Reliable Knowledge-Based Adaptive Tests by Credal Networks

Francesca Mangili, Claudio Bonesana, and Alessandro Antonucci

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
Lugano, Switzerland
{francesca,claudio,alessandro}@idsia.ch

Abstract. An *adaptive* test is a computer-based testing technique which adjusts the sequence of questions on the basis of the estimated ability level of the test taker. We suggest the use of *credal networks*, a generalization of Bayesian networks based on sets of probability mass functions, to implement adaptive tests exploiting the knowledge of the test developer instead of training on databases of answers. Compared to Bayesian networks, these models might offer higher expressiveness and hence a more reliable modeling of the qualitative expert knowledge. The counterpart is a less straightforward identification of the information-theoretic measure controlling the question-selection and the test-stopping criteria. We elaborate on these issues and propose a sound and computationally feasible procedure. Validation against a Bayesian-network approach on a benchmark about German language proficiency assessments suggests that credal networks can be reliable in assessing the student level and effective in reducing the number of questions required to do it.

1 Introduction

The use of communication and information technologies in education is actually growing. Both online (e.g., MOOCs) and classroom courses are urgently asking for more flexible and sophisticated e-learning and e-testing tools [1]. AI-based approaches such as intelligent tutoring systems, adapting the interaction with the student on the basis of his/her knowledge and/or psychological profile, represent an important direction to improve the quality of the (e-)learning experience [2].

Bayesian networks (BNs) [3] have been used to model the knowledge driving such intelligent systems [4]. However, collecting large sets of reliable data in educational domains may be difficult and time consuming (e.g., a course with few students, or taught for the first time), and the quantification should be based on expert knowledge only. To elicit a Bayesian network, an expert might face questions like: "*which is the probability of a student with a particular knowledge level giving the right answer to a question?*". Giving sharp probabilities for questions of this kind can be problematic for an expert, whose knowledge is mostly qualitative (e.g., "*a right answer is very unlikely*"). Fuzzy linguistic approaches represent a viable, non-numerical, way to address these issues [5]. To stick within

the probabilistic framework, verbal-numerical probability scales associated with sharp values [6] or intervals [7] have been also proposed.

In this paper we show how to conjugate an interval-valued probabilistic elicitation of expert knowledge with the BN framework. This means to cope with a *credal network* (CN) [8], a generalization of BNs based on the imprecise probability theory [7], where local parameters are defined by set-valued probabilities. This simplifies the elicitation process and offers a more reliable handling of the related uncertainty. Moving from BNs to CNs implies two main issues: (i) numerical inferences will be interval-valued too, thus making debatable both the decision criterion [9] an the information measures [10] to adopt; and (ii) inference tasks in CNs typically belongs to higher complexity classes than their Bayesian counterparts [11]. Both these issues are addressed by defining a computationally feasible procedure based on CNs to be used for practical implementation of intelligent systems solely specified by expert knowledge. To the best of our knowledge this is the first attempt to perform e-testing with models of this kind.

We focus on the application of CNs to *computer adaptive testing* (CAT), i.e., an approach to e-testing that adjusts the sequence and the number of questions to the ability level of the test taker. CATs have the potential to make the test an individualised experience that challenges and does not discourage the test takers, as most of the questions are near their ability levels. Building upon *item response theory* [12], the common background underpinning CATs, graphical modeling (such as BNs and CNs) offers a powerful language for describing complex multivariate dependencies between skills and rich tasks. Several researchers have exploited the potential of BNs both in adaptive and non-adaptive educational assessment [13,14]. These authors focus on applications for which data are available to learn the model parameters. We regard this point as a serious limitation, possibly hindering CATs adoption by many teachers and instructors.

We start from a CAT procedure based on BNs that uses *entropy* as the information-theoretic measure driving the question selection and the stopping criteria (Sect. 2). Our goal is to improve this procedure by using CNs to better describe the pervasive uncertainty characterizing the model. A direct extension of the Bayesian framework to CNs would require the computation of bounds for the conditional entropy with respect to the CN specification. This corresponds to a non-linear non-convex optimization task. We therefore propose a number of simplifying assumptions to overcome this problem at the price of accepting sub-optimal question selection schemes (Sect. 3). The approach is tested on a real-world benchmark about German language proficiency assessment (Sect. 5). The results are promising: CAT based on CNs is effective in reducing the number of questions while maintaining a high accuracy in the evaluation and the approximations introduced do not compromise the procedure's effectiveness.

