadas, bem como se existe alguma abordagem multiobjetivo para essas aplicações. A seguir apresenta-se o protocolo com as diretrizes que nortearam esse estudo.

2.5.1 Questões de Pesquisa de Mapeamento Sistemático

Buscando direcionar esforços durante o levantamento e compreensão dos Estudos Primários (EP) sobre o uso de Redes nas fases de Mineração de Regras de Associação, a seguinte questão de pesquisa guiou este trabalho. *QP: Quais os tipos de Redes que estão sendo utilizadas nos processos de Mineração de Regras de Associação?* Essa questão motivou a elaboração das seguintes sub questões:

- a. Existem trabalhos relevantes que utilizam Redes que implementam Regras de Associação com análise de mais de um item objetivo simultaneamente?
- b. Quais são as tarefas de Mineração de Regras de Associação que são auxiliadas pelas Redes?
- c. Em qual(is) fase(s) da mineração as Redes são utilizadas com as Regras de Associação (pré-processamento, extração, pós-processamento)?
- d. Quais os processos que são implementados com o auxílio de Redes em Mineração de Regras de Associação?
- e. Em qual(is) fase(s) da mineração as Redes são utilizadas com as Regras de Associação (pré-processamento, extração, pós-processamento)?
- f. Existem abordagens multiobjetivo do uso de Rede de Regras de Associação (ARN), conforme Pandey *et al.* (2009) indica como continuidade de sua pesquisa?

Baseando-se nas questões de pesquisa, utilizou-se o critério PICOC (do inglês - *Population, Intervention, Comparison, Outcomes, Context*) (PETTICREW; ROBERTS, 2006) para elaboração da *String* de Pesquisa utilizada para encontrar os estudos primários. Entretanto, não foi utilizado o critério de comparação, pois o foco do estudo não é comparar intervenções. Foram utilizados os critérios população, intervenção, resultados e contexto descritos a seguir:

- **População:** pesquisadores que atuam na área de Mineração de Regras de Associação.
- **Intervenção:** abordagens que utilizam Redes para auxiliar em alguma fase da Mineração de Regras de Associação.
- **Resultados:** a qualidade dos resultados e o esforço necessário para extração do conhecimento em Mineração de Regras de Associação.
- **Contexto:** projetos que demandam o uso de Redes em alguma fase da Mineração de Regras de Associação com um objetivo definido

2.5.2 Estratégia e String de Pesquisa

A estratégia de pesquisa se inicia por meio da formação da *string* de pesquisa que é utilizada para realizar as buscas nas bibliotecas digitais. As etapas realizadas nesse MSE serão discutidas a seguir.

O processo de construção da *string* de pesquisa foi realizado em 3 etapas: (E1) inicialmente foram identificados os termos chaves contidos em População e Intervenção (PICOC); (E2) foram identificados sinônimos ou palavras alternativas para cada termo encontrado em E1 e foi usado o conectivo *OR* para ligar os termos; (E3) após a realização de E2 os termos de População foram conectados com cada termo de Intervenção e os termos resultantes foram ligados por intermédio do conectivo *AND*. A tabela 2 apresenta o resultado de cada uma dessas etapas.

Foi realizada uma pesquisa piloto nas bibliotecas digitais *Compendex*, *Scopus* e *Web of Science* com o objetivo de verificar se a pesquisa abrangeria trabalhos relevantes da área. Com isso, observou-se que novos termos poderiam ser incluídos na *string* de pesquisa para torná-la mais robusta.

Os termos escolhidos para as tarefas de mineração da Intervenção em (E2) estão embasados em Zaki e Meira (2013) que utilizam essa divisão para orientar o objetivo da Mineração de Regras de Associação.

Durante as pesquisas piloto, verificou-se uma série de ruídos na busca que foram corrigidos restringindo a linguagem dos trabalhos para a língua inglesa e a área de conhecimento para a computação, por se tratar de buscas de técnicas e abordagens de implementação.

Tabela 2 – Construção da *String* de Pesquisa

Etapa	Resultado
E1	População: “Association Rule” População: “Network” Intervenção: “Association rules mining tasks”
E2	População: “Association Rule” OR “Association Rules” População: “Network” OR “Networks” OR “Graph” OR “Graphs” Intervenção: “Analysis” OR “Classification” OR “Regression” OR “Clustering”
E3	(“Association Rule” OR “Association Rules”) AND (“Network” OR “Networks” OR “Graph” OR “Graphs”) AND (“Analysis” OR “Classification” OR “Regression” OR “Clustering”)

Fonte: Elaborada pelo autor.

2.5.3 Processo de Busca

O processo de busca por estudos primários foi realizado aplicando a *string* de pesquisa em diferentes bibliotecas digitais utilizando os próprios mecanismos de busca avançada de cada uma delas. Wazlawick (2014) relata as principais bases de dados para pesquisa científica em computação. Entre as mais importantes estão: “*Web of Knowledge*” e “*Scopus*”. Essas duas bases de dados indexam importantes periódicos da área de computação segundo a classificação dos veículos de publicação (Qualis-CAPES)¹. O mapeamento sistemático foi realizado utilizando essas duas bases, como também as bases “*Compendex*” (*Engineering Village*), por se tratar de uma importante base para Engenharia da Computação, e “*DBLP Computer Science Bibliography*”, por conter conteúdos mais atualizados de conferências e congressos relacionados a pesquisas em Computação. Essas bases cobrem as publicações mais importantes no que diz respeito à Ciência da Computação como “*IEEE Xplore*”, “*ACM Digital Library*”, “*Elsevier*” e “*Springer*”. O período de tempo foi limitado entre os anos de 2011 e 2018 para ter uma visão de como estão os estudos mais recentes, além da velocidade exponencial de atualização das áreas de computação.

Como resultado dessa etapa de busca foram encontrados 2.371 trabalhos (Figura 14). A ferramenta *StArt* (HERNANDES *et al.*, 2012) foi utilizada para remoção dos trabalhos duplicados, bem como a *TheEnd* (BRAGA *et al.*, 2018) para realizar o primeiro refinamento. Esse primeiro refinamento contabilizou o número de itens da *string* de busca encontrados nos títulos, resumos e palavras-chaves (limiar), eliminando os artigos que possuíam um ou nenhum desses elementos no conjunto universo pesquisado (limiar < 2). Com isso obteve-se um total de 1269 candidatos a estudos primários. Na Figura 15 é descrita a distribuição dos artigos selecionados por ano de publicação.

¹ <<https://qualis.capes.gov.br>>

Figura 14 – Distribuição dos trabalhos por base de busca

Fonte: Elaborada pelo autor.

Figura 15 – Distribuição dos trabalhos por ano de publicação

Fonte: Elaborada pelo autor.

2.5.4 Critérios de Seleção de Estudos

Os critérios para inclusão e exclusão de estudos primários foram definidos em conformidade com os objetivos e questões de pesquisa.

Como critérios de inclusão foram considerados neste mapeamento os trabalhos que possuíam as seguintes características:

- Foram escritos em inglês;

- Possuíam palavras da expressão em seu título e/ou resumo;
- Foram publicados nos últimos 8 anos (De 2011 a 2018);
- Utilizavam Redes para auxiliar alguma fase do processo de Mineração de Regras de Associação;
- Possuíam uma comprovação de qual tipo de tarefa utilizou as Redes no auxílio da Mineração de Regras de Associação.

Como critérios de exclusão foram removidos desta pesquisa os trabalhos que possuíam pelo menos 1 dessas características:

- Possuíam menos de 2 páginas;
- Retornaram erros nos processos de buscas;
- Estavam hospedados em serviços pagos não financiados pela universidade;
- Não utilizavam Redes para auxiliar em alguma fase da Mineração de Regras de Associação;
- Eram baseados nos mesmos experimentos já considerados (o artigo de maior impacto foi mantido, medido pelo número de citações).

2.5.5 Processo de Seleção dos Estudos

Primeiramente foi realizada uma leitura de títulos e resumos. Nessa etapa, foram aplicados os critérios de inclusão e exclusão em todos os 1269 trabalhos candidatos a estudos primários identificados durante a pesquisa. Desses 1269 trabalhos, 487 foram aceitos porque satisfaziam os critérios de inclusão. Em seguida, foi efetuada a leitura das seções de introdução e conclusão dos 487 trabalhos selecionados na etapa anterior para um maior refinamento dos estudos primários, resultando em 144 artigos.

Na etapa final de seleção, foi realizada a leitura completa dos trabalhos. Nessa etapa, dos 144 trabalhos aceitos, 97 foram selecionados em definitivo, por serem os únicos a conter em sua totalidade as informações solicitadas nesse mapeamento.

2.5.6 Processo de Extração dos Dados

A extração das informações relevantes foi realizada com o auxílio das ferramentas *StArt* (HERNANDES *et al.*, 2012) e *TheEnd* (BRAGA *et al.*, 2018) contendo campos sobre a extração propriamente dita (pesquisador e data da extração), campos gerais (título, autores, ano, local de publicação, etc.) e contendo campos específicos para cada questão de pesquisa como:

- (a) A Base de Dados está disponível?

- (b) A significância estatística foi justificada?
- (c) É realizada uma discussão teórica sobre o método aplicado?
- (d) É utilizada uma validação objetiva?
- (e) O trabalho é comparado com trabalhos anteriores?
- (f) Utiliza Redes de Regras de Associação?
- (g) Qual a tarefa de mineração do uso de Redes em mineração das Regras de Associação?
- (h) Qual o tipo de Rede utilizado com a Mineração das Regras de Associação?
- (i) Em qual fase da mineração são utilizadas as Redes?

2.5.7 Resultados

Nesta Seção são apresentados os resultados que foram baseados nos 97 trabalhos incluídos como estudos primários da pesquisa.

A primeira análise realizada nos trabalhos selecionados foi sobre o local de publicação dos mesmos, com o intuito de verificar centros de pesquisas que concentram-se na área de Mineração de Regras de Associação, bem como observar a atuação dos pesquisadores nesta temática ao redor do mundo. Na Figura 16 é mostrada a distribuição dos artigos pelo local de publicação. Infere-se que as pesquisas em Mineração de Regras de Associação que utilizam Redes em alguma fase ainda estão dispersas, apesar de cinco locais se destacarem (China, Índia, EUA, França e Taiwan), o número de publicações ainda é pequeno quando verificado o período no qual os artigos foram disponibilizados (2011 a 2018).

O endereço e/ou link da base de dados utilizada nas pesquisa está disponível para uso por outros pesquisadores em apenas 36% dos artigos estudados (Figura 17). As principais bases de dados utilizadas nos artigos são as disponibilizadas pela *UCI Machine Learning Repository*². Os trabalhos utilizam essas bases para seus experimentos (WANG *et al.*, 2018; WANG; GUO; LI, 2017; USMAN; USMAN, 2016; PADUA; CARVALHO; REZENDE, 2016; NGUYEN; NGUYEN, 2015; IQBAL *et al.*, 2014; PADUA; REZENDE; CARVALHO, 2014; NIRMALA; PALANISAMY, 2013; LIU; ZHAI; PEDRYCZ, 2012; ZHAO; ZHANG; ZHANG, 2011) a fim de validar as abordagens propostas e comparar com resultados já existentes na literatura.

Nadi *et al.* (2014) utilizaram dois *datasets*, “Connect” e “Chess”. Trabalhos relacionados a dados de Redes Sociais utilizaram *datasets* em comum como o “Pokec”, “YAGO2”, “GraMi” e “Web graph” (WANG; XU, 2018; FAN; WU; XU, 2016). A pesquisa de Choobdar, Silva e Ribeiro (2011) apresenta dois *datasets* relacionados a modelos de Redes.

² <<http://archive.ics.uci.edu/ml>>

Figura 16 – Distribuição dos artigos selecionados pelo local de publicação

Fonte: Elaborada pelo autor.

No trabalho de [Sheu et al. \(2016\)](#), foi utilizado em sua avaliação experimental, dados de *spam*, “SpamAssassin” e “Enron-Spam”, todos disponíveis on-line.

Outros repositórios foram encontrados como o “SCADA” ([ANAND; VINODCHANDRA, 2018](#)), “CISE UFL” e “Standfor” ([AL-KHASSAWNEH; BAKAR; ZAINUDIN, 2012](#)). Foram também disponibilizados *datasets* de imagens: “Corel5k” e “IAPR” ([LI et al., 2017](#)), dados médicos e de vigilância em saúde: “MEPS Dataset” ([BELYI et al., 2016](#)) “BioGRID”, “MINT” e “InterPro” ([HENDRICKX et al., 2015](#)), além de “GSE15852” e “GSE31699” ([LEE; SHIN, 2017](#)) relacionados a pesquisas de câncer.

Datasets de pesquisa em mineração de textos com Regras de Associação também foram verificados. A pesquisa de [Namaki et al. \(2017\)](#) utiliza o “Citation” que consiste em dados de Rede de citações, o “Panama” que contém um total de 839K entidades *offshore*. [Su \(2017\)](#) utilizou dados de patentes americanas e [Sheu \(2018\)](#) palavras chaves coletadas do *website SafeSquid*.

Redes de Regras de Associação possibilitam o uso de análises estatísticas para auxiliar na extração do conhecimento, portanto, verificou-se quais trabalhos também executam esta análise, justificando os métodos utilizados. Na Figura 18 é demonstrado que os trabalhos ainda não possuem total preocupação com esse tipo de análise, abrangendo apenas 39% das pesquisas. Por falta

Figura 17 – Disponibilidade de Base de Dados nos artigos estudados

Fonte: Elaborada pelo autor.

de recursos, ou mesmo por não conseguir mensurar algumas métricas específicas os artigos não efetuaram a demonstração desse processo. Para dados relacionados à saúde (DELGADO *et al.*, 2018; MINELGA *et al.*, 2017; INDHUMATHY; NABHAN; ARUMUGAM, 2017; TSAI *et al.*, 2017; KIM *et al.*, 2017; TURGEMAN; MAY; SCIULLI, 2017; MARTÍNEZ-BALLESTEROS *et al.*, 2017; BELYI *et al.*, 2016; KIM *et al.*, 2016; HERNANDEZ-LEAL *et al.*, 2013), a validação estatística transcreve os índices de acurácia dos métodos, além de possibilitar um comparativo com outras abordagens já existentes na literatura. Outros trabalhos que utilizam Redes Bayesianas (DELGADO *et al.*, 2018; CAI *et al.*, 2017; NEMATİ; SANT'ANNA; NOWACZYK, 2016; IQBAL *et al.*, 2014; ODEH; AL-NAJDAWI, 2013; ZHAO; ZHANG; ZHANG, 2011; VEDULA; THATAVARTI, 2011) também executam o estudo estatísticos, pois são a base para a construção desse tipo de Rede.

A utilização de uma validação estatística foi possível em pesquisas relacionadas a Grafos e Redes (FAN; HU, 2017; NAMAKI *et al.*, 2017; ADARICHEVA; NATION, 2017; CAI *et al.*, 2017; FAN; WU; XU, 2016; LIU; SHEN, 2016; HENDRICKX *et al.*, 2015) por fazerem uso de Medidas de Rede que podem auxiliar no processo de mineração de dados e geração de conhecimento.

Estruturas de árvores de decisão foram utilizadas nas pesquisas de Li *et al.* (2017) e Sheu *et al.* (2016), além do trabalho de Wang *et al.* (2018) que utiliza árvores na validação estatística dos *itemsets* frequentes.

Outros trabalhos também realizaram a validação estatística com o uso de métricas específicas (YU *et al.*, 2017; SHENG *et al.*, 2016; KIM *et al.*, 2016; LI-XIONG *et al.*, 2015) e em etapas de pós-processamento de Regras de Associação (PADUA; CARVALHO; REZENDE, 2016; PADUA; REZENDE; CARVALHO, 2014).

Figura 18 – Execução de validação estatística na pesquisa

Fonte: Elaborada pelo autor.

Outro aspecto importante e de relevância científica é a descrição do método proposto na pesquisa a fim de que os experimentos possam ser replicados e o conhecimento propagado. Sendo assim, no estudo deste mapeamento foram encontrados 74% de trabalhos que apresentam a metodologia (Figura 19), possibilitando uma análise mais detalhada do processo de mineração. Destacam-se os trabalhos de Dasgupta e Saha (2016), por apresentar de modo detalhado um *framework* para descoberta de padrões importantes no processo de mineração paralela de proteínas, Wang *et al.* (2018), que apresenta um *framework* para Mineração de Regras de Associação temporais, e Fan e Hu (2017) por demonstrar a metodologia de construção de um *framework* para análise de dependências em Grafos. Estas são as únicas pesquisas a apresentarem *frameworks* com descrição do método proposto.

Os trabalhos que apresentam metodologias detalhadas foram aplicados em diversas áreas do conhecimento como por exemplo, sistemas de suporte à decisão (HONG *et al.*, 2018; MINELGA *et al.*, 2017; GAO *et al.*, 2017a; KIM *et al.*, 2016), pesquisas relacionadas a mineração de dados nas nuvens e com dados oriundos da web (SHEU, 2018; ZHANG *et al.*, 2018; SLAIMI *et al.*, 2017; SHAHMOHAMMADI; KHADANGI; BAGHERI, 2016; LI-XIONG *et al.*, 2015) e pesquisas em distribuição de energia elétrica (SHENG *et al.*, 2016; NEMATİ; SANT'ANNA; NOWACZYK, 2016).

No que tange à validação, 58% dos trabalhos buscaram uma validação (Figura 20), alguns inclusive com especialistas na área. A pesquisa de Baralis *et al.* (2013) utiliza os conceitos de sumarização com Grafos para implementar uma nova abordagem de detecção de correlações em múltiplos termos, sendo validada por especialistas. A avaliação por especialistas foi detectada principalmente em dados relacionados à saúde (HONG *et al.*, 2018; Özgür Cingiz; BIRICIK; DIRI, 2017; TSAI *et al.*, 2017; FU *et al.*, 2017).

Figura 19 – Descrição do método proposto aplicado na pesquisa

Fonte: Elaborada pelo autor.

Os processos de validação apresentados foram realizados com o uso de várias estruturas e métricas como Redes Bayesianas (DELGADO *et al.*, 2018; CAI *et al.*, 2017; NEMATI; SANT'ANNA; NOWACZYK, 2016) e árvores de decisão (FU *et al.*, 2017; LI *et al.*, 2017; SHEU *et al.*, 2016).

Figura 20 – Realização de validação objetiva

Fonte: Elaborada pelo autor.

Em 55% dos artigos é efetuada uma comparação (Figura 21) do método proposto com outros na literatura. Destaca-se a pesquisa de Ding *et al.* (2013) que trata de um tema muito comum, detecção de *malwares* de forma rápida. Os autores efetuam os experimentos e comparam seus métodos aos já tradicionais.

Outros trabalhos realizam otimizações em processos já existentes (ANAND; VINOD-CHANDRA, 2018; VALLE; RUZ; MORRÁS, 2018; WANG *et al.*, 2018; ABAY; MUTLU; KARAGOZ, 2016; FAN; WU; XU, 2016; CHEN; XING, 2016; USMAN; USMAN, 2016; PADUA; CARVALHO; REZENDE, 2016; ACHARYA; MAHALI; MOHAPATRA, 2015; PADUA; REZENDE; CARVALHO, 2014; NADI *et al.*, 2014; ALZOUBI; OMAR; BAKAR, 2011; VEDULA; THATAVARTI, 2011). Nestas pesquisas o foco maior é na técnica de mineração de dados e não em uma aplicação específica. Mostrando a importância de comparações entre métodos diferentes para comprovação da inovação e dos objetivos em cada trabalho.

Algumas pesquisas apresentam técnicas de aprendizado de máquina com o intuito de validar e comparar os resultados das técnicas compostas com heurísticas já aplicadas. No trabalho de Turgeman, May e Sciulli (2017), os resultados são comparados com técnicas de Redes de Função de Base Radial (RBF) e *Support Vector Machine* (SVM). Martínez-Ballesteros *et al.* (2017) comparam os resultados por meio de ontologias, algoritmos genéticos e árvores de decisão.

Figura 21 – Trabalhos que realizam comparações metodológicas com pesquisas anteriores

Fonte: Elaborada pelo autor.

A principal questão de pesquisa foi sobre os tipos de Redes utilizadas na Mineração de Regras de Associação, levando a questionar como essas Redes são utilizadas e qual tarefa de mineração as mesmas auxiliam. Na Figura 22 explicita-se a distribuição dos tipos de Redes encontradas, bem como suas comparações com as fases da mineração e o tipo de tarefa.

O principal tipo de Rede encontrada foi aquela relacionada aos Grafos Simples. Os Grafos foram utilizados em todas as fases de mineração como pré-processamento (ZHANG *et al.*, 2018; KIM *et al.*, 2017; SLAIMI *et al.*, 2017; ABAY; MUTLU; KARAGOZ, 2016; FAN; WU; XU, 2016; LIU; SHEN, 2016; HENDRICKX *et al.*, 2015; ACHARYA; MAHALI; MOHAPATRA, 2015; CHOBDAR; SILVA; RIBEIRO, 2011), extração (WANG; XU, 2018;

Figura 22 – Mapa de resultados da combinação das questões de pesquisa

Fonte: Elaborada pelo autor.

INDHUMATHY; NABHAN; ARUMUGAM, 2017; FAN; HU, 2017; NAMAKI *et al.*, 2017; ADARICHEVA; NATION, 2017; LEE; SHIN, 2017; YU *et al.*, 2017; WANG; GUO; LI, 2017; RAWAT *et al.*, 2016; SHAHMOHAMMADI; KHADANGI; BAGHERI, 2016; NADI *et al.*, 2014; RANJAN; VAISH; COULL, 2014; NGUYEN *et al.*, 2013; RAJPUT; THAKUR; THAKUR, 2012; AL-KHASSAWNEH; BAKAR; ZAINUDIN, 2012; ALZOUBI; OMAR; BAKAR, 2011) e pós-processamento (CHEN *et al.*, 2018; SU, 2017; GAO *et al.*, 2017a; MAKINO; KATO; TANIMOTO, 2017; WANG *et al.*, 2017; CHEN; XING, 2016; LIU; SHEN, 2016; RUSH *et al.*, 2016; PADUA; CARVALHO; REZENDE, 2016; SIMARD; ST-PIERRE; BISKRI, 2016; ZHENWEI; LINGYUN; LIZHU, 2016; KIM *et al.*, 2016; WILKE *et al.*, 2016; USMAN; USMAN, 2016; DU *et al.*, 2015; LI; DU; YANG, 2015; LI-XIONG *et al.*, 2015; PADUA; REZENDE; CARVALHO, 2014; BARALIS *et al.*, 2013; DING *et al.*, 2013; PETELIN *et al.*, 2013; RODRIGUES; GAMA; FERREIRA, 2012).

Na fase de pré-processamento destacam-se os trabalhos de Zhang *et al.* (2018) e Slaimi *et al.* (2017) por apresentarem processos de organização dos dados em estruturas de Rede a fim de que sejam estabelecidas conexões para aplicações web. De modo similar, a pesquisa de Kim *et al.* (2017) apresenta um Grafo para estabelecer as conexões entre doenças e genes de modo a possibilitar a extração das Regras de Associação.

Para extração das regras, os Grafos foram utilizados como estruturas de apoio para que os *itemsets* fossem formados e as medidas de suporte e confiança calculados. Destacam-se nessa fase as pesquisas de Wang e Xu (2018), por estabelecer Regras de Associação diretamente ligadas à estrutura do Grafo, o trabalho de Yu *et al.* (2017), que apresenta a extração de padrões espaciais em Redes e a pesquisa de Namaki *et al.* (2017), por estabelecer regras de associações temporais com o uso de Redes de eventos heterogêneas.

O pós-processamento das Regras de Associação com o uso de Grafos é bem difundido (Figura 22). O trabalho de [Chen et al. \(2018\)](#) merece destaque por apresentar um sistema intuitivo de construção do conhecimento pela observação direta nos Grafos. [Makino, Kato e Tanimoto \(2017\)](#) apresentam a possibilidade de inferência de estruturas imprevisíveis, ampliando a elaboração da geração de hipóteses candidatas no processo de extração do conhecimento. Na pesquisa de [Liu e Shen \(2016\)](#), é apresentado o uso de Redes com a finalidade de análise de associação para exploração visual de conjuntos de dados científicos multivariados, sendo uma abordagem de fácil compreensão ao usuário.

Após os Grafos, o tipo de Rede mais apresentado é a árvore. Foi detectado o uso dessa estrutura em todas as fases da mineração. No pré-processamento ([VALLE; RUZ; MORRÁS, 2018](#); [THONGKAM; SUKMAK, 2017](#); [SHEU et al., 2016](#); [ZHANG et al., 2013](#)), as pesquisas relatam a criação de árvores para auxílio na formação dos *itemsets* frequentes, sendo um passo fundamental e uma forma eficaz de identificar as inter-relações inerentes e desconhecidas entre as características de dados de espectros celestes e suas propriedades físico-químicas.

A maioria dos artigos que utilizaram árvores, o fizeram na fase de extração ([ANAND; VINODCHANDRA, 2018](#); [SHEU, 2018](#); [WANG et al., 2018](#); [GAO et al., 2017b](#); [MENG; SHA, 2017](#); [OLIINYK et al., 2017](#); [TURGEMAN; MAY; SCIULLI, 2017](#); [MONDAL; BHATTACHARYA; MONDAL, 2016](#); [NGUYEN; VO; HONG, 2015](#); [WEN; ZHONG; WANG, 2015](#); [NIRMALA; PALANISAMY, 2013](#)). As árvores foram utilizadas como estrutura base, nas quais os nós eram os *itemsets* frequentes e as regras eram obtidas percorrendo a árvore.

Na fase de pós-processamento da mineração ([CHERN-TONG; AZIZ, 2018](#); [MARTÍNEZ-BALLESTEROS et al., 2017](#); [LI et al., 2017](#); [MINELGA et al., 2017](#); [MEI; JIANG, 2016](#); [NGUYEN et al., 2015](#); [NGUYEN; NGUYEN, 2015](#); [PAPER et al., 2015](#); [SELMANE; BOUSSAID; BENTAYEB, 2015](#)), a estrutura em árvore foi utilizada a fim de facilitar a extração do conhecimento por meio da análise das relações entre as regras extraídas dos dados.

Outra estrutura de destaque nesse mapeamento são as Redes Bayesianas, que possuem uma abordagem estatística em sua formação. As Redes Bayesianas foram aplicadas nas fases de extração e pós-processamento ([DELGADO et al., 2018](#); [CAI et al., 2017](#); [SHENG et al., 2016](#); [NEMATI; SANT'ANNA; NOWACZYK, 2016](#); [HERNANDEZ-LEAL et al., 2013](#); [ODEH; AL-NAJDAWI, 2013](#); [VEDULA; THATAVARTI, 2011](#); [ZHAO; ZHANG; ZHANG, 2011](#)).

Redes Semânticas ([MALLAT et al., 2015](#); [KE et al., 2014](#)) também tiveram suas abordagens evidenciadas, além de Redes Biológicas ([HONG et al., 2018](#); [FU et al., 2017](#); [TSAI et al., 2017](#); [Özgür Cingiz; BIRICIK; DIRI, 2017](#); [DASGUPTA; SAHA, 2016](#); [BELYI et al., 2016](#); [SOROKINA; MEDIGUE; VALLENET, 2015](#); [MORE et al., 2011](#)) que embora seja uma aplicação e não uma modelagem proposta, ainda é definida como uma Rede.

Alguns artigos utilizam Redes em mais de uma fase ([KIM et al., 2016](#); [IQBAL et al., 2014](#); [KARIMI-MAJD; MAHOOTCHI, 2014](#)) aplicando as técnicas nas fases de extração e

pós-processamento ou mesmo utilizando uma abordagem de Redes nas três fases da mineração (RASHID; GONDAL; KAMRUZZAMAN, 2015).

Outra questão importante neste mapeamento diz respeito aos tipos de tarefa em que as Redes são utilizadas em apoio ao processo de mineração de dados. Destacam-se os Grafos Simples nas tarefas de agrupamento e análise, e as árvores nas tarefas de classificação. Como o mapeamento foi orientado para Regras de Associação, nenhuma pesquisa encontrada nas tarefas de regressão foi obtido, o que atendeu aos resultados esperados.

Quanto às tarefas de agrupamento (VALLE; RUZ; MORRÁS, 2018; ZHANG *et al.*, 2018; WANG; XU, 2018; WANG; GUO; LI, 2017; YU *et al.*, 2017; TSAI *et al.*, 2017; FU *et al.*, 2017; MARTÍNEZ-BALLESTEROS *et al.*, 2017; WANG *et al.*, 2017; SLAIMI *et al.*, 2017; MONDAL; BHATTACHARYA; MONDAL, 2016; ZHENWEI; LINGYUN; LIZHU, 2016; RUSH *et al.*, 2016; CHEN; XING, 2016; LI-XIONG *et al.*, 2015; RODRIGUES; GAMA; FERREIRA, 2012; AL-KHASSAWNEH; BAKAR; ZAINUDIN, 2012; RAJPUT; THAKUR; THAKUR, 2012; CHOBDAR; SILVA; RIBEIRO, 2011; ALZOUBI; OMAR; BAKAR, 2011), várias abordagens fazem uso das Redes na etapa de pós-processamento da mineração. Destaca-se os trabalhos de Karimi-Majd e Mahootchi (2014) que efetua, na etapa de pós-processamento da mineração, o uso de Redes para auxílio no agrupamento e mapeamento das regras. Os autores geram para cada *cluster*, um novo tipo de Grafo, conhecido como sub-grafo neste artigo. Cada sub-grafo tem mensagens exclusivas que podem contribuir para a concepção de novos serviços estudados na pesquisa.

Valle, Ruz e Morrás (2018) apresentam uma metodologia para análise de cesta de compras baseada em árvores, que complementa a busca de Regras de Associação. Na pesquisa é demonstrada a construção de uma estrutura de árvore hierárquica que encontrar fortes interdependências entre os produtos da mesma categoria formando os grupos por correlação direta.

Para as tarefas de classificação, os trabalhos apresentaram uma maior variedade de tipos de Rede (HONG *et al.*, 2018; CHERN-TONG; AZIZ, 2018; SHEU, 2018; DELGADO *et al.*, 2018; CHEN *et al.*, 2018; TURGEMAN; MAY; SCIULLI, 2017; MINELGA *et al.*, 2017; OLINYK *et al.*, 2017; GAO *et al.*, 2017a; THONGKAM; SUKMAK, 2017; LI *et al.*, 2017; MAKINO; KATO; TANIMOTO, 2017; PADUA; CARVALHO; REZENDE, 2016; NEMATI; SANT'ANNA; NOWACZYK, 2016; SIMARD; ST-PIERRE; BISKRI, 2016; MEI; JIANG, 2016; KIM *et al.*, 2016; SHENG *et al.*, 2016; USMAN; USMAN, 2016; SHEU *et al.*, 2016; ACHARYA; MAHALI; MOHAPATRA, 2015; NGUYEN *et al.*, 2015; PADUA; REZENDE; CARVALHO, 2014; KE *et al.*, 2014; ZHAO; ZHANG; ZHANG, 2011), dando destaques ao uso de Regras de Associação em classe, bem como Regras de Associação temporais e contínuas.

A classificação é uma tarefa muito importante na Mineração de Regras de Associação, destaca-se o trabalho de Chern-tong e Aziz (2018) que propõe uma nova técnica de poda para árvore de decisão usando algoritmo genético. Nguyen *et al.* (2015) propuseram uma nova

abordagem para a Mineração de Regras de Associação em classe (CAR - do inglês *Class Association Rule*) relevantes e que considera restrições à regra consequente. Uma estrutura de árvore para armazenar o conjuntos de *itemsets* frequentes do conjunto de dados foi projetada. Em seguida, alguns teoremas para a poda dos nós da árvore que não pode gerar regras que satisfaçam as restrições de classe são fornecidos e provados. Finalmente, um algoritmo eficiente para mineração de CARs é apresentado, além de comparar com outros métodos já existentes.

Nas tarefas de análise exploratória vários tipos de Redes foram utilizadas, bem como a forma na qual cada uma delas se apresentou (WANG *et al.*, 2018; ANAND; VINODCHANDRA, 2018; SU, 2017; Özgür Cingiz; BIRICIK; DIRI, 2017; KIM *et al.*, 2017; INDHUMATHY; NABHAN; ARUMUGAM, 2017; CAI *et al.*, 2017; NAMAKI *et al.*, 2017; FAN; HU, 2017; MENG; SHA, 2017; ADARICHEVA; NATION, 2017; LEE; SHIN, 2017; GAO *et al.*, 2017b; SHAHMOHAMMADI; KHADANGI; BAGHERI, 2016; BELYI *et al.*, 2016; LIU; SHEN, 2016; FAN; WU; XU, 2016; WILKE *et al.*, 2016; DASGUPTA; SAHA, 2016; RAWAT *et al.*, 2016; ABAY; MUTLU; KARAGOZ, 2016; DU *et al.*, 2015; LI; DU; YANG, 2015; HENDRICKX *et al.*, 2015; NADI *et al.*, 2014; RANJAN; VAISH; COULL, 2014; LIU; ZHAI; PEDRYCZ, 2012; MORE *et al.*, 2011; VEDULA; THATAVARTI, 2011).

Como tarefa de análise dos dados destaca-se o trabalho de Liu, Zhai e Pedrycz (2012) que apresenta uma forma de Grafos Direcionados de *itemsets* para armazenar as informações de conjuntos de *itemsets* frequentes de bancos de dados de transação, além de fornecer a estrutura da lista de armazenamento para o Grafo. Além disso, os autores desenvolveram o algoritmo de mineração de conjuntos de *itemsets* frequentes máximos baseadas nesta estrutura. Rawat *et al.* (2016) apresentam uma abordagem para a Mineração de Regras de Associação multinível por meio de técnicas de busca *And-Or* que efetuam a poda dos nós com um custo computacional reduzido.

Apenas o trabalho de Rodrigues, Gama e Ferreira (2012) apresentou uma Rede de Regras de Associação, porém com uma abordagem diferente da apresentada por Pandey *et al.* (2009). Neste trabalho, a Rede é construída para demonstrar apenas os relacionamentos entre os *itemsets*. A elaboração da Rede é feita baseada em análise de Redes Sociais além de não ser construída de modo direcionado a um item objetivo.

Um *gap* de pesquisa foi detectado no Mapeamento realizado:

- Nenhum trabalho apresentou uma abordagem multialvo utilizando algum tipo de Rede no auxílio da Mineração de Regras de Associação. Embora pode-se considerar que as abordagens de classificação designam os atributos em, no mínimo, duas categorias, nenhum trabalho abordou diretamente mais de uma regra como objetivo dos estudos realizados.

Com isso percebeu-se um vasto campo de pesquisa ainda a ser explorado com o uso de técnicas otimizadas de ARNs e abordagens multiobjetivo.

2.5.8 Ameaça à Validade do Mapeamento Sistemático

Existem muitos fatores que podem ameaçar a validade de um Mapeamento Sistemático de Estudos. Alguns desse fatores foram identificados com o objetivo de mitigar suas influências nos resultados desse estudo.

A primeira ameaça refere-se à identificação dos estudos primários, pois a utilização dos mecanismos de busca automatizada não garante a inclusão de todos os trabalhos relevantes, e ainda há o problema da indexação de trabalhos recentes. Essa ameaça foi mitigada com a realização de uma pesquisa piloto para verificar a consistência dos resultados obtidos nas bibliotecas digitais e a análise das referências dos artigos selecionados para possível inclusão de novos trabalhos.

A formação da *string* de pesquisa também pode ser considerada como uma ameaça à validade deste mapeamento sistemático. Há uma vasta quantidade de termos que poderiam ser incluídos para caracterizar uma fase da Mineração de Regras de Associação, como “*frequent itemsets*”, por exemplo. Entretanto, ao realizar uma pesquisa piloto nas bibliotecas digitais, verificou-se que o resultado tornaria a pesquisa inviável devido à grande quantidade de trabalhos retornados. Dessa forma, decidiu-se por termos centrados em “*Association Rules*”. Essa decisão pode ter influenciado a não inclusão de trabalhos de pré-processamento na Mineração de Regras de Associação, entretanto foi possível mapear o uso das Redes nas etapas de pós-processamento, possibilitando uma adequação maior dos objetivos da MSE com os deste doutorado.

Outro fator que pode ameaçar a validade dessa pesquisa diz respeito a seleção dos estudos e a extração dos dados. Essas etapas podem ser enviesadas pela opinião pessoal dos pesquisadores que estão executando o processo, por isso buscou-se critérios objetivos de inclusão e exclusão para minimizar o impacto da subjetividade dos pesquisadores.

2.6 Considerações Finais

A tarefa de associação tem como premissa básica encontrar elementos que implicam na presença de outros elementos em uma mesma transação, ou seja, encontrar relacionamentos ou padrões frequentes entre conjuntos de dados. O termo transação indica quais itens foram consultados em uma determinada operação de consulta.

A descoberta de Regras de Associação é uma técnica de extração de conhecimento com bastante aplicabilidade a questões concretas, mas que coloca algumas dificuldades ao nível da sua análise, principalmente se as regras não sofrerem nenhum tipo de adequação na fase de pós-processamento.

As Regras de Associação podem ser aplicadas em várias tarefas da mineração de dados, portanto, a adequação dos processos para obtenção dessas regras deve ser considerada, já que o custo computacional, velocidade de processamento e tempo de resposta são essenciais.

Neste capítulo foram discutidos temas relacionados à Mineração de Regras de Associação. Foram apresentados os conceitos básicos sobre Regras de Associação, que são essenciais para a contextualização deste trabalho. Foi dado enfoque especial para o uso de Redes em alguma etapa da Mineração de Regras de Associação, tema de estudo deste trabalho, e em especial às Redes de Regras de Associação (ARNs) para a extração de conhecimento por meio da elaboração de hipóteses. Um dos desafios da Mineração de Regras de Associação é a garantia da extração de padrões de interesse ao usuário. Isso foi destacado com a breve apresentação da estrutura das ARNs.

