Learning Bayesian Networks for Student Modeling

Eva Millán*, Guiomar Jiménez, María-Victoria Belmonte & José-Luis Pérez-de-la-Cruz

ETSI Informática, University of Málaga, Campus de Teatinos, 29080 Málaga (Spain)

*eva@lcc.uma.es

Abstract: In the last decade, there has been a growing interest in using Bayesian Networks (BN) in the student modelling problem. This increased interest is probably due to the fact that BNs provide a sound methodology for this difficult task. In order to develop a Bayesian student model, it is necessary to define the structure (nodes and links) and the parameters. Usually the structure can be elicited with the help of human experts (teachers), but the difficulty of the problem of parameter specification is widely recognized in this and other domains. In the work presented here we have performed a set of experiments to compare the performance of two Bayesian Student Models, whose parameters have been specified by experts and learnt from data respectively. Results show that both models are able to provide reasonable estimations for knowledge variables in the student model, in spite of the small size of the dataset available for learning the parameters.

Keywords: student modeling, machine learning, Bayesian networks.

1 Introduction

In the field of student modelling, Bayesian Networks (in what follows, BNs) have been proposed to represent and compute student's features, and their use is now well established [1,2,3,4,5]. In the last decade, a number of *Bayesian Student Models* (BSMs) have been developed in a number of educational applications [6].

In order to define a BSM for a given field or task it is necessary to define both its structure (a graph) and a set of numerical parameters (prior probabilities of root nodes and conditional probabilities for the rest). While the structure can be easy to elicit, numerical parameters can become very difficult to estimate by teachers or human experts. A natural alternative is then to learn the parameters from a set of experimental data by means of machine learning techniques [7]. In fact, such techniques have already been successfully used in the context of the student modelling problem [8]. However, which approach will be more accurate? That is the research question addressed by this work: whether or not the performance of "learnt" BSMs is comparable to the performance of BSMs whose parameters are given by human "experts". The purpose of our research question is therefore to determine whether human judgment could safely be replaced by machine learning procedures in the task of specifying parameters for a Bayesian student model.

To our knowledge, this problem has not been discussed before. In an interesting review of student modeling approaches in the last decade [6] it is shown that machine learning techniques have been used for different aspects of the student model, but "no adaptive and/or personalized tutoring system has used a compound student model which brings together an overlay model with machine learning algorithms or Bayesian networks".

Our work is inscribed inside the Mathematics Education Project (Proyecto Matematica Ensino, PMatE) (http://pmate4.ua.pt/pmate/), which was launched in 1990 by the University of Aveiro (Portugal) with the aim of improving scientific education in schools [9]. The goal of this project was to invest in new information technologies for teaching and learning as a way to enrich, enhance and boost scientific education in Portugal. To this end, computer tools and contents for several areas of knowledge (especially for mathematics) have been developed. Since 1990 the project has been available in the web and includes materials both for formative and competition purposes. Every year PMatE promotes six National Mathematical Competitions in several science areas: one for each school degree, from Primary to Higher Education, and one for all school degrees in the network.

Besides of the motivational gains obtained by taking these tests, they can be a powerful tool for diagnosing students' skills and competences. From 2011 to 2013 a research project was conducted to determine whether or not the diagnostic capabilities of computerized tests in PMatE could be improved by using approximate reasoning techniques, namely Bayesian Networks. In this previous work, a BSM for first-degree equations was implemented and evaluated [10]. It was an instance of the generic BSM defined in a former research work [11]. The parameters of the network were computed according to the formula proposed in the generic model from a set of parameters that was specified by teachers (difficulty, guessing factor, slip factor, discrimination). In what follows we will call this model the *expert BSM*, to account for the fact that it relies entirely on expert (human) judgment. The performance of this expert BSM was evaluated with 152 students from two different schools in

Portugal. Results of the study showed that it was able to provide accurate diagnosis at various levels of granularity.

In the work presented here we have now used the data obtained in the former evaluation with the 152 real students to learn the parameters of the BN. The structure of the original BSM of the former study [10] had to be simplified, because otherwise it would have been impossible to learn such a big number of parameters (110) with so little data (152 students). The four simplified structures are based in granularity levels, and have been kept as similar as possible to the original structure. The proposed four structures have been used to diagnose student´s knowledge with the parameters of the network as specified by the experts participating in the former study, and with the parameters that have been learned in the study presented here. Then the results have been compared using the same techniques of the original study, namely Bland-Altman plots and confidence intervals.

