CONSENSUS EXTENSION TO ACTIVE LEARNING: THEORY AND APPLICATIONS

Pattern Recognition in Bioinformatics 2010
Scott Doyle and Anant Madabhushi
Rutgers University, Department of Biomedical Engineering, Piscataway, NJ
Overview

- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Active Learning (AL) Overview

- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Consensus of Ambiguity: Combining AL methods

- Theory of CoA

- Experimental Results

- Concluding Remarks
Supervised Classification Paradigm

Dataset \( x_i \in X \)  \rightarrow \text{Randomly Sampled Training Points}  \rightarrow \text{Expert Annotated Training Points}  \rightarrow C(x_i)
Building Training with Random Learning

- Each sample is an observation of that sample’s class
- With random sampling or learning (RL): more samples are better (more complete class model)
- Problem: Training samples are difficult to obtain!
Building Training with Random Learning

High accuracy requires training that is:
- Accurate – Correctly labeled
- Representative – Contains class information
- Discriminative – Captures class differences

Building Training with Random Learning

- Expert medical knowledge is required
- Large images (1-2 GB): tedious, time-consuming to obtain detailed contours
- Each training image requires a great deal of effort
Active Learning (AL) Overview

- Active Learning (AL):
  - Selectively choose only informative samples for training
  - “Informative”: samples that are likely to increase classifier performance

![Diagram showing active learning process]

- Informative Points
- Uninformative Points
- Actively Sampled Training Points
- Expert Annotations
- Supervised Classifier

\[ C(x_i) \]
Active Learning (AL) Overview

- How do we find “informative” annotation samples?
- Concept of “sample ambiguity”
The more ambiguous a sample is, the more likely it is informative (should be selected for annotation)

Overview

- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Active Learning (AL) Overview

- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Consensus of Ambiguity: Combining AL methods

- Theory of CoA

- Experimental Results

- Concluding Remarks
Measuring Sample Ambiguity

  - Distance to the decision hyperplane
Measuring Sample Ambiguity

- Bayes’ Likelihood:
  - Based on likelihood of class membership

![Graph showing posterior probability of class membership](image)
## Measuring Sample Ambiguity

- **Seung (1992), Freund (1997): Query-by-Committee**
  - Based on disagreement among weak bagged classifiers

<table>
<thead>
<tr>
<th>Weak Classifiers</th>
<th>Average Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2/M = 0.2</td>
</tr>
<tr>
<td></td>
<td>3/M = 0.3</td>
</tr>
<tr>
<td></td>
<td>2/M = 0.2</td>
</tr>
<tr>
<td></td>
<td>5/M = 0.5</td>
</tr>
</tbody>
</table>

Combining AL Methods

- AL methods use one description of ambiguity each
- Ensemble methods combine multiple algorithms:
  - Variance is exploited to yield optimal results
  - Consensus classification: sample classification
  - Consensus Active Learning: sample ambiguity
- Consensus of Ambiguity (CoA)
Advantage of CoA vs. AL

- RL: All samples to be annotated
Advantage of CoA vs. AL

- RL: All samples to be annotated
Advantage of CoA vs. AL

- RL: All samples to be annotated
- AL: Subset of samples eligible for annotation
Advantage of CoA vs. AL

- RL: All samples to be annotated
- AL: Subset of samples eligible for annotation
- CoA: FEWER eligible samples for annotation
Advantage of CoA vs. AL

- RL: All samples to be annotated
- AL: Subset of samples eligible for annotation
- CoA: FEWER eligible samples for annotation
Overview

- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Active Learning (AL) Overview

- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Combining different AL methods

- Theory of CoA

- Experimental Results

- Concluding Remarks
CoA Theory: Specific Properties

- **CoA Properties:**
  - Multiple AL algorithms reduce ambiguous samples
  - Additional algorithms increase benefit of CoA

- **Necessary Components:**
  - General definition of ambiguous sample
  - Consensus among multiple algorithms (consensus ratio)
  - Identifying “strongly” ambiguous samples
Goal of the training algorithm: Build $S^{tr}$ from unlabeled samples contained in $X$.

