数据分析与数据挖掘第三次作业第一大题 – e问分析文档

数据分析与数据挖掘

DAM COURSE, SPRING 2018

BY

1552674 李　源

同濟大學
TONGJI UNIVERSITY

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。对于 top k- 各自的 MS 主基站的具体结果，在修正前后的情况如下：

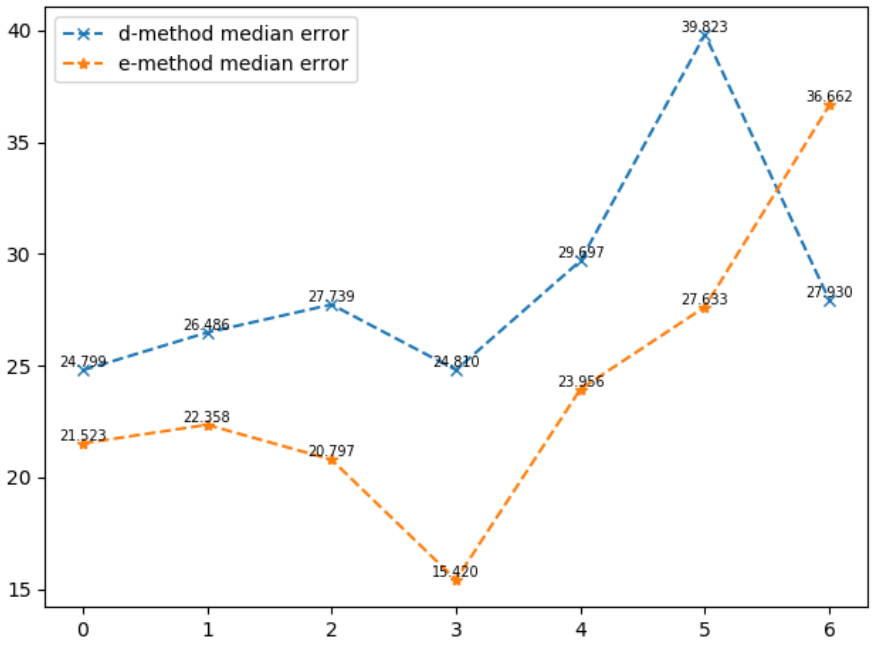

Figure 1.1: 不同的 MS 主基站修正前后中位误差分布

我将所有的 top k- 的 MS 基站的点重新合并到一起，做出了总体的 CDF 曲线，以及与未修正的 CDF 曲线比较如下：

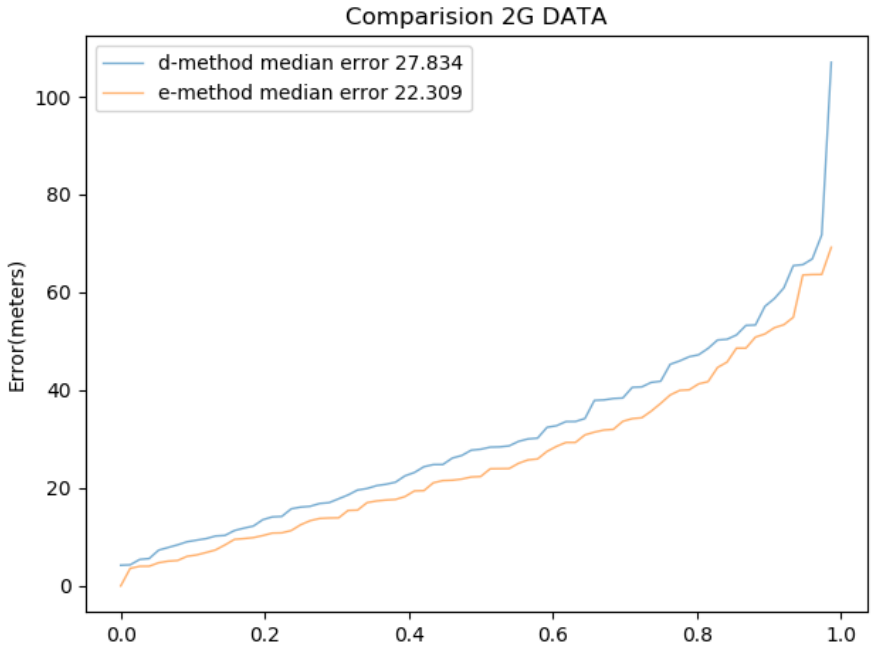

Figure 1.2: 修正与未修正的 CDF 曲线

2 分析讨论

这里主要对 e 问的修正方法和 d 问和 e 问的差异进行分析。

2.1 e 问的修正方法

e 问的修正办法与 d 问的不同点在于，d 问是简单地考虑了预测结果的好坏，然后将 topk+ 分组对应的 MS 基站的数据融入到了 topk- 的分组当中。而 e 问进一步考虑了数据之间的相似性，即在融合之前，首先要考虑将基站进行聚类，然后在同一组当中，再进行 topk+ 融入 topk- 的操作。

