COARSE-TO-FINE, COST-SENSITIVE CLASSIFICATION OF E-MAIL

Jay Pujara  jay@cs.umd.edu
Lise Getoor  getoor@cs.umd.edu

12/10/2010
Parallel Coarse-to-Fine Problems

- **Structure in output**
  - Labels naturally have a hierarchy from coarse-to-fine

- **Structure in input**
  - Features may have an order or systemic dependency
  - Acquisition costs vary: cheap or expensive features

- Exploit structure during classification

- Minimize costs
E-mail Challenges: Spam Detection

- Most mail is spam
- Billions of classifications
- Must be incredibly fast
E-mail Challenges: Categorizing Mail

- E-mail does more, tasks such as:
  - Extract receipts, tracking info
  - Thread conversations
  - Filter into mailing lists
  - Inline social network response

- Computationally intensive processing
- Each task applies to one class
Features have costs & dependencies

IP is known at socket connect time, is 4 bytes in size
Features have costs & dependencies

The Mail From is one of the first commands of an SMTP conversation. From addresses have a known format, but higher diversity.
Features have costs & dependencies

The subject, one of the mail headers, occurs after a number of network exchanges. Since the subject is user-generated, it is very diverse and often lacks a defined format.
Coarse task is constrained by feature cost

Feature Structure

- IP
  - Derived features
- Mail From
  - Derived features
- Subject
  - Derived features
- Body
  - Derived features

Class Structure

- Ham
  - Business
  - Social Network
- Spam
  - Personal
  - Newsgroup

Cost

Granularity
Fine task is constrained by misclassification cost

Feature Structure:
- Derived features
- IP
- Derivative features
- Mail From
- Subject
- Body

Class Structure:
- Ham
- Spam
  - Business
  - Social Network
  - Personal
  - Newsgroup

Cost:
- $  
- $$$

Granularity:
- 

Derived features
Approach: Granular Cost Sensitive Classifier

Training:
- Loss functions of form: \( L = \alpha FC + (1-\alpha) MC \)
- Choose \( \alpha_c \) and \( \alpha_f \) for coarse and fine tasks
- Calculate margin threshold where feature acquisition decreases loss across training data

Test:
- Compute decision margin with available features
- Acquire features until margin above threshold
- Classify instance
Experimental Setup

- Data from 1227 Yahoo! Mail messages from 8/2010
- Feature costs calculated from network + storage cost

<table>
<thead>
<tr>
<th>Class</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spam</td>
<td>531</td>
</tr>
<tr>
<td>Business</td>
<td>187</td>
</tr>
<tr>
<td>Social Network</td>
<td>223</td>
</tr>
<tr>
<td>Newsletter</td>
<td>174</td>
</tr>
<tr>
<td>Personal/Other</td>
<td>102</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>.168</td>
</tr>
<tr>
<td>MailFrom</td>
<td>.322</td>
</tr>
<tr>
<td>Subject</td>
<td>.510</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Feature Cost</th>
<th>Misclass Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coarse</td>
<td>Fine</td>
</tr>
<tr>
<td>Fixed: IP+MailFrom</td>
<td>.490</td>
<td>.098</td>
</tr>
<tr>
<td>GCSC: $\alpha_c=.3$, $\alpha_f=.05$</td>
<td>.479</td>
<td>.091</td>
</tr>
<tr>
<td>Fixed: IP+MailFrom+Subject</td>
<td>1.00</td>
<td>.090</td>
</tr>
<tr>
<td>GCSC: $\alpha_c=.15$, $\alpha_f=.01$</td>
<td>.511</td>
<td>.088</td>
</tr>
</tbody>
</table>

- Evaluated NB & SVM base classifiers, NB results shown
- Compare fixed features vs. GCSC with 10-fold L1O CV
- Same feature cost, decrease misclassification cost
- Decrease feature cost, same misclassification cost
Dynamics of choosing $\alpha_c$ and $\alpha_f$

As $\alpha_c$ increases, disparity in costs for different values of $\alpha_f$ widens.
Conclusion

- Examine a problem setting with coarse-to-fine structure in both input and output
- Propose a classifier, mapping input to output
  - at different granularities
  - sensitive to feature and misclassification costs
- Demonstrate results superior to baseline

Questions?

Research funded by Yahoo! Faculty Research Engagement Program