Evaluating Learning Factors Analysis

Michael Lipschultz, Diane Litman, Pamela Jordan, and Sandra Katz

Learning Research & Development Center, University of Pittsburgh
{lipschultz,litman}@cs.pitt.edu, {pjordan,katz}@pitt.com

Abstract. Learning Factors Analysis (LFA), a form of student modeling, is used to predict whether a student can correctly answer a tutor question. Existing evaluations of LFA rely on metrics like the cross-validated root mean squared error (RMSE). However, the LFA output can be a binary classification (the student will be correct or not), so we can use classification metrics, such as precision and recall, to evaluate LFA models. In this paper, we show that this finer-grained analysis can lead to different conclusions than relying on only RMSE.

Keywords: evaluation, enhancing learning outcomes, e-learning

1 Introduction

Computer-based tutoring systems often decide instructional acts based on whether a student is predicted to know certain knowledge components (KCs). One such approach to modeling student knowledge is through Learning Factors Analysis (LFA). While training, it weights problem solving and student proficiency features to make predictions about student correctness. Examples of LFA include Additive Factors Models, which use student turn counts [1], and Performance Factors Models, which count correct and incorrect student turns separately [3].

Previous work with LFA models evaluate model performance using summary metrics, such as the cross-validation root-mean-square error (RMSE) [2], which measures the average difference between a model's prediction and the actual value. It summarizes the overall performance of a model, but does not provide information on performance for a particular kind of outcome. For example, we are interested in predicting when a student will answer incorrectly since these instances are where the student may need help. In this paper, we split the numeric output of LFA models into two classes and evaluate model performance using class-level classification metrics. We show that this method of evaluation can lead to different conclusions than relying on RMSE alone.

2 LFA Analysis

Our data is from a previous study [4] using a typed natural-language physics tutoring system. In it, the student solved a work-energy problem, then discussed physics concepts involved in the solution. Each of the 64 students solved and

$$\ln\frac{p_{ij}}{1-p_{ij}} = \theta_i + \sum_k \beta_k KC_{jk} + \sum_k KC_{jk}(\mu_k \ln(C_{ik}) + \rho_k \ln(I_{ik}))$$

Fig. 1. Performance Factors Model. i represents student i. j represents turn j. k represents KC k. p_{ij} is the probability that student$_i$ would be correct on turn$_j$. θ_i is the coefficient for proficiency of student$_i$. β_k is coefficient for difficulty of KC k. μ_k and ρ_k are coefficients representing how useful preceding correctness counts (C_{ij}) and incorrectness counts (I_{ij}) are for predicting current turn correctness.

		Correct Class			*Incorrect* Class			Unweighted Average		
Model	RMSE	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
AFM	0.478	0.657	0.909	0.752	0.434	0.132	0.180	0.545	0.520	0.466
PFM	0.471	0.656	0.913	0.756	0.407	0.118	0.162	0.532	0.516	0.459

Table 1. Results for AFM and PFM. All metrics are averaged over 64 folds. Smaller RMSE is better. Larger precision, recall, and F1 are better.

discussed 7 physics problems. There were 4458 problem-solving turns, with 2811 tagged *correct* and 1647 *incorrect*. Of the 2756 post-problem discussion turns, 1883 were *correct* and 873 were *incorrect*. The turns were tagged for eight KCs.

In order for our student model to predict whether a student's response to a tutor question will be correct, we extracted four sets of features, similar to previous LFA work using natural language tutoring [2]. The first set is whether each of the eight KCs occurs in the current student turn: $KC_k = 1$ if KC_k occurs in the current turn, otherwise 0. The second set (N_k) counts the preceding post-problem discussion (PPD) student turns where KC_k occurs. The third set (C_k) counts the correct preceding PPD student turns where KC_k occurs. The fourth set (I_k) counts the incorrect preceding PPD student turns where KC_k occurs.

