Heterogeneous Risk Minimization

problem

OOD问题

给定训练数据集:$D = \{D^e\}_{e\in \text{supp}(\mathcal{E}_{tr})}$,其中$D^e = \{(x_i^e, y_i^e)\}_{i=1}^{n_e}$,学习预测器$f() : \mathcal{X} \to \mathcal{Y}$,能够达到:

$$\arg \min_{e\in \text{supp}(\mathcal{E})} \mathcal{L}(f|e) \tag{1}$$

其中$\text{supp}(\mathcal{E}_{tr}) \subset \text{supp}(\mathcal{E})$

解决OOD泛化问题,常常需要借助以下两个假设:

> **Assumption 2.1**存在随机变量$\Phi^*(X)$,满足
>
> *a.* Invariance property: *for all* $e, e' \in \text{supp}(\mathcal{E})$, *we have* $P^e(Y|\Phi^*(X)) = P^{e'}(Y|\Phi^*(X))$ *holds.*
>
> *b.* Sufficiency property: $Y = f(\Phi^*) + \epsilon,\ \epsilon \perp X.$

Heterogeneous Risk Minimization.

基于Assumption2.1,作者给出了关于训练数据中的异质性假设:

> **Assumption 2.2.** Heterogeneity Assumption.
> *For random variable pair* (X, Φ^*) *and* Φ^* *satisfying Assumption 2.1, using functional representation lemma (El Gamal & Kim, 2011), there exists random variable* Ψ^* *such that* $X = X(\Phi^*, \Psi^*)$, *then we assume* $P^e(Y|\Psi^*)$ *can arbitrary change across environments* $e \in \text{supp}(\mathcal{E})$.

同时比较直观能想到**当$|\mathcal{I}_{\mathcal{E}}|$越小说明更强的异质性,有更多的变量被排除在外**.基于此作者给出了**HRM问题**

> **Problem 1** Heterogeneous Risk Minimization.
>
> $$\textit{Given heterogeneous dataset } D = \{D^e\}_{e\in \mathrm{supp}(\mathcal{E}_{latent})} \textit{ without environment labels, the task is to generate environments } \mathcal{E}_{tr} \textit{ with minimal } |\mathcal{I}_{\mathcal{E}_{tr}}| \textit{ and learn invariant model under learned } \mathcal{E}_{tr} \textit{ with good OOD performance.}$$
>
> 简单来说就是要找出能让$|\mathcal{I}_{\mathcal{E}}|$最小的数据集,并在这个数据集上进行模型训练.

HRM问题本质上是存在反馈作用的,使用更小的集合$\mathcal{I}_{\mathcal{E}}$获得$\mathcal{E}_{tr}$,可以帮助我们获得更加具有Invariant属性的预测器$\Phi(X)$,从而对数据中的Invariant部分会有更加明确的了解从而促进$\mathcal{E}_{tr}$的确定.

related work

基于这两个假设,有很多寻找maximal Invariant predictor的工作,例如Invariant Rationalization,maximal invariant predictor等.
关于maximal Invariant predictor以及不变集合的定义如下:

Definition 2.1. *The invariance set $\mathcal{I}$ with respect to $\mathcal{E}$ is defined as:*

$$\begin{aligned} \mathcal{I}_{\mathcal{E}} &= \{\Phi(X) : Y \perp \mathcal{E} | \Phi(X)\} \\ &= \{\Phi(X) : H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}]\} \end{aligned} \tag{2}$$

where $H[\cdot]$ is the Shannon entropy of a random variable. The corresponding maximal invariant predictor (MIP) of $\mathcal{I}_{\mathcal{E}}$ is defined as:

$$S = \arg\max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi) \tag{3}$$

where $I(\cdot;\cdot)$ measures Shannon mutual information between two random variables.

