Validation of Decision Tree Algorithm for Initialization into Dynamic Assessment

Chen & Clark

Key words: (1) dynamic assessment, (2) Bayesian data analysis, (3) alternate assessment

Abstract:

Dynamic assessment, built upon an underlying learning map structure, presents unique items matched to each student's knowledge, skills, and abilities. This paper seeks to validate a decision tree approach to initially place students into the assessment at an appropriate level of complexity using regression and Bayesian analyses.

Background:

Like adaptive assessment, dynamic assessment presents unique items matched to each individual student's knowledge, skills, and abilities. One key difference is that dynamic assessment is built upon an underlying learning map structure consisting of nodes, or fine-grained learning targets. While student response data informs subsequent routing through the learning map, it is critical that a student's initial entry into the map provides an optimal match of the student to the complexity of the item, especially when assessing students with significant cognitive disabilities. Should initial items be too challenging upon entry into the assessment system, little information is gained from student responses. Furthermore, student frustration or disengagement may abound. As such, this paper explores the extent that a decision tree approach can be validated for use in initial placement of students into the assessment system.

Methods:

After extensive consultation with content and special education experts, a decision tree was developed to facilitate student entry into the learning map for English language arts and mathematics content areas. Data for each decision tree, and subsequent placement into the learning map, is obtained from teacher entered responses from a survey about each individual student's academic and expressive communication abilities. These responses are required for participation in the assessment due to their use for initial placement.

Data was collected across multiple testing events to validate the decision tree algorithm: a fall 2013 pilot of approximately 1400 students, and three spring 2014 field test events, each with roughly 10,000 students. During the pilot event, all students were assigned to fixed forms spanning multiple levels of complexity in order to collect data from students at varying knowledge, skill, and ability levels.

Using response data from the pilot, a series of regression models were fit to determine the extent that student achievement and expressive communication ability could be used to predict total score and performance within each set of items over increasingly complex levels. A Bayesian clustering algorithm was also fit to the data to classify students into four ordinal categories. These categories were then compared to the four categories determined by the decision tree algorithm, and decision consistency was evaluated.

Item difficulty analysis from the field test data was used to further evaluate the effectiveness of the decision tree to place students in the map. Items were examined within and

across bands to confirm fit of items to students. This included examining percent correct by node (learning target) and item difficulty and slippage parameters from a Bayesian model.

Results

Because fixed forms were administered in the pilot, percent correct could be directly compared across students in the four complexity bands. Findings indicated that as the tested level of the item increased (became more cognitively complex), items became more challenging for students classified to the lower bands by the decision tree, as expected. Similarly, at the highest band, items at the lowest level were answered correct by nearly all the students. The table below provides the percent correct across items in three testlets of increasing difficulty for one grade band assessment.

Band	Low Item 1	Low Item 2	Low Item 3	Mid Item 1	Mid Item 2	Mid Item 3	Target Item 1	Target Item 2	Target Item 3
0 - low (N=37)	38%	32%	32%	32%	30%	30%	22%	22%	19%
1 (N=92)	76%	70%	68%	37%	43%	43%	40%	43%	28%
2 (N=68)	87%	90%	87%	66%	62%	60%	54%	59%	44%
3 - high (N= 8)	88%	88%	88%	88%	88%	88%	75%	75%	75%

The OLS regression model using student the academic variables from the decision tree significantly predicted total score for ELA and mathematics across all three grade-band assessments. The RMSE values indicate a sample standard deviation of around 1.0 for all models. The addition of expressive communication variables to be the models resulted in a slightly smaller RMSE value for both content areas, and more conservative classification to linkage levels. Therefore, overall, the regression findings suggest that the academic and communication variables selected for learning map initialization were successful predictors of performance on the pilot assessment. Similar findings were obtained for an ordinal and logistic regression model predicting optimal performance at each of the three tested levels.

The machine learning portion of this study using data from the pilot is underway and results will be included in the final paper.

Field test data was collected to provide further support for use of the decision tree. After examining percent correct within node by band, findings indicate that items were well matched to students. Should items appear to be too rigorous ($p < .35$) by band, there would be evidence that the algorithm is not effectively matching students to items. The table below provides a summary of flags from field tests 1 and 2. Less than 5% of items were flagged as being too challenging.

Grade	ELA Flag	ELA #	ELA %	Math Flag	Math #	Math %
3rd	1	166	0.6%	3	230	1.3%
4th	4	177	0.5%	3	179	1.7%
5th	0	156	0.0%	8	160	5.0%
6th	5	137	3.6%	6	127	4.7%
7th	23	191	12.0%	4	247	1.6%
8th	12	149	8.1%	3	124	2.4%
9th - 10th	4	103	3.9%	1	148	0.7%
11th - 12th	8	195	4.1%	--	--	--
Total	57	1274	4.5%	28	1215	2.3%

Implication

The research presented here provides a unique marriage of content expertise and statistical information to provide an optimal first match of students to items in a dynamic assessment environment. Furthermore it contributes valuable findings regarding initialization into a dynamic assessment on a large scale basis that will inform not only the operational program examined but the larger research community.