Network Representation Learning: A Survey

15352204 lind8@mail2.sysu.edu.cn

March 18, 2018

1 Motivation

Network Representation Learning (NRL), also called network embedding, is a popular research topic nowadays. Network is commonly seen in our daily life, e.g. social network, citation network, webpage network. However, using a discrete adjacency matrix to represent a large-scale network is difficult to store in distributed system, and lack of information of more complex relations like common neighbors. NRL aims to embed a network to a low-dimensional vector space, by preserving the important information. Challenges includes:

1) Preserving structure information including both global and local similarity. E.g. first-order proximity, second-order proximity, high-order proximity, intra-community proximity;

2) Preserving the attributes of nodes, and heterogeneity from different sources;

3) Data sparsity due to the massive nodes but few links;

4) Scalability due to the massive nodes.

2 Contributions

Based on the existing network embedding techniques, the authors build a new categorization of NRL algorithm and summarize basic methods.

3 NRL methods

3.1 Unsupervised NRL

DeepWalk Random walk + skip gram model Referring to word2vec, nodes words, node sequence - word sequence

node2vec node2vec: Difference from DeepWalk, using DFS and BFS random walk.

LINE Minimize the difference between the joint distribution and the empirical distribution, 1st proximity and 2nd proximity

GraRep Perform SVD on k-step transition probability matrix

SDNE Second-order proximity, deep-learning based methods (deep autoencoder model), minimize the loss between the input vertex and reconstructing vertex.

HOPE High-order proximity, considering the direct network, each node is represented as source node and target node, matrix factorization

APP Similar to HOPE, using softmax;

M-NMF Preserve 1st proximity and 2nd proximity, preserve modularity by matrix factorization.

3.2 Unsupervised Content Augmented NRL

TADW MF DeepWalk + attributes

HSCA TADW + penalizes

pRBM RBM, binary attributes, weighted network

UPP-SNE DeepWalk + non-linear mapping

3.3 Semi-supervised NRL

Unsupervised methods + Label information, e.g. node classification methods

DDRW DeepWalk + L2-loss SVM classifier

MMDW MF version DeepWalk + SVM classifier

TLINE LINE + multiple SVM classifier

GENE Maximize the conditional probability of co-occurring vertices given vertex labels, inspire by DeepWalk and document modeling

3.4 Semi-Supervised Content Augmented NRL

TriDNR Paragraph vector model + DeepWalk

LDE Similar to TriDNR, modeling word-word-document, document-document-document, document-label

DMF TADW + linear classifier

Planetoid Constructing vertex embedding, deep neural network

LANE Embeds joint proximity, vertex attributes, vertex label into a unified embedding representation.

3.5 Evaluation methods

Vertex classification, vertex clustering, link prediction, network reconstruction.

4 My Thinking

We can consider existing NRL methods as combination of several sub-methods. Unsupervised content augmented NRL fuse Unsupervised NRL and vertex attributes. Semi-supervised NRL methods Almost all NRL focus on static networks, i.e. off-line methods. Nonetheless, most real-world networks are in dynamic addition. For example, in social network, users build a new connections or new user login in. Dynamic network consider the following changes: 1) Addition/deletion of nodes. 2) Addition/deletion of edges. 3) Attribute value changes (only in augmented network) The difficulties lies in how to represent the changes of network by modify the embedding result in latent space of those modified vertex and remaining the embedding result of unchanged nodes. Compared to off-line methods, such online method will reduce cost and improve efficiency.