Augmenting the Novice-Expert Overlay Model in an Intelligent Tutoring System: Using Confidence-Weighted Linear Classifiers

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*— In BioWorld, a medical intelligent tutoring system, novice physicians are tasked with solving virtual patient cases. Whilst the importance of modeling and predicting clinical reasoning is recognized, an important aspect of the learner contribution remains unexplored - the written case summary prepared by the learner. The premise of investigating the case summaries is that it captures the thought and process of the learners in solving the cases; since, the case summaries hold important reasoning information, it makes sense to incorporate it as part of the novice-expert overlay model. In this paper, case summaries written by novices and experts were considered as an addendum to the existing novice-expert overlay model in the BioWorld system. Toward this goal, using a promising new classification method called confidence-weighted linear classifiers, this paper proposes a way to augment the novice-expert overlay model in BioWorld.**

Keywords- data mining, confidence-weighted linear classifiers, novice-expert overlay, medical education, computer-based learning environments, intelligent tutoring systems, clinical reasoning

I. Introduction

BioWorld is a computer-based learning environment designed as a cognitive tool to train novice physicians in diagnosing virtual patients and receiving timely feedback [1]. Ongoing feedback is provided to learners pertaining to their prioritized list of evidence items that supports their diagnostic process as it compares to evidence taken by an expert to solve the case. In an attempt to tailor the content of the feedback to the specific needs of different novices, BioWorld relies on a novice-expert overlay approach to assess learning during problem solving [2].

A novice-expert overlay model is a process which consists of three components: a) the user interactions of novices captured during problem solving; b) an expert solution path; and c) the feedback delivery system. In order to adapt the nature of the feedback to the specific needs of different novice physicians, BioWorld compares the evidence items identified as pertinent by novices and experts with the aim of highlighting similarities and differences in the context of the feedback palette.

A more comprehensive approach would include several types of user interactions during the analysis performed by the learner model. In past studies, our examinations have included other types of user interactions that have been shown to be critical to diagnostic performance, including help-seeking, information-seeking, diagnostic laboratory testing, and case summary writing [3]. The evidence from the investigations of the case summaries suggests that there are significant differences between case summaries written by novices as opposed to experts. The linguistic features extracted from the case summaries have been used successfully to discriminate between the type and correctness of case summaries. However, there is a pressing need to build text classification algorithms that perform efficiently within the server database.

This paper addresses this issue by evaluating the performance of confidence-weighted linear classifiers, an online learning method for Natural Language Processing (NLP) problems. In particular, this research seeks to address the following questions: a) what is the accuracy of confidence-weighted linear classifiers in the recognition of case summaries written by novices and experts?; and b) how does the addition of a stepwise feature selection algorithm affect the accuracy of the text classification model?

