Machine Learning Methods To Identify A User's Social Circle Automatically

Wang Xiaojuan[1],Hao Xiaolong[2],FU Chunyan[1,a*]

[1]College of Information Science & Electronoic Technology, JIAMUSI University, Jiamusi 154007, China

[2] JIAMUSI University, Jiamusi 154007, China

[a]jmsfu@126.com

Keywords: social circle, overlap, hierarchical nested circle

Abstract. Uers organize networks and exchanges with the help of social networks, clasiffy the friend to the so-called social circle of friends. Construction of such a social circle will take the user a great deal of time. This paper presents a machine learning methods, he mutual networking between friends is as points clustering problem on the user's personal network. Through the relational model of multi-point created by overlap social circle we can analyze and measure the similarity of user-specific information to find hierarchical nested circles. In this paper, we obtain real data from the Weixin, Google and Renren to validate performance of proposed method. The method can explain why the nodes belong to some gathering at the same time of improving the accuracy.

Introduction

At present, Social networking sites allow users to manually assign their friends to each social circle ("circle of friends" in Weixinin, and "friends" in Renren), or determine friend by common attributes. the former is not only a waste of time but also can not automatically updated when the friends of user's is increased, the latter can't capture the individual information of the groups, the characteristics to identify friends may be lost when the personal information is missing or need to keep.

To solve this problem, two data sources can be used, the first is the side collection of individual network, we hope the gathering circles are constituted by the variable points set that dense contact[1]. However, different circle of friends are seriously overlapped, variable point can belong to more than one friend circle[2,3], and many gathering circles are hierarchical nested inside a larger circle, therefore the establishment of a variable point belong to multiple gathering rings model is very important. Secondly, each circle is not only closely linked, but also often have common attributes or characteristics[4] between its members, and therefore need to explicitly construct different dimensions of user information on each gathering circle.

This paper presents an unsupervised learning method to determine which dimensions of similarity would constitute aggregation thatclosely linked. According to the latent variables of variable and similarity to structure affiliation of the aggregation, and as a common configuration information.

The basic idea of this method is: reference thoughts of Blau spatial, allows different information similar according to different aggregation,it means a aggregation circle may be formed by friends from the same school, and the other aggregation circle is formed by friends from the

same region. Then modeling by members and similarity function of the gathering point,so to explain the observed data in the best way.

Unsupervised learning methods of model parameters

The parameters for algorithm operating are shown in table 1:

Table1 The parameters for algorithm operating

Meaning	Parameter
The input of the model is a personal network	G=(V，E)
Information of each user	v(v∈ V)
The center of the personal network	U
Each aggregation set in private network	C={C1…Ck}，Ck $\subseteq$V
how aggregation occurs	θ_k
the information of user is encoded into a tuple features	$f(x, y)$

Consider the aggregation C as potential variables, the goal is to find $\hat{\Theta} = \{\hat{q}, \hat{a}\}$:

$$\hat{\Theta}, \hat{C} = \arg\max_{\Theta, C} l_{\Theta}(G; C) - l\,\Omega(q) \tag{1}$$

To solve by Θ and C:

$$C^t = \arg\max_{C} l_{\Theta^t}(G; C) \tag{2}$$

$$\Theta^{t+1} = \arg\max_{\Theta} l_{\Theta}(G; C^t) - l\,\Omega(q) \tag{3}$$

Use the gradient increased to optimize (3), the partial derivative can be given as follows:

$$\frac{\partial l}{\partial q_k} = \sum_{e \in V \times V} -d_e(k) f(e)_k \frac{e^{\Phi(e)}}{1 + e^{\Phi(e)}} + \sum_{e \in E} d_k(e) f(e)_k - \frac{\partial \Omega}{\partial q_k}$$

$$\frac{\partial l}{\partial a_k} = \sum_{e \in V \times V} d(e \notin C_k) < f(e), q_k > \frac{e^{\Phi(e)}}{1 + e^{\Phi(e)}} - \sum_{e \in E} d(e \notin C_k) < f(e), q_k >$$

For certain $C \setminus C_i$,solving $\arg\max_{C_i} l_q(G; C \setminus C_i)$ can be expressed as the Pseudo Boolean model optimization problem to the junction of the image.it can be written as:

$$C_k = \arg\max_{C} \sum_{(x,y) \in V \times V} E_{(x,y)}(d(x \in C), d(y \in C)) \tag{4}$$