Results of our study with the four structures show that the performance of the two models is acceptable but the expert BSM gets more accurate results than the learnt BSM, when compared to the estimations of the knowledge provided by the teachers. However, even with such a little set of data, it is possible to learn the parameters and obtain acceptable estimations of the knowledge of the student.

The paper is structured as follows: in the next section we will present the materials and methods used in this study: the expert BSM developed in previous works and the validation methods used (Bland-Altman plots [12] and confidence intervals [13]). Section 3 describes the results of our research, while Section 4 discusses their implications. The paper finishes with some conclusions and future lines of research.

2 Materials and Methods

2.1 Previous Work: the Expert Bayesian Student Model

In order to understand the present study, we need to briefly present the results obtained in the previous study [10]. The domain of application was first-degree equations. Figures 1 and 2 represent the expert BSM that was used in this study.

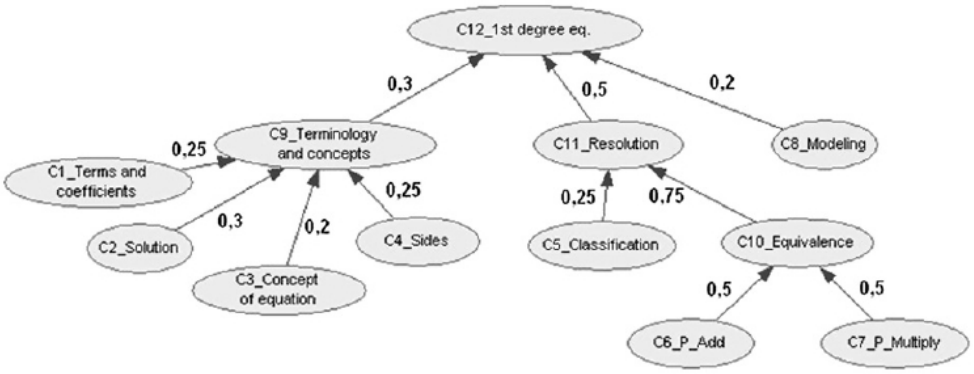

Fig. 1. Granularity relationships the expert BSM

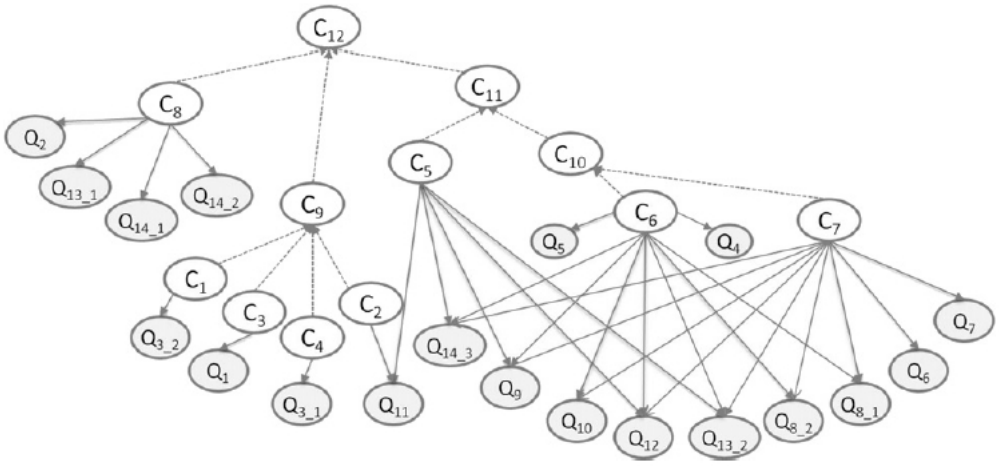

Fig. 2. Complete structure of the expert BSM (granularity + questions)

This BSM was developed by human experts, and includes two different types of variables:

