STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **6**, June 2018, pp 159–177.
Published online in International Academic Press (www.IAPress.org)

Hybridized Support Vector Machine and Recursive Feature Elimination with Information Complexity

Hamparsum Bozdogan [1], Seung Hyun Baek [2,*]

[1] *Department of Business Analytics & Statistics, University of Tennessee, Knoxville, USA.*
[2] *Division of Business Administration, Hanyang University, ERICA, Korea.*

Abstract In statistical data mining research, datasets often have nonlinearity and at the same time high-dimensionality. It has become difficult to analyze such datasets in a comprehensive manner using traditional statistical methodologies. In this paper, a novel wrapper method called SVM-$ICOMP_{PERF}$-RFE based on a hybridized support vector machine (SVM) and recursive feature elimination (RFE) with information-theoretic measure of complexity (ICOMP) is introduced and developed to classify high-dimensional data sets and to carry out subset selection of the features in the original data space for finding the best subset of features which are discriminating between the groups. Recursive feature elimination (RFE) ranks features based on information complexity (ICOMP) criterion. ICOMP plays an important role not only in choosing an optimal kernel function from a portfolio of many other kernel functions, but also in selecting important subset(s) of features. The potential and the flexibility of our approach are illustrated on two real benchmark data sets, one is ionosphere data which includes radar returns from the ionosphere, and another is aorta data which is used for the early detection of atheroma most commonly resulting heart attack. Also, the proposed method is compared with other RFE based methods using different measures (i.e., weight and gradient) for feature rankings.

Keywords Feature Selection, Support Vector Machine, Recursive Feature Elimination, Information Complexity Criterion, ICOMP

AMS 2010 subject classifications 68T10,68T20

DOI: 10.19139/soic.v6i2.327

1. Introduction

In many classification problems there are very high-dimensional input datasets and finding the best subset of the original input features or variables which mostly contribute to the separation of the classes or groups is a challenge. Therefore, the problem of feature selection is a difficult combinatorial problem in Machine Learning and it has very of high practical importance in many applications.

Kernel-based methods have gained popularity for classification, clustering, and regression analysis in machine learning since the introduction of support vector machine (SVM) during the early 1990s, after obtaining support vectors (SVs) to classify a data set, questions such as: *"How do we know which features are more responsible for, and important to, the classification?"* has often been raised. This is due to the fact that the mapping is not one-to-one and onto in SVM. The application of a kernel function is thus an uninvertible process, and there is no way to go from the feature space back to the original space. Because of this geometry, SVM does not land itself

[*] Correspondence to: Seung Hyun Baek (Email: sbaek4@hanyang.ac.kr). Division of Business Administration, Hanyang University, ERICA. 55 Hanyangdaehak-ro, Sangnok, Ansan, Gyeonggi-do, Korea (15588).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

in an automated internal relevant feature selection easily. Hence algorithms for feature selection play an important role in SVM .

In the literature of Machine Learning, as discussed in [11] in detail, there are two main approaches to solve the feature selection problem: (a) the filter approach, and (b) the wrapper approach. Both approaches differ in the way they evaluate a given feature subset. The filter method uses some relevance measure, which is independent of the performance of the learning algorithm. On the other hand, in the wrapper method each feature subset is taken into consideration with the classifier. That is, the features are evaluated by estimating the generalization performance (i.e. the expected risk) of the learning machine.

In this paper, the wrapper method called SVM-$ICOMP_{PERF}$-RFE, which combines recursive feature elimination and an information-theoretic measure of complexity (ICOMP) criterion especially designed for SVM based on feature selection developed by [12] is considered and emphasized. In the usual RFE, backward feature elimination is performed to find say, m, features which lead to the largest margin of class separation. This combinatorial problem is solved in a greedy fashion. In the two-class case the RFE algorithm begins with the set of all features and successively eliminates the feature which induces the smallest change based on sensitivity analysis for an appropriately defined cost function which is a measure of predictive ability (and is inversely proportional to the margin). Then, the RFE algorithm at each step eliminates the feature which keeps this quantity small. Assuming that the change of the set of support vectors when removing only one feature is negligible.

An information-theoretic measure of complexity (ICOMP) criterion [3, 4, 5, 6, 7] is used in RFE rankings of the features as an effective measure. ICOMP plays an important role not only in choosing an optimal kernel function from a portfolio of many other kernel functions, but also in selecting important subset(s) of features. It takes into account both the badness of fit? or the lack of fit? and the model complexity at the same time in one criterion function.

The proposed method is compared with two different RFE based methods [12, 32, 10] with two real benchmark data sets.

The rest of the organization of this paper is as follows. In Section 2, the background of SVM [30, 27, 9] and the several forms of the kernel functions are presented. In Section 3, the information complexity (ICOMP) criterion to choose the optimal kernel function and to select the best subset of features using HSVM-RFE is introduced. Section 4 discusses recursive feature elimination (RFE) technique and provides algorithms for three RFE based methods. In Section 5, numerical examples are provided to study the efficiency of the proposed method with two real benchmark data sets. The proposed method and two existing RFE based methods are compared in Section 6. This paper is concluded by Section 7.

2. Support Vector Machine (SVM)

Consider the case of classifying a set of linearly separable data into two groups. Assume a set of training data is given by $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i$ is an input vector, $y_i \in (-1, 1)$ is a binary class index, and n is the size of training data. SVM finds optimal separating hyperplane that maximizes the margin between the classes [30]. Then, a decision boundary (i.e. classifier) that partitions the underlying vector space into two classes can be represented by the following hyperplane:

$$\mathbf{w}^T\mathbf{x} + b = 0,$$

where $\mathbf{w}$ is the weight vector and b is the bias. The objective of SVM is to find maximum margin(M) decision boundary between two parallel hyperplanes, $\mathbf{w}^T\mathbf{x} + b = 1$ and $\mathbf{w}^T\mathbf{x} + b = -1$. An example of SVM is illustrated in Figure 1. Since the margin is given by $2/\|\mathbf{w}\|$, the corresponding optimization problem can be written as follows:

$$\begin{array}{ll} Minimize & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{Subject to} & \left\{ \begin{array}{c} y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, 2, \ldots, n \\ \xi_i \geq 0, i = 1, 2, \ldots, n \end{array} \right. \end{array}$$

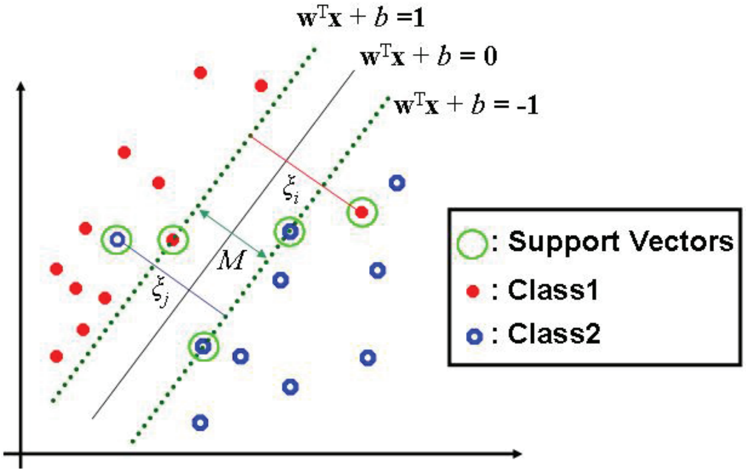

$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$
ξ_i
M
ξ_j
○: Support Vectors
●: Class1
○: Class2

Figure 1. Illustration of linear SVM for nonlinearly separable case.

where ξ_i is the positive slack variable and $C\,(>0)$ is a pre-defined regularization coefficient. The linearly-constrained optimization problem can be solved as a dual problem that maximizes the following function:

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right),$$

subject to the constraint

$$\sum_{i=1}^{n} \alpha_i \mathbf{x}_i = 0, , i = 1, 2, \cdots, n$$
$$0 \leq \alpha_i \leq C, i = 1, 2, \cdots, n.$$

Once the optimum values (α^*, b^*) are obtained, based on the training set of points, a new point $\mathbf{x}_{new}$ of the test data set is classified by the following decision rule:

$$\begin{array}{lll} Class1 & if & D(\mathbf{x}_{new}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_{new}) + b^* < 0 \\ Class2 & if & D(\mathbf{x}_{new}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_{new}) + b^* > 0, \end{array}$$

where $D(\cdot)$ is a classifier based on the training data set. $K(\mathbf{x}_i, \mathbf{x}_{new})$ is the kernel trick proposed by [1]. The kernel trick maps input data in the original space with nonlinearly into a high-dimensional feature space. The Table 1 presents some common kernel functions.

