Improvements to Platt’s SMO Algorithm for SVM Classifier Design

ssk@guppy.mpe.nus.edu.sg shirish@csa.iisc.ernet.in cbchiru@csa.iisc.ernet.in murthy@csa.iisc.ernet.in

Technical Report CD-99-14

Control Division
Dept. of Mechanical and Production Engineering
National University of Singapore
Singapore-119260
Ph: (65)-874-4684

1 A revised version of this report is under preparation for submission to a journal. We welcome any comments and suggestions for improving this report.
Abstract

This paper points out an important source of confusion and inefficiency in Platt’s Sequential Minimal Optimization (SMO) algorithm that is caused by the use of a single threshold value. Using clues from the KKT conditions for the dual problem, two threshold parameters are employed to derive modifications of SMO. These modified algorithms perform significantly faster than the original SMO on all benchmark datasets tried.

1 Introduction

In the past few years, there has been a lot of excitement and interest in Support Vector Machines[16, 2] because they have yielded excellent generalization performance on a wide range of problems. Recently, fast iterative algorithms that are also easy to implement have been suggested[9,4,7,3,6]. Platt’s Sequential Minimization Algorithm (SMO)[9,11] is an important example. A remarkable feature of SMO is that it is also extremely easy to implement. Comparative testing against other algorithms, done by Platt, have shown that SMO is often much faster and has better scaling properties.

In this paper we enhance the value of SMO even further. In particular, we point out an important source of confusion and inefficiency caused by the way SMO maintains and updates a single threshold value. Getting clues from optimality criteria associated with the KKT conditions for the dual, we suggest the use of two threshold parameters and devise two modified versions of SMO that remove the confusion associated with SMO and are much more efficient than the original SMO. Computational comparison on a number of benchmark datasets shows that the modifications perform significantly faster than the original SMO in most situations. The ideas mentioned in this paper can also be applied to the SMO regression algorithm[13]. We will report the results of that extension in another paper[12].

The paper is organized as follows. In section 2 we briefly discuss the SVM problem formulation, the dual problem and the associated KKT optimality conditions. We also point out how these conditions lead to proper criteria for terminating algorithms for designing SVM classifiers. Section 3 gives a short summary of Platt’s SMO algorithm. In section 4 we point out the problem associated with the way SMO uses a single threshold value, and describe the modified algorithms in section 5. Computational comparison is done in section 6. The appendix gives the pseudo-codes for our
SMO modifications. These pseudo-codes are very similar to those for the SMO given by Platt in [9]. They are short, and, it is very easy to develop a working code for SVM design using them.

2 The SVM Problem and Optimality Conditions.

The basic problem addressed in this paper is the two category classification problem. The tutorial by Burges[2] gives a good overview of the solution of this problem using SVMs. Throughout the paper we will use $x$ to denote the input vector of the support vector machine and $z$ to denote the feature space vector which is related to $x$ by a transformation, $z = \phi(x)$. As in all SVM designs, we do not assume $\phi$ to be known; all computations will be done using only the kernel function, $k(x, \hat{x}) = \phi(x) \cdot \phi(\hat{x})$, where “.” denotes inner product in the $z$ space. Let $\{(x_i, y_i)\}$ denote the training set, where $x_i$ is the $i$-th input pattern and $y_i$ is the corresponding target value; $y_i = 1$ means $x_i$ is in class 1 and $y_i = -1$ means $x_i$ is in class 2. Let $z_i = \phi(x_i)$. The optimization problem solved by the support vector machine is:

$$\min \frac{1}{2} \|w\|^2 + C \sum_i \xi_i$$  \hspace{1cm} (1a)

subject to : $y_i(w \cdot z_i - b) \geq 1 - \xi_i \ \forall \ i; \ \xi_i \geq 0 \ \forall \ i$  \hspace{1cm} (1b)

This problem is referred to as the primal problem. The Lagrangian for this problem is:

$$L = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i + \sum_i \alpha_i [1 - \xi_i - y_i(w \cdot z_i - b)] - \sum_i \pi_i \xi_i$$

The KKT optimality conditions are given by:

$$\nabla_w L = w - \sum_i \alpha_i y_i z_i = 0; \ \frac{\partial L}{\partial b} = \sum_i \alpha_i y_i = 0; \ \frac{\partial L}{\partial \xi_i} = C - \alpha_i - \pi_i = 0 \ \forall i;$$

$$\alpha_i \geq 0, \ \alpha_i [1 - \xi_i - y_i(w \cdot z_i - b)] = 0, \ \pi_i \geq 0, \ \pi_i \xi_i = 0 \ \forall i.$$  \hspace{1cm}

We will refer to the $\alpha_i$’s as Lagrange multipliers. Let us define

$$w(\alpha) = \sum_i \alpha_i y_i z_i$$

Using Wolfe duality theory[16,2] it can be shown that the $\alpha_i$’s are obtained by solving the following Dual problem:

$$\max W(\alpha) = \sum_i \alpha_i - \frac{1}{2} w(\alpha) \cdot w(\alpha)$$  \hspace{1cm} (2a)

2
subject to $0 \leq \alpha_i \leq C, \sum_{i} \alpha_i y_i = 0$ (2b)

Once the $\alpha_i$’s are obtained, the other primal variables, $w$, $b$, $\xi$ and $\pi$ can be easily determined by using the KKT conditions mentioned earlier. It is possible that the solution is non-unique; for instance, when all $\alpha_i$s take the boundary values of 0 and $C$, it is possible that $b$ is not unique.

The numerical approach in SVM design is to solve the dual (instead of the primal) because it is a finite-dimensional optimization problem. (Note that $w(\alpha) \cdot w(\alpha) = \sum_i \sum_j y_i y_j \alpha_i \alpha_j k(x_i, x_j)$.) To derive proper stopping conditions for algorithms which solve the dual, it is important to write down the optimality conditions for the dual. The Lagrangian for the dual is:

$$L = \frac{1}{2} w(\alpha) \cdot w(\alpha) - \sum_i \alpha_i - \sum_i \delta_i \alpha_i + \sum_i \mu_i (\alpha_i - C) - \beta \sum_i \alpha_i y_i$$

Define

$$F_i = w(\alpha) \cdot z_i - y_i = \sum_j \alpha_j y_j k(x_i, x_j) - y_i$$

The KKT conditions for the dual problem are:

$$\frac{\partial L}{\partial \alpha_i} = (F_i - \beta) y_i - \delta_i + \mu_i = 0, \quad \delta_i \geq 0, \quad \delta_i \alpha_i = 0, \quad \mu_i \geq 0, \quad \mu_i (\alpha_i - C) = 0 \quad \forall i$$

These conditions can be simplified by considering three cases.

