Sparse Fisher Discriminant Analysis for Computer Aided Detection

M. Murat Dundar*
Glenn Fung*
Jinbo Bi*
Sandilya Sathyakama*
Bharat Rao*

Abstract

We describe a method for sparse feature selection for a class of problems motivated by our work in Computer-Aided Detection (CAD) systems for identifying structures of interest in medical images. Typical CAD data sets for classification are large (several thousand candidates) and unbalanced (significantly fewer than 1% of the candidates are "positive"). To be accepted by physicians, CAD systems must generalize well with extremely high sensitivity and very few false positives. In order to find the features that can lead to superior generalization, researchers typically generate a large number of experimental features for each candidate. The reason for such a large number of features is that there are no definitive methods for capturing the shape and image-based characteristics that correspond to the diagnostic features used by physicians to identify structures of interest in the image - for example, cancerous polyps in a CT (computed tomography) volume of a patient's colon. Thus several (100+) shape, texture, and intensity based features may be generated for each candidate at various levels of resolution. We propose a sparse formulation for Fisher Linear Discriminant (FLD) that scales well to large datasets; our method inherits all the desirable properties of FLD, while improving on handling large numbers of irrelevant and redundant features. We demonstrate that our sparse FLD formulation outperforms conventional FLD and two other methods for feature selection from the literature on both an artificial dataset and a real-world Colon CAD dataset.

Keywords: fisher linear discriminant, sparse formulation, feature selection

1 Problem Specification.

Over the last decade, Computer-Aided Detection (CAD) systems have moved from the sole realm of academic publications, to robust commercial systems that are used by physicians in their clinical practice to help detect early cancer from medical images. The growth has been fueled by the Food and Drug Administration's (FDA) decision to grant approval in 1998 for a CAD system that detected breast cancer lesions from mammograms (scanned x-ray images) [1]. Since then a number of CAD systems have received FDA approval. Virtually all these commercial CAD systems, focus on detection (or more recently diagnosis [2]) of breast cancer lesions for mammography.

The typical workflow for a CAD system when used to identify structures in a new patient image is:

1. *Identify candidate structures in the image*: Most medical images, particularly image volumes generated by high-resolution computed tomography (CT), are very large. Typically, a very efficient image processing algorithm considers each pixel (or voxel) in the image as a potential candidate "seed", and selects a small fraction of the seeds as candidates. Even this small fraction is necessarily very large, in order to maintain high sensitivity. High sensitivity (ideally very close to 100%) is essential, because any cancers missed at this stage can never be found by the CAD system.

2. *Extract features for each candidate*: Unlike the previous step, the image processing algorithms to extract the features may be compuationally expensive. Thus, sparse feature selection (while building the classifier) is important in order to ensure a relatively small number of features in the deployed CAD system.

3. *Classify candidates as positive or negative*: A previously-trained classifier is used to label each candidate.

4. *Display a positive candidates*: Typically, the digitized image is displayed with marks for inspection by the physician.

*Computer Aided Diagnosis and Therapy, Siemens Medical Solutions Inc,USA

Typically, CAD systems are used as "second readers" – the physician views the image to identify potential cancers (the "first read"), and then reviews the CAD marks to determine if any additional cancers can be found. In order to receive clinical acceptance and to actually be used in the daily practice of a physician, it is immediately obvious that CAD systems must be efficient (for instance, completing the detections in the minutes taken by the physician during the "first read") and have very high sensitivity (the whole point of CAD is to boost the physician's sensitivity, which is already fairly high – 80%-90% for colon cancer – to the high 90's). What is not immediately obvious is that these systems must also have extremely high specificity, at best, only a few (5 or fewer) false marks per image is acceptabe. There are a number of reasons for this, including the risk of introducing unnecessary biopsies (a concern of the FDA before clinical approval is granted), liability issues, but most of all, every false mark increases the time needed to review the image, and this is particularly unacceptable in the US because of the financial pressures on physicans.

It is immediately obvious that the same CAD paradigm can be applied to detect not only breast cancer but other cancers, and also to analyze imaging modalities that can provide more resolution than a 2-dimensional x-ray image. The applications that are likely to constitute the next wave of CAD systems in clinical are detection of colon cancer and of lung cancer from 3-dimensional computed tomography (CT) volumes. As with breast cancer, to be successful ColonCAD and LungCAD systems must be efficient, have extremely high sensitivity, and introduce very few false positives per volume. As with BreastCAD, this performance must be demonstrated to the FDA in a clinical trial, on new (as yet unseen) CT volumes. The choice of features and the classifier in the CAD system is critical to success.

