Exploring the Link Between Initial and Final Diagnosis in a Medical Intelligent Tutoring System

Tenzin Doleck[1], Ram B. Basnet[2], Eric Poitras[3], Susanne Lajoie[1]
[1]McGill University, Montreal, Canada
{tenzin.doleck, susanne.lajoie}@mcgill.ca
[2]Colorado Mesa University, Grand Junction, Colorado
rbasnet@coloradomesa.edu
[3]University of Utah, Salt Lake City, Utah
assistlaboratory@gmail.com

***Abstract*— A constant topic in medical education is clinical reasoning: how do learners solve cases? Learner interactions with Intelligent Tutoring Systems yield fine-grained data that are useful in generating meaningful information and illuminating understanding about learner behaviors and outcomes. We examine and analyze the log files generated by BioWorld, an Intelligent Tutoring System for the medical domain. More specifically, to further our understanding of the nature of reasoning employed by learners while solving virtual patient cases in BioWorld, one important step is to examine the initial list of selected diagnostic hypotheses before any other learner action is taken in diagnosing a case. By exploring the link between initial selected hypotheses and final submitted hypothesis, a better understanding of the learners' reasoning might be achieved.**

Index Terms— data mining, decision trees, medical education, computer-based learning environments, clinical reasoning, intelligent tutoring systems, assessment, learning

I. INTRODUCTION

As educational data becomes increasingly important, the access to ever-increasing data creates new possibilities as well as challenges in their analysis. In recent years, the field of educational data mining has significantly grown and soft computing techniques can and are being applied with much success on educational data [1, 2]. This study is anchored in the context of medical education via an Intelligent Tutoring System for clinical reasoning. For advancing our understanding of learners' clinical reasoning, we investigate the link between initial selected hypotheses and final submitted hypothesis by learners. Toward this goal we utilize a common data mining technique, namely, decision trees for our experiment.

BioWorld is a computer-based learning environment designed to train novice physicians in medical diagnostic reasoning [3, 5]. Patient cases are solved by novices through tools that are embedded in the system interface, allowing novices to order lab tests, manage hypotheses, highlight symptoms, search a library, and ask for help. An important component of BioWorld is the feedback palette, where novices receive a report that highlights areas of improvement on the basis of an expert path to solving the problem. In sum, these tools are designed to shape how novices reason about the cases by creating explicit representations of medical diagnostic reasoning.

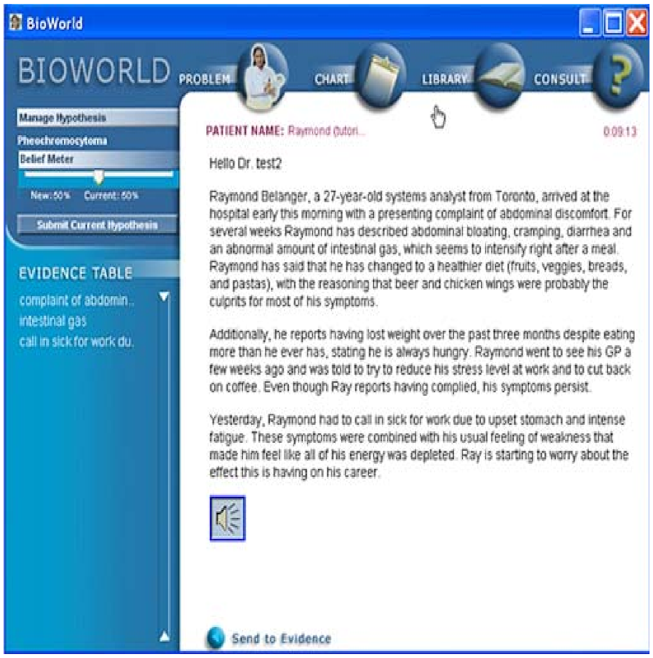
A key aspect of diagnostic reasoning is a physicians' confidence towards their main hypothesis for a case. As a result of making progress in solving the case or reaching an impasse, physicians may change their own diagnosis of the patient condition. Lines of diagnostic reasoning are made up of physicians' actions, but are delimited by changes in confidence, which correspond with the strength of conviction that a diagnosis is correct or incorrect.

