Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 27, Num. 3 (2020) 40-49

RESEARCH ARTICLE

Automatic identification of knowledge related to dengue cases in the state of Piauí in public databases using Filtered-Association Rules Networks

Identificação automática de conhecimento relacionado a casos de dengue no estado do Piauí em banco de dados públicos utilizando Redes de Regras de Associação Filtradas

Joan D. S. Silva[1], Jâina Carolina Meneses Calçada[2], Solange Oliveira Rezende[3], Dario Brito Calçada[1,3]*

Abstract: Dengue is an endemic disease in Brazil since the 1980s and since 1996 in Piauí. The number of cases increases each year, with the incidence of more severe symptoms. This research aimed to evaluate the use of an automatic knowledge identification technique in factors related to the number of dengue occurrences. We built a dataset formed by data available in the Information System for Notifiable Diseases (SINAN) and meteorological data of the municipalities of the coastal plain of Piauí. The technique used was that of Filtered Association Rules Networks, which allows visual analysis of knowledge through the use of network structures and rules filtering. As a main result, we confirmed the understanding that the most significant number of cases occurs in May, as it is the moment when the rainfall indexes are decreasing, besides that socio-cultural and race factors do not interfere in the identification of the population of higher risk. This research presents the innovation of the use of a computational technique of automatic knowledge discovery that can assist in the elaboration of prevention actions by epidemiological surveillance.
Keywords: Association Rules — Dengue — Epidemiological Surveillance — Knowledge Discovery — Networks

Resumo: A dengue é uma doença endêmica no Brasil desde a década de 1980 e desde 1996 no Piauí. O número de casos aumenta a cada ano com a incidência de sintomas mais graves. Esta pesquisa teve como objetivo avaliar o uso de uma técnica automática de identificação de conhecimento em fatores relacionados ao número de ocorrências de dengue. Foi construído um *dataset* formado por dados disponíveis no Sistema de Informação de Agravos de Notificação (SINAN) e dados meteorológicos dos municípios da planície litorânea piauiense. A técnica utilizada foi a de Redes de Regras de Associação Filtradas que possibilita uma análise visual do conhecimento por meio do uso de estruturas de rede e filtragem das regras. Como resultado principal, foi confirmado o conhecimento que o maior número de casos ocorre no mês de maio, pois é o instante quando os índices pluviométricos estão se reduzindo, além de que fatores sócio-culturais e de raça não interferem para a identificação de população de maior risco. Esta pesquisa apresenta a inovação do uso de uma técnica computacional de descoberta automática de conhecimento que pode auxiliar na elaboração de ações de prevenção pela vigilância epidemiológica.
Palavras-Chave: Dengue — Descoberta do Conhecimento — Redes — Regras de Associação — Vigilância Epidemiológica

[1] *Universidade Estadual do Piauí (UESPI) Campus Alexandre Alves de Oliveira – Parnaíba, PI – Brazil*
[2] *Universidade Estadual Vale do Acaraú (UVA) Centro de Ciências da Saúde – Sobral, CE – Brazil*
[3] *Instituto de Ciências Matemáticas e de Computação (ICMC) Universidade de São Paulo (USP) – São Carlos, SP – Brazil*
***Corresponding author**: dariobcalcada@gmail.com
DOI: https://doi.org/10.22456/2175-2745.99849 • **Received:** 24/01/2020 • **Accepted:** 17/05/2020

1. Introdução

A dengue é uma doença viral transmitida pelo mosquito *Aedes aegypti*, sendo ela uma das mais endêmicas do mundo, cuja a incidência entre os ano de 2010 e 2016 saltou de 0.5 milhões para 3.34 milhões de casos reportados a organização mundial da saúde [1]. Estima-se que 3.9 bilhões de pessoas estejam em risco de infecção em mais de 128 países e que ocorrem

anualmente entre 284–528 milhões de infecções em todo o globo[2].

No Brasil, há registro da doença desde o final do século XIX, em Curitiba no estado do Paraná. Na década de 80 houve epidemias nos Estados de Roraima, Minas Gerais, São Paulo, Bahia, Pernambuco, Ceará, Alagoas e Rio de Janeiro. No estado do Rio de Janeiro ocorreu uma epidemia em 1986, na qual a dengue adquiriu importância epidemiológica. A doença logo atingiu a Região Nordeste tornando-se endêmica no país[3].

No Piauí, a presença do *Aedes aegypti* foi confirmada em 1986. Já em 1994, levantamentos entomológicos realizados pela Fundação Nacional de Saúde (Funasa) confirmaram a presença do mosquito no Município de Teresina-PI. Nesse mesmo ano, foram notificados os primeiros casos autóctones de dengue, confirmando-se a primeira epidemia em 1996. No ano de 2012 foi detectada a maior epidemia, com registro de 12236 casos e seis óbitos. [4].

A ocorrência de dengue é influenciada por uma mistura de fatores como a rápida urbanização e crescente densidade populacional, variáveis meteorológicas e políticas de controle do vetor pela vigilância sanitária. Como não existe uma disponibilidade de vacina ou droga contra dengue, o controle do vetor e a eliminação do mosquito adulto e das larvas é realizado por meio da redução de focos de crescimento, sendo estas as únicas ações efetivas no controle da transmissão da enfermidade [5].

Com o aumento constante de pessoas infectadas, os serviços públicos de saúde possuem a importante missão do controle e prevenção de doenças, como a dengue. Para auxiliar nesse complexo procedimento, a informação sobre a doença é um ativo extremamente significativo, tanto no processo de tomada de decisão, quanto na criação de políticas na área da saúde e aumento da qualidade de vida. Um sistema de aviso prévio é uma peça essencial para auxiliar tais ações. Nos últimos anos, variáveis meteorológicas como temperatura e nível de chuva, tem sido estudadas pelo seu potencial como ferramentas de aviso prévio no combate de doenças infecciosas sensíveis ao clima, como malária, dengue e febre do nilo ocidental [6, 7, 8].

