Explorando Bases de Dados Longitudinais via Análise Triádica de Conceito - Um Estudo de Caso na Percepção da COVID-19

João P. Santos[1], **Mark A. J. Song**[1], **Luis E. Zárate**[1]

[1]Instituto de Ciências Exatas e Informática
Laboratório de Inteligência Computacional Aplicada - LICAP
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Caixa Postal 1.686 — 30.535-901 -– Belo Horizonte — MG — Brasil

santosjp167@gmail.com, {song, zarate}@pucminas.br

Abstract. *Longitudinal studies and databases are commonly applied in the health area. These databases record observations from the same sample of individuals during consecutive periods of time called waves. In this work we propose to apply Triadic Concept Analysis to obtain triadic rules, which correspond to rules of association with conditions, to describe the temporal relationships existing between the waves of the study. The results show a promising strategy for describing databases of this type. As a case study, it is considered a database about the attitudes and perception of individuals in a population when facing the COVID-19 pandemic.*

Resumo. *Bases de dados longitudinais são comumente encontradas na área da saúde. Essas bases registram observações de uma mesma amostra de indivíduos durante períodos de tempo consecutivos denominados ondas. Neste trabalho propomos aplicar a Análise Triádica de Conceito para obter regras triádicas, que correspondem a regras de associação com condições, para descrever as relações temporais existentes entre as ondas do estudo. Os resultados mostram uma estratégia promissora para descrever bases de dados deste tipo. Como estudo de caso, foi utilizado uma base de dados a respeito das atitudes e percepções de indivíduos de uma população durante o enfrentamento da pandemia do COVID-19.*

1. Introdução

Estudos longitudinais na área da saúde registram observações de uma mesma amostra ao longo do tempo, permitindo acompanhar a evolução clínica ou psicológica dos indivíduos [Diggle 1994]. Essas bases são valiosas para descrever e validar procedimentos, além de apoiar a tomada de decisão por meio da extração de informações úteis sobre os pacientes ao longo de diferentes fases de tratamento ou exposição [Ribeiro et al. 2017].

Este trabalho propõe o uso da Análise Triádica de Conceitos (ATC) [Lehmann and Wille. 1995], uma extensão da Análise Formal de Conceitos (AFC) [Ganter and Wille 2012], como abordagem para analisar essas bases longitudinais. A ATC introduz um terceiro elemento, chamado de condição, que permite representar relações entre objetos e atributos em diferentes ondas de estudo, possibilitando a extração

de regras triádicas que revelam padrões entre sintomas em um ou mais momentos distintos.

Apesar do crescente uso da AFC em mineração de dados [Kim et al. 2015], a ATC ainda é pouco explorada nesse contexto da saúde. Este trabalho aplica a ATC a um estudo de caso com dados da Bélgica durante a pandemia de COVID-19, com o objetivo de identificar relações entre ações de saúde pública e a propagação do vírus. A abordagem mostra-se promissora para compreender como a relação entre ações básicas de higiene e intervenções governamentais podem interferir na contaminação e disseminação de uma pandemia como a COVID-19.

2. Fundamentação teórica

2.1. Bases de dados longitudinais

Uma base de dados longitudinal pode ser representada como uma estrutura organizada que armazena dados coletados repetidamente ao longo do tempo sobre os mesmos indivíduos, permitindo a análise da evolução de variáveis específicas. Em vez de registrar apenas um momento (como nas bases transversais), ela acompanha cada sujeito em diferentes pontos temporais, registrando eventos, medições clínicas, tratamentos, comportamentos ou qualquer outro atributo que seja interessante para a análise.

Na área da saúde, essa estrutura é particularmente útil para monitorar o progresso de doenças, avaliar respostas a intervenções médicas ou estudar o impacto de fatores ambientais e comportamentais na saúde ao longo dos anos. Essa forma de representação permite análises estatísticas e geração de *insights* interessantes que podem contribuir no diagnóstico ou ainda no tratamento de doenças.

2.2. Análise Formal de Conceitos e Análise Triádica

Análise Formal de Conceitos (AFC) é um campo da matemática aplicada, é fundamentada na teoria dos reticulados conceituais [Ganter and Wille 2012]. A AFC identifica conceitos formais a partir da relação de incidência entre objetos e atributos, presentes em um conjunto de dados.

