Bayesian Nonexhaustive Learning for Online Discovery and Modeling of Emerging Classes

Murat Dundar\textsuperscript{a}
Assistant Professor

in collaboration with
Ferit Akova\textsuperscript{a}, Yuan Qi\textsuperscript{b}, Bartek Rajwa\textsuperscript{b}

\textsuperscript{a} Computer & Information Science Department, Indiana University – Purdue University, Indianapolis (IUPUI)
\textsuperscript{b} Computer Science Department, Purdue University, West Lafayette, IN
\textsuperscript{c} Discovery Park, Purdue University, West Lafayette, IN
Problem Statement

• A training data is not representative if the list of classes is not complete, i.e., non-exhaustive

• Samples of unrepresentative, i.e., unobserved classes will be misclassified (into one of the observed classes) with a probability one.

Non-exhaustive list of classes $\rightarrow$ Ill-defined classification problem!
What contributes to non-exhaustiveness?

- Some of the classes may not be in existence at the time of training
- Classes may exist but may not be known
- Existence of classes may be known but samples are simply not obtainable

Is this a real problem?
Applications from Information Retrieval

• Topical classification of published literature, news articles, documents, web sites etc.
• Image annotation
• Object categorization
• ...

New classes emerge on a continuous basis!
Bio-detection and Bio-surveillance

- Classify microorganisms based on their phenotypes
- An exhaustive training library of microorganisms is not practical
  - High mutation rate new classes can emerge anytime
  - Tested samples are from most prevalent classes

Our research is driven by Pathogen Detection
Simple optical interrogation can generate unique phenotypic fingerprints

(A) *Listeria monocytogenes* 7644, (B) *E. coli* ETEC O25, (C) *Staphylococcus aureus* P103, (D) *Vibrio cholerae* O1E
Non-exhaustive Learning

- Operates in a multiclass setting with observed and unobserved classes
- Evaluates new samples sequentially
- Performs online inference in the presence of labeled and unlabeled samples
- Identifies “interesting” class formations for follow-up analysis

Main Contributions
Our Approach in a Nutshell

• A Dirichlet process prior (DPP) for online modeling of classes
• Normal data model with a bivariate Normal x Inverted Wishart prior
• Sequential importance sampling-resampling (SISR) approach to perform online inference
Some Concepts

• Observed classes: classes initially available in the training library. These are verified classes.
• Unobserved classes: classes not represented in the training library
• New classes: classes discovered online. These are unverified classes.
• A new class could be a super or a sub-class of an unobserved class
Our Notation

Training Library: \( D = \{x_i, y_i\}_{i=1}^{n^\ell}, \quad x_i \in X = \mathbb{R}^d, \quad y_i \in Y = \{1, \ldots, J\} \)

\( n^\ell \) : total number of labeled samples

\( n^\ell_j \) : number of samples in class \( j \)

J: number of observed classes

\( p(\cdot | \phi_j), \quad \phi = \{\phi_j\}_{j=1}^{J} \)

Defining a DPP over classes is equivalent to modeling the prior distribution of \( \theta_i, \quad \theta_i \in \phi \), by a Dirichlet process.
Prediction with Observed Classes

- An incoming sample \( x \)
- Does \( x \) belongs to one of the observed classes or to a new one?
- Evaluate the posterior:

\[
p(y | x, y, x) \propto \frac{\alpha}{\alpha + n^\ell} p(x) \delta_{j+1} + \frac{1}{\alpha + n^\ell} \sum_{j=1}^{J} n_j^\ell p(x | \phi_j) \delta_j
\]

\( x = \{x_i\}_{i=1}^{n^\ell} \): the set of all training samples

\( y = \{y_i\}_{i=1}^{n^\ell} \): the corresponding set of observed labels.

An easy task when performed with observed classes only
Prediction with Observed & New Classes

• Two types of samples available during online execution:
  • samples in the training dataset with known labels
  • samples observed online with temporary class labels

\[ \tilde{x}^{n''} = \{\tilde{x}_i\}_{i=1}^{n''} : \text{the set of } n'' \text{ samples sequentially observed online} \]
\[ \tilde{y}^{n''} = \{\tilde{y}_i\}_{i=1}^{n''} : \text{the corresponding set of unknown class labels} \]
\[ \tilde{y}_i \in \{1, \ldots, \tilde{J} + J\} \]
\[ \tilde{J} \text{ being the number of new classes associated with these } n'' \text{ samples.} \]
Inference in Online Setting

- Sequential Importance Sampling Resampling
- Also known as particle filters
- At any given time, the sampler only depends on a set of particles and their corresponding weights
- Weights are efficiently updated in a sequential manner each time a new sample is observed without the need for the past samples.