Samples chosen according to a Training Function: $\Phi(x_i)$ which measures sample ambiguousness.
**Definition 1.** A sample \( x_i \in X \) is considered ambiguous if \( a < \Phi(x_i) < b \) where \( a \), \( b \) are lower and upper bounds for \( \Phi(x_i) \), respectively.
CoA Theory: Multiple Algorithms

- CoA employs multiple training algorithms:
  \[ \Phi_j, j \in \{1, 2, \ldots, M \} \]
- Each algorithm returns a corresponding set of ambiguous (i.e. eligible-to-annotate) samples:
  \[ S_1^E, S_2^E, \ldots, S_M^E \]
Definition 2. Given nonempty sets of ambiguous samples, \( S_j^E, j \in \{1, 2, \cdots, M\} \), consensus ratio is defined as \( R = U/V \) where \( U = \left| \bigcap_{j=1}^{M} S_j^E \right| \) and \( V = \left| \bigcup_{j=1}^{M} S_j^E \right| \).
**Definition 2.** Given nonempty sets of ambiguous samples, $S_j^E, j \in \{1, 2, \ldots, M\}$, consensus ratio is defined as $R = U/V$ where $U = |\cap_{j=1}^{M} S_j^E|$ and $V = |\cup_{j=1}^{M} S_j^E|$.
Definition 2. Given nonempty sets of ambiguous samples, \( S_j^E, j \in \{1, 2, \cdots, M\} \), consensus ratio is defined as

\[
\mathcal{R} = \frac{U}{V} \quad \text{where} \quad U = |\bigcap_{j=1}^{M} S_j^E| \quad \text{and} \quad V = |\bigcup_{j=1}^{M} S_j^E| .
\]
Definition 2. Given nonempty sets of ambiguous samples, $S_j^E, j \in \{1, 2, \cdots, M\}$, consensus ratio is defined as $R = U/V$ where $U = |\cap_{j=1}^{M} S_j^E|$ and $V = |\cup_{j=1}^{M} S_j^E|$. 
Definition 2. Given nonempty sets of ambiguous samples, $S_j^E, j \in \{1, 2, \cdots, M\}$, consensus ratio is defined as $R = U/V$ where $U = | \bigcap_{j=1}^{M} S_j^E |$ and $V = | \bigcup_{j=1}^{M} S_j^E |$.
Definition 2. Given nonempty sets of ambiguous samples, \( S^E_j, j \in \{1, 2, \cdots, M\} \), consensus ratio is defined as \( R = U/V \) where \( U = |\cap_{j=1}^M S^E_j| \) and \( V = |\cup_{j=1}^M S^E_j| \).
Definition 2. Given nonempty sets of ambiguous samples, $S_j^E, j \in \{1, 2, \cdots, M\}$, consensus ratio is defined as $R = U/V$ where $U = |\cap_{j=1}^{M} S_j^E|$ and $V = |\cup_{j=1}^{M} S_j^E|$. If they overlap completely, then $U = V$ and $R = 1$. 
Definition 2. Given nonempty sets of ambiguous samples, \( S_j^E, j \in \{1, 2, \cdots, M\} \), consensus ratio is defined as
\[
\mathcal{R} = \frac{U}{V} \quad \text{where} \quad U = |\bigcap_{j=1}^{M} S_j^E| \quad \text{and} \quad V = |\bigcup_{j=1}^{M} S_j^E|.
\]

If they are independent, then \( U = 0 \), and \( \mathcal{R} = 0 \).

Low ratios: greater benefit from the consensus scheme
High ratios: algorithms perform the same, so less benefit
CoA Theory: Consensus Ratio

Consensus among three different AL algorithms

Plateau at 0.2: relatively little consensus
Motivates the use of ensemble approach
Definition 3. A sample $x_i \in X$ will be considered strongly ambiguous if $x \in \hat{S}^E = \bigcap_{j=1}^{M} S_j^E$; that is, if the sample is designated as ambiguous by $\Phi_j$, for all algorithms $j \in \{1, 2, \cdots, M\}$. 
Definition 3. A sample \( x_i \in X \) will be considered strongly ambiguous if \( x \in \hat{S}^E = \cap_{j=1}^M S_j^E \); that is, if the sample is designated as ambiguous by \( \Phi_j \), for all algorithms \( j \in \{1, 2, \ldots, M\} \).
Definition 3. A sample \( x_i \in X \) will be considered strongly ambiguous if \( x \in \hat{S}^E = \bigcap_{j=1}^{M} S_j^E \); that is, if the sample is designated as ambiguous by \( \Phi_j \), for all algorithms \( j \in \{1, 2, \cdots, M\} \).
Definition 3. A sample $x_i \in X$ will be considered strongly ambiguous if $x \in \hat{S}^E = \bigcap_{j=1}^{M} S_j^E$; that is, if the sample is designated as ambiguous by $\Phi_j$, for all algorithms $j \in \{1, 2, \cdots, M\}$.
CoA Theory: Strong Ambiguity