2 Adaptive Testing by Bayesian Networks

Skills modeling. We describe the knowledge level of a student as a collection of categorical variables, say $\boldsymbol{X} := (X_1, \ldots, X_n)$, called *skills*. A joint *probability*

mass function (PMF) $P(\boldsymbol{X})$ describes the uncertainty about the actual values of the skills. A compact specification of such multivariate model can be achieved by a BN [3]. This corresponds to: (i) a directed acyclic graph whose nodes are in one-to-one correspondence with the variables of $\boldsymbol{X}$; and (ii), for each $X_i \in \boldsymbol{X}$, a collection of conditional PMFs $P(X_i|\pi_{X_i})$, one for each value π_{X_i} of the joint variable Π_{X_i} denoting the *parents* (i.e., the immediate predecessors) of X_i. The *Markov condition* for BNs assumes every variable conditionally independent of its non-descendants non-parents given the parents. Accordingly, the joint PMF associated with a BN is such that $P(\boldsymbol{x}) := \prod_{i=1}^{n} P(x_i|\pi_{X_i})$, for each $\boldsymbol{x}$, where the values of x_i and π_{X_i} are those consistent with $\boldsymbol{x}$.

Questions modeling. The above joint probabilistic model describes the uncertainty about the skills of a student prior to his/her answers to the questions. To evaluate the student we formulate a number of *questions*, described as a collection of variables $\boldsymbol{Y} := (Y_1, \ldots, Y_m)$. We assume these variables to be Boolean, with the true value corresponding to the correct answer.[1] We call *background* of a question the set of skills "required" to answer it. This can be regarded as a conditional independence statement: given the background skills, the answer to the question is independent of the other skills and of the other questions. Following the Markov condition, this can be modeled by representing each question as a leaf node whose parents are the background skills. Such augmented graph requires the quantification, for each $Y_j \in \boldsymbol{Y}$, of a conditional PMF $P(Y_j|\pi_{Y_j})$ for each value π_{Y_j} of the background skills Π_{Y_j}. This procedure defines a BN over the skills and the questions, and hence a joint PMF $P(\boldsymbol{X}, \boldsymbol{Y})$.

Non-adaptive testing. Let $\boldsymbol{Y} = \boldsymbol{y}$ denote a student's answers to the test. In the above considered framework, the posterior knowledge about the skills is modeled by the joint PMF $P(\boldsymbol{X}|\boldsymbol{y})$. By running standard BN updating algorithms, the most probable level $\tilde{x}_i$ of skill X_i can be therefore evaluated as $\tilde{x}_i := \arg\max_{x_i} P(x_i|\boldsymbol{y})$, for each $X_i \in \boldsymbol{X}$. This reflects a non-adaptive, probabilistic approach to student evaluation.

Adaptive testing. To add adaptiveness to the above approach, every question should be chosen on the basis of the previous answers. As the goal is to gather information about the student skills, we evaluate the expected *information gain* (IG, i.e., the change in information entropy) associated with each possible new question, and pick the one maximizing this measure. The entropy of a BN over $\boldsymbol{X}$ can be computed as $H(\boldsymbol{X}) := \sum_{i=1}^{n} H(X_i|\Pi_{X_i})$ [3], where $H(X_i|\Pi_{X_i}) := \sum_{\pi_{X_i}} H(X_i|\pi_{X_i})P(\pi_{X_i})$ is the *conditional entropy* for X given its parents and $H(X_i|\pi_{X_i})$ is the entropy of the conditional PMF $P(X_i|\pi_{X_i})$.[2] Let $\boldsymbol{Y} = \boldsymbol{y}$ denote

[1] Extension to non Boolean answers is trivial as all answers Y_i are *manifest* variables, and, thus, Y_i can be always regarded as a binary variable with the two values denoting the observed answer y_i and its negation [15].

