|  |  |
| --- | --- |
|  |  |
| 项目报告D |  |
|  |  |
|  | 课程：数据挖掘姓名：姜其升 |
|  | 学号：1752058指导教师：饶卫雄教授 |

[项目报告D 1](#_Toc39142553)

[1 距离分布情况比较 3](#_Toc39142558)

[2 聚类结果及质量比较 4](#_Toc39142559)

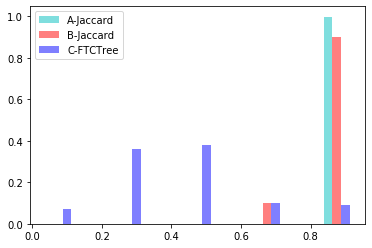
[3 优缺点比较 5](#_Toc39142560)

[3.1 方法A： 5](#_Toc39142561)

[3.2 方法B 5](#_Toc39142562)

[3.3 方法C 5](#_Toc39142563)

## 1 距离分布情况比较



对于A题：上图为Jaccard距离的分布情况，我们可以看到，分布较为集中，99%的顾客间距离分布在0.8~1这个区间之内，分布较为集中，按照此方法计算顾客相似度较小，距离较大，分布过于集中，不利于聚类，聚类结果可能较差，区分度不大。

对于B题：上图为Jaccard距离的分布情况，我们可以看到，分布较为集中，超过90%的顾客间距离分布在0.8~1这个区间之内，剩余不到10%则主要分布于0.5~0.8区间，其余部分几乎没有分布，分布较为集中，按照此方法计算顾客相似度较小，距离较大，分布过于集中，不利于聚类，聚类结果可能较差，区分度不大。

对于C题：上图为FTCTree距离的分布情况，我们可以看到，FTCTree距离的分布较为分散，大部分的顾客分布在0.2~0.6的区间之内，占比超过70%，且在这两个区间内的分布情况也较为平均。在其余区间内的分布也较为平均，每个占比不到10%。按照此方法计算顾客相似度适中，有梯度，分布较为松散，有利于聚类，可以得到较好的聚类结果，区分度较好。

综合比较：Jaccard系数无论采用何种方式计算其分布情况都较为集中，顾客距离较远，这可能时因为顾客在购买东西时差异较大，即使种类相似，但金额差异仍可能存在使得距离较远，可能导致聚类结果不理想，虽然在一个类中，但偏好差异明显或虽然在不同类中，但偏好差异并不十分明显。而对于TFCTree距离来说则避免了这一点，他的分布较为分散，不同距离中都有对应的分布，他考虑了顾客购买的频率而非金额，加上时间推移的因素，因而距离较为合适。

## 2 聚类结果及质量比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法 | 距离 | K | SC | CP |
| A | Jaccard距离 | 17 | 0.026 | 0.731 |
| B | Jaccard距离 | 16 | 0.063 | 0.669 |
| C | FTCTree距离 | 4 | 0.189 | 0.289 |

由于A和B方法中，顾客的距离较远，因此，SC的值较小和CP的值较大；C方法中，顾客距离适中，分布均匀，因此SC的值较大和CP的值较小，均表现出良好的性能。

在A和B方法中，寻找最优K值时，由于SC和CP的变化并不十分明显，因此较难寻找，可以采用其他的计算SC方式（如，b(i)代表据簇外的点的最小距离）描述其内聚程度或使用其他更优、变化更为明显的指标进行寻找。

在C方法中，最终的K值为4，每个簇中都有大量的顾客，如果想要进一步细分，则需要改进BIC的计算方式，以此获得更多的簇。也可以采用自定义K大小的方式进行聚类。

## 3 优缺点比较

### 3.1 方法A：

优点：相似度计算公式简单，速度较快；

缺点：距离分布过于集中，区分度不佳，需要多次迭代求的最优解，K的求解过分依赖经验和初始点的设置。

### 3.2 方法B

优点：相似度计算公式简单；

缺点：相似度需要多次计算，比较耗时，距离分布过于集中，区分度不佳，需要多次迭代求的最优解，K的求解过分依赖经验和初始点的设置。

### 3.3 方法C

优点：无需设置初始参数K，可自动得出K值，距离分布均匀，有辨识度；

缺点：树的构建和距离计算方式较为复杂。