# Detecting Outlier Behaviors in Student Progress Trajectories Using a Repeated Fuzzy Clustering Approach

Colm P. Howlin
Realizeit
Dublin, Ireland
colm.howlin@realizeitlearning.com

Charles D. Dziuban
University of Central Florida
Orlando, FL
charles.dziuban@ucf.edu

## ABSTRACT
Clustering of educational data allows similar students to be grouped, in either crisp or fuzzy sets, based on their similarities. Standard approaches are well suited to identifying common student behaviors; however, by design, they put much less emphasis on less common behaviors or outliers. The approach presented in this paper employs fuzzing clustering in the identification of these outlier behaviors. The algorithm is an iterative one, where clustering is applied, outliers identified, the data restricted to the outliers, and the process repeated. This approach produces a clustering that is crisp between each iteration and fuzzy within. It arose as a consequence of trying to cluster student progress trajectories in an adaptive learning platform. Included are results from applying the repeated fuzzy clustering algorithm to data from multiple courses and semesters at the University of Central Florida, (N=5,044).

## 1. INTRODUCTION
Personalization holds the promise of making learning more engaging and effective for students. Each student can receive personalized feedback and guidance based on their interaction with the learning material and their current needs and goals. Key to being able to provide this is an understanding of the full range of learning behaviors that students can exhibit, and the driving forces behind them. Truly personalized learning needs to understand not just the most common behaviors, but also those that are more atypical or outliers.

A variety of techniques have been employed to uncover student behaviors in different learning contexts [22]. Clustering is a common approach with a considerable range in both the applications and the algorithm employed [25]. Applications have included adapting question delivery, promoting group-based collaboration, and the characterization of atypical student behavior.

This work presents a clustering approach to automatically detect and quantify the range of behaviors, including the outliers, that are evident in student progress data, in order to provide feedback to instructors on their student's behaviors. This goal throws up two restrictions on our approach. First, the clustering of behaviors must be fully automated. Not all instructors will have the required knowledge to make decisions such as picking the parameters of the clustering algorithm. Therefore these decisions need to be handled by the algorithm. Second, the output from the clustering must be readily interpretable by an instructor, including both understanding what makes a cluster a cluster, but also easily understand the differences between clusters. These two restrictions provide a means of measuring the ultimate effectiveness of the algorithm and the quality of the clusters that it produces.

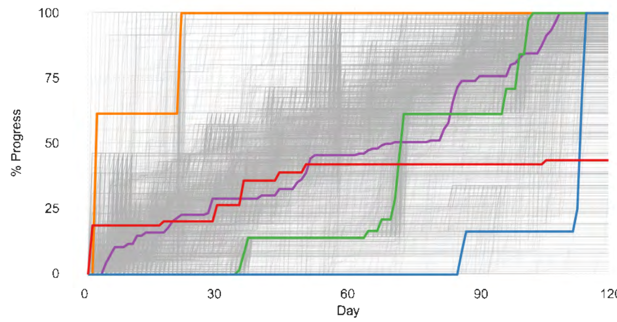


**Figure 1: Student progress trajectories. The gray lines show the trajectories for all 5,044 students at UCF. The colored lines highlight several individual trajectories.**

In [11], the authors examined student progress data against time for an online course delivered at the University of Central Florida (UCF) through the Realizeit adaptive learning platform. The course was self-paced with students free to set their rate of progress. While most set a steady, consistent pace over the 15-week term, some students set a very different pace. These outliers roughly fall into two categories: students who race ahead of the rest, and those who fall behind, leaving all their learning to the last minute.

Figure 1 provides an understanding of the challenges when clustering these progress trajectories. The x-axis represents time in days, and the y-axis is progress measured as the percentage of concepts mastered. The progress trajectories for 5,044 students across 51 online course instances in 9 terms

Colm Howlin and Charles Dziuban "Detecting Outlier Behaviors in Student Progress Trajectories Using a Repeated Fuzzy Clustering Approach" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 742 - 747

at UCF are shown in gray. Each line represents a single student. Patterns are difficult to distinguish, but the consistent trajectory of most students through their course is evident along the diagonal. Five progress trajectories (colored lines) have been singled out to highlight the range of possible behaviors. The challenge in clustering this data comes from the fact that clustering algorithms, by design, attempt to group the data into as few clusters as possible and therefore put much less emphasis on outliers. They seek the most common patterns. Our goal is to find both common and outlier behaviors.

