Support Vector Machine via Nonlinear Rescaling Method

Roman Polyak
Department of SEOR and Department of Mathematical Sciences
George Mason University
4400 University Dr. Fairfax, VA 22030
rpolyak@gmu.edu

Shen-Shyang Ho
Department of Computer Science
George Mason University
4400 University Dr. Fairfax, VA 22030
sho@gmu.edu

Igor Griva
Department of Mathematical Sciences
and
Department of Computational and Data Sciences
George Mason University
4400 University Dr. Fairfax, VA 22030
igriva@gmu.edu

Received: date / Accepted: date

Abstract

In this paper we construct the linear support vector machine (SVM) based on the nonlinear rescaling (NR) methodology (see [9, 11, 10] and references therein). The formulation of the linear SVM based on the NR method leads to an algorithm which reduces the number of support vectors without compromising the classification performance compared to the linear soft-margin SVM formulation. The NR algorithm computes both the primal and the dual approximation at each step. The dual variables associated with the given data-set provide important information about each data point and play the key role in selecting the set of support vectors. Experimental results on ten benchmark classification problems show that the NR formulation is feasible. The quality of discrimination, in most instances, is comparable to the linear soft-margin SVM while the number of support vectors in several instances were substantially reduced.

Keywords: Support Vector Machine; Convex Optimization; Duality ; Lagrange Multipliers

1 Introduction

In the past decade, the Support Vector Machine (SVM) [13] was among the most widely used tools in statistical learning. Both the primal and dual SVM formulations lead to solving quadratic programming (QP) problems in order to find a separating hyperplane. The normal vector of the separating hyperplane w is conveniently represented as a linear combination of the support vectors [13]. The reduction of the support vectors in this representation leads to the reduction of classification time and therefore highly desirable [2, 8, 14]. In case when the given binary-class data-set can be separated, solving a QP problem finds a hyperplane (hyper-surface) that separates the two classes with maximum margin [1]. In the non-separable case, there is a trade-off between the margin size and the number of data points in the data-set which cannot be separated. In the case of the standard soft-margin SVM formulation, the Lagrange multipliers (dual variables) corresponding to the data points inside the margin have a fixed value which is equal to the penalty parameter [4]. All the data points within the margin are support vectors. They all have the same Lagrange multipliers in the representation of the normal vector of the separating hyperplane equal to the a priori given penalty parameter. It leads to a large number of support vectors and thus compromises the sparsity of the representation of w. Several methods have been suggested to reduce the number of support vectors [2, 8, 14]. However, these approaches are still based on the soft-margin SVM and therefore include the upper bounds on the dual variables in the formulations, a restrictive factor for finding a sparse representation of w.

The main contribution of this paper is the nonlinear rescaling (NR) formulation of the SVM that substantially reduced the number of support vectors without compromising the quality of discrimination. Moreover, this formulation does not require a pre-defined penalty parameter, which is a critical factor in the soft-margin SVM formulation. When the classification problem is separable, the solution from the NR formulation is identical to the optimal margin SVM.

The distinct characteristic of the NR theory [9, 11] is the use of the Lagrange multipliers as the main driving force which insures the convergence of NR methods for solving constrained optimization problems. The positive scaling parameter can be either fixed or increased from step to step. By increasing the scaling parameter, one can improve the rate of convergence. The fundamental differences between NR methods and the interior point methods [7] is that the NR methods do not require finding an interior starting point and they do not keep the primal sequence inside the feasible set. Moreover, the NR methods are exterior point methods by nature in which the Lagrange multipliers carry important information throughout the computational process.

In the NR formulation of SVM, the Lagrange multipliers characterize the "cost" of the "non-separability". The "large" Lagrange multipliers that stand

out among all the Lagrange multipliers correspond to the data points that are "most certain" on the "wrong side" of the separating hyperplane. In fact, sometimes one can consider such data points as "noise" which have to be eliminated from the input data [5]. On the other hand, the "small" Lagrange multipliers identify the data points which has practically no impact on the separating hyperplane. They can also be eliminated. To this end, the Lagrange multipliers enable us to identify the data points which are critical in defining the discrimination rule and at the same time to reduce substantially the number of support vectors. The experimental results show that in most cases the SVM based on the NR method (NR-SVM) reduces the number of support vector substantially without compromising the quality of discrimination compared to the linear soft-margin SVM.