[2] To have entropy levels between zero and one, we define the entropy of the PMF $P(X)$ as $H(X) := -\sum_x P(x) \log_b P(x)$, with b number of states of X.

the answers to the questions already asked and $\boldsymbol{Y}'$ the set from which the next question should be picked. If the answer to every question $Y' \in \boldsymbol{Y}'$ would be known, and denoted by y', the question $\tilde{Y}' \in \mathbf{Y}'$ to choose would be the one leading to the largest IG. Yet, as the decision has to be made before the student's answer, conditional entropy should be considered instead, i.e.,

$$\tilde{Y}' := \arg\max_{Y' \in \boldsymbol{Y}'} \left[H(\boldsymbol{X}|\boldsymbol{y}) - H(\boldsymbol{X}|Y', \boldsymbol{y}) \right] . \tag{1}$$

CATs should also decide when to stop asking questions. Again, entropy can be used as a measure to decide when the current evaluation is sufficiently informative, i.e., we stop the test if the skills entropy given the answers is below some threshold $\tilde{H}$. The overall approach is depicted in Fig. 1.

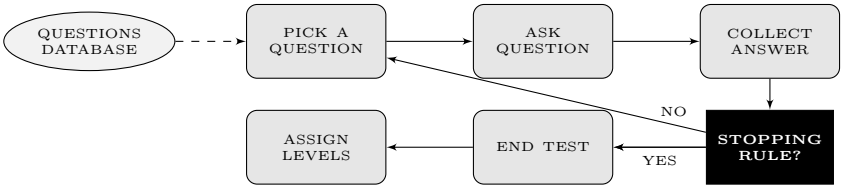

Fig. 1. CAT procedure

3 Adaptive Testing by Credal Networks

Credal sets and credal networks. A set of PMFs over X_i is called here *credal set* (CS) and denoted as $K(X_i)$. We always remove the inner points (i.e., those corresponding to convex combinations of the others) of a CS. CNs [16] are generalized BNs whose local PMFs are replaced by CSs. The BN defined in the previous section over the skills $\boldsymbol{X}$ and the questions $\boldsymbol{Y}$ becomes a CN if we replace with CSs the skill-to-skill and skill-to-question conditional PMFs. A joint CS $K(\boldsymbol{X}, \boldsymbol{Y})$ is consequently obtained as the collection of all the joint PMFs induced by BNs whose parameters take their values from the corresponding CSs, i.e.,

$$K(\boldsymbol{X}, \boldsymbol{Y}) := \left\{ P(\boldsymbol{X}, \boldsymbol{Y}) \left| \begin{smallmatrix} P(\boldsymbol{x},\boldsymbol{y}) := \prod_{i=1}^{n} P(x_i|\pi_{X_i}) \cdot \prod_{j=1}^{m} P(y_j|\pi_{Y_j}), \\ P(X_i|\pi_{X_i}) \in K(X_i|\pi_{X_i}),\, P(Y_j|\pi_{Y_j}) \in K(Y_j|\pi_{Y_j}) \end{smallmatrix} \right. \right\} , \tag{2}$$

where the values of x_i, π_{X_i}, y_j, π_{Y_j} are those consistent with $\boldsymbol{x}$ and $\boldsymbol{y}$.

Expert knowledge modeling. For a reliable expert knowledge modeling, we use CSs induced by *probability intervals*. Qualitative judgments about the probability of a state are converted in interval constraints such as $l \leq P(x) \leq u$, with the interval $[l, u]$ capturing the expert knowledge behind the judgment in a more reliable way than a sharp assessment. The CS consistent with these constraints is eventually obtained by standard polyhedral algorithms. Verbal to interval-numeric scales such as that Tab. 2 are used. For instance, if for the probability of the true state of the Boolean variable Y the expert judgment is "very likely", the corresponding linear constraint is $.2 \leq P(Y = \text{true}) \leq .4$.

Non-adaptive testing. Given the answers $\boldsymbol{y}$ to the questions $\boldsymbol{Y}$, we evaluate the student as in the previous section by updating the marginal probabilities of each skill. With CNs, these posterior values are set-valued and their characterization can be provided by lower and upper bounds, say $\underline{P}(X_i|\boldsymbol{y})$ and $\overline{P}(X_i|\boldsymbol{y})$ for each $X_i \in \boldsymbol{X}$. CN updating algorithms can eventually compute these bounds. The task displays higher complexity than in the case of BNs (e.g., exact inference in non-binary singly-connected CNs is NP-hard [11]), but approximate techniques can be considered when exact inference is unfeasible [17].