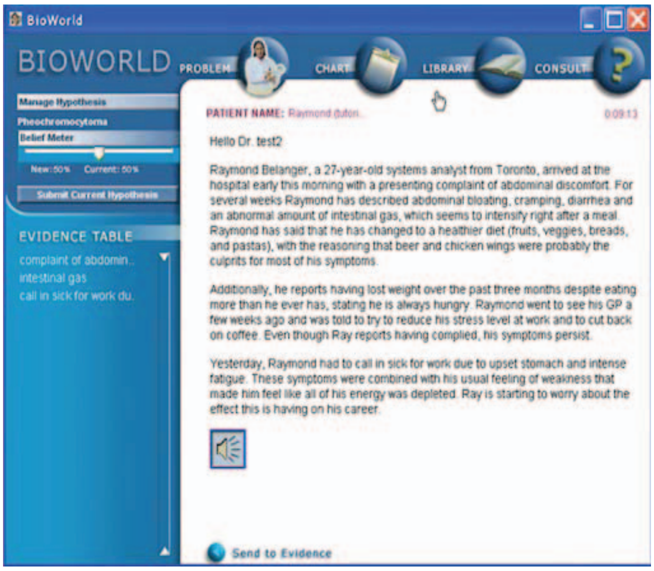
II. The Context: BioWorld

BioWorld (Fig. 1) is an Intelligent Tutoring System for the medical domain that is designed to support novice physicians in practicing diagnostic reasoning skills while receiving feedback [1]. The system was created using a cognitive apprenticeship framework [4] where learners practice realistic clinical reasoning tasks and are scaffolded in the context of their learning with expert models. In BioWorld, novice physicians learn clinical reasoning by diagnosing virtual patient cases by identifying relevant evidence/symptoms, ordering lab-tests, seeking help via the embedded library, and reasoning about the nature of the underlying disease [5].

BioWorld employs a novice-expert overlay system [2] to assess clinical reasoning during learning. The system compares similarities and differences in learner solution path against an expert solution path. For instance, after novices submit their

978-1-4799-3975-6/14/$31.00 ©2014 IEEE

final hypothesis, learners are able to compare their solution with an expert's solution. Along with a comparison of the novice-expert on diagnosis and evidence, the system also provides a detailed explanation of an expert's reasoning. However, the current version of this user model omits an important aspect of the novices' clinical reasoning, i.e., the final written case summary. The written case summaries contain highlights apropos the symptoms, vital signs, and lab-tests that were germane to diagnosing the patient case. In order to address this gap and to augment the current novice-expert overlay model, this paper represents the first steps in exploring the efficacy of text categorization algorithms in differentiating between novice and expert case summaries written in BioWorld to augment the current novice-expert overlay model in BioWorld.



PATIENT NAME: Raymond Belan... 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

The participants for this study were recruited through advertisements and newsletter. A total of 30 volunteer undergraduate students agreed and participated in the study. The participants were compensated $20 at the completion of the study. The sample comprised of 19 women (63%) and 11 men (37%), with an average age of 23 (SD = 2.60). All 30 participants were registered in the same classes at a large Northeastern Canadian Research University (where 28 were medical students and 2 were dental students).

B. Procedure

Participants were tasked with solving three virtual patient cases in BioWorld on an individual basis for a total duration of 2 hours. While solving the patient cases, participants also engaged in thinking aloud. The three virtual patient cases were Amy, Cynthia, and Susan-Taylor. The correct diagnosis for the cases Amy, Cynthia, and Susan-Taylor was diabetes mellitus (type1), pheochromocytoma, and hyperthyroidism respectively.

C. Measures

The BioWorld system generates log files of user actions. There are three types of performance metrics 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). 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). Along with the aforementioned data elements, the case summaries written in the BioWorld System are also logged in the log files.

IV. CONFIDENCE-WEIGHTED LINEAR CLASSIFIERS

In recent years, text mining has become an important means of knowledge-based discovery and has assumed a central method with multifarious potential applications in varied fields ranging from commerce to education. The field of educational data mining has significantly grown and data mining has been applied with much success on educational data [6, 7]. When investigating the potential for augmenting the novice-expert overlay model in BioWorld, one natural idea was to consider the use of machine learning techniques for this task. Although a survey of the literature will yield a gamut of machine learning algorithms for solving text classification problems, the scope of this study has been limited to an online algorithm. Since our study is an initial attempt for augmenting the novice-expert overlay model in the BioWorld system, as such, for the purpose of simplicity, the constrained nature of our study is appropriate. Future extensions of this study will explore other algorithms.

This study investigates the efficacy of Confidence-weighted linear classifiers (CWLC). CWLC proposed by Dredze et al. [8], is an online learning method for Natural Language Processing (NLP) problems. Online learning algorithms process input piece-by-piece serially and operate in rounds. Similar to popular algorithms like SVMs, the prediction rules for CW are linear classifiers as well. For a detailed description of the CW algorithm, see [8]. Online learning algorithms are a class of highly promising approach. Dredze et al. [8] have employed CWLC on a number of NLP problems and have demonstrated the efficacy of the CWLC over other online and batch methods, and the ability to learn faster in online settings.

