Comparisons of BKT, RNN and LSTM for Predicting Student Learning Gains

Chen Lin and Min Chi

The Department of Computer Science,
North Carolina State University, USA
clin12, mchi@ncsu.edu

Abstract. The objective of this study is to develop effective computational models that can predict student learning gains, preferably as early as possible. We compared a series of Bayesian Knowledge Tracing (BKT) models against vanilla RNNs and Long Short Term Memory (LSTM) based models. Our results showed that the LSTM-based model achieved the highest accuracy and the RNN based model have the highest F1-measure. Interestingly, we found that RNN can achieve a reasonably accurate prediction of student final learning gains using only the first 40% of the entire training sequence; using the first 70% of the sequence would produce a result comparable to using the entire sequence.

Keywords: LSTM, RNN, BKT, Learning Gain Prediction

1 Introduction

A number of studies have shown that the effectiveness of any learning environment varies greatly based on individual differences such as motivation, aptitude, and incoming competence, etc [1]. Thus, it is essentially important to track whether a student has embarked upon a unprofitable learning experience and to identify such an individual as early as possible so adaptive remediation can be offered. However, it is often very hard to do so because many factors may impact whether a student would learn, yet those factors are not fully understood. As a result, much student modeling research focused on modeling student knowledge competence level over time such as Bayesian Knowledge Tracing (BKT) [2]. Nonetheless, a student with high knowledge level does not mean that the student benefited from the learning environment; the student may already have high competence before he/she starts use the system.

To fully honor the promise of the learning environment, the main objective in this study is to predict student learning gains on Intelligent Tutoring Systems (ITSs). As far as we know, little research has done in this direction, probably because predicting learning gains is much more challenging than predicting student student knowledge level. While it is often reasonable to assume that a student who have done well so far would continue to do well in the final exam, it is unclear what factors impact student learning gains. Previous research used $Learning\ Gain\ (LG) = post - pre$ or $NLG = \frac{(post-pre)}{(1-pre)}$ (pre and post are defined as a students' pretest score before training and posttest score after training respectively; 1 is the maximum score one can get) to measure learning gain. Both LG and NLG are biased against students with High pretest scores in that it is often harder for them to obtain the same LG and NLG scores as their peers with Low pretest scores. Therefore, we proposed Quantized Learning Gain (QLG) by dividing students into "low", "medium" and "high" groups using 33rd and 66th percentile based on their pretest and posttest scores. *Low* QLG students are defined as those who either down-graded its group from pre- to posttest, or remained in "low" or "medium" groups even though their performance can be improved; the rest are labeled as "High" QLG.

2 Methods & Training Data

Proposed Methods include the classic BKT models, Recurrent Neural Network (RNN) [4] and Long Short Term Memory (LSTM) [5]. BKT [2] is the classic approach for student modeling. It leverages student performance (i.e., correct, incorrect) over time to updates estimations of student knowledge level. Intervention-BKT is a variation of BKT by incorporating instructional interventions within its framework and it has been shown to outperform conventional BKT in various prediction tasks [3]. For BKT family variation models, we have tried the basic BKT model and the mixed model combining Intervention-BKT and BKT. For both models, we used either performance, response time, or a combination of both as observations. Thus we have a total of six BKT models.

Compared to BKT-based models, RNN exhibits greater flexibility: it allows multivariate inputs and does not require any explicit encoding of domain concepts, thus requiring near-to-zero human expert involvement. LSTM is a special type of RNN that contains a system of gating units that controls the flow of information. LSTM has been shown to learn long-term dependencies more easily than vanilla RNN [6]. It has been shown that both RNN and LSTM out-performed the conventional BKT model [7]. For RNN and LSTM, the system-student interaction tuples were converted into a sequence of input vectors. The system learns and passes information across many steps to predict QLG at the final step. A sequential *target replication* (TR) technique inspired by [8] was implemented in our study, where the final target was copied at each sequence step, providing a local error signal. Models with TR were named RNN-TR and LSTM-TR respectively. For both RNN and LSTM, Different combinations of layers (1 or 2) and nodes (50, 100, 150, or 200) were tested. As LSTM contains more parameters, a dropout rate of 0.2 was applied between a LSTM layer and a final dense layer. All of these experiments were conducted using Kera's implementation and trained to 100 epochs with a 100 mini-batch size.

Training Dataset was collected from training 524 students on a probability tutor called Pyrenees. The training was assigned as their final homework in the undergraduate Discrete Mathematics course at the Department of Computer Science at North Carolina State University from 2014-2016. Students were required to complete 4 phases: 1) pre-training, 2) pretest, 3) training and 4) post-test. All students received the same 12 training problems in the same order. Pretests and

Table 1. Prediction Results for All 11 Models

Model	Accuracy	Precision	Recall	F1-measure
1 Majority	0.528	0.587	0.595	0.591
2 Best BKT Family	0.642	0.648	0.821	0.724
3 RNN	0.669	0.668	0.841	0.744
4 RNN-TR	0.667	0.658	**0.874***	**0.750***
5 LSTM	**0.670***	0.670	0.837	0.744
6 LSTM-TR	0.648	**0.678***	0.734	0.705

best model marked in **bold** and *

post-tests were graded in a double-blind manner by a single experienced grader. The scores were normalized using a range of [0,1].

