6th International Conference on Higher Education Advances (HEAd'20)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/HEAd20.2020.11137

On the Use of Bayesian Probabilistic Matrix Factorization for Predicting Student Performance in Online Learning Environments

Jinho Kim[1], Jung Yeon Park[2], Wim Van Den Noortgate[1]
[1]Faculty of Psychology and Educational Sciences and ITEC, imec research group, KU Leuven, Belgium, [2]School of Education, George Mason University, USA.

Abstract

Thanks to the advances in digital educational technology, online learning (or e-learning) environments such as Massive Open Online Course (MOOC) have been rapidly growing. In the online educational systems, however, there are two inherent challenges in predicting the performance of students and providing personalized supports to them: sparse data and the cold-start problem. To overcome such challenges, this article aims to employ a pertinent machine learning algorithm, the Bayesian Probabilistic Matrix Factorization (BPMF) that can enhance the prediction by incorporating background information on the side of students and/or items. An experimental study with two prediction scenarios and 24 experimental conditions was conducted to study the BPMF based on real online learning data. The results show that the lower rate of missingness and the appropriate dimensionality of latent features provided better prediction accuracy in both prediction scenarios. The use of side information enhanced the prediction accuracy but the effect was diminished for the high dimensional latent features when the data are sparse. The methodological value, applicability, and practical implications of the BPMF and side information to the online educational systems were also discussed.

Keywords: *digital educational technology; online learning; online educational systems; machine learning; Bayesian Probabilistic Matrix Factorization; student performance prediction.*

1. Introduction

Digital educational technology has advanced considerably over the last few decades. Thanks to the advances, particularly, online learning (or e-learning) environments such as Massive Open Online Course (MOOC) have been rapidly growing and getting attention. Such online educational systems have promising advantages in helping students access more easily to the qualified instructions and resources as well as in allowing them to manage their learning process flexibly (Zhang & Chang, 2015). Moreover, students would have more benefits from the adaptive assessment tailored to the behaviors and needs of individual students in the online learning environments. In the online educational systems, however, there are two inherent challenges in predicting the future performance of students and providing personalized supports to them: sparse data and the cold-start problem. First, as widely known in general online recommender systems, student-item interaction data are often inevitably sparse (in this context, *sparse* means that many elements of the data matrix are empty); there is a considerable number of students who respond to only a certain subset of all possible items in the online systems. Second, the online educational systems suffer from an inability to recommend and/or provide appropriate items for new students who have started the online assessment for the first time. The systems are often incapable of correctly matching the new students with the incipient items due to the lack of background information for each student, which results in inaccuracy of item recommendations (and hence a lot of dropout) at the beginning of online learning, which is called the 'cold-start' problem (Bobadilla et al., 2012).

To overcome such challenges, a pertinent machine learning algorithm for the online educational systems, the Bayesian Probabilistic Matrix Factorization (BPMF; Salakhutdinov & Mnih, 2008), can be employed. By incorporating background information on the side of students and/or items, it enhances the student performance prediction. This article aims to examine the methodological value, applicability, and practical implications of the BPMF and side information to the real online learning data, so that it would help facilitate a personalized learning for students, develop adaptive assessment, instructional strategies, and course curriculum for teachers and developers. An experimental study is conducted to apply the BPMF to the logging data of an online learning environment, Statistics Online. The experiment is designed to study the prediction in two challenging scenarios encountered in online educational systems: (a) existing students take only a part of the all items (some student-item interaction data are sparse), and (b) the data are entirely missing for new students who do not take any items yet. Given the two prediction scenarios, several experimental conditions on the rate of data missingness to be filled and the dimensionality of latent features to be used in the BPMF are considered. Then, background information variables on the student and/or item sides of the Statistics Online data are incorporated into the BPMF to see if either or both of them enhance the accuracy of predicting students' performance in both prediction scenarios.

