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

N. Gisolfi*, M. Fiterau*, A. Dubrawski*, S. Ray*, S. Labov**, K. Nelson**,
* Auton Lab, Carnegie Mellon University, ** Lawrence Livermore National Laboratory

Problem Formulation and Approach. We consider an incident classification task in a radiation threat detection and adjudication system. As vehicles travel across international borders, they may be scanned for sources of harmful radiation, such as improperly contained medical or industrial isotopes, or nuclear devices. A substantial number of potential threats flagged by radiation measurement devices that may be used in such applications are actually non-threatening artifacts due to naturally occurring radioactive materials (e.g. ceramics, marble, cat litter, or potash). We have been using machine learning methodology to dismiss alerts that are confidently explainable by non-threatening natural causes, without increasing the risk of neglecting actual threat [1].

A robust alert adjudication system must be trained and validated on data that includes the actual threats. However, such data is (luckily) hard to come by. Therefore, it is practical and common to place the bulk of the available empirically gathered positive incident examples into a testing data set, and create training data using benign measurements mixed with a carefully chosen selection of simulated threat. Nonetheless, the volumes and complexities of the feature space in data typically encountered in radiation measurement applications makes synthesising a robust, sufficiently large, and (most importantly) comprehensive set of training data difficult and prone to omissions.

We present an engineering framework that facilitates data quality audits by automatically detecting gaps in training data coverage. It highlights areas of discrepancy between training and testing samples. It also pinpoints the areas of feature space where the observed performance of the threat adjudication system appears suboptimal. The findings are presented in the form of human-readable, low-dimensional projections of data, in order to ensure interpretability of results and to simplify planning of corrective actions. The resulting iterative data improvement procedure boosts threat adjudication accuracy while reducing the required workload of data engineers and application domain experts, when compared to using uninformed data gathering process.

The proposed process involves: (1) Building a threat classifier (any plausible type of a classification model can be used, we employ the random forest method primarily due to its scalability to highly-dimensional feature spaces, but also because of the computable on-the-fly metrics that diagnose reliability of predictions being made, which it provides), (2) Gap Retrieval Module (GRM), and (3) Human-driven procedure of addressing the identified gaps. Of particular relevance are two metrics that attempt to characterize reliability of predictions made by our random forest classifier: Dot-Product-Sum (DPS) and In-Bounds Score (IBS). DPS measures consistency of predictions made independently by the individual trees in the forest. IBS is perfect if for each node of a classification tree, the query fits within the range of the bounding box of the training data. Otherwise, it returns the value proportional to the fraction of nodes where the query was in-bounds. GRM identifies where the original threat classification model performs well and where it performs poorly. It does this in one of two ways: (a) By finding low-dimensional projections where the testing and training data distributions differ significantly, and (b) By finding low-dimensional regions of data space where the original classifier experiences considerably low accuracy. The GRM leverages a previously published algorithm called Regression for Informative Projection Retrieval (RIPR) [2]. This algorithm discovers a small set of low-dimensional projections of possibly highly multivariate data which reveal specific low-dimensional structures in data, if such structures exist. RIPR's primary application is to improve understandability of classification, regression, or clustering tasks by explaining their results in a human-readable form. Here, we primarily leverage its ability to detect low-dimensional patterns of unexpected discrepancy between training and testing data, as well as low-dimensional structures of low performance areas, in order to facilitate improvements in training data generation. As a result of executing GRM, the resulting low-dimensional subspaces are visualized and the domain experts and data engineers 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 expanded training data will reflect these changes in the next machine learning iteration, and the process can continue until the training set is shaped into a faithful reflection of the test set, and the performance of the threat adjudication system is optimized.

