Feature Selection for Shape-Based Classification of Biological Objects*

Paul Yushkevich† Sarang Joshi
Stephen M. Pizer
Medical Image Display and Analysis Group
The University of North Carolina at Chapel Hill

November 15, 2002

Abstract

Feature selection methodology from machine learning literature is adapted and applied to the problem of statistical shape-based classification of biological objects. The feature selection paradigm is used to discover the regions of objects where the difference between classes is most pronounced. This also improves the generalization ability and statistical significance of shape-based analysis. A feature selection algorithm based on support vector machines is extended to take advantage of special relationships between neighboring features, which are inherently present in geometric object representations. Performance analysis using simulated data is presented.

1 Introduction

Recent advances in medical imaging and image processing techniques have enabled medical researchers to link changes in the shape of human organs to the progress of some long-term diseases. For instance, differences in the shape of the hippocampus between schizophrenia patients and healthy subjects have been reported [8, 25]. Results of this nature are powerful because they promise to enable early diagnosis of serious diseases, and because they may reveal the nature of the biological processes responsible for the diseases.

However, the methods for characterizing how and where a given disease affects an organ can still be improved. Such characterization is made difficult by the fact that shape differences between healthy and diseased organs are often less prominent that the inter-population variability. Furthermore, standard statistical techniques may not be reliable in these kinds of problems, due to small

*Submitted to CVPR 2003.
†Corresponding author, email *pauly@cs.unc.edu*

sample sizes and large numbers of features that are needed to describe human organs.

This paper explores the ways in which a machine learning technique called *feature selection* can be used to improve shape characterization. Feature selection is used to reduce the dimensionality of classification problems by finding the subset of features that best captures the differences between classes. Classifiers restricted to the selected subset of features are less affected by sampling noise and tend to generalize better than the classifiers trained on the entire feature set. Feature selection has been shown to dramatically improve the generalization ability of classifiers in high-dimensional problems [3, 29].

The potential benefit of using feature selection algorithms in shape characterization problems extends beyond the improved generalization ability. Examination of features deemed most relevant by such algorithms may reveal the areas of organs that are most affected by a disease, leading to improved localization and understanding of the biological processes responsible for the disease.

Feature selection algorithms in machine learning literature usually address general classification problems and make minimal assumptions as to where the different features come from and how they may be related. In shape classification problems, where features are usually derived from dense geometrical object representations, there exist special relationships between neighboring features. By incorporating our knowledge of these relationships into feature selection algorithms, we can improve their performance and stability when applied to shape classification. The two properties of shape features that are particularly useful for improving feature selection are *structure* and *locality*.

We use the term *structure* to indicate the importance of the order in which the features are arranged in a classification problem. In many problems, the order of the features is arbitrary, as is the case, for example, when all the features describe different physical properties of an object, such as its height, weight, age or density. However, when the features are measurements regularly sampled from a lattice, as is the case in many geometrical object representations, the order of the features is important, as nearby features are more likely to be correlated than the far-away features.

A biological process responsible for variability in the shape of an anatomical object exhibits *locality* if it affects the object at one or at most a few locations, which are consistent across the population of objects. In reference to a feature set, we use the term locality to mean that some components of the statistical variability in the data can be localized to one or more subsets the features.

In the absence of structure and locality, the feature selection problem is purely combinatorial, since in the set of n features there are 2^n possible subsets and all of them are considered *a priori* to be equally worthy candidates for feature selection. The properties of structure and locality constitute prior knowledge about the kinds of feature subsets that ought to be selected. Feature sets consisting of one or a few contiguous subsets are more likely candidates than feature sets in which the selected features appear scattered. By assuming that shape features exhibit structure and locality, we can reduce the number of possible solutions of a feature selection algorithm.

The paper is organized as follows. In Section 2 we provide an overview of the related literature in the fields of machine learning and shape characterization. In Section 3 we incorporate the prior knowledge about structure and locality of shape features into an existing feature selection method by Bradley and Mangasarian [3]. In Section 4.1 we analyze the performance of the algorithm in simulated data examples, where the features are normally distributed. In Section 4.2 we apply the method to a simulated shape classification problem.

