Using Pattern Recognition Techniques to Analyze Educational Data

Camilo Vieira[1], Alejandra J. Magana[1], Mireille Boutin[2]
Computer and Information Technology[1] , School of Electrical and Computer Engineering[2]
Purdue University
West Lafayette, IN
cvieira@purdue.edu

***Abstract*—This paper proposed a workshop to introduce the use of computational tools and methods to analyze educational data. The workshop will demonstrate three different contexts in which these tools can be used to visualize and characterize patterns within educational data, and validate them using statistical techniques. Participants in this workshop will have the opportunity to learn how to implement these methods using R programming language.**

Keywords—pattern recognition; education; cluster; visualization; education

I. GOAL

This special session will introduce novel approaches to analyze educational data by taking advantage of the affordances of pattern recognition techniques. Participants in this special session will have a hands-on experience in using computational tools and methods to analyze qualitative or quantitative educational data and supplement the analysis with pattern recognition methods. By the end of this session, the participants will be able to describe different approaches to analyze educational data, and the conditions under which these approaches work.

II. JUSTIFICATION

Computational science has been described as "the application of computer science to solve problems across a range of disciplines" [1]. Science and engineering disciplines are now taking advantage of computers' capabilities to process large amounts of data, create complex visualizations of data, and simulate complex phenomena. As this shift occurs, educational researchers have started to identify different ways in which they can use computation to support their inquiry process to supplement traditional educational research methods in the social sciences. This special session will provide participants with hands-on experiences on how to use pattern recognition methods in the context of engineering education research.

III. INTERACTION

The activities will be divided in three parts:

A. Introduction

- The organizers will introduce the session and themselves, and will provide a context for the workshop. (10 minutes)
- Participants will be asked to discuss in small groups their own experiences using computational tools and methods for data analysis (5 minutes), and will share the most interesting and innovative with the whole group (5 minutes)

B. Implementation of the Methods (50 minutes)

The participants will work collaboratively to understand, implement, and evaluate three different computational methods to analyze qualitative and quantitative educational data: (1) clustering qualitative data; (2) add-on preferential groups; and (3) creating visualizations of educational data.

- The organizers can provide a maximum of eight laptops with the required software installed: the most recent version of R and R Studio. The participants are welcome to bring their own computer with the software already installed.
- The organizers will provide sample data sets and scripts for the participants to implement these three methods.

C. Reflection (10 minutes)

Participants reflect and share with the group how they can use the methods in their own practices.

IV. DESCRIPTION

This special session will be focused on the use of pattern recognition techniques to visualize and characterize educational data. For instance, different clustering methods can be used to group students based on their grades or the perceived usefulness of a set of instructional methods. This approach allows researchers to understand the educational phenomenon beyond summative values (e.g. mean and standard deviation), or correlational values (e.g. Pearson correlation).

978-1-5090-5920-1/17/$31.00 ©2017 IEEE

Figure 1 shows the distribution of students' responses to two different scale questions. Each axis corresponds to the perceived usefulness of a given instructional approach. A larger circle represents a larger number of students choosing that value for the two scales.

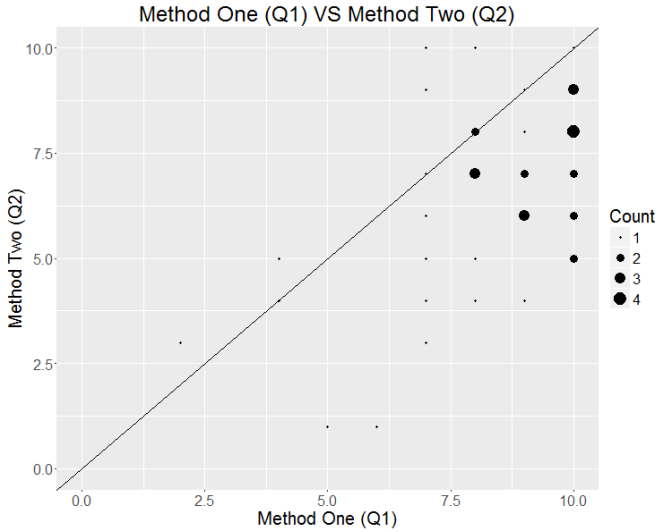

Fig. 1. Sample visualizations of distribution of students' response for two scale questions

This distribution can be represented as groups of students with different preferences. Figure 2 shows how we can compare students' perceived usefulness and preferences about two instructional methods, and identify six different clusters of students. For instance, students in the lower right corner (17.86%), found Method One more useful than Method Two. Conversely, students in the higher left corner (1.19%) found Method Two more useful than Method One. However, there is still a group of students in the lower left corner (4.76%+2.38%) that did not consider either method as useful. Furthermore, we can validate these clusters using non-traditional statistical techniques such as the permutation test [2].

