## 4.5 限制性样本问题

限制性样本问题是说我们可获得的训练样本存在某种限制，比如：正负样本不均衡、只有少量正样本、不同分组的样本存在分布差异性等等，这类问题目前已有一些较为成熟的算法可以解决。比如，正负样本不均衡问题，可以通过欠采样、过采样或SMOTE算法来解决；只有少量正样本问题可以通过PU-learning方法来解决；因果推断中实验组和对照组的协变量分布存在差异性可以采用Propensity Score Analysis和Entropy balancing来解决。下面就各种限制性样本问题，给出几种参考的解决方案。

### 4.5.1 PU-learning

#### 4.5.1.1 简介

正例和无标记样本学习（Learning from Positive and Unlabled Example）简称PU-learning，是一种半监督的二元分类模型，通过标注过的正样本和大量未标注的样本训练出一个二元分类器[11]。

与普通分类问题不同，PU问题中P的规模通常相当小，扩大正样本集合也比较困难；而U的规模通常很大，比如在网页分类中，未标识的网页资源可以非常廉价、方便的从网络中获取。引入U的目的就是降低人工分类的预备工作量，同时提高精度，尽可能达到自动分类的效果。

PU-learning方法在许多领域内都有应用[13]，比如：

1. 检索：从大量无标注的样本中选取特定的样本，比如人脸标注。
2. 异常检测：异常点检测，恶意url检测，致病基因检测等。
3. 序列数据检测：负样本的分布随着时间改变，这样传统的分类将不再适合，PU 只需要更新未标注样本，这样的花销更小，比如垃圾邮件检测，由于存在对抗，负样本（垃圾邮件）的形式一直在变，而非垃圾则一般相对稳定状态。

#### 4.5.1.2 实现方法

目前 PU-learning方法可分为两类：

* 直接法：利用正样本集和未标记样本集直接训练分类器，对最后得到的概率进行校正。在满足一定条件假设下，通过数学公式推导可以证明，由正样本+未标记样本训练出的分类器和正样本+负样本训练出的分类器存在一个常数概率转换关系。
* 两步法：从未标注数据中识别出有效的负样本，然后利用正负样本迭代训练分类器。

这里我们重点介绍两步法的使用，两步法的两个步骤如图4.28所示，具体概括为：

* 第1步：根据已标注过的正样本P在未标注样本集U中找出可靠的负样本集合(Reliable Negative Examples，简称RN)，将PU问题转化为二分类的问题；
* 第2步：利用正负样本通过迭代训练得到一个二元分类器。

理论上已经证明：如果最大化未标注样本集U中负样本的个数，同时保证正样本被正确分类，则会得到一个性能不错的分类器。

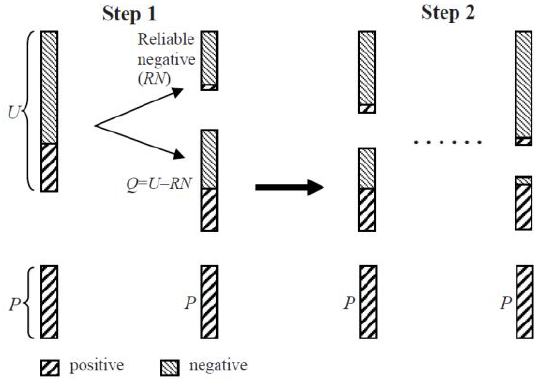


图4.28 PU-learning的两步法

上述两个步骤中，找出RN以及训练二元分类器都有很多方法可以选择，下面对这些方法进行简单的介绍[12]。

#### 4.5.1.3 计算 RN

（1）朴素贝叶斯分类器

使用朴素贝叶斯（Naive Bayesian，NB）分类方法计算 RN，可以简单参考以下步骤：

* 把 P 中的每个样本标记为类别 1；
* 把 U 中的每个样本标记为类别-1；
* 使用 P 和 U 训练得到贝叶斯分类器；
* 对 U 中的每个样本使用上述分类器进行分类，如果分类结果为-1，则把该样本加入 RN。