A. Data Set

Our data for the experiments come from the BioWorld system. The case summaries were extracted from the log files generated by BioWorld. The data for the novice-expert classification problem included a total of 74 case summaries, with 60 cases written by novices and 14 cases written by experts.

A sample of a case summary written by a student is presented below:

> Patient has elevated T3, T4; low TSH, and elevated thyroid stimulating antiglobulin. This is very suggestive of hyperthyroidism due to an autoimmune process. Listed symptoms (anxiety, weight loss, elevated HR, BP, tremor, sweating) all support this diagnosis.

A sample of a case summary written by an expert is presented below:

> 37 year old female, presenting after starting high blood pressure pill with episodes of palpitations, flushing and sweating. On exam, hypertensive, relative tachycardia. Labs revealed: normal TSH, T4, T3, glucose. Elevated free urinary cathecolamines but normal total. CT abdo normal.

B. Experimental Setup

The case summaries were extracted from the log files generated by the BioWorld System. The methodology employed in our study is displayed in Fig. 2. In our study, the WEKA [9] toolkit (Fig. 3) was used to run our experiments. WEKA is a comprehensive workbench for machine learning algorithms for data mining tasks ranging from data preprocessing to classification.

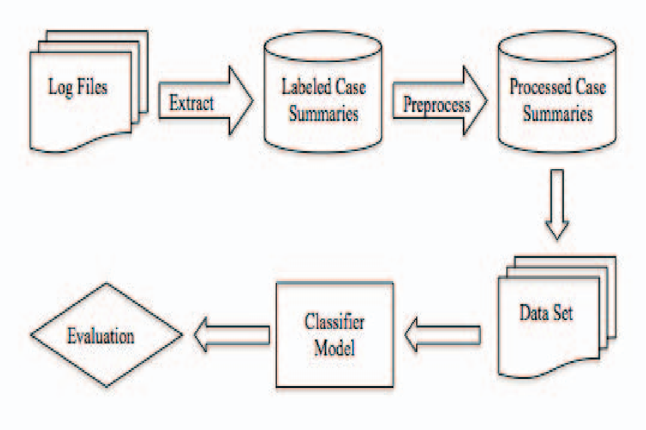


Fig. 2. Experimental Setup

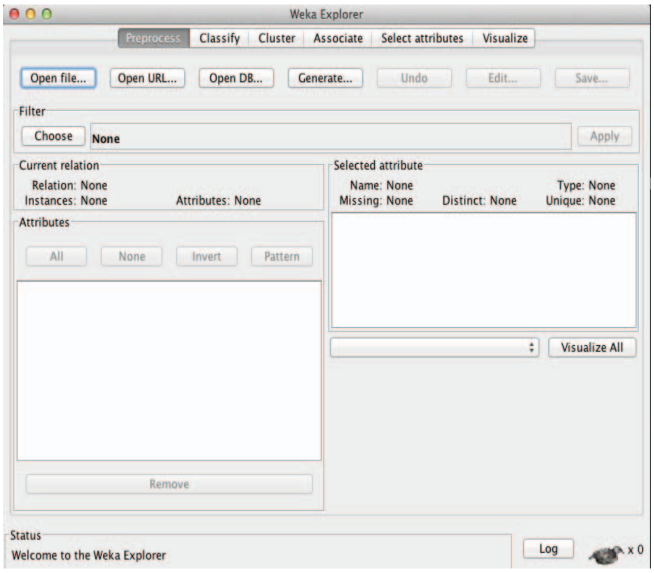


Fig. 3. WEKA Screenshot

The preprocessing filter *StringToWordVector* provided by WEKA was used to extract feature vector from the summary texts. The default values were used except for the parameter that converted texts into lowercase. Stemmer is commonly used to reduce the feature size in text classification problems where the feature size tends to be in the order of tens of thousands. Because our dataset is fairly small in terms of both the size of each summary as well as the total number of summaries provided by expert and novices in the field, the stemmer was not utilized. Stopwords were not removed; they were kept as the part of feature vector. Using the aforementioned preprocessing technique, 823 word tokens were extracted as features. The resulting *arff* file format was then converted into libsvm format supported by CWLC tool out of the box.

