F-003

不変リスク最小化による不確実性の較正

吉田 晃太朗[1,a)]　多田 圭吾[♯2,b)]　長沼 大樹[♯3,4,c)]
Kotaro Yoshida[1,a)]　Tada Keigo[2,b)]　Naganuma Hiroki[3,4,c)]

1. はじめに

近年の深層ニューラルネットワークによる予測では高い精度を実現できるものの、不確実性の較正における課題 (Calibration) が指摘されている [2]。一方、分布外汎化も実応用における喫緊の課題である。これは、モデルの評価環境におけるデータの分布が学習環境とは異なる場合、推論精度が低下してしまう問題である。その対策手法の一つとして、学習環境から不変的な特徴量を抽出し未知のデータにおいてもより頑健なモデルを学習することを目的として設計された不変リスク最小化 (IRM: Invariant Risk Minimization) [1] が知られている。上記二つの問題の関連として、不確実性の較正と分布外汎化性能は相関関係にあることが示唆されており、特に、マルチドメインでの不確実性の較正は IRM の特殊ケースとみなせることから、不確実性の較正手法により分布外汎化性能が改善されることが報告されている [5]。本稿では、不確実の較正のシナリオに IRM が有効かという仮説を検証する。

2. 背景

分布外汎化: 分布外汎化のシナリオにおいて代表的な手法の一つである IRM は [1]、以下を最小化することを目的とする。

$$R^{OOD}(f) = \max_{e \in \varepsilon_{all}} R^e(f) \tag{1}$$

ここで、$R^{OOD}(f)$ は評価環境で考えられるすべての環境の中で誤差が最大となるようなものを最小化するものである。しかし、実際には学習時にアクセスできる環境には制約があり、IRM の定式化においては、前述の制約下で不変特徴量を抽出することを目指す。

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

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

ここで、$\mathcal{X}$ は各データの入力空間、$\widehat{\mathcal{H}}$ はすべての環境に共通した特徴量空間、そして $\mathcal{Y}$ はラベル空間となる。つまり、Φ が任意の入力から環境に依存しない特徴量を抽出し、その後、それらの特徴量を用いて ω が分類を行うということを目指している。

しかし、上記の定式化ではバイレベルの最適化問題を解くことになり実用上困難であるため、一般的には IRMv1 と呼ばれる以下の近似式が用いられる。本稿の実験においても IRMv1 を用いている。

$$\min_{\Phi:\mathcal{X}\to\widehat{\mathcal{H}}} \sum_{e\in\varepsilon_{train}} R^e(\Phi) + \lambda \cdot \| \nabla_{\omega|\omega=1.0} R^e(\omega\cdot\Phi) \|^2 \tag{3}$$

不確実性の較正: 不確実性が正しく較正されていること (例えば、モデルによる 80%の予測が 100 事例存在するとき、その中で予測が正しいものは 80 事例存在するべきである) は、医療画像診断や、自動運転などのシナリオにおいて、極めて重要である。近年の深層ニューラルネットワークでは、モデル容量の増加などが原因で、その不確実性が正しく較正されない問題が指摘されており、その較正のための多くの手法が提案されている [2]。不確実性の較正の良し悪しは、ECE として定量化され一般に用いられており、以下のように定義される [3]。

$$ECE = \sum_{i=1}^{M} \frac{|B_m|}{N} |Acc(B_m) - Conf(B_m)| \tag{4}$$

0 から 1 までの予測確率を M 分割し各ビンに対して、Acc をビン内の予測精度、$Conf$ をビン内の予測確率 (確信度) の平均としている。

3. 関連研究

十分な種類を持つマルチドメインでの不確実性の較正は分布外汎化性能の改善につながることが示されている [5]。二値分類器 $f : \mathcal{X} \to [0, 1]$ の予測の不確実性が正しく較正されているとは、全ての $e \in \mathbb{E}_{train}$ 及び、e 上で f が出力可能な全ての α に対して以下が成り立つことと定義される。

$$\mathbb{E}[Y^e | f(X^e) = \alpha] = \alpha \tag{5}$$

上記の不確実性の較正の定義はマルチドメインでの不変性の定義に結びつけることができる。f が E_{train} において不変であることの定義は、$e, e' \in E_{train}$ 及びそれぞれの環境で f が出力可能な α に対して以下が成り立つことである。

$$\mathbb{E}[Y^e | f(X^e) = \alpha] = \mathbb{E}[Y^{e'} | f(X^{e'}) = \alpha] \tag{6}$$

1 東京工業大学 / Tokyo Institute of Technology
2 立命館大学 / Ritsumeikan University
3 モントリオール大学 / Université de Montréal
4 Mila / Montreal Institute for Learning Algorithms
a) yoshida.k.bl@m.titech.ac.jp
b) is0463hx@ed.ritsumei.ac.jp
c) naganuma.hiroki@mila.quebec
♯ denotes equal contribution

このとき、(5) が成り立つ場合、明らかに (6) も成立することから、全ての $e \in E_{train}$ において f の不確実性が正しいことが、E_{train} において f が不変であることの十分条件になることが分かる。よって、マルチドメインの不確実性の較正がモデルの分布外汎化性能を改善することが示される。

