Clustering Students Based on Their Prior Knowledge

Nisrine Ait Khayi
Institute for Intelligent Systems
The University of Memphis
Memphis, TN, 38152, USA
+19012473589
ntkhynyn@memphis.edu

Vasile Rus
Institute for Intelligent Systems
The University of Memphis
Memphis, TN, 38152, USA
+19016785259
vrus@memphis.edu

ABSTRACT
In this paper, we applied a number of clustering algorithms on pre-test data collected from 264 high-school students. Students took the pre-test at the beginning of a 5-week experiment in which they interacted with an intelligent tutoring system. The primary goal of this work is to identify clusters of students exhibiting similar knowledge patterns. In particular, we show that the DP-means clustering algorithm yields very good results using binary response data. Other clustering algorithms such as k-modes have demonstrated better results when using categorical response data.

Keywords
Clustering, intelligent tutoring systems, students modelling.

1. INTRODUCTION
Assessment is a key element in education in general and in Intelligent Tutoring Systems (ITSs) in particular because fully adaptive tutoring presupposes accurate assessment [2, 12].

Capturing students' knowledge state, our focus, and other learner characteristics that are important for learning such as their emotional state is critical to facilitate learning through adaptivity, i.e., tailoring instruction to each individual learner [11]. It should be noted that adaptivity can be thought of at two levels: macro-adaptivity which means selecting appropriate instructional tasks and micro-adaptivity which implies offering appropriate scaffolding while students work on a task (also called within-task adaptivity). Our work presented here could inform both micro- and macro-adaptivity. For instance, understanding the knowledge gaps of students in a particular cluster could inform what instructional tasks to choose for these students, i.e., it informs macro-adaptivity.

Indeed, an important preliminary step in creating an ITS that is sensitive to student misconceptions and individual learning trajectories is to first understand the various levels of mastery with respect to a target domain, for instance, physics. For example, important questions that need to be answers are: What are the predominant misconceptions they hold? and Are these misconceptions evenly distributed across topics and level of course taught? Using the clustering method proposed here will help answer such important questions. To this end, in this paper, we document for each group of students identified by our clustering algorithm, the major misconceptions exhibited by that group.

In this study, we applied clustering on a pretest data collected at the beginning of an experiment in which high-school students interacted with a dialogue-based ITS. Our goal was to identify student groups and analyze them as a group in terms of misconceptions and mastered concepts. The identified groups could then be used to inform the authoring of instructional tasks and within-task instructional strategies and feedback for each group as opposed to each learner, which would be a much more expensive process. Learning such individualized strategies for each learner would be possible using automated methods, such as reinforcement learning, but they require substantially more experimental data which it is not have available.

The main clustering algorithm used in this study is the DP-means algorithm [4]. Its main advantage is identifying the number of clusters using a Dirichlet Process Mixture Model. After briefly presented related work and the context of our own work, what follows is a description of the DP-means algorithm. We then present details of the experiments and results. We conducted experiments using two types of data: binary and categorical responses. In addition, other clustering algorithms were employed to compare the results with those obtained with DP-means. We evaluated the performance of the resulted clusters using intrinsic, e.g., based on the silhouette index which measures the compactness of each cluster, and extrinsic methods, e.g., based on students' post-test scores derived from post-test responses which were not used to generate the clusters.

2. RELATED WORK
Clustering has been used in the past for analyzing education data as indicated by the research studies presented next. Bouchet and colleagues [1] have applied the Expectation-Maximization clustering algorithm on data collected from the MetaTutor ITS. MetaTutor scaffolds student's metacognitive skills while learning about the human circulatory system. The main objective of their clustering was reinforcing self-regulated learning via student profiling. The results consisted of three distinct clusters of students in terms of performance. The results have been analyzed using a MANOVA approach.

Rodrigo and colleagues [8] have applied k-means clustering on data collected from students interacting with Aplusix, an ITS for Algebra. The main research goal of their work was identifying students' behaviors through an analysis of interaction logs. The results have demonstrated the existence of two clusters of students

Nisrine Ait Khayi and Vasile Rus "Clustering Students Based on Their Prior Knowledge" 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. 246 - 251

associated with differing behavior and affective states. The first cluster reflected more collaborative work whereas the second cluster reflected more solitary work.

