Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-20 (2015) 481–486

A Bayesian Approach to Model-Development: Automatic Learning for Tuning Predictive Performance

Logan Ward* Steen Andreassen*,**

**Centre for Model-based Medical Decision Support, Aalborg University, Aalborg, Denmark*
***Treat Systems A/S, Aalborg, Denmark*

Abstract: The value of manually constructed and tuned Bayesian networks has been demonstrated empirically, however this informal process is limited in terms of what can be reasonably achieved. This paper presents the application of a formal machine learning process, EM learning, to a manually constructed CPN for the assessment of the severity of sepsis. Through learning, the model is tuned to predict 30-day mortality, and displays a significant improvement in discriminatory ability assessed by area under the ROC curve (previous model AUC = 0.647, new model AUC = 0.739, $p<0.001$).

Key Words: Machine learning; Biological and medical system modelling; System identification and validation

1. INTRODUCTION

Bayesian networks are a set of probabilistic models and can be used to create diagnostic models for diseases (Andreassen et al. 1996; Sadeghi et al. 2006; Schurink et al. 2007; Kariv et al. 2011). These models can also provide advice on treatment selection, provided they are accompanied by decision theory and utility functions (Hejlesen et al. 1997; Andreassen et al. 1999; Leibovici et al. 2000).

A Bayesian network can be represented graphically by a set of nodes, linked together by arrows. The nodes themselves represent stochastic variables. The arrows represent causal relationships between the variables, a requirement for the network to provide plausible reasoning (Pearl 1988), and the reason they are also referred to as Causal Probabilistic Networks or CPNs (Andreassen et al. 1991). Numerically, a CPN consists of a set of conditional probability tables defining the relationships between a node and its parent(s). The task of constructing a CPN therefore consists of specifying the graphical structure and the set of associated conditional probabilities. Nodes are not limited to representing observable events such as blood pressure or temperature measurements, but can also represent latent concepts such as diagnoses or prognoses which are not observed, but still of interest. Once constructed, the CPN is used to update the probability distributions for the unobserved variables when evidence is inserted into the CPN.

CPNs are ideal models for the fusion of data and knowledge, which may be represented by patient databases and the combination of expert opinion and reports in the scientific literature, respectively. Any or all of these sources of evidence may be used in the construction of a CPN. Throughout the construction process, the conditional probabilities themselves may be considered stochastic variables. The value of the semi-formal approach of using knowledge to assign a priori distributions has been demonstrated empirically through the success of the Treat decision support system (Andreassen et al. 2005; Paul et al. 2006). Treat aids in decision-making regarding diagnosis and optimal treatment of acute infections.

The CPN model of Treat is large with close to 6000 nodes. The severity of a patient's illness is assessed by a small section of the model, approximately 40 nodes. Figure 1 presents a framework for the development of this network, referred to as the "Sepsis CPN". The individual phases are described in the literature; the initial specification of the model (Figure 1, phase I) where all observable nodes were discrete stochastic variables (Leibovici et al. 2000), known as the Discrete Sepsis CPN (D-Sepsis CPN), and the subsequent development of model with continuous variables (Figure 1, phase II), the Continuous Sepsis CPN (Ward et al. 2014). Although the conversion to continuous variables was able to solve some of the shortcomings of the discretization in the D-Sepsis model, the model requires tuning. The C-Sepsis CPN has been tuned manually, using a combination of knowledge gleaned from the literature and expert opinion, however this process is limited in terms of what can be reasonably achieved.

The C-Sepsis CPN can be further improved by supplementing the manual methods used in its development with machine learning from case databases. In this case, we take the sub-network of the C-Sepsis CPN that does not include respiratory parameters. We recognize in the Treat network that oxygen saturation, shortness of breath and respiratory rate are affected differently by lung- and other infections, and that without incorporating any knowledge of the site of infection, it does not make sense to include these parameters. The purpose of this paper is to present a method for tuning the sepsis CPN to predict all-cause 30-day mortality using a database of real patient cases. The new model is internally validated by testing its ability to predict 30-day mortality.

