Research on Semi-Supervised SVMs

(半监督支持向量机的研究)

Yu-Feng Li

National Key Laboratory for Novel Software Technology,

Nanjing University, China

URL: http://lamda.nju.edu.cn/liyf/

Email: liyf@lamda.nju.edu.cn

南京大學

NANJING UNIVERSITY 1902

LAMDA

Learning And Mining from DatA

http://lamda.nju.edu.cn

MLA’13, Shanghai

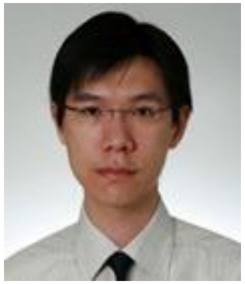
Joint work with

Zhi-Hua Zhou

Nanjing University

James Kwok

Hong Kong University
of Science and Technology



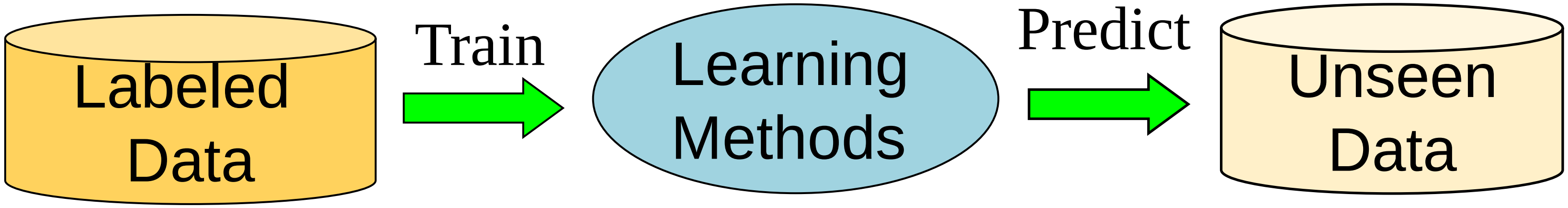
Ivor Tsang

Nanyang Technological
University

Acknowledge

Teng Zhang, Linli Xu and Kai Zhang

Supervised Learning



Labeled Data

Train

Learning Methods

Predict

Unseen Data

In order to have a good generalization performance, supervised learning methods often assumes that a large amount of labeled data are available.

Labeled Data Is Expensive

- However, labeling process is **expensive** in many real tasks
  - Disease diagnosis
  - Drug detection
  - Image classification
  - Text categorization
  - …

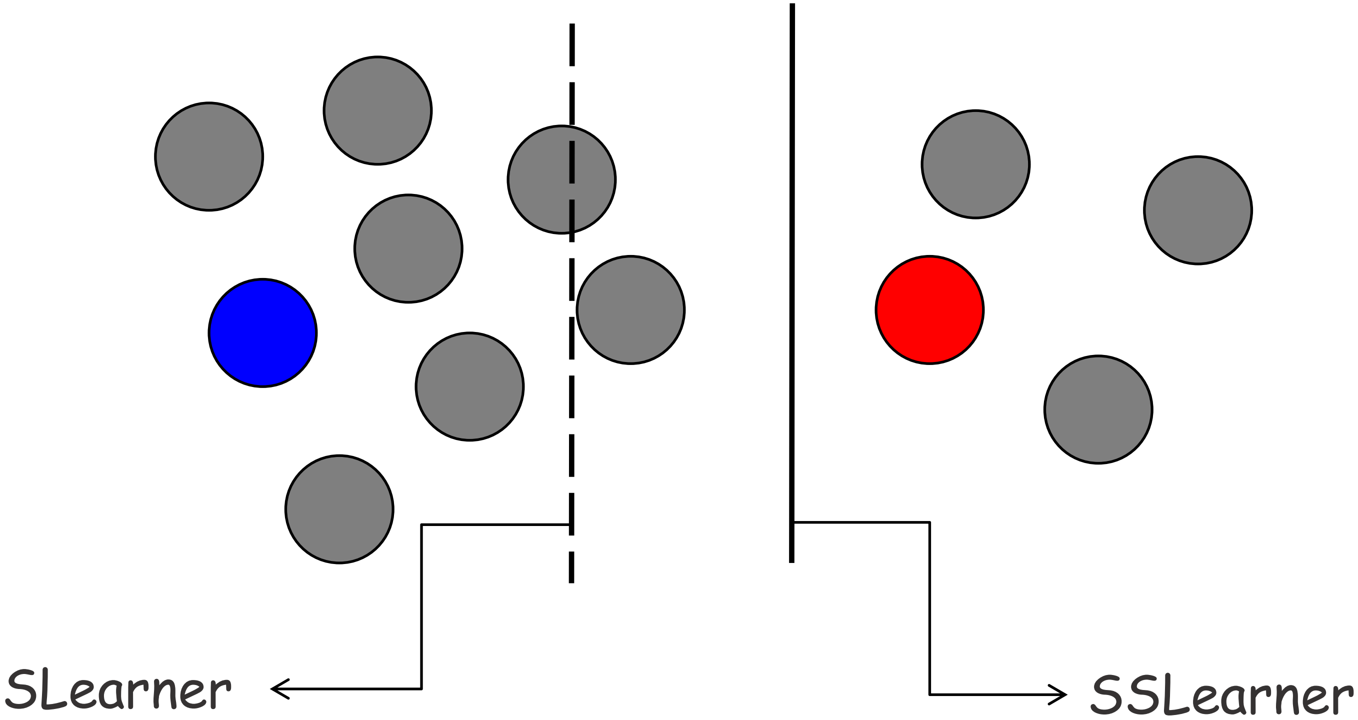
Human efforts and material resources

Exploiting Unlabeled Data

- Collection of unlabeled data is usually cheaper

- Two popular schemes for exploiting unlabeled data to help supervised learning
  - Semi-supervised learning: the learner tries to exploit the unlabeled examples by itself.
  - **Active learning**: the learner actively selects some unlabeled examples to query from an oracle

Semi-Supervised Learning



- Several Surveys and Books
  - O. Chapelle et al. *Semi-supervised learning*. MIT Press Cambridge, 2006.
  - X. Zhu and A. Goldberg. *Introduction to semi-supervised learning*. Morgan & Claypool Publishers, 2009.
  - Z.-H. Zhou and M. Li. *Semi-supervised learning by disagreement*. Knowledge and Information Systems, 24(3):415–439, 2010.
  - 周志华. 基于分歧的半监督学习, 特邀综述. 自动化学报. 2013年11月.

Four Major Paradigms of SSL

- **Generative methods** [Miller & Uyar, 1997; Nigam et al., 2000; Cozman & Cohen, 2002]
- **Co-training/Disagreement-based methods** [Blum & Mitchell, 1998; Balcan et al., 2005; Zhou & Li, 2010]

The seminal work [Blum & Mitchell, 1998] has won the '10-year best paper' award in the 25th International Conference on Machine Learning (ICML'08).

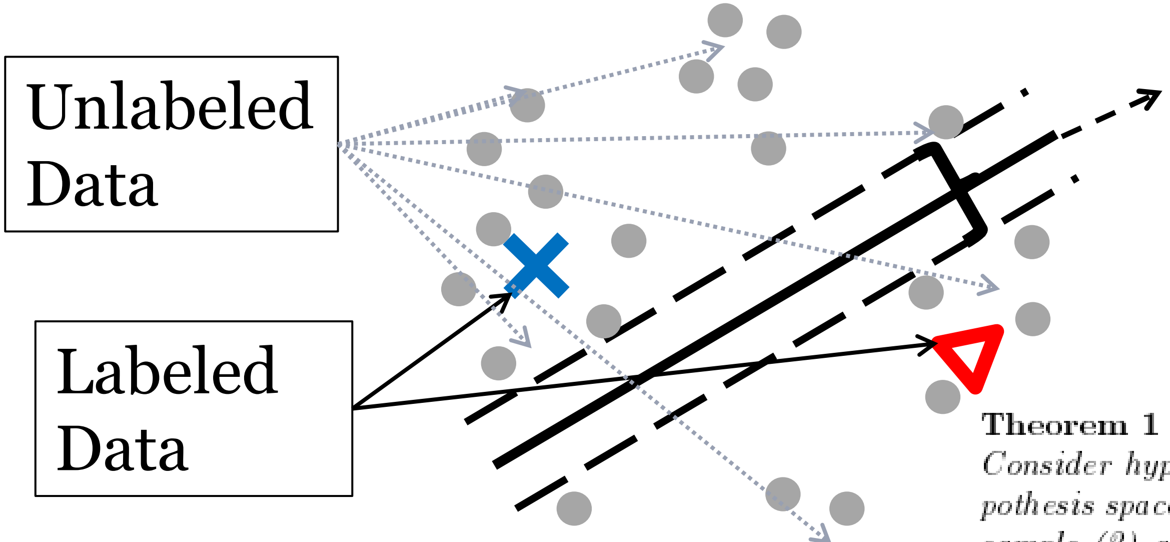
- **Graph-based methods** [Blum & Chawla, 2001; Zhu et al., 2003; Zhou et al., 2005; Belkin et al., 2006]

The seminal work [Zhu et al., 2003] has won the '10-year best paper' award in the 30th International Conference on Machine Learning (ICML'13).

- **Semi-supervised support vector machines (S3VMs)** [Vapnik, 1998; Bennett & Demiriz, 1999; Joachims, 1999; Chapelle & Zien, 2005]

The seminal work [Joachims, 1999] has won the '10-year best paper' award in the 26th International Conference on Machine Learning (ICML'09).

S3VMs

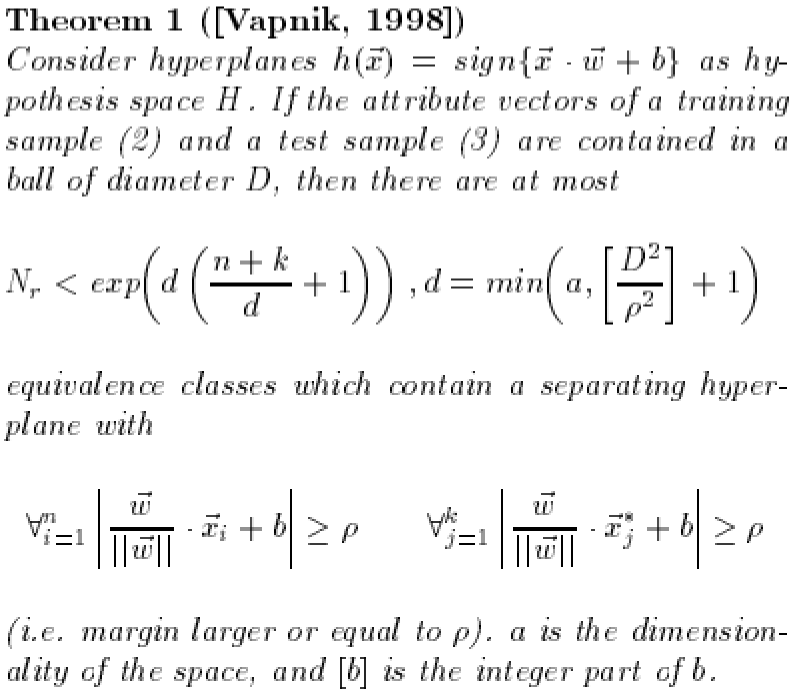


Unlabeled Data

Labeled Data

Large-margin separator
(or, low-density separator)

In [Vapnik, SLT'98], it is shown that large margin could help improve the generalization learning bound.



Theorem 1 **([Vapnik, 1998])**
Consider hyperplanes $h(\vec{x}) = sign\{\vec{x} \cdot \vec{w} + b\}$ *as hypothesis space* H. *If the attribute vectors of a training sample (2) and a test sample (3) are contained in a ball of diameter* D, *then there are at most*

$$N_r < exp\left(d\left(\frac{n+k}{d}+1\right)\right), d = min\left(a, \left[\frac{D^2}{\rho^2}\right]+1\right)$$

equivalence classes which contain a separating hyperplane with

$$\forall_{i=1}^{n} \left|\frac{\vec{w}}{||\vec{w}||} \cdot \vec{x}_i + b\right| \geq \rho \qquad \forall_{j=1}^{k} \left|\frac{\vec{w}}{||\vec{w}||} \cdot \vec{x}_j^* + b\right| \geq \rho$$

(i.e. margin larger or equal to ρ*).* a *is the dimensionality of the space, and* $[b]$ *is the integer part of* b.

S3VMs: Formulation

Control model complexity

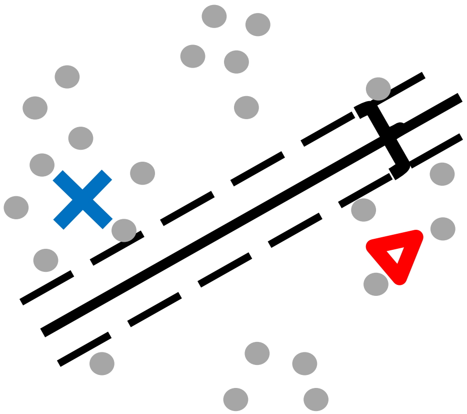
Losses on labeled and unlabeled data

$$\begin{aligned}
\min_{\hat{y}_{l+1},\ldots,\hat{y}_N} \quad \min_{\mathbf{w},\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{j=l+1}^{N}\xi_j \\
\text{s.t.} \quad & y_i\mathbf{w}'\mathbf{x}_i \geq 1-\xi_i,\ \ \xi_i \geq 0. \\
& \hat{y}_j\mathbf{w}'\mathbf{x}_j \geq 1-\xi_j,\ \ \xi_j \geq 0. \\
& \hat{y}_j \in \{+1,-1\}. \\
& -\beta \leq \frac{\sum_{j=l+1}^{N}\hat{y}_j}{N-l} - \frac{\sum_{i=1}^{l}y_i}{l} \leq \beta. \\
& i = 1,\ldots,l,\ j = l+1,\ldots,N.
\end{aligned}$$

Both labeled and unlabeled data have large margin

The label of unlabeled data are unknown, and need to be optimized

Balance constraint



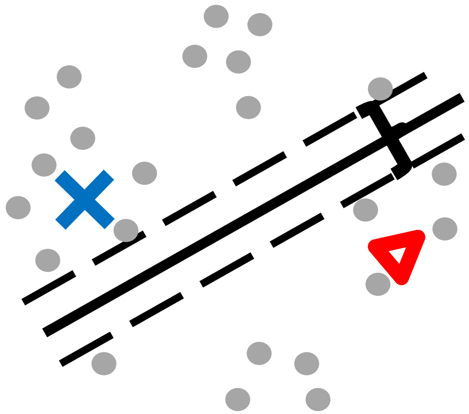
S3VMs: Formulation

$$\boxed{\min_{\hat{\mathbf{y}}\in\mathcal{B}}}\ \boxed{\min_{\mathbf{w}}\ \Omega(\mathbf{w}) + C_1\sum_{i=1}^{l}\ell(\mathbf{w},\mathbf{x}_i,y_i) + C_2\sum_{j=l+1}^{N}\ell(\mathbf{w},\mathbf{x}_j,\hat{y}_j)}$$

SVM

Prior knowledge

S3VMs are an mixed-integer program, thus intractable in general.



S3VMs: Applications

- Text Categorization [Joachims 1999; Joachims, 2002]
- Email Classification [Kockelkorn et al., 2003]
- Image Retrieval [Wang et al., 2003]
- Bioinformatics [Kasabov & Pang, 2004]
- Named Entity Recognition [Goutte et al., 2002]
- ...

Outline

- Scalability of S3VMs “多”
  - WellSVM [Li et al., JMLR13]

- Efficiency of S3VMs “快”
  - MeanS3VM [Li et al., ICML09]

- Safeness of S3VMs “好”
  - S4VM [Li and Zhou, ICML11]

- Cost sensitivity of S3VMs “省”
  - CS4VM [Li et al., AAAI10]

Outline

- Scalability of S3VMs “多”
  - WellSVM [Li et al., JMLR13]
- Efficiency of S3VMs “快”
  - MeanS3VM [Li et al., ICML09]
- Safeness of S3VMs “好”
  - S4VM [Li and Zhou, ICML11]
- Cost sensitivity of S3VMs “省”
  - CS4VM [Li et al., AAAI10]

Related Works

- Global optimization
  - Branch-and-Bound [Chepelle et al., NIPS2006]
  - Deterministic Annealing [Sindhwani et al., ICML2006]
  - Continuation Method [Chepelle et al., ICML2006]

- Pro: good performance on very small data sets
- Con: poor scalability (i.e., could not handle with more than several hundred examples)

Related Works

- Local optimization
  - Local Conbinatorial Search [Joachims, ICML1999]
  - Alternating Optimization [Zhang et al., ICML2009]
  - Constrained Convex-Concave Procedure (CCCP) [Collobert et al., JMLR2006]

- Pro: good scalability
- Con: suffer from local optima, suboptimal performance

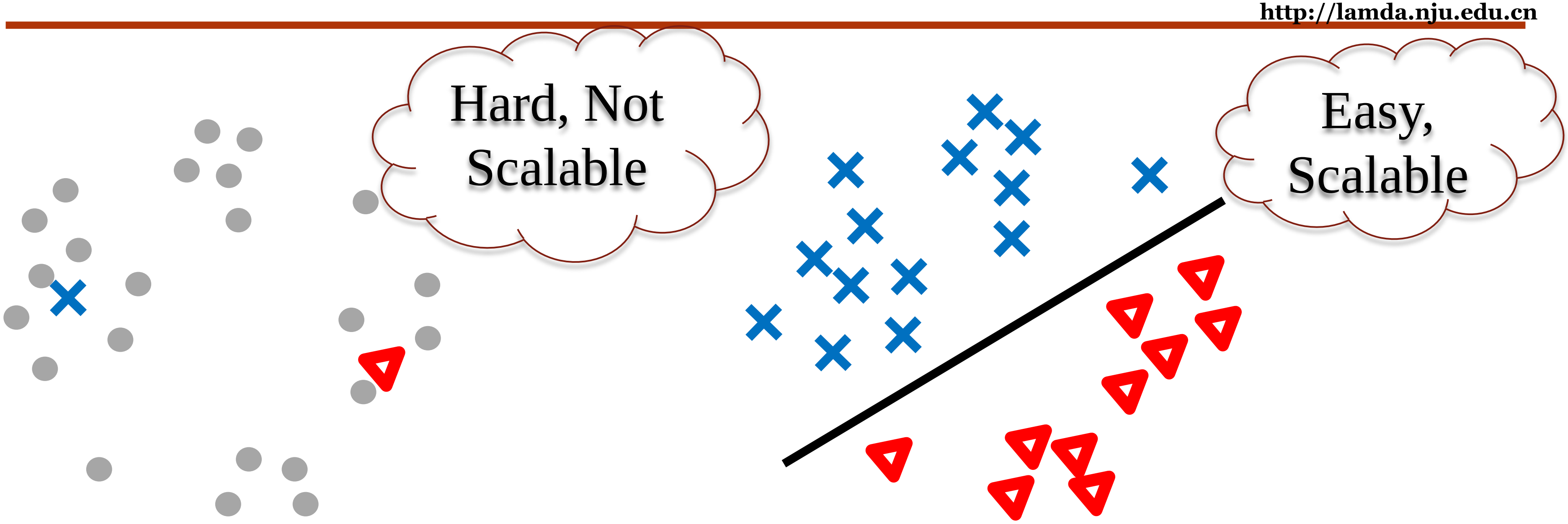
Related Works

- SDP convex relaxation [Xu et al., 2005; De Bie and Cristianini, 2006]
  - Relax S3VMs as convex Semi-Definite Programming (SDP)
  - SDP typically scales $O(n^{6.5})$ where n is the sample size [Zhang et al., TNN2011].

- Pro: promising performance
- Con: poor scalability (i.e., could not handle with more than several thousand examples)

Can we have a scalable and convex S3VM?

