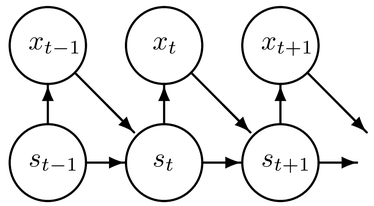
# Semisupervised Classification of Texts Using Particle Learning for Probabilistic Automata

Ana Paula Sales, Christopher Challis, Ryan Prenger, and Dan Merl
Computational Engineering Division

## Abstract

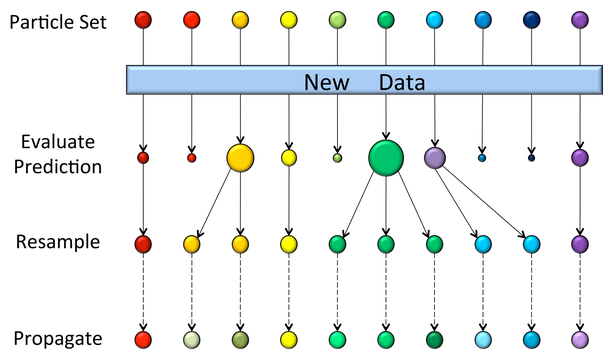
Some of the key statistical problems underlying many current national security applications stem from a common need for continuously deployable, self-adapting learning systems. This is especially true for adversarial classification problems, one example of which is spam detection, where intuitively the boundary between the positive and negative classes is constantly evolving as the adversary improves its attack vector. Here, we present a particle learning approach for efficient online estimation of simple text grammars as represented by probabilistic automata, and the development of a semisupervised classification system based on a flexible class of composite mixture models.

## Probabilistic Automata




- Strict subclass of HMMs: state-space model characterized by probabilistically generated observations and deterministic state transitions conditional on both the current state and the current observation
- Potentially computationally attractive substitute for HMMs for some natural language applications

## Particle Learning

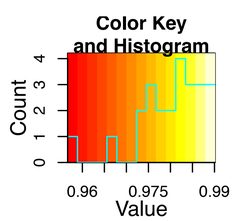


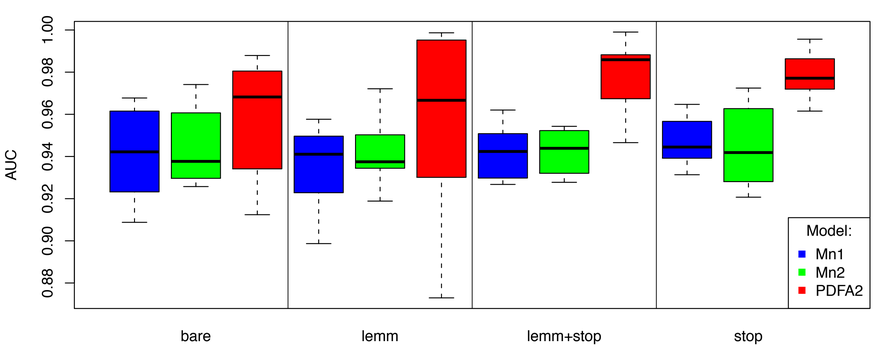

- Stochastic "survival of the fittest" algorithm
- Smoothing and learning of streaming data on the fly
- No need to store entire dataset on memory at once
- Full Bayesian posterior inference
- Highly parallelizable
- Automatic anomaly detection

## Case Study: Email Spam Detection

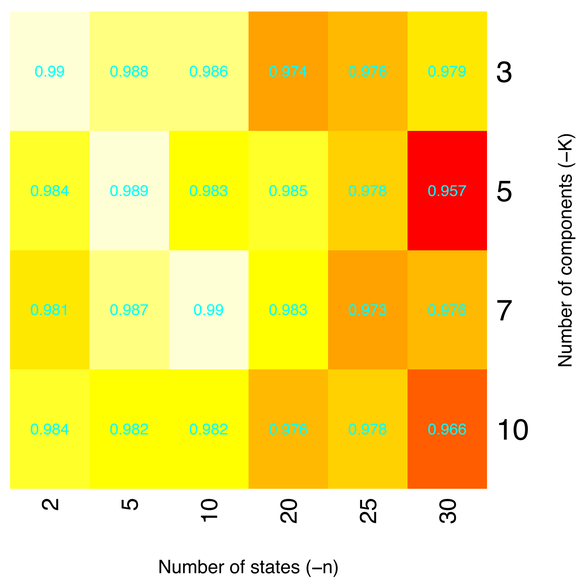
- We demonstrate the utility of this approach in the context of email classification into spam or not spam, using the public email corpus PU1.
- PU1 consists of 481 spam and 618 legitimate email messages.
- PU1 is available in 4 versions, obtained by enabling or disabling a lemmatizer and a stop list.
- We compare the predictive performance of three particle filter models, summarized in the table below:

| Model | Email Subject | Email body | Distribution |
|---|---|---|---|
| Mn1 | yes | no | Multinomial |
| Mn2 | yes | yes | Multinomial |
| PDFA2 | yes | yes | PDF |




Distribution of AUC measurements obtained from a 10 fold cross validation trial to compare the performance of models Mn1, Mn2, and PDFA2.




Sensitivity analysis of the number of components in the mixtures and the number of states in the PDFA using the lemm+stop version of the PU1 dataset and including only the subject of the email messages. The colors denote the mean AUC from 10-fold cross validation trial, and the number within the cells denote one standard deviation from the mean. The color key to the left provides a mapping of the color to AUC value, and a histogram showing the of AUC values.

## How Can This Method Be Useful to Your Problem?

- Implemented in high-performance C++
- Seamless and straightforward instantiation of new models
  → The composite mixture model structure allows specification of a joint probability model for heterogeneous collection of independent variables without requiring complex embeddings.
- Specification of arbitrary configurations of exponential family distributions
- Currently implemented distributions: Gaussian, exponential, Poisson, beta, multinomial, and PDFA
- Our general approach is flexible and applicable to a vast array of domains, such as real-time threat detection, real-time video surveillance, real-time energy distribution, and bio-surveillance.
- Applicable to regression and multiclass classification problems
- Automatic anomaly detection

## It's Only Getting Better and Faster

We are currently pursing the following improvements to this framework:

- Parallelization of particle filters to increase computational performance
- Investigation of the computational vs. predictive tradeoff
- Nonstationary mixtures and drift detection
- Adaptive sampling for increased performance
  - Data down-sampling
  - Data distribution among parallel particle filters

## Contact

Ana Paula Sales apsales@llnl.gov (925) 423-8358

Lawrence Livermore National Laboratory

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-POST-557920