（2）Rocchio 技术

Rocchio 是一种早期的文档分类技术，其基本思想是：每个样本可以用一组特征向量来表示，特征值可以使用TF-IDF方式计算得到。

设全部样本集合为D，类别为的训练样本集合为。通过对每个类别构造一个原型向量，可以得到 Rocchio 分类器：



其中，α和β分别调节与类别相关及不相关类别的权重。

对一个待分类的样本t，使用余弦相似度计算其与每个类别的原型向量的相似距离，取距离最小的类别作为该样本的类别。

使用Rocchio算法与上述NB分类器计算RN的步骤很类似，只要把上述算法中第3步的分类器替换为 Rocchio 分类器即可。

（3）Spy算法

Spy的基本思想是从P中划分出一个子集S，将S中的样本放到U中，从而得到新的正样本集P-S和未标识样本集U+S。使用P-S作为正样本，U+S作为负样本，利用迭代的EM算法进行分类，当分类结束后，利用对那些「间谍」样本的标识，确定一个参数阈值th，再对U中的文档进行划分得到可靠的反样本集合RN。其中，从P中划分子集S的数量比例一般为15%。算法步骤描述如下：

* RN集合置空；
* 从P中随机选取子集S，得到新的正样本集PS=P-S和未标识样本集US=U+S，记 PS 中各个样本类别为1，US各个样本类别为-1；
* PS和US作为训练集，用I-EM算法训练得到一个贝叶斯分类器；
* 使用子集S确定出一个概率阈值th；
* 对US中的每个样本d使用贝叶斯分类器计算其属于正类别的概率P(1|d)，如果小于阈值概率th，则把其加入RN集合。

（4）1-DNF算法

1-DNF算法基本思想是：对于每个特征，如果其在P集合中的出现频次大于N集合，记该特征为正特征 (Positive Feature, PF)，所有满足该条件的特征组成一个PF集合。对U中的每个样本，如果其完全不包含PF集合中的任意一个特征，则该样本应加入RN。算法步骤描述如图4.29所示：

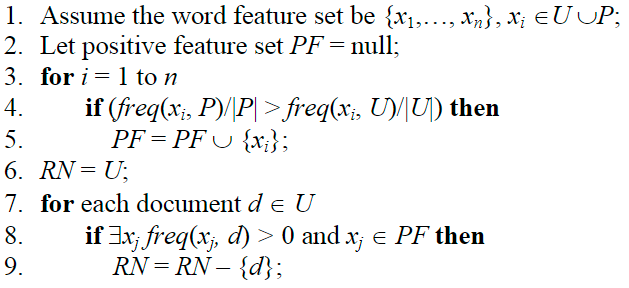


图4.29 1-DNF技术

#### 4.5.1.4 训练分类器

（1）SVM

使用 SVM 直接对 P 和 RN 进行训练得到分类器。

（2）S-EM

EM 算法主要由 Expectation 和 Maximization 两个步骤组成。前者对缺失标签的数据打上标签；后者则用全部数据一起计算模型参数。算法步骤描述如下：

* 对 P 中的每个样本标记为类别 1；
* 对 RN 中的每个样本标记为类别-1；
* Q=U-RN 中的样本起初没有任何类别标签，在 EM 算法第一次迭代完成时，这些数据将会具有一个基于概率的类别标签。在后续的迭代中，都使用上一轮更新完类别标签的数据集 Q，直至 EM 算法收敛。

在上述流程中，每次迭代使用 Naive Bayesian 算法修正Q集合中的数据标签。

（3）PEBL 算法

PEBL算法主要思想是使用SVM迭代地从U-RN中抽取出更多的负样本，并把它们放到RN集合中，直至U-RN中没有可以被分为负样本的数据为止。算法步骤如下：

* 对 P 中的每个样本标记为类别 1；
* 对 RN 中的每个样本标记为类别-1；
* 令 i=1，Q=U-RN，开始以下的循环：
  + 使用 P 和 RN 训练一个 SVM 分类器 Si；
  + 使用 Si 对 Q 中的样本进行分类，把其中所以分类为-1的样本记为W；
  + 如果W为空，则结束循环；否则：Q = Q-W, RN = RN ∪ W, i = i + 1。

（4）Roc-SVM 算法

PEBL 算法中得到的最后一个分类器不一定是最优分类器，为此，对该算法进行一些改进，得到了Roc-SVM算法。算法步骤如下：

* 使用 PEBL 算法直至收敛，记最后一个分类器为S\_last；
* 使用S\_last对P进行分类；
* 如果P中超过8%的样本被分为负样本，则选择S1作为最终的分类器；否则，选择S\_last作为最终分类器。