Uma importante fonte do conhecimento usada no processo de vigilância epidemiológica no Brasil é o SINAN (Sistema de Informação de Agravos de Notificação). O SINAN é alimentado principalmente pela notificação e investigação de casos de doenças e agravos que constam da lista nacional de patologias de notificação compulsória. Mesmo sendo um fator essencial para ações de prevenção, é facultado aos estados e municípios a inclusão de outros problemas de saúde importantes. Este sistema fornece as informações usadas na formulação de políticas em todas as esferas administrativas (municipal, estadual e federal) para o controle e prevenção de doenças notificáveis [9].

Com o propósito de melhorar a descoberta de padrões relevantes e, possivelmente inovadores, em grandes bases de dados como o SINAN, o uso de técnicas de Descoberta de Conhecimento em Base de Dados (KDD, do inglês *Knowledge-discovery in Databases*) é uma alternativa. Em relação a serviços de saúde, o KDD tem sido usado para a extração automática de conhecimento, que pode auxiliar na prevenção de doenças, diagnósticos mais precisos, tratamentos, detecção de anomalias, prognóstico, controle de infecções hospitalares e pesquisa epidemiológica. [10, 11, 12].

Uma das técnicas utilizadas no KDD é a Mineração de Regras de Associação (ARM - do inglês *Association Rules Mining*), que visa a identificação de padrões em *datasets*. O processo de Mineração de Regras de Associação pode ser visto como um conjunto de ações genéricas que devem ser realizadas de acordo com os dados disponíveis [13]. Um exemplo de uso da Mineração de Regras de Associação para a descoberta de conhecimento é a sua utilização na análise de dados referente a compras em um supermercado, que pode gerar uma regra como: $\{\text{feijão, couve}\} \Rightarrow \{\text{linguiça}\}$. Essa regra é utilizada para gerar a hipótese de que "clientes que compram feijão e couve tendem também a comprar linguiça". O exemplo ilustra uma das características mais atrativas das Regras de Associação, na qual cada regra é expressa de uma forma muito fácil de ser compreendida quando formulada por *itemsets* de tamanho reduzido [14, 15].

A quantidade de Regras de Associação extraídas está diretamente ligada ao número de itens que formam a base de dados. Em um *dataset*, com uma quantidade elevada de elementos, o conjunto de Regras de Associação geradas torna-se cada vez maior, inviabilizando a análise de todas as regras. Por exemplo, em um conjunto de apenas 100 elementos gera-se 9.900 regras com *itemsets* unitários e 98.000.100 regras se os *itemsets* possuírem dois elementos, um crescimento exponencial.

A utilização de redes para a análise das regras geradas é de grande auxílio no processo de identificação dos padrões no conjunto de dados. As Redes de Regras de Associação (ARN - do inglês *Association Rules Network*) tem como ideia central sintetizar, podar e integrar, no contexto dos objetivos específicos da pesquisa, as Regras de Associação descobertas pelo algorítimo de mineração.

O objetivo principal neste trabalho foi validar o uso de uma técnica de extração automática de conhecimento a fim de identificar o comportamento da dengue nos municípios da planície litorânea piauiense formado pelas cidades de Bom Princípio do Piauí, Buriti dos Lopes, Cajueiro da Praia, Caraúbas do Piauí, Caxingó, Cocal, Cocal dos Alves, Ilha Grande, Luís Correia, Murici dos Portelas, Parnaíba, Piracuruca, São João da Fronteira e São José do Divino[1] , utilizando dados disponíveis no SINAN dos anos de 2007 a 2014 e dados meteorológicos fornecidos pela EMPRAPA-Meio Norte. Foi utilizado o método de KDD com a técnica de Redes de Regras de Associação Filtradas (*Filtered*-ARN), tendo como objetivo identificar conhecimento que pode auxiliar a vigilância epidemiológica no processo de tomada de decisão e formulação de políticas preventivas.

[1] ⟨ftp://geoftp.ibge.gov.br/organizacao_do_territorio/estrutura_territorial/divisao_territorial/2016/DTB_2016_v2.zip⟩

Este artigo está dividido da seguinte maneira. Trabalhos relacionados são descritos na Seção 2. Na Seção 3, é apresentada a fundamentação teórica da pesquisa abordando os principais conceitos utilizados. A metodologia usada nos experimentos de estudo e extração de conhecimento pelo uso das *Filtered*-ARNs está detalhada na Seção 4. Os resultados da mineração de dados com o conjunto de dados epidemiológicos e meteorológicos são apresentados e avaliados na Seção 5. Finalmente, todas as conclusões e trabalhos futuros estão listados na Seção 6.

2. Trabalhos Relacionados

Em [16], observa-se o uso da técnica de mineração de regras de associação Fuzzy. O método foi aplicado para a extração de regras em dados endemiológicos, socioeconômicos, ambientais e climáticos. Foram construídos modelos preditivos que tiveram como saída a incidência dos casos de dengue em uma determinada província das Filipinas, com quatro semanas de antecedência.

A relação entre dengue e fatores climáticos, foi estudada em [17], que utilizando a técnica de SVM(*Support Vector Machines*) demonstrou que fatores climáticos, como a temperatura média e precipitação de chuva, são determinantes para a previsão de surtos de dengue nas cidades estudadas. Foi observado também que o maior número de casos ocorre entre os meses de março e maio.

Em [18], a técnica *wavelet transform* foi aplicada a dados epidemiológicos e climáticos do estado da Paraíba, coletados entre 2007 e 2015. As análises realizadas mostraram que o casos de dengue começaram a aumentar a partir dos primeiros meses após o período de chuva, no entanto, a temperatura não se mostrou um fator relevante para a predição de surtos.

[19] utiliza regressão de Poisson e DLNM(*Distributed Lag Non-linear Model*) para avaliar e comparar a relação da umidade absoluta semanal, temperatura máxima, miníma e media, índice de chuva umidade relativa e velocidade do vento, com os casos de dengue de 2001 a 2009 em Singapura. Os resultados mostraram que a umidade absoluta teve um impacto mais estável na incidência de dengue do que a temperatura quando fatores virológicos foram levados em consideração.

