Deterministic Annealing for Semi-supervised Kernel Machines

Vikas Sindhwani$^1$, Sathiya Keerthi$^2$, Olivier Chapelle$^3$

$^1$University of Chicago  
$^2$Yahoo! Research  
$^3$Max Planck Institute, Tübingen

ICML 2006
V. Vapnik’s *Transductive* SVM idea

Suppose, for a binary classification problem, we have

- $l$ labeled examples $\{x_i, y_i\}_{i=1}^l$, $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$
- $u$ unlabeled examples $\{x'_j\}_{j=1}^u$

Denote $y' = (y'_1 \ldots y'_u)$ as the unknown labels.

**Train an SVM while optimizing unknown labels**

Solve, over $f \in \mathcal{H}_K : \mathcal{X} \rightarrow \mathcal{R}$ and $y' \in \{-1, +1\}^u$,

$$
\min_{f,y'} \underbrace{\frac{\lambda}{2} \| f \|^2_K}_{\text{regularizer}} + \underbrace{\frac{1}{l} \sum_{i=1}^l V(y_i, f(x_i))}_{\text{labeled loss}} + \underbrace{\frac{\lambda'}{u} \sum_{j=1}^u V(y'_j, f(x'_j))}_{\text{unlabeled loss}}
$$

subject to: $\frac{1}{u} \sum_{j=1}^u \max(0, y'_j) = r$ (positive class ratio)
V. Vapnik’s *Transductive* SVM idea

Suppose, for a binary classification problem, we have

- \( l \) labeled examples \( \{x_i, y_i\}_{i=1}^l, \ x_i \in \mathcal{X}, y_i \in \{-1, +1\} \)
- \( u \) unlabeled examples \( \{x'_j\}_{j=1}^u \)

Denote \( y' = (y'_1 \ldots y'_u) \) as the unknown labels.

Train an SVM while optimizing unknown labels

Solve, over \( f \in \mathcal{H}_K : \mathcal{X} \to \mathcal{R} \) and \( y' \in \{-1, +1\}^u \),

\[
\min_{f,y'} \underbrace{\frac{\lambda}{2} \|f\|_K^2}_{\text{regularizer}} + \underbrace{\frac{1}{l} \sum_{i=1}^l V(y_i, f(x_i))}_{\text{labeled loss}} + \underbrace{\frac{\lambda'}{u} \sum_{j=1}^u V(y'_j, f(x'_j))}_{\text{unlabeled loss}}
\]

subject to:

\[
\frac{1}{u} \sum_{j=1}^u \max(0, y'_j) = r \quad \text{(positive class ratio)}
\]
Equivalent Continuous Optimization Problem

Optimization Problem

\[
\min_{f, y'} J(f, y') = \frac{\lambda}{2} \|f\|_K^2 + \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i)) + \frac{\lambda'}{u} \sum_{j=1}^{u} V(y'_j, f(x'_j))
\]

\[
\min_f J(f) = \frac{\lambda}{2} \|f\|_K^2 + \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i)) + \frac{\lambda'}{u} \sum_{j=1}^{u} \min \left[ V\left(+1, f(x'_j)\right), V\left(-1, f(x'_j)\right) \right]
\]

effective loss

\[
\frac{\lambda'}{u} \sum_{j=1}^{u} \min \left[ V\left(+1, f(x'_j)\right), V\left(-1, f(x'_j)\right) \right]
\]

\[
\left[ \text{effective loss} \right] V'(f(x'_j))
\]
Effective Loss Function Over Unlabeled Examples

(a) Hinge Loss

(b) Quadratic Hinge Loss

(c) Squared Loss

(d) Logistic Loss

- Non-convex
- Penalty if decision surface gets too close to unlabeled examples.
This idea implements a common assumption for SSL...

Low-Density Separation Assumption

The true decision boundary passes through a region containing low volumes of data. Implements the prior knowledge/assumption that

$$\int_{B(f)} P(x) dx$$ is small

where $$B(f) = \{x : |f(x)| < 1\}$$

Cluster Assumption

Points in a data cluster belong to the same class.
### Solution Strategies

#### JTSVM [Joachims, 98]
- Label unlabeled data using supervised SVM. Alternate
  - Optimize $f$ given current $y'$
  - Optimize $y'$ by switching a pair of labels

#### TSVM [Chapelle and Zien, 05]
- Use differentiable losses – quadratic hinge loss over labels and a gaussian loss over unlabeled examples.
- Apply gradient descent.

