Feature Selection for Shape-Based Classification of Biological Objects

Paul Yushkevich[1⋆], Sarang Joshi[1], Stephen M. Pizer[1], John G. Csernansky[2], and Lei E. Wang[2]

[1] Medical Image Display and Analysis Group,
University of North Carolina, Chapel Hill, NC, USA
[2] Department of Psychiatry,
Washington University School of Medicine, St. Louis, MO, USA

Abstract. This paper introduces a method for selecting subsets of relevant statistical features in biological shape-based classification problems. The method builds upon existing feature selection methodology by introducing a heuristic that favors the geometric locality of the selected features. This heuristic effectively reduces the combinatorial search space of the feature selection problem. The new method is tested on synthetic data and on clinical data from a study of hippocampal shape in schizophrenia. Results on clinical data indicate that features describing the head of the right hippocampus are most relevant for discrimination.

1 Introduction

Recent advances in medical imaging and image processing techniques have enabled clinical researchers to link changes in shape of human organs with the progress of long-term diseases. For example, it has been reported that the shape of the hippocampus is different between schizophrenia patients and healthy control subjects [5, 8, 6, 22]. Results of this nature help localize the effects of diseases to specific organs and may subsequently lead to better understanding of disease processes and potential discovery of treatment. This paper addresses the problem of further localizing the effects of diseases to specific *regions* of objects.

Like a number of other methods (e.g., [5, 15, 24, 28, 6, 16, 9]), our approach uses statistical classification to gain insight into the differences in the shape of biological objects between distinct classes of subjects. We enhance classification by using *feature selection* as a tool for localizing inter-class shape differences and for improving the generalization ability of classifiers. The difference between the feature selection method proposed in this paper and the more traditional approaches to dimensionality reduction, such as principal components analysis (PCA), is that the feature subsets yielded by our method have local support in the shape representation, while features such as PCA coefficients have global support. Local feature support makes it possible to identify regions of objects where differences between classes are most significant.

⋆ Corresponding author. Email *pauly@cs.unc.edu*.

The main contribution of this paper is the extension of an existing feature selection method [2] in a way that takes advantage of special properties of features that describe shape. The extended algorithm, called *window selection*, searches for subsets of features that are both highly relevant for classification and are localized in shape space. Window selection takes advantage of a heuristic that the relevance of a given feature for classification correlates with the relevance of the features that describe neighboring locations. This heuristic effectively reduces the otherwise combinatorial search space of feature selection.

The performance analysis of window selection, as compared to feature selection without locality, is reported in this paper for simulated and clinical data. In the synthetic experiments, classes of normally distributed data are generated in a way that simulates the locality of shape features. The ability of the selection algorithms to correctly detect relevant features and the ability to generalize well to new data are compared. The clinical data comes from a study of hippocampal shape in schizophrenia [6], and it is used to compare the results of window selection with previous findings of the relevant regions of the hippocampus.

This paper is organized in five sections. Section 2 describes the details of the window selection algorithm. Sections 3 and 4 present experimental results using simulated and clinical data, respectively. Finally, Sec. 5 discusses the work planned for the future.

2 Methods

2.1 Feature Selection

Feature selection is a machine learning methodology that reduces the number of statistical features in high-dimensional classification problems by finding subsets of features that are most relevant for discrimination (e.g., [19, 14, 20, 13, 25]). Classifiers constructed in the subspace of the selected features tend to generalize better to new data than do classifiers trained on the entire feature set. This paper extends a feature selection method developed by Bradley and Mangasarian [2, 3]. Their method uses elements from support vector machine theory and formulates feature selection as a smooth optimization problem that can be expressed as a sequence of linear programming problems.