2. Algorithm of Bayesian Probabilistic Matrix Factorization

The BPMF is an advanced version of matrix factorization that uses a collaborative filtering approach in machine learning algorithms (Salakhutdinov & Mnih, 2008). Compared to the previous matrix factorization methods, the BPMF can use additional information for students and/or items when making predictions on the two latent feature vectors, and hence it can implement more accurate factorization, resulting in better predictions especially for the entities that have few or no observations. Its prediction process is formulated in mathematical terms as follows. Let Y denote a response data matrix and Y_{ij} is the binary response value (0 = incorrect response; 1 = correct response) of student i on item j; U denotes a D-dimensional latent feature matrix for students and U_i is the i-th student-specific vector in U; V denotes a D-dimensional latent feature matrix for items and V_j is the j-th item-specific vector in V. Note that D-dimension represents a size of the latent feature vectors, U_i and V_j. In this setup, an optimization algorithm enables to predict the incomplete entries in the observed Y, which minimizes the sum of the squared differences between Y_{ij} and $U_i^T V_j$, the dot product of i-th row in U and the j-th column in V. Specifically, the following optimization problem is solved in terms of a loss function L of U and V:

$$L(U,V) = \sum_{(i,j)\in I_Y}(Y_{ij} - U_i^T V_j)^2 + \lambda_U \|U\|_F^2 + \lambda_V \|V\|_F^2 , \quad (1)$$

where I_Y is the set of observed entries in Y and $\| . \|_F$ represents the Frobenius norm, which is the square root of the sum of the absolute squares of the elements in the matrix. Regularization parameters, $\lambda_U > 0$ and $\lambda_V > 0$, are derived from Gaussian priors on the U_i and V_j and a Gaussian noise model on the observed Y_{ij}.

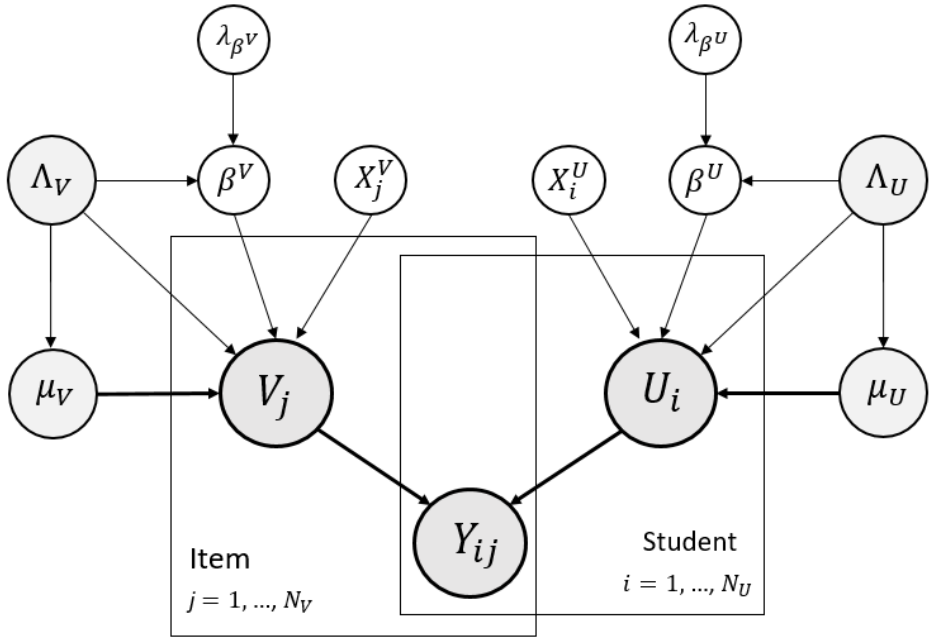

Figure 1. Model parameters and hyperparameters for the BPMF.

For the model parameters used in the BPMF (see Figure 1), each latent feature vector is assumed to follow a multivariate normal distribution. That is, $U_i \sim MVN(\mu_U + \beta_U X_i^U,\ \Lambda_U)$, where μ_U is a vector of prior means for U_i, Λ_U is a variance covariance matrix for U_i and β_U is a vector of weights for the i-th student's background variable X_i^U. Similarly, $V_j \sim MVN(\mu_V + \beta_V X_j^V,\ \Lambda_V)$, where μ_V is a vector of prior means for V_j, Λ_V is a variance covariance matrix for V_j, and β_V is a vector of weights for the j-th item's background variable X_j^V. Particularly, model parameters and hyperparameters are integrated out using Markov chain Monte Carlo methods, which enables to control the model complexity automatically based on the training data. Based on the discovered latent feature matrices, the missing entries are predicted on a probability scale between 0 and 1 by computing the dot products of the two latent feature vectors of the corresponding rows and columns, U_i and V_j, respectively.

