Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images

Carlton Chu a,1, Ai-Ling Hsu b,1, Kun-Hsien Chou c, Peter Bandettini a, ChingPo Lin b,c,*

ABSTRACT

There are growing numbers of studies using machine learning approaches to characterize patterns of anatomical difference discernible from neuroimaging data. The high-dimensionality of image data often raises a concern that feature selection is needed to obtain optimal accuracy. Among previous studies, mostly using fixed sample sizes, some show greater predictive accuracies with feature selection, whereas others do not. In this study, we compared four common feature selection methods. 1) Pre-selected region of interests (ROIs) that are based on prior knowledge. 2) Univariate t-filtering. 3) Recursive feature elimination (RFE), and 4) t-test filtering constrained by ROIs. The predictive accuracies achieved from different sample sizes, with and without feature selection, were compared statistically. To demonstrate the effect, we used grey matter segmented from the T1-weighted anatomical scans collected by the Alzheimer’s disease Neuroimaging Initiative (ADNI) as the input features to a linear support vector machine classifier. The objective was to characterize the patterns of difference between Alzheimer’s disease (AD) patients and cognitively normal subjects, and also to characterize the difference between mild cognitive impairment (MCI) patients and normal subjects. In addition, we also compared the classification accuracies between MCI patients who converted to AD and MCI patients who did not convert within the period of 12 months. Predictive accuracies from two data-driven feature selection methods (t-test filtering and RFE) were no better than those achieved using whole brain data. We showed that we could achieve the most accurate characterizations by using prior knowledge of where to expect neurodegeneration (hippocampus and parahippocampal gyrus). Therefore, feature selection does improve the classification accuracies, but it depends on the method adopted. In general, larger sample sizes yielded higher accuracies with less advantage obtained by using knowledge from the existing literature.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Common questions that investigators have about analyzing image data often concern how best to represent it. For morphometric analysis, it is common for researchers to ask whether they should use voxel-based morphometry, cortical thickness mapping, or any one of a vast number of other possible representations. No one approach is optimal in all situations. One that works well for differentiating among some populations may not work so well when used for others. The empirical way to determine which is “better” involves assessing how well the various representations are able to actually discriminate among the populations. Studies that apply predictive and classification models are based on the predictive validity (Forster, 2002). If the classification accuracy is higher than chance, it implies that the input features (i.e. could be the whole brain or a small brain region) contain some information to discriminate among the populations (Ashburner and Koppel, 2011). There is increasing interest in applying multivariate pattern analysis (MVPA) to study the patterns of neurodegenerative diseases and mental disorders using anatomical MRI, including Alzheimer’s diseases (AD), Huntington’s Diseases (HD), major depression disorders (MDD), schizophrenia, autism spectrum disorder (ASD) (Chu, 2009; Costafreda et al., 2009; Eger et al., 2008; Fan et al., 2005; Koppel et al., 2009; Koppel et al., 2008), and many other disorders or diseases, which had been applied to

* Corresponding author at: Institute of Neuroscience, National Yang-Ming University, 155, Nong st., Sec. 2, Peitou, Taipei, Taiwan, ROC. E-mail address: cplin@ym.edu.tw (C. Lin).

1 These authors contributed equally to this work.

2 Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). 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. ADNI investigators include (complete listing available at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf).
Voxel Based Morphometry (VBM), especially studies on classification of AD have grown significantly in recent years.

The high dimensionality of the input feature space in comparison with the relatively small number of subjects (curse of dimensionality) is a widespread concern, so some form of feature selection is often applied (Costafreda et al., 2009; Fan et al., 2007; Guyon and Elisseeff, 2003; Pelaez-Coca et al., 2010). However, many studies have used support vector machine (SVM) and other kernel methods. These approaches search for a solution in the kernel space, so the optimized parameters are bounded by the number of training samples, rather than the dimension of the input features. In other words, kernel methods implicitly reduce the dimensionality of the problem to that of the number of samples (subjects in this case). Besides, neuroimaging data exhibit high correlations among features, implying a much lower “effective dimensionality” than the actual number of voxels, and samples (images) are likely to occur on a lower-dimensional manifold. Nevertheless, feature selection does have benefits other than alleviating the effect of the curse of dimensionality. One practical benefit is that it speeds up the testing process. Some also claim that it can make interpretation easier. It is often simpler to explain results from those subsets of the data that are sufficient to maintain equivalent classification accuracies (Marquand et al., 2011). The main benefit claimed for feature selection, which is the main focus in this manuscript, is that it increases classification accuracy. It is believed that removing non-informative signal can reduce noise, and can increase the contrast between labelled groups. There are some researchers who think it is necessary to apply feature selection. Some reviewers also make the papers difficult to publish if some feature selection has not been used. However, there is little evidence to support the claim that feature selection is always necessary.

