Course recommendations in MOOCs using collaborative filtering and survival analysis (Invited Paper)

Alireza Gharahighehi[1,2,*], Michela Venturini[1,2] and Celine Vens[1,2]

[1]*KU Leuven, Campus Kulak, Department of Public Health and Primary Care*
[2]*Itec, imec research group at KU Leuven*

Abstract

Massive Open Online Courses (MOOCs) are becoming a complementary, or even preferred, method of learning compared to traditional education among learners. While MOOCs enable learners to access a wide range of courses from various disciplines, anytime and anywhere, a significant number of course enrollments in MOOCs end up in dropouts. To increase learners' engagement in MOOCs, they need to interact with the courses that match their preferences and needs. A course Recommender System (RS) models learners' preferences and recommends courses based on their previous interactions within the MOOC platform. Dropout events in MOOCs, like other time-to-event predictions, can be effectively modeled using survival analysis methods. The objective of this talk is to illustrate the benefits of employing survival analysis in enhancing the performance of collaborative filtering-based course recommendations in MOOCs.

Keywords

recommendation systems, survival analysis, collaborative filtering, massive open online courses

1. Introduction

Massive Open Online Courses (MOOCs) platforms offer a diverse selection of online courses to learners worldwide, promoting the concept of equitable learning by removing barriers of location and time. Despite its considerable advantages, a significant portion of MOOC enrollments end up in dropouts. It has been reported that dropout rates for courses offered by prestigious institutions like MIT and Harvard can be as high as 90% [2]. While dropouts may result from various reasons, such as accessing only the free portions of the courses, finding the course or topic irrelevant, or insufficient competencies, this information is valuable for modeling users' preferences in MOOCs and would provide more infromed recommendations.

Recommender Systems (RSs) are intelligent information retrieval algorithms that utilize users' past interactions to suggest the most relevant items to them. Generally, RSs can be categorized into two main types: Content-based filtering and collaborative filtering. Content-based filtering RSs recommend items whose features match those of items that the target user has previously liked. On the other hand, collaborative filtering RSs model users' preferences based on similarities between the past interactions of users and items.

In a MOOC platform, a collaborative filtering-based RS can be applied to recommend courses to users based on their previous enrollments in the platform. While previous enrollments are informative to model users' preferences, the dropout information is still missing in this kind of recommendations. The dropout event in MOOCs is crucial as a significant portion of enrollments result in dropout. This additional information about user-course interactions can be useful to better model users' preferences or needs regarding the courses in MOOC platform. Survival analysis (SA) comprises a set of statistical methods that model the time until an instance experiences a specific event such as death or machine failure [3]. The key characteristic of survival data is that some instances have unobserved events,

ALL'24: Workshop on Adaptive Lifelong Learning, July 08–12, 2024, Recife, Brazil [1]

[*] Corresponding author.

alireza.gharahighehi@kuleuven.be (A. Gharahighehi); michela.venturini@kuleuven.be (M. Venturini); celine.vens@kuleuven.be (C. Vens)
0000-0003-1453-1155 (A. Gharahighehi)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3836/paper5.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

referred to as censored data. The most common form of censoring in SA is right-censoring, where the target event is not observed during follow-up or the instance is lost before the end of the follow-up period. The main strength of SA is its utilization of such partial information during the learning process by considering instances with censored events, which are usually discarded in classification and regression tasks. We believe that time to dropout is highly informative in modeling users' preferences in the context of course recommendations, as it provides valuable insights regarding students' engagement in MOOCs [4].

In a previous study [5], the authors demonstrated that SA can improve the performance of a specific RS, namely Bayesian Personalized Ranking (BPR), when the predictions of a SA method, trained based on time to dropouts, are embedded in the BPR algorithm. In this invited talk, we discuss how to generalize the usage of SA in any type of collaborative filtering-based RS. In the next section, we briefly report the existing literature around dropout in MOOCs and then in Section 3, elaborate on the research questions that could be tackled by researchers regarding enhancing MOOC recommendations using dropout information. Finally, in Section 4, we illustrate the possible experimental setup to conduct such research studies.