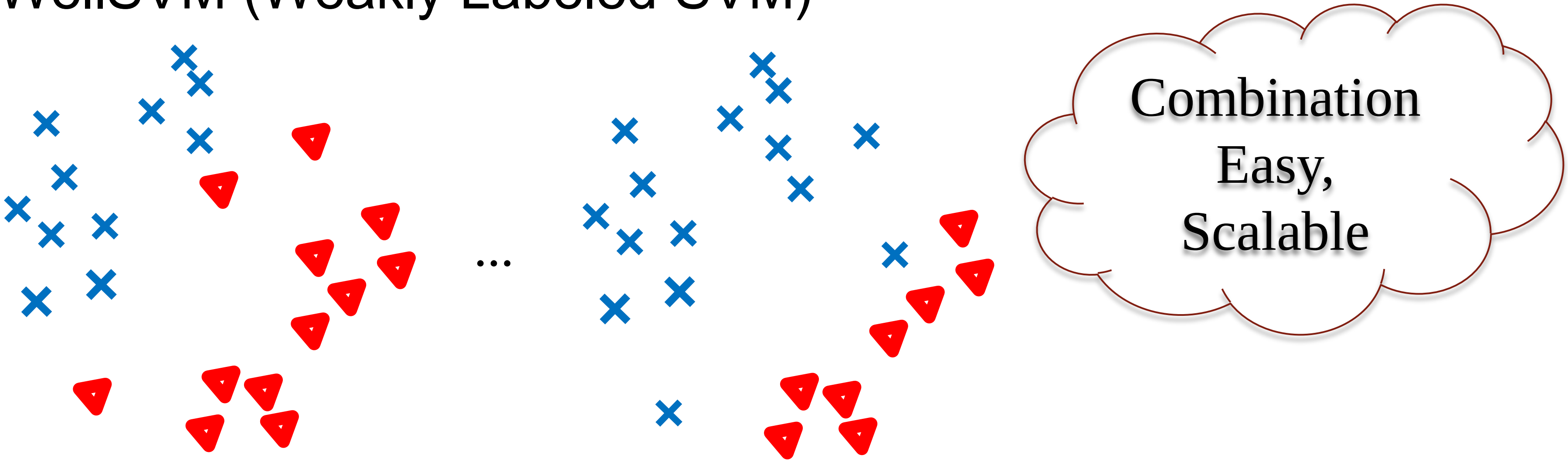
Observation



Hard, Not Scalable

Easy, Scalable

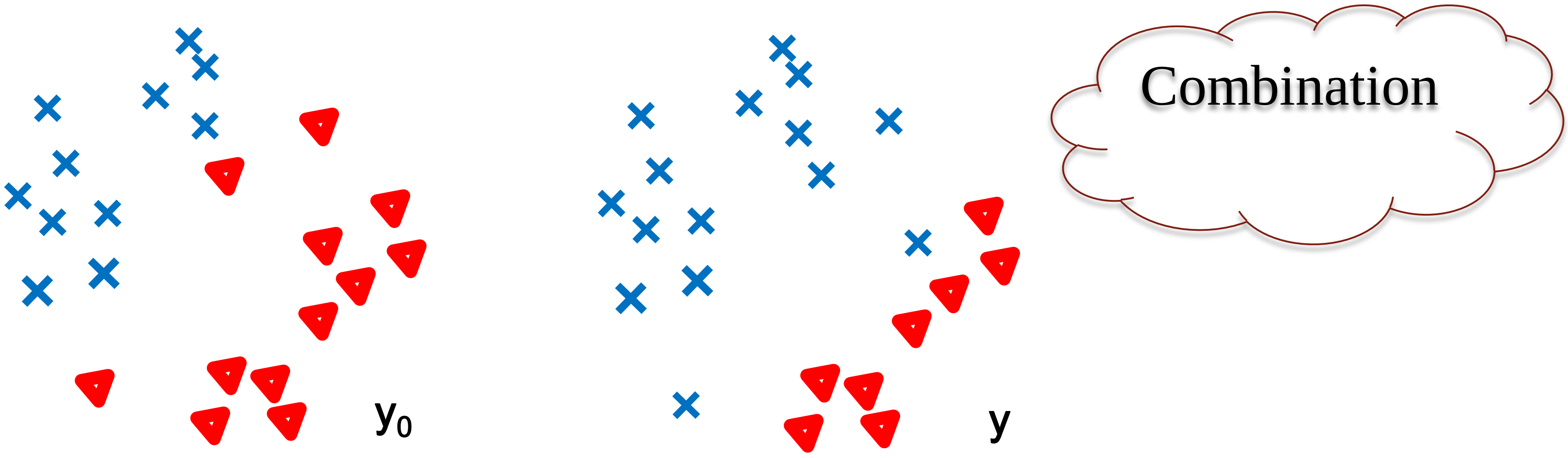
- WellSVM (Weakly Labeled SVM)



Combination Easy, Scalable

WellSVM Algorithm

- Step 1: Initialize a label assignment $\mathbf{y}_0$ for unlabeled data and set the working set $\mathbf{C} = \mathbf{y}_0$.
- Step 2: Generate an informative label assignment $\mathbf{y}$ and update $\mathbf{C} = \mathbf{C} \cup \mathbf{y}$.
- Step 3: Learn an optimal combination for the label assignments in $\mathbf{C}$ such that the margin is maximized.
- Step 4: Repeat Steps 2-3 until convergence.



WellSVM [Li et al., JMLR13]

S3VMs and Its Dual Form

- S3VMs Primal

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w}} \Omega(\mathbf{w}) + C_1 \sum_{i=1}^{l} \ell(\mathbf{w}, \mathbf{x}_i, y_i) + C_2 \sum_{j=l+1}^{N} \ell(\mathbf{w}, \mathbf{x}_j, \hat{y}_j)$$

- S3VMs Dual

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) := \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\Big(\mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}'\Big)\boldsymbol{\alpha}.$$

WellSVM: Main Results

- Minimax Relaxation

WellSVM $$\boxed{\max_{\boldsymbol{\alpha}\in\mathcal{A}}}\boxed{\min_{\hat{\mathbf{y}}\in\mathcal{B}}} G(\boldsymbol{\alpha},\hat{\mathbf{y}}) := \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\Big(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\Big)\boldsymbol{\alpha}.$$

S3VMs $$\boxed{\min_{\hat{\mathbf{y}}\in\mathcal{B}}}\boxed{\max_{\boldsymbol{\alpha}\in\mathcal{A}}} G(\boldsymbol{\alpha},\hat{\mathbf{y}}) := \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\Big(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\Big)\boldsymbol{\alpha}.$$

- Advantages of WellSVM
  - A tight and convex relaxation of S3VMs
    - At least as tight as existing convex SDP relaxations
  - Can make use of state-of-the-art SVM softwares
    - Scalable

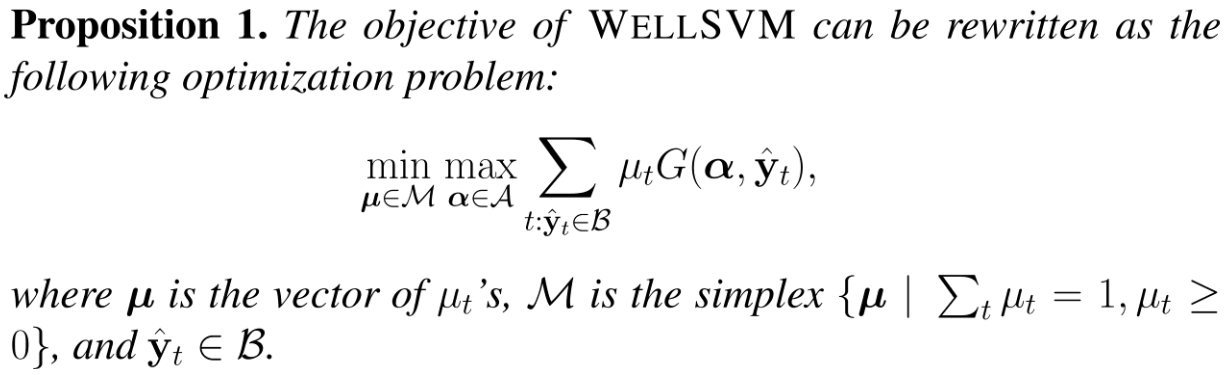
WellSVM [Li et al., JMLR13]

Relax

- Rewritten as

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max\nolimits_\theta \ \theta \qquad \text{s.t.} \ \ G(\boldsymbol{\alpha},\hat{\mathbf{y}}_t) \geq \theta,\ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\},$$

WellSVM is a convex relaxation of S3VMs

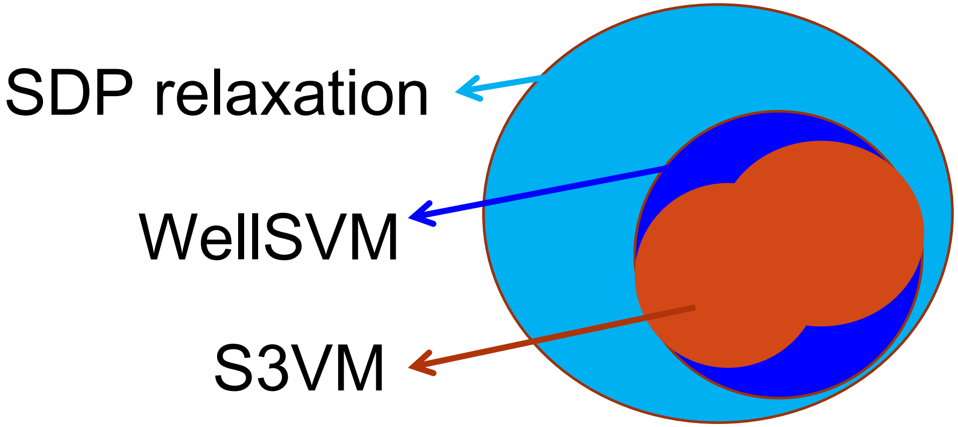


Proposition 1. *The objective of* WELLSVM *can be rewritten as the following optimization problem:*

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha},\hat{\mathbf{y}}_t),$$

where $\boldsymbol{\mu}$ *is the vector of* μ_t*'s,* $\mathcal{M}$ *is the simplex* $\{\boldsymbol{\mu} \mid \sum_t \mu_t = 1, \mu_t \geq 0\}$*, and* $\hat{\mathbf{y}}_t \in \mathcal{B}$.

WellSVM is at least as tight as SDP convex relaxations.



WellSVM [Li et al., JMLR13]

Optimization

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_\theta \ \theta \right.$$
$$\left. \text{s.t.} \ \ G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\},$$

- exponential number of constraints, direct optimization computationally intractable

- Typically not all these constraints are active at optimality
  - Including only a subset of them: a very good approximation
  - Cutting-Plane method
    - Generate a violated label assignment

$$\mathbf{y}^* = \text{argmax}_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}' \left(\mathbf{K} \odot \boldsymbol{\alpha}\boldsymbol{\alpha}'\right) \bar{\mathbf{y}}.$$

Can be solved by sorting.

WellSVM [Li et al., JMLR13]

Optimization

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \ \theta \right.$$
$$\left. \text{s.t.} \ \ G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\},$$

- exponential number of constraints, direct optimization computationally intractable

- Typically not all these constraints are active at optimality
  - Including only a subset of them: a very good approximation
  - Cutting-Plane method
    - Optimal combination

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\Big(\textstyle\sum_{t=1}^{T} \mu_t \mathbf{K} \odot \hat{\mathbf{y}}_t \hat{\mathbf{y}}_t' \Big)\boldsymbol{\alpha},$$

Multiple Kernel Learning, can make use of state-of-the-art SVM software

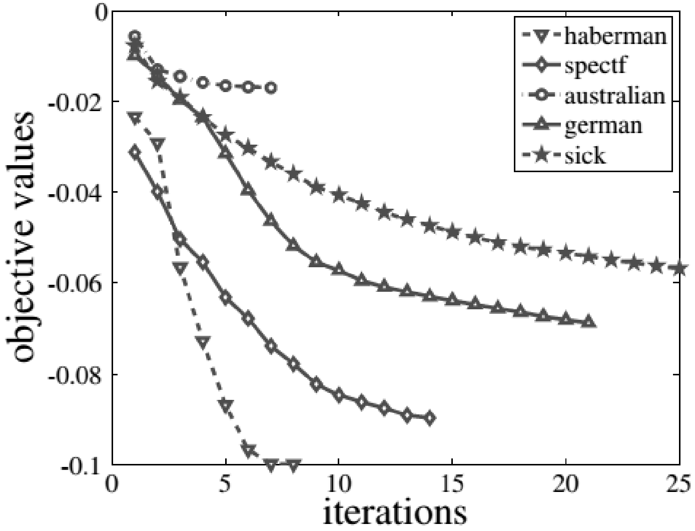
WellSVM [Li et al., JMLR13]

Properties

- $-G(\boldsymbol{\alpha}, \mathbf{y})$ is λ-strongly convex and M-Lipschitz.
- Let $p^{(t)}$ be the optimal objective value of at the t-th iteration.

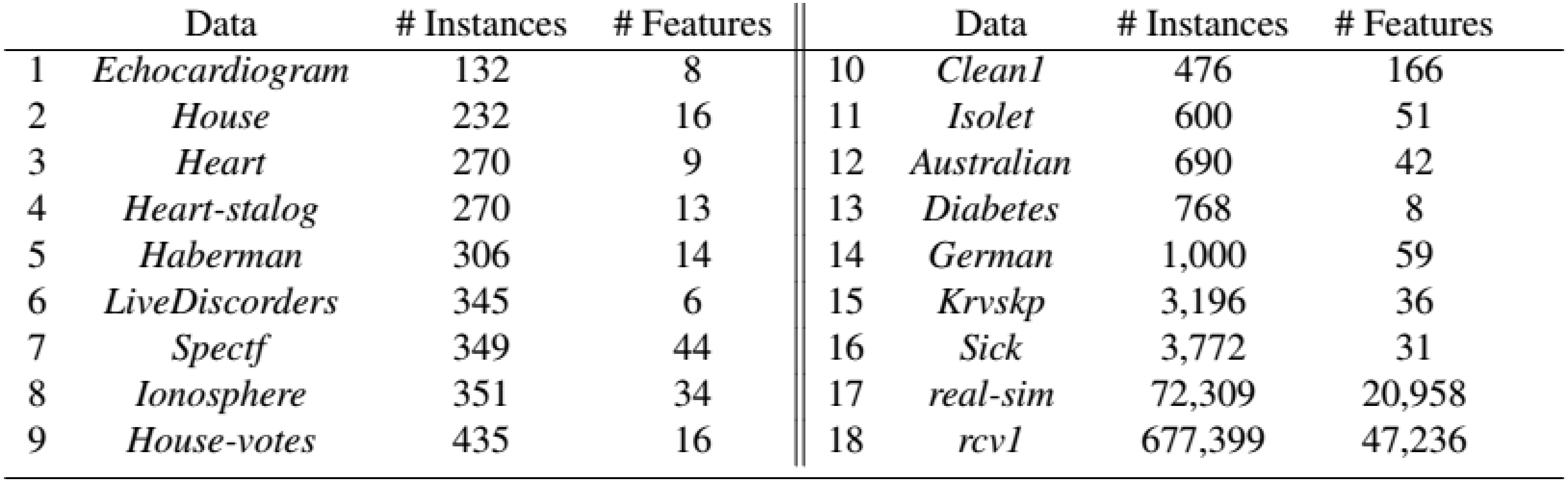
$$p^{(t+1)} \leq p^{(t)} - \eta, \text{ where } \eta = \left(\frac{-c+\sqrt{c^2+4\epsilon}}{2}\right)^2, \text{ and } c = M\sqrt{2/\lambda}.$$

The algorithms converges in no more than $\frac{p^{(1)}-p^*}{\eta}$ iteration.



the more effort spent on generating a violated label, the faster the convergence

Experiments



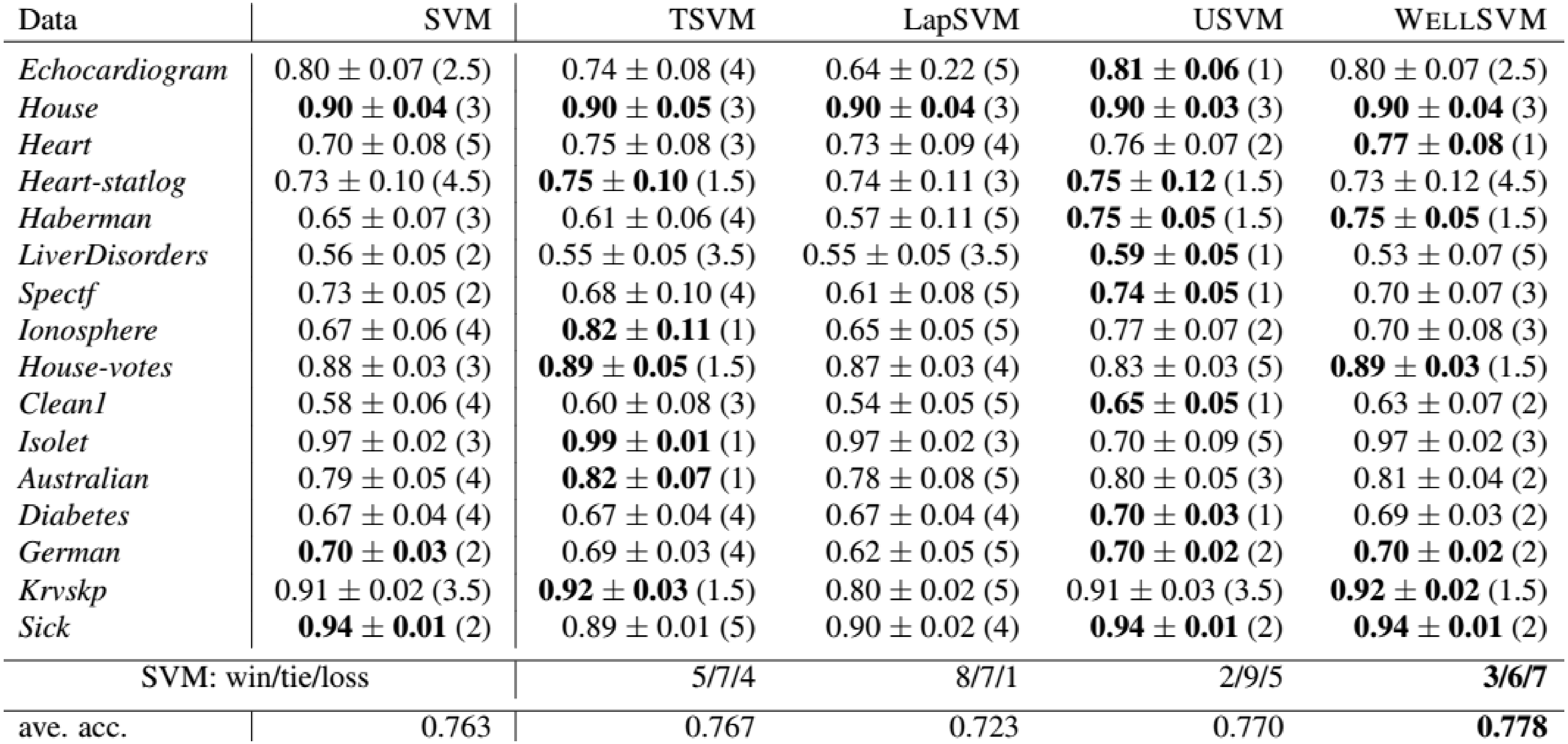
	Data	# Instances	# Features
1	*Echocardiogram*	132	8
2	*House*	232	16
3	*Heart*	270	9
4	*Heart-stalog*	270	13
5	*Haberman*	306	14
6	*LiveDiscorders*	345	6
7	*Spectf*	349	44
8	*Ionosphere*	351	34
9	*House-votes*	435	16
10	*Clean1*	476	166
11	*Isolet*	600	51
12	*Australian*	690	42
13	*Diabetes*	768	8
14	*German*	1,000	59
15	*Krvskp*	3,196	36
16	*Sick*	3,772	31
17	*real-sim*	72,309	20,958
18	*rcv1*	677,399	47,236

- 75% for training, 25% for testing
- WellSVM (LIBSVM for non-linear kernel, LIBLINEAR for linear kernel) vs
  - Standard SVM (using labeled data only)
  - Transductive SVM (TSVM) [Joachims, 1999]
  - Laplacian SVM (LapSVM) [Belkin et al., 2006]
  - UniverSVM (USVM) [Collobert et al., 2006]
  - SVMlin [Sindhwani and Keerthi, 2006]
- SDP-based S3VMs [Xu et al., NIPS2005; De Bie et al., SSL book, 2006]: cannot converge after 3 hours on the smallest data sets

WellSVM [Li et al., JMLR13]

