International Journal of Approximate Reasoning 51 (2010) 167–178

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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

"I can name that Bayesian network in two matrixes!"

Russell G. Almond [1]

Educational Testing Service (ETS), Constructed Response Scoring, Rosedale Rd, Princeton, NJ 08541, United States

ARTICLE INFO

Article history:
Received 11 March 2008
Received in revised form 23 April 2009
Accepted 30 April 2009
Available online 20 May 2009

Keywords:
Bayesian networks
Elicitation
Q-matrix
Assessment design
Covariance selection models

ABSTRACT

For a number of situations, a Bayesian network can be split into a core network consisting of a set of latent variables describing the status of a system, and a set of fragments relating the status variables to observable evidence that could be collected about the system state. This situation arises frequently in educational testing, where the status variables represent the student proficiency and the evidence models (graph fragments linking competency variables to observable outcomes) relate to assessment tasks that can be used to assess that proficiency. The traditional approach to knowledge engineering in this situation would be to maintain a library of fragments, where the graphical structure is specified using a graphical editor and then the probabilities are entered using a separate spreadsheet for each node. If many evidence model fragments employ the same design pattern, a lot of repetitive data entry is required. As the parameter values that determine the strength of the evidence can be buried on interior screens of an interface, it can be difficult for a design team to get an impression of the total evidence provided by a collection of evidence models for the system variables, and to identify holes in the data collection scheme. A *Q*-matrix – an incidence matrix whose rows represent observable outcomes from assessment tasks and whose columns represent competency variables – provides the graphical structure of the evidence models. The *Q*-matrix can be augmented to provide details of relationship strengths and provide a high level overview of the kind of evidence available. The relationships among the status variables can be represented with an inverse covariance matrix; this is particularly useful in models from the social sciences as often the domain experts' knowledge about the system states comes from factor analyses and similar procedures that naturally produce covariance matrixes. The representation of the model using matrixes means that the bulk of the specification work can be done using a desktop spreadsheet program and does not require specialized software, facilitating collaboration with external experts. The design idea is illustrated with some examples from prior assessment design projects.

[1] Lou DiBello provided a lot of the impetus for "simplifying Bayesian networks" which spurred this research. Debbie Pisacreta and Holly Knott did some early prototypes of a *Q*-matrix tool. Peggy Redman and Lisa Hemat have both served in the capacity of ECD librarian and their feedback has been a valuable source of input. Lisa additionally was an early user of many of the R functions. Interactions with John Sabatini, Richard Roberts and Aurora Graf, each experts in their respective fields, has taught me much about communicating ECD designs and methodology with others. Finally, the experiences of the ACED design team, Val Shute, Aurora Graf, Jody Underwood, David Williamson, and Peggy Redman have been very influential in this reconception of the ECD tools. ACED development and data collection was sponsored by National Science Foundation Grant No. 0313202.

E-mail addresses: ralmond@ets.org, almond@acm.org
URLs: http://ralmond.net

0888-613X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2009.04.005

1. Problem

One important application of Bayesian networks is to infer the state of an unknown system from a collection of observable evidence. In educational testing, the target of inference is the state of one or more proficiency variables, and the observable evidence is the outcomes from multiple tasks designed to require the targeted proficiencies. In a more classical diagnosis setting, the unknowns might be the state of health of a patient or piece of equipment and the observables come from the outcomes from various tests performed by a doctor or technician. To avoid the use of the word "test" which has different connotations in these different applications, I will call an activity designed to produce observable evidence about the system state a *task*, and a collection of tasks used for the purposes of inferring the state of the system an *assessment*.

As the number of tasks grows large, so does the size of the Bayesian network model for the assessment. In the educational testing setting, the form of an assessment is often chosen from among a large pool of potential tasks. In this case, knowledge engineering techniques are required to manage the volume of information. Several authors [16,14,13,19] have proposed dividing such large networks up into a library of graph fragments, which are then assembled to create the a Bayes net to answer a specific query. These methods try to take advantage of natural design patterns in network construction to produce prototype networks that are applicable to a wide collection of data gathering tasks. Networks based on these design patterns typically share the same structure and differ only in parameters relating to the strength of the relationships among the variables. Almond and Mislevy [3] note that when the goal of the Bayesian network is to infer the state of an unknown system, the library of fragments has a special form. A core system model, or *student proficiency model*, captures the relationship among the status variables for the subject as a complete Bayesian network. A library of evidence model fragments relate the observable outcomes from a particular assessment task to the core status variables. As many tasks are close variants, the system can rely on design patterns to produce evidence models that differ only in the value of the parameters.

In psychological measurement applications, the proficiency (status) variables are typically latent, and therefore expert knowledge is required to label the variables and their states, even if the model parameters are later refined through data. Evidence-centered assessment design (ECD; [18]) is a knowledge engineering method applicable to building Bayesian network models for educational assessment. It stores the complete network for an assessment pool as a central core *student proficiency model* and a library of *evidence models* corresponding to the tasks. Section 2 describes our first ECD design repository and its limitations.

Although the library of fragments representation is a complete description of the problem space, it is not necessarily the most convenient representation for every purpose. In particular, asking how much potential evidence is available in an assessment pool for a particular hypothesis about student proficiency requires scanning all of the fragments in the library. A more compact view is convenient for answering these kinds of queries. Note that the graphical structure of the network may also be expressed through an incidence matrix, a matrix whose rows and columns correspond to nodes in the graph and where a positive value indicates an edge between the corresponding nodes. This paper proposes using two matrixes as views of the assessment: the *Q*-Matrix (Section 3) describes the relationship between proficiency variables and observable outcome variables, and a correlation matrix (Section 4) expresses the relationship among the latent proficiency variables. These two matrixes not only provide a good overview of the model, but they also can be specified using the spreadsheet programs commonly available on personal computers, and hence do not require specialized software. This suggests a lighter weight, more nimble procedure for knowledge engineering (Section 5) useful for models for tracking the unknown state of a system.

2. The evidence-centered design data repository

In ECD, a complete design for an educational assessment consists of a number of design objects called models [18]. The four central models for the assessment lay out the basic evidentiary basis for the assessment as follows:

(1) *Student proficiency model*: Identifies the aspects of student knowledge, skill and ability about which the assessment will make claims. (This corresponds to the system status model in other applications.)
(2) *Evidence model*: Identifies observable evidence for the student having (or not having) the targeted competencies.
(3) *Task model*: Designs situations which provide the student with opportunities to provide that evidence. (These do not correspond to Bayes net fragments, but rather describe data gathering procedures for which particular evidence model fragments are appropriate.)
(4) *Assembly model*. Describes rules for how many of what kinds of tasks will constitute a valid form of the assessment.

Consider a reading assessment. The reading experts have identified four aspects of reading proficiency (e.g., recognizing word meanings, synthesizing information) that they want to report on. These are labeled $S_1, \ldots, S_4$. The proficiency model is the joint distribution over these variables in the population of test takers. The reading proficiencies cannot be observed directly; instead, the assessment designers write a series of reading tasks (consisting of a reading assignment followed by a short question or questions about what was read) to assess the examinee's reading proficiency level. Many reading tasks follow a common design shell; the design shells are the task models. Tasks from a given task model usually require a subset of the proficiencies to solve. We can represent this by drawing a graph fragment linking the appropriate proficiency variables to the observable outcome variables (that represent scored responses to the tasks). Note that the psychometric properties (such