Skill Set Profile Clustering Based on Student Capability Vectors Computed From Online Tutoring Data

Elizabeth Ayers[1], Rebecca Nugent[1], and Nema Dean[2]
{eayers, rnugent}@stat.cmu.edu, {nema}@stats.gla.ac.uk
[1]Department of Statistics, Carnegie Mellon University
[2]Department of Statistics, University of Glasgow

Abstract. In educational research, a fundamental goal is identifying which skills students have mastered, which skills they have not, and which skills they are in the process of mastering. As the number of examinees, items, and skills increases, the estimation of even simple cognitive diagnosis models becomes difficult. To address this, we introduce a capability matrix showing for each skill the proportion correct on all items tried by each student involving that skill. We apply variations of common clustering methods to this matrix and discuss conditioning on sparse subspaces. We demonstrate the feasibility and scalability of our method on several simulated datasets and illustrate the difficulties inherent in real data using a subset of online mathematics tutor data. We also comment on the interpretability and application of the results for teachers.

1 Introduction

In educational research, a fundamental goal is identifying which skills students have mastered, which skills they have not, and which skills they are in the process of mastering. A variety of cognitive diagnosis models [12] address this problem using information from a student response matrix and an expert-elicited assignment matrix of the skills required for each item. However, even simple models [10] become more difficult to estimate as the number of skills, items, and examinees grows [1]. Any procedure used should be able to handle the missing values that arise in assessment; students may not have time to finish all items, for example, or they might intentionally skip items. Moreover, data is often missing by design in assessment embedded in online tutoring systems.

In response, we introduce a *capability matrix* showing for each skill the proportion correct on all items tried by each student involving that skill, expanding on the sum-score work of [8]. We apply clustering methods to the capability matrix to identify groups of students with similar skill set profiles, similar to [14] which clusters students based on their collaborative behavior. In addition, we propose a conditional clustering heuristic that takes advantage of obvious group separation in one or more dimensions. These methods are faster than common cognitive diagnosis models, provide a unique visualization tool of students' skill mastery, and scale well to large datasets. We show that these methods also add flexibility in the assignment of skill mastery; we are also able to determine the students' skills for which mastery is uncertain, a conservative classification scheme that does not force a hard skill mastery assignment of yes or no. For illustrative purposes, we demonstrate our method on three datasets simulated from the DINA model [10], a common cognitive diagnosis model, and on a small subset of data obtained from the ongoing IES Assistment Project [7]. Finally we conclude with comments on current and future work.

2 The Capability Matrix

We begin by assembling the skill dependencies of each item into a Q-matrix [2,3,15]. The Q-matrix, also referred to as a transfer model or skill coding, is a $J \times K$ matrix where $q_{jk} = 1$ if item j requires skill k and 0 if it does not, J is the total number of items, and K is the total number of skills. The Q-matrix is usually an expert-elicited assignment matrix. This paper assumes the given Q-matrix is known and correct.

$$Q = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,K} \\ \vdots & \ddots & & \vdots \\ q_{J,1} & q_{J,2} & \cdots & q_{J,K} \end{bmatrix}, \quad Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,J} \\ \vdots & \ddots & & \vdots \\ y_{N,1} & y_{N,2} & \cdots & y_{N,J} \end{bmatrix}$$

Student responses are assembled in a $N \times J$ response matrix Y where y_{ij} indicates both if student i attempted item j and whether or not they answered item j correctly and N is the total number of students. If student i did not answer item j then $y_{ij} = NA$. The indicator $I_{y_{ij} \neq NA} = 0$ expresses this missing value. If student i attempted item j ($I_{y_{ij} \neq NA} = 1$), then $y_{ij} = 1$ if they answered correctly, or 0 if they answered incorrectly.

