Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Multi-Label Regularized Generative Model for Semi-Supervised Collective Classification in Large-Scale Networks

Qingyao Wu [a], Jian Chen [a,*], Shen-Shyang Ho [b], Xutao Li [b], Huaqing Min [a], Chao Han [a]

[a] *School of Software Engineering, South China University of Technology, China*
[b] *School of Computer Engineering, Nanyang Technological University, Singapore*

ARTICLE INFO

Article history:
Received 19 January 2015
Received in revised form 6 April 2015
Accepted 6 April 2015
Available online xxxx

Keywords:
Collective classification
Generative model
Semi-supervised learning
Multi-label learning
Large-scale sparsely labeled networks

ABSTRACT

The problem of *collective classification* (CC) for large-scale network data has received considerable attention in the last decade. Enabling CC usually increases accuracy when given a fully-labeled network with a large amount of labeled data. However, such labels can be difficult to obtain and learning a CC model with only a few such labels in large-scale sparsely labeled networks can lead to poor performance. In this paper, we show that leveraging the unlabeled portion of the data through *semi-supervised collective classification* (SSCC) is essential to achieving high performance. First, we describe a novel data-generating algorithm, called *generative model with network regularization* (GMNR), to exploit both labeled and unlabeled data in large-scale sparsely labeled networks. In GMNR, a network regularizer is constructed to encode the network structure information, and we apply the network regularizer to smooth the probability density functions of the generative model. Second, we extend our proposed GMNR algorithm to handle network data consisting of multi-label instances. This approach, called the *multi-label regularized generative model* (MRGM), includes an additional label regularizer to encode the label correlation, and we show how these smoothing regularizers can be incorporated into the objective function of the model to improve the performance of CC in multi-label setting. We then develop an optimization scheme to solve the objective function based on EM algorithm. Empirical results on several real-world network data classification tasks show that our proposed methods are better than the compared collective classification algorithms especially when labeled data is scarce.

1. Introduction

Networks have become ubiquitous in many application domains such as Internet, social, economical and scientific fields. Researchers in these fields have shown that systems of different nature can be represented as network data for which instances are interrelated. For example, web pages are connected to each other by hyperlinks, and telephone accounts are linked by calls. Generally, network data contain nodes (instances) interconnected with each other by edges reflects the relation or dependence between the nodes. Information on the nodes is provided as a set of attribute features (e.g., words present in the web page). Such network data are obviously not independent and identically distributed, and the class membership of an instance may influence the class membership of a related instance. Furthermore, many network data are large-scale and often involve the scenario where each node can be assigned a set of multiple labels simultaneously.

The problem of learning from large-scale network data is a challenging issue that has attracted growing attention from both academia and industry [1–3] due to its importance of related applications, ranging from web page classification to spatial data analysis and social network analysis. The *collective classification* (CC) is a task to jointly classifying interrelated instances in the network [4]. Enabling CC usually improves the performance of predictive models on network data as inference outcome for one instance can be used to improve inferences on related instances. However, such a performance improvement usually requires a fully-labeled network which contains a sufficiently large amount of labeled instances. For many large-scale network data, it is extremely expensive and time-consuming to obtain such labels especially when each instance has multiple class labels. In particular, the number of possible label assignments for an instance is exponential to the number of possible labels in a multi-label setting, which is extremely large even with a small number of possible labels. On the other hand, there are often large amount of unlabeled data available in the networks. Hence, it is of interest to develop learning algorithms that are able to utilize the large amount of unlabeled data together with the limited amount of labeled data in the large-

[*] Corresponding author. Tel.:/fax: +86 020 39380295.
E-mail address: ellachen@scut.edu.cn (J. Chen).

http://dx.doi.org/10.1016/j.bdr.2015.04.002
2214-5796/© 2015 Elsevier Inc. All rights reserved.

scale sparsely labeled network data to avoid the expensive datelabeling effort and to enhance the learning performance.

In this paper, we study the problem of *semi-supervised collective classification* (SSCC) when one is given only with limited number of labeled data, which is common case in large-scale networks. Recently, various researchers have considered to examine the SSCC task using some forms of semi-supervised learning to improve the performance of CC [5,6]. It has been shown that leveraging the unlabeled portion of the data is essential to achieving high performance. The main aim of this paper is to find a generative representation for network data classification by exploiting information from both labeled and unlabeled data. To achieve this, we propose a new data generative algorithm, called *generative model with network regularization* (GMNR), based on the *probabilistic latent semantic analysis* (PLSA) method, and incorporate the network structure into it. In GMNR, a network regularizer is constructed to encode the network structure, and we apply the network regularizer to smooth the label probability distributions of the generative model. We find that the GMNR method is able to achieve a robust classification performance even in the paucity of labeled data.

