# A Transfer Learning approach for applying Matrix Factorization to small ITS datasets

Lydia Voß
Information Systems and Machine Learning Lab
Universitätsplatz 1, 31141
Hildesheim, Germany
lvoss@ismll.uni-hildesheim.de

Carlotta Schatten
Information Systems and Machine Learning Lab
Universitätsplatz 1, 31141
Hildesheim, Germany
schatten@ismll.uni-hildesheim.de

Claudia Mazziotti
Institute of Educational Research, Ruhr-University Bochum
Universitätsstraße 150, 44780
Bochum, Germany
claudia.mazziotti@rub.de

Lars Schmidt-Thieme
Information Systems and Machine Learning Lab (ISMLL)
Universitätsplatz 1, 31141
Hildesheim, Germany
schmidt-thieme@ismll.uni-hildesheim.de

## ABSTRACT

Machine Learning methods for Performance Prediction in Intelligent Tutoring Systems (ITS) have proven their efficacy; specific methods, e.g. Matrix Factorization (MF), however suffer from the lack of available information about new tasks or new students. In this paper we show how this problem could be solved by applying Transfer Learning (TL), i.e. combining similar but not equal datasets to train Machine Learning models. In our case we obtain promising results by combining data collected of German fractions' tasks (517 interactions, 88 students, 20 tasks) with their non-exact translation of a previously American US version (140 interactions, 14 students, 16 tasks). In order to do so we also analyze the performance of MF based predictors on smaller ITS' samples evaluating their usefulness.

## Keywords

Transfer Learning, Intelligent Tutoring Systems, Matrix Factorization, Vygotsky Policy Sequencer

## 1. INTRODUCTION

One of the main uses of Educational Data Mining in Intelligent Tutoring Systems (ITS) is Performance Prediction, which aims to ameliorate the student's model by understanding whether a student mastered a specific set of skills or not. Specific methods, e.g. Matrix Factorization (MF), suffer from the lack of available information about new ITS tasks or new students imposing challenging requirements on organizing trials. This happens because the algorithm is personalized, i.e. there is one model for each student interacting with the system and one for each task one can practice with. If no data are available for one task or for one student no prediction can be computed, this problem is called the cold-start problem. Moreover, first data for new tasks in ITS applications are obligatorily collected in a specific sequence, which is generally fixed or rule-based. As a consequence more interaction data are available for the first tasks in the sequence whereas just a few are available for the last ones making the prediction for specific tasks more challenging. In the FP7 iTalk2Learn project[1] we developed a domain independent sequencer [9] for one of our use cases based on MF Performance Prediction. One of this use cases is a German translation of Fraction Tutor (FT) a web-based Cognitive Tutor for fractions developed by Carnegie Mellon University[2]. Our data collection for the German version (88 students, 20 tasks, 517 interactions) represents, to the best of our knowledge, one of the smallest dataset used to train a MF based recommender for Performance Prediction in ITS. We also possess the data collected with the original US American version (16 tasks, 14 students and 140 interactions), which, according to common practice, should be discarded. In this paper we want to:

- Show, that we can use two different but comparable datasets (the German and English ones) to ameliorate Performance Prediction.
- Analyze in detail the effects of a small dataset on the performances of MF used as performance predictor.
- Propose a practical solution to the data collection to reduce data sparsity.

The paper is structured as follows. the second and third section describe the state of the art and the theory behind the performance predictors we used. In Sec. 4 the data collection, translation and preprocessing is described. In the Experiment Section we discuss the usefulness and measure

[1] www.iTalk2Learn.eu
[2] https://mathtutor.web.cmu.edu/

the performances of MF based predictors. Then we conclude the Section combining the English and German datasets to evaluate the feasibility of Transfer Learning approaches to exploit generally discarded data in ITS.