3. Information-Theoretic Measure of Complexity

An information-theoretic measure of complexity called ICOMP has been proposed by Bozdogan [3, 4, 5, 7] as a decision rule for model selection such as AIC [2], and BIC [24]. The development and construction of ICOMP is based on a generalization of the covariance complexity index originally introduced by [29]. Instead of penalizing the number of free parameters directly, ICOMP penalizes the covariance complexity of the model. It is defined by

$$ICOMP = -2\log L(\hat{\theta}_k) + 2C(\hat{\mathbf{\Sigma}}_{Model}),$$

Table 1. Kernel functions.

Function	$K(\mathbf{X},\mathbf{Y})$	Parameters
Linear	$\left(\mathbf{X}^T\mathbf{Y}+b\right)^a$	$a=1, b=0$
Polynomial (degree=2)	$\left(\mathbf{X}^T\mathbf{Y}+b\right)^a$	$a=2, b=1$
Polynomial (degree=3)	$\left(\mathbf{X}^T\mathbf{Y}+b\right)^a$	$a=3, b=1$
Gaussian	$\exp\left(-\left(\frac{1}{a^b}\left\|\mathbf{X}-\mathbf{Y}\right\|^2\right)^c\right)$	$a=2, b=c=1$
Cauchy	$\left(\mathbf{1}+\frac{1}{a}\left\|\mathbf{X}-\mathbf{Y}\right\|^2\right)^{-1}$	$a=1$
Inverse Multi-Quadric	$\left(\left\|\mathbf{X}-\mathbf{Y}\right\|^2+a^2\right)^{-\frac{1}{2}}$	$a=1$

where $L(\hat{\theta}_k)$ is the maximized likelihood function, $\hat{\theta}_k$ is the maximum likelihood estimate of the parameter vector θ_k under the model M_k, and C represents a real-valued complexity measure and $\widehat{Cov}(\hat{\theta}_k) = \hat{\mathbf{\Sigma}}_{Model}$ represents the estimated covariance matrix of the parameter vector of the model. ICOMP should not be confused with the stochastic complexity (SC) or the minimum description length (MDL) of Rissanen [21, 22, 23], although they both use the notion of complexity of a model class based on coding theory. The detailed information-theoretic measure of complexity (ICOMP) is recapitulated in the subsections for the benefit of the readers who may not be familiar with ICOMP criterion.

3.1. *Mutual Information in High Dimensions*

For a random vector, the complexity is defined as follows.
Definition: The complexity of a random vector is a measure of the interdependency between its components.

A continuous p-variate distribution is used with joint density function $f(\mathbf{x}) = f(x_1, ..., x_p)$ and marginal density functions $f_j(x_j), j = 1, .., p$. Following [15], and [13], the *information measure of dependence* is defined as follows:

$$\begin{aligned} I(\mathbf{x}) &= I(x_1,...,x_p) = E_f[\log \frac{f(x_1,\ldots,x_p)}{f_1(x_1)\cdots f_p(x_p)}] \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(x_1,\ldots,x_p)\log\frac{f(x_1,\ldots,x_p)}{f_1(x_1)\cdots f_p(x_p)}dx_1\cdots dx_p \end{aligned}$$

where $I(\mathbf{x})$ is the Kullback-Leibler information divergence [16] against independence. The properties of the Kullback-Leibler information divergence are as follows:

- $I(\mathbf{x}) \equiv I(x_1,\ldots,x_p) \geq 0$ i.e., the expected mutual information is nonnegative.
- $I(\mathbf{x}) \equiv I(x_1,\ldots,x_p) = 0$ if and only if $f(x_1,\ldots,x_p) = f_1(x_1)\cdots f_p(x_p)$ for every p-tuple $(x_1,\ldots,x_p)$, i.e., if and only if the random variables $x_1,\ldots,x_p$ are mutually statistically independent.

The *KL divergence* is related to Shannon's entropy [25] by the important identity

$$I(\mathbf{x}) \equiv I(x_1,\ldots,x_p) = \sum_{j=1}^{p} H(x_j) - H(x_1,\ldots,x_p) \tag{1}$$

where $H(x_j)$ is the marginal entropy, and $H(x_1,\ldots,x_p)$ is the global or joint entropy.

[31] calls this latter quantity the strength of structure and a *measure of inter-dependence*.

To define the information-theoretic measure of complexity of a multivariate distribution, let $f(\mathbf{x}) = f(x_1, \ldots, x_p)$ be a multivariate Gaussian density function given by

$$\begin{aligned} f(\mathbf{x}) &= f(x_1, \ldots, x_p) \\ &= (2\pi)^{-\frac{p}{2}} \left|\mathbf{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}, \end{aligned}$$

where $\mu = (\mu_1, \mu_2, ..., \mu_p)^T$, $-\infty < \mu_j < \infty$, $j = 1, 2, ..., p$ and $\mathbf{\Sigma} > 0$ (positive definite).

As a short hand, let

$$\mathbf{x} \sim N_p(\mu, \mathbf{\Sigma}).$$

Then the joint entropy $H(\mathbf{x}) = H(x_1, ..., x_p)$ from equation (1) for the case in which $\mu = \mathbf{0}$ is given by

$$\begin{aligned} H(\mathbf{x}) &= H(x_1, ..., x_p) = -\int_{\mathrm{R}^p} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \qquad (2) \\ &= \int_{\mathrm{R}^p} f(\mathbf{x}) \left[\frac{p}{2}\log(2\pi)\left|\mathbf{\Sigma}\right| + \frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right] d\mathbf{x} \\ &= \frac{p}{2}\log(2\pi)\left|\mathbf{\Sigma}\right| + \frac{1}{2}tr\left[\int_{\mathrm{R}^p} f(\mathbf{x})\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T d\mathbf{x}\right]. \end{aligned}$$

Then, since $E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T] = \mathbf{\Sigma}$, the joint entropy is

$$\begin{aligned} H(\mathbf{x}) &= H(x_1, ..., x_p) = \frac{p}{2}\log(2\pi) + \frac{p}{2} + \frac{1}{2}\log\left|\mathbf{\Sigma}\right| \\ &= \frac{p}{2}\left[\log(2\pi) + 1\right] + \frac{1}{2}\log\left|\mathbf{\Sigma}\right|. \end{aligned}$$

From equation (2), the marginal entropy $H(x_j)$ is

$$\begin{aligned} H(x_j) &= -\int_{-\infty}^{+\infty} f(x_j)\log f(x_j)dx_j \\ &= \frac{1}{2}\log(2\pi) + \frac{1}{2} + \frac{1}{2}\log(\sigma_j^2), j = 1, 2, \ldots, p, \end{aligned}$$

where σ_j^2 is the variance of the j^{th} variable.

