Building Sparse Large Margin Classifiers

Mingrui Wu, Bernhard Schölkopf, Gökhan Bakir
Max Planck Institute for Biological Cybernetics

August 10, 2005

Kernel Classifiers

- binary classification problem: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, $\mathbf{x}_i \in \mathcal{X} \subseteq \mathcal{R}^d$, $y_i \in \{-1, 1\}$
- classification function $\hat{f}(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$

$$f(\mathbf{x}) = \sum_{i=1}^{N_{XV}} \hat{\alpha}_i K(\hat{\mathbf{x}}_i, \mathbf{x}) + b \tag{1}$$

- for positive definite kernel function (Schölkopf & Smola, 2002) K: feature space $\mathcal{F}$,
  implicit map $\phi : \mathcal{X} \rightarrow \mathcal{F}$, $\mathbf{x} \rightarrow \phi(\mathbf{x})$
  $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$

$$f(\mathbf{x}) = \langle \Psi, \phi(\mathbf{x}) \rangle + b \tag{2}$$

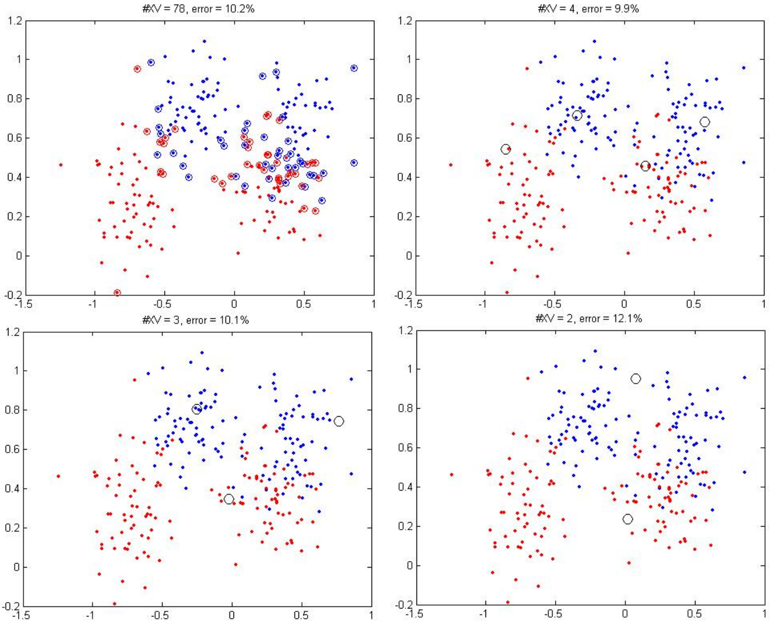
$$\Psi = \sum_{i=1}^{N_{XV}} \hat{\alpha}_i \phi(\hat{\mathbf{x}}_i) \tag{3}$$

Sparse: Small N_{XV} is Desirable

- Reduced Set (RS) method (Burges, 1996), given $N_z << N_{XV}$, find $\mathbf{z}_1, \ldots, \mathbf{z}_{N_z}$ and $\beta_1, \ldots, \beta_{N_z}$ such that

$$\| \Psi - \sum_{j=1}^{N_z} \beta_j \phi(\mathbf{z}_j) \|^2 \tag{4}$$

  is minimized
- Reduced Support Vector Machine (RSVM) (Lee & Mangasarian, 2001)
- Relevance Vector Machine (RVM) (Tipping, 2001)


#XV = 78, error = 10.2%
#XV = 4, error = 9.9%
#XV = 3, error = 10.1%
#XV = 2, error = 12.1%
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-1.5
-1
-0.5
0
0.5
1

Building Sparse Large Margin Classifiers (SLMC)

- objective: given $N_z > 0$, build a kernel classifier, such that $N_{XV} = N_z$ and the margin (Vapnik, 1995) of the classifier is maximized.
- $$\begin{aligned}
\min_{\mathbf{w},\boldsymbol{\xi},b,\boldsymbol{\beta},\mathbf{Z}} \quad & \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{N} \xi_i && (5)\\
\text{subject to} \quad & y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad \forall i && (6)\\
& \xi_i \geq 0 \quad \forall i && (7)\\
& \mathbf{w} = \sum_{i=1}^{N_z} \beta_i \phi(\mathbf{z}_i) && (8)
\end{aligned}$$
- $$G(b, \boldsymbol{\beta}, \mathbf{Z}) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{N} \xi_i \qquad (9)$$

