Modeling Metacognitive and Cognitive Processes in Data Science Problem Solving (Student Abstract)

Maryam Alomair[1, 2], **Shimei Pan**[1], **Lujie Karen Chen**[1]
University of Maryland, Baltimore County[1],
King Faisal University[2]
maryama4@umbc.edu[1], shimei@umbc.edu[1], lujiec@umbc.edu[1], mmalomair@kfu.edu.sa[2]

Abstract

Data Science (DS) is an interdisciplinary topic that is applicable to many domains. In this preliminary investigation, we use *caselet*, a mini-version of a case study, as a learning tool to allow learners to practice data science problem solving (DSPS). Using a dataset collected from a real-world classroom, we performed correlation analysis to reveal the structure of metacognitive and cognitive processes. We also explored the similarity of different DS knowledge components based on students' performance. In addition, we built predictive models to characterize the relationship between metacognition, cognition and learning gain.

Introduction

Solving DS problems requires decision-making skills. The data science courses are often heavily biased toward teaching component skills, including a conceptual understanding of methods and procedure skills, which prepares learners to know WHAT and HOW. However, this is quite different from what learners are required to do in the real world, where they solve real-world data science problems by answering WHAT, WHEN, and WHY. This is a giant and challenging gap traditionally filled by project-based learning or an internship where learners work on a mentored course project or through an internship. Those two pedagogical learning approaches have drawbacks (e.g., a lack of diversity where learners are limited exposure to one type of problem in one project.) DSPS differs from declarative and procedure knowledge, which requires different knowledge organizations. Thus we need explicit instruction in DSPS. As a learning science principle, design rationale for caselet (Chen and Dubrawski 2018), with various characteristics such as adaptive expertise, apprenticeship learning, and deliberate practice, is crucial. Caselets are scalable deliberate practice tool for DSPS, which requires strategic decision-making skills. Each caselet includes problem context, data profile, close-ended multiple-choice questions, and feedback with explanations. All caselet questions are categorized into knowledge components linked to DS competencies (Koedinger, Corbett, and Perfetti 2012). Besides, researchers have recently shown increased interest in the metacognition process while learners solve problems (Winne and Azevedo 2014), which plays a role in this learning process, so it has been applied in this study. This study analyzes DSPS practice to designate the correlation between metacognitive and the cognitive processes, inspect the similarity of DS knowledge components and predict learning gain using the metacognitive and cognitive processes as predictors.

Learning Context & Data Collection

We collected the data from a graduate-level data science course in a public university, with 24 graduate students enrolled. Seven learners are female, and four of the learners are Ph.D. students. The course covers DS techniques and utilizes various learning pedagogical approaches, including problem-based and case-based learning (Allchin 2013). Moreover, learners gain knowledge throughout the semester through declarative knowledge. Besides, we introduce the learners to the DSPS practice by solving seven real-world problems in caselet format. The data includes three assessment stages: pre-assessment, process, and post-assessment. Learners keep tracking their metacognition processes while and after answering caselets. In addition, learners evaluated their metacognitive awareness inventory (MAI) as a self-reported metacognitive assessment (Young and Fry 2008).

Exploratory Data Analysis

To reveal any potential non-linear relationship between the learning gain and different factors, we plotted locally weighted scatter plot smoothing (LOWESS). It exposes a negative relationship between pre-assessment and learning gain, i.e., high prior knowledge is correlated with lower learning gain since the course is not scaled based on the learners' prior knowledge. The LOWESS plot shows two negative values of learning gain, which may indicate some learners may not attain knowledge from the whole course, which requires more investigation. For an exploration purpose, a graphic representation was applied to pre-assessment, the caselets, and the self-reported metacognition, only edges corresponding to significant pairwise correlation are retained. The graph (Figure 1) shows that cognitive processes are highly correlated, and metacognitive assessments (knowledge about regulation (KC) and regulation of cognition (RC)) are also highly correlated with some correlations between metacognition and cognition.