The input to this feature selection algorithm consists of a training set of objects that fall into two classes of sizes m and k. Each object is represented by an n-dimensional feature vector. The classes are represented by the feature matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{k\times n}$. We wish to find the set of features, i.e., a subset of columns of $\mathbf{A}$ and $\mathbf{B}$, that are most relevant for discriminating between the two classes. The idea of [2] is to look for a relevant subset of features by finding a hyperplane

$$P = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{w}^T\mathbf{x} = \gamma\right\} \tag{1}$$

that optimally separates the two classes, while lying in the minimal number of dimensions, as formulated by the energy minimization problem

$$P = \arg\min_{\gamma,\mathbf{w}} E_{\text{sep}}(\gamma, \mathbf{w}) + \lambda E_{\text{dim}}(\mathbf{w}) \ . \tag{2}$$

The term E_{sep} measures how well the hyperplane P separates the elements in $\mathbf{A}$ from the ones in $\mathbf{B}$. It is expressed as

$$E_{\text{sep}}(\gamma, \mathbf{w}) = \frac{1}{m} \left\| (-\mathbf{A}\mathbf{w} + \mathbf{e}\gamma + \mathbf{e})_+ \right\|_1 + \frac{1}{k} \left\| (\mathbf{B}\mathbf{w} - \mathbf{e}\gamma + \mathbf{e})_+ \right\|_1 \tag{3}$$

where $\mathbf{e}$ represents a vector of appropriate size whose elements are all equal to 1, and $(\bullet)_+$ is an operation that replaces the negative elements of $\bullet$ with zero.

Let P^- and P^+ be a pair of hyperplanes parallel to P, whose distance to P is $1/\|\mathbf{w}\|$. Then, E_{sep} measures the distance to P^+ of those elements of $\mathbf{A}$ that lie on the 'wrong side' of P^+, as well as the distance to P^- of the elements of $\mathbf{B}$ that lie on the 'wrong side' of P^-. By wrong side, we mean that half-space of P^- or P^+ which contains the hyperplane P.

The energy term E_{dim} in (2) is used to reduce the number of dimensions in which the hyperplane P lies. It has the general form

$$E_{\text{dim}}(\mathbf{w}) = \mathbf{e}^T I(\mathbf{w}), \tag{4}$$

where $I(\mathbf{w})$ is an indicator function that replaces each non-zero element of $\mathbf{w}$ with 1. However, since indicator functions are inherently combinatorial and badly suited for optimization, Bradley and Mangasarian suggest approximating the indicator function with a smooth function

$$I\left(\{w_1 \dots w_n)\}\right) = \left\{1 - \varepsilon^{-\alpha|w_1|}, \dots, 1 - \varepsilon^{-\alpha|w_n|}\right\}, \tag{5}$$

which, according to [1], yields the same solutions as the binary indicator function for finite values of the constant α .

2.2 Window Selection for Shape Features

General feature selection algorithms make minimal assumptions about the nature and the properties of features. For instance, the same algorithm may be used for classifying documents on the basis of word frequency or for breast cancer diagnosis. Without prior knowledge of feature properties, the feature selection problem is purely combinatorial, since in a set of n features there are 2^n possible subsets and all of them are considered to be equally worthy candidates for selection.

In shape classification problems, features are typically derived from dense geometrical object representations [4, 23, 18, 21, 10, 9, 7, 15], and special relationships exist between features derived from neighboring locations in the objects. We hypothesize that by incorporating the heuristic knowledge of these relationships into a feature selection algorithm, we can improve its performance and stability when applied to shape classification.

Features that describe shape are geometric in nature and the concept of distance between two features can be defined, usually in terms of geometric distance between locations described by the features. Furthermore, natural biological processes exhibit *locality*: geometric features capturing shape of anatomical objects

that are close together are likely to be highly correlated. General features, such as word frequencies in documents, may not exhibit this property of locality.

Locality makes it possible to impose a prior on the search space of a feature selection algorithm. Locality implies that feature sets consisting of one or a few clusters are more likely candidates than feature sets in which the selected features are isolated. To reward locality, the energy minimization in (2) is expanded to include an additional *locality energy* term $E_{\text{loc}}(\mathbf{w})$:

$$P = \arg\min_{\gamma,\mathbf{w}} E_{\text{sep}}(\gamma, \mathbf{w}) + \lambda E_{\text{dim}}(\mathbf{w}) + \eta E_{\text{loc}}(\mathbf{w}) \ . \tag{6}$$

$E_{\text{loc}}(\mathbf{w})$ estimates the number of clusters formed by the features selected by $\mathbf{w}$, thus rewarding the locality of the selected features. Let $J \subset \{1 \dots n\}$ be the set of non-zero features in $\mathbf{w}$. To measure how clustered the components of J are, we define an 'alphabet' of structured subsets of $\{1 \dots n\}$ called *windows*, and measure the most compact description needed to express J using this alphabet.

