Predicting Multiple Responses with Boosting and Trees

Ping Li[1,2] and John Abowd[2]

Department of Statistics & Biostatistics

Department of Computer Science

Rutgers University (1)

and

Department of Statistical Science

Cornell University (2)

November 4, 2013

Motivation: Multi-Label Learning

- Traditional classification methods only deal with a single response (label) for each example. For example, handwritten digit recognition (0, 1, 2, ..., 9).

- In many practical problems, however, one example may involve multiple responses (labels). In scene classification, an image might be both "Mountain" and "Beach". In Census survey forms, one can choose to declare multiple races, for example, both "American Indian" and "White".

- Multi-label learning is more challenging. Our working progress demonstrates that it is very promising to use boosting and trees for this type of problems.

History and Progress

- **LogitBoost** (Fridman et. al.,2000) is a well-known work on boosting in statistics. It is also known that the original version had numerical problem.
- **MART** (Fridman, 2001) avoided the numerical problem by using only the first-order information to build the trees (the base learner for boosting). The algorithm is extremely popular in industry.
- **ABC-MART, ABC-LogitBoost** (Ping Li, 2009, 2010) substantially improved MART and logitboost by writing the traditional derivatives of logistic regression in a different way, for the task of multi-class (not multi-label) classification.
- **Robust LogitBoost** (Ping Li, 2010) derived the new tree-split criterion for logitboost and fully solved the numerical issue. (Robust) Logitboost is often more accurate than MART due to the use of second-order information.
- Our idea is to extend multi-class boosting algorithms to multi-label settings, using essentially the same (logistic regression) framework.

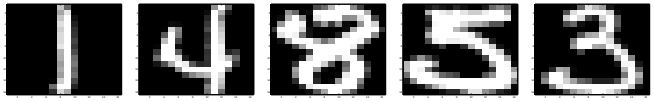
Why Tree-Based Boosting Algorithms Are Popular in Industry?

- Scale up easily to large datasets.
- No need to clean / transform / normalized / kernelize the data.
- Few parameters and parameter tuning is simple.

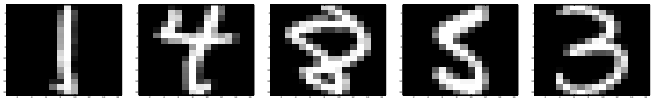
What is Classification?

An Example: USPS Handwritten Zipcode Recognition

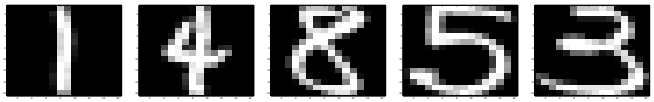
Person 1:



Person 2:



Person 3:



The task: Teach the machine to automatically recognize the 10 digits.

Multi-Class Classification

Given a training data set

$$\{y_i,\ X_i\}_{i=1}^{N},\qquad X_i \in \mathbb{R}^p,\qquad y_i \in \{0, 1, 2, ..., K-1\}$$

the task is to learn a function to predict the class label y_i from X_i.

- $K = 2$: binary classification
- $K > 2$: multi-class classification

Many important practical problems can be cast as (multi-class) classification.

For example, Li, Burges, and Wu, NIPS 2007

Learning to Ranking Using Multiple Classification and Gradient Boosting.

Logistic Regression for Classification

First learn the class probabilities

$$\hat{p}_k = \mathbf{Pr}\left\{y = k|X\right\}, \qquad k = 0, 1, ..., K-1,$$

$$\sum_{k=0}^{K-1} \hat{p}_k = 1, \qquad \left(\text{only } K-1 \text{ degrees of freedom}\right).$$

Then assign the class label according to

$$\hat{y}|X = \underset{k}{\text{argmax}}\ \hat{p}_k$$

Multinomial Logit Probability Model

$$p_k = \frac{e^{F_k}}{\sum_{s=0}^{K-1} e^{F_s}}$$

where $F_k = F_k(\mathbf{x})$ is the function to be learned from the data.

Classical logistic regression:

$$F(\mathbf{x}) = \beta^{\mathsf{T}}\mathbf{x}$$

The task is to learn the coefficients β.

Flexible additive modeling:

$$F(\mathbf{x}) = F^{(M)}(\mathbf{x}) = \sum_{m=1}^{M} \rho_m h(\mathbf{x}; \mathbf{a}_m),$$

$h(\mathbf{x}; \mathbf{a})$ is a pre-specified function (e.g., trees).

The task is to learn the parameters ρ_m and $\mathbf{a}_m$.

————————-

Both **LogitBoost** (Friedman et. al, 2000) and **MART** (Multiple Additive Regression Trees, Friedman 2001) adopted this model.

Learning Logistic Regression by Maximum Likelihood

Seek $F_{i,k}$ to maximize the mutlinomial likelihood: Suppose $y_i = k$,

$$Lik \propto p_{i,0}^0 \times ... \times {p_{i,k}}^1 \times ... \times p_{i,K-1}^0 = p_{i,k}$$

or equivalently, maximizing the log likelihood:

$$\log Lik \propto \log p_{i,k}$$

Or equivalently, minimizing the negative log likelihood loss

$$L_i = -\log p_{i,k}, \qquad (y_i = k)$$

The Negative Log-Likelihood Loss

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left\{ - \sum_{k=0}^{K-1} r_{i,k} \log p_{i,k} \right\}$$

$$r_{i,k} = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{otherwise} \end{cases}$$

Two Basic Optimization Methods for Maximum Likelihood

1. **Newton's Method**

   Uses the first and second derivatives of the loss function.

   The method in LogitBoost.

2. **Gradient Descent**

   Only uses the first order derivative of the loss function.

   ——————————-

   MART used a creative combination of gradient descent and Newton's method.