Approaches of feature selection have shown many promising results in the field of bioinformatics (Saeng et al., 2007). A lot of feature selection methods and techniques originated or were specifically developed for bioinformatics. However, the properties of input features in bioinformatics are very different from features in neuroimaging data, especially as there are often more independent features in bioinformatics data. Among previous studies of diseases classification using imaging data, mostly using a fixed sample size, some show higher classification accuracies with feature selection. However, a recent study (Cuingnet et al., 2010) using 509 subjects showed no accuracy improvement with feature selection. We hypothesize that any improvement due to feature selection depends on sample size and the prior knowledge that the feature selection method utilizes. A priori, we believe that when the training sample is sufficiently large, feature selection would have a tiny, or even negative, impact on performance. Based on Occam’s razor (the principle of parsimony) we should prefer the simpler method when the performance is equivalent to more sophisticated ones. That is to say, perhaps it is preferable not to apply feature selection when no improvements of classification performance are shown.

To test our hypothesis, we compared the accuracy of discriminating AD from normal controls (NC) and discriminating mild cognitive impairment (MCI) from NC using four common feature selection methods and different training set sizes. The imaging data came from the Alzheimer’s disease Neuroimaging Initiative (ADNI), which have been used by many studies of AD classification (Filipovych and Davatzikos, 2011; Gerard et al., 2009; Pelaez-Coca et al., 2011; Stonnington et al., 2010; Yang et al., 2011; Zhang et al., 2011). To obtain the maximum amount of training data of the same modality, we used a single T1-weighted image from each subject. The pre-processing pipeline was the same as is often used in voxel-based morphometry (VBM) studies, and involved using Jacobian-scaled, spatially normalized gray matter (GM) as the input features. Cross-validation and resampling techniques were applied to estimate the accuracies. The performance, with and without feature selection, was compared statistically.

2. Methods

This section describes the methods of feature selection and machine learning applied to the ADNI data, and the pre-processing of the image data.

2.1. ADNI dataset

Data used in the preparation of this article were obtained from the Alzheimer’s disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI), which was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers and clinical and neuropsychological assessment can be combined to measure the progression of amnestic MCI and early probable AD. The determination of sensitive and specific biomarkers of early AD and disease progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

Although ADNI includes various image modalities, we only used the more abundant T1-weighted images obtained using MPRAGE or equivalent protocols with varying resolutions (typically 1.25 × 1.25 mm in-plane spatial resolution and 1.2 mm thick sagittal slices). Only images obtained using 1.5 T scanners were used in this study, and we used the first time point if there are multiple images of the same subject acquired at different times. The sagittal images were pre-processed according to a number of steps detailed under the ADNI website, which corrected for field inhomogeneities and image distortion, and were resliced to axial orientation. We visually inspected the gray matter (GM) map of all the subjects (native space) after tissue segmentation. Images that could not be segmented correctly were removed (list of removed subjects are in the Supplementary material). The dataset, which we used in the comparisons of feature selection, contained 260 MCI (86 females, 174 males), 188 NC (96 females, 92 males), and 131 AD (63 females, 68 males) (Table 1). The list of subjects used in this manuscript can be found in the Supplementary material.

In addition, a subset of MCI patients who had at least three longitudinal scans (baseline image, and two subsequent images six months and twelve months after) was selected. The converters were defined as subjects who had MMSE scores between 20 and 26 (inclusive), and CDR of 0.5 or 1.0 in the second or the third scan. The non-converters were defined as subjects who had MMSE scores between 24 and 30 (inclusive), and CDR of 0.5 in all three scans. These criteria are consistent with the criteria of AD and MCI in ADNI.

2.2. Image pre-processing

All individual T1-weighted images were pre-processed with SPM8 (http://www.fil.ion.ucl.ac.uk/spm) running under a Linux MATLAB 7.10 platform (The MathWorks, Natick, MA, USA). Before tissue

Table 1 Demographic of the dataset.

<table>
<thead>
<tr>
<th></th>
<th>AD</th>
<th>MCI</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>131</td>
<td>261</td>
<td>188</td>
</tr>
<tr>
<td>Female/male</td>
<td>63/68</td>
<td>86/174</td>
<td>96/92</td>
</tr>
<tr>
<td>Age (Mean±SD)</td>
<td>78.53±4.91</td>
<td>77.92±4.97</td>
<td>76.51±4.58</td>
</tr>
<tr>
<td>*MMSE (Mean±SD)</td>
<td>23.24±2.08</td>
<td>26.98±1.81</td>
<td>29.16±0.96</td>
</tr>
<tr>
<td>*CDR</td>
<td>0.5, 1.0</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>Memory box score</td>
<td>1.0</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