It means we hope the side of great weights (less than q_k) appears in C_k, the side of small weights do not appears in C_k . Defined $o_k(e) = \sum_{C_k \in C \setminus C_i} d_k(e) \langle f(e), q_k \rangle$, the value of E_e in $o_k(e) = \sum_{C_k \in C \setminus C_i} d_k(e) \langle f(e), q_k \rangle$, the formula (4) can be expressed as:

$$E_e^k(0,0) = E_e^k(0,1) = E_e^k(1,0) = \begin{cases} o_k(e) - a_k < f(e), q_k > -\log(1 + e^{o_k(e) - a_{k < f(e), qk >}}), & e \in E \\ -\log(1 + e^{o_k(e) - a_{k < f(e), qk >}}), & e \notin E \end{cases}$$

$$E_e^k(1,1)=\begin{cases} o_k(e)+<f(e),q_k>-\log(1+e^{O_{k(e)+<f(e),qk>}}), & e\in E \\ -\log(1+e^{O_{k(e)+<f(e),qk>}}), & e\notin E \end{cases}$$

We may use the public Pseudo Boolean Optimization Software "QPBO" [5] and approximate the question in (4) accurately, each Ck will be get by the calculation of formula (4), repeat the optimization in formula (2) and (3) until convergence, it means until Ct+1=Ct. Adjust formula (1) with l1 paradigm will lead to thinning parameter. Because personal network is relatively small, theproposed algorithm can deal with the problem of this size. Weixin, for example, there is an average of 190 or so nodes in individual network, although the maximum network encountered has 4964 nodes. Since this method is non-supervised, the inference of each network is carried out independently. This method can be used in all graphics in Weixin, gathering of each user's are conducted independently, and there is only contain hundreds of nodes in typically individual network.

In order to select the optimal number of social circles, the value of K is got according to the Bayesian Information Criterion (BIC) [6],

$$\hat{K}=\arg\min_K BIC(K;\Theta^K) \tag{5}$$

Θ^K is a set of prediction parameters in order to obtain a certain number K,

$$BIC(K;\Theta^K) \quad -2l_{\Theta^K}(G;C)+|\Theta^K|\log|E| \tag{6}$$

Regularization parameter $l\in\{0,1,10,100\}$ is determined by the results of cross-validation(leave-one-out cross-validation LOOCV), although there is no significant impact on the experiment.

Experiments

The paper access to personal networks and real data from three major social networks: Weixin, Google and Renren. including 190 friends circle and 4020 users. The tree structure of the information on two users such as x and y in Wechat is shown in figure 2, we can compare the building characters through the tree branches.

All information data sets can be represented as a tree and each layer codes show more and more specific information in the tree. For the data form the Google+, we collected data from six aspects(gender, name, title, organization, universities, and place of residence). For the data form Micro letter, collect data from 26 aspects, including native place, birthday, colleagues, political landscape etc. As for renren, simply collecting data from two aspects, that set of labels and tips the user used within two weeks. "Category" corresponds to the parent node of the leaf node in the outline tree, shown in Figure 1.

Firstly,let us describe how to use a different vector to code for relations between two users. Assuming that each user $v\in V$ has an associated information tree, and $l\in T_v$ are the tree leaves. $s_{x,y}$,defined as the difference vector of the user x and y, is a binary indicator that reflects the difference between them:

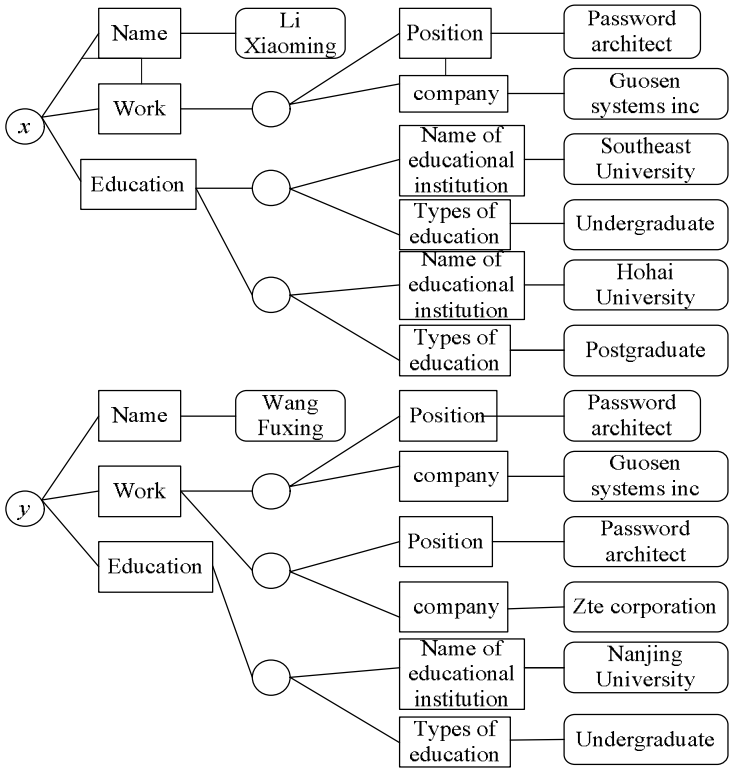

Fig1 The feature structure of user information on data set form Wechat

$$s_{x,y}[l] = d((l \in T_x) \neq (l \in T_y)) \tag{7}$$

There are certain advantages for the above different vector in terms of information coding granularity, but its drawback is that it has too many dimensions (up to 4122 dimensions). One way to solve this problem is to form a different vector based on the parent node of the leaf node. To encode for the two common categories of user information, regardless of the specific values. For example, just concern about how many co-owned label the two users have, but do not care about which one it is:

