CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

TITULO

Discente

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Orientador

Rio de Janeiro,

Setembro de 2022

Titulo

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Discente

Aprovada por:

Presidente, Prof. nome orientador, D.Sc. (orientador)

---

nome coorientador, D.Sc. (coorientador)

Membro 1, D.Sc.

Membro 2, D.Sc.

Rio de Janeiro,

Setembro de 2022

FICHA CATALOGRÁFICA A SER SOLICITADA NA BIBLIOTECA DO CEFET/RJ APÓS A REVISÃO FINAL DO TEXTO.

MAIS INFORMAÇÕES EM:
https://eic.cefet-rj.br/ppcic/index.php/ficha-catalografica-nada-consta/

DEDICATÓRIA

DEDICATÓRIA texto

AGRADECIMENTOS

Agradeço aos ...

Por fim, agradeço ao órgão de fomento, responsável por fomentar esta pesquisa.

RESUMO

Titulo

Discente

Orientadores:
Orientador

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Resumo

Palavras-chave:

Rio de Janeiro,

Setembro de 2022

ABSTRACT

Titulo

Discente

Advisors:
Orientador

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

FAZER

Key-words:
Abstract

Rio de Janeiro,

Setembro de 2022

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

Capítulo I Introdução

Dentre as diversas possibilidades para minerar dados, a mineração de padrões frequentes desempenha um papel relevante para o levantamento de associações, correlações e muitas outras relações interessantes entre os dados [Han et al., 2011]. Os itens frequentes em um *dataset* podem ser expressos por regras de associação. As regras de associação funcionam de forma que apresentam itens frequentes na posição de antecedente levando a um item frequente na posição de consequente [Lodhi, 2013]. Desta forma, os itens no antecedente são as condições necessárias para se chegar ao item do consequente.

Capítulo II Metodologia

II.1 Parâmetros

Os tamanhos máximo e mínimo das regras determinam quantos itens frequentes devem aparecer na regra de associação, contando tanto os itens no antecedente quanto o do consequente. O tamanho mínimo de regras determina o menor número de itens enquanto o máximo determina o maior número de itens para geração das regras. Se ambos os tamanhos são definidos por 2, apenas regras com um item no antecedente e outro no consequente são criadas, assim como a regra dada como exemplo no parágrafo anterior. A Figura II.1 mostra exemplos de regras com tamanhos diferentes. Os parâmetros de tamanho mínimo e máximo devem ser pensados com cuidado. As regras com muitos itens podem ser difíceis de interpretar e trazer redundância, além de demandar um esforço computacional maior para gerar mais regras. Por outro lado, diminuir o tamanho máximo pode limitar a investigação e omitir correspondências importantes.

Tabela II.1: Exemplos de regras de associação com tamanhos (ordem) de 2 a 5

Regras	Tamanho
{rs.notificacao=Codo} ⇒ {resultado.exame=Negativo}	2
{rs.notificacao=Central; resultado.exame=Vivax} ⇒ {tipo.deteccao=Passiva}	3
{rs.notificacao=Centro Norte; mes.notificacao=12; ano.notificacao=2014} ⇒ {resultado.exame=Negativo}	4
{rs.notificacao=Area Norte; tipo.deteccao=Ativa; mes.notificacao=02; ano.notificacao=2009} ⇒ {resultado.exame=Negativo}	5

O suporte define a frequência do item dentro do *dataset* [Berzal et al., 2002]. Ao se definir um suporte de 50%, por exemplo, está-se estipulando que somente itens que aparecem em pelo menos metade das transações são considerados como frequentes. Dado um conjunto de dados D, o suporte (sup) para um item X é o percentual de ocorrências de X em relação a D. Compreendendo-se as ocorrências de X em D como sendo um evento, a Equação II.1 apresenta o suporte de do item X como sendo a probabilidade de X ocorrer em D.

$$sup(X) = P(X) \tag{II.1}$$

Analisando-se uma regra de associação do tipo $X \Rightarrow Y$, tem-se que $sup(X \Rightarrow Y)$ é a probabilidade de ambos eventos X e Y ocorrerem juntos, *i.e.*, $sup(X \Rightarrow Y) = P(X \cap Y)$[1].

