A label ranking approach for selecting rankings of collaborative filtering algorithms

Conference Paper · April 2018

DOI: 10.1145/3167132.3167418

3 authors:

Tiago Cunha
Institute for Systems and Computer Engineering of Porto (INESC Porto)
20 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Carlos Soares
University of Porto
235 PUBLICATIONS **2,349** CITATIONS
SEE PROFILE

Andre de Carvalho
University of São Paulo
373 PUBLICATIONS **3,293** CITATIONS
SEE PROFILE

A Label Ranking approach for selecting rankings of Collaborative Filtering algorithms

Tiago Cunha
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal
tiagodscunha@fe.up.pt

Carlos Soares
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal
csoares@fe.up.pt

André C.P.L.F. de Carvalho
Universidade de São Paulo, ICMC
São Carlos, São Paulo, Brasil
andre@icmc.usp.br

ABSTRACT

The large amount of Recommender System algorithms makes the selection of the most suitable algorithm for a new dataset a difficult task. Metalearning has been successfully used to deal with this problem. It works by mapping dataset characteristics with the predictive performance obtained by a set of algorithms. The models built on this data are capable of predicting the best algorithm for a new dataset. However, typical approaches try only to predict the best algorithm, overlooking the performance of others. This study focus on the use of Metalearning to select the best ranking of CF algorithms for a new recommendation dataset. The contribution lies in the formalization and experimental validation of using Label Ranking to select a ranked list of algorithms. The experimental procedure proves the superior performance of the proposed approach regarding both ranking accuracy and impact on the baselevel performance. Furthermore, it draws and compares the knowledge regarding metafeature importance for both classification and Label Ranking tasks in order to provide guidelines for the design of algorithms in the Recommender System community.

CCS CONCEPTS

• **Information systems → Recommender systems**; *Data mining*; • **Computing methodologies → Machine learning**;

KEYWORDS

Collaborative Filtering, Metalearning, Label Ranking

ACM Reference Format:
Tiago Cunha, Carlos Soares, and André C.P.L.F. de Carvalho. 2018. A Label Ranking approach for selecting rankings of Collaborative Filtering algorithms. In *SAC 2018: SAC 2018: Symposium on Applied Computing , April 9–13, 2018, Pau, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3167132.3167418

1 INTRODUCTION

Among the several open challenges in Collaborative Filtering (CF), the choice of the best CF algorithm(s) for a new dataset is still an undeveloped topic. Since training and evaluating all algorithms for a new dataset requires a prohibitive amount of resources, automatic solutions based on prior knowledge are of the utmost importance. Such solution would allow to predict the best algorithms for an endless amount of datasets, without the necessity of running empirical studies for each specific case. This study investigates a Metalearning (MtL)-based approach for CF algorithm selection.

MtL is concerned with discovering the effect of characteristics of the problems on the behavior of algorithms [3]. It has been extensively used for algorithm selection [3]. The algorithm selection problem can be viewed as a learning task. It uses a metadataset, where each meta-example corresponds to a dataset. For each meta-example, the predictive features are characteristics (metafeatures) extracted from the corresponding dataset and the targets are the performance (metatargets) of a set of algorithms for the dataset [2].

The algorithm selection problem for CF has received considerable attention recently [1, 4–6, 9, 11, 13]. The related work investigated the effect of different metafeatures, meta-algorithms and metatargets, for different baselevel algorithms and datasets. However, none studied the selection of a ranking of algorithms, having focused only on the prediction of the best algorithm. This limits the decision making process, by not knowing how other methods are expected to perform. By tackling the algorithm selection problem using rankings, one has a sorted predicted utility for all algorithms [3]. Furthermore, it also allows the extraction of further metaknowledge.

If one considers the CF algorithms to be labels in the classification problem, then one can use ranking-based techniques to tackle the problem of selecting a ranking of algorithms. However, one must also provide a ranking containing all the candidate algorithms, since the system is unaware of which recommended algorithms the practitioner will actually choose [3]. This motivates the hypothesis that the usage of Label Ranking (LR) is a suitable and potentially efficient solution for this task, since it fulfills both requirements.

The main contribution of this paper is the formalization and experimental validation of a LR approach for the selection of rankings of CF algorithms. This approach builds on the state of the art by formalizing the problem and experimentally validating the hypothesis. The proposed approach is evaluated using ranking accuracy measures and its performance compared against the performance of the classification approach used so far. The experimental results show the benefits of the approach proposed in both accounts.