Gradient based Approach

- $$G(b, \boldsymbol{\beta}, \mathbf{Z}) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{N} \xi_i \tag{10}$$
- at any given $\mathbf{Z}$, G becomes a function of b and $\boldsymbol{\beta}$, denoted by $G(b, \boldsymbol{\beta} \mid \mathbf{Z})$
- $$W(\mathbf{Z}) = \min_{b \in \mathcal{R}, \boldsymbol{\beta} \in \mathcal{R}^{N_z}} G(b, \boldsymbol{\beta} \mid \mathbf{Z}) \tag{11}$$
- for any $\mathcal{A} \subseteq \mathcal{R}^{d \times N_z}$ we have

$$\min_{\mathbf{Z} \in \mathcal{A}} W(\mathbf{Z}) = \min_{b, \boldsymbol{\beta}, \mathbf{Z} \in \mathcal{A}} G(b, \boldsymbol{\beta}, \mathbf{Z}) \tag{12}$$

- minimize $W(\mathbf{Z})$: compute $W(\mathbf{Z})$ and $\nabla W(\mathbf{Z})$

Computing $W(\mathbf{Z})$ and β: Original Problem

$$\min_{\mathbf{w},\boldsymbol{\xi},b,\boldsymbol{\beta}} \quad \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{N}\xi_i \tag{13}$$

$$\text{subject to} \quad y_i(\mathbf{w}^\top\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad \forall i \tag{14}$$

$$\xi_i \geq 0 \quad \forall i \tag{15}$$

$$\mathbf{w} = \sum_{i=1}^{N_z}\beta_i\phi(\mathbf{z}_i) \tag{16}$$

Computing $W(\mathbf{Z})$ and β: Dual Problem

- dual problem

$$\max_{\boldsymbol{\alpha}\in\mathcal{R}^N} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \hat{K}_z(\mathbf{x}_i,\mathbf{x}_j) \tag{17}$$

$$\text{subject to} \quad \sum_{i=1}^{N} y_i\alpha_i = 0 \tag{18}$$

$$\text{and} \quad 0 \le \alpha_i \le C \quad \forall i \tag{19}$$

- modified kernel function $\hat{K}_z$

$$\hat{K}_z(\mathbf{x}_i,\mathbf{x}_j) = \psi_z(\mathbf{x}_i)^\top(\mathbf{K}^z)^{-1}\psi_z(\mathbf{x}_j) \tag{20}$$

$$\psi_z(\mathbf{x}_i) = [K(\mathbf{z}_1,\mathbf{x}_i),\ldots,K(\mathbf{z}_{N_z},\mathbf{x}_i)]^\top \tag{21}$$

Computing $W(\mathbf{Z})$ and β: conclusion

- given $\mathbf{Z}$, computing the expansion coefficients of SVM with kernel function K is equivalent to training an SVM with a modified kernel function $\hat{K}_z$.
- function value

$$W(\mathbf{Z}) = \sum_{i=1}^{N} \alpha_i^z - \frac{1}{2} \sum_{i}^{N} \sum_{j=1}^{N} \alpha_i^z \alpha_j^z y_i y_j \hat{K}_z(\mathbf{x}_i, \mathbf{x}_j) \tag{22}$$

- expansion coefficients

$$\boldsymbol{\beta} = (\mathbf{K}^z)^{-1} \sum_{i=1}^{N} \alpha_i^z y_i \psi_z(\mathbf{x}_i) = (\mathbf{K}^z)^{-1} (\mathbf{K}^{zx}) \mathbf{Y} \boldsymbol{\alpha}^z \tag{23}$$

Computing $\nabla W(\mathbf{Z})$

- function value

$$W(\mathbf{Z}) = \sum_{i=1}^{N} \alpha_i^z - \frac{1}{2} \sum_{i}^{N} \sum_{j=1}^{N} \alpha_i^z \alpha_j^z y_i y_j \hat{K}_z(\mathbf{x}_i, \mathbf{x}_j) \tag{24}$$

- gradient (Chapelle et al., 2002)

$$\frac{\partial W}{\partial \mathbf{z}_{uv}} = -\frac{1}{2} \sum_{i,j=1}^{N} \alpha_i^z \alpha_j^z y_i y_j \frac{\partial \hat{K}_z(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{z}_{uv}} \tag{25}$$

