A first approximation in order to define a Difficulty Factor of the bi-classification in a dataset by using SVMs*

L. Gonzalez-Abril[a] and C. Angulo[b]
[a] Applied Economics I Dept., Seville University, 41018 Seville, Spain
[b] Automatic Control Dept., Universitat Politècnica de Catalunya.
08800 - Vilanova i la Geltrú, Spain
luisgon@us.es, cecilio.angulo@upc.edu

Abstract

The main aim in this paper is to analyze the complexity of a Support Vector Machine –SVM– in the construction of a classifier for a bi-classification problem on a specific dataset. Hence, an index is defined in terms of both, the Lagrange multipliers and the number of support vectors. Experimentation for checking the defined index is carried out with a well-known dataset, the Glass Identification Database.

1 Introduction

SVMs are learning machines which implement the structural risk minimization inductive principle [5]. This theory was developed on the basis of a separable binary classification problem. Complexity is a property in the machine learning domain with multiple definitions. We will define the dual "problem-solution difficulty" referred to:

- Difficulty of the problem, mainly related to 'linear separability'. Hence, a linearly separable problem is the most simple problem.
- Difficulty associated to the available information, usually the training set. Facing the same problem, data availability can convert the binomial problem-solution in either, more complex or simple.
- Complexity of the solution. In front of the same dataset (even whether they come from different problems), simpler / more complex structures can be chosen.

For linearly separable problems, linear solutions should have a similar difficulty according to the defined index. For non-linearly separable problems, it must be checked which kernel is being used. By using a Gaussian kernel the problem reduces to a separable one in the feature space, so a different concept should be applied to measure difficulty, probably related with kernel parameters and the regularization term.

The remainder of this paper is arranged as follows: Section 2 presents the standard SVM approach. Section 3 puts forward the justification for the definition of a *difficulty factor* which quatifies how difficult is for a SVM to carry out a classification in a bi-classification problem, based on a specific dataset. An experiment is carried out in Section 4 in order to show the usefulness of this difficulty factor. Finally, conclusions and future work are drawn.

*This work has been partially supported by the Spanish Ministry of Science and Innovation R&D under the project TIN2009-14378-C02-01 and the Spanish Ministry of Economy and Competitiveness under the project TIN2012-38416-C03-01.

2 Standard SVM Approach

Let $\mathcal{Z} = \{z_i\}_{i=1}^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set, with $x_i \in \mathcal{X}$ as the input space and $y_i \in \{+1, -1\}$ the output space. Let $\phi : \mathcal{X} \to \mathcal{F}$ be a feature mapping with a dot product denoted by $\langle \cdot, \cdot \rangle$. A linear classifier $f_{\mathrm{w}}(x) = \langle \mathrm{x}, \mathrm{w} \rangle + b$ is sought in $\mathcal{F}$, with $b \in \mathbb{R}$. Outputs are obtained in the form $h_{\mathrm{w}}(x) = \mathrm{sign}(f_{\mathrm{w}}(x))$.

For the standard primal SVM 2-norm formulation [5], the optimization problem becomes

$$\begin{array}{l} \min\limits_{\mathrm{w} \in \mathcal{F}, b \in \mathcal{R}} \frac{1}{2}\|\mathrm{w}\|^2 + C \sum_i \xi_i \\ \text{s.t. } y_i \left(\langle \mathrm{x_i}, \mathrm{w} \rangle + b\right) + \xi_i \geq 1, \quad \xi_i \geq 0, \quad z_i \in \mathcal{Z} \end{array} \tag{1}$$

where C is the regularization term and ξ_i are slack variables.

The solution of the optimization problem (1) can be written as $\mathrm{w} = \sum_{i=1}^n \alpha_i y_i \mathrm{x_i}$ where α_i are Lagrange multipliers for the dual problem of (1). Furthermore,

$$\sum_{i=1}^n \alpha_i y_i = 0 \tag{2}$$

$$0 \leq \alpha_i \leq C, \quad i = 1, \cdots, n \tag{3}$$

$$\alpha_i (y_i \left(\langle \mathrm{x_i}, \mathrm{w} \rangle + b\right) - 1 + \xi_i) = 0, \quad i = 1, \cdots, n \tag{4}$$

Thus, a vector x_i is called a support vector (SV) when $\alpha_i \neq 0$. The bias, term b, is calculated a posteriori [2] from (4). Hence, the classifier can be written as

$$f(x) = \sum_i \alpha_i y_i K(x_i, x) + b \tag{5}$$

where $K(\cdot, \cdot)$ is the well-known kernel function [4], $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathcal{R}$, which is defined as $K(x, y) = \langle \mathrm{x}, \mathrm{y} \rangle = \langle \phi(x), \phi(y) \rangle$.

