논문: https://dl.acm.org/doi/10.5555/3692070.3694221

Exploiting Human-AI Dependence for Learning to Defer

Dependent Cross-Entropy, DCE

박세현
LAMDA Lab in SKKU

Exploiting Human-AI Dependence for Learning to Defer

- Learning to Defer, L2D
  - 모델이 직접 결정할지 사람(의사결정자)에게 위임할지를 학습
  - Human-AI 협업 성능을 최적화
- 모델의 posterior confidence vs 전문가가 맞을 확률
  - 모델과 전문가의 개별 비교

  ⇒ 실제로는 독립적이지 않다 (의존성이 존재)

"언제 위임이 이득인가?"

⇒ 모델이 틀릴 때 전문가가 맞는 사건 vs 모델이 맞을 때 전문가가 틀린 사건

Exploiting Human-AI Dependence for Learning to Defer

- 기존 방식(규칙)에서 더 해석 가능한 의존성 형태로 재표현

전문가가 맞을 확률이 더 큰가?

위임으로 얻는 이득이 위임으로 생기는 손해보다 큰가?

Problem Setting

- 변수/분포/데이터셋
  - 입력 공간: $X \subseteq \mathbb{R}^d$
  - 레이블 공간: $Y = [K] = \{1, 2, \ldots, K\}$
  - 전문가 예측: $M \in Y$

$$x \in X, y \in Y, m \in M \Rightarrow (x, y, m)$$

$$\{(x_i, y_i, m_i)\}_{i=1}^{n} \Rightarrow \; joint\ distribution\ p(x, y, m)$$

Problem Setting

- defer 옵션 분류기

$$f: X \to Y_{\perp}, \qquad Y_{\perp} = Y \cup \{\perp\}$$

  - $f(x) \neq \perp$: 모델 예측을 그대로 사용
  - $f(x) = \perp$: 전문가에게 위임

❖ $\perp$: Defer

Problem Setting

- 0-1-deferral loss: '틀리면 1, 맞으면 0' + 위임 규칙 반영

$$L_{01}^{\perp}(f(x), y, m) = \mathbb{I}_{f(x)\neq y}\mathbb{I}_{f(x)\neq\perp} + \mathbb{I}_{m\neq y}\mathbb{I}_{f(x)=\perp}$$

❖ $\mathbb{I}$: 밑(명제)이 참이면 1, 거짓이면 0

Problem Setting

- 목표 리스크 최소화
  - L 을 최소화 하는것이 목표

$$R_{01}^{\perp}(f) = \mathbb{E}_{p(x,y,m)}[L_{01}^{\perp}((f(x), y, m))]$$

최적해 f^* $(Bayes\ optimality)$

Problem Setting

- Bayes Optimality

$$f^*(\boldsymbol{x}) = \begin{cases} \perp, & \mathbb{P}(Y = M|\boldsymbol{x}) > \max\limits_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}), \\ \arg\max\limits_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}), & \text{otherwise.} \end{cases}$$

  - $\max \eta_y(x)$: 모델이 그 샘플에서 맞을 최대 확률
  - $\mathbb{P}(Y = \mathrm{m}|x)$: 전문가가 맞을 확률

❖ $\eta_y(x) = \mathbb{P}(Y = y|x)$

Consistent Surrogate Losses for L2D

- Surrogate loss
  - 0-1-deferral loss는 불연속 + 비볼록
  - $R_{01}^{\perp}$를 직접 최소화 어려움

  $\Rightarrow$ 연속 surrogate loss $L^{\perp}$ 설계

- 점수함수 g와 의사결정 f의 연결
  - f를 최적화 대신, 점수 벡터로 defer여부를 결정
    - 점수함수: $g : X \rightarrow \mathbb{R}^{K+1}$
    - defer 점수: $s_{\perp} = s_{K+1}$

$$\varphi(g(\boldsymbol{x})) = \begin{cases} \perp, & g_{\perp}(\boldsymbol{x}) > \max\limits_{y \in \mathcal{Y}} g_y(\boldsymbol{x}) \\ \arg\max\limits_{y \in \mathcal{Y}} g_y(\boldsymbol{x}), & \text{otherwise.} \end{cases}$$

