Web Intelligence and Agent Systems: An International Journal 12 (2014) 249–265
DOI 10.3233/WIA-140295
IOS Press

Predicting personality traits of microblog users

Shuotian Bai [a], Sha Yuan [a], Bibo Hao [a] and Tingshao Zhu [b,*]
[a] *University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing, China*
E-mail: baishutian10@ucas.ac.cn
[b] *Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road Chaoyang District, Beijing, China*

Abstract. Personality can be defined as a set of characteristics which makes a person unique. Psychological theory suggests that people's behavior is a reflection of personality. Therefore, it is feasible to predict personality through behavior. Conventional personality assessment is performed by self-report inventory. Participants need to fill in a tedious inventory to get their personality scores. In the large-scale investigation, every returned inventory needs manual computation, which costs much manual efforts and cannot be done in real time. In order to avoid these shortages, this research aims to objectively predict the Big-Five personality from the usage records of Sina Microblog. Since its initial launch in December, 2005, Sina Microblog has been the leading microblogging service provider in China. Millions of users upload and download resources via microblogging status everyday. Therefore, by conducting an online user survey of 444 active users, this paper analyzes the relation modes between personality and online behavior. Furthermore, this research proposes multi-task regression and incremental regression to predict the Big-Five personality from online behaviors. The results indicate that correlation factors are significant between different personality dimensions. Besides, our training data set is reliable enough and multi-task regression performs better than other modeling algorithms.

Keywords: Big-five personality, Sina Microblog, regression, prediction

1. Introduction

Personality uniquely characterizes an individual, and profoundly influences user's mental status and social behaviors [2,36]. In psychological definition, personality is the particular combination of emotional, attitudinal, and behavioral response patterns of an individual [9]. One of the mostly influential and generally accepted personality theories is the big-five personality theory [17,23], which includes five basic traits: Agreeableness (Agre.), Conscientiousness (Cons.), Extraversion (Extr.), Neuroticism (Neur.), and Openness (Open.), to form human personality [18].

Psychological theory suggests that people's behavior is a reflection of personality. Therefore, it is feasible to predict personality through behavior [30]. From the point of view of behavior, agreeableness reflects the individual behavioral characteristics, such as conducting help, cooperation and sympathy for others. Conscientiousness includes elements of self-discipline, organization and thoroughness of planning, as well as the need for achievement. Extraversion is directly related to social skills, talkative ability and personal charm. Neuroticism reflects the degree of emotional stability, and has a close link to mental health (depression and anxiety). Openness reflects the richness of the individual imagination, aesthetic feelings, degree of dedication, and curiosity about new things [18].

Conventional personality assessments use self report inventory [23]. Although inventory method can accurately return the personality score of the participant with a profound theoretical basis, it still has many deficiencies. Firstly, self-report method is inefficient, as it costs a great deal of manpower and material resources in large-scale experiments. The answers of returned inventories need to be manually inputted into

*Corresponding author. E-mail: tszhu@psych.ac.cn.

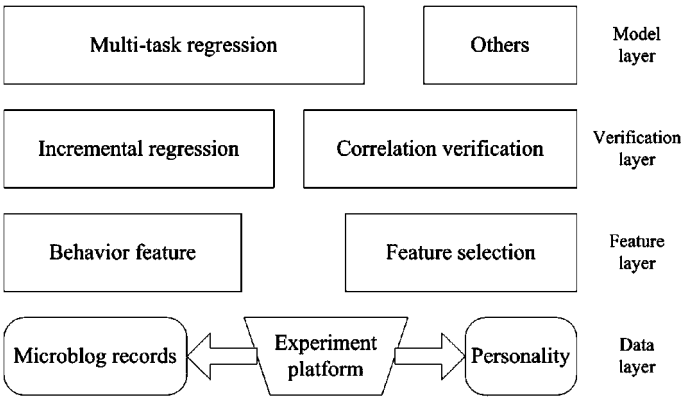

Fig. 1. Hierarchical treatment scheme.

the calculators. To avoid the manual mistakes in the typing work, usually, the third party review is needed to check out the whole process. Secondly, it takes long time to complete inventory survey. The returned questionnaires require manual processing and labeling calculation. No matter how fast the manual handlings are done, it still needs several days to get the results in hundred-user survey. Hence, the inventory method cannot return the results in time.

To address these issues, much research focuses on the online environment [3,4,16,20], such as microblog sites. The development of information technology provides a new method to conduct the research on personality and behavior [32]. Microblog becomes a popular tool for social communication [1,38]. According to China Internet Network Information Center, the number of Chinese microblog registered users is 370 million until July, 2013 [14]. Since its initial launch in December 2005, Sina Microblog has been the leading microblogging service provider in China, and has become the destination for both uploading and downloading resources by status and photo/video-sharing. Sina Microblog, launched by Sina.com, is a microblog site which provides a Twitter like microblogging service. Users can upload pictures or messages via the web pages, WAP pages, mobile client, Short Message Service (SMS), Multimedia Message Service (MMS). Sina Microblog can be understood as a "micro-blog" or "one sentence blog". Users can compile what they see, hear or think to sentences or a picture which can be shared anytime, anywhere via computer or mobile phone to a friends. Meanwhile, as a social networking sites, users can follow friends, and watch the messages released by their friends in real time. Besides, Sina Microblog also releases its Application Program Interfaces (APIs) [37] for third party application development (http://open.weibo.com). After acquiring the authorization of APIs, users' behavior data can be downloaded automatically.

Current personality related research [33,47] pays attention to variation relationship between personality and other variables. In classical personality research, people try to find its influence on behavior [46]. That is to conclude the special kind of behavior of some people with special personality. Recently, the development of information technology brings a new idea for psychology. Web behavior, a special behavior, attracts many scholars. However, the automatical prediction of personality is still in the beginning.

Therefore, this paper proposes to predict the big-five personality traits from user's microblogging behaviors by means of hierarchical treatment scheme shown in Fig. 1. Using Sina Microblog APIs, a large scale of user behavior data with users' personality labels is downloaded in data layer. There are 29 Microblogging based features designed to predict and compute user's big-five personality traits in feature layer. In verification layer, in order to evaluate the quality of the dataset, this paper compares the predicting effect between incremental regression and linear regression, and proves that our processed dataset is reliable. In the practical application of machine learning, it is very often to encounter multiple related tasks learning problem. The tasks share the same training set, and relatively independent but with some contact. The traditional idea trains model respectively for each task. This method considers only the information on each task, and ignores the correlation or shared information between tasks. In order to improve learning outcomes, the multi-task learning [11] is used in this paper. On the dataset, relevant factors exist between the five dimensions of personality [26], hence, a personality pre-

diction model based on multi-task regression is established which works better than the other models in model layer. The main contributions of this paper are as follows:

- This paper manage to predict Sina Microblog users' big five personality through analyzing online behavioral characteristics. Setting user's online behavior as independent variable while personality score as dependent variable, this paper proposes a novel personality prediction method.
- The multi-task regression algorithm is used to predict the Big Five personality, and improve the prediction accuracy. Since big five personality is a multi-dimensional vector, this paper compares different algorithms and proves that multi-task regression works better than sing task regression.

The following part of this paper is organized as follows: Section 2 narrates the collection of online data and personality traits in detail. Section 3 presents the algorithm flow of incremental regression and multi-task regression. The experiment results are shown in Section 4. In Section 5, a detailed discussion is made for the whole steps. Section 6 is the related work by other research groups. Finally in Section 7, we make the conclusion about this research and look forward to the future work.

This paper is an extended version of our paper at the 2013 IEEE/WIC/ACM International Conference on Web Intelligence [7]. In this paper, the details of the experiment are described, such as the selection strategy of the participants and the selection of microblogging features. Besides, more discussion have been done on the experimental results, including how the algorithm parameters are selected, how the online features are related with personality, the practical application of personality prediction and so on.

2. Material and data preparation

The hierarchical treatment scheme shown in Fig. 1 is the core frame of this research. In this research, the whole system can be separated into four layers according to the different functions.

