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

南京大学
NANJING UNIVERSITY 1902

The 20th ACM SIGKDD International Conference
on Knowledge Discovery and Data Mining

Large Margin Distribution Machine

Teng Zhang and Zhi-Hua Zhou

LAMDA Group

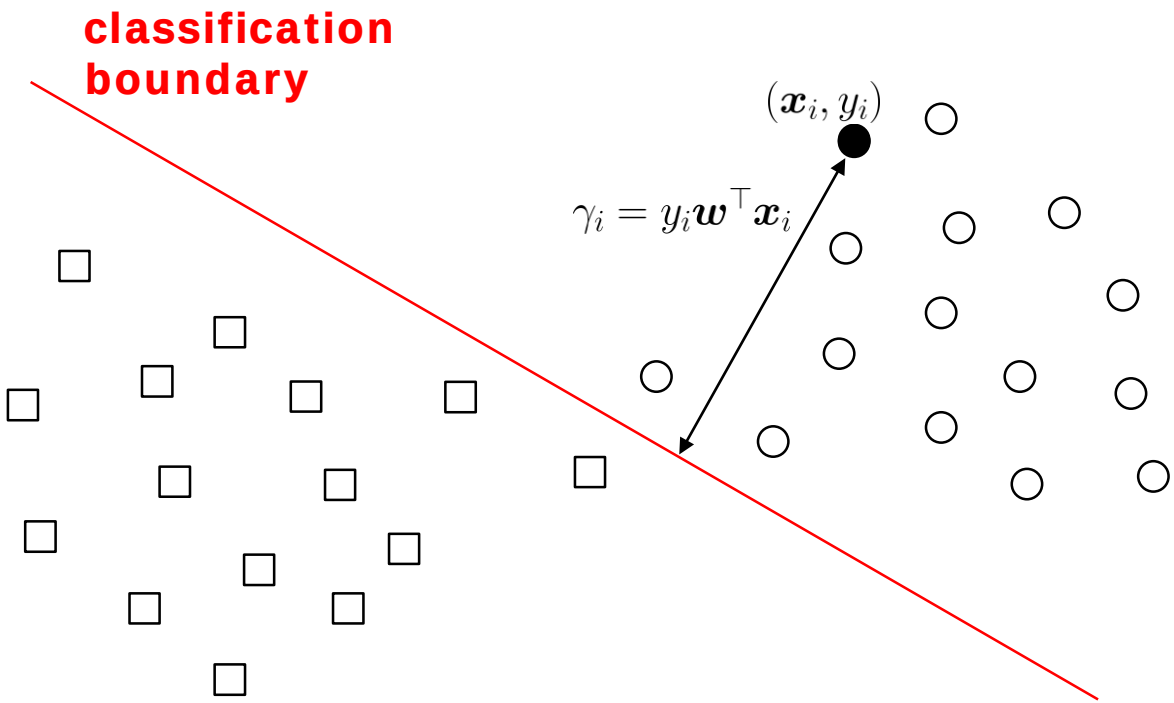
National Key Laboratory for Novel Software Technology

Nanjing University, China

{zhangt, zhouzh}@lamda.nju.edu.cn

Margin

Given m training examples $\{(\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\})\}_{i=1}^m$, the margin of instance $(\boldsymbol{x}_i, y_i)$ is formulated as $\gamma_i = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$.



Margin Theory

Boosting

[Schapire et al., ASJ'98] — Adaboost doesn't overfit since it tends to enlarge the margin

[Breiman, NCJ'98] — Arc-gv optimize the minimum margin with a poor performance

[Reyzin&Schapire, ICML'06] — margin distribution rather than minimum margin is more crucial

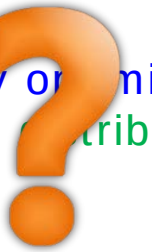
[Gao&Zhou, AIJ'13] — both the margin mean and margin variance are important

SVM

[Cortes&Vapnik, ML'95] — SVMs optimize the minimum margin

[Garg&Roth, ICML 2003
Pelckmans et al, NIPS 2007
Aiolli et al, ICANN 2008] — optimize the weighted combination of margins

directly optimize the margin distribution



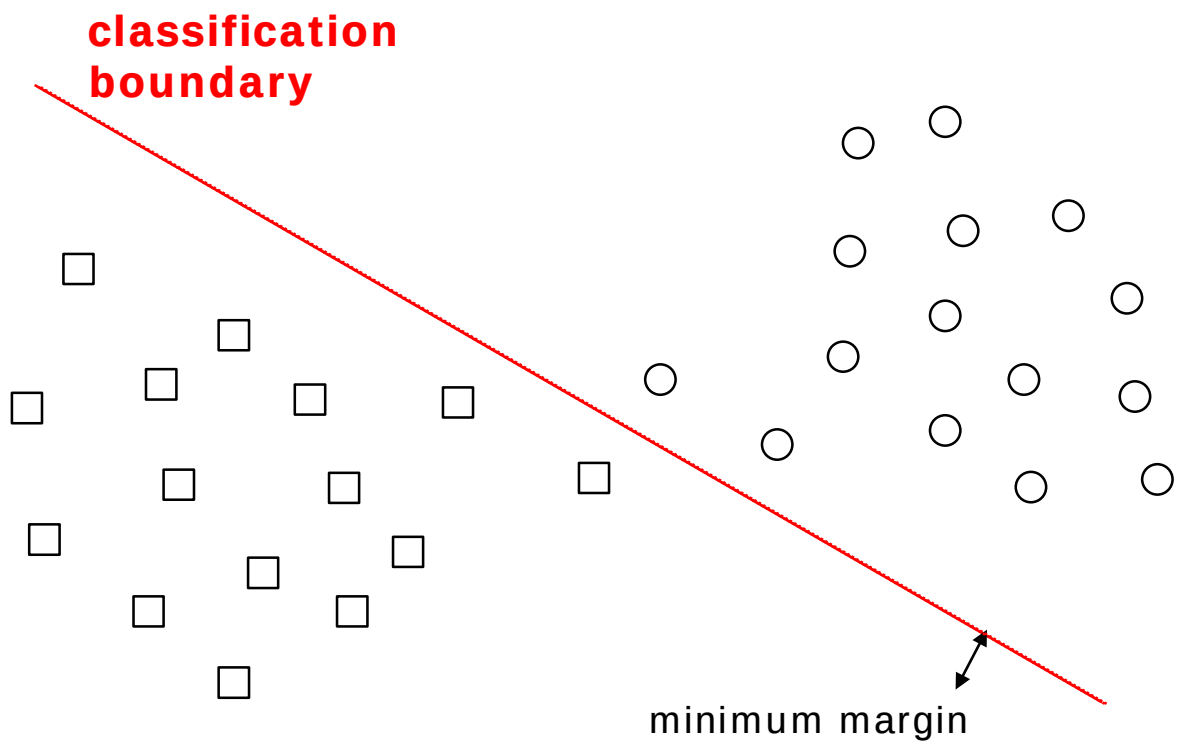
Outline

- SVM and its extensions
- Our Method
- Experiments
- Conclusion

SVM

Try to maximize the minimum margin

$$\max \{\min\{\gamma_i\}_{i=1}^m\} \qquad \gamma_i = y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i), \forall i = 1, \ldots, m.$$



SVM ignores the whole margin distribution!

SVMs beyond minimum margin maximization

- MDO [Garg&Roth, ICML 2003]

  Optimize the weighted margin combination, however, the setting of weights is heuristic, and the objective function is non-convex.

All the methods consider the margin mean but ignore the influence of margin variance. However, margin mean can not characterize the margin distribution well!

size.

- KM-OMD [Aiolli et al, ICANN 2008]

  Optimize the weighted margin combination. It considers hard-margin only.

Outline

- SVM and its extensions
- Our Method
- Experiments
- Conclusion

Basic Idea

Inspired by [Gao&Zhou, AIJ'13], the mean and the variance of the margin are important.