5% labeled examples

Learning And Mining from DatA
http://lamda.nju.edu.cn



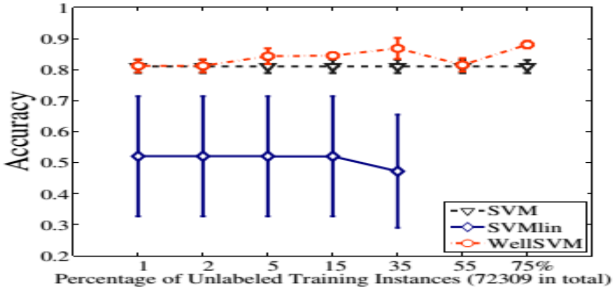
Data	SVM	TSVM	LapSVM	USVM	WellSVM
Echocardiogram	0.80 ± 0.07 (2.5)	0.74 ± 0.08 (4)	0.64 ± 0.22 (5)	**0.81** ± **0.06** (1)	0.80 ± 0.07 (2.5)
House	**0.90** ± **0.04** (3)	**0.90** ± **0.05** (3)	**0.90** ± **0.04** (3)	**0.90** ± **0.03** (3)	**0.90** ± **0.04** (3)
Heart	0.70 ± 0.08 (5)	0.75 ± 0.08 (3)	0.73 ± 0.09 (4)	0.76 ± 0.07 (2)	**0.77** ± **0.08** (1)
Heart-statlog	0.73 ± 0.10 (4.5)	**0.75** ± **0.10** (1.5)	0.74 ± 0.11 (3)	**0.75** ± **0.12** (1.5)	0.73 ± 0.12 (4.5)
Haberman	0.65 ± 0.07 (3)	0.61 ± 0.06 (4)	0.57 ± 0.11 (5)	**0.75** ± **0.05** (1.5)	**0.75** ± **0.05** (1.5)
LiverDisorders	0.56 ± 0.05 (2)	0.55 ± 0.05 (3.5)	0.55 ± 0.05 (3.5)	**0.59** ± **0.05** (1)	0.53 ± 0.07 (5)
Spectf	0.73 ± 0.05 (2)	0.68 ± 0.10 (4)	0.61 ± 0.08 (5)	**0.74** ± **0.05** (1)	0.70 ± 0.07 (3)
Ionosphere	0.67 ± 0.06 (4)	**0.82** ± **0.11** (1)	0.65 ± 0.05 (5)	0.77 ± 0.07 (2)	0.70 ± 0.08 (3)
House-votes	0.88 ± 0.03 (3)	**0.89** ± **0.05** (1.5)	0.87 ± 0.03 (4)	0.83 ± 0.03 (5)	**0.89** ± **0.03** (1.5)
Clean1	0.58 ± 0.06 (4)	0.60 ± 0.08 (3)	0.54 ± 0.05 (5)	**0.65** ± **0.05** (1)	0.63 ± 0.07 (2)
Isolet	0.97 ± 0.02 (3)	**0.99** ± **0.01** (1)	0.97 ± 0.02 (3)	0.70 ± 0.09 (5)	0.97 ± 0.02 (3)
Australian	0.79 ± 0.05 (4)	**0.82** ± **0.07** (1)	0.78 ± 0.08 (5)	0.80 ± 0.05 (3)	0.81 ± 0.04 (2)
Diabetes	0.67 ± 0.04 (4)	0.67 ± 0.04 (4)	0.67 ± 0.04 (4)	**0.70** ± **0.03** (1)	0.69 ± 0.03 (2)
German	**0.70** ± **0.03** (2)	0.69 ± 0.03 (4)	0.62 ± 0.05 (5)	**0.70** ± **0.02** (2)	**0.70** ± **0.02** (2)
Krvskp	0.91 ± 0.02 (3.5)	**0.92** ± **0.03** (1.5)	0.80 ± 0.02 (5)	0.91 ± 0.03 (3.5)	**0.92** ± **0.02** (1.5)
Sick	**0.94** ± **0.01** (2)	0.89 ± 0.01 (5)	0.90 ± 0.02 (4)	**0.94** ± **0.01** (2)	**0.94** ± **0.01** (2)
SVM: win/tie/loss		5/7/4	8/7/1	2/9/5	**3/6/7**
ave. acc.	0.763	0.767	0.723	0.770	**0.778**

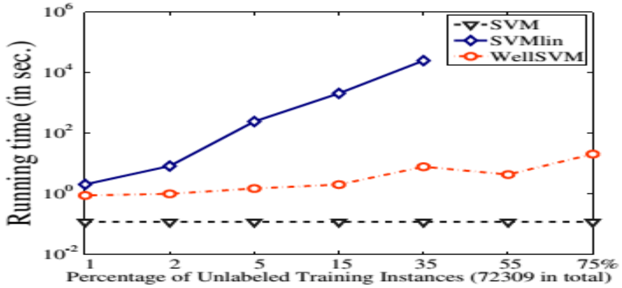
- WellSVM is highly competitive

WellSVM [Li et al., JMLR13]

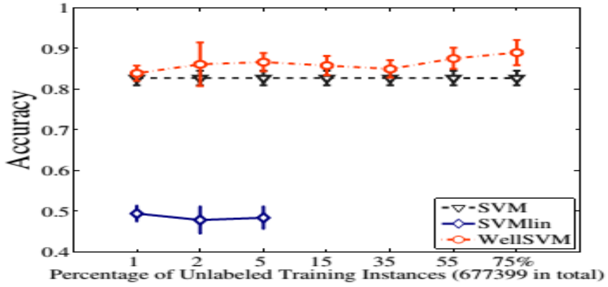
Larger Data Sets

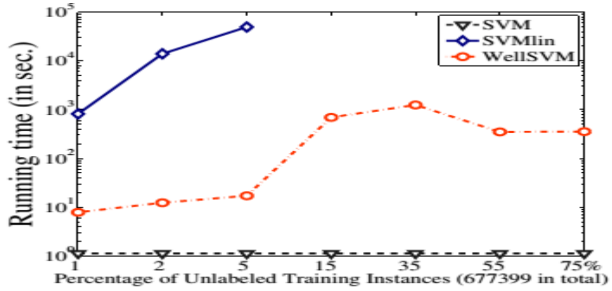
- Real-sim: 20,958 features, 72,309 instances





- RCV1: 47,236 features, 677,399 instances





- WellSVM is always more accurate than SVMlin
- For RCV1, SVMlin can not converge in 24 hours when >5% examples are used for training.

Outline

- Scalability of S3VMs “多”
  - WellSVM [Li et al., JMLR13]
- Efficiency of S3VMs “快”
  - MeanS3VM [Li et al., ICML09]
- Safeness of S3VMs “好”
  - S4VM [Li and Zhou, ICML11]
- Cost sensitivity of S3VMs “省”
  - CS4VM [Li et al., AAAI10]

MeanS3VM [Li et al., ICML09]

Related Work

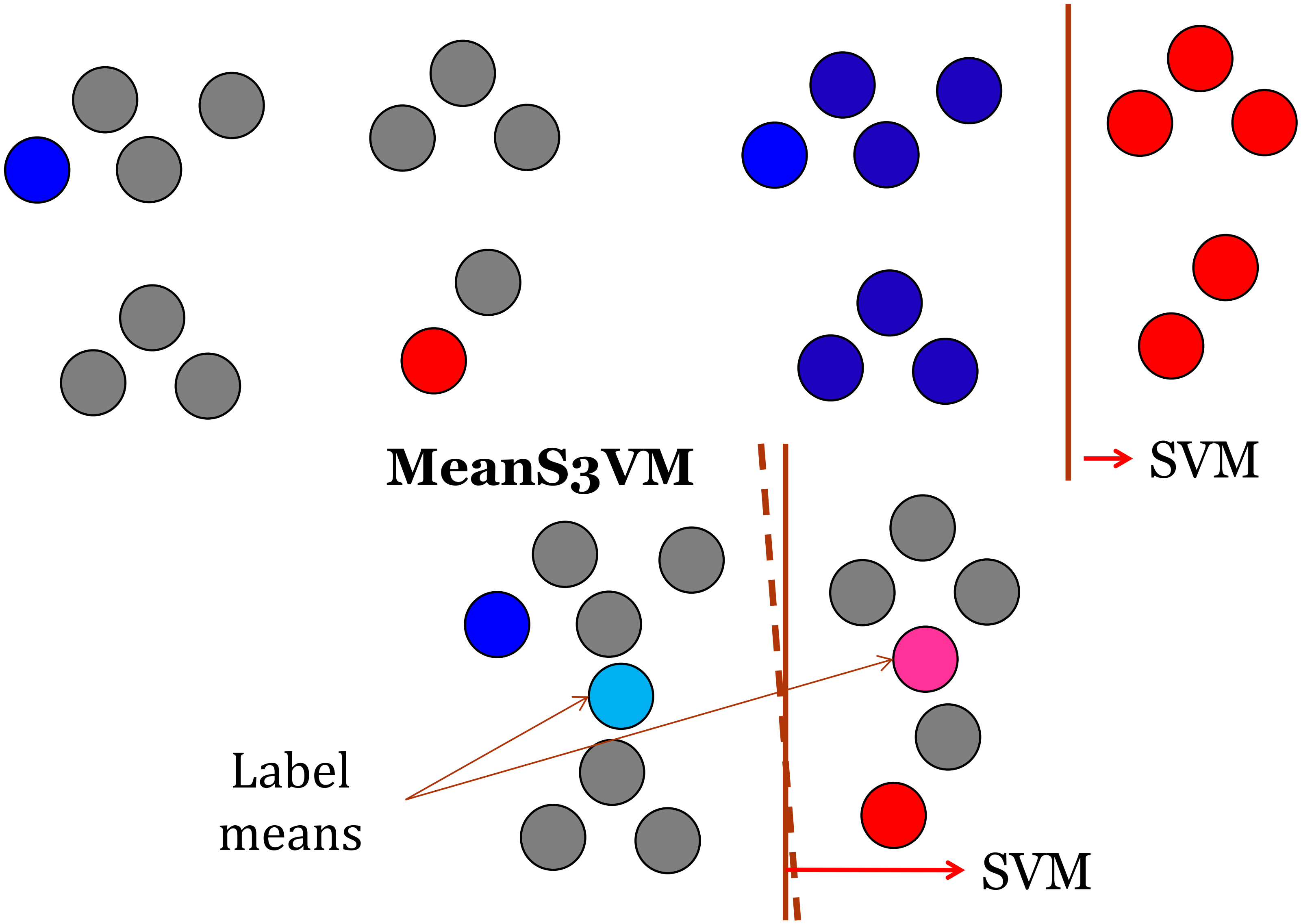
$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w}} \Omega(\mathbf{w}) + C_1 \sum_{i=1}^{l} \ell(\mathbf{w}, \mathbf{x}_i, y_i) + C_2 \sum_{j=l+1}^{N} \ell(\mathbf{w}, \mathbf{x}_j, \hat{y}_j)$$

- State-of-the-art S3VMs typically aim at on optimizing the objective function of S3VMs, which has to estimate a label assignment for all the unlabeled data. This is computational inefficient, especially when there are a large amount of unlabeled data.

approximate algorithms. Specifically, simpler sufficient statistics might be useful to approximate a good performance

MeanS3VM [Li et al., ICML09]

Observation



SVM

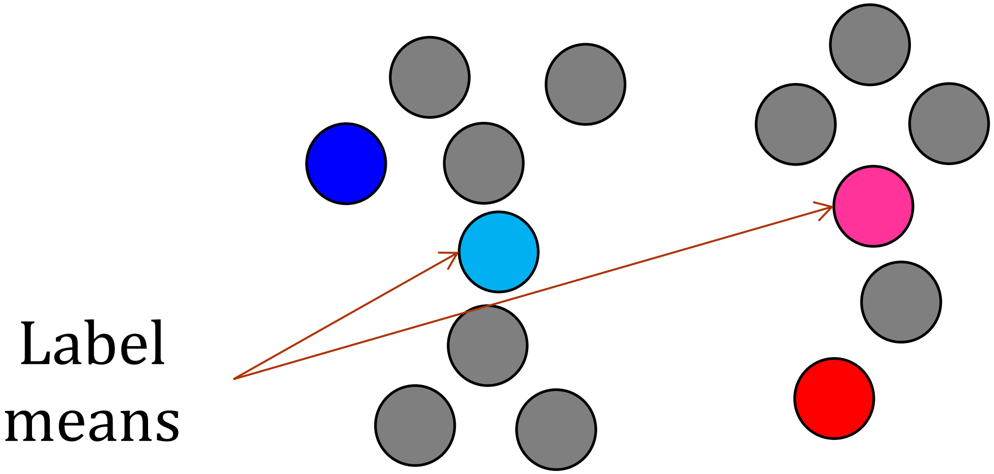
MeanS3VM

Label means

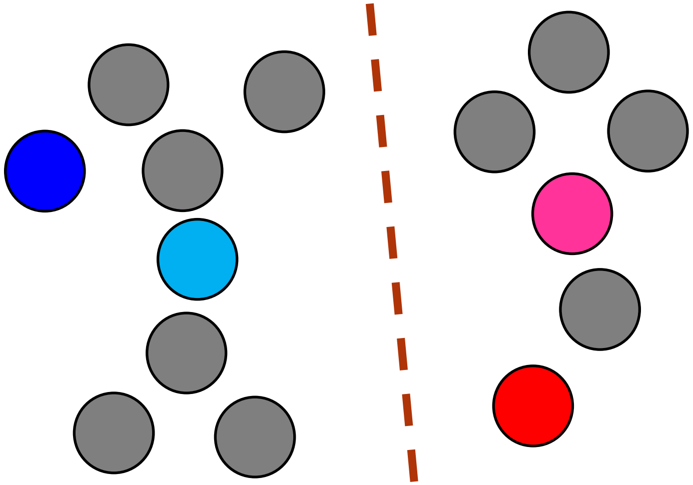
SVM

MeanS3VM Algorithm

- Step 1: Estimate the label means of unlabeled data.



- Step 2: Train an SVM with the use of estimated label means.



MeanS3VM [Li et al., ICML09]

Usefulness of Label Mean

We consider following optimization problem

$$
\begin{aligned}
\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{p}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i\in\mathcal{I}_l} \xi_i + C_2 \sum_{i\in\mathcal{I}_u} (\xi_i + p_{i-l} - |f(\mathbf{x}_i)|) \\
\text{s.t.} \quad & y_i(\mathbf{w}'\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i \in \mathcal{I}_l, \\
& \mathbf{w}'\phi(\mathbf{x}_i) + b \leq p_{i-l}, \quad -\mathbf{w}'\phi(\mathbf{x}_i) - b \leq p_{i-l}, \\
& p_{i-l} \geq 1 - \xi_i, \quad i \in \mathcal{I}_u; \quad \xi_i \geq 0, \; i \in \mathcal{I}_l \cup \mathcal{I}_u, \\
& \sum_{i\in\mathcal{I}_u} \mathrm{sgn}(\mathbf{w}'\phi(\mathbf{x}_i) + b) = r.
\end{aligned}
$$

Lemma 1. *Let $(\mathbf{w}^*, b^*, \boldsymbol{\xi}^*, \boldsymbol{p}^*)$ be the optimal solution. Then, for $i \in \mathcal{I}_u$,*

$$
\xi_i^* + p_{i-l}^* = \begin{cases} 1 & |f(\mathbf{x}_i)| \leq 1, \\ |f(\mathbf{x}_i)| & \textit{otherwise.} \end{cases}
$$

It is equivalent to S3VMs

MeanS3VM [Li et al., ICML09]

Usefulness of Label Mean

We consider following optimization problem

$$\begin{aligned}
\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{p}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i\in\mathcal{I}_l} \xi_i + C_2 \sum_{i\in\mathcal{I}_u} (\xi_i + p_{i-l} - \boxed{|f(\mathbf{x}_i)|)} \\
\text{s.t.} \quad & y_i(\mathbf{w}'\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i \in \mathcal{I}_l, \\
& \mathbf{w}'\phi(\mathbf{x}_i) + b \leq p_{i-l}, \quad -\mathbf{w}'\phi(\mathbf{x}_i) - b \leq p_{i-l}, \\
& p_{i-l} \geq 1 - \xi_i, \quad i \in \mathcal{I}_u; \quad \xi_i \geq 0, \; i \in \mathcal{I}_l \cup \mathcal{I}_u,
\end{aligned}$$

$$\boxed{\sum_{i\in\mathcal{I}_u} |f(\mathbf{x}_i)| - (u_- - u_+)b = \mathbf{w}'\left(\sum_{i\in\mathcal{I}_u, f(\mathbf{x}_i)\geq 0} \phi(\mathbf{x}_i) - \sum_{i\in\mathcal{I}_u, f(\mathbf{x}_i)<0} \phi(\mathbf{x}_i)\right) = u_+\mathbf{w}'_+ - u_-\mathbf{w}'_-}$$

$$\left.\begin{aligned}
\hat{\mathbf{m}}_+ &= \tfrac{1}{u_+} \textstyle\sum_{i\in\mathcal{I}_u, f(\mathbf{x}_i)\geq 0} \phi(\mathbf{x}_i) \\
\hat{\mathbf{m}}_- &= \tfrac{1}{u_-} \textstyle\sum_{i\in\mathcal{I}_u, f(\mathbf{x}_i)<0} \phi(\mathbf{x}_i)
\end{aligned}\right\}$$

are the estimates of the label means

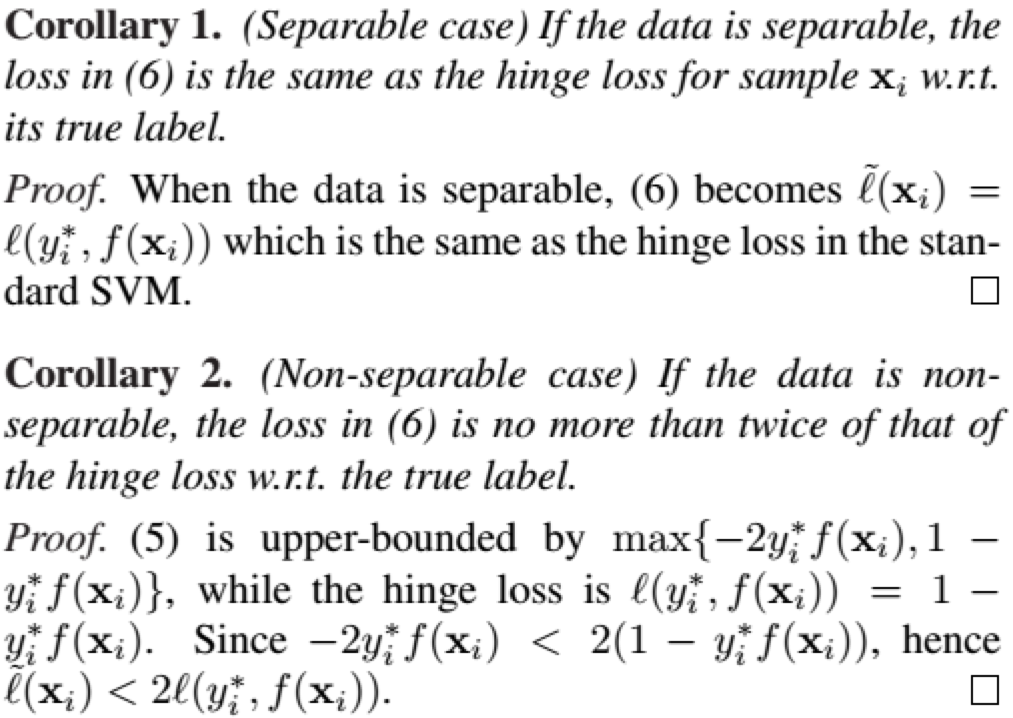
MeanS3VM

$$
\begin{aligned}
\min_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{p}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i \in \mathcal{I}_l} \xi_i + C_2 \sum_{i \in \mathcal{I}_u} (\xi_i + p_{i-l}) - C_2(u_+ \mathbf{w}' \boxed{\mathbf{m}_+} - u_- \mathbf{w}' \boxed{\mathbf{m}_-} + (u_+ - u_-) b) \\
\text{s.t.} \quad & y_i(\mathbf{w}'\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i \in \mathcal{I}_l, \\
& \mathbf{w}'\phi(\mathbf{x}_i) + b \leq p_{i-l}, \quad -\mathbf{w}'\phi(\mathbf{x}_i) - b \leq p_{i-l}, \\
& p_{i-l} \geq 1 - \xi_i, \quad i \in \mathcal{I}_u; \quad \xi_i \geq 0, \; i \in \mathcal{I}_l \cup \mathcal{I}_u, \\
& \sum_{i \in \mathcal{I}_u} \operatorname{sgn}(\mathbf{w}'\phi(\mathbf{x}_i) + b) = r.
\end{aligned}
$$

- Input is only related to label means, rather than the labels
- Can MeanS3VM be a good approximation?