This document is organized as follows: Section 2 presents related work on algorithm selection for CF and LR; Section 3 formalizes the algorithm selection problem using LR and explains the experimental setup. In Section 4, several aspects of the proposed approach are evaluated and discussed. Section 5 presents the conclusions and suggests directions for future work.

SAC 2018, April 9–13, 2018, Pau, France

ACM ISBN 978-1-4503-5191-1/18/04.
https://doi.org/10.1145/3167132.3167418

2 RELATED WORK

2.1 Algorithm selection for CF

Earlier work in algorithm selection for CF relies on the analysis of simple NN and MF algorithms on a hand-full of datasets, evaluated with common error based measures. The metafeatures used focus on the ratings distribution, matrix structure and neighborhood statistics. The metamodels are built using regression algorithms, evaluated using correlations and/or error metrics [1, 9, 11, 13]. Recently, the problem has been investigated in more detail [4]. Both the base- and metalevels experimental setups were significantly extended. In particular, the metafeatures considered, which systematize the vast majority of data characteristics used in earlier work, are of added importance. Hence, this work replicates such metafeatures to be used in the LR approach proposed. Furthermore, the metamodels are also maintained in the experimental procedure in order to allow comparison of results.

In order to understand the systematic metafeatures, one must consider first the framework used. It requires three main elements: object o, function f and post-function pf. The framework applies a function to an object and, afterwards, the post-function to the outcome in order to derive the final metafeature. Thus, any metafeature can be represented using the following notation: $\{o.f.pf\}$ [14].

Consider a matrix R, with rows U and columns I. The objects to be used in the framework are R, U and I. The functions f considered to characterize these objects are: original ratings (*ratings*), count the number of elements (*count*), mean value (*mean*) and sum of values (*sum*). The post-functions pf are maximum, minimum, mean, standard deviation, median, mode, entropy, Gini index, skewness and kurtosis. Additionally, it includes the number of users, items, ratings and the matrix sparsity. This results in 74 metafeatures which were reduced by feature selection, ending up with the following set: *R.ratings.kurtosis*, *R.ratings.sd*, *I.count.kurtosis*, *I.count.minimum*, *I.mean.entropy*, *I.sum.skewness*, *nusers*, *sparsity*, *U.mean.minimum*, *U.sum.kurtosis*, *U.mean.skewness* and *U.sum.entropy*.

2.2 Label Ranking

LR aims to predict for any instance a preference relationship among a finite set of labels or alternatives [12, 17]. Let us consider a finite set of labels $L = \{l_1, l_2, ..., l_n\}$ for which predictions will be made, where n is the total number of labels available. Consider also that a binary preference relation $\succ_x \subseteq L \times L$ allows to dictate the preference associated to an instance $x \subset X$ regarding sets of two labels. For instance, $l_1 \succ_x l_2$ means l_1 is preferred over l_2, when associated with instance x. When all possible preference relations are specified for an instance, then a total strict order (i.e. a ranking) of L is obtained. This ranking, $\pi_x \subset \Omega$, can be seen as a permutation of $\{1, ..., n\}$, where n is the number of labels. Such a permutation requires to fix the positions of all algorithms (for instance by alphabetical order) and to obtain the respective ranking positions. As an example, for $n = 4$, the permutation associated with instance x for the fixed label ordering $\{l_1, l_2, l_3, l_4\}$ is $\{1, 3, 4, 2\}$ and it means that $l_1 \succ_x l_4 \succ_x l_2 \succ_x l_3$. In LR, each instance x is associated with a ranking π_x. The goal for a LR learning algorithm is to find the mapping $g : X \rightarrow \Omega$, such that a loss function in Ω is minimized. Typically, ranking accuracy measures, such as Kendall's tau and Spearman's rank, are used for this purpose [7].

3 LR FOR CF ALGORITHM SELECTION

3.1 Problem Formalization

Our algorithm selection problem requires 4 search spaces [15]:

- the problem space P: set of CF datasets;
- the feature space F: set metafeatures;
- the algorithm space A: set of CF algorithms;
- the performance space Y: performance of all algorithms in a set of suitable evaluation measures.

Using LR, the set of labels $L \in \Omega$, for which predictions will be made, is given by the names of all algorithms $a \in A$. Recall that in order to create the rankings π, the predictive performance of all CF algorithms is assessed regarding a specific evaluation measure. The preference relations $\succ$, which are the basis to the rankings π, are established based on those performance estimates. Therefore, the algorithm selection problem for CF using LR can be defined as follows: for every dataset $p \in P$, with features $f(p) \in F$ associated with the respective rankings π_p, find the selection mapping $g(f(p))$ into the permutation space Ω, such that the selected ranking of algorithms π_p maximizes the performance mapping $y(\pi_p) \in Y$.

