ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1742

Cross-Domain Error Correction in Personality Prediction

Işıl Doğa Yakut Kılıç [1] and **Shimei Pan** [2]

Abstract. In this paper, we analyze domain bias in automated text-based personality prediction, and proposes a novel method to correct domain bias. The proposed approach is very general since it requires neither retraining a personality prediction system using examples from a new domain, nor any knowledge of the original training data used to develop the system. We conduct several experiments to evaluate the effectiveness of the method, and the findings indicate a significant improvement of prediction accuracy.

1 Introduction

Recently, an array of automated text-based personality analysis tools and services has emerged as the amount of user generated data, such as social media posts, have increased significantly [3]. Such tools analyze textual data authored by an individual (e.g., one's social media posts), and generate a personality profile based on the results (i.e. IBM's personality insights [2]). These results can be used to infer consumer behavior patterns or brand preferences [4] which can be used by marketing and public relations teams in their decision making processes [1]. Varying to contradictory results due to domain difference (e.g., social media posts versus emails) however render the tool ineffective and untrustworthy. Typically, personality traits are measured by standard psychometric surveys (e.g., IPIP survey) which are questionnaire-based, independent of domain, situation and text. Since personality traits obtained from psychometric survey are frequently used as the ground truth to train and evaluate an automated personal trait prediction system, the discrepancy in the predicted results is mainly due to domain overfitting in machine learning (ML) rather than situation-dependent personality change [5]. Thus, if an automated systems could predict personality accurately, in principle they will output the same personality values as those obtained from psychometric surveys, regardless of the situation. In reality, most personality tools are developed and trained using text samples from one particular domain (e.g., IBM's Personality Insights was trained using tweets [5]). Since most machine learning algorithms work under the assumption that the test data will be drawn from the same population as the training data, when an application domain is very different from the training domain —as is often the case in real-world applications— accuracy suffers.

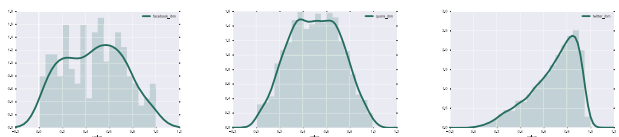
This short paper introduces a novel approach that identifies and reduces domain bias in text-based personality prediction systems. Since users of a personality prediction tool (e.g., a retailer) often do not have access to the training data (e.g., the dataset used to train IBM's Personality Insights), the proposed method employs a black-box approach that assumes access to neither the training data used to develop the tool nor training data from the new target domain (e.g., the application domain).

[1] University of Maryland, Baltimore County, USA, email: yakut1@umbc.edu
[2] University of Maryland, Baltimore County, USA, email: shimei@umbc.edu

2 Assessing Domain Bias in Personality Predictions

Domain difference in personality predictions can be evaluated at two levels: individual level and population level. At the individual level, for each person, we compute his trait scores based on his writing samples (e.g., social media posts). We compare the differences between the inferred trait scores and the ground truth personality of the same person. The larger the differences, the more severe the domain bias. We use Mean Square Error (MSE) as individual-level evaluation measure. Aside from observations at the individual level, the discrepancy at the population level can be shown between the distribution of predicted traits and the ground truth of a population. We use Kolmogorov-Smirnov test of equality between two distributions as the population-level evaluation metrics. Figure 1 shows the distributions of the Big Five Personality traits of the users from three datasets: Facebook, Quora, and Twitter. As can be seen, all three distributions are very different. The discrepancy of the predicted traits between Twitter and Quora (Figures 1b and 1c) is even more disturbing since the same set of individuals were used in collecting both datasets.

Figure 1: The Distributions of the Derived Conscientiousness Scores from three datasets: Facebook, Quora and Twitter.



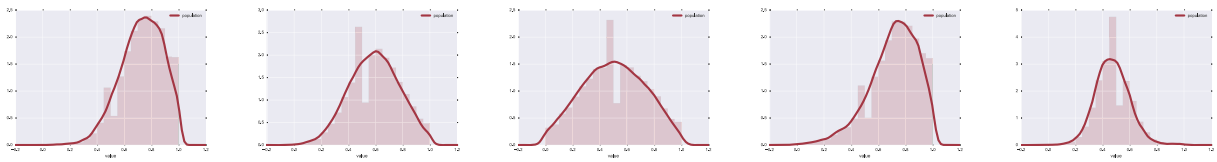
(a) Facebook (b) Quora (c) Twitter

3 Domain Bias Correction

Our method consists of two key operations: distribution parameter estimation and domain weight estimation based on domain similarity. Using these two processes, the method creates a linear transformation model based on the similarity between the application and the reference ground truth distributions.

