Neurocomputing 149 (2015) 683–689

Contents lists available at ScienceDirect

Neurocomputing

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

Letters

Learning a hyperplane classifier by minimizing an exact bound on the VC dimension[1]

Jayadeva

Department of Electrical Engineering, Indian Institute of Technology, Delhi, Hauz Khas, New Delhi 110016, India

ARTICLE INFO

Article history:
Received 25 January 2014
Received in revised form
14 July 2014
Accepted 29 July 2014
Communicated by R.W. Newcomb
Available online 12 August 2014

Keywords:
Machine learning
Support vector machines
VC dimension
Complexity
Generalization
Sparse

ABSTRACT

The VC dimension measures the complexity of a learning machine, and a low VC dimension leads to good generalization. While SVMs produce state-of-the-art learning performance, it is well known that the VC dimension of a SVM can be unbounded; despite good results in practice, there is no guarantee of good generalization. In this paper, we show how to learn a hyperplane classifier by minimizing an exact, or Θ bound on its VC dimension. The proposed approach, termed as the Minimal Complexity Machine (MCM), involves solving a simple linear programming problem. Experimental results show, that on a number of benchmark datasets, the proposed approach learns classifiers with error rates much less than conventional SVMs, while often using fewer support vectors. On many benchmark datasets, the number of support vectors is less than one-tenth the number used by SVMs, indicating that the MCM does indeed learn simpler representations.

1. Introduction

Support vector machines are amongst the most widely used machine learning techniques today. The classical SVM [1] has evolved into a multitude of diverse formulations with different properties. The most commonly used variants are the maximum margin L_1 norm SVM [1], and the least squares SVM (LSSVM) [2], both of which require the solution of a quadratic programming problem. In the last few years, SVMs have been applied to a number of applications to obtain cutting edge performance; novel uses have also been devised, where their utility has been amply demonstrated [3–24]. SVMs were motivated by the celebrated work of Vapnik and his colleagues on generalization, and the complexity of learning. It is well known that the capacity of a learning machine can be measured by its Vapnik–Chervonenkis (VC) dimension. The VC dimension can be used to estimate a probabilistic upper bound on the test set error of a classifier. A small VC dimension leads to good generalization and low error rates on test data.

In his widely read tutorial, Burges [25] states that SVMs can have a very large VC dimension, and that "*at present there exists no theory which shows that good generalization performance is guaranteed for SVMs*". This paper shows how to learn a classifier with large margin, by minimizing an exact (Θ) bound on the VC dimension. In other words, the proposed objective linearly bounds the VC dimension from both above and below. We show that this leads to a simple linear programming problem. This approach is generic, and it suggests numerous variants that can be derived from it – as has been done for SVMs. Experimental results provided in the sequel show that the proposed Minimal Complexity Machine outperforms conventional SVMs in terms of test set accuracy, while often using far fewer support vectors. That the approach minimizes the machine capacity may be gauged from the fact that on many datasets, the MCM yields better test set accuracy while using less than 1/10−th the number of support vectors obtained by SVMs.

The motivation for the MCM originates from some sterling work on generalization [26–29]. We restrict our attention in this paper to a given binary classification dataset for which a hyperplane classifier needs to be learnt. Consider such a binary classification problem with data points $x^i, i=1,2,\ldots,M$, and where samples of class $+1$ and -1 are associated with labels $y_i=1$ and $y_i=-1$, respectively. We assume that the dimension of the input samples is n, i.e. $x^i=(x_1^i, x_2^i, \ldots, x_n^i)^T$. For the set of all gap tolerant hyperplane classifiers with margin $d \geq d_{min}$, Vapnik [28] showed that the VC dimension γ is bounded by

$$\gamma \leq 1+\min\left(\frac{R^2}{d_{min}^2}, n\right) \tag{1}$$

where R denotes the radius of the smallest sphere enclosing all the training samples. Burges, in [25], stated that "*the above arguments strongly suggest that algorithms that minimize R^2/d^2 can be expected*

E-mail address: jayadeva@ee.iitd.ac.in
URLS: http://www.jayadeva.net, http://ee.iitd.ernet.in/people/jayadeva.html

[1] For commercial use of the MCM and its variants, please contact FITT, IIT Delhi.

http://dx.doi.org/10.1016/j.neucom.2014.07.062
0925-2312/

to give better generalization performance. Further evidence for this is found in the following theorem of (*Vapnik, 1998*), *which we quote without proof*". We follow this line of argument and show, through a constructive result, that this is indeed the case.

The remainder of this paper is organized as follows. Section 2 outlines the proposed optimization problem for a linear hyperplane classifier in the input space. Section 3 discusses the extension of the Minimum Complexity Machine to the kernel case. Section 4 is devoted to a discussion of results obtained on selected benchmark datasets. Section 5 contains concluding remarks. In Appendix A, we derive an exact bound for the VC dimension of a hyperplane classifier. Appendix B deals with the formulation of the hard margin MCM.

