Finding Meaningful Gaps to Guide Data Acquisition for a Radiation Adjudication System

Nicholas Gisolfi, Madalina Fiterau, Artur Dubrawski
Auton Lab, Carnegie Mellon University, Pittsburgh, PA

Motivation

- Machine learning methodologies for radiation threat detection.
- Training data generated synthetically because too few true threats are observed in the field.
- High dimensional data is prone to omissions of meaningful information.
- We aim to provide a framework which presents insufficiencies of training data in a user-friendly manner, allowing data engineers to inject data needed to fill gaps in the feature space.

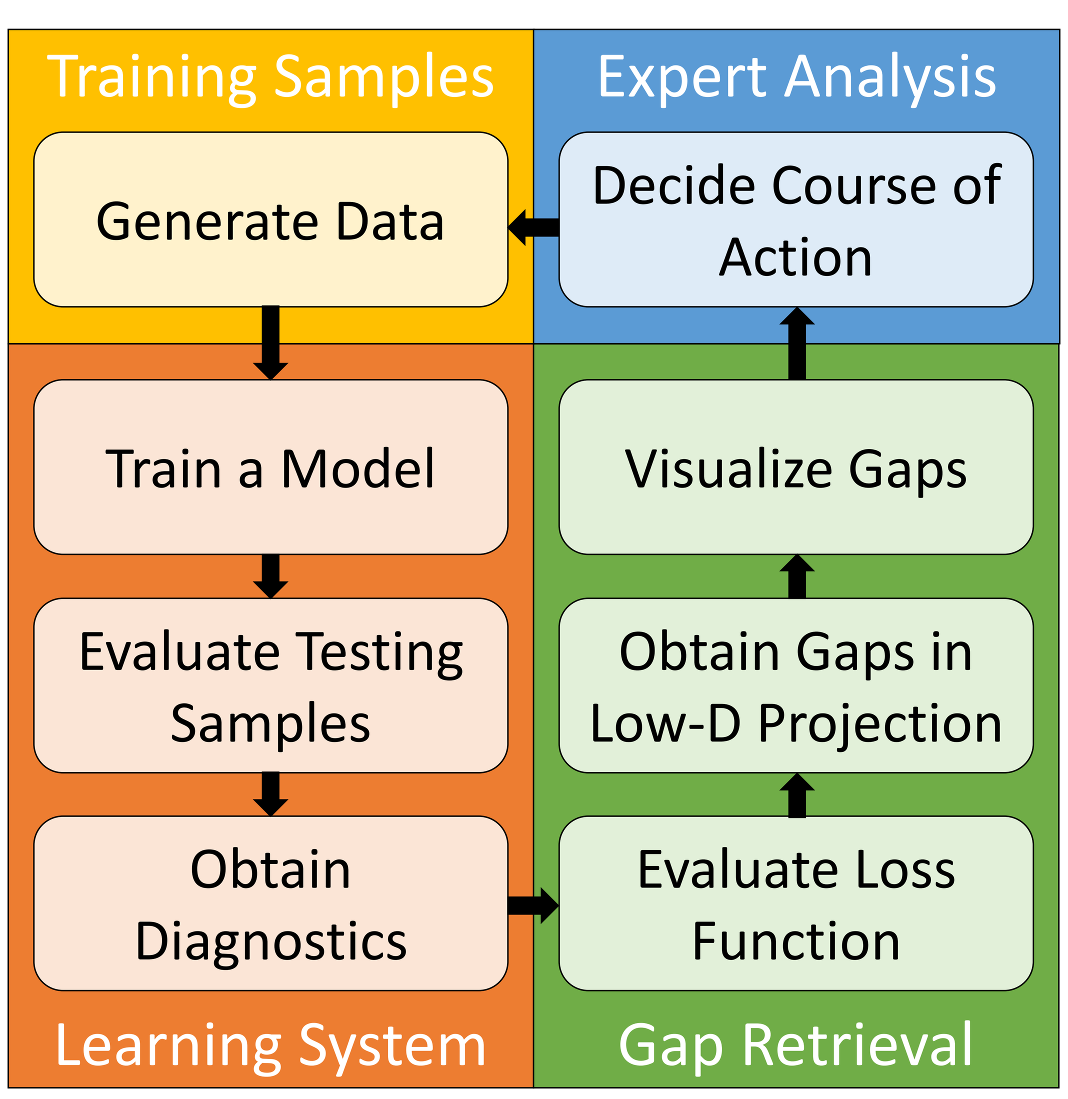
Data

- 114 Features
- Over 50K Samples
- Semi-Synthetic
- Multiple Folds

Classes include:

1. Non-Emitting sources
2. Emitting sources posing a threat
3. Emitting sources explainable by naturally occurring radioactive materials

Iterative Build Process



Learning System

Contains a learner and an evaluation procedure which characterizes performance diagnostics on the test data.

Gap Retrieval System

Finds low-dimensional projections where the testing and training data differ significantly, or the performance diagnostics indicate considerable loss of accuracy.

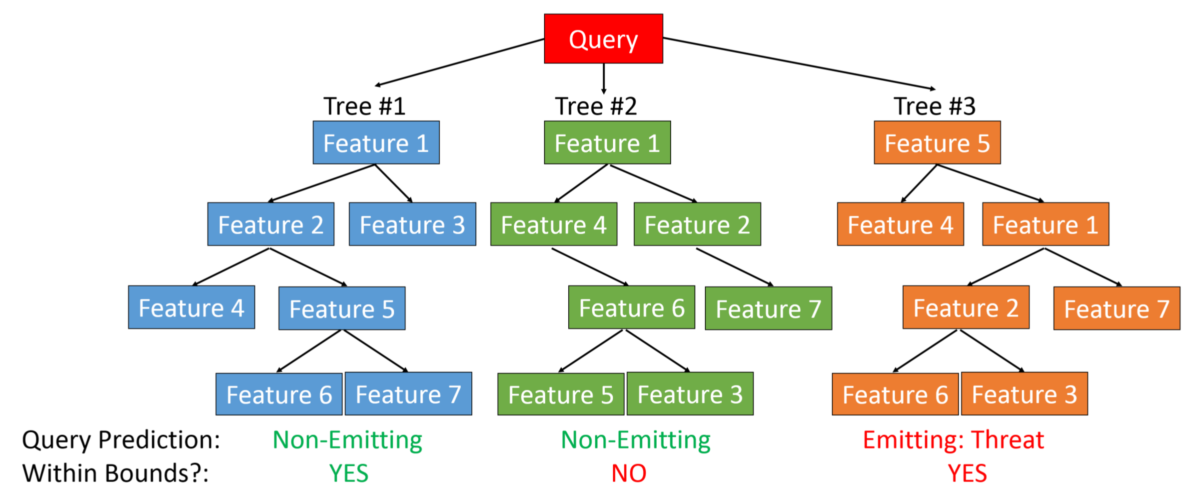
Expert Analysis of Training Data

Experts gain intuition for what data may be missing from the training set and decide which parts of the feature space would most benefit from additional samples. The training samples in the next iteration will reflect these changes.

Acknowledgements: This work has been partially supported by the DHS/DNDO under competitively awarded Exploratory Research Project HSHQDC-12-X-00218 and grant 2010-DN-077-ARI040-02, and by the NSF under awards 0911032 and 1320347. This support does not constitute an express or implied endorsement on the part of the US Government. Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the DoE/NNSA under Contract DE-AC52-07NA27344.

Model – Random Forest

Build random forest using k-fold cross validation which admits diagnostics



Diagnostic 1 – Agreement Score

- Describes the extent to which predictions made by all trees agree.
- Optimally, all trees in the forest reaching the same classification label for a given sample.

Diagnostic 2 – Inbounds Score

- Quantifies whether or not a query falls within a range of values that has been observed by a tree in the random forest during training.
- Optimally, all trees have seen a sample of similar feature values during training.

