Neurocomputing 226 (2017) 192–199

Contents lists available at ScienceDirect

Neurocomputing

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

A parameter selection method of the deterministic anti-annealing algorithm for network exploring

Bianfang Chai[a,c,*], Jinghong Wang[b], Jian Yu[c]

[a] Department of Information Engineering, Hebei GEO University, Hebei 050031, China
[b] College of Information Technology, Hebei Normal University, Shijiazhuang 050024, China
[c] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Communicated by Zidong Wang

Keywords:
Mixture model
Community detection
Deterministic anti-annealing EM algorithm
Convergence rate
Jacobian matrix

ABSTRACT

The traditional expectation maximization (EM) algorithm for the mixture model can explore the structural regularities of a network efficiently. But it always traps into local maxima. A deterministic annealing EM (DAEM) algorithm is put forward to solve this problem. However, it brings about the problem of convergence speed. A deterministic anti-annealing expectation maximization (DAAEM) algorithm not only prevents poor local optima, but also improves the convergence speed. Thus, the DAAEM algorithm is used to estimate parameters of the mixture model. This algorithm always sets its initial parameter β_0 by experience, which maybe get trapped into meaningless results due to too small β_0, or converge to local maxima more frequently due to too large β_0. A parameter selection method for β_0 is designed. In our method, the convergence rate of the DAAEM algorithm for mixture model is first derived from Jacobian matrix of the posterior probabilities. Then the theoretical lower bound of β_0 is computed based on the convergence rate at meaningless points. In our experiments we select β_0 by rounding up the lower bound to the nearest tenth. Experiments on real and synthetic networks demonstrate that the parameter selection method is valid, and the performance of the DAAEM algorithm beginning from the selected parameter is better than the EM and DAEM algorithms for mixture model. In addition, we find that the convergence rate of the DAAEM algorithm is affected by assortative mixing by degree of a network.

1. Introduction

Networks have gained significant attention for representing complex systems, and analyzing them helps us understand the systems. When we have no prior on networks, it is necessary to analyze them by automatics. Many analysis techniques for these networks have emerged in the past few years, and community detection [1] is a popular one. It has become useful for many reasons, such as suppressing the complexity of the whole network and identifying the key nodes in networks, etc.

Up to now, a huge amount of methods for the task of community detection have been developed, including hierarchical clustering, divisive clustering, modularity-based methods, etc. They just focus on detecting tightly connected subgraphs. But complex networks may have many other types of structures, including core-periphery, hierarchical, multipartite structures, or the mixture of them, etc. Recently, some models have been provided to detect a more wide variety of structures besides tightly connected subgraphs. These models are mainly classified by two categories. One category is ones based on the stochastic block model (SBM) [2], whose algorithms estimate parameters by the Gibbs sampling method [3], the variational EM algorithm [4–6], the variational Bayes methods [7], the belief propagation method [8], etc. The time complexities of these algorithms are approximately $O(mc^2)$, where m and c respectively denote the number of edges and clusters. The other category is mixture model [9] for network exploring, whose time complexity is $O(mc)$. By contrast, the EM algorithm for mixture model (EMMM) [9] is more efficient than the algorithms for models based on the SBM.

However, it is well known that the traditional EM algorithm always converges to poor local maxima. The DAEM algorithm [10] has been provided to overcome local maximum problem. It starts with $\beta = \beta_0 \simeq 0$ and slowly increases β to 1. At each β, the DAEM algorithm executes the EM algorithm. This increases the convergence time, especially on data with skewed mixing coefficients and large overlap among clusters. The anti-annealing EM algorithm [11] starts from $\beta > 1$ and slowly decreases it down to 1. It improves the speed by restricting the amount of overlaps, but trends to converge to poor local optima more frequently. The DAAEM algorithm [11] not only prevents the EM algorithm from getting trapped into local optima, but also

* Corresponding author at: Department of Information Engineering, Hebei GEO University, Hebei 050031, China.
E-mail address: chaibianfang@163.com (B. Chai).

http://dx.doi.org/10.1016/j.neucom.2016.11.050
Received 14 September 2015; Received in revised form 16 August 2016; Accepted 27 November 2016
Available online 30 November 2016

improves the speed of convergence. It starts from a small parameter $\beta = \beta_0 \simeq 0$, and slowly increases β beyond 1, up to a chosen upper bound, and finally slowly decreases β down to 1.

