Evaluating the Performance of the Model Selection with Average ECE and Naive Calibration in Out-of-Domain Generalization Problems for Binary Classifiers

Anxian Liu (A.Liu-2@student.tudelft.nl)
Supervised by Rickard Karlsson, Stephan Bongers
Responsible Professor: Jesse Krijthe

1. Introduction

- **Out-of-domain (OOD) generalization problem:** learn a model from one or more domain(s) that can be used in an unknown test domain.
- **Solution:** Multi-domain calibration
- **Naive calibration** and **model selection with average expected calibration error (ECE) across training domains** are two of the approaches to optimize models, so they achieve this type of calibration.

2. Motivation

- Both are **easy to apply** but **limited in their power** to learn a model that is truly well-calibrated across multiple domains [1]

3. Research question

- How well does naive calibration and model selection with average ECE perform in the out-of-domain (OOD) generalization problem for binary classifiers?
- RQ1: Does naive calibration improve average prediction performance, as measured in the accuracy or AUROC[1], across unseen domains?
- RQ2: Does OOD Accuracy[2] improve as the number of training domains grows?
- RQ2: Is model selection with average ECE a reasonable model selection strategy in the OOD generalization problem?

4. Methods

- **Experiment A:**
  - 200 datasets
  - Train and calibrate seven binary classifiers
  - Calculate the difference in OOD accuracy/OOD AUROC[3] before and after naive calibration
  - Bootstrapping hypothesis test
- **Experiment B**:
  - 10 datasets
  - Train and calibrate seven binary classifiers
  - A positive linear relationship between the number of training domains and OOD accuracy?
- **Experiment C**:
  - 3 datasets
  - Train 400 neural networks on each dataset
  - A linear relationship between OOD accuracy and average ECE? And how strong is it?

5. Data generation

- Causal relation:

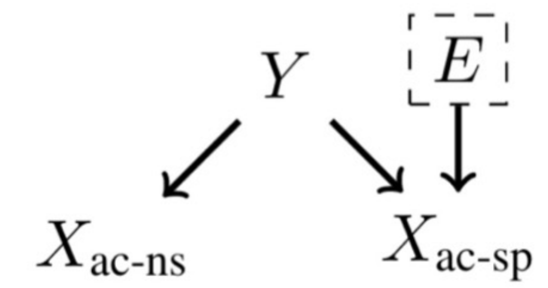

Figure 1: The causal diagram of the synthetic data [1]

- Illustration:

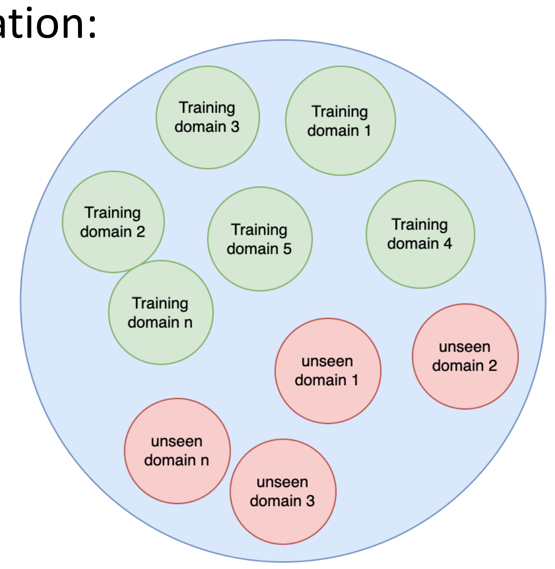

Figure 2: The Illustration of a dataset

6. Results

	Avg Diff OOD ACC	P-value	Confidence interval of the mean
Logistic Regression	0.032	0.0	(0.024, 0.041)
Linear SVM	0.021	0.0	(0.014, 0.031)
Decision Tree	0.009	0.056	(-0.001, 0.021)
Random Forest	0.010	0.086	(-0.004, 0.023)
Neural Network	0.015	0.0	(0.011, 0.019)
AdaBoost	0.005	0.0008	(0.0033, 0.010)
Naive Bayes	0.001	0.371	(-0.004, 0.005)

Table 1: Results of Experiment A

- There are similar results for OOD AUROC
- The models that have a statistically significant improvement in OOD accuracy are in **bold**

	PCC between the number of training domains and OOD ACC	PCC between the number of training data and OOD ACC	the Partial Correlation
Logistic Regression	0.85	-0.94	0.81
Linear SVM	0.88	0.49	0.85
Decision Tree	0.92	0.17	0.92
Random Forest	0.90	0.31	0.89
Neural Network	0.86	-0.88	0.81
AdaBoost	0.37	0.04	0.37
Naive Bayes	0.90	0.21	0.90

Table 2: Results of Experiment B

- A **positive linear correlation** between the number of training domains and OOD accuracy
- PCC: Pearson correlation coefficient

	PCC between ECE and OOD accuracy	PCC between validation accuracy and OOD accuracy	the Partial Correlation
Dataset A	-0.84	0.37	-0.82
Dataset B	-0.64	0.37	-0.56
Dataset C	-0.70	0.31	-0.71

Table 3: Results of Experiment C

- A relatively strong **negative linear correlation** between average ECE and OOD accuracy

7. Conclusion

- Naive calibration can improve OOD accuracy and OOD AUROC of some binary classifiers. At least, It does not make the model worse.
- For most classifiers, training the model on data from more training domains leads to higher OOD accuracy.
- Average ECE is a **reasonable** metric for selecting a model, and it is **better** than validation accuracy in the OOD generalization problem.

8. Limitations

- All experiments are based on synthetic data.
- Isotonic regression is the only method to implement naive calibration.
- PCC and the partial correlation only measure linear relationships.

9. Future work

- Use real-world datasets.
- Try another method to implement naive calibration, such as Bayesian Binning into Quantiles [2].
- Conduct Experiments B and C on more datasets and analyze results with statistical tools.

10. References

[1] Wald, Y., Feder, A., Greenfeld, D., & Shalit, U. (2022). On Calibration and Out-of-domain Generalization. *ArXiv:2102.10395 [Cs]*. http://arxiv.org/abs/2102.10395

[2] Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (n.d.). *Obtaining Well Calibrated Probabilities Using Bayesian Binning*. 7.

1: the area under the receiver operating characteristic 2: average accuracy across the unseen domains 3: average area under the ROC Curve across unseen domains