CLASSIFICATION USING KERNEL SVM

PARASCHIV-MUNTEANU, Iuliana
University of Bucharest, Romania
Faculty of Mathematics and Computer Science
pmiulia@fmi.unibuc.ro

STATE, Luminiţa
University of Piteşti, Romania
Faculty of Mathematics and Computer Science
lstate@clicknet.ro

Abstract

The paper aims to present the results of the research toward improving the performance expressed in accuracy and time complexity of SVM implementations. The implementation of the search process for soft margin hyperplane uses a slight modification of Sequential Minimal Optimization (SMO) algorithm introduced by Platt in 1998-1999, to solve the quadratic programming problem involved in the learning process. The SMO is a simple algorithm that quickly solves the SVM problem by decomposing the overall quadratic programming problem into smaller quadratic programming sub-problems without any extra matrix storage and without invoking an iterative numerical routine for each sub-problem. The proposed method was tested on simulated data generated randomly from normal 2-dimensional distributions.

Keywords: *Support Vector Machine, classification, kernel functions*

ACM Classification: I.5.2, I.5.3.

AMS Classification: 68T10.

1. Introduction

The Support Vector Machines (SVM) is a pattern classification technique introduced by Vladimir Vapnik ([9]). Basically SVM is a parametric method, where the parameters are given by the solution of a constrained Quadratic Programming (QP) problem with linear inequality and equality constraints, in a number of variables equal to the size of learning data.

In this paper we present the results of the research aiming to improve the performance of SVM implementation expressed, in accuracy and time complexity. In our developments we used the 'kernel trick', that is we used

kernels that allow to increase the dimension of the input space without increasing the computational complexity hoping that in the higher dimensional space (feature space) the initially nonlinearly separable problem becomes an almost linearly separable one. The implementation of the search process for soft margin hyperplane uses a slight modification of Sequential Minimal Optimization (SMO) algorithm introduced by Platt ([7]) in 1998-1999, for solving the resulted QP problem.

2. **Overview of support vector machines**

Let us assume that the available data $\mathcal{S}$ is represented by examples coming from two classes such that for each example the provenance class is known,

$$\mathcal{S}=\left\{ (x_i, y_i) \mid x_i = \left(x_i^{(1)}, \ldots, x_i^{(d)}\right)^T \in \mathbf{R}^d \,,\ y_i \in \{-1, 1\} \,,\ i = \overline{1, N} \right\} . \quad (1)$$

We denote by f a discriminant (classification) function $f : \mathbf{R}^d \longrightarrow \mathbf{R}$ and consider the decision rule

$$d_f(x) = \begin{cases} 1 & , \text{ if } f(x) > 0 \\ -1 & , \text{ if } f(x) < 0 \\ 0 & , \text{ if } f(x) = 0 \end{cases}$$

where $d_f(x) = 0$ means that the example x can not be classified.

The optimal separating hyperplane

We say that the data is linearly separable if there exists a linear discriminant function $f : \mathbf{R}^d \longrightarrow \mathbf{R}$, $f(x) = w^T x + b$, such that for any $(x_i, y_i) \in \mathcal{S}$, $y_i f(x_i) > 0$ holds.

In a SVM-based approach, the search for a solution (w, b) is usually formulated as a constraint optimization problem

$$\begin{cases} \min\limits_{w \in \mathbf{R}^d,\, b \in \mathbf{R}} \left(\dfrac{1}{2} \|w\|^2 \right) \\ \\ y_i \left(w^T x_i + b \right) \geq 1 \,, \quad i = \overline{1, N} \,. \end{cases} \quad (2)$$

Using the Lagrange multipliers method ([1]), (2) reduces to the QP problem

$$\begin{cases} \max\limits_{(\alpha_1, \ldots, \alpha_N) \in \mathbf{R}^N} \left(-\dfrac{1}{2} \displaystyle\sum_{i=1}^{N} \sum_{k=1}^{N} \Big(\alpha_i \alpha_k y_i y_k \left(x_i^T x_k \right) \Big) + \sum_{i=1}^{N} \alpha_i \right) \\ \displaystyle\sum_{i=1}^{N} \alpha_i y_i = 0 \,; \quad \alpha_i \geq 0 \,, \quad \forall\, 1 \leq i \leq N \,. \end{cases} \quad (3)$$

where $\alpha_1, \ldots, \alpha_N$ are the Lagrange multiplies.

