White box RBF classifiers with component selection for clinical prediction models

Vanya Van Belle[1,2] and Paulo Lisboa[2]

[1] ESAT-SCD / iMinds-KU Leuven Future Health Department, KU Leuven, Leuven, Belgium
[2] Department of Mathematics and Statistics, Liverpool John Moores University, Liverpool, UK

Abstract. Support vector machines for classification are very powerful methods to obtain classifiers for complex problems. Although the performance of these methods is consistently high and non-linearities and interactions between variables can be handled efficiently when using non-linear kernels such as the Radial Basis Function (RBF) kernel, their use in domains where interpretability is an issue is hampered by their lack of transparency. Many feature selection algorithms have been developed to allow for some interpretation but the impact of the different input variables on the prediction still remains unclear. Alternative models using additive kernels are restricted to main effects, reducing their usefulness in many applications. This paper proposes a new approach to expand the RBF kernel into interpretable and visualizable components, including main and two-way interaction effects. In order to obtain a sparse model representation, an iterative l_1-regularized parametric model using the interpretable components as inputs is proposed. Results on toy problems illustrate the ability of the method to select the correct contributions and an improved performance over standard RBF classifiers in the presence of irrelevant input variables. The method is illustrated on two real life UCI datasets.
Keywords: Support vector machines, RBF kernel, white box methods, interpretability, sparsity

1 Introduction

Machine learning methods [46, 41, 40] are increasingly used to classify data. They are specifically powerful in higher dimensions and when the effects of the variables are assumed to be non-linear or interacting with each other. A disadvantage of these methods is their inherent black-box nature and as such the resulting models do not reveal any information on the contribution of each specific input variable on the predicted outcome. In many applications, such as medical and financial decision making, interpretability of the prediction model is considered more important than best performance. The use of standard machine learning methods in practice is therefore hampered in these domains.

Interpretability of prediction models can have different meanings. In this work we will concentrate on two parts of interpretable models. Firstly, unnecessary variables should be discarded in the final model. Secondly, the impact of the value of the different input variables on the prediction should be clear. Both of these requirements have been studied in the literature, but weaknesses in the proposed approaches still remain and methods simultaneously tackling both aspects are rare. Different feature selection methods for support vector machines (SVM) and in extension for least-squares support vector machines (LS-SVM) have been proposed. Three main approaches can be identified. A first approach filters irrelevant inputs out before building the classifier on the selected set. One possibility is to rank inputs according

to some criterion, e.g. Fisher's criterion, Pearson correlation or mutual information criteria [16, 14]. More advanced approaches such as RELIEF and FOCUS have been proposed in [23, 39, 1]. Although filter approaches are very efficient w.r.t. computation, this approach might not be optimal [24, 18]. A second approach involves wrappers that use the performance of a specific classifier to rank subsets of variables. The least informative input (or set of inputs) is removed in an iterative procedure until convergence. One example is the recursive feature elimination SVM (SVM-RFE) [19], that iteratively eliminates the input with the lowest difference in the margin when calculating the kernel matrix without this input. Similar approaches using different ranking functions were proposed in [49, 34]. More recent work has focused on the embedding of feature selection within the classifier. Many of these approaches solve the feature selection task by replacing the 2-norm in standard SVMs by a 0-norm, a 1-norm or approximations and combinations of these [13, 3, 48, 4, 31]. A drawback of these approaches is that feature selection is performed in the primal model formulation, restricting its use to linear models. Several methods are reported to deal with feature selection in the dual formulation. However, these methods most often result in sparsity in the features determined in feature space and not in the input space. Since the resulting features can not be interpreted in function of the input variables, these methods are not suitable for applications where interpretability is an issue. Only some approaches study the combination of feature selection in input space while optimizing the dual problem formulation as (a relaxation of) mixed integer programming problems [28, 43]. [27] proposed to learn an anisotropic kernel, where the bandwidth w.r.t. the different inputs was varied and inputs with a large bandwidth are subsequently eliminated.

In order to allow for an explanation of the model's prediction, models are often restricted to be additive [20, 32]. Thanks to the additive structure, the contribution of each input variable to the prediction is clear. However, several classification problems can not be solved using a sum of main effects. The use of anova models [38], extending the additive structure to incorporate a number of predefined interaction terms, offers a solution to this problem. In its general form, the anova decomposition is composed as the sum of the main effects and all possible combinations of inputs. For most practical applications demanding an interpretable prediction model, reducing this decomposition to main and two-way interaction effects is sufficient. An additional advantage of this approach is the possibility to visualize the effects and thus enable validation of the resulting models by experts in the domain. Anova models for component selection where proposed in [5, 26, 35]. The kernel approach taken in [17] for regression problems is most strongly related to the work presented here for classification. They replace the kernel by means of a weighted sum of kernels. The problem is then solved by iteratively solving two convex optimization problems: (i) solve the problem in the Lagrange multipliers, fixing the weights in the sum of kernels; and (ii) solve the problem in the weights, fixing the Lagrange multipliers. Their approach is restricted to kernels without hyperparameters to reduce computational load.

The goal of this work is to combine component selection with SVMs using the Radial Basis Function (RBF) kernel in order to obtain flexible but interpretable models. We propose to replace the RBF kernel by a truncated version, containing only main and two-way interaction effects. Using this kernel, a standard SVM is solved. In a second step, the different contributions to the prediction of the SVM classifier are calculated and used as input variables for a

linear and iteratively reweighted l_1-regularized SVM. The result is a white box RBF classifier with component selection.

The remainder of the paper is organized as follows. Section 2 starts with introducing the notations used throughout the paper and summarizes support vector machines for classification. In Section 2.2 we illustrate how the RBF kernel can be represented as a sum of kernels evaluated on subsets of the input variables. Section 2.3 proposes a method to obtain sparse results. Section 3 discusses the model selection aspects of this work. Our approach is illustrated on toy problems and real life classification problems in Section 4. Section 5 summarizes some final conclusions.

2 A white box RBF classifier

In this Section, we propose a novel approach to obtain sparse and interpretable classifiers that are able to select relevant (non-)linear and interaction effects. The standard RBF kernel is truncated to only include main and two-way interaction effects. These effects are then combined in a sparse way by solving an iteratively reweighted l_1-regularized SVM in primal space.

