Learning parameters for a knowledge diagnostic tools in orthopedic surgery

S. LALLÉ
Joseph Fourier University, France

AND

V. LUENGO
Joseph Fourier University, France

We propose and illustrate a methodology for taking into account data for a knowledge diagnostic tool in orthopaedical surgery, using Bayesian networks and machine learning techniques. We aim to make the conception of the diagnostic system less time-consuming and subjective, particularly in complex and ill-defined domain such as medicine. As in our domain, conceptions may be either empirical or tacit (i.e. instinctive, or learned but hard to express) so that the diagnostic often is unsure even for the experts. Thus, a first Bayesian network was built like an expert system for modeling this incertitude, where experts (in didactic and surgery) provide both the structure and the probabilities. However, learning the probability distributions of the variables allows going from an expert network toward a more data-centric one. We compare and analyze here various learning algorithms with regard to experimental data, and compute different recognition rates. Then we point out some crucial issues like the lack of data.

Key Words and Phrases: Knowledge diagnostic, machine learning, Bayesian network

1. INTRODUCTION

TELEOS (Technology Enhanced Learning Environment for Orthopaedical Surgery) is an Intelligent Tutoring System designed for the percutaneous orthopedic surgery [Vandcard and Luengo 2004]. In this platform students can practice different interventions with a simulator, which propose 3D and haptic feedbacks. The project focused on minimally invasive procedures that imply two different notions: gain access to the bones (the gesture) and indirect observation through image guidance. The TELEOS’ simulator recreates these two aspects for every kind of operation [Larcher et al. 2010].

Using the data record by the simulator, TELEOS performs a knowledge diagnostic in order to identify a set of conceptions and provide a personalized feedback to the student (like hints or another exercise). This paper is focused on the knowledge diagnostic system. In our ill-defined domain, knowledge may be either empirical or tacit (i.e. instinctive, or learned but hard to express) so that the diagnostic often is unsure even for the experts. Thus, this incertitude was modeled in TELEOS with a dynamic Bayesian network [Minh Chieu et al. 2010]. The network was built like an expert system, where experts (in didactic and surgery) provide both the structure and the probabilities. In this paper, we try to build the network in a more automated way, using machine learning methods for computing the probabilities (the parameters of the Bayesian network), in spite of our complex domain.

Numerous advantages of using such automatic techniques may be highlight. First, get probabilities that are more objective and less dependent on the experts, especially on their
various points of view since learning algorithms will blend them. It seems particularly interesting in unformalized domain like surgery, the activity of the student may show misinterpretations. The learning usually relies on data annotated by the experts, but the latter are absolutely not involved in the design of the diagnostic tool itself. More, machine learning methods can deal with misses or lack of data.

Secondly, this method is less time-consuming. Indeed, our Bayesian network is very large with a lot of nodes. In such cases, experts can roughly estimate all the parameters and are prone to error or approximation. Such mistakes are hard to fix without the help of another experts. Automatic methods allow to test and compare various approaches too, for example different student model or different kind of data. In particular, we wish take into account raw data coming from the robot that may provide many hints about the knowledge state of the students. Finally, we hope we could refine the diagnostic.

2. STATE OF THE ART
Model the reasoning and the knowledge of students with Bayesian networks (i.e. causal networks) already exist in the field of Intelligent Tutoring System. We can restrict three general approaches for designing these models, like propose by Mayo and Mitrovic [2001].

First, in the expert-centric approach an expert specifies either directly or indirectly the complete structure and conditional probabilities of the Bayesian student model. The pointed disadvantage is that the resulting models may include so many variables that it becomes infeasible to evaluate the network effectively on-line. Another classical disadvantage is the difficulty to define conditional probabilities. For example, tractability testing was an important issue in the initial evaluation of DT-Tutor [Murray and VanLehn 2000]. In our case, we can observe the difficulty to calibrate the system with this kind of approach.