Margin mean:

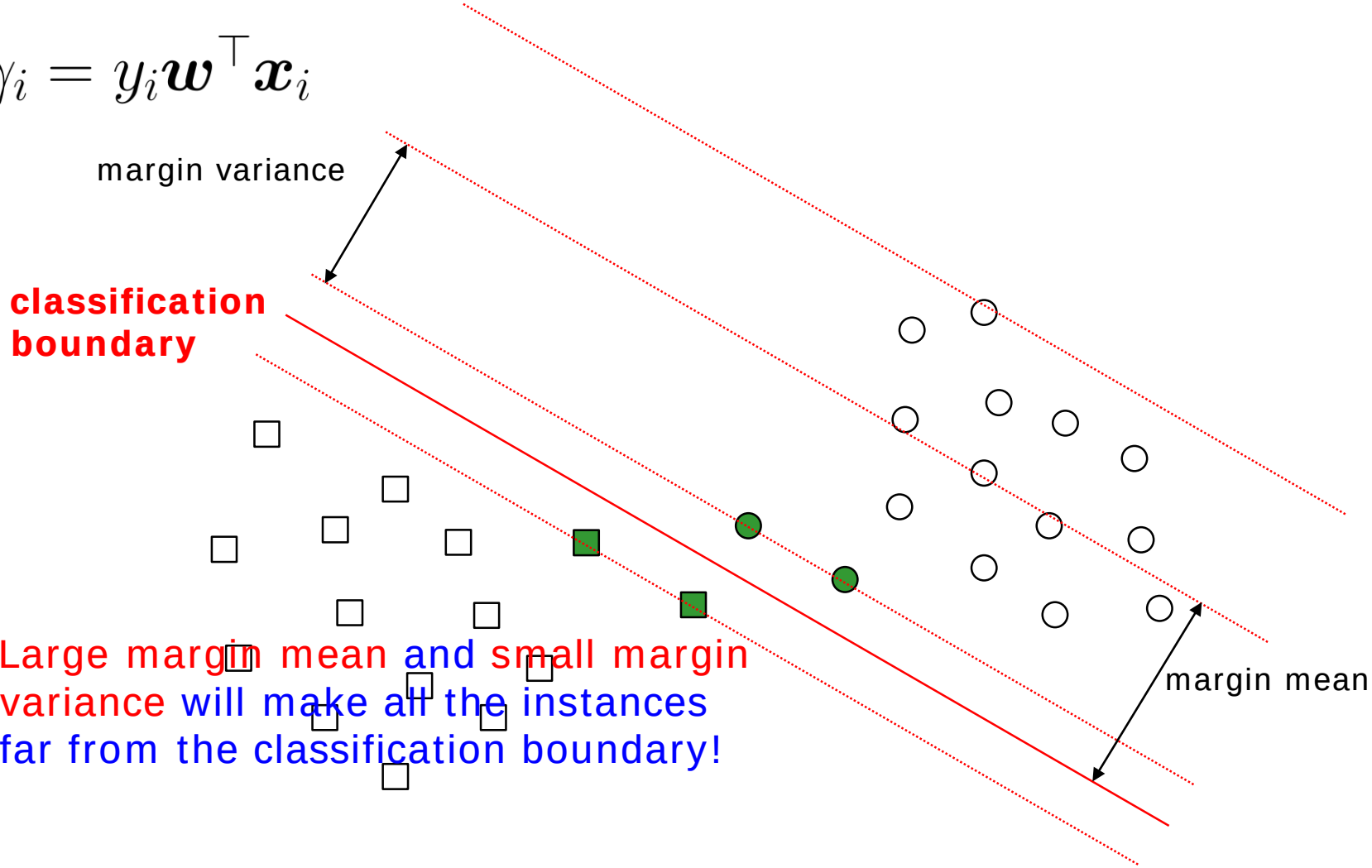
$$\bar{\gamma} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i)$$

Margin variance:

$$\hat{\gamma} = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}(y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) - y_j \boldsymbol{w}^\top \phi(\boldsymbol{x}_j))^2$$

Intuitive Interpretation

$$\gamma_i = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$$



Large margin mean and small margin variance will make all the instances far from the classification boundary!

The LDM Method

LDM: Large margin Distribution Machine

Idea: maximize the margin mean and minimize the margin variance simultaneously.

Formulation:

margin variance

$$\min_{\boldsymbol{w},\boldsymbol{\xi}} \quad \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + \frac{2\lambda_1}{m^2}(m\boldsymbol{w}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{w} - \boldsymbol{w}^\top\boldsymbol{X}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{X}^\top\boldsymbol{w})$$

margin mean

$$-\frac{\lambda_2}{m}(\boldsymbol{X}\boldsymbol{y})^\top\boldsymbol{w} + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \ i = 1, \ldots, m.$$

The LDM Method (cont.)

With representer theorem, the dual of LDM is

$$\min_{\boldsymbol{\beta}} \quad f(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{H}\boldsymbol{\beta} + \left(\frac{\lambda_2}{m}\boldsymbol{H}\boldsymbol{e} - \boldsymbol{e}\right)^\top \boldsymbol{\beta},$$
$$\text{s.t.} \quad 0 \leq \beta_i \leq C, \; i = 1, \ldots, m.$$

- convex quadratic programming
- decoupled box constraint

It can be solved by the coordinate descent method efficiently since a closed-form solution can be achieved in each iteration.

$$\beta_i^{new} = \min\left(\max\left(\beta_i - \frac{[\nabla f(\boldsymbol{\beta})]_i}{h_{ii}}, 0\right), C\right)$$

Large Scale Kernel LDM

For large scale problems, we can solve the prime LDM directly by average stochastic gradient descent (ASGD).

Key idea of ASGD: utilize unbiased estimation of the gradient

THEOREM 2. *If two examples* $(\boldsymbol{x}_i, y_i)$ *and* $(\boldsymbol{x}_j, y_j)$ *are sampled from training set randomly, then*

$$\nabla g(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{x}_j) = 4\lambda_1 \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{w} - 4\lambda_1 y_i \boldsymbol{x}_i y_j \boldsymbol{x}_j^\top \boldsymbol{w} + \boldsymbol{w} - \lambda_2 y_i \boldsymbol{x}_i - mC \begin{cases} y_i \boldsymbol{x}_i & i \in \boldsymbol{I}, \\ 0 & otherwise, \end{cases} \tag{15}$$

where $\boldsymbol{I} \equiv \{i \mid y_i \boldsymbol{w}^\top \boldsymbol{x}_i < 1\}$ *is an unbiased estimation of* $\nabla g(\boldsymbol{w})$.

the unbiased estimation of the gradient of the objective function can be obtained by sampling two examples randomly in each iteration.

Theoretical Analysis

Based on leave-one-out cross-validation estimate, we can derive a bound on the expectation of error for LDM.

THEOREM 3. *Let* $\boldsymbol{\alpha}$ *denote the optimal solution of (18), and* $E[R(\boldsymbol{\alpha})]$ *be the expectation of the probability of error, then we have*

$$E[R(\boldsymbol{\alpha})] \leq \frac{E[h\sum_{i\in I_1}\alpha_i + |I_2|]}{m}, \tag{19}$$

where $I_1 \equiv \{i \mid 0 < \alpha_i < C\}$, $I_2 \equiv \{i \mid \alpha_i = C\}$, $h = \max\{diag\{\boldsymbol{H}\}\}$ *and* $\boldsymbol{H} = \boldsymbol{Y}\boldsymbol{X}^\top\boldsymbol{Q}^{-1}\boldsymbol{X}\boldsymbol{Y}$.

Observation

- A similar bound also holds for SVM [Vapnik, 1995], and the only difference is that $\boldsymbol{H} = \boldsymbol{Y}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{Y}$.
- $\boldsymbol{Q}$ encodes the information of the margin distribution with the result that the value of h for LDM is much smaller than SVM.

Outline

- Introduction
- Our Method
- Experiments
- Conclusion

Data

20 regular scale data sets and 12 large scale data sets

Scale	Dataset	#Instance	#Feature	Dataset	#Instance	#Feature
regular	*promoters*	106	57	*haberman*	306	14
	planning	182	12	*vehicle*	435	16
	colic	188	13	*clean1*	476	166
large	*farm-ads*	4,143	54,877	*ijcnn1*	141,691	22
	news20	19,996	1,355,191	*skin*	245,057	3
	adult-a	32,561	123	*covtype*	581,012	54
	w8a	49,749	300	*rcv1*	697,641	47,236
	cod-rna	59,535	8	*url*	2,396,130	3,231,961
	real-sim	72,309	20,958	*kdd2010*	8,407,752	20,216,830

The size of data set ranges from 106 to more than 8,000,000, and the dimensionality ranges from 2 to more than 20,000,000, covering a broad range of properties.

Settings

Compared methods:

LDM, MDO, MAMC, KM-OMD, SVM

Parameters:

selected by 5-fold cross validation

Evaluation:

30 times hold-out tests, 1/2 training, 1/2 testing

Results

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

Accuracy on twenty regular scale data sets with linear kernel

Dataset	SVM	MDO	MAMC	KM-OMD	LDM
promoters	0.723±0.071	0.713±0.067	0.520±0.096○	**0.736±0.061**	0.721±0.069
planning-relax	0.683±0.031	0.605±0.185○	**0.706±0.034●**	0.479±0.050○	**0.706±0.034●**
colic	0.814±0.035	0.781±0.154	0.661±0.062○	0.813±0.028	**0.832±0.026●**
parkinsons	0.846±0.038	0.732±0.270○	0.764±0.035○	0.814±0.024○	**0.865±0.030●**
colic.ORIG	0.618±0.027	0.624±0.040	0.623±0.027	**0.635±0.045●**	0.619±0.042
sonar	0.725±0.039	0.734±0.035	0.533±0.045○	**0.766±0.033●**	0.736±0.036
vote	0.934±0.022	0.587±0.435○	0.884±0.022○	0.957±0.013●	**0.970±0.014●**
house	0.942±0.015	0.943±0.015	0.883±0.029○	0.957±0.020●	**0.968±0.011●**
heart	0.799±0.029	0.826±0.026●	0.537±0.057○	**0.836±0.026●**	0.791±0.030
breast-cancer	0.717±0.033	0.710±0.031	0.706±0.027	0.696±0.031○	**0.725±0.027●**
haberman	0.734±0.030	0.728±0.029	**0.738±0.020**	0.667±0.040○	**0.738±0.020**
vehicle	0.959±0.012	0.956±0.012	0.566±0.160○	**0.960±0.010**	0.959±0.013
clean1	0.803±0.035	0.798±0.031	0.561±0.025○	**0.821±0.027●**	0.814±0.019●
wdbc	0.963±0.012	0.966±0.010	0.623±0.020○	0.968±0.009●	**0.968±0.011●**
isolet	0.995±0.003	0.501±0.503○	0.621±0.207○	0.995±0.003	**0.997±0.002●**
credit-a	0.861±0.014	0.862±0.013	0.596±0.063○	0.863±0.013	**0.864±0.013●**
austra	0.857±0.013	0.842±0.055	0.567±0.044○	0.858±0.013	**0.859±0.015**
australian	0.844±0.019	0.842±0.020	0.576±0.049○	0.858±0.016●	**0.866±0.014●**
fourclass	0.724±0.014	0.377±0.238○	0.641±0.020○	**0.736±0.014●**	0.723±0.014
german	0.711±0.030	0.737±0.014●	0.697±0.017○	0.729±0.017●	**0.738±0.016●**
Ave. accuracy	0.813	0.743	0.650	0.807	0.823
LDM: w/t/l	12/8/0	9/10/1	17/3/0	10/5/5	

