Batch and On-line Spam Filter Comparison

Gordon V. Cormack

University of Waterloo

Andrej Bratko

Jozef Stefan Institute

On-line vs Off-line Evaluation

TREC – Text Retrieval Conference (On-line)

- chronological order, immediate feedback
- real email messages (and filters!)
- soft classification: *spamminess score*
- Receiver Operating Characteristic (ROC)

Classical Evaluation (Batch)

- *k*-fold cross validation
- contrived email messages (and filters!)
- hard classification: *spam* or *ham*
- accuracy, weighted accuracy, Total Cost Ratio (TCR)

Test Methods and Corpora

- TREC 2005 Public Corpus
  - on-line test (TREC methodology)
  - 10-fold cross validation (random splits)
  - 9:1 chronological split
    - on-line test sequence
    - batch test set
  - tokenized, obfuscated versions of same corpora
- Ling Spam & PU1 Corpora
  - 10-fold cross validation
  - splits, tokenization, obfuscation defined by corpora

Subject Filtering Methods

X^2 *(Graham/Robinson)*

- Bogofilter *(Relson, Louis et al.)*

Support Vector Machine *(Vapnik)*

- SVM^{light} *(Joachims)*

Logistic Regression *(Fisher)*

- LR-TRIRLS *(Komarek)*

Prediction by Partial Matching *(Cleary & Witten)*

- Adaptive PPM-D Classifier *(Bratko)*

Dynamic Markov Modeling *(Cormack & Horspool)*

- Adaptive DMC Classifier *(Cormack)*

Prediction by Partial Matching

For each class:

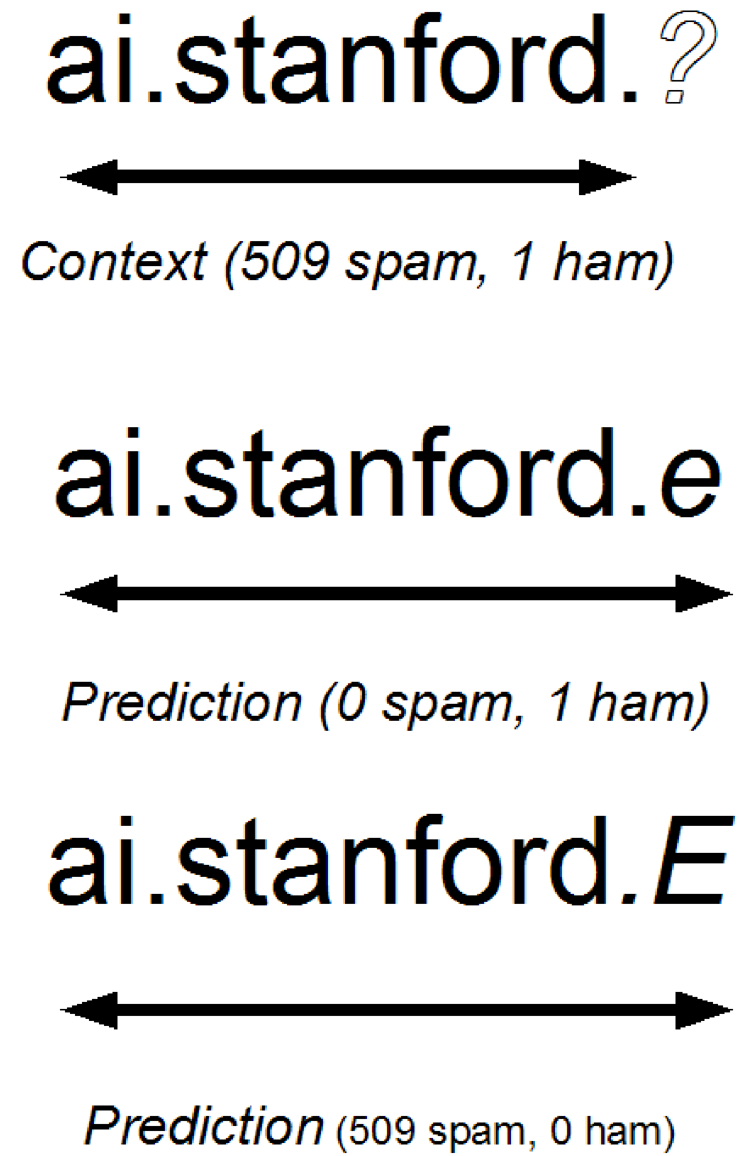
- left context occurrences
- left context+prediction
- *log-likelihood* estimate
  - compressed length

Smoothing/backoff:

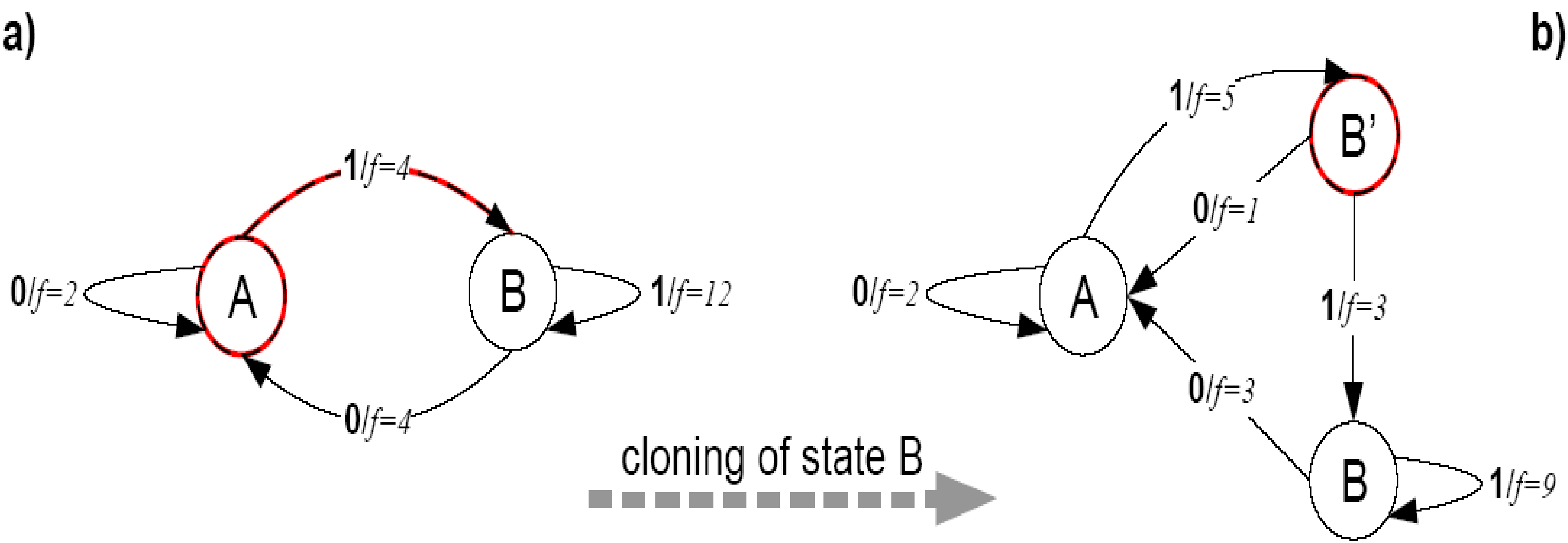
- zero occurrence problem

Adaptation:

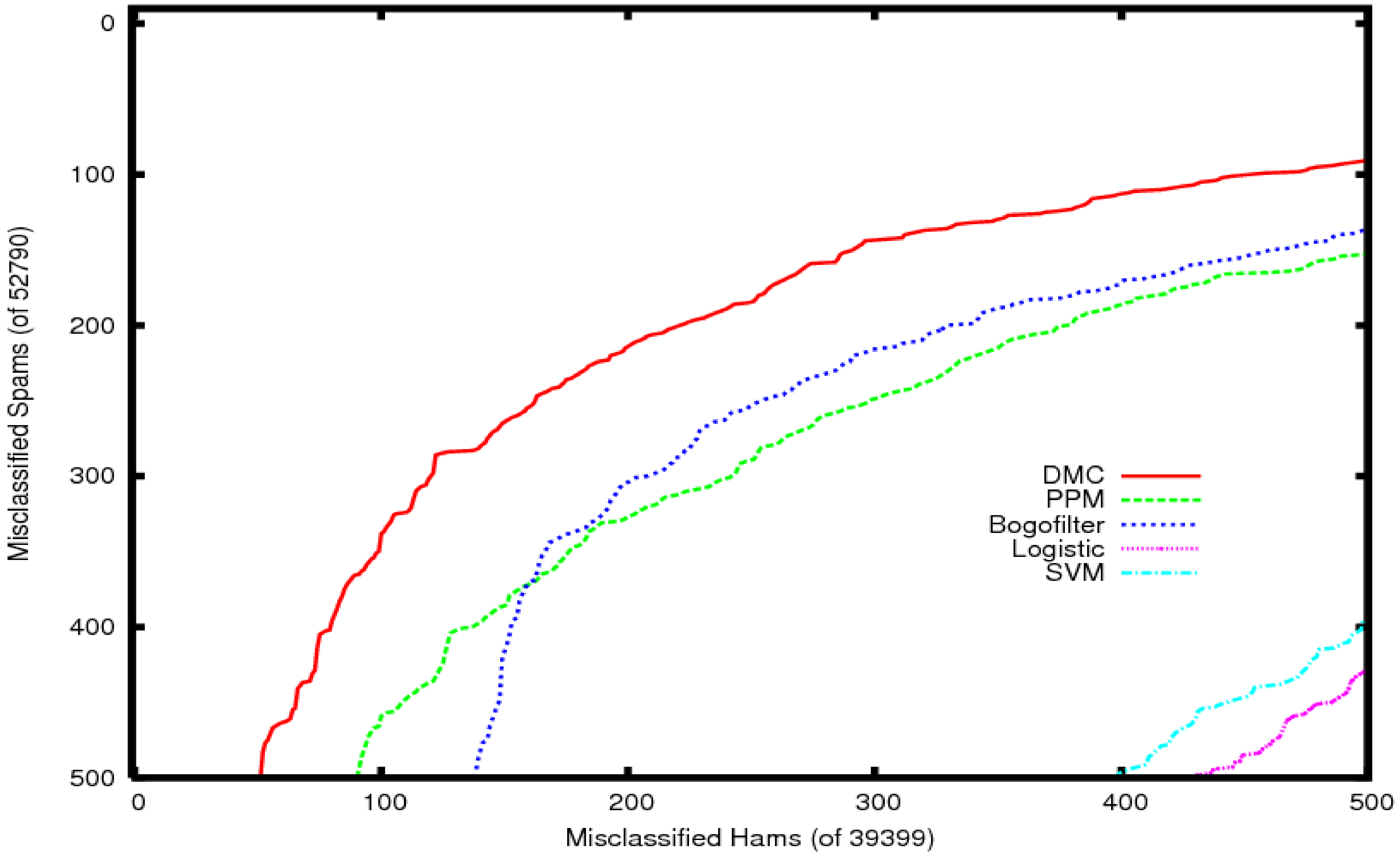
- increment counts assuming in-class



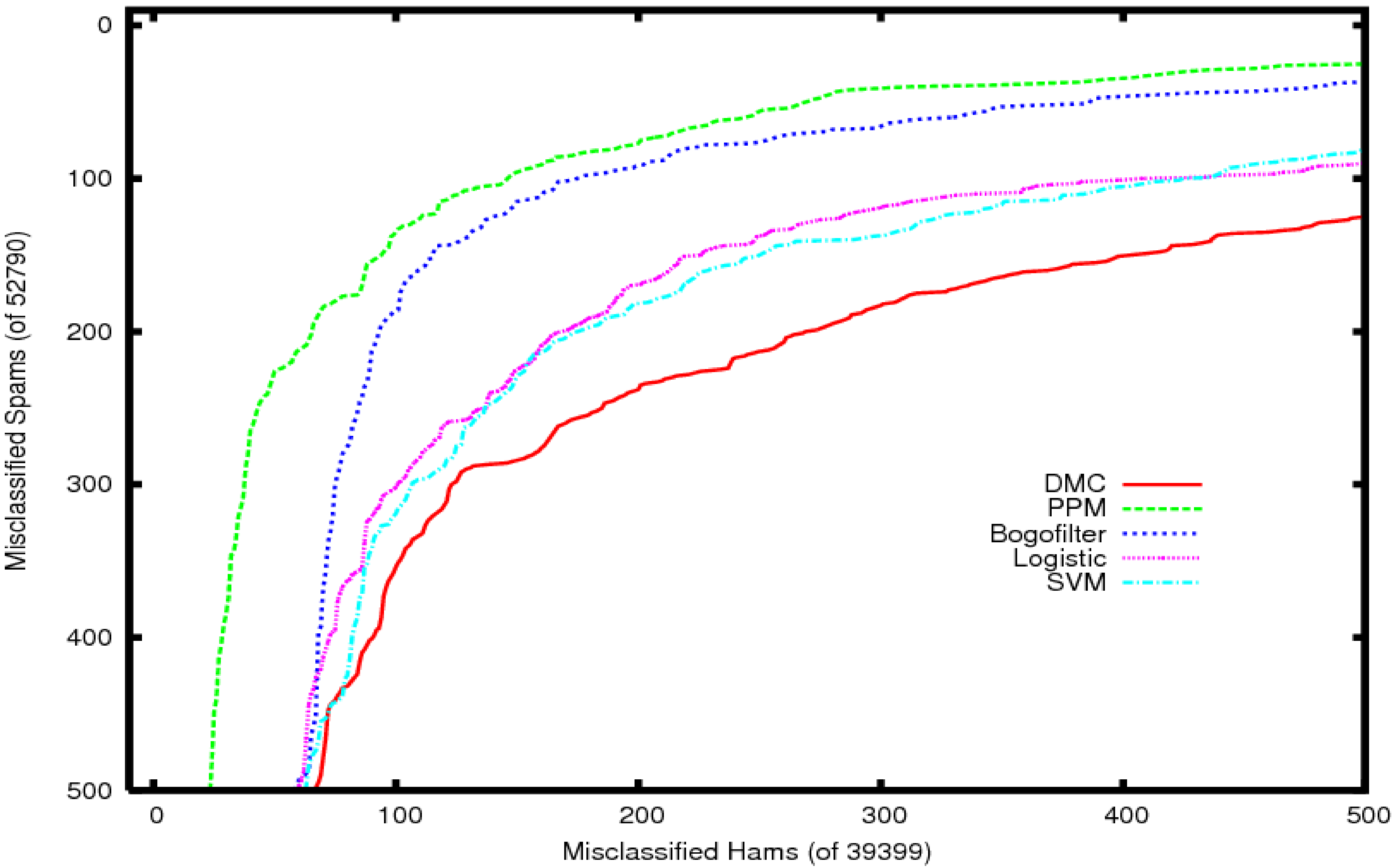
DMC State Cloning



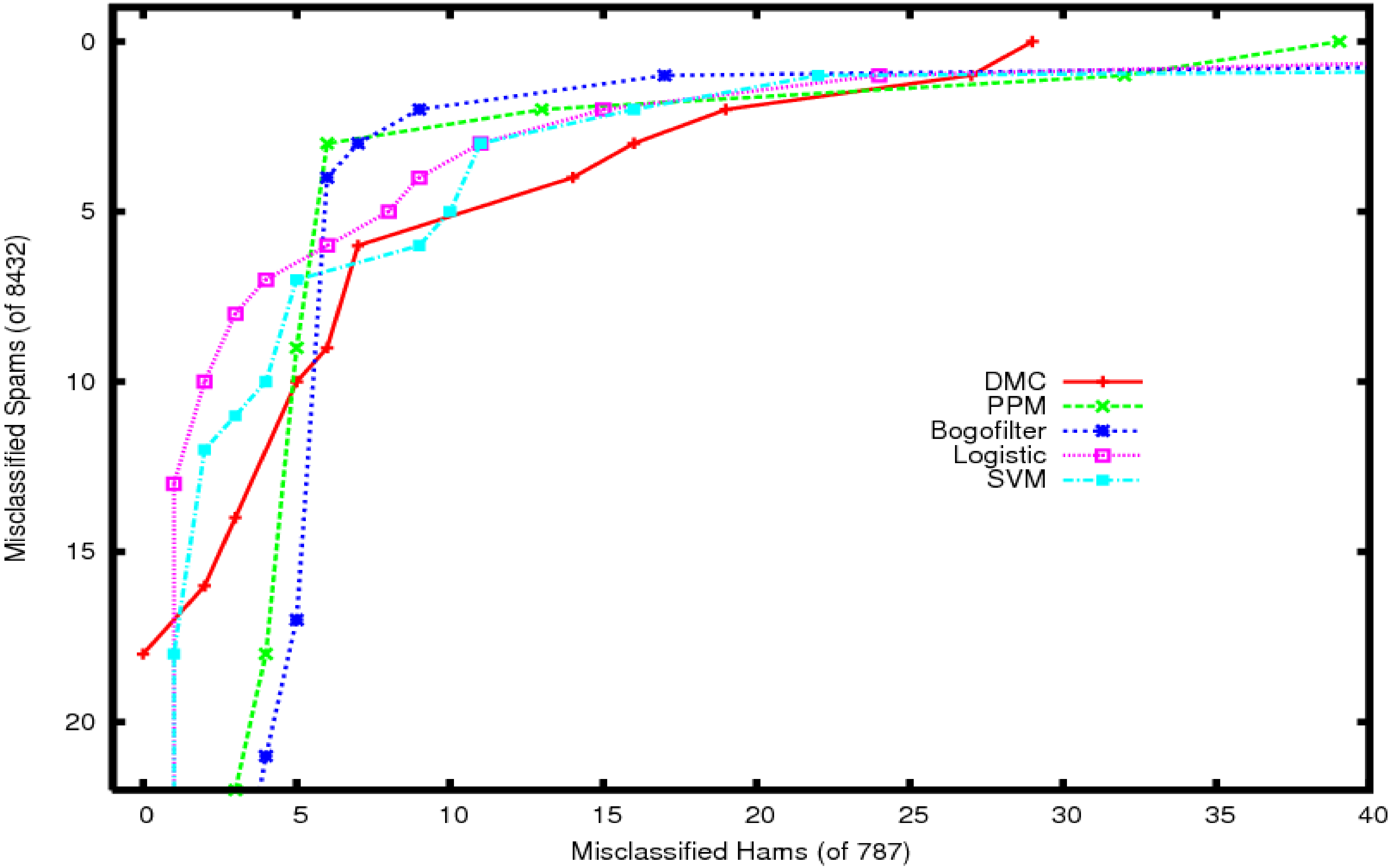
TREC Corpus, On-line



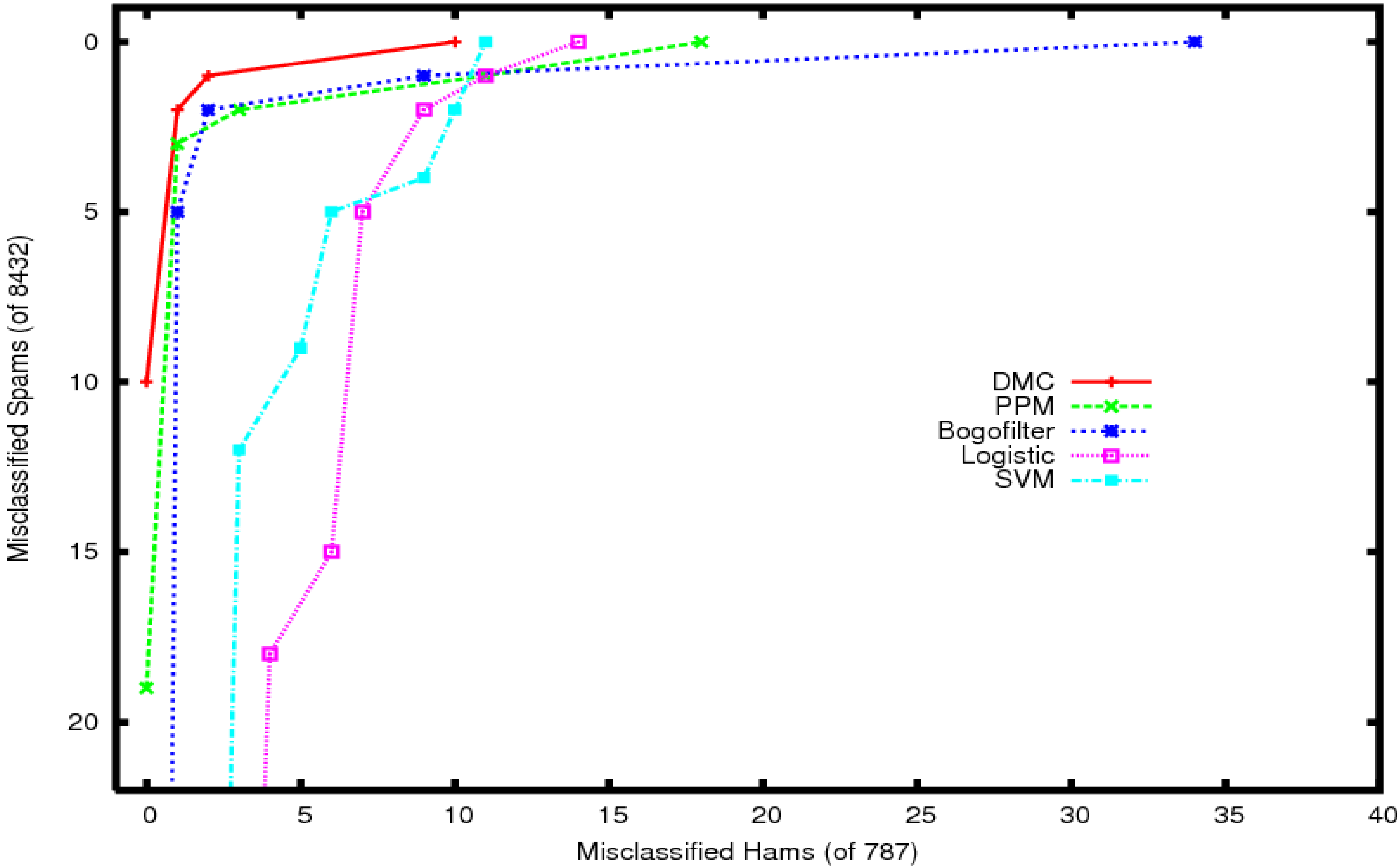
10-Fold Cross Validation



9:1 Chronological, Batch



9:1 Chronological, On-line



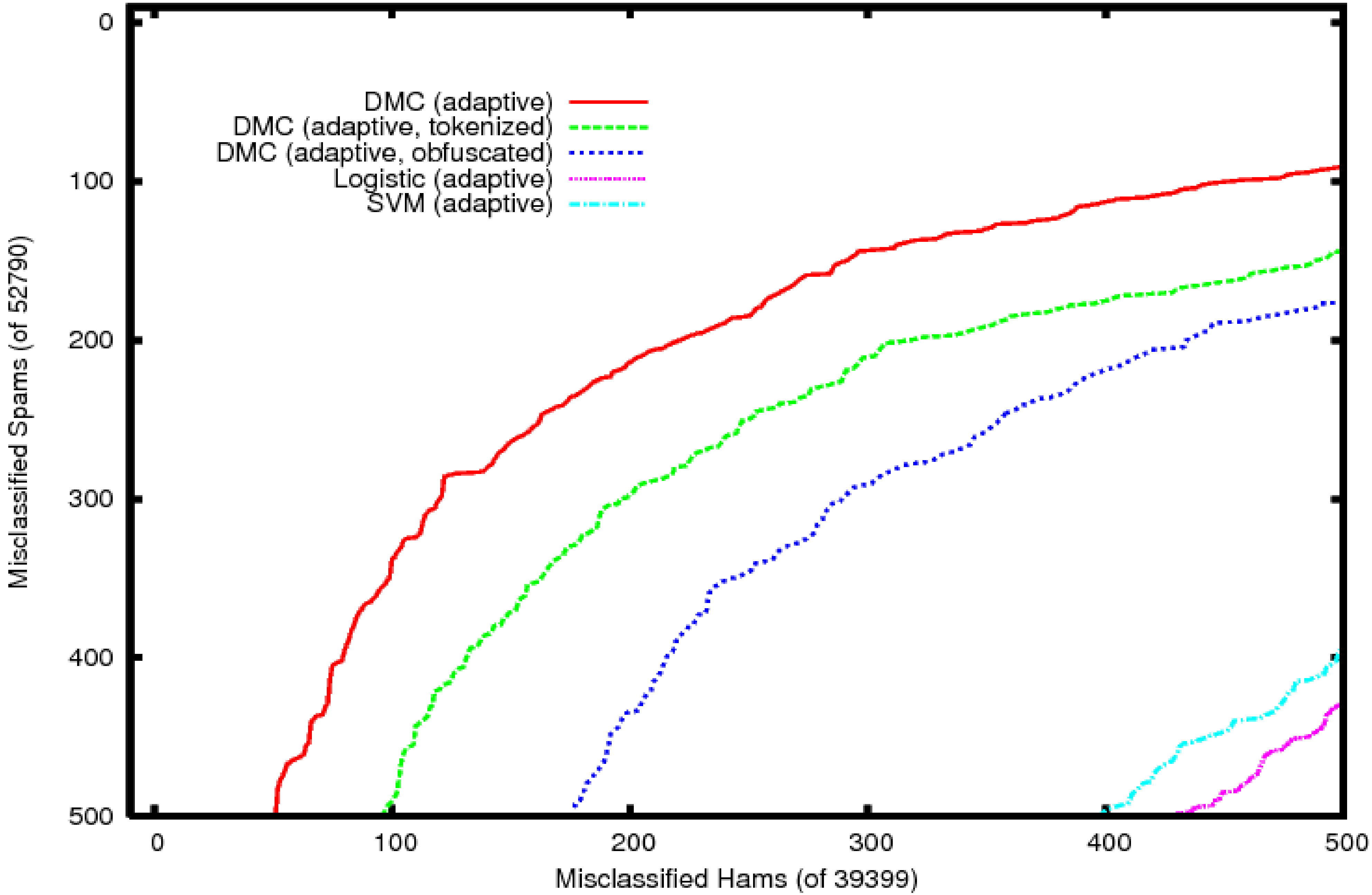
Batch, On-line (1-ROCA)%

