Extending Bayesian Personalized Ranking with Survival Analysis for MOOC Recommendation

Alireza Gharahighehi
alireza.gharahighehi@kuleuven.be
KU Leuven, Campus Kulak, Department of Public Health and Primary Care
Belgium
Itec, imec research group at KU Leuven
Kortrijk, Belgium

Michela Venturini
michela.venturini@kuleuven.be
KU Leuven, Campus Kulak, Department of Public Health and Primary Care
Belgium
Itec, imec research group at KU Leuven
Kortrijk, Belgium

Achilleas Ghinis
achilleas.ghinis@student.kuleuven.be
KU Leuven, Department of Mathematics
Leuven, Belgium

Frederik Cornillie
frederik.cornillie@kuleuven.be
KU Leuven, Department of Linguistics
Kortrijk, Belgium
Itec, imec research group at KU Leuven
Kortrijk, Belgium

Celine Vens
celine.vens@kuleuven.be
KU Leuven, Campus Kulak, Department of Public Health and Primary Care
Belgium
Itec, imec research group at KU Leuven
Kortrijk, Belgium

ABSTRACT

Massive Open Online Courses (MOOCs) have recently attracted students and professionals as complementary tools to academic education. Despite the number of advantages MOOCs provide, such as openness and flexibility regarding learning pace, such courses are characterized by a consistently higher dropout rate than conventional classrooms. A crucial factor that influences dropout is the choice of the appropriate course, hence the need for effective course recommendations. A course recommendation system (RS) that uses dropout information can mitigate course withdrawal and user dissatisfaction. In this paper, an extension of Bayesian Personalized Ranking, which is a learning-to-rank RS, is proposed that uses the pseudo-labels extracted by survival analysis based on dropout information to recommend courses in the context of MOOCs. The proposed approach performs the best compared to six competing RSs on three MOOCs datasets.

CCS CONCEPTS

• **Information systems → Recommender systems**; *Collaborative filtering*; • **Mathematics of computing → Survival analysis**.

KEYWORDS

Bayesian personalized ranking

UMAP '23 Adjunct, June 26–29, 2023, Limassol, Cyprus

ACM ISBN 978-1-4503-9891-6/23/06.
https://doi.org/10.1145/3563359.3597394

ACM Reference Format:
Alireza Gharahighehi, Michela Venturini, Achilleas Ghinis, Frederik Cornillie, and Celine Vens. 2023. Extending Bayesian Personalized Ranking with Survival Analysis for MOOC Recommendation. In *UMAP '23 Adjunct: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23 Adjunct), June 26–29, 2023, Limassol, Cyprus.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3563359.3597394

1 INTRODUCTION

Despite the growing popularity of Massive Open Online Courses (MOOCs), both in contexts of independent (self-directed) learning and in formal educational programs, learning in MOOCs is characterized by a very high dropout rate. Average dropout rates of up to 90% have been reported in courses offered by premier institutions like MIT and Harvard [2]. Tackling the challenge of dropout in MOOCs requires both an empirical understanding of the phenomenon, informed by theory in the learning sciences (e.g. [9]), and the design of appropriate and effective interventions that can help learners in setting their goals and in completing them. In this context, focusing on better course recommendations might help decrease the chances of dropout and guide users toward the appropriate choices. As MOOCs potentially offer universal access to education, improving the quality of recommendations in such settings is also critical in promoting equitable quality education. Recommender Systems (RSs) are intelligent filtering algorithms that model users' preferences and recommend items that fit these preferences. Generally, there are two main types of RSs, content-based filtering, and collaborative filtering. While content-based filtering RSs only use the target user's interactions to build a user profile and recommend items that best match the user profile, collaborative filtering RSs utilize collaborative information, i.e., the interactions of other users, to infer the target user's preferences.

RSs use users' feedback to infer their preferences. Usually, instead of users' explicit feedback, i.e., ratings or like/dislike, only users' implicit feedback, such as clicking on a link, watching a video, reading an article, or adding an item to their basket is available. Implicit feedback, also called one-class feedback [16], only represents users' positive feedback. Therefore, users' negative and missing feedback is not distinguishable. The students' feedback in MOOCs is often implicit. Bayesian Personalized Ranking (BPR) [17] is a learning-to-rank collaborative filtering RS that models users' preferences based on implicit feedback. The main assumption in BPR is that users prefer items that they have observed to the ones that they have not observed. In this paper, we propose an extension of BPR based on different types of user feedback in the context MOOCs.

