Predicting Student Performance: The Case of Combining Knowledge Tracing and Collaborative Filtering

Solmaz Abdi
The University of Queensland
solmaz.abdi@uq.edu.au

Hassan Khosravi
The University of Queensland
h.khosravi@uq.edu.au

Shazia Sadiq
The University of Queensland
shazia@itee.uq.edu.au

ABSTRACT

In the past few years, many competing learning models have been proposed for improving the accuracy of predicting student performance (PSP). A well-studied subclass of algorithms focused on PSP uses temporal models to determine the knowledge state of users. Bayesian Knowledge Tracing (BKT), as one of the leading models in this subclass, uses Hidden Markov Models to capture the student knowledge states. An emerging new subclass of algorithms focused on PSP uses collaborative filtering, which is used primarily by recommender systems. Matrix Factorization (MF), a leading model in this subclass, can be presented as a rating prediction problem where students, tasks, and performance information are mapped to users, items and ratings, respectively. BKT and MF complement each other's strengths and limitations quite effectively. In particular, BKT relies on four skill-specific parameters for learning the sequential behavior of learners on each concept, but it does not capture the similarities among users and items. In contrast, MF uses latent factors to exploit the similarities among users and items from learner-item performance, but disregards any temporal effect in modeling student learning. In this paper, we aim to investigate the effect of combining variations of BKT and MF using a proposed algorithm that exploits the power of MF in modeling the implicit similarities among learners and items while using the explicit parametrization of BKT towards improving PSP. Our results on four benchmark educational datasets show that our approach outperforms the base classes as well as traditional techniques such as linear regression, logistic regression and Neural Networks for combining BKT and MF.

1. INTRODUCTION

Heavily studied in the community of educational data mining (EDM), the problem of predicting student performance (PSP) uses observations from students' behavior to find a model that predicts their future performance on unseen learning tasks [3].

Temporal models have been used extensively for PSP and determining the knowledge state of users. They rely on the sequential behavior of learners to model their learning. In these models, the students' performance on the next task is predicted using their performance on their prior test items [11] and a Q-matrix [1], which is a binary matrix that shows the relationship between test items and underlying concepts. One of the leading temporal models for PSP is Bayesian Knowledge Tracing (BKT) [3]. BKT uses Hidden Markov Models for capturing students' knowledge states as a set of binary variables. While BKT has received significant attention and improvement since it was first proposed, it is unable to capture similarities among learners or items, which has shown to be an important aspect in improving PSP [14].

Applying collaborative filtering (CF) techniques is another promising approach for PSP. One of the most successful collaborative filtering techniques is the factorization method based on the matrix or tensor decomposition [2]. As shown by [8], applying matrix factorization (MF) can lead to improved prediction results in PSP compared to traditional PSP methods. MF predicts student performance by extracting similarities among learners and items from the learner-item performance data in form of latent factors. MF creates two matrices with latent factors for each of learners and items, so there is no need to explicitly encode Q-matrix or other parameters such as Slip and Guess [14]. In addition, MF is very effective in dealing with insufficient data as it effectively captures and uses the similarities among learners and items [14]. The main limitation of MF is its lack of temporal effect as MF discards any temporal information and learns the typical performance of students at one time. Tensor Factorization overcomes this limitation; however, the running time of tensor factorization is significantly longer than MF [13], so in practice it is not used as frequently.

In this paper, we introduce a new approach called MBKT that combines BKT and MF for the task of PSP. Traditional models of combining where the predictions results of individual algorithms are stacked, would require MF to learn an implicit Q-matrix and latent factors incorporating Slip and Guess from Scratch. To fully exploit the advantages of combining BKT and MF, MBKT first utilizes BKT to capture the temporal effects of the student model using an explicit Q-matrix and parameters referring to Slip and Guess. This information is then passed on to MF, which enables the latent factor of MF to be tuned for capturing the similarities between students and items.