Derivatives Used in LogitBoost and MART

The loss function:

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left\{ - \sum_{k=0}^{K-1} r_{i,k} \log p_{i,k} \right\}$$

The first derivative:

$$\frac{\partial L_i}{\partial F_{i,k}} = - \left(r_{i,k} - p_{i,k} \right)$$

The second derivative:

$$\frac{\partial^2 L_i}{\partial F_{i,k}^2} = p_{i,k} \left(1 - p_{i,k} \right) .$$

The Original LogitBoost Algorithm

1: $F_{i,k} = 0,\ p_{i,k} = \frac{1}{K}, \quad k = 0$ to $K-1,\ i = 1$ to N

2: For $m = 1$ to M Do

3: For $k = 0$ to $K-1$, Do

4: $w_{i,k} = p_{i,k}\left(1 - p_{i,k}\right), \qquad z_{i,k} = \frac{r_{i,k} - p_{i,k}}{p_{i,k}\left(1 - p_{i,k}\right)}.$

5: Fit the function $f_{i,k}$ by a weighted least-square of $z_{i,k}$ to $\mathbf{x}_i$ with weights $w_{i,k}$.

6: $F_{i,k} = F_{i,k} + \nu \frac{K-1}{K}\left(f_{i,k} - \frac{1}{K}\sum_{k=0}^{K-1} f_{i,k}\right)$

7: End

8: $p_{i,k} = \exp(F_{i,k}) / \sum_{s=0}^{K-1} \exp(F_{i,s}), \quad k = 0$ to $K-1,\ i = 1$ to N

9: End

The Original MART Algorithm

1: $F_{i,k} = 0,\ p_{i,k} = \frac{1}{K},\quad k = 0$ to $K-1,\ i = 1$ to N

2: For $m = 1$ to M Do

3: For $k = 0$ to $K-1$ Do

4: $\{R_{j,k,m}\}_{j=1}^{J} = J$-terminal node regression tree from $\{r_{i,k} - p_{i,k},\ \mathbf{x}_i\}_{i=1}^{N}$

5: $\beta_{j,k,m} = \frac{K-1}{K} \frac{\sum_{\mathbf{x}_i \in R_{j,k,m}} r_{i,k} - p_{i,k}}{\sum_{\mathbf{x}_i \in R_{j,k,m}} \left(1-p_{i,k}\right) p_{i,k}}$

6: $F_{i,k} = F_{i,k} + \nu \sum_{j=1}^{J} \beta_{j,k,m} 1_{\mathbf{x}_i \in R_{j,k,m}}$

7: End

8: $p_{i,k} = \exp(F_{i,k}) / \sum_{s=0}^{K-1} \exp(F_{i,s}),\quad k = 0$ to $K-1,\ i = 1$ to N

9: End

The Numerical Issue in LoigtBoost

4: $w_{i,k} = p_{i,k}\left(1 - p_{i,k}\right)$, $z_{i,k} = \frac{r_{i,k} - p_{i,k}}{p_{i,k}\left(1 - p_{i,k}\right)}$.

5: Fit the function $f_{i,k}$ by a weighted least-square of $z_{i,k}$ to $\mathbf{x}_i$ with weights $w_{i,k}$.

6: $F_{i,k} = F_{i,k} + \nu \frac{K-1}{K}\left(f_{i,k} - \frac{1}{K}\sum_{k=0}^{K-1} f_{i,k}\right)$

The "instability issue":

When $p_{i,k}$ is close to 0 or 1, $z_{i,k} = z_{i,k} = \frac{r_{i,k} - p_{i,k}}{p_{i,k}(1-p_{i,k})}$ may approach infinity.

Robust LogitBoost avoids this **pointwise thresholding** and is essentially free of numerical problems.

Tree-Splitting Using the Second-Order Information

Feature values: x_i, $i = 1$ to N. Assume $x_1 \leq x_2 \leq ... \leq x_N$.

Weight values: w_i, $i = 1$ to N. **Response values**: z_i, $i = 1$ to N.

We seek the index s, $1 \leq s < N$, to maximize the gain of weighted SE:

$$\begin{aligned} Gain(s) =& SE_T - (SE_L + SE_R) \\ =& \sum_{i=1}^{N} (z_i - \bar{z})^2 w_i - \left[\sum_{i=1}^{s} (z_i - \bar{z}_L)^2 w_i + \sum_{i=s+1}^{N} (z_i - \bar{z}_R)^2 w_i \right] \end{aligned}$$

where $\bar{z} = \frac{\sum_{i=1}^{N} z_i w_i}{\sum_{i=1}^{N} w_i}$, $\bar{z}_L = \frac{\sum_{i=1}^{s} z_i w_i}{\sum_{i=1}^{s} w_i}$, $\bar{z}_R = \frac{\sum_{i=s+1}^{N} z_i w_i}{\sum_{i=s+1}^{N} w_i}$.

After simplification, we obtain

$$Gain(s) = \frac{\left[\sum_{i=1}^{s} z_i w_i\right]^2}{\sum_{i=1}^{s} w_i} + \frac{\left[\sum_{i=s+1}^{N} z_i w_i\right]^2}{\sum_{i=s+1}^{N} w_i} - \frac{\left[\sum_{i=1}^{N} z_i w_i\right]^2}{\sum_{i=1}^{N} w_i}$$

$$= \frac{\left[\sum_{i=1}^{s} r_{i,k} - p_{i,k}\right]^2}{\sum_{i=1}^{s} p_{i,k}(1-p_{i,k})} + \frac{\left[\sum_{i=s+1}^{N} r_{i,k} - p_{i,k}\right]^2}{\sum_{i=s+1}^{N} p_{i,k}(1-p_{i,k})} - \frac{\left[\sum_{i=1}^{N} r_{i,k} - p_{i,k}\right]^2}{\sum_{i=1}^{N} p_{i,k}(1-p_{i,k})}.$$

Recall $w_i = p_{i,k}(1-p_{i,k}), \quad z_i = \frac{r_{i,k}-p_{i,k}}{p_{i,k}(1-p_{i,k})}$.

This procedure is numerically stable.

MART only used the first order information to construct the trees:

$$MARTGain(s) = \frac{1}{s}\left[\sum_{i=1}^{s} r_{i,k} - p_{i,k}\right]^2 + \frac{1}{N-s}\left[\sum_{i=s+1}^{N} r_{i,k} - p_{i,k}\right]^2 - \frac{1}{N}\left[\sum_{i=1}^{N} r_{i,k} - p_{i,k}\right]^2.$$

Which can also be derived by letting weights $w_{i,k} = 1$ and response $z_{i,k} = r_{i,k} - p_{i,k}$.