For the efficiency-Centric methodology the model is this time partially specified or restricted in some way, and domain knowledge is “fitted” to the model. The restrictions are generally chosen to maximise some aspect of efficiency, such as the amount of
numeric specification required and/or the evaluation time. Like introduced by the authors, the disadvantage is that in general, restrictions to increase efficiency can introduce incorrect simplifying assumptions about the domain.

Finally, in the data-Centric approach, explained by Mayo and Mitrovic, the structure and conditional probabilities of the network are learned primarily from data. This class of model dispenses with attempting to model unobserved student states, such as their domain mastery, and instead concentrates to modeling the relationships between observed variables to predict student performance. In their works they observed that the problem is that the definition of observables for the model is relevant only with a large number and variety of observables. The data centered approach is still recent in ITS [Mayo and Mitrovic 2001] because it imply data mining and learning methods whose validation is complicate. However, the advantages are showed in several non vocational domains.

Nevertheless, one may want to model hidden variables as well as perform a deep analysis of the domain or reuse a similar analysis for favoring semantic senses of the network. The data-centric approach studied by Mayo and Mitrovic doesn’t really fit this more classical methodology. Lastly, machine learning techniques and data are rarely used for an expert system in order to go towards a more data-centric approach. We presumably can explain that by the need of data before the implementation. But as data are nowadays shared, structured, enriched and finally crucial, take them into account during the build step of the diagnostic tools along with a domain model is the main idea of our work.

In addition, the robot itself records haptic information, like the speed and position of the instrument in the bones, which was not taken into account in the previous network. Consider them seems, however, interesting, given that we may be able to use those uninterpreted haptic data for diagnosing the underlying cognitions. This study so consider a refined network, which take into account complex haptic information. Similar works were performed by Fournier-Viger and Nkambou [2008] within a simulation-based tutoring system to teach astronauts how to manipulate a robot deployed on the International Space Station. In particular, the tutor learns patterns of actions and does some clustering for actions with parameters (angles of rotation for instance). However, as far as we know, it is something new in the field of student modeling and epistemic knowledge diagnostic.

In this paper, we first briefly introduce the TELEOS project, particularly the diagnostic, in the following section. Then we detail the parameters learning algorithms for Bayesian network. At least, we present the data used for the learning, which are very expensive in our domain. Then, we present some results, which lead to a discussion and the conclusion.

3. CONTEXT: THE DIAGNOSTIC IN TELEOS

The project is principally aiming to promote the learning of percutaneous surgery. There are three types of knowledge that are at stake during this learning activity: declarative, pragmatic (often empirical) and perceptivo-gestural. The objective of the system is to let the learner train himself freely on a percutaneous orthopedic surgery in order to give him an epistemic feedback according to his actions. The feedback accompanies the subject in the learning process, by provoking reinforcements, destabilizations, hints, scaffolding, etc.

By investigating the empirical, descriptive and gestural cognitions, a student model that fits the domain was built as detailed in a previous work [Minh Chieu 2010]. Compute the knowledge state allows determining an appropriate feedback [Luengo et al. 2009]. On the other hand, we need a student model in order to diagnose the knowledge. In TELEOS, a model was built based on a didactic analysis of the domain and formalized using the
cKc model. This one is mainly concerned by the representation of the conceptions throughout four sets [Balacheff 1995]:

- A set P of problem
- A set O of operator. The students use them for solving a problem of P (in our domain, operators always are actions)
- A set L of representation systems (for example, X-rays)
- A set $\sum$ of controls (or cognitions). A control is used by the student in order to check the result of an operator.

We diagnose the controls in order to determine if they are brought into play in a valid way or not. An example is given in the table 1 for illustrating the cKc model.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Operator</th>
<th>Representation</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix a thoracic vertebra</td>
<td>Take a front X-ray</td>
<td>X-rays (2D)</td>
<td>The pedicles must be centered on the X-rays.</td>
</tr>
<tr>
<td>Fix a lumbar vertebra</td>
<td>Impact the vertebra</td>
<td>Anatomy (3D)</td>
<td>If the surgical instrument is below the pedicles, then it damages the foramens.</td>
</tr>
</tbody>
</table>

We get an expert-centric network fully based on the didactic analysis (figure 2). The probabilities themselves were given by the experts by mean of questionnaires.

