A CLASSIFICATION SCHEME FOR 'HIGH-DIMENSIONAL-SMALL-SAMPLE-SIZE' DATA USING SODA AND RIDGE-SVM WITH MICROWAVE MEASUREMENT APPLICATIONS

Yinan Yu, Tomas McKelvey, Senior Member, IEEE, Sun-Yuan Kung, Fellow, IEEE

Chalmers University of Technology, Gothenburg, Sweden
Princeton University, NJ, USA

ABSTRACT

The generalization performance of SVM-type classifiers severely suffers from the 'curse of dimensionality'. For some real world applications, the dimensionality of the measurement is sometimes significantly larger compared to the amount of training data samples available. In this paper, a classification scheme is proposed and compared with existing techniques for such scenarios. The proposed scheme includes two parts: (i) feature selection and transformation based on Fisher discriminant criteria and (ii) a hybrid classifier combining Kernel Ridge Regression with Support Vector Machine to predict the label of the data. The first part is named Successively Orthogonal Discriminant Analysis (SODA), which is applied after Fisher score based feature selection as a preliminary processing for dimensionality reduction. At this step, SODA maximizes the ratio of between-class-scatter and within-class-scatter to obtain an orthogonal transformation matrix which maps the features to a new low dimensional feature space where the class separability is maximized. The techniques are tested on high dimensional data from a microwave measurements system and are compared with existing techniques.

Index Terms— Feature extraction, SODA, Ridge-SVM, Microwave measurements

1. INTRODUCTION

With the rapid development of data capture and storage technologies, the 'curse of dimensionality' [1] becomes an extremely common issue. Therefore, feature selection is critical in many pattern recognition and machine learning applications such as image processing, computer vision, mobile computing [2], etc. In some challenging applications we are facing the 'high dimensionality and small-sample-size' problem. For instance, in this work we use a microwave measurement system with 10 microwave transceivers to measure the scattering parameters of the object under study for the purpose of detecting existence of anomalies within the object. Vectorization of the raw feature space results in about 20,000 complex numbers. Due to the difficulties of obtaining independent objects to measure, the amount of training data is very limited in comparison to the size of the feature space. Techniques such as Principal Component Analysis [3] and Least Absolute Shrinkage and Selection Operator (LASSO) [4] are commonly employed to overcome such issues. Furthermore, when the sample size is small, the distribution of the data is difficult to estimate and hence no optimal classifier is guaranteed.

In this paper, we propose a classification scheme dealing with 'high dimensionality small-sample-size' problems. The first part of the scheme is called feature selection and extraction including (i) evaluating Fisher score for preliminary feature selection, (ii) Successively Orthogonal Discriminant Analysis (SODA) technique, which finds an orthogonal transformation that maps the feature vectors to a low dimensional space where the class separability is maximized with respect to Fisher discriminant criteria. Existing techniques for finding such transformations, e.g. Orthogonal Linear Discriminant Analysis (OLDA) [5], heavily depend on the number of classes which determines the rank of the between-class-scatter matrix $\boldsymbol{S}_B$. In binary classification, where $\mathrm{rank}(\boldsymbol{S}_B) = 1$, this type of techniques do not apply. These two steps (i) and (ii) are both based on maximizing the class separability according to Fisher discriminant criteria and are therefore highly compatible and complementary to each other.

The second part of the proposed scheme is a Ridge-SVM [6] based classifier where a support vector machine (SVM) [7] based on Kernel Ridge Regression (KRR) [8] is applied to the feature space produced by SODA for label prediction. Ridge-SVM is a hybrid classifier which deals with unknown data distribution [9]. For comparison, a LASSO regression is employed instead of SODA. Under the sparsity assumption, the feature space is shrunk to a dense subset where only the relevant features are selected with respect to the $l1$ constraint. Other classic feature reduction and transformation techniques are also applied and tested empirically.