2. Related work

The task of dropout prediction in the context of MOOCs has been mainly modeled as a classification task [6, 2]. While in this studies the task was predicting the event of dropout they ignored the time information, i.e., time to dropout, in their predictions. SA can be used to incorporate the time information in modeling dropout in MOOCs and there are some promising examples in the literature. In [7] survival analysis was used to model dropout risk in the context of MOOCs and unveil social and behavioral feature impacts on the outcome. Xie [8] utilized survival analysis to examine the hazard function of dropout, employing the learner's course viewing duration on a course in MOOCs. Labrador et al. [9] specified the fundamental factors attached to learners' dropout in an online MOOCs platform using Cox Proportional Hazard regression. Wintermute et al. [10] applied Weibull survival function to model the certificate rates of learners in a MOOCs platform, assuming that learners "survive" in a course for a particular time before stochastically dropping out. In [11] a more sophisticated SA deep learning approach was proposed to tackle volatility and sparsity of the data, that moderately outperformed the Cox model.

While SA has been already applied in literature to model dropout in MOOCs, to the best of our knowledge, such time to dropout from courses has never been incorporated in course recommendations in the context of MOOCs. The research gap this invited talk aims to investigate is whether utilizing this time-to-event information would enhance the performance of typical collaborative filtering approaches, and how it would do so.

3. Research questions

In this invited talk, the focus is to discuss the merits of SA in course recommendations when the time-to-events information is available in the context of MOOCs. The following research questions would be interesting to tackle:

1. Does a SA method trained based on time to dropout have a positive impact on the performance of course recommendations in MOOCs when combined with a regular RS? In a previous paper [5], authors followed the most straightforward approach to combine the SA method and the RS. They enhanced the training data of the RS using the predictions of the SA method. What would be the other possible ways of combining the SA method and the RS?
    a) Is it possible to generalize the proposed augmentation idea in [5] to the other possible learning-to-rank recommendation approaches such as Weighted Approximate-Rank Pairwise (WARP) [12]?

Table 1
Datasets descriptions

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

b) Is it possible to use the predictions of SA directly for recommendations, i.e., sort the items based on their risk scores predicted by SA? Can SA directly compete collaborative filtering approaches?

c) How effective it is if the predictions of SA being used to post-process the output of a collaborative filtering method, i.e., to adjust the predictions of RS based on the predictions of SA?

d) Does it make sense to model the whole prediction task in MOOCs as a multi-task learning problem where the tasks are predicting time-to-dropout or the risk scores for the SA method and ranking the courses for the recommendation task? Do these two mentioned tasks benefit from a partially shared learning process?

e) To select the SA method, which objective should be followed? Should the SA method be selected/tuned based on SA objectives such as c-index, or RS objectives such as recall, or a combination of both?

f) Can we define other events of interests in the context of MOOCs such as course completion or milestone achievement?

g) How to recommend a completely new published course to the learners? How to address the item cold-start problem [13]?

2. How to generalize the discussed idea to other contexts and applications?

a) To the best of our knowledge, there are only three publicly available MOOC datasets (described in Table 1) that can be used to validate course RSs. Is it possible to apply the same idea to other applications, for instance series recommendations, using the time to stop watching the series?

b) Is it possible to extend the current idea with multiple time to events, for instance course completion and course dropout in the context of MOOCs? How to make an ensemble of recommendations [14] based on these different methods?

4. Experimental setup

To conduct research studies to answer the raised research questions in the previous section, powerful collaborative filtering RSs should be employed as the competing approaches. Following the findings of the award winning paper [15], traditional collaborative filtering methods such as BPR [16], WARP [12], User- and Item-based k Nearest Neighbors (UKNN [17] and IKNN [18]) and Sparse Linear Methods (SLIM [19]) outperform more recent neural network based RSs and therefore can be used as baselines. These baselines can be used to benchmark performance of a SA-based RS, a collaborative filtering RS enhanced with SA post-processing, or a multi-task method that learns both the time-to-event prediction task and the recommendation task. Researchers could apply the competing methods on the three publicly available datasets, namely Xuentangx [20], Canvas [21] and KDD-CUP [20] and possibly an additional dataset from another domain such as series recommendations. The competing methods shall be evaluated based on the typical evaluation measures for RSs such as NDCG and recall.

