

Letters

Learning a hyperplane classifier by minimizing an exact bound on the VC dimension¹



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.

© 2014 Elsevier B.V. All rights reserved.

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 $x^i, i = 1, 2, \dots, M$, 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. $x^i = (x_1^i, x_2^i, \dots, x_n^i)^T$. 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

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

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.inURLs: <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.

to give better generalization performance. Further evidence for this is found in the following theorem of (Vapnik, 1998), which we quote without proof⁶. 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

$$u^T x + v = 0. \quad (2)$$

Let us denote

$$h = \frac{\max_{i=1,2,\dots,M} \|u^T x^i + v\|}{\min_{i=1,2,\dots,M} \|u^T x^i + v\|}. \quad (3)$$

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

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

and we show that there exist constants $\alpha, \beta > 0$, $\alpha, \beta \in \mathfrak{R}$ such that

$$\alpha h^2 \leq \gamma \leq \beta h^2, \quad (5)$$

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

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

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

$$\min_{w,b,h} h \quad (7)$$

$$h \geq y_i \cdot [w^T x^i + b], \quad i = 1, 2, \dots, M \quad (8)$$

$$y_i \cdot [w^T x^i + b] \geq 1, \quad i = 1, 2, \dots, M, \quad (9)$$

where $w \in \mathfrak{R}^n$, and $b, h \in \mathfrak{R}$. 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

$$f(x) = w^T x + b \quad (10)$$

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

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

$$h \geq y_i \cdot [w^T x^i + b] + q_i, \quad i = 1, 2, \dots, M \quad (12)$$

$$y_i \cdot [w^T x^i + b] + q_i \geq 1, \quad i = 1, 2, \dots, M \quad (13)$$

$$q_i \geq 0, \quad i = 1, 2, \dots, M. \quad (14)$$

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 \mathfrak{R}^n to \mathfrak{R}^l , where $l > n$. The separating hyperplane in the image space is given by

$$u^T \phi(x) + v = 0. \quad (15)$$

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

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

$$h \geq y_i \cdot [w^T \phi(x^i) + b] + q_i, \quad i = 1, 2, \dots, M \quad (17)$$

$$y_i \cdot [w^T \phi(x^i) + b] + q_i \geq 1, \quad i = 1, 2, \dots, M \quad (18)$$

$$q_i \geq 0, i = 1, 2, \dots, M. \quad (19)$$

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

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

Therefore,

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

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

$$K(p, q) = \phi(p)^T \phi(q). \quad (22)$$

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

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

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

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

$$q_i \geq 0, \quad i = 1, 2, \dots, M. \quad (26)$$

Once the variables $\lambda_j, j = 1, 2, \dots, M$, and b are obtained, the class of a test point x can be determined by evaluating the sign of

$$f(x) = w^T \phi(x) + b = \sum_{j=1}^M \lambda_j K(x, x^j) + b. \quad (27)$$

Note that in (20), the $\phi(x^j)$'s for which the corresponding λ_j 's are non-zero constitute the support of the vector w . Hence, $\phi(x^j)$'s for which the λ_j 's are non-zero may be termed as support vectors. The maximum number of support vectors is the number of training samples M , as in the case of the SVM.

4. Experimental results

The MCM was coded in MATLAB [30]. Fig. 1 provides a flowchart illustrating the MCM implementation; as may be seen, we use the *linprog* function in MATLAB to solve the optimization problem comprising the MCM.

In order to compare the MCM with SVMs, we chose a number of benchmark datasets from the UCI machine learning repository [31]. Table 1 summarizes information about the number of samples and features of each dataset. Some of the benchmark datasets are multi-class ones. These have been learnt by adopting a one-versus-rest approach. Accuracies, CPU times, and the number of support vectors have been averaged across all the classifiers learnt in a one-versus-rest approach.

Table 2 summarizes five fold cross validation results of the linear MCM on the datasets listed in Table 1. Accuracies are indicated as mean \pm standard deviation, computed over the five folds. The table compares the linear MCM with the SVM using a linear kernel. Note that the MCM solves the primal problem and not the dual. The linear SVM is solved using the dual, by using a linear kernel. This is done because the SVM primal is numerically

Table 1

Characteristics of the benchmark datasets used.

