Volume 2 No.1, January 2012 ISSN 2224-3577

International Journal of Science and Technology

http://www.ejournalofsciences.org

Proximal Support Vector Machine for Disease Classification

[1]P.Venkatesan, [2]G. Meena Devi

[1] National Institute for Research in Tuberculosis, Indian Council of Medical Research, Chetpet, Chennai-31.

[2] Department of Mathematics, St.Joseph's College of Engineering, Chennai.-119

ABSTRACT

Parameter selection is one of the important steps involved in any model fitting. In this paper we have used Uniform Design Tables to choose the parameters for PSVM and SVM to classify the data. UD is one of the efficient space filling designs, which spreads the combination of parameters in the space uniformly scattered and generalizes the performance of the model efficiently. This paper compares performance of Proximal SVM and SVM with respect to prediction accuracy, using clinical trial outcomes on spinal tuberculosis treated patients. The PSVM performs slightly better than SVM in the training set, but the performance is as good as SVM in the test set.

Key words: *Uniform Design tables, SVM, Proximal SVM (PSVM)*

1. INTRODUCTION

In real world there are problems which cannot be solved using mathematical model or by classical programming techniques. In these areas, machine learning approach plays a vital role. As a broad subfield of artificial intelligence, machine learning is concerned with the design and development of algorithms and techniques that allow computers to learn [1]. The support vector machine (SVM) is widely used machine learning algorithm for pattern classification problem. It maps the data into high dimensional input space and constructs an optimal separating hyperplane in this space. SVM is based on the structural minimization principle. The quality of the solution does not depend directly on the dimensionality of the input space.

2. SUPPORT VECTOR MACHINES

SVMs[2,3,4,5,6,7] are one of the recently developed machine learning algorithm under the supervised learning approach, from the statistical learning theory implementing the structural risk minimization (SRM) principle. It has been successful in many real world classification problems like handwritten recognition, object recognition, text categorization, image recognition, classification of gene expression and many more. Unlike neural networks, SVMs minimize the estimation error keeping the training error fixed [5]. There are many variants of SVM ever since its existence in the literature. Proximal support vector machines (PSVM) was introduced recently as a variant of SVM for binary classifications in [8] and [9]. Theoretically, both PSVM and SVM target the optimal Bayes rule asymptotically, which explains their comparable performance in most studies. However, the PSVM solves a system of linear equations using an extremely fast and simple algorithm and thus demands much less computational effort than the SVM. SVMs are basically binary classifiers but can be extended to multi classification task using either one against one or one against all approaches. This next section deals with the brief overview of the theory and formulation of SVM.

3. Linear SVM

3.1 The separable case

Let us consider the training set D be $\{(x_i, y_i)\}_{i=1}^{N}, x_i \in R^m$ and the output label $y_i \in \{1,-1\}$, which is separable into two classes. Let w denote the normal vector to the separating hyperplane and $|b|/\|w\|$ is the perpendicular distance from the origin. Thus the hyperplane is given by:

$$w \cdot x + b = 0 \tag{1}$$

In other words, the basic problem becomes predicting the pair w and b such that the hyperplane is optimal (ie) with maximal margin on either sides of it.

Since the training set is linearly separable, if $y_i = 1$ then $w \cdot x_i + b \geq 1$ and if $y_i = -1$ then $w \cdot x_i + b \leq -1$ for all $x_i \in D$.Combining both

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i \tag{2}$$

The boundary of each class can be given as for class +,

$$w \cdot x + b = 1 \tag{3}$$

for class - , $\quad w \cdot x + b = -1$ (4)

The training points that lie on these hyperplanes are called support vectors (fig 1).

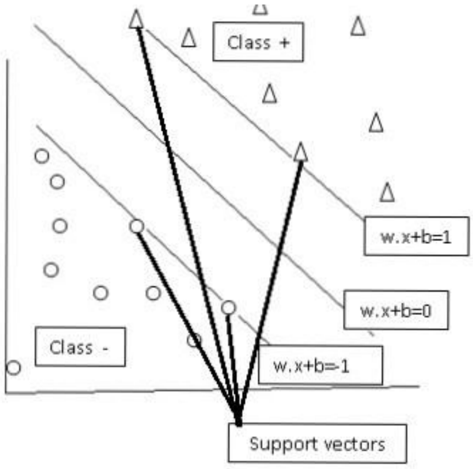

Fig 1. Linearly separable data with support vectors

We wish to find the hyperplane such that we can separate the point x_i, according to the function

$$f(x_i) = sign(w \cdot .x_i + b) = \begin{cases} 1, \; if \; y_i = 1, \\ -1, \; if \; y_i = -1 \end{cases} \quad (5)$$

Hence the SVM algorithm tries to find the optimal separating hyperplane with the maximum margin.