Furthermore, we extend the GMNR method to the multi-label learning setting such that instances of the network data have multiple class labels. The new algorithm, called *multi-label regularized generative model* (MRGM) utilizes an additional label regularizer to explicitly encode the label correlation. This approach is able to capture the knowledge of the underlying network structure and the label correlation observed in the data to smooth the label probability density functions when learning the generative model. As a result, the predictions ensure consistency among interlinked instances and related labels. Intuitively, an instance connected to neighbors with high probabilities of related class labels also has a high chance for these class labels. In summary, the main contributions of this paper are as follows.

- A framework that utilizes the generative model that takes into account the network structure and label correlation for the collective classification problem in a semi-supervised and multi-label learning setting;
- An effective *expectation–maximization* (EM) algorithm to solve the maximum likelihood estimation problem in the proposed methods, and to compute the label probability distributions of the instances for classification;
- A theoretical discussion on the convergence of the proposed algorithm using an auxiliary function similar to that used in [7].
- An extensive statistical evaluation of the effectiveness of the proposed approach using various real-world network datasets.

This paper extends our preliminary work [8] which considers single-label learning problem in CC. In this paper, we focus on the multi-label learning problem in semi-supervised CC. We propose a novel data generative model which is able to exploit the network information and label correlation simultaneously for handling network data consisting of multi-label instances. Both of this work and the preliminary work are based on the PLSA data generative model, however, this paper significantly extends and upgrades the work presented there. The extensions and differences of this work and the preliminary work are summarized as follows:

1. Motivation of multi-label collective classification problem and collective inference techniques from the semi-supervised learning perspective is given.
2. An extensive discussion of the related work, including collective classification, semi-supervised learning, multi-label learning, as well as the PLSA model, is given.
3. We consider the CC task in a multi-label formulation, and extend the PLSA model for multi-label learning via incorporating a novel label regularizer into the model for smoothing the resulting label probability.
4. A new thresholding method for separating the relevant and irrelevant labels for a given multi-label instance is presented.
5. Additional experiments using new multi-label data networks are conducted. Experimental results with the Friedman and Nemenyi tests to assess the statistical significance of the differences in performance are reported.
6. A theoretical discussion on the convergence properties of the proposed method is given.

The rest of the paper is organized as follows. Section 2 describes the background and the related work. Section 3 presents the proposed methodology and its derivation in detail. Section 4 discusses the datasets, the experimental setup and experimental results. Finally, conclusions are drawn in Section 5.

2. Related work

2.1. Mining network data

Numerous approaches have been designed for learning from network data and label predicting for the unlabeled nodes. These approaches have been mainly studied in the research fields of collective inference, active inference, semi-supervised learning and multi-label learning. Details on these related works are described below.

Macskassy and Provost [9] provide a brief review of the previous work of collective classification in network data. Generally, the collective classification methods can be categorized into three groups: local classifier-based methods, global formulation-based methods, and relational-only methods. i) A local classifier-based method is based on an iterative process. The local classifier is trained for prediction using the attribute features derived from the content and additional relational features by aggregating the labels of neighbors. One example is the iterative classification algorithm (ICA) [4] which has been reported to be a fairly accurate method with robust performance to different network datasets. Gibbs sampling [1] is further used in the ICA framework to enrich the statistical properties of the algorithm. In recent years, there is a lot of work proposed to use a similar schema as ICA but with different local classifiers or different scenarios [1,10]. ii) A global method trains a classifier to optimize a global objective function for prediction. It is often based on a graphical model such as the loopy belief propagation of the relaxation labeling [10]. iii) A relational-only method uses only relational information for classification. Typically, the algorithm computes a new label distribution for an instance by averaging the current distributions of its neighbors. Weighted-vote relational neighbor with relaxation labeling (wvRN+RL) [9] is one such method. Recently, Macskassy and Provost [9] show that wvRN+RL performed very well in some cases. In fact, it should be considered as a baseline for CC evaluations. Sen et al. [2] provide an empirical study to analyze the strengths and weaknesses of different CC methods. One of the major disadvantages of these CC methods is that they are mainly studied in the scenario where there are a large amounts of labeled data in the network. However, it is difficult and time-consuming to acquiring such labels in many practical applications. On the other hand, there are usually large amount of unlabeled data available. As pointed out in [5], when the labeled data are limited, the performance of collective classification may be degraded due to the insufficient number of labeled neighbors [11].

In response, recent studies have examined semi-supervised collective classification methods on sparsely-labeled networks, using