Perform the diagnostic requires to know what the student is doing. In addition to the four kinds of vertices derivate from the cKc model figure the situation variables (in green on the figure 2) which were identified during the didactical analysis. They provide discrete information such as “is the chirurgical instrument outside of the vertebras?”

![Fig. 2. Detail of the Bayesian network for the knowledge diagnostic. The cognitions (or controls) are in grey, the problem in yellow, the operator in blue and the situation variables in green. At least, the representation systems figure in pink.](image-url)
More recently, we have modified the structure of our network for including haptic information, concerning the speed, the force and the position of the instrument. It may be interesting to take them into account, because their accuracy is high and also because this data can help to understand the tacit knowledge in orthopedic surgery. For example, the body of a vertebrae has a very low density compared to its frontier; thus, if the student progress too fast, he can "slide" and cause damages. However, these features are continuous in the time, and present a lack of genericity (depending on the patient or the pathologies). Concretely, we discretized them thanks to data-mining methods and add haptic nodes in the network (see figure 3) [Lallé and Luengo 2011]. More specifically, we compute clusters for the force and the speed, and a grid-segmentation of the vertebrae for the trajectory. Even if the results of the diagnostics with and without haptic data are substantially the same, the haptic network allows to anticipate (predict) the state of the students’ cognitions. For example, if the student applies a strong force in a vertebra with low bone density, the probability that he will make a mistake further is higher; haptic information may give significant help for this kind of situation.

Also, we don’t need expert analysis for building the haptic network with this methodology. However, because the experts were not involved at this level and didn’t estimate the parameters, the result of our data-mining approach is of little semantic meaning and increases the network complexity (due to the large domain of the variables). It would probably be hard for the experts to estimate the corresponding probabilities in the network. Thus, it is an other motivation for us for studying machine learning tools.

The haptic vertices are bound to cognitions which depend on the haptic behavior (for instance, the student should not force in low-density parts of vertebrae). The experts were not involved at this level and didn’t estimate the parameters. Indeed, the result of our data-mining approach is of little semantic meaning and increases the network complexity. Learning parameters appears here to be a pertinent idea again. On the other hand, that also may reduce misinterpretations, increase the flexibility and refined the diagnostic

![Figure 3. Global structure of the network improved with haptic nodes](image)

**4. LEARNING PARAMETERS**

Given data, the parameters of a Bayesian network can be learned, i.e. computed from a base of observations. Thus, we can find in the literature some algorithms designed for learning the parameters of a Bayesian network. However, they present different
characteristics and have both strong and weak points – no guarantee of results can be offered anyway [Heckerman 1995]. The idea here is to compare these methods.

Learning for Bayesian network mainly consists in count observations in the database. For instance, suppose that the students didn’t align the pedicles on the X-rays in 60% of the time. Then we basically assume that the corresponding cognitions in the network would be brought into play in an incorrect way with a probability of 0.6. Just counting facts in the database is the easier way to learn parameters. More formally, the distribution $\theta$ is express in function of the database $D$ as following for Bayesian networks [Heckerman 1995]:

$$\theta = \arg\max \theta \ p(D \mid \theta)$$

$$\Leftrightarrow \theta = \frac{Ni, j, k}{\sum_j (Ni, j, k)}$$

with $Ni, j, k$ the number of entries in the dataset $D$ where the node $i$ is in his state $j$, and knowing that his parents are in the configuration $k$. Thus actually, it is just a count. Moreover, the degrees of belief collected alongside the data are involved here. Indeed, each entry of the dataset $D$ is weighted according to its rate, so that the count $Ni, j, k$ is not absolutely regular.

The algorithm implementing the formula above is called Maximum Likelihood (ML). However, if a fact doesn’t appear in the database (miss), the formula will return a probability of 0, which is presumably not correct. As data are expensive and hard to collect, we got misses in the database and this algorithm might be not relevant in all cases. That’s why we studied two other classic algorithms which can deal with this issue. But as ML is the most classic and easiest method, we still implement it in our comparison.

