Computer Experiments for Meta-learning of Machine Learning Models

Anna Flowers, Justin Krometis, Robert B. Gramacy, Christopher T. Franck
DATAWorks 2025

MOTIVATION

It is desirable to learn about a model's **operating envelope,** the range of values for which the model performs well. The operating envelope can be determined by answering:

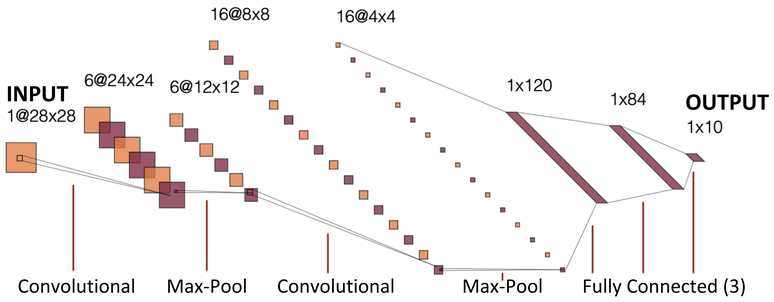
- What does the model get right?
- What does the model get wrong?
- What outcomes have a high degree of uncertainty?
- How might the training stage differ from the operational environment?

We would like to create a methodology to estimate a model's operating envelope at each phase of testing and collect data accordingly. This has been studied for relatively simple models, but we wish to extend it to more complex Machine Learning models.

MODELS: GAUSSIAN PROCESSES AND NEURAL NETWORKS

We utilize a neural network with the following structure:

Neural Network – LeNet-5



Let X_r be a d-dimensional matrix of inputs and Y_r be a corresponding vector of outputs. Assuming a traditional **Gaussian Process (GP)** prior implies that $Y_r \sim \mathcal{N}\left(0, \Sigma(X_r)\right)$, where $\Sigma(X_r)^{ij} = \Sigma\left(x_i, x_j\right) = \tau^2 \exp\left\{-\sum_{k=1}^{d} \frac{\|x_{ik} - x_{jk}\|^2}{\theta_k}\right\}$.

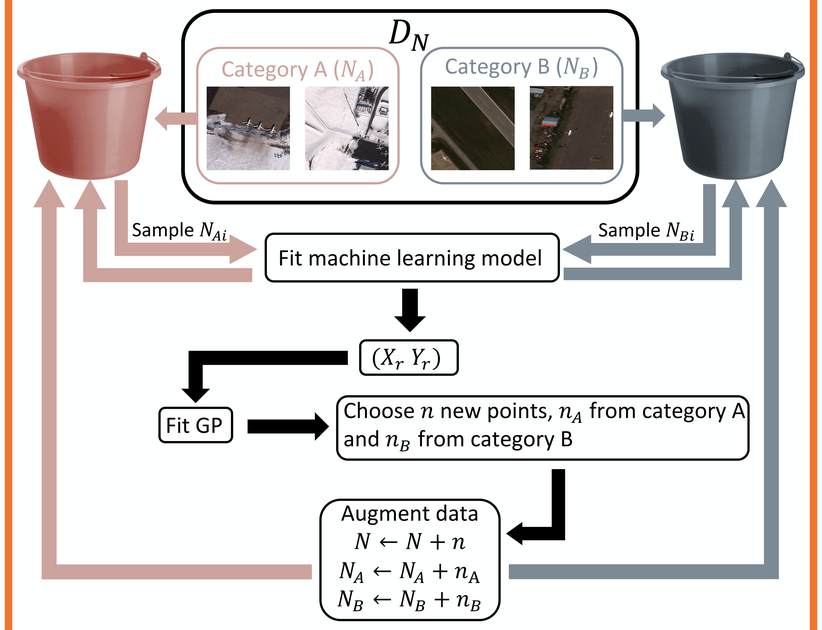
For new data $\mathcal{X}$, the predictive distribution of $\mathcal{Y}(\mathcal{X})$ conditional on (X_r, Y_r) is

$$\mathcal{Y}(\mathcal{X}) \mid X_r, Y_r \sim \mathcal{N}\left(\mu_r(\mathcal{X}), \Sigma_r(\mathcal{X})\right) \equiv \mathrm{GP}(\mathcal{X}; X_r, Y_r),$$
$$\mu_r(\mathcal{X}) = \Sigma(\mathcal{X}, X_r)\Sigma(X_r)^{-1} Y_r$$
$$\Sigma_r(\mathcal{X}) = \Sigma(\mathcal{X}) - \Sigma(\mathcal{X}, X_r)\Sigma(X_r)^{-1}\Sigma(\mathcal{X}, X_r)^{\mathsf{T}}.$$

