Predict Quit Rate of Group from User Behavioral and Social Information

Chang Tu, Peng Cui, and Shiqiang Yang

***Abstract*—There is a common intuition that the user behavioral pattern and social information of a group may influence its attraction to users. In this paper, we employ user behavioral and social information to predict the user quit rate of social groups and validate the link between social behavioral pattern and group quit rate on a large scale real dataset — Tencent QQ groups. We routinely model this task as a regression problem, and generate 97 features from user behavioral and social information. Then we use an improved Scalable Orthogonal Regression (iSOR) method to predict the quit rate of QQ group. Our study shows that the quit rate of a group can be predicted with high accuracy, furthermore, the iSOR method selected several import features from the total 97 social behavioral features.**

***Index Terms*—Quit rate, social group, social information, user behavior.**

I. Introduction

Understanding the social and user behavioral information of groups has been an attractive topic in social network analysis. Most of the existing works have focused on the social community detection (e.g. [1]-[3]), group profiling (e.g. [4], [5]) and group stability analysis (e.g. [6]-[10]) using social behavioral information. But little attention is paid to the link between social behavioral pattern and user quit rate of a group, which has great significance to both social science and the industry. For social science, it helps to understand the human collective behaviors. For industry, the social network operators can serve the customers more intelligently. In this paper, we extract 97 features for quit rate prediction and we explore the relation between social behavioral information and quit rate of a group.

There exist various forms of groups in many online social network services [11]. In Tencent QQ groups, for example, user can find groups of her interest by searching keywords or group ID. She can join a group when permitted by the group manager or just quit a group she joined without any restriction. All these groups are created and maintained by QQ users. Users in a group can send messages that all group members can see. The continuous records of the user behavioral and social information in a group provide us the opportunity to study the relation between social behavioral information and quit rate of a group.

In this paper, we extract lots of features from social behavioral information and predict the quit rate using an improved linear regression model, which could select significant orthogonal features automatically [12]. Our experiment result shows that the quit rate of a group can be predicted precisely using social behavioral pattern without knowing the conversation content, which validates the link between social behavioral information and group attraction. We also show that the behavioral pattern of group conversation is important to attract users staying in a group.

The main contributions of this paper are:

1) Extract useful social features and user behavioral features for quit rate prediction.

We totally extract 97 features from user behavioral and social information for our regression problem. All these features can be used for other social group research problems.

2) Validate the link between quit rate and social behavioral information.

We employ 97 features for prediction and the experiment result shows that the quit rate is to some extend related to the user behavioral pattern and social relation in a group.

3) Incorporate the historical data while selecting orthogonal features for regression.

We improve the scalable orthogonal regression (SOR) [12] method by incorporating historical quit rate value. The improved method gives better result while selecting orthogonal features.

The rest of the paper organizes as follows. We first briefly review the works related to social group. We then give an overview of the Tencent QQ group dataset, followed by the extracted social and behavioral features. After that, we introduce the improved Scalable Orthogonal Regression (iSOR) model for our quit rate prediction problem. The later section shows the experiment result of our regression problem. We conclude with an outline of future work.

II. Related Work

The extensive use of online social network service is reshaping people's daily interaction with each other, and so is the research direction of some researchers. Social group, one of the most import functionalities in social network, attracts lots of users consuming time on it [13]. Amount of research have focused on social group analysis. Most of the existing works have paid attention to community detection, group stability analysis and group profiling problems in social networks [14].

Some community detection method based on the social structure. For example, S. Wasserman and K. Faust evaluated the importance of an edge and detected community by incrementally adding edge in a decreasing order of

Manuscript received February 15, 2015; revised June 2, 2015.

The authors are with the Department of Computer Science and Technology, Tsinghua University, Beijing 10084, China (e-mail: tuc12@mails.tsinghua.edu.cn, cuip@tsinghua.edu.cn, yangshq@tsinghua.edu.cn).

DOI: 10.7763/IJKE.2015.V1.17

importance in the early years [15]. In recent years, modularity-based method was frequently used to community detection problem. L. Duan and W. N. Street connected modularity based method with correlation analysis by subtly reformatting their math formula [2], which provided a more effective way to solve resolution problem of the original modularity-based method.

