Encontrando Regras de Associação sem Especificar Suporte Mínimo e Confiança Mínima

Oto Antonio Lopes Cunha FIlho[1], **João Batista Rocha-Junior**[1]

[1]Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Feira de Santana (UEFS) - Feira de Santana - Bahia - Brasil

otoengc@gmail.com, joao@uefs.br

Abstract. *The extraction of information and knowledge in databases has been assuming a relevant role in aiding decision making. One of the main areas of research is association rule mining. This area makes possible to capture the relationships among the attributes present in a database. Most algorithms used to extract association rules use support and confidence as parameters. Support represents the proportion of a given rule in the database and confidence represents the validity of this rule. Thus, professionals responsible for data analysis need to identify and define support and confidence thresholds (minimum support and minimum confidence, respectively) to obtain association rules. However, in certain contexts, it is difficult to identify good values for support and confidence in order to obtain the desired rules. In these situations, it may be necessary to run several queries with different values of support and confidence until the desired rules are obtained. The purpose of this research is to examine association rules mining techniques and algorithms capable of obtaining association rules without the need of specifing support and confidence, to propose new algorithms and analyze these algorithms in terms of performance and quality of the rules obtained.*

Informações Gerais

Nível: Mestrado;
Orientador: João Batista Rocha-Junior;
Universidade e programa de pós-graduação: Universidade Estadual de Feira de Santana (UEFS). Programa de pós-graduação em ciência da computação (PGCC);
Mês e ano de ingresso no programa: agosto de 2020;
Mês e ano previstos para término no programa: dezembro de 2022;
Etapas concluídas e datas de conclusão:

- Disciplina: Metodologia da Pesquisa I (dezembro 2020);
- Disciplina: Análise e Projeto de Algoritmos (dezembro 2020);
- Disciplina: Engenharia de Software (dezembro 2020);
- Disciplina: Inteligência artificial (dezembro 2020);
- Proposta de Dissertação (julho 2021);
- Disciplina: Metodologia da Pesquisa II (julho 2021);
- Disciplina: Engenharia de Software Experimental (julho 2020);
- Disciplina: Redes e Sistemas de Computação (julho 2021);
- Exame de Qualificação (abril 2022).

Etapas em andamento:

- Pesquisa Orientada

1. Introdução

Muitos dados são produzidos diariamente na sociedade, que por sua vez não pode se limitar a produção destes, mas utilizá-los em prol da economia e do povo [Santos 2017]. Diversas empresas veem estes dados como privilégios legais, estratégicos e imprescindíveis para maior autonomia em suas ações [Galvão and Marin 2009]. Todavia, devido ao fluxo e forma como as informações são criadas e armazenadas, o ser humano é naturalmente incapaz de explorar, extrair e interpretar este contingente de informação para gerar um conhecimento válido.

Introduzida por [Agrawal et al. 1994], a mineração de conjuntos frequentes de itens é um subcampo da mineração de dados, que consiste em extrair eventos, padrões ou itens de ocorrência frequente nos dados. A análise desses padrões e eventos oferecem benefícios significativos nos processos de tomada de decisão [Fournier-Viger et al. 2017].

As medidas de interesse mais utilizadas na mineração de conjuntos frequentes e suas regras de associação são suporte e confiança. Suporte indica quão frequente um conjunto de itens aparece no conjunto dos dados [Agrawal et al. 1993]. Confiança representa a "força" de uma regra de associação, ou seja, refere-se à correlação entre os dois conjuntos de itens que constituem uma regra (antecedente e consequente) [Agrawal et al. 1993].

A maioria dos algoritmos desenvolvidos para extrair regras de associação recebe como parâmetro valores mínimos de suporte e confiança e retornam todas as regras que possuem suporte e confiança maiores ou iguais aos valores mínimos passados como parâmetro. Levando em consideração que a maioria dos dados são dinâmicos (seu conteúdo e relações se alteram com o tempo), esta passagem de parâmetros exige um estudo prévio do conjunto de dados para cada momento em que a análise for realizada. Desta forma, é difícil selecionar bons valores de suporte e confiança para se obter as regras desejadas. Por exemplo, os valores de suporte e confiança que retornam poucas regras em uma base, podem retornar milhares de regras em outra. Portanto, extrair informações utilizando esses parâmetros é custoso e demorado, visto que várias consultas precisam ser realizadas até obter as regras pretendidas [Moslehi and Haeri 2020].