A principal definição da AFC é o contexto formal diádico, a qual corresponde a uma tupla da forma $K := (G, M, I)$, onde G corresponde a um conjunto de objetos, M a um conjunto de atributos, e I indica uma relação de incidência ($I \subseteq \mathrm{G}\times M$), indicando quais objetos possuem determinados atributos. A relação de incidência pode ser também representada por (g,m) $\in$ I, e lê-se "o objeto g possui o atributo m", e dentro do contexto de saúde "o paciente g possui o sintoma m".

A partir do contexto formal é possível extrair regras de implicação. Desta forma, uma implicação é da forma $P \rightarrow Q$, tal que P e Q são conjuntos de atributos e $P' \subseteq Q'$. Em outras palavras, cada objeto que possui os atributos de P possui também os atributos de Q.

Para alguns problemas, como o considerado neste trabalho, é conveniente identificar regras de implicação associadas a uma condição. Na teoria da Análise Conceitual Triádica (ATC), introduzida por [Lehmann and Wille. 1995], é proposto o contexto triádico, formalmente definido como uma quádrupla $\mathcal{K} = (K1, K2, K3, Y)$, onde $K1$, $K2$ e $K3$ são conjuntos, e Y corresponde a uma relação ternária de incidência entre $K1$, $K2$

e $K3$. Em outras palavras, $Y \subseteq K1 \times K2 \times K3$, onde os elementos de $K1$, $K2$, e $K3$ são chamados objetos, atributos e condições, respectivamente. Esta relação pode ser representada por $(g, m, b) \in Y$, e pode ser lido como: "o objeto g possui o atributo m sob a condição b".

2.3. Regras de associação

De acordo com a literatura, [Biedermann 1997] foi o primeiro trabalho a lidar com o problema de extração de regras de implicação a partir de um contexto triádico. A partir desses contextos é possível extrair dois tipos de regras que serão utilizadas neste trabalho. Essas regras são denominadas *Biedermann Conditional Attribute Association Rule* (BCAAR) e *Biedermann Attributional Condition Association Rule* (BACAR) [Missaoui and Kwuida. 2011], que são formalizadas a seguir:

- **BCAAR**: $(A1 \rightarrow A2)C(sup, conf)$, onde $A1$, $A2 \subset K2$ e $C \subset K3$, ou seja, quando $A1$ ocorrer sob todas as condições C, $A2$ também ocorrerá nas mesmas condições com suporte (sup) e confiança ($conf$).

- **BACAR**: $(C1 \rightarrow C2)A(sup, conf)$, onde $C1$, $C2 \subset K3$ e $A \subset K2$, ou seja, quando $C1$ ocorrer para todos os atributos A, então a condição $C2$ também irá ocorrer nos mesmos atributos.

Neste trabalho foram utilizadas as métricas de *Suporte* e *Confiança* para avaliar as regras de associação triádicas. É importante notar que o valor do suporte não deve ser unicamente considerado para avaliar se uma implicação é mais relevante do que outra (uma interpretação comum quando é considerada a mineração de itens frequentes). A métrica de suporte mede a proporção de casos de variáveis de interesse presentes na regra de associação, para uma determinada onda, ou entre ondas. Por exemplo, um valor alto de suporte indica que a relação entre variáveis é uma observação recorrente na população do estudo. Por outro lado, quando o valor é baixo, pode indicar que a relação é uma exceção e não deve ser desconsiderada do ponto de vista clínico.

3. Trabalhos relacionados

Desde que foi formalizada, a AFC tem sido constantemente desenvolvida e empregada em diversas aplicações, como Recuperação de Informação, Análise de segurança, Mineração de textos, Detecção de tópicos, Análise de redes sociais, Mecanismos de busca, entre outras [Singh et al. 2016]. Por exemplo, uma aplicação de AFC no contexto da saúde, teve como objetivo encontrar redundâncias entre os diversos testes de exames clínicos, e propor substituições por testes mais baratos [Gupta et al. 2007].