Survival Analysis (SA) is a set of statistical and machine learning methods where the outcome of interest is the time to a certain event [3]. The main peculiarity of survival data is the presence of censoring, that is, missing information regarding the time of the event of interest; right-censoring is the most common form and happens when the event of interest is not observed during follow-up or the individual is lost before the end of the follow-up. The main advantage of SA is its ability to handle such partial information during the learning process, by properly encoding censored observations, which are usually discarded in classification or regression tasks. We believe that time to dropout represents valuable information in the context of course recommendations as it provides knowledge regarding students' engagement with courses [18]. In our work, we propose a novel method that exploits time to dropout information using SA to provide better course recommendations in MOOCs context. Specifically, we use SA to enrich the set of available student-course interactions by adding pseudo-labeled interactions that are integrated into the BPR framework as an additional type of implicit feedback. As illustrated in Fig 1, BPR only distinguishes between seen (positive feedback) and unseen courses (negative feedback). In the context of MOOCs, a user may complete courses or dropout from them. While both types of events are seen interactions, they do not reflect a similar preference level. Furthermore, time-to-dropout from courses can be used by SA to predict other courses with high risk of dropout (red "D" in the figure) as pseudo-labeled interactions. Given this additional feedback level, a more elaborated sampling strategy can be applied in TBPR[1] to better model user preferences.

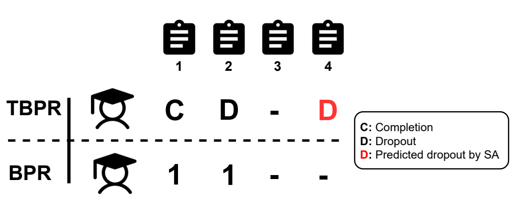

Figure 1: A toy example to illustrate the proposed approach

[1] Bayesian Personalized Ranking with time-to-dropout

2 RELATED WORK

Although the original BPR [17] is based on binary implicit feedback, several studies have extended BPR to deal with non-binary implicit feedback. Lerche and Jannach [11] proposed an extension of BPR (GBPR) by adding pair-wise preferences between graded positive feedback, for instance, more clicks reflect more preference. Loni et al. [13] proposed multi-channel BPR (MBPR) where they introduced a biased sampler to sample a positive and a negative item for a user based on different positive feedback levels. They [13] also proposed another extension of BPR (MBPR-P) where the popularity of items is considered in the biased sampler. In [6], another extension of BPR (EBPR) is proposed where users' consumption behavior such as reading a news article or listening to a music track is used to model users' preferences.

Dropout prediction in MOOCs has been extensively studied as a classification problem, and several Machine Learning models were used [2, 4]. However, in such examples, time information was mostly discarded. SA is well suited to model time to dropout information and the literature provides some promising examples. Gitinabard et al. [8] used survival analysis to model dropout risk and uncover social and behavioral feature impact on the outcome. Xie [20] employed survival analysis to model the hazard function of dropout by using the learner's viewing duration on a course. Labrador et al. [10] performed a prospective study on an online MOOCs platform and used Cox Proportional Hazard regression to uncover the most important factors related to student dropout. Wintermute et al. [19] modeled the certificate rates of MOOC users with a *Weibull* survival function, following the intuition that students "survive" in a course for a particular time before stochastically dropping out. Pan et al. [15] proposed a more sophisticated SA deep learning approach to address volatility and sparsity of the data, that moderately outperformed Cox. However, to the best of our knowledge, such time to dropout has never been incorporated in MOOC recommendations.

