$\propto$SVM for Learning with Label Proportions

Felix X. Yu[†] Dong Liu[†] Sanjiv Kumar[§] Tony Jebara[†] Shih-Fu Chang[†]

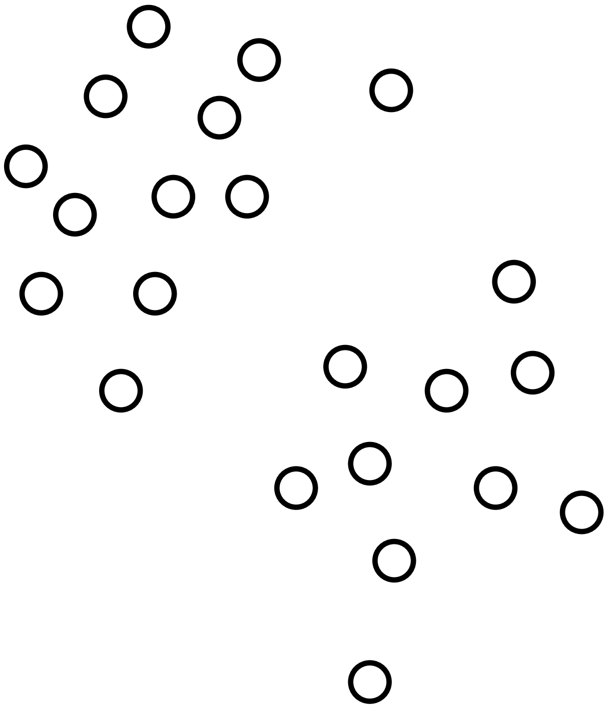
[†]Columbia University, New York, NY 10027
[§]Google Research, New York, NY 10011

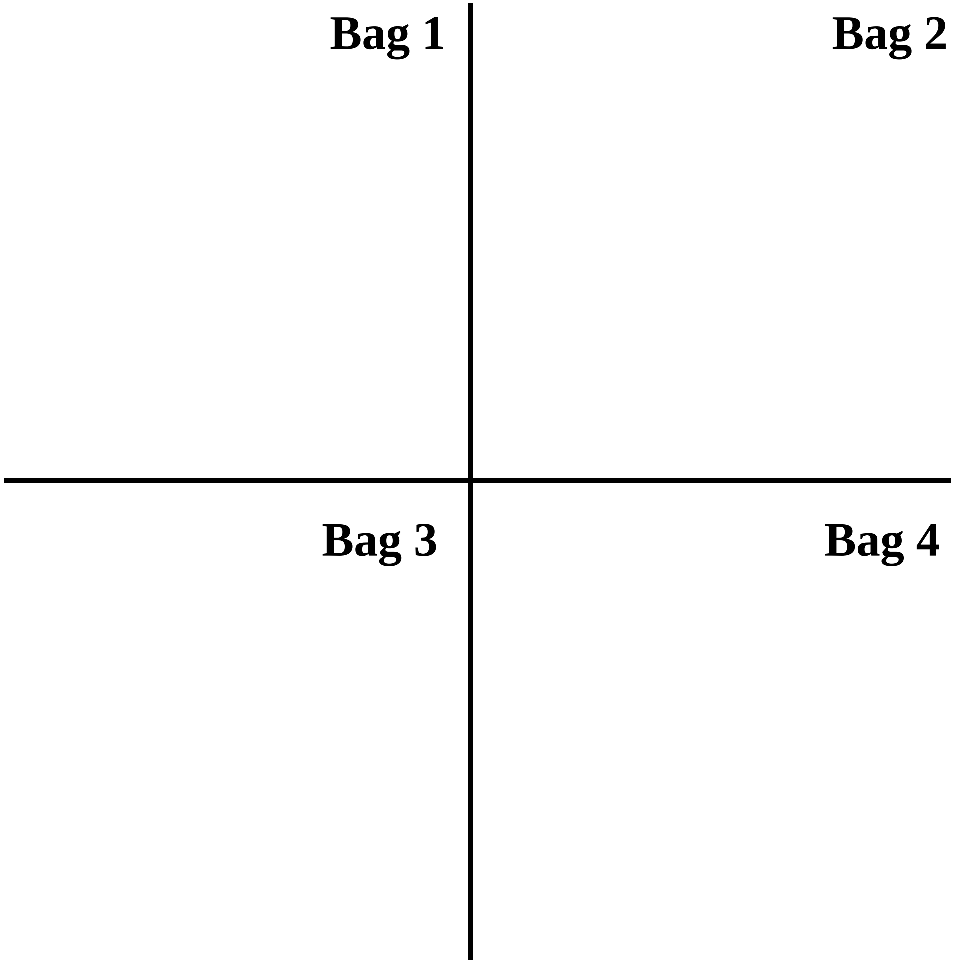
$\propto$ is the symbol for "proportional-to".

Outline

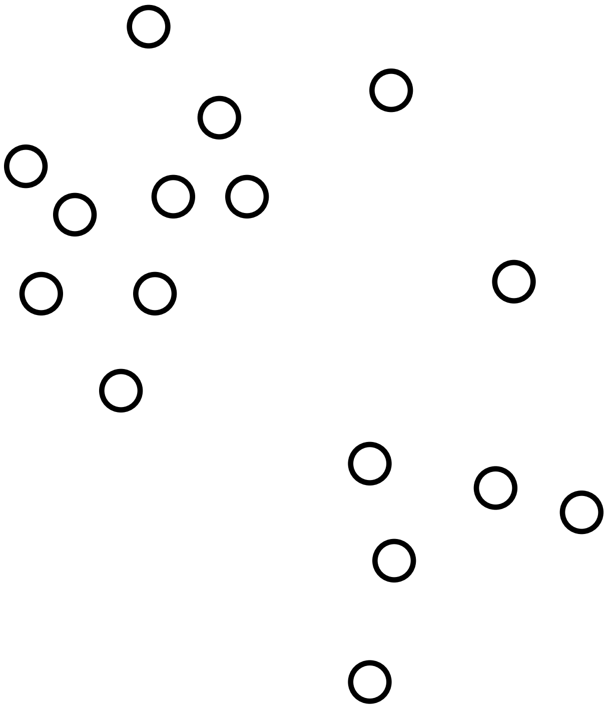
- Learning Setting and Applications
- Related Works
- Formulation
- Algorithms
- Experiments

