

Augmenting the DonorsChoose.org Corpus for Meta-Learning^{*}

Gordian Edenhofer^{1,2}, Joeran Beel^{1,2}, Andrew Collins¹, and Akiko Aizawa²

¹ Trinity College Dublin, School of Computer Science & Statistics, ADAPT Centre, Ireland

² National Institute of Informatics Tokyo, Digital Content and Media Sciences Division, Japan
beelj@tcd.ie, collinsa@tcd.ie, aizawa@nii.ac.jp

Abstract. The DonorsChoose.org dataset of past donations provides a big and feature-rich corpus of users and items. The dataset matches donors to projects in which they might be interested in and hence is intrinsically about recommendations. Due to the availability of detailed item-, user- and transaction-features, this corpus represents a suitable candidate for meta-learning approaches to be tested. This study aims at providing an augmented corpus for further recommender systems studies to test and evaluate meta-learning approaches. In the augmentation, metadata of collaborative and content-based filtering techniques is amended to the corpus. It is further extended with aggregated statistics for users and transactions and an exemplary meta-learning experiment. The performance in the learning subsystem is measured via the recall of recommended items in a Top-N test set. The augmented dataset and the source code are released into the public domain at [GitHub:BeelGroup/Augmented-DonorsChoose.org-Dataset](https://github.com/BeelGroup/Augmented-DonorsChoose.org-Dataset).

Keywords: recommender systems · meta-learning · dataset augmentation · ensemble learning · hybrid recommenders.

1 Introduction

Meta-Learning is the process of applying machine learning algorithms on metadata as is generated by a subsystem composed of machine learners. The goal of meta-learning systems is to improve existing approaches by combining the strengths of several single machine learning systems. In recommender systems it is an emerging field of interest as meta-learning promises to better sort through item, user and transaction information as to retrieve the best possible match of item and user.

One challenge for researchers interested in meta-learning is that suitable datasets are time-consuming and cumbersome to construct. Creating a system of individual machine learning experiments and amending their metadata to the original corpus takes time and resources away from addressing more important challenges in meta-learning for recommender systems. Multiple examples of manual and repetitious data augmentation may be found in the scientific literature, for instance in [1, 2, 3].

Our objective is to address the data augmentation part in a comprehensive way as to allow researchers to focus on the actual meta-learning evaluation. We are presenting a

^{*} This publication emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 13/RC/2106.

public domain dataset which to our best knowledge has not been used for meta-learning before in the scientific community, yet due to its feature-rich, vast corpus is a suitable candidate. The raw dataset is augmented with metadata of several machine learning algorithms from a variety of fields (cf. [2]). It is further amended with aggregated statistics about users and items and metadata of an exemplary experiment.

2 Methodology

2.1 The Dataset

DonorsChoose.org is an organization which enables teachers to file project requests for resources for their school. Collectively the nonprofit has raised \$685 million for classrooms in the United States of America. Their dataset composed of past donations features 4.69 million transactions performed by 2.12 million users having donated to 1.11 million projects. The corpus provides significantly more transactions with more detailed item and transaction information compared to, e.g., the popular MovieLens dataset. An exemplary itemized view of the augmented dataset is shown in Table 1.

The corpus features comprehensive information on the users (donors), items (projects) and transactions respectively interactions (donations). Transactions are represented by an interaction of a user with an item at a specific point in time and further include a strength (amount donated). Users feature information about their location and a boolean value indicating whether they are a teacher. Items feature multiple textual descriptions, funding goals, a list of associated categories and types, information about the creator and details about the location of the school which the item is about.

In a learning subsystem step, two collaborative filtering techniques and two content-based filtering techniques are applied. The performance measures of the single algorithms are amended to the table. Next, aggregated user and item statistics are calculated and added to the table. Lastly, four exemplary meta-learners are applied and the performance measures of the individual approaches are added to the corpus.

Table 1: Illustrative example of the overall design of the augmented transaction table with amended learning subsystem performance scores, statistics and meta-learner information.

Transaction		Donor Details		Project Details		Rating	Learning Subsystem Performance		Statistics		Meta-Learner Performance	
Donor ID	Project ID	Location	Teacher	Location	...		Collab	Content	Mean by User	Item	User-Clustering	...
		Zip ...		Zip		SVD ...	fastText ...	ZipIsEqual	K-Means	...
309bb	00002	495 ...	True	495	2	6 ...	20 ...	1.0	...	20	...
A8fa8	c0001	28 ...	False	200	1	11 ...	4 ...	0.5	...	4	...
38298	e0006	195 ...	False	856	3	9 ...	1 ...	0.0	...	1	...