3 Result

Comparisons among BKT, RNN and LSTM: Two RNNs and two LSTMs were compared against a series of BKT family models and a majority baseline on predicting QLG. Table 1 shows our 10-fold cross validation accuracy, precision, recall and F1-measure results. Note that for our task, detecting *Low* QLP is more important. Our results showed that all the BKT-based, RNN, and LSTM models outperformed the majority baselines. Given the limited space, only the best model among BKT family models were reported as shown in row 2 in Table 1. Generally speaking, the RNN, RNN-TR, LSTM models outperformed the best BKT-based models while LSTM-TR has a worse recall and F1-measure than the BKT family models. Among RNN family models, LSTM achieves the highest accuracy; LSTM-TR has the highest precision rate, and RNN-TR achieves the best recall and F1-measure. Without the replicating target technique, both RNN and LSTM have similar results; whereas using target replication technique improves the recall and F1-measure for RNN, but not for LSTM.

Early Detection of Low QLG Given that RNN-TR achieves the highest Recall and F1-measure, we investigated its performance in early detection. In Figure 1, the line with triangular points represents the F1-measure and the line with circular points represents accuracy. They are both measured at every 10% increment of the sequence length. From 10% to 40%, both accuracy (0.52 to 0.64) and the F1-measure (0.65 to 0.72) increase significantly; from 40% to 70%, the increase becomes moderate (0.64 to 0.66 for accuracy; 0.72 to 0.74 for F1-measure); from 70% to the remainder of the sequence, the increase is only slight (0.66 to 0.67 for accuracy; 0.74 to 0.75 for F1-measure). The results indicate that a good prediction of QLG can be achieved by using the first 40% of the entire sequence and that using the first 70% of the entire sequence is as good as using the entire sequence.

4 Discussion

This study compared a majority baseline model, a series of BKT family models and four RNN/LSTM family models to predict students' QLG. Our results sug-

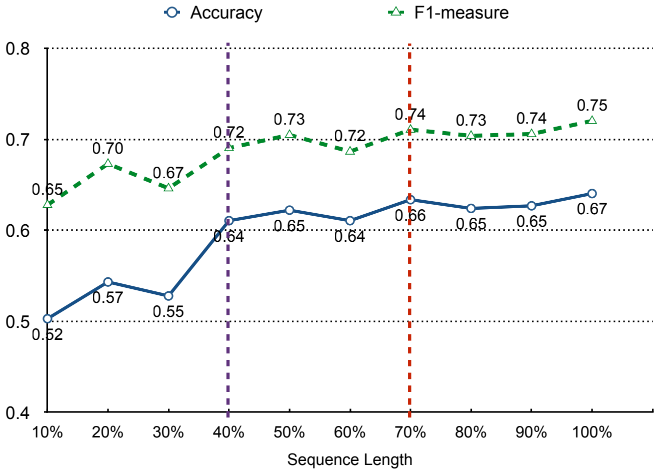

Fig. 1. Accuracy and F1-measure for RNN Prediction Using Partial Sequence

gest that both all models outperformed Majority baseline and LSTM achieved the highest accuracy whereas RNN-TR, a RNN model with target replicate technique, achieved the highest recall and F1-measure. Furthermore, the performance of using RNN-TR to perform early detection of students who may have a *low* QLG was explored. We found that the model can achieve reasonably good result when using only the first 40% of the entire student log sequence.

References

1. Merrill, D. C., et al (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. The Journal of the Learning Sciences, 2(3), 277-305.
2. Corbett, A. T., Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4(4), 253-278.
3. Lin, C., Chi, M. (2016, June). Intervention-BKT: Incorporating Instructional Interventions into Bayesian Knowledge Tracing. In International Conference on Intelligent Tutoring Systems (pp. 208-218). Springer International Publishing.
4. Mikolov, T., Karafit, M., Burget, L., Cernock, J., Khudanpur, S. (2010, September). Recurrent neural network based language model. In Interspeech (Vol. 2, p. 3).
5. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J.,Sohl-Dickstein, J. (2015). Deep knowledge tracing. In Advances in Neural Information Processing Systems (pp. 505-513).
6. LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
7. Piech, C., et al (2015). Deep knowledge tracing. In Advances in Neural Information Processing Systems (pp. 505-513).
8. Lipton, Z. C., et al (2015). Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677.