Our approach draws inspiration from He et al. [16, 17] who used clustering to search for hidden communities in social networks. In their work, they first used clustering to discover the most apparent communities. They then decreased the weights on the edges in the social network that represented these communities. Repeating the clustering uncovers previously hidden communities. Our approach, repeated fuzzy clustering (RFC) uses a similar technique where clustering is applied, outliers identified, the data restricted to the outliers, and the process repeated. The purpose of this paper is to describe and demonstrate the RFC algorithm.

Algorithmic clustering methods are essentially "blind" in that there is no linguistic functioning in their process. The categories identified are impervious to shared characteristics that ground themselves in cultural beliefs. However, the linguistic and algorithmic categorization processes do have common intersections. Linguistically, Rosch, [23, 24], described this as prototype theory where through any number of cultural and societal processes what is the best representational icon of a category is formed by our preconceived notions. Adaptive learning provides diverse paths to success, many of which may not align with our preconceived notions of what constitutes successful or unsuccessful behavior. Clearly, clustering algorithms have assumptions built into them a priori but once built are not influenced by preconceptions. The questions we ultimately wish to address involves whether or not the clustering of student trajectories can provide a foundation for category characteristics through the multiple lenses of methods, education, linguistics, and prototype theory and should they make educational sense how can we use them to improve learning? [20].

# 2. APPROACH

Here we provide an outline of the RFC algorithm. In the following subsections, we provide the specifics on our implementation of each function, although it is possible to alter these to suit other needs or implementations. The algorithm proceeds by first grouping students using fuzzy clustering for a range of values of $k$ (the number of clusters) - lines $5 \rightarrow 7$. Validity indices are calculated for each solution, and the most appropriate number of clusters is chosen - line 8. The algorithm then proceeds by identifying outliers and removing them from the data. The algorithm then reapplies the clustering creating a more compact solution. This part of the process repeats until the algorithm identifies no new outliers - line 10. The data is then limited to the previously identified outliers on this loop - lines $11 \rightarrow 12$. The whole process then repeats with the data filtered to the outliers.

**Algorithm** Repeated Fuzzy Clustering

1: $D$ is the student data
2: $Outliers =$ All students
3: $i = 0$
4: **while** $|Outliers| > tol \;\&\; i < M$ **do**
5: **for** $k$ in $1 : k_{max}$ **do**
6: $F_k = FuzzyCluster\,(k, D)$
$V_k = ValidityIndices\,(F_k)$
7: **end for**
8: Select $k$ using $V$
9: $i = i + 1$
10: $FC_i = RefineClustering\,(F_k)$
11: $Outliers = IdentifyOutliers\,(FC_i, D)$
12: $D = D \setminus Outliers$
13: **end while**

There are three parameters to the algorithm: $k_{max}$ is the maximum number of clusters to consider at each repetition; $tol$ is limit on the number of outliers that must be present for the algorithm to repeat; $M$ is the maximum number of repetitions. There are four functions within the algorithm where choice is possible. These enable the tailoring of the algorithm to specific needs or implementations. The choices here can lead to the introduction of additional parameters.

## 2.1 Fuzzy Clustering

Fuzzy clustering is used to determine the grouping of students within a loop. The choice of fuzzy, as opposed to crisp, is because it provides a membership value for each student in each cluster. This is relied upon to determine outliers, S2.4. In this implementation fuzzy k-means [10] is used, although it would be possible to use any other fuzzy clustering algorithm in its place [14]. An effect of using fuzzy clustering in our approach is that the algorithm produces crisp divisions between loops and fuzzy divisions within.

## 2.2 Validity Indices

Validity indices provide a quantitative measure of cluster validation. Their calculation is a fundamental part of the clustering process and provides guidance when deciding on $k$, the number of clusters. There is a huge range of cluster validity indices [2] with a large subset focused on fuzzy clustering [26]. In this implementation, we use the six available in the FClust R package [12]. These include the Silhouette index [19], Fuzzy silhouette index [5], Partition coefficient [3], Modified partition coefficient [8], Partition entropy [4],and Xie and Beni index [27]. For each clustering solution, we record the value of $k$ recommended by each validity index. The final value of $k$ is the mode of these recommendations. In the case of two possible values for $k$, we chose the smallest.