	On-line		Batch	
Method	Full Corpus	9:1 Chronological	10-fold C.V.	9:1 Chronological
DMC	**.013 (.010-.018)**	**.0003 (.0000-.003)**	.015 (.012-.018)	**.003 (.001-.006)**
PPM	.017 (.014-.021)	.0007 (.0001-.005)	**.006 (.004-.009)**	.003 (.001-.008)
Bogofilter	.048 (.038-.062)	.002 (.0001-.041)	.020 (.012 - .033)	.009 (.003-.029)
LR	.068 (.058-.079)	.020 (.003-.135)	.016 (.012-.021)	.12 (.001-10.1)
SVM	.075 (.064-.088)	.007 (.0015-.033)	.021 (.015-.029)	.13 (.003-5.6)

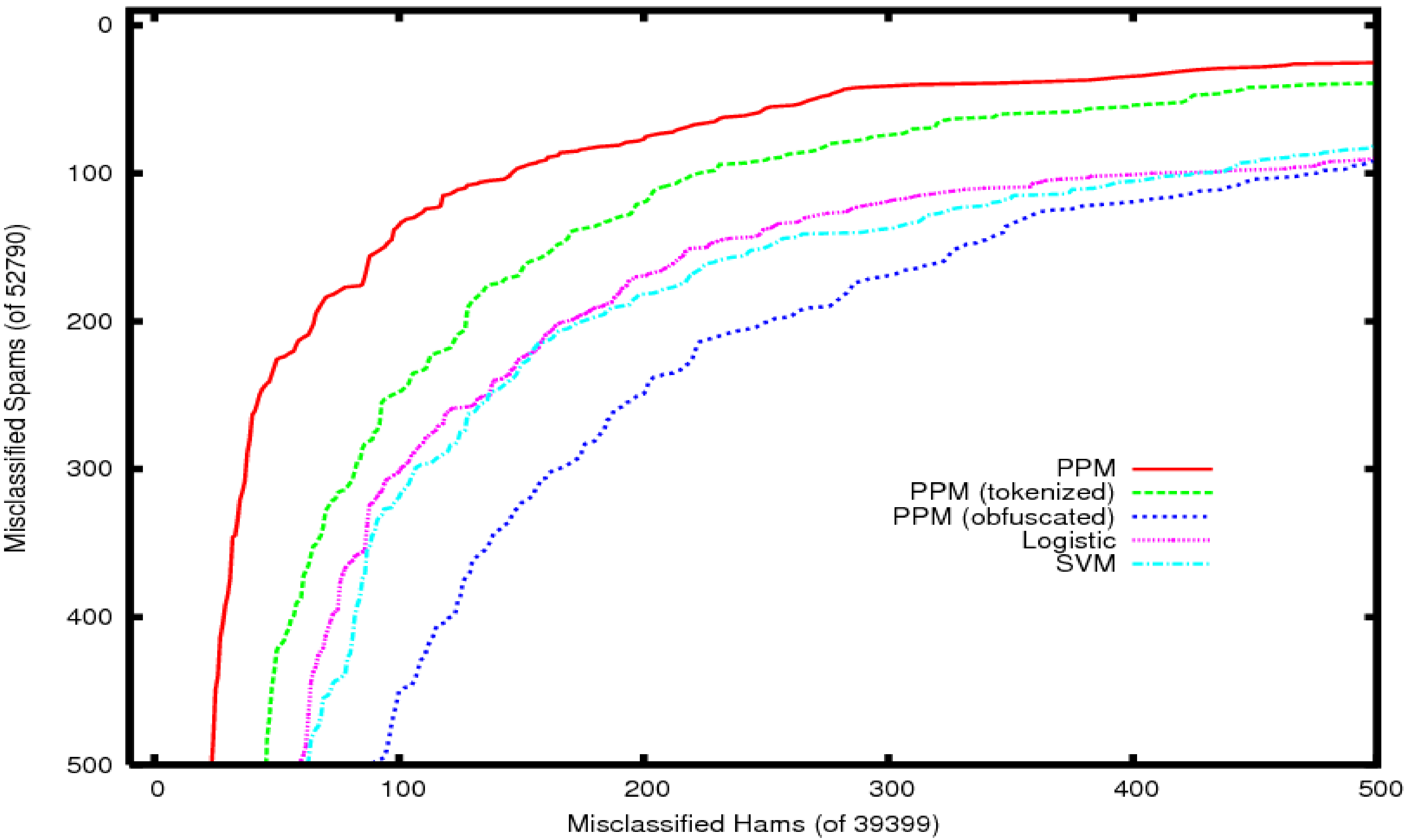
Effect of Order/Adaptation

Filter	Training Regimen	Testing Regimen: On-line Random Order	On-line Corpus Order	Batch
DMC	Random Order	.01 (.006-.017)	.007 (.004-.011)	.009 (.006-.015)
DMC	Corpus Order	.035 (.026-.047)	.037 (.024-.057)	.31 (.25-.37)
PPM	Batch	.0052 (.003-.01)	.0053 (.003-.009)	.0055 (.003-.01)