Dataset	Size (samples \times features \times classes)
Blogger	100 \times 6 \times 2
Fertility diagnosis	100 \times 9 \times 2
Promoters	106 \times 57 \times 2
Echocardiogram	132 \times 12 \times 2
Teaching assistant	151 \times 5 \times 3
Hepatitis	155 \times 19 \times 2
Hayes	160 \times 5 \times 3
Plrx	182 \times 12 \times 2
Seed	210 \times 7 \times 3
Glass	214 \times 10 \times 6
Heartstatlog	270 \times 13 \times 2
Horsecolic	300 \times 27 \times 2
Haberman	306 \times 3 \times 2
Ecoli	336 \times 8 \times 3
House voters	435 \times 16 \times 2
Wholesale customer	440 \times 8 \times 2
IPLD	583 \times 10 \times 2
Balance	625 \times 4 \times 3
Australian	690 \times 14 \times 2
Crx	690 \times 15 \times 2
Transfusion	748 \times 5 \times 2
Tic tac toe	958 \times 9 \times 2
Sorlie	85 \times 456 \times 2
Secom	1567 \times 591 \times 2
Tian	173 \times 12626 \times 2

less attractive owing to the ill conditioning of the Hessian. Hence, there is a subtle difference between a linear MCM and a SVM (linear kernel). The values of C were determined for the MCM by performing a grid search; tuning C carefully is important for the MCM. The table also provides a comparison between the CPU times taken by the SVM and the MCM. For the implementation of the SVM, we use a small MATLAB code that solves the quadratic programming dual form of the classical L_1 norm SVM by calling the *quadprog* command in MATLAB. This has been done to enable a comparison of the CPU times taken by the MCM and the SVM. Faster implementations based on the active set methods are possible for the MCM; such methods exist for the SVM and include Platt's SMO [32] and the 1SMO algorithm [33].

Table 3 summarizes five fold cross validation results of the kernel MCM on a number of datasets. A Gaussian or Radial Basis Function (RBF) kernel was chosen for both the MCM and the SVM. The width of the MCM Gaussian kernel and the value of C were chosen by using a grid search.

The table shows test set accuracies and the number of support vectors for both the kernel MCM, and the classical SVM with a Gaussian kernel. The results indicate that the kernel MCM yields better generalization than the SVM. In the case of kernel classification, the MCM uses fewer support vectors, generally about one-third the number used by SVMs. In the case of many of the datasets, the MCM uses less than one-tenth the number of support vectors required by a SVM. The code for the MCM classifier would be available from the author's website. The large difference with the SVM results indicates that despite good performance, SVM solutions may still be far from optimal. Vapnik [1] showed that

$$E(P_{\text{error}}) \leq \frac{E(\# \text{support vectors})}{\# \text{training samples}}, \quad (28)$$

where $E(P_{\text{error}})$ denotes the expected error on test samples taken from the general distribution, $\# \text{training samples}$ denotes the number of training samples, and $E(\# \text{support vectors})$ denotes the expected number of support vectors obtained on training sets of the same size. Although the bound was shown for linearly separable datasets, it does indicate that the number of support vectors is also related to the prediction error. An examination of the table indicates that the

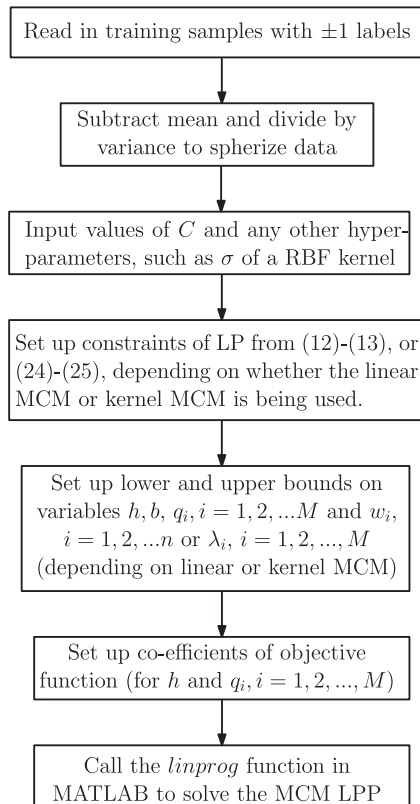


Fig. 1. Illustration of the MATLAB flowchart for the MCM code.

Table 2
Linear MCM results.