The group stability analysis problem mainly focused on predicting the changing trend of group size. A. Patil and J. Liu gave a definition of stable and shrinking groups, and they extracted features from social behavioral information to classify the stable and shrinking groups [8]. Their experiment results on two dataset show that the stability can be predicted with high accuracy. However, they modeled the problem as a two-class problem, which cannot predict the group size changing degree precisely.

The group profiling research often focused on completing the group information, such as group theme and group labels. It can be used for group recommendation, social marketing, and trending analysis etc. L. Tang and X. Wang explored several group-profiling strategies to construct a group description [4], which can be used to better understanding of group formation. P. Cui and T. Zhang proposed a regularized mixed regression model to infer group themes from collective social and behavioral information of group members [5]. Their model can be used to infer hierarchical semantics, such as group category and labels, which was important to group profiling. Sociologists also proposed some theory to explain how group forms [15]. Przemyslaw A. employed the common identity and bond theory to classify the groups as topical or social [16], which gave new opportunities for group profiling.

Our work focuses on the quit rate prediction of social group. We adopt the QQ group dataset and extract 97 features from user behavioral and social information and use an improved scalable orthogonal regression method for prediction.

III. Dataset and Features

In this section, we will give an overview of the dataset, and introduce the features we extract from social and user behavioral information of group.

A. Data Description

Tencent QQ is one of the largest social network provider containing more than 500 million active users and more than 100 million online groups created and administrated by QQ users. We randomly select 10,000 online groups with total 3 million anonymized users. Each QQ user can create a QQ group, and select a category for it. Group owner can invite her friends to join the group, so that they can have conversation in a group. Users can find the group she interested by searching keywords or group ID number. Each member in the group can participate in group conversation.

We collect the group data from Tencent QQ platform during January 1st to March 1st 2014. We get each user's main profile(e.g. age, sex, geo-location, etc.), and the social relation in QQ platform between users in the same group (note that any pair of group members may or may not have friendship relation in QQ platform). We also collect the message sending log (without content) of users for one month. Besides, we have each group's main information (e.g. name, category, group size, etc.) and group member's join/quit log in two months(e.g. the member joined or quited the group, and the timestamp of the action). We totally have 3,018,087 user join/quit logs, 2,974,869 user profiles, 47,110,795 social relation, and 59,791,066 user participating behaviors.

B. Quit Rate Definition

As we have the user join/quit log of each group, we can conclude the group size at each timestamp. We defined the quit rate of group as follows:

$$quit\_rate = \frac{\#quit\_users}{\#group\_size + \#new\_users} \qquad (1)$$

In this definition, the group size is the total number of group members at the beginning of this month, so the value of quit rate falls into the interval [0, 1.0]. Fig. 1 shows the distribution of quit rate. This figure shows that the quit rate of groups mainly concentrated in the interval [0, 0.2].

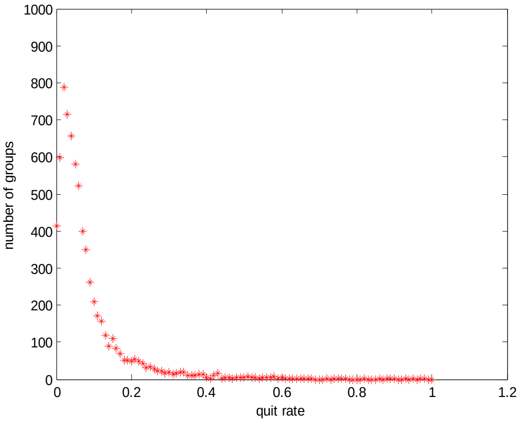

Fig. 2. Quit rate distribution.

The quit rate difference (Fig. 2) between January 2014 and February 2014 shows that simply using this month's quit rate to predict the future is improper. So we need to incorporate the user behavioral and social information to raise the prediction accuracy.