Reyes-Gonzalez and colleagues [7] have used the LC-Conceptual clustering algorithm from logical combinatorial pattern recognition for student modeling in an ITS. This algorithm is based on two phases: the first phase consists of building groups of objects based on their similarity and a grouping criterion. The second phase is called the intentional structure phase where the distinctive features of each resulted cluster are determined. Fang and colleagues [3] have used k-means clustering to capture learning patterns in over 250 students who used AutoTutor to gain reading comprehension skills. The average response times per question and performance across lessons have been used to cluster the students' learning behavior. The results showed the convergence of four types of learners: proficient readers, struggling readers, conscientious readers and disengaged readers. Classifying readers can improve the adaptivity of AutoTutor ITS by providing a proactive feedback and intervention based on the learning behaviors.

Similar to those other approaches, our intention was to discover groups of students with similar knowledge states as characterized by their responses to the multiple-choice pre-test. Each incorrect choice in the pre-test is associated with a major misconception and therefore students that pick similar choices should be assigned to the same cluster. The centroid of the cluster could then be used to interpret the strengths and weaknesses of students in that cluster and appropriate interventions designed for that group.

3. CONTEXT OF THIS WORK

Our work was conducted in the context of an experiment in which high-school students interacted with a dialogue-based intelligent tutoring systems that tutors students on science topics through problem-solving. The system encourages students to self-explain solutions to complex science problems and only offers help, in the form of hints, when needed, e.g., when the student is floundering. That is, during a typical tutorial session, the system challenges students to solve a number of problems that are carefully selected by the system in order to optimize student learning (macro-adaptivity). When working on a particular problem, students are first asked to provide a solution that must include a justification based on concepts and principles of the target domain, which was Newtonian Physics in the case of our study presented here. All other things equal, low knowledge students will most likely struggle to provide solid self-explanations and most likely to articulate misconceptions which would lead to more scaffolding dialogue moves in terms of hints and correcting misconceptions on the part of the computer tutor (micro-adaptation). High knowledge students would need less scaffolding and therefore the corresponding dialogues should be shorter.

Before students start interacting with the system, they took a pre-test to assess their initial knowledge state. The tool elected to assess students' initial knowledge state was an enhanced version of the Force Concept Inventory (FCI). The Force Concept Inventory (FCI) is a 30-item multiple-choice "test" designed to assess student understanding of the most basic concepts in Newtonian mechanics (Halloun, Hake, and Mosca, 1995). The FCI presents students with various situations and ask them to choose between Newtonian explanations for the phenomena, versus common-sense alternatives (Hestenes, Wells, & Swackhamer, 1992). The FCI has been widely used to measure learning in introductory physics courses. For example, Hake (1998) reported FCI data from 6,000 high school and university students. Coletta and Phillips (2005) combined their data with data collected by Hake (1998) and in combination used the FCI to measure learning in 73 university and college introductory physics classes. The data we have is based on an augmented version of the FCI consisting of 35 multiple-choice questions. The augmented FCI adds a number of questions for certain Newtonian topics which were not covered enough in the original FCI test.

We administered the augmented Force Concept Inventory (aFCI) to students at three public and two private high schools in the mid-south region, including six teachers and 26 classrooms. The pretest was administered in classroom. Students completed the aFCI via provided scantron sheets, which were then collated and processed. The results of the scantron sheets were then compared to direct markings on the actual aFCI test in the case of blank or unidentifiable scantron responses. The data collection process was quite successful, resulting in 444 students with complete pretest data. We only used a subset of 265 students in our experiments because post-test data, used for extrinsic evaluation of our clustering, was available only for those 265 (the rest of the students either missed a tutoring session, or the post-test, or both).

It should be noted that the data is very diverse in terms of student prior knowledge of physics because students were recruited from a large variety of physics-related courses including introduction to physics, honors physics, and AP physics. This should allow us to draw general conclusions

4. DP-means BASED CLUSTERING

The DP-means algorithm, as described by Kullis & Jordan [4], is a hard-clustering approximation of nonparametric Bayesian models. Under the assumption that the DP-means is derived from a Dirichlet Process Mixture Model, there exists a lambda value α such that when used by the algorithm, the number of clusters k is identified. The DP-means algorithm is similar to the k-means clustering algorithm except that a new cluster is generated when the distance from a data point to the nearest cluster is larger than the threshold α.