If $\alpha^* = (\alpha_1^*, \ldots, \alpha_N^*)$ is a solution of (3), we say that x_i is a *support vector* if $\alpha_i^* \neq 0$. If $\mathcal{S}_1$ is the set of support vectors then the solution of (2) is

$$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i \,, \quad b^* = \frac{1}{|\mathcal{S}_1|} \sum_{x_i \in \mathcal{S}_1} \left(y_i - \sum_{x_j \in \mathcal{S}_1} \alpha_j^* y_j \left(x_i^T x_j \right) \right). \tag{4}$$

The soft margin hyperplane

The developments presented below were generalized by Cortes and Vapnik ([3], [9]) yielding to the concept of soft margin hyperplane. The method is essentially a regularization technique that introduces the effect of classification error set expressed by $\Phi_\sigma(\xi_1, \ldots, \xi_N) = \sum_{i=1}^{N} \xi_i^\sigma$, where σ is a positive constant and the non-negative variables ξ_i , $1 \leq i \leq N$ are called *slack variables*.

The parameters of the hyperplane that minimizes the classification error (soft margin hyperplane) are given by the solution of the QP constrained problem

$$\begin{cases} \min\limits_{w \in \mathbf{R}^d,\, b \in \mathbf{R},\, \xi \in \mathbf{R}^N} \left(\dfrac{1}{2} \|w\|^2 + c\, F\left(\sum\limits_{i=1}^{N} \xi_i^\sigma \right) \right) \\ \\ y_i \left(w^T x_i + b \right) \geq 1 - \xi_i \,, \; \forall\, 1 \leq i \leq N \,; \quad \xi_i \geq 0 \,, \; \forall\, 1 \leq i \leq N \,, \end{cases} \tag{5}$$

where c is a positive constant and F is a monotone increasing convex function such that $F(0) = 0$.

In case when $F(u) = u$ and $\sigma = 1$, using the Lagrange multipliers method the parameters of a hyperplane that minimizes the miss-classification errors in separating the data, are given by a solution of the constrained QP problem like (3), where $0 \leq \alpha_i \leq c$, $i = \overline{1, N}$. The computation for solving the QP problem like (3) is carried out by algorithms $SVM1$ and $SVM2$ ([8]).

Mapping to a higher dimensional space

Being given the non-separability can be due to the poor quality of the selected attributes in discriminating between the classes, a way to ameliorate the non-separability could be the increase of the data dimensionality by adding non-linear combinations of the measured attributes.

A symmetric function $K : R^d \times \mathbf{R}^d \to \mathbf{R}$ is a positive semi-definite kernel ([5]) if for any positive natural number M and for any $(h_1, \ldots, h_M) \in \mathbf{R}^M$, $\{x_1, \ldots, x_M\} \subset \mathbf{R}^d$, $\sum\limits_{i,j=1}^{M} h_i h_j K(x_i, x_j) \geq 0$ holds.

Let $g : \mathbf{R}^d \longrightarrow \mathbf{R}^m$ be a non-linear mapping where $m > d$. Then $K(x, x') = (g(x))^T g(x')$, $\forall\, x, x' \in \mathbf{R}^d$, is a positive semi-definite kernel. According to the Mercer theorem the explicit functional expression of the mapping g becomes hidden by the particular selected kernel and, moreover, the needed computations in the feature space are carried out in the initial space of attributes.

The dual problem of determining the soft margin hyperplane in the higher dimensional space $\mathbf{R}^m$, using the Lagrange multiplies method is

$$\begin{cases} \max\limits_{\alpha\in\mathbf{R}^N}\left(\sum\limits_{i=1}^{N}\alpha_i-\frac{1}{2}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\alpha_i\alpha_j y_i y_j K\left(x_i\,,x_j\right)\right) \\ \sum\limits_{i=1}^{N}\alpha_i y_i=0\,;\quad 0\le\alpha_i\le c\,,\quad i=\overline{1,N}\,. \end{cases} \tag{6}$$