3.2. Initial Definition of Covariance Complexity

[29, p. 61] provides a reasonable initial definition of complexity of a covariance matrix $\mathbf{\Sigma}$ for the multivariate Gaussian distribution. This measure is given by:

$$I(x_1, ..., x_p) \equiv C_0(\mathbf{\Sigma}) = \sum_{j=1}^{p} H(x_j) - H(x_1, ..., x_p)$$

$$= \sum_{j=1}^{p}\left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma_{jj}) + \frac{1}{2}\right] - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log\left|\mathbf{\Sigma}\right| - \frac{p}{2}.$$

This reduces to

$$C_0(\mathbf{\Sigma}) = \frac{1}{2}\sum_{j=1}^{p}\log(\sigma_{jj}) - \frac{1}{2}\log|\mathbf{\Sigma}|\,, \tag{3}$$

where $\sigma_{jj} \equiv \sigma_j^2$, is the variance of the j^{th} variable, and is the j^{th} diagonal element of $\mathbf{\Sigma}$.
The characteristics of covariance complexity C_0 are as follows:

- $C_0\,(\mathbf{\Sigma}) = 0$ if and only if Σ is a diagonal matrix.
- $C_0\,(\mathbf{\Sigma}) = \infty$ if and only if $|\mathbf{\Sigma}| = 0$.
- The first term of equation (3) is not invariant under orthonormal transformations.

As pointed out by [29], the result in equation (3) is not an effective measure of the amount of complexity in the covariance matrix $\mathbf{\Sigma}$, since:

- $C_0(\mathbf{\Sigma})$ depends on the coordinates of the original random variables $x_1, ..., x_p$.
- The first term of $C_0(\mathbf{\Sigma})$ in equation (3) would change under orthonormal transformations.

3.3. Definition of Maximal Covariance Complexity

To improve upon $C_0(\mathbf{\Sigma})$ in equation (3), we propose the following.

Proposition: A maximal information theoretic measure of complexity of a covariance matrix $\mathbf{\Sigma}$ of a multivariate Gaussian distribution is defined as follows:

$$\begin{aligned} C_1(\mathbf{\Sigma}) &= \max_T C_0(\mathbf{\Sigma}) = \max_T \{H(x_1) + \cdots + H(x_p) - H(x_1, ..., x_p)\} \\ &= \frac{p}{2}\log\left[\frac{tr(\mathbf{\Sigma})}{p}\right] - \frac{1}{2}\log|\mathbf{\Sigma}| \\ &= \frac{p}{2}\log\frac{\overline{\lambda}_a}{\overline{\lambda}_g}, \end{aligned}$$

where the maximum is taken over the orthonormal similarity transformation, T of the overall coordinate systems $x_1, ..., x_p$ and $\overline{\lambda}_a$ and $\overline{\lambda}_g$ are arithmetic and geometric means of the eigenvalues. The properties of maximal information-theoretic measure of complexity are as follows:

- $C_1\,(\mathbf{\Sigma})$ is the log ratio between the arithmetic and geometric mean of the eigenvalues.
- $C_1\,(\mathbf{\Sigma})$ incorporates the two most basic scalar measures of multivariate scatter - *trace* and *determinant*.
- $C_1\,(\mathbf{\Sigma}) \to 0$ as $\mathbf{\Sigma} \to \mathbf{I}_p$.
- As interaction between variables increases, so does $C_1\,(\mathbf{\Sigma})$.

3.4. Modified Maximal Covariance Complexity

Following [29], the geometric definition of covariance complexity is defined by the Frobenius norm given by

$$C_F(\mathbf{\Sigma}) = \frac{1}{s}\|\mathbf{\Sigma}\|^2 - \left(\frac{tr(\mathbf{\Sigma})}{s}\right)^2,$$

where $\|\mathbf{\Sigma}\|^2 = tr(\mathbf{\Sigma}^T\mathbf{\Sigma})$, the square of the Frobenius norm of $\mathbf{\Sigma}$.

In terms of the *eigenvalues* (or *singular values*), $C_F(\mathbf{\Sigma})$ reduces to

$$C_F(\mathbf{\Sigma}) = \frac{1}{s}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2,$$

where s is the rank of $\mathbf{\Sigma}$, λ_j is the j^{th} eigenvalue of $\mathbf{\Sigma} > 0, j = 1, 2, \dots, s$ and $\overline{\lambda}_a$ is arithmetic mean of the eigenvalues. Note that $C_F(\mathbf{\Sigma}) \geq 0$ with $C_F(\mathbf{\Sigma}) = 0$ only when all $\lambda_j = \overline{\lambda}_a$.

$C_1(\mathbf{\Sigma})$ can be approximated in terms of the eigenvalues $\lambda_j, j = 1, 2, \dots, s$ by

$$C_1(\mathbf{\Sigma}) \cong \frac{1}{4}\sum_{j=1}^{s}(\frac{\lambda_j - \overline{\lambda}_a}{\overline{\lambda}_a})^2.$$

Since in the feature space we are dealing with orthonormal matrices, to prevent the C_1 complexity not to go to zero, we relate C_1 and C_F as a second order equivalent measure of complexity denoted by C_{1F}. Hence, the modified maximal entropic complexity $C_{1F}(\mathbf{\Sigma})$ is defined as follows:

$$C_{1F}(\mathbf{\Sigma}) \quad = \quad \frac{s}{4}\frac{C_F(\mathbf{\Sigma})}{(\frac{tr(\mathbf{\Sigma})}{s})^2} = \frac{s}{4}\frac{\frac{1}{s}\left\|\mathbf{\Sigma}\right\|^2 - \left(\frac{tr(\mathbf{\Sigma})}{s}\right)^2}{(\frac{tr(\mathbf{\Sigma}))}{s})^2}.$$

In terms of the eigenvalues, $C_{1F}(\mathbf{\Sigma})$ is given by

$$\begin{aligned} C_{1F}(\mathbf{\Sigma}) &= \frac{s}{4}\frac{\frac{1}{s}tr(\mathbf{\Sigma}^T\mathbf{\Sigma}) - \left(\frac{tr(\mathbf{\Sigma})}{s}\right)^2}{(\frac{tr(\mathbf{\Sigma})}{s})^2} \\ &= \frac{s}{4}\frac{1}{s\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2 \\ &= \frac{1}{4\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2. \end{aligned}$$

where $s = rank(\mathbf{\Sigma})$. The properties of the modified maximal entropic complexity C_{1F} are as follows:

- $C_{1F}(\mathbf{\Sigma})$ is scale-invariant and $C_{1F}(\mathbf{\Sigma}) \geq 0$ with $C_{1F}(\mathbf{\Sigma}) = 0$ only when all $\lambda_j = \overline{\lambda}_a$.
- $C_{1F}(\mathbf{\Sigma})$ measures the relative variation in the eigenvalues rather than absolute variation of the eigenvalues.

3.5. *ICOMP as a Performance Measure:* $ICOMP_{PERF}$

Singularity of the estimated *covariance matrix* is a common problem that has recently attracted many researchers' work. Because of this, many methods have been proposed to make the covariance matrix *well-conditioned* so that we can estimate the covariance matrix. The usual response to *singular* or *ill-conditioned* covariance matrix estimates is the *"naive" ridge regularization*, $\hat{\mathbf{\Sigma}^*} = \left[\hat{\mathbf{\Sigma}} + \alpha\mathbf{I}_p\right]$, which works to counteract the ill-conditionedness by adjusting the eigenvalues of $\hat{\mathbf{\Sigma}}$. The ridge parameter, α, is typically chosen to be very small. This, of course, begs the questions

- *How large of a perturbation do we need?*
- *How small a perturbation can we get away with?*

This is a case where simplicity is not necessarily a good thing; it does not solve the problem with many real datasets. Yet another approach that does not seem to work well in practice is to augment $\hat{\mathbf{\Sigma}}$ with a multiple of the *kernel matrix*, as suggested by [17]. After much experimentation with a variety of different methods to improve the condition of the covariance matrix, a stabilization method [28] is applied to resolve the *ill-conditioning* of a covariance matrix. After the stabilization procedure, the two-stage *stabilization* and *smoothing* process is applied to provide a *well-conditioned* covariance matrix which is both nonsingular and positive definite.