3 METHODOLOGY

In this paper, we extend BPR with different types of implicit feedback that exist in the context of MOOCs. BPR is a learning-to-rank RS based on binary implicit feedback that learns model parameters using users' pairwise preferences between observed (positive class) and unobserved items (negative class). For instance, in the case of matrix factorization, the model parameters are users' and items' latent features, i.e., embeddings. In each training round of BPR, a sample (u, i, j) is drawn from the interaction data, where u is a user, i is an item that user i has observed, i.e., an item from the positive class, and j is an item that the same user has not observed, an item from the negative class. Given this sample, BPR updates the embeddings of u, i, and j using Stochastic Gradient Descent (SGD).

3.1 BPR extension with additional feedback

The users' implicit feedback in MOOCs has different types. The obvious form is users' enrolments in courses. Given this type of feedback, BPR assumes that users prefer courses that they have enrolled in over the courses that they have not enrolled in. As mentioned, dropout is an important event in the user experience in MOOCs and therefore should be considered in modeling users' preferences. The enrollments can result in course completion or dropout.

3.2 Pseudo-labelled interaction extraction

Our approach is based on the enrichment of the training set with additional pseudo-labeled interactions, namely unobserved student-course interactions that are likely to result in dropout events. In order to do so, we trained *xgboost* for SA, using the observed interactions in the training set, considering time-to-dropout as the outcome, thus encoding course completion as censoring. Subsequently, we used the model to predict pseudo-labels for the unseen interactions from the training set and obtain an additional feedback class to be used to train BPR. Details about the implementation are given in Section 4. Using the given users' implicit feedback and the extracted pseudo-labeled interactions by SA, we propose an extension of BPR that exploits this additional possible feedback to learn the model parameters, i.e., user and item latent features. The extracted pseudo-labeled interactions are the ones that have a high risk of dropout. Algorithm 1 presents the proposed approach:

Algorithm 1: Bayesian Personalized Ranking with time-to-dropout (TBPR)

Input: User feedback D
Output: Learned parameters Θ (user and item latent features)
Initialize parameters Θ;
Extract pseudo-labelled dropout interactions in D using SA as a new feedback class;
Repeat
 draw a positive class + from $p(+)$;
 draw a positive interaction (u, i) from $p(u, i|+)$;
 draw a negative class − from $p(-|u, +)$;
 draw a negative item j from $p(j|u, -)$;
 update Θ using (u, i, j) and SGD based on the BPR update rule [17];
Until *Convergence*;

To draw samples from the empirical distributions ($p(+)$ and $p(-|u, +)$) mentioned in Algorithm 1, the four hyperparameters defined in Table 1 should be specified[2]. For instance, if the positive class is "completion" (with the probability of α), the negative class is "dropout", "pseudo-labelled", or "missing", with the probabilities of β, γ, or $1-\beta-\gamma$, respectively. Positive interactions ((u, i) from $p(u, i|+)$) and negative items (j from $p(j|u, -)$) are uniformly sampled form possible candidates.

4 EXPERIMENTAL DESIGN

4.1 Datasets prepossessing and description

Publicly available datasets generated from MOOCs are scarce and most of them are described by Lohse et al. in [12]. We evaluated our approach by using three widely used publicly available datasets, namely XuetangX [5], KDDCUP [5], and Canvas [14]. Both KDD-CUP and Xuetangx anonymized datasets are provided by XuetangX platform[3]. The Canvas dataset contains de-identified data from Canvas Network[4] open courses from January 2014 to September 2015. Table 2 describes the three (preprocessed) publicly available datasets relating to MOOCs that were used to evaluate the proposed approach. The raw JSON files containing logs with all interactions an individual had with a course for the XuentangX and KDDCUP datasets were processed to extract the first and last interactions a user had with a given course and the time-to-event variable was defined as the difference between the dates of these actions. Canvas dataset was in tabular format and it already contained that information.