First, the EM algorithm is a variation of the Maximum Likelihood which tries to estimate the missing values and iterate until a stable estimation is reached. Secondly, the Maximum A Posteriori (MAP) works in the same way that the ML algorithm, but takes into account a prior distribution for the variables. Several references (as Heckerman 1995) present these two methods in more details, if needed. Yet, notice than frameworks or software already implement these algorithms, like Hughins$^1$.

5. EVALUATION

We need quality data, in term of quantity, coverage and representativeness, for learning the parameters in a pertinent way. We performed experimentations in the Grenoble’s hospital for collecting the required data. First, one surgeon and six interns realized a set of six exercises on our simulator, each of them presenting various characteristics and difficulties. During all the surgery they had to explain what they are doing, and how they took decisions. Then, the experimental team on one hand, and a second expert on the other both played the part of the diagnostic tools and diagnosed the knowledge states, based on the observation of the students’ activity (for instance, the X-rays that were taken). They also gave a degree of belief from 0 (not sure) to 4 (absolutely sure). As beginners took part in this experimentation, we hopefully got errors and unsatisfying tries. On the other hand, our data only are a brief snapshot of the expertise of the subjects, whereas the evolution of the students is highly important in ITS. Fortunately, students of various skilled and surgeons were involved in the test, so that this

$^1$ http://www.hugin.com
bias is reduced. After having study the collecting data, we notice that they suffer from two major issues:

- first, their cost, because we collected them directly at the hospital, during several depth interviews with students and experts

- secondly, their quantity: as these data are expensive, we can't obtain a lot of them, which is a crucial problem for learning.

Therefore, we obtain a table containing the following value:

\[
\text{entry} := \{\text{Operator, problem, representation, control, diagnostic of the team, diagnostic of the expert, situation variables}\}
\]

The annotations fit the overall model. The range of each variable also fit the domain of the nodes in the Bayesian network. A control (or cognition) can only be brought into play in a valid way, an invalid way or not brought into play whereas it should. However, we sometimes don't have the value of some parameters (as the position of the X-rays amplifier), so we are not able to compute all the situation variables.

A more detailed statistic analysis on the data allows us to extract information. Since we know that we have few data, it is important to evaluate their quality before the learning.

Here is given a first report:

- Number of controls: 45
- Number of annotations: 2695
- Number of annotations per controls (mean): 79.2
- Number of controls with less than fifteen annotations: 28

The figure 4 shows the graphical representation of this information: the controls are shown on the abscissa, and on the ordinate is the current amount of annotations for each one. As a result, we can note that for almost 1/2 of the controls, the learning won't be pertinent because of the lack of annotations (less than ten). However, a substantial sample of knowledge offers better quality, mainly for empirical and declarative one. We consider that at least fifty entries (either diagnosed by the experts or the experimental team) per controls in the dataset is relevant. During the validation step, we worked about this subset of data in order to see if our learning methods are useful.

We want now evaluate the efficiency of our learning algorithms and compare them. In the field of Artificial Intelligence, the learning usually is assessed by computing the recognition rate, i.e. checking if the network is able to find the same diagnostic than a reference set of data (for instance the data already diagnosed by the expert). It simply is the ratio of good prediction report to bad prediction.

In addition, the cross-validation method seems to be suitable for us, especially because we have few data. The main idea is to divide the dataset into two subsets: one for learning the network and one for evaluating its. As we have two different sets of data (either annotated by the expert or by the experimental team), we can also cross them. Thus, we divided the data in two ways:
— Validation 1: two thirds of the expert data for learning, the rest plus the experimental data for validating.
— Validation 2: the experimental data for learning and the expert ones for validating.

The following tables give the recognition rate for each learning algorithms.

