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], **Francisco das C. Rocha**[1]

[1]Universidade Estadual do Piauí (UESPI)
Campus Alexandre Alves de Oliveira – Parnaíba, PI – Brazil

joandavi98@hotmail.com

Abstract. *Dengue is an endemic disease in Brazil since the 1980s and since 1996 in Piauí. The number of cases increases eachyear with the incidence of more severe symptoms. This research aimed to perform the automatic identificationof factors related to the number of dengue occurrences. A dataset was built consisting of data available in the Information System for Notifiable Diseases (SINAN) and meteorological data provided by Embrapa Meio-Norteregarding the municipalities of the Piauí coast. The technique used was that of Filtered Association RulesNetworks (Filtered-ARN) that allows a visual analysis of knowledge through the use of network structures, inaddition to filtering the rules by selecting those that have mathematical evidence of influence. As a main result,it was generated the knowledge that the greatest number of cases occurs in the month of May and when thepluviometric indexes are decreasing, in addition to the fact that socio-cultural and race factors do not interferein the identification of the population with the highest risk of contagion. The innovation brought the use of acomputational technique of automatic knowledge discovery that can assist in the formation of prevention actionsby the authorities responsible for epidemiological surveillance.*

Resumo. *A dengue é uma doença viral 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 realizar a identificação automática do conhecimento dos 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 fornecidos pela Embrapa Meio-Norte referente aos municípios do litoral piauiense. A principal técnica utilizada foi a de Redes de Regras de Associação Filtradas (Filtered-ARN) que possibilita uma análise visual do conhecimento por meio do uso de estruturas de rede, além de realizar a filtragem das regras selecionando as que possuem comprovação matemática de influência. Como resultado principal, foi gerado o conhecimento que o maior número de casos ocorrem no mês de maio e quando os índices pluviométricos estão se reduzindo, além de que fatores socioculturais e de raça não interferem em nada para identificação de população com maior risco de contágio. A inovação trouxe o uso de uma técnica computacional de descoberta automática*

de conhecimento que pode auxiliar na formação de ações de prevenção pelas autoridades responsáveis pela vigilância epidemiológica.

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 anos de 2010 e 2016 saltou de 0,5 milhão para 3,34 milhões de casos reportados à Organização Mundial da Saúde [WHO 2019]. 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[Bhatt et al. 2013, Brady et al. 2012].

No Brasil, os primeiros registro da doença ocorreram na cidade de Curitiba-PA, no final do século XIX. Na década de 80 houve epidemias nos Estados de Roraima, Minas Gerais, São Paulo, Bahia, Pernambuco, Ceará, Alagoas e Rio de Janeiro. Neste último, ocorreu uma epidemia em 1986, na qual a dengue adquiriu importância epidemiológica. A doença logo atingiu a Região Nordeste tornando-se endêmica no país[Braga and Valle 2007].

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 [Monteiro et al. 2009].

A ocorrência de dengue é influenciada por uma complexa mistura de fatores como, a rápida urbanização e crescente densidade populacional, capacidade dos sistemas de saúde, variáveis meteorológicas e eficiência no controle do vetor pela vigilância sanitária. Como não existe uma 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 [Hii et al. 2012].

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, bem como o aumento da expectativa de vida. Para auxiliar nesse complexo procedimento a informação 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 [Thomson et al. 2005, Degallier et al. 2010, Wang et al. 2011].

Uma importante fonte do conhecimento usado 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 uma fator essencial para ações de prevenção, é facultado aos estado e municípios a inclusão de outros problemas de saúde importantes na sua região. 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

[SINAN 2016].

Com o propósito de melhorar a descoberta de padrões relevantes e, possivelmente, inovadores em grandes bases de dados, como o SINAN, o uso do KDD (sigla em inglês para descoberta de conhecimento em base de dados) 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. [TRINDADE 2005, Ohsaki et al. 2002, Fathima et al. 2011].

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 e o conhecimento que se espera obter [Calçada et al. 2018].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 compra linguiça". O exemplo ilustra uma das características mais atrativas das Regras de Associação, ela é expressa em uma forma muito fácil de ser compreendidas quando formuladas por *itemsets* de tamanho reduzido [Weng 2016, Calçada 2019].

A quantidade de Regras de Associação extraídas está diretamente relacionada ao número de itens que formam a base de dados. Em um conjunto de dados 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 9900 regras com *itemsets* de um elemento e 98000100 regras se o os *itemsets* possuírem dois elementos, um crescimento exponencial.