*MMSE: Mini-Mental State Examination (Scores 0–30) AD(20–26); MCLNC(24–30).
*CDR: Clinical Dementia Rating (0.5 — very mind, 1 — mild, 2 — moderate, 3 — severe).
segmentation, the images were re-oriented such that the origin of their coordinate systems was manually set to be close to the anterior commissure. These images were then segmented into grey matter (GM) and white matter (WM) maps in the "realigned" space using the "new segment" toolbox with default settings. "New segmentation" is an extension of the conventional "unified segmentation" (Ashburner and Friston, 2005). It provides more accurate tissue segmentation and deals with voxels outside the brain by including extra tissue probability maps (skull, scalp, etc.). The realigned GM and WM maps were also converted to 1.5 mm isotropic. Next, the inter-subject alignment was refined using the Dartel toolbox (Ashburner, 2007). Briefly, this involves creating averages of the GM and WM images of all subjects, which is used as an initial template. Images were then warped into closer alignment with this template, and a new template constructed by averaging the warped images (Ashburner and Friston, 2009). This procedure was repeated for a fixed number of iterations, using all the default settings. The GM images were warped into alignment with the average-shaped template using the estimated deformations. Total GM volume was preserved by multiplying each voxel of the warped GM with the Jacobian determinant of the deformation field (a.k.a. "modulation"). No smoothing was applied as we found no significant improvement of classification accuracy using whole brain GM with smoothing. This finding also confirmed the previous study (Kloppel et al., 2008). To avoid possible edge effects between tissue boundaries, we excluded voxels in the GM population template that had values less than 0.2. These masked, warped and modulated images, which have 299,477 voxels in each image, served as the input features to the following feature selection and classification procedures.

### 2.3. Support vector machine (SVM)

The support vector machine (SVM) is one of the more popular supervised learning algorithms that have been applied to neuroimaging data in recent years (Craddock et al., 2009; Fan et al., 2005; Kloppel et al., 2008; Mourao-Miranda et al., 2005). SVM demonstrates good classification performance, and is computationally efficient for training with high dimensional data. SVM is also known as the maximum margin classifier (Cristianini and Shawe-Taylor, 2000). The framework of SVM was motivated by statistical learning theory (Vapnik, 1995), in which the decision boundary is selected to maximize the separation between two classes rather than minimizing the empirical loss. It is a sparse method, which means the decision boundary may be determined by a subset of training samples, called the support vectors (SV). Intuitively, SVM ignores trivial training samples, which are far away from the opposite class, and the SV are the samples which are closest to the decision boundaries (for hard-margin SVM). The equations of SVM can be found in many literatures including Wikipedida, therefore we do not repeat in this manuscript. In this study, we used both the hard-margin SVM and soft-margin SVM implementation in LIBSVM (Chang and Lin, 2001). The difference between soft-margin and hard-margin SVM is that soft-margin SVM penalizes the training error, and the optimization in the primal form tries to maximize the margin and minimize the training errors. The free parameter C controls the weighting of the errors, it can also be realized as the inverse of regularization (smaller C, more regularization, vice versa). In the dual form, the objective functions in both soft-margin and hard-margin are the same, except C is the upper bound of the Lagrange multipliers for soft-margin SVM. In other words, if the C in soft-margin SVM is larger or equal to the largest Lagrange multiplier in the hard-margin SVM, both formulations would yield the same solution. This is an important issue, as some studies that claimed using a soft-margin SVM, however, the C values may be large enough to result a hard-margin SVM training. The effect of C value also depends on the number of features and the scale of the features. When more features are used or when the scales of the features increase, the C value would need to decrease to achieve the same regularization. In Table 2, we show the C values used in this study and the corresponding number of voxels. The C values are all powers of 2, and the power of the maximum C value (C0) for each corresponding number of voxels was defined by rounding down the log of base two of the maximum Lagrange multiplier from training AD vs. NC using the ROI with the corresponding voxels with the whole dataset. The subsequent C value is computed by dividing the previous C value. (i.e. C1 = C0/4, C2 = C1/4, etc.)

SVM is a kernel algorithm that means the input of the algorithm is a kernel rather than the original input features. For an input matrix, $X = [x_1, x_2, ..., x_n]^T$, each row of $X$ is one sample with $d$ features, and the linear kernel matrix is defined by $K = XX^T$. Alternatively, we can also formulate the kernel as a linear combination of sub-kernels, $K = K_1 + K_2 + ... + K_d$. Each sub-kernel is computed from each feature (voxel) individually. In this perspective, feature selection can be viewed as finding the optimal combination of kernels or a binary version of multi-kernel learning (MKL).