$$s'_{x,y}[p] = \sum_{l \in children(p)} s_{x,y}[l] \tag{8}$$

The advantage of this approach is that it only needs a fixed number of dimensions, regardless of the size of personal networks (as mentioned above, there are 26 micro-channel, Google + has six, all networks have two).

Now let us describe how to get the edge feature $f(x,y)$ based on the different vector $s_{x,y}$ (and $s'_{x,y}$).The first property should be a common relationship between the members and each other in the circle:

$$f^1(x,y) = (1; -s_{x,y}) \tag{9}$$

The second property is the common relationship between circle members and the individual networks :

$$f^2(x,y) = (1; -|s_{x,u} - s_{y,u}|) \tag{10}$$

These two parameters allow the assessment of which mechanism is more suitable for capturing user subjective definition of aggregation. Both properties have a constant feature of "1" and it is used to control the possibility of the same circle of friends, or to measure which one is composed by friends in a greater extent. It is important, even if the user does not have a personal information, he can still be easily predicted relationship between him and the other users in accordance with the connection mode. Similarly, for "compressed" different vector, define:

$$y^{1}(x,y) = (1; -s'_{x,y})$$

$$y^{2}(x,y) = (1; -|s'_{x,u} - s'_{y,u}|) \tag{11}$$

So far there are four identified ways to represent different aspects of two users' personal information. The two of them is to construct different vectors($s_{x,y}$ and $s'_{x,y}$),the other two is to capture the compatibility of one pair of information($f(x,y)$ and $y(x,y)$)

Experimental results

The experimental results show that the two characteristics put forward from this method (the features of friend to friend $\phi 1$ and the features of user to a friend $\phi 2$) play the same role. The two schemes encode the similar information on the whole.It is easy to understand, because the user and his friends have the similar information. Using compression characteristics $\psi 1$ and $\psi 2$ does not significantly affect the performance because they reduce the dimension of the completed feature. The accuracy of the test shows that compression features describe enough user common attribute category.

This article also implements a more detailed testing model experiment of personal network operation on the micro letter to verify whether the method can correctly identify overlapping sets and subsets. The method runs in the cases of full feature $\phi 1$、BER=0.78. The experimental results show that the algorithm includes an alumni community living in big cities, which do not involve the details of the personal identity information. The model determines the social dimension based on the aforementioned method.

Experiment finds that all the algorithms operated in micro letter are better than in Google+ and renren. Possible reasons are as follows: 1)The data on the micro letter is completed in a sense. Every participant's personal network community circle has been marked, but in other data sets, only publicly visible circle of friends can be observed, which may not be the latest. 2)26 categories on micro letter are more detailed than the 6 categories on Google+ and the basis data of renren. The more fundamental difference is the nature of the network itself: the edge of the micro letter represents the relationship of each other and the edges of Google+ and renren represent the affiliation. That changes the role relationship. Whether the algorithm uses the edge information or personal information should not get good performance.

Conclusion

The paper proposes the method can learn in completely unsupervised condition. The method also can explain why the nodes belong to some gathering at the same time of improving the accuracy.The experimental result shows that this paper presents a machine learning methods is obviously better than the natural selection and the currently popular method.

Acknowledgements

This work was financially supported by the Surface Project on Science and Technology Research of the Education Department of Heilongjiang Province (12541811).

References

[1]Chao Huaihu, Zhu Jianming, Pan Yun, Li Qingfeng. Situational awareness of the P2P mobile social network structure and algorithm [J]. Chinese Journal of Computers. 2012,35(6):1223-1234.

[2] J. Yang and J. Leskovec. Community-affiliation graph model for overlapping community detection[C]. In ICDM, 2013.

[3] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society[C]. Nature 2005.

[4]Li Zhi, Li Qianmu, Zhang Hong, Liu Fengyu. Based on recent social social time delay tolerant network routing strategy [J]. Journal of Computer Research and Development. 2012, 49(6):1185-1195.

[5] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary MRFs via extended roof duality[C]. In CVPR, 2007.

[6] M. Handcock, A. Raftery, and J. Tantrum. Model-based clustering for social networks[C]. Journal of the Royal Statistical Society Series A, 2007.