To cluster students by their skill set profiles, we need a summary statistic of their skill performance. We define an $N \times K$ *capability matrix* B, where B_{ik} is the proportion of correctly answered items involving skill k that student i attempted. If student i did not attempt any items with skill k, we assign a value of 0.5, an uninformative probability of skill mastery. That is, if $\sum_{j=1}^{J} I_{Y_{ij} \neq NA} \cdot q_{jk} = 0$, $B_{ik} = 0.5$. Otherwise,

$$B_{ik} = \frac{\sum_{j=1}^{J} I_{Y_{ij} \neq NA} \cdot Y_{ij} \cdot q_{jk}}{\sum_{j=1}^{J} I_{Y_{ij} \neq NA} \cdot q_{jk}} \qquad \text{where } i = 1, 2, ...N; k = 1, 2, ..., K \tag{1}$$

where Y_{ij} and q_{jk} are the corresponding entries from the response matrix Y and Q-matrix.

There are several benefits of using a summary statistic of this form. The statistic scales for the number of items in which the skill appears as well as for missing data. If a student has not seen all or any of the items requiring a particular skill, we still derive an estimate based on the available information. Also, the values naturally fall onto a skill mastery scale. For each skill, zero indicates that a student has not mastered that skill, one indicates that they have, and 0.5 indicates uncertainty, partial mastery, or no information.

Moreover, the vectors $B_i = \{B_{i1}, B_{i2}, ..., B_{iK}\}$ for $i = 1, 2, ..., N$, lie in a K-dimensional unit hyper-cube where each skill corresponds to a dimension and each corner is one of the 2^K natural skill set profiles $C_i = \{C_{i1}, C_{i2}, ..., C_{iK}\}$, $C_{ik} \in \{0, 1\}$. In Figure 1, we show the corresponding hyper-cubes for $K = 2, 3$ with selected profile locations labeled. For example, if $K = 3$ and a student has the first two skills but not the third, their true skill set profile would be $\{1, 1, 0\}$, the triangle in the bottom right back corner in Figure 1(b). The B_i's map each student into the unit hyper-cube. Ideally, students would be represented with a point mass at each of the 2^K corners. However in practice, students will not map directly to the corners due to error, they may guess without having the skills or they may have the skills and slip. In the capability matrix and the corresponding hyper-cube, values near or at zero and one indicate certainty about skill mastery (no/yes). We are less certain about skill mastery for values near 0.5. Note that multiple students may map to the same locations.
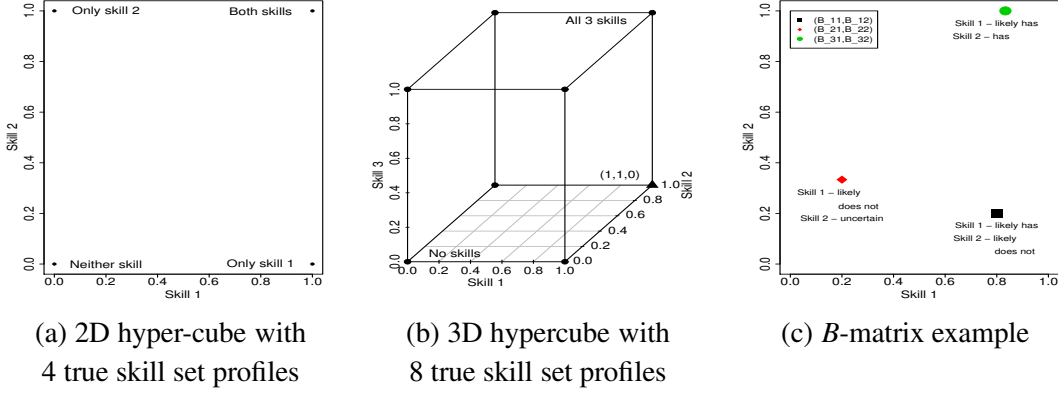
(a) 2D hyper-cube with   (b) 3D hypercube with   (c) B-matrix example
4 true skill set profiles   8 true skill set profiles

Figure 1: Examples of 2D and 3D hyper-cubes with skill set profiles and a 2D B-matrix.