2 Background

2.1 Feature Selection

In classification problems, the generalization ability of a classifier can be improved by removing features that do not contribute to classification, i.e. are *irrelevant*. A number of algorithms for automatic feature selection have been developed in the machine learning literature.

Feature selection methods can be categorized into *filter* methods and *wrapper* methods. Filter methods deal with the feature selection task independently of the classification: they find and remove irrelevant features first, and pass the rent on to the classification [21]. Wrapper methods use classification as a subtask; as they try different subsets of features, they perform classification and cross-validation on each subset, until an optimum is found [17, 22]. Wrapper methods generally perform better than filter methods, but are much more time consuming, as each iteration of the method requires an execution and testing of the classification method.

Feature selection methods can also be categorized as *exhaustive*, *randomized* or *sequential*, based on the search algorithm that they employ for finding the optimal feature subset [1, 14]. Exhaustive methods search for an optimal subset of n features using either a combinatorial search of all the 2^n possible subsets, or using AI techniques, such as the branch and bound algorithm [23]. Sequential feature selection methods achieve polynomial time complexity by iteratively adding and subtracting features to a subset in a greedy fashion. Randomized methods employ stochastic search techniques, such as simulated annealing and genetic algorithms. Comparisons of a number of popular feature selection techniques can be found in paper by Aha and Bankert [1], and by Jain and Zongker [15]. According to [7, 14], only the exhaustive search procedure can be guaranteed to produce the globally optimal feature subset.

The feature selection algorithm by Bradley and Mangasarian [3, 4], which is used and extended in this paper, does not directly fall into one of the above categories. It can be said to be a wrapper method because it uses the classification algorithm as a component. It also falls somewhere between the randomized and sequential categories as it has a stochastic component and uses hill climbing techniques. It formulates the feature selection as a smooth optimization problem and finds an optimum by solving a sequence of linear programming problems. The method falls into a broader category of techniques that integrate support

vector machine methodology with feature selection [29, 16].

2.2 Shape-Based Classification

Shape classification methods can be characterized by the object representation that they use to yield statistical features and by the statistical methods employed to analyze the features.

Landmark methods [2] use corresponding points of special anatomic or geometric significance as features. Since the number of such landmarks is relatively small, the properties of structure and locality, defined in Section 1, do not apply to features derived from landmarks. However, methods based on landmarks can benefit from the existing feature selection methodology.

The feature selection method presented in this paper is best suited for the shape analysis approaches that use dense object representations. This rich class of representations includes boundary point distribution models [6], parametric boundary models based on Fourier and spherical harmonic basis decomposition [26, 20], discrete and continuous medial representations [24, 31, 12], as well as functional object representations, such as distance transforms [13], and deformation fields based on a warping of a template to each object in the training set [8, 9, 18]. The features yielded by these representations exhibit structure and locality because they are densely and regularly sampled.

Many of the methods in statistical shape analysis focus on estimating the probability distribution on the shape space [6, 20, 26, 18]. Such probability distributions can be used as priors for deformable segmentation and can be sampled to visualize shape variability. Principal component analysis, which is used by many of the methods to estimate the shape distribution, is related to feature selection, as it reduces the dimensionality of the data to a linear combination of the original features.

While the issues of representation and correspondence have been the focus of extensive research in shape characterization literature, the application of feature selection paradigm to the shape characterization problem have received less attention. However, there has been considerable work in the literature in using classification techniques to detect, localize and describe the anatomical differences in the shape between different populations [11, 10, 27, 8, 18, 19, 32].

3 Methods

This section contains the details of the new feature selection method for shape classification, which is an extension of an existing feature selection algorithm by Bradley and Mangasarian [3]. The novelty of our method is that it searches for an optimal set of *windows* of features, as described in the following sections. For clarity, we will use the term *feature selection* to refer to the original algorithm, while referring to our extended version using the term *window selection*.

