Classification of Structural Images via High-Dimensional Image Warping, Robust Feature Extraction, and SVM

Yong Fan, Dinggang Shen, and Christos Davatzikos

Section of Biomedical Image Analysis, Department of Radiology,
University of Pennsylvania, Philadelphia, PA 19104
{yong.fan, dinggang.shen, christos.davatzikos}@uphs.upenn.edu

Abstract. This paper presents a method for classification of medical images, using machine learning and deformation-based morphometry. A morphological representation of the anatomy of interest is first obtained using high-dimensional template warping, from which regions that display strong correlations between morphological measurements and the classification (clinical) variable are extracted using a watershed segmentation, taking into account the regional smoothness of the correlation map which is estimated by a cross-validation strategy in order to achieve robustness to outliers. A Support Vector Machine-Recursive Feature Elimination (SVM-RFE) technique is then used to rank computed features from the extracted regions, according to their effect on the leave-one-out error bound. Finally, SVM classification is applied using the best set of features, and it is tested using leave-one-out. The results from a group of 61 brain images of female normal controls and schizophrenia patients demonstrate not only high classification accuracy (91.8%) and steep ROC curves, but also exceptional stability with respect to the number of selected features and the SVM kernel size.

1 Introduction

Morphological analysis of medical images is performed commonly in a variety of research and clinical studies. Region of Interest volumetry (ROI) has been traditionally used to obtain regional measurement of anatomical volumes and investigate abnormal tissue structures with disease [1]. However, in practice, *a priori* knowledge about abnormal regions is not always available. Even when good *a priori* hypotheses can be made about specific ROIs, a region of abnormality might be part of an ROI, or span multiple ROIs, thereby potentially reducing statistical power of the underlying morphological analysis significantly. These limitations can be effectively overcome by methods falling under the general umbrella of High-Dimensional Morphological Analysis (we will refer to these methods as HDMA), such as voxel-based and deformation-based morphometric analysis methods, e.g. [2-5]. However, a voxel-by-voxel analysis is limited by noise, registration errors, and excessive inter-individual variability of measurements that are too localized, such as voxel-wise displacement fields, Jacobians, or residuals. Most importantly, voxel-by-voxel mass-univariate analysis of transformations or residuals does not capture multi-variate relationships in the data.

J. Duncan and G. Gerig (Eds.): MICCAI 2005, LNCS 3749, pp. 1 – 8, 2005.

Linear methods, such as PCA [6] are not effective in capturing complex relationships in high-dimensional spaces.

In order to overcome these limitations, pattern classification methods have begun to emerge in the recent years in the field of computational anatomy [7-11], aiming at capturing nonlinear multivariate relationships among many anatomical regions, to more effectively characterize group differences. A major challenge in these methods has been the sheer dimensionality of HDMA-related measurements, which is often in the millions, coupled with the relatively small number of training samples, which is at best in the hundreds, and often just a few dozens. Accordingly, extracting a small number of most informative features from the data has been a fundamental challenge. A main emphasis of this paper is the extraction of distinctive, but also robust features from high-dimensional morphological measurements obtained from brain MR images, which are used in conjunction with nonlinear support vector machines (SVM) for classification. The proposed method is tested on classifying normal controls from schizophrenia patients in female participants.

The key elements of the proposed approach are: 1) Regional volumetric information is first extracted from a template warping transformation; herein we focus entirely on volumetric information, such as atrophy, and don't consider higher order shape characteristics. 2) The anatomy of interest (the brain, herein) is partitioned into a number of regions, via a watershed algorithm applied to the correlation map between clinical status and regional volumetric measurements; various techniques are applied to estimate the correlation map, in order to achieve robustness to outliers. 3) An SVM-RFE technique is applied to the previously rank-ordered features that are computed from the extracted regions, in order to select the most important feature set for classification. 4) A nonlinear SVM classifier is applied and tested via cross-validation and ROC analysis. 5) Group differences are visually displayed via a *discriminative direction method* [7,11]. We now detail the methodology.

2 Methods

Our classification method involves three steps: feature extraction, feature selection, and nonlinear classification, which are detailed next.