Datasets	Linear MCM		Linear SVM	
	Accuracy	Time (s)	Accuracy	Time (s)
Blogger	69.00 ± 17.15	0.0012 ± 6.64e−5	58.00 ± 20.40	6.17 ± 2.51
Fertility diagnosis	88.00 ± 9.27	0.0013 ± 5.27e−5	86.00 ± 9.01	8.12 ± 1.47
Promoters	74.08 ± 10.88	0.0014 ± 4.53e−5	67.78 ± 10.97	0.85 ± 0.03
Echocardiogram	90.88 ± 5.75	0.0014 ± 4.62e−5	86.38 ± 4.50	0.72 ± 0.36
Teaching assistant	66.27 ± 6.77	0.0013 ± 3.46e−5	64.94 ± 6.56	16.07 ± 4.17
Hepatitis	68.38 ± 6.26	0.0014 ± 3.82e−5	60.64 ± 7.19	1.90 ± 0.56
Hayes	76.32 ± 9.25	0.0012 ± 2.73e−5	73.56 ± 7.73	7.19 ± 3.81
Plrx	71.83 ± 7.49	0.0015 ± 3.37e−5	71.42 ± 7.37	4.35 ± 0.78
Seed	97.61 ± 1.51	0.0015 ± 3.97e−5	90.95 ± 4.09	12.37 ± 4.51
Glass	99.06 ± 1.16	0.0042 ± 5.56e−3	98.12 ± 1.75	11.83 ± 3.44
Heartstatlog	84.81 ± 3.87	0.0018 ± 1.47e−5	82.59 ± 2.22	9.43 ± 4.25
Horsecolic	81.00 ± 4.03	0.0021 ± 7.17e−5	80.26 ± 4.63	41.39 ± 13.93
Haberman	73.89 ± 3.71	0.0019 ± 4.34e−5	72.56 ± 3.73	13.74 ± 6.63
Ecoli	96.73 ± 1.96	0.0023 ± 1.3e−4	96.73 ± 1.96	18.41 ± 2.57
House voters	95.63 ± 1.84	0.0031 ± 1.87e−4	94.48 ± 2.46	15.77 ± 2.19
Wholesale customer	92.26 ± 1.97	0.0033 ± 1.07e4	91.13 ± 1.95	32.11 ± 8.29
IPLD	71.35 ± 2.93	0.0065 ± 4e−5	71.35 ± 2.93	12.30 ± 8.26
Balance	95.26 ± 1.02	0.0077 ± 1.3e−3	95.20 ± 1.01	8.37 ± 1.03
Australian	85.73 ± 2.04	0.0076 ± 9.75e−5	84.49 ± 1.18	407.97 ± 167.73
Crx	69.56 ± 2.79	0.0095 ± 1.36e−3	67.79 ± 3.47	498.04 ± 35.22
Transfusion	78.19 ± 3.25	0.0082 ± 8.21e−4	77.13 ± 2.26	173.06 ± 44.12
Tic tac toe	74.22 ± 5.50	0.038 ± 4.9e−2	73.91 ± 6.11	24.13 ± 6.81
Sortie ^a	94.084 ± 1.54	0.165 ± 0.15	90.19 ± 2.47	187.50 ± 1.37
Secom ^a	87.87 ± 1.88	957.00 ± 87.29	86.04 ± 0.82	6359.78 ± 15.93
Tian ^a	81.71 ± 1.43	1.39 ± 0.67	80.92 ± 1.39	7832.76 ± 6.31

^a Marked datasets were run on shared computing platforms, and compute times may not be representative.

Table 3
Kernel MCM results.

