Învățare Automată - Laboratorul 3
Arbori de decizie. Algoritmul ID3

Tudor Berariu
Laboratorul AIMAS
Facultatea de Automatică și Calculatoare

4 martie 2013

1 Introducere

Scopul acestui laborator îl reprezintă înțelegerea conceptului de *arbore de decizie* și implementarea unui algoritm de clasificare supervizată care construiește astfel de structuri: algoritmul ID3.

Consultați cursul pentru pseudocod și explicații detaliate. Paragrafele de mai jos reprezintă un sumar concentrat al noțiunilor necesare laboratorului de astăzi.

2 Arbori de decizie

Problema pe care o abordăm este una de clasificare supervizată. Fiind dat un set de exemple aparținând a două clase, pentru fiecare dintre aceste exemple fiind cunoscute valorile pentru un set de atribute discrete și clasa din care face parte, să se construiască un clasificator care să distingă între obiectele celor două clase.

Un arbore de decizie este un astfel de clasificator și are următoarele proprietăți:

- fiecare nod care nu este frunză reprezintă un test pentru un atribut, având câte un arc (și un subarbore) corespunzător fiecărei valori posibile;

- fiecare frunză este etichetată cu o clasă.

Pentru a clasifica obiecte noi se pornește de la nodul rădăcină și se *coboară* din fiecare nod care nu este frunză în subarborele corspunzător valorii pe care o are obiectul pentru atributul corespunzător nodului.

3 Algoritmul ID3

Algoritmul ID3 primește un set de date (obiecte valori pentru un număr de atribute și etichetate cu numele clasei din care fac parte) și construiește un arbore de decizie pe baza acestuia.

Algoritmul funcționează astfel:

1. dacă toate exemplele aparțin aceleiași clase, atunci algoritmul întoarce un singur nod etichetat cu acea clasă
2. dacă nu mai există atribute, se construiește un nod etichetat cu cea mai frecventă clasă
3. altfel:
   (a) se alege atributul A_{best} care aduce cel mai mare câștig informațional și se construiește un nod corespunzător acestuia;
   (b) pentru fiecare valoare posibilă v_i a atributului A_{best} se construiește o submulțime S_i a setului de date ce conține acele exemple care au valoarea v_i pentru atribului A_{best}
   (c) pentru fiecare valoare posibilă v_i a atributului A_{best} se construiește un subarbore aplicând recursiv algoritmul ID3 pentru mulțimea S_i.

Câștigul informațional

Entropia Pentru a alege atributul care aduce cel mai mare câștig informațional se folosește conceptul de *entropie* din teoria informației. Pentru un set de date S și mulțimea de clase C, calculăm entropia astfel:

$$H(S) = -\sum_{c \in C} \frac{|S_c|}{|S|} \cdot \log_2 \left(\frac{|S_c|}{|S|} \right) \tag{1}$$

unde S_c reprezintă submulțimea lui S formată din exemplele etichetate cu clasa c.

Câștigul informațional

$$Gain(S, A_i) = H(S) - \sum_{v_j \in A_i} \frac{|S_j|}{|S|} \cdot H(S_j) \tag{2}$$

unde S_j reprezintă submulțimea lui S ce conține obiectele care au valoarea v_j pentru atributul A_i

4 Cerințe

Implementați algoritmul ID3 și afișați arborii construiți pentru cele două seturi de date din arhivă (`meteo.txt` și `invest.txt`, thanks to Andrei Olaru).