Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A semi-supervised social relationships inferred model based on mobile phone data

Chen Yu [a,*], Namin Wang [a], Laurence T. Yang [a,b], Dezhong Yao [a], Ching-Hsien Hsu [c], Hai Jin [a]

[a] Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, China

[b] Department of Computer Science, St Francis Xavier University, Canada

[c] Department of Computer Science, Chung Hua University, Hsinchu, 300, Taiwan, ROC

HIGHLIGHTS

- We extracted the mobile phone communication features from the network.
- We used principal component analysis to achieve the dimensionality reduction.
- We used the co-training style semi-supervised algorithm to train two classifiers.
- We used the classifiers to obtain the relationship labels.

ARTICLE INFO

Article history:
Received 14 October 2015
Received in revised form
7 November 2016
Accepted 21 November 2016
Available online xxxx

Keywords:
Mobile phone data
Social relationship mining
Semi-supervised learning
Graphic structure

ABSTRACT

Exploring the relationships of humans is an important study in the mobile communication network. But the relationship prediction accuracy is not good enough when the number of known relationship labels (e.g., “friend” and “colleague”) is small, especially when the number of different relation classes are imbalanced in the mobile communication network. To deal with issues, we present a semi-supervised social relationships inferred model. This model can infer the relationships based on a large amount of unlabeled data or a small amount of labeled data. The model is a co-training style semi-supervised model which is combined with the support vector machine and naive Bayes. The final relationship labels are decided by the two classifiers. The proposed model is evaluated by a real mobile communication network dataset and the experiment results show that the model is effective in relationship mining, especially when the relationship network is in a stable state.

1. Introduction

With the increase in mobile subscriptions, it has inevitably brought about a sharp increase in the amount of communication data, such as sensor data records, search records, social records and so on. Using these information to mine users' behavior patterns [1,2] and social relationships [3–5] has become a hot topic in the pervasive computing. Social relationships are important part of individual in a social network. During the past decade, some researchers use proximity sensor to mining location based social networks [6,7] and some use online social network service to mining communities [8].

During those social networks, knowing the relationships among users in the mobile phone communication network can bring great benefits. It can be used as personalized service recommendations based on relationships, better understanding of the changes in the dynamics of the social structure, automatic group phone contact, and so on [9]. Nowadays, each mobile phone has the functionality to group contacts, but this functionality is hardly used. A survey by Grob et al. [10] showed that only 16% of mobile phone users create any contact groups. In addition, users of social network sites are laborious to construct social groups (e.g. ‘circles’ on Google+, and ‘lists’ on Facebook and Twitter) [11]. We think the reason for this phenomenon is that categorizing the relationship with contacts is time-consuming and a waste of effort. So how to classify the user's relationships is a worthy study subject [12]. Because the relationship labels in the communication network are seldom known, the challenge we face is how to use the

* Corresponding author.
E-mail address: yuchen@hust.edu.cn (C. Yu).

http://dx.doi.org/10.1016/j.future.2016.11.027

small labeled relationships set and the large amount of mobile phone communication information to infer the huge unlabeled relationships set in the mobile phone communication network.

To solve this challenge, we propose a semi-supervised model to solve these problems, the model can only use a small labeled relationships set, and a large amount of mobile phone communication information to infer social relationships with high accuracy. First, we extracted mobile phone communication features from the mobile phone communication network. Second, we used principal component analysis (PCA) to achieve the dimensionality reduction. Third, we used the co-training style [13] semi-supervised algorithm to train two classifiers. Finally we used these classifiers and the structure of relationship network (for example social balance [14]) to obtain the final relationship labels. The contributions of this paper can be summarized as below:

- A co-training style semi-supervised social relationship inferred model is proposed.
- We evaluated our model on the real dataset: MIT Reality Mining [15].
- The average accuracy is improved than the supervised model when the labeled dataset is small, demonstrating that our method is more stable.
- Our model achieved a greater improvement than other semi-supervised models, especially when the relation network is in a stable state.

The rest of this paper is organized as follows. We review the related works in the next section. In Section 3, we introduce the model framework of inferred relationships based on mobile communication network data. In Section 4, we express the model formally and introduce the inference process specifically. In Section 5, we test and verify the model on a real dataset and give the analysis of experimental results. Finally, in Section 6, we conclude our work and mention future work.

