Physica A 438 (2015) 454–468

Contents lists available at ScienceDirect

Physica A

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

An online expectation maximization algorithm for exploring general structure in massive networks

Bianfang Chai [a,b], Caiyan Jia [a,*], Jian Yu [a,*]

[a] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China
[b] Department of Information Engineering, Shijiazhuang University of Economics, Hebei 050031, China

HIGHLIGHTS

- An efficiency online variational EM algorithm (noted as onlineVEM) was designed.
- The objective function of the onlineVEM was proved to be additive.
- Model parameters were estimated by differentiating the objective function.
- The onlineVEM was tested to be robust, accurate and efficient.

ARTICLE INFO

Article history:
Received 1 October 2014
Received in revised form 18 April 2015
Available online 14 July 2015

Keywords:
Complex networks
General structure exploration
Online expectation maximization algorithm

ABSTRACT

Mixture model and stochastic block model (SBM) for structure discovery employ a broad and flexible definition of vertex classes such that they are able to explore a wide variety of structure. Compared to the existing algorithms based on the SBM (their time complexities are $O(mc^2)$, where m and c are the number of edges and clusters), the algorithms of mixture model are capable of dealing with networks with a large number of communities more efficiently due to their $O(mc)$ time complexity. However, the algorithms of mixture model using expectation maximization (EM) technique are still too slow to deal with real million-node networks, since they compute hidden variables on the entire network in each iteration. In this paper, an online variational EM algorithm is designed to improve the efficiency of the EM algorithms. In each iteration, our online algorithm samples a node and estimates its cluster memberships only by its adjacency links, and model parameters are then estimated by the memberships of the sampled node and old model parameters obtained in the previous iteration. The provided online algorithm updates model parameters subsequently by the links of a new sampled node and explores the general structure of massive and growing networks with millions of nodes and hundreds of clusters in hours. Compared to the relevant algorithms on synthetic and real networks, the proposed online algorithm costs less with little or no degradation of accuracy. Results illustrate that the presented algorithm offers a good trade-off between precision and efficiency.

1. Introduction

Many systems in the biological, physical, computer and social sciences are widely modeled as complex networks due to a concise mathematical representation of the topology of systems. These networks possess large size, complex structure

* Corresponding authors.
E-mail addresses: chaibianfang@163.com (B. Chai), cyjia@bjtu.edu.cn (C. Jia), jianyu@bjtu.edu.cn (J. Yu).

http://dx.doi.org/10.1016/j.physa.2015.07.001
0378-4371/

and volatile topology. Social scientists have repeatedly pointed out that identifying features underlying networks often helps people understand and utilize them. One of the most important feature is the structural feature that nodes can be partitioned into clusters such that nodes in each cluster have similar patterns of connection to other nodes. Community structure is one typical example of this structural feature.

Most well known community detection algorithms mainly detect tightly knit clusters, in which the pattern of connection between nodes refers that nodes in a cluster connect to nodes in the same cluster more possibly than ones in different clusters. This type of structure is also called homophily or assortative mixing [1,2]. But in many cases, we do not know what types of structure are there in a network, whether one or several types of structure exist, such as hierarchical structure [3], assortative mixing [4], disassortative mixing [1], a blend of assortative mixing and disassortative mixing [5–7], star structure [8]. Thus, community detection approaches are invalid for detecting structural feature when a network has no community structure or has other types of structure.

Enormous effective models have been employed to detect general structure to circumvent the above issues [1,9–14,7]. The existing models are classified as two categories. One category is the stochastic block model (SBM) [10–12] and its variants, including the degree-corrected SBM [13], the general stochastic block (GSB) model [14], the popularity and productivity stochastic block (PPSB) model [7]. The other category is mixture models for exploring a very broad range of types of structure [1,9]. All of these probabilistic models define a flexible and general similar pattern of connection to encompass an infinite variety of structural types, under the condition that nodes belong to a cluster if they connect to other nodes similarly. The algorithms for models based on the SBM have the $O(mc^2)$ time complexity. But algorithms for mixture models [1,9] have the $O(mc)$ time complexity. In real applications, many real networks are large sizes and have a lot of clusters. The models based on the SBM under the $O(mc^2)$ time complexity are not scalable to deal with these large networks with a large number of clusters. On the contrary, the algorithm of mixture models is better due to the linear time complexity on the number of communities. But it still cannot handle networks with millions of nodes. This inspires us to improve the EM algorithm of mixture models such that it scales both in the number of communities and in the number of edges.

Three kinds of algorithms that scale to the number of edges have been developed to perform structure detection on large networks, including fast community detection algorithms by removing nodes or edges [15,16], stochastic variational inference algorithms by using stochastic optimization [17–19], and online EM algorithms by updating model parameters sequentially [20,21]. The goal of these algorithms aims to improve the efficiency of structure detection with no or less accuracy degradation. The first kind of algorithms sparsify networks by removing some nodes or edges and run the algorithms on the incomplete networks [15,16]. However, these algorithms are only designed for detecting communities and are unable to provide patterns of connection between clusters directly. The second kind of algorithms are based on variational Bayes, which approximates posterior distributions of parameters by a newly developed stochastic optimization method, named stochastic variational inference [22]. These algorithms include the parsimonious triangular model (PTM) [17], the assortative mixed-membership stochastic blockmodel (a-MMSB) [18], the assortative MMSB with node popularities (AMP) [19]. Although these algorithms are fast, they aim to improve variational Bayes algorithm and are all used to detect traditional densely-connected communities. The third kind of algorithms are recursive online versions for classical batch EM algorithms with the assumption that the data set grows over time. The algorithms are also used to deal with large static data incrementally. In this situation, the data are processed object by object instead of using the entire data set to infer hidden variables. In other words, the hidden variables are inferred by only using their related local data. Existed online EM algorithms for networked data exploring are based on traditional online algorithms for large i.i.d. data set [23,24]. Several classical online EM algorithms have been presented to deal with large networks, including the online classification EM algorithm for affiliation model (online CEM) [20], the online SAEM algorithm and the online variational EM algorithm for the SBM [21]. Among these online EM algorithms, the online variational EM method for the SBM has been testified to provide the best performance in terms of precision and efficiency [21]. The batch EM algorithm for mixture models has been demonstrated to have the same ability of general structure detection as the SBM [25], while its complexity is less than the one of EM algorithms for the SBM. As far as we know, there is no online algorithm of mixture models for exploring general structure. Thus, it is necessary to design an online algorithm to further speed up the EM algorithm of mixture models.

The rest of this paper is organized as follows. In Section 2, we describe mixture models for exploring network structure, and provide the inference for parameter estimation. In Section 3, we propose an online variational EM algorithm for mixture models. In Section 4, a criterion is provided to choose the number of communities. In Section 5, the proposed online algorithm is tested on some synthetic networks and real networks. Finally, the conclusions are discussed in Section 6.

2. Mixture models for general structure detection

In this section, we describe the original mixture model provided by Newman and Leicht [1] and its extended version provided by Ramasco and Mungan [9].

2.1. Original mixture model

The mixture model provided by Newman and Leicht [1] aims to deduce the cluster assignments of nodes in a network by fitting the model to an observed directed network. Given a network with N fixed nodes, the network is partitioned into c