- Stage 1. Stabilization algorithm [28]:

1. Perform spectral decomposition of $\hat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{V}$ is the matrix with eigenvectors and $\mathbf{\Lambda}$ has eigenvalues on the diagonal.
2. Calculate the mean eigenvalue $\overline{\lambda} = (\sum\limits_{i=1}^{p} \lambda_i)/p$.
3. Form a new matrix of eigenvalues as

$$\mathbf{\Lambda}^* = \begin{bmatrix} \max(\lambda_1, \bar{\lambda}) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \max(\lambda_p, \bar{\lambda}) \end{bmatrix}$$

4. Finally, recompose the new stabilized matrix

$$\hat{\mathbf{\Sigma}}_{STA} = \mathbf{V}\mathbf{\Lambda}^*\mathbf{V}^T.$$

- Stage 2: Compute the Stabilized and Smoothed Convex Sum Covariance Estimator

The second step is to feed the *stabilized* covariance matrix into a *smoothed* convex sum covariance matrix estimator (CSE) was proposed based on the quadratic loss function used by [20] and later by [8]. The stabilized and smoothed convex sum covariance estimator (STA-CSE) is as follows:

$$\hat{\mathbf{\Sigma}}_{STA\_CSE} = \frac{n}{n+m}\hat{\mathbf{\Sigma}}_{STA} + (1 - \frac{n}{n+m})\hat{\mathbf{D}}_{STA},$$

where $\hat{\mathbf{D}}_{STA} = (\frac{1}{p}tr(\hat{\mathbf{\Sigma}}_{STA}))\mathbf{I}_p$. For $p \geq 2$, m is chosen to be

$$0 < m < \frac{2[p(1+\beta) - 2]}{p - \beta},$$

where

$$\beta = \frac{(tr(\hat{\mathbf{\Sigma}}_{STA}))^2}{tr(\hat{\mathbf{\Sigma}}_{STA}^2)}.$$

This estimator improves upon $\hat{\mathbf{\Sigma}}_{STA}$ by shrinking all the estimated eigenvalues of $\hat{\mathbf{\Sigma}}_{STA}$ toward their common mean. The motivation of using both stabilization and smoothing of the covariance matrix in the ranking process of RFE subset selection is to extract more information since a reduced rank problem occurs in the kernel based methods. To remedy the current existing problems in the usual kernel based methods, the use of both stabilization and smoothing the covariance matrix is an attractive approach.

The choice of the best mapping function is not so simple and automatic. In the literature, a valid method for selecting the appropriate *kernel function* does not yet exist. The goal of SVM is to minimize the probability of misclassification error. Intuitively, then, the penalty term for a poorly-fitting model would be based on the *classification error rate*. In SVM problems, the error variance, σ^2 is estimated by the mean squared difference between actual group labels (y_i) and predicted group labels ($\hat{y}_i$) given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Now following the work of [14], the information-measure of complexity as performance measure of SVM is defined as follows:

$$ICOMP_{PERF} = n\log 2\pi + n\log(\hat{\sigma}^2) + n + 2C_{1F}(\hat{\mathbf{\Sigma}}_{STA\_CSE}),$$

where $\hat{\mathbf{\Sigma}}_{STA\_CSE}$ is the *stabilized and smoothed convex sum covariance matrix estimator* (STA-CSE) given by

$$\hat{\mathbf{\Sigma}}_{STA\_CSE} = \frac{n}{n+m}\hat{\mathbf{\Sigma}}_{STA} + (1-\frac{n}{n+m})\hat{\mathbf{D}}_{STA}, \quad \hat{\mathbf{D}}_{STA} = (\tfrac{1}{p}tr(\hat{\mathbf{\Sigma}}_{STA}))\mathbf{I}_p,$$

and

$$C_{1F}(\hat{\mathbf{\Sigma}}_{STA\_CSE}) = \frac{1}{4\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2.$$

First, the hybrid covariance estimate is calculated, and then the diagonal matrix of the largest singular values as a reduced rank approximation of $\hat{\mathbf{\Sigma}}_{STA\_CSE}$ is computed. By minimizing $ICOMP_{PERF}$, the classification error is minimized under the best fitting model. Also, $ICOMP_{PERF}$ is used to choose an optimal kernel function. One of the major motivations of introducing the information measure of complexity (ICOMP) criterion is based on the fact that in SVM-RFE subset selection problems the number of features is same from one subset to another. In such cases the models in terms of the number of parameters are considered to be equivalent. In equivalent models, AIC, BIC, or MDL type criteria do not have provision of distinguishing one equivalent model from another. Since their penalty terms are fixed, and not varying. In the literature cross-validation-based criteria has been used for feature selection. These type of criteria due to the high-dimensionality of the feature space are too time-consuming. The proposed method shortens the feature selection time.

4. Recursive Feature Elimination (RFE)

A feature selection method based on RFE has been developed by [12] which is called SVM-RFE. SVM-RFE is an application of a recursive feature elimination based on sensitivity analysis using an appropriately defined cost function ($\mathbf{w}: weight$). The SVM-Gradient-RFE method [32, 10] used the gradient as the cost function. In this paper, our cost function that we would like to use is the $ICOMP_{PERF}$. In our approach, the least sensitive feature which has the minimum value of the $ICOMP_{PERF}$ is eliminated first. This eliminated feature becomes ranking p (p: number of features). Later, the machine is retrained on the remaining $p-1$ features and then the feature with the minimum value of $ICOMP_{PERF}$ is eliminated. The process continuous in an iterative fashion until no feature is left in that subset. This means that at the end of this iterative ranking scheme all the features are ranked according to $ICOMP_{PERF}$ criterion. This is different than the Guyon's ranking scheme [12] where only weights have been considered without taking into account the model fit and the complexity of the model. This eliminated feature becomes ranking $p-1$. By doing this process repeatedly until no feature is left, the features will be ranked.

SVM-RFE Algorithm

Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ be a training set with $\mathbf{y} = (y_1, \cdots, y_n)^T$.

1. Construct a training model $\mathbf{X} = \mathbf{X}\,(:, \mathbf{s})$, where $\mathbf{s}$ is the subset of features; $\mathbf{s} = 1, 2, \cdots, p$.

2. Until all values of the cost function are obtained with the number of non-ranked features, compute the cost function for all subset

$$C(i) = (1/2)\alpha^T\mathbf{H}\alpha - (1/2)\alpha^T\mathbf{H}_{(-i)}\alpha,$$

where $\mathbf{H} = y_i y_j K\,(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{H}_{(-i)}$ is $\mathbf{H}$ matrix without the i^{th} feature.

3. Find the feature k with the smallest cost function value, and add k into the ranked subset, $\mathbf{r}$ and remove k from a subset, $\mathbf{s}$.

4. Repeat 1-3 until subset, $\mathbf{s}$ is empty.

SVM Gradient-RFE Algorithm

Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ be a training set with $\mathbf{y} = (y_1, \cdots, y_n)^T$.

1. Construct a training model $\mathbf{X} = \mathbf{X}\,(:, \mathbf{s})$, $\mathbf{s}$ is the subset of features; $\mathbf{s} = 1, 2, \cdots, p$.