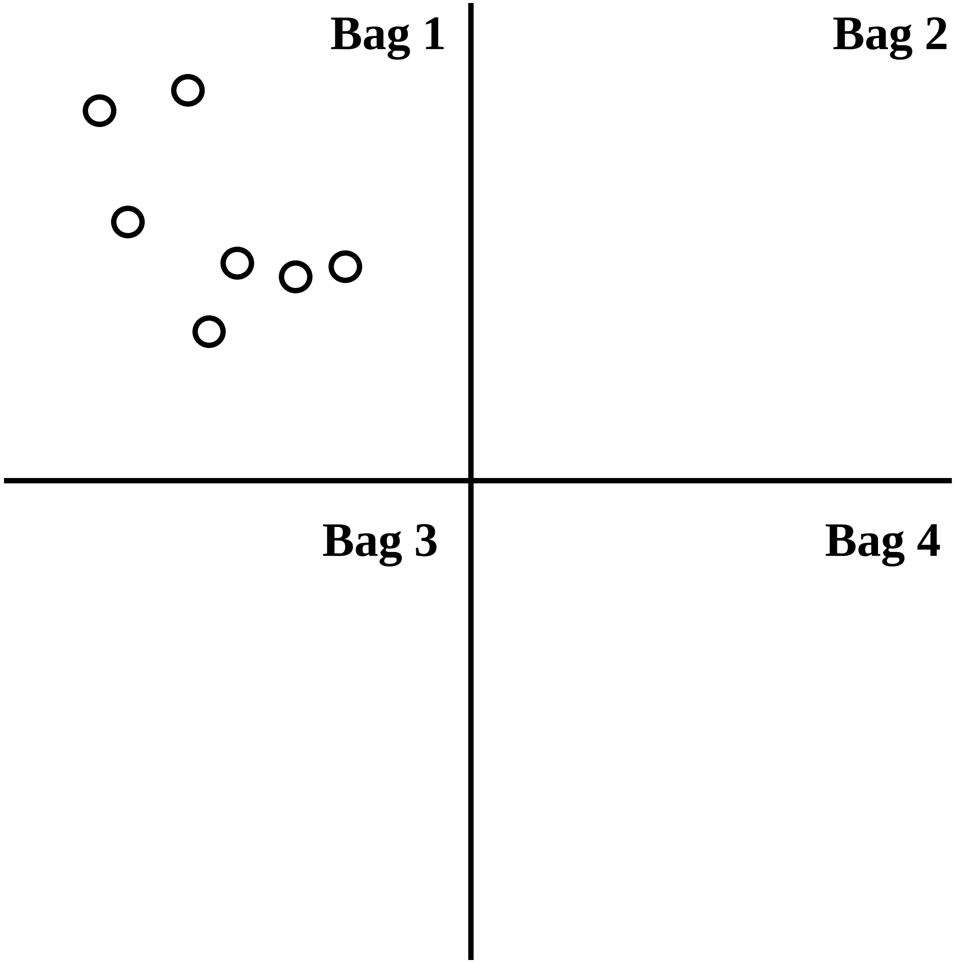
Learning with label proportions



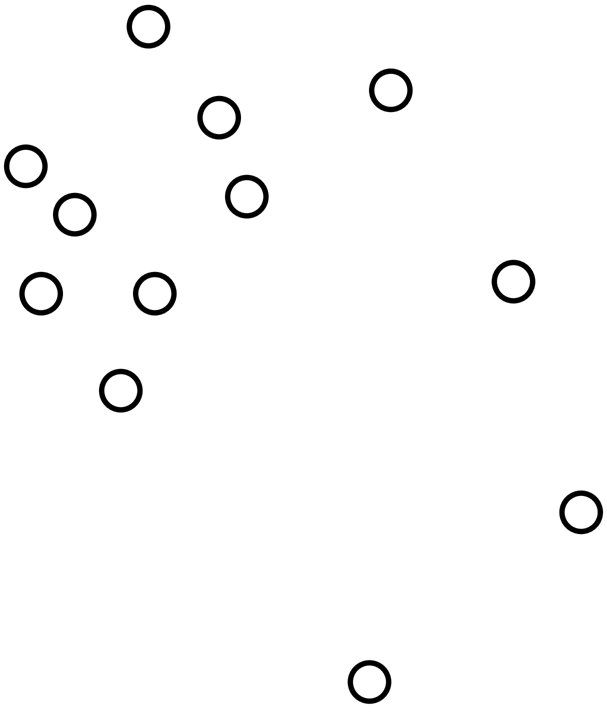


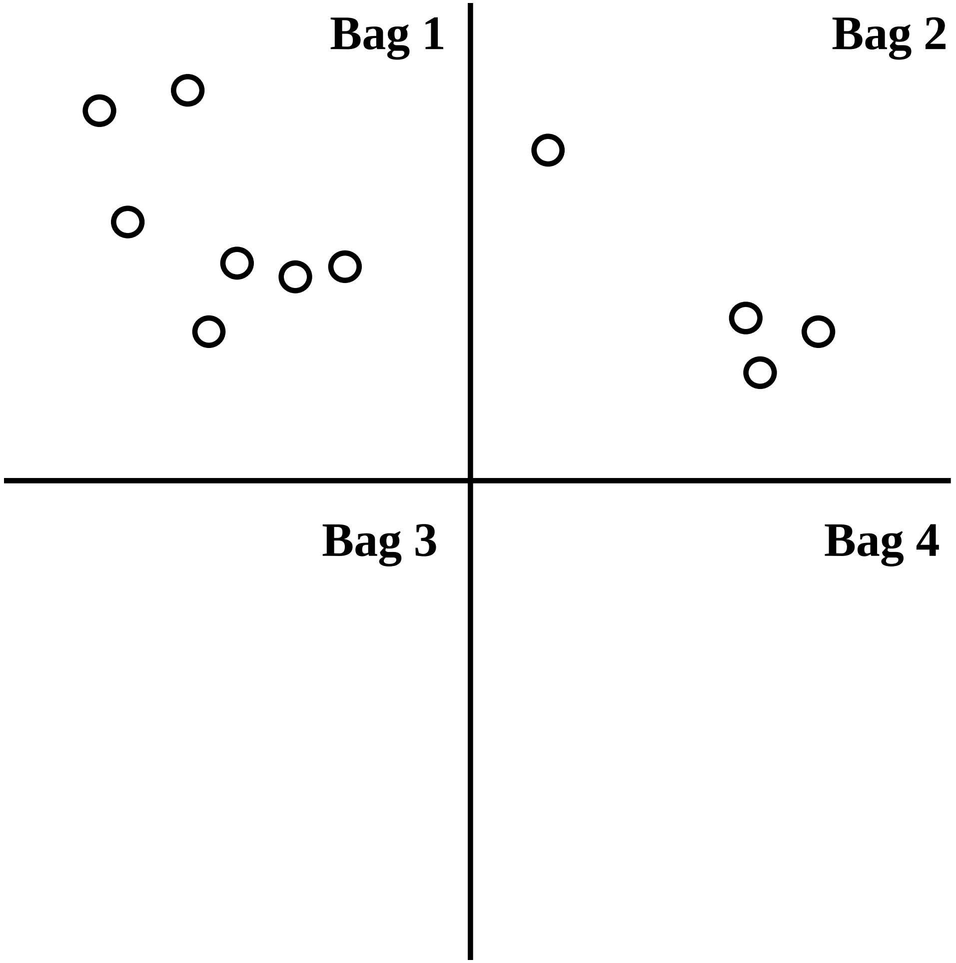
Learning with label proportions



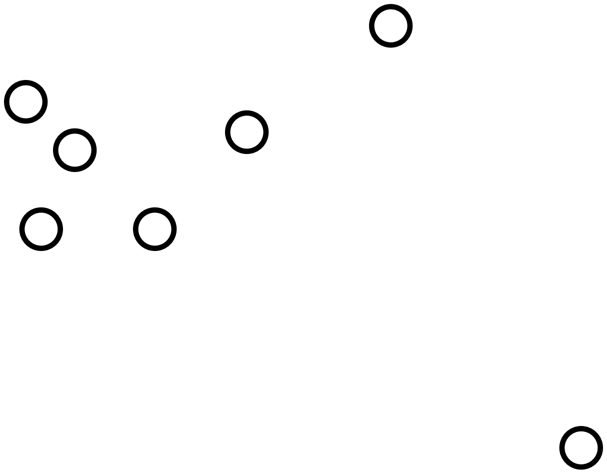


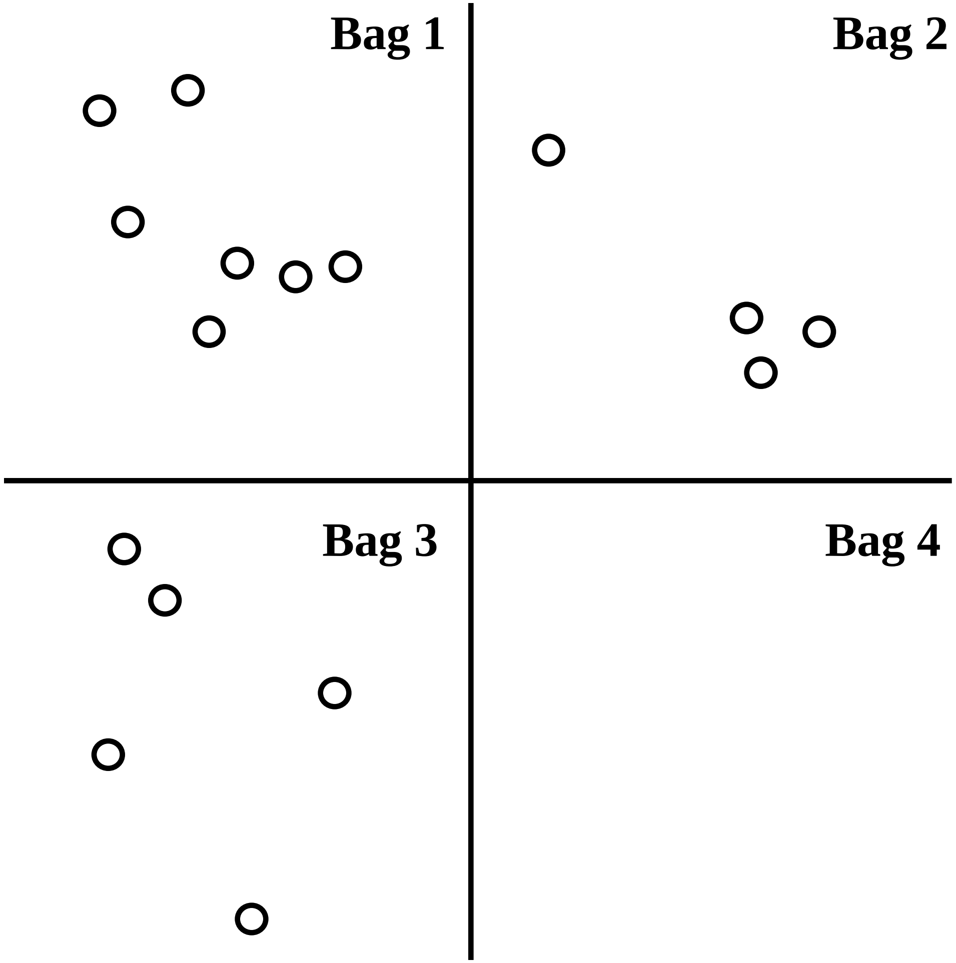
Learning with label proportions



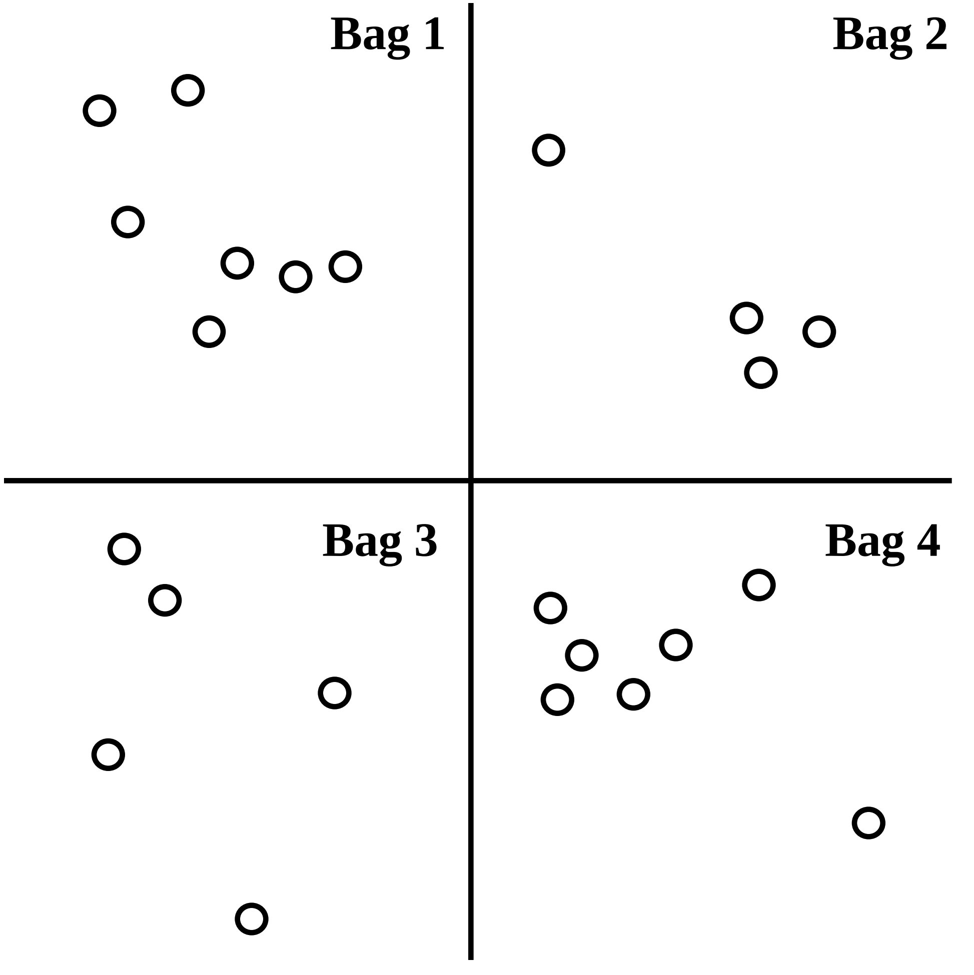


Learning with label proportions

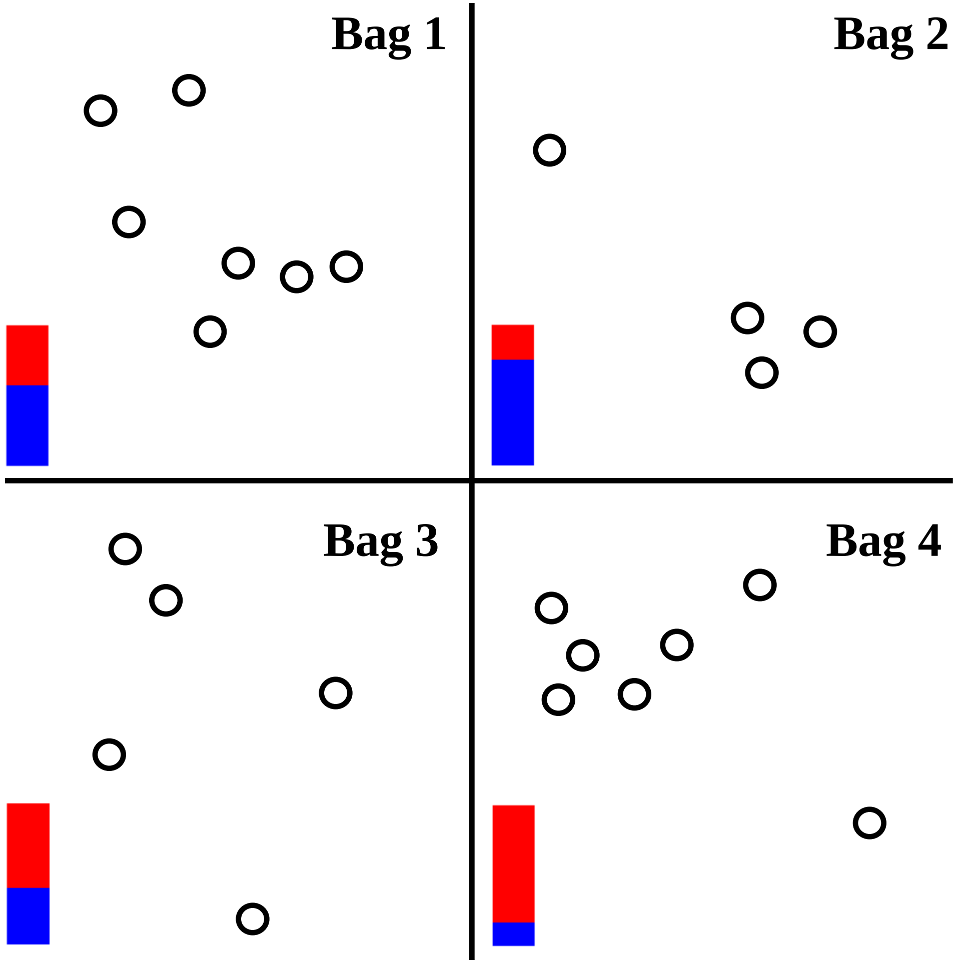




