Sequence Analysis of Learning Behavior in Different Consecutive Activities.

Independent Study

By:
Abdulelah Abuabat

Advised By:
Peter Brusilovsky

School of Computing and Information
University of Pittsburgh

Pittsburgh, USA

June 7, 2018

1. Introduction

Day after day, research shows how vital sequential analysis is, especially since many settings depend on it. One of these settings is studying user behavior (student, customer…etc.) and trying to discover if there are concrete results and indications that can be used to help this user to gain more benefits from the system. My interest lies in analyzing both independent and dependent activities sequentially, and check what knowledge can be acquired, and then how to apply this knowledge to help the user at the end.

In this study, I am interested in examining sequence independent activities in an educational system, called Mastery Grids[1], and check if users can be identified by their behavior, and how stable this behavior is overall. Additionally, I applied the hierarchal analysis technique to see if there are common behavior traits among students, and what it looks like. I got inspired by the work of Guerra et al. (2014), where they study the behavior of the students in one type of activity, which in this case is parameterized exercises. Here, I attempted to include different type of activities and see if the final outcomes are still consistent with [2] findings.

2. Dataset

In this work, I used the two copies of the *Final_IS0017Fall2016_RawActivity*. The logs stored the students' activities from opening the session, in the *Mastery Grids*, till the student quit. I included all the students, who had the following three activates: animated example, example (WEBEX), and parameterized exercises (QUIZJET). So, I ended with a total of 44 students. The features that I considered from these two logs were: student username, session, topic, and the activity itself. The maximum number of session in these logs were 42, and the number of topics were 21.

[1] A platform to help student to practice examples and exercises for Java programming.

2.1 Activities Labeling[2]

In each sequence, I considered the duration of each activity, and the correctness if it is an exercise, in (see figure 1). An example of a sequence is: "AnEx ex f P p", which means that the student spent more than the median time in doing the animated example activity "AnEx" then he/she spent less than the median time in reviewing the example "ex" ... etc. And I used the median since the values were not normally distributed. The sequence started and ended alongside the system session. Thus, if a student shifts between topics, while he/she is in the same system session; for each topic the sequence construction will not be interrupted. At the end, I got around 650 sequences.

Moreover, I tested different approaches in defining the start and end point for the sequence, and I will mention three examples. First, the time among activities must be less than the median of all the activities in order to make sure that the student was working without interrupting and disturbing. Second, each sequence should contain exercises besides the examples, not just one type of activity. Third, each sequence should end with exercises in order to see how valuable and influential the examples and the animated example for the exercise results was. However, I did not get worthwhile results. The reason for that I believe is that these kinds of definitions reduce the amount of sequences pointedly, which affect the pattern mining process. If the dataset was massive and contained more students and observations, I believe that the mentioned definitions can lead to more interesting results.

---

[2] labeling abbreviations:
ex = spend more than the median in the constant example (WEBEX).
Ex = spend less than the median in the constant example (WEBEX).
AnEx = spend more than the median in the animated example.
anex = spend less than the median in the animated example.
P = quiz - Pass - user spends more than the median.
F = quiz - Fail - user spends more than the median.
p = quiz - Pass - user spends less than the median.
f = quiz - Fail - user spends less than the median.

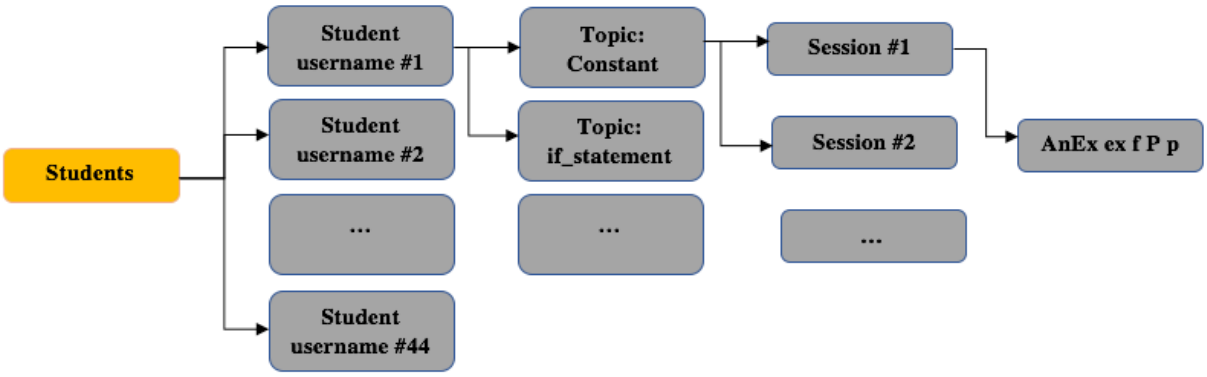

Figure 1: Labeling Structure.