The key requirement of the CAD classifier, is its ability to generalize well. Namely, it should correctly label as yet unseen datasets. Although generalization is a fundamental problem of machine learning commonly encountered in all domains, the inherent nature of the data collection and feature extraction process makes this problem more challenging to deal with in CAD algorithms. The typical process for off-line training of the classifier used in Step 3 of the CAD workflow, is to gather a database of patient images, within which structures have been labelled by a panel of expert physicians, generate candidates from these images with high sensitivity (i.e., ideally there should exist candidates corresponding to each "positive" label), generate a features for each candidate, and train the classifier. The biggest problem is the choice of features, both in the training set, and in the deployed classifier.

Physicians detect cancers by visually extracting shape and texture based features, that are often qualitative rather than quantitative from the images (hereafter, "image" and "volume" are used interchangably in this document). However, there are usually no definitive image-processing algorithms that exactly correspond to the precise, often subtle, features used intuitively by physicians. To achieve high sensitivity and specificity, CAD researchers must necessarily consider a very large number of experimental image processing features. Therefore, a typical training dataset for a CAD classifier is extremely unbalanced (significantly less than 1% of the candidates are positive), contains a very large number of candidates (several thousand), each described by many features (100+), most of which redundant and irrelevant. Note that because of the efficiency requirement of the deployed CAD system, as few features as possible should be used by the CAD classifier.

The rest of this paper is organized as follows. In the next section, we discuss the need for a linear classifier and briefly review the Fisher Linear Discriminant (FLD). We also introduce our notion of spare FLD, where we seek to eliminate the redundant and irrelevant features from the original training set using a wrapper approach. In Section 3 we review the concept and formulation of FLD. In Section 4 we modify the conventional FLD problem so as to achieve sparseness and propose an iterative feature selection algorithm based on our the sparse formulation. Finally we present experimental results on an artificial dataset and a ColonCAD dataset, and compare our approach with conventional FLD and also with two well-known methods from the literature for feature selection. We empirically demonstrate that unlike some of the conventional methods for feature selection, our sparse FLD formulation not only inherits all the desirable generalization properties of FLD, it is especially efficient on problems characterized by large datasets with many features,

2 Feature selection with FLD

As discussed, a typical training dataset for a CAD classifier is large, extremely unbalanced, and has many features (100+), most of which redundant and irrelevant. Because of the efficiency requirement of the deployed CAD system, as few features as possible should be used by the CAD classifier.

A large number of features provides more control over the discriminant function. However, even with our "large" training sample, the high-dimensional feature space is mostly empty [5]. This allows us to find many classifiers that perform well on the training data, but it is well-known that few of these will generalize well. This is particularly true of nonlinear classifiers that represent more complex discriminant functions. Furthermore, many computationally expensive nonlinear classification algorithms (e.g. nonlinear SVM, neural networks, kernel-based algorithms) do not scale well to large datasets. When the potential pitfalls of designing a classifier and the characteristics of the data are considered, it appears safer to train a CAD system with a linear classifier. This is empirically demonstrated in our previous study [6] where we compare the generalization capability of some linear and nonlinear classification algorithms on a CAD dataset.

Fisher Linear Discriminant (FLD) [7] is a well-known classification method that projects high-dimensional data onto a line and performs classification in this one dimensional space. This projection is obtained by maximizing the ratio of between and within class scatter matrices – the so called *Rayleigh quotient*. As a linear classifier it is rather robust against feature redundancy and noise and has an order of complexity $O\left(ld^2\right)$ (l is the number of training samples in the dataset and d is the number of features in the feature set). This linear dependence on data size permits multiple fast runs of the algorithm even with large data sets. In addition to these properties, FLD is closely connected to the Bayes classifier. More specifically, when the classes are normally distributed with equal covariance matrices, the discriminants obtained through FLD and the Bayes classifier are in the same direction and with the choice of an appropriate threshold Bayes error can be achieved. Although it relies on heavy assumptions which are not true in most practical cases, FLD has proven very powerful in a wide variety of challenging real-world applications.

In this study we propose a sparse formulation of FLD where we seek to eliminate the irrelevant and redundant features from the original dataset within a *wrapper* framework [8]. To achieve sparseness, earlier studies focused on direct optimization of an objective function consisting of two terms: the goodness of fit and the regularization term. In order to avoid overfitting by excessively maximizing the goodness of fit, a regularization term commonly expressed as $\ell_0 - norm$ [9], [10] or $\ell_1 - norm$ [11], [12] of the discriminant vector is added to the objective function. Optimization of this objective function generates sparse solutions, i.e. a solution that depends only on a subset of the features.

