Comparison of Selection Criteria for Multi-Feature Hierarchical Activity Mining in Open Ended Learning Environments

Yi Dong
Institute for Software Integrated Systems
Vanderbilt University
1025 16th Ave S, Nashville, TN 37212
yi.dong@vanderbilt.edu

John S. Kinnebrew
BRIDJ
283 Newbury St, Boston, MA 02115
john.kinnebrew@gmail.com

Gautam Biswas
Institute for Software Integrated Systems
Vanderbilt University
1025 16th Ave S, Nashville, TN 37212
gautam.biswas@vanderbilt.edu

ABSTRACT

This paper extends our previous work on a Multi-Feature Hierarchical Sequential PAttern Mining (MFH-SPAM) algorithm for deriving students' behavior patterns from their activity logs in an Open-Ended Learning Environment(OELE). The new algorithm is computationally efficient, and we compare the results generated by the two algorithms.

1. INTRODUCTION

Open-Ended Learning Environments [2, 5] present students with a challenging problem-solving task, along with information resources and tools for solving the task. The complexity of the learning task drives a need for dynamic and adaptive scaffolding to help novice students become effective learners. Learner models and formative assessments need to include representations that capture students' problem-solving processes in addition to their knowledge and performance in the task domain. The wealth of data that can be collected from computer-based environments provides opportunities for developing algorithms to accurately model, understand, assess students' learning behaviors and strategies.

In past work, we have developed a hierarchical sequence mining methods [3] for assessing and comparing students' learning strategies and behaviors from their interaction traces collected from OELEs. We then applied a classifier wrapper method [4] to discover smaller subsets of mined patterns that better differentiate students behavior patterns between two groups of students [7]. To address the computational complexity problems with this method while retaining the advantages of the hierarchical approach, this paper applies another selection criteria: Information Gain [6] to derive differential patterns. We conduct experimental studies to analyze student behaviors and compare the two methods.

2. BACKGROUND: MFH-SPAM

Sequential PAttern Mining (Sequential PAttern Mining) algorithm performs a Depth First Search (DFS) traversal to find all possible patterns that exceed a pre-defined frequency threshold from a data set that contains sequences of item sets [1]. SPAM employs a bitmap representation for the patterns and data sequences, which makes it easy to (1) derive pattern extensions and (2) find pattern matches in data sequences during traversal. The DFS search proceeds by extending action sequences with (1)*Sequence-extension* step (*S-step*), which extends pattern by adding a new itemset to the **end** of current pattern sequence, and (2)*itemset-extension* step (*I-step*), which adds a new item to the **last** itemset of a current sequence as an extension.

The MFH-SPAM algorithm further extends the original SPAM algorithm by adding two steps: (1) the *hierarchical-extension* step (*H-step*), which provides a way to get into more details for given actions by bringing in hierarchical representations, and (2) the *feature-extension* step (*F-step*) which makes patterns more informative by associating features with corresponding actions [7]. As a result, MFH-SPAM finds many more patterns compared to the SPAM algorithm. MFH-SPAM also allows for gaps between items(actions) that make up a pattern [3] to accommodate noise tolerance in the action sequences.

In general, even for reasonably-sized domains, the basic MFH-SPAM algorithm returns thousands of patterns, and this presents challenges in extracting the more important patterns that best characterize and differentiate student behavior. Given the computational complexities of the classifier-wrapper method used earlier [7], this paper develops a new selection criterion based on information gain [6] to identify activity patterns that distinguishes students based on their pre- to post-test learning gains measured outside of the system. The information gain for a given pattern P_1 is computed from the reduction in *Shannon entropy* when P_1 becomes known, where *Shannon entropy* for a sample data is a measure of its homogeneity. We focus on analyzing patterns with high information gain that are good differentiators between student groups.

3. CASE STUDY AND RESULTS

We run our case study on a dataset that was generated from an experiment we ran with 98 middle school students who used a learning by teaching environment, *Betty's Brain*, in a science class for a period of approximately six weeks. Learner are tasked to construct a correct causal map of a science process by reading resources, and use the knowledge learned to construct and assess the correctness of their causal map during the study. In one of our current study, students worked on a thermoregulation unit.

The students' learning gains from pre- to post-test provided us with two equally distributed groups: 49 high performers in Group 1, and 49 low performers in Group 2. We then ran the two versions of the MFH-SPAM algorithm: (1) with the classifier wrapper method, and (2) with the information gain methods to select the top 10 patterns that best differentiate the two groups. The results, presented in Tables 1 and 2 respectively, list the mean frequency of usage and the standard deviation for each selected pattern.

