RESEARCH NOTEBOOK

# The problems to be solved

* 1. In OCIL algorithm, the parameter of different attributes is fixed. According to the idea of K-prototype, should we introduce a weight of numerical and categorical attributes? And we can define the different importance of two attributes.
  2. The OCIL algorithm( even the k-prototype) didn’t consider the ordinal attributes. Generally, we always turn the ordinal attributes into the numerical attributes. However, some latest algorithm( like clustMD) revealed the new calculation of the ordinal attributes. So could we improve the original algorithm and reconsider the ordinal attributes?
  3. When we talking about the improvement of the algorithm’s performance, we should find a method to measure the performance of the clustering methods specifically. Many researches had explored the measurement, and we could compare the different algorithm under a certain framework.
  4. We always preassign the parameter k and weight of different attributes and improve those by iterative calculation which will consume a lot of computing resource. So could we have some methods to Improve the efficiency of iteration?

# Reference

1. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data
2. Model based clustering for mixed data: clustMD
3. A k-mean clustering algorithm for mixed numeric and categorical data, 2006, Amir Ahmad, Lipika Dey
4. 混合型数据聚类方法的比较, 2019, 刘 超, 姚清华

# Project Plan

谢谢教授能有这次research project的机会，经过这两天的学习从无到有对混合数据的聚类算法有了一些感性的认知，在聚类这个大框架下我感觉混合数据的聚类算法有两个方向比较重要，第一个是对不同种类数据之间权重的取舍，第二个就是样本点与聚类中心之间距离/相似度的计算。不同数据集之间对权重的要求也是不一样的，同时距离的计算方法也会影响权重的作用。接下来的初步的研究计划如下，非常感谢老师的指导

1. 搭建项目的测试环境，用代码复现OCIL原算法，编写测试代码，对聚类算法性能的几个重要指标（内部、外部、相对度量指标）进行计算。
2. 尝试加入权重和顺序型变量，用不同的数据集进行训练，比较改进后的聚类性能。
3. 比较最新的聚类算法，尝试使用不同的距离/相似度计算方法。
4. 研究K 和权重在迭代过程中的变化过程，尝试一个简单的优化方法或者一个简单的K的预计算方法，减少迭代次数。