The proposed approach is similar in nature to the sparse formulation introduced in [11] for Kernel Fisher Discriminant with quadratic loss and linear regularizer. The technique in [11] is well-formulated providing some algorithmic advantages and reveals some pleasing theoretical connections of Fisher Discriminant with Support Vector Machines and Relevance Vector Machines [13]. Although this technique scales well to high-dimensional feature sets the constraints on each sample in the training set does not permit the algorithm to scale well to very large datasets.

Our approach achieve sparseness by introducing regularity constraints into the problem of finding FLD. Since we maintain the original formulation of FLD as we introduce the regularization constraints, the proposed technique can scale to very large datasets (on the order of hundred thousand samples). Casting this problem as a biconvex programming problem provides us a more direct way of controlling the size of the feature subset selected. This problem is iteratively solved and once the algorithm stops the nonzero elements of the solution indicates features that are relevant to classification task at hand, and their value quantifies the degree of this relevancy. The proposed algorithm inherits all desirable characteristics of FLD while improving on handling large number of redundant and irrelevant features. This makes the algorithm numerically more stable and improve its prediction performance.

3 Fisher's Linear Discriminant

Let $X_i \in R^{d\times l}$ be a matrix containing the l training data points on d-dimensional space and l_i the number of labeled samples for class w_i, $i \in \{\pm\}$. FLD is the projection α, which maximizes,

$$J\left(\alpha\right) = \frac{\alpha^T S_B \alpha}{\alpha^T S_W \alpha} \tag{3.1}$$

where

$$S_B = \left(m_+ - m_-\right)\left(m_+ - m_-\right)^T$$

$$S_W = \sum_{i\in\{\pm\}} \frac{1}{l_i}\left(X_i - m_i e_{l_i}^T\right)\left(X_i - m_i e_{l_i}^T\right)^T$$

are the between and within class scatter matrices respectively and

$$m_i = \frac{1}{l_i} X_i e_{l_i}$$

is the mean of class w_i and e_{l_i} is an l_i dimensional vector of ones.

Transforming the above problem into a convex quadratic programming problem provides us some algorithmic advantages. First notice that if α is a solution to (3.1), then so is any scalar multiple of it. Therefore to avoid multiplicity of solutions, we impose the constraint $\alpha^T S_B \alpha = b^2$, which is equivalent to $\alpha^T (m_+ - m_-) = b$ where b is some arbitrary positive scalar. Then the optimization problem of (3.1) becomes,

$$\begin{array}{lll} \mathit{Problem}\ 1: & \min_{\alpha \in R^d} & \alpha^T S_W \alpha \\ & s.t. & \alpha^T (m_+ - m_-) = b \end{array}$$

For binary classification problems the solution of this problem is $\alpha^* = \frac{b S_W^{-1}(m_+ - m_-)}{(m_+ - m_-)^T S_W^{-1}(m_+ - m_-)}$. Note that each element of the discriminant vector is a weighted sum of the difference between class mean vectors where the weighting coefficients are rows of $\frac{b S_W^{-1}}{(m_+ - m_-)^T S_W^{-1}(m_+ - m_-)}$. According to this expansion since S_W^{-1} is positive definite unless the difference of the class means along a given feature is zero all features contributes to the final discriminant.

If a given feature in the training set is redundant, its contribution to the final discriminant would be artificial and not desirable. As a linear classifier FLD is well-suited to handle features of this sort provided that they do not dominate the feature set, that is, the ratio of redundant to relevant features is not significant. Although the contribution of a single redundant feature to the final discriminant would be negligible when several of these features are available at the same time, the overall impact could be quite significant leading to poor prediction accuracy. Apart from this impact, in the context of FLD these undesirable features also pose numerical constraints on the computation of S_W^{-1} especially when the number of training samples is limited. Indeed, when the number of features, d is higher than the number of training samples, l, S_W becomes ill-conditioned and its inverse does not exist. Hence eliminating the irrelevant and redundant features may provide a two-fold boost on the performance.

In what follows we propose a sparse formulation of FLD. The proposed approach incorporates a regularization constraint on the conventional algorithm and seeks to eliminate those features with limited impact on the objective function.