2.1 Feature Extraction

As mentioned in the Introduction, the features used for brain classification are extracted from automatically generated regions, which are determined from the training data. Several issues are taken into consideration here. First, morphological changes of brain structures resulting from pathological processes usually don't occur in isolated spots, but rather they occur in regions that can have irregular shapes and are not known *a priori*. Second, noise, registration errors and inter-individual anatomical variations necessitate the collection of morphological information from regions much larger than the voxel size, which must additionally be distinctive of the pathology of interest. Third, multivariate classification methods are most effective and generalizable when applied to a small number of reliable and discriminative features. Accordingly, features irrelevant to classification must be eliminated.

In the following, we detail the procedure for automatically generating adaptive regions from a training dataset, by first introducing the method to extract local morphological features, then defining the criteria for adaptively clustering voxels into regions, and finally extracting overall features from each region.

Construction of a morphological profile of the brain. In order to obtain the morphological profile from an individual brain image, warping the image into a template space is often a first step, leading to various morphological measurements (e.g. deformation field, Jacobian determinant, tissue density maps, or residuals) that are in the same space and therefore directly comparable across individuals. Herein we follow the framework that was proposed in [5], and which is based on a mass-preserving shape transformation framework; a similar method is used within the SPM software and is often referred to as "Jacobian modulation". We have used this approach because it is robust to registration errors, due to the mass preservation principle, in contrast to the determinant of the Jacobian that is directly affected by registration errors.

The approach in [5] uses images that are first segmented into three tissues, namely gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF)--we used the segmentation method of [12] in our experiments. Then, by using a high-dimensional image warping method [13] these segmented images are spatially transformed into a template space, by preserving the total tissue mass; this is achieved by increasing the respective density when a region is compressed, and vice versa. As a result, three tissue density maps, $f_0(u)$, $f_1(u)$, $f_2(u)$, are generated in the template space, each reflecting local volumetric measurements corresponding to GM, WM, and CSF, respectively, at location u. These three tissue density maps give a quantitative representation of the spatial tissue distribution. Regional atrophy is reflected by reduction in the respective tissue density map. Representative tissue density maps are shown in Fig.1.

Learning-based generation of adaptive regions. Brain regions are generated by spatial clustering of morphological features of similar classification power. For each morphological feature, its classification power is highly related to its discriminative power and reliability. The discriminative power of a feature can be quantitatively measured by its *relevance* to classification as well as its *generalizability* for classification. The relevance of a feature to classification can be measured by the correlation between this feature and the corresponding class label in a training dataset (e.g. normal or pathologic). In machine learning and statistics for relevance analysis, the correlation measures can be broadly divided into linear correlation and non-linear correlation. Most non-linear correlation measures are based on the information-theoretical concept of entropy, such as mutual information, computed by probability estimation. For continuous features, probability density estimation is a hard task especially when the number of available samples is limited. On the other hand, linear correlation measures are easier to compute even for continuous features and are robust to overfitting, thus they are widely used for feature selection in machine learning. Here, we used the absolute Pearson correlation coefficient, closely related to T-test [14] in the context of extracting group differences, to measure the relevance of each feature to classification. Given an image location u, the Pearson correlation between a feature $f_i(u)$ of tissue i, and class label y, is defined as

$$\rho_i(u) = \frac{\sum_j \left(f_{ij}(u) - \overline{f_i(u)}\right)\left(y_j - \bar{y}\right)}{\sqrt{\sum_j \left(f_{ij}(u) - \overline{f_i(u)}\right)^2 \sum_j \left(y_j - \bar{y}\right)^2}}, \tag{1}$$

where j denotes the j th sample in the training dataset. Thus, $f_{ij}(u)$ is a morphological feature of tissue i in the location u of j th sample (the tissue density map, here), and $\overline{f_i(u)}$ is the mean of $f_{ij}(u)$ over all samples. Similarly, y_j is a class label of the j th sample, and $\bar{y}$ is the mean of y_j over all samples. In addition to the relevance, the *generalizability* of a feature is equally important for classification. A bagging strategy [15] is adopted to take the generalization ability into account, when measuring the discriminative power of a feature by absolute Pearson score. That is, given n training samples, a leave-one-out procedure is used to measure the discriminative power of each feature $f_i(u)$ by a conservative principle that selects the minimal absolute Pearson score from n resulted scores as this feature's discriminative power, defined as

$$P_i(u) = \min_{1 \le k \le n} \left|\rho_{ik}(u)\right|, \tag{2}$$

where $\rho_{ik}(u)$ is the k th leave-one-out Pearson correlation at location u of tissue map i. Maximizing the measure in (2) would select features that maximize the margin from 0, i.e., from no discrimination at all.

