Modelos probabilísticos gráficos aplicados à identificação de doenças

Probabilistic graphic models applied to identification of diseases

Renato Cesar Sato[1], Graziela Tiemy Kajita Sato[2]

RESUMO

A tomada de decisões é um aspecto fundamental na conduta de um diagnóstico ou tratamento. A ampla difusão dos sistemas computacionais e dos bancos de dados permite sistematizar, por meio do uso da inteligência artificial, parte dessa tomada de decisão. Neste texto, é apresentada, de modo básico, a possibilidade de uso dos modelos gráficos probabilísticos como ferramenta de análise na causalidade das condições de saúde. Essa metodologia vem sendo utilizada para diagnósticos da doença de Alzheimer, apneia do sono e doenças cardiológicas.

Descritores: Modelos estatísticos; Gerenciamento clínico; Teorema de Bayes; Técnicas de apoio para a decisão

ABSTRACT

Decision-making is fundamental when making diagnosis or choosing treatment. The broad dissemination of computed systems and databases allows systematization of part of decisions through artificial intelligence. In this text, we present basic use of probabilistic graphic models as tools to analyze causality in health conditions. This method has been used to make diagnosis of Alzheimer´s disease, sleep apnea and heart diseases.

Keywords: Models, statistical; Disease management; Bayes theorem; Decision support techniques

INTRODUÇÃO

Parte das atividades realizadas nas organizações de saúde está relacionada com o processo de obtenção de informações e tomada de decisões. Colocando essa questão de outra maneira, o gerenciamento das atividades de saúde apoia-se no processo de tomar conhecimento daquilo que acredita ser verdade e como agir com base nesse conhecimento obtido. Para ilustrar essa situação, corriqueiramente as organizações de saúde vivenciam a situação em que o profissional da saúde obtém informações sobre um determinado paciente (seus sintomas, características físicas, histórico etc.) e, com base nessas informações, chega a uma determinada conclusão sobre a condição de saúde e qual a melhor conduta a ser tomada. Assim apresentamos, neste artigo, como os modelos gráficos probabilísticos (MGP), em especial as redes bayesianas, podem e são utilizadas no apoio da tomada de decisão na área da saúde.

Os MGP possuem uma ampla aplicação nas atividades relacionadas com inteligência artificial, sendo justamente nessa área que essa metodologia ganhou força e se desenvolveu nas últimas décadas. Podemos entender o MGP como um gráfico em que os nós representam as variáveis, e os arcos (direcionais ou não direcionais) representam as dependências que existem entre as variáveis. Essa estrutura permite montar o conjunto das distribuições probabilísticas, sejam elas conjuntas ou condicionais entre as variáveis.[1] Dentro dos MGP, podemos encontrar as chamadas "redes bayesianas". Essas redes surgiram por volta dos anos 1980 como modelos probabilísticos para lidar com a incerteza no contexto da inteligência artificial. No entanto, a evolução computacional e as possibilidades de aplicações fizeram com que em pouco tempo esse tema passasse a ser explorado dentro das universidades e de grandes empresas. Algumas das principais aplicações das redes bayesianas na área de saúde estão relacionadas com os sistemas de diagnósticos, modelagem de interações dos

[1] Universidade Federal de São Paulo, São Paulo, SP, Brasil.

[2] Centro Técnico Aeroespacial, São José dos Campos, SP, Brasil.

Autor correspondente: Renato Cesar Sato – Instituto de Ciência e Tecnologia, Unidade Parque Tecnológico, Universidade Federal de São Paulo – Avenida Cesare Mansueto Giulio Lattes, 1.201, sala 114 Eugênio de Mello – CEP: 12247-014 – São José dos Campos, SP, Brasil – Tel.: (12) 3921-9598 – E-mail: rcsato@gmail.com

Data de submissão: 17/6/2014 – Data de aceite: 22/2/2015

DOI: 10.1590/S1679-45082015RB3121

genes, e detecção e quantificação das influências causais no contexto epidemiológico.

Apesar dessas aplicações e vantagens obtidas pelo uso das redes bayesianas, esse tema ainda é disperso nos textos e manuais para profissionais da saúde. Isso torna o tema relativamente restrito ou apresentado de maneira superficial. As redes bayesianas oferecem uma abordagem de modelar os problemas enfrentados pelas organizações de saúde e que vêm chamando a atenção na última década. Um exemplo de aplicação é o caso da medicina personalizada, que envolve a previsão do progresso da doença com base na interpretação dos dados dos paciente por meio de um modelo de doença.[2] Neste texto, faremos, a seguir, uma apresentação sobre o funcionamento dessa metodologia no contexto das doenças, bem como suas vantagens, desvantagem, limitações e estrutura de implementação.