LogitBoost used more information and could be more accurate in many datasets.

Robust LogitBoost

1: $F_{i,k} = 0$, $p_{i,k} = \frac{1}{K}$, $k = 0$ to $K-1$, $i = 1$ to N

2: For $m = 1$ to M Do

3: For $k = 0$ to $K-1$ Do

4: $\{R_{j,k,m}\}_{j=1}^{J} = J$-terminal node regression tree from $\{r_{i,k} - p_{i,k},\ \mathbf{x}_i\}_{i=1}^{N}$,

: with weights $p_{i,k}(1-p_{i,k})$.

5: $\beta_{j,k,m} = \frac{K-1}{K} \frac{\sum_{\mathbf{x}_i \in R_{j,k,m}} r_{i,k} - p_{i,k}}{\sum_{\mathbf{x}_i \in R_{j,k,m}} (1-p_{i,k}) p_{i,k}}$

6: $F_{i,k} = F_{i,k} + \nu \sum_{j=1}^{J} \beta_{j,k,m} 1_{\mathbf{x}_i \in R_{j,k,m}}$

7: End

8: $p_{i,k} = \exp(F_{i,k}) / \sum_{s=0}^{K-1} \exp(F_{i,s}), \quad k = 0$ to $K-1,\ i = 1$ to N

9: End

Experiments on Binary Classification

(Multi-class classification is even more interesting!)

Data

IJCNN1: 49990 training samples, 91701 test samples

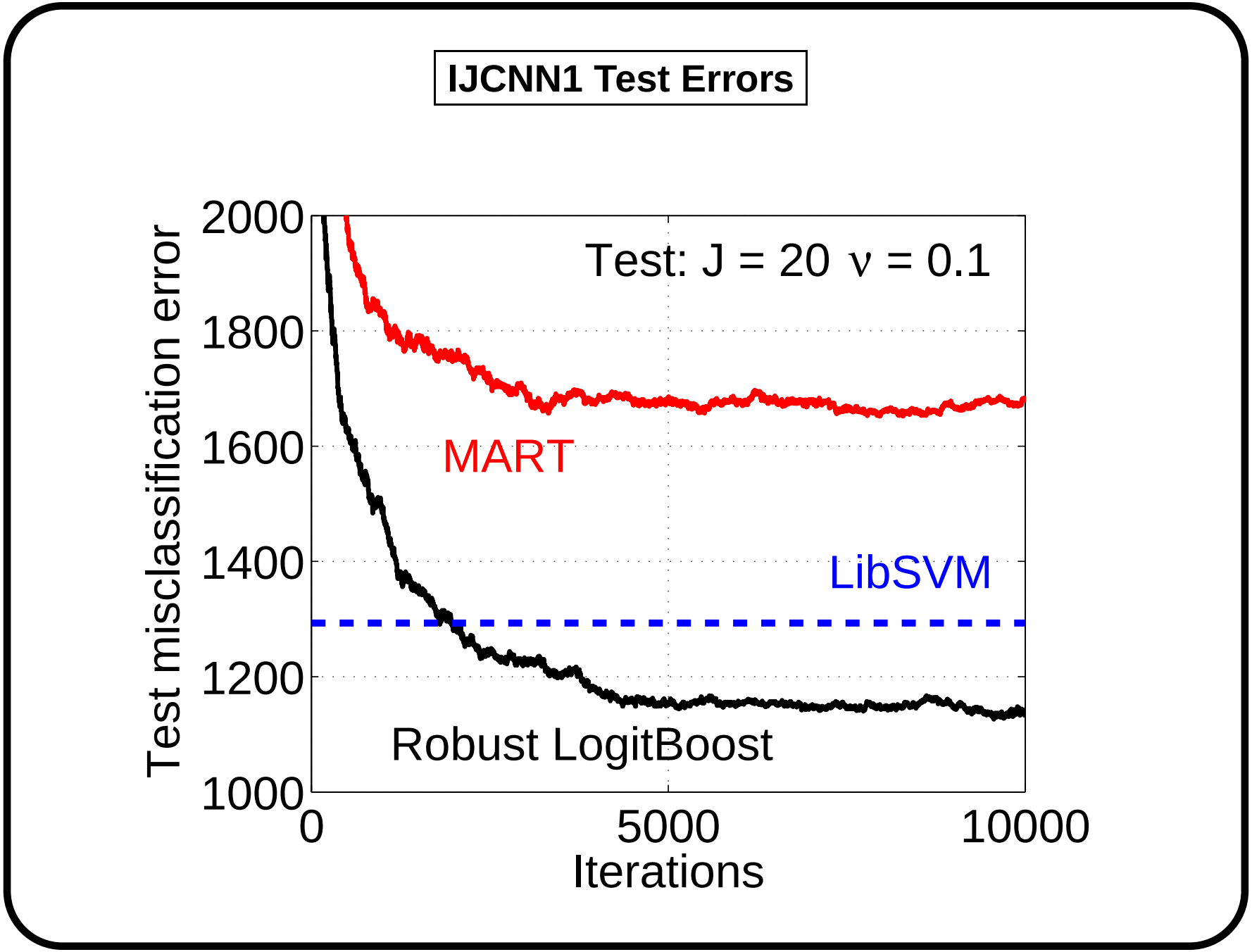
This dataset was used in a competition. LibSVM was the winner.