Em geral, a AFC tem sido utilizada em análise de dados e representação de conhecimento, onde associações e dependências são identificados a partir de uma relação binária de incidência entre objetos e atributos. Diversos artigos [Wille 2001, Stumme et al. 2002] também discutiram o uso da extração de regras de associação por meio da AFC.

Considerando o melhor do nosso conhecimento, atualmente não existem efetivos ou representativos trabalhos mostrando formalmente a aplicação de regras triádicas, em especial na área da saúde. Neste trabalho, o objetivo é mostrar a potencialidade da análise triádica para analisar uma base longitudinal contendo as reações psicológicas e de

cuidados durante o enfrentamento à pandemia do COVID-19 na população da Bélgica. As regras triádicas podem servir para analisar dados longitudinais incorporando o aspecto temporal presente nesse tipo de estudo.

4. Metodologia

Nesta seção os procedimentos necessários para aplicação da análise triádica em bases de dados de estudos longitudinais são apresentados.

1) Restrições para a base de dados longitudinal: Para viabilizar a análise é necessário colocar as seguintes restrições à base de dados:

1. Os registros (indivíduos da amostra) devem ser os mesmos para todas as ondas consideradas;
2. As categorias de variáveis em cada onda devem ser as mesmas para todas as ondas consideradas;
3. As variáveis por categoria devem ser as mesmas para todas as ondas consideradas;

Após considerar as restrições apresentadas, o conjunto de dados considerado para a análise triádica deve possuir os mesmos registros, as mesmas variáveis (observações clínicas/sintomas) e não possuir dados ausentes, para todas as ondas analisadas.

2) Caracterização da análise triádica: É importante ressaltar que a análise triádica não é uma tarefa de classificação. Caso os registros considerados (indivíduos/pacientes) caracterizem populações distintas, a análise triádica deve ser aplicada sob uma das populações específicas. Assim, a análise triádica descreve por meio de regras de associação a população de uma classe de registros. Por exemplo, ao analisar registros da evolução de pacientes que foram submetidos à um tratamento específico, deve-se primeiro retirar os registros de pacientes com tratamento placebo e trabalhar apenas com a população que se deseja descrever.

Em relação às variáveis de interesse, devem ser selecionadas aquelas que representam a população em análise. Para isso, é de grande importância ter conhecimento de especialistas de domínio que possam auxiliar na escolha dessas variáveis.

3) Definindo a relação triádica: Para marcar a incidência binária da relação entre objetos e atributos (pacientes e variáveis) é importante considerar *thresholds* como valores de referência para cada variável de interesse. Valores acima de um *threshold* pode indicar condições clínicas favoráveis ou desfavoráveis do paciente em relação a uma doença.

Neste trabalho, é considerado que as relações de incidência devem ocorrer para condições favoráveis ou desfavoráveis para um estado de saúde, e não com a mistura de ambas condições, dentro de um mesmo contexto formal triádico. Essa restrição deve-se ao fato dos algoritmos de extração de regras de associação, dentro da teoria da análise formal de conceitos, não serem capazes de diferenciar as situações de saúde quando ambas condições são consideradas dentro do contexto.

4) Extração de regras de associação triádicas: Após a transformação inicial da base para um contexto triádico, convertemos esse contexto para um formato de arquivo para servir de entrada à ferramenta *LatticeMiner* [Missaoui and Emamirad 2017], para a geração das regras triádicas. Para isso, convertemos o contexto para um arquivo JSON contendo, mais explicitamente, as relações entre objetos, atributos e condições. A partir do *LatticeMiner*, são obtidas as regras BCAARs e BACARs com o suporte e confiança mínimos desejados.

5. Aplicação da análise triádica

Esta seção mostra a aplicação da análise triádica em uma base de dados longitudinal contendo as reações psicológicas e comportamentais de uma população durante o enfrentamento à pandemia COVID-19.

Materiais

A base de dados utilizada provém de uma pesquisa realizada através de entrevistas telefônicas com uma amostra de residentes da Bélgica entre os primeiros meses da pandemia do COVID-19 [De Coninck et al. 2022]. O estudo longitudinal contou com a participação de 1.646 entrevistados, e foi dividido em cinco entrevistas entre março de 2020 e abril de 2021. O estudo faz uma análise relacionando comportamentos como lavar as mãos e usar máscaras em público com a propagação da epidemia, visando auxiliar na proposta de ações de contenção e prevenção à contaminação do vírus.