The paper is organized as follows. In the next section, we review the basic SVM problem and show that due to the problem formulation the Lagrange multipliers for all the data points which cannot be separated have the same value equal to the a priori chosen penalty parameter. In Section 3, we describe the NR method and review the basic convergence results. In Section 4 we introduce the SVM formulation based on NR theory and applied the NR method to solve the SVM problem. In Section 5, we compare the performance of NR-SVM with the linear soft-margin SVM on ten benchmark problems.

2 Background

For a given set of labeled data points

$$\{(a_1, y_1), \cdots, (a_n, y_n)\}$$

where $y_i \in \{-1, 1\}$ and $a_i \in \Re^m$, the soft-margin SVM problem [13] consists of finding the triple $(w^*, \xi^*, b^*) \in \Re^m \times \Re^n \times \Re$ that minimizes

$$u = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \tag{1}$$

subject to the constraints

$$y_i(w \cdot a_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad i = 1, 2, \cdots, n \tag{2}$$

where w is the normal vector for the "separating" hyperplane, $(w, x) + b = 0$, the vector $\xi = (\xi_1, \cdots, \xi_n)$ defines the constraints violation and $C > 0$ is an empirically defined penalty parameter, which is used to penalize the constraint violations.

Very often instead of (1)–(2) the dual problem [13] is used. The dual QP consists of maximizing

$$v = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (a_i \cdot a_j) + \sum_{i=1}^{n} \alpha_i \tag{3}$$

subject to

$$\sum_{i=1}^{n} y_i \alpha_i = 0, \tag{4}$$

$$0 \leq \alpha_i \leq C, \quad i = 1, \cdots, n. \tag{5}$$

From the primal solution (w^*, ξ^*, b^*) and the dual solution $\alpha^* = (\alpha_1^*, \cdots, \alpha_n^*)$, we have

$$u^* = v^*, \tag{6}$$

and the following complementarity conditions are satisfied

$$\begin{aligned} \xi_i^* > 0 \quad &\Rightarrow \quad \alpha_i^* = C, \\ \xi_i^* = 0 \quad &\Leftarrow \quad 0 \leq \alpha_i^* < C, \end{aligned} \tag{7}$$

and

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i a_i. \tag{8}$$

For the data points on the margin, the corresponding components in the dual vector $\alpha^* = (\alpha_1^*, \cdots, \alpha_n^*)$ are between 0 and C. The dual values of the data points outside the margin are zero. It follows from (7) that all non-separable (i.e. within the margin) data points have the same dual value, which is equal to the a priori chosen penalty parameter C. In other words, all Lagrange multipliers in the representation (8) which correspond to the data points within the margin have the same C value.

In the next section, we describe the general NR methods and in section 4, we specify the NR approach for the SVM which does not require an a priori chosen penalty parameter C. We would like to emphasize that the NR method converges for any fixed scaling parameter $k > 0$ due to the Lagrange multipliers update [6, 11]. Therefore there is no need to predefine the penalty parameter. The Lagrange multipliers characterize the "cost" of the constraint violation. At each step the Lagrange multipliers provide extra information about the non-separability of the data points and at the same time they indicate the data points that do not affect the discrimination rule and can be eliminated.

3 Nonlinear Rescaling Method

Let $-\infty < t_0 < 0 < t_1 < \infty$. We consider a class Ψ of twice continuously differentiable functions $\psi : (t_0, t_1) \to \Re$, which satisfy the following properties:

1. $\psi(0) = 0, \qquad \psi'(0) = 1;$
2. $\psi'(t) > 0;$

3. $\psi''(t) < 0$.

