Synthesis of clustering techniques in educational data mining

Mr. Doipayan Roy, Purdue University
Prof. Peter Bermel, Purdue University, West Lafayette (College of Engineering)

DR. PETER BERMEL is an assistant professor of Electrical and Computer Engineering at Purdue University. His research focuses on improving the performance of photovoltaic, thermophotovoltaic, and nonlinear systems using the principles of nanophotonics. Key enabling techniques for his work include electromagnetic and electronic theory, modeling, simulation, fabrication, and characterization.

Dr. Bermel is widely-published in both scientific peer-reviewed journals and publications geared towards the general public. His work, which has been cited over 4400 times, for an h-index value of 24, includes the following topics: * Understanding and optimizing the detailed mechanisms of light trapping in thin-film photovoltaics * Fabricating and characterizing 3D inverse opal photonic crystals made from silicon for photovoltaics, and comparing to theoretical predictions * Explaining key physical effects influencing selective thermal emitters in order to achieve high performance thermophotovoltaic systems

Dr. Kerrie Anna Douglas, Purdue University, West Lafayette (College of Engineering)

Dr. Douglas is an Assistant Professor in the Purdue School of Engineering Education. Her research is focused on methods of assessment and evaluation unique to engineering learning contexts.

Prof. Heidi A. Diefes-Dux, Purdue University, West Lafayette (College of Engineering)

Heidi A. Diefes-Dux is a Professor in the School of Engineering Education at Purdue University. She received her B.S. and M.S. in Food Science from Cornell University and her Ph.D. in Food Process Engineering from the Department of Agricultural and Biological Engineering at Purdue University. She is a member of Purdue’s Teaching Academy. Since 1999, she has been a faculty member within the First-Year Engineering Program, teaching and guiding the design of one of the required first-year engineering courses that engages students in open-ended problem solving and design. Her research focuses on the development, implementation, and assessment of modeling and design activities with authentic engineering contexts.

Mr. Michael Richey, The Boeing Company

Michael Richey is an Associate Technical Fellow currently assigned to support workforce development and engineering education research. Michael is responsible for leading learning science research, which focuses on learning ecologies, complex adaptive social systems and learning curves. Michael pursues this research agenda with the goal of understanding the interplay between innovation, knowledge transfer and economies of scale as they are manifested in questions of growth, evolvability, adaptability and sustainability.

Additional responsibilities include providing business leadership for engineering technical and professional educational programs. This includes topics in advanced aircraft construction, composites structures and product lifecycle management. Michael is responsible for leading cross-organizational teams from academic, government focusing on how engineering education must acknowledge and incorporate this new information and knowledge to build new methodologies and paradigms that engage these developments in practice. The objective of this research is focused on achieving continuous improvement and sustainable excellence in engineering education.

Dr. Krishna Madhavan, Purdue University, West Lafayette (College of Engineering)

Dr. Krishna Madhavan is an Associate Professor in the School of Engineering Education. In 2008 he was awarded an NSF CAREER award for learner-centric, adaptive cyber-tools and cyber-environments using learning analytics. He leads a major NSF-funded project called Deep Insights Anytime, Anywhere (http://www.dia2.org) to characterize the impact of NSF and other federal investments in the area of STEM education. He also serves as co-PI for the Network for Computational Nanotechnology (nanoHUB.org)
that serves hundreds of thousands of researchers and learners worldwide. Dr. Madhavanserved as a Visiting Researcher at Microsoft Research (Redmond) focusing on big data analytics using large-scale cloud environments and search engines. His work on big data and learning analytics is also supported by industry partners such as The Boeing Company. He interacts regularly with many startups and large industrial partners on big data and visual analytics problems.

Siddharth Shah
Synthesis of clustering techniques in educational data mining

Introduction:

With the increasing demand for high quality education, coupled with the geographic and logistical limitations of the traditional in-class education system, educational institutions are turning to alternative forms of knowledge dissemination through online learning environments - such as Massive Open Online Courses (MOOCs). These learning environments are producing a tremendous amount of data that can provide deep insights into learning processes and learner behaviors. The large amounts of generated data require careful processing to convert them into actionable insights. The process of educational data mining (EDM) is concerned with developing methods for exploring data coming from educational settings and using those methods to better understand students and the settings in which they learn [37]. Online educational platforms are different from classroom settings in that they allow students to register without regard for geographic location, financial or academic status. Online platforms also allow students to participate in and drop out of courses whenever they want with very little consequences. Learner motivation and behavior on such platforms is more diverse and very different from those in a traditional educational setting. Investigating types of learners and their behavioral traits is very important for devising effective pedagogical strategies for online learning [25].

