

Modifying A Linear Support Vector Machine for Microarray Data Classification

Prof. Rosemary A. Renaut

Dr. Hongbin Guo & Wang Juh Chen

Department of Mathematics and Statistics, Arizona State University

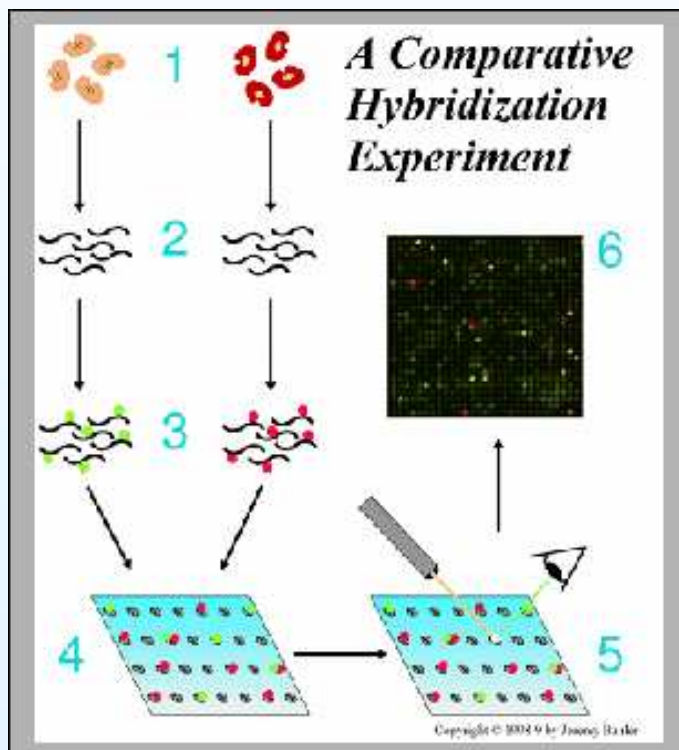
May 2005

OUTLINE

- Goal : Develop an efficient and accurate classifier for microarray data
- Methods : A modified linear support vector machine
- Content:
 - Microarray datasets and the classification problem
 - Support vector machine
 - New algorithm: development
 - Experimental data and design
 - Results and Discussion
 - Conclusions

Microarray Data

Microarrays are designed to understand how genomes are functioning. On a glass slide, thousands of spots, each related to a single gene, are used to simultaneously detect gene expression levels.



<http://www.math.tau.ac.il/~rshamir> A comparative hybridization experiment with cDNA microarrays (Prof. Ron Shamir):

1. Control Cells (left) and Target Cells (right)
2. Harvesting mRNA from both cell groups
3. Tagging the mRNA with green and red dye
4. Applying the mRNA to the cDNA microarray
5. Reading the result using a laser
6. A false-color composite representing the results

Classification and Characteristics of the Data

Classification: Assume m samples with known type, normal or cancerous tissue. For each sample microarray data are collected yielding n gene expression data for each sample. The classifier is based on the n features from the gene expression data: equivalently we have $m \times n$ matrix \mathbf{X} .

$m \ll n$: Many traditional classification methods, say Fisher's discriminant, require $m > n$. Support Vector Machine (SVM), [Vapnik95], was shown to provide good prediction accuracy, [Brown00].

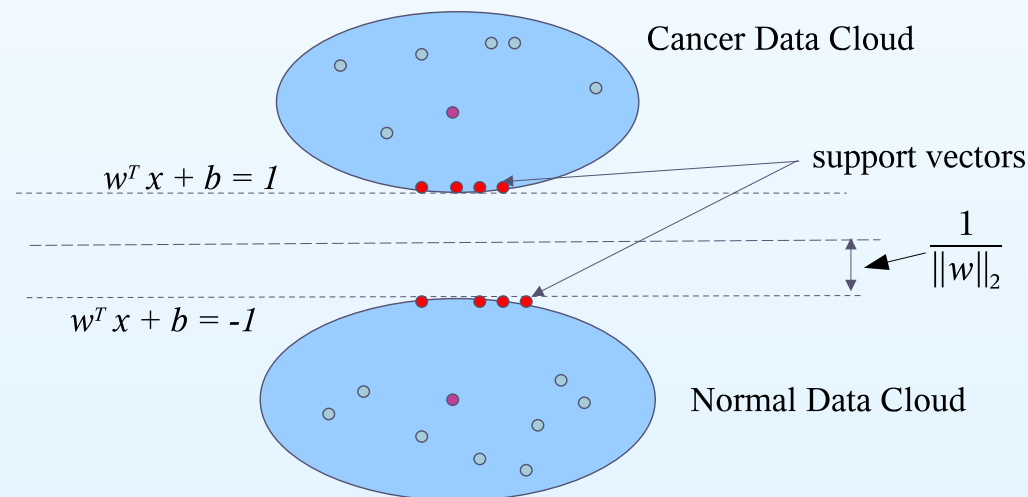
Noise: To better account for noisy data, we modify the SVM in a way similar to regularized total least squares, [GoHaOI99, RenGuo05].

