Looking Good With Flickr Faves: Gaussian Processes for Finding Difference Makers in Personality Impressions

Xiaoyu Xiong
School of Computing Science
University of Glasgow
Sir A.Williams Bldg.
G12 8QQ Glasgow, UK
x.xiong.1@
research.gla.ac.uk

Maurizio Filippone
Department of Data Science
EURECOM
Campus SophiaTech
06410 Biot, FRANCE
maurizio.filippone@
eurecom.fr

Alessandro Vinciarelli
School of Computing Science
University of Glasgow
Sir A.Williams Bldg.
G12 8QQ Glasgow, UK
vincia@dcs.gla.ac.uk

ABSTRACT

Flickr allows its users to generate galleries of "faves", i.e., pictures that they have tagged as favourite. According to recent studies, the faves are predictive of the personality traits that people attribute to Flickr users. This article investigates the phenomenon and shows that faves allow one to predict whether a Flickr user is perceived to be above median or not with respect to each of the Big-Five Traits (accuracy up to 79% depending on the trait). The classifier - based on Gaussian Processes with a new kernel designed for this work - allows one to identify the visual characteristics of faves that better account for the prediction outcome.

Keywords

Personality; Social Media; Computational Aesthetics

1. INTRODUCTION

Every trace left on social media - pictures, posts, videos, comments, etc. - reaches a large number of unacquainted observers: "*[...] the audience layer sits beyond the weak ties layer. It is made up of strangers [that] can play constructive roles when they are activated*" [7]. Employers gathering information about job candidates are a typical case. According to the Harvard Business Review, the outcome of the interviews depends, to a significant extent, on the impression that the employers develop by watching the online material posted by the candidates [2]. Such an example shows that it is important to investigate the interplay between, on the one hand, the observable traces people leave online and, on the other hand, the impressions that these traces convey. For this reason, this article investigates the relationship between "*faves*" - the pictures that Flickr users tag as favourite - and personality impressions.

The experiments of this work show that faves can be used to predict whether a Flickr user is perceived to be above median or not with respect to each of the Big-Five personality traits [10]. The accuracies range between 58% and 79% depending on the particular trait. The task has been performed with a classifier based on Gaussian Processes [9]. The main novelty of the approach is the *Group Automatic Relevance Determination* (G-ARD) kernel. Its accuracies are comparable, if not superior, to those achieved with Support Vector Machines, a widely applied state-of-the-art classifier. However, the most important advantage of the G-ARD kernel is that its parameter set includes weights - set automatically during the training process - capable of identifying the feature groups that better account for the classification outcome. In this respect, the G-ARD kernel is inspired by the *Automatic Relevance Determination* (ARD) one [6]. The main difference is that this latter has a weight for each individual feature and, therefore, the number of its parameters tends to be larger. For this reason, the G-ARD kernel appears to be more suitable when the amount of training material is limited like in the case of this work.

In the experiments, the analysis of the G-ARD weights shows that the classification outcome depends, to a significant extent, on the following characteristics of the faves: presence of human faces, composition, textural properties and, finally, number and size of visually homogeneous regions. Furthermore, weight differences across personality assessors of different national origins provide indications about cultural effects. The prediction of personality traits has been addressed extensively in the literature (see [12] for an extensive survey). However, this is the first work, to the best of our knowledge, that tries to go beyond simple classification to provide insights about the actual interplay between data and attributed traits. This applies, in particular, to previous results obtained over the publicly available data used in this work [3, 11].

The rest of this article is organised as follows: Section 2 describes the data and the personality model adopted in this work, Section 3 presents the features and the G-ARD approach, Section 4 reports on experiments and results and the final Section 5 draws some conclusions.

2. DATA AND PERSONALITY

The experiments of this work have been performed over *PsychoFlickr*, a publicly available corpus of 60,000 pictures tagged as favourite by 300 Flickr users (200 faves per user), the subjects hereafter [3, 11]. For every subject, the Corpus includes two personality assessments (see below): the

MM '16, October 15–19, 2016, Amsterdam, The Netherlands.

DOI: http://dx.doi.org/10.1145/2964284.2967253

first is the average of the traits attributed by 11 British assessors, the second is the average of the traits attributed by 11 Asian assessors. This makes it possible to investigate cultural effects. The personality assessments are presented in terms of the *Big Five Traits* (BF). These are five behavioural dimensions that are known to capture most individual differences and are recognised as the most effective personality model proposed so far [10]. The BFs are as follows: *Openness* (tendency to have wide interests and to be intellectually curious), *Conscientiousness* (tendency to be responsible, thorough and planful), *Extraversion* (tendency to be active and establish social relationships), *Agreeableness* (tendency to act according to the benefit of others) and *Neuroticism* (tendency to experience only the negative side of life).

From a computing point of view, the main advantage of the BF is that it represents personalities as five-dimensional vectors, a format particularly suitable for computer processing. Each component of the vector is a score that accounts for how well the behaviour of an individual fits the tendencies associated to a particular trait. The scores can be obtained with questionnaires designed for personality assessment. In the experiments of this work, the questionnaire is the *Big Five Inventory 10* (BFI-10) [8]. Once the BF scores are available for all persons in a corpus, it is possible to estimate the median for every trait. In this way, the subjects can be split into two classes, namely those who are above the median versus the others. Hereafter, the two classes are referred to as *high* and *low*, respectively.

3. THE APPROACH

The approach proposed in this work includes two main steps, namely feature extraction and classification into *high* or *low* (see Section 2). The rest of this section presents both steps in detail.

3.1 Feature Extraction

Every picture of the corpus is represented with a set of 82 features inspired by Computational Aesthetics, the domain aimed at predicting whether people consider an image visually appealing or not [5]. The main reason behind this choice is that these features capture the visual appearance of the faves, the only information that the assessors have at disposition to attribute personality traits to the 300 subjects of the Corpus. Furthermore, the features have been shown to be effective in tasks similar to the one addressed in this work [3, 11]. In view of the G-ARD approach (see Section 3.2), the features have been split into 9 groups corresponding to the main visual properties of a picture (see [11] for a full description of the features):

- *G1: Faces.* Number of human faces (1 feature);
- *G2: Colour Properties.* Statistics on the distribution of Hue, Saturation and Value (5 features), valence, arousal and dominance of the emotions elicited by the colours (3 features), variety of colours (1 feature);
- *G3: Colour Distribution.* Fraction of pixels that can be mapped into each of the 11 basic colour categories (red, yellow, pink, etc.) (11 features);
- *G4: Homogeneous Regions.* Amount of edge pixels, number and size of homogeneous regions, image size (4 features);
- *G5: Composition.* Depth of field and use of the rule of thirds (2 features);
- *G6: Texture Wavelets.* Wavelets coefficients (12 features);
- *G7: GIST filters.* Coefficients of GIST filters (24 features);
- *G8: Gray Level Co-occurrence Matrix.* Statistics of pixel values co-occurrences in 3×3 patches (12 features);
- *G9: Texture Statistics.* Tamura features and gray level distribution entropy (4 features).

The feature set is designed to account for content independent visual characteristics and, hence, cope with the wide semantic variability of the pictures posted online. The only content dependent feature is the number of human faces (Group G1) because these are ubiquitous in pictures. A subject of the PsychoFlickr Corpus is represented with the average of the 200 feature vectors extracted individually from every fave. In this way, the whole PsychoFlickr Corpus is represented by 300 vectors, one per subject (see Section 2).

3.2 Trait Classification

The classification approach proposed in this work is based on Gaussian Processes (GPs) [9]. These share with Support Vector Machines - the classifier that achieves state-of-the-art results in most tasks addressed in the literature - the important property of being nonparametric. However, GPs have at least two major advantages. The first is that an appropriate definition of their kernel allows one to explain the role played by the different feature groups in the classification (without explicit knowledge of the mapping between features and labels). The second is that GPs are formulated in probabilistic terms and, hence, a Bayesian treatment allows one to incorporate confidence levels when making predictions.

Consider the vector $\mathbf{y}$ where the i^{th} component y_i is the actual class of training vector $\mathbf{x}_i$ ($y_i = 1$ when the class is *high* and $y_i = -1$ when the class is *low*). The assumption underlying GP classifiers is that the y_i's are independently Bernoulli distributed conditioned on a set of N latent variables f_i evaluated in correspondence of the training vectors $\mathbf{x}_i$, where N is the size of the training set. As a result, the expression of the likelihood is as follows: $p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{N} p(y_i \mid f_i(\mathbf{x}_i))$, where $p(y_i \mid f_i(\mathbf{x}_i))$ is equal to the normal cumulative distribution $\Phi(y_i f_i(\mathbf{x}_i))$.

Under the GP assumption, the latent values $\mathbf{f}$ are jointly Gaussian distributed with $p(\mathbf{f} \mid \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K}$ is the kernel matrix and $\boldsymbol{\theta}$ is the parameter vector of the kernel function. The main novelty of this work is the *Group-Automatic Relevance Determination* (G-ARD) kernel function, a new kernel parametrisation designed to quantify the role played by the feature groups identified in Section 3.1 or any other meaningful partitions of the features:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp\left\{ -\sum_{r=1}^{N_g} \frac{1}{N_r {\tau_r}^2} \left[\sum_{s \in \mathcal{G}_r} \left(\mathbf{x}_{i(s)} - \mathbf{x}_{j(s)}\right)^2 \right] \right\}, \quad (1)$$

where σ is the marginal variance of the latent values, τ_r is the length-scale parameter for group r (it ensures, on the one hand, that the weights do not depend on the number of features in the groups and, on the other hand, that different

weights are comparable even if the respective groups include different numbers of features), N_r is the number of features in group r, N_g is the number of groups, $\mathbf{x}_{i(s)}$ is the s^{th} component of vector $\mathbf{x}_i$, and $\mathcal{G}_r$ is the set of the indexes of the features that belong to group r. Compared to the ARD kernel, the G-ARD allows one to reduce the number of weights from 82 (the number of features) to 9 (the number of groups).

3.3 Fully Bayesian Inference and Predictions

Let $\boldsymbol{\theta}$ denote the parameters of the G-ARD kernel. For a new input vector $\mathbf{x}_\star$ with latent function $f_\star$, a fully Bayesian treatment of the $y_\star$ prediction requires solving:

$$p(y_\star \mid \mathbf{y}) = \int p(y_\star \mid f_\star) p(f_\star \mid \mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y}) df_\star d\mathbf{f} d\boldsymbol{\theta} \quad (2)$$

where $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y})$ is the posterior over $(\mathbf{f}, \boldsymbol{\theta})$. The integral with respect to f_* is univariate and is not problematic, whereas integration over $\mathbf{f}$ and $\boldsymbol{\theta}$ is intractable. Common practice is to approximately integrate out $\mathbf{f}$ and replace the integral over $\boldsymbol{\theta}$ with a fixed point estimate $\boldsymbol{\theta}$, usually obtained by optimising an approximation to the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. While this yields a tractable formulation of the model, it can potentially underestimate uncertainty or cause inaccurate evaluation of the relative influence of different features [4]. We propose to account for the posterior $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y})$, which encodes the uncertainty in model parameters, enabling us to gain an understanding of the importance of different features with an associated level of confidence.

Sampling from $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y})$ requires the design of highly nontrivial Markov Chain Monte Carlo (MCMC) algorithms. In this work, predictions are carried out using an adaptive importance sampling-based approach, and in particular the Pseudo-Marginal Adaptive Multiple Importance Sampling (PM-AMIS) proposed in [14], which has been shown to be faster than state-of-the-art MCMC approaches at computing predictions for GP models. The intuition is that the algorithm adaptively constructs an approximate posterior over $\boldsymbol{\theta}$ that is used to build an increasingly more accurate importance sampling estimator of the predictive distribution (2). The importance weights have the following form:

$$w_i^t = p(\boldsymbol{\theta}_i^t) / \frac{1}{\sum_{t=0}^{T-1} N_t} \sum_{t=0}^{T-1} N_t q_t(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}_t}), \quad (3)$$

where T is the total number of iterations, $p(.)$ denotes the posterior of $\boldsymbol{\theta}$ up to a constant, $q_t(.)$ denotes the importance density at iteration t with sequentially updated parameters $\widehat{\boldsymbol{\gamma}_t}$, and $\boldsymbol{\theta}_i^t$ are samples drawn from $q_t(.)$ with $0 \leq t \leq T-1$, $1 \leq i \leq N_t$. At each iteration of PM-AMIS, all the importance weights get updated, including those computed at previous iterations. Because in GP classification the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ cannot be computed analytically, PM-AMIS resorts to an unbiased estimate of the marginal likelihood using a "nested" importance sampling estimation procedure. Even though the computation of the weights is now approximate, because of the fact that $p(\mathbf{y} \mid \boldsymbol{\theta})$ is estimated unbiasedly, it can be shown that PM-AMIS does not introduce any bias in predictions [14].

4. EXPERIMENTS AND RESULTS

The experiments address the task of predicting whether a subject belongs to class *high* or *low* for the Big-Five (see Section 2). The main reason behind this choice is that it corresponds to the natural tendency to compare others in terms of who is higher or lower along a given dimension: "*a compelling argument can be made for emphasising comparisons among individuals, which we do in everyday life (Who is more assertive? Who is more responsible?) and which is useful for such practical purposes as deciding whom to hire for a particular job*" [1].

Table 1: Prediction accuracy.

	Ope	Con	Ext	Agr	Neu
GPs (UK)	65%	58%	71%	73%	79%
SVM (UK)	59%	62%	71%	74%	77%
GPs (Asia)	68%	52%	74%	68%	69%
SVM (Asia)	68%	47%	68%	69%	70%

The experiments have been performed using a k-fold validation approach ($k = 10$) and the same subject never appears in training and test set. The classification has been performed with both the G-ARD approach (see section 3.2) and an SVM with radial basis functions kernel (one parameter). The SVM classifier optimises the kernel parameters by minimising the cross-validation error across a set of candidate values, which is generally not feasible for large parameter sets and / or small amounts of data like the one adopted in this work (300 subjects in total). In contrast, the proposed G-ARD GPs can integrate the uncertainty in the kernel parameters by means of a Bayesian approach when they make predictions. As a result, Table 1 shows that the G-ARD is competitive with the SVM even if the number of its kernel parameters is larger. The accuracy differences across the traits are in line with results typically observed in Personality Computing [12], where different traits can be predicted with different degrees of accuracy depending on the particular data. This parallels the psychological concept of *relevance* according to which the traits emerge with different evidence in different contexts (e.g., Extraversion is easier to observe at a party than at a funeral) [13].

Besides achieving high accuracies, the G-ARD provides information about the feature groups that better account for the classification outcomes. Figure 1 shows the G-ARD weights for both Asian and British assessors. Overall, the presence of human faces (group G1) plays the most important role for all traits and both cultures. The only exception is Extraversion, where the role of G1 is significant, but comparable to those of other groups. The probable reason is that in the case of this trait, strongly associated to social interactions, it is important not only that there are other faces, but also in what type of image they appear (e.g., the face is the main element in a portrait, but it is just a detail in the picture of a crowded public space). Overall, faces appear to be more important for British assessors than for Asian ones for all traits except Neuroticism (the difference between the G1 weights is always statistically significant). The other feature groups for which the weights are large are those that correspond to high level aspects of a picture, namely amount and size of visually homogeneous regions (G4), composition (G5) and textural properties (G9). The other groups have a non-negligible role, but appear to be less important. One possible reason of these results is that visual features accessible at first glance, like those included