w/t/l counts: after t-tests (95% SI)

bold: best

●/○: significantly better/worse than SVM

LDM achieves the best accuracy on 13 data sets

LDM outperforms SVM, MDO, MAMC and KM-OMD for 12, 9 ,17, 10 times respectively

Results (cont.)

Accuracy on twenty regular scale data sets with RBF kernel

Dataset	SVM	MDO	MAMC	KM-OMD	LDM
promoters	0.684±0.100	N/A	0.638±0.121○	0.701±0.085	**0.715±0.074●**
planning-relax	**0.708±0.035**	N/A	0.706±0.034	0.683±0.031○	0.707±0.034
colic	0.822±0.033	N/A	0.623±0.037○	0.825±0.024	**0.841±0.018●**
parkinsons	**0.929±0.029**	N/A	0.852±0.036○	0.906±0.033○	0.927±0.029
colic.ORIG	0.638±0.043	N/A	0.623±0.027	0.621±0.039	**0.641±0.044**
sonar	0.842±0.034	N/A	0.753±0.052○	0.821±0.051○	**0.846±0.032**
vote	0.946±0.016	N/A	0.913±0.019○	0.930±0.029○	**0.968±0.013●**
house	0.953±0.020	N/A	0.561±0.139○	0.938±0.022○	**0.964±0.013●**
heart	0.808±0.025	N/A	0.540±0.043○	0.805±0.048	**0.822±0.029●**
breast-cancer	0.729±0.030	N/A	0.706±0.027○	0.691±0.024○	**0.753±0.027●**
haberman	0.727±0.024	N/A	**0.742±0.021●**	0.676±0.042○	0.731±0.027
vehicle	0.992±0.007	N/A	0.924±0.025○	0.988±0.008○	**0.993±0.006**
clean1	0.890±0.020	N/A	0.561±0.025○	0.772±0.043○	**0.891±0.024**
wdbc	0.951±0.011	N/A	0.740±0.042○	0.941±0.040	**0.961±0.010●**
isolet	0.998±0.002	N/A	0.994±0.004○	0.995±0.003○	**0.998±0.002**
credit-a	0.858±0.014	N/A	0.542±0.032○	0.845±0.029○	**0.861±0.013**
austra	0.853±0.013	N/A	0.560±0.018○	0.854±0.017	**0.857±0.014●**
australian	0.815±0.014	N/A	0.554±0.015○	**0.860±0.014●**	0.854±0.016●
fourclass	**0.998±0.003**	N/A	0.791±0.014○	0.838±0.014○	0.998±0.003
german	0.731±0.019	N/A	0.697±0.017○	0.742±0.017●	**0.743±0.016●**
Ave. accuracy	0.844	N/A	0.701	0.822	0.854
LDM: w/t/l	10/10/0	N/A	18/1/1	15/5/0	

w/t/l counts: after t-tests (95% SI)

bold: best

●/○: significantly better/worse than SVM

LDM achieves the best accuracy on 15 data sets

LDM outperforms SVM, MAMC and KM-OMD for 10, 18, 15 times respectively

Results (cont.)

Accuracy on twelve large scale data sets with linear kernel

Results (cont.)

Accuracy on twelve large scale data sets with linear kernel

Dataset	SVM	MDO	MAMC	KM-OMD	LDM
farm-ads	0.880±0.007	0.880±0.007	0.759±0.038∘	N/A	**0.890±0.008•**
news20	0.954±0.002	0.948±0.002∘	0.772±0.017∘	N/A	**0.960±0.001•**
adult-a	0.845±0.002	0.788±0.053∘	0.759±0.002∘	N/A	**0.846±0.003•**
w8a	0.983±0.001	**0.985±0.001•**	0.971±0.001∘	N/A	0.983±0.001
cod-rna	**0.899±0.001**	0.774±0.203	0.667±0.001∘	N/A	0.899±0.001
real-sim	0.961±0.001	0.955±0.002∘	0.744±0.004∘	N/A	**0.971±0.001•**
ijcnn1	0.921±0.003	**0.921±0.002**	0.904±0.001∘	N/A	0.921±0.002
skin	0.934±0.001	0.929±0.003∘	0.792±0.000∘	N/A	**0.934±0.001**
covtype	0.762±0.001	0.760±0.003∘	0.628±0.002∘	N/A	**0.763±0.001**
rcv1	0.969±0.000	0.959±0.000∘	0.913±0.000∘	N/A	**0.977±0.000•**
url	**0.993±0.006**	0.993±0.006	0.670±0.000∘	N/A	0.993±0.006
kdd2010	0.852±0.001	N/A	0.853±0.000•	N/A	**0.881±0.001•**
Ave. accuracy	0.913	0.899	0.786	N/A	0.919
LDM: w/t/l	6/6/0	7/3/1	12/0/0	N/A	

Results (cont.)

Accuracy on twelve large scale data sets with linear kernel

Dataset	SVM	MDO	MAMC	KM-OMD	LDM
farm-ads	0.880±0.007	0.880±0.007	0.759±0.038∘	N/A	**0.890±0.008•**
news20	0.954±0.002	0.948±0.002∘	0.772±0.017∘	N/A	**0.960±0.001•**
adult-a	0.845±0.002	0.788±0.053∘	0.759±0.002∘	N/A	**0.846±0.003•**
w8a	0.983±0.001	**0.985±0.001•**	0.971±0.001∘	N/A	0.983±0.001
cod-rna	**0.899±0.001**	0.774±0.203	0.667±0.001∘	N/A	0.899±0.001
real-sim	0.961±0.001	0.955±0.002∘	0.744±0.004∘	N/A	**0.971±0.001•**
ijcnn1	0.921±0.003	**0.921±0.002**	0.904±0.001∘	N/A	0.921±0.002
skin	0.934±0.001	0.929±0.003∘	0.792±0.000∘	N/A	**0.934±0.001**
covtype	0.762±0.001	0.760±0.003∘	0.628±0.002∘	N/A	**0.763±0.001**
rcv1	0.969±0.000	0.959±0.000∘	0.913±0.000∘	N/A	**0.977±0.000•**
url	**0.993±0.006**	0.993±0.006	0.670±0.000∘	N/A	0.993±0.006
kdd2010	0.852±0.001	N/A	0.853±0.000•	N/A	**0.881±0.001•**
Ave. accuracy	0.913	0.899	0.786	N/A	0.919
LDM: w/t/l	6/6/0	7/3/1	12/0/0	N/A	

w/t/l counts: after t-tests (95% SI)

bold: best

•/∘: significantly better/worse than SVM

Results (cont.)

Accuracy on twelve large scale data sets with linear kernel

Dataset	SVM	MDO	MAMC	KM-OMD	LDM
farm-ads	0.880±0.007	0.880±0.007	0.759±0.038∘	N/A	**0.890±0.008•**
news20	0.954±0.002	0.948±0.002∘	0.772±0.017∘	N/A	**0.960±0.001•**
adult-a	0.845±0.002	0.788±0.053∘	0.759±0.002∘	N/A	**0.846±0.003•**
w8a	0.983±0.001	**0.985±0.001•**	0.971±0.001∘	N/A	0.983±0.001
cod-rna	**0.899±0.001**	0.774±0.203	0.667±0.001∘	N/A	0.899±0.001
real-sim	0.961±0.001	0.955±0.002∘	0.744±0.004∘	N/A	**0.971±0.001•**
ijcnn1	0.921±0.003	**0.921±0.002**	0.904±0.001∘	N/A	0.921±0.002
skin	0.934±0.001	0.929±0.003∘	0.792±0.000∘	N/A	**0.934±0.001**
covtype	0.762±0.001	0.760±0.003∘	0.628±0.002∘	N/A	**0.763±0.001**
rcv1	0.969±0.000	0.959±0.000∘	0.913±0.000∘	N/A	**0.977±0.000•**
url	**0.993±0.006**	0.993±0.006	0.670±0.000∘	N/A	0.993±0.006
kdd2010	0.852±0.001	N/A	0.853±0.000•	N/A	**0.881±0.001•**
Ave. accuracy	0.913	0.899	0.786	N/A	0.919
LDM: w/t/l	6/6/0	7/3/1	12/0/0	N/A	

w/t/l counts: after t-tests (95% SI)

bold: best

•/∘: significantly better/worse than SVM

LDM achieves the best accuracy on 8 data sets

Results (cont.)

