Learning from Inconsistent and Unreliable Annotators

by Ping Zhang and Zoran Obradovic
Center for Data Analytics and Biomedical Informatics, Temple University

September, 2011
Athens, Greece
Typical supervised classification

Learn a classification function

\[ f : \mathbb{R}^d \rightarrow \mathcal{Y} \]

<table>
<thead>
<tr>
<th>Instance</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>1</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0</td>
</tr>
<tr>
<td>$x_4$</td>
<td>1</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>$x_N$</td>
<td>1</td>
</tr>
</tbody>
</table>
Golden ground truth

- How to obtain the labels for training?

\[ y_i \in \mathcal{Y} = \{0, 1\} \]

- Getting the actual golden ground truth can be
  - Expensive
  - Potentially dangerous
  - Could be impossible
Subjective ground truth from multiple annotators

- Getting golden ground truth is hard, so we use opinion from an annotator.
- An annotator provides his/her subjective version of the truth.
- Error prone/noisy/unreliable.
- Use multiple annotators who label the same example.
Annotations from multiple annotators

Each radiologist is asked to annotate whether a lesion is malignant (1) or not (0).

<table>
<thead>
<tr>
<th>Lesion ID</th>
<th>Radiologist 1</th>
<th>Radiologist 2</th>
<th>Radiologist 3</th>
<th>Radiologist 4</th>
<th>Truth Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x</td>
</tr>
<tr>
<td>02</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>x</td>
</tr>
<tr>
<td>03</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>x</td>
</tr>
<tr>
<td>04</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>x</td>
</tr>
<tr>
<td>05</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>x</td>
</tr>
<tr>
<td>06</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>x</td>
</tr>
<tr>
<td>07</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>x</td>
</tr>
</tbody>
</table>

In practice there is a substantial amount of disagreement.

We have no knowledge of the actual golden ground truth.

Getting absolute ground truth (e.g. biopsy) can be expensive.
We are interested in

- Building a model to answer questions
  - How to train a classifier?
  - How to evaluate annotators?
  - How to estimate the actual ground truth?
How to judge an annotator? (1)

Sensitivity

$$\alpha^j = \Pr[y^j = 1 | y = 1]$$

Label assigned by annotator j

True label

Specificity

$$\beta^j = \Pr[y^j = 0 | y = 0]$$
How to judge an annotator? (2)

Good annotators should have high sensitivity and high specificity.
Varying effectiveness on types of data

- In many cases annotator knowledge can fluctuate considerably depending on the groups of input instances.
- Build data-dependent model based on the intuition that inconsistent annotators have different sensitivity and specificity for different regions of the feature space.
- How to find the fittest model to approximate the distribution of the instances?
How to approximate the distribution of the instances?

- **Gaussian mixture model (GMM):** Linear superposition of Gaussians components
- Well-studied statistical inference techniques are available (EM algorithm)
- A "soft" group assignment is available. E-step evaluates the probability that an observation $x_i$ belongs to component $k$ as $\tau_{ik}$
- Choose the model and the number of components by **Bayesian Information Criterion (BIC)**
Problem Statement

- **Input:** Given N instances with annotations from R annotators

\[ D = \{ x_i, y_{i1}, \ldots, y_{iR} \}_{i=1}^N \]

- **Output:**
  - Sensitivities at each component
  - Specificities at each component
  - Estimates of true labels \( y_1, \ldots, y_N \)
If we know the true labels

- We can learn a classifier
- To model the data-dependent behavior of annotators, we hypothesize that each annotator has its own sensitivity and specificity for each mixture component

\[
\alpha_k^j = \Pr(y_i^j = 1 \mid y_i = 1, k\text{-th Gaussian mixture component generates } x_i)
\]

\[
\beta_k^j = \Pr(y_i^j = 0 \mid y_i = 0, k\text{-th Gaussian mixture component generates } x_i)
\]

\[
\alpha_k^j = \frac{\sum_{i=1}^{N} z_{ik} y_i^j}{\sum_{i=1}^{N} z_{ik}}
\]

\[
\beta_k^j = \frac{\sum_{i=1}^{N} (\tau_{ik} - z_{ik})(1 - y_i^j)}{\sum_{i=1}^{N} (\tau_{ik} - z_{ik})}
\]

\(Z_i\) is a soft label (probability that the label is 1) and \(z_{ik} = z_i \tau_{ik}\)
How to find the unknown true labels (1)

- Hypothesize the behavior of annotator: Given an instance $x_i$ to label, the annotator finds the mixture component which most likely generates that instance. Then the annotators generate labels with their sensitivities and specificities at the most likely component.

$$z_i = \frac{\Pr[y_i^1, \ldots, y_i^R | y_i = 1, \phi] \cdot \Pr[y_i = 1 | x_i, \phi]}{\Pr[y_i^1, \ldots, y_i^R | \phi]}$$