Case 1. $\alpha_i = 0$

$$\delta_i \geq 0, \quad \mu_i = 0 \quad \Rightarrow (F_i - \beta) y_i \geq 0 \quad (3a)$$

Case 2. $0 < \alpha_i < C$

$$\delta_i = 0, \quad \mu_i = 0 \quad \Rightarrow (F_i - \beta) y_i = 0 \quad (3b)$$

Case 3. $\alpha_i = C$

$$\delta_i = 0, \quad \mu_i \geq 0 \quad \Rightarrow (F_i - \beta) y_i \leq 0 \quad (3c)$$

Define the following index sets at a given $\alpha$: $I_0 = \{i : 0 < \alpha_i < C\}$; $I_1 = \{i : y_i = 1, \alpha_i = 0\}$; $I_2 = \{i : y_i = -1, \alpha_i = C\}$; $I_3 = \{i : y_i = 1, \alpha_i = C\}$; and $I_4 = \{i : y_i = -1, \alpha_i = 0\}$. Note that these index sets depend on $\alpha$. The necessary conditions in (3a)-(3c) can be rewritten as

$$\beta \leq F_i \forall i \in I_0 \cup I_1 \cup I_2; \quad \beta \geq F_i \forall i \in I_0 \cup I_3 \cup I_4 \quad (4)$$

Define:

$$b_{up} = \min\{F_i : i \in I_0 \cup I_1 \cup I_2\} \quad \text{and} \quad b_{low} = \max\{F_i : i \in I_0 \cup I_3 \cup I_4\} \quad (5)$$
Then optimality conditions will hold at some $\alpha$ iff

$$b_{\text{low}} \leq b_{\text{up}}$$  \hspace{1cm} (6)$$

It is easy to see the close relationship between the threshold parameter $b$ in the primal problem and the multiplier, $\beta$. In particular, at optimality, $\beta$ and $b$ are identical. Therefore, in the rest of the paper $\beta$ and $b$ will denote one and the same quantity.

We will say that an index pair $(i, j)$ defines a violation at $\alpha$ if one of the following sets of conditions holds:

$$i \in I_0 \cup I_3 \cup I_4, \quad j \in I_0 \cup I_1 \cup I_2 \quad \text{and} \quad F_i > F_j$$ \hspace{1cm} (7a)$$

and

$$i \in I_0 \cup I_1 \cup I_2, \quad j \in I_0 \cup I_3 \cup I_4 \quad \text{and} \quad F_i < F_j$$ \hspace{1cm} (7b)$$

Note that optimality conditions will hold at $\alpha$ iff there does not exist any index pair $(i, j)$ that defines a violation.

Since, in numerical solution, it is usually not possible to achieve optimality exactly, there is a need to define approximate optimality conditions. The condition (6) can be replaced by

$$b_{\text{low}} \leq b_{\text{up}} + 2\tau$$ \hspace{1cm} (8)$$

where $\tau$ is a positive tolerance parameter. (In the pseudo-codes given in [9] and the appendix of this paper, this parameter is referred to as $\text{tol}$.) Correspondingly, the definition of violation can be altered by replacing (7a) and (7b) by:

$$i \in I_0 \cup I_3 \cup I_4, \quad j \in I_0 \cup I_1 \cup I_2 \quad \text{and} \quad F_i > F_j + 2\tau$$ \hspace{1cm} (9a)$$

and

$$i \in I_0 \cup I_1 \cup I_2, \quad j \in I_0 \cup I_3 \cup I_4 \quad \text{and} \quad F_i < F_j - 2\tau$$ \hspace{1cm} (9b)$$

Hereafter in the paper, when optimality is mentioned it will mean approximate optimality.

Since $\beta$ can be placed halfway between $b_{\text{low}}$ and $b_{\text{up}}$, approximate optimality conditions will hold iff there exists a $\beta$ such that (3a)-(3c) are satisfied with a $\tau$-margin, i.e.,

$$ (F_i - \beta)y_i \geq -\tau \quad \text{if} \quad \alpha_i = 0$$ \hspace{1cm} (10a)$$

$$ \left|(F_i - \beta)y_i\right| \leq \tau \quad \text{if} \quad 0 < \alpha_i < C$$ \hspace{1cm} (10b)$$

and

$$ (F_i - \beta)y_i \leq \tau \quad \text{if} \quad \alpha_i = C$$ \hspace{1cm} (10c)$$

4
(10a)-(10c) are the approximate optimality conditions employed by Platt[9], Joachims[4] and others. In [6] we have argued the soundness of using the above approximate conditions as a stopping criterion for dual algorithms.

3 Platt’s SMO Algorithm.

A number of algorithms have been suggested for solving the dual problem. Traditional quadratic programming algorithms such as the active set method[5] and interior point algorithms[13] are not suitable for large size problems because of the following reasons. First, they require that the kernel matrix $k(x_i, x_j)$ be computed and stored in memory. This requires extremely large memory. Second, these methods involve expensive matrix operations such as the Cholesky decomposition of a large submatrix of the kernel matrix. Third, for practitioners who would like to develop their own implementation of an SVM classifier, coding these algorithms is very difficult.

Several attempts have been made to develop methods that overcome some or all of these problems. Vapnik[15] made the observation that if the number of support vectors is small and they are known beforehand, then one could directly solve the reduced problem involving only the support vectors and thereby deal with significantly larger datasets. Since the support vectors are not known, a beginning set of vectors is chosen and chunked into memory and the resulting problem is solved. Then the remaining vectors are tested for optimality and those that violate are included. The process is repeated until a solution is obtained. This is referred to as the chunking algorithm.

If the number of support vectors itself is large then the chunking algorithm is also unsuitable. Osuna et.al.[8] suggested the use of only a subset of the vectors as a working subset and optimize on the corresponding $\alpha_i$’s while freezing the others. Though the arguments given by Osuna et.al. about the convergence of the algorithm are incorrect, it is expected that the algorithm will converge asymptotically as the number of steps goes to infinity. Joachims[4] has developed an efficient algorithm for SVM by building upon the basic idea given in [8].

Recently Platt suggested an algorithm[9] – Sequential Minimal Optimization (SMO) – that puts the subset selection in Osuna et.al’s algorithm to the extreme by iteratively selecting subsets only of size 2. Note that, because of the presence of the equality constraint (see (2b)), at least two variables need to be chosen for optimization so as to take a step. Platt’s computational experiments[9,11]
have shown SMO to be very much faster than the chunking algorithm; it also scales much better as problem size grows. The SMO algorithm also fares better than Joachim's algorithm[4].