The main function of the bottom data layer is to build a large set of users' microblogging records. Luckily, Sina Microblog opens the APIs for programmer automatically collecting users' microblogging data. Besides, users' personality situations are essential to label their records. Therefore, this research develops an online experiment platform to meet the above requirements. In feature layer, 45 Microblogging related features are designed according to other related studies [16,24]. Meanwhile, stepwise regression is used to conduct feature selection. At the end, 29 features are selected as the characteristic space. Verification layer is designed to estimate the quality of the dataset and verify the precondition of using multi-task regression. For the microblogging data, the comparison of predicting effect between incremental regression and linear regression can prove whether our processed dataset is reliable or not. For the personality tasks, they share the same training set, and relatively independent but with some under verified contact. Once the contact is proved, multi-task regression can be suitable for big five personality prediction. In model layer, a personality prediction model based on multi-task regression is established which works better than the other models.

2.1. Survey design

An online experiment platform is implemented for the survey. The flow chart of the platform is shown in Fig. 2. The platform is a Web Access Connection to Sina Microblog, in which participants could log in the platform with Sina Microblog account. Respondents are invited through the "@" function of microblogging status. They accessed the questionnaire through their Sina Microblog accounts, and firstly are asked to consent to participate in the study and assert that they are 18 years or older. Due to the ethical issues of survey study, participants need to inform consent for the experimental process. In our study, the purpose of the experiment is told to all participants and they can choose to take part in the survey freely. Figure 3 is the display of our informed consent. Once participants click the button "Agree", that means they agree to take part in the experiment, and authorize our experiment platform to download their Microblog behavior data (profile and status) with the usage of APIs which can be used to describe their Microblog usages. They are also instructed to complete the Big-Five Inventory to get their personality scores.

This research download user's basic online behavior records with user API "users/show". This API can return user's information with user's Sina ID number. The returned information is a long string in Java Script Object Notation (JSON) format. A demo for the usage of this API is "http://open.weibo.com/tools/console?uri=users/show".

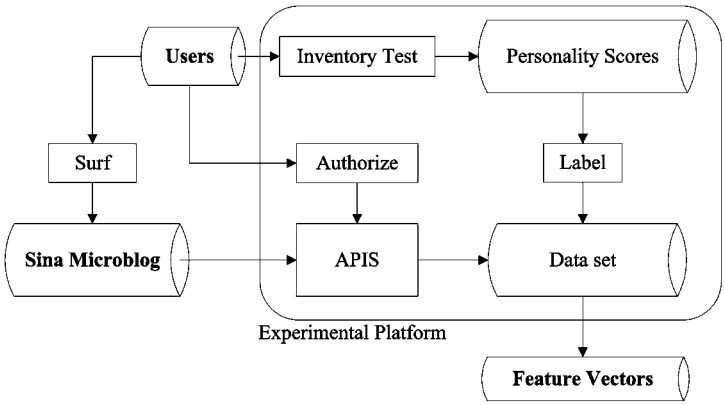

Fig. 2. Flowchart of data collection.

知情同意书

- **研究目的**：您今天参加的是中国科学院心理研究所关于微博使用的问卷研究。
- **研究程序**：整个问卷包括10部分，大五人格测试、心理健康测试、主观幸福感测试、情绪智力测试、微博使用问卷、微博使用动机问卷、微博使用持续意愿问卷、个人信息调查、手机使用情况问卷等。问卷填写总共需要大约三十分钟。在实验过程中您需要集中注意，按照自己的实际情况认真填写。
- **报酬**：认真完成所有问卷10部分内容，并经过研究人员核实数据为真实有效后，会得到30元手机话费作为酬谢，话费将通过支付宝或其他电子商务平台充入您指定的手机账户（您填写的问卷通过审核之后，工作人员会在微博上邀请您来填写手机号码）。
- **风险/不适**：本实验是在网上填写问卷，不会对您的身体造成任何伤害。但是考虑到未成年人隐私的保护，以及未成年人不具有完全民事行为能力，本实验暂不接收未成年人的参加。
- **参与的自愿性与机密性**：参与本测试是完全自愿的，您可以在任何时候按自己的意愿退出测试。
  实验所得到的科学资料，可能会出版在科学界的文章里，但这些文章里不会含有您的姓名或任何有关您身份的信息。为了保护您的隐私，所有资料都以编号，而不以姓名存档。所有记录的机密性，将依法达到最高程度。若您的个人信息被用作它途的话，由中科院心理所承担相关法律责任。
- **协议书文件**：您在网上通过点击"同意"按钮，与在您亲笔签名的协议具有同等效力。当您点击下面的"同意"按钮时候，即表明您已经申明以下内容：
  1. 我已经阅读并知道了协议书的全部内容。
  2. 我已经知道了参与该项测试是完全自愿的，我可以在任何时候退出测试而不受到任何惩罚。
  3. 我已经知道了从本研究中收集的信息将完全保密，并不会以任何能识别个人身份的方式公开，除非得到我个人的同意或国家法律需要。

关于支付报酬的说明

1、您写问卷后应获得报酬，将由中国科学院心理研究所向您以手机话费的形式支付；

2、请您填写完成本应用中的问卷，我们将依照下面的规则向您支付报酬：

a、合格地填写**完成全部的10个问卷**后，您将可以获得价值**30元**的报酬；

b、如果未填写完全部10个问卷，您将得不到任何报酬；

c、本实验有明确的时间限制，实验开始时间为：2012年5月1日；有效参与实验的时间为从被试收到本课题组的邀请开始计算，**一周内**完成有效，并且最迟截止至2012年5月31日24时，如逾期不完成规定的实验任务，则**不会**得到**任何**的被试报酬；

d、对于经过研究人员审核，认定为合格填写问卷的用户，本应用会通过微博已@的方式邀请用户填写手机号，话费将会通过支付宝或其他电子商务平台充入用户填写的手机账号。

3、有偿填写问卷只针对本应用邀请的用户，其余用户不享受有偿填写；任何用户（自然人）最多获得一次填写报酬；以任何黑客手段对本应用进行攻击，尝试将自己加入邀请名单的用户，不享受填写报酬，并且我们保留追究攻击者责任的所有权利；

4、我们邀请您填写问卷的目的是获取有效研究数据，支付报酬是以您认真填写真实数据为前提的；关于支付报酬的条款，最终解释权归支付方所有。

我已阅读并同意**心理测试知情协议**，并且知晓在本页勾选和点击"同意"按钮的效力等同于填写纸质"知情同意书"。

同意 决回

Fig. 3. Informed consent.

2.2. *Participants*

The survey was conducted during April 18, 2012 to May 12, 2012. Participants were invited to log in our platform online and finish the personality inventory test. From the 562 Sina Microblog participants, 444 (171 males and 273 females, 18 years old or elder) with an average age of 23.8 were recruited. The basic microblog usage of participants is shown in Table 1.

Some samples were dropped because of the inactive usage for Sina Microblog or carelessness in finishing the inventory.

- Disqualified users, such as online "water army" or "waistcoat" accounts, which have few pieces of original status published, are eliminated from our experiment. These accounts are normally registered to brush the popularity of some topics or some celebrities, and show little personal intentions. This research defines qualified participants as those that have new status published within the last one month.
- VIP users shall be also removed, since most of these accounts are named by public characters or social organizations but managed by their brokers. This kind of accounts publish statuses fre-

Table 1
Basic microblog usage of participants

Item	Interval	Count	Proportion
Follower count	<100	50	11.26
	100–300	155	34.91
	301–600	141	31.76
	>600	98	22.07
Friend count	<100	58	13.06
	100–300	242	54.50
	301–600	102	22.97
	>600	42	9.46
Status count	<1000	66	14.86
	1001–2000	177	39.86
	2001–3000	122	27.48
	>3000	79	17.79

quently with special objectives. The content of these massages are about the entertainment news, policy decisions or advertisements, which have nothing do with personality.

– The platform can record the exact time point participants submitting the answer. With this function, little care in answering the inventory, such as finishing the inventory in an extremely short time, can be detected. These samples are removed as well to keep the reliability of the dataset.