Analysis of $\hat{K}_z$: feature space $\mathcal{F}_z$

- kernel function K, feature space $\mathcal{F}$, $\phi(\mathbf{z}_1), \ldots, \phi(\mathbf{z}_{\mathbf{N_z}})$
- orthonormalization of $\phi(\mathbf{z}_1), \ldots, \phi(\mathbf{z}_{\mathbf{N_z}})$

$$\mathbf{T} = (\mathbf{K}^z)^{-\frac{1}{2}} \tag{26}$$

$$\mathbf{U}^z = [\phi(\mathbf{z}_1), \ldots, \phi(\mathbf{z}_{N_z})]\mathbf{T}^\top \tag{27}$$

$$(\mathbf{U}^z)^\top \mathbf{U}^z = \mathbf{T}\mathbf{K}^z\mathbf{T}^\top = \mathbf{I} \tag{28}$$

- Subspace $\mathcal{F}_z$

$$\begin{aligned} (\mathbf{U}^z)^\top \phi(\mathbf{x}) &= \phi_z(\mathbf{x}) \\ \hat{K}_z(\mathbf{x}_i, \mathbf{x}_j) &= \langle \phi_z(\mathbf{x}_i), \phi_z(\mathbf{x}_j) \rangle \end{aligned}$$

- find $\mathcal{F}_z$ such that margin of $\{\phi_z(\mathbf{x}_i), y_i\}_{i=1}^N$ is maximized

Comparison with Related Approaches

- RS method, given $\mathbf{Z}$, computing $\boldsymbol{\beta}$ to minimize
  -
  $$\| \Psi - \sum_{j=1}^{N_z} \beta_j \phi(\mathbf{z}_j) \|^2 \tag{29}$$
  -
  $$\boldsymbol{\beta} = (\mathbf{K}^z)^{-1}(\mathbf{K}^{zx})\mathbf{Y}\boldsymbol{\alpha} \tag{30}$$
  - modified RS method
  $$\boldsymbol{\beta} = (\mathbf{K}^z)^{-1}(\mathbf{K}^{zx})\mathbf{Y}\boldsymbol{\alpha}^z \tag{31}$$
- RSVM
  - XVs $\mathbf{Z}$ is randomly selected from the training data
  - modified RSVM
- RVM: XVs are always a subset of the training data

Experimental Settings

- approaches to be compared: SVM, RS method, modified RS method (MRS), RSVM, modified RSVM (MRSVM), RVM and SLMC
- data sets: Banana, Breast Cancer, Waveform, German and Image

Numerical Results

Table: Results on five classification benchmarks. (Gunnar Rätsch)

Dataset		Banana	B.Cancer	Waveform	German	Image
SVM	N_{SV}	86.7	112.8	158.9	408.2	172.1
	Error	11.8	28.6	9.9	22.5	2.8
5%	RS	39.4	28.8	**9.9**	22.9	37.6
	MRS	27.6	28.8	10.0	22.5	19.4
	RSVM	29.9	29.5	15.1	23.6	23.6
	MRSVM	28.1	29.4	14.7	23.9	20.7
	SLMC	**16.5**	**27.9**	**9.9**	**22.3**	**5.2**
10%	RS	21.9	**27.9**	10.0	22.9	18.3
	MRS	17.5	29.0	**9.9**	**22.6**	6.9
	RSVM	17.5	31.0	11.6	24.5	14.2
	MRSVM	16.9	30.3	11.8	23.7	12.7
	SLMC	**11.0**	**27.9**	**9.9**	22.9	**3.6**
RVM	$N_z/N_{SV}(\%)$	13.2	5.6	9.2	3.1	20.1
	Error	**10.8**	29.9	10.9	**22.2**	3.9

Summary

- SLMC, discriminating subspace $\mathcal{F}_z$
- given XVs, computing the expansion coefficients, modified RS method, modified RSVM, try on other methods
- add one more constraint to build other sparse learning algorithms: KFD, KPCA, one-class SVM, regression, etc.

References

Burges, C. J. C. (1996). Simplified support vector decision rules. *Proc. 13th International Conference on Machine Learning* (pp. 71–77). Morgan Kaufmann.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, *46*, 131–159.

Lee, Y.-J., & Mangasarian, O. L. (2001). RSVM: reduced support vector machines. *CD Proceedings of the First SIAM International Conference on Data Mining*. Chicago.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: The MIT Press.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *1*, 211–214.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer Verlag.