Neurocomputing 148 (2015) 498–511

Contents lists available at ScienceDirect

Neurocomputing

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

Efficient binary classification through energy minimisation of slack variables

Margarita Kotti [a,*], Konstantinos I. Diamantaras [b]

[a] Department of Surgery and Cancer, Department of Bioengineering, Faculty of Medicine, Imperial College London, Charing Cross Hospital, London W6 8RF, United Kingdom

[b] Information Technology Department, ATEI of Thessaloniki, Sindos 57400, Greece

ARTICLE INFO

Article history:
Received 2 February 2014
Received in revised form
23 April 2014
Accepted 6 July 2014
Communicated by Haowei Liu
Available online 21 July 2014

Keywords:
Slack minimisation
Binary classification
Kernel methods
Genetic optimisation

ABSTRACT

Slack variables are utilized in optimisation problems in order to build soft margin classifiers that allow for more flexibility during training. A robust binary classification algorithm that is based on the minimisation of the energy of slack variables, called the Mean Squared Slack (MSS), is proposed in this paper. Initially, the algorithm is analysed for the linear case, where the minimum mean squared slack is attained as a separating vector. Next, the kernel trick is exploited to facilitate computation of non-linear separating hyperplanes. For this paper, two kernels are tested, namely the radial basis function (RBF) and the polynomial kernel. In order to ensure a time and memory efficient system that converges in a few iterations four strategies are applied so as to withhold just a subset of feature vectors that are misclassified during training. Aiming to the automatic optimisation of the kernel parameters a modern combination of particle swarm optimisation (PSO) with artificial immune system (AIS) is tested. The aforementioned evolutionary methods are combined in a parallel architecture. Four datasets of diverse nature are exploited for performance evaluation, namely the iris, the SPECTheart, the vertebral column, and the wine quality datasets. Simulation experiments demonstrate high classification accuracy in a number of benchmark datasets.

1. Introduction

Binary classification is a central problem in machine learning. Boser et al. [1] were the first to use kernels to construct a non-linear estimation algorithm, which is the hard margin predecessor of Support Vector Machines (SVM) [2]. The substitution of kernels for dot products transforms a linear geometric algorithm into a non-linear one. This way, hyperplane classifiers evolved to SVMs [3]. SVMs are binary maximum margin classifiers that try to find the hyperplane which optimally separates the data. The feature vectors at the margin are called support vectors and they define the classifier's hyperplane. SVMs present the ability to generalize well even with a limited number of training data. In this paper we employ an SVM alternative which minimises the energy of the slack variables directly, even though that solution may not necessarily yield the maximum margin classifier.

With respect to applications, SVMs are commonly used for example in bio-informatics and natural language processing. This may be partially attributed to the fact that both fields deal with high-dimensional problems, such as micro array processing tasks, fault diagnosis, and categorisation. Additionally, SVMs have been employed in a variety of applications including speech and speaker recognition, emotion classification, e-learning, database marketing, intrusion detection, geo- and environmental sciences, finance time series forecasting, and high energy physics.

Many of the aforementioned fields exploit non-linear SVMs. Non-linear SVMs transform the feature space into a higher dimensional one using a set of non-linear basis functions. Hopefully, in the higher dimension feature space the training feature vectors may be separated linearly. An advantage of the SVM is that it is not necessary to explicitly implement this transformation. Instead a kernel representation can be used, where the solution is written as a weighted sum of the values of certain kernel function evaluated at the support vectors. Most recent methods exploit the idea of constructing kernel algorithms where the starting point is a linear criterion instead of a linear algorithm [3]. For example, a linear criterion may be that two samples have identical means or two random variables present zero covariance. Other alternatives aim to improved scalability by utilizing parallel SVM (PSVM) [4]. Parallel SVMs loads only essential data to each machine, which reduces memory use through performing a row-based, approximate matrix factorisation. Additional recent theoretical advances include the bounds generalisation based on Rademacher complexity theory for model selection and error estimation [5]. Furthermore, probably approximately correct (PAC) Bayesian theory is

* Corresponding author.
E-mail addresses: m.kotti@imperial.ac.uk, mkotti@it.teithe.gr (M. Kotti), kdiamant@it.teithe.gr (K.I. Diamantaras).

http://dx.doi.org/10.1016/j.neucom.2014.07.013
0925-2312/

utilized to compute a dimension-independent bound of the generalisation error [6].

This paper presents a binary classification algorithm which is based on the minimisation of the energy of the slack variables. A slack variable is defined as zero if the training feature vector is classified correctly and as a small positive value if the training feature vector is classified incorrectly. A maximum margin classifier, such as an SVM seeks to put a soft penalty on the sum of the slack variables, whereas in the approach presented in this paper we attack directly the slack variables of the misclassified patterns. Since many patterns may be classified incorrectly during training, four different strategies are presented in this paper in order to sustain just a subset of the aforementioned training feature vectors. This way time and memory efficiency is achieved. The first strategy retains a subset of the misclassified training feature vectors in a "first come-first kept" approach, the second one retains those patterns in a stochastic manner, the third aims to retain only the "worst" patterns, whereas the final one sustains those patterns whose slack variables attain values within a predefined range.