USO DAS REDES BAYESIANAS NA IDENTIFICAÇÃO DE DOENÇAS

Podemos entender uma rede bayesiana como um modelo no qual a causalidade é importante, porém não é raro termos um entendimento incompleto sobre o que está acontecendo e, por isso, tentarmos descrever probabilisticamente essas relações. Desse modo, o aspecto probabilístico ganha uma importante dimensão nessa rede de relações na tentativa de superar a limitação do conhecimento e auxiliar o processo de tomada de decisão.

As redes bayesianas são gráficos direcionais acíclicos (GDA), nos quais os nós são variáveis aleatórias e as premissas de independência entre as variáveis são mantidas. Os nós do gráfico representam as variáveis aleatórias da rede bayesiana e podem variar quanto à sua natureza. Isso oferece a flexibilidade de incluir dados como quantidades observadas, variáveis latentes, parâmetros desconhecidos ou até mesmo hipóteses levantadas pelo pesquisador.

Por um diagrama, podemos obter uma representação explícita sobre o que poderá acontecer em determinada situação e tentar inferir sobre as causas dos efeitos que estão sendo observados.

Tomemos como exemplo o diagrama representado pela figura 1 para identificar fatores de risco de doenças cardiovasculares como sendo tabagismo, sedentarismo e estresse. As variáveis latentes, ou seja, aquelas que não podem ser observadas diretamente, mas que possuem um papel importante nas causalidades também podem ser incluídas durante a modelagem, podendo requerer uma etapa adicional para listar esse tipo de variáveis.[3] Por exemplo, o *status* socioeconômico, compreendido pelo nível educacional, qualificações profissionais e ocupação, pode não atuar diretamente

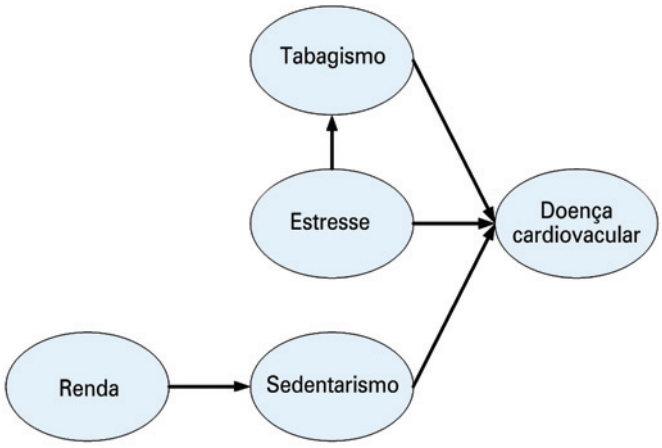

Figura 1. Premissas básicas do modelo

no desenvolvimento de doença cardiovascular, porém estudos demonstrados na Europa Oriental, Estados Unidos e Japão correlacionam menor incidência de doenças cardiovasculares na população de *status* mais elevado.[4] Desse modo, a renda de uma pessoa pode não influenciar diretamente no desenvolvimento de uma doença cardiovascular, mas fatores de estresse oriundos da condição de estresse podem promover aumento de fatores de risco, como maior consumo de bebidas alcoólicas e *status* socioeconômico (compreendidos como nível educacional, qualificações profissionais e ocupação).

Para fins de simplificação, o modelo descrito na figura 1 parte de fortes premissas. Nesse caso, estamos supondo que apenas a renda é um fator latente do sedentarismo, porém, em um modelo mais completo, temos que considerar que pode haver uma relação da renda e dos outros fatores do modelo também.

Os dados na área de saúde possuem grande quantidade de variáveis, o que aumenta a dificuldade no estudo e na avaliação dos fenômenos a eles relacionados. Outro agravante ocorre mediante a presença de situações em que uma variável é dependente de uma ou mais variáveis. Ao supormos a situação de um indivíduo acometido por uma doença cardiovascular e indicando a doença pela letra "D", para fins de modelagem, podemos associar outras variáveis, como tabagismo, sedentarismo, etilismo representados, respectivamente, pelas letras A, B e C. Graficamente podemos facilmente apresentar essa representação conforme a figura 2.

