山东大学

SHANDONG UNIVERSITY

山东大学机器学习课程

实验报告

——实验七：ID-3 决策树的设计与实现

姓名：刘梦源

学院：计算机科学与技术学院

班级：计算机 14.4

学号：201400301007

一、实验目的：

（1）学习 ID3 分类器，并用 ID3 模型分类数据。
（2）学习用 matlab 对数据结构程序的编写

二、实验环境：

（1）硬件环境：
英特尔® 酷睿™ i7-7500U 处理器
512 GB PCIe® NVMe™ M.2 SSD
8 GB LPDDR3-1866 SDRAM
（2）软件环境：
Windows10 家庭版 64 位操作系统
Matlab R2016a

三、实验内容

算法如下：（参考西瓜书）

分析与设计：

决策树学习的基本算法：

输入：训练集D = $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
属性集A = $\{a_1, a_2, \ldots, .a_d\}$;

过程：函数 TreeGenerate(D,A)

1. 生成结点 node；
2. if D 中样本全属于同一类别 C then
3. 将 node 标记为 C 类叶节点；return
4. end if
5. if A=∅ OR D 中样本在 A 上取值相同 then
6. 将 node 标记为叶节点，其类别标记为 D 中样本最多的类；return
7. end if
8. 从 A 中选择最优划分属性a_*；
9. for a_* 的每一个值a_*^v do
10. 为 node 生成一个分支；令D_v表示 D 中在a_*上取值为a_*^v的样本子集；
11. if D_v为空 then
12. 将分支结点标记为叶节点，其类别标记为 D 中样本最多的类；return
13. else
14. 以 TreeGenerate(D_v,A\{a_*})为分支结点
15. end if
16. end for

输出：以 node 为根结点的一棵决策树

主要公式：（1）、（2）摘录自西瓜书 P75，（3）摘自许老师 PPT

$$\mathrm{Ent(D)} = -\sum_{k=1}^{|\gamma|} p_k \log_2 p_k \tag{1}$$

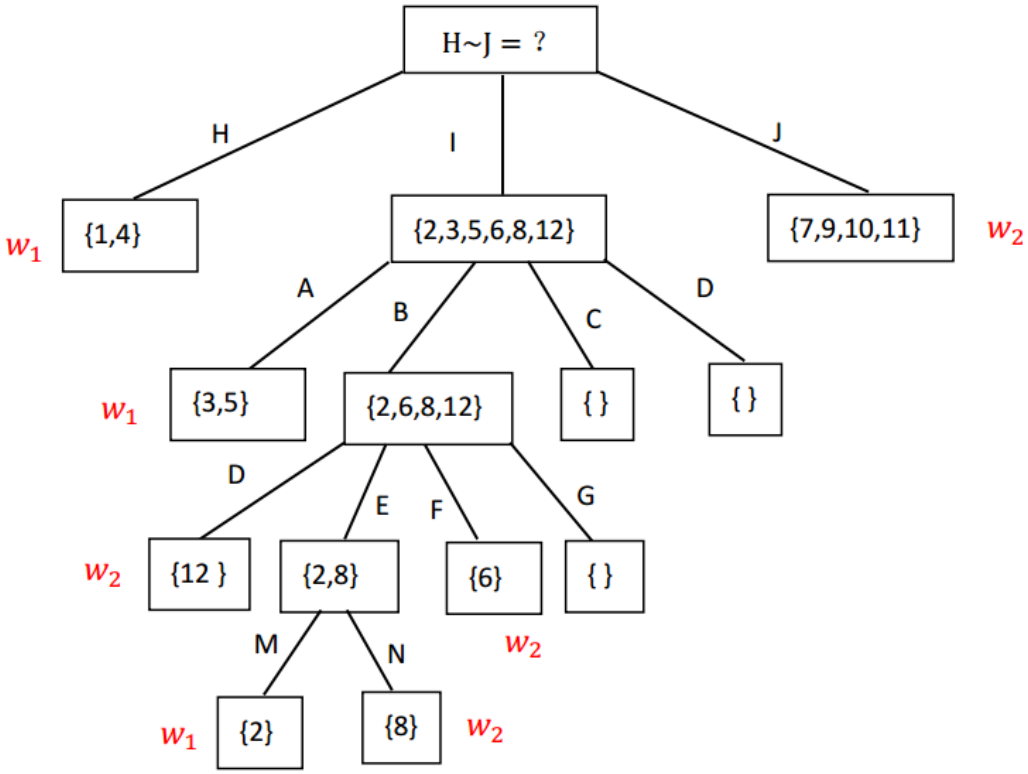
$$\mathrm{Gain(D,a)} = \mathrm{Ent(D)} - \sum_{v=1}^{V} \frac{|D^v|}{|D|}\mathrm{Ent}(D^v) \tag{2}$$

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$$

$$SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|} \tag{3}$$

四、实验结果

（a）树的训练



注：错误数据处理方式为：就认为那是一个新特征 D

（b）对数据进行的分类

第一个数据{B,G,I,K,N}无法用这个决策树分类，因为 G 对应的分支为空集无法判定类别。

第二个数据{C,D,J,L,M},可以通过这个决策树判定它属于 w2 类。

(c)

（b）的逻辑表达式：

$$\{B, G, I, K, N\} = (\text{'H}\sim\text{J'} = I)\text{AND}(\text{'A}\sim\text{D'} = B)\text{AND}(\text{'E}\sim\text{G'} = G)$$
$$\{C, D, J, L, M\} = (\text{'H}\sim\text{J'} = J)$$

（d）w1 w2 的逻辑表达式

$$w_1 = H\ or\ (I\ and\ A)\ or\ (I\ and\ B\ and\ E\ and\ M)$$
$$w_2 = J\ or\ (I\ and\ B\ and\ D)\ or\ (I\ and\ B\ and\ F) or (I\ and\ B\ and\ E\ and\ N)$$

这个实验比较简单，但两点会导致大家的结果出现差异，第一，ID3 是否改进，是按照信息增益进行计算还是信息增益率进行计算；第二，错误数据点的处理方式有差异。但我在我的假设条件下，取得了较为理想的实验结果。