[Bennett & Demirez, 98], [Fung & Mangasarian, 01], [Collobert, Sinz, Weston, Bottou, 05], [Gartner, Le, Burton, Smola, Vishwanathan, 05]
Non-convexity can hurt empirical performance
Non-convexity can hurt empirical performance
Non-convexity can hurt empirical performance
Non-convexity can hurt empirical performance
Non-convexity can hurt empirical performance
Non-convexity can hurt empirical performance
Non-convexity can hurt empirical performance
Non-convexity can hurt empirical performance

Error rates on COIL6: SVM 21.9, JTSVM 21.2, TSVM 21.6
Deterministic Annealing: Intuition

**Question**

What should the shape of the loss function be so that the decision boundary locally evolves in a desirable manner?

**Key Idea**

Deform the loss function (objective) as the optimization proceeds...somehow!
Deterministic Annealing: Intuition

**Question**

What should the shape of the loss function be so that the decision boundary locally evolves in a desirable manner?

**Key Idea**

Deform the loss function (objective) as the optimization proceeds...somehow!
Deterministic Annealing: Intuition

**Question**

What should the shape of the loss function be so that the decision boundary locally evolves in a desirable manner?

**Key Idea**

Deform the loss function (objective) as the optimization proceeds...somehow!
Deterministic Annealing as a Homotopy Method

- Work with a family of objective functions $J_T$.
- Smoothly deform an “easy” (convex) function $J_{T_1}$ to the given “hard” function $J_{T_2} = J$ by varying $T$.
- Track minimizers along the deformation path.
- DA is a specific implementation of this idea.
Another Equivalent Continuous Optimization Problem

"Relax" $y'$ to $p = (p_1 \ldots p_u)$ where $p_j$ is like the prob that $y'_j = 1$.

$$J(f, p) = E_p J(f, y') = \frac{\lambda}{2} \|f\|_K^2 + \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i))$$

$$+ \frac{\lambda'}{u} \sum_{j=1}^{u} \left[ p_j V(+1, f(x'_j)) + (1 - p_j) V(-1, f(x'_j)) \right]$$

Family of Objective Functions: Avg Cost - T Entropy

$$J_T(f, p) = E_p J(f, y') - \underbrace{T H(p)}_{-\frac{T}{u} \sum_{j=1}^{u} [p_j \log p_j + (1 - p_j) \log (1 - p_j)]}$$
Deterministic Annealing for Semi-supervised SVMs

Full Optimization problem at $T$

$$\min_{f,p} J_T(f, p) = \frac{1}{2} \|f\|_K^2 + \frac{1}{T} \sum_{i=1}^l V(y_i, f(x_i)) +$$

$$\frac{\lambda'}{u} \sum_{j=1}^u \left[ p_j V(+1, f(x'_j)) + (1 - p_j) V(-1, f(x'_j)) \right] +$$

$$\frac{T}{u} \sum_{j=1}^u \left[ p_j \log p_j + (1 - p_j) \log p_j \right] \quad \text{s.t} \quad (1/u) \sum_{j=1}^u p_j = r$$

- **Deformation**: $T$ controls non-convexity of $J(f, p)$. At $T = 0$, reduces to the original non-convex objective function $J(f, p)$.
- **Optimization at** $T$ $(f^*_T, p^*_T) = \arg\min_{f,p} J_T(f, p)$
- **Annealing**: Return: $f^* = \lim_{T \to 0} f^*_T$
- **Balance constraint**: $\frac{1}{u} \sum_{j=1}^u p_j = r$
Alternating Convex Optimization

At any T, optimize f keeping p fixed

- Representer theorem:
  \[ f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) + \sum_{j=1}^{u} \alpha'_j K(x, x'_j) \]
- Minimize weighted regularized loss using standard tricks.

At any T, optimize p keeping f fixed

- \[ p_j^* = \frac{\frac{1}{g_j - \nu}}{1 + e^{-\frac{\nu}{T}}} \quad g_j = \lambda' \left[ V(f(x'_j)) - V(-f(x'_j)) \right] \]
- Obtain \( \nu \) by solving \( \frac{1}{u} \sum_{j=1}^{u} \frac{1}{g_j - \nu} = r \)

Stopping Conditions

- At any T, alternate until \( KL(p_{\text{new}} | p_{\text{old}}) < \varepsilon \). Obtain \( p_T^* \).
- Reduce T, Seed old \( p_T^* \), until \( H(p_T^*) < \varepsilon \).
How effective Loss deforms as a function of $T$

(a) Hinge Loss

(b) Quadratic Hinge Loss

(c) Squared Loss

(d) Logistic Loss
Effective Loss in JTSVM, $\nabla$ T SVM wrt $\lambda'$. 

\[ J_{\lambda'}(f) = \frac{1}{2} \|f\|^2_K + \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i)) + \frac{\lambda'}{u} \sum_{j=1}^{u} V'(f(x'_j)) \]

Unlabeled examples outside the margin do not influence the decision boundary!
Deterministic Annealing: Some Quick Comments

- **Smoothing**: At high $T$, spurious & shallow local min are smoothed away.

