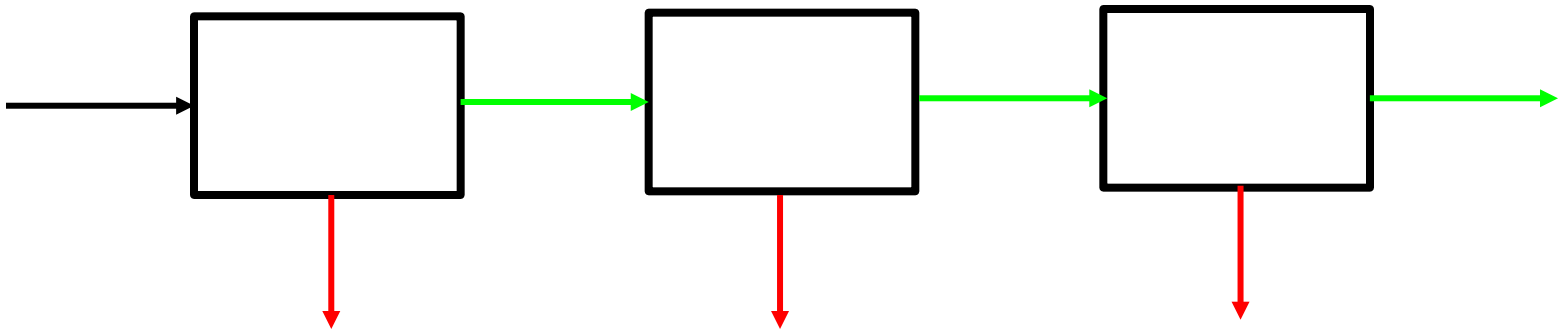


Designing Efficient Cascaded Classifiers: Tradeoff between Accuracy and Cost

Vikas Raykar, **Balaji Krishnapuram,** Shipeng Yu

Siemens Healthcare

KDD 2010

Strategies for minimizing different types of costs

- Traditional classifier design minimizes misclassification costs
- Costs associated with collecting labels minimized by active label acquisition and semi-supervised learning
- Feature acquisition is not as well explored
  - There is some preliminary work on active acquisition: but can be slow at run time, not yet widely used (Foster, Prem, Yu)
  - Reinforcement learning & multi-arm bandit: not successful in practice yet (Shihao Ji, many others)
  - Cascaded architectures are the dominant solution till now (Viola-Jones face detection cascades using Adaboost)

Reducing Feature cost: Needs

Acquisition cost can be either

- Computational | fast detectors
- Financial | expensive medical tests
- Human discomfort | biopsy

Requirements:

- Features are acquired on demand.
- A set of features can be acquired as a group.
- Each feature group incurs a certain cost.
- Need knob to tradeoff accuracy vs cost

