A New Method for Measuring Similarity Between Educational Items from Response Data

Tanya Nazaretsky[1], Sara Hershkovitz[2], Giora Alexandron[1]

[1]Weizmann Institute of Science, [2]The Center for Educational Technology

Introduction

- A main goal of Educational Data Mining (EDM) is developing methods for exploring large-scale data that come from interactive learning environments, and using those methods for improving learning outcomes [1]
- A fundamental question within EDM (and *Psychometrics*) is identifying groups of items (questions) that require the same set of skills.
- Standard statistical methods that are used for that are based on the assumption that student's performance on items that require the same skill should be similar (see for example in [2]). This holds if the latent trait is relatively fixed during the activity being measured, as in the context of *testing*
- However, this assumption *does not* hold in the context of *learning*, which means that the latent trait changes rapidly
- We propose a novel similarity measure, termed *Kappa Learning*, which aims to identify similarity between items under the assumption that the latent trait can change, namely, that students can acquire new skills during the activity

Cohen's Kappa and Kappa Learning

Cohen's Kappa: An index that measures inter-rater agreement for qualitative items

- We consider Items as Raters, Learners as Subjects to classify, and learner answers as classification results
- The raters 'agree' if a student gives the same answer to the pair of items
- We use the term Knowledge Component (KC) to denote a set of items that require that same skill. We assume that each item requires one skill
- Value of 1/0 means a learner has mastered/not mastered the KC that the item belongs to

$$P_o = \frac{a+d}{n}, P_e = \frac{(a+b)(a+c)+(b+d)(c+d)}{n^2}$$

Kappa Learning: A new measure of similarity that assumes learning

- It accommodates learning by giving a different interpretation to the notion of 'agreement' in Cohen's Kappa formula and taking into account possible improvement of students' skills
- In case a student got the first item incorrect and the second item correct, we interpret this as *learning*, namely, mastering the skill underlying the item (*guess* and *slip* can occur, but are not modeled explicitly). **This is an additional case of agreement, and is where our measure differs from Cohen's Kappa.** We then get the following definitions for P_o,P_e:

$$P_o = \frac{a+\mathbf{b}+d}{n}, P_e = \frac{(a+b)(a+c)+\mathbf{(a+b)(b+d)}+(b+d)(c+d)}{n^2}$$

Kappa Definition

$$Kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o = observed level of agreement
P_e = expected level of agreement

Notation

Let's define a contingency table. Assume Q_1, Q_2 is a pair of items.

- a - number of students answered both Q_1 and Q_2 correctly
- b - number of students answered Q_1 incorrectly and Q_2 correctly
- c - number of students answered Q_1 correctly and Q_2 incorrectly
- d - number of students answered both Q_1 and Q_2 incorrectly

n=a+b+c+d - total number of students

	Q_1 +	Q_1 -	
Q_2 +	a	b	a + b
Q_2 -	c	d	c + d
	a + c	b + d	n

Procedure

1. From students' response data, compute *user-based* item similarity matrix for Kappa Learning and the 3 reference measures (Cohen's Kappa, Yule, and Pearson)
2. Compute *item-based* Pearson distance matrix from the user-based similarity matrix
3. Run K-Means and Ward's Hierarchical clustering on the *item-based* distance matrix. The number of clusters is derived from the ground truth tagging supplied by the subject matter experts
4. Per clustering, use Adjusted Rand Index to measure the goodness-of-fit against ground truth

Empirical Settings

- The data come from a Computerized Tutor that teaches Fractions for 4th grade
- The data contain the response data of 594 students on 551 items
- The subject matter experts identified 83 KCs and tagged each item with the corresponding KC

Results

Similarity measure	Adjusted Rand Index
Kappa Learning	**0.36**
Kappa	0.25
Yule	0.29
Pearson	0.29

Table 1: Hierarchical Clustering for each similarity measure

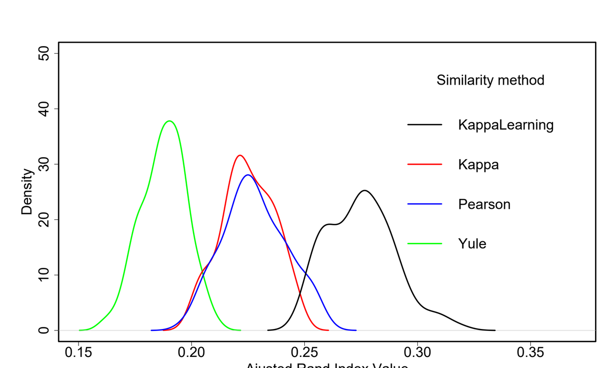

Figure 1: K-means clustering (100 iterations per similarity measure)

Conclusions

Kappa Learning outperforms other similarity measures in terms of goodness-of-fit against ground truth (experts' mapping of the items into the KCs)

References

[1] International Educational Data Mining Society. `http://educationaldatamining.org/`.

[2] Jiri Rihak and Radek Pelanek. Measuring Similarity of Educational Items Using Data on Learners' Performance. *Proceedings of EDM'17*, 2017.