This study aims to model lines of diagnostic reasoning, particularly with regard to novices' changes in hypotheses while solving patient cases. The rationale is to gain better understanding of the temporal unfolding of diagnoses during problem-solving. We expect that the relationship between the initial diagnosis that was selected and the final diagnosis that was submitted can be represented as a decision tree classifier. In doing so, the analytical techniques benefits the broader research community by identifying commonly held misconceptions about patient cases, which can be used by system designers to adapt instruction to the specific needs of struggling learners.

II. BIOWORLD: AN INTELLIGENT TUTORING SYSTEM

Developing expertise in clinical reasoning is a focal topic in medical education, and technology-rich environments such as Intelligent Tutoring Systems can be used to provide learners with opportunities for gaining clinical reasoning skills. BioWorld (figure 1) is an example of an Intelligent Tutoring System for the medical domain that was designed to engage and support novice physicians in practicing and developing realistic clinical reasoning skills. The system was created using a cognitive apprenticeship framework [4] where learners practice realistic tasks and are scaffolded in the context of their learning with expert models. In BioWorld, novice physicians learn clinical reasoning skills by diagnosing virtual patient cases by identifying relevant symptoms, ordering lab-tests, developing diagnostic hypotheses, prioritizing and summarizing evidence, and reasoning about the nature of the underlying disease [3, 5, 6].

978-1-4799-6876-3/14/$31.00 ©2014 IEEE



Pheochromocytoma

New: 50% Current: 50%

complaint of abdomin...
intestinal gas
call in sick for work du...

PATIENT NAME: Raymond (tutori... 0:09:13

Hello Dr. test2

Raymond Belanger, a 27-year-old systems analyst from Toronto, arrived at the hospital early this morning with a presenting complaint of abdominal discomfort. For several weeks Raymond has described abdominal bloating, cramping, diarrhea and an abnormal amount of intestinal gas, which seems to intensify right after a meal. Raymond has said that he has changed to a healthier diet (fruits, veggies, breads, and pastas), with the reasoning that beer and chicken wings were probably the culprits for most of his symptoms.

Additionally, he reports having lost weight over the past three months despite eating more than he ever has, stating he is always hungry. Raymond went to see his GP a few weeks ago and was told to try to reduce his stress level at work and to cut back on coffee. Even though Ray reports having complied, his symptoms persist.

Yesterday, Raymond had to call in sick for work due to upset stomach and intense fatigue. These symptoms were combined with his usual feeling of weakness that made him feel like all of his energy was depleted. Ray is starting to worry about the effect this is having on his career.

Fig. 1. BioWorld Interface

III. METHOD OVERVIEW

A. Participants

In this study, participants were recruited via advertisements and newsletters. Thirty undergraduate students volunteered to participate in the study and were compensated $20 for their participation at the completion of a 2-hour session. The gender breakdown of the participants was as follows: 19 women (63%) and 11 men (37%), with an average age of 23 (SD = 2.60). All 30 participants (28 medical students and 2 dental students) were registered in the same classes at a large Northeastern Canadian Research University.

B. Procedure

Participants were tasked with solving three virtual patient cases (endocrinology cases) in BioWorld on an individual basis for a total duration of 2 hours. The three cases were Amy, Cynthia, and Susan-Taylor. The correct diagnosis for the cases Amy, Cynthia, and Susan-Taylor were diabetes mellitus (type 1), pheochromocytoma, and hyperthyroidism respectively.

C. Measures

Like many computer-based learning environments, the BioWorld system also generates log files of user actions. Three types of performance metrics are recorded in the log files, namely, diagnostic efficacy (e.g., accuracy, count of matches with experts, and percentage of matches with experts), efficiency (e.g., number of tests ordered and time to solve the case), and affect (e.g., confidence). Furthermore, information saved in the log-file included the attempt identifier (participant and case ID), a timestamp, the BioWorld space (e.g., chart), the specific action taken (e.g., add test), and details in relation to the action (e.g., Thyroid Stimulating Hormone (TSH) Result: 0.2 mU/L). For this study, we use the logs that contain the user actions recorded by the system while the participants solved the three patient cases.

