Probabilistic graphic models applied to identification of diseases

Modelos probabilísticos gráficos aplicados à identificação de doenças

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

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

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

INTRODUCTION

Data collection and decision making are part of the activities pertaining to healthcare organizations. In other words, management of health care activities is based on grasping what is thought to be true and deciding how to acting accordingly. To illustrate this point, healthcare organizations often face scenarios where health care professionals collect data on a given patient (*i.e*. symptoms, physical characteristics, history, etc.) that ultimately support diagnostic and therapeutic decisions. This paper describes how probabilistic graphical models (PGM), and Bayesian networks in particular, may contribute to decision making in health care.

The PGM are widely used in activities involving artificial intelligence, having progressed and gained momentum in this particular area over the last decades. Probabilistic graphical models can be understood as graphs where nodes correspond to variables, while uni- or bidirectional arcs represent the intervariable dependences, a structure that enables the assembly of sets of joint or conditional probabilistic distributions.[1] The term "Bayesian network" was coined in the 1980s to describe PGM intended for uncertainty management in artificial intelligence. However, advancements in computer sciences and wide applicability potential fostered the adoption of such networks by universities and large business enterprises. Diagnostic systems, gene interaction modelling and detection and quantification of causal relations in epidemiology are among the major applications of Bayesian networks in health care.

In spite of its potential applications and benefits, Bayesian networks are briefly mentioned in health-related texts and handbooks and are therefore presented

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

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

Corresponding author: Renato Cesar Sato – Instituto de Ciência e Tecnologia, Unidade Parque Tecnológico, Universidade Federal de São Paulo – Avenida Cesare Mansueto Giulio Lattes, 1,201, room 114 Eugênio de Mello – Zip code:12247-014 – São José dos Campos, SP, Brazil – Phone: (55 12) 3921-9598 – E-mail: rcsato@gmail.com

Received on: June 17, 2014 – Accepted on: Feb 22, 2015

DOI: 10.1590/S1679-45082015RB3121

in a limited or superficial form to health professionals. Bayesian networks represent a modeling approach to issues of increasing concern to health organizations over the last decade. Personalized medicine involves the prediction of disease progression based on interpretation of patient data in the light of a disease model[2] and is one potential area of application of these networks. This paper introduces the application of Bayesian networks in the context of disease; related advantages and disadvantages, limitations and deployment structure are also discussed.

APPLICATION OF BAYESIAN NETWORKS IN DISEASE RECOGNITION

Bayesian networks are models in which causality plays a fundamental role. However, incomplete understanding of the scenario being analyzed is common and probabilistic approaches are often used in an effort to describe causal relations. Probabilistic aspects therefore take on an important dimension in this network of relations to overcome limited knowledge and to help in the decision making process.

Bayesian networks are directed acyclic graphs (DAGs) where nodes represent random variables, and assumptions of intervariable independence are maintained. Nodes in graphs correspond to Bayesian network random variables and may vary in nature. Quantitative data, latent variables, unknown parameters and even research hypotheses may therefore be included in the model.

One diagram can provide explicit representation of the potential progression of a given situation and may allow inferences about the likely causes of observed effects.

Let us examine the diagram represented by figure 1 to identify risk factors for cardiovascular disease - in this case, smoking, sedentary lifestyle and stress. Latent variables (*i.e.* not amenable to direct observation but playing an important role in causality) may also be included in the model, although an additional step may be required for this purpose.[3] For example, socioeconomic status, which combines level of education, professional qualifications and occupation, may not have a direct impact on cardiovascular disease development. However, lower incidence of cardiovascular disease has been correlated with higher socioeconomic status in Eastern European, North American and Japanese studies.[4] Hence, while income may not impact the development of cardiovascular disease directly, stress factors derived from stress conditions may indeed increase related risk factors, such as alcohol abuse and socioeconomic status

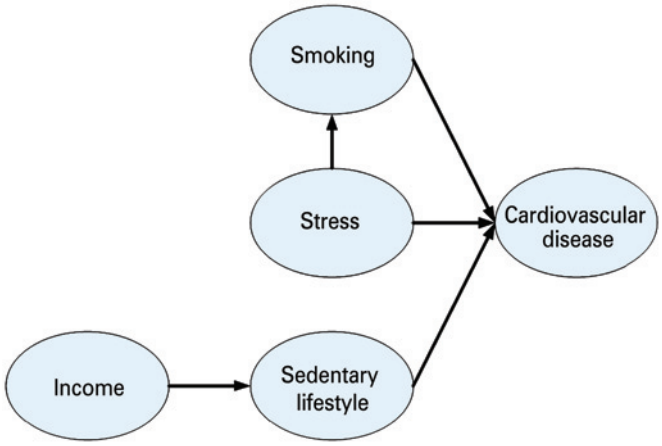

Figure 1. Basic assumptions in the model

(*i.e.* level of education, professional qualifications and occupation).

For simplification purposes, the model displayed in figure 1 starts with strong assumptions. We are assuming income alone to be a latent factor of sedentary lifestyle, whereas a more complete model should also account for relations with remaining factors in the model.

Large numbers of variables in healthcare data add to the difficulties involved in investigation and assessment of related phenomena. Scenarios where one variable depends upon one or more variables represent additional complicating factors. In a model assuming a patient suffering from a cardiovascular condition and indicated by the letter "D", some variables, such as smoking, sedentary lifestyle and alcohol abuse (letters A, B and C, respectively) may be included. This scenario can be graphically reproduced as in figure 2.

