Relevância do Entendimento do Domínio de Problema na Construção de Modelos Computacionais de Aprendizado

Ligia F. de Carvalho Gonçalves[1], Daniel Rocha Franca[2], Luis Enrique Zárate[3]

[123]Departamento de Ciência da Computação, Ciência de Dados – Pontifícia Universidade Católica de Minas Gerais (PUC Minas), MG – Brazil

{ligiacarv.goncalves@gmail.com, danielfrancamdt@gmail.com,zarate@pucminas.br}

Abstract. *The objective of this work is to confirm the relevance of prior understanding of the problem domain for data science projects, specifically for building learning models. As case studies we will consider three problem domains in the health area, and as the main source of data, we will consider the recent National Health Survey, PNS 2019 prepared by IBGE. The experiments show that prior understanding of the problem domain, and its representation through conceptual models, are useful for applying a conceptual attribute selection process in the search for more assertive learning models.*

Resumo. *O objetivo deste trabalho é confirmar a relevância do entendimento prévio do domínio de problema para projetos em ciência de dados, especificamente para construção de modelos de aprendizado. Como estudos de caso consideraremos três domínios de problemas na área da saúde, e como fonte principal de dados, consideraremos a recente Pesquisa Nacional em Saúde, PNS 2019 elaborada pelo IBGE. Os experimentos mostram que o entendimento prévio do domínio de problema, e sua representação por meio de modelos conceituais, são úteis para aplicação de um processo de seleção conceitual de atributos na busca por modelos de aprendizado mais assertivos.*

1. Introdução

A Ciência de dados é uma área interdisciplinar de soluções para os mais diversos contextos e domínios de problemas. A principal tarefa é inferir conhecimento a partir de fatos e evidências expressos a partir de dados. Para isso, deve-se pressupor a identificação dos principais atributos envolvidos no domínio, e a disponibilidade de dados representativos para construção dos modelos computacionais de aprendizado.

Tipicamente, os modelos de aprendizado procuram por conhecimento aplicando um processo de descoberta de conhecimento em bases de dados. Esse processo considera seis etapas: seleção, pré-processamento, transformação, mineração de dados, validação e interpretação. Em relação à etapa de seleção, o entendimento acerca do domínio de problema para obter modelos mais representativos é o alvo deste trabalho.

Na prática de mercado, é possível constatar que responsáveis pela construção de modelos de aprendizado, muitas vezes se apresam por encontrar padrões e construir modelos por meio da aplicação de ferramentas disponíveis no mercado. Embora, um

processo de descoberta de conhecimento, executado cuidadosa e criteriosamente requer sempre um tempo maior. Na prática, não é gasto tempo suficiente na etapa de entendimento do domínio de problema, e por consequência, o conhecimento extraído a partir do modelo de aprendizado pode não estar correto, ser óbvio ou não ser relevante. Em [Guyon et al. 2018], foi realizada uma análise das soluções apresentadas pelos participantes durante desafios em AutoML (NIPS 2015, ICML 2016, PAKDD 2018). Os autores identificaram que o pré-processamento não foi alvo dos participantes. Segundo as análises, os participantes mais bem colocados, não aplicaram um processo para seleção de atributos, e 2/3 dos participantes ignoraram atributos irrelevantes.

A construção de modelos de aprendizado, deveria começar pela aquisição de conhecimento e entendimento do domínio, e depois pela obtenção das fontes de dados que o representem. Infelizmente essa prática é negligenciada e os autores atualmente não tem discutido efetivamente esse problema. O objetivo seria identificar atributos que podem compor ou enriquecer o conjunto de dados utilizado para construção de modelos de aprendizado, contribuindo para que o conhecimento extraído seja relevante.

Nos trabalhos de [Ribeiro e Zárate 2019], e [Araújo et. al. 2022] os autores têm incorporado como parte das suas metodologias de mineração de dados, uma etapa inicial de entendimento de domínio do problema com participação de especialistas de domínio. Os autores têm proposto modelos conceituais com o objetivo de utilizá-los como um guia para selecionar conceitualmente atributos relacionados ao domínio de estudo. Por exemplo em Ribeiro e Zárate (2019) e Araújo et. al. (2022), os autores consideraram fontes de dados de alta dimensionalidade, chegando a 3.500 e 10.000 atributos respectivamente. A manipulação dessa quantidade de dados poderia tornar-se inviável computacionalmente, o que demandou um processo de seleção conceitual de atributos. De acordo com os autores, a etapa de entendimento foi relevante para aumentar a representatividade e assertividade dos modelos propostos.