Na tentativa de reduzir alguns destes obstáculos diversas pesquisas com foco em mineração de dados foram surgindo. Neste contexto, este trabalho tem como objetivo examinar algumas das técnicas e algoritmos de mineração de regras de associação capazes de obter regras de associação sem a necessidade de especificar, previamente, suporte e confiança mínima. Além disso, esta pesquisa apresenta novos algoritmos e análises destes em termos de performance e qualidade das regras obtidas (o acompanhamento da qualidade ocorre utilizando uma função de pontuação sobre a regra de associação).

Com as organizações cada vez mais competitivas, faz-se necessário algoritmos que auxiliem na tomada de decisão, entregando bons resultados no menor tempo possível. Algoritmos de mineração de regras de associação fazem parte deste processo. Estes são capazes de extrair relações presentes numa base de dados. Por exemplo, é possível adquirir relações entre doenças, pacientes e remédios. Também é possível utilizá-los para identificar vínculos de itens comprados num supermercado [Devi 2012] ou até mesmo utilizá-los em sistemas de recomendação de produtos ou serviços online.

No entanto, quanto mais informações nas bases de dados, mais demorado fica o processo de configuração e execução destes algoritmos. Em termos de análise combi-

natória, é possível afirmar que para um banco de dados com apenas 10 itens distintos podem ser geradas mais de 50 mil regras. Desta forma, os profissionais responsáveis pelas análises dos dados precisam gastar tempo tentando encontrar os parâmetros adequados daquele contexto, para então começarem a trabalhar sobre as informações adquiridas.

Nessa conjuntura, faz-se necessário estudos a respeito de melhorias e/ou novos modelos de algoritmos. Os resultados desta pesquisa podem contribuir para criação de mecanismos que permitam simplificar o processo de mineração de conjuntos frequentes e itens e regras de associação. Assim, seu potencial é voltado tanto para facilitar a seleção das regras de interesse, independente do conjunto de dados em questão, quanto para a qualidade das regras adquiridas num tempo hábil.

Com este estudo será validado nossa hipótese. Está consiste em dizer que com a solução proposta é possível localizar k regras válidas ao buscador (conforme uma função de aptidão) para qualquer conjunto de dados que possua pelo menos k regras.

Este trabalho está estruturado da seguinte forma: a Seção 1 contém a definição do problema, bem como a justificativa, objetivos, motivação e relevância deste estudo em questão; na Seção 2 são apresentados os trabalhos relacionados com as suas respectivas lacunas; a Seção 4 contém uma visão geral de como o estudo é conduzido; a Seção 3 contém a solução proposta; Na Seção 5 são apresentados os resultados parciais da pesquisa; a Seção 6 contém as considerações finais deste artigo e por fim as referências.

2. Principal Diferencial

Nesta seção são apresentados, numa visão geral, alguns trabalhos que contém elementos semelhantes à proposta apresentada nesta pesquisa. Estes trabalhos são apontados na Tabela 1 juntamente com as lacunas que a nossa proposta tenta sanar e/ou mitigar.

Tabela 1. Resumo dos trabalhos relacionados.

Referência	Método	Limitações
[Qodmanan et al. 2011]	Algoritmo genético.	Gera regras redundantes e inválidas.
[Moslehi and Haeri 2020]	Algoritmo genético para regras quantitativas.	Gera regras redundantes e inválidas.
[Huang and Kao 2004]	Lógica Fuzzy com derivação de suporte e confiança.	Gera regras inválidas e necessidade de análise semântica.
[Djenouri and Comuzzi 2017]	Algoritmos genéticos e enxames de partículas.	Necessita do parâmetro de suporte mínimo.
[Nguyen et al. 2018]	Top-k com duas etapas de filtragem dos itens candidatos.	Gera regras redundantes e necessita do limiar de confiança.
[Zaki et al. 1997]	Top-k com amostragem progressiva.	Gera regras falsas.
[Riondato and Upfal 2015]	Amostragem aleatória.	Gera regras falsas e alta complexidade na condição de parada.
[Kameya and Sato 2012]	Combina Top-k e FP-Growth com busca *branch-and-bound*.	Exclusivo para padrões discriminativos.

