

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. 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.

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).

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 interchangeably 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 are redundant and irrelevant.

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 [3]. 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 [4] where we compare the generalization capability of some linear and nonlinear classification al-

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

gorithms on a CAD dataset.

Fisher Linear Discriminant (FLD) [5] 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 $O(ld^2)$ (l is the number of training samples in the dataset and d is the number of features in the feature set).

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 [6]. 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 ℓ_0 – *norm* [7], [8] or ℓ_1 – *norm* [9], [10] 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.

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.

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 experi-

mental 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.

2 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,

$$(2.1) \quad J(\alpha) = \frac{\alpha^T S_B \alpha}{\alpha^T S_W \alpha}$$

where

$$S_B = (m_+ - m_-)(m_+ - m_-)^T$$

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

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 (2.1), 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 (2.1) becomes,

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

For binary classification problems the solution of this problem is $\alpha^* = \frac{b S_W^{-1} (m_+ - m_-)}{(m_+ - m_-)^T S_W^{-1} (m_+ - m_-)}$. 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.

3 Sparse Fisher Discriminant Analysis

If we require α to be nonnegative, the 1-norm of α can be calculated as $\alpha^T e_l$. With the new constraints Problem 1 can be updated as follows,

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

We denote the feasible set associated with Problem 1 by $\Omega_1 = \{\alpha \in R^d, \alpha^T (m_+ - m_-) = b\}$ and that associated with Problem 2 by $\Omega_2 = \{\alpha \in R^d, \alpha^T (m_+ - m_-) = b, \alpha^T e_l \leq \gamma, \alpha \geq 0\}$ 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 $i = \{1, \dots, d\}$. 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.

We define a linear transformation such that $x \mapsto Dx$. With this transformation Problem 2 takes the following form,

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

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 $\alpha^* = [1, \dots, 1]^T$, 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.

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

where $*$ is an element-wise product. We first initialize $\alpha = [1, \dots, 1]^T$ 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.

4 The Iterative Feature Selection Algorithm

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

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

For each iteration i do the following:

- (i) Select the d^i features with α_j^i values greater than $\epsilon, d^i \leq d^{i-1}$.
- (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_j^i , for $j = 1, 2, \dots, d^i$ 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_j^i would be equal to one. Therefore the algorithm is guaranteed to stop at the latest when $d^i \leq \gamma$.

5 Experimental Results and Discussion

5.1 A Toy Example This experiment is adapted from [11]. The probability of $y = 1$ or $y = -1$ is equal. The first three features x_1, x_2, x_3 are drawn as $x_i = yN(i, 5)$. 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 $x_i = N(0, 20)$. 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, 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.

5.2 Example 2: Colon Cancer

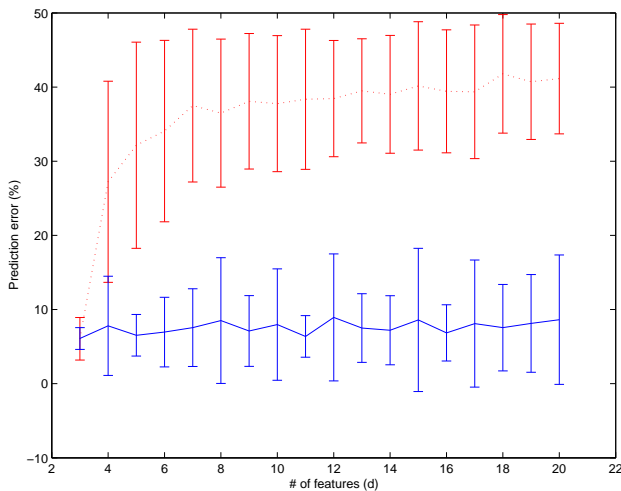


Figure 1: Testing Error vs l for the Artificial Data. Full dimensionality and two-dimensional feature subset compared. The dotted curve corresponds to Conventional FLD, the solid curve corresponds to proposed sparse approach

5.2.1 Data Sources and Domain Description

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.

5.2.2 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 [9] for Kernel Fisher Discriminant with linear loss and linear regularizer (SKFD) and a greedy sequential forward-backward feature selection algorithm [12] implemented with FLD (GFLD).

5.3 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.

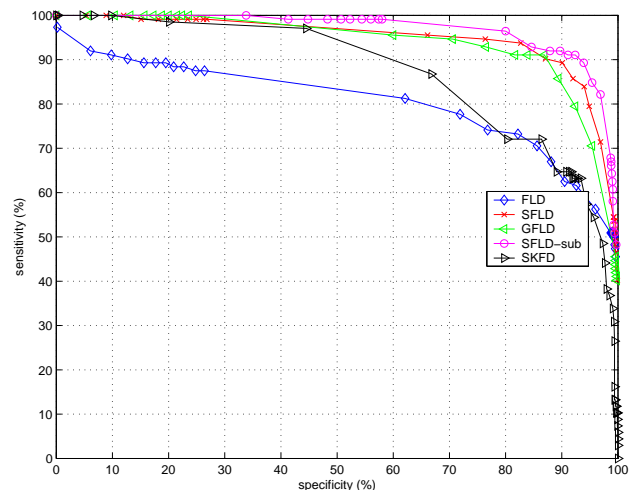


Figure 2: ROC curves for Training Results (LOPO results)

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

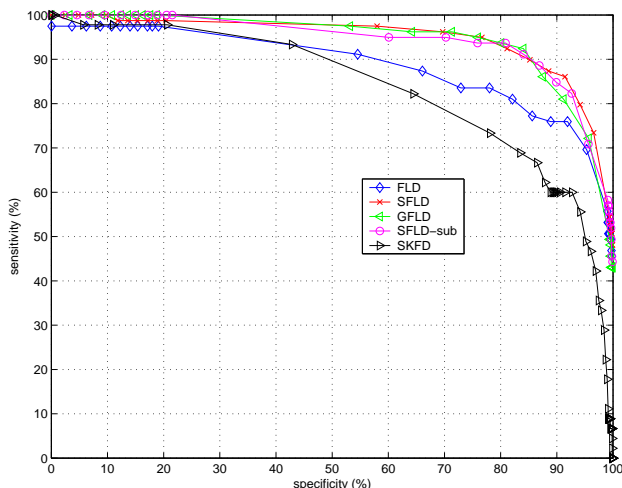


Figure 3: ROC curves for Testing Results

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)

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 .

6 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 for 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] C. Lee and D. Landgrebe *Analyzing High Dimensional Multispectral Data*, IEEE Transactions on Geoscience and Remote Sensing, 31 (1993), pp 792–800.
- [4] 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.
- [5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Progress, San Diego, CA, 1990.
- [6] G. John, R. Kohavi, K. Pfleger, *Irrelevant Features and the Subset Selection Problem*, In Proc. of ICML, (1994).
- [7] 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.
- [8] 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.
- [9] S. Mika, G. Ratsch, K. Muller *A Mathematical Programming Approach to the Kernel Fisher Algorithm*, Proc. NIPS 13, (2001), pp. 591-597.
- [10] 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.
- [11] 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.
- [12] J. Kittler, *Feature Set Search Algorithms*, Pattern Recognition and Signal Processing, Sijhoff and Noordhoff, the Netherlands, 1978.