Neste sentido a utilização de redes para a análise das regras geradas é de grande auxílio nesse processo. As Redes de Regras de Associação (ARN - do inglês *Association Rules Network*), tem com ideia central sintetizar, podar e integrar no contexto dos objetivos específicos da pesquisa as Regras de Associação descobertas pelo algoritmo de mineração.

O problema abordado neste trabalho foi o de identificar o comportamento da dengue nos municípios da planície litorânea piauiense formado pelos municípios 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. Este artigo está organizado da seguinte forma: Nada Seção 2 é apresentada a fundamentação teórica da pesquisa abordando os principais conceitos utilizados, na Seção

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

3 é apresentada a metodologia seguida por essa pesquisa, a Seção 4 aborda os resultados obtidos e suas consequências, por fim, na Seção 5 são relatados as conclusões do trabalho e sugestões de trabalhos futuros.

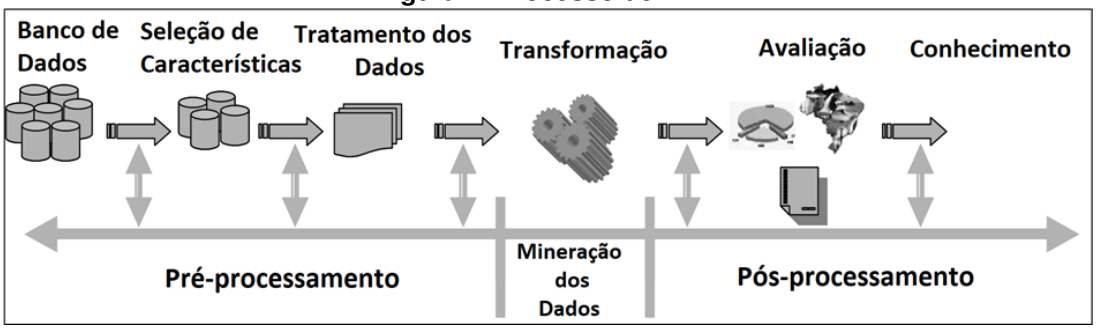
2. Fundamentação teórica

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

2.1. 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 [Fayyad et al. 1996]. O processo completo de KDD é dividido em 3 etapas: pré-processamento, mineração dos dados e pós-processamento (Figura 1).

Figura 1. Processo de KDD



Fonte: [TRINDADE 2005]

O pré-processamento consiste na aquisição e manipulação dos dados que podem conter informações úteis para o problema abordado. Após a obtenção dos dados brutos, são selecionados os atributos de interesse e os dados são tratados, valores duplicados são removidos, e ainda pode ser feita a conversão de alguns tipos de dados, como, por exemplo, converter um dado simbólico para um dado numérico, além de processos de normalização, redução de dimensões, identificação e tratamento de *outliers* (valores que não seguem o mesmo padrão de distribuição dos dados) e tratamento de dados faltantes. Essas etapas são necessárias para preparar os dados para a etapa de mineração[Olaru et al. 1999].

Na etapa de mineração, são aplicadas técnicas para extração de padrões relevantes para aplicação estudada. Na etapa de mineração, métodos que fazem uso de inteligência artificial, estatística ou pesquisa operacional podem ser utilizados a fim de que o conhecimento seja identificado de forma otimizada. Alguns algoritmos fazem uso de Redes para a visualização dos resultados. Neste trabalho foi utilizado o processo de Mineração de Regras de Associação [Calçada 2019].

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 [Calçada et al. 2018]. Neste trabalho foram construídas as Redes de Regras de Associação Filtradas para otimização do processo de identificação automática de conhecimento de modo que os padrões gerados tenham maior probabilidade de serem relevantes ao domínio estudado.

2.2. 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. Por exemplo, observando os dados de venda de um supermercado, sabe se que 80% dos clientes que compram o produto *Q* também adquirem, na mesma ocasião, o produto *W* [Calçada et al. 2018].

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, 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, 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 – *Right Hand Side*)[Agrawal et al. 1994].

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) [Agrawal et al. 1994].

$$sup(LHS \Rightarrow RHS) = P(LHS \cup RHS) \tag{1}$$

$$conf(LHS \Rightarrow RHS) = P(RHS|LHS) \tag{2}$$

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. 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 [Calçada 2019].

