# 软件学院 数据分析/挖掘课程作业之3

1. [分类和回归算法] 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. sk-learn 提供随机森林分类器(可参考[SKLearn RandomForestClassifier](http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html))和随机森林回归(可参考[SKLearn RandomForestRegressor](http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html" \l "sklearn.ensemble.RandomForestRegressor))两种方法，其中随机森林回训练MR数据“信号强度特征”与“对应GPS经纬度数值坐标”的回归模型、直接预测测试MR数据的未知GPS经纬度位置(即GPS经纬度数值坐标)；而随机森林分类器则训练MR数据“信号强度特征”和“对应GPS经纬度所在栅格”的分类模型、预测MR测试数据未知GPS经纬度位置所在栅格、最后计算预测的栅格中心位置作为该MR记录的位置。
   2. 随机选取80%的数据记录作为训练集，余下20%作为测试集合，计算预测位置和证实位置的误差（采用欧式距离），按照计算误差从小到大进行排序。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。本次作业要求将随机森林分类器和随机森林回归这两种定位方法进行对比和讨论。
   3. 本次作业提供2G GSM和4G LET网络的数据集，要求对比和讨论这两个不同数据集合的定位结果。
2. [频繁集挖掘] 利用上述MR数据对应的GPS位置所在栅格进行频繁集挖掘，手机标识符对应于频繁集挖掘中的交易ID，GPS位置所在栅格作为Item商品ID，这样可以将相同手机标识符的所有GPS位置所在栅格看成是一个完整的交易记录。具体如下：
   1. 首先不考虑MR数据的时序情况，利用Apiori和FPGrowth方法进行频繁集挖掘，要求support>=3，输出所有的频繁集结果，对比二者计算时间。
   2. 考虑MR数据的时序情况，利用GSP 和SPADE方法进行时序模式挖掘，其中要求support>=3，输出所有的频繁集结果，对比二者计算时间。
   3. [SPMF](http://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php)提供参考频繁集挖掘代码实现。
3. 提交日期：2017/05/02日 23：59PM
4. 附录[数据格式]MR数据格式及样例

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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** |  |
| 31.21702 | 99129 | 1 | -10.88 | -87.5 | 37900 | 153 |  | 1 |  |

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。**

数据集位置：101.37.20.204 ：/home/student/DataMining/GSM

/home/student/DataMining/LTE