TO BE PRESENTED AT MICCAI 2014

A Universal and Efficient Method to Compute Maps from Image-based Prediction Models

Mert R. Sabuncu*

A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA

Abstract. Discriminative supervised learning algorithms, such as Support Vector Machines, are becoming increasingly popular in biomedical image computing. One of their main uses is to construct image-based prediction models, e.g., for computer aided diagnosis or "mind reading." A major challenge in these applications is the biological interpretation of the machine learning models, which can be arbitrarily complex functions of the input features (e.g., as induced by kernel-based methods). Recent work has proposed several strategies for deriving maps that highlight regions relevant for accurate prediction. Yet most of these methods either rely on strong assumptions about the prediction model (e.g., linearity, sparsity) and/or data (e.g., Gaussianity), or fail to exploit the covariance structure in the data. In this work, we propose a computationally efficient and universal framework for quantifying associations captured by black box machine learning models. Furthermore, our theoretical perspective reveals that examining associations with predictions, in the absence of ground truth labels, can be very informative. We apply the proposed method to machine learning models trained to predict cognitive impairment from structural neuroimaging data. We demonstrate that our approach yields biologically meaningful maps of association.

Keywords: Machine learning, image-based prediction

1 Introduction

Broadly, there are two approaches in statistical data analysis [1]: generative (i.e., model based or classical) and discriminative (i.e., prediction oriented). While the former offers more interpretable models, the latter can yield more accurate predictions [1]. Over the last decade, discriminative supervised learning models have been widely adopted to analyze biomedical image data, for example to demonstrate that one can accurately predict a clinical diagnosis form imaging measurements, e.g. [2–6]. The main challenge in these studies is the biological interpretation of image-based prediction models.

One way to gain biological insight is to derive maps of association, which have traditionally been obtained via mass-univariate techniques, such as voxel-based

* Supported by NIH NIBIB 1K25EB013649-01 and a BrightFocus grant (AHAF-A2012333). Data used were obtained from ADNI: http://tinyurl.com/ADNI-main.

morphometry [7]. Motivated by this approach, recent studies have employed various strategies to compute such maps based on multivariate discriminative models, e.g. [8–11]. These techniques attempt to quantify the statistical relevance of voxel-level features with respect to the predicted variable. Several methods to compute feature relevance, or variable importance, have also been proposed in the machine learning literature, e.g. [12–16]. Yet, as we elaborate in the next section, most of these methods suffer from drawbacks, such as being specific to a type of algorithm/model (non-universality).

In this paper, we present a universal and computationally efficient method to examine associations captured by black box machine learning models. Our method does not rely on knowledge about the learning algorithm. Furthermore, we do not make any strong distributional assumptions about the data. In its simplest form, the proposed method simply uses a dataset, on which predictions have been computed. Our theoretical framework demonstrates that, even in the absence of ground truth labels, the associations we quantify can be informative about the underlying biology. We apply the proposed method to compute maps of association from discriminative models trained to predict a clinical or behavioral condition from structural brain magnetic resonance imaging (MRI) scans.

2 Theory

2.1 Motivation

A popular approach for interpreting a linear discriminative model is to examine the weights, e.g. [17]. Yet, as recently pointed out [11], this interpretation can be misleading. Furthermore, it is not clear whether directly examining the estimated model parameters provides any insight about the underlying biology. This is because the model can be arbitrarily inaccurate and thus model parameters alone might provide little information about the target variable. Sampling strategies, e.g. [9, 13, 15] address this issue by randomly perturbing the data and examining the variation in model parameters and/or predictions. This approach, however, typically requires repeatedly running the computationally expensive training step or resorting to approximation strategies. Moreover, it assumes a particular model structure, e.g. linear, sparse, or a tree.

More general methods applicable to any black box prediction model have recently been proposed, e.g. [14, 18]. Yet these techniques often make strong assumptions about the data (e.g. binary or Gaussian) to offer practical solutions. Our goal in this work is to propose a technique for measuring feature relevance that is universal (i.e. applicable to any black box model), computationally efficient, and robust with respect to the data and the algorithm. Furthermore, we would like the proposed method to capture nonlinear relationships, as well. To achieve this, we build on the theoretical framework that was recently used to derive a generalized measure of correlation [19].

2.2 Proposed Feature Relevance Measure and Its Properties

Let's assume a black box predictive model. We will use capital letters to denote random variables. Let $\boldsymbol{X}$ be the input data vector, which is typically high dimensional, e.g., images, P be the prediction produced by the model, and Y be the target variable that we aim to predict, e.g., clinical status. Note P is typically a non-random function of $\boldsymbol{X}$ and we denote the i'th component of $\boldsymbol{X}$ as $\boldsymbol{X}_i$.

A generalized measure of correlation (GMC) between two random variables, say P and Y, can be derived based on the well-known variance decomposition formula [19]:

$$V(Y) = V(E(Y|P)) + E(V(Y|P)), \tag{1}$$

where V denotes (conditional) variance and E denotes (conditional) expectation, defined over appropriate random variables. The first term on the right, $V(E(Y|P))$, can be interpreted as the explained variance of Y by P. Thus the GMC between Y and P, which we denote as $\gamma(Y|P)$, can be defined as:

$$\gamma(Y|P) = \frac{V(E(Y|P))}{V(Y)}. \tag{2}$$

The GMC is a measure of correlation that quantifies both linear and non-linear dependencies [19] and ranges from 0 (no correlation) to 1 (max. correlation).

We expand Eq. 1 by applying another variance decomposition to $V(E(Y|P))$:

$$V(Y) = V(E(E(Y|P)|\boldsymbol{X}_i)) + E(V(E(Y|P)|\boldsymbol{X}_i) + E(V(Y|P)), \tag{3}$$

where $\boldsymbol{X}_i$ is an input variable and $V(E(E(Y|P)|\boldsymbol{X}_i))$ can be viewed as the explained variance of Y by $\boldsymbol{X}_i$, *as captured* by the model's prediction P. Thus, we define the *captured correlation* as:

$$\kappa(Y|P|\boldsymbol{X}_i) = \frac{V(E(E(Y|P)|\boldsymbol{X}_i))}{V(Y)}. \tag{4}$$

Some of the properties of $\kappa(Y|P|\boldsymbol{X}_i)$ are as follows (Proofs of P1-4 are omitted due to space constraints). Note ρ denotes Pearson's correlation.

P1. $0 \le \kappa(Y|P|\boldsymbol{X}_i) \le \gamma(Y|P) \le 1$.
P2. If P and $\boldsymbol{X}_i$ *or* P and Y are independent, then $\kappa(Y|P|\boldsymbol{X}_i) = 0$.
P3. If $\exists f$ s.t. $f(P) = Y$, then $\kappa(Y|P|\boldsymbol{X}_i) = \gamma(Y|\boldsymbol{X}_i)$.
P4. If $\exists g$ s.t. $g(\boldsymbol{X}_i) = P$, then $\kappa(Y|P|\boldsymbol{X}_i) = \gamma(Y|P)$.
P5. If $\rho(\boldsymbol{X}_i, \boldsymbol{X}_j) = \pm 1$, then $\kappa(Y|P|\boldsymbol{X}_i) = \kappa(Y|P|\boldsymbol{X}_j)$.
Proof: If $\exists a \neq 0, b$ s.t. $\boldsymbol{X}_i = a\boldsymbol{X}_j + b$, then, for any Z, $E(Z|\boldsymbol{X}_i) = E(Z|\boldsymbol{X}_j)$. Thus, $V(E(E(Y|P)|\boldsymbol{X}_i)) = V(E(E(Y|P)|\boldsymbol{X}_j))$, where we use $Z \triangleq E(Y|P)$.
P6. If $\rho(E(Y|P), P) = \pm 1$, then $\kappa(Y|P|\boldsymbol{X}_i) = \gamma(Y|P)\gamma(P|\boldsymbol{X}_i)$.
Proof: Define $Z \triangleq E(Y|P)$. If $\exists a \neq 0, b$ s.t. $Z = aP + b$, then $a^2V(P) = V(Z)$ and $\frac{V(E(Z|\boldsymbol{X}_i))}{a^2} = V(E(P|\boldsymbol{X}_i))$. Then $\kappa(Y|P|\boldsymbol{X}_i) = \frac{V(E(Z|\boldsymbol{X}_i))}{V(Y)}\frac{V(P)a^2}{V(P)a^2} = \frac{V(E(Z|\boldsymbol{X}_i))}{a^2V(Y)}\frac{a^2V(P)}{V(P)} = \frac{V(E(P|\boldsymbol{X}_i))}{V(Y)}\frac{V(Z)}{V(P)} = \gamma(Y|P)\gamma(P|\boldsymbol{X}_i)$.

The first five properties summarize the general behavior of κ as a dependency measure. For example, it is zero if the model's prediction is independent of the variable $\boldsymbol{X}_i$. If the model is perfectly accurate, κ reduces to the GMC between Y and $\boldsymbol{X}_i$. Moreover, as **P5** suggests, the captured correlation is indifferent to whether a variable is directly used in the prediction or correlated alternatives are. Thanks to this property, captured correlation will not highlight an arbitrary subset among correlated variables, the way sparse models do.

P6 is a particularly interesting property, which states that under a specific condition, the captured correlation is proportional to the GMC between the prediction P and input variable $\boldsymbol{X}_i$. We note that, in fact $E(Y|P) = P$ is a common modeling assumption that seems to hold in many practical problems. For example, many regression models (where Y is continuous), assume a zero-mean independent additive Gaussian noise model. Or, in binary classification, P can be the probability of class 1. In both examples, these models imply $E(Y|P) = P$ and thus $\kappa(Y|P|\boldsymbol{X}_i) \propto \gamma(P|\boldsymbol{X}_i)$. In this case, the ranking of variables with respect to their captured correlations is the same as their ranking with respect to their GMC with the prediction. This is a critical observation. It suggests that, in the absence of ground truth data, examining the associations between input variables and the model's predictions can be informative about the relationships with the (ground truth) target variable.

2.3 A Non-parametric Estimator

We propose to employ a non-parametric strategy, which relies on the mild distributional assumption of finite first and second order moments, to estimate the correlation measures κ and γ. Here, we assume that we have access to N independent samples of (X, P, Y), where for notational simplicity we have replaced $\boldsymbol{X}_i$ with X. We denote these samples as $\{x_j, p_j, y_j\}$, where lower case letters represent observations, indexed by subscripts. We use the well-known Nadaraya-Watson estimator:

$$E(Y|P) \approx \frac{\sum_{j=1}^{N} k_P(p_j - P) y_j}{\sum_{l=1}^{N} k_P(p_l - P)} = \sum_{j=1}^{N} \bar{k}_P(p_j - P) y_j, \tag{5}$$

where k_P is an appropriate kernel function and $\bar{k}_P(p_j - P) = \frac{k_P(p_j - P)}{\sum_{l=1}^{N} k_P(p_l - P)}$. Similarly, we can write:

$$E(E(Y|P)|X) \approx \sum_{k=1}^{N} \bar{k}_X(x_k - X) \sum_{j=1}^{N} \bar{k}_P(p_j - p_k) y_j, \tag{6}$$

where $\bar{k}_X(x_k - X) = \frac{k_X(x_k - X)}{\sum_{l=1}^{N} k_X(x_l - X)}$ and k_X is an appropriate kernel. Now, let's concatenate the observations into length N column vectors $\{\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{y}\}$ and define two $N \times N$ matrices K_X and K_P, the (j, k)'th entries of which are $\bar{k}_X(x_j - x_k)$

and $\bar{k}_P(p_j - p_k)$, respectively. Given the above, an estimate of κ is:

$$\kappa(Y|P|X) = \frac{\hat{V}(K_X K_P \boldsymbol{y})}{\hat{V}(\boldsymbol{y})}, \tag{7}$$

where $\hat{V}$ denotes the sample variance, defined as $\hat{V}(\boldsymbol{y}) = \frac{1}{N-1}\sum_{j=1}^{N}(y_j - \bar{y})^2$ with $\bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j$. Similarly, an estimate of γ is: $\gamma(P|X) = \frac{\hat{V}(K_X \boldsymbol{p})}{\hat{V}(\boldsymbol{p})}$.

In our implementation, we employed Gaussian kernel functions for k_X (and similarly k_P). I.e., $k_X(x_j - x_k) = \exp(-\frac{(x_j - x_k)^2}{h_X})$. Based on Silverman's rule of thumb we set the bandwidth as: $h_X = \hat{V}(\boldsymbol{x})/N^{0.2}$. Note that, this choice also ensures that the estimates are invariant to rescaling a variable.

3 Experimental Results

Data: We analyzed data from two public datasets, OASIS (oasis-brains.org) and ADNI (adni.loni.usc.edu), which contain brain MRI scans from healthy and demented subjects. We processed the T1-weighted structural brain MRIs using FreeSurfer (FS v5.1, surfer.nmr.mgh.harvard.edu) to obtain thickness measurements across the entire cortex, resampled onto a common template, *fsaverage*. FS also provides estimates of volumes for a range of cortical and sub-cortical structures, such as the hippocampus. The target variable we used was mini mental state exam (MMSE) score, which measures cognitive impairment and is associated with dementia, including Alzheimer's disease (AD). The OASIS sample consisted of young healthy subjects (YCN, N=200, 26.8 ± 9.7 years, 55% Female), old cognitively normal (OCN, clinical dementia rating, CDR, zero) subjects (N=135, 69.1±13.8 y, 72% F) and AD patients (CDR> 0, N=100, 76.8±7.1 y, 59% F). We subdivided the OASIS OCN+AD sample (N=235) into five partitions (of equal size) for cross-validation. We call this the OASIS cross-validation sample. The entire ADNI sample contained N=810 (75.2 ± 6.9 y, 42% F) CN subjects, subjects with mild cognitive impairment and AD patients.

Machine Learning Algorithms: We explored two classes of publicly available algorithms to predict MMSE from brain MRI measurements. The first one is the Relevance Voxel Machine[1] [8] (RVoxM), which is an adaptation of a sparse Bayesian model, customized to handle image data. The second algorithm was the Support Vector Machine (SVM) with a radial basis function kernel[2]. We trained RVoxM and SVM to predict MMSE, based on FS-computed cortical thickness data. We also trained a separate SVM only on volumes of brain structures (saved as FS file *aseg.stats*), which we call SVM-aseg. We performed 5-fold cross-validation on the OASIS sample, where each of the five partitions was treated as the test sample in each fold, with the remaining subjects used for training. Thus, each OASIS subject was treated as a test case once, during

[1] downloaded from people.csail.mit.edu/msabuncu/sw/RVoxM/index.html
[2] downloaded from csie.ntu.edu.tw/ cjlin/libsvm

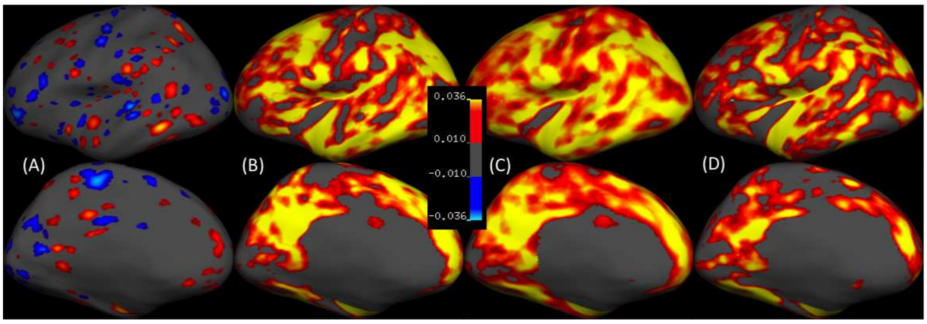

Fig. 1. All visualizations (in color) are on inflated *fsaverage* surface, a population average representation of the human cerebral cortex. Top and bottom rows show lateral and medial views, respectively. Only left hemispheres are shown. (A) Weights of RVoxM model trained on entire OASIS OCN+AD sample to predict MMSE from cortical thickness data. Note that most regions have no contribution to the model, i.e., have zero weight (shown in gray). (B) Captured correlation (κ) computed based on RVoxM's MMSE predictions on OASIS cross-validation sample. (C) κ-map for SVM's MMSE predictions on OASIS cross-validation sample (trained on cortical thickness). (D) κ-map for SVM-aseg's MMSE predictions on OASIS cross-validation sample

which the ("out-of-bag") image-based prediction was computed. The Pearson correlation between out-of-bag predictions and ground truth values were 0.46, 0.53, and 0.35 (all P$< 10^{-10}$) for RVoxM, SVM and SVM-aseg, respectively. The ADNI data were only used for training to obtain prediction models.

Results: Fig. 1A visualizes the weights of the RVoxM model trained to predict MMSE from cortical thickness data on the OASIS sample. Because of RVoxM's sparsity assumption, most cortical regions have zero contribution to the model. We argue that this fact, along with the issues associated with interpreting the parameters of a discriminative model [11] makes it hard to make biological sense of this map and the SVM models. Moreover, we could not visualize the (non-linear) kernel SVM models, since there is no well-accepted strategy to do so. Fig. 1B-D illustrate maps of captured correlation (κ) computed with three different models trained on the OASIS cross-validation samples (based on out-of-bag predictions). All these maps bear a striking resemblance to AD-associated thinning maps reported in prior work [20]. Note MMSE is a variable strongly correlated with and used to clinically diagnose AD. The right-most map was actually computed based on an SVM model trained on the *aseg* features, which do not include regional cortical thickness measurements (although there is a variable that measures global cortical volume). There is strong agreement between these three κ-maps (pairwise Pearson correlations > 0.84, see Fig. 2D), which suggests that the captured correlation measure is robust to the variation in prediction algorithm and utilized image features.

Fig. 2A-C illustrate maps of GMC, ($\gamma(P|X)$, which ignores the ground truth variables Y) between cortical thickness values and the RVoxM predictions of

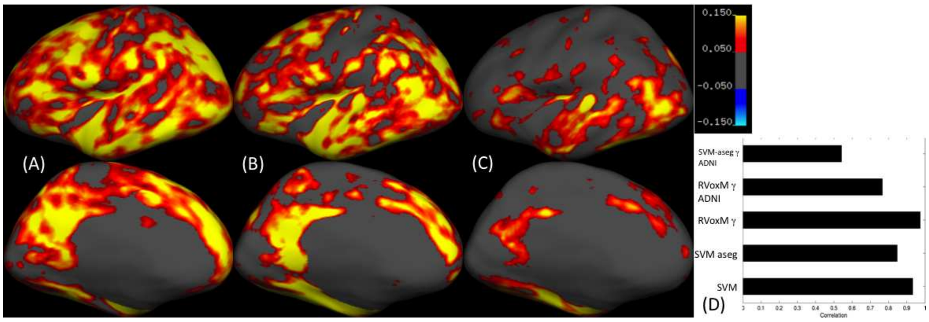

Fig. 2. (A) GMC (γ) between RVoxM's MMSE predictions and cortical thickness values computed on OASIS cross-validation sample. (B) γ-map between the ADNI RVoxM model's MMSE predictions and cortical thickness values computed on OASIS AD+CN sample. (C) γ-map between the ADNI RVoxM model's MMSE predictions and cortical thickness values computed on OASIS CN sample. (D) Pearson correlations of different maps with the RVoxM κ-map computed on OASIS cross-validation (shown in Fig. 1B). SVM, SVM-aseg, RVoxM γ, RVoxM γ ADNI, and SVM-aseg γ ADNI refer to maps of Fig. 1C, 1D, 2A, 2B, and 2C, respectively. For further details see caption of Fig.1.