2.1 Support vector classifier

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a set of observations, with $x_i \in \mathbb{R}^d$ the input variables of observation i and $y_i \in \{-1, 1\}$ the corresponding class label. The standard SVM for classification [46] is then formulated as

$$\begin{aligned} \min_{w,b,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^N \epsilon_i \\ \text{subject to} \quad & \begin{cases} y_i \left(w^T \varphi(x_i) + b \right) \geq 1 - \epsilon_i, & \forall\, i = 1, \ldots, N \\ \epsilon_i \geq 0, & \forall\, i = 1, \ldots, N\,. \end{cases} \end{aligned} \tag{1}$$

In this notation, $\varphi(\cdot)$ represents a feature map, mapping the input variables into a (possibly infinite) feature space; $w \in \mathbb{R}^{d_\varphi}$ is a coefficients vector and γ is a strict positive regularization parameter making the trade-off between smoothness and correct classification of the training data. When solving this problem in primal space, the feature map needs to be specified explicitly and a prediction for a new point $x_\star$ is obtained from

$$\hat{y} = \text{sign}(w^T \varphi(x_\star) + b)\,.$$

Defining the Lagrangian of problem (1), and deriving the Karush-Kuhn-Tucker conditions yields the dual problem formulation

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i,j=1}^N y_i y_j \varphi(x_i)^T \varphi(x_j) \alpha_i \alpha_j - \sum_{i=1}^N \alpha_i \\ \text{subject to} \quad & \begin{cases} \displaystyle\sum_{i=1}^N \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq \gamma, \quad \forall\, i = 1, \ldots, N\,. \end{cases} \end{aligned} \tag{2}$$

An advantage of this approach is that the feature map $\varphi(x)$ does not need to be constructed explicitly. Any continuous function $K(x, x_\star)$ for any points x and $x_\star$ satisfying Mercer's condition [29] can be expressed as an inner product

$$K(x, x_\star) = \varphi(x)^T \varphi(x_\star) \,.$$

The classifier then becomes

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x_\star) + b \right) \,.$$

In many applications, the Radial Basis Function (RBF) is chosen as the kernel since it is able to model non-linearities and interactions between variables automatically and is bounded. A drawback of using a non-additive kernel like the RBF is that the resulting classifier is a black-box model, not revealing any information on the way the predictions are obtained. In the next Section, it is shown how the RBF kernel can be approximated to obtain a white box classifier.

2.2 Truncated radial basis functions

Several additive kernels, such as the polynomial and clinical kernel [7], can be used to enable interpretability. However, in practice not all problems can be solved by main effects. Anova kernels offer a solution to this problem [38], but prior knowledge is needed in order to define which terms should be included in the anova decomposition.

In this work, we propose to expand the RBF kernel and to truncate its contributions to main and two-way interaction effects as follows. The RBF kernel is defined as $K_{\text{RBF}}(x, z) = \exp\left(-\frac{||x-z||_2^2}{\sigma^2}\right)$, with x and $z \in \mathbb{R}^d$. Using the Taylor expansion of the exponential function $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, the RBF kernel can be written as

$$K_{\text{RBF}}(x, z) = \sum_{n=0}^{\infty} \frac{(-1)^n (||x - z||_2^2)^n}{n! \sigma^{2n}} \,.$$

Using the multinomial theorem

$$(x^1 + x^2 + \cdots + x^d)^n = \sum_{k_1 + \cdots + k_d = n} \binom{n}{k_1, \ldots, k_d} \prod_{1 \leq p \leq d} (x^p)^{k_p} \,,$$

with x^p the p^{th} variable of x, this becomes

$$
\begin{aligned}
K_{\text{RBF}}(x,z) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \left[\sum_{p=1}^{d} (x^p - z^p)^{2n} + \sum_{\substack{\sum_{l=1}^{d} k_l = n \\ k_l \neq n}} \binom{n}{k_1, \ldots, k_d} \prod_{1 \leq p \leq d} (x^p - z^p)^{2k_p} \right] \\
&= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{p=1}^{d} (x^p - z^p)^{2n} \\
&\quad + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{\substack{k_p + k_q = n \\ k_p, k_q \neq n}} \binom{n}{k_p, k_q} (x^p - z^p)^{2k_p} (x^q - z^q)^{2k_q} \qquad (3) \\
&\quad + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{\substack{\sum_{l=1}^{d} k_l = n \\ k_l \neq n \\ k_l + k_m \neq n}} \binom{n}{k_1, \ldots, k_d} \prod_{1 \leq p \leq d} (x^p - z^p)^{2k_p} .
\end{aligned}
$$

The first term in (3) represents the contributions of single input variables (main effects), the second term represents all two-way interaction effects and the last term represents all interaction effects with more than two variables involved. In order for the results to be interpretable and explainable, we will focus on the first two terms since these can be visualized. For most applications where interpretability is an issue it suffices to take two-way interactions into account. Using equation (3), the RBF kernel evaluated on two 2-dimensional vectors $x^{p,q}$ and $z^{p,q}$ can be expressed as

$$
\begin{aligned}
K_{\text{RBF}}(x^{p,q}, z^{p,q}) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \left[(x^p - z^p)^{2n} + (x^q - z^q)^{2n} \right] \qquad (4) \\
&\quad + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{\substack{k_p + k_q = n \\ k_p, k_q \neq n}} \binom{n}{k_p, k_q} (x^p - z^p)^{2k_p} (x^q - z^q)^{2k_q} ,
\end{aligned}
$$

and contains the main effects of both input variables and their interaction effect. The truncated RBF kernel is then defined as a combination of RBF kernels evaluated in each pair of input variables:

$$
K^{\text{tr}}_{\text{RBF}}(x,z) = \frac{2}{d(d-1)} \sum_{p=1}^{d} \sum_{q>p}^{d} K_{\text{RBF}}(x^{p,q}, z^{p,q}) .
$$

Replacing the RBF kernel with its truncated version, the prediction of the classifier for a new point $x_\star$ is obtained from

$$\begin{aligned}
\hat{y} &= \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K_{\text{RBF}}^{\text{tr}}(x_i, x_\star) + b\right) \\
&= \text{sign}\left(\frac{2}{d(d-1)} \sum_{i=1}^{N} \alpha_i y_i \left(\sum_{p=1}^{d}\sum_{q>p}^{d} K_{\text{RBF}}(x_i^{p,q}, x_\star^{p,q})\right) + b\right) \\
&= \text{sign}\left(\frac{2}{d(d-1)} \sum_{p=1}^{d}\sum_{q>p}^{d}\sum_{i=1}^{N} \alpha_i y_i K_{\text{RBF}}(x_i^{p,q}, x_\star^{p,q}) + b\right) \\
&= \text{sign}\left(\sum_{p=1}^{d}\sum_{q>p}^{d} \hat{y}^{p,q} + b\right) .
\end{aligned}$$

2.3 Parsimonious RBF classifiers

The partial contributions $\hat{y}^{p,q}$ are a weighted sum of RBF kernels that are evaluated in 2-dimensional vectors. As such, the partial contribution $\hat{y}^{p,q}$ contains the main effects of both variables and an interaction effect. In order to be able to select all of these effects separately, $\hat{y}^{p,q}$ is extracted in three components: (i) $\hat{y}^p$ which is built upon the first term in equation (4), (ii) $\hat{y}^q$ which is built upon the second term in equation (4), and (iii) a contribution of the interaction expressed as $\hat{y}^{p,q} - \hat{y}^p - \hat{y}^q$. Note that this extraction can not be made on the level of the kernel due to the necessity of the kernel in the SVM classifier to be positive semidefinite.