4 Sparse Fisher Discriminant Analysis

Blindly fitting classifiers without appropriate regularization conditions is guaranteed to give badly over-fitted models. Methods for controlling model complexity are essential in modern data analysis. Especially when the number of features available is large, an appropriate regularization can dramatically reduce the dimensionality and produces better generalization performance which is supported by learning theory [16]. For linear models of the form $\alpha^T \mathbf{x}$ as considered here, well-established regularization conditions include the 2-norm penalty and 1-norm penalty on the weight vector α. The generic regularized model fitting problem is written as:

$$\hat{f} = \min_f \left(\text{error}(f) + \lambda P(f) \right). \tag{4.2}$$

where λ is called the regularization parameter. Due to the sparsity and favorable computational properties of the 1-norm penalty, the 1-norm regularization has drawn a lot of attention in the statistical community. We thus adopt the 1-norm penalty $P(f) = \sum |\alpha_i|$ in our sparse FLD formulation, which generates sparser feature subsets than 2-norm penalty. As analyzed in [16], the regularized model fitting formulation 4.2 has an important equivalent formulation as

$$\hat{f} = \min_f \{\text{error}(f),\ \text{subj. to:} P(f) \leq \gamma\}. \tag{4.3}$$

where the parameter γ plays a similar role to the regularization parameter λ in (4.2) to trade off between the training error and the penalty term.

If we require α to be nonnegative, the 1-norm of α can be calculated as $\alpha^T e_l$. We thus obtain the following optimization problem.

With the new constraints Problem 1 can be updated as follows,

$$\begin{array}{lll} \mathit{Problem}\ 2: & \min_{\alpha \in R^d} & \alpha^T S_W \alpha \\ & s.t. & \alpha^T (m_+ - m_-) = b \\ & & \alpha^T e_l \leq \gamma,\ \alpha \geq 0 \end{array}$$

We denote the feasible set associated with Problem 1 by $\Omega_1 = \left\{\alpha \in R_d,\ \alpha^T (m_+ - m_-) = b\right\}$ and that associated with Problem 2 by $\Omega_2 = \left\{\alpha \in R_d,\ \alpha^T (m_+ - m_-) = b,\ \alpha^T e_l \leq \gamma,\ \alpha \geq 0\right\}$ and observe that $\Omega_2 \subset \Omega_1$. Then we define $\delta_{max} = max_i \frac{b}{(m_+ - m_-)_i}$ and $\delta_{min} = min_i \frac{b}{(m_+ - m_-)_i}$ where $i = \{1, \ldots, d\}$. The set Ω_2 is empty whenever $\delta_{max} < 0$ or $\delta_{min} > \gamma$. In addition to the feasibility constraints $\gamma < \delta_{max}$ should hold in order to achieve a sparse solution. In what follows we introduce a linear transformation which will ensure $\delta_{max} > 0$ and standardize the sparsity constraint.

For the sake of simplicity and without loss of generality we assume that S_W is a diagonal matrix with elements $\lambda_i,\ i = 1, \ldots, d$ where λ_i are the eigenvalues of S_W. Under this scenario the solution to Problem 1 is $\alpha^* = \bar{b} \left[\frac{(m_+ - m_-)_1}{\lambda_1}, \ldots, \frac{(m_+ - m_-)_d}{\lambda_d} \right]^T$ where $\bar{b} = \frac{b}{\sum_{i \in \{\pm\}} \frac{(m_+ - m_-)_i^2}{\lambda_i}}$. Next we define a linear transformation $D = diag\left(\alpha_1^*, \ldots, \alpha_d^*\right) = \bar{b}\, diag\left(\frac{(m_+ - m_-)_1}{\lambda 1}, \ldots, \frac{(m_+ - m_-)_d}{\lambda d} \right)$ such that $x \mapsto Dx$

where *diag* indicates a diagonal matrix. With this transformation Problem 2 takes the following form,

$$\begin{array}{lll} Problem\ 3: & \min_{\alpha \in R^d} & \alpha^T D S_W D \alpha \\ & s.t. & \alpha^T D \left(m_+ - m_-\right) = b \\ & & \alpha^T e_l \leq \gamma,\ \alpha \geq 0 \end{array}$$

We redefine $\bar{\delta}_{max} = max_i \frac{b\lambda_i}{\bar{b}(m_+ - m_-)_i^2}$ and $\bar{\delta}_{min} = min_i \frac{b\lambda_i}{\bar{b}(m_+ - m_-)_i^2}$ where $i = \{1, \ldots, d\}$. Note that both $\bar{\delta}_{min}$ and $\bar{\delta}_{max}$ are nonnegative and hence both feasibility constraints are satisfied when $\gamma > \bar{\delta}_{min}$. For $\gamma > d$ the globally optimum solution α^* to Problem 3 is $\alpha^* = [1, \ldots, 1]^T$, i.e nonsparse solution. For $\gamma < d$ sparse solutions can be obtained. Unlike Problem 2 where the upper bound on γ depends on mean vectors here the upper bound is d, i.e. the number of features.