- *Knowledge variables* (those labelled as C): these binary variables represent student's knowledge (1= the student knows, 0= the student does not know). They are not directly observable. As shown in Figure 1, they are grouped at four levels of granularity. In the study it was shown that the probability of a student knowing a concept could be used as an estimator of his/her degree of knowledge about that concept.
- *Evidence variables* (those labelled as Q): these variables represent student's answers to questions that were posed to students in a real exam (i.e., these variables are observable. Each variable is binary (1=correct answer, 0= wrong answer). For example, an actual question administered in the exam is:

$$Q_5\text{: Solve } 2 + x = 3.$$

Relations among variables are of two kinds: *aggregation relations* and *causal relations*. Aggregation arcs go from lower-level knowledge variables to higher-level knowledge variables. Figure 1 shows the knowledge variables and their aggregation relations. There are 12 knowledge variables, 8 at the lowest or elementary level, one at the top level (global knowledge of the whole subject, namely first degree equations) and three variables at intermediate levels. Causal relations go from lowest-level knowledge variables to evidence variables (answers). There are 19 evidence variables. The complete structure of this expert BSM is shown in Fig. 2.

Real answers from 152 students from two different schools were input to and processed by the expert BSM. In this way, the expert BSM was able to compute for each student the probability of knowing/not knowing each elementary or aggregated knowledge concept.

Independently of this BN computation, the student´s answers to the exam were also submitted to three different human experts (teachers). From their answers each human expert provided a grade (a number between 0 and 1) for each of the concepts involved. A high degree of agreement between experts was reported in this study [10], so the average of such values was taken as a reliable measure of the hidden variables (the student's state of knowledge for each concept).

To evaluate the performance of this model, the two measures were compared. That is, the posterior probability of knowing each concept was compared to the (average) grade estimated by the three teachers for the same concept. The validation method used in this study was the one proposed by Bland and Altman for clinical measurements of continuous variables [12]. Bland–Altman plots provide a good graphical visualization of the inter-rater agreement, and, together with confidence intervals for the mean difference, a good method to measure it [13].

2.2 Methodology for Learning the Parameters

Now we present the new experiments performed in the work presented here. As explained before, we have defined four different BN structures for learning the parameters in this experiment. All these four structures are simplifications of the expert BSM so that that all knowledge variables lie at the same level of granularity. That is, there are no aggregation relations. This simplification was taken for two reasons: a) the structure of the expert BSM implied the learning of a great number of parameters (110), while the number of observations was relatively small (152 students) and b) aggregation relationships in the original network do no represent causal rela-

tionships, therefore though their have their role in the original study, it makes no sense to learn them. The BN structures we have defined are:

- Structure 1 (lower level of granularity). Elementary knowledge variables (8 nodes) and all its causal arcs.
- Structure 2 (higher level of granularity). Global knowledge node and causal arcs to all question nodes.
- Structure 3 (intermediate level of granularity). Intermediate knowledge nodes (C_i) and causal arcs from C_i to every question node that depends on C_i or on any of its children.
- Structure 4 (intermediate level of granularity). Similar to structure 3, but choosing a different set of four intermediate knowledge variables.

The four structures are shown in Figures 3-6.