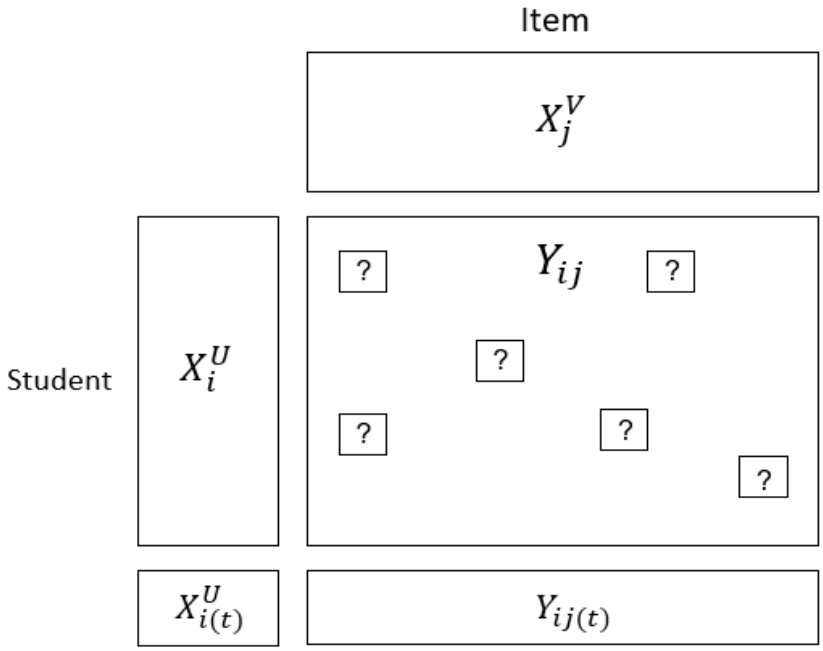

Figure 2. Illustration of the prediction scenarios.

3. Design and Evaluation

3.1. Prediction Scenarios

To consider the practical challenges in the online learning environments, two prediction scenarios for the BPMF are taken into account in an experiment as follows:

(a) Scenario 1 (sparse data); existing students have sparse response data because they take only a part of the all items,

(b) Scenario 2 (cold-start problem); new students have no response data because they did not take any items yet.

Figure 2 illustrates the two prediction scenarios in which the BPMF is implemented. For the first scenario, missing entries (indicated with a question mark) in the sparse data matrix Y for existing students will be predicted. For the second scenario, missing entries in the unobserved data matrix $Y_{(t)}$ for new students will be predicted, whereas background information $X_{i(t)}^U$ on the new student i may be available.

3.2. Experimental Conditions

To investigate the best working conditions of the BPMF for each of the two prediction scenarios, and to study whether the use of student and/or item information variables can enhance the prediction accuracy, several experimental conditions are considered:

(1) the rate of data missingness to be filled is varied by 10% or 50%,

(2) the dimensionality of latent features to be used is varied by 1, 5, 10, or 20 dimensions,

(3) none, only student information variables, or both student and item information variables are incorporated into the BPMF.

In total, 2 x 4 x 3 = 24 experimental conditions are considered in each prediction scenario.

3.3. Evaluation Method

For each data set, the prediction performance of the BPMF is evaluated by a 10-fold cross validation (CV). Note that *k*-fold CV means that the benchmark dataset is randomly selected by *k* subsets of equal size. Then, one of the *k* subsets is defined as a "test" subset for the evaluation of the predictions. The remaining *k*-1 subsets are used for "training" the model. This procedure is repeated by *k* times, each time a different subset is used as the test subset. Each subset (fold) is used only once as a test subset. In the end, the final result is calculated by the average of the *k* different results obtained from the *k*-fold CVs. To evaluate the accuracy of the predictions made by our system, we employed Receiver Operating Characteristic (ROC) curves and Area under the Receiver Operating Characteristic (AUROC). Note that a ROC curve represents the relation between true positive rates (true positive/(true positive+false nagative)) and false positive rates (false positive/(false positive+true negative)) at various thresholds. A precision recall curve is defined as the precision (true positive/(true positive+false positive)) against the recall (true positive/(true positive+ false negative)) at various thresholds. The true positive rate is the same as recall, and is also denoted as sensitivity, while the false positive rate is also denoted as (1-specificity). In case of totally random predictions the AUROC is approximately equal to 0.5 and Area Under the Precision–Recall (AUPR) is equal to the frequency of the positive class.