2. LASSO FEATURE SELECTION: EXPLORE THE FEATURE SPARSITY

Theoretical analysis [10] shows that classifiers such as SVMs become less efficient when there are many irrelevant features but only a few training instances for each class. Therefore, LASSO type regression is a reasonable choice for selecting relevant features in such cases.

In some similar methods, e.g. FVM [11], the feature vectors are selected using LASSO to produce a linear classifier. Instead, we consider the LASSO regression as a preliminary step to select a subset from the original feature space. Each selected dimension is then weighted with the associated nonzero $l1$ coefficient to obtain the new feature space.

Given training measurements, the data matrix X is constructed by placing all the data vectors as its columns:

$$\boldsymbol{X} = \left[\quad \boldsymbol{x}_1^+, \boldsymbol{x}_2^+ \cdots \boldsymbol{x}_{N_+}^+, \boldsymbol{x}_1^-, \boldsymbol{x}_2^- \cdots \boldsymbol{x}_{N_-}^- \quad \right], \qquad (1)$$

where the $+$ and $-$ indicate two classes and the m dimensional vectors $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ are sampled from the two classes respectively.

Under the assumption of sparsity, the feature subset selection problem can be re-formulated as a $l1$ constraint optimization:

This work was sponsored by the Swedish Research Council (VR) which is gratefully acknowledged

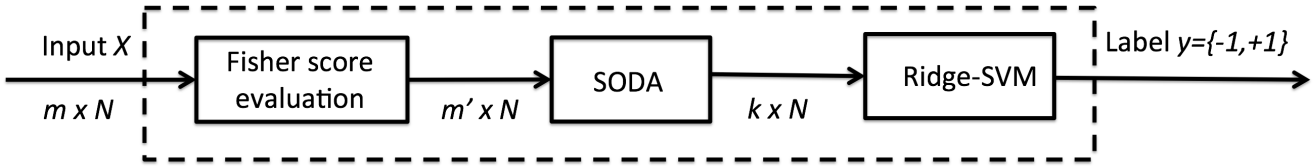

Fig. 1. The proposed classification scheme.

$$\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad \|\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\alpha}\|_{l2} + \lambda\|\boldsymbol{\alpha}\|_{l1}. \tag{2}$$

where $\boldsymbol{Y}$ is the label vector containing $\{0, 1\}$ for all the involved training samples. Given λ, this well studied optimization structure can be solved by convex optimization techniques. The solution vector $\boldsymbol{\alpha}$ of size m containing only a few (m') non-zero elements which is then used as a weighting vector for the features. Namely, for a testing sample $\boldsymbol{x}$, we have:

$$\boldsymbol{x}_{\text{selected}} = \{x_j : x_k\mathbb{I}\{\alpha_k \neq 0\}, k = 1, ..., m, j_k = 1, ..., m'\}$$
$$\boldsymbol{x}_{\text{weighted}} = diag(\alpha_{k_1}, \cdots, \alpha_{k'_m})\boldsymbol{x}_{\text{selected}}$$

where $\mathbb{I}$ is the indicator function, coefficients $\alpha_{k_1}...\alpha_{k_{m'}}$ are the nonzero elements of the vector $\boldsymbol{\alpha}$. Note that in future discussions, the feature space is denoted by $\mathcal{X}$ and the feature vectors $\boldsymbol{x}$ for convenience.

3. SUCCESSIVELY ORTHOGONAL DISCRIMINANT ANALYSIS (SODA)

In this section, a new feature extraction technique called Successively Orthogonal Discriminant Analysis (SODA) is developed. First, feature selection by evaluating Fisher score is used to reduce the dimensionality, and then SODA is applied to find an orthogonal matrix containing a set of orthogonal vector $\boldsymbol{w}_i$'s, which provides a map $\boldsymbol{x} \rightarrow \boldsymbol{x}'$ such that in the new space consisting of vectors $\boldsymbol{x}' = \boldsymbol{W}^T\boldsymbol{x}$, the class separability is maximized. Figure 1 illustrate the sequence of processing steps. The two steps are presented below.

