Information Sciences 181 (2011) 3000–3016

Contents lists available at ScienceDirect

Information Sciences

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

Evolution strategies based adaptive L_p LS-SVM

Liwei Wei [a], Zhenyu Chen [b], Jianping Li [c,*]

[a] China National Institute of Standardization, Beijing 100088, China
[b] School of Business Administration, Northeastern University, Shenyang 110819, China
[c] Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100080, China

ARTICLE INFO

Article history:
Received 18 December 2007
Received in revised form 23 February 2011
Accepted 24 February 2011
Available online 16 March 2011

Keywords:
Adaptive penalty
Least squares support vector machine
Classification
Evolution strategies

ABSTRACT

Not only different databases but two classes of data within a database can also have different data structures. SVM and LS-SVM typically minimize the empirical ϕ-risk; regularized versions subject to fixed penalty (L_2 or L_1 penalty) are non-adaptive since their penalty forms are pre-determined. They often perform well only for certain types of situations. For example, LS-SVM with L_2 penalty is not preferred if the underlying model is sparse. This paper proposes an adaptive penalty learning procedure called evolution strategies (ES) based adaptive L_p least squares support vector machine (ES-based L_p LS-SVM) to address the above issue. By introducing multiple kernels, a L_p penalty based nonlinear objective function is derived. The iterative re-weighted minimal solver (IRMS) algorithm is used to solve the nonlinear function. Then evolution strategies (ES) is used to solve the multi-parameters optimization problem. Penalty parameter$p$, kernel and regularized parameters are adaptively selected by the proposed ES-based algorithm in the process of training the data, which makes it easier to achieve the optimal solution. Numerical experiments are conducted on two artificial data sets and six real world data sets. The experiment results show that the proposed procedure offer better generalization performance than the standard SVM, the LS-SVM and other improved algorithms.

1. Introduction

1.1. Progress of SVM and LS-SVM

Support vector machines (SVM), proposed by Vapnik [56,57], had been used in a range of problems including pattern classification, bioinformatics and text categorization due to their good generalization performance [8,11,30,35,55]. SVM mapped the input space into a higher dimensional feature space to make two classes of data linearly separable; the curse of dimensionality can be avoided with the use of kernels for nonlinear transformations. Based on the structured risk minimization principle, SVM, to simultaneously minimize empirical classification error and maximize the margin between two classes, tried to find a separating hyper-plane in the higher dimensional feature space. Therefore SVM optimized the generalization error and outperformed other traditional learning machines in many applications. It solved quadratic optimization (QP) problems in the dual space. The final classifier obtained by the SVM depended only on a small proportion of training samples (i.e. support vectors), which was good for implementation.

Despite the excellent properties of SVM, there are still some drawbacks, including selection of hyper-parameters [10,43] and the high computational cost of the QP problem [48]. Many modified versions of SVM had been proposed to deal with

* Corresponding author. Tel.: +86 10 5935 8805.
E-mail addresses: weilw@cnis.gov.cn (L. Wei), zychen@mail.neu.edu.cn (Z. Chen), ljp@casipm.ac.cn (J. Li).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2011.02.029

these problems [2,15,41,48]. Least squares support vector machine (LS-SVM) was one of the solutions, presented by Suykens [54]. In LS-SVM, a squared loss function was subjected to equality instead of inequality constraints so as to obtain a linear set of equations instead of a QP problem in the dual space [24]. Efficient iterative methods for solving large scale linear systems were available in numerical linear algebra [25]. A conjugate gradient algorithm based iterative process had been developed for solving the related Karush–Kuhn–Tucker system [54]. But there are some potential drawbacks of LS-SVM, one of which is that the use of a squared loss function might lead to loss of sparse solutions and less robust estimates with respect to outliers.

Recently, many improved models of LS-SVM have been proposed to make up for the above shortcomings. Suykens et al. [52,53] proposed a simple approach to address sparseness by sorting the support value spectrum (SVS), i.e. the absolute value of the solution of LS-SVM. Kruif and Vries [32] presented a more sophisticated pruning mechanism that omitted samples bearing the least errors. Another method [31] obtained sparse solutions by deleting some columns of the coefficient matrix through some defined measure. Li et al. [34] proposed another improved pruning method that selected training points that were nearer to the class boundary. Ref. [9] presented a new sparse hybrid classifier, using the reduced remaining subset (RRS) with LS-SVM. The new hybrid classifier was considered sparse because it was able to detect support vectors. However, these algorithms required solving a set of linear equations (slowly decreasing in size) many times, which incurred high computational costs.