## 2. RELATED WORK

As we did not have access to the required skills information in [7, 8], MF and the VPS sequencer presented in [9] are used for Performance Prediction. MF has many applications, its most common use is for Recommender Systems and recently this concept was extended to Performance Prediction and to sequencing problems in ITS [10, 9], but all experiments were done with simulated students' interactions or offline experiments. In [7], we showed how the VPS sequencer could be integrated and worked in a large commercial ITS. A similar analysis on MF was done in [5] where Performance Prediction was tested on a small dense dataset (each student saw each task). The performance predictors were standard Collaborative Filtering techniques, where the best one performing resulted to be Biased Matrix Factorization (see Section 3.1 for more details). In this paper, we possess even less interactions. Not only the students did not interact with all available tasks, but sometimes they also solved less than three tasks. We try to solve this problem with Transfer Learning (TL)[3]. In contrast to classical Machine Learning methods, TL methods exploit the knowledge accumulated from auxiliary data to facilitate predictive modeling consisting of different but similar patterns in the current data [2]. Auxiliary data could mean additional information describing the state of the system and/or data collected with a second slightly modified version of the same system (e.g. using equal movies from different movie rating datasets and transfer the knowledge [4]). In this case correctly done transfer of knowledge, i.e. using similar but not equal datasets, is required and could improve the performance of predictors in classification and regression tasks ([4]) by considering previously unused data. This approach becomes particularly helpful when recollection is expensive or impossible. However TL was never applied to ITS data. Consequently, in Sec. 5.3 we evaluate the feasibility of applying TL to our use case to get a better Performance Prediction.

## 3. MATRIX FACTORIZATION BASED PREDICTORS

We use MF to predict the students performance. The matrix $Y \in \mathbb{R}^{S \times T}$ can be seen as an incomplete table of $T$ tasks and $S$ students. This matrix is used to train the system. MF is the approximation of this incomplete matrix by decomposing it in two smaller matrices $W \in \mathbb{R}^{S \times K}$ and $H \in \mathbb{R}^{T \times K}$. The elements of the two matrices are called *latent features* and are learned with gradient descend.

Using the available entries (e.g. the score recorded from previous tasks) the missing entries can be computed by means of very fast optimization algorithms. In our experiments we use MF and a simple variation of MF, the Biased Matrix Factorization (BMF) which uses three additional variables: the global average performance $\mu$, the student (user) bias $b_s$ and the task (item) bias $b_t$. For predicting students performance the following equation is used (for MF without the bold variables):

$$p_{t,s} = \boldsymbol{\mu} + \boldsymbol{b_s} + \boldsymbol{b_t} + \sum_{k=1}^{K} w_{sk} h_{tk}, \tag{1}$$

$t$ represents a task, $s$ a student, $k$ the latent features and $K$ represents the total number of latent features. The optimization function is represented by:

$$\min_{w_s, h_t, \boldsymbol{b_t}, \boldsymbol{b_s}} \sum_{s,t \in \mathcal{D}} (y_{ts} - \hat{y}_{ts})^2 + \lambda \left( \|W\|^2 + \|H\|^2 + \|\boldsymbol{b_t}\|^2 + \|\boldsymbol{b_s}\|^2 \right) \tag{2}$$

with $\mathcal{D}$ the set of collected task student interactions. The final goal of the algorithm is to minimize the Root Mean Squared Error (RMSE) on the set of known scores.

In order to evaluate the performances of BMF and MF generally simple models like Global Average (GA, using the Global Average Score (GAS) of the students as prediction value) are used. To check which is the contribution of the Biases of the BMF to the performance of the MF we use the model called Biases, which has Eq. 2 as optimization function and Eq. 1 as prediction function, but with $K = 0$.

## 4. DATA COLLECTION AND ITS CHARACTERISTICS

In this section we describe the ITS we used, the data collection and what was done to connect Fraction Tutor and MF approaches.

### 4.1 Data collection and sequencing

We have carefully translated the English/US American FT tasks into child-friendly German and iteratively adapted to German students' needs. As a result of the translation and adaption process the US American and the German tasks are not 100% identical and we are using TL according to the definition in Sec. 2 and exploiting the knowledge from the auxiliary Englis dataset to ameliorate the German Performance Prediction.

We used three different sequences to have an equal number of interactions for each task, each sequence using a different order of task categories (6 categories). The interleaved sequence starts with one task of each category (hierarchically) and repeats this process. The second sequence refers to the so called blocked practice sequence where first all tasks of category I need to be solved, then category II and so on. Last is the mixed sequence that has a coincidental order.

In order to collect log data and train the MF for the FT we conducted a study with students (i.e. fifth graders) in classrooms (i.e. 21-28 students per class) in Germany. Students of three classes (88 students) of a German Gymnasium could interact with FT which was integrated in the iTalk2Learn platform [4].