2.3. Measures

In this study, big-five personality is measured as ground truth labels. The score of each dimension is linearly mapped within -1 to $+1$. This research uses the Chinese version of John O.'s 44-item Big Five Inventory (http://www.ocf.berkeley.edu/∼johnlab/bfi.htm). It contains five subsets, measuring Agreeableness (Agre.) with 9 items, Conscientiousness (Cons.) with 9 items, Extraversion (Extr.) with 8 items, Neuroticism (Neur.) with 8 items, and Openness (Open.) with 10 items, respectively.

BFI is one of the most generally used brief measures of the big-five personality [25]. The questionnaire has been generally used for its high reliability and validity. The items in the inventory consist of short and easy understanding phrases to assess the prototypical traits of the big-five dimensions. Participants evaluate themselves on each question with a 5-point likert scale, ranging from Disagree strongly (1 point) to Agree strongly (5 points). When the participant finishes the scale, the scores of the five subsets can be obtained. Then, participants five dimension personality can be represented with the average score of each subset. The boxplot in Fig. 4 is the distribution of participants' personality scores of each dimension in this study.

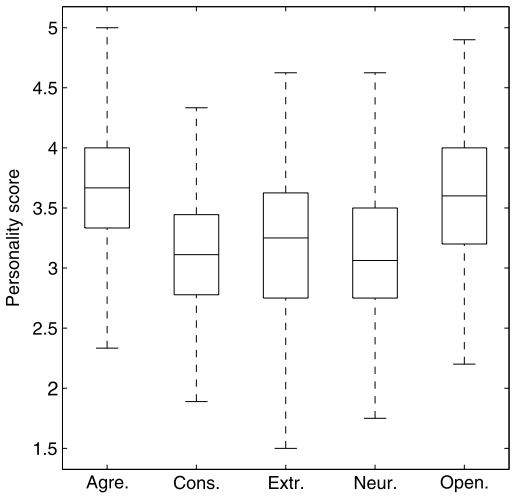

Fig. 4. Personality score.

Take item one as an example, "I think I am talkative". According to the description, participants assess the level of agreement or disagreement of themselves from "A. Disagree strongly", "B. Disagree", "C. Neutral", "D. Agree", and "E. Agree strongly". During subsequent processing, answers of the scales are coded for scoring. In detail, the answers A to E are coded as 1 to 5, respectively. For each subset, the average score can be used to stand for the personality trait of the user. Furthermore, this research linearly maps personality traits to the interval $[-1, 1]$ for experimental analysis.

2.4. Behavioral features

Forty five features are extracted initially, of which twenty nine are selected using stepwise regression. The twenty nine features can be categorized into 4 groups. The detailed information of features are shown in Table 2.

The first group includes 4 features about the *profile* information of participants. This research extracts user's gender, hometown information, and register date. For gender, the feature values are coded as 0 for females and 1 for males. Hometown, namely province and city, are coded as first-tier cities, second-tier city and third-tier city. This research uses the life (at May 12, 2012) of the account to stand for register date.

Table 2
Online features in Sina Microblog

Group	Count	Example
Profile	4	hometown ID (province ID, city level), gender, register date
Self-presentation	7	length of screen name and self-description, whether user's domain is the same as his blog address
Security settings	7	comments available or not, type of verification, whether user's domain is default
Social networking	11	number of friends, followers, mutual followers and statuses, and the proportion of the original status

Seven features are used to figure up how the user presents to others and form the *self-presentation* group. In this group, feature values of length of screen name and self-description are the count of characters in screen name and self-description. For them, the proportions of Chinese characters are extracted. This study also extracts whether the participant uses "I" (Chinese word "wo") in description or not. Another feature is designed to express whether user's domain is the same as his blog address. The length of domain is extracted as well.

Group *security settings* contains 7 features about the privacy settings. This study extracts user's availability of comments and private letter. Moreover, whether the user is verified, reason and type of verification are extracted. Others, such as whether user's domain and large avatar is default are extracted and coded with binary number. As a whole, this group describes the openness and sense of curiosity of the user.

At last, the *social networking* group, having 11 features, is defined as the online interaction of the user. In this group, number of friends, followers, mutual followers, tags, trends and statuses are extracted directly. Microblogging status can be the original message written by the user, or the republished massage from other users. Therefore, the proportion of the original status is extracted. This study also extracts the features on whether the friends, followers, and republished status is classified and grouped or not as well as whether the participant has Sina blog or not.

3. Regression models

The prediction of personality can be regarded as a model fitting process. In this paper, two modeling approaches are adopted, incremental regression and multi-task regression. Through sorting the sample points into an array, incremental regression initially builds a local prediction model. Once a next sample comes, the error rate of the sample in the existing model can be calculated. According to the magnitude of the error, the algorithm has two choices: one is to amend the existing local model using the new sample, the other is to create a new model from this sample. Nevertheless, big five personality labels are multivariate vectors. There exists weak correlation between tasks. Therefore, multi-task regression learning [5] is also used to predict the personality. Different from typical regression, multi-task regression puts all labels together, and tries to model all tasks in a whole. As a result, the model consists both the specific information of each task and the shared information among tasks.

3.1. Incremental regression model

Incremental regression is a linear regression which can be used to fit complex non-linear problems as shown in Algorithm 1.

Before training, this regression algorithm needs some pre-process on the training dataset. Specifically, each feature dimension needs to be normalized before sample sorting. Facing to the different situations, the sorting strategy varies a lot. Normally, it sorts the samples according to the normalized magnitude of each sample vector from small to large. Then it starts with one end of the sorted sample set and builds a simplest local regression model with fewest (n) samples, where n is selected as the minimum sample amount for modeling. For example, if the target is a regression problem in two-dimensional space, n will be set as 2. Next, it tests the error of the model with the next sample in the sample queue. If the error is less than the threshold, the algorithm will refit the model with this new sample. Otherwise, the local model together with its domain will be saved into the line array. Meanwhile, a new local model will be built again with the next fewest samples from the upper test sample.

3.2. Multi-task regression model

The main objective of the multi-task learning is to use multiple task modeling strategies to improve performance beyond the single-task learning in the same scene.

Algorithm 1 Incremental regression algorithm

Require:
Ranked points, Min dimension n, Error threshold ε
Initialize point array, line array, dataset (size N)
Preparation
Ranked points = Sort(Normalize dataset)
repeat
Put first n samples from ranked points into point array
Model = Line Regression(point array)
while error $> \varepsilon$ **do**
test sample $TS = (Point\ array)_{n+1}$
Put the next one sample into point array
error $=< Label(TS), Model(TS) >$
$n = n + 1$
end while
Put the last one sample back to the ranked points
Save the model into line array
Delete point array from ranked points
Clear point array
until No samples exist in ranked points

Assume there are T tasks and N instances. Each instance is represented as a column vector $\mathbf{x}^T \in \mathcal{R}^m$ (m features) and paired with a multi-dimensional output vector $\mathbf{y}^T \in R^T$ (T tasks). Therefore,

$$\mathbf{X} = \left[{\mathbf{x}_1}^T, {\mathbf{x}_2}^T, \ldots, {\mathbf{x}_N}^T\right] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mN} \end{bmatrix} \tag{1}$$

and

$$\mathbf{Y} = \left[{\mathbf{y}_1}^T, {\mathbf{y}_2}^T, \ldots, {\mathbf{y}_N}^T\right] = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ y_{T1} & y_{T2} & \cdots & y_{TN} \end{bmatrix} \tag{2}$$

The objective is to find an optimal coefficient matrix $\mathbf{W}$

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{T1} & w_{T2} & \cdots & w_{Tm} \end{bmatrix} \tag{3}$$

such that

$$\mathbf{W} = \arg\min_W \{\hat{Y} - WX\} \tag{4}$$

and

$$y_{ij} = w_{i\cdot} \bullet x_{\cdot j} = \sum_{k=1}^{m} w_{ik} \bullet x_{kj} \tag{5}$$

Starting with an multi-task learning formulation that jointly considers T regressors, the object function is in a way similar to the primal form of the loss function as

$$\min_W \big(L(X, Y, W; 1:T) + \lambda\Omega(W)\big) \tag{6}$$

where $L(X, Y, W; 1:T)$ denotes the empirical loss function, $\Omega(W)$ is the regularization term, and λ is a trade-off constant.