Let $\tilde{y}^p$ be the normalized version of $\hat{y}^p$ with zero mean and a standard deviation of 1 and $\tilde{y}^{p,q}$ the normalized version of $\hat{y}^{p,q} - \hat{y}^p - \hat{y}^q$ and denote these as the partial contributions or components of the predictor. These partial contributions are then used as inputs for a linear and iteratively reweighted l_1-regularized SVM classifier [4] with non-negative coefficients.

$$\min_{\beta, b^*, \epsilon^*} \sum_{p=1}^{d} \chi^p \beta^p + \sum_{p=1}^{d}\sum_{q>p}^{d} \chi^{p,q} \beta^{p,q} + \gamma^* \sum_{i=1}^{N} \epsilon_i^*$$

subject to

$$\begin{cases}
y_i \left(\sum_{p=1}^{d} \beta^p \tilde{y}^p + \sum_{p=1}^{d}\sum_{q>p} \beta^{pq} \tilde{y}^{pq} + b^*\right) \geq 1 - \epsilon_i^*, & \forall\, i = 1, \ldots, N \\
\epsilon_i^* \geq 0, & \forall\, i = 1, \ldots, N \\
\beta^p \geq 0, & \forall\, p = 1, \ldots, d \\
\beta^{pq} \geq 0, & \forall\, p = 1, \ldots, d; q = p+1, \ldots, d\,,
\end{cases} \tag{5}$$

where χ^p equals 1 in the first iteration and is defined as

$$\chi^p = \frac{1}{\varepsilon + c\beta^p}\,, \tag{6}$$

in the next iterations. Here, ε is a small, predefined constant (e.g. 0.005) and c a parameter to control the sparsity of the solution [6]. Method (5) is iterated until the average of the absolute value of the difference between the β-vectors in two iterations is less then 10^{-8}. The 1-norm penalty was first introduced by [44] as the Least Absolute Shrinkage and Selection Operator in the context of linear regression. In equation (5) the coefficients are restricted to be non-negative since all the components are assumed to correlate with the outcome. Remark the link with the non-negative garrote estimator [5, 52], which was originally proposed to shrink the estimates from least-squares regression.

The procedure to obtain an interpretable and sparse classifier is summarized in Algorithm 1. A description on how the different parameters are tuned in our experiments follows in Section 3. The results can be further improved by iterating until the selected set of components remains unchanged. In practical applications, this is achieved after two to four iterations.

Algorithm 1 Procedure to obtain a sparse white box RBF classifier.

1: Determine the optimal tuning parameters γ and σ for the truncated RBF kernel in (2).
2: Given the optimal value of γ and σ, solve equation (2) to obtain α.
3: Given α, estimate the partial contributions $\tilde{y}^p$ and $\tilde{y}^{p,q}$.
4: Given the partial contributions, determine the optimal value of γ^* in equation (5) with a fixed value of $c = 1$.
5: Given γ^*, determine the optimal value of c in equation (5).
6: Given c and γ^*, solve equation (5) to obtain the sparse model representation.
7: Obtain the prediction as

$$\hat{y} = \text{sign}\left(\sum_{p=1}^{d} \beta^p \tilde{y}^p + \sum_{p=1}^{d}\sum_{q\geq p}^{d} \beta^{p,q}\tilde{y}^{p,q} + b^*\right)\,.$$

3 Tuning of the parameters

The performance of the proposed approach depends on the value of several parameters. In addition to the tuning of the parameters involved in standard SVMs, other parameters need to be set to an appropriate value and the optimal value of some of them are related. In the experiments, the parameters were tuned according to the following scheme.

Tuning of the bandwidth of the (truncated) RBF and the regularization parameter γ in equation (2) is performed by means of coupled simulated annealing [50]. The parameter values were randomly initialized, where σ was scaled with $\sqrt{d}$. The procedure started from 10 different initializations. The parameter combination leading to the best 10-fold cross-validation area under the receiver operating characteristic curve (AUC) was selected.

The value of c in equation (5) is dependent on the value of γ^*. A high value of γ^* inhibits a sparse solution, whatever the value of c. Tuning both parameters simultaneously would necessitate the use of a risk measure capturing the trade-off between sparsity and performance.

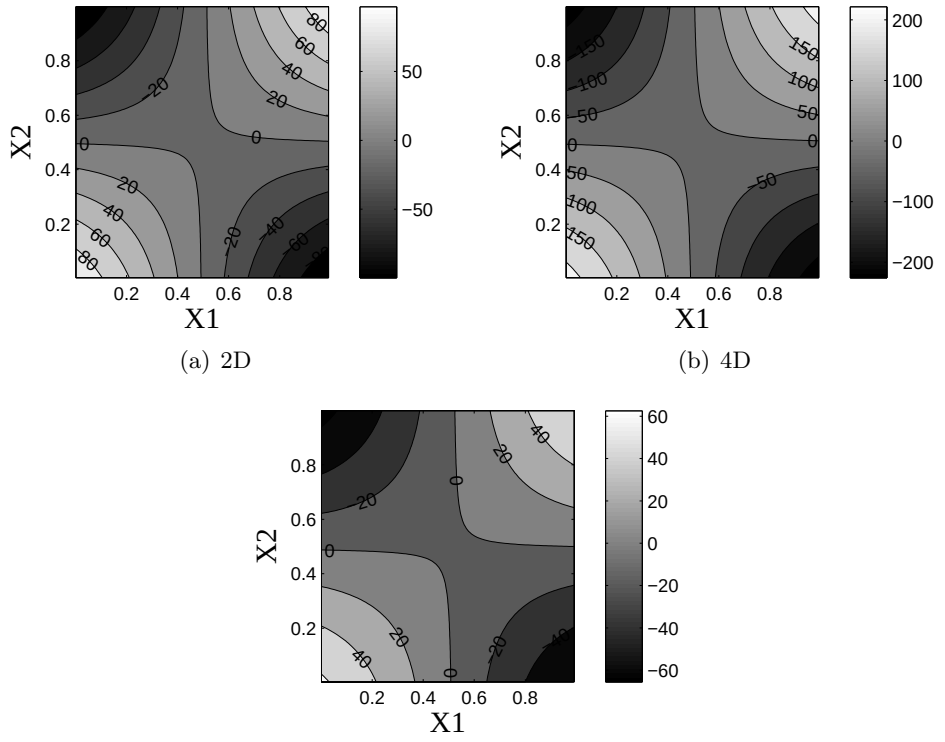
Since it is not clear in advance which trade-off is realistic, this choice is left open for discussion. In the experiments, the value of γ^* was tuned by means of 5-fold cross-validation, with $c = 1$. The grid over which γ^* was varied was defined as an exponential grid on $[0.01, 1000]$. The AUC was used as model selection criterion. Using the tuned value of γ^*, the value of c was varied, and the 5-fold cross-validation AUC was reported. To reduce computational load, the range of values over which c is varied was restricted to values for which the resulting coefficients vector w yielded 1 to $3d$ non-zero elements. The optimal value of c was defined as the lowest value yielding an AUC on 5-fold cross validation that did not lead to a significant reduction in AUC ($p>0.05$) according to the test of DeLong [9]. In order to be able to compare the results over folds, a logistic regression model was trained in each training-test split, converting uncalibrated latent variables to calibrated probabilities [33]. Other selection schemes are possible and might result in slightly different results. However, thanks to the visualization of the different components, interpretation of the selected components is possible and irrelevant terms can often be detected.