我在这里的做法是，首先运用 KMeans 聚类方法，通过距离来将 MS 基站进行聚合。在完成聚合之后，我再考虑将数据融入到 topk- 的分组当中。同时，我在融入的时候，进行了随机抽样，尝试看这种方法能不能对更好地提高预测的准确性。

KMeans 的聚类结果如下：

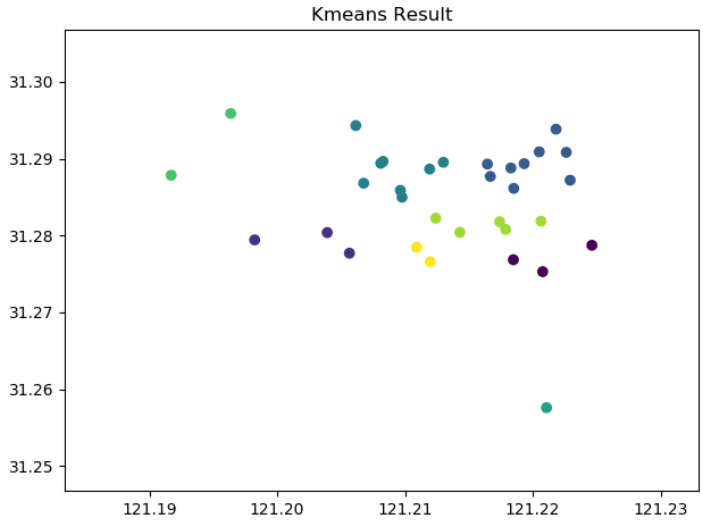

Figure 2.1: KMeans 聚类结果

2.2 d 问和 e 问的差异

这里我首先将 topk- 分组中的 MS 基站的原始数据分布和融入了的 topk+ 分组中的数据分布画了出来，进行分析，这里我将 d 问和 e 问的结果画在了一起，进行比较。