Datasets	Kernel MCM			Kernel SVM		
	Accuracy	CPU time (s)	#SV	Accuracy	CPU time (s)	#SV
Blogger	88.00 ± 4.00	0.32 ± 0.03	22.20 ± 5.91	81.00 ± 10.20	2573 ± 49.2	51.20 ± 3.06
Fertility diagnosis	89.00 ± 2.00	0.18 ± 0.09	9.80 ± 19.60	88.00 ± 9.27	8.03 ± 1.95	38.20 ± 1.60
Promoters	84.93 ± 1.56	0.45 ± 0.39	82.40 ± 2.73	75.59 ± 7.63	4.40 ± 1.33	83.80 ± 0.98
Echocardiogram	89.34 ± 4.57	0.31 ± 0.01	12.00 ± 0.00	87.14 ± 7.27	8.58 ± 1.91	48.00 ± 2.10
Teaching assistant	74.83 ± 2.60	0.39 ± 0.13	26.60 ± 32.43	68.88 ± 6.48	4192 ± 162	86.00 ± 3.22
Hepatitis	85.80 ± 8.31	0.44 ± 0.02	20.00 ± 0.00	82.57 ± 6.32	3561 ± 4392	72.20 ± 4.31
Hayes	81.82 ± 7.28	0.31 ± 0.05	3.23 ± 1.11	79.57 ± 6.60	1427 ± 54.7	84.20 ± 2.04
Plrx	71.99 ± 5.81	0.41 ± 0.10	4.40 ± 8.80	71.41 ± 6.04	144.21 ± 5816	116.2 ± 6.14
Seed	97.13 ± 0.95	0.79 ± 0.01	11.20 ± 5.71	91.90 ± 2.86	3362 ± 85.1	51.80 ± 1.72
Glass	96.23 ± 2.77	1.69 ± 0.50	36.00 ± 11.49	90.64 ± 5.09	20 475 ± 832	64.80 ± 2.40
Heartstatlog	84.44 ± 3.21	1.32 ± 0.76	10 ± 2.23	83.7 ± 1.54	1547 ± 324.52	124.6 ± 4.15
Horsecolic	82.33 ± 4.03	3.84 ± 2.31	36.60 ± 17.70	81.33 ± 4.14	13 267 ± 2646	187.2 ± 3.27
Haberman	73.49 ± 3.85	1.23 ± 0.32	8.50 ± 7.00	72.81 ± 3.51	2087 ± 750	138.2 ± 3.27
Ecoli	97.32 ± 1.73	3.47 ± 0.30	24.00 ± 1.41	96.42 ± 2.92	11 829 ± 248	57.00 ± 4.65
House voters	95.87 ± 1.16	4.24 ± 0.83	17.80 ± 8.91	95.42 ± 2.04	8827 ± 349	93.60 ± 3.93
Wholesale customer	92.72 ± 1.54	7.31 ± 0.93	39.00 ± 10.64	90.90 ± 1.90	9243 ± 362	123.40 ± 2.15
IPLD	72.03 ± 3.20	4.06 ± 5.02	23.40 ± 30.50	70.15 ± 2.24	9743 ± 322	311.60 ± 5.31
Balance	97.64 ± 1.32	8.78 ± 1.32	14.60 ± 0.49	97.60 ± 0.51	15 442 ± 651	143.00 ± 4.23
Australian	85.65 ± 2.77	103.45 ± 18.04	108.80 ± 1.60	84.31 ± 3.01	94 207 ± 4476	244.8 ± 4.64
Crx	69.56 ± 2.90	5.95 ± 2.55	3.40 ± 6.80	69.27 ± 2.62	19 327 ± 5841	404.4 ± 8.69
Transfusion	77.00 ± 2.84	7.08 ± 0.69	6.00 ± 3.52	76.73 ± 2.88	18 254 ± 1531	302.20 ± 7.55
Tic tac toe	98.32 ± 0.89	12.55 ± 0.56	10.00 ± 0.00	93.94 ± 2.10	18 674 ± 973	482.60 ± 3.93
Sortie ^a	98.82 ± 2.35	0.44 ± 0.15	50 ± 4.77	97.644 ± 2.88	78.63 ± 9.81	68.95 ± 3.72
Secom ^a	94.11 ± 2.23	1521 ± 75.5	382.8 ± 44.23	92.29 ± 0.82	38 769.25 ± 8.87	593.2 ± 17.22
Tian ^a	97.09 ± 3.83	2.05 ± 0.199	70.4 ± 3.26	95.188 ± 4.26	88.97 ± 3.26	75.6 ± 1.01

^a Marked datasets were run on shared computing platforms, and compute times may not be representative.

proposed approach shows a lower test set error, and also uses a smaller number of support vectors.

Table 4 shows the values of h for the linear MCM and the kernel MCM on the benchmark datasets. The values have been indicated as mean ± standard deviation, computed over the five folds in a five-

fold cross validation setting. Note that h^2 is an exact bound on the VC dimension γ , and thus an approximate measure of the capacity of the learning machine. The table indicates that the values of h for the kernel MCM are generally smaller than that for the linear MCM. This also reflects the better generalization achieved by the kernel MCM.

Table 4Values of h for the linear and the kernel MCM.