## 2.3 Refining Clusters

Refining the clustering solution is an optional step that enhances the compactness of the final clusters on each loop. Given a solution, outliers once identified are removed from the data. The clustering procedure is then rerun with the same value of $k$ to derive a tighter clustering solution that better represents that data and students that remain. This process repeats as required until a stable solution emerges and no outliers are present.

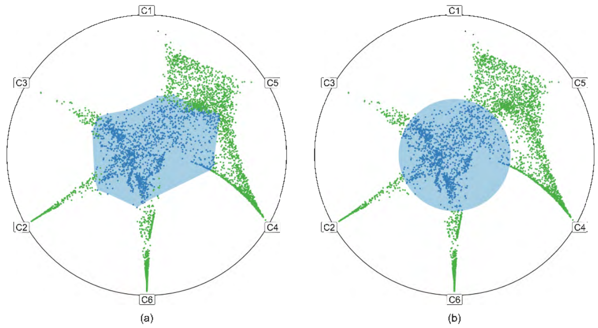


**Figure 2: Radviz of membership for a solution with $k = 6$. Outliers are shown in blue. (a) Outliers as identified by (1). (b) Outliers as identified by (2).**

## 2.4 Identifying Outliers

Identifying outliers is the most crucial step in the RFC algorithm. This process places the split in the data for each loop of the algorithm. Many strategies are possible with the choice depending on the application and chosen clustering procedure. [21], [15] and [13] all explore identifying outliers as part of the k-means clustering process. This is done for several reasons including creating more compact clusters. These process generally rely on distance measures to identify the outliers in the data.

The method presented here identifies outliers using the membership values from fuzzy clustering with two versions considered below. The rationale behind both of these approaches is that they seek a solution which places observations mostly within one or two clusters. Any observation split among three or more clusters is an outlier. For an instructor having a student predominately within only one or two clusters should help simplify the task of interpreting their behavior.

The first and simplest version of identifying outliers uses the maximum membership value $m_i$ and the sum of the two highest membership values $s_i$ for an individual observation $i$. The condition classifies an observation as an outlier if the values of $m_i$ or $s_i$ fall below some limit. Equation (1) places a limit of $\frac{1}{2}$ on the value of $m_i$ and a limit on $s_i$ that is increases slowly from $\frac{1}{2}$ with increasing number of clusters $k$.

$$Outliers = \left\{ i \mid m_i < \frac{1}{2} \vee s_i < \frac{1}{2} + \frac{1}{k} \right\} \tag{1}$$

This condition will not work for $k = 2$ as $s_i$ will always equal 1, and the condition on $m_i$ will never be satisfied. In this case, one possible solution would be to place a stricter limit on $m_i$ and drop condition on $s_i$.

The second approach makes use of the Radviz method of visualizing fuzzy cluster membership [18]. Radviz represents each cluster by a dimensional anchor and distributes each dimensional anchor evenly on a unit circle. Each observation corresponds to a point. The visualization connects each point to each anchor by a spring whose stiffness corresponds to that observation's cluster membership for the associated anchor. The position is where the spring's tension is at its minimum. Imagine each anchor pulls on a data point with a strength equal to the cluster membership. The higher the membership value, the stronger the pull and the closer the data point to that anchor. The ordering of the anchors is essential, and work has been completed to determine the optimum position [9].

The advantage of this method is that it makes observations which are evenly split among multiple clusters evident as these will be close to the center of the visualization since they get equally pulled in all directions. Observations that are a member of a small number of clusters will generally be further from the center. An example of a Radviz, created using [1], from one stage of implementing the RFC algorithm with six clusters, can be seen in Figure 2. In part (a), outliers, as defined by (1) with $k = 6$, are colored in blue and are visible in the center of the graph.