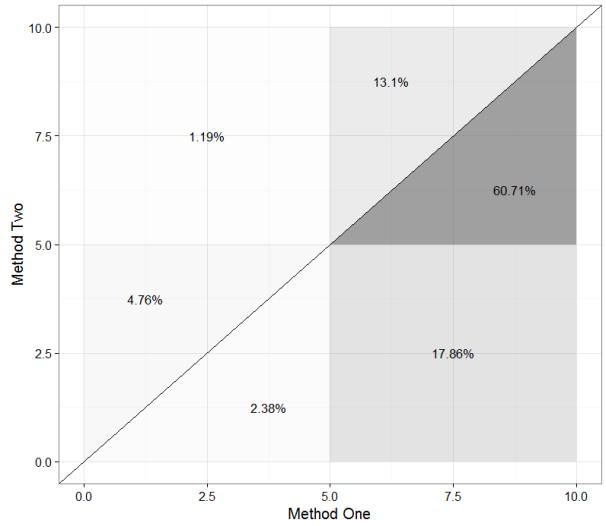

Fig. 2. Sample visualizations of clusters of students based on quantitative (a) educational data

We can also use clusters to group students based on a qualitative coding process, which often becomes a high dimensional problem. For instance, Figure 3 shows how students (rows) were grouped based on the characteristics of their written comments within a sample Python code. Students' in-code comments were first analyzed using a coding scheme that involved 21 categories (columns). Then, using a binary distance between students, we conducted a hierarchical cluster analysis to identify the different ways in which students explained this code. The colors in the plot correspond to each type of explainer that we identified.

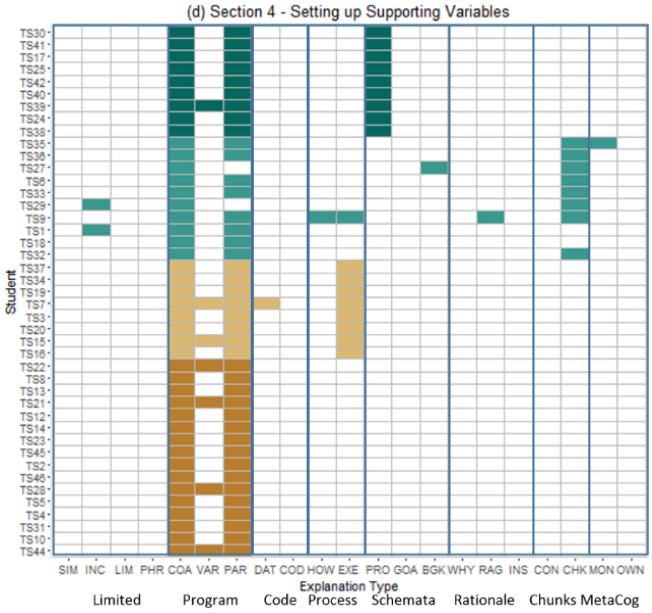

Fig. 3. Sample visualizations of clusters of students based on quantitative (a) educational data

Both examples presented here were developed using the statistical software R. The special session will include hands-on activities with R programming, where the participants will conduct their own analysis and create their own visualizations of qualitative and quantitative educational data. The R code that is required to create these and other visualizations will be made freely available to the participants through [3-4].

V. Time – Outcomes/Future Work

The special session will last for 80 minutes. During the first 20 minutes, the organizers will provide background information about the topics to be discussed during the session, its relevance, and sample contexts in which this could be applied. During the next 50 minutes, the participants will work collaboratively on their own computers, analyzing sample data provided by the organizer. By the end of the session, participants will be able to describe different pattern recognition techniques and how they can be used to analyze quantitative and qualitative educational data. Participants will also be able to create a simple program in R to analyze this data.

VI. Personnel

Dr. Camilo Vieira is a postdoctoral researcher in the Department of Computer and Information Technology at Purdue University. His research focuses on integrating computational science and engineering within the engineering curricula, as well as using computational tools and methods to support educational research and educational practices. He is the author of the R code that will be shared during the session.

Dr. Alejandra Magana is an Associate Professor in the Department of Computer and Information Technology and Engineering Education. She conducts research on affordances for cyberlearning and engagement as well as computing technologies for STEM learning.

Dr. Mireille Boutin is an Associate Professor in the School of Electrical and Computer Engineering at Purdue University. Dr. Boutin has extensive expertise in research and teaching of machine learning, signal processing and big data analytics. She and her students have developed an automatic summarization method for large bodies of text documents and a method for automatic clustering of high-dimensional data.

Acknowledgment

This work has been partially supported by the NSF grant award EEC#1544244 .

References

[1] ACM/IEEE-CS Joint Task Force on Computing Curricula, "Computer Science Curricula 2013," ACM Press and IEEE Computer Society Press, 2013. *http://dx.doi.org/10.1145/2534860*

[2] M. D. Ernst, "Permutation methods: a basis for exact inference." Statistical Science. 2004; vol 19, pp. 676-685.

[3] C. Vieira, "R Code for analyzing students' explanations of programming code", 2017, https://doi.org/10.5281/zenodo.555934

[4] C. Vieira, "R Code for the APG Model", 2017, https://doi.org/10.5281/zenodo.555933