Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Redes Bayesianas Aplicadas ao Diagnóstico de Covid-19

Sergio Floquet[1]
Colegiado de Engenharia Civil/Univasf, Juazeiro, BA - IMECC/Unicamp, Campinas, SP
Rodolfo C. Pacagnella[2]
FCM/Unicamp, Campinas, SP
Cristiano Torezzan[3]
IMECC/Unicamp, Limeira, SP

O procedimento de determinar o diagnostico médico de um dado paciente pode ser visto sob a ótica Bayesiana, como um processo que inicia-se com a coleta de um conjunto X de evidências, como sintomas e resultados de exames, com o intuito de estimar a probabilidade de ocorrência de um possível diagnóstico Y, ou seja, deseja-se estimar $P(Y|X)$. Sob essa ótica, trabalhos como [2] e [4] buscam encontrar semelhanças entre o raciocínio clínico e o pensamento Bayesiano. Nesse contexto, um desafio científico relevante consiste em desenvolver modelos matemáticos e computacionais que considerem o raciocínio clínico na busca de identificar padrões em bases de dados que possam auxiliar os profissionais de saúde na tomada de decisão. Essa é uma das áreas de aplicação mais proeminentes de aprendizado de máquina supervisionado.

Como alternativa a modelos complexos, ditos como *black-box*, as Redes Bayesianas (RB) têm recebido atenção na área médica pois permitem extrair informações interpretáveis dos resultados [3]. Uma RB permite a representação do problema em grafo, onde cada nó da rede é associado a um dos atributos do problema e cada aresta representa a dependência direta entre atributos [3]. Para isso, um Grafo Acíclico Dirigido (DAG) é construído, de forma que uma aresta direcionada que conecta $A \to B$ informa que B é dependente de A. Desta forma, cada nó da rede depende dos seus antecedentes e gera influência sobre os seus descendentes.

A tarefa de *treinar* uma RB pode ser dividida em duas etapas, a primeira consiste em aprender a estrutura da rede por meio de algum algoritmo ou pela experiência de um especialista na área e a segunda etapa consiste em determinar os parâmetros das distribuições do modelo.

Neste trabalho aplicamos as RBs na análise dos dados reais anonimizados de pacientes suspeitos de Covid-19, disponíveis em [1]. O conjunto contém dados de $N = 64174$ pacientes da cidade do Rio de Janeiro - RJ, coletados entre 28/04/2020 e 16/07/2020. Sintomas como Febre, Tosse, Dor de Garganta, Coriza, Mialgia, Enjoo, Diarreia, Perda de Olfato e Falta de Ar, além do Gênero e se havia Caso de Covid Confirmado em Casa foram utilizados para a construção de uma RB do tipo *Tree-Augmented Naive Bayes* (TAN), onde definimos o atributo raiz como a Covid-19, de forma que todos os outros atributos tem relação com a doença neste modelo. Na figura 1 apresentamos a rede TAN gerada por meio do pacote *bnlearn* [5], em linguagem R.

Pela disposição de árvore da rede, podemos associar os atributos a partir da organização estruturada dos nós pais, o que nos leva a perceber que os sintomas com um maior grau de importância são: {Dor de Garganta}, {Tosse} e {Febre} respectivamente. A interpretabilidade do modelo devido a disposição da rede, oriunda somente dos dados, associa uma relação de dependência para alguns sintomas: Caso Covid Confirmado em Casa → {Falta de Ar}, Mialgia → {Enjoo e Diarreia}, Febre → {Caso Covid Confirmado em Casa, Mialgia e Perda de Olfato} e Tosse → {Febre e Coriza}.

