**《数据挖掘概论》作业1**

学号姓名：18373722 朱英豪 作业提交截止时间：2020年10月21日

1. 某个地方的公司销售主管与你联系，他相信他已经设计出了一种评估顾客满意度的完整方法。他这样解释他的方案：“这太简单了，我简直不敢相信，以前竟然没有人想到我只是记录顾客对每一种产品的抱怨次数，我在数据挖掘书中读到计数具有比率属性，因此我的产品满意度度量必定具有比率属性。但是当我根据客户满意度度量评估产品，并拿给老板看时，他说，我忽略了显而易见的东西，说我的度量毫无价值，我想他简直是疯了，没发现我们的畅销产品满意度极差，因为对它的抱怨最多，你能帮我摆平他吗？“
2. 谁是对的？销售主管还是他的老板？如果你的回答是他的老板，你需要做些什么来修正满意度度量？
3. 对于原来的产品满意度度量的属性类型，你的想法是什么？

**【答】**

**a）我认为二人互有对错。**

**销售主管：抱怨的数量多的计数不代表其满意度差。因为畅销产品售出的数量多，即基数大，故即使一个很小的抱怨比率，抱怨数量都会比较大。**

**老板：使用抱怨次数的比率是有意义的，并非毫无价值——抱怨次数与总售出次数之比可以更好地衡量满意度。**

**如果我是老板，我会通过计算抱怨次数与总售出次数之比来衡量满意度。**

**b）满意度度量是关于次数的函数，其属性类型：比率类型，是数值类型；单纯的次数亦是离散属性。**

**对于仅统计抱怨次数的满意度度量模型，正如我先前所提的，会出现因为畅销产品售出的数量多，即基数大，故即使一个很小的抱怨比率，抱怨数量都会比较大的情况，但该抱怨数量之多不能代表其满意度差，相反，其抱怨比率之低更能显示其满意度高。**

1. 几个月之后。上一题中提到的那个销售主管又同你联系，这次他设计了一个更好的方法，用于评估顾客喜爱一种产品超过喜爱其他类似产品的程度。他解释说：在开发一种新产品时，我们通常创建一些变种，并评估顾客更喜欢哪一种。我们的标准做法是同时散发所有的产品变种，并要求他们根据喜爱程度对产品的变种划分等级，然而我们的评测题目很不明确，当有两个以上产品时，尤其如此，这种测试占用了很长时间。我建议对产品逐对比较，然后使用这些比较来划分等级，这样，如果我们有三个产品变种，我们就让顾客比较变种1和2，然后2和3，最后3和1。使用我们的方法，评测时间是原来的1/3，但是进行评测的雇员抱怨说，他们不能从评测结果得到一致的等级评定，昨天我的老板想要知道最新的产品评估，另外我还得告诉你老的产品评估方法就是他提出的，你能帮助我吗？
   1. 销售主管是否陷入困境？他的方法能够根据顾客的喜好产品产生产品变种的有序等级吗？解释你的观点。
   2. 是否有办法修正销售主管的方法？对于基于逐对比较创建序数度量，你作何评价？
   3. 对于原来的产品评估方案，每个产品变种的总等级，通过计算所有测评题目上的平均值得到，你是否认为这是一种合理的方法，你会采取哪种方法？

**【答】**

**a）销售主管陷入了困境。他的方法不一定能够产生产品变种的有序等级。**

**因为是对产品的两两比较，得出的是如产品1>产品2或产品1<产品2这样仅好坏的二元结果，可能存在这种情况：****产品1<产品2，产品2<产品3，产品3<产品1。即此时，对于三元产品组(i,j,k)，不存在i，使得i>j且i>k，即找不到最优的一个产品。**

**b）有办法进行修正。如在评测各个产品时，进行等级赋值。如产品1与产品2进行比较时，进行评分。如满分10分，给产品1打5分，给产品2打7分。对于3个产品变种的情况，两两比较后各个产品都会有2个评分值，可对该评分值进行取平均/取和等操作，再进行比较。**

**对于题中的基于逐对比较创建序数度量的方法，我的评价是：存在如先前所提到的死循环而得不出最佳产品的情况，是其弊一；对于很多产品，每次比较时只选取2个，评价时过于注重局部，忽略了整体情况，存在局限性，是其弊二；且当产品很多时，众多的仅优劣的比较结果使得到所有产品排序很困难，很可能产生若干产品优于其他产品数量相同的情况，而无法得出最佳产品。优点为比较2个产品会比较简单，比较次数为n(n-1)/2比较少。**

**c）我认为这是一种比较合理的方法。较之题中原来的方案（单纯比较高低），我会选择该计算平均分值的方法。但正如在b）中我所提及的，二元比较存在局限性，我认为可适度提高每次比较产品的数量，如在需要评价10个产品变种时，每次比较时可为3/4个产品同时比较，而非仅两两比较。**

**另外，选用计算中位数也可能更好，中位数能避免异常值（如某个分特别高，导致平均值比绝大多数的值都高的反常情况）带来的影响。**

1. 你能想象一种情况，标识号对于预测是有用的吗？

**【答】对于定期（最好是时间间隔相同，定周期采样的）时序数据，其序号与时间有很强的相关性——标识号从小到大，相应记录的时间也是从早到晚，在这种情况下标识号对于预测或是有用的。但这样的标识号实际上仍是代表了时间，与时间数据重复，实际上还是时间数据在预测。**

**考虑以下一种情况：在某中学，学生学号是按照入学考试时的考试成绩的排名来得到的。则此时若要预测未来某次考试时学生的成绩排名，学生学号对于预测是有用的。**