Suppose that we have the response matrix Y, the Q-matrix, and corresponding B-matrix shown below. Figure 1(c) illustrates the corresponding mapping of the three students into the two-dimensional hyper-cube. For student 1 at $\{0.8, 0.2\}$, we might say they likely have skill 1 but likely do not have skill 2. For student 2 at $\{0.2, 0.33\}$ we might say they likely do not have skill 1 or skill 2, but we are less certain about their skill 2 status. Finally, for student 3 at $\{0.83, 1\}$, we would say that they likely have skill 1 and definitely have skill 2.

$$Y = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & NA & 1 \\ 0 & 0 & NA & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad B = \begin{pmatrix} B_{11} = \frac{4}{5} & B_{12} = \frac{1}{5} \\ B_{21} = \frac{1}{5} & B_{22} = \frac{2}{6} = \frac{1}{3} \\ B_{31} = \frac{5}{6} & B_{32} = \frac{6}{6} = 1 \end{pmatrix}$$

$$Q^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

3 Clustering Methods

To identify groups of students with similar skill set profiles, we cluster the rows of the B matrix (and subsequently partition the hyper-cube) using two commonly used clustering procedures: k-Means and model-based clustering. While other methods are available, characteristics of these clustering procedures make them natural choices for this problem.

3.1 k-Means (With Empty Clusters)

k-Means [6] is a popular iterative descent algorithm for data $X = \{\underline{x}_1, \underline{x}_2 ..., \underline{x}_n\}$, $\underline{x}_i \in \mathfrak{R}^K$. It uses squared Euclidean distance as a dissimilarity measure and tries to minimize within-cluster distance and maximize between-cluster distance. For a given number of clusters G, k-Means searches for cluster centers m_g and assignments A that minimize the criterion

$$\min_A \sum_{g=1}^{G} \sum_{A(i)=g} \|\underline{x}_i - m_g\|^2.$$

The algorithm alternates between optimizing the cluster centers for the current assignment (by the current cluster means) and optimizing the cluster assignment for a given set of

3

cluster centers (by assigning to the closest current center) until convergence (i.e. cluster assignments do not change). It tends to find compact, spherical clusters and requires *a priori* both the number of clusters G and a starting set of cluster centers. The final cluster assignment can be sensitive to the choice of centers; a common method for initializing k-Means is to randomly choose G observations. However, in our hyper-cube, we have a natural set of starting cluster centers, the 2^K skill set profiles at the corners. If students map closely to their profile corners, k-Means should locate the groups affiliated with the corners quickly. However, if we are missing students from one or more skill set profiles in our population, forcing $G = 2^K$ clusters will split some clusters unnecessarily. We modify the k-Means algorithm to allow for empty clusters (or absent skill set profiles) in the following way:

1. Set the starting cluster centers m_g to the corners of the K-dim hyper-cube (2^K centers).
2. Create the cluster assignment vector A by assigning each B_i to the closest m_g.
3. For all clusters g, if no B_i is assigned to m_g, i.e. $\sum I_{A(i)=g} = 0$, then m_g remains the same. Else, $m_g = \frac{\sum_{i=1}^n I_{A(i)=g} \cdot B_i}{\sum_{i=1}^n I_{A(i)=g}}$.
4. Alternate between 2) and 3) until the cluster assignment vector A does not change.

This flexible k-Means variation allows for empty clusters or fewer clusters than originally requested and removes the constraint that there be one cluster per skill set profile.

3.2 Model-based Clustering

Model-based clustering [4,5,11] is a parametric statistical approach that assumes: the data $X = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_n\}$, $\underline{x}_i \in \Re^K$ are an independently and identically distributed sample from some unknown population density $p(\underline{x})$; each population group g is represented by a (often Gaussian) density $p_g(\underline{x})$; and $p(\underline{x})$ is a weighted mixture of these density components, i.e.

$$p(\underline{x}) = \sum_{g=1}^G \pi_g \cdot p_g(\underline{x}; \theta_g)$$

where $\sum \pi_g = 1$, $0 < \pi_g \leq 1$ for $g = 1, 2, ..., G$, and $\theta_g = (\mu_g, \Sigma_g)$ for Gaussian components. The method finds estimates for the number of clusters G as well as their centers and variances (μ_g, Σ_g) that maximize a chosen information criterion. Essentially, it finds the weighted combination of Gaussian densities that "best fits" the data. While it may require the groups to have Gaussian densities, it is very flexible (unlike k-Means) on the shape, volume, and orientation of the densities. This freedom allows model-based clustering to fit a wide array of student groups of different shapes and sizes.

