Computational Statistics and Data Analysis 54 (2010) 1484–1504

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Optimized fixed-size kernel models for large data sets

K. De Brabanter [a,*], J. De Brabanter [a,b], J.A.K. Suykens [a], B. De Moor [a]

[a] Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[b] Hogeschool KaHo Sint-Lieven, (Associatie K.U. Leuven), Departement Industrieel Ingenieur, B-9000, Gent, Belgium

ARTICLE INFO

Article history:
Received 13 October 2009
Received in revised form 7 January 2010
Accepted 19 January 2010
Available online 28 January 2010

Keywords:
Kernel methods
Least squares support vector machines
Classification
Regression
Plug-in estimate
Entropy
Cross-validation

ABSTRACT

A modified active subset selection method based on quadratic Rényi entropy and a fast cross-validation for fixed-size least squares support vector machines is proposed for classification and regression with optimized tuning process. The kernel bandwidth of the entropy based selection criterion is optimally determined according to the solve-the-equation plug-in method. Also a fast cross-validation method based on a simple updating scheme is developed. The combination of these two techniques is suitable for handling large scale data sets on standard personal computers. Finally, the performance on test data and computational time of this fixed-size method are compared to those for standard support vector machines and ν-support vector machines resulting in sparser models with lower computational cost and comparable accuracy.

1. Introduction

Support vector machines (SVM) (Vapnik, 1995, 1999) and least squares support vector machines (LS-SVM) (Suykens and Vandewalle, 1999; Suykens et al., 2002) are state of the art learning algorithms for pattern recognition and function estimation. Typically a quadratic programming (QP) problem has to be solved in dual space in order to determine the SVM model. The formulation of the optimization problem in the primal space associated with this QP problem involves inequality constraints in the form of box constraints and an additional equality constraint.

Unfortunately, the designs of QP solvers, e.g. MINOS and LOQO, assume that the full kernel matrix is readily available. To overcome this difficulty, decomposition methods (Osuna et al., 1997a,b; Saunders et al., 1998; Joachims, 1999) were designed. A particular case of the decomposition method is iterative chunking where the full scale problem is restricted to a small subset of training examples called the working set. An extreme form of chunking is sequential minimal optimization (SMO) proposed by Platt (1999). SMO uses the smallest possible working set size, i.e. two elements. This choice greatly simplifies the method. Due to this, SMO is considered as the current state of the art QP solver for solving medium scale as well as large scale SVM.

In the LS-SVM formulation the inequality constraints are replaced by equality constraints and a sum of squared errors (SSE) cost function is used. Due to the use of equality constraints and the L_2 cost function in LS-SVM the solution is found by solving a linear system instead of quadratic programming. To tackle large scale problems with LS-SVM, Suykens et al. (1999) and Van Gestel et al. (2004) effectively employed the Hestenes–Stiefel conjugate gradient algorithm (Golub and Van Loan, 1989; Suykens et al., 1999). This method is well suited for problems with a larger number of data (up to about 10,000 data points). As an alternative, an iterative algorithm for solving large scale LS-SVM was proposed by Keerthi and Shevade (2003). This method is based on the solution of the dual problem using an idea similar to that of the SMO algorithm, i.e. using Wolfe duality theory, for SVM.

* Corresponding author. Tel.: +32 16 328658.
E-mail addresses: kris.debrabanter@esat.kuleuven.be (K. De Brabanter), jos.debrabanter@kahosl.be (J. De Brabanter), johan.suykens@esat.kuleuven.be (J.A.K. Suykens), bart.demoor@esat.kuleuven.be (B. De Moor).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.01.024

The vast majority of textbooks and articles discussing SVM and LS-SVM first state the primal optimization problem and then go directly to the dual formulation (Vapnik, 1995; Suykens and Vandewalle, 1999). A successful attempt at solving LS-SVM in primal weight space resulting in a parametric model and sparse representation, introduced by Suykens et al. (2002), is referred to as using *fixed-size least squares support vector machines* (FS-LSSVM) and was also applied in Espinoza et al. (2007). In this method an explicit expression for the feature map or an approximation to it is required. A procedure for finding this approximated feature map is based on the Nyström method (Nyström, 1930; Baker, 1977). Williams and Seeger (2001) used the Nyström method to speed up Gaussian processes (GP) (Williams and Barber, 1998). The Nyström method is related to finding a low rank approximation to the given kernel matrix by choosing m rows or columns of the kernel matrix. Many ways of selecting those m rows or columns of the kernel matrix can be found in the literature (Suykens et al., 2002; Achlioptas et al., 2002; Drineas and Mahoney, 2005). Smola and Schölkopf (2000) presented a sparse greedy approximation technique for constructing a compressed representation of the kernel matrix. This technique approximates the kernel matrix by the subspace spanned by a subset of its columns. The basis vectors are chosen incrementally to minimize an upper bound of the approximation error. A comparison of some of the above mentioned techniques can be found in Hoegaerts et al. (2004). Suykens et al. (2002) proposed searching for m rows or columns while maximizing the quadratic Rényi entropy criterion and estimate in the primal space leading to a sparse representation. This criterion will be used in the remainder of the paper.

