Maximum Margin Semi-supervised Learning with Irrelevant Data

Irwin King with Haiqin Yang

king@cse.cuhk.edu.hk
http://www.cse.cuhk.edu.hk/~king

Department of Computer Science & Engineering
The Chinese University of Hong Kong
Example Datasets

USPS

MNIST

Maximum Margin Semi-supervised Learning with Irrelevant Data, Irwin King, MLA'09, Nanjing, China, November 6-8, 2009
Other Applications

- Website categorization
  - Classify “sports news” vs. “financial news” with crawling
- Medical diagnosis
  - Classify one disease against another with the presence of other disease types
- Binary classification in multi-class classification tasks
Why Semi-supervised Learning?

• **Labeling** data are rare, costly, and time consuming to obtain

• Many **unlabeled** data are easy to collect and may provide useful information

Consider to learn from both **labeled** and **unlabeled** data simultaneously!
Support Vector Machines (SVMs)
S$^3$VMs w/ Clean Unlabeled Data

Data

S$^3$VM Illustration

Maximum Margin Semi-supervised Learning with Irrelevant Data, Irwin King, MLA’09, Nanjing, China, November 6-8, 2009
**S³VM w/ Unclean Unlabeled Data**

**Data**

- +1 Class
- -1 Class
- Unlabeled Data

**S³VM Illustration**

- +1 Class
- -1 Class
- Unlabeled Data
- S³VM
Classifiers

Maximum Margin Semi-supervised Learning with Irrelevant Data, Irwin King, MLA’09, Nanjing, China, November 6-8, 2009
SSL Assumptions

• Unlabeled data are from the same distribution as the labeled data

• What if they are not?

• Unlabeled data may be a mixture of relevant and irrelevant data!

• How can we utilize this irrelevant data information to improve performance in SSL?
Setup of Tri-Class SVM (3C-SVM)

\[ \mathcal{L} = \{(x_i, y_i)\}_{i=1}^L \]
\[ x_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \{-1, 0, 1\} \]
\[ \mathcal{U} = \mathcal{U}_R \cup \mathcal{U}_0 = \{x_i\}_{i=1}^U \]

**Objective:** seek

\[ f_\vartheta(x) = w^T \phi(x) + b, \vartheta = (w, b) \]

to separate the binary class data correctly with the help of (mixed) unlabeled data.
Function Definition

• **Objective function:**

$$\min_{\vartheta} \frac{1}{2} \| w \|^2 + \sum_{x_i \in L} r_i \ell_L(f_\vartheta(x_i), y_i) + \sum_{x_i \in U} r_i \ell_U(f_\vartheta(x_i)),$$

**Margin**

**Empirical Risk**

**Labeled Data**

**Empirical Risk**

**Unlabeled Data**

• **Facts:** if $f_\vartheta(x_i) \gg 0$, more confident on $+1$-class

**Facts:** if $f_\vartheta(x_i) \ll 0$, more confident on $-1$-class

• **Principle:** rely more on labeled and relevant data (symmetrical hinge loss)

**Principle:** ignore irrelevant data ($\epsilon$-insensitive loss)
Function Definition

\[
\min_\vartheta \quad \frac{\lambda}{2} \| \mathbf{w} \|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}} r_i \ell_L(f_\vartheta(\mathbf{x}_i), y_i) + \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \ell_U(f_\vartheta(\mathbf{x}_i)) ,
\]

\[
\min_\vartheta \quad \frac{\lambda}{2} \| \mathbf{w} \|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}_{\pm 1}} r_i H_1(y_i f_\vartheta(\mathbf{x}_i)) + \sum_{\mathbf{x}_i \in \mathcal{L}_0} r_i I_\varepsilon(f_\vartheta(\mathbf{x}_i)) \\
+ \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \min\{H_1(\|f_\vartheta(\mathbf{x}_i)\|), I_\varepsilon(\|f_\vartheta(\mathbf{x}_i)\|)\} .
\]

\[H_1(z) = \max\{0, 1 - z\}, \quad I_\varepsilon(z) = \max\{0, |z| - \varepsilon\}.\]
Loss Functions

