Predicting Psychology Attributes of a Social Network User

Rustem M. Khayrullin, Ilya Makarov, and Leonid E. Zhukov

National Research University Higher School of Economics, Moscow, Russia
rmkhayrullin@edu.hse.ru, iamakarov@hse.ru, lzhukov@hse.ru

Abstract. Nowadays, the number of people using social network site increases every day. The social networking sites, such as Facebook or Twitter, are sources of human interaction, where users are allowed to create and share their activities, thoughts and place different information about themselves. However, most of this information remains unnoticed. In this work, we propose a machine learning approach to predict Big-Five personality using information from users accounts from the social network. The predictions can be used in different areas such as psychology, business, marketing.

Keywords: Social Networks, Machine Learning, Psychology, Big Five Personality, Shwartz Human Values

1 Introduction

Today we cannot imagine our life without social media resources. We communicate with our friends, share information with them or just spend our free time in social networks, for example on Facebook. Eventually, this process results in the accumulation of a huge amount of information, which can tell almost everything about the person, since people tend to write a lot about themselves. In our work, we will use information from Vkontakte accounts (a social network, which is popular in Russia, the analogue of Facebook) to predict information about their owners. These predictions can be used in various areas of our lives, such as psychology, business or marketing.

In the present work, we will use the psychological model of the Big Five [2] and the psychological model of Schwartz [8] to predict the portrait of a person. This work is relevant because it allows to solve a lot of problems and has a clear practical application. It is no coincidence that similar works have been conducted for a long time and have become especially popular in recent years. In [10], IBM published the research report on predicting the socio-psychological portrait on the scale of the Big Five based on the analysis of the owner's Twitter account logs. In [6], the authors published a paper on the construction of correlations between the signs of a person in a social network and his psychological characteristics on the "Big Five" scale. However, in the most works character recognition is based on analysing the semantics of the text at the moment (for example, posts on Twitter or statuses in social networks), and the accuracy of

the predictions was extremely low. In the present work, we want to use the various characteristics of a person, which he or she points out in his or her own account in the social network to predict the portrait of the person. The disadvantage of this approach is that very often people do not write full information about themselves or knowingly indicate false information about themselves. In case of a large number of noisy data, the construction of a good predictive model is a complex task. We use machine learning methods to implement predictions. In particular, we have used such methods as Random forest, Gradient Boosting, SVM.

Chapter 2 presents the results achieved to date in this field. In Chapter 3 we give the main definitions and describe predictions of the psychological scales of Big-Five and the psychological scale of Schwartz. In Chapter 4 we give the main conclusions on the work done.

2 Related Work

There are a lot of articles related to the study of the Big-Five scale. One of the earliest ones is the work written by S.D. Gosling et al. [3] examines the psychological scale of Big-Five and the correlation with measurements on it for different psychological tests. Their result noticed the fact that different surveys do not have a high correlation with respect to the values of the characteristics among themselves. Such a result can lead to understanding why person's psychological portrait is difficult to predict. There were also many works on the direct prediction of the psychological portrait, for example, in F. Mairesse et al. [5], the prediction of a psycho-portrait was based on the linguistic analysis of the Twitter posts. The prediction used a decision tree as a model while dividing into several classes of the original Big-Five scale. We also used similar division into several classes to improve quality of prediction model.

Later on, in [9], a model for predicting the characteristics of a Twitter account has been constructed. Their work was divided into 2 parts. In the first one they tried to use the data from the account (some statistical information) and in the second one they tried to predict the signs for certain linguistic variables received during the processing of texts posts. Such work was conducted by IBM and described in research report [10].

Similar research was conducted in [1], in which S. Bai et al. used information from the popular social network in China to predict the psychological portrait based on the Big-five scale. They used a less accurate version of this scale consisting of 2 components (0, 1) with decision trees prediction models only. Correlation tables for some of the signs from users' accounts on the Facebook network and the current scale of signs were built in [6].

More advanced machine learning were applied in [4], in which the authors tried to predict the characteristics of a person on the Big-Five scale using neural networks, but also for reduced binary scale version from [1]. It should also be noted that, in the work, the generated signs for the analysis of Twitter posts were used as signs for the neural network.

3 Model Description

In our work, we use the Big-Five scales and the Schwartz scale. Big-Five scale contains 5 dimensions:

- Extraversion vs. Introversion
- Agreeableness vs. Antagonism
- Conscientiousness vs. Lack of direction
- Neuroticism vs. Sensible stability
- Openness vs. Closeness to experience

Schwartz human values scale contains 10 dimensions:

- Self-Direction
- Stimulation
- Hedonism
- Achievement
- Power
- Security
- Conformity
- Tradition
- Benevolence
- Universalism

First of all, we consider the correlation between the initial signs and the resulting values of psychological traits. Correlation tables were constructed for the data, the most significant results are shown in Table 1.

