Neural Networks 53 (2014) 69–80

Contents lists available at ScienceDirect

Neural Networks

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

Kernel learning at the first level of inference

Gavin C. Cawley *, Nicola L.C. Talbot

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

ARTICLE INFO

Article history:
Received 28 March 2013
Received in revised form 22 January 2014
Accepted 24 January 2014

Keywords:
Kernel methods
Model selection
Regularisation
Over-fitting
Automatic relevance determination

ABSTRACT

Kernel learning methods, whether Bayesian or frequentist, typically involve multiple levels of inference, with the coefficients of the kernel expansion being determined at the first level and the kernel and regularisation parameters carefully tuned at the second level, a process known as *model selection*. Model selection for kernel machines is commonly performed via optimisation of a suitable model selection criterion, often based on cross-validation or theoretical performance bounds. However, if there are a large number of kernel parameters, as for instance in the case of automatic relevance determination (ARD), there is a substantial risk of over-fitting the model selection criterion, resulting in poor generalisation performance. In this paper we investigate the possibility of learning the kernel, for the Least-Squares Support Vector Machine (LS-SVM) classifier, at the first level of inference, i.e. parameter optimisation. The kernel parameters and the coefficients of the kernel expansion are jointly optimised at the first level of inference, minimising a training criterion with an additional regularisation term acting on the kernel parameters. The key advantage of this approach is that the values of only two regularisation parameters need be determined in model selection, substantially alleviating the problem of over-fitting the model selection criterion. The benefits of this approach are demonstrated using a suite of synthetic and real-world binary classification benchmark problems, where kernel learning at the first level of inference is shown to be statistically superior to the conventional approach, improves on our previous work (Cawley and Talbot, 2007) and is competitive with Multiple Kernel Learning approaches, but with reduced computational expense.

1. Introduction

The training procedures for artificial neural networks (Bishop, 1995; MacKay, 1992), kernel learning methods (Schölkopf & Smola, 2002) and Gaussian process classifiers (MacKay, 1998; Rasmussen & Williams, 2006; Williams & Barber, 1998), can be viewed as multi-level optimisation problems (Guyon, Saffari, Dror, & Cawley, 2009). The model parameters are optimised at the first level of inference, for instance the weights of an artificial neural network, or the coefficients of the kernel expansion of a kernel machine. However, there are normally a number of *hyper-parameters* that must be determined, for example the number of hidden layer units in a multi-layer perceptron network, the choice of kernel and the values of any associated kernel parameters for a kernel machine, or regularisation parameters controlling the complexity of the model. These hyper-parameters are normally optimised at a second level of inference, a process known as *model selection* (Guyon, 2009). The division between parameters and hyper-parameters typically arises due to computational considerations. The dual parameters of a kernel machine, for example, are generally given by the solution of a convex optimisation problem, for which computationally efficient algorithms are available (Boyd & Vandenberghe, 2004). It is therefore computationally convenient to alternate between optimising the coefficients of the kernel expansion at the first level of inference and optimising the kernel and regularisation parameters at the second level of inference, taking advantage of the simple mathematical structure of the problem at the first level of inference.

In the case of kernel learning methods, the convex nature of the optimisation problem at the first level of inference implies a single, global optimum, thus avoiding the potential pitfall of multiple local minima that complicates the application of multi-layer perceptron networks. However, in order to maximise generalisation performance in practical applications, the values of a small number of regularisation and kernel parameters must also be carefully tuned during model selection (Chapelle, Vapnik, Bousquet, & Mukherjee, 2002). This is most often achieved via minimisation of

* Corresponding author. Tel.: +44 1603 593258.
E-mail addresses: G.Cawley@uea.ac.uk, gcc@cmp.uea.ac.uk (G.C. Cawley), nlct@cmp.uea.ac.uk (N.L.C. Talbot).

http://dx.doi.org/10.1016/j.neunet.2014.01.011
0893-6080/© 2014 Elsevier Ltd. All rights reserved.

a cross-validation estimate of generalisation performance, using grid search, Nelder–Mead simplex (Nelder & Mead, 1965) or gradient descent-based methods (Chapelle et al., 2002). This approach has been shown to be highly effective for kernel machines with a small number of hyper-parameters (e.g. Cawley, 2006). However, as the number of hyper-parameters becomes large, there is an increasing risk of over-fitting the model selection criterion, resulting in poor performance (Cawley & Talbot, 2007, 2010). Chapelle (2002) suggests that the additional estimation error might reasonably be expected to grow with the square root of the number of hyper-parameters. This danger has been observed previously (Bengio, 2000), and is especially evident in studies involving Automatic Relevance Determination (ARD), where the kernel includes separate scaling parameters for each feature. It is also well understood that the model selection criterion should not be also used for performance estimation as its direct optimisation during model selection will introduce an optimistic bias, and hence procedures such as nested cross-validation are necessary (Cawley & Talbot, 2010; Cherkassky & Mulier, 1998; Hastie, Tibshirani, & Friedman, 2001). While over-fitting of the model selection criterion is clearly a significant problem, research towards a potential solution appears to have received relatively little attention. Cawley and Talbot (2007) propose the addition of a regularisation term to the model selection criterion penalising large values of the kernel parameters, and thus promoting a relatively smooth model. Regularisation of the kernel parameters is shown to be effective in some cases, however the problem of over-fitting in model selection is far from solved. The use of automatic relevance determination has several distinct benefits, including (cf. Chapelle et al., 2002):

- The potential for improved generalisation performance—it is intuitively reasonable to expect that suppressing irrelevant attributes should result in improvements in accuracy.
- Explanation of the data—determination of which attributes have useful explanatory power, and which do not, is often a useful scientific finding.
- Reduced cost of data collection—if redundant attributes can be identified and eliminated, there is no need to determine the values of that attribute in operation. In some applications (such as medical diagnosis, where some screening tests are more expensive to conduct than others), the cost of evaluating the attributes may be an important practical consideration.