Note that if α_i^+ and α_j^- are multipliers associated to the vectors of $\mathcal{Z}_{(+)} = \{(x_i, y_i) \in \mathcal{Z}, \ y_i = +1\}$ and $\mathcal{Z}_{(-)} = \{(x_j, y_j) \in \mathcal{Z}, \ y_j = -1\}$, respectively, then, from (2), $\sum_i \alpha_i^+ = \sum_j \alpha_j^- > 0$, since otherwise $\mathrm{w} = 0$.

Actas de las XV Jornadas de ARCA, JARCA 2013
Sistemas Cualitativos y sus Aplicaciones en
Diagnosis, Robótica e Inteligencia Ambiental
I. Sanz, Ll. Museros, J. A. Ortega, A. Fernández-Montes (eds.)

ISBN: 978-84-616-7622-4 © 2013

From the previous results, it can be agreed that the solution vector w depends on the training set $\mathcal{Z}$, the regularization term C and the kernel $K(\theta)$, where θ is the parameter vector for the kernel K. Therefore,

$$\mathrm{w} = \mathrm{w}(\mathcal{Z}, C, K(\theta))$$

Let us denote $\alpha = \sum_i \alpha_i$. It has been shown in [1] that the α value can be interpreted as the strength that support vectors must attain in order to obtain good accuracy in generalization. Hence, it is important to take into account that α is in strong relationship with the "difficulty" of the optimization problem (1). Another expression for α is that given in [3], as follows:

$$\alpha = \|\mathrm{w}\|^2 + \sum_{i \in SV} \alpha_i \xi_i \quad (6)$$

where $SV = \{x_i \in \mathcal{X} | \alpha_i = 0\}$. This new expression (6) for α is very useful since in it appeares the relathionship between the solution, the Lagrange multipliers, the number of support vectors and the slack variables of the problem (1).

It can be noted, as a direct corollary, that if the problem (1) is linearly separable, that is, $\xi_i = 0, \forall i$, then $\alpha = \|\mathrm{w}\|^2$.

Let $N_{(+)}^{SV} = \#\{\alpha_i^+, \alpha_i^+ \neq 0\}$ be the number of SVs for the positive class, $N_{(-)}^{SV} = \#\{\alpha_i^-, \alpha_i^- \neq 0\}$ be the number of SVs for the negative class, and $N^{SV} = N_{(+)}^{SV} + N_{(-)}^{SV}$ be the total number of SVs. Therefore, lower and upper bounds for the value of α can be also obtained from [3],

$$0 < \|\mathrm{w}\|^2 \leq \alpha \leq 2 \cdot C \cdot \min\left\{N_{(+)}^{SV}, N_{(-)}^{SV}\right\} \quad (7)$$

The value of C is given a priori in the problem (1), therefore, higher is the value for α, higher value is for $\min\left\{N_{(+)}^{SV}, N_{(-)}^{SV}\right\}$, which usually provides a high number of support vectors N^{SV}. Similarly, looking into the lower bound, a small value for α implies that the margin separating $\mathcal{Z}_{(+)}$ and $\mathcal{Z}_{(-)}$, $\frac{2}{\|\mathrm{w}\|^2}$, is large. Hence, the solution will provide good generalization performance as well as being smooth (small VC-dimension), and therefore its reliability is better than for a sharp solution (α value is high).

3 Introducing a Difficulty Factor

Let us suppose that a company is in charge of a project involving to solve a task by means of a certain number of workers (the training set $\mathcal{Z}$). The company has a technical expertise (the kernel function and its parameters, $K(\theta)$), and let's suppose that all the N workers are equally qualified[1].

In order to carry out the project, each worker can spend a maximum of C resources to complete its corresponding work. In this point, it is possible to consider that C is the number maximum in hours that a worker spends in the project. This condition is given by (3).

In order to construct an index about how difficult is the project in hands, two criterion should be considered:

Criterion 1: A project is more difficult than another one if a higher number of workers is needed in its ejecution.

Let us remember that all workers are equally qualified. Hence, a factor to consider is $P = \frac{N_{SV}}{N}$, where N_{SV} denotes the number of workers needed in finishing the project, that is, the number of SVs (number of instances with Lagrange multiplier nonzero). Therefore, P denotes the proportion of workers needed to complete the project, and it is clear that $0 < P \leq 1$.

Criterion 2: A project is more difficult than another one if a higher number of resources (hours) is needed in its ejecution.