Specifically, the DP-means algorithm is derived from a Dirichlet Process Mixture Model (DPMM) as illustrated below:

- $\mu_1, \dots, \mu_k \sim G_0$
- $\pi \sim Dir\ (k, \pi_0)$
- $z_1, \dots, z_n \sim Discrete\ (\pi)$
- $x_1, \dots, x_n \sim N(\mu_{z_i},\ \sigma I)$
- The Dirichlet prior of dim k is placed using some π_0

where:

- μ is the mean of each of the clusters, drawn from some base distribution G_0, which is the prior distribution over the means.
- $\pi = (\pi_1, \pi_2 \dots)$ corresponds to the vector of probabilities of being in a cluster.
- z_i is an indicator of cluster assignment.
- x_i is a data point

The corresponding clustering algorithm is described in Figure 1. The input consists of data instances $x_1, \dots, x_n$, where x_i represents the vector of pre-test answer choices of the i^{th} student. Since the pre-test contains 35 questions, each such response vector x_i contains 35 entries corresponding to each answer choice picked by student i. The clustering algorithm begins by initializing a single cluster whose mean is the global centroid. Then, it initializes a set of cluster indicators: $z_i = 1$ for all $i = 1, \ldots, n$ where $z_i = k$ means that the student x_i belongs to the k^{th} cluster as denoted by l_k.

In step 3, the algorithm computes the distances between each data point and the existing centroids. It then compares the minimum of these distances with α. If the minimum is larger than the threshold α, a new cluster is generated, and its centroid is assigned the current data point x_i. Otherwise, the cluster indicator of the current data point is set to the $argmin$ of the distances. After looping over all data points, the number of clusters k and the clusters indicators are computed. Finally, the DP-means algorithm generates the clusters l_j and their centroids μ_j for $j = 1, \ldots, k$. Step 3 is repeated until the algorithm converges.

Algorithm: DP-means

Input: $x_1, \ldots, x_n$: input data, α : cluster penalty parameter

Output: Clustering $l_1, \ldots, l_k$ and number of clusters k

1. Init. $k = 1, l_1 = \{x_1, \ldots, x_n\}$ and μ_1 the global mean.
2. Init. Cluster indicators $z_i = 1$ for all $i = 1, \ldots, n$
3. Repeat until convergence
   - For each point x_i
     - Compute $d_{ic} = \|x_i - \mu_c\|^2$ for $c = 1, \ldots, k$
     - If $d_{ic} > \alpha$, set $k = k + 1, z_i = k, and\ \mu_k = x_i$
     - Otherwise, set $z_i = argmin_c\, d_{ic}$
   - Generate clusters $l_1, \ldots, l_k$ based on $z_1, \ldots, z_k$: $l_j = \{x_i \mid z_i = j\}$
   - For each cluster l_j , compute $\mu_j = \frac{1}{|l_j|} \sum_{x \in l_j} x$.

Figure 1. DP-means algorithm

4. EXPERIMENTS AND RESULTS

4.1 Dataset

The data used in our experiments consists of pre-test answers collected from 264 high-school students who took the aFCI pre-test, went through a 5-week training period, and then took a post-test. Furthermore, after each training sessions students took a short post-test (6 questions). In all our experiments, we will use this post-test after the very first training session as the extrinsic evaluation criterion as it is closest in time (among all post-tests) to the pre-test and therefore is a good estimate of students' early knowledge states as best captured by the pre-test. The pretest includes 35 multiple choice questions that have the same weight. Two types of data have been used in our experiments: 5-way response data and binary response data. The categorical data consists of the actual answer choices students picked for the 35 multiple choice questions coded as A, B, C, D and E. For each question, one those choices is the correct answer. The binary data represents the same data coded as binary correctness values: 0 – incorrect, i.e.., the student picked any of the incorrect answer choices, and 1 – correct, i.e., the student picked the correct answer choice.

Tables 1 and 2 illustrate the data representation for the two tables. As described below, the columns represent the 35 questions and the rows represent individual students' responses.