Support Vector Machine: Background

The set of training vectors belonging to two separate classes

$$T = \{(x^1, y^1), \dots, (x^m, y^m)\}, \quad x^i \in R^n, y^i \in \{-1, 1\}$$

is separated by the hyperplane $w^T x + b = 0$. Scaling w and b the margin is written as $|w^T x + b| = 1$.



Support Vector Machine - Primal

Cost Functional:

$$\min_{\mathbf{w}, b, \xi} J_P(\mathbf{w}, \xi) = \frac{1}{2} \|\mathbf{w}\|_2^2 + \mu \sum_{i=1}^m \xi^i,$$

$$\text{subject to } y^i [\mathbf{w}^T \mathbf{x}^i + b] \geq 1 - \xi^i, \quad (1)$$

$$\xi^i \geq 0, \quad i = 1, \dots, m,$$

where ξ^i are *slack variables*

μ is a real positive constant.

Solution yields the classifier $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$

Support Vector Machine: Dual

The dual quadratic programming problem is :

$$\min_{\alpha} \frac{1}{2} \alpha^T A A^T \alpha - \mathbf{e}^T \alpha$$

With constraints

$$0 \leq \alpha_i \leq \mu, \quad i = 1, 2, \dots, m$$
$$\mathbf{y}^T \alpha = 0,$$

where $\alpha = (\alpha_1, \dots, \alpha_m)^T$ are Lagrange multipliers and

$$\mathbf{y} = (y^1, y^2, \dots, y^m)^T,$$

$$A = \text{diag}\{\mathbf{y}\} \cdot \mathbf{X}$$

$$\mathbf{e} = (1, 1, \dots, 1)^T$$

Least Squares SVM - Primal

Suykens and Vandewalle developed the Least Squares SVM (LS-SVM), [Suykens99], as follows

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} J_P(\mathbf{w}, \xi) &= \frac{1}{2} \|\mathbf{w}\|_2^2 + \frac{\mu}{2} \|\xi^i\|_2^2, \\ \text{subject to } y^i [\mathbf{w}^T \mathbf{x}^i + b] &= 1 - \xi^i, i = 1, \dots, m. \end{aligned} \quad (2)$$

Note:

- Inequalities in SVM are replaced by equalities in LS-SVM.
- ξ^i need not be positive in LS-SVM.

Least Squares SVM : Dual

The dual variables, Lagrange multipliers α , of LS-SVM problem are obtained from solution of the following linear system:

$$\begin{pmatrix} 0 & \mathbf{y}^T \\ \mathbf{y} & AA^T + I/\mu \end{pmatrix} \begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{e} \end{pmatrix}. \quad (3)$$

A conjugate method for efficient solution of large scale problems was proposed in [[Suykensetal](#)].

Modified SVM (MSVM) Primal

The constraints of both SVM and LS-SVM measure the distance of $f(\mathbf{x}^n)$ to the margins $f(\mathbf{x}) = 1$ and $f(\mathbf{x}) = -1$. We propose that the distance be measured in the feature space as follows,

$$\min_{\mathbf{w}, b, E} J_P(\mathbf{w}, E) = \frac{\mu}{2} \|\mathbf{w}\|_2^2 + \frac{1}{2} \|E\|_F^2, \quad (4)$$

$$\text{subject to } (A + E)\mathbf{w} + b\mathbf{y} = \mathbf{e},$$

where μ is a positive constant.

Modified SVM (MSVM)- Dual

With standard reduction and appropriate rearrangement yields obtain

$$\begin{pmatrix} \frac{1}{\nu}AA^T & \mathbf{y}^T & \mathbf{e}^T \\ \mathbf{y} & \lambda(\alpha) & 0 \\ \mathbf{e} & 0 & \lambda(\alpha)(\nu - \|\alpha\|^2 - 1) \end{pmatrix} \begin{pmatrix} \alpha \\ b \\ -1 \end{pmatrix} = \lambda(\alpha) \begin{pmatrix} \alpha \\ b \\ -1 \end{pmatrix}, \quad (5)$$

where α are the Lagrange multipliers

$$\begin{aligned} \nu &= \|\alpha\|^2 - \mu \\ \lambda(\alpha) &= \frac{\mathbf{e}^T \alpha}{\nu - \|\alpha\|^2}. \end{aligned}$$

It is easy to verify that solution α^* satisfies $\lambda(\alpha^*) = \|w\|^2$, thus by the requirements for the problem we seek the minimum eigenvalue that satisfies this system.