Definition 3. A sample \( x_i \in X \) will be considered strongly ambiguous if \( x \in \hat{S}^E = \cap_{j=1}^M S_j^E \); that is, if the sample is designated as ambiguous by \( \Phi_j \), for all algorithms \( j \in \{1, 2, \ldots, M\} \).
CoA Theory: Addition of Algorithms

- Proposition. As the number of algorithms being combined $M$ increases, the consensus ratio $R$ will either remain the same or will decrease.

- Analogy to a sieve: as you add more layers of filtering, fewer samples will “get through”.

- Remember: Small consensus ratio means better motivation for using consensus algorithm
Overview

- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Active Learning (AL) Overview

- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Combining different AL methods

- Theory of CoA

- Experimental Results

- Concluding Remarks
Evaluating the Training Set

- Training set evaluation: Probabilistic Boosting Tree
- Two medical image analysis databases:
  - Prostate cancer detection from histopathology
  - Breast cancer grading from histopathology
- Three training algorithms:
  - Query-By-Committee (QBC)
  - Bayes Likelihood (BAY)
  - Support Vector Machine Distance (SVM)
Experiment 1: Prostate Dataset

- Experiment 1 – Prostate Histopathology
  - 30 x 30 pixel grid on prostate biopsy samples
  - 14 texture features extracted from each ROI
  - 12,000 ROIs classified
Experiment 1: Prostate Dataset

Prostate Dataset - Accuracy

Prostate Dataset - AUC
Experiment 1: Eligible Sample Size

The diagram illustrates the average eligible sample set size for different methods, with error bars indicating variability. The methods compared are QBC AL, BAY AL, SVM AL, and CoA AL. The y-axis represents the size of eligible sample set, while the x-axis lists the methods. The bars show the comparison of these methods across different scenarios or conditions.
Experiment 2: Breast Grading

- Experiment 2 – Breast Histopathology Grading
  - 9,000 ROIs of homogeneous tissue (500 x 500 pixels)
  - Graph-based features to describe nuclear arrangement

Diagram:
- **Low-Grade**
  - Original Image
  - Delaunay Graph
  - Voronoi Graph
- **High-Grade**
  - Original Image
  - Delaunay Graph
  - Voronoi Graph
Experiment 2: Breast Grading

Breast Dataset - Accuracy

Breast Dataset - AUC
Experiment 2: Breast Grading

Average Eligible Sample Set Size

Size of Eligible Sample Set

QBC AL | BAY AL | SVM AL | CoA AL

Bar chart showing the average eligible sample set size for different algorithms.
Overview

- Classification and Training
  - Supervised Classification Paradigm
  - Building Training with Random Learning
  - Overview of Active Learning (AL)

- Extending Active Learning
  - Ambiguity as a measure of sample usefulness
  - Combining different AL methods

- Theory of CoA

- Experimental Results

- Concluding Remarks
Concluding Remarks

- **CoA**: Using multiple AL algorithms reduces set of informative samples, making annotation easier.
- Additional methods increase CoA benefit if the consensus ratio decreases (but is still $>0$).
- Generalizable to any supervised classification problem where:
  - Data are costly, difficult to annotate
  - Target class is complex, RL requires many samples
  - Multiple AL algorithms can be leveraged simultaneously.
Acknowledgments

- Wallace H. Coulter Foundation
- New Jersey Commission on Cancer Research
- National Cancer Institute
  - R01CA140772-01
  - R01CA136535-01
  - R21CA127186-01
  - R03CA143991-01
- Cancer Institute of New Jersey
- US Department of Defense (W81XWH-08-1-0145)
- Bioimagene, Inc.