MeanS3VM [Li et al., ICML09]

Properties

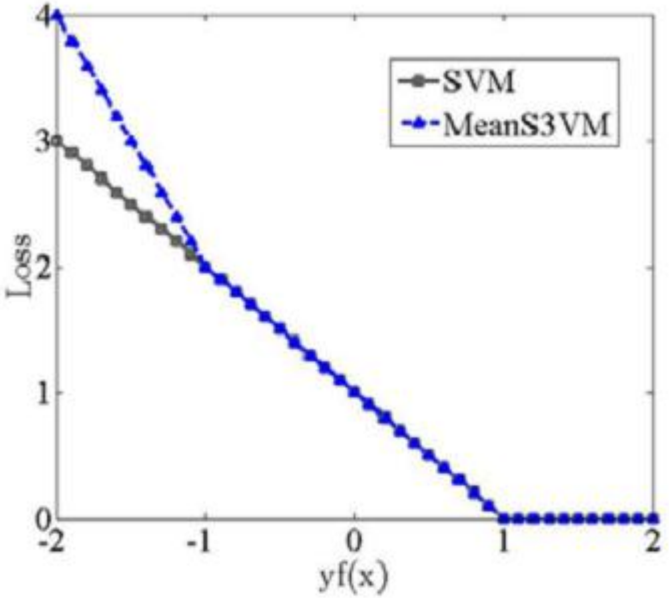


Corollary 1. *(Separable case) If the data is separable, the loss in (6) is the same as the hinge loss for sample $\mathbf{x}_i$ w.r.t. its true label.*

Proof. When the data is separable, (6) becomes $\tilde{\ell}(\mathbf{x}_i) = \ell(y_i^*, f(\mathbf{x}_i))$ which is the same as the hinge loss in the standard SVM. □

Corollary 2. *(Non-separable case) If the data is non-separable, the loss in (6) is no more than twice of that of the hinge loss w.r.t. the true label.*

Proof. (5) is upper-bounded by $\max\{-2y_i^* f(\mathbf{x}_i), 1 - y_i^* f(\mathbf{x}_i)\}$, while the hinge loss is $\ell(y_i^*, f(\mathbf{x}_i)) = 1 - y_i^* f(\mathbf{x}_i)$. Since $-2y_i^* f(\mathbf{x}_i) < 2(1 - y_i^* f(\mathbf{x}_i))$, hence $\tilde{\ell}(\mathbf{x}_i) < 2\ell(y_i^*, f(\mathbf{x}_i))$. □



With the knowledge of label means, MeanS3VM is closely related to supervised SVM with the knowledge of all the labels.

MeanS3VM [Li et al., ICML09]

Estimate the Label Means

- Large margin approach

$$
\begin{aligned}
\min_{\mathbf{d}\in\Delta}\ \min_{\mathbf{w},b,\rho,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C_1\sum_{i=1}^{l}\xi_i - C_2\rho \\
\text{s.t.} \quad & y_i(\mathbf{w}'\phi(\mathbf{x}_i)+b) \geq 1-\xi_i,\ \ i=1,\ldots,l, \\
& \frac{1}{u_+}\left(\mathbf{w}'\sum_{j=l+1}^{l+u} d_{j-l}\phi(\mathbf{x}_j)\right)+b \geq \rho, \\
& \frac{1}{u_-}\left(\mathbf{w}'\sum_{j=l+1}^{l+u}(1-d_{j-l})\phi(\mathbf{x}_j)\right)+b \leq -\rho.
\end{aligned}
$$

We proposed two algorithms to solve it, one is based on minimax convex relaxation proposed in [Li et al., JMLR13] and the other is based on alternating optimization.

- Note that it has much fewer constraints than S3VM, which greatly reduces the time complexity of the optimization.
- It can also be explained in terms of MMD [Gretton et al., NIPS'06] which aims to separate the distribution of each class with large margin

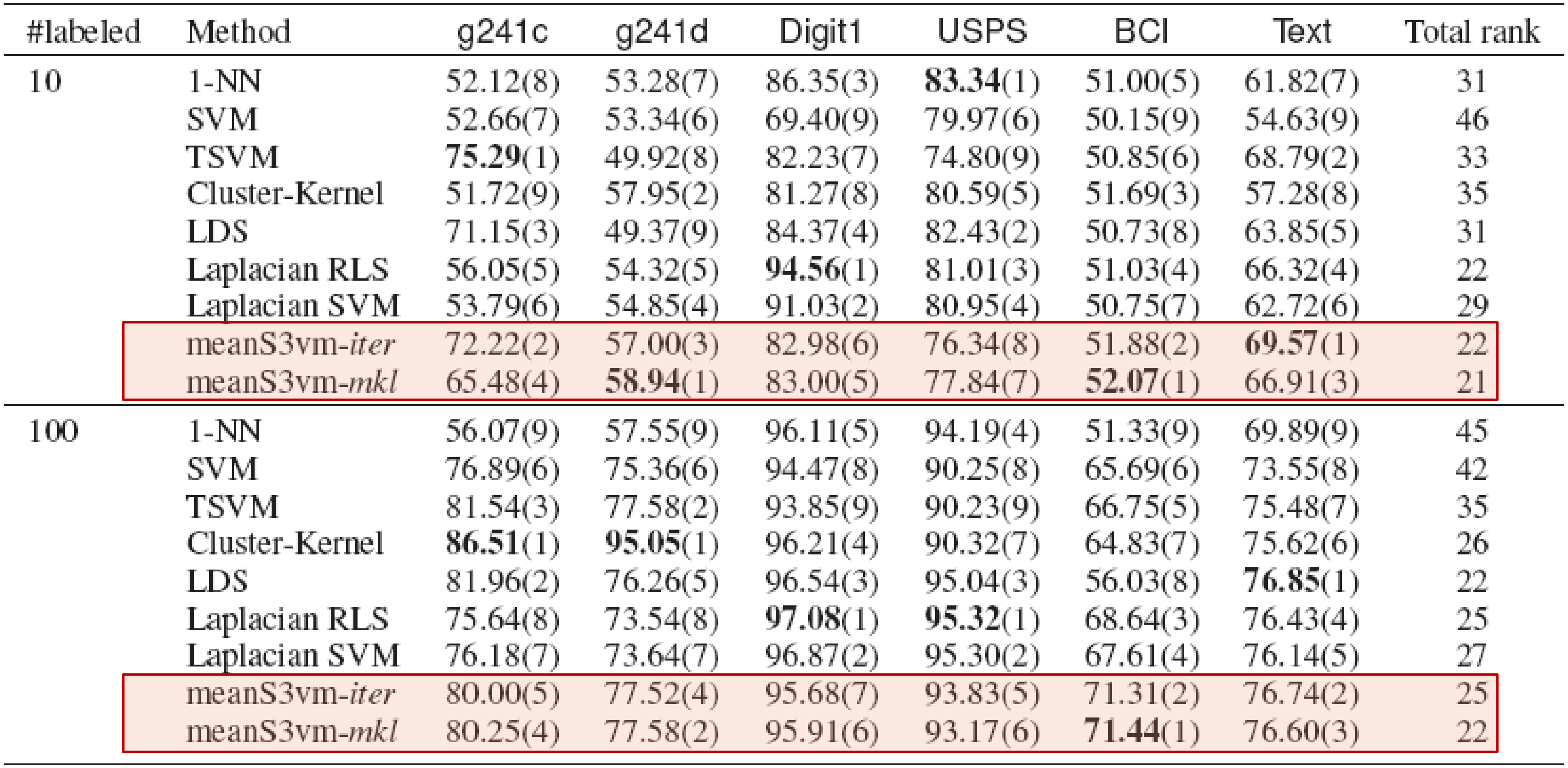
Experiment

- Benchmark Data Sets
- UCI Data Sets
- Text Categorization
- CPU Time
    - We call our MeanS3VM using convex relaxation as MeanS3VM-mkl and the one using alternating optimization as MeanS3VM-iter.

MeanS3VM [Li et al., ICML09]

Benchmark Tasks

Following the same setup in SSL-book 2006



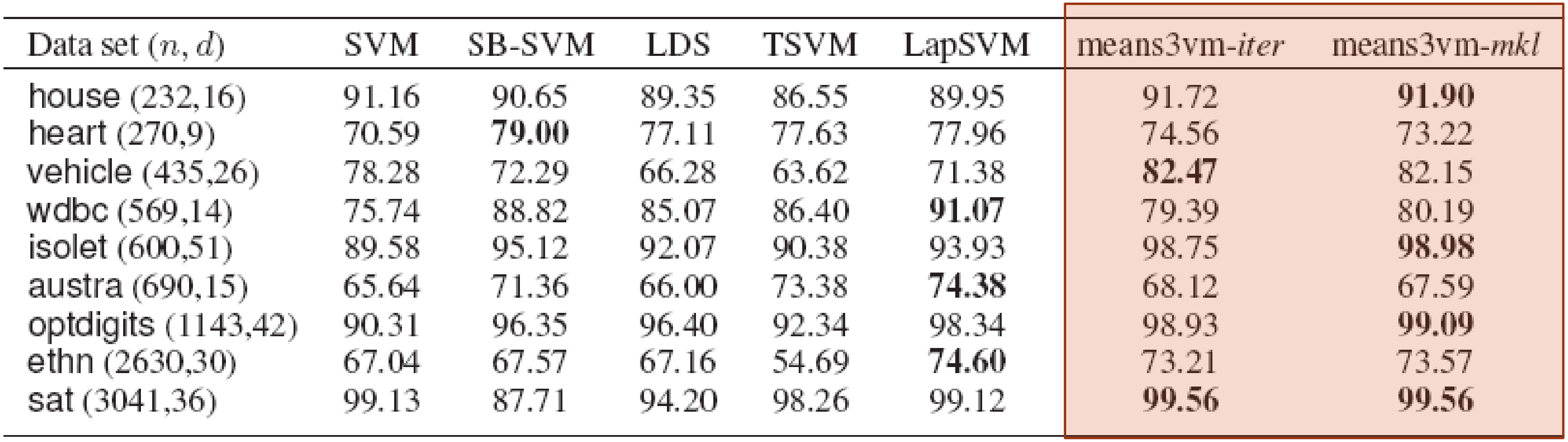
#labeled	Method	g241c	g241d	Digit1	USPS	BCI	Text	Total rank
10	1-NN	52.12(8)	53.28(7)	86.35(3)	**83.34**(1)	51.00(5)	61.82(7)	31
	SVM	52.66(7)	53.34(6)	69.40(9)	79.97(6)	50.15(9)	54.63(9)	46
	TSVM	**75.29**(1)	49.92(8)	82.23(7)	74.80(9)	50.85(6)	68.79(2)	33
	Cluster-Kernel	51.72(9)	57.95(2)	81.27(8)	80.59(5)	51.69(3)	57.28(8)	35
	LDS	71.15(3)	49.37(9)	84.37(4)	82.43(2)	50.73(8)	63.85(5)	31
	Laplacian RLS	56.05(5)	54.32(5)	**94.56**(1)	81.01(3)	51.03(4)	66.32(4)	22
	Laplacian SVM	53.79(6)	54.85(4)	91.03(2)	80.95(4)	50.75(7)	62.72(6)	29
	meanS3vm-*iter*	72.22(2)	57.00(3)	82.98(6)	76.34(8)	51.88(2)	**69.57**(1)	22
	meanS3vm-*mkl*	65.48(4)	**58.94**(1)	83.00(5)	77.84(7)	**52.07**(1)	66.91(3)	21
100	1-NN	56.07(9)	57.55(9)	96.11(5)	94.19(4)	51.33(9)	69.89(9)	45
	SVM	76.89(6)	75.36(6)	94.47(8)	90.25(8)	65.69(6)	73.55(8)	42
	TSVM	81.54(3)	77.58(2)	93.85(9)	90.23(9)	66.75(5)	75.48(7)	35
	Cluster-Kernel	**86.51**(1)	**95.05**(1)	96.21(4)	90.32(7)	64.83(7)	75.62(6)	26
	LDS	81.96(2)	76.26(5)	96.54(3)	95.04(3)	56.03(8)	**76.85**(1)	22
	Laplacian RLS	75.64(8)	73.54(8)	**97.08**(1)	**95.32**(1)	68.64(3)	76.43(4)	25
	Laplacian SVM	76.18(7)	73.64(7)	96.87(2)	95.30(2)	67.61(4)	76.14(5)	27
	meanS3vm-*iter*	80.00(5)	77.52(4)	95.68(7)	93.83(5)	71.31(2)	76.74(2)	25
	meanS3vm-*mkl*	80.25(4)	77.58(2)	95.91(6)	93.17(6)	**71.44**(1)	76.60(3)	22

MeanS3vms are highly competitive

MeanS3VM [Li et al., ICML09]

UCI Data Sets

9 data sets, 10 labeled data, 50% train / 50% test, 20 runs



Data set (n, d)	SVM	SB-SVM	LDS	TSVM	LapSVM	means3vm-*iter*	means3vm-*mkl*
house (232,16)	91.16	90.65	89.35	86.55	89.95	91.72	**91.90**
heart (270,9)	70.59	**79.00**	77.11	77.63	77.96	74.56	73.22
vehicle (435,26)	78.28	72.29	66.28	63.62	71.38	**82.47**	82.15
wdbc (569,14)	75.74	88.82	85.07	86.40	**91.07**	79.39	80.19
isolet (600,51)	89.58	95.12	92.07	90.38	93.93	98.75	**98.98**
austra (690,15)	65.64	71.36	66.00	73.38	**74.38**	68.12	67.59
optdigits (1143,42)	90.31	96.35	96.40	92.34	98.34	98.93	**99.09**
ethn (2630,30)	67.04	67.57	67.16	54.69	**74.60**	73.21	73.57
sat (3041,36)	99.13	87.71	94.20	98.26	99.12	**99.56**	**99.56**

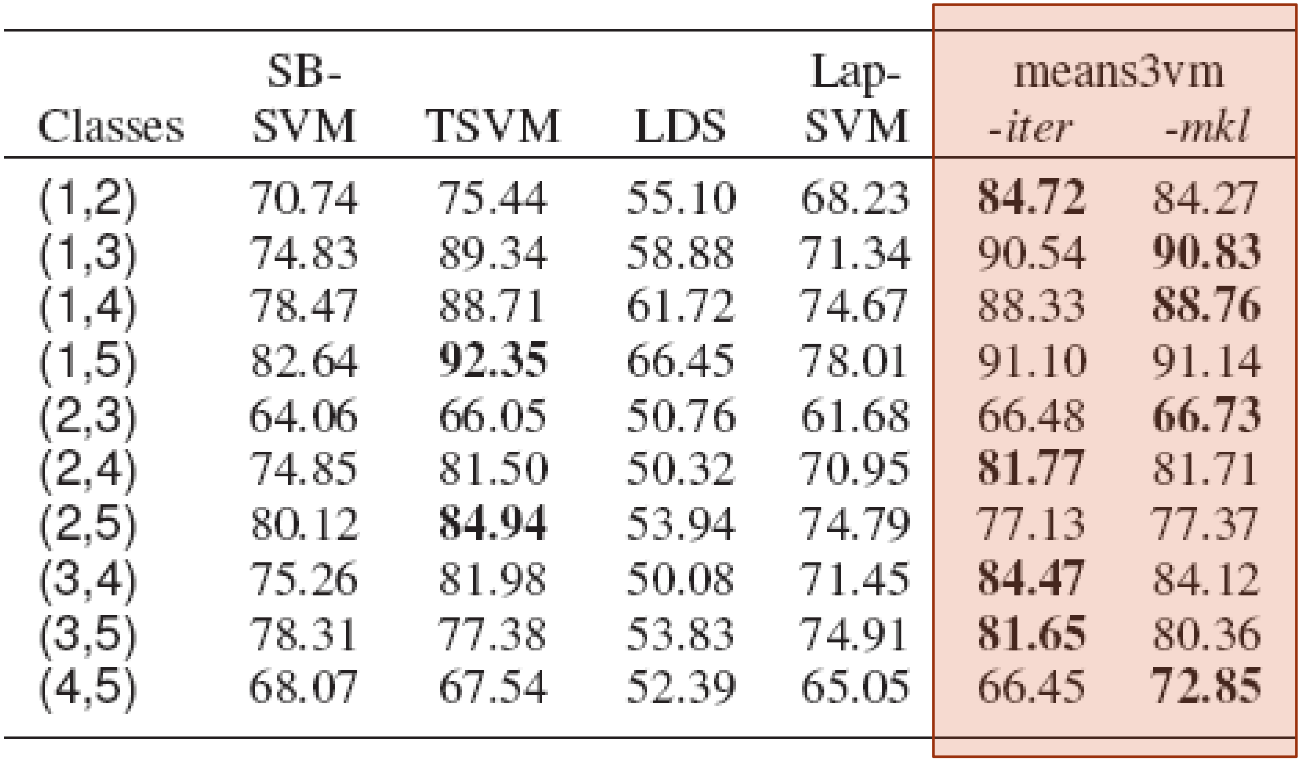
MeanS3VMs are highly competitive.

In particular, they achieve the best performance in 6 of 9 tasks.

MeanS3VM [Li et al., ICML09]

Text Categorization

10 binary tasks: 2 labeled data, 50% train / 50% test, 20 runs



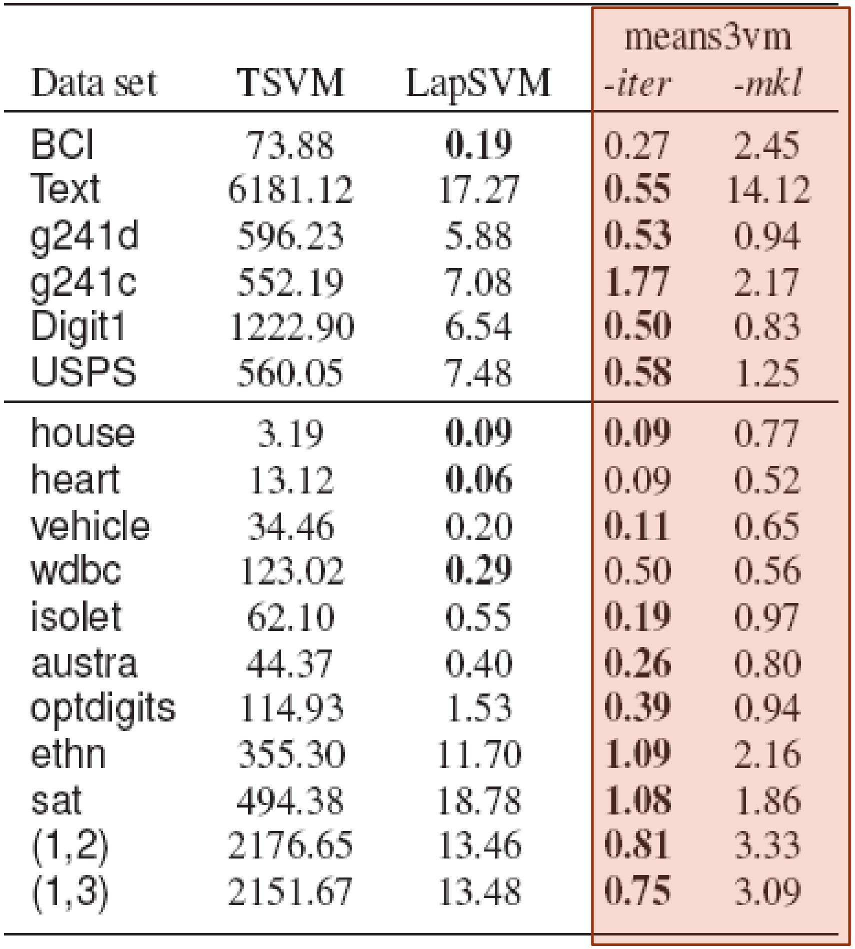
	SB-			Lap-	means3vm	
Classes	SVM	TSVM	LDS	SVM	*-iter*	*-mkl*
(1,2)	70.74	75.44	55.10	68.23	**84.72**	84.27
(1,3)	74.83	89.34	58.88	71.34	90.54	**90.83**
(1,4)	78.47	88.71	61.72	74.67	88.33	**88.76**
(1,5)	82.64	**92.35**	66.45	78.01	91.10	91.14
(2,3)	64.06	66.05	50.76	61.68	66.48	**66.73**
(2,4)	74.85	81.50	50.32	70.95	**81.77**	81.71
(2,5)	80.12	**84.94**	53.94	74.79	77.13	77.37
(3,4)	75.26	81.98	50.08	71.45	**84.47**	84.12
(3,5)	78.31	77.38	53.83	74.91	**81.65**	80.36
(4,5)	68.07	67.54	52.39	65.05	66.45	**72.85**

Means3vms are highly competitive.
In particular, they achieve the best performance in 8 of 10 tasks.

MeanS3VM [Li et al., ICML09]

CPU Time

MeanS3VM-*iter* is almost the fastest method. On larger data sets, MeanS3VM-*iter* is 10 times faster than Laplacian SVM, 100 times faster than TSVM.