Let us give a brief description of the SMO algorithm. Because the working set is only of size 2 and the equality constraint can be used to eliminate one of the two Lagrange multipliers, the optimization problem at each step is a quadratic minimization in just one variable. It is straightforward to write down an analytic solution for it. Complete details are derived in [9]. The procedure, takeStep (which is a part of the pseudocode given there) gives a clear description of the implementation. There is no need to recall all details here. We only make one important comment on the role of the threshold parameter, $\beta$. As in [9] define the output error on the $i$-th pattern as

$$E_i = F_i - \beta$$

Consistent with the pseudocode of [9] let us call the indices of the two multipliers chosen for optimization in one step as $i_2$ and $i_1$. A look at the details in [9] shows that to take a step by varying $\alpha_{i_1}$ and $\alpha_{i_2}$, we only need to know $E_{i_1} - E_{i_2} = F_{i_1} - F_{i_2}$. Therefore a knowledge of the value of $\beta$ is not needed to take a step.

The method followed to choose $i_1$ and $i_2$ at each step is crucial for efficient solution of the problem. Based on a number of experiments Platt came up with a good set of heuristics. He employs a two loop approach: the outer loop chooses $i_2$; and, for a chosen $i_2$, the inner loop chooses $i_1$. The outer loop iterates over all patterns violating the optimality conditions, first only over those with Lagrange multipliers neither on the upper nor lower boundary, and once all of them are satisfied, over all patterns violating the optimality conditions to ensure that the problem has indeed been solved. Clearly, the algorithm spends a large fraction of its time adjusting the multipliers which take non-boundary values and only a small amount of time with the multipliers that take boundary values. Appropriately, therefore, Platt maintains and updates a cache for $E_i$ values for indices $i$ corresponding to non-boundary multipliers. The remaining $E_i$ are computed as and when needed.

Let us now see how the SMO algorithm chooses $i_1$. The aim is to make a large increase in the objective function. Since it is expensive to try out all possible choices of $i_1$ and choose the one that gives the best increase in objective function, the index $i_1$ is chosen to maximize $|E_{i_2} - E_{i_1}|$. (If we define $\rho(t) = W(\alpha(t))$ where $t$ is a real parameter that denotes the change in the values of $y_{i_1} \alpha_{i_1}$ and $-y_{i_2} \alpha_{i_2}$, and $\alpha(t)$ is the corresponding Lagrangian multiplier vector, then $|\rho'(0)| = |E_{i_1} - E_{i_2}|$.)
Since $E_i$ is available in cache for non-boundary multiplier indices, only such indices are initially used in the above choice of $i_1$. If such a choice of $i_1$ does not yield sufficient progress, then the following steps are taken. Starting from a randomly chosen index, all indices corresponding to non-bound multipliers are tried as choices for $i_1$, one by one. If still sufficient progress is not possible, all indices are tried as choices for $i_1$, one by one, again starting from a randomly chosen index. Thus the choice of random seed affects the running time of SMO; see, for example, the computational costs mentioned in section 5.

Although a value of $\beta$ is not needed to take a step, it is needed if (10a)-(10c) are employed for checking optimality. In the SMO algorithm $\beta$ is updated after each step. If, after a step involving $(i_1, i_2)$, one of $\alpha_{i_1}$, $\alpha_{i_2}$ (or both) takes a non-boundary value then (3b) is exploited to update the value of $\beta$. In the rare case that this does not happen, there exists a whole interval, say, $[\beta_{\text{low}}, \beta_{\text{up}}]$, of admissible thresholds for $\alpha_{i_1}$ and $\alpha_{i_2}$. In this situation SMO simply chooses: $\beta = (\beta_{\text{low}} + \beta_{\text{up}})/2$. In the next section we will see the problems caused by such a choice.

4 Problems with SMO Algorithm.

SMO is a carefully organized algorithm which has excellent computational efficiency. However, because of its way of computing and using a single threshold value it can get into a confused end state and can also become inefficient. Let us illustrate the first issue using a numerical example.

**Example 1.** Consider the following example where there are 3 patterns:

$$y_1 = -1, \ y_2 = y_3 = +1, \ C = \frac{1}{4}, \ \text{Kernel Matrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 6 \end{bmatrix}$$

Suppose we start from $\alpha_1 = \alpha_2 = \alpha_3 = 0$ (the usual point where SMO starts). Calculating $F_i$ we get $F_1 = 1, \ F_2 = F_3 = -1$. All three indices violate the optimality conditions. (Note that $b_{\text{low}} = 1$ and $b_{\text{up}} = -1$; SMO uses $\beta = 0$ to check optimality conditions.) Suppose SMO chooses indices 1 and 2 for optimization, keeping $\alpha_3$ fixed at 0. It is easy to check that this leads to the point, $\alpha_1 = \alpha_2 = C, \ \alpha_3 = 0$. At this new point we have $F_1 = 3/4, \ F_2 = -3/4, \ F_3 = -1/2$. Note that $b_{\text{low}} = -3/4$ and $b_{\text{up}} = -1/2$ and hence optimality conditions are satisfied. SMO chooses $\beta = (F_1 + F_2)/2 = 0$. If this value of $\beta$ is used to check optimality, the third training pattern
shows a violation of the optimality criterion employed by Platt (i.e., (10)), but actually there is no violation! Note that any \( \beta \) chosen from the interval, \([-3/4, -1/2]\) would have ensured the verification of (10).

This example clearly sums up our first concern. Because SMO constrains itself unnecessarily to a particular single choice of the threshold, \( \beta \), it gets into trouble, especially at termination.

The issue raised here appears to be somewhat pathological since the presence of even a single index \( i \) with \( 0 < \alpha_i < C \) forces \( \beta \) to be unique and so there is really no serious problem. (Note that unless \( C \) takes certain extreme values, there is little possibility of not having an index \( i \) with \( 0 < \alpha_i < C \).) But we would like to point out that there is still a practical problem of inefficiency. At any instant, the SMO algorithm fixes \( \beta \) based on the current two indices which are being optimized. However, while checking whether the remaining examples violate optimality or not, it is quite possible that a different, shifted choice of \( \beta \) may do a better job. So, in the SMO algorithm it is quite possible that, even though \( \alpha \) has reached a value where optimality is satisfied (i.e., (8)), SMO hasn’t detected this because it has not identified the correct choice of \( \beta \). It is also quite possible that, a particular index may appear to violate the optimality conditions because (10) is employed using an “incorrect” value of \( \beta \) although this index may not be able to pair with another to define a violation. In such a situation the SMO algorithm does an expensive and wasteful search looking for a second index so as to take a step. We believe that this is a major source of inefficiency in the SMO algorithm.