加えて、上記の定義を用いて、分類器 f が全ての $e \in E_{train}$ で不変なとき、ある g が存在し、f の各環境における予測誤差を上回らずに、$g \circ f$ が f よりも正しい不確実性を達成できることも導かれている [5]。これは以下のように導かれる。まず、E_{train} において f が不変であるとき、一般に以下が成り立つ。

$$\mathbb{E}[(Y^e - f(X^e))^2] \geq \min_{g:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(Y^e - g \circ f(X^e))^2]$$

つまり、良い $g : \mathbb{R} \to \mathbb{R}$ を選べば必ず、全ての $e \in E_{train}$ において $g \circ f$ は f の二乗誤差を上回らない。ここで、証明は省略するが、g が右辺を達成することと以下が成り立つことは必要十分である。

$$g(\hat{\alpha}) = \mathbb{E}[Y^e | f(X^e) = \hat{\alpha}] \tag{7}$$

この g を選ぶと $g \circ f(X)$ が (5) の不確実性の正しさを満足する。加えて、f の不変性から (6) も成り立つため、全ての $e \in E_{train}$ に対して同時に $g(\hat{\alpha}) = \mathbb{E}[Y^e | f(X^e) = \hat{\alpha}]$ を満たすような g が唯一存在することが示される。

ここで、IRM において (2) の不変量抽出器 Φ は、全ての $e, e' \in E_{train}$ に対して $\mathbb{E}[Y^e | \Phi(X^e) = \hat{h}] = \mathbb{E}[Y^{e'} | \Phi(X^{e'}) = \hat{h}]$ を満たすことを要請され [1]、これは $\Phi(X) = f(X)$ とした場合 (6) と一致する。つまり、マルチドメインにおける不確実性の較正はオリジナルの IRM を達成する一つの特殊ケースであると考えられ、実際に IRMv1 の正則化項をマルチドメインでのキャリブレーションを目的とした CLOvE 項とすることで、学習環境でのキャリブレーション性能と分布外汎化性能を同時に向上させることも検証により示されている [5]。ここではモデルの分布外汎化を目的として不確実性の較正が用いられているが、本稿では、不確実性の較正を目的として IRM を用いる。

4. 実験

IRM のマルチドメインにおけるキャリブレーション性能を検証するため、IRMv1 と ERM を用いて対照実験を行った。データセットシフト下での分布外汎化性能の劣化が ECE の劣化につながることが示されている [4] ことから、その影響を抑えて両者で正当な比較をするため、分布外汎化性能 (OOD-Test Acc) に閾値 (RotatedMNIST : 94%, PACS : 70%) を設け、学習の際に閾値を初めて超えた時点で学習を停止させた。また、(3) の IRMv1 のハイパラメータである λ は Arjovsky ら [1] に倣った。データセットは RotatedMNIST と PACS を用いた。RotatedMNIST は手書き文字 MNIST を回転させたもので、その回転角度別に環境を作成し、[0°, 15°, 30°, 45°, 60°] を学習に、[75°] を評価に用いた。PACS は photo、cartoon、painting、sketch の 4 つのスタイル（環境）で描かれた 7 つのクラスの動物や人間の画像のデータセットで、[photo, cartoon, painting] を学習に、[sketch] を評価に用いた。

	RotatedMNIST		PACS	
	ERM	IRMv1	ERM	IRMv1
ID-Train Acc	99.0±0.2	98.8±0.1	95.7±0.1	94.2±0.2
ID-Valid Acc	98.2±0.2	98.1±0.2	93.2±0.2	92.2±0.1
OOD-Test Acc	**94.5±0.4**	94.4±0.4	**72.0±0.3**	71.3±0.1
ID-Train ECE	1.49±0.32	1.64± 0.13	5.16±1.23	6.13±1.25
ID-Valid ECE	2.00±0.25	2.15± 0.13	5.92±1.02	6.90±0.59
OOD-Test ECE	1.99±0.359	**1.93± 0.44**	12.4±2.00	**10.8±0.39**

表 1: IRMv1 と ERM のキャリブレーション性能比較

表 1 に各データセットで行った実験結果を示す。注目すべき点は、どちらのデータセットにおいても分布外汎化性能 (OOD-Test Acc) は ERM の方が高いものの、キャリブレーション性能 (OOD-Test ECE) では IRMv1 の方が高い性能を示した点である。これは IRMv1 のキャリブレーション手法としての有効性を示唆している。RotatedMNIST においては、OOD-Test ECE は IRMv1 の方が改善しているものの ERM との差は非常に小さいが、これは RotatedMNIST がマルチドメインのデータセットとして非常に簡易でありアルゴリズムとしての差が発揮されなかった可能性がある。

5. おわりに

本稿では不確実性の較正のシナリオで、IRM のキャリブレーション性能を検証した。結果として、IRM がキャリブレーションのシナリオで有用であることが示唆された。しかし、IRM の実装の際の IRMv1 への近似が OOD 汎化に悪影響がある可能性も指摘されていることから [5]、計算可能なレベルのスケールにおいてオリジナルの IRM を検証する必要もある。

参考文献

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

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

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

[4] 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).

[5] 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).