The *spatial consistency* of a feature is another important issue in classification, since morphological features are locally extracted and thus might not be reliable due to registration errors and inter-individual anatomical variations. A feature is spatially consistent if it is similar to other features in its spatial neighborhood, implying that small registration errors will not significantly change the value of this feature. For each feature $f_i(u)$, we measure its spatial consistency, $R_i(u)$, by an intra-class correlation coefficient that is computed from all features in its spatial neighborhood and all samples in the training dataset [16]. In our applications, the value of $R_i(u)$ is constrained to lie between 0 and 1.

For each feature, its discriminative power score $P_i(u)$ and its spatial consistency score $R_i(u)$ are both non-negative, with high score indicating better feature for classification. We combine these two measurements into one by the following equation,

$$s_i(u) = P_i(u)^p R_i(u)^r, \quad p, r > 0, \tag{3}$$

thus obtaining a single score $s_i(u)$ for each feature $f_i(u)$, which reflects the classification power of this feature for the particular classification problem. Three score maps are produced for GM, WM and CSF, respectively.

As we mentioned above, the disease-affected brain regions generally occur in the clusters of spatially contiguous voxels. Therefore, a watershed segmentation method [17] is employed to partition a brain into different regions according to the scores $s_i(u)$, and finally to obtain separate partitions for each tissue. A typical brain region partition result, with all regions generated from three tissue density maps, is shown in Fig. 1, on the right.

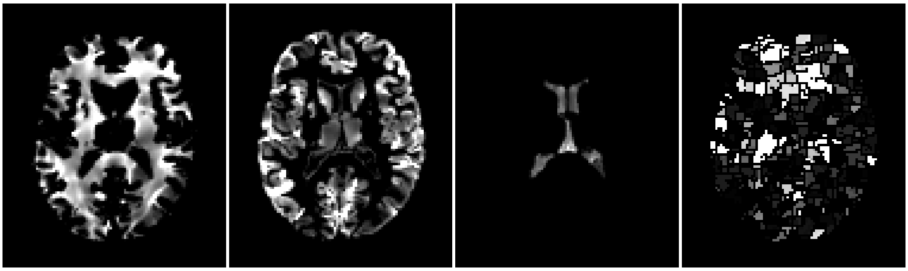

Fig. 1. Typical tissue density maps (GM, WM, CSF, from left to right) and automatically generated brain regions in which high grey-levels indicate discriminative power

Feature extraction from generated brain regions. For each region generated as described above, its corresponding volumetric measure is computed by summing up all tissue density values in this region, which effectively calculates the volumes of the corresponding regions in individual anatomies. Volumetric measures from all WM, GM, and CSF regions constitute an attribute vector to represent morphological information of the brain. Although currently we focus on local tissue volumetric information, other types of information could also be considered.

2.2 Feature Selection via SVM-RFE

Although the number of regions determined in Sec. 2.1 is dramatically smaller than the original number of brain voxels, measures obtained from many regions are less effective, irrelevant and redundant for classification. This requires a feature selection method to select a small set of the most informative features for classification. We have experimented with several feature selection methods, and determined that the SVM-RFE algorithm has the best performance. SVM-RFE is a feature subset selection method based on SVM, initially proposed for a cancer classification problem [18]. It was later extended by introducing SVM-based leave-one-out error bound criteria in [19]. The goal of SVM-RFE is to find a subset of size r among d variables ($r < d$), which optimizes the performance of the classifier. This algorithm is based on a backward sequential selection that removes one feature at a time. At each time, the removed feature makes the variation of SVM-based leave-one-out error bound smallest, compared to removing other features. In order to apply this subset selection to our problem in a reasonable time cost and to avoid local optima, we first remove the most irrelevant features by the feature ranking method [14] in which the rank score is computed by a Pearson correlation based bagging strategy as we described above, and then apply the SVM-RFE algorithm on the set of remaining features.

2.3 SVM-Based Classification

The nonlinear support vector machine is a supervised binary classification algorithm [20]. SVM constructs a maximal margin linear classifier in a high (often infinite) dimensional feature space, by mapping the original features via a kernel function. The Gaussian radial basis function kernel is used in our method.