4 Results

This Section illustrates the use of the presented method on artificial and real-life data. Toy problems illustrate the ability of the model to detect the relevant components, whilst being as performant as standard SVMs. Two datasets from the UCI machine learning depository [12] are used to compare our results with results from other methods described in the literature. In all the experiments γ and γ^* were scaled with $\frac{N}{N_+}$ and $\frac{N}{N_-}$, where N_* indicates the number of observations in class $*$, for elements belonging to the positive and negative class respectively. The method was iterated until the set of selected components did no longer change or the maximal number of iterations (here 10) was exceeded. In each iteration the components selected in the previous iteration where taken into account in addition to the main effects of the variables involved in a selected interaction effect. The reported confidence intervals where calculated by means of the bias corrected and accelerated percentile method using 1000 bootstraps samples.

4.1 Artificial example 1: the XOR problem

In this first experiment, the XOR problem is considered in three different dimensions: 2D, 4D and 10D. In all three settings, only the first two variables are relevant. All variables are independently drawn from a uniform distribution. The proposed method (l_1-SVM-RBF$^{\text{tr}}$) selects a single component in all three cases: the interaction between the first and second input variable (see Figure 1). Table 1 compares the results of the presented method with two standard SVMs using an RBF kernel: one using all variables, and one using those variables that were selected by our method. Note that when our method selects a single interaction effect, the use of the standard RBF kernel involves using that effect together with both main effects. Since our method is able to build a classifier only using a selected set of variables, the performance does not drop when increasing the number of irrelevant features. The standard SVM classifier suffers from overfitting when more irrelevant features are included. The performance increases again when restricting the used feature set to the ones selected by our method.



(a) 2D

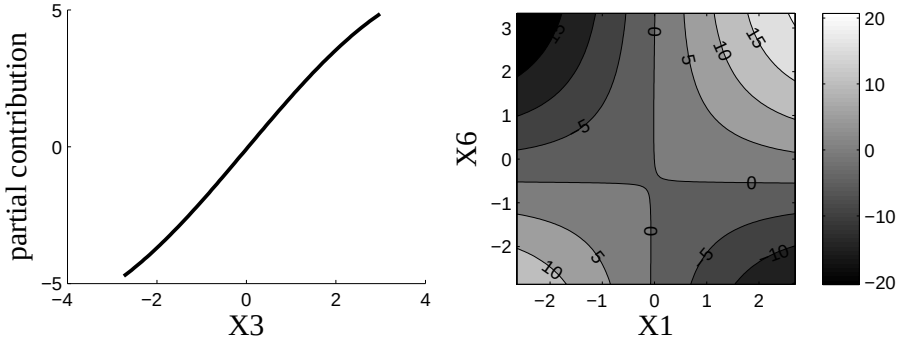
(b) 4D

(c) 10D

Fig. 1. Illustration of the selected effects for the XOR problem.

Table 1. Comparison of the proposed method with standard SVMs using an RBF kernel on the XOR problems in three different settings. For every dataset, only the first two input variables contribute to the class labels.

method	AUC (95% CI)	ACC (95% CI)	BER (95% CI)
2-dimensional problem			
SVM-RBF (all)	1.000 (0.999-1.000)	0.996 (0.972-1.000)	0.004 (0.000-0.024)
SVM-RBF (subset)	1.000 (0.999-1.000)	0.996 (0.980-1.000)	0.004 (0.000-0.028)
l_1-SVM-RBF$^{\text{tr}}$	1.000 (0.999-1.000)	0.996 (0.980-1.000)	0.004 (0.000-0.018)
4-dimensional problem			
SVM-RBF (all)	0.995 (0.989-0.998)	0.956 (0.919-0.972)	0.046 (0.022-0.073)
SVM-RBF (subset)	1.000 (0.999-1.000)	0.996 (0.980-1.000)	0.004 (0.000-0.028)
l_1-SVM-RBF$^{\text{tr}}$	1.000 (1.000-1.000)	0.996 (0.976-1.000)	0.004 (0.000-0.027)
10-dimensional problem			
SVM-RBF (all)	0.947 (0.917-0.966)	0.852 (0.800-0.888)	0.147 (0.109-0.196)
SVM-RBF (subset)	1.000 (0.998-1.000)	0.992 (0.972-0.996)	0.007 (0.000-0.027)
l_1-SVM-RBF$^{\text{tr}}$	0.997 (0.990-0.999)	0.976 (0.944-0.988)	0.027 (0.007-0.040)



(a) main effects (b) interaction effects

Fig. 2. Artificial example 2. Selected main and interaction effects.