IV. DECISION TREE ANALYSIS

Recently, there has been an upsurge in the interest and application of educational data mining techniques for eliciting fine-grained information from log files [1, 2]. The log files generated by Intelligent Tutoring Systems hold important information about learner behavior and outcomes, and, are thus important in that such data holds the potential to improve the system functionality in providing adaptive scaffolding and tutoring. Using the information from the log files generated by the BioWorld system, this study sought to test the hypothesis that learners would be more likely to pick the correct hypothesis while diagnosing easier patient case and vice versa. In a previous study by Gauthier et al. [7], they highlighted the difficulty levels of the various patient cases; the difficulty levels (i.e., anticipated accuracy) of the three relevant cases are summarized in Table 1. Driving our study is the goal to examine our initial line of thinking where we hypothesize that learners would be more likely to pick the correct diagnosis in the initial set of hypotheses for the Amy case, while the converse would be true for the Cynthia case, the more difficult case.

TABLE I. ANTICIPATED ACCURACIES FOR THE THREE CASES

Case	Correct Diagnosis	Anticipated Accuracy
Amy	Diabetes Mellitus (type1)	94%
Susan Taylor	Hyperthyroidism	78%
Cynthia	Pheochromocytoma	33%

We looked at the initial hypothesis that learner's select and wanted to investigate the relation to the final submitted diagnosis. For this investigation, we decided to use the decision tree analysis, one of the most common data mining techniques, because of the fact that it is easy to implement, understand, interpret, and visualize via a graphical rendition. While solving a patient case in Bioworld, since, learners have the liberty to pick any number of hypothesis initially, we wanted to limit the list to the first three hypothesis selected. Furthermore, since the number of instances (in this case the number of participants, n=30) is relatively small, coupled with the fact that the number of attributes (hypothesis) is large, delimiting the number of attributes is warranted. Having the number of attributes larger than the number of instances prevents us from taking advantage of the decision tree analysis. Thus, to apply decision tree analysis, we generated a few rules for delimiting the initial list of hypotheses: 1). Select hypothesis (first three) right after learners 'add evidence', 2). Select hypothesis before any lab tests are ordered, library is accessed, or any other learner actions apart from 'add evidence', 3). If learner immediately deletes the hypothesis then exclude the hypothesis. Based on these rules we generated a dataset to which we applied decision tree analysis for the three cases.

A significant body of recent research exalts the affordances of educational data mining [1, 2]. Decision trees, a classification process of a given input, are popular classification methods and are regularly used in varied data mining tasks. Moreover, Decision trees, along with providing a means of classification, also provide a visualization of the flowchart-like tree structure. We used the C4.5 algorithm, which uses a greedy technique to induce decision trees, implemented as J48 classifier [8] in WEKA [9]. The WEKA toolkit is a comprehensive workbench for machine learning algorithms for data mining tasks ranging from data preprocessing to classification. The methodology employed in our study is highlighted in figure 2. In the following section, we provide the results of the J48 classifier for the 3 cases (Amy, Cynthia, and Susan Taylor). After running the algorithm on the dataset, the output generated also includes the percent of correctly and incorrectly classified instances.