### 2.4. Feature selection methods

The goal of feature selection is to select a subset of features (voxels) from the original input features. We applied four different feature selection methods. One is based on using prior knowledge about regions of brain atrophy found in previous studies, and two methods are purely data-driven approaches. We also created a hybrid method, which combined the known parcellation of anatomical regions and the data-driven approach. To compare the feature selection methods, we selected equal numbers of features in all methods with different level of selected features, except the hybrid method. The methods are described in the following:

1. **Atlas-based region of interest (ROI)**. The ROIs are defined by the LONI Probabilistic Brain Atlas (LPBA40) (http://www.loni.ucla.edu/Atlases/LPBA40) (Shattuck et al., 2008), which were generated from anatomic regions of MRI delineated manually from 40 subjects. The delineated structures were all transformed into the

### Table 2

<table>
<thead>
<tr>
<th>Number of voxels</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole brain</td>
<td>2−12</td>
<td>2−11</td>
<td>2−10</td>
<td>2−9</td>
<td>2−8</td>
<td>2−7</td>
</tr>
<tr>
<td>82,260</td>
<td>2−9</td>
<td>2−8</td>
<td>2−7</td>
<td>2−6</td>
<td>2−5</td>
<td>2−4</td>
</tr>
<tr>
<td>71,229</td>
<td>2−9</td>
<td>2−8</td>
<td>2−7</td>
<td>2−6</td>
<td>2−5</td>
<td>2−5</td>
</tr>
<tr>
<td>25,355</td>
<td>2−7</td>
<td>2−6</td>
<td>2−5</td>
<td>2−4</td>
<td>2−3</td>
<td>2−3</td>
</tr>
<tr>
<td>12,754</td>
<td>2−7</td>
<td>2−6</td>
<td>2−5</td>
<td>2−4</td>
<td>2−3</td>
<td>2−3</td>
</tr>
<tr>
<td>11,031</td>
<td>2−6</td>
<td>2−5</td>
<td>2−4</td>
<td>2−3</td>
<td>2−3</td>
<td>2−3</td>
</tr>
<tr>
<td>9,542</td>
<td>2−6</td>
<td>2−5</td>
<td>2−4</td>
<td>2−3</td>
<td>2−3</td>
<td>2−3</td>
</tr>
<tr>
<td>7,522</td>
<td>2−5</td>
<td>2−4</td>
<td>2−3</td>
<td>2−2</td>
<td>2−2</td>
<td>2−2</td>
</tr>
<tr>
<td>6,463</td>
<td>2−4</td>
<td>2−3</td>
<td>2−2</td>
<td>2−2</td>
<td>2−2</td>
<td>2−2</td>
</tr>
<tr>
<td>4,568</td>
<td>2−3</td>
<td>2−2</td>
<td>2−2</td>
<td>2−2</td>
<td>2−2</td>
<td>2−2</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>ROI Full name</th>
<th>Abbreviation</th>
<th>Number of voxels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cingulate gyrus</td>
<td>CG</td>
<td>12,754</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>HC</td>
<td>4568</td>
</tr>
<tr>
<td>Parahippocampal gyrus</td>
<td>PHG</td>
<td>6463</td>
</tr>
<tr>
<td>Fusiform gyrus</td>
<td>FG</td>
<td>9542</td>
</tr>
<tr>
<td>Precuneus</td>
<td>PCu</td>
<td>7522</td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>MTG</td>
<td>25,355</td>
</tr>
<tr>
<td>Temporal lobe</td>
<td>TL</td>
<td>71,229</td>
</tr>
<tr>
<td>Hippocampus + parahippocampal gyrus</td>
<td>HC + PHG</td>
<td>11,031</td>
</tr>
<tr>
<td>Temporal lobe + hippocampus + parahippocampal gyrus</td>
<td>TL + HC + PHG</td>
<td>82,260</td>
</tr>
<tr>
<td>Whole brain</td>
<td>All</td>
<td>299,477</td>
</tr>
</tbody>
</table>
atlas space. We used the SPM5 space, which is the template space in SPM5. There are a total of 56 structures in the atlas. Most regions are left/right pairs e.g. left cuneus, right cuneus, except for cerebellum and brainstem. To generate the parcellation, we converted the LPBA40 atlas into an image of discrete regions, where each voxel was assigned a label corresponding to the most probable of the 56 structures. We combined both left and right structures into one, hence reducing the total number of ROIs to 27. Because the data were spatially normalized into population average space (details see Image pre-processing), the maximum likelihood map in LPBA40 were additionally affine-registered to the population average, and resampled using nearest neighbor interpolation. Based on findings from previous studies (Bastos Leite et al., 2004; Chetelat et al., 2002; Frisoni et al., 2002; Pennanen et al., 2005), we selected the nine ROIs listed in Table 3 (also see Fig. 1). Using knowledge from the literatures, we would expect hippocampus and parahippocampal gyrus to achieve the best performance. Note that hippocampus defined in LPBA40 includes amygdale.