3.1. Step 1: Fisher score feature selection

For computational reasons, a feature selection technique based on Fisher score evaluation [12] is employed to preliminarily reduce the dimensionality. For data $\boldsymbol{x}$, the Fisher score of feature j is defined as follows:

$$fs_j(\boldsymbol{x}) = \frac{1}{(\sigma^j)^2}\sum_{k\in\{+,-\}} n_k(\mu_k^j - \mu^j)^2 \tag{3}$$

where k denotes the class index and takes the value of $+$ or $-$ and n_k is number of samples in the corresponding class. The scalars μ_k^j, σ_k^j are the mean value and variance of feature j from class k while μ^j is the mean value for feature j of the whole training set. Finally, $(\sigma^j)^2$ is defined as:

$$(\sigma^j)^2 = \sum_{k\in\{+,-\}} n_k(\sigma_k^j)^2 \tag{4}$$

After the Fisher score for each feature is computed, a predefined number of features with largest scores are selected. As a preliminary feature selection technique, it has some advantages such as: 1) computationally simple; 2) efficiently reduces the feature space with discriminant information preserved with respect to Fisher criteria; and 3) compatible with other 'Fisher type' techniques. However, this classic and straightforward approach has some drawbacks including: 1) no optimality is guaranteed due to its heuristic nature; 2) no combinations of the features are taken into account and 3) redundant features cannot be handled. This motivates us to develop a second step to compensate for these drawbacks.

3.2. Step 2: SODA for feature transformation

To further enhance the discriminant ability after the feature subset selection, we develop a new technique called Successively Orthogonal Discriminant Analysis (SODA). The attempt of this approach is to construct a linear transformation $\boldsymbol{W}^T$ which takes the data points from the feature space $\mathcal{X}$ (with dimension m') to a new space $\mathcal{X}'$ (dimension k, $k < m'$) where the class separability is maximized on the training samples. The separability is measured over the 'between-class scatter matrix' $\boldsymbol{S}_B$ and the 'within-class scatter matrix' $\boldsymbol{S}_W$, which are respectively defined as:

$$\begin{aligned}\boldsymbol{S}_W &= \frac{1}{N_+}\sum_{i=1}^{N_+}(\boldsymbol{x}_i^+ - \boldsymbol{\mu}^+)(\boldsymbol{x}_i^+ - \boldsymbol{\mu}^+)^T \\ &+ \frac{1}{N_-}\sum_{j=1}^{N_-}(\boldsymbol{x}_j^- - \boldsymbol{\mu}^-)(\boldsymbol{x}_j^- - \boldsymbol{\mu}^-)^T\end{aligned}$$

$$\boldsymbol{S}_B = (\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)^T \tag{5}$$

where $\boldsymbol{\mu}^+$ and $\boldsymbol{\mu}^-$ are the mean vector estimated from the corresponding class $+$ and $-$. In LDA, we want to find a vector $\boldsymbol{w}$ which maximizes the following Fisher score

$$\frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}. \tag{6}$$

The vector which maximizes (6) is the generalized eigenvector corresponding to the largest generalized eigenvalue of the problem $\boldsymbol{S}_B\boldsymbol{w} = \boldsymbol{S}_W\boldsymbol{w}\lambda$. In the case of a singular $\boldsymbol{S}_W$ the solution can be determined as the eigenvector of the Fisher matrix $\boldsymbol{F} = \boldsymbol{S_W}^+\boldsymbol{S_B}$, corresponding to the largest eigenvalue [13].

Instead of a one dimensional vector, we want to find a transformation matrix $\boldsymbol{W}$ with k column vectors. However, for a binary classification problem, the matrix $\boldsymbol{S}_B$ has rank one and therefore only one non-zero eigenvalue can be obtained. Therefore, a new formulation and its solution are proposed as follows.