3. Method

In this phase I examined the results of labeling those 650 sequences by conducting a sequential pattern mining. From the sequential pattern mining results, I tried to study two things: 1) The ability to recognize the students based on their pattern in doing different activities. To study this, I checked the pattern stability for each student. 2) Clustering the students into two clusters and checking which patterns are common in each cluster.

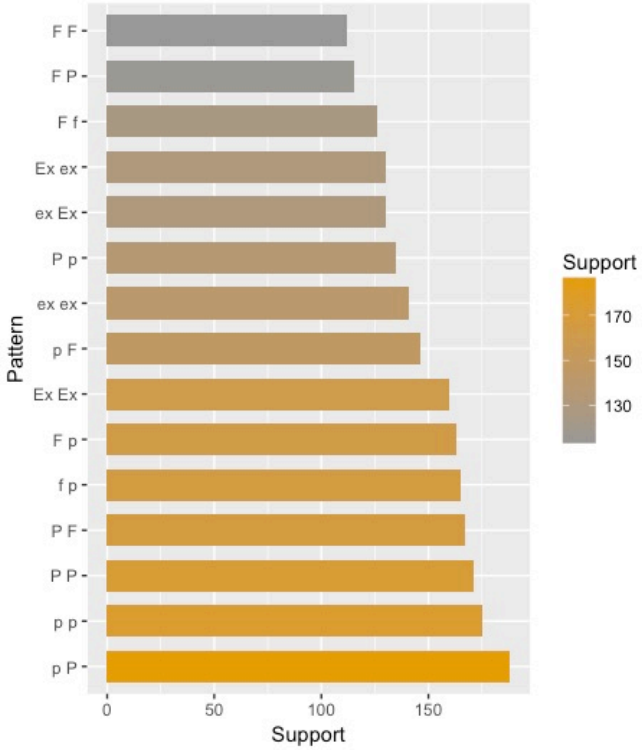

Figure 2: The Common 15 Patterns.

3.1 Sequential Pattern Mining Algorithm

In order to discover the most frequent sequence and pattern among my sequences, I have used the SPAM algorithm [1], which is a sequential pattern mining using a bitmap representation [3]. In this algorithm, I have to decide two things: the minimum support "*minsup*" and the maximum gap. Minimum support is the support of a pattern χ which is the percentage of sequences of the total sequences in the dataset. This contains χ as a sequence or a sub- sequence. After several tests, and based on the results that I got, I found 4% was a practical choice for the minimum support, and 1 for the maximum gap. Lastly, I only considered patterns with 2 as the minimum length in order to better understanding the patterns' results (see figure 2).

3.2 Pattern Stability

To have enhanced insight about the results, I checked how many times each pattern occurred for each student (see figure 3). Then, I normalized the numbers. And if a certain pattern did not occur, I smoothed it by storing it as a very small number, 0.0001, so I can get more sense and precise results in calculating the distance in the next step.

Regarding the stability test, I followed Guerra *et al.* (2014) by splitting the sequences activities for each student randomly, and these two parts were represented as vectors. Next, I checked the distance between each student and his/her other part, and the distance between each student and other students' parts. If the pattern is stable, then the distance between each student's parts will be closer than the distance between each student with other students' parts. To do so, I have used the symmetric version of Kullback-Leibler, which is Jensen-Shannon (JS) divergence [4] to measure the distance between the two distributions. The results showed that the pattern was stable, and the pattern did not happen arbitrarily, and the *self-distance* ($M = 0.370$) is considerably less than the *distance-to-other* ($M = 0.514$). Lastly, to check how significant the different between the mean of the two populations: *self-distance and distance-to-other*, I applied a paired samples t-test, and the results ($t = -7.84$, $p < 0.001$) confirmed that the differences were not at random.

