数据分析与数据挖掘第三次作业第一大题 – c问分析文档

数据分析与数据挖掘

DAM COURSE, SPRING 2018

BY

1552674 李　源

Tongji University
School of Software Engineering

1 代码运行结果

这里我参照了文档要求，选择了预测结果转为绝对经纬度后的中位误差作为评价指标。其各自的 MS 主基站的具体结果分布如下：

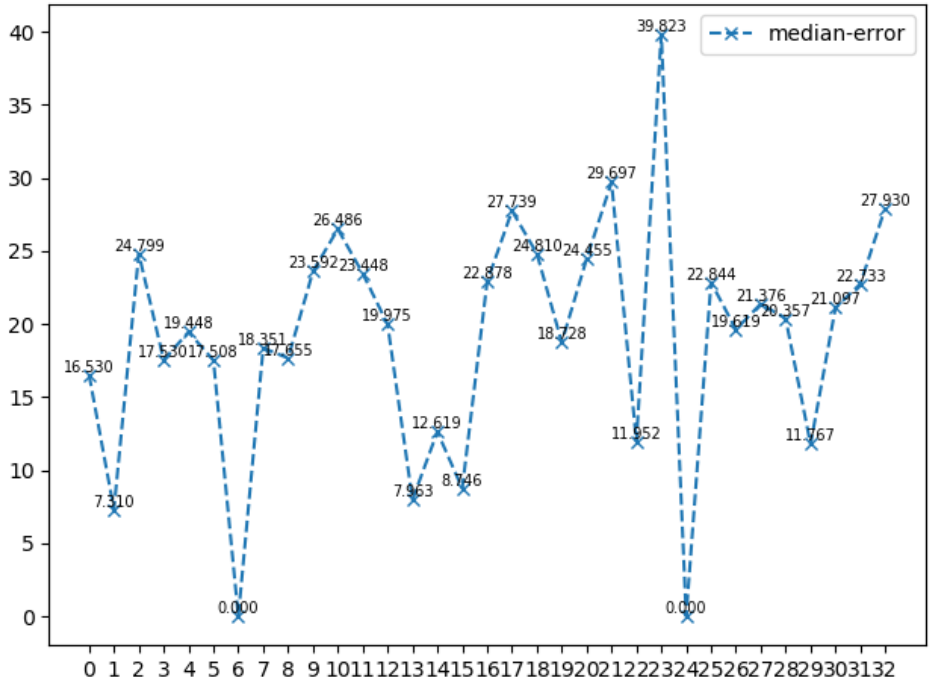

Figure 1.1: 不同的 MS 主基站中位误差分布

我将所有的点重新合并到一起，做出了总体的 CDF 曲线，以及与 a 问的 RandomForestClassifier 结果比较如下：

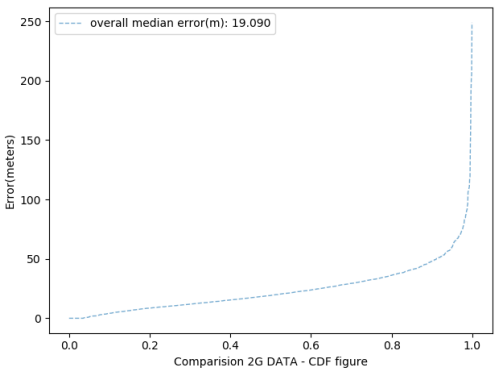

Figure 1.2: c 问 CDF 图

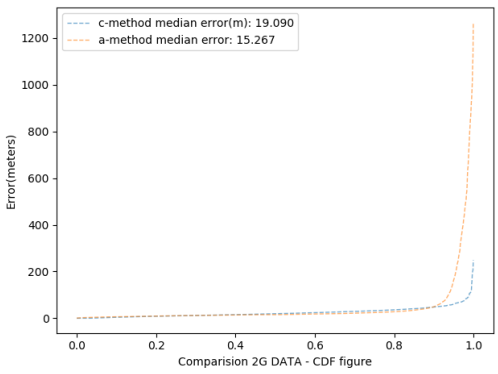

Figure 1.3: a 问与 c 问比较结果

2 分析讨论

这里主要对不同基站的预测结果，以及 c 问与 a 问的差异这二个部分进行分析讨论。

2.1 不同基站的预测结果

首先我想在这里分析的是，查看预测结果可以发现，有两个 MS 基站的中位误差为 0.000，也就是说其预测完全正确。针对这样的预测结果，我首先考虑到的是，这两个 MS 基站对应的数据集较少，于是我尝试将每个基站的中位误差和总的数据集个数输出，其结果如下（为了节省篇幅，这里只列出部分结果）：

数据集个数	中位误差
2	0.000
3	11.952
8	7.310
9	17.655
18	21.096
.....	
62	17.508
63	21.376
70	0.000
.....	
267	8.746
301	18.727
323	11.766
.....	

