第二周读书报告

学号 71117203
姓名：袁佳怡

一、自己提出的问题的理解

1、提出的问题 1：如何理解

A point to note is that for a table data set, the join step of the candidate generation function (Fig. 2.3) needs to be slightly modified in order to ensure that it does not join two itemsets to produce a candidate itemset containing two items from the same attribute.

讨论后的理解：因为每个 item 包含了 attribute 和 value，应该要避免产生的候选项集中存在两个项是同一个属性不同的 value 值，虽然对于 attribute 和 value 的 pair 而言，出现 value 不同那么就是不同的 pair，但是不能有两个同一属性的两个项目在同一个候选项集内。举例：比如有一个年龄和身高的表格，不能得出 年龄：11 和年龄：41 在一起出现的最多的结论，这是不合常理的，所以在生成候选集的时候就应该直接被剪枝。

二、别人提出的问题的理解（选择几个问题罗列，并给出理解）：

2、问题 2：为什么在 MScandidate 函数中每次循环中使|sup(ik-1) − sup(i'k−1)| ≤ ϕ 就能令一个 itemset 中 maxsup 与 minsup 的差值在 ϕ 的范畴内。

自己的理解：因为由两个频繁项目集 f1 和 f2 生成的候选集 c，对于F_{k-1}而言，其中的任意两个的 item 都已经满足|sup(x)-sup(y)|<=φ，那么生成的候选项目 c 中唯一可能不满足的就是|sup(ik-1) − sup(i'k−1)| ≤ ϕ，所以只要考虑这个就可以了。

3、问题 3：算法中为什么记录了 f – {a}就能确保所需的非 frequent 的 condition 的 count 被记录下来

自己的理解：因为它是迭代产生的，对于每一个频繁项目集减去其中 MIS 值最小的 count 都进行了统计，使得在 rule 生成中，对于前件非频繁项目集的 count 有记录。

4、问题 4：为什么对于使用 MIS 的 item 来说，向下闭包性质不再适用。

自己的理解：因为对于一个项目集而言，它是频繁的是要满足真实的 support 满足集合中项目的最小 MIS 即可，但假设这个项目集有一个子集，他的真实 support 大于等于原项目集的 support，但子集中最小的 MIS 大于原先集合的，因为它不包含原先集合中 MIS 最小的那个项，那么此时子集的真实 support 和它其中所有项的最小 MIS 之间，可能是小于关系，即子集非频繁项目集，此时向下闭包性质不再适用。当然这可以依据书上给出的一个反例进行理解。

5、问题 5:如何理解 Fig.2.8 的 9-11 行?

自己的理解：这是一个剪枝的过程，每一个候选的项目集 c，对于 c 的每个（k-1）子集 s，如果 s 不在F_{k-1}中，且 s 包含 c[1]，也即 c 中拥有最低 MIS 值的项目，则就能从C_k中删除。因为如果 s 包含了 c[1]但它不在F_{k-1}中，说明 s 的真实 support 小于了 MIS[1]，那么 c 的真实 support 必然小于等于 s 的真实 support 也必定小于 MIS[1]，所以此时能

够确定这个 c 一定不会是频繁项目集，所以可以剪枝。但这里如果这个 s 不包含 c[1]，则即使 s 不在F_{k-1}中，我们也不能把 c 删除，因为 s 不包含 c[1]且 s 不在F_{k-1}中，说明 s 的真实 support 小于 MIS[2]，但并不知道 s 的真实 support 与 MIS[1]之间的关系，此时 c 的真实 support 小于等于 s 的真实 support 小于 MIS[2]，但不能判断 c 的真实 support 是否大于 MIS[1]，所以暂时不能将这样的一个 c 给剪枝掉，需要放到后面的步骤中进行筛选。

6、 问题 6：如何理解在 Level2-candidate-gen 函数中，第五行 sup(l)与 sup(h)的大小关系？
自己的理解：这里的 l 和 h 只是代表了两个 item 之间 MIS 值的大小 MIS(l)小于等于 MIS(h)，但并不能确定 sup(l)与 sup(h)的大小关系，所以这里需要添加一个绝对值。

三、读书计划
1、本周完成的内容章节：2.3-2.5
2、下周计划：2.6 和第三章