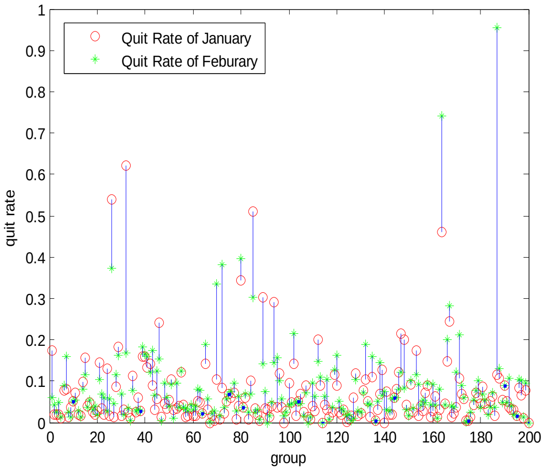

Fig. 2. Quit rate difference (200 groups).

C. Behavioral Features

There are two types of groups in QQ platform, e.g. topical groups and social groups. The group members in social group mainly have some kind of relationship in real world, such as colleagues, friends, classmates etc. However, the members in topical groups may not know each other in the real world. They gather in the same group as they have some common interest [17].

In the QQ groups, a user's experience in group may be influenced by others' behavioral in the group [18]. For example, if someone sends a message in a group but no one replies him, he will leave this group probably.

Here we give the definition of some behavioral features.

1) Conversation features

One of QQ group's main functionalities is to provide a convenient platform for lots of users' conversation. In QQ groups, one can send a message at any time and receive every message sent by the other members. The conversation pattern may have impact to some user's experience in group. Here we list some conversation features.

Number of Conversations. Here we define a conversation as containing more than 10 messages and the time interval between two successive messages less than 5 minutes.

Number of Hot Conversations. A conversation is hot conversation if it contains more than 30 messages. Here we use the number of hot conversations as a feature to depict the active degree of a group.

Number of No Reply Messages. A message is no reply message if there is no one sending a message in half an hour after the message sent into group.

Message Affinity. We first count the three users which send the three largest numbers of messages in a group, then we sum their message numbers and calculate the ratio of this sum to the total number of messages in this group this month.

Average Time Length of Conversation. We first get each conversation's time length, and then we calculate the average value of all time lengths.

Initiator Ratio. A user is an initiator if she initiated a conversation in group. We first get the number of initiators and then calculate the ratio of this number to the total number of group members.

2) Group owner behavioral features

Group owner(creator) is a special role in QQ group. The owner of a group can either invite her friends to join the group or kick a user out of the group if she does not like.

Owner Participated Conversation Ratio. We first count the number of conversations that the owner participated in, and then calculate the ratio of this number to the number of conversations.

Number of Owner Initiated Conversations. A conversation is owner initiated if the first message of a conversation sent by the group owner. We use the number of owner initiated conversations to depict the owner's initiative in group.

3) Other behavioral features

There are some other behavioral features that cannot be included in the above features.

User Login Days' Variance. We first get the number of each user's login days in this month, and then we calculate the stand deviation of these numbers as the login days' variance.

Mobile Phone User Ratio. If a user sending more messages by mobile phone than by computer, we call this user a mobile phone user. We calculate the ratio of the number of mobile phone users to group size.

Inactive User Ratio. A user is inactive if she never post a message in group this month. We calculate the ratio of the number of inactive users to the group size.

D. Social Features

We suppose that the social information of a group is to some extend related to the group attractiveness to members, which influences one's decision to quit or not. Therefore, we extract some features from social information.

We categorize social features into three different groups according to their information sources, and we list some important features' (according to the experiment result) definition.

1) Group features

Group information such as group size, group name and category sometimes may influence the user to join a group. Here we list some import features extracted from group information.

Group Size. The number of group members in group at the end of this month. As group size is used for the calculation of quit rate, we believe that current group size is an important feature to predict the future quit rate.

Group Category. There are total 7 categories in our dataset, and we use a 7 dimension vector to indicate the category of a group. It is a common intuition that the relational group (one of the 7 categories) may have lower quit rate.

2) User features

The construction of group members sometimes defines the group's attribute. The common age and location may attract the users to stay longer in the group.

Average Age. Here we use the average value of group members' age to depict the group's main age of users.

Male Rate. We calculate the male rate by the ratio of male members' number to the group size.

Number of Young Female Users. A female user is young if her age is between 15 and 25. We use the number of young female user as a feature.

Location Affinity. We first get the province that includes the largest number of group members, and then calculate the ratio of the number of group members in that province to the group size.