Accuracy on twelve large scale data sets with linear kernel

Dataset	SVM	MDO	MAMC	KM-OMD	LDM
farm-ads	0.880±0.007	0.880±0.007	0.759±0.038∘	N/A	**0.890±0.008•**
news20	0.954±0.002	0.948±0.002∘	0.772±0.017∘	N/A	**0.960±0.001•**
adult-a	0.845±0.002	0.788±0.053∘	0.759±0.002∘	N/A	**0.846±0.003•**
w8a	0.983±0.001	**0.985±0.001•**	0.971±0.001∘	N/A	0.983±0.001
cod-rna	**0.899±0.001**	0.774±0.203	0.667±0.001∘	N/A	0.899±0.001
real-sim	0.961±0.001	0.955±0.002∘	0.744±0.004∘	N/A	**0.971±0.001•**
ijcnn1	0.921±0.003	**0.921±0.002**	0.904±0.001∘	N/A	0.921±0.002
skin	0.934±0.001	0.929±0.003∘	0.792±0.000∘	N/A	**0.934±0.001**
covtype	0.762±0.001	0.760±0.003∘	0.628±0.002∘	N/A	**0.763±0.001**
rcv1	0.969±0.000	0.959±0.000∘	0.913±0.000∘	N/A	**0.977±0.000•**
url	**0.993±0.006**	0.993±0.006	0.670±0.000∘	N/A	0.993±0.006
kdd2010	0.852±0.001	N/A	0.853±0.000•	N/A	**0.881±0.001•**
Ave. accuracy	0.913	0.899	0.786	N/A	0.919
LDM: w/t/l	6/6/0	7/3/1	12/0/0	N/A	

w/t/l counts: after t-tests (95% SI)

bold: best

•/∘: significantly better/worse than SVM

LDM outperforms SVM, MDO and MAMC for 6, 7, 12 times respectively

LDM achieves the best accuracy on 8 data sets

Results (cont.)

Accuracy on twelve large scale data sets with linear kernel

Dataset	SVM	MDO	MAMC	KM-OMD	LDM
farm-ads	0.880±0.007	0.880±0.007	0.759±0.038○	N/A	**0.890±0.008●**
news20	0.954±0.002	0.948±0.002○	0.772±0.017○	N/A	**0.960±0.001●**
adult-a	0.845±0.002	0.788±0.053○	0.759±0.002○	N/A	**0.846±0.003●**
w8a	0.983±0.001	**0.985±0.001●**	0.971±0.001○	N/A	0.983±0.001
cod-rna	**0.899±0.001**	0.774±0.203	0.667±0.001○	N/A	0.899±0.001
real-sim	0.961±0.001	0.955±0.002○	0.744±0.004○	N/A	**0.971±0.001●**
ijcnn1	0.921±0.003	**0.921±0.002**	0.904±0.001○	N/A	0.921±0.002
skin	0.934±0.001	0.929±0.003○	0.792±0.000○	N/A	**0.934±0.001**
covtype	0.762±0.001	0.760±0.003○	0.628±0.002○	N/A	**0.763±0.001**
rcv1	0.969±0.000	0.959±0.000○	0.913±0.000○	N/A	**0.977±0.000●**
url	**0.993±0.006**	0.993±0.006	0.670±0.000○	N/A	0.993±0.006
kdd2010	0.852±0.001	N/A	0.853±0.000●	N/A	**0.881±0.001●**
Ave. accuracy	0.913	0.899	0.786	N/A	0.919
LDM: w/t/l	6/6/0	7/3/1	12/0/0	N/A	

w/t/l counts: after t-tests (95% SI)

