Direct Zero-norm Optimization for Feature Selection

Kaizhu Huang[1], Irwin King[2], Michael R. Lyu[2]

[1] Department of Engineering Mathematics
University of Bristol
[2]Department of Computer Science & Engineering
The Chinese University of Hong Kong

December 16, 2008
ICDM 2008, Pisa, Italy

Problem

Zero-norm Definition

Zero-norm $||\mathbf{w}||_0^0$: Number of non-zero elements in a vector $\mathbf{w}$

$$||\mathbf{w}||_0^0 = card\{w_i | w_i \neq 0\}$$

Problem Definition

Zero-norm Feature Selection

$$\min_{\mathbf{w},b} ||\mathbf{w}||_0^0 + C\sum_{i=1}^{l}\xi_i$$

s.t. $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i$,

$\mathbf{x}_i (i = 1, \ldots, l)$: training samples

$y_i \in \{-1, +1\}$: category label of $\mathbf{x}_i$

- Challenges
  - Zero-norm is non-convex and discontinuous
  - Minimizing zero-norm is combinatorially very difficult problem [Amaldi & Kann 1998]
- Previous Solution: Optimizing a surrogate term
  - $||\mathbf{w}||_0^0 \approx \sum_i 1 - \exp\{-\alpha|w_i|\}$ [Bradley et al. 1998]
  - $||\mathbf{w}||_0^0 \approx \sum_i \ln(\epsilon + |w_i|)$ [Weston et al. 2003]

Problem

Zero-norm Definition

Zero-norm $||\mathbf{w}||_0^0$: Number of non-zero elements in a vector $\mathbf{w}$

$$||\mathbf{w}||_0^0 = card\{w_i | w_i \neq 0\}$$

Problem Definition

Zero-norm Feature Selection

$$\min_{\mathbf{w},b} ||\mathbf{w}||_0^0 + C \sum_{i=1}^{l} \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i,$$

$\mathbf{x}_i (i = 1, \ldots, l)$: training samples

$y_i \in \{-1, +1\}$: category label of $\mathbf{x}_i$

- Challenges
  - Zero-norm is non-convex and discontinuous
  - Minimizing zero-norm is combinatorially very difficult problem [Amaldi & Kann 1998]
- Previous Solution: Optimizing a surrogate term
  - $||\mathbf{w}||_0^0 \approx \sum_i 1 - \exp\{-\alpha|w_i|\}$ [Bradley et al. 1998]
  - $||\mathbf{w}||_0^0 \approx \sum_i \ln(\epsilon + |w_i|)$ [Weston et al. 2003]

Problem

Zero-norm Definition

Zero-norm $||\mathbf{w}||_0^0$: Number of non-zero elements in a vector $\mathbf{w}$

$$||\mathbf{w}||_0^0 = card\{w_i | w_i \neq 0\}$$

Problem Definition

Zero-norm Feature Selection

$$\min_{\mathbf{w},b} ||\mathbf{w}||_0^0 + C \sum_{i=1}^{l} \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i,$$

$\mathbf{x}_i (i = 1, \ldots, l)$: training samples

$y_i \in \{-1, +1\}$: category label of $\mathbf{x}_i$

- Challenges
  - Zero-norm is non-convex and discontinuous
  - Minimizing zero-norm is combinatorially very difficult problem [Amaldi & Kann 1998]
- Previous Solution: Optimizing a surrogate term

Problem

Zero-norm Definition

Zero-norm $||\mathbf{w}||_0^0$: Number of non-zero elements in a vector $\mathbf{w}$

$$||\mathbf{w}||_0^0 = card\{w_i | w_i \neq 0\}$$

Problem Definition

Zero-norm Feature Selection

$$\min_{\mathbf{w},b} ||\mathbf{w}||_0^0 + C \sum_{i=1}^{l} \xi_i$$
$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i,$$

$\mathbf{x}_i (i = 1, \ldots, l)$: training samples

$y_i \in \{-1, +1\}$: category label of $\mathbf{x}_i$

- Challenges
  - Zero-norm is non-convex and discontinuous
  - Minimizing zero-norm is combinatorially very difficult problem [Amaldi & Kann 1998]
- Previous Solution: Optimizing a surrogate term
  - $||\mathbf{w}||_0^0 \approx \sum_i 1 - \exp\{-\alpha |w_i|\}$ [Bradley et al. 1998]
  - $||\mathbf{w}||_0^0 \approx \sum_i \ln(\epsilon + |w_i|)$ [Weston et al. 2003]