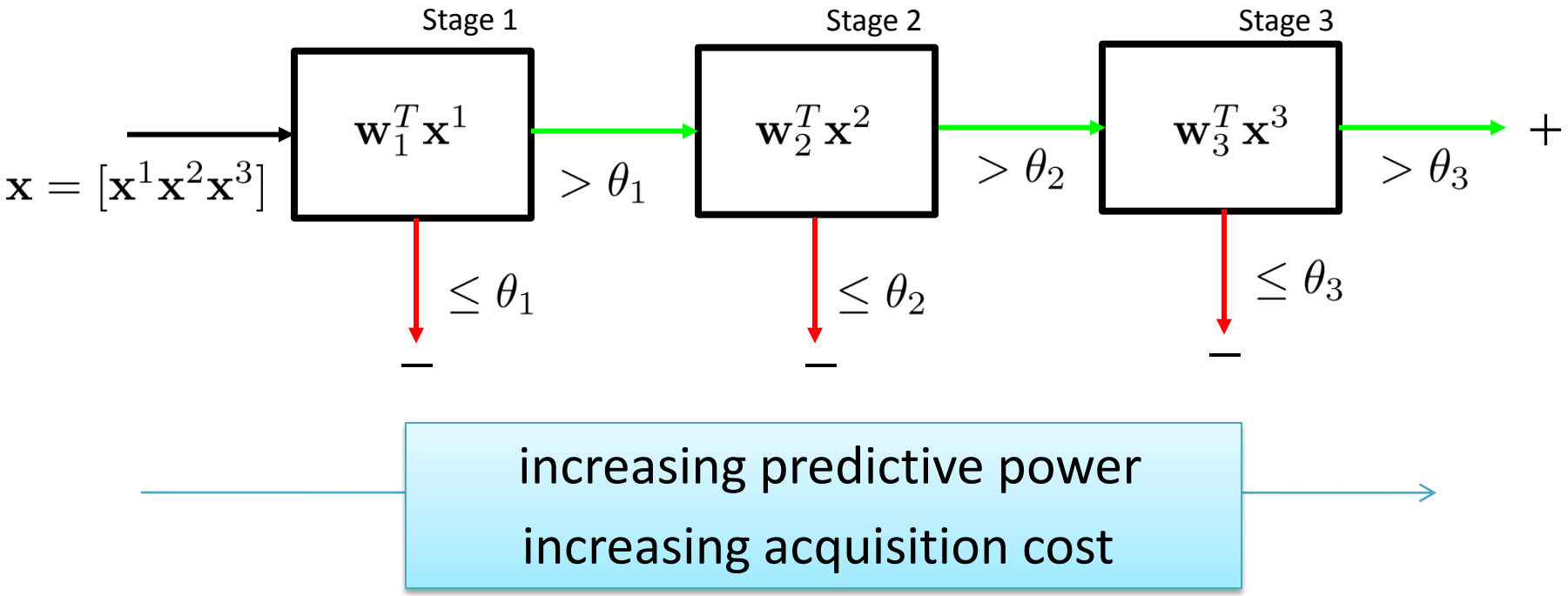
Example: Survival Prediction for Lung Cancer

- 2-year survival prediction for lung cancer patients treated with chemo/radiotherapy

	Feature Group	Number of features	examples	Cost
1	clinical features	9	gender, age	0 no cost
2	features before therapy	8	lung function creatinine clearance	1
3	imaging /treatment features	7	gross tumor volume treatment dose	2
4	blood bio-markers	21	Interleukin-8 Osteopontin	5 expensive

increasing predictive power … increasing acquisition cost

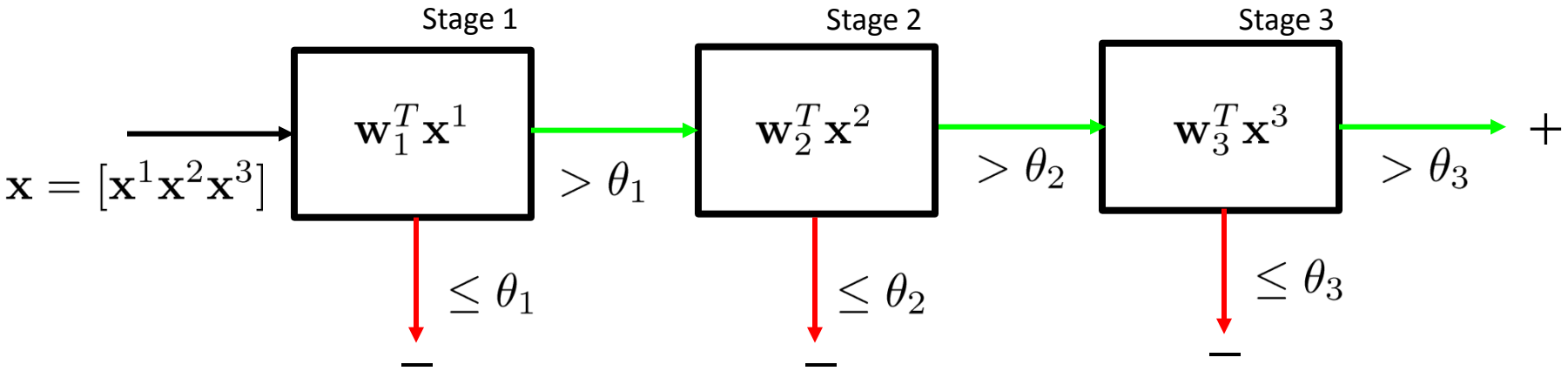
A cascade of linear classifiers



- Training each stage of the cascade
- Choosing the thresholds for each stage

Sequential Training of cascades

- Conventionally each stage is trained using only examples that pass through all the previous stages.
- Training depends on the choice of the thresholds.
- For each choice of threshold we have to retrain.