The function $\psi \in \Psi$ is used to transform the constraints of a given constrained optimization problem into an equivalent set of constraints.

Let $f : \Re^m \to \Re$ be convex, and $c_i : \Re^m \to \Re, i = 1, \cdots, n$ be concave functions. We consider the following convex optimization problem

$$x^* \in X^* = Arg\min\{f(x)|x \in \Omega\} \tag{9}$$

where $\Omega = \{x \in \Re^m : c_i(x) \geq 0, i = 1, \cdots, n\}$.

It follows from properties 1.–3. that for any given scaling parameter $k > 0$, we have

$$\Omega = \{x : k^{-1}\psi(kc_i(x)) \geq 0, i = 1, \cdots, n\}$$

Therefore, for any $k > 0$, the following problem

$$\begin{aligned} & x^* \in X^* \\ & = Arg\min\{f(x)|k^{-1}\psi(kc_i(x)) \geq 0, i = 1, \cdots, n\} \end{aligned} \tag{10}$$

is equivalent to the original convex optimization problem (9).

The classical Lagrangian $\mathcal{L} : \Re^m \times \Re^n_+ \times \Re_{++} \longrightarrow \Re$

$$\mathcal{L}(x, \lambda, k) = f(x) - k^{-1}\sum_{i=1}^{n} \lambda_i \psi(kc_i(x)), \tag{11}$$

which corresponds to problem (10) is the main tool in developing NR methods for solving the constrained optimization problem.

In our experiments, we use the shifted logarithmic barrier function $\psi(t) = \ln(t + 1)$, which leads to the modified barrier functions theory and methods [9]. Each step of the NR method alternates finding an unconstrained minimizer of $\mathcal{L}(x, \lambda, k)$ in $\Re^m$ and the Lagrange multipliers update. The scaling parameter can be fixed or one can change k at each iteration. We consider the version of the NR method with a fixed scaling parameter.

Let $\lambda^0 \in \Re^n_{++}$ be the initial Lagrange multiplier vector and the positive scaling parameter k is fixed. Let us assume that the primal-dual pair $(x^s, \lambda^s) \in \Re^m \times \Re^n_{++}$ has been found already. One step of NR method consists of finding:

$$\begin{aligned} & x^{s+1} : \nabla_x \mathcal{L}(x^{s+1}, \lambda^s, k) \\ = \quad & \nabla f(x^{s+1}) - \sum_{i=1}^{n} \psi'(kc_i(x^{s+1}))\lambda_i^s \nabla c_i(x^{s+1}) \\ = \quad & 0 \end{aligned} \tag{12}$$

and updating the Lagrange multipliers by the formula:

$$\lambda_i^{s+1} = \psi'(kc_i(x^{s+1}))\lambda_i^s, \qquad i = 1, \cdots, n. \tag{13}$$

From (12)–(13), we have

$$\nabla_x \mathcal{L}(x^{s+1}, \lambda^s, k) = \nabla_x L(x^{s+1}, \lambda^{s+1}) = 0 \tag{14}$$

where $L(x, \lambda) = f(x) - \sum \lambda_i c_i(x)$ is the classical Lagrangian for the original problem (9).

Therefore,

$$x^{s+1} = \arg\min\{L(x, \lambda^{s+1}) | x \in \Re^m\}$$

and

$$d(\lambda^{s+1}) = L(x^{s+1}, \lambda^{s+1})$$

where $d(\lambda) = \inf_{x \in \Re^m} L(x, \lambda)$ is the dual function.

The NR method (12)–(13) solves simultaneously the primal problem (9) and the following dual problem

$$d(\lambda^*) = \arg\max\{d(\lambda) | \lambda \in \Re^n_+\}. \tag{15}$$

The following theorems establish the convergence properties of the NR method (12)–(13).