Table 1.Categorial data

	Q1	**Q2**	**…**	**Q35**
Dh001	A	B	…	C
Dh002	C	D	…	C
…	…	…	…	…
DH356	C	D	...	B

Table 2. Binary data

	Q1	**Q2**	**…**	**Q35**
Dh001	0	0	…	0
Dh002	1	0	…	1
…	…	…	…	…
DH356	1	0	...	0

4.2 Experiments: Binary data

A first set of experiments have been conducted using the binary data as input for the DP-means algorithm. Since we have binary data and DP-mean is based on the Euclidean distance, we have applied Principal Component Analysis (PCA) to convert the binary values to continuous ones. For this purpose, numerous values of n (number of components) have been tested. The value 35 led to a convergence state of 10 clusters in which several clusters are redundant, i.e., using the extrinsic criterion based on the overall post-test score. For example, the average of the post test score for clusters 6, 7 and 9 is 3.0. Thus, we have tested randomly several values. The value 24 led to better clustering results in terms of splitting well the clusters based on the extrinsic criterion. Thus, we used those components to represent our data points for the rest of the experiments. On the binary data, a Manhattan distance could be used which we tried and didn't lead to better results than the above method of using PCA.

The α distance parameter has not been defined a priori. To select a suitable value of this parameter, we followed first the procedure described by Kulis and colleagues [4] as in the following: given k=3 as the desired number of clusters, we first initialize a set A with the global mean of the data. Then iteratively we calculate the maximum distance to A (the distance to A is the smallest distance among points in A). We repeat this k (=3) times and assign to α the value of the maximum distance to A. In our work, we got the value of 3.26. Testing the DP-means with this value led to the convergence of two clusters of students. To reach the desired number of clusters which is 3, we have tried other values in a close interval of [3.26, 2.8] as described in Table 3.

Thus, various values have been tested and compared. The evaluation of the resulted clusters has been done using the following measures:

- Silhouette index: Its value measures how similar a student response vector (her set of responses to the pre-test questions) is to its own cluster (cohesion) relative to the other clusters (separation). The silhouette index is a value within the [-1, +1] interval. A high value of the silhouette index indicates that the student is well matched with the other students in the same cluster. The following metric distances have been tested:

Euclidean distance, Manhattan distance and cosine similarity. The obtained results have shown that the Euclidean distance led to better results as demonstrated in Table 3.

- Mean of post test score: The data collected from the interaction of the students with the ITS includes post test scores for the 264 students. Since the post test is taken at the end of the experiment, weeks after the students took the pre-test, and since it has not been used in the cluster, it can be used as an extrinsic measure of cluster validity and interpretation of the resulting clusters. Indeed, this measure is used by us to assess the mastery level of each resulted cluster of students. In addition, it has been used as a way to check the separation of the clusters. The maximum and minimum values of the post test score in this collected data are 6.0 and 0.0 respectively.
- Mean of pretest score: The data collected includes the pretest performance of each student based on of the correct answers. The highest value is 35 and the lowest value is 0.

Table 3 offers a set of results of DP-means clustering using different types of distances.

Table 3. clustering results with different types of distances

Distance	α	Number of clusters
Manhattan	2.9	255
	3.0	255
	3.1	255
Euclidian	2.9	5
	3.0	3
	3.1	2
Cosine	2.9	1
	3.0	1
	3.1	1

Table 4. DP-means clustering results with different values of α

α	Clusters	Mean pretest score	Mean post-test score	Number of students
2.9	C1	15.26	3.31	207
	C2	31.28	5.66	36
	C3	6.47	1.89	19
	C4	25.0	5.0	1
	C5	14.0	2.0	1
3.0	C1	17.68	3.44	195
	C2	31.11	5.64	37
	C3	8.83	1.62	32
3.1	C1	13.87	3.18	277
	C2	29.54	5.64	37