2. Univariate t-test filtering (t-test). This is one of the most commonly used approaches. It is equivalent to performing a simple mass-univariate analysis (statistical parametric map) of the data, and thresholding at some value to obtain a mask. A two-sample t-test at each voxel of the training set was performed. Because training should not be informed by the labels of the test set in any way, the test set was excluded from this mass-univariate analysis. We then determined various thresholds for the absolute t-value.

---

**Fig. 1.** Overlay of the selected ROIs from the affine-registered maximum likelihood map in LPBA40 onto the population grey matter average.

**Fig. 2.** Classification accuracies vs. sample size for classifying AD & NC and classifying MCI & NC using the whole brains. We also fitted a Von Bertalanffy growth function, \( f(x) = a \times (1 - \exp(-b \times x)) \).
based on the number of voxels we intended to select at each level (shown in Table 3). In other words, a smaller set (higher threshold) was always a subset of a larger set (lower threshold). It can also be seen as a combination of kernels computed from voxels at different range of absolute t-values $K = K_{a1} + K_{a2} + \ldots + K_{an}$. As the threshold lowering down, more kernels were added.

3. Averaged absolute t-value in ROIs (t-test + ROI). The method combines the standard univariate t-test filtering with predefined ROIs from the atlas. We first performed mass-univariate t-tests on the training samples, from which we computed the averaged absolute t-value in the 27 ROIs defined by LPBA40 (combining left and right structures). The ROIs were ranked according to their averaged absolute t-value. We selected the top 1, 2, 3, 4, and 5 ROIs, and there were 5 levels of inclusion, i.e., top one means using the 1st ranked ROI, top 2 means the combination of the 1st and 2nd ranked ROI, and top 3 means the combination from the 1st to the 3rd ranked ROI, etc. This was the only method that did not have a fixed number of voxel at each level of selection, because different ROIs have different numbers of voxels. For this, we only varied the “number of combined ROIs”.

4. Recursive feature elimination (RFE). Recursive feature elimination is an iterative feature selection algorithm developed specifically for SVM (Guyon et al., 2002). It is also a popular feature selection method for fMRI data (Craddock et al., 2008; De Martino et al., 2009; Marquand et al., 2010). The method starts by selecting the most relevant features, which would potentially yield the largest margin in classification. For a linear kernel, it means the feature with the lowest absolute feature weight. An approximation is usually made for high-dimensional data, whereby a small set of features is removed at each iteration. In our implementation, we removed 3000 voxels (around 1%) at each iteration, sometimes adjusting this number to obtain the same number of features as used by the other methods (i.e., the number of voxels in Table 3).

2.5. Cross-validation and resampling

To estimate the classification accuracies with different training size, we first selected a subset of subjects from the whole dataset. To reduce the gender confound, we balanced the number of males and females in both control and patient groups, such that both groups had equal proportions of males and females. To reduce the age confound, we further matched the age in each gender group (e.g., we matched the age of male control and male patients, etc.). We not only matched the mean of the age, we also tried to match the distribution in both populations. To achieve this, we first randomly selected a subset from the group with fewer samples (e.g., fewer samples in female NC than female MCI). We then estimated the distribution of age by smoothing the histogram (bin size 1 year) with Gaussian smoothing with 3 year FWHM. The new samples were then drawn from the group with more samples using a variation of rejection sampling (Bishop, 2006).

After the subset was selected, we applied leave-one-out cross-validation (loocv) to the subset for the hard-margin SVM. In each cross-validation trial, one sample was left out, and we applied feature

Fig. 3. Classification accuracies obtained from data-driven feature selection. The dark line is the accuracy obtained using the whole brain (no feature selection). Neither t-test nor RFE achieved significantly better classification than the whole brain. The table of $p$-values can be found in the Supplementary material. For t-tests, the accuracies were calculated from averaging 200 resamples, and for RFE, the accuracies were calculated from averaging 50 resamples. To be compatible, we also averaged different numbers of resamples for the whole brain. That is why the lines of whole brain classification (black lines) are different between (a) and (b), and are different between (c) and (d).
Fig. 4. Graph on the left shows classification accuracies obtained using different ROIs for classifying AD and NC with different sample sizes (resample 200 times). The dark line is the accuracy obtained by using the whole brain (no feature selection). Table on the right shows results from paired t-tests of classification accuracies between each ROI and whole brain (no feature selection). 1 indicates the ROI is significantly better than whole brain, 0 indicates not significantly better, and −1 indicates whole brain is significantly better than that ROI. The rows are different sample sizes and the columns are different ROIs. The tables of the original p-values are in the Supplementary material.