5 Modifications of the SMO Algorithm.

In this section we suggest two modified versions of the SMO algorithm, each of which overcomes the problems mentioned in the last section. As we will see in the computational evaluation of section 6, these modifications are almost always better than the original SMO algorithm and, in most situations they give quite a remarkable improvement in efficiency when tested on several benchmark problems.

In short, the modifications avoid the use of a single threshold value \( \beta \) and the use of (10) for checking optimality. Instead, two threshold parameters, \( b_{up} \) and \( b_{low} \) are maintained and (8) (or (9)) is employed for checking optimality. The two modifications are adequately described by the
pseudo-codes given in the appendix. We only give some additional pointers that will help to give an easy understanding of the pseudo-codes. We assume that the reader is familiar with [9] and the pseudo-codes given there.

1. Suppose, at any instant, \( F_i \) is available for all \( i \). Let \( i_{\text{low}} \) and \( i_{\text{up}} \) be indices such that

\[
F_{i_{\text{low}}} = b_{\text{low}} = \max\{F_i : i \in I_0 \cup I_3 \cup I_4\} \tag{11a}
\]

\[
F_{i_{\text{up}}} = b_{\text{up}} = \min\{F_i : i \in I_0 \cup I_1 \cup I_2\} \tag{11b}
\]

Then checking a particular \( i \) for optimality is easy. For example, suppose \( i \in I_1 \cup I_2 \). We only have to check if \( F_i < F_{i_{\text{low}}} - 2\tau \). If this condition holds then there is a violation, and, in that case SMO’s \texttt{takeStep} procedure can be applied to the index pair, \((i, i_{\text{low}})\). Similar steps can be given for indices in the other sets. Thus, in our approach, the checking of optimality of \( i_2 \) and the choice of the second index, \( i_1 \) go hand in hand, unlike the original SMO algorithm. As we will see below, we compute and use \((i_{\text{low}}, b_{\text{low}})\) and \((i_{\text{up}}, b_{\text{up}})\) via an efficient updating process.

2. To be efficient, we would, like in the SMO algorithm, spend much of the effort altering \( \alpha_i \), \( i \in I_0 \); cache for \( F_i \), \( i \in I_0 \) are maintained and updated to do this efficiently. And, when optimality holds for all \( i \in I_0 \), only then examine all indices for optimality.

3. Some extra steps are added to the \texttt{takeStep} procedure. After a successful step using a pair of indices, \((i_2, i_1)\), let \( \tilde{I} = I_0 \cup \{i_1, i_2\} \). We compute, partially, \((i_{\text{low}}, b_{\text{low}})\) and \((i_{\text{up}}, b_{\text{up}})\) using \( \tilde{I} \) only (i.e., use only \( i \in \tilde{I} \) in (11)). Note that these extra steps are inexpensive because cache for \( \{F_i : i \in I_0\} \) is available and updates of \( F_{i_1}, F_{i_2} \) are easily done. A careful look shows that, since \( i_2 \) and \( i_1 \) have been just involved in a successful step, each of the two sets, \( \tilde{I} \cap (I_0 \cup I_1 \cup I_2) \) and \( \tilde{I} \cap (I_0 \cup I_3 \cup I_4) \), is non-empty; hence the partially computed \((i_{\text{low}}, b_{\text{low}})\) and \((i_{\text{up}}, b_{\text{up}})\) will not be null elements. Since \( i_{\text{low}} \) and \( i_{\text{up}} \) could take values from \( \{i_2, i_1\} \) and they are used as choices for \( i_1 \) in the subsequent step (see item 1 above), we keep the values of \( F_{i_1} \) and \( F_{i_2} \) also in cache.

4. When working only with \( \alpha_i \), \( i \in I_0 \), i.e., a loop with \texttt{examineAll=0}, one should note that, if (8) holds at some point then it implies that optimality holds as far as \( I_0 \) is concerned. (This is because, as mentioned in item 3 above, the choice of \( b_{\text{low}} \) and \( b_{\text{up}} \) are influenced by all indices in \( I_0 \).) This gives an easy way of exiting this loop.

5. There are two ways of implementing the loop involving indices in \( I_0 \) only (\texttt{examineAll=0}).

\textbf{Method 1.} This is in line with what is done in SMO. Loop through all \( i_2 \in I_0 \). For each \( i_2 \),
Method 2. Always work with the worst violating pair, i.e., choose $i_2 = i_{\text{low}}$ and $i_1 = i_{\text{up}}$.

Depending on which one of these methods is used, we call the resulting overall modification of SMO as SMO-Modification 1 and SMO-Modification 2.

6. When optimality on $I_0$ holds, as already said we come back to check optimality on all indices ($\text{examineAll}=1$). Here we loop through all indices, one by one. Since $(b_{\text{low}}, i_{\text{low}})$ and $(b_{\text{up}}, i_{\text{up}})$ have been partially computed using $I_0$ only, we update these quantities as each $i$ is examined. For a given $i$, $F_i$ is computed first and optimality is checked using the current $(b_{\text{low}}, i_{\text{low}})$. For example, if $i \in I_1 \cup I_2$ and $F_i < b_{\text{low}} - 2\tau$, then there is a violation, in which case we take a step using $(i, i_{\text{low}})$. On the other hand, if there is no violation, then $(i_{\text{up}}, b_{\text{up}})$ are modified using $F_i$, i.e., if $F_i < b_{\text{up}}$ then we do: $i_{\text{up}} := i$ and $b_{\text{up}} := F_i$.

7. Suppose we do as described above. What happens if there is no violation for any $i$ in a loop having $\text{examineAll}=1$? Can we conclude that optimality holds for all $i$? The answer is: YES. This is easy to see from the following argument. Suppose, by contradiction, there does exist one $(i, j)$ pair such that they define a violation, i.e., they satisfy (9). Let us say, $i < j$. Then $j$ would not have satisfied the optimality check in the above described implementation because $F_i$ would have, earlier than $j$ is seen, affected either the calculation of $b_{\text{low}}$ and/or $b_{\text{up}}$ settings. In other words, even if $i$ is mistakenly taken as having satisfied optimality earlier in the loop, $j$ will be detected as violating optimality when it is analysed. Only when (8) holds it is possible for all indices to satisfy the optimality checks. Furthermore, when (8) holds and the loop over all indices has been completed, the true values of $b_{\text{up}}$ and $b_{\text{low}}$, as defined in (5) would have been computed since all indices have been encountered.