由于 SVM 算法对噪声很敏感，如果在迭代过程中，把Q中的一些正样本错分为-1而划分到RN中，那么会导致最终的分类器 S\_last 性能很差，这也是 PEBL算法的一个缺陷。为此，需要对S\_last的分类性能进行评估，看是否选择其作为最终分类器。选择8%作为判断阈值也是一个保守的做法，以防选择一个性能不好的分类器。

上述的选择S1或S\_last 的做法其实还是欠妥，因为这两个分类器可能性能都很差。S1性能差是因为初始的RN集合中可能包含很少的负样本，不能体现出负样本的整体分布情况；S\_last性能差则是由于PEBL算法在某个迭代过程中把一些正样本错分到RN中。为此，我们可以选择使用Spy或Rocchio算法得到初始的RN，这样可以使S1更加稳健。有一点要注意的是：多数情况下，最佳分类器可能并不是 S1 或 S\_last，而是在迭代过程中产生的某一个分类器，然而，这个最佳分类器却是很难被获取的。

### 4.5.2 PSM

#### 4.5.2.1 PSA简介

PSA (Propensity Score Analysis)，称作倾向性分析，其基本原理是将多个混杂因素的影响用一个综合的倾向性评分来表示，从而降低了协变量的纬度，减少了自变量的个数，有效的克服了分层分析和多因素调整分析中要求自变量个数不能太多的短板。常用的方法有：

* 倾向性评分匹配法 (Propensity Score Matching)
* 倾向性评分分层法
* 倾向性评分校正法
* 倾向性评分加权法 (Propensity Score Weighting)

总体来说，倾向性分析的方法，是通过计算出每个研究对象的倾向性评分，从而可以用倾向性评分一个指标来集中体现多个混杂因素的综合影响，然后再使用分层、匹配、校正或加权等多种方法进行分析，以达到控制混杂因素的目的。

#### 4.5.2.2 PSM原理

PSM，即倾向性评分匹配法 (Propensity Score Matching)，是倾向性分析（PSA）中应用最为广泛的一种方法。

为了研究某项措施或者某个行为对人群的影响（例如吸烟对健康的影响，读研对收入的影响），或者互联网中某项措施对于用户的影响，最直接有效的评估方法是我们从大量的样本中随机选择对照组（control group）和实验组（treated group），保证这些用户在实验前的一致性（通过aa test），然后进行AB test，之后来评估效果。

而在现实中，由于一些因素，并不能有效的开展上述实验，例如不能强迫可以读研的人不去读研吧，以及组织实验者来吸烟吧。在这种情况下，我们期望能够引入一种方法来找到有效的对照组，以此来评估该项措施的效果。

因此，PSM的主要思想就是：首先，对实验组和对照组的每个样本计算一个综合的倾向性评分，从小到大进行排序；然后，对实验组的每个样本，从对照组中选取与其倾向性评分最为接近的所有个体，并从中随机抽取一个或N个研究对象作为匹配对象，直至所有的研究对象均匹配完毕，未匹配上的研究对象则进行舍去；最后，保存匹配后的数据，供后续分析使用，总体流程如图4.30所示。其关键步骤有2个：

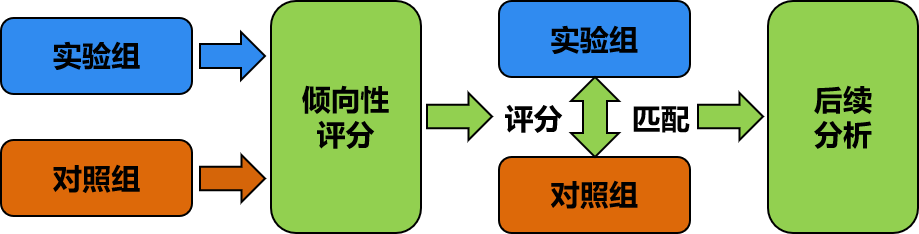


图4.30 PSM的流程

（1）估计倾向性评分

估计评分的方法有判别分析（discriminant analysis）、逻辑回归（logistic regression）和随机森林（random forests）等。最好的方法有待讨论，但比较流行的方法之一是逻辑回归。