- **Simulated Annealing**: Stochastic search allowing “uphill” moves depending on $T$. Associated Markov process converges *slowly* to Gibbs distribution at equilibrium which minimizes $E_p J - TH(p)$ (free energy). As $T \rightarrow 0$ very *slowly*, global solution guaranteed (in prob). DA retains annealing but avoids stochastic search by directly optimizing $E_p J - TH(p)$ for $p$.

- **Maximum Entropy**: $E_p J - TH(p)$ is the Lagrangian of: $\max_p S(p)$ subject to $E_p J = \beta$.

- **Proven Heuristic**: Very strong record of empirical success, including in clustering, classification, compression problems. For SSL, has been applied with EM in [Nigam, 2001].
Deterministic Annealing: Some Quick Comments

- **Smoothing:** At high $T$, spurious & shallow local min are smoothed away.

- **Simulated Annealing:** Stochastic search allowing “uphill” moves depending on $T$. Associated Markov process converges *slowly* to Gibbs distribution at equilibrium which minimizes $E_p \mathcal{J} - TH(p)$ (free energy). As $T \to 0$ *very slowly*, global solution guaranteed (in prob). DA retains annealing but avoids stochastic search by directly optimizing $E_p \mathcal{J} - TH(p)$ for $p$.

- **Maximum Entropy:** $E_p \mathcal{J} - TH(p)$ is the Lagrangian of: $\max_p S(p)$ subject to $E_p \mathcal{J} = \beta$.

- **Proven Heuristic:** Very strong record of empirical success, including in clustering, classification, compression problems. For SSL, has been applied with EM in [Nigam, 2001].
Deterministic Annealing: Some Quick Comments

- **Smoothing:** At high $T$, spurious & shallow local min are smoothed away.

- **Simulated Annealing:** Stochastic search allowing “uphill” moves depending on $T$. Associated Markov process converges *slowly* to Gibbs distribution at equilibrium which minimizes $E_p \mathcal{J} - TH(p)$ (*free energy*). As $T \to 0$ very *slowly*, global solution guaranteed (in prob). DA retains annealing but avoids stochastic search by directly optimizing $E_p \mathcal{J} - TH(p)$ for $p$.

- **Maximum Entropy:** $E_p \mathcal{J} - TH(p)$ is the Lagrangian of: $\max_p S(p)$ subject to $E_p \mathcal{J} = \beta$.

- **Proven Heuristic:** Very strong record of empirical success, including in clustering, classification, compression problems. For SSL, has been applied with EM in [Nigam, 2001].
Deterministic Annealing: Some Quick Comments

- **Smoothing**: At high $T$, spurious & shallow local min are smoothed away.

- **Simulated Annealing**: Stochastic search allowing “uphill” moves depending on $T$. Associated Markov process converges slowly to Gibbs distribution at equilibrium which minimizes $E_p J - TH(p)$ (*free energy*). As $T \to 0$ very slowly, global solution guaranteed (in prob). DA retains annealing but avoids stochastic search by directly optimizing $E_p J - TH(p)$ for $p$.

- **Maximum Entropy**: $E_p J - TH(p)$ is the Lagrangian of: $\max_p S(p)$ subject to $E_p J = \beta$.