SVM is not only empirically demonstrated to be one of the most powerful pattern classification algorithms, but also has provided many theoretic bounds on the leave-one-out error to estimate its capacity, for example, the radius/margin bound, which could be utilized in feature selection. Another reason for us to select SVM as a classifier is its inherent sample selection mechanism, i.e., only support vectors affect the decision function, which may help us find subtle differences between groups.

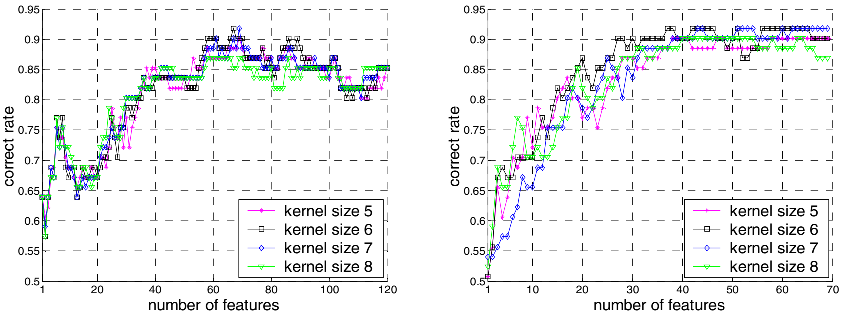

Fig. 2. Performance of ranking based feature selection (left plot) and SVM-RFE feature selection (right plot). Plotted are the average classification rates for different SVM kernel sizes and different feature numbers. Notably, the SVM-RFE algorithm starts selecting subsets of features from 69 features, which are top-ranked features, selected by ranking-based feature selection method. The SVM-RFE algorithm performs a robust selection of features and leads to stable performance.

3 Results

We tested our approach on MR T1 brain images, in order to compare the brain differences between female schizophrenia patients (N = 23) and normal controls (N = 38).

A full leave-one-out cross-validation is performed in our experiments. In each leave-one-out validation experiment, one subject was first selected as testing subject, and the remaining subjects are used for the entire adaptive regional feature extraction, feature selection and training procedure described in Section 2. Then, the classification result on the testing subject using the trained SVM classifier was compared with the ground-truth class label, to evaluate the classification performance. Absolutely *all* feature selection and training steps were cross-validated, i.e. the testing image had no influence on the construction of the classifier. By repeatedly leaving each subject out as testing subject, we obtained the average classification rate from 61 leave-one-out experiments. Finally, these experiments were repeated for different numbers of features, in order to test the stability of the results. The best average correct classification rate was 91.8% by using 37 features, selected by SVM-RFE algorithm, as shown in Fig. 2. Although a reasonably good performance was achieved just via the feature ranking method according to the scores computed by a bagging strategy [15], as shown in Fig. 2-left, more stable performance was achieved by incorporating the SVM-RFE method (Fig. 2-right), since simple feature ranking does not consider correlations between features. Furthermore, these plots also indicate that the described algorithm is quite robust with respect to the SVM Gaussian kernel. The ROC curve of the classifier that yields the best classification result is also shown in Fig. 3, which indicates that our classifier has large area under ROC curve.

Besides using a classifier to determine the abnormality of brains, we can also use it for detecting group difference. In [7,11], the *discriminative direction method* was used to estimate the group difference from the classification function. Here, we utilized a similar method to estimate the group difference. The group differences are overlaid on the template brain, highlighting the most significant and frequently detected group differences in our leave-one-out experiments (Fig. 4).

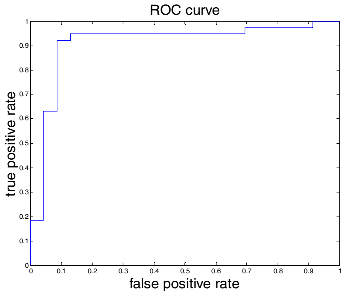

Fig. 3. ROC curve

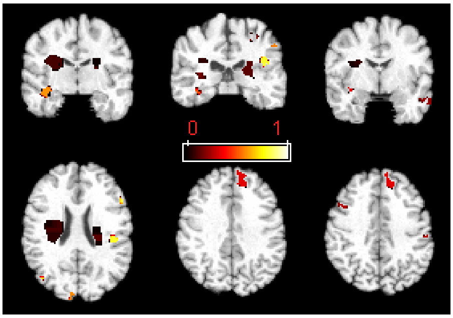

Fig. 4. Regions of most representative of the group differences, found via decision function gradient.(high light indicates more significant)

4 Discussions and Conclusions

We have presented a statistical classification method for identification of brain abnormality based on regional morphological information. The classifier is built on adaptive regional feature extraction and feature selection. In particular, brain regions are generated automatically by grouping local morphological features with similar classification power. This adaptive regional feature extraction method aims at overcoming the limitations of the traditional ROI methods that need prior knowledge of what specific regions might be affected by disease, and the limitations of the voxel based morphometry (VBM) methods that use an identical isotropic filter to collect regional morphological information in all brain locations. The robust feature selection method used in this paper further removes features that are irrelevant and redundant to classification, thus improving the classification performance. The experimental results indicate that this method can achieve high classification rate in a schizophrenic study.

