2V-03

不確実性の較正評価による不変リスク最小化近似手法の比較理解

Understanding Variants of Invariant Risk Minimization from the Perspective of Calibration

吉田 晃太朗 ♯*1　長沼 大樹 ♯*2*3
Kotaro Yoshida　Hiroki Naganuma
*1東京工業大学 Tokyo Institute of Technology
*2モントリオール大学 Université de Montréal
*3Mila Mila - Quebec AI Institute

1. はじめに

学習データとは異なるデータ分布 (環境) に対してモデルの推論性能が低下してしまう分布外汎化の問題に対し、データの分布不変な特徴量を捉えるよう制約を化した不変リスク最小化 (Invariant Risk Minimization, IRM) [Arjovsky 19] が提案されている。しかしながら、IRM は複雑な bi-level 最適化問題を含んでおり実現困難であるため、IRM を実装可能にするため IRMv1 を含む様々な近似手法が提案されている [Arjovsky 19] [Ahuja 22] [Ahuja 20] [Chen 23] [Lin 22]。しかし、これら近似手法は分布外汎化に関するデータセット上でのモデルの推論精度によってのみ評価されており、どの程度不変的な特徴量を獲得しているかは明らかでない。

本研究では、IRM が複数の環境にわたるモデルの不確実性の較正として一般化できる [Wald 21] ことに着目し、実際に計算可能なモデルの不確実性の較正を指標とし、IRM の近似手法の比較評価を行った。

2. 不変リスク最小化 (IRM)

モデル $f : \mathcal{X} \to \mathcal{Y}$ が環境の集合 E において不変であることは以下で定義される [Wald 21]。

$$\mathbb{E}[Y^e|f(X^e)] = \mathbb{E}[Y^{e'}|f(X^{e'})], \forall e, e' \in E \tag{1}$$

IRM は (1) の実現によってモデルをデータ分布の変化に対して頑健にすることを目的とし、以下で定義される [Arjovsky 19]。

$$\min_{\substack{\Phi:\mathcal{X}\to\widehat{\mathcal{H}} \\ \omega:\widehat{\mathcal{H}}\to\mathcal{Y}}} \sum_{e\in\varepsilon_{train}} R^e(\omega \circ \Phi) \tag{2}$$

$$subject\ to \quad \omega \in \arg\min_{\bar{\omega}:\widehat{\mathcal{H}}\to\mathcal{Y}} R^e(\bar{\omega} \circ \Phi),\ for\ all\ e \in \varepsilon_{train}$$

ここで、$\mathcal{X}$ は各データの入力空間、$\widehat{\mathcal{H}}$ は抽出された不変な特徴量空間、そして $\mathcal{Y}$ はモデルの出力空間となる。IRM は $f = \omega \circ \Phi$ とし、Φ が入力データを $\widehat{\mathcal{H}}$ へ写像、その抽出された不変特徴量を基に最終的な予測をすることで分布外汎化として一貫した予測の実現を目指している。

また、IRM は複数の環境におけるモデルの不確実性の較正の一般化であることが知られている [Wald 21]。不確実性の較正は特に分類タスクにおいて近年注目されている問題であり、モデルの予測確率 (確信度) と実際の精度は実用上一致していることが望ましいが、近年の大規模なモデルはそれが一致していないことが指摘されている [Guo 17]。環境の集合 E においてモデルの不確実性が正しく較正されていることは、f の任意の予測確率 α に対して以下で定義される。

$$\mathbb{E}[Y^e|f(X^e) = \alpha] = \alpha, \forall e \in E \tag{3}$$

ここで (3) は (1) の十分条件となり、環境を跨いだ不確実性の較正が IRM を達成する一つの特殊ケースであると見做せる。

連絡先: yoshida.k.bl@m.titech.ac.jp, ♯ equal contribution

3. 不変リスク最小化の近似手法

(2) の定式化は、厳密に解くことが困難であるため、これまで実装可能な様々な近似手法が提案されてきた。

IRMv1

IRMv1 [Arjovsky 19] は (2) における ω を線形写像として制約を課し $\Phi' = (\omega \cdot \Phi)$, $\omega' = 1.0$ という変形を可能にし、以下のような 1 変数の最適化問題に緩和している。

$$\min_{\Phi':\mathcal{X}\to\mathcal{Y}} \sum_{e\in\varepsilon_{train}} R^e(\Phi') + \lambda \parallel \nabla_{\omega'|\omega'=1.0} R^e(\omega' \cdot \Phi') \parallel^2 \tag{4}$$

Information Bottleneck based IRM(IB-IRM)

IB-IRM [Ahuja 22] は情報ボトルネック法を応用し Φ の持つ情報量を圧縮することで、不変的な予測に悪影響を及ぼす特徴量への過度な依存を防ぎ IRMv1 を改良することを目指している。正則化項として Φ によって各データから抽出された特徴量の分散を追加している。

$$\lambda \parallel \nabla_{\omega'|\omega'=1.0} R^e(\omega' \cdot \Phi') \parallel^2 + \gamma Var(\Phi) \tag{5}$$

