Instance-Dependent Positive and Unlabeled Learning With Labeling Bias Estimation

List of Symbols

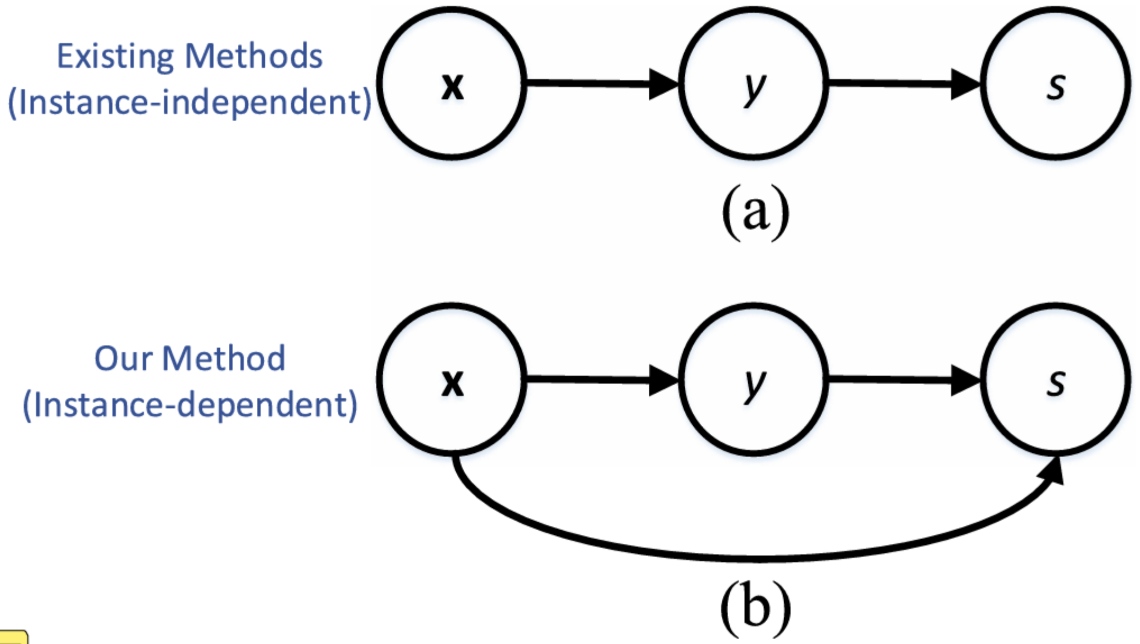
- s - 样例是否被表注 1标注 0非标
- y - 样本标签 1正 0负
- x - 样本特征
- k - the sizes of positive set
- n - the sizes of entire training set
- $S_P = \{x_i\}_{i=1}^{k}$ 正例集合
- $S_U = \{x_i\}_{i=k+1}^{n}$ 未标记集合
- η - 表示被观测到的概率，
  $P(s=1|y=1,x) \neq P(s=1|y=1)$ 并且
  $p(s=1|y=1,x) = \eta(x)$
- θ_1 - $P(y=1|x;\theta_1)$,给定随机变量x和固定参数θ_1的条件下，$y=1$的概率
- θ_2 - $\eta(x;\theta_2)$,给定随机变量x和固定参数θ_2的条件下，被标记为正例的概率
- $h(x)$ - 概率得分函数，$sgn(h(x)-0.5)$ 大于0.5记为正类

PU问题

SCAR假设 & SAR假设：

- 在SCAR假设下，正样本是完全随机从所有正样本中选取的，这意味着每一个正样本被选中作为标记样本的概率是相同的，与其特征无关。
- SAR假设认为，虽然正样本是从所有正样本中选取的，但选取的概率可能与某些属性相关，即正样本的选择不完全随机，可能依赖于实例的特征。

这两种情况可以通过以下的结构说明:



η 表示被观测到的概率

SCAR假设下：$P(s=1|y=1,x)=P(s=1|y=1)=\eta$

即，

$$P(s=1|y=1)=\eta=\frac{P(s=1,y=1)}{P(y=1)}=\frac{P(s=1)}{P(y=1)}$$

$P(s=1)\&P(y=1)$ 可以直接从数据中估计出来

SAR假设下：$P(s=1|y=1,x) \neq P(s=1|y=1)$，并且定义$p(s=1|y=1,x)=\eta(x)$

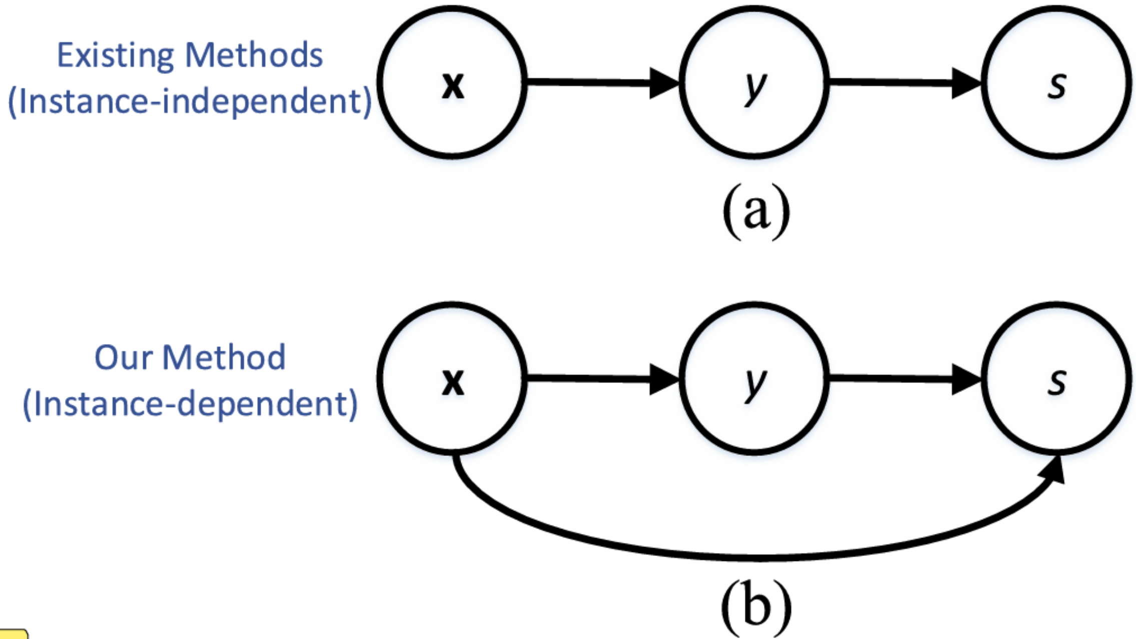
$\eta(x)=p(s=1|y=1,x)=\frac{P(S=1,y=1|x)}{P(y=1|x)}=\frac{P(s=1|x)}{P(y=1|x)}$

在这里 $\eta(x)$ 和后验概率$P(y=1|x)$ 共现，所以这篇文章重点是为了找到一个方法来联合估计这两个概率。

θ_1 - $P(y=1|x;\theta_1)$ - $h(x)$ - score function

θ_2 - $\eta(x;\theta_2)$ - labeling model

回到结构图



可以得到：

$P(y,s|x)=P(y|x)P(s|y,x)$ - s的分布依赖于x和y的值，而y的分布只依赖于x

因为所有样本都是独立抽取的，所以又可以表示为：

$$P(y,s|x)=\prod_{i=1}^{n}P(y_i,s_i|x_i)$$

$$=\prod_{i=1}^{n}P(y_i|s_i,x_i)\cdot P(s_i|x_i)$$

回到SAR假设下的PU问题中，存在

- $P(s=0|y=1,x)=1$
- $P(s=1|y=0,x)=0$
- $P(s=1|y=1,x)=\eta(x;\theta_2)$
- $P(s=0|y=1,x)=1-\eta(x;\theta_2)$

合并一下：

$$P(s=s'|y,x)=\begin{cases}(1-\eta(x;\theta_2))^{1-s'}\eta(x;\theta_2)^{s'}, & y=1\\ 1-s', & y=0\end{cases}$$

那么问题转为，如何估计θ

通过propensity score 把 $h(x)$ 和$\eta(x)$ 定义为

$h(x;\theta_1)=P(y=1|x)=(1+exp(-\theta_1^Tx))^{-1}$

$\theta(x;\theta_2)=P(s=1|y=1,x)=(1+exp(-\theta_x^Tx))^{-1}$

为什么可以这样定义

Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data

AND DONALD B. RUBIN
University of Chicago, Chicago, Illinois, U.S.A.

SUMMARY

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Both large and small sample theory

实际上倾向性得分是一个特殊的二分类模型

目标是最大化这个函数：

$$\arg\max_{\theta}\prod_{i=1}^{n}P(s_i|x_i;\theta)=\arg\max_{\theta}\prod_{i=1}^{n}\sum_{y_i}P(s_i,y_i|x_i;\theta).$$

取对数

$$\arg\max_{\theta}\mathcal{L}(\theta)=\sum_{i=1}^{n}\log\sum_{y_i}P(s_i,y_i|x_i;\theta).$$

通过观察法，1.取对数 2.有参数 有隐变量y

通过EM算法求解θ