Thus, even if the use of automatic relevance determination does not give a performance advantage over the more basic RBF kernel, it is worth developing methods to avoid over-fitting in model selection so that the second and third benefits of ARD can be obtained more fully and reliably. In many applications, especially where data are in limited supply, a simple but incorrect model will out-perform a more correct, but more complex model because the parameters of the model can be estimated more reliably. A common example is the use of naive Bayes in text classification, where the assumption of independence is clearly not justified. If explaining the data is an important concern, the correct model should be used, and methods developed to allow the parameters to be estimated more accurately and reliably.

The approach presented in this paper seeks to minimise the risk of over-fitting in model selection by minimising the number of hyper-parameters to be optimised during model selection, hence minimising the degrees of freedom available to over-fit the model selection criterion. This is achieved by demoting the selection of kernel parameters from the second level of inference to the first, such that they are jointly optimised with the dual model parameters, minimising a single regularised training criterion. An additional regularisation term is used to penalise values of the kernel parameters likely to result in poor generalisation performance. As the values of only two regularisation parameters need then be determined in model selection, it is reasonable to expect the chance of over-fitting the model selection criterion to be substantially reduced, even when many kernel parameters are used. The optimisation of kernel parameters at the first level of inference is similar to the design of radial basis function networks via gradient descent methods (Webb & Shannon, 1998); however the addition of a regularisation term is required to maintain generalisation performance.

The remainder of this paper is structured as follows: Section 2 describes a training algorithm for kernel ridge regression with optimisation of the kernel parameters at the first level of inference. Results obtained on a suite of synthetic and real-world benchmark datasets is presented in Section 3. Section 4 provides discussion, including suggestions for further research and recommendations for practical applications. Finally, the work is summarised and conclusions drawn in Section 5.

2. Kernel learning at the first level of inference

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$, represent the training sample, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of explanatory variables describing the ith example, and $y_i \in \{-1, +1\}$, is the corresponding desired response indicating the class to which the example belongs. The Least-Squares Support Vector Machine (LS-SVM) classifier (Suykens, Van Gestel, De Brabanter, De Moor, & Vanderwalle, 2002) constructs a linear classifier, $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b$, in a feature space, $\mathcal{F}$, defined via a fixed transformation $\phi : \mathcal{X} \to \mathcal{F}$. However, rather than define the feature space directly, it is instead induced by a positive definite kernel function, $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, giving the inner product between points in the feature space, such that $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$. In this study, we adopt the simple Gaussian Radial Basis Function (RBF) kernel,

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \exp\left(-\theta_1 \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right), \tag{1}$$

where θ_1 is a kernel parameter controlling the sensitivity of the kernel, and the automatic relevance determination (ARD) or feature scaling variant of the RBF kernel (Chapelle et al., 2002),

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \exp\left(-\sum_{i=1}^{d} \theta_i [\boldsymbol{x}_i - \boldsymbol{x}_i']^2\right), \tag{2}$$

where θ_i are kernel parameters allowing the sensitivity of the kernel with respect to each of the explanatory variables to be tuned independently. Ideally, the kernel parameters associated with irrelevant features will adopt very small values, implementing a form of Automatic Relevance Determination (ARD) (MacKay, 1994; Neal, 1996). For fixed $\boldsymbol{\theta}$, the primal model parameters, $(\boldsymbol{w}, b)$, are given by the minimiser of a convex training criterion

$$\mathcal{L}(\boldsymbol{w}, b) = \sum_{i=1}^{\ell} c\left(y_i, f(\boldsymbol{x}; \boldsymbol{w}, b)\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2,$$

where $c(\cdot, \cdot)$ is a convex loss (in this case, the squared loss $c(y, f) = 0.5(y - f)^2$) representing the data misfit and λ is a regularisation parameter controlling the bias–variance trade-off (Geman, Bienenstock, & Doursat, 1992). It can be shown that the vector of model parameters, $\boldsymbol{w}$, can be expressed as an expansion over the training examples, such that

$$\boldsymbol{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i) \Longrightarrow f(\boldsymbol{x}; \boldsymbol{w}, b) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b,$$

where $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^{\ell}$ is a vector of *dual* model parameters. For a fixed value of the regularisation parameter, λ, the optimal dual model parameters are given by the solution of a system of linear equations,

$$\begin{bmatrix} \boldsymbol{K} + \lambda \boldsymbol{I} & \boldsymbol{1} \\ \boldsymbol{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} \tag{3}$$