In this paper, we investigate the use of clustering techniques to identify learner types in MOOCs. We describe one popular clustering technique - the K-means algorithm and discuss how it may be used for learner categorization in an online learning environment. We present a method for applying the K-means algorithm for learner type identification within the more constrained context of a highly technical and advanced MOOC on nanotechnology. We investigate different types of learner behavior that emerge from the above-mentioned clustering and the ways in which each group of learners is distinct. Finally, we assign labels to each user group per their dominant behavioral characteristics and use hypothesis testing to show that the difference in learner behavior across groups is statistically significant.

Literature Review:

Learning platforms such as MOOCs provide the means for knowledge dissemination without regard to geographic, social and financial barriers [1] and hold the potential to better support learning and create opportunities for novel forms of assessment [5]. MOOCs tend to have enrollment patterns that are different from traditional higher education learning environments [2]. The reasons for enrolling in MOOCs vary by user; some users take courses to enhance or as a part of academic programs; others enroll for professional development; others take courses purely out of curiosity [14]. Analysis of learner behavior in MOOCs is therefore different from analysis of traditional forms of knowledge dissemination due to the diversity of participants with respect to motivation [3], course usage patterns and completion rates [4] and other similar factors. As stated by Douglas, et al. [20], MOOCs offer open access and free registration to everyone, allowing learners to enter and leave the MOOC at will.

Understanding learners is essential for good instructional design [25]. There are many ways of characterizing learner behavior in the context of EDM. In this paper, we define behavior as the
pattern in which a learner accesses resources available on the learning platform. Douglas, et al. [20] and Hicks, et al. [21] characterized learners based on usage patterns using K-means clustering. Hicks, et al. [21] focused almost entirely on Fully Engaged Learners and integrated survey data collected before the start of the course to compare the motivation and behavior of this group with other groups. In contrast, Douglas, et al. [20] investigated the behavior of all learner groups in detail and used the Kruskal-Wallis test to show that the differences in performance (as measured by assessment scores) and participation across clusters was statistically significant. At the heart of both papers was the use of K-means clustering algorithm to group the learner population into clusters based on similarity in usage patterns.

The K-means algorithm is a very popular partition-based clustering algorithm in data mining [8]. Li, et al. [27] observed that manual construction of student models is often time consuming and proposed an innovative method for automated identification of student models using K-means clustering along with other popular machine learning techniques like Principal Component Analysis. Trivedi, et al. [28] pointed out some theoretical limitations of the K-means algorithm and demonstrated the use of spectral clustering to obtain improved prediction of student performance. This paper addressed one of the most severe limitations of the K-means algorithm – namely, that the algorithm can converge at some shallow local minimum and result in clustering that is not optimal. It further demonstrated that spectral clustering can be used to obtain the global minimum, resulting in an improvement of prediction of student performance. This is a classic example of clustering as a preprocessing tool before supervised learning techniques are applied for extracting patterns from data.

Apart from partition-based clustering techniques (like K-means), hierarchical clustering and density-based clustering are two other approaches in data mining literature. DeFrietas, et al. [29] presented a comparative study of these three clustering paradigms- partition-based (K-means), density-based (DBSCAN) and hierarchy-based (BIRCH) in the context of Learning Management Systems (LMS). Santhisree, et al. [33] demonstrated the use of DBSCAN clustering for identifying user behavior from sequential access of web data. Golding and Donaldson [34] used hierarchical clustering on questionnaire data to map out approaches to teaching profiles of teachers in higher education on the basis of their scores on the Approaches to Teaching Inventory (ATI) [26]. Lee, et al. [32] described a visualization technique for understanding the ways in which learning outcomes are related to learner access patterns by combining hierarchical cluster analysis with heat maps.