C. Results

The CWLC algorithm was used to learn to classify the novice-expert dataset. CWLC was applied on the generated dataset using 10-fold cross-validation on the entire dataset. The initial results with all the features are presented in Table 1. Further experiments were conducted with feature selection to ascertain if the classifier results could be improved. Feature selection techniques are commonly applied in data mining problems to extract the most discriminative features to reduce the feature size as well as to improve classifier performance. Correlation-based Feature Subset Selection (CfsSubsetEval) technique provided by WEKA (we used the default values for the parameters) was utilized; this technique evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them [10].

The *GreedyStepwise* forward search technique was employed through the space of attribute subsets to search for the most discriminative feature set and rank them based on their individual predictive ability. The process selected 31 top features. A new dataset was generated using these 31 features, and CWLC was applied to this dataset. Having applied the aforementioned feature selection technique to our dataset, there was a marked improvement in the classifier's results (from 89.19% to 97.30%) when the selected features were used in Novice-Expert classification. The improved classification results are presented in Table 2.

TABLE I. NOVICE-EXPERT CLASSIFICATION (USING ALL FEATURES)

Accuracy	**89.19%**
Error	10.81%
FP Rate	57.14%
FN Rate	0.00%
True Positive Rate	100.00%
True Negative Rate	42.86%
Precision	88.24%
Recall	100.00%
FMeasure	93.75%
Matthews Correlation Coefficient	0.615
Balanced Error Rate (BER)	28.57%

TABLE II. NOVICE-EXPERT CLASSIFICATION (USING TOP FEATURES)

Accuracy	**97.30%**
Error	2.70%
FP Rate	21.43%
FN Rate	0.00%
True Positive Rate	96.67%
True Negative Rate	100.00%
Precision	95.08%
Recall	100.00%
FMeasure	97.48%
Matthews Correlation Coefficient	0.885
Balanced Error Rate (BER)	10.71%

V. CONCLUSION

In this study, the case summaries written by experts and novices in the BioWorld system were analyzed. The use of machine learning in augmenting the current novice-expert overlay model in the system was considered. The current study used the confidence-weighted linear classier for the novice-expert classification task. The results from this study support the idea of employing text classification in augmenting the novice-expert overlay model in BioWorld. The implications of the current study are promising in that the study demonstrated via the experimental results the efficacy of confidence-weighted linear classifier in the novice-expert classification task. The CWLC shows significant promise; with the classifier reaching an accuracy level of 97.30% after the feature selection techniques were incorporated. Going forward with the effort towards augmenting the novice-expert overlay model, several opportunities for improving the BioWorld system are available. Future work will explore other classifiers useful in text mining tasks. A research problem that deserves attention is the automatic scoring of text, i.e., the case summary in our problem space. Future studies will explore this line of research.

REFERENCES

[1] 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.

[2] V. J. Shute, and D. Zapata-Rivera, “Adaptive educational systems,” in Adaptive technologies for training and education, P. Durlach, Ed. New York, NY: Cambridge University Press, 2012, pp. 7-27.

[3] B. Goldberg, R. Sottilare, I. Roll, S. Lajoie, E. Poitras, G. Biswas, J. Segedy, J. Kinnebrew, E. Wiese, Y. Long, V. Aleven, K. Koedinger, and P. Winne, “Enhancing self-regulated learning through metacognitively-aware intelligent tutoring systems,” in Proceedings of the 11th International Conference of the Learning Sciences, Colorado, USA.

[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] 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.

[7] 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.

[8] M. Dredze, K. Crammer, and F. Pereira, “Confidence-WeightedLinear Classification,” in Proceedings of the International Conferenceon Machine Learning (ICML), Omnipress, 2008, pp. 264-271.

[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] M. A. Hall, “Correlation-based Feature Subset Selection for Machine Learning,” Hamilton, New Zealand, 2008.