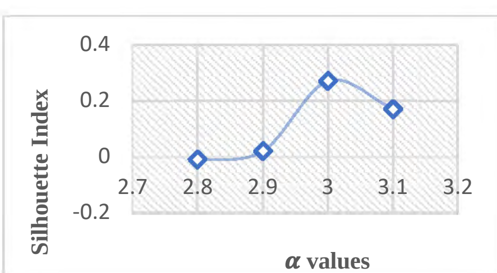

Figure 2. Quality of the DP-means algorithm using different values of α

The results in Figure 2 show that the value 3.0 of parameter α led to the highest value of the Silhouette index (0.27). In addition, this α value resulted in three distinct clusters, well separated (Figure 3), in terms of students' performance in the post test and pretest (as described in table 4). The first cluster contains 195 students. The mean post test score is 3.44 and the mean pretest score is 17.68 which are average scores. Students who belong to this cluster can be described as average performers or learners. The second cluster contains 37 students. The mean post test score is 5.66 and the mean pretest score is 31.11 which are high scores. The students in this cluster can be described as high performers or learners of Physics. The third cluster contains 32 students. The mean post test score is 1.625 and the mean pretest score is 8.83 which are very low scores. The students of this cluster can be describing as struggling ones.

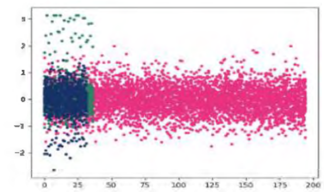

Figure 3. DP-means visualization with $\alpha = 3$

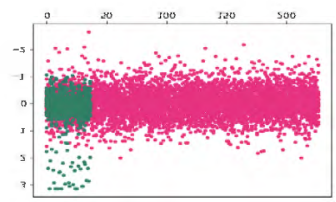

Figure 4. DP-means visualization with $\alpha = 3.1$

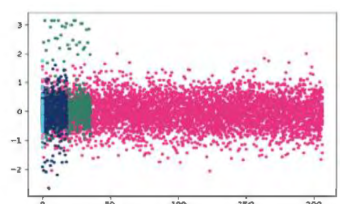

Figure 5. DP-means visualization with $\alpha = 3.2$

To compare the results of the DP-means algorithm with other clustering algorithms, we have also applied the k-means and agglomerative clustering algorithms on the same binary data. Since the best result of the DP-means was for an α value 3.0, we ran the k-means algorithm using k =3 and the agglomerative clustering using the same number of clusters (=3). Tables 5 and 6 present the results for k-means and agglomerative clustering, respectively.

Table 5. k-means results

Clusters	Mean Post-Test Score	Mean Pretest Score	Number of students
C1	2.28	9.44	97
C2	5.21	28.84	52
C3	3.83	17.22	115

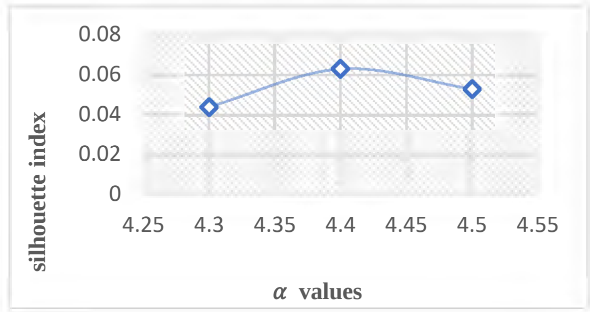

Figure 6. Quality of the DP-means algorithm using different values of α.

Table 6. Agglomerative clustering results

Clusters	Mean Post Test Score	Mean Pretest Score	Number of students	
C1	3.55	16.46	165	
C2	5.45	30.0	42	
C3	2.07	8.28	57	

The results depicted in Table 4 show that the DP-means algorithm with α=3.0 outperforms the k-means and the agglomerative algorithms as described in tables 5 and 6 respectively. The difference, in the mean post test score and the mean of pretest score, between the clusters of DP-means is larger than the difference using the other clustering algorithms. This indicates the convergence of well separated groups of students, in terms of learning level and prior knowledge, when applying the DP-means.

A detailed analysis of the top 10 students closest to the centroids of each of the three clusters found by DP-means, revealed that students in cluster 1 struggled mostly with questions related to Newton's third and first laws, whereas students in cluster two struggled with questions related to Newton's third law. Students in cluster 3 struggled the most and they showed weaknesses across all major topics in Newtonian Physics. Since in this experiment we used just correctness values for each pre-test question it is not possible to provide a more detailed analysis in terms of specific misconceptions, e.g., assuming faster velocity implies a larger force in an action-reaction pair, students in each clusters exhibit.

