BÁO CÁO ĐỒ ÁN 2

BÀI TẬP 2: THUẬT TOÁN ID3

Tên: Đỗ Thành Nhơn – 1512387
Nguyễn Thành Tân– 1512491
Lớp: 15CNTN
GVHD: Nguyễn Ngọc Thảo – Lê Ngọc Thành

PHỤ LỤC

I. Xây dựng chương trình.

1. Đọc và biểu diễn dữ liệu.

- Các thuộc tính của từng mẫu sẽ được lưu trữ trong mảng 2 chiều có kích thước MxN với M là số dòng dữ liệu, N là số thuộc tính.
- Tên của từng loại thuộc tính sẽ được lưu trữ trong một mảng 1 chiều với kích thước là số thuộc tính.
- Dữ liệu được đọc từ file sẽ được lưu trữ trong 2 cấu trúc dữ liệu trên.

2. Các hàm quan trọng.

2.1. Hàm tính Entropy và Infomation Gian.

- Công thức tính:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$E(T, X) = \sum_{c \epsilon X} P(c)E(c)$$

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

2.2.Hàm lấy ra thuộc tính có Infomation Gian lớn nhất.

- Ta duyệt qua hết từng loại thuộc tính để tính thông số Infomation Gian.
- Chọn thuộc tính có Infomation Gian lớn nhất để phân chia cây quyết định.
- Ứng với mỗi thuộc tính C và lớp con C_i, trích ra những mẫu chứa C_i.

2.3.Hàm xây dựng cây quyết định.

- Cấu trúc dữ liệu của mỗi Node bao gồm:
  - Name: chứa tên của thuộc tính đang xét.
  - Label: chứa nhãn của thuộc tính đang xét.
  - Vector<Node*>: chứa danh sách các node con của node hiện tại.
- Hàm xây dựng cây được cài đặt theo giải thuật đệ qui dựa theo hướng dẫn trong giáo trình Cơ sở trí tuệ nhân tạo.

2.4.Hàm thực hiện truy vấn.

- Ứng với mỗi dòng dữ liệu test chưa có nhãn, ta sẽ sử dụng những thuộc tính và dựa vào cây quyết định để tìm nhãn.
- Được cài đặt theo phương pháp duyệt cây theo chiều sâu DFS.
- Đến khi nào tới Node lá thì từng và trả về nhãn tại Node lá.
- Nếu không tìm được thì trả về chuỗi rỗng.

2.5.Hàm in cây quyết định.

- Hàm in cây quyết định được cài đặt bằng giải thuật duyệt cây theo chiều sâu. Ứng với mỗi Node sẽ in ra thông tin tên thuộc tính và nhãn của thuộc tính.

II. Demo chương trình.

- Phần Demo sử dụng tập dữ liệu Zoo.
- Do tập dữ liệu này có thuộc tính là Animal name, do đó khi dùng thuộc tính này để phân loại thì thuộc tính sẽ được dùng đầu tiên, do đó cây chỉ phân loại 1 lần và không có mang tính tổng quát.
- Do đó khi chạy tập dữ liệu Zoo cần loại bỏ thuộc tính này để cây quyết định có tính tổng quát cao hơn.

```
Dos-MacBook-Pro:Source thanhnhon1997$ ./main
Attribute name
0. animal
1. hair
2. feathers
3. eggs
4. milk
5. airborne
6. aquatic
7. predator
8. toothed
9. backbone
10. breathes
11. venomous
12. fins
13. legs
14. tail
15. domestic
16. catsize
17. type
Nhap vao thuoc tinh ban muon loai bo, nhap -1 neu ban khong muon loai bo thuoc tinh nao: 0
```

```
+legs = 0
|       +fins = false
|       |       +toothed = false
|       |       |       +type = invertebrate
|       |       +toothed = true
|       |       |       +type = reptile
|       +fins = true
|       |       +eggs = false
|       |       |       +type = mammal
|       |       +eggs = true
|       |       |       +type = fish
+legs = 2
|       +hair = false
|       |       +type = bird
|       +hair = true
|       |       +type = mammal
+legs = 4
|       +hair = false
|       |       +aquatic = false
|       |       |       +type = reptile
|       |       +aquatic = true
|       |       |       +toothed = false
|       |       |       |       +type = invertebrate
|       |       |       +toothed = true
|       |       |       |       +type = amphibian
|       +hair = true
|       |       +type = mammal
+legs = 5
|       +type = invertebrate
+legs = 6
|       +aquatic = false
|       |       +type = insect
|       +aquatic = true
|       |       +type = invertebrate
+legs = 8
|       +type = invertebrate
```

III. Đánh giá độ chính xác.

1. Phân chia dữ liệu.

- Tập dữ liệu Zoo sẽ được chia làm 2 phần:
  - Tập dữ liệu huấn luyện: bao gồm 90 mẫu.
  - Tập dữ liệu kiểm tra: bao gồm 10 mẫu đã được bỏ nhãn.

2. Đánh giá.

- Kết quả cho thấy độ chính xác ứng với bộ dữ liệu huấn luyện và kiểm tra đạt là 6/10, tương đương 60%.

```
Input: slug,false,false,true,false,false,false,false,false,false,true,false,false,0,false,false,false,?
Output invertebrate
Result invertebrate

Input: sole,false,false,true,false,false,true,false,true,true,false,false,true,0,true,false,false,?
Output invertebrate
Result fish

Input: sparrow,false,true,true,false,true,false,false,false,true,true,false,false,2,true,false,false,?
Output bird
Result bird

Input: squirrel,true,false,false,true,false,false,false,true,true,true,false,false,2,true,false,false,?
Output bird
Result mammal

Input: stingray,false,false,true,false,false,true,true,true,true,false,true,true,0,true,false,true,?
Output invertebrate
Result fish

Input: swan,false,true,true,false,true,true,false,false,true,true,false,false,2,true,false,true,?
Output bird
Result bird

Input: termite,false,false,true,false,false,false,false,false,false,true,false,false,6,false,false,false,?
Output insect
Result insect

Input: toad,false,false,true,false,false,true,false,true,true,true,false,false,4,false,false,false,?
Output reptile
Result amphibian

Input: tortoise,false,false,true,false,false,false,false,false,true,true,false,false,4,true,false,true,?
Output reptile
Result reptile

Input: wren,false,true,true,false,true,false,false,false,true,true,false,false,2,true,false,false,?
Output bird
Result bird

Da test 10 mau, ket qua chinh xac 6 mau.
Do chinh xac: 60 %
```