Theorem 1 *[9] If the standard second order optimality conditions are satisfied and f, $c_i, i = 1, \cdots, n$ are smooth enough then there is $k_0 > 0$ large enough that for any $k \geq k_0$, the following bounds hold*

$$\begin{aligned} &a) \quad ||x^{s+1} - x^*|| \leq ck^{-1}||\lambda^s - \lambda^*|| \\ &b) \quad ||\lambda^{s+1} - \lambda^*|| \leq ck^{-1}||\lambda^s - \lambda^*|| \end{aligned} \tag{16}$$

and the constant $c > 0$ is independent of k.

Theorem 2 *[11] If (10) is a convex programming problem, Slater's conditions are satisfied and X^* is a bounded set, then for any $k > 0$ the NR method (12)–(13) generates the primal-dual sequence $\{x^s, \lambda^s\}$ such that:*

1. $\lim_{s \to \infty} \lambda^s = \lambda^*$,
2. $\lim_{s \to \infty} f(x^s) = \lim_{s \to \infty} d(\lambda^s) = f(x^*) = d(\lambda^*)$,
3. *for any converging subsequence $\{x^{s_e}\}$,*

$$\lim_{s_e \to \infty} x^{s_e} = x^* \in X^*.$$

The NR method (12)–(13) requires finding an unconstrained minimizer x^{s+1} of $\mathcal{L}(x, \lambda^s, k)$ at each step which is generally speaking an infinite procedure. To make the NR method (12)–(13) practical we replace the minimizer x^{s+1} by its approximation $\bar{x}^{s+1}$, which one can find using the stopping criterion introduced in [9]. The approximation $\bar{x}^{s+1}$ can be found in finite number of Newton's steps applied for minimization of $\mathcal{L}(x, \lambda^s, k)$ in x. Replacing x^{s+1} by $\bar{x}^{s+1}$ does not compromise both the convergence and the rate of convergence of the NR method.

In the next section, we introduce the NR formulation for the SVM and describe the NR method for solving the SVM problem.

4 Nonlinear Rescaling SVM (NR-SVM) Formulation

For a given set of labeled data points $\{(a_i, y_i) \in \Re^{m+1}\}, i \in I = \{1, \cdots, n\}, y_i \in \{-1, 1\}$, to construct a SVM means to find a hyperplane $h = h(w, b) = \{x : (w, x) - b = 0\}$ such that the sets $I_+ = \{i : (a_i, 1)\}$ and $I_- = \{i : (a_i, -1)\}$ will be separated with a maximum margin.

For every $i \in I_+$ in the "positive" halfspace, we consider the distance $d(a_i, h) = (w, a_i) - b \geq 0$ from $a_i, i \in I_+$ to the hyperplane h and for every $i \in I_-$ in the "negative" halfspace, we consider the distance $d(a_i, h) = -(w, a_i) + b \geq 0, i \in I_-$. To find the hyperplane h, which separates set I_+ from I_- with maximum margin, one has to solve the following problem:

$$\Delta^* = \max_{||w||^2=1, b \in \Re} \min_{i \in I} d(a_i, h)$$

By introducing $\Delta = \min_{i \in I} d(a_i, h)$, one can rewrite the problem of finding Δ^* as follows:

$$\Delta \to \max \tag{17}$$

subject to

$$c_i(x) \equiv c_i(w, b, \Delta) = (w, a_i) - b - \Delta \geq 0, i \in I_+ \tag{18}$$

$$c_i(x) \equiv c_i(w, b, \Delta) = -(w, a_i) + b - \Delta \geq 0, i \in I_-, \tag{19}$$

$$||w||^2 = 1 \tag{20}$$

where I_+ and I_- consist of positively and negatively labeled data points respectively.

To describe the NR method for solving the problem (17)–(20), we consider an equivalent problem. For any given positive parameters $k > 0, \tau > 0$ and a transformation $\psi \in \Psi$, the following problem:

$$-\tau\Delta \to \min \tag{21}$$

subject to

$$k^{-1}\psi(\cdot) = k^{-1}\psi(kc_i(x)) \geq 0, \quad i \in I_+ \tag{22}$$

$$k^{-1}\psi(\cdot) = k^{-1}\psi(kc_i(x)) \geq 0, \quad i \in I_- \tag{23}$$

$$\frac{1}{2}\left(||w||^2 - 1\right) = 0 \tag{24}$$

is equivalent to (17)–(20).