Problem

Zero-norm Definition

Zero-norm $||\mathbf{w}||_0^0$: Number of non-zero elements in a vector $\mathbf{w}$

$$||\mathbf{w}||_0^0 = card\{w_i | w_i \neq 0\}$$

Problem Definition

Zero-norm Feature Selection

$$\min_{\mathbf{w},b} ||\mathbf{w}||_0^0 + C \sum_{i=1}^{l} \xi_i$$
$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i,$$

$\mathbf{x}_i (i = 1, \ldots, l)$: training samples

$y_i \in \{-1, +1\}$: category label of $\mathbf{x}_i$

- Challenges
  - Zero-norm is non-convex and discontinuous
  - Minimizing zero-norm is combinatorially very difficult problem [Amaldi & Kann 1998]
- Previous Solution: Optimizing a surrogate term
  - $||\mathbf{w}||_0^0 \approx \sum_i 1 - \exp\{-\alpha |w_i|\}$ [Bradley et al. 1998]
  - $||\mathbf{w}||_0^0 \approx \sum_i \ln(\epsilon + |w_i|)$ [Weston et al. 2003]

Problem

Zero-norm Definition

Zero-norm $||\mathbf{w}||_0^0$: Number of non-zero elements in a vector $\mathbf{w}$

$$||\mathbf{w}||_0^0 = card\{w_i | w_i \neq 0\}$$

Problem Definition

Zero-norm Feature Selection

$$\min_{\mathbf{w},b} ||\mathbf{w}||_0^0 + C \sum_{i=1}^{l} \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i,$$

$\mathbf{x}_i (i = 1, \ldots, l)$: training samples

$y_i \in \{-1, +1\}$: category label of $\mathbf{x}_i$

- Challenges
  - Zero-norm is non-convex and discontinuous
  - Minimizing zero-norm is combinatorially very difficult problem [Amaldi & Kann 1998]
- Previous Solution: Optimizing a surrogate term
  - $||\mathbf{w}||_0^0 \approx \sum_i 1 - \exp\{-\alpha |w_i|\}$ [Bradley et al. 1998]
  - $||\mathbf{w}||_0^0 \approx \sum_i \ln(\epsilon + |w_i|)$ [Weston et al. 2003]

Problem

Zero-norm Definition

Zero-norm $||\mathbf{w}||_0^0$: Number of non-zero elements in a vector $\mathbf{w}$

$$||\mathbf{w}||_0^0 = card\{w_i | w_i \neq 0\}$$

Problem Definition

Zero-norm Feature Selection

$$\min_{\mathbf{w},b} ||\mathbf{w}||_0^0 + C \sum_{i=1}^{l} \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i,$$

$\mathbf{x}_i (i = 1, \ldots, l)$: training samples

$y_i \in \{-1, +1\}$: category label of $\mathbf{x}_i$

- Challenges
  - Zero-norm is non-convex and discontinuous
  - Minimizing zero-norm is combinatorially very difficult problem [Amaldi & Kann 1998]
- Previous Solution: Optimizing a surrogate term
  - $||\mathbf{w}||_0^0 \approx \sum_i 1 - \exp\{-\alpha |w_i|\}$ [Bradley et al. 1998]
  - $||\mathbf{w}||_0^0 \approx \sum_i \ln(\epsilon + |w_i|)$ [Weston et al. 2003]

Contributions

- A direct zero-norm optimization is achieved for feature selection
- A Bayesian interpretation or justification
- More accurate and faster than surrogate approaches
- A variation of our proposed method is strictly equivalent to [Weston et al. 2003] (not elaborated in the talk)

Contributions

- A direct zero-norm optimization is achieved for feature selection
- A Bayesian interpretation or justification
- More accurate and faster than surrogate approaches
- A variation of our proposed method is strictly equivalent to [Weston et al. 2003] (not elaborated in the talk)

Contributions

- A direct zero-norm optimization is achieved for feature selection
- A Bayesian interpretation or justification
- More accurate and faster than surrogate approaches
- A variation of our proposed method is strictly equivalent to [Weston et al. 2003] (not elaborated in the talk)

Contributions

- A direct zero-norm optimization is achieved for feature selection
- A Bayesian interpretation or justification
- More accurate and faster than surrogate approaches
- A variation of our proposed method is strictly equivalent to [Weston et al. 2003] (not elaborated in the talk)

Bayesian Viewpoint on Classifiers (I)

- The output z of classifiers $\{\mathbf{w}, b\}$ is corrupted by a zero-mean and unit-variance Gaussian distribution o.

$$z(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{h}(\mathbf{x}) + o$$

b is incorporated into $\mathbf{w}$;

$$\mathbf{h}(\mathbf{x}) = \begin{cases} \text{Linear case:} & [1, \mathbf{x}]' \\ \text{Kernel case:} & [1, k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_l)]' \end{cases}$$

- Given a prior probability of $\mathbf{w}$, EM can be used to find the optimal $\mathbf{w}$ (in the sense of MAP).
- Jeffery priors: S_1: $p(w_i|\tau_i) = \mathcal{N}(w_i|0, \tau_i)$. S_2: $p(\tau_i) \propto 1/\tau_i$ will motivate the zero-norm implementation.