An alternative to the hard clustering techniques discussed so far (in which cluster assignments are unique - every observation belongs to a single cluster) is fuzzy means clustering (or soft clustering). Pechenizkiy, et al. [35] used C-means clustering (a fuzzy clustering technique) to identify significant variables that affect the performance of undergraduate students. Inyang and Enobong [38] used fuzzy clustering for predicting student performance and remarked upon the fact that students that performed poorly found it difficult to advance their careers and find employment.
It can be said that a significant amount of research has been done using clustering as the sole tool for pattern analysis, and there is an equally vast body of research involving clustering alongside other popular machine learning techniques. In this work, the K-means algorithm was used for categorization of MOOC learners based on their access patterns.

K-means Clustering:

The K-means algorithm takes a set of observations and number of clusters (K) as input and attempts to partition the input space into K cells by minimizing a cost function that penalizes assignment of every observation to a specific cluster. All observations that lie within a particular cell collectively form a cluster. The K-means algorithm is a centroid-based clustering technique, meaning that the centroid of observations belonging to a cluster is considered representative of the entire cluster [11].

The assignment of observations to clusters is based on similarity. The goal is to group observations that are similar (as measured by some pre-defined quantity) into a single cluster. This requires us to define a method for quantifying ‘similarity’ between observations. The performance of a clustering algorithm depends heavily on the similarity metric being used [6]. One common implementation of the K-means algorithm uses Euclidean distance between observations as the measure for similarity [22]. If the observations belong to a multidimensional vector space, the L₂ norm of the difference between two observations is used for quantifying similarity. The closer two observations are in the L₂ norm sense, the more similar they are.

At the heart of the K-means algorithm is a penalty (or cost) function that penalizes the assignment of every observation to a cluster [11,13]. The more similar an observation is to the members of the cluster, the smaller the penalty is of assigning the observation to that particular cluster. The penalty of assigning an observation to a cluster is taken as the Euclidean distance of the observation from the cluster centroid. Assigning an observation to a far-off cluster incurs a large penalty (and vice-versa). Minimizing such a cost function ensures that similar observations are grouped in the same cluster. Mathematically, an observation \(x_i\) is assigned to cluster \(k\) in the following way:

\[ k = \arg\min_j \|x_i - \mu_j\|^2; \quad j = 1, 2, 3 \ldots K \]

\(\mu_j\) is the centroid of the \(j^{th}\) cluster. Assigning observations to clusters in this way minimizes the following cost function [11]:

\[ C = \sum_{j=1}^{K} n_j \sum_{i=1}^{n_j} \|x_{ij} - \mu_j\|^2 \]

Here, \(n_j\) is the number of observations in the \(j^{th}\) cluster and \(x_{ij}\) is the \(i^{th}\) observation of \(j^{th}\) cluster.

Minimizing this cost function is NP-hard in general, since there exists no known closed form solution [7,18]. Cluster assignment is done iteratively, with cluster centroids being initialized randomly at the first iteration [8]. Every observation is assigned to a unique cluster at each iteration and all cluster centroids are recalculated at the end of the iteration. This process is continued until a convergence criterion is satisfied. The iterative procedure with random
initializations guarantee convergence at some local minimum of the cost function [8,23]. There is no guarantee that the algorithm will converge at the best possible choice of clusters (i.e. the global minimum of the cost function). This is a common problem encountered in optimization problems involving cost functions with multiple local minima. This is one of the major theoretical limitations of the K-means algorithm. This problem can be partially overcome by running the algorithm multiple times with random initializations and taking the cluster assignment that has the minimum total penalty as measured by the cost function. All runs of the algorithm converge at a local minimum, running it several times with random initializations is done to find a good (deep) local minimum.

Graphical clustering methods such as spectral clustering can be used to obtain the global minimum of the cost function, but those methods come with limitations of their own [39]. Spectral clustering uses a similarity matrix in which the \((i,j)\) element of the matrix quantifies the similarity (as measured by a metric, such as the Euclidean distance) between the \(i^{th}\) and \(j^{th}\) data points. In the context of this work, the size of this matrix depends on the number of students registered for the MOOC. Since spectral clustering depends on factorization of the similarity matrix, the technique becomes unsuitable for situations where the size of the dataset is very large, which is typically the case in MOOCs where registration can be in the tens of thousands. Additionally, typical behavioral patterns in MOOCs give rise to access distributions in the form of globular clusters. K-means perform well when the aim is to detect globular clusters, but spectral clustering performs better when clusters form connected contours. In our context, where the datasets are typically very large and clusters tend to have globular shapes, K-means is the better approach, given it’s ease of implementation, simplicity, efficiency, and empirical success [31].