Equivalently, in this case the idea is to:

Minimize $\frac{1}{2} \| w \|^2$

subject to $y_i(w \cdot x_i + b) \geq 1 \quad \forall i$ (6)

With Larangian multipliers α_i, the objective function becomes

$$\text{Minimize} \quad L_P = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{N} \alpha_i [y_i(w \cdot x_i + b) - 1] \quad (7)$$

The Wolfe dual formulation [10] of the above is

Maxi

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \quad (8)$$

which is an unconstrained quadratic optimization problem. The solution becomes $w = \sum_{i=1}^{N_s} \alpha_i y_i x_i$ where N$_S$ is the number of support vectors (training points with $\alpha_i \neq 0$).

3.2. The non separable case

If D is not separable, we shall use the soft margin optimization. The important and inherent constraint in linear separable case is that it assumes that there are no training errors [6], which is impossible in many of the real world problems. To address this issue, the slack variables ξ_i's are introduced to allow the slight violation of the margin constraint to manage the noisy data

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \quad (9)$$

Thus it can be stated as

$$\text{Min} \quad \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i$$

$$\text{Subject to} \quad \begin{array}{ll} y_i(w \cdot x_i + b) \geq 1 - \xi_i & \\ \xi_i \geq 0 & \forall i \end{array} \quad (10)$$

where C is constant, which is a free parameter and balances the margin maximization and classification violation. In this way it can be regarded as a regularization parameter. It is fixed by the user. Let α_i's be the Larangian multipliers for $y_i(w \cdot x_i + b) - 1 + \xi_i \geq 0$ and μ_i's be the Larangian multipliers for $\xi_i \geq 0$. The Larangian primal objective function is

$$Min \quad L_P = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i [y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{N} \mu_i \xi_i \quad (11)$$

The dual of this is

$$\text{Maximize} \quad L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

$$\text{subject to} \; 0 \leq \alpha_i \leq C \quad and \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (12)$$

which is same as (8) except the upper bound for the Larangian multipliers. The advantage of the concept of the slack variables ξ_i's is that ξ_i's does not appear in the dual formulation. Therefore, in this case also the solution becomes $w = \sum_{i=1}^{N_s} \alpha_i y_i x_i$ where N$_S$ is the number of support vectors, whereas the bias b can be found using Karush Kuhn Tucker conditions for the primal. Therefore optimal hyperplane $w \cdot x + b = 0$ has been constructed. The decision function is

$$\begin{aligned} f(x) &= sign(w \cdot .x + b) \\ &= sign(\sum_{i=1}^{N_S} \alpha_i y_i x_i \cdot x + b) \end{aligned} \quad (13)$$

4. Non Linear SVM

In case of nonlinear separable data, the SVM first maps the input to high dimensional feature space and wherein the data are separated. This mapping is done through the functions called kernels. Let $z = \phi(x)$ be the function defined on the input x such that:

$$z_i \cdot z_j = \phi(x_i) \cdot \phi(x_j) = K(x_i, x_j) \qquad (14)$$

from R^m to the feature space Z, where K is the kernel satisfies the Mercer's conditions. Now the idea is to find the hyperplane $w \cdot z + b = 0$ so that we can separate the point in the feature space with this hyperplane. Incorporating (14), the nonlinear separating hyperplane is the solution of

Maximize

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j)$$

subject to $0 \le \alpha_i \le C \quad and \quad \sum_{i=1}^{N} \alpha_i y_i = 0$. (15)

And the decision function is:

$$f(x) = sign(w \cdot z + b)$$

$$= sign[\sum_{i=1}^{N_S} \alpha_i y_i K(x_i \cdot x) + b] \qquad (16)$$

5. THE LINEAR PROXIMAL SVM[9]

The standard SVM problem can be considered as classifying m points into two classes in the n-dimensional real space R^n, represented by the mxn matrix A. Let D be mxm diagonal matrix with +1 or -1 along its diagonal according to the class of the corresponding to the data instance. With the regularization parameter ϑ, in matrix notation the problem can be stated as

$$\underset{(w,\vartheta,y)\in R^{n+1+m}}{Min} \; \vartheta e'y + \tfrac{1}{2} w'w \qquad (17)$$
$$Such \;\; that \quad D(Aw - e\gamma) + y \ge e, \quad y \ge 0$$

Changing 1-norm of y to a 2-norm squared and appending γ^2 to $w'w$,

$$\underset{(w,\vartheta,y)\in R^{n+1+m}}{Min} \; \vartheta \frac{\|y\|^2}{2} + \tfrac{1}{2}\left(w'w + \gamma^2\right) \qquad (18)$$
$$Such \;\; that \quad D(Aw - e\gamma) + y \ge e$$

If the inequality constraint is replaced with equality constraint,

$$\underset{(w,\vartheta,y)\in R^{n+1+m}}{Min} \; \vartheta \frac{\|y\|^2}{2} + \tfrac{1}{2}\left(w'w + \gamma^2\right) \qquad (19)$$
$$Such \;\; that \quad D(Aw - e\gamma) + y = e$$

