数据结构与算法I 思考题7

2019201409 于倬浩

问题

V-Optimal Histogram

给出n个点对(i,f_i)，要求将这n个点对分成b组，第i组的起点、终点分别为s_i,e_i，且组内所有值的平均值是h_i，满足$\sum_{i=1}^{b}\sum_{j=s_i}^{e_i}(h_i-f_j)^2$最小。

解法

考虑使用动态规划算法。首先，我们考虑使用当前分组数量划分阶段。

设$hist(p,t)$表示将前p个数字，分成t组的最优解，即满足$\sum_{i=1}^{t}\sum_{j=s_i}^{e_i}(h_i-f_j)^2$最小。假设第t组的起点为$s_t$，终点为$p$，那么$hist(p,t)$的子问题为$hist(s_t-1,t-1)$，即满足$\sum_{i=1}^{t-1}\sum_{j=s_i}^{e_i}(h_i-f_j)^2$的解。假设这个子问题存在更优的解，那么直接将解换成更优的，然后加上`i=t`时的一项和式，即可使得$hist(p,t)$得到改善。因此，按照这种方法划分子问题具有最优子结构。

按照上述方法，定义状态：设`f[i][j]`表示前`i`个数字，分成`j`组的最小方差和。那么，有如下状态转移方程：

$f[i][j]=min_{k=j-1}^{i-1}(f[k][j-1]+W(k+1,i))$。

其中$W(L,R)=\sum_{j=L}^{R}(\frac{\sum_{i=L}^{R}f_i}{R-L+1}-f_j)^2=\sum_{i=L}^{R}f_i^2-(R-L+1)(\sum_{i=L}^{R}f_i)^2$。

因此，我们需要$\Theta(n)$的时空复杂度预处理$sum[i] = \sum_{k=1}^{i} f_k$，$sum2[i] = \sum_{k=1}^{i} f_k^2$，即可使用$\Theta(1)$的时间复杂度，算出$W(L, R) = sum2[R] - sum2[L-1] - (R - L + 1)(sum[R] - sum[L-1])^2$。

因此，单次转移的代价即为$\Theta(n)$，状态数为$\Theta(nb)$，总时间复杂度即为$\Theta(n^2b)$，空间复杂度为$\Theta(nb)$。另外，实际上当b>n时，由于序列最多分成n段，因此多余的状态无需考虑，即时间复杂度为$\Theta(min(n^2b, n^3))$。

如果需要输出方案，只需记录`h[i][j]=k`表示转移到`f[i][j]`的状态是`f[k][j - 1]`，倒推一遍即可算出决策点。

核心代码如下：

```
for(int i = 1; i <= n; ++i) {
    sum[i] = sum[i - 1] + f[i];
    sum2[i] = sum2[i - 1] + f[i] * f[i];
}

for(int i = 1; i <= n; ++i) {
    for(int j = 1; j <= min(i, m); ++) {
        f[i][j] = inf;
        for(int k = j - 1; k <= i - 1; ++k) {
            f[i][j] = min(f[i][j], f[k][j - 1] + sum2[i] - sum2[k]
                           - (i - k + 1) * (sum[i] - sum[k]) *
(sum[i] - sum[k]));
        }
    }
}
```