The K-means algorithm is as follows [10,17]:

1) Randomly initialize \(K\) cluster centroids: \(\mu_{0,1}, \mu_{0,2}, \ldots, \mu_{0,K}\). Set \(t=1\).
2) Assign observation \(x_i\) to cluster \(k\) such that the following condition is satisfied:

\[
  k = \text{argmin}_j \|x_i - \mu_{t-1,j}\|_2^2; \quad j = 1, 2, 3 \ldots K
\]

Use this rule to assign all samples \(x_1, x_2, \ldots, x_N\) to unique clusters.
3) Calculate cluster centroids as follows:

\[
  \mu_{t,j} = \frac{\sum_{i=1}^{n_{t,j}} x_{ij}}{n_{t,j}}
\]

\(n_{t,j}\) is the number of samples assigned to \(j^{th}\) cluster at iteration \(t\), \(x_{ij}\) is the \(i^{th}\) sample of cluster \(j\) and \(\mu_{t,j}\) is the centroid for cluster \(j\) at the end of iteration \(t\).
4) Calculate the following quantity:

\[
  \delta = \max \left\{ \|\mu_{t-1,j} - \mu_{t,j}\|_2^2 \right\} \quad \text{for} \quad j = 1, 2, \ldots, K
\]
5) If \(\delta < \text{threshold}\), stop. Otherwise, \(t=t+1\) and go to Step 2 (next iteration if convergence criteria is not satisfied).

Finally, any description of the K-means algorithm would be incomplete without describing the methodology for selecting the number of clusters given a particular set of observations. The number of clusters in a dataset is often ambiguous [16]. The best choice depends on the shape and distribution of samples as well as the resolution demanded by the user and/or the application. One of the ways of determining the optimal number of clusters is by considering the amount of
variance in the data that is explained by using a certain number of clusters. The model must explain as much of the observed variance as possible using a reasonably low number of clusters. Elbow plots [24] or gap statistics [9] are usually used for determining the number of clusters in a dataset. Elbow graphs plot the percentage of data variance explained by the cluster model versus the number of clusters. The percentage of variance explained by a cluster model is taken as the ratio of between-cluster variance to total variance. A typical elbow plot is illustrated in Figure 1.

The percentage of variance explained by the cluster model increases as the number of clusters is increased. This is expected— as increasing the number of clusters increases the modeling power and thus allows the model to better capture patterns in the data. But increasing the number of clusters beyond a certain point produces vanishing gains in modeling data variance. This point is characterized by a distinctive ‘elbow’ in the graph (in Figure 1, the elbow occurs at K=4). The number of clusters at which this elbow occurs is the taken as the optimal number of clusters (with minor user or theoretical context dictated adjustments). Taking a larger number of clusters does not cause any significant increase in the amount of variance explained by the model. Taking a lesser number of clusters is not optimal either, since that model will leave considerable unexplained variance.
Raw Data:

We collected course usage patterns for characterization of learners in two highly advanced MOOCs on nanotechnology - Thermoelectricity and Nanobiosensors. Both courses were offered by Purdue University on edX. Learner usage pattern were extracted from clickstream data. Clickstream data is defined by Madhavan and Richey [36] as a proxy for user behavior that not only records that a click was made by the user on an electronic object, but also captures the associated metadata, such as the system time when the click was made (for standardization), the object on which the click event occurred, the dwell time on the clicked entity, and a few other navigational details. Clickstreams not only embody behavior, they are also a manifestation of learner intent [36]. Two types of course materials were made available to learners every week: learning materials like lectures, homeworks and tutorials; and assessment materials like quizzes and exams. Analysis of access patterns is a common approach for understanding learner behavior and interaction. For example, Hecking, et al. [12] investigate characteristic patterns of resource access in two graduate level courses to gain deeper insights into the usage of learning materials

Clustering:

In this study, learners were characterized by the number of distinct course modules accessed from beginning to end of the course. Each learner was represented by the number of distinct course resources accessed during the course. The input space for K-means clustering was the set of positive integers ranging from 0 (representing learners who did not access any course resources) to number of course modules available in the course (representing learners who accessed all course materials).

Learners with similar behavior were those that accessed almost the same number of course modules. Conversely, dissimilar learners were those that were very different with regard to number of course materials accessed. This approach resulted in learners with similar number of courseware accesses being grouped in the same cluster.