2.2 Dataset Preparation

To prepare the dataset for evaluations, duplicate interactions are merged as the recommender systems shall not recommend items to which the user has already donated to. Internal stop words in description strings of items are stripped. Furthermore, transactions containing no information about the user’s location are dropped. Thereby enabling the consistent usage of a user’s location information in new approaches.

Additionally, users having interacted less than twice are removed as well. Although, this limits the target group, the step is unavoidable as the employed validation process requires at least one interaction per user for testing and one for training. The evaluation requires user profiles and hence training-transactions in order to recommend new items which in turn need to be validated against test-transactions. Lastly, the donated amount is transformed to a transaction strength in the range of unity to five in analogy to a rating score. This step reduces the influence of outliers respectively unusual donated amounts.

2.3 Learning Subsystem

The learning subsystem is a recommender system itself. SVD and KNN are used as representatives for collaborative filtering techniques and TF-IDF and word-embedding for content-based filters [4, 5]. Hereby, the otherwise intrinsic inductive bias is reduced. The word-embedding relies on fastText and utilizes pre-trained vectors [6]. Vector representations of items are the normalized sum of the vectors of the normalized embeddings of the individual words. In order to work with the data in an efficient way, transactions are sampled. For this study a sample-size of 100,000 was chosen. This represents a fair balance between required computing time and ability to generalize.

A classical 5-fold cross-validation on the users is applied for training the collaborative filters. Using the recall in a Top-N test set, a score of how good an algorithm is able to recommend an item is calculated. The algorithm is provided with all the items with which the user from the test set has interacted with except for one and is requested to create a ranking of 100 random items with which the user has not interacted with plus the one with which an interaction has taken place. An ordering is assumed to be implied by the magnitudes of the values in the reconstructed matrix of the decomposition respectively the distance to the neighbor. The position of the interacted item in the set of 101 items is then used as score for the user-item transaction.

The validation process of the content-based approaches uses leave-one-out. For each transaction a profile of the user is constructed which excludes this transaction. Afterwards, the algorithm is requested to create a ranking of 100 items with which the user has not interacted with and the one with which a transaction has taken place. The ranking is performed using the cosine similarity of an item's vector and the user's profile.

The choice of evaluation method ensures that it can be applied to a vast set of different algorithms as long as a ranking of items can be produced and therefore allow for a high variety within the set of algorithms in the learning subsystem. It further ensures that not just the ability to decompose the data is measured but the ability to recommend new items to users [7]. Hereby, the learning subsystem is kept extensible while still providing useful metadata to meta-learning algorithms.

2.4 Statistics and Application of Learning Subsystem

The meta-learning system processes the metadata from the learning subsystem and further statistics. Its performance is evaluated via a holdout split on the transaction table using 80% for training and the rest for testing. Transactions are expanded to contain more detailed information about the time, the user's location and aggregated statistics about the mean of the categories, location and interaction-strength of items with which

the user has interacted with. Statistics are solely based on the training set. Values for unknown users are filled with the mean in the training data.

Four different meta-learning approaches are discussed which address the Algorithm Selection Problem (ASP) via a switching hybrid ensemble. The aim of the system is to predict the algorithm with the lowest recall-position for each transaction respectively each row in Table 1. The overall performance of a meta-learner is measured using the mean recall-position for when the recommended algorithm for each transaction is used.

The first approach aims at predicting the algorithm which will best describe the transaction via a classification based on the given meta-features using a decision tree. The second approach employs a gradient boosting regressor to predict the position of the recall for each transaction and for each algorithm in the learning subsystem separately. As best algorithm, the one with the lowest predicted recall-position is chosen.

The third meta-learner is a classical stacking ensemble using a decision tree which is given the prediction from the learning subsystem as additional input. Based on the metadata and the output of the learning subsystem, it then performs a transaction classification. The final approach aims at solving the ASP via clustering transactions based on the meta-features using K-Means and assigns clusters an overall best algorithm via majority voting.

3 Results

The preparation of the dataset preserves a little more than half of the data with the most notable drop in the number of transactions being introduced by requiring users to have donated at least twice. The sample of 100,000 transactions contains 18,735 unique users and 88,100 unique projects.