The above sparse formulation is indeed a biconvex programming problem.

$$\begin{array}{lll} Problem\ 4: & \min_{\alpha,\, a \in R^d} & \alpha^T \left(S_W * \left(a a^T\right)\right) \alpha \\ & s.t. & \alpha^T \left(\left(m_+ - m_-\right) * a\right) = b \\ & & \alpha^T e_l \leq \gamma,\ \alpha \geq 0 \end{array}$$

We first initialize $\alpha = [1, \ldots, 1]^T$ and solve for a^*, i.e. the solution to Problem 1, then we fix a^* and solve for α^*, i.e. the solution to Problem 3.

5 The Iterative Feature Selection Algorithm

Successive feature elimination can be obtained by iteratively solving the above biconvex programming problem.

(0) Set $\alpha^0 = e_n$, $d^0 = d$, $\gamma << d$

For each iteration i do the following:

(i) Select the d^i features with α^i_j values greater than ϵ, $d^i \leq d^{i-1}$.

(ii) Calculate the class scatter matrices and means in the $d^i - dimensional$ feature space.

(iii) Solve Problem 4 to obtain a^i.

(iv) Fix a to a^i and update the class scatter matrices and means.

(v) Solve Problem 4 to obtain α^i.

Stop when all α^i_j, for $j = 1, 2, \ldots, d^i$ are greater than $\epsilon = 1e - 16$.

Since at each iteration we truncate α the above algorithm is not guaranteed to converge. However at any iteration i when $d^i \leq \gamma$ no sparseness would be achieved and hence all α^i_j would be equal to one. Therefore the algorithm is guaranteed to stop at the latest when $d^i \leq \gamma$.

6 Experimental Results and Discussion

6.1 A Toy Example This experiment is adapted from [14]. Using an artificial data we demonstrate that the performance of conventional FLD suffers from the presence of too many irrelevant features whereas the proposed sparse approach produces a better prediction accuracy by successfully handling these irrelevant features.

The probability of $y = 1$ or $y = -1$ is equal. The first three features x_1, x_2, x_3 are drawn as $x_i = yN\left(i, 5\right)$. Note that only one of these features is relevant for discriminating one class from the other, the other two are redundant. The rest of the features are drawn as $x_i = N\left(0, 20\right)$. Note that these features are noise. The noise features are added to the feature set one by one allowing us to observe the gradual change in the prediction capability of both approaches.

We initialize $d = 3$, i.e. start with the first three features and proceed as follows. We generate 200 samples for training and 1000 samples for testing. Then we train and test both approaches and record the corresponding prediction errors. Next we increase d by one and repeat the above procedure until we reach $d = 20$. For the proposed approach we select the best two features. The error bars in Figure 1 are obtained by repeating the above process 100 times for each d each time using a different training and testing set.

Looking at the results, at $d = 3$ with two redundant features the prediction accuracy of the conventional FLD is decent. With the same two redundant features at $d = 3$ the standard deviation in prediction error is slightly smaller with the proposed formulation indicating the elimination of one or both of the redundant features. As d gets larger and noise features are added to the feature set the performance of the conventional FLD deteriorates significantly whereas the average prediction error for the proposed formulation remains around its initial level with some increase in the standard deviation. Also 90% of the time the proposed formulation selects feature two and three together. These are the two most powerful features in the set.

6.2 Example 2: Colon Cancer

6.2.1 Data Sources and Domain Description Colorectal cancer is the third most common cancer in both men and women. It is estimated that in 2004, nearly 147,000 cases of colon and rectal cancer will be diagnosed in the US, and more than 56,730 people would die from colon cancer [3]. While there is wide consensus that screening patients is effective in decreasing advanced disease, only 44% of the eligible population undergoes any colorectal cancer screening.