No estudo de [20], foi desenvolvido um modelo relacionando a dinâmica populacional do mosquito *aedes aegypti* e a precipitação de chuva a fim de estudar como diferentes cenários de chuva influenciam a taxa de sobrevivência do mosquito na área de Taiwan. Foi descoberto uma relação não linear entre a duração da estação seca e a probabilidade de extinção dos mosquitos transmissores.

Em todos os trabalhos estudados, observa-se o estudo do relacionamento de dados epidemiológicos com meteorológicos. Embora o universo dos dados seja similar, nenhumas das pesquisas apresentou o uso de uma técnica de identificação automática de conhecimento pelo processo de formulação de hipóteses com o uso de redes.

3. Fundamentação teórica

Nesta seção são abordados conceitos chaves utilizados na execução da presente pesquisa. Sendo eles: KDD, Regras de associação e Redes de Regras de Associação Filtradas.

3.1 KDD

O KDD é um campo que se preocupa com o desenvolvimento de técnicas e métodos para trazer sentido aos dados. O problema básico abordado pelo KDD é o de transformar dados brutos e de baixo nível (que tipicamente são grandes demais para serem interpretados) em uma forma mais compreensível e de melhor interpretação. Para isso o núcleo desse processo é a aplicação de técnicas de mineração de dados para a extração e descoberta de padrões [21]. O processo completo de KDD é dividido em 3 etapas: pré-processamento, mineração dos dados e pós-processamento (Figura 1).

O pré-processamento consiste na aquisição e manipulação dos dados que podem conter informações úteis para auxiliar a descoberta de conhecimento. Após a obtenção dos valores brutos, selecionam-se os atributos de interesse e é feito um tratamento dos dados. Nesta etapa podem ocorrer a remoção de valores duplicados, a conversão de dados simbólicos para numéricos, além de processos de normalização, redução de dimensões, identificação e tratamento de *outliers* (valores que não seguem o mesmo padrão de distribuição do conjunto) e dados faltantes. Essas etapas são necessárias para preparar o *dataset* para a mineração propriamente dita[22].

Na etapa de mineração, são aplicadas técnicas para extração de padrões relevantes para aplicação estudada. Métodos de inteligência artificial, estatística ou pesquisa operacional podem ser utilizados a fim de que o conhecimento seja identificado de forma otimizada [15]. Neste trabalho foi utilizado o processo de Mineração de Regras de Associação.

Por fim, na etapa de pós-processamento, os resultados obtidos na parte de mineração são analisados. Várias técnicas podem ser usadas a fim de que o conhecimento seja identificado de modo mais eficiente, como o uso de Redes [13]. Neste trabalho foram construídas as Redes de Regras de Associação Filtradas (*Filtered*-ARNs) para otimização do processo de identificação automática de conhecimento de modo que os padrões gerados tenham maior probabilidade de serem relevantes ao domínio estudado.

3.2 Regras de Associação

Uma regra de associação caracteriza o quanto a presença de um conjunto de elementos em uma base de dados tem como consequência a presença de algum outro conjunto distinto de elementos nos mesmos registros. Desse modo, o objetivo das regras de associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados [13].

Definição 1: seja $i\{i_1, i_2, ..., i_n\}$ 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,

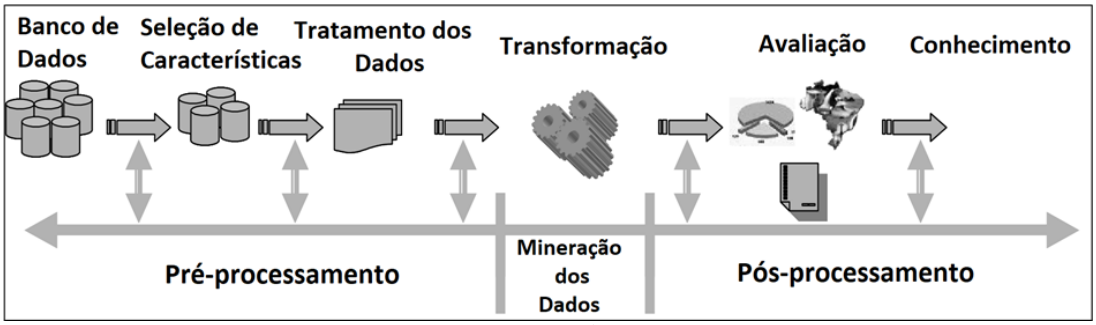

Figure 1. Processo de KDD [10]

em que cada transação D corresponde a um conjunto 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, fizeram 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 – *Left Hand Side*)[23].

Definição 2: para cada regra (LHS⇒RHS), extraída de um conjunto de transações T, é calculado um valor de suporte (*sup*), apresentado na Equação 1, que verifica a força de associação entre LHS e RHS (probabilidade da ocorrência de LHS ∪ RHS); e um valor de confiança (*conf*), apresentado na Equação 2, que mede a força da implicação lógica da regra (probabilidade condicional de RHS dado LHS) [23].

$$sup(LHS \Rightarrow RHS) = P(LHS \cup RHS) \quad (1)$$

$$conf(LHS \Rightarrow RHS) = P(RHS|LHS) \quad (2)$$

3.3 Redes de Regras de Associação

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 ou não consegue ser aplicado. Sendo assim, a busca de métodos que produzam um resultado de fácil interpretação é de extrema importância em aplicações [13]. Nesta pesquisa o método proposto utiliza estrutura de redes para facilitar a interpretação dos resultados.

A ideia central das ARNs 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. Resumindo, Redes de Regras de Associação possuem as seguintes características [24]:

- Poda no contexto: uma rede de regras de associação para podar no contexto de um objetivo específico. Alterando o objetivo resultara na poda de regas diferentes.
- Estrutura de rede: redes de regras de associação fornecem um mecanismo para determinar a relação entre as variáveis relevantes e o objetivo utilizando a construção de uma rede. Isso pode ajudar na analise dos efeitos de mudanças ocorridas de modo direto e indireto na mineração das regras de associação.
- Redes de regras de associação pode servir como uma ponte entre as saídas geradas pela mineração de regras de associação e sua avaliação.