Recentemente em [Zárate et al. 2023] foi proposto o método CAPTO para captura do conhecimento tácito, experiência do especialista de domínio. O método é baseado em modelos de gestão do conhecimento, e junto com o conhecimento explícito disponível/adquirido, propõe uma estratégia para construção de modelos conceituais para representação de domínios de problemas. Os modelos são constituídos a partir da identificação de dimensões (perspectivas), aspectos e atributos que podem ser relevantes ao domínio. O método descrito consiste em cinco etapas primordiais: Socialização mapeamento, Combinação, Focalização e Congruência que são principalmente focadas nas discussões acerca do tema para compreender visões divergentes a fim de identificar novas perspectivas. Ademais outro ponto relevante é a consulta com o especialista do domínio para a verificação da veracidade das informações que foram adicionadas ao modelo conceitual.

De acordo com os autores, o método CAPTO pode ser utilizado como uma etapa inicial para seleção conceitual de atributos. Este procedimento pode diminuir o tempo gasto no processo de descoberta, tipicamente incremental, aplicado na busca por conhecimento em bases de dados. Além disso, o entendimento formal acerca de um domínio de problema pode auxiliar na construção de "Data-Lakes" orientados a domínio.

Frente ao exposto, o objetivo deste trabalho é confirmar a relevância do entendimento prévio do domínio de problema (utilizando o método CAPTO) para projetos em ciência de dados, especificamente para construção de modelos de

aprendizado. Como estudos de caso serão considerados três domínios de problemas na área da saúde. Como fonte principal de dados, consideraremos a recente Pesquisa Nacional em Saúde, PNS 2019 elaborada pelo IBGE.

Sendo o objetivo mostrar a relevância do entendimento prévio a aplicação de algoritmos de aprendizado de máquina, os modelos de aprendizado serão construídos considerando a base de dados original e após aplicado um processo de seleção conceitual de atributos. Tarefas de preparação e pré-processamento dos conjuntos de dados, que podem melhorar o desempenho dos modelos, não são aplicados. Esta estratégia se faz necessária para diminuir os graus de liberdade para comparação dos resultados. Os experimentos mostram que o entendimento prévio do domínio de problema, e sua representação por meio de modelos conceituais, são úteis para aplicação de um processo de seleção conceitual de atributos na busca por modelos de aprendizado mais assertivos.

2 Trabalhos Relacionados

O trabalho de S.Brandy et al, publicado em 2020, aborda sobre o desenvolvimento de modelos conceituais como guia para pesquisas, práticas e políticas na saúde pública. Os autores sumarizam a criação de um modelo conceitual em três etapas a identificação de recursos relacionados ao conceito geral, a consideração de risco e fatores protetivos e a seleção dos fatores que serão inclusos no modelo conceitual. Ademais, reafirmam a relevância que entender o problema em questão tem para as áreas trabalhadas dentro dos campos especificados em seu trabalho.

No artigo de B.C. Sally et al. (2018), os autores identificam três principais evidências que indicam a utilização da modelagem conceitual em modelos híbridos. Primeiramente, destacam a importância da discussão detalhada da situação-problema e dos objetivos do modelo onde a modelagem conceitual descreve os objetivos, entradas, saídas, suposições e simplificações, fornecendo uma abstração clara do sistema proposto, independentemente do software utilizado. Em segundo lugar a análise dos submodelos que compõem o modelo híbrido é fundamental. Esses submodelos, como aqueles de Simulação Discreta de Eventos (DES) e Dinâmica de Sistemas (SD), podem afetar de diferentes formas seja de maneira sequencial, dinâmica ou integrada. A classificação dos tipos de hibridização é usada para descrever essas interações e a estrutura geral do modelo híbrido. Por fim, a terceira evidência envolve a representação conceitual do modelo utilizando notação gráfica. Essa representação é crucial para visualizar e documentar as interações entre os componentes do modelo, assegurando que a estrutura do modelo seja bem compreendida.