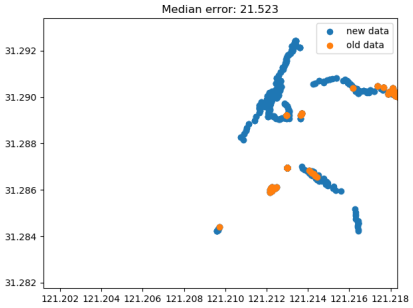

Figure 2.2: 基站1 e 问结果

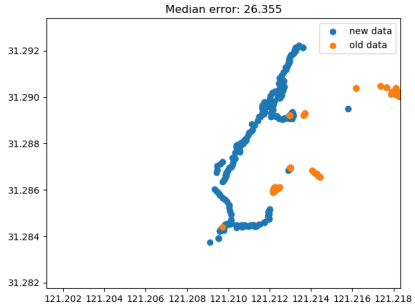

Figure 2.3: 基站1 d 问结果

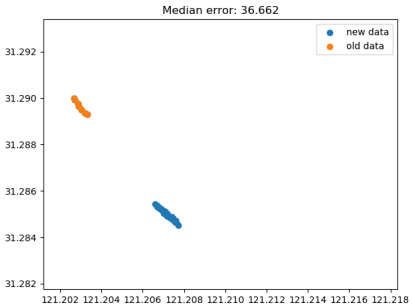

Figure 2.4: 基站2 e 问结果

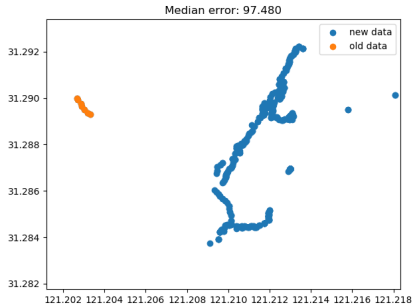

Figure 2.5: 基站2 d 问结果

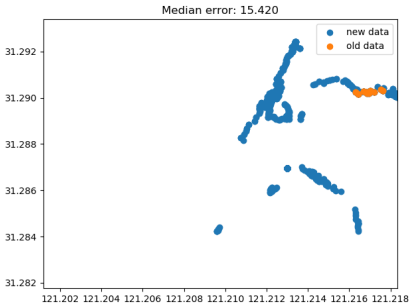

Figure 2.6: 基站3 e 问结果

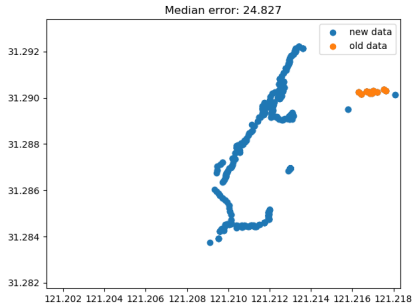

Figure 2.7: 基站3 d 问结果

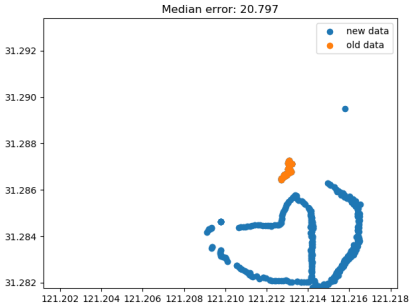

Figure 2.8: 基站4 e 问结果

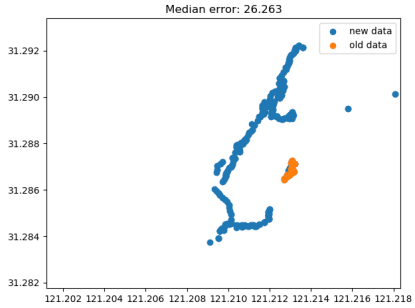

Figure 2.9: 基站4 d 问结果

这里很明显的可以发现，e 问的加入的新数据集与原始的相似程度要明显的高于 d 问的结果，比如对于基站 1，原始的数据是一些离散的点，而在经过 e 问的修正之后，其数据集在一定程度上变成了连续的点，并且相较于 d 问来说，其相似性更高。再比如基站 3 也是类似的情况。

而对于基站 2 的情况来说，e 问的方法能够在一定程度上减少太多的距离较远的数据，这些数据对于训练可能会出现负影响。

总的来说，e 问相较于 d 问的优点是，e 问在补充数据集的同时，考虑到了 topk+ 和 topk- 数据集之间的相似性，而不是简单的将 topk+ 的数据直接融入进去。在 c 问的时候，我们就可以发现，训练效果的优异是与数据集的分布相关的。 e 问的方法可以说是在一定程度上解决了这个问题。

2.3 一些其他的尝试

在融入数据的时候，我还尝试了将融入数据进行随机抽样，以及判断融入数据的 MS 主基站的位置与原始数据中 MS 主基站的位置之间距离，并将距离较远的数据删除掉。其结果如下：

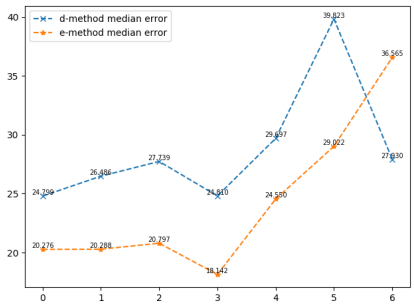

Figure 2.10: 删除距离较远数据

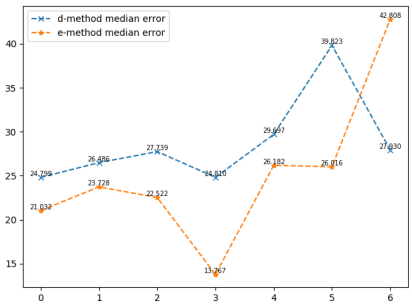

Figure 2.11: 随机抽样

随机抽样的结果可以说是有好有坏，毕竟随机删除一些数据也没有什么理论支持，这只是我的一次尝试。而删除距离较远数据之后，与不删除相比，也没有取得更好的结果。这两种方法的尝试，我主要是想进一步的删除一些仍然距离较远的数据。

3 性能比较

两种方法的耗时如下，e 问因为需要再次预测，所以时间会相较于 c 问更久，不过 e 问取得较大的改进，因此时间的增加并取得好的效果。

方法	时间
e 问	13.37s
c 问	9.42s