Na Figura 2 é representado um exemplo de ARN, no 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 ⇒ A) e (D ⇒ A). Assim, os itens "B" e "D" são modelados no nível 1 da ARN e passam a ser 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" será uma das eliminadas no processo de poda, pois mesmo possuindo o item "C" como consequente, o item "B" já estava inserido na ARN em um nível abaixo, inviabilizando esta regra.

3.4 Redes de Regras de Associação Filtradas

Para que as regras geradas tenham uma maior probabilidade de representar um conhecimento verdadeiro, [13] elaborou a construção das Redes de Regras de Associação Filtradas, que consiste é um grafo direcionado construído para modelar todas as regras a fim de descrever um item selecionado. O resultado é um gráfico que explica o item e permite que o usuário construa hipóteses com base nesse item 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,

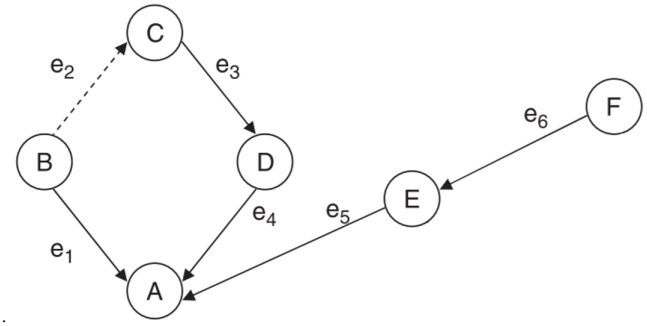

Figure 2. Exemplo de ARN com hiper-aresta reversa [15]

(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 em que não existe influência entre o antecedente e o consequente da regra. A poda é feita de acordo com o item objetivo, no qual a rede é modelada considerando apenas as regras que estão correlacionadas direta ou indiretamente a esse item. Nesse trabalho, para a filtragem da regras de associação, foram utilizadas as medidas *Added Value* e *Gain*.

- ***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 [25]. 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 for nulo, tem se uma coincidência aleatória, ou seja a frequência de LHS não altera a frequência de RHS.

$$AV = Conf(LHS \Rightarrow RHS) - sup(RHS) \quad (3)$$

- ***Gain*[0..1]:** É uma medida que da um *trade-off* entre suporte e confiança, auxiliando na seleção das regras de acordo com a frequências da mesma em relação à confiança mínima [26].

$$Gain = sup(LHS \cap RHS) - minconf.sup(LHS) \quad (4)$$

A utilização de medidas assimétricas como *Added Value* e *Gain* na filtragem das regras é umas das diferenças principais entre a *Filtered*-ARN e a ARN sendo que, na *Filtered*-ARN, o usuário pode visualizar um conjunto de itens que provem uma influência estatística em vez de elementos que apenas se relacionam com o item objetivo. Assim, a *Filtered*-ARN apresenta regras que indicam hipóteses com comprovação de dependência entre antecedente e consequente [13].

4. Metodologia

Nesta pesquisa, foi realizado todo o processo de mineração de dados, compreendendo desde a contrução do *dataset*, pré-processamento do conjunto de dados, extração e análise das regras de associação e construção das *Filtered*-ARNs para obteção de *insights* das características relacionadas a pacientes acometidos de dengue e parâmetros meteorológicos (Figura 3). A coleta dos dados foi realizada manualmente no site Tabnet (uma versão web do software TabWin - programa para a tabulação de dados desenvolvido pelo Datasus).

As variáveis disponíveis para a seleção eram diferentes para cada ano informado no sistema, então foram selecionadas apenas variáveis que possuíam dados em todos os anos estudados. As variáveis que formaram o *dataset* foram: município de notificação (apenas os municípios da planície litorânea piauiense), mês dos primeiros sintomas, raça, sexo, faixa etária, classificação final e critério de confirmação. A coleta foi feita seguindo os seguintes passos: i) selecionou-se os municípios que compõem a planície litorânea piauiense, o ano e o mês dos primeiros sintomas, ii) posteriormente, foram escolhidas as variáveis faixa etária, raça, sexo, classificação final e critério de confirmação por serem encontradas em todos os anos do banco de dados disponível. Deste modo, o site retornou uma tabela com o total de casos nos municípios selecionados e organizados pelos valores disponíveis da variável escolhida. Esse processo foi realizado para todos os meses de 2007 a 2014.

Para enriquecimento do *dataset*, também foram utilizados dados meteorológicos cedidos pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), colhidos pelo Instituto Nacional de Meteorologia (INMET). O conjunto de dados foi disponibilizado em forma de tabelas, no qual cada tabela correspondia a um mês do ano. Essas tabelas foram agrupadas sendo adicionada a quantidade de chuva em *mm* do respectivo mês e removendo as entradas que possuíam dados faltantes. Foram utilizados os mesmos anos de 2007 a 2014 para realização da Mineração de Regras de Associação.

Todas as variáveis numéricas foram categorizadas conforme a Tabela 1. A categorização foi feita a fim de que os grupos de valores pudessem representar um mesmo elemento e que as regras de associação representem o comportamento de influência entre as variáveis de modo mais eficiente.

A partir desses dados foram extraídas as regras de associação com o uso do algorítimo *Apriori-TID* implementado em Java®. Foram geradas apenas regras de tamanho igual a dois, com conjuntos LHS e RHS unitários, a fim de que fossem construídas as *Filtered*-ARNs. Os valores mínimos de confiança e suporte foram, 0.01 e 0, respectivamente, para que todas as regras possíveis fossem obtidas.

