Personality Modeling: Potential and Pitfalls

David N. Chin
University of Hawai'i at Mānoa
Dept. of Information & Computer Sciences
1680 East West Rd, POST 317
Honolulu, HI 96822 USA
chin@hawaii.edu

ABSTRACT
Personality affects the responses of users to many things, so would be very useful for user adaptation. Since assessing personality requires long questionnaires that may not be practical or can be falsified, there is a need for techniques that infer personality from other user artifacts. This research field is too new to have established best practice research procedures. Pitfalls include overfitting data and how to correctly compare different classifiers or multiple regression models for statistical significance of accuracy differences.

Keywords
Personality; user modeling.

1. INTRODUCTION
Personality is a psychological theory meant to help explain individual differences among people in patterns of behavior, thinking, and feeling. The predominant model, the Big Five, has 5 personality traits: openness to new experiences, conscientiousness (self-discipline), extraversion (outgoing), agreeableness (cooperative), and neuroticism (easiness to anger, anxiety, depression, or vulnerability).

Personality affects the response of users to many things from user interfaces [2] to success in medical school [4] and pair programming [9]. Because the typical method for measuring personality by lengthy questionnaires often is not practical or can be cheated, there is a need and potential for other indirect means of recognizing user personality. However as in any new field, there are also many potential pitfalls to be avoided.

2. POTENTIAL
The typical Big Five personality questionnaire has 50-items (IPIP) with a slightly shorter 20-item version (Mini-IPIP) available and the Myers-Briggs Type Indicator (MBTI) standard questionnaire is a 93-item assessment with longer (144 and 222 items) versions available. In many applications, users either will not or cannot take the time to answer that many questions. In other applications, users can easily cheat on the personality self-assessments, invalidating the results. This is particularly problematic in high-stakes contexts such as job applications and applications to medical schools.

By using indirect indicators to infer personality, user modeling systems can bypass the drudgery of answering many questions in a personality assessment. For example, [10] infers personality in the massively multiplayer online game (MMOG) World of Warcraft from player actions such as the ratio of dungeon-based achievements versus all achievements and the ratio of need rolls versus greed rolls, features of guild/character names such as the number of negative or positive words in the name, and social network measures such as degree centrality and frequency of playing with different numbers of other characters.

Also if these indicators are collected from user artifacts produced before their applications for jobs or medical school, the likelihood of cheating can be greatly minimized. Even if the indicators are collected as part of the application process, the indirect indicators may prove more difficult for applicants to trick. For example many researchers have found correlations between text and personality [7], 11, [12]. If an essay is included in a medical school application, that text could be analyzed to predict personality and cheating on the essay to masquerade as a different personality would likely be much more difficult than picking different answers on a personality questionnaire. Also the predictions from the text analysis could be compared to results from a personality questionnaire to catch cheaters.

3. PITFALLS
Inferring personality is still a very new research area, so researchers have yet to establish best practice research procedures. The most common methodology is to use machine learning to train a classifier or derive a multiple regression equation for each personality trait. As with all machine learning tasks, care must be taken to avoid overfitting the data, which will typically happen when the number of possible training features gets close to the number of data points (users with personality profiles). Even the largest known dataset of personality profiles, the myPersonality Facebook dataset [3] with over 6M personality profiles is eclipsed by the number of possible words (from an estimated 20K for daily newspapers to over 1M in comprehensive dictionaries), bigrams (# words squared), and trigrams (# words cubed) in English.

As always, test datasets should be strongly segregated from training and tuning datasets so that no test set data is ever used for anything other than testing, including feature selection. To avoid overfitting, the number of features should be trimmed using a cutoff unrelated to their predictive value (e.g., information theoretic measures such as pointwise mutual information). For example, features could be trimmed based purely on their frequency in the training dataset. Although there is no commonly agreed upon ratio for the number of features relative to the sample size (number of users), [6] has found that regression equations stabilize only after reaching a ratio of 100 users per predictor feature.