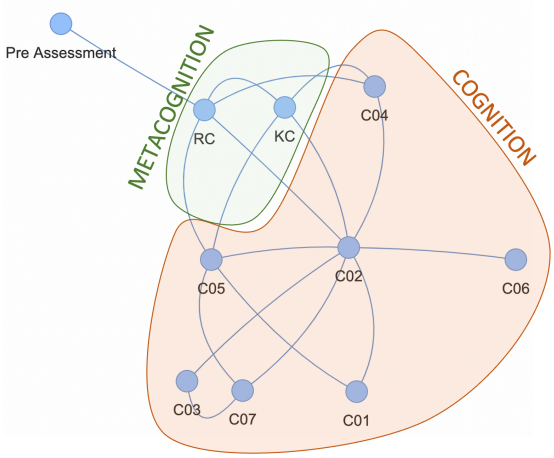

Figure 1: Pairwise correlation among metacognition and cognition assessments, represented as a graph

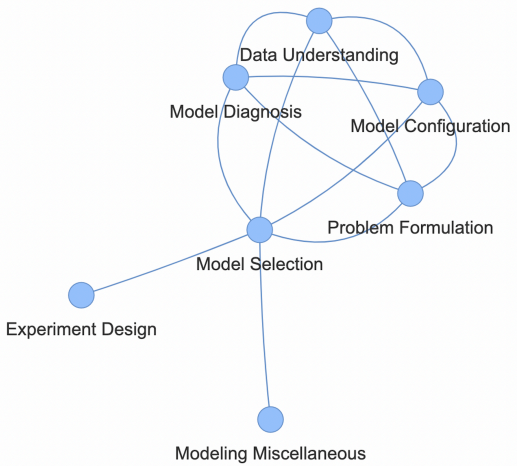

Figure 2: Graph showing similarity among caselets' knowledge components, based on cosine similarity calculated from caselet performances, only edges with high cosine similarity are retained.

To identify contributing factors to learning gain, we opted for mutual information gain (Ross 2014) which measures how much a variable can tell us about the learning gain. It also shows the reduction in uncertainty about learning gain given knowledge of other assessments. We ended up with pre-assessment, caselet score, and assignment score, respectively, as the most informative predictors.

To disclose KCs' structure, we used cosine similarity (Figure 2) to demonstrate how KCs are analogous based on learners' caselet performance. Model diagnosis, data understanding, and model configuration knowledge components are the most similar components.

Predictive Model

We explored various regression models to predict the learning gain using metacognition (MAI) and cognition (pre-assessment and DSPS) as predictors and using Leave-One-Learner-Out to avoid over-fitting. Starting with linear regression (LR) as a linear baseline model, we noted that pre-assessment and MAI were the best predictors, with a 12.4% mean absolute percentage error (MAPE). Moving to non-linear models, we applied support vector machine (SVM) and random forest (RF), where SVM shows an increase in the MAPE. However, RF has an outstanding performance compared to LR and SVM, where pre-assessment has an 8.9% MAPE. Besides, adding more predictors to RF did not improve the performance, where the high correlation among pre-assessment, MAI, and caselets may be the reason.

Conclusion

The current phase of this research is a pilot study of DSPS deliberate practice (caselet), evaluating its effectiveness and measuring learning gain in a real-world teaching context, considering metacognition and cognition. Exploratory data analysis reveals interesting multivariate and non-linear relationships among metacognition, cognition, and learning gains. In addition, we run machine learning experiments to predict the learning gain using metacognitive and cognitive factors. It demonstrates that random forest was able to model the non-linearity better than the linear baseline model and thus has better predictive performance. In future work, we aim to collect more data and overcome various limitations because of the small sample size. We are in the process of designing an interactive learning environment to support large-scale caselet practices that allow us to collect cognitive and metacognitive trace data on a large scale. We are also investigating the approaches to model learning growth at the knowledge component level to gain deeper understanding of the learning process.

References

Allchin, D. 2013. Problem-and Case-Based Learning in Science: An Introduction to Distinctions, Values, and Outcomes. *merican Association for the Advancement of Science*, 12.

Chen, L.; and Dubrawski, A. 2018. Accelerated Apprenticeship: Teaching Data Science Problem Solving Skills at Scale. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–4. New York, NY: Association for Computing Machinery.

Koedinger, K. R.; Corbett, A. T.; and Perfetti, C. 2012. The Knowledge Learning Instruction Framework: Bridging the Science Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, 36.

Ross, B. C. 2014. Mutual Information Between Discrete and Continuous Data Sets. *Public Library of Science one*, 9.

Winne, P. H.; and Azevedo, R. 2014. Metacognition. *Cambridge University Press*, 63–87.

Young, A.; and Fry, J. D. 2008. Metacognitive Awareness and Academic Achievement in College Students. *Journal of the Scholarship of Teaching and Learning*, 8.