**2019/02/16读书报告**

学号 71117228 姓名 李泓烨

读书进度：3.3

1. **读书报告内容**

**1. 自己提出问题的理解**

① 为什么在Level2-candidate-gen()函数中，第五行sup(l)不会比sup(h)小？

讨论后的理解：这里只是说sup(l)可能比sup(h)少，并不一定是绝对的，这里我看的是英文版的书，翻译的时候理解有误

② 怎么理解MScandidate-gen()函数中的pruning step？为什么会存在那个exception？

讨论后的理解：对于c的每个(k-1)-子集子集s，如果s不在中，且s包含c[1]，**也即c中拥有最低MIS值的项目**，则c就能从中删除。而如果这个s不包含c[1]，即使s并不在中，我们也不能把c删除，因为我们不能断定s会否满足MIS(c[1])，尽管我们知道s不满足MIS(c[2])，除非MIS(c[1]) = MIS(c[2])(算法第9行)

**2. 别人提出问题的理解**

1. **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？

**讨论后的理解**：看出现的比重，不同应用中不同类型的数据占比的区别可能很大，排序之后考虑。因为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，可能是入侵数据的单独研究

2. **提出的问题2**：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？

**自己的理解**：在书上的例子中，recall很低，所以分类不是很完整，从而说明模型存在一些问题，而如果单纯考虑准确率的话，可能这个问题不明显

3. **提出的问题3**：对于决策树生成算法而言，如何计算sorting of a continuous attribute takes |D|log|D|的时间复杂度？

**自己的理解**：这个取决于不同的排序算法，排序算法一般情况都是Dlog|D|的时间复杂度

补充：按attribute属性值对实例进行排序？->比如有一个属性，顺序可能不对，所以需要排序；把属性的值理解为是连续的区间，不过选择阈值的时候只会选择一个点

4. **提出的问题4**：如何理解that interpolation is needed if such a point cannot be found？

**自己的理解**：插值——这是针对数据缺失人工补足的一种方法，在书后面可能会提到，暂时还没有看到

5. **提出的问题5**：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？

**自己的理解**：剪枝过程有两个，一个是利用training set来进行，还有一个是利用validation set来进行，区别就是validation set不能用来进行训练。按理说来，子树的错误应该更小，但是如果分割出来的子树错误更大，那么也就没有必要再分割了，可以进行剪枝。不用validation set，利用training set本身出现的error也可以判断estimated error

6. **提出的问题6**：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？

**自己的理解**：当满足多条规则的时候，某些例子必然会满足多条规则，此时需要对规则进行排序。但是规则的排序不一定要严格按照一个标准，只要保证例子可以符合其中的一个规则即可，也可以像第二章根据它的support或者confidence

7. **提出的问题7**：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(P22)

**自己的理解**：计算k个accuracy取平均值，就是相当于对整个数据集计算了accuracy

8. **提出的问题8**：在计算entropy时为什么需要用log，还有log的底数是否要和最后类的数量相同。

**自己的理解**：这并不能解释，超出了自己的知识范围，可能在信息论领域中，可以解释信息熵的公式是怎么得到的吧

9. **提出的问题9**：为什么一个attribute的值越多，其增益就越大，感觉好像是这样的，有没有数学证明?

**自己的理解**：这并不能严格地证明吧，书上也只是说了有一个趋势而已，不过一般情况下一个数据集属性值越多增益就是越大的，极端情况可以考虑每一个数据对应的属性值都是不同的

10. **提出的问题10**：Breakeven Point的计算对于分类器评估的意义何在？

**自己的理解**：这也只是一种标准，一个模型如果好的话，recall和precision都应该尽量的高

11. **提出的问题11**：在Rule Pruning的过程中，例10里面，为什么Rule 1可以删除条件Y>2.6?难道就是根据例9？

**自己的理解**：这是书上的一个例子，这样做只是更加合理一些

补充：书上没讲，感觉就是随缘，期待着我们能够发现更加具有标准化的剪枝方法

12. **提出的问题12**：为什么选择使得gainRatio达到最大的属性，这是怎么证明的？

**自己的理解**：并不一定要选择gainRatio达到最大的属性。正确的顺序应该如下：先选信息增益高于平均的，再选信息增益率最高的。利用数据集的相对于属性值分布的熵归一化信息增益，已修正信息增益的偏袒性

13. **提出的问题13**：在连续属性值的问题中，根据阈值修改算法8到11行究竟是怎么修改的？

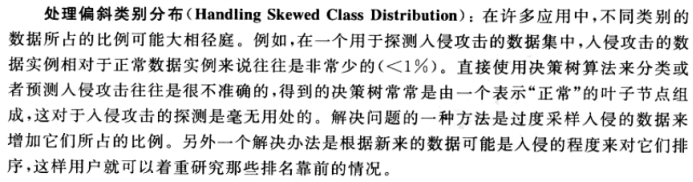
**自己的理解**：因为已经对数据集中的属性值进行了排序，个人认为在已经排好序的数据集上选择阈值会更加地简单和合理，代码层面的话也比较容易实现

14. **提出的问题14**：在连续属性值的问题中，根据阈值修改算法决策树的生成是否可用其他方法例如搜索回溯？

**自己的理解**：不行

15. **提出的问题15**：在3.2.3节中，他说任何的threshold value在vi和vi+1之间的有同样的效果，帮我解释一下？

**自己的理解**：只要把这个区间划分成两个部分就可以，具体怎么切都行。虽然属性是连续的，但是数据集中的具体值还是离散的

16. **提出的问题16**：

最后一句如何理解？

**自己的理解**：和一开始的问题相同

**3. 读书计划**

① 本周完成的内容章节：看完了2.6和第3章

② 下周计划：看完第4章