Automating Neurological Disease Diagnosis Using Structural MR Brain Scan Features

Allan Raventós and Moosa Zaidi

Stanford University

I. Introduction

Nine percent of those aged 65 or older and about one third of those aged 85 or older have Alzheimer's disease.[1] The incidence of Alzheimer's is expected to triple from 2010 to 2050.[2] 1.1% of American adults are Schizophrenic.[3] Currently, both of these diseases are diagnosed primarily through a clinical mental health exam. However, Alzheimer's Disease and more recently Schizophrenia have been shown to have a strong neuroanatomical footprint that appears in a Magnetic Resonance (MR) scan.[4,5]

Figure 1 shows a healthy brain along with one afflicted with advanced stage Alzheimer's Disease (both from the OASIS dataset, described later). Severe tissue loss and important structural changes can be seen in the latter. This guides our intuition that machine learning models based on structural features should prove very effective. Our objective in this research project is to develop tools to automate (or at least assist) diagnosis and screening of these diseases using structural MR brain scan features.

Then in a hospital setting, a brain MR scan could be obtained, its structural features automatically generated as we will describe later, and then fed into the models we develop in order to assess whether a patient has Schizophrenia or Alzheimer's Disease.

II. Related Work

Sabuncu and Konukoglu[6] at the Athinoula A. Martinos Center for Biomedical Imaging run a survey of machine learning algorithms on the datasets we use here. Their work is helpful in providing benchmarks for the kinds of prediction accuracies that can be achieved for each disease-algorithm combination. However, their work does not include robust feature selection or hyperparameter fitting, which we carry out.

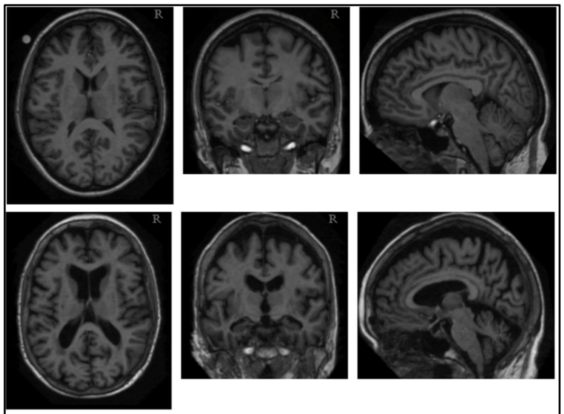

Figure 1: Top row: healthy brain without Alzheimer's Disease. Bottom row: brain with advanced stage of Alzheimer's Disease (Case 2 in OASIS dataset). We can clearly see cortical shrinkage, hippocampus shrinkage and enlarged ventricles in Row 2 as compared to Row 1.

Orru et al.[7] perform a review of the current state of research in using Support Vector Machines (SVMs) to identify biomarkers for various psychiatric diseases, and conclude that, although harder to implement in a hospital setting, the tools show promise. Their work provides us with some of the top benchmarks for SVM classification (including the possibility of predicting Alzheimer's Disease with upwards of 80% accuracy), allowing us to gauge the quality of our results.

III. Dataset and Features

Our data was provided by the Athinoula A. Martinos Center for Biomedical Imaging.[6] It consists of Schizophrenia data from MIND Clinical Imaging Consortium (MCIC) and Alzheimer's data from Open-Ended Series of Imaging Studies (OASIS). The OASIS data consists of Case 1, a mild Alzheimer's (defined by a Clinical Dementia Rating > 0), and Case 2, advanced stage Alzheimer's (Clinical Dementia Rating ≥ 1).

Furthermore, for each, we consider two sets of features. F1 contains the volumes of 45 anatomical structures (e.g. cerebral cortex, lateral ventricle), as well as 68 thicknesses of cortical parcellations (e.g. anterior frontal). F2 contains 20,484 values of cortical thickness smoothed with a Gaussian kernel. F1 and F2 were each extracted from 150 three-dimensional brain scans for each disease using FreeSurfer Computer Vision software. This feature extraction was carried out by [6].

Both F1 and F2 contain features that differentiate between the left hemisphere and right hemisphere of the brain. F1 contains, for example, metrics for left and right brain white matter, whereas F2 is a less-preprocessed concatenation of left hemisphere and right hemisphere brain data. One thing we intended to examine is any major differences in the usefulness of left and right brain features, given that Alzheimer's Disease is reported to more strongly affect left brain structures than right brain structures.[8]

IV. Algorithms

We employed various algorithms in our attempt to develop the most appropriate model for each disease.