Considerando-se que LHS = CAFÉ e RHS = PÃO, pode se calcular o suporte e a confiança para a regra (CAFÉ ⇒ PÃO) tendo como resultados *sup*(CAFÉ ∪ PÃO) = 0,3

Tabela 1. Compras diárias

	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: [Calçada 2019]

ou 30% e $conf$(CAFÉ $\Rightarrow$ PÃO) = 1 ou 100%. Este resultado implica em 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.

2.3. Redes de Regras de Associação

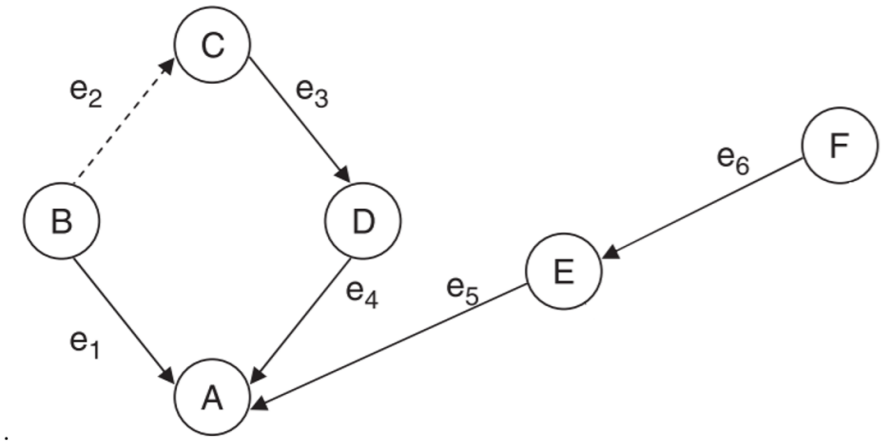
Como os algoritmos de regras de associação são capazes de extrair todas as regras de associação de acordo com um suporte mínimo e valor mínimo de confiança, o número de regras extraídas geralmente supera a capacidade de interpretação do usuário. O conhecimento gerado muitas vezes é inconclusivo ou não consegue ser aplicado. Sendo assim, a busca de métodos que produzam um resultado de fácil interpretação é de extrema importância em aplicações [Calçada et al. 2018]. Nesta pesquisa o método proposto utiliza estrutura de redes para facilitar a interpretação dos resultados.

A ideia central das Redes de Regras de Associação (ARN - do inglês *Association Rules Network*) 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 [Pandey et al. 2009]:

- 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 análise 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 $\Rightarrow$ A) e (D $\Rightarrow$ 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.

Figura 2. Exemplo de ARN com hiper-aresta reversa



Fonte:[Calçada 2019]

2.4. Redes de Regras de Associação Filtradas

Para que as regras geradas tenham uma maior probabilidade de representar um conhecimento verdadeiro, [Calçada et al. 2018] elaborou a construção das Redes de Regras de Associação Filtradas (*Filtered*-ARN – do inglês *Filtered-Association Rules Network*), que consiste em 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, (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 [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 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 [Fukuda et al. 1996].