Elbow plots were used to establish the optimal number of clusters. Figure 2 and Figure 3 are the elbow plots for the Thermoelectricity and Nanobiosensors MOOCs, respectively. It is clear that both curves saturate completely after K=4. Hence, the number of clusters was taken as 4 for both courses.
Figure 2- Elbow plot for Thermoelectricity MOOC

Figure 3- Elbow plot for Nanobiosensors MOOC
Cluster plots:

Figures 4 and 5 are the cluster plots for Thermoelectricity and Nanobiosensors MOOCs, respectively. Each learner cluster was given a different color.

![Figure 4- Cluster plot for Thermoelectricity MOOC](image)
Clusters were plotted with learners along Y-axis and course modules over time along X-axis. The course modules were arranged on the X-axis in order of how they appear in the course. As we move from left to right on the X-axis, we go from modules at the beginning of the course to modules in the middle weeks, and then to the final modules of the course. The coordinate \((i,j)\) of the cluster plot was marked with a circular marker if the \(j\)th learner on the Y-axis accessed the \(i\)th course module. Additionally, the size of the marker is directly related to how many times the particular learner accessed the module. If the number of times a course material was accessed is \(n\), the radius of the marker was taken as \(r = \frac{\log(n+1)}{\alpha}\), where \(\alpha\) is a constant determined by trial-and-error experimentation. \(\alpha\) is chosen such that the size of markers become reasonable after application of the scaling function, it’s value had to be determined separately for each course. The scaling function is monotonically increasing in \(n\) - a marker with greater radius implies that the particular learner has accessed the module more times. The use of this logarithmic scaling function made the plots more readable and prevented some markers from becoming excessively big. Finally, the clusters were arranged in decreasing order of learner activity - clusters comprising the heaviest users (as measured by number of distinct module accesses) appear on top and clusters comprising less active users appear at the bottom.
The cluster plots not only illustrate the usage patterns of learners, they also demonstrate how frequently learners in different clusters revisit course modules. A large number of big markers within a cluster is indicative of learners who frequently revisit modules after accessing it for the first time. This issue is discussed in detail in the next section. Additionally, these plots show that some modules are more popular than others - for example, the final module of both courses was heavily accessed by learners of all clusters. In contrast, there are a few modules towards the end of the second week in Thermoelectricity that were not accessed by a majority of learners in the course.

Results and Discussion:

1) Number of learner groups- The elbow plots for both courses indicate that the best choice for number of clusters is four. Identifying the optimal number of clusters can be difficult at times when the elbow in the plot is not distinct. Table 1 illustrates the percentage change in amount of variance explained by the clustering as the number of clusters is increased:

<table>
<thead>
<tr>
<th>Number of clusters</th>
<th>Thermoelectricity BCSS/TSS</th>
<th>Thermoelectricity Percentage increase from K=4</th>
<th>Nanobiosensors BCSS/TSS</th>
<th>Nanobiosensors Percentage increase from K=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.7994</td>
<td>-16.87%</td>
<td>0.8330</td>
<td>-12.65%</td>
</tr>
<tr>
<td>3</td>
<td>0.9098</td>
<td>-5.39%</td>
<td>0.9389</td>
<td>-1.55%</td>
</tr>
<tr>
<td>4</td>
<td>0.9617</td>
<td>0.00%</td>
<td>0.9537</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>0.9752</td>
<td>1.40%</td>
<td>0.9750</td>
<td>2.22%</td>
</tr>
<tr>
<td>6</td>
<td>0.9776</td>
<td>1.65%</td>
<td>0.9838</td>
<td>3.16%</td>
</tr>
<tr>
<td>7</td>
<td>0.9796</td>
<td>1.86%</td>
<td>0.9826</td>
<td>3.00%</td>
</tr>
</tbody>
</table>

Table 1- Percentage changes in explained variance versus number of clusters

There is a steep rise in both plots up to K=4, and both plots saturate after K=4. We conclude that using more than four clusters will not produce results substantially better than K=4, given the potential risk of overfitting at large values of K. Also, using less than four clusters is not optimal either: that will handicap the model to an extent that will leave a considerable portion of variance unexplained. Thus, K=4 is the optimal choice for both courses.

