ADAPQUEST: A Software for Web-Based Adaptive Questionnaires based on Bayesian Networks

Claudio Bonesana , Francesca Mangili , Alessandro Antonucci
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA) - Lugano, Switzerland
{claudio.bonesana,francesca.mangili,alessandro}idsia.ch

Abstract

We introduce ADAPQUEST, a software tool written in Java for the development of adaptive questionnaires based on Bayesian networks. Adaptiveness is intended here as the dynamical choice of the question sequence on the basis of an evolving model of the skill level of the test taker. Bayesian networks offer a flexible and highly interpretable framework to describe such testing process, especially when coping with multiple skills. ADAPQUEST embeds dedicated elicitation strategies to simplify the elicitation of the questionnaire parameters. An application of this tool for the diagnosis of mental disorders is also discussed together with some implementation details.

1 Introduction

A questionnaire is called *adaptive* when its question sequence is dynamically driven by the answers of the taker. The typical goal is to optimally estimate an aspect of interest of the test taker described by a set of target variables (e.g., his/her skills) while also reducing as much as possible the number of questions.

Algorithm 1 depicts a standard workflow for adaptive questionnaires. The best question (Q^*) to ask in a particular stage of the questionnaire is picked from an item pool ($\boldsymbol{Q}$) by a function (`Pick`) of the previous answers ($\boldsymbol{e}$). The answer of the test taker (σ) is consequently collected (`Answer`) and the process iterated unless some (`Stopping`) condition, still based on the previous answers, is achieved. Finally a function (`Evaluate`) returns a grade based on all the answers collected before the end of the questionnaire.

Trading off accuracy and the number of questions is the typical challenge with real adaptive systems. Algorithms to drive the selection mechanism are extremely important to improve the quality and the reliability of the evaluation process in modern interactions with users. This has to be supported by flexible interfaces able to provide such adaptiveness and interact with portable implementations of the above algorithms.

An important field of application of adaptive questionnaires is education, where they can be used both for training and assessment. In classical assessment tests, tools to achieve some form of adaptiveness by simple deterministic rules have been considered [DeVellis, 2006]. More successful results can be achieved by the latent modelling of the skill level of the taker [Courville, 2004].

Algorithm 1 Adaptive questionnaire workflow: given student σ and item pool $\boldsymbol{Q}$, a grade based on answers $\boldsymbol{e}$ is returned.

$\boldsymbol{e} \leftarrow \emptyset$
while not `Stopping(e)` **do**
  $Q^* \leftarrow$ `Pick`$(\boldsymbol{Q}, \boldsymbol{e})$
  $q^* \leftarrow$ `Answer`(Q^*, σ)
  $\boldsymbol{e} \leftarrow \boldsymbol{e} \cup \{Q^* = q^*\}$
  $\boldsymbol{Q} \leftarrow \boldsymbol{Q} \setminus \{Q^*\}$
end while
return `Evaluate(e)`

Item response theory (IRT) is the most popular approach of this kind [Embretson and Reise, 2013]. The probability of a correct answer is described by a logistic model with a small number of parameters. Under standard independence assumptions, this allows for a simple updating process, thus making also very easy the implementation of adaptive strategies. As a matter of fact, implementing IRT, but also more sophisticated techniques, such as the Rasch model [Brinkhuis and Maris, 2020], in a computer system is relatively straightforward and a huge number of tools for e-learning tools currently embed these algorithms.[1]

Despite its popularity and advantages at the implementation level, IRT might be unable to properly cope with questionnaires aiming to evaluate multiple target variables in the same moment [Millán *et al.*, 2000]. The IRT independence assumptions in those cases might be unrealistic, and the model consequently performs poorly. In order to overcome these weaknesses, other formalisms have been considered. Among them, Bayesian networks (BNs, [Koller and Friedman, 2009]) emerged as a sensible choice able to guarantee an accurate selection of the items [Vomlel, 2004], but also a good explainability of the actions [Almond *et al.*, 2015].

However, in spite of a huge amount of adaptive tools based on IRT, BNs are much less used in this area. To the best of our knowledge, the software we are presenting, called ADAPQUEST[2] is the first mature contribution of this kind.

[1] E.g., `concertoplatform.com`.

[2] See `github.com/IDSIA/adapquest`.

We see two possible explanations for such situation. First, although BN inference is nowadays a standard technique, the number of freely available libraries for this task is limited and their embedding in other software projects might be not smooth, while an implementation from scratch would require dedicated efforts not always compatible with an application project. Second, the number of parameters to be tuned for a BN approach might be large and typically higher than those needed for IRT. As the target variables are often regarded as latent variables, learning them from data might not be possible (see [Plajner and Vomlel, 2020] for dedicated data approaches) and elicitation techniques should be considered instead. This might be time consuming and also complicated for practitioners not confident with probabilistic graphical models, thus preventing a widespread diffusion of those flexible approaches.