Average QQ Age. QQ age depicts the time length from the user's registration day to now. We calculate the average of all group members' QQ age.

3) Social relation features

If a user has some friends stays in the same group, she may stay longer and act more active by social influence. Here we list some useful features related to social relation in the QQ platform.

Friendship Density. We first calculate the number of member pairs that have friendship relation in QQ platform, and then we calculate the ratio of this number to the total number of member pairs we could have in the group.

Owner's Friends Ratio. It's calculated by the ratio of the number of group owner's friends in group to the total number of group members.

E. Feature Correlation

We totally extract 97 features from user behavioral and social information. We calculate their correlations with the quit rate of next month. Here we list 20 representative features and their correlation coefficients. From Table I, we can see that the social behavioral features of last month are closely related to the quit rate of next month.

TABLE I: CORRELATION COEFFICIENT BETWEEN FEATURE AND QUIT RATE

Feature Category	Feature	Correlation
Behavioral Features	Inactive User Ratio	-0.4572
	Number of Initiators	0.3535
	Number of Hot Conversations	0.3171
	Number of Conversations	0.3144
	Number of Mobile Messages	0.2868
	Std of Participator Size	0.2857
	Average Message Number	0.2702
	Std of 3 Participators	0.2684
	Average Participator Size	0.2644
	Average Duration	0.2195
Social Features	Average QQ Age	-0.2612
	Young Member Ratio	0.1912
	Number of Young Female Users	0.1780
	Average Age	-0.1726
	Young Female Ratio	0.1528
	Number of Cities	0.1139
	Number of Female Users	0.1081
	Number of Provinces	0.0994
	Owner's Age	-0.0947
	Minimum Age	-0.0885
Historical Quit Rate	Quit Rate of This Month	0.6177

IV. QUIT RATE PREDICTION

In this section, we will present the iSOR method for quit rate prediction in social group. First we introduce some symbols and notations that will be used in this paper.

A. Notations

We use a matrix X to denote the QQ group data matrix containing n groups and m features ($X \in R^{nxm}$), so the column vector $X_{.j}$ can be regarded as the *j*-th feature of data. For later convenience, we normalize each feature $X_{.j}$ by dividing its L2-norm, so that $\|X_{.j}\|_2 = 1$ .Since our target is to predict the quit rate of group, we use a column vector $y \in R^n$ as the corresponding response vector (e.g. quit rate of next month), and $y' \in R^n$ as the historical value (e.g. quit rate of last month).

We suppose that the features and the quit rate are to some extend linearly related. Then we model the quit rate prediction problem as a linear regression problem,

$$f(x) = y_i' + \sum_{j=1}^{m} \omega_j \times X_{ij} = y_i' + X_{i.}\omega \tag{2}$$

where the column vector $\omega \in R^n$ is the regression coefficient.

B. Problem Formulation

Now we can see that the quit rate prediction problem is transformed into a linear regression problem. We use square loss which takes the following form:

$$J(\omega) = \frac{1}{2}\sum_i (f(X_{i.}) - y_i)^2 = \frac{1}{2}\|y' + X\omega - y\|^2 \tag{3}$$

In order to raise the prediction accuracy, we also need to consider two other aspect of the quit rate prediction problem:

1) The number of powerful features should be limited

There are many factors that influence the group attraction to users, however, only a small number of features highly impact the users. What's more, for a deep understanding of group attraction to users, we need to limit the number of features selected. We can achieve this by adding a L1 regularizer on ω like [12]. So we get the following objective:

$$J(\omega) = \frac{1}{2}\|y' + X\omega - y\|^2 + \alpha\|\omega\|_1 \tag{4}$$

Here, α is a model parameter controlling the sparsity.

2) The selected features should be complementary

We want the features are complementary to each other, so we can get little redundancy on all features like [12], [19]. Mathematically, this can be satisfied by adding an orthogonal regularization term on social behavioral features vectors. This term is as follows:

$$R(\omega) = \sum_{i,j}^{m} (\omega_i X_{.i} X_{.j} \omega_j)^2 \tag{5}$$

Adding the orthogonal regularizer, we need to minimize the following function:

$$J(\omega) = \frac{1}{2}\|y' + X\omega - y\|^2 + \alpha\|\omega\|_1 + \frac{\beta}{4}R(\omega) \tag{6}$$

Here, β is another model parameter which controls the redundancy of features.