Bayesian Viewpoint on Classifiers (I)

- The output z of classifiers $\{\mathbf{w}, b\}$ is corrupted by a zero-mean and unit-variance Gaussian distribution o.

$$z(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{h}(\mathbf{x}) + o$$

b is incorporated into $\mathbf{w}$;

$$\mathbf{h}(\mathbf{x}) = \begin{cases} \text{Linear case:} & [1, \mathbf{x}]' \\ \text{Kernel case:} & [1, k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_l)]' \end{cases}$$

- Given a prior probability of $\mathbf{w}$, EM can be used to find the optimal $\mathbf{w}$ (in the sense of MAP).
- Jeffery priors: S_1: $p(w_i|\tau_i) = \mathcal{N}(w_i|0, \tau_i)$. S_2: $p(\tau_i) \propto 1/\tau_i$ will motivate the zero-norm implementation.

Bayesian Viewpoint on Classifiers (I)

- The output z of classifiers $\{\mathbf{w}, b\}$ is corrupted by a zero-mean and unit-variance Gaussian distribution o.

$$z(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{h}(\mathbf{x}) + o$$

b is incorporated into $\mathbf{w}$;

$$\mathbf{h}(\mathbf{x}) = \begin{cases} \text{Linear case:} & [1, \mathbf{x}]' \\ \text{Kernel case:} & [1, k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_l)]' \end{cases}$$

- Given a prior probability of $\mathbf{w}$, EM can be used to find the optimal $\mathbf{w}$ (in the sense of MAP).
- Jeffery priors: S_1: $p(w_i|\tau_i) = \mathcal{N}(w_i|0, \tau_i)$. S_2: $p(\tau_i) \propto 1/\tau_i$ will motivate the zero-norm implementation.

Bayesian Viewpoint on Classifiers (II)(Jeffery priors)

- M-step (Maximize the following w.r.t. $\mathbf{w}$)

  $$\log p(\mathbf{w}|\mathbf{y},\mathbf{z}) \propto \log p(\mathbf{z}|\mathbf{w}) + \log p(\mathbf{w}) \propto -||\mathbf{H}\mathbf{w}-\mathbf{z}||^2 - \mathbf{w}^T\mathbf{\Lambda}\mathbf{w},$$

  where $\mathbf{\Lambda} = \mathrm{diag}(1/\tau_1,\ldots,1/\tau_l)$.

- E-step (Calculate the Expectation of missing variables z_i and $1/\tau_i$)

  $$E[z_i|\widehat{w}_{(t)},\mathbf{y}] = \begin{cases} \mathbf{w}^T\mathbf{h}(\mathbf{x}_i) + \frac{\mathcal{N}(\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)}{1-\mathcal{S}(-\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)} & \text{if} \quad y_i = 1 \\ \mathbf{w}^T\mathbf{h}(\mathbf{x}_i) - \frac{\mathcal{N}(\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)}{\mathcal{S}(-\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)} & \text{if} \quad y_i = -1 \end{cases}$$

Bayesian Viewpoint on Classifiers (II)(Jeffery priors)

- M-step (Maximize the following w.r.t. **w**)

  $$\log p(\mathbf{w}|\mathbf{y},\mathbf{z}) \propto \log p(\mathbf{z}|\mathbf{w}) + \log p(\mathbf{w}) \propto -||\mathbf{H}\mathbf{w}-\mathbf{z}||^2 - \mathbf{w}^T\mathbf{\Lambda}\mathbf{w},$$

  where $\mathbf{\Lambda} = \mathrm{diag}(1/\tau_1, \ldots, 1/\tau_l)$.

- E-step (Calculate the Expectation of missing variables z_i and $1/\tau_i$)

$$\mathrm{E}[z_i|\widehat{w}_{(t)},\mathbf{y}] = \begin{cases} \mathbf{w}^T\mathbf{h}(\mathbf{x}_i) + \frac{\mathcal{N}(\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)}{1-\mathcal{S}(-\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)} & \text{if} \quad y_i = 1 \\ \mathbf{w}^T\mathbf{h}(\mathbf{x}_i) - \frac{\mathcal{N}(\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)}{\mathcal{S}(-\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)} & \text{if} \quad y_i = -1 \end{cases}$$