Pareto IRM(PAIR)

PAIR [Chen 23] は、一般的に経験損失最小化 (ERM) と分布外汎化はトレードオフの関係にありそれらを適切に管理する必要があることに着目し、多目的最適化の観点からそれらのパレート最適解を求めることでよりロバストなモデルを実現している。実際には ERM、IRMv1 のペナルティ、そして vREx [Krueger 21] のペナルティに対する多目的最適問題として実装している。

$$\min_{\Phi':\mathcal{X}\to\mathcal{Y}} (\mathcal{L}_{ERM}, \mathcal{L}_{IRMv1}, \mathcal{L}_{vREx}) \tag{6}$$

IRM Game

IRM GAME [Ahuja 20] は、IRM に各環境間におけるゲーム理論の枠組み導入し、推論精度の観点からナッシュ均衡を達成することを目的とする。学習データの各環境に独自の識別器 ω^e を割り当てそれぞれが各環境で最適になるように学習し、全ての識別器のアンサンブルを最終的な ω とする。環境固有の識別器を学習することで IRMv1 でなされた線形性への制限を排除し、より (2) に近い実装を目指している。

Bayesian IRM(BIRM)

BIRM [Lin 22] では、学習データが不十分である場合 IRMv1 は学習環境への過適合の恐れがあるためベイズ推定によるアプローチを導入した。モデルが不変的な特徴量を獲得できている場合、Φ によって写像されたデータと正解ラベルが与えられたときの事後確率 $p(\omega^e|\Phi(X^e), Y^e)$ は全ての環境で不変であることに注目し IRMv1 のペナルティを変更している。点推定ではなくベイズ推定を用いることでデータの不確実性を考慮し、モデルの汎化性能の向上を目指している。

4. 実験

前節までの IRM の近似手法と ERM の計 6 手法を用いて、分布外汎化データセット上での不確実性の較正性能を比較した。

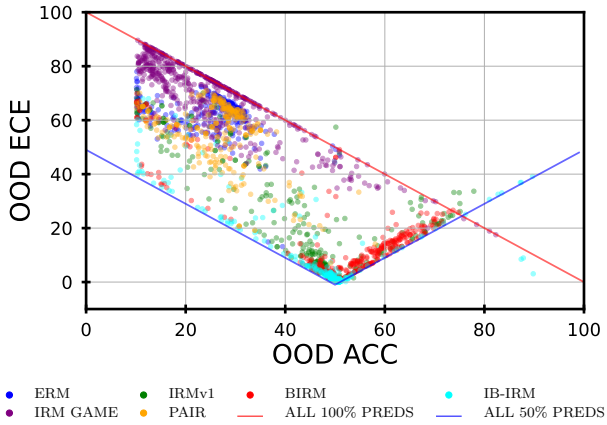

図 1: CMNIST における分布外での推論精度 (横軸) と ECE(縦軸) の関係比較。赤実線はモデルの予測確率が全て 100%、青実線はモデルの予測確率が全て 50%の場合の理論値。比較的高い推論精度を達成している IRMv1,IB-IRM,BIRM は青線近くに分布し、これらは予測を曖昧にすることで非不変的な特徴量への過度な依存を緩和している。

		ERM	IRMv1	IB-IRM	BIRM	IRM GAME	PAIR
RMNIST	ACC	91.1±1.0	91.9±1.0	90.9±0.9	89.1±0.4	90.8±1.0	87.4±0.8
	ECE	5.27±1.11	3.98±0.69	**1.58±0.44**	6.93±0.56	7.17±0.98	7.88±0.79
PACS	ACC	73.6±2.0	74.9±1.9	75.9±0.9	71.5±1.7	73.2±1.3	76.3±1.2
	ECE	**11.99±1.86**	13.35±1.45	12.85±0.60	13.73±1.93	12.50±1.45	12.46±0.95
VLCS	ACC	70.3±1.0	70.5±1.2	69.4±0.7	68.8±0.8	69.9±1.2	71.1±1.2
	ECE	12.53±1.42	11.90±1.26	**7.84±1.55**	10.57±1.81	8.70±0.93	10.34±1.11
Avg.	ACC	78.3±1.3	79.1±1.4	78.8±1.12	76.5±0.8	78.1±1.2	78.3±1.1
	ECE	9.93±1.46	9.74±1.13	**7.42±0.86**	10.41±1.43	9.46±1.12	10.23±0.95

表 1: RMNIST、PACS 及び VLCS における分布外での推論精度と ECE の比較。同等の推論性能下で平均的に IB-IRM の ECE が低い。

データセットとして CMNIST、RMNIST、PACS、VLCS を採用し、不確実性の較正の指標として ECE [Johansson 21] を用いた。ECE は小さいほど望ましい。