Após a extração das regras, foi executado o algoritmo de construção das *Filtered*-ARNs e com isso, ocorreu a filtragem das regras. Nessa etapa as regras que possuíam valores de *Added Value* = 0 e *Gain* $\leq$ 0.001 foram excluídas do conjunto. Com as regras filtradas, uma *Filtered*-ARN foi gerada com "[totaldecasos]=(328.6-365.0]" como nó alvo por tratar-se da categoria com maior número de casos ocorridos naquele período. A rede foi construída visualmente com o uso do software *Gephi* [27]. O *Gephi* é uma aplicação *open source* específica para a construção de redes e está disponível on-

mes1ºsintoma(s)	janeiro, fevereiro, março, abril, maio, junho, julho, agosto, setembro, outubro, novembro, dezembro
mmdechuva	(0-40.7], (40.7-81.4], (81.4-122.1], (122.1-162.8], (162.8-203.5], (203.5-244.2], (244.2-284.9], (284.9-325.6], (325.6-366.3], (366.3-407.0]
ano	2007, 2008, 2009, 2011, 2012, 2013, 2014
municipiodenotificacao	Bom_Principio_do_Piaui, Buriti_dos_Lopes, Cajueiro_da_Praia, Caraubas_do_Piaui, Caxingó, Cocal, Cocal_dos_Alves, Ilha_Grande, Luis_Correia, Murici_dos_Portelas, Parnaiba, Piracuruca, Sao-_Jose_do_Divino, Sao_Joao_da_Fronteira
totaldecasos	(1-37.4], (37.4-73.8], (73.8-110.2], (146.6-183.0], (219.4-255.8], (292.2-328.6], (328.6-365.0],
faixa_etaria:_<1ano	0, 1, 2, 3, 4, 6, 8, 9
faixa_etaria:_1-4	(0-1.7], (1.7-3.4], (3.4-5.1], (5.1-6.8], (6.8-8.5], (8.5-10.2], (10.2-11.9], (13.6-15.3], (15.3-17.0]
faixa_etaria:_5-9	(0-3.2], (3.2-6.4], (6.4-9.6] ,(9.6-12.8], (12.8-16.0], (16.0-19.2], (19.2-22.4], (28.8-32.0]
faixa_etaria:_10-14	(0-3.2], (3.2-6.4], (6.4-9.6], (9.6-12.8], (12.8-16.0], (16.0-19.2], (19.2-22.4], (25.6-28.8], (28.8-32.0]
faixa_etaria:_15-19	(0-3.6], (3.6-7.2], (7.2-10.8], (10.8-14.4], (14.4-18.0], (18.0-21.6], (25.2-28.8], (28.8-32.4], (32.4-36.0]
faixa_etaria:_20-39	(0-11.6],(11.6-23.2], (23.2-34.8], (34.8-46.4], (46.4-58.0], (58.0-69.6], (69.6-81.2],(92.8-104.4], (104.4-116.0]
faixa_etaria:_40-59	(0-9.3], (9.3-18.6], (18.6-27.9], (27.9-37.2], (37.2-46.5], (46.5-55.8], (55.8-65.1], (83.7-93.0]
faixa_etaria:_60-64	0, 1, 2, 3, 5, 7, 9, 10, 13
faixa_etaria:_65-69	0, 1, 2, 3, 4, 6, 9
faixa_etaria:_70-79	0, 1, 2, 3, 4, 6, 8, 10, 17
faixa_etaria:_80e+	0, 1, 2, 3, 7
raca:_ign/branco	(0-4.5], (4.5-9.0], (9.0-13.5], (13.5-18.0], (18.0-22.5], (22.5-27.0], (40.5-45.0]
raca:_branca	(0-9.0], (9.0-18.0], (18.0-27.0], (27.0-36.0], (36.0-45.0],(45.0-54.0], (81.0-90.0]
raca:_parda	(0-24.5], (24.5-49.0], (49.0-73.5], (73.5-98.0], (122.5-147.0], (171.5-196.0], (220.5-245.0]
raca:_preta	(0-2.7], (2.7-5.4], (5.4-8.1], (8.1-10.8], (10.8-13.5], (13.5-16.2], (16.2-18.9], (24.3-27.0]
raca:_amarela	0, 1, 2, 3, 5, 6, 8
raca:_indigena	0, 1, 2, 4
sexo:embranco	0, 1
sexo:masculino	(0-15.4], (15.4-30.8], (30.8-46.2], (46.2-61.6], (61.6-77.0], (77.0-92.4], (92.4-107.8], (123.2-138.6], (138.6-154.0]
sexo:_feminino	(0-21.1],(21.1-42.2],(42.2-63.3], (84.4-105.5], (105.5-126.6], (147.7-168.8], (168.8-189.9], (189.9-211.0]
class.final:ign/branco	0, 1, 2, 3, 7, 10
class.final:dengueclassico	(0-34.3], (34.3-68.6], (68.6-102.9], (137.2-171.5], (205.8-240.1], (274.4-308.7], (308.7-343.0]
class.final:denguecomcomplicacoes	0, 1, 2, 5, 7
class.final:dengue	(0-3.3], (3.3-6.6], (6.6-9.9] ,(9.9-13.2], (13.2-16.5], (19.8-23.1], (29.7-33.0]
class.final:febrehemorragicadodengue	0, 1, 2, 4, 5
class.final:inconclusivo	(0-3.6], (3.6-7.2], (7.2-10.8], (10.8-14.4], (21.6-25.2], (28.8-32.4], (32.4-36.0]
criterioconf.:ign/branco	(0-3.6], (3.6-7.2], (7.2-10.8], (10.8-14.4], (21.6-25.2], (28.8-32.4], (32.4-36.0]
criterioconf.:eminvestigacao	0, 1, 2, 3, 4 ,6, 7, 8, 11, 12
criterioconf.:laboratorial	(0-7.7], (7.7-15.4], (15.4-23.1], (23.1-30.8], (30.8-38.5], (38.5-46.2], (46.2-53.9], (53.9-61.6], (61.6-69.3], (69.3-77.0]
criterioconf.:clinico-epidemiologico	(0-30.2], (30.2-60.4], (90.6-120.8], (151.0-181.2], (241.6-271.8], (271.8-302.0]