C. Optimization Algorithm

In order to minimize the objective $J(\omega)$, we propose an auxiliary function,

$$F(\omega) = \frac{1}{2}\|y' + X\omega - y\|^2 + \frac{\beta}{4}R(\omega) \tag{7}$$

The derivative of $F(\omega)$ with respect to ω is

$$\begin{aligned}\nabla \mathrm{F}(\omega) = & -X^T y + X^T y' + X^T X\omega \\ & + \beta((\omega\omega^T) \odot (X^T X) \odot (X^T X))\omega\end{aligned} \tag{8}$$

$\odot$ is Hadamard product. The derivative function is obviously continuous and differentiable, so it is locally Lipschitz continuous [20]. According to this, $\forall \theta \in R^{\mathrm{n}}$, there exists T>0 satisfying $\|\omega - \theta\| < \mathrm{T}$, such that

$$F(\omega) \le F(\theta) + (\omega-\theta)^T \nabla F(\theta) + \frac{T}{2}\|\omega-\theta\|^2 \qquad (9)$$

Now we introduce another auxiliary function,

$$G(\omega,\theta) = F(\theta) + (\omega-\theta)^T \nabla F(\theta) + \frac{\mathrm{T}}{2}\|\omega-\theta\|^2 + \alpha\|\omega\|_1 \qquad (10)$$

We can easily find out that,

$$J(\omega) = \mathrm{G}(\omega,\omega) \le \mathrm{G}(\omega,\theta) \qquad (11)$$

Based on this inequality, we can design the following iterative optimization strategy [21]:

1) Set $\omega^0 = 0$, where **0** is an all zero vector.

2) Update $\omega^{k+1} = \arg\min_{\omega} G(\omega,\omega^k)$ for *k* = 1, 2, ….

For the above inequality, we can get the following property about ω^k series:

$$J(\omega^{k+1}) \le G(\omega^{k+1},\omega^k) \le G(\omega^k,\omega^k) \le J(\omega^{\mathrm{k}}) \qquad (12)$$

So the value of objective $J(\omega)$ will be monotonically decreasing with the iterative optimization strategy. We can transform min $G(\omega,\omega^k)$ to the following form,

$$min(\omega-\omega^k)^T \nabla F(\omega^k) + \frac{T}{2}\left\|\omega-\omega^k\right\|^2 + \alpha\|\omega\|_1 \qquad (13)$$

As $\nabla F(\omega^k)$ is a constant with respect to ω, the above objective is equivalent to the following form,

$$min(\omega-\omega^k)^T \nabla F(\omega^k) + \frac{T}{2}\left\|\omega-\omega^k\right\|^2 + \alpha\|\omega\|_1 + \frac{1}{2T^2}\left\|\nabla F(\omega^k)\right\|^2 \qquad (14)$$

According to the method used in [12], we can easily get the solution of ω is

$$\omega_j^{k+1} = (\left|\omega_j^k - \frac{\tau}{T}\right| - \frac{\alpha}{T})_+ sign(\omega_j^k - \frac{\tau}{T}) \qquad (15)$$

where $\tau = (\nabla F(\omega))_j$, and *j* = 1, 2, 3, … , *m*.

The whole procedure of iSOR has been summarized in algorithm 1. In this algorithm, the parameter γ is used to increase T so as to satisfy the Lipschitz condition.

Algorithm 1 improved Scalable Orthogonal Regression

Require: Model parameter $\alpha > 0, \beta > 0$, Lipschitz parameter $\mathrm{T} > 0$, group feature X, the quit rate of this month y’ , the quit rate of next month y and the increasing rate γ.

```
Initialize the coefficient $\omega^0 \leftarrow 0$
compute $\mathrm{F}^0 \leftarrow \mathrm{F}(\omega^0)$ using (6)
while not converged do
    compute $\nabla \mathrm{F}(\omega)$ using (8)
    update $\omega^{k+1}$ using (15)
    compute $\mathrm{F}^{k+1} \leftarrow \mathrm{F}(\omega^{k+1})$ using (6)
    if $\mathrm{F}^{k+1} < \mathrm{F}^k$ then
        k←k+1
    else
        $\mathrm{T} \leftarrow \gamma\mathrm{T}$
    end if
end while
return the final coefficient $\omega$
```

V. Experiments

In this section, we will present the experiment results on iSOR method. We first give the experiment setting, and then we introduce the baseline methods and the evaluation metric, after that is the experiment result, and we conclude with feature analysis on different methods.