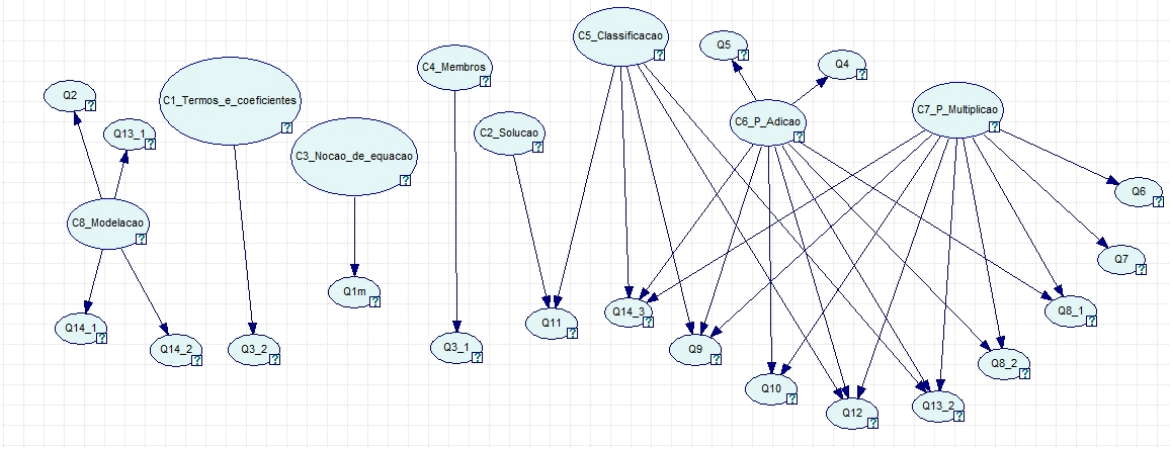

Fig. 3. Structure 1

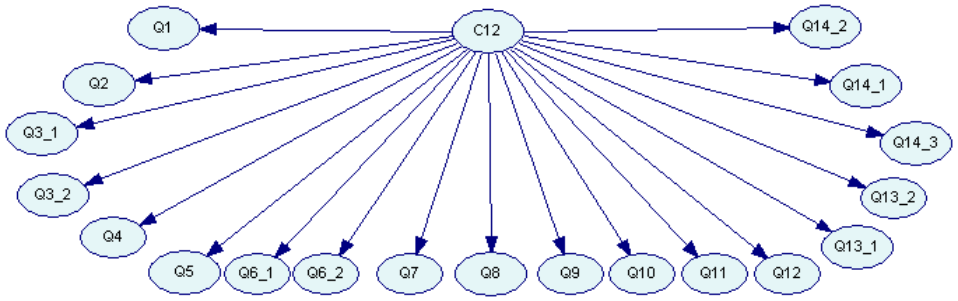

Fig. 4. Structure 2

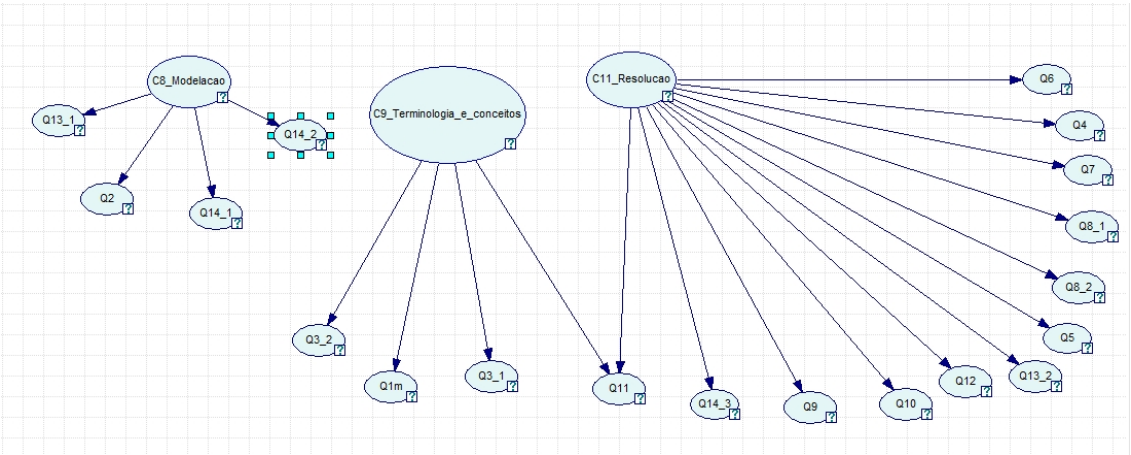

Fig. 5. Structure 3

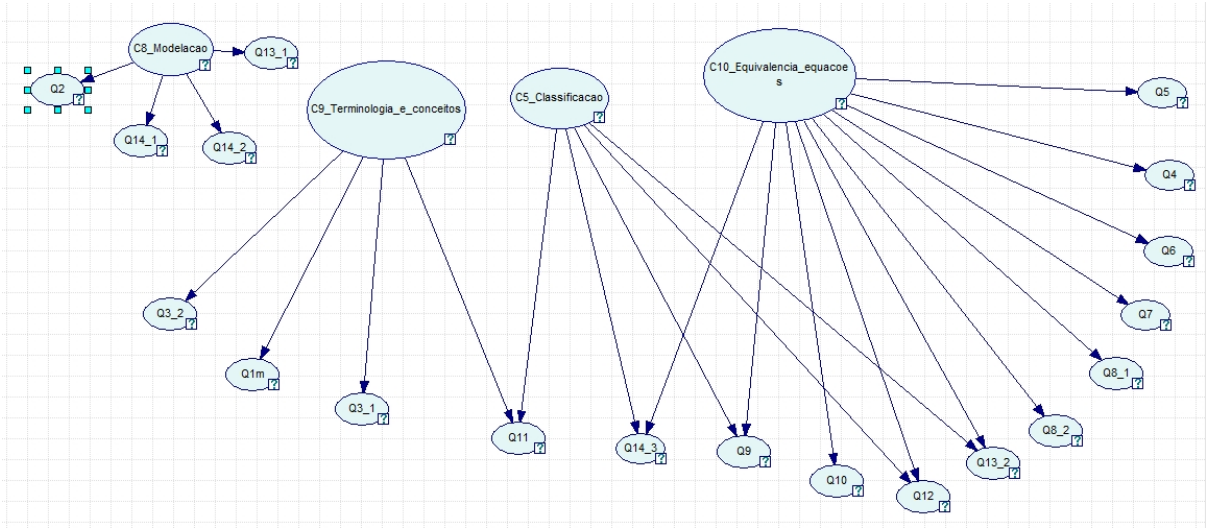

Fig. 6. Structure 4

For each of these structures we have learnt the parameters. The number of parameters to be learnt in each case is 77, 39, 43 and 52, respectively. They were computed by applying the EM algorithm [14] to the data available (152 students). The implementation of EM used in the experiment was the one provided by the free SMILE environment (http://genie.sis.pitt.edu/), developed at the University of Pittsburgh.

3 Results

In this section we present the results obtained in the experiments performed with each structure, replicating the experiment in the original study [10]. In each experiment i, the same structure i has been used for the two BSMs (i =1,2,3,4). The first BSM uses the parameters of the original study (expert-BSM), while the second BSM uses the parameters estimated by the EM algorithm (learnt-BSM).

Then, the student´s answers have been used as input for both BSMs (expert and learnt), and the posterior probabilities of knowing each concept computed. These probabilities have been then compared to scores obtained by averaging the three estimations given by teachers (the only reliable measure available for the unobservable variable "state of knowledge").

Therefore, and, as in the original study, the variables to be compared in this study are continuous with values from 0 to 1. In order to obtain a global measure of the performance of the BSMs we have computed the difference of the two values (deviations) and then the mean value, standard deviation and confidence intervals for such values.