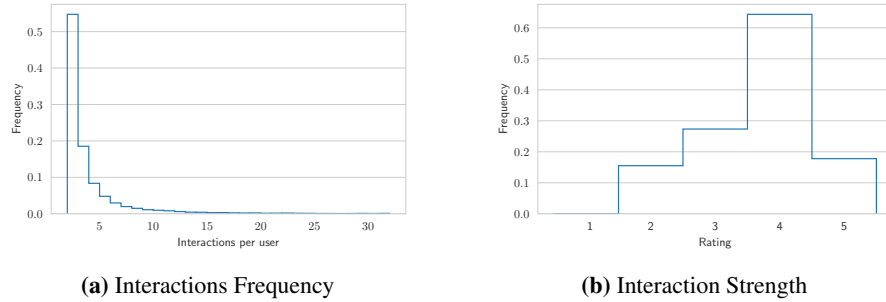


Fig. 1: Frequency of the number of interactions per user without outliers on the left and the distribution of the transformed interaction strength on the right.

The requirement enforcing users to have interacted at least twice results in a distribution of the number of users interacting with items which decays exponentially as depicted in Figure 1a. This is to be expected as the time users can invest in researching items and their total interaction strength (budget) is limited. The dataset is dominated by users which interacted exactly two times.

The distribution of the transformed interaction strength in analogy to a rating score is shown in Figure 1b. It peaks at the second highest score. Interactions with a weak to medium interaction strength and ones with the strongest interaction strength are about equally likely. Interactions with the lowest strength value are very infrequent.

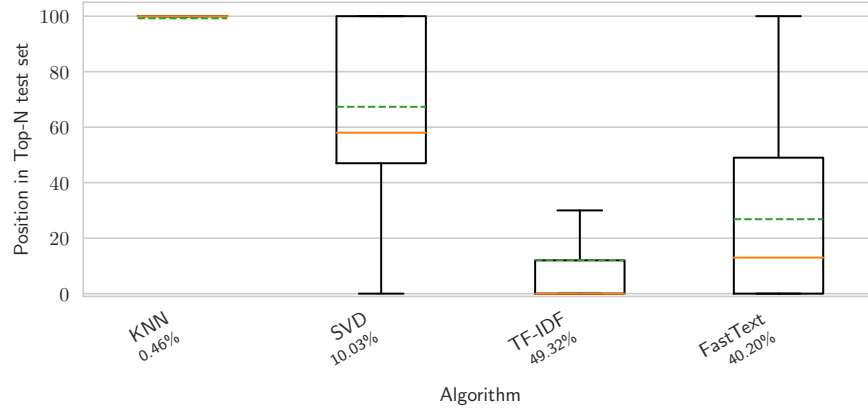


Fig. 2: Statistical features of the distributions of the recall-positions of the learning subsystem. The dashed green lines indicate the mean, the straight orange lines the median. The percentage below each algorithm indicates how often it is the best possible algorithm. To resolve a draw deterministically, the “best” algorithm is decided in alphabetic order.

Statistical information about the performances are visualized in Figure 2. It can be observed that the performance of the learning subsystem depends highly on the employed filtering techniques. Overall, content-based filters are significantly better in achieving a low recall-position (average position in Top-N set; content-based filtering: 19.48, collaborative filtering: 83.28). The algorithm with the poorest performance is KNN. SVD performs better but is still unable to reliably yield good recommendations. Furthermore, grouping users by shared item interactions has no observable positive impact on the recommendation of new items.

Considering the low average number of interactions per user and the vast number of items available, KNN apparently fails to find appropriate neighbors. SVD seems to struggle with the sparsity of the input matrix as looking at the reconstruction of the decomposed matrix it is revealed that most values are well below unity and very similar to each other. SVD yields recall-positions below 10 infrequently.

Content-based filtering techniques achieve better performance scores on the dataset. User profiles based on TF-IDF and fastText each yield recall-positions below 10 for roughly half of all transactions with the simpler TF-IDF performing slightly better than fastText. This could be at least in part due to users explicitly searching for specific terms instead of exploring all possible items manually.

Between the different meta-learning approaches, significant changes in the mean recall-position can be noted as seen in Figure 3. The worst performer is the user-clustering using K-Means (average position in Top-N: 15.56). The stacking decision tree performs

best (average position in Top-N: 9.02). Of the three non-stacking approaches the gradient booster performs best (average position in Top-N: 11.21) and achieves a better score than the single best algorithm, i.e., TF-IDF (average position in Top-N: 11.96). The classifying decision tree (average position in Top-N: 15.27) outperforms the user-clustering but is still notably worse than the overall best algorithm.