- Hinge loss
- Symmetrical hinge loss
- $\varepsilon$-insensitive loss ($\varepsilon=0.1$)
Our Loss Function

- Hinge loss
- Symmetrical hinge loss
- $\varepsilon$-insensitive loss ($\varepsilon=0.1$)

Unlabeled Data
Labeled Data

Maximum Uncertainty

Unlabeled Data
Labeled Data

Maximum Margin Semi-supervised Learning with Irrelevant Data, Irwin King, MLA’09, Nanjing, China, November 6-8, 2009
Model Relationships

\[
\min_{\vartheta} \frac{\lambda}{2} \|w\|^2 + \sum_{x_i \in \mathcal{L}_{\pm 1}} r_i H_1(y_i f_{\vartheta}(x_i)) + \sum_{x_i \in \mathcal{L}_0} r_i I_\varepsilon(f_{\vartheta}(x_i)) + \sum_{x_i \in \mathcal{U}} r_i \min \{ H_1(||f_{\vartheta}(x_i)||), I_\varepsilon(||f_{\vartheta}(x_i)||) \}.
\]

### 3C-SVM

<table>
<thead>
<tr>
<th>$\mathcal{L}$</th>
<th>$-1$</th>
<th>$0$</th>
<th>$1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{U}$</td>
<td>$-1$</td>
<td>$0$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

### SVM

<table>
<thead>
<tr>
<th>$\mathcal{L}$</th>
<th>$-1$</th>
<th>$1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{U}$</td>
<td>$-1$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

### $S^3$VM

<table>
<thead>
<tr>
<th>$\mathcal{L}$</th>
<th>$-1$</th>
<th>$1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{U}$</td>
<td>$-1$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

### $\mathcal{U}$-SVM

<table>
<thead>
<tr>
<th>$\mathcal{L}$</th>
<th>$-1$</th>
<th>$0$</th>
<th>$1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{U}$</td>
<td>$-1$</td>
<td>$1$</td>
<td></td>
</tr>
</tbody>
</table>
Theorem

- Objective function:

\[
\min_{\vartheta} \lambda \frac{1}{2} \| w \|^2 + \sum_{x_i \in \mathcal{L}_{\pm 1}} r_i H_1(y_i f_{\vartheta}(x_i)) + \sum_{x_i \in \mathcal{L}_0} r_i I_{\varepsilon}(f_{\vartheta}(x_i)) \\
+ \sum_{x_i \in \mathcal{U}} r_i \min\{H_1(|f_{\vartheta}(x_i)|), I_{\varepsilon}(|f_{\vartheta}(x_i)|)\}.
\]

- 3C-SVM

Suppose \( r_i = \infty \) for unlabeled data and \( \varepsilon = 0 \).

Unlabeled data \( x_j \) satisfies

(a) \( |w^T \phi(x_j) + b| \geq 1 \Rightarrow \) data lie on or out of the margin gap,

or

(b) \( w^T \phi(x_j) + b = 0 \Rightarrow w^T (\phi(x_j) - \phi(x_0)) = 0, \ x_j, x_0 \in \mathcal{U}_0 \)
Removing Min-Terms

\[
\min_{\vartheta, d} \frac{\lambda}{2} \|w\|^2 + \sum_{x_i \in L_{\pm 1}} r_i H_1(y_i f_\vartheta(x_i)) + \sum_{x_i \in L_0} r_i I_\varepsilon(f_\vartheta(x_i))
\]

\[
+ \sum_{x_{k+L} \in U} r_{k+L} \left( \underbrace{H_1(|f_\vartheta(x_i)| + D(1-d_k))}_{Q_1} + \underbrace{I_\varepsilon(|f_\vartheta(x_i)| - Dd_k)}_{Q_2} \right),
\]