Tables 1 and 2 show the results of the global comparison between "real" knowledge values (those given by the average of three teachers) and knowledge values as computed by the expert BSMs (Table 1) and learnt BSMs (Table 2) for each given structure. For every student and every knowledge variable, the difference between "real" and computed variables have been calculated, and the table shows the mean and standard deviation and the 0.05 confidence interval of such differences:

Experiment	μ (mean)	σ (st. deviation)	Confidence interval	Size of the conf. interval
1	0.01	0.18	(0.003, 0.017)	0.014
2	0.03	0.07	(0.020, 0.040)	0.020
3	0.03	0.17	(0.010, 0.050)	0.040
4	0.01	0.18	(-0.003, 0.026)	0.029

Table 1. Results for the expert BSMs.

Experiment	μ (mean)	σ (st. deviation)	Confidence interval	Size of the conf. interval
1	-0.05	0.54	(-0.084, 0.023)	0.107
2	-0.04	0.3	(-0.086, 0.009)	0.095
3	-0.25	0.34	(-0.286, -0.224)	0.062
4	0.04	0.34	(0.016, 0.070)	0.054

Table 2. Results for the learnt BSMs.

As in the original work, we have also used Bland–Altman plots [12] to obtain a visual representation of the degree of agreement between both estimations. In Bland-Altman plots, the x-axis value represents the average of the two scores, while the y-axis represents the difference. In this way, the closer the points are to the horizontal axis, the best results. Also, the lower and higher values of the abscissa represent concepts with lower and higher level of knowledge, respectively.

Figures 7-10 show Bland Altman plots for the sets of differences corresponding to the four BSM. The left half of every figure shows the diagram for the expert BSM and the right half shows it for the learnt BSM.