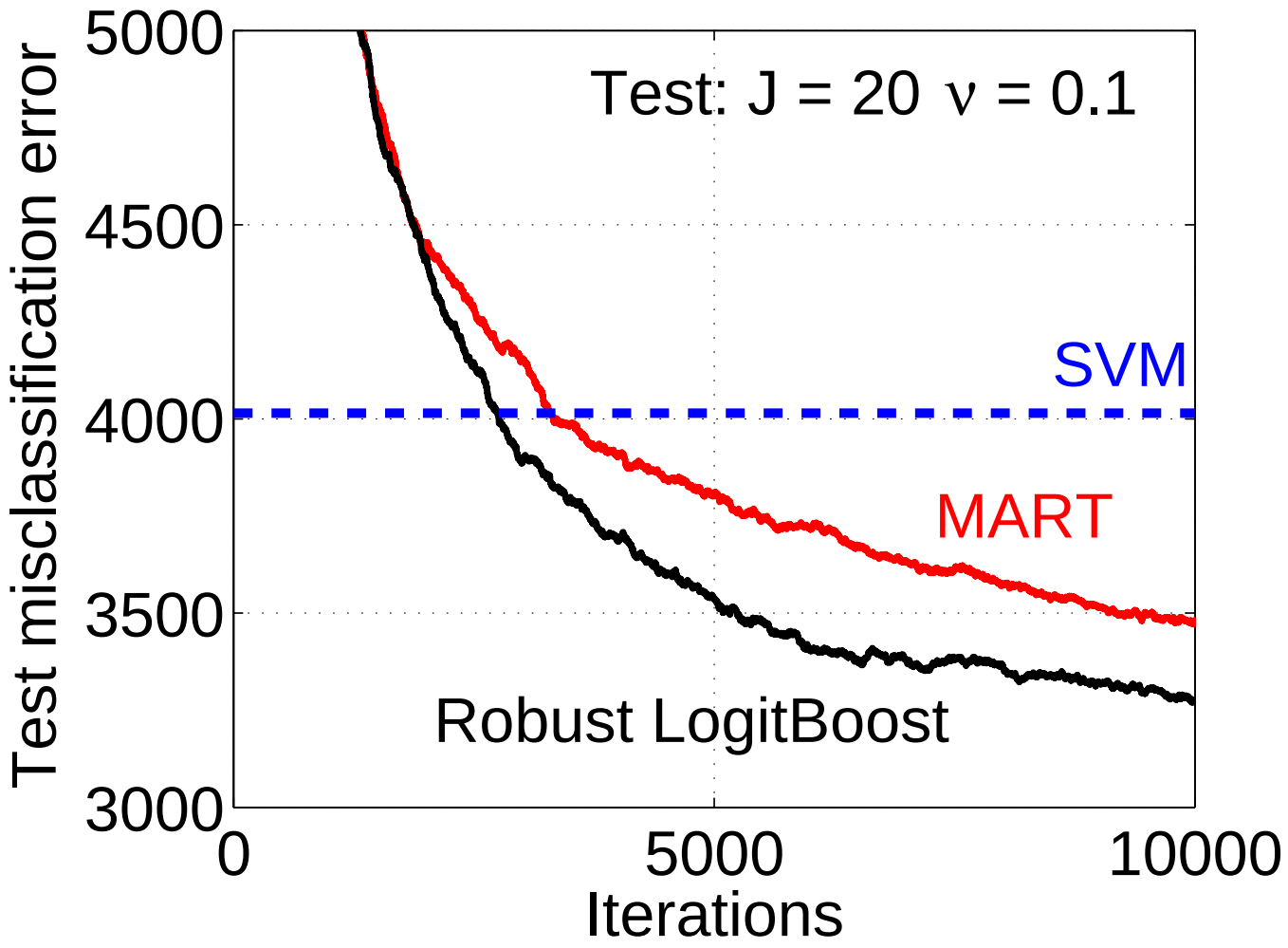
Forest100k: 100000 training samples, 50000 test samples

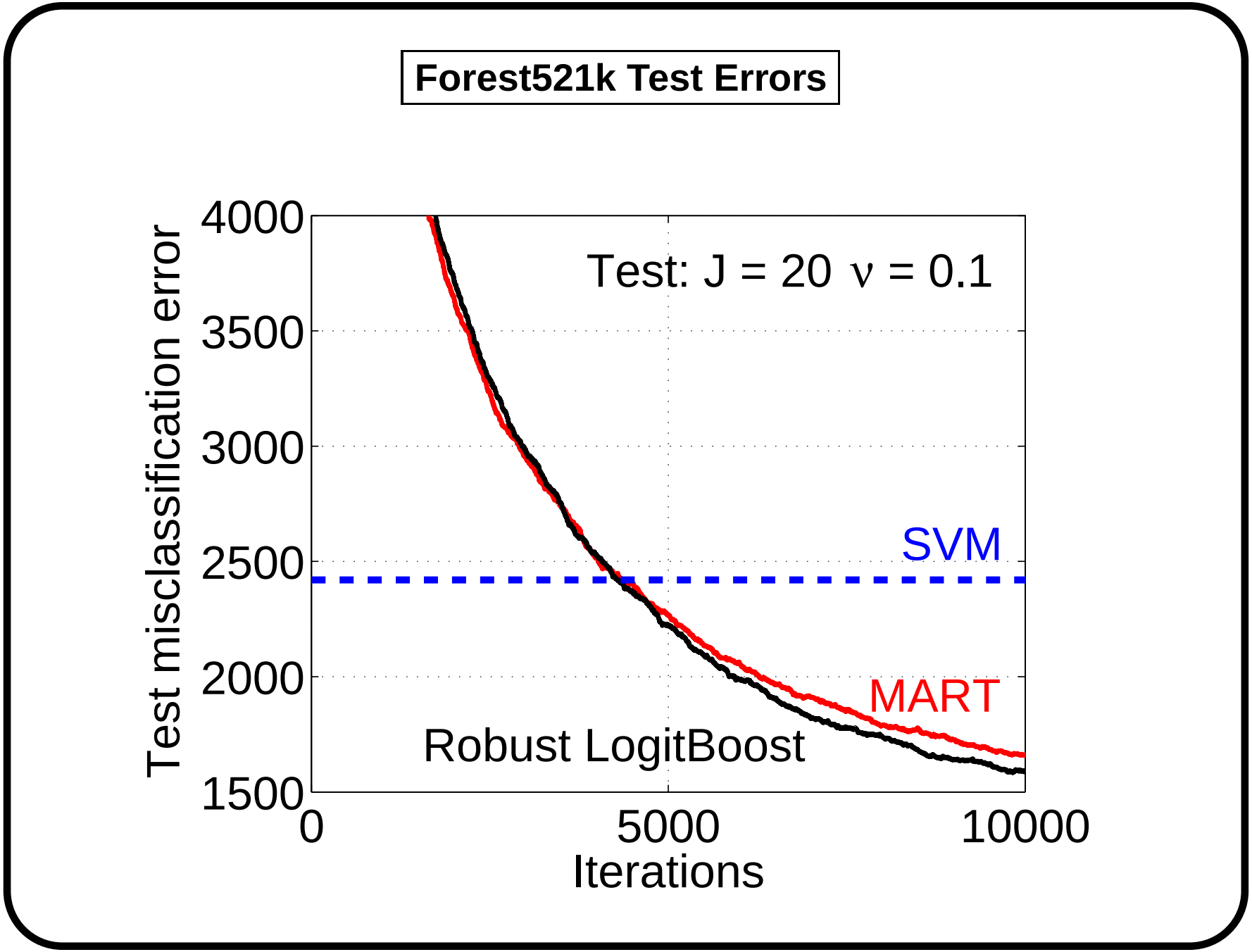
Forest521k: 521012 training samples, 50000 test samples