Métodos

1) Pré-processamento da base de dados

Entrevistados que não participaram de todas as 5 ondas foram removidos do conjunto de dados usado neste estudo, restando 348 entrevistados. A base longitudinal não apresenta dados ausentes em relação aos 348 entrevistados que participaram em todas as cinco ondas do estudo, portanto não foi necessária nenhuma remoção ou tratamento destes dados.

As perguntas consideradas para a análise são mostradas na Tabela 1. As seguintes perguntas foram consideradas para a análise: "PVD11 - Minhas mãos parecem sujas após tocar em dinheiro", "PVD12 - É provável que eu pegue um resfriado, uma gripe ou outra doença, especialmente se estiver circulando", "PVD13 - Me sinto ansioso estando perto de pessoas doentes", "government_crisis - O entrevistado acha que o governo está lidando bem com a crise".

2) Construção do contexto triádico

As respostas das perguntas da entrevista são valores numéricos entre 1 e 7, sendo 1 "discordo completamente"e 7 "concordo completamente". Para fins de discretização, foi considerado que uma resposta de 5, 6, ou 7 representa que o indivíduo concorda com a afirmação, enquanto uma resposta de 1, 2 ou 3 representa que o indivíduo não concorda.

Na construção do contexto triádico, os entrevistados correspondem aos objetos, enquanto as perguntas da entrevista são considerados como atributos e, por fim, as ondas de entrevistas são tratadas como condições.

3) Extração das regras triádicas

Após a construção do contexto triádico e aplicação da ferramenta *Lattice Miner*, foram geradas 287 regras do tipo BACAR e 134 regras do tipo BCAAR.

A Tabela 2 apresenta as principais regras de associação do tipo BACAR geradas. Estas regras relacionam as condições (ondas de entrevistas) que possuem as mesmas respostas na entrevistas. Em outras palavras, estas regras de associação relacionam as ondas de entrevistas que possuem respostas semelhantes, ou seja, podemos analisar quantos indivíduos mantiveram as mesmas respostas com o passar do tempo.

Tabela 1. Variáveis e condições

Variável	**Descrição da variável**
PVD11	"Minhas mãos parecem sujas após tocar em dinheiro"
PVD12	"É provável que eu pegue um resfriado, uma gripe ou outra doença, especialmente se estiver circulando"
PVD13	"Me sinto ansioso estando perto de pessoas doentes"
government_crisis	O entrevistado acha que o governo está lidando bem com a crise
Condição	**Descrição da condição**
Onda1	Primeira onda de entrevistas, entre 17 e 22 de Março de 2020
Onda2	Segunda onda de entrevistas, entre 6 e 18 de Abril de 2020
Onda3	Terceira onda de entrevistas, entre 17 de Maio e 5 de Junho de 2020
Onda4	Quarta onda de entrevistas, entre 18 e 31 de Agosto de 2020
Onda5	Quinta onda de entrevistas, entre 17 de Março e 5 de Abril de 2021

Tabela 2. Regras de associação BACAR

N	**Regra**	**Suporte**	**Confiança**
1	(Onda1 → Onda2) PVD11	24,7%	64,7%
2	(Onda2 → Onda3) PVD11	27,3%	67,9%
3	(Onda3 → Onda4) PVD11	24,1%	63,2%
4	(Onda4 → Onda5) PVD11	25,3%	67,2%
5	(Onda1 → Onda2) PVD12	35,3%	69,1%
6	(Onda2 → Onda3) PVD12	31,0%	59,0%
7	(Onda3 → Onda4) PVD12	28,7%	67,1%
8	(Onda4 → Onda5) PVD12	30,5%	68,8%
9	(Onda1 → Onda2) PVD13	28,4%	76,2%
10	(Onda2 → Onda3) PVD13	30,2%	60,0%
11	(Onda3 → Onda4) PVD13	29,3%	68,9%
12	(Onda4 → Onda5) PVD13	26,1%	62,8%
13	(Onda1 → Onda2) government_crisis	50,0%	82,5%
14	(Onda2 → Onda3) government_crisis	35,6%	56,4%
15	(Onda3 → Onda4) government_crisis	18,7%	47,4%
16	(Onda4 → Onda5) government_crisis	16,1%	67,5%