Experimental Results. To find data gaps directly, our algorithm simply looks for mismatches between the training and testing data distributions in all 2 or 3 dimensional projections of data, to enable visually interpretable output. In this scenario, the algorithm returns the most prominent gap, even if it is located in a projection that yields relatively little information to support model predictions. Our results show that GRM is able to identify potentially irregularly shaped areas of mismatch between the training and test sets. The set up of our experiments involves the selection of two random samples: one of an arbitrary number of data points in the training set composed of semi-synthetic

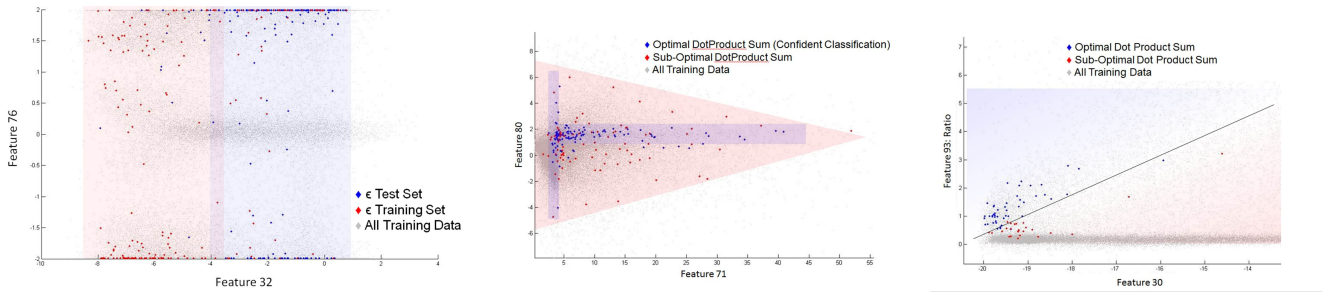

Figure 1: Example projections retrieved using direct (left), and diagnostic nonparametric (middle) and parametric approaches.

data and another similarly sized sample of data from the testing set. By taking these uniformly random samples, any mismatch we find is representative of the entire dataset with high probability, as the process does not change the training and testing data distributions. The leftmost graph in Figure 1 shows an overlay mismatch where the test set seems to simply be a shifted version of the training set. After conferring with the data engineers who built the data, we determined that the cause of the overlap is actually a single scalar parameter that was changed between two successive artificial data builds. Visualization provided by our framework allows data engineers to easily gather succinct information about the variations of the underlying structure of data.

Next, we tied in a cost function that determines which gaps are more meaningful in terms of the impact they may have on the threat classification performance. We can achieve this by incorporating diagnostic measures resulting from the original classifier performance evaluation, as observed on the test samples. The middle graph in Figure 1 shows a projection retrieved using a nonparametric loss estimator. We see that our random forest makes the most confident predictions (high DPS) for blue points which occupy a densely packed T-shaped space in the projection. Red points, which correspond to predictions which were not fully consistent among the trees in the forest (low DPS), indicate test data which may benefit from additional nearby training samples. They are far more spread out within the projection, and often reside near the edges of the gray point cloud which represents all of the training data.

Humans are good at understanding how to fill in gaps in low dimensional projections that retain some sort of a regular structure (i.e. a box or triangle), which is why we also devised a parametric loss estimator. It enables extraction of projections that contain regularly-shaped gaps which may cause considerable loss of threat classification performance. In the rightmost graph in Figure 1, we use linear Support Vector Machine model to separate high- and low-performance areas. Our goal here is to find projections of data where misclassified queries occupy one side of the classification boundary, while correctly classified queries occupy the other side. This is a useful type of a gap to look at because it identifies sets of features that jointly emphasize a controversy on how test data should be classified.

To prove our framework inceases model accuracy, we train random forest models using different subsets of training data. We start by taking our original data set and removing samples which fall within a certain region of a 2D projection, thus creating an artifical hole in the data. The random forest trained from this data set achieves 75.0% classification accuracy. We then run RIPR which identifies this gap and we add excluded samples back to the training set, which fills the gap that RIPR identified. Now training a new random forest, we achieve 75.7% accuracy. This shows we are able to improve model performance by filling in gaps that the GRM identifies.

Conclusion. We presented a framework by which data engineers and application domain experts can identify shortcomings in their semi-synthetic data, so that they could make effective changes when collecting and generating additional data in an iterative build process, as well as glean insights regarding the underlying structure of high-dimensional feature spaces. The approach has been successfully used to boost performance of a radiation threat adjudication system.

Acknowledgements. This work has been supported by the U.S. Department of Homeland Security, Domestic Nuclear Detection Office, under competitively awarded contract/IAA HSHQDC-12-X-00218, and by the National Science Foundation under awards 0911032 and 1320347. This support does not constitute an express or implied endorsement on the part of the Government. Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344.

[1] Artur Dubrawski, Saswati Ray, Peter Huggins, Simon Labov, and K 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.