Our results on four benchmark datasets obtained from the DataShop platform [9] indicate that using MBKT for combining various variations of BKT and MF for PSP outperforms the base models. We also show that MBKT outperforms traditional methods of combining the results of BKT and MF using linear regression, logistic regression and Neural Networks.

2. RELATED WORK

The problem of combining different algorithms for improvement in PSP has been well studied, with contradicting results. To evaluate the effect of ensemble techniques in Intelligent Tutoring System (ITS), Baker et al. [4] selected nine different PSP individual algorithms and combined them using logistic and linear regression on a genetic dataset. Their experimental results showed that the accuracy of ensembling is mixed and slightly different from the individual algorithms. They argued that there may be three explanations for this lack of improvement: (1) use of only simple models of ensembling like linear and logistic regression, (2) use of small datasets with a limited number of learner interactions, and use of similar ensemble techniques on learning models with slight differences. Pardos et al. [10] reported that ensembling on large enough datasets will lead to promising improvements even with similar base models. However, in practice, student models rely on small datasets for training, so the results of ensemble techniques on large datasets cannot be applied directly to ITS. More recently, [12] used a knowledge graph representation to identify feasible activity scopes, which were combined to predict student performance on a learning objective in an ensemble.

Despite development of various ensembling algorithms on PSP, to the best of our knowledge, collaborative filtering algorithms have not been used in conjunction with knowledge tracing algorithms in the previous studies. Given that these two complement each other on many fronts, we attempt to extend the work of previous studies by primarily investigating the impact of combining MF as a leading collaborative filtering algorithms with knowledge tracing for PSP.

3. COMBINING KNOWLEDGE TRACING AND MATRIX FACTORIZATION

As mentioned in the previous sections, the characteristics of BKT and MF complement each other quite well. BKT utilizes the temporal behavior of learners to model their learning, while MF does so by capturing the similarities among learners and items. In addition, BKT uses an explicit Q-matrix to find the parameters related to learners including their initial knowledge of skills, the mastery probability of skills and Slip and Guess parameters. In contrast, MF uses latent factors to implicitly learn a Q-matrix and the mentioned learner-related parameters. In this paper, we propose a new model called MBKT for combining BKT [3] and MF [14] for PSP that takes advantage of how these models complement each other. We also considered two other variations of BKT as described in [6]. The first variation, BKT-CGS (Contextual Guess and Slip) model, is a variation in which Guess and Slip properties are no longer learned per skill but rather averaged across all skills and actions. The second variation, BKT-PPS (Prior Per Student) assumes a personalized prior knowledge per student. In our experiments, we used a simplified version of this model that divides students to high-performance and low-performance groups as proposed by [6]. Using MBKT, the predicted performance of leaner u on item i is predicted as follows:

In the first step, the BKT model is utilized to predict student performance using the following formula

$$O^{BKT}_{N\times M} = BKT(train\_set),$$

where o^{BKT}_{ui} presents the computed probability of the BKT model on user u answering item i correctly based on the last opportunity of u on the topic related to i.

In the second step, the error of BKT predictions for the learner-item performance is computed as follows

$$E^{BKT}_{N\times M} = O^{train}_{N\times M} - O^{BKT}_{N\times M},$$

where o^{train}_{ui} is 1 if user u has answered question i correctly in their final attempt, 0 if user u has answered question i incorrectly and Null otherwise. e^{BKT}_{ui} is the computed error of the BKT model for user u on question i.

In the third step, the error of BKT predictions for the learner-item performance is passed on to MF as input to predict the BKT prediction error using the following formula

$$O^{MF}_{N\times M} = MF(E^{BKT}_{N\times M}),$$

where o^{MF}_{ui} presents the approximated error of the BKT model on the final opportunity of user u on answering question i.

Finally, the outcome of MBKT is computed by summing the BKT predictions and predicted error of MF for BKT using

$$O^{MBKT}_{N\times M} = O^{BKT}_{N\times M} + O^{MF}_{N\times M},$$

where o^{MBKT}_{ui} represents the predicted performance of user u on question i, which is computed by MBKT.

