A Probabilistic Method to Predict Classifier Accuracy on Larger Datasets given Small Pilot Data

Ethan Harvey[1], Wansu Chen[2], David M. Kent[3], Michael C. Hughes[1]

[1]Department of Computer Science, Tufts University
[2]Department of Research and Evaluation, Kaiser Permanente Southern California
[3]Predictive Analytics and Comparative Effectiveness Center, Tufts Medical Center

When a large dataset of labeled images is not available, research projects often have a common trajectory:

(1) gather a small "pilot" dataset of images and corresponding class labels,
(2) train classifiers using this available data, and then
(3) plan to collect an even larger dataset to further improve performance.

When gathering more labeled data is expensive, practitioners face a key decision in step 3: *given that the classifier's accuracy is Y% at the current size x, how much better might the model do at 2x, 10x, or 50x images?*

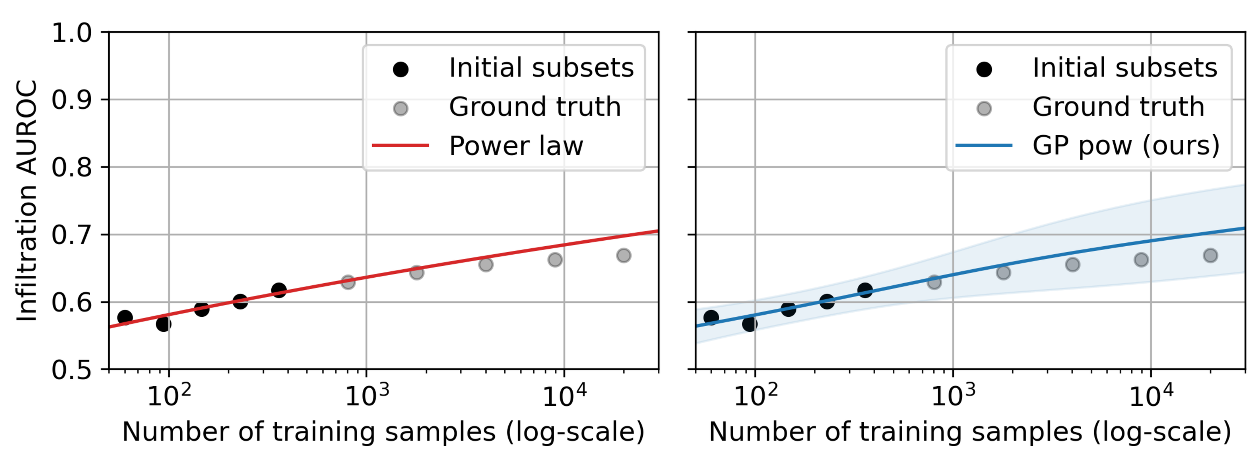

Figure 1: Example learning curves for AUROC of predicting infiltration in chest X-rays. Left: Power law fit with MSE. Right: Our Gaussian process with a power law mean function and 95% confidence interval for uncertainty.

Goals

What: A probabilistic method that can model a *range* of plausible curves to extrapolate classifier accuracy to larger datasets.

Why:

- Recent approaches have focused almost entirely on estimating one single "best-fit" curve.
- Errors are large when extrapolating from small dataset sizes (Rosenfeld et al., 2020; Mahmood et al., 2022).

How:

- A Gaussian process (GP) that can match existing curve-fitting approaches in terms of error while providing additional uncertainty estimates.

Probabilistic Method

GP extrapolation model

$$p(\mathbf{f} \mid \mathbf{x}) = \mathcal{N}(\mathbf{f} \mid m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$
$$p(\mathbf{y}|\mathbf{f}, \mathbf{x}) = \textstyle\prod_{r=1}^{R} \mathcal{N}(y_r | f_r, \tau^2)$$

Power law mean

$$m^{\text{pow}}(x) = (1-\varepsilon) - \theta_1 x^{\theta_2}$$

Arctan mean

$$m^{\text{arc}}(x) = \frac{2}{\pi} \arctan\left(\theta_1 \frac{\pi}{2} x + \theta_2\right) - \varepsilon$$

Covariance function

$$k(x, x') = \sigma^2 \exp\left(-\frac{(\log(x) - \log(x'))^2}{2\lambda^2}\right)$$

We pay particular attention to the mean, offering two concrete choices, a power law and an arctan, inspired by the best performing methods from prior work.