Table 1: Classifier Wrapper method.

Pattern	Mean(STD)	Mean(STD)
	High Group	Low Group
editlink;quiztaken	25.9(21.9)	10.6(13.3)
editmap-eff-sup	24.1(17.6)	12.3(11.5)
quiz;editmap	14.0(18.5)	7.5(12.7)
editmap-eff;quiz;expl	11.2(9.7)	5.9(8.6)
quiz;editlink;read	6.1(7.6)	2.3(2.5)
read-shrt;read;editmap;linkadd	4.0(3.3)	2.4(1.7)
read-long	19.8(30.2)	34.0(29.2)
read-shrt;editlink	13.8(9.1)	19.7(10.1)
editmap;quizview	6.8(5.6)	9.7(10.9)
editmap-ineff-unsup;read	5.6(5.1)	8.5(6.4)

Table 2: Patterns with High Information Gain

Pattern	Mean(STD)	Mean(STD)
	High Group	Low Group
quiz	95.3(51.2)	72.9(51.1)
expl	90.4(75.8)	70.0(68.9)
editlink;quiztaken	25.9(21.9)	10.6(13.3)
editmap-eff-sup	24.1(17.6)	12.3(11.5)
editmap-ineff;quiz	20.3(16.3)	14.6(12.7)
editlink;quiz;editmap	16.7(21.4)	7.2(16.1)
quiz;editmap;read	6.4(7.7)	2.8(2.9)
quiz;editlink;read	6.1(7.6)	2.3(2.5)
read-long	19.8(30.2)	34.0(29.2)
take-notes	9.5(11.3)	23.9(24.4)

Both methods find patterns that are good differentiators between the two groups of students. For example, *read-long* (sufficiently long duration read actions) has a high to low performer use ratio of 1 : 2. On the other hand, the quiz followed by an edit link followed by a resource read (*quiz;editlink;read*) has a high to low performer use ratio of 2.75 : 1. Another pattern *editlink;quiztaken* (high to low performer use ratio of 2.5 : 1) found by both methods indicates high performers are better able to use the quizzes(*quiztaken*) to check the correctness of their maps, and to direct their information seeking activities. The classifier wrapper method applying cross validation where decision tree is built multiple times for each chosen pattern, results in larger amount of calculations for information gain, whereas our new method which theoretically finds patterns with highest information gain based on i-frequency, calculates information gain only once for each pattern. Moreover, the new method tends to find shorter patterns because that shorter patterns occupying fewer bits in action sequences for i-frequency based information gain calculation, have lower value of pattern entropy which lead to higher information gain compare to longer patterns with similar usage ratio [6].

4. DISCUSSION AND CONCLUSIONS

In this paper, we have extended an initial version of MFH-SPAM by developing additional selection criteria for pattern selection and also allowing for gaps in the pattern generation from action sequences. The new method is computationally efficient than the previous approach(running time reduced from 28 seconds to 16 seconds) while retaining the strength of finding frequent patterns that are good differentiators.

In future work, we will perform more systematic analysis of the differences between groups using hypothesis testing methods. In addition, we will use correlational analysis to study in more depth the relations between behaviors and performance. We will also work toward using the patterns derived to detect learner behaviors online, and develop scaffolding and hint mechanisms that combine behavior and performance analysis to help students become better learners in OELEs.

5. ACKNOWLEDGMENTS

This work is supported by NSF IIS grant number # 1548499.

6. REFERENCES

[1] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 429–435. ACM, 2002.

[2] G. Clarebout, J. Elen, W. L. Johnson, and E. Shaw. Animated pedagogical agents: An opportunity to be grasped? *Journal of Educational multimedia and hypermedia*, 11(3):267–286, 2002.

[3] J. S. Kinnebrew, K. M. Loretz, and G. Biswas. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*, 5(1), 2013.

[4] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, Dec. 1997.

[5] S. Land. Cognitive requirements for learning with open-ended learning environments. *Educational Technology Research and Development*, 48(3):61–78, 2000.

[6] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, Mar. 1986.

[7] C. Ye, J. S. Kinnebrew, J. R. Segedy, and G. Biswas. Learning behavior characterization with multi-feature, hierarchical activity sequences. In *8th International Conference on Educational Data Mining*, June 2005.