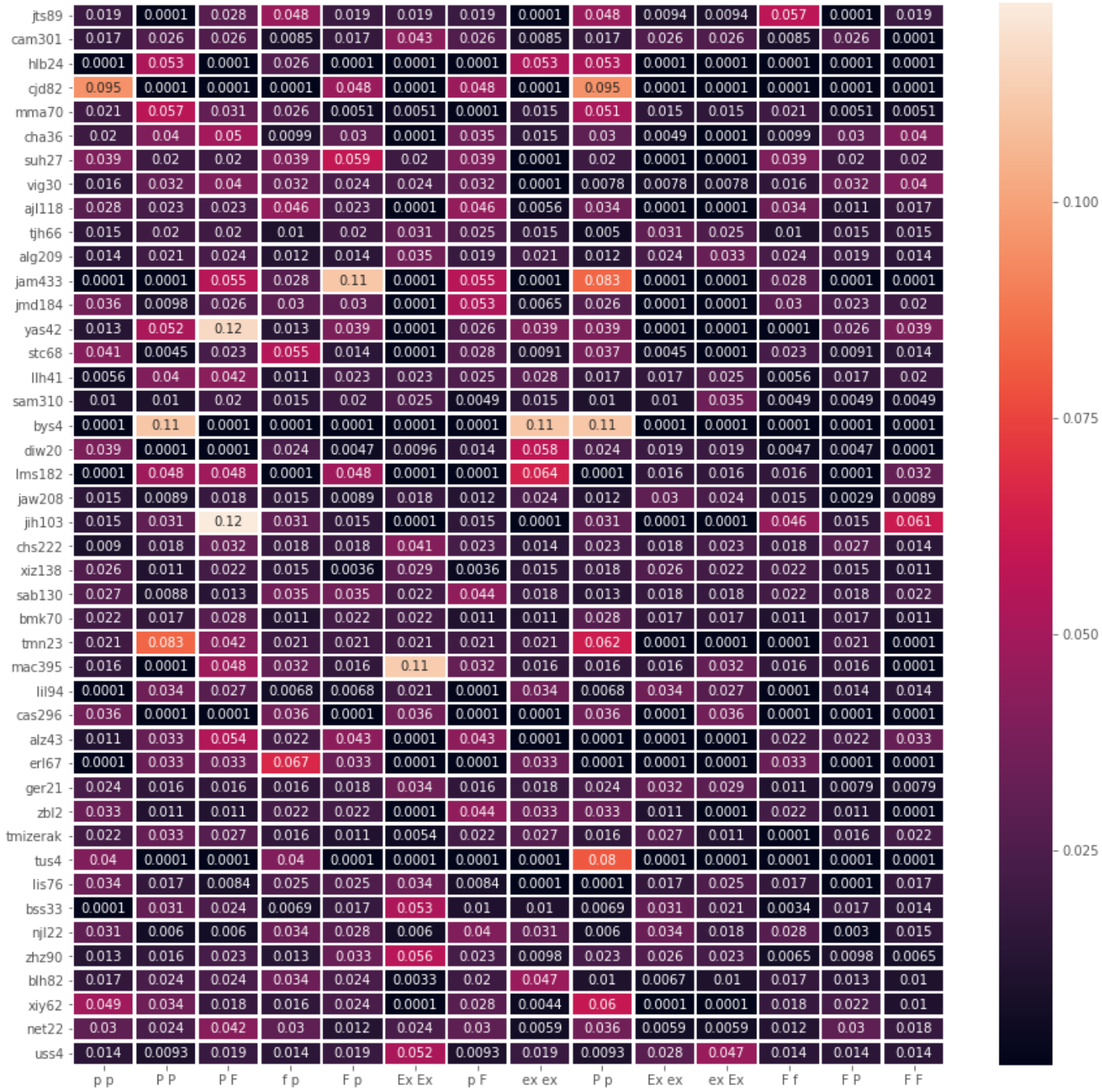

Figure 3: The Heat Map of the Common 15 Patterns for each Student. The numbers have been normalized.

3.3 Clustering Students Based on their Pattern

Another way to look at the dataset is by dividing the students into different clusters, because this might help to infer if there is a successful pattern or not, and if we can distinguish between a pattern that can lead to positive results and a pattern which can lead to undesirable results. Regrettably, the dataset is not that big that it can help me come to these kinds of conclusions. However, I applied an unsupervised machine learning technique to give me at least a glance about the dataset, and to show me the general behavior for each group. In this context, I applied a *Hierarchical Clustering*, using *Ward* method, with $k = 2$. (See figure 4 and 5).