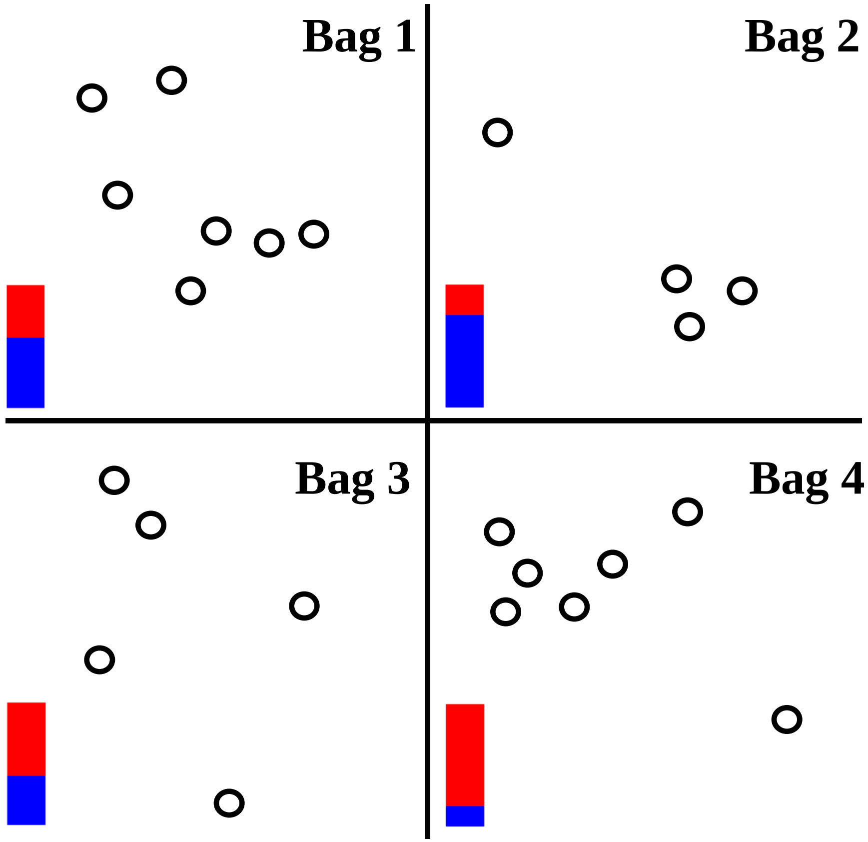
Learning with label proportions



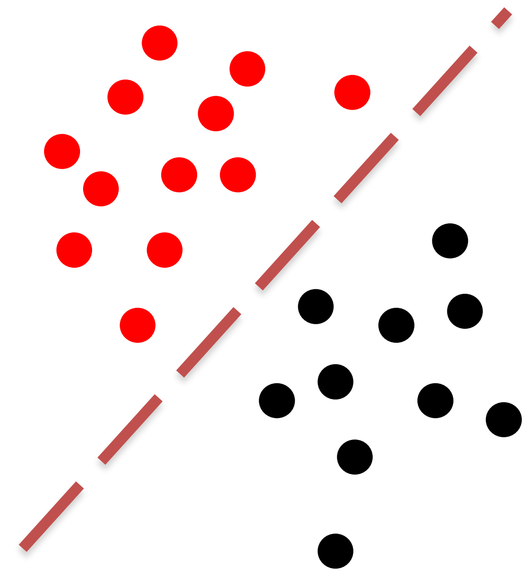
Learning with label proportions



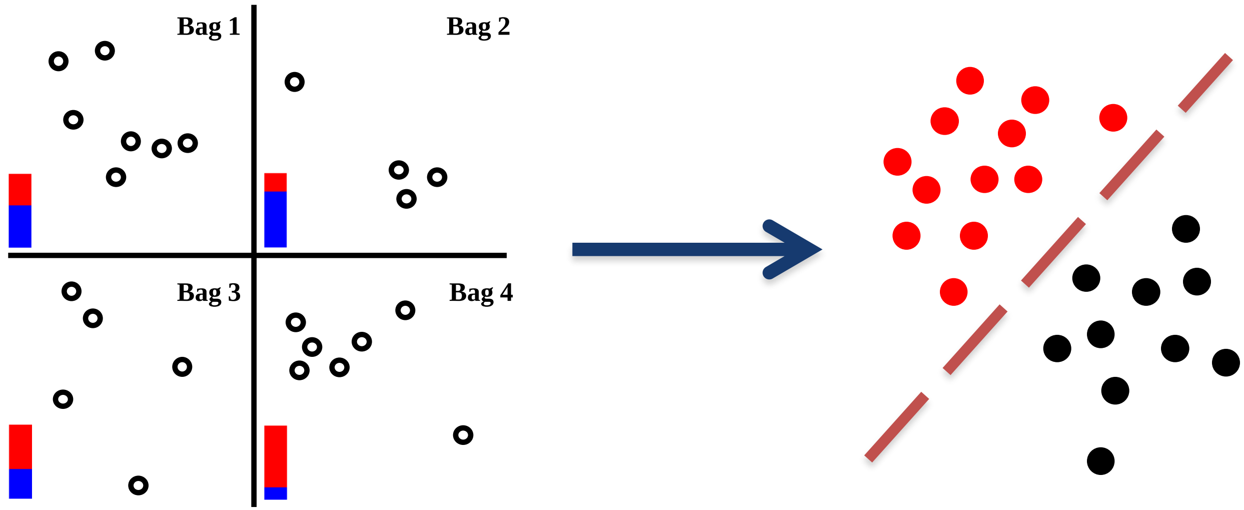
Learning with label proportions



Learning with Label Proportions

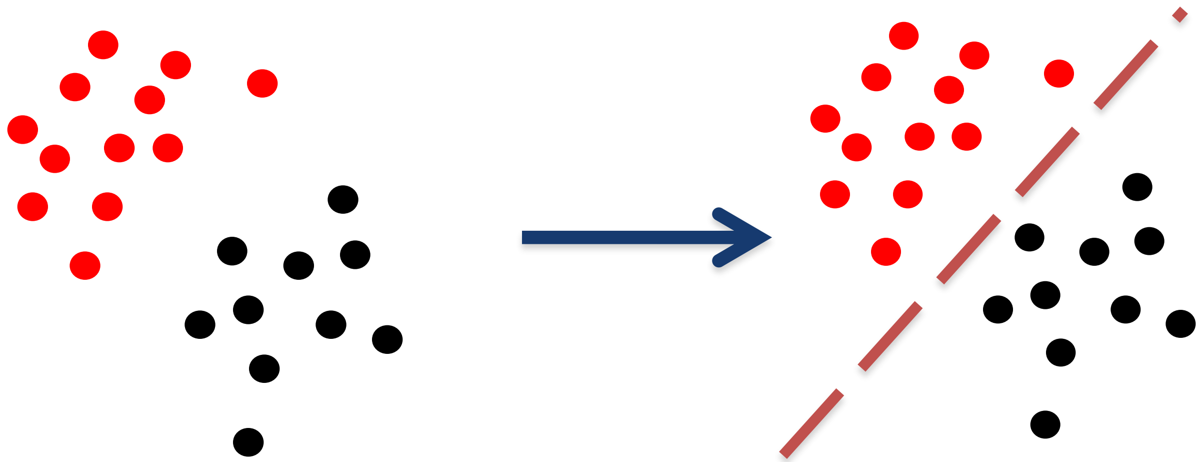


Learning with label proportions



Learning with Label Proportions

Supervised Learning



Applications