3. Algoritmos Desenvolvidos

Este capítulo contém os conceitos dos algoritmos e estratégias desenvolvidas durante a pesquisa. Dentre estes, temos o *Baseline*, um algoritmo Genético que atua como base para observação do comportamento dos dados sobre aspectos da inteligência computacional (Seção 3.1). E a estratégia/solução proposta pelos autores desta pesquisa (Seção 3.2).

Todos os algoritmos seguintes utilizam uma função fitness (ou de aptidão), Equação 1, que é responsável por atribuir uma pontuação (score) para cada regra levando em consideração as medidas de suporte e confiança da mesma.

$$f(X \Rightarrow Y) = \frac{(1 + Sup(X \cup Y))^2}{1 + Sup(X)} \tag{1}$$

[Qodmanan et al. 2011] indica que a Equação 1 vai de $0,5$ à 2 e afirma que esta função cresce conforme o crescimento do suporte e/ou confiança, ou seja, uma função que relaciona positivamente ambas as medidas de interesse.

3.1. *Baseline*

O algoritmo *Baseline* utilizado neste estudo é uma replica do algoritmo Genético proposto por [Qodmanan et al. 2011]. Seguindo a estratégia de [Qodmanan et al. 2011], este algoritmo possui as etapas de codificação das regras em cromossomos, inicialização da população, seleção, cruzamento, mutação e decisão. Além disso este algoritmo necessita de alguns parâmetros probabilísticos com valores de 0 à 1, são eles: probabilidade da seleção - *sp* (*selection probability*); probabilidade de cruzamento - *cp* (*crossover probability*); e probabilidade de mutação - *mp* (*mutation probability*). Estes parâmetros são utilizados nas etapas de seleção, cruzamento e mutação respectivamente.

O resultado deste algoritmo é a "melhor" população. Esta população resultante é encontrada ao selecionar, dentro das gerações (laço de repetição do algoritmo), a população que tem a maior média e menor desvio padrão dos scores de seus indivíduos (cromossomos). Tais scores são definidos utilizando a Equação 1.

3.2. Amostragem (Iterativo)

Ao utilizar técnicas de amostragem probabilística na lógica de algoritmos de mineração de conjuntos frequentes de itens e suas regras de associação é possível reduzir o espaço de busca (trabalhando sobre as amostras) e em seguida realizar inferências sobre o conjunto total de itens (população). Neste contexto, o algoritmo Iterativo (In) tem como objetivo encontrar as Top-k regras de associação ao reduzir seu campo de busca através da amostragem e realizar uma inferência a respeito do suporte mínimo que deve ser utilizado na busca operando sobre toda base de dados. Este algoritmo consiste em duas etapas: a primeira etapa chamada de amostragem e a segunda de busca global.

A primeira etapa deste algoritmo tem uma variação utilizando o Skyline (Is) das regras da amostra para extração da regra resultante. De maneira geral, a primeira etapa utiliza a Equação 2 para fornecer um valor n que representará o tamanho da amostra. As variáveis N, Z, p, q e E representam, respectivamente, o tamanho da base de dados, o valor que especifica o nível de confiança da amostra, a proporção da população que tem

uma determinada característica, a proporção da população que não tem uma determinada característica e a precisão das proporções amostrais.

$$n = \frac{N \cdot Z^2 \cdot p \cdot q}{E^2 \cdot (N-1) + Z^2 \cdot p \cdot q} \tag{2}$$

A estratégia base para a primeira etapa do algoritmo (amostragem) consiste em localizar todos os conjuntos frequentes de itens sem a utilização de nenhum limiar e em seguida gerar todas as regras possíveis para então ordená-las pelo valor do score e então selecionar a regra que seja um ponto representativo [Vlachou et al. 2022]. Na etapa de busca global, ou segunda etapa, a regra selecionada pela primeira etapa é mapeada na base de dados completa a fim de se obter seu valor "real" de suporte. E então, este valor de suporte é utilizado como suporte mínimo para obtenção dos conjuntos frequentes e em seguida obtém-se as regras. Assim o resultado deste algoritmo ocorre ao aplicar o método Top-k sobre estas regras de associação.