In this study, $L(X, Y, W; 1:T)$ is set as the least square loss and the regularizer is set as Frobenius norm. That is,

$$L(x, y, \mathbf{W}; 1:T) = \sum_{t=1}^{T} \sum_{n=1}^{N} \left(y_{tn} - \sum_h w_{th} x_{hn}\right)^2 \tag{7}$$

$$\Omega(W) = tr\big(W^T W\big) \tag{8}$$

Substitute into Eq. 6, it has a unique optimizer,

$$W^* = \big(X^T X + \lambda I\big)^{-1} \big(X^T Y\big) \tag{9}$$

The trade-off constant λ can be selected in many ways. Here the bias-variance decomposition [10] is chosen which minimizes the expected loss $(bias)^2 + variance$. This model with the optimal predictive capability is the one that leads to the best balance between bias and variance. In this paper, there are totally 29 features. As a result, m is set as 29. While there are five personality prediction tasks, T is set as 5. Therefore, the coefficient matrix $\mathbf{W}$ is a 5×29 matrix.

4. Results

This paper experiments on how personality reflects Sina Microblog behaviors. On one hand, the associated modes of users' personality and network characteristics are found. On the other hand, through different machine learning algorithms, computational models of personality based on network characteristics are

Table 3
Pearson correlations of personality traits

	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
Agreeableness	–	.25**	.11*	−.41**	.18**
Conscientiousness	.25**	–	.26**	−.40**	.29**
Extraversion	.11*	.26**	–	−.35**	.32**
Neuroticism	−.41**	−.40**	−.35**	–	−.16**
Openness	.18**	.29**	.32**	−.16**	–

**$p < 0.01$. Correlation is significant at the 0.01 level (2-tailed).
*$p < 0.05$. Correlation is significant at the 0.05 level (2-tailed).
$N = 444$.

built. In order to test the performance of the different models, 5-fold cross validation is used for modeling, linear regression is chosen as the baseline, and Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) [41] are selected as the assessment criteria.

4.1. Correlation analysis

Multi-task regression takes all personality dimensions as a whole to build a novel model. However, a basic premise is that there exists correlation (even weak correlation) between the big five personality dimensions. Therefore, it is necessary to analyze the Pearson Correlation Coefficient between personality dimensions which is shown in Table 3. In an intuitive understanding, Pearson Correlation Coefficient describes the degree of tightness between two fixed variables and defined as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (10)$$

where n is the total sample size, X_i and Y_i are observations, $\bar{X}$ and $\bar{Y}$ are mean values.

The Pearson Correlation Coefficient r describes the degree of linear correlation between the two variables. The value of r ranges between -1 and $+1$. If r is positive, the two variables are positively correlated (the greater the value of one variable, the greater the value of another variable). If r is negative, the two variables are inversely related (the smaller the value of a variable, the greater the value of another variable). The absolute value of r stands for the strength of the correlation.

Table 3 shows that agreeableness, conscientiousness, extraversion, and openness pairwise have a significant positive correlation, especially for extraversion and openness ($r = 0.32$). However, neuroticism is significantly negative correlated with the other four dimensions, especially for the correlation factors with agreeableness and conscientiousness (above 0.4).

4.2. Personality traits and online features

This study extracts 29 microblogging features. To find their degree of correlation, Pearson Correlation Coefficient is computed in Table 4.

From the table, agreeableness has significant ($r = 0.11$) positive association with the proportion of original status significantly which is the same for conscientiousness ($r = 0.15$, $p < 0.01$). Besides, conscientiousness has positive association with follower count ($r = 0.09$). Males ($r = 0.15$, $p < 0.01$) and verified ($r = 0.10$) users are significantly more conscientious. Extraversion has positive association with original status proportion ($r = 0.15$), the length of screen name ($r = 0.11$), description ($r = 0.11$) and domain ($r = 0.12$, $p < 0.01$), as well as the count of friends ($r = 0.19$, $p < 0.01$) and mutual followers ($r = 0.12$, $p < 0.01$) significantly. However, it has negative association with the Chinese word proportion in screen name ($r = -0.10$). Neuroticism has negative association with domain length ($r = -0.10$) and is more significant in female users ($r = 0.13$, $p < 0.01$). Finally, openness has positive association with screen name length ($r = 0.11$), tag count ($r = 0.12$, $p < 0.01$), classification of friends ($r = 0.18$, $p < 0.01$) and followers ($r = 0.10$), and the proportion of original status ($r = 0.18$, $p < 0.01$).

4.3. Regression accuracy

This paper tries to build the personality prediction model. To achieve this, 29 microblogging usage features are extracted. Using incremental regression shown in Algorithm 1, at least 30 samples are needed to train a local model, as a result, n is set as 30. According to the normalized magnitude of each sample,

Table 4
Pearson correlations between personality traits and online features

	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
Province ID	−0.03	−0.04	−0.02	0.03	0.04
City level	−0.03	0.00	0.00	0.01	0.02
Gender	0.06	0.15**	0.01	−0.13**	0.05
Life of account	0.07	−0.02	0.00	0.04	0.05
Screen name length	0.02	0.04	0.11*	−0.04	0.11*
Description length	0.00	−0.04	0.11*	−0.04	0.03
Chinese word proportion of screen name	−0.04	−0.09	−0.10*	0.06	−0.01
Chinese word proportion of description	0.07	0.07	0.05	−0.07	0.01
Does description have "I"	0.03	0.02	0.04	−0.01	0.07
Is domain same as blog address	0.02	0.09	−0.01	−0.02	−0.01
Domain length	−0.02	0.01	0.12**	−0.10*	0.05
Comment availability	0.02	0.04	0.03	0.06	−0.02
Private letter availability	0.01	0.02	0.01	−0.04	0.03
Verified or not	0.02	0.10*	0.04	−0.07	0.03
Verification reason ID	0.02	0.08	0.03	−0.06	0.02
Verification type ID	0.08	0.06	0.08	−0.04	−0.05
Default domain or not	−0.04	0.09	−0.01	0.08	0.04
Default avatar or not	0.02	0.00	0.06	−0.07	−0.03
Friend count	0.03	0.06	0.19**	−0.05	0.04
Follower count	−0.02	0.09*	−0.03	−0.01	0.01
Mutual follower count	0.09	0.08	0.12*	−0.08	0.03
Tag count	0.05	0.02	0.01	−0.03	0.12**
Trend count	0.08	0.08	0.05	0.01	0.05
Status count	0.03	0.06	−0.01	0.03	0.07
Have Sina blog or not	0.05	0.07	0.09*	−0.02	−0.01
Proportion of original status	0.11*	0.15**	0.19**	−0.04	0.18**
Are friends classified	0.01	0.06	0.06	0.03	0.18**
Are followers classified	0.06	0.06	0.05	−0.07	0.10*
Are republished statuses classified	−0.02	0.08	0.06	−0.03	0.02

**$p < 0.01$. Correlation is significant at the 0.01 level (2-tailed).
*$p < 0.05$. Correlation is significant at the 0.05 level (2-tailed).
$N = 444$.

the samples in dataset are sorted. The error threshold is set as 0.10. A local model with the first 30 samples is built which is used to test the error of the following sample. If the error is smaller than the threshold, these 31 samples are grouped together to build a new local model. Repeat this process until the error of a sample is larger than the threshold. At this time, both the local model and its domain are saved. After traversing all the samples, a group of local model will be returned and compose a global prediction model. Table 5 shows the MAE and RMSE of predicting personality on Sina Microblog.

The above analysis proved that personality dimensions have moderate correlation relationship with each other. When predicting psychological properties, this relationship needs to be taken into account to improve the performance. Therefore, multi-task regression is used here. Two comparable prediction systems are set up. One is the single task linear regression model which uses the online behaviors of participants to predict personality. This is the baseline of the experiment. The other one is the multi-task regression model which can figure up the common information between tasks as well as the special information of each task. Minimizing $(bias)^2 + variance$, λ is selected as 0.0446 ($\ln(\lambda) = -3.11$) for each dimension shown in Figs 5 and 6. The MAE and RMSE are shown in Table 5 and Fig. 7.