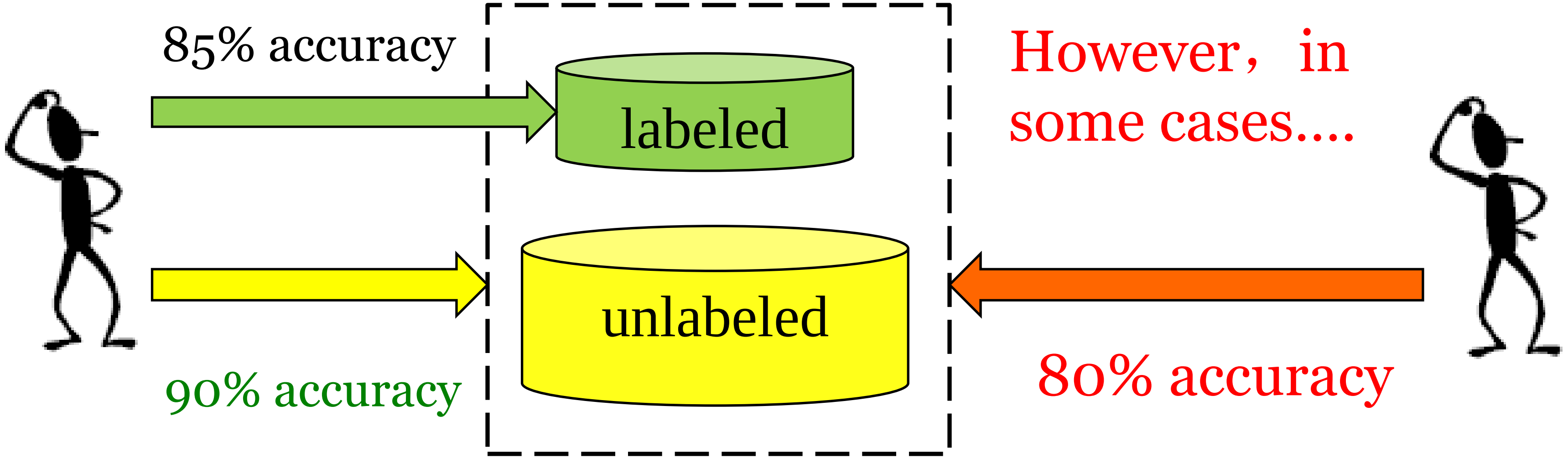
Data set	TSVM	LapSVM	means3vm *-iter*	 *-mkl*
BCI	73.88	**0.19**	0.27	2.45
Text	6181.12	17.27	**0.55**	14.12
g241d	596.23	5.88	**0.53**	0.94
g241c	552.19	7.08	**1.77**	2.17
Digit1	1222.90	6.54	**0.50**	0.83
USPS	560.05	7.48	**0.58**	1.25
house	3.19	**0.09**	**0.09**	0.77
heart	13.12	**0.06**	0.09	0.52
vehicle	34.46	0.20	**0.11**	0.65
wdbc	123.02	**0.29**	0.50	0.56
isolet	62.10	0.55	**0.19**	0.97
austra	44.37	0.40	**0.26**	0.80
optdigits	114.93	1.53	**0.39**	0.94
ethn	355.30	11.70	**1.09**	2.16
sat	494.38	18.78	**1.08**	1.86
(1,2)	2176.65	13.46	**0.81**	3.33
(1,3)	2151.67	13.48	**0.75**	3.09

Outline

- Scalability of S3VMs “多”
  - WellSVM [Li et al., JMLR13]
- Efficiency of S3VMs “快”
  - MeanS3VM [Li et al., ICML09]
- Safeness of S3VMs “好”
  - S4VM [Li and Zhou, ICML11]
- Cost sensitivity of S3VMs “省”
  - CS4VM [Li et al., AAAI10]

S4VM [Li and Zhou, ICML11]

Make Unlabeled Data Never Hurt



85% accuracy

labeled

unlabeled

90% accuracy

However， in some cases....

80% accuracy

SSL works well!!

Unlabeled data may *hurt* the performance.

How to develop **safe** SSL methods which do not *significantly* degenerate the performance?

Related Works

- Generative method: [Cozman et al., 2003] conjectured that the performance degeneration is caused by incorrect model assumption. However, it is very difficult to make a correct model assumption without sufficient domain knowledge.

- Co-training method: Incorrect pseudo-labels may mislead the learning process. One possible solution is to employ data editing process [Li and Zhou, 2005]. However, it only works for dense data.

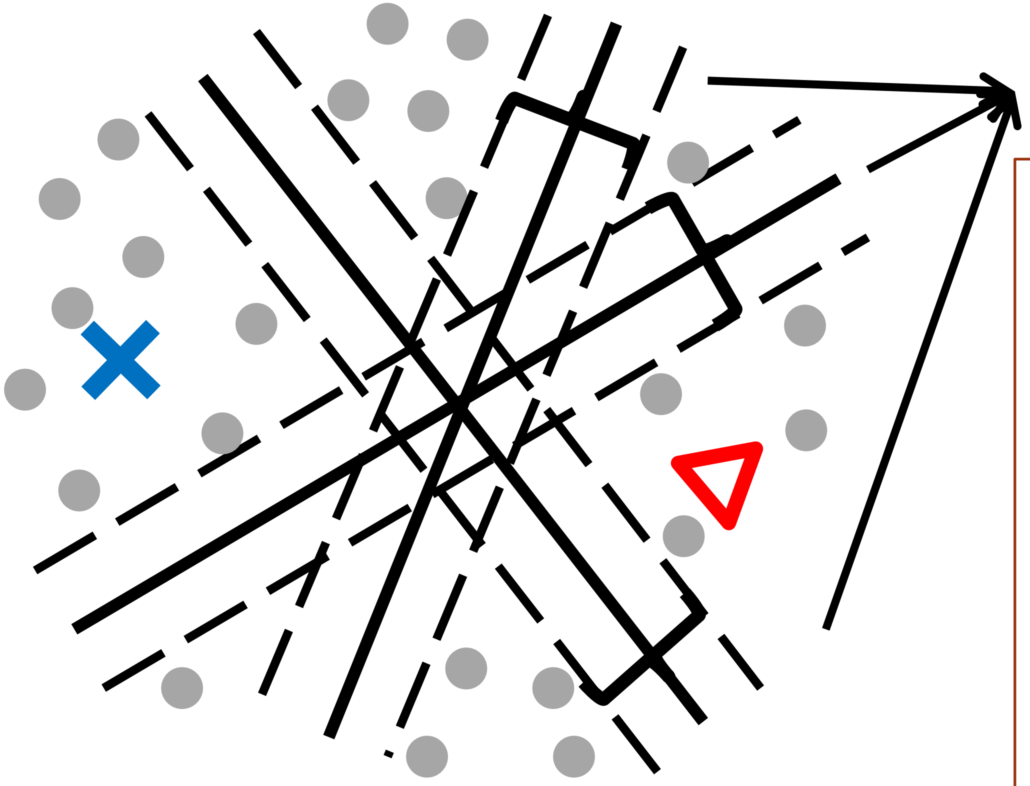
- Graph-based method: Graph construction is the crucial problem. However, how to develop a good graph in general situations remains an open problem.

Related Works

- S3VMs: The correctness of S3VMs has been studied on very small data sets [Chapelle et al., 2008]. However, there is no clear solution to avoid performance degeneration using unlabeled data.
- There are also some general discussions from a theoretical perspective [Balcan and Blum, 2010; Ben-David et al., 2008; Singh et al., 2009].
- To our best knowledge, few safe SSL approaches have been proposed.

S4VM [Li and Zhou, ICML11]

Observation



Large Margin Separator

S4VMs
(Safe S3VMs)

i) **More than one** Large Margin Separators!!

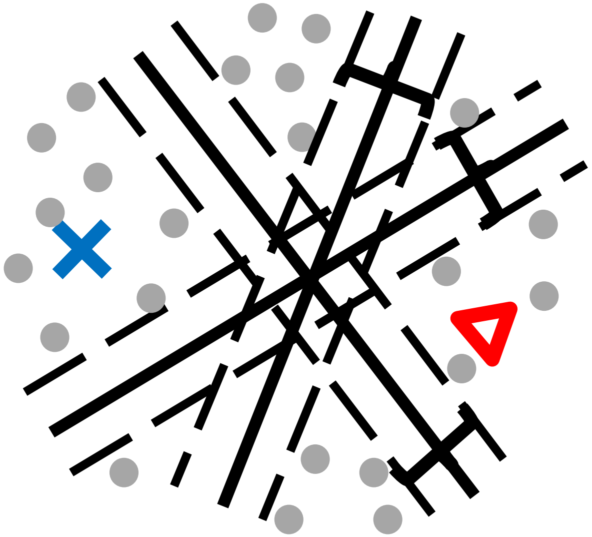
ii) Current S3VMs **randomly** select one of them as the output.

iii) Large Margin Separators are usually **diverse**.

iv) **Incorrect selection** degenerates the performance!

S4VM Algorithm

- Step 1: Generate a pool of large-margin separators (LMS).



- Step 2: Construct S4VM by optimizing the performance improvement in the worst-case for any separator.

S4VM [Li and Zhou, ICML11]

Construct S4VM from a pool of LMS

- Maximize accuracy

$$\max_{\mathbf{y}\in\{\pm1\}^u} \quad J(\mathbf{y},\mathbf{y}^*,\mathbf{y}^{svm}) = gain(\mathbf{y},\mathbf{y}^*,\mathbf{y}^{svm}) - \lambda\, loss(\mathbf{y},\mathbf{y}^*,\mathbf{y}^{svm})$$

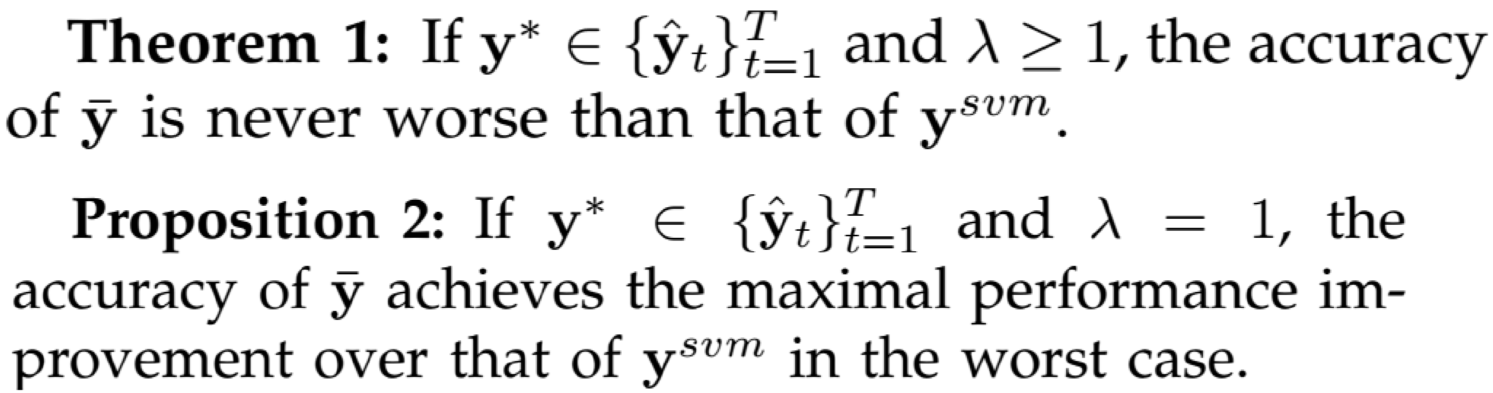
  - gain(): gained accuracy against inductive SVMs
  - loss(): lost accuracy against inductive SVMs
  - λ : measure the risk that user would like to undertake
  - $\mathbf{y}^*$: ground-truth label assignment
  - Difficulty: The ground-truth is unknown.

- Note that ground-truth is a LMS, we assume that $\mathbf{y}^* \in \{\hat{\mathbf{y}}_t\}_{t=1}^T$
- Maximize the worst-case accuracy

$$\bar{\mathbf{y}} = \arg\max_{\mathbf{y}\in\{\pm1\}^u} \min_{\hat{\mathbf{y}}\in\{\hat{\mathbf{y}}_t\}_{t=1}^T} \quad gain(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm}) - \lambda\, loss(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm})$$

S4VM [Li and Zhou, ICML11]

Properties



Theorem 1: If $\mathbf{y}^* \in \{\hat{\mathbf{y}}_t\}_{t=1}^T$ and $\lambda \geq 1$, the accuracy of $\bar{\mathbf{y}}$ is never worse than that of $\mathbf{y}^{svm}$.

Proposition 2: If $\mathbf{y}^* \in \{\hat{\mathbf{y}}_t\}_{t=1}^T$ and $\lambda = 1$, the accuracy of $\bar{\mathbf{y}}$ achieves the maximal performance improvement over that of $\mathbf{y}^{svm}$ in the worst case.

Under the assumption employed in S3VMs, that is the ground-truth is realized by a large-margin separator, S4VM is provable safe and able to achieve the largest performance improvement.

S4VM [Li and Zhou, ICML11]

Optimization

$$\bar{\mathbf{y}} = \arg\max_{\mathbf{y}\in\{\pm1\}^u} \min_{\hat{\mathbf{y}}\in\{\hat{\mathbf{y}}_t\}_{t=1}^T} gain(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm}) - \lambda\, loss(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm})$$

- Integer linear programming. Because

$$\begin{aligned} gain(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm}) &= \sum\nolimits_{j=1}^{u} I(y_j=\hat{y}_j)I(\hat{y}_j\neq y_j^{svm}) = \sum\nolimits_{j=1}^{u} \frac{1+y_j\hat{y}_j}{2}\frac{1-y_j^{svm}\hat{y}_j}{2}, \\ loss(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm}) &= \sum\nolimits_{j=1}^{u} I(y_j\neq\hat{y}_j)I(\hat{y}_j = y_j^{svm}) = \sum\nolimits_{j=1}^{u} \frac{1-y_j\hat{y}_j}{2}\frac{1+y_j^{svm}\hat{y}_j}{2}. \end{aligned}$$

  - Linear functions of $\hat{\mathbf{y}}$

Proposition 1: If $\mathbf{y}^* \in \{\hat{\mathbf{y}}_t\}_{t=1}^T$ and $\lambda \geq 1$, the accuracy of any $\mathbf{y}$ satisfying $\min_{\hat{\mathbf{y}}\in\mathcal{M}} J(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm}) \geq 0$, is never worse than that of $\mathbf{y}^{svm}$.

- Simple convex relaxation method is employed.
- The safeness of the solution is guaranteed.

S4VM [Li and Zhou, ICML11]

Generate a pool of LMS

Large margin, large diversity

- Objective

$$\min_{\{f_t,\hat{\mathbf{y}}_t\in\mathcal{B}\}_{t=1}^{T}} \quad \sum\nolimits_{t=1}^{T} \underline{h(f_t,\hat{\mathbf{y}}_t)} + M\underline{\Omega(\{\hat{\mathbf{y}}_t\}_{t=1}^{T})},$$

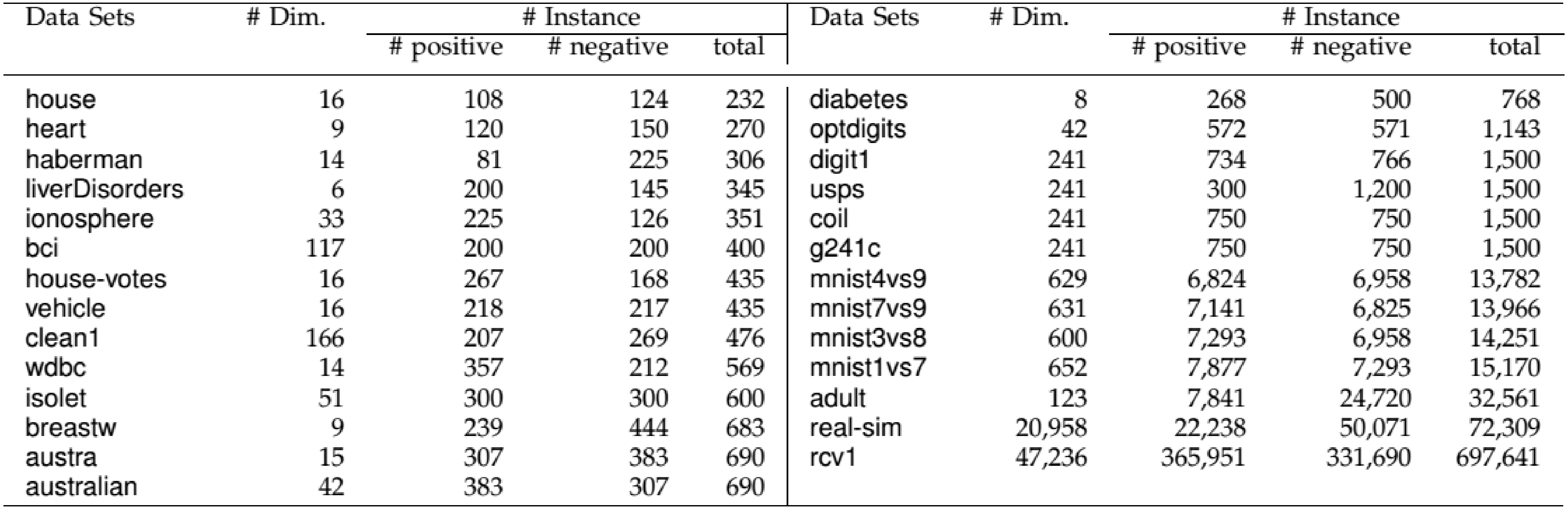
objective function of S3VM

A quantity of penalty about the diversity of separators, e.g.,

$$\sum\nolimits_{1\leq t\neq\tilde{t}\leq T}\mathbf{I}\left(\frac{\hat{\mathbf{y}}_t'\hat{\mathbf{y}}_{\tilde{t}}}{u}\geq 1-\epsilon\right)$$

- Two implementations
  - global simulated annealing search
  - simple and efficient sampling

Experiments

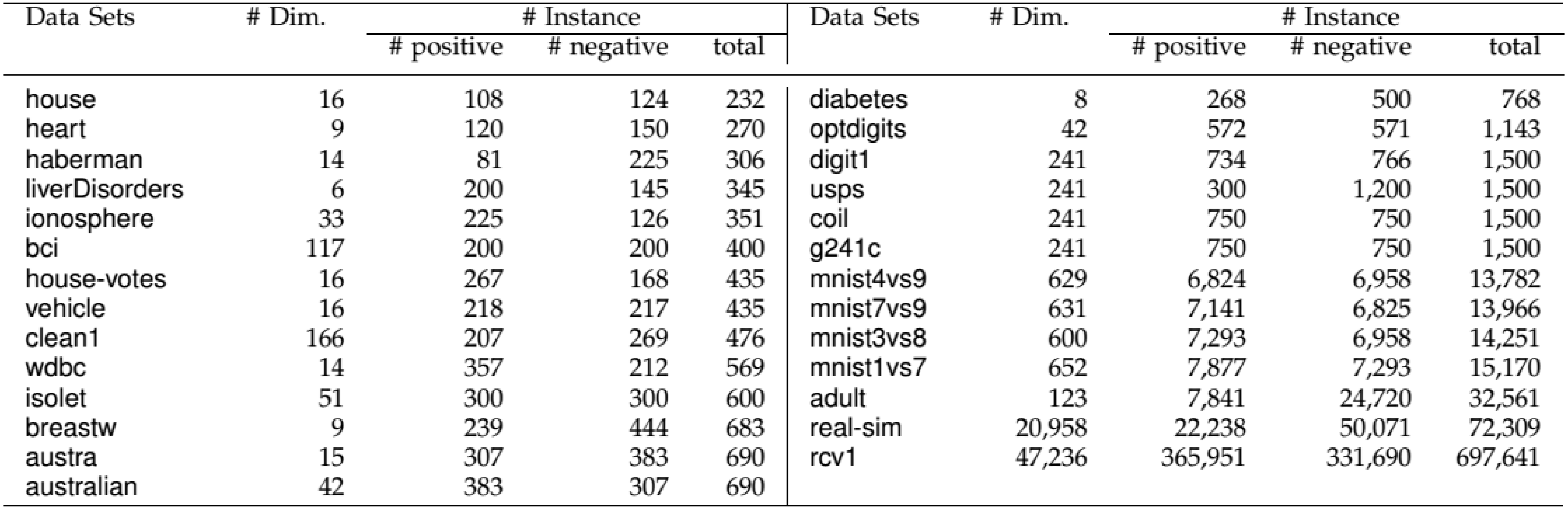


Data Sets	# Dim.	# Instance			Data Sets	# Dim.	# Instance		
		# positive	# negative	total			# positive	# negative	total
house	16	108	124	232	diabetes	8	268	500	768
heart	9	120	150	270	optdigits	42	572	571	1,143
haberman	14	81	225	306	digit1	241	734	766	1,500
liverDisorders	6	200	145	345	usps	241	300	1,200	1,500
ionosphere	33	225	126	351	coil	241	750	750	1,500
bci	117	200	200	400	g241c	241	750	750	1,500
house-votes	16	267	168	435	mnist4vs9	629	6,824	6,958	13,782
vehicle	16	218	217	435	mnist7vs9	631	7,141	6,825	13,966
clean1	166	207	269	476	mnist3vs8	600	7,293	6,958	14,251
wdbc	14	357	212	569	mnist1vs7	652	7,877	7,293	15,170
isolet	51	300	300	600	adult	123	7,841	24,720	32,561
breastw	9	239	444	683	real-sim	20,958	22,238	50,071	72,309
austra	15	307	383	690	rcv1	47,236	365,951	331,690	697,641
australian	42	383	307	690					

- 10 instances are used for training (satisfying balance constraint), the rest for testing.
- Inductive SVM and S4VMs (LIBSVM for small and medium data sets with linear and non-linear kernel; LIBLINEAR for larger data sets with linear kernel)
- S3VM (TSVM for small and medium data sets with linear and non-linear kernel; USVM for larger data sets with linear kernel)