The SMO algorithm for solving the dual problem

The SMO algorithm was introduced by Platt ([7]) as an iterative method for solving constrained optimization problems of the type

$$\begin{cases} \min\limits_{\alpha\in\mathbf{R}^N} W(\alpha) \\ \sum\limits_{i=1}^{N}\alpha_i y_i=0\,,\quad 0\le\alpha_i\le c\,,\quad i=\overline{1,N}\,. \end{cases} \tag{7}$$

In our tests we applied the SMO algorithm to minimize the function

$$W(\alpha)=-\sum_{i=1}^{N}\alpha_i+\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K\left(x_i\,,x_j\right)\,.$$

The SMO is a simple algorithm that solves the QP problem by decomposing the overall quadratic programming problem into smaller QP sub-problems. Briefly, the idea of the SMO algorithm is that at each iteration the smallest still unsolved QP sub-problems is solved. At every step, SMO chooses two Lagrange multipliers to jointly optimize, finds the optimal values for these multipliers and updates the SVM to reflect the new optimal values.

3. **Experimental results**

The potential and performances in solving classification problems of the presented approach described in the previous section were tested on the labeled training data were

$$\mathcal{S}=\left\{\left(\begin{pmatrix}1\\1\end{pmatrix},-1\right),\left(\begin{pmatrix}-1\\-1\end{pmatrix},-1\right),\left(\begin{pmatrix}-1\\1\end{pmatrix},1\right),\left(\begin{pmatrix}1\\-1\end{pmatrix},1\right)\right\}$$

and on simulated data generated from two-dimensional Gaussian distributions. In our tests we used polynomial and exponential kernels.

Test 1: In order to solve the well-known XOR problem we used the kernel $K\left(x,x'\right)=\left(x^Tx'+1\right)^6$ and the solution is showed in Figure 1*a*.

In tests **2**, and **3** we aimed to discriminate between examples coming from two normal classes generated from two-dimensional Gaussian distributions $\mathcal{N}\left(\mu_i,\Sigma_i\right)$, $i=1,2$, of volumes N_1 and N_2.

Test 2:
$N_1=N_2=25$, $\mu_1=\begin{pmatrix}1\\1\end{pmatrix}$, $\Sigma_1=\begin{pmatrix}1&0\\0&1\end{pmatrix}$, $\mu_2=\begin{pmatrix}4\\3\end{pmatrix}$, $\Sigma_2=\begin{pmatrix}0.5&0\\0&0.5\end{pmatrix}$.

The results obtained using polynomial and exponential kernels are showed in Figure 2.

Test 3:
$N_1 = 50$, $N_2 = 25$, $\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 4 \\ 3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}$.

The results obtained using polynomial and exponential kernels are showed in Figure 3 and Figure 4.