- Healthcare
  - Medical record
- Social science
  - Voting behavior
  - Census data
  - Energy consumption
  - Marketing
- Computer vision
  - Facial attributes (“90 percent of Asians have black hair”)

Applications

Privacy issues

- Healthcare
  - Medical record
- Social science
  - Voting behavior
  - Census data
  - Energy consumption
  - Marketing
- Computer vision
  - Facial attributes ("90 percent of Asians have black hair")

Applications

- Healthcare
  - Medical record
- Social science
  - Voting behavior
  - Census data
  - Energy consumption
  - Marketing
- Computer vision
  - Facial attributes ("90 percent of Asians have black hair")

Easier to get label proportions

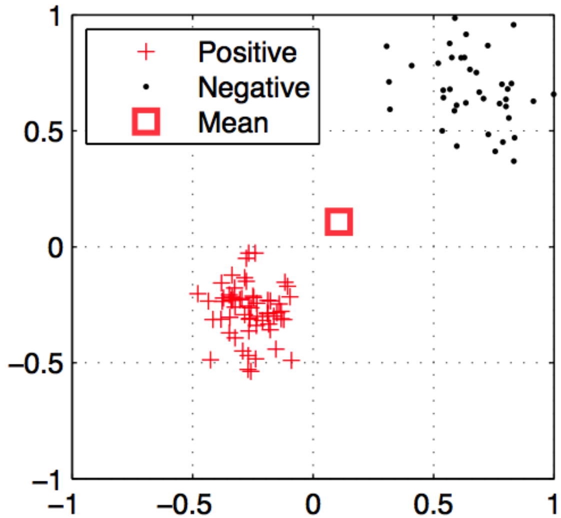
Outline

- Learning Setting and Applications
- Related Works
- Formulation
- Algorithms
- Experiments

