International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

SELP: Semi-supervised evidential label propagation algorithm for graph data clustering ☆,☆☆

Kuang Zhou [a,b,*], Arnaud Martin [c], Quan Pan [b], Zhunga Liu [b]

[a] School of Natural and Applied Sciences, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[b] School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[c] DRUID, IRISA, University of Rennes 1, Rue E. Branly, 22300 Lannion, France

ARTICLE INFO

Article history:
Received 24 February 2017
Received in revised form 24 September 2017
Accepted 27 September 2017
Available online xxxx

Keywords:
Semi-supervised learning
Label propagation
Theory of belief functions
Uncertainty
Community detection

ABSTRACT

With the increasing size of social networks in the real world, community detection approaches should be fast and accurate. The label propagation algorithm is known to be one of the near-linear solutions which is easy to implement. However, it is not stable and it cannot take advantage of the prior information about the network structure which is very common in real applications. In this paper, a new Semi-supervised clustering approach based on an Evidential Label Propagation strategy (SELP) is proposed to incorporate limited domain knowledge into the community detection model. The main advantage of SELP is that it can effectively use limited supervised information to guide the detection process. The prior information about the labels of nodes in the graph, including the labeled nodes and the unlabeled ones, is initially expressed in the form of mass functions. Then the evidential label propagation rule is designed to propagate the labels from the labeled nodes to the unlabeled ones. The communities of each node can be identified after the propagation process becomes stable. The outliers can be identified to be in a special class. Experimental results demonstrate the effectiveness of SELP on both graphs and classical data sets.

1. Introduction

As described in [2], communities (also called clusters or modules) are groups of nodes (vertices) which probably share common properties and/or play similar roles within the graph (or network).[1] Identifying communities may offer insight on how the network is organized [3], and it is often the precondition for the structural and functional analysis of the networked systems. Community detection for networks has attracted considerable attention crossing many areas from physics, biology, and economics to sociology [3]. It can be seen as the task of clustering on graph data, which consists of a finite set of nodes, together with a set of unordered pairs of these vertices. These pairs are known as edges in the graph.

As the size of real-world networks grows rapidly, the community detection algorithms need to be fast and efficient. The Label Propagation Algorithm (LPA), which was first investigated by Raghavan et al. [4], has the benefits of nearly-linear

☆ This paper is part of the Virtual special issue on Belief Functions: Theory and Applications, edited by Jirina Vejnarova and Vaclav Kratochvil.
☆☆ This paper is an extension and revision of [1].
* Corresponding author at: School of Natural and Applied Sciences, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China.
E-mail addresses: kzhoumath@163.com (K. Zhou), Arnaud.Martin@univ-rennes1.fr (A. Martin), quanpan@nwpu.edu.cn (Q. Pan), liuzhunga@nwpu.edu.cn (Z. Liu).
[1] In this work, "graph" and "network" are considered as synonyms.

https://doi.org/10.1016/j.ijar.2017.09.008
0888-613X/

running time and easy implementation. But the original LPA is not stable due to randomness. Different communities may be detected in different runs over the same network. Moreover, by assuming that a node always adopts the label of the majority of its neighbors, LPA ignores any other structural information existing in the neighborhood. In real applications, there is often some prior information about the network structure. For instance, in co-authorship networks, the communities related to some famous scholars are easy to know. In the movie network, the types of some special films may be clear to us. If such kind of prior information could be fused effectively in the unsupervised community detection models, the performance could be improved.

Supervised classification is one of the most popular techniques in machine learning. Generally, the goal of supervised learning is to train a classifier that reliably approximates a classification task based on a set of labeled examples from the problem of interest. The performance of the learned classifier highly depends on the proportion of labeled samples. However, in many practical applications of pattern classification, it is usually difficult to get abundant labeled samples since the task of manual labeling is time consuming and often requires expensive human labor. On the contrary, there are usually a large number of unlabeled samples which are easier to obtain. Consequently, Semi-Supervised Learning (SSL), which aims to effectively combine the unlabeled data with labeled data, has been developed to perform the classification task when there are not enough training data.

Some semi-supervised community detection approaches have already be proposed [5–7]. The supervised information in these models are mainly two types: 1. The labels of some nodes are given in advance; 2. There are some must-link and/or cannot-link pair-wise constrains between some node pairs. In this paper we focus on the former type, *i.e.,* some nodes in the graph are assumed to be labeled in advance. There are some problems when dealing with the information about node labels among the existing semi-supervised community detection methods, such as:

- If there are some outliers in the graph, the performance of the community detection model will become worse.
- If the labeled objects are located in the overlapping region between or among communities, the same label will be propagated to more than one class and, consequently, the accuracy of the detection results will be low.

The theory of belief functions is very effective in dealing with uncertain information, and it has already been applied in many fields, such as data classification [8–12], data clustering [13–16], complex networks [17–20], data fusion [21] and statistical estimation [22–24]. In this work, we try to address the above problems in semi-supervised community detection models using the theory of belief functions. The Semi-supervised Evidential Label Propagation (SELP) algorithm will be proposed to take advantage of the prior information in the graph. The initial knowledge about node labels is expressed in the form of Bayesian categorical mass functions, while the labels of the unlabeled nodes are represented by vacuous mass functions. The evidential label propagation rule is designed to propagate the labels from the labeled nodes to the unlabeled ones iteratively. The basic belief assignments about each nodes' classes are obtained after convergence of the algorithm. Experimental results show that SELP can improve the accuracy of the detected communities compared with the unsupervised version. This result confirms that limited supervised information is of great value for the community detection task.

The rest of this paper is organized as follows. In Section 2, some basic knowledge and the rationale of our method are briefly introduced. In Section 3, the proposed SELP algorithm will be presented in detail. In order to show the effectiveness of the proposed community detection approaches, in Section 4 we test the SELP algorithm on different artificial and real-world data sets and compare it with related partitioning methods. Finally, we conclude and present some perspectives in Section 5.

2. Background

In this section some related preliminary knowledge, including the theory of belief functions and the classical label propagation algorithm, will be presented.

2.1. Theory of belief functions

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ be the finite domain of X, called the discernment frame. The belief functions are defined on the power set 2^Ω. Function $m : 2^\Omega \to [0, 1]$ is said to be a Basic Belief Assignment (bba) on 2^Ω, if it satisfies:

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

Every $A \in 2^\Omega$ such that $m(A) > 0$ is called a focal element. The credibility and plausibility functions can be defined, respectively, as

$$Bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B) \ \forall A \subseteq \Omega, \tag{2}$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \ \ \forall A \subseteq \Omega. \tag{3}$$

Please cite this article in press as: K. Zhou et al., SELP: Semi-supervised evidential label propagation algorithm for graph data clustering, Int. J. Approx. Reason. (2017), https://doi.org/10.1016/j.ijar.2017.09.008