Artificial Intelligence in Medicine 60 (2014) 53–64

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

White box radial basis function classifiers with component selection for clinical prediction models

Vanya Van Belle[a,b,*], Paulo Lisboa[b]

[a] Department of Electrical Engineering/iMinds Future Health Department, KU Leuven, Kasteelpark Arenberg 10/2446, 3001 Leuven, Belgium
[b] Department of Mathematics and Statistics, Liverpool John Moores University, Byrom Street, Liverpool L3 5UX, UK

ARTICLE INFO

Article history:
Received 28 November 2012
Received in revised form 4 October 2013
Accepted 8 October 2013

Keywords:
Interpretable support vector machines
Radial basis functions
White box methods
Feature selection
Clinical decision support

ABSTRACT

Objective: To propose a new flexible and sparse classifier that results in interpretable decision support systems.
Methods: Support vector machines (SVMs) 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: 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. For a 10-dimensional x-or problem, an SVM using the standard RBF kernel obtains an area under the receiver operating characteristic curve (AUC) of 0.947, whereas the proposed method achieves an AUC of 0.997. The latter additionally identifies the relevant components. In a second 10-dimensional artificial problem, the underlying class probability follows a logistic regression model. An SVM with the RBF kernel results in an AUC of 0.975, as apposed to 0.994 for the presented method. The proposed method is applied to two benchmark datasets: the Pima Indian diabetes and the Wisconsin Breast Cancer dataset. The AUC is in both cases comparable to those of the standard method (0.826 versus 0.826 and 0.990 versus 0.996) and those reported in the literature. The selected components are consistent with different approaches reported in other work. However, this method is able to visualize the effect of each of the components, allowing for interpretation of the learned logic by experts in the application domain.
Conclusions: This work proposes a new method to obtain flexible and sparse risk prediction models. The proposed method performs as well as a support vector machine using the standard RBF kernel, but has the additional advantage that the resulting model can be interpreted by experts in the application domain.

1. Introduction

Machine learning methods [1–3] 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

[*] Corresponding author at: Department of Electrical Engineering/iMinds Future Health Department, KU Leuven, Kasteelpark Arenberg 10/2446, 3001 Leuven, Belgium. Tel.: +32 16 32 10 65; fax: +32 16 32 19 86.
E-mail address: vanya.vanbelle@esat.kuleuven.be (V. Van Belle).

0933-3657/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2013.10.001

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 (SVMs) and in extension for least-squares support vector machines 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 [4,5]. More advanced approaches such as relief and focus have been proposed in [6–8]. Although filter approaches are very efficient w.r.t. computation, this approach might not be optimal [9,10]. 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 [11], 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 [12,13]. 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 [14–18]. 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 cannot 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 [19,20]. Maldonado et al. [21] 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.

Another approach that is often used to enable interpretation of SVMs are rule extraction methods [22,23]. However, the approach of these methods is quite different from the one presented in this manuscript. The learned rules give an explanation of the model but they are not equal to the model. The rules only mimic the original model and are thus an approximation of the learned logic of the SVM. Decision rules are a binary approximation to the smooth response function. Our method makes the response function explicit in its variable specific components and for pairwise interactions. Additionally, there is no mechanism controlling the difference in performance between the original model and the learned rules. The intention of this work is to provide flexible methods that are interpretable by design, and contain an explicit control mechanism on the performance.

In order to allow for an explanation of the model's prediction, models are often restricted to be additive [24,25]. Thanks to the additive structure, the contribution of each input variable to the prediction is clear. However, several classification problems cannot be solved using a sum of main effects. The use of ANOVA models [26], 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 [27,28]. 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 application domain. ANOVA models for component selection where proposed in [29–31]. The kernel approach taken by Gunn and Kandola [32] 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. In this work, we explicitly choose to restrict the components to main and two-way interaction effects to facilitate the visualization of the effect of the different components on the prediction. In most clinical research, main effects are considered and when assumed necessary, interactions are added [27,28].

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 2.4 indicates how the results can be interpreted in clinical practice. 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 [1] 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

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