As a first attempt, we consider Naïve Bayes (NB). NB assumes the conditional distributions of the features with respect to the label to be independent, and predicts labels as

$$\hat{y} = \arg \operatorname*{MAX}_{y} P(y) \prod_{i=1}^{n} P(x_i|y)$$

The parameters are chosen to maximize the joint likelihood of the observed data. Taking the conditional distributions to be Bernoulli and including Laplace smoothing, these maximum likelihood estimates of the parameters are

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\left\{x_j^{(i)} = 1, y^{(i)} = 1\right\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} + 2}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\left\{x_j^{(i)} = 1, y^{(i)} = 0\right\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} + 2}$$

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{\, y^{(i)} = 1\}}{m}$$

Because our feature are continuous we need to discretize them in order to apply NB. We do this by splitting features at their median value. We investigated splitting at the mean, and we also investigated splitting into more intervals and taking the features' conditional distributions to be Multinomial. However, we found splitting at the median to give highest accuracy using cross-validation. Splitting at the median was likely particularly effective because we had balanced datasets with an equal number of examples in each class.

As an alternative to needing to discretize features we could instead modify NB to model the conditional distributions as Gaussian rather than Bernoulli/Multinomial. This modification results in Gaussian Naïve Bayes (GNB). Our maximum likelihood estimates of the parameters of the conditional distributions then become

$$\mu_{j,y=k} = \frac{\sum_{i=1}^{m} x_j^{(i)} 1\{y^{(i)} = k\}}{\sum_{i=1}^{m} 1\{y^{(i)} = k\}}$$

$$\sigma_{j,y=k}^2 = \frac{\sum_{i=1}^{m} (x_j^{(i)} - \mu_{j,y=k})^2 1\{y^{(i)} = k\}}{\sum_{i=1}^{m} 1\{y^{(i)} = k\}}$$

where k is the label. Intuitively we are simply taking the average mean and average variance of feature across examples.

We also consider the Support Vector Machine (SVM) under a Gaussian kernel. That is, we take the following primal optimization problem

$$\min_{w,b,\xi} \frac{1}{2}w^T w + C\sum_{i=1}^{m} \xi_i$$
$$s.t.\ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

with the kernel
$K\left(x^{(i)}, x^{(j)}\right) = \exp\left(-\gamma\left\|x^{(i)} - x^{(j)}\right\|^2\right)$.

This optimization problem attempts to separate test samples with the disease from those without the disease in an infinite dimensional feature space. (We can see that we are using infinite features by noticing that the kernel itself is an infinite sum over polynomials.) We decided that a Gaussian kernel would be better suited to separate brain scan features than a linear one, since especially in the case of F2, which is a rawer, almost voxel-level dataset, there is no reason to expect that the data would be linearly separable in its initial dimension.

We also consider the ν-SVM, an alternative formulation of the SVM presented above, again using a Gaussian kernel. ν parametrizes the fraction of allowed errors and minimum number of support vectors. The parameter ν is restricted to (0,1], whereas the C in the standard SVM can take any positive value. The primal optimization problem is as follows:

$$\min_{w,b,\xi,\rho} \frac{1}{2}w^T w - \nu\rho + \frac{1}{m}\sum_{i=1}^{m} \xi_i$$
$$s.t.\ y^{(i)}\left(w^T x^{(i)} + b\right) \geq \rho - \xi_i$$
$$\xi_i \geq 0, \rho \geq 0$$

Finally, we also consider Random Forests (RF). A decision tree classifier classifies an input vector by traversing a tree where each node is a feature and each edge one of the possible values of the preceding node. Usually the tree is learned by splitting greedily. Random Forests extend decision tree classifiers to make them less susceptible to fitting. This is done by having multiple trees, providing each tree only a random sample of the training examples, requiring each tree to only consider a random subset of the features at each split, and finally having the trees vote to determine the overall result.

V. Results

For the purposes of evaluating our algorithms, we use five-fold cross-validation, since this was the standard set by the Machine Learning Challenge 2014, which used a subset of the dataset we consider here.

Where computationally feasible we use full backwards feature search with cross-validation. When choosing features for the huge F2 set, however, we use a variance-thresholding approach. First, we rank all 20,484 features in descending order based on their variance across all samples. We then run the algorithm first on the highest variance feature, and then move in steps, adding the next 50 highest variance features at each step. We chose 50 experimentally, by observing that with such "raw" data, predictive ability didn't increase substantially by adding one feature at a time. Finally, we choose the number of features that achieves the highest predictive success across all tests performed.