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

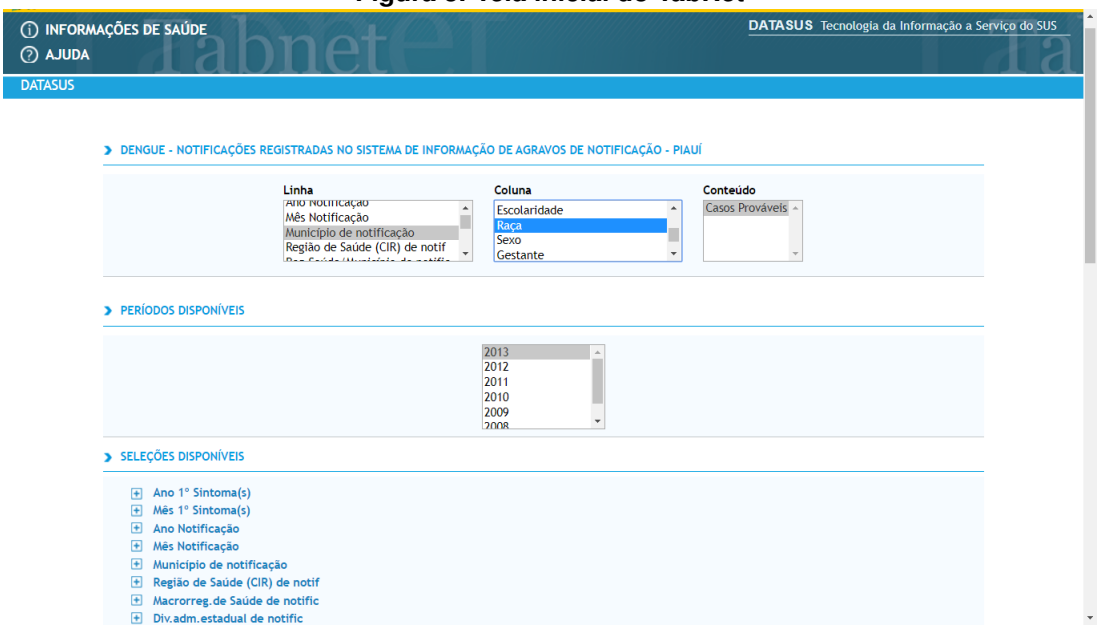
A utilização de medidas assimétricas como *Added Value* e *Gain* na filtragem das regras é umas das diferenças principais entre a *Filtered*-ARN e a ARN sendo que, na *Filtered*-ARN, o usuário pode visualizar um conjunto de itens que provem uma influência estatística em vez de elementos que apenas se relacionam com o item objetivo. Assim, a *Filtered*-ARN apresenta regras que indicam hipóteses com comprovação de dependência entre antecedente e consequente [Calçada et al. 2018].

3. Metodologia

Todo o processo de KDD foi realizado. Desde a coleta de dados e construção do *dataset*, o pré-processamento do conjunto de dados, bem como a extração das regras de associação, filtragem e a construção das *Filtered*-ARNs para identificação de conhecimento.

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). Na figura 3 é apresentada a interface principal do sistema.

Figura 3. Tela inicial do TabNet

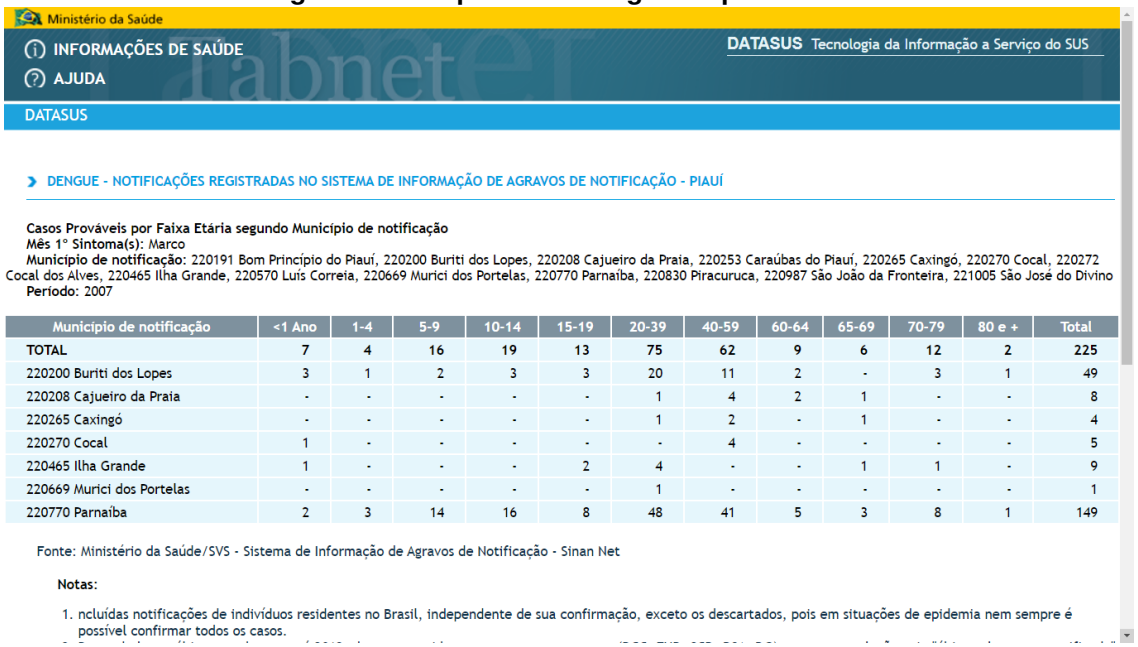


Fonte: Elaborada pelo autor

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* elaborado neste trabalho 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 (um exemplo de uma tabela gerada pode ser visto na Figura 4). Esse processo foi realizado para todos os meses de 2007 a 2014.

