**2020.12.16-2020.12.30 周报**

**这周的总结：**

这两周主要准备和进行了几门考试和结课报告：**运筹学**、**通信网规划与优化**与**统计信号处理**，在科研方面进行了通信网与数据挖掘之间关系的学习，主要学习了一篇关于实验过程的文章。

其中主要讲述了具体的实验过程第一部分是进行数据的处理。按照数据挖掘的流程，在获取到基本的数据后，系统需要对其进行初步的处理，确定引入的几个特征变量、去除掉重复以及不相关的特征。基于网络规划知识数据库，对实测数据进行初步清洗与分析，去除具有大量重复和缺省值的属性，并且通过分析提取影响信号覆盖质量的相关特征，将其输入到学习模型中进行训练，从而获得最终的基站部署。

第二部分是模型的选取。学习模型分为预测模型和规划模型，基站站址的选定是根据每一次基站部署好后网络的空间覆盖情况进行调整，由于每一次基站调整，会使整个网络的信号覆盖情况发生变化，因此，需要针对网络信号覆盖的变化，利用相应的数据学习得到预测模型，对每一次的网络拓扑下的覆盖情况进行预测。而规划模型则是根据预测所得的覆盖情况，确定合适的基站部署。整个规划的总目标是通过每一次基站调整逐渐缩小信号覆盖差的区域，并且使区域中信号覆盖质量接近所要求的标准。

另外介绍了两种算法，一种是提升回归树（BRT，boosting regression tree） 算法属于集成学习方法中的一种，通过集成多个基学习器共同完成学习任务。相比于单一的回归算法，如线性回归、对数几率回归算法等，BRT 算法以组合多个决策树的方式，能够获取更加优越的泛化能力，从而提升了模型的预测精度。

另一种是数据挖掘中优化过的K聚类算法。典型的 K 均值聚类算法是将数据集 X={x1,...,xn}中的 n 点进行划分，把原来独立的n 个点通过设定距离相似度划分进 K 个簇当中，簇集合 C={c1,...,ck}。一般会以两点之间的欧式距离作为相似性度量，把数据点划分进距离较近的簇中心所在的簇中。算法一般是以最小化簇内位置误差平方和（SSE, sum of the squared error）为目标函数。

在经典的 K-means 算法中，每个数据点对定位簇中心的位置有着同样的重要性。然而，优化后的K算法把基站位置的选择当作基于覆盖分布空间模式的加权问题进行处理，即认为空间中的每一个点不再对簇中心有等价的影响，引入权重来衡量数据点对基站位置的影响度，从而提出加权 K-centroids 算法。分配好数据点以后，本算法需要对簇中心的位置即站址的位置进行迭代调节，主要从距离影响度和覆盖权重两方面考虑。对于距离影响度来说，由于基站要尽可能覆盖远处的终端，对于离基站近的终端自然信号会相对好一些，所以它离基站的距离对于覆盖权重来说，优化基站位置是希望在基站覆盖内终端的接收信号能尽可能好，所以需要重点关注覆盖质量差的终端，赋予它对基站位置调整更大的影响权重。当迭代到基站的位置不再发生变化时候，就说明找到了最优的基站选址，此时结束迭代，就可以得到最佳的簇中心位置，即基站站址。

**下周的计划：**

结课报告完成后，继续对于箔条的理论学习，争取寒假阶段能够厘清自己对于这个领域的认识。