（2）评分匹配

匹配算法也有多种，例如：精确匹配 (exact matching), 完全匹配(full matching), 遗传迭代匹配 (genetic matching), 最邻近匹配(nearest neighbor matching), 最优匹配(optimal matching), 和子分类匹配(subclassification) 等。

另外，需要注意的是，有多少研究对象可以成功匹配，常常与选择匹配的比例和匹配的标准有关。匹配的比例最常见的为1:1匹配，需要根据两组人群的数量来决定合适的匹配比例，建议不要超过1:4匹配。

对于匹配标准，如果匹配的标准很高，则能够成功匹配的对象就可能会少，甚至出现匹配不上的现象，造成研究对象信息的浪费，如果匹配的标准很宽泛，则匹配的效果就会较差，有可能出现两组人群在匹配后依然存在混杂因素分布不均衡的现象。

例如某个个体的倾向性评分为0.8，如果设定匹配标准为±0.02，则需要为其寻找倾向性评分在0.78-0.82之间的对照进行匹配，匹配范围太窄就可能出现匹配不上的情况；如果设定匹配标准为±0.2，则需要为其寻找倾向性评分在0.8-1.0之间的对照进行匹配，匹配范围太宽则可能降低匹配的效果。

#### 4.5.2.3 示例

PSM解决的是样本的选择偏误会带来的内生性问题，比如：在比较读研对工资的影响时，要选择能力、智商、家庭背景、工作单位等都差不多的样本进行比较，所以需要样本匹配，PSM方法就可以平衡掉干扰协变量的影响。

就上面这个例子而言，我们想探究小明读研和没读研的工资差距是多少？但这是一个反事实的问题，因为事实上他已经读了。而且这种问题是不能做实验比较的，因为我们不能让一个考上研究生的不去读研，来帮我们做对照组。另外，随机选择读研和没读研的人，然后比较他们的工资也是不合理的，因为这两组人群的其他因素也会造成影响。

因此，我们使用倾向评分匹配，从一大堆没读研的人（样本子集）中，对每个人读研的概率进行估计（比如采用逻辑回归），找到与小明有差不多读研概率但没读的小华，作为小明的对照。所以，这个问题的操作步骤为：

1. 对总体样本进行逻辑回归，估计出每一个观测对象读研的概率；
2. 根据读研概率，把读研的和没读研的配对起来，得到实验组和对照组；
3. 针对两组样本，比较是否读研对工资的影响（其实目的就是消除其他变量的影响）。

### 4.5.3 Entropy balancing

#### 4.5.3.1 简介

熵均衡（Entropy balancing）方法同倾向性评分加权法很像，就是通过调节混杂因素（协变量）构成的权重来调整两组观察效应的平均水平，从而消除两组样本之间由于内部混杂因素分布不同对效应值的影响。

两种方法的区别在于，倾向性评分加权法是通过倾向性评分来获取每个样本的权重，而熵均衡方法是基于信息熵来获取每个样本的权重。当每一个样本的权重计算出来之后，就可以使用加权回归的方法来进行因果推断。下面我们来看一下如何通过熵均衡方法来获取样本权重。

#### 4.5.3.2 原理

如前所述，我们希望调整实验组和对照组的样本权重，消除两组样本协变量之间的差异性，换句话说，我们想让两组样本的协变量保持一致，而这个目的可以通过平衡协变量的矩来实现。算法的基本思路就是：调整对照组样本的权重，让实验组和对照组的协变量的一阶矩、二阶矩、三阶矩乃至更高阶矩保持一致，意味着两组样本的期望、方差、偏斜度、峰度等保持一致，即信息量一致。这样就可以消除协变量的影响。下面是算法过程[14]：

我们的优化的目标函数为：

 (1)

约束条件为：

 (2)

 (3)

 (4)

其中，每个样本i属于两个不同的组，表示样本i在实验组，表示样本i在对照组；是为对照组的每个样本选择的权重；是距离度量方法；是重加权对照组的协变量矩的R阶平衡约束，是实验组的协变量的R阶矩；

因此，熵均衡求解权重的方法转化为一个有约束的优化问题，可以使用拉格朗日乘数法解决。对于上面的熵均衡方法的重加权机制有三点需要说明：

第一点，损失函数中的是距离度量方法，这里我们采用KL散度 来估计权重和原始权重（KL散度是用来度量使用基于Q的编码来编码来自W的样本平均所需的额外的比特个数。典型情况下，W表示数据的真实分布，Q表示数据的理论分布）。所以，损失函数衡量的是：对照组估计的权重的分布与原始的权重之间的距离，其中。注意：损失函数要是非负值，从而使W越来越接近Q。