Um Mapeamento Sistemático da literatura realizado com o objetivo de obter uma visão geral sobre uso de Redes aliados às Regras de Associação, cujos resultados foram apresentados neste capítulo, indica a diversidade de Redes envolvidas no processo de Mineração de Regras de Associação, bem como as suas aplicabilidades em várias tarefas e nas três etapas (pré-processamento, extração, pós-processamento) da mineração. Este Mapeamento Sistemático de Estudos teve por objetivo principal verificar como as Redes podem auxiliar nas fases de Mineração de Regras de Associação, bem como em quais tarefas de mineração elas são aplicadas. Os tipos de Redes foram listados em conformidade com as características extraídas de cada trabalho, com isso foi possível identificar o estado da arte em pesquisas nessa área. No entanto, esse estudo da literatura também mostrou que os trabalhos se concentram em tarefas específicas de estudo geral, sem o direcionamento a um ou mais itens alvos, nem com o destaque objetivo das medidas utilizadas nas tarefas de mineração.

Finalmente, acredita-se que o uso da Rede de Regras de Associação (ARN) possa auxiliar as diversas fases da Mineração de Regras de Associação, principalmente no pós-processamento, etapa na qual as Redes são aplicadas na maioria dos trabalhos, bem como serem utilizadas em diversas tarefas de mineração. Além disso, uma abordagem multialvo pode ser implementada a fim de potencializar ainda mais a extração do conhecimento

Com o objetivo de apresentar as tarefas de Mineração de Regras de Associação, neste capítulo foi identificada uma lacuna relevante que deu origem à proposta desta tese. Foram analisados aspectos de validação e como estruturas de Rede podem auxiliar na geração do conhecimento (hipóteses) obtido nas tarefas de Mineração de Regras de Associação. A geração de hipóteses apresenta fatores subjetivos e sem a garantia de que os resultados sejam otimizados.

A fim de resolver tais problemas, é necessário que as abordagens de mineração façam uso de técnicas objetivas para mensuração da relevância e veracidade das hipóteses, além de garantir estudos direcionados a mais de um item alvo. Neste trabalho, são propostos e avaliados diferentes técnicas de Mineração de Regras de Associação com uso de Redes. Os modelos propostos e a avaliação dos mesmos são apresentados nos próximos capítulos desta tese.

RELAÇÃO ENTRE MEDIDAS DE CENTRALIDADE E MEDIDAS DE REGRAS DE ASSOCIAÇÃO

3.1 Considerações Iniciais

Medidas de interesse são muito importantes para a Mineração de Regras de Associação. Como os conjuntos de dados geralmente estão aumentando, enumerar todas as combinações possíveis de LHS e RHS e depois verificar sua correlação não é computacionalmente viável (YAN; ZHANG; ZHANG, 2009). O número de regras descobertas pode ser tão elevado, que quase transforma a sua interpretação em um novo problema de mineração (DOMINGUES; REZENDE, 2005).

A poda de regras que não satisfazem o objetivo desejado é necessária em muitos casos. Para isso deve-se escolher métricas capazes de avaliar a qualidade e classificar as Regras de Associação extraídas. Assim, muitas pesquisas têm sido dedicadas a definir medidas apropriadas de interesse e métodos para realizar a melhor escolha dos valores destas mesmas medidas (BALCÁZAR, 2013).

Como os algoritmos de Regras de Associação são capazes de extrair as regras de acordo com um suporte mínimo (*minsup*) e valor mínimo de confiança (*minconf*), a escolha dos valores desses parâmetros é de suma importância. Várias abordagens foram propostas para guiar o usuário na exploração de regras (AGRAWAL; IMIELINSKI; SWAMI, 1994; AGRAWAL; SRIKANT, 1994; AGRAWAL; SHAFER, 1996; GRAHNE; ZHU, 2005). No entanto, os principais algoritmos de extração de Regras de Associação se concentram na exploração de acordo com um processo subjetivo de escolha de valores para os parâmetros de seleção, sem a garantia de produção dos melhores resultados. A escolha dos parâmetros limitadores de suporte e confiança mínimos continuam não sendo justificados, tornando uma das etapas do processo totalmente

dependente da escolha do usuário, mesmo em abordagens que utilizam Medidas de Rede como auxiliares no processo de Mineração de Regras de Associação (WANG; XU, 2018).

Com estas observações levantou-se as questões de pesquisa: quais as relações que existem entre métricas relacionadas a Regras de Associação e medidas relacionadas a estruturas de Rede? Existe alguma relação que possa ser utilizada para obtenção de uma abordagem objetiva para seleção das Regras de Associação?

Com o intuito de responder a estas questões, neste capítulo, é proposta uma metodologia para avaliação dos valores das principais medidas utilizadas no processo de Mineração de Regras de Associação (suporte e confiança mínimos) com o uso de Redes de Regras de Associação e Medidas de Centralidade. Redes de Regras de Associação (ARNs) possuem uma estrutura para sintetizar, podar, e analisar um conjunto de Regras de Associação para a construção de hipóteses candidatas, porém o processo de seleção das Regras continua sendo feito de modo subjetivo e sem relação nenhuma com o conjunto de dados que está sendo analisado.

De uma perspectiva de descoberta de conhecimento, a ARN possibilita uma análise orientada para o contexto, o que facilita o entendimento dos relacionamentos e a utilidade das informações dos dados (PANDEY *et al.*, 2009), portanto é proposto nesse trabalho o uso de Medidas de Centralidade para auxiliar no processo de avaliação das medidas de suporte e confiança mínimos utilizados na extração das Regras que irão formar as ARNs, bem como uma nova medida de avaliação das próprias ARNs, a fim de garantir uma geração de hipóteses otimizada. A ideia central da ARN é que as Regras de Associação descobertas pelo algoritmo de mineração podem ser sintetizadas, podadas, e integradas no contexto de objetivos específicos da pesquisa. Conceitualmente, se houver um atributo alvo de interesse, pode-se formar uma Rede com as variáveis mais relevantes.

Na Seção 3.2 é fornecida uma visão geral de trabalhos que relacionam Redes a processos de Mineração de Regras de Associação, bem como a metodologia aplicada para responder as questões de pesquisa. É apresentada também a proposta de uma nova medida de Relevância aplicada a Redes de Regras de Associação (ARNs). Na Seção 3.3 é descrita a proposta do uso de medidas de Rede para avaliação das métricas de suporte e confiança mínimos, bem como a avaliação experimental com 3 *datasets* oriundos da UCI. Por fim, na Seção 3.4 todo o processo da Prova de Conceito da proposta com uma base de dados real é demonstrado.

3.2 Modelo Proposto para Cálculo da Relevância de Redes de Regras de Associação

Em geral, uma Rede R é representada como $R = (V, E)$, na qual V é um conjunto de vértices (ou nós) e E é um conjunto de arestas (ou links), que ligam pares de vértices em V . Estatisticamente, uma Rede pode ser caracterizada por valores derivados, como o grau médio

dos nós e o comprimento médio (caminho) entre os nós (NEWMAN, 2010). De acordo com a aplicação da Rede, outras características adicionais como diâmetro da rede, número de triângulos e o coeficiente de agrupamento, também podem ser analisadas (NETTLETON, 2013).

Liu, Zhai e Pedrycz (2012) afirmam que a etapa da Mineração de Regras de Associação que recebe maior atenção para o uso de Redes pela comunidade de pesquisadores é a de pré-processamento. Esta etapa visa utilizar a modelagem em Rede para remover itens e conexões não interessantes, ou até mesmo para otimizar a formação dos *itemsets* frequentes, reduzindo a quantidade de itens a serem explorados pelo usuário. No entanto, na fase seguinte de Extração das Regras, não existe um método objetivo de seleção das medidas de suporte mínimo e confiança mínima a fim de garantir uma melhor extração dos padrões.

Na pesquisa de Yan, Zhang e Zhang (2005), são utilizadas estratégias de mineração evolutivas para direcionar a pesquisa das medidas de acordo com as regras geradas, a fim de mitigar os problemas relacionados com a escolha das medidas de suporte e confiança mínimos. Posteriormente, Yan, Zhang e Zhang (2009) propuseram uma nova medida de confiança simples para medir o interesse das Regras de Associação. Essa medida distingue-se de outras medidas de confiança existentes em termos de significado probabilístico. O cálculo da medida é realizado com o uso da medida *Lift* de probabilidade das Regras de Associação a fim de examinar as correlações positivamente complementares entre os conjuntos de itens frequentes. Porém a medida *Lift* não consegue representar a influência entre antecedente e consequente da regra por ser uma Medida Objetiva Simétrica.

Conforme mencionado na Seção 2.4, As Redes de Regras de Associação (ARNs) efetuam a união da utilização de Redes com o processo de Mineração de Regras de Associação (PANDEY *et al.*, 2009). Utilizando a estrutura das ARNs e as característica das Medidas Objetivas Assimétricas, é proposta neste trabalho a medida de Relevância da ARN por meio da utilização da medida *Added Value* (ver Equação 2.3), com a finalidade de otimizar a geração de hipóteses por meio das regras com relações de influência comprovada.

Pandey *et al.* (2009) apresenta as ARNs como Redes Direcionadas que são construídas de acordo com um item objetivo do conjunto de dados que está sendo analisado. Desta forma, opta-se pelo uso de Medidas Objetivas Assimétricas, em especial a medida de *Added Value*, por considerarem a ordem dos elementos da regra ($AV(A \Rightarrow B) \neq AV(B \Rightarrow A)$).

A fim de se obter um método quantitativo para comparação das Redes, é proposta nesta tese a medida de Relevância de uma ARN. A Relevância da Rede é calculada com uso da medida *Added Value* de todas as arestas conectadas diretamente ao nó objetivo da ARN.

Conjectura 1 - Relevância de ARNs: a Relevância de uma ARN é estabelecida como o somatório dos valores absolutos de *Added Value* (AV) calculados para todas as arestas conectadas diretamente ao nó objetivo multiplicado pelo número de arestas com $AV \neq 0$

(Equação 3.1).

$$Relevancia = \left(\sum_{i=1}^n |AV(i)| \right) * N \quad (3.1)$$

em que $AV(i)$ indica o valor de *Added Value* de cada aresta i conectada ao nó objetivo da ARN, n representa o grau do nó objetivo e N o número de arestas com AV não nulo.

Observando, por exemplo, a ARN da Figura 23, pode-se inferir que a Relevância da Rede é 8,262. O somatório dos módulos das medidas *Added Value* conectadas ao item objetivo “[lenses]=no” formam $Relevancia = (|0,125| + |0,042| + |-0,125| + |-0,042| + |1,11E-12| + |0,042| + |0,125| + |-0,042| + |0,375|) * 9 = 8,262$.

Figura 23 – Medidas *Added Value* (ver Equação 2.3) de arestas conectadas ao item objetivo “[lenses]=no” da ARN

Fonte: Elaborada pelo autor.

Conforme apresentado na Seção 2.2.2, a medida de *Added Value* mede o ganho de RHS na presença de LHS, podendo apresentar resultados negativos e positivos para as Regras. Desta forma, a Equação 3.1 é formulada pelo módulo de *Added Value* ($|AV|$) para que todas as influências sejam acrescidas na Relevância da Rede. Como fator de ponderação (peso) optou-se pela quantidade de arestas com $AV \neq 0$, pois são as conexões que apresentam uma influência real entre o antecedente e o consequente da Regra.

Exemplo Uma ARN R possui apenas duas arestas conectadas ao item alvo. As arestas apresentam valores $AV = 0,315$ e $AV = -0,315$, nesse caso a Relevância da Rede seria nulo pelo uso dos valores das medidas *Added Value* apresentados. Utilizando os valores absolutos das medidas, a Relevância da Rede torna-se $Relevancia = 1,26$ ($Relevancia = (|0,315| + |-0,315|) * 2$), descrevendo melhor a influência relacionada ao nó alvo.

Com o auxílio da medida de Relevância da ARN, foram feitas as análises das Medidas de Rede encontradas nas ARNs. As correlações com as medidas de suporte e confiança mínimos geraram a proposta apresentada na Seção 3.4 deste trabalho e responde diretamente as questões de pesquisa iniciais.

Com o objetivo de analisar o comportamento das Medidas de Rede foram realizados alguns experimentos em *datasets* conhecidos. Esses experimentos possuíram a finalidade de estudar o comportamento dessas medidas de acordo com a variação do suporte e confiança mínimos, gerando uma compreensão completa da relação entre estes itens.

3.3 Estudo das Medidas de Rede Associados a Suporte e Confiança Mínimos - Avaliação Experimental

Para a escolha dos conjuntos de dados que foram utilizados nos experimentos desta pesquisa, foi elaborada uma proposta para o cálculo da Taxa de Complexidade de *datasets* utilizando os dados encontrados em Gupta (2015). Na presente pesquisa o autor realizou a comparação de 9 algoritmos de aprendizado de máquina com 11 *datasets* oriundos da UCI. Com os resultados gerados neste trabalho (Apêndice B), foi proposto um cálculo que possibilita comparar o grau de complexidade de cada base de dados de acordo com a definição a seguir.

Proposta 2 - Taxa de Complexidade: a taxa de complexidade de um *dataset* é calculada pela razão entre a média aritmética do Tempo de Complexidade (T_c) (tempo entre o início e o fim da execução) gerado por algoritmos de Aprendizado de Máquina e a média aritmética da Acurácia (Acc) destes mesmos algoritmos multiplicado pelo fator de correção 1000 (Equação 3.2).

$$TxComplexidade = \frac{T_c}{Acc} * 1000 \quad (3.2)$$

Assim, os 11 *datasets* da UCI foram classificadas de acordo com a Tabela 3.

Nos experimentos realizados nesse trabalho, as regras foram extraídas com o uso do algoritmo *Apriori-TID* implementado em Java®. As medidas de suporte mínimo foram definidas como {0,0, 0,1, 0,2, 0,3, 0,4, 0,5} e de confiança mínimos {0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9}. A variação do suporte mínimo é menor porque valores acima de 0,5 geravam um número muito pequeno de Regras. Como foram selecionados 6 valores para *minsup* e 9 para *minconf*, pelo uso do princípio fundamental da contagem foram totalizados 54 experimentos para cada conjunto de dados.

Após a extração das regras, foram construídas as ARNs, pelo método de Pandey et al. (2009), utilizando cada medida de suporte e confiança mínimos considerados. Com as Redes estruturadas, foram calculadas as seguintes Medidas: Grau Médio, Grau Ponderado Médio

Tabela 3 – Taxa de Complexidade

<i>Dataset</i>	Taxa de Complexidade
LENSES	1,804
LABOR	2,305
IRIS	3,065
LUNG CANCER	4,765
VOTE	5,355
HAYES-ROTH	5,686
TEACHING ASSISTANT	6,030
STATLOG	8,256
GLASS IDENTIFICATION	8,897
SOYBEAN LARGE	123,469
ABALONE	1192,960

Fonte: Elaborada pelo autor.

(utilizando-se a confiança da regra como peso da aresta), Diâmetro, Densidade e Comprimento Médio do Caminho, além de duas medidas relacionadas com o nó objetivo (Autovetor e *Page-Rank*), bem como a Relevância da ARN.

Foi elaborada uma Avaliação Experimental com 3 *datasets* para estudo da relação entre métricas limitadoras de suporte e confiança em processos de extração de Regras de Associação e Medidas de Rede. Foram selecionados três *datasets*, um de baixa complexidade, um de média complexidade e outro de alta complexidade (Tabela 3). Para cada estudo foram variados os valores de suporte e confiança mínimos e calculadas as Medidas de Rede de cada Rede de Regra de Associação construída.

No primeiro experimento realizado, foi utilizado o *dataset Lenses*¹, disponível na UCI. Nesse conjunto de dados, cada instância representa os atributos de um paciente e a lente de contato que foi prescrita para ele. O objetivo principal do *dataset* é descrever quais características implicam na prescrição de cada tipo de lente de contato.

O segundo experimento foi feito com o *dataset Hayes-roth*², também disponível na UCI. O conjunto de dados traz informações geradas sobre características pessoais como idade e nível de escolaridade. Essa base de dados contém 132 instâncias e 6 atributos de valor numérico, sendo o primeiro apenas um índice, o qual foi excluído. O último atributo divide as instâncias em 3 classes.

O terceiro estudo foi feito no *dataset Soybean (Large)*³ por se tratar de um conjunto de dados com alto grau de complexidade. O *dataset* relata os resultados de uma pesquisa de características de doenças em plantações de soja. Existem 19 classes, apenas as 15 primeiras foram usadas, pois as últimas quatro classes são pouco exploradas por apresentarem poucos

¹ <<https://archive.ics.uci.edu/ml/datasets/lenses>>

² <<https://archive.ics.uci.edu/ml/datasets/Hayes-Roth>>

³ <<https://archive.ics.uci.edu/ml/datasets/Soybean+%28Large%29>>

exemplos. Existem 35 atributos categóricos, alguns nominais e alguns ordenados. Os valores dos atributos são codificados numericamente, com o primeiro valor codificado como “0”, o segundo como “1”, e assim por diante. O *dataset* possui campos “vazios”.

3.3.1 Lenses dataset

O *dataset* “Lenses” possui apenas 24 instâncias e 5 atributos, sendo que um deles indica o tipo de lente de contato indicado ao paciente. As regras foram extraídas variando as medidas de suporte e confiança mínimos conforme descritos anteriormente. O tamanho da regra foi especificado em dois, considerando um item no LHS e um item no RHS conforme a metodologia para construção de Redes de Regras de Associação. Usando essa configuração, foram obtidas Regras de Associação para o *dataset* “Lenses” e construídas as ARNs com nó objetivo “[lenses]=no” por se tratar da classe com o maior número de instâncias.

Para análise do impacto de *minsup*, foram fixados os valores da confiança mínima e variados os valores do suporte mínimo de 0,0 a 0,5. As Relevâncias das ARNs foram calculadas (Tabela 4), bem como as outras Medidas de Rede.

A Rede com maior Relevância (Equação 3.1) e menor suporte mínimo foi obtida para o valor de *minsup* = 0,0. Ao analisar as Relevâncias das ARNs, percebe-se que a variação foi nula. Então, a Medida de Relevância não sofre uma variabilidade direta pela mudança do suporte mínimo na fase de extração das Regras. A seleção do menor *minsup* justifica-se para que o maior número de regras possa ser analisado.

Tabela 4 – Relevância x Suporte Mínimo (*Lenses dataset*)

<i>minsup</i>	Relevância
0,0	6,3312
0,1	6,3312
0,2	6,3312
0,3	6,3312
0,4	2,7083
0,5	2,7083

Fonte: Elaborada pelo autor.

As Redes resultantes apresentaram comportamentos diferentes para as Medidas de Rede. As variabilidades das medidas de Diâmetro da Rede e Centralidade de Autovetor do nó objetivo para cada ARN foram irrelevantes, sendo em sua maioria nula. Desta forma, essas medidas não apresentam relação com métricas de suporte e confiança mínimos. Na Figura 24 é apresentado um gráfico com os resultados para as medidas de Grau Médio, Grau Ponderado Médio (utilizando-se a confiança da regra como peso da aresta), Densidade, Comprimento Médio do Caminho e *PageRank* do nó alvo para o valor de *minconf* = 0,3.

Figura 24 – Relação entre Medidas de Rede e Suporte Mínimo com $minconf = 0,3$ (*Lenses dataset*)

Fonte: Elaborada pelo autor.

Com a finalidade de verificar o impacto de $minconf$ nas Medidas de Rede, foi fixado o suporte mínimo em 0,0 e foram variados os valores de confiança mínima de 0,1 a 0,9. Para cada Rede resultante foram calculados os valores das Medidas de Rede, com exceção de Diâmetro e Autovetor do nó alvo por não apresentarem variação significativa.

As ARN resultantes apresentaram as Relevâncias descritas na Tabela 5. A variação da confiança mínima não interfere diretamente nos valores da Relevância uma vez que confianças mínimas diferentes proporcionam Redes de mesma Relevância.

Tabela 5 – Relevância x Confiança Mínima (*Lenses dataset*)

$minconf$	Relevância
0,1	1,1664
0,2	1,1664
0,3	6,3312
0,4	6,3312
0,5	6,3312
0,6	2,916
0,7	1,0
0,8	0,375
0,9	0,375

Fonte: Elaborada pelo autor.

Com as outras Medidas de Rede, as Redes resultantes descreveram as variações mostradas na Figura 25. Observa-se que as Medidas de Grau Ponderado Médio e o Comprimento Médio do Caminho coincidem exatamente no ponto da Rede de Relevância máxima e menor $minconf$.

Figura 25 – Relação entre Medidas de Rede e Confiança Mínima com $minsup = 0,0$ (*Lenses dataset*)

Fonte: Elaborada pelo autor.

3.3.2 Hayes-roth dataset

Foram extraídas as regras de tamanho dois variando os valores de suporte e confiança mínimos. Após a etapa de geração das regras, foram construídas ARNs com nó objetivo “[class]=3” por se tratar da classe com o maior número de instâncias. Foram calculadas as respectivas Medidas de Rede estudadas nesta pesquisa.

A Relevância das ARNs com a alteração dos suportes mínimos não apresentou variação, portanto não se pode estabelecer uma relação entre a Medida de Relevância e $minsup$ nesse caso.

Com o intuito de verificar o impacto de $minconf$ nas Medidas de Rede, foi fixado o suporte mínimo em 0,0 e foram variados os valores de confiança mínima de 0,1 a 0,9. Para cada Rede resultante foram calculados os valores das Medidas de Rede, com exceção de Diâmetro e Autovetor do nó alvo por não apresentarem variação significativa.

A ARN com maior Relevância foi obtida quando $minconf = 0,1$ (Tabela 6). Para o *dataset Hayes-roth* as Redes com Relevância mais elevadas foram obtidas com valores baixos de confiança mínima. A partir $minconf = 0,3$, todas as Redes apresentaram o mesmo valor de Relevância, portanto não pode-se relacionar a medida de confiança mínima com a medida de Relevância de ARNs.

Na Figura 26 é apresentado o comportamento das medidas de Rede para $minsup = 0,0$. Devido ao comportamento da medida de Relevância, são apresentados os resultados apenas das ARNs que possuíam $minconf \leq 0,3$.

Observado a variabilidade das Medidas de Rede, infere-se que todas apresentaram um intervalo homogêneo entre si, com exceção da medida de Grau Médio. Todas as medidas estudadas apresentaram intervalos de aproximação similares. Estudando o comportamento das medidas destacadas no experimento com o conjunto de dados *Hayes-roth*, percebe-se uma

Tabela 6 – Relevância x Confiança Mínima (*Haye-roth dataset*)

<i>minconf</i>	Relevância
0,1	54,0818
0,2	27,7317
0,3	6,9545
0,4	6,9545
0,5	6,9545
0,6	6,9545
0,7	6,9545
0,8	6,9545
0,9	6,9545

Fonte: Elaborada pelo autor.

Figura 26 – Relação entre Medidas de Rede e Confiança Mínima com *minsup* = 0,0 (*Hayes-roth dataset*)

Fonte: Elaborada pelo autor.

aproximação das medidas de Grau Ponderado Médio e Comprimento Médio do Caminho quando *minconf* = 0,1. A diferença entre essas Medidas de Rede aumenta para *minconf* = 0,2, esses valores de confiança mínima correspondem às Redes de maiores Relevâncias (Tabela 6).

3.3.3 Soybean Large dataset

Uma vez que o *dataset* possui 19 classes, 36 atributos e 347 instâncias, sendo algumas com campos vazios, foi feita a exclusão de 40 instâncias com valores nulos. Foram extraídas as Regras de Associação com tamanho dois, considerando um item no LHS e um item no RHS. Usando essa configuração, foram obtidas Regras de Associação e construídas as ARNs com nó objetivo “[class]=frog-eye-leaf-spot” por se tratar de uma das classes com o maior número de instâncias.

Executando o experimento de análise do suporte mínimo e as Medidas de Rede, percebeu-se que a medida de Relevância das ARNs não sofreu variação com a alteração da medida de *minsup*. A Rede de maior Relevância e menor suporte mínimo foi encontrada com *minsup* = 0,1. Sendo assim, definiu-se esse valor para o estudo da medida de confiança mínima.

Após esta etapa, foram extraídas as Regras variando os valores de *minconf*. As Redes resultantes apresentaram as Relevâncias apresentadas na Tabela 7. Observa-se que as Redes com *minconf* $\geq$ 0,6 apresentaram valores iguais ou nulos de Relevância e que a ARN com maior Relevância foi construída para *minconf* = 0,3. Percebe-se que a medida de Relevância não possui um comportamento padrão para a variação da medida de confiança mínima.

Tabela 7 – Relevância x Confiança Mínima (*Soybean Large dataset*)

<i>minconf</i>	Relevância
0,1	24,414
0,2	24,420
0,3	24,762
0,4	4,0545
0,5	2,16
0,6	0,6772
0,7	0,6772
0,8	0,6772
0,9	0,0

Fonte: Elaborada pelo autor.

Na Figura 27 pode-se observar o resultado da variação das Medidas de Rede. Apenas ARNs com *minconf* $\leq$ 0,6 são apresentados devido ao comportamento da medida de Relevância. As Medidas de Rede apresentaram comportamentos distintos de variação, porém Grau Ponderado Médio e o Comprimento Médio do Caminho apresentaram valores mais próximo quando a ARN possui maior Relevância. O mesmo comportamento foi observado nos experimentos anteriores com os *datasets Lenses* e *Hayes-roth*.

Nos experimentos realizados, o suporte mínimo, não apresentou relação direta com as Medidas de Rede. O suporte mínimo ideal é o que forma Redes de maior Relevância com o menor valor adotado, pois o maior número de correlações é estabelecido. Verificou-se que o valor ideal de *minsup* coincide com a porcentagem da classe objetivo em relação ao total de instâncias do *dataset* nos experimentos realizados. Por exemplo, no *dataset "Soybean"*, o suporte mínimo ideal foi de 0,1 (10%), o que é aproximadamente a porcentagem da quantidade de instâncias que possuem o nó objetivo selecionado ("[class]=frog-eye-leaf-spot") em relação ao total, nesse caso 40 instâncias do nó objetivo de 307 no total. O mesmo comportamento ocorre com os outros *datasets* estudados.

Ao se analisar o comportamento das Medidas de Rede com a variação da confiança mínima percebe-se que a medida de Relevância não sofre nenhum tipo de influência. As Redes

Figura 27 – Relação entre Medidas de Rede e Confiança Mínima com $minsup = 0,1$ (Soybean Large dataset)

Fonte: Elaborada pelo autor.

construídas após a extração de regras com confianças mínimas diferentes apresentam o mesmo valor de Relevância. A confiança mínima ideal é a que gera uma ARN de maior Relevância com o menor valor possível, porque considera um maior número de regras, que contém o item alvo, com influência comprovada ($AV \neq 0$). A confiança mínima ideal proporciona a criação de Redes com maior possibilidade de gerar hipóteses verdadeiras, otimizando a extração do conhecimento.

O Grau Ponderado Médio e o Comprimento Médio do Caminho são medidas que apresentam um indicativo de comportamento padrão e podem ser utilizados para avaliação da confiança mínima, mesmo em *datasets* mais complexos. Os resultados foram estudados e uma proposta de avaliação das medidas de $minsup$ e $minconf$ para extração de Regras de Associação foi formulada.

3.4 Método Proposto para Avaliação de $minsup$ e $minconf$ com o uso de Medidas de Centralidade

Com os resultados da Avaliação Experimental descrita na Seção 3.3, observa-se a possibilidade de se utilizar Medidas de Rede para criar critérios objetivos de seleção e avaliação dos parâmetros de suporte e confiança mínimos.

Proposta 3 - Método de seleção de $minsup$: o suporte mínimo ideal é calculado pelo suporte

da classe objetivo a ser estudada (Equação 3.3).

$$\text{minsup_ideal} = \text{sup}(\text{item_objetivo}) \quad (3.3)$$

Proposta 4 - Método de avaliação de *minconf*: a confiança mínima é avaliada pela diferença *DIF* (Equação 3.4) entre as medidas de Grau Ponderado Médio, com valores de confiança como pesos, e o Comprimento Médio do Caminho. Quando esta diferença aumenta, o valor de *minconf* ideal é o valor calculado na etapa anterior (Equação 3.5).

$$\text{DIF} = \text{Grau_Ponderado_Medio} - \text{Comprimento_Medio_do_Caminho} \quad (3.4)$$

$$\text{DIF}(i+1) > \text{DIF}(i) \rightarrow \text{minconf_ideal} = \text{minconf}(i) \quad (3.5)$$

Para o *dataset Lenses*, o valor de *DIF* (Equação 3.4) aumentou quando *minconf* passou de 0,3 para 0,4. Então a confiança mínima que gera uma ARN com maior probabilidade de gerar hipóteses verdadeiras é 0,3.

Avaliando os resultados obtidos por meio da Avaliação Experimental com o *dataset Hayes-roth*, a diferença entre as Medidas de Rede em questão aumenta quando *minconf* passa de 0,1 para 0,2, portanto o valor ideal da confiança mínima é 0,1. Para o *dataset Soybean Large*, a diferença aumentou quando *minconf* passou de 0,3 para 0,4, sendo que a ARN que possui maior possibilidade de uma extração de conhecimento otimizada foi gerada exatamente com a confiança mínima de 0,3.

Portanto, para a seleção do suporte mínimo é suficiente um estudo do número de instâncias do *dataset* e da quantidade referente ao item objetivo que será selecionado para a criação da ARN. Para a avaliação da confiança mínima e futura geração gráfica da Rede, pode-se utilizar as Medidas de Centralidade de Grau Ponderado Médio, com as confianças como pesos, e Comprimento Médio do Caminho.

3.5 Prova de Conceito

Para validação das propostas de seleção de suporte mínimo e avaliação de confiança mínima, foi executado uma Prova de Conceito com o uso de dados relacionados à agricultura orgânica (*Green Manure*). Todo o processo de mineração de dados foi realizado, desde a preparação e pré-processamento do *dataset*, bem como a extração de Regras de Associação e a construção das Redes para obtenção de conhecimento. O experimento também teve o objetivo de analisar quais as melhores condições para obtenção de adubos orgânicos com menor taxa de decomposição. As ARNs foram elaboradas a fim de que os padrões obtidos fossem comparados.

3.5.1 Geração de dataset

A coleta dos dados foi realizada no segundo semestre de 2015, na Embrapa Meio-Norte/UEP Parnaíba (3°05'S; 41°46'W e 46,8m altitude). O clima da região é do tipo C1dA'a' (THORNTWHAITE; MATHER, 1955), caracterizado como sub-úmido seco, megatérmico, com pequeno excedente hídrico e concentração de 29,7% de evapotranspiração potencial nos trimestres de outubro, novembro e dezembro, com umidade relativa média do ar de 77,5%, precipitação média de 1107mm e temperatura média de 27,6°C (BASTOS; JUNIOR; RODRIGUES, 2012).

O solo da área experimental foi descrito como um Latossolo Amarelo texturizado de textura média, fase de caatinga litorânea (MELO *et al.*, 2004). As amostras de solo foram coletadas em diferentes pontos da área experimental, a fim de obter uma maior uniformidade do substrato. A partir do material coletado, uma amostra composta foi coletada e posteriormente analisada no laboratório de água e solo da Embrapa Meio-Norte, UEP Parnaíba, para a adequada análise e caracterização química do solo.

Os dados de precipitação para o período do experimento foram registrados pela estação climatológica do INMET, localizada no Campo Experimental da Embrapa Meio-Norte em Parnaíba, estado do Piauí (Figure 28).

Figura 28 – Médias de precipitação pluviométrica em Parnaíba/PI(2015)

Fonte: Elaborada pelo autor.

O experimento foi conduzido em área experimental livre de adubação química nos últimos cinco anos. O preparo do solo consistiu de uma lavoura e uma colheita superficial com subsequente nivelamento manual.

O delineamento experimental foi inteiramente causalizado, com sete tratamentos e quatro repetições, em parcelas de 2,0m x 2,0m, contendo cinco fileiras de plantas espaçadas de 0,50m. Alta densidade de sementes (aproximadamente o dobro do número de sementes correspondentes a cada tratamento) foi usada, e a população de plantas foi posteriormente ajustada por afinamento aos 20 dias após a semeadura (DAS) (Tabela 8). O controle de ervas daninhas foi feito manualmente. O sistema de irrigação utilizado foi um aspersor convencional, com um turno de irrigação de dois dias, com duração média de 1,0 (uma) hora.

As espécies de clima tropical anual utilizadas nos tratamentos foram: T1 - Feijão guandu-

Tabela 8 – Espaçamento e densidade de sementeira de plantas utilizadas como adubos verdes

Nome Comum	Profundidade (cm)	Espaçamento (cm)	Sementes/m (linear)
Crotalária breviflora	2 to 3	0.50	30 to 35
Crotalária juncea	2 to 3	0.50	22 to 27
Crotalária mucronata	2 to 3	0.50	40 to 45
Feijão de porco	2 to 5	0.50	4 to 5
Guandu Fava Larga	2 to 3	0.50	20 to 25
Guandu Anão IAPAR 43	2 to 3	0.50	20 to 25
Tefrósia	2 to 3	1.00	5

Fonte: Elaborada pelo autor.

anão (*Cajanus cajan*) cultivar IAPAR 43; T2 - Crotalária mucronata (*Crotalaria mucronata*); T3 - Crotalária juncea (*Crotalaria juncea*); T4 - Feijão de porco (*Canavalia ensiformis* L.); T5 - Crotalária breviflora (*Crotalaria breviflora*); T6 - Feijão guandu (*Cajanus cajan*) cultivar Caqui/Fava larga; and the perennial tropical climate species T7 - Tefrósia (*Tephrosia candida*).

Após 100 dias de instalação do teste, os cortes finais de todos os tratamentos foram realizados a 0,05m acima da superfície do solo com o uso de tesouras de poda. As linhas centrais foram consideradas um enredo útil, descartando as bordas, totalizando uma área de 0,5m² (1,00m linear por 0,5m entre linhas). As raízes das plantas também foram colhidas manualmente com uma ferramenta do tipo “hoe”, a uma profundidade de 0,20m, para obtenção de biomassa fresca e seca para as espécies utilizadas no estudo. As determinações de altura de plantas (AP) foram realizadas; massa fresca da parte aérea (MFPA); massa seca da parte aérea (MSPA); massa fresca de raiz (MFR) e massa seca de raiz (MSR). Germinação (G), floração (F) e formação de vagens (VF) também foram avaliadas. Um florescimento de 50% das plantas foi considerado como pleno.

Todo material botânico, tanto caule como raiz, foi pesado para obtenção de biomassa fresca (MFPA e MFR), em escala de precisão, e posteriormente colocado em sacos de papel para ser embalado em estufa de ventilação forçada a 65°C por 48 horas até atingir um peso constante para a determinação da produção de biomassa seca (MSPA e MSR).

Além das medidas mencionadas, também foram mensurados os dados relacionados ao diâmetro do coletor, assim como o cálculo dos meios de emergência (E), floração (F), altura de plantas no florescimento (AP) e amadurecimento de vagens (V). Depois de coletar todos os valores, uma tabela foi elaborada e o pré-processamento dos dados foi executado.

3.5.2 Pré-processamento

Os dados coletados na fase de plantio foram organizados e categorizados de acordo com os valores obtidos. Todos os parâmetros foram distribuídos em 6 classes de acordo com a Tabela 9. A divisão em classes foi estabelecida pelo método estatístico de categorização

(BARBETTA; REIS; BORNIA, 2010) por meio do cálculo do intervalo (I) dos valores de cada variável (Equação 3.6), sendo Max e Min os valores máximo e mínimo encontrados.

$$I = \frac{Max - Min}{6} \quad (3.6)$$

Com o intervalo definido, a primeira classe começa em Min e vai até $Min + I$ e a última classe termina em Max .

Tabela 9 – Distribuição dos parâmetros em 6 classes

Variável	Classe = 1	Classe = 2	Classe = 3	Classe = 4	Classe = 5	Classe = 6
MFPA (t/ha)	9.60 to 22.47	22.47 to 35.34	35.34 to 48.20	48.20 to 61.07	61.07 to 73.93	73.93 to 86.80
MSPA (t/ha)	5.00 to 9.27	9.27 to 13.53	13.53 to 17.8	17.80 to 22.07	22.07 to 26.33	26.33 to 30.60
MFR (t/ha)	0.40 to 1.53	1.53 to 2.67	2.67 to 3.80	3.80 to 4.93	4.93 to 6.07	6.07 to 7.20
MSR (t/ha)	0.24 to 0.63	0.63 to 1.03	1.03 to 1.42	1.42 to 1.81	1.81 to 2.21	2.21 to 2.60
E (DAS)	2.00 to 2.50	2.50 to 3.00	3.00 to 3.50	3.50 to 4.00	4.00 to 4.50	4.50 to 5.00
F (DAS)	57.00 to 75.33	75.33 to 93.67	93.67 to 112.00	112.00 to 130.33	130.33 to 148.67	148.67 to 167.00
AP (m)	0.65 to 0.87	0.87 to 1.08	1.08 to 1.30	1.30 to 1.51	1.51 to 1.72	1.72 to 1.94
V (DAS)	73.00 to 91.00	91.00 to 109.00	109.00 to 127.00	127.00 to 145.00	145.00 to 163.00	163.00 to 181.00
DC (cm)	0.10 to 0.22	0.22 to 0.33	0.33 to 0.45	0.45 to 0.57	0.57 to 0.68	0.68 to 0.80
R/plant	1.00 to 4.50	4.50 to 8.00	8.00 to 11.50	11.50 to 15.00	15.00 to 18.50	18.50 to 22.00

Fonte: Elaborada pelo autor.