The classical Lagrangian

$$\begin{aligned} \mathcal{L}(\cdot) &= \mathcal{L}(w, \Delta, \lambda, \gamma, \tau) \\ &= -\tau\Delta - k^{-1}\Sigma_{i \in I_+}\lambda_i\psi(kc_i(x)) \\ &\quad -k^{-1}\Sigma_{i \in I_-}\lambda_i\psi(kc_i(x)) + \gamma\frac{1}{2}\left(||w||^2 - 1\right) \end{aligned} \tag{25}$$

for the problem (21)–(24) is our basic tool. We use the Lagrangian $\mathcal{L}(\cdot)$ to describe the NR-SVM.

The NR method for solving the problem (21)–(24) consists of finding the minimum of the Lagrangian (25) for the equivalent problem in $x = (w, b, \Delta)$, and then update the Lagrange multipliers $\lambda = (\lambda_1, \cdots, \lambda_n)$ and τ. The scaling parameter k can be fixed or updated at any iteration. Let $\epsilon > 0$ be small enough. We describe one step of the NR method for solving (21)–(24) given a fixed positive scaling parameter k.

1. Find

$$\widehat{x} = \arg\min\{\mathcal{L}(x, \lambda, \gamma, \tau, k) | x \in \Re^{m+2}\} \tag{26}$$

which is equivalent to solving the following system of equations:

$$\nabla_w \mathcal{L}(\cdot) = -\Sigma_{i \in I_+} \lambda_i \psi'(\cdot) a_i + \Sigma_{i \in I_-} \lambda_i \psi'(\cdot) a_i + \gamma w = 0 \tag{27}$$

$$\nabla_\Delta \mathcal{L}(\cdot) = -\tau + \Sigma_{i \in I_+} \lambda_i \psi'(\cdot) + \Sigma_{i \in I_-} \lambda_i \psi'(\cdot) = 0 \tag{28}$$

$$\nabla_b \mathcal{L}(\cdot) = \Sigma_{i \in I_+} \lambda_i \psi'(\cdot) - \Sigma_{i \in I_-} \lambda_i \psi'(\cdot) = 0 \tag{29}$$

2. Update the Lagrange multipliers by the formula:

$$\widehat{\lambda}_i = \lambda_i \psi'(\cdot), i \in I_+ \bigcup I_- \tag{30}$$

3. Find $\widehat{\gamma}$ from $||\widehat{w}||^2 = 1$ where

$$\widehat{w} = \gamma^{-1} \left(\Sigma_{i \in I_+} \widehat{\lambda}_i a_i - \Sigma_{i \in I_-} \widehat{\lambda}_i a_i \right) \tag{31}$$

4. Compute

$$\widehat{\tau} \quad = \quad \Sigma_{i \in I_+} \widehat{\lambda}_i + \Sigma_{i \in I_-} \widehat{\lambda}_i \tag{32}$$

5. Set

$$\widehat{\lambda} \quad := \quad (\widehat{\lambda}_i \widehat{\tau}^{-1}, i = 1, \cdots, n) \tag{33}$$