Datasets	Kernel MCM h	Linear MCM h
Blogger	3.73 ± 1.90	2.53 ± 0.78
Fertility diagnosis	1.00 ± 0.00	2.72 ± 1.56
Promoters	1.00 ± 0.00	35.77 ± 18.58
Echocardiogram	8.00 ± 2.94	35.76 ± 18.58
Seed	8.32 ± 4.50	9.46 ± 3.28
Hepatitis	3.02 ± 1.49	2.02 ± 1.32
Teaching assistant	1.38 ± 0.47	2.12 ± 0.94
Plrx	1.00 ± 0.00	1.00 ± 0.00
Hayes	3.23 ± 1.11	5.00 ± 0.00
Glass	8.80 ± 7.96	13.64 ± 3.32
Heartstatlog	2.12 ± 1.87	3.79 ± 1.08
Horsecolic	1.20 ± 0.40	1.00 ± 0.00
Haberman	1.18 ± 0.24	1.06 ± 0.07
Ecoli	2.41 ± 1.02	4.85 ± 1.93
House voters	4.24 ± 0.83	4.41 ± 4.21
Wholesale customer	2.36 ± 0.97	19.79 ± 6.55
IPLD	6.46 ± 3.22	1.00 ± 0.00
Balance	8.73 ± 7.21	16.60 ± 0.80
Australian	1.00 ± 0.00	1.86 ± 0.43
Crx	1.00 ± 0.00	1.00 ± 0.00
Transfusion	2.07 ± 1.36	1.57 ± 0.71
Tic tac toe	1.52 ± 0.06	3.07 ± 1.04
Sorlie	1.00 ± 0.00	1.12 ± 0.17
Secom	8.32 ± 0.71	1.00 ± 0.00
Tian	1.25 ± 0.27	5.31 ± 2.26

5. Conclusion

In this paper, we propose a way to build a hyperplane classifier, termed as the Minimal Complexity Machine (MCM), that attempts to minimize a bound on the VC dimension. The classifier can be found by solving a linear programming problem. Experimental results show that the learnt classifier outperforms the classical SVM in terms of generalization accuracies on a number of selected benchmark datasets. At the same time, the number of support vectors is less, often by a substantial factor. It has not escaped our attention that the proposed approach can be extended to least squares classifiers, as well as to tasks such as regression; in fact, a large number of variants of SVMs can be re-examined with the objective of minimizing the VC dimension.

Acknowledgment

The author would like to thank Prof. Suresh Chandra of the Department of Mathematics, IIT Delhi, for his valuable comments and a critical appraisal of the manuscript. The extensive simulations were the result of the untiring efforts of Siddharth Sabharwal and Sanjit Singh Batra. Early simulations on the linear model were done by Prasoon Goel.

Appendix A. An exact bound on the VC dimension γ

We derive an exact or tight (Θ) bound on the VC dimension γ . Vapnik [28] showed that the VC dimension γ for fat margin hyperplane classifiers with margin $d \geq d_{\min}$ satisfies

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

where R denotes the radius of the smallest sphere enclosing all the training samples. We first consider the case of a linearly separable dataset. By definition, there exists a hyperplane $u^T x + v = 0$ with positive margin d that can classify these points with zero error. We

can always choose $d_{\min} = d$; for all further discussion we assume that this is the case. Without loss of generality, we consider hyperplanes passing through the origin. To see that this is possible, we augment the co-ordinates of all samples with an additional dimension or feature whose value is always 1, i.e. the samples are given by $\hat{x}^i \leftarrow \{x^i; 1\}$, $i = 1, 2, \dots, M$; also, we assume that the weight vector is $(n+1)$ -dimensional, i.e. $\hat{u} \leftarrow \{u; v\}$.

Then, the margin, which is the distance of the closest point from the hyperplane, is given by

$$d = \min_{i=1,2,\dots,M} \frac{\|\hat{u}^T \hat{x}^i\|}{\|\hat{u}\|} \quad (\text{A.2})$$

$$\frac{R}{d} = \frac{\max_{i=1,2,\dots,M} \|\hat{x}^i\|}{\min_{i=1,2,\dots,M} \frac{\|\hat{u}^T \hat{x}^i\|}{\|\hat{u}\|}} = \frac{\max_{i=1,2,\dots,M} \|\hat{u}\| \|\hat{x}^i\|}{\min_{i=1,2,\dots,M} \|\hat{u}^T \hat{x}^i\|} \quad (\text{A.3})$$