The kernel representation of the quadratic Rényi entropy, established by Girolami (2002) and related to density estimation and principal component analysis, requires a bandwidth specific to the entropy criterion. Numerous bandwidth selection methods for density estimation exist, e.g. least squares cross-validation (LSCV) (Rudemo, 1982; Bowman, 1984), biased cross-validation (BCV) (Scott and Terrel, 1987), smoothed bootstrap (SB) (Taylor, 1989; Faraway and Jhun, 1990), plug-ins (Hall, unpublished manuscript; Sheather, 1986; Sheather and Jones, 1991), reference rules (Deheuvels, 1977; Silverman, 1986). In this paper we use the *solve-the-equation plug-in method* (Sheather and Jones, 1991) which is related to the plug-in family. The rationale for using this method is based on the fact that the calculation can be done efficiently using the improved fast Gauss transform (IFGT) (Yang et al., 2003) and hence it is computationally more efficient than LSCV, BCV and SB. Also it has better convergence rates than the above mentioned methods (Sheather, 2004).

Kernel based methods require the determination of tuning parameters including a regularization constant and kernel bandwidth. A widely used technique for estimating these parameters is cross-validation (CV) (Burman, 1989). A simple implementation of v-fold cross-validation trains a classifier/regression model for each split of the data and is thus computationally expensive when v is large, e.g. in leave-one-out (LOO) CV. An extensive literature exists on reducing the computational complexity of v-fold CV and LOO-CV; see e.g. (Vapnik and Chapelle, 2000; Wahba et al., 2000) for SVM, (Ying and Keong, 2004; An et al., 2007) for LS-SVM and (Cawley and Talbot, 2004) for sparse LS-SVM. Using the fact that the FS-LSSVM training problem has a closed form, we apply a simple updating scheme to develop a fast v-fold CV suitable for large data sets. For typical 10-fold CV, the proposed algorithm is 10 to 15 times faster than the simple implementation. Experiments also show that the complexity of the proposed algorithm is not very sensitive to the number of folds.

A typical method for estimating the tuning parameters would define a grid (grid-search) over these parameters of interest and perform v-fold CV for each of these grid values. However, three disadvantages come up with this approach (Bennett et al., 2006). A first disadvantage of such a grid-search CV approach is the limitation of the desirable number of tuning parameters in a model, due to the combinatorial explosion of grid points. A second disadvantage of this approach is their practical inefficiency; namely, they are incapable of assuring the overall quality of the solution produced. A third disadvantage in grid-search is that the discretization fails to take into account the fact that the tuning parameters are continuous. Therefore we propose an alternative for finding better tuning parameters. Our strategy is based on the recently developed coupled simulating annealing (CSA) method with variance control proposed by Xavier de Souza et al. (2006, in press). Global optimization methods are typically very slow. For many difficult problems, ensuring convergence to a global optimum might mean impractical running times. For such problems, a reasonable solution might be enough in exchange for a faster convergence. Precisely for this reason, many simulated annealing (SA) algorithms (Ingber, 1989; Rajasekaran, 2000) and other heuristic based techniques have been developed. However, due to speed-up procedures, these methods often get trapped in poor optima. The CSA method used in this paper is designed to easily escape from local optima and thus improves the quality of solution without compromising too much the speed of convergence. To better understand the underlying principles of these classes of methods consider the work of Suykens et al. (2001). One of the largest differences from SA is that CSA features a new form of acceptance probability functions that can be applied to an ensemble of optimizers. This approach considers several current states which are coupled together by their energies in their acceptance function. Also, in contrast with the case for classical SA techniques, parallelism is an inherent characteristic of this class of methods.

In this paper we propose a fast cross-validation technique suitable for large scale data sets. We modify and apply the solve-the-equation plug-in method for entropy bandwidth selection. Finally, we combine a fast global optimization technique with a simplex search in order to estimate the tuning parameters (regularization parameter and kernel bandwidth).

This paper is organized as follows. In Section 2 we give a short introduction concerning LS-SVM for classification and regression. In Section 3 we discuss the estimation in the primal weight space. Section 4 explains the active selection of a subsample based on the quadratic Rényi entropy together with a fast optimal bandwidth selection method using the *solve-the-equation plug-in* method. Section 5 describes a simple heuristic for determining the number of PV (prototype vectors). Section 6 discusses the proposed v-fold CV algorithm for FS-LSSVM. In Sections 7 and 8 the different algorithms are successfully demonstrated on real-life data sets. Section 9 states the conclusions of the paper. Finally, Appendix gives a detailed discussion of CSA.