2. Related work

The social relationship is the core construct of sociology. Relational ties among the social network are channels for the transfer and "flow" of resources [16]. However, the known relationships that exist in a social network are sparse and it is necessary to infer relationships from some of the information observed in the social network [17,18]. Traditionally, tie strength prediction [17,19–21] and specific semantic relationships inference [22,23,15] have been two aspects that have been studied in the research into social network relationships.

Tie strength has been investigated for decades as a social science topic. Gilbert and Karrie [19] built a model to distinguish strong and weak ties, and on a 2000 social media ties dataset the model showed over 85% accuracy. Xiang et al. [20] proposed an unsupervised model to distinguish strong relationships from weak relationships using profile similarity and interaction activity. The model was a latent variable model based on homophile sociology. A Cross community sensing and mining framework is proposed to discovery connections between heterogeneous network [8]. Some game theory models are introduced to evaluate relationships among devices, service organizers and users [24,25].

The above studies have one common characteristic: they are based on virtual online social network information when classifying tie strength. But Zhang et al. [21] used the real world interactions of mobile phone call records to quantify and predict social tie strengths. They thought that relationships change over time so they added a time factor to their model and achieved prediction performance with 95.2% accuracy. Zhang et al. [26] proposed a mobility prediction system using mobile call patterns. Mirisaee et al. [27] analyzed the voice and SMS logs of mobile phones to mining social relationships. However, those methods are all supervised models.

Many methods inferred semantic relationships are also studied, such as friend, family, and colleague, using real world interaction information, not just focusing on binary or vague strength ties [28,2]. The encounter information between two users, such as encounter time, duration, and frequency, has the potential power to infer the semantic relationships. Mobile phone call and message records also have a stronger ability to predict semantic relations. For example, user *a* always call user *b* during working hours at the office, the relationship between them is likely to be colleagues. Eagle et al. [15] exploited periodical patterns of location and proximity context to infer users' relationships (workplace colleagues, outside friends or close friends). Using the same dataset, relationship evolution can also be observed [29]. To ascertain the approach to study collaboration and communication within organizations, Choi et al. [23] used machine learning techniques to infer social relationship types among co-workers who stay in same organization. Min et al. [22] classified contacts into three life facets, i.e. "Family", "Work", and "Social" using mobile phone call and text message logs. They defined 153 features based on these records and used three machine learning methods to classify relations: a rule-based model (decision tree C4.5), a probabilistic model (naive Bayes (NB)) and a multi-classified support vector machine (SVM). Their results show that the SVM is much better than the others and achieved 90% accuracy.

In summary, the previous inference works mainly based on labeled relationships in mobile communication network, but few pay attention to the unsupervised, and semi-supervised models. We mainly focus on the unlabeled relations inferred model in this paper.

The machine learning models, supervised, unsupervised, and semi-supervised, are all defined according to the input of the training set adapted to different scenarios. Supervised machine learning uses labeled samples to train a classifier, unsupervised uses unlabeled samples to train a classifier and semi-supervised trains the classifier according to both labeled and unlabeled samples [30]. Actually, the relationship labels are difficult to obtain [10,11], so we used the large amount of unlabeled sample information to reinforce the power of the classifier and propose a semi-supervised model to classify relations. Co-training is the most important paradigm in semi-supervised learning algorithm. The original co-training algorithm is proposed in 1998 by A. Blum and T. Mitchell [31], they assume that the dataset has two sufficient and redundant views. However, the sufficient and redundant views are often difficult to meet in the real problem. S. Goldman and Y. Zhou [13] proposed a co-training algorithm do not need sufficient and redundant view. The idea of them is use two different classification algorithm to train two different classifiers. The model we built is the second co-training style algorithm based on the classical SVM and NB algorithms. In the final result, we join social balance theory to optimize our model.

3. Model overview

From the mobile communication network shown in Fig. 1(a), we can obtain the mobile phone call pattern, message sent and received pattern, encounter pattern (inferred from Bluetooth), location pattern, and so on. In this paper all these patterns are called mobile communication pattern (MCP). According to the MCP, we can infer the relationships between two users shown in Fig. 1(b). For example, if user *a* and user *b* stayed in the same office for a long time every day and the relationship between them is most likely to be colleagues; if user *a* and user *c* always go for dinner together they may be friends. In the network there is a small amount of known relationships and if we can ascertain all the

Please cite this article in press as: C. Yu, et al., A semi-supervised social relationships inferred model based on mobile phone data, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.11.027