Tokenization, On-line



Tokenization, Batch



Tokenization/Obfuscation

Method	On-line Full Corpus	On-line 9:1 Chronological	Batch 10-fold C.V.	Batch 9:1 Chronological
DMC	*.013 (.010-.018)*	*.0003 (.0000-.003)*	*.015 (.012-.018)*	*.003 (.001-.006)*
tokenized	.025 (.020-.032)	.0006 (.0001-.006)	.025 (.019-.033)	.001 (.000-.013)
obfuscated	.037 (.030-.045)	.0004 (.0000-.0042)	.029 (.023-.037)	.002 (.001-.006)
PPM	.017 (.014-.021)	*.0007 (.0001-.005)*	*.006 (.004-.009)*	*.003 (.001-.008)*
tokenized	.038 (.033-.045)	.0016 (.0003-.009)	.012 (.009-.016)	.005 (.002-.012)
obfuscated	.075 (.066-.084)	.0046 (.0016-.013)	.020 (.014-.027)	.015 (.006-.035)
Bogofilter	*.048 (.038-.062)*	*.002 (.0001-.041)*	*.020 (.012 - .033)*	*.009 (.003-.029)*
obfuscated	.13 (.11-.15)	.024 (.004-.14)	.055 (.045-.068)	.036 (.012-.11)

Other Methods

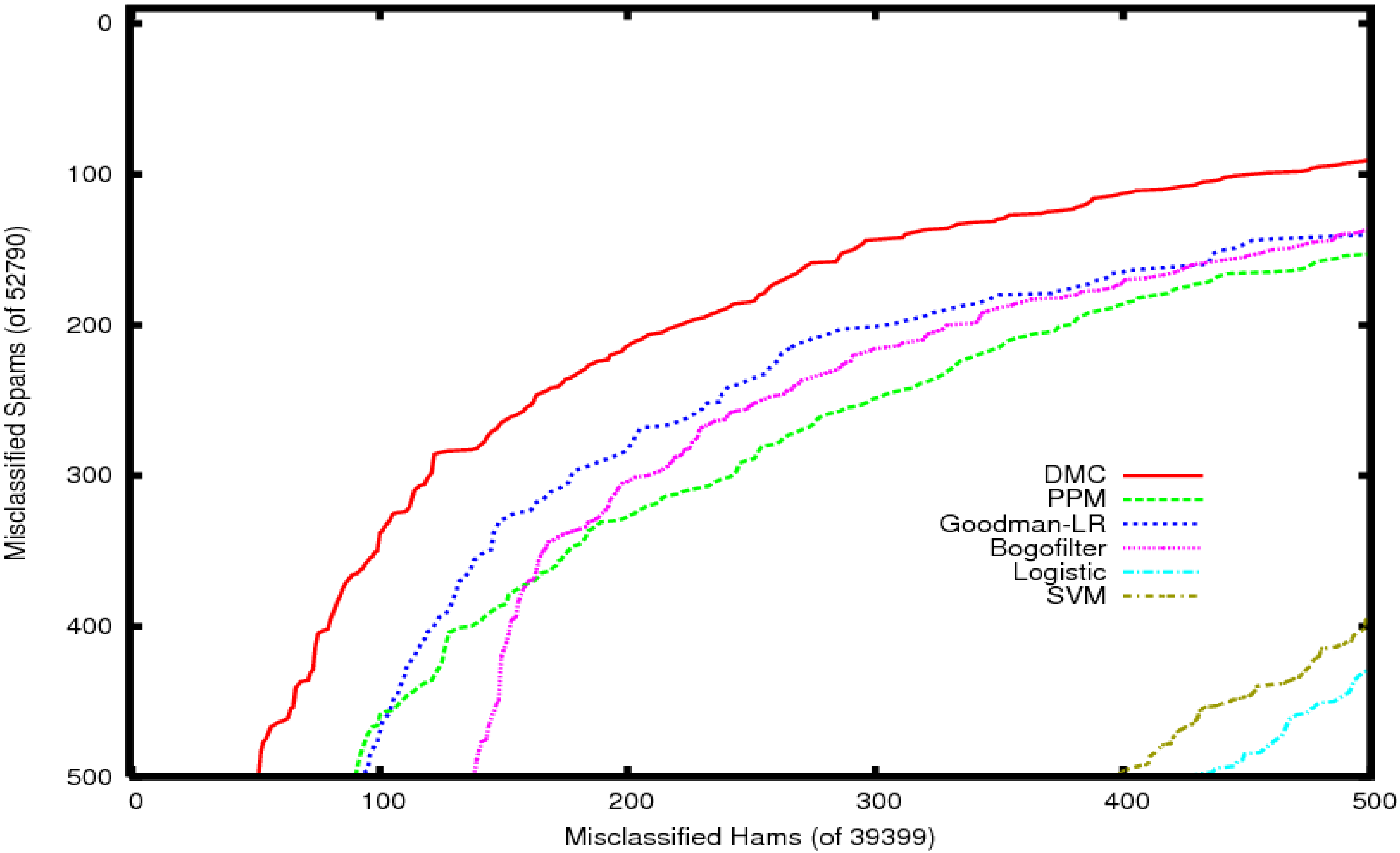
Gradient Descent Logistic Regression *(Goodman)*

