Learning from Noise:
Applying Sample Complexity for Political Science Research

Perry Carter & Dahyun Choi
Ph.D. Candidates
Department of Politics
Princeton University
Correspondence: dahyunc@princeton.edu

PRINCETON UNIVERSITY

Overview

- Social science concepts are multidimensional and inherently noisy.
- A tool for guaranteeing the sample size necessary to achieve a minimum level of accuracy with a precise level of confidence
- Researcher-specified bounds on conceptual complexity and labeling error

Probably Approximately Correct (PAC) Model

Data-Generating distribution

- We employ the notation of domain χ , label set $\mathbf{Y}$, and (binary) concept classes $\mathbf{C}$. We consider a probability distribution $\mathbf{D}$ (unknown) over χ.
- A labeled set of training examples $\mathbf{S} = \{(x_1, y_1), ..., (x_m, y_m)\}$ is generated by taking $x_i \sim D$ i.i.d

True Error

- Consider a data-generating distribution D and the true labeling concept c. The *true error* of a classification rule h with respect to D is the probability that h makes a mistake.

$$err_D(h) = Pr_{x \sim D}[h(x) \neq c(x)] \tag{1}$$

Empirical Error

- Given a sample set S, the empirical error of a concept h with respect to S is the fraction of instances in S that are incorrectly labeled by h.

$$err_s(h) = \frac{1}{m}\sum_{i=1}^{m} 1(h(x_i) \neq y_i)) \tag{2}$$

Intuition

- Assuming that $\mathbf{S}$ is coming from a fixed but unknown distribution D, the goal is to use the sample set S to learn a concept h that has a small true error on D.
- We assume that there is an unknown concept $c \in C$ that truly labels instances in distribution D. We also assume that we have access to another set of concepts H from which we have to choose the concept. For ease of representation, we often call H the class of hypotheses.

Sample Complexity Bounds (SCB)

- Sample complexity characterizes the number of examples used or required by a PAC learning algorithm to attain error rate greater than ϵ with probability bounded by δ, given noisy labels with probability $\eta < 1/2$.
- We provide three tools for researchers to explicitly characterize the sample size needed to guarantee desired accuracy, based on researcher-specified assumptions.
- Combining [5] with [1], a general lower bound on sample complexity (SCB) is given by

$$\Omega(\frac{VC(\mathbf{C})}{\epsilon(1-2\eta)^2} + \frac{log(1/\delta)}{\epsilon(1-2\eta)^2}) \tag{3}$$

where $VC(\mathbf{C})$ indicates the Vapnik–Chervonenkis dimension, which measures the underlying complexity of the target concept.

Estimating Vapnik–Chervonenkis dimension (VCD) for complexity bounds

- Calculating VCD analytically is challenging for most concepts [4].
- Solution: estimate empirically based on known relationship between worst-case generalization error and $VC(\mathbf{C}) = d$:

$$f(d;n) = \begin{cases} 1 & n < \frac{d}{2} \\ a\frac{log\frac{2n}{d}+1}{\frac{d}{d}-a''}(\sqrt{1+\frac{a'(\frac{n}{d}-a'')}{log\frac{2n}{d}+1}}+1) & \text{else} \end{cases}$$

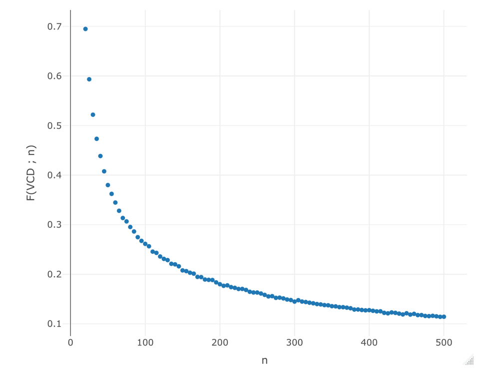

Figure 1: Simulation for Estimating the Risk Bounds

The y-axis gives the estimated bound on the relationship between empirical risk and sample size for a given classifier. Since the functional form of this relationship is known up to a constant given the true VCD, we can then estimate the VCD of any classifier through non-linear regression [6]. Moreover [4] shows that this estimate is consistent in the number of simulations.

Simulation-based Analysis

Step 1: Decide on desired accuracy parameters and concept definition
Step 2: Calculate the VCD of the chosen model using the above estimation procedure [4]
Step 3: Generate a fine grid of points over the k-dimensional feature space
Step 4: Classify these points according to the pre-defined concept
Step 5: Generate observed labels by adding independent random noise with probability η
Step 6: Calculate sample complexity bounds empirically for a range of acceptable error rates
Step 7: Repeat the process according to a range of values of "optimism" parameter (analytic bound corresponds to worst-case sampling).