S4VM [Li and Zhou, ICML11]

Experiments

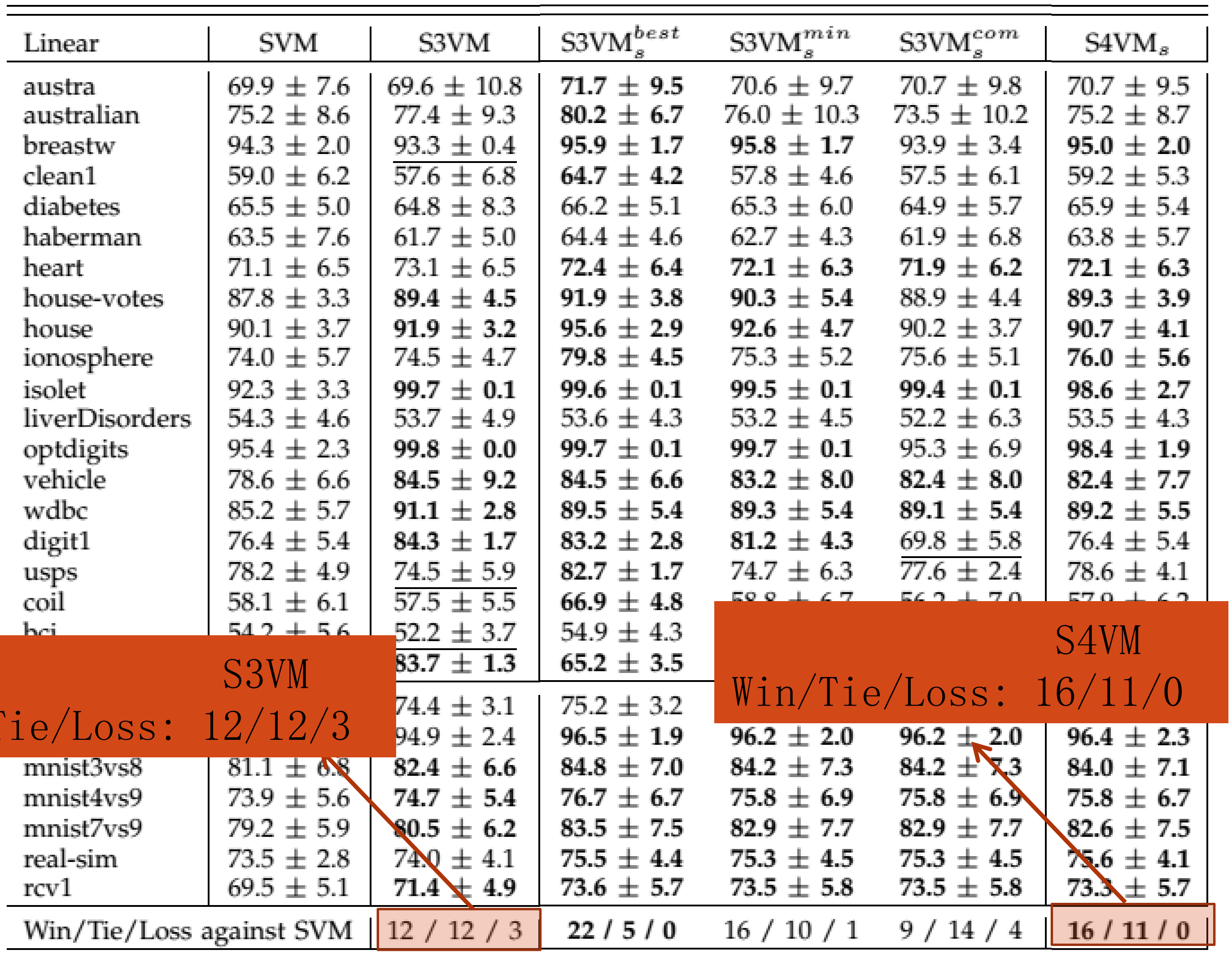


Data Sets	# Dim.	# Instance			Data Sets	# Dim.	# Instance		
		# positive	# negative	total			# positive	# negative	total
house	16	108	124	232	diabetes	8	268	500	768
heart	9	120	150	270	optdigits	42	572	571	1,143
haberman	14	81	225	306	digit1	241	734	766	1,500
liverDisorders	6	200	145	345	usps	241	300	1,200	1,500
ionosphere	33	225	126	351	coil	241	750	750	1,500
bci	117	200	200	400	g241c	241	750	750	1,500
house-votes	16	267	168	435	mnist4vs9	629	6,824	6,958	13,782
vehicle	16	218	217	435	mnist7vs9	631	7,141	6,825	13,966
clean1	166	207	269	476	mnist3vs8	600	7,293	6,958	14,251
wdbc	14	357	212	569	mnist1vs7	652	7,877	7,293	15,170
isolet	51	300	300	600	adult	123	7,841	24,720	32,561
breastw	9	239	444	683	real-sim	20,958	22,238	50,071	72,309
austra	15	307	383	690	rcv1	47,236	365,951	331,690	697,641
australian	42	383	307	690					

- S4VM vs
  - S3VM^{best}: the best performance among the multiple LMS
  - S3VM^{min}: the LMS with minimum objective values
  - S3VM^{com}: combine LMS using uniform weights

S4VM [Li and Zhou, ICML11]

Linear kernel



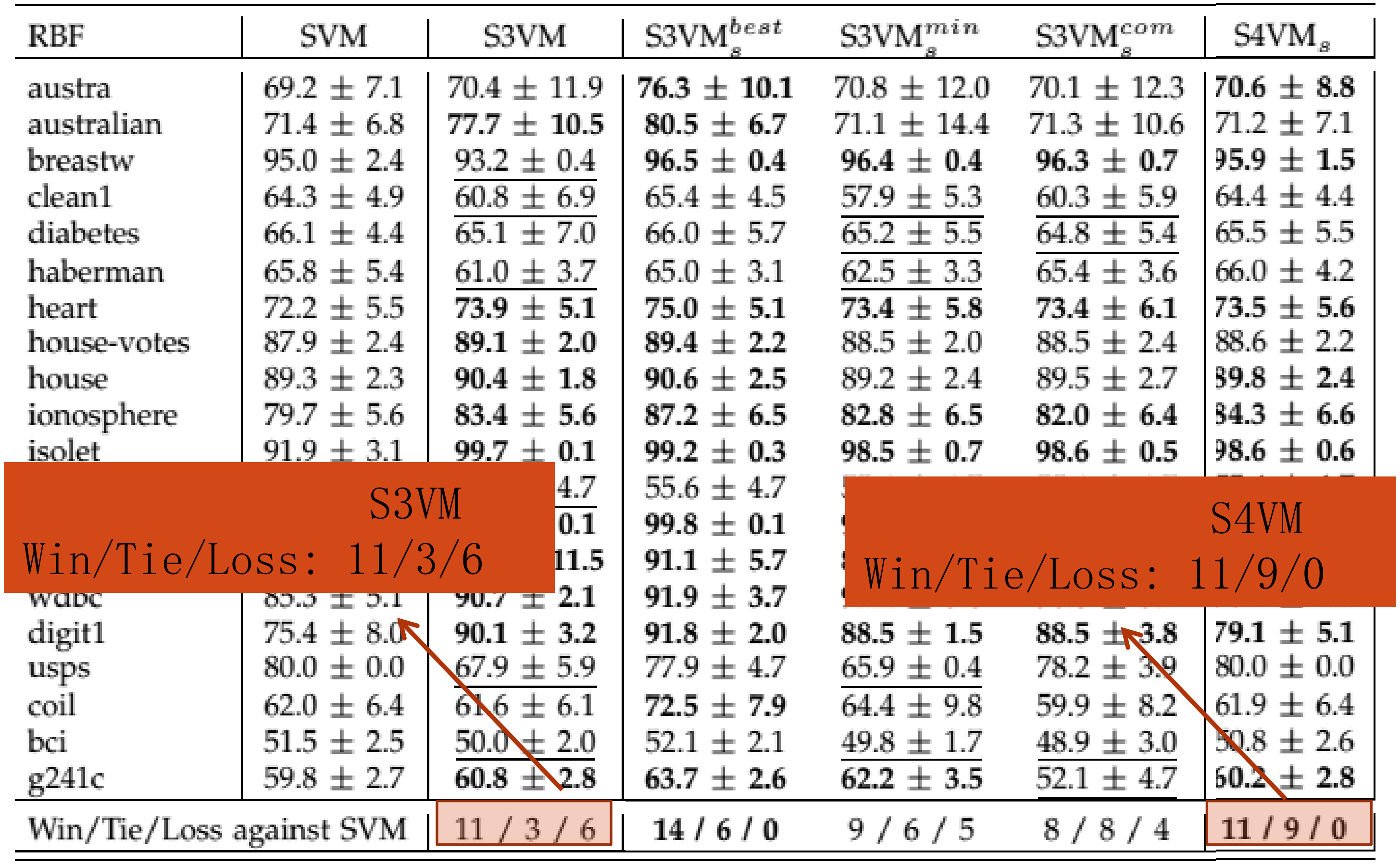
Linear	SVM	S3VM	S3VM$_s^{best}$	S3VM$_s^{min}$	S3VM$_s^{com}$	S4VM$_s$
austra	69.9 ± 7.6	69.6 ± 10.8	**71.7 ± 9.5**	70.6 ± 9.7	70.7 ± 9.8	70.7 ± 9.5
australian	75.2 ± 8.6	77.4 ± 9.3	**80.2 ± 6.7**	76.0 ± 10.3	73.5 ± 10.2	75.2 ± 8.7
breastw	94.3 ± 2.0	93.3 ± 0.4	**95.9 ± 1.7**	**95.8 ± 1.7**	93.9 ± 3.4	**95.0 ± 2.0**
clean1	59.0 ± 6.2	57.6 ± 6.8	**64.7 ± 4.2**	57.8 ± 4.6	57.5 ± 6.1	59.2 ± 5.3
diabetes	65.5 ± 5.0	64.8 ± 8.3	66.2 ± 5.1	65.3 ± 6.0	64.9 ± 5.7	65.9 ± 5.4
haberman	63.5 ± 7.6	61.7 ± 5.0	64.4 ± 4.6	62.7 ± 4.3	61.9 ± 6.8	63.8 ± 5.7
heart	71.1 ± 6.5	73.1 ± 6.5	**72.4 ± 6.4**	**72.1 ± 6.3**	**71.9 ± 6.2**	**72.1 ± 6.3**
house-votes	87.8 ± 3.3	**89.4 ± 4.5**	**91.9 ± 3.8**	**90.3 ± 5.4**	88.9 ± 4.4	**89.3 ± 3.9**
house	90.1 ± 3.7	**91.9 ± 3.2**	**95.6 ± 2.9**	**92.6 ± 4.7**	90.2 ± 3.7	**90.7 ± 4.1**
ionosphere	74.0 ± 5.7	74.5 ± 4.7	**79.8 ± 4.5**	75.3 ± 5.2	75.6 ± 5.1	**76.0 ± 5.6**
isolet	92.3 ± 3.3	**99.7 ± 0.1**	**99.6 ± 0.1**	**99.5 ± 0.1**	**99.4 ± 0.1**	**98.6 ± 2.7**
liverDisorders	54.3 ± 4.6	53.7 ± 4.9	53.6 ± 4.3	53.2 ± 4.5	52.2 ± 6.3	53.5 ± 4.3
optdigits	95.4 ± 2.3	**99.8 ± 0.0**	**99.7 ± 0.1**	**99.7 ± 0.1**	95.3 ± 6.9	**98.4 ± 1.9**
vehicle	78.6 ± 6.6	**84.5 ± 9.2**	**84.5 ± 6.6**	**83.2 ± 8.0**	**82.4 ± 8.0**	**82.4 ± 7.7**
wdbc	85.2 ± 5.7	**91.1 ± 2.8**	**89.5 ± 5.4**	**89.3 ± 5.4**	**89.1 ± 5.4**	**89.2 ± 5.5**
digit1	76.4 ± 5.4	**84.3 ± 1.7**	**83.2 ± 2.8**	**81.2 ± 4.3**	69.8 ± 5.8	76.4 ± 5.4
usps	78.2 ± 4.9	74.5 ± 5.9	**82.7 ± 1.7**	74.7 ± 6.3	77.6 ± 2.4	78.6 ± 4.1
coil	58.1 ± 6.1	57.5 ± 5.5	**66.9 ± 4.8**	58.8 ± 6.7	56.2 ± 7.0	57.9 ± 6.2
bci	54.2 ± 5.6	52.2 ± 3.7	54.9 ± 4.3	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	**83.7 ± 1.3**	**65.2 ± 3.5**	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	74.4 ± 3.1	75.2 ± 3.2	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	94.9 ± 2.4	**96.5 ± 1.9**	**96.2 ± 2.0**	**96.2 ± 2.0**	**96.4 ± 2.3**
mnist3vs8	81.1 ± 6.8	**82.4 ± 6.6**	**84.8 ± 7.0**	**84.2 ± 7.3**	**84.2 ± 7.3**	**84.0 ± 7.1**
mnist4vs9	73.9 ± 5.6	**74.7 ± 5.4**	**76.7 ± 6.7**	**75.8 ± 6.9**	**75.8 ± 6.9**	**75.8 ± 6.7**
mnist7vs9	79.2 ± 5.9	**80.5 ± 6.2**	**83.5 ± 7.5**	**82.9 ± 7.7**	**82.9 ± 7.7**	**82.6 ± 7.5**
real-sim	73.5 ± 2.8	74.0 ± 4.1	**75.5 ± 4.4**	**75.3 ± 4.5**	**75.3 ± 4.5**	**75.6 ± 4.1**
rcv1	69.5 ± 5.1	**71.4 ± 4.9**	**73.6 ± 5.7**	**73.5 ± 5.8**	**73.5 ± 5.8**	**73.3 ± 5.7**
Win/Tie/Loss against SVM		12 / 12 / 3	**22 / 5 / 0**	16 / 10 / 1	9 / 14 / 4	**16 / 11 / 0**

S3VM
Win/Tie/Loss: 12/12/3

S4VM
Win/Tie/Loss: 16/11/0

S4VM [Li and Zhou, ICML11]

Non-linear Kernel

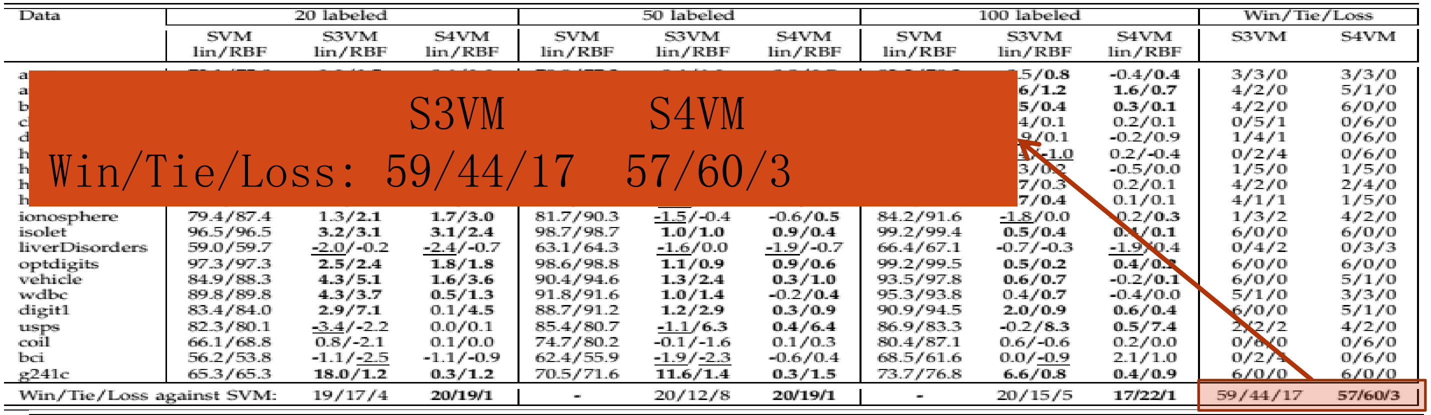


RBF	SVM	S3VM	S3VM$_s^{best}$	S3VM$_s^{min}$	S3VM$_s^{com}$	S4VM$_s$
austra	69.2 ± 7.1	70.4 ± 11.9	**76.3 ± 10.1**	70.8 ± 12.0	70.1 ± 12.3	**70.6 ± 8.8**
australian	71.4 ± 6.8	**77.7 ± 10.5**	**80.5 ± 6.7**	71.1 ± 14.4	71.3 ± 10.6	71.2 ± 7.1
breastw	95.0 ± 2.4	93.2 ± 0.4	**96.5 ± 0.4**	**96.4 ± 0.4**	**96.3 ± 0.7**	**95.9 ± 1.5**
clean1	64.3 ± 4.9	60.8 ± 6.9	65.4 ± 4.5	57.9 ± 5.3	60.3 ± 5.9	64.4 ± 4.4
diabetes	66.1 ± 4.4	65.1 ± 7.0	66.0 ± 5.7	65.2 ± 5.5	64.8 ± 5.4	65.5 ± 5.5
haberman	65.8 ± 5.4	61.0 ± 3.7	65.0 ± 3.1	62.5 ± 3.3	65.4 ± 3.6	66.0 ± 4.2
heart	72.2 ± 5.5	**73.9 ± 5.1**	**75.0 ± 5.1**	**73.4 ± 5.8**	**73.4 ± 6.1**	**73.5 ± 5.6**
house-votes	87.9 ± 2.4	**89.1 ± 2.0**	**89.4 ± 2.2**	88.5 ± 2.0	88.5 ± 2.4	88.6 ± 2.2
house	89.3 ± 2.3	**90.4 ± 1.8**	**90.6 ± 2.5**	89.2 ± 2.4	89.5 ± 2.7	**89.8 ± 2.4**
ionosphere	79.7 ± 5.6	**83.4 ± 5.6**	**87.2 ± 6.5**	**82.8 ± 6.5**	**82.0 ± 6.4**	**84.3 ± 6.6**
isolet	91.9 ± 3.1	**99.7 ± 0.1**	**99.2 ± 0.3**	**98.5 ± 0.7**	**98.6 ± 0.5**	**98.6 ± 0.6**
[illegible]	[illegible]	[illegible]	55.6 ± 4.7	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	**99.8 ± 0.1**	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	**91.1 ± 5.7**	[illegible]	[illegible]	[illegible]
wdbc	85.3 ± 5.1	**90.7 ± 2.1**	**91.9 ± 3.7**	[illegible]	[illegible]	[illegible]
digit1	75.4 ± 8.0	**90.1 ± 3.2**	**91.8 ± 2.0**	**88.5 ± 1.5**	**88.5 ± 3.8**	**79.1 ± 5.1**
usps	80.0 ± 0.0	67.9 ± 5.9	77.9 ± 4.7	65.9 ± 0.4	78.2 ± 3.9	80.0 ± 0.0
coil	62.0 ± 6.4	61.6 ± 6.1	**72.5 ± 7.9**	64.4 ± 9.8	59.9 ± 8.2	61.9 ± 6.4
bci	51.5 ± 2.5	50.0 ± 2.0	52.1 ± 2.1	49.8 ± 1.7	48.9 ± 3.0	50.8 ± 2.6
g241c	59.8 ± 2.7	**60.8 ± 2.8**	**63.7 ± 2.6**	**62.2 ± 3.5**	52.1 ± 4.7	**60.2 ± 2.8**
Win/Tie/Loss against SVM		11 / 3 / 6	**14 / 6 / 0**	9 / 6 / 5	8 / 8 / 4	**11 / 9 / 0**

S3VM
Win/Tie/Loss: 11/3/6

S4VM
Win/Tie/Loss: 11/9/0

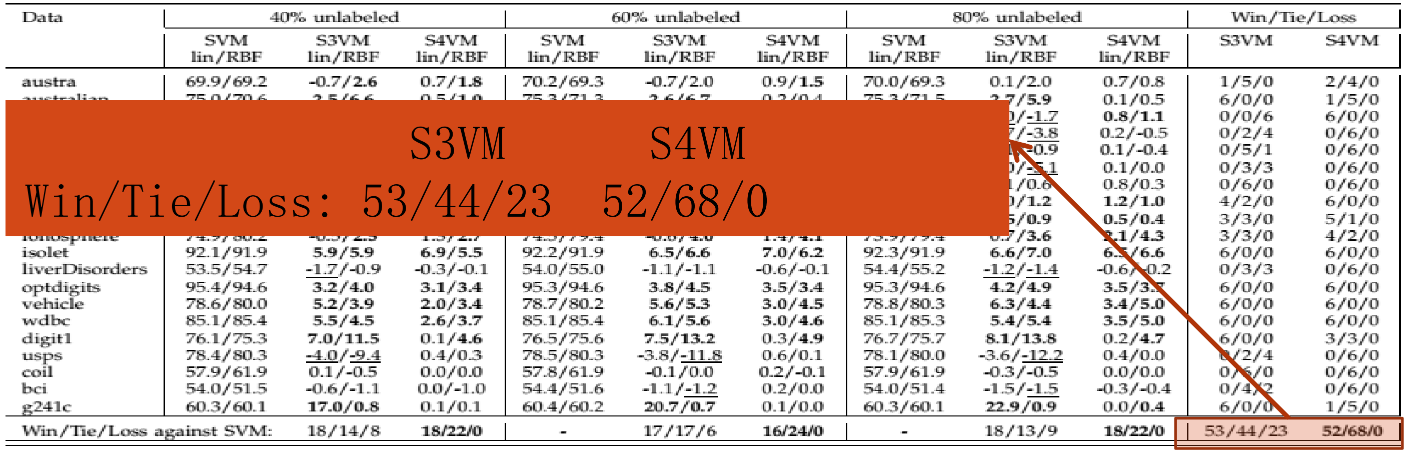
Influence of the amount of labeled and unlabeled data



