Reproduce “Private traits and attributes are predictable from digital records of human behavior”

JING CHEN, University of Pittsburgh, Pittsburgh, PA, jic118@pitt.edu  
EASON, University of Pittsburgh, Pittsburgh, PA, \*\*\*@pitt.edu

KEYWORDS

data mining, private traits, facebook likes, SVD, machine learning

Abstract

The work we will do is reproducing research results in the paper “Private traits and attributes are predictable from digital records of human behavior” [1].

1 INTRODUCTION

The motivations we try to reproduce the result of this paper can be conclude as follow. First, following the paper instruction and reproducing the whole research process allows us to verify its result, whether the result is reproducible by the procedures given in the paper, how the difference of dataset can affect the result and wthether the methods they implemented in this paper is reasonable. Second, by reproducing the study, we can put data mining theory and technical skills learned in class into practice. Also, we will try to find whether we can make some adjustments that can improve the result and provide a detailed discussion on the adjustments. We believe this work can give us a deeper understanding of knowledge in this field. Last, this problem focus by the paper is meaningful and profound, nowadays, we are exposed to social media environment, there are a huge number of easily accessible digital records of behavior on such social media, which can be used to automatically and accurately predict a range of highly sensitive personal attributes. Data mining technique skills like SVD, linear regression model, logistic regression model, cross-validation, ROC, AUC, and Pearson product-moment correlation had been used in our work.

2 DATASETS

The dataset used in our work download from the myPersonality Project[[1]](#footnote-2) database. This dataset includes three csv files, users (110K), likes (1.58 million), and userslikes (10 million). Users has the information about 110K users, including userid, gender, age, political parties, openness, conscientiousness, extraversion, agreeableness, and emotional stability. Likes has two columns, likeid and name. Users-likes has two columns, userid and likeid.

3 FINISHED WORK

The first step is to preprocess users-likes datasets, according to the support information of this paper, that requires to remove Likes which associate with fewer than 20 users, as well as users with fewer than two Likes. After doing that, we got 105,262 users and 54,457 Likes. Then we create a sparse matrix (105262\*54457) which generated by a pakage called Matrix in R. This sparse matrix only has a small size of 87Mb (the dense matrix about 40Gb).

Next, dimensionality reduction technique SVD has been implemented to the matrix and got two matrixes (one has 100 top SVD components, one has 30 compoents).

For the next phase, we applied cross validation on the dataset, in which, logistic regression model has been used for the dichotomous variables, linear regression model for the numerical variables.

In order to evaluate prediction performance of each model, we express the prediction accuracy of dichotomous variables in terms of the area under the receiver-operating characteristic curve (AUC). The prediction accruracy of numeric expressed by the the Pearson product-moment correlation coefficient between the actual and predicted

values.

4 RESULTS AND DISCUSSION

The preliminary prediction results of numeric variables are showed in [Fig. 1](#fig2). As we can see, the highest correlation was age(r=0.42). The following are the traits “Openness” (r=0.34). However，the performance on each trait was inferior to the results given in the original paper. This chart shares the same tendency with original result chart, that means we are close to the result the paper given.

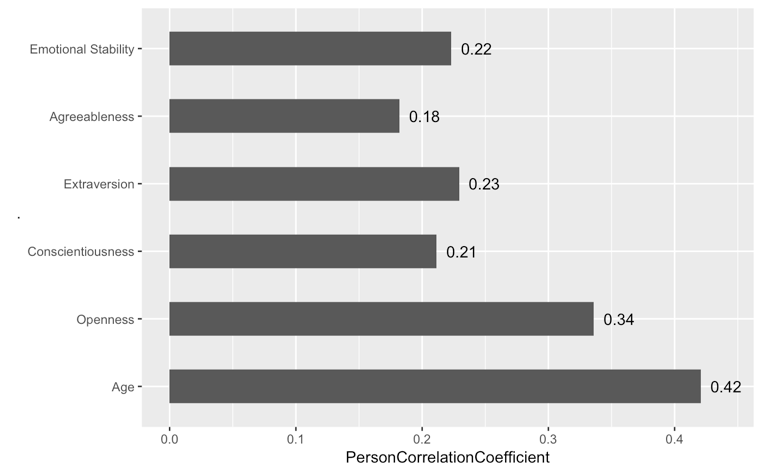


Fig. 1. Prediction accuracy of regression for numeric attributes and traits expressed by the Pearson correlation coefficient between predicted and actual attribute values.

The prediction results of dichotomous variables are presented in [Fig. 2](#fig3). Gender were classified in 86% cases, Democrat and Republican were classified in 78% cases.

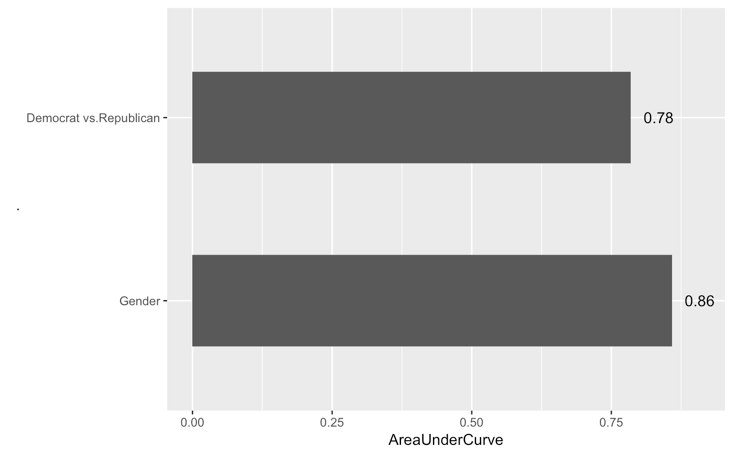


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

4 CONCLUSIONS

In

ACKNOWLEDGMENTS

This work was partially supported by the MIUR-PRIN 2010–11 Project 2010ECA8P3 “DyNanoMag” and by the National Research Foundation, Prime Minister's office, Singapore under its Competitive Research Programme (CRP Award No. NRF-CRP 10-2012-03).

REFERENCES

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

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. 2002. Wireless Sensor Networks: A Survey. *Comm. ACM* 38, 4 (2002), 393–422.

[3] David A. Anisi. 2003. *Optimal Motion Control of a Ground Vehicle*. Master’s thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.

[4] P. Bahl, R. Chancre, and J. Dungeon. 2004. SSCH: Slotted Seeded Channel Hopping for Capacity Improvement in IEEE 802.11 Ad-Hoc Wireless Networks. In *Proceeding of the 10th International Conference on Mobile Computing and Networking* (MobiCom’04). ACM, New York, NY, 112–117.

[5] Kenneth L. Clarkson. 1985. Algorithms for Closest-Point Problems (Computational Geometry). Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171.

[6] Jacques Cohen (Ed.). 1996. Special Issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996).

[7] Bruce P. Douglass. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. DOI: http://dx.doi.org/10.1007/3-540-65193-429

[8] Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. DOI: http://dx.doi.org/10.1007/3-540-09237-4

Journal Name:ACM TRANS. GRAPH.

Volume:9

Issue:4

Year:2010

Month:March

Article Seq:39

DOI:10.1145/1234

Copyright Year:2017

Copyright Statement:rightsretained

ISSN:0730-0301

Price:$15.00

1. Data deposition: The data reported in this paper have been deposited in the myPersonality Project database (www.mypersonality.org/wiki). [↑](#footnote-ref-2)