The two largest datasets from Bordes et al. JMLR 2005,

Fast Kernel Classifiers with Online and Active Learning


IJCNN1 Test Errors
Test misclassification error
2000
1800
1600
1400
1200
1000
0
5000
10000
Iterations
Test: J = 20 ν = 0.1
MART
LibSVM
Robust LogitBoost


Forest100k Test Errors
Test misclassification error
5000
4500
4000
3500
3000
0
5000
10000
Iterations
Test: J = 20 ν = 0.1
SVM
MART
Robust LogitBoost


Forest521k Test Errors
Test misclassification error
4000
3500
3000
2500
2000
1500
0
5000
10000
Iterations
Test: J = 20 ν = 0.1
SVM
MART
Robust LogitBoost

ABC-Boost for Multi-Class Classification

ABC = Adaptive Base Class

ABC-MART = ABC-Boost + MART

ABC-LogitBoost = ABC-Boost + (Robust) LogitBoost

The key to the success of ABC-Boost is the use of “better” derivatives.

Review Components of Logistic Regression

The multinomial logit probability model:

$$p_k = \frac{e^{F_k}}{\sum_{s=0}^{K-1} e^{F_s}}, \qquad \sum_{k=0}^{K-1} p_k = 1$$

where $F_k = F_k(\mathbf{x})$ is the function to be learned from the data.

The sum-to-zero constraint:

$$\sum_{k=0}^{K-1} F_k(\mathbf{x}) = 0$$

is commonly used to obtain a unique solution (only $K - 1$ degrees of freedom).

Why the sum-to-zero constraint?

$$\frac{e^{F_{i,k}+C}}{\sum_{s=0}^{K-1} e^{F_{i,s}+C}} = \frac{e^C e^{F_{i,k}}}{e^C \sum_{s=0}^{K-1} e^{F_{i,s}}} = \frac{e^{F_{i,k}}}{\sum_{s=0}^{K-1} e^{F_{i,s}}} = p_{i,k}.$$

For identifiability, one should impose a constraint.

One popular choice is to assume $\sum_{k=0}^{K-1} F_{i,k} = const$, equivalent to

$$\sum_{k=0}^{K-1} F_{i,k} = 0.$$

This is the assumption used in many papers including LogitBoost and MART.

The negative log-Likelihood loss

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left\{ - \sum_{k=0}^{K-1} r_{i,k} \log p_{i,k} \right\}$$

$$r_{i,k} = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{otherwise} \end{cases} \qquad \sum_{k=0}^{K-1} r_{i,k} = 1$$

Derivatives used in LogitBoost and MART:

$$\frac{\partial L_i}{\partial F_{i,k}} = -\left(r_{i,k} - p_{i,k}\right)$$

$$\frac{\partial^2 L_i}{\partial F_{i,k}^2} = p_{i,k}\left(1 - p_{i,k}\right),$$

which could be derived without imposing any constraints on F_k.

Derivatives Under Sum-to-zero Constraint

The loss function:

$$L_i = -\sum_{k=0}^{K-1} r_{i,k} \log p_{i,k}$$

The probability model and sum-to-zero constraint:

$$p_{i,k} = \frac{e^{F_{i,k}}}{\sum_{s=0}^{K-1} e^{F_{i,s}}}, \qquad \sum_{k=0}^{K-1} F_{i,k} = 0$$

Without loss of generality, we assume $k = 0$ is the base class

$$\boxed{F_{i,0} = -\sum_{i=1}^{K-1} F_{i,k}}$$

New derivatives:

$$\frac{\partial L_i}{\partial F_{i,k}} = (r_{i,0} - p_{i,0}) - (r_{i,k} - p_{i,k}),$$

$$\frac{\partial^2 L_i}{\partial F_{i,k}^2} = p_{i,0}(1 - p_{i,0}) + p_{i,k}(1 - p_{i,k}) + 2p_{i,0}p_{i,k}.$$

MART and LogitBoost used:

$$\frac{\partial L_i}{\partial F_{i,k}} = -(r_{i,k} - p_{i,k}), \qquad \frac{\partial^2 L_i}{\partial F_{i,k}^2} = p_{i,k}(1 - p_{i,k}).$$

Datasets

- **UCI-Covertype** Total 581012 samples.

  Two datasets were generated: **Covertype290k**, **Covertype145k**

- **UCI-Poker** Original 25010 training samples and 1 million test samples.

  Poker25kT1, **Poker25kT2**, **Poker525k**, **Poker275k**, **Poker150k**, **Poker100k**.

- **MNIST** Originally 60000 training samples and 10000 test samples.

  MNIST10k swapped the training with test samples.

- **Many variations of MNIST** Original MNIST is a well-known easy problem. (`www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/DeepVsShallowComparisonICML2007`) created a variety of much more difficult datasets by adding various background (correlated) noise, background images, rotations, etc.

