Meta-Path Graphical Lasso for Learning Heterogeneous Connectivities

Yao Zhang[1,2] Yun Xiong[1,2] Xinyue Liu[3] Xiangnan Kong[3] Yangyong Zhu[1,2]
{zhang_yao15;yunx;yyzhu}@fudan.edu.cn {xliu4;xkong}@wpi.edu

1 Shanghai Key Laboratory of Data Science, Shanghai, China
2 School of Computer Science, Fudan University, Shanghai, China
3 Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, USA

Dataology
上海市数据科学重点实验室
Shanghai Key Laboratory of Data Science

BACKGROUND

Undirected graphical models have gained prominence since they provide a natural way to model the complex interactions among a set of random variables. However, in many applications, the structure of the graph is unknown and must be inferred from limited observations.

One of the widely used graphical models is the Gaussian Graphical Model (GGM), which assumes the variables follow a multivariate Gaussian distribution. In the framework of GGM, the problem of learning the structure of a graphical model is equivalent to estimating the inverse of the covariance matrix, also referred to as the precision or concentration matrix, since the non-zero pattern of this precision matrix corresponds to the edges in the underlying graph structure. To get an interpretable graph and maintain the low model complexity, some researchers considered the sparse inverse covariance matrix estimation problem, which is also known as Graphical Lasso (GLasso).

OBJECTIVE

Most learning algorithms suffer from a very high computational complexity and are impractical when the number of nodes exceeds tens of thousands. Pathway Graphical Lasso (PathGLasso) is a recently proposed method, and it gains great acceleration and gives a more meaningful result by taking advantage of the domain knowledge that in the biological system, a pair of genes will not be connected if they do not participate together in any of the cellular processes, typically referred to as pathways.

Conventional approaches for graphical lasso mainly focus on learning one type of relation from the node activities. However, in many real-world application, the node activities can usually be explained by multiple types of relations among the nodes.

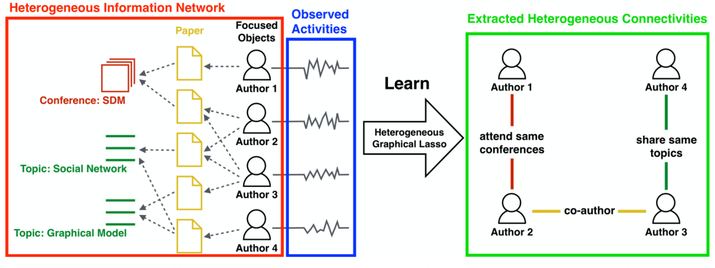

Figure 1. An illustration of learning heterogeneous connections from the observed activities of the focused objects (nodes) by incorporating a heterogeneous information network.

Thus the objective of this paper is to propose a method that

(a). can extract heterogeneous connectivities from observations with the help of side information;

(b). can be solved efficiently.

METHOD

With the recent advance in data collection techniques, many real-world applications are facing large scale heterogeneous information networks (HIN) with multiple types of objects interconnected through multiple types links, which involves a significant amount of information.

To achieve the above objective, we incorporate meta paths extracted from a HIN into the conventional graphical lasso framework. We present a novel method called Heterogeneous Graphical Lasso (HeteGLasso), and solve it using the non-convex alternating direction method of multipliers (ADMM).

$$\min_{\Theta^{(1)},\ldots,\Theta^{(K)}} \sum_{k=1}^{K}\left(-l(S,\Theta^{(k)}) + \|\Theta^{(k)}\|_{1,\Lambda^{(k)}}\right)$$

Learn multiple precision matrices simultaneously.

$$\text{s.t.} \quad \Theta_{ij}^{(k)} = 0, \quad \forall (i,j) \notin \mathcal{P}^{(k)} \qquad \text{meta path constraints}$$

$$\mathbf{card}(\Theta_{ij}^{(\cdot)}) \leqslant 1, \quad \forall i \neq j \qquad \text{exclusive constraints}$$

HeteGLasso estimates multiple precision matrices simultaneously by incorporating the objects' individual activities and the external heterogeneous information network. Each learned graph corresponds to a certain kind of relations. There are two kinds of constraints:

Meta path constraints force some elements to be zero when the corresponding objects are not connected trough a certain kind of meta path in HIN.