Each worker spends in the project a total of α_i resources (hours) in his/her work from the avalaible C hours. Thus, the total of resources needs in the project is $\alpha = \sum_i \alpha_i$.

Furthermore, using (7), the optimum value of the resources needs to complete the project is $\|\mathrm{w}\|^2$. In the case that this value is not reached, it will be due to the difficulties found during the project execution.

Hence, the quotient between α and $\|\mathrm{w}\|^2$, $Q = \frac{\alpha}{\|\mathrm{w}\|^2}$, denotes the number of times that the optimum value has been exceeded, and $1 \leq Q < +\infty$. The interpretation of this quotient is as follows: the higher Q, the greater the difficulty of the problem.

In this point, it is possible to interpret that the optimal policy is to attain that $\alpha = \|\mathrm{w}\|^2$, however it is not true since if $C = +\infty$ there are a higher probability to attain this result. This situation is not realistic since in this case $\|\mathrm{w}\|^2$ can be greater and, maybe, a better solution could be obtained with a lower value of C, that is, a lower cost to complete the project. This statement is confirmed in the next section.

From the previous considerations, an index, called Complexity Factor and denoted by $DFactor$, is defined, in order to quantify the difficulty of a classification in a biclassification problem with SVM, as follows:

$$DFactor = DFactor(\mathcal{Z}, C, K(\theta)) = \frac{\alpha}{\|\mathrm{w}\|^2} \cdot \frac{N_{SV}}{N}$$

In a similar form to the optimization problem (1), this index depends on the training set $\mathcal{Z}$, the regularization term C and the used kernel $K(\theta)$, where θ is the parameter vector for the kernel K.

It is worth noting that the $DFactor$ coeficient is adimensional, and therefore, it is useful to compare different classification problems on different datasets. Furthermore, it can be checked that $0 < DFactor < +\infty$.

4 Experimental Results

The use of the just introduced difficulty factor is conducted on the widely used Glass Identification Database from the UCI Repository[2]. A summary of the characteristics of this dataset is: 214 instances, 6 classes, 9 features and 70, 76, 17, 13, 9, 29 instances per class.

The difficulty factor will be evaluated on models from learned SVM using the linear kernel, which is chosen as a baseline for the empirical evaluation, and C is explored on a one-dimensional grid with the following values: $C = [2^{-5}, 2^{-4}, \ldots, 2^8, 2^9]$.

[1] This restriction can be relaxed for the general case.

[2] Available at http://www.ics.uci.edu/~mlearn/MLRepository.html

For each classe C_i, $i = 1, 2, \cdots, 6$, a $i - v - r$ SVM is considered, where the positive class is the i-class. The performance for the $1 - v - r$ SVM is also evaluated in the form of the accuracy rate, iin order to check the relationship between the Complexity Factor and the Accuracy.

The Glass Identification Database is randomly partitioned by stratified sampling into a training set (with 107 instances, that is, the 50% of the dataset) and a test set. This procedure is repeated 50 times in order to ensure good statistical behaviour.

The value of C, the square of the norm of w, the value of α, the number of support vectors, the Complexity factor and the Accuracy are reported in Tables 1 and 2. Let us indicate that that, except to the C-paramenter, the rest of values are given by the mean of the 50 times that the experiment is carried out. Furthermore, the correlation coeficient between the Complexity Factor and the Accuracy, denoted by ρ, for each values of the C parameter is calculated per classes.

Some conclusions can be drawn from the experimentation carried out:

- It can be seen that there is a low correlation between the $DFactor$ and the Accuracy (see the coefficient ρ in the upper-right corner for each class considered). That is, a large difficulty does not imply a large Accuracy, which is a logical result.
- Furthemore, sometimes the correlation is positive (classes 1, 2, 4 and 6) and in other cases is negative (classes 3 and 5).
- With respect to the imbalance in the instances of the dataset, it can be seen that this fact does not increase the difficulty in the classification problem. Thus, the $DFactor$ for the 1-, 2- and 3-classes (70, 76 and 17 positive instances, respectively) is greater than the $DFactor$ for the 4- and 5-class (13 and 9 positive instances, respectively). Nevertheless, the $DFactor$ for the 4-class (with 13 positive instances) is greater than the $DFactor$ for the 6-class with 29 positive instances.