2) Identifying learner groups- We identify and label four categories of learners that emerge from access-based clustering using the K-means algorithm. These clusters are very similar to the ones observed by Douglas, et al. [20] and Hicks, et al. [21]. Thus, we use similar nomenclature for identifying the clusters:

(i) Fully Engaged Learners- This group of learners is the black cluster in Figure 4 and the green cluster in Figure 5. Members of this group accessed almost all study materials in the course and attempted most quizzes and exams. This group was active throughout the life of the course and accessed all modules regularly.
(ii) Consistent Viewers- This group of learners is the blue cluster in Figure 4 and red cluster in Figure 5. Like Fully Engaged Learners, this group was also active throughout the offering of the course. The difference lies in how members of this group accessed almost all study materials (video lectures, tutorials, et cetera) but did not access much of the assessment materials. The assessment materials come towards the end of each week. We found that members of this group had a lot of activity at the beginning of every week but almost no activity towards the end, where most of the modules were assessment-based and were intended to test the learner’s understanding of the week’s materials.

(iii) Sporadic Learners- This group of learners is the green cluster in Figure 4 and black cluster Figure 5. Members in this group did not have any well defined access patterns. Some learners in this cluster accessed resources regularly at the beginning of the course, but their activity became inconsistent in the later weeks. Some of them had activity at the beginning and end of the course but very little activity in between. Some of these learners accessed both study materials and assessment-based resources, while others accessed only study materials and almost no assessment-based modules. We identified some Sporadic Learners in both courses who attempted all week exams, but did not access a single study resource in the course.

(iv) One Week Engaged Learners- This group of learners is the red cluster in Figure 4 and blue cluster in Figure 5. These learners fully access the first week’s materials, but did very little in the course afterwards. The cluster-wise distribution of students is given in Table 2.

<table>
<thead>
<tr>
<th>Group</th>
<th>Thermoelectricity Percentage of total learners</th>
<th>Nanobiosensors Percentage of total learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Engaged Learners</td>
<td>26.75%</td>
<td>13.09%</td>
</tr>
<tr>
<td>Consistent Viewers</td>
<td>25.47%</td>
<td>20.23%</td>
</tr>
<tr>
<td>Sporadic Learners</td>
<td>29.93%</td>
<td>26.19%</td>
</tr>
<tr>
<td>One-Week Engaged Learners</td>
<td>17.83%</td>
<td>40.47%</td>
</tr>
</tbody>
</table>

Table 2- Distribution of learners

3) Revisiting a course module- We also studied the number of times learners revisited modules after the first access. Clickstream data records the first time an object is accessed as well as all future access of the object by a particular learner. We investigated how many times learners in each group revisited course modules. The results were consistent for both courses. Table 3 shows the statistics for average revisits per learner per module across clusters.

<table>
<thead>
<tr>
<th>Group</th>
<th>Thermoelectricity Average revisits (per learner per module)</th>
<th>Nanobiosensors Average revisits (per learner per module)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Engaged Learners</td>
<td>0.6958</td>
<td>0.5357</td>
</tr>
<tr>
<td>Consistent Viewers</td>
<td>1.8757</td>
<td>1.2796</td>
</tr>
<tr>
<td>Sporadic Learners</td>
<td>2.7820</td>
<td>2.3787</td>
</tr>
<tr>
<td>One-Week Engaged Learners</td>
<td>1.3984</td>
<td>0.7320</td>
</tr>
</tbody>
</table>

Table 3- Average revisits across clusters

We conducted hypothesis tests (two sample t-tests [15, 19]) to determine if the observed differences in average revisit rates are statistically significant. Table 4 shows the cluster
parameters for performing two sample t-tests and Table 5 illustrates the results of hypothesis testing on average revisit rates.

<table>
<thead>
<tr>
<th>Group</th>
<th>Thermoelectricity</th>
<th>Thermoelectricity</th>
<th>Nanobiosensors</th>
<th>Nanobiosensors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of samples</td>
<td>Standard deviation</td>
<td>Number of samples</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>Full Engaged</td>
<td>41</td>
<td>0.3916</td>
<td>11</td>
<td>0.2080</td>
</tr>
<tr>
<td>Consistent Viewers</td>
<td>40</td>
<td>0.4267</td>
<td>17</td>
<td>0.2204</td>
</tr>
<tr>
<td>Sporadic Learners</td>
<td>48</td>
<td>0.3777</td>
<td>22</td>
<td>0.2267</td>
</tr>
<tr>
<td>One-Week Engaged</td>
<td>28</td>
<td>0.3862</td>
<td>34</td>
<td>0.2125</td>
</tr>
</tbody>
</table>

Standard deviation in this table refers to the standard deviation of average revisits of learners in each cluster.