Table 5
MAE and RMSE of different algorithms for each dimension

Dimension	Single task regression		Incremental regression		Multi-task regression	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Agreeableness	0.1881	0.2306	0.1785	0.2203	0.1034	0.1285
Conscientiousness	0.1899	0.2375	0.1680	0.2153	0.1352	0.1667
Extraversion	0.2086	0.2585	0.1989	0.2517	0.1285	0.1598
Neuroticism	0.2196	0.2656	0.2017	0.2537	0.1252	0.1605
Openness	0.2021	0.2521	0.1768	0.2204	0.1319	0.1639
Average	**0.2017**	**0.2489**	**0.1848**	**0.2323**	**0.1248**	**0.1559**

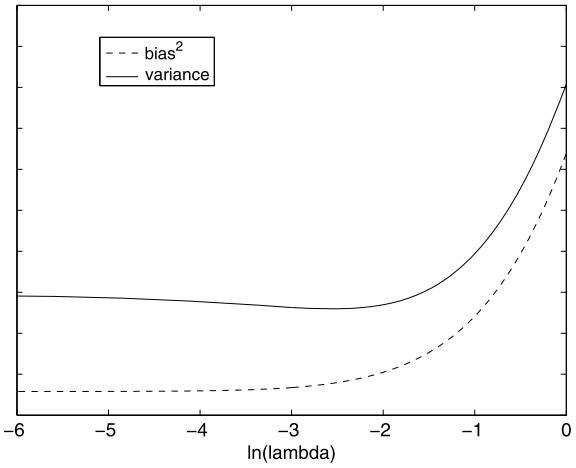

Fig. 5. Bias and variance.

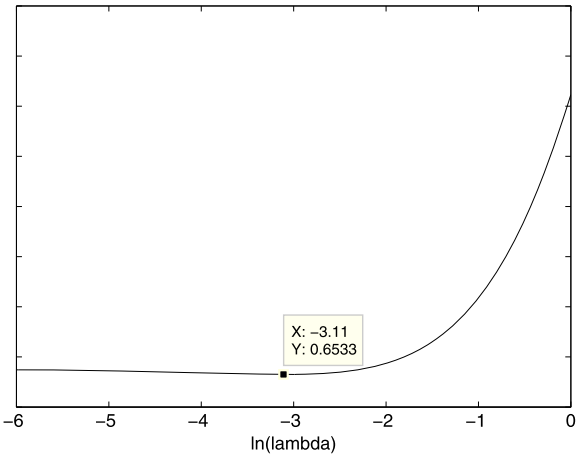

Fig. 6. Optimal parameter selection.

5. Discussion

The previous experiment tries to predict personality on Sina Microblog. Three models are established with single task as the prediction baseline and other two model are our novel methods.

5.1. Data reliability

As shown in the last row in Table 5, the simplest single task linear regression works worst (percentage MAE is 20.17). Although incremental regression is still a broad linear regression, it can train and build several local optimal models, which can be more powerful than the single task regression model (percentage MAE is 18.86). However, incremental regression has its own limitations. First, it relies too much on the sorting strategy. Since sample sorting is the first step of the algorithm, and the following local model establishment depends on the sorted sample, without a better sorting method, the algorithm will not get a responsible result. Second, the algorithm is quite sensitive with the noisy samples. When a noisy sample is coming to the test of the local model, the test error will get an extremely large value. Even though the next samples are all regular, the domain of this local model will still get an end.

Figure 8 shows an example to illustrate noise sensibility of incremental regression towards linear regression. In this example, samples are in two dimensional coordinate system with $\langle x_i, y_i \rangle$ pairs. The purpose is to find a best fitting for $y = f(x)$. The two sub-pictures in the upper row demonstrate the difference between the linear regression and incremental regression in an ideal dataset. In the left upper graph, although there is no noisy sample intuitionally, linear regression can still not learn the pattern and fit badly. The reason for the bad performance is that independent variable X is not linearly related with dependent variable Y. Yet, incremental regression focuses on different local optimums, and uses several linear model to fit the nonlinear pattern, which works relatively better than simple linear regression. Even so, incremental regression is too much sensitive with noisy samples as shown in

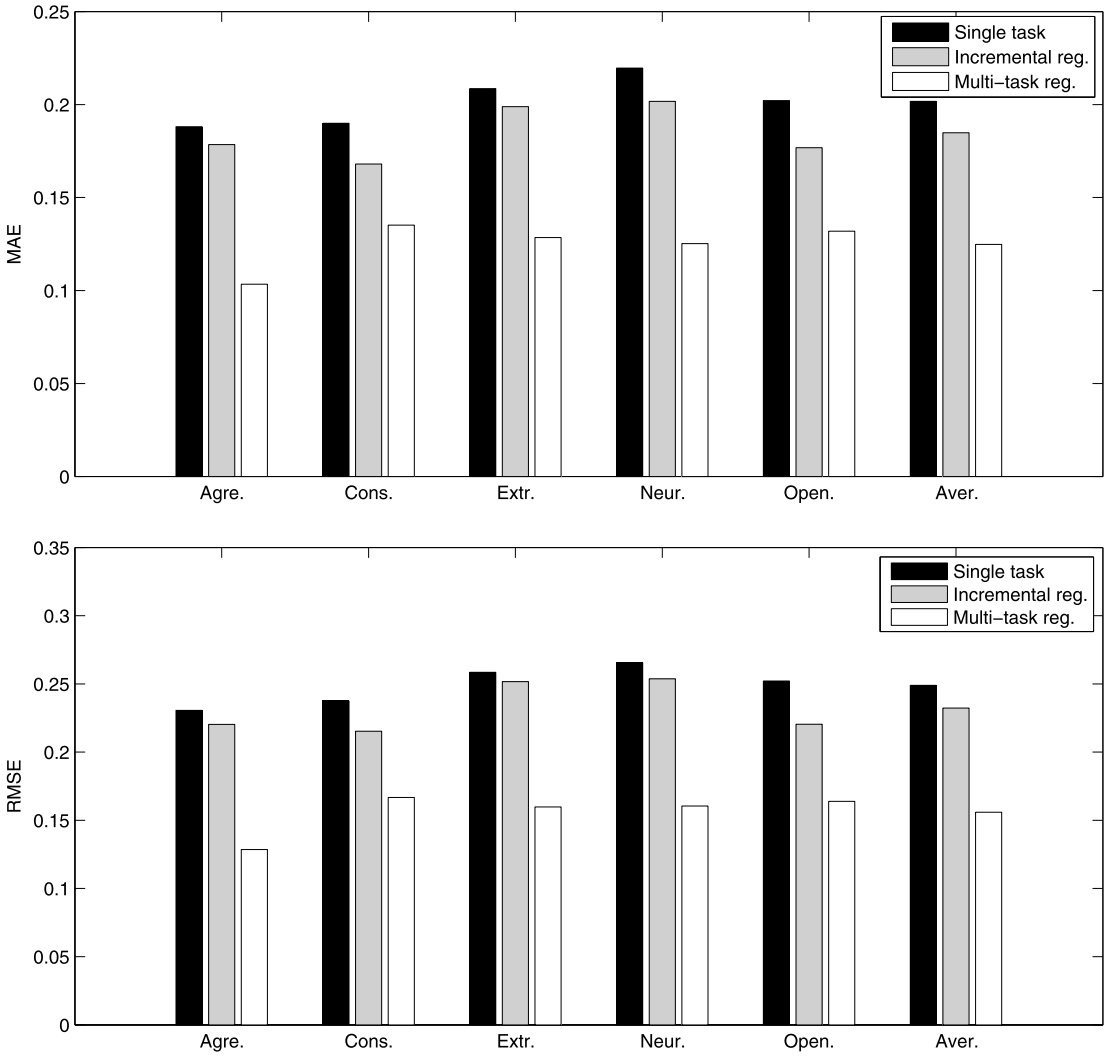

Fig. 7. MAEs and RMSEs of different different algorithms on big five personality dimensions.

the lower two sub-pictures in Fig. 8. For linear regression, some noise sample can be tolerated and has little influence for the model. However, this is not the same for incremental regression, since noise sample directly influences the local optimums and brings a relatively larger error.