To compare the posterior intervals and decide the actual level of the student we might adopt the (conservative) *interval dominance* criterion [9], which rejects a level if its upper probability is smaller than the lower probability of some other level. Overlaps between intervals might therefore induce a situation of *indecision* between two or more levels. This is a so-called *credal* classification of the student level [18], and it represents the fact that students answers are somehow contradictory or not informative enough to provide a sharp decision. Interval dominance can return unnecessarily imprecise results. *Maximality* is a more refined criterion that rejects the levels which are less probable than another level for all the elements of the CS [7]. Maximality can be reduced to multiple updating tasks on auxiliary binary leaf nodes defined for each pair of states [17].

Adaptive testing. To achieve CAT with CNs using entropy as measure of informativeness for PMFs, as in the BN approach of Sect. 2, computation of entropies should be extended to CSs. This topic has been the subject of much discussion [10]. A cautious approach [19] consists in taking the upper entropy $\overline{H}(\boldsymbol{X})$, i.e., the entropy of the most entropic PMF in the convex closure $\overline{K}(\boldsymbol{X})$ of $K(\boldsymbol{X})$. In our framework, we should, then, look for maximum values of conditional entropies, such as $\overline{H}(X_i|Y', \boldsymbol{y})$ or $\overline{H}(X_i|\Pi_{X_i})$, as conditional entropies are required to compute both: (i) the joint (unconditional) entropy $H(\boldsymbol{X})$ (and its posterior values); and (ii) the conditional entropies involved in the question selection in Eq. (1). By definition a conditional entropy is a convex combination (whose weights are the elements of a marginal PMF) of convex functions (the entropies). The objective function might, then, be non-convex, as the weights are also optimization variables.[3]

Then, to bypass this non-convex optimization task, we compute (i) by separately considering the entropies of each skill $X_i \in \boldsymbol{X}$. This is analogous to the marginal approach commonly considered in multi-label classification to minimize Hamming losses [20]. The issue (ii) is more challenging. We consider the following upper approximation of $\overline{H}(X_i|Y', \boldsymbol{y})$:

$$\overline{\overline{H}}(X_i|Y', \boldsymbol{y}) = \max_{P(y'|\boldsymbol{y}) \in \{\underline{P}(y'|\boldsymbol{y}), \overline{P}(y'|\boldsymbol{y})\}} \sum_{y' \in \{\text{true,false}\}} \overline{H}(X_i|\boldsymbol{y}, y') P(y'|\boldsymbol{y}) \,, \quad (3)$$

[3] E.g., if $f(x)$ and $g(x)$ are convex functions of x, $h(x, y) := yf(x) + (1-y)g(x)$ is not convex even for $0 \leq y \leq 1$.

where the bounds of $P(y'|\boldsymbol{y})$ are obtained by standard CN updating algorithms. The problem thus reduces to the computation of upper entropies as

$$\overline{H}(X_i|\boldsymbol{y}) := \sup_{P(X_i|\boldsymbol{y})\in\overline{K}(X_i|\boldsymbol{y})} H(X_i|\boldsymbol{y})\,, \tag{4}$$

where $\overline{K}(X_i|\boldsymbol{y})$ is the posterior CS after conditioning on the observed answers $\boldsymbol{y}$. If $\overline{K}(X_i|\boldsymbol{y})$ has a finite number of non-inner points, this is a linearly-constrained convex optimization whose solution typically corresponds to either the uniform PMF or a non-inner point on the frontier of $\overline{K}(X_i|\boldsymbol{y})$. A numerical solution can be easily found by a simple iterative approach in the special case of CS specified by probability intervals [19]. We have therefore computed the posterior lower and upper bounds of $P(X_i|\boldsymbol{y})$, and then maximized the entropy with respect to those bounds. The procedure induces an outer approximation of $\overline{K}(X_i|\boldsymbol{y})$, and hence the upper approximation of the maximum entropy $\overline{\overline{H}}(X|\boldsymbol{y}) \geq \overline{H}(X|\boldsymbol{y})$. Finally, to generalise Eq. (2) to CNs, we define the information gain provided by a question Y' for its background skill $X_{Y'}$ as $\overline{\overline{H}}(X_{Y'}|\boldsymbol{y}) - \overline{\overline{H}}(X_{Y'}|Y',\boldsymbol{y})$ and select the question $\tilde{Y}'$ leading to the maximum information gain, i.e.,

$$\tilde{Y}' := \arg\max_{Y'\in\boldsymbol{Y}'} \left[\overline{\overline{H}}(X_{Y'}|\boldsymbol{y}) - \overline{\overline{H}}(X_{Y'}|Y',\boldsymbol{y})\right]. \tag{5}$$