4.2 Artificial example 2

In this second experiment, a classification problem with an underlying logistic regression function is used to illustrate the ability of the model to select the correct relevant variables. A dataset with 10 variables is created by means of a multivariate Gaussian distribution. The variables are uncorrelated except for the first three variables, whose correlation matrix is

$$\begin{bmatrix} 1 & 0.8 & 0.2 \\ 0.8 & 1 & 0.1 \\ 0.2 & 0.1 & 1 \end{bmatrix}.$$

The probability of an observation to belong to the positive class is modeled by

$$P(\text{class } 1|\mathrm{x}_1, \ldots, \mathrm{x}_{10}) = \frac{\exp(5\mathrm{x}_1 + 5\mathrm{x}_3 + 10\mathrm{x}_1\mathrm{x}_6)}{1 + \exp(5\mathrm{x}_1 + 5\mathrm{x}_3 + 10\mathrm{x}_1\mathrm{x}_6)}.$$

The method detects a main effect for x_3 and an interaction effect for x_1 and x_6. The main effect of x_1 is not selected. This is due to the fact that the split in main and interaction effects, without specification of the form of these effects, is not unique in an additive model. Additionally, the effect size of x_1 is smaller than the effect sizes for both other relevant components, and might not be reflected in the AUC. Figure 2 illustrates the selected effects. Table 2 compares the results with the standard approach using all features and the selected subset (x_1, x_3, x_6). The performance of all methods are comparable but l_1-SVM-RBF$^{\text{tr}}$ offers a way to interpret the results.

Table 2. Comparison of the test set performance (artificial example 2) of the presented method with a standard SVM using an RBF kernel on the whole set of variables and the selected subset.

method	AUC (95% CI)	ACC (95% CI)	BER (95% CI)
SVM-RBF (all)	0.975 (0.954-0.987)	0.908 (0.860-0.936)	0.084 (0.062-0.135)
SVM-RBF (subset)	0.995 (0.984-0.998)	0.956 (0.920-0.976)	0.040 (0.025-0.082)
l_1-SVM-RBF$^{\text{tr}}$	0.994 (0.983-0.998)	0.940 (0.900-0.960)	0.054 (0.037-0.099)

4.3 Stability analysis

In a last artificial example, the stability of the selected components and obtained performance is tested. The setting is the same as in the previous example, but the underlying model is now defined as

$$P(\text{class } 1|\mathrm{x}_1,\ldots,\mathrm{x}_{10}) = \frac{\exp(5\mathrm{x}_2 + 10\mathrm{x}_1\mathrm{x}_3)}{1+\exp(5\mathrm{x}_2 + 10\mathrm{x}_1\mathrm{x}_3)}\,.$$

The dataset is split into a training and test set. Ten different initializations of the parameters γ and σ and different splits into folds are used to investigate the stability of the method. Variation in the parameters γ^* and c will be less since they are evaluated on a fixed grid. Their optimal value will vary by their dependence on the split in folds of the training set. The results are summarized in Table 3. In six out of ten initializations, the selected components are x_2 and x_1x_3. In the remaining four initializations, the correct components are selected but a subset of $\{x_1$,x_3,$x_2x_3\}$ is also selected. Due to correlations between variables and the non-unique split between main and interaction effects in an additive model, the method is not always able to select the components we expect. The method can therefor be stabilized by repeated subsampling of the training data. A final model can then be built on the complete training set, only including those components that are selected in the majority of the subsamples.

Table 3. Comparison of the test set performance in the stability analysis of the presented method with a standard SVM using an RBF kernel on the whole set of variables and the selected subset.

method	AUC (std)	ACC (std)	BER (std)
SVM-RBF (all)	0.954 (0.004)	0.862 (0.009)	0.143 (0.008)
SVM-RBF (subset)	0.982 (0.000)	0.944 (0.000)	0.058 (0.000)
l_1-SVM-RBF$^{\text{tr}}$	0.975 (0.004)	0.921 (0.014)	0.078 (0.013)

4.4 The Pima Indians Diabetes dataset

This dataset contains information on eight continuously measured variables for 768 females, aged 21 or more, of Pima Indian heritage. The goal is to predict whether these women have diabetes. Observations with a zero value for plasma glucose, body mass index or blood pressure (n=44) were assumed to be missing values and were removed from the dataset. The proposed method was performed on ten randomizations between training (two thirds of the data) and test set (one third of the data). The results are compared with a standard SVM using an RBF kernel with all inputs and the selected subset of inputs in Table 4. The proposed method is competitive with a standard SVM using an RBF kernel, but offers an interpretable model representation. Plasma glucose and body mass index are selected in all ten randomizations. Age is selected in 9 out of ten randomizations. In four cases, other variables are selected as well. Given these results, we trained a model on all the data, restricting the components to the components that were selected in more than five randomization: the main effects of plasma glucose, body mass index and age. The estimated effects of the selected components are illustrated in Figure 3.

Table 4. Comparison of the test set performance (mean and std) for the Pima Indians Diabetes dataset of the presented method with a standard SVM using an RBF kernel on the whole set of variables and the selected subset. The results illustrate that the presented method (l_1-SVM-RBF$^{\text{tr}}$) is competitive with the standard SVM with the additional advantage of being interpretable.

method	AUC	ACC	BER
SVM-RBF (all)	0.826 (0.014)	0.759 (0.019)	0.280 (0.019)
SVM-RBF (subset)	0.826 (0.020)	0.762 (0.021)	0.271 (0.026)
l_1-SVM-RBF$^{\text{tr}}$	0.826 (0.022)	0.767 (0.020)	0.269 (0.022)