We make our code and models publicly available at:
github.com/tufts-ml/extrapolating-classifier-accuracy-to-larger-datasets

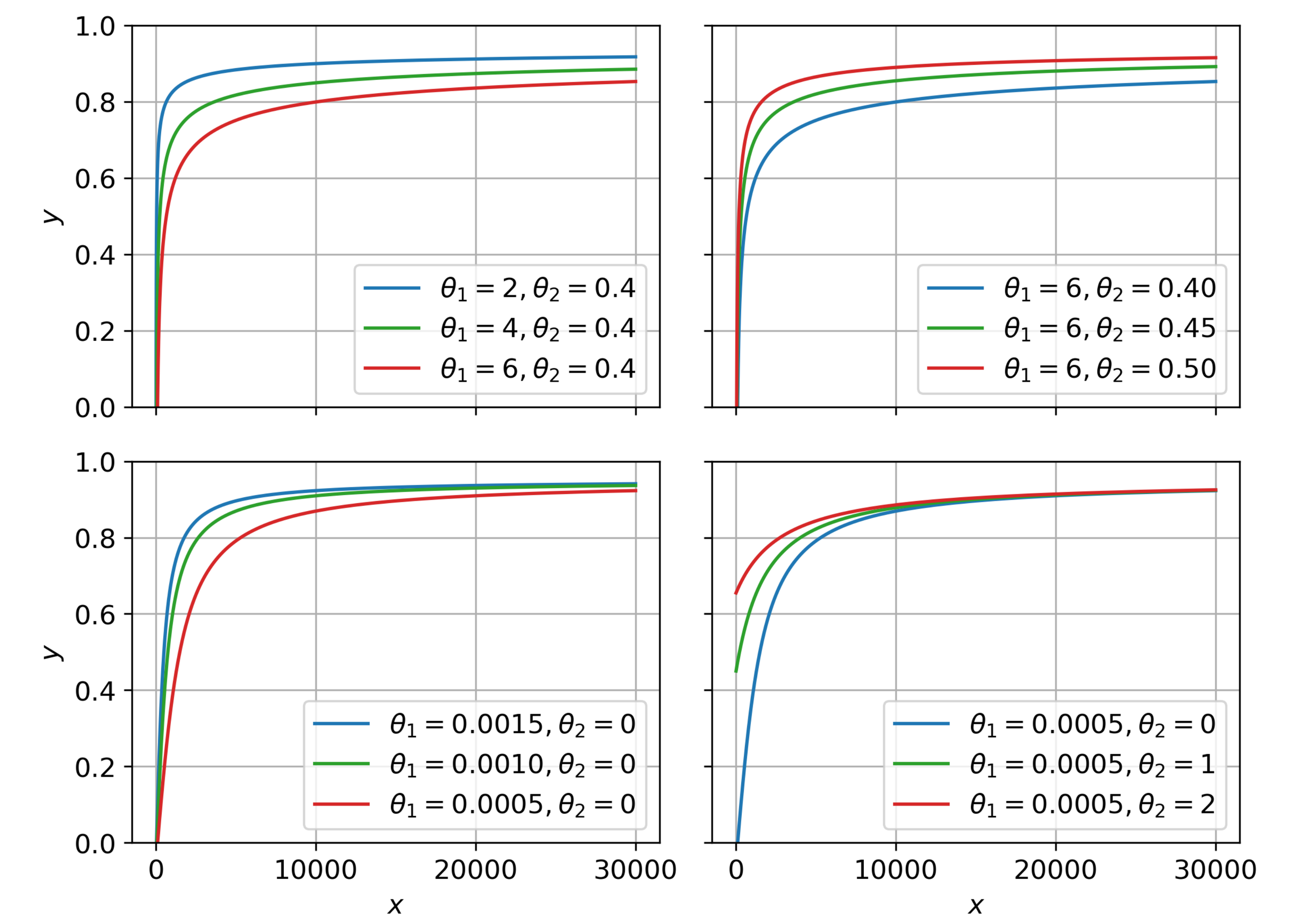

Figure 2: Example varying parameters for the power law (left) and arctan function (right) with $\varepsilon = 0.05$.

Priors

Our GP model has two kinds of parameters. Let $\eta = \{\tau, \sigma, \lambda, \varepsilon\}$ denote the parameters that control *uncertainty* (in the likelihood or the GP covariance function) or *asymptotic behavior* (e.g., ε sets the saturation value of $m(x)$). We develop prior distributions for each parameter in η.

