软件学院 数据分析挖掘-编程作业之 2

1. [手机信号数据集] 本次作业编程作业 1（b）部分之上，设计 MR 聚类算法，并利用聚类结果重新更新作业 1 之定位模型。MR 记录按照相同的主基站分组之后，:

   a) 针对每个主基站分组后的 MR 数据，计算 MR 样本参考于该主基站的相对位置。利用所有 MR 数据的相对位置，设计和实现 K-Means 聚类方法，使得同簇内 MR 数据相对位置之间的相似度高、跨簇 MR 数据相对位置之间的相似度低，要求计算一个最优 K 值及其聚类结果；

   b) 针对每个主基站分组后的 MR 数据，将 MR 数据进行如何简化处理：对每个 MR 样本中的基站编号替换为对应的顺序号，比如主服务器基站对应的顺序号为 1，次服务基站顺序号为 2…，举例如下图所示。然后在处理后的 MR 数据根据信号强度值进行聚类，要求设计一个合理的 MR 数据信号强度的距离计算公式，使得在该距离计算公式基础之上的聚类结果，尽可能与步骤 a)聚类结果接近。

MRTime	2018/4/23 9:20	IMSI	xxx	SRNCID	6188	BestCellID	26051	LCS BIT	300
RNCID_1	6188	CellID_1	26051	AsuLevel_1	27	SignalLevel_1	4	RSSI_1	-74.5
RNCID_2	6188	CellID_2	27394	AsuLevel_2	10	SignalLevel_2	3	RSSI_2	-84.88
RNCID_3	6188	CellID_3	27377	AsuLevel_3	18	SignalLevel_3	4	RSSI_3	-85.13
RNCID_4	6188	CellID_4	27378	AsuLevel_4	12	SignalLevel_4	4	RSSI_4	-85.87
RNCID_5	6182	CellID_5	41139	AsuLevel_5	8	SignalLevel_5	3	RSSI_5	-88.88
RNCID_6	6188	CellID_6	27393	AsuLevel_6	9	SignalLevel_6	3	RSSI_6	-90.22
RNCID_7	6182	CellID_7	44754	AsuLevel_7	9	SignalLevel_7	3	RSSI_7	-95

MRTime	2018/4/23 9:20	IMSI	xxx	SRNCID	6188	BestCellID	26051	LCS BIT	300
RNCID_1	6188	CellID_1	26051	AsuLevel_1	27	SignalLevel_1	4	RSSI_1	-74.5
RNCID_2	6188	CellID_2	27394	AsuLevel_2	10	SignalLevel_2	3	RSSI_2	-84.88
RNCID_3	6188	CellID_3	27377	AsuLevel_3	18	SignalLevel_3	4	RSSI_3	-85.13
RNCID_4	6188	CellID_4	27378	AsuLevel_4	12	SignalLevel_4	4	RSSI_4	-85.87
RNCID_5	6182	CellID_5	41139	AsuLevel_5	8	SignalLevel_5	3	RSSI_5	-88.88
RNCID_6	6188	CellID_6	27393	AsuLevel_6	9	SignalLevel_6	3	RSSI_6	-90.22
RNCID_7	6182	CellID_7	44754	AsuLevel_7	9	SignalLevel_7	3	RSSI_7	-95

   提示：给定两个 MR 样本记为 R1 和 R2，如果二者所包括的基站数量并不相同，比如 R1 包含 7 个基站对应的信号强度，即基站最高顺序号为 7，而 R2 基站最高顺序号为 2,则仅考虑 R1 基站顺序号 1 和 2 对应信号与 R2 基站顺序号 1 和 2 对应信号之间的距离，而无需考虑 R1 基站顺序号 3…7 对应信号。

   c) 根据上述 b)聚类结果，将同簇的 MR 样本构建一个对应的定位模型（参考作业 1b），对比本次定位测试定位误差与作业 1b，并解释为什么。

提交日期：2022/05/22 日 23：59PM,提交内容发送至 tongjidam18@163.com，提交内容包括：

1、 每个作业提交内容以**学号+hw2.zip** 作为文件命名方法，并以**学号+hw2.zip** 作为邮件主题发送；其中包括每个小题的子目录，命名方式分别为对应小题的序号，每个子目录包括对应目的代码和 word 报告。其中报告包括 1）代码运行结果屏幕拷贝；2）讨论分析部分；3）性能比较图表