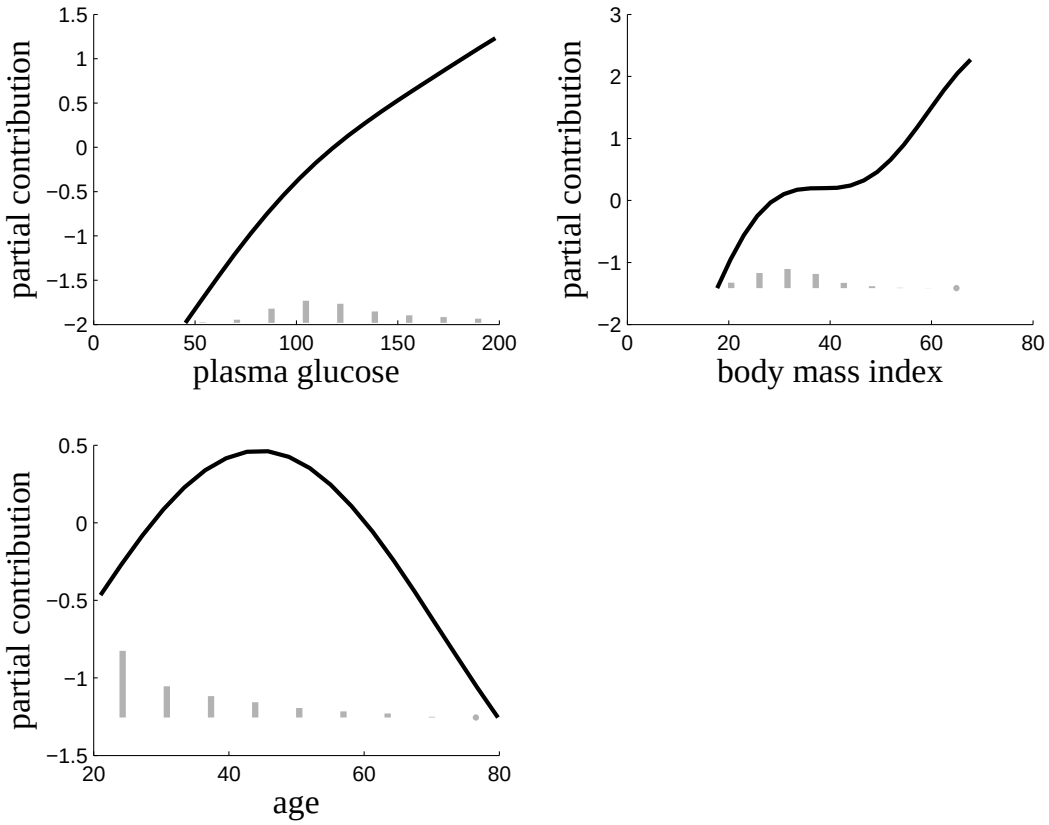


Fig. 3. Illustration of the selected features and their effects on the prediction of diabetes in the Pima Indian Diabetes dataset. The gray bars indicate the number of observations in the corresponding variable range.

To validate the feature selection process, the results are compared with different results reported in the literature. Table 5 shows that the selected features were also identified as important features by different other types of feature selection and/or ranking methods.

Table 5. Comparison of the feature selection results on the Pima Indian Diabetes problem with results from the literature. For ranking methods, we indicated the set of variables with an equal number of variables as detected by the proposed method (l_1-SVM-RBF$^{\text{tr}}$).

method	origin of results	nb pregnant	plasma glucose	blood pressure	skin fold	serum insulin	body mass index	pedigree function	age
Wang and Wang (2009)	[47]	X	X						X
SUD: Dash et al. (1997) [8]	[47]	X				X			X
Relief-F: Kononenko (1994)[25]	[47]	X	X						X
K-means: Girolami and He (2003) [15]	[47]			X	X				X
MHCC: Yacob (2012) [51]	[51]	X	X		X		X	X	X
Zhou and Dillon (1991) [53]	[51]		X	X	X	X	X	X	X
Hwang and Rim (2002) [21]	[51]	X				X	X		X
Mojammado and Gharehpetian (2009)[30]	[51]	X	X						X
decision tree (C4.5): Quinlan (1993)	[22]		X	X			X	X	X
genetic algorithm: Goldberg (1989)	[22]		X			X	X		X
fast corr.-based filtering: Balakrishnan and Narayanaswamy (2009)	[2]		X				X	X	X
l_1-SVM-RBF$^{\text{tr}}$			X				X		X

Figure 4 illustrates the sparsity performance (AUC) trade-off made by means of the value of c in equation (6) for the first randomization between training and test set. Using three components is as performing as using 19 components. The same pattern is seen in the other randomizations. Figure 5 visualizes the decision boundary.

4.5 The Wisconsin Breast Cancer dataset (original)

This dataset contains information on 699 women of whom 683 had complete information on all 9 variables. The variables in this dataset are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image and are all integers ranging from 1 to 10. Due to their ordinal nature, all variables were considered to be continuous. We used 10 randomizations between training (two thirds of the data) and the test set (one third of the data). In 7 cases, uniform shape and bare nuclei were selected; one case selected uniform shape and chromatin; 1 case selected bare nuclei and chromatin; one case selected uniform size and bare nuclei. Based on these results, a final model containing uniform shape and bare nuclei was trained on the complete data set. The results are summarized in Table 6. The results are comparable but the presented method has the advantage of interpretability and sparsity in the number of components. The estimated effects are illustrated in Figure 6. An illustration of the decision surface (see Figure 7) shows that the estimated boundary separates both classes nearly perfectly.

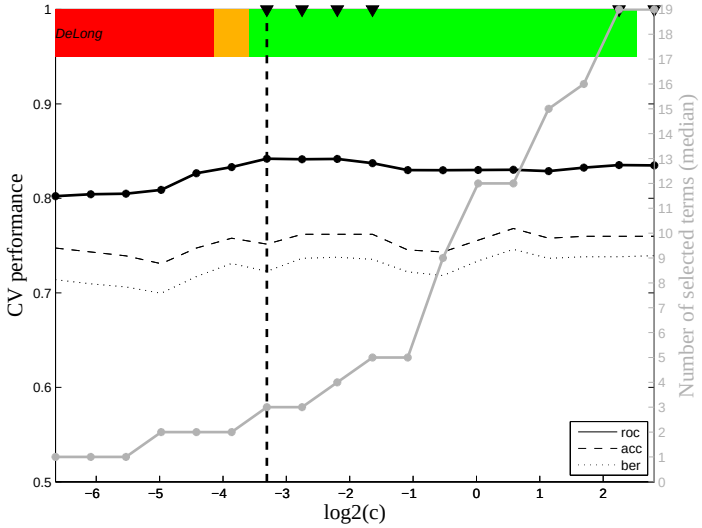


Fig. 4. Illustration of the sparsity performance trade-off by means of the value of c in equation (6). The upper bar indicates the p-value calculated by means of the method of DeLong. The model with the highest $c-$value (corresponding to the model with the most components) is the reference model. Every model with a smaller $c-$value is compared with this reference. A sparser model obtaining a higher AUC than the currect reference model becomes the new reference model (indicated by means of the triangles at the top). A green color indicates no significant differences between the AUCs of this and the reference model. An orange bar indicates a p-value between 0.01 and 0.05. A red color indicates a p-value less than 0.01. The automated procedure, selecting the sparsest model where the DeLong p-value is larger than 0.05, selects 3 input variables. Inspection of this Figure shows that selecting these variables results in a cross validation performance equal to the one obtained using 19 components.

Table 6. Comparison of the test set performance (mean and std) on the Wisconsin Breast Cancer dataset of the presented method with a standard SVM using an RBF kernel on the whole set of variables and the selected subset for 10 randomizations of training and test set. The results illustrate that the presented method is competitive with the standard SVM with the additional advantage of being interpretable.

method	AUC	ACC	BER
SVM-RBF (all)	0.996 (0.001)	0.968 (0.008)	0.037 (0.011)
SVM-RBF (subset)	0.990 (0.004)	0.954 (0014)	0.054 (0.018)
l_1-SVM-RBF$^{\text{tr}}$	0.990 (0.004)	0.956 (0.010)	0.051 (0.013)

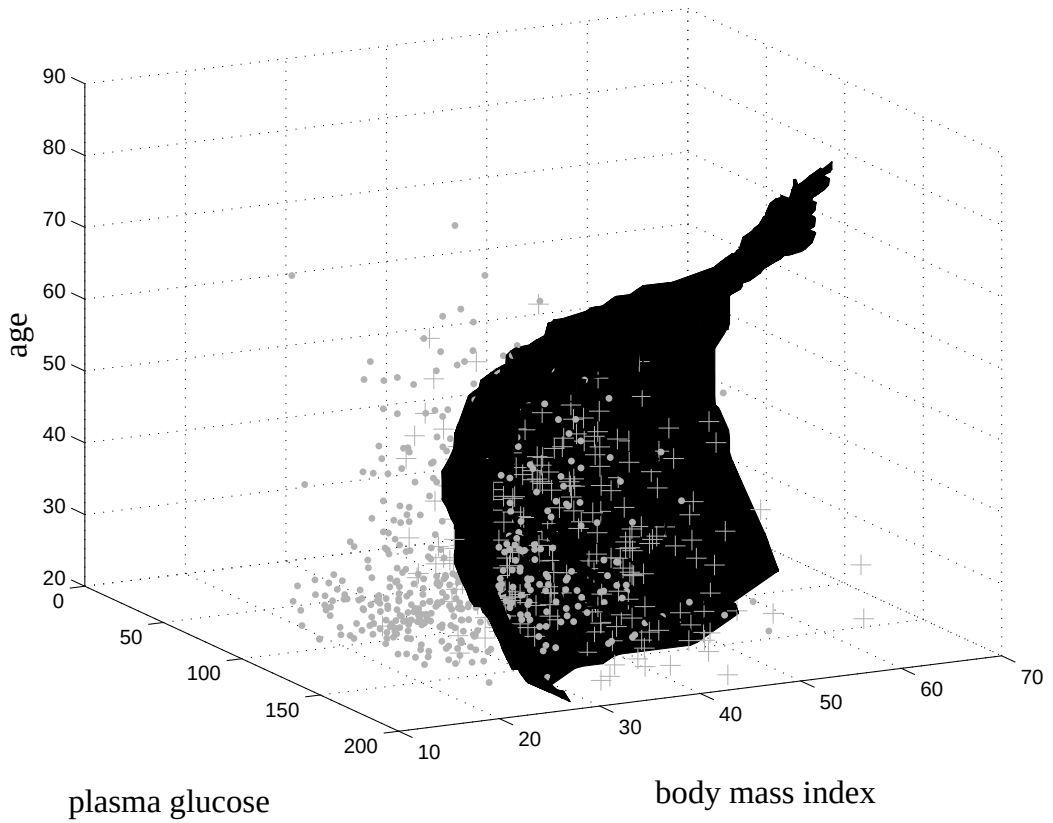


Fig. 5. Decision boundary for the Pima Indian Diabetes problem. Both classes are well separated by the hyperplane. The dots and pluses indicate the observations for healthy patients and patients with diabetes, respectively.

To validate the feature selection process, the results are compared with different results reported in the literature. Table 7 shows that the selected features are among those selected by other feature selection methods.

5 Conclusions

This work proposed a novel approach to enable the use of support vector machines with RBF kernels in domains where interpretability of the resulting classifiers is an issue. An expansion of the RBF kernel in components that are visualizable allows validation of the estimated effects of the input variables by experts in the domain of the application. It was shown how the extracted components could be shrunk to obtain a sparse model representation. Results on toy and artificial problems illustrate the ability of the model to select relevant main and two-way interaction effects. Comparison of the results on two benchmark datasets illustrates that the proposed method is competitive with other classifiers, but has the advantage of being interpretable.

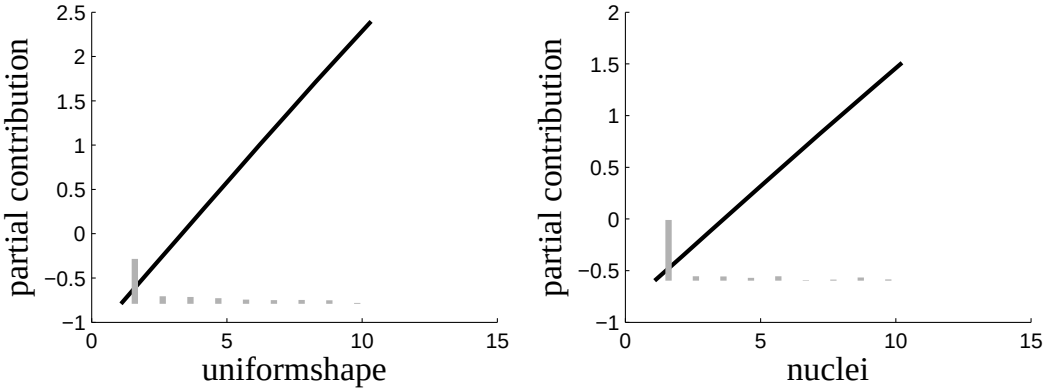


Fig. 6. Illustration of the selected features and their effects on the prediction of malignancy in the Wisconsin breast cancer dataset. The gray bars indicate the number of data points with the corresponding value of the input variable.

Table 7. Comparison of the feature selection results on the Wisconsin breast cancer dataset with results from the literature.

method	origin of results	clump thickness	uniformity of cell size	uniformity of cell shape	marginal adhesion	epithelial cell size	bare nuclei	bland chromatin	normal nucleoli	mitoses
OSRE [11]	[11]			X	X		X		X	
Neurorule (set 1) [36]	[11]	X	X		X		X		X	
Neurorule (set 2) [36]	[11]	X	X				X		X	
BIO-RE [42]	[11]	X		X				X	X	X
information geometric [10]	[10]	X		X	X		X			
NeuroLinear [37]	[10]			X	X		X		X	X
ensemble based [45]	[10]	X		X	X		X		X	
l_1-SVM-RBF$^{\text{tr}}$				X			X			

Acknowledgements

Research supported by Research Council KUL: GOA MaNet, PFV/10/002 (OPTEC), several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, G.0108.11 (Compressed Sensing), G.0869.12N (Tumor imaging) , IWT: TBM070706-IOTA3, PhD Grants; iMinds; Belgian Federal Science Policy Office: IUAP P7/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017); EU: RECAP 209G within INTERREG IVB NWE programme, EU HIP Trial FP7-HEALTH/ 2007-2013 (n. 260777), ERC AdG A-DATADRIVE-B. VVB is a postdoctoral fellow of the Research Foundation - Flanders (FWO).

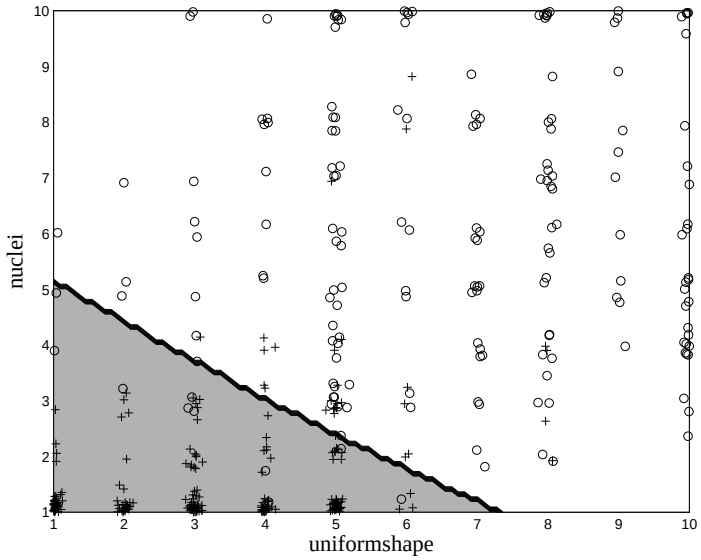


Fig. 7. Decision boundary for the Wisconsin breast cancer dataset. Both classes are well separated by the hyperplane. In order to improve the visualization, a random disturbance term is added to the variables, being integers in the dataset. The circles and pluses indicate the observations for benign and malignant tumors, respectively.

References

1. H. Almuallim and T. G. Dietterich. Learning with many irrelevant features. In *In Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 547–552. AAAI Press, 1991.
2. S. Balakrishnan and R. Narayanaswamy. Feature selection using fcbf in type ii diabetes databases. *International Journal of the Computer, the Internet and the Management*, 17(SP 1):50.1–50.8, 2009.
3. A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.
4. P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML, pages 82–90, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.
5. L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.
6. E. J. Candès, M. B. Wakin, and S. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.
7. A. Daemen and B. De Moor. Development of a kernel function for clinical data. In *Proceedings of the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 5913–5917. IEEE, Piscataway, 2009.
8. M. Dash, H. Liu, and J. Yao. Dimensionality reduction of unsupervised data. *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, 0:532–539, 1997.
9. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.
10. A. Eleuteri, R. Tagliaferri, and L. Milano. A novel information geometric approach to variable selection in mlp networks. *Neural Networks*, 18(10):1309–1318, 2005.
11. T. A. Etchells and P. J. G. Lisboa. Orthogonal search-based rule extraction (osre) for trained neural networks: a practical and efficient approach. *IEEE Transactions on Neural Networks*, 17(2):374–384, 2006.
12. A. Frank and A. Asuncion. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2010.
13. G. Fung and O. L. Mangasarian. A feature selection newton method for support vector machine classification. Technical Report 02-03, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, September 2002. `ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/02-01.ps`.

14. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.
15. M. Girolami and C. He. Probability density estimation from optimally condensed data samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1253–1264, 2003.
16. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, MGaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.
17. S. R. Gunn and J. S. Kandola. Structural modelling with sparse kernels. *Machine Learning*, 48:137–163, 2002.
18. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, Mar. 2003.
19. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
20. T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.
21. Y.-S. Hwang and H.-C. Rim. Decision tree decomposition-based complex feature selection for text chunking, 2002.
22. A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277, 2010.
23. K. Kira and L. A. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.
24. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324, 1997.
25. I. Kononenko. Estimating attributes: Analysis and extensions of relief. *Machine Learning ECML94*, 784:171–182, 1994.
26. Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297, 2006.
27. S. Maldonado, R. Weber, and J. Basak. Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, 181(1):115–128, 2011.
28. O. L. Mangasarian and G. Kou. Feature selection for nonlinear kernel support vector machines. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, ICDMW, pages 231–236, 2007.
29. J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–446, 1909.
30. M. Mohammadi and G. B. Gharehpetian. Application of core vector machines for on-line voltage security assessment using a decision-tree-based feature selection algorithm. *IET Generation Transmission Distribution*, 3(8):701, 2009.
31. J. Neumann, C. Schnörr, and G. Steidl. Combined svm-based feature selection and classification. *Machine Learning*, 61(1-3):129–150, Nov. 2005.
32. K. Pelckmans, I. Goethals, J. De Brabanter, J. A. K. Suykens, and B. De Moor. *Componentwise Least Squares Support Vector Machines*, chapter Support Vector Machines: Theory and Applications, pages 77–98. (Wang L., ed.), Springer, 2005.
33. J. C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74, 1999.
34. A. Rakotomamonjy. Variable selection using svm based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.
35. P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.
36. R. Setiono. Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 18(3):205–219, 2000.
37. R. Setiono and H. Liu. Neurolinear: From neural networks to oblique decision rules. *Neurocomputing*, 17(1):1–24, 1997.
38. M. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, and J. Weston. *Support vector regression with* ANOVA *decomposition kernels*, chapter Advances in kernel methods: support vector learning, pages 285–291. MIT Press, Cambridge, MA, USA, 1999.

39. Y. Sun. Iterative relief for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1035 –1051, june 2007.
40. J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.
41. J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.
42. I. Taha and J. Ghosh. Three techniques for extracting rules from feedforward networks. In *In Intelligent Engineering Systems Through Artificial Neural Networks ANNIE*, pages 23–28. ASME Press, 1996.
43. M. Tan, L. Wang, and I. W. I. Tsang. Learning sparse svm for feature selection on very high dimensional datasets. In *Proceedings of the 27th International Conference on Machine Learning*, ICML, 2010.
44. R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):267–288, 1997.
45. P. van de Laar. Input selection based on an ensemble. *Neurocomputing*, 34(1-4):227–238, 2000.
46. V. Vapnik. *Statistical Learning Theory*. Wiley and Sons, New York, 1998.
47. X. Wang and S. Wang. Feature ranking by weighting and ise criterion of nonparametric density estimation. *Journal of Applied Sciences*, 9(6):1014–1024, 2009.
48. J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, Mar. 2003.
49. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In *Advances in Neural Information Processing Systems 13*, pages 668–674. MIT Press, 2000.
50. S. Xavier de Souza, J. A. K. Suykens, J. Vandewalle, and D. Bolle. Coupled simulated annealing. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 40(2):320–335, 2010.
51. Y. M. Yacob, H. A. Mat Sakim, and N. A. Mat Isa. Decision tree-based feature ranking using manhattan hierarchical cluster criterion. *International Journal of Engineering and Physical Sciences*, 6, 2012.
52. M. Yuan and L. Lin. On the non-negative garrote estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:143–161, 2007.
53. X. J. Zhou and T. S. Dillon. A statistical-heuristic feature selection criterion for decision tree induction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):834–841, 1991.