Com os dados distribuídos em categorias, foi elaborado um *array* com seus parâmetros e classes, em que cada instância representava uma classe e cada coluna um atributo. O *dataset* foi processado para que cada transação tivesse os valores no formato “atributo = valor”. Usando o atributo “MFPA”, por exemplo, seus valores foram alterados para “MFPA = 1”, nos casos em que o valor estava dentro dos limites da classe “1” no atributo MFPA, “MFPA = 2”, para valores da segunda classe, etc., executando o mesmo procedimento para todos os atributos e classes.

Após a obtenção de todos os parâmetros, o tempo de meia-vida para cada amostra de planta foi calculado nos intervalos de 30, 60, 90 e 120 dias. A decomposição dos resíduos segue o modelo exponencial simples usado em (REZENDE *et al.*, 1999) (Equação 3.7), em que X representa a quantidade de matéria seca após um período de tempo t , X_0 é a quantidade inicial de matéria seca, k é a constante de decomposição e t é o tempo em dias.

$$X = X_0 \cdot e^{-k \cdot t} \quad (3.7)$$

Reorganizando os termos da equação 3.7, é possível calcular a constante de decomposição ou o valor k (Equação 3.8).

$$k = \frac{\ln \frac{X}{X_0}}{t} \quad (3.8)$$

O tempo de meia-vida ($t_{\frac{1}{2}}$) foi o parâmetro utilizado na avaliação da decomposição de resíduos vegetais, expressando o tempo, em dias, necessário para que metade do material se

decomponha. De acordo com (REZENDE *et al.*, 1999), é possível calcular o tempo de meia-vida por meio da constante de decomposição k (Equação 3.9):

$$t_{\frac{1}{2}} = \frac{\ln(2)}{k} \quad (3.9)$$

De acordo com a metodologia descrita, os tempos de meia-vida ($t_{\frac{1}{2}}$) foram calculados escolhendo os intervalos de 30, 60, 90 e 120 dias a partir da data de plantio. Esses intervalos foram selecionados para padronização do comportamento de decomposição das plantas estudadas. O *dataset* gerado está disponível on-line⁴ no formato “arff”.

Os tempos de meia-vida obtidos nos intervalos de 30, 60, 90 e 120 dias são apresentados na Tabela 10.

Tabela 10 – Categorização dos tempos de meia vida para coleções de 30, 60, 90 e 120 dias.

Classe	$t_{\frac{1}{2}}$ 30 (dias)	$t_{\frac{1}{2}}$ 60 (dias)	$t_{\frac{1}{2}}$ 90 (dias)	$t_{\frac{1}{2}}$ 120 (dias)
1	94.44 to 877.37	40.57 to 404.13	71.70 to 1017.70	88.98 to 264.16
2	877.37 to 1660.30	404.13 to 767.70	1017.70 to 1963.69	264.16 to 439.35
3	1660.30 to 2443.22	767.70 to 1131.26	1963.69 to 2909.68	439.35 to 614.53
4	2443.22 to 3226.14	1131.26 to 1494.82	2909.68 to 3855.67	614.53 to 789.71
5	3226.14 to 4009.07	1494.82 to 1858.38	3855.67 to 4801.66	789.71 to 964.89
6	4009.07 to 4792.00	1858.38 to 2221.94	4801.66 to 5747.65	964.89 to 1140,08

Fonte: Elaborada pelo autor.

A categorização em 6 classes dos tempos de meia vida obtidos em cada intervalo determinado foi realizada. Os valores de 120 dias foram escolhidos para a continuidade dos estudos, pois são os maiores intervalos de tempo e isso possibilita melhor mensuração do uso em grandes plantações.

3.5.3 Avaliação de minsup e minconf ideais

Uma vez que o *dataset* possui 6 classes relacionados à meia-vida de decomposição de leguminosas, 11 atributos e 28 instâncias, sendo os atributos formados por categorias de valores referente à pesquisa de campo executada pela EMBRAPA Meio Norte. As regras foram extraídas com o uso do algoritmo *Apriori-TID* implementado em Java®. As medidas de suporte mínimo foram definidas como {0,0, 0,1, 0,2, 0,3, 0,4, 0,5} e de confiança mínimos {0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9}. As ARNs foram geradas considerando como item objetivo “[HalfLife] = 6” o que significa leguminosas com maiores taxa de meia-vida por possuírem uma taxa de decomposição elevada.

A medida de suporte mínimo não gerou impacto na Relevância das Redes de Regras de Associação geradas, devido a não variação de seus resultados. Seguindo a proposta, o valor

⁴ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/datasets/Green_Manure.zip>

de *minsup* ideal é 0,07, porém com o número reduzido de instâncias o suporte mínimo ideal a ser adotado deve ser 0 (zero), indicando que todos os atributos devem ser estudados para esse *dataset*. Portanto, o valor de suporte mínimo ideal é muito próximo do suporte da classe alvo estudada, validando a **proposta 3**.

As Redes resultantes construídas apresentaram as medidas de Relevância apresentadas na Tabela 11. Observa-se que a ARN com maior relevância é a de confiança mínima 0,3.

Tabela 11 – Relevância x Confiança Mínima (*Green Manure dataset*)

<i>minconf</i>	Relevância
0,1	41,744
0,2	43,941
0,3	44,062
0,4	4,0545
0,5	2,16
0,6	0,3857
0,7	0,0
0,8	0,0
0,9	0,0

Fonte: Elaborada pelo autor.

Para estudo das Medidas de Rede, foram analisadas apenas as ARNs com *minconf* $\leq$ 0,6, pois as Redes com confianças mínimas superiores apresentaram Relevância nula (Tabela 11). Pode-se verificar, por meio da Figura 29, o resultado da variação das Medidas de Rede.

Figura 29 – Relação entre Medidas de Rede e Confiança Mínima com *minsup* = 0,1 (*Green Manure dataset*)

Fonte: Elaborada pelo autor.

De acordo com o resultado observado na Figura 29 e a **proposta 4**, a Rede de maior Relevância deve ser a de *minconf* igual a 0,3, por apresentar um aumento na diferença entre

as medidas de Grau Ponderado Médio, com valores de confiança como peso das arestas, e Comprimento Médio do Caminho (DIF) quando $minconf = 0,4$, confirmando a proposta e servindo de base de validação para extração otimizada de conhecimento.

Portanto, pode-se afirmar que existem relações entre Medidas de Centralidade e medidas referentes a seleção de Regras de Associação (suporte e confiança mínimos). Essas relações podem ser utilizadas como critério de avaliação da medida de $minconf$ a fim de que a extração do conhecimento seja otimizada por meio da elaboração de hipóteses com maior probabilidade de serem verdadeiras. O suporte mínimo deve ser definido por meio de um estudo das instâncias do *dataset*.

3.6 Considerações Finais

Neste capítulo foi apresentada a proposta de uma medida para cálculo de Relevância de Redes de Regras de Associação e uma metodologia para avaliação das medidas de suporte e confiança mínimos no processo de Mineração de Regras de Associação. Fazendo uso de Redes de Regras de Associação, pode-se calcular Medidas de Centralidade para auxiliar na avaliação do atributo de $minconf$ que geram Redes com maior possibilidade de uma extração de conhecimento otimizada. A medida de suporte mínimo não apresentou nenhuma relação com as Medidas de Rede estudadas neste trabalho. Também foi proposta uma Taxa de Complexidade para conjuntos de dados oriundos da UCI.

Uma avaliação experimental foi desenvolvida para estudar o impacto das medidas de suporte e confiança mínimos nas Medidas de Rede, bem como verificar como as Medidas de Centralidade poderiam ser utilizadas na avaliação de $minsup$ e $minconf$. Os *datasets* explorados, *Lenses*, *Hayes-roth* e *Soybean Large* possuem complexidades diferentes dando um caráter mais abrangente aos estudos. O objetivo foi descrever o comportamento das medidas e verificar como as mesmas poderiam ser selecionadas para otimizar a extração do conhecimento.

Para validação das propostas, foi executado uma Prova de Conceito com um *dataset* real formado por dados referentes a adubação orgânica. Os resultados demonstraram a veracidade das propostas para obtenção de Redes que geram hipóteses com maiores possibilidades de serem verdadeiras, portanto pode-se utilizar as Medidas de Centralidade para auxiliar em processos de Mineração de Regras de Associação.

Os resultados demonstraram que o uso das Medidas de Grau Ponderado Médio, com as confianças como pesos, e Comprimento Médio do Caminho, pode auxiliar na avaliação e seleção do melhor valor para $minconf$. Quando a diferença entre essas medidas sofre um aumento no estudo de medidas crescentes de $minconf$, percebe-se que o valor anterior da confiança mínima é a ideal. Para escolha do valor de $minsup$ basta uma exploração simples do *dataset* a ser estudado. As medidas de suporte mínimo ideal e confiança mínima ideal encontrados nos experimentos realizados também podem servir de referência para pesquisas futuras.

Os resultados relatados neste capítulo possibilitaram as respostas para a questão de pesquisa **Q3** (*Quais as relações que existem entre métricas relacionadas a Regras de Associação e medidas referentes a estruturas de Rede?*) e **Q6** (*Qual o impacto nos resultados da geração de hipóteses em processos de Mineração de Regras de Associação com o uso de Redes?*), atendendo ao objetivo desta pesquisa.

REDES DE REGRAS DE ASSOCIAÇÃO FILTRADAS

4.1 Considerações Iniciais

Os conjuntos de dados estão aumentando, tanto em quantidade quanto em dimensionalidade. Tendo presente, como exemplo, dois conjuntos A e B , enumerar todas as combinações possíveis dos elementos de A e B e depois verificar sua correlação não é uma tarefa computacionalmente viável. Assim, a Mineração de Regras de Associação pode ser vista como um mecanismo para oferecer hipóteses candidatas para serem validadas.

A pesquisa em algoritmos de mineração de dados tornou possível filtrar eficientemente grandes quantidades de dados e gerar padrões candidatos que são interessantes, muitas vezes surpreendentes e em certos casos até mesmo mais confiáveis. Um dos desafios das técnicas de mineração de dados é que elas geralmente geram um grande número de “padrões” ou hipóteses, o que torna extremamente difícil para um pesquisador decidir quais hipóteses são confiáveis e se valem a pena serem analisadas. O que é proposto nessa tese, é um método para superar esse desafio associado especificamente a Regras de Associação - uma técnica de mineração de dados bem conhecida ([ZAKI; MEIRA, 2013](#)).

Algoritmos que descobrem Regras de Associação usam medidas de interesse que são capazes de avaliar a qualidade de uma regra. Entre essas medidas, o suporte e a confiança se destacam, embora *Lift*, *Gain*, *Certainty Factor*, *Added Value* ou *Leverage* também sejam indicadores que fornecem informações úteis sobre as regras extraídas ([MARTÍNEZ-BALLESTEROS et al., 2014](#)).

Como os algoritmos de Regra de Associação são capazes de extrair todas as Regras de Associação de acordo com um suporte mínimo e valor mínimo de confiança, o número de regras extraídas geralmente supera a capacidade de exploração do usuário. Várias abordagens foram

propostas para guiar o usuário na exploração de regras. No entanto, a grande maioria dessas abordagens se concentra na exploração com o uso de todas as regras geradas e não considerando explorações que possam se concentrar em um pequeno conjunto de itens ou em um item alvo.

Por outro lado, para facilitar a extração do conhecimento, muitos processos de mineração utilizam técnicas de Redes para visualização dos dados (KIM *et al.*, 2011). A fim de aliar Medidas Objetivas com estrutura de Rede para visualização, é proposta nesta tese a Rede de Regras de Associação Filtrada (*Filtered-ARN* - do inglês *Filtered-Association Rules Network*). A *Filtered-ARN* faz uso de Medidas Objetivas Assimétricas (Definição 3 - Seção 2.2.2 - $M(A \Rightarrow B) \neq M(B \Rightarrow A)$) aliados à Rede de Regras de Associação (ARN) proposta por Pandey *et al.* (2009) para estruturar e auxiliar na análise das regras extraídas em um *dataset*. A estrutura da Rede é a mesma que a ARN, porém a seleção de regras com influência estatística comprovada pode promover a identificação de hipóteses com maior probabilidade de serem verdadeiras.

O objetivo central da *Filtered-ARN* é apresentar um Grafo com Regras de Associação que possuam uma maior chance de serem interessantes. As hipóteses geradas podem ser validadas por outros métodos, como uma avaliação realizada por um especialista do domínio. A principal diferença entre a *Filtered-ARN* e a ARN convencional é que, na *Filtered-ARN*, o usuário pode visualizar um conjunto de itens que promovem uma influência estatística em vez de elementos que apenas se relacionam com o item objetivo. Ao permitir isso, a *Filtered-ARN* apresenta regras que indicam hipóteses com comprovação de dependência entre antecedente e consequente.

A fim de validar as *Filtered-ARN*, foi realizada uma Avaliação Experimental com 3 *datasets* e os resultados foram comparados com a ARN convencional e com um algoritmo de árvore de decisão, pois os mesmos podem ser utilizados para visualização de graus de dependência entre elementos de um *dataset*. Três provas de conceito também foram realizadas com o intuito de comprovar a eficácia da técnica proposta na mineração de dados com *datasets* reais. Os resultados demonstraram que a *Filtered-ARN* é capaz de descrever os elementos que influenciam o item objetivo de modo mais conciso se comparado a ARN, permitindo que o usuário observe os casos em que um item interfere estatisticamente em um item objetivo. Além disso, a *Filtered-ARN* foi comparada a um algoritmo de árvore de decisão, com o objetivo de analisar a explicação dos dados. Tal comparação demonstra que a *Filtered-ARN* apresenta uma estrutura mais eficiente que a árvore de decisão, tornando mais fácil para o usuário entender o conhecimento extraído.

Na próxima seção é apresentada a proposta da *Filtered-ARN*, com suas definições e propriedades. A Avaliação Experimental com 3 *datasets* oriundos da UCI são apresentados na Seção 4.3. As Provas de Conceito com bases de dados reais são descritos e seus resultados analisados na Seção 4.4.

4.2 Modelo Proposto para Geração de Hipóteses de Maior Confiabilidade

Existem casos em que a Mineração de Regras de Associação é feita com o objetivo de explicar itens predeterminados. A Rede de Regras de Associação (ARN) apresenta uma exploração guiada por um único item objetivo. Esta exploração remove todas as regras que não são interessantes no contexto do item objetivo, de acordo com as métricas mínimas de suporte e confiança, mostrando ao usuário apenas as regras relevantes, porém sem a certeza da dependência estatística entre os elementos das regras (PANDEY *et al.*, 2009).

Para exemplificar, considere o *dataset Lenses*¹. Se o usuário construir uma ARN com o atributo “[lenses]=hard”, com o objetivo de descobrir quais sintomas levam ao paciente que utiliza uma lente rígida, o atributo “[prescription]=myope” aparece conectado diretamente ao nó objetivo (Nível = 1). Esse conhecimento pode direcionar o usuário a formular a hipótese que “um paciente com miopia tem maior probabilidade de usar uma lente rígida”. No entanto, a ARN pode, às vezes, apresentar relações que não possuem influência entre os elementos da regra. Ao calcularmos o valor *Added Value* da regra “[prescription]=myope $\Rightarrow$ [lenses]=hard” encontramos $AV = 0$, o que afere uma total independência entre os elementos constituintes dessa regra, portanto a hipótese “um paciente com miopia tende a utilizar lentes rígidas” apresenta um caráter equivocado, pois os dados não possuem influência direta.

Com o objetivo de possibilitar uma exploração completa, levando em conta a relação entre um conjunto de itens objetivos, nesta tese é proposta a Rede de Regras de Associação Filtrada (*Filtered-ARN*), que permite a exploração de um item objetivo com análise de dependência entre os elementos das regras por meio do uso de filtros de Medidas Objetivas Assimétricas na etapa de seleção das Regras de Associação extraídas.

Definição 1: Dado um conjunto de Regras de Associação R , contendo regras de *itemsets* unitários, e um item objetivo Z , a *Filtered-ARN* é uma Rede Direcionada (Rede Direcionada (RD)) que modela todas as regras relacionadas ao item em Z , como:

1. Cada aresta modela uma regra $r \in R$.
2. A partir de qualquer ponto da Rede, é sempre possível alcançar pelo menos 1 vértice representando um item Z .
3. Dado um vértice $v \in \text{Filtered-ARN}$, como $v \notin Z$. Não há caminho de qualquer item Z para v .
4. Se existe uma regra r como $RHS(r) \subset Z$, então a regra $r \in \text{Filtered-ARN}$.

No algoritmo de construção da *Filtered-ARN*, faz-se uso de filtros com medidas objetivas assimétricas (*Added Value* e *Gain*), construindo o gráfico de acordo com as regras selecionadas.

¹ <<http://archive.ics.uci.edu/ml/datasets/Lenses>>

O Algoritmo para geração da *Filtered-ARN* pode ser descrito em 3 passos:

- (1) Passo A: similar ao primeiro passo de todos os processos de Mineração de Regras de Associação. Extração das regras com corte por suporte e confiança mínimos.
- (2) Passo B: efetuar o cálculo das medidas objetivas assimétricas *Added Value* e *Gain*, realizando a exclusão de todas as regras com $AV = 0$ e valores inferiores ao ganho mínimo (*mingain*).
- (3) Passo C: escolher um item frequente Z , que será representado no conjunto de regras como o nó objetivo, e construir um *B-graph* que flui de forma recursiva para Z .

A primeira etapa consiste na fase de Mineração de Regras de Associação. A única restrição adicionada a essa etapa, se comparada a uma Mineração de Regras de Associação convencional, é que as regras devem possuir conjuntos unitários no antecedente e consequente ($|LHS| = 1$ e $|RHS| = 1$). Essa restrição foi adicionada para facilitar a modelagem da *Filtered-ARN*.

O segundo passo é a filtragem das regras, que consiste na seleção das regras que possuem elementos com dependência estatística e definição do ganho mínimo de influência (*mingain*). Esta etapa guiará toda a exploração, pois definirá as regras de interesse que serão utilizadas com o item objetivo com as quais a Rede será construída.

No último passo, o usuário deve selecionar o item que deseja entender no *dataset*. Em seguida, é efetuada a construção da *Filtered-ARN*. Esta etapa é responsável por obter todas as regras que estão direta ou indiretamente relacionadas ao item objetivo e modelá-las. A construção da *Filtered-ARN* é feita recursivamente. Primeiro, o item selecionado como item objetivo é modelado no gráfico (Nível = 0). Então, todas as regras que apresentam o item LHS não inserido no gráfico e também possuem o item do RHS no nível 0 são modelados na Rede. O mesmo processo é feito para todos os itens no nível 1, nível 2 e assim por diante. Até que não haja mais regras a serem modeladas. Além disso, a *Filtered-ARN* é construída de acordo com os níveis de seus vértices.

Definição 2: O nível de um determinado vértice $v \in \text{Filtered-ARN}$ é o número de arestas necessárias para acessar o item objetivo.

Por exemplo, o item objetivo Z tem nível zero (Nível = 0), pois ele não precisa passar por nenhuma aresta para alcançar o próprio item Z . Itens na parte LHS de regras que possuem $RHS \subset Z$ terá o nível um (Nível = 1). Eles estão uma aresta longe dos itens em Z (Figura 30).

Conforme descrito, a *Filtered-ARN* é construída conectando vértices em um gráfico de acordo com a Regra de Associação que está representando. A *Filtered-ARN* é estabelecida como sendo uma RD, a prova é a seguinte:

Figura 30 – Níveis dos nós de uma *Filtered-ARN*

Fonte: Elaborada pelo autor.

Prova: a *Filtered-ARN* é construída conectando vértices do nível X aos vértices no nível $X - 1$. Assim, todas as conexões na *Filtered-ARN* são direcionadas. Suponha que haja um ciclo $A \rightarrow B \rightarrow C \rightarrow A$, de acordo com as regras de construção da *Filtered-ARN*, $\text{Nível}(A) = \text{Nível}(B) + 1$, $\text{Nível}(B) = \text{Nível}(C) + 1$ e $\text{Nível}(C) = \text{Nível}(A) + 1$. $\text{Nível}(C) = \text{Nível}(A) - 2$, portanto, não é possível criar um ciclo.

Outra propriedade importante é que, a partir de qualquer vértice da *Filtered-ARN*, é possível atingir um vértice no nível 0 (item objetivo).

Prova: suponha que exista um vértice v_x , no nível $N > 0$, que não esteja conectado a nenhum vértice nível zero ($\text{Nível} = 0$). A *Filtered-ARN* é construída do nível zero aos níveis mais altos, sempre modelando as regras que têm o RHS em um nível inferior.

Assim, o vértice v_x só será modelado se um vértice em um nível inferior àquele que v_x está conectado também estiver conectado a um vértice no nível 0. Então, todos os nós na *Filtered-ARN* estão conectados a, pelo menos, um vértice em nível 0.

A exclusão das regras de dependência estatística nula ($AV = 0$) é uma propriedade fundamental das *Filtered-ARNs*, pois garantem que todos os itens que participam da Rede gerem algum tipo de influência no item objetivo.

Prova: a medida objetiva *Added Value* (Equação 2.3) é o resultado da diferença entre a confiança $\text{conf}(LHS \Rightarrow RHS)$ e o $\text{sup}(RHS)$. Essa medida torna-se zero ($AV = 0$) quando

a confiança da regra e o suporte do consequente são iguais. Como a confiança da regra é uma probabilidade condicionada à presença de RHS e o suporte é a probabilidade de RHS, significa que o item antecedente não influencia em nenhum momento na presença do consequente, provando a total independência estatística dos mesmos.

Todas essas propriedades são importantes para a análise da *Filtered-ARN*, pois garantem que todos os itens modelados sempre apontem para o item objetivo selecionado com algum tipo de relação de influência. As propriedades garantem que toda a *Filtered-ARN* estará explorando o item definido em Z , permitindo que o usuário entenda sua ocorrência e construção de hipóteses a partir dos dados.

Para validação da *Filtered-ARN*, foram utilizados três *datasets* da UCI² (*Lenses*, *Hayes-roth* e *SoyBean Large*), além de 3 Provas de Conceito com *datasets* reais. Para a seleção das regras foi estabelecido um valor de ganho mínimo (*mingain*) igual a 0,01, efetuando-se a filtragem das regras que possuem $AV = 0$. Para a construção da Rede foi usado o processo recursivo descrito anteriormente.

Como o objetivo da *Filtered-ARN* é analisar as correlações no *dataset* de modo a apoiar o usuário na obtenção de informações fidedignas e construir hipóteses com grande probabilidade de serem verdadeiras, a *Filtered-ARN* é avaliada com dois outros métodos. Primeiro, faz-se uma comparação entre a *Filtered-ARN* e a ARN convencional, analisando as diferenças entre elas e as vantagens e desvantagens de usar cada uma dessas abordagens. Depois disso, a *Filtered-ARN* é comparada a um algoritmo de árvore de decisão. O objetivo dessa comparação é confrontar os resultados gráficos e discutir qual deles produz uma estrutura melhor para análise do *dataset* de acordo com um conjunto de itens.

4.3 Avaliação Experimental

Com o objetivo de validar a abordagem *Filtered-ARN* e apresentar sua capacidade de exploração de *datasets*, foram realizados alguns experimentos em bases de dados conhecidas. Os experimentos foram focados em apresentar as diferenças dos métodos, mostrando como a *Filtered-ARN* é capaz de possibilitar uma exploração mais ampla dos dados, dando ao usuário uma compreensão completa dos mesmos. Além disso, um algoritmo de árvore de decisão foi utilizado nos *datasets* a fim de comparar seu resultado gráfico com a saída da *Filtered-ARN*.

O primeiro estudo utilizou o *dataset Lenses*, disponível na UCI³. Nessa base de dados, cada linha representa os atributos de um paciente e a lente de contato que foi prescrita para ele. O objetivo principal do *dataset* é descrever quais características implicam na prescrição de cada tipo de lente de contato. É importante lembrar que este *dataset* é uma versão simplificada do

² <<http://archive.ics.uci.edu/ml/>>

³ <<https://archive.ics.uci.edu/ml/datasets/lenses>>

problema, os *insights* obtidos a partir dele podem não representar as correlações reais do cenário de prescrição de lentes de contato.

O segundo estudo foi realizado no *dataset Hayes-roth*, disponível on-line⁴. O conjunto de dados traz informações geradas por dados pessoais, como idade e nível de escolaridade. Esse banco de dados contém 6 atributos de valor numérico, sendo o primeiro apenas um índice, que foi excluído. O último atributo divide as instâncias em 3 classes.

O terceiro estudo foi feito no *dataset Soybean (Large)*, também disponível na UCI⁵. O *dataset* relata os resultados de uma pesquisa de características de doenças em plantações de soja. Existem 19 classes, apenas as 15 primeiras foram usadas em experimentos anteriores. As últimas quatro classes são pouco exploradas, pois têm poucos exemplos. Existem 35 atributos categóricos, alguns nominais e alguns ordenados. Os valores dos atributos são codificados numericamente, com o primeiro valor codificado como “0”, o segundo como “1,” e assim por diante. O *dataset* possui campos “vazios” que foram excluídos.

Portanto, para a realização dos experimentos apresentados neste trabalho, foram selecionados três *datasets*, sendo um de baixa complexidade (LENSES), um de média complexidade (HAYES-ROTH) e outro de alta complexidade (SOYBEAN LARGE), de acordo com a taxa de complexidade proposta nesta tese (Apêndice B), apresentada no Capítulo 3. Com as 3 bases de dados, foram realizadas as mesmas etapas para a construção das *Filtered-ARNs*.

Nesses experimentos, as regras foram extraídas com o uso do algoritmo *Apriori-TID* implementado em Java®. As regras extraídas foram filtradas pela medida *Added Value*, sendo excluídas todas aquelas que possuem valor nulo ($AV = 0$), que indicam ausência de influência estatística, e também foi selecionado o ganho mínimo de 0,01, para que o maior número de regras possível pudesse ser analisado. As Redes de Regras foram construídas graficamente com o uso do software *Gephi* (BASTIAN; HEYMANN; JACOMY, 2009). O *Gephi* é uma aplicação *open source* específica para construção de Redes e está disponível on-line⁶. As árvores de decisão foram geradas pelo algoritmo J48, disponível no *software* Weka⁷, usando a configuração padrão. Nas Seções seguintes são explanados os resultados com cada *dataset*.

4.3.1 Lenses dataset

Como o *dataset* tem um pequeno número de atributos, o suporte mínimo foi definido como 0, portanto, todos os valores foram considerados. A confiança mínima foi definida para 0,25, possibilitando o estudo de classes que ocorreram pelo menos 1 em cada 4 vezes. Além disso, o tamanho da regra foi especificado em dois, considerando um item no LHS e um item no RHS. Usando essa configuração, foram obtidas 99 Regras de Associação candidatas para o

⁴ <<https://archive.ics.uci.edu/ml/datasets/Hayes-Roth>>

⁵ <<http://archive.ics.uci.edu/ml/datasets/Soybean+%28Large%29>>

⁶ <<https://gephi.org/users/download/>>

⁷ <<http://www.cs.waikato.ac.nz/ml/weka/>>

dataset “Lenses”. Após a etapa de filtragem restaram 60 regras.

A *Filtered*-ARN foi gerada considerando como item objetivos: “[lenses]=hard”. O resultado pode ser visto na Figura 31. A Rede gerada possui 4 níveis: nível 0, nível 1, nível 2 e nível 3. A *Filtered*-ARN tem apenas 2 itens relacionados ao item objetivo, que são: “[tear]=normal” e “[astigmatic]=yes”. Estas regras podem ser capazes de influenciar no item “[lenses]=hard” e é interessante investigar, pois são os únicos parâmetros que geram uma influência no item objetivo, portanto com um alto grau de probabilidade de geração de hipóteses verdadeiras.

Figura 31 – *Filtered*-ARN com “[lenses]=hard” como item alvo

Fonte: Elaborada pelo autor.

Pode-se observar na *Filtered*-ARN gerada que os nós de nível 2 são as outras classes que indicam o tipo de lente “[lenses]=soft” e “[lenses]=no”, indicando que os elementos do nível 1 sofrem interferência desses itens. Quando analisa-se os nós de nível 3, percebe-se que algumas regras se destacam como: “[tear]=reduced” $\Rightarrow$ “[lenses]=no”, “[age]=presbyopic” $\Rightarrow$ “[lenses]=no”, “[age]=young” $\Rightarrow$ “[lenses]=no” e “[prescription]=myope” $\Rightarrow$ “[lenses]=no”. Estas regras possuem conexão apenas com a classe “[lenses]=no”, o que gera hipóteses para a construção de uma nova *Filtered*-ARN com “[lenses]=no” como item alvo, provocando um novo direcionamento na exploração do conhecimento.

Na Figura 32 é apresentada a ARN com “[lenses]=hard” como item objetivo. Pode-se observar que a Rede possui apenas 3 níveis, porém com uma estrutura totalmente diferente da *Filtered*-ARN. Como a construção das Redes é diretamente induzida pelo item RHS das regras, percebe-se que regras do nível 1 sem influência comprovada ($AV = 0$) aparecem, como “[prescription]=myope” $\Rightarrow$ “[lenses]=hard” e “[age]=young” $\Rightarrow$ “[lenses]=hard”, o que leva a geração de hipóteses equivocadas a respeito do uso de lentes rígidas.

Outra diferença notória é o aumento do número de nós de nível 3, sem a distinção de dependências entre os mesmos, o que impossibilita o direcionamento dos estudos, pois a Rede demonstra o mesmo grau de importância para todos os itens do mesmo nível.

Figura 32 – ARN com “[lenses]=hard” como item alvo

Fonte: Elaborada pelo autor.

A árvore de decisão é apresentada na Figura 33. O algoritmo J48 obteve 20/24 classificações corretas (83,33%) e perdeu 4/24 exemplos (16,67%).

Figura 33 – Árvore de Decisão para o dataset *Lenses*

Fonte: Elaborada pelo autor.

Comparando a saída da *Filtered-ARN* com a estrutura da árvore de decisão (Figura 33), é possível perceber diferenças na explicação dos itens objetivos. Ambos têm o “[tear]=reduced” e “[prescription]=hypermetrope” conectados diretamente ao “[lenses]=no”, porém na *Filtered-ARN* outras regras surgem como a condição de idade ([age]).

Na árvore, é explicado que “[prescription]=myope” se liga diretamente à “[lenses]=hard”, o que é um equívoco por se tratar de uma condição sem influência ($AV = 0$), o que pode ser inferido na *Filtered-ARN*.

4.3.2 Hayes-roth dataset

O dataset *Hayes-roth* tem 6 atributos, sendo que os 2 primeiros foram gerados aleatoriamente, então eles foram removidos. O tamanho da regra foi especificado em dois, considerando 1

item no LHS e 1 item no RHS. Usando esta configuração, Regras de Associação foram obtidas e as Redes com o nó objetivo “[class]=3” foram construídos porque é a classe com o maior número de instâncias.

Como a base de dados tem apenas 162 instâncias, o suporte mínimo foi definido como 0, portanto, todos os valores foram considerados. A confiança mínima foi definida para 0.25, possibilitando o estudo de classes que ocorreram pelo menos 1 em cada 4 vezes. Usando essa configuração, foram obtidas 167 Regras de Associação candidatas para o *dataset* “Hayes-roth”. Após a etapa de filtragem restaram 123 regras.

Na Figura 34 é apresentada a *Filtered-ARN* com “[class]=3” como item objetivo. Observa-se a presença de apenas 4 nós em toda a Rede, sendo 1 nó alvo e os outros 3 nós ligados diretamente ao nó alvo.

Figura 34 – *Filtered-ARN* com “[class]=3” como item alvo e nós de nível 1

Fonte: Elaborada pelo autor.

Com a análise da *Filtered-ARN* da Figura 34, infere-se que o perfil das pessoas da classe 3 do *dataset* possuem idades maiores ([age]=4), níveis educacionais mais elevados ([educational_level]=4) e status matrimonial do tipo 4 ([marital_status]=4). Essas informações ficam claras na *Filtered-ARN* e representam conexões com alta probabilidade de serem verdadeiras.

Foi construída a ARN (Figura 35) para comparação com a *Filtered-ARN*. A estrutura da ARN é totalmente diferente, com 3 níveis de nós. Várias conexões sem influência aparecem, acarretando a geração de hipóteses com alto nível de chance de serem falsas. Conexões com o item alvo como “[hobby]=1”, “[hobby]=2” e “[hobby]=3” são inseridas na ARN, porém essas regras possuem AV = 0, o que denota a não influência do atributo “hobby” na seleção de pessoas da classe 3 do *dataset*.

A árvore de decisão gerada (Figura 36) para validação possui 19 folhas e tamanho 25. Comparando a saída da *Filtered-ARN* com a árvore de decisão, é possível perceber diferenças na explicação dos itens objetivos. Ao analisar a árvore, percebe-se as mesmas conexões da *Filtered-ARN* ligadas diretamente à “classe = 3” do *dataset*, “[age]=4”, “[educational_level]=4” e “[marital_status]=4”, porém surgem dependências do nível educacional e do estado matrimonial com outras categorias de idade, o que pode gerar inconsistência no conhecimento extraído e a

Figura 35 – ARN com “[class]=3” como item alvo e nós de nível 1 (vermelho)

Fonte: Elaborada pelo autor.

elaboração de hipóteses falsas.

Figura 36 – Árvore de Decisão para o *dataset Hayes-roth*

Fonte: Elaborada pelo autor.

4.3.3 Soybean Large dataset

Para continuar a validação dos benefícios do uso da *Filtered-ARN*, algumas experiências foram realizadas no *dataset* de plantações de soja (Soybean Large), por se tratar de um *dataset* com alto grau de complexidade (Tabela 3).

Uma vez que o *dataset* possui 19 classes, 36 atributos e 307 instâncias, sendo algumas com campos vazios, o suporte mínimo foi definido como 0,03 para evitar o aparecimento de classes com informações incompletas. A confiança mínima foi definida para 0,45, para a geração de uma proporção mais coerente com o item objetivo “[classe] = rhizoctonia”, o qual

foi selecionado, em relação aos itens das demais classes. Além disso, o tamanho da regra foi especificado em dois, considerando 1 item no LHS e 1 item no RHS. Usando essa configuração, foram obtidas 4.223 Regras de Associação candidatas para o *dataset* “Soybean”. Após a etapa de filtragem restaram 4.019 regras.

A *Filtered-ARN* foi gerada considerando como item objetivo “[*classe*] = rhizoctonia”, o que significa plantas com um tipo específico de fungo. Um destaque do resultado da Rede com os nós de nível 1 e nível 2 pode ser visto na Figura 37. A *Filtered-ARN* completa está disponível on-line⁸.

Analisando-se os itens de nível 1 da *Filtered-ARN*, percebe-se que apenas 1 elemento provoca influência direta no item objetivo, “[*stem-cankers*] = 1”. Esta relação pode ser capaz de descrever o item objetivo com um alto grau de dependência, ocasionando a formação de uma hipótese com grande probabilidade de ser verdadeira.

Figura 37 – *Filtered-ARN* com “[*class*]=rhizoctonia” como item alvo e nós de nível 1 (laranja)

Fonte: Elaborada pelo autor.

Na Figura 38 é apresentado um corte da ARN com “[*class*]=rhizoctonia” como item objetivo e nós de nível 1, nível 2 e nível 3 sem antecessores. A Rede completa está disponível on-line⁹. Pode-se observar que a Rede possui um número bem maior de itens conectados diretamente ao objetivo. Além do item “[*stem-cankers*] = 1”, outros 4 elementos formam o nível 1 da Rede, porém sem nenhuma garantia de dependência, o que pode ocasionar o levantamento de hipóteses falsas.

Os elementos “[*canker-lesion*] = 1” e “[*severity*] = 2”, que fazem parte do nível 1 da ARN, são observados no nível 2 na *Filtered-ARN*, deduzindo-se com isso que eles podem afetar o item objetivo, mas de um modo mais indireto que o indicado pela ARN. Os outros itens observados no nível 1 da ARN, “[*leaves*] = 0” e “[*fruit-pods*] = 3”, fazem parte do nível 3 da *Filtered-ARN*,

⁸ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/capitulos/capitulo4.zip>

⁹ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/capitulos/capitulo4.zip>

o que diminui drasticamente a probabilidade de uma hipótese ser gerada diretamente com o item objetivo.

Na *Filtered-ARN* pode-se observar outros itens de nível 2, “[fruit-spots] = 4” e “[date] = 0”, que influenciam na única condição ligada diretamente ao item objetivo.

Figura 38 – ARN com “[class]=rhizoctonia” como item alvo e nós de nível 1 (laranja)

A árvore de decisão gerada para validação possui 69 folhas e obteve 87,58% de precisão. Como a árvore é muito longa, não foi gerada imagem integral da mesma. O arquivo contendo a saída completa está disponível on-line¹⁰. Ao analisar a árvore, é difícil entender até mesmo o comportamento do item “[class]=rhizoctonia”. A árvore que explica as classes ficou mais complexa de se entender do que a Rede gerada pela *Filtered-ARN*.

4.4 Provas de Conceito

Com o intuito de validar os benefícios do uso da *Filtered-ARN*, foram formuladas três provas de conceito com dados reais. O primeiro *dataset* de Adubo Orgânico (*Green Manure*) coletado junto a EMBRAPA Meio Norte. A segunda base de dados é relacionada ao processo de geração automática de conteúdo processual (PCG), uma técnica para produção automatizada de elementos de jogos de vídeo-game, nesta pesquisa sendo mais específica ao jogo do *Angry Birds*. O terceiro *dataset* descreve parâmetros de exames mamográficos e como os mesmos são classificados de acordo com um padrão internacional médico chamado BI-RADS (do inglês, *Breast Imaging Reporting and Data System*). Para a seleção das regras foi estabelecido um valor de ganho mínimo (*mingain*) igual a 0,01, efetuando-se a filtragem das regras que possuem AV = 0. Para a construção das Redes foi usado o processo descrito na Seção 4.2.