The US American data were collected when students (14 of one class) interacted with the US American version of FT [3]. To these students tasks were proposed in a single sequence. All of them completed at least half of the sequence.

### 4.2 Dataset characteristics

[3]From now on we will refer to Machine Learning's Transfer Learning as TL in order not to mix it with the students' transfer learning

[4]The iTalk2Learn platform is a Plug-In platform used to integrate different components. In our case: FT tasks, database, and simple fixed sequencer.

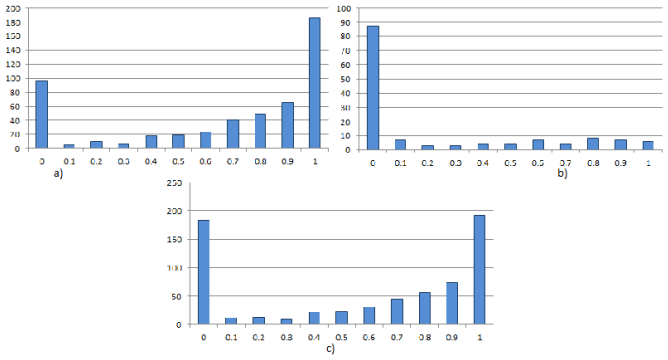

**Figure 1: a) German scores b) English scores, c) combined German and English scores**

For exploring the task cold-start problem for the German and English datasets (described in Sec. 1) we assigned to each task IDs from 0 to 23, where German and English tasks' (0-15) translations have the same ID. As a result we have: 14 interactions for IDs $0-6$, 11 for ID 7 $((7;11))$, $(8;10)$, $(9;8)$, $(10;6)$, $(11;2)$, $(12;2)$, $(13;1)$, $(14;1)$, $(15;1)$. For the German data the interactions are more spread out because of the three different sequences which were used: $(0;38)$,$(1;59)$, $(2;36)$, (3;0), (4;73), (5;47), (6;5), (7;0), (8;22), (9;29), (10;3), (11;0), (12;22), (13;32), (14;0),(15;0), (16;24), (17;32), (18;12), (19;26), (20;29), (21;28), (22;0), $(23,2)$. There are IDs only used in the English data: $(3,7,11,15)$. The tasks (11, 14, 15, 22, 23) have less than 2 interactions for the German and English datasets and are removed in the preprocessing. Thanks to the different sequences we have a sufficient number ([6]) of interactions for most tasks. For the English experiments we removed the last tasks, since there were too few intercations.

For the students' cold-start problem the dataset can be considered as sparse. The English dataset should be less influenced by the students' cold-start problem, because each student interacted at least with 7 tasks.

In order to have a continuous score measure as we had in [9] we used following equation to compute the score:

$$\text{score} = 1 - \left( \frac{\#hints}{\#totalnumhints} + (\#incorrect * 0.1) \right) \quad (3)$$

If the score is less than zero we set the score to 0 avoiding negative scores. For the German (a)), English (b)) and German+English (c)) data we computed the score Histogram to measure how much the data is unbalanced (See Fig. 1). Both datasets are very unbalanced but by combining the two datasets we can achieve a more balanced distribution. We will explain in the Experiment Section how this is influencing the models' performances.

# 5. EXPERIMENTS

To split the data in test and train set we used Leave One Out (LOO) for each student; which is a common approach to split for small datasets (here we used the last task seen by the student). To evaluate the error we measure the RMSE averaged over five experiments to avoid the influence of the random initialization of the model parameters on the model performances. The standard deviation of the error for the models prediction lies around $10^{-3}$, which is normal for

| HL | GA | Biases | MF | BMF |
|---|---|---|---|---|
| $\geq 3$ | 0.337 | 0.390 | 0.405 | 0.386 |
| $\geq 4$ | 0.336 | 0.371 | 0.385 | 0.370 |
| $\geq 5$ | 0.325 | 0.325 | 0.337 | 0.334 |
| $\geq 6$ | 0.319 | 0.321 | 0.328 | 0.322 |
| $\geq 7$ | 0.333 | 0.358 | 0.355 | 0.355 |
| $\geq 8$ | 0.345 | 0.298 | 0.296 | 0.292 |

a)

| HL | GA | Biases | MF | BMF |
|---|---|---|---|---|
| $\geq 7$ | 0.285 | 0.240 | 0.235 | 0.235 |
| $\geq 8$ | 0.295 | 0.241 | 0.229 | 0.218 |

b)