MMSE. The correlation between the RVoxM-derived κ and γ-maps (Fig. 1-B and Fig. 2-A) is 0.97, providing evidence that the associations with the predicted values are informative about associations with the ground truth (thanks to property **P6** of captured correlation). Fig. 2B-C were in fact computed using models trained on a separate dataset (ADNI). The map of Fig. 2C is particularly intriguing, as it was computed on healthy subjects (the OASIS young and old cognitively normal sample). Since this sample does not include subjects with dementia, there is little variation in the MMSE values (29.1 ± 1.1). However, the γ-map with the predicted MMSE scores demonstrate that, even in this healthy cohort, regions of potentially significant association with cognitive impairment can be detected. There is a correlation of 0.54 (P$< 1e - 10$) between the map of Fig. 2C and the benchmark map of Fig. 1B. This result demonstrates the robustness of the proposed measure with respect to substantial variation in the data, since both the training and testing data are different between the analyses.

4 Conclusion

We proposed a novel measure, called captured correlation, to quantify associations between input features and the target variable, as captured by the prediction model. We applied this measure to image-based prediction models and demonstrated that captured correlation yields biologically meaningful maps that are robust to the choice of learning algorithm. We showed that under certain assumptions, captured correlation is proportional to the association between features and predictions. Intriguingly, this perspective provides a theoretical justification for examining associations with predictions, in the absence of ground truth labels. For example, one can analyze large, unlabeled datasets in order to identify potentially relevant areas, which could then be further interrogated on labeled datasets. Our approach can also be used to examine and prioritize multivariate relationships, such as the association between multiple image features and the target variable. Future work will pursue these interesting directions.

References

1. L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.
2. Y. Kawasaki et al. Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *Neuroimage*, 34(1), 2007.
3. C. Davatzikos et al. Individual patient diagnosis of ad and ftd via high-dimensional pattern classification of mri. *Neuroimage*, 41(4):1220–1227, 2008.
4. S. Klöppel et al. Automatic detection of preclinical neurodegeneration presymptomatic huntington disease. *Neurology*, 72(5):426–431, 2009.
5. D. Zhang et al. Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.
6. C Plant *e*t al. Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease. *Neuroimage*, 50(1):162–174, 2010.
7. J. Ashburner and K.J. Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11(6):805–821, 2000.
8. M.R. Sabuncu and K. Van Leemput. The Relevance Voxel Machine (RVoxM): A self-tuning bayesian model for informative image-based prediction. *IEEE Transactions on Medical Imaging*, 2012.
9. B. Gaonkar and C. Davatzikos. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. *NeuroImage*, 78:270–283, 2013.
10. E. Konukoglu et al. On feature relevance in image-based prediction models: An empirical study. In *LNCS v. 8184: Mach.Learn. in Med. Im.*, pages 171–178. 2013.
11. S. Haufe et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014.
12. P. Golland. Discriminative direction for kernel classifiers. *Advances in Neural Information Processing Systems*, 1:745–752, 2002.
13. C Strobl *e*t al. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.
14. A. Zien et al. The feature importance ranking measure. In *Machine Learning and Knowledge Discovery in Databases*, pages 694–709. Springer, 2009.
15. Meinshausen and Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
16. A Goldstein *e*t al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J of Comp and Graph Stat*, 2014.
17. N.U.F. Dosenbach et al. Prediction of individual brain maturity using fMRI. *Science*, 329(5997):1358, 2010.
18. S Sonnenburg et al. Poims: positional oligomer importance matricesunderstanding support vector machine-based signal detectors. *Bioinformatics*, 24(13):i6–i14, 2008.
19. S Zheng et al. Generalized measures of correlation for asymmetry, nonlinearity, and beyond. *J of the American Statistical Association*, 107(499):1239–1252, 2012.
20. B. Dickerson et al. The cortical signature of alzheimer's disease. *Cerebral cortex*, 19(3):497–510, 2009.