Using Ensemble Decisions and Active Selection to Improve Low-Cost Labeling for Multi-View Data

Umaa Rebbapragada, Kiri Wagstaff
Jet Propulsion Laboratory
California Institute of Technology

July 2, 2011
ICML 2011 Workshop on Combining Learning Strategies for Reducing Label Cost
Bellevue, WA
Classification of Sensor Network Data

- Node-level classification (in situ)
- Each node collects unique “view”
- Limited availability of labeled data
- Continuous stream of unlabeled data
- Nodes may communicate
## Combining Learning Strategies

<table>
<thead>
<tr>
<th>Example Selection</th>
<th>Labeling</th>
<th>Learning Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-confidence</td>
<td>Low-cost (self)</td>
<td>Bootstrapping Co-training</td>
</tr>
<tr>
<td>Low-confidence</td>
<td>High-cost (oracle)</td>
<td>Active learning</td>
</tr>
<tr>
<td>Low-confidence</td>
<td>Low-cost (ensemble)</td>
<td>?</td>
</tr>
</tbody>
</table>
Co-training

- Multi-view
- Semi-supervised
- Self-labels
Co-training\textsuperscript{1}

- Each learner classifies its unlabeled pool
- Each learner selects its most confident predictions
- All selected examples are moved to $L_1$ and $L_2$

[1] Blum and Mitchell, Combining Labeled and Unlabeled Data with Co-training, COLT, 1998
Very Long Baseline Array (VLBA) Data

• Geographically dispersed antennas
• Time series observations of pulsar PSR B0329+54
• 21 observations per example
• Classify pulse/non-pulse (680 pos, 680 neg)
• Created 4-, 6- and 9-view data sets. Results shown on 4-view data set
Co-training Produces Unreliable Labels

- Self-labeling introduces label noise
- Sensitive to base learner
- Noted by Pierce & Cardie, 2001
Co-training Produces Unreliable Labels

- Self-labeling introduces label noise
- Sensitive to base learner
- Noted by Pierce & Cardie, 2001
Talk Outline

• Low-cost ways to improve label reliability?

• Can we combine low-cost labeling with low-confidence example selection?
Multi-view Ensemble Labeling

- Learner classifies $U_1$, and selects an example
- Learner queries the other learners for a label
- Learner receives responses, and unifies them
- Example is added to $L_1$, $L_2$, $L_3$ and $L_4$
Ensemble Labeling

• Strategies for unifying neighbor predictions

• Abstain if cannot unify prediction with high confidence

• Majority Vote
  – Choose prediction with most votes
  – Abstain if at least half the ensemble did not make this prediction

• Consensus Vote
  – Choose unanimous prediction, otherwise abstain
Ensemble Labeling

Label Reliability

Test Set Accuracy

Logistic Regression
Low Confidence (LC) Example
Selection with Oracle Labeling

Logistic Regression

SVM
Pairing Low-Confidence Example Selection with Low-Cost Labeling

Label Reliability

Accuracy

Logistic Regression – p-values fail significance tests!
Collaborative Learning

- Each node of sensor network contains a classifier
- Classifier initialized with small amount of labeled data
- Classifier labels incoming data
- Collaborative learning learners to collaborate via queries to its nearest neighbors for examples and labels
Future Work

• Experimental results on 4-, 6-, 9-view VLBA data, and two 4-view data sets created from UCI repository
• Improve abstention policies => improve label reliability
• Goal: Confirm low-cost labeling and low-confidence example selection are compatible