$$\mathrm{E}[\tau_i^{-1}|\widehat{\mathbf{w}}_{(t)},\mathbf{y}] = \frac{\int_0^{+\infty}\frac{1}{\tau_i}p(\tau_i|\widehat{\mathbf{w}}_{(t)},\mathbf{y})d\tau_i}{\int_0^{+\infty}p(\tau_i|\widehat{\mathbf{w}}_{(t)},\mathbf{y})d\tau_i} = \frac{\int_0^{+\infty}\frac{1}{\tau_i}p(\tau_i)p(\widehat{\mathbf{w}}_{(t)}|\tau_i)d\tau_i}{\int_0^{+\infty}p(\tau_i)p(\widehat{\mathbf{w}}_{(t)}|\tau_i)d\tau_i} = |\widehat{w}_{i,(t)}|^{-2}.$$

Bayesian Viewpoint on Classifiers (II)(Jeffery priors)

- M-step (Maximize the following w.r.t. $\mathbf{w}$)

  $$\log p(\mathbf{w}|\mathbf{y},\mathbf{z}) \propto \log p(\mathbf{z}|\mathbf{w}) + \log p(\mathbf{w}) \propto -||\mathbf{H}\mathbf{w}-\mathbf{z}||^2 - \mathbf{w}^T\mathbf{\Lambda}\mathbf{w},$$

  where $\mathbf{\Lambda} = \mathrm{diag}(1/\tau_1,\dots,1/\tau_l)$.
- E-step (Calculate the Expectation of missing variables z_i and $1/\tau_i$)

$$E[z_i|\widehat{w}_{(t)},\mathbf{y}] = \begin{cases} \mathbf{w}^T\mathbf{h}(\mathbf{x}_i) + \frac{\mathcal{N}(\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)}{1-\mathcal{S}(-\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)} & \text{if} \quad y_i = 1 \\ \mathbf{w}^T\mathbf{h}(\mathbf{x}_i) - \frac{\mathcal{N}(\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)}{\mathcal{S}(-\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)} & \text{if} \quad y_i = -1 \end{cases}$$

$$E[\tau_i^{-1}|\widehat{\mathbf{w}}_{(t)},\mathbf{y}] = \frac{\int_0^{+\infty}\frac{1}{\tau_i}p(\tau_i|\widehat{\mathbf{w}}_{(t)},\mathbf{y})d\tau_i}{\int_0^{+\infty}p(\tau_i|\widehat{\mathbf{w}}_{(t)},\mathbf{y})d\tau_i} = \frac{\int_0^{+\infty}\frac{1}{\tau_i}p(\tau_i)p(\widehat{\mathbf{w}}_{(t)}|\tau_i)d\tau_i}{\int_0^{+\infty}p(\tau_i)p(\widehat{\mathbf{w}}_{(t)}|\tau_i)d\tau_i} = |\widehat{w}_{i,(t)}|^{-2}.$$

Bayesian Viewpoint on Classifiers (II)(Jeffery priors)

- M-step (Maximize the following w.r.t. **w**)

  $$\log p(\mathbf{w}|\mathbf{y},\mathbf{z}) \propto \log p(\mathbf{z}|\mathbf{w}) + \log p(\mathbf{w}) \propto -||\mathbf{H}\mathbf{w}-\mathbf{z}||^2 - \mathbf{w}^T\mathbf{\Lambda}\mathbf{w},$$

  where $\mathbf{\Lambda} = \mathrm{diag}(1/\tau_1, \ldots, 1/\tau_l)$.
- E-step (Calculate the Expectation of missing variables z_i and $1/\tau_i$)

$$\mathrm{E}[z_i|\widehat{w}_{(t)},\mathbf{y}] = \begin{cases} \mathbf{w}^T\mathbf{h}(\mathbf{x}_i) + \frac{\mathcal{N}(\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)}{1-\mathcal{S}(-\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)} & \text{if} \quad y_i = 1 \\ \mathbf{w}^T\mathbf{h}(\mathbf{x}_i) - \frac{\mathcal{N}(\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)}{\mathcal{S}(-\mathbf{w}^T\mathbf{h}(\mathbf{x}_i)|0,1)} & \text{if} \quad y_i = -1 \end{cases}$$

$$\mathrm{E}[\tau_i^{-1}|\widehat{\mathbf{w}}_{(t)},\mathbf{y}] = \frac{\int_0^{+\infty} \frac{1}{\tau_i} p(\tau_i|\widehat{\mathbf{w}}_{(t)},\mathbf{y})d\tau_i}{\int_0^{+\infty} p(\tau_i|\widehat{\mathbf{w}}_{(t)},\mathbf{y})d\tau_i} = \frac{\int_0^{+\infty} \frac{1}{\tau_i} p(\tau_i)p(\widehat{\mathbf{w}}_{(t)}|\tau_i)d\tau_i}{\int_0^{+\infty} p(\tau_i)p(\widehat{\mathbf{w}}_{(t)}|\tau_i)d\tau_i}$$
$$= |\widehat{w}_{i,(t)}|^{-2}.$$