Again, prior probability by classifier

$$\Pr[y_i^1, \ldots, y_i^R | y_i = 1, \alpha] = \Pr[y_i^1, \ldots, y_i^R | y_i = 1, \alpha^R]$$

where $q = \arg \max_{k=1,\ldots,K} (\tau_{ik})$

$$= \prod_{j=1}^R \Pr[y_i^j | y_i = 1, \alpha_q^j] = \prod_{j=1}^R [\alpha_q^j]^{y_i^j} [1 - \alpha_q^j]^{1-y_i^j}$$
How to find the unknown true labels (2)

- Therefore, if we know annotators’ sensitivities and specificities at each component, the estimation of the hidden true label is:

\[
  z_i = \frac{a_i p_i}{a_i p_i + b_i (1 - p_i)}
\]

where

\[
p_i = \Pr[y_i = 1 | x_i, w] = \sigma(w^T x_i)
\]

\[
a_i = \prod_{j=1}^{R} [\alpha_q^j]^{y_j^i} [1 - \alpha_q^j]^{1-y_j^i}
\]

\[
b_i = \prod_{j=1}^{R} [1 - \beta_q^j]^{y_j^i} [\beta_q^j]^{1-y_j^i}
\]

\[
q = \arg \max_{k=1,...,K} (\tau_{ik})
\]
GMM-MAPML Algorithm

Find the fittest model to approximate the distribution of the instances

If we know how good each predictor is, we can estimate the true label

\[ z_i = \frac{\sigma(w^T x_i) \prod_{j=1}^{R} [\alpha_q^j]^{y_i^j} [1-\alpha_q^j]^{1-y_i^j}}{\sigma(w^T x_i) \prod_{j=1}^{R} [\alpha_q^j]^{y_i^j} [1-\alpha_q^j]^{1-y_i^j} + (1-\sigma(w^T x_i)) \prod_{j=1}^{R} [\beta_q^j]^{1-y_i^j} [1-\beta_q^j]^{y_i^j}} \]

Iterate until convergence

Initialize using majority-voting

If we know the true label we can estimate how good each predictor is at each component

\[ \alpha_k^j = \frac{\sum_{i=1}^{N} z_{ik} y_i^j}{\sum_{i=1}^{N} z_{ik}} \]
\[ \beta_k^j = \frac{\sum_{i=1}^{N} (\tau_{ik} - z_{ik})(1-y_i^j)}{\sum_{i=1}^{N} (\tau_{ik} - z_{ik})} \]

Learn a classifier
Analysis of the model

\[
\text{logit}(z_i) = \ln \frac{z_i}{1 - z_i} = \ln \frac{\Pr[y_i = 1 | y_i^1, \ldots, y_i^R, x_i, \phi]}{\Pr[y_i = 0 | y_i^1, \ldots, y_i^R, x_i, \phi]}
\]

\[
= w^T x_i + \sum_{j=1}^{R} y_i^j [\text{logit}(\alpha_q^j) + \text{logit}(\beta_q^j)] + c
\]

Observations \quad \text{Annotators’ labels} \quad \text{Constant}

Consider both sensitivity and specificity as weight.

\[
q = \arg \max_{k=1, \ldots, K} (\tau_{ik}) \quad \text{indicates data-dependent.}
\]
Emotional speech classification

Why to study emotional speech?
- Recognition (e.g., Interface optimization in call centers)
- Generation (e.g., TTS, games)

Acted emotional utterance
- Semantically neutral
- Four acted emotions: happy, neutral, sad, angry
Dataset: EMA database from University of Southern California

- Golden ground truth is known: 568 utterances were chosen as best emotional utterances
- 39-element feature vectors were extracted from the speech signal (WAV file) by using VOICEBOX
- Binary labels: \{happy, neutral\} were assigned to positive emotion (0), \{sad, angry\} were assigned to negative emotion (1)
- Multiple annotators: 5 annotators with different academic background. Most of them are non-native English speakers. Noisy/unreliable annotators
Experiment Results: ROC comparisons

ROC Curve for the classifier

- LR Concatenation: AUC=0.792
- Majority Voting: AUC=0.798
- MAP-ML: AUC=0.825
- GMM-MAPML: AUC=0.858
- LR Ground Truth: AUC=0.881
## Experiment Results: GMM-MAPML based estimates of annotators’ accuracy

### First Component

<table>
<thead>
<tr>
<th>Listeners</th>
<th>Estimated Sensitivity</th>
<th>Estimated Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>0.902</td>
<td>0.891</td>
</tr>
<tr>
<td>Listener 2</td>
<td>0.843</td>
<td>0.862</td>
</tr>
<tr>
<td>Listener 3</td>
<td>0.784</td>
<td>0.802</td>
</tr>
<tr>
<td>Listener 4</td>
<td>0.756</td>
<td>0.744</td>
</tr>
<tr>
<td>Listener 5</td>
<td>0.719</td>
<td>0.698</td>
</tr>
</tbody>
</table>