These modification changes the nature of the optimization problem significantly, as the explicit exact solution can be given as $w = A'Du; \;\; \gamma = -e'Du; \;\; y = \frac{u}{\vartheta}$ with $u = \left(\frac{1}{\vartheta} + HH'\right)^{-1} e$ where H = $D[A$- $e]$ using

Larangian formulation and KKT optimality conditions. A similar theory of the nonlinear version can be done. The only change in the final solution is the matrix A is replaced by the kernel matrix. [9]

7. MATERIALS AND METHODS

The data consists of 108 spinal tuberculosis patients allocated to a randomized clinical trial [11]. The patients had tuberculosis involving the thoracic or lumbar spine, allocated randomly to one of the three treatment series. The three treatment series were:

Rad 6: Modified 'Hong Kong' operation of radical resection of the lesion and insertion of autologus bone grafts. In addition, had isonaized plus rifampicin for 6 months.

Amb 6: Ambulant from the start of chemotherapy and received isonaized plus rifampicin for 6 months.

Amb 9: Ambulant from the start of chemotherapy and received isonaized plus rifampicin for 9 months.

The variables considered are Age, Gender (M-1, F-0), treatment(Rad-1, Amb6-2, Amb9-3) Fusion of bones in months, the angle of kyposis, Site of disease (Thoracic -1, Thoracic lumbar-2, Lumbar-3), Number of vertebrae involved, Total vertebral body loss, Sinuses and abscesses. Details can be found in [11].

The objective is to classify whether patients are favorable response or not. That is the disease was radio graphically quiescent or not. We have used PSVM Matlab code [12] and libSVM code available in [13] to build the model for this classification. Matlab 7.5 has been used for both the models. The parameter selection for nonlinear PSVM is done through uniform designs tables [14,15,16]. The results are given below in the following table.

Formulation		Classification rates (%)	
		Training Set	Testing Set
PSVM	linear	89.09	87.82
	nonlinear	90.64	86.18
SVM	linear	100	91.18
	nonlinear	100	85.29

8. RESULTS AND DISCUSSIONS

From the results the linear SVM has better accuracy of classification rates in the testing set, whereas the PSVM has approximately 87% in both cases. The classification rates are obtained with 10-fold cross validation as performance measure. However, the time to train the model is much lesser in the case of PSVM. This approach of diagnosing the disease with the selection of parameters through uniform designs table can be thought of a promising one. The nonlinear PSVM have used to diagnosis Ischemic heart disease and the training accuracy is 100%. [17] SVMs lend themselves particularly well to the analysis of broad patterns of disease classification from clinical data. They can easily deal with a large number of features and a small number of training patterns (dozens of patients). They integrate pattern selection and feature selection in a single consistent framework. SVM method has the potential in distinguishing cured patients from non cured and therefore may help in the early response prediction in spinal tuberculosis patients under treatment.

REFERENCES

[1] Ethem Alpaydin(2006), Introduction to Machine Learning, Prentice – Hall of India.

[2] Boser, B., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory.

[3] Vapnik V(1998), Statistical learning theory, John Wiley, New-York.

[4] Vapnik.V (1995), The Nature of Statistical Learning Theory. Springer.

[5] Kecman .V(2004), Learning and soft computing: Support Vector Machines, Neural Networks and Fuzzy Logic models.

[6] Burges.C.J.C(1998), A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 2, 121-167.

[7] Cristianini N., Shawe-Taylor J. (2000), An Introduction to Support Vector Machines, Cambridge University Press.

[8] Suykens. J. A. K. , Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural Processing Letters, 9(3), 293-300

[9] Fung.G, Mangasarian. O. L(2001),Proximal Support Vector Machine Classifiers , Proceedings of Knowledge Discovery and Data Mining, San Francisco.

[10] Bennett K.P, Bredensteiner. E(2000), Geometry in learning. In Geometry at Work : a collection of papers showing applications of geometry, Mathematical Association of America.

[11] An ICMR/BMRC Working Party Study in Madras(1989), A controlled trail of short-course regimens of chemotherapy in patients receiving ambulatory treatment of undergoing radical surgery for tuberculosis of the spine, Journal of Tuberculosis,36,1-21.

[12] http://research.cs.wisc.edu/dmi/svm/psvm/

[13] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[14] Fang.K.T , Dennis K.J. Lin (2003), Uniform Experimental designs and their applications in Industry, Hand Book of Statistics, 22, Elsevier Science B.V, 131-167

[15] Huang.C.M, Lee.Y.J, Dennis.K.J.Lin, Huang.S.Y (2007), Model Selection for Support Vector Machines via Uniform Design, Computational Statistics and Data analysis, 52, 335-346

[16] http://www.stat.psu.edu/~rli/DMCE/UniformDesign/

[17] Soman.K.P, Shyam.D.M, Madhavdas. P (2003), Efficient classification and analysis of ischemic heart disease using proximal support vector machines based decision trees. Conference on Convergent Technologies for Asia-Pacific Region