bold: best

●/○: significantly better/worse than SVM

LDM outperforms SVM, MDO and MAMC for 6, 7, 12 times respectively

LDM achieves the best accuracy on 8 data sets

LDM is significantly better than the other methods

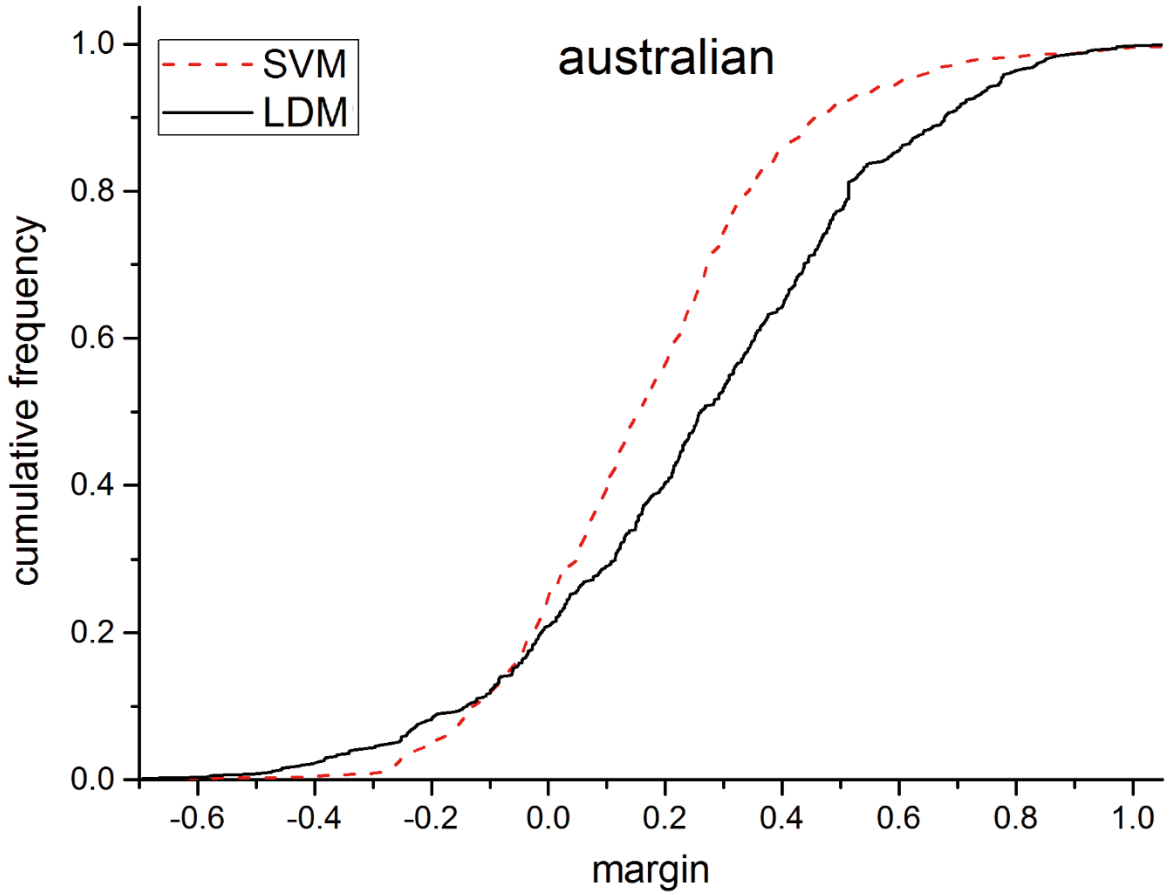
Margin Distribution

Cumulative frequency (y-axis) with respect to margin (x-axis) of SVM and LDM

Margin Distribution

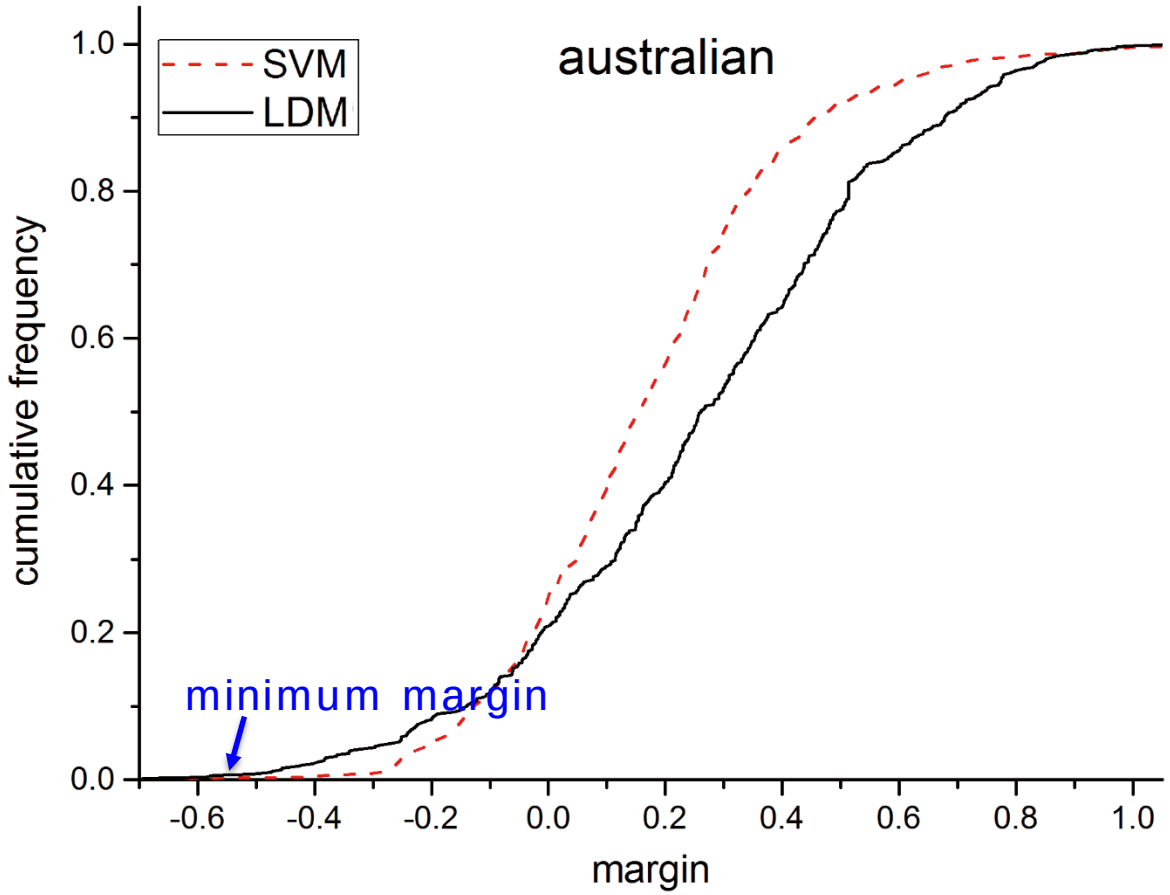
Learning And Mining from DatA
http://lamda.nju.edu.cn