This section is organized as follows. Subsection 3.1 formulates the feature selection problem in general terms of energy minimization. Subsection 3.2 ex-

pands this formulation by adding an energy term that favors feature sets that exhibit locality and structure. The concept of minimal window cover is used to measure how localized and structured a feature set is. Subsection 3.3 describes how the feature selection problem can be formulated and solved using linear programming.

3.1 Feature Selection

The input to a 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 Bradley and Mangasarian [3] 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 elements 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 ones of appropriate size, 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 [5], yields the same solutions as the binary indicator function for finite values of the constant α .

3.2 Locality Term for Feature Selection

We hypothesize that features derived from dense geometrical object representations of biological objects to exhibit the properties of structure and locality, which are described in Section 1. These properties imply that that if a certain feature strongly contributes to the separation between two classes of objects, then the neighboring features are also likely to strongly contribute to the separation.

The formulation of the feature selection problem in (2) rewards good separation between classes in a small number of features, but does not take structure and locality of the features into account. As discussed in Section 1, structure and locality constitute a prior term for feature selection that can reduce the complexity of its solution space.

To reward locality, we expand the energy minimization formulation (2) to include an additional energy term:

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

The term $E_{\text{loc}}(\mathbf{w})$ rewards selection of neighboring features, by requiring that the non-zero elements of $\mathbf{w}$ be ordered in a structured manner.

Let $J \subset \{1 \dots n\}$ be the set of features for which $\mathbf{w}$ is non-zero. To measure how structured J is, we define an 'alphabet' of structured subsets of $\{1 \dots n\}$ that we call *windows*, and measure the most compact description needed to express J using this alphabet.

The neighborhood relationships between the features in the set $\{1 \dots n\}$ depend on the structure of the space from which the features are sampled. Typically, as in the case of parametric shape descriptions, the underlying structure of a feature set is a lattice of one or two dimensions.

In order to define an alphabet of windows over the feature set $\{1 \dots n\}$, we use a metric $d(i, j)$ that assigns a non-negative distance to every pair of features i, j. A set $W \subset \{1 \dots n\}$ is defined to be a *window of size* q if (i) $d(i, j) \leq q$ 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.

The distance function allows us to define windows on arbitrarily organized features. For instance, when features are organized in a one-dimensional lattice, and the distance function is $d(i, j) = |i - j|$, the windows are contiguous subsets of features. By letting $d(i, j) = |i - j| \mod n$, we can allow for wrap-around windows, which are useful for periodic features, such as features sampled along the boundary of a closed object. On higher-dimensional lattices, different distance functions such as Euclidean distance and Manhattan distance generate differently shaped windows. For the features sampled from vertices on a mesh, windows can be constructed using the transitive distance function, which counts the smallest number of edges on a mesh 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

$$J \subset \bigcup_{i \in \alpha} W_i \ . \tag{7}$$

We take the locality energy component $E_{\text{loc}}(\mathbf{w})$ to be equal to the size of the minimal window cover of the set of non-zero features in the vector $\mathbf{w}$. While such a formulation is combinatorial in nature, in the following sections we formulate it the context of linear programming and derive an elegant implementation.

3.3 Linear Programming Formulation

Bradley and Mangasarian express the energy minimization problem in (2) as a series of linear programming problems [3]. This section briefly summarizes their approach and extends it to include the formulation (6).