Another pitfall is how to compare classifiers and multiple regression models. Researchers building personality classifiers seem to have settled on binary classifiers that divide the population evenly into high and low classes for each personality trait based on above and below the mean. Occasionally three-class models divide the population into high/medium/low at one standard deviation above and below the mean. Usually, because different datasets are used, classifiers and regression models cannot be directly compared. A higher classification accuracy or a smaller root mean squared error (RMSE) does not mean

anything if they are from two different datasets and even worse if those two datasets are from totally different domains.

Even when the same dataset is used, it is still problematic comparing two classifiers or regression equations. It may very well be that using a different dataset from the same general population would reverse the accuracy orderings. One would like to know if the difference in accuracies are statistically significant. The recommended practice is to use pairwise comparisons. For example, [8] recommend comparing only those datapoints that either classifier got right and the other got wrong using a Binomial test since a *t*-test is the wrong statistical test because a *t*-test assumes independence of the datasets for each treatment (each classifier), which obviously is false since the classifiers are being tested on the same test dataset. For multiple regression models, [1] also recommend pairwise comparisons of RMSE for each datapoint. Thus to test the statistical significance of differences in accuracy between classifiers or regression equations, researchers need not only access to the same datasets, but also either the prediction (classifiers) or RMSE (regression) for each datapoint in the test set.

4. REFERENCES

[1] A. Feelders and W. Verkooijen. 1996. On the statistical comparison of inductive learning methods. In D. Fisher and H.-J. Lenz (Eds.), *Learning from Data: Artificial and Intelligence V*, pages 271–279. Springer-Verlag

[2] Arvid Karsvall. 2002. Personality preferences in graphical interface design. In *Proceedings of the second Nordic conference on Human-computer interaction* (NordiCHI '02). ACM, New York, NY, USA, 217-218. DOI=http://dx.doi.org/10.1145/572020.572049

[3] Kosinski, M., Matz, S., Gosling, S., Popov, V. & Stillwell, D. (2015) Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. *American Psychologist* 70(6). 543-556. http://dx.doi.org/10.1037/a0039210

[4] Lievens, F., Ones, D.S., and Dilchert, S. 2009. Personality scale validities increase throughout medical school. *J. Appl. Psychol* Nov; 94(6): 1514-35. DOI=10.1037/a0016137

[5] Oded Nov, Ofer Arazy, Claudia López, and Peter Brusilovsky. 2013. Exploring personality-targeted UI design in online social participation systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13). ACM, New York, NY, USA, 361-370. DOI=http://dx.doi.org/10.1145/2470654.2470707

[6] Osborne, Jason W. (2000). Prediction in multiple regression. *Practical Assessment, Research & Evaluation*, 7(2). Retrieved May 24, 2016 from http://PAREonline.net/getvn.asp?v=7&n=2

[7] Pennebaker, J. and King, L. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296

[8] Steven L. Salzberg and Usama Fayyad. 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, pages 317–328

[9] Sfetsos, P., Stamelos, I., Angelis, L. and Deligiannis, I., 2006. Investigating the impact of personality types on communication and collaboration-viability in pair programming–an empirical study. In *Extreme programming and agile processes in software engineering* (pp. 43-52). Springer Berlin Heidelberg

[10] Shen, J., Brdiczka, O., Ducheneaut, N., Yee, N., and Begole, B. 2012. Inferring Personality of Online Gamers by Fusing Multiple-View Predictions. In *User Modeling, Adaptation, and Personalization: 20th International Conference*, UMAP 2012, Montreal, Canada, July 16-20, 2012, Proceedings, pages 261–273. Springer Berlin Heidelberg

[11] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9), e73791. http://doi.org/10.1371/journal.pone.0073791

[12] Wright, W.R., and Chin, D.N. 2014. Personality Profiling from Text: Introducing Part-of-Speech N-Grams. In V. Dimitrova. T. Kuflik, D. Chin, F. Ricci, P. Dolog, G.-J. Houben (Eds.), *User Modeling, Adaptation, and Personalization, 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014 Proceedings*, pp. 243-253