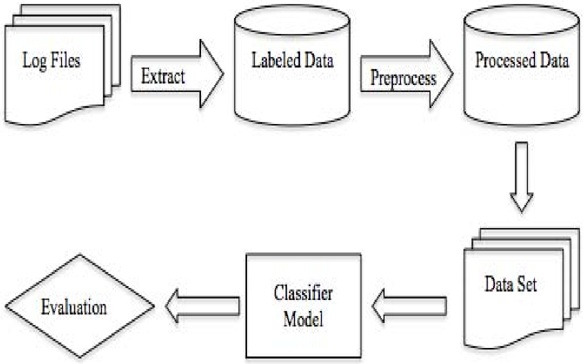


Fig. 2. Experimental Setup

A. *Case 1: Amy*

The results of the decision tree analysis on the Amy case are presented below. The decision tree can be expressed in rule format. For the case of Amy, if one of the initial selected Hypotheses is Diabetes Mellitus (type 1), then in 23 of the 30 instances, such a selection yields the correct final hypothesis.

```
J48 unpruned tree
------------------
DiabetesMellitus(typeI) = TRUE
|  DiabetesMellitus(typeII) = TRUE: Correct (7.0)
|  DiabetesMellitus(typeII) = FALSE
|  |  Anorexia = TRUE: Correct (3.0)
|  |  Anorexia = FALSE
|  |  |  RenalDysfunction = TRUE: Correct (3.0)
|  |  |  RenalDysfunction = FALSE: Correct (10.0/1.0)
DiabetesMellitus(typeI) = FALSE
|  RenalDysfunction = TRUE: Correct (2.0/1.0)
|  RenalDysfunction = FALSE: Incorrect (5.0/1.0)

Number of Leaves  :      6
Size of the tree :   11
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         27              90     %
Incorrectly Classified Instances        3              10     %
Kappa statistic                         0.7059
Mean absolute error                     0.1668
Root mean squared error                 0.2995
Relative absolute error                49.3471 %
Root relative squared error            73.4555 %
Coverage of cases (0.95 level)         96.6667 %
Mean rel. region size (0.95 level)     80      %
Total Number of Instances              30
```

B. *Case 2: Susan Taylor*

The results of the decision tree analysis on the Susan Taylor case are presented below. The decision tree can be expressed in rule format. For the case of Susan Taylor, if one of the initial selected Hypotheses is Hyperthyroidism, then in 17 of the 30 instances, such a selection yields the correct final hypothesis.

```
J48 unpruned tree
------------------
Hyperthyroid(Gravesdisease) = TRUE: Correct (17.0)
Hyperthyroid(Gravesdisease) = FALSE
|  BarbiturateIntoxication = TRUE: Correct (3.0)
|  BarbiturateIntoxication = FALSE
|  |  PanicAttack = TRUE
|  |  |  Anorexia = TRUE: Correct (2.0/1.0)
|  |  |  Anorexia = FALSE
|  |  |  |  Arrhythmia = TRUE: Incorrect (2.0)
|  |  |  |  Arrhythmia = FALSE: Correct (2.0/1.0)
|  |  PanicAttack = FALSE
|  |  |  Arrhythmia = TRUE: Correct (2.0/1.0)
|  |  |  Arrhythmia = FALSE: Correct (2.0)

Number of Leaves  :      7
Size of the tree :   13
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         24              80     %
Incorrectly Classified Instances        6              20     %
Kappa statistic                         0.28
Mean absolute error                     0.2056
Root mean squared error                 0.4271
Relative absolute error                68.7821 %
Root relative squared error           111.9816 %
Coverage of cases (0.95 level)         83.3333 %
Mean rel. region size (0.95 level)     55      %
Total Number of Instances              30
```

C. *Case 3: Cynthia*

The results of the decision tree analysis on the Cynthia case are presented below. The decision tree can be expressed in rule format. For the case of Cynthia, if one of the initial selected Pheochromocytoma, then in 11 of the 30 instances, such a selection yields the correct final hypothesis.

```
J48 unpruned tree
------------------
```

```
Pheochromocytoma = TRUE: Correct (11.0)
Pheochromocytoma = FALSE
|  PanicAttack = TRUE
|  |  Hyperthyroid(Gravesdisease) = TRUE: Correct (3.0/1.0)
|  | Hyperthyroid(Gravesdisease) = FALSE: Correct (4.0/2.0)
|  PanicAttack = FALSE
|  |  Hyperthyroid(Gravesdisease) = TRUE: Incorrect (7.0)
|  |  Hyperthyroid(Gravesdisease) = FALSE
|  |  |  Arrhythmia = TRUE: Incorrect (3.0)
|  |  |  Arrhythmia = FALSE: Correct (2.0/1.0)

Number of Leaves  :          6
Size of the tree :   11
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         21              70      %
Incorrectly Classified Instances        9              30      %
Kappa statistic                         0.3946
Mean absolute error                     0.2806
Root mean squared error                 0.5002
Relative absolute error                55.7268 %
Root relative squared error            99.1195 %
Coverage of cases (0.95 level)         80      %
Mean rel. region size (0.95 level)     56.6667 %
Total Number of Instances              30
```