Table 1 reports the top results found for each algorithm and dataset. First of all, we notice a predominantly higher performance when predicting Alzheimer's rather than Schizophrenia (SCZ). This makes sense since Alzheimer's takes a much higher toll on brain structure. We also see almost exclusively stronger performance when running on advanced stage Alzheimer's (AD2) as compared to mild Alzheimer's (AD1), which is intuitive since the former is expected to have a noticeable effect on physical brain structures.

Naïve Bayes had unusually high prediction accuracy throughout when running on five-fold cross-validation. It also performed consistently better than Gaussian Naïve Bayes. The fact that ordinary Naïve Bayes consistently outperformed its Gaussian counterpart suggests that the GNB assumption that the features follow a Gaussian conditional distribution is not accurate.

When comparing SVM and the ν-SVM we see that we are able to achieve better fitting when using the ν-SVM. However, both models lead to high prediction accuracy, which would suggest

	SCZ – F1	SCZ – F2	AD 1 – F1	AD 1 – F2	AD 2 – F1	AD 2 – F2
NB	0.72	0.73	0.79	0.77	0.90	0.93
GNB	0.62	0.60	0.63	0.67	0.76	0.74
SVM	0.66	0.59	0.71	0.73	0.80	0.84
ν-SVM	0.80	0.60	0.78	0.73	0.92	0.84
RF	0.66	0.64	0.74	0.65	0.76	0.77

Table 1: Top results achieved for each algorithm (row) and dataset (column).

that the data becomes highly separable when using the Gaussian kernel.

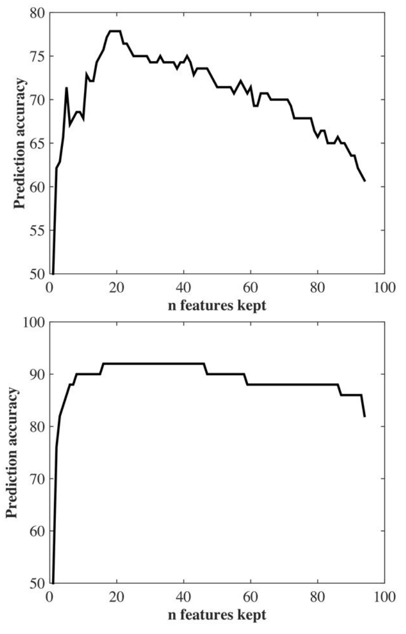

Figure 2: Backwards search for nu-SVM run on AD1-F1 (above) and AD2-F1 (below)

We note the confusion matrix for the ν-SVM operating on AD2-F1 (total instances averaged out over the five folds of cross validation):

$$\begin{bmatrix} 4.2\ t.p. & 0.8\ f.n. \\ 0\ f.p. & 5\ t.n. \end{bmatrix}$$

We observe that we get no false positives, which means that a positive output from the algorithm carries a lot of weight. At the same time, we do note that the ν-SVM results in a small number of false negatives. In a clinical context, this implies that when the algorithm provides a negative output, we should take care to further assess the patient's condition.

SVM with a Gaussian kernel is prone to overfitting on features since we are operating in an infinite dimensional space and are trading off an increase in variance for a reduction in bias. We try to control for this in mainly two ways. Firstly we make sure that our feature selection curves are not very noisy or oscillating near our choice of number of features. This smoothness criterion is a more qualitative metric, but serves to show that the result is consistent, and not one that shoots up and down when single features are added. Figure 2 shows the feature selection curves for running ν-SVM on AD1-F1 and AD2-F1. As we can see one peaks and the other is fairly constant, yet both are generally smooth.

Another way in which we try to alleviate the risk of overfitting in SVM is that after completing feature selection, we split the dataset into training and testing subsets, train the testing set on the smaller number of features and ensure that its performance remains strong.

