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

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

increasing predictive power ... increasing acquisition cost
A cascade of linear classifiers

\[ x = [x^1 x^2 x^3] \]

1. **Stage 1**
   - \[ w_1^T x^1 \]
   - \( w_1^T x^1 > \theta_1 \)
   - \( w_1^T x^1 \leq \theta_1 \) → -

2. **Stage 2**
   - \[ w_2^T x^2 \]
   - \( w_2^T x^2 > \theta_2 \)
   - \( w_2^T x^2 \leq \theta_2 \) → -

3. **Stage 3**
   - \[ w_3^T x^3 \]
   - \( w_3^T x^3 > \theta_3 \)
   - \( w_3^T x^3 \leq \theta_3 \) → -

Increasing predictive power
Increasing acquisition cost

- 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.

\[ x = [x^1 x^2 x^3] \]

\[ w_1^T x^1 \geq \theta_1 \]  
\[ w_2^T x^2 \geq \theta_2 \]  
\[ w_3^T x^3 \geq \theta_3 \]  

+
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 $x^j$

$K$ stage cascade $[C_1, C_2, \ldots, C_K]$

$x = [x^1, x^2, \ldots, x^K]$

Each stage is of the form $f_{C_j}(x) = w_j^T x^j$

$y = 1$ if $f_{C_j}(x) > \theta_j$ and $y = 0$ if $f_{C_j}(x) \leq \theta_j$

Logistic Regression $p_{C_j}(y = 1|x, w) = \sigma(f_{C_j}(x))$ $\sigma(z) = 1/(1 + e^{-z})$
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.

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

• 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

\[ \hat{w}_{MAP} = \arg \max_w L(w) \]

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

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

\[ E_{p(x)}[\mathcal{T}(x)] \leq c \]

where \( \mathcal{T} \) is the cost for a new instance \( x \)

• 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

<table>
<thead>
<tr>
<th>Stage 1</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>$t_1$</td>
</tr>
<tr>
<td>Stage 2</td>
<td>$t_2 \sigma(f_{C_1}(x_i))$</td>
</tr>
<tr>
<td>Stage 3</td>
<td>$t_3 \sigma(f_{C_1}(x_i)) \sigma(f_{C_2}(x_i))$</td>
</tr>
</tbody>
</table>

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

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

$\beta$ 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

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

C. Dehing-Oberije, D. De Ruyscher, 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)

<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Number of features</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Clinical features</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2 CT/PET scan features before treatment</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3 CT/PET scan features after treatment</td>
<td>2</td>
<td>10</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Lung Cancer</th>
<th>Testing set</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Single stage classifier</td>
<td>0.79[± 0.12]</td>
<td>✓</td>
<td>1.00[± 0.00]</td>
<td></td>
</tr>
<tr>
<td>(2) Proposed soft cascade $\beta = 0$</td>
<td>✓ 0.72[± 0.11]</td>
<td>✓</td>
<td>0.37[± 0.08]</td>
<td></td>
</tr>
<tr>
<td>(3) Sequential Training via Logistic Regression</td>
<td>0.71[± 0.09]</td>
<td>✓</td>
<td>0.45[± 0.08]</td>
<td></td>
</tr>
<tr>
<td>(4) Sequential Training via AdaBoost</td>
<td>✓ 0.63[± 0.05]</td>
<td>✓</td>
<td>0.68[± 0.08]</td>
<td></td>
</tr>
<tr>
<td>(5) Sequential Training via LDA</td>
<td>0.70[± 0.03]</td>
<td>✓</td>
<td>0.66[± 0.08]</td>
<td></td>
</tr>
<tr>
<td>(6) Proposed soft cascade $\beta = 10N$</td>
<td>0.73[± 0.12]</td>
<td></td>
<td>0.35[± 0.12]</td>
<td></td>
</tr>
<tr>
<td>(7) Proposed soft cascade $\beta = 100N$</td>
<td>0.70[± 0.11]</td>
<td></td>
<td>0.35[± 0.11]</td>
<td></td>
</tr>
<tr>
<td>(8) Proposed soft cascade $\beta = 1000N$</td>
<td>0.70[± 0.11]</td>
<td>✓</td>
<td>0.27[± 0.10]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rectum Cancer</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Single stage classifier</td>
<td>0.83[± 0.06]</td>
<td>✓</td>
<td>1.00[± 0.00]</td>
<td></td>
</tr>
<tr>
<td>(2) Proposed soft cascade $\beta = 0$</td>
<td>✓ 0.79[± 0.06]</td>
<td>✓</td>
<td>0.59[± 0.09]</td>
<td></td>
</tr>
<tr>
<td>(3) Sequential Training via Logistic Regression</td>
<td>0.76[± 0.09]</td>
<td>✓</td>
<td>0.70[± 0.10]</td>
<td></td>
</tr>
<tr>
<td>(4) Sequential Training via AdaBoost</td>
<td>0.73[± 0.10]</td>
<td>✓</td>
<td>0.68[± 0.09]</td>
<td></td>
</tr>
<tr>
<td>(5) Sequential Training via LDA</td>
<td>✓ 0.71[± 0.09]</td>
<td></td>
<td>0.63[± 0.12]</td>
<td></td>
</tr>
<tr>
<td>(6) Proposed soft cascade $\beta = 10N$</td>
<td>0.79[± 0.06]</td>
<td></td>
<td>0.57[± 0.08]</td>
<td></td>
</tr>
<tr>
<td>(7) Proposed soft cascade $\beta = 100N$</td>
<td>0.77[± 0.04]</td>
<td>✓</td>
<td>0.50[± 0.07]</td>
<td></td>
</tr>
<tr>
<td>(8) Proposed soft cascade $\beta = 1000N$</td>
<td>0.76[± 0.08]</td>
<td>✓</td>
<td>0.48[± 0.08]</td>
<td></td>
</tr>
</tbody>
</table>
Computer aided diagnosis

- Motivation here is to reduce the computational cost
- 196 CT scans with 923 positive candidates and 544,555 negative candidates.

<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Number of features</th>
<th>Average Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>1.07 secs</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>3.10 secs</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>20.7 secs</td>
</tr>
</tbody>
</table>

Test set FROC Curves

- Proposed soft cascade: \( \beta=0 \) cost = 0.095
- Sequential Training: Adaboost Cost = 0.195
- Proposed soft cascade: \( \beta=1000 \) cost = 0.065
- Sequential Training: Proposed Classifier Cost = 0.194
- Sequential Training: LDA Cost = 0.131
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