Such situation motivated us to present ADAPQUEST, as a new freely available Java software tool embedding BN inference and modelling features implemented for the design of adaptive tests and their practical implementation through web interfaces. ADAPQUEST supports state-of-the-art techniques for both the elicitation process, intended to make as simple and as smooth as possible such elicitation process, and the adaptive selection of the items intended to guarantee the necessary explainability of the process [Antonucci *et al.*, 2021]. The tool is directly built on the top of a recently developed library for probabilistic graphical models, that takes care of the necessary inference tasks [Huber *et al.*, 2020].

The paper is organised as follows. In Section 2 we discuss the basic ideas of adaptive testing based on BNs and the tools used for explainability and elicitation. In Section 3 we give some technical information related to the development of ADAPQUEST. In Section 4 we consider a case study, already implemented in ADAPQUEST, and freely available to the community as a demonstrative project. A discussion about possible outlooks concludes the paper in Section 5.

2 Adaptive Questionnaires by Bayesian Nets

Bayesian networks (BNs) [Koller and Friedman, 2009] are a popular class of probabilistic graphical models designed for a compact specification of joint, generative, probability distributions. To implement adaptive questionnaires with BNs we regard the set of *skills* to be evaluated during the questionnaire as a joint variable $\boldsymbol{S}$, and we similarly regard the item pool $\boldsymbol{Q}$ as a set of variables. Here we only consider discrete variables. The BN uses a directed acyclic graph over these variables as a model of the conditional independence relations among them. This allows for a compact specification of the distribution $P(\boldsymbol{S}, \boldsymbol{Q})$. Given such a generative model, standard algorithms for BNs can be used to answer queries about the model variables. E.g., $P(\boldsymbol{S}|\boldsymbol{e})$ is the posterior probability for the skills given the answers $\boldsymbol{e}$ to the questions in $\boldsymbol{Q}$ properly asked to the taker. This distribution can be used to grade the taker at the end of the questionnaire, but also to decide whether or not to keep asking questions. For the latter task information-theoretic measures, such as the entropy $H(\boldsymbol{S}|\boldsymbol{e})$ of the posterior distribution over the skills given the answers received so far, are used and we typically stop the questionnaire when this entropy level goes below a threshold H^*. The selection of the optimal question to ask to the taker is slightly more involved: as the actual answer to a candidate question Q is not known, expectation obtained by a weighted average over the probability for the possible answers (i.e., the conditional entropy $H(\boldsymbol{S}|Q, \boldsymbol{e})$) should be considered instead. To detect the optimal question we maximize the *information gain*, i.e., the difference between the current entropy and the conditional one for the candidate question. Algorithm 2 depicts a typical workflow for BNs. The final grade is also an expectation, based on the posterior distribution of a function f able to precisely grade the taker when no uncertainty about the skills is present.

Algorithm 2 Adaptive questionnaire workflow based on a BN over the questions $\boldsymbol{Q}$ and the skills $\boldsymbol{S}$: given the taker profile $\boldsymbol{s}_\sigma$, the algorithm returns an evaluation corresponding to the expectation of an evaluation function f with respect to the posterior for the skills given the answers $\boldsymbol{e}$.

$\boldsymbol{e} = \emptyset$
while $H(\boldsymbol{S}|\boldsymbol{e}) > H^*$ **do**
  $Q^* \leftarrow \arg\max_{Q \in \boldsymbol{Q}} [H(\boldsymbol{S}|\boldsymbol{e}) - H(\boldsymbol{S}|Q, \boldsymbol{e})]$
  $q^* \leftarrow \texttt{Answer}(Q^*, \boldsymbol{s}_\sigma)$
  $\boldsymbol{e} \leftarrow \boldsymbol{e} \cup \{Q^* = q^*\}$
  $\boldsymbol{Q} \leftarrow \boldsymbol{Q} \setminus \{Q^*\}$
end while
return $\mathbb{E}_{P(\boldsymbol{S}|\boldsymbol{e})}[f(\boldsymbol{S})]$

The procedure is extremely easy to achieve, provided that a reliable BN inference engine to compute $P(\boldsymbol{S}|\boldsymbol{e})$ and $P(Q|\boldsymbol{e})$ is available. In terms of explainability the model offers high transparency: the numerical values leading to a particular question selection, to the stopping condition and to a grade might be reported online during the test execution. Moreover, the techniques recently proposed in [Antonucci *et al.*, 2021] allow to associate with such quantitative information the modal state of the variables, this providing a qualitative summary of the different actions.