6 Computational Comparison.

In this section we compare the performance of our modifications against the original SMO algorithm. We implemented all these methods in Fortran and ran them using f77 on a 200 MHz Pentium machine. The value, $\tau = 0.001$ was used for all experiments. The following standard problems were used in our testing: Wisconsin Breast Cancer data[1, 17]; Two Spirals data[14]; Checkers
data[5]; UCI Adult data[10]; and Web page classification data[10,4]. Except for Checkers data, for which we created a random set of points on a $4 \times 4$ checkers grid (see [6]), all other data sets were downloaded from the sites mentioned in the above references and were used in full for training, i.e., no division into training/validation/test sets was made. In the case of Adult data set, the inputs are represented in a special binary format, as used by Platt in his testing of SMO. To study scaling properties as training data grows, Platt did staged experiments on the Adult and Web data. We have used only the data from the first, fourth and seventh stages. The gaussian kernel,

$$k(x_i, x_j) = \exp(-0.5 \|x_i - x_j\|^2 / \sigma^2)$$

was used in all experiments. The $\sigma^2$ values employed, together with $n$, the dimension of the input, and $m$, the number of training points, are given in Table 1. The $\sigma^2$ values given in the table were chosen as follows. For the Adult and Web data the $\sigma^2$ values are the same as those used by Platt in his experiments on SMO; for other data, we chose $\sigma^2$ suitably to get good generalization.

When a particular method is used for SVM design, the value of $C$ is usually unknown, and it has to be chosen by trying a number of values and using a validation set. Therefore, good performance of a method over a range of $C$ values is important. Therefore for each problem we have tested the algorithms over an appropriate range of $C$ values.

The cost of updating the cache for $F_i$ is the dominant part of the computational cost. Hence
the total number of kernel evaluations is a very good indicator of the computing cost. Since such a measure is pretty much independent of the computing environment used, it is easy for others developing new algorithms to compare their methods against the ones studied in this paper, without actually running these methods again. In Tables 2-10 we have given the total number of kernel evaluations for the various problems tried. To point out the effect of the choice of random seed on the cost associated with the original SMO algorithm, we have reported costs for two random seeds. (We haven’t done this for the Web data since, for that data, change of random seed had no effect on the computational cost.) Our SMO modifications do not require any random seed.

It is very clear that the modifications outperform the original SMO algorithm. In many situations the improvement in efficiency is remarkable. Between the two modifications, the second one fares better overall.

7 Conclusion.

In this paper we have pointed out an important source of inefficiency in Platt’s SMO algorithm that is caused by the operation with a single threshold value. We have suggested two modifications of the SMO algorithm that overcome the problem by efficiently maintaining and updating two threshold parameters. Our computational experiments show that these modifications speed up the SMO algorithm considerably in many situations. Platt has already established the SMO algorithm to be one of the fastest algorithms for SVM design. The modified versions of SMO presented in this paper enhance the value of the SMO algorithm even further. The ideas mentioned in this paper for SVM classification can also be extended to the SMO regression algorithm[13]. We will report the results of that extension in another paper[12].
<table>
<thead>
<tr>
<th>C</th>
<th>SMO Random Seed 1</th>
<th>SMO Random Seed 2</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>47.855</td>
<td>36.822</td>
<td>1.193</td>
<td>3.500</td>
</tr>
<tr>
<td>0.04</td>
<td>2.936</td>
<td>2.671</td>
<td>2.005</td>
<td>1.725</td>
</tr>
<tr>
<td>0.06</td>
<td>2.114</td>
<td>2.648</td>
<td>2.035</td>
<td>1.950</td>
</tr>
<tr>
<td>0.10</td>
<td>1.627</td>
<td>1.824</td>
<td>1.860</td>
<td>1.680</td>
</tr>
<tr>
<td>0.20</td>
<td>1.647</td>
<td>2.045</td>
<td>1.775</td>
<td>1.404</td>
</tr>
<tr>
<td>0.40</td>
<td>1.720</td>
<td>1.372</td>
<td>1.362</td>
<td>1.255</td>
</tr>
<tr>
<td>0.50</td>
<td>1.613</td>
<td>1.618</td>
<td>1.265</td>
<td>1.183</td>
</tr>
<tr>
<td>0.70</td>
<td>1.653</td>
<td>1.377</td>
<td>1.339</td>
<td>1.065</td>
</tr>
<tr>
<td>1.00</td>
<td>1.531</td>
<td>1.560</td>
<td>1.474</td>
<td>1.210</td>
</tr>
<tr>
<td>2.00</td>
<td>1.516</td>
<td>1.686</td>
<td>1.331</td>
<td>1.019</td>
</tr>
<tr>
<td>3.00</td>
<td>1.625</td>
<td>1.690</td>
<td>1.314</td>
<td>0.990</td>
</tr>
</tbody>
</table>

Table 2: Wisconsin Breast Cancer data: Number of Kernel evaluations \( \times 10^{-6} \)

References


<table>
<thead>
<tr>
<th>C</th>
<th>SMO Random Seed 1</th>
<th>SMO Random Seed 2</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0.093</td>
<td>0.100</td>
<td>0.096</td>
<td>0.096</td>
</tr>
<tr>
<td>0.04</td>
<td>0.110</td>
<td>0.134</td>
<td>0.097</td>
<td>0.097</td>
</tr>
<tr>
<td>0.06</td>
<td>0.135</td>
<td>0.469</td>
<td>0.097</td>
<td>0.097</td>
</tr>
<tr>
<td>0.10</td>
<td>0.116</td>
<td>0.117</td>
<td>0.097</td>
<td>0.097</td>
</tr>
<tr>
<td>0.20</td>
<td>0.099</td>
<td>0.148</td>
<td>0.097</td>
<td>0.097</td>
</tr>
<tr>
<td>0.40</td>
<td>0.255</td>
<td>0.179</td>
<td>0.172</td>
<td>0.171</td>
</tr>
<tr>
<td>0.50</td>
<td>0.198</td>
<td>0.241</td>
<td>0.284</td>
<td>0.323</td>
</tr>
<tr>
<td>0.70</td>
<td>0.457</td>
<td>0.445</td>
<td>0.240</td>
<td>0.210</td>
</tr>
<tr>
<td>1.00</td>
<td>0.559</td>
<td>0.548</td>
<td>0.571</td>
<td>0.443</td>
</tr>
<tr>
<td>2.00</td>
<td>3.343</td>
<td>6.055</td>
<td>2.900</td>
<td>1.520</td>
</tr>
<tr>
<td>3.00</td>
<td>4.128</td>
<td>2.911</td>
<td>2.905</td>
<td>1.710</td>
</tr>
<tr>
<td>10.0</td>
<td>3.343</td>
<td>4.413</td>
<td>3.043</td>
<td>1.690</td>
</tr>
</tbody>
</table>