Class	1	2	3	4	5	6
$DFactor$	5.1	56.2	2702.5	0.6	0.07	0.3

- If the $DFactor$ for each i-class is considered as a function of C, that is, $DFactor = DFactor(C)$, then it can be seen that the behavior of the $DFactor$ is similar to a parabola, that is, it is starting like a decreasing funtion, next it is an increasing function. Therefore, it could be a good approcah to calculate the minimum value for this coefficient.
- It is worth noting that, if the number of support vectors is considered as a function of C, that is, $N_{SV} = N_{SV}(C)$, then $N_{SV}(C)$ is an decreasing function. This is a coherent result if the toy example is considered. Thus, if the number of hours (C) given to a worker is high, then less workers are neccesary to finish the project.

5 Conclusions and Future Work

A preliminar study has been carried out in order to analyze the complexity of a Support Vector Machine –SVM– in the construction of a classifier in a bi-classification problem on a dataset. For this end, an difficulty index, called Difficulty Factor $DFactor$, has been defined in terms of the Lagrange multipliers and the number of support vectors.

The results of the experimentation are very promising. Nevertheless, a more extensive experimentation must be carried out with other datasets, as well as using other kernels, like the Gaussian Kernel. In the future, a comparative with other approaches must be also developed.

Furthermore, a more extensive theoretical study on this index is necessary in order to justify the behaviour of the square of the norm of w, the value of α, the number of support vectors and the Difficulty factor as a function of the C parameter, the kernel and its parameters.

References

[1] L. González, C. Angulo, F. Velasco, and A. Català. Dual unification of bi-class support vector machine formulations. *Pattern Recognition*, 39(7):1325–1332, 2006.

[2] L. Gonzalez-Abril, C. Angulo, F. Velasco, and J.A. Ortega. A note on the bias in SVMs for multi-classification. *IEEE Transactions on Neural Networks*, 19(4):723–725, January 2008.

[3] L. Gonzalez-Abril, F. Velasco, C. Angulo, and J.A. Ortega. A study on output normalization in multiclass svms. *Pattern Recognition Letters*, 34:344 – 348, 2013.

[4] B. Scholkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002.

[5] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc, 1998.

Table 1: Results of the experiment for Glass dataset (214 instances, 6 classes and 50% of trainning set) (I).

Class 1		70 positive instances			$\rho = 0.297$
C	$\|\|w\|\|^2$	α	N_{SV}	$DFactor$	$Accuracy$
0,03125	0,15003	2,18750	72,54	14,63597	67,66355
0,0625	0,48345	4,32947	72,06	7,67926	68,95327
0,125	1,09126	8,35777	69,88	5,47608	69,79439
0,25	2,15407	15,94600	67,24	**5,13683**	70,93458
0,5	4,01776	30,29983	64,06	5,15374	72,78505
1	6,54621	57,31328	61,22	5,89962	74,03738
2	9,70047	108,81071	58,80	7,06987	74,46729
4	14,34320	208,32346	56,72	8,78393	74,39252
8	22,35422	406,65687	55,76	11,09728	74,20561
16	35,76606	797,06584	54,46	14,20072	74,54206
32	60,06251	1540,82363	53,80	18,35958	74,56075
64	89,03755	3060,47578	53,30	25,86930	74,44860
128	126,71379	6080,18781	52,76	40,40407	74,50467
256	154,95411	12041,03911	52,44	72,30871	74,48598
512	177,33032	23932,98678	52,28	136,83068	74,22430
Class 2		76 positive instances			$\rho = 0.581$
C	$\|\|w\|\|^2$	α	N_{SV}	$DFactor$	$Accuracy$
0,03125	0,06041	2,36770	79,38	464,81064	64,05607
0,0625	0,14997	4,70325	79,16	369,61696	63,70093
0,125	0,36311	9,32261	78,68	239,15487	63,14019
0,25	0,78845	18,41372	77,92	155,51249	63,73832
0,5	1,49787	36,25420	76,88	97,83539	63,64486
1	2,71356	71,37772	75,68	70,20220	63,51402
2	4,59433	140,62285	74,92	59,43410	63,47664
4	8,81604	277,97082	74,18	**56,24490**	63,55140
8	19,53419	550,64070	73,38	58,59186	64,00
16	43,85392	1089,28432	72,64	57,14370	64,48598
32	82,11256	2145,65345	71,96	80,00543	64,26168
64	137,24884	4224,86787	70,94	137,01672	64,85981
128	184,48199	8325,67611	70,58	255,19730	64,69159
256	231,07883	16479,65711	69,64	490,59029	65,02804
512	256,42297	32735,22822	69,20	949,63084	65,34579
Class 3		17 positive instances			$\rho = -0.660$
C	$\|\|w\|\|^2$	α	N_{SV}	$DFactor$	$Accuracy$
0,03125	0,00017	0,53750	27,12	2793,48565	92,14953
0,0625	0,00056	1,07500	27,68	3769,80355	92,14953
0,125	0,00205	2,15000	27,14	**2702,50815**	92,14953
0,25	0,00773	4,30000	26,96	2843,29992	92,14953
0,5	0,03000	8,60000	26,60	3040,71169	92,14953
1	0,11740	17,20000	26,36	5530,12748	92,13084
2	0,46654	34,40000	25,62	9383,22044	91,98131
4	1,78116	68,76392	25,44	19443,16103	91,57009
8	5,79765	137,01548	25,68	38591,26219	91,14019
16	13,65781	270,57655	26,08	235792,85466	90,99065
32	27,78309	531,92295	25,88	201069,10037	90,67290
64	52,99602	1043,64106	26,06	342287,72468	90,56075
128	90,15940	2045,05415	25,92	987594,66576	90,22430
256	123,54940	4008,99558	25,84	1723266,24672	90,22430
512	179,21435	7912,59284	26,04	3933230,81072	90,24299