- **Proven Heuristic**: Very strong record of empirical success, including in clustering, classification, compression problems. For SSL, has been applied with EM in [Nigam, 2001].
First Experiments

Successes in 10 trials. $l=2$

<table>
<thead>
<tr>
<th>Dataset → Algorithm ↓</th>
<th>2MOONS</th>
<th>2CIRCLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>JTSVM ($l_2$)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>JTSVM ($l_1$)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\nabla$TSMV</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>DA ($l_2$)</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>DA ($l_1$)</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Used RBF Kernels, Optimal parameters.
Real-world Datasets

\( \lambda' = 1; \lambda, \sigma \) optimized; avg. over 10 random splits. \( T = \frac{10}{1.5^i} \) \( i = 0, 1, \ldots \)

<table>
<thead>
<tr>
<th>Unlab</th>
<th>USPS2</th>
<th>COIL6</th>
<th>PCMAC</th>
<th>ESET2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>7.5</td>
<td>21.5</td>
<td>18.9</td>
<td>19.4</td>
</tr>
<tr>
<td>JTSVM</td>
<td>7.6</td>
<td>19.9</td>
<td>10.4</td>
<td>9.2</td>
</tr>
<tr>
<td>( \triangledown )TSVM</td>
<td>6.9</td>
<td>21.4</td>
<td>5.4</td>
<td>8.7</td>
</tr>
<tr>
<td>DA((l^2))</td>
<td>6.4</td>
<td>13.6</td>
<td>5.3</td>
<td>8.1</td>
</tr>
<tr>
<td>DA((sqr))</td>
<td>5.7</td>
<td>13.8</td>
<td>5.4</td>
<td>9.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>7.8</td>
<td>21.9</td>
<td>17.9</td>
<td>19.7</td>
</tr>
<tr>
<td>JTSVM</td>
<td>7.2</td>
<td>21.2</td>
<td>7.0</td>
<td>8.9</td>
</tr>
<tr>
<td>( \triangledown )TSVM</td>
<td>7.1</td>
<td>21.6</td>
<td>4.5</td>
<td>9.1</td>
</tr>
<tr>
<td>DA((l^2))</td>
<td>6.3</td>
<td>15.0</td>
<td>4.8</td>
<td>8.5</td>
</tr>
<tr>
<td>DA((sqr))</td>
<td>6.3</td>
<td>15.2</td>
<td>4.7</td>
<td>9.4</td>
</tr>
</tbody>
</table>
Large Scale Text Categorization

UseNet articles from two discussion groups: Auto-vs-Aviation. Used special primal routines for linear kernels, [Keerthi and Decoste, 2005]. More results in [SK,SIGIR 06] #features=20707, #training=35543, #test=35587.
Large Scale Text Categorization

UseNet articles from two discussion groups: Auto-vs-Aviation. Used special primal routines for linear kernels, [Keerthi and Decoste, 2005]. More results in [SK,SIGIR 06] #features=20707, #training=35543, #test=35587.

![Graph showing test error vs. number of labeled examples]

- Test error
- Number of Labeled Examples
UseNet articles from two discussion groups: Auto-vs-Aviation. Used special primal routines for linear kernels, [Keerthi and Decoste, 2005]. More results in [SK, SIGIR 06] #features=20707, #training=35543, #test=35587.
## Importance of Annealing

<table>
<thead>
<tr>
<th></th>
<th>USPS2</th>
<th>COIL6</th>
<th>PC-MAC</th>
<th>ESET2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unlab</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DA</td>
<td>6.4</td>
<td>13.6</td>
<td>5.3</td>
<td>8.1</td>
</tr>
<tr>
<td>T=0.1</td>
<td>6.6</td>
<td>20.0</td>
<td>5.7</td>
<td>7.8</td>
</tr>
<tr>
<td>T=0.01</td>
<td>7.6</td>
<td>20.1</td>
<td>7.1</td>
<td>8.1</td>
</tr>
<tr>
<td>T=0.001</td>
<td>7.9</td>
<td>20.3</td>
<td>9.1</td>
<td>8.8</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DA</td>
<td>6.3</td>
<td>15.0</td>
<td>4.8</td>
<td>8.5</td>
</tr>
<tr>
<td>T=0.1</td>
<td>6.8</td>
<td>21.0</td>
<td>4.7</td>
<td>8.0</td>
</tr>
<tr>
<td>T=0.01</td>
<td>7.0</td>
<td>21.3</td>
<td>5.7</td>
<td>8.5</td>
</tr>
<tr>
<td>T=0.001</td>
<td>7.2</td>
<td>21.5</td>
<td>7.3</td>
<td>8.8</td>
</tr>
</tbody>
</table>
Summary and Open Questions

Summary

- New optimization method that better approaches global solution for TSVM-like SSL.
- “Easy” to “Hard” approach.
- Can use off-the-shelf optimization subroutines.

Open Questions

- Intriguing connections between annealing behaviour, loss function and regularization.
- Annealing sequence? Detailed experimental studies.

Also see: A Continuation method for Semi-supervised SVMs, O. Chapelle, M. Chi, A. Zien, ICML 2006.