2. Until all values of the average sum of the angles are obtained with the number of non-ranked features,

(i) compute the gradient, $\nabla_{(-i)} g\,(\mathbf{x})$ without i^{th} feature

$$\nabla_{(-i)} g\,(\mathbf{x}) = \sum_{m \in SV} \alpha_m y_m \nabla_{(-i)} K\,(\mathbf{x}_m, \mathbf{x})\,.$$

(ii) compute the sum of angles between $\nabla_{(-i)} g(\mathbf{x})$ and $\mathbf{e}_m, \gamma$

$$\gamma(i) = \sum_{m \in SV} \angle(\nabla_{(-i)} g(\mathbf{x}), \mathbf{e}_m),$$

where $(-i)$ means without the i^{th} feature, $\mathbf{e}_m$ is unit vectors, and
$\angle\left(\nabla_{(-i)} g\,(\mathbf{x})\,, \mathbf{e}_m\right) = min_{\beta \in \{0,1\}} \{\beta\pi + (-1)^{\beta} \arccos(\frac{\langle \nabla_{(-i)} g(\mathbf{x}) \cdot \mathbf{e}_m \rangle}{\|\nabla_{(-i)} g(\mathbf{x})\|})\}.$

(iii) compute the average sum of the angles $A(i) = 1 - \frac{2}{\pi} \cdot \frac{\gamma(i)}{|SV|}$.

3. Find the feature k with the smallest the average sum of the angle $A(i)$, add k into the ranked subset, $\mathbf{r}$ and remove k from subset, $\mathbf{s}$.

4. Repeat 1-3 until subset, $\mathbf{s}$ is empty.

Proposed SVM-$ICOMP_{PERF}$-RFE Algorithm

Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ be a training set with $\mathbf{y} = (y_1, \cdots, y_n)^T$.

1. Construct a training model $\mathbf{X} = \mathbf{X}\,(:, \mathbf{s})$, $\mathbf{s}$ is the subset of features; $s = 1, 2, \cdots, p$.

2. Until all $ICOMP_{PERF}$ values are obtained with the number of non-ranked features, compute $ICOMP_{PERF}$ based on the error rate obtained from SVM. The $ICOMP_{PERF}$ is given by

$$ICOMP_{PERF} = n \log 2\pi + n \log(\hat{\sigma}^2_{(-i)}) + n + 2C_{1F}(\hat{\mathbf{\Sigma}}_{STA\_CSE(-i)}),$$

where $\hat{\sigma}^2_{(-i)}$ is the estimated error variance without the i^{th} feature and $\hat{\mathbf{\Sigma}}_{STA\_CSE(-i)}$ is the stabilized and smoothed convex sum covariance matrix estimator without the i^{th} feature in the model.

3. Find the feature k with the smallest $ICOMP_{PERF}$, add k into the ranked subset, $\mathbf{r}$ and remove k from subset, $\mathbf{s}$.

4. Repeat 1-3 until subset, $\mathbf{s}$ is empty.

5. Numerical Results

In data mining literature, data partitioning is an important issue to find proper models for new data sets. In general, one can use different data partitioning to get different results. Most of such data partitioning schemes do not take into account of randomness that may affect the performance of the results which can be different. In the analysis, to avoid partitioning dependency, the data is randomly partitioned into 20% as one set and 80% as another set based on Pareto's principle [18]. Two experiments are performed with two different sets; $20\%/80\%$ and vice versa as training/test sets. The feature rankings corresponding to kernel functions are determined and reported for those different sets. Also, the smallest value of $ICOMP_{PERF}$, and the 95% confidence intervals (CIs) given by $\bar{X}_{error} \pm 1.96\hat{\sigma}_{error}$ for the training and test errors are reported. Ionosphere and Aorta datasets are used for these experiments. Figures 2 illustrates radar refraction by ionosphere and heart anatomy.

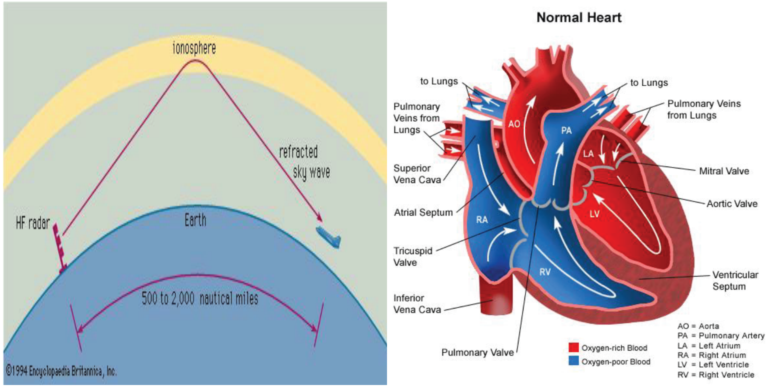


Figure 2. Illustration of radar refraction by ionosphere and heart anatomy.

Ionosphere Data

The ionosphere data is radar data which was collected by a system in Goose Bay, Labrador [26]. The system measures radar returns from the ionosphere. The data consists of 351 observations and 34 features with binary classes; good and bad returns. Figure 3 shows the scatter plots of the data with groups identified by blue (circle) and red (cross) colors. As shown in Figure 3, the separation in dimension 5 against dimensions 13, 19 and dimensions 18, 29 are quite poor. Tables 2 and 3 show performances of experiments based on $ICOMP_{PERF}$. In Table 2, the polynomial kernel with degree 3 on the 20% set shows a narrower confidence interval than other kernel functions for both training and test sets. As shown in Tables 2 and 3, the smallest $ICOMP_{PERF}$ values are obtained at the polynomial kernel with degree 3 for the 20% set and for the 80% set. Tables 4 and 5 show the best subset selection based on the smallest $ICOMP_{PERF}$ values. The training and test errors of the best subsets in both partitioned sets are within the 95% error confidence intervals.

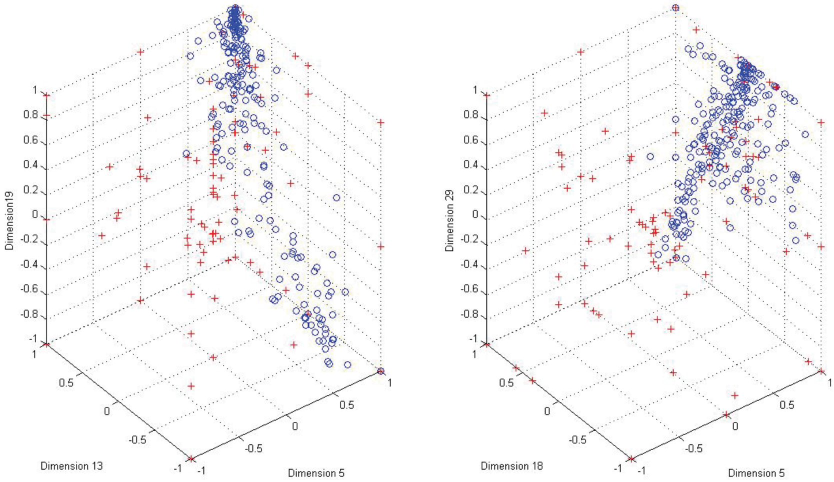


Figure 3. Grouped scatter plots for ionosphere data.

Table 2. Top subset features selected with 20% set using SVM-RFE ranking.