Algorithm

Design algorithm similar to iterative eigenvalue method for regularized total least squares.

Algorithm: Given μ , guess $\nu^{(0)}$ and $\lambda^{(0)} > 0$, set $k = 0$ and iterate.

1. While not converged

Do

(a) Calculate smallest eigenvalue, $\lambda^{(k)}$, and corresponding eigenvector.

(b) Scale the eigenvector to calculate $\alpha^{(k)}$.

(c) Test for convergence. If converged **Break** else $k = k + 1$.

(d) Update $\nu^{(k)}$ and $\lambda^{(k)}$

End Do.

2. $\alpha^* = \alpha^{(k)}$.

Experimental Datasets

Three public microarray datasets

Dataset name	# of samples	# of genes	# of normal samples	# of cancer samples
Lymphoma	72	7129	47	25
Ovarian	31	87558	17	14
Myeloma	105	7129	31	74

Experiments for Testing the Approach

- Test 1:** Perform leave-one-out cross validation for all three datasets and compare MSVM with SVM.
- Test 2:** Add additional noise to Lymphoma dataset and compare MSVM with SVM.
- Test 3:** Split Lymphoma dataset to *training* and *testing* data groups such that training data has 38 samples (27 normal and 11 cancer) and testing data has 34 samples (20 normal and 14 cancer). Classify testing data with classifier generated from training data.
- Test 4:** For Ovarian dataset perform leave-one-out cross validation for MSVM and SVM on reduced dataset with selected features. Gene selection is based on the measurement $F_j = \left| \frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-} \right|$, where μ_j^\pm and σ_j^\pm denote the mean and standard deviation for data associated with j -th gene and belonging to normal and cancer groups respectively.

Results of Test 1

Leave-one-out cross validation

Dataset name	Error rate of MSVM	Error rate of SVM
Lymphoma	2.8%	9.7%
Ovarian	29.0%	54.8%
Myeloma	1.0%	4.8%

For Ovarian dataset SVM Kernel matrix is ill-conditioned. Using an *optimal* regularization parameter, the error rate of SVM can be reduced to 25.8%, [Furey00].

Test 2: Noise added

Leave-one-out cross validation with added noise

Dataset	Noise Level	Error rate of MSVM	Error rate of SVM
Lymphoma	0%	2.8%	9.7%
	5%	2.8%	11.1%
	10%	4.2 %	8.3%
	50%	8.3 %	22.2%

MSVM is more robust to noise in data.

Test 3

Training and test data sampled from Lymphoma dataset

Noise Level	Error rate of MSVM	Error rate of SVM
0%	26.5%	29.4%
5%	26.5%	29.4%
50%	26.5 %	41.2%

MSVM outperforms SVM.

Results of Test 4

Leave-one-out cross validation using selected features: Ovarian dataset

# of Genes	Error rate of MSVM	Error rate of SVM
87558	29.0%	54.8%
1000	16.1%	16.1%
100	25.8 %	25.8%

Determining number of relevant genes (features) is an open problem.

Conclusions

1. Modified SVM method for microarray dataset classification has been proposed and tested.
2. MSVM is more robust to noise in data than SVM (Method of Steve Gunn).
3. MSVM outperforms SVM in most cases for the tested microarray datasets. Only in 3.5% cases does SVM perform best.

Future Work

1. Improve the algorithm by more rigorous numerical analysis.
2. Examine the feature selection problem.
3. Test the MSVM for datasets from other fields.
4. Optimize for cross validation.

References

- [Vapnik95] V. N. Vapnik, “The Nature of Statistical Learning Theory”, 1998.
- [Brown00] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. S. Furey, M. Ares, Jr. and D. Haussler, “Knowledge-based Analysis of Microarray Gene Expression Data Using SVMs”, *PNAS*, **97**, 262-267, 2000.
- [Furey00] T. S. Furey, N. Duffy, N. Cristianini, D. Bednarski, M. Schummer and D. Haussler, “SVM Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data”, *Bioinformatics*, **16**, 10, 906-914 2000.
- [GoHaOI99] G. H. Golub, P. C. Hansen and D. P. O’Leary, “Tikhonov Reg. and total least squares”, *Num. Lin. Alg.. App.*, **21**, 185-194, 1999.
- [RenGuo05] R. A. Renaut and H. Guo, “Efficient Solution of Reg. Total Least Squares”, *SIAM J. Matrix Anal. App.*, **26**, 457-476, 2005.
- [Suykensetal] J.A.K. Suykens, L. Lucas, P. Van Dooren, B. de Moor and J. Vandewalle, “LS SVM Classifiers: A Large Scale Algorithm”.
- [Suykens99] Suykens J.A.K. and J. Vandewalle, “LS SVM classifiers”, *Neural Processing Letters* **9(3)**, 293-300, 1999.