An alternative to (1) is to use the position of each observation on the Radviz graph. Here outliers are defined as being those at the center of the graph within some circle of radius $r$ and where $x_i$ and $y_i$ are the Cartesian coordinates of the position of the observation $i$ in the visualization. The parameter $r$ has a similar role to $m$ in the fuzzy k-means algorithm in that it controls the fuzziness of the clusters. The larger $r$, the crisper the clustering.

$$Outliers = \left\{ i \mid x_i^2 + y_i^2 < r^2 \right\} \tag{2}$$

This method has the advantage of also working without modification for the case where $k = 2$, as points become spaced along a straight line. In this case, the condition 2 reduces to $m < r + \frac{1}{2}$. Figure 2(b) displays the outliers as identified by (2) using $r = 0.4$. We can see a significant overlap of points using both conditions.

# 3. EXPERIMENTAL RESULTS

## 3.1 Dataset

The data used to test the algorithm is from UCF's use of the Realizeit platform. The data encompasses $N = 5044$ students across 51 online and blended course deliveries across nine terms from 2015 to 2018. Both spring and fall terms last 15 weeks and the summer term is 12 weeks. The courses cover a range of disciplines including Psychology, Spanish, College Algebra, various Computing courses, and Nursing. UCF uses the platform in a variety of different contexts and the student learning in the platform contributes a more significant element of their final grade in some course than others. The data only contains first-time students; repeat students are filtered out.

## 3.2 Features

It is possible to define a distance metric for the progress trajectories in their raw form and to use the RFC algorithm. However, we can obtain more easily interpretable results for an instructor by extracting features from the trajectories that capture the key behavioral aspects. Through testing and iteration the following six were selected:

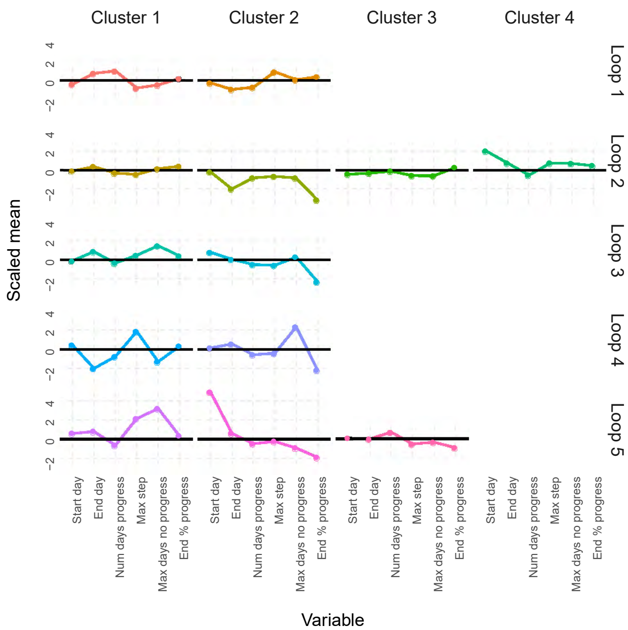


**Figure 3: The center (mean) of each cluster on each of the key features for the normalized UCF data.**

- **Start day** - The first day on which the student made progress.
- **End day** - The last day on which the student made some progress. The day on which the student reached their final % progress value.
- **End % progress** - The percentage of concepts mastered by the student by the end of the course.
- **Num days progress** - The number of days on which the student made progress.
- **Max step** - The single largest jump in progress on a single day.
- **Max days no progress** - Between the start and end day, the largest number of consecutive days on which the student made no progress.

One weakness of the chosen features is that they are only observable after the course is finished making an early prediction of behaviors difficult. Note that the trajectories do not capture all activities completed by the student, just those that increase their progress. For example, practices, revisions, or assessments are not evident in this data.

## 3.3 Clustering

The RFC algorithm made use of the fuzzy k-means algorithm with the fuzzy parameter set at $m = 2$. Note that the data was normalized before using fuzzy clustering. We set $k_{max} = 10$, $M = 10$, and $tol = 0.05N \approx 250$. The algorithm completed 5 loops and automatically produced 13 clusters in total. The breakdown of clusters per loop and the weight of each cluster is provided in Table 1. We use weight since a student belongs only partially to any one cluster.