We define feature windows are structured sets of 'neighboring features'. The neighborhood relationships between the features in the set $\{1 \dots n\}$ depend on the topology of the underlying space that is being described by the features. For instance, if features are computed at points that are regularly sampled from a boundary manifold, then two features are neighbors if the geodesic distance between the points from which they are computed is small.

Let d_{ij} be a metric that assigns a non-negative distance to every pair of features $i, j \in \{1 \dots n\}$. This distance metric is used to define feature windows. A set $W \subset \{1 \dots n\}$ is called a *window of size* q if (i) $d_{ij} \leq q$ for all $i, j \in W$, and (ii), there does not exist a superset of W in $\{1 \dots n\}$ for which the condition (i) holds. An alphabet of windows is just a set of all possible windows of sizes $1, \dots, w_{\max}$.

The distance metric allows us to define windows on arbitrarily organized features. For instance, when features are organized in a one-dimensional lattice, the distance metric $d_{ij} = |i - j|$ yields windows that are contiguous subsets of features, while $d_{ij} = |i - j| \bmod n$ allows for wrap-around windows, which are useful when features are sampled along a closed curve. On higher-dimensional lattices, different distance metrics such as Euclidean or Manhattan distance generate differently shaped windows. For features computed at vertices in a mesh, windows can be constructed using transitive distance, which counts the smallest number of edges that separate a pair of vertices.

Let $\mathbf{W} = \{W_1 \dots W_N\}$ be a set of windows of various sizes over the feature set $\{1 \dots \text{n}\}$. The *minimal window cover* of a feature subset J is defined as the smallest set $\alpha \subset \{1 \dots N\}$ for which $J = \bigcup_{i \in \alpha} W_i$. The locality energy component $E_{\text{loc}}(\mathbf{w})$ is defined as the size of the minimal window cover of the set J of non-zero features in the vector $\mathbf{w}$. While such a formulation is combinatorial in nature, in the following section it is elegantly expressed in terms of linear programming.

2.3 Linear Programming Formulation

According to Bradley and Mangasarian [2], the feature selection problem (2) can be formulated as the following smooth non-linear program:

$$\begin{array}{ll} \underset{\gamma,\mathbf{w},\mathbf{y},\mathbf{z},\mathbf{v}}{\text{minimize}} & \frac{\mathbf{e}^T\mathbf{y}}{m} + \frac{\mathbf{e}^T\mathbf{z}}{k} + \lambda \mathbf{e}^T I(\mathbf{v}), \\ \text{subject to} & \begin{array}{c} -\mathbf{A}\mathbf{w} + \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{y} \\ \mathbf{B}\mathbf{w} - \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{z} \\ \mathbf{y} \geq 0,\ \mathbf{z} \geq 0\ , \\ -\mathbf{v} \leq \mathbf{w} \leq \mathbf{v}\ . \end{array} \end{array} \tag{7}$$

This formulation does not directly minimize the objective function (2), but rather it minimizes positive vectors $\mathbf{y}, \mathbf{z}$, and $\mathbf{v}$, which constrain the components of the objective function. Such a transformation of the minimization problem is frequently used in support vector methodology in order to apply linear or quadratic programming to energy minimization problems.

The vector $\mathbf{v}$ constraints $\mathbf{w}$ from above and below and thus eliminates the need for using the absolute value of $\mathbf{w}$ in the objective function, as is done in (3). The non-zero elements of $\mathbf{v}$ correspond to selected features.

In order to introduce the locality energy E_{loc} into the linear program, we can express the non-zero elements of $\mathbf{v}$ as a union of a small number of windows, and penalize the number of windows used. Let $W_1 \ldots W_N$ be an 'alphabet' of windows, as defined in Sec. 2.2. Let $\mathbf{\Omega}$ be an $n \times N$ matrix whose elements ω_{ij} are equal to 1 if the feature i belongs to the window W_j, and are equal to 0 otherwise. Let $\mathbf{u}$ be a sparse positive vector of length N whose non-zero elements indicate a set of selected windows. Then the non-zero elements of $\mathbf{\Omega u}$ indicate a set of features that belong to the union of the windows selected by $\mathbf{u}$.