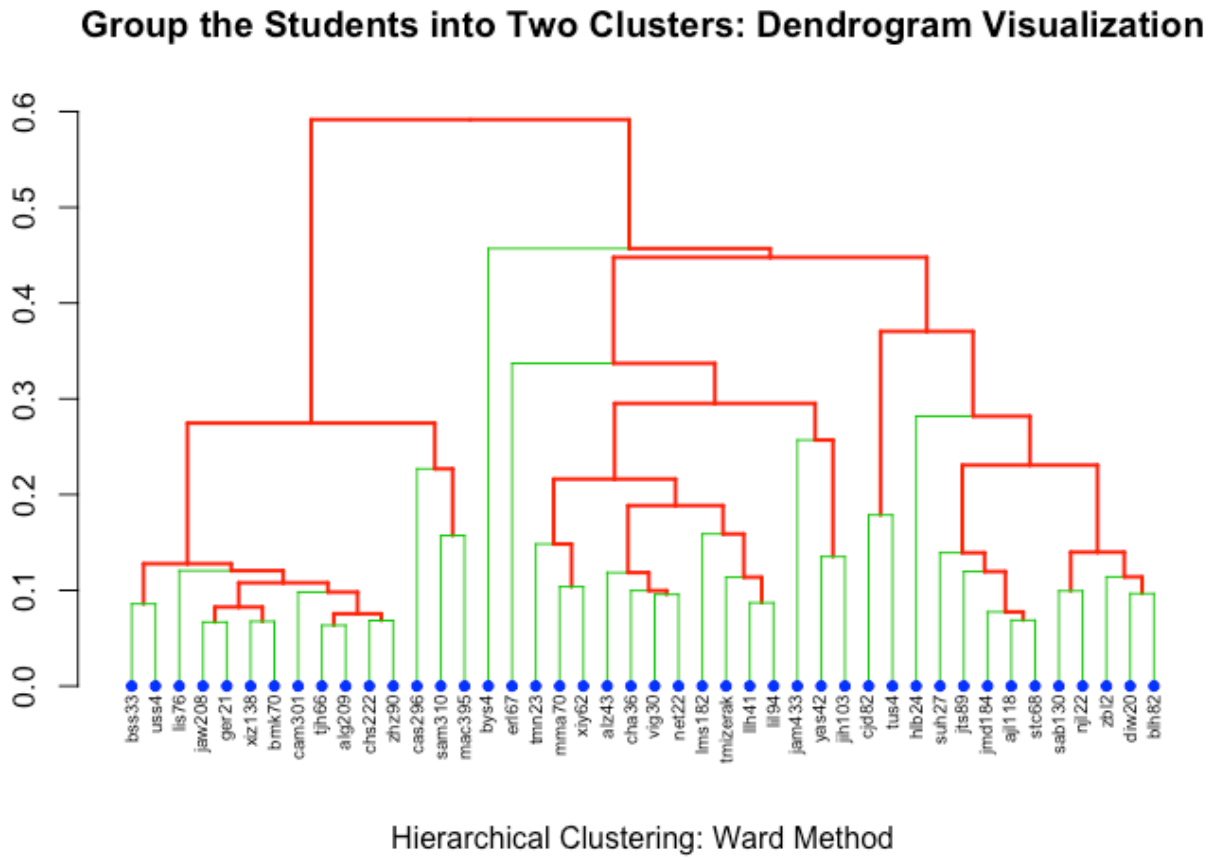

Figure 4: Dendrogram Visualization for the two clusters by using Ward method.

