Information Sciences 405 (2017) 175–189

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Link sign prediction by Variational Bayesian Probabilistic Matrix Factorization with Student-t Prior[☆]

Yisen Wang[a], Fangbing Liu[a], Shu-Tao Xia[a,b], Jia Wu[c,*]

[a] Department of Computer Science and Technology, Tsinghua University, China
[b] Graduate School at Shenzhen, Tsinghua University, China
[c] Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

ARTICLE INFO

Article history:
Received 7 April 2016
Revised 9 April 2017
Accepted 11 April 2017
Available online 12 April 2017

Keywords:
Signed networks
Matrix factorization
Student-t distribution

ABSTRACT

In signed social networks, link sign prediction refers to using the observed link signs to infer the signs of the remaining links, which is important for mining and analyzing the evolution of social networks. The widely used matrix factorization-based approach – Bayesian Probabilistic Matrix Factorization (BMF), assumes that the noise between the real and predicted entry is Gaussian noise, and the prior of latent features is multivariate Gaussian distribution. However, Gaussian noise model is sensitive to outliers and is not robust. Gaussian prior model neglects the differences between latent features, that is, it does not distinguish between important and non-important features. Thus, Gaussian assumption based models perform poorly on real-world (sparse) datasets. To address these issues, a novel Variational Bayesian Probabilistic Matrix Factorization with Student-t prior model (TBMF) is proposed in this paper. A univariate Student-t distribution is used to fit the prediction noise, and a multivariate Student-t distribution is adopted for the prior of latent features. Due to the high kurtosis of Student-t distribution, TBMF can select informative latent features automatically, characterize long-tail cases and obtain reasonable representations on many real-world datasets. Experimental results show that TBMF improves the prediction performance significantly compared with the state-of-the-art algorithms, especially when the observed links are few.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Signed social networks have increasingly become an actively researched field in social networks [6,44–46,48,49]. Each link of the signed network has a sign, expressing the attitude from the generator to the receiver, which can be positive (stating trust or approval) or negative (representing distrust or disapproval). Link sign prediction is defined as using the observed link signs to predict the unknown signs of the remaining links [15]. For an online retailer, *e.g.*, Amazon, the link sign can represent whether the user likes the item or not, so link sign prediction is significant for recommending items to users. It is the same as the video website, *e.g.*, Netflix and Hulu. The relations between users can also be more informative

☆ The first two authors contributed equally to this study. This research is supported in part by the National Natural Science Foundation of China (61371078), and the R&D Program of Shenzhen (Nos. JCYJ20140509172959977, JSGG20150512162853495, ZDSYS20140509172959989).

* Corresponding author.
E-mail addresses: wangys14@mails.tsinghua.edu.cn (Y. Wang), lfb13@mails.tsinghua.edu.cn (F. Liu), xiast@sz.tsinghua.edu.cn (S.-T. Xia), jia.wu@uts.edu.au (J. Wu).

http://dx.doi.org/10.1016/j.ins.2017.04.014
0020-0255/© 2017 Elsevier Inc. All rights reserved.

