## 2020.2.16 第三次读书报告

09018330 孙毅远

### 一、自己提出的问题的理解

#### 1.决策树的生成是否可用其他方法例如搜索回溯？

理解:应该不能使用,回溯的话复杂度过高不便处理

### 二、别人提出的问题的理解

#### 1. 3.2.4节中，处理决策树额外问题中的Handling Skewed Class Distribution的第二种方法Another solution is to rank the new cases according to how likely they may be intrusions.要如何理解？如何对新的样例进行排序？排序后又对调整模型有什么好处？

理解:我认为需要看出现的比重，因为不同应用中不同类型的数据占比区别可能很大，因为是入侵数据的比率很小，决策树做出来就是\一个叶子结点，所以对新的数据，可能是入侵数据的进行单独研究，在决策树不适用的情况下需要采用别的方法去进行研究

#### 2. 比起单纯看模型的accuracy来说，计算recall和precision可以提供哪些额外的信息来进一步评估模型？

理解:需要看具体应用的特性决定使用哪种标准，因为这两种标准测试的偏向并不相同，需要结合实际情况看

#### 3. The general idea of post-pruning is to estimate the error of each tree node. If the estimated error for a node is less than the estimated error of its extended sub-tree, then the sub-tree is pruned.这段话怎么理解？不利用validatoin set怎么判断estimated error。

理解:post-pruning是在当前结点错误比子树少的情况下就把子树删去的后剪枝，是在整棵树建立完成的情况下剪枝

#### 4. 以cross-validation为例，按理来说k次交叉验证后应得到的k个模型分别有不同的accuracy，那取这k个accuracy的平均值有何意义？

理解:k次交叉验证为了更精确地检测当前方法是否适用，防止单次出现偶然数据

### 三、读书计划

#### 本周 3.1-3.3

#### 下周 4.1-4.3，5.1

### 四、读书笔记

#### 3.1 Basic Concepts

* 分类模型
* 算法用 进行学习，得到模型，再用 来评测模型的 (精准度)。
* 训练数据应与测试数据一致
* 学习的过程

#### 3.2 Decision Tree Induction

* 决策树学习算法所得到的分类模型是一棵树的形式，称之为决策树（）
* 和
* 一个决策节点内包含针对数据实例（）某个属性的测试，一个叶子节点代表一个类标（）
* 每个叶子结点下的（）代表到达此叶子节点的 个训练样例里有 个样例的 与叶子节点标识一致
* 对于一个数据集，可构建的决策树并非是唯一的，在实际应用中，希望得到一棵尽可能小且准确的决策树
* 在实际的数据集中，到达某个特定叶子节点的样例实际类标可能不同，也就是 ，实际上，这个值表示的就是关联规则挖掘的 ， 反映的是
* 构建一棵好的决策树的问题是一个NP完全问题，目前所有构建算法都是基于启发式算法的
* 学习算法（）
* 学习算法是使用分治（），递归对训练数据进行分隔，从而构造决策树，
* 递归结束的条件是产生了一个 ，分为三种情况
  + 一是数据集 中只含有唯一的类别 ，说明这一个划分已经完成，令该处为标为 的；
  + 二是，遍历过所有的 进行判断，但仍不能确定数据集 唯一的类别，则将该处标为数据集中数量最多的类的；
  + 三是 ，说明虽然还有 作为进一步分类的依据，但对提高集合纯度的已经没有作用，因此忽略不计，将该处标为数据集中数量最多的类标的
* 递归过程中计算得出最能提高这个集合纯度的 ，使它为此处的 ，再依次递归
* 混杂度函数（）
  + 信息增益（）基于熵（）
  + 是 类在数据集 中的概率，且定义 ，熵的单位是位（）
  + 熵可以作为数据混杂度或者混乱度的衡量指标
  + 由公式可知，当数据越来越纯净时，熵的值越来越小，基于此，考虑熵的信息增益
  + 对于给定的数据集 ，首先计算 的混杂度，表示成 （即 中第7行）
  + 再把所有属性计算一次（即 中第8-10行），找到哪个属性来划分 可以使得混杂度减少最多
  + 设属性 可取 个值，假设用 来划分 ，划分后 的熵变为：
  + 由此，属性 的信息增益就是：
  + 信息增益率（）
  + 信息增益率利用数据集的相对于属性值的分布的熵归一化信息增益，用以修正极端数据情况下无意义的划分（例如划分成多个单元素集合）
  + 其中 是属性 可能取值的数目，选择可使 达到最大的属性来扩展决策树
* 处理连续属性（）
* 之前讨论的都是离散属性，对于连续属性的应用，在一个树节点中可以将属性 的值划分为两个区间（），只要找到一个合适的分割阈值（）

#### 3.3 Classifier Evaluation

* 分类精度评估方法
* 查准率，查全率， 和平衡点