4. Metodologia

Os procedimentos seguidos para este trabalho se iniciam com a etapa de coleta e construção de base de dados representa a aquisição das bases de dados reais e construção das bases de dados sintéticas. As bases de dados reais servem para avaliar o comportamento da solução proposta em um cenário real, enquanto que as bases de dados sintéticas servem para avaliar o comportamento da solução proposta em cenários hipotéticos, garantindo maior confiabilidade aos resultados obtidos. As bases de dados reais requerem uma etapa de pré-processamento, que serve para retirar registros indesejados, bem como homogeneizar o formato dos dados para que possam ser processados.

Paralelamente ao processo de coleta do banco de dados, tem-se a etapa de propor solução. Nessa fase, algoritmos são planejados e propostos a fim de sanar ou mitigar alguns dos desafios abordados nesta pesquisa. Com a proposta de solução definida, tem-se início a fase de desenvolvimento. Nesta fase são implementados algoritmos base e avançados. Os algoritmos base (*Baseline*) são desenvolvidos a partir do estado da arte e são utilizados para servir de comparação com os algoritmos mais avançados.

Por fim, são realizados experimentos controlados usando os artefatos adquiridos até então. Na fase de avaliação experimental, tanto o *Baseline* quanto os algoritmos propostos são submetidos a casos de testes, criados durante esta etapa e relacionados a cada tipo de base de dados (real e sintética). Em seguida os resultados destes testes são analisados e avaliados com o objetivo de quantificar os ganhos obtidos com a solução proposta.

5. Resultados Parciais

Para a execução dos algoritmos foram utilizadas as seguintes bases de dados:

- `Visited`: uma base de dado da empresa Sanar Saúde[1], contendo itens visitados por clientes. Estes dados foram extraídos entre setembro de 2020 à fevereiro de 2021 e contém 27.098 transações com 942 itens distintos;
- `Purchases`: também uma base de dados fornecida pela empresa Sanar Saúde, no entanto as transações se referem aos itens comprados por cliente. Estes dados foram extraídos entre setembro de 2020 à fevereiro de 2021 e contém 1.631 transações com 315 itens distintos.

[1]https://www.sanarsaude.com/

A configuração de execução do algoritmo Baseline foi $sp = 0.95$, $cp = 0.85$, $mb = 0.01$, $ruleLen = 3$, $alfa = 0.01$, $minGen = 25$, $maxGen = 1000$, $sizePop = 10$. Onde sp, cp e mp correspondem respectivamente às taxas de seleção, cruzamento e mutação do algoritmo. O parâmetro $ruleLen$ representa o tamanho da regra (utilizado na codificação do cromossomo); $alfa$ representa o erro máximo aceitável entre o melhor e pior cromossomo de uma população. E os atributos $minGen$, $maxGen$ e $sizePop$ correspondem respectivamente ao mínimo e máximo permitido para o ciclo de gerações e o tamanho da população. Para o algoritmo Iterativo, tem-se $k = 10$.

Dada a natureza não exata dos algoritmos deste experimento, os mesmos foram executados cinco vezes em cada base de dados para fornecer os resultados das Tabelas 2 e 3. As informações presentes nestas tabelas contém os algoritmo utilizados, sendo In e Is o algoritmo Iterativo e sua variação e *Bas* o Baseline, respectivamente. Além disto, as tabelas também possuem a média das medidas de quantidade de regras válidas retornadas, suporte, confiança e score das regras, memória alocada pelas estruturas de dados dos algoritmos e o tempo de execução dos mesmos.

Tabela 2. Resultados para 5 execuções dos algoritmos na base purchases.

Alg.	Qtd. Regras	Sup	Conf	Score	Mem.	Tempo
In	100%	0.0023	0.9400	1.0019	3GB	118.20s
Is	100%	0.0033	0.6137	1.0006	2210MB	121.93s
Bas	15%	0.0002	0.1061	0.1226	3.7GB	298.00s

Tabela 3. Resultados para 5 execuções dos algoritmos na base visited.