3. Metodologia

3.1. Materiais para os experimentos

Neste trabalho utiliza-se como fonte de dados a Pesquisa Nacional de Saúde (PNS), pelo IBGE em parceria com o Ministério da Saúde no ano de 2019. A pesquisa analisa a percepção do estado de saúde, estilo de vida, doenças crônicas e saúde bucal da população brasileira. A PNS é uma base para o retrato da saúde da população e para propostas de políticas públicas implementadas pelo Estado Brasileiro. A base de dados

original da PNS-2019 possui 1.088 atributos organizados em 26 módulos, e 293.726 registros devidamente anonimizados. A pesquisa foi aprovada pela Comissão Nacional de Ética em Pesquisa – CONEP pelo parecer Parecer: 3.529.376 de Agosto de 2019.

Dentro do contexto da PNS 2019, foi realizado um corte para as doenças crônicas <Hipertensão (HAS)>, <Doença Pulmonar Obstrutiva Crônica (DPOC)> e <Acidente Vascular Cerebral (AVC)> contendo 23.851, 1279, e 3950 registros de indivíduos diagnosticados clinicamente com as doenças, respectivamente. De forma a balancear o conjunto de dados, foi adicionada a mesma quantidade de registros de indivíduos não diagnosticados com as doenças crônicas.

3.2 Método Experimental

1) Construção dos Modelos Conceituais:

Para a construção dos modelos conceituais, foi aplicado o método CAPTO. Neste contexto, os domínios de problema considerados foram a descrição do perfil dos indivíduos que apresentam as doenças crônicas: hipertensão, DPOC e AVC.

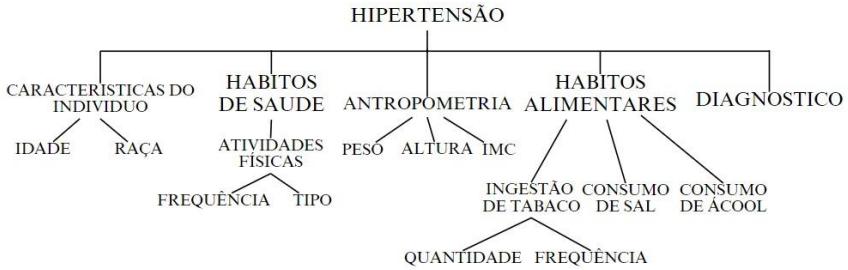

Figura 1. Modelo conceitual unificado para o domínio da Hipertensão

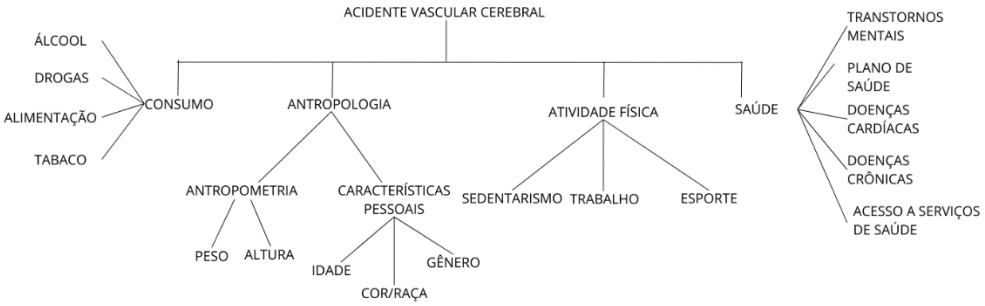

Figura 2. Modelo conceitual unificado para o domínio do AVC

Seguindo as etapas do método, foram formados três grupos de trabalho compostos por especialistas de domínio e um cientista de dados. Em sequência, através do processo de captura do conhecimento tácito/explícito, desenvolvido por meio de pesquisas e diálogos entre cada grupo, foram identificadas dimensões e aspectos relevantes de cada domínio. O primeiro resultado dos grupos de trabalho correspondeu à produção de modelos conceituais, que representam as diferentes dimensões e aspectos associados aos domínios de estudo. Para minimizar a diversidade de termos, todos os grupos consideraram o mesmo dicionário de dados, a partir das fontes de dados disponíveis

para o PNS-2019[1]. Como resultado desse processo, as Figuras 1 e 2 mostram respectivamente, os modelos conceituais para os domínios da hipertensão e AVC.