### Second Component

<table>
<thead>
<tr>
<th>Listeners</th>
<th>Estimated Sensitivity</th>
<th>Estimated Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>0.925</td>
<td>0.951</td>
</tr>
<tr>
<td>Listener 2</td>
<td>0.814</td>
<td>0.799</td>
</tr>
<tr>
<td>Listener 3</td>
<td>0.779</td>
<td>0.792</td>
</tr>
<tr>
<td>Listener 4</td>
<td>0.877</td>
<td>0.861</td>
</tr>
<tr>
<td>Listener 5</td>
<td>0.728</td>
<td>0.736</td>
</tr>
</tbody>
</table>

### Graph

- **First Component**
- **Second Component**

---

**Male Utterances**

**Female Utterances**

- 1st Component
- 2nd Component

**% of utterances in components**
Protein Disorder Prediction

- **Lock and Key Paradigm:**
  AA seq $\rightarrow$ **3-D Structure** $\rightarrow$ Function

- **Definition:** A part of the protein or the whole protein doesn't have a fixed tertiary structure

- **Importance:** Involved in many important functions and in various diseases.
117 experimentally characterized targets (=26083 residues) were analyzed containing: 9.30% disordered residues and 90.70% ordered residues

Golden ground truth is known: either X-ray or NMR experimental characterization

20-element feature vectors (19 amino acid composition features and 1 sequence complexity feature) were extracted from the protein sequences

Multiple annotators: Labels by 15 predictors developed at different institutions

Disordered segments <4 residues were not considered
### CASP9 Assessment Scores

<table>
<thead>
<tr>
<th>Predictor Name</th>
<th>Institution</th>
<th>ACC</th>
<th>Sw</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-MAPML</td>
<td></td>
<td>0.785</td>
<td>0.527</td>
<td>0.874</td>
</tr>
<tr>
<td>MAP-ML</td>
<td></td>
<td>0.764</td>
<td>0.513</td>
<td>0.859</td>
</tr>
<tr>
<td>MAJORITY VOTING</td>
<td></td>
<td>0.735</td>
<td>0.496</td>
<td>0.776</td>
</tr>
<tr>
<td>PRDOS2</td>
<td>Tokyo Tech, Japan</td>
<td>0.754</td>
<td>0.509</td>
<td>0.855</td>
</tr>
<tr>
<td>MULTICOM-REFINE</td>
<td>University of Missouri, USA</td>
<td>0.75</td>
<td>0.5</td>
<td>0.822</td>
</tr>
<tr>
<td>BIOMINE_DR_PDB</td>
<td>University of Alberta, Canada</td>
<td>0.741</td>
<td>0.483</td>
<td>0.821</td>
</tr>
<tr>
<td>GSMETADISORDERMD</td>
<td>IIMCB in Warsaw, Poland</td>
<td>0.738</td>
<td>0.476</td>
<td>0.816</td>
</tr>
<tr>
<td>MASON</td>
<td>George Mason University, USA</td>
<td>0.736</td>
<td>0.473</td>
<td>0.743</td>
</tr>
<tr>
<td>ZHOU-SPINE-D</td>
<td>IU School of Medicine, USA</td>
<td>0.731</td>
<td>0.462</td>
<td>0.832</td>
</tr>
<tr>
<td>DISTILL-PUNCH1</td>
<td>UCD Dublin, Ireland</td>
<td>0.726</td>
<td>0.453</td>
<td>0.8</td>
</tr>
<tr>
<td>OND-CRF</td>
<td>Umea University, Sweden</td>
<td>0.706</td>
<td>0.412</td>
<td>0.737</td>
</tr>
<tr>
<td>UNITED3D</td>
<td>Kitasato University, Japan</td>
<td>0.704</td>
<td>0.412</td>
<td>0.781</td>
</tr>
<tr>
<td>CBRC_POODLE</td>
<td>CBRC, Japan</td>
<td>0.694</td>
<td>0.405</td>
<td>0.83</td>
</tr>
<tr>
<td>MCGUFIN</td>
<td>University of Reading, UK</td>
<td>0.688</td>
<td>0.402</td>
<td>0.817</td>
</tr>
<tr>
<td>ISUNSTRUCT</td>
<td>IPR RAS, Russia</td>
<td>0.679</td>
<td>0.396</td>
<td>0.742</td>
</tr>
<tr>
<td>DISOPRED3C</td>
<td>University College London, UK</td>
<td>0.67</td>
<td>0.391</td>
<td>0.853</td>
</tr>
<tr>
<td>ULG-GIGA</td>
<td>University of Liege, France</td>
<td>0.585</td>
<td>0.341</td>
<td>0.726</td>
</tr>
<tr>
<td>MEDOR</td>
<td>Aix-Marseille University, France</td>
<td>0.579</td>
<td>0.338</td>
<td>0.688</td>
</tr>
</tbody>
</table>
GMM-MAPML based estimates of CASP9 disorder predictors’ accuracy
Thank you! | Questions?

Ping Zhang: ping@temple.edu
Zoran Obradovic: zoran.obradovic@temple.edu