**Figure 2: a) RMSE German, b) RMSE English**

| HL | $\geq 3$ | $\geq 4$ | $\geq 5$ | $\geq 6$ | $\geq 7$ | $\geq 8$ |
|---|---|---|---|---|---|---|
| GAS | 0.673 | 0.673 | 0.673 | 0.673 | 0.673 | 0.673 |
| # test students | 88 | 72 | 53 | 38 | 26 | 22 |
| # interactions | 517 | 501 | 482 | 467 | 455 | 451 |

**Table 1: GA and test size German data**

movie recommender datasets and small datasets. For each experiment we used the models described in sec. 3 (GA, MF, BMF). For finding the best hyperparameters we used Grid Search (learning rate: $[0.01, 0.09]$ stepsize 0.01; regularization: $[0.001, 0.009]$ stepsize 0.001, $[0.01, 0.09]$ stepsize 0.01, $[0.1, 0.9]$ stepsize 0.1; num. iterations: $100-300$ stepsize 20; num. latent features: $2-100$ stepsize 10). Moreover for each experiment we computed the performance Global Average Score (GAS) and report the number of students whose data are used.

## 5.1 Cold-start problem, MF Utility and Intra-Student Variance

For our experiments we studied different History Lengths (HL), i.e. the number of interactions the student had with the ITS, and we deleted the students with a HL less than 2. Starting with HL $\geq 3$ we continued removing the students with $HL \leq 4$, $HL \leq 5$, etc. until $HL \leq 8$. We kept the same train data and just removed the test data, so the test set shrinks while increasing the HL requirements. GAS and number of test students are reported in Tab. 1. Table a) in Fig. 2 lists the RMSE for the German dataset.

The performances as well as the behavior of Biases, BMF and MF are coherent with the one reported in [10]. For $HL \leq 5$ Biases, MF and BMF have not sufficient informationto predict the performances (see a) in fig. 2). Keeping students with $HL \leq 5$ in the train influenced BMF negatively. The small gain between BMF and Biases can be explained with the performances of MF which are almost always worse than GA ones. This is coherent with MF and BMF behaviors where generally Biases give a strong contribution to the model performances. We can say that the Performance Prediction of GA was positively influenced by having all data in the train set, since it can be computed on a more robust statistic. BMF and MF are in general influenced by data of students with short history negatively at the beginning, although, for students with a longer history, these data can be used to ameliorate performances.

Next we evaluate the performances of Biases/MF/BMF on an even smaller dataset: the English one. The performances also of GA are quite good, although Biases, MF, and BMF clearly outperform it (see b) in Fig. 2). GA prediction ability is due to the fact that the dataset is highly unbalanced; with a majority of samples with 0 score the probability that a sample of this dataset is similar to the GAS is higher. Fig. 2 shows that BMF outperforms the Biases and the re-

| HL | GA | Biases | MF | BMF |
|---|---|---|---|---|
| $\geq 3$ | 0.506 | 0.391 | 0.407 | 0.389 |
| $\geq 4$ | 0.500 | 0.375 | 0.389 | 0.375 |
| $\geq 5$ | 0.522 | 0.333 | 0.342 | 0.331 |
| $\geq 6$ | 0.516 | 0.322 | 0.336 | 0.321 |
| $\geq 7$ | 0.523 | 0.346 | 0.362 | 0.344 |
| $\geq 8$ | 0.514 | 0.293 | 0.285 | 0.288 |

a)

| HL | GA | Biases | MF | BMF |
|---|---|---|---|---|
| $\geq 7$ | 0.564 | 0.277 | 0.288 | 0.273 |
| $\geq 8$ | 0.564 | 0.283 | 0.310 | 0.275 |

b)

**Figure 3: a) RMSE GerEng, b) RMSE EngGer**

sults are better than the German ones. According to our previous experience, we think that the difference in the performances (comparing experiments with same HL to avoid the cold-start problem contribution) is due to the variance between the different elements of the students' population under study. In our previous work [1] we showed the negative impact of intra-class variance in the performance of classifiers with small data samples. This applies in our opinion to the case because the intra-student variance of the German data, collected in three classes from different schools, should be higher than the intra-student variance of the English dataset that was collected in one class only.