In order to implement window selection as a smooth non-linear program, the terms $\mathbf{u}$ and $\mathbf{\Omega u}$ are used in place of $\mathbf{v}$ in the objective function. The resulting formulation penalizes both the number of selected windows and the number of features contained in those windows:

$$\begin{array}{ll} \underset{\gamma,\mathbf{w},\mathbf{y},\mathbf{z},\mathbf{u}}{\text{minimize}} & \frac{\mathbf{e}^T\mathbf{y}}{m} + \frac{\mathbf{e}^T\mathbf{z}}{k} + \left(\lambda \mathbf{e}^T\mathbf{\Omega} + \eta \mathbf{e}^T\right) I(\mathbf{u}), \\ \text{subject to} & \begin{array}{c} -\mathbf{A}\mathbf{w} + \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{y} \\ \mathbf{B}\mathbf{w} - \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{z} \\ \mathbf{y} \geq 0,\ \mathbf{z} \geq 0\ , \\ -\mathbf{\Omega u} \leq \mathbf{w} \leq \mathbf{\Omega u}\ . \end{array} \end{array} \tag{8}$$

This formulation of the objective function is identical to the energy minimization formulation (6) if none of the windows selected by $\mathbf{u}$ overlap. In case of an overlap, the penalty assessed on the combined number of features in all of the selected windows, and not on the total number of windows in the vector $\mathbf{w}$.

We use a fast successive linear approximation algorithm outlined in [2] to solve the program (8). The algorithm is randomly initialized and iteratively solves a linear programming problem in which the concave term $I(\mathbf{u})$ is approximated using the Taylor series expansion. The algorithm does not guarantee a

global optimum but does converge to a minimum after several iterations. The resulting vector $\mathbf{u}$, whose non-zero elements indicate the selected windows, is very sparse. The *Sequential Object-Oriented Simplex Class Library (SoPlex)*, developed by Roland Wunderling [26], is used for solving the linear programming problems.

The parameters λ and η affect the numbers of features and windows (and hence the sizes of the windows) selected by the window selection algorithm. Larger values of λ yield fewer features, and similarly, larger values of η yield fewer windows. When both parameters are zero, the algorithm performs no feature selection and acts as a linear support vector machine classifier. The number of features yielded in this case is bounded only by the size of the training set.

Bradley and Mangasarian [2] suggest reserving a small portion of the training set and using it to search for the value of parameter λ that leads to optimal cross-validation performance. In the synthetic data experiments described below, we have found that cross-validation performance is poorly suited for finding optimal parameters because of its low signal-to-noise ratio. Parameters yielded by optimization seldom correctly identified the relevant sets of features [27]. However, if in a particular application one roughly knows how many relevant features and windows are desired, then the parameter values needed to produce such windows can be determined experimentally and a search for optimal values is unnecessary.

3 Results on Simulated Data

This section summarizes a simulated data experiment is described in full in [27]. In this experiment window selection and feature selection were compared in a setting where the features are normally distributed, and the relevant features are clustered.

In each variation of the experiment, two training classes were randomly sampled from pairs of 15-dimensional normal distributions with identity covariance matrices and with means that differ in only 6 of the 15 dimensions. The relevant dimensions in one case are arranged into a single contiguous block, and in another case they form two disjoint blocks of 3 features. Feature selection and window selection with windows defined using distance metric $d_{ij} = |i - j|$ were applied to the training samples. Classifiers were constructed in the subspaces defined by the selected features, and their expected generalization ability was computed empirically. The experiment was repeated for different sizes of the training set (30, 60, 90, and 120), and for each training set size, it was repeated 40 times, with the average generalization rate recorded.

Figure 1 shows the results of these experiments: classifiers based on window selection outperformed the classifier based on feature selection, especially in the first case when the relevant features are arranged into a single block. Both selection schemes resulted in better classifiers than the classifier constructed on the entire feature set. Also, window selection correctly identified the relevant sets features with significantly higher frequency than feature selection (see [27]).