Essa representação gráfica também pode ser descrita em termos das probabilidades condicionais.

$$p(A,B,C,D)=p(A)p(B)p(C)p(D|A,B,C)$$

As ligações entre os pares de variáveis representam as dependências condicionais; os nós que não são co-

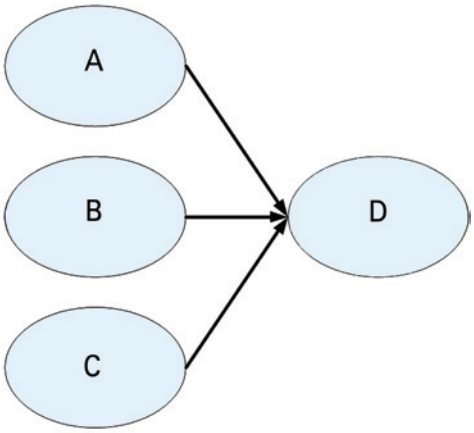

Figura 2. Estrutura condicional

nectados representam a independência condicional entre eles. Estes são termos para definir a associação da função probabilidade a partir de um conjunto de valores. Um conjunto de dados com N variáveis diz que existem 2^N modelos de redes disponíveis e que podemos utilizar o teorema de Bayes para selecionar o modelo mais adequado para os conjunto de dados apresentados. As vantagens computacionais passam a ser visíveis com o aumento no número de variáveis. Por exemplo, um conjunto com 10 variáveis possui 1.024 redes possíveis, enquanto que, para um conjunto com 15 variáveis, o número de possibilidades aumenta para 32.768. Percebe-se, então, a necessidade de aplicar uma metodologia dessa natureza para modelar os problemas de saúde, em razão da velocidade com que a complexidade da rede aumenta conforme tentamos aprimorar nosso modelos adicionando mais variáveis.

A tentativa de buscar a rede otimizada pode ser dividida três etapas.[5] Para isso, é necessário encontrar os nós-pais otimizados. Dizemos que um nó é pai quando outros nós estão associados a eles, e esses nós associados são chamados de "nós-filhos".

APLICAÇÕES NA MEDICINA

Na medicina, as redes bayesianas vêm sendo utilizadas para modelagem da incerteza.[6] Um exemplo de aplicação é o apoio a tomada de decisão quanto a um diagnóstico. O diagnóstico de doença de Alzheimer[7,8] pode ser beneficiado utilizando esse método, que consiste em classificar respostas de questionários específicos para diagnóstico. Estes, por sinal, auxiliam na construção de matrizes de julgamento e na construção de escalas de valores para cada ponto de vista fundamental, já previamente eleitos pelo clínico.[8] O modelo classifica com mais acuidade o perfil que diagnostica a doença de Alzheimer. Outros exemplos dessas aplicações podem ser a apneia do sono[9,10] e as doenças cardiovasculares.[11,12]

ESTRUTURA DE MODELAGEM DE UMA REDE BAYESIANA

O processo de modelagem de uma rede bayesiana pode ser definido em cinco etapas (Figura 3).[13]

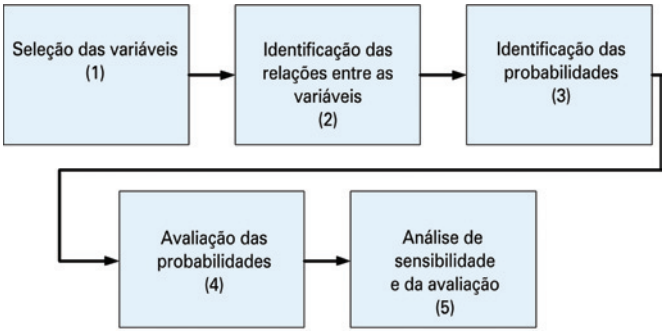

Figura 3. Estágios da modelagem de uma rede bayesiana

Seleção de variáveis relevantes

É necessário fazer um levantamento de todas as possíveis variáveis que fazem parte do problema. Para isso, costumam-se realizar entrevistas com o especialista na área do processo.

Identificação do relacionamento entre as variáveis

Após a identificação das variáveis, é necessário verificar como as mesmas se relacionam, ou seja, definir a causa e o efeito que levam uma variável a interferir em outra. Essas casualidades também estão relacionadas com o conhecimento do especialista sobre o acontecimento de determinados eventos.

Identificação das probabilidades qualitativas e restrições lógicas

Identificar o tipo de distribuição das probabilidades requeridas para a construção da rede. A restrição lógica objetiva limitar o universo de probabilidades que devem ser avaliadas. Geralmente, essa etapa consiste em mapear uma base de dados.

Avaliação das probabilidades

Nesse estágio, a distribuição de probabilidade é atribuída a cada nó da rede. Podem-se obter estimativas qualitativas e utilizar uma escala predeterminada, ou, senão, tentar visualizar a probabilidade de um evento como uma área. Em ambos os casos, esse processo de

estimação é longo e suscetível a erros, promovendo resultados que podem não ser confiáveis.[1]

Análise da sensibilidade e avaliação

Com a rede já modelada, é necessário verificar sua validade. O autor enfatiza que, na avaliação, deve-se, a partir de dados reais, submeter em outros sistemas probabilísticos para comparar os resultados.