From the Cauchy–Schwarz inequality, we have

$$\|\hat{u}^T \hat{x}^i\| \leq \|\hat{u}\| \|\hat{x}^i\| \quad (\text{A.4})$$

$$\Rightarrow \max_{i=1,2,\dots,M} \|\hat{u}^T \hat{x}^i\| \leq \max_{i=1,2,\dots,M} \|\hat{u}\| \|\hat{x}^i\| \quad (\text{A.5})$$

Therefore, from (A.3), we have

$$\frac{\max_{i=1,2,\dots,M} \|\hat{u}^T \hat{x}^i\|}{\min_{i=1,2,\dots,M} \|\hat{u}^T \hat{x}^i\|} \leq \frac{R}{d} \quad (\text{A.6})$$

or, in terms of the original variables u and v ,

$$\frac{\max_{i=1,2,\dots,M} \|u^T x^i + v\|}{\min_{i=1,2,\dots,M} \|u^T x^i + v\|} \leq \frac{R}{d} \quad (\text{A.7})$$

Denoting

$$h = \frac{\max_{i=1,2,\dots,M} \|u^T x^i + v\|}{\min_{i=1,2,\dots,M} \|u^T x^i + v\|}, \quad (\text{A.8})$$

we can write

$$h \leq \frac{R}{d}. \quad (\text{A.9})$$

$$\Rightarrow h^2 \leq \left(\frac{R}{d}\right)^2 < 1 + \left(\frac{R}{d}\right)^2. \quad (\text{A.10})$$

Assuming that the dimension n of the data samples is sufficiently large, we have, from (1),

$$\gamma \leq 1 + \left(\frac{R}{d}\right)^2. \quad (\text{A.11})$$

Hence, $\exists \beta \in \mathbb{R}$, $\beta > 0$, such that

$$\gamma \leq \beta h^2. \quad (\text{A.12})$$

We also note that

$$h^2 \geq 1, \quad (\text{A.13})$$

the minimum being achieved when all samples are equidistant from the separating hyperplane.

Also note that the VC dimension γ satisfies

$$\gamma \geq 1. \quad (\text{A.14})$$

In short, both h^2 and γ have the same lower and upper bounds, i.e. they are of the same order. Therefore, from (A.12)–(A.14), we note that there exist constants $\alpha, \beta > 0$, $\alpha, \beta \in \mathbb{R}$ such that

$$\alpha h^2 \leq \gamma \leq \beta h^2, \quad (\text{A.15})$$

or, in other words, h^2 constitutes a tight or exact (θ) bound on the VC dimension γ .

Since the dataset has been assumed to be linearly separable, we have

$$u^T x^i + v \geq 0 \quad \text{if } y_i = 1 \text{ (Class 1 points)} \quad (\text{A.16})$$

$$u^T x^i + v \leq 0 \quad \text{if } y_i = -1 \text{ (Class -1 points)} \quad (\text{A.17})$$

We also have

$$\|u^T x^i + v\| = \begin{cases} u^T x^i + v & \text{if } u^T x^i + v \geq 0 \\ -(u^T x^i + v) & \text{if } u^T x^i + v \leq 0 \end{cases} \quad (\text{A.18})$$

Therefore, from (A.16)–(A.18), we can write

$$\|u^T x^i + v\| = y_i \cdot [u^T x^i + v], \quad i = 1, 2, \dots, M. \quad (\text{A.19})$$

Therefore,

$$h = \frac{\max_{i=1,2,\dots,M} y_i (u^T x^i + v)}{\min_{i=1,2,\dots,M} y_i (u^T x^i + v)}. \quad (\text{A.20})$$

Appendix B. The hard margin MCM formulation

In this appendix, we derive the hard margin MCM formulation in the input space. We begin with the optimization problem in (6), which was obtained from the exact bound on γ derived in Appendix A. 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; hence, no other constraints are present in the optimization problem (6). For the convenience of the reader, (6) [also (A.20)] is reproduced as