The term $E_{\text{sep}}(\mathbf{w}, \gamma)$ in (3) is linear. The global minimum of E_{sep} can be found by solving the following linear programming problem:

$$\begin{array}{ll} \underset{\gamma, \mathbf{w}, \mathbf{y}, \mathbf{z}}{\text{minimize}} & \frac{\mathbf{e}^T \mathbf{y}}{m} + \frac{\mathbf{e}^T \mathbf{z}}{k} \ , \\ & -\mathbf{A}\mathbf{w} + \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{y} \\ \text{subject to} & \mathbf{B}\mathbf{w} - \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{z} \\ & \mathbf{y} \geq 0, \ \mathbf{z} \geq 0 \ . \end{array} \tag{8}$$

The feature selection problem in (2) can be formulated as a smooth non-linear problem

$$\begin{array}{ll} \underset{\gamma, \mathbf{w}, \mathbf{y}, \mathbf{z}, \mathbf{v}}{\text{minimize}} & \frac{\mathbf{e}^T \mathbf{y}}{m} + \frac{\mathbf{e}^T \mathbf{z}}{k} + \lambda \mathbf{e}^T I(\mathbf{v}), \\ & -\mathbf{A}\mathbf{w} + \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{y} \\ \text{subject to} & \mathbf{B}\mathbf{w} - \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{z} \\ & \mathbf{y} \geq 0, \ \mathbf{z} \geq 0 \ , \\ & -\mathbf{v} \leq \mathbf{w} \leq \mathbf{v} \ . \end{array} \tag{9}$$

Bradley and Mangasarian call this problem *Feature Selection Concave (FSV)*. Notice, that the absolute value of $\mathbf{w}$ from (3) does not appear in this formulation; instead the positive vector $\mathbf{v}$ is used to clamp $\mathbf{w}$ from above and below.

The non-zero elements of the vector $\mathbf{v}$ correspond to the selected features. In order to introduce the locality energy E_{loc} into the linear program, we 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 $W_1 \dots W_N$ be an 'alphabet' of windows, as defined in section 3.2. Let $\mathbf{\Omega}_{n \times N}$ be a matrix whose elements are defined as

$$\omega_{ij} = \begin{cases} 1 & \text{if } i \in W_j, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Each column of this matrix represents a window in the alphabet. Let $\mathbf{u}$ be a sparse positive vector of length N. Then the non-zero elements of the vector

$\mathbf{\Omega u}$ constitute the union of windows corresponding to the non-zero elements of $\mathbf{u}$.

In the following linear program, we introduce $\mathbf{u}$ into the objective of the linear programming formulation, in a manner that penalizes both the number of windows used and the number of features contained in those windows:

$$\begin{array}{ll} \underset{\gamma,\mathbf{w},\mathbf{y},\mathbf{z},\mathbf{u}}{\text{minimize}} & \frac{\mathbf{e}^T\mathbf{y}}{m} + \frac{\mathbf{e}^T\mathbf{z}}{k} + \left(\lambda \mathbf{e}^T \mathbf{\Omega} + \eta \mathbf{e}^T\right) I(\mathbf{u}), \\ \text{subject to} & \begin{array}{c} -\mathbf{A}\mathbf{w} + \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{y} \\ \mathbf{B}\mathbf{w} - \mathbf{e}\gamma + \mathbf{e} \leq \mathbf{z} \\ \mathbf{y} \geq 0,\ \mathbf{z} \geq 0\ , \\ -\mathbf{\Omega u} \leq \mathbf{w} \leq \mathbf{\Omega u}\ . \end{array} \end{array} \tag{11}$$

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 [3] to solve the program (11). 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.

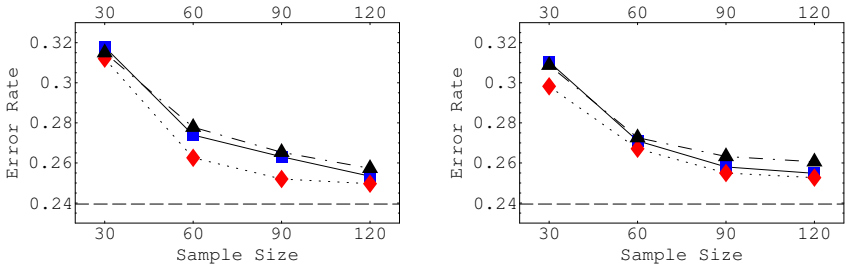
Different values of the parameters λ and η result in different numbers of windows and features being selected. The feature selection algorithm is repeated over a range of parameter values, and the feature subset that generalizes best is reported.