For the stopping criterion, as we do not consider the joint entropy over the skills, we separately require each $\overline{\overline{H}}(X_i|\boldsymbol{y})$ to be smaller than a threshold $\tilde{H}$. To be consistent with this choice, we remove from the set of questions to be selected, those whose background skills already satisfy this condition.

Note that the use of an outer approximation of the upper entropy affects only the question selection process (eventually making it sub-optimal), whereas it has no effect on the student evaluation given a set of answers.

4 Application to Language Assessment

Before the academic year begins, the students of the University of Applied Sciences and Arts of Southern Switzerland (SUPSI) are asked to take an online German language placement test with 95 questions. In years 2015 and 2016, the answers of 451 students to all the questions have been collected. This benchmark is used to simulate CATs based on BNs and CNs as described in Sects. 2 and 3.

Model elicitation. Four skills are assessed: *Wörtschatz* (X_1, vocabulary), *Kommunikation* (X_2, communication), *Hören* (X_3, listening), and *Lesen* (X_4, reading). For each skill the student is assigned to a knowledge level compliant with EU guidelines.[4] Levels A1, A2, B1, and B2 are considered, and skills are therefore modeled as quaternary variables. Teachers associate each question with a single skill, which is set as the unique background skill of the question. The

[4] `http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf`.

number of questions associated with $X_1/X_2/X_3/X_4$ is 26/24/30/15. The current evaluation method assigns levels by setting thresholds on the percentage γ of correct answers on each skill (A1 if $\gamma < 35\%$, A2 up to 55%, B1 up to 75%).[5]

We first elicit from the teachers the structure of the BN/CN graph over the skills. The result is a chain, which is augmented by leaf nodes modeling the questions, each having its background skill as single parent. Overall, a tree-shaped topology as in Fig. 2 is obtained. This makes exact inference in the BN fast, while in the CN a variable elimination might be slow (a minute for query in our setup). A faster approximate CN algorithm is therefore used [17].

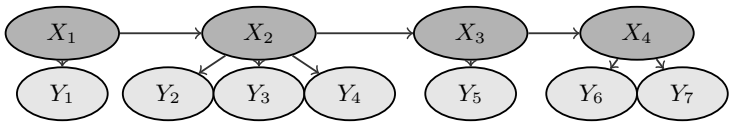

Fig. 2. A directed graph for CAT.

Teachers report their knowledge about the unconditional states of X_1 and the conditional states of X_i given X_{i-1}, for $i = 2, 3, 4$, as qualitative judgments (top of Tab. 1). To simplify the elicitation, the probabilities $P(X_i|X_{i-1})$ are given the same verbal judgment for all $i = 2, 3, 4$. A more detailed model could provide more accurate evaluations but it would be very hard for the domain expert to elicit it in a reliable way. Also, questions are divided by the teachers in three groups, corresponding to different difficulty levels. Questions in the same group are quantified in the same way, irrespective of their background skill, giving the judgments reported in the bottom part of Tab. 1.

X_1	$P(X_1)$
A1	*improbable*
A2	*uncertain*
B1	*uncertain*
B2	*improbable*

$P(X_i\|X_{i-1})$	$X_{i-1}=$A1	$X_{i-1}=$A2	$X_{i-1}=$B1	$X_{i-1}=$B2
$X_i=$A1	*fifty-fifty*	*uncertain*	*improbable*	*impossible*
$X_i=$A2	*uncertain*	*fifty-fifty*	*uncertain*	*improbable*
$X_i=$B1	*improbable*	*uncertain*	*fifty-fifty*	*uncertain*
$X_i=$B2	*impossible*	*improbable*	*uncertain*	*fifty-fifty*

$P(Y=T\|X)$	$X=$A1	$X=$A2	$X=$B1	$X=$B2
Easy	*uncertain*	*fifty-fifty*	*expected*	*probable*
Medium	*improbable*	*uncertain*	*fifty-fifty*	*expected*
Difficult	*impossible*	*improbable*	*uncertain*	*fifty-fifty*

Table 1. Expert judgements.