Como o objetivo da *Filtered-ARN* é analisar as correlações no *dataset* de modo que o usuário possa obter informações fidedignas e construir hipóteses com grande probabilidade de serem verdadeiras, a *Filtered-ARN* é avaliada com dois outros métodos. Primeiro, faz-se uma comparação entre a *Filtered-ARN* e a ARN convencional, analisando as diferenças entre

¹⁰ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/capitulos/capitulo4.zip>

elas e as vantagens e desvantagens de se usar cada uma dessas abordagens. Depois disso, a *Filtered-ARN* é comparada a um algoritmo de árvore de decisão. O objetivo dessa comparação é confrontar os resultados e discutir qual deles produz uma estrutura melhor para análise do *dataset* de acordo com um conjunto de itens. A ferramenta Gephi (BASTIAN; HEYMANN; JACOMY, 2009) foi utilizada na construção gráfica de todas as Redes, nas quais foram aplicados algoritmos de Layout diferenciados para obter uma melhor percepção dos dados, bem como extrair o conhecimento das relações definidas pela Rede.

A primeira prova de Conceito foi avaliada por especialistas da EMBRABA Meio Norte. Na segunda e terceira provas de conceito, foram construídos algoritmos genéticos (AGs) nos quais as informações geradas por cada Rede pudessem ser inseridas no processo de seleção da população. Sendo assim, foram gerados 4 (quatro) AGs, o primeiro utilizando apenas os conceitos básicos de algoritmos evolutivos, o segundo configurado com os dados extraídos da árvore de decisão, o terceiro com os dados extraídos da ARN e, por fim, o quarto configurado com os dados extraídos da *Filtered-ARN*. A interpretação dos resultados é feita diretamente na estrutura da Rede. Para comparação dos resultados, foi calculado a acurácia relacionada aos indivíduos da população final de cada AG, a fim de que se comprovar a otimização dos classificadores por meio da extração de conhecimento extraído com o uso das *Filtered-ARNs*.

4.4.1 Green Manure dataset

Com o intuito de validar uma aplicação das Redes de Regras de Associação Filtradas, foi feita uma Prova de Conceito com o uso de dados relacionados à agricultura orgânica. Os experimentos realizados que proporcionaram a geração do *dataset* estão descritos na Seção 3.5. Uma ARN e uma Árvore de Decisão foram elaboradas a fim de que os padrões obtidos fossem comparados. Os resultados foram validados¹¹ por especialistas da EMBRAPA Meio Norte/UEP Parnaíba-PI.

4.4.1.1 Filtered-ARN

Uma vez que o *dataset* possui 6 classes relacionados à meia-vida de decomposição de leguminosas, 11 atributos e 28 instâncias, sendo os atributos formados por categorias de valores referente à pesquisa de campo executada pela EMBRAPA Meio Norte, o suporte mínimo foi definido como 0,3 e a confiança mínima foi definida para 0,5 para aferição comparativa com os estudos de Calçada, Rezende e Teodoro (2017). Além disso, o tamanho da regra foi especificado em dois, considerando um item no LHS e um item no RHS. Usando essa configuração, foram obtidas 64 Regras de Associação candidatas para o *dataset* “Green Manure”. Após a etapa de filtragem restaram 50 regras.

¹¹ A validação dos resultados foi realizada pela avaliação do conhecimento identificado nas Redes por especialistas

A *Filtered-ARN* foi gerada considerando como item objetivos “[HalfLife]=6” o que significa leguminosas com maiores taxa de meia-vida por possuírem uma taxa de decomposição elevada. O resultado da Rede pode ser visto na Figura 39. O arquivo contendo a saída completa pode ser visto on-line¹².

Analisando-se os itens de nível 1 da *Filtered-ARN*, percebe-se que 5 itens estão relacionados ao item objetivo, destacando-se “[AP] = 4”, parâmetro relacionado a altura da planta no florescimento, que não possui nenhum antecessor. Estas regras podem ser capazes de descrever quais as prováveis características das plantas com maior tempo de meia-vida, sendo os únicos parâmetros que geram uma influência no item objetivo, portanto gerando hipóteses com um alto grau de probabilidade de serem verdadeiras.

Figura 39 – *Filtered-ARN* com “[HalfLife]=6” como item objetivo e nós de nível 1 (azul)

Fonte: Elaborada pelo autor.

4.4.1.2 ARN

Na Figura 40 é apresentada a ARN com “[HalfLife]=6” como item objetivo. Pode-se observar que a Rede possui uma estrutura diferente da *Filtered-ARN*. O número de itens conectados diretamente ao objetivo aumentou para 7. Percebe-se que alguns elementos são

¹² <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/greenmanure/>

4.4.1.3 Árvore de Decisão

Além de comparar com a ARN, a *Filtered-ARN* também foi comparada a uma árvore de decisão. A árvore de decisão gerada pelo algoritmo J48 é apresentada na Figura 41. O algoritmo J48 obteve apenas 39,28% de classificações corretas formando uma árvore com 16 folhas.

Figura 41 – Árvore de decisão construída com o algoritmo J48 e o *dataset* de adubos orgânicos

Fonte: Elaborada pelo autor.

Comparando a saída da *Filtered-ARN* com a árvore de decisão (Figura 41), é possível perceber diferenças na explicação dos itens objetivos. Para o objetivo “[HalfLife] = 6.0” é observado na árvore apenas a possibilidade direta do parâmetro [F] referente à floração da planta, sendo obtidos os valores “[F] = 1.0” quando “[MSPA] = 1.0” e “[F] = 1.0 ou 2.0 ou 4.0 ou 5.0” quando “[MSPA] = 2.0”. As outras condições paramétricas para obtenção de meias-vidas mais elevadas são totalmente ignoradas pelo algoritmo J48. Desta forma, a aplicação do algoritmo *Filtered-ARN* possibilita um estudo mais amplo de parâmetros, otimizando a extração de conhecimento por meio do estudo de hipóteses mais relevantes e com maior probabilidade de serem verdadeiras.

Todos os resultados foram avaliados por especialistas da EMBRAPA Meio Norte e apenas as informações da *Filtered-ARN* foram validadas em sua totalidade. Os especialistas informaram ainda que a relação entre diâmetro do coletor ([DC]) e taxa de meia-vida ([Half-life]) das plantas com menor índice de decomposição é considerada apenas como uma hipótese suspeita. Este resultado incentiva positivamente a realização de novos experimentos agrícolas para comprovação do conhecimento fornecido pela mineração de dados com uso das *Filtered-ARNs*.

4.4.2 Angry Birds dataset

Ferreira e Zhao (2016) introduziram um algoritmo genético (AG) como técnica para geração de nível de procedimento (PLG - do inglês *Procedural Level Generation*) de fases do jogo *Angry Birds*. Primeiramente, o método inicializa aleatoriamente uma população, gerando vários níveis que são representados como indivíduos. Os indivíduos são avaliados, em seguida, usando um clone dos jogos *Angry Birds*, que simula uma partida para estimar se um nível é

estável e jogável. A população evolui com a aplicação de operadores genéticos, como seleção, cruzamento e mutação. O processo evolutivo é realizado até que um número de gerações seja alcançado ou a melhor aptidão não melhore após 10 gerações. O operador de *crossover* gera dois novos indivíduos pela aplicação de um *Crossover* Uniforme e a mutação pode ser aplicada em cada novo indivíduo gerado após o cruzamento. A mutação muda aleatoriamente cada elemento do indivíduo (bloco ou porco). O AG é totalmente descrito em [Ferreira e Zhao \(2016\)](#).

Um indivíduo no AG proposto representa um nível de *Angry Birds* codificado como uma matriz de colunas. Cada elemento em uma coluna é um bloco elementar, um porco ou um bloco composto. Um bloco composto é uma estrutura pré-definida feita por blocos elementares. A análise dos indivíduos está relacionada a “Estabilidade” e “Jogabilidade”, portanto o algoritmo avaliará apenas a qualidade das estruturas geradas.

O indivíduo (matriz de colunas do nível) é decodificado como um nível completo. No entanto, as pilhas de blocos gerados, com ou sem porcos, não podem ser estáveis após o processo de decodificação baseado na simulação física do clone do jogo. Esta simulação retornará o nível de movimento de cada bloco, permitindo medir o nível de estabilidade. Por exemplo, a partir da medida mencionada é possível saber quando uma pilha ou parte dela entrou em colapso. A viabilidade é medida quando o indivíduo decodificado não gerou um nível colapsado. Neste ponto, um robô simula o lançamento das aves, usando diferentes ângulos de inclinação, até que todos os porcos sejam mortos. Além disso, sabemos que tal nível pode ser finalizado pelo jogador.

A população inicial é gerada de forma randômica. Para avaliar os indivíduos, o algoritmo deve criar um descritor de nível *xml* contendo todos os indivíduos da população atual. A coordenada *x* dos blocos começa no centro da primeira coluna. Os elementos na mesma coluna têm a mesma coordenada *x*, que são baseados nos valores de distância associados às colunas. A coordenada *y* de um elemento depende da altura dos elementos abaixo dele ([FERREIRA; TOLEDO, 2014](#)). Este arquivo é analisado e os resultados da simulação são usados para calcular a função de *fitness* descrita pela Equação 4.1.

$$fitness(x) = |(b_n * B) - B_u| + |(l_n * L) - L_b| + p_f + s \quad (4.1)$$

Nesta função, *B* é um parâmetro de entrada que define a quantidade máxima de aves permitida em um nível, enquanto *B_u* é a quantidade de aves usada durante a simulação a partir do nível gerado. O parâmetro de entrada *L* define a quantidade máxima de blocos permitidos em um nível e *L_b* é a quantidade de blocos no início da simulação para o nível atual. Os demais termos da equação avaliam o número de porcos no final da simulação *p_f* e a estabilidade *s* dos blocos dentro do nível. O nível é considerado estável e viável quando essas duas variáveis são zero.

Nesta experiência, a etapa de inicialização do AG é executada para criar 1000 níveis para criação do *dataset*. Em seguida, esses 1000 níveis são avaliados usando a função de adequação

da Equação 4.1 com os parâmetros $B = 5$ pássaros e $L = 100$ blocos. No final do processo, 1000 níveis foram gerados aleatoriamente para o jogo Angry Birds e, a partir dos valores obtidos para p_f e s após a avaliação de aptidão (simulação), é possível saber se os níveis são viáveis e estáveis. Foram produzidos 39% de níveis estáveis e 69% de níveis viáveis. Ao analisar simultaneamente os dois aspectos, o número de níveis estáveis e viáveis é de 27%.

Após a obtenção dos indivíduos (fases), foi desenvolvido um *script* para obtenção de suas características e construção de dois *datasets*. Uma base de dados refere-se à característica de “Estabilidade” apresentada na Tabela 12, enquanto a outra base lida com a “Viabilidade” ou “Jogabilidade”, conforme mostrado na Tabela 13. O atributo referente a quantidade de pássaros apenas é relacionada à jogabilidade da fase, pois a estabilidade independe deste parâmetro.

Tabela 12 – Atributos do *dataset* “Stability”

Atributos	Tipo de dado
Número de porcos	Inteiro
Número de blocos	Inteiro
Número de tipos de blocos	Inteiro
“Stability”	Binário

Fonte: Elaborada pelo autor.

Tabela 13 – Atributos do *dataset* “Viability”

Atributos	Tipo de dado
Número de porcos	Inteiro
Número de blocos	Inteiro
Número de tipos de blocos	Inteiro
Número de pássaros	Inteiro
“IsPlayable”	Binário

Fonte: Elaborada pelo autor.

Dos dois *datasets*, foram então avaliados os parâmetros extraídos aplicando a árvore de Decisão, ARN e *Filtered*-ARN. Estes parâmetros deram suporte ao *game designer* para configurar o AG conforme as hipóteses geradas, visando à geração de níveis estáveis e viáveis. Os conjuntos de dados gerados estão disponíveis on-line (“Stability”¹³ e “Viability”¹⁴)

4.4.2.1 Árvore de Decisão

Foi executado o experimento com o uso do algoritmo J48 implementado no software Weka com cada base de dados. A acurácia para a base “Estabilidade” foi de 65.062%, porém a finalidade deste experimento era a extração de parâmetros pelo resultado gráfico (árvore) para

¹³ <http://sites.labic.icmc.usp.br/dariobalcada/tese/dariobalcada_files/datasets/Stability.zip>

¹⁴ <http://sites.labic.icmc.usp.br/dariobalcada/tese/dariobalcada_files/datasets/Viability.zip>

auxiliar a configuração do AG. Com a base “Viabilidade” a acurácia foi de 93.0376%. Ambas as árvores estão disponíveis on-line¹⁵.

Os parâmetros foram analisados e o resultado (Tabela 14) foi utilizado para configuração das restrições utilizadas pelo AG.

Tabela 14 – Conhecimento extraído com uso da Árvore de Decisão

Atributo	Stability	Viability	Stability e Viability
Porcos	1	1 – 5	1
Blocos	8 – 10	< 23, > 31	8 – 10
Tipos de Blocos	1, 2	< 9	1, 2
Pássaros	-	1, 2, > 14	-

Fonte: Elaborada pelo autor.

No caso da estabilidade, a forma circular dos porcos explica seu número reduzido quando se analisa a estabilidade dos mesmos em uma pilha. Além disso, o número reduzido de blocos e os tipos de blocos ajudam a aumentar a chance de níveis estáveis. Por outro lado, a viabilidade do nível aumenta quando temos mais porcos e mais blocos, por serem facilmente colapsados ao lançar pássaros que ajudam a matar os porcos. Assim, considerando a estabilidade e viabilidade, temos os níveis com uma grande quantidade de blocos, mas com apenas um porco e dois tipos de bloco.

A partir da análise anterior, o próximo passo foi definir a aptidão da função do AG com $B = 2$ pássaros e $L = 10$ blocos e gerar aleatoriamente 1000 novos níveis. O cálculo de acurácia foi feito e constatou-se que a porcentagem de níveis estáveis diminuiu a 10,6% e os níveis viáveis também foram reduzidos para 5,6%. Quando se analisa os dois aspectos simultaneamente, o número de fases estáveis e jogáveis foi de apenas 2%. Assim, a árvore de decisão não produz uma análise confiável de parâmetros, pois leva a uma pior geração de níveis pelo procedimento de inicialização do AG.

4.4.2.2 ARN

O algoritmo *Apriori* foi utilizado nos *datasets* “Estabilidade” e “Viabilidade” para extrair as Regras de Associação, definindo os parâmetros de $minsup = 0.2$ e $minconf = 0.7$, pois foram os valores com melhor número de regras extraídas, o que levou à geração de um total de 347 e 497 regras para os *datasets* de “Estabilidade” e “Viabilidade”, respectivamente. Após a extração das regras, foram selecionados os itens alvo de cada base de dados para a construção de dois ARNs. O item objetivo em cada *dataset* foi “Stability = 1.0”(true) e “IsPlayable = 1.0”(true).

Depois que a ARN foi construída, foram coletadas informações sobre os intervalos de valores para cada parâmetro que possuem maior probabilidade de influência na geração de níveis

¹⁵ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/angrybirds/>

estáveis e jogáveis. Essas configurações foram aplicadas como restrições na abordagem evolutiva do AG.

A ARN para a estabilidade é mostrada na Figura 42 por um Grafo no qual o conhecimento extraído é renderizado visualmente. A ARN está disponível on-line ¹⁶.

Figura 42 – ARN com “Stability = 1.0”(true) como item alvo e nós de nível 1

Fonte: Elaborada pelo autor.

Considerando o item alvo “Stability = 1.0”(true) e analisando apenas o nível 1 da Rede, pode-se inferir que o número de porcos nas fases mais estáveis é igual a 3 (três), e esta é a quantidade máxima de porcos encontrados nas Regras de Associação geradas. Quando o número de porcos é menor que 3 (três), percebe-se que pode haver uma estabilidade da fase, mas esses elementos são precedidos por uma instabilidade ou “Stability = 0.0”(false).

O número de blocos que têm uma associação direta e nenhum antecessor para uma fase estável varia entre 13 e 18 (treze e dezoito) blocos. Isso indica que estruturas mais complexas com essas quantidades de blocos tendem a ser estáveis na geração de nível. A estrutura com 20 (vinte) blocos também apresenta estabilidade e, como se poderia esperar, a ausência de blocos ou a presença de um único bloco o torna um nível estável, uma vez que não possui elementos acima do solo da fase.

Outro fator essencial é a diversidade de blocos na fase, desta forma, foi contado o número de diferentes tipos de blocos que foram utilizados na construção das simulações. Os níveis que mantêm a variabilidade usando um número de 9 a 11 tipos diferentes de bloco são mais propensos a fornecer estabilidade, enquanto um número acima de 12 tipos diferentes de bloco não gerou Regras de Associação com suporte suficiente e confiança a serem consideradas.

¹⁶ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/angrybirds/>

Além dos parâmetros estudados para estabilidade (número de porcos, número de blocos e número de tipos de blocos), foi adicionada a quantidade de aves disponíveis ao jogador em cada nível a fim de avaliar a viabilidade das fases. A ARN foi definida com o objeto alvo “IsPlayable = 1.0”(true), indicando que a fase gerada é possível de ser executada pelo jogador com os recursos disponíveis. A ARN gerada possui um alto nível de densidade, portanto apenas uma parte da Rede é apresentada na Figura 43. A ARN completa está disponível on-line ¹⁷.

Figura 43 – Nível 1 sem predecessores de ARN com “IsPlayable = 1.0” (true) como item alvo e nós de nível 1

Fonte: Elaborada pelo autor.

Percebe-se que o número de porcos dos níveis mais prováveis a serem reproduzidos (viáveis) são aqueles que possuem 1, 2 ou 10 porcos. Extrapolando a observação para nós de nível 1 com o número de porcos igual a 3 (três) e 4 (quatro), porém eles têm como predecessor o nó “IsPlayable = 0.0”(false). Isto significa que níveis contendo 3 e 4 porcos também estão associados a fases nas quais a jogabilidade não pode ser garantida.

A quantidade de blocos foi contada e infere-se que os níveis com alto grau de jogabilidade estão entre aqueles com um número de blocos variando de 3 a 25 (três a vinte e cinco), assim como níveis com 27 (vinte e sete), 61 (sessenta e um), 64 (sessenta e quatro) e 68 (sessenta e oito). Desta forma, o AG foi configurado de modo que níveis com um número de blocos entre 28 e 60 podem se tornar impraticáveis em termos de jogabilidade.

¹⁷ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/angrybirds/>

A quantidade diferente de tipos de blocos está diretamente conectada à viabilidade da fase e os níveis mais prováveis de serem viáveis compreendem uma diversidade de blocos que variam de 2 a 11 tipos. As outras variações têm um índice de baixa reprodutibilidade, algumas das quais não são encontradas na ARN construída devido à baixa confiança ou porque são arestas inversas que são podadas na etapa de construção da Rede.

Um parâmetro adicionado ao estudo de viabilidade dos níveis gerados pelo simulador é a quantidade de aves disponíveis. Observa-se por meio da ARN que o número de aves necessário para tornar o nível viável é de 5 (cinco) e entre 8-13.

Todo o conhecimento gerado pela observação das ARNs é mostrado na Tabela 15). A formulação automática de hipóteses está diretamente relacionada às Regras de Associação conectadas ao item objetivo. A série de valores importantes inferidos pode permitir a configuração de algoritmos AG mais eficientes para geração de níveis.

Tabela 15 – Conhecimento extraído com uso da ARN

Atributo	Stability	Viability	Stability e Viability
Porcos	1,2,3,4	1,2,3,4,10	1,2,3,4
Blocos	0,1,13-18,20	3-25,27,61,64,68	13-18,20
Tipos de Blocos	9-11	2-11	9-11
Pássaros	-	5, 8-13	-

Fonte: Elaborada pelo autor.

Para validar a análise fornecida pela ARN, definiu-se a função de adequação do AG com $L = 20$ blocos e $B = 13$ aves para orientar a rotina de inicialização do AG. Os parâmetros sugeridos pela ARN geraram 43,14% de níveis estáveis. O número de níveis viáveis é ainda melhor, com 96,08%. Levando-se em conta a estabilidade e a viabilidade simultaneamente, o algoritmo gera 41,18% de níveis estáveis e jogáveis. Pode-se dizer que houve um ganho computacional com o uso de ARNs no processo de Mineração de Regras de Associação, quando comparado com métodos que não utilizam o conhecimento gerado por este tipo de análise.

4.4.2.3 Filtered-ARN

Para validação da abordagem *Filtered-ARN*, na elaboração da Rede filtrada, foram utilizadas as mesmas Regras de Associação geradas para a construção das ARNs. O processo de extração das Regras de Associação é similar às duas metodologias. Para a construção das *Filtered-ARNs*, foram utilizados os filtros de *Added Value* nulo e atribuído *mingain* = 0,01.

Logo após a construção das *Filtered-ARNs*, informações sobre os intervalos de valores para cada parâmetro foram coletadas para geração de níveis estáveis e jogáveis. Essas restrições possuem uma maior probabilidade de serem verdadeiras do que aquelas retiradas das ARNs devido ao uso de medidas objetivas auxiliares. Todas as restrições foram aplicadas no algoritmo

genético gerador das fases. A *Filtered*ARN para a estabilidade é mostrada na Figura 44, sendo também disponível on-line ¹⁸.

Figura 44 – *Filtered*-ARN com “Stability = 1.0”(true) como item alvo e nós de nível 1

Fonte: Elaborada pelo autor.

Considerando o item alvo “Stability = 1.0”(true) e analisando apenas o nível 1 da Rede, pode-se inferir que o número de porcos nas fases mais estáveis é igual a 2 (dois). Ocorrendo uma mudança em relação à ARN, devido a característica de poda da Rede em função do item objetivo.

O número de blocos que têm uma associação direta e nenhum antecessor para uma fase estável varia com um número fixo de 5 e no intervalo entre 11 e 17 blocos. Isso indica que estruturas mais complexas com essas quantidades de blocos tendem a ser estáveis na geração de nível. Do mesmo modo que a ARN, e por razões lógicas, a ausência de blocos ou a presença de um único bloco torna um nível estável.

A diversidade de blocos na fase também foi analisada, desta forma, foi contado o número de diferentes tipos de blocos que foram utilizados na construção das simulações. Os níveis mais estáveis eram os que apresentam 8, 9 ou 11 tipos diferentes.

Para o *dataset* de “Viability” foi construída a *Filtered*-ARN com o objeto alvo “IsPlayable = 1.0”(true), indicando que a fase gerada é possível de ser executada pelo jogador. A *Filtered*-ARN gerada possui um alto nível de densidade, portanto apenas uma parte da Rede é apresentada na Figura 45. A *Filtered*-ARN completa está disponível on-line ¹⁹.

¹⁸ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/angrybirds/>

¹⁹ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/angrybirds/>

Figura 45 – Nível 1 sem predecessores de *Filtered*-ARN com “IsPlayable = 1.0”(true) como item alvo e nós de nível 1

Fonte: Elaborada pelo autor.

De modo similar a ARN, percebe-se que o número de porcos dos níveis mais prováveis a serem reproduzidos (viáveis) são aqueles que possuem 1, 2 ou 10 porcos. Em relação a quantidade de blocos, verificou-se uma grande quantidade de intervalos e valores que podem ser observados na Tabela 16. Desta forma, o AG foi configurado levando-se em consideração apenas os valores encontrados pelo uso da Rede. A variabilidade dos tipos de blocos compreendem valores que variam de 3 a 6 e de 8 a 11 tipos, trazendo uma novidade em relação ao resultado encontrado na ARN e na Árvore de Decisão. Observa-se também, por meio da *Filtered*-ARN, que o número de aves necessário para tornar o nível jogável compreende-se nos intervalos de 7 a 9 e de 11 a 13. As informações encontradas por meio das *Filtered*-ARNs é mostrado na Tabela 16. Os valores inferidos foram utilizados na configuração do algoritmo genético.

Tabela 16 – Conhecimento extraído com uso da *Filtered*-ARN

Atributo	Stability	Viability	Stability e Viability
Porcos	2	1,2,10	2
Blocos	0,1,5,11-17	3-8,10-17,19-27,29-31,36,45,56,57,59,64	5,11-17
Tipos de Blocos	0,8,9,11	3-6,8-10	8,9
Pássaros	-	7-9,11-13	-

Fonte: Elaborada pelo autor.

A rotina de inicialização do AG foi orientada por uma função de adequação, alterando

apenas o número máximo de blocos para $L = 64$ blocos. Neste caso, foi encontrado um total de 82,80% de níveis estáveis e o número de níveis viáveis atingiu 100,00%. Pode-se dizer que o procedimento de inicialização do AG produziu apenas níveis com uma jogabilidade viável. Assim, decidiu-se alterar também o número de aves para $B = 13$ para avaliar níveis estáveis e viáveis simultaneamente. O número de níveis estáveis e viáveis atingiu 98,80%.

Com a geração de um total de 1000 níveis aleatoriamente em cada versão do algoritmo genético, foi feito o cálculo da porcentagem de fases estáveis e viáveis. Os resultados alcançados são apresentados na Figura 46. Demonstra-se, assim, que com o uso da *Filtered-ARN*, formulam-se hipóteses mais verdadeiras e que podem ser aplicadas no desenvolvimento de outras técnicas computacionais, gerando um ganho de processamento e de resultados mais promissores.

Figura 46 – Acurácias do *Angry Birds* para a abordagem com Algoritmo Genético

Fonte: Elaborada pelo autor.

4.4.3 Mammography dataset

A aquisição dos dados foi realizada a partir das imagens adquiridas no repositório digital de imagens desenvolvido pela Universidade do Sul da Flórida, que possui em sua coleção uma ampla gama de imagens digitalizadas e devidamente caracterizadas (DDSM). O *dataset* fornece uma estrutura de quadros de imagens, padronizada no tamanho de 230 x 350, discriminando as características de margens, contornos e classificação BI-RADS de cada exame, identificando nas imagens regiões de interesse (Regiões de Interesse - do inglês *Regions Of Interest* (ROI) - do inglês *Regions Of Interest*).

A caracterização do ROI extraído foi realizada pela construção de um *script*. Esse *script* serviu para identificar as bordas encontradas e extrair o tamanho (área em cm^2) das ROIs em cada imagem. As imagens foram padronizadas nas dimensões 230 x 350, conforme estabelecido pelo DDSM.

Ao final do procedimento de identificação de ROIs para calcificações individuais delimitando a área de interesse, foram distribuídas, no total, 150 amostras entre as 5 possibilidades de

classificação analisadas no escopo deste trabalho, de acordo com a nomenclatura e os critérios de diferenciação do BI-RADS (ACR, 2005).

Os parâmetros selecionados para a construção do *dataset* final foram: tamanho, margem, forma, densidade e classificação BI-RADS dos nódulos, além da idade do paciente. O *dataset* foi estruturado em um arquivo do tipo “.arff” disponível on-line²⁰.

No caso dos parâmetros de forma e margem, uma escala foi estabelecida para quantificar os dados de acordo com o grau dos mesmos atributos atribuídos ao BI-RADS. A Tabela 17 apresenta a escala adotada.

Tabela 17 – Quantificação de forma e margem.

Forma	Margem	Valor
Redondo	Circunscrita	1
Ovulado	Turvo	2
Lobular	Microlobulado	3
Irregular	Mal Definido	4
Distorção Arquitetônica	Cravada	5

Fonte: Elaborada pelo autor.

Foi construído o sistema híbrido proposto por Calçada *et al.* (2017), no qual é elaborada a combinação de um algoritmo genético (AG) formado por uma população de Redes Neurais Artificiais (RNA). A implementação deste sistema híbrido adota RNAs no modelo *MultiLayer Perceptron* (MLP) para classificação das mamografias.

Para a evolução das RNAs, foi implementado um AG com cromossomos vetoriais do tamanho definido pela Equação 4.2. Cada entrada neste cromossomo define um peso de RNA que deve, inicialmente, ser um valor contido no intervalo [-10,0; 10,0].

$$W_n = (I_n * H_n) + (H_n * O_n) + H_n + O_n \quad (4.2)$$

W_n representa o número total de genes no cromossomo (pesos totais da RNA), I_n o número de parâmetros de entrada da RNA, H_n o número de neurônios na camada oculta da RNA e O_n a quantidade de neurônios na camada de saída. O valor de *fitness* dos indivíduos da população foi obtido por meio do cálculo do erro quadrático médio, em que o indivíduo mais apto é aquele com menor valor de *fitness*.

Uma população inicial de 20 indivíduos é gerada com pesos selecionados aleatoriamente. A seleção para reprodução usa a estratégia de torneio entre 3 cromossomos, escolhidos aleatoriamente da população, na qual aquele com o melhor valor de aptidão é escolhido para se reproduzir. Após a seleção de dois cromossomos pelo torneio, o operador de *crossover* é executado para criar dois novos cromossomos. O *crossover* com ponto de corte foi aplicado,

²⁰ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/datasets/Breast_Cancer.zip>

no qual os pontos são definidos de forma aleatória. O operador de mutação é então aplicado com uma taxa de 10% alterando os valores nos genes de cada novo cromossomo criado. Na composição da nova população em cada interação, um novo cromossomo criado aleatoriamente e inserido no lugar do pior indivíduo da população. A introdução desse novo indivíduo busca manter a diversidade da população (MCCAFFREY, 2014). A seleção para sobrevivência usa uma população intermediária (*Offspring*), na qual as crianças são inseridas. Em seguida, os 20 (vinte) melhores indivíduos entre a população atual (pais) e filhos são selecionados. O processo evolutivo é executado até que um número de 2000 gerações seja alcançado ou o *fitness* do melhor indivíduo não sofra alteração após 10 gerações (critério de parada).

Ao fim do processo, foi selecionada a RNA de melhor *fitness* (melhor indivíduo) para cálculo da taxa de acerto ($100\% - \text{fitness}$). O valor obtido nessa configuração foi de 96%. Mais detalhes sobre essa implementação do AG podem ser encontrados em Calçada *et al.* (2017).

4.4.3.1 Árvore de Decisão

Foi realizado um estudo dos parâmetros por meio da aplicação do algoritmo de árvore de decisão. Neste estudo, foi utilizada a ferramenta Weka por meio do algoritmo de árvore de decisão J48. A aplicação da técnica de mineração de dados por meio da árvore de decisão possibilitou a identificação do parâmetro de densidade com a maior probabilidade de importância, porque esse elemento forma a raiz da árvore e não é encontrado em nenhum nó folha. A árvore de decisão está disponível on-line²¹. Desta forma, decidiu-se inserir uma nova entrada para as RNAs que formam os indivíduos do AG formada pelos valores da densidade. Desta forma o número de entradas das RNAs passaram de 5 para 6, sendo a densidade inserida duas vezes.

Com a inserção da nova entrada, ao fim do processo, foi selecionada a RNA de melhor *fitness*. Com essa configuração houve um aumento na taxa de acerto para 97,33%. Demonstrando que o sistema é sensível a configurações que podem melhorar seu desempenho.

4.4.3.2 ARN

Foi executada a geração das Regras de Associação com o uso do algoritmo *Apriori*, utilizando-se *minsup* nulo para que todos os parâmetros fossem aceitos e *minconf* = 0,1 por se tratar de um *dataset* relativamente pequeno com 150 instâncias. Em seguida foi construída a ARN com item objetivo “[BI-RADS] = 5” por se tratar do grau mais grave desta escala (ACR, 2005). O resultado parcial pode ser observado na Figura 47 e a ARN completa disponível on-line²².

A análise foi feita observando-se os elementos ligados diretamente ao nó objetivo, totalizando 50 hipóteses de influência para o diagnóstico de câncer. Com esse resultado o algoritmo genético foi configurado com uma penalidade no *fitness* dos indivíduos que não possuíam nenhuma dessas característica, obtendo uma taxa de acerto de 97,486%.

²¹ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/mammography/>

²² <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/mammography/>

Figura 47 – ARN com “[BI-RADS] = 5” como item alvo e nós de nível 1 (vermelho)

Fonte: Elaborada pelo autor.

Comparando esse resultado com as execuções anteriores do AG, percebe-se uma melhoria na acurácia, o que demonstra uma otimização do processo pela aquisição do conhecimento obtido com a mineração dos dados com uso de Redes.

4.4.3.3 Filtered-ARN

Fazendo uso das regras obtidas na etapa anterior, foi construída a *Filtered-ARN*, na qual todas as regras com *Added Value* nulo ($AV = 0$) e com $mingain \leq 0,01$ são excluídas. O resultado parcial pode ser observado na Figura 48 e a *Filtered-ARN* completa on-line²³.

São observados 13 itens conectados diretamente ao nó objetivo, os quais foram utilizados para a configuração do algoritmo genético. O resultado foi uma melhora na acurácia para 98,731%, sendo ainda maior que o AG configurada pelos parâmetros obtidos na ARN, indicando um melhor estudo de atributos que se relacionam diretamente ao diagnóstico de câncer de mama.

Ao se comparar todos os resultados encontrados em cada etapa (Figura 49), percebe-se que o uso de processos de Mineração de Regras de Associação pode melhorar os sistemas de auxílio a diagnósticos, sendo promissora a aplicação de tais técnicas com o uso de Redes por possibilitar uma geração de conhecimento/hipóteses mais compreensíveis, principalmente quando são adotados métodos matemáticos para a seleção das regras, como nesta abordagem com *Filtered-ARNs*.

²³ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/mammography/>

Figura 48 – *Filtered-ARN* com “[BI-RADS] = 5” como item alvo e nós de nível 1 (vermelho)

Fonte: Elaborada pelo autor.

Figura 49 – Acurácia obtida em cada etapa dos experimentos com o dataset de mamografias

Fonte: Elaborada pelo autor.

4.5 Considerações Finais

Neste capítulo foi apresentada a proposta da Rede de Regras de Associação Filtrada (*Filtered-ARN*), um método capaz de modelar as Regras de Associação previamente selecionadas, por meio de Medidas Objetivas Assimétricas, de acordo com um item objetivo previamente definido. As regras selecionadas para a construção da Rede são aquelas que possuem uma

dependência estatística comprovada pela medida *Added Value* (AV). O item objetivo é utilizado como norteador da exploração e é escolhido de acordo com a problemática sobre a qual se deseja formular hipóteses. A *Filtered-ARN* cria um Hipergrafo Direcionado, modelando as Regras de Associação que têm o item objetivo no RHS recursivamente. Ele visa explicar a correlação entre os itens no *dataset* com o item objetivo.

Uma avaliação experimental foi desenvolvida com 3 *datasets* para validar a capacidade da *Filtered-ARN* de auxiliar na geração de hipóteses otimizadas e aplicáveis. *Datasets* disponíveis on-line foram explorados: *Lenses*, *Hayes-roth* e *Soybean Large*. Foram realizadas também três provas de conceito com *datasets* reais. O primeiro com dados sobre adubação orgânica (*Green Manure*) obtido na EMBRAPA Meio-Norte, o segundo relacionado ao processo de geração automática de conteúdo processual (PCG) e o terceiro *dataset* com características de exames mamográficos classificados de acordo com um padrão internacional médico. Além da *Filtered-ARN*, foram aplicados os algoritmos J48 e a ARN para comparação. O objetivo foi descrever a ocorrência de regras que influenciam estatisticamente a classe desejada, visando encontrar os itens que melhor expliquem a ocorrência do item da classe.

Também foi comparado o algoritmo *Filtered-ARN* ao J48, que é um algoritmo de árvore de decisão disponível no Weka. Como a árvore de decisão realiza apenas a classificação dos elementos, sua estrutura não é muito boa em explicar as relações. A árvore de decisão gerada no *dataset Soybean* é um exemplo disso, pois a árvore conseguiu 87,58% de precisão, porém com um tamanho elevado (69 folhas), o que dificultou o entendimento do usuário com relação às correlações. A *Filtered-ARN* obteve bons resultados, descrevendo os dados por meio do uso das Regras de Associação extraídas e selecionadas e mostrando ao usuário os itens que são inteiramente responsáveis pela ocorrência do item objetivo, gerando influência no mesmo.

Os resultados mostraram que a ARN é capaz de descrever relações com um item objetivo, no entanto, ele não mostra ao usuário quais os casos em que os itens verdadeiramente influenciam estatisticamente o item objetivo. É importante destacar esses casos porque algumas explorações são feitas com item objetivo e hipóteses equivocadas são elaboradas ocasionando prejuízos na extração do conhecimento, como a não veracidade deste. A *Filtered-ARN* é bem sucedida em apresentar esses casos, permitindo que o usuário veja os itens que produzem variações no item objetivo.

Para o *dataset Green Manure*, os especialistas validaram os resultados. Apenas as hipóteses formuladas pela observação da estrutura da *Filtered-ARN* foram confirmadas em sua totalidade, sendo inclusive verificada uma nova possibilidade de estudo agrícola para os pesquisadores envolvidos.

Os resultados relatados neste capítulo possibilitaram as respostas para as questões de pesquisa **Q2** (*Como é feita a validação da extração do conhecimento nos processos de Mineração de Regras de Associação?*), **Q4** (*Quais as Medidas Objetivas utilizadas nos processos de Mineração de Regras de Associação que fazem uso de estruturas de Rede?*) com suas questões

secundárias e **Q6** (*Qual o impacto nos resultados da geração de hipóteses em processos de Mineração de Regras de Associação com o uso de Redes?*), atendendo ao objetivo desta pesquisa.

REDES DE REGRAS DE ASSOCIAÇÃO MULTIALVO

5.1 Considerações Iniciais

A descoberta de novos conhecimentos em *datasets* com mais de uma classe pelo uso de técnicas de mineração de dados é crucial para fazer um uso efetivo dos dados armazenados e fornecer uma descrição clara e compreensível dos padrões existentes. Assim, a semântica das Regras de Associação possibilita uma melhor leitura da informação, porém, quanto maior a quantidade de elementos dos conjuntos de *itemsets* que formam uma regra, menor a capacidade de interpretação da mesma (DELGADO *et al.*, 2018).