可以发现，中位误差的结果，和数据集个数的多少并不存在明显的联系。于是我进行了进一步的统计，针对每一个 MS 基站，我将其所有的点在地图上进行了标注，得到了如下的结果：

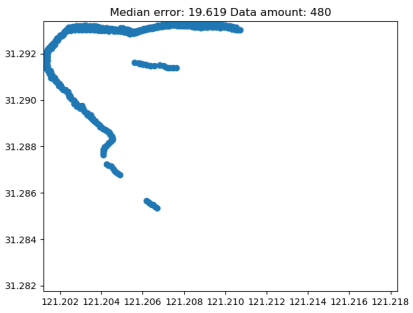

Figure 2.1: 中位误差 19.619米

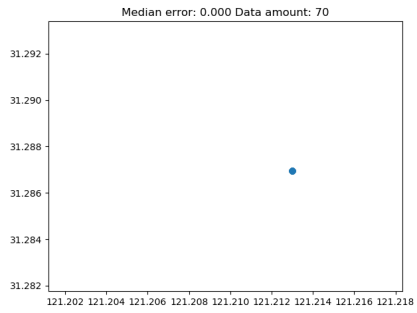

Figure 2.2: 中位误差 0.000米

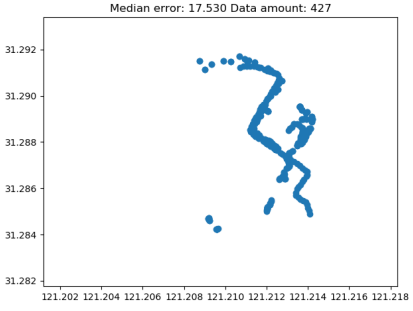

Figure 2.3: 中位误差 17.530米

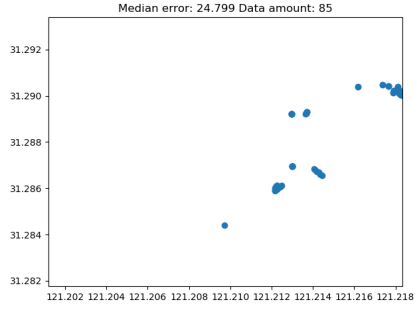

Figure 2.4: 中位误差 24.799米

这里我选择了四个 MS 基站的结果进行详细分析。可以发现，针对 Figure 2.2 的结果，虽然其一共有 70 条数据，但是都集中在十分相近的位置，所以其数据集相当于是将一条数据复制了很多次后得到的，其最终的预测结果当然会更好。

而对于 Figure 2.1 和 Figure 2.3 的 MS 基站，其数据集分布比较均匀，从直观的角度来看，是连续的行走路径上的一些离散的点，而我们的数据集也是人行走路径上的一些离散的点构成的，那么我可以认为，这两者的数据集是相似的，那么最终预测的中位误差应该相近。而其最终的预测结果比较接近总的平均预测结果，是符合我的预期的。

而 Figure 2.4 中的 MS 基站，其数据集是分布比较分散的一些离散的点，而最终的中位误差相较于总的中位误差要更大。

总结上面的情况，我可以认为对于 c 问中的所有 MS 基站的预测结果，主要受到了 MS 基站包含的**数据集数量**和**数据集分布情况**这两个因素。如果一个 MS 基站的数据集过少，其预测结果会比较优秀，但这不是我们所期待的，因为这样很大程度上会存在过拟合的情况。而如果数据集分布过于分散，这样的基站的预测结果会存在较大误差。最理想的情况应该是，拥有一定数量的数据集，且这些数据集分布比较均匀，如 Figure 2.1 和 Figure 2.3 的情况。