- **UCI-Letter** Total 20000 samples.

dataset	K	# training	# test	# features
Covertype290k	7	290506	290506	54
Covertype145k	7	145253	290506	54
Poker525k	10	525010	500000	25
Poker275k	10	275010	500000	25
Poker150k	10	150010	500000	25
Poker100k	10	100010	500000	25
Poker25kT1	10	25010	500000	25
Poker25kT2	10	25010	500000	25
Mnist10k	10	10000	60000	784
M-Basic	10	12000	50000	784
M-Rotate	10	12000	50000	784
M-Image	10	12000	50000	784
M-Rand	10	12000	50000	784
M-RotImg	10	12000	50000	784
M-Noise1	10	10000	2000	784
M-Noise2	10	10000	2000	784
M-Noise3	10	10000	2000	784
M-Noise4	10	10000	2000	784
M-Noise5	10	10000	2000	784
M-Noise6	10	10000	2000	784
Letter15k	26	15000	5000	16
Letter4k	26	4000	16000	16
Letter2k	26	2000	18000	16

Summary of test mis-classification errors

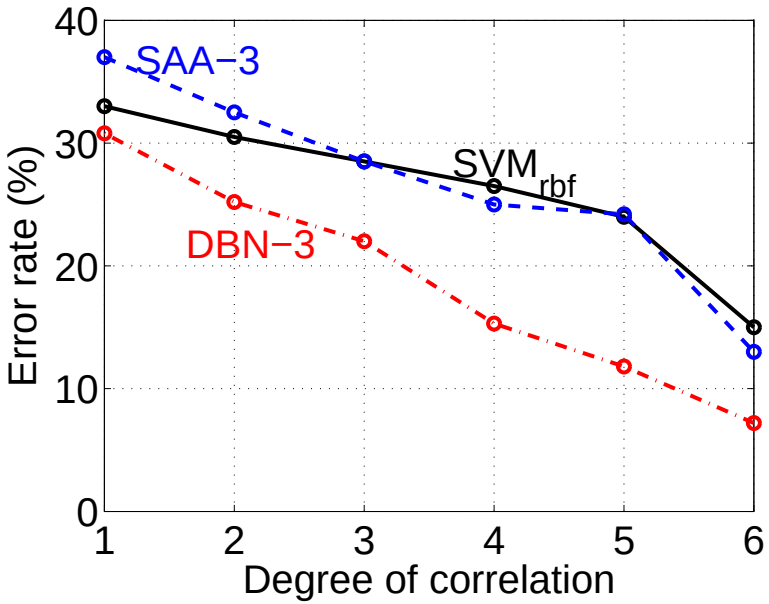
Dataset	**mart**	**abc-mart**	**logitboost**	**abc-logitboost**	logistic regression	# test
Covertype290k	11350	10454	10765	9727	80233	290506
Covertype145k	15767	14665	14928	13986	80314	290506
Poker525k	7061	2424	2704	1736	248892	500000
Poker275k	15404	3679	6533	2727	248892	500000
Poker150k	22289	12340	16163	5104	248892	500000
Poker100k	27871	21293	25715	13707	248892	500000
Poker25kT1	43575	34879	46789	37345	250110	500000
Poker25kT2	42935	34326	46600	36731	249056	500000
Mnist10k	2815	2440	2381	2102	13950	60000
M-Basic	2058	1843	1723	1602	10993	50000
M-Rotate	7674	6634	6813	5959	26584	50000
M-Image	5821	4727	4703	4268	19353	50000
M-Rand	6577	5300	5020	4725	18189	50000
M-RotImg	24912	23072	22962	22343	33216	50000
M-Noise1	305	245	267	234	935	2000
M-Noise2	325	262	270	237	940	2000
M-Noise3	310	264	277	238	954	2000
M-Noise4	308	243	256	238	933	2000
M-Noise5	294	244	242	227	867	2000
M-Noise6	279	224	226	201	788	2000
Letter15k	155	125	139	109	1130	5000
Letter4k	1370	1149	1252	1055	3712	16000
Letter2k	2482	2220	2309	2034	4381	18000

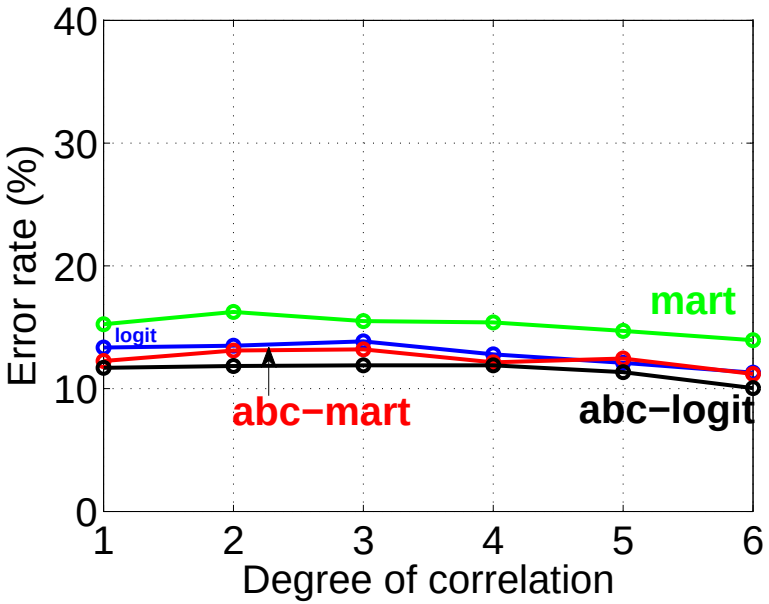
Comparisons with SVM and Deep Learning

Datasets: M-Noise1 to M-Noise6

Results on SVM, Neural Nets, and Deep Learning are from

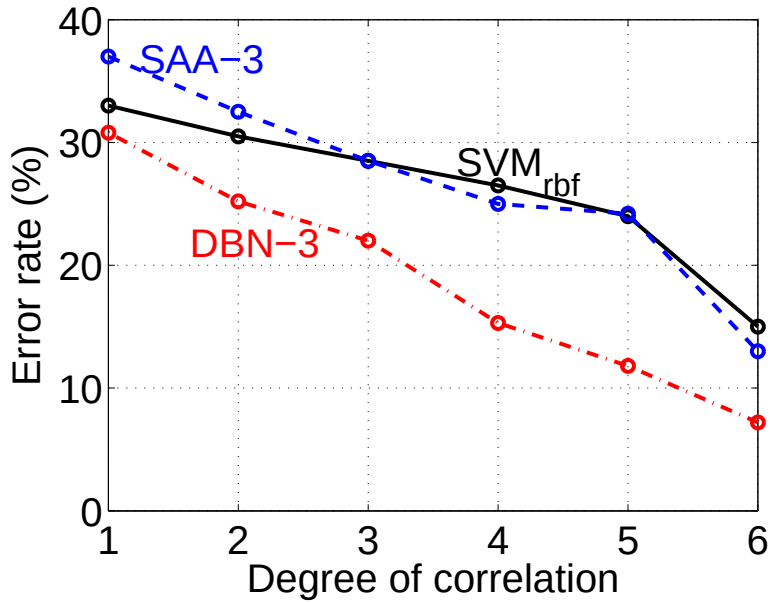
`www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/DeepVsShallowComparisonICML2007`