Table 3: Two Spirals data: Number of Kernel evaluations $\times 10^{-6}$

<table>
<thead>
<tr>
<th>C</th>
<th>SMO Random Seed 1</th>
<th>SMO Random Seed 2</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.180</td>
<td>1.012</td>
<td>0.810</td>
<td>0.670</td>
</tr>
<tr>
<td>5.0</td>
<td>1.320</td>
<td>1.165</td>
<td>1.275</td>
<td>1.044</td>
</tr>
<tr>
<td>10.0</td>
<td>1.387</td>
<td>1.624</td>
<td>1.453</td>
<td>1.113</td>
</tr>
<tr>
<td>50.0</td>
<td>3.241</td>
<td>2.584</td>
<td>2.353</td>
<td>1.739</td>
</tr>
<tr>
<td>$10^2$</td>
<td>6.027</td>
<td>5.038</td>
<td>4.578</td>
<td>2.119</td>
</tr>
<tr>
<td>$5 \times 10^2$</td>
<td>20.187</td>
<td>9.970</td>
<td>7.556</td>
<td>4.607</td>
</tr>
<tr>
<td>$10^3$</td>
<td>17.518</td>
<td>16.943</td>
<td>7.321</td>
<td>8.569</td>
</tr>
<tr>
<td>$5 \times 10^3$</td>
<td>62.729</td>
<td>96.136</td>
<td>49.270</td>
<td>38.660</td>
</tr>
<tr>
<td>$10^4$</td>
<td>60.202</td>
<td>68.392</td>
<td>52.000</td>
<td>17.274</td>
</tr>
<tr>
<td>$5 \times 10^4$</td>
<td>34.093</td>
<td>44.377</td>
<td>28.380</td>
<td>26.450</td>
</tr>
</tbody>
</table>

Table 4: Checkers data: Number of Kernel evaluations $\times 10^{-6}$
<table>
<thead>
<tr>
<th>C</th>
<th>SMO Random Seed 1</th>
<th>SMO Random Seed 2</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1.129</td>
<td>1.077</td>
<td>0.324</td>
<td>0.325</td>
</tr>
<tr>
<td>0.20</td>
<td>0.917</td>
<td>1.042</td>
<td>0.569</td>
<td>0.570</td>
</tr>
<tr>
<td>0.40</td>
<td>0.752</td>
<td>0.751</td>
<td>0.546</td>
<td>0.545</td>
</tr>
<tr>
<td>0.50</td>
<td>0.846</td>
<td>0.734</td>
<td>0.543</td>
<td>0.539</td>
</tr>
<tr>
<td>0.70</td>
<td>0.834</td>
<td>0.944</td>
<td>0.545</td>
<td>0.541</td>
</tr>
<tr>
<td>1.00</td>
<td>0.723</td>
<td>0.728</td>
<td>0.547</td>
<td>0.647</td>
</tr>
<tr>
<td>2.00</td>
<td>0.891</td>
<td>0.868</td>
<td>0.630</td>
<td>0.610</td>
</tr>
<tr>
<td>3.00</td>
<td>0.888</td>
<td>0.863</td>
<td>0.727</td>
<td>0.696</td>
</tr>
<tr>
<td>5.00</td>
<td>1.053</td>
<td>1.082</td>
<td>0.845</td>
<td>0.749</td>
</tr>
<tr>
<td>10.0</td>
<td>2.041</td>
<td>2.089</td>
<td>1.428</td>
<td>1.198</td>
</tr>
<tr>
<td>20.0</td>
<td>3.921</td>
<td>3.904</td>
<td>2.463</td>
<td>1.946</td>
</tr>
<tr>
<td>50.0</td>
<td>7.915</td>
<td>8.446</td>
<td>4.740</td>
<td>3.402</td>
</tr>
<tr>
<td>100.0</td>
<td>13.315</td>
<td>12.358</td>
<td>6.543</td>
<td>4.502</td>
</tr>
<tr>
<td>200.0</td>
<td>16.656</td>
<td>19.692</td>
<td>9.382</td>
<td>5.588</td>
</tr>
<tr>
<td>500.0</td>
<td>24.019</td>
<td>25.676</td>
<td>14.715</td>
<td>6.942</td>
</tr>
</tbody>
</table>

Table 5: Adult 1 data: Number of Kernel evaluations $\times 10^{-7}$
<table>
<thead>
<tr>
<th>C</th>
<th>SMO Random Seed 1</th>
<th>SMO Random Seed 2</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>9.812</td>
<td>9.290</td>
<td>6.856</td>
<td>5.819</td>
</tr>
<tr>
<td>0.20</td>
<td>10.074</td>
<td>8.145</td>
<td>4.506</td>
<td>4.499</td>
</tr>
<tr>
<td>0.40</td>
<td>7.739</td>
<td>7.745</td>
<td>4.336</td>
<td>4.330</td>
</tr>
<tr>
<td>0.50</td>
<td>9.472</td>
<td>7.657</td>
<td>5.233</td>
<td>4.341</td>
</tr>
<tr>
<td>0.70</td>
<td>6.706</td>
<td>6.700</td>
<td>4.388</td>
<td>4.352</td>
</tr>
<tr>
<td>1.00</td>
<td>6.715</td>
<td>7.588</td>
<td>4.498</td>
<td>4.467</td>
</tr>
<tr>
<td>2.00</td>
<td>7.163</td>
<td>7.200</td>
<td>5.034</td>
<td>4.923</td>
</tr>
<tr>
<td>3.00</td>
<td>7.901</td>
<td>6.939</td>
<td>5.638</td>
<td>5.446</td>
</tr>
<tr>
<td>5.00</td>
<td>8.980</td>
<td>9.631</td>
<td>7.204</td>
<td>5.880</td>
</tr>
<tr>
<td>10.0</td>
<td>16.431</td>
<td>15.086</td>
<td>11.711</td>
<td>9.310</td>
</tr>
<tr>
<td>20.0</td>
<td>33.564</td>
<td>33.288</td>
<td>20.864</td>
<td>15.386</td>
</tr>
<tr>
<td>50.0</td>
<td>77.886</td>
<td>71.813</td>
<td>42.554</td>
<td>29.409</td>
</tr>
<tr>
<td>100.0</td>
<td>128.383</td>
<td>126.491</td>
<td>66.100</td>
<td>48.257</td>
</tr>
<tr>
<td>200.0</td>
<td>207.332</td>
<td>217.001</td>
<td>112.869</td>
<td>78.402</td>
</tr>
<tr>
<td>500.0</td>
<td>384.589</td>
<td>393.042</td>
<td>216.034</td>
<td>122.202</td>
</tr>
</tbody>
</table>