10.1016/j.ifacol.2015.10.187

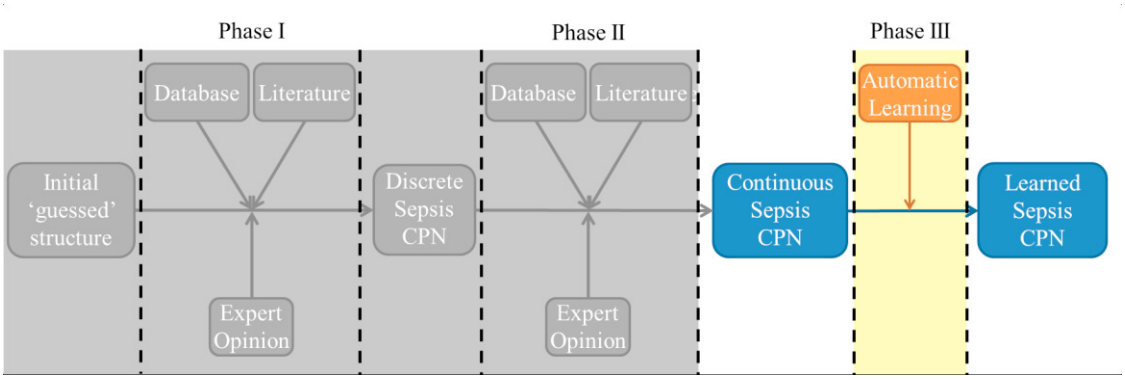

Fig. 1 Sepsis CPN development framework. Phase I describes the development of the discrete sepsis CPN (D-Sepsis CPN), phase II the continuous sepsis CPN (C-Sepsis CPN) and phase III the development of the learned sepsis CPN (L-Sepsis CPN) through formal learning methods - the subject of this paper

2. METHODS

In this paper, we describe the modification of the C-sepsis CPN into the Learned Sepsis CPN (L-Sepsis CPN) (Figure 1, Phase III of the development framework). This is the final step of our sepsis CPN development framework, with the result being the L-Sepsis CPN, or from the network constructor's perspective: the posterior distributions. For the purpose of this paper, the Continuous (C-) Sepsis CPN is regarded as the specification of a prior conditional probability distribution for the observable variables. Figure 2 shows an overview of the L-sepsis CPN. The non-infectious systemic inflammatory response syndrome (NISIRS) and sepsis represent two syndromes, the severity of which we describe using five states; no, mild, moderate, severe and critical. These states can also be thought of as the degree of activation of the immune system. Each of these severities is associated with a mortality rate. The NSIRS and sepsis nodes are linked to the infection variables, which we describe with individual parameter distributions, through a set of factor nodes. The specific structure of the sepsis CPN is described in the literature (Leibovici et al. 2000; Andreassen et al. 2005; Ward et al. 2014).

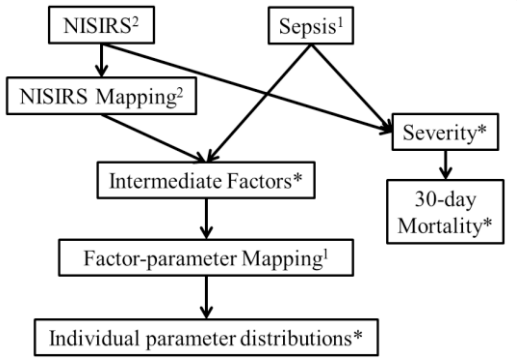

Fig. 2 Schematic view of the sepsis CPN identifying where automatic learning is to take place. 1 Learning the weights for the composite parameter distributions. 2 Learning the weights of the NISIRS severities across the intermediate factors. * denotes conditional probability tables defined prior to learning based on the C-Sepsis CPN and/or the literature. NSIRS: Non-infectious Systemic Inflammatory Response Syndrome

To prepare the L-Sepsis CPN for learning, structural changes were made. One issue with the C-Sepsis CPN is that some of the literature-derived distributions overlap greatly, which mean that they are difficult to use for classification. Additionally, the individual Gaussian distributions defined for each severity state meant that very large odds ratios were seen for outlying parameter values. Our previous attempts at learning have taught us that it is difficult to learn individual Gaussian distributions for each severity state of sepsis. Instead of doing this, we create a semi-discrete environment where a set of Gaussian curves roughly corresponding to pathophysiological states covers the region of interest for a given variable. Instead of learning the distributions themselves, we learn how each sepsis state spreads itself over the set of defined distributions, creating multi-modal or composite distributions.

Our learning process can be defined as partially supervised. We cannot observe the states of the NISIRS and sepsis nodes as such; however we can observe something to which the severity states of both are linked: 30-day mortality. The explicit definition and unobservable nature of the non-infectious SIRS also creates identifiability issues when it comes to learning. To overcome this issue, we choose to learn in a stepwise fashion, learning first the distributions for patients with infection, and then those without infection. Learning is carried out using the Expectation-Maximisation (EM) method (Lauritzen 1995).

A 10-fold cross-validation is performed as an internal validation in order to ensure that the learning method is robust. The learned network is assessed for its discriminative ability using the area under the receiver operating characteristic (ROC) curve. The performance of the L-Sepsis CPN is compared to that for Treat with the C-Sepsis CPN. Calibration of the full learned model is assessed using the Hosmer-Lemeshow statistic and calibration curve.

Descriptive statistics and significance testing for the data was carried out using SPSS (Version 22, IBM Corporation). Continuous variables were analysed using one-way ANOVA, and categorical variables with the Pearson Chi-squared statistic. Automatic learning was performed using the EM learning algorithm within Hugin (Version 7.6 (x64), Hugin Expert A/S). ROC-analysis was also undertaken in SPSS.