In order to make the parameter estimating algorithm for mixture model [9] converge to a better local optimum or an approximate global optimum efficiently, we use the framework of the DAAEM algorithm to estimate parameters of mixture model. The DAAEM algorithm for mixture model is noted as DAAEMMM for short. As the DAEM algorithm, the DAAEM algorithm needs to set an initial β_0. In the literature [10,11], the initial β_0 is always set experimentally. Ueda et al. [10] confirmed that $\beta_0 = 0.1$ may be small enough, and β in the new iteration is set as constant (constant is always set as $1.1 \sim 1.5$) times of β in the last iteration. Naim et al. [11] thought that it was necessary to select a smaller β_0 for complex data, while a larger β_0 for simple data. There is no study on how to set it theoretically. If we set a too small β_0, the DAAEMMM algorithm may divide nodes into meaningless clusters [12], such as the partition that each node belongs to each cluster with equal probabilities. Once the algorithm converges to this case, it is unable to escape from this point in subsequent iterations. If we set a too large β_0, the DAAEMM algorithm is easier to converge to local maxima. From the aspect of convergence property, the algorithm beginning with the selected initial parameter β_0 should not converge to a meaningless clustering result point, which should be an unstable fixed point of the DAAEMMM algorithm.

Some scientists have contributed to study fixed points of algorithms by convergence analysis. The convergence rate is used to measure whether a fixed point is stable. Parameters embedded in the convergence rate are estimated by making meaningless fixed points be unstable. Hessian matrix and Jacobian matrix are two popular tools to compute the convergence rate. Lei Xu and Michael I. Jordan presented Hessian matrix of the log-likelihood function for Gaussian mixture with respect to the collection of mixture parameters [13]. Jian Yu et al. [14] executed the optimality test by computing Hessian matrix of parameter mapping of the Fuzzy c-means algorithm. The formulas of the Hessian matrix are too complicated to analyze the convergence properties. Jacobian matrix has the same ability as the Hessian matrix to judge whether an algorithm converges to a local maximum. Chaomurilige et al. [15] provided a theoretical method for selecting the fuzziness parameter in the Gustafson-Kessel algorithm using Jacobian matrix analysis. It has been demonstrated that Jacobian matrix is a more effective tool for convergence analysis, which is used to select the parameters of an algorithm by making the convergence rate at meaningless fixed points be not less than 1. But there is no study on how to select β_0 parameter of the DAAEM or DAEM algorithm using Jacobian matrix. Here we provide a theoretical parameter selection method for the initial parameter β_0 of the DAAEMMM algorithm, which is also suitable to the DAEM algorithm for mixture model (DAEMMM). In our method, the convergence rate of the DAAEMMM algorithm is computed based on the Jacobian matrix of the posterior probabilities. Then, a theoretical lower bound of parameter β_0 is captained by analyzing the stability of fixed points according to the convergence rate. The initial parameter β_0 is set according to this theoretical lower bound value. In our experiments, β_0 are selected by rounding up the theoretical lower bound to the nearest tenth.

In this paper, we first use the DAAEM algorithm to estimate parameters of mixture model for network exploring. Then we design a method to get the theoretical minimal β_0 for the DAAEMMM algorithm based on Jacobin matrix analysis. In outline, the paper is organized as follows. In Section 2, we give a brief review of the EMMM algorithm for exploratory analysis in networks. In Section 3, we analyze the disadvantages of the EMMM algorithm on several real networks, and then present a DAAEMMM algorithm for general structure detection. In Section 4, we design a theoretical parameter selection method for β_0 based on the convergence rate. Finally, experiments on synthetic and real networks validate the selected parameter according to the theoretical lower bound of β_0 from our provided method and demonstrate the performance of the DAAEMMM algorithm. In addition, we test the relations between the convergence rate and the structure of a network measured by assortative mixing by degree. Finally, the conclusions are stated.