Both methods return a set of cluster centers and variances and an assignment vector mapping each B_i to a cluster. They do not, however, automatically assign a natural skill set profile (hyper-cube corner) to each cluster. Ideally, we have 2^K clusters, each closest to a unique corner. In reality, some corners will have no students nearby. The k-Means algorithm has been altered to allow for this option; model-based clustering estimates centers in high-frequency areas and should not put a center near an empty corner. We do not advocate a one-to-one mapping of clusters to corners; clusters near areas of uncertainty in the hyper-cube should be identified as such. If a cluster of students is centered at $\{0.12, 0.88, 0.55\}$,

they should be labeled as likely not having skill 1, likely having skill 2, and uncertain on skill 3. This conservative classification will help teachers avoid misclassifying students. To classify a new student, we calculate the capability vector and assign to the nearest cluster.

3.3 Subspace Clustering

If few items require skill k, B_{ik} only take a few unique values. For example, if three items need skill k, $B_{ik} \in \{0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\}$. Clustering on the K-dimensional hyper-cube may not perform well as students will map to only a few (K-1)-dimensional hyper-cubes. Instead we recommend conditioning on the coarsely gridded dimension (skill k, where students are already well-separated) and clustering on the (K-1)-dimensional conditional subspaces (repeating as needed).

4 Examples

For our simulated data, we use the deterministic inputs, noisy "and" gate model (DINA; [10]) a conjunctive cognitive diagnosis model. The DINA model item response form is

$$P(Y_{ij} = 1 \mid \eta_{ij}, s_j, g_j) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}$$

where $\alpha_{ik} = I_{\{\text{Student } i \text{ has skill } k\}}$ indicates if student i possesses skill k, $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ indicates if student i has all skills needed for item j, for item j, $s_j = P(Y_{ij} = 0 \mid \eta_{ij} = 1)$ is the slip parameter and $g_j = P(Y_{ij} = 1 \mid \eta_{ij} = 0)$ is the guess parameter. If a student is missing any of the required skills, the probability that they will answer an item correctly drops due to the conjunctive assumption.

When simulating data from the DINA model, we first fix skill difficulties and inter-skill correlation and generate true skill set profiles C_i for each student. If skills are of equal difficulty with little or no inter-skill correlation, students are evenly spread among the 2^K natural skill set profiles. If skill difficulty varies, skill set profiles with only "easy" skills will have more students than those including the "hard" skills. High inter-skill correlation pulls students toward the no mastered skills and all mastered skills corners ($C_i = \{\underline{0}\}, \{\underline{1}\}$). Next we draw slip and guess parameters from a random uniform distribution ($s_j \sim \text{Unif}(0,0.30)$; $g_j \sim \text{Unif}(0,0.15)$). Given profiles and slip/guess parameters, we generate the student response matrix Y. Prior to clustering, we remove 10% of the responses completely at random.

For these examples we know the true underlying skill set profiles C_i and can calculate their agreement with the clustering partitions using the Adjusted Rand Index (ARI; [9]), a common measure of agreement between two partitions. The expected value of the ARI is zero and the maximum value is one, with larger values indicating better agreement.

4.1 Simulated DINA Data

In Example 1, we generated response data for $N = 250$ students for $J = 30$ items, $K = 2$ skills. The Q-matrix contains only single skill items, 15 items per skill. The skills are equal difficulty with an inter-skill correlation of 0.25. Figure 2(a) shows the results. Clusters are number/color coded with triangle centers. We asked k-Means for $2^K = 4$ clusters; all students were clustered correctly (ARI = 1). Model-based clustering chooses five clusters

5

(ARI = 0.926). The "extra" high frequency area near {1, 1} results from the close proximity or identical locations of the 19 students in Cluster 5. Teachers could interpret these results as two groups with similar skill 2 mastery but different skill 1 mastery.