A Tabela 3 apresenta as principais regras de associação do tipo BCAAR geradas. Estas regras relacionam os atributos (respostas das entrevistas) que estão presentes nas mesmas condições (ondas de entrevistas). Neste caso, estas regras de associação relacionam as respostas de diferentes perguntas dentro das mesmas ondas de entrevistas.

Tabela 3. Regras de associação BCAAR

N	**Regra**	**Suporte**	**Confiança**
1	(PVD12 → PVD13) Onda1	23,3%	45,5%
2	(PVD12 → PVD13) Onda2	31,6%	60,1%
3	(PVD12 → PVD13) Onda3	26,7%	62,4%
4	(PVD12 → PVD13) Onda4	24,7%	55,8%
5	(PVD12 → PVD13) Onda5	24,1%	52,8%
6	(PVD12 → government_crisis) Onda1	29,3%	57,3%
7	(PVD12 → government_crisis) Onda2	32,2%	61,2%
8	(PVD12 → government_crisis) Onda3	18,1%	42,3%
9	(PVD12 → government_crisis) Onda4	11,5%	26,0%
10	(PVD12 → government_crisis) Onda5	15,8%	34,6%
11	(PVD12, PVD13 → government_crisis) Onda1	13,8%	59,3%
12	(PVD12, PVD13 → government_crisis) Onda2	19,3%	60,9%
13	(PVD12, PVD13 → government_crisis) Onda3	11,5%	43,0%
14	(PVD12, PVD13 → government_crisis) Onda4	6,0%	24,4%
15	(PVD12, PVD13 → government_crisis) Onda5	8,9%	36,9%

Análise dos resultados

Analisando as regras BACAR, podemos acompanhar o comportamento de um atributo através das ondas. Neste caso de estudo, podemos observar como as respostas a uma pergunta da entrevista mudaram ou se mantiveram ao longo das ondas de entrevistas realizadas. Por exemplo, as regras 13, 14, 15, 16 mostram a evolução da variável "*govern-*

ment_crisis" desde a primeira, até a última onda. Através dos valores de confiança da regra 13, é possível perceber que 82,5% dos entrevistados que acreditavam que o governo estava lidando bem com a crise na primeira onda, continuaram acreditando na segunda onda. Este valor diminuiu para 56,4% entre a segunda e terceira onda, diminuiu novamente para 47,4% entre a terceira e quarta onda, e aumentou para 67,5% entre a quarta e a última onda. Estes valores podem ser interessantes por refletirem a opinião da população em relação às medidas governamentais a respeito da pandemia.

Já as regras BCAAR mostram a relação entre diferentes atributos em um mesmo conjunto de ondas. Neste caso de estudo, podemos observar como os indivíduos entrevistados concordaram com diferentes afirmações na mesma onda de entrevistas. Por exemplo, as regras 1, 2, 3, 4 e 5 mostram a relação entre as variáveis "PVD12"e "PVD13"ao longo das ondas de entrevistas. Através dos valores de suporte, é possivel perceber que 23,3% dos entrevistados na primeira onda se sentiam suscetíveis a doenças e também se sentiam ansiosos estando perto de pessoas doentes. Este valor aumentou para 31,6% na segunda onda, diminuiu para 26,7% na terceira, diminuiu novamente para 24,7% na quarta e, finalmente, terminou em 24,1% na última onda. É interessante notar que, nestas regras, o valor de suporte aumentou na segunda onda de entrevistas. Isso pode ser um indicativo de que na época em que a segunda onda de entrevistas foi realizada, havia uma maior preocupação em relação à pandemia, de maneira geral.