2) Seleção Conceitual de Atributos:

A partir dos modelos conceituais, foi procurado associar atributos da PNS para cada aspecto. As Tabelas 1-3, sintetizam os atributos considerados relevantes para cada domínio de acordo com a base de dados PNS 2019.

Tabela 1: Dimensões, Aspectos e Atributos vinculados ao domínio DPOC.

Dimensão: Hábitos de saúde
Tabagismo: Módulo P – Estilos de Vida: P50 e P52
Dimensão: Condições Físicas e Mentais
Doenças crônicas: Módulo Q – Doenças Crônicas: Q116a e Q74
Dimensão: Condições Socioeconômicas
Moradia: Módulo A - Informações do Domicílio: A16a
Trabalho: Módulo E – Características de trabalho das pessoas 14 anos ou mais de idade e Módulo M – Características do trabalho e apoio social: E12, M9, M11Aa, M11ag e V0026;
Dimensão: Características do indivíduo
Idade: Módulo C – Características gerais dos moradores: C8;
Dimensão: Antropometria
Peso e altura: Módulo P – Estilos de Vida – P4A

Tabela 2: Dimensões, Aspectos e Atributos vinculados ao domínio Hipertensão.

Dimensão: Hábitos de saúde
Consumo de Tabaco: Módulo P – Estilos de Vida: P50, P51, P53, P05403, P05406, P05802, P05901;
Atividade Física: Módulo P – Estilos de Vida: P35, P37, P39c, P39d, P40a, P41, P42, P43;
Dimensão: Hábitos alimentares
Dieta - Ingestão de alimentos: Módulo P – Estilos de Vida: P9a, P15, P18 e P26a
Dieta - Ingestão de bebidas: Módulo P – Estilos de Vida: P16a e P28a;
Dimensão: Condições Físicas e Mentais
Doenças crônicas: Módulo Q – Doenças Crônicas: Q30a e Q55a;
Dimensão: Características do indivíduo
Raça: Módulo C – Características gerais dos moradores: C9;
Idade: Módulo C – Características gerais dos moradores: C8;
Dimensão: Antropometria
Peso e altura: Módulo P – Estilos de Vida: P1a, P4a;

Tabela 3: Dimensões, Aspectos e Atributos vinculados ao domínio AVC

Dimensão: Consumo
Dieta - Ingestão de alimentos: Módulo P – Estilos de Vida: P6, P9a, P10a, P11a, P13, P15, P16a, P18, P19, P20a, P20b, P21a, P21b, P23, P25a, P26a e P26b;
Dieta - Ingestão de bebidas: Módulo P – Estilos de Vida: P27, P28a e P29;
Tabagismo: Módulo P – Estilos de Vida: P50, P51, P52, P53, P54a, P54b, P54c, P54d, P54e, P54f, P54g, P58 e P59;
Consumo de Drogas: Não há informação disponível na base de dados PNS
Dimensão: Saúde
Acesso a Serviços de Saúde: Módulo I – Cobertura de Plano de Saúde: I1b;
Plano de Saúde: Módulo I – Cobertura de Plano de Saúde: I1b e I6;
Saúde Mental: Módulo J - Utilização dos serviços de saúde e Módulo Q – Doenças Crônicas: Q92 e Q110a;
Doenças crônicas: Módulo Q – Doenças Crônicas: Q30a, Q60,
Doenças Cardíacas: Módulo Q – Doenças Crônicas: Q2a e Q63a;
Dimensão: Atividade Física
Sedentarismo: Módulo P – Estilo de Vida: P45a e P45b
Esporte: Módulo P – Estilo de Vida: P34, P35, P36, P37, P38, P39, P39c, P39d, P40, P40a, P41, P42, P43, P44, P44a e P44c
Trabalho: Módulo E – Características de trabalho das pessoas 14 anos ou mais de idade e Módulo M – Características do trabalho e apoio social: E11, E14a, E17, E19, M5c, M5d, M6, M7 e M8;
Dimensão: Antropologia
Características Pessoais: Módulo C – Características gerais dos moradores: C6, C7 e C8;
Antropometria: Módulo P – Estilos de Vida: P1a e P4a;