Exclusive constraints ensure that there is at most one nonzero element among all matrices at each off-diagonal position. This means multiple precision matrices and corresponding graphs can be easily combined as a graph with multiple types of edges.

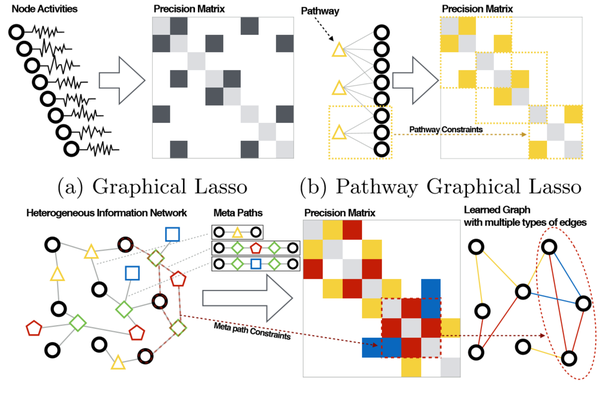

Figure 2: Comparison of three different methods for graphical lasso. (a) Glasso only uses the observed activities and outputs a sparse precision matrix. (b) PathGLasso further considers the pathway-based prior. Pathways constrain the nonzero pattern in the precision matrix. (c) HeteGLasso incorporates a HIN to learn multiple types of connections simultaneously. Meta path constraints regulate at which position elements can be nonzero, while exclusive constraints ensure the unique type of the edge between a pair of objects.

INNER SUBPROBLEM

By using non-convex ADMM, the primal optimization problem can be solved in an iterative fashion. In each iteration, we need to solve subproblems taking the same form:

$$\min_{\Theta} \quad -l(S,\Theta) + \|\Theta\|_{1,\Lambda} + \frac{\tau}{2}\|\Theta - W + V\|_F^2$$
$$\text{s.t.} \quad \Theta_{ij} = 0, \qquad \forall (i,j) \notin \mathcal{P},$$

which is a graphical lasso problem with additional Frobenius norm and meta path constraints.

We solve it by iteratively updating the parameters that correspond to one meta path group, with all of the other parameters held fixed. After re-arranging the variables, Θ takes the form:

$$\Theta = \begin{bmatrix} A & B & 0 \\ B^T & C & D \\ 0 & D^T & E \end{bmatrix}$$

parameters in the current path group
parameters in the overlapping part
parameters in the rest of the path groups

To update A, B, C with D, E fixed, the subproblem can be reduced to a smaller unconstrained problem, which can be solved efficiently. Thus the entire problem can be solved efficiently.

CASE STUDY

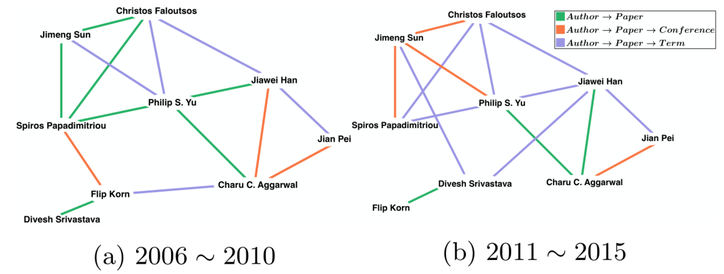

Figure 3: A subgraph of DBLP identified by HeteGLasso.

Detailed quantitative results are shown in the paper, so here we use a case study to demonstrate the effectiveness of HeteGLasso. We split the DBLP dataset into two periods: 2006~2010 and 2011~2015 and apply HeteGLasso to these two parts of dataset.

We can observe that the relations between authors had changed. For example, Jiawei Han and Charu C. Aggarwal were connected via author → paper → conference because they had attended the same conferences, but recently they were connected via author → paper due to their more frequent cooperation. This means HeteGLasso successfully captured the relations between authors in different periods.