All You Need is Ratings: A Clustering Approach to Synthetic Rating Datasets Generation

Diego Monti[a], Giuseppe Rizzo[b] and Maurizio Morisio[a]

[a]Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy
[b]LINKS Foundation, Via Pier Carlo Boggio 61, 10138 Turin, Italy

Introduction

Because of the shortage of public datasets, practitioners have started to rely on synthetic ratings in order to conduct their offline experiments [1]. An obvious advantage of such an approach is that it enables the creation of rating datasets with an arbitrary number of users and items at a limited cost of dataset acquisition. However, the results obtained from such experiments may be questionable, as the generated datasets are usually not capable of capturing the characteristics of a particular domain of interest [2]. For example, different generative approaches rely on descriptive statistics, like mean and standard deviation. In this work, we propose a novel approach for automatically generating synthetic datasets with a configurable number of users leveraging on a reference dataset that is used as the seed of the process and that encodes the peculiarities of a domain of interest. Such a generative method can be exploited to create different rating datasets containing users that exhibit behaviors similar to the ones available in the reference dataset. However, the synthetic users do not have a direct relation with the real users and, therefore, no private or commercially sensible information is leaked. At the same time, because the number of synthetic users is configurable, the generated dataset can be exploited to conduct scalability tests in a realistic way and to train recommendation algorithms using reinforcement learning methods.

Dataset Generation

Our approach for generating synthetic datasets starting from a reference dataset consists of two steps. In the first one, it is necessary to analyze an existing collection of user preferences in order to obtain an accurate representation of the domain of interest. Then, in the second one, it is possible to exploit such a representation for creating different generated datasets.
We argue that only relying on a few statistical distributions computed empirically at a global level from an existing dataset or specified by a researcher is not sufficient to realistically simulate the individual tastes of human beings. Such methods would lead to the creation of datasets with users having no individual preferences, thus making the task of any recommender system nearly impossible.
For this reason, we included a preliminary clustering phase as part of the first step in order to group the users in a fixed number of communities. The individual rating behaviors, represented by different statistical distributions, are learned for each community and then exploited during the sampling phase.
For simplicity, we assume that each user can only express positive preferences about the items available in the system. However, this approach can also be exploited to simulate datasets with ratings expressed on a more complex scale by repeating these steps for each rating value and then by merging the results.

Algorithms

The user clustering and distribution learning process is formalized in Algorithm 1. We represent each user $v \in \mathcal{U}$ from the reference dataset as a vector with length equal to the number of items $|\mathcal{I}|$. Given this data structure, we decided to apply the K-means clustering algorithm [3] to group together users who liked a similar set of items in K different clusters. Every cluster identifies a different community of users.
We create the following empirical distributions from the reference ratings:

- P^C, how users are distributed in K clusters;
- P_k^U, how ratings are distributed in $|\mathcal{U}|$ users for each cluster;
- P_k^I, how ratings are distributed in $|\mathcal{I}|$ items for each cluster.

Starting from the empirical distributions obtained from Algorithm 1, it is possible to generate a synthetic dataset by applying to them a sampling function σ. The rating sampling procedure is formalized in Algorithm 2.

Require: $\mathcal{U} \neq \{\emptyset\} \wedge K > 0 \wedge K \leq |\mathcal{U}|$
$\mathcal{C} \leftarrow \text{K-means}(\mathcal{U}, K)$
$P^C \leftarrow P(v \in \mathcal{C}_k)$
for all $k \in \{1, \ldots, K\}$ **do**
  $P_k^U \leftarrow P(\rho_v | v \in \mathcal{C}_k)$
  $P_k^I \leftarrow P(\rho_\iota | \iota \in \mathcal{I}_v \wedge v \in \mathcal{C}_k)$
end for
return P^C, P_k^U, P_k^I

Algorithm 1: User clustering and distribution learning.

Require: $U > 0, P^C, P_k^U, P_k^I$
$\mathcal{R} \leftarrow \{\emptyset\}$
for all $u \in \{1, \ldots, U\}$ **do**
  $k \leftarrow \sigma(P^C)$
  $I \leftarrow \sigma(P_k^U)$
  for all $i \in \{1, \ldots, I\}$ **do**
    $\rho_{u,i} \leftarrow \hat{\sigma}(P_k^I)$
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{\rho_{u,i}\}$
  end for
end for
return $\mathcal{R}$

Algorithm 2: Rating sampling.

Experimental Results

We compared the results obtained from the evaluation of different recommenders conducted on popular datasets typically exploited in literature with the ones computed in the same experimental conditions using various collections of synthetic preferences generated starting from them using multiple techniques.
The results obtained with MovieLens 100K are available in Table 1. We observe that the relative order of the measures is the same between the generated and the reference datasets.

Table 1: The results obtained with the baseline, generated, and reference versions of MovieLens 100K.

	Baseline dataset			Generated dataset			Reference dataset		
Algorithm	Precision	Recall	NDCG	Precision	Recall	NDCG	Precision	Recall	NDCG
Random	0.009416	0.008877	0.009841	0.009847	0.008977	0.010022	0.007743	0.006300	0.008183
Most Popular	0.060065	0.053209	0.064384	0.099672	0.083875	0.110229	0.112759	0.102804	0.130632
User KNN	0.055952	0.050587	0.058744	0.154158	0.135917	0.169499	0.205234	0.221684	0.233362
BPRMF	0.045346	0.033628	0.048740	0.122538	0.106186	0.129742	0.182770	0.186838	0.198869
WRMF	0.047078	0.042876	0.048104	0.164114	0.144272	0.173916	0.221592	0.233235	0.250386

Conclusion and Future Work

In this paper, we have discussed a method for generating synthetic datasets with an arbitrary number of users starting from existing collections of preferences. Differently from the approaches already available in literature, we propose to first model user communities in order to generate more realistic ratings that can be successfully exploited during an evaluation campaign.
As future work, we would like to explore additional methods for creating synthetic datasets. We believe that Generative Adversarial Networks (GANs) could be successfully exploited for this task, as they are already used to generate fake images. Such approaches would require the definition of a way for representing the preferences of a user similarly to an image.

References

[1] Hsiang-Fu Yu, Cho-Jui Hsieh, Si Si, and Inderjit Dhillon. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. In *2012 IEEE 12th International Conference on Data Mining*, pages 765–774, New York, NY, USA, dec 2012. IEEE.

[2] Nikos Manouselis and Constantina Costopoulou. Preliminary study of the expected performance of MAUT collaborative filtering algorithms. In *The Open Knowlege Society. A Computer Science and Information Systems Manifesto*, pages 527–536. Springer Publishing, New York, NY, USA, 2008.

[3] John A. Hartigan and Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100, 1979.