第九章作业

陈镜融
14307130118@fudan.edu.cn

May 7, 2018

1 9.4

9.4 Compare the advantages and disadvantages of eager classification (e.g., decision tree, Bayesian, neural network) versus lazy classification (e.g., k -nearest neighbor, case-based reasoning).

eager learning训练速度慢，但在分类上比lazy learning算法更快。 lazy learning用更少的时间训练，但更多的时间来预测。 随着数据集和模型的增大，eager learning训练的难度和时间大大增加，因此算法的选择需要在数据量和训练时间上做权衡。[1]

2 9.5

9.5 Write an algorithm for k-nearest-neighbor classification given k, the nearest number of neighbors, and n, the number of attributes describing each tuple.

假设数据已经经过了数值化和归一化处理。距离采用欧几里得距离。要实现K近邻，可以采用KD-Tree的方法[1]。

主要分两个过程，1. 建树 2. 查询k近邻

建树的过程，就是每次在一个维度d上，在当前点集合中查找这个维度的中位数点，将该点拎作树跟，将点集按维度d分为小于中位数点和大于中位数点的两个集合。接着选择另一维度，递归往下处理即可。 建树的过程是每次取中位数的过程，利用类似快排的快速选择算法，或者C++ STL中的std::nth_element函数，可以做到O(n)选取中位数，因此整个建树的复杂度是O(nlogn)的。

另外，为了做到k近邻查询，需要处理出树上每个节点对应的区域边界，仔细思考发现该点对应的边界，就是在父亲节点维度上的最小值和最大值，再加上父亲节点坐标，三个点划分出来的两个区间。因此在建树时可以顺便预处理。

有了每个节点对应的边界，可以考虑对于给定点，如何查询最邻近点。 做法是维护一个当前最小值，从树根进行深度优先遍历，直接利用查询点到当前树上节点所代表的区域，在该维度上作一条垂直线，一定有这条垂线的长度小于等于区间内（也就是子树内）

[1] https://gopalcdas.com/2017/05/24/construction-of-k-d-tree-and-using-it-for-nearest-neighbour-search/

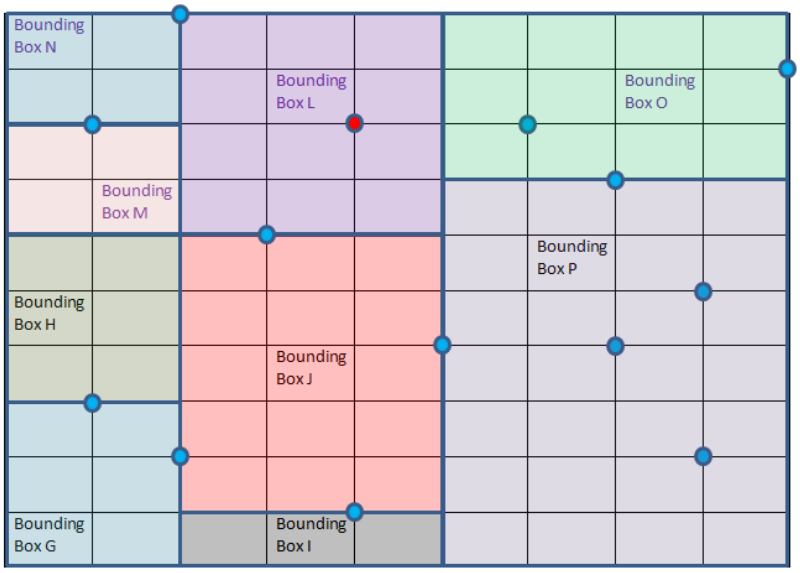

Figure 1:

任何点到目标点的距离。因此该距离若大于当前最小值，这个子树就可以不用遍历，因此降低了查询时间。

该方法的时间复杂度低于暴力查找的复杂度。但我不太会算

对于k近邻，用一样的思路，把当前维护的最小值替换成用一个k个点的集合，每次更新这k个点的集合即可。用优先队列可以高效维护。

算法实现的代码在`https://github.com/crazyboycjr/knn`可以找到，kdtree相关的代码在kdtree.h里。

Reference

[1] Fentie, S. G., and Alemu, A. D. A comparative study on performance evalution of eager versus lazy learning methods.