[5] These data as well as the software used for the simulations are freely available at `http://ipg.idsia.ch/software.php?id=138`.

For the CN, those judgements are translated in interval constraints for the corresponding events on the basis of the verbal-numerical scale in Tab. 2. Different probability intervals are considered for skills and questions as they refer to events of different type. For instance, when the expert considers "*impossible*" for an A1 level student to know the answer to a difficult question, the student is assigned a probability between .175 and .2 of answering correctly, as the questions offer only four choices plus the option of giving no answer. Notice that, by doing so, we are not anymore assuming that all questions in the same difficulty group share exactly the same conditional PMFs (as done by the BN model), as PMFs of different questions can vary independently in the given intervals. This seems a more sensible assumption than that of the precise model. For the BN, the PMFs corresponding to the centers of mass of the CSs defining the CN are used. Numerical inferences in the BN are consequently included in the intervals computed with the CN.

Judgement	*impossible*	*improbable*	*uncertain*	*fifty-fifty*	*expected*	*probable*
Skills	1-10%	10-20%	20-40%	30-50%	-	-
Questions	17.5-20%	22.5-25%	30-35%	60-65%	75-80%	95-97.5%

Table 2. A verbal-numerical scale for probability-intervals elicitation.

Experimental results. BN and CN methods in their non-adaptive (NA) and adaptive (AD) versions are considered. *Accuracy*, i.e., the proportion of students to whom the test assigns the same level of the current evaluation method, describes BN performances. This measure cannot be used for the set-valued outputs of CN methods. In this case the u_{65} measure can provide a comparison with the accuracy [18]. If $\mathcal{L}$ is the set of levels assigned by the CN on a skill and L its cardinality, a *discounted* accuracy gives $1/L$ if $\mathcal{L}$ includes the true level and zero otherwise. The u_{65} is a concave reinforcement of this score based on risk-adverse arguments. Its underlying assumption is that acknowledging the indecision between more levels has larger utility than randomly choosing one of them (e.g., the teacher could set up further assessments in the undecided cases). Tab. 3 shows the NA comparison. In Fig. 3 (left), the BN-NA accuracy is separately evaluated on the determinate (light bars) and indeterminate (dark bars) instances, i.e. those for which, respectively, a single level or multiple levels are returned by the CN model. On average, CN-NA returns single levels in 37.25% of the cases and, if this is not the case, an average of 2.36 levels (3.22 with interval dominance) are returned.

In the AD case we also track the average number of asked questions. Results are in Fig. 3 (right). CN-AD (circles) is tested for different thresholds over the entropy (labels of the markers) against a version of BN-AD based on the joint entropy (triangles). Similar values are obtained by coping with marginal entropies.

Algorithm	Average	X_1	X_2	X_3	X_4
BN-NA (acc)	63.09%	67.56%	60.85%	75.84%	48.10%
CN-NA (u_{65})	65.37%	67.71%	66.67%	70.33%	56.76%

Table 3. Non-adaptive tests results.

We also allow the BN-AD method to return multiple levels by maximizing the expected u_{65} utility over any possible set of levels. This variant is called BN-AD' and the corresponding u_{65} measure is reported (squares).

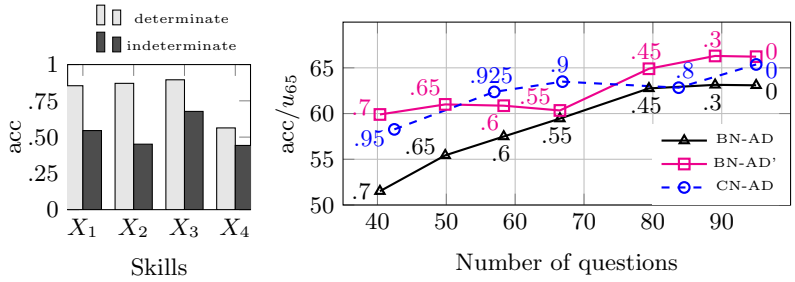

Fig. 3. Non-adaptive (left) and adaptive (right) tests performance.

