# 软件学院 数据分析/挖掘课程作业之4

1. [改进版定位算法] 本次作业利用手机与基站连接信号强度的测量报告数据(measurement report: MR)来预测手机所在GPS经纬度位置，其中通过信号强度数据提取相关特征信息，训练sk-learn 的RandomForest模型，预测GPS经纬度位置; 通过绘制误差概率分布图，确定中位误差。要求如下：
   1. 按照每个MR记录对应的主基站对MR记录进行分组，使得每组MR记录都有相同的主基站，总的的分组个数即为主基站个数。假定某主基站的经纬度坐标为<x0, y0>，该分组中的某MR记录对应的经纬度坐标为<x, y>，则计算该MR记录与主基站的相对位置为：<x‘ y‘>=< x-x0, y-y0>。在完成每个分组的MR记录相对位置计算之后，针对每个分组构建一个对应的随机森林回归模型（与作业3相似），不过该模型是以MR记录与主基站的相对位置作为标签。
   2. 使用处理好的训练集用于训练模型，测试数据集用于测试统计，通过上述随机森林模型预测测试数据的相对位置<x‘’，y‘‘>，然后计算还原为原始位置：<x‘’+ x0，y‘‘+ y0>。与作业3要求类似计算定位误差并排序。重复10次训练集/测试集的选择和误差计算，求平均误差，绘制平均误差概率分布图，其中x-轴为排序编号，y-轴为对应的平均误差，中位误差为x-轴50%排序编号的y轴对应平均误差；并在该图上，重现绘制作业3的回归定位算法定位误差，并比较和讨论这两个方法之间的优劣。
   3. 假定存在K个主基站，也即存在K个对应的MR记录分组和随机森林回归模型。考虑每个分组，记为Gi，（其中0<=i<=k-1），对应的主基站位置为<xi, yi>，利用上述步骤通过Gi对应的随机森林模型RFi预测Gi对应的20%测试数据MRi的原始位置<x‘’+ xi，y‘‘+ yi>；与此同时，尝试利用其他K-1个分组对应的随机森林模型RFj (其中0<=j<=k-1，且j≠i)对Gi中的测试数据MRi进行定位，假如某随机森林模型RFj对MRi预测的相对位置是<xji, yji>，然后可预测得该MR的原始位置<xji + xi, yji + yi>，比较通过K-1个RFj预测可得的原始位置<xji + xi, yji + yi>和通过RFi预测可得的原始位置<x‘’ + xi, y‘’+ yi>。存在K个分组Gi，对每个Gi重复上述步骤，比较K-1个RFj预测结果中和RFi预测结果最为接近的分组序号，并加以讨论和分析。
2. [HMM模型]使用训练数据集作为训练数据，预测测试集的经纬度位置：
   1. 单层HMM：构建一个单层HMM模型，直接预测测试数据经纬度。
   2. 双层HMM：首先对地图进行栅格化处理（参考作业三的多分类方法），第一层HMM用于预测MR数据对应的栅格，第二层HMM用于从栅格到纬度值位置。
   3. 讨论和比较这两个方法的优缺点。
3. 2人一个小组，按照小组完成本次作业，提交日期：2017/06/26日 23：59PM，答辩日期：2017/06/27日，78节课，每组6分钟。

数据说明：

2g数据集位置：/home/student/DataMining/GSM/final\_2g\_tr.csv

/home/student/DataMining/GSM/final\_2g\_te.csv

4g数据集位置：/home/student/DataMining/LTE/final\_4g\_tr.csv

/home/student/DataMining/LTE/final\_4g\_te.csv

工参表位置：/home/student/DataMining/GSM/final\_2g\_gongcan.csv

/home/student/DataMining/LTE/final\_4g\_gongcan.csv

数据集格式说明：

其中添加了GridID代表栅格，Grid\_center\_x,Grid\_center\_y为该栅格的中心坐标。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **MRTime** | **IMSI** | **SRNCID** | **BestCellID** | **SRNTI** | **RAB** | **Delay** | **UE\_TXPower** | **LCS\_BIT** | **Longitude** |
| 16:34.0 | 1234567890 | 99129 | 1 | -999 | -999 | -999 | -999 | 300 | 121.502 |
| **Latitude** | **RNCID\_1** | **CellID\_1** | **EcNo\_1** | **RSCP\_1** | **RTT\_1** | **UE\_Rx\_Tx\_1** | **….** | **GridID** | **Grid\_center\_x** |
| 31.21702 | 99129 | 1 | -10.88 | -87.5 | 37900 | 153 |  | 1 | 121.492345 |
| Grid\_center\_y | | | | | | | | | |
| 31.20893 | | | | | | | | | |

基站的坐标位于工参表中，为了匹配数据集中主基站的位置，需要匹配工参表中LAC和CI这两列到数据集的RNCID\_1和CellID\_1中，当这两列完全一致时，可以确认在工参表中找到了唯一的主基站，该基站在工参表中的所有信息都可用于建模。

其中**MRTime** 记录该MR记录的时间；**IMSI**为手机唯一标识符；**RNCID**为服务网络控制器(serving network controller)ID，**BestCellID**为最优基站编号，SRNCID和BestCellID一起构成一个唯一标识服务主基站编号；**Longitude**和**Latitude**为手机所在经纬度，GridID为该GPS位置所在栅格编号。每个MR记录包括最多6个非空连接基站信息，分别对应**RNCID\_[1-6], CellID\_[1-6]，EcNo\_[1-6]，RSCP\_[1-6]，RTT\_[1-6], UE\_Rx\_Tx\_[1-6]**，其中手机连接信号强度**RSSI=RSCP – EcNo.**