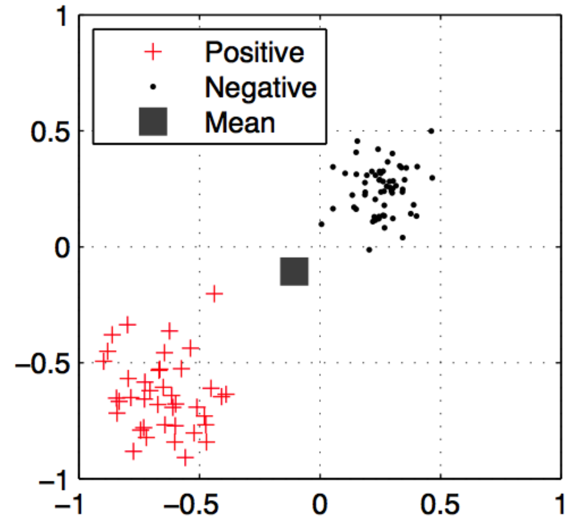
Related Works

- Related learning settings: semi-supervised learning, clustering, multi-instance learning etc.

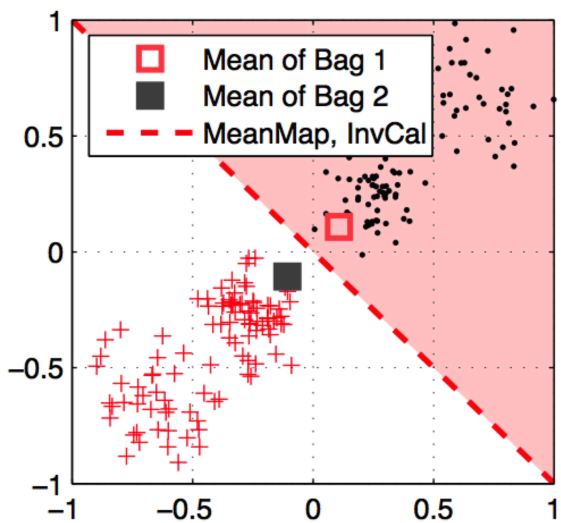
- Learning with label proportions: former works rely heavily on the “mean of each bag”
  - MeanMap (Quadranto et al., 2009)
    - exponential model
    - class-conditional distribution of data is independent of the bag
  - Inverse Calibration (Rueping, 2011)
    - large-margin regression
    - mean of each bag has a soft label corresponding to its label proportion



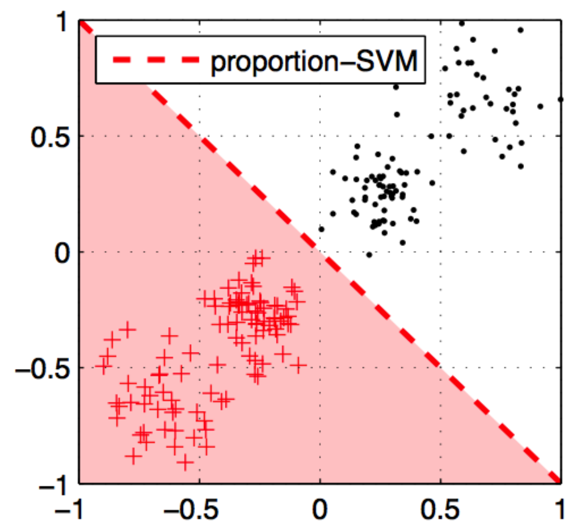
(a) Bag 1, with $p_1 = 0.6$



(b) Bag 2, with $p_2 = 0.4$



(c) MeanMap and InvCal with 0% accuracy.



(d) ∝SVM with 100% accuracy.

Contributions

- Introduce $\propto$SVM which explicitly models the unknown instance labels.
- Alleviates the need for making restrictive assumptions on the data.
- Two optimization algorithms based on alternating minimization and convex relaxation.
- Outperforms existing methods under various settings/datasets.

Outline

- Learning Setting and Applications
- Related Works
- Formulation
- Algorithms
- Experiments

Formulation (Learning Setting)

- The training set $\{\mathbf{x}_i\}_{i=1}^N$ is given in the form of K non-overlapping bags:

$$\{\mathbf{x}_i | i \in \mathcal{B}_k\}_{k=1}^K, \quad \cup_{k=1}^K \mathcal{B}_k = \{1 \cdots N\}.$$

- The k-th bag is with label proportion p_k:

$$\forall_{k=1}^K, \quad p_k := \frac{|\{i | i \in \mathcal{B}_k, y_i^* = 1\}|}{|\mathcal{B}_k|}.$$

$y_i^* \in \{1, -1\}$: the *unknown* ground-truth label of $\mathbf{x}_i$, $\forall_{i=1}^N$

Formulation

- Prediction model:

$$f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T \varphi(\mathbf{x}) + b).$$

- Explicitly model the unknown instance labels as

$$\mathbf{y} = (y_1, \cdots, y_N)^T. \; y_i \in \{1, -1\}, \forall_{i=1}^N.$$

- The label proportion of the k-th bag can be modeled as

$$\tilde{p}_k(\mathbf{y}) = \frac{|\{i|i \in \mathcal{B}_k, y_i = 1\}|}{|\mathcal{B}_k|} = \frac{\sum_{i \in \mathcal{B}_k} y_i}{2|\mathcal{B}_k|} + \frac{1}{2}.$$

∝SVM Formulation

- Large-margin framework:

$$\min_{\mathbf{y},\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} L(y_i, \mathbf{w}^T\varphi(\mathbf{x}_i) + b) + C_p\sum_{k=1}^{K} L_p\left(\tilde{p}_k(\mathbf{y}), p_k\right)$$
$$\text{s.t.} \quad \forall_{i=1}^{N}, \quad y_i \in \{-1, 1\}.$$

∝SVM Formulation

- Large-margin framework:

$$\min_{\mathbf{y},\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} L(y_i, \mathbf{w}^T\varphi(\mathbf{x}_i) + b) + C_p \sum_{k=1}^{K} L_p\left(\tilde{p}_k(\mathbf{y}), p_k\right)$$
$$\text{s.t.} \quad \forall_{i=1}^{N}, \quad y_i \in \{-1, 1\}.$$

- Generalizes the classic SVM.
- Naturally spans supervised/semi-supervised learning and clustering.

∝SVM Formulation

- Large-margin framework:

$$\min_{\mathbf{y},\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} L(y_i, \mathbf{w}^T\varphi(\mathbf{x}_i) + b) + C_p\sum_{k=1}^{K} L_p\left(\tilde{p}_k(\mathbf{y}), p_k\right)$$
$$\text{s.t.} \quad \forall_{i=1}^{N}, \quad y_i \in \{-1, 1\}.$$

- Generalizes the classic SVM.
- Naturally spans supervised/semi-supervised learning and clustering.