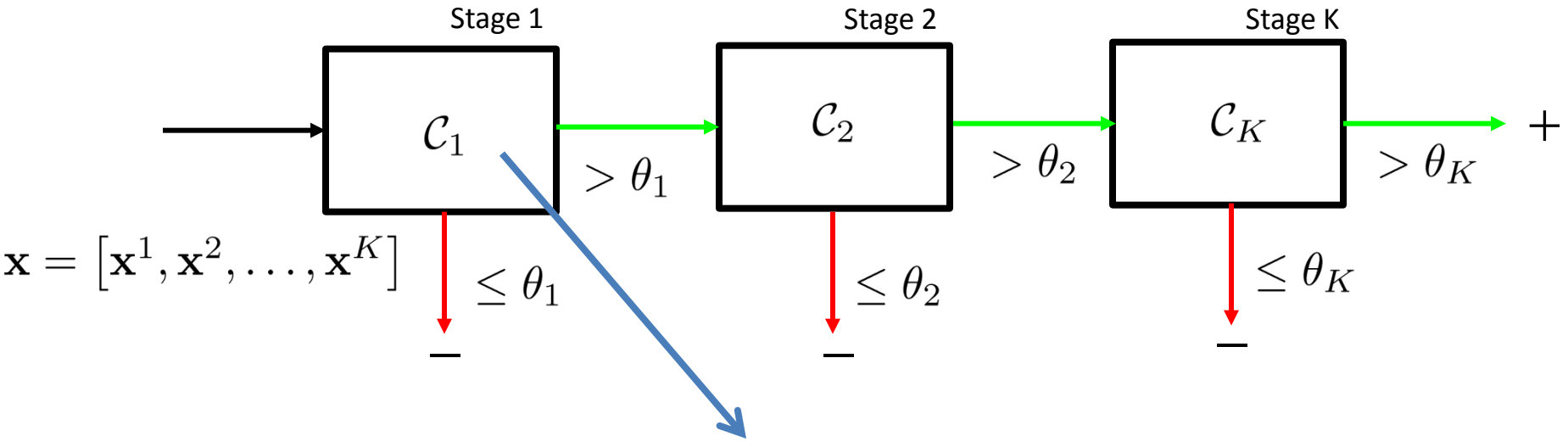
Contributions of this paper

- Joint training of all stages of the cascade.
  - Notion of probabilistic soft cascades
- A knob to control the tradeoff between accuracy vs cost
  - Modeling the expected feature cost
- Decoupling the classifier training and threshold selection.
  - Post-selection of thresholds

Notation

t_j estimate of the cost it takes to acquire/compute $\mathbf{x}^j$

K stage cascade $[\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K]$



Each stage is of the form $f_{\mathcal{C}_j}(\mathbf{x}) = \mathbf{w}_j^\top \mathbf{x}^j$

$y = 1$ if $f_{\mathcal{C}_j}(\mathbf{x}) > \theta_j$ and $y = 0$ if $f_{\mathcal{C}_j}(\mathbf{x}) \leq \theta_j$

Logistic Regression $p_{\mathcal{C}_j}(y = 1|\mathbf{x}, \mathbf{w}) = \sigma\left(f_{\mathcal{C}_j}(\mathbf{x})\right)$ $\sigma(z) = 1/(1 + e^{-z})$

Soft Cascade

- Probabilistic version of the hard cascade.
- An instance is classified as positive if **all the K stages** predict it as positive.

$$p(y=1|\mathbf{x},\mathbf{w}) = \prod_{j=1}^{K} \sigma\left(f_{C_j}(\mathbf{x})\right)$$

- An instance is classified as negative if **at least one** of the K classifiers predicts it as negative.

$$p(y=0|\mathbf{x},\mathbf{w}) = 1 - \prod_{j=1}^{K} \sigma\left(f_{C_j}(\mathbf{x})\right)$$

Some properties of soft cascades

- Sequential ordering of the cascade is not important.
- Can only deploy hard cascade (Order definitely matters during deployment)
- Primarily a mathematical idea to ease the training process.
- We use a maximum a-posteriori (MAP) estimate with Laplace prior on the weights.