[1] https://www.ibge.gov.br/estatisticas/sociais/saude/29540-2013-pesquisa-nacional-de-saude.html

3) Composição dos Conjuntos de Dados para Experimentos:

a) Conjuntos de dados original sem seleção de atributos (DOSSA): Para todas as doenças consideradas foram utilizados os 1087 atributos em sua composição sem qualquer alteração a sua estrutura original. Desse modo, os seguintes conjuntos de dados[2] DOSSA-1, DOSSA-2 e DOSSA-3 referem-se respectivamente as doenças DPOC, Hipertensão e AVC.

b) Conjuntos de dados com seleção de atributos DCSA: Estes conjuntos de dados são resultado da aplicação do método CAPTO. Desta forma todas os atributos não citados na Tabela 1, Tabela 2, e Tabela 3 foram descartados. Os seguintes conjuntos foram construídos: DCSA-1 (DPOC), DCSA-2 (Hipertensão) e DCSA-3 (AVC) com 13, 30 e 95 atributos selecionados respectivamente.

4) Construção dos Modelos de Aprendizado

Com o objetivo de descrever o perfil dos indivíduos com as doenças DPOC, Hipertensão e AVC, foi adotada a tarefa de classificação utilizando árvores de decisão. A construção dos modelos de aprendizado foi realizada na plataforma workflow KNIME. As árvores de decisão foram parametrizadas com: Número mínimo de registros por nó = 100, Medida de qualidade = Gini index. Os conjuntos de dados foram particionados em 70% para treinamento, e 30% para teste. Adicionalmente, os mesmos conjuntos de registros foram utilizados no procedimento de teste, garantindo consistência nos resultados.

2. Experimentos e Análise dos Resultados

Para avaliar os modelos para cada domínio, foram usadas duas métricas de avaliação: precisão, e acurácia. A Tabela 4 evidencia os resultados para o teste por doença para cada classe {1: com diagnóstico para doença, 2: sem diagnóstico para doença}.

Tabela 4: Quantidade de registros por Conjunto de Dados

Domínio	Conjunto de dados	Pr-C1	Pr-C2	AC	Matriz de confusão		
DPOC	DOSSA-1	1.0	1.0	1.0		**1**	**2**
					1	391	0
					2	0	377
DPOC	DCSA-1	0.747	0.719	0.733		**1**	**2**
					1	281	110
					2	95	282
HAS	DOSSA-2	1.0	1.0	1.0		**1**	**2**
					1	7105	
					2	0	7206
HAS	DCSA-2	0.611	0.631	0.62		**1**	**2**
					1	4179	2978
					2	2443	4676
AVC	DOSSA-3	0.456	0	0.436		**1**	**2**
					1	514	54
					2	612	0
AVC	DCSA-3	0.643	0.673	0.657		**1**	**2**
					1	378	179
					2	0	368

[2] https://github.com/licapLaboratory/DataBase-and-KnimeProject-CAPTO

Para os conjuntos de dados DOSSA-1, DOSSA-2 (sem seleção conceitual), os modelos apresentaram resultados extremamente altos, para treinamento e teste, alcançando 100% de precisão e acurácia, ou seja, obtiveram resultados irrealistas e totalmente enviesados. Porém para a base DOSSA-3 o resultado do modelo foi consideravelmente baixo 43% de acurácia. Este resultado ocorre devido ao uso de forma integral da grande quantidade de atributos fornecidos pela base da PNS 2019, adicionando ruído ao modelo e tornando-o ineficaz para prever a classe alvo correspondente ao AVC. Para os modelos treinados com os conjuntos DCSA-1, DCSA-2 e DCSA-3 (com seleção conceitual de atributos) os resultados foram mais consistentes, alcançando respectivamente 73,3%, 60,9% e 65.7% de acurácia, um resultado mais realista em comparação com o teste sem a remoção de atributos dominantes e atributos consequentes ressaltando as vantagens da pré-seleção de atributos.

