

软件学院 数据分析挖掘-编程作业之 1

1. [手机信号数据集] 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 MR 信号强度提取相关特征信息, 训练 sk-learn 的分类和回归模型, 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:
 - a) [8 分]利用 sk-learn 提供的决策树分类器 [DecisionTreeClassifier](#)、高斯朴素贝叶斯分类器 Gaussian Naive Bayes (GaussianNB)、K 近邻分类器 KNeighborsClassifier、及随机森林分类器(可参考 [SKLearn RandomForestClassifier](#))共计 4 个分类器来训练 MR 数据“信号强度特征”与和“对应 GPS 经纬度所在栅格”的分类器模型, 预测 MR 测试数据所在栅格、最后计算预测的栅格中心位置作为该 MR 记录的位置。通过随机选取 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。本次作业提供 2G GSM 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, 讨论和比较上述 4 个训练模型的精度。其中 MR 信号强度数据特征信息可参考如下工作
 - i. 参考文献 1: <https://dl.acm.org/citation.cfm?doid=2983323.2983345>: Fangzhou Zhu, Chen Luo, Mingxuan Yuan, Yijian Zhu, Zhengqing Zhang, Tao Gu, Ke Deng, Weixiong Rao, Jia Zeng: City-Scale Localization with Telco Big Data. CIKM 2016: 439-448
 - ii. 参考文献 2: Yukun Huang, Weixiong Rao, Fangzhou Zhu, Ning Liu, Mingxuan Yuan, Jia Zeng, Hua Yang: Experimental Study of Telco Localization Methods. MDM 2017: 299-306, <https://ieeexplore.ieee.org/document/7962466/>
 - b) [7 分]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, 总的分组个数即为主基站个数。假定某主基站的经纬度坐标为 $\langle x_0, y_0 \rangle$, 该分组中的某 MR 记录对应的经纬度坐标为 $\langle x, y \rangle$, 则计算该 MR 记录与主基站的**相对位置**为: $\langle x', y' \rangle = \langle x - x_0, y - y_0 \rangle$ 。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 MR 定位模型, 不过该模型是以 MR 记录与主基站的相对位置作为标签。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 $\langle x'', y'' \rangle$, 然后计算还原为原始位置: $\langle x'' + x_0, y'' + y_0 \rangle$; 比较和讨论 a)和 c) 方法之间的优劣。

Tips: 划分栅格:

1. 利用所有 MR 数据的 GPS labels, 确定一个整体的位置范围, 并将该范围转换成一个大的矩形(所有数据的位置标签均落入该矩形)。
再将该矩形划分成若干个等大的小正方形栅格(边长为 20m 左右较为适合)。分类器训练时, 将所有 MR 数据的 GPS labels 转换为栅格 ID labels, 并划分训练集和测试集进行训练。测试时, 分类器输出对每一条 MR 数据预测的栅格 ID。
2. 计算误差时, 将分类器预测的栅格 ID 转换成该栅格的中心点坐标, 并用该中新点坐标代表预测出的位置。最后用真实的 GPS label 进行误差计算。

提交日期: 2022/04/17 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+hw1.zip作为文件命名方法, 并以学号+hw1.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