这个关于maximal Invariant predictor的定义也正好对应关于$\Phi^*(X)$的假设

作者证明了以下的关于$\Phi^*(X)$的定理:

Theorem 2.1. *(Informal) For predictor $\Phi^*(X)$ satisfying Assumption 2.1, Φ^* is the maximal invariant predictor with respect to $\mathcal{E}$ and the solution to OOD problem in equation 1 is $\mathbb{E}_Y[Y|\Phi^*] = \arg\min_f \sup_{e \in \text{supp}(\mathcal{E})} \mathbb{E}[\mathcal{L}(f)|e]$.*

但theorem要发挥作用往往依赖于一个假设:**从$\mathcal{E}_{tr}$学习到的Invariant set $\mathcal{I}_{\mathcal{E}_{tr}}$与$\mathcal{E}$下对应的$\mathcal{I}_{\mathcal{E}}$**完全相同.但这实际上很难保证,即:

Theorem 2.2. $\mathcal{I}_{\mathcal{E}} \subseteq \mathcal{I}_{\mathcal{E}_{tr}}$

并且很多情况下,一般environment labels往往不是available的

> **Theorem 2.3.** *Given set of environments* $\text{supp}(\hat{\mathcal{E}})$*, denote the corresponding invariance set* $\mathcal{I}_{\hat{\mathcal{E}}}$ *and the corresponding maximal invariant predictor* $\hat{\Phi}$*. For one newly-added environment* e_{new} *with distribution* $P^{new}(X, Y)$*, if* $P^{new}(Y|\hat{\Phi}) = P^{e}(Y|\hat{\Phi})$ *for* $e \in \text{supp}(\hat{\mathcal{E}})$*, the invariance set constrained by* $\text{supp}(\hat{\mathcal{E}}) \cup \{e_{new}\}$ *is equal to* $\mathcal{I}_{\hat{\mathcal{E}}}$*.*

Method

本文中作者主要关注$X = [\Phi^*, \Psi^*]^T \in \mathbb{R}^d$,即在原特征之上寻找.主要的过程由$\mathcal{M}_c$(异质数据识别)与$\mathcal{M}_p$(不变预测).结构如下:

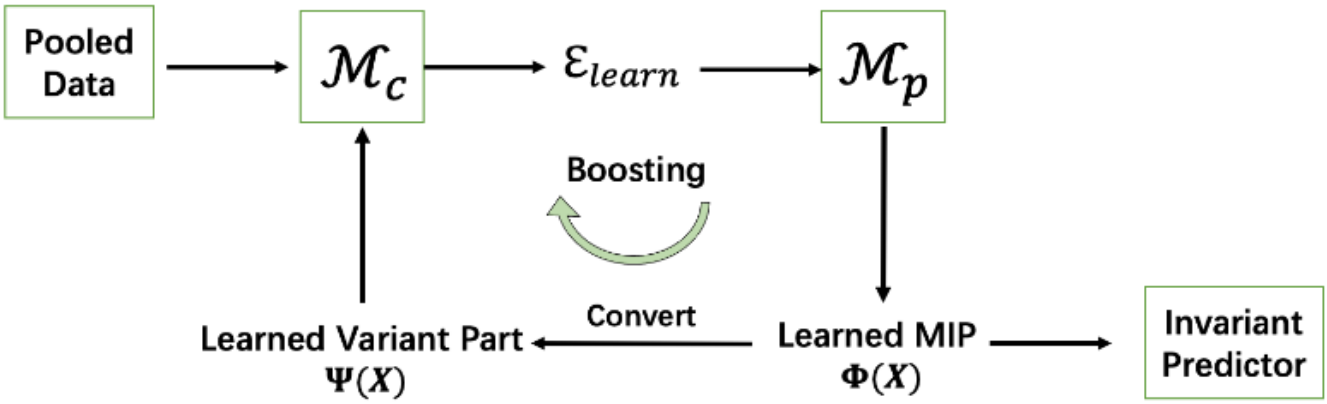

Figure 1. The framework of HRM.

