# Clasificarea Datelor Arbori de Decizie - Alg ID3

Cristian-Andrei Purcaru

UVT Master IACD An 1, 2016

Data Mining: PROF. DR. DANIELA ZAHARIE

## Problematica

Pe baza analizei unui set initial de exemple, se doreste predictia rezultatului unei noi probe.

Pentru a putea prezice rezultatul unui eveniment, ne vom folosi de baza de cunostinte. Baza de cunostinte este totalitatea evenimentelor inregistrate si a rezultatelor lor, in contextul in care fiecare a avut loc, numite si 'