THE 10TH ANNUAL MLSP COMPETITION: SECOND PLACE

Alexander V. Lebedev MD [1,2]
alexander.vl.lebedev@gmail.com

[1] University of Bergen
Department of Clinical Medicine (Bergen, Norway)
[2] Stavanger University Hospital
Centre for Age-Related Medicine (Stavanger, Norway)

ABSTRACT

The goal of the MLSP 2014 Classification Challenge was to automatically detect subjects with schizophrenia and schizoaffective disorder based on multimodal features derived from the magnetic resonance imaging (MRI) data. The patients with age range of 18-65 years were diagnosed according to DSM-IV criteria. The training data consisted of 46 patients and 40 healthy controls. The test set included 119 748 subjects with unknown labels. In the present solution, we implemented so-called "feature trimming", consisting of: 1) introducing a random vector into the feature set, 2) calculating feature importance based on mean decrease of the Gini-index derived by running Random Forest classification, and 3) removing the features with importance below the "dummy variable". Support Vector Machine with Gaussian Kernel was used to run final classification with reduced feature set achieving test set AUC of 0.923.

Index Terms — Schizophrenia, MRI, Random Forest, Support Vector Machines, Feature Trimming

1. INTRODUCTION

Schizophrenia (Sch) is a devastating neuropsychiatric disorder affecting around 0.3–0.7% of the population throughout the world [1]. Its etiology is largely unknown, but likely multifactorial with substantial contribution of genetic and prenatal factors [2].

Since the 19th Century, after E. Kraepelin's description of "dementia praecox" and then introduction of the term "schizophrenia" by E. Bleuler, this concept undergone many revisions, and is still a matter of hot debates [3].

Meanwhile, even considering various clinical manifestations of the disorder and absence of any biomarkers that would be implemented into the diagnostic criteria, recent studies have clearly demonstrated that schizophrenia is a brain disease, possibly manifesting as a temporo-limbic and prefrontal dysconnectivity syndrome that affects circuits involved in cognitive integration [4, 5]. Moreover, recent imaging studies employing methods of multivariate statistics and machine learning have revealed an opportunity not only to detect schizophrenia using biological features [6], but also to successfully predict the disease onset in subjects who are at risk for psychosis but yet do not meet the clinical criteria [7].

All of the above encourages multidisciplinary research of image-based computer-aided diagnostic tools that will hopefully improve early diagnosis of schizophrenia. The later is very important for optimal patient management and is associated with better clinical outcome.

The MLSP 2014 Classification Challenge was focused on automated detection of subjects with schizophrenia and schizoaffective disorder based on multimodal features derived from the magnetic resonance imaging (MRI) data. In the present paper, we are reporting our 2nd place solution.

2. METHODS

2.1 Data

The whole sample included the training set (46 patients with schizophrenia and schizoaffective disorder and 40 healthy controls [HCs]), and the testing data (119 748 subjects with unknown labels). The patients were diagnosed according to DSM-IV criteria for schizophrenia and schizoaffective disorder during structured interview [8].

Structural magnetic resonance images (sMRI) and resting state functional MRI (rs-fMRI) data were acquired on a 3T MRI scanner at the Mind Research Network (Albuquerque, New Mexico).

Image preprocessing was performed using Statistical Parametric Mapping software, version 5 (SPM5):

(http://www.fil.ion.ucl.ac.uk/spm). Further feature extraction was done using independent component analysis, as implemented in the GIFT Toolbox (http://mialab.mrn.org/software/gift/), yielding source-based morphometric (SBM) loadings and Functional Network Connectivity (FNC) features for sMRI and rs-fMRI, correspondingly.
For acquisition and preprocessing details of structural and functional imaging data, including feature extraction protocol, see [9, 10]. For the present competition, SBM and FNC features were already available for the participants.

2.2 Implemented Solution

"Feature Trimming"

At the first step, after feature concatenation we performed the procedure, which will be further called "feature trimming". Its steps are straightforward and in very simple words can be described as: 1) introducing a random vector into the feature set, 2) calculating feature importance based on mean decrease of the Gini-index derived by running Random Forest (RF) classification, and 3) removing the features with importance below the "dummy variable". More detailed description follows below.

The Random Forest algorithm is formally defined as a collection of tree-structured classifiers:

$$f(x,\ \theta_k), k = 1, 2, \dots, K\ ,$$

where θ_k are random i.i.d. vectors (independent and identically distributed) [11] and each tree provides a vote for the most popular label at input x [12]. For classification problems, the forest prediction output is the majority vote. The algorithm converges with large number of trees [12].
Here we are given a set of training data $D = \{(v^i, c^i)\}_{i=1,\dots,n}$ (to be defined below). The classification task is then to learn a general mapping from previously unseen test data to their corresponding diagnostic labels, i.e. $c: R^d \to C$. More specifically, adopting the notation in [13], let $v = (x_1, \dots, x_d) \in R^d$ denote the input data feature vector (predictor), and let $c \in$ C denote the output diagnostic label (response). In our case, x_i is a measure (SBM or FNC feature) derived from the ICA analysis briefly mentioned above (d = number of features; e.g., 41 – for volumetric data), and C = {Sch, HC}. The RF algorithm incorporate a collection of binary classification trees indexed by $t = 1, \dots, n_{tree}$. Each classification tree is characterised by its input root node, internal split nodes, and its leaf terminal nodes containing class labels.
In this setting, the RF algorithm can be briefly described as follows: (i) Draw n_{tree} samples from the original data D, using random sampling with replacement; (ii) For each bootstrap sample, grow a classification tree such that for each node: randomly sample m_{try} of the predictor variables and chose the "best split" according to the Gini criterion defined below from among those feature variables ($1 < mtry \ll d$). The largest tree possible is grown and is not pruned. Using only m_{try} of the predictor variables selected at random is in contrast to standard tree classification (CART), where each node is split using the best split among all d variables; (iii) the forest consists of n_{tree} trees. Each tree gives a classification for a given data point. Predict new data point x by putting x down each of the n_{tree} trees and make a majority vote for classification across the forest.
For a given tree, let S_0 denote the set of input predictor data vectors that is fed into the root node, S_j be the subset of data points reaching node j in the binary split tree, and $\{S_j^L, S_j^R\}$ denote the subset of data points that reaches the left and right child, respectively, of node j, where $S_j^L \cup S_j^R = S_j$ and $S_j^L \cap S_j^R = \theta$. In the "off-line" tree training, each split node j is associated with a parameter vector θ_j that is trained by optimizing an objective function I (defined below), i.e.:

$$\theta_j = \arg max_{\theta \in T}\, I(S_j, \theta)$$

In this notation, a binary-valued test function $h(v, \theta_j) : R^d \times T \to \{0, 1\}$ is applied at each split node j. Here, 0 and 1 denote "false" and "true", respectively, and the data point v arriving at split node j is sent to its left (0) or right (1) child node, accordingly. T is the set of all possible split function parameters, and $T_j \subseteq T$ is the subset of parameters available at node j. We thus have $S_j^L(S_j, \theta) = \{(v, c) \in S_j \mid h(v, \theta) = 0\}$ and $S_j^R(S_j, \theta) = \{(v, c) \in S_j \mid h(v, \theta) = 1\}$.
The objective function used is the Gini index, i.e.:

$$I = i(\tau) = 1 - \sum_{c \in C} p_c^2,$$

measuring the likelihood that a data point would be incorrectly labeled if it was randomly classified according to the distribution of class labels within the node. The optimal binary split is then the one that maximises the improvement (mean decrease) in the Gini index, which was used as a metric in our approach. To be more specific, at every split node τ one of the m_{try} variables, say x_k, is used to form the split and there is a resulting decrease in the Gini index. The mean decrease of the Gini index, $\Delta i(\tau)$ was used as a metric, i.e.:

$$\Delta i(\tau) = i(\tau) - (p_L i(\tau^L) + p_R i(\tau^R))$$

where $i(\tau) = 1 - \sum_{c \in C} p_c^2$ is the Gini index at node τ, $p_L = \frac{|S_j^L|}{|S_j|}$ and $p_R = \frac{|S_j^R|}{|S_j|}$ are the probabilities of sending a data point to the left and the right node, respectively.
This metric reflects the contribution of a variable x_k to the node homogeneity of τ. Thus, a higher mean decrease of the

Gini index for a particular feature means that the variable is present more often in nodes with higher purity among all trees in the forest (overall). The sum of all decreases in the forest due to a given variable x_k, normalized by the number of trees, therefore gives an estimate of its Gini importance (Eq. 3), i.e.:

$$I_G(x_k) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \sum_{\tau} \Delta i_{x_k}(\tau, t)$$

Therefore, the Gini importance $I_G(x_k)$ indicates how frequent the particular feature x_k was selected in a split node, and how large its overall discriminative value was for the classification task. Finally, if you introduce a random "dummy" feature and calculate its mean decrease of the Gini index, you can then exclude ("trim") everything with importance below this value.

Final Model

The reduced feature set was then used to run final classification employing Support Vector Machine (SVM) with Gaussian Kernel [14]. The optimization problem:

$$max\left\{\sum_{i=1}^{N} \alpha - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j, \sigma)\right\}$$

with $K(x_i, x_j, \sigma) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

The reason for not using tree-based ensembles (Random Forest and boosted trees) was empirical – because SVM resulted in superior cross-validated accuracy (of note, for RF both out-of-bad estimation and cross-validation were assessed, achieving similar values)

3. RESULTS

The original dataset contained 410 features (32 for SBM and 378 for FNC). After the feature trimming, we ended up with 122 variables.

Next, we estimated hyperparameter σ (sigma, width parameter) for the Gaussian Radial Basis kernel and tuned C parameter (the penalty factor, controlling trade-off between model complexity and proportion of nonseparable instances) using leave-one-out cross validation for the final SVM classifier. The resulted test set area under the receiver operating characteristic curve (AUC) was 0.923.

Of note, cross-validated performance of various models that had been tested (RF, boosted trees, neural network, SVM) varied around 0.8 and 0.85 (for overall accuracy) and the public scores that we were receiving after the submissions were unstable. Therefore, we decided not to implement more complex solutions (ensembling, hierarchical models) and stopped on a relatively simple model.

4. REPRODUCIBILITY

The data were accessed and downloaded via the MLSP-2014 Schizophrenia Classification Challenge webpage:

- https://www.kaggle.com/c/mlsp-2014-mri

Step-by-step instructions with the code descibing our solution can be found at:

- https://github.com/alex-lebedev/Kaggle-MLSP-2014

5. ADDITIONAL COMMENTS

In general, it was somewhat difficult to evaluate performance of the models, since the mismatch between cross-validated accuracies and the feedback that we were receiving during submissions was very big (with private AUC scores varying around 0.65). It was the main reason why we stopped and did not try to improve our model further. Meanwhile, additional feature selection (e.g., recursive feature elimination, sparsity-based approaches) and/or classifier ensembling could potentially result in a superior performance.

6. COPYRIGHT FORM

The MIT License (MIT) file can be found at:

- https://github.com/alex-lebedev/Kaggle-MLSP-2014

7. REFERENCES

[1] J. van Os and S. Kapur, "Schizophrenia," *Lancet,* vol. 374, pp. 635-45, Aug 22 2009.

[2] C. A. Ross, R. L. Margolis, S. A. Reading, M. Pletnikov, and J. T. Coyle, "Neurobiology of schizophrenia," *Neuron,* vol. 52, pp. 139-53, Oct 5 2006.

[3] A. Jablensky, "The diagnostic concept of schizophrenia: its history, evolution, and future prospects," *Dialogues Clin Neurosci,* vol. 12, pp. 271-87, 2010.

[4] M. E. Shenton, T. J. Whitford, and M. Kubicki, "Structural neuroimaging in schizophrenia: from methods to insights to treatments," *Dialogues Clin Neurosci,* vol. 12, pp. 317-32, 2010.

[5] W. Pettersson-Yeo, P. Allen, S. Benetti, P. McGuire, and A. Mechelli, "Dysconnectivity in

schizophrenia: where are we now?," *Neurosci Biobehav Rev,* vol. 35, pp. 1110-24, Apr 2011.

[6] Y. Takayanagi, Y. Kawasaki, K. Nakamura, T. Takahashi, L. Orikabe, E. Toyoda*, et al.*, "Differentiation of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables," *Prog Neuropsychopharmacol Biol Psychiatry,* vol. 34, pp. 10-7, Feb 1 2010.

[7] N. Koutsouleris, E. M. Meisenzahl, C. Davatzikos, R. Bottlender, T. Frodl, J. Scheuerecker*, et al.*, "Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition," *Arch Gen Psychiatry,* vol. 66, pp. 700-12, Jul 2009.

[8] M. First, R. Spitzer, M. Gibbon, and J. Williams, " Structured Clinical Interview for DSM-IV Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP) New York, NY: Biometrics Research, New York State Psychiatric Institute.," 2002.

[9] J. M. Segall, E. A. Allen, R. E. Jung, E. B. Erhardt, S. K. Arja, K. Kiehl*, et al.*, "Correspondence between structure and function in the human brain at rest," *Front Neuroinform,* vol. 6, p. 10, 2012.

[10] M. S. Cetin, F. Christensen, C. C. Abbott, J. M. Stephen, A. R. Mayer, J. M. Canive*, et al.*, "Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia," *Neuroimage,* vol. 97, pp. 117-26, Aug 15 2014.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*: Wiley-Interscience, 2006.

[12] L. Breiman, "Random Forests," *Machine Learning,* vol. Volume 45, Number 1, pp. 5-32, 2001.

[13] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis* vol. XIX: Springer, 2013.

[14] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 1995.