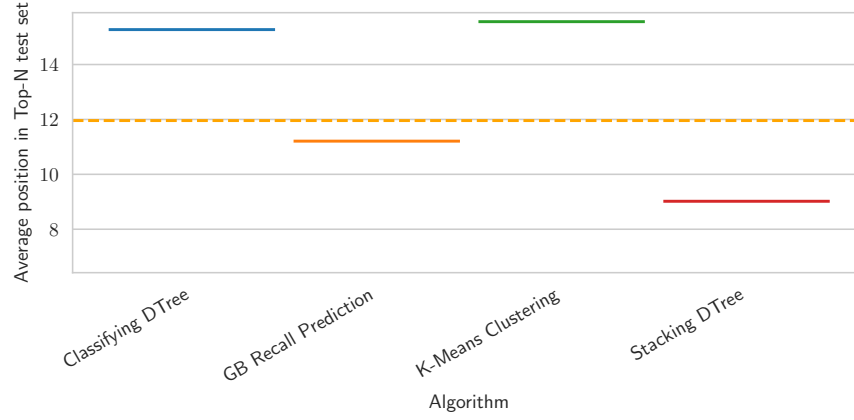


Fig. 3: Performance of the meta-learner system addressing the ASP. The dashed orange line represents the overall best algorithm, i.e., TF-IDF.

That the stacking ensemble performs best is unsurprising as it is provided significantly more information than the other algorithms. The performance of the gradient booster is encouraging as it successfully addresses the ASP. From the poor performance of the user-clustering it can be reasoned that similar transactions do not necessarily yield similar results using the same recommender algorithm. Hence, the similarity of meta-feature does not imply that the performance of recommendations is similar.

The classification and error prediction approach apparently suffer from the level of indirection which is introduced by predicting a single algorithm and disregarding the penalty which is introduced if a bad performing one is chosen. Even a high accuracy in algorithm selection does not guarantee a low mean recall-position. This is partly mitigated by the error prediction approach but still not completely suppressed.

4 Conclusion

The study provides an extensively augmented dataset based on the transaction data published by DonorsChoose.org. The corpus is amended with metadata of individual recommender algorithms. Collaborative and content-based filtering techniques are used in the process and their performance is evaluated via the recall in a Top-N test set. Aggregated user and item statistics are amended to the table. Metadata of four switching hybrid ensemble meta-learners is amended to the dataset. The augmented public domain dataset lays the groundwork that future evaluations of existing and novel meta-learning approaches can build upon.

References

- [1] Tiago Cunha, Carlos Soares, and André C.P.L.F. de Carvalho. “Metalearning and Recommender Systems: A literature review and empirical study on the algorithm selection problem for Collaborative Filtering”. In: *Information Sciences* 423 (2018), pp. 128–144. DOI: <https://doi.org/10.1016/j.ins.2017.09.050>. URL: <http://www.sciencedirect.com/science/article/pii/S0020025517309702>.
- [2] Andrew Collins, Jöran Beel, and Dominika Tkaczyk. “One-at-a-time: A Meta-Learning Recommender-System for Recommendation-Algorithm Selection on Micro Level”. In: *CoRR* abs/1805.12118 (2018). arXiv: [1805.12118](https://arxiv.org/abs/1805.12118). URL: <http://arxiv.org/abs/1805.12118>.
- [3] Michael Ekstrand and John Riedl. “When Recommenders Fail: Predicting Recommender Failure for Algorithm Selection and Combination”. In: *Proceedings of the Sixth ACM Conference on Recommender Systems*. RecSys ’12. Dublin, Ireland: ACM, 2012, pp. 233–236. DOI: [10.1145/2365952.2366002](https://doi.org/10.1145/2365952.2366002). URL: <http://doi.acm.org/10.1145/2365952.2366002>.
- [4] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [5] Radim Řehůřek and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora”. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. <http://is.muni.cz/publication/884893/en>. Valletta, Malta: ELRA, May 2010, pp. 45–50.
- [6] Piotr Bojanowski et al. “Enriching Word Vectors with Subword Information”. In: *CoRR* abs/1607.04606 (2016). arXiv: [1607.04606](https://arxiv.org/abs/1607.04606). URL: <http://arxiv.org/abs/1607.04606>.
- [7] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. “Performance of Recommender Algorithms on Top-n Recommendation Tasks”. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. RecSys ’10. Barcelona, Spain: ACM, 2010, pp. 39–46. ISBN: 978-1-60558-906-0. DOI: [10.1145/1864708.1864721](https://doi.org/10.1145/1864708.1864721). URL: <http://doi.acm.org/10.1145/1864708.1864721>.