Como os algoritmos de Regras de Associação são capazes de extrair todas as Regras de Associação de acordo com um suporte mínimo e valor mínimo de confiança, o número de regras extraídas geralmente supera a capacidade de interpretação do usuário. O conhecimento gerado muitas vezes é inconclusivo. Sendo assim, a busca de métodos que produzam um resultado de fácil interpretação é de extrema importância em aplicações (LIPTON, 2016). Nesta pesquisa propõe-se utilizar estruturas de Redes para facilitar a interpretação dos resultados.

A Rede de Regras de Associação Filtradas (*Filtered-ARN*), proposta no Capítulo 4, é um Grafo Direcionado construído para modelar todas as regras a fim de descrever um item selecionado. O resultado é um gráfico que explica as influências a um determinado item e possibilita que o usuário construa hipóteses com base nesse elemento selecionado. Este item selecionado é chamado de “item objetivo”, pois se torna o alvo da exploração do *dataset*. A *Filtered-ARN* oferece as seguintes características: (1) filtragem das regras, (2) poda no contexto, (3) estrutura de Rede e (4) geração de hipóteses para avaliação. A filtragem é realizada com o uso de Medidas Objetivas Assimétricas a fim de excluir as regras nas quais não existe influência entre o antecedente e o conseqüente da regra. A poda é feita de acordo com o item objetivo,

na qual a Rede é modelada considerando apenas as regras que estão correlacionadas direta ou indiretamente a esse item. A estrutura de Rede fornece um mecanismo para determinar o relacionamento entre os itens que é mais fácil para o usuário entender.

A *Filtered-ARN*, no entanto, só pode analisar um item por execução, o que pode ser uma séria limitação em algumas bases de dados. Em *datasets* de pacientes com problemas cardíacos, deseja-se por exemplo, identificar os fatores que geram complicações e quais destes estão diretamente ligados ao óbito, ou seja dois itens a serem analisados simultaneamente.

Para possibilitar uma exploração mais completa, mostrando ao usuário quais fatores influenciam em um conjunto de dois itens, esta tese propõe a Rede de Regras de Associação Multialvo (MTARN - do inglês *Multi-Target Association Rules Network*), que permite a exploração de dois itens objetivo em conjunto. A abordagem multialvo da MTARN apresenta uma explicação de quais fatores influenciam diretamente no conjunto analisado. A MTARN expande os recursos presentes na *Filtered-ARN*, adicionando algumas propriedades interessantes.

A fim de validar a MTARN, foi realizada uma avaliação experimental com 3 *datasets* e os resultados foram comparados com duas ARNs, uma para cada item objetivo, bem como com duas *Filtered-ARNs*. Uma Prova de Conceito também foi realizada com o intuito de comprovar a eficácia da técnica proposta na mineração de dados com *datasets* reais. Os resultados demonstraram que a MTARN é capaz de descrever os elementos que influenciam dois itens objetivos de modo mais conciso se comparado a ARN e *Filtered-ARN*. Isso permite que o usuário observe os casos em que mais de um elemento interfere estatisticamente em um conjunto objetivo.

Este capítulo é dividido da seguinte forma. Na Seção 5.2, é descrito o processo de Mineração de Regras de Associação com o uso de Redes Multialvo, bem como a metodologia para construção das Redes de Regras de Associação Multialvo (MTARNs). Na Seção 5.3, é detalhada a Avaliação Experimental com 3 *datasets* oriundos da UCI, bem como toda a metodologia realizada nos experimentos de estudo da extração do conhecimento pelo uso da MTARN. Na Seção 5.4, é apresentada e avaliada a Prova de Conceito e os resultados da mineração com o *dataset* real de pacientes submetidos a cirurgias de revascularização miocárdica.

5.2 Modelo Proposto para Geração de Hipóteses em Problemas Multialvo

Existem casos em que a Mineração de Regras de Associação é feita com o objetivo de explicar itens predeterminados. A *Filtered-ARN* apresenta uma exploração guiada por um único item objetivo. Esta exploração remove todas as regras que não são interessantes no contexto do item objetivo, de acordo com as métricas mínimas de suporte e confiança, e também com as Medidas Assimétricas *Added Value* e *Gain*, mostrando ao usuário apenas as regras relevantes.

Embora as *Filtered-ARNs* promovam a identificação dos fatores relacionados a um determinado problema, alguns estudos necessitam da análise simultânea de dois fatores em conjunto. Sendo assim, a *Filtered-ARN* não consegue apresentar a relação de influência direcionada a mais de um objetivo simultaneamente.

Com o objetivo de possibilitar uma exploração completa e considerando as relações existentes entre conjuntos de itens objetivo, nesta pesquisa é proposta a Redes de Regras de Associação Multialvo (MTARN), que permite a exploração de dois itens objetivo simultaneamente com análise de dependência entre os elementos das regras.

No algoritmo de construção da MTARN, faz-se uso de Regras de Associação com consequente de tamanho fixo igual a 2 e antecedentes de tamanhos 1 ou 2, além de filtros com medidas objetivas assimétricas (*Added Value* e *Gain*), construindo o gráfico de acordo com as regras selecionadas.

O Algoritmo de construção da MTARN é descrito em 3 passos:

- (1) Passo A: similar ao primeiro passo de todos os processos de Mineração de Regras de Associação, com a restrição de formação de regras com consequentes formados por 2 itens ($|RHS| = 2$) e antecedentes formados por 1 ou 2 itens ($|LHS| \leq 2$).
- (2) Passo B: efetuar o cálculo das Medidas Objetivas Assimétricas *Added Value* e *Gain*, realizando a exclusão de todas as regras com $AV = 0$ e que possuam $Gain < mingain$.
- (3) Passo C: escolher um item frequente Z de tamanho igual a 2, que será representado no conjunto de regras como o nó objetivo. Construir uma Rede Direcionada que flui de forma recursiva para Z .

Definição 1: dado um conjunto de Regras de Associação R , contendo regras de *itemsets* de tamanho igual a 2 no consequente, e um item objetivo Z , a MTARN é uma Rede Direcionada que modela todas as regras relacionadas ao item em Z (Multialvo), como:

1. Cada aresta modela uma regra $r \in R$.
2. A partir de qualquer ponto da Rede, é sempre possível alcançar pelo menos 1 vértice representando um item Z .
3. Dado um vértice $v \in MTARN$, com $v \notin Z$. Não há caminho de qualquer item Z para v .
4. Se existe uma regra r como $RHS(r) \subset Z$, então a regra $r \in MTARN$.

Além disso, a MTARN é construída de acordo com os níveis de seus vértices.

Definição 2: O nível de um determinado vértice $v \in MTARN$ é o número de arestas necessárias para acessar o item Z .

Por exemplo, o item alvo Z tem nível zero ($Nível = 0$), pois ele não precisa passar por nenhuma aresta para alcançar o próprio item Z . Itens na parte LHS de regras que possuem $RHS \subset Z$ terá o nível um ($Nível = 1$). Eles estão a uma aresta de distância dos itens em Z .

A exclusão das regras de dependência estatística nula ($AV = 0$) é uma propriedade fundamental das MTARNs, pois garante que todos os itens que participam da Rede geram algum tipo de influência no item objetivo.

Na Figura 50 é representado um exemplo de MTARN, na qual foi selecionado o item “B,C” como objetivo. Seleciona-se então todas as regras que possuem “B,C” como consequente, neste caso apenas as regras ($A \Rightarrow B,C$), ($A,D \Rightarrow B,C$) e ($D,E \Rightarrow B,C$). Assim, os itens “A”, “A,D” e “D,E” são modelados no nível 1 da MTARN e passam a ser os objetivos, passando a executar o algoritmo de forma recursiva até que não restem mais itens como objetivos nos níveis mais altos da abordagem. Neste caso, foram modeladas as regras que possuem “D,E” como objetivo, sendo inseridos as regras ($F \Rightarrow D,E$) e ($F,G \Rightarrow D,E$). Nesse exemplo, não existem regras que possuem “A,D” ou “F,G” como consequente. A abordagem Multialvo formada por itemsets com $|RHS| = 2$ e $|LHS| \leq 2$ não permite a formação de regras com os nós “A” e “F” como consequentes, portanto estes elementos nunca possuirão antecedentes.

Figura 50 – Exemplo de uma MTARN com “B,C” como itens-alvo

Fonte: Elaborada pelo autor.

Com o objetivo de validar a abordagem MTARN e apresentar sua capacidade de exploração de *datasets*, foram realizados alguns experimentos em bases de dados conhecidas, bem como uma prova de conceito com um *dataset* real.

5.3 Avaliação Experimental

Os experimentos realizados tiveram como objetivo apresentar as diferenças entre métodos com apenas 1 item alvo e com 2 itens simultâneos, mostrando como a MTARN é capaz de possibilitar uma exploração mais ampla dos dados, dando ao usuário uma compreensão completa dos mesmos. Além disso, a abordagem multialvo da MTARN possibilita o estudo de parâmetros de influência simultâneos em dois itens objetivo. A fim de comparar seus resultados com a saída da MTARN, foram construídas também ARNs e *Filtered-ARNs*.

Para validação da MTARN, foram utilizados três *datasets* da UCI ¹ (*Iris*, *Hayes-roth* e *SoyBean Large*), além de uma Prova de Conceito com *dataset* real (*Coronary dataset*) coletado junto aos pesquisadores do Instituto do Coração Dante Pazzanese, na cidade de São Paulo, estado de São Paulo. Para a seleção das regras foi estabelecido um valor de ganho mínimo (*mingain*) igual a 0,01, efetuando-se a filtragem das regras que possuem $AV = 0$. Para a construção da Rede foi usado o processo recursivo descrito anteriormente.

Para a escolha dos *datasets*, foi consultada a Tabela 3 de Taxa de Complexidade, já descrita no Capítulo 3 desta tese. Para a realização dos experimentos apresentados neste trabalho, foram selecionados três conjuntos de dados, sendo um de baixa complexidade (IRIS), um de média complexidade (HAYES-ROTH) e outro de alta complexidade (SOYBEAN LARGE). Com os 3 *datasets* foram realizadas as mesmas etapas para construção das MTARNs.

O primeiro estudo foi utilizando o *dataset Iris*, disponível na UCI². Este é um dos *datasets* mais conhecidos encontrado na literatura sobre reconhecimento de padrões. O conjunto de dados contém 3 classes de 50 instâncias cada, no qual cada classe se refere a um tipo de planta de íris. Uma classe é linearmente separável das outras duas; os últimos não são linearmente separáveis um do outro.

O segundo estudo foi realizado no *dataset Hayes-roth*, disponível on-line³. O conjunto de dados traz informações geradas por dados pessoais, como idade e nível de escolaridade. Esse banco de dados contém 6 atributos de valor numérico, sendo o primeiro apenas um índice, que foi excluído. O último atributo divide as instâncias em 3 classes.

O terceiro estudo foi feito no *dataset Soybean (Large)*, também disponível na UCI⁴. O *dataset* contém os resultados de uma pesquisa de características de doenças em plantações de soja. Existem 19 classes referentes às características das plantas doentes. Em razão do número limitado de exemplos de determinadas classes, apenas as 15 primeiras foram usadas nos experimentos realizados. As últimas quatro classes são pouco exploradas, pois têm poucos exemplos. Existem 35 atributos categóricos, alguns nominais e alguns ordenados. Os valores dos atributos são codificados numericamente, com o primeiro valor codificado como “0”, o segundo como “1”, e assim por diante. O *dataset* possui campos “vazios”.

Nos experimentos realizados, as regras foram extraídas com o uso do algoritmo *Apriori-TID*, implementado em Java®. Para construção das ARNs e *Filtered-ARNs* foram extraídas regras de tamanho 2, com apenas 1 item em RHS e 1 item em LHS. Para a construção das MTARNs um novo processo de extração de Regras de Associação foi executado para que fossem obtidas regras de tamanho ≤ 4 com consequente (RHS) de tamanho igual a 2 (Definição 1). Os parâmetros de *minsup* e *minconf* foram os mesmos utilizados na construção das ARNs e

¹ <<http://archive.ics.uci.edu/ml/>>

² <<https://archive.ics.uci.edu/ml/datasets/iris>>

³ <<https://archive.ics.uci.edu/ml/datasets/Hayes-Roth>>

⁴ <<http://archive.ics.uci.edu/ml/datasets/Soybean+%28Large%29>>

Filtered-ARNs.

Após esta etapa as regras foram filtradas considerando a medida *Added Value*, tal que foram desconsideradas todas aquelas que possuem valor nulo ($AV = 0$), por indicar ausência de influência estatística. Além disso, foi selecionado o ganho mínimo de 0,01, para que o maior número de regras possível pudesse ser analisado, sendo então executado o filtro dessa medida, o qual descarta todas as regras cujo $Gain < mingain$. Com as regras filtradas foram construídas as *Filtered-ARNs* e *MTARNs*.

Todas as Redes foram construídas graficamente com o uso do software *Gephi* (BASTIAN; HEYMANN; JACOMY, 2009). O *Gephi* é uma aplicação *open source* específica para construção de Redes que possui vários algoritmos de visualização (*Layout*) já implementados e está disponível on-line⁵.

5.3.1 Iris dataset

Como o *dataset* tem um pequeno número de atributos, o suporte mínimo foi definido como 0, portanto, todos os valores foram considerados. A confiança mínima foi definida para 0,25, possibilitando o estudo de classes que ocorreram pelo menos 1 em cada 4 vezes. Para construção da *MTARN*, foram extraídas regras de tamanho ≤ 4 com consequente (RHS) de tamanho igual a 2 ($|RHS| = 2$). Usando esta configuração, foram obtidas 2.286 regras candidatas para o *dataset* “*Iris*”. Após a etapa de filtragem restaram 1.762 regras.

Para construção das *ARNs* e *Filtered-ARNs*, o tamanho da regra foi especificado em dois, considerando um item no LHS e um item no RHS. Usando essa configuração, foram obtidas 281 Regras de Associação candidatas para o *dataset* “*Iris*”. Após a etapa de filtragem restaram 266 regras.

Foram geradas duas *ARNs* com itens alvo “[class]=iris-virginica” (Figura 51) e “[petal-width]=12” (Figura 52), respectivamente. As Redes geradas possuem graus diferentes do nó alvo, sendo $Grau = 21$ para a *ARN* com objetivo “[class]=iris-virginica” e $Grau = 8$ para a *ARN* com alvo “[petal-width]=12”. Foram analisados apenas os nós ligados diretamente ao item objetivo (Nível = 1).

Observando as duas *ARNs*, infere-se que o item objetivo de uma está presente no nível 1 da outra. Todos os nós de nível 1 da *ARN* com “[petal-width]=12” como item alvo estão no nível 1 da *ARN* com “[class]=iris-virginica” como item alvo. Embora já seja conhecido que *ARNs* podem apresentar arestas com regras sem dependência entre os itens, a geração dessas Redes é importante para estudo do impacto que o uso das *MTARNs* provocam na extração do conhecimento com múltiplos elementos em um mesmo *itemset*.

As *Filtered-ARNs* foram geradas considerando os mesmos itens alvos das *ARNs*, “[class]=iris-virginica” (Figura 53) e “[petal-width]=12” (Figura 54). As Redes geradas apresen-

⁵ <<https://gephi.org/users/download/>>

Figura 51 – ARN com “[class]=iris-virginica” como item alvo

Fonte: Elaborada pelo autor.

taram grau uma unidade menor no nó alvo. Na *Filtered-ARN* com “[class]=iris-virginica” como item alvo, o nó “[sepal-width]=10” não é apresentado. Da mesma forma para a *Filtered-ARN* com “[petal-width]=12” como item alvo, não é apresentado o nó “[petal-length]=12”. Isso ocorre porque esses dois nós não influenciam diretamente os respectivos itens objetivo ($AV = 0$).

A MTARN foi gerada com “[petal-width]=12.[class]=iris-virginica” como item objetivo, devido a estes elementos formarem regras com confiança de 100% entre si. Na Rede resultante que pode ser observada na Figura 55, o nó alvo ficou com um grau 32. Analisando os nós de nível 1 sem predecessores, percebe-se a presença de 6 itens unitários.

Os nós de nível 1 sem antecedentes, formados por apenas 1 elemento, confirmam o conhecimento já observado nas *Filtered-ARNs* com a presença dos itens “[petal-length]=9”, “[petal-length]=10”, “[petal-length]=11”, “[sepal-length]=9”, “[sepal-length]=12” e “[sepal-width]=7”. Infere-se portanto que a MTARN mantém o conhecimento sobre itens que individualmente conseguem influenciar diretamente o conjunto alvo formado por 2 elementos.

Ao se observar itens com dois elementos em nós do nível 1 da MTARN, percebe-se o surgimento de novos conhecimentos, além da consolidação de outros. Os nós “[sepal-length]=7.[petal-length]=11” e “[sepal-length]=8.[sepal-width]=7” possuem informações de parâmetros que influenciam individualmente o item alvo com o complemento de outro elementos que podem fortalecer esta influência.

Figura 52 – ARN com “[petal-width]=12” como item alvo

Fonte: Elaborada pelo autor.

O nó “[sepal-width]=4.[petal-length]=12” apresenta o item “[petal-length]=12” que individualmente não gera influência em nenhum dos itens alvo ($AV = 0$), mas que pode ter uma influência quando associado ao parâmetro “[sepal-width]=4”.

Nós como “[sepal-length]=7.[sepal-width]=8” e “[sepal-length]=10.[sepal-width]=9” apresentam conjuntos de itens formados por elementos da ARN com outros inéditos, como é o caso de “[sepal-width]=8” e “[sepal-width]=9”. O número de possibilidades de geração de hipóteses aumenta consideravelmente ao se estudar os padrões ligados diretamente a um nó objetivo formado por mais de um item.

Além de elementos já conhecidos das ARNs e *Filtered-ARNs*, observa-se que a MTARN gera conhecimento inédito de padrões. Por exemplo, o nó “[sepal-length]=6.[sepal-width]=5” é formado por elementos que individualmente não são apresentados em nenhuma das Redes que analisam os padrões com apenas 1 item objetivo. Esta informação possibilita afirmar que o uso da MTARN viabiliza ao usuário novos conhecimentos de modo a produzir novas hipóteses com grande probabilidade de serem verdadeiras.

Figura 53 – *Filtered-ARN* com “[class]=iris-virginica” como item alvo

Fonte: Elaborada pelo autor.

5.3.2 Hayes-roth dataset

O *dataset Hayes-roth* tem 6 atributos, sendo que os 2 primeiros foram gerados aleatoriamente, então eles foram removidos. Como o *dataset* tem apenas 162 instâncias, o suporte mínimo foi definido como 0, portanto, todos os valores foram considerados. A confiança mínima

Figura 54 – *Filtered-ARN* com “[petal-width]=12” como item alvo

Fonte: Elaborada pelo autor.

Figura 55 – MTARN com “[petal-width]=12.[class]=iris-virginica” como item alvo

Fonte: Elaborada pelo autor.

foi definida para 0.2, possibilitando o estudo de classes que ocorreram pelo menos 1 em cada 5 vezes.

Como mencionado na Seção 5.3, para construção da MTARN, foram extraídas regras de tamanho ≤ 4 com consequente (RHS) de tamanho igual a 2. Usando esta configuração, foram obtidas 956 regras candidatas para o dataset “Hayes-roth”. Após a etapa de filtragem restaram 834 regras. Além disso, para a construção das ARNs e *Filtered*-ARNs, o tamanho da regra foi especificado em dois, considerando 1 item no LHS e 1 item no RHS. Usando essa configuração, foram obtidas 167 Regras de Associação candidatas para o dataset “Hayes-roth”. Após a etapa de filtragem restaram 123 regras. Foram escolhidos elementos que geram uma regra com confiança de 100% entre si. Sendo assim os itens objetivos selecionados foram “[class]=3” e “[marital_status]=3”.

Na Figura 56 é apresentada a ARN com “[class]=3” como item objetivo e na Figura 57 é apresentada a ARN com “[marital_status]=3” como item objetivo. As Redes apresentam graus diferentes para os itens alvo, sendo $Grau = 9$ e $Grau = 3$, respectivamente. Foram analisados apenas os nós ligados diretamente ao item objetivo (Nível = 1).

Observando as duas ARNs, infere-se que o item objetivo de uma está presente no nível 1 da outra. Apenas o item “[hobby]=1” é apresentado no nível 1 de ambas as ARNs, o que pode ser um indício de que apenas essa característica pode gerar um padrão de intersecção entre os itens alvo.

As *Filtered*-ARNs foram geradas considerando os mesmos itens alvos das ARNs, “[class]=3” (Figura 58) e “[marital_status]=3” (Figura 59). As Redes resultantes apresentaram graus diferentes para seus nós alvos em relação a cada ARN, por retirar as regras sem

Figura 56 – ARN com “[class]=3” como item alvo

Fonte: Elaborada pelo autor.

influência comprovada ($AV = 0$). Na *Filtered*-ARN com “[class]=3” como item alvo, o grau para o nó objetivo caiu para *Grau* = 3 e na *Filtered*-ARN com “[marital_status]=3” como item alvo, o nó objetivo apresentou *Grau* = 2. Percebe-se que todos os nós com o parâmetro “[hobby]” foram excluídos, portanto esta informação não gera nenhum tipo de influência nos itens alvo de modo individual.

A MTARN foi gerada com o item objetivo “[marital_status]=3.[class]=3”. Na Rede resultante, que pode ser observada na Figura 60, verifica-se que o nó alvo ficou com um grau 8. Analisando os nós de nível 1, percebe-se que nenhum item ficou sem antecedentes, portanto não existem nós com *itemsets* unitários no nível 1 da Rede.

Analisando os nós de nível 1 com grau de entrada igual a 1, observa-se 3 conjuntos, “[age]=2.[educational_level]=4”, “[age]=4.[educational_level]=1” e “[age]=4.[educational_level]=2”. Estes nós indicam que os parâmetros relacionados a idade e nível educacional são os que mais influenciam simultaneamente na classificação = 3 em conjunto com o *status* matrimonial = 3. Este conhecimento pode ser observado nas *Filtered*-ARNs, portanto a MTARN consegue manter a informação de padrões vistos anteriormente, porém com os itens em conjunto.

O nó “[age]=1.[educational_level]=4” apresenta grau de entrada igual a 2 na MTARN, indicando que este nó sofre influência de outros dois conjuntos de elementos. O parâmetro “[age]=1” é totalmente inédito no estudo dos padrões realizado pelas ARNs e *Filtered*-ARNs. Portanto a MTARN possibilitou a geração da hipótese que “pessoas com idade na categoria 1 e nível educacional igual a 4 são classificadas na classe 3 com *status* matrimonial 3”. Essa hipótese não poderia ser formulada sem a formação de Redes Multialvo.

Figura 57 – ARN com “[marital_status]=3” como item alvo

Fonte: Elaborada pelo autor.

Figura 58 – *Filtered-ARN* com “[class]=3” como item alvo

Fonte: Elaborada pelo autor.

A presença de nós com o parâmetro “[hobby]” também ocorrem no nível 1 da MTARN. Os *itemsets* “[hobby]=3.[age]=4”, “[hobby]=1.[age]=4”, “[hobby]=1.[educational_level]=4” e “[hobby]=2.[educational_level]=4” promovem a construção de hipóteses com itens que individualmente não influenciam diretamente a nenhum dos elementos que formam o item alvo. Sendo assim, a MTARN promove a formação de novos conhecimentos ao final do processo de Mineração de Regras de Associação.

5.3.3 Soybean Large dataset

O *dataset Soybean Large* possui 19 classes, 36 atributos e 307 instâncias. Como algumas instâncias possuem campos vazios, o suporte mínimo foi definido em 0,03 de forma a evitar o aparecimento de classes com informações incompletas. A confiança mínima foi definida para

Foram geradas duas ARNs com itens alvo “[class]=rhizoctonia” e “[fruit-pods]=3”. Alguns trechos relevantes das Redes podem ser observados nas Figuras 61 e 62 respectivamente. As ARNs completas estão disponíveis on-line⁶.

Figura 61 – ARN com “[class]=rhizoctonia” como item alvo

Figura 62 – ARN com “[fruit-pods]=3” como item alvo

Fonte: Elaborada pelo autor.

Analisando as duas ARNs, observa-se que o item alvo de uma está contido no nível 1 da outra. A ARN com “[class]=rhizoctonia” como item alvo possui um grau para o nó objetivo igual a 6, e a ARN com “[fruit-pods]=3” como item alvo tem um grau igual a 17 para seu nó objetivo. Todos os nós de nível 1 da ARN com “[class]=rhizoctonia” como item alvo estão inseridas no nível 1 da ARN com “[class]=rhizoctonia” como item alvo.

As *Filtered*-ARNs foram geradas com os mesmos itens alvos utilizados na construção das ARNs, “[class]=rhizoctonia” e “[fruit-pods]=3”. Alguns trechos mais relevantes das Redes geradas são apresentados nas Figuras 63 e 64, respectivamente. Essas Redes apresentaram graus diferentes para os itens objetivos quando comparadas às ARNs. Na *Filtered*-ARN com “[class]=rhizoctonia” como nó alvo, o grau desse elemento caiu para 3 e na *Filtered*-ARN com “[fruit-pods]=3” como item alvo, o grau do nó objetivo foi igual a 12. As *Filtered*-ARNs completas estão disponíveis on-line⁷.

Algumas mudanças estruturais são observadas em relação aos níveis dos nós. Por exemplo, os itens “[severity]=2” e “[temp]=0” subiram de nível da ARN para a *Filtered*-ARN com “[fruit-pods]=3” como item alvo. Todas as características relacionadas a manchas de folhas da

⁶ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/capitulos/capitulo5.zip>

⁷ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/capitulos/capitulo5.zip>

Figura 63 – *Filtered-ARN* com “[class]=rhizoctonia” como item alvo

Fonte: Elaborada pelo autor.

Figura 64 – *Filtered-ARN* com “[fruit-pods]=3” como item alvo

Fonte: Elaborada pelo autor.

soja (“[leafspots-marg]=2”, “[leafspot-size]=2”, “[leafspots-halo]=0”), bem como o parâmetro “[date]=0”, foram excluídas para a construção da *Filtered-ARN* por não apresentarem influência em relação ao item alvo.

Para a construção da MTARN, foi definido como item objetivo “[class]=rhizoctonia.[fruit-pods]=3”. Na Figura 65 é apresentado o nó alvo com os nós de nível 1. A MTARN formulada para o dataset “Soybean” obteve uma grande densidade, sendo o grau do nó objetivo igual a 90. A Rede completa está disponível on-line⁸.

Os elementos “[leaves] = 0” e “[stem-cankers] = 1” estão inseridos no nível 1 da MTARN. Com estas conexões, infere-se que a MTARN mantém as informações mais pertinentes conectadas aos elementos que formam o nó alvo, uma vez que a mesma influência aparece nas *Filtered-ARN*s.

Das conexões de nível 1 geradas, destaca-se o item “[date]=1.[temp]=0”. A conexão apresenta o parâmetro “[date]=1”, que havia sido excluído pelo filtro *Added Value* de forma individual, gerando influência no item alvo quando sua ocorrência é simultânea ao item “[temp]=0”.

⁸ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/capitulos/capitulo5.zip>

Figura 65 – Nós de nível L = 1 da MTARN com “[class]=rhizoctonia.[fruit-pods]=3” como item alvo

Fonte: Elaborada pelo autor.

Verificando os antecedentes desta regra, percebe-se a incidência da classe “[class]=rhizoctonia” unida com outros parâmetros, formulando-se assim, novas hipóteses que não são possíveis sem a abordagem multialvo.

Novos conhecimentos são formulados, mesmo com parâmetros que não apareciam individualmente na ARN nem na *Filtered-ARN*, como é o caso dos nós “[area]=1.[extern-decay]=1” e “[canker]=1.[fruiting]=0”. Além disso, parâmetros relacionados a folhas de plantas, que haviam sido excluídas na filtragem das regras de consequente de tamanho 1, associados a outros elementos geram influência no item alvo, por exemplo pode se destacar “[temp]=0.[leafspots-marg]=2”, “[leafspot-size]=2.[canker-lesion]=1”.

Embora o impacto da abordagem multialvo já fosse comprovado com a avaliação experimental realizada, que foi apresentada na Seção 5.3, há a necessidade de comprovação da aplicação da MTARN em dados do mundo real. Para isso foi formulada a Prova de Conceito apresentada na Seção seguinte.

5.4 Prova de Conceito

Na formulação da Prova de Conceito para validação da MTARN, todo o processo de mineração de dados foi realizado. Desde coleta de dados e construção do *dataset*, o pré-

processamento do conjunto de dados, bem como a extração de Regras de Associação, filtragem e a construção das MTARNs e *Filtered*-ARNs para obtenção de conhecimento. A validação⁹ do método proposto foi realizada por especialistas do Instituto Dante Pazzanese de Cardiologia, localizado na cidade de São Paulo, estado de São Paulo (Figura 66).

Figura 66 – Metodologia aplicada para a realização da Prova de Conceito

Fonte: Elaborada pelo autor.

5.4.1 Coronary Dataset

Para a construção do *Coronary dataset*, foram coletados os dados referentes às cirurgias de revascularização miocárdica realizados nos pacientes do Instituto Dante Pazzanese de 1999 a 2015. As informações coletadas foram tabuladas e divididas em três categorias, a primeira é referente aos antecedentes pessoais do paciente (ap); a segunda foi elaborada pelo estudo hemodinâmico atual; e a terceira pelo diagnóstico clínico cardíaco. Foi acrescentada também um atributo referente ao sexo do paciente. O arquivo “arff”, disponível on-line¹⁰, foi construído com dados relacionados a aspectos clínicos, pré-operatórios e pós-operatórios.

O estudo foi dirigido para a verificação de quais os parâmetros geram complicações em pacientes que foram submetidos a procedimentos cirúrgicos de revascularização miocárdica e quais desses parâmetros levam os pacientes a óbito. Na Tabela 18 pode-se observar os atributos do dataset.

O atributo de área corpórea do *dataset* é calculado pela Equação 5.1 (DUBOIS; DUBOIS, 1916).

$$areaCat = \sqrt{\frac{pesoCat * alturaCat}{3600}} \quad (5.1)$$

O *dataset* é formado por duas classes, a primeira (Complicacoes) representa a ocorrência ou não de intercorrências, e a segunda (Obito) representa o falecimento ou não do paciente em decorrência da cirurgia realizada, na fase hospitalar.

⁹ A validação dos resultados foi realizada pela avaliação do conhecimento identificado nas Redes por especialistas

¹⁰ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/datasets/Coronary.zip>

Tabela 18 – Variáveis - *Coronary dataset*

Atributo	Descrição
apHAS	Hipertensão Arterial Sistêmica
apDM	Diabete melito
apAf	Antecedentes familiares
apTabagismo	Tabagista; ou interrompeu a mais de 1 ano ou não fuma
apDPOC	Doença pulmonar obstrutiva crônica
apCreatinina	Creatinina > 1,5 (quadro de Insuficiência renal)
apColesterol	Colesterol > 200 mg/dl (Dislipidemia)
apTriglicerides	Triglicérides > 200 mg/dl (Dislipidemia)
apVasculo	Vasculopatia Arterial Periférica
apIAMPrevio	Teve infarto agudo do miocárdio anteriormente
eha_ventriculograma	Função ventricular esquerda
dc_diag2	Diagnóstico clínico atual
lesao	Número de artérias acometidas (Uniarterial, biarterial ou triarterial)
Pontes_Mamarias	Revascularização utilizando enxertos com artéria mamária
QNP	Quadro neurológico prévio
idadeCat	Até 60 anos (jovem adulto); 60-70 Intermediário e > 70 idoso
pesoCat	Massa corpórea
alturaCat	Altura
areaCat	Área de superfície corpórea
Sexo	Masculino ou Feminino
Complicacoes	Apresentou complicações em até 1 ano após a cirurgia
Obito	Paciente veio a óbito

Fonte: Elaborada pelo autor.

O algoritmo *Apriori-TID* foi usado para extrair as Regras de Associação. O conjunto de dados tem 2 classes relacionadas a saúde do paciente após processo de revascularização miocárdica, 20 atributos e 8.672 instâncias. Cada instância representa um paciente atendido no Instituto Dante Pazzanese de Cardiologia.

Foi atribuído um valor mínimo de suporte de 0,0, de modo que todas as situações possíveis de formação de regras fossem inicialmente consideradas. A confiança mínima foi fixada em 0,10 para considerar uma influência de 10% do LHS, de modo que as condições de estudo de complicações e óbitos em pacientes fosse descrito de forma mais ampla.

Seguindo a metodologia descrita no Capítulo 4 desta tese, a aplicação do filtro da medida *Added Value* retirou as regras sem influência entre seus *itemsets*. O valor atribuído para o ganho mínimo (*mingain*) foi 0,01 para que o estudo fosse o mais amplo possível, por se tratar de dados relacionados à saúde de pacientes. Então, foram construídas duas *Filtered-ARNs*, tendo a primeira como item alvo “[Complicacoes]=Sim”, e a segunda como objetivo a classe “[Obito]=Sim”. As Redes completas estão disponíveis on-line¹¹.

As *Filtered-ARNs* foram plotadas com o uso do software *Gephi* (BASTIAN; HEY-

¹¹ <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/coronary/>

MANN; JACOMY, 2009). Os itens do nível 1 (Nível = 1) e nível 2 (Nível = 2) da Rede foram selecionados, por serem os elementos que apresentam influência mais próxima dos itens alvo, para que o conhecimento fosse avaliado e validado por especialistas.

Para a construção das MTARNs, um novo processo de extração de Regras de Associação foi executado para que fossem obtidas regras com consequente de tamanho igual a 2 (Definição 1). Os parâmetros de *minsup*, *minconf* e *mingain* foram os mesmos utilizados na construção das *Filtered-ARNs*, bem como o filtro *Added Value*.

A MTARN foi construída tendo como item alvo a junção das classes em estudo “[Complicacoes]=Sim.[Obito]=Sim” para que a análise dos fatores que levam ao conjunto de complicações e óbito de pacientes com síndrome coronariana aguda fossem descobertos. A MTARN foi plotada com o uso do software *Gephi* (BASTIAN; HEYMANN; JACOMY, 2009) e a versão completa da Rede está disponível on-line¹². Os itens do nível 1 foram selecionados, por serem os que influenciam diretamente nos itens alvo, para que o conhecimento fosse avaliado e validado por especialistas.

5.4.2 Resultados e Discussão

Após a construção da *Filtered-ARN* com alvo “[Obito]=Sim” (Figura 67), foram analisados nós de nível um (Nível = 1). Apenas dois nós foram encontrados “[Complicacoes]=Sim” e “[Pontes_Mamarias]=Sem_mamaria”. Com esta informação foram elaboradas duas hipóteses: i) “todos os pacientes que foram a óbito tiveram algum tipo de complicação”, conhecimento óbvio que valida a eficácia da técnica utilizada; e ii) “os pacientes com maior probabilidade de falecimento são os que não possuem revascularização utilizando enxertos com artéria mamária”. O conhecimento gerado foi confirmado pelos especialistas validando a rede, no entanto alguns itens relacionados com óbitos em pacientes submetidos a cirurgia de revascularização não aparecem na rede conectados ao item alvo como Vasculopatia Arterial Periférica (apVasculo) e sexo feminino.

Como o número de nós de nível um é pequeno, foram observados os nós de nível dois (Nível = 2). Os elementos encontrados no nível 2 da Rede totalizaram 6, e todos foram confirmados pelos especialistas, mas sem a confirmação dos itens de Vasculopatia Arterial Periférica e sexo feminino.

Com a análise dos nós de nível um (Nível = 1) da *Filtered-ARN* com alvo “[Complicacoes]=Sim” (Figura 68), foram obtidos 47 possibilidades de influência direta em complicações pós-operatórias, das quais 5 não possuem nó antecessor, o que denota uma importância maior a estes itens.

Os itens de nível um sem antecedentes são apresentados a seguir. Todos eles, exceto a “[Lesao]=Uniarterial” que representa uma falsa conexão, foram confirmados por especialistas.

¹² <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/coronary/>

Figura 67 – Destaque de *Filtered-ARN* com alvo “[Obito]=Sim” e nós de nível um (Nível = 1) e dois (Nível = 2)

Figura 68 – Destaque de *Filtered-ARN* com alvo “[Complicacoes]=Sim” e nós de nível um (Nível = 1)

- “[apDPOC]=Sim”: representa que o paciente possui doença pulmonar obstrutiva crônica que foi diagnosticada previamente ou no ato da internação
- “[apVasculo]=Sim”: representa um paciente com Vasculopatia Arterial Periférica
- “[Lesao]=Uniarterial”: representa um tipo lesão arterial encontrada no paciente
- “[Obito]=Sim”: paciente veio a óbito
- “[QNP]=Sim”: paciente com isquemia cerebral prévia

Além dos itens sem antecedentes, outros nós de nível um (Nível = 1) foram estudados e os especialistas informaram que muitos itens são inconclusivos. Por exemplo, tanto pacientes do sexo feminino como masculino são conectados a óbito. Atributos como [apColesterol], [apCreatinina], [apDm], [eha_ventriculograma], [idadeCat], [Lesao] e [Pontes_Mamarias] apresentam todas as possibilidades de valores conectadas ao item alvo, tornando inviável qualquer tipo

de afirmação a respeito desses fatores, além da conexão direta de “[Obito]=Nao” ao item de complicações. Sendo assim, a rede não consegue fornecer mais informações concisas a respeito de complicações em pacientes de revascularização coronariana.