	city	sibling	grandparent	parent	child	sex	age
Self-Direction	-0.05	**-0.19**	-0.06	**-0.15**	-0.09	**-0.22**	-0.03
Stimulation	**-0.14**	0.05	-0.02	-0.03	**-0.22**	**-0.16**	**-0.27**
Hedonizm	**-0.11**	0.05	0.05	0.01	**-0.28**	-0.00	**-0.27**
Achievement	-0.02	0.07	-0.08	0.02	**-0.16**	-0.08	**-0.26**
Power	-0.04	-0.04	-0.11	-0.16	-0.14	-0.19	-0.03
Security	-0.06	-0.08	0.02	-0.02	-0.09	0.04	0.05
Conformity	-0.04	0.02	0.12	-0.03	-0.14	0.08	-0.14
Tradition	0.09	-0.05	0.14	0.08	**0.21**	0.08	**0.23**
Benevolence	0.03	-0.02	0.07	**0.21**	**0.18**	**0.38**	0.08
Universalism	0.05	-0.11	0.02	-0.06	**0.18**	**0.30**	0.10
Extraversion	-0.04	**-0.17**	0.05	-0.09	0.00	**0.20**	0.00
Agreeableness	0.12	**-0.18**	-0.05	-0.12	-0.06	0.04	-0.08
Conscientiousness	0.12	-0.10	-0.12	-0.10	-0.09	-0.02	0.04
Neuroticism	0.06	**-0.17**	**-0.15**	**-0.15**	-0.11	-0.01	0.05
Openness	-0.09	**-0.24**	-0.05	0.08	0.05	0.15	0.02

Table 1. Correlation Table.

Further, the scale data were considered from the point of view of regression and classification for constructing the predictive model. We used algorithms of machine learning, such as Gradient boosting, Random forest, SVM, Linear / Logistic Regression from Scikit-learn [7]. The results were obtained by 5- and 10- fold cross-validation.

For regression algorithms, we compare the results presented for R^2 and MSE, and accuracy, f1-score for classification algorithms. For each of the problems, we consider some of the best features that could be predicted (three in each case). In the case of the classification problem, we split our initial scale into 5 classes and try to predict the hit of an element in 1 of the classes. The results obtained are presented in Tables 2 and 3.

It can be noted that even the best results for regression are extremely poorly predicted by the significance of the features, which is indicated by Negative value of R^2. In the case of the classification problem, the results are encouraging, but still the feature selection based on only social profile is not sufficient. It is also worth noting that the results obtained for the remaining characteristics do not provide sufficient grounds for stating the conclusions about successful classification.

As a result of machine learning, the best methods were SVR, Random Forest - in the case of a regression problem, and Gradient Boosting, Logistic regression - in the case of a classification problem. Also, attempts were made to improve the result using SVM (the use of different kernels, but this transformation did not yield tangible results).

Hedonizm	mean R^2	std R^2	mean MSE	std MSE
Linear Regression	-0.94	0.50	211.38	59.34
Random Forest	-1.04	0.41	218.51	45.44
Gradient Boosting	-1.10	0.33	235.83	74.73
SVR	-0.36	0.25	138.26	32.13
Power	mean R^2	std R^2	mean MSE	std MSE
Linear Regression	-0.30	0.23	132.43	46.63
Random Forest	-0.59	0.16	157.42	46.47
Gradient Boosting	-0.73	0.42	170.33	46.30
SVM	-0.59	0.18	158.78	46.11
Univeralizm	mean R^2	std R^2	mean MSE	std MSE
Linear Regression	-0.72	0.39	164.03	48.98
Random Forest	-0.51	0.22	146.75	44.97
Gradient Boosting	-0.60	0.43	148.79	39.61
SVM	-0.54	0.14	149.60	51.89

Table 2. Regression results - Hedonizm, Power, Univeralizm.

Self-Direction	mean f1-score	std f1-score	mean accuracy	std accuracy
Logistic Regression	0.23	0.11	0.30	0.09
Random Forest	0.24	0.05	0.27	0.06
Gradient Boosting	0.20	0.05	0.23	0.05
SVM	0.17	0.06	0.29	0.07
Power	mean f1-score	std f1-score	mean accuracy	std accuracy
Logistic Regression	0.24	0.03	0.66	0.01
Random Forest	0.23	0.05	0.63	0.02
Gradient Boosting	0.22	0.04	0.61	0.03
SVM	0.22	0.05	0.69	0.03
Conformity	mean f1-score	std f1-score	mean accuracy	std accuracy
Logistic Regression	0.35	0.06	0.50	0.07
Random Forest	0.37	0.13	0.47	0.10
Gradient Boosting	0.32	0.07	0.43	0.07
SVM	0.27	0.06	0.51	0.09

Table 3. Classification results - Self-Direction, Power, Conformity.