Discussion. Using the traditional models of combining where the prediction results of individual algorithms are stacked, MF needs to learn the Q-matrix and latent factors from scratch using a random initialization. This makes the combination unlikely to fully exploit the advantage of combining BKT as a temporal model and MF as a model to draw out the similarities among learners and items. In MBKT, instead of directly stacking the prediction results of BKT and MF, BKT is utilized as the underlying algorithm to predict student performance. Then the prediction error of BKT is passed to MF as input to learn the BKT error. Insinuating the outcome of the BKT model in the input of MF enables MBKT to benefit from BKT's explicit parameterization of the learners and items including the initial knowledge, the mastery probability of skills and Slip and Guess concepts. This, in turn, would enable the latent factors of MF to further focus on modeling similarities among learners instead of trying to incorporate those parameters.

4. EXPERIMENTS

In this paper, we have discussed the benefits of combining knowledge tracing and collaborative filtering algorithms for PSP using MBKT. In this section, we aim to investigate whether use of MBKT leads to improved PSP. Our evaluation has been guided by the following two research questions.

- RQ1: Does MBKT improve the performance of PSP compared to the base models?
- RQ2: Does MBKT improve the performance of PSP compared to traditional techniques of stacking the results of BKT and MF?

For the experiments, we utilize LearnSphere [9] to find the parameters of each BKT variation using 10 fold cross-validation with Baum-Welch solver. To find the latent factors related to each MF variation, we use MyMediaLite library [5] with again 10 fold cross-validation.

4.1 Dataset

We use four data sets that are commonly used for PSP from DataShop [9] in our evaluation. The total number of interactions and students of each dataset is described in table 1.

Table 1: DataSets

Data Set	#transactions	#students
Geometry Area	6,778	59
Intelligent Writing Tutor	6,625	120
Writing 1	12,568	31
Writing 2	11,347	54

These are the results of learners' interactions with the tutoring system. As learners engage in the system, all interactions such as their success or failure, time spent on each step, etc are recorded. In these experiments, the unique interaction between learners and system is the step, which belongs to the hierarchy of *unit*, *section* and *problem*. *KC* defines different knowledge components for each step in the hierarchy and *Opportunity* determines the total number of times that a leaner has had on the *KC* related to the step. In these datasets *FirstAttempt* is considered as the outcome of the interaction: *correct* means success and *incorrect* and *hint* show failure in that interaction.

4.2 Methods and Evaluation Metric

In our experiments, standard BKT (BKT) [3], Contextualized Guess and Slip BKT (BKT-CGS) [10], Prior Per Student BKT (BKT-PPS) [10], Standard Matrix Factorization (MF) and Biased Matrix Factorization (BMF) as described in [14] are used as the base methods.

The BKT and MF variations are combined using logistic regression (LogReg), linear regression (LinReg), Neural Networks (NN) and MBKT.

Evaluation Metric. As commonly used in evaluating the PSP algorithms, Root Mean Squared Error (RMSE) is utilized to measure the error as follows:

$$RMSE = \sqrt{\frac{1}{|D|^{\mathsf{test}}} \sum_{(u,i)\in D^{\mathsf{test}}} (o_{ui}^{\mathsf{test}} - o_{ui}^{\mathsf{prediced}})^2},$$

where $o_{ui}^{\mathsf{prediced}}$ is the predicted probability, o_{ui}^{test} is the real output of the instance and D^{test} is the total number of instances.

4.3 Results

Table 2 compares the RMSE of the model fit statistics related to each model for the task of PSP. In this table, Geo, IntW, HW1, and HW2 refer to Geometry Area, Intelligent Writing, Hand Writing 1, and Hand Writing 2 datasets respectively. Based on the experimental results for all datasets, there is no superiority among different BKT variations. Among the two MF variations, BMF significantly outperforms MF both as an individual algorithm and in combination with the BKT variations. For instance, the average RMSE for BMF and MF as an individual algorithm on all datasets is 0.412 and 0.427 respectively. A similar difference is observed in the combinational models. So, for the rest of discussions, we only concentrate on BMF as the collaborative filtering algorithm.