Table 4- Parameters for two sample t-tests

<table>
<thead>
<tr>
<th>Groups compared</th>
<th>Hypothesis</th>
<th>Thermoelectricity p-value</th>
<th>Thermoelectricity Conclusion</th>
<th>Nanobiosensors p-value</th>
<th>Nanobiosensors Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Engaged vs. Consistent Viewers</td>
<td>$\mu_1 &lt; \mu_2$</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
</tr>
<tr>
<td>Fully Engaged vs. Sporadic</td>
<td>$\mu_1 &lt; \mu_3$</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
</tr>
<tr>
<td>Fully Engaged vs. One-Week Engaged</td>
<td>$\mu_1 &lt; \mu_4$</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
<td>0.05</td>
<td>Difference is significant</td>
</tr>
<tr>
<td>Consistent Viewers vs Sporadic</td>
<td>$\mu_2 &lt; \mu_3$</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
</tr>
<tr>
<td>Consistent Viewers vs. One-Week Engaged</td>
<td>$\mu_2 &gt; \mu_4$</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
</tr>
<tr>
<td>Sporadic vs. One-Week Engaged</td>
<td>$\mu_3 &gt; \mu_4$</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
<td>Less than 0.05</td>
<td>Difference is significant</td>
</tr>
</tbody>
</table>

$\mu_1, \mu_2, \mu_3, \mu_4$ are the average revisits per learner per module for Fully Engaged, Consistent, Sporadic and One-Week learner groups (respectively).

Table 5- Result of hypothesis testing across all pairs of clusters

Taking the standard threshold of $p=0.05$, we reject the null hypothesis in all tests. We arrive at two conclusions from Table 5. First, the ranking of learner groups in both courses on the basis of average revisit rates is significant from a statistical standpoint. Second, the ranking is identical in both courses. Sporadic learners had the highest average revisit rates in both courses, followed by Consistent Viewers, followed by One-Week Engaged learners, followed by Fully Engaged learners.

4) Fully Engaged learners vs. Consistent Viewers- Members of both these groups stayed active throughout the course. The difference lies in how Fully Engaged learners accessed both study materials and assessment resources throughout the course whereas Consistent Viewers accessed only study materials. We used two sample t-tests to show that this difference in behavior is statistically significant. Hypothesis testing is crucial in this context because if the difference in access patterns is not sufficiently significant, one may argue that these two groups should be
merged into a single cluster. In the absence of the statistical tests that we present, it would be difficult to counter this argument. Except for how these two groups accessed (or did not access) assessment resources, their usage patterns were identical.

<table>
<thead>
<tr>
<th>Group</th>
<th>Thermoelectricity Average visits to study resources</th>
<th>Thermoelectricity Average visits to assessments</th>
<th>Nanobiosensors Average visits to study resources</th>
<th>Nanobiosensors Average visits to assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Engaged</td>
<td>0.7020 ($\mu_{S,FE,T}$)</td>
<td>0.6834 ($\mu_{A,FE,T}$)</td>
<td>0.5891 ($\mu_{S,FE,N}$)</td>
<td>0.4993 ($\mu_{A,FE,N}$)</td>
</tr>
<tr>
<td>Consistent Viewers</td>
<td>2.340 ($\mu_{S,CV,T}$)</td>
<td>0.358 ($\mu_{A,CV,T}$)</td>
<td>1.451 ($\mu_{S,CV,N}$)</td>
<td>0.305 ($\mu_{A,CV,N}$)</td>
</tr>
</tbody>
</table>

Table 6- Average visits to study materials (S) and assessment resources (A)

The hypothesis tests comparing Fully Engaged (FE) and Consistent Viewers (CV) groups’ average visits to assessments (A) were as follows:

Test 1:
Null hypothesis: $\mu_{A,CV,T} = \mu_{A,FE,T}$
Alternate hypothesis: $\mu_{A,CV,T} < \mu_{A,FE,T}$

Test 2:
Null hypothesis: $\mu_{A,CV,N} = \mu_{A,FE,N}$
Alternate hypothesis: $\mu_{A,CV,N} < \mu_{A,FE,N}$

These tests yield p-values of 0.001 for Thermoelectricity and 0.036 for Nanobiosensors. Taking the standard threshold of p=0.05, we reject the null hypothesis in both tests and conclude that Consistent Viewers group access assessments less frequently as compared to Fully Engaged learners in both courses. The fact that we can accept both alternate hypotheses substantiates the claim that Consistent Viewers and Fully Engaged learners differ in how they accessed assessment materials and that Consistent Viewers did not access many assessment materials as compared to Fully Engaged learners.

