Signal Processing 124 (2016) 93–102

Contents lists available at ScienceDirect

Signal Processing

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

Supervised sampling for networked data

Meng Fang [a,b,*], Jie Yin [b], Xingquan Zhu [c]

[a] University of Technology, Sydney, NSW, Australia
[b] CSIRO, Sydney, NSW, Australia
[c] Florida Atlantic University, FL, USA

ARTICLE INFO

Article history:
Received 7 May 2015
Received in revised form
29 September 2015
Accepted 29 September 2015
Available online 17 October 2015

Keywords:
Supervised sampling
Random walks
Information network

ABSTRACT

Traditional graph sampling methods reduce the size of a large network via uniform sampling of nodes from the original network. The sampled network can be used to estimate the topological properties of the original network. However, in some application domains (e.g., disease surveillance), the goal of sampling is also to help identify a specified category of nodes (e.g., affected individuals) in a large network. This work therefore aims to, given a large information network, sample a subgraph under a specific goal of acquiring as many nodes with a particular category as possible. We refer to this problem as *supervised sampling*, where we sample a large network for a specific category of nodes. To this end, we model a network as a Markov chain and derive supervised random walks to learn stationary distributions of the sampled network. The learned stationary distribution can help identify the best node to be sampled in the next iteration. The iterative sampling process ensures that with new sampled nodes being acquired, supervised sampling can be strengthened in turn. Experiments on synthetic as well as real-world networks show that our supervised sampling algorithm outperforms existing methods in obtaining target nodes in the sampled networks.

1. Introduction

Many research works have been carried out on networked data to study the problem of node classification at various levels, including the Web, citation networks, and online social networks. The large size of these networks and other restrictions, such as privacy, make learning from the entire network become extremely computational expensive or even impossible. For example, discovering a specific community in the DBLP citation network would require searching all the HTML pages and downloading terabyte-level data, which is most likely impractical. Therefore, research studies have attempted to address the problem of acquiring a smaller, but representative, subset of samples from a large graph [1,2] and then proceed with subsequent network mining tasks.

Currently, most graph sampling algorithms have been mainly focused on generating a uniform sample of nodes and edges at random from the original graph. Assuming that the node and edge information is readily observable, they usually operate on an entirely, static graph. These methods are characterized by the order in which the nodes are visited (or traversed), for example, Bread-First Search (DFS), Depth-First Search (DFS), forest fire, and snowball sampling. They typically start at a seed node, and recursively visit (one, some or all) its neighbors. These methods are varied and distinct with each other because of different ordering strategies of visiting the nodes. Although some research works have shown that these methods are biased

* Corresponding author.
E-mail addresses: Meng.Fang@student.uts.edu.au, Meng.Fang@csiro.au (M. Fang), Jie.Yin@csiro.au (J. Yin), xzhu3@fau.edu (X. Zhu).

http://dx.doi.org/10.1016/j.sigpro.2015.09.040
0165-1684/© 2015 Elsevier B.V. All rights reserved.

towards high-degree nodes [3], they are still found to be very popular and widely used for sampling nodes in real-world large networks.

In reality, however, real-world networks may not be immediately accessible until each node and its connections are progressively crawled. For example, in a citation network, papers need to be read or preprocessed so as to find their citations, as well as categories, general terms, keywords, and authors. Thus, collecting a paper's detailed information or identifying a paper's research topic incurs a cost. It would be desirable to minimize the cost by collecting a small portion of the network instead of the entire network. Similar issues may also occur in large online social networks such as Facebook or Twitter, where one may be interested in identifying a specific group of users with certain professions or hobbies. In addition, real-world networks often have imbalanced node distributions where the majority of nodes belong to one class and very few nodes belong to the minority class. As a result, uniform sampling may fail to include the nodes belonging to the minority class because these nodes often have low degree and few connections. For example, in disease surveillance, there may exist very few affected individuals in a large population network. Due to the fact that the nodes' attribute information is not considered, as well as their bias towards high-degree nodes, traditional sampling methods are not effective for sampling nodes of minority category in large networks.

Motivated by the above observations, in this paper, we propose a new strategy for obtaining a biased sample of nodes by carrying out network sampling under supervision. We refer to this class of problems as *supervised sampling*, where we aim to identify nodes belonging to a specific category (i.e., positive instances) that may comprise only a small portion of the overall network. We provide practical implementations of supervised sampling, where given a large graph and a specific category, the goal is to iteratively sample a subgraph from the original graph under the requirements related to the category. To tackle this problem, we model a graph as a Markov chain, where nodes are considered as interior states and edges are chains between states. We design a supervised random walk to compute the stationary distributions of nodes, which indicate the probability of nodes being positive, by using nodes' attribute information. Unlike uniform sampling, we iteratively choose the best nodes to be sampled in the next iteration based on their probabilities of being positive. At each iteration, the sampling process is guided by a supervised random walk that is more likely to visit positive nodes in the neighborhood. Once a node is visited, the sampled network is expanded to include the node itself, its neighboring nodes, as well as new edges between them. After a node is sampled, the genuine label of the node is also revealed. All such information can be used to update the stationary distribution of the sampled network, which will strengthen supervised sampling at the next iteration.

The main contribution of this work is twofold: first, we introduce a new supervised sampling problem on large networks; second, we present a novel unified framework to perform supervised sampling for a given task through formulating a supervised random walk as an optimization problem. Experiments on synthetic and real-world networks show that our proposed algorithm achieves a higher recall of positive nodes while sampling large networks than baseline methods, especially for networks having imbalanced class distributions.

2. Related work

In recent years, there have been many research efforts on studying information networks, such as node classification [4], link prediction [5,6], active learning [7], transfer learning [8,9], personalized recommendation [10,11], and so on. These studies are different from traditional instance-based learning problems because both instance content and network structure information are available for learning. Sen et al. [4] introduced a classification framework for networked data as collective classification. Collective classification is a combined classification of a set of interlinked objects using correlations between node labels and node content (i.e., attributes), and information of each node's neighborhood. Even when the instances are not explicitly linked to form a network, the use of the correlations between instances is also beneficial for improving the classification performance (e.g., [12]). Link prediction is also a fundamental problem in the network settings [5], which aims to predict the presence of links between network nodes. Backstrom and Leskovec [5] proposed to combine network structure information with rich node and edge attributes. Ye et al. [6] adopted Nonnegative Matrix Tri-Factorization (NMTF) to learn latent topological features from network structure, and use them to enhance nodes' features. Bilgic et al. [7] proposed an active learning algorithm for node classification based on collective classification.

Sampling techniques have also been extensively studied on very large scale information networks. Traditional graph sampling techniques can be roughly classified into two categories: *graph traversals* and *random walks* [3]. For graph traversals, nodes are sampled without replacement; once a node is visited, it is never revisited again. Depending on the order in which nodes are visited, these methods include Breadth-First Search (BFS), Depth-First Search (DFS), forest fire, and snowball sampling [13–15]. Among others, BFS has been popularly used for sampling social networks, which has been studied extensively [14–18]. However, existing research has shown that BFS is biased towards high-degree nodes in real-world networks [19,20]. When using graph traversals for sampling, the sampling process terminates after a fraction of graph nodes are collected.

Random walks fall into the other category of sampling techniques, which usually start at any specific node and initiate a random walk by proceeding to the next node selected at random from the neighbors of the current node. It is found that random walks are biased towards high degree nodes in the graph [21]. Some methods have been proposed to correct the bias of random walks. For example, Gjoka et al. [3] proposed a Metropolis-Hastings algorithm to collect an unbiased sample of *Facebook* users.