The feature to predict is student correctness on the current turn. This is a binary feature, with the student being either correct or incorrect. Since LFA makes numeric predictions, we must convert correctness into a numeric value. Following previous work [2], we convert *correct* into 1 and *incorrect* into 0.

In the LFA literature, there are two modeling techniques that use some of the feature sets above. Additive Factors Model (AFM) uses the KC_k and N_k feature sets. Performance Factors Model (PFM) uses KC_k, C_k, and I_k feature sets. To examine whether AFM or PFM perform better when predicting correctness in our data, we performed leave-one-student-out cross-validation for both AFM and PFM. The PFM formula can be seen in Figure 1.

Table 1 lists the results for AFM and PFM. Since the correctness predictions (p_{ij}) of the LFA models are real values between 0 and 1, we need to convert it into binary values to evaluate the models using the classification metrics. We use as the split point the middle of the range, 0.5, with all values less than 0.5 being classified as *incorrect* and the rest classified as *correct*.

In the table, we see that PFM has the better RMSE value. This is consistent with the literature, which has found that including features examining prior correctness improves performance on RMSE [2]. However, when we look at the classification metrics, we see instances where AFM performs better. For both

the minority class (*incorrect*) and for the unweighted average[1], we see that AFM outperforms PFM across all three metrics. For the *correct* class, PFM performs better on recall and F1, but AFM performs slightly better on precision. Overall, AFM performs better on seven of the nine classification metrics, suggesting that AFM is the model to use, particularly since we are interested in predicting when the student will answer incorrectly. However, had we not examined the classification metrics and only looked at RMSE, we would have chosen PFM.

3 Discussion & Future Work

We examined the performance of two common LFA methods, AFM and PFM, on classification metrics in addition to RMSE. Consistent with the LFA literature, we find that PFM has the better RMSE value, suggesting PFM is the better model. However, when we examine the classification metrics, we find that AFM typically outperforms PFM. From this finding, we believe that it is important to examine classification metrics when choosing an LFA model for predicting student correctness on a dialogue turn.

In this work, we split the correctness predictions into binary values using 0.5 as the split point, but any value between 0 and 1 can be used, depending on the application. In future work, we will use ROC curves and area under precision-recall curves to identify the best thresholds and LFA models. Otherwise, when we convert the numeric output of an LFA model into a binary classification, uncertainty information would not be well-utilized.

Finally, in future work we will also examine other datasets to determine whether our evaluation findings generalize and we will use classification metrics to compare other student modeling techniques to LFA.

Acknowledgments. The research was supported by IES, U.S. DoE, Grant R305A100163 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent views of the Institute or the DoE.

References

1. Cen, H., Koedinger, K.R., Junker, B: Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. ITS. 164–175 (2006)
2. Chi, M., Koedinger, K., Gordon, G., Jordan, P., VanLehn, K.: Instructional Factors Analysis: A Cognitive Model for Multiple Instructional Interventions. EDM. (2011)
3. Pavlik, P.I., Cen, H., Koedinger, K.R.: Performance factors analysis – a new alternative to knowledge tracing. AIED. 531–538 (2009)
4. Chi, M., VanLehn, K., Litman, D., Jordan, P.: An Evaluation of Pedagogical Tutorial Tactics for a Natural Language Tutoring System: A Reinforcement Learning Approach. IJAIED. 83–113 (2011)

[1] We report unweighted averages because we are interested in the minority class; weighted averages give preference to the majority class.

Evaluating Learning Factors Analysis

Michael Lipschultz, Diane Litman, Pamela Jordan, and Sandra Katz

Motivation

Tutor: If the man pushed on the cart with twice the force, how would the work change?
Student: double *[knowledge component (KC): DefWork, correct: True]*
Tutor: Right, the work would double! Now, …

>> Which question is student more likely to benefit from?
- how much work is done if the man pushed only half the distance? *[KC: DefWork]*
- what is the work done by the crate on the man? *[KCs: DefWork, Newtons3Law]*

>> Use good student model to predict which question student is more likely to get wrong
>> How do we define "good" student model?
Standard metrics can be misleading!