Next, the kernel trick is utilized in order to facilitate the computation of non-linear separating hyperplanes. Specifically, 2 types of kernels are tested for this paper: the radial basis function (RBF) kernel and the polynomial one. It is widely accepted that the parameters which are related to the aforementioned kernels – that is σ for the RBF kernel and power/offset for the polynomial kernel – may have crucial influence on the classification efficiency. Obviously, the optimal classification accuracy is obtained by optimal parameters setting. Aiming to reach the best performing parameters in an automatic manner, an evolutionary algorithm is employed. The aforementioned algorithm is the combination of particle swarm optimisation (PSO) [7] and artificial immune systems (AIS) [8]. The combination is achieved in a manner of a parallel network. Specifically, in every iteration a pool of u memory cells is constructed and their respective affinity (testing accuracy) is calculated. Best half $u/2$ of the memory cells are selected and given as input to PSO and AIS, independently. New memory cells are produced by each evolutionary algorithm and the best-performing of them are added to the memory cell pool. Also, at each iteration random memory cells are generated to ensure that the size of population is u. The evolutionary algorithm terminates after a user-defined number of iterations. The proposed approach is tested on four datasets of different nature to verify its robustness using a 5-fold cross-validation experimental protocol. High classification accuracy is achieved, 99.2% for the iris dataset, 80.140%, for the SPECTheart dataset, and 85.769% for the vertebral column dataset.

In summary, our contributions are as follows:

- A novel binary classification algorithm is presented which attacks the slack variables directly.
- It is demonstrated that in the linearly separable case the minimum mean squared slack is attained at a separating vector.
- It is proven that the minimiser in the linearly non-separable case is bounded, but not zero.
- The algorithm is an EM algorithm, so there is convergence, at least in the local sense. Additionally, the evolutionary nature of the parameter estimation algorithm facilitates the escape from local minima. Moreover, one of the proposed strategies is based on the stochastic selection of the subset which also aims at the same direction.
- The algorithm is time and memory efficient since it converges in just a few iterations.
- The algorithm is stable, regardless of the subset retainment strategy that is employed.
- A hybrid evolutionary system is tested, as a parallel network of PSO and AIS in order to select the kernel parameters in an automatic manner.
- The algorithm proves to handle efficiently datasets of highly diverse nature.

The rest of this paper is organized as follows. In Section 2 the proposed algorithm is analysed. In Section 4 we extend from the linear case to the kernel case and the subset retainment strategies are detailed. The evolutionary algorithm for the automatic selection of kernel parameters is detailed in Section 5. Experimental evaluation in demonstrated in Section 6, whereas the presented results are discussed in Section 7. Finally, conclusions are drawn in Section 8.

2. Mean squared slack minimisation

2.1. Problem formulation

Let us consider the classification task for a set of training data

$$\mathcal{X} = \{(\mathbf{x}(i),\ t(i))\ |i = 1, \ldots, N\} \tag{1}$$

where $\mathbf{x}(i) \in \mathbb{R}^n$ is a feature vector, $t(i) \in \{-1, 1\}$ is the class label of $\mathbf{x}(i)$, and N is the size of $\mathcal{X}$. The classification task is described as the search for a proper weight vector $\mathbf{w} \in \mathbb{R}^n$ and bias b that solve the set of inequalities:

$$t(i)y(i) \geq \gamma, \quad i = 1, \ldots, N \tag{2}$$

where

$$y(i) = \mathbf{w}^T\mathbf{x}(i) + b \tag{3}$$

is the output of the classifier for pattern i.

In general, there may not exist any feasible solution for (2). In this case, it is useful to define a slack variable $\xi(i)$, associated with pattern i,

$$\xi(i) = \max\{\gamma - t(i)y(i), 0\} \tag{4}$$

so that

$$\xi(i) = 0 \quad \text{iff } t(i)y(i) \geq \gamma, \tag{5}$$

$$\xi(i) > 0 \quad \text{iff } t(i)y(i) = \gamma - \xi(i) < \gamma. \tag{6}$$

Thus, the slack variable is positive only for the misclassified patterns, i.e. those with output y less than γ. Typically a maximum margin classifier, such as an SVM, seeks to minimise the norm of the weight vector $\mathbf{w}$ while putting a soft penalty on the sum of the slack variables. Nevertheless, the computational complexity of the resulting quadratic problem may be high.

An alternative approach would be to attack the slack variables directly [9]. Since for misclassified patterns ξ is positive, we can, in fact, define a whole family of cost functions of the form

$$J_p = \tfrac{1}{2}\overline{E}\{\xi^p | \xi > 0\} \tag{7}$$

where $\overline{E}\{X|Y\}$ is the empirical average of the sequence $X(i)$ under condition Y:

$$\overline{E}\{X|Y\} = \frac{1}{N_Y} \sum_{\substack{\text{all } i \text{ where} \\ Y \text{ is true}}} X(i) \tag{8}$$

and N_Y is the number of instances where Y is true. It is not difficult to see that for $p=1$ we obtain the Perceptron cost function J_1 [10, Chapter 5].

Defining

$$S = \{i : \xi(i) > 0\}, \tag{9}$$

to be the set of indexes of the patterns with positive margin we obtain

$$J_p = \frac{1}{2|S|}\sum_{i \in S} \xi(i)^p.$$