SODA formulation. *The matrix $\boldsymbol{W} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_k]$ defines a map $\boldsymbol{x} \rightarrow \boldsymbol{x}'$, whose columns satisfy:*

$$\begin{aligned} \underset{\boldsymbol{w}_i}{\textit{maximize}} \quad & \frac{\boldsymbol{w}_i^T \boldsymbol{S}_B \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_W \boldsymbol{w}_i} \\ \textit{subject to} \quad & \boldsymbol{w}_i \perp \boldsymbol{w}_{1,\ldots,i-1} \\ & \boldsymbol{w}_i^T \boldsymbol{w}_i = 1 \\ & \boldsymbol{w}_i \in \textit{Span}(\boldsymbol{S}_W) \end{aligned} \tag{7}$$

where Span$(\boldsymbol{S}_W)$ denotes the range space of matrix $\boldsymbol{S}_W$. □

Algorithm SODA

- Let $\boldsymbol{S}_W^{(0)} = \boldsymbol{S}_W$
  $\boldsymbol{F}^{(1)} = (\boldsymbol{S}_W^{(0)})^+ \boldsymbol{S}_B$
- For $i = 1 : k$
  - Solve for $\boldsymbol{F}^{(i)} \boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$
    where λ_i is the largest and only eigenvalue of $\boldsymbol{F}^{(i)}$.
  - Let $\boldsymbol{D}^{(i)} = \boldsymbol{I}_{m \times m} - \boldsymbol{w}_i \boldsymbol{w}_i^T$ be the deflation matrix
    $\boldsymbol{S}_W^{(i)} = \boldsymbol{D}^{(i)} \boldsymbol{S}_W^{(i-1)} \boldsymbol{D}^{(i)}$
    $\boldsymbol{F}^{(i+1)} = (\boldsymbol{S}_W^{(i)})^+ \boldsymbol{S}_B$
- Form matrix: $\boldsymbol{W} = \left[\boldsymbol{w}_1 \cdots \boldsymbol{w}_k\right]$
- Transformation of the features: $\boldsymbol{x}' = \boldsymbol{W}^T \boldsymbol{x}$

Theorem. *The columns of the matrix $\boldsymbol{W}$ obtained from Algorithm SODA solves the optimization problem stated in (7).*
Proof. The proof consists of two parts: 1) $\boldsymbol{w}_i$ is in the column space of $(\boldsymbol{S}_W)^{(i)}$, and 2) $\boldsymbol{w}_1 \perp \boldsymbol{w}_2 \cdots \perp \boldsymbol{w}_k$.

1) It suffices to show that $\boldsymbol{w}_1 \in$ Span$(\boldsymbol{S}_W^{(1)})$. It is obvious that $\boldsymbol{w}_1 \in$ Span$((\boldsymbol{S}_W^{(1)})^+)$. Let $\boldsymbol{S}_W^{(1)} = \boldsymbol{U_1} \boldsymbol{\Sigma} \boldsymbol{U_1}^T$, we have $(\boldsymbol{S}_W^{(1)})^+ = \boldsymbol{U_1} \boldsymbol{\Sigma}^{-1} \boldsymbol{U_1}^T$. Therefore, $(\boldsymbol{S}_W^{(1)})^+$ and $\boldsymbol{S}_W^{(I)}$ share the same column space and hence it follows that $\boldsymbol{w}_1 \in$ Span$(\boldsymbol{S}_W^{(1)})$.
2) For $i = 1$, the first components $\boldsymbol{w}_1$ is set to be equal to normalized eigenvector associated with the largest (and the only) eigenvalue of $(\boldsymbol{S}_W^{(1)})^+ \boldsymbol{S}_B$, which obviously maximizes the Fisher score of step $i = 1$.