Table 2: RMSE of different learning models

Methods		**Geo**	**IntW**	**HW1**	**HW2**
BKT		0.422	0.438	0.431	0.408
BKTPPS		0.421	0.422	0.412	0.392
BKTCGS		0.419	0.438	0.431	0.407
MF		0.427	0.453	0.433	0.396
BMF		0.418	0.433	0.407	0.390
BKT -MF	LogReg	0.419	0.447	0.440	0.397
	LinReg	0.424	0.447	0.450	0.397
	NN	0.420	0.449	0.451	0.4
	MBKT	0.428	0.44	0.432	0.395
BKT -BMF	LogReg	0.417	0.421	0.406	0.391
	LinReg	0.415	0.422	0.406	0.390
	NN	0.420	0.421	0.406	0.391
	MBKT	0.411	0.418	**0.404**	0.387
BKTPPS -MF	LogReg	0.419	0.431	0.428	0.395
	LinReg	0.424	0.427	0.433	0.395
	NN	0.420	0.435	0.438	0.396
	MBKT	0.424	0.423	0.417	0.391
BKTPPS -BMF	LogReg	0.417	0.412	0.406	0.388
	LinReg	0.416	**0.411**	0.407	0.390
	NN	0.420	0.412	0.407	0.387
	MBKT	0.415	**0.411**	0.406	**0.386**
BKTCGS -MF	LogReg	0.420	0.447	0.44	0.397
	LinReg	0.430	0.447	0.405	0.397
	NN	0.421	0.449	0.452	0.4
	MBKT	0.422	0.435	0.431	0.394
BKTCGS -BMF	LogReg	0.416	0.421	0.406	0.391
	LinReg	0.415	0.422	0.406	0.399
	NN	0.421	0.421	0.406	0.391
	MBKT	**0.408**	0.418	0.405	0.387

RQ1. The results of cross-validated RMSE on all datasets indicates that for all combinations of BKT variations and BMF, MBKT achieves the best RMSE. As presented in Table 2, MBKT outperforms its base models by $\approx 10\%$. To evaluate the statistical significance of the improvements in predictions, Ttest is used. For each dataset, we applied Ttest on the RMSE of the best individual model and the best combination of BKT and MF using MBKT. For all four datasets, the difference between the results of the individual algorithms and MBKT was statistically significant with the computed p values smaller than 0.01.

RQ2. To answer this research question, we used the traditional stacking techniques including linear regression, logistic regression, and Neural Network to combine each of the BKT variations with BMF. Our experimental results on all datasets indicate that for each combination of BKT variations and BMF using MBKT and other stacking techniques, MBKT always outperforms the traditional stacking techniques, except for IntW where linear regression achieves the same RMSE as MBKT when combining BKTPPS and BMF. To evaluate the statistical significance of the mod-

els, we limited our comparisons to the combinations with the same base models. Our results on the four datasets indicate that with BKTPPS and BMF as the base models, MBKT and linear regression were not significantly different from one another for both Geometry Area and Intelligent Writing Tutor datasets. For the renaming 10 combinations, MBKT improve PSP with statistical significance ($p < 0.01$) compared to traditional stacking techniques.

In addition, MBKT always outperforms its base models and achieves $\approx 10\%$ improvement in the predictive power compared to its underlying BKT model. This is a significant improvement for a predicting model. In contrast, applying the traditional combining models do not always improve the predictions over those of the base models. For example, for Hand Writing 2, using logistic regression or Neural Network for combining BKT or BKT-CGS with BMF leads to poorer RMSE than BMF itself. This lack of success for traditional combining models reflects the same result is presented by [4].