6. Conclusões, contribuições e trabalhos futuros

Este trabalho demonstrou o potencial da análise triádica na descrição de estudos longitudinais, especialmente no contexto da percepção do cidadão durante uma pandemia. Mesmo com um conjunto reduzido de dados, a técnica mostrou-se promissora ao auxiliar na tomada de decisões e formulação de políticas públicas, destacando sua aplicabilidade em bases maiores, maior precisão na caracterização dos pacientes e possibilidade de previsão de desfechos futuros de tratamentos.

Em bases de dados na área de saúde, as regras triádicas oferecem *insights* valiosos sobre a evolução clínica dos pacientes, permitindo identificar associações entre sintomas e padrões de risco ao longo do tempo. Esses insights não apenas ampliam o entendimento sobre o comportamento dos dados, mas também fortalecem o apoio à tomada de decisões clínicas, viabilizando intervenções mais rápidas e eficazes que aumentam as chances de sucesso nos tratamentos.

Os resultados obtidos sugerem direções importantes para pesquisas futuras, como a experimentação com diferentes valores de *threshold* para as variáveis de interesse. A participação de especialistas da área de estudo é essencial, tanto na escolha das variáveis quanto na definição das regras mais relevantes, reforçando o caráter multidisciplinar necessário à análise triádica.

Além disso, estudos futuros poderiam almejar explorar a análise triádica sendo aplicada a outros tipos de domínios, como na nutrição (monitoramento de indivíduos durante uma dieta ou treinamento), no controle epidemiológico (acompanhamento de indivíduos ao longo do tempo após a vacinação), entre outros. Acredita-se que a análise triádica pode ser aplicada a qualquer base de dados longitudinal que atenda aos critérios descritos na metodologia.

Agradecimentos

Os autores agradecem o apoio recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Processo No 303133/2021-0, e da Fundação de Ampara à Pesquisa do Estado de Minas Gerais (FAPEMIG), Processo PCE-00349-25.

Referências

Biedermann, K. (1997). How triadic diagrams represent conceptual structures. In Lukose, D., Delugach, H., Keeler, M., Searle, L., and Sowa, J., editors, Conceptual Structures: Fulfilling Peirce's Dream, pages 304–317, Berlin, Heidelberg. Springer Berlin Heidelberg.

De Coninck, D., d'Haenens, L., Molenberghs, G., Declercq, A., Delecluse, C., Van Roie, E., and Matthijs, K. (2022). Updating 'perceptions and opinions on the covid-19 pandemic in flanders, belgium' with data of two additional waves of a longitudinal study. Data in Brief, 42:108010.

Diggle, P. J. (1994). Analysis of longitudinal data. Technometrics, 45:181 – 181.

Ganter, B. and Wille, R. (2012). Formal concept analysis: mathematical foundations. Springer Science & Business Media.

Gupta, A., Kumar, N., and Bhatnagar, V. (2007). Analysis of Medical Data using Data Mining and Formal Concept Analysis.

Kim, E.-H., Kim, H.-G., Hwang, S.-H., and Lee, S.-I. (2015). Farm: An fca-based association rule miner. Knowledge-Based Systems, 85:277–297.

Lehmann, F. and Wille., R. (1995). A triadic approach to formal concept analysis. conceptual structures: ap-plications, implementation and theory. Springer.

Missaoui, R. and Emamirad, K. (2017). Lattice miner-a formal concept analysis tool. In 14th International Conference on Formal Concept Analysis, page 91.

Missaoui, R. and Kwuida., L. (2011). Mining triadic associa-tion rules from ternary relations. In Inter. Conf. on Formal Concept, Springer, pages 204–218.

Ribeiro, C. E., Brito, L. H. S., Nobre, C. N., Freitas, A. A., and Zárate, L. E. (2017). A revision and analysis of the comprehensiveness of the main longitudinal studies of human aging for data mining research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(3):e1202.

Singh, P. K., Kumar, C. A., and Gani, A. (2016). A comprehensive survey on formal concept analysis, its research trends and applications. 26(2):495–516.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., and Lakhal, L. (2002). Computing iceberg concept lattices with titanic. Data Knowledge Engineering, page 189–222.

Wille, R. (2001). Why can concept lattices support knowledge discovery in databases? Proceedings of the concept lattices based knowledge discovery in databases workshop, pages 7–20.