(a) Ex 1: 2 skills; all single skill items

(b) Ex 2: 3 skills; all single skill items

(c) Ex 3: 3 skills; single, double and triple skill items
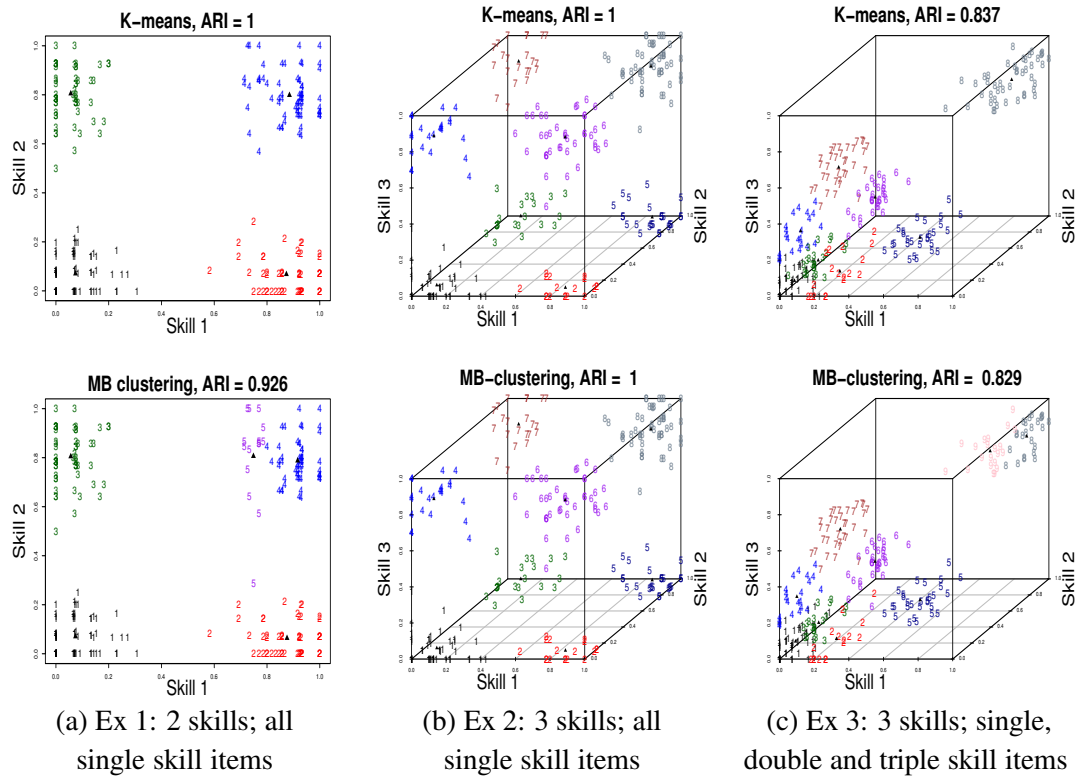
Figure 2: Simulated data examples for $K = 2, 3$ skills for single skill and multiple skill items, 10% missing responses. Clusters are color/number coded, centers denoted by black triangles.

In Example 2, we simulated as in Example 1 but increased the number of skills to $K = 3$. Again the Q-matrix was designed to only include single skill items, 10 items per skill. Here, both k-Means and model-based clustering recovered the true skill set profiles (ARI=1). Figure 2(b) shows the clustering results for both methods.

For Example 3, we simulated as in Example 2 but used a balanced design Q-matrix including multiple skill items where each skill appeared by itself in four items, in four double skill items with each of the other two skills, and in three triple skill items. Results are in Figure 2(c). Both methods find clusters of students showing mastery of all three skills in the back upper right corner near the {1, 1, 1} skill set profile. However, the remaining students are pulled toward the front lower left corner (the {0, 0, 0} skill set profile), a direct result of the combination skill items. If a student incorrectly answers a multiple skill item, all skills required by that item are penalized (not just the unmastered skills). We have seen that a balanced design negates the penalty effect (ARI = 0.837, 0.829); the remaining clusters are effectively scaled and maintain their separation.