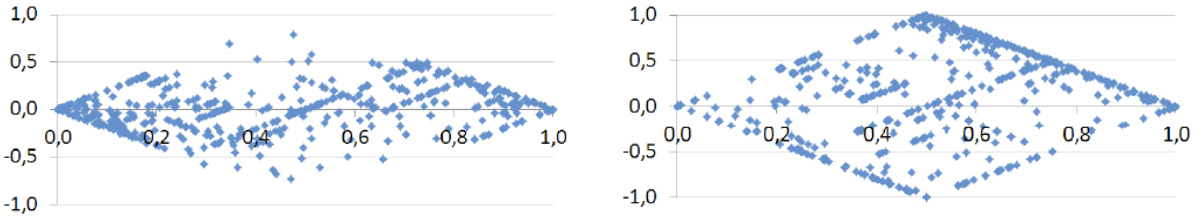

Fig. 7. Bland-Altman plots for Experiment 1

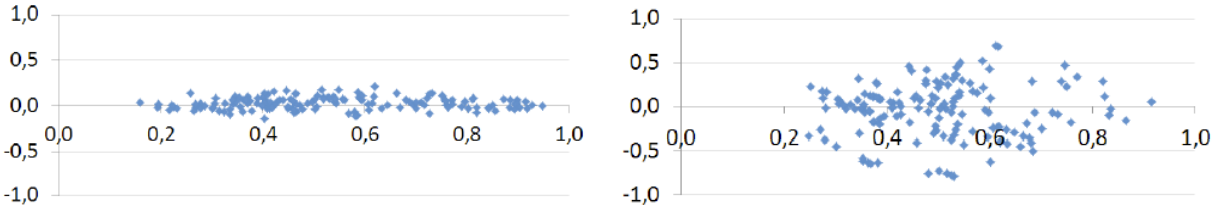

Fig. 8. Bland-Altman plots for BSMs of experiment 2

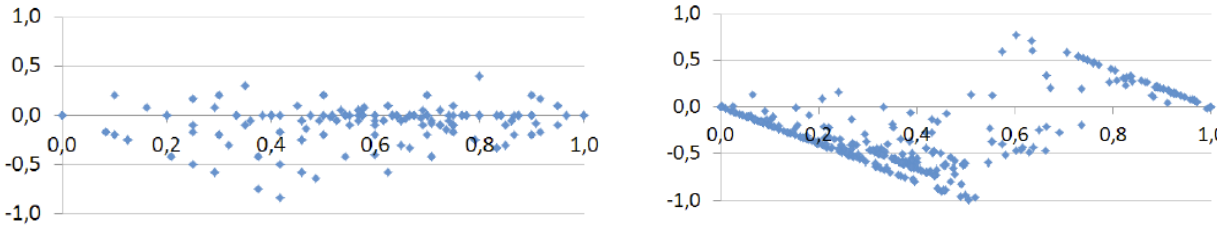

Fig. 9. Bland-Altman plots for BSMs of experiment 3

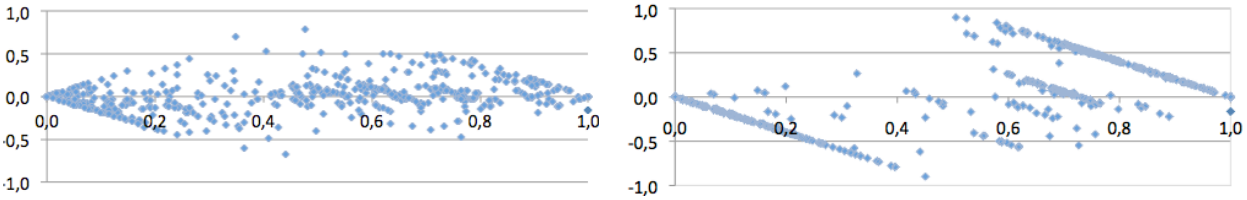

Fig. 10. Bland-Altman plots for BSMs of experiment 4

4 Discussion

In this section we will discuss the results presented in previous section.