In this example, there is only one noise sample, but brings big error for modeling. If more noise samples exist in training dataset, the performance of incremental regression will descend susceptibly. Therefore, in the personality prediction problem, it is extremely important to identity the noisy samples and remove them out of the dataset.

In the actual process of conducting online survey, the time nodes of answering each question of the personality inventory are recorded. Time nodes can be used for noisy samples checking which can not be achieved by the traditional offline psychological investigations. In this research, the following operation are carried out on the raw data:

- If the time spent for one question is less than one second, remove the sample. Usually, it is impossible to answer the question in such a short interval.
- If the answer distribution or answer array shows a clear regular pattern, remove it. The answer option are denoted with English letters A to E, this research removes the samples with "AAAA…", "AABBCC…", "ABCD…", "DCBA" or "ABCDCBA" patterns.

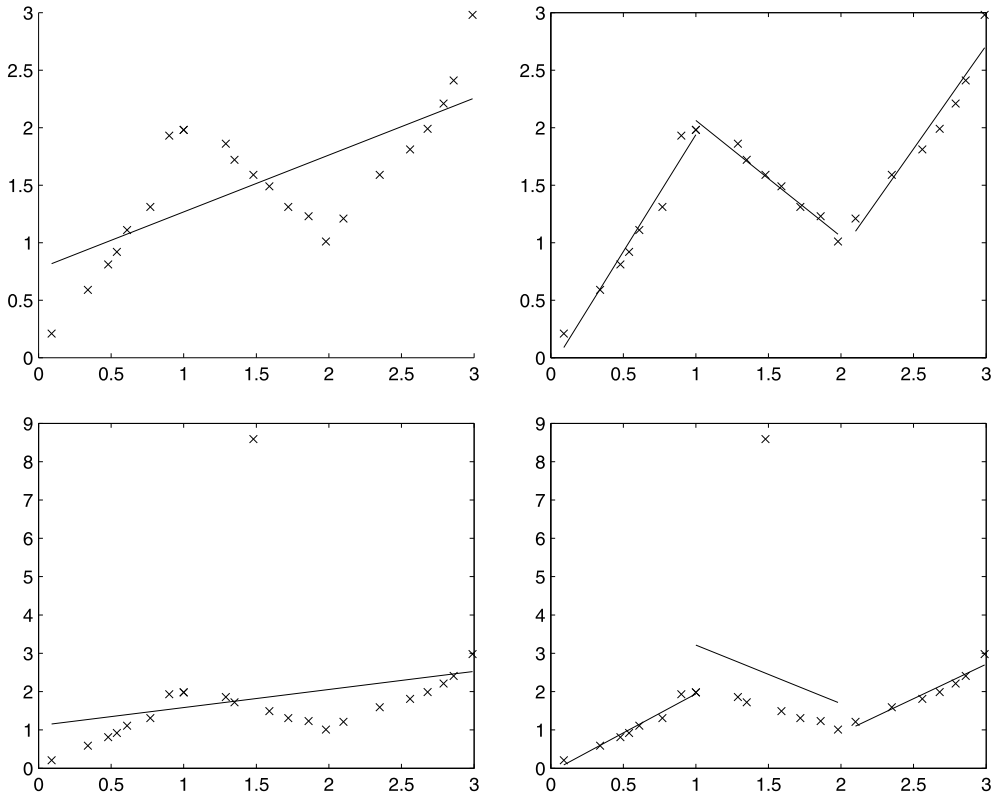

Fig. 8. Incremental regression and linear regression.

In order to keep the quality of the samples, participants are informed to consent to our demands. Once they decide to take part in our experiment, their answers will be carefully checked and a fee will be only given to the participants who passed the noise check. The MAE of the incremental regression gets a nearly 2 percent reduction which proves that our dataset is reliable.

5.2. Personality and online features

Since personality is multi-dimensional, and there exist moderate correlations among the dimensions, multi-task regression is used to build the personality prediction model. Results show that the average MAE of multi-task regression model is 12.48%, which gets more than 6 percentage points reduction compared to incremental regression. In this research, before training the model, each dimension of the features in dataset is normalized. Table 6 lists part of the regression coefficient matrix of the model (not all the coefficients are meaningful). The sign of the coefficient stands for the co-variation between personality and features (if A is bigger, B will be bigger or smaller), since the personality scores are linearly mapped into interval $[-1, 1]$. The absolute value of the coefficient stands for the degree of the covariation (if A is bigger, how bigger B will be), since all the features are normalized in this discussion. Figure 9 shows the distribution of the features in Table 6.

From the table, agreeableness is positively related with number of mutual followers (if A is the follower of B and B is also the follower of A, they are mutual follows), the mutual followers proportion in all followers and original status percentage [31]. That means users with high degree of agreeableness tend to be the mutual followers of other users. If one user follows them, they will follow back to this user more likely. At the same time, they also like to update status more often and most of their statuses are written by themselves (original statuses percentage, 0.1013) not the retweeting others' statuses.

Conscientiousness can be regarded as self-discipline. When reflecting to Sina Microblog, users with high degree of conscientiousness have more mutual followers (0.4508) and more friends online (0.1157), although their follower number is not in large scale

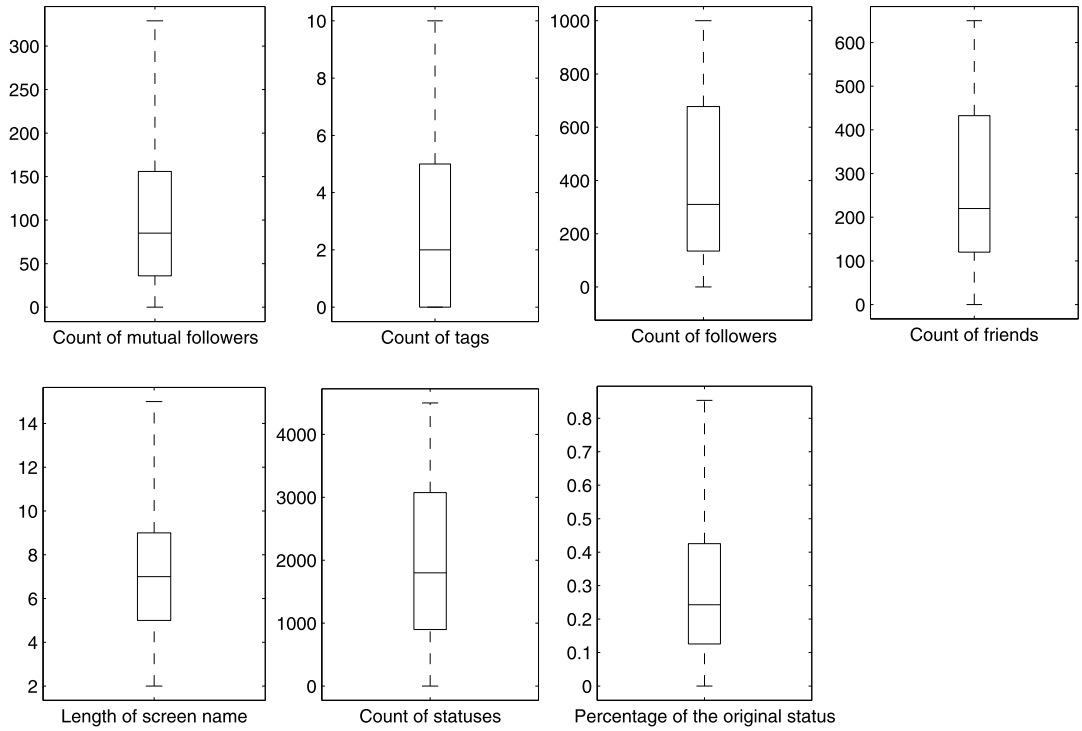

Fig. 9. Distribution of seven contributory features.