Table 1. Categorias

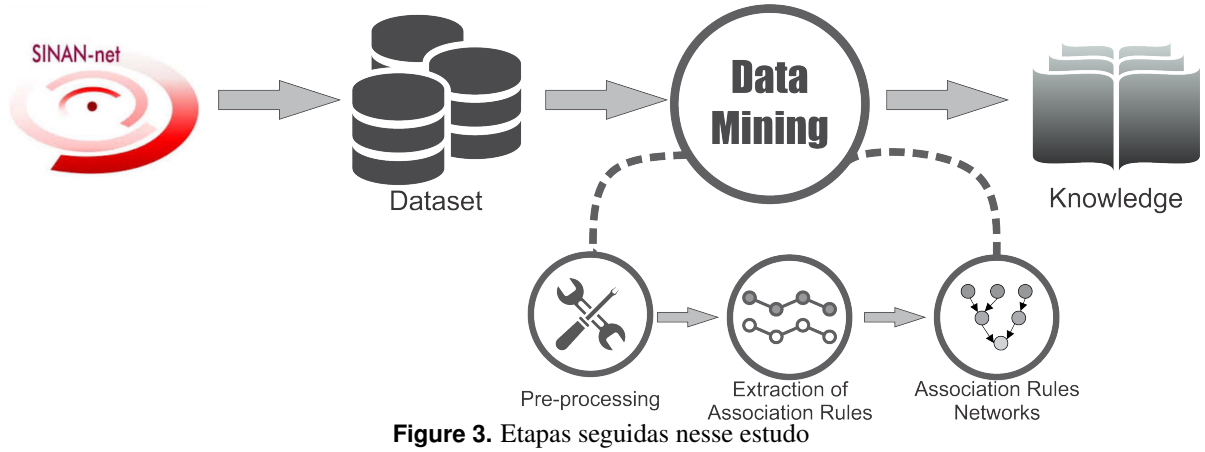

Figure 3. Etapas seguidas nesse estudo

line[2].

5. Resultados e Discussões

O algoritmo *Apriori* fez a extração de 19.427 regras de tamanho dois, sendo um número relativamente alto para que algum tipo de padrão fosse identificado. Com o algorítimo de construção da *Filtered*-ARN, após a filtragem com uso das medidas objetivas assimétricas, restaram 16.052 regras para construção da Rede. A *Filtered*-ARN construída foi elaborada como o nó alvo "[totaldecasos]=(328.6-365.0]" e é formada por 268 nós e 1027 arestas e está disponível on-line em ⟨http://bit.ly/2G6ZHPz⟩.

Para estudo, foram analisadas apenas os nós de nível 1 da Rede, i.e. os nós que estão conectados diretamente ao nó alvo. os nós de nível 1 são aqueles que influenciam diretamente o nó estudado. O nó alvo e os nós de nível 1 estão destacados na Figura 4. Foram detectados 30 (trinta) itens relacionados diretamente ao item objetivo.

Analisando os nós conectados ao item alvo, percebe-se a existência de todas as categorias das variáveis de faixa etária, raça e sexo. Portanto o número de casos independe de idade, raça e se os indivíduos são do sexo masculino ou feminino. Isso demonstra que todos os indivíduos estão sujeitos a contaminação, corroborando com o comportamento endêmico de todas as arboviroses.

Observa-se na *Filtered-ARN* (figura 4) que o maior número de casos sempre é encontrado no mês de maio ("[mes1ºsintoma(s)]=maio"), o que indica uma relação periódica da doença, assim como o demostrado em [18]. Essa informação é de grande valor para que as autoridades responsáveis possam agir em processos de prevenção da doença e tomada de decisão da vigilância epidemiológica.

Também é interessante destacar o item "[mmdechuva]=(81.4-122.1]", que de acordo com a tabela 1 e a figura 5 representa um valor intermediário dos índices pluviométricos da região estudada. O número de casos tende a aumentar quando o volume de chuvas está diminuindo o que também possibilita a intervenção direta no processos de prevenção à doença e, por consequência, a redução do número de casos.

Embora, os resultados alcançados já sejam evidenciados por outros trabalhos, a técnica demonstrou-se eficaz para identificação de conhecimento relevante. Sendo assim, pode-se adotar esta metodologia para extração de conhecimento em outras áreas de pesquisa, bem como com o acréscimo de parâmetros relacionados a dengue. Um fator limitante do trabalho ocorreu na fase de coleta dos dados, pois o SINAN é um *dataset* de grande proporções, porém carece de usabilidade e o usuário enfrenta dificuldades para a obtenção completa da informação nele contida.

6. Conclusões e trabalho futuros

Apesar das limitações da pesquisa e dos dados disponíveis, foi possível a validação da técnica pela obtenção de conhecimento verdadeiro de forma automática. Durante a fase de coleta dos dados foi possível observar que o TabNet é uma enorme fonte de dados, porém carece de usabilidade, o usuário enfrenta dificuldades no aprendizado para poder tirar proveito da plataforma, o que produz transtornos para a disseminação da informação nele contida. Em uma única pesquisa, não é possível ter acesso a todas as informações disponíveis na base de dados sobre uma determinada doença, o que provoca a necessidade da construção de ferramentas que possam fazer a união das informações desejadas.

Com a construção do *dataset*, foi possível realizar a descoberta do conhecimento, a qual foi otimizada pelo uso de uma estrutura em rede. Pelo uso das *Filtered*-ARNs pôde-se observar todos os principais fatores ligados diretamente ao maior número de casos de dengue. As hipóteses geradas já são reconhecidas como verdadeiras [17, 18], portanto, a metodologia foi validada como eficiente na geração de hipóteses e podem ser utilizadas em tarefas diretamente ligadas a vigilância epidemiológica.

Para trabalhos futuros, sugere-se que a mesma metodologia utilizada nesta pesquisa pode ser escalada para uma maior área geográfica, e para outras doenças. Há possibilidade de cruzamento dos resultados com dados socioeconômicos do

[2] ⟨https://gephi.org/users/download/⟩

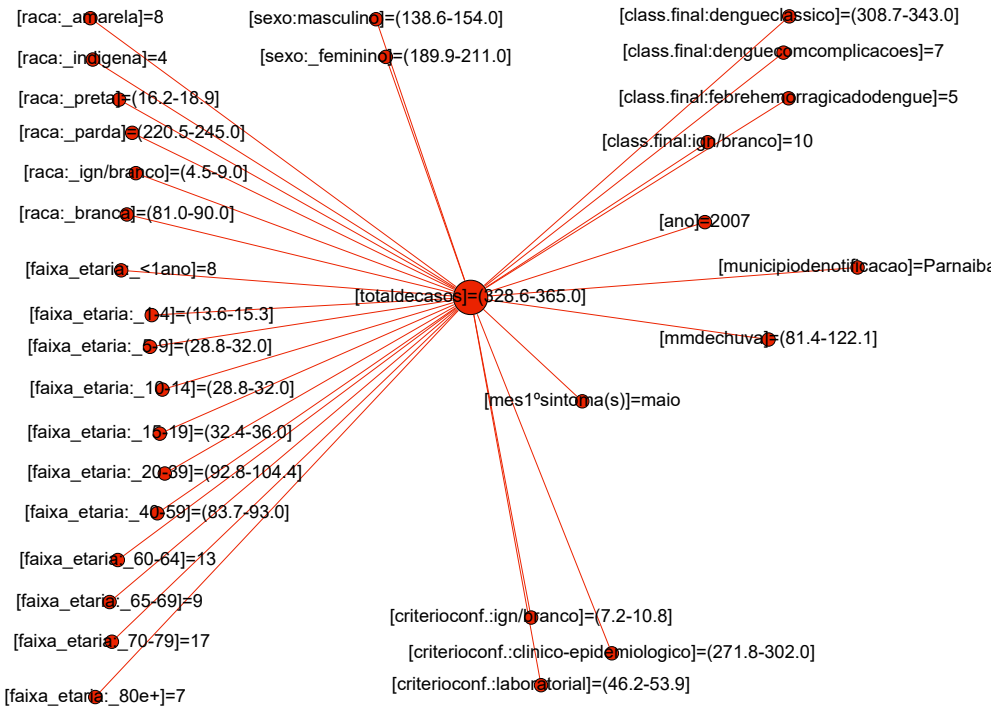

Figure 4. *Filtered*-ARN com "[totaldecasos]=(328.6-365.0]" como item alvo e nós de nível 1

local em estudo, inclusão de outras variáveis meteorológicas além do índice pluviométrico, como temperatura e taxa de evaporação, além de fazer uso de outras técnicas de extração de conhecimento.

Agradecimentos

À Universidade Estadual do Piauí pela estrutura fornecida aos estudos iniciais e ao Grupo de Estudo e Desenvolvimento de Aplicações Inteligentes (GEDAI) por toda troca de experiências e conhecimentos trocados durante a elaboração da pesquisa.

Contribuições dos autores

Joan D. S. Silva, Bacharel em Ciência da Computação, participou de todo o planejamento, execução e confecção do manuscrito. Jâina Carolina Meneses Calçada, Enfermeira e Mestre em Saúde da Família, participou de todo o planejamento e da aquisição dos dados epidemiológicos, da análise dos resultados e confecção do manuscrito. Solange Oliveira Rezende, Profª. Drª., participou na elaboração da técnica computacional utilizada e na análise experimental completa. Dario Brito Calçada, Prof. Dr., foi orientador e supervisor do projeto e participou de todas as etapas do mesmo. Todos os autores participaram da construção científica do manuscrito, considerando as etapas de leitura e revisão.

References

[1] WHO. *Dengue and severe dengue*. 2019. Disponível em: ⟨https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue⟩.

[2] FARES, R. C. G. et al. Epidemiological Scenario of Dengue in Brazil. *BioMed Research International*, Hindawi Limited, v. 2015, p. 1–13, 2015. Disponível em: ⟨https://doi.org/10.1155/2015/321873⟩.

[3] BRAGA, I. A.; VALLE, D. Aedes aegypti: histórico do controle no Brasil. *Epidemiologia e Serviços de Saúde*, scielo, v. 16, p. 113 – 118, 06 2007. Disponível em: ⟨http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742007000200006&nrm=iso⟩.

[4] MONTEIRO, E. S. C. et al. Aspectos epidemiológicos e vetoriais da dengue na cidade de Teresina, Piauí-Brasil, 2002 a 2006. *Epidemiologia e Serviços de Saúde*, scielo, v. 18, p. 365 – 374, 12 2009. Disponível em: ⟨http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742009000400006&nrm=iso⟩.

[5] HII, Y. L. et al. Forecast of dengue incidence using temperature and rainfall. *PLOS Neglected Tropical Diseases*, Public Library of Science, v. 6, n. 11, p. 1–9, 11 2012. Disponível em: ⟨https://doi.org/10.1371/journal.pntd.0001908⟩.

[6] THOMSON, M. C. et al. Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana. *The American journal of tropical medicine and hygiene*, ASTMH, v. 73, n. 1, p. 214–221, 2005.