4.1 Experiment 1

The average deviation of the learnt BSM is $\mu = -0.05$. That means that it BSM performs globally well, both in absolute terms and in relation with the expert BSM (average deviation $\mu = 0.01$) (Tables 1 and 2). However, this optimistic analysis is contradicted when we consider the values of σ and corresponding confidence intervals. For the learnt BSM it is $\sigma = 0.54$, while it is $\sigma = 0.18$ for the expert BSM. Also, the size of the confidence interval is much smaller for the expert BSM. Both measures confirm that dispersion of the learnt BSM is much bigger. If we consider the Bland-Altman plots (Figure 7), we can see that deviations are similar for concepts with extreme values (next to 0 or 1). However, for intermediate values (those whose probability of being known is 0.5, which are usually the more difficult to diagnose) the learnt BSM shows greater deviations.

This result is somehow discouraging for the learnt BSM, as we would have expected given the small number of data available. Probably the reason is that the number of parameters to be estimated is 77 and the data available is 152 cases (students), so it is difficult to obtain very accurate results for the "difficult" cases.

4.2 Experiment 2

The average deviation of the learnt BSM is $\mu = -0.04$. Again the learnt BSM performs globally well, both in absolute terms and in relation with the expert BSM (average deviation $\mu = 0.03$) (Tables 1 and 2). For the learnt BSM, the value of σ is now much better ($\sigma = 0.30$) but above that of the expert BSM ($\sigma = 0.07$, also below the previous one). Confidence intervals for the two measures are also relatively small. If we consider the Bland-Altman plots (Figure 8), we cannot see a clear pattern. Deviations are now great for some intermediate and for some extreme values.

As we would have expected, the performance of the learnt BSM is now "good enough". A posible reason for this better result that is the network is computing only one value (C12, which represents the global level of knowledge of each student). Another possible reason is that the number of parameters to be estimated is substantially smaller (39).

4.3 Experiment 3

The average deviation of the learnt BSM is $\mu = -0.25$. This value is very high and can be cosnsidered as an outlier when compared with the other experiments. On the contrary, for the expert BSM the average deviation has a value $\mu = 0.03$ similar to those of the other experiments (Tables 1 and 2). Concerning the values of σ, for the learnt BSM it is similar to that of experiment 2 and for the expert BSM it is similar to that of experiment 1. The sizes of the confidence intervals are again relatively small. The performance of this learnt BSM is a little surprising. But, if we consider the Bland-Altman plot (Figure 9) we can observe that some lines are formed, which is even a bigger surprise: we can see that the learnt BSM systematically underestimates variables "poorly known" (values below 0.5) and systematically overestimates variables "well known" (values over 0.5). And, since $\sigma<0$, underestimating is more intense that overestimating. However, notice that this performance is not necessarily undesirable. The learnt BSM is magnifying something that is really inside the student's answers.

4.4 Experiment 4

The average deviation of the learnt BSM is $\mu = 0.04$, similar to those of experiments 1 and 2. For the expert BSM the average deviation has a value $\mu = 0.01$, similar to those of the other experiments (Tables 1 and 2). Concerning the values of σ, for the learnt BSM it is $\sigma = 0.34$ (similar to those of experiments 2 and 3) and for the expert BSM it is $\sigma = 0.18$ (similar to those of experiments 1 and 3). Confidence interval sizes are quite small. If we consider the Bland-Altman plot (Figure 10), we can see a similar effect to the reported in experiment 4: the learnt BSM also underestimates variables "poorly known" (values below 0.5) and overestimates variables "well known" (values over 0.5). However, since $\sigma \sim 0$, underestimating and overestimating seem to be of similar strength. Again notice that this performance is not necessarily undesirable and, in this case, the magnifying effect is more "fair" that in experiment 3.

5 Conclusions and Future Work.

In this paper we present the results of a research concerning the automatic estimation of parameters for student models based on Bayesian Networks. We have defined four Bayesian Networks and, by using the algorithm EM and the software tool SMILE, we have performed four experiments of parameter learning. We have compared the diagnosis provided by the models with those provided by a pool of three human graders. We have

also compared these results with the performance of another Bayesian Network whose parameters were not automatically learned, but directly given by a team of human experts.