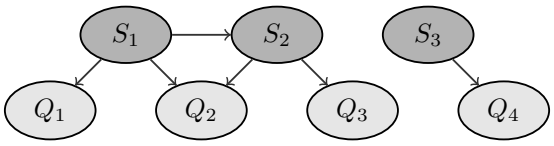

Figure 1: BN for questionnaires with three skills and four questions.

The only critical part of such workflow is the learning of the BN structure and its parameters. In principle, given a data set of observations for $\boldsymbol{S}$ and $\boldsymbol{Q}$, standard statistical learning techniques could be used. This typically requires complete data, but algorithms to cope with partially incomplete data are also available. Yet, the skills $\boldsymbol{S}$ are typically represented by latent variables and their observations unavailable. This particular situation when modelling adaptive questionnaires have been investigated, and specialised learning techniques have been developed (e.g., see [Plajner and Vomlel, 2020]).

In ADAPQUEST, we assume the design of the questions and the quantification of the BN to take place in the same

time. Thus, when no complete data are available, we assume the quantification process to be based on an elicitation process from a domain expert (e.g., the instructor). Regarding the structure, the natural interpretability of the directed graph underlying a BN makes this task simple: first a graph over $\boldsymbol{S}$ is elicited in order to reflect the dependencies between the skills (e.g., in Figure 1, the first two skills are connected and hence dependent, while the third is independent from the first two). Regarding the questions, as we assume them to correspond to the children of the skills, the only effort is to identify the skills relevant to properly answer a particular question.

Once this qualitative part has been achieved, the quantification of the parameters in a BN corresponds to assessing probabilities for single nodes/variables given all the values of the parents. Typical questions are about the probability of a correct answer if the taker has the necessary skills. Or the probability of having skill S_2 while not having S_1, and so on.

Alternative parametrisations has been proposed to make such elicitation easier [Antonucci *et al.*, 2021]. Consider for instance a Boolean skill S and a Boolean question Q. Two probabilities such as $p := P(q|s)$ and $p' := P(q|\neg s)$ corresponding to the probability of a correct answer given that the taker has or has not the skill are sufficient to quantify the relation. A possibly simpler parametrisation is provided by $\delta := p - p'$ and $\gamma := 1 - \frac{p+p'}{2}$, the two numbers being still normalised between zero and one and giving the discriminative power of the question and its difficulty. These parameters can be extended to the general case, and make faster and easier the elicitation efforts when adding new items to the pool.

3 The ADAPQUEST Software

ADAPQUEST is a REST micro-service written in Java using the Spring Boot framework.[3] The structure is extremely flexible and the configuration of a new survey/test can be done via API, via code, or by using simple JSON files. If needed, the tool can also connect to an already configured and compatible database where the item pool and the model are stored. The simple hierarchy of classes and a state-of-the-art architecture allows the system to be easily expanded with new features and adaptive criteria to cover further explorations on the field.

From a developer point of view, each questionnaire/survey is composed by three parts: (i) a BN model used to perform inference in order to find the next best question based on the history of given answers as in Algorithm 2; (ii) a pool of questions, associated with their BN nodes; (iii) and an highly configurable part that act on the adaptive engine. The BN inferences are based on the CreMA library [Huber *et al.*, 2020].

The tool is intended to be a back-end service that needs a custom front-end web-application to show the questions. Yet, a demonstrative web interface is already available. This can be used to test the functioning of the tool itself, of new models, and the questions flow. An exchange library can be used to query and manage a remote tool from a client, allowing standard CRUD operations on the surveys and the stored answers. This also allows to integrate the tool as a library inside other projects and easily interact with it. This is especially important when running extensive simulations.

[3] See `spring.io/projects/spring-boot`.

4 A Case Study on Mental Disorder Diagnosis

Finally, let us present a case study involving the practical use of ADAPQUEST in the development of a survey. The code for the specification of the model used for that is freely available in the ADAPQUEST repository and can be used as a guidance for the development of new applications within this framework. Although here the focus is on a survey, the implementation of a test would be identical. We refer the reader to [Mangili *et al.*, 2017] for an educational application whose items cannot be disclosed for the sake of confidentiality.

Following the empirical evidence corroborating the relation between job quality and mental health [Bracci and Riva, 2020], ADAPQUEST was applied to the development of a questionnaire for the early detection of employees at risk of mental health problems. Data for the development of the model were taken from the Swiss Household Panel (SHP) [Tillmann *et al.*, 2016]. The association between job quality and health being the focus of the questionnaire, mostly job related questions were selected from the panel. General demographic information (e.g., age, region, education, etc.) and questions about the current perceived mental states (e.g., self assessed degree of anger, happiness, etc. on a scale from 0 to 10) were also included. Overall, the questions database contains 48 questions based on which the risk of developing a stated of distress, experiencing lack of happiness or running into psychological disorders within one year are estimated. Those three events of interest are described by the binary variables *distress*, *lack*, *disorder*.

