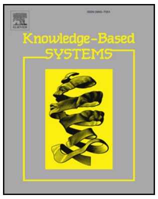
Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Collective behavior learning by differentiating personal preference from peer influence

Zan Zhang[a,b], Lin Liu[b], Hao Wang[*,a], Jiuyong Li[b], Daning Hu[c], Jiaqi Yan[d], Rene Algesheimer[e], Markus Meierer[e]

[a] *School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui, China*
[b] *School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, Australia*
[c] *Department of Informatics, University of Zurich, Zurich, Switzerland*
[d] *School of Information Management, Nanjing University, Nanjing, Jiangsu, China*
[e] *Department of Business Administration, University of Zurich, Zurich, Switzerland*

ARTICLE INFO

Keywords:
Classification with networked data
Causal analysis
Collective behavior
Propensity score

ABSTRACT

Networked data, generated by social media, presents opportunities and challenges to the study of collective behaviors in a social networking environment. In this paper, we focus on multi-label classification on networked data, for which behaviors are represented as labels and an individual can have multiple labels. Existing relational learning methods exploit the connectivity of individuals and they have shown better performance than traditional multi-label classification methods. However, an individual's behavior may be influenced by other factors, particularly personal preference. Hence, we propose a novel approach that integrates causal analysis into multi-label classification to learn collective behaviors. We employ propensity score matching and causal effect estimation to distinguish the contributions of peer influence and personal preference to collective behaviors and incorporate the findings into the design of the classifier. We further study behavior heterogeneity across subgroups in social networks, as people with different demographic features may behave differently due to different impacts of peer influence and personal preference. We estimate conditional average causal effects to analyze the impacts of peer influence and personal preference in different subgroups in social networks. Experiments on real-world datasets demonstrate that our proposed methods improve classification performance over existing methods.

1. Introduction

The advancement in social networks has produced massive amount of networked data. Increasing attention has been paid to the learning of human collective behaviors from networked data. For example, given some individuals' behaviors (e.g. adoption of certain products), how to infer the others' behaviors in the same social network. This can be considered as a classification problem where individuals' behaviors are the labels and the task is to learn a classifier from the labeled individuals, which then can be used to predict the behaviors of the other individuals.

A key challenge to networked data classification is that instances in the data are not independently identically distributed (i.i.d.) [1]. Individuals in a social network interconnect through different types of links. Conventional approaches, which usually assume that the individuals or instances are i.i.d., often have unsatisfactory performances with the data [2]. Relational learning (RL) has been proposed to address this problem by utilizing the connectivity between individuals [3,4]. Many studies have shown that the RL methods have better performance than traditional classifiers [5–7].

However, some existing RL methods only consider *peer influence*, without taking into account other factors. Peer influence is defined as how one's behaviors change with the change of his/her friends' behaviors [8]. In a networked dataset, an individual's friends are those directly connected to the individual in the network. However, peer influence may only provide partial information for correct labeling, since other factors, particularly an individual's *personal preference* can play an important role in their behaviors. In this paper, we use the term personal preference to represent the tendency of a person to have certain behavior (i.e. class label) as a result of his/her characteristics or personality. For instance, some people buy iPhones because they are Apple fans, instead of just being influenced by their friends. We consider

[*] Corresponding author.
E-mail address: jsjxwangh@hfut.edu.cn (H. Wang).

https://doi.org/10.1016/j.knosys.2018.06.027
Received 18 June 2017; Received in revised form 11 June 2018; Accepted 14 June 2018

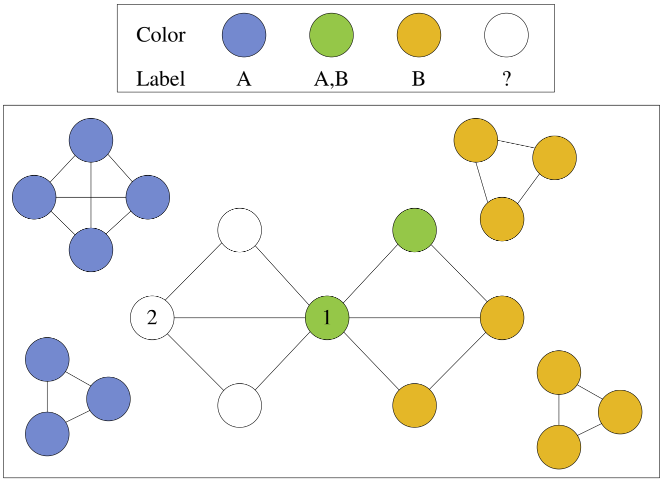


Fig. 1. A simple example of classification on networked data.

individuals with similar personal preference due to similar personality or characteristics tend to behave similarly, and the more similar two individuals are in their personal preference, the more likely they have the same behavior.

Although some existing RL methods consider both peer influence and the effects of personal preference, they do not distinguish the impacts of these two factors and consider that the contributions of these two factors are equal. Therefore, it is necessary to develop new method to consider both factors and distinguish their contributions. We use the example in Fig. 1 to show the effects of peer influence and personal preference on classification. There are two labels, A and B, and a node/individual's characteristics are indicated by node color. In this example, Node 1 has both two labels A and B. Node 2 connects to Node 1 and two other unlabeled nodes. Assume that Node 2 has one label only (label A or B), and our task is to label (classify) Node 2. Connectivity-based methods would classify Node 2 to have the same labels as Node 1 (both A and B) based on the connectivity, because in Node2's neighbors, only Node1's labels are known. However, an individual's behavior is not only a result of peer influence, but also due to personal preference. We can use personal preference to provide extra information for classification. Assume that Node2's characteristics are more similar to those of the nodes with label A than the nodes with label B. We can infer that Node 2 should have higher probability to be assigned label A, because Node2's personal preference is more similar to the nodes with label A.