The datasets presented are missing 10% of the responses; we compare their results to those for only students not missing any responses. In educational data mining, we commonly use case-wise deletion of students to generate a complete dataset. This method is

impractical here as it leaves us with 11, 10, and 15 students respectively. Instead we use the original generated response matrices prior to removing responses at random. The B-matrices are re-calculated and clustered. Only Example 3 had different ARIs. When using the complete data set, the ARI for k-Means increases from 0.837 to 0.880, for model-based clustering, 0.829 to 0.946. These jumps are expected as the lack of missingness increases the number of items seen (and the fineness of the grid) and decreases the relative effect of the penalty associated with incorrectly answering a multiple skill item; the resulting clusters are less removed from the corners.

A higher dimensional example with $N = 1000$ students, $J = 80$ items, and $K = 20$ skills was also explored. In this case there were 425 unique latent classes used to generate the data. Model-based clustering found 424 clusters and had an ARI of 0.99. Giving k-means 2^{20} starting centers is unreasonable; we're currently developing methods to systematically and appropriately choose a smaller set of starting centers.

4.2 Assistment Data

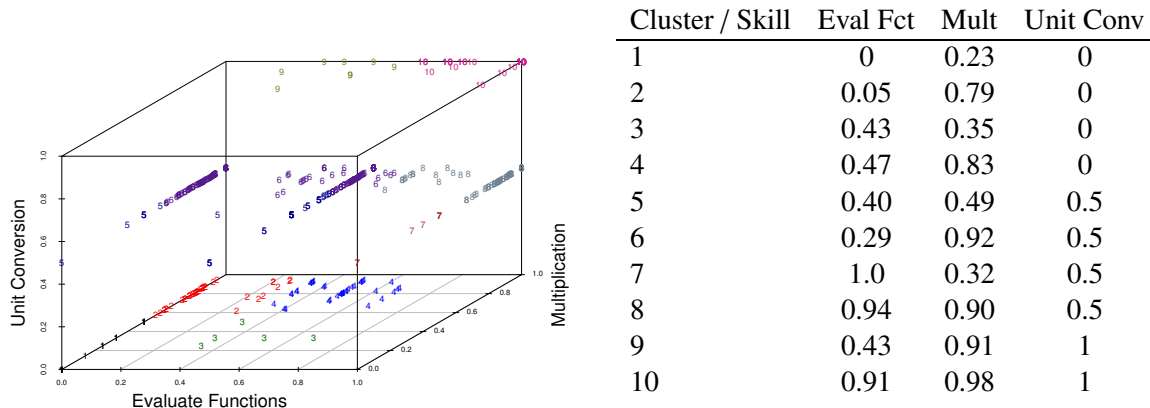| Cluster / Skill	Eval Fct	Mult	Unit Conv
1	0	0.23	0
2	0.05	0.79	0
3	0.43	0.35	0
4	0.47	0.83	0
5	0.40	0.49	0.5
6	0.29	0.92	0.5
7	1.0	0.32	0.5
8	0.94	0.90	0.5
9	0.43	0.91	1
10	0.91	0.98	1

Figure 3: Assistment System example of conditional k-Means clustering on the B-matrix; clusters are color/number coded. The table shows the cluster centers.