   For the next stage, i.e. $i = 2$, the deflation operator completely removes $\boldsymbol{w}_1$ component from $\boldsymbol{S}_W^{(1)}$, forcing $(\boldsymbol{S}_W^{(1)})^+ \boldsymbol{S}_B$ to contain only components orthogonal to $\boldsymbol{w}_1$. It follows that the optimal solution $\boldsymbol{w}_2$ for the maximal Fisher score of step $i = 2$ must be orthogonal to $\boldsymbol{w}_1$. By induction, it eventually leads to $\boldsymbol{w}_1 \perp \boldsymbol{w}_2 \cdots \perp \boldsymbol{w}_k$. Thus, the proof is completed. □

4. KERNEL RIDGE REGRESSION AND SVM

In the previous sections, the feature space has been shrunk and transformed in order to enhance the class separability. These feature selection and extraction techniques are considered as pre-processing steps to produce the input of the classifier. In this section, a hybrid classifier based on support vector machine (SVM) using kernel ridge regression (KRR) is presented as the final step in the classification scheme [14]. As we know, in data driven classification techniques, the performance of a classifier heavily depends on the distribution of the data. When the distribution is not known in advance, it is difficult to choose a proper type of classifier which best solves the problem. In particular, for the case when only training data with a limited sample size is available, the distribution of the data set cannot be easily estimated. This motivates us to propose a unified hybrid classifier named Ridge-SVM. The classifier is very versatile as it covers existing classifiers, including KDA, KRR, and SVM, as special cases [9]. Given a kernel function $\boldsymbol{K}$, the formulation of KRR and SVM are compared to make sense of the concept:
(1) KRR:

$$\begin{aligned} \underset{\boldsymbol{a}}{\text{maximize}} \quad & \left\{\boldsymbol{a}^T \boldsymbol{Y} - \tfrac{1}{2} \boldsymbol{a}^T \left[\boldsymbol{K} + \rho \boldsymbol{I}\right] \boldsymbol{a}\right\} \\ \text{Subject to} \quad & \Sigma_i a_i = 0 \end{aligned} \tag{8}$$

where $\boldsymbol{Y}$ contains the labels $\{0, 1\}$. Parameter ρ penalizes the weak and vulnerable components in the spectral space to avoid overfitting problem.
(2) SVM:

$$\begin{aligned} \underset{\boldsymbol{a}}{\text{maximize}} \quad & \left\{\boldsymbol{a}^T \boldsymbol{Y} - \tfrac{1}{2}\, \boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a}\right\} \\ \text{Subject to} \quad & \Sigma_i a_i = 0 \\ & 0 \leq \alpha_i \leq C, \text{where } \alpha_i = a_i Y_i \end{aligned} \tag{9}$$

Parameter C controls the size of the participating components. When C is sufficiently small, the SVM is expected to be more robust. The parameters ρ and C are complementary to each other which also motivates the development of the hybrid classifier. The Ridge-SVM can be then formulated as follows:

$$\begin{aligned} \underset{\boldsymbol{a}}{\text{maximize}} \quad & \left\{\boldsymbol{a}^T \boldsymbol{Y} - \tfrac{1}{2} \boldsymbol{a}^T \left[\boldsymbol{K} + \rho \boldsymbol{I}\right] \boldsymbol{a}\right\} \\ \text{Subject to} \quad & \Sigma_i a_i = 0 \\ & C_{min} \leq \alpha_i \leq C_{max}, \text{where } \alpha_i = a_i Y_i \end{aligned} \tag{10}$$

and the discriminant function is:

$$f(\boldsymbol{x}) = \Sigma_{i=1} \boldsymbol{\alpha}_i \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{11}$$

for some $b \in \mathbb{R}$ with the decision boundary $f(\boldsymbol{x}) = 0$.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The measured signals from the microwave system are the complex scattering parameters [15] with a frequency range from 100 MHz to 3.0 GHz. Each signal contains 401 frequency points, and 55 different channels are involved corresponding to the reflection and transmission channels implied by the 10 element array antenna. By considering the real and imaginary parts of the complex signal, we get a 44,110-dimensional feature space. In this study, the main interest is to use the measurements to classify an object into one of two classes. Our experiment involves 27 samples class + and 45 class −. As usual, class + contains objects with an anomaly which the classifier should detect at a low false alarm rate. To assess the performance of the tested methods a leave-one-out validation approach is applied. Figure 2 shows the distribution of the two first principal components calculated by PCA for the available data. Clearly, a linear classifier based on the first two principal components is not a viable alternative and show why kernel based methods can be useful.