Fitting to data via MAP estimation

We optimize the following maximum a-posteriori (MAP) objective to obtain point estimates of θ and η:

$$\hat{\theta}, \hat{\eta} = \operatorname*{argmax}_{\theta, \eta} \ \log p_{\theta,\eta}(\mathbf{y}|\mathbf{x}) + \log p(\eta)$$

Extrapolation via the predictive posterior

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{y}, \mathbf{x}) = \mathcal{N}(\mathbf{y}_*|\mu, \mathbf{\Sigma}),$$
$$\mu = \mathbf{m}_* + \mathbf{K}_*^T(\mathbf{K} + \tau^2 I_R)^{-1}(\mathbf{y} - \mathbf{m})$$
$$\mathbf{\Sigma} = \mathbf{K}_{**} + \tau^2 I_Q - \mathbf{K}_*^T(\mathbf{K} + \tau^2 I_R)^{-1}\mathbf{K}_*$$

Results

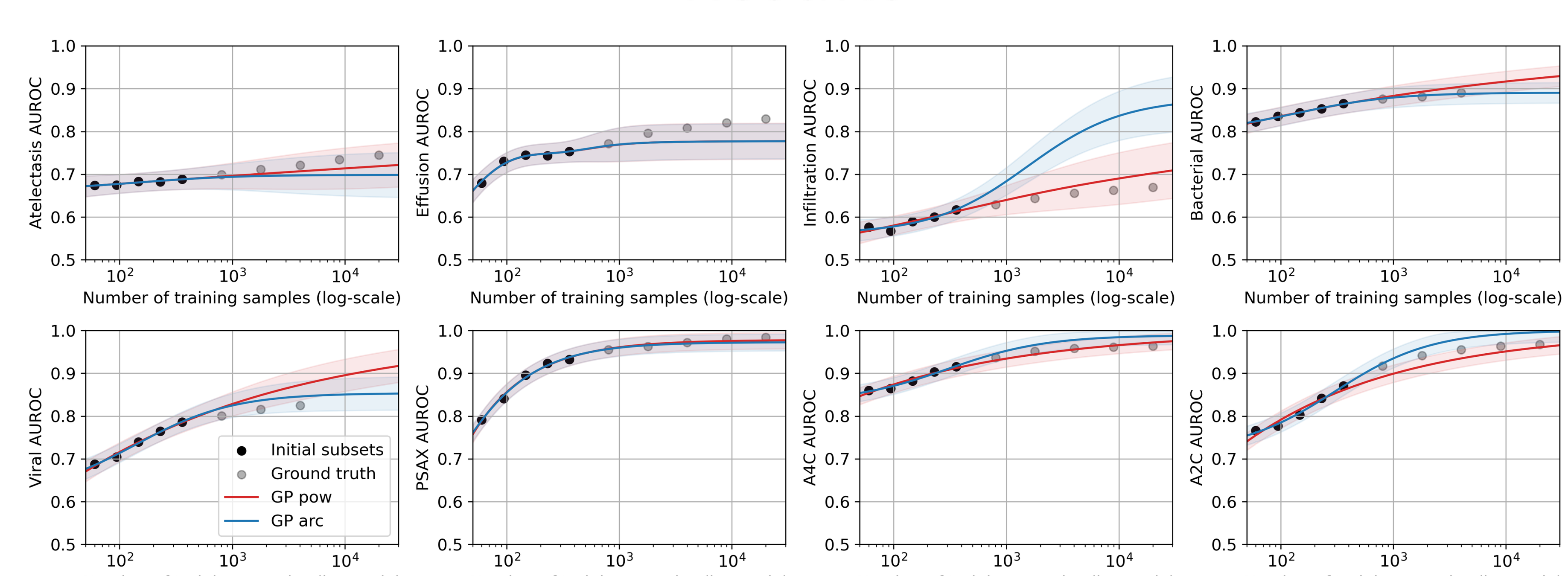

Figure 3: Long-range extrapolation results for AUROC of predicting atelectasis, effusion, and infiltration from the ChestX-ray14 dataset; bacterial and viral pneumonia from the Chest X-Ray dataset; and PSAX, A4C, and A2C from the TMED-2 dataset.

Outlook. We hope our approach provides a useful tool for practitioners in medical imaging and beyond to manage uncertainty when assessing data adequacy.

References

Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sanja Fidler, and Marc T. Law. How Much More Data Do I Need? Estimating Requirements for Down-Stream Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A Constructive Prediction of the Generalization Error Across Scales. In *International Conference on Learning Representations (ICLR)*, 2020.