4.4 Experiments: Clustering categorical data

A second set of experiments have been conducted using categorical data and the DP-means and K-modes clustering algorithms. That is, in this case, we used the actual answer choices picked by students for the pre-test questions in order to find the clusters.

To this end, first, we have converted the categorical responses to numerical ones using one-hot encoding. Basically, each answer choice becomes a dimension in a vector space representation. A value of 1 is assigned to that dimension for a given question in the pre-test if a student picked the choice corresponding to the dimension as their answer choice. This results in an encoding of categorical integer features as a one-hot numeric array. The encoder derives the categories based on the unique values in each feature.The output of the one-hot-encoding is fed into the clustering algorithms.

The results presented in Table 7 reflect a decrease in quality of the DP-means clustering using categorical data. The silhouette index, as described in Figure 6, has decreased in comparison with the DP-means based on binary data. The highest value was 0.06 when using the value 4.4 of α. The different values of α didn't lead to a good split of students in terms of the performance. For example, in the case of α = 4.4, cluster C2 and C3 can be merged in one cluster since their mean post test and pretest scores are very close. For α = 4.3, there is redundancy in the resulted clusters. For example, C3 and C6 can be merged in one cluster.

Table 7. DP-means clustering results (categorical data) with different values of α

α	Clusters	Mean Post Test Score	Mean Pretest Score	Number of students
4.3	C1	4.17	20.39	68
	C2	1.5	5.5	67
	C3	1.0	10.0	1
	C4	2.0	4.5	54
	C5	2.0	6.0	40
	C6	1.66	6.0	22
	C7	2.16	9.35	12
4.4	C1	4.12	20.14	187
	C2	1.0	10.0	1
	C3	1.66	6.0	3
	C4	2.15	9.0	73
4.5	C1	3.55	16.87	263
	C2	1.0	10.0	1

In order to overcome this drawback of DP-means when applied to categorial data, we have applied the k-modes clustering algorithm [5, 6]. The k-modes algorithm is based on defining the dissimilarity measure between objects. This dissimilarity between two objects A and B can be defined by the total mismatches of the corresponding attribute categories of the two objects. The smaller the number of mismatches is, the more similar the two objects. The dissimilarity measure is calculated using the following equation:

$$d(X,Y) = \sum_{j=1}^{m} \delta(xj, yj) \qquad (1)$$

where:

$$\delta(xj, yj) = \begin{cases} 0 \ (xj = yj\ ) \\ 1 (xj \neq yj) \end{cases} \qquad (2)$$

The following are the results with k-modes when using k = 3.

Table 8. Kmodes results with k = 3

Clusters	Mean Post Test Score	Number of students
C1	3.72	120
C2	5.03	59
C3	2.23	85

The results listed in Table 8 demonstrate that the k-modes outperforms the DP-means when using categorical data. The resulted clusters reflect a good split between clusters in terms of performance in the post test. The C1 cluster reflects an average knowledge level of students. C2 reflects a high level of knowledge of students. And C3 reflects a low level of learning. A more detailed analysis indicates the same overall conclusions reached using the correctness data, e.g., students in cluster one struggle mostly with Newton's second and third laws. However, when using the categorical data, we can further pinpoint which aspects of Newton's third law for instance, students struggle with. For instance, many students in cluster C1 seem to struggle with the misconception that in an interaction between two objects the more massive one will act with a bigger force on the smaller one which is not true. According to Newton's third law, to each action there is an equal and opposite reaction. Therefore, this analysis suggests that when a new student uses a Physics ITS, after they take the pre-test and their answer patterns place him closer to the centroid of cluster C1, i.e., in cluster C1, then appropriate instructional tasks that have been designed for students in that cluster should be activated in order to overcome major gaps students in that cluster exhibit.

5. CONCLUSIONS

In this work, DP-means clustering algorithm has been applied on the pretest data of 264 students collected from their interaction with DeepTutor ITS. Various values of α have been tested. The results demonstrated that 3.0 is the best value and three distinct clusters of students have been converged. These clusters reflect three distinct levels of learning which has been assessed using post test scores. The first cluster of students correspond to an average level of learning, the second cluster represents students with a high level of learning and the third cluster of students those with a low level of learning. Results have demonstrated also that DP-means outperforms k-means and Agglomerative clustering in terms of splitting well students based on their performance in the post test. Another finding is that the quality of DP-means algorithm, measured by the silhouette index, decreases when we use the categorical data in comparison with the binary data. To overcome this drawback, we have used the k-modes clustering.