The BN model used in the adaptive questionnaire is based on a naive Bayes classifier [Koller and Friedman, 2009], the classes being represented by the joint state of the three variables *distress*, *lack* and *disorder*, which is described by a single node called *target*. After collecting a sufficient number of answers, the system computes from the posterior probabilities of the *target* node the marginal probabilities of the three health-related variables being true. Such probabilities are taken as a measure of risk for the mental well-being of the test taker. The naive Bayes model adopts the questionable assumption of independence of the question nodes given the target node, however, the predictive performance of that model exceeded that of more complex networks. Namely, with a 10-fold cross validation over the SHP data set (including 57422 instances) we estimated an AUC of 0.78 (standard deviation = 0.01) 0.88 (s.d. = 0.01) and 0.78 (s.d. = 0.03), for the *distress*, *lack* and *disorder* variables, respectively.

5 Conclusions and Outlooks

We presented a new software tool for the design and implementation of adaptive questionnaires and surveys based on Bayesian networks. The tool is freely available to the community. As a necessary future work we intend to support *credal* networks [Piatti *et al.*, 2010] for the design of adaptive questionnaires. This can be based on the ideas outlined in [Mangili *et al.*, 2017] and later extended by [Antonucci *et al.*, 2021]. This would allow to support interval-valued elicitation, thus providing higher realism in the modelling step.

Acknowledgments

The case study has been realised using the data collected by the SHP, which is based at the Swiss Centre of Expertise in the Social Sciences FORS. The project is financed by the Swiss National Science Foundation

References

[Almond *et al.*, 2015] Russell G Almond, Robert J Mislevy, Linda S Steinberg, Duanli Yan, and David M Williamson. *Bayesian networks in educational assessment*. Springer, 2015.

[Antonucci *et al.*, 2021] Alessandro Antonucci, Francesca Mangili, Claudio Bonesana, and Giorgia Adorni. A new score for adaptive tests in Bayesian and credal networks. In *Proceedings of the Sixteenth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2021)*, 2021.

[Bracci and Riva, 2020] Anna Bracci and Egidio Riva. Perceived job insecurity and anxiety. a multilevel analysis on male and female workers in European countries. *Frontiers in Sociology*, 5:75, 2020.

[Brinkhuis and Maris, 2020] Matthieu J.S. Brinkhuis and Gunter Maris. Dynamic estimation in the extended marginal Rasch model with an application to mathematical computer-adaptive practice. *British Journal of Mathematical and Statistical Psychology*, 73(1):72–87, 2020.

[Courville, 2004] Troy G. Courville. *An empirical comparison of item response theory and classical test theory item/person statistics*. PhD thesis, Texas A&M University, 2004.

[DeVellis, 2006] Robert F. DeVellis. Classical test theory. *Medical care*, pages S50–S59, 2006.

[Embretson and Reise, 2013] Susan E. Embretson and Steven P. Reise. *Item Response Theory*. Psychology Press, 2013.

[Huber *et al.*, 2020] David Huber, Rafael Cabañas, Alessandro Antonucci, and Marco Zaffalon. CREMA: a Java library for credal network inference. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models (PGM 2020)*, Proceedings of Machine Learning Research, Aalborg, Denmark, 2020. PMLR.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

[Mangili *et al.*, 2017] Francesca Mangili, Claudio Bonesana, and Alessandro Antonucci. Reliable knowledge-based adaptive tests by credal networks. In A. Antonucci, L. Cholvy, and O. Papini, editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty. ECSQARU 2017*, volume 10369 of *Lecture Notes in Computer Science*, pages 282–291. Springer, Cham, 2017.

[Millán *et al.*, 2000] Eva Millán, Mónica Trella, José-Luis Pérez-de-la Cruz, and Ricardo Conejo. Using Bayesian networks in computerized adaptive tests. In *Computers and Education in the 21st Century*, pages 217–228. Springer, 2000.

[Piatti *et al.*, 2010] Alberto Piatti, Alessandro Antonucci, and Marco Zaffalon. *Building Knowledge-based Expert Systems: A Tutorial*, volume 11, chapter 2. Nova Science Publishers, New York, 2010.

[Plajner and Vomlel, 2020] Martin Plajner and Jiří Vomlel. Monotonicity in practice of adaptive testing. *arXiv preprint arXiv:2009.06981*, 2020.

[Tillmann *et al.*, 2016] Robin Tillmann, Marieke Voorpostel, Ursina Kuhn, Florence Lebert, V-A Ryser, Oliver Lipps, Boris Wernli, and Erika Antal. The swiss household panel study: Observing social change since 1999. *Longitudinal and Life Course Studies*, 7(1):64–78, 2016.

[Vomlel, 2004] Jiří Vomlel. Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12(supp01):83–100, 2004.