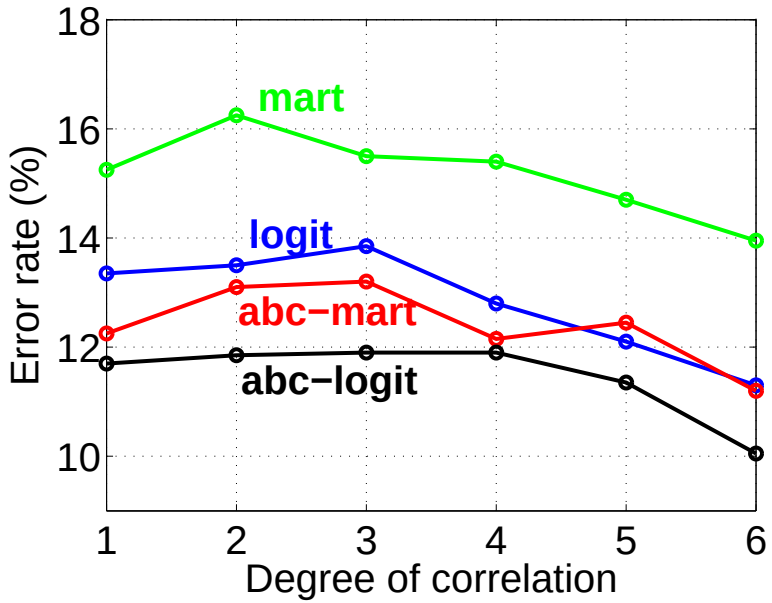




Comparisons with SVM and Deep Learning

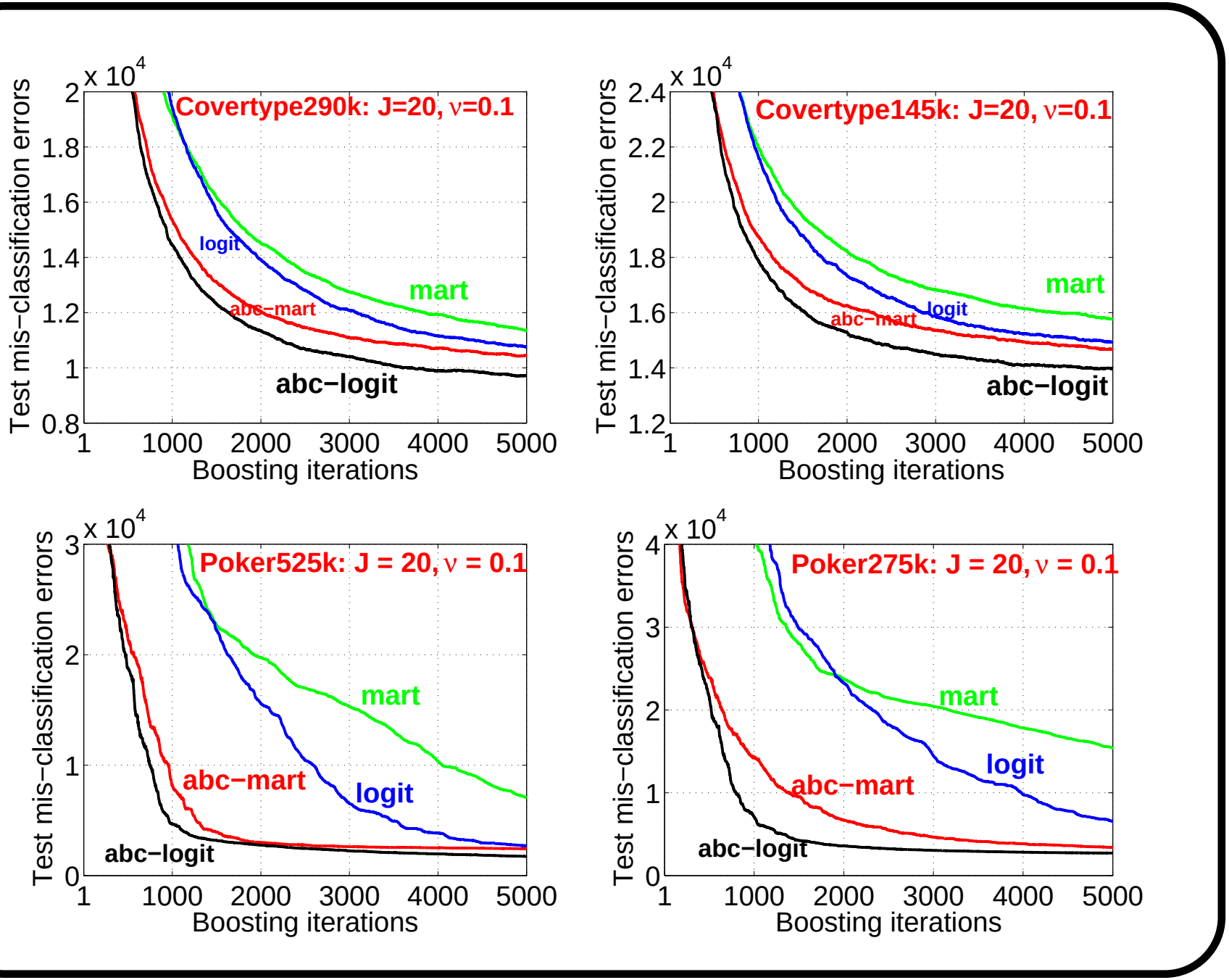
Datasets: M-Noise1 to M-Noise6





More Comparisons with SVM and Deep Learning

	M-Basic	M-Rotate	M-Image	M-Rand	M-RotImg
SVM-RBF	**3.05%**	11.11%	22.61%	14.58%	55.18%
SVM-POLY	3.69%	15.42%	24.01%	16.62%	56.41%
NNET	4.69%	18.11%	27.41%	20.04%	62.16%
DBN-3	3.11%	**10.30%**	16.31%	**6.73%**	47.39%
SAA-3	3.46%	**10.30%**	23.00%	11.28%	51.93%
DBN-1	3.94%	14.69%	16.15%	9.80%	52.21%
mart	4.12%	15.35%	11.64%	13.15%	49.82%
abc-mart	3.69%	13.27%	9.45%	10.60%	46.14%
logitboost	3.45%	13.63%	9.41%	10.04%	45.92%
abc-logitboost	3.20%	11.92%	**8.54%**	9.45%	**44.69%**