References

[1] A. Gharahighehi, R. Van Schoors, P. Topali, J. Ooge, Adaptive lifelong learning (all), in: International Conference on Artificial Intelligence in Education, Springer, 2024, pp. 452–459. doi:10.1007/978-3-031-64312-5_57.

[2] J. Chen, B. Fang, H. Zhang, X. Xue, A systematic review for mooc dropout prediction from the perspective of machine learning, Interactive Learning Environments (2022) 1–14. doi:10.1080/10494820.2022.2124425.

[3] T. G. Clark, M. J. Bradburn, S. B. Love, D. G. Altman, Survival analysis part i: basic concepts and first analyses, British journal of cancer 89 (2003) 232–238. doi:10.1038/sj.bjc.6601118.

[4] M. Rõõm, M. Lepp, P. Luik, Dropout time and learners' performance in computer programming moocs, Education Sciences 11 (2021) 643.

[5] A. Gharahighehi, M. Venturini, A. Ghinis, F. Cornillie, C. Vens, Extending bayesian personalized ranking with survival analysis for mooc recommendation, in: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, 2023, pp. 56–59. doi:10.1007/978-3-031-64312-5_57.

[6] F. Dalipi, A. S. Imran, Z. Kastrati, Mooc dropout prediction using machine learning techniques: Review and research challenges, in: 2018 IEEE global engineering education conference (EDUCON), IEEE, 2018, pp. 1007–1014.

[7] N. Gitinabard, F. Khoshnevisan, C. F. Lynch, E. Y. Wang, Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features, arXiv preprint arXiv:1809.00052 (2018).

[8] Z. Xie, Modelling the dropout patterns of mooc learners, Tsinghua Science and Technology 25 (2019) 313–324.

[9] M. M. Labrador, G. R. G. Vargas, J. Alvarado, M. Caicedo, Survival and risk analysis in moocs, Turkish Online Journal of Distance Education 20 (2019) 149–159. doi:10.17718/tojde.640561.

[10] E. H. Wintermute, M. Cisel, A. B. Lindner, A survival model for course-course interactions in a massive open online course platform, PloS one 16 (2021) e0245718. doi:10.1371/journal.pone.0245718.

[11] F. Pan, B. Huang, C. Zhang, X. Zhu, Z. Wu, M. Zhang, Y. Ji, Z. Ma, Z. Li, A survival analysis based volatility and sparsity modeling network for student dropout prediction, PloS one 17 (2022) e0267138.

[12] J. Weston, S. Bengio, N. Usunier, Wsabie: Scaling up to large vocabulary image annotation, in: Twenty-Second International Joint Conference on Artificial Intelligence, 2011.

[13] A. Gharahighehi, K. Pliakos, C. Vens, Addressing the cold-start problem in collaborative filtering through positive-unlabeled learning and multi-target prediction, IEEE Access 10 (2022) 117189–117198.

[14] A. Gharahighehi, C. Vens, K. Pliakos, An ensemble hypergraph learning framework for recommendation, in: Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24, Springer, 2021, pp. 295–304. doi:10.1007/978-3-030-88942-5_23.

[15] M. F. Dacrema, P. Cremonesi, D. Jannach, Are we really making much progress? a worrying analysis of recent neural recommendation approaches, in: Proceedings of the 13th ACM conference on recommender systems, 2019, pp. 101–109.

[16] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, 2009, pp. 452–461.

[17] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 10th international conference on World Wide Web, 2001, pp. 285–295.

[18] P. Lops, M. d. Gemmis, G. Semeraro, Content-based recommender systems: State of the art and trends, Recommender systems handbook (2011) 73–105.

[19] X. Ning, G. Karypis, Slim: Sparse linear methods for top-n recommender systems, in: 2011 IEEE 11th international conference on data mining, IEEE, 2011, pp. 497–506. doi:10.1109/ICDM.2011.134.

[20] W. Feng, J. Tang, T. X. Liu, Understanding dropouts in moocs, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 517–524. doi:10.1609/aaai.v33i01.3301517.

[21] C. Network, Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Open Dataset, Harvard Dataverse, 2016. doi:10.7910/DVN/1XORAL.