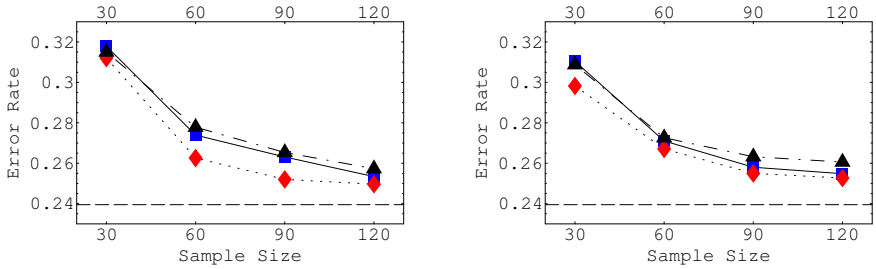

Fig. 1. Performance of window and feature selection on Gaussian data. Relevant features are arranged into one block (left plot) and two blocks (right block). Plotted are the expected error rates of the window selection algorithm (diamond, dotted line), the feature selection algorithm (square, dashed line), and global discriminant analysis (triangle, solid line) versus training sample size.

4 Results on Clinical Hippocampus Data

The window and feature selection algorithms were applied to the study of the shape of the hippocampus in schizophrenia using the data set that is identical to the one reported in [6]. The data set consists of 117 subjects, 52 of whom are schizophrenia patients, and the remaining 65 are matched healthy controls. The left and right hippocampi of each subject are described using boundary meshes that consist of 6,611 vertices and 13,218 triangular faces. These segmentations were obtained using large-deformation diffeomorphic image matching described in [15, 12, 5, 6].

Hippocampus is not a homogenous structure but rather consists of many identifiable sub-regions, which may be affected differently by schizophrenia. Indeed, [6] stipulates that "the pattern of shape abnormality suggested a neuroanatomical deformity of the head of the hippocampus, which contains neurons that project to the frontal cortex". However, the statistical methodology employed in [6] is based on the eigenshape formulation that does not allow local specificity of shape variation. The motivation for applying feature and window selection to this data set is to find the regions of the hippocampus where the shape differences associated with schizophrenia are most significant.

In order to use window and feature selection to produce regions large enough to cover 10%-20% of the hippocampal surface, we reduced the number of features from nearly $40,000$ that result from using the x, y, z coordinates of each mesh vertex as features, to 160 summary features, which describe small patches on the surface of the hippocampus. The reduction was necessary because window selection and feature selection algorithms yield fewer features than there are subjects in the training set and because of the prohibitive computational cost of using so many features.

Patch features were computed as follows. We aligned the sets of 117 left and 117 right meshes using the Generalized Procrustes algorithm [11] restricted to translation and orientation. In the process, we computed the mean left and right

Table 1. Results of leave-one-out experiments with feature selection and window selection on clinical data with patch summary features. Each column represents one set of 117 experiments. Legend: λ and η are the parameters from (6) that affect the number of selected features and windows, $\bar{N}_{\text{win}}$ is the average number of selected windows, $\bar{N}_{\text{feat}}$ is the average number of selected features, and R is the leave-one-out correct classification rate, in percent.

η	λ	$\bar{N}_{\text{win}}$	$\bar{N}_{\text{feat}}$	R (%)	η	λ	$\bar{N}_{\text{win}}$	$\bar{N}_{\text{feat}}$	R (%)
0.0	0.04		22.9	55.6	0.08	0.04	10.8	28.1	61.5
0.0	0.08		16.4	65.0	0.08	0.08	5.7	13.5	64.1
0.0	0.12		7.5	65.0	0.08	0.12	2.8	6.1	64.1
0.0	0.16		4.6	68.4	0.08	0.16	1.6	2.9	59.0
0.04	0.04	11.8	28.7	68.4	0.12	0.04	9.2	24.5	68.4
0.04	0.08	8.5	19.3	69.2	0.12	0.08	3.9	9.9	62.4
0.04	0.12	4.2	8.5	62.4	0.12	0.12	2.1	4.7	57.3
0.04	0.16	2.1	4.0	54.7	0.12	0.16	1.4	2.8	61.4

hippocampal meshes. We subdivided each mesh into 80 patches of roughly equal area using METIS graph partitioning software [17] on a graph whose vertices correspond to the mesh triangles and are weighted by the average areas of the triangles. The partitioned left and right mean meshes are shown in the top row of Fig. 2. Each patch was represented by a single summary feature, which measures the average inward/outward deformation of the patch with respect to the mean mesh. The use of a single feature per location makes is easier to define a distance metric between features, as having multiple features per location would either require defining the distance between them to be zero, which would result in them always being selected together, or it would require two distance functions, one for features from different locations and another for features from the same location.

