Multi-Task Active Learning

Yi Zhang
Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL’ 08)
  - Image Classification (CVPR’ 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories
Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL’ 08)
  - Image Classification (CVPR’ 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories
Active Learning

- Select samples for labeling
  - Optimize model performance given the new label

\[ U : \quad x_i \quad \ldots \quad x \quad \ldots \quad x' \quad \ldots \quad x_n \]

\[ \hat{p} = \hat{p}(Y \mid X = x) \]

Label

Samples

Model(s)
Active Learning

- Uncertainty sampling

\[
\arg\max_{x \in U} \left[ - \sum_{y} \hat{p}(Y = y | x) \log_2 \hat{p}(Y = y | x) \right]
\]

- Maximize: the reduction of model entropy on \( x \)
Active Learning

- Query by committee (e.g., vote entropy)

\[
\arg\max_{x \in U} \left[ - \sum_y \hat{p}_C(Y = y|x) \log_2 \hat{p}_C(Y = y|x) \right]
\]

- Maximize: the reduction of version space
Active Learning

- Density-weighted entropy

\[
\arg\max_{x \in U} \left[ \hat{P}_U(x) \cdot - \sum_y \hat{p}(Y = y|x) \log_2 \hat{p}(Y = y|x) \right]
\]

- Maximize: approx. entropy reduction over \( U \)
Active Learning

- Estimated error (uncertainty) reduction

\[ \arg\min_{x \in U} \left[ \sum_{y} \hat{p}(Y = y|x) \sum_{x' \in U} \text{Uncertain}^+(x,y)(x') \right] \]

- Maximize: reduction of uncertainty over $U$
Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL’ 08)
  - Image Classification (CVPR’ 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories
The Problem

- Select a sample $\rightarrow$ labeling all tasks

$X_1 \ldots X \ldots X' \ldots X_n$

$Y_1 \vdots Y_i \vdots Y_m$

$\vdots \vdots \vdots \vdots \vdots \vdots \vdots \vdots$

$\vdots \vdots \vdots \vdots \vdots \vdots \vdots \vdots$

$\vdots \vdots \vdots \vdots \vdots \vdots \vdots \vdots$

Samples

Tasks
Methods

- Alternating selection
  - Iterate over tasks, sample a few from each task
Methods

- Rank combination
  - Combine rankings/scores from all single-task ALs
Experiments

- Learning two (dissimilar) tasks
  - Named entity recognition: CRFs
  - Parsing: Collins’ parsing model
- Competitive AL methods
  - Random selection
  - One-side active learning: choose samples from one task, and require labels for all tasks
    - Separate AL in each task is not studied (!)
  - Alternating selection
  - Ranking combination
Unanswered Questions

- Why “choose-one, labeling-all”? 
  - Authors: annotators may prefer to annotate the same sample for all tasks
- Why learning two dissimilar tasks together? 
  - Outputs of one task may be useful for the other 
  - Not studied in the paper
Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL’ 08)
  - Image Classification (CVPR’ 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories
The Problem: Multi-Label Image Classification

- Select any sample-label pair for labeling
Proposed Method

\[
\arg\max_{x \in D, y_s \in U(x)} \left[ \sum_{i=1}^{m} MI(y_i, y_s | y_{L(x)}, x) \right]
\]

- **D**: the set of samples
- **x**: a sample in **D**
- **U(x)**: unknown labels of **x**
- **L(x)**: known labels of **x**
- **m**: number of tasks
- **y_s**: a selected label from **U(x)**
- **y_i**: the label of the **i**\(^{th}\) task (for a sample **x**)

[Diagram showing samples and tasks with labels]
Proposed Method

- Why maximizing Mutual Information?
  - Connecting Bayes (binary) classification error to entropy and MI (Hellman and Raviv, 1970)

\[
\mathcal{E} \left( y_i \mid y_s, y_L(x), x \right) \leq \frac{1}{2} H \left( y_i \mid y_s, y_L(x), x \right)
\]
Proposed Method

