Q4&Q5 分析报告

一、 我的理解

方法 1、方法 2、方法 3 根据算法可以分为两类，方法 1 和方法 2 为第一类算法，方法 3 为第二类算法。

第一类算法可以看作自定义距离公式的 kmeans，Jaccard 距离公式无疑是最适合这种行为相似度分析的，通过商品编号 **pluno**、商品金额 **amt** 两个属性对不同用户的购买行为进行相似性分析，最终将用户归为不同的类别中。

方法 1 和方法 2 的区别在于判断用户是否拥有同一类购物偏好的标准不同，方法 1 通过判断两个用户是否购买同一件商品以及购买的力度判断相似度，而方法 2 将商品分类有粗到细分为四类，越粗的分类包含的商品越多，因此不同用户在越粗的分类中相似度越高，算法最后取一个平均作为最终相似度。理论上来看方法 2 更为科学合理，因为其标准考虑到多个维度。

方法 3 则是为了利用上用户购买的时间 **sldatime**，这个属性能够帮助我们更为准确地判断具体阶段用户的偏好，为此方法 3 设计了一个树形结构利用树的深度来表达用户购买时间属性，同时将辅助判断的商品金额 **amt** 换成购买频数 **freq**，效果上没有太大差距。算法在计算 Jaccard 相似度时因为引入树形存储结构，对交集和并集进行的符合逻辑的设计能够较好判断用户相似性。最后利用 BIS 公式对聚类结果能否再次分裂进行判断，解决了传统 kmeans 算法调试 k 参数的问题，整个算法整体上很完整对购买记录数据具有针对性处理。

二、 优缺点比较

算法	优点	缺点
使用一级 Jaccard 距离的 kmeans	1.数据处理简单处理速度快	1.相似度标准高容易造成用户距离普遍很高；使

	2.无监督聚类	用的用户购买属性较少 2.初始点选择随机导致聚类质量不稳定 3.参数 k 很难设定
使用四级 Jaccard 距离的 kmeans	1. 判断标准维度增加准确性提高，属性使用更加合理 2. 无监督聚类	1.只是单一属性的维度增加，使用的属性不过全面 2.初始点选择随机导致聚类质量不稳定 3.参数 k 很难设定
使用树形结构的新型 kmeans	1.不仅使用一个属性的多个维度，还使用了更多的属性提升聚类质量 2.解决传统 kmeans 算法 k 值选择问题，聚类结果不需要监督 3.距离公式更加有针对	1.初始点选择随机导致聚类质量不稳定 2.用户购物信息还有很多可挖掘的属性没有使用

	性使得距离分布更加平均	

三、 我的思考&可能的改进

经过三个 kmeans 算法的实现我发现了 kmeans 对初始点的选择十分依赖，初始点选择的好坏直接决定聚类最终的质量。因此在相似度和距离公式无法改进的情况下，初始点的选择算法可以说是一个很好的突破点，如何选择初始点能够更好聚类。

虽然使用了 BIS 公式判断簇是否分裂是一很好的自动调参手段，但我感觉，购物记录这种带有较强的个人情感喜好的数据或许不仅可以通过数据本身判断还可以通过数据的组合判断用户购物喜好、习惯等方面更深层次地判断不同用户是否相似。例如：通过购买数量、是否打折、是否促销、促销类型组合判断用户是即买即用型、打折囤货型等等。同时性别和年龄也是判断一个用户购买行为的重要信息，男女的购物习惯普遍差距大。

我的另一个想法是现在已有的判断标准都是单层的，有点非此即彼的意思。我设想可以构件一个层次分类标准，一层层进行分类越深的层次划分标准越细，这样不会出现一刀切的情况，对不同层级的数据进行相应的分析使得最后得出的结果更加符合逻辑。