A MTARN construída e plotada é apresentada na Figura 69. Analisando os nós de nível um (Nível = 1) foram obtidos 29 itens, dos quais 7 não possuem nós antecessores, o que significa atributos de maior importância. Os itens de nível um sem antecedentes são:

Figura 69 – Destaque de MTARN com alvo “[Complicacoes]=Sim.[Obito]=Sim” e nós de nível um (Nível = 1)

- “[Pontes_Mamarias]=Sem_mamaria”: representa pacientes que não passaram por revascularização utilizando enxertos com artéria mamária
- “[apCreatinina]=Sim.[apVasculo]=Sim”: representa pacientes com quadro de insuficiência renal e Vasculopatia Arterial Periférica
- “[apCreatinina]=Sim.[idadeCat]=[71-95]”: pacientes com insuficiência renal e idade entre 71 e 95 anos
- “[Pontes_Mamarias]=Sem_mamaria.[alturaCat]=[1.31-1.56]”: pacientes que não passaram por revascularização utilizando enxertos com artéria mamária e possuem altura entre 1.31m e 1.56m
- “[dc_diag2]=Angina_Instavel.[Pontes_Mamarias]=Sem_mamaria”: representa pacientes com angina instável e não possuem enxertos com artéria mamária
- “[SEXO]=F.[apCreatinina]=Sim”: pacientes do sexo feminino com insuficiência renal
- “[eha_ventriculograma]=Severo.[idadeCat]=[71-95]”: pacientes com função ventricular esquerda severa e idade entre 71 e 95 anos

Todos os itens foram validados pelos especialistas, os quais enfatizaram a possibilidade de análise em conjunto dos principais fatores conhecidos de complicação e óbito em pacientes que passaram por cirurgia de revascularização coronariana. O conhecimento identificado pela análise da MTARN foi considerado pelos especialistas de alta complexidade e todos de caráter verdadeiro.

Verificando os itens de nível um (Nível = 1) com antecedentes, os especialistas informaram que todos os itens que possuem o atributo “[Pontes_Mamarias]=Sem_mamaria” já podem ser considerados verdadeiros por se tratar de uma condição já comprovada de complicação e óbito. Os especialistas destacaram ainda que os itens com inconsistência na *Filtered-ARN* (Figura 68) foram apresentados na MTARN de forma concisa e em conjunto com outros, como “[Sexo]=F” que indica que a incidência de complicações e óbitos está diretamente conectada a pacientes do sexo feminino.

A observação mais aprofundada da MTARN possibilitou a geração de uma hipótese que será estudada futuramente no Instituto Dante Pazzanese. O fator de suspeita é a não conexão entre tabagismo e complicações pós operatórias ([apTabagismo]=Nao).

A MTARN apresentou um conhecimento mais amplo e validado pelos especialistas como verdadeiro, o que gerou mais confiança à Rede e principalmente, auxiliou no estudo de pacientes submetidos a cirurgia de revascularização miocárdica.

5.5 Considerações Finais

Neste capítulo foi apresentada a proposta da Rede de Regras de Associação Multialvo (MTARN), um método capaz de modelar as Regras de Associação, previamente selecionadas por meio de Medidas Objetivas Assimétricas, de acordo com dois itens objetivo previamente definidos. As regras selecionadas para a construção da Rede são aquelas que possuem uma dependência estatística comprovada pela medida *Added Value*. O item objetivo é utilizado como norteador da exploração e é escolhido de acordo com a problemática que se deseja formular hipóteses. A MTARN cria um Hipergrafo Direcionado, modelando as Regras de Associação que tem o item objetivo no RHS de tamanho fixo igual a 2, recursivamente. Ele visa explicar a correlação entre os itens no *dataset* com o item objetivo formado por dois elementos.

Uma Avaliação Experimental com 3 *datasets* foi desenvolvida para validar a capacidade da MTARN de auxiliar na geração de hipóteses otimizadas e aplicáveis. *Datasets* artificiais foram explorados: *Iris*, *Hayes-roth* e *Soybean Large*. Foi realizada também uma Prova de Conceito com *dataset* real relacionado a pacientes que fizeram cirurgia de revascularização miocárdica de 1999 a 2015. Além da MTARN, foram aplicados os algoritmos ARN e *Filtered-ARN* para comparação. O objetivo foi descrever a ocorrência de regras que influenciam estatisticamente o conjunto desejado, visando encontrar os itens que melhor explicam a ocorrência dos itens objetivos.

A MTARN possibilitou a geração de hipóteses diferentes da abordagem *Filtered-ARN* e com a mesma comprovação matemática de influência pelo uso dos filtros *Added Value* e de ganho mínimo (*mingain*). Parâmetros analisados individualmente que eram excluídos da abordagem ARN pelos filtros, reapareceram associados a outros, comprovando a necessidade de estudos em conjunto dos itens de cada *dataset*.

Na Prova de Conceito com o uso de MTARNs, foi possível otimizar a descoberta de um conhecimento diretamente ligado aos fatores de complicações e óbitos em pacientes submetidos a cirurgia cardíaca para revascularização miocárdica. As constatações feitas usando a técnica de Mineração de Regras de Associação por meio da MTARN foram confirmadas por especialistas do Instituto Dante Pazzanese de Cardiologia, dando uma maior credibilidade ao uso da técnica de mineração em problemas com dois atributos classes. Embora, com o uso da *Filtered-ARN*, tenha sido possível gerar hipóteses que puderam ser confirmadas pelos especialistas, muitas observações foram consideradas inconclusivas ou até mesmo equivocadas para cada classe de modo individual. O uso das Redes de Regras de Associação Multialvo possibilitou uma análise gráfica mais ampla do conteúdo dos dados explorados gerando hipóteses que foram comprovadamente verdadeiras, bem como a elaboração de novas hipóteses que poderão ser estudadas em trabalhos futuros.

Os resultados relatados neste capítulo possibilitaram as respostas para as questões de pesquisa **Q2** (*Como é feita a validação da extração do conhecimento nos processos de Mineração de Regras de Associação?*), **Q5** (*Em problemas multiobjetivo, como podem ser utilizadas estruturas de Rede nos processos de Mineração de Regras de Associação?*) e **Q6** (*Qual o impacto nos resultados da geração de hipóteses em processos de Mineração de Regras de Associação com o uso de Redes?*), atendendo ao objetivo desta pesquisa.

CONCLUSÕES

6.1 Contribuições Científicas

As contribuições deste trabalho estão relacionadas às questões de pesquisa e aos objetivos apresentados na introdução desta tese. Essas questões são sintetizadas nesta seção, apresentando os resultados obtidos e as abordagens propostas e desenvolvidas.

Questão Q1 (Q1.1 a Q1.4): *Como o uso das Redes pode auxiliar nos processos de Mineração de Regras de Associação? Quais os tipos de Rede são utilizados nos processos de Mineração de Regras de Associação? Quais tarefas de mineração utilizam Redes nos processos de Mineração de Regras de Associação? Quais fases da Mineração de Regras de Associação utilizam Redes? Como é feita a geração de hipóteses nos processos de Mineração de Regras de Associação que utilizam Redes?*

Com o objetivo de mapear os trabalhos realizados na comunidade de Mineração de Regras de Associação com o uso de Redes, foi conduzida uma Revisão Sistemática da literatura desse amplo tema. Esse Mapeamento Sistemático da literatura seguiu um protocolo bem definido, realizando a busca de estudos nas principais bibliotecas digitais e fontes de pesquisa da área de computação. A aplicação da expressão de busca nas quatro fontes de pesquisa resultou na identificação de 2.371 estudos, dos quais 97 estudos foram selecionados após verificação dos critérios de inclusão e exclusão estabelecidos. O Mapeamento Sistemático teve uma amplitude temporal maior (8 anos) devido a formulação de dois ciclos de pesquisa. O primeiro em setembro de 2016, com o intuito de identificação das lacunas de pesquisa e elaboração do projeto do doutorado, e o segundo em dezembro de 2018 a fim de consolidação das informações e validação do *gap* que é preenchido por esta pesquisa.

Os resultados obtidos com esse mapeamento confirmaram alguns indicativos prévios dos pesquisadores sobre o tema de pesquisa e apontaram outras características interessantes da

área. Em relação aos tipos de Redes que são utilizados no auxílio da Mineração de Regras de Associação (Q1.1), verificou-se a predominância de Grafos Simples, com destaque também ao uso de estruturas de árvores em todas as etapas da mineração. Ao se considerar as tarefas nas quais as Redes são utilizadas para auxílio na Mineração de Regras de Associação (Q1.2), encontrou-se uma predominância em aplicações mais genéricas de análise, porém tarefas de agrupamento foram bem consolidadas com o uso de Redes. As tarefas de natureza não supervisionada, como o agrupamento, são implementadas com Redes pela vantagem estrutural que este tipo de abordagem propicia. Em relação às fases da mineração que utilizam Redes como auxiliares (Q1.3), verificou-se que todas as fases podem ser implementadas com o uso de algum tipo de Rede, porém a etapa de pós-processamento recebe destaque por se utilizar do resultado gráfico gerado pelas Redes e, consequentemente, a otimização da extração do conhecimento. Desta forma, considerando a geração de hipóteses (Q1.4), destaca-se o uso de técnicas de pós-processamento, como avaliação por consulta, e principalmente, técnicas de visualização, pois as Redes geram resultados gráficos que auxiliam diretamente na extração do conhecimento. Nesta tese, esse trabalho foi apresentado no [Capítulo 2, Seção 2.4](#) - Página 60.

Questão Q2: *Como é feita a validação da extração do conhecimento nos processos de Mineração de Regras de Associação?*

O processo de Mineração de Regras de Associação é dividido em 3 etapas bem definidas: pré-processamento, extração de padrões e pós-processamento. A extração do conhecimento ocorre após a terceira etapa (pós-processamento). Com os estudos primários analisados no Mapeamento Sistemático, percebeu-se que o conhecimento muitas vezes é tratado de modo subjetivo e está atrelado ao tipo de aplicação ou conjunto de domínio no qual os dados estão vinculados.

Em muitas pesquisas utilizam-se critérios para avaliação do conhecimento gerado, como confiabilidade e aplicabilidade, porém totalmente dependentes de interpretação e explanação (NAMAKI *et al.*, 2017). A documentação e visualização dos resultados está conectada ao conhecimento pré-existente do usuário em relação ao *dataset* explorado, ou ao domínio em que os dados estão inseridos.

Foram detectadas algumas pesquisas que utilizam a elaboração de hipóteses de forma semi-automática, por fazerem uso de Redes Bayesianas. A etapa de construção de Redes Bayesianas, em processos de Mineração de Regras de Associação, faz uso de conjuntos de *itemsets* frequentes e cálculos estatísticos para conexão dos nós da Rede (DELGADO *et al.*, 2018). Sendo assim, medidas estatísticas são utilizadas para validação das ligações estabelecidas entre os elementos da Rede a fim de que as relações sejam comprovadas, gerando então as hipóteses.

O trabalho de Pandey *et al.* (2009), que serviu como base para este doutorado, executa a

formulação de hipóteses diretamente pelas arestas formadas na construção das Redes de Regras de Associação (ARNs). O conhecimento é estabelecido a partir de um item objetivo que direciona toda a construção da Rede e, por conseguinte, a geração das hipóteses. A ARN é dividida em níveis, sendo o nó alvo o nível 0 (zero), e os demais níveis a distância mais curta do nó ao item alvo. Embora a geração de hipóteses seja realizada de modo direto por meio de uma simples leitura, não existe comprovação matemática de que as mesmas hipóteses são reais. Portanto, a ARN não produz uma validação do conhecimento gerado, necessitando assim de uma outra análise para sua consolidação.

Neste trabalho, buscou-se tratar esta lacuna pela construção de filtros de Regras de Associação para seleção das regras extraídas por meio de Medidas Assimétricas. Foram selecionadas as medidas *Added Value*, que promove uma comprovação matemática da influência entre o item antecedente e o item consequente da regra, e a medida *Gain*, que promove uma normalização da medida de confiança, comumente utilizada nos processos de Mineração de Regras de Associação. Nesta tese, estes filtros são apresentados no [Capítulo 4, Seção 4.2](#) - Página 105.

Questão Q3: *Quais as relações que existem entre métricas relacionadas a Regras de Associação e medidas relacionadas a estruturas de Rede?*

Estruturas de Rede possibilitam uma melhor visualização da informação e das relações entre os itens de um *dataset*. [Hahsler e Karpienko \(2017\)](#) apresentaram um método de visualização interativa, por meio de uma representação de matriz agrupada, que permite explorar e interpretar intuitivamente cenários altamente complexos. Os grupos de regras geradas são selecionadas com o uso da medida *Lift* e aninhados, formando uma hierarquia que pode ser explorada interativamente até a regra individual.

No trabalho de [Deng et al. \(2014\)](#) é demonstrado o uso de classificadores associativos que consistem em um conjunto de regras ordenado e representado como um modelo de árvore. Além disso, Deng também propôs um algoritmo para transformar uma árvore em um conjunto de regras ordenadas.

Outra abordagem com o uso de Redes foi proposta por [Valle, Ruz e Morrás \(2018\)](#) que utiliza-se estruturas de árvores para estudo das regras geradas para análise de cestas de compras. Devido à estrutura de árvore hierárquica das distâncias, a Rede de Associação possibilita encontrar fortes interdependências entre os produtos da mesma categoria e encontrar produtos que servem como acessos ou pontes para um conjunto de outros produtos com alta correlação entre si.

Os trabalhos apresentados processam e modelam as regras de modo a facilitar o entendimento do usuário. Esses trabalhos podem reduzir o número de regras a serem analisadas pelo usuário, mas não avaliam a definição dos parâmetros das medidas de seleção das regras. Métricas

de seleção (suporte e confiança mínimos) influenciam totalmente a exploração e construção de hipóteses sobre os dados.

Com a proposta de se avaliar a influência de medidas relacionadas a Redes com as medidas de suporte e confiança mínimos em Redes de Regras de Associação, foi efetuado uma Avaliação Experimental a fim de que fossem definidos parâmetros objetivos de avaliação da definição dos valores de suporte e confiança mínimos. Para comprovação da proposta uma Prova de Conceito foi organizada com um conjunto de dados real.

Geralmente os valores de *minsup* e *minconf* em processos de Mineração de Regras de Associação são escolhidos de modo subjetivo, sem um critério matemático de comprovação para que os melhores resultados sejam obtidos, e o conhecimento seja extraído de maneira otimizada. Os resultados demonstraram que com o uso das Medidas de Grau Ponderado Médio, com as confianças como pesos, e Comprimento Médio do Caminho, pode-se avaliar e selecionar o melhor valor para *minconf*. Para a avaliação do suporte mínimo, aplica-se uma análise direta no conjunto de dados a ser estudado, relacionando a classe com a quantidade de instâncias relativa (suporte) que ela apresenta. Os detalhes da avaliação experimental e os resultados obtidos são apresentados nesta tese, no [Capítulo 3](#), Seções 3.3 e 3.4 - Página 87 e 94.

Questão Q4 (Q4.1 e Q4.2): *Quais as Medidas Objetivas utilizadas nos processos de Mineração de Regras de Associação que fazem uso de estruturas de Rede? Como as Medidas Objetivas interferem na geração do conhecimento pela construção de hipóteses? Como as Medidas Objetivas podem ser utilizadas na validação da extração de conhecimento?*

Com o intuito de responder a questão **Q4** e suas sub-questões, foram investigadas e avaliadas as principais Medidas Objetivas Assimétricas relacionadas a Regras de Associação. A escolha de apenas Medidas Assimétricas foi estabelecido pela natureza direcionada das Redes de Regras de Associação.

Os trabalhos apresentados no Mapeamento Sistemático ([Capítulo 2](#)), funcionam processando as regras e modelando-as de maneira a facilitar o entendimento do usuário. No entanto, às vezes o usuário quer analisar o comportamento de um item específico. Esses trabalhos podem reduzir o número de regras a serem analisadas pelo usuário, mas não explicam como um item específico interage com o *dataset* inteiro. Este item de exploração pode ser extremamente útil na construção de hipóteses sobre os dados.

Com o objetivo de possibilitar uma exploração completa e levando em conta a relação entre um conjunto de itens a um elemento alvo, nesta tese é proposta a Rede de Regras de Associação Filtrada (*Filtered-ARN* - do inglês *Filtered-Association Rules Network*), que permite a exploração de um item objetivo com análise de dependência entre os elementos das regras. A abordagem com a *Filtered-ARN*, mescla a técnica de visualização da regras da ARN com a análise matemática das Medidas Objetivas *Added Value* e *Gain*. Com isso, obteve-se uma

extração de conhecimento com validação matemática, possibilitando a geração de hipóteses com maior probabilidade de serem verdadeiras. Foi feita uma Avaliação Experimental com *datasets* encontrados no Repositório UCI, e três Provas de Conceito, a primeira com uma base de dados real relacionada a adubação orgânica, a segunda referente ao desenvolvimento automático de jogos, e a terceira associada a diagnóstico de exames mamográficos. A proposta está descrita nesta tese no [Capítulo 4, Seção 4.2](#) - Página 105.

Questão Q5: *Em problemas multialvo, como podem ser utilizadas estruturas de Rede nos processos de Mineração de Regras de Associação?*

Para análise de vários fatores simultaneamente, pesquisas são elaboradas com o uso de Redes Bayesianas por serem formuladas tendo como base o raciocínio probabilístico ([DELGADO et al., 2018](#); [NEMATI](#); [SANT'ANNA](#); [NOWACZYK, 2016](#)). A principal vantagem de raciocínio probabilístico sobre raciocínio lógico é o fato de que agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará. Sendo assim, pode-se elaborar hipóteses de relações entre os elementos de um *dataset*. Embora sejam uma opção, as Redes Bayesianas são construídas pela formação de um conjunto de itens e de relações de acordo com as probabilidades calculadas, portanto não são comumente utilizadas em problemas multialvo. Problemas multialvo são aqueles em que se deve analisar a influência dos itens de um conjunto de dados direcionado a mais de uma classe ou atributos de uma classe.

Verificou-se por meio do Mapeamento Sistemático ([Capítulo 2](#)) uma lacuna na área de pesquisa de Mineração de Regras de Associação com o uso de Redes. Nenhum trabalho apresentou uma abordagem multialvo utilizando algum tipo de Rede no auxílio da Mineração de Regras de Associação. Embora pode-se considerar que as abordagens de classificação designam os atributos em, no mínimo, duas categorias, nenhum trabalho abordou diretamente mais de uma regra como objetivo dos estudos realizados.

Com a finalidade de responder a questão **Q5**, foi elaborada uma abordagem multialvo para Mineração de Regras de Associação com o uso de Redes. Com o objetivo de possibilitar uma exploração completa e levando em conta a relação entre um conjunto de itens objetivo, nesta pesquisa é proposta a Rede de Regras de Associação Multialvo (MTARN - do inglês *Multi-Target Association Rules Network*), que permite a exploração de dois itens objetivo simultaneamente com análise de dependência entre os elementos das regras. A abordagem MTARN, faz uso dos filtros de medidas *Added Value* e *Gain* para seleção das regras com *itemsets* dependentes. A dependência do consequente da regra pelo antecedente é comprovada matematicamente. As regras geradas possuem consequente de tamanho igual a 2 ($|RHS| = 2$) e antecedentes com tamanho ≤ 2 ($|LHS| \leq 2$), essa configuração garante uma ampliação das análises e extração do conhecimento. Foi feita uma Avaliação Experimental com *datasets* encontrados no Repositório UCI, e uma Prova de Conceito com uma base de dados real relacionada a pacientes que sofreram

cirurgias de revascularização miocárdica. A proposta está descrita nesta tese no [Capítulo 5, Seção 5.2](#) - Página 136.

Questão Q6: *Qual o impacto nos resultados da geração de hipóteses em processos de Mineração de Regras de Associação com o uso de Redes?*

O uso de Redes direcionadas como a *Filtered-ARN* e a *MTARN* promovem uma maior interpretabilidade dos resultados. A validação é possível por meio da aplicação das hipóteses em algoritmos de classificação ou por intermédio de especialistas, gerando um impacto positivo na resolução dos problemas. Todas as abordagens propostas foram comprovadas com a elaboração de Provas de Conceito com *datasets* reais. Em todas as aplicações foi possível destacar a facilidade de visualização das dependências, bem como a elaboração simplificada de hipóteses.

A principal inovação deste trabalho está na proposta de métodos capazes de gerar conhecimento de forma otimizada, matematicamente comprovada e por intermédio da geração facilitada de hipóteses. A extração de conhecimento foi otimizada por meio do resultado gráfico da Rede. Mesmo em Redes com alta densidade, é possível extrair os nós diretamente ligados ao item alvo (Nível = 1) e com isso, elaborar hipóteses com grande probabilidade de serem verdadeiras. Na abordagem *MTARN* ainda é possível verificar itens que individualmente não geram influência direta no problema estudado, mas que associados a outro proporcionam relações importantes para a compreensão dos dados. Todos os métodos propostos foram comparados àqueles utilizados por pesquisadores na área de Mineração de Regras de Associação, gerando sempre um impacto positivo no conhecimento extraído.

6.2 Publicações

Durante o desenvolvimento deste trabalho, as contribuições obtidas foram divulgadas por meio de publicação de artigos em periódicos e publicação e apresentação de artigos em conferências. Essas publicações são listadas a seguir, apresentando a relação de cada uma com este trabalho e indicando aquelas que estão diretamente relacionadas às questões de pesquisas estabelecidas.

Artigos publicados

CALÇADA, D. B.; REZENDE, S. O.; PADUA, R. de. *Asymmetric Objective Measures applied to Filter Association Rules Networks*. In: Latin American Computing Conference - CLEILACLO. São Paulo, 2018. p. 1–10.

Neste artigo são descritas as propostas de construção de filtros de seleção de Regras de Associação com Medidas Assimétricas, abordagem de Redes de Regras de Associação

Filtradas e também da Taxa de Complexidade de *datasets* oriundos da UCI. O trabalho apresentado está relacionado às questões **Q1**, **Q2** e **Q6**. (QUALIS B3)

PADUA, R.; CALÇADA, D. B.; CARVALHO, V. O. de; REZENDE, S. O. *Exploring the data using Extended Association Rule Network*. In: 7th Brazilian Conference on Intelligent Systems BRACIS. São Paulo, 2018.

Neste artigo é apresentada uma abordagem Multialvo relacionada a atributos de uma mesma classe. As regras geradas são do mesmo tamanho das Redes de Regras de Associação. O autor desta tese participou dos experimentos e escrita do artigo em conjunto com Renan de Padua que também é aluno de doutorado no grupo de pesquisa. O trabalho apresentado está relacionado às questões **Q1**, **Q5** e **Q6**. (QUALIS B2)

CALÇADA, D. B.; REZENDE, S. O.; TEODORO, M. S. *Analysis of decomposition parameters of green manure in the Brazilian Northeast with Association Rules Networks*. **Computers and Electronics in Agriculture**, v.159, p.34-41, 2019.

Neste artigo, que é a extensão do resumo publicado na *I International Conference on Agro BigData and Decision Support Systems in Agriculture*, é apresentada a aplicação das Redes de Regras de Associação aplicadas no estudo de parâmetros relacionados a adubação orgânica de forma completa, desde a geração do *dataset* até a validação do conhecimento extraído. O trabalho apresentado está relacionado às questões **Q1** e **Q6**. (QUALIS A2)

Resumo publicado

CALÇADA, D. B.; REZENDE, S. O.; TEODORO, M. S. *Analysis of decomposition parameters of green manure in the Brazilian Northeast with Association Rules Networks*. In: *I International Conference on Agro BigData and Decision Support Systems in Agriculture*. Montevideo, Uruguay, 2017. p. 63–65.

Neste artigo foi apresentada a aplicação das Redes de Regras de Associação aplicadas no estudo de parâmetros relacionados a adubação orgânica. Este trabalho foi elaborado em parceria com um pesquisador da EMBRAPA Meio Norte. Este trabalho resultou de uma exploração inicial da área pesquisa. O trabalho apresentado está relacionado às questões **Q1** e **Q6**.

Relatório Técnico publicado

CALÇADA, D. B.; REZENDE, S. O.; *Medidas objetivas assimétricas aplicadas na construção de Redes de Regras de Associação*. Relatório Técnico No 428, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2018.

Nesse relatório técnico, disponível on-line¹, é apresentado o estudo das Medidas Objetivas Assimétricas e o impacto de cada um na construção de Redes de Regras de Associação. Estes experimentos serviram de base para elaboração da proposta do uso de filtro de medidas para seleção de Regras de Associação. O trabalho apresentado nesse relatório está relacionado às questões **Q1, Q2, Q4 e Q6**.

Artigos submetidos

CALÇADA, D. B.; REZENDE, S. O. *Filtered-ARN: Asymmetric objective measures applied to filter Association Rules Networks*. **CLEI Electronic Journal**. [SUBMETIDO em 2018]

Este artigo é a extensão do trabalho publicado na *Latin American Computing Conference - CLEILACLO*, com o acréscimo de uma avaliação experimental e uma prova de conceito para validação da abordagem de Redes de Regras de Associação Filtradas. O trabalho apresentado está relacionado às questões **Q1, Q2, Q4 e Q6**. (QUALIS B5)

CALÇADA, D. B.; TÓRTORO, L.; TOLEDO, C. F. M.; REZENDE, S. O. *Automatic discovery of knowledge using Filtered-Association Rule Networks (Filtered-ARNs)*. **Applied Soft Computing**. [SUBMETIDO em 2018]

Este artigo apresenta aplicações com *datasets* reais das *Filtered-ARNs* propostas nesta tese. A abordagem é aplicada em dados relacionados a aplicações diferentes e validados por meio do uso de algoritmos genéticos. Esta pesquisa é fruto da parceria com o Professor Cláudio Toledo do ICMC durante a disciplina de Algoritmos Evolutivos. O trabalho apresentado está relacionado às questões **Q1, Q2, Q4 e Q6**. (QUALIS A1)

CALÇADA, D. B.; REZENDE, S. O. *Use of network measures for objective evaluation of minimum support and confidence in association rules mining processes*. **International Journal of Data Mining, Modelling and Management**. [SUBMETIDO em 2019]

Este artigo apresenta um estudo do impacto entre a seleção das medidas de suporte e confiança mínimos de Regras de Associação com as Medidas de Centralidade relacionadas a estruturas de Rede. O resultado é uma avaliação objetiva das medidas de *minsup* e *minconf* que podem auxiliar na seleção de Redes de Regras de Associação com resultados otimizados. O trabalho apresentado está relacionado às questões **Q1 e Q3**. (QUALIS B2)

CALÇADA, D. B.; REZENDE, S. O.; NETO, C. de M. C.; AMATO, V. L. *Study of complication factors in patients undergoing coronary artery bypass surgery: An approach with Multi-Target Association Rules Networks*. **Artificial Intelligence in Medicine**. [SUBMETIDO em 2019]

Este trabalho apresenta uma proposta de extração de conhecimento multialvo com o uso de Redes. São apresentadas as Redes de Regras de Associação Multialvo e sua aplicação em

¹ <<http://repositorio.icmc.usp.br/handle/RIICMC/6882>>

um estudo de atributos de pacientes atendidos no Instituto Dante Pazzanese de Cardiologia, instituição parceira desta pesquisa. O trabalho apresentado está relacionado às questões Q5 e Q6. (QUALIS A2)

6.3 Limitações e Trabalhos Futuros

Nessa seção são discutidas algumas características e limitações das abordagens propostas e das avaliações experimentais realizadas, bem como direções de trabalhos futuros.

Uma característica das abordagens propostas refere-se à construção de Redes com regras selecionadas por filtros de Medidas Objetivas Assimétricas. O uso de Medidas Objetivas é bastante difundido em pesquisas com Mineração de Regras de Associação (AGGARWAL, 2015), portanto pode-se questionar o processo de seleção das medidas *Added Value* e *Gain*, além das Medidas Objetivas Simétricas que não foram estudadas por não caracterizarem o sentido da influência entre os *itemsets* da regra.

Especificamente para o modelo *Filtered-ARN*, limitações em relação à aquisição de conhecimento podem impactar na qualidade da representação. Podem-se ter casos em que a quantidade de nós de nível um (Nível = 1) é muito reduzida, formando assim poucas hipóteses com elementos de alta influência ao nó alvo. Nesses casos, pode-se utilizar os valores das confianças das regras para ponderar o nível dos nós, formando assim um novo processo de cálculo de importância do nó. Portanto, mesmo nós de níveis mais altos, podem ter importância elevada em relação ao item objetivo estudado.

Com base nas limitações e em possibilidades de explorações identificadas durante o desenvolvimento deste trabalho, as seguintes áreas podem ser exploradas futuramente:

Uso de processos automáticos para seleção de suporte e confianças mínimos. Uma outra perspectiva que pode ser investigada é o uso de processos automáticos para seleção de suporte e confianças mínimos, uma vez que foi formulada a proposta de relação entre essas medidas de Regras de Associação e Medidas de Centralidade de Redes. Uma das limitações para esse procedimento é o alto custo computacional para processamento de grandes *datasets* (Big Data Mining), bem como a sua aplicação em processos de geração de dados em tempo real, como em *datasets* web, por exemplo. Para trabalhos futuros, pode-se realizar o estudo do impacto das Medidas de Centralidade da Rede em outras Medidas Objetivas das regras e como as mesmas podem ser utilizadas para seleção de Regras de Associação com maior grau de interesse.

Uso das Redes em processos de Mineração de Textos. Uma linha de trabalho futuro promissora é a investigação do uso das Redes propostas em processos de Mineração de Textos, principalmente em processos de seleção de atributos para classificação automática de textos. Com a abordagem *Filtered-ARN* pode-se formular uma estrutura que pode auxiliar

em classificadores automáticos. As MTARNs podem contribuir na classificação de textos que representem tanto os problemas de organização por tópico quanto os de organização semântica de modo simultâneo. As Redes também podem se utilizadas como oráculos em sistemas de aprendizado dinâmico.

Preparação de novos *datasets* para ampliação dos experimentos. As avaliações experimentais realizadas apresentam uma limitação relacionada ao uso de *datasets* de baixa complexidade. Nesse sentido, um importante trabalho futuro é a seleção, coleta e preparação de novos experimentos que possibilitem uma análise mais ampla das Redes propostas em diversas aplicações. Com uma quantidade maior de *datasets*, poderá ser realizada uma avaliação mais abrangente das abordagens propostas.

Aplicação das Redes propostas em tarefas de Seleção de Atributos. Apesar de neste trabalho o foco ter sido na extração do conhecimento pela geração de hipóteses, os resultados obtidos podem ser aplicados em outras tarefas, como a seleção de atributos. Assim, em relação à expansão dessa pesquisa, tem-se como trabalho futuro a evolução das abordagens propostas considerando-se aspectos de seleção de atributos e uso das Redes em tarefas de Inteligência Artificial Explicável. Entre essas tarefas, pode-se citar o aprendizado não supervisionado, recuperação e extração de informação, além de diferentes tarefas de análise de sentimentos e sistemas de recomendação em mineração de textos. Nesse sentido, pretende-se inicialmente investigar a aplicação das abordagens propostas na seleção de palavras relacionadas ao aspecto de tópico e semântico de textos, e utilizar a explicação gerada pela Redes na melhoria de classificadores automáticos.

REFERÊNCIAS

- ABAY, N. C.; MUTLU, A.; KARAGOZ, P. A path-finding based method for concept discovery in graphs. In: **IISA 2015 - 6th International Conference on Information, Intelligence, Systems and Applications**. [S.l.: s.n.], 2016. p. 1–6. ISBN 9781467393119. Citado nas páginas 76 e 80.
- ACHARYA, A. A.; MAHALI, P.; MOHAPATRA, D. P. Model Based Test Case Prioritization Using Association Rule Mining. In: **Computational Intelligence in Data Mining**. [S.l.: s.n.], 2015. v. 3, p. 429–440. ISBN 978-81-322-2201-9. Citado nas páginas 76 e 79.
- ACR. **BI-RADS: Sistema de laudos e registro de dados de imagem da mama**. 1st. ed. São Paulo: American College of Radiology, 2005. Citado nas páginas 129, 130 e 189.
- ADAMO, J. M. Data Mining for Association Rules and Sequential Patterns. **Springer-Verlag, New York, NY**, 2001. Citado nas páginas 46 e 49.
- ADARICHEVA, K.; NATION, J. B. Discovery of the D-basis in binary tables based on hypergraph dualization. **Theoretical Computer Science**, Elsevier B.V., v. 658, p. 307–315, 2017. ISSN 03043975. Citado nas páginas 73, 76, 77 e 80.
- AGARWAL, R.; AGGARWAL, C.; PRASAD, V. Depth first generation of long patterns. **S.J. Simoff & O.R. Zaiarie (Eds.), Proceedings of ACM SIGKDD 2000**, p. 108–118, 2000. Citado na página 45.
- _____. A tree projection algorithm for generation of frequent itemsets [Special issue]. **Journal of Parallel and Distributed Computing**, v. 61, n. 3, p. 350–371, 2001. Citado na página 45.
- AGGARWAL, C. C. **Data Mining: The Textbook**. 1st. ed. New York, USA: Springer, 2015. 746 p. ISSN 18684394. ISBN 9783319141411. Citado nas páginas 27, 38 e 167.
- AGGARWAL, C. C.; PROCOPIUC, C.; YU, P. S. Finding localized associations in market basket data. **IEEE Transactions on Knowledge and Data Engineering**, v. 14, n. 1, p. 51–62, 2002. ISSN 10414347. Citado nas páginas 38 e 41.
- AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining Association Rules between Sets of Items in Large Databases. **Special Interest Group on Management of Data**, v. 22, n. 2, p. 207–216, 1994. ISSN 01635808. Citado nas páginas 28, 37, 38, 39, 41 e 83.
- AGRAWAL, R.; SHAFER, J. C. Parallel mining of association rules. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 962–969, 1996. ISSN 10414347. Citado nas páginas 37, 43, 61 e 83.
- AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. **Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proceedings of Twentieth International Conference on Very Large Data Bases (VLDB)**, p. 487–499, 1994. Citado nas páginas 37, 39, 41, 48, 61 e 83.

- AL-KHASSAWNEH, Y. A. J.; BAKAR, A. A.; ZAINUDIN, S. Triangle Counting Approach for graph-based Association Rules Mining. In: **Proceedings - 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2012**. [S.l.: s.n.], 2012. p. 661–665. ISBN 9781467300223. Citado nas páginas 72, 76, 77 e 79.
- ALTAF, W.; SHAHBAZ, M.; GUERGACHI, A. Applications of association rule mining in health informatics: a survey. **Artificial Intelligence Review**, Springer Netherlands, v. 47, n. 3, p. 313–340, 2017. ISSN 15737462. Citado na página 27.
- ALZOUBI, W. A.; OMAR, K.; BAKAR, A. A. An efficient mining of transactional data using graph-based technique. In: **Conference on Data Mining and Optimization**. [S.l.: s.n.], 2011. p. 74–81. ISBN 9781612842127. ISSN 21556938. Citado nas páginas 76, 77 e 79.
- ANAND, H. S.; VINODCHANDRA, S. S. Association rule mining using treap. **International Journal of Machine Learning and Cybernetics**, Springer Berlin Heidelberg, v. 9, n. 4, p. 589–597, 2018. ISSN 1868808X. Citado nas páginas 72, 76, 78 e 80.
- BALCÁZAR, J. L. Formal and Computational Properties of the Confidence Boost of Association Rules. **ACM Transactions on Knowledge Discovery from Data**, v. 7, n. 4, p. 01–41, 2013. ISSN 15564681. Citado na página 83.
- BARALIS, E.; CAGLIERO, L.; MAHOTO, N.; FIORI, A. GraphSum: Discovering correlations among multiple terms for graph-based summarization. **Information Sciences**, Elsevier Inc., v. 249, p. 96–109, 2013. ISSN 00200255. Citado nas páginas 74 e 77.
- BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. **Estatística para cursos de engenharia e informática**. 3rd. ed. São Paulo: Atlas, 2010. 416 p. Citado na página 98.
- BASTIAN, M.; HEYMAN, S.; JACOMY, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. In: **Third International AAAI Conference on Weblogs and Social Media**. [S.l.: s.n.], 2009. p. 361–362. ISBN 978-1-57735-421-5. ISSN 14753898. Citado nas páginas 54, 55, 109, 116, 140 e 153.
- BASTOS, E. A.; JUNIOR, A. S. d. A.; RODRIGUES, B. H. N. **Boletim agrometeorológico de 2011 para o município de Parnaíba, Piauí**. Teresina, 2012. 37 p. Citado na página 96.
- BELI, E.; GIABBANELLI, P. J.; PATEL, I.; BALABHADRAPATHRUNI, N. H.; ABDALLAH, A. B.; HAMEED, W.; MAGO, V. K. Combining association rule mining and network analysis for pharmacosurveillance. **Journal of Supercomputing**, Springer US, v. 72, n. 5, p. 2014–2034, 2016. ISSN 15730484. Citado nas páginas 28, 72, 73, 78 e 80.
- BERZAL, F.; BLANCO, I.; SÁNCHEZ, D.; VILA, M.-A. Measuring the accuracy and interest of association rules: A new framework. **Intelligent Data Analysis**, v. 6, p. 221–235, 2002. ISSN 1088467X. Citado na página 48.
- BONACICH, P. Eigenvector-like measures of centrality for asymmetric relations. **Social Networks**, v. 23, p. 191–201, 2001. Citado na página 58.
- BRAGA, R.; OLIVEIRA, P. A.; SOUZA, M.; NETO, P. S.; RABÊLO, R.; BRITTO, R. Ferramentas para desenvolvimento de sistemas baseados em Inteligência computacional - um mapeamento sistemático. **XII Simpósio Brasileiro de Automação Inteligente (SBAI)**, p. 384, 2015. Citado na página 64.

BRAGA, R.; OLIVEIRA, P. A.; SOUZA, M.; ANDRADE, G.; BRITTO, R.; RABÊLO, R.; NETO, P. S. **TheEnd - Systematic Mapping Tool**. 2018. Disponível em: <<http://easii.ufpi.br/theend>>. Citado nas páginas 68 e 70.

BRANDES, U. A faster algorithm for betweenness centrality. **Journal of Mathematical Sociology**, v. 25, n. 2, p. 163–177, 2001. ISSN 0022250X. Citado na página 59.