## 5.2 Transfer Learning

To test the possibility to use English data to ameliorate the German prediction performances, we combined the English and German datasets as follows. In this experiment the data from an English task and its translation are considered by the MF as the same task. When combining the German and English datasets (See Table a) in Fig. 3), the performances of GA drop to approximately 0.5 because the most samples are almost equally distributed between 0 and 1 with a GAS around 0.56. To prove feasibility of TL we ran more experiments starting with the best results of the previous Sections. We added the English data to the German train set Table a) in Fig. 3), where the addition of the English data in training is always taking to a contribution for $HL \geq 6$.
The same amelioration cannot be seen when adding the German data to the English train, since adding the German data increases the intra-student variance worsening the English model performances (Table b) in Fig. 3, and Tab. 2).

| BMF + HL | $\geq 3$ | $\geq 4$ | $\geq 5$ | $\geq 6$ | $\geq 7$ | $\geq 8$ |
|---|---|---|---|---|---|---|
| German | 0.386 | 0.370 | 0.334 | 0.322 | 0.355 | 0.292 |
| GerEng | 0.389 | 0.375 | 0.331 | 0.321 | 0.344 | 0.288 |
| English | / | / | / | / | 0.235 | 0.218 |
| EngGer | / | / | / | / | 0.273 | 0.275 |

**Table 2: Comparison of BMFs perfromances for all experiments.**

# 6. CONCLUSIONS

In this paper we proposed a practical solution to the data collection to reduce data sparsity, by proposing tasks with different sequences. Moreover, we analyzed in detail the effects of a small dataset on the performances of MF used as performance predictor. Thanks to these analyses it was also possible to determine the utility of MF based performance predictors and sequencing in new ITS' tasks. Considering the Utility of BMF in comparison to GA, before having at least 7 interactions for a student it would be better to use GA as performance predictor. With using TL we already get better results for BMF with HL $\geq 5$. This should hold theoretically also for the use of the VPS, although an experiment with online model update is required for a full evaluation. Finally, we proposed to exploit generally discarded data exploiting the concept of TL. As future work we will investigate more advanced methods to perform TL on small datasets and try to ameliorate performances of the first BMF predictions ($HL \leq 5$).

# 7. ACKNOWLEDGMENTS

The research leading to the results reported here has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 318051 - iTalk2Learn project (www.iTalk2Learn.eu). Special thanks goes to Jenny Olsen and Martina Rau who agreed to give us access to their versions of Fractions Tutor. Special thanks further goes to Brett Leber and Jonathan Sewall who helped us by adapting the Fractions Tutor tasks into German. Last but not least, we want to thank Hamza Sati, Sebastian Strauss, Jörg Striewski, and Richard Hesse for supporting us when conducting the classroom study in Germany.

# 8. REFERENCES

[1] R. Janning, C. Schatten, and L. Schmidt-Thieme. Hnnp-a hybrid neural network plait for improving image classification with additional side information. In *ICTAI*. IEEE, 2013.

[2] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 2015.

[3] J. Olsen, D. Belenky, V. Aleven, and N. Rummel. Using an intelligent tutoring system to support collaborative as well as individual learning. In *ITS*, volume 8474 of *LNCS*, pages 134–143. Springer, 2014.

[4] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, Oct 2010.

[5] Š. Pero and T. Horváth. Comparison of collaborative-filtering techniques for small-scale student performance prediction task. In *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*. Springer, 2015.

[6] I. Pilászy and D. Tikk. Recommending new movies: Even a few ratings are more valuable than metadata. RecSys, 2009.

[7] C. Schatten, R. Janning, and L. Schmidt-Thieme. Integration and evaluation of a machine learning sequencer in large commercial its. In *AAAI*. Springer, 2015.

[8] C. Schatten, M. Mavrikis, R. Janning, and L. Schmidt-Thieme. Matrix factorization feasibility for sequencing and adaptive support in its. In *EDM*, 2014.

[9] C. Schatten and L. Schmidt-Thieme. Adaptive content sequencing without domain information. In *CSEDU*, 2014.

[10] N. Thai-Nghe, L. Drumond, T. Horvath, A. Krohn-Grimberghe, A. Nanopoulos, and L. Schmidt-Thieme. Factorization techniques for predicting student performance. *IGI Global*, 2011.