Fig. 5. Graph on the left shows classification accuracies obtained using different ROIs for classifying AD and NC with different sample sizes (resample 200 times). The dark line is the accuracy obtained by using the whole brain (no feature selection). Table on the right shows results from paired t-tests of classification accuracies between each ROI and whole brain (no feature selection). 1 indicates the ROI is significantly better than whole brain, 0 indicates not significantly better, and −1 indicates whole brain is significantly better than that ROI. The rows are different sample sizes and the columns are different ROIs. The tables of the original p-values are in the Supplementary material.
size of the subset. For classification between AD and NC, the minimum size of the subset was 14 (seven in each group, three female, and four male), and the maximum size of the subset was 262. The increment between each size-varied subsets was eight subjects (two male AD, two female AD, two male NC, and two female NC). However, different feature selection methods had different computational costs. For methods that required longer computation, we omitted some population sizes. For discrimination between MCI and NC, the subsets ranged from 20 to 356 subjects, with increments of eight. Again, the increments varied according to the computationally intensive nature of the feature selection approaches.

We have also compared the difference between loocv and 10 fold cross-validation. When we averaged more than 50 resamples, the results were not significantly different between 10 fold cross-validation and loocv, especially when the sample size was large. Details are in the Supplementary material. Because 10 fold cross-validation and loocv have minimum difference with large resamples, we applied 10 fold cross-validation to the soft-margin SVM to accelerate the computation.

2.6. Statistical analysis

One of the objectives of this study is to examine whether feature selection could improve the classification accuracies. Because we used the same resampled sets for each sample size across different feature selection methods, we could compare the accuracies of loocv, with and without the use of feature selection, using paired t-test.

3. Results

Regardless of feature selection method, larger sample sizes yielded better performance. For classifying AD and NC, whole brain without feature selection achieved an average accuracy of 84.3% at

Fig. 6. Classification accuracies and the results of paired t-tests of the hybrid method (t-test + ROI) for classifying AD & NC. The dark line is the accuracy obtained using whole brain (no feature selection). ROI – 1 means the top ROI, and ROI – 2 means the combination of the top two ROIs, etc. The rows in the table are for different sample sizes, and the columns are different numbers of top ranked ROIs combined.

Fig. 7. Classification accuracies and the results of paired t-tests for the hybrid method (t-test + ROI) for classifying MCI & NC. The dark line is the accuracy obtained using whole brain (no feature selection). ROI – 1 means the top ROI, and ROI – 2 means the combination of the top two ROIs, etc. The rows in the table are for different sample sizes, and the columns are different numbers of top ranked ROIs combined.
the maximum training size of 262, with no sign of reaching a plateau. We used Von Bertalanffy growth function, \( f(x) = a \times (1 - \exp(-b \times x)) \), to fit the data points of accuracies vs. sample size and found out that the actual accuracies were higher with larger sample size than the fitted curve (see Fig. 2).

For classifying MCI and NC, whole brain without feature selection achieved an average accuracy of 67.3% at the maximum training size of 356. We also fitted the accuracies curve with Von Bertalanffy function, and found similar results as classifying AD and NC that the actual accuracies were higher than the fitted curve at large training size (see Fig. 2).

None of the data-driven feature selection methods (t-test filtering and RFE) achieved significantly superior performance than whole brain (no feature selection) for classifying MCI and NC or classifying AD and NC (see Fig. 3). On the other hand, the atlas based ROI method showed significantly (\( p < 0.001 \)) better accuracies than whole brain (no feature selection) using parahippocampus (PHG) ROI or parahippocampus + hippocampus (PHG + HC) ROIs for classifying MCI and NC and classifying AD and NC. On average, the use of PHG + HC was better than PHG alone (see Fig. 4). However, for classifying AD and NC, HC was also significantly better than whole brain when the training size was less than or equal to 166. The results were also similar for classifying AD and NC that the actual accuracies were only significantly better when the sample size was over 260, and 5 levels of inclusion was employed (see Fig. 7). The tables of the original p-values are in the Supplementary material.

We also used soft-margin SVM with different levels of regularization (i.e. different C) with feature selection. In general, lowering the C (i.e. increasing regularization) did not improve the accuracies for most ROIs, except those that already yielded better results than whole brain e.g. HC, PHG, and HC + PHG (see Fig. 8). The improvement of accuracies is most prominent in MCI vs. NC with larger sample size. HC became significantly better than whole brain using the optimal C value. For t-test and RFE, the best accuracies with the optimal C were still not significantly better than whole brain using hard-margin SVM (see Fig. 9).

The ROIs that yielded better classification accuracy than whole brain for classifying MCI converters vs. MCI non-converters are similar to the result of AD vs. NC, except that FG was significantly better.