For our real data, we use a subset of 26 items requiring three skills (for easy visualization) from the Assistment System online mathematics tutor [7]. The Q-matrix is unbalanced; Skill 1 (Evaluating Functions) appears in eight items, Skill 2 (Multiplication) in 20 items, and Skill 3 (Unit Conversion) in two items. Overall, 551 students answered at least one item, however there is a large amount of missing data (57%). Recall, if student i did not see any items requiring skill k, $B_{ik} = 0.5$. Since Unit Conversion appears in only two items, $B_{iUC} \in \{0, \frac{1}{2}, 1\}$. The three corresponding planes are visible in Figure 3. We condition the unique B_{iUC} values and apply our k-Means variation (Section 3.1) to each plane. The final cluster centers are in the table in Figure 3. k-Means is preferable here because the limited number of unique values in the Evaluate Functions skill dimension leads to instability in the more flexible model-based clustering models. The planes corresponding to $B_{iUC} = 0$ and 0.5 each have four clusters; the plane for $B_{iUC} = 1$ has two. There are natural interpretations for each of the clusters. For example, a teacher might interpret Cluster 9 as students who know Unit Conversion and Multiplication, but are uncertain on Evaluating Functions. Cluster 10 could be interpreted as the students who have mastered all three skills.

7

5 Conclusions and Future Work

We derived a capability matrix to summarize student skill mastery for use in clustering algorithms. In simulated datasets, the method performed well (i.e., high values of ARI). In the Assistments data the method responded well to missing data, allowing us to draw conclusions for the skills that students have seen and distinguish the skills that require more assessment. Early results suggest that the Q-matrix design plays a large role in the location and interpretation of the clusters. Finally, we visually presented examples with $K = 2$ and $K = 3$ skills and showed the method scales to a larger number of skills.

Currently, we are comparing our results to other student skill knowledge estimates. For example, using WinBUGS [13], the DINA model estimates produce essentially the same profile clusters for the simulated datasets; however, it runs around 700 times more slowly.

References

[1] Anozie, N.O. and Junker, B. W. (2007). *Investigating the utility of a conjunctive model in Q-matrix assessment using monthly student records in an online tutoring system.* National Council on Measurement in Education (NCME-07), April 12, 2007, Chicago, IL.

[2] Barnes, T.M. (2003). *The Q-matrix Method of Fault-tolerant Teaching in Knowledge Assessment and Data Mining.* Ph.D. Dissertation, Department of Computer Science, North Carolina State University.

[3] Embretson, S.E. (1991). A Multidimensional Latent Trait Model for Measuring Learning and Change. *Psychometrika*, 56, 495-515.

[4] Fraley, C. and Raftery, A. (1998). How many clusters? which clustering method? - answers via model-based cluster analysis. *The Computer Journal*, 41, 578-588.

[5] Fraley, C. and Raftery, A. (1999). Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16, 297-306.

[6] Hartigan, J. and Wong, M.A. (1979). A k-means clustering algorithm. *Applied Statistics*, 28, 100-108.

[7] Heffernan, N.T., Koedinger, K.R. and Junker, B.W. (2001). *Using Web-Based Cognitive Assessment Systems for Predicting Student Performance on State Exams.* Research proposal to the Institute of Educational Statistics, US Department of Education. Department of Computer Science at Worcester Polytechnic Institute, Worcester County, Massachusetts.

[8] Henson, J., Templin, R., and Douglas, J. (2007). Using efficient model based sum-scores for conducting skill diagnoses. *Journal of Education Measurement*, 44, 361-376.

[9] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193-218.

[10] Junker, B.W. and Sijtsma K. (2001). Cognitive Assessment Models with Few Assumptions and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, 25, 258-272.

[11] McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*.

[12] Nichols, P.D., Chipman, S.F., and Brennan, R.L. (1995). *Cognitively Diagnostic Assessment.* Lawrence Erlbaum Associates.

[13] Spiegelhalter, D.J., Thomas, A. and Best, N.G. (2003). *WinBUGS: Bayesian Inference Using Gibbs Sampling, Manual Version 1.4.* Cambridge: Medical Research Council Biostatistics Unit.

[14] Talavera, L., and Gaudioso, E. (2004). Mining student data to characterize similar behaviour groups in unstructured collaboration spaces. *Proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI 2004.* Valencia, Spain.

[15] Tatsuoka, K.K. (1983). Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory. *Journal of Educational Measurement*. Vol. 20, No. 4, 345-354.