4 Experimental Results

4.1 Normally distributed features

This section presents the experiments used to analyze the performance of the window selection, comparing it to the original feature selection algorithm without locality and to classification without feature selection. The experiments are applied in a situation where the true distributions of the classes are known, only a number of features is relevant for classification, and the relevant features are arranged in a structured and localized manner. Two similar experiments are performed. In the first, the relevant features are arranged sequentially in a single contiguous block, and in the second, the relevant features are arranged into two disjoint contiguous blocks.

The two samples $\mathbf{A}$ and $\mathbf{B}$ in both experiments are drawn from the multivariate normal distribution, with means μ and $-\mu$, and identity covariance matrices. The 15-dimensional vector μ has 9 elements that are equal to zero and 6 non-zero elements. The non-zero elements are arranged into contiguous



Experiment I. Experiment II.

Figure 1: A comparison of the average expected error rates of the window selection algorithm (red diamond, dotted line), the feature selection algorithm (blue square, dashed line), and global discriminant analysis (black triangle, solid line). The plots show the average error rates achieved with each algorithm for sample sizes ranging from 30 to 120. The results shown were computed using $\lambda = 0.03$ in both experiments, $\eta = 0.16$ in Experiment I and $\eta = 0.02$ in Experiment II.

groups: in Experiment I, they form one group of 6 elements, in Experiment II they form two groups of 3 elements each:

$$\begin{aligned} \mu_{\mathrm{I}} &= \{1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0\}/(2\sqrt{6}) \\ \mu_{\mathrm{II}} &= \{1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0\}/(2\sqrt{6}) \end{aligned}$$

Each experiment is performed using training sets of sizes 30, 60, 90, and 120. For each size, 40 random training sets are drawn from the normal distributions described above. For each training set, the two feature and window selection algorithms are applied using different values of the energy modulation parameters λ and η. The parameter λ, which modulates the penalty on the number of selected features, takes values is the range $0.01, 0.02, \ldots, 0.2$. The parameter η, which modulates the penalty on the number of windows takes values $0.02, 0.04, \ldots, 0.2$. For each set of parameter values, the algorithm is applied 10 times with different random initializations, and the best of the 10 results is recorded.

For every subset of features selected by the two algorithms under different parameters and on different training sets, we compute the expected generalization performance of a classifier trained on that subset. A parametric classifier formed by the Fisher linear discriminant is used; it is a natural choice for this type of an experiment where the data is sampled from normal distributions with equal covariances. This choice also underlines the fact that the classifier used as a part of the feature selection algorithm need not be the same as the classifier used for the subsequent discrimination.

Let $\mathbf{w}$ be the Fisher linear discriminant of unit length, and let $\{\mathbf{x} : \mathbf{w}^T\mathbf{x} = \gamma\}$ define its decision boundary. Since the underlying normal distributions are

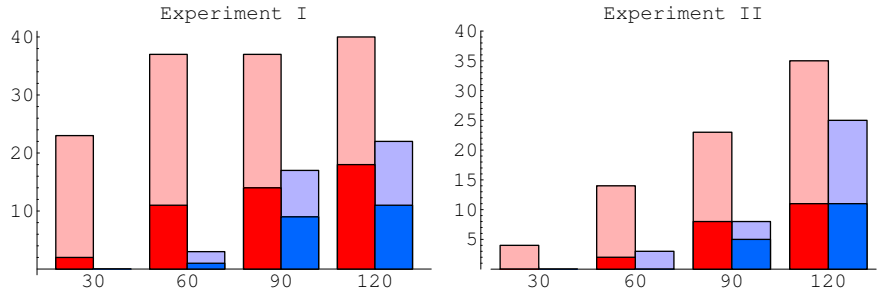


Figure 2: These charts show how many times out of 40 the two feature selection algorithms selected the correct subset of features, for a range of sample sizes. The red double bars (left) describe the window selection algorithm and the blue double bars (right) describe the feature selection algorithm. The top part of each double bar shows how many times the right feature subset was selected for *some* value of the parameters. The bottom part shows how often the right feature set was deemed optimal by cross validation.