2. The Linear Minimal Complexity Machine

We first consider the case of a linearly separable dataset. By definition, there exists a hyperplane that can classify these points with zero error. Let the separating hyperplane be given by

$$u^T x + v = 0. \tag{2}$$

Let us denote

$$h = \frac{\max_{i=1,2,\ldots,M} \| u^T x^i + v \|}{\min_{i=1,2,\ldots,M} \| u^T x^i + v \|}. \tag{3}$$

In Appendix A, we show that h may also be written as

$$h = \frac{\max_{i=1,2,\ldots,M} y_i(u^T x^i + v)}{\min_{i=1,2,\ldots,M} y_i(u^T x^i + v)}, \tag{4}$$

and we show that there exist constants $\alpha, \beta > 0$, $\alpha, \beta \in \Re$ such that

$$\alpha h^2 \le \gamma \le \beta h^2, \tag{5}$$

or, in other words, h^2 constitutes a tight or exact (θ) bound on the VC dimension γ. An exact bound implies that h^2 and γ are close to each other.

Therefore, the machine capacity can be minimized by keeping h^2 as small as possible. Since the square function $(\cdot)^2$ is monotonically increasing, we can minimize h instead of h^2. We now formulate an optimization problem that tries to find the classifier with smallest machine capacity that classifies all training points of the linearly separable dataset correctly; this problem is given by

$$\underset{u,v}{\text{minimize}}\ h = \frac{\max_{i=1,2,\ldots,M} y_i(u^T x^i + v)}{\min_{i=1,2,\ldots,M} y_i(u^T x^i + v)} \tag{6}$$

Note that in deriving the exact bound in Appendix A, we assumed that the separating hyperplane $u^T x + v = 0$ correctly separates the linearly separable training points; consequently, no other constraints are present in the optimization problem (6).

In Appendix B, we show that the optimization problem (6) may be reduced to the problem

$$\min_{w,b,h} h \tag{7}$$

$$h \ge y_i \cdot [w^T x^i + b], \quad i = 1, 2, \ldots, M \tag{8}$$

$$y_i \cdot [w^T x^i + b] \ge 1, \quad i = 1, 2, \ldots, M, \tag{9}$$

where $w \in \Re^n$, and $b, h \in \Re$. We refer to the problem (7)–(9) as the hard margin Linear Minimum Complexity Machine (Linear MCM).

Note that the variable h in (7) and that in (5) refer to the same functional. By minimizing h in (7), we are minimizing an exact bound on γ, the VC dimension of the classifier. Once w and b have been determined by solving (7)–(9), the class of a test sample x may be determined from the sign of the discriminant function

$$f(x) = w^T x + b \tag{10}$$

In general, datasets will not be linearly separable. The soft margin equivalent of the MCM is obtained by introducing additional slack variables, and is given by

$$\min_{w,b,h} h + C \cdot \sum_{i=1}^{M} q_i \tag{11}$$

$$h \ge y_i \cdot [w^T x^i + b] + q_i, \quad i = 1, 2, \ldots, M \tag{12}$$

$$y_i \cdot [w^T x^i + b] + q_i \ge 1, \quad i = 1, 2, \ldots, M \tag{13}$$

$$q_i \ge 0, \quad i = 1, 2, \ldots, M. \tag{14}$$

Here, the choice of C allows a tradeoff between the complexity (machine capacity) of the classifier and the classification error.

Once w and b have been determined, the class of a test sample x may be determined as before by using the sign of $f(x)$ in (10). In the sequel, we show how to extend the idea to nonlinearly separable datasets.

3. The kernel MCM

We consider a map $\phi(x)$ that maps the input samples from $\Re^n$ to $\Re^l$, where $l > n$. The separating hyperplane in the image space is given by

$$u^T \phi(x) + v = 0. \tag{15}$$

Following (11)–(13), the corresponding optimization problem for the kernel MCM may be shown to be

$$\min_{w,b,h,q} h + C \cdot \sum_{i=1}^{M} q_i \tag{16}$$

$$h \ge y_i \cdot [w^T \phi(x^i) + b] + q_i, \quad i = 1, 2, \ldots, M \tag{17}$$

$$y_i \cdot [w^T \phi(x^i) + b] + q_i \ge 1, \quad i = 1, 2, \ldots, M \tag{18}$$

$$q_i \ge 0, i = 1, 2, \ldots, M. \tag{19}$$

The image vectors $\phi(x^i), i = 1, 2, \ldots, M$ form an overcomplete basis in the empirical feature space, in which w also lies. Hence, we can write

$$w = \sum_{j=1}^{M} \lambda_j \phi(x^j). \tag{20}$$

Therefore,

$$w^T \phi(x^i) + b = \sum_{j=1}^{M} \lambda_j \phi(x^j)^T \phi(x^i) + b = \sum_{j=1}^{M} \lambda_j K(x^i, x^j) + b, \tag{21}$$

where $K(p, q)$ denotes the Kernel function with input vectors p and q, and is defined as

$$K(p, q) = \phi(p)^T \phi(q). \tag{22}$$

Substituting (21) into (16)–(18), we obtain the following optimization problem:

$$\min_{w,b,h,q} h + C \cdot \sum_{i=1}^{M} q_i \tag{23}$$

$$h \ge y_i \cdot \left[\sum_{j=1}^{M} \lambda_j K(x^i, x^j) + b \right] + q_i, \quad i = 1, 2, \ldots, M \tag{24}$$

$$y_i \cdot \left[\sum_{j=1}^{M} \lambda_j K(x^i, x^j) + b \right] + q_i \ge 1, \quad i = 1, 2, \ldots, M \tag{25}$$

$$q_i \ge 0, \quad i = 1, 2, \ldots, M. \tag{26}$$