S3VM S4VM

Win/Tie/Loss: 59/44/17 57/60/3

Data	20 labeled			50 labeled			100 labeled			Win/Tie/Loss	
	SVM lin/RBF	S3VM lin/RBF	S4VM lin/RBF	SVM lin/RBF	S3VM lin/RBF	S4VM lin/RBF	SVM lin/RBF	S3VM lin/RBF	S4VM lin/RBF	S3VM	S4VM
a	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.5/0.8	-0.4/0.4	3/3/0	3/3/0
a	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6/1.2	1.6/0.7	4/2/0	5/1/0
b	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5/0.4	0.3/0.1	4/2/0	6/0/0
c	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4/0.1	0.2/0.1	0/5/1	0/6/0
d	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	9/0.1	-0.2/0.9	1/4/1	0/6/0
h	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4/-1.0	0.2/-0.4	0/2/4	0/6/0
h	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3/0.2	-0.5/0.0	1/5/0	1/5/0
h	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7/0.3	0.2/0.1	4/2/0	2/4/0
h	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7/0.4	0.1/0.1	4/1/1	1/5/0
ionosphere	79.4/87.4	1.3/2.1	1.7/3.0	81.7/90.3	-1.5/-0.4	-0.6/0.5	84.2/91.6	-1.8/0.0	0.2/0.3	1/3/2	4/2/0
isolet	96.5/96.5	3.2/3.1	3.1/2.4	98.7/98.7	1.0/1.0	0.9/0.4	99.2/99.4	0.5/0.4	0.4/0.1	6/0/0	6/0/0
liverDisorders	59.0/59.7	-2.0/-0.2	-2.4/-0.7	63.1/64.3	-1.6/0.0	-1.9/-0.7	66.4/67.1	-0.7/-0.3	-1.9/0.4	0/4/2	0/3/3
optdigits	97.3/97.3	2.5/2.4	1.8/1.8	98.6/98.8	1.1/0.9	0.9/0.6	99.2/99.5	0.5/0.2	0.4/0.2	6/0/0	6/0/0
vehicle	84.9/88.3	4.3/5.1	1.6/3.6	90.4/94.6	1.3/2.4	0.3/1.0	93.5/97.8	0.6/0.7	-0.2/0.1	6/0/0	5/1/0
wdbc	89.8/89.8	4.3/3.7	0.5/1.3	91.8/91.6	1.0/1.4	-0.2/0.4	95.3/93.8	0.4/0.7	-0.4/0.0	5/1/0	3/3/0
digit1	83.4/84.0	2.9/7.1	0.1/4.5	88.7/91.2	1.2/2.9	0.3/0.9	90.9/94.5	2.0/0.9	0.6/0.4	6/0/0	5/1/0
usps	82.3/80.1	-3.4/-2.2	0.0/0.1	85.4/80.7	-1.1/6.3	0.4/6.4	86.9/83.3	-0.2/8.3	0.5/7.4	2/2/2	4/2/0
coil	66.1/68.8	0.8/-2.1	0.1/0.0	74.7/80.2	-0.1/-1.6	0.1/0.3	80.4/87.1	0.6/-0.6	0.2/0.0	0/6/0	0/6/0
bci	56.2/53.8	-1.1/-2.5	-1.1/-0.9	62.4/55.9	-1.9/-2.3	-0.6/0.4	68.5/61.6	0.0/-0.9	2.1/1.0	0/2/4	0/6/0
g241c	65.3/65.3	18.0/1.2	0.3/1.2	70.5/71.6	11.6/1.4	0.3/1.5	73.7/76.8	6.6/0.8	0.4/0.9	6/0/0	6/0/0
Win/Tie/Loss against SVM:		19/17/4	20/19/1	-	20/12/8	20/19/1	-	20/15/5	17/22/1	59/44/17	57/60/3



S3VM S4VM

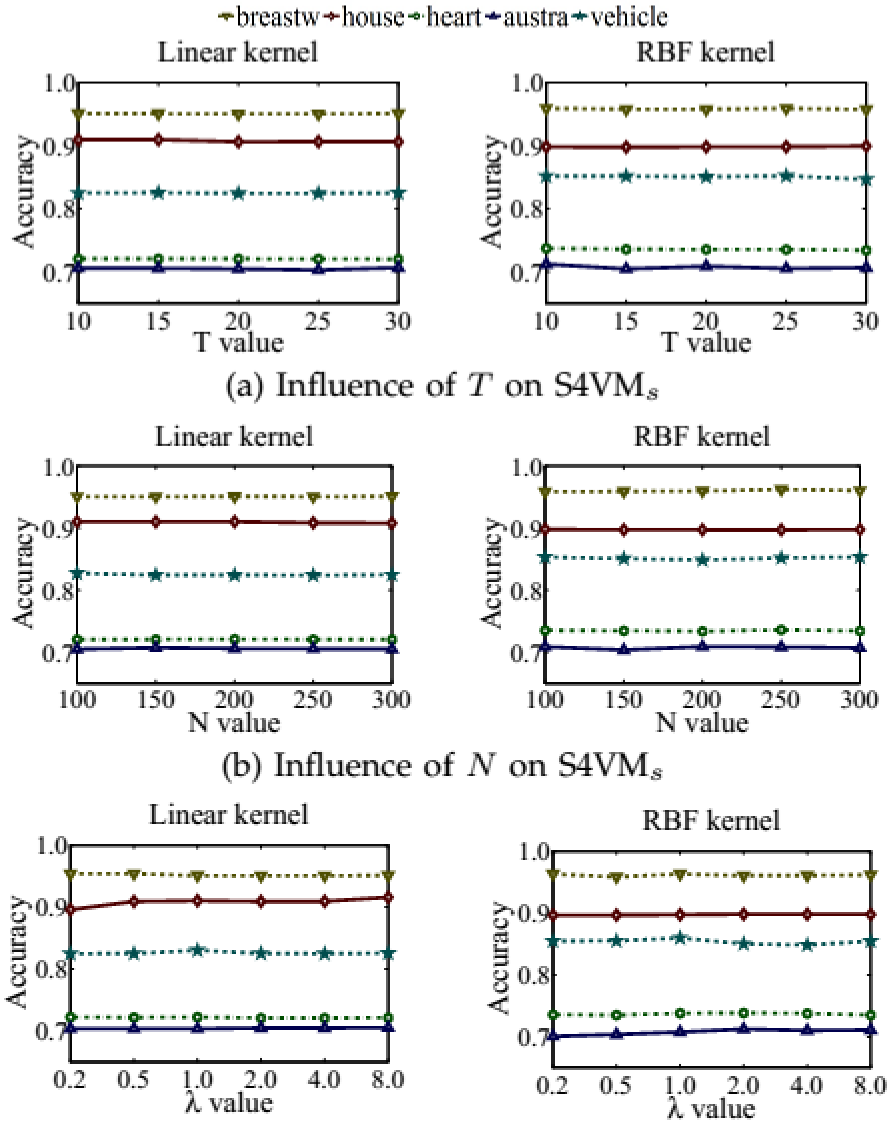
Win/Tie/Loss: 53/44/23 52/68/0

Data	40% unlabeled			60% unlabeled			80% unlabeled			Win/Tie/Loss	
	SVM lin/RBF	S3VM lin/RBF	S4VM lin/RBF	SVM lin/RBF	S3VM lin/RBF	S4VM lin/RBF	SVM lin/RBF	S3VM lin/RBF	S4VM lin/RBF	S3VM	S4VM
austra	69.9/69.2	-0.7/2.6	0.7/1.8	70.2/69.3	-0.7/2.0	0.9/1.5	70.0/69.3	0.1/2.0	0.7/0.8	1/5/0	2/4/0
australian	75.0/70.6	2.5/6.6	0.5/1.0	75.3/71.3	2.6/6.7	0.2/0.4	75.3/71.5	2.7/5.9	0.1/0.5	6/0/0	1/5/0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0/-1.7	0.8/1.1	0/0/6	6/0/0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7/-3.8	0.2/-0.5	0/2/4	0/6/0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.9	0.1/-0.4	0/5/1	0/6/0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0/-5.1	0.1/0.0	0/3/3	0/6/0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1/0.6	0.8/0.3	0/6/0	0/6/0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0/1.2	1.2/1.0	4/2/0	6/0/0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5/0.9	0.5/0.4	3/3/0	5/1/0
ionosphere	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.7/3.6	2.1/4.3	3/3/0	4/2/0
isolet	92.1/91.9	5.9/5.9	6.9/5.5	92.2/91.9	6.5/6.6	7.0/6.2	92.3/91.9	6.6/7.0	6.5/6.6	6/0/0	6/0/0
liverDisorders	53.5/54.7	-1.7/-0.9	-0.3/-0.1	54.0/55.0	-1.1/-1.1	-0.6/-0.1	54.4/55.2	-1.2/-1.4	-0.6/-0.2	0/3/3	0/6/0
optdigits	95.4/94.6	3.2/4.0	3.1/3.4	95.3/94.6	3.8/4.5	3.5/3.4	95.3/94.6	4.2/4.9	3.5/3.7	6/0/0	6/0/0
vehicle	78.6/80.0	5.2/3.9	2.0/3.4	78.7/80.2	5.6/5.3	3.0/4.5	78.8/80.3	6.3/4.4	3.4/5.0	6/0/0	6/0/0
wdbc	85.1/85.4	5.5/4.5	2.6/3.7	85.1/85.4	6.1/5.6	3.0/4.6	85.1/85.3	5.4/5.4	3.5/5.0	6/0/0	6/0/0
digit1	76.1/75.3	7.0/11.5	0.1/4.6	76.5/75.6	7.5/13.2	0.3/4.9	76.7/75.7	8.1/13.8	0.2/4.7	6/0/0	3/3/0
usps	78.4/80.3	-4.0/-9.4	0.4/0.3	78.5/80.3	-3.8/-11.8	0.6/0.1	78.1/80.0	-3.6/-12.2	0.4/0.0	0/2/4	0/6/0
coil	57.9/61.9	0.1/-0.5	0.0/0.0	57.8/61.9	-0.1/0.0	0.2/-0.1	57.9/61.9	-0.3/-0.5	0.0/0.0	0/6/0	0/6/0
bci	54.0/51.5	-0.6/-1.1	0.0/-1.0	54.4/51.6	-1.1/-1.2	0.2/0.0	54.0/51.4	-1.5/-1.5	-0.3/-0.4	0/4/2	0/6/0
g241c	60.3/60.1	17.0/0.8	0.1/0.1	60.4/60.2	20.7/0.7	0.1/0.0	60.3/60.1	22.9/0.9	0.0/0.4	6/0/0	1/5/0
Win/Tie/Loss against SVM:		18/14/8	18/22/0	-	17/17/6	16/24/0	-	18/13/9	18/22/0	53/44/23	52/68/0

- S4VMs are highly competitive with S3VMs on varied amounts of labeled and unlabeled data.
- S4VMs are inferior to inductive SVM on only 3 over the 240 cases; whereas S3VM degenerates performance on 40 over the 240 cases.

S4VM [Li and Zhou, ICML11]

Influence of Parameters



(a) Influence of T on S4VM$_s$

(b) Influence of N on S4VM$_s$

(c) Influence of λ on S4VM$_s$

- S4VMs are quite insensitive to parameters
- This property makes S4VMs even more attractive, especially when the number of labeled examples is too few to afford a reliable model selection.

Outline

- Scalability of S3VMs “多”
  - WellSVM [Li et al., JMLR13]

- Efficiency of S3VMs “快”
  - MeanS3VM [Li et al., ICML09]

- Safeness of S3VMs “好”
  - S4VM [Li and Zhou, ICML11]

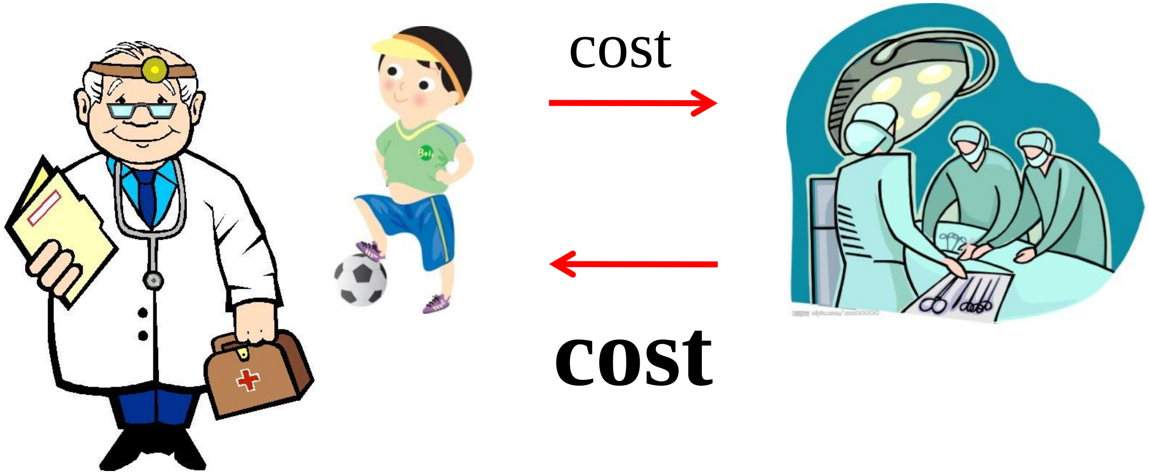
- Cost sensitivity of S3VMs “省”
  - CS4VM [Li et al., AAAI10]

CS4VM [Li et al., AAAI10]

Cost-Sensitive with Unlabeled Data

- In many applications, two phenomena may occur simultaneously
  - Different errors are associated with different cost.
  - Many training data are unlabeled

- Disease Diagnosis



cost

cost

Unlabeled instances

More application: Fraud detection.

Related Works

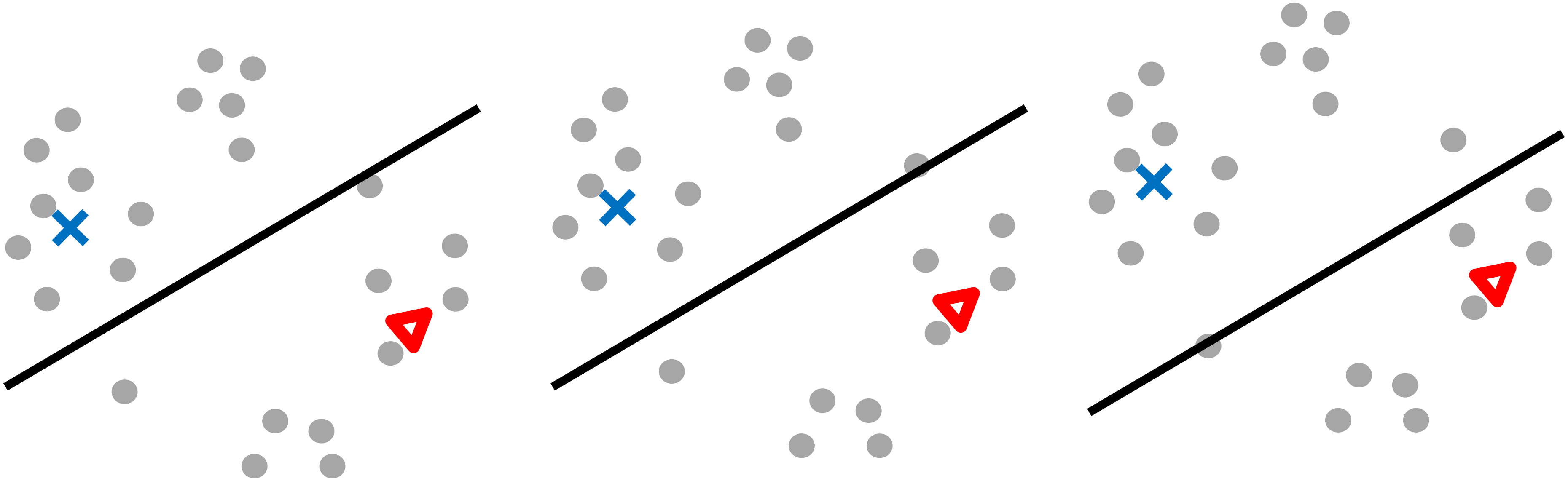
- The use of unlabeled data in cost-sensitive learning has been considered in a few studies [Greiner, Grove, and Roth 2002; Margineantu 2005; Liu, Jun, and Ghosh 2009; Qin et al. 2008].

- Most of which try to involve human feedback on informative unlabeled instances and then refine the cost sensitive model using the queried labels.

- To our best knowledge, S3VMs with unequal costs of unlabeled data have not been studied before.

Observation

Cost: ▽ = 1



Cost: × = 0.5

Cost: × = 1

Cost: × = 2

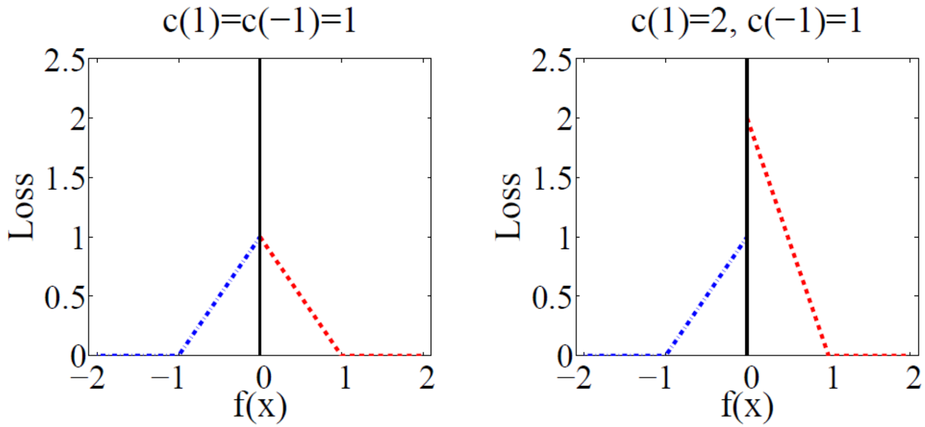
CS4VM [Li et al., AAAI10]

CS4VM (Cost-Sensitive S3VM)

$$\begin{aligned} \min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i\in\mathcal{I}_l} c(y_i)\xi_i + C_2 \sum_{i\in\mathcal{I}_u} c(\hat{y}_i)\xi_i, \\ \text{s.t.} \quad & y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i, i \in \mathcal{I}_l, \\ & \hat{y}_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i, i \in \mathcal{I}_u, \end{aligned}$$

- $c(y)$: cost of label y

S3VM ⟺



When $c(1) \neq c(-1)$, the loss is no longer continuous.