Joint cascade training

- Once we have a probabilistic cascade we can write the log-likelihood.

$$\log p(\mathcal{D}|\mathbf{w}) = \sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log(1 - p_i)$$

$$p_i = \prod_{j=1}^{K} \sigma\left(f_{C_j}(\mathbf{x}_i)\right)$$

- We impose a Laplacian prior.

$$p(w_i|\gamma) = \frac{\sqrt{\gamma}}{2} \exp\left(-\sqrt{\gamma}|w_i|\right)$$

- Maximum a-posteriori (MAP) estimate

$$\widehat{\mathbf{w}}_{\mathrm{MAP}} = \arg\max_w L(\mathbf{w})$$

$$L(\mathbf{w}) = \left[\sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log(1 - p_i)\right] - \sqrt{\gamma}\|\mathbf{w}\|_1$$

Accuracy vs Cost

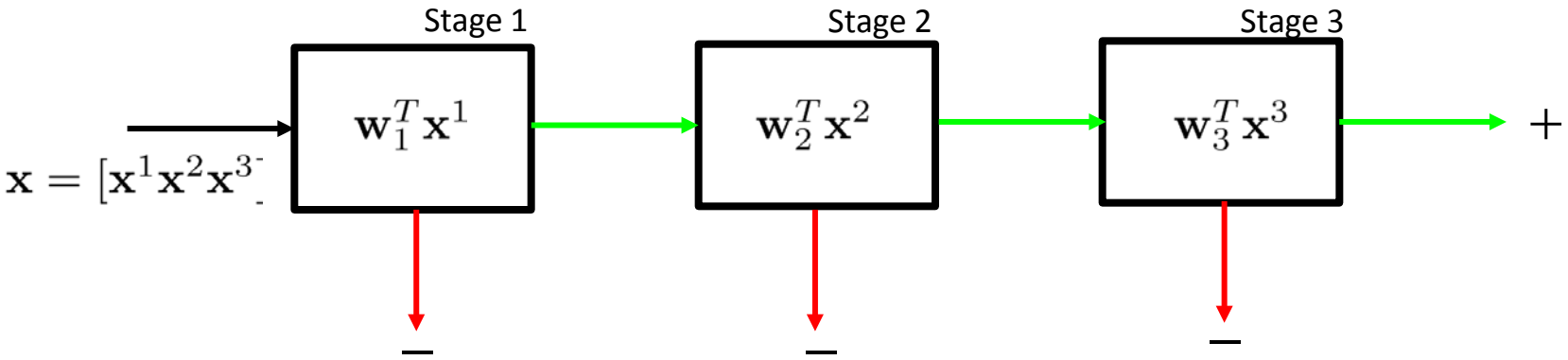
- We would like to find the MAP estimate subject to the constraints on the expected cost for a new instance

$$E_{p(\mathbf{x})}\left[\mathcal{T}(\mathbf{x})\right] \leq c$$

where $\mathcal{T}$ is the cost for a new instance $\mathbf{x}$

- The expectation is over the unknown test distribution.
- Since we do not know the test distribution we estimate this quantity based on the training set.

Modeling the expected cost



For a given instance	Cost
Stage 1	t_1
Stage 2	$t_2\sigma\left(f_{\mathcal{C}_1}(\mathbf{x}_i)\right)$
Stage 3	$t_3\sigma\left(f_{\mathcal{C}_1}(\mathbf{x}_i)\right)\sigma\left(f_{\mathcal{C}_2}(\mathbf{x}_i)\right)$

$$\mathcal{T}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[t_1 + \sum_{j=2}^{K} t_j \prod_{l=1}^{j-1}\sigma\left(f_{\mathcal{C}_l}(\mathbf{x}_i)\right)\right]$$

$$\widehat{\mathbf{w}}_{\text{MAP}} = \arg\max_w L(\mathbf{w}) - \beta\mathcal{T}(\mathbf{w})$$

β controls the tradeoff between accuracy and cost

We optimize using cyclic coordinate descent

Experiments

- Medical Datasets
  - Personalized medicine
    - Survival prediction for lung cancer
    - Tumor response prediction for rectal cancer
  - Computer aided diagnosis for lung cancer

Survival Prediction for Lung Cancer

- 2-year survival prediction for advanced non-small cell lung cancer (NSCLC) patients treated with chemo/radiotherapy.
- 82 patients treated at MAASTO clinic among which 24 survived two years

	Feature Group	Number of features	examples	Cost
1	clinical features	9	gender, age	0 no cost
2	features before therapy	8	lung function creatinine clearance	1
3	imaging /treatment features	7	gross tumor volume treatment dose	2
4	blood bio-markers	21	Interleukin-8 Osteopontin	5 expensive

C. Dehing-Oberije, D. De Ruysscher, and et al. Tumor volume combined with number of positive lymph node stations is a more important prognostic factor than TNM stage for survival of non-small-cell lung cancer patients treated with (chemo)radiotherapy. *Int J Radiat Oncol Biol Phys.*, 70(4):1039–1044, 2007.