An alphabet of windows was defined over the patch summary features using the transitive distance function, which counts the number of patch edges that separate any two patches. Under this function, single patches form windows of size 0 and sets of mutually adjacent patches form windows of size 1. For computational efficiency, windows of larger size were not included in the alphabet.

Feature selection and window selection algorithms were applied to patch summary features in a series of leave-one-out cross-validation experiments. In each leave-one-out iteration, one subject was removed from the data set, the selection algorithm was applied to the remaining subjects, an $L1$ support vector classifier was constructed in the subspace spanned by the selected features, the left out subject was assigned a class label by the classifier, and this class label was compared to the true class label of the left out subject. The average correct classification over 117 leave-one-out iterations was recoded. The feature selection and window selection experiments were repeated for different values of modulation parameters λ and η. Table 1 shows the results of these experiments.

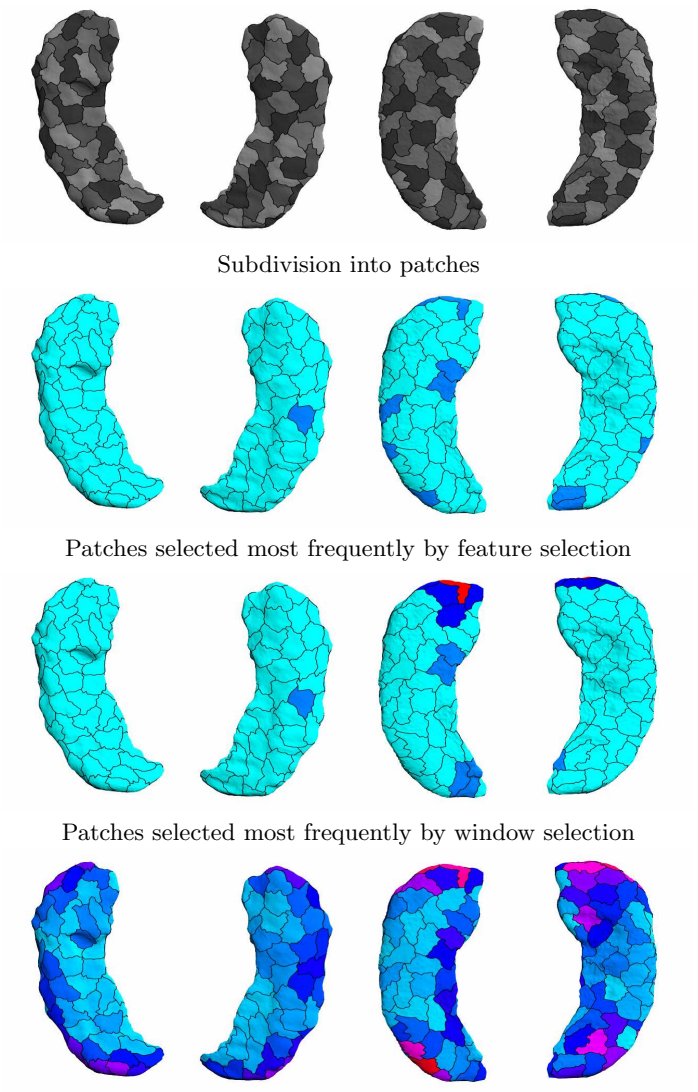

Fig. 2. Top row: mean left and mean right hippocampal meshes partitioned into 80 patches each. The meshes are shown from superior and anterior viewpoints. Second row: ten patches that were selected most frequently during leave-one-out validation of feature selection. Third row: ten windows that were selected most frequently during leave-one-out validation of feature selection (some of the windows overlap, and patches that belong to more than one window are shaded darker on the cyan-red hue scale). Bottom row: p-values of the mean difference tests computed at each patch; the negative logarithm of the p-values is displayed using the cyan-red hue scale (cyan = no significance, red = high significance).

