规则学习 Rule Learning

2020 年 4 月 12 日

1 15.1 基本概念 Basic Concepts

1.1 规则 If-then

$$\oplus \leftarrow f_1 \wedge f_2 \wedge \cdots \wedge f_L \tag{1}$$

1.2 覆盖 Cover

规则 1 覆盖西瓜数据集 2.0 中的样本 1，2，3，4，5
规则 1：好瓜 ← 根蒂 = 蜷缩 ∧ 脐部 = 凹陷
规则 2：not 好瓜 ← 纹理 = 模糊
我们将上面两条规则组成的规则集记为 R

1.2.1 覆盖可能造成的两个问题

1. 冲突：同一示例被不同规则覆盖，且判别结果不同。冲突消解：投票法、排序法、元规则法等
投票法：判别结果相同的规则数最多的为最终结果
排序法：定义一种排序，最靠前的为最终结果
元规则法：根据领域知识设定一些“规则的规则”，即指导规则的使用
2. 不完全覆盖：规则集不能完全覆盖整个数据集
增加默认规则：“未被规则 1，2 覆盖的都不是好瓜”

1.3 规则的表达能力

1. 命题规则：原子命题 + 逻辑连接词，例：R

2. 一阶规则：原子公式 + 量词 + 逻辑连接词，例：

$$\forall X(N(\sigma(X))) \leftarrow N(X),\ where\ \sigma(X) = X + 1 \tag{2}$$

3. 命题规则是一阶规则的特殊形式，一阶规则表达能力比命题规则要强

2 15.2 序贯覆盖（以命题规则学习为例）

2.1 基本思想

逐条归纳，通过每次训练生成一条仅覆盖正例的规则，就去除掉那些覆盖的样例，用剩下的继续训练

2.2 以西瓜数据集 2.0 的训练集为例的序贯覆盖

1. 好瓜 $\leftarrow$ 色泽 = 青绿 覆盖了 1、4、6，10、13、17, 并不是全部都是正例

2. 好瓜 $\leftarrow$ 色泽 = 青绿 $\wedge$ 根蒂 = 蜷缩 覆盖了 1、4，17，并不是全部都是正例

3. 好瓜 $\leftarrow$ 色泽 = 青绿 $\wedge$ 根蒂 = 蜷缩 $\wedge$ 敲声 = 浊响 覆盖了 1，此时都是正例，就把这条规则加到规则集中，然后继续用剩下的样例进行训练

缺陷：基于穷尽搜索，在属性和样例数量较多的时候可能会组合爆炸

2.3 自顶向下与自底向上

自顶向下 top down：也就是上面这个过程，从一般到特殊，覆盖范围从大到小，其泛化性能更好（理由是更容易产生能较短的能覆盖更多正例的规则）

自底向上 bottom up：从特殊到一般，覆盖范围从小到大，适用于一阶学习这种假设空间比较复杂的情况

2.4 以西瓜数据集 2.0 的训练集为例的自顶向下

详见 TopDown 函数的输出

3 15.3 剪枝优化

3.0.1 基本思想

通过剪枝前后发生的性能变化来判断是否进行剪枝，可以缓解过拟合的风险

规则生成本质上是一个贪心搜索过程，也就是说，每一步都去寻找一个局部最优解，分阶段去逼近最优解。例如上面的分析过程，我们每次只在一些规则里面找到最好的规则（而不是去找全部的可能的规则进行比较），然后把它加进规则集里面，最后我们得到的规则集不一定是最好的解，但可能会是一个可行解。我们把这个局部最优解当成全局最优解来使用，可能会造成过拟合，于是通过剪枝来提高规则集的泛化性能。

预剪枝是指生长过程中剪枝，后剪枝是指规则产生后剪枝。

3.0.2 CN2 算法

在预剪枝时，假设用规则集预测必须显著优于直接用训练集的后验概率分布（也就是训练集正反例数量的比率）进行猜测。使用了似然率统计量 LRS 来表示两种预测方法之间的差别。

$$LRS = 2(\hat{m}_+ \log_2 \frac{(\frac{\hat{m}_+}{\hat{m}_+ + \hat{m}_-})}{(\frac{m_+}{m_+ + m_-})} + \hat{m}_- \log_2 \frac{(\frac{\hat{m}_-}{\hat{m}_+ + \hat{m}_-})}{(\frac{m_-}{m_+ + m_-})}) \tag{3}$$

当 LRS 越大的时候，说明规则集预测与直接用训练集分布进行猜测的差别越大。在数据量比较大的现实任务中，通常设置 LRS 很大（例如 0.99）时才停止。

3.0.3 REP

减错剪枝 REP 是后剪枝常用的策略，其基本做法是：将样例集划分为训练集 T 和验证集 V，从 T 上学得规则集 R 后进行多轮剪枝，每一轮穷举可能的剪枝操作，包括删除规则中的某个或多个文字，删除整条规则等，然后用 V 对剪枝产生的规则集进行评估，保留最好的规则集进行下一轮剪枝，直到无法通过剪枝提高验证集上的性能为止。

3.0.4 RIPPER 算法