- \( d_k = 0 \Rightarrow Q_1 = 0 \),
- \( d_k = 1 \Rightarrow Q_2 = 0 \),
- \( H_1(|z| + a) \): non-convex, approximated by ramploss,
  \( H_{1-a}(z) - H_\kappa(z) + H_{1-a}(-z) - H_\kappa(-z) \),
- \( I_\varepsilon(|z| - b) = H_{-\varepsilon-b}(-z) + H_{-\varepsilon-b}(z) \),
- \( H_1(|z| + a) \) and \( I_\varepsilon(|z| - b) \) are symmetrical loss.
Concave-Convex Procedure

- Objective function: \( Q^\kappa(\vartheta, d) = Q_{\text{vex}}(\vartheta, d) + Q_{\text{cav}}^\kappa(\vartheta) \)

- Each step

\[
\vartheta^{t+1} = \arg \min_{\vartheta} \left( Q_{\text{vex}}(\vartheta, d^t) + \frac{\partial Q_{\text{cav}}^\kappa(\vartheta^t)}{\partial \vartheta} \cdot \vartheta \right),
\]

\[
\begin{align*}
\max_{\alpha, \alpha^*} & -\frac{1}{2} \| w(\alpha, \alpha^*) \|^2 + \varrho(\alpha, \alpha^*) \\
\text{s.t.} & \quad A_e[\alpha; \alpha^*] = \mu^T Y \cdot U, \\
& \quad A[\alpha; \alpha^*] \leq 0, \\
& \quad 0 \leq \alpha, \alpha^* \leq r.
\end{align*}
\]

\[
d_k = \begin{cases} 
1 & \text{if } \xi_k \leq \xi_k^* \\
0 & \text{otherwise}
\end{cases}, \quad \xi_k = H_1(|f_\vartheta(x_{k+L})|), \quad \xi_k^* = I_\varepsilon(|f_\vartheta(x_{k+L})|), \quad k=1,\ldots,U.
\]
Algorithm 1: CCCP for 3C-SVMs

Initialization:
\[ t = 0; \]
Calculate \( \vartheta^0 = (w^0, b^0) \) from a \( \mathcal{U} \)-SVM solution on the labeled/unlabeled data;
Compute
\[ \mu_i^0 = \begin{cases} r_i & \text{if } y_i f_{\vartheta^0} (x_i) < \kappa \text{ and } i \geq L + 1; \\ 0 & \text{otherwise} \end{cases} \]
repeat
\[ t \leftarrow t + 1; \]
Solve the optimization in (6) to obtain \( \vartheta^t \);
Update \( d^t \) from (4);
Update \( \mu^t \) from (5);
if \( Q^\kappa (\vartheta^t, d^t) > Q^\kappa (\vartheta^{t-1}, d^{t-1}) \) then
\[ \text{Let } d^t = d^{t-1}; \]
Solve the optimization in (6) to obtain \( \vartheta^t \) by fixing \( d^{t-1} \);
Update \( \mu^t \) from (5);
end if
until \( |\mu^{t+1} - \mu^t| \leq \epsilon. \)
3C-SVM Demo
3C-SVM Result

Demo for 3C-SVM

- +1 Class
- -1 Class
- Unlabeled
- 3C-SVM
- 0 Class

Maximum Margin Semi-supervised Learning with Irrelevant Data, Irwin King, MLA’09, Nanjing, China, November 6-8, 2009
Experimental Set-up

Comparing Algorithms

- SVMs
- S³VMs
- U-SVMs
- 3C-SVMs

Platform Used

- Matlab 7.3
- MOSEK 5.0
Data Generation

- Follow scheme from Sinz et al., 2008.
- $\pm 1$-class: $c_i^{\pm} = \pm 0.3$, $i = 1, \ldots, 50$, $\sigma_{1,2}^2 = 0.08$, $\sigma_{3,\ldots,50}^2 = 10$.
- Two Gaussians with the Bayes risk being approximately 5%.
- First $U_0$: zero mean, $\sigma_{1,2}^2 = 0.1$, $\sigma_{3,\ldots,50}^2 = 10$.
- Second $U_0$: variance values are the same as $\pm 1$-class data, mean is $t \cdot c^+$, $t = 0.5$. 
Test Procedure