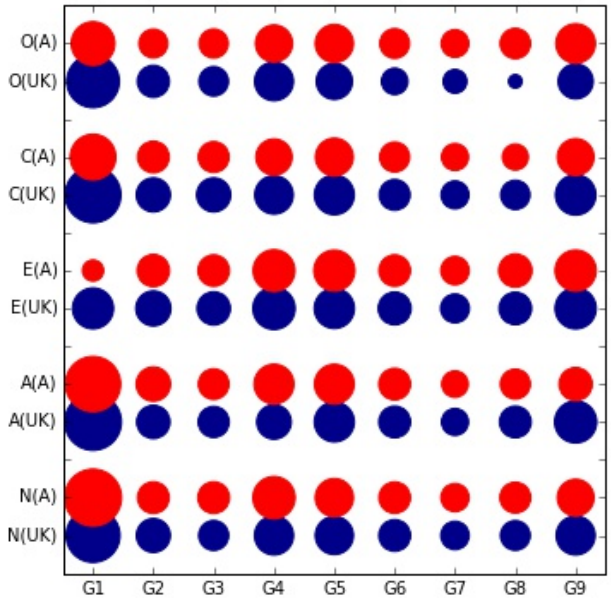

Figure 1: The plot shows the coefficients of the G-ARD for the five traits (O,C,E,A,N) and the two cultures, namely Asia (A) and UK.

in the groups above, are probably more likely than others to drive the personality impressions of the assessors.

The difference between the weights resulting from British and Asian assessors is always statistically significant except in the case of G5 for Neuroticism ($p < 0.01$ after Bonferroni Correction according to a weighted two-sample t-test). These results suggest that there is a cultural effect on personality perception. The largest differences can be observed for G1 (see above). Furthermore, British assessors are less sensitive to number and size of visually homogeneous regions for Agreeableness and Neuroticism while they are more sensitive to the texture properties for Conscientiousness and Agreeableness (conversely for Openness). Overall, Figure 1 suggests that UK and Asian assessors are sensitive to the same visual characteristics, but with different relative importance. One probable explanation is that there is no cultural difference for the physiological aspects - hence all assessors are sensitive to the same visual features - but there are cultural differences when it comes to the association between visual features and personality traits.

5. CONCLUSIONS

This article has shown that Flickr faves can be used to predict whether a Flickr user is perceived to be above median with respect to the Big-Five traits. The results show that the new G-ARD kernel designed for the experiments of this work allows a GP based classifier to achieve comparable accuracies as state-of-the-art SVMs. Furthermore, the parameters of of the G-ARD kernel allow one to identify the groups of features that better account for the classification outcome while detecting cultural differences between UK and Asian personality assessors.

The classification accuracies, well above chance for all traits, show that the weights of the G-ARD kernel provide reliable information about the interplay between low-level, visual characteristics of faves and attribution of personality traits. According to recent sociological investigations [7], this is important because the impression people convey online can change the outcome of important issues like, e.g., getting or not getting a job [2]. For this reason, future work will concentrate on how to use the information provided by the G-ARD weights to ensure that items posted online do not convey a wrong impression, whether it comes to faves or other types of online material.

6. REFERENCES

[1] S. Cloninger. Conceptual issues in personality theory. In P.J. Corr and G. Matthews, editors, *The Cambridge handbook of personality psychology*, pages 3–26. Cambridge University Press, 2009.

[2] D. Coutu. We googled you. *Harvard Business Review*, 85(6):1–8, 2007.

[3] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina. Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 213–222, 2013.

[4] M. Filippone and M. Girolami. Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226, 2014.

[5] F. Hoenig. Defining computational aesthetics. In *Proceedings of the Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, pages 13–18, 2005.

[6] David J.C. MacKay. Bayesian non-linear modelling for the prediction competition. In *In ASHRAE Transactions, V.100, Pt.2*, pages 1053–1062. ASHRAE, 1994.

[7] L. Rainie and B. Wellman. *Networked*. MIT Press, 2012.

[8] B. Rammstedt and O.P.P John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007.

[9] C.E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[10] G. Saucier and L.R. Goldberg. The language of personality: Lexical perspectives on the five-factor model. In J.S. Wiggins, editor, *The Five-Factor Model of Personality*, pages 21–50. Guilford Press, 1996.

[11] C. Segalin, A. Perina, M. Cristani, and A. Vinciarelli. The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. *IEEE Transactions on Affective Computing (to appear)*, 2016.

[12] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[13] A. Wright. Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing*, 5(3):292–296, 2014.

[14] X. Xiong, V. Šmídl, and M. Filippone. Adaptive Multiple Importance Sampling for Gaussian Processes. *eprint arXiv:1508.01050v2*, 2016.