É importante ressaltar que em relação aos conjuntos de dados DCSA-1, DCSA-2 e DCSA-3, embora não exista melhora substancial nas métricas de avaliação, o método pode auxiliar na elaboração de modelos mais representativos e generalizáveis, após um processo de preparação de dados mais efetivo. Os modelos construídos a partir de DOSSA-1 e DOSSA-2 e DOSSA-3 evidentemente sofrem de sobreajuste e são incapazes de assimilar os atributos associados ao problema.

3. Conclusões

Os resultados alcançados pela aplicação dos modelos conceituais evidenciam o impacto que o entendimento do domínio de problema com a seleção conceitual de atributos pode ter sobre a construção de um modelo de aprendizado representativo e realista. Note que este processo, de seleção conceitual, seria uma etapa inicial de um processo de preparação de dados para descoberta de conhecimento. Esta nova etapa é denominada neste trabalho como Domain-Driven Learning Models (DDLM).

Como mencionado, o entendimento do domínio de problema como uma etapa anterior a construção e aplicação de algoritmos de machine learning é normalmente negligenciado, e isso pode levar à descoberta de padrões óbvios, não relevantes, e inclusive errôneos. É importante ressaltar que a falta de atributos para representar o modelo conceitual, é um indicativo que a descoberta de conhecimento pode ter restrições, ou que inclusive o projeto pode ser abortado. Note que todo modelo conceitual, acerca de um domínio de problema é sempre uma aproximação da realidade, daí a descoberta de conhecimento corresponderá a uma interpretação ou visão acerca do domínio de problema em estudo. Como trabalhos futuros pretende-se avaliar outras bases de dados na área da saúde, disponíveis pela PNS 2019, de forma a confirmar a nossa visão da necessidade do entendimento do domínio de problema para projetos em ciência de dados.

Agradecimentos

Os autores agradecem o apoio recebido do Fundo de Incentivo à Pesquisa (FIP) da PUC Minas por meio do Processo No 30914-1S/2024.

Referências

Guyon, I.; et. al. Analysis of the AutoMl Challenge series 2015-2018. Frank Hutter; Lars Kotthoff; Joaquin Vanschoren (eds). AutoML: Methods, Systems, Challenges, Springer Verlag, In: press, The Springer Series on Challenges in Machine Learning. 2019. doi: 10.1007/978-3-030-05318-5_10

Ribeiro, C. E.; Zárate, L. E. Classifying longevity profiles through longitudinal data mining, Expert Systems with Applications, v. 117, p. 75-89, 2019. DOI: 10.1016/j.eswa.2018.09.035

Araújo, A. S.; Silva, A. R.; Zárate, L .E. Extreme precipitation prediction based on neural network model – A case study for southeastern Brazil, Journal of Hydrology, V. 606, 127454 2022. doi: 10.1016/j.jhydrol.2022.127454.

Zarate, L., Petrocchi , B. ., Dias Maia, C. ., Felix, C., & Gomes, M. P. . (2023). CAPTO - A method for understanding problem domains for data science projects: CAPTO - Um método para entendimento de domínio de problema para projetos em ciência de dados. Concilium, 23(15), 922–941. https://doi.org/10.53660/CLM-1815-23M33.

Teece, D.J. (2013). Nonaka's Contribution to the Understanding of Knowledge Creation, Codification and Capture. In: von Krogh, G., Takeuchi, H., Kase, K., Cantón, C.G. (eds) Towards Organizational Knowledge. The Nonaka Series on Knowledge and Innovation. Palgrave Macmillan, London. https://doi.org/10.1057/9781137024961_2

Brady SS, Brubaker L, Fok CS, et al. Development of Conceptual Models to Guide Public Health Research, Practice, and Policy: Synthesizing Traditional and Contemporary Paradigms. *Health Promot Pract*. 2020;21(4):510-524. doi:10.1177/1524839919890869

Sally C. Brailsford, Tillal Eldabi, Martin Kunc, Navonil Mustafee, Andres F. Osorio, Hybrid simulation modelling in operational research: A state-of-the-art review, European Journal of Operational Research, Volume 278, Issue 3, 2019, Pages 721-737, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2018.10.025.(https://www.sciencedirect.com/science/article/pii/S0377221718308786)