因为是基于原特征的所以寻找$\Phi(X)$等价于找一个mask$M$：$\Phi(X) = M \odot X, \Psi(X) = (1 - M) \odot X$于是就可以直接进行convert.

Implementation of Mp

在这一部分作者将feature selection与Invaraint Learning结合,**尝试学习即Invariant且含有最大信息量的feature**

$$\begin{aligned}\mathcal{L}^e(M \odot X, Y; \theta) &= \mathbb{E}_{P^e}[\ell(M \odot X^e, Y^e; \theta)] \\ \mathcal{L}_p(M \odot X, Y; \theta) &= \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e] + \lambda \text{trace}(\text{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e))\end{aligned} \tag{2}$$

其中需要学习的参数就是θ, M,其中$\lambda \text{trace}(\text{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e))$就类似于IRM的约束项用于确保学到的$\theta$能在多个环境中达到最优.

由于hard feature selection会引起很大的方差,所以考虑soft feature selection:

$$m_i = \max\{0, \min\{1, \mu_i + \epsilon\}\} \tag{3}$$

其中$\mu = \left[\mu_1, ..., \mu_d\right]^T$,$\epsilon \in \mathcal{N}(0, \delta^2)$.于是单一环境中的损失函数为:

$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{Pe}\mathbb{E}_M[\ell(M \odot X^e, Y^e; \theta) + \alpha||M||_0] \tag{4}$$

其中$||M||_0 = \sum_{i\in[d]} \bar{\text{CDF}}(\mu_i/\sigma)$类似于L1正则化确保选择的特征是稀疏的.

最终损失函数的形式为

$$\min_{\theta,\mu} \mathcal{L}_p(\theta; \mu) = \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e(\theta, \mu)] + \lambda \text{trace}(\text{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e))(8) \qquad \text{where} \tag{5}$$

$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e}\mathbb{E}_M \left[\ell(M \odot X^e, Y^e; \theta) + \alpha \sum_{i\in[d]} \text{CDF}(\mu_i/\sigma)\right] \tag{6}$$

Implementation of Mc

$\mathcal{M}_c$将单一数据集作为输入,输出多环境的数据集组合.作者通过聚类算法来实现.**聚类的目标是让$P(Y|\Psi)$更加多样化.于是根据$P(Y|\Psi)$进行聚类**.初始阶段,$\Psi(X) = X$

作者使用混合高斯模型对数据点进行聚类,首先获取各个样本点所属类别构成的分布:$\hat{P}_N = \frac{1}{N}\sum_{i=1}^N \delta_i(\Psi, Y)$,其中

$$\delta_i(\Psi, Y) = \begin{cases} 1, & \text{if } \Psi = \psi_i \; and \; Y = y_i \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

混合高斯分布表示为$\mathcal{Q} = \{Q|Q = \sum_{j\in[K]} q_j h_j(\Psi, Y), \mathbf{q} \in \Delta_K\}$,于是使用KL散度来尽量接近这两个分布:

$$\min_{Q\in\mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \tag{8}$$

KL散度可以简化为

$$\min_{\Theta,\mathbf{q}}\left\{\mathcal{L}_c=-\frac{1}{N}\sum_{i=1}^{N}\log\left[\sum_{j=1}^{K}g_j h_j(\psi_i,y_i)\right]\right\} \tag{9}$$

最后依赖贝叶斯公式确定数据点属于哪个分布,或者说来自哪个类:

$$P(e_j|\Psi,Y)=q_j h_j(\Psi,Y)/\left(\sum_{i=1}^{K}q_i h_i(\Psi,Y)\right) \tag{10}$$

想法

1. **尝试学习即Invariant且含有最大信息量的feature**这个其实在IR模型中也可以使用.
2. 这篇文章相较于一般的IRM就只做了一方面的改进:利用聚类算法以及迭代的手段来划分出最佳的数据集供Invariant Learning来使用.