A New Formulation of the Bayes Optimality

$$f^*(\boldsymbol{x}) = \begin{cases} \perp, \mathbb{P}(Y \neq r, Y = M|\boldsymbol{x}) > \mathbb{P}(Y = r, M \neq Y|\boldsymbol{x}), \\ \arg\max_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}), \qquad \text{otherwise}, \end{cases}$$

$$\textit{where } r = \arg\max_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}).$$

- $y \neq m$일 때 모델이 만든 예측을 수락
- $y \neq q,\ y = m$일 때 전문가에게 예측을 위임
- $y = q,\ y = m$일 때 위임 여부를 결정하지 않음

❖ $q = \text{argmax}_{y \in Y} g_y(x)$

Proposed Dependent Surrogate Loss

- Model inputs: 정답 y, 전문가 예측 m, 모델의 현재 예측 q
- Model outputs: K개 클래스(q) + defer(logit)
- Dependent Cross-Entropy, DCE
  - 분류 성능 Up (올바른 예측)
  - 위임이 유리한 경우 defer가 모델 예측을 이기도록 학습 (위임 결정)

$$L^{\perp}_{\mathrm{DCE}}(g(\boldsymbol{x}), y, m) = -\mathbb{I}_{y \neq m} \log(\psi^{y}_{\mathcal{Y}^{\perp}}(g(\boldsymbol{x}))) - \mathbb{I}_{y=m}\big(\log(\psi^{y}_{\mathcal{Y}}(g(\boldsymbol{x}))) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/q}(g(\boldsymbol{x})))\big)$$

- 전문가가 틀린 케이스는 모델이 책임
- 모델이 틀리고 전문가가 맞는 케이스는 위임
- 둘 다 맞는 케이스는 전이 신호 도출

❖ ψ: softmax 변환

Experiments

- 기존 loss L (Baseline)
  - Crose-Entropy 기반 CE (Mozannar & Sontag, ICML 2020)
  - One vs All 기반 OvA (Verma & Nalisnick, ICML 2022)
  - Asymmetric Softmax 기반 A-SM (Cao et al., NeurIPS 2023)

- Dataset
  - CIFAR-100N (Wei et al., ICLR 2021)
  - ImageNet-16H (노이즈 타입 110, 125) (Kerrigan et al., NeurIPS 2021)

* 전문가 예측 Amazon Mechanical Turk의 주석

Experiments

- 평가 지표
  - Error
  - Budgeted Error (10%, 20%, 30%)
  - Expected Calibration Error, ECE

$$\mathrm{ECE}(\widehat{\mathbb{P}}(Y = M|\cdot)) = \mathbb{E}_{p(\boldsymbol{x})}[|\mathbb{P}(Y = M|\widehat{\mathbb{P}}(Y = M|\boldsymbol{x}) = c) - c|].$$

* 추정된 신뢰도가 전문가가 올바른 예측을 할 실제 가능성과 얼마나 잘 일치하는지를 측정

Result

Table 1. Test performance of each method on CIFAR-100 for 5 trials with $p = 94\%$. The mean(%)(standard error(%)) of related metrics are reported. The best method for the misclassification error and budgeted errors are highlighted in boldface.

Method	Expert	Error	Budgeted Error			Coverage
			10%	20%	30%	ECE
CE	20	22.31(0.54)	27.64(1.08)	22.44(0.62)	22.31(0.54)	79.21(1.13) 5.45(0.40)
	40	20.65(0.98)	36.35(2.24)	29.27(2.34)	22.00(1.55)	66.93(2.83) 9.61(1.28)
	60	16.22(0.19)	49.57(1.87)	42.17(1.78)	34.31(1.93)	48.21(2.00) 11.09(0.32)
OvA	20	24.33(2.01)	24.33(2.01)	24.33(2.01)	24.33(2.01)	93.09(0.48) 4.53(0.37)
	40	25.82(2.23)	29.00(3.10)	25.82(2.24)	25.82(2.23)	82.90(2.35) 8.36(0.95)
	60	19.33(1.86)	28.78(2.87)	21.85(2.63)	19.33(1.86)	74.81(1.78) 7.64(0.60)
A-SM	20	21.94(0.24)	21.94(0.24)	21.94(0.24)	21.94(0.24)	98.35(0.11) 4.17(0.21)
	40	21.22(0.77)	**21.34(0.90)**	21.22(0.77)	21.22(0.77)	90.64(0.99) 5.57(0.50)
	60	18.40(0.74)	**22.20(1.31)**	18.40(0.74)	18.40(0.74)	83.95(0.91) 5.11(0.23)
DCE (Proposed)	20	**21.21(0.23)**	**21.44(0.22)**	**21.21(0.23)**	**21.21(0.23)**	88.15(0.21) 1.68(0.27)
	40	**19.09(0.37)**	22.71(0.40)	**19.10(0.38)**	**19.09(0.37)**	80.95(0.86) 4.59(0.35)
	60	**15.81(0.31)**	24.89(0.62)	**18.20(0.57)**	**15.81(0.31)**	74.59(0.70) 5.57(0.23)