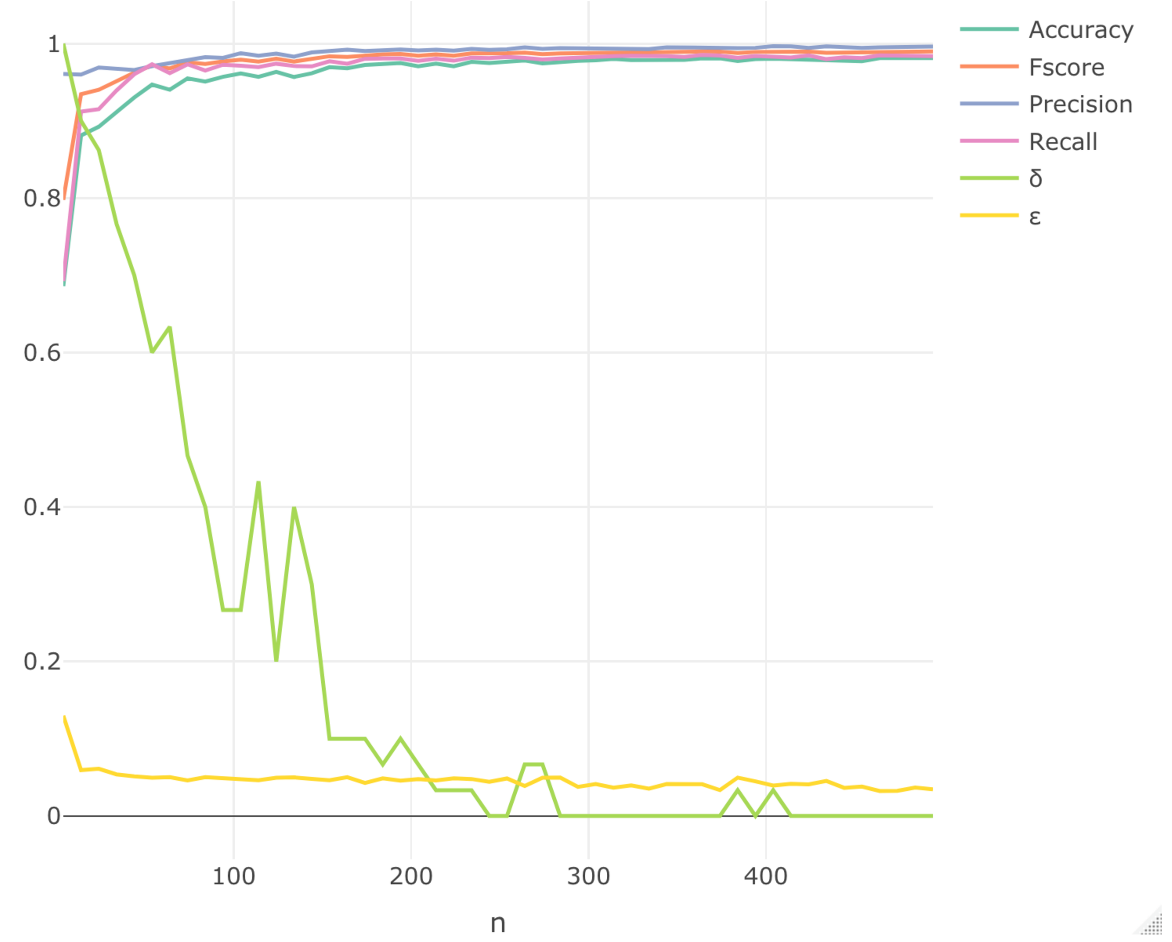

Figure 2: Learning to Classify Polyarchies

1. A stylized version of the well-known model of "polyarchy" proposed by in [2] - an unusually well-defined concept.
2. Empirical research on democracy is hampered by small sample size.
3. Values are calculated by fixing η = 0.05 and either ϵ = 0.05 or δ = 0.01
4. Theoretical bound gives 188 cases as required minimum sample size assuming perfectly square classification region.
5. This corresponds closely to simulation results under "pessimistic" sampling regime corresponding to Figure 2 (observations that provide less discriminant value are more likely).

Application to Predicting Recidivism [3]

- Comparing the overall accuracy and bias in human assessment with the algorithmic assessment of COMPAS
- 20 human coders recruited through Amazon's Mechanical Turk
- 7 Features (e.g., age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior crimes, crime degree, and crime charge) are used.
- Linear discriminant analysis (as in original paper) trained on a random 80% of training and 20% testing split, with VC dimension of 8.
- Best achievable accuracy with high confidence is approximately 35%, but additional benefit of sample size above 500 is minimal.
- Highlights concept formation problem: advantages of big data are dependent on precise specification of target concept.