II.1.1 Algoritmo Apriori

O algoritmo Apriori é o pioneiro e talvez o mais compreensivo algoritmo de mineração de conjuntos de padrões frequentes [Zheng et al., 2001]. Foi proposto em 1994 por R. Agrawal e R. Srikant e tem o objetivo de produzir eficientemente as regras de associação [Agrawal et al., 1994]. Para isso, o Apriori faz uma abordagem iterativa conhecida como pesquisa de nível e se baseia em seu princípio conceitual. O princípio é chamado *propriedade do Apriori* e se trata da seguinte constatação: "todos os subconjuntos não vazios de um conjunto de itens frequentes também devem ser frequentes" [Agrawal et al., 1994]. O algoritmo Apriori também envolve a criação de regras de associação. Essa abordagem utiliza apenas um item no consequente e todos os outros (κ-1 itens, em um padrão de tamanho κ) no antecedente. A regra $X \Rightarrow Y$ é presente em o conjunto de transações desde que satisfaça a condição de confiança mínima determinada [Han et al., 2000]. A Figura II.1 apresenta o diagrama das etapas principais executadas pelo algoritmo Apriori para geração de regras de associação.

Figura II.1: Diagrama das etapas do Apriori

O Algoritmo 1 apresenta o algoritmo Apriori. Determinam-se os itens candidatos dentro do suporte a partir de uma varredura nos dados. Essa varredura é iterativa, variando para cada tamanho κ de conjunto de itens. Nota-se a chamada de duas funções dentro da função principal (*apriori_gen* e *subset*).

[1]Nos artigos clássicos de padrões frequentes utilizam $X \cup Y$ como sendo a união dos itemsets [Han et al., 2011; Gadár and Abonyi, 2019]. Nesta dissertação, a união dos itemsets se traduz na interseção dos seus respectivos eventos estatísticos [Larsen and Marx, 2005], ou seja, $P(X \cap Y)$.

Algorithm 1 Algoritmo Apriori

1: **Entrada:** Uma base de dados D e o valor do suporte mínimo (min_{sup})
2: **Saída:** O conjunto L com todos os conjuntos de itens frequentes
3: **function** $apriori(D, min_{sup})$
4: **for** $\kappa = 2; L_{k-1} \neq \emptyset; k++$ **do**
5: $C_k = apriori\_gen(L_{k-1})$
6: **for each** $t \in D$ **do**
7: $C_t = subset(C_k, t)$;
8: **for each** $c \in C_t$ **do**
9: $c.count++t$;
10: **end for**
11: **end for**
12: $L_k = c \in C_k | c.count++ \geq min_{sup}$;
13: **end for**
14: return $L = \sqcup_k L_k$
15: **end function**

Capítulo III Conclusões

A riqueza do banco de dados estudado, principalmente após o pré-processamento, tem potencial de trazer conhecimento interessante a partir da aplicação bem empregada de mineração de padrões frequentes. As informações levantadas apontam ocorrências relevantes e coerentes sobre a malária no contexto das regiões de saúde. Para alcançar esse objetivo, a abordagem para obtenção de regras divergentes (ARD) desenvolvida neste trabalho se provou útil na descoberta de conhecimento, já que, por meio dela, foi possível levantar as informações relevantes sobre a malária que não se mostravam claras durante a análise exploratória de dados.

Referências

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

Berzal, F., Blanco, I., Vila, M., and others. Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6(3):221–235, 2002.

Gadár, L. and Abonyi, J. Frequent pattern mining in multidimensional organizational networks. *Scientific Reports*, 9(1), 2019.

Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Haryana, India; Burlington, MA, 3 edition, 2011.

Han, J., Pei, J., and Yin, Y. Mining Frequent Patterns Without Candidate Generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 1–12, New York, NY, USA. ACM, 2000.

Larsen, R. J. and Marx, M. L. *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, Upper Saddle River, N.J, 4 edition, 2005.

Lodhi, K. Survey on frequent pattern mining. *International Journal of Engineering, Science and Mathematics*, 2(3):64, 2013.

Zheng, Z., Kohavi, R., and Mason, L. Real world performance of association rule algorithms. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 401–406. ACM, 2001.