In [6], using a 10-fold cross-validation methodology, a similar classification rate of 68.4% is reported. The methods in [6] are based on eigenanalysis of the entire set of $40,000$ features. The results in Table 1 show that with intelligent feature selection a similar classification rate can be achieved with only 160 summary features. The feature selection methodology also specifies the local regions of the hippocampus that are significant for discrimination.

The second row of Fig. 2 shows the ten patches that were selected most frequently in the 117 leave-one-out experiments conducted with the feature selection algorithm with $\lambda = 0.16$. The third row of Fig. 2 shows the ten most frequently selected patch windows in the window selection experiment with $\lambda = 0.12$ and $\eta = 0.08$. Window selection results in fewer isolated features than feature selection. For reference, the bottom row of Fig. 2 plots the p-values of mean difference hypothesis tests computed at each patch. No correction for the repeated nature of tests has been applied. While the pattern of patches selected by the window and feature selection algorithms closely resembles the pattern of patches with low p-values, the selected patches do not correspond to the patches with lowest p-values. As stipulated in [6], the head of the right hippocampus was shown by window selection to be most relevant for discrimination.

5 Discussion and Conclusions

It is unlikely that a classification technique will one day make it possible to accurately diagnose schizophrenia on the basis of hippocampal shape. Therefore, our goal in developing the window selection algorithm was not so much to build a better classifier but rather to find the regions of the hippocampus that are significant for discrimination. With respect to this goal, the results presented in this paper are encouraging. However, these results require further validation using a different hippocampal data set. We plan to perform this validation in the future.

We also plan to perform window and feature selection on hippocampal patches selected manually on the basis of biological homogeneity and function. The use of anatomically significant patches in the selection algorithms could open new insights into schizophrenia. On the theoretical front, we plan to extend this paper's framework to select features in a hierarchical manner. Selected patches would be further partitioned into smaller patches, and the selection algorithms would be performed again on the residuals, resulting in a high-resolution set of selected features. Hierarchical feature selection would eliminate the information loss incurred by reduction to patch summary features.

In conclusion, we have presented a framework for using feature selection in shape characterization, developed a new window selection algorithm for handling localized shape features, and applied feature and window selection to synthetic and clinical data. The results on clinical data confirm an earlier finding from [6] that the head of the hippocampus is significant in respect to schizophrenia and suggest that the framework does provide useful locality and effective discrimination.

Acknowledgements

The research reported in this paper was carried out under partial support of the NIH grant P01 CA47982 and the Silvio Conte Center at Washington University School of Medicine grants MH56584 and MH62130. Dr. Guido Gerig, Dr. J.S. Marron, Dr. Dr. James Damon, Dr. Keith E. Muller, Sean Ho, P. Thomas Fletcher, and other participants of the Statistics of Shape Seminar held at the University of North Carolina have contributed to this research through constructive criticism and advise. We thank Dr. Adam Cannon at Columbia University for his help in stimulating this research.