4 Dataset

As a training dataset we used a small database of users, which were manually marked by experts. Additional verification was obtained via psychological tests processed by users who allowed access to their social network profiles when passing the tests. Information from accounts in the social network we will get by using VKontakte API. The information we will get contains the following characteristics:

- 'status', 'about', 'wall comments', 'relation',
- 'has mobile', 'has photo', 'inspired by',
- 'people main', 'life main', 'political',
- 'tv', 'movies', 'games', 'music',
- 'smoking', 'alcohol', 'religion',
- 'home town', 'city', 'languages',
- 'universities','education',
- 'activities', 'interests',
- 'occupation', 'career',
- 'followers count',
- 'sex', 'age'.

We separate "complex" characteristics, such as

- 'people main', 'life main', 'inspired by',
- 'alcohol', 'smoking', 'political', 'religion',
- 'languages', 'city', 'followers count'

For the items of this group we considered the value that forms persons attitude to the given attribute (each characteristic is associated with a scale). Other attributes we associated with 0,1-values, such that 1 corresponds to the indicated attribute in the profile, while 0 means that characteristic is not specified. We also added signs of sex and age in addition to numerical data. After that we form data based on the characteristics.

The distribution of result features is shown in Fig.1

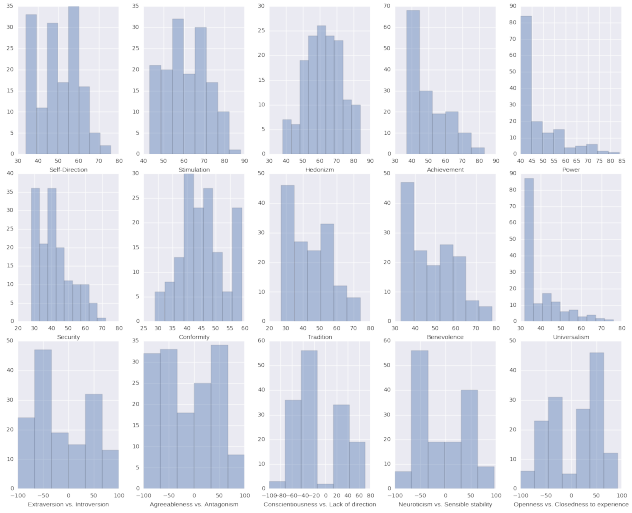

Fig. 1. Histograms of features

5 Discussion

First of all, a correlation table was constructed for the signs, which revealed significant values for some psychological signs and background data from social networks. However, for the regression problem we have received unsatisfactory results, regardless of the machine learning model. For the classification problem our results can be considered satisfactory. We have received predictions on certain grounds (in particular, Self-Direction, Power, Conformity). It is also worth noting that such results can be a consequence of a small sample, so we cannot unequivocally answer the question of the significance of the results obtained. As possible explanations for this result, there are three main reasons. Firstly, this can mean that in the future it is necessary to consider a more carefully selected set of characteristics. Secondly, it should be noted that such a poor result can be a consequence of a subjective scale of psychological signs, because tests cannot accurately determine the psychological portrait of a person. Thirdly, which is

also important, the presence of unverified information in social accounts makes it difficult to build a working model.

6 Conclusion

During the work, a set of characteristics for the prediction of an account from a social network was generated, 2 models of prediction of features (regression model, classification model) were constructed. Predictive models were SVM, SVR, Logistic Regression, Random Forest, Gradient Boosting. As a result, significant correlations were found in some of the psychological traits with signs from social networks, and the model of prediction with classification was successfully reviewed. Adding information from semantics of posted text information should benefit our research in the future work.

Acknowledgements. The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

References

1. Bai, S., Zhu, T., Cheng, L.: Big-five personality prediction based on user behaviors at social network sites. arXiv preprint arXiv:1204.4809 (2012)
2. Goldberg, L.R.: An alternative" description of personality": the big-five factor structure. Journal of personality and social psychology 59(6), 1216 (1990)
3. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in personality 37(6), 504–528 (2003)
4. Kalghatgi, M.P., Ramannavar, M., Sidnal, N.S.: A neural network approach to personality prediction based on the big-five model. International Journal of Innovative Research in Advanced Engineering (IJIRAE) 2(8), 56–63 (2015)
5. Mairesse, F., Walker, M., et al.: Words mark the nerds: Computational models of personality recognition through language. In: Proceedings of the Cognitive Science Society. vol. 28 (2006)
6. Marshall, T.C., Lefringhausen, K., Ferenczi, N.: The big five, self-esteem, and narcissism as predictors of the topics people write about in facebook status updates. Personality and Individual Differences 85, 35–40 (2015)
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830 (Nov 2011)
8. Schwartz, S., Bilsky, W.: (1987). toward a universal psychological structure of human values. Journal of Personality and Social Psychology 53(3), 550–562 (1987)
9. Sumner, C., Byers, A., Boochever, R., Park, G.J.: Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In: Machine learning and applications (icmla), 2012 11th international conference on. vol. 2, pp. 386–393. IEEE (2012)
10. Yang, H., Li, Y.: Identifying user needs from social media. IBM Research Division, San Jose p. 11 (2013)