5. CONCLUSION AND FUTURE WORK

In this paper, we investigated the effect of combing time-aware knowledge tracing algorithms with matrix factorization as a time-invariant collaborative filtering algorithm for PSP. Variations of Bayesian Knowledge Tracing (BKT) and Matrix Factorization (MF) were used for this task. These models complement each other's strengths and limitations quite effectively. BKT captures temporal changes in learners' behavior using an explicit Q-matrix and BKT parameters such as Slip and Guess. In contrast, MF captures similarities among learners using latent variables that implicitly encode a Q-matrix as well as learners' initial knowledge, skill mastery probability, Slip and Guess Parameters. We introduced an algorithm for combining MF and BKT, where instead of directly combining the prediction result of each individual algorithm, it first utilizes BKT as the underlying algorithm to predict student performance. It then passes the error, true values - predicted values, from BKT predictions as input to MF. Incorporating the outcome of the BKT model in the input of MF enables it to benefit from BKT's explicit parameterization including Slip and Guess concepts. This, in turn, would enable the latent factors of MF to further focus on modeling similarities among learners instead of trying to incorporate Slip and Guess parameters.

Our results on four benchmark datasets from the Datashop platform indicates that using MBKT for combining variations of BKT and MF leads to as much as 10% improvement over the base models for PSP on unseen datasets. In addition, MBKT generally provides statistically significant improvements over traditional techniques such as linear regression, logistic regression and Neural Networks for combining BKT and MF again, for PSP on unseen dataset.

There are several interesting directions to pursue in future work. Primarily, we are working on integrating our approach into an open-source, student facing adaptive learning environment called Recommendation in Personalized Peer Learning Environments (RiPPLE) [7]. Our goal is to use the proposed algorithm for predicting student performance, which in turn, is used for recommending personalized questions based on learners' current knowledge gaps.

6. REFERENCES

[1] Tiffany Barnes. The q-matrix method: Mining student response data for knowledge. In *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*, pages 1–8, 2005.

[2] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[3] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[4] Ryan SJ d Baker, Zachary A Pardos, Sujith M Gowda, Bahador B Nooraei, and Neil T Heffernan. Ensembling predictions of student knowledge within intelligent tutoring systems. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 13–24. Springer, 2011.

[5] Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. MyMediaLite: A free recommender system library. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011)*, 2011.

[6] SM Gowda, RSJD Baker, Z Pardos, and NT Heffernan. The sum is greater than the parts: ensembling student knowledge models in assistments.

[7] Hassan Khosravi. Recommendation in personalised peer-learning environments. *arXiv preprint arXiv:1712.03077*, 2017.

[8] Hassan Khosravi, Kendra Cooper, and Kirsty Kitto. Riple: Recommendation in peer-learning environments based on knowledge gaps and interests. *JEDM-Journal of Educational Data Mining*, 9(1):42–67, 2017.

[9] Kenneth R Koedinger, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. A data repository for the edm community: The pslc datashop. *Handbook of educational data mining*, 43, 2010.

[10] Zachary A Pardos, Sujith M Gowda, Ryan SJd Baker, and Neil T Heffernan. The sum is greater than the parts: ensembling models of student knowledge in educational software. *ACM SIGKDD explorations newsletter*, 13(2):37–44, 2012.

[11] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820. ACM, 2010.

[12] Martin Stapel, Zhilin Zheng, and Niels Pinkwart. An ensemble method to predict student performance in an online math learning environment. In *EDM*, pages 231–238, 2016.

[13] Nguyen Thai-Nghe, Lucas Drumond, and Tomás Horváth. Matrix and tensor factorization for predicting student performance.

[14] Nguyen Thai-Nghe, Lucas Drumond, Tomáš Horváth, Artus Krohn-Grimberghe, Alexandros Nanopoulos, and Lars Schmidt-Thieme. Factorization techniques for predicting student performance. *Educational recommender systems and technologies: Practices and challenges*, pages 129–153, 2011.