Table 2: Results of the experiment for Glass dataset (214 instances, 6 classes and 50% of trainning set) (II).

Class 4		13 positive instances			$\rho = 0.384$
C	$\|\|w\|\|^2$	α	N_{SV}	$DFactor$	$Accuracy$
0,03125	0,05122	0,39052	15,60	4,96336	94,44860
0,0625	0,07849	0,73820	15,18	3,14286	94,42991
0,125	0,16698	1,42739	14,74	2,29412	94,18692
0,25	0,42744	2,77050	14,54	1,72759	93,98131
0,5	1,07557	5,31616	14,30	1,30854	93,96262
1	2,58171	10,03674	13,96	0,96095	93,77570
2	5,81971	18,48299	13,38	0,75971	93,75701
4	11,10862	32,72208	12,64	**0,69914**	94,01869
8	19,10062	56,59447	11,98	0,76634	94,35514
16	31,99856	98,14128	11,24	0,80549	94,46729
32	50,03263	170,36773	10,76	0,80299	94,24299
64	82,62073	299,76689	10,72	0,99012	94,05607
128	116,62966	527,10655	10,64	1,27920	94,05607
256	157,20173	949,40599	10,54	2,13086	94,05607
512	194,99134	1751,65450	10,58	3,83468	94,05607
Class 5		9 positive instances			$\rho = -0.673$
C	$\|\|w\|\|^2$	α	N_{SV}	$DFactor$	$Accuracy$
0,03125	0,01861	0,28125	12,00	4,00858	95,79439
0,0625	0,06455	0,55834	12,00	2,00510	95,81308
0,125	0,15364	1,07914	11,92	1,11219	95,92523
0,25	0,40226	2,08659	11,60	0,68066	95,94393
0,5	1,19133	4,02233	11,52	0,43729	95,88785
1	3,23076	7,49134	11,60	0,28608	96,56075
2	7,49622	12,99237	10,72	0,19302	97,55140
4	13,50557	20,28361	9,90	0,14724	97,71963
8	21,40755	29,29908	8,96	0,11928	97,83178
16	31,79014	40,27344	8,34	0,09954	97,62617
32	43,68375	51,89686	7,98	0,08608	97,43925
64	59,03985	64,78150	7,94	0,07794	97,19626
128	73,70016	73,78895	7,92	0,07405	97,15888
256	73,84336	73,86108	7,92	**0,07403**	97,15888
512	73,84336	73,86108	7,92	0,07403	97,15888
Class 6		29 positive instances			$\rho = 0.563$
C	$\|\|w\|\|^2$	α	N_{SV}	$DFactor$	$Accuracy$
0,03125	0,33592	0,63975	23,06	0,42353	96,05607
0,0625	0,48026	0,98561	18,76	0,37872	96,00000
0,125	0,69045	1,55173	15,78	0,36205	95,73832
0,25	1,05316	2,52502	13,90	0,35725	95,92523
0,5	1,70843	4,19311	12,56	0,35199	95,81308
1	2,84269	7,01936	11,48	0,36681	95,71963
2	4,87162	11,84844	10,60	0,36494	95,66355
4	8,15177	19,78895	10,24	0,36622	95,47664
8	14,28531	33,25863	9,72	0,37848	95,25234
16	23,21205	55,18516	9,48	0,38052	95,19626
32	38,18519	91,50995	9,16	0,34115	95,04673
64	61,55302	151,63704	9,14	**0,31592**	94,87850
128	85,68839	247,87621	9,10	0,33891	94,89720
256	118,93421	420,19334	9,02	0,31809	94,89720
512	197,24269	747,02958	8,90	0,38135	94,85981