The performance of our learnt models has been different depending on the experiment. Best results were obtained for the model with just one hidden (or knowledge) variable, but reasonably good results were also obtained in other cases. In the case of the structures representing the intermediate levels of the granularity hierarchy, it seems that the learnt BSMs tend to polarize the results when compared to the more uniform estimation provided by the expert BSM. All in one we think that the answer to our research question is positive: the performance of "learnt" BSMs is comparable to the performance of BSMs whose parameters are given by human "experts".

However we must recognize that none of the four learnt models exhibited better performance than the human-adjusted models. Probably this result is due to the limitations inherent to this study. The first limitation is the size of the training set (152 cases) compared to the number of parameters to be learnt (ranging from 39 to 77 depending on the experiment). But it is difficult to obtain larger datasets to work with, given that each student had to take the exam and that three different tutors graded each exam. The second limitation is the problem that arises in some of the structures, when a hidden variable depends on a single evidence variable that takes just one value in the training set; in this case, it is impossible that an algorithm can learn the parameters. Finally, another limitation is that the real state of knowledge of the student is unknown; therefore we are trying to infer variables that are intrinsically hidden. To avoid this problem we made the same assumption than in the former study, which assumed that the average of the scores given by the three teachers that independently could provide a reliable estimation of the "real" values for the hidden (knowledge) variables. However, tough this is a convenient assumption, perhaps this is not entirely true.

All in one, we think that results presented here can be considered as a first encouraging step towards the use of machine learning techniques in overlay models, which is still an open problem. Even with such a small dataset the parameters can be learned and the learnt BSMs provide reasonable estimations of student's knowledge. If the same techniques are applied to larger datasets, it seems reasonable to assume that the results would be much better.

Regarding future work, we plan to use large databases in which we can apply this kind of techniques for parameter learning. If such datasets are large enough, techniques of structural learning can be also applied to this kind of student models.

6 References

1. Collins, J. A., Greer, J. E., and S. H. Huang: Adaptive assessment using granularity hierarchies and Bayesian nets. In Proceedings of ITS'96. LNCS 1086: 569–577 (1996)
2. Jameson, A.: Numerical uncertainty management in user and student modelling: an overview of systems and issues. User Modelling and User-Adapted Interaction, 5:193–251 (1996)
3. Conati, C., Gertner, A., VanLehn, K., and M. Druzdzel: On-line student modelling for coached problem solving using Bayesian networks. In Proceedings of UM'97: 231–242 (1997)
4. VanLehn, K., Niu, Z., Siler, S., and A. S. Gertner: Student modelling from conventional test data: A Bayesian approach without priors. In Proceedings of ITS'98. LNCS 1452: 434–443 (1998)
5. Millán, E., Loboda, T., and J. L. Pérez-de-la-Cruz: Bayesian networks for student model engineering. Computers and Education 55(4): 1663–1683 (2010)
6. Chrysafiadi, K. and Virvou, M. Student modelling approaches: A literature review for the last decade. Expert Systems with Applications, 40(11): 4715-4722 (2013)
7. Neapolitan, R. E.: Learning Bayesian Networks. Upper Saddle River, NJ, Prentice-Hall (2003)
8. Arroyo, I, and Woolf, B.P. Inferring learning and attitudes from a Bayesian Network of log file data. Proceedings of AIED'05: 33-40 (2005)
9. Sousa Pinto, J., Oliveira, P., Anjo, B., Vieira, S. I., Isidro, R. O., and M. H. Silva: TDmat-mathematics diagnosis evaluation test for engineering sciences students. International Journal of Mathematical Education in Science and Technology, 38(3), 283–299 (2007)
10. Millán, E, Descalco, L., Castillo, G., Oliveira, P., and Diogo, S.: Using Bayesian networks to improve knowledge assessment. Computers and Education 60(1): 436–447 (2013)
11. Millán, E., and J. L. Pérez-de-la-Cruz: A Bayesian Diagnostic Algorithm for Student Modelling and its Evaluation. User Modelling and User-Adapted Interaction 12: 281–330 (2002)
12. Bland, J. M., and D.G. Altman: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 8476(327): 307–10 (1986)
13. C. Hamilton, C, Stamey, J. Using Bland-Altman to assess agreement between two medical devices – don't forget the confidence intervals. Journal Clinical Monitoring and Computing: 331–333 (2007)
14. Dempster, A. P., Laird, N. M., and Rubin, A.D.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B: 39(1): 1–38 (1977)