![Fig. 8. Classification accuracies obtained using different ROIs for classifying AD & NC and MCI & NC with different sample sizes (resample 200 times) by soft-margin SVM. The dark line is the accuracy obtained by using the whole brain (no feature selection). C* indicates large C value that is identical to a hard-margin SVM. The exact C value for the corresponding feature size can be found in Table 2.](image-url)
than whole brain when the sample sizes are small (see Fig. 10). Overall, the classification accuracies were very low.

In addition, we compared the features selected from two datadriven techniques (t-test and RFE). The features selected using the two methods were very different. For illustrative purpose, we generated the frequency maps of features selected in the re-resamples from classifying AD and NC with 11,031 features for both methods (see Fig. 11). The correlation between the frequency maps is corr(t-test, RFE) = 0.31. T-tests selected the features consistently in each resampling trials, and most features were in the HC, PHG, and TL. RFE resulted in a more distributed selection, with the selected features scattered across the whole brain. However, features in HC were also consistently selected.

4. Discussion

Generally speaking, the classification accuracies were higher with larger sample size than the estimates from fitting a Von Bertalanffy function. This implies that the improvement of classification accuracies due to increasing samples plateaus slower than the exponential growth function. If the accuracies keep following the projection of the linear fit, prediction accuracies of 90% for separating AD and NC may be achieved using around 500 training samples. However, the main focus in this study was comparing the performance, with and without feature selection. From the results, we found that appropriate methods of feature selection improved the classification accuracies regardless of the sample size. However, data-drive methods without prior knowledge either did not improve the accuracies or made them worse. We were surprised to see such results, and speculate that it was because there was no additional information in the training samples from which those feature selection methods can extract and SVM cannot, especially when features were highly correlated. In our dataset, there may be hardly any non-informative voxels, but only less-informative voxels. The information encoded in the multivariate patterns also increased the difficulty of ranking features. Nevertheless, data driven approaches could still reduce the less-informative features, and maintain the most-informative voxels. Nevertheless, data driven approaches could still reduce the less-informative features, and maintain the most-informative voxels without any prior knowledge. This may be useful for people who rely on feature-weight map to make interpretation because feature selection reduces the chance of interpreting unreliable voxels in the weight-map.

In general, as the training size increased, the difference between the accuracies, with and without feature selection, was reduced. This implies that any method of feature selection would not improve the accuracies when the training size is sufficient. In the case of classifying AD and NC, because of higher contrasts between the two classes, the required samples are small. On the other hand, the contrast between MCI and NC is weaker and the patterns are less homogenous (i.e. larger effective dimensionality), thus more samples are required before whole brain can achieve the same accuracies as specific ROIs. In this study, even 356 samples were not sufficient.

It seemed that making use of prior knowledge resulted in the best accuracies. However, this method depends on the reliability of the information used. For example, although functional studies...
(Ikonomovic et al., 2011) show a strong difference between patterns of brain activity in AD and NC in precuneus, the use of the precuneus ROI (GM density) resulted in very poor classification accuracies. Also, the most well known marker of brain atrophy in AD is HC, but our results showed that combining HC and PHG resulted in higher accuracies than either of them alone. This implies that covariance between information encoded in the ROIs may help classification. This effect makes selecting the optimal combination of ROIs difficult, and we can only claim that the selected ROIs are sub-optimal. Perhaps the best method of feature selection is cross-validation with pre-defined ROIs. From our results, we found that the ranks of the classification accuracies of each ROI were quite consistent across different sample sizes, especially for classifying AD and NC. Although we only used a small validation set, we should still have a high chance of identifying the best ROI through cross-validation. Alternatively, we can also use the hybrid method (t-test + ROI) that uses the ROI as spatially constrain and t-test as the ranking of features. The hybrid method did show significant, but small, improvements of classification accuracy in NC vs. AD and NC vs. MCI (Figs. 7 and 8). However, the selected ROIs in each resample were not always consistent, although hippocampus ROI was often ranked the top.

The choice of hard-margin or soft-margin SVM did not affect our conclusion, as methods of feature selection and ROIs that did not yield better solutions still not perform better after optimizing the soft-margin SVM. In practice, people often use three-way-split cross-validation to estimate the test accuracies by determining the C with the cross-validation in the inner loops. This procedure would greatly increase the computation. We took a pragmatic approach by presenting the accuracies with all the C values we tested. The best accuracies obtained from one of the C should be the optimistic measure i.e. higher than the testing accuracy of three-way-split CV. Because the optimistic accuracies, which should be more conservative in our comparison, did not change our conclusions, we did not use the time consuming three-way-split CV.
In this study, we applied feature selection as a separate procedure to the classifier. There are also classifiers with a built-in capability of feature selection during the training. These include lasso regression (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), automatic relevance determination (Chu et al., 2010; MacKay, 1995; Tipping, 2001), and other sparse models (Carroll et al., 2009; Grosenick et al., 2008; Ryali et al., 2010; Yamashita et al., 2008). However, those methods would require a lot of memory to select features from 300,000 voxels, which is often impractical. In practice, efficient methods of feature selection are often applied to eliminate the majority of less-informative features, before applying the classifiers that can select features during training.