Data

64 students
7 work-energy physics problems
Typed Dialogues with computer tutor
- Solve problem
  - 2,811 student turns *correct*
  - 1,647 *incorrect*
- Discuss concepts used in solution
  - 1,883 *correct*
  - 873 *incorrect*

Tagged for 8 KCs

Learning Factors Analysis (LFA): Additive Factors & Performance Factors

Logit regression models predicting student correctness given:
- Student proficiency
- KCs relevant for current turn

… and model-specific features

Features calculated on only post-problem discussion turns

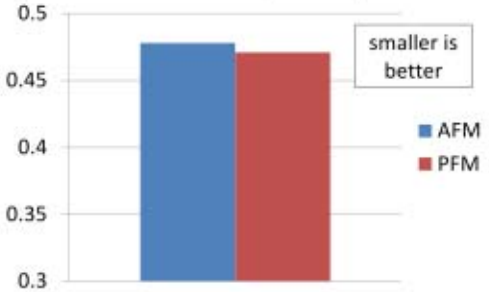
Additive Factors Model (AFM):
- # Prior turns for relevant KCs

Performance Factors Model (PFM):
- # Prior correct turns for relevant KCs
- # Prior incorrect turns for relevant KCs

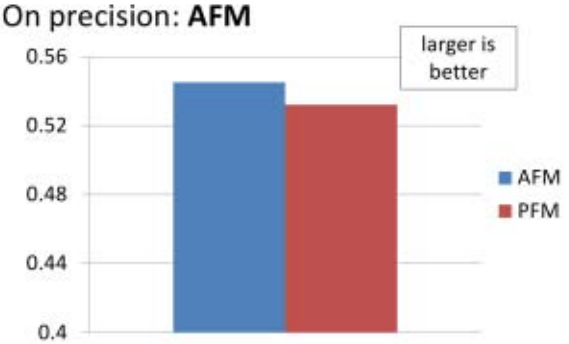
AFM vs. PFM: Which is better?

Leave-one-student-out cross-validation on all turns

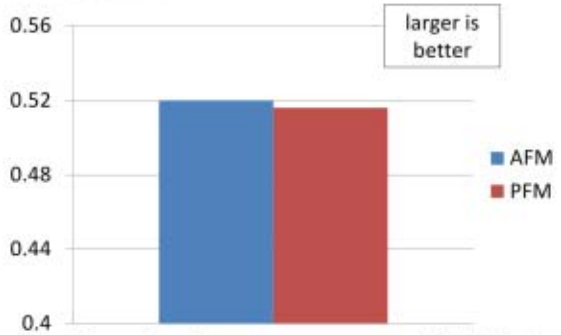
By standard metric (RMSE): **PFM**



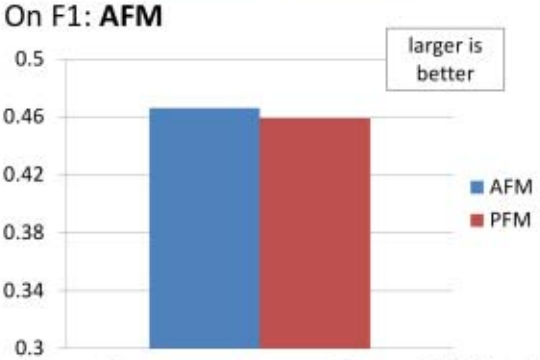
On precision: **AFM**



On recall: **AFM**



On F1: **AFM**



Standard measure says PFM is best, but other measures show AFM to be better.

Only unweighted average is reported here. Results similar on *incorrect* class.
See paper for performance on *correct* and *incorrect* classes.

Conclusion

Important to examine precision, recall, F1

Future Work

Utilize uncertainty information
Results generalize to other data sets?
LFA vs. Other Modeling Techniques