**Table 1: Cluster and Loop weights $W$, % and deviation from average $\delta$.**

| Cluster | $W$ | % | $\delta$ | Total $W$ | Total % | $\bar{\delta}$ |
|---|---|---|---|---|---|---|
| C1L1 | 1685.6 | 33.42 | 0.60 | 3244 | 64.31 | 0.57 |
| C2L1 | 1558.4 | 30.90 | 0.55 | | | |
| C1L2 | 280.6 | 5.56 | 0.29 | 915 | 18.14 | 0.70 |
| C2L2 | 125.3 | 2.48 | 1.25 | | | |
| C3L2 | 223.7 | 4.44 | 0.38 | | | |
| C4L2 | 285.3 | 5.66 | 0.87 | | | |
| C1L3 | 202.0 | 4.00 | 0.61 | 293 | 5.81 | 0.68 |
| C2L3 | 91.0 | 1.80 | 0.75 | | | |
| C1L4 | 129.4 | 2.57 | 1.12 | 340 | 6.74 | 1.08 |
| C2L4 | 210.6 | 4.18 | 1.04 | | | |
| C1L5 | 63.3 | 1.25 | 1.29 | 252 | 5.00 | 1.07 |
| C2L5 | 76.9 | 1.52 | 1.52 | | | |
| C3L5 | 111.8 | 2.22 | 0.44 | | | |

The first loop captures the standard approach of applying the fuzzy k-means algorithm once and stopping (if the refinement step is excluded). It is the clusters on loop two to five that are new, and it is here that we find the outlier behaviors that would be missed by the standard approach. Notice that the number of students clustered on each loop generally decreases as the loop count increases. Another point is that these “outliers” account for over 30% of the students.

Figure 3 visualizes the center (mean of the normalized data) of each cluster for each feature. Figure 4 displays the trajectories belonging to each cluster with a membership greater than 0.5. The students with the highest membership values for each cluster are shown in black, and these can be taken as prototypes for each cluster to help interpretation. Note that some of the trajectories in each cluster vary considerably from the prototypes due to the fuzzy nature of the clusters and likely have membership values close to 0.5. The noise present in the clusters on the final loop suggests that perhaps the algorithm stopped too early and allowing additional loops could uncover new behaviors.

From an examination of these graphs, we can see that some outlier behaviors are entirely different from the most common behaviors found on the first loop. There are certain similarities in some cases but enough of a difference to make them worthy of being categorized as separate behaviors.

The clusters found on loop one represent more successful behaviors in that the students generally finish over 50% of the concepts. The first that represents unsuccessful behavior appears on loop two, with more appearing on later loops. Below we provide notes on some of the individual behaviors. A detailed analysis is beyond the scope of this paper.

- Students in cluster 1 on loop 4 (C1L4) master all the concepts in a short period right at the start of the course.
- C2L5 are the students who generally did too little too late.
- C3L5 are students who start well but for some reason stopped with about a month to go.