However:

- A non-convex integer programming problem.

Outline

- Learning Setting and Applications
- Related Works
- Formulation
- Algorithms
- Experiments

The alter-∝SVM Algorithm

- For a fixed $\mathbf{y}$, the optimization $w.r.t$ $\mathbf{w}$ and b is an SVM problem.

- When $\mathbf{w}$ and b are fixed:

$$\begin{aligned} \min_{\mathbf{y}} \quad & \sum_{i=1}^{N} L(y_i, \mathbf{w}^T \varphi(\mathbf{x}_i) + b) + \frac{C_p}{C} \sum_{k=1}^{K} L_p\left(\tilde{p}_k(\mathbf{y}), p_k\right) \\ \text{s.t.} \quad & \forall_{i=1}^{N}, \quad y_i \in \{1, -1\}. \end{aligned}$$

The alter-∝SVM Algorithm

$$\min_{\mathbf{y}} \quad \sum_{i=1}^{N} L(y_i, \mathbf{w}^T \varphi(\mathbf{x}_i) + b) + \frac{C_p}{C} \sum_{k=1}^{K} L_p\left(\tilde{p}_k(\mathbf{y}), p_k\right)$$
$$\text{s.t.} \quad \forall_{i=1}^{N}, \quad y_i \in \{1, -1\}.$$

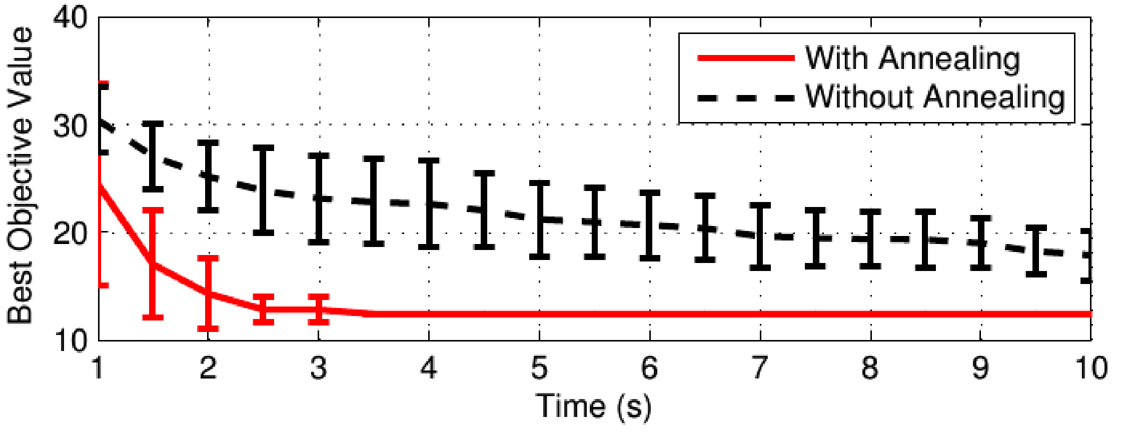
- Consider each bag separately.
- For the k-th bag: sorting, $\mathcal{O}(|\mathcal{B}_k| \log |\mathcal{B}_k|)$ time).

Proposition:
The above can be solved in $\mathcal{O}(N \log(J))$ time, $J = \max_{k=1\cdots K} |\mathcal{B}_k|$.

The alter-∝SVM Algorithm

To alleviate the problem of local solutions:

- Multiple initilizations.
- An additional annealing loop to gradually increase C.



The conv-$\propto$SVM Algorithm

- Does not require multiple initializations.
- Motivated by large-margin clustering (Xu et al., 2004) (Li et al., 2009).

The conv-∝SVM Algorithm

- Reformulation:

$$\min_{\mathbf{y}\in\mathcal{Y},\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} L(y_i, \mathbf{w}^T\varphi(\mathbf{x}_i))$$

$$\mathcal{Y} = \left\{\mathbf{y} \middle| |\tilde{p}_k(\mathbf{y}) - p_k| \leq \epsilon, y_i \in \{-1, 1\}, \forall_{k=1}^{K}\right\}$$

- Write the inner problem as its dual (with hinge loss):

$$\min_{\mathbf{y}\in\mathcal{Y}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}^T\left(\boldsymbol{\mathcal{K}} \odot \mathbf{y}\mathbf{y}^T\right)\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\mathbf{1}$$

$$\boldsymbol{\alpha} \in \mathbb{R}^N$$
$$\mathcal{A} = \{\boldsymbol{\alpha} | 0 \leq \boldsymbol{\alpha} \leq C\}$$

The conv-∝SVM Algorithm

$$\min_{\mathbf{y}\in\mathcal{Y}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}^T\left(\boldsymbol{\mathcal{K}}\odot\mathbf{y}\mathbf{y}^T\right)\boldsymbol{\alpha}+\boldsymbol{\alpha}^T\mathbf{1}$$

$$\boldsymbol{\alpha}\in\mathbb{R}^N$$
$$\mathcal{A}=\{\boldsymbol{\alpha}|0\leq\boldsymbol{\alpha}\leq C\}$$

- Convex in $\boldsymbol{M} := \mathbf{y}\mathbf{y}^T$.
- Relax the feasible space of $\boldsymbol{M}$ to get a convex problem.

$$\mathcal{M}_0=\{\mathbf{y}\mathbf{y}^T|\mathbf{y}\in\mathcal{Y}\}$$

↓ Relaxation

$$\mathcal{M}=\left\{\sum_{\mathbf{y}\in\mathcal{Y}}\mu_{(\mathbf{y})}\mathbf{y}\mathbf{y}^T\middle|\boldsymbol{\mu}\in\mathcal{U}\right\},$$
$$\mathcal{U}=\{\boldsymbol{\mu}|\sum_{\mathbf{y}\in\mathcal{Y}}\mu_{(\mathbf{y})}=1,\mu_{(\mathbf{y})}\geq 0\}$$

The conv-∝SVM Algorithm

- Solving the relaxed $\boldsymbol{M}$ is identical to finding $\boldsymbol{\mu}$:

$$\min_{\boldsymbol{\mu}\in\mathcal{U}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}^T\left(\sum_{\mathbf{y}\in\mathcal{Y}} \mu_{(\mathbf{y})}\boldsymbol{\mathcal{K}} \odot \mathbf{y}\mathbf{y}^T\right)\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\mathbf{1}.$$

- Multiple Kernel Learning (MKL).
- $|\mathcal{Y}|$ is very large. Not tractable to solve directly.
- Primal variables $\rightarrow$ dual constraints.
- Cutting plane method (Li et al., 2009) (Joachims et al., 2009).