Cumulative frequency (y-axis) with respect to margin (x-axis) of SVM and LDM



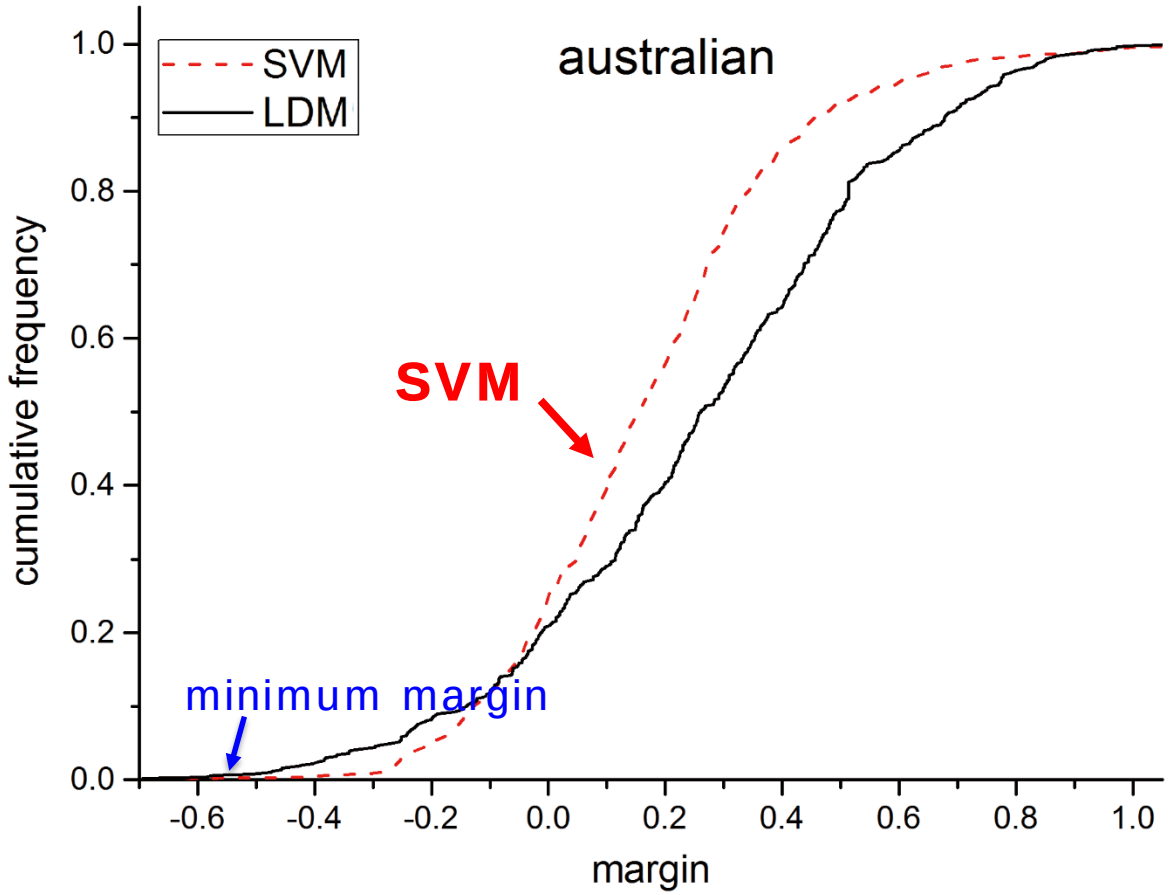
Margin Distribution

Cumulative frequency (y-axis) with respect to margin (x-axis) of SVM and LDM



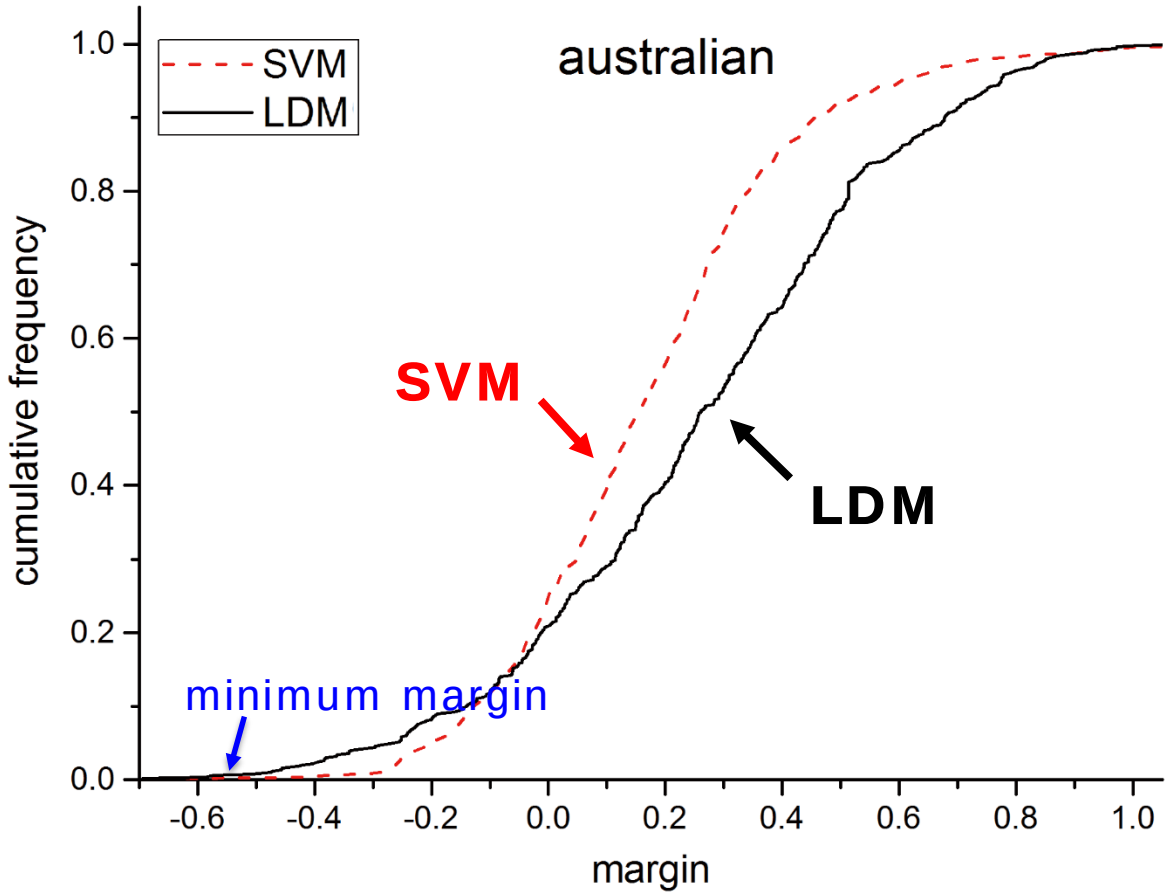
Margin Distribution

Cumulative frequency (y-axis) with respect to margin (x-axis) of SVM and LDM



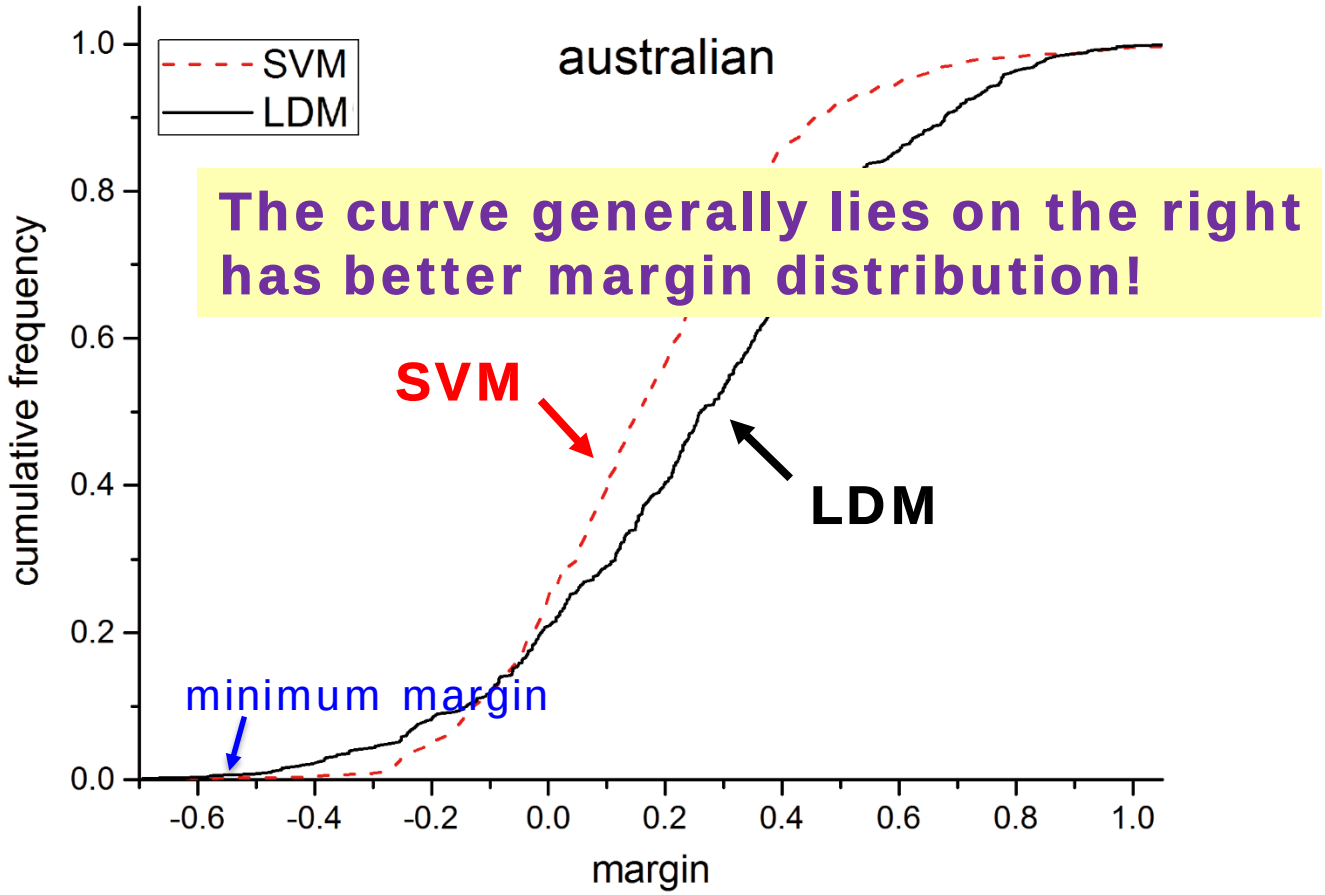
Margin Distribution

Cumulative frequency (y-axis) with respect to margin (x-axis) of SVM and LDM



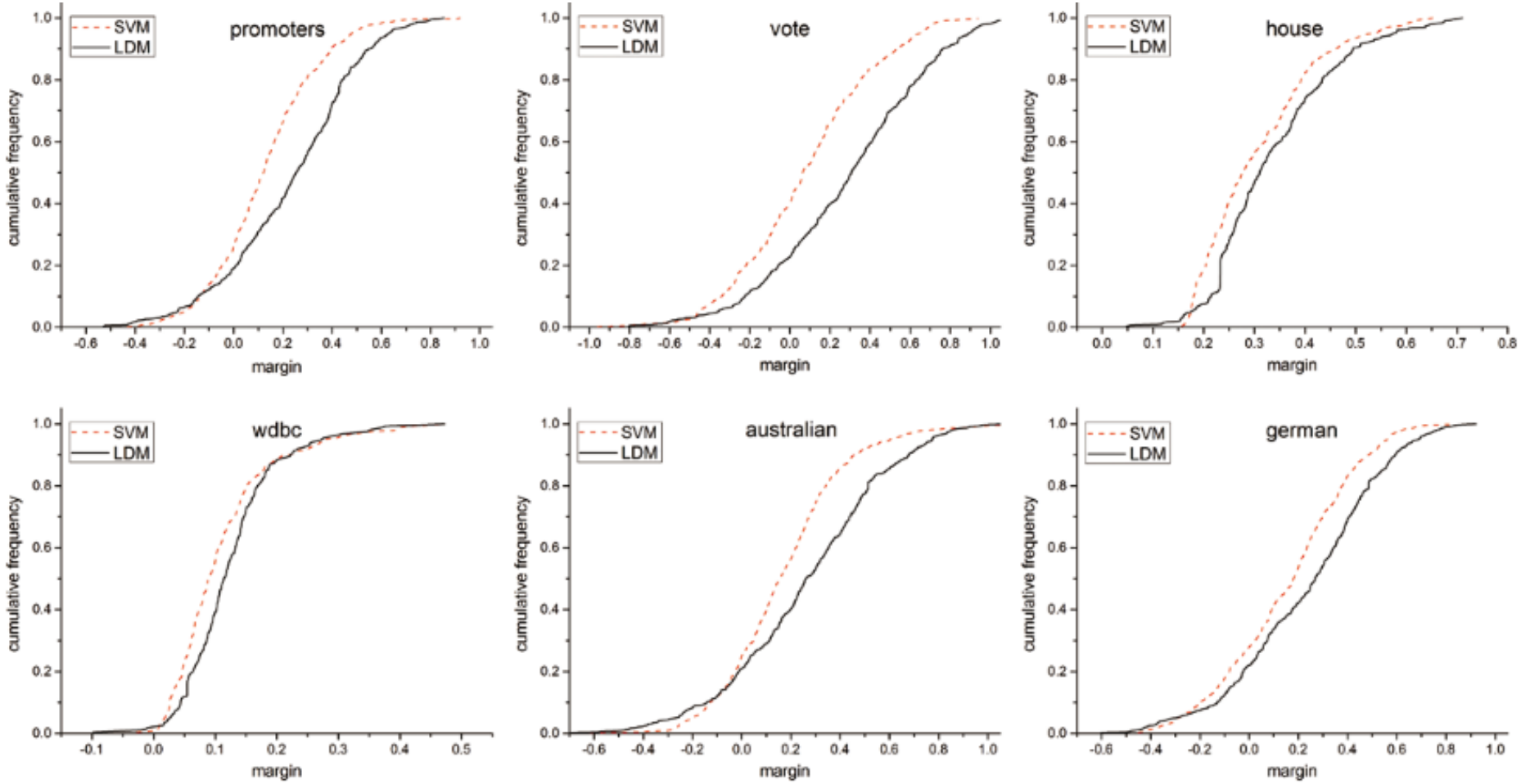
Margin Distribution

Cumulative frequency (y-axis) with respect to margin (x-axis) of SVM and LDM



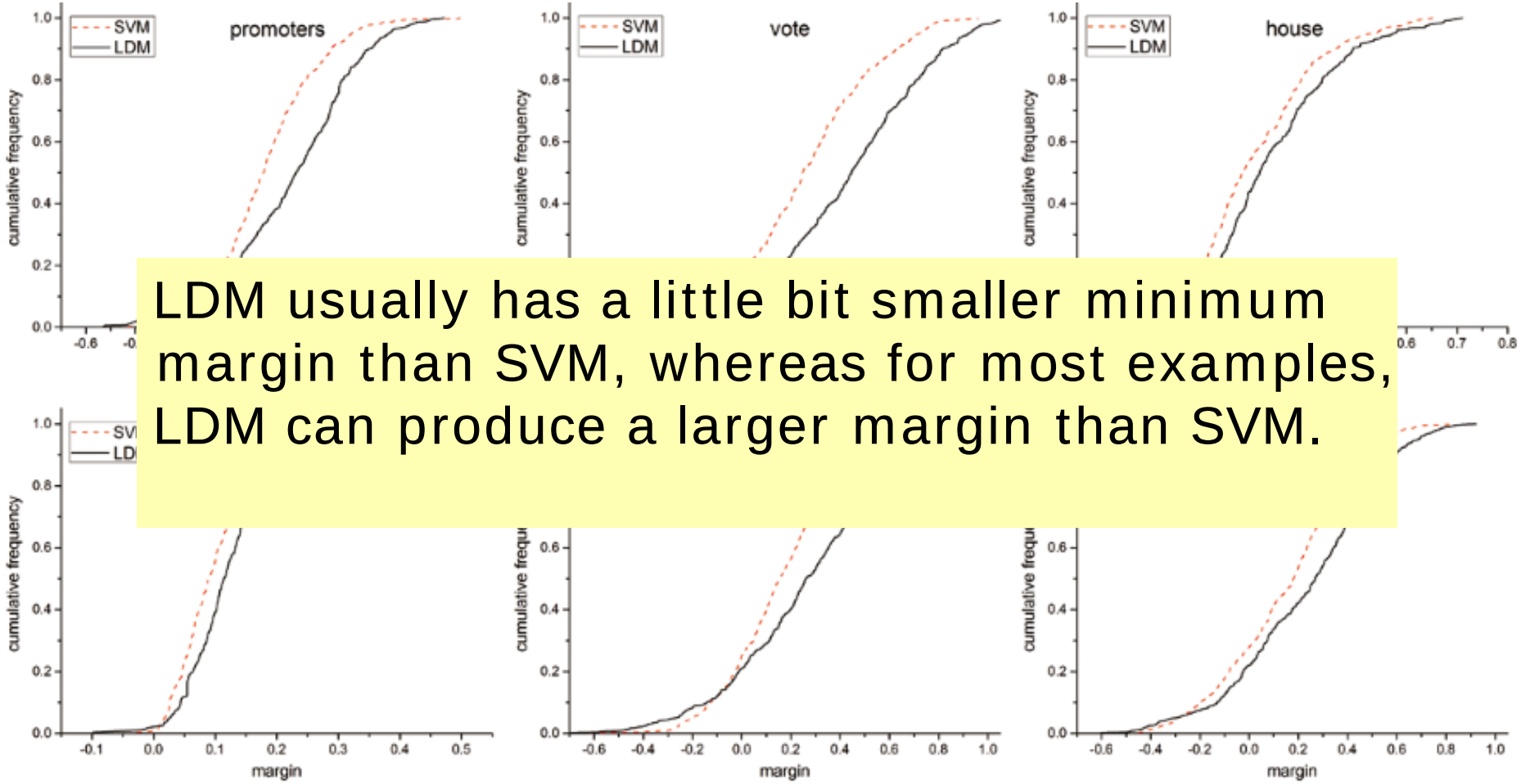
Margin Distribution

Cumulative frequency (y-axis) with respect to margin (x-axis) of SVM and LDM



Margin Distribution

Cumulative frequency (y-axis) with respect to margin (x-axis) of SVM and LDM



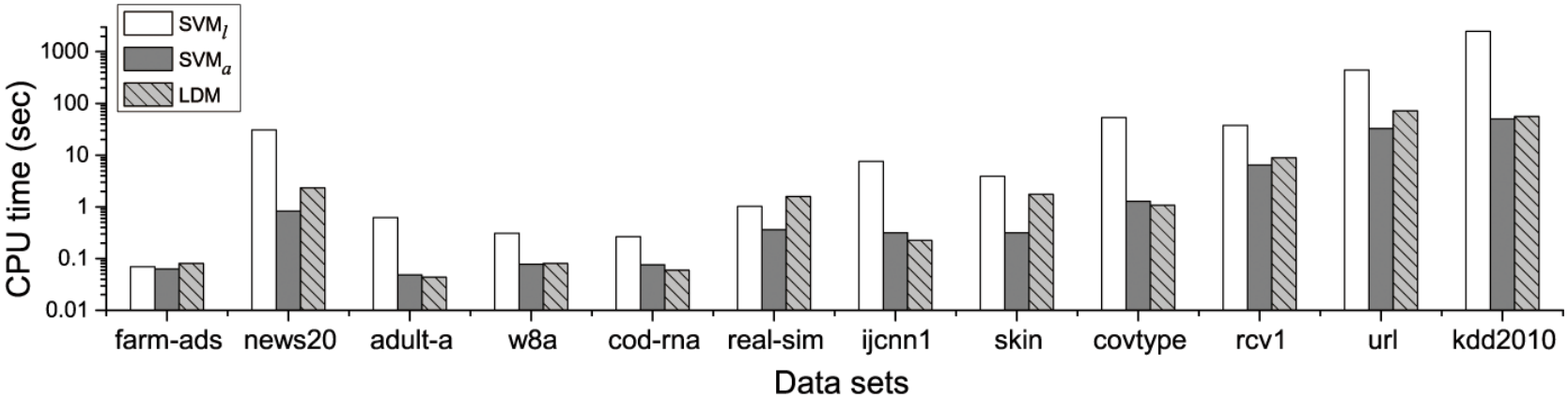