Alg.	Qtd. Regras	Sup	Conf	Score	Mem.	Tempo
In	40%	0.0020	0.0968	0.3943	75.00GB	859.00s
Is	100%	0.0022	0.4235	0.9992	64.21GB	428.74s
Bas	7%	0.0004	0.2676	0.3472	80.05GB	552.00s

6. Considerações Finais

Esta pesquisa realizou um estudo experimental em alguns algoritmos de mineração de regras de associação. Os resultados parciais apontam que o algoritmo *Baseline* necessita de uma análise prévia das bases de dados para a extração de regras válidas. Para as bases de dados utilizadas e parâmetros "padrões", o algoritmo Baseline não foi capaz de extrair as k regras válidas na maioria dos casos de teste, apesentando baixa qualidade. Em contrapartida, os algoritmos Iterativo e sua variação permitiram encontrar regras para todas as bases de dados do experimento em todos os casos de teste.

Para melhorias do experimento, serão utilizadas mais duas bases de dados e mais um algoritmo baseado no algoritmo Apriori. Além disso, a formas de captura de memória utilizada processamento será mais refinada dentro dos algoritmos. Por fim, mais testes serão realizados realizadas nos algoritmos em questão, bem como aperfeiçoamentos no algoritmo Iterativo e suas variações. Tais aperfeiçoamentos dizem respeito à redução da quantidade de recursos utilizados no algoritmo e à função de amostragem.

Agradecimentos

Gostaria de agradecer a empresa **Sanar Saúde** pelo fornecimento de dados fundamentais para a pesquisa. Também agradeço ao Programa Interno de Auxílio Financeiro aos Pro-

gramas de Pós-Graduação Stricto Sensu (**AUXPPG**) da Universidade Estadual de Feira de Santana (UEFS), ao Programa de Apoio à Pós-Graduação (**PROAP**) da CAPES e à equipe do **SBBD 2022 - WTDBD** pelo apoio e incentivo à este trabalho.

Referências

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD*.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *VLDB*. Citeseer.

Devi, M. R. (2012). Applications of association rule mining in different databases. *Journal of Global Research in Computer Science*.

Djenouri, Y. and Comuzzi, M. (2017). Combining Apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem. *Information Sciences*.

Fournier-Viger, P., Lin, J. C.-W., Vo, B., Chi, T. T., Zhang, J., and Le, H. B. (2017). A survey of itemset mining. *Wiley Interdisciplinary Reviews: DMKD*.

Galvão, N. D. and Marin, H. d. F. (2009). Técnica de mineração de dados: uma revisão da literatura. *Acta Paulista de Enfermagem*.

Huang, Y.-P. and Kao, L.-J. (2004). Using fuzzy support and confidence setting to mine interesting association rules. In *AMFI*. IEEE.

Kameya, Y. and Sato, T. (2012). Rp-growth: top-k mining of relevant patterns with minimum support raising. In *SIAM*.

Moslehi, F. and Haeri, A. (2020). A genetic algorithm-based framework for mining quantitative association rules without specifying minimum support and minimum confidence. *Scientia Iranica*.

Nguyen, L. T., Vo, B., Nguyen, L. T., Fournier-Viger, P., and Selamat, A. (2018). ETARM: an efficient top-k association rule mining algorithm. *Applied Intelligence*.

Qodmanan, H. R., Nasiri, M., and Minaei-Bidgoli, B. (2011). Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. *ESWP*.

Riondato, M. and Upfal, E. (2015). Mining frequent itemsets through progressive sampling with rademacher averages. In *SIGKDD*.

Santos, M. A. S. d. (2017). Estudo comparativo de algoritmos exaustivos para mineração de padrões discriminativos em bases de dados biomédicas. Master's thesis, UFP.

Vlachou, A., Doulkeridis, C., Rocha-Junior, J. B., and Nørvåg, K. (2022). On decisive skyline queries. In *ICBDAKD*. Springer.

Zaki, M. J., Parthasarathy, S., Li, W., and Ogihara, M. (1997). Evaluation of sampling for data mining of association rules. In *International Workshop on RIDE. High Performance Database Management for Large-Scale Applications*. IEEE.