References

1. Giuliania, N.R., Calhoun, V.D., Pearlson, G.D., Francisd, A., Buchanan, R.W.: Voxel-based morphometry versus region of interest: a comparisonof two methods for analyzing gray matter differences in schizophrenia.Schizophrenia Research 74 (2005) 135—147
2. Thompson, P.M.; MacDonald, D.; Mega, M.S.; Holmes, C.J.; Evans, A.; Toga, A.W.: Detection and mapping of abnormal brain structure with a probabilistic atlas of cortical. Journal of Computer Assisted Tomography 21 (1997) 567-581
3. Ashburner, J., Friston, K.J.: Voxel-based morphometry--the methods. NeuroImage 11 (2000) 805—821
4. Chung, M.K., Worsley, K.J. Paus, T., Cherif, C., Collins, D.L., Giedd, J.N., Rapoport, J.L., Evanst, A.C.: A unified statistical approach to deformation-based morphometry. NeuroImage 14 (2001) 595-606
5. Davatzikos, C., Genc, A., Xu, D., Resnick, S.M.: Voxel-based morphometry using the ravens maps: Methods and validation using simulated longitudinal atrophy. NeuroImage 14 (2001) 1361—1369
6. Miller, M.; Banerjee, A.; Christensen, G.; Joshi, S.; Khaneja, N.; Grenander, U.; Matejic, L.: Statistical methods in computational anatomy. Statistical Methods in Medical Research 6 (1997) 267-299
7. Golland, P., Grimson, W.E.L., Shenton, M.E., Kikinis, R.: Deformation analysis for shape based classification. In: IPMI. (2001) 517—530
8. Gerig, G.; Styner, M.; Lieberman, J.: Shape versus Size: Improved understanding of the morphology of brain structures. In: MICCAI (2001) 24-32
9. Yushkevich, P.A., Joshi, S.C., Pizer, S.M., Csernansky, J.G., Wang, L.E.: Feature selection for shape-based classification of biological objects. In Taylor, C.J., Noble, J.A., eds.: IPMI. Volume 2732 of Lecture Notes in Computer Science., Springer (2003) 114—125
10. Liu, Y., Teverovskiy, L., Carmichael, O.T., Kikinis, R., Shenton, M.E., Carter, C.S., Stenger, V.A., Davis, S., Aizenstein, H., Becker, J.T., Lopez, O.L., Meltzer, C.C.: Discriminative MR image feature analysis for automatic schizophrenia and alzheimer's disease classification. In: MICCAI (1). (2004) 393—401
11. Lao, Z., Shen, D., Xue, Z., Karacali, B., M.Resnick, S., Davatzikos, C.: Morphological classification of brains via high-dimensional shape transformations and machine learning methods. NeuroImage 21 (2004) 46—57
12. Pham, D.L., Prince, J.L.: Adaptive fuzzy segmentation of magnetic resonance images. IEEE Transactions on Medical Imaging 18 (1999) 737—752
13. Shen, D., Davatzikos, C.: HAMMER: Hierarchical attribute matching mechanism for elastic registration. IEEE Transactions on Medical Imaging 21 (2002) 1421—1439.
14. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3 (2003) 1157—1182
15. Breiman, L.: Bagging predictors. Machine Learning 24 (1996) 123—140
16. McGraw, K.O., Wong, S.P.: Forming inferences about some intraclass correlation coefficients. Psychological Methods 1 (1996) 30—46
17. Vincent, L., Soille, P.: Watersheds in digital spaces: An efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 583—589
18. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46 (2002) 389—422
19. Rakotomamonjy, A.: Variable selection using SVM-based criteria. Journal of Machine Learning Research 3 (2003) 357—1370
20. Vapnik, V.N.: Statistical Learning Theory. Wiley (1998)