Table 6
Part of regression coefficients of multi-task regression for each dimension

Features	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
Count of mutual followers	**0.1921**	**0.4508**	**0.4965**	−0.0720	**0.6041**
Count of tags	−0.0473	**−0.1337**	−0.0121	−0.0470	**0.5999**
Count of followers	−0.0351	**−0.5066**	**−0.1208**	0.0726	**−0.3900**
Count of friends	−0.0116	**0.1157**	−0.0409	−0.0345	0.0122
Length of screen name	0.0699	**−0.1949**	**0.1082**	0.0254	0.0471
Count of statuses	0.0786	**−0.1310**	**−0.1979**	**0.1823**	0.0930
Percentage of the original status	**0.1013**	0.0331	0.0582	0.0244	−0.0172

(−0.5066) [45]. They are well self-controlled and can carefully publish statuses without "watering" online (usually have short screen name, −0.1949), therefore, their status number is relatively small (−0.1310).

Extravert has more mutual followers (0.4965) online and a relatively long screen name (0.1082) [47]. However, since extraverts are more eager to show their charms, they tend to have a more strong sociability. Therefore, extraverts have more friends offline (online friends number is relatively small, −0.1208) and don't need to publish more statues to attract others (−0.1979).

Neuroticism is the stability of emotion. People with high degree in neuroticism are more likely to suffer mental health and do a lot of repetitive operations, therefore, they will update their status quite often (0.1823) [40].

Openness shows the richness of individual imagination and curiosity about new things. People with high score in openness tend to follow others very much and listen to what other users are talking about (0.6041) in spite of the shortness in followers (−0.3900). There are more resources in their tags (0.5999) for their curiosity about online information around them [15].

5.3. Practical application

Personality prediction have broad application prospects in practice. For Internet service provider, per-

sonality prediction can be applied for recommendation system. In existing theories, psychological traits stand for personal preference on resources [28,39,43]. These studies indicate that the desire of online purchase, musical taste, or friend making can be affected by the personality trait of the user. With these foundations, social network service providers can even mine potential community. Therefore, improving user's profile with personality traits could enhance the performance of personalized recommendation system and attract more users.

Previous studies on personality and crime [42] pointed out that the structure of personality traits correlate closely with antisocial behavior [34]. Some of the unusual personality has the tendency to be antisocial or psychopathic. To some extent, personality prediction can help network supervision department find hidden dangerous users and prevent network security events.

Another application prospect is in job hunting. Nowadays, most employers receive resumes through online recruitment platform. Using personality prediction techniques, employers can get a deep understanding of the job hunter which helps employers find the qualified personnel they really need.

6. Related work

This paper is an extended version of our paper at the 2013 IEEE/WIC/ACM International Conference on Web Intelligence [7]. In the previous paper, we aim to measure the Big-Five personality from the usages of Sina Microblog objectively. By conducting a user study with 444 users, multi-task regression and incremental regression algorithms are proposed to predict the Big-Five personality from online behaviors. The results indicate that personality can be predicted with a high accuracy through online Microblog usage. Based on that version, this paper moves forward a step to find the correlation between personality traits and microblogging behaviors. The detailed steps of our experiment are revealed. We believe that due to the popularity of the network, social networking has become a social reality corresponding to online virtual society.

Andrew et al. [44] analyzed 700 million words and phrases collected from the Facebook volunteers, who also took standard personality tests. They built personality classification models based on user's language information. They found that neurotic people disproportionately use the phrase "sick of" and the word "depressed" and males use the possessive "my" when mentioning their "wife" or "girlfriend" more often than females use "my" with "husband" or "boyfriend".

Kosinski et al. [29] researched on psychological factors on Facebook. Their analysis was based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. They used logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminated between homosexual and heterosexual men. They showed that the easily accessible digital records of behavior in Facebook can be used to automatically and accurately predict a range of highly sensitive personal attributes such as personality, happiness and so on.

Gosling et al. [24], conducted experiments towards the manifestations of personality in Facebook usage. They delivered a mapping between personality and SNS online behaviors, and examined the personality with self-reported Facebook usage and observable profile information. They provided the correlation factor between personality and online behaviors. In their experiment, 11 features were used, including the number of friends, weekly usage and other features. Although their research verified the correlation of online characteristics and personality labels, they did not further establish prediction model of personality, or give the quantitative indicators of personality and online behaviors.

Junco [27] studied the relationship between Facebook use and student engagement. He found that Facebook use was negatively predictive of engagement scale score and positively predictive of time spent on SNS. However, his work was based on user's statistic information, such as common friend count, familiar shared resources, time spent on SNS or information checked frequency which considers user's SNS usage instead of her inner preferences and personality.

Correa et al. [16] researched the relationship between use of social networking media and the user's big five personality. By calculating the Pearson correlation factors, they found the use of social networking media was significant positive correlated with openness and extroversion, and negatively related with neuroticism. Their study did not further consider the prediction of personality based on the social media usage.

Moore et al. [35] worked on the influence of social networking usage based on users' Big Five personality. They invited 219 graduate students as a group, and collected user data through self-presentation record-

ing. Results were meaningful, but the self-presentation of online behavior brought the Subjective orientation of the identification of online behavior. It was not strict enough for setting up the sample set in such a way.

7. Conclusion and future work

7.1. Implications

This paper analyzes the personality on Sina Microblog based on the big-five theory. The co-variation of personality and microblogging features is found from the aspects of Pearson Correlation Coefficient and regression coefficient. In order to find a quantitative index for the reliability of the dataset, incremental regression model is used for modeling. Compared with linear regression, incremental regression performances better which proves that the filtering of dataset is reasonable. Meanwhile, with the purpose of improving the prediction accuracy, multi-task regression model is used which enhances the predicting accuracy for more than 30%.

This study finds that personality trait agreeableness is positively related with mutual follower count and the original status proportion. Conscientiousness is positively related with mutual follower count and friend count, but negatively related with tag count, follower count, status count and screen name length. Extraversion is positively related with mutual follower count and screen name length, but negatively related with follower count and status count. Neuroticism is positively related with status count. Openness is positively related with mutual follower count and tag count, but negatively related with follower count.

7.2. Limitations and future work

This research leaves some blanks to be desired in future work. First, we will continue to collect users' data in Sina Microblog, and invite more participants to get a larger dataset. To achieve this, some interesting functions will be added to the platform application. In feature, once the participant finish the inventory, the application will give some feedback information to the user, such as the advice for psychological care or friend recommendation on identical personality. We will try our best to make the application from research to productization.

For features, we plan to design and extract other user network characteristics on Microblog. In this research, all the features are behavior features which describe the online activity of the users. There is no feature for the status text. Therefore, we will work on text mining in Chinese environment. This research field has already attracted many researchers, especially for the study about personality manifestation in text content [6,8,19]. In text mining, we will use text analysis software Linguistic Inquiry and Word Count (LIWC [13,21,22]) to extract content related features. To make it function well in Chinese environment, we will establish the Chinese dictionary ourselves.

At the same time, with the promotion of Sina Microblog, the microblogging status is no longer pure text messages. More and more users upload pictures, videos or other multi-media statuses. We will also work on multi-media mining and try to extract features on multi-media resources.

Another idea of our research is to work on the behavior patterns of other psychological attributes, such as mental health and social attitude. Different from personality, mental health and social attitude [12] are continuously changing variables. The objective for personality prediction is to find its impact on behaviors. However, mental health disorders are illnesses. The purpose for prediction should be the intervention for mental health problems. For social attitude, it can be useful to detect the psychological tendency of the public which can help the administrator to make the decision. We will also try other typical multi-task learning algorithms to tune the performance of the predicting model.

Acknowledgment

The authors gratefully acknowledges the generous support from National High-tech R&D Program of China (2013AA01A606), NSFC (61070115), Institute of Psychology (113000C037), Key Research Program of CAS (KJZD-EW-L04), Strategic Priority Research Program (XDA06030800) and 100-Talent Project (Y2CX093006) from Chinese Academy of Sciences.

References

[1] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, Analysis of topological characteristics of huge online social networking services, in: *Proc. of the 16th International Conference on World Wide Web*, ACM, 2007, pp. 835–844.