Table 1: Sampling hyperparameters

			Negative class $P(-\|u,+)$		
			Dropout	Pseudo-labelled	Missing
Positive class $P(+)$	Completion	α	β	γ	$1-\beta-\gamma$
	Dropout	$1-\alpha$	-	δ	$1-\delta$

Table 2: Datasets descriptions

	XuentangX	KDDCUP	Canvas
# Users	2417	1944	959
# Items	246	39	193
Sparsity	95.5%	87.1%	95.4%

4.2 Experimental setup

Each dataset is split into three disjoint sets: training, validation, and test sets. Test and validation sets contain one interaction (course completion) per user. The rest of the interactions are used for training. Apart from the proposed approach (TBPR), we consider six competing approaches in our experiments: the original BPR [17], MBPR [13], MBPR with popularity bias (MBPR-P) [13], GBPR [11], EBPR [6] and the popularity-based (Pop) RS as a simple baseline. To evaluate the performance of RSs, two performance measures, namely *recall* and *Normalized Discounted Cumulative Gain (NCDG)*, are considered. *Recall* evaluates the RS in predicting the relevant courses in the top@k recommendation list, i.e., the ranked list of k courses that the user has not interacted with. *NDCG* is a rank-sensitive measure that penalizes the score of recommendations if the relevant items appear in the lower ranks in the recommendation list. There are some hyperparameters to be tuned on the validation sets, including the ones for BPR, GBPR, EBPR, MBPR, and α, β, γ and δ to draw samples of the proposed approach (TBPR[5]). The hyperparameters are selected based on *NDCG*.

To predict the pseudo-labeled interactions used to augment the training set, we modeled the tabular feature set for *xgboost* considering the known interactions in the training set for each course and each user. Additionally, we applied *PCA* for dimensionality reduction. We trained *xgboost* for SA using default parameters[6] and we selected the 10% of observations with the highest predicted risk score of dropout among the unseen interactions.

[2]The hyperparameter tuning procedure is explained in Section 4.

[3]https://www.xuetangx.com/
[4]https://www.canvas.net/
[5]Time-to-event BPR
[6]https://xgboost.readthedocs.io/en/stable/parameter.html

Table 3: Results based on top@3 recommendations.

	Xuentangx		Canvas		KDDCUP	
	ndcg@3	recall@3	ndcg@3	recall@3	ndcg@3	recall@3
Pop	0.048	0.055	0.057	0.063	0.188	0.207
BPR	0.143	0.156	0.139	0.153	0.447	0.485
GBPR	0.100	0.108	0.078	0.092	0.362	0.392
EBPR	0.153	0.167	0.124	0.140	0.458	0.495
MBPR	0.157	0.174	0.125	0.137	0.473	0.513
MBPR-P	0.067	0.075	0.107	0.119	0.320	0.352
TBPR	**0.171**	**0.183**	**0.146**	**0.161**	**0.480**	**0.522**

5 RESULTS AND DISCUSSIONS

The results of applying the proposed and the competing approaches on the three datasets described in Section 4 are reported in Table 3. The values in Table 3 are based on *top@3* recommendation lists. The proposed approach, TBPR, consistently outperforms the other competing approaches in all datasets based on both performance measures.

The MBPR-P approach performs inferior compared to MBPR which indicates that adding the popularity bias to the model does not help in the context of MOOCs. GBPR performs worse compared to BPR. A possible reason is that the sampling approach in GBPR is not flexible enough to express the differences between different types of implicit feedback. The proposed approach (TBPR) is more expressive as it is trained with additional information generated by SA and therefore performs better than the competing approaches which do not use this information.

6 CONCLUSION

The main contribution of this paper is to apply survival analysis (SA) to model time-to-dropout in the context of Massive Open Online Courses recommendations and to extend Bayesian Personalized Ranking (BPR) with the additional information, i.e., pseudo-labels, generated from the SA model. The proposed approach performs better compared to BPR and four other BPR extensions in three datasets based on Normalized Discounted Cumulative Gain (NCDG) and Recall for top@3 recommendations.

The experiments are still ongoing and therefore the results in this paper are preliminary. There are several possible directions for future work. While we showed that using the pseudo-labels generated by the SA method has positive impact on the performance of BPR, we believe there is still room for improvement as the models (the SA method and BPR) are optimized separately. A possible future work is to cast the problem as a multi-task learning setting where the tasks are time-to-event prediction and course ranking. Another promising direction for future work is to consider the choice of SA model as a hyperparameter and tune it based on the dataset. For instance, one could consider the use of a cure survival model [1] which might provide less biased dropout prediction in our setting. Furthermore, one can apply survival analysis to model time-to-completion and further extend the training set. In this paper, the proposed approach and baselines are evaluated only based on relevance performance measures. Other performance criteria, such as diversity [7], can be considered to evaluate models based on different perspectives. Finally, the proposed approach can be extended by including pair-wise comparisons within a class, for instance, two courses that resulted in dropout but with different time-to-dropout.