References

1. P. Bradley, O. Mangasarian, and J. Rosen. Parsimonious least norm approximation. Technical Report 97-03, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, March 1997.
2. P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proc. 15th International Conf. on Machine Learning*, pages 82–90. Morgan Kaufmann, San Francisco, CA, 1998.
3. P. S. Bradley, O. L. Mangasarian, and W. N. Street. Feature selection via mathematical programming. *INFORMS Journal on Computing*, 10:209–217, 1998.
4. T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 1(61):38–59, 1994.
5. J. Csernansky, S. Joshi, L. Wang, J. Haller, M. Gado, J. Miller, U. Grenander, and M. Miller. Hippocampal morphometry in schizophrenia via high dimensional brain mapping. In *Proc. National Academy of Sciences*, volume 95, pages 11406–11411, 1998.
6. J. G. Csernansky, L. Wang, D. Jones, D. Rastogi-Cruz, J. A. Posener, G. Heydebrand, J. P. Miller, and M. I. Miller. Hippocampal deformities in schizophrenia characterized by high dimensional brain mapping. *Am. J. Psychiatry*, 159:2000–2006, 2002.
7. C. Davatzikos, M. Vaillant, S. Resnick, J. Prince, S. Letovsky, and R. Bryan. A computerized approach for morphological analysis of the corpus callosum. *Journal of Computer Assisted Tomography*, 20:207–222, 1995.
8. G. Gerig, M. Styner, M.E. Shenton, and J. Lieberman. Shape versus size: Improved understanding of the morphology of brain structures. In W Niessen and M Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 2208, pages 24–32, New York, October 2001. Springer.
9. P. Golland, B. Fischl, M. Spiridon, N. Kanwisher, R. L. Buckner, M. E. Shenton, R. Kikinis, A. M. Dale, and W. E. L. Grimson. Discriminative analysis for image-based studies. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 1, pages 508–515. Springer, 2002.
10. P. Golland, W.E.L. Grimson, and R. Kikinis. Statistical shape analysis using fixed topology skeletons: Corpus callosum study. In *International Conference on Information Processing in Medical Imaging*, LNCS 1613, pages 382–388. Springer Verlag, 1999.
11. J.C. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

12. J.W. Haller, A. Banerjee, G.E. Christensen, M. Gado, S. Joshi, M.I. Miller, Y.I. Sheline, M.W. Vannier, and J.G. Csernansky. Three-dimensional hippocampal MR morphometry by high-dimensional transformation of a neuroanatomic atlas. *Radiology*, 202:504–510, 1997.
13. Tony S. Jebara and Tommi S. Jaakkola. Feature selection and dualities in maximum entropy discrimination. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, pages 291–300, San Francisco, CA, 2000. Morgan Kaufmann Publishers.
14. George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129, 1994.
15. S. Joshi, U. Grenander, and M. Miller. On the geometry and shape of brain sub-manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:1317–1343, 1997.
16. S. Joshi, S. Pizer, P.T. Fletcher, P. Yushkevich, A. Thall, and J.S. Marron. Multiscale deformable model segmentation and statistical shape analysis using medial descriptions. *Invited submission to IEEE-TMI*, page t.b.d., 2002.
17. G. Karypis and V. Kumar. *MeTiS – A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices – Version 4.0*. University of Minnesota, 1998.
18. András Kelemen, Gábor Székely, and Guido Gerig. Elastic model-based segmentation of 3D neuroradiological data sets. *IEEE Transactions on Medical Imaging*, 18:828–839, October 1999.
19. K. Kira and L. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Tenth National Conference Conference on Artificial Intelligence (AAAI-92)*, pages 129–134. MIT Press, 1992.
20. Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.
21. S.M. Pizer, D.S. Fritsch, P. Yushkevich, V. Johnson, and E. Chaney. Segmentation, registration, and measurement of shape variation via image object shape. *IEEE Transactions on Medical Imaging*, 18:851–865, October 1999.
22. M.E. Shenton, G. Gerig, R.W. McCarley, G. Szekely, and R. Kikinis. Amygdala-hippocampus shape differences in schizophrenia: The application of 3D shape models to volumetric mr data. *Psychiatry Research Neuroimaging*, pages 15–35, 2002.
23. L.H. Staib and J.S. Duncan. Boundary finding with parametrically deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(11):1061–1075, November 1992.
24. M. Styner. *Combined Boundary-Medial Shape Description of Variable Biological Objects*. PhD thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2001.
25. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In *Advances in Neural Information Processing Systems 13*, pages 668–674. MIT Press, 2001.
26. Roland Wunderling. *Paralleler und Objektorientierter Simplex-Algorithmus*. PhD thesis, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, 1996. ZIB technical report TR 96-09.
27. P. Yushkevich. *Statistical Shape Characterization using the Medial Representation*. PhD thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2003.
28. P. Yushkevich, Pizer S.M., S. Joshi, and Marron J.S. Intuitive, localized analysis of shape variability. In *International Conference on Information Processing in Medical Imaging*, pages 402–408, Berlin, Germany, 2001. Springer-Verlag.