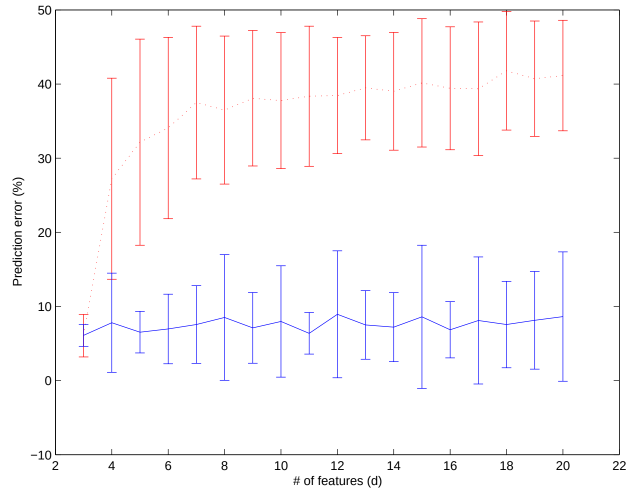

Figure 1: Testing Error vs l for the Artificial Data. Full dimensionality and two-dimensional feature subset compared. The dotted curve corresponds to Conventional FLD, the solid curve corresponds to proposed sparse appraoch

There are many factors for this, Multiple reasons have been identified for non-compliance, key being: patient comfort, bowel preparation and cost. Non-invasive virtual colonoscopy derived from computer tomographic (CT) images of the colon holds great promise as a screening method for colorectal cancer, particularly if CAD tools are developed to facilitate the efficiency of radiologists' efforts in detecting lesions. In over 90% of the cases colon cancer progressed rapidly is from local (polyp adenomas) to advanced stages (colorectal cancer), which has very poor survival rates. However, identifying (and removing) lesions (polyp) when still in a local stage of the disease, has very high survival rates [4], thus illustrating the critical need for early diagnosis.

The database of high-resolution CT images used in this study were obtained from NYU Medical Center, Cleveland Clinic Foundation, and two EU sites in Vienna and Belgium. The 163 patients were randomly partitioned into two groups: training (n=96) and test (n=67). The test group was sequestered and only used to evaluate the performance of the final system.

Training Data Patient and Polyp Info: There were 96 patients with 187 volumes. A total of 76 polyps were identified in this set with a total number of 9830 candidates.

Testing Data Patient and Polyp Info: There were 67 patients with 133 volumes. A total of 53 polyps were identified in this set with a total number of 6616 candidates. A combined total of 207 features are extracted for each candidate by three imaging scientists.

6.2.2 Feature Selection and Classification: In this experiment we consider three feature selection algorithms in a wrapper framework and compare their prediction performance on the Colon Dataset. These techniques are namely, the sparse formulation proposed in this study (SFLD), the sparse formulation introduced in [11] for Kernel Fisher Discriminant with linear loss and linear regularizer (SKFD) and a greedy sequential forward-backward feature selection algorithm [15] implemented with FLD (GFLD). In what follows we present a brief overview of these algorithms and discuss possible design issues.

Sparse Fisher Linear Discriminant (SFLD): The choice of γ plays an important role on the generalization performance of our algorithm. It regularizes the algorithm by seeking a balance between the "goodness of fit", i.e. *Rayleigh Quotient* and the number of features used to achieve this performance.

We estimate the value of this parameter by cross validation. We adopt Leave-One-Patient-Out (LOPO) cross validation approach. In this scheme, we leave-out both views, i.e. the supine and the prone views, of one patient from the training data. The classifier is trained using the patients from the remaning set, and tested on both views of the "left-out" patient. LOPO is superior to other cross-validation metrics such as leave-one-volume-out, leave-one-polyp-out or k-fold cross-validation because it simulates the actual use, wherein the CAD system processes both volumes for a new patient. For instance, with any of the above alternative methods, if a polyp is visible in both views, the corresponding candidates could be assigned to different folds; thus a classifier may be trained and tested on the same polyp (albeit in different views).

In order to find the optimum value of γ, we run the algorithm in Section 5 for varying sizes of $\gamma \in [1\ d]$. For each value of γ we obtain the Receiver Operating Characteristics (ROC) curve by evaluating the Leave One Patient Out (LOPO) Cross Validation performance of the algorithm and then compute the area under this curve. We choose the optimum value of γ as the value that results in the largest area.

Kernel Fisher Discriminant with linear loss and linear regularizer (SKFD): In this approach there is a set of contraints for every data point on the training set which leads to large optimization problems. In order to alleviate the computational burden on mathematical programming formulation for this approach we choose to use Laplacian models for both the loss function and the regularizer as suggested in [11]. This choice leads to linear programming formulation instead of the more conventional and more computationally expensive quadratic programming formulation that is obtained when a gaussian model is assumed for both the loss function and the regularizer.