The same graphical representation can be described in terms of conditional probabilities.

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

The links between pairs of variables represent conditional dependencies, while nodes that are not connected represent variables, which are conditionally independent of each other. Such terms describe the association of the probability function from a set of values. A data set containing N variables states there are 2^N available network models and that Bayes' theorem can be used to select the best suited model for that particular dataset. The advantages of computerized systems stand out as the number of variables increase. To illustrate this point, let us compare two data sets containing 10 and 15 variables respectively. While 1,024 networks would apply to the first, 32,768 possibilities would be available to the latter. We can clearly see the need for this type of methodology in disease modeling:

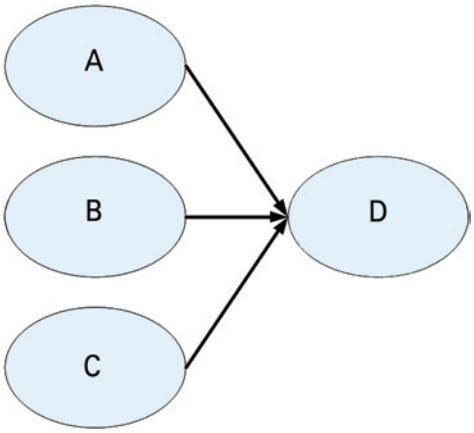

Figure 2. Conditional structure

as we try to improve our models by including new variables, the complexity of the network increases.

The search for optimized networks can be divided into three steps[5] and requires the identification of optimized parent nodes. Parent nodes are those that have nodes below them (child nodes).

APPLICATIONS IN MEDICINE

Bayesian networks have been used to model uncertainty in medicine.[6] Diagnostic decision making support is one among other applications. The method is based on categorization of responses to specific diagnostic questionnaires and as such may contribute to the diagnosis of Alzheimer's disease.[7,8] These questionnaires help to construct judgment matrices and value scales for each fundamental viewpoint previously selected by the clinician in charge.[8] The model is able to more accurately categorize the diagnostic profile of Alzheimer's disease. Sleep apnea[9,10] and cardiovascular diseases[11,12] are other potential applications.

STRUCTURE OF A BAYESIAN NETWORK MODEL

The process involved in a Bayesian network model can be described in five steps (Figure 3).[13]

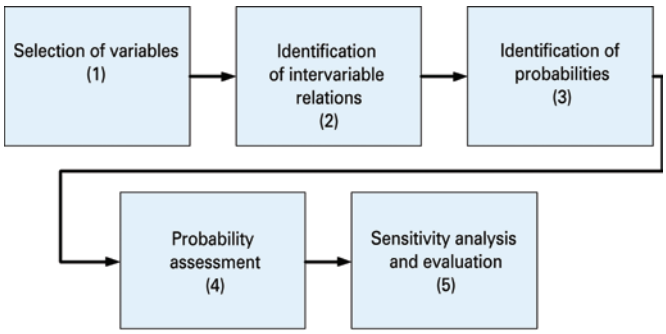

Figure 3. Steps in a Bayesian network model

Selection of relevant variables

All variables pertaining to the problem must be investigated. This is usually done by means of surveys by an expert in the process area.

Identification of intervariable relations

Identification of variables is followed by investigation of intervariable relations (*i.e.*, definition of cause-and-effect relations between variables). These causalities are also related to expert's knowledge of given events.

Identification of qualitative probabilities and logical restrictions

Identification of probability distributions required for network construction. Logical restriction aims to limit the range of probabilities that must be assessed. This step usually consists of database mapping.

Probability assessment

In this phase, probability distributions are assigned to each node in the network. Qualitative estimates can be obtained and a predetermined scale employed, or attempts can be made at visualizing the probability of an event as an area. In both cases the estimation process is long and susceptible to error; therefore outcomes may not be reliable.[1]

Sensitivity analysis and evaluation

Network models must be validated. Real data must be submitted to different probabilistic systems and outcomes compared.

CONCLUSION

Better structuring of databases from healthcare and other organizations has led to improvement of health-related causal models. This, in turn, has fostered greater interaction between systems supporting diagnostic, prognostic and therapeutic decision-making, as well as investigation of functional interactions, and medical knowledge, probability and computing. The advent of the so-called "big data" presents a good opportunity to further extend the application of this type of analysis to healthcare. In this context, Bayesian networks stand out as important tools to help overcome limitations imposed by common uncertainties in the field of health.

Bayesian networks are graphical models that enable the investigation of intervariable relations. Cause-and-effect relations can be established and probability theory

applied to uncertainty issues due to network directional nature. However, isolated application of an algorithm does not provide the best diagnostic structure, and overseeing by expert professionals is required. Potential limitations of Bayesian networks include violations of probability distributions employed for system structuring and limited system update possibilities in the face of the need for novel information. Nevertheless, the major limitation to bear in mind concerns the difficulties pertaining to analysis of an unknown network and calculation of probabilities of all possible pathways, which may actually not be feasible in healthcare scenarios where diagnosis is based on clinical experience and subjective data. Also, outcomes will reflect the quality of *a priori* data and model selection. Therefore, Bayesian networks should be viewed as ancillary tools rather than substitutes for decision-making processes.

REFERENCES

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 (São 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.