Predictive performance of prevailing approaches to skills assessment techniques: Insights from real vs. synthetic data sets

Behzad Beheshti
Polytechnique Montreal
behzad.beheshti@polymtl.ca

Michel C. Desmarais
Polytechnique Montreal
michel.desmarais@polymtl.ca

1. INTRODUCTION

A number of skills assessment models have recently emerged, and others have been around for decades. Their predictive performance have often been compared on a pairwise basis, but few studies have taken a comprehensive approach to compare them on a common basis. In this study, we apply a methodology that adopts both synthetic and real data for the purpose of this comparison. Synthetic data is generated from the underlying model of the different skills assessment techniques. The results show wide differences of performances between the skills assessment methods over synthetic data sets. They create a kind of "signature" for each specific data. If this signature is unique, it might reveal the latent structure of the skills. We discuss the potential benefits and the limits of the methodological approach that consists in exploring the performance of skills assessment methods based on the comparison of real and synthetic data.

2. SKILLS ASSESSMENT COMPARISON

This work compares a number of skills assessment techniques over real and synthetic data: the well known single skill Item Response Theory (IRT), the DINA and DINO models that rely on slip and guess factors [3], matrix factorization approaches based on conjunctive and disjunctive Q-matrices [1], and the POKS approach based on the Knowledge Space theory Falmagne [2], which does not directly attempt to model underlying skills but instead rely on observable items only. For baseline comparison, the expected value and majority class performances are also reported. The performance comparison relies solely on each approach's ability to predict item outcome, not on the skills assessment directly which is not possible with real data.

The synthetic data sets are generated according to the each technique's underlying model. We naturally expect to obtain the highest performance when the technique and the synthetic data underlying model are aligned, but of particular interest is the relative performance of the techniques over the different types of synthetic data. An interesting hypothesis is whether the performance patterns of the different techniques over a synthetic data set is unique and the extent to which it represents a "signature" of the underlying skills model ground truth of a data set.

3. RESULTS

Figure 1 and 2 show the performance of each technique over the synthetic and real data sets. We report the predictive accuracy of each method, along with the average success rate of the data set as a comparison point (last column), which constitutes the performance of predicting the majority class (or $(1-\text{perf})$ when perf is below 0.5). An error bar of 1 standard deviation is reported and computed over the 10 random sampling simulation runs and provides an idea of the variability of the results. Also reported is the performance of random data with a 0.75 average success rate.

4. DISCUSSION

The results do show wide differences in the performance of the techniques for different synthetic data sets. For real data sets, the differences are smaller, though still significant.

An interesting finding is that the relative performance of the different skills modeling approaches create signatures over data sets. According to these signatures, the Vomlel real data set is closest to the linear compensatory simulated data set. As could be expected, random data does have a unique signature of its own: all methods converge towards the score of the majority class. The EPCE data set is close to this signature.

Another finding is the small relative differences between the techniques for the Fraction 2/3 data set compared to the other Fraction data sets and the Vomlel data set. This data has the peculiarity that the items were chosen based on a small number of single skill per item.

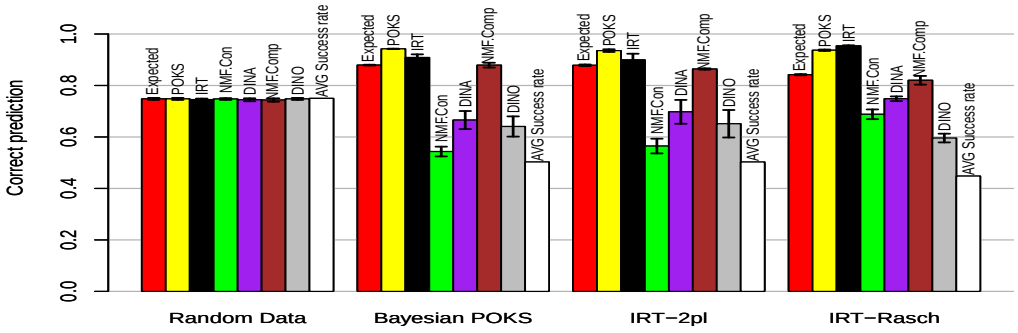
Future work will aim to establish if the findings generalize and the extent to which performance patterns generalize, but the approach of comparing these patterns of multiple models and techniques over real and synthetic data sets appears promising.