LDM usually has a little bit smaller minimum margin than SVM, whereas for most examples, LDM can produce a larger margin than SVM.

Scalability

The time cost of LDM and SVM on the twelve large scale data sets

Scalability

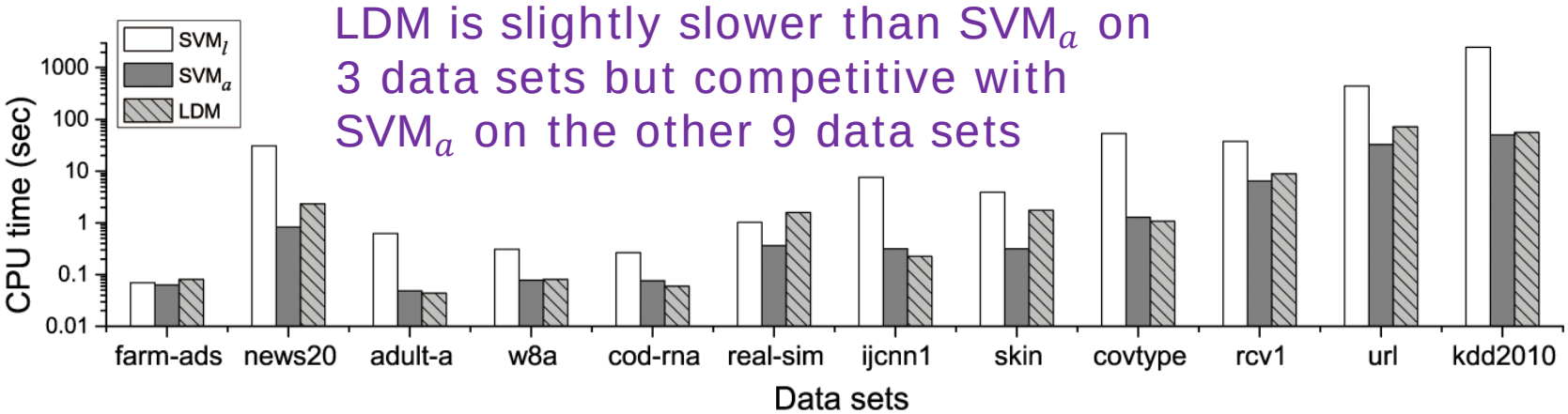
The time cost of LDM and SVM on the twelve large scale data sets



- SVM$_l$: SVM implemented by the LIBLINEAR package
- SVM$_a$: SVM implemented by ASGD
- LDM: LDM implemented by ASGD

Scalability

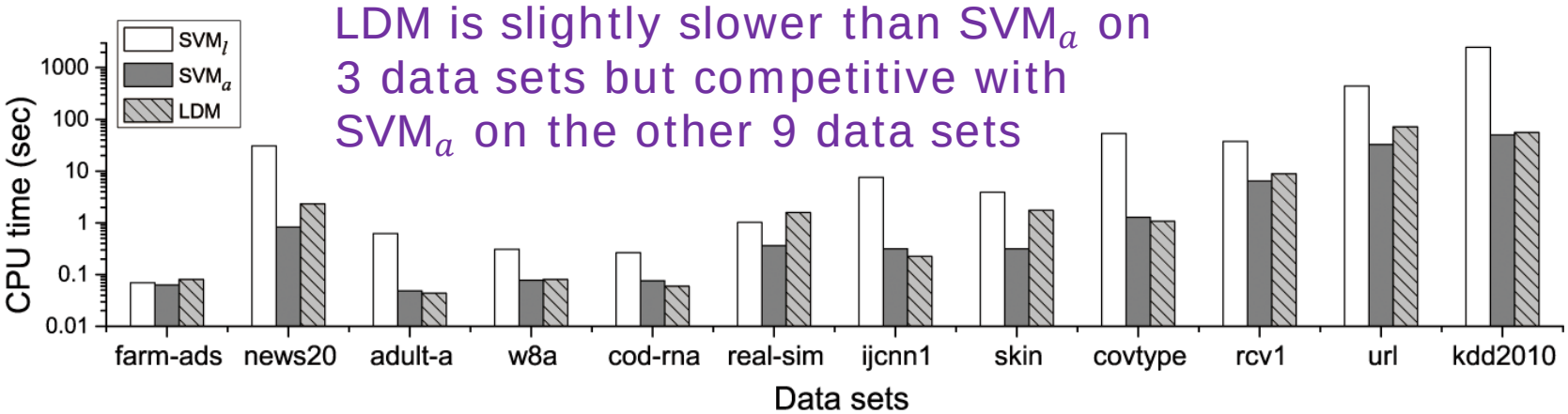
The time cost of LDM and SVM on the twelve large scale data sets



- SVM$_l$: SVM implemented by the LIBLINEAR package
- SVM$_a$: SVM implemented by ASGD
- LDM: LDM implemented by ASGD

Scalability

The time cost of LDM and SVM on the twelve large scale data sets



- SVM$_l$: SVM implemented by the LIBLINEAR package
- SVM$_a$: SVM implemented by ASGD
- LDM: LDM implemented by ASGD