Main Results & Bayesian Interpretation

Equivalence between a hierarchy model & $||\mathbf{w}||_0^0$

Proposition 1. *The 2-level hierarchical-Bayes model $p(w_i|\tau_i) = N(w_i|0, \tau_i)$, $p(\tau_i) = 1/\tau_i$, $\tau_i > 0$ over w_i is equivalent to the zero-norm regularized classifier asymptotically.*

Proof Sketch: In the M-step, we maximize

$$-\underbrace{||\mathbf{Hw} - \mathbf{z}||^2}_{\text{Error}} \qquad -\underbrace{\mathbf{w}^T\mathbf{\Lambda w}}_{||w||_0^0,\ \text{if } t \to \infty}$$

$$\because \mathbf{\Lambda}_{ii} = |\widehat{w}_{i,(t)}|^{-2}$$

(obtained in the E-step)

$||\mathbf{w}||_0^0$ & $\mathbf{w}^T\mathbf{\Lambda w}$

Proposition 2. *The prior assumed in zero-norm is only related to the term $\mathbf{w}^T\mathbf{\Lambda w}$ as defined in the EM process, where $\mathbf{\Lambda} = diag(1/\tau_1, \ldots, 1/\tau_l)$, $1/\tau_i$ $(i = 1, \ldots, l)$ can be iteratively updated by $|\widehat{w}_{i,(t)}|^{-2}$ for the zero-norm regularization.*

Main Results & Bayesian Interpretation

Equivalence between a hierarchy model & $||\mathbf{w}||_0^0$

Proposition 1. *The 2-level hierarchical-Bayes model $p(w_i|\tau_i) = N(w_i|0, \tau_i)$, $p(\tau_i) = 1/\tau_i$, $\tau_i > 0$ over w_i is equivalent to the zero-norm regularized classifier asymptotically.*

Proof Sketch: In the M-step, we maximize

$$-\underbrace{||\mathbf{Hw}-\mathbf{z}||^2}_{\text{Error}} \qquad -\underbrace{\mathbf{w}^T\mathbf{\Lambda w}}_{\substack{||w||_0^0,\ \text{if } t\to\infty \\ \because \mathbf{\Lambda}_{ii} = |\widehat{w}_{i,(t)}|^{-2} \\ \text{(obtained in the E-step)}}}$$

$||\mathbf{w}||_0^0$ & $\mathbf{w}^T\mathbf{\Lambda w}$

Proposition 2.*The prior assumed in zero-norm is only related to the term $\mathbf{w}^T\mathbf{\Lambda w}$ as defined in the EM process, where $\mathbf{\Lambda} = diag(1/\tau_1, \ldots, 1/\tau_l)$, $1/\tau_i$ $(i = 1, \ldots, l)$ can be iteratively updated by $|\widehat{w}_{i,(t)}|^{-2}$ for the zero-norm regularization.*

Achieving zero-norm adaptively

Asymptotically True Zero-norm for feature selection

$$\{\mathbf{w}^{(t)}, b^{(t)}\} = \arg\min_{w,b} C \sum_{i=1}^{m} \xi_i + \mathbf{w}^T \Lambda^{(t-1)} \mathbf{w}$$

$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \ldots, l$$

$$\Lambda^{(t)} = diag(1/|w_1^{(t-1)}|^2, \ldots, 1/|w_n^{(t-1)}|^2).$$

- The process is very similar to the EM process–It converges rapidly.
- $\mathbf{w}^T \Lambda^{(t-1)} \mathbf{w}$ iteratively achieves zero-norm
- It is a standard Quadratic Programming problem at each iteration–The whole optimization can be solved in polynomial time.

Achieving zero-norm adaptively

Asymptotically True Zero-norm for feature selection

$$\{\mathbf{w}^{(t)}, b^{(t)}\} = \arg\min_{w,b} C \sum_{i=1}^{m} \xi_i + \mathbf{w}^T \Lambda^{(t-1)} \mathbf{w}$$

$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \dots, l$$

$$\Lambda^{(t)} = diag(1/|w_1^{(t-1)}|^2, \dots, 1/|w_n^{(t-1)}|^2).$$

- The process is very similar to the EM process–It converges rapidly.
- $\mathbf{w}^T \Lambda^{(t-1)} \mathbf{w}$ iteratively achieves zero-norm
- It is a standard Quadratic Programming problem at each iteration–The whole optimization can be solved in polynomial time.