BRIN, S.; MOTWANI, R.; ULMAN, J. D.; TSUR, S. Dynamic itemset counting and implication rules for market basket data. In: **Proc. of the ACM SIGMOD Intl. Conf. on Management of Data**. [S.l.: s.n.], 1997. p. 255–264. Citado nas páginas 48 e 49.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer Networks and ISDN Systems**, v. 30, p. 107–117, 1998. Citado na página 59.

CAI, R.; LIU, M.; HU, Y.; MELTON, B. L.; MATHENY, M. E.; XU, H.; DUAN, L.; WAITMAN, L. R. Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. **Artificial Intelligence in Medicine**, Elsevier B.V., v. 76, p. 7–15, 2017. ISSN 18732860. Citado nas páginas 28, 73, 75, 78 e 80.

CALÇADA, D. B.; REZENDE, S. O.; TEODORO, M. S. Analysis of decomposition parameters of green manure in the Brazilian Northeast with Association Rules Networks. In: **I International Conference on Agro BigData and Decision Support Systems in Agriculture**. Montevideo, Uruguay: [s.n.], 2017. p. 63–65. Citado na página 116.

CALÇADA, D. B.; TOLEDO, C. F. M.; CHAGAS, P. V. C. das; ALMEIDA, C. C. Classification of Microcalcifications in Mammographic Exams: optimization with the use of hybrid Artificial Neural Network-Genetic Algorithm (GA-ANN). In: **XIV Encontro Nacional de Inteligência Artificial e Computacional - ENIAC**. [S.l.: s.n.], 2017. p. 307–318. Citado nas páginas 129 e 130.

CALDARELLI, G. Scale-free networks: complex webs in nature and technology. **Scitech Book News**, v. 31, n. 3, 2007. Citado na página 53.

CARVALHO, V. O. D. Generalização de regras de associação utilizando conhecimento de domínio e avaliação do conhecimento generalizado. **Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2007**, 2007. Citado na página 50.

CHAWLA, S. Feature Selection, Association Rules Network and Theory Building. **JMLR: Workshop and Conference Proceedings - The Fourth Workshop on Feature Selection in Data Mining**, v. 10, p. 14–21, 2010. Citado nas páginas 61 e 64.

CHAWLA, S. C. S.; DAVIS, J.; PANDEY, G. P. G. On local pruning of association rules using directed hypergraphs. **Proceedings. 20th International Conference on Data Engineering**, n. January, 2004. ISSN 1063-6382. Citado nas páginas 30 e 60.

CHEN, C.; XING, Z. Mining Technology Landscape from Stack Overflow. In: **Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement - ESEM '16**. [S.l.: s.n.], 2016. p. 1–10. ISBN 9781450344272. ISSN 19493789. Citado nas páginas 76, 77 e 79.

CHEN, P.; LU, Y.; ZHENG, V. W.; CHEN, X.; YANG, B. KnowEdu: A System to Construct Knowledge Graph for Education. **IEEE Access**, v. 6, p. 31553–31563, 2018. ISSN 21693536. Citado nas páginas 30, 77, 78 e 79.

- CHERN-TONG, H.; AZIZ, I. A. A Performance Evaluation of Chi-Square Pruning Techniques in Class Association Rules Optimization. In: **Applied Computational Intelligence and Mathematical Methods**. [S.l.: s.n.], 2018. v. 662, p. 196–203. ISBN 978-3-319-67620-3. Citado nas páginas 30, 78 e 79.
- CHOOBDAR, S.; SILVA, F.; RIBEIRO, P. Network node label acquisition and tracking. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [S.l.: s.n.], 2011. v. 7026 LNAI, p. 418–430. ISBN 9783642247682. ISSN 03029743. Citado nas páginas 71, 76 e 79.
- COLEMAN, T. F.; MORÉ, J. J. Estimation of Sparse Jacobian Matrices and Graph Coloring Blems. **SIAM Journal on Numerical Analysis**, v. 20, n. 1, p. 187–209, 1983. ISSN 0036-1429. Citado na página 59.
- DASGUPTA, S.; SAHA, B. A Framework for Discovering Important Patterns Through Parallel Mining of Protein-Protein Interaction Network. In: **Proceedings of the Second International Conference on Computer and Communication Technologies**. [S.l.: s.n.], 2016. v. 381, p. 397–406. ISBN 978-81-322-2525-6. Citado nas páginas 74, 78 e 80.
- DELGADO, E. R.; RODRÍGUEZ-MAZAHUA, L.; PELÁEZ-CAMARENA, S. G.; GUZMÁN, J. A. P.; LÓPEZ-CHAU, A. Association Analysis of Medical Opinions About the Non-realization of Autopsies in a Mexican Hospital. **Scientific Programming**, v. 2018, p. 233–251, 2018. ISSN 10589244. Citado nas páginas 73, 75, 78, 79, 135, 160 e 163.
- DENG, H.; RUNGER, G.; TUV, E.; BANNISTER, W. CBC: An associative classifier with a small number of rules. **Decision Support Systems**, Elsevier B.V., v. 59, n. 1, p. 163–170, 2014. ISSN 01679236. Citado na página 161.
- DING, Y.; YUAN, X.; TANG, K.; XIAO, X.; ZHANG, Y. A fast malware detection algorithm based on objective-oriented association mining. **Computers & Security**, Elsevier Ltd, v. 39, Part B, n. 2008, p. 315–324, 2013. ISSN 0167-4048. Citado nas páginas 75 e 77.
- DOMINGUES, M. A.; REZENDE, S. O. Post-processing of Association Rules using Taxonomies. **Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA 2005)**, p. 192–197, 2005. Citado nas páginas 39, 46 e 83.
- DU, X.; CHUNXIAO, X.; SHAO, F.; SUN, R. A novel marine big data analysis framework based on complex network theory. In: **2015 IEEE 16th International Conference on Communication Technology (ICCT)**. [S.l.: s.n.], 2015. p. 186–191. ISBN 978-1-4673-7004-2. Citado nas páginas 77 e 80.
- DUBOIS, D.; DUBOIS, E. F. A formula to estimate the approximate surface area if height and weight be known. **Arch. Intern. Med.**, v. 17, p. 863–871, 1916. Citado na página 151.
- FAN, W.; HU, C. Big Graph Analyses: From Queries to Dependencies and Association Rules. **Data Science and Engineering**, Springer Berlin Heidelberg, v. 2, n. 1, p. 36–55, 2017. ISSN 2364-1185. Citado nas páginas 73, 74, 76, 77 e 80.
- FAN, W.; WU, Y.; XU, J. Adding Counting Quantifiers to Graph Patterns. In: **Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16**. [S.l.: s.n.], 2016. p. 1215–1230. ISBN 9781450335317. ISSN 07308078. Citado nas páginas 71, 73, 76 e 80.

FELIZARDO, K. R.; NAKAGAWA, E. Y.; MACDONELL, S. G.; MALDONADO, J. C. A visual analysis approach to update systematic reviews. **Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. ACM, 4.**, 2014. Citado na página 64.

FERREIRA, L.; TOLEDO, C. A search-based approach for generating Angry Birds levels. In: **IEEE Conference on Computational Intelligence and Games, CIG.** [S.l.: s.n.], 2014. ISBN 9781479935468. ISSN 23254289. Citado na página 120.

FERREIRA, L. N.; ZHAO, L. Time series clustering via community detection in networks. **Information Sciences**, Elsevier Ltd., v. 326, n. 2016, p. 227–242, 2016. ISSN 00200255. Citado nas páginas 30, 119 e 120.

FERREIRA, N. R.; MEDEIROS, R. B.; SOARES, G. L. G. Potencial Alelopático do Capim-Annoni-2 (*Eragrostis plana* Nees) na germinação de sementes de gramíneas perennes estivais. **Revista Brasileira de Sementes**, v. 30, n. 2, p. 043–050, 2008. ISSN 0101-3122. Nenhuma citação no texto.

FISHER, D. Iterative Optimization and Simplification of Hierarchical Clusterings. **Journal of Artificial Intelligence Research**, v. 4, p. 147–179, 1996. Citado na página 50.

FOURNIER-VIGER, P.; LIN, J. C. W.; VO, B.; CHI, T. T.; ZHANG, J.; LE, H. B. A survey of itemset mining. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 7, n. 4, p. 1–18, 2017. ISSN 19424795. Citado nas páginas 27 e 52.

FREITAS, L. Q. de. **Medidas de centralidade em grafos**. 103 p. Tese (Doutorado) — UFRJ, 2010. Citado na página 58.

FU, X.; SONG, X.; LI, X.; WONG, K. K.; LI, J.; ZHANG, F.; WANG, C.; WANG, Z. Phylogenetic Tree Analysis of the Cold-Hot Nature of Traditional Chinese Marine Medicine for Possible Anticancer Activity. **Evidence-based Complementary and Alternative Medicine**, v. 2017, p. 1–11, 2017. ISSN 17414288. Citado nas páginas 74, 75, 78 e 79.

FUKUDA, T.; MORIMOTO, Y.; MORISHITA, S.; TOKUYAMA, T. Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization. In: **Proceedings of the 1996 ACM SIGMOD international conference on Management of data.** [S.l.: s.n.], 1996. p. 13–23. Citado na página 51.

GAO, Y.; XU, A.; HU, P. J. H.; CHENG, T. H. Incorporating association rule networks in feature category-weighted naive Bayes model to support weaning decision making. **Decision Support Systems**, Elsevier B.V., v. 96, p. 27–38, 2017. ISSN 01679236. Citado nas páginas 74, 77 e 79.

GAO, Y.; ZHANG, Q.; CHEN, L.; WANG, K.; CHEN, N.; LIU, H. Research on application of FP-tree based association rule mining on test sequence in train control system. In: **Chinese Control Conference, CCC.** [S.l.: s.n.], 2017. p. 9993–9997. ISBN 9789881563934. ISSN 21612927. Citado nas páginas 78 e 80.

GENG, L.; HAMILTON, H. J. Interestingness Measures for Data Mining: A Survey. **ACM Computing Surveys**, v. 38, n. 3, p. 1–32, 2006. ISSN 03600300. Citado na página 48.

GONÇALVES, E. C. Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas Objective and Subjective Measures for Association Rules. **INFOCOMP Journal of Computer Science**, v. 4, n. 1, p. 26–35, 2005. Citado na página 47.

- GOODMAN, R. M.; SMYTH, P. Rule Induction Using Information Theory. **Knowledge Discovery in Databases**, n. 9, 1991. Citado na página 50.
- GRAHNE, G.; ZHU, J. Fast algorithms for frequent itemset mining using FP-trees. **IEEE Trans Knowl Data Eng**, v. 17, p. 1347–1362, 2005. Citado nas páginas 43, 61 e 83.
- GUPTA, A. Classification of Complex UCI Datasets Using Machine Learning Algorithms Using Hadoop. **International Journal of Scetific & Techology Research**, v. 4, n. 05, p. 85–94, 2015. Citado nas páginas 87, 193, 194, 195, 196, 197, 198 e 199.
- HAHSLER, M.; KARPIENKO, R. Visualizing association rules in hierarchical groups. **Journal of Business Economics**, Springer Berlin Heidelberg, v. 87, n. 3, p. 317–335, 2017. ISSN 18618928. Citado nas páginas 28, 53 e 161.
- HAN, E.-H.; KARYPIS, G.; KUMAR, V.; MOBASHER, B. Clustering based on association rule hypergraphs. **SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, (SIGMOD-DMKD'97)**, 1997. Citado na página 64.
- HAN, J.; KAMBER, M. **Data Mining Concepts and Techniques**. Second. San Francisco, CA: Morgan Kaufmann, 2006. 772 p. Citado nas páginas 42 e 43.
- HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. Waltham, MA, USA: Morgan Kaufmann, 2012. 745 p. ISSN 1469-994X. ISBN 978-0-12-381479-1. Citado na página 47.
- HAN, J.; PEI, J.; YIN, Y.; MAO, R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. **Data Mining and Knowledge Discovery**, v. 8, n. 1, p. 53–87, 2004. ISSN 13845810. Citado na página 44.
- HENDRICKX, T.; CULE, B.; MEYSMAN, P.; NAULAERTS, S.; LAUKENS, K.; GOETHALS, B. Mining association rules in graphs based on frequent cohesive itemsets. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [S.l.: s.n.], 2015. v. 9078, p. 637–648. ISBN 9783319180311. ISSN 16113349. Citado nas páginas 52, 72, 73, 76 e 80.
- HERNANDES, E.; ZAMBONI, A.; FABBRI, S.; THOMMAZO, A. D. Using gqm and tam to evaluate start-a tool that supports systematic review. **CLEI Electronic Journal**, v. 15, n. 1, p. 3–3, 2012. Citado nas páginas 68 e 70.
- HERNANDEZ-LEAL, P.; RIOS-FLORES, A.; ÁVILA-RIOS, S.; REYES-TERÁN, G.; GONZALEZ, J. A.; FIEDLER-CAMERAS, L.; ORIHUELA-ESPINA, F.; MORALES, E. F.; SUCAR, L. E. Discovering human immunodeficiency virus mutational pathways using temporal Bayesian networks. **Artificial Intelligence in Medicine**, Elsevier B.V., v. 57, n. 3, p. 185–195, 2013. ISSN 09333657. Citado nas páginas 73 e 78.
- HONG, C. F.; CHEN, Y. C.; CHEN, W. C.; TU, K. C.; TSAI, M. H.; CHAN, Y. K.; YU, S. S. Construction of diagnosis system and gene regulatory networks based on microarray analysis. **Journal of Biomedical Informatics**, Elsevier, v. 81, n. October 2017, p. 61–73, 2018. ISSN 15320464. Citado nas páginas 74, 78 e 79.
- HOUTSMA, M.; SWAMI, A. Set-oriented mining of association rules. **Research Report RJ 9567, IBM Almaden Research Center, San Jose, California**, 1993. Citado na página 41.

- INDHUMATHY, M.; NABHAN, A.; ARUMUGAM, S. A weighted association rule mining method for predicting HCV-human protein interactions. **Current Bioinformatics**, v. 12, n. 1, p. 1–12, 2017. ISSN 15748936. Citado nas páginas [73](#), [76](#), [77](#) e [80](#).
- IQBAL, K.; YIN, X. C.; HAO, H. W.; ILYAS, Q. M.; YIN, X. A central tendency-based privacy preserving model for sensitive XML association rules using Bayesian networks. **Intelligent Data Analysis**, v. 18, n. 2, p. 281–303, 2014. ISSN 1088467X. Citado nas páginas [71](#), [73](#) e [78](#).
- KAKKAD, U. K.; ALUVALU, R. Association rule mining using Apriori algorithm: a survey. **IOSR Journal of Computer Engineering (IOSR-JCE)**, v. 16, n. 2, p. 112–118, 2013. Citado na página [42](#).
- KARIMI-MAJD, A.-M.; MAHOOTCHI, M. A new data mining methodology for generating new service ideas. **Information Systems and e-Business Management**, Springer Berlin Heidelberg, v. 13, n. 3, p. 421–443, 2014. ISSN 1617-9846. Citado nas páginas [78](#) e [79](#).
- KE, C.-K.; CHANG, S.-F.; JEN, C.-Y.; LIAO, J. C. Software Recommendation of Application Market by Semantic Network and Multi-Criteria Decision Analysis. **Multimedia and Ubiquitous Engineering, Lecture Notes in Electrical Engineering**, v. 308, p. 317–322, 2014. ISSN 18761119. Citado nas páginas [78](#) e [79](#).
- KIM, C.; LEE, H.; SEOL, H.; LEE, C. Identifying core technologies based on technological cross-impacts: An association rule mining (ARM) and analytic network process (ANP) approach. **Expert Systems with Applications**, Elsevier Ltd, v. 38, n. 10, p. 12559–12564, 2011. ISSN 09574174. Citado na página [104](#).
- KIM, H.; ISHAG, M.; PIAO, M.; KWON, T.; RYU, K. A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries. **Symmetry**, v. 8, n. 6, p. 47, 2016. ISSN 2073-8994. Citado nas páginas [73](#) e [78](#).
- KIM, J.; BANG, C.; HWANG, H.; KIM, D.; PARK, C.; PARK, S. IMA: Identifying disease-related genes using MeSH terms and association rules. **Journal of Biomedical Informatics**, Elsevier, v. 76, n. November, p. 110–123, 2017. ISSN 15320464. Citado nas páginas [73](#), [76](#), [77](#) e [80](#).
- KIM, J.; HAN, M.; LEE, Y.; PARK, Y. Futuristic data-driven scenario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map. **Expert Systems with Applications**, Elsevier Ltd, v. 57, p. 311–323, 2016. ISSN 09574174. Citado nas páginas [73](#), [74](#), [77](#) e [79](#).
- KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. **Technical Report. EBSE Technical Report EBSE-2007-01. Keele University and Durham University Joint Report**, v. 2, p. 1051, 2007. ISSN 00010782. Citado nas páginas [64](#) e [65](#).
- KITCHENHAM, B.; DYBÅ, T.; JØRGENSEN, M. Evidence-based Software Engineering. **Proceedings of the 26th International Conference on Software Engineering, (ICSE '04), IEEE Computer Society, Washington DC, USA**, p. 273 – 281, 2004. Citado na página [65](#).
- KITCHENHAM, B.; PRETORIUS, R.; BUDGEN, D.; BRERETON, O. P.; TURNER, M.; NIAZI, M.; LINKMAN, S. Systematic literature reviews in software engineering: A tertiary study. **Information and Software Technology**, v. 8, n. 52, p. 792–805, 2010. Citado na página [64](#).

- KLEMETTINEN, M.; MANNILA, H.; RONKAINEN, P.; TOIVONEN, H.; VERKAMO, A. Finding interesting rules from large sets of discovered association rules”. In: NABIL, R. (Ed.). **Proceedings of 3rd International Conference on Information and Knowledge Management**. [S.l.: s.n.], 1994. p. 401–407. Citado na página [53](#).
- LE, T.; VO, B. The lattice-based approaches for mining association rules: a review. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 6, n. 4, p. 140–151, 2016. ISSN 19424795. Citado nas páginas [37](#) e [45](#).
- LEE, H.; SHIN, M. Mining pathway associations for disease-related pathway activity analysis based on gene expression and methylation data. **BioData Mining**, BioData Mining, v. 10, n. 1, p. 1–13, 2017. ISSN 17560381. Citado nas páginas [72](#), [76](#), [77](#) e [80](#).
- LI-XIONG, Z.; XIAO-LIN, X. U.; JIA, L. I.; LU, Z.; XUAN-CHEN, P. A. N.; ZHI-YUAN, M. A.; LI-HONG, Z. Malicious URL Prediction based on Community Detection. In: **2015 International Conference on Cyber Security of Smart cities, Industrial Control System and Communications (SSIC) Malicious**. [S.l.: s.n.], 2015. v. 2, p. 1–7. ISBN 9781467379779. Citado nas páginas [73](#), [74](#), [77](#) e [79](#).
- LI, Y.; DU, X.; YANG, B. Application of Association Analysis and Visualization Methods in Car Parts Repair. In: **Joint International Mechanical, Electronic and Information Technology Conference (JIMET 2015)**. [S.l.: s.n.], 2015. p. 1136–1141. ISBN 2010092600000. Citado nas páginas [77](#) e [80](#).
- LI, Z.; LI, L.; YAN, K.; ZHANG, C. Automatic image annotation using fuzzy association rules and decision tree. **Multimedia Systems**, Springer Berlin Heidelberg, v. 23, n. 6, p. 679–690, 2017. ISSN 09424962. Citado nas páginas [30](#), [72](#), [73](#), [75](#), [78](#) e [79](#).
- LIPTON, Z. C. The Mythos of Model Interpretability. In: **ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)**. New York, NY, USA: [s.n.], 2016. p. 1–9. ISBN 1532-821X (Electronic). ISSN 1532-821X; 0003-9993. Citado na página [135](#).
- LIU, B.; HSU, W.; MA, Y. Pruning and summarizing the discovered associations. **Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data mining**, p. 125–134, 1999. Citado na página [48](#).
- LIU, X.; SHEN, H.-W. Association analysis for visual exploration of multivariate scientific data sets. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 1, p. 955–964, 2016. ISSN 1077-2626. Citado nas páginas [73](#), [76](#), [77](#), [78](#) e [80](#).
- LIU, X.; ZHAI, K.; PEDRYCZ, W. An improved association rules mining method. **Expert Systems with Applications**, Elsevier Ltd, v. 39, n. 1, p. 1362–1374, 2012. ISSN 09574174. Citado nas páginas [31](#), [60](#), [71](#), [80](#) e [85](#).
- MAKINO, Y.; KATO, K.; TANIMOTO, S. The Unpredictable Structure of Risk Chains Using Association Rule Mining. In: **2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)**. [S.l.: s.n.], 2017. p. 507–511. ISBN 9781509055043. Citado nas páginas [30](#), [77](#), [78](#) e [79](#).
- MALLAT, S.; HKIRI, E.; MARAOUI, M.; ZRIGUI, M. Semantic Network Formalism for Knowledge Representation:. **International Journal on Semantic Web and Information Systems**, v. 11, n. 4, p. 64–85, 2015. ISSN 1552-6283. Citado na página [78](#).

MARTÍNEZ-BALLESTEROS, M.; GARCÍA-HEREDIA, J. M.; NEPOMUCENO-CHAMORRO, I. A.; RIQUELME-SANTOS, J. C. Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different biological sources. **Information Fusion**, v. 36, p. 114–129, 2017. ISSN 15662535. Citado nas páginas 30, 73, 76, 78 e 79.

MARTÍNEZ-BALLESTEROS, M.; MARTÍNEZ-ÁLVAREZ, F.; TRONCOSO, A.; RIQUELME, J. C. Selecting the best measures to discover quantitative association rules. **Neurocomputing**, Elsevier, v. 126, p. 3–14, 2014. ISSN 09252312. Citado na página 103.

MCCAFFREY, J. **Neural Network How-To: Code an Evolutionary Optimization Solution**. 2014. 2 p. Disponível em: <<https://visualstudiomagazine.com/Articles/2014/03/01/Code-an-Evolutionary-Optimization-Solution.aspx>>. Citado na página 130.

MEI, X.; JIANG, Y. Association rule-based feature selection for credit risk assessment. In: **Proceedings of 2016 IEEE International Conference of Online Analysis and Computing Science, ICOACS 2016**. [S.l.: s.n.], 2016. p. 301–305. ISBN 9781467377546. Citado nas páginas 30, 78 e 79.

MELANDA, E. A. Pos-processamento de Regras de Associação. **Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004**, 2004. Citado na página 29.

MELO, F. d. B.; CAVALCANTE, A. C.; Andrade Junior, A. S. de; BASTOS, E. A. **Levantamento Detalhado dos Solos da Área da Embrapa Meio-Norte/UEP de Parnaíba et al.** Teresina, 2004. Citado na página 96.

MENG, Q.; SHA, J. Tree-based frequent itemsets mining for analysis of life-satisfaction and loneliness of retired athletes. **Cluster Computing**, Springer US, v. 20, n. 4, p. 3327–3335, 2017. ISSN 15737543. Citado nas páginas 78 e 80.

MEO, R.; PSAILA, G.; CERI, S. A new SQL-like operator for mining association rules. In: VIJAYARAMAN, T. (Ed.). **Proceedings of the 22nd International Conference on very Large Data Bases**. [S.l.: s.n.], 1996. p. 122–123. Citado na página 52.

MINELGA, J.; VERIKAS, A.; VAICIUKYNAS, E.; GELZINIS, A.; BACAUSKIENE, M. A Transparent Decision Support Tool in Screening for Laryngeal Disorders Using Voice and Query Data. **Applied Sciences**, v. 7, n. 12, p. 1096, 2017. ISSN 2076-3417. Citado nas páginas 30, 73, 74, 78 e 79.

MOGHADDAM, F. A.; LAGO, P.; GROSSO, P. Energy-Efficient Networking Solutions in Cloud-Based Environments: A Systematic Literature Review. **ACM Comput. Surv.** **47, 4, Article 64 (May 2015)**, 32 pages, v. 47, n. 4, 2015. Citado na página 64.

MONDAL, K. C.; BHATTACHARYA, S.; MONDAL, A. S. A Suffix Tree based Parallel Approach for Association Rule Mining and Biclustering. In: **2016 International Conference on Computer, Electrical & Communication Engineering (ICCECE)**. [S.l.: s.n.], 2016. p. 1–8. Citado nas páginas 78 e 79.

MORE, S.; VIDYA, M.; SUJANA, N.; SOUMYA, H. D. Association Rule Mining for the Identification of Activators from Gene Regulatory Network. In: **Advances in Computing and Communications**. [S.l.: s.n.], 2011. p. 361–370. Citado nas páginas 78 e 80.

- NADI, F.; HORMOZI, S. G.; FOROOZANDEH, A.; SHAHRAKI, M. H. A new method for mining maximal frequent itemsets based on graph theory. In: **Proceedings of the 4th International Conference on Computer and Knowledge Engineering, ICCKE 2014**. [S.l.: s.n.], 2014. p. 183–188. ISBN 9781479954865. Citado nas páginas [71](#), [76](#), [77](#) e [80](#).
- NAMAKI, M. H.; WU, Y.; SONG, Q.; LIN, P.; GE, T. Discovering Graph Temporal Association Rules. In: **Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17**. [S.l.: s.n.], 2017. p. 1697–1706. ISBN 9781450349185. Citado nas páginas [27](#), [52](#), [72](#), [73](#), [76](#), [77](#), [80](#) e [160](#).
- NANDI, J. C. B.; PEREIRA, R. M.; FELIPPE, G.; MADEIRA, K.; BARBOSA, A. C. G.; GARCIA, M. C. d. M. O Algoritmo de Associação Frequent Pattern-Growth na Shell Orion Data Mining Engine. **Congresso Sul Brasileiro de Computação (SULCOMP)**, v. 7, 2014. Citado nas páginas [43](#) e [44](#).
- NEMATI, H. M.; SANT'ANNA, A.; NOWACZYK, S. Bayesian Network representation of meaningful patterns in electricity distribution grids. In: **2016 IEEE International Energy Conference, ENERGYCON 2016**. [S.l.: s.n.], 2016. p. 1–6. ISBN 9781467384636. Citado nas páginas [73](#), [74](#), [75](#), [78](#), [79](#) e [163](#).
- NETTLETON, D. F. Data mining of social networks represented as graphs. **Computer Science Review**, Elsevier Inc., v. 7, n. 1, p. 1–34, 2013. ISSN 15740137. Citado nas páginas [54](#) e [85](#).
- NEVES, J. M. P. M. das. Ambiente de pós-processamento para regras de associação. p. 129, 2002. Citado na página [48](#).
- NEWMAN, M. **Networks: An introduction**. 1st. ed. New York, USA: Oxford University Press, 2010. v. 55. 772 p. ISSN 00218774. ISBN 9780470749838. Citado nas páginas [30](#), [53](#), [54](#), [55](#), [56](#), [57](#), [58](#) e [85](#).
- NGUYEN, D.; NGUYEN, L. T. T.; VO, B.; HONG, T. P. A novel method for constrained class association rule mining. **Information Sciences**, Elsevier Inc., v. 320, p. 107–125, 2015. ISSN 00200255. Citado nas páginas [30](#), [78](#) e [79](#).
- NGUYEN, K. N. T.; CERF, L.; PLANTEVIT, M.; BOULICAUT, J. F. Discovering descriptive rules in relational dynamic graphs. **Intelligent Data Analysis**, v. 17, n. 1, p. 49–69, 2013. ISSN 1088467X. Citado nas páginas [76](#) e [77](#).
- NGUYEN, L.; VO, B.; HONG, T.-p. CARIM : An Efficient Algorithm for Mining Class-Association Rules with Interestingness Measures. **The International Arab Journal of Information Technology**, v. 12, n. 6, p. 627–634, 2015. Citado nas páginas [60](#) e [78](#).
- NGUYEN, L. T. T.; NGUYEN, N. T. Updating mined class association rules for record insertion. **Applied Intelligence**, v. 42, n. 4, p. 707–721, 2015. ISSN 0924669X. Citado nas páginas [30](#), [38](#), [71](#) e [78](#).
- NIRMALA, M.; PALANISAMY, V. An efficient prediction of missing itemset in shopping cart. **Journal of Computer Science**, v. 9, n. 1, p. 55–62, 2013. ISSN 15493636. Citado nas páginas [71](#) e [78](#).
- NOURANI, V.; MOLAJOU, A. Application of a hybrid association rules/decision tree model for drought monitoring. **Global and Planetary Change**, Elsevier, v. 159, p. 37–45, 2017. ISSN 09218181. Nenhuma citação no texto.

ODEH, F.; AL-NAJDAWI, N. ACNB: Associative Classification Mining Based on Naïve Bayesian Method. **International Journal of Information Technology and Web Engineering**, v. 8, n. 1, p. 23–35, 2013. ISSN 1554-1045. Citado nas páginas [73](#) e [78](#).

OLIINYK, A.; SKRUPSKY, S.; SUBBOTIN, S.; KOROBICHUK, I. Parallel method of production rules extraction based on computational intelligence. **Automatic Control and Computer Sciences**, v. 51, n. 4, p. 215–223, 2017. ISSN 1558108X. Citado nas páginas [78](#) e [79](#).

Özgür Cingiz, M.; BIRICIK, G.; DIRI, B. ARNetMiT R Package: Association rules based gene co-expression networks of miRNA targets. **Cellular and Molecular Biology**, v. 63, n. 3, p. 18–25, 2017. ISSN 1165158X. Citado nas páginas [74](#), [78](#) e [80](#).

PADUA, R. D.; CARVALHO, V. O. D.; REZENDE, S. O. Post-processing Association Rules : A Network Based Label Propagation Approach. In: **SOFSEM 2016: Theory and Practice of Computer Science**. [S.l.]: Springer, Berlin, Heidelberg, 2016. v. 9587, p. 580–591. ISBN 9783662491928. Citado nas páginas [28](#), [71](#), [73](#), [76](#), [77](#) e [79](#).

PADUA, R. D.; REZENDE, S. O.; CARVALHO, V. O. D. Post-processing association rules using networks and transductive learning. In: **Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014**. [S.l.: s.n.], 2014. p. 318–323. ISBN 9781479974153. Citado nas páginas [28](#), [60](#), [71](#), [73](#), [76](#), [77](#) e [79](#).

PANDEY, G.; CHAWLA, S.; POON, S.; ARUNASALAM, B.; DAVIS, J. G. Association Rules Network: Definition and Applications. **Statistical Analysis and Data Mining**, v. 1, n. 4, p. 260–179, 2009. ISSN 19321864. Citado nas páginas [30](#), [31](#), [39](#), [60](#), [61](#), [62](#), [63](#), [66](#), [80](#), [84](#), [85](#), [87](#), [104](#), [105](#) e [160](#).

PAPER, C.; SID, S.; UNIVERSIT, A.; LYON, L.; UNIVERSIT, F. B.; LYON, L.; UNIVERSIT, O. B.; LYON, L. P-Triar: Personalization Based on Triadic Association Rules. **Advances in Databases and Information Systems, ADBIS 2014, Springer International Publishing**, v. 9282, n. September 2014, p. 234–247, 2015. Citado nas páginas [30](#) e [78](#).

PEDRO, R. W. D.; NUNES, F. L. S.; MACHADO-LIMA, A. Using Grammars for Pattern Recognition in Images: A Systematic Review. **ACM Comput. Surv.** **46**, **2**, Article 26 (Nov. 2013) Pages., **34**, v. 46, n. 2, 2013. Citado na página [64](#).

PETELIN, B.; KONONENKO, I.; MALAČIČ, V.; KUKAR, M. Multi-level association rules and directed graphs for spatial data analysis. **Expert Systems with Applications**, v. 40, n. 12, p. 4957–4970, 2013. ISSN 09574174. Citado na página [77](#).

PETTICREW, M.; ROBERTS, H. **Systematic Reviews in the Social Sciences: A Practical Guide**. [S.l.]: Blackwell Publishing, 2006. 336 p. ISSN 1016-9040. ISBN 9781405121101. Citado na página [67](#).

PIATETSKY-SHAPIO, G. Discovery, Analysis and Presentation of Strong Rules. **Knowledge Discovery in Databases, AAAI/MIT Press**, p. 229–248, 1991. Citado na página [48](#).

PIRI, S.; DELEN, D.; LIU, T.; PAIVA, W. Development of a new metric to identify rare patterns in association analysis: The case of analyzing diabetes complications. **Expert Systems with Applications**, Elsevier Ltd, v. 94, p. 112–125, 2018. ISSN 09574174. Citado na página [52](#).

RAEDER, T.; CHAWLA, N. V. **Market basket analysis with networks**. [S.l.: s.n.], 2011. v. 1. 97–113 p. ISSN 18695469. ISBN 1327801000. Citado nas páginas [64](#) e [65](#).

- RAJPUT, D. S.; THAKUR, R. S.; THAKUR, G. S. An Integrated Approach and Framework for Document Clustering Using Graph Based Association Rule Mining. In: **Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012)**. [S.l.: s.n.], 2012. v. 236, n. SocProS 2012, p. 1421–1437. ISBN 978-81-322-1601-8. ISSN 21945357. Citado nas páginas 76, 77 e 79.
- RANJAN, P.; VAISH, A.; COULL, N. Social network ties dynamic measurement framework. In: **2014 Students Conference on Engineering and Systems**. [S.l.: s.n.], 2014. v. 211012, p. 1–5. ISBN 978-1-4799-4939-7. Citado nas páginas 76, 77 e 80.
- RASHID, M.; GONDAL, I.; KAMRUZZAMAN, J. Mining Associated Patterns from Wireless Sensor Networks. **IEEE TRANSACTIONS ON COMPUTERS**, v. 64, n. 7, p. 1998–2011, 2015. ISSN 00189340. Citado nas páginas 39, 60 e 79.
- RAWAT, P.; KANT, S.; PANT, B.; CHAUDHARY, A. A Better Approach for Multilevel Association Rule Mining. In: **Proceedings of Fifth International Conference on Soft Computing for Problem Solving**. [S.l.: s.n.], 2016. v. 437, p. 597–604. ISBN 978-981-10-0450-6. Citado nas páginas 28, 76, 77 e 80.
- REZENDE, C. d. P.; CANTARUTTI, R. B.; BRAGA, J. M.; GOMIDE, J. A.; PEREIRA, J. M.; FERREIRA, E.; TARRÉ, R.; MACEDO, R.; ALVES, B. J. R.; URQUIAGA, S.; CADISCH, G.; GILLER, K. E.; BODDEY, R. M. Litter deposition and disappearance in *Brachiaria* pastures in the Atlantic Forest region of the south of Bahia, Brazil. **Nutrients Cycling in Agroecosystems**, v. 54, p. 99–112, 1999. Citado nas páginas 98, 99 e 185.
- REZENDE, S. O.; PUGLIES, J. B.; MELANDA, E. A.; PAULA, M. F. Mineração de Dados. In S. O. Rezende (Ed.), **Sistemas Inteligentes: Fundamentos e Aplicações (1 ed.)**. Editora Manole, n. 12, p. 307–335, 2003. Citado nas páginas 27 e 29.
- RIBEIRO, M. X.; VIEIRA, M. T. P.; TRAINA, A. J. M. Mineração de Regras de Associação Usando Agrupamentos. In: **I Workshop sobre Algoritmos de Mineração de Dados (WAMD'2005) em conjunto com SBBD'2005, 2005. Anais do SBBD/SBES 2005**. Uberlândia, MG: [s.n.], 2005. p. 9–16. Citado na página 46.
- RODRIGUES, M.; GAMA, J.; FERREIRA, C. A. Identifying relationships in transactional data. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [S.l.: s.n.], 2012. v. 7637 LNAI, p. 81–90. ISBN 9783642346538. ISSN 16113349. Citado nas páginas 28, 77, 79 e 80.
- RUSH, E. N.; HARRIS, B.; ALTIPARMAK, N.; TOSUN, A. S. Dynamic Data Layout Optimization for High Performance Parallel I/O. In: **2016 IEEE 23rd International Conference on High Performance Computing (HiPC)**. [S.l.: s.n.], 2016. p. 132–141. ISBN 978-1-5090-5411-4. Citado nas páginas 77 e 79.
- SAHAR, S. **What Is Interesting: Studies on Interestingness in Knowledge Discovery**. 200 p. Tese (Doutorado) — Tel-Aviv University, 2003. Citado na página 49.
- SANTOS, B. R. M.; CARVALHO, D. R. Visualization of Association Rules. **Iberoamerican Journal of Applied Computing**, v. 7, n. 1, p. 1–13, 2017. Citado na página 53.
- SELMANE, S. A.; BOUSSAID, O.; BENTAYEB, F. Towards Collaborative Multidimensional Query Recommendation with Triadic Association Rules. **International Journal of Decision**

Support System Technology, v. 7, n. 3, p. 17–35, 2015. ISSN 1941-6296. Citado nas páginas 30 e 78.

SHAHMOHAMMADI, A.; KHADANGI, E.; BAGHERI, A. Presenting new collaborative link prediction methods for activity recommendation in Facebook. **Neurocomputing**, Elsevier, v. 210, p. 217–226, 2016. ISSN 18728286. Citado nas páginas 74, 76, 77 e 80.

SHENG, G.; HOU, H.; JIANG, X.; CHEN, Y. A Novel Association Rule Mining Method of Big Data for Power Transformers State Parameters Based on Probabilistic Graph Model. **IEEE Transactions on Smart Grid**, v. 9, n. 2, p. 1–1, 2016. ISSN 1949-3053. Citado nas páginas 73, 74, 78 e 79.

SHEU, J.-j. Learning Association Rules and Tracking the Changing Concepts on Webpages : An Effective Pornographic Websites Filtering Approach. **JOURNAL OF ELECTRONIC SCIENCE AND TECHNOLOGY**, v. 16, n. 1, p. 24–36, 2018. Citado nas páginas 72, 74, 78 e 79.

SHEU, J.-J.; CHEN, Y.-K.; CHU, K.-T.; TANG, J.-H.; YANG, W.-P. An intelligent three-phase spam filtering method based on decision tree data mining. **SECURITY AND COMMUNICATION NETWORKS**, v. 9, p. 4013–4026, 2016. ISSN 09739769. Citado nas páginas 72, 73, 75, 78 e 79.