3.2 Experimental setup

3.2.1 Baselevel. The baselevel component is concerned with the CF datasets, algorithms and evaluation measures. The 38 datasets used come from different domains, namely Amazon Reviews, Flixter, BookCrossing, Jester, MovieLens, MovieTweetings, Tripadvisor, Yahoo! and Yelp. It should be noted that a domain may contain multiple datasets. The experiments were carried out with MyMediaLite [10]. Two types of CF tasks were addressed: Rating Prediction (RP) and Item Recommendation (IR). The following algorithms were used in this work for RP: MF, Biased MF, Latent Feature Log Linear Model, SVD++, 3 variants of Sigmoid Asymmetric Factor Model, User Item Baseline and Global Average. In the case of IR, the algorithms were BPRMF, Weighted BPRMF, Soft Margin Ranking, WRMF and Most Popular. In IR, the algorithms are evaluated using NDCG and AUC. In RP, the algorithms are evaluated using RMSE and NMAE. All experiments use 10-fold cross-validation. No parameter optimization was conducted. By not tuning the parameters, one prevents biasing the performance results in favor of any algorithm.

3.2.2 Metalevel. To build the metadataset, this work employs the metafeatures (described in Section 2.1) to all 38 CF datasets in order to create the independent variables. The baselevel metrics (RMSE, NMAE, NDCG and AUC) are used to create the metatargets. A different ranking of algorithms for each dataset can be created using each evaluation measure. In essence, since 4 baselevel measures are used, it means that there are 4 different problems. The LR algorithms used are adaptations of well-known learning algorithms: KNN and Naive Bayes (NB) [16], Ranking Tree (RT) and Label Ranking Random Forest (RF) [7] and the baseline Average Rankings. Meta-level performance is measured using Kendall's Tau coefficient with leave-one-out cross-validation. The impact on the baselevel performance is assessed by comparing the performance of the predicted CF algorithms on average for all datasets. Thresholds t are considered to compare a cumulative performance of the methods. Hence, for $t = 1$ only the first algorithm is compared; for $t = 2$, both the first and second algorithms are used, etc.

4 RESULTS AND DISCUSSION

4.1 Metalevel evaluation

Figure 1 presents the Kendall's Tau performance. A plot is presented for each of the four different metatargets and each column represents the performance of a meta-algorithm. It also contains the predictive performance of the baseline Average Rankings for each metatarget, represented as horizontal lines.

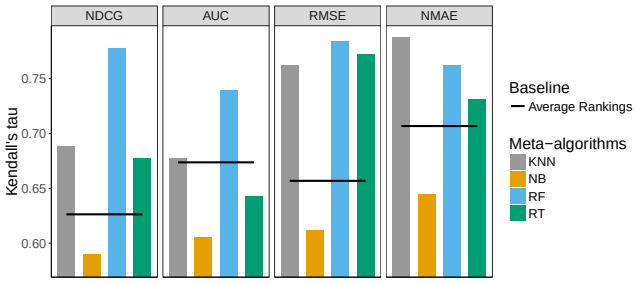


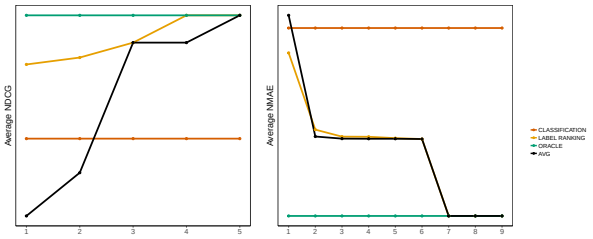
Figure 1: Kendall's Tau ranking accuracy.

The following observations can be made:

- **RF is overall the best algorithm**: except on NMAE.
- **KNN performs mostly well**: except on AUC.
- **RT has inconsistent results**: either it beats the baseline by a large margin or it fails poorly.
- **NB performs the worst**: never beats the baseline.

The statistical significance of algorithm comparison [8] confirms NB is the worst algorithm, while RF is indeed the best algorithm. However, there is no statistical significance in the difference between RT and KNN with regards to the baseline. Thus, RF is the only metamodel which is significantly better than the baseline.