Achieving zero-norm adaptively

Asymptotically True Zero-norm for feature selection

$$\{\mathbf{w}^{(t)}, b^{(t)}\} = \arg\min_{w,b} C \sum_{i=1}^{m} \xi_i + \mathbf{w}^T \Lambda^{(t-1)} \mathbf{w}$$

$$\text{s.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \ldots, l$$

$$\Lambda^{(t)} = diag(1/|w_1^{(t-1)}|^2, \ldots, 1/|w_n^{(t-1)}|^2).$$

- The process is very similar to the EM process–It converges rapidly.
- $\mathbf{w}^T \Lambda^{(t-1)} \mathbf{w}$ iteratively achieves zero-norm
- It is a standard Quadratic Programming problem at each iteration–The whole optimization can be solved in polynomial time.

Reduce Support Vectors in the dual space

Primal space

$\min_{\mathbf{w},b} C\sum_{i=1}^{m}\xi_i + \mathbf{w}^T\Lambda^{(t-1)}\mathbf{w}$
s.t. $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i,$

Target: Feature selection by minimizing $||w||_0^0$
Decision Function:
$f(\mathbf{w}, b) = \mathbf{w}\cdot\mathbf{x} + b$

SV reduction in Dual space

$\min_{\alpha,b} C\sum_{i=1}^{l}\xi_i + \alpha^T\Lambda^{(t-1)}\alpha,$
s.t. $y_i(\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$

Target: SV selection by minimizing $||\alpha||_0^0$
Decision function:
$f(\alpha, b) = \sum_{i=1}^{l}\alpha_i k(\mathbf{x}_i, \mathbf{x}) + b$
Reduce the number of SVs by 10 times while maintaining the accuracy

Reduce Support Vectors in the dual space

Primal space

$\min_{\mathbf{w},b} C\sum_{i=1}^{m}\xi_i + \mathbf{w}^T\Lambda^{(t-1)}\mathbf{w}$

s.t. $\quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i,$

Target: Feature selection by minimizing $||w||_0^0$

Decision Function:

$f(\mathbf{w}, b) = \mathbf{w}\cdot\mathbf{x} + b$

SV reduction in Dual space

$\min_{\alpha,b} C\sum_{i=1}^{l}\xi_i + \alpha^T\Lambda^{(t-1)}\alpha,$

s.t. $\quad y_i(\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$

Target: SV selection by minimizing $||\alpha||_0^0$

Decision function:

$f(\alpha, b) = \sum_{i=1}^{l}\alpha_i k(\mathbf{x}_i, \mathbf{x}) + b$

Reduce the number of SVs by 10 times while maintaining the accuracy

Extensions to arbitrary-norm

$||\mathbf{w}||_p^p$

Proposition 3. *The priors assumed in $||\mathbf{w}||_p^p$ ($0 \leq p \leq 2$ or $p = \infty$) are only related to the term $\mathbf{w}^T \mathbf{\Lambda} \mathbf{w}$ as defined in the EM process, where $\mathbf{\Lambda} = diag(1/\tau_1, \ldots, 1/\tau_l)$, $1/\tau_i$ ($i = 1, \ldots, l$) can be iteratively updated by $\gamma |\widehat{w}_{i,(t)}|^{-(2-p)}$ respectively.*

1. Arbitrary Norm can be achieved without knowing the priors!
2. ∞-norm defined as $||\mathbf{w}||_\infty = \max_i |w_i|$ can be even achieved: $\Lambda = diag(0, \ldots, 0, 1/w_{max,(t)}, 0, \ldots, 0)$ with $w_{max,(t)} = \max_i w_{i,(t)}$

Details can be seen in our Neural Computation 08 paper.

Extensions to arbitrary-norm

$||\mathbf{w}||_p^p$

Proposition 3. *The priors assumed in $||\mathbf{w}||_p^p$ ($0 \leq p \leq 2$ or $p = \infty$) are only related to the term $\mathbf{w}^T \mathbf{\Lambda} \mathbf{w}$ as defined in the EM process, where $\mathbf{\Lambda} = diag(1/\tau_1, \ldots, 1/\tau_l)$, $1/\tau_i$ ($i = 1, \ldots, l$) can be iteratively updated by $\gamma|\widehat{w}_{i,(t)}|^{-(2-p)}$ respectively.*

1. Arbitrary Norm can be achieved without knowing the priors!
2. ∞-norm defined as $||\mathbf{w}||_\infty = \max_i |w_i|$ can be even achieved: $\Lambda = \text{diag}(0, \ldots, 0, 1/w_{i_{max},(t)}, 0, \ldots, 0)$ with $w_{i_{max},(t)} = \max_i w_{i,(t)}$

Details can be seen in our Neural Computation 08 paper.