From this example, we see that it is important to distinguish personal preference from peer influence and use both for classification. However, it is challenging to model and quantify the impacts of the two factors in networked data classification. For instance, in the context of adoption of iPhones, peer influence is associated with the presence of iPhone adopters in one's friends (called *adopter friends* hereafter). Personal preference is associated with having similar personal preference with other people. However, as the impacts of peer influence and personal preference are intertwined, it is difficult to estimate how much one's behavior is due to the influence of adopter friends and how much is a result of personal preference only.

Furthermore, the impacts of peer influence and personal preference vary across different subpopulations in social networks. There has been some work studying the behavioral heterogeneity [9–12]. For instance, political scientists and campaign professionals have conducted randomized experiments to investigate whether phone calls or in-person conversations are more effective at increasing candidate support. They considered research questions related to heterogeneity of subpopulations, e.g. "Do phone calls increase candidate support more from the female subpopulation than from the male subpopulation?" and "How does the effectiveness of phone calls change across subpopulations at different ages?" [13].

However, no study has been done on such heterogeneity for collective behavior learning from networked data. In learning collective behaviors in social networks, we are interested in similar questions regarding the heterogeneity in different subgroups. For example, for a female, is her friends' adoption of iPhones more likely to increase the chance for her to adopt an iPhone than for a male? Different people with different demographic features may behave differently due to different impacts of peer influence and personal preference. Therefore considering the heterogeneity of causal effects of peer influence and personal preference in different subgroups can help with accurate identification of their contributions to collective behaviors.

In this paper, we present **MCPP**, the **M**ulti-label **C**lassification algorithm which distinguishes **P**eer influence and **P**ersonal preference. We innovatively apply propensity score matching to identify and quantify the causal effect of peer influence on a node's labeling and thus to obtain the weights of peer influence and personal preference regarding their respective contributions to the labeling of a node. The weights are then used in the design of a multi-label relational classifier. We further propose (**MCPPS**), the **M**ulti-label **C**lassification algorithm which distinguishes **P**eer influence and **P**ersonal preference in **S**ubgroups to learn collective behavior while taking heterogeneity of subgroups into consideration. We use real social network datasets in our experiments. The results demonstrate that our proposed approaches can improve the performance of networked data classification.

The principal contributions of this paper as be summarized as follows:

- We propose a causal analysis approach to distinguishing the contributions of peer influence and personal preference to the collective behaviors in a social network environment, and we provide a method to examine the heterogeneity of peer influence and personal preference by estimating the conditional average causal effect in different subgroups.
- We design two multi-label classification algorithms based on the findings of the causal analyses. That is, we use the estimated causal effects of peer influence and personal preference to weight their respective contributions to the class membership probabilities (whereas existing methods either only consider a single factor or use equal weights for the two factors). We also show the effectiveness of the algorithms by making a comparative study with the state-of-the-art methods for networked data classification.

2. Problem definition

Let $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{C}, \boldsymbol{F})$ represent a social network, where $\boldsymbol{V}$ is the set of nodes denoting individuals and $\boldsymbol{E}$ the set of undirected edges denoting the relationships between the nodes; $\boldsymbol{C}$ is the set of labels each for a behavior in $\mathcal{G}$; and $\boldsymbol{F}$ is the set of features describing an individual. For a node $v \in \boldsymbol{V}$, $\boldsymbol{N} \subset \boldsymbol{V}$ denotes the set of neighbor nodes directly linked to v.

The behaviors studied here refer to the collective behaviors shared by a group of individuals in a social network, e.g. buying a product. For $\boldsymbol{C}$={$c_1, c_2, ..., c_m$}, the behaviors of an individual $v \in \boldsymbol{V}$ can be described by a binary vector, $\boldsymbol{l}$= ($l^{c_1}, l^{c_2}, ..., l^{c_m}$), where $l^{c_k} = 1$ if $c_k \in \boldsymbol{C}$ is a label of v; otherwise $l^{c_k} = 0$. For instance, if $\boldsymbol{C}$= {c_1, c_2, c_3}, representing the three behaviors considered in a social network, e.g. buying an iPhone, a Samsung or Sony phone, then $\boldsymbol{l}$= (0, 1, 0) indicates that v bought a Samsung phone.

Our goal is to predict individuals' behaviors based on the observed behaviors of other individuals in the same social network. The major problem addressed in this paper can be defined as follows.

Problem Definition. Given $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{C}, \boldsymbol{F})$, and assume that $\forall v' \in \boldsymbol{V'}$ where $\boldsymbol{V'} \subset \boldsymbol{V}$, its behavior vector $\boldsymbol{l'}$ is known. The goal of this paper is to predict the behavior vector $\boldsymbol{l}$ for each $v \in (\boldsymbol{V} \setminus \boldsymbol{V'})$.

We consider that peer influence and personal preference are the two major factors impacting individuals' behaviors, and in our design of the