- $L = 20, 50, 200, 500$
- $U = 500 = (\tau U, (1 - \tau)U)$, $\tau = 0.1, 0.5, 0.9$
- Labeled + Unlabeled/500 Test, ten-run average
- Hyperparameters
  - Linear kernel
  - Regularized parameters, forward tuning
    
    \[
    \begin{array}{ccccc}
    & C_L & C_U & \varepsilon & \kappa \\
    \hline
    SVM & \checkmark & \times & \times & \times \\
    U-SVM & - & \checkmark & \checkmark & \times \\
    \end{array}
    \]
  - Further tune on $S^3$VM
  - 3C-SVM uses the same parameters of other models
Accuracy

![Graphs showing accuracy over the number of labeled data points for different values of $\sigma^2$ and $\tau$. The graphs compare SVM, $S^3$VM, U-SVM, and 3C-SVM.]
Real-World Data Description

- Datasets:
  - Small size: USPS
  - Large size: MNIST

- Setup
  - ±1-class: Digits “5” and “8”
  - $U_0$: Other digits
  - $L = 20$
  - $U = 500 = (\tau U, (1 - \tau)U)$, $\tau = 0.1, 0.5, 0.9$
  - RBF kernel: $K(x, y) = \exp(-\gamma \|x - y\|^2)$, $\gamma = \frac{1}{0.3d}$
  - Other hyperparameters are set similar to those in the synthetic datasets
## Accuracy Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Algorithm</th>
<th>$\tau = 0.1$</th>
<th>$\tau = 0.5$</th>
<th>$\tau = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>USPS</td>
<td>SVM</td>
<td>72.4± 15.9 (0.7)</td>
<td>72.4± 15.9 (9.5)</td>
<td>72.4± 15.9 (53.1)</td>
</tr>
<tr>
<td></td>
<td>S$^3$VM</td>
<td>63.6± 8.9 (0.0)</td>
<td>68.2± 8.0 (2.2)</td>
<td>73.2± 7.0 (9.5)</td>
</tr>
<tr>
<td></td>
<td>U-SVM</td>
<td>83.1± 2.5 (0.0)</td>
<td>73.4± 4.4 (0.0)</td>
<td>64.2± 3.6 (0.0)</td>
</tr>
<tr>
<td></td>
<td>3C-SVM</td>
<td>87.2±2.3</td>
<td>80.6±4.8</td>
<td>75.4±7.3</td>
</tr>
<tr>
<td>MNIST</td>
<td>SVM</td>
<td>70.9± 11.4 (0.3)</td>
<td>70.9± 11.4 (0.8)</td>
<td>70.9± 11.4 (13.6)</td>
</tr>
<tr>
<td></td>
<td>S$^3$VM</td>
<td>70.9± 10.5 (0.7)</td>
<td>72.4± 10.1 (1.0)</td>
<td>75.7± 9.1 (9.8)</td>
</tr>
<tr>
<td></td>
<td>U-SVM</td>
<td>84.2± 2.2 (0.2)</td>
<td>80.0± 4.6 (0.9)</td>
<td>75.0± 3.9 (1.0)</td>
</tr>
<tr>
<td></td>
<td>3C-SVM</td>
<td>85.3±1.6</td>
<td>82.8±2.9</td>
<td>77.6±3.9</td>
</tr>
</tbody>
</table>
Accuracy on Detecting 0-class

Accuracy vs. \( \varepsilon \) on USPS

FPR vs. TPR on USPS

Accuracy vs. \( \varepsilon \) on MNIST

FPR vs. TPR on MNIST

Maximum Margin Semi-supervised Learning with Irrelevant Data, Irwin King, MLA'09, Nanjing, China, November 6-8, 2009
Conclusions

• A novel maxi-margin classifier, 3C-SVM, can distinguish data into -1, +1, and 0, three categories.