In conclusion, our results confirmed the previous study (Cuingnet et al., 2010) that features selection did not improve the performance if there was no prior knowledge. That is to say, if the right methods and the right prior knowledge are used, feature selection does improve classification performance. Although, we demonstrated the effect using anatomical MRI, our experience in Pittsburgh brain activity interpretation competition (PBAIC 2007 (Chu et al., 2011) also led to the same conclusion. Teams using data driven feature selection to select fMRI features did not perform well in the competition than teams applied prior knowledge in feature selection.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, NIMH, National Science Council of Taiwan (NSC 100-2628-E-010-002-MY3, NSC 100-2627-B-010-004-) and National Health Research Institute (NHRI-EX100-9813EC). JA is funded by the Welcome Trust. This study utilized the high performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG016573). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contributions from the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, 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 Foundation for the National Institutes of Health. Other ADNI contributing institutions include: Arizona State University, Atlanta, GA; Christian Albrechts University, Kiel, Germany; Cleveland Clinic, Cleveland, OH; Columbia University, New York, NY; Dartmouth College, Hanover, NH; Emory University, Atlanta, GA; Fisher Center for Alzheimer’s Research Foundation, New York, NY; Florida Institute of Technology, Melbourne, FL; Foundation for Instituto de Investigaciones Biomédicas ‘Augusto Roa Bastos’ (FIB), La Paz, Bolivia; Geisinger Health System, Danville, PA; Georgetown University, Washington, DC; Grandview University, Des Moines, IA; Harvard Medical School, Boston, MA; Hiroshima University, Hiroshima, Japan; Humboldt University, Berlin, Germany; Indiana University, Indianapolis, IN; Institute for Neuro-Imaging and Informatics, West Virginia University, Morgantown, WV; Jack and Judy Barrickman Alzheimer’s Research Center, Ann Arbor, MI; John Hopkins University, Baltimore, MD; Kaiser Permanente Southern California, Los Angeles, CA; King’s College, London, UK; Kure University, Japan; Leyden University, Leiden, The Netherlands; Max Planck Institute of Psychiatry, Munich, Germany; Mayo Clinic, Rochester, MN; Michigan State University, East Lansing, MI; Mind motion GmbH, Munich, Germany; MGH and Harvard Medical School, Boston, MA; National Institute of Environmental Health Sciences, Research Triangle Park, NC; National Institute of Aging and National Institute on Deafness and Other Communication Disorders, National Institute of Health, Rockville, MD; National Institute of Mental Health, Bethesda, MD; Neurologica, Inc., Cambridge, MA; Neurologica, Inc., Melbourne, FL; Neurology, Inc., Malvern, PA; Northwestern University, Chicago, IL; Oregon Health and Science University, Portland, OR; Oxford University, Oxford, UK; Pennsylvania State University, University Park, PA; Pacific University, Hillsboro, OR; Pennsylvania University, Philadelphia, PA; Pelotas, Brazil; Paris Brain Institute, Paris, France; Penn State University, Hershey, PA; Philadelphia Neuroimaging Center, Philadelphia, PA; University of California, San Diego. ADNI data are disseminated by the Foundation for the National Institutes of Health. For more information please visit www.adni-info.org.

In conclusion, our results confirm the previous study (Cuingnet et al., 2010) that features selection did not improve the performance if there was no prior knowledge. That is to say, if the right methods and the right prior knowledge are used, feature selection does improve classification performance. Although, we demonstrated the effect using anatomical MRI, our experience in Pittsburgh brain activity interpretation competition (PBAIC 2007 (Chu et al., 2011) also led to the same conclusion. Teams using data driven feature selection to select fMRI features did not perform well in the competition than teams applied prior knowledge in feature selection.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, NIMH, National Science Council of Taiwan (NSC 100-2628-E-010-002-MY3, NSC 100-2627-B-010-004-) and National Health Research Institute (NHRI-EX100-9813EC). JA is funded by the Welcome Trust. This study utilized the high performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG016573). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co.,Medpace, Inc.,Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffman-La Roche, Schering-Plough,Synarc, Inc., as well as non-profit partners the Alzheimer’s Association and Alzheimer’s Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI 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, 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 California, Los Angeles.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2011.11.066.

References