Covertype290k: J=20, ν=0.1
Test mis−classification errors
x 10^4
Boosting iterations
logit
mart
abc−mart
abc−logit
Covertype145k: J=20, ν=0.1
Poker525k: J = 20, ν = 0.1
Poker275k: J = 20, ν = 0.1

Extending Multi-Class to Multi-Label Learning

Multi-Class Learning: Suppose $y_i = k$,

$$Lik \propto p_{i,0}^0 \times ... \times p_{i,k}{}^1 \times ... \times p_{i,K-1}^0 = p_{i,k}$$

Multi-Label Learning: Suppose $y_i \in S_i = \{0, k\}$,

$$Lik \propto p_{i,0}^1 \times ... \times p_{i,k}{}^1 \times ... \times p_{i,K-1}^0 = p_{i,0} p_{i,k}$$

There are actually more than one ways to determine the weights. For example, we can choose the following loss function:

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left\{ - \sum_{k=0}^{K-1} w_{i,k} \log p_{i,k} \right\}, \quad w_{i,k} = \begin{cases} 1/|S_i| & \text{if } y_i \in S_i \\ 0 & \text{otherwise} \end{cases}$$

Combining Multi-Label Model with Boosting and Trees

- We need to modify the existing boosting algorithms (MART, LogitBoost, ABC-MART, ABC-LogitBoost) to incorporate the new models.

- For each example, the algorithm will again output a vector of class probabilities. We need to a criterion to truncate the list to assign class labels.

- We need a good evaluation criterion to assess the quality of multi-label learning.

Evaluation Criteria

Using our model and boosting, we learn the set of class probabilities for each example and sort them in descending order:

$$\hat{p}_{i,(0)} \geq \hat{p}_{i,(1)} \geq \cdots \geq \hat{p}_{i,(K-1)}$$

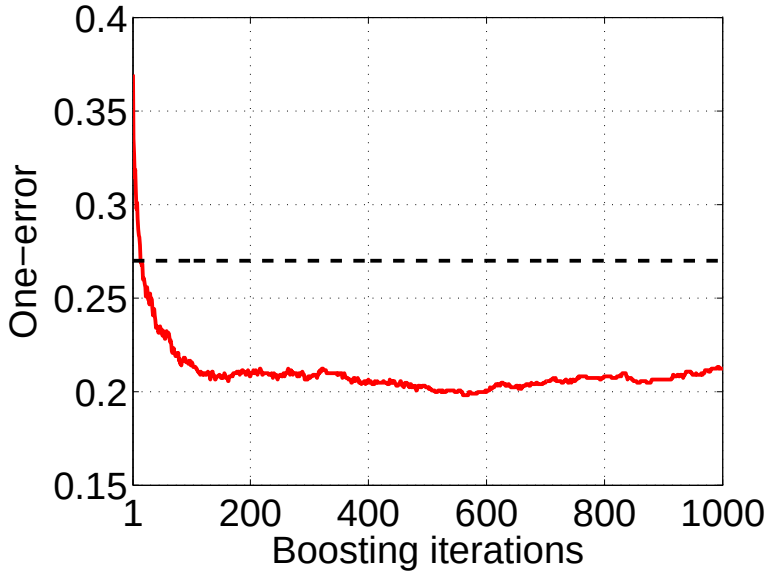
We consider three criteria:

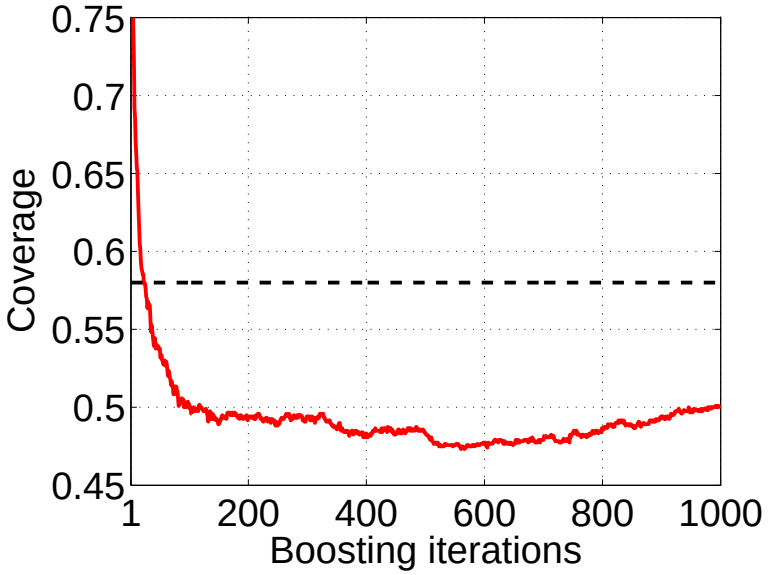
- **One-error**: How many times the top-ranked label is not in the true labels.
- **Coverage**: How far one needs, on average, to go down the list of labels in order to cover all the ground truth labels.
- **Precision**: A more comprehensive ranking measure borrowed from information retrieval (IR) literature.

Experiments and Comparisons

We implemented our method with MART (other implementations are forthcoming).

We compared our result with an existing publication on the same dataset.





Our method with boosting and trees (red curves) is substantially better than published results (dashed horizontal line).

Our precision is about 87% but the other paper did not report.

Ongoing Work

- Test our (and others') multi-label algorithms on Census data.
- Experiment with various multi-label probability models.
- Implement (Robust) LogitBoost for multi-label learning
- Implement ABC-MART and ABC-LogitBoost for multi-label learning