The linear programming formulation used in this experiment is

$$
\begin{aligned}
& \min_{(\alpha,\beta,\epsilon)\in R^{n+1+m}} && \nu\|\epsilon\|_1 + \|\alpha\|_1 \\
(6.4) \qquad & \text{s.t.} && A\alpha + \beta = y + \epsilon \\
& && e_i'\epsilon_i = 0 \text{ for } i \in \{+\} \\
& && e_i'\epsilon_i = 0 \text{ for } i \in \{-\}
\end{aligned}
$$

Where $e_\pm$ is vector of ones of size the number of points in class ±. The final classfier for an unseen data point x is given by $sign(\alpha^T x - \beta)$. The regularization parameter ν is estimated by LOPO.

Greedy sequential forward-backward feature selection algorithm with FLD (GFLD):

This approach starts with an empty subset and performs a forward selection succeeded by a backward attempt to eliminate a feature from the subset. During each iteration of the forward selection exactly one feature is added to the feature subset. To determine which feature to add, the algorithm tentatively adds to the candidate feature subset one feature that is not already selected and tests the LOPO performance of a classifier built on the tentative feature subset. The feature that results in the largest area under the ROC curve is added to the feature subset. During each iteration of the backward elimination the algorithm attempts to eliminate the feature that results in the largest ROC area gain. This process goes on until no or negligible improvement is gained. In this study the algorithm stops when the increase on the ROC area after a forward selection is less than 0.005. A total of 17 features is selected before this constraint is met.

6.3 Results and Discussion: Even though we choose the computationally least expensive model for SKFD this approach failed to run with the original training set. Thus we were forced to run SKFD on a smaller subset of the training dataset where we included all the positive candidates and a random subset of size 1000 of the negative candidates. The 5 algorithms we ran were

1. SFLD on the original training set.

2. GFLD on the original training set.

3. Conventional on the original training set.

4. SKFD on the subset training set.

5. SFLK on the subset training set (denoted as SFLD-sub).

The ROC curves in Figure 2 demonstrates the LOPO performance of the each algorithm and those in Figure 3 show the performance on the test data set. Table 1 shows the number of features selected (d), the area of the ROC curve scaled by 100 (Area) and the sensitivity corresponding to 90% specificity (Sens) for all algorithms considered in this study.

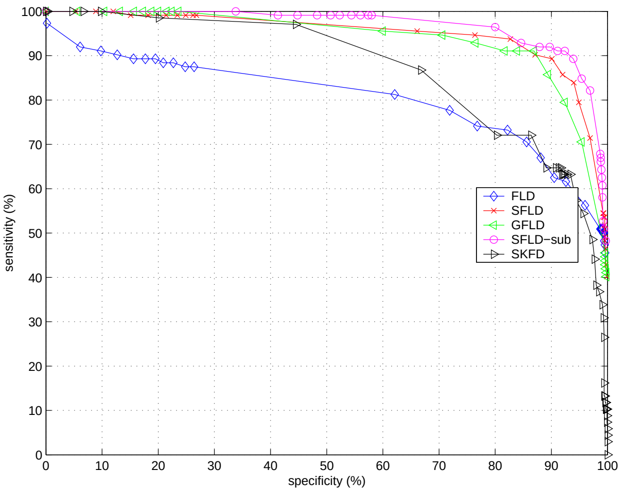

Figure 2: ROC curves for Training Results (LOPO results)

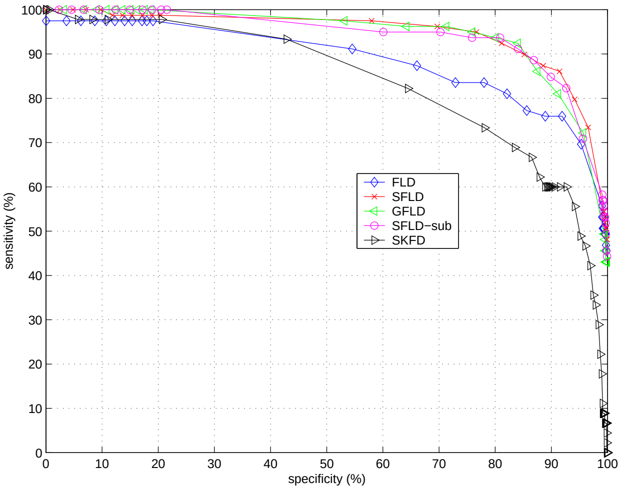

Figure 3: ROC curves for Testing Results

Table 1: The number of features selected (d), the area of the ROC curve scaled by 100 (Area) and the sensitivity corresponding to 90% specificity (Sens) is shown for all algorithms considered in this study. The values in parenthesis show the corresponding values for the testing results.