Kernel	Best Subset	$ICOMP_{PERF}$	CI for Training	CI for Test
Linear	$\{27, 12\}$	121.14	[0.03046, 0.33089]	[0.12241, 0.38356]
Cauchy	$\{1-9, 11-34\}$	87.61	[0.08101, 0.36773]	[0.25495,0.42540]
Polynomial(d=2)	$\{2-20, 22-30, 32-34\}$	-47953.45	[0, 0.23670]	[0.06150, 0.31593]
Polynomial(d=3)	$\{2, 3, 8, 12-14, 18, 20, 22, 24-32\}$	$\mathbf{-47957.44}$	[0, 0.14278]	[0.10669, 0.21464]

Table 3. Top subset features selected with 80% set using SVM-RFE ranking.

Kernel	Best Subset	$ICOMP_{PERF}$	CI for Training	CI for Test
Linear	$\{7\}$	606.94	[0.09676, 0.23190]	[0.09271, 0.26610]
Cauchy	$\{3\}$	441.65	[0.02329, 0.20342]	[0,0.38375]
Polynomial(d=2)	$\{5, 14\}$	454.56	[0,0.13966]	[0, 0.18645]
Polynomial(d=3)	$\{5\}$	$\mathbf{441.51}$	[0,0.09553]	[0.02858,0.02696]

Aorta Data

The aorta data are from medical imaging for a study of heart tissue. Hardening of the arteries is the leading cause of death and debility in the industrial world. Nuclear magnetic resonance (NMR) imaging has a role in diagnosing of arteries for prognosis of heart attack. The NMR aorta data was used by [19]. The dataset sampled from 418 patients on 20 different characteristics. The first group consists of 194 patients who exhibited early atheroma, and the second group consists of 224 patients who were healthy. Figure 4 shows grouped scatter plots for the poor separation of dimension 3 against dimensions 13, 19 and against dimensions 10, 20 (group1: blue, group2: red), respectively. Tables 6 and 7 show that the best subset based on $ICOMP_{PERF}$ is obtained at the Cauchy kernel in the 20% set and inverse multi-quadratic kernel in the 80% set. The confidence intervals are obtained based on $ICOMP_{PERF}$. The confidence intervals, are significantly narrow intervals in both of the sets. Tables 8 and 9 show the best subset selected based on $ICOMP_{PERF}$.

6. Comparison with Other RFE Based Methods

To compare three different RFE based methods; SVM-RFE, SVM-Gradient-RFE, SVM-$ICOMP_{PERF}$-RFE, the ionosphere and aorta datasets are used with the same kernel functions that are used in Tables 2, 3, 6, and 7. The datasets are randomly partitioned into two cases; $20\%/80\%$ and $80\%/20\%$ as training/test sets. Tables 10 and 11 present comparisons of three RFE based methods using the ionosphere data with four different kernel functions in two different cases. The average error rate represents misclassification error rate for test set. The SVM-$ICOMP_{PERF}$-RFE is the clear winner for most kernel functions except the linear kernel in the $80\%/20\%$ case. The best performance is obtained using the Cauchy kernel in the two cases with 88.12% and 93.28% accuracies. Tables 12 and 13 present comparisons of the three RFE based methods using the aorta data with four different kernel functions in two different cases. As shown in Tables 12 and 13, the SVM-$ICOMP_{PERF}$-RFE is the best method for the polynomial kernel (degree=2) with 99.99% accuracy for the $20\%/80\%$ case, the polynomial kernel (degree=2) with 99.88% accuracy for the $80\%/20\%$ case, and the inverse multi-quadratic kernel with 100% accuracy for the $80\%/20\%$ case.

Table 4. Subset selection based on $ICOMP_{PERF}$ with 20% set (Polynomial: degree=3).

Rank	Feature	$ICOMP_{PERF}$	Training Error	Test Error
1	3	185.9418	0.2	0.19217
2	14	163.3763	0.2143	0.14235
3	24	125.9411	0.1	0.1566
4	26	185.4274	0.1143	0.15302
5	13	190.4090	0.0714	0.15658
6	28	158.2199	0.0571	0.2171
7	2	254.1286	0.1	0.1637
8	8	171.2137	0.0571	0.1459
9	20	123.1558	0.0286	0.1495
10	30	143.0001	0.0143	0.1424
11	12	-47854.7927	0	0.1886
12	18	48313.953	0.0286	0.1708
13	27	136.8903	0.0143	0.1068
14	31	273.9907	0.0429	0.1388
15	25	-47934.01	0	0.1174
16	29	201.188	0	0.1352
17	32	48348.9985	0.0571	0.1779
18	**22**	$\mathbf{-47957.4425}$	**0**	**0.1708**
19	6	48366.0982	0.0571	0.1779
20	16	96.5792	0.0143	0.1851
21	5	-47867.1294	0	0.1566
22	4	48260.6735	0.0143	0.1495
23	11	-47852.1665	0	0.2242
24	10	196.314	0	0.1566
25	34	200.772	0	0.1388
26	1	48249.2818	0.0143	0.1459
27	19	-47847.3168	0	0.1957
28	33	185.951	0	0.121
29	21	205.575	0	0.1637
30	7	204.57	0	0.2135
31	23	208.216	0	0.1388
32	9	188.548	0	0.1744
33	17	48266.3703	0.0143	0.1388
34	15	-47870.1323	0	0.1566

Figure 5 shows line plots of error rates for the test set with the Cauchy kernel function which gives smallest average error rates using the ionosphere data shown in Tables 10 and 11. Figure 6 shows line plots of error rates for the test set with the polynomial kernel (degree=2) and inverse multi-quadratic kernel functions, which give smallest average error rates using the aorta data shown in Tables 12, and 13. The SVM-$ICOMP_{PERF}$-RFE is competitive with both SVM-RFE and SVM-Gradient-RFE as shown in Figure 5. Also, SVM-$ICOMP_{PERF}$-RFE outperforms both SVM-RFE and SVM-Gradient-RFE with few features as shown in Figure 6.

Table 5. Subset selection based on $ICOMP_{PERF}$ with 80% set (Polynomial: degree=3).

Rank	Feature	$ICOMP_{PERF}$	Training Error	Test Error
$\mathbf{1}$	$\mathbf{5}$	$\mathbf{441.5118}$	$\mathbf{0.1708}$	$\mathbf{0.1714}$
2	4	541.7953	0.0890	0.1143
3	14	698.4002	0.0676	0.1714
4	34	838.3473	0.0819	0.0857
5	33	717.6374	0.0605	0.1143
6	30	754.7581	0.0534	0.0857
7	18	752.3821	0.0463	0.1286
8	22	769.0320	0.0427	0.0857
9	6	772.8447	0.0391	0.0571
10	16	768.1328	0.0356	0.0857
11	31	697.4870	0.0249	0.0714
12	32	795.0805	0.0249	0.1143
13	26	834.1837	0.0285	0.0857
14	25	603.7533	0.0142	0.1571
15	10	950.3118	0.0249	0.0429
16	12	640.0070	0.0142	0.1429
17	8	717.9700	0.0107	0.1286
18	20	797.0560	0.0107	0.0429
19	2	679.8700	0.0071	0.0714
20	29	801.4970	0.0071	0.1286
21	24	911.7650	0.0107	0.1143
22	28	682.2560	0.0071	0.1143
23	21	907.2940	0.0107	0.0571
24	3	689.8410	0.0071	0.0714
25	27	911.3660	0.0107	0.1000
26	7	501.0110	0.0036	0.1286
27	23	994.9170	0.0071	0.1000
28	19	817.5010	0.0071	0.1143
29	13	612.9350	0.0036	0.0857
30	17	1008.7330	0.0071	0.0429
31	11	808.4890	0.0071	0.0714
32	15	623.0110	0.0036	0.0857
33	1	1001.9020	0.0071	0.0429
34	9	628.2170	0.0036	0.1143

Table 6. Top subset features selected with 20% set using SVM-RFE ranking.