A. *Experiment Setting*

We totally get 10,000 groups from Tencent platform. After data preprocessing, we extract 97 features from the user behavioral and social information. In our experiment, we select 7000 groups for training and testing. We divide the 7000 groups into 5 parts for 5-fold cross validation.

In our iSOR model for quit rate prediction, there are 2 model parameters α and β which controls the sparsity and the orthogonality of features. In the training process, we tune the value of α and β from [0.001 0.002 0.005 0.01 0.02 0.05 0.1 0.2 0.5], and we get $\alpha = 0.1$ and $\beta = 0.02$ as the optimal values. As *T* is Lipschitz parameter, we set *T* = 1.0, and the increasing rate $\gamma = 1.05$.

In our hypothesis, the quit rate of a group is to some extend related to the user behavioral information and social information in the group. So we separately use user behavioral features and social features running the regression model, and then we combine them together for prediction.

B. *Baseline and Evaluation Metrics*

In order to demonstrate the advantages of the iSOR method, we implement the following methods as baseline:

1) Linear regression with L2 norm (Ridge regression)

This is a representative model for linear regression using square loss and L2 norm as regularizer. This method can avoid over-fitting by weight decay.

2) Linear regression with L1 norm (Lasso regression)

Linear regression with Lasso has the advantage for feature selection [22]. As we have extracted 97 features from data, we can use Lasso Regression to select important features automatically. Here we use the standard l1_ls method for experiment.

3) Scalable orthogonal regression (SOR)

This is a linear regression model using square loss with lasso and orthogonal regularizer. Our iSOR model is based on SOR. We implement the SOR algorithm [12].

As our quit rate prediction problem is a regression problem, we use MAE (mean absolute error) and Kendall Tau to evaluate prediction performance of the proposed method and the baseline method. Here are the definitions:

$$MAE = \frac{1}{n}\sum_{i} \left| f(X_{i.}) - y_i \right| \tag{16}$$

$$\tau = \frac{\#\text{concordant\_pair-}\#\text{discordant\_}pair}{\frac{1}{2}n(n-1)} \tag{17}$$

C. Experiment Result

As discussed above, we do experiments with four baselines models and our iSOR model. We first employ three different feature sets for iSOR method, e.g. social features, user behavioral features and the combined feature set.

TABLE II: Different Feature Set Prediction Performance

Feature Set	MAE	τ
Social Features	0.04045	0.54478
Behavioral Features	0.04038	0.54489
Combined Features	**0.04037**	**0.54581**

TABLE III: Prediction Performance of Different Methods

	Method	MAE	τ
1	Ridge	0.0431	0.5438
	Lasso	0.0428	0.5493
	SOR	0.0426	0.5471
	iSOR	**0.0420**	**0.5641**
2	Ridge	0.0408	0.5513
	Lasso	0.0405	**0.5569**
	SOR	0.0402	0.5536
	iSOR	**0.0402**	0.5541
3	Ridge	0.0416	0.5339
	Lasso	0.0412	0.5377
	SOR	0.0403	0.5395
	iSOR	**0.0400**	**0.5516**
4	Ridge	0.0425	0.5244
	Lasso	**0.0424**	**0.5282**
	SOR	0.0425	0.5221
	iSOR	0.0431	0.5252
5	Ridge	0.0394	0.5187
	Lasso	0.0392	0.5224
	SOR	0.0387	0.5232
	iSOR	**0.0365**	**0.5341**
average	Ridge	0.0415	0.5344
	Lasso	0.0412	0.5389
	SOR	0.0409	0.5371
	iSOR	**0.0404**	**0.5458**

Table II is the experiment result with three different feature sets. We can see that the user behavioral features possess more predictive power than social features, and the combined feature set can reach the best performance. The group user's action of quit or staying in group is influenced by the interactive behavioral patterns and social attributes.