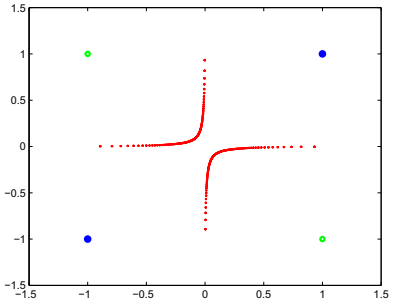

Figure 1. *XOR solution by kernel SVM*

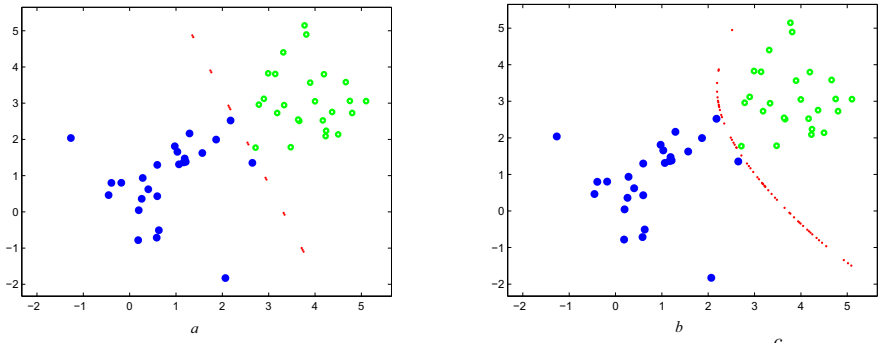

Figure 2. *a)* $K(x, x') = x^T x'$: *b)* $K(x, x') = (x^T x' + 1)^6$

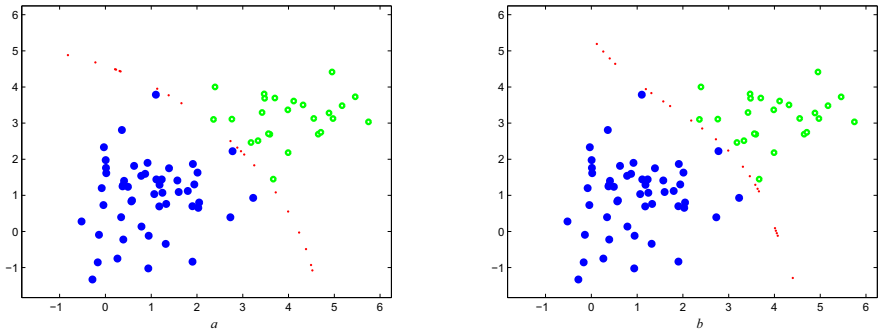

Figure 3. *a)* $K(x, x') = (x^T x' + 1)^2$; *b)* $K(x, x') = (x^T x' + 1)^6$

4. Conclusions

The paper concerns with using SVM-based techniques for two class classification. Since the design of the SVM involves the solutions of QP problems, a natural problem that arises is to find fast algorithms to solve them. In our developments we used the SMO algorithm introduced by Platt ([7]).

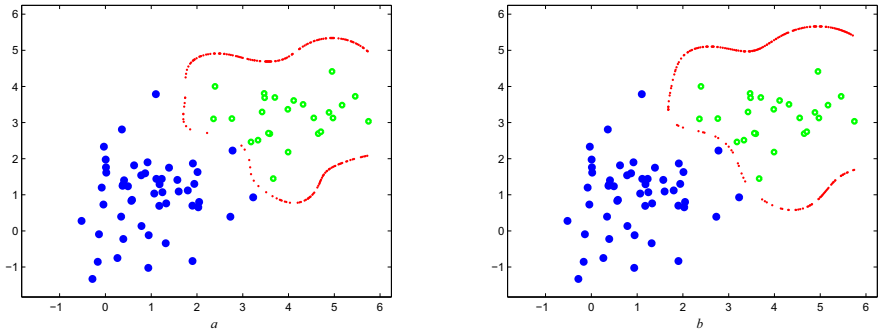

Figure 4. *a)* $K(x,x')=e^{-2(x-x')^T(x-x')}$; *b)* $K(x,x')=e^{-(x-x')^T(x-x')}$

Some work aiming to combine different types of classifiers with the SVM is still in progress.

References

1. Abe, S., *Support Vector Machines for Pattern Classification*, "Springer", London, 2005.
2. Burges, C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition*, "Data Mining and Knowledge Discovery", 2, 121 - 167, 1998.
3. Cortes, C., Vapnik, V.N., *Support-vector Networks*, "Machine Learning", 20(3), 273-297, 1995.
4. Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, "Cambridge University Press", 2000.
5. Mercer, J., *Functions of Positive and Negative Type, and Their Connection with the Theory of Integral Equations*, "Philosophical Transactions of the Royal Society of London", Series A, Vol. 209, 415-446, 1909.
6. Minh, H.Q., Niyogi, P., Yao, Y., *Mercer's Theorem, Feature Maps and Smoothing*, "Proc. of Computational Learning Theory (COLT'06)", 154-168, 2006.
7. Platt, J.C., *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*, in *Advances Kernel Methods: Support Vector Machines*, edited by B. Schölkopf, C.J.C. Burges and A.J. Smola, "The MIT Press", 1999.
8. State, L., Paraschiv-Munteanu, I., *A Comparative Analysis on the Potential of SVM and K-means in Solving Classification Tasks*, "Proceedings of the First International Conference on Modelling and Development of Intelligent Systems MDIS-2009", Sibiu, Romania, 244-253, 2009.
9. Vapnik, V.N., *Statistical Learning Theory*, "Wiley-Interscience", New York, 1998.