4.2 Baselevel evaluation



(a) NDCG metatarget **(b) NMAE metatarget**

Figure 2: Top-K evaluation for different metatargets

Figure 2 presents the results of the impact on the baselevel performance for the NMAE and NDCG metatargets. For better interpretability, only two LR algorithms are presented: the baseline average rankings and RF. Furthermore, the performance of the LR approach is compared with the classification approach [4]. Given that classification approaches focus on a single label, rather than the whole ranking, one can only assess the impact on the baselevel performance for the first threshold. The remaining thresholds must therefore contain the same value. The results show that:

- **LR has a positive impact**: there is always at least a threshold whose performance beats the baseline. For NDCG, this holds for $t \leq 2$, while for NMAE, only for $t = 1$.
- **LR outperforms the classification approach**: this happens in all metatargets. This result confirms the importance of the proposed approach.

5 CONCLUSIONS AND FUTURE WORK

This work extends the state of the art on algorithm selection for CF by formalizing and experimentally validating an approach for the prediction of the best ranking of algorithms using Label Ranking. Experimental results show that the metamodels created using Label Ranking Random Forest obtain good results in terms of ranking accuracy, with a statistically significant difference with regards to the baseline Average Rankings. They also show that the ranking of CF algorithms predicted has a positive impact on the baselevel performance when compared to those algorithms recommended by the baseline Average Rankings and the traditional classification approach. Possible future work tracks include the choice of other ranking strategies to compare with Label Ranking, exploration of different metafeatures and to study the algorithm selection problem using consensus ranking to obtain an unique metatarget.

Acknowledgments. This work is financed by the Portuguese funding institution FCT - Fundação para a Ciência e a Tecnologia through grant SFRH/BD/117531/2016 and by CNPq and FAPESP.

REFERENCES

[1] Gediminas Adomavicius and Jingjing Zhang. 2012. Impact of data characteristics on recommender systems. *ACM Information Systems* 3, 1 (2012), 1–17.
[2] Pavel Brazdil, Christophe Giraud-Carrier, Carlos Soares, and Ricardo Vilalta. 2009. *Metalearning: Applications to Data Mining* (1 ed.). Springer.
[3] Pavel Brazdil, Carlos Soares, and Joaquim da Costa. 2003. Ranking Learning Algorithms : Using IBL and Meta-Learning on Accuracy and Time Results. *Machine Learning* 50, 3 (2003), 251–277.
[4] Tiago Cunha, Carlos Soares, and André C.P.L.F. de Carvalho. 2016. Selecting Collaborative Filtering algorithms using Metalearning. In *European Conference on Machine Learning and Knowledge Discovery in Databases*. 393–409.
[5] Tiago Cunha, Carlos Soares, and André C.P.L.F. de Carvalho. 2018. Metalearning and Recommender Systems: A literature review and empirical study on the algorithm selection problem. *Information Sciences* (2018), 128–144.
[6] Tiago Cunha, Carlos Soares, and Andre C. P. L. F. de Carvalho. 2017. Recommending Collaborative Filtering algorithms using subsampling landmarkers. In *Discovery Science*. 189–203.
[7] Cláudio Rebelo de Sá, Carlos Soares, Arno Knobbe, and Paulo Cortez. 2016. Label Ranking Forests. *Expert Systems* (2016).
[8] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research* 7 (2006), 1–30.
[9] Michael Ekstrand and John Riedl. 2012. When Recommenders Fail: Predicting Recommender Failure for Algorithm Selection. *ACM RecSys* (2012), 233–236.
[10] Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2011. MyMediaLite. In *ACM RecSys*. 305–308.
[11] Josephine Griffith, Colm O'Riordan, and Humphrey Sorensen. 2012. Investigations into user rating information and predictive accuracy in a collaborative filtering domain. In *ACM Symposium on Applied Computing*. 937–942.
[12] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. 2008. Label ranking by learning pairwise preferences. *Artificial Intelligence* 172, 16-17 (2008), 1897–1916.
[13] Pawel Matuszyk and Myra Spiliopoulou. 2014. Predicting the Performance of Collaborative Filtering. In *Web Intelligence, Mining and Semantics*. 38:1–38:6.
[14] Fábio Pinto, Carlos Soares, and João Mendes-Moreira. 2016. Towards automatic generation of Metafeatures. In *PAKDD*. 215–226.
[15] John Rice. 1976. The Algorithm Selection Problem. *Advances in Computers* 15 (1976), 65–118.
[16] Carlos Soares. 2015. *labelrank: Predicting Rankings of Labels*. https://cran.r-project.org/package=labelrank
[17] Shankar Vembu and Thomas Gärtner. 2010. Label ranking algorithms: A survey. In *Preference Learning*. 45–64.