Efficient and highly scalable online inference
Particle Filters

- Approximate the posterior mean by a discrete summation of weighted particles
- Particles are sampled from an importance function
- Importance Function: \( q(\tilde{y}^{n+1} | \tilde{x}^{n+1}, x, y) \)

\[
E_{p(\tilde{y}^{n+1} | \tilde{x}^{n+1}, x, y)} \left[ \tilde{y}^{n+1} \right] = \int_{\tilde{y}^{n+1}} \int_{\tilde{x}^{n+1}, x, y} p(\tilde{y}^{n+1} | \tilde{x}^{n+1}, x, y) \, d\tilde{y}^{n+1}
= \int_{\tilde{y}^{n+1}} \int_{\tilde{x}^{n+1}, x, y} w_{n+1}(\tilde{y}^{n+1}) q(\tilde{y}^{n+1} | \tilde{x}^{n+1}, x, y) \, d\tilde{y}^{n+1}
\approx \sum_{m=1}^{M} \tilde{y}^{n+1}_m w^{n+1}_m (\tilde{y}^{n+1}_m) \delta_{\tilde{y}^{n+1}_m} \]
Sampling and Resampling

Every time a new sample is observed:

• First $R$ new particles are sampled for each of the $M$ particle using the importance function
• Then weights are updated for all $M \times R$ particles
• Finally resampling, stratified on the particle weights, is performed to select $M$ particles out of $M \times R$ ones.
Data Model

\[p(x | \phi_j) = N(\mu_j, \Sigma_j), \quad \phi_j = \{\mu_j, \Sigma_j\}\]

\[G_0 = p(\mu, \Sigma) = N\left(\mu | \mu_0, \frac{\Sigma}{\kappa}\right) \times W^{-1}(\Sigma | \Sigma_0, m)\]

\[p(\mu | \Sigma)\]

\[p(\Sigma)\]

\[\mu_0 : \text{the prior mean}\]
\[\kappa : \text{a scaling constant that controls the deviation of the mean vectors from the prior mean.}\]
\[\Sigma_0 : \text{encodes our prior belief about the expected } \Sigma.\]
\[m : \text{is a scalar that is negatively correlated with the degrees of freedom. In other words the}\]
\[\text{larger the } m \text{ is the less } \Sigma \text{ will deviate from } \Sigma_0 \text{ and vice versa.}\]

Normally distributed classes

A Bivariate Normal -Inverted Wishart G_0
Parameter Estimation

- The parameters $\Sigma_0$, $m$, $\mu_0$, $\kappa$ can be estimated from the data by Empirical Bayes.
- The parameter $\alpha$ can be estimated by Empirical Bayes to encode our prior belief about the odds of encountering a new class.
An Illustration

- 23 Normally-distributed classes are generated according to our data model discussed previously.
- 3 out of 23 classes are randomly chosen and considered unrepresented.
- The training set contains 20 classes (2000 samples), the test set contains 23 classes (2300 samples).
- Test samples are sequentially evaluated using the proposed approach.
(a) True class distributions
(b) 100/2300 classified
(c) 300/2300 classified
(d) 2300/2300 classified
A Biodetection Application

• A total of 2054 samples from 28 classes each representing a different bacteria serovar were considered in this study.
• These are the type of serovars most commonly found in food samples.
• Each serovar is represented by between 40 to 100 samples.
• Samples are the forward-scatter patterns characterizing the phenotype of a bacterial colony obtained by illuminating the colony surface by a laser light.
• Each scatter pattern is a gray level image characterized by a set of 50 features.
Samples are randomly split into two as train and test, with 70% of the samples going into the training set and the remaining 30% in the test.

Four of the classes are considered unknown and all of their samples are moved from the training set to the test set.

Training set contains 24 classes, testing set contains 28 classes
Evaluation

- The performance of the proposed algorithm is evaluated on three fronts:
  - classification accuracy for observed classes
  - classification accuracy for unobserved classes
  - number of clusters discovered for each of the unobserved class
- Results are compared against the exhaustive scenario (when both train and test sets were exhaustively defined) as well as a previously developed technique (Bayes-NoDe)
## Results

<table>
<thead>
<tr>
<th></th>
<th>observed classes (overall accuracy)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhaustive case</td>
<td>accuracy (%)</td>
<td>94.0</td>
<td>88.5</td>
<td>94.1</td>
<td>100.0</td>
</tr>
<tr>
<td>Bayes-NoDe</td>
<td>accuracy (%)</td>
<td>91.2</td>
<td>60.5</td>
<td>82.8</td>
<td>87.6</td>
</tr>
<tr>
<td></td>
<td>(0.2)</td>
<td>(3.6)</td>
<td>(13.4)</td>
<td>(8.1)</td>
<td>(3.4)</td>
</tr>
<tr>
<td>NEL-SIR</td>
<td>avg. # of clusters</td>
<td>-</td>
<td>2.7</td>
<td>5.5</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>accuracy (%)</td>
<td>91.1</td>
<td>75.7</td>
<td>90.7</td>
<td>99.5</td>
</tr>
<tr>
<td></td>
<td>(0.4)</td>
<td>(1.2)</td>
<td>(8.0)</td>
<td>(0.5)</td>
<td>(0.7)</td>
</tr>
<tr>
<td></td>
<td>avg. # of clusters</td>
<td>-</td>
<td>2.0</td>
<td>1.3</td>
<td>1.7</td>
</tr>
</tbody>
</table>

- All four unobserved classes are successfully discovered
- Reasonable number of clusters generated per class
- Performs equally well with the exhaustive case for well-separated classes
- When there is overlap between classes some loss in classifier accuracy as compared to the exhaustive case is expected
Conclusions

- Exhaustive training set is not realistic for many real-world datasets
- A new approach to online learning with observed and unobserved classes is introduced
- Promising results are obtained with respect to classification accuracy and class discovery on an important biodetection problem.