SHORTLIFFE, E. H.; BUCHANAN, B. G. A Model of Inexact Reasoning in Medicine. **Mathematical Biosciences**, v. 23, p. 351–379, 1975. Citado na página 49.

SILVA, T. S. A. **Um Estudo de Medidas de Centralidade e Confiabilidade em Redes**. 60 p. Tese (Doutorado) — Centro Federal de Educação Tecnológica do Rio de Janeiro, 2010. Citado nas páginas 58 e 59.

SIMARD, F.; ST-PIERRE, J.; BISKRI, I. Mining and visualizing robust maximal association rules on highly variable textual data in entrepreneurship. In: **Proceedings of the 8th International Conference on Management of Digital EcoSystems - MEDES**. [S.l.: s.n.], 2016. p. 215–222. ISBN 9781450342674. Citado nas páginas 28, 52, 77 e 79.

SIQUEIRA, G. M.; JÚNIOR, W. M.; PRADO, T. A. D. K. L. de; CARVALHO, M. L. B. de. iFP-growth : um Algoritmo Incremental para Determinar Regras de Associação. **XVII Simpósio Brasileiro de Banco de Dados**, p. 293–304, 2002. Citado na página 45.

SLAIMI, F.; SELLAMI, S.; BOUCELMA, O.; HASSINE, A. B. Leveraging Track Relationships for Web Service Recommendation. In: **Proceedings - 13th IEEE International Conference on E-Business Engineering, ICEBE 2016 - Including 12th Workshop on Service-Oriented Applications, Integration and Collaboration, SOAIC 2016**. [S.l.: s.n.], 2017. p. 220–225. ISBN 9781509061198. Citado nas páginas 74, 76, 77 e 79.

SMALDON, J.; FREITAS, A. A. A New Version of the Ant-Miner Algorithm Discovering Unordered Rule Sets. In: **Proceedings of the 8th annual conference on Genetic and evolutionary computation GECCO 06**. [S.l.: s.n.], 2006. p. 43. ISBN 1595931864. Citado nas páginas 50 e 51.

SOROKINA, M.; MEDIGUE, C.; VALLENET, D. A new network representation of the metabolism to detect chemical transformation modules. **BMC Bioinformatics**, BMC Bioinformatics, v. 16, n. 1, p. 385, 2015. ISSN 1471-2105. Citado na página 78.

- SU, H. N. Global interdependence of collaborative R&D-typology and association of international co-patenting. **Sustainability**, v. 9, n. 4, p. 1–28, 2017. ISSN 20711050. Citado nas páginas 72, 77 e 80.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. Association Analysis: Basic Concepts and Algorithms. In: **Introduction to Data mining**. [S.l.: s.n.], 2005. p. 327–414. ISBN 978-0321321367. Citado na página 48.
- TEODORO, M. S.; CASTRO, K. N. d. C.; MAGALHÃES, J. A. Assessment of legumes with potential use as green manure in the Coastal Tablelands of Piauí States, Brazil. **Caatinga**, 2018. Nenhuma citação no texto.
- THOMAS, S.; BODAGALA, S.; ALSABTI, K.; RANKA, S. An efficient algorithm for the incremental updation of association rules in large databases. **Knowledge Discovery and Data Mining**, p. 263–266, 1997. Citado na página 45.
- THONGKAM, J.; SUKMAK, V. Enhancing the performance of association rule models by filtering instances in colorectal cancer patients. **Engineering and Applied Science Research**, v. 44, n. 2, p. 76–83, 2017. Citado nas páginas 78 e 79.
- THORNTHWAITE, C. W.; MATHER, J. R. The water balance. In: TECHNOLOGY, C. D. I. of (Ed.). **Publication in climatology**. 1. ed. [S.l.]: Drexel Institute of technology, 1955. p. 104. Citado na página 96.
- TOIVONEN, H. Sampling large databases for association rules. **The VLDB Journal**, p. 134–145, 1996. Citado na página 52.
- TSAI, P. J.; LIN, Y. H.; CHEN, J. L.; YANG, S. H.; CHEN, Y. C.; CHEN, H. Y. Identifying Chinese Herbal Medicine Network for Endometriosis: Implications from a Population-Based Database in Taiwan. **Evidence-based Complementary and Alternative Medicine**, v. 2017, p. 1–10, 2017. ISSN 17414288. Citado nas páginas 73, 74, 78 e 79.
- TURGEMAN, L.; MAY, J. H.; SCIULLI, R. Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. **Expert Systems with Applications**, Elsevier Ltd, v. 78, p. 376–385, 2017. ISSN 09574174. Citado nas páginas 73, 76, 78 e 79.
- USMAN, M.; USMAN, M. Multi-level mining and visualization of informative association rules. **Journal of Information Science and Engineering**, v. 32, n. 4, p. 1061–1078, 2016. ISSN 10162364. Citado nas páginas 28, 46, 71, 76, 77 e 79.
- VALLE, M. A.; RUZ, G. A.; MORRÁS, R. Market basket analysis: Complementing association rules with minimum spanning trees. **Expert Systems with Applications**, Elsevier Ltd, v. 97, p. 146–162, 2018. ISSN 09574174. Citado nas páginas 27, 76, 78, 79 e 161.
- VALVERDE-REBAZA, J. C.; De Andrade Lopes, A. **Link prediction in online social networks using group information**. 31–45 p. Tese (Doutorado) — Universidade de São Paulo, 2014. Citado na página 54.
- VASCONCELOS, L. M. R. de; CARVALHO, C. L. de. Aplicação de Regras de Associação para Mineração de Dados na Web. **Instituto de Informática da Universidade Federal de Goiás**, p. 20, 2004. Citado na página 40.

- VEDULA, V.; THATAVARTI, S. Binary Association Rule Mining Using Bayesian Network. In: **Proceedings of International Conference on Information and Network Technology (ICINT 2011)**. [S.l.: s.n.], 2011. v. 4, p. 171–176. ISBN 9199661454. Citado nas páginas 73, 76, 78 e 80.
- VINAYA, M.; SHAH, K. Performance Evaluation of Distributed Association Rule Mining Algorithms. **Procedia - Procedia Computer Science**, Elsevier Masson SAS, v. 79, p. 127–134, 2016. ISSN 1877-0509. Citado na página 37.
- WANG, L.; MENG, J.; XU, P.; PENG, K. Mining temporal association rules with frequent itemsets tree. **Applied Soft Computing Journal**, Elsevier B.V., v. 62, p. 817–829, 2018. ISSN 15684946. Citado nas páginas 71, 73, 74, 76, 78 e 80.
- WANG, L. E.; GUO, Y.; LI, X. Transparent link: A framework of anonymizing msa-dataset based on probabilistic graphical model. In: **Proceedings - 3rd IEEE International Conference on Big Data Computing Service and Applications, BigDataService 2017**. [S.l.: s.n.], 2017. p. 204–209. ISBN 9781509063185. Citado nas páginas 71, 76, 77 e 79.
- WANG, X.; BAI, Y.; ZHANG, R.; WEI, X. The Research of Entertainers Collaboration Relation Based on Association Rules. In: **2017 10th International Symposium on Computational Intelligence and Design (ISCID)**. [S.l.: s.n.], 2017. p. 26–29. ISBN 978-1-5386-3675-6. Citado nas páginas 77 e 79.
- WANG, X.; XU, Y. Mining Graph Pattern Association Rules. In: **Database and Expert Systems Applications**. [S.l.: Springer International Publishing, 2018. v. 11030, p. 223–235. ISBN 978-3-319-98811-5. Citado nas páginas 30, 71, 76, 77, 79 e 84.
- WAZLAWICK, R. **Metodologia de Pesquisa Para Ciência Da Computação**. 2014. 159 p. Citado na página 68.
- WEN, J.; ZHONG, M.; WANG, Z. Activity recognition with weighted frequent patterns mining in smart environments. **Expert Systems with Applications**, Elsevier Ltd, v. 42, n. 17-18, p. 6423–6432, 2015. ISSN 09574174. Citado na página 78.
- WENG, C. H. Identifying association rules of specific later-marketed products. **Applied Soft Computing Journal**, Elsevier B.V., v. 38, p. 518–529, 2016. ISSN 15684946. Citado na página 28.
- WILKE, G.; EMMENEGGER, S.; LUTZ, J.; KAUFMANN, M. Merging Bottom-Up and Top-Down Knowledge Graphs for Intuitive Knowledge Browsing. In: **Flexible Query Answering Systems 2015**. [S.l.: s.n.], 2016. v. 400, p. 445–459. ISBN 978-3-319-26153-9. ISSN 21945357. Citado nas páginas 77 e 80.
- YAN, X.; ZHANG, C.; ZHANG, S. ARMGA: Identifying interesting association rules with genetic algorithms. **Applied Artificial Intelligence**, v. 19, n. 7, p. 677–689, 2005. ISSN 08839514. Citado na página 85.
- _____. Confidence metrics for association rule mining. **Applied Artificial Intelligence**, v. 23, n. 8, p. 713–737, 2009. ISSN 08839514. Citado nas páginas 83 e 85.
- YU, W.; AI, T.; HE, Y.; SHAO, S. Spatial co-location pattern mining of facility points-of-interest improved by network neighborhood and distance decay effects. **International Journal of Geographical Information Science**, Taylor & Francis, v. 31, n. 2, p. 280–296, 2017. ISSN 13623087. Citado nas páginas 73, 76, 77 e 79.

- ZAKI, M. J.; MEIRA, M. J. **Data Mining and Analysis: Fundamental Concepts and Algorithms**. [S.l.]: Cambridge University Press, 2013. 562 p. ISBN 9780521766333. Citado nas páginas 31, 38, 67 e 103.
- ZANIN, M.; PAPO, D.; SOUSA, P. A.; MENASALVAS, E.; NICCHI, A.; KUBIK, E.; BOCCALETTI, S. Combining complex networks and data mining: Why and how. **Physics Reports**, Elsevier B.V., v. 635, p. 1–44, 2016. ISSN 03701573. Nenhuma citação no texto.
- ZHANG, C.; LI, Z.; LI, T.; HAN, Y.; WEI, C.; CHENG, Y.; PENG, Y. P-CSREC: A New Approach for Personalized Cloud Service Recommendation. **IEEE Access**, IEEE, v. 6, p. 35946–35956, 2018. ISSN 2169-3536. Citado nas páginas 74, 76, 77 e 79.
- ZHANG, J.; ZHAO, X.; ZHANG, S.; YIN, S.; QIN, X. Interrelation analysis of celestial spectra data using constrained frequent pattern trees. **Knowledge-Based Systems**, v. 41, n. 2013, p. 77–88, 2013. ISSN 09507051. Citado na página 78.
- ZHAO, W. Q.; ZHANG, Y. F.; ZHANG, S. L. Bayesian Network with Association Rules Applied in the Recognition of Handwritten Digits. **Advanced Materials Research**, v. 187, p. 7–12, 2011. ISSN 1662-8985. Citado nas páginas 71, 73, 78 e 79.
- ZHAO, Y.; ZHANG, C.; CAO, L. **Post-Mining of Association Rules : Techniques for Effective Knowledge Extraction**. [S.l.: s.n.], 2009. 394 p. ISBN 9781605664040. Citado na página 52.
- ZHENJIANG, H.; CHIN, W.; TAKEICHI, M. Calculating a New Data Mining Algorithm for Market Basket Analysis. **Second International Workshop on Practical Aspects of Declarative Languages (PADL' 00)**, Boston, Massachusetts, Janeiro 17-18 2000. LNCS 1753, Springer Verlag., p. 169–184., 2001. Citado na página 41.
- ZHENWEI, Y.; LINGYUN, W.; LIZHU, W. Research on association rules based on Complex Networks.pdf. In: **2nd Workshop on Advanced Research and Technology in Industry Applications (WARTIA 2016)**. [S.l.: s.n.], 2016. p. 1629–1634. Citado nas páginas 28, 77 e 79.

REPOSITÓRIO DE *DATASETS*

Com o objetivo de viabilizar a avaliação das propostas deste trabalho com *datasets* reais, foram gerados conjuntos de dados referentes a diferentes aplicações. Os *datasets* gerados estão descritos e todos disponibilizados on-line¹.

A.1 *Green Manure dataset*

1. Título: *Green Manure dataset*

2. Origem: A coleta dos dados para construção do *dataset* foi realizada em experimentos agrícola na cidade de Parnaíba, estado do Piauí, junto a pesquisadores da EMBRAPA Meio Norte no ano de 2015. Foram coletadas informações sobre 7 espécies de plantas utilizadas como adubo orgânico.

3. Informações Relevantes: Os dados coletados na fase de plantio foram organizados e categorizados de acordo com os valores obtidos. Todos os parâmetros foram distribuídos em 6 classes. Na Tabela 19, pode-se observar todos os atributos do *dataset*. O tempo de meia-vida para cada amostra de planta foi calculado nos intervalos de 30, 60, 90 e 120 dias para formação das classes. A decomposição dos resíduos segue o modelo exponencial simples usado em (REZENDE *et al.*, 1999) (Equação A.1), em que X representa a quantidade de matéria seca após um período de tempo t , X_0 é a quantidade inicial de matéria seca, k é a constante de decomposição e t é o tempo em dias.

$$X = X_0 \cdot e^{-k \cdot t} \quad (\text{A.1})$$

¹ <<http://sites.labc.icmc.usp.br/dariobcalcada/tese/>>

Tabela 19 – Atributos - *Green Manure dataset*

Atributo	Descrição
MFPA	Massa Fresca da Parte Aérea (t/ha)
MSPA	Massa Seca da Parte Aérea (t/ha)
MFR	Massa Fresca da Raiz (t/ha)
MSR	Massa Seca da Raiz (t/ha)
Emerg	Emergência da Planta (DAS)
Flores	Florescimento (DAS)
AltFlores	Altura da Planta no Florescimento (m)
Vagem	Formação das Vagens (DAS)
DiamCol	Diâmetro do Coletor (cm)
NumRamos	Número de Ramos/planta

Fonte: Elaborada pelo autor.

Reorganizando os termos da equação A.1, é possível calcular a constante de decomposição ou o valor k (Equation A.2).

$$k = \frac{\ln \frac{X}{X_0}}{t} \quad (\text{A.2})$$

O tempo de meia-vida ($t_{\frac{1}{2}}$) foi o parâmetro utilizado na avaliação da decomposição de resíduos vegetais, expressando o tempo, em dias, necessário para que metade do material se decomponha.

$$t_{\frac{1}{2}} = \frac{\ln(2)}{k} \quad (\text{A.3})$$

As classes também estão categorizadas de 1 a 6.

4. Número de Instâncias: 28

5. Número de Atributos: 10

6. Informações dos Atributos: 4 classes : Trinta, Sessenta, Noventa, CentoVinte - {1,2,3,4,5,6}

Atributos

MFPA - {1,2,3,4,5,6}

MSPA - {1,2,3,4,5,6}

MFR - {1,2,3,4,5,6}

MSR - {1,2,3,4,5,6}

Emerg - {1,2,3,4,5,6}

Flores - {1,2,3,4,5,6}

AltFlores - {1,2,3,4,5,6}

Vagem - {1,2,3,4,5,6}

DiamCol - {1,2,3,4,5,6}

NumRamos - {1,2,3,4,5,6}

7. Distribuição das classes: Tabela 20

Tabela 20 – Distribuição das Classes por Categoria - *Green Manure dataset*

Classe	1	2	3	4	5	6
Trinta	8	3	6	3	5	3
Sessenta	9	2	1	10	4	2
Noventa	9	4	10	0	2	3
CentoVinte	6	7	8	2	3	2

Fonte: Elaborada pelo autor.

8. **Download:** <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/datasets/Green_Manure.zip>

A.2 Angry Birds datasets

1. **Título:** *Stability dataset* e *Viability dataset*

2. **Origem:** Os dados foram gerados por meio de um algoritmo genético (AG) como técnica para geração de nível de procedimento de fases do jogo *Angry Birds*. Após a obtenção dos indivíduos (fases), suas características foram utilizadas na construção de dois *datasets*. Uma base de dados refere-se à característica de “Estabilidade” (*Stability*), enquanto a outra base lida com a “Viabilidade” ou “Jogabilidade” (*Viability*).

3. **Informações Relevantes:** Os parâmetros selecionados estão diretamente relacionados aos níveis do jogo *Angry Birds* e contam com atributos relacionados a quantidade de porcos disponíveis para o jogador, quantidade de blocos presentes na fase, quantidade de tipos de blocos (variabilidade) na fase e quantidade de pássaros disponíveis no caso da base de dados *Viability*. Os atributos estão descritos nas Tabelas 21 e 22 a seguir.

Tabela 21 – Atributos do *dataset* “Stability”

Atributos	Tipo de dado
Número de porcos	Inteiro
Número de blocos	Inteiro
Número de tipos de blocos	Inteiro
“Stability”	Binário

Fonte: Elaborada pelo autor.

Tabela 22 – Atributos do *dataset* “Viability”

Atributos	Tipo de dado
Número de porcos	Inteiro
Número de blocos	Inteiro
Número de tipos de blocos	Inteiro
Número de pássaros	Inteiro
“IsPlayable”	Binário

Fonte: Elaborada pelo autor.

A.2.1 Stability

4. Número de Instâncias: 5.567

5. Número de Atributos: 3

6. Informações dos Atributos: Uma classe : “Stability” - {0,1}

Atributos

pigs - Quantidade de porcos

block - Quantidade de Blocos

block_type - Quantidade de tipos de blocos

7. Distribuição das classes: Tabela 23

Tabela 23 – Distribuição das Classes por Categoria - *Stability dataset*

Classe	Quantidade de Instâncias
Stability = 0	2.804
Stability = 1	2.763

Fonte: Elaborada pelo autor.

8. Download: <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/datasets/Stability.zip>

A.2.2 Viability

4. Número de Instâncias: 4.553

5. Número de Atributos: 4

6. Informações dos Atributos: Uma classe : “IsPlayable” - {0,1}

Atributos

pigs - Quantidade de porcos

block - Quantidade de Blocos

block_type - Quantidade de tipos de blocos

birds - Quantidade de pássaros

7. Distribuição das classes: Tabela 24

Tabela 24 – Distribuição das Classes por Categoria - *Stability dataset*

Classe	Quantidade de Instâncias
IsPlayable = 0	3.491
IsPlayable = 1	1.062

Fonte: Elaborada pelo autor.

8. **Download:** <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/datasets/Viability.zip>

A.3 Mammography dataset

1. **Título:** *Mammography dataset*

2. **Origem:** A aquisição dos dados foi realizada a partir das imagens adquiridas no repositório digital de imagens desenvolvido pela Universidade do Sul da Flórida (DDSM). Foi realizado uma extração das regiões de interesse (ROIs) e calculado a área de cada nódulo. Os ROIs foram classificados de acordo com os critérios internacionais para avaliação de nódulos mamários (BI-RADS) (ACR, 2005).

3. **Informações Relevantes:** Os parâmetros selecionados para construção do *dataset* final foram: tamanho, margem, forma, densidade e classificação BI-RADS dos nódulos, além da idade do paciente. No caso dos parâmetros de forma e margem, uma escala foi estabelecida para quantificar os dados de acordo com o grau dos mesmos atributos atribuídos ao BI-RADS (Tabela 25). Todos os atributos são numéricos.

Tabela 25 – Quantificação de forma e margem.

Forma	Margem	Valor
Redondo	Circunscrita	1
Ovulado	Turvo	2
Lobular	Microlobulado	3
Irregular	Mal Definido	4
Distorção Arquitetônica	Cravada	5

Fonte: Elaborada pelo autor.

4. Número de Instâncias: 150

5. Número de Atributos: 5

6. Informações dos Atributos: Uma classe : “BI-RADS” - {1,2,3,4,5}

Atributos

tamanho - Perímetro do ROI em centímetros

forma - {1,2,3,4,5}

margem - {1,2,3,4,5}

densidade - Valos da densidade do ROI

idade - Idade do paciente

7. Distribuição das classes: Tabela 26

Tabela 26 – Distribuição das Classes por Categoria - *Mammography dataset*

Classe	Quantidade de Instâncias
BI-RADS = 1	24
BI-RADS = 2	16
BI-RADS = 3	22
BI-RADS = 4	60
BI-RADS = 5	28

Fonte: Elaborada pelo autor.

8. Download: <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/datasets/Breast_Cancer.zip>

A.4 Coronary dataset

1. Título: *Coronary dataset*

2. Origem: A coleta dos dados para construção do *dataset* foi realizada em pacientes que foram submetidos a cirurgia de revascularização miocárdica no Instituto Dante Pazzanese nos anos de 1999 a 2015.

3. Informações Relevantes: As informações coletadas foram tabuladas e divididas em três categorias, a primeira é referente aos antecedentes pessoais do paciente (ap); a segunda foi elaborada pelo estudo hemodinâmico atual; e a terceira pelo diagnóstico clínico cardíaco. Foi acrescentada também um atributo referente ao sexo do paciente. Esta base de dados foi construída com dados relacionados a aspectos clínicos, pré-operatórios e pós-operatórios. O estudo foi dirigido para a verificação de quais os parâmetros geram complicações

em pacientes que foram submetidos ao procedimento cirúrgico e quais desse levam os pacientes a óbito. Na Tabela 27, pode-se observar todos os atributos do *dataset*.

Tabela 27 – Atributos - *Coronary dataset*

Atributo	Descrição
apHAS	Hipertensão Arterial Sistêmica
apDM	Diabete melito
apAf	Antecedentes familiares
apTabagismo	Tabagista; ou interrompeu a mais de 1 ano ou não fuma
apDPOC	Doença pulmonar obstrutiva crônica
apCreatinina	Creatinina > 1,5 (quadro de Insuficiência renal)
apColesterol	Colesterol > 200 mg/dl (Dislipidemia)
apTriglicerides	Triglicérides > 200 mg/dl (Dislipidemia)
apVasculo	Vasculopatia Arterial Periférica
apIAMPrevio	Teve infarto agudo do miocárdio anteriormente
echa_ventriculograma	Função ventricular esquerda
dc_diag2	Diagnóstico clínico atual
lesao	Número de artérias acometidas (Uniarterial, biarterial ou triarterial)
Pontes_Mamarias	Revascularização utilizando enxertos com artéria mamária
QNP	Quadro neurológico prévio
idadeCat	Até 60 anos (jovem adulto); 60-70 Intermediário e > 70 idoso
pesoCat	Massa corpórea
alturaCat	Altura
areaCat	Área de superfície corpórea
Sexo	Masculino ou Feminino
Complicacoes	Apresentou complicações em até 1 ano após a cirurgia
Obito	Paciente veio a óbito

Fonte: Elaborada pelo autor.

O atributo de área corpórea do *dataset* é calculado pela Equação A.4 (Fórmula de Du Bois). A base de dados é formado por duas classes, a primeira (“Complicacoes”) representa o ocorrimto de problemas com o paciente, e a segunda (“Obito”) representa o falecimento ou não do paciente em decorrência da patologia cardíaca.

$$areaCat = \sqrt{\frac{pesoCat * alturaCat}{3600}} \quad (A.4)$$

4. Número de Instâncias: 8.672

5. Número de Atributos: 20

6. Informações dos Atributos: 2 classes : “Complicacoes” e “Obito” - Sim, Não

Valores dos Atributos

SEXO - {M,F}

apHas - {Sim,Nao}

apDm - {Sim,Nao}
 apAf - {Sim,Nao}
 apTabagismo - {Atual,Interropido,Nao}
 apDPOC - {Sim,Nao}
 apCreatinina - {Sim,Nao}
 apColesterol - {Sim,Nao}
 apTriglicerides - {Sim,Nao}
 apVasculo - {Sim,Nao}
 apIAMPrevio - {Sim,Nao}
 eha_ventriculograma - {Discreto,Moderado,Normal,Severo}
 dc_diag2 - {Angina_Estavel,Angina_Instavel,Isquemia_Anatomica,Isquemia_Silenciosa,Indicacao_Anatomica}
 Lesao - {Triarterial,Biarterial,Uniarterial,Tronco,Nenhuma}
 Pontes_Mamarias - {Com_mamaria,Sem_mamaria}
 QNP - {Sim,Nao}
 idadeCat - {[18-54],[54-60],[60-65],[65-71],[71-95]}
 pesoCat - {[37-62.5],[62.5-70],[70-76],[76-85],[85-148]}
 areaCat - {[1.22-1.66],[1.66-1.78],[1.78-1.88],[1.88-2],[2-11.9]}
 alturaCat - {[1.31-1.56],[1.56-1.62],[1.62-1.67],[1.67-1.72],[1.72-1.8]}

7. Distribuição das classes: Tabela 28

Tabela 28 – Distribuição das Classes por Categoria - *Coronary dataset*

Classe	SIM	NAO
Complicacoes	3.607	5.065
Obito	425	8.247

Fonte: Elaborada pelo autor.

8. Download: <http://sites.labc.icmc.usp.br/dariobcalcada/tese/dariobcalcada_files/datasets/Coronary.zip>

TAXA DE COMPLEXIDADE DE *DATASETS* DO REPOSITÓRIO UCI

Muitas pesquisas são elaboradas com o uso de *datasets* oriundos do repositório UCI¹, porém os critérios de escolha de qual base de dado será usada é muito subjetivo e não relata a complexidade que cada *dataset* possui.

Para uma escolha objetiva de *datasets*, foi elaborada, neste trabalho, uma proposta para o cálculo da Taxa de Complexidade utilizando os dados encontrados em Gupta (2015). Na presente pesquisa o autor realizou a comparação de 9 algoritmos de Aprendizado de Máquina com 11 *datasets* oriundos da UCI. Foram calculados os valores de Acurácia para cada algoritmo de Aprendizado de Máquina, bem como um Tempo de Complexidade. Com os resultados gerados neste trabalho, foi proposto um cálculo que permite comparar o grau de complexidade de cada *dataset* de acordo com a definição.

Proposta de Taxa de Complexidade: a taxa de complexidade de um *dataset* é calculado pela razão entre a média aritmética do Tempo de Complexidade (T_c) gerado por algoritmos de Aprendizado de Máquina e a média aritmética da Acurácia (Acc) destes mesmos algoritmos multiplicado pelo fator de correção 1000 (Equação B.1).

$$TxComplexidade = \frac{T_c}{Acc} * 1000 \quad (B.1)$$

Assim, os 11 *datasets* da UCI foram classificadas e os resultados podem ser observados na Tabela 29.

Esta classificação foi a base para a seleção dos conjuntos de dados desta tese. Os valores podem servir como parâmetro de seleção de *datasets* para outras pesquisas que utilizam o repositório UCI como fonte.

¹ <archive.ics.uci.edu/ml>

Tabela 29 – Taxa de Complexidade

<i>Dataset</i>	Taxa de Complexidade
LENSES	1,804
LABOR	2,305
IRIS	3,065
LUNG CANCER	4,765
VOTE	5,355
HAYES-ROTH	5,686
TEACHING ASSISTANT	6,030
STATLOG	8,256
GLASS IDENTIFICATION	8,897
SOYBEAN LARGE	123,469
ABALONE	1192,960

Fonte: Elaborada pelo autor.

Os dados retirados da pesquisa de Gupta (2015) são apresentados nas Tabelas a seguir.

Na Tabela 30 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “IRIS”². Este conjunto de dados possui 150 instâncias, 4 atributos que apresentam comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala, além de 3 classes de flores, sendo íris setosa, íris virgínica e íris versicolor.

Tabela 30 – Resultados de Acurácia e Tempo Computacional *dataset* IRIS

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	96,00%	503
Decision Table	92,70%	676
Logistic	96,00%	1196
Multilayer Perceptron	96,00%	2416
NaiveBayes	96,00%	450
Random Forest	96,00%	1790
VFI	96,67%	406
ZeroR	33,33%	382
Genetic Programming	96,00%	16657

Fonte: Adaptada de Gupta (2015).

Na Tabela 31 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “LABOR”³. Os dados incluem todos os argumentos coletivos alcançados no setor de negócios e serviços pessoais para as regiões com pelo menos 500 membros. Tem 57 instâncias, 16 atributos e 2 classes que classificam as instâncias como boas e ruins.

Na Tabela 32 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado

² <<http://archive.ics.uci.edu/ml/datasets/Iris>>

³ <<http://archive.ics.uci.edu/ml/datasets/Labor+Relations>>

Tabela 31 – Resultados de Acurácia e Tempo Computacional *dataset* LABOR

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	77,19%	742
Decision Table	75,44%	983
Logistic	85,96%	1171
Multilayer Perceptron	85,96%	3570
NaiveBayes	91,28%	577
Random Forest	89,47%	1933
VFI	80,70%	533
ZeroR	64,91%	430
Genetic Programming	66,67%	6599

Fonte: Adaptada de [Gupta \(2015\)](#).

de Máquina no *dataset* “ABALONE”⁴. O conjunto de dados apresenta a previsão da idade de conchas abalone de medições físicas. Tem 4.177 instâncias, 8 atributos com informações de sexo, comprimento, diâmetro, altura, altura total, peso sem concha, peso da víscera, peso da concha e 29 classes numeradas de 1-29.

Tabela 32 – Resultados de Acurácia e Tempo Computacional *dataset* ABALONE

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	19,20%	7206
Decision Table	26,52%	5076
Logistic	30,00%	351456
Multilayer Perceptron	26,86%	551456
NaiveBayes	24,77%	2181
Random Forest	22,60%	350904
VFI	13,81%	1578
ZeroR	16,50%	713
Genetic Programming	16,52%	1076914

Fonte: Adaptada de [Gupta \(2015\)](#).

Na Tabela 33 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “LENSES”⁵. É o conjunto de dados que apresenta a montagem de lentes de contato. Tem 24 instâncias, 4 atributos: idade, prescrição de óculos, astigmatismo, taxa de produção de lágrimas e 3 classes: *soft*, *hard* e *no*.

Na Tabela 34 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “SOYBEAN”⁶. O conjunto de dados apresenta a descrição da doença de plantações de soja. Tem 683 instâncias, 35 atributos e 19 classes.

⁴ <<http://archive.ics.uci.edu/ml/datasets/Abalone>>

⁵ <<http://archive.ics.uci.edu/ml/datasets/Lenses>>

⁶ <<http://archive.ics.uci.edu/ml/datasets/Soybean+%28Large%29>>

Tabela 33 – Resultados de Acurácia e Tempo Computacional *dataset* LENSES

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	70,83%	438
Decision Table	70,83%	625
Logistic	66,67%	716
Multilayer Perceptron	66,67%	1262
NaiveBayes	70,83%	502
Random Forest	66,67%	1145
VFI	70,83%	445
ZeroR	62,50%	338
Genetic Programming	75,00%	5726

Fonte: Adaptada de Gupta (2015).

Tabela 34 – Resultados de Acurácia e Tempo Computacional *dataset* SOYBEAN

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	90,34%	1553
Decision Table	81,41%	4370
Logistic	92,98%	78999
Multilayer Perceptron	93,70%	100257
NaiveBayes	92,09%	1166
Random Forest	92,97%	99876
VFI	86,09%	1188
ZeroR	13,47%	766
Genetic Programming	81,91%	606928

Fonte: Adaptada de Gupta (2015).

Na Tabela 35 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “LUNG CANCER”⁷. Os dados descrevem três tipos de câncer de pulmão. Tem 32 instâncias, 56 atributos e 3 classes numeradas de 1 a 3.

Na Tabela 36 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “HAYES-ROTH”⁸. Tem 160 instâncias, 5 atributos divididos em: Nome, Passatempo, Idade, Nível de Educação, Estado Civil e 3 classes numeradas de 1 a 3.

Na Tabela 37 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “GLASS IDENTIFICATION”⁹. O conjunto de dados possui 1 atributo de índice e 9 atributos (índice de refração, teor de sódio, potássio, magnésio, alumínio, cálcio, silício, bário e ferro) e consiste em 214 instâncias de 7 classes diferentes distribuídas em: janelas de construção com flocos de vidro processado, janelas de vidro sem flocos de vidro processado, janelas para veículos com flocos de vidro processado, janelas para veículos sem flocos de vidro

⁷ <<http://archive.ics.uci.edu/ml/datasets/Lung+Cancer>>

⁸ <<http://archive.ics.uci.edu/ml/datasets/Hayes-Roth>>

⁹ <<http://archive.ics.uci.edu/ml/datasets/Glass+Identification>>

Tabela 35 – Resultados de Acurácia e Tempo Computacional *dataset* LUNG CANCER

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	87,50%	450
Decision Table	84,38%	359
Logistic	93,75%	812
Multilayer Perceptron	90,63%	4137
NaiveBayes	96,88%	250
Random Forest	87,50%	2129
VFI	84,38%	581
ZeroR	81,25%	250
Genetic Programming	85,00%	28739

Fonte: Adaptada de [Gupta \(2015\)](#).

Tabela 36 – Resultados de Acurácia e Tempo Computacional *dataset* HAYES-ROTH

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	80,30%	810
Decision Table	56,82%	828
Logistic	53,79%	855
Multilayer Perceptron	71,21%	2948
NaiveBayes	75,00%	615
Random Forest	80,30%	2233
VFI	61,36%	584
ZeroR	37,88%	497
Genetic Programming	81,82%	24653

Fonte: Adaptada de [Gupta \(2015\)](#).

processado, recipientes, utensílios de mesa e faróis.

Tabela 37 – Resultados de Acurácia e Tempo Computacional *dataset* GLASS IDENTIFICATION

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	68,69%	859
Decision Table	68,22%	1265
Logistic	63,08%	2953
Multilayer Perceptron	65,89%	9048
NaiveBayes	47,66%	547
Random Forest	80,37%	4484
VFI	54,67%	500
ZeroR	35,51%	406
Genetic Programming	66,82%	28955

Fonte: Adaptada de [Gupta \(2015\)](#).

Na Tabela 38 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado

de Máquina no *dataset* “TEACHING ASSISTANT”¹⁰. Os dados consistem em avaliações do desempenho do ensino ao longo de três semestres regulares e dois semestres de verão de 151 atribuições de professores assistentes (TA - do inglês *Teaching Assistant*) de ensino no Departamento de Estatística da Universidade de Wisconsin-Madison. As pontuações foram divididas em 3 categorias aproximadamente iguais (“baixa”, “média” e “alta”) para formar a variável de classe. Ele contém 151 instâncias, 5 atributos que apresentam os dados: TA é um falante nativo de inglês, curso de instrutor, curso, verão ou semestre regular, tamanho de classe e 3 classes numeradas de 1 a 3 onde 1 indica baixo, 2 indica médio e 3 indica alto.

Tabela 38 – Resultados de Acurácia e Tempo Computacional *dataset* TEACHING ASSISTANT

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	41,06%	442
Decision Table	35,01%	775
Logistic	41,06%	835
Multilayer Perceptron	38,41%	2049
NaiveBayes	41,06%	442
Random Forest	41,06%	882
VFI	41,06%	518
ZeroR	34,44%	448
Genetic Programming	37,75%	2351

Fonte: Adaptada de [Gupta \(2015\)](#).

Na Tabela 39 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “STATLOG (*Australian Credit Approval*)”¹¹. Este conjunto de dados refere-se a aplicativos de cartão de crédito. Todos os nomes e valores de atributos foram alterados para símbolos sem sentido para proteger a confidencialidade dos dados. Este *dataset* é interessante porque existe uma boa combinação de atributos contínuos, nominais com pequenos números de valores e nominais com números maiores de valores. Há também alguns valores ausentes. Contém 690 instâncias, 14 atributos e 2 classes.

Na Tabela 40 é apresentado os resultados obtidos pelo uso dos algoritmos de Aprendizado de Máquina no *dataset* “VOTE”¹². Este conjunto de dados inclui votos para cada um dos congressistas da Câmara dos Deputados dos EUA sobre os 16 votos chave identificados pelo CQA (*Congressional Quarterly Almanac*). O CQA lista nove tipos diferentes de votos: votados, emparelhados e anunciados (estes três simplificados para sim), votados contra, emparelhados e anunciados contra (estes três simplificados para não), votados presentes, votados presentes para evitar conflito de interesses, e não votou ou de outra forma fez uma posição conhecida (estes três simplificados a uma disposição desconhecida). Tem 435 instâncias, 16 atributos e 2 classes referentes aos partidos americanos (democrata e republicano).

¹⁰ <<http://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation>>

¹¹ <<http://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29>>

¹² <<http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records>>

Tabela 39 – Resultados de Acurácia e Tempo Computacional *dataset* STATLOG

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	85,80%	1391
Decision Table	85,07%	1892
Logistic	86,67%	1687
Multilayer Perceptron	85,80%	8158
NaiveBayes	75,36%	833
Random Forest	86,09%	5108
VFI	85,36%	398
ZeroR	55,51%	724
Genetic Programming	87,10%	40305

Fonte: Adaptada de [Gupta \(2015\)](#).

Tabela 40 – Resultados de Acurácia e Tempo Computacional *dataset* VOTE

Algoritmo	Acurácia	Tempo de Complexidade(ms)
J48	96,09%	725
Decision Table	95,63%	1099
Logistic	96,09%	1439
Multilayer Perceptron	95,86%	6106
NaiveBayes	96,02%	590
Random Forest	95,17%	2582
VFI	95,86%	600
ZeroR	61,38%	416
Genetic Programming	61,38%	28937

Fonte: Adaptada de [Gupta \(2015\)](#).

Muito outros *datasets* são encontrados no repositório UCI. Pode-se então aplicar a mesma metodologia realizada por [Gupta \(2015\)](#) para obtenção das variáveis de Tempo de Complexidade (T_c) e Acurácia (Acc) gerados pelos algoritmos de Aprendizado de Máquina. Com os valores obtidos, aplica-se a Equação [B.1](#) e se obtém a Taxa de Complexidade do conjunto de dados. Desta forma, pode-se utilizar a Taxa de Complexidade como um critério objetivo e conciso para seleção de *datasets*.