Two types of preliminary feature selections are used in our experiments: LASSO and a Fisher-Score-based method. In our study, the number of features for LASSO is set at $m' = 40$. Figure 3 illustrates how the detection rate depends on the

Feature Selection	Feature Transformation	# Features	Classifier	Detection Rate
LASSO (40 features)	×	40	Ridge SVM (RBF Kernel)	73%
LASSO (40 features)	Linear scaling	40	Ridge SVM (RBF Kernel)	88%
LASSO (40 features)	SODA	10	Ridge SVM (RBF Kernel)	78%
LASSO (40 features)	SODA	10	SVM (RBF Kernel)	72%
Fisher Score (400 features)	×	400	LDA	74%
Fisher Score (400 features)	PCA	10	Ridge SVM (RBF Kernel)	68%
Fisher Score (400 features)	SODA	10	Ridge SVM (RBF Kernel)	92%
×	×	44,110	SVM (RBF Kernel)	75%

Table 1. Comparison of rate of detection of class 1 with a constant false alarm rate of 20% for the different scenarios tested, where for Ridge-SVM $C_{min} = -1$, $C_{max} = 1$ and $\rho = 0.4$. For SVM, $C_{min} = 0$, $C_{max} = 1$ and $\rho = 0$. For parameters achieving better performance (for SVM and Ridge-SVM), see Table 2.

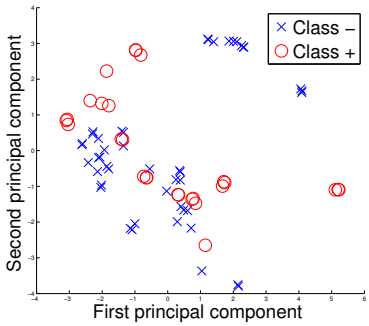

Fig. 2. Two dimensional visualization of the first two principal components of the dataset calculated using PCA.

number of Fisher-score based features. Note that the rates saturate for feature numbers over 400.

Table 1 illustrates our experimental results with various feature types, feature numbers, feature transformations, and classifiers. The numbers in the last column reflect the detection rates of class + at a constant false alarm rate of 20% estimated as the empirical probabilities based on the leave-one-out validation. Note that, reduction from 40 LASSO features into 10 SODA outputs actually results in a performance deterioration (from 88% down to 78%). On the other hand, the combination of the Fisher-Score-based method and SODA can yield very high detection rates. Using 400 Fisher-Score based features as input to SODA and again using 10 as the SODA's output dimension, the performance can increase to 92%, far higher than the conventional PCA and LASSO approaches (68% and 88%, respectively). In summary, when combined with Fisher-score based feature selection, SODA can effectively reduce the dimensionality needed for classifiers while still retaining very high performance. Such a good combination may be due to the fact that they are both tied with FDA.

Our study also suggests that Ridge-SVM is promising in enhanced generalization ability for datasets with unknown distribution [6, 9]. Based on the microwave dataset of class +, the detection rates of SVM Ridge-SVM, with various learning parameters C_{min}, C_{max} and ρ, are summarized in Table 2. We can see that Ridge-SVM's 95.2% clearly output performs SVM's 86.8% and 92.1% (with $\rho = 0.4$ and $\rho = 0.4$, respectively).