As a comment, CNs seem to identify hard-to-evaluate students as those for which multiple levels are provided. In fact, the agreement between the BN and the traditional tests is larger when the CN test is determinate. As a consequence, the CN u_{65} measure is, on average, larger than the BN accuracy. A limitation of the CN test is the large fraction of indeterminate evaluations. One can interpret this result as a lack of robustness of the BN model, as even small variations in the model specifications can result in different decisions. Results also show that, both BN-AD and CN-AD approaches reduce the number of questions asked without significantly affecting the accuracy. BN-AD performances are improved by the "credal" variant BN-AD'. The results becomes very similar to those of the CN-AD. Yet, the latter method appears to be a more principled and suitable approach for a direct modeling of qualitative expert knowledge.

5 Conclusions and Outlooks

A procedure for adaptive testing built solely on expert knowledge has been proposed based on credal networks. The procedure has been validated on a real dataset about a German language test. Results are promising, as the credal approach simplifies the model elicitation, recognizes when a sharp decision about the student level should not be made (that is, when the traditional and precise

Bayesian evaluations disagree) and achieves an accuracy comparable to that of an indecisive Bayesian approach maximizing the expected u_{65} measure. However, the fraction of instances where CNs issue multiple levels remains rather large, therefore further research is needed to make CN-based CATs a viable solution for adaptive testing solely based on expert knowledge.

References

1. E. Pollard, J. Hillage, Exploring e-learning, Inst. for Empl. Studies Brighton, 2001.
2. H. Burns, C. A. Luckhardt, J. W. Parlett, C. L. Redfield, Intelligent tutoring systems: Evolutions in design, Psychology Press, 2014.
3. D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
4. R. G. Almond, R. J. Mislevy, L. Steinberg, D. Yan, D. Williamson, Bayesian networks in educational assessment, Springer, 2015.
5. M. Badaracco, L. Martínez, A fuzzy linguistic algorithm for adaptive test in intelligent tutoring system based on competences, Expert Syst. Appl. 40 (8) (2013) 3073–3086.
6. S. Renooij, C. Witteman, Talking probabilities: communicating probabilistic information with words and numbers, Int. J. Approx. Reason. 22 (3) (1999) 169–194.
7. P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, 1991.
8. A. Piatti, A. Antonucci, M. Zaffalon, Building knowledge-based systems by credal networks: a tutorial, in: A. R. Baswell (Ed.), Advances in Mathematics Research, Vol. 11, Nova Science Publishers, New York, 2010.
9. M. Troffaes, Decision making under uncertainty using imprecise probabilities, Int. J. Approx. Reason. 45 (1) (2007) 17–29.
10. G. Klir, M. Wierman, Uncertainty-based information: elements of generalized information theory, Vol. 15, Springer Science & Business Media, 1999.
11. D. Mauá, C. de Campos, A. Benavoli, A. Antonucci, Probabilistic inference in credal networks: new complexity results, J. Artif. Intell. Res. 50 (2014) 603–637.
12. R. K. Hambleton, H. Swaminathan, Item response theory: Principles and applications, Vol. 7, Springer Science & Business Media, 1985.
13. J. Vomlel, Building adaptive tests using Bayesian networks, Kybernetika 40 (3) (2004) 333–348.
14. M. Plajner, J. Vomlel, Bayesian network models for adaptive testing, arXiv preprint arXiv:1511.08488.
15. A. Antonucci, A. Piatti, Modeling unreliable observations in Bayesian networks by credal networks, in: Scalable Uncertainty Management (SUM 2009), Proceedings, Vol. 5785 of Lecture Notes in Computer Science, 2009, pp. 28–39.
16. F. G. Cozman, Credal networks, Artificial Intelligence 120 (2000) 199–233.
17. A. Antonucci, C. de Campos, M. Zaffalon, D. Huber, Approximate credal network updating by linear programming with applications to decision making, Int. J. Approx. Reason. 58 (2014) 25–38.
18. M. Zaffalon, G. Corani, D. Mauá, Evaluating credal classifiers by utility-discounted predictive accuracy, Int. J. Approx. Reason. 53 (8) (2012) 1282–1301.
19. J. Abellan, S. Moral, Maximum of entropy for credal sets, Int. J. Uncertain. Fuzz. 11 (05) (2003) 587–597.
20. A. Antonucci, G. Corani, The multilabel naive credal classifier, Int. J. Approx. Reason. 83 (in press) (2016) 320–336.