known, the expected error rate ϵ of the classifier can be computed as

$$2\epsilon = \int_{-\infty}^{\gamma} \phi(t - \mu^T \mathbf{w})\, \mathrm{d}t + \int_{\gamma}^{\infty} \phi(t + \mu^T \mathbf{w})\, \mathrm{d}t, \tag{12}$$

where $\phi(t)$ is the standard normal probability density. Figure 1 shows the average expected error rate achieved by each feature selection approach for each sample size. For comparison, the error rate of a classifier based on the Fisher linear discriminant is also plotted.

Figure 2 shows the frequency with which both algorithms find the correct feature subset in each experiment. For each sample size, it shows the number of times the feature subset was found for some parameter value, as well as the number of times that such a parameter value yielded the smallest cross-validation error.

The window selection algorithm outperforms the original feature selection algorithm in both experiments. However, the feature selection algorithm has a lower computational complexity, because it's formulation as a linear programming problem has an $O(m+k+n)$ variables and $O(m^2+k^2+n^2)$ inequalities, while the window selection algorithm generates a problem with $O(m+k+n+N)$ variables and $O(m^2 + k^2 + n^2 + n \cdot N)$ inequalities. In these experiments, $N = n(n+1)/2$, as the window set contained all the windows possible The linear programming problems were solved using the *Sequential object-oriented simplex class library (SoPlex)*, developed by Roland Wunderling. [30].

4.2 Synthetic shape example

We use synthetic shape data to analyze the performance of the feature window selection algorithm in a shape classification problem. The classification is performed on two classes of artificial objects, which are constructed using a point

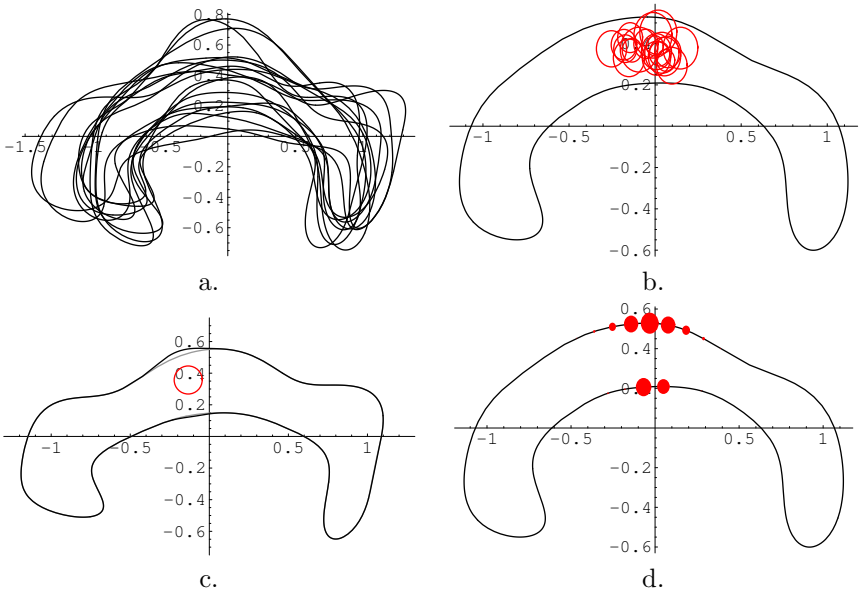


Figure 3: The generation of synthetic shapes. a. Samples from the Gaussian distribution that describes global shape variability common to both classes. b. A distribution of random tumors in relation to the average corpus callosum shape. c. The effect of a embedding a tumor into the object: the grey outline is sampled from the global shape distribution, the black outline in the deformed outline, the circle indicates the location of the warped tumor. d. The relevance measure on the features, as indicated by the relative sizes of the red dots on the boundary of the corpus callosum.

distribution model of 71 corpus callosum boundaries. The point distribution model, which was graciously provided to us by Prof. G. Gerig, consists of 64 boundary points that have been regularly sampled from Fourier contour segmentations of the corpus callosum. The segmentation technique [28] aligns the objects in space and ensures correspondence.