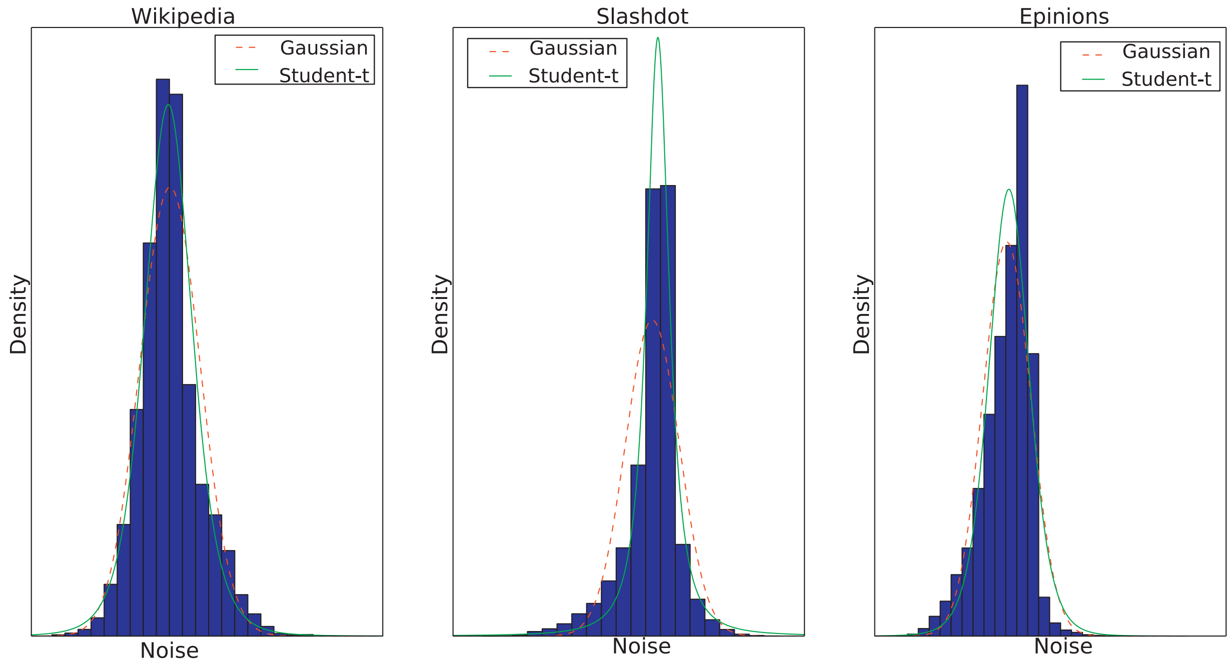

Fig. 1. Noise fitting results of different distributions. Compared to Gaussian distribution, Student-t distribution has higher kurtosis and owns better representation ability for noise. The figure is best viewed in color.

by allowing users to label each other as friends or enemies, compared to traditional social networks which only contain friend relations. Generally, there are two kinds of methods for solving link sign prediction, *i.e.*, the node similarity-based method [13,36] and the matrix factorization-based method [1,12]. The former not only relies on the completeness of the network structure, but also depends on high level features derived from social psychology theory [7,15,16,40]. The matrix factorization-based method overcomes the incompleteness of the network, and eliminates the need to extract structure features from limited observed links [4,26], which will be the focus of this work.

A signed social network is modeled in a matrix format $S \in \Re^{N \times M}$ with N users and M items. Each entry s_{ij} is defined as the attitude from the generator i to the receiver j, where $s_{ij} = 1$ for positive attitude, $s_{ij} = -1$ for negative attitude and $s_{ij} = 0$ for unknown cases. That is to say, s_{ij} is nonzero if the link sign is observed and zero otherwise. In matrix factorization models, the sign matrix $S \in \Re^{N \times M}$ can be factorized into the user latent feature matrix $U \in \Re^{K \times N}$ and the item latent feature matrix $V \in \Re^{K \times M}$, where $K \ll N, M$. The goal is to find appropriate U and V to minimize the discrepancy between S and $U^T V$, therefore, the noise is defined as the absolute value of $s_{ij} - u_i^T v_j$. A number of algorithms have been proposed to solve this problem, such as Variational Bayesian Matrix Factorization (VBMF) [17], Probabilistic Matrix Factorization (PMF) [30], Bayesian Probabilistic Matrix Factorization (BMF) [29] and Robust Bayesian Matrix Factorization (RBMF) [14]. We find that they all assume that the noise is Gaussian noise, while in real-world datasets, Gaussian distribution is not robust/reasonable and cannot characterize long-tail cases, which can be illustrated by the following experiments. On three widely used datasets (*i.e.*, *Wikipedia*, *Slashdot* and *Epinions*) in the signed social network literature [16], we plot the histogram of the noise $|s_{ij} - u_i^T v_j|$ fitted by different distributions. Fig. 1 shows that the solid curves have higher peaks and heavier tails than the dashed curves, meaning that the solid curves can better fit the noise on real-world datasets. In other words, Student-t distribution is more suitable and more robust in fitting the noise than Gaussian distribution [37].

The fraction of the observed links with regard to the above three datasets is less than 1% (refer to Section 6.1), which is very sparse. As for the prior assumption of latent features, VBMF, PMF and BMF all adopt the multivariate Gaussian prior, which does not distinguish the importance of the latent features, making them perform poorly on the sparse datasets. Moreover, a study of the posteriors obtained by BMF on standard movie recommendation benchmarks also shows that the tails are significantly heavier than Gaussian distribution [14]. Thus, heavy-tailed distribution should be considered for the prior of latent features. To this end, RBMF has tried to assume that the prior of latent features is multivariate Student-t distribution, which obtains significant performance improvement. However, RBMF fixes the mean of Student-t distribution as 0, and still assumes that the noise is Gaussian noise.

To address the above issues, we propose a Variational Bayesian Probabilistic Matrix Factorization with Student-t prior model (TBMF), using Student-t distribution to describe both the noise and the prior of latent features. A univariate Student-t distribution is used to fit the prediction noise, and a multivariate Student-t distribution is adopted for the prior of latent features. A variational Bayesian inference is employed to estimate the parameters and find the solution that maximizes the posterior probability. In summary, our contributions are as follows:

Download English Version:

https://daneshyari.com/en/article/4944473

Download Persian Version:

https://daneshyari.com/article/4944473

Daneshyari.com