4. An Experimental Study

4.1. Data

The data set contains item responses of 2,044 students at KU Leuven (University of Leuven), Belgium to a total of 20 assessment items in the university's online course for regression analysis. The observed responses are all dichotomous indicating whether the student has answered the item correctly (= 1) or not (= 0). In addition to the student-item responses, there

are various side information about students and item themselves. Specifically, there are 23 student-side information variables that describe students' background (e.g., the status of dyslexia, dyscalculia, AD(H)D, ASS, another language problem, school type, resident area, and so on) and 6 item-side information variables for the items' properties (e.g., question type, attainment target, and so on). Although the initial response data and side information variables do not have any missing observations, they are adjusted based on the experimental conditions we considered for the purpose of an experimental study.

4.2. Analysis

For data analysis, 'Macau' (Simm et al., 2015), a Python package that provides wrapper functions for the BPMF was used. The MCMC sampling with 100,000 iterations (first half as a burn-in) was used to factorize a student-item matrix. The final estimates of the model parameters were obtained by taking the mean of the posterior samples after the burn-in periods of 50,000 iterations.

4.3. Results

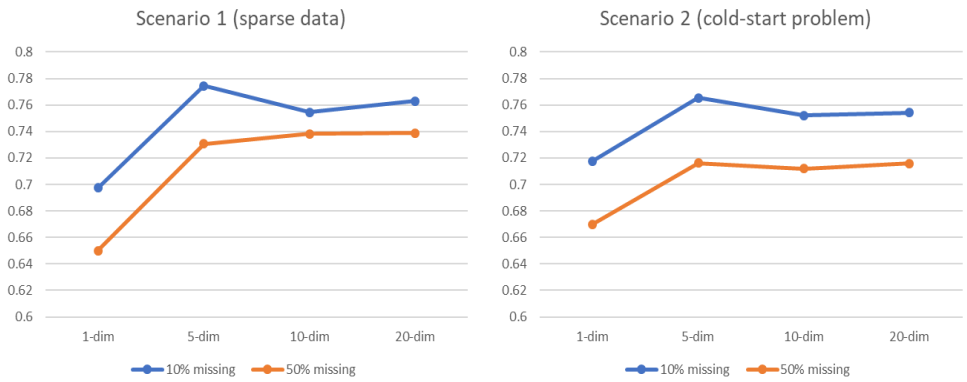

Figure 3. AUROC plots showing the effects of missingness rate and latent feature dimensionality in the BPMF.

As can be seen in the AUROC plots (see Figure 3), the results show that AUROC was greater in dataset with 10% missing values compared to the one with 50% missing values in both prediction scenarios, regardless of the number of latent feature dimensions. In both prediction scenarios, the prediction accuracy increased dramatically between 1 and 5 dimensional latent features, but afterward any changes look very minor even with a higher number of latent feature dimensions. The BPMF with 10 and 20 dimensions may overfit the current dataset.

