周报（10/09/16）

本周进度：

1、论文总结：在《Box drawings for learning with imbalanced data》中作者提出了两个解决不平衡数据的方法。首先提出的是精确方盒算法(Exact Boxes)，利用平行线对不平衡数据中的少数部分的每一维进行划分。由于参数中夹杂着若干整数，解决的问题便由线性规划升级为混合整数线性规划(MILP)问题。利用了Gurobi Optimization 的软件对问题进行了求解。然后作者基于第一种最求精度忽略计算成本的方法进行改进继而提出快速方盒算法(Fast Boxes)。该算法大体上分为三步(1)：少数样本点的聚类。(2)：确定对边界影响的样本点。(3)：利用推导的公式对边界扩充，这样的算法更为直观与易解释。

2、改进：我想是否可以利用椭圆来包围我们关心的样本。改进的具体过程如下：本文的作者是进行了每一维的线性化分，由线到面然后再到体的循序渐进从而导出方程。结合二维的椭圆方程与三维的椭球面方程大胆演算出更为有效的算法。主体公式如下：

$$\sum_{j=1}^{m} \frac{\left(\boldsymbol{X}_{j,k} - \boldsymbol{O}_{j,k}\right)^2}{\left(\boldsymbol{R}_{j,k}\right)^2} = 1$$

$\boldsymbol{X}_{j,k}$代表着X在j维属于第k个聚类的样本的数值，$\boldsymbol{O}_{j,k}$表示中心点，$\boldsymbol{R}_{j,k}$表示j维第k 个聚类的伪半径。剩余的思想较为相似。聚类，划分几个不同的区间来对中心点以及伪半径的调整。不同的是在划分区间的时候，我们有了三种不同的情况。最后的分类器的作用方式是将数据代入等式的左边然后与1进行比较。

3、难点：个人认为新颖的地方是将问题直接用非线性的方式进行逼近求解。现在唯一的理论难度是选择对参数的总体调整还是分维度调整。如果类似本论文的分维度进行参数调整，优点在于简单易懂易操作，缺点是对于我想表达的一些关键点对椭球面的拉伸效果不容易去理论证明。

下周计划：

1、完善与论证提出的想法。利用代码数据进行实践验证。

2、调整学习态度，保证每天的有效学习时间。

本周结果：

Table 1: Algorithm Table

Algorithm	Positive True	Negative False
$fastboxes$	49/50	93/4950
$fasteclipse$	48/50	56/4950

由于时间的关系以及新算法的完善，本周的工作仅仅实现了在本周进度中提到的简单想法。也就是在调整中心点与伪半径的时候，依旧采用的是对每一维度的分别调试。代码效果还不错，有进一步提升的空间。

PS: $fasteclipse$是新的算法。