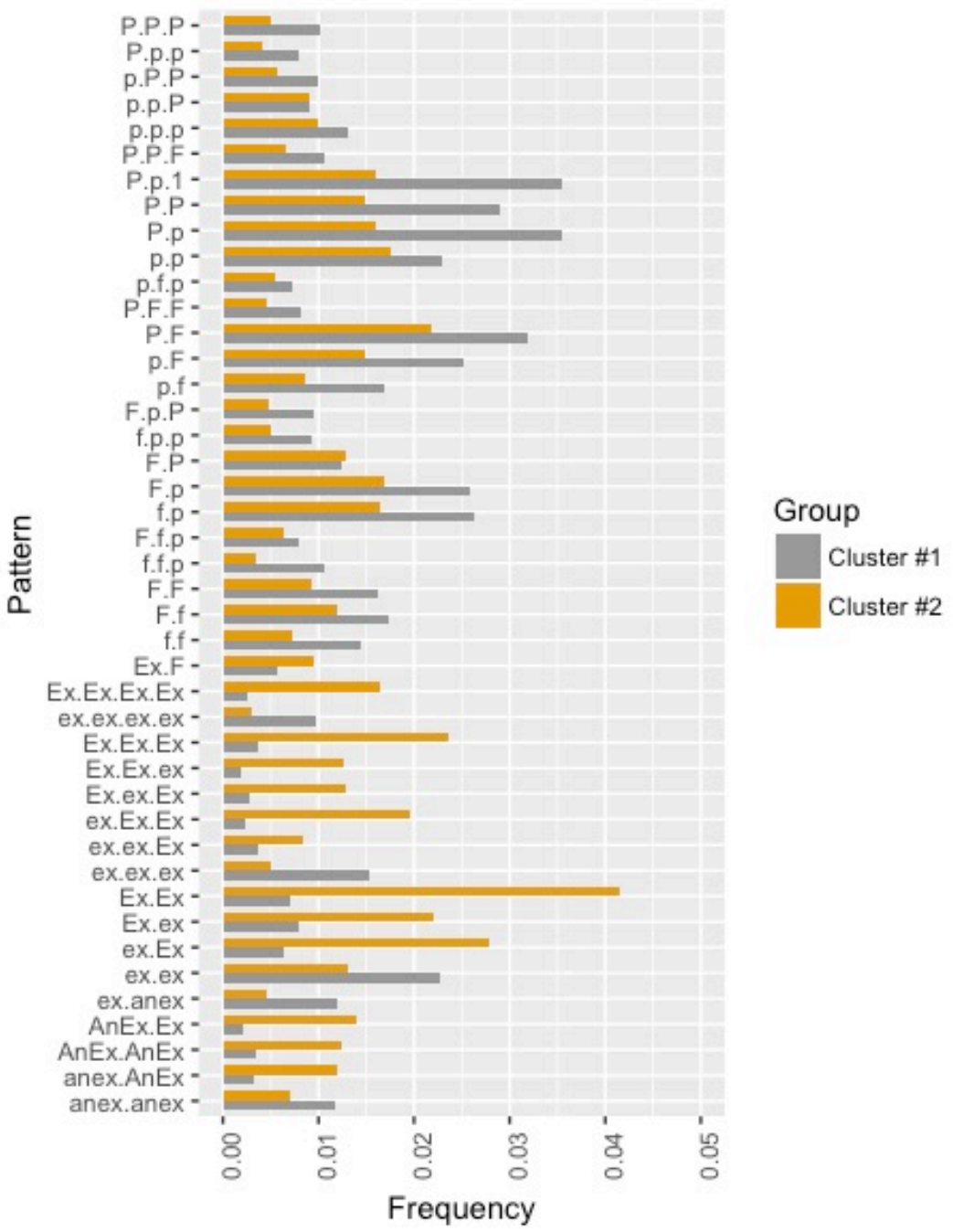

Figure 5: A Bar chart that shows the frequency for each pattern in each cluster.

As you can see, it seems that Cluster #1 spent more time in doing the parameterized exercises and also tended to repeat the exercises and gain positive results. On the other hand, Cluster #2 looks like they paid more attention to the examples and the animated examples.

4. Discussion and Conclusion

The overall results of my study are promising, and it is possible to get more insightful and tangible knowledge regarding user behavior in such a system. It can also lead to the level of a heuristic system, which can guide the students in a more successful and effective manner. However, the dataset needs additional observations and extra features such as the final grade for the course in order to reach this level of intelligence.

In my future work, I would like to focus my study on the extent to which the user can be identified by his/her behavior, especially when performing a different type of activity. Furthermore, I would like to examine if it is possible to change the user behavior and steered in a different direction.

5. Acknowledgment

I would like to thank Jordan Ariel for his great assistance and guidance, and for his help to review my results.

References

[1] Ayres, J., Gehrke, J., Yiu, T., & Flannick, J. (n.d.). Sequential PAttern Mining using A Bitmap Representation. Retrieved from http://www.philippe-fournier-viger.com/spmf/SPAM.pdf

[2] Guerra, J., Sahebi, S., Brusilovsky, P., & Lin, Y.-R. (2014). The Problem Solving Genome: Analyzing Sequential Patterns of Student Work with Parametrerized Exercises. *Proceedings of the 7th International Conference on Educational Data Mining*, (June), 153–160.

[3] http://www.philippe-fournier-viger.com/spmf/SPAM.php

[4] Majtey, A. P., Lamberti, P. W., & Prato, D. P. (2005). Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. Retrieved from https://arxiv.org/pdf/quant-ph/0508138.pdf

[5] Kinnebrew, J. S., & Biswas, G. (n.d.). Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. Retrieved from https://pdfs.semanticscholar.org/8dd9/3074f9cbb846c0a9951b2555190b84599748.pdf