CMNIST

図 1 は横軸を分布外での推論精度、縦軸を分布外での ECE とし、それぞれの手法で複数のハイパーパラメータで学習し散布図として示したものである。赤の実線はモデルの予測確率が全て 100%で確信度が極めて高い場合、青の実線は予測確率が全て 50%で予測の確信度が極めて低く曖昧な場合の理論値である。推論精度が比較的高い手法 (**IRMv1**,**IB-IRM**,**BIRM**) は青の実線に沿っており、予測を曖昧にすることで非不変的な特徴量への過度な適合を克服できていることが分かる。

RMNIST, PACS, VLCS

不確実性の較正はモデルの分布外における推論性能に影響を与える [Ovadia 19] ことを考慮し、より公平な ECE 比較のために分布外汎化性能と合わせた ECE の比較評価を行った。学習環境での推論精度に閾値を設け、達成した段階で、学習の早期停止を行い、同等の推論精度の条件とした。結果は表 1 に示しており、平均的に IB-IRM の ECE が低い結果となった。また、図 2 では学習環境とテスト環境の ECE の関係を可視化しており、赤の実線は両環境において ECE が等しい場合を示す。環境を跨いだ不確実性の較正の観点からするとこの赤線上に近いことが望ましい。特に図 2a と図 2c において IB-IRM が他の手法に比べ、学習環境への過適合が見られず、より理想的な特徴量学習が行われていることが示唆された。

5. おわりに

本稿では、計算不可能な IRM を焦点とし、近似手法を比較理解するため、計算可能な不確実性の較正指標を用いることで比較評価を行った。実験の結果、情報ボトルネック法を IRM に応用した IB-IRM が不確実性の較正の観点から優れている

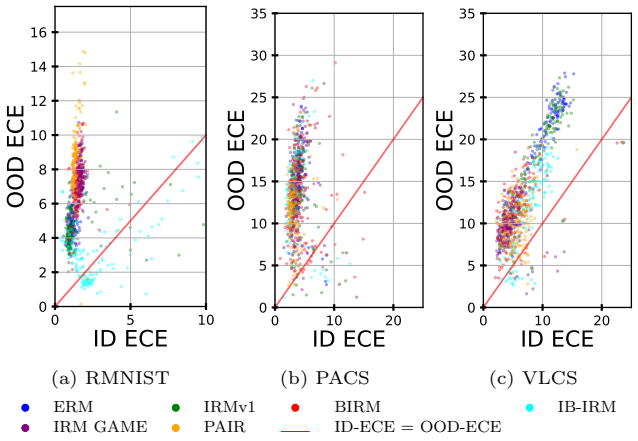

図 2: 学習環境 (横軸) 及びテスト環境 (縦軸) での ECE の関係比較。赤実線は両環境で ECE が等しい場合を示す。IB-IRM(水色) が他手法に比べ、学習環境に過適合しない傾向が観測された。

ことが示唆された。IB-IRM は特徴抽出器 $\Phi(X)$ の複雑性を抑えることで、環境特有の特徴量への依存に起因する過度に高い確信度を防いでいることが要因であると推察される。

今後の課題として、不変特徴量学習の程度が学習環境外データの推論精度だけでは測りきれないことが明らかとなったため、広い意味での汎化指標や測定方法の確立が必要である。また、分布外汎化・不確実性の較正の関係理解、及びそれに基づいた新たな IRM の近似手法の開発が期待される。

謝辞

本研究は、令和 5 年度 東京工業大学 TSUBAME より若い世代の利用者支援制度の支援を受けたものである。

参考文献

[Ahuja 20] Ahuja, K., Shanmugam, K., Varshney, K. R., and Dhurandhar, A.: Invariant Risk Minimization Games (2020)

[Ahuja 22] Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I.: Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization (2022)

[Arjovsky 19] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D.: Invariant risk minimization, *arXiv preprint arXiv:1907.02893* (2019)

[Chen 23] Chen, Y., Zhou, K., Bian, Y., Xie, B., Wu, B., Zhang, Y., Ma, K., Yang, H., Zhao, P., Han, B., and Cheng, J.: Pareto Invariant Risk Minimization: Towards Mitigating the Optimization Dilemma in Out-of-Distribution Generalization (2023)

[Guo 17] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q.: On calibration of modern neural networks, in *International Conference on Machine Learning*, pp. 1321–1330PMLR (2017)

[Johansson 21] Johansson, U., Löfström, T., Boström, H., and Nguyen, K.: Calibrating multi-class models, in *International Symposium on Conformal and Probabilistic Prediction with Applications* (2021)

[Krueger 21] Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A.: Out-of-Distribution Generalization via Risk Extrapolation (REx) (2021)

[Lin 22] Lin, Y., Dong, H., Wang, H., and Zhang, T.: Bayesian Invariant Risk Minimization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16021–16030 (2022)

[Ovadia 19] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, *Advances in neural information processing systems*, Vol. 32, (2019)

[Wald 21] Wald, Y., Feder, A., Greenfeld, D., and Shalit, U.: On calibration and out-of-domain generalization, *Advances in Neural Information Processing Systems*, Vol. 34, (2021)