Pathological Complete Response (pCR) Prediction for Rectal Cancer

- Predict tumor response after chemo/radiotherapy for locally advanced rectal cancer
- 78 patients (21 had pCR)

	Feature Group	Number of features	Cost
1	Clinical features	6	0
2	CT/PET scan features before treatment	2	1
3	CT/PET scan features after treatment	2	10

C. Capirci, L. Rampin, and et al. Sequential FDG-PET/CT reliably predicts response of locally advanced rectal cancer to neo-adjuvant chemo-radiation therapy. *Eur J Nucl Med Mol Imaging*, 34:1583–1593, 2007.

Methods compared

- Single stage classifier
- Proposed soft cascade
  - With beta = 0
  - Varying beta
- Sequential Training
  - Logistic Regression
  - AdaBoost [Viola-Jones cascade]
  - LDA

Evaluation Procedure

- 70 % for training 30 % for testing
- Area under the ROC Curve
- Normalized average cost per patient
  - Using all the features has a cost of 1
- Results averages over 10 repetitions
- Thresholds for each stage chosen using a two-level hierarchical grid search

Results

	Testing set AUC mean[± std]	Testing set Cost mean[± std]
Lung Cancer		
(1) Single stage classifier	0.79[± 0.12]	✓ 1.00[± 0.00]
(2) Proposed soft cascade $\beta = 0$	⋆ 0.72[± 0.11]	⋆ 0.37[± 0.08]
(3) Sequential Training via Logistic Regression	0.71[± 0.09]	✓ 0.45[± 0.08]
(4) Sequential Training via AdaBoost	✓ 0.63[± 0.05]	✓ 0.68[± 0.08]
(5) Sequential Training via LDA	0.70[± 0.03]	✓ 0.66[± 0.08]
(6) Proposed soft cascade $\beta = 10N$	0.73[± 0.12]	0.35[± 0.12]
(7) Proposed soft cascade $\beta = 100N$	0.70[± 0.11]	0.35[± 0.11]
(8) Proposed soft cascade $\beta = 1000N$	0.70[± 0.11]	✓ 0.27[± 0.10]
Rectum Cancer		
(1) Single stage classifier	0.83[± 0.06]	✓ 1.00[± 0.00]
(2) Proposed soft cascade $\beta = 0$	⋆ 0.79[± 0.06]	⋆ 0.59[± 0.09]
(3) Sequential Training via Logistic Regression	0.76[± 0.09]	✓ 0.70[± 0.10]
(4) Sequential Training via AdaBoost	0.73[± 0.10]	✓ 0.68[± 0.09]
(5) Sequential Training via LDA	✓ 0.71[± 0.09]	0.63[± 0.12]
(6) Proposed soft cascade $\beta = 10N$	0.79[± 0.06]	0.57[± 0.08]
(7) Proposed soft cascade $\beta = 100N$	0.77[± 0.04]	✓ 0.50[± 0.07]
(8) Proposed soft cascade $\beta = 1000N$	0.76[± 0.08]	✓ 0.48[± 0.08]