• The model incorporates standard SVMs, S^3VMs, and U-SVMs as specific cases.

• Introduce a min-loss function which combines loss for relevant and irrelevant data.

• Present transformations so it can be solved by CCCP, a high efficiency algorithm.

• Effectiveness and efficiency are demonstrated through synthetic and real-world data.
Future Work

• Model speed-up
• Multi-class extension
• Theoretical analysis, generalization bound, etc.
References


On-Going Research

Machine Learning

- Heavy-Tailed Symmetric Stochastic Neighbor Embedding (NIPS’09)
- Adaptive Regularization for Transductive Support Vector Machine (NIPS’09)
- Direct Zero-norm Optimization for Feature Selection (ICDM’08)
- Semi-supervised Learning from General Unlabeled Data (ICDM’08)
- Learning with Consistency between Inductive Functions and Kernels (NIPS’08)
- An Extended Level Method for Efficient Multiple Kernel Learning (NIPS’08)
- Semi-supervised Text Categorization by Active Search (CIKM’08)
- Transductive Support Vector Machine (NIPS’07)
- Global and local learning (ICML’04, JMLR’04)
On-Going Research

Web Intelligence/Information Retrieval

• A Generalized Co-HITS Algorithm and Its Application to Bipartite Graphs (KDD’09)
• Entropy-biased Models for Query Representation on the Click Graph (SIRIR’09)
• Effective Latent Space Graph-based Re-ranking Model with Global Consistency (WSDM’09)
• Formal Models for Expert Finding on DBLP Bibliography Data (ICDM’08)
• Learning Latent Semantic Relations from Query Logs for Query Suggestion (CIKM’08)
• RATE: a Review of Reviewers in a Manuscript Review Process (WI’08)
• MatchSim: link-based web page similarity measurements (WI’07)
• Diffusion rank: Ranking web pages based on heat diffusion equations (SIGIR’07)
• Web text classification (WWW’07)
On-Going Research

**Recommender Systems/Collaborative Filtering**

- Learning to Recommend with Social Trust Ensemble (SIRIR’09) Semi-Nonnegative Matrix Factorization with Global Statistical Consistency in Collaborative Filtering (CIKM’09)

- Recommender system: accurate recommendation based on sparse matrix (SIGIR’07)

- SoRec: Social Recommendation Using Probabilistic Matrix Factorization (CIKM’08)

**Human Computation**

- An Analytical Study of Puzzle Selection Strategies for the ESP Game (WI’08)

- An Analytical Approach to Optimizing The Utility of ESP Games (WI’08)
Acknowledgments

• Prof. Michael R. Lyu

• Dr. Hongbo Deng (Ph.D.)

• Baichuan Li (Ph.D.)

• Zhenjiang Lin (Ph.D.)

• Hao Ma (Ph.D.)

• Mingzhen Mo (M.Phil.)

• Dingyan Wang (M.Phil.)

• Wei Wang (M.Phil.)

• Haiqin Yang (Ph.D.)

• Connie Yuen (Ph.D.)

• Xin Xin (Ph.D.)

• Chao Zhou (Ph.D.)

• Yi Zhu (Ph.D.)
VeriGuide

- Similarity text detection system
- Developed at CUHK
- Promote and uphold academic honesty, integrity, and quality
- Support English, Traditional and Simplified Chinese
- Handle .doc, .txt, .pdf, .html, etc. file formats
- Generate detailed originality report including readability
VeriGuide Free Trial

http://www.cse.cuhk.edu.hk/~king

Maximum Margin Semi-supervised Learning with Irrelevant Data, Irwin King, MLA’09, Nanjing, China, November 6-8, 2009
Q & A

http://www.cse.cuhk.edu.hk/~king