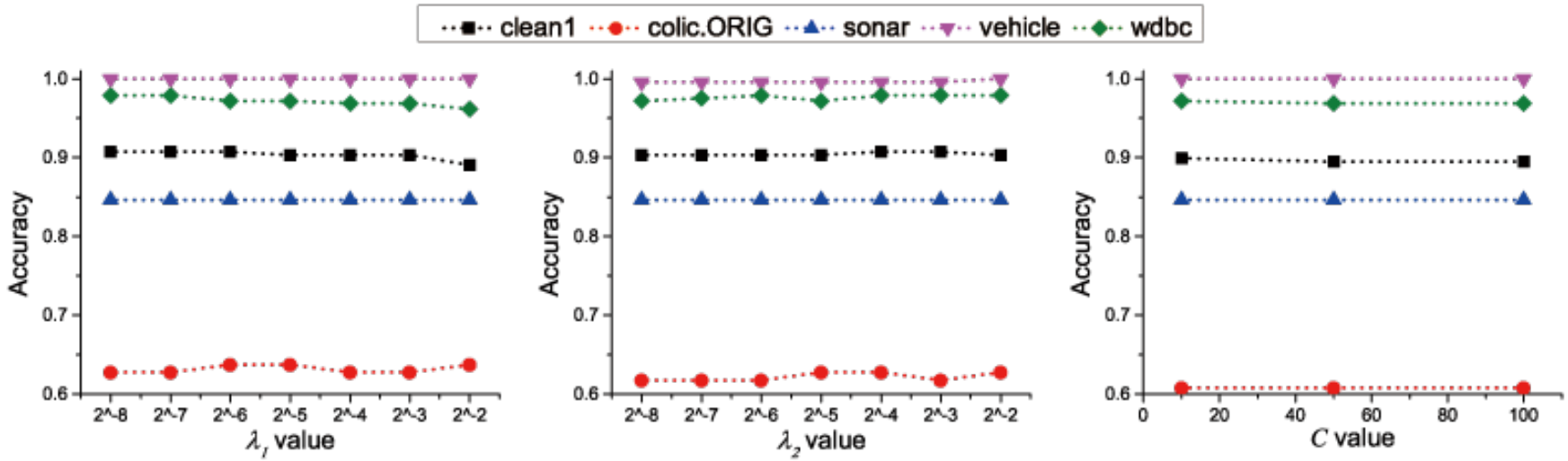
LDM is also computationally efficient.

Parameter Influence

LDM has three parameters: λ_1, λ_2 and C

Parameter Influence

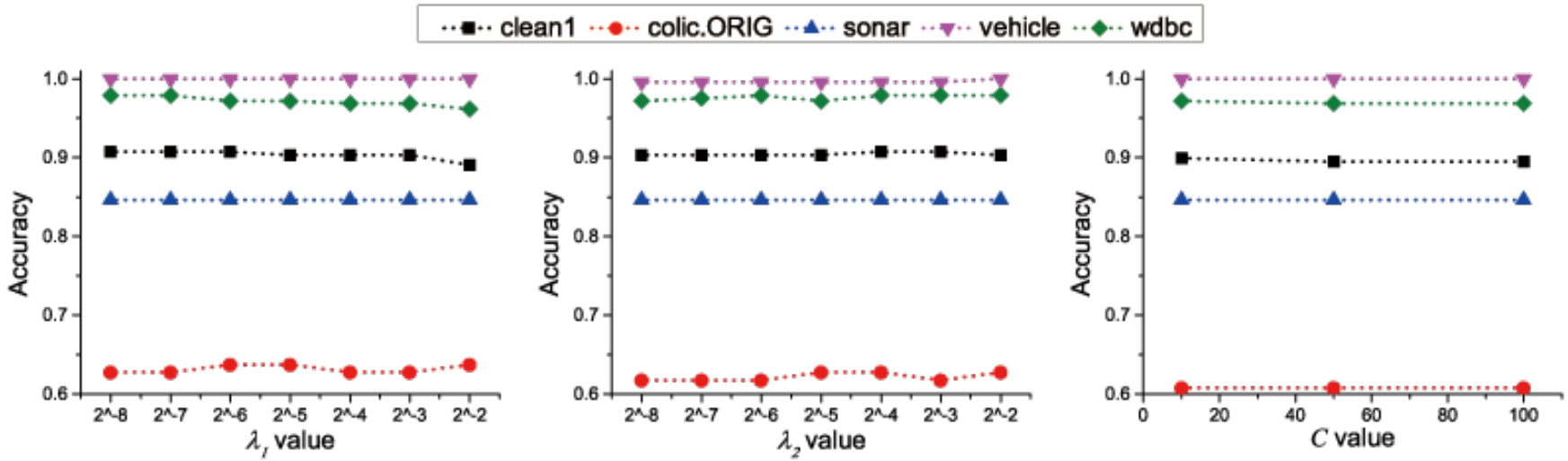
LDM has three parameters: λ_1, λ_2 and C



(a) Influence of λ_1 on LDM (b) Influence of λ_2 on LDM (c) Influence of C on LDM

Parameter Influence

LDM has three parameters: λ_1, λ_2 and C



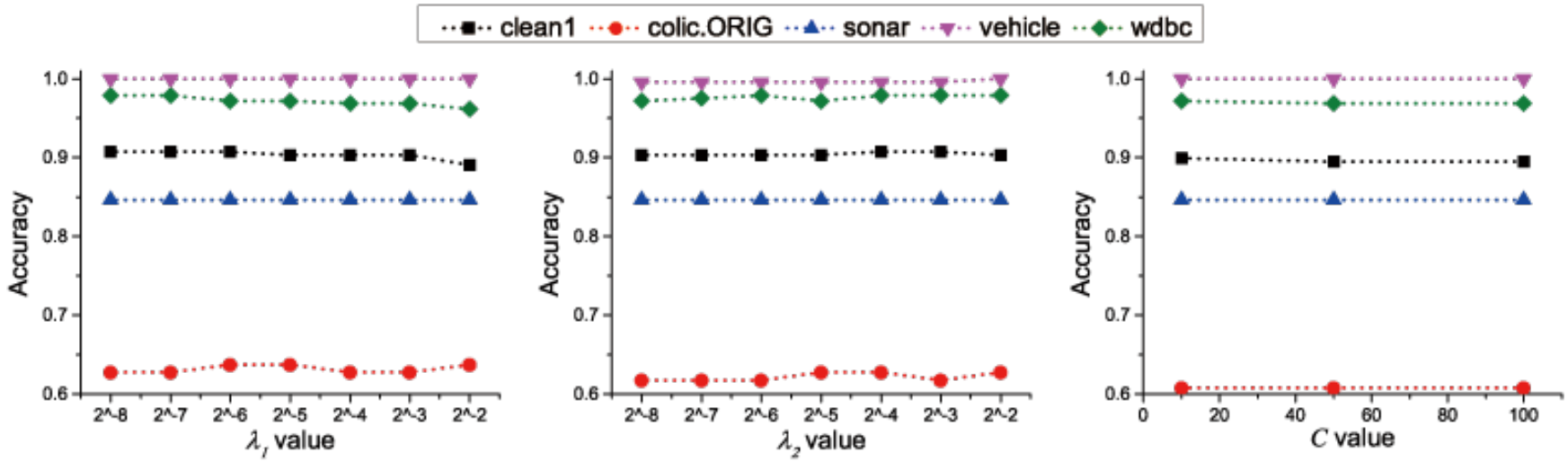
(a) Influence of λ_1 on LDM (b) Influence of λ_2 on LDM (c) Influence of C on LDM

- Vary one parameter and fix the other two as the value suggested by cross validation.
- λ_1 and λ_2 take value of $\{2^{-8}, \ldots\ldots, 2^{-2}\}$, C takes value of {10, 50, 100}

Parameter Influence

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

LDM has three parameters: λ_1, λ_2 and C



(a) Influence of λ_1 on LDM (b) Influence of λ_2 on LDM (c) Influence of C on LDM

- Vary one parameter and fix the other two as the value suggested by cross validation.
- λ_1 and λ_2 take value of $\{2^{-8}, \ldots\ldots, 2^{-2}\}$, C takes value of {10, 50, 100}

LDM is not very sensitive to the setting of the parameters

Outline

- Introduction
- Our Method
- Experiments
- Conclusion

Conclusion

Main contribution:

- ✓ The first study on directly optimizing margin distribution for SVM
  - ✓ We show that directly optimizing the margin distribution for SVM is beneficial.
  - ✓ We propose an effective method LDM that optimizes margin distribution. (The code package for LDM method http://lamda.nju.edu.cn/code_LDM.ashx)

Future work:

- ✓ To generalize LDM to other learning settings
- ✓ . . .

Conclusion

Main contribution:

- ✓ The first study on directly optimizing margin distribution for SVM
  - ✓ We show that directly optimizing the margin distribution for SVM is beneficial.
  - ✓ We propose an effective method LDM that optimizes margin distribution. (The code package for LDM method http://lamda.nju.edu.cn/code_LDM.ashx)

Future work:

- ✓ To generalize LDM to other learning settings
- ✓ . . .

Thanks!