[2] Y. Amichai-Hamburger, N. Lamdan, R. Madiel, and T. Hayat, Personality characteristics of wikipedia members, *CyberPsychology and Behavior* **11**(6) (2008), 679–681.

[3] Y. Amichai-Hamburger, G. Wainapel, and S. Fox, "On the Internet no one knows I'm an introvert": Extroversion, neuroticism, and Internet interaction, *CyberPsychology and Behavior* **5**(2) (2002), 125–128.

[4] T. Amiel and S.L. Sargent, Individual differences in Internet usage motives, *Computers in Human Behavior* **20**(6) (2004), 711–726.

[5] A. Argyriou, T. Evgeniou, and M. Pontil, Convex multi-task feature learning, *Machine Learning* **73**(3) (2008), 243–272.

[6] S. Bai, R. Gao, and T. Zhu, Determining personality traits from Renren status usage behavior, in: *Proc. of Computational Visual Media*, Springer, 2012, pp. 226–233.

[7] S. Bai, B. Hao, A. Li, S. Yuan, R. Gao, and T. Zhu, Predicting big five personality traits of microblog users, in: *Proc. of Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences*, IEEE, 2013, 1, pp. 501–508.

[8] S. Bai, Y. Ning, S. Yuan, and T. Zhu, Predicting reader's emotion on Chinese web news articles, in: *Proc. of Pervasive Computing and the Networked World*, Springer, 2013, pp. 16–27.

[9] S. Bai, T. Zhu, and L. Cheng, Big-five personality prediction based on user behaviors at social network sites, arXiv preprint, (2002) 1204.4809.

[10] M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[11] R. Caruana, Multitask learning, *Machine Learning* **28**(1) (1997), 41–75.

[12] L.S. Chen, H.H. Tu, and E.T. Wang, Personality traits and life satisfaction among online game players, *CyberPsychology and Behavior* **11**(2) (2008), 145–149.

[13] N. Cheng, R. Chandramouli, and K.P. Subbalakshmi, Author gender identification from text, *Digital Investigation* **8**(1) (2011), 78–88.

[14] CNNIC, The 32nd Chinese internet development statistics report, 7, 2013.

[15] D.A. Cobb-Clark and S. Schurer, The stability of big-five personality traits, *Economics Letters* **115**(1) (2012), 11–15.

[16] T. Correa, A.W. Hinsley, and H.G. De Zuniga, Who interacts on the web?: The intersection of users' personality and social media use, *Computers in Human Behavior* **26**(2) (2010), 247–253.

[17] P.T. Costa and R.R. MacCrae, Revised NEO personality inventory (NEO PI-R) and NEO Five-Factor inventory (NEO FFI): Professional manual, Psychological Assessment Resources, 1992.

[18] B. De Raad, *The Big Five Personality Factors: The Psycholexical Approach to Personality*, Hogrefe & Huber Publishers, 2000.

[19] R. Gao, B. Hao, S. Bai, L. Li, A. Li, and T. Zhu, Improving user profile with personality traits predicted from social media content, in: *Proc. of the 7th ACM Conference on Recommender Systems*, ACM, 2013, pp. 355–358.

[20] V.P. Goby, Personality and online/offline choices: Mbti profiles and favored communication modes in a Singapore study, *CyberPsychology and Behavior* **9**(1) (2006), 5–13.

[21] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, Predicting personality from Twitter, in: *Proc. of Privacy, Security, Risk and Trust (Passat), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (Socialcom)*, IEEE, 2011, pp. 149–156.

[22] J. Golbeck, C. Robles, and K. Turner, Predicting personality with social media, in: *Proc. of CHI'11 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2011, pp. 253–262.

[23] L.R. Goldberg, The structure of phenotypic personality traits, *American Psychologist* **48**(1) (1993), 26.

[24] S.D. Gosling, A.A. Augustine, S. Vazire, N. Holtzman, and S. Gaddis, Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information, *CyberPsychology, Behavior, and Social Networking* **14**(9) (2011), 483–488.

[25] S.D. Gosling, P.J. Rentfrow, and W.B. Swann Jr, A very brief measure of the big-five personality domains, *Journal of Research in Personality* **37**(6) (2003), 504–528.

[26] Y.A. Hamburger and E. Ben-Artzi, The relationship between extraversion and neuroticism and the different uses of the Internet, *Computers in Human Behavior* **16**(4) (2000), 441–449.

[27] R. Junco, The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement, *Computers and Education* **58**(1) (2012), 162–171.

[28] A. Karsvall, Personality preferences in graphical interface design, in: *Proc. of the Second Nordic Conference on Human-Computer Interaction*, ACM, 2002, pp. 217–218.

[29] M. Kosinski, D. Stillwell, and T. Graepel, Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the National Academy of Sciences* **110**(15) (2013), 5802–5805,

[30] R.N. Landers and J.W. Lounsbury, An investigation of big five and narrow personality traits in relation to Internet usage, *Computers in Human Behavior* **22**(2) (2006), 283–293.

[31] L. Li, Y. Yang, and L. Mingxin, The relationship between adolescents' neuroticism, Internet service preference, and Internet addiction, *Acta Psychological Sinica* **3** (2006), 008.

[32] B. Marcus, F. Machilek, and A. Schutz, Personality in cyberspace: Personal web sites as media for personality expressions and impressions, *Journal of Personality and Social Psychology* **90**(6) (2006), 1014.

[33] M.P. McCreery, K.S. Kathleen, P.G. Schrader, and R. Boone, Defining the virtual self: Personality, behavior, and the psychology of embodiment, *Computers in Human Behavior* **28**(3) (2012), 976–983.

[34] J.D. Miller and D. Lynam, Structural models of personality and their relation to antisocial behavior: A meta-analytic review, *Criminology* **39**(4) (2001), 765–798.

[35] K. Moore and J.C. McElroy, The influence of personality on Facebook usage, wall postings, and regret, *Computers in Human Behavior* **28**(1) (2012), 267–274.

[36] L.C. Morey and S.E. Lowmaster, Personality assessment inventory, *Corsini Encyclopedia of Psychology*, 2010.

[37] F. Morstatter, J. Pfeffer, H. Liu, and K.M. Carley, Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose, in: *Proc. of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013, pp. 400–408.

[38] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, and J. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, *Science* **328**(5980) (2010), 876–878.

[39] E. Neustadt, T. Chamorro-Premuzic, and A. Furnham, The relationship between personality traits, self-esteem, and attach-

ment at work, *Journal of Individual Differences* **27**(4) (2006), 208–217.

[40] D.M. Olvet and G. Hajcak, The error-related negativity relates to sadness following mood induction among individuals with high neuroticism, *Social Cognitive and Affective Neuroscience* **7**(3) (2012), 289–295.

[41] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, Our Twitter profiles, our selves: Predicting personality with Twitter, in: *Proc. of In Privacy, Security, Risk and Trust (Passat), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (Socialcom)*, IEEE, 2011, pp. 180–185.

[42] J.A. Reid, Crime and personality: Personality theory and criminality examined, *Student Pulse* **3**(01) (2011).

[43] P.J. Rentfrow and S.D. Gosling, The do re mi's of everyday life: The structure and personality correlates of music preferences, *Journal of personality and social psychology* **84**(6) (2003), 1236.

[44] H.A. Schwartz, J.C. Eichstaedt, M. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, and M.E. Seligman, Personality, gender, and age in the language of social media: The open vocabulary approach, *PloS one* **8**(9) (2013), e73791.

[45] S. Stieger, C. Burger, M. Bohn, and M. Voracek, Who commits virtual identity suicide? Differences in privacy concerns, internet addiction, and personality between Facebook users and quitters, *Cyberpsychology, Behavior, and Social Networking* **16**(9) (2013), 629–634.

[46] C. Sumner, A. Byers, and M. Shearing, Determining personality traits and privacy concerns from Facebook activity, in: *Black Hat Briefings*, 2011, pp. 1–29.

[47] K. Wilson, S. Fornasier, and K.M. White, Psychological predictors of young adults' use of social networking sites, *CyberPsychology, Behavior, and Social Networking* **13**(2) (2010), 173–177.