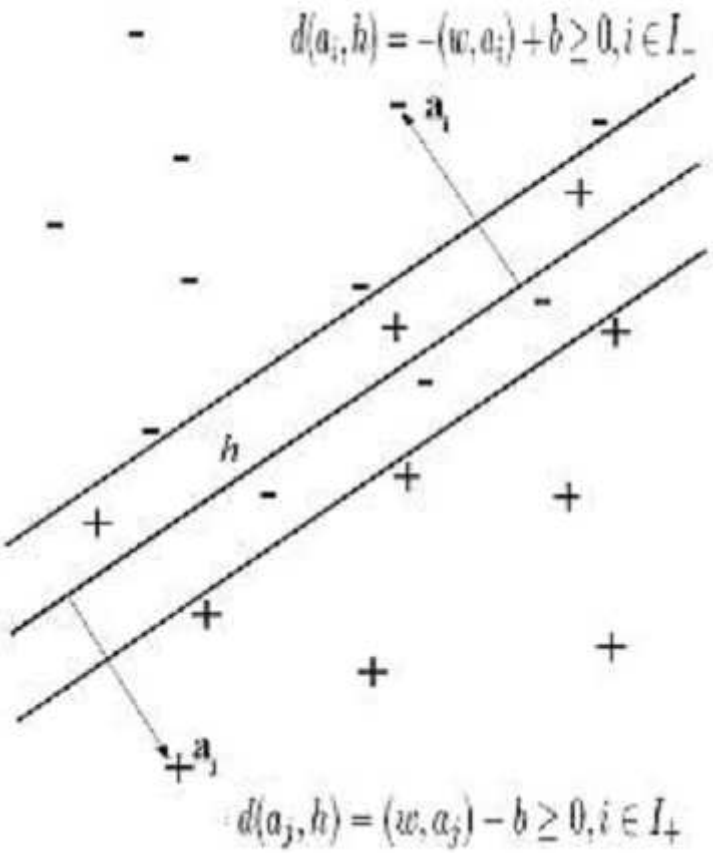
6. If $||\lambda - \widehat{\lambda}|| > \epsilon$, then set $(x, \lambda, \gamma, \tau) := (\widehat{x}, \widehat{\lambda}, \widehat{\gamma}, \widehat{\tau})$ and go to step 1.
   Else $x^* = x, \lambda^* = \lambda$.

We use the Lagrange multipliers $\lambda^* \in \Re^n$ to select the support vectors by eliminating vector a_i when $0 < \lambda_i \leq \epsilon$. The NR method (26)–(33) is the basis for the NR-SVM algorithm.

We would like to point out that if the training set I is such that the subsets $(a_i, y_i), i \in I_-$ and $(a_i, y_i), i \in I_+$ can be separated, then it follows from the formulation (17)–(20) that $\Delta^* > 0$ and the maximal margin is $2\Delta^*$. The discriminating rule is identical to the classical SVM.

If the subsets $(a_i, y_i), i \in I_-$ and $(a_i, y_i), i \in I_+$ cannot be separated, then $\Delta^* < 0$.

In such case the classical SVM allows finding a hyperplane, which partially "separates" the sets I_+ and I_- (see Figure **??**) and all vectors within the margin will have the same Lagrange multipliers $\lambda_i = C > 0$. The NR-SVM formulation provides a particular Lagrange multiplier for each vector. Moreover, $\lim_{s \to \infty} \lambda^s = \lambda^*$. Therefore, the Lagrange multipliers $\lambda_i^s \to 0$ can be eliminated.


$d(a_i, h) = -(w, a_i) + b \geq 0, i \in I_-$
a_i
h
a_j
$d(a_j, h) = (w, a_j) - b \geq 0, i \in I_+$

Dataset	Size	Dimension	NR-SVM		Linear C-SVM		
			SV	Error	C	SV	Error
Banana	400	2	4.19	0.44	10	365.77	0.45
Breast-Cancer	200	9	44.70	0.31	1	121.26	0.29
Diabetis	468	8	29.04	0.26	1000	241.49	0.23
German	700	20	102.00	0.29	100	378.62	0.24
Heart	170	13	17.64	0.20	1	61.87	0.17
Ringnorm	400	20	37.08	0.29	1	221.86	0.25
Splice	1000	60	339.45	0.17	1	347.80	0.16
Thyroid	140	5	24.08	0.15	100	33.99	0.10
Titanic	150	3	35.64	0.24	1	68.43	0.23
Waveform	400	21	34.00	0.15	1	110.57	0.13

Table 1: Comparison of NR-SVM and Linear C-SVM on ten benchmark problems. (SV (Number of support vectors) and Error (Number of testing examples wrongly classified/Number of testing examples) are averaged over 100 trials.)