Extensions to arbitrary-norm

$||\mathbf{w}||_p^p$

Proposition 3. *The priors assumed in* $||\mathbf{w}||_p^p$ $(0 \leq p \leq 2$ *or* $p = \infty)$ *are only related to the term* $\mathbf{w}^T \mathbf{\Lambda} \mathbf{w}$ *as defined in the EM process, where* $\mathbf{\Lambda} = diag(1/\tau_1, \ldots, 1/\tau_l)$, $1/\tau_i$ $(i = 1, \ldots, l)$ *can be iteratively updated by* $\gamma|\widehat{w}_{i,(t)}|^{-(2-p)}$ *respectively.*

1. Arbitrary Norm can be achieved without knowing the priors!
2. ∞-norm defined as $||\mathbf{w}||_\infty = \max_i |w_i|$ can be even achieved: $\Lambda = \mathrm{diag}(0, \ldots, 0, 1/w_{i_{max},(t)}, 0, \ldots, 0)$ with $w_{i_{max},(t)} = \max_i w_{i,(t)}$

Details can be seen in our Neural Computation 08 paper.

Extensions to arbitrary-norm

$||\mathbf{w}||_p^p$

Proposition 3. *The priors assumed in $||\mathbf{w}||_p^p$ ($0 \leq p \leq 2$ or $p = \infty$) are only related to the term $\mathbf{w}^T \mathbf{\Lambda} \mathbf{w}$ as defined in the EM process, where $\mathbf{\Lambda} = diag(1/\tau_1, \ldots, 1/\tau_l)$, $1/\tau_i$ ($i = 1, \ldots, l$) can be iteratively updated by $\gamma |\widehat{w}_{i,(t)}|^{-(2-p)}$ respectively.*

1. Arbitrary Norm can be achieved without knowing the priors!
2. ∞-norm defined as $||\mathbf{w}||_\infty = \max_i |w_i|$ can be even achieved: $\mathbf{\Lambda} = \mathrm{diag}(0, \ldots, 0, 1/w_{i_{max},(t)}, 0, \ldots, 0)$ with $w_{i_{max},(t)} = \max_i w_{i,(t)}$

Details can be seen in our Neural Computation 08 paper.

Extensions to arbitrary-norm

$||\mathbf{w}||_p^p$

Proposition 3. *The priors assumed in $||\mathbf{w}||_p^p$ $(0 \leq p \leq 2$ or $p = \infty)$ are only related to the term $\mathbf{w}^T \mathbf{\Lambda} \mathbf{w}$ as defined in the EM process, where $\mathbf{\Lambda} = diag(1/\tau_1, \ldots, 1/\tau_l)$, $1/\tau_i$ $(i = 1, \ldots, l)$ can be iteratively updated by $\gamma|\widehat{w}_{i,(t)}|^{-(2-p)}$ respectively.*

1. Arbitrary Norm can be achieved without knowing the priors!
2. ∞-norm defined as $||\mathbf{w}||_\infty = \max_i |w_i|$ can be even achieved: $\mathbf{\Lambda} = \text{diag}(0, \ldots, 0, 1/w_{i_{max},(t)}, 0, \ldots, 0)$ with $w_{i_{max},(t)} = \max_i w_{i,(t)}$

Details can be seen in our Neural Computation 08 paper.

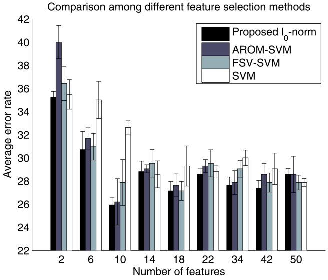
Experimental Setup

- Comparison Algorithms
  - FSV [Bradley et al. 1998]
  - AROM [Weston et al. 2003]
  - SVM

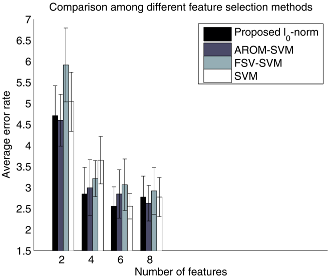
  W
- Data Set
  - Two UCI data
  - Two microarray Gene data
- Data set descriptions

Data set	Dimension	# Sample
Sonar	60	208
Breast	9	683
Colon	2000	62
Lymphoma	4026	96

Accuracy (I)