The hypothesis tests comparing Consistent Viewers’ (CV) average visits to assessments (A) versus study materials (S) were as follows:

Test 3:
Null hypothesis: $\mu_{A,CV,T} = \mu_{S,CV,T}$
Alternate hypothesis: $\mu_{A,CV,T} < \mu_{S,CV,T}$

Test 4:
Null hypothesis: $\mu_{A,CV,N} = \mu_{S,CV,N}$
Alternate hypothesis: $\mu_{A,CV,N} < \mu_{S,CV,N}$

Both these tests yield p-values of the order $10^{-4}$. Again, using the standard threshold of p=0.05, we reject the null hypothesis in both tests and conclude that Consistent Viewers had a tendency to access study resources more than assessment resources. The fact that average visits to study
resources is significantly higher confirms our claim that Consistent Viewers had a tendency to exclusively use study resources and ignore assessment materials.

The hypothesis tests comparing Fully Engaged (FE) learners’ average visits to assessments (A) versus study materials (S) were as follows:

Test 5:
Null hypothesis: $\mu_{A,FE,T} = \mu_{S,FE,T}$
Alternate hypothesis: $\mu_{A,FE,T} < \mu_{S,FE,T}$

Test 6:
Null hypothesis: $\mu_{A,FE,N} = \mu_{S,FE,N}$
Alternate hypothesis: $\mu_{A,FE,N} < \mu_{S,FE,N}$

Tests 5 and 6 yielded p-values of 0.83 and 0.323, respectively. Using $p=0.05$ as the threshold, we cannot reject the null hypothesis in either case. This shows that the difference in access of assessment and study materials by Fully Engaged learners is not statistically significant. Unlike the Consistent Viewers group, they did not prefer the study materials over assessments.

To sum up, we found that Consistent Viewers accessed study materials more than assessment materials. This is not the case for Fully Engaged learners; this group did not seem to have any preference for study materials over assessment. Additionally, we found that Fully Engaged learners access assessment materials more than Consistent Viewers. Thus, we can safely claim that the two groups are different when it comes to usage of course materials.

Conclusion:

In this work, we presented a method for applying the K-means clustering algorithm to obtain learner groups in highly technical and advanced engineering MOOCs. We did a detailed study of the access and revisit patterns of learners across groups and found that the difference in behavior is statistically significant. We observed that the cluster plots provide a good visual illustration of access and revisit patterns of learner groups. Identification of learner groups and the dominant behavioral traits of each group is the first step to understanding how learners interact with online course materials in a MOOC. Using learner behavior for better instructional design of MOOCs is an open research topic, our research group is actively exploring ways by which insights into learner behavior can be used for optimizing design of online courses.

Future Work:

We cannot claim that this is an exhaustive study of all aspects of learner behavior in online learning environments. First, the courses in this study are not necessarily representative of all MOOCs and online courses in general. Both courses in this study were highly advanced engineering MOOCs—taking this type of courses require a lot of domain knowledge in nanotechnology and semiconductor physics. Thermoelectricity and Nanobiosensors are graduate level courses and are intended for advanced engineering students and industry professionals working in nanotechnology and related fields. In short, these courses are meant for people who
already have technical knowledge at least equivalent to a bachelor’s degree in a STEM field. This research work can be extended by considering courses in soft sciences, humanities and in general, courses that do not require previous domain knowledge.

Second, we did not consider assessment scores and final grades in this study. It will be interesting to see if grades depend on consistency of the learner in accessing course resources. Common wisdom dictates that highly consistent learners are expected to have better grades (in general) as compared to less consistent learners. This study can be extended to address the question of whether (and to what extent) performance of a learner (as measured by assessment scores and/or course grade) is dependent on consistency in accessing course resources.

Lastly, we did not consider learner demographics in this work. This study can be extended by segregating learners based on geographic location and carrying out analysis of learner behavior and performance for each demographic group.

References:


