Journal of Biomedical Informatics 48 (2014) 28–37

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Not just data: A method for improving prediction with knowledge

Barbaros Yet [a,*], Zane Perkins [b], Norman Fenton [a], Nigel Tai [c], William Marsh [a]

[a] School of Electronic Engineering and Computer Science, Queen Mary, University of London, UK
[b] Centre for Trauma Science, Queen Mary, University of London, UK
[c] The Royal London Hospital, London, UK

ARTICLE INFO

Article history:
Received 27 June 2013
Accepted 24 October 2013
Available online 2 November 2013

Keywords:
Latent variables
Knowledge engineering
Bayesian networks

ABSTRACT

Many medical conditions are only indirectly observed through symptoms and tests. Developing predictive models for such conditions is challenging since they can be thought of as 'latent' variables. They are not present in the data and often get confused with measurements. As a result, building a model that fits data well is not the same as making a prediction that is useful for decision makers. In this paper, we present a methodology for developing Bayesian network (BN) models that predict and reason with latent variables, using a combination of expert knowledge and available data. The method is illustrated by a case study into the prediction of acute traumatic coagulopathy (ATC), a disorder of blood clotting that significantly increases the risk of death following traumatic injuries. There are several measurements for ATC and previous models have predicted one of these measurements instead of the state of ATC itself. Our case study illustrates the advantages of models that distinguish between an underlying latent condition and its measurements, and of a continuing dialogue between the modeller and the domain experts as the model is developed using knowledge as well as data.

1. Introduction

Purely data-driven approaches are currently accepted as the primary, if not the only, way of developing predictive models. Because of the impressive results achieved with such approaches by organizations like Amazon and Google, it is often assumed that this success is repeatable in other domains as long as a large enough amount of data is available. However, a purely data-driven approach can only predict the type of values recorded in a dataset, such as measurements made, decisions taken or outcomes recorded. Even when large volumes of data exist, purely data driven machine learning methods may not provide either accurate predictions or the insights required for improved decision-making. In this paper, we consider the common real-world situation in which successful decision making depends on inferring underlying or latent information that is not – and can never be – part of the data. In such a situation a predictive model for decision support will contain latent variables representing this underlying state but the values of these variables will not be present in the data. We therefore need to depend on domain expertise to identify the important latent variables and to model relations between them and the observed variables.

Domain experts do not just substitute guesswork for data. They may have access to information that is not machine-readable and they should back up any judgements by reference to published research whenever possible. Yet, such expert knowledge is usually avoided in data-driven approaches using arguments such as 'avoiding subjectivity' and 'using facts based on the data' [1,2]. The use of latent variables is also limited: some data-driven approaches, such as regression modelling, do not include latent variables at all. Other approaches contain latent variables but these are estimated only from data values, so that the use of latent variables in these methods does not escape the limits of the data. The objectivity of data-driven approaches holds only so far as the prediction of observed values really serves the needs of users. When this is not the case, erroneous results may follow. In this paper, we show some examples of these errors and how they are avoided by appropriate and rigorous use of domain knowledge.

We propose a pragmatic methodology to develop Bayesian network (BN) models with latent variables. Our method integrates domain expertise with the available data to develop and refine the model systematically through a series of expert reviews. We illustrate the application and results of this method with a clinical case study of a problem for which purely data-driven approaches have been tried but have not been considered to be successful by clinicians. Our case study shows some possible reasons for these past failures. It is beyond of the scope of the paper to provide full details of the developed model, but the details can be found at the ATC BN website [3].

* Corresponding author. Address: Risk and Information Management (RIM) Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK.
E-mail address: barbaros@eecs.qmul.ac.uk (B. Yet).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.10.012

The remainder of this paper is as follows: Section 2 presents the overview of our methodology. The case study is introduced in Section 3 and developed further in Sections 4 (learning and review) and 5 (model refinement). We present our conclusion in Section 6.

2. Method overview

The limitations of data for making predictions useful to a decision-maker can be summarised in three points:

1. *Measurement errors*: a dataset contains measurements of variables, but measurement errors mean that the true state of each variable differs from the measured data. In some domains, including clinical diagnosis, this introduces significant uncertainty about the true value, so that a data-driven model cannot accurately predict the underlying state even if it can accurately predict the associated measurement values.
2. *Sub-optimal decisions*: the objective of a decision-support model is to enable a decision-maker to determine the optimal decisions given the observed situation. A dataset may contain a 'decision' variable, that is, one that reflects the decision made (e.g. a treatment given by a clinician). A model that predicts the value of a decision variable can be useful if all the past decision-makers had similar utilities and they were completely rational in evaluating utilities with their underlying uncertainties. However, there is usually no information about the utilities involved in past decisions, and the data may have records of some decisions that were incorrect at the time or, although correct at the time, were made on outdated understanding. A model that predicts the value of a decision variable is therefore limited in its performance even if the prediction is highly accurate. Moreover, a model can only be used for 'what if' analysis – exploring the consequences of decision alternatives – if it is causal; choosing one of the decision alternatives erases the factors that influenced past decisions [4]. Although these problems are well known, models that are developed to fit past decisions are common in scientific literature (see Section 3.1).
3. *Causes of outcomes*: an 'outcome' variable records what happened. But outcomes can have many causes, only some of which may be recorded in the dataset (for example, in medical applications not all interventions and treatments are recorded). A prediction based on only some causes may be useful – the missing causes simply add uncertainty – but understanding of the scope of the causes included is important to the correct application of the model. A purely data-driven approach does not resolve this problem; only an expert can detect if the data omits known causes of the outcome. If omitted causes can be identified, this information can be used either to improve the model or to clarify its scope and to assess its performance within the scope of the causes modelled.

The main aim of our method (illustrated in the flow diagram in Fig. 1) is to overcome these limitations. We show how to develop BNs that predict and reason with latent variables using a training dataset including measurements of these variables, but not including their true state. Domain expertise is used both at the start of the development to discover latent variables and then later to refine the model in a series of expert reviews; it is during these reviews that discrepancies between knowledge and data are revealed. Expert knowledge can be used in various degrees when deriving the structure of a BN [5]. In our method, the structure of the BN is developed with domain experts by using small BN fragments for commonly occurring reasoning types as building-blocks to form the complete BN structure [6]. The advantage of experts deriving the model's structure, rather than learning it from data, is to ensure causal coherence: latent variables influence measurements and decision variables influence outcomes. Hybrid approaches that combine expert knowledge and data can also be used at this stage for deriving the BN structure [7,8]. Moreover, structure learning methods can be used as a complementary approach to evaluate and refine a BN structure developed by experts [9]. Of course, all causal assumptions need to be supported by the best available evidence, such as experimental results or expert consensus. Lack of knowledge of true causal relationship is a problem and affects both expert and data-led modelling (aside from the limited capabilities of algorithms such as inductive causation (IC) [10]) alike. Equally, not all causal relationships are uncertain: it is clear that an object's temperature causes the thermometer reading rather than the other way around.

The next step is to label the latent variables in the training dataset, overcoming the problem that their values are unknown. The first label is derived from measurement data using deterministic (but not necessarily complete) rules defined by domain experts; the second uses data clustering. The experts' rules can be of any form, but are typically derived from current practice. For example, if the related measurements are continuous, these rules are threshold values for the measurements. For clustering, we use the standard Expectation–Maximisation (EM) for BNs with known structure [11]. EM is an iterative algorithm that is used for learning the parameters of a BN from a dataset with missing values. Each iteration of EM has two steps: the E-step completes the data by calculating the expected values of unobservable variables based on the current set of parameters; the M-step learns a new set of parameters from the maximum likelihood estimate of this completed data. When EM is used for parameter learning, the M-step of the final iteration calculates the BN parameters. When it is used for clustering, the unobserved variables are labelled according to the values in the E-step of the final iteration. In our method, all of the values of the unobserved variable are missing from the dataset and we are using EM for clustering the unobserved values. Although EM can also be used for structure learning [12,13] this is not required in our method as the BN structure is developed with domain experts. Extensions of EM that builds upon the information bottleneck [14], variational Bayesian [15] and hierarchical [16] frameworks have been proposed for learning latent variables. Van der Gaag et al. [17] presents a similar approach to labelling with expert rules where they represent combinations of multiple observations with latent variables.

We now have two labels for each latent variable: one from clinical measurements and the experts' rules, the other from EM clustering. A final label is achieved by combining the two labels in cases where the labels are the same and by expert review of cases where there is a difference between the two labelling methods. We prepare a list of cases where the labels differ. Domain experts then decide the final label for each data record in this list. The experts can review other data including information that is not machine-readable and cite relevant research to support this decision. We also include a random subset of cases that were labelled consistently in the review to assess the experts' consistency with the labelling by measurements and clustering approaches. This combination of expert review and data has a number of advantages. It allows for the possibility of errors in measurement, and it uses the experts efficiently. Expert review is a costly resource and using it for every single case in the data is usually not feasible, especially if the dataset is large. Therefore, our method aims to use it only for ambiguous cases, where the labels from measurements conflict with the results of the clustering on our data.