Kernel	Best Subset	$ICOMP_{PERF}$	CI for Training	CI for Test
Cauchy	$\{4\}$	$\mathbf{-57785.1}$	[0,0]	[0,0.00767]
Gaussian	$\{12, 13, 14, 17\}$	-57071	[0,0.1171]	[0,0.2881]
Polynomial(d=2)	$\{4\}$	-57679	[0,0]	[0,0.0270]
Inv. Multi-Quadratic	$\{7, 15, 17, 20\}$	-57414.62	[0,0.0434]	[0,0.2467]

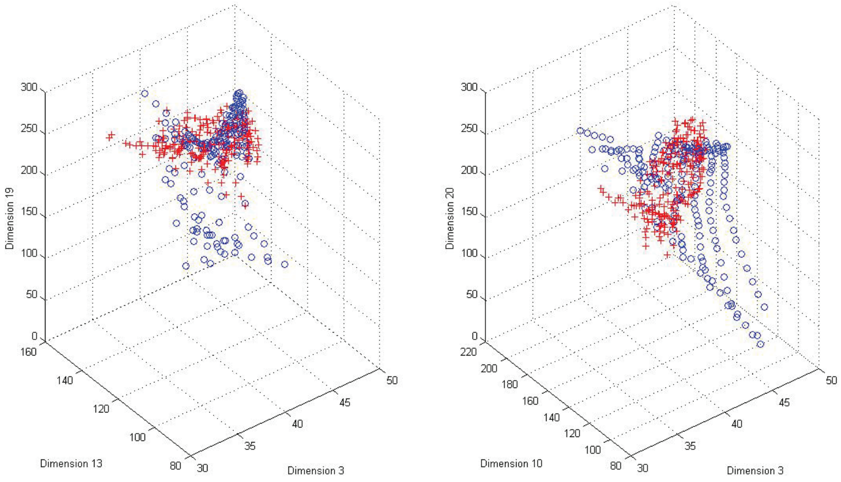
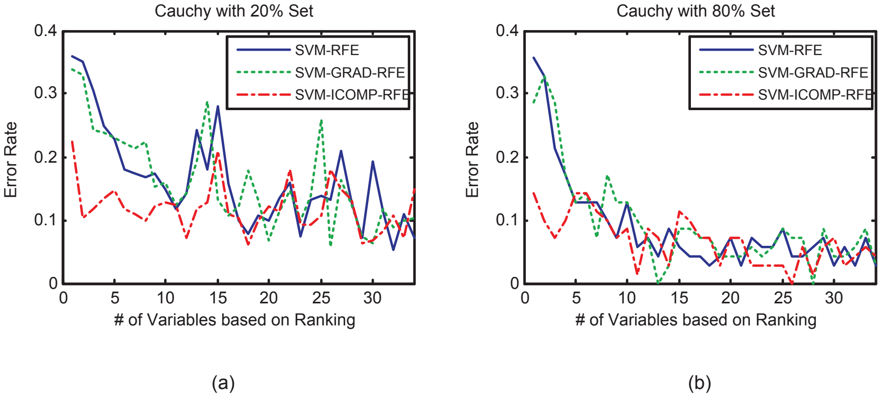


Figure 4. Grouped scatter plots for aorta data.

Table 7. Top subset features selected with 80% set using SVM-RFE ranking.

Kernel	Best Subset	$ICOMP_{PERF}$	CI for Training	CI for Test
Cauchy	$\{7, 15, 20\}$	-228526.1	[0,0.1254]	[0,0.2610]
Gaussian	$\{2\}$	-229734.4	[0,0]	[0,0.0103]
Polynomial(d=2)	$\{4\}$	-229608	[0,0]	[0,0]
Inv. Multi-Quadratic	$\{4\}$	$-\mathbf{229759.2}$	[0,0]	[0,0]



(a)

(b)

Figure 5. Best results of SVM-$ICOMP_{PERF}$-RFE using ionosphere data: (a) Cauchy kernel function with 20% set (b) Cauchy kernel function with 80% set.

7. Conclusion and Discussion

In this paper, a novel SVM-$ICOMP_{PERF}$-RFE method is proposed using an information complexity ($ICOMP_{PERF}$) criterion. SVM-RFE is used in conjunction with $ICOMP_{PERF}$ not only to choose an optimal

Table 8. Subset selection based on $ICOMP_{PERF}$ with 20% set (Cauchy).

Rank	Feature	$ICOMP_{PERF}$	Training Error	Test Error
1	**4**	$\mathbf{-57785.101}$	**0**	**0**
2	14	236.839	0	0
3	20	238.263	0	0
4	5	238.381	0	0
5	12	238.382	0	0
6	10	238.382	0	0
7	11	238.381	0	0.006
8	13	238.382	0	0.003
9	17	238.382	0	0.006
10	9	238.381	0	0
11	1	238.382	0	0
12	19	238.382	0	0
13	18	238.381	0	0
14	16	238.382	0	0
15	3	238.382	0	0
16	6	238.381	0	0.003
17	2	238.382	0	0
18	8	238.382	0	0.012
19	15	238.381	0	0
20	7	238.382	0	0

Table 9. Subset selection based on $ICOMP_{PERF}$ with 80% set (Inv. Multi-Quadratic).

Rank	Feature	$ICOMP_{PERF}$	Training Error	Test Error
1	**4**	$\mathbf{-229759.22}$	**0**	**0**
2	7	941.35	0	0
3	15	945.22	0	0
4	20	946.32	0	0
5	16	947.29	0	0
6	5	947.53	0	0
7	17	947.71	0	0
8	10	947.77	0	0
9	14	947.82	0	0
10	6	947.8	0	0
11	8	947.84	0	0
12	18	947.85	0	0
13	11	947.83	0	0
14	13	947.85	0	0
15	1	947.84	0	0
16	12	947.86	0	0
17	9	947.84	0	0
18	2	947.84	0	0
19	19	947.84	0	0
20	3	947.85	0	0

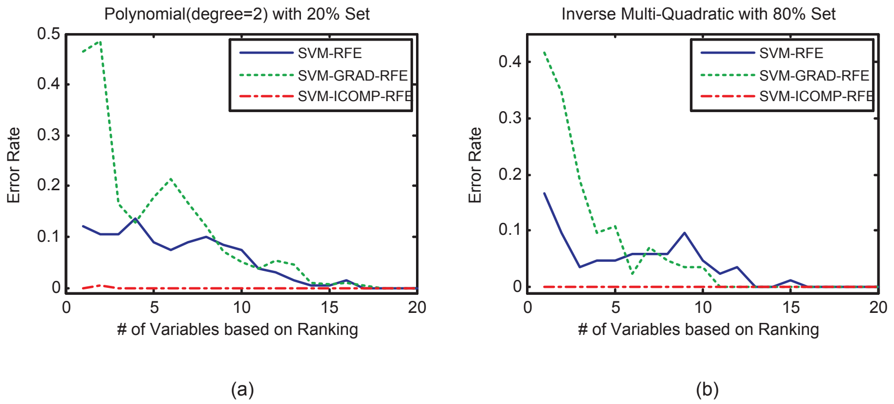


Figure 6. Best results of SVM-$ICOMP_{PERF}$-RFE using aorta data: (a) Polynomial kernel (degree=2) function with 20% set (b) Inverse Multi-Quadratic kernel function with 80% set.

Table 10. Comparison using ionosphere data with $20\%/80\%$.

	SVM-RFE	SVM-Gradient-RFE	SVM-$ICOMP_{PERF}$-RFE
	Average Error Rate	Average Error Rate	Average Error Rate
Linear	0.22273	0.19552	**0.19510**
Cauchy	0.16381	0.16140	**0.11880**
Polynomial(d=2)	0.19992	0.18903	**0.17522**
Polynomial(d=3)	0.21572	0.21195	**0.18830**

Table 11. Comparison using ionosphere data with $80\%/20\%$.