Table 6: Adult 4 data: Number of Kernel evaluations × 10⁻⁷

<table>
<thead>
<tr>
<th>C</th>
<th>SMO Random Seed 1</th>
<th>SMO Random Seed 2</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>193.680</td>
<td>206.289</td>
<td>31.590</td>
<td>31.590</td>
</tr>
<tr>
<td>0.40</td>
<td>90.758</td>
<td>90.618</td>
<td>45.742</td>
<td>45.720</td>
</tr>
<tr>
<td>1.00</td>
<td>90.091</td>
<td>71.677</td>
<td>47.993</td>
<td>47.490</td>
</tr>
<tr>
<td>5.00</td>
<td>103.471</td>
<td>103.198</td>
<td>77.732</td>
<td>70.700</td>
</tr>
<tr>
<td>20.0</td>
<td>370.405</td>
<td>380.250</td>
<td>224.689</td>
<td>153.932</td>
</tr>
</tbody>
</table>

Table 7: Adult 7 data: Number of Kernel evaluations × 10⁻⁷
<table>
<thead>
<tr>
<th>C</th>
<th>SMO</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.679</td>
<td>0.633</td>
<td>0.632</td>
</tr>
<tr>
<td>0.20</td>
<td>1.197</td>
<td>0.755</td>
<td>0.727</td>
</tr>
<tr>
<td>0.40</td>
<td>1.215</td>
<td>0.893</td>
<td>0.971</td>
</tr>
<tr>
<td>0.50</td>
<td>1.013</td>
<td>0.912</td>
<td>0.918</td>
</tr>
<tr>
<td>0.70</td>
<td>0.963</td>
<td>1.070</td>
<td>1.032</td>
</tr>
<tr>
<td>1.00</td>
<td>1.206</td>
<td>1.063</td>
<td>0.988</td>
</tr>
<tr>
<td>2.00</td>
<td>1.365</td>
<td>1.260</td>
<td>1.142</td>
</tr>
<tr>
<td>3.00</td>
<td>1.449</td>
<td>1.308</td>
<td>1.270</td>
</tr>
<tr>
<td>5.00</td>
<td>1.252</td>
<td>1.178</td>
<td>1.242</td>
</tr>
<tr>
<td>10.0</td>
<td>1.421</td>
<td>1.397</td>
<td>1.348</td>
</tr>
<tr>
<td>20.0</td>
<td>1.570</td>
<td>1.364</td>
<td>1.221</td>
</tr>
<tr>
<td>50.0</td>
<td>1.621</td>
<td>1.373</td>
<td>1.363</td>
</tr>
<tr>
<td>100.0</td>
<td>1.666</td>
<td>1.301</td>
<td>1.250</td>
</tr>
<tr>
<td>200.0</td>
<td>1.336</td>
<td>1.366</td>
<td>1.257</td>
</tr>
<tr>
<td>500.0</td>
<td>1.378</td>
<td>1.442</td>
<td>1.420</td>
</tr>
</tbody>
</table>

Table 8: Web 1 data: Number of Kernel evaluations × 10⁻⁷
<table>
<thead>
<tr>
<th>C</th>
<th>SMO</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>6.714</td>
<td>4.126</td>
<td>3.578</td>
</tr>
<tr>
<td>0.20</td>
<td>7.686</td>
<td>3.937</td>
<td>4.010</td>
</tr>
<tr>
<td>0.40</td>
<td>9.155</td>
<td>4.333</td>
<td>4.972</td>
</tr>
<tr>
<td>0.50</td>
<td>8.808</td>
<td>5.102</td>
<td>5.185</td>
</tr>
<tr>
<td>0.70</td>
<td>9.146</td>
<td>6.199</td>
<td>5.059</td>
</tr>
<tr>
<td>1.00</td>
<td>8.154</td>
<td>6.156</td>
<td>6.061</td>
</tr>
<tr>
<td>2.00</td>
<td>8.494</td>
<td>7.436</td>
<td>6.594</td>
</tr>
<tr>
<td>3.00</td>
<td>9.887</td>
<td>8.331</td>
<td>9.092</td>
</tr>
<tr>
<td>5.00</td>
<td>10.826</td>
<td>8.749</td>
<td>9.464</td>
</tr>
<tr>
<td>10.0</td>
<td>9.685</td>
<td>11.193</td>
<td>12.402</td>
</tr>
<tr>
<td>20.0</td>
<td>12.162</td>
<td>9.795</td>
<td>10.265</td>
</tr>
<tr>
<td>50.0</td>
<td>10.733</td>
<td>10.973</td>
<td>10.305</td>
</tr>
<tr>
<td>100.0</td>
<td>12.155</td>
<td>11.314</td>
<td>11.821</td>
</tr>
<tr>
<td>200.0</td>
<td>12.169</td>
<td>10.177</td>
<td>10.907</td>
</tr>
<tr>
<td>500.0</td>
<td>12.792</td>
<td>11.121</td>
<td>10.661</td>
</tr>
</tbody>
</table>

Table 9: Web 4 data: Number of Kernel evaluations \( \times 10^{-7} \)
<table>
<thead>
<tr>
<th>C</th>
<th>SMO</th>
<th>SMO Modification 1</th>
<th>SMO Modification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>134.575</td>
<td>33.218</td>
<td>38.725</td>
</tr>
<tr>
<td>0.20</td>
<td>143.039</td>
<td>40.536</td>
<td>46.877</td>
</tr>
<tr>
<td>0.40</td>
<td>132.594</td>
<td>48.952</td>
<td>40.187</td>
</tr>
<tr>
<td>0.50</td>
<td>115.698</td>
<td>50.801</td>
<td>45.382</td>
</tr>
<tr>
<td>0.70</td>
<td>106.148</td>
<td>42.707</td>
<td>53.282</td>
</tr>
<tr>
<td>1.00</td>
<td>120.296</td>
<td>49.265</td>
<td>48.310</td>
</tr>
<tr>
<td>2.00</td>
<td>146.941</td>
<td>56.402</td>
<td>64.235</td>
</tr>
<tr>
<td>3.00</td>
<td>112.226</td>
<td>58.735</td>
<td>69.329</td>
</tr>
<tr>
<td>5.00</td>
<td>115.890</td>
<td>82.549</td>
<td>69.308</td>
</tr>
<tr>
<td>10.0</td>
<td>113.551</td>
<td>85.744</td>
<td>86.436</td>
</tr>
<tr>
<td>20.0</td>
<td>103.516</td>
<td>95.809</td>
<td>93.830</td>
</tr>
<tr>
<td>50.0</td>
<td>129.473</td>
<td>93.215</td>
<td>89.486</td>
</tr>
<tr>
<td>100.0</td>
<td>136.820</td>
<td>91.090</td>
<td>110.006</td>
</tr>
<tr>
<td>200.0</td>
<td>148.265</td>
<td>93.362</td>
<td>94.349</td>
</tr>
<tr>
<td>500.0</td>
<td>125.315</td>
<td>94.553</td>
<td>105.505</td>
</tr>
</tbody>
</table>

Table 10: Web 7 data: Number of Kernel evaluations × 10^{-7}


Appendix. Pseudo-Codes for Modified SMO Algorithms.