1.2. Penalty function and the SVM or LS-SVM performance

However, SVM and LS-SVM with the L_2 penalty were not preferred if the underlying model was sparse [18]. One of the main advantages of L_1 was that it was less sensitive to outliers [28]. Therefore, many researchers chosen the L_1 penalty in their models to improve generalization performance, examples being the L_1 SVM [2,59] and LS-SVM-LP (L_1-LS-SVM) [58]. The L_1 SVM might have some advantages over the standard L_2 SVM, especially when there were redundant noise features [59]. Therefore, to improve the robustness of LS-SVM, they proposed a L_1 penalty based LS-SVM model (L_1-LS-SVM) [58]. L_1-LS-SVM can select features and do re-sampling. Compared with the standard LS-SVM, L_1-LS-SVM achieved better performance on data sets having redundant samples or outliers. But L_1 SVM and L_1-LS-SVM are not preferred if the data set have a non-sparse structure.

From the statistical viewpoint, [18] approached the problem of variable selection and featured extraction using a unified framework: penalized likelihood methods [18]. They addressed issues related to the choice of penalty functions for each field. These include areas like computational biology, health studies, financial engineering and risk management. Fan's work does not adaptively select the optimal penalty for every data set. [19] advocated penalty functions with three properties: sparseness, unbiasedness and continuity [19]. However, none of the L_p penalties ($p \in [0, +\infty]$) satisfies the above three conditions simultaneously. Friedman et al. [20] showed that SVM with L_1 penalty was preferred if the underlying model was sparse, while SVM with L_2 penalty performed better if most predictors contributed to the response. So it could be concluded that L_p penalty with $1 < p < 2$ satisfied sparseness condition less than L_1 penalty but more than L_2 penalty. Therefore, a learning machine with fixed penalty might fit best for some data structures [36]. We have introduced the adaptive penalty learning procedure which is data-driven for all data sets.

An adaptive L_q SVM was proposed by Liu et al. in which the Grid Search (GS) method was used to obtain the optimal L_q penalty ($q > 0$) [36]. The L_q SVM minimized the hinge loss function subject to L_q penalty and adaptively selects optimal data-driven penalty. Liu et al. used the local quadratic approximation (LQA) to directly solve the L_qSVM without needing dual Lagrangian transformation. This model adopted the linear kernel. Consequently, it was not suitable for the nonlinear separable problem. But it could give promising performance for the linear separable classification problem.

1.3. Our work

To solve the above issues, the adaptive L_p LS-SVM model is proposed to improve the performance of LS-SVM. Different from L_q SVM, the proposed adaptive L_p LS-SVM introduces multiple kernels and solves a linear equation set with deficient ranks in the dual space. In the processing signal field, a similar work, a generalized basis selection (GBS) framework, had been proposed, which unifies various methods for efficient representation of signals [4,7]. To obtain sparse solutions, GBS introduced a range of parameters p ($p \in [0,2]$), which could be solved by the affine scaling transformation (AST) method for $0 \leqslant p \leqslant 1$ and the iterative re-weighted minimal solver (IRMS) method for $1 \leqslant p \leqslant 2$. The GBS was only a combination of several signal representation methods; it did not select the appropriate p according to the data set. However, it proposed a promising IRMS algorithm to solve the nonlinear programming problem, which included the nonlinear objective function and a series of linear equation constraints.

This study aims to extend the L_1-LS-SVM model to solve the above mentioned problems. The adaptive L_pLS-SVM model introduces the convex combination of single feature basic kernels. This is equal to introducing the Lagrange parameter $\alpha'_{i,d}$ for the dth component of sample i in LS-SVM. A linear equation set with deficient ranks like the over-complete problem in Basis Pursuit (BP) has been derived. A L_p penalty based objective function is minimized and the optimal parameter$p$ is adaptively selected according to various data structures.

There are three parameters that need to be optimized in the learning procedure: penalty, kernel and the regularized parameter. Considering the computational complexity, GS is suitable for adjustment of only a few parameters. ES and