	SVM-RFE	SVM-Gradient-RFE	SVM-$ICOMP_{PERF}$-RFE
	Average Error Rate	Average Error Rate	Average Error Rate
Linear	0.15546	**0.15420**	0.16177
Cauchy	0.08908	0.09454	**0.06723**
Polynomial(d=2)	0.16933	0.13445	**0.13277**
Polynomial(d=3)	0.17941	0.15840	**0.13656**

Table 12. Comparison using aorta data with $20\%/80\%$.

	SVM-RFE	SVM-Gradient-RFE	SVM-$ICOMP_{PERF}$-RFE
	Average Error Rate	Average Error Rate	Average Error Rate
Cauchy	**0.00374**	0.13488	0.04880
Gaussian	**0.05749**	0.13084	0.10195
Polynomial(d=2)	0.05404	0.11033	**0.00015**
Inv. Multi-Quadratic	**0.02784**	0.12590	0.05434

Table 13. Comparison using aorta data with $80\%/20\%$.

	SVM-RFE	SVM-Gradient-RFE	SVM-$ICOMP_{PERF}$-RFE
	Average Error Rate	Average Error Rate	Average Error Rate
Cauchy	**0.01548**	0.07738	0.04167
Gaussian	**0.02083**	0.06310	0.03393
Polynomial(d=2)	0.03095	0.05	**0.00119**
Inv. Multi-Quadratic	0.03929	0.06845	**0**

kernel function from a portfolio of many other kernel functions, but also to select important subset(s) of features. The numerical examples on two benchmark datasets show that the proposed hybridized method exhibits a promising performance for feature subsetting and the optimal kernel selection. This method provides a unification of both $ICOMP_{PERF}$ as the feature selection criterion and RFE as the search algorithm. In this framework, $ICOMP_{PERF}$ is a key cost function. Furthermore, the hybrid covariance matrix known as stabilized and smoothed convex sum covariance estimator (STA-CSE) is used to avoid the singularity in the kernel based methods. In the literature related to recursive feature elimination such stabilization issues have not been addressed before. As shown in Tables 10, 11, 12, and 13, the comparisons of feature ranking methods demonstrate that SVM-$ICOMP_{PERF}$-RFE is a promising way to obtain the best subset of features. Further research is being currently carried out to extend these new results from binary SVM to multi-class SVM environment. The results of this research findings will be reported under a separate paper in different application areas.

REFERENCES

1. M. Aizerman, E. Braverman, and L. Rozonoer, *Theoretical foundations of the potential function method in pattern recognition learning*, Automation and Remote Control, vol. 25, pp. 821–837, 1964.
2. H. Akaike, *Information theory and an extension of the maximum likelihood principle*, in Second international symposium on information theory, edited by B.N. Petrov, and B.F. Csaki, Academiai Kiado, Budapest, pp. 267–281, 1973.
3. H. Bozdogan, *ICOMP: a new model-selection criteria*, in Classification and related methods of data analysis, edited by H.H. Bock, North-Holland, Amsterdam, 1988.
4. H. Bozdogan, *The theory and applications of information-theoretic measure of complexity (ICOMP) as a new model selection criterion*, Unpublished Report, The Institute of Statistical Mathematics, Tokyo, Japan, and the Department of Mathematics, University of Virginia, Charlottesville, VA, USA, 1988.
5. H. Bozdogan, *On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models*, Communications in Statistics Theory and Methods, vol. 19, pp. 221–278, 1990.
6. H. Bozdogan, *Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity*, in Multivariate Statistical Modeling, edited by H. Bozdogan, Academic Publishers, Dordrecht, Netherland, pp. 69–113, 1994.
7. H. Bozdogan, *Akaike's information criterion and recent develpments in information complexity*, Journal of Mathematical Psychology, vol. 44, pp. 62–91, 2000.
8. M. Chen, *Estimation of covariance matrices under a quadratic loss function*, Technical Report S–46, Department of Mathematics, SUNY at Albany, 1976.
9. P. Chen, C. Lin, and B. Scholkopf, *A tutorial on v-support vector machine*, Applied Stochastic Models in Business and Industry, vol. 21, no. 2, pp. 111–136, 2005.
10. H. Cho, S.H. Baek, E. Youn, M.K. Jeong, and A. Taylor, *A two-stage classification procedure for near-infrared spectra based on multi-scale vertical energy wavelet thresholding and SVM-based gradient-recursive feature elimination*, Journal of the Operational Research Society, vol. 60, no. 8, pp. 1107–1115, 2009.
11. H. Frohlich, *Feature selection for support vector machines by means of genetic algorithms*, M.S. Thesis, University of Tuebingen, 2002.
12. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Gene selection for cancer classification using support vector machines*, Machine Learning, vol. 46, no. 1/3, pp. 389–422, 2002.
13. C.J. Harris, *An information theoretic approach to estimation*, in Recent Theoretical Developments in Control, edited by M.J. Gregson, Academic Press, London, pp. 563–590, 1978.
14. J. Howe, and H. Bozdogan, *Regularized SVM classification with a new complexity-driven stochastic optimizer*, European Journal of Pure and Applied Mathematics, vol. 9, no. 2, pp. 216–230, 2016.
15. S. Kullback, *Information theory and statistics*, Dover Publications, New York, 1968.
16. S. Kullback, and R. Leiber, *On information and sufficiency*, Annals of Mathematical Statistics, vol. 22, pp. 79–86, 1951.
17. S. Mika, *Kernel fisher discriminants*, Ph.D. Dissertation, Technical University of Berlin, 2002.

18. V. Pareto, *Manual of political economy*, Kelly, New York, 1909.
19. J. Pearlman, *Nuclear magnetic resonance spectral signatures of liquid crystals in human atheroma as basis for multi-dimensional digital imaging of atherosclerosis*, Ph.D. Dissertation, University of Virginia, Charlottesville, VA, 1986.
20. S. Press, *Estimation of a normal covariance matrix*, Technical Report P–5436, The Rand Corporation, Santa Monica, CA, 1975.
21. J. Rissanen, *Stochastic complexity and modeling*, Annals of Statistics, vol. 14, pp. 1080–1100, 1986.
22. J. Rissanen, *Stochastic complexity*, Journal of Royal Statistical Society: Series B, vol. 49, no. 3, pp. 223–239 252–265, 1987.
23. J. Rissanen, *Stochastic complexity in statistical inquiry*, World Scientific Publishing Company, New Jersey, 1989.
24. G. Schwarz, *Estimating the dimension of a model*, Annals of Statistics, vol. 6, pp. 461–464, 1978.
25. C.E. Shannon, *A mathematical theory of communication*, Bell Systems Technology Journal, vol. 27, pp. 279–423, 1948.
26. V. Sigillito, S. Wing, L. Hutton, and K. Baker, *Classification of radar returns from the ionosphere using neural networks*, Johns Hopkins APL Technical Digest, vol. 10, pp. 262–266, 1989.
27. A. Smola, and B. Scholkopf, *A tutorial on support vector regression*, Statistics and Computing, vol. 14, pp. 199–222, 2004.
28. C. Thomaz, *Maximum entropy covariance estimate for statistical pattern recognition*, Ph.D. Dissertation, Imperial College London, 2004.
29. M.H. Van Emden, *An analysis of complexity*, Mathematisch Centrum, Amsterdam, 1971.
30. V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, 1995.
31. S. Watanabe, *Pattern recognition: human and mechanical*, Wiley, New York, 1985.
32. E. Youn, *Feature selection in support vector machines*, M.S. thesis, University of Florida, 2002.