Combining random walks and nonparametric topic model for network community detection

Ruimin Zhu

Northwestern University

ruiminzhu2014@u.northwestern.edu

May 2, 2017

Outline

Network community detection

- matrix factorization methods
- optimization methods
- generative models
- other methods

Where do the inspirations come from

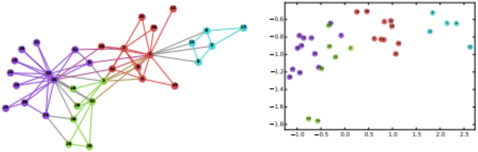
RW-HDP combines random walks and topic model for community detection.

Conducting random walks is a way of aggregating information. Each random walker is an agent who explores a local part of the network. Combining the information they collected properly, we get a big picture of the network.

Topic models are generative models that generally used for documents analysis.

Deepwalk

1. Conduct short random walks on the network
2. Treat them as sentences
3. Deep learning for word (vetex) embedding
4. Use embedding for other tasks



(a) Input: Karate Graph (b) Output: Representation

SSN-LDA

...

Basic idea

1. Conduct random walks on the network. Treat vertexes as words, communities as topics, and random walks as documents
2. Use topic model to find the topic structure of each document
3. Use Bayes theorem to find the probabilities of a vertex belonging to different topics and classify it to the largest one

Random walks

Let N_d be the length of the d^{th} random walks. The length can either be fixed or is a Poisson random variable.

Randomly sample D vertexes from the network as starting points of random walks.

Conduct D random walks and treat them as documents.

Stick breaking construction

...

Graph representation

...

Conditional distributions

Using Markov blanket we get the following conditional distributions

$$p(z_{dn}^i = 1|\pi_d, \beta_{1:\infty}, w_{dn}, c_d) \propto \exp\{\log \sigma_i(\pi_d) + \sum_{k=1}^{\infty} c_{di}^k \log \beta_{k,w_{dn}}\}$$

$$p(c_{di}^k = 1|\nu, \beta_{1:\infty}, w_d, z_d) \propto \exp\{\log \sigma_k(\nu) + \sum_{n=1}^{N} \log \beta_{k,w_{dn}}\}$$

$$p(\pi_{di}|z_d) \sim \text{Beta}(1 + \sum_{n=1}^{N} z_{dn}^i, \alpha + \sum_{n=1}^{N} \sum_{j>i} z_{dn}^j)$$

$$p(v_k|c) \sim \text{Beta}(1 + \sum_{d=1}^{D} \sum_{i=1}^{\infty} c_{di}^k, \omega + \sum_{d=1}^{D} \sum_{i=1}^{\infty} \sum_{j>k} c_{di}^j)$$

$$p(\beta_k|z, c, w) \sim \text{Dirichlet}(\eta + \sum_{d=1}^{D} \sum_{i=1}^{\infty} c_{di}^k \sum_{n=1}^{N} z_{dn}^i w_{dn}).$$

Notice that all of them are in Exponential families.

Stochastic variational inference

Based on the conditional distributions, we using the following variational family under the mean field assumption

$$q(\beta, \nu, z, \pi) = \Big(\prod_{k=1}^{K} q(\beta_k|\lambda_k) q(\nu_k|a_k) \Big) \times \Big(\prod_{d=1}^{D} \prod_{i=1}^{T} q(c_{di}|\zeta_{di}) q(\pi_{di}|\gamma_{di}) \prod_{n=1}^{N} q(z_{dn}|\phi_{dn}) \Big)$$

The latent variables c_{di}, π_{di}, z_{dn} depends only on a single document, while the global variables v_k, β_k depends on all documents. To efficiently update the proxy, we can use Stochastic Variational Inference method.

Bayes theorem for community assignment

...

Pros and Cons

Pros

1. Nonparametric topic model allow community number auto detection
2. Soft-clustering
3. High accuracy compared to other generative models
4. Can be extended to online setting

Cons

1. The inference of probabilistic model is always slow, even SVI is used

Future works

1. Include teleportation to allow single agent to explore a larger area of the network
2. Hyperparameters tuning
3. Ground truth benchmarks comparison

References

...