On-line Filter Fusion *(Lynam & Cormack)*

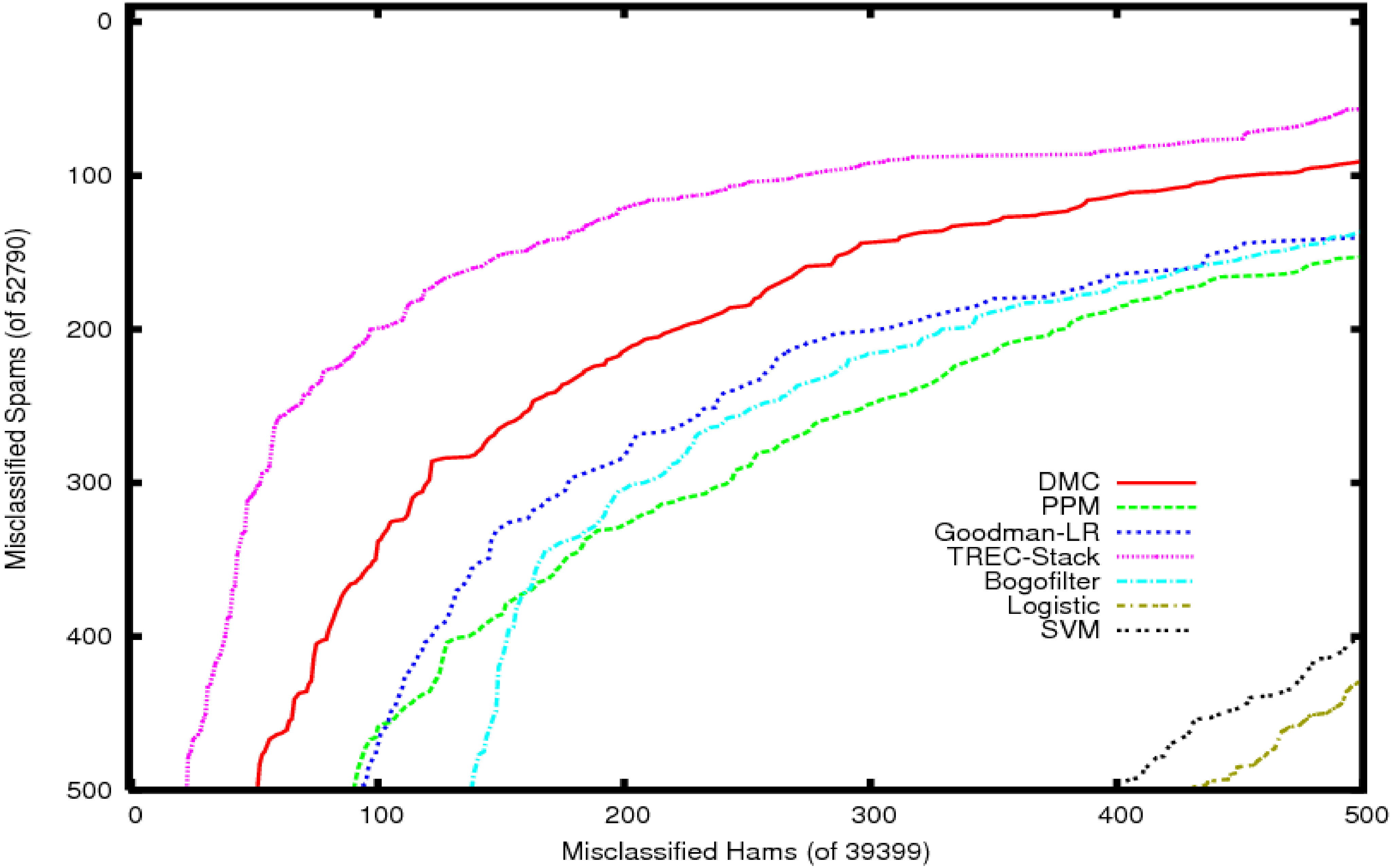
Classical machine learning *(Cast of thousands)*

- Naïve Bayes (which naïve Bayes?)
- kNN
- Perceptron
- Winnow
- Decision trees
- Boosting
- Stacking

