International Journal of Approximate Reasoning 95 (2018) 62–76

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Collaborative Topic Model for Poisson distributed ratings ☆

Hoa M. Le[a], Son Ta Cong[b], Quyen Pham The[b], Ngo Van Linh[a], Khoat Than[a,*]

[a] *School of Information & Communication Technology, Hanoi University of Science and Technology, No. 1, Dai Co Viet road, Hanoi, Viet Nam*
[b] *VCCorp, Hanoi, Viet Nam*

ARTICLE INFO

Article history:
Received 30 May 2017
Received in revised form 24 January 2018
Accepted 4 February 2018
Available online 7 February 2018

Keywords:
Recommender system
Content-based collaborative filtering
Probabilistic matrix factorization
Topic models

ABSTRACT

We present Collaborative Topic Model for Poisson distributed ratings (CTMP), a hybrid and interpretable probabilistic content-based collaborative filtering model for recommender system. The model enables both content representation by admixture topic modelling, and computational efficiency from Poisson factorization living together under one tightly coupled probabilistic model, thus addressing the limitation of previous methods. CTMP excels in predictive performance under different real-world recommendation contexts, and easily scales to big datasets, while recovering interpretable user profiles. Moreover, our empirical study also shows strong evidence that sparsity in the estimates of topic mixture can be recovered via learning, despite not being specified in the model. The sparse representation derived from CTMP would allow efficient storage of the item contents, consequently providing a computational advantage for other tasks in industrial settings.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

Recommender systems have taken an important part in daily life and business. User data, commonly come in the form of explicit or implicit ratings (e.g. likes, clicks), enable companies to analyze their customer preferences more thoroughly, but at the same time are also massive. Thus, it is required for any recommender systems to *scale* efficiently with the big and highly sparse data in real-world. Moreover, the *interpretability* of the recommendation is also not to be ignored, in which valuable insight can be gained from the interpretation to address downstream analyses.

Traditionally, there are two frameworks in “recommendation from user data”: content-based and collaborative filtering (CF). On their own, methods from CF framework has long been the industry standard [1], and among those, probabilistic matrix factorization (PMF) [2] gained the most success. Nevertheless, all CF methods cannot recommend items that are not rated by any users i.e. cannot address cold-start problem, while the purely content-based methods do not have such an issue. Therefore, hybrid methods that can combine the strength of both frameworks are desirable.

Incorporating a (probabilistic) model of content to matrix factorization has risen as the most promising hybrid methods [3–6]. In particular, by representing item content with, for example, topic models [7], the methods can also enjoy the interpretable semantics of the latent space characterized by the topic mixtures, which in turn makes the semantics of

☆ This paper is part of the Virtual special issue on Uncertainty Management in Machine Learning Applications, Edited by Van-Nam Huynh.

* Corresponding author.

E-mail addresses: hoamle@outlook.com (H.M. Le), congson1293@gmail.com (S. Ta Cong), phamthequyen93@gmail.com (Q. Pham The), linhnv@soict.hust.edu.vn (N. Van Linh), khoattq@soict.hust.edu.vn (K. Than).

https://doi.org/10.1016/j.ijar.2018.02.001
0888-613X/© 2018 Elsevier Inc. All rights reserved.

item latent factor more interpretable. However, there still exists limitation in predictive performance and/or computational expense.

In this paper, we propose ***C**ollaborative **T**opic **M**odel for **P**oisson* distributed ratings (CTMP), a probabilistic hybrid model with scalability and interpretability, and make the following contributions.

1. We alleviate the flexibility of latent Dirichlet allocation (LDA) [8], and scalability from Poisson factorization [9] to model ratings under one unified model, which relevant works had only partially address. CTMP has been evaluated in different contexts, including real-world recommendation tasks in the industry, and performs substantially better than existing models. The predictive performance excels in scientific article and commercial product recommendation, while is also competitive with news data.
2. We develop a co-ordinate ascent algorithm that can fit our non-conjugate model, which is fast and scalable.
3. Our empirical studies provide strong evidence showing that sparse estimates of topic mixtures can also be achieved via learning, even though the model specification does not encourage so. Recovering sparse topic mixtures is desirable in text modelling, for a document only talks about a handful number of salient topics in real life [10]. Moreover, this sparsity is particularly important in industrial settings, since number of topics can goes up to order of thousands, accordingly to the data a company has [11]. Finally, sparsity provides a compact and interpretable content representation which is efficient to store. It largely reduces memory requirement, thus is beneficial for other recommendation tasks, such as recommending products under the same topics/categories in which the users are interested in near real-time.

In the rest of this paper, we briefly review related work in Section 2; followed by the details of CTMP formalization, learning, and prediction in Section 3; empirical evaluation in Section 4; and lastly conclude in Section 5.

2. Related work

Jointly modelling content into matrix factorization has been an active study in recommender system literature. Agarwal and Chen [3] first introduced topic models in matrix factorization with fLDA, in which the item latent factor takes the role of topic proportion in the LDA [8] representation. However, the setting has a limitation in distinguish items that have similar topic mixture, but content details (which topic mixture cannot capture) are of interest to different groups of user. CTR [4] addresses the issue by making the item latent factor be an offset from topic proportion, thus enables the item latent factor to also capture contribution of user ratings. CTR has shown significant improvement over fLDA, and there have been numbers of variants that applied CTR's treatment to item factor (e.g. [12,5,6]).

Still, CTR has a severe computational limitation. CTR assumes that the ratings follow the Gaussian distribution, and therefore the training requires iterating all of the entries in the rating matrix. It is highly inefficient, especially when working with massive and highly sparse real-world datasets. To address the problem, Gopalan et al. [9,5] extended Poisson factorization with CTPF, and modelled the ratings by Poisson distribution. CTPF only concerns non-zero ratings when training, and is much more efficient and scalable. However, CTPF models content generation by standard mixtures of Gamma, in order to ensure that the augmented model is conditionally conjugate and has closed-form updates.

In this paper, we address CTR computational limitation by modelling Poisson ratings as CTPF does, while modelling contents with LDA. Also, in CTPF, the authors endorsed *sparsity* in content representation via the model parameter priors, while we achieve that via learning. In CTR, while it's also possible to endorse sparsity in the content representation, the authors, however, used a hyperparameter setting that discouraged the property.

It is worth noting that there are recent works that modelled contents with deep neural networks, and attempted to represent topic intensities by a hidden layer (e.g. [12,6]). However, alongside with expensive training, established and comprehensive studies of the representation semantics and/or sparsity are not presented in any deep-learning-based methods at the time of this writing. Therefore, we do not include the methods in the scope of this paper.

3. A Collaborative Topic Model for Poisson distributed ratings

In this section, we introduce a new generative model for both implicit rating and content. Then, we explicitly describe learning and prediction phases, and key properties of CTMP.

Recommender systems often use feedbacks (e.g., views, clicks, likes, purchase histories) from users to make recommendations of items (e.g., news, products, songs, clips,...). The feedbacks may be explicitly or implicitly provided by users (see Table 1(a)). Some systems also use the textual content, such as product's description or news' content, to further understand the user's preference and then make accurate recommendation. Table 1 provides an example of users' feedbacks and item contents. Collaborative filtering based systems mainly use users' feedbacks alone, whereas a hybrid system can use both feedbacks and contents. Our model (CTMP) will use both kinds of data for accurate recommendation.

Download English Version:

https://daneshyari.com/en/article/6858801

Download Persian Version:

https://daneshyari.com/article/6858801

Daneshyari.com