V. Conclusion

The work described in this study represents a first step in examining learners initial selected hypotheses while solving virtual patient cases in BioWorld. In this study, we employed decision tree analysis to understand the relation between the initial selected hypotheses and the final submitted hypothesis; the results from this study indicate that the initial selected hypotheses predict the final submitted hypothesis. Particular patterns as hypothesized emerged. For the case of Amy, if one of the initial selected Hypotheses is Diabetes Mellitus (type 1), then in 23 of the 30 instances, such a selection yields the correct final hypothesis. For the case of Susan Taylor, if one of the initial selected Hypotheses is Hyperthyroidism, then in 17 of the 30 instances, such a selection yields the correct final hypothesis. For the case of Cynthia, if one of the initial selected Pheochromocytoma, then in 11 of the 30 instances, such a selection yields the correct final hypothesis. Taken together, these results suggest that the novices were more likely to select the correct final hypothesis if they had selected it as one of their initial hypotheses based simply off of the patient summary; this could have resulted from the fact that some novices were able to generate the correct hypothesis from just reading the patient summary. This finding is in line with previous studies that have shown the propensity of physicians to generate hypotheses immediately from the patient's symptoms [10]. As highlighted earlier, previous study by Gauthier et al. [7] outlined the varying difficulty of the three cases; the anticipated accuracies were 94%, 78%, and 33% for Amy, Susan Taylor, and Cynthia, respectively. This study demonstrated that there is a higher probability of selecting the correct hypothesis in the initial set of hypotheses for the easier case (Amy). While the probability of selecting the correct hypothesis in the initial set of hypotheses for the Susan Taylor case was slightly lower. Finally, in the more difficult case (Cynthia), learners were less likely to select the correct hypothesis in the initial set of hypotheses. Thus, the takeaway was that the likelihood of the picking the correct hypothesis in the initial list of hypotheses was dependent on the difficulty of the case. In addition to demonstrating a link between the initial selected hypotheses and the final submitted hypothesis, the knowledge elicited from this study also has implications for the design of scaffolding (prompts). Interventions could be designed for the more difficult cases, to provide scaffolding when the learners do not include the correct hypothesis in their initial list of hypotheses.

References

[1] R.S.J.D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3-17, 2009.

[2] C. R. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol.40, no. 6, pp. 601–618, 2010.

[3] S. P. Lajoie, L. Naismith, E. Poitras, Y.-J. Hong, I. Cruz-Panesso, J. Ranellucci, S. Mamane, and J. Wiseman, "Technology-rich tools to support self-regulated learning and performance in medicine," in International Handbook of Metacognition and Learning Technologies, vol. 28, R. Azevedo and V. Aleven, Eds. New York: Springer, 2013, pp. 229-242.

[4] A. Collins, "Cognitive apprenticeship," in Cambridge Handbook of the Learning Sciences, R. K. Sawyer, Ed. Cambridge UK: Cambridge University Press, 2006, pp. 47-60.

[5] S. P. Lajoie, "Developing professional expertise with a cognitive apprenticeship model: Examples from Avionics and Medicine," in Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments, K. A. Ericsson, Ed. Cambridge UK: Cambridge University Press, 2009, pp. 61-83.

[6] S. P. Lajoie, E. G. Poitras, T. Doleck, and A. Jarell, "Modeling Metacognitive Activities in Medical Problem-Solving with BioWorld," In Peña-Ayala (Ed.), Metacognition: Fundamentals, Applications, and Trends. Springer Series: Intelligent Systems Reference Library, 2015.

[7] G. Gauthier, S. P. Lajoie, L. Naismith, and J. Wiseman, "Using expert decision maps to promote reflection and self-assessment in medical case-based instruction," In Proceedings of Workshop on the Assessment and Feedback in Ill-Defined Domains, Intelligent Tutoring Systems, Montreal, Canada, pp. 68-80, 2008.

[8] J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann, San Mateo, CA, 1993.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann,and I. H. Witten, "The weka data mining software: An update," SIGKDD Explorations, vol. 11, pp. 10-18, 2009.

[10] E. Berner and M. Graber, 'Overconfidence as a Cause of Diagnostic Error in Medicine', The American Journal of Medicine, vol. 121, no. 5, pp. S2-S23, 2008.