第三次读书报告

侯宇捷

一．自己提出的问题

1. 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validation set怎么判断estimated error。

讨论后的解答：可以使用train set来进行预估，同样可以用估计错误。

2.    为避免rule pruning后一个例子符合多条rule而对rule进行排序，此时如何排序？按照什么准则？

讨论后的解答：根据讨论，可以使用leaf node处的confidence，也可以人工根据要求划分优先级。

3. 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？它并不能代表k个模型中的任意一个。

讨论后的解答：可以估算出这个生成方法的整体准确率，可与其他非决策树方法生成的rule的准确率进行比较选取方法。

二．他人提出的问题

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？

讨论后的解答：有人提出可以根据人工或其他算法进行排序。以网络入侵为例，排序后我们所需的是网络入侵的例子增加，训练时可以集体投入，针对训练，以免出现全是非网络入侵的例子在训练集中，导致分类出来规则全与非网络入侵这一class有关，缺少实际需要的rule。

2. 为什么选择使得gain Ratio达到最大的属性，这是怎么证明的？

讨论后的解答：并非全看gain ratio，考虑到gain ratio对可取值数目较少的属性有所偏向，而gain对可取值数目较多的属性偏向，因而可用启发式算法，先找出gain高于平均水准的，再取其中gain ratio最高的。

3.决策树的生成是否可用其他方法例如搜索回溯？

讨论后的解答：就书上的内容来看，决策树是一种贪心算法，不考虑回溯操作。

三．读书计划

本周完成的内容章节：第三章全部要求内容

下周计划：第四章全部要求内容，视时间来读第五章

四．读书摘要：

决策树中的一些基本概念：accuracy，purity，entropy，gain，gain ratio

对于本章决策树的内容与逻辑梳理：

首先是决策树一步步向下延伸的关键——purity。让训练数据在每个分支下对应的class种类越少，称这个越pure，而这也是决定决策树每一个decision node以何attribute来划分的基础。也就是说，我们应该选划分后让决策树更加pure的属性来做分支。

此处面对第一个问题：如何量化并比较每一个属性在某一次划分时对决策树purity的贡献。此时引出第一个entropy（D）来表示这个决策树在划分这一步之前的impurity，并用entropy（Ai，D）来表示如果用第i个attribute来划分这一层划分后的impurity，由此推出gain=entropy（D）-entropy（Ai，D）来衡量划分用的attribute。如果gain越大，说明减少的impurity越多，由此选造成的gain最大的attribute来划分这一层的decision node。然而，考虑到这样的gain较为倾向于value数更多的attribute，面对如id这样的特别多的且不到特定时刻用处不大的attribute也会划分，因此又引入了gain ratio的概念。其中，gain ratio计算式的分母被称为split info，而它可能为零或极小，此处可引入启发式算法来解决该问题（具体书中未多提）。这样层层递归下去，某一层直到训练数据都只有一个class了，或者划分用的attribute都用完了才停止（此时直接用最多的class来表示这里的leaf node）。

解决完上述问题后，决策树的算法基本成型，然而还有一些问题需要考虑：首先是如果attribute的value种类并非如“对、错”， “多、少”这样是离散的，而是如数字型是连续的，该如何处理？此时可采取离散化的方法，将数字分成一个个区间（而在C4.5中经常采取二分即可）。考虑到虽然数字是连续的，但value是有限的n个，则有n-1种方法将这n个value二分，依此计算其gain ratio即可。而要注意的是，如果是离散型的attribute，决策树以之划分好一层后，显然第二次在用它划分是无意义的。然而，连续型的因有多种划分方式，之后还可以继续以这attribute来划分。

其次，要考虑到过度拟合的问题。如果训练数据中有分类错误，或者考虑到本身分类中存在的随机性与复杂度，在上述算法不用完属性或彻底在branch下只有一个class不停止的情况下，有可能造成过度拟合，即对训练数据万分匹配，却与unseen data set匹配极低。因此，对决策树本身进行prune。其中，简单分pre-pruning与post-pruning两种。后者主要是依靠估计误差与子树误差，还有利用validation data set的方法。此外，还需对rule进行prune，但要考虑到prune后的rule可能导致一个例子的数据对应多条rule或无对应的rule，此时设置default class与对rule进行排序。

还有，在实际情况下可能产生一个例子中有些属性的数据未被采集到的情况与倾斜类别分布问题。前者对离散值而言，可采取用频繁值代替或标注unknown特殊值来处理；对连续值来说，可采取平均值来代替。后者则可人工筛选出所需的positive class对应的数据投入训练集中，或为可能符合的数据进行排序，投入训练集中。

最后，对于classifier效果的评估。为避免positive class过少，虽然accuracy极高，但对positive class预估准确的极少的情况，引入precision与recall与breakeven point的概念。此外，引入multiple random sampling与cross-validation的方法以评估classifier的效能，以确定最后选取由什么方法生成的模型。