Gap Retrieval System - RIPR

Regression-Based Informative Projection Recovery algorithm searches subspaces to find projections where data is most separable

Overview of Algorithm

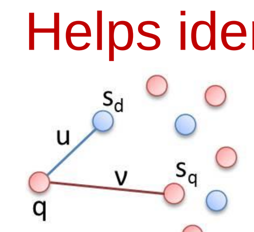
for each 2D subspace in the feature space
- Train classification model

for sample in training set
- evaluate loss function

Associate each point with ideal projection
Visualize most populated projections

Non-Parametric Loss Estimator

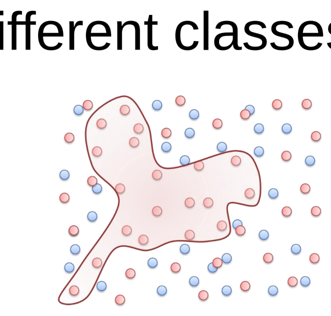
- Ratio of distances between a query sample and samples of similar and different classes
- Helps identify irregular gaps



Visualization of non-parametric formulation

$$\hat{l}_q = N\frac{v}{u}$$

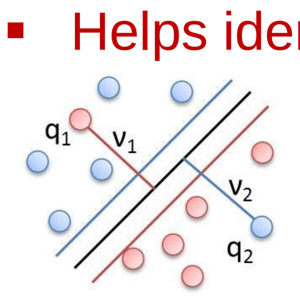
Point-Wise Loss Function



Non-Parametric loss estimation reveals irregular gaps in data

Parametric Loss Estimator

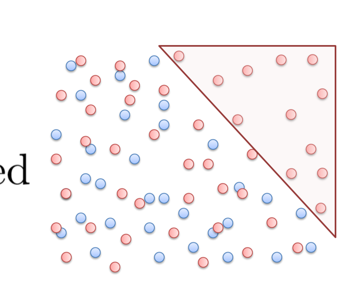
- Distance to a decision boundary
- Helps identify structured gaps



Visualization of parametric formulation

$$\hat{l}_q = \begin{cases} ||v|| & \text{if } q \text{ mislabeled} \\ 0 & \text{else} \end{cases}$$

Point-Wise Loss Function



Parametric loss estimation reveals structured gaps in data

Experimental Objective

Direct Gap Finding

Finds density mismatches between two sets of data. Predict which set a sample belongs to.

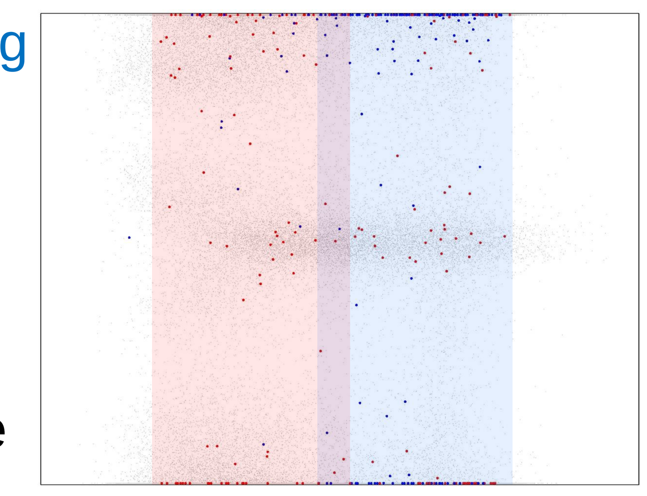
Diagnostic Gap Finding

Finds areas where predictions are confident. Predict the confidence of the classification prediction for each point.

Experiment Results

Non-Parametric, Direct Gap Finding

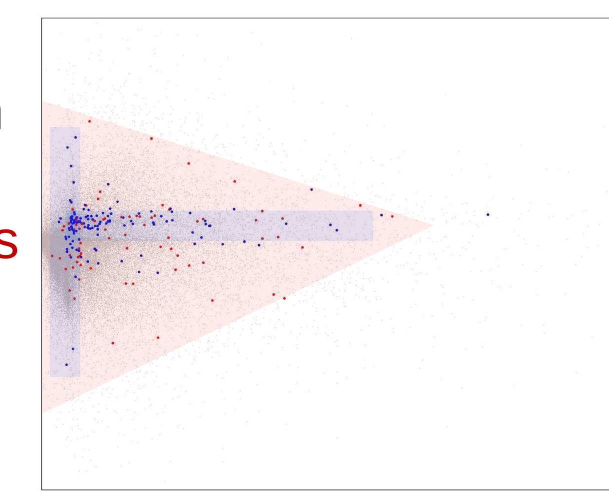
- Distribution of testing samples are shifted from training samples
- Due to changing a single coefficient between successive data builds