- Why maximizing Mutual Information?
  - Connecting Bayes (binary) classification error to entropy and MI (Hellman and Raviv, 1970)

\[
\mathcal{E} \left( y_i \mid y_s; y_L(x), x \right) \leq \frac{1}{2} H \left( y_i \mid y_s; y_L(x), x \right)
\]

\[
\mathcal{E} \left( y \mid y_s; y_L(x), x \right)
\]

\[
\overset{(1)}{=} \frac{1}{m} \sum_{i=1}^{m} \mathcal{E} \left( y_i \mid y_s; y_L(x), x \right)
\]

\[
\overset{(2)}{\leq} \frac{1}{2m} \sum_{i=1}^{m} H \left( y_i \mid y_s; y_L(x), x \right)
\]

\[
\overset{(3)}{=} \frac{1}{2m} \sum_{i=1}^{m} \left\{ H \left( y_i \mid y_L(x), x \right) - MI \left( y_i; y_s \mid y_L(x), x \right) \right\}
\]
Proposed Method

- Compare: maximize the reduction of entropy
Modeling Joint Label Probability

\[ \max_{x \in D, y_s \in U(x)} \left\{ \sum_{i=1}^{m} MI(y_i, y_s | y_L(x), x) \right\} \]

- But how to compute:

\[ MI(y_i, y_s | y_L(x), x) \]

- Need the joint conditional probability of labels

\[ \hat{p}(y|x) = \hat{p}(y_1, y_2, \ldots, y_m | x) \]
Modeling Joint Label Probability

- Linear maximum entropy model
  \[ \hat{P}(y|x) = \frac{1}{Z(x)} \exp (y^T (b + Ry + Wx)) \]

- Kernelized version
  \[ \hat{P}(y|x) = \frac{1}{Z(x)} \exp (y^T (b + Ry) + y^T K(W, x)) \]

- EM for incomplete labels
Experiments

- **Data**
  - Image scene classification
  - Gene function classification

- **Two competitive AL methods**
  - Random selection of sample-label pairs
  - Choose one sample, labeling all tasks for it
    - Separate AL in each task is not studied (!)
Discussion

- Maximizing the joint mutual information is reasonable
- Directly estimate the joint label probability
  - Recognize the correlation between labels
  - Need more labeled examples
  - What if # tasks is large?
  - Cannot use specialized models for each task
  - Can we use external knowledge to couple tasks?
Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL’ 08)
  - Image Classification (CVPR’ 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories
Constraint-Driven Multi-Task Active Learning

- Multiple tasks $Y_1, Y_2, \ldots, Y_m$
- Learners for each task $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m$
- A set of constraints $C$ among tasks
- May have new tasks to launch
Value of Information (VOI) for Active Learning

- Single-task AL
  - Value of information (VOI) for labeling a sample \( x \)

\[
VOI(Y, x) = VOI(x) = \sum_{y \in Dom(Y)} P(Y = y|x)R(Y = y, x)
\]
Value of Information (VOI) for Active Learning

- Single-task AL
  - Value of information (VOI) for labeling a sample $x$
    \[
    VOI(Y, x) = VOI(x) = \sum_{y \in \text{Dom}(Y)} P(Y = y|x)R(Y = y, x)
    \]
  - Reward $R(Y=y, x)$, e.g., how surprising it is?
    \[
    R(Y = y, x) = -\log_2 \hat{p}(Y = y|x)
    \]
Value of Information (VOI) for Active Learning

- Single-task AL
  - Value of information (VOI) for labeling a sample $x$
    \[ VOI(Y, x) = VOI(x) = \sum_{y \in \text{Dom}(Y)} P(Y = y|x)R(Y = y, x) \]
  - Reward $R(Y=y, x)$, e.g., how surprising it is?
    \[ R(Y = y, x) = -\log_2 \hat{p}(Y = y|x) \]
  - Finally, replace $P(Y=y|x)$ with $\hat{p}$
    \[ VOI(x) = \sum_{y \in \text{Dom}(Y)} -\hat{p}(Y = y|x) \log_2 \hat{p}(Y = y|x) \]
Constraint-Driven Active Learning

- Multiple tasks with constraints

\[ VOI(Y_i, x) = \sum_{y_i \in \text{Dom}(Y_i)} P(Y_i = y_i|x)R(Y_i = y_i|x) \]

\[ x \in D, Y_i \in U(x) \]

- Probability estimate of outcomes

\[ P(Y_i = y_i|x) = \begin{cases} \hat{p}_i(Y_i = y_i|x) & \text{if } Y_i \text{ is a learned task} \\ \frac{1}{|\text{Dom}(Y_i)|} & \text{if } Y_i \text{ is a new task} \end{cases} \]
Constraint-Driven Active Learning