The two classes are generated in such a way that the differences between them are restricted to a single area of the corpus callosum. The differences between the classes are made difficult to detect by addition of a large global variability component that is common to the two classes.

The global variability component is simulated by randomly sampling members of both classes from a single Gaussian probability distribution, which is fitted to the corpus callosum point distribution model using principal component analysis, following the methodology of Cootes et al.[6]. Fig. 3a shows a few of the randomly generated outlines.

A local difference between the classes is induced by applying a randomized deformation to the outlines in Class II, using the following four-step procedure:

1. A random outline O_{rnd} is sampled from the Gaussian distribution. This is the same step used to generate the outlines in Class I.

2. A circular 'tumor', described by a point $\mathbf{c}$ and radius r, is randomly generated near a predetermined location inside of the average corpus callosum outline O_{avg}. Fig. 3b shows the outline O_{avg} and the distribution of the randomly sampled tumors.

3. A circular outward deformation field

$$\phi(\mathbf{x}) = \frac{\mathbf{x} - \mathbf{c}}{\|\mathbf{x} - \mathbf{c}\|^{\frac{3}{2}}} r \tag{13}$$

   is applied to each point of O_{avg}, yielding a new outline O_{tumor}.

4. A thin plate spline warp field, which interpolates the mapping from O_{avg} to O_{rnd}, is applied O_{tumor}, yielding the final outline O_{def}, which becomes a member of Class II. Fig. 3c shows the effect of the tumor, by superimposing the outlines O_{def} and O_{rnd}.

A single feature measuring the distance to the origin is computed at each boundary point. The approach used to produce the outlines in Class II makes it possible to measure the relevance of each feature. The relevance is computed as the average difference in distance to the origin between corresponding points on the final outlines O_{def} and intermediate outlines O_{rnd}. Fig. 3d shows the relevance of each feature in the experiment.

The two classes were generated with 100 objects in each. Another 1000 objects were generated for testing the generalization ability of the selected feature sets. The feature selection algorithms with and without locality were applied to the classes using a range of modulation parameters λ and η. The non-parametric linear classifier (8) was used for measuring the generalization ability of feature subsets.

Of the all the feature subsets computed for the different parameter values, the ones that generalized best to the test data are reported. Fig. 4a shows the best feature subset for the algorithm without locality. This subset generalizes to the test data with the error rate of 0.310. Fig 4b shows the best result for window selection, which achieves the error rate of 0.300. Without any feature selection the error rate of the classifier is 0.386. While the optimal set of features yielded by window selection includes some irrelevant features, this result is encouraging taking into account the small sample size, high dimensionality, and relatively large inter-population variability.

These results show that both feature selection methods improve the generalization ability of classifier and that while both algorithms detect some of the relevant features, the algorithm that uses locality is more accurate.

5 Conclusions

We have adapted a feature selection method from the machine learning literature to the problem of shape classification by defining an additional energy

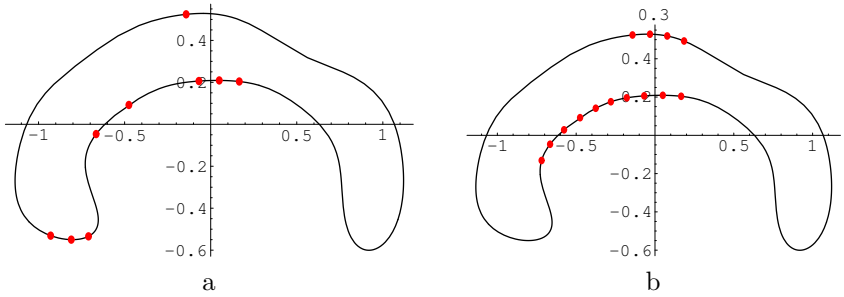


Figure 4: The best generalizing subset of features selected by the feature selection algorithm without locality (a) and by the window selection algorithm (b). The red dots on the boundary indicate selected features.