In the quit rate prediction problem, we extract 97 features from dataset. Some of the features have much redundancy, and Neither Ridge or Lasso can reduce the redundancy in our feature set. In SOR method, the historical quit rate can be treated only as a common feature. However, iSOR method treat historical quit rate as a special feature and the other features can be used to tune the predicting value.

To demonstrate that the iSOR model is more suitable for our prediction problem, we experiment with different baseline methods on QQ group dataset, and Table III shows the compared result. Overall, the iSOR model using the combined feature set performs best.

In Fig. 3, we present a scatterplot to show the relation of the true value and predicted value (using iSOR method) of quit rate. We discover that for the high value of quit rate there are more instances with high errors. From the distribution of quit rate (Fig. 1), we can infer this may be caused by the small amount of training data for the higher quit rate.

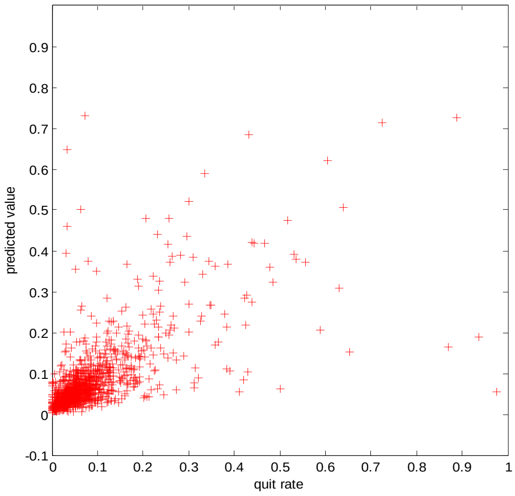

Fig. 3. Scatterplot of predicted and true quit rate.

D. Feature Analysis

TABLE IV: Features Importance of Different Methods

Ridge	Lasso	SOR	iSOR
Historical quit rate	Historical quit rate	Historical quit rate	Historical quit rate
Average QQ Age	Average QQ Age	Ratio of Active Users	Ratio of Inactive Users
Average User Login Days	Variance of User Login Days	Young Member Ratio	Group Category 1
Number of Provinces	Average User Login Days	Ratio of Mobile Message	Ratio of Mobile Message
Variance of User Login Days	Number of Provinces	Variance of User Login Days	Location Affinity
Number of Hot Conversations	Number of Initiators	Number of Initiators	User Ratio of 3 most Province
Number of Initiators	Average Message Number	Message Ratio of 3 most Active User	User Ratio of 3 most City

Different method may value different features. In Table IV, we list the top 7 features according their weights. We can see that the historical quit rate is selected by these methods. There are some common features in ridge and lasso, so is the SOR and iSOR methods. However, SOR and iSOR select smaller number of features than Ridge and Lasso. This may be caused by the feature orthogonality of SOR and iSOR methods.

VI. Conclusions

In this paper, we aim to predict the quit rate of social group by user behavioral and social information and try to understand the link between quit rate of group and social behavioral pattern. We totally extract 97 features from QQ group dataset, and separately using the user behavioral and social features for regression. Our experiment result shows that user behavioral feature have more predictive power than social features, which demonstrates that the user behavior of a group is more influential to group members. We also employ an improved SOR method, which effectively select several import features from the combined feature set. So we can deduce that these selected features influences the users' quit action. However, the group's attraction to users and the user's action may also be influenced by some influential member in group [23]. How to find out the influential member in a group and study the link between influential member and group attraction will be our future work.

Acknowledgment

This work was supported by the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," in *Proc. the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[2] L. Duan, W. N. Street, Y. Liu, and H. Lu, "Community detection in graphs through correlation," in *Proc. the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1376-1385, New York, August 2014.

[3] S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486, pp. 75–174, 2010.

[4] L. Tang, X. Wang, and H. Liu, "Group profiling for understanding social structures," *ACM Transaction on Intelligent Systems and Technology*, vol. 3, no. 1, p. 15, 2012.

[5] P. Cui, T. Zhang, F. Wang, and P. He, "Perceiving group themes from collective social and behavioral information," presented the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[6] Y. C. Wang, R. Kraut, and J. M. Levine, "To stay or leave? The relationship of emotional and informational support to commitment in online health support groups," in *Proc. CSCW '12*, pp. 833–842, 2012.