- Reward function $R(y, x)$ in:

$$VOI(Y_i, x) = \sum_{y_i \in Dom(Y_i)} P(Y_i = y_i | x) R(Y_i = y_i, x)$$
Constraint-Driven Active Learning

- Propagate rewards via constraints

$$R(Y_i = y_i, x) = \sum_{(Y_j = y_j) \in Prop(Y_i = y_i, x, C)} -\log_2 \hat{p}_j (Y_j = y_j | x)$$

$$Prop(Y_i = y_i, x, C) = \{(Y_j = y_j) | (Y_i = y_i) \rightarrow_c (Y_j = y_j), Y_j \in U(x)\}$$
Constraint-Driven Active Learning

- Multi-task AL with constraints

\[ VOI(Y_i, x) = \sum_{y_i \in Dom(Y_i)} \hat{p}_i(Y_i = y_i | x) \sum_{(Y_j = y_j) \in Prop(Y_i = y_i, C)} - \log_2 \hat{p}_j(Y_j = y_j | x) \]

\[ x \in D, Y_i \in U(x) \]

- Recognize *inconsistency* of among tasks
- Launch new tasks
- Favor poorly performed tasks, and “pivot” tasks
- Density-weighted measure?
- Use state-of-the-art learners for single tasks
Experiments

- Four named entity recognition tasks
  - “Animal”
  - “Mammal”
  - “Food”
  - “Celebrity”

- Constraints
  - 1 inheritance, 5 mutual exclusion
  - Lead to 12 propagation rules (plus 1 identity rule)
Experiments

- Competitive methods for AL
  - VOI of sample-task pairs with constraints
  - VOI of sample-task pairs without constraints
  - Single-task AL

\[
VOI(Y_i, x) = \sum_{y_i \in \text{Dom}(Y_i)} \hat{p}_i(Y_i = y_i | x) \sum_{(Y_j = y_j) \in \text{Prop}(Y_i = y_i, C)} - \log_2 \hat{p}_j(Y_j = y_j | x)
\]
Experiments

- Results: MAP on animal, food and celebrity
Experiments

- Results: MAP on all four tasks
Experiments

- **Analysis**
  - True labels from the NNLL system
  - 90% precision for “mammal”
    - 10% label noise on the task “mammal”
  - Tasks are generally “easy”
    - Positive examples are highly homogenous
Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL’ 08)
  - Image Classification (CVPR’ 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories
Cost-Sensitive Active Learning Across Tasks

- Which scenario is reasonable?
  - Choose one sample, label all tasks
  - Arbitrary sample-label pairs
Cost-Sensitive Active Learning Across Tasks

- Costs for labeling multi tasks on a sample $x$
  - $x$ is a long document
Cost-Sensitive Active Learning Across Tasks

- Costs for labeling multi tasks on a sample $x$
  - $x$ is a word or an image
Cost-Sensitive Active Learning Across Tasks

- Learn a more realistic cost function?
- Active learning aware of labeling costs?
Outline

● Active Learning
● Multi-Task Active Learning
  ● Linguistic Annotations (ACL’ 08)
  ● Image Classification (CVPR’ 08)
● Current Work and Discussions
  ● Constraint-Driven Active Learning Across Tasks
  ● Cost-Sensitive Active Learning Across Tasks
  ● Active Learning of Constraints and Categories
Active Constraint Learning

- New constraints/rules are highly valuable
- Find significant rules and avoid false discovery
  - Oversearching (Quinlan, et al. IJCAI’ 95)
  - Multiple comparisons (Jensen, et al. MLJ’ 00)
  - Statistical tests (Webb, MLJ’ 06)
- Combining first-order logic with graphical models
  - Bayesian logic programs (logic + BN)
  - Markov logic networks (logic + MRF)
  - Structure sparsity on graphs?
Active Category Detection

- Automatically detect new categories
- Clustering
  - High-dimensional space
  - Co-clustering/bi-clustering
  - Local search vs. global partition
- Subgraph/community detection
  - A huge bipartite graph
  - Optimize modularity of the graph
  - Overlapping communities?
Thanks!

- Questions?