The reference personality ground truth dataset is created to provide a domain-independent personality ground truth data from which we draw statistics to support domain bias correction. For this purpose, we use a personality ground truth dataset obtained from psychometric surveys (e.g., IPIP survey). The dataset is relatively big and contains 20,000+ people in total. Figure 2 shows the reference ground truth distributions of the Big Five personality traits. As one would expect, all of the personality traits have approximately normal distributions based on D'Augustino and Pearson's Normality test.

Figure 2: Reference ground truth distributions by aggregating the ground truth personality scores containing more than 20,000 data points. The data is aggregated from three separate sources: Personality Questionnaire results of individuals who are Facebook users, Mechanical Turkers and general public.



(a) Openness (b) Consciencious (c) Extraversion (d) Agreeableness (e) Emotional Range

Distribution Parameter Estimation In order to correct the predicted trait scores so that the inferred distributions fit the reference ground truth distributions, we first need to parametrize the distributions. Parameters of distributions can be used to scale and shift one distribution to fit another.Distribution parameters are calculated by first using Box-Cox Transformation and then applying Maximum Likelihood Estimation. Box-Cox transformation is a process that creates a normal distribution for given data using power functions. Given any distribution, the Box-Cox transformation will find an appropriate exponent λ and transform it into a normal distribution using the following formula:

$$y = \frac{(x^{\lambda} - 1)}{\lambda} \qquad (for\ \lambda > 0) \qquad (1)$$

$$log(x) \qquad (for\ \lambda = 0) \qquad (2)$$

where λ is the value that maximizes the log-likelihood function. Using the resulting stabilized data, we apply Maximum Likelihood Estimation to get the parameters.

The Maximum Likelihood Estimation (MLE) estimation of the distribution parameters μ (expectation) and σ^2 (variance) are calculated using mean *m* and standard deviation *s* of the data samples.

Domain Weight Estimation In cases where trait analysis results from multiple domains are available, we propose to weight the distribution parameters. The motivation for weighting application domains comes from the intuition that as the difference between predictions from an application domain and the reference ground truth increases, the prediction power of a system on that particular application domain decreases. To measure domain similarity between an application domain and the reference ground truth, we used the Kolmogorov-Smirnov statistic (KS statistic). The KS-statistic between two samples is given by

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)| \qquad (3)$$

where F_1 and F_2 are the *empirical distribution functions* of the two samples of size n and n', with μ and σ as their mean and standard deviation, and sup is the *supremum function* (or the Least Upper Bound).

Linear Transformation With distribution parameters and associated weight for each domain acquired in previous steps, we can create a linear transformation function to map a source value to a corrected value. The linear transformation function is unique to the application datasets at hand. The goal of this transformation is to make the distributions of the corrected values more similar to the reference ground truth distribution.

Specifically, using the estimated reference population parameters and domain weights, we can now design the following linear function to derive the corrected trait values V_{*t}

$$V_{*t} = \sum_{i=1}^{n} [((S^i_{*t} - \mu_{S^i})/\sigma_{S^i} * \sigma_G + \mu_G) * D_{S^i,S^G}] \qquad (4)$$

where S^i is one of the n application datasets. The μ and σ values are calculated mean and standard deviation of an application dataset, and μ_G and σ_G are the ground truth parameters. The D_{S^i,S^G} is the domain weight calculated using Equation 3 and normalizing them among the application datasets. The formula calculates the corrected values by transforming application datasets and calculating the weighted means.

4 Results

We conducted various evaluations to demonstrate the effective of our method. At the individual level, the relative MSE reduction ranging from 23% to 30% on different datasets. At the population level, our results indicate that the corrected trait distributions are much more similar to the ground truth distributions than those before bias reduction for all the traits on all the test domains.

5 Conclusion

This short paper presented an analysis of domain difference on personality prediction results, and proposed a method to correct the bias that require no knowledge about the training data. The algorithm uses parameter estimations, domain weighting, and linear transformations to correct the domain bias. The effectiveness of the method has been demonstrated based on both individual-level and population-level evaluation metrics. The proposed method is very general and can be used to correct other domain-bias problems.

REFERENCES

[1] Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen, 'Personalized persuasion: Tailoring persuasive appeals to recipients personality traits', *Psychological Science*, **23**(6), 578–581, (2012).
[2] IBM. Personality insights, IBM Watson developer cloud, 2015.
[3] Jianqiang Shen, Oliver Brdiczka, and Juan Liu, 'Understanding email writers: Personality prediction from email messages', in *User Modeling, Adaptation, and Personalization*, 318–330, Springer, (2013).
[4] Chao Yang, Shimei Pan, Jalal Mahmud, Huahai Yang, and Padmini Srinivasan, 'Using personal traits for brand preference prediction'.
[5] Michelle X Zhou, Fei Wang, Tom Zimmerman, Huahai Yang, Eben Haber, and Liang Gou, 'Computational discovery of personal traits from social multimedia', in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pp. 1–6. IEEE, (2013).