[7] E. Zheleva, H. Sharara, and L. Getoor, "Co-evolution of social and affiliation networks," in *Proc. the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1007–1016, 2009.

[8] A. Patil, J. Liu, and J. Gao, "Predicting group stability in online social networks," in *Proc. the 22nd International Conference on World Wide Web*, pp. 1021-1030, 2013.

[9] C. Budak and R. Agrawal, "On participation in group chats on twitter," in *Proc. the 22nd international conference on World Wide Web*, pp. 165-176, 2013.

[10] S. R. Kairam *et al*., "The life and death of online groups: Predicting group growth and longevity," in *Proc. the Fifth ACM International Conference on Web Search and Data Mining*, pp. 673–682, 2012.

[11] L. Backstrom *et al*., "Group formation in large social networks: membership, growth, and evolution," in *Proc. the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 44–54, 2006.

[12] D. Luo, F. Wang, J. Sun, M. Markatou, J. Hu, and S. Ebadollahi, "Sor: Scalable orthogonal regression for low-redundancy feature selection and its healthcare applications," *SDM*, pp. 576–587, 2012.

[13] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *Proc. the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 44–54, 2006.

[14] A. Mislove *et al*., "Measurement and analysis of online social networks," in *Proc. the 7th ACM SIGCOMM Conference on Internet Measurement*, pp. 29–42, ACM, 2007.

[15] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994, ch. 4.

[16] A. Przemyslaw *et al*., "Distinguishing Topical and social groups based on common identity and bond theory," in *Proc. the Sixth ACM International Conference on Web Search and Data Mining*, pp. 627-636, 2013.

[17] L. Backstrom, R. Kumar, C. Marlow, J. Novak, and A. Tomkins, "Preferential behavior in online groups," in *Proc. International Conference on Web Search and Web Data Mining*, pp. 117–128, New York, NY, USA, 2008.

[18] M. E. Ireland *et al*., "Language style matching predicts relationship initiation and stability," *Psychological Science*, vol. 22, pp. 39–44, Jan. 2011.

[19] P. Cui *et al*., "Cascading outbreak prediction in networks: a data-driven approach," in *Proc. the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 901-909, 2013.

[20] N. Garofalo and D. Nhieu, "Lipschitz continuity, global smooth approximations and extension theorems for sobolev functions in carnot-carath'eodory spaces," *Journal d'Analyse Math Ematique*, vol. 74, no. 1, pp. 67–97, 1998.

[21] D. P. Bertsekas, *Nonlinear Programming*, MIT Press, 1999.

[22] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "A method for large-scale l1-regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[23] P. Cui, F. Wang, and S. Yang, "Item-level social influence prediction with probabilistic hybrid factor matrix factorization," presented at Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.

Chang Tu received the B.E. degree from the School of Naval Architecture and Ocean Engineering of Huazhong University of Science and Technology in 2008. He is pursuing the M.E. degree from the Department of Computer Science and Technology of Tsinghua University, and his main research interests include data mining and social network analysis.

Peng Cui received the Ph.D. degree in computer science in 2010 from Tsinghua University and he is an associate professor at Tsinghua. He has vast research interests in data mining, multimedia processing, and social network analysis. Until now, he has published more than 20 papers in conferences such as SIGIR, AAAI, ICDM, etc. and journals such as IEEE TMM, IEEE TIP, DMKD, etc. Now his research is sponsored by National Science Foundation of China, Samsung, Tencent, etc. He also serves as a guest editor, co-chair, PC member, and reviewer of several high-level international conferences, workshops, and journals.

Shiqing Yang received the B.E. and M.E. degrees from the Department of Computer Science and Technology of Tsinghua University in 1977 and 1983, respectively. He is now a professor at Tsinghua University. His research interests include multimedia technology and systems, video compression and streaming, content-based retrieval for multimedia information, multimedia content security, and digital right management. He has published more than 100 papers and MPEG standard proposals. Prof. Yang has organized many conferences as program Chair or TPC member, including PCM05, PCM06 Workshop On ACM Multimedia05, MMM06, ICME06, MMSP05, ASWC06, etc.