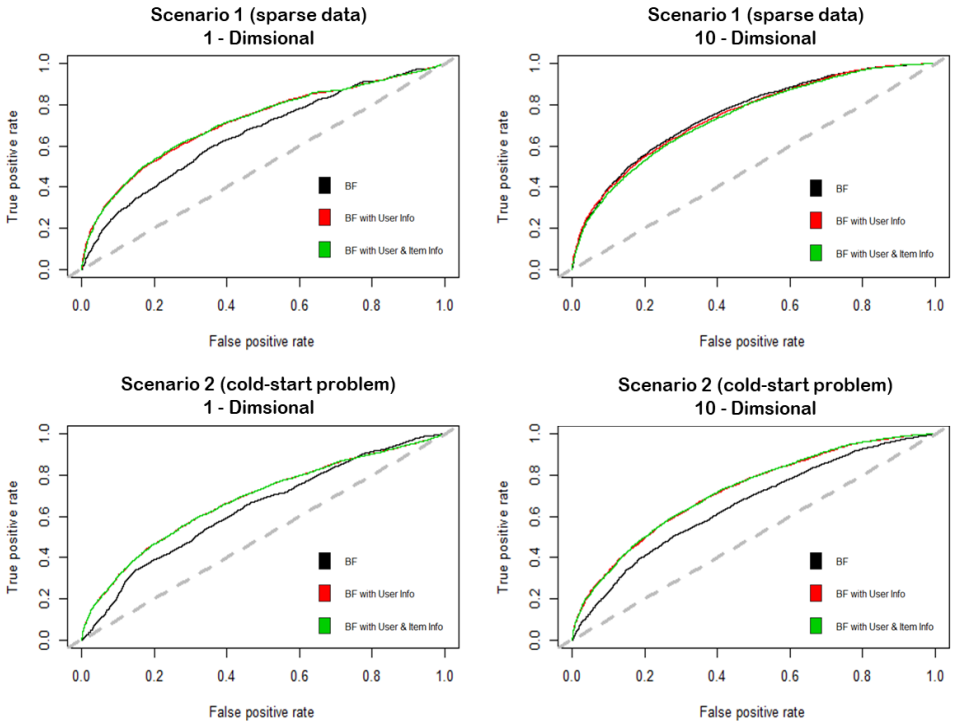

Figure 4. ROC curves showing the effect of using side information variables in the BPMF.

Figure 4 visualizes the relations between true positive rate and false positive rate using ROC curves. The top two panels show the results of Scenario 1 when predicting the existing student's performance on the unsolved items; it is found that compared to the BPMF without using any side information, using side information provided better prediction accuracy (top left). But the effect of using the side information was diminished when 10 dimensional latent features were used (top right). On the other hand, the bottom two panels show the results of Scenario 2 when predicting the new student's performance on the existing items; it is found that using side information provided better prediction accuracy for both 1 and 10 dimensional latent features. Additionally, there was almost no gain by adding item-side information on top of student-side information in both prediction scenarios.

5. Conclusion and Discussion

An experimental study to predict student performance using the BPMF was designed with the two prediction scenarios (Scenario 1 for the sparse data, Scenario 2 for the cold-start problem) and 24 experimental conditions. The Statistics Online data were adjusted and used for the experimental study to consider the reality in the online learning environments. The

results show that the lower rate of missingness (10%) and the appropriate dimensionality of latent features (5 dimensions) provided better prediction accuracy in both prediction scenarios. The use of student and/or item side information variables in the BPMF enhanced the prediction accuracy; the effect was diminished for the high dimensional latent features (10 dimensions) in Scenario 1 but it was kept for both of the low and high dimensional latent features in Scenario 2. Thus, considering the two challenging scenarios, this study helps us find optimal conditions on the use of the BPMF to predict student performance more accurately, which has a predictive value in a methodological perspective.

Moreover, this study sheds light on the applicability and practical implications of the BPMF and side information to the online educational systems. It can help tearchers and developers forecast individual or grouped students' performance on the online course, which allows them to develop and improve personalized or grouped instructional strategies, course curriculum, adaptive assessment, and other elements of the online educational systems. It can also help facilitate a personalized learning for the existing students as well as provide a suitable recommendation of the items and courses for the new students.

Lastly, more extensive simulations are desired for future studies to identify the effects of latent feature dimensions where different ratios of the number of students and items are given and the strength of relation between latent features and side information.

References

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, 26, 225-238.

Salakhutdinov, R., & Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In W. Cohen, A. McCallum, & S. Roweis (Eds.), *Proceedings of the 25th international conference on Machine learning* (pp. 880-887). Association for Computing Machinery, New York: NY, USA.

Simm, J., Arany, A., Zakeri, P., Haber, T., Wegner, J. K., Chupakhin, V., Ceulemans, H., and Moreau, Y. (2015). Macau: Scalable Bayesian multi-relational factorization with side information using MCMC. arXiv e-prints:1509.04610v2 [stat.ML].

Zhang, S., & Chang, H. H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, 1(1), 67-92.