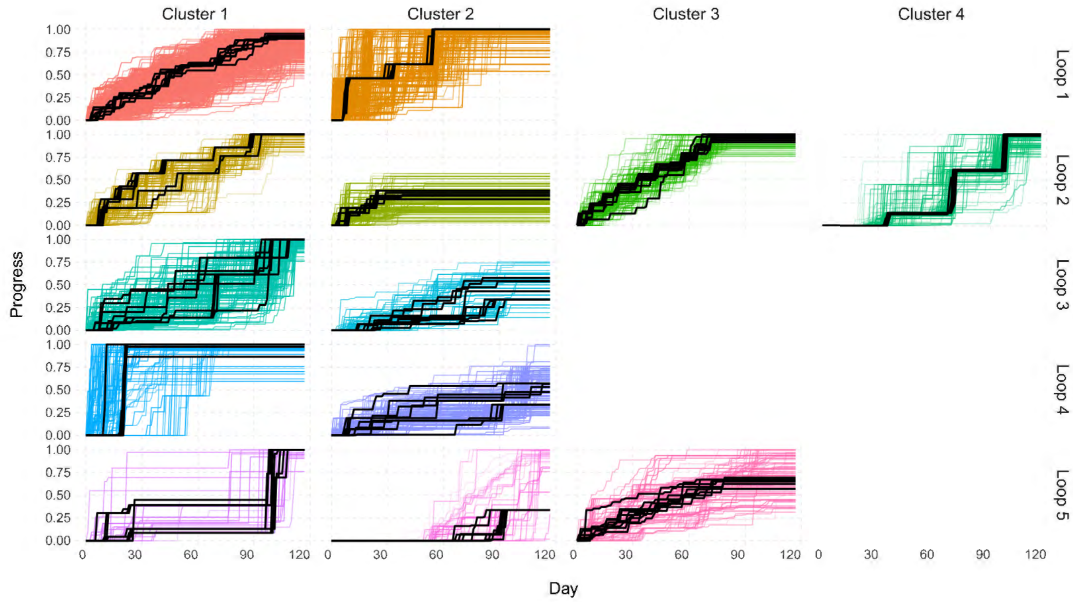


**Figure 4: The trajectories belonging to each cluster in the UCF data. The most representative (highest membership values) members of each cluster are shown in black.**

- C2L1 and C4L2 are similar in that they make their progress is a small number of large steps. The difference is when in the course that progress takes place.
- C1L5 are students who have a long dormant period in the middle of the course and leave everything to the last minute.

As expected the clusters found on the early loops tend to capture behaviors that are close to the "average," whereas later loops have clusters that are more different. We demonstrate this by examining the cluster centers displayed in Figure 3. The deviation of a cluster center from the mean (solid black line) is an indication of how far the behavior is from the average. Table 1 provides this deviation, calculated as the mean absolute difference, for each cluster and loop. We see that in general later clusters capture more extreme behaviors. The cluster closest to the average is C1L2, but only represents about 5.5% of the students. The cluster furthest from the average is C2L5 and represents about 1.5% of students. What makes these students stand out is their late start time and low level of progress.

## 3.4 Comparison

To highlight the limitations of standard approaches, we applied both crisp and fuzzy k-means to the UCF dataset. In summary, these algorithms produce a much smaller number of clusters and do not capture the same range of outlier behavior as those captured by the RFC algorithm. Table 2 display the results from fuzzy k-means for various values of the fuzziness parameter $m$. For each value of $m$, the table provides the validity indices, the selected number of cluster $k$, and the number of outliers based on (2). The value of $m = 2$ is the default and corresponds to applying just one loop of the RFC without refinement. We observe that we get more clusters and fewer outliers as $m \to 1$. Indeed the validity indices suggest that $m = 1.01$ is the best solution

**Table 2: Results of fuzzy k-means for various values of $m$ including the validity indices and number of outliers.**

| m | k | SIL.F | SIL | PC | PE | MPC | XB | Out. |
|---|---|---|---|---|---|---|---|---|
| 1.01 | 5 | .58 | .58 | 1.0 | .00 | 1.0 | .29 | 1 |
| 1.2 | 4 | .59 | .56 | .94 | .11 | .92 | .36 | 66 |
| 1.4 | 3 | .58 | .51 | .83 | .31 | .75 | .46 | 366 |
| 1.6 | 3 | .60 | .50 | .73 | .50 | .59 | .49 | 864 |
| 1.8 | 3 | .62 | .49 | .63 | .66 | .44 | .52 | 1438 |
| 2.0 | 2 | .52 | .44 | .69 | .48 | .37 | .54 | 1606 |

of those presented. However, with this solution, we only get five clusters, and these contain high levels of noise and are therefore can be challenging for instructors to interpret. This low number of clusters does not accurately capture the full range of behaviors apparent in the data.

With the solution improving as $m \to 1$ the logical step to take is to set $m = 1$ and perform simple crisp clustering using the k-means algorithm. We performed this using the NBClust R package [7] which provides a collection of 23 appropriate validity indices to help with the choice of $k$. Of these, 7 proposed 3 clusters, followed by 5 indices proposing 7 clusters. Both values lead to the same conclusion as we arrived at with fuzzy clustering; that is, the number of clusters does not adequately capture the full range of behaviors.

# 4. CONCLUSIONS AND FUTURE WORK

The RFC algorithm has allowed us to uncover outlier behaviors that are in some cases very different to the most common behaviors found on loop 1, and in other cases appear visually similar but represent a very different type of learning behavior. The behavioral clusters found here are by no means an exhaustive list. Adjusting the parameters of the algorithm, for example, by changing the parameters that control the fuzziness of the clustering, would possibly allow more outlier behaviors to emerge.

