International Journal of Approximate Reasoning 54 (2013) 1300–1309

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Refining a Bayesian Network using a Chain Event Graph ☆

L.M. Barclay *, J.L. Hutton, J.Q. Smith

Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Received 13 December 2012
Received in revised form 3 April 2013
Accepted 14 May 2013
Available online 27 May 2013

Keywords:
Chain Event Graphs
Bayesian Networks
Bayesian model selection
Causality

ABSTRACT

The search for a useful explanatory model based on a Bayesian Network (BN) now has a long and successful history. However, when the dependence structure between the variables of the problem is asymmetric then this cannot be captured by the BN. The Chain Event Graph (CEG) provides a richer class of models which incorporates these types of dependence structures as well as retaining the property that conclusions can be easily read back to the client. We demonstrate on a real health study how the CEG leads us to promising higher scoring models and further enables us to make more refined conclusions than can be made from the BN. Further we show how these graphs can express causal hypotheses about possible interventions that could be enforced.

1. Introduction

The Bayesian Network (BN) is the most widely used graphical model [7,16] which expresses the relationship between the variables of the system in terms of conditional independence statements. Its graphical structure makes it a particularly useful tool to feed conclusions back to the client and it has therefore been employed in many real-world applications. However, in certain cases the BN does not provide a rich enough structure to incorporate all information obtainable from the data set. This is the case, for example, when the conditional independence statements of the problem are asymmetric or only certain combinations of variables affect another variable and this cannot be represented simply by the directed edges between variables in the BN [21]. To take these features into account extensions to the BN have been proposed, mostly in the form of tables or tree-like structures which are added to the graph, leading to the context-specific Bayesian Network [5,21,11]. However, these methods focus primarily on efficient propagation and learning and lose the benefit of the BN's expressiveness for the client. An interesting related graphical model is the Recursive Probability Tree [6] which also focuses on efficient computation of context-specific independencies.

The Chain Event Graph (CEG) is a new flexible class of graphical models which can represent asymmetric structures directly in its topology. It is related to the Probability Decision Graph (PDG) [19,15,18] and retains the framework of a probability tree in a more compact graph. Because of its graphical derivation it inherits many of the benefits of a BN. For example, we can read off conditional independence statements directly from the topology of the graph [24,26], carry out model selection on CEGs [10] and run fast propagation algorithms [27]. It also admits a causal extension [28,25]. However, it is a more general class than either the PDG or the discrete BN. While the PDG and BN are proven to be distinct classes of models, neither containing the other [14], the CEG contains each of these classes as a special case [24]. Therefore, it is very straightforward to exploit and develop the technologies originally designed for the BN to this much richer class of CEGs. In this paper we illustrate how we use the CEG to refine an initial BN model description, demonstrating how this can provide a

* Corresponding author.
E-mail addresses: L.M.Barclay@warwick.ac.uk (L.M. Barclay), J.L.Hutton@warwick.ac.uk (J.L. Hutton), J.Q.Smith@warwick.ac.uk (J.Q. Smith).

0888-613X/$ – see front matter
http://dx.doi.org/10.1016/j.ijar.2013.05.006

more detailed but still transparent explanation of an underlying process, which seamlessly enhances an original BN analysis. The application in this paper explores a particular study of the Christchurch Health and Development Study [1], looking at the effect of social and family factors on children's health in a New Zealand birth cohort [9].

We begin the paper with a description of the programme of study forming the basis of this analysis. In Section 3 we report how an initial routine BN model search enabled us to elaborate the conclusions made in the original study. In Sections 4 and 5 we show how we then made more detailed exploration using the CEG, discovering a much better fitting model that – like a BN – could be read back to the client in narrative style. We also illustrate in Section 6 how the fitted model can be given a causal interpretation which, if valid, allows us to perform various 'what–if' analyses under various policy controls. In the final section we propose further extensions to the present analysis.

2. The Christchurch Health and Development Study

The Christchurch Health and Development Study (CHDS) is carried out by a research group at Otago University, led by Professor David Fergusson. It is a cohort study which has followed up 1265 children born in mid 1977 in Christchurch, New Zealand, for over 30 years. The CHDS has explored the children's development from childhood to adulthood regarding their education, behaviour and health with respect to a wide range of social, economic and family factors.

The running example used in this paper reanalyses an early subset of the CHDS discussed in [9] which studies the first five years of the Christchurch cohort looking at the effect the family's social background, the economic status and the number of family life events have on the child's health which is measured by rates of hospital admission.

Based on previous study of the data set, Fergusson et al. [9] concluded to only consider admissions due to illness and accidents as these were the only reasons for admissions that are sensitive to social and family situation. To describe the family's social background the CHDS group collected information about the mother's education and age at birth, the family's socioeconomic status and ethnic origin, and whether the child grew up in a single or two parent family. These variables were then combined using factor analysis to give a single measure of the social background (see [8] for details). Similarly, the economic status was measured as a function of the family income, possible financial difficulties and by rating the standard of living and the quality of the accommodation of the child. Again these were simplified into a single measure of the overall economic situation. Of particular interest in this study was whether the effect of adverse life events in the children's lives might be associated with an increased rate of illness. This has provoked controversy and continues to be a subject of research [29]. Twenty events were classed as life events, based on a variation of the Holmes and Rahe Social Readjustment Rating Scale [13], in which the mother of the child was interviewed and reported on the events that occurred. These included the experience of moving house, the husband changing job, the death of a close friend or relative, serious financial problems within the family, divorce, or a serious illness or accident within the family [3].

Data was collected through a variety of sources: Mothers were interviewed every year on the family life events, while the economic situation was predominantly rated every year by an interviewer. Hospital admissions were also obtained by interviewing the mother of every child at each year and further by asking them to keep a health diary record. The results were additionally compared with the central records of the Christchurch Hospital.

Rates of hospital admissions were first compared for the various levels of each covariate separately using one-way analysis of variance, concluding that the more socially or economically disadvantaged the child's background (without adjusting for the other factors) the higher the hospital admission rate tended to be. In addition to this the probability of an admission increased significantly with the number of family life events. For further analysis a Cox proportional hazards model estimating the risk of hospital admission over the five years was fitted. Results from [9] showed that according to this model the family's economic status did not influence the risk of admission significantly once adjusting for the other covariates, suggesting that in this type of population financial problems were not the main reason for health problems. In contrast, family life events and social background both appeared to have a significant impact on the admission rates, even once having adjusted for the other covariates. In particular, the most significant association was found between the hospital admissions rate and the number of family life events. These interesting conclusions motivated the following analysis. We demonstrate that our graphical methods not only allow us to disentangle the association between the variables but enable us to present our results in a transparent way.

Complete data was available to us for 890 children and so our analysis was carried out on these. To construct the four variables of interest we aimed to follow as far as possible the methodology of Fergusson et al. [9]. However, as the variables describing the social and economic background are discrete, predominantly with few categories, we slightly adapted the methods of Fergusson et al. [9], who use a factor analysis. Instead we fitted a latent-class model to construct a categorical variable for the social background and the economic situation. The predicted classes of each child were used to describe its social and economic situation. For simplicity, we here assumed binary latent classes throughout, distinguishing between a high or low social background and a high or low economic situation. We further made a distinction between 'no hospital admission' and 'at least one hospital admission' since the counts for more than one admission are sparse. Similarly, we divided the life events into the three approximately equal sized categories: 'low' (0–5 events over the five years), 'average' (6–9 events) and 'high' (at least 10 events). There were 169 (19%) children with at least one admission and the proportion of admission overall ranged from 0.12 to 0.26 (Table 1).