ACKNOWLEDGMENTS

This research was funded by the Research Fund Flanders (project G0A2120N). The authors also acknowledge the Flemish Government (AI Research Program).

REFERENCES

[1] Mailis Amico and Ingrid Van Keilegom. 2018. Cure models in survival analysis. *Annual Review of Statistics and Its Application* 5 (2018), 311–342.

[2] Jing Chen, Bei Fang, Hao Zhang, and Xia Xue. 2022. A systematic review for MOOC dropout prediction from the perspective of machine learning. *Interactive Learning Environments* (2022), 1–14.

[3] Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. 2003. Survival analysis part I: basic concepts and first analyses. *British journal of cancer* 89, 2 (2003), 232–238.

[4] Fisnik Dalipi, Ali Shariq Imran, and Zenun Kastrati. 2018. MOOC dropout prediction using machine learning techniques: Review and research challenges. In *2018 IEEE global engineering education conference (EDUCON)*. IEEE, 1007–1014.

[5] Wenzheng Feng, Jie Tang, and Tracy Xiao Liu. 2019. Understanding dropouts in MOOCs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 517–524.

[6] Alireza Gharahighehi and Celine Vens. 2020. Extended bayesian personalized ranking based on consumption behavior. In *Benelux Conference on Artificial Intelligence, Belgian Dutch Conference on Machine Learning*. Springer, 152–164.

[7] Alireza Gharahighehi and Celine Vens. 2023. Diversification in session-based news recommender systems. *Personal and Ubiquitous Computing* 27, 1 (2023), 5–15.

[8] Niki Gitinabard, Farzaneh Khoshnevisan, Collin F Lynch, and Elle Yuan Wang. 2018. Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. *arXiv preprint arXiv:1809.00052* (2018).

[9] Judy Tzu Huang. 2012. Learning Retention. In *Encyclopedia of the Sciences of Learning*. Springer US, 1964–1965.

[10] Manuel Medina Labrador, Gustavo R Garcia Vargas, Jorge Alvarado, and Mario Caicedo. 2019. Survival and risk analysis in MOOCs. *Turkish Online Journal of Distance Education* 20, 4 (2019), 149–159.

[11] Lukas Lerche and Dietmar Jannach. 2014. Using graded implicit feedback for bayesian personalized ranking. In *Proceedings of the 8th ACM Conference on Recommender systems*. 353–356.

[12] James J Lohse, Christine A McManus, and David A Joyner. 2019. Surveying the MOOC data set universe. In *2019 IEEE Learning With MOOCS (LWMOOCS)*. IEEE, 159–164.

[13] Babak Loni, Roberto Pagano, Martha Larson, and Alan Hanjalic. 2016. Bayesian personalized ranking with multi-channel user feedback. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 361–364.

[14] Canvas Network. 2016. Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Open Dataset. https://doi.org/10.7910/DVN/1XORAL

[15] Feng Pan, Bingyao Huang, Chunhong Zhang, Xinning Zhu, Zhenyu Wu, Moyu Zhang, Yang Ji, Zhanfei Ma, and Zhengchen Li. 2022. A survival analysis based volatility and sparsity modeling network for student dropout prediction. *PloS one* 17, 5 (2022), e0267138.

[16] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 502–511.

[17] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.

[18] Marili Rõõm, Marina Lepp, and Piret Luik. 2021. Dropout time and learners' performance in computer programming MOOCs. *Education Sciences* 11, 10 (2021), 643.

[19] Edwin H Wintermute, Matthieu Cisel, and Ariel B Lindner. 2021. A survival model for course-course interactions in a Massive Open Online Course platform. *PloS one* 16, 1 (2021), e0245718.

[20] Zheng Xie. 2019. Modelling the dropout patterns of MOOC learners. *Tsinghua Science and Technology* 25, 3 (2019), 313–324.