The purpose of this paper was to describe and demonstrate the RFC algorithm in its current form. Many possible improvements and extensions could be carried out. Once the RFC algorithm has finished, it is possible that clusters on a later loop could better capture a student that belongs to some clusters on an earlier loop. One extension could be to carry out a refinement process moving students from earlier to later clusters. Potentially we can achieve further improvements by including additional features that capture other aspects of behaviors or by applying a weighting to features that are considered more critical.

Lakoff [20] puts the clustering process this way, "Categorization is not to be taken lightly. There is nothing more basic than categorization to our thought, perception, action and speech" ([20] pg. 5). Identifying these student trajectories as either subordinate, superordinate of basic level create a substantial educational responsibility in the adaptive learning environment where students have control time, pace and feedback. If John Carroll [6] was correct in that learning is a function of time spent and time needed then the question is what resources do various student cohorts require. We argue that the clustering process can help in the better understanding of what it will take to help larger numbers of students become successful. As we explore these procedures several questions emerge. If and when will the process become excessively granular and dysfunctional how can these processes be integrated into the educational environment? Can these methods contribute to resolving achievement inequality? Finally, the question remains about whether the clusters exhibit a categorical structure with meaningful prototypes that respond to instructional interventions.

# 5. REFERENCES


[1] Y. Abraham. *Radviz: Project Multidimensional Data in 2D Space*, 2016. R package version 0.7.0.

[2] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. nigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[3] J. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3:58–73, 1974.

[4] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[5] R. Campello and E. Hruschka. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 157:2858–2875, 2006.

[6] J. B. Carroll. A model of school learning. *Teachers College Record*, 64:723–733, 1963.

[7] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36, 2014.

[8] R. Dave. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters*, 17:613–623, 1996.

[9] L. Di Caro, V. Frias-Martinez, and E. Frias-Martinez. Analyzing the role of dimension arrangement for data visualization in radviz. In *Advances in Knowledge Discovery and Data Mining*, pages 125–132. Springer, 2010.

[10] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[11] C. Dziuban, C. Howlin, J. Constance, and P. Moskal. An adaptive learning partnership. *Educause Review*, Dec. 2017.

[12] M. Ferraro and P. Giordani. A toolbox for fuzzy clustering using the r programming language. *Fuzzy Sets and Systems*, 279:1–16, 2015.

[13] G. Gan and M. K.-P. Ng. k-means clustering with outlier removal. *Pattern Recognition Letters*, 90:8–14, 2017.

[14] A. Gosain and S. Dahiya. Performance analysis of various fuzzy clustering algorithms: A review. *Procedia Computer Science*, 79:100–111, 2016.

[15] V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen, and P. Fränti. Improving k-means by outlier removal. In H. Kalviainen, J. Parkkinen, and A. Kaarna, editors, *Image Analysis*, pages 978–987, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[16] K. He, Y. Li, S. Soundarajan, and J. E. Hopcroft. Hidden community detection in social networks. *CEUR Workshop Proceedings*, 1828:89–93, Feb. 2017.

[17] K. He, S. Soundarajan, X. Cao, J. Hopcroft, and M. Huang. Revealing multiple layers of hidden community structure in networks. *http://arxiv.org/abs/1501.05700v1*, 2015.

[18] P. Hoffman, G. G. Grinstein, and D. Pinkney. Dimensional anchors: A graphic primitive for multidimensional multivariate information visualizations. In *Workshop on New Paradigms in Information Visualization and Manipulation*, pages 9–16, 1999.

[19] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

[20] G. Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press, 1987.

[21] H. Liu, J. Li, Y. Wu, and Y. Fu. Clustering with outlier removal. *http://arxiv.org/abs/1801.01899*, 2018.

[22] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007.

[23] E. Rosch. Natural categories. *Cognitive Psychology*, 4:328–350, 1973.

[24] E. Rosch. Cognitive reference points. *Cognitive Psychology*, 7:532–557, 1975.

[25] A. Vellido, F. Castro, and A. Nebot. *Clustering Educational Data. In Handbook of Educational Data Mining*, pages 72–92. CRC Press, first edition, 2 2011.

[26] W. Wang and Y. Zhang. On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, 158(19):2095–2117, 2007.

[27] X. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:841–847, 1991.