Result

Table 2. Test performance of each method on CIFAR-100 for 5 trials with $p = 75\%$. The mean(%)(standard error(%)) of related metrics are reported. The best method for the misclassification error and budgeted errors are highlighted in boldface.

Method	Expert	Error	Budgeted Error			Coverage
			10%	20%	30%	ECE
CE	20	23.57(0.38)	24.50(0.40)	23.57(0.38)	23.57(0.38)	85.07(0.38) 6.12(0.13)
	40	23.21(0.80)	29.47(1.48)	24.39(1.16)	23.21(0.80)	76.58(3.28) 11.88(0.59)
	60	21.18(0.18)	32.08(0.43)	26.25(0.30)	21.19(0.19)	70.47(0.42) 15.97(0.12)
OvA	20	23.63(0.35)	24.54(0.35)	23.63(0.35)	23.63(0.35)	85.05(0.32) 5.91(0.38)
	40	22.73(0.42)	28.86(1.05)	23.78(0.79)	22.73(0.42)	76.38(1.53) 12.00(0.77)
	60	21.18(0.27)	32.40(0.48)	26.53(0.56)	21.24(0.28)	70.09(0.66) 16.06(0.34)
A-SM	20	22.21(0.16)	22.21(0.16)	22.21(0.16)	22.21(0.16)	99.50(0.07) 3.85(0.32)
	40	22.49(0.47)	**22.49(0.47)**	22.49(0.47)	22.49(0.47)	96.42(0.37) 5.06(0.10)
	60	21.06(0.54)	**21.06(0.54)**	21.06(0.54)	21.06(0.54)	92.54(0.22) 5.20(0.45)
DCE (Proposed)	20	**22.12(0.27)**	**22.13(0.26)**	**22.12(0.27)**	**22.12(0.27)**	90.08(0.29) 1.70(0.15)
	40	**21.28(0.48)**	22.91(0.61)	**21.28(0.48)**	**21.28(0.48)**	84.19(0.47) 6.88(0.65)
	60	**19.81(0.85)**	24.08(1.14)	**19.84(0.89)**	**19.81(0.85)**	80.27(0.59) 11.43(1.81)

Result

Table 3. Test performance of each method on CIFAR-100N for 5 trials. The mean(%)(standard error(%)) of related metrics are reported. The best method for the misclassification error and budgeted errors are highlighted in boldface.

Method	Error	Budgeted Error			Coverage
		10%	20%	30%	ECE
CE	25.61(0.52)	31.97(0.60)	25.94(0.54)	25.61(0.52)	79.69(1.13) 27.29(0.50)
OvA	32.07(0.53)	54.32(1.31)	46.97(1.33)	40.24(1.17)	55.35(2.07) 32.23(0.25)
A-SM	28.50(0.34)	49.06(1.12)	41.66(1.06)	34.91(0.99)	58.38(1.53) 31.73(0.37)
DCE	**21.34(0.34)**	**26.94(0.41)**	**21.88(0.48)**	**21.34(0.34)**	78.74(0.63) 21.94(0.42)

Conclusion

- L2D에 대한 새로운 Bayes optimality 공식 제공
- 인간-모델 의존성 패턴을 기반으로 위임 원칙 제시
- L2D를 위한 새로운 손실 dependent cross-entropy (DCE) 제안
- 제안된 방법이 여러 벤치마크에 대하여 베이스라인보다 우수함 입증