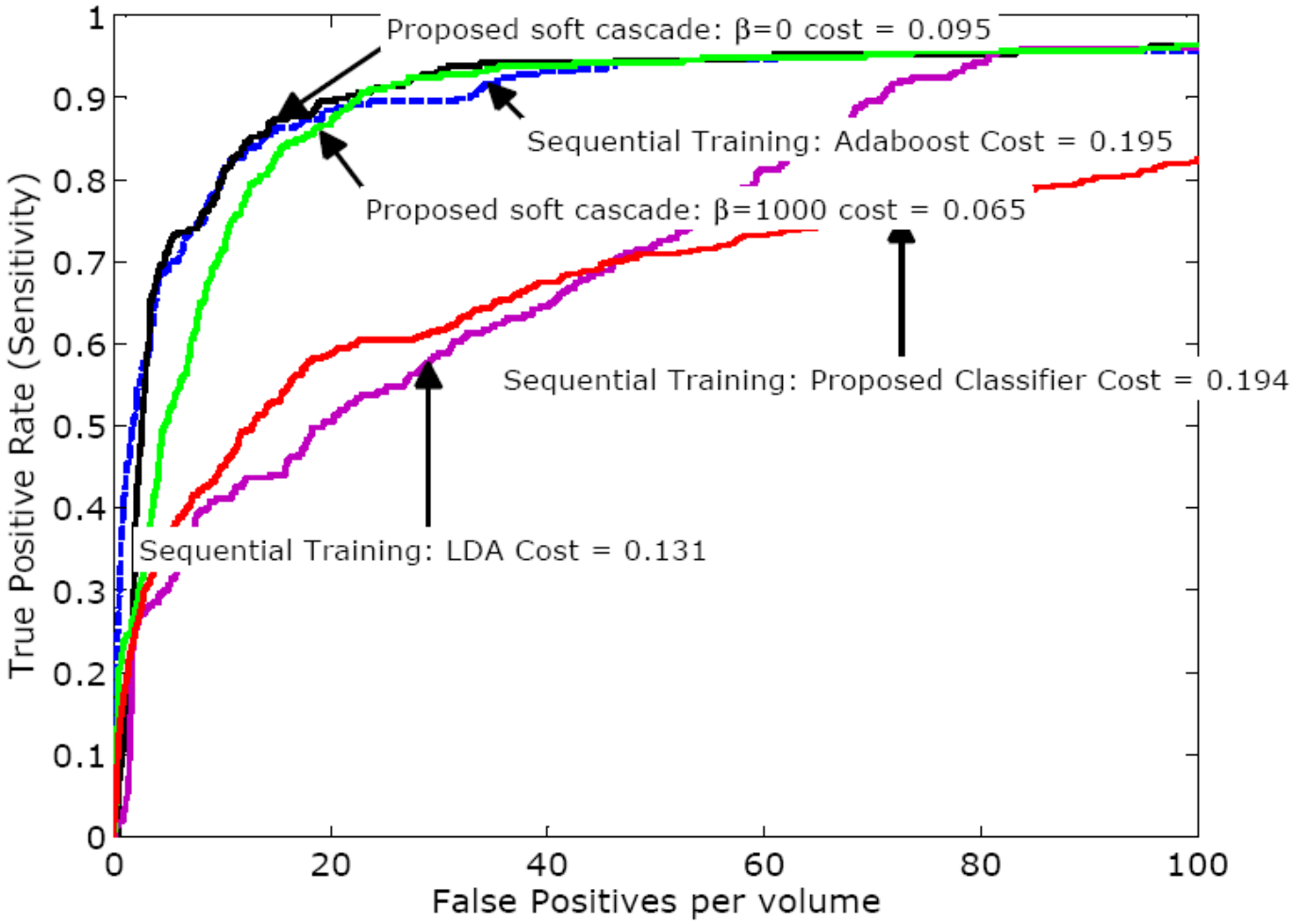
Computer aided diagnosis

- Motivation here is to reduce the computational cost
- 196 CT scans with 923 positive candidates and 54455 negative candidates.

Feature Group	Number of features	Average Cost
1	9	1.07 secs
2	23	3.10 secs
3	25	20.7 secs

R. B. Rao, J. Bi, G. Fung, M. Salganicoff, N. Obuchowski, and D. Naidich. LungCAD: a clinically approved, machine learning system for lung cancer detection. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1033–1037, 2007.

Test set FROC Curves



Conclusions

- Joint training of all stages of the cascade.
  - Notion of probabilistic soft cascades
- A knob to control the tradeoff between accuracy vs cost
  - Modeling the expected feature cost
- Open issues:
  - Order of the cascade
  - The accuracy vs cost knob is not sensitive in all problem domains

Related work

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 511–518, 2001.

L. Bourdev and J. Brandt. Robust object detection via soft cascade. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 236–243, 2005.

M. Dundar and J. Bi. Joint Optimization of Cascaded Classifiers for Computer Aided Detection. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

S. Ji and L. Carin. Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40(5):1474–1485, 2007.

P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. An expected utility approach to active feature-value acquisition. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 745–748, 2005.

V. S. Sheng and C. X. Ling. Partial example acquisition in cost-sensitive learning. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 638–646, 2007.