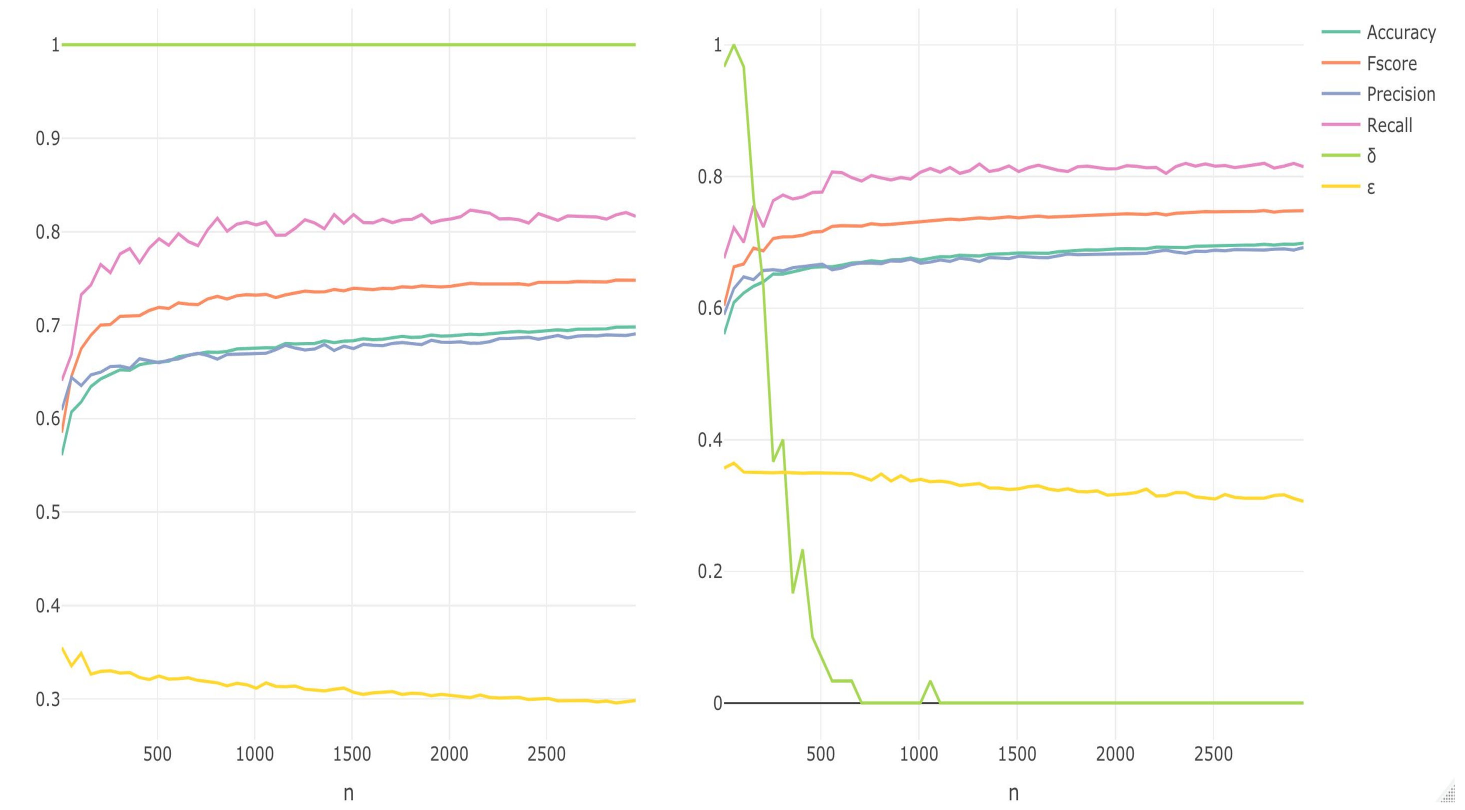

Figure 3: Simulation Analysis When ϵ = 0.05 (Left) & ϵ = 0.35 (Right)

Acknowledgements

We would like to thank Brandon Stewart and Kosuke Imai for their discussion at various stages of this project. A (R) package is on the way. Suggestions and contributions are welcome!

References

[1] Javed A Aslam and Scott E Decatur. On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195, 1996.

[2] Robert A Dahl. *Polyarchy: Participation and opposition*. Yale university press, 2008.

[3] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[4] Daniel J McDonald, Cosma Rohilla Shalizi, and Mark Schervish. Estimated vc dimension for risk bounds. *arXiv preprint arXiv:1111.3404*, 2011.

[5] Hans Ulrich Simon. General bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 402–411, 1993.

[6] Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the vc-dimension of a learning machine. *Neural computation*, 6(5):851–876, 1994.