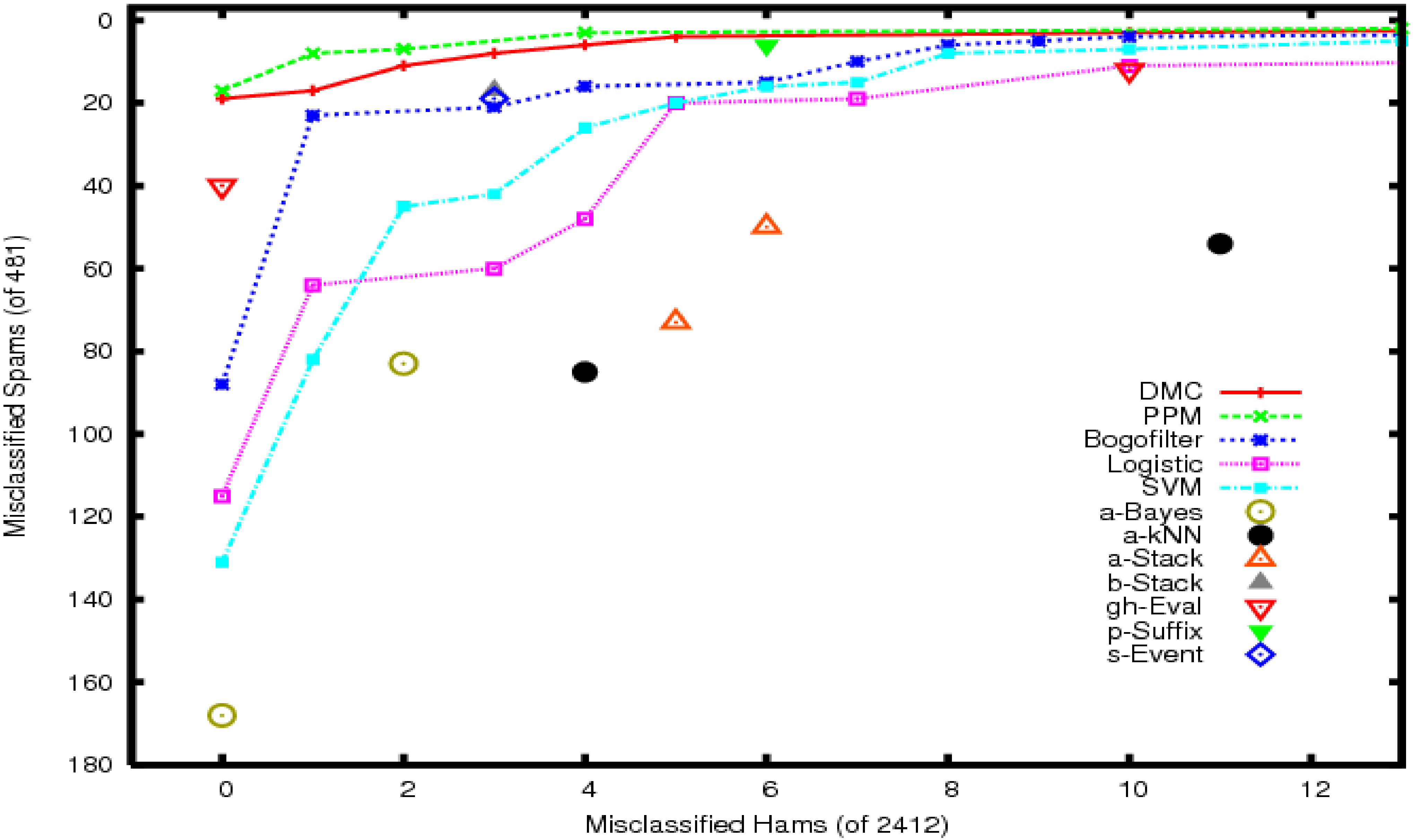
Goodman's Gradient Descent LR



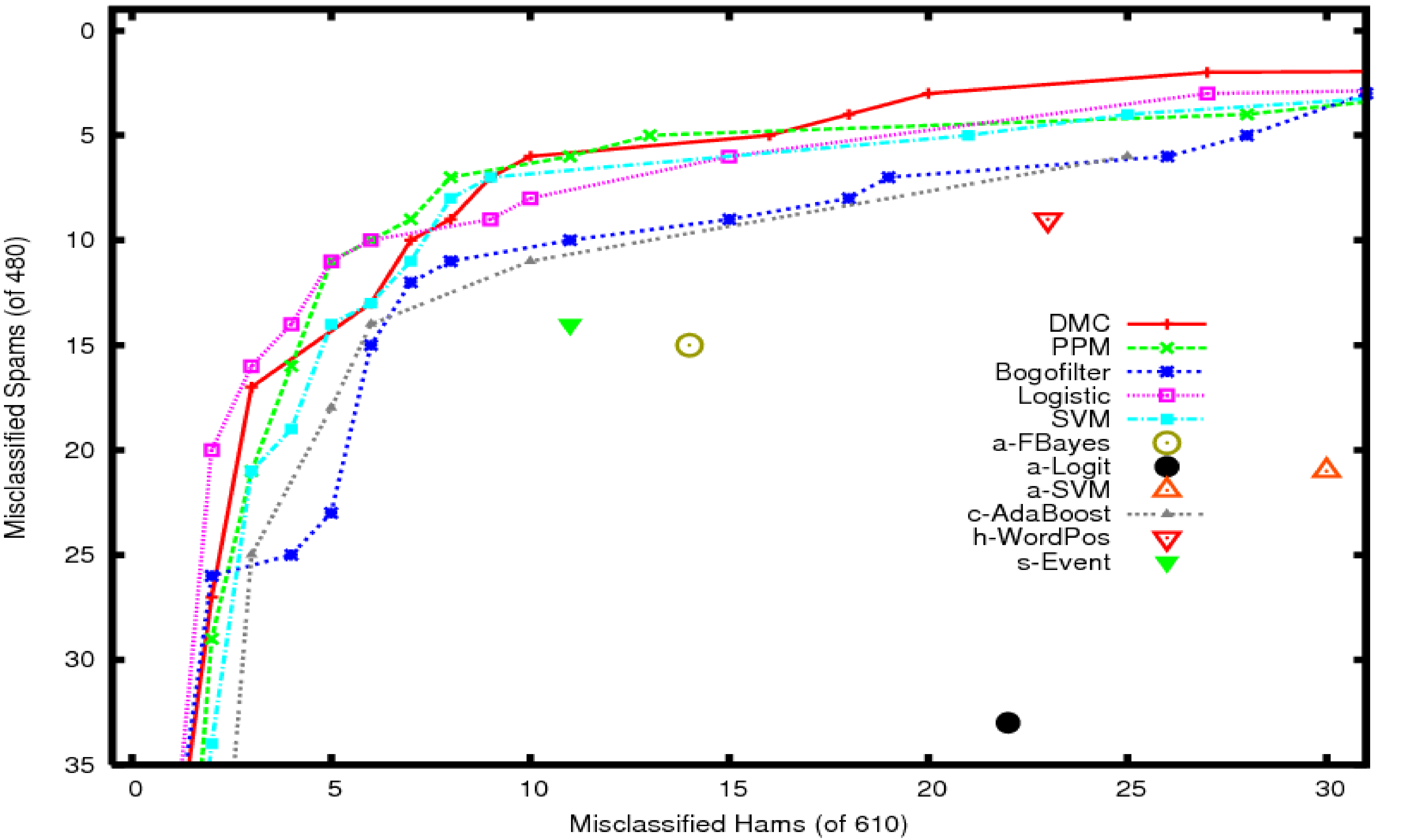
Stacking – 53 TREC Filters



Ling Spam Corpus



PU1 Corpus



Misclassified Spams (of 480)
Misclassified Hams (of 610)
0
5
10
15
20
25
30
35
0
5
10
15
20
25
30
DMC
PPM
Bogofilter
Logistic
SVM
a-FBayes
a-Logit
a-SVM
c-AdaBoost
h-WordPos
s-Event

Conclusions

- Batch and on-line are different
  - good filters can be adapted to do both well
- Feature engineering is important
  - email is not just a bag or sequence of tokens
- Real filters beat contrived ones
  - even on contrived corpora
- PPM and DMC effectively filter spam
  - fast (100s of messages/sec)
  - voracious appetite for RAM (0.5 – 2.0 GB)