Voxel importance in classifier ensembles based on sign consistency patterns: Application to sMRI

Vanessa Gómez-Verdejo
Emilio Parrado-Hernández
Department of Signal Processing and Communications
Universidad Carlos III de Madrid
Spain.

Jussi Tohka
Department of Bioengineering and Aerospace Engineering
Universidad Carlos III de Madrid
Spain

for the Alzheimer's Disease Neuroimaging Initiative[0]

***Abstract*—This paper investigates a new measure of voxel importance based on analysing the sign consistency of voxels in an ensemble of linear SVM classifiers. The ensemble is endowed with a significant degree of diversity since the training set for each individual classifier is a random subsample of the initial training set. The importance of a voxel is proportional to the number of times that the voxel's weight has the same sign in all the classifiers of the ensemble. The multivariate nature of the method yields a robust importance pattern formed by clusters of voxels. The method is demonstrated with a MCI vs. control subjects classification task using the ADNI data.**

I. Introduction

The analysis of structural magnetic resonance images (sMRI) with machine learning is attracting the interest of the neuroscientific community as a powerful tool to characterize diseases that are reflected as alterations in the brain structure. Given a training set of brain images and the associated class information, here a diagnosis of the subject, supervised machine learning algorithms can learn the voxel-wise model that generated the class information based on the brain images. This has direct applications to the design of imaging biomarkers, but the inferred models can additionally be considered as multivariate, discriminative representations of the disease of interest. This representation is fundamentally different from the ordinary (thresholded) t-statistic map returned by massively univariate general linear model [1]. An important problem in using voxel-based supervised classification algorithms for brain imaging applications is that the dimensionality of data (the number of voxels in the images of a single subject) far exceeds the number of training subjects available. This has led to a number of papers studying feature selection within brain imaging (see [2] for a review). However, in addition to selecting a set of important features, it can be interesting to rank and study their importance to the classification. This problem, termed variable/feature/voxel importance determination, has received significantly less attention and it is the topic of this paper.

The simplest approach to voxel importance is to study its correlation to the class label, for example, via a t-test. This is exactly what the massively univariate analysis does. In the machine learning setting, if the features have been properly standardized, the weights of a linear classifier can be considered as measures of voxel importance (see, e.g, [3], [4]). The first strategy considers voxels independently of others and, therefore, may miss complex interactions, indeed to be meaningful for the classification, a voxel does not have to be correlated with the class label [1]. The downsides of the second approach are less obvious and some of them are outlined in [5]. In the machine learning community, the most widely used variable importance measures are based on Random Forest (RFs) classifiers [6]. RFs offer two ways for assessing variable importance: one based on Gini importance and one based on the analysis of out-of-bag samples (permutation importance). Both measures have found applications in brain imaging: [7] studied voxel selection based on Gini importance, [8] ranked the different types of variables (imaging, psychological test scores) for MCI-to-AD conversion prediction based on the out-of-bag variable importance and [9] ranked the importance of cortical ROI volumes to schizophrenia classification.

In this paper, we introduce and study a new variable importance measure based on sign consistency of the weights in an ensemble of linear support vector machines (SVMs). This procedure was applied to voxel selection in [10]. Here, instead, we compare the voxel importances returned by the method to the voxel importances resulting from the application of the univariate t-test and an elastic net penalized logistic regression. We also measure the stability of the voxel importances by using split-half sampling akin to [11]. More specifically, we demonstrate the approach using Mild Cognitive Impairment (MCI), which is a transitional stage between age-related cognitive decline and Alzheimers disease (AD), versus normal control (NC) classification with structural MRI data from ADNI. This problem was selected because a large data set was available and the MCI vs. NC classification is markedly more challenging than AD vs. NC classification [11].

II. Materials and methods

A. ADNI data

Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003

[0]Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). For up-to-date information, see www.adni-info.org.

We use MRIs from 404 MCI subjects and 231 normal controls (NC) for whom baseline MRI data (T1-weighted MP-RAGE sequence at 1.5 Tesla, typically 256 x 256 x 170 voxels with the voxel size of 1 mm x 1 mm x 1.2 mm) were available. The MRIs were preprocessed into gray matter tissue images in the stereotactic space as described in [12], [8], smoothed with 8-mm FWHM Gaussian kernel, resampled to 4 mm spatial resolution and masked into 29852 voxels.

B. Voxel importance via sign consistency bagging

1) Voxel importance with ensembles of linear SVMs: This paper builds on the voxel selection method of [10], that we call here sign consistency bagging (SCB). This method assumes that the voxel values are always positive; if this requirement is not satisfied naturally, it can be always ensured by adding a suitable constant to voxel values. First, one trains a few thousand linear SVMs, each with a different subset of training data selected at random without replacement (here we select subsets with the number of original samples halved, and the subsampling rate is not critical [10]). The parameter C for all the SVMs trained in this work is selected large enough to ensure complete separation since in sMRI the number of samples is orders of magnitude smaller than the number of voxels; notice that this assumption still holds even after performing SCB voxel selection. The SVM s in the ensemble is described by the weights $[w_1^s, \ldots, w_N^s]$, where N is the number of voxels and $s = 1, \ldots, S, S = 20000$ here. Once the ensemble is trained, the voxels can be sorted in descending order according to the sign consistency observed in their corresponding weights in all the classifiers that form the ensemble. Voxels whose weights show the same sign in all the classifiers are placed at the top of the list. The voxel selection method picks up voxels whose sign consistency exceedes a threshold r. The sign consistency score provides also an importance score $I(j)$ for a voxel j: Define $n(j) = |\{s : w_j^s < 0\}|$,$p(j) = |\{s : w_j^s > 0\}|$, then $I(j) = 2*\max(n(j), p(j))/S - 1$. There is a strong correlation between the sign consistency of the voxel and its discriminative capacity. A voxel that systematically appears with the same sign in most of the classifiers of the ensemble presents a robust discriminative power: its probability of being gray matter is indicative of one of the classes. On the other hand, the sign fluctuations of the non-consistent voxels (showing both signs in significant proportions) indicate that they are not relevant for the classification or that their relevance depend on the value of the consistent voxels in each particular subject, what leads to conclude that their importance is less. Moreover, since the members of the ensemble have been trained with an L_2 norm regularization that does not enforce sparsity in the primal, the voxel importance is computed at the ensemble level, leading to more robust results than those voxel importances computed at the individual classifiers level.

Furthermore, the L_2 norm regularization deals with brain areas formed by highly correlated voxels by splitting the magnitude of the weights among all the correlated voxels, thus preserving the regional organization of the signal. This causes that the selected voxels appear in disjoint compact clusters with all voxels in a same cluster having the same sign, what forms a voxel importance pattern.

Finally, notice that learning a few thousand classifiers does not involve a dramatic computational load since the L_2 norm SVM may be optimized in the dual space (in which the variables are the training samples), and in MRI data the number of training instances is in the order of tens to hundreds.

2) Transductive refinement of the voxel selection and importance: Classification tasks in sMRI are ultimately related to localized alterations of the brain structure. This means that most voxels in a brain scan are not related to the disease. In fact, most voxels in a brain scan contribute to separate that brain from the others. In [10] the identification of relevant voxels is enhanced by borrowing certain ideas from transductive learning and conformal analysis. Transduction refers to learning scenarios in which one has access to the observations, but not the labels, of the test set (see [13] on why this does not lead to testing on training data problem). Conformal analysis relies on a nonconformity measure, which measures how unusual each possible label for a test sample looks relative to the training labels. Therefore, one decides for the label that best conform the other labeled data.

These ideas refine the voxel importance $I(j)$ in the following way, resulting in the SCBconf importance measure. The selection of voxels (by thresholding the importance values) is run twice for every test sample. During the first run, the test sample has the label 1, and we obtain a subset of selected voxels V_+ with the importance greater than a threshold decided by an inner cross-validation loop. During the second run, the test sample has the label -1, and we similarly obtain a subset of selected voxels V_-. Finally, V_+ and V_- are intersected to arrive at the final set of selected voxels for a particular test sample. The intuition is that voxels appearing only in one of the subsets, but not in the intersection, strongly depend on the particular labeling, and should therefore not be selected as they are not relevant for disease discrimination but rather to separate individual subjects from the rest of the dataset.

The transductive refinement filters out voxels that obtain spuriously high importance after analyzing the ensemble. It has been adapted to the split-half sampling and the inner ten fold cross-validation framework of this paper in the following manner. The training set used to construct every ensemble (each classifier is trained sampling from this overall training set) is augmented by adding 10% of the test samples selected at random and with random labels. The augmented training set is processed with SCB to yield a set of voxels. Repeating this procedure 20 times yields 20 different sets of voxels. The final set of voxels are those that appear in the intersection of the 20 sets. Then, the importance of each of these voxels is computed by averaging its importance over the 20 repetitions.

III. Experiments and Results

A. Variable importance compared to t-test and elastic net

We compared the voxel importances obtained based on SCB and SCBconf methods to those obtained by a massively

univariate t-test and those by an elastic-net penalized logistic regression (ENET) [14]. With the t-test, we defined the voxel importance score as the absolute value of the corresponding t-statistic value. T-tests were computed without assuming equal variances. ENET models were trained using the GLMNET package. We considered two different α-values (we use the notation defined in[14]): $\alpha = 0.5$, balancing the sparsity enforcing L1-penalty and shrinkage favoring L2 penalty and $\alpha = 0.05$ placing more weight on the shrinkage term, hopefully leading to more dense (and stable) models. The value $\alpha = 0.05$ produces a strong grouping effect, forcing joint selection of the correlated voxels [15]. The parameters λ for the ENETs were optimized using 10-fold stratified cross-validation using mis-classification rate as the performance measure. The voxel-importances were defined as the absolute values of the weights for a voxel in the resulting linear classifier.

Figure 1 shows the voxel importance measures by different methods across an axial slice at Hippocampus and mid-temporal cortices. The basic characteristics of voxel importances by different methods can be well seen in Figure. SCB methods produced dense voxel importances, with a number of voxels receiving the normalized importance score greater than 0.2 (the threshold of 0.2 was heuristically selected to match the FDR corrected t-test threshold for visualization purposes). Also, the t-test gave an importance score to every voxel and for a number of voxels, this score exceeded the FDR-corrected threshold at $q = 0.05$. Instead, as expected, both ENET models produced only a sparse pattern of voxels with non-zero weights, 17 % of voxels received a non-zero importance with $\alpha = 0.05$. This is a clear disadvantage if a voxel ranking is desired. However, the pattern of the highest importances were more similar SCBs and ENETs than with SCBs and t-statistic importances. In particular, notice the lack of t-statistic based importance in the area in the orbitofrontal cortex, which both ENETs and SCB ranked highly important for the classification. Joint density plots of voxel importances are shown in Fig. 2, where it can be seen that t-test produced a very different voxel importance map than either ENET or SCBconf, whereas the main difference between ENET and SCBconf was that the former associates a large majority of voxels with zero importance. Interestingly, some voxels with maximal SCBconf importance received a zero-weight in ENET. To make a quantitative argument, the Spearman correlation (with a tie adjustment) between the t-test and SCBconf voxel importances was 0.17 while it was 0.51 for ENET-0.05 and SCBconf.

B. Reproducibility analysis

We analysed the reproducibility of the voxel importances derived based on SCB and SCBconf. For this, we adapted the split-half procedure of [11]. We sampled without replacement $N = 100$ subjects from each of the two classes. This procedure was repeated 100 times. We denote the two subject samples (split halves; train and test) A_i and B_i for the iteration $i = 1, \ldots, 100$. The sampling was without replacement so that the split-half sets A_i and B_i were always non-overlapping and are considered as independent train and test sets. SCB was trained on the split A_i and tested on the split B_i and, vice versa, trained on B_i and tested on A_i. The value of the sign consistency threshold (r) was determined in an inner 10 fold cross validation for each train/test partition.

Both SCB and SCBconf based voxel selection followed by SVM on selected voxels achieved an average classification accuracy (ACC) of 76 %. This generalization accuracy is roughly equivalent to one of a standard SVM in [11]. The embedded feature selection schemes such as ENET reached better accuracies in [11]. The average absolute difference in ACC between the two split-halves (ACC when trained on A_i and tested on B_i versus ACC when trained on B_i and tested on A_i) was 3.60 % with SCB and 3.27 % with SCBconf, thus the transduction stage reduced the variation in the ACC. The absolute difference, averaged over the voxel and runs, between voxel importances (normalized to the range [0,1]) was 0.147 (standard deviation across runs 0.004) for SCBconf and 0.320 (standard deviation across runs 0.004) for SCB, thus indicating that the transduction step improves the reproducibility of voxel importances.

IV. Conclusion

We have introduced and evaluated a new voxel importance measure based on sign consistency on the classifier ensembles, rooted in the voxel selection mechanism introduced in [10]. The experiments demonstrated that the approach was able to produce robust voxel importance estimates that were quantitatively different from ones provided by massively univariate hypothesis testing and elastic-net penalized logistic regression. While the ideas of random subsampling and random relabeling are widely used for variable importance and selection, for example, in out-of-bag variable importances of Random Forests [6], the idea of sign consistency is much less exploited and novel in brain imaging. Finally, it is important to note that the voxel importances derived based on voxel weights of multivariate classifiers cannot be interpreted in the same way as the statistic maps in the massively univariate analysis; a voxel can be important for the classification albeit showing very modest correlation with the class label.

Acknowledgments: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education,

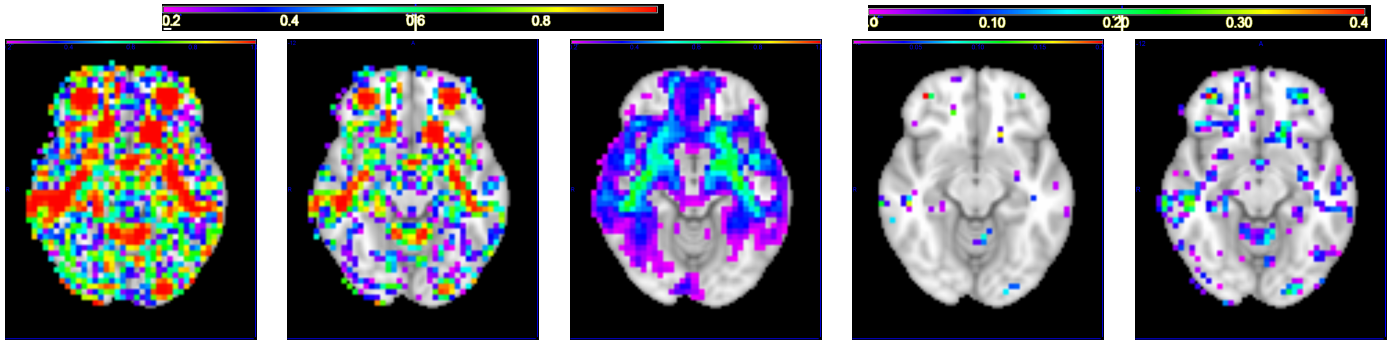

Fig. 1. Voxel importances using different methods. Axial slices at MNI coordinates $z = -12mm$ are shown. All importance scores are normalized by the maximum importance score. The threshold for the SCB and SCB-conf were set to 0.2 for visualization purposes. T-test importances were thresholded at FDR corrected $q = 0.05$ threshold (the threshold being 0.1967). ENET importances contain all voxels with non-zero weight in the optimal model decided by CV. The range of the colorbar is from 0.2 to 1 with SCB, SCBconf, and t-test and from 0 to 0.4 with ENETs.

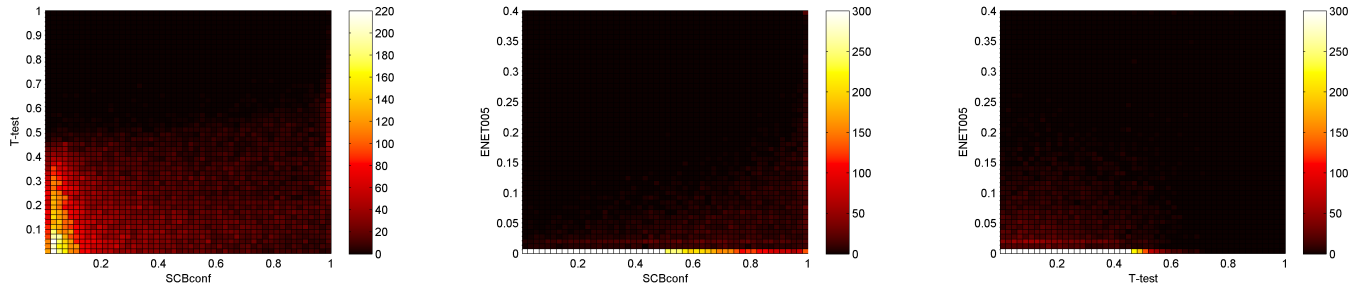

Fig. 2. Density plots comparing normalized voxel importances obtained with different methods. The figure shows that there were voxels with very high SCBconf importances that failed to reach statistical significance with t-test ($q = 0.05$ threshold 0.1967) and received the zero weight in the ENET with a low value of parameter α. Also, there were voxels that were significant in the t-test that were not important for the classification according to SCBconf or ENET.

and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Authors acknowledge support from Spain MINECO (grant TEC2014- 52289R).This project has received funding from the Universidad Carlos III de Madrid, the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement nr 600371, el Ministerio de Economa y Competitividad (COFUND2013-40258) and Banco Santander.

REFERENCES

[1] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *Neuroimage*, vol. 87, pp. 96–110, 2014.

[2] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014.

[3] J. R. Cohen, R. F. Asarnow, F. W. Sabb, R. M. Bilder, S. Y. Bookheimer, B. J. Knowlton, and R. A. Poldrack, "Decoding developmental differences and individual variability in response inhibition through predictive analyses across individuals," *The developing human brain*, p. 136, 2010.

[4] B. S. Khundrakpam, J. Tohka, and A. C. Evans, "Prediction of brain maturity based on cortical thickness at different spatial resolutions," *NeuroImage*, vol. 111, pp. 350–359, 2015.

[5] U. Grömping, "Variable importance assessment in regression: linear regression versus random forest," *The American Statistician*, 2012.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] G. Langs, B. H. Menze, D. Lashkari, and P. Golland, "Detecting stable distributed patterns of brain activation using gini contrast," *NeuroImage*, vol. 56, no. 2, pp. 497–507, 2011.

[8] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects," *Neuroimage*, vol. 104, pp. 398–412, 2015.

[9] D. Greenstein, J. D. Malley, B. Weisinger, L. Clasen, and N. Gogtay, "Using multivariate machine learning methods and structural mri to classify childhood onset schizophrenia and healthy controls," *Front Psychiatry*, vol. 3, p. 53, 2012.

[10] E. Parrado-Hernández, V. Gómez-Verdejo, M. Martínez-Ramón, J. Shawe-Taylor, P. Alonso, J. Pujol, J. M. Menchón, N. Cardoner, and C. Soriano-Mas, "Discovering brain regions relevant to obsessive–compulsive disorder identification through bagging and transduction," *Medical image analysis*, vol. 18, no. 3, pp. 435–448, 2014.

[11] J. Tohka, E. Moradi, and H. Huttunen, "Comparison of feature selection techniques in machine learning for anatomical brain mri in dementia," *Neuroinformatics*, p. in press, 2016.

[12] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, A. D. N. Initiative *et al.*, "BrainAGE in mild cognitive impaired patients: predicting the conversion to alzheimers disease," *PloS ONE*, vol. 8, no. 6, p. e67346, 2013.

[13] A. Gammerman, V. Vovk, and V. Vapnik, "Learning by transduction," in *AISTATS98*. Morgan Kaufmann Publishers Inc., 1998, pp. 148–155.

[14] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Software*, vol. 33, no. 1, pp. 1–22, 2010.

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc.: Series B*, vol. 67, no. 2, pp. 301–320, 2005.