References

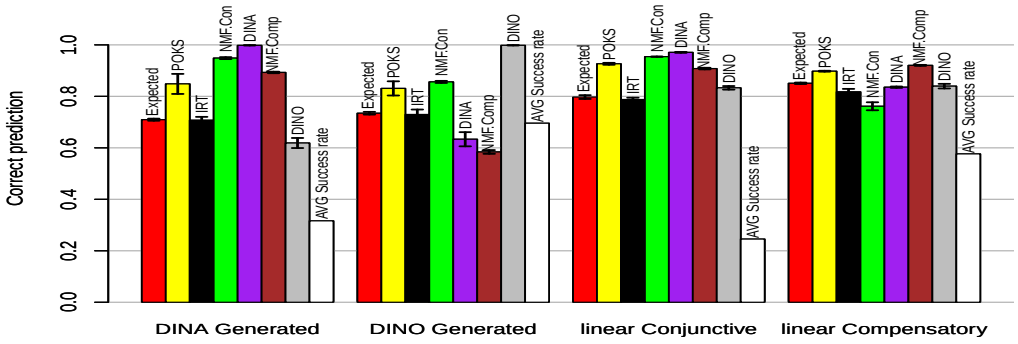
[1] M. Desmarais. Conditions for effectively deriving a Q-matrix from data with non-negative matrix factorization. In *4th International Conference on Educational Data Mining, EDM*, pages 41–50, 2011.

[2] M. C. Desmarais, P. Meshkinfam, and M. Gagnon. Learned student models with item to item knowledge structures. *User Modeling and User-Adapted Interaction*, 16(5):403–434, 2006.

[3] B. W. Junker and K. Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272, 2001.

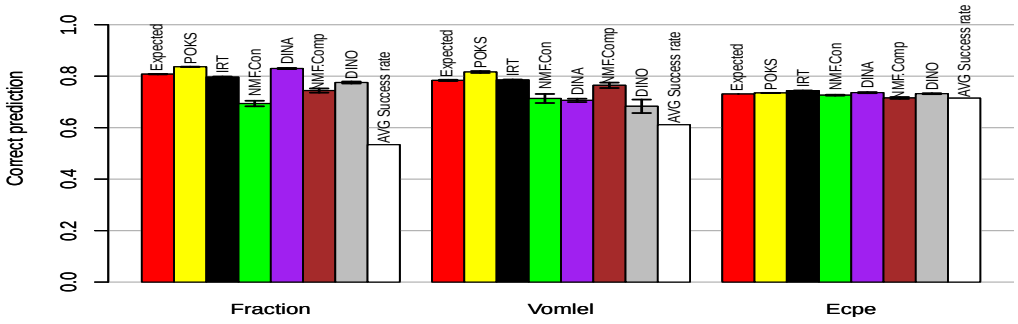


(a) Non Q-matrix based

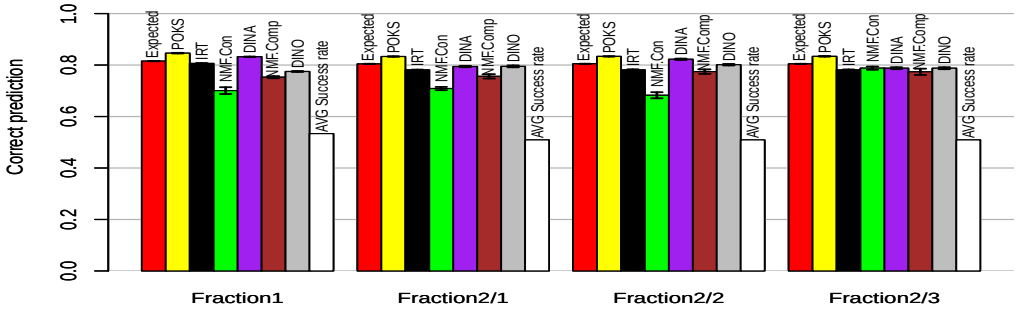


(b) Q-matrix based

Figure 1: Item outcome prediction accuracy results of **synthetic data sets**



(a) Independent data sets



(b) Subsets of the Fraction data set

Figure 2: Item outcome prediction accuracy results of **real data sets**