第二点，重加权机制引入了等式(2)的平衡约束条件。这个式子可以平衡实验组和重加权对照组的协变量分布的矩（假设相关矩存在），一个典型的平衡约束是：包含实验组的给定协变量的R阶矩，对照组的矩函数定义为或者。

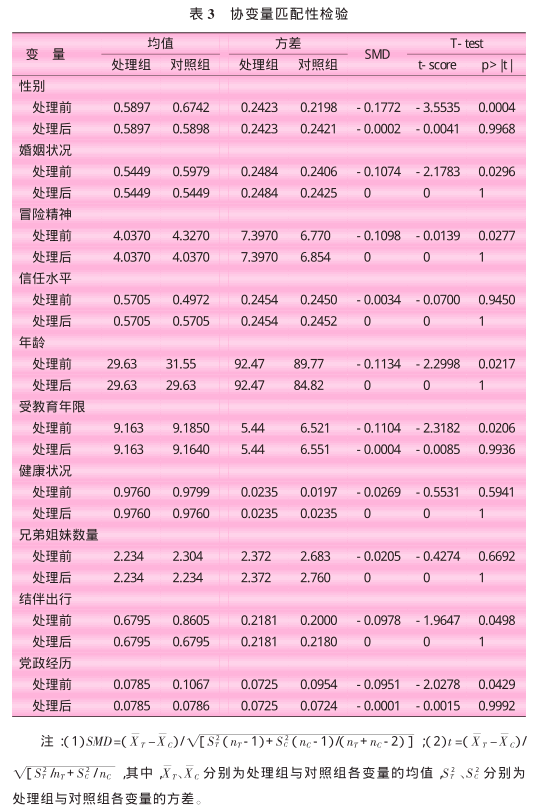
第三点，等式(3)和(4)的标准化约束条件。第一个条件要求权重的和为1，第二个条件要求权重的非负性，因为距离度量不能定义负的权重值。

熵均衡方法可以理解为对传统倾向性加权方法的推广。传统倾向性加权方法首先用逻辑回归估计单位权重，然后进行平衡检验，看估计的权重是否确实使协变量分布相等；而熵均衡方法从反面来解决这个权重调整问题，通过预先定义的平衡约束来直接估计权重。研究人员并不认为精确估计的逻辑回归评分就能随机的平衡协变量，而是认为直接利用他们对样本矩的认识，预先指定一组潜在的平衡约束，这些约束可以使得重加权对照组中的样本矩与实验组中的相应矩完全匹配。熵均衡方法搜索一组可以满足平衡约束的权重，但同时又要尽可能接近(在熵意义上)该组原始的平均权重，从而可以保持后续分析的效率。这种方法的主要优点是：它直接将单位权重调整为已知的样本矩，从而在有限的样本中获得精确的矩匹配。因此，传统意义上的平衡检查不再需要。

#### 4.5.3.3 示例

陆文聪，谢昌财（2017）在社会关系、信息网络对新农民工收入的影响一文中使用了熵均衡数据处理方法[17]。在探索“社会关系”对新农民工的收入影响效应时（新农民工在其务农城市“有亲戚关系”与“没有亲戚关系”的收入差别为多少），对其协变量性别、婚姻状况、冒险精神、信任水平、年龄、受教育年限、健康状况、兄弟姐妹数量、结伴出行、党政经历采用熵均衡进行处理，处理前后的均值、方差及匹配性检验结果如表4-9。

表4-9 协变量匹配性检验



匹配前协变量在一阶矩和二阶矩上均有较大差异，经熵均衡法匹配后，处理组与对照组协变量的均值、方差基本一致。为进一步验证熵均衡结果的可靠性，可以计算熵均衡处理前后处理组与控制组的标准化均数差（SMD），并对均值差异进行t检验。经熵均衡调整后所有协变量p值为1（或无限逼近于1），表明处理组与对照组各个协变量数据已经实现100%匹配。数据预处理完成后，接下来就可以对重加权后的数据进行重新回归，得到的因果推断结果更具有可信度。