Finding Gaps in Data to Guide Development of a Radiation Threat Adjudication System

Nicholas Gisolfi[1], Madalina Fiterau[1], Artur Dubrawski[1], Saswati Ray[1], Simon Labov[2], Karl Nelson[2]

[1]Auton Lab, Carnegie Mellon University, Pittsburgh, PA [2]Lawence Livermore National Laboratory, Livermore, CA

Motivation

Modeling nuclear threats in synthetic data challenges human data engineers to include all relevant niches of the feature space. High-dimensional synthetic data is prone to omissions of meaningful information which may be used to improve a trained model's accuracy at a classification task.

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.

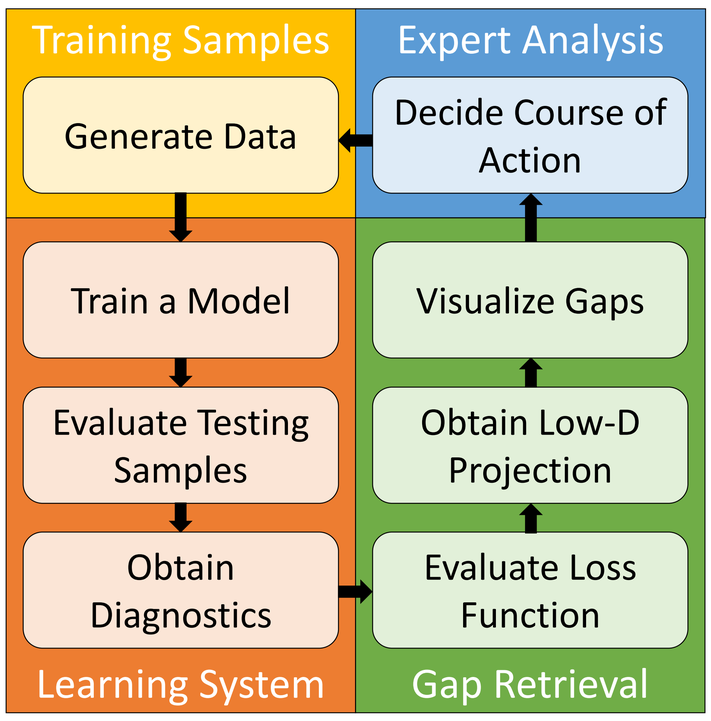
Data

- Multiple Classes
- 113 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

The Iterative Build Process



Learning System

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

Gap Retrieval System

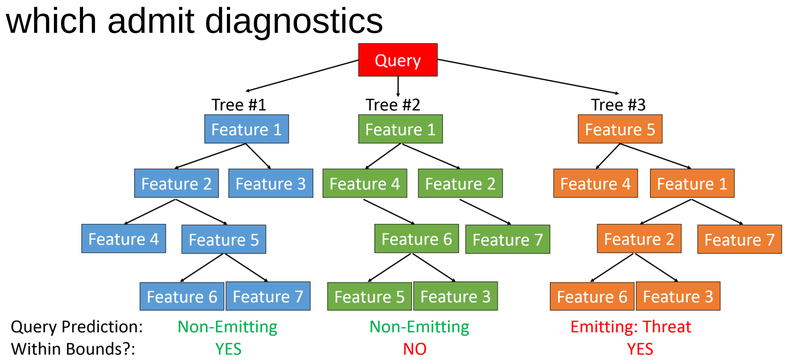
Find low-dimensional projections where the testing and training data differ significantly, or the performance diagnostics indicate considerable loss of accuracy.

Expert analysis of training data

Experts gain intuition as to 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 and the process continues until the training samples are a faithful representation of the test set.

Learning System – Random Forest

Build random forest using k-fold cross validation which admit 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.

Regression Based Informative Projection Recovery

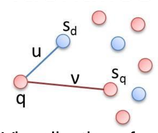
Search subspaces to find projections where data is most separable

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

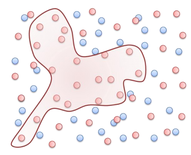
- Ratio of distances between a query sample and samples of similar and different classes
- Identifies irregular gaps



Visualization of non-parametric formulation

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

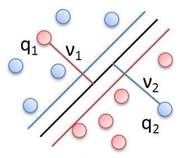
Point-Wise Loss Function



Non-Parametric loss estimation reveals irregular gaps in data

Parametric Loss Estimator

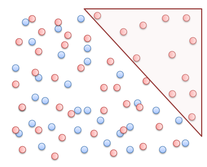
- Distance to a decision boundary
- Identifies structured gaps



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

Used for finding density mismatches between two sets of data. Predict which set a sample belongs to.

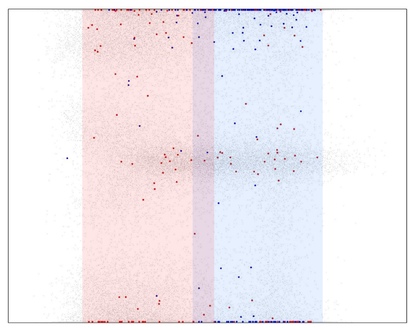
Diagnostic Gap Finding

Used to determine areas where predictions are confident or not. Predict the confidence of the trained model at each point.

Experiment Results

Non-Parametric, Direct Gap Finding

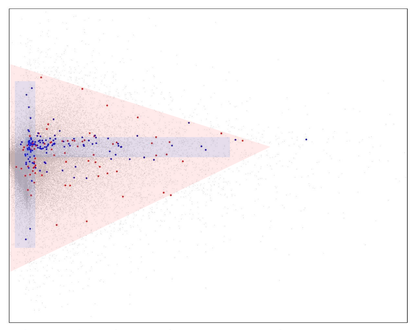
- Distribution of testing samples are shifted from training samples
- Due to changing a single coefficient between successive data builds



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

Non-Parametric, Diagnostic Gap Finding

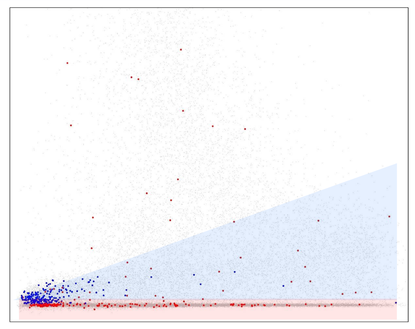
- Most confident predictions reside in T-shape while less confident predictions reside outside this region
- Recovered irregular shaped gap in data



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

Parametric, Direct Gap Finding

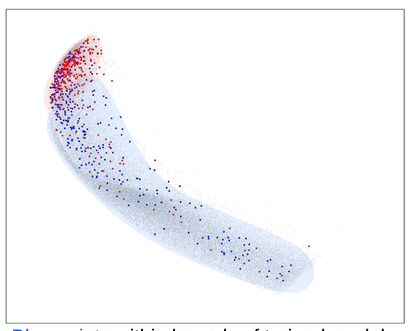
- A linear bound separates samples from testing set and training set.
- Distribution of testing samples differs significantly from that of training samples



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

Conclusions

- Framework generates visualizations which allow engineers to make changes necessary to improve synthetic data generation.
- By resolving gaps in training data, model classification performance may improve.
- Nonparametric loss function finds irregular gaps.
- Parametric loss approach reveals structured gaps and divides in the data, allowing users to easily identify adjustments to data generation that will improve model 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.