term which rewards selection of structured and localized subsets of features. We have tested the new approach on normally distributed synthetic data and showed that it performs better that the original algorithm in cases when the underlying probability densities exhibit structure and locality. We have also tested the algorithm on synthetic shape data and demonstrated that it can localize the relevant set of features in a high dimensional small sample size problem where the differences between classes are less evident than the inter-population variability. In the future, we plan to apply the method to three-dimensional clinical data as well as to other classification problems where structure and locality are present.

Acknowledgements

We thank Prof. G. Gerig and his graduate advisee Sean Ho at UNC-CH for providing the Fourier boundary models of the corpus callosum, as well as general support and advice. The corpus callosum data was originally collected at the Harvard Medical School and Brigham and Women's Hospital by M. Frumin and M.E.Shenton. We thank Prof. J.S. Marron and Prof. K.E. Muller at UNC-CH for their help and advice. We thank Prof. A. Cannon of the Department of Computer Science at Columbia University for exposing us to the feature selection literature.

The research reported in this paper was performed under support from the NCI grant P01 CA47982.

References

[1] D. W. Aha and R. L Bankert. A comparative evaluation of sequential feature selection algorithms. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 1–7. Springer-Verlag, New York, NY, 1995.

[2] F. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:567–585, 1989.

[3] Paul S. Bradley and Olvi 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.

[4] P. S. Bradley, O. L. Mangasarian, and W. N. Street. Feature selection via mathematical programming. *INFORMS Journal on Computing*, 10:209–217, 1998.

[5] 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.

[6] 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.

[7] T. Cover and J. Campenhout. The possible orderings in the measurment selection problem. *IEEE Transactions Systems, Man and Cybernetics*, 7(9):657–661, 1977.

[8] 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.

[9] 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.

[10] 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.

[11] Polina Golland, Bruce Fischl, Mona Spiridon, Nancy Kanwisher, Randy L. Buckner, Martha Elizabeth Shenton, Ron Kikinis, Anders M. Dale, and W. Eric L. Grimson. Discriminative analysis for image-based studies. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 1, pages 508–515. Springer, 2002.

[12] 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.

[13] P. Golland, W.E.L. Grimson, M.E. Shenton, and R. Kikinis. Deformation analysis for shaped based classification. In *International Conference on Information Processing in Medical Imaging*, Berlin, Germany, 2001. Springer-Verlag.

[14] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[15] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[16] 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.

[17] 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.

[18] 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.

[19] S Joshi, S Pizer, PT Fletcher, P Yushkevich, A Thall, and JS Marron. Multi-scale deformable model segmentation and statistical shape analysis using medial descriptions. *Invited submission to IEEE-TMI*, page t.b.d., 2002.

[20] 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.

[21] 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.

[22] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[23] Patrenahalli M. Narendra and Keinosuke Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 26(9):917–922, 1977.

[24] S. Pizer, D. 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.

[25] 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.

[26] 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.

[27] M. Styner. *Combined Boundary-Medial Shape Description of Variable Biological Objects*. PhD thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2001.

[28] G. Székely, A. Kelemen, Ch. Brechbühler, and G. Gerig. Segmentation of 2-D and 3-D objects from MRI volume data using constrained elastic deformations of flexible Fourier contour and surface models. *Medical Image Analysis*, 1(1):19–34, 1996.

[29] 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.

[30] Roland Wunderling. *Paralleler und Objektorientierter Simplex-Algorithmus*. PhD thesis, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, 1996. ZIB technical report TR 96-09.

[31] Paul Yushkevich, P. Thomas Fletcher, Sarang Joshi, Andrew Thall, and Stephen M. Pizer. Continuous medial representations for geometric object modeling in 2d and 3d. Technical report TR02-003, University of North Carolina, Chapel Hill, 2002.

[32] 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.