Blue points come from testing set
Red points come from training set

Non-Parametric, Diagnostic Gap Finding

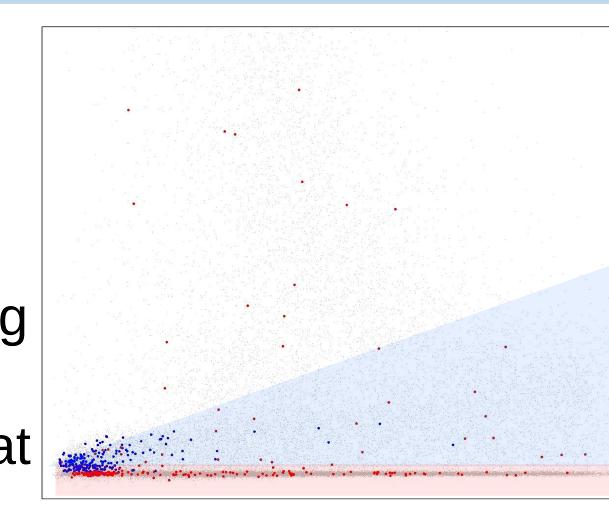
- Most confident predictions reside in T-shape while less confident predictions reside outside this region
- Recovered irregular shaped gap in data



Blue points in total agreement between trees
Red points indicate non-uniform consensus

Parametric, Direct Gap Finding

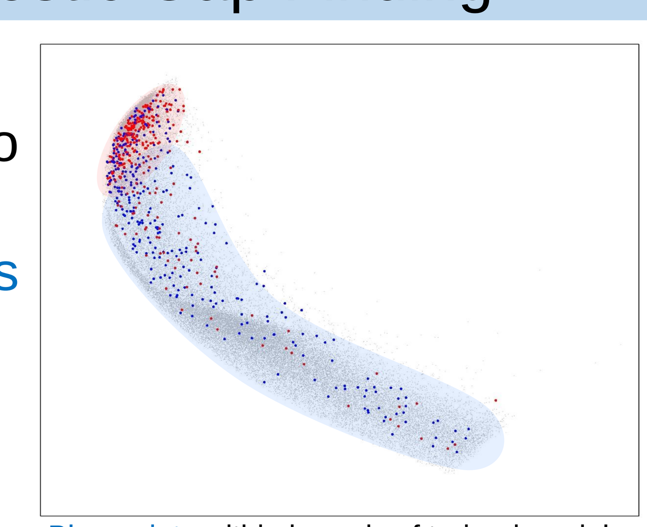
- A linear bound separates samples from testing set and training set.
- Distribution of testing samples differs significantly from that of training samples



Blue points come from testing set
Red points come from training set

Parametric, Diagnostic Gap Finding

- Less confident predictions cluster to a small region while confident predictions are spread.
- This region is easy to interpret by data engineers



Blue points within bounds of trained model
Red points outside bounds of trained model

Effect of Filling Gaps on Model Accuracy

Baseline With Gap	Add Data	Fill the Gap
75.0%	75.2%	75.7%

- Targeting gaps with additional data boosts model accuracy more efficiently than adding samples which may or may not cover the gap.

Conclusions

- Visualizations allow engineers to make changes necessary to improve synthetic data generation.
- By resolving gaps in training data, model classification performance improves.
- Nonparametric loss function finds irregular gaps.
- Parametric loss function reveals structured gaps in data, allowing users to identify adjustments in data generation that will improve accuracy.

References

[1] Artur Dubrawski, Saswati Ray, Peter Huggins, Simon Labov, and Karl Nelson. Diagnosing Machine Learning-Based Nuclear Evaluation System. In *Proceedings of the IEEE Nuclear Science Symposium*, 2012.
[2] Madalina Fiterau and Artur Dubrawski. Informative projection recovery for classification, clustering and regression. In *International Conference on Machine Learning and Applications*, Volume 12, 2013.