The pseudo-codes for the improved SMO algorithms are presented below. Here, statements starting with “%” denote comments.

target = desired output vector
point = training point matrix
fcache = cache vector for Fi values

% Note: Our definition of Fi is different from the Ei in Platt’s SMO algorithm. Our Fi does not subtract any threshold.

procedure takeStep(i1,i2)
% Much of this procedure is same as that in Platt’s SMO pseudo-code.
    if (i1 == i2) return 0
    alph1 = Lagrange multiplier for i1
    y1 = target[i1]
    F1 = fcache[i1]
    s = y1*y2
    Compute L, H
    if (L == H)
        return 0
    k11 = kernel(point[i1],point[i1])
    k12 = kernel(point[i1],point[i2])
    k22 = kernel(point[i2],point[i2])
    eta = 2*k12-k11-k22
    if (eta < 0)
{ 
    a2 = alph2 - y2*(F1-F2)/eta 
    if (a2 < L) a2 = L 
    else if (a2 > H) a2 = H 
} 
else 
{
    Lobj = objective function at a2=L 
    Hobj = objective function at a2=H 
    if (Lobj > Hobj+eps) 
        a2 = L 
    else if (Lobj < Hobj-eps) 
        a2 = H 
    else 
        a2 = alph2 
} 
if (|a2-alph2| < eps*(a2+alph2+eps)) 
    return 0 
    a1 = alph1+s*(alph2-a2) 
Update weight vector to reflect change in a1 & a2, if linear SVM 
Update fcache[i] for i in I_0 using new Lagrange multipliers 
Store a1 and a2 in the alpha array 
% The update below is simply achieved by keeping and updating information 
% about alpha_i being at 0, C or in between them. Using this together with 
% target[i] gives information as to which index set i belongs. 
    Update I_0, I_1, I_2, I_3 and I_4 
% Compute updated F values for i1 and i2... 
    fcache[i1] = F1 + y1*(a1-alph1)*k11 + y2*(a2+alph2)*k12 
    fcache[i2] = F2 + y1*(a1-alph1)*k12 + y2*(a2+alph2)*k22 
Compute (i_low, b_low) and (i_up, b_up) by applying equations (11a) and
(11b), using only i1, i2 and indices in I_0; see item 3 of section 5.
return 1
endprocedure

procedure examineExample(i2)
y2 = target[i2]
alph2 = Lagrange multiplier for i2
if (i2 is in I_0)
{
    F2 = fcache[i2]
}
else
{
    compute F2 = F_i2 and set fcache[i2] = F2
    % Update (b_low, i_low) or (b_up,i_up) using (F2,i2)...
    if ((i2 is in I_1 or I_2) && (F2 < b_up) )
        b_up = F2, i_up = i2
    else if ((i2 is in I_3 or I_4) && (F2 > b_low) )
        b_low = F2, i_low = i2
}
% Check optimality using current b_low and b_up and, if
% violated, find an index i1 to do joint optimization with i2...
    optimality = 1
    if (i2 is in I_0, I_1 or I_2)
    {
        if (b_low-F2 > 2*tol)
            optimality = 0, i1 = i_low
    }
    if (i2 is in I_0, I_3 or I_4)
    {

23
if (F2-b_up > 2*tol)
    optimality = 0, i1 = i_up
}
if (optimality == 1)
    return 0

% For i2 in I_0 choose the better i1...
if (i2 is in I_0)
{
    if (b_low-F2 > F2-b_up)
        i1 = i_low
    else
        i1 = i_up
}
if takeStep(i1,i2)
    return 1
else
    return 0
endprocedure

main routine for Modification 1:
initialize alpha array to all zero
initialize b_up = -1, i_up to any one index of class 1
initialize b_low = 1, i_low to any one index of class 2
set fcache[i_low] = 1 and fcache[i_up] = -1
numChanged = 0;
examineAll = 1;
while (numChanged > 0 | examineAll)
{
    numChanged = 0;
    if (examineAll)
{  
loop I over all training examples  
   numChanged += examineExample(I)  
}  
else  
{  
   loop I over I_0  
   numChanged += examineExample(I)  
}% It is easy to check if optimality on I_0 is attained...  
   if (bup > blow - 2*tol) at any I  
      exit the loop after setting numChanged = 0  
}  
if (examineAll == 1)  
      examineAll = 0  
else if (numChanged == 0)  
      examineAll = 1  
}

main routine for Modification 2:  
initialize alpha array to all zero  
initialize b_up = -1, i_up to any one index of class 1  
initialize b_low = 1, i_low to any one index of class 2  
set fcach[low] = 1 and fcach[i_up] = -1  
numChanged = 0;  
examineAll = 1;  
while (numChanged > 0 | examineAll)  
{  
   numChanged = 0;  
   if (examineAll)
{  
    loop I over all training examples 
    numChanged += examineExample(I) 
  }

else
% The following loop is the only difference between the two SMO
% modifications. Whereas, in modification 1, the inner loop selects
% i2 from I_0 sequentially, here i2 is always set to the current
% i_low and i1 is set to the current i_up; clearly, this corresponds
% to choosing the worst violating pair using members of I_0 and some
% other indices.
{
  inner_loop_success = 1;
  do until ( (bup > blow - 2*tol) | inner_loop_success = 0 )
  {
    i2 = i_low  
    y2 = target(i2)
    alph2 = Lagrange multiplier for i2
    F2 = fcache[i2]
    i1 = i_up  
    inner_loop_success = takeStep(i_up,i_low)
    numChanged += inner_loop_success
  }
  numChanged = 0
}
if (examineAll == 1)
  examineAll = 0
else if (numChanged == 0)
  examineAll = 1
}