2. The EM algorithm of mixture model for network exploring

In this section, we first describe the mixture model for exploratory analysis in networks. Then the EM algorithm of its parameter estimation (EMMM) in directed networks [9] is inferred.

Structure detection based on mixture model [9] aims to deduce the assignments of nodes in a network by fitting a model to an observed network. Here we focus on the mixture model for directed networks, and it is easy to extend to the case of undirected or weighted networks. A network with N nodes is represented by an adjacency matrix A with element $A_{ij} = 1$ if there is an edge from node i to node j and 0 otherwise. Suppose that nodes of a network fall into c communities and model parameters are specified as the triplet $(\{g_i\}, \{\pi_r\}, \{\theta_{ri}\})$, where the hidden variable g_i indicates the group assignment of node i, π_r the fraction of nodes in group r, and θ_{ri} the probability that there is a directed edge from nodes of group r to node i. The model parameters $\{\pi_r\}$ and $\{\theta_{ri}\}$ satisfy the normalization conditions $\sum_{r=1}^{c} \pi_r = 1$ and $\sum_{i=1}^{N} \theta_{ri} = 1$.

Assume that edges of a network are generated by a mixture of underlying probability distribution. Each edge from i to j is generated independently. First, the begin node i of a edge $<i, j>$ selects its group g_i by a probability π_{g_i}. Then a node i in group g_i links to end node j by a probability $\theta_{g_i j}$. The likelihood of an observed network A can be written as:

$$\begin{aligned} Pr(A|\pi, \theta) = \prod_{ij} Pr(A_{ij}|\pi, \theta) &= \prod_{i} \sum_{r=1}^{c} \{Pr(g_i = r|\pi) \\ \prod_{j} Pr(A_{ij}|g_i = r, \theta)\} &= \prod_{i} \sum_{r=1}^{c} [\pi_r \prod_{j} (\theta_{rj})^{A_{ij}}]. \end{aligned} \tag{1}$$

The logarithm of the likelihood in Eq. (1) is:

$$L = \log Pr(A|\pi, \theta) = \sum_{i} \log \sum_{r=1}^{c} \left[\pi_r \prod_{j} (\theta_{rj})^{A_{ij}} \right]. \tag{2}$$

We often use the EM algorithm to estimate the parameters by maximizing the log-likelihood L. The lower bound of L can be computed by Jesen inequality from Eq. (2), noted as $\overline{L}$.

$$\overline{L} = \sum_{ir} q_{ir} \left[\log \pi_r + \sum_{j} A_{ij} \log \theta_{rj} \right] - \sum_{ir} q_{ir} \log q_{ir}, \tag{3}$$

where q_{ir} is the posterior probability that node i belongs to cluster r, and it satisfies the normalization condition $\sum_r q_{ir} = 1$.

By introducing the Lagrange multiplier to incorporate the constraint q in E step, the objective function becomes:

$$\overline{L'} = \overline{L} + \lambda \left(1 - \sum_{r} q_{ir} \right). \tag{4}$$

By letting the derivative of $\overline{L'}$ by q_{ir} be zero in E step, the updating equation for q_{ir} is inferred as follows:

$$q_{ir} = \frac{Pr(A_{i.}, g_i = r|\pi, \theta)}{Pr(A_{i.}|\pi, \theta)} = \frac{\pi_r \prod_j \theta_{rj}^{A_{ij}}}{\sum_s \pi_s \prod_j \theta_{sj}^{A_{ij}}}. \tag{5}$$

By introducing the Lagrange multipliers to incorporate the normal-