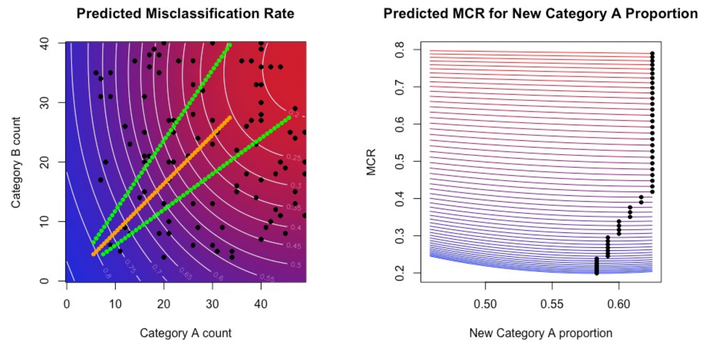
METHODS

We assume the metadata of a data set D_N is known and that it partitions D_N into two data sets, D_{N_A} and D_{N_B}, for categories A and B.

We consider using this metadata to inform new data acquisitions. In particular, we try to learn the optimal metadata composition by training varying subsets of the data using a machine learning model and saving the model's performance for that subset of the data.



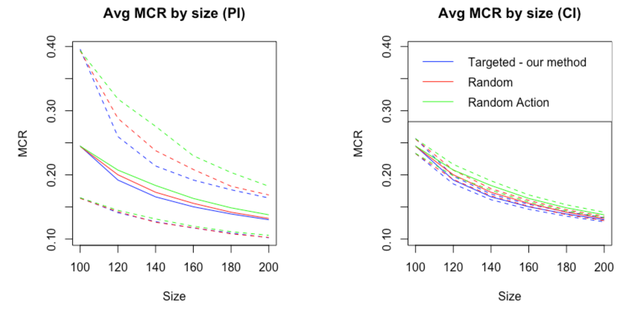
Once enough samples have been collected, a GP is fit and used to find the best composition of metadata to add.



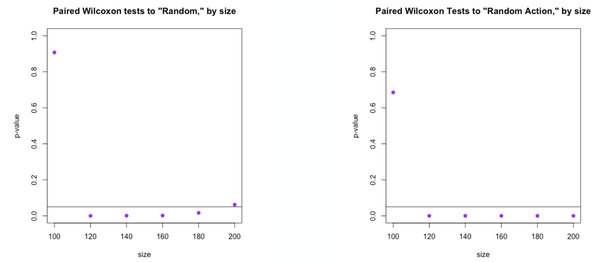
RESULTS

This data acquisition method was used to sequentially add points, 20 at a time, to increase a data set from size 100 to size 200.

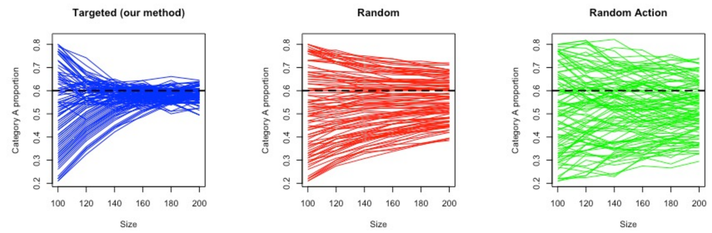
We present two competitor methods:
1. Random: new images are selected at random from the corpus of images
2. Random action: the composition of metadata in the next batch of images is selected at random (e.g. 6 from category A, 14 from category B)



Statistical testing shows there is a significant difference between our method, in blue, and both other methods after a single acquisition.



We also find that, although model performance improvement is minimal, our method succeeds in quickly identifying the optimal metadata balance.



This method is currently being tested on other data sets. One obstacle to implementing this method is the computational cost of fitting many machine learning models, although parallelization is likely possible.

HUME CENTER FOR NATIONAL SECURITY AND TECHNOLOGY
VIRGINIA TECH™