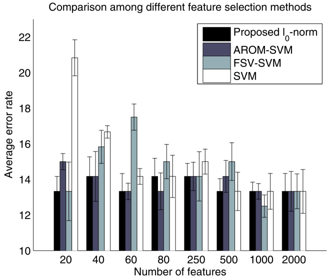


Sonar

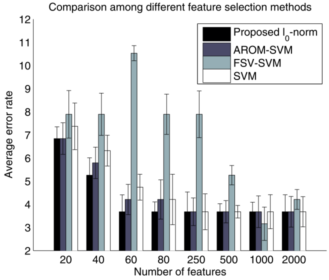


Breast

Accuracy (II)



Colon



Lymphoma

Computational Time

Data Set	Proposed Algorithm	AROM SVM	FSV SVM	SVM
Sonar	0.8061 ± 0.02	6.1431 ± 1.05	2.2888 ± 0.41	0.0146 ± 0.00
Breast	0.3203 ± 0.01	0.6247 ± 0.06	290.4822 ± 13.27	0.0461 ± 0.00
Colon	0.0223 ± 0.00	1.3558 ± 0.29	2.6941 ± 0.25	0.0018 ± 0.00
Lymphoma	0.1766 ± 0.01	2.3809 ± 0.21	23.640 ± 3.16	0.0057 ± 0.00

1. SVM is fastest because it chooses features naively.
2. The proposed algorithm cost much less time than the other two methods.
3. FSV is especially slow in Colon and Lymphoma because it scales against the number of features, while the other three scales against number of samples.

Computational Time

Data Set	Proposed Algorithm	AROM SVM	FSV SVM	SVM
Sonar	0.8061 ± 0.02	6.1431 ± 1.05	2.2888 ± 0.41	0.0146 ± 0.00
Breast	0.3203 ± 0.01	0.6247 ± 0.06	290.4822 ± 13.27	0.0461 ± 0.00
Colon	0.0223 ± 0.00	1.3558 ± 0.29	2.6941 ± 0.25	0.0018 ± 0.00
Lymphoma	0.1766 ± 0.01	2.3809 ± 0.21	23.640 ± 3.16	0.0057 ± 0.00

1. SVM is fastest because it chooses features naively.
2. The proposed algorithm cost much less time than the other two methods.
3. FSV is especially slow in Colon and Lymphoma because it scales against the number of features, while the other three scales against number of samples.

Computational Time

Data Set	Proposed Algorithm	AROM SVM	FSV SVM	SVM
Sonar	0.8061 ± 0.02	6.1431 ± 1.05	2.2888 ± 0.41	0.0146 ± 0.00
Breast	0.3203 ± 0.01	0.6247 ± 0.06	290.4822 ± 13.27	0.0461 ± 0.00
Colon	0.0223 ± 0.00	1.3558 ± 0.29	2.6941 ± 0.25	0.0018 ± 0.00
Lymphoma	0.1766 ± 0.01	2.3809 ± 0.21	23.640 ± 3.16	0.0057 ± 0.00

1. SVM is fastest because it chooses features naively.
2. The proposed algorithm cost much less time than the other two methods.
3. FSV is especially slow in Colon and Lymphoma because it scales against the number of features, while the other three scales against number of samples.

Performance in Dual Space

Data set	Proposed Algorithm		SVM		RVM	
	TSA	#SVs	TSA	#SVs	TSA	#SVs
Twonorm	97.81	16.60	97.70	537.40	97.47	39.20
Titanic	78.82	256.70	78.86	1981.00	77.81	1768.92

- Notes:
  - TSA: Test Set Accuracy
  - RVM: Relevance Vector Machine,a state-of-the-art sparse classifier

Conclusion and Future Work

- Overcome the combinatorially difficult problem & Achieve the direct zero-norm optimization asymptotically
- Computationally efficient
  - can be solved in polynomial time
  - much faster than the approximating methods
- Can be used in dual space for reducing SVs.

Conclusion and Future Work

- Overcome the combinatorially difficult problem & Achieve the direct zero-norm optimization asymptotically
- Computationally efficient
  - can be solved in polynomial time
  - much faster than the approximating methods
- Can be used in dual space for reducing SVs.

Conclusion and Future Work

- Overcome the combinatorially difficult problem & Achieve the direct zero-norm optimization asymptotically
- Computationally efficient
  - can be solved in polynomial time
  - much faster than the approximating methods
- Can be used in dual space for reducing SVs.

Conclusion and Future Work

- Overcome the combinatorially difficult problem & Achieve the direct zero-norm optimization asymptotically
- Computationally efficient
  - can be solved in polynomial time
  - much faster than the approximating methods
- Can be used in dual space for reducing SVs.

Conclusion and Future Work

- Overcome the combinatorially difficult problem & Achieve the direct zero-norm optimization asymptotically
- Computationally efficient
  - can be solved in polynomial time
  - much faster than the approximating methods
- Can be used in dual space for reducing SVs.