Furthermore, such clustering could offer a good trade-off between adaptivity and authoring costs. For instance, macro-adaptation can be expensive if the number of unique student knowledge states is very large as it requires selecting a unique set of tasks for each such unique knowledge state. Concretely, if using a 5-way/choice 35 multiple-choice question pre-test, the number of possible combinations of 35 answers is 5^{35}, an extremely large space. That is, if each student's knowledge state is described by the 35 responses we end up with 5^{35} student knowledge states or student models which, by comparison, is much larger than the world's population which is a bit over 5^{14}. Considering each of these potential knowledge states and selecting for each corresponding learner a unique set of tasks becomes a computationally and authoring challenge. An alternative, for instance, would be to group students into clusters of similar mental models and then select and author tasks for each such clusters. That is, grouping students into similar mental model groups can offer a good trade-off between adaptivity and authoring costs.

We plan to further investigate the resulting clusters for a better understanding of the characteristics of the students in each cluster. For instance, we do have information about the Physics class (intro, honors, AP) each student took and therefore a detailed analysis for students in each cluster based on their class type can be performed in order to understand what are the major misconceptions students in each class struggle with. Not only this could inform an ITS for Physics, but this information can be shared with teachers in order to help them better plan their lessons plans to address major misconceptions their students may have.

6. ACKNOWLEDGMENTS

This research was partially sponsored by the University of Memphis and the Institute for Education Sciences under award R305A100875 to Dr. Vasile Rus.

7. REFERENCES

[1] Bouchet, F., Harley, J. & Trevors, G. & Azevedo, R. (2013). Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning. Journal of Educational Data Mining. 5. 104-146.

[2] Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. Cognitive Science, 25, 471–533 (2010)

[3] Fang, Y., Shubeck, K., Lippert, A. & Cheng, Q. & Shi, G. & Feng, S. & Gatewood, J. & Chen, S. & Cai, Z., Pavlik, Jr, P., Frijters, J., Greenberg, D. & Graesser, A. (2018). Clustering the Learning Patterns of Adults with Low Literacy Skills Interacting with an Intelligent Tutoring System.

[4] Kulis, B. & Jordan, M. I. (2012). Revisiting k-means: New algorithms via Bayesian nonparametrics. *In Proceedings of the 23rd International Conference on Machine Learning, 2012*.

[5] Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values, Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, pp. 21-34, 1997.

[6] Huang, Z. (1998). Extensions to the k-modes algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2(3), pp. 283-304, 1998.

[7] Reyes-González, Y., Martínez-Sánchez, N., Díaz-Sardiñas, A & Patterson-Peña, M.C. (2018). Conceptual clustering: A new approach to student modeling in Intelligent Tutoring Systems. Revista Facultad de Ingenieria. 70-76. 10.17533/udea.redin.n87a09.

[8] Rodrigo, Ma.M.T., A. Anglo, E., O. Sugay, J., & Baker, R. (2008). Use of Unsupervised Clustering to Characterize Learner Behaviors and Affective States while Using an Intelligent Tutoring System. *Proceedings - ICCE 2008: 16th International Conference on Computers in Education. 49-56.*

[9] Rus, V., D'Mello, S. K., Hu, X., & Graesser, A. C. (2013). *Recent advances in intelligent tutoring systems with conversational dialogue*, AI Magazine, 34(3), 42-54.

[10] Rus, V., Niraula, N.B., and Banjade, R. (2015). DeepTutor: an effective, online intelligent tutoring system that promotes deep learning. *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15). AAAI Press 4294-4295.*

[11] Shute, V.J., Zapata-Rivera, D. (2012). Adaptive educational systems. Adaptive technologies for training and education, P. DURLACH, Ed. New York, NY: Cambridge University Press, 7-27.

[12] Woolf, B.: Building Intelligent Interactive Tutors, Student-Centered Strategies for Revolutionizing E-Learning, Elsevier & Morgan Kaufmann Publishers, (2008).