Figura 4. Exemplo de tabela gerada pelo TabNet



Ministério da Saúde

INFORMAÇÕES DE SAÚDE

AJUDA

DATASUS Tecnologia da Informação a Serviço do SUS

DATASUS

DENGUE - NOTIFICAÇÕES REGISTRADAS NO SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO - PIAUÍ

Casos Prováveis por Faixa Etária segundo Município de notificação
Mês 1° Sintoma(s): Março
Município de notificação: 220191 Bom Princípio do Piauí, 220200 Buriti dos Lopes, 220208 Cajueiro da Praia, 220253 Caraúbas do Piauí, 220265 Caxingó, 220270 Cocal, 220272 Cocal dos Alves, 220465 Ilha Grande, 220570 Luís Correia, 220669 Murici dos Portelas, 220770 Parnaíba, 220830 Piracuruca, 220987 São João da Fronteira, 221005 São José do Divino
Período: 2007

Município de notificação	<1 Ano	1-4	5-9	10-14	15-19	20-39	40-59	60-64	65-69	70-79	80 e +	Total
TOTAL	7	4	16	19	13	75	62	9	6	12	2	225
220200 Buriti dos Lopes	3	1	2	3	3	20	11	2	-	3	1	49
220208 Cajueiro da Praia	-	-	-	-	-	1	4	2	1	-	-	8
220265 Caxingó	-	-	-	-	-	1	2	-	1	-	-	4
220270 Cocal	1	-	-	-	-	-	4	-	-	-	-	5
220465 Ilha Grande	1	-	-	-	2	4	-	-	1	1	-	9
220669 Murici dos Portelas	-	-	-	-	-	1	-	-	-	-	-	1
220770 Parnaíba	2	3	14	16	8	48	41	5	3	8	1	149

Fonte: Ministério da Saúde/SVS - Sistema de Informação de Agravos de Notificação - Sinan Net

Notas:

1. ncluídas notificações de indivíduos residentes no Brasil, independente de sua confirmação, exceto os descartados, pois em situações de epidemia nem sempre é possível confirmar todos os casos.

Fonte: Elaborada pelo autor

Para enriquecimento do *dataset*, também foram utilizados dados meteorológicos cedidos pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), colhidos pelo Instituto Nacional de Meteorologia (INMET). Foram disponibilizados em forma de tabelas, onde 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 2. 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 apenas um item em LHS quanto em RHS a fim de que fossem cons-

Tabela 2. Categorias

mes1°sintoma(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]

Fonte: Elaborada pelo autor

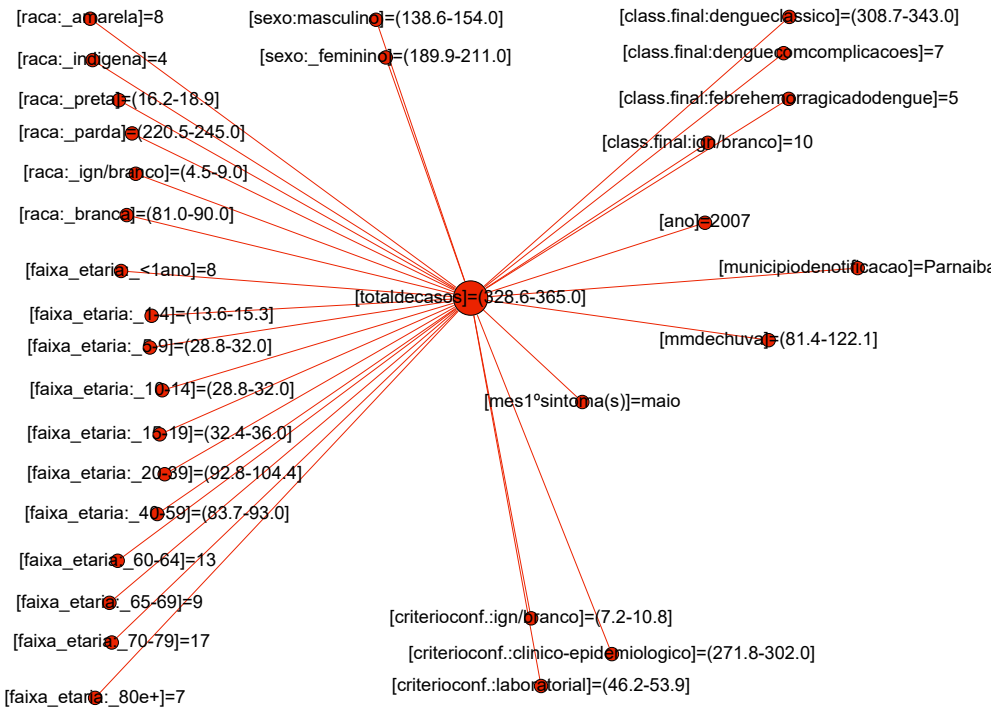
truídas as *Filtered*-ARNs. Os valores mínimos de confiança e suporte foram, 0,01 e 0, respectivamente, para que todas as regras possíveis fossem obtidas.

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* [Bastian et al. 2009]. O *Gephi* é uma aplicação *open source* específica para a construção de redes e está disponível on-line[2].

4. 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 algoritmo 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`.

Figura 5. *Filtered*-ARN com "[totaldecasos]=(328.6-365.0]"como item alvo e nós de nível 1



Fonte: Elaborada pelo autor

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

[2]`https://gephi.org/users/download/`

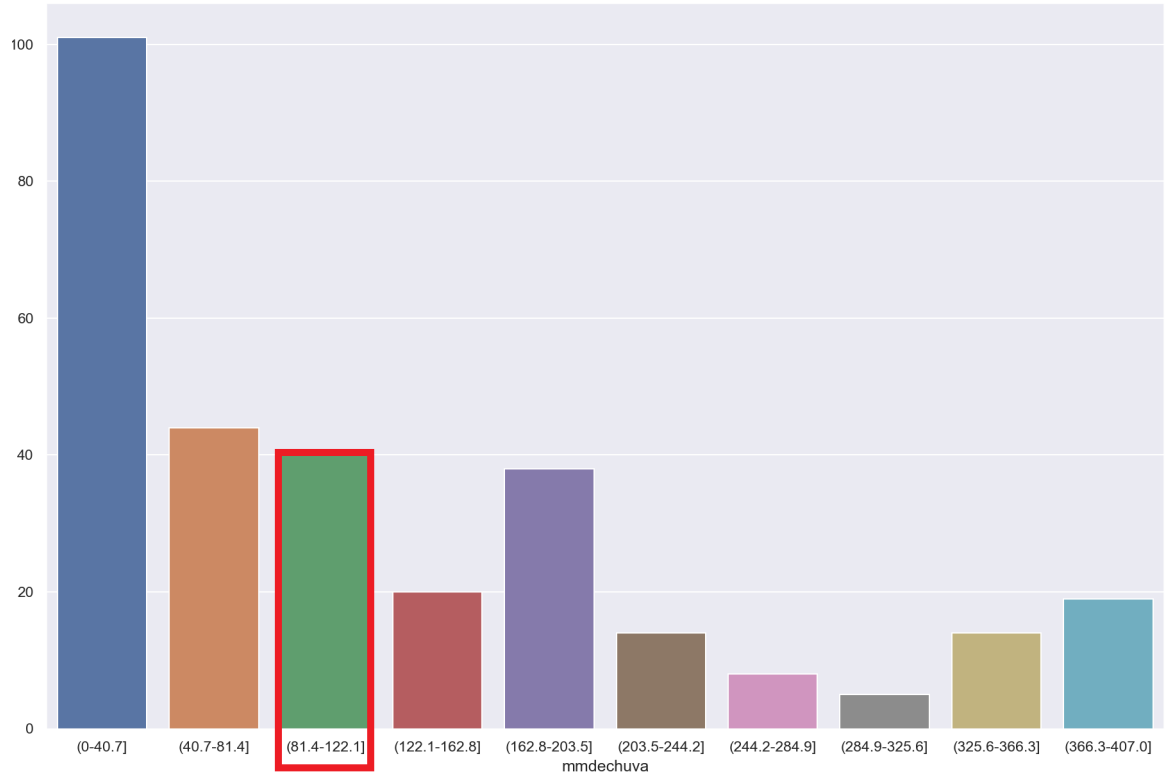
diretamente o nó estudado. O nó alvo e os nós de nível 1 estão destacados na Figura 5. Foram detectados 30 (trinta) itens relacionados diretamente ao item objetivo.

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* que o maior número de casos sempre é encontrado no mês de maio ("[mes1°sintoma(s)]=maio"), o que indica uma relação periódica da doença. Essa informação é de grade 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 2 e a Figura 6 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.

Figura 6. Distribuição do índice pluviométrico



Fonte: Elaborada pelo autor

5. Conclusões e trabalho futuros

Apesar das limitações da pesquisa e dos dados disponíveis, foi possível a geração de resultados interessantes. 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. Estas informações foram consideradas de grande relevância 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 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.

Referências

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., Myers, M. F., George, D. B., Jaenisch, T., Wint, G. R. W., Simmons, C. P., Scott, T. W., Farrar, J. J., and Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, 496(7446):504–507.

Brady, O. J., Gething, P. W., Bhatt, S., Messina, J. P., Brownstein, J. S., Hoen, A. G., Moyes, C. L., Farlow, A. W., Scott, T. W., and Hay, S. I. (2012). Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Neglected Tropical Diseases*, 6(8):e1760.

Braga, I. A. and Valle, D. (2007). Aedes aegypti: histórico do controle no Brasil. *Epidemiologia e Serviços de Saúde*, 16:113 – 118.

Calçada, D. B. (2019). *Redes de regras de associação filtradas e multialvo*. PhD thesis, Universidade de São Paulo.

Calçada, D. B., de Padua, R., and Rezende, S. O. (2018). Asymmetric Objective Measures applied to Filter Association Rules Networks. In *XLIV Latin American Computer Conference (CLEI) Asymmetric*, pages 258–267, São Paulo.

Degallier, N., Favier, C., Menkes, C., Lengaigne, M., Ramalho, W. M., Souza, R., Servain, J., and Boulanger, J.-P. (2010). Toward an early warning system for dengue prevention: modeling climate impact on dengue transmission. *Climatic Change*, 98(3-4):581–592.

Fathima, A. S., Manimegalai, D., and Hundewale, N. (2011). A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue. *International Journal of Computer Science Issues (IJCSI)*, 8(6):322.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Fukuda, T., Morimoto, Y., Morishita, S., and Tokuyama, T. (1996). Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. *Acm Sigmod Record*, 25(2):13–23.

Hii, Y. L., Zhu, H., Ng, N., Ng, L. C., and Rocklöv, J. (2012). Forecast of dengue incidence using temperature and rainfall. *PLOS Neglected Tropical Diseases*, 6(11):1–9.

Monteiro, E. S. C., Coelho, M. A. E., Cunha, I. S. d., Cavalcante, M. d. A. S., and Carvalho, F. A. A. d. A. (2009). Aspectos epidemiológicos e vetoriais da dengue na cidade de Teresina, Piauí- Brasil, 2002 a 2006. *Epidemiologia e Serviços de Saúde*, 18:365 – 374.

Ohsaki, M., Sato, Y., Yokoi, H., and Yamaguchi, T. (2002). A rule discovery support system for sequential medical data, in the case study of a chronic hepatitis dataset. In *Workshop Notes of the International Workshop on Active Mining, at IEEE International Conference on Data Mining*, page 121.

Olaru, C., Geurts, P., and Wehenkel, L. (1999). Data mining tools and application in power system engineering. In *Proceedings of the 13th Power System Computation Conference, PSCC99*, pages 324–330. Trondheim, Norway.

Pandey, G., Chawla, S., Poon, S., Arunasalam, B., and Davis, J. G. (2009). Association Rules Network: Definition and Applications. *Statistical Analysis and Data Mining*, 1(4):260–179.

Sahar, S. (2003). What is interesting: studies on interestingness in knowledge discovery. *Phd Thes, Tel-Aviv University The*.

SINAN (2016). Sistema de informação de agravos de notificação.

Thomson, M. C., Mason, S. J., Phindela, T., and Connor, S. J. (2005). Use of rainfall and sea surface temperature monitoring for malaria early warning in botswana. *The American journal of tropical medicine and hygiene*, 73(1):214–221.

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

Wang, J., Ogden, N. H., and Zhu, H. (2011). The impact of weather conditions on culex pipiens and culex restuans (diptera: Culicidae) abundance: a case study in peel region. *Journal of medical entomology*, 48(2):468–475.

Weng, C.-H. (2016). Identifying association rules of specific later-marketed products. *Applied Soft Computing*, 38:518–529.

WHO (2019). Dengue and severe dengue.