5 Experimental Results

We perform experiments to compare the NR-SVM and the linear soft-margin SVM in terms of the reduction in the number of support vectors and the classification error. We used the Matlab interface of LIBSVM 2.81 [3] for the SVM implementation (C-SVM) with C taking the values: 1, 10, 100, and 1000, without kernels, i.e. linear SVM. We compare NR-SVM performance with the best performance of linear C-SVM using the various C values. We use ten binary classification problems from [12] to evaluate and compare the performance of our NR-SVM with the linear C-SVM. For each benchmark problem, there are 100 realizations each.

The experimental results are shown in Table 1. The quality of discrimination (based on the test error rate), in most instances, is comparable to the linear C-SVM while the number of support vectors in several instances were substantially reduced. Similar to our experimental results, it has been observed in [8, 14] that a reduction in the number of support vectors increases the test error rate slightly.

6 Conclusions

In this paper we construct the linear support vector machine (SVM) based on the NR methodology. The formulation of the NR-SVM leads to the algorithm which reduces the number of support vectors without compromising the classification performance compared to the linear soft-margin SVM formulation. In particular, the NR-SVM does not require a predefined penalty parameter. One notes that vectors which have very small Lagrange multipliers, can be removed at each NR step to improve the computational efficiency. Moreover, when a vector point has a much higher Lagrange multiplier, one can suspect that either the vector point is "noise" or it is wrongly labeled.

The distinct characteristics of the NR method is the ability to associate with each vector point a Lagrange multiplier, which measures the "non-separability" of this vector point. It allows the use of the SVM approach for medical diagnostic and drug discovery purposes. In particular, when it comes to medical diagnostic we use the given vector points together with the vector point that represents a new medical case. One assigns a positive label to the new vector point and solves the NR-SVM. Then one solves the NR-SVM again when the vector point is assigned with a negative label. In the case when the Lagrange multipliers of this new vector point are substantially different for the two labels, for example "very small" value for positive label and "very large" value for negative label, then we have a double conformation that the medical case under consideration belongs to the positive set.

References

[1] Boser, B. E., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *COLT* (pp. 144–152).

[2] Burges, C. J. C. (1996). Simplified support vector decision rules. *ICML* (pp. 71–77).

[3] Chang, C.-C., & Lin, C.-J. (2001). Libsvm: a library for support vector machines.

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

[5] Guyon, I., Matic, N., & Vapnik, V. (1996). Discovering informative patterns and data cleaning. In *Advances in knowledge discovery and data mining*, 181–203.

[6] Jensen, D., & Polyak, R. (1994). The convergence of a modified barrier method for convex programming. *IBM Journal of Research and Development*, *38*, 307–321.

[7] Nesterov, Y., & Nemirovskii, A. (1994). *Interior-point polynomial algorithms in convex programming.* Philadelphia : Society for Industrial and Applied Mathematics.

[8] Nguyen, D., & Ho, T. B. (2005). An efficient method for simplifying support vector machine. *ICML* (pp. 617–624).

[9] Polyak, R. (1992). Modified barrier functions (theory and methods). *Math. Program.*, *54*, 177–222.

[10] Polyak, R. (2002). Nonlinear rescaling vs smoothing technique in convex optimization. *Math. Program. Ser. A*, *92*, 197–235.

[11] Polyak, R., & Teboulle, M. (1997). Nonlinear rescaling and proximal-like methods in convex optimization. *Math. Program.*, *76*, 265–284.

[12] Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for adaboost. *Machine Learning*, *42*, 287–320.

[13] Vapnik, V. N. (2000) *The nature of statistical learning theory*, Springer. 2nd edition.

[14] Wu, M., Scholkopf, B., & Bakir, G. (2006). A Direct Method for Building Sparse Kernel Learning Algorithms, *Journal of Machine Learning Research*, Vol. 7, 603–624.