Table II. Recognition rates with the first validation method

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>59.3%</td>
</tr>
<tr>
<td>MAP</td>
<td>63.4%</td>
</tr>
<tr>
<td>EM</td>
<td>59.8%</td>
</tr>
</tbody>
</table>

Table III. Recognition rates with the second validation method

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>51.3%</td>
</tr>
<tr>
<td>MAP</td>
<td>58.5%</td>
</tr>
<tr>
<td>EM</td>
<td>51.6%</td>
</tr>
</tbody>
</table>

As the rates are close, we also have tested the algorithms with 3-fold method. We randomly assign all the observations to three partitions such that the partitions are near-equal size. Then each fold is successively used for computing the recognition rate. As we have few data, we use a small number of folds.

Table IV. 3-folds cross validation

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fold 1</th>
<th>Fold 2</th>
<th>Fold 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>58.33%</td>
<td>60.2%</td>
<td>57.75%</td>
</tr>
<tr>
<td>MAP</td>
<td>63.3%</td>
<td>63.7%</td>
<td>66.18%</td>
</tr>
<tr>
<td>EM</td>
<td>59.36%</td>
<td>60.3%</td>
<td>59.97%</td>
</tr>
</tbody>
</table>

According to these results, the MAP estimation gives the best recognition rate, with a prior distribution based on our background knowledge. Since there is few miss in our subset, EM algorithm gives almost the same results than the Maximum Likelihood, which was used for the maximization step. We also notice that the results are close for each fold. There is no evidence of sensibility to hard instances or to the repartition of the data with our dataset. However, the rates are not perfect, probably due to the lack of data and the difference between the two kinds of annotation. Indeed, strict separation of expert and experimental data (table III) gives the worst results. On the other hand, various data reinforce the confidence in the rates, as it reduces the machine learning bias. However, drawing further conclusions is somewhere hard since the results still are very close. We can yet notice that new data are going to be available sooner in order to perform more tests.

Compute the recognition rate is a very common way for validating learning methods in the field of machine learning, but it clearly is a kind of quantitative research. In complex and vague domains such as knowledge diagnosing, qualitative methods may give lots of results, for instance by testing out the network with real students and verifying if the final global diagnosing is consistent. Yet this kind of evaluation remains expensive.

6. DISCUSSION AND CONCLUSION

We have bring out a methodology allowing to learn from data the parameters of a Bayesian network for knowledge diagnostic that keep the deep didactical model of the domain. The method needs data that fit the didactical model and the structure of the Bayesian network. Thus our methodology can be reused or studied with respect to these formalisms. Concerning the learning itself, we show that the algorithms efficiency
depends on the quantity of data. In complex and ill defined domains like surgery where data are very expensive, this remains a crucial issue. Here one can notice that we need more data to learn the entire network and refine the evaluation step. In particular, a major limit of our work lies in the fact that we have not be able to learn parameters for haptic nodes, due to a hardware issue. As learning may be a suitable approach for such complex raw data though, this problem shall be address in the future. Alongside with real experimentations, a better evaluation could be realized in the future.

Another point to clarify is to determine the domain for which this approach is suitable. Data-centric approach usually is defined as suitable for ill-defined domain (like surgery for us). But as parameter learning usually stays independent from the domain, this issue is bypassed.

According to Mayo and Mitrovic’s classification, our resulting system is situated between data-centric and expert-centric approach, because we don’t learn the structure of the network. But even in Mayo and Mitrovic’s approach, the structure is never updated once it was learned, contrary to the parameters. Indeed, we already emphasized that it is relatively fast to apply those machine learning techniques for estimating the probabilities, given data. Consequently, we show a way to refine causal models that keeps the prior expert knowledge, and besides, we also use them for the MAP algorithm.

Our network has to be evaluating in a deeper way, as suggest above. Nevertheless, we already emphasized that it is relatively fast and useful to apply those machine learning techniques, given data. Nowadays, the knowledge diagnostic tools of new ITS often comes from a large set of classical methods. On the other hand, we rarely find comparison between these methods due to their cost and huge time consumption [Mitrovic et al. 2003]. Consequently, no evidence that the best one was chosen can be provide. The use of raw data alongside machine learning techniques may favor this process in future works. A first comparison between two approaches seems all-interesting.

REFERENCES