Algorithm	d	Area	Sens (%)
SFLD	25	94.8 (94.9)	89 (87)
SFLD-sub	17	94.7 (94.1)	92 (85)
GFLD	17	94.3 (94.7)	85 (83)
SKFD	18	88.0 (82.0)	65 (60)
FLD	207	80.3 (89.1)	63 (77)

These results show that Sparse (SFLD) and SFLD-sub clearly outperform the greedy and conventional FLD and SKFD both on the training and testing datasets. Although SFLD-sub performs better than SFLD on the training data, SFLD generalizes slightly better on the testing data. This is not surprising because SFLD-sub uses a subset of the original training data. GFLD performs almost equally well with SFLD-sub and SFLD algorithms but the difference is hidden in the computational cost required to select the features in GFLD. The computational cost of GFLD is proportional to d^3 whereas that of SFLD is proportional to d^2.

7 Conclusions

In this study we proposed a sparse formulation of famous Fisher Linear Discriminant and applied this technique to a Colon dataset. Experimental results favor the proposed algorithm over two other feature selection/regularization techniques implemented in the FLD framework both in terms of prediction accuracy and the computational cost fir large data sets. Future study will focus on obtaining sparse solutions in an iterative scheme without truncating the discriminant vector which will in turn guarantee convergence.

References

[1] J. Roehrig, *The Promise of CAD in Digital Mammography*, European Journal of Radiology, 31 (1999), pp. 35-39.

[2] S. Buchbinder, I. Leichter, R. Lederman, B. Novak, P. Bamberger, M. Sklair-Levy, G. Yarmish, and S. Fields *Computer-aided Classification of BI-RADS Category 3 Breast Lesions1*, Radiology, 230 (2004), pp. 820-823.

[3] D. Jemal, R. Tiwari, T. Murray, A. Ghafoor, A. Saumuels, E. Ward, E. Feuer, M. Thun *Cancer Statistics 2004*, CA Cancer J. Clin., 54 (2004), pp. 8–29.

[4] L. Bogoni, P. Cathier, M. Dundar, A. Jerebko, S. Lakare, J. Liang, S. Periaswamy, M. Baker and M. Macari *CAD for Colonography: A tool to address a growing need*, To Appear in British Journal of Radiology.

[5] C. Lee and D. Landgrebe *Analyzing High Dimensional Multispectral Data*, IEEE Transactions on Geoscience and Remote Sensing, 31 (1993), pp 792–800.

[6] M. Dundar, G. Fung, L. Bogoni, M. Macari, A. Megibow, B. Rao *A Methodology for Training and Validating a CAD System and Potential Pitfalls*, In Proc. CARS, (2004), pp. 1010–1014.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Progress, San Diego, CA, 1990.

[8] G. John, R. Kohavi, K. Pfleger, *Irrelevant Features and the Subset Selection Problem*, In Proc. of ICML, (1994).

[9] J. Weston, A. Elisseeff, B. Scholkopf, M. Tipping *Use of the Zero-Norm with Linear Models and Kernel Methods*, Journal of Machine Learning Research, 3 (2003), pp. 1439–1461.

[10] P. Bradley and O. Mangasarian *Feature Selection via Concave Minimization and Support Vector Machines*, Proc. of 15th International Conference on Machine Learning, (1998), pp. 82–90.

[11] S. Mika, G. Ratsch, K. Muller *A Mathematical Programming Approach to the Kernel Fisher Algorithm*, Proc. NIPS 13, (2001), pp. 591-597.

[12] J. Bi, K. Bennett, M. Embrechts, C. Breneman, M. Song *Dimensionality Reduction via Sparse Support Vector Machines*, Journal of Machine Learning Research, 3 (2003), pp. 1229–1243.

[13] M. Tipping *The relevance vector machine*, Advances in Neural Information Processing Systems, pp. 652–668, MIT Press, 2000.

[14] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio and V. Vapnik, *Feature Selection for SVMs*, Advances in Neural Information Processing Systems., 13, pp. 668–674.

[15] J. Kittler, *Feature Set Search Algorithms*, Pattern Recognition and Signal Processing, Sijhoff and Noordhoff, the Netherlands, 1978.

[16] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 1998.