It is interesting to note that feature selection using variance thresholding (used with SVMs on F2) led to much greater gains in accuracy for the Alzheimer's datasets than for the Schizophrenia dataset. For AD1-F2 and AD2-F2 we saw accuracy gains of 8 to 12 percent, whereas for SCZ-F2 the gain is zero (and projecting the features onto the right singular vectors of the data matrix does not help). We interpret this as the features that contribute to better prediction being more clearly separable on the basis of greater variance from features that introduce background noise for the AD data. This is reasonable: given that Alzheimer's damages brain structures more strongly, we would expect for the smoothed voxels in the affected structures to have a much higher variance resulting from greater variation between healthy and diseased brains compared to background variation in the smoothed pixels independent of the AD classification

In regards to our hypothesis that left brain features would mores strongly predict Alzheimer's, we find no considerable preference for left hemisphere features in our feature selected models. When running on F2 the features chosen when running our variance thresholding algorithm are consistently about half right-brain and half left-brain. The same seems to hold for F1, where if a left brain feature is selected, its right brain analog is usually selected as well. These results suggests which structure a region of the brain belongs to is far more important than which hemisphere it belongs to in relation to its connection with Alzheimer's.

Random Forest (RF) had moderate performance compare to the other algorithms. Since the results of RF are inherently random we averaged several trials in reporting results. Testing on the same training data we were able to achieve over 95% accuracy with as few as 5 trees and consistently 100% accuracy with 10 trees. As seen in the Table 1 the accuracy with 5-fold cross validation was significantly lower, however. For cross validation we saw improvement of accuracy increasing the number of trees up to around 100, for further orders of magnitude, further gains were minimal and computation time increased dramatically. All these results suggest that RF has very low bias but continues to struggle with the high variance inherited from decision trees despite randomization. Increasing the number of trees reduces variance but has diminishing returns.

VI. Conclusions / Future Work

We are able to obtain very high prediction accuracies using the features extracted from MR brain scans. Of particular note are ν-SVM achieving 80% accuracy on Schizophrenia F1 and 92% accuracy on Alzheimer's F1, both of which improve on results found in previous papers, as well as the >90% results of Naïve Bayes on AD2.

Our results suggest it is possible to diagnose Schizophrenia and Alzheimer's with high accuracy using machine learning applied to structural brain MR scans. Our results confirm prior findings of the significant neuroanatomical footprint of theses disease. Our work also suggests that further increasing diagnosis accuracy is a promising direction for future work. Even if automated MRI based diagnosis of Alzheimer's and Schizophrenia is not used standalone, the technology could assist doctors making a diagnosis or flag at-risk patients. In fact, implementing a machine learning system that examines structural MRI data along with additional clinical data could be another promising direction for future work. For example, including age would likely greatly increase prediction of Alzheimer's. What is remarkable about these results and those of related studies is the ability to achieve high accuracy from just the images.

For future work, we would certainly like to expand this study to other datasets we found, including one for ADHD. We would further like to contact doctors in order to see how this kind of method could actually be employed in practice.

VII. Acknowledgements

We would like to thank our project mentor, Irene Kaplow, for her support throughout the project, as well as Mert Sabuncu of the Martinos Center for Biomedical Imaging for corresponding with us and providing data and advice.

VIII. References

1. Alzheimer’s Association, 2014 Alzheimer’s Disease Facts and Figures, Alzheimer’s & Dementia, Volume 10, Issue 2

2. Hebert, L. E., Weuve, J., Scherr, P. A., & Evans, D. A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*, *80*(19), 1778–1783. http://doi.org/10.1212/WNL.0b013e31828726f5

3. McGrath, J., Saha, S., Chant, D., & Welham, J. (2008). Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic reviews*, *30*(1), 67-76.

4. Jack, C. R., et al. (2008), The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging, 27: 685–691. doi: 10.1002/jmri.21049

5. Haukvik, U. K., Hartberg, C. B., & Agartz, I. (2013). Schizophrenia--what does structural MRI show?. *Tidsskrift for den Norske laegeforening: tidsskrift for praktisk medicin, ny raekke*, *133*(8), 850-853.

6. Sabuncu, M. R., Konukoglu, E. & Alzheimer’s Disease Neuroimaging Initiative. (2015). Clinical prediction from structural brain MRI scans: a large-scale empirical study. *Neuroinformatics*, *13*(1), 31-46.

7. Orru, G. et al. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev*. 2012 Apr;36(4):1140-52.

8. NIH, 2011-2012 Alzheimer's Disease Progress Report, Understanding the Biology of Alzheimer’s Disease and the Aging Brain 2012, https://www.nia.nih.gov/alzheimers/publication/2011-2012-alzheimers-disease-progress-report/understanding-biology-alzheimers.

9. Chih-Chung Chang and Chich-Jen Lin, LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1—27:27,2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

10. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.