[1]sergio.floquet@univasf.edu.br
[2]rodolfop@unicamp.br
[3]torezzan@unicamp.br

© 2025 SBMAC

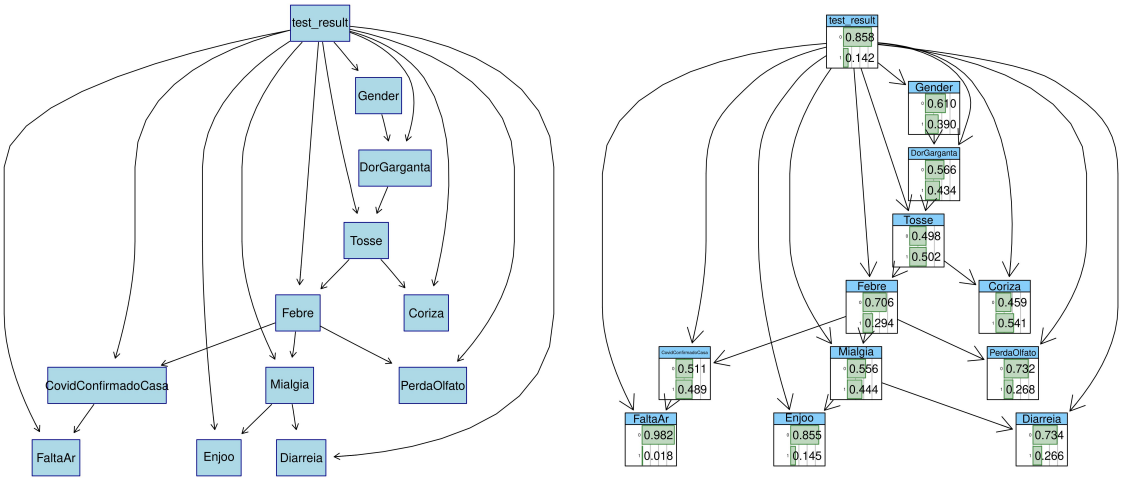

Figura 1: Resultado da *Tree-Augmented Naive Bayes* (esquerda) e as probabilidades associadas (direita). Fonte: Os autores.

Para analisar a capacidade de previsão do modelo TAN delimitamos a nossa análise a 18274 casos balanceados entre positivos e negativos, separamos 80% e 20% para treinamento e teste, respectivamente e aplicamos para 100 *seeds* distintos, tomando a média dos resultados. A título de comparação implementamos também os métodos Random Forest (RF), a Regressão Logística (RL), o Naive Bayes (NB) e eXtreme Gradient Boosting (XGB) [1]. Os resultados são apresentados na Tabela 1 e mostram que o desempenho do TAN é compatível com os demais métodos, com a vantagem de fornecer informações interpretáveis, como a relação de dependência entre os sintomas.

Tabela 1: Métricas de avaliação para classificação dos modelos RL, RF, XGB, NB e TAN.

	AUC	MCC	Acurácia	Sensibil.	Especif.	PPV	NPV	F1 Score
RL	68,42%	37,13%	68,42%	74,64%	62,20%	66,37%	71,04%	70,26%
RF	68,48%	37,10%	68,48%	72,68%	64,28%	67,05%	70,19%	69,74%
XGB	68,14%	36,36%	68,13%	71,59%	64,68%	66,96%	69,50%	69,19%
NB	66,61%	33,23%	66,60%	68,32%	64,89%	66,05%	67,20%	67,16%
TAN	67,75%	35,52%	67,75%	69,51%	65,99%	67,14%	68,41%	68,29%

Referências

[1] L. F. *et al.* Dantas. "App-based symptom tracking to optimize SARS-CoV-2 testing strategy using machine learning". Em: **PLoS One** 16.3 (2021), e0248920. DOI: 10.1371/journal.pone.0248920.

[2] C. J Gill, L. Sabin e C. H. Schmid. "Why clinicians are natural bayesians". Em: **The BMJ** 330.7499 (2005), pp. 1080–1083. DOI: 10.1136/bmj.330.7499.1080.

[3] K. B. Korb e A. E. Nicholson. **Bayesian artificial intelligence**. 2a. ed. Boca Raton: CRC press, 2010. ISBN: 9781439815915.

[4] A. M. Lipsky e R. J. Lewis. "Placing the Bayesian network approach to patient diagnosis in perspective". Em: **Annals of emergency medicine** 45.3 (2005), pp. 291–294. DOI: 10.1016/j.annemergmed.2004.10.006.

[5] M. Scutari. "Learning Bayesian Networks with the bnlearn R Package". Em: **Journal of Statistical Software** 35.3 (2010), pp. 1–22. DOI: 10.18637/jss.v035.i03.

© 2025 SBMAC