CONCLUSÃO

A maior estruturação dos bancos de dados das instituições de saúde e demais organizações passou a permitir o aprimoramento dos modelos de causalidades das condições de saúde. Esse fenômeno abriu um importante espaço para que os sistemas de apoio à decisão nas áreas de análise diagnóstica e prognóstica, decisão sobre tratamentos e estudo das interações funcionais ganhassem espaço na interação dos conhecimentos da medicina, probabilidade e computação. O advento do chamado "*big data*" surge como uma área promissora para estender ainda mais esse tipo de análise dentro da área de saúde. As redes bayesianas surgem, então, como importante ferramenta, capaz ajudar a superar as limitações impostas pelas incertezas tão normalmente presentes na área da saúde.

Podemos considerar que as redes bayesianas são modelos gráficos em que é possível tentar obter as relações entre as variáveis. Devido à sua característica direcional, elas permitem estabelecer claramente a relação de causa-efeito e podem lidar com a incerteza a partir da teoria da probabilidade. No entanto, o uso isolado de um algoritmo não oferece a melhor estrutura diagnóstica, devendo ele ser supervisionado por profissionais especializados. Dentre as principais limitações do uso das redes bayesianas estão associadas as possíveis violações das distribuições de probabilidades, nas quais o sistema foi estruturado, e a limitação do sistema em atualizar seus objetivos diante da necessidade de novas informações. No entanto, a principal limitação a ser considerada diz respeito à dificuldade de analisar uma rede desconhecida, bem como calcular as probabilidades de todos os caminhos possíveis. Isso pode ser, inclusive, impossível de realizar em determinadas situações na área da saúde, nas quais o diagnóstico baseia-se em experiência clínica e depende de informações de natureza subjetiva coletada. Além disso, as respostas obtidas dependem da qualidade das informações *a priori*, bem como a seleção do modelo que deve ser considerado. Portanto, as redes bayesianas devem ser vistas como uma complementaridade, e não uma substituição da tomada de decisão.

REFERÊNCIAS

1. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. The MIT Press; 2009.
2. Velikova M, van Scheltinga JT, Lucas PJ, Spaanderman M. Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. Int J Approx Reasoning. 2014;55(1,Part 1):59-73.
3. Sato RC, Zouain DM. Factor analysis for the adoption of nuclear technology in diagnosis and treatment of chronic diseases. einstein (Sao Paulo). 2012; 10(1):62-6.
4. Enderlein G, Heinemann LA, Stark H. The risk factor concept in cardiovascular disease. In: Stellman JM. Encyclopaedia of occupational health and safety. International Labour Organization; 1998.
5. Sarkar IN, editor. Bayesian methods in biomedical data analysis. New York: Academic Press; 2013.
6. Maglogiannis I, Zafiropoulos E, Platis A, Lambrinoudakis C. Risk analysis of a patient monitoring system using Bayesian Network modeling. J Biomed Inform. 2006;39(6):637-47.
7. Wu X, Li R, Fleisher AS, Reiman EM, Guan X, Zhang Y, et al. Altered default mode network connectivity in Alzheimer's disease -a resting functional MRI and Bayesian network study. Human Brain Mapp. 2011;32(11):1868-81.
8. Pinheiro PR, Castro A, Pinheiro M, editors. A Multicoloria Model Applied in the Diagnosis of Alzheimer's Disease: A Bayesian Network. Computational Science and Engineering, 2008 CSE'08 11th IEEE International Conference [Internet] 2008: IEEE [cited 2015 May 27]. Available from: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4578211&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4578211
9. Bock J, Gough DA. Toward prediction of physiological state signals in sleep apnea. IEEE Trans Biomed Eng. 1998;45(11):1332-41.
10. Fontenla-Romero O, Guijarro-Berdiñas B, Alonso-Betanzos A, Moret-Bonillo V. A new method for sleep apnea classification using wavelets and feedforward neural networks. Artif Intell Med. 2005;34(1):65-76.
11. Díez FJ, Mira J, Iturralde E, Zubillaga S. DIAVAL, a Bayesian expert system for echocardiography. Artif Intell Med. 1997;10(1):59-73.
12. Sciarretta S, Palano F, Tocci G, Baldini R, Volpe M. Antihypertensive treatment and development of heart failure in hypertension: a Bayesian network meta-analysis of studies in patients with hypertension and high cardiovascular risk. Arch intern Med. 2011;171(5):384-94. Review.
13. Lucas PJ, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artif Intell Med. 2004;30(3):201-14.