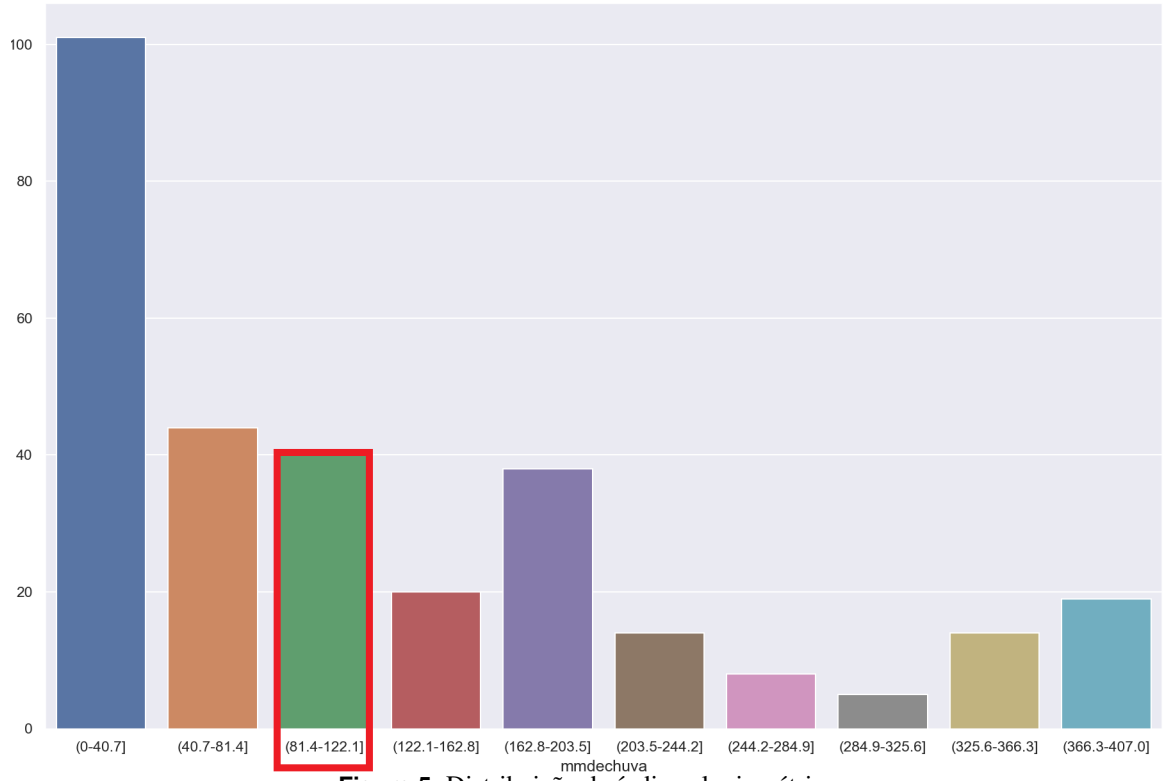

Figure 5. Distribuição do índice pluviométrico

[7] DEGALLIER, N. et al. Toward an early warning system for dengue prevention: modeling climate impact on dengue transmission. *Climatic Change*, Springer, v. 98, n. 3-4, p. 581–592, 2010.

[8] WANG, J.; OGDEN, N. H.; ZHU, H. The impact of weather conditions on culex pipiens and culex restuans (diptera: Culicidae) abundance: a case study in peel region. *Journal of medical entomology*, Oxford University Press Oxford, UK, v. 48, n. 2, p. 468–475, 2011.

[9] SINAN. *Sistema de Informação de Agravos de Notificação*. 2016. Disponível em: ⟨http://www.portalsinan.saude.gov.br⟩.

[10] TRINDADE, C. M. *Identificação do Comportamento das Hepatites Virais a partir da exploração de bases de dados de Saúde Pública. 2005, 139f.* Tese (Doutorado) — Dissertação (Mestrado em Tecnologia em Saúde)-Pontifícia Universidade Católica do Paraná, PUCPR, 2005., 2005.

[11] ANGUERA, A. et al. Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry. *Computational and Structural Biotechnology Journal*, Elsevier BV, v. 14, p. 185–199, 2016. Disponível em: ⟨https://doi.org/10.1016/j.csbj.2016.05.002⟩.

[12] FATHIMA, A. S.; MANIMEGALAI, D.; HUNDEWALE, N. A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue. *International Journal of Computer Science Issues (IJCSI)*, Citeseer, v. 8, n. 6, p. 322, 2011.

[13] CALÇADA, D. B.; PADUA, R. de; REZENDE, S. O. Asymmetric Objective Measures applied to Filter Association Rules Networks. In: *XLIV Latin American Computer Conference (CLEI) Asymmetric*. São Paulo: [s.n.], 2018. p. 258–267.

[14] WENG, C.-H. Identifying association rules of specific later-marketed products. *Applied Soft Computing*, Elsevier, v. 38, p. 518–529, 2016.

[15] CALÇADA, D. B. *Redes de regras de associação filtradas e multialvo*. 199 p. Tese (Doutorado) — Universidade de São Paulo, 2019.

[16] BUCZAK, A. L. et al. Prediction of High Incidence of Dengue in the Philippines. *PLoS Neglected Tropical Diseases*, Public Library of Science (PLoS), v. 8, n. 4, p. e2771, abr. 2014. Disponível em: ⟨https://doi.org/10.1371/journal.pntd.0002771⟩.

[17] STOLERMAN, L. M.; MAIA, P. D.; KUTZ, J. N. Forecasting dengue fever in Brazil: An assessment of climate conditions. *PLOS ONE*, Public Library of Science (PLoS), v. 14, n. 8, p. e0220106, ago. 2019. Disponível em: ⟨https://doi.org/10.1371/journal.pone.0220106⟩.

[18] SANTOS, C. A. G. et al. Correlation of dengue incidence and rainfall occurrence using wavelet transform for João Pessoa city. *Science of The Total Environment*, Else-

vier BV, v. 647, p. 794–805, jan. 2019. Disponível em: ⟨https://doi.org/10.1016/j.scitotenv.2018.08.019⟩.

[19] XU, H.-Y. et al. Statistical Modeling Reveals the Effect of Absolute Humidity on Dengue in Singapore. *PLoS Neglected Tropical Diseases*, Public Library of Science (PLoS), v. 8, n. 5, p. e2805, maio 2014. Disponível em: ⟨https://doi.org/10.1371/journal.pntd.0002805⟩.

[20] VALDEZ, L.; SIBONA, G.; CONDAT, C. Impact of rainfall on Aedes aegypti populations. *Ecological Modelling*, Elsevier BV, v. 385, p. 96–105, out. 2018. Disponível em: ⟨https://doi.org/10.1016/j.ecolmodel.2018.07.003⟩.

[21] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996.

[22] OLARU, C.; GEURTS, P.; WEHENKEL, L. Data mining tools and application in power system engineering. In: TRONDHEIM, NORWAY. *Proceedings of the 13th Power System Computation Conference, PSCC99*. [S.l.], 1999. p. 324–330.

[23] AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499.

[24] PANDEY, G. et al. Association Rules Network: Definition and Applications. *Statistical Analysis and Data Mining*, v. 1, n. 4, p. 260–179, 2009.

[25] SAHAR, S. What is interesting: studies on interestingness in knowledge discovery. *Phd Thes, Tel-Aviv University The*, Citeseer, 2003.

[26] FUKUDA, T. et al. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. *Acm Sigmod Record*, ACM, v. 25, n. 2, p. 13–23, 1996.

[27] BASTIAN, M.; HEYMANN, 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.