	C_{min}	C_{max}	Detection Rate
SVM with $\rho = 0$	0	1	74.6%
	0	10	86.8%
	0	100	67.9%
SVM with $\rho = 0.4$	0	1	87.2%
	0	10	92.6%
Ridge-SVM with $\rho = 0.4$	-0.1	0.1	91.2%
	0.1	1	87.3%
	-1	1	92.9%
	-1	10	87.6%
	1	10	91.4%
	-10	10	95.2%
	-10	100	85.6%
	10	100	89.0 %
	-100	100	94.2 %

Table 2. In this experiment, the input vector of SVM or Ridge-SVM classifiers comprises 10 SODA outputs reduced from 200 Fisher-score based features. Shown here are detection rates of class + for different C, C_{min}, and C_{max}, again with 20% false alarm rate. In our study, setting $C = 10$ appears to give the best SVM result. For Ridge-SVM, on the other hand, setting $(C_{min}, C_{max}) = (-10, 10)$ produces a much higher rate than $(-1, 1)$ or $(-100, 100)$.

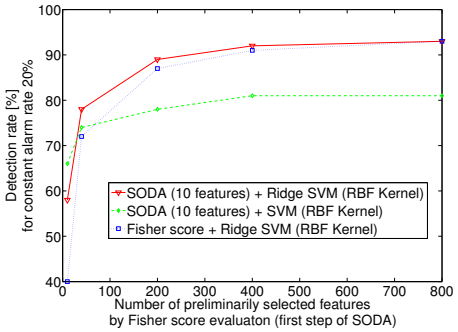

Fig. 3. SODA comparison: different numbers of selected features versus detection rate for a 20% false alarm rate.

6. REFERENCES

[1] Bellman, R., *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.

[2] Kung, S.Y., Wu Pei-Yuan, *On efficient learning and classification kernel methods.* Acoustics, Speech and Signal Processing (ICASSP), pp. 2065 - 2068, 2012.

[3] Jolliffe I.T. *Principal Component Analysis.* Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002.

[4] Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of Royal Statistics Society: Series B, Vol. 58, No. 1, pages 267-288, 1996.

[5] Ye, J, Xiong, T, *Computational and Theoretical Analysis of Null Space and Orthogonal Linear Discriminant Analysis.* Journal of Machine Learning Research vol. 7, pp. 11831204, 2006.

[6] Kung S. Y., *Kernel Approaches to Unsupervised and Supervised Machine Learning.* Proc. PCM'2009, Bangkok, Thailand, Lecture Notes in Computer Science, pp 1-32, vol. 5879, Springer-Verlag, 2009.

[7] Olivier Chapelle, Patrick Haffner, and Vladimir N. Vapnik, *Support Vector Machines for Histogram-Based Image Classification.* IEEE Transactions on Neural Networks, VOL. 10, NO. 5, Sep. 1999.

[8] Arthur E. Hoerl, Robert W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, Vol. 12, No. 1, pp. 55-67 Feb., 1970.

[9] Kung S. Y., *Kernel Methods and Machine Learning.* Cambridge Press, 2013.

[10] Ng, A. Y., *Feature selection, L1 vs. L2 regularization, and rotational invariance.* 21st ICML, New York, USA, ACM, 2004.

[11] Li, F., Yang, Y., Xing, E. P., *From Lasso regression to Feature vector machine.* NIPS. 2003.

[12] Duda, R.O. and Hart, P.E. and Stork, D.G., *Pattern Classification*, 2nd Edition, John Wiley & Sons, New York, 2011.

[13] Krzanowski, W. J., Jonathan, P., McCarthy, W. V. and Thomas, M. R., *Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data.* Appl. Statist., 44, 101115, 1995.

[14] Kung, S.Y. and Mak, M.W., *PDA-SVM Hybrid: A Unified Model for Kernel-Based Supervised Classification.* Journal of VLSI Signal Processing Systems - Special Issue on Advances in Multimedia Computing, Communications and Applications, vol. 65, May, 2011.

[15] Pozar, David M., *Microwave Engineering.* Third Edition, John Wiley & Sons, pp 170-174, 2005.