Outline

- Learning Setting and Applications
- Related Works
- Formulation
- Algorithms
- Experiments

Experiments

- Performance of different techniques on 12 datasets from the UCI/LibSVM repository.

Dataset	Size	Attributes	Classes
heart	270	13	2
heart-c	303	13	2
colic	366	22	2
vote	435	16	2
breast-cancer	683	10	2
australian	690	14	2
credit-a	690	15	2
breast-w	699	9	2
a1a	1,605	119	2
dna	2,000	180	3
satimage	4,435	36	6
cod-rna.t	271,617	8	2

- Follow the experimental setting of (Rueping, 2011):
  - Random bag generation (with different bag sizes). Performance of 5-fold cross validation.
  - Linear and RBF kernels.

Experiments

Dataset	Method	2	4	8	16	32	64
heart	MeanMap	82.69±0.71	80.80±0.97	79.65±0.82	79.44±1.21	80.03±2.05	77.26±0.85
	InvCal	**83.15±0.56**	81.06±0.70	80.26±1.32	79.61±3.84	76.36±3.72	73.90±3.00
	alter-∝SVM	83.15±0.85	**82.89±1.30**	**81.51±0.54**	80.07±1.21	79.10±0.96	78.63±1.85
	conv-∝SVM	82.96±0.26	82.20±0.52	81.38±0.53	**81.17±0.55**	**80.94±0.86**	**78.87±1.37**
colic	MeanMap	82.45±0.88	81.38±1.26	81.71±1.16	79.94±1.33	76.36±2.43	77.84±1.69
	InvCal	82.20±0.61	81.20±0.87	81.17±1.74	78.59±2.19	74.09±5.26	72.81±4.80
	alter-∝SVM	**83.28±0.50**	**82.97±0.39**	**82.03±0.44**	**81.62±0.46**	**81.53±0.21**	**81.39±0.34**
	conv-∝SVM	82.74±1.15	81.83±0.46	79.58±0.57	79.77±0.84	78.22±1.19	77.31±1.76
vote	MeanMap	91.15±0.33	90.52±0.62	91.54±0.20	90.28±1.63	89.58±1.09	89.38±1.33
	InvCal	95.68±0.19	94.77±0.44	93.95±0.43	**93.03±0.37**	87.79±1.64	86.63±4.74
	alter-∝SVM	**95.80±0.20**	**95.54±0.25**	**94.88±0.94**	92.44±0.60	**90.72±1.11**	**90.93±1.30**
	conv-∝SVM	92.99±0.20	92.01±0.69	90.57±0.68	88.98±0.35	88.74±0.43	88.62±0.60
australian	MeanMap	85.97±0.72	85.88±0.34	85.34±1.01	83.36±2.04	83.12±1.52	80.58±5.41
	InvCal	**86.06±0.30**	86.11±0.26	**86.32±0.45**	84.13±1.62	82.73±1.70	81.87±3.29
	alter-∝SVM	85.74±0.22	85.71±0.21	86.26±0.61	**85.65±0.43**	83.63±1.83	**83.62±2.21**
	conv-∝SVM	85.97±0.53	**86.46±0.23**	85.30±0.70	84.18±0.53	**83.69±0.78**	82.98±1.32
dna-1	MeanMap	91.53±0.25	90.58±0.34	86.00±1.04	80.77±3.69	77.35±3.59	68.47±4.30
	InvCal	89.32±3.39	92.73±0.53	87.99±1.65	81.05±3.14	74.77±2.95	67.75±3.86
	alter-∝SVM	**95.67±0.40**	**94.65±0.52**	**93.71±0.47**	**92.52±0.63**	**91.85±1.42**	**90.64±1.32**
	conv-∝SVM	93.36±0.53	86.75±2.56	81.03±3.58	75.90±4.56	76.92±5.91	77.94±2.48
dna-2	MeanMap	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	InvCal	89.65±4.05	93.12±1.37	89.19±1.17	83.52±2.57	77.94±2.82	72.64±3.89
	alter-∝SVM	**95.63±0.45**	**95.05±0.75**	**94.25±0.50**	**93.95±0.93**	**92.74±0.93**	**92.46±0.90**
	conv-∝SVM	94.06±0.86	90.68±1.18	87.64±0.76	87.32±1.55	85.74±1.03	85.33±0.79
satimage-2	MeanMap	97.08±0.48	96.82±0.38	96.50±0.43	96.45±1.16	95.51±0.73	94.26±0.22
	InvCal	97.53±1.33	98.33±0.13	98.38±0.23	[illegible]	[illegible]	[illegible]
	alter-∝SVM	**98.83±0.36**	**98.69±0.37**	**98.62±0.27**	[illegible]	[illegible]	[illegible]
	conv-∝SVM	96.55±0.13	96.45±0.19	96.45±0.39	[illegible]	[illegible]	[illegible]

Horizontal: bag size

Dataset	Method	2	4	8	16	32	64
australian	MeanMap	85.97±0.72	85.88±0.34	85.34±1.01	83.36±2.04	83.12±1.52	80.58±5.41
	InvCal	**86.06±0.30**	86.11±0.26	**86.32±0.45**	84.13±1.62	82.73±1.70	81.87±3.29
	alter-∝SVM	85.74±0.22	85.71±0.21	86.26±0.61	**85.65±0.43**	83.63±1.83	**83.62±2.21**
	conv-∝SVM	85.97±0.53	**86.46±0.23**	85.30±0.70	84.18±0.53	**83.69±0.78**	82.98±1.32
dna-1	MeanMap	91.53±0.25	90.58±0.34	86.00±1.04	80.77±3.69	77.35±3.59	68.47±4.30
	InvCal	89.32±3.39	92.73±0.53	87.99±1.65	81.05±3.14	74.77±2.95	67.75±3.86
	alter-∝SVM	**95.67±0.40**	**94.65±0.52**	**93.71±0.47**	**92.52±0.63**	**91.85±1.42**	**90.64±1.32**
	conv-∝SVM	93.36±0.53	86.75±2.56	81.03±3.58	75.90±4.56	76.92±5.91	77.94±2.48

- Our methods outperform MeanMap and InvCal.
- The gains from ∝SVM are typically even more significant when the bag size is large.

  -- *on the dna-1 dataset, with RBF kernel and bag size 64, alter-∝SVM outperforms the former works by 22%.*

- Also shown in the paper: less sensentive to bag proportion variations.

Conclusion

- $\propto$SVM
- Two optimization algorithms
- State-of-the-art result

Future works/ Open issues

- Robustness to proportion noise.
- Bag generation, bags with overlappings.
- $\propto$SVM for semi-supervised learning, and learning with label errors.