而最糟糕的情况则是，数据集数量过少，且分布比较分散的话，则会导致效果最差。

2.2 c 问与 a 问的差异

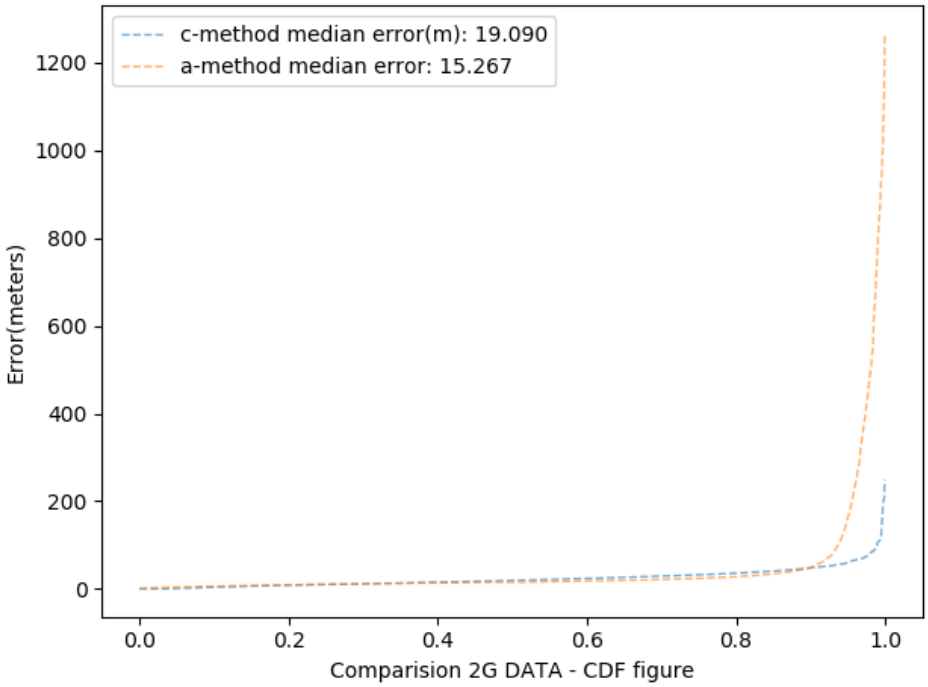

Figure 2.5: c 问与 a 问的 CDF 曲线

这里我做出了 c 问与 a 问的 CDF 曲线。这张图有两个比较明显的特点：一是 a 问的中位误差15.267m要低于 c 问的中位误差19.090m；二是这两条 CDF 曲线，其在前半部分 a 问对应的曲线的 y 值都要低于 c 问对应的曲线的值，但都较小，而在最后 0.9 - 1.0 的部分，a 问的 CDF 曲线出现了一个很明显的陡峭上升。这两个特点反应了 c 问与 a 问的差异性。

接下来我针对这两个特点的产生进行分析。

2.2.1 分类还是回归?

很显然的，a 问是一个多分类问题，而 c 问是一个回归问题。我们可以认为回归问题是用来解决预

测一个值的情况，而分类问题则是考虑将事物打上一个标签。因此，也有一种说法是回归得到的结果是连续的，而分类得到的结果是离散的。

对于 a 问来说，相当于是先将整个区域划分成 $20m * 20m$ 的很多栅格，根据某条记录的基站信息，来预测该记录是否在某一个栅格中。这种方法带来的一个显著的误差，是在计算中位误差时，某一条记录对应的位置是该栅格的中心点，这样的一个转换，必然会带来一定程度上的误差。

而 c 问则是，先将数据按照基站分类，我认为可以理解为，将整个区域按照基站划分成很多小的区域，再根据某条记录的基站信息，来预测该记录在其对应的 MS 主基站区域中的某个位置。这种方法的误差，与 a 问相比，不会存在转换带来的误差。

在这个情况上，c 问的方法更胜一筹。

2.2.2 数据集的多少?

在这个角度考虑，a 问的数据集是全体的，而 c 问的数据集因为需要按照 MS 基站分组，数据集显然没有 a 问多。这样来看，a 问更占优势。

2.2.3 极端误差

a 问还有一个问题在于，可以认为其预测可能出现的标签是在整体的范围内，那么可能出现的极端误差可能会很大。而 c 问相对来说限制了其预测结果的范围在了对应的 MS 基站的范围内，极端误差不会过大。

反应在 CDF 曲线上，也可以发现 c 问最后的极端误差要明显地小于 a 问，a 问最大的误差可能达到了 $1200m$，而 c 问则最多只有 $200m$ 左右。

总的来看，虽然 a 问的中位误差要较小于 c 问，然而当 $0.9 - 1.0$ 的时候，a 问的预测结果误差太大。因此 c 问的方法实际上要略优于 a 问。

3 性能比较

两种方法的耗时如下，c 问稍微优于 a 问。多个分类器而每个分类器数据集较少（ c 问），相较一个分类器而数据集较多（ a 问），取得了稍微的优势。

方法	时间
a 问	10.34s
c 问	9.42s