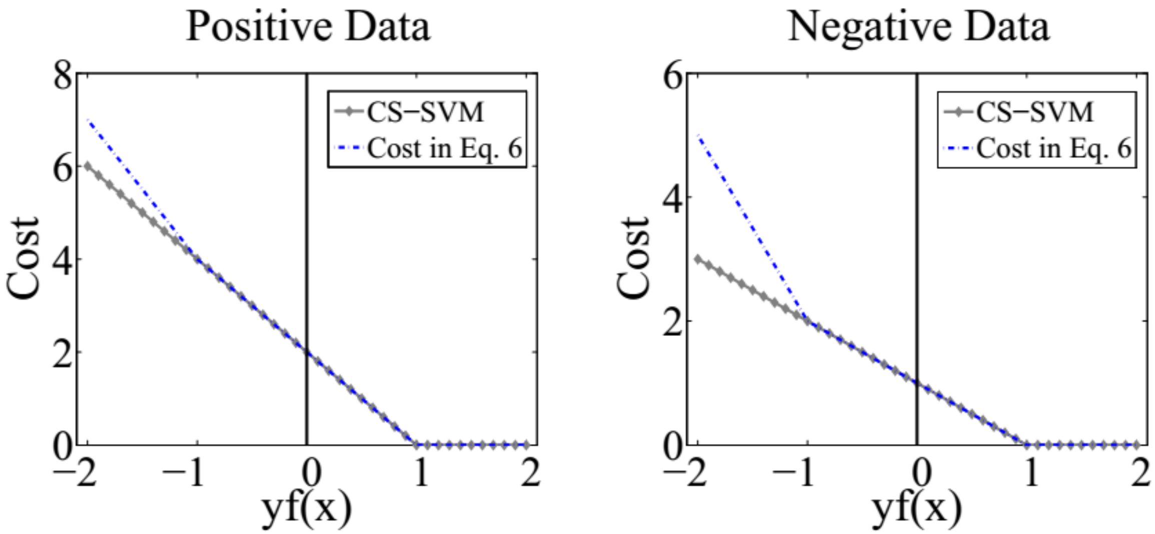
CS4VM [Li et al., AAAI10]

Usefulness of Label Means

Learning And Mining from DatA

http://lamda.nju.edu.cn

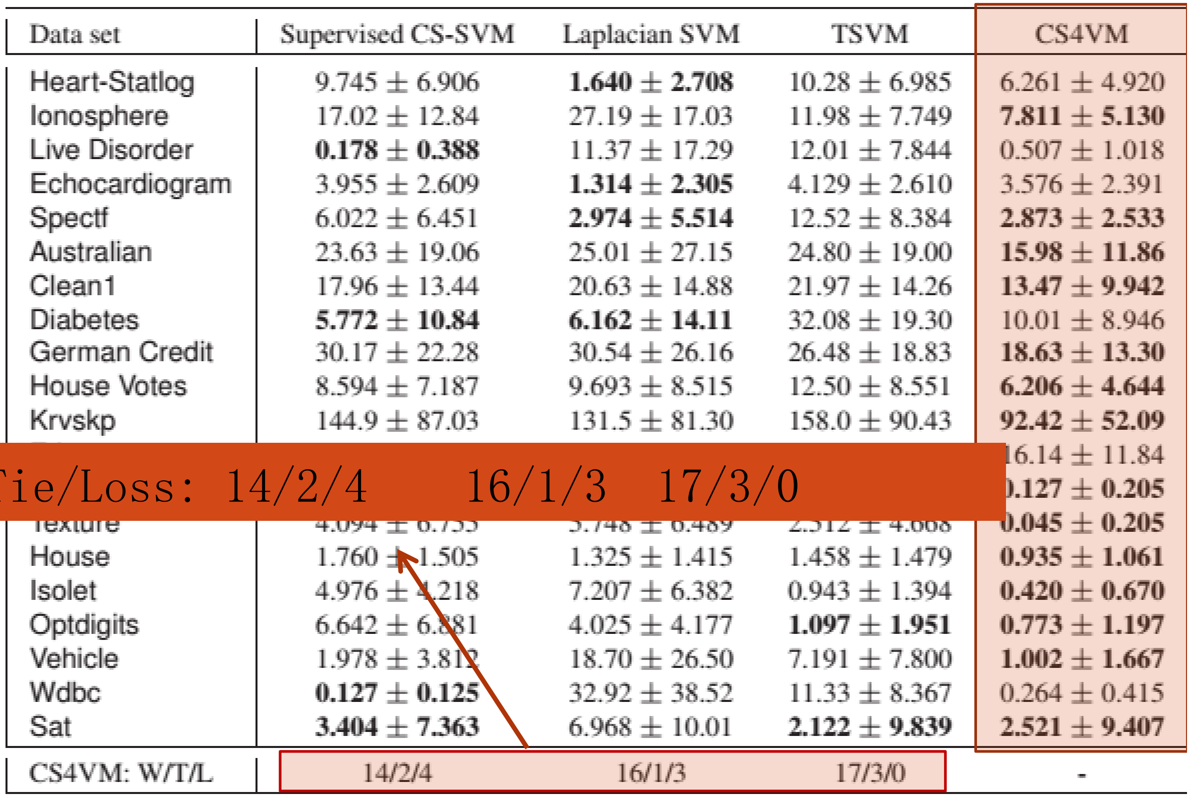
Suppose that f^* *is the optimal solution of CS4VM. When all the unlabeled data do not suffer from large loss, i.e.,* $y_i^* f^*(\mathbf{x}_i) \geq -1, \forall i \in \mathcal{I}_u$, *CS4VM is equivalent to the CS-SVM. Otherwise, let* $\hat{\ell}(\mathbf{x}_i)$ *be the loss for the unlabeled instance* $\mathbf{x}_i$ *in CS4VM. Then,* $\hat{\ell}(\mathbf{x}_i) \leq \frac{c(1)+c(-1)}{c(y_i^*)} \ell(y_i^*, f(\mathbf{x}_i))$.



- With the knowledge of label means, CS4VM is closely related to supervised CS-SVM with the knowledge of all the labels.

CS4VM [Li et al., AAAI10]

Experiments



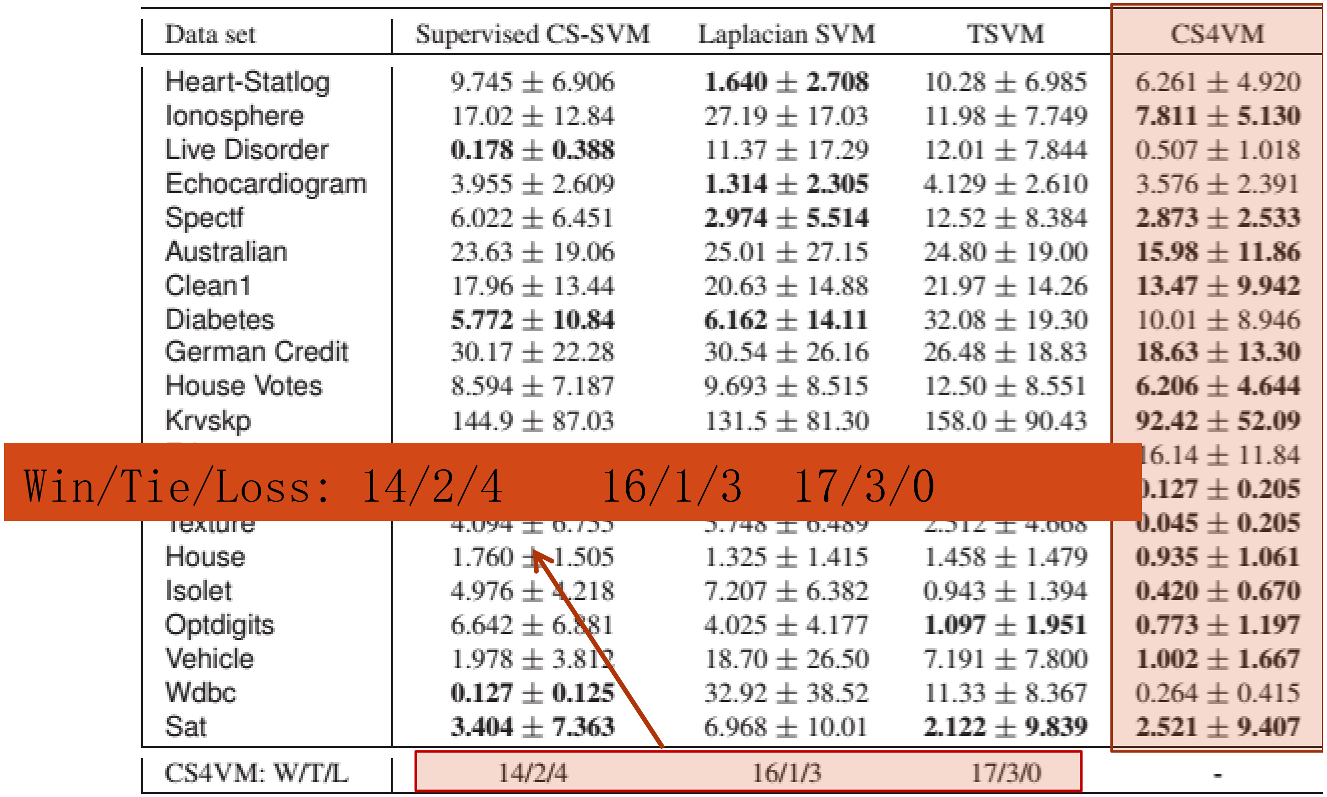
Data set	Supervised CS-SVM	Laplacian SVM	TSVM	CS4VM
Heart-Statlog	9.745 ± 6.906	**1.640 ± 2.708**	10.28 ± 6.985	6.261 ± 4.920
Ionosphere	17.02 ± 12.84	27.19 ± 17.03	11.98 ± 7.749	**7.811 ± 5.130**
Live Disorder	**0.178 ± 0.388**	11.37 ± 17.29	12.01 ± 7.844	0.507 ± 1.018
Echocardiogram	3.955 ± 2.609	**1.314 ± 2.305**	4.129 ± 2.610	3.576 ± 2.391
Spectf	6.022 ± 6.451	**2.974 ± 5.514**	12.52 ± 8.384	**2.873 ± 2.533**
Australian	23.63 ± 19.06	25.01 ± 27.15	24.80 ± 19.00	**15.98 ± 11.86**
Clean1	17.96 ± 13.44	20.63 ± 14.88	21.97 ± 14.26	**13.47 ± 9.942**
Diabetes	**5.772 ± 10.84**	**6.162 ± 14.11**	32.08 ± 19.30	10.01 ± 8.946
German Credit	30.17 ± 22.28	30.54 ± 26.16	26.48 ± 18.83	**18.63 ± 13.30**
House Votes	8.594 ± 7.187	9.693 ± 8.515	12.50 ± 8.551	**6.206 ± 4.644**
Krvskp	144.9 ± 87.03	131.5 ± 81.30	158.0 ± 90.43	**92.42 ± 52.09**
[illegible]	[illegible]	[illegible]	[illegible]	16.14 ± 11.84
[illegible]	[illegible]	[illegible]	[illegible]	**0.127 ± 0.205**
Texture	4.094 ± 6.755	3.748 ± 6.489	2.312 ± 4.668	**0.045 ± 0.205**
House	1.760 ± 1.505	1.325 ± 1.415	1.458 ± 1.479	**0.935 ± 1.061**
Isolet	4.976 ± 4.218	7.207 ± 6.382	0.943 ± 1.394	**0.420 ± 0.670**
Optdigits	6.642 ± 6.881	4.025 ± 4.177	**1.097 ± 1.951**	**0.773 ± 1.197**
Vehicle	1.978 ± 3.812	18.70 ± 26.50	7.191 ± 7.800	**1.002 ± 1.667**
Wdbc	**0.127 ± 0.125**	32.92 ± 38.52	11.33 ± 8.367	0.264 ± 0.415
Sat	**3.404 ± 7.363**	6.968 ± 10.01	**2.122 ± 9.839**	**2.521 ± 9.407**
CS4VM: W/T/L	14/2/4	16/1/3	17/3/0	-

Win/Tie/Loss: 14/2/4 16/1/3 17/3/0

- c(-1)=1, c(1) is chosen randomly from [0, 1000] in a uniform distribution.
- Experiments repeat for 100 times.

CS4VM [Li et al., AAAI10]

Experiments

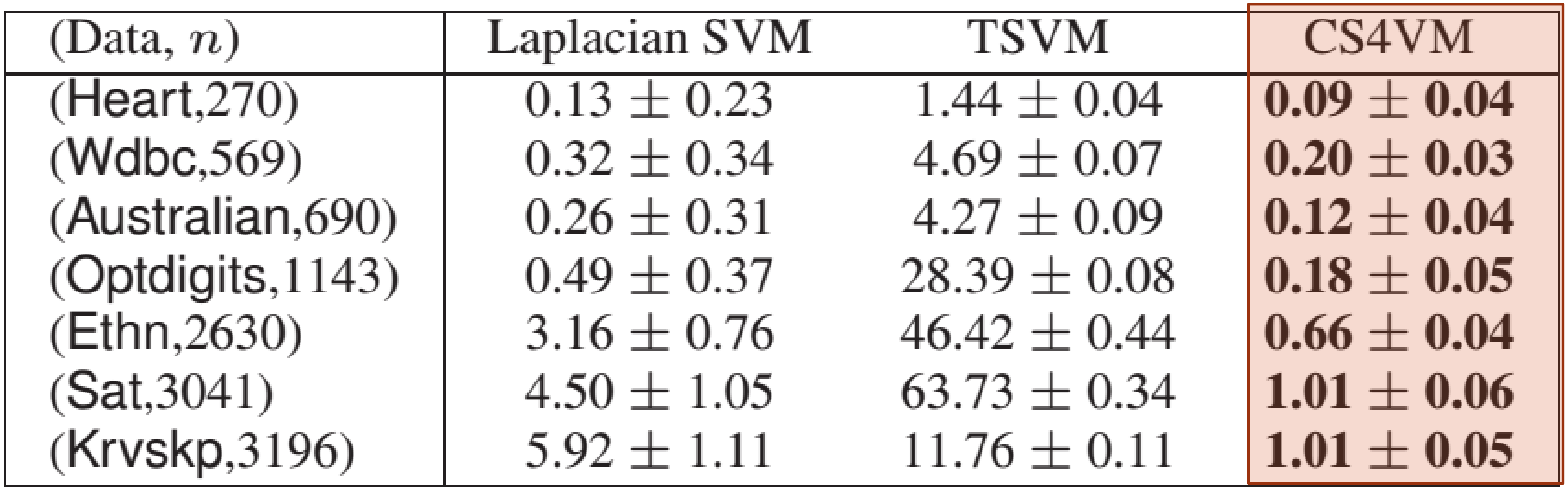


Data set	Supervised CS-SVM	Laplacian SVM	TSVM	CS4VM
Heart-Statlog	9.745 ± 6.906	**1.640 ± 2.708**	10.28 ± 6.985	6.261 ± 4.920
Ionosphere	17.02 ± 12.84	27.19 ± 17.03	11.98 ± 7.749	**7.811 ± 5.130**
Live Disorder	**0.178 ± 0.388**	11.37 ± 17.29	12.01 ± 7.844	0.507 ± 1.018
Echocardiogram	3.955 ± 2.609	**1.314 ± 2.305**	4.129 ± 2.610	3.576 ± 2.391
Spectf	6.022 ± 6.451	**2.974 ± 5.514**	12.52 ± 8.384	**2.873 ± 2.533**
Australian	23.63 ± 19.06	25.01 ± 27.15	24.80 ± 19.00	**15.98 ± 11.86**
Clean1	17.96 ± 13.44	20.63 ± 14.88	21.97 ± 14.26	**13.47 ± 9.942**
Diabetes	**5.772 ± 10.84**	**6.162 ± 14.11**	32.08 ± 19.30	10.01 ± 8.946
German Credit	30.17 ± 22.28	30.54 ± 26.16	26.48 ± 18.83	**18.63 ± 13.30**
House Votes	8.594 ± 7.187	9.693 ± 8.515	12.50 ± 8.551	**6.206 ± 4.644**
Krvskp	144.9 ± 87.03	131.5 ± 81.30	158.0 ± 90.43	**92.42 ± 52.09**
[illegible]	[illegible]	[illegible]	[illegible]	16.14 ± 11.84
[illegible]	[illegible]	[illegible]	[illegible]	**0.127 ± 0.205**
Texture	4.094 ± 6.755	3.748 ± 6.489	2.312 ± 4.668	**0.045 ± 0.205**
House	1.760 ± 1.505	1.325 ± 1.415	1.458 ± 1.479	**0.935 ± 1.061**
Isolet	4.976 ± 4.218	7.207 ± 6.382	0.943 ± 1.394	**0.420 ± 0.670**
Optdigits	6.642 ± 6.881	4.025 ± 4.177	**1.097 ± 1.951**	**0.773 ± 1.197**
Vehicle	1.978 ± 3.812	18.70 ± 26.50	7.191 ± 7.800	**1.002 ± 1.667**
Wdbc	**0.127 ± 0.125**	32.92 ± 38.52	11.33 ± 8.367	0.264 ± 0.415
Sat	**3.404 ± 7.363**	6.968 ± 10.01	**2.122 ± 9.839**	**2.521 ± 9.407**
CS4VM: W/T/L	14/2/4	16/1/3	17/3/0	-

Win/Tie/Loss: 14/2/4 16/1/3 17/3/0

- CS4VM outperms Laplician SVM, TSVM and Supervised CS-SVM.
- Wilcoxon sign tests (at 95% significance level) show that CS4VM is always significantly better than Laplician SVM, TSVM and Supervised CS-SVM.

CPU Time



(Data, n)	Laplacian SVM	TSVM	CS4VM
(Heart,270)	0.13 ± 0.23	1.44 ± 0.04	**0.09 ± 0.04**
(Wdbc,569)	0.32 ± 0.34	4.69 ± 0.07	**0.20 ± 0.03**
(Australian,690)	0.26 ± 0.31	4.27 ± 0.09	**0.12 ± 0.04**
(Optdigits,1143)	0.49 ± 0.37	28.39 ± 0.08	**0.18 ± 0.05**
(Ethn,2630)	3.16 ± 0.76	46.42 ± 0.44	**0.66 ± 0.04**
(Sat,3041)	4.50 ± 1.05	63.73 ± 0.34	**1.01 ± 0.06**
(Krvskp,3196)	5.92 ± 1.11	11.76 ± 0.11	**1.01 ± 0.05**

- CS4VM is faster than Laplacian SVM and TSVM.

Summary

- Scalability of S3VMs

  “多”

  - WellSVM [Li et al., JMLR13]
    - A tight and convex relaxation of S3VMs
      - As least as tight as SDP convex relaxations
    - Can make use of state-of-the-art SVM software
      - Scalable
    - Empirical studies validate its promising performances and good scalability.
    - http://lamda.nju.edu.cn/code_WellSVM.ashx
- Efficiency of S3VMs

  “快”

  - MeanS3VM [Li et al., ICML09]
    - An approximation of S3VM
    - MeanS3VM + label means ≈ SVM + all the labels
      - Develop efficient algorithms
    - Empirical studies validate its promising performances and good computational efficiency.
    - http://lamda.nju.edu.cn/code_meanS3VM.ashx

Summary

- Safeness of S3VMs “好”
  - S4VM [Li and Zhou, ICML11]
    - Study safe S3VMs
    - Under the assumption employed in S3VMs, S4VMs are provably safe and able to achieve the largest performance improvement
    - Comprehensive empirical studies validate their highly competitive performances and safeness.
    - http://lamda.nju.edu.cn/code_S4VM.ashx
- Cost sensitivity of S3VMs
  - CS4VM [Li et al., AAAI10] “省”
    - Study cost-sensitive S3VMs
    - CS4VM + label means ≈ CS-SVM + all the labels
    - Empirical studies validate its encouraging performances and good computational efficiency.
    - http://lamda.nju.edu.cn/code_CS4VM.ashx

Future work

- The four methods are now proposed for individual goals. How to integrate the advantages of them into one method?
- How to have a guarantee of label mean estimations?
- The connection between safeness and generalization is unclear. How to relax the safeness assumption of S4VMs? How to develop safe graph-based methods?
- How to develop multi-class or multi-label cost-sensitive S3VMs?

Related References

- Yu-Feng Li, Ivor Tsang, James Kwok and Zhi-Hua Zhou. **Convex and Scalable Weakly Labeled SVMs**. Journal of Machine Learning Research, 14:2151-2188, 2013.
- Yu-Feng Li, James T. Kwok and Zhi-Hua Zhou. **Semi-supervised learning using label mean**. In: *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*, Montreal, Canada, 2009, pp.633-640.
- Yu-Feng Li and Zhi-Hua Zhou. **Towards making unlabeled data never hurt**. In: *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, Bellevue, WA, 2011, pp.1081-1088.
- Yu-Feng Li, James T. Kwok, and Zhi-Hua Zhou. **Cost-sensitive semi-supervised support vector machine**. In: *Proceedings of the 24th AAAI Conference on Artificial Intelligences (AAAI'10)*, Atlanta, GE, 2010, pp.500-505.
- Yu-Feng Li and Zhi-Hua Zhou. **Improving semi-supervised support vector machines through unlabeled instances selection**. In: *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, San Francisco, CA, 2011, pp.386-391.
- Yu-Feng Li, Ivor W. Tsang, James T. Kwok, and Zhi-Hua Zhou. **Tighter and convex maximum margin clustering**. In: *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS'09)*, Clearwater Beach, FL, 2009, pp.328-335.
- Yu-Feng Li, James T. Kwok, Ivor W. Tsang, and Zhi-Hua Zhou. **A convex method for locating regions of interest with multi-instance learning**. In: *Proceedings of the 20th European Conference on Machine Learning (ECML'09)*, Bled, Slovenia, 2009, pp.17-32.
- Yu-Feng Li, Ju-Hua Hu, Yuang Jiang and Zhi-Hua Zhou. **Towards discovering what patterns trigger what labels**. In: *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, Toronto, Canada, 2012, pp.1012-1018.

Thanks!