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

The problem (6) may be rewritten as

$$\min_{u,v,g,l} \frac{g}{l} \quad (\text{B.1})$$

$$g \geq y_i \cdot [u^T x^i + v], \quad i = 1, 2, \dots, M \quad (\text{B.2})$$

$$l \leq y_i \cdot [u^T x^i + v], \quad i = 1, 2, \dots, M \quad (\text{B.3})$$

This is a linear fractional programming problem [34]. We apply the Charnes–Cooper transformation [see 34, p. 463]. This consists of introducing a variable $p = 1/l$, which we substitute into (B.1)–(B.3) to obtain

$$\min_{u,v,g,p,l} h = g \cdot p \quad (\text{B.4})$$

$$g \cdot p \geq y_i \cdot [p \cdot u^T x^i + p \cdot v], \quad i = 1, 2, \dots, M \quad (\text{B.5})$$

$$l \cdot p \leq y_i \cdot [p \cdot u^T x^i + p \cdot v], \quad i = 1, 2, \dots, M \quad (\text{B.6})$$

$$p \cdot l = 1 \quad (\text{B.7})$$

Denoting $w \equiv p \cdot u$, $b \equiv p \cdot v$, and noting that $p \cdot l = 1$, we obtain the following optimization problem:

$$\min_{w,b,h} h \quad (\text{B.8})$$

$$h \geq y_i \cdot [w^T x^i + b], \quad i = 1, 2, \dots, M \quad (\text{B.9})$$

$$1 \leq y_i \cdot [w^T x^i + b], \quad i = 1, 2, \dots, M \quad (\text{B.10})$$

which may be written as

$$\min_{w,b,h} h \quad (\text{B.11})$$

$$h \geq y_i \cdot [w^T x^i + b], \quad i = 1, 2, \dots, M \quad (\text{B.12})$$

$$y_i \cdot [w^T x^i + b] \geq 1, \quad i = 1, 2, \dots, M \quad (\text{B.13})$$

We refer to the problem (B.11)–(B.13) as the hard margin Linear Minimum Complexity Machine (Linear MCM). Note that h^2 is an exact bound on γ , the VC dimension of the classifier.

References

- [1] C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.* 20 (3) (1995) 273–297.
- [2] J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural Process. Lett.* 9 (3) (1999) 293–300.
- [3] L. Yang, S. Yang, R. Zhang, H. Jin, Sparse least square support vector machine via coupled compressive pruning, *Neurocomputing* 131 (2014) 77–86.
- [4] S. Ding, X. Hua, Recursive least squares projection twin support vector machines for nonlinear classification, *Neurocomputing* 130 (2014) 3–9.
- [5] D. Wu, L. Shao, Multi-max-margin support vector machine for multi-source human action recognition, *Neurocomputing* 127 (2014) 98–103.
- [6] B. Li, S. Song, K. Li, A fast iterative single data approach to training unconstrained least squares support vector machines, *Neurocomputing* 115 (2013) 31–38.
- [7] Z. Wu, H. Zhang, J. Liu, A fuzzy support vector machine algorithm for classification based on a novel (PIM) fuzzy clustering method, *Neurocomputing* 125 (2014) 119–124.
- [8] W. Wenjian, G. Husheng, J. Yuanfeng, B. Jingye, Granular support vector machine based on mixed measure, *Neurocomputing* 101 (2013) 116–128.
- [9] X. Peng, D. Xu, Bi-density twin support vector machines for pattern recognition, *Neurocomputing* 99 (2013) 134–143.
- [10] A. Jalalian, S.K. Chalup, Gdtw-p-svms: variable-length time series analysis using support vector machines, *Neurocomputing* 99 (2013) 270–282.
- [11] X. Peng, D. Xu, Norm-mixed twin support vector machine classifier and its geometric algorithm, *Neurocomputing* 99 (2013) 486–495.
- [12] X. Huang, S. Mehrkanoon, J.A. Suykens, Support vector machines with piecewise linear feature mapping, *Neurocomputing* 117 (2013) 118–127.
- [13] T.A. Gomes, R.B. Prudêncio, C. Soares, A.L. Rossi, A. Carvalho, Combining meta-learning and search techniques to select parameters for support vector machines, *Neurocomputing* 75 (1) (2012) 3–13.
- [14] T. Ni, W.-Z. Gu, J. Zhai, An inexact smoothing-type algorithm for support vector machines, *Neurocomputing* 129 (2014) 127–135.
- [15] X. Peng, D. Xu, Geometric algorithms for parametric-margin vector machine, *Neurocomputing* 99 (2013) 197–205.
- [16] G. Yin, Y.-T. Zhang, Z.-N. Li, G.-Q. Ren, H.-B. Fan, Online fault diagnosis method based on incremental support vector data description and extreme learning machine with incremental output structure, *Neurocomputing* 128 (2014) 224–231.
- [17] A.L. Chau, X. Li, W. Yu, Convex and concave hulls for classification with support vector machine, *Neurocomputing* 122 (2013) 198–209.
- [18] Y. Kumar, M. Dewal, R. Anand, Epileptic seizure detection using [DWT] based fuzzy approximate entropy and support vector machine, *Neurocomputing* 133 (2014) 271–279.
- [19] A. Ruano, G. Madureira, O. Barros, H. Khosravani, M. Ruano, P. Ferreira, Seismic detection using support vector machines, *Neurocomputing* 135 (2014) 273–283.
- [20] H. Liu, S. Li, Decision fusion of sparse representation and support vector machine for [SAR] image target recognition, *Neurocomputing* 113 (2013) 97–104.
- [21] Y.-P. Zhao, J. Zhao, M. Zhao, Twin least squares support vector regression, *Neurocomputing* 118 (2013) 225–236.
- [22] R. Ji, Y. Yang, W. Zhang, Incremental smooth support vector regression for takagi-sugeno fuzzy modeling, *Neurocomputing* 123 (2014) 281–291.
- [23] Y.-P. Zhao, J.-G. Sun, Z.-H. Du, Z.-A. Zhang, Y.-C. Zhang, H.-B. Zhang, An improved recursive reduced least squares support vector regression, *Neurocomputing* 87 (2012) 1–9.
- [24] L. Sun, C. de Visser, Q. Chu, J. Mulder, A novel online adaptive kernel method with kernel centers determined by a support vector regression approach, *Neurocomputing* 124 (2014) 111–119.
- [25] C.J. Burges, A tutorial on support vector machines for pattern recognition, *Data Min. Knowl. Discov.* 2 (2) (1998) 121–167.
- [26] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, A framework for structural risk minimisation, in: *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, ACM Desenzano sul Garda, Italy, June 28–July 01, 1996, pp. 68–76.
- [27] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, *IEEE Trans. Inf. Theory* 44 (5) (1998) 1926–1940.
- [28] V.N. Vapnik, V. Vapnik, *Statistical Learning Theory*, vol. 2., Wiley, New York, 1998.
- [29] B. Schölkopf, A.J. Smola, *Learning with Kernels*, “The” MIT Press Cambridge, MA, USA, 2002.
- [30] MATLAB, version R2012a, The MathWorks Inc., Natick, Massachusetts, 2012.
- [31] K. Bache, M. Lichman, UCI Machine Learning Repository, 2013, URL: (<http://archive.ics.uci.edu/ml>).
- [32] J. Platt, et al., *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*.

- [33] S. Joshi, Jayadeva, G. Ramakrishnan, S. Chandra, Using sequential unconstrained minimization techniques to simplify svm solvers, *Neurocomputing* 77 (1) (2012) 253–260.
- [34] S. Chandra, Jayadeva, A. Mehra, *Numerical Optimization with Applications*, Alpha Science International, Oxford, UK., 2009.



Jayadeva obtained his B.Tech and Ph.D degrees from the Department of Electrical Engineering, IIT Delhi, in 1988 and 1993 respectively. He is currently Microsoft Chair Professor in the same department. He has been a speaker of the IEEE Computer Society Distinguished Visitor Programme, and is a recipient of the Young Engineer Award from the Indian National Academy of Engineering, the Young Scientist Award from the Indian National Science Academy, and the BOYSCAST Fellowship from the Department of Science and Technology, Govt. of India. He was a URSI Young Scientist at the General Assembly in Lille, France (1996), and received the Sir J.C. Bose Young Scientist title from the Indian

Council of the URSI. He visited the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology in 1997 as a BOYSCAST fellow. He spent a sabbatical year as a visiting Researcher at IBM India Research Laboratory, from July 2006. One of his papers in Neurocomputing was listed on the Top25 hotlist; he is a recipient of best paper awards from the IETE Journal of Research, and two other conference papers. He holds a US Patent on A/D conversion, another on assessing pronunciation abilities, and is the co-author of the book "Numerical Optimization and Applications". He has served on the Steering and Program Committees of several international conferences. He was a keynote speaker at the International Workshop on Mobile Technologies for Pervasive Healthcare, Philip Island, Australia, in 2007, was tutorial chair for PrEMI 2009, delivered an invited talk at ICICS, Singapore, 2011, and was a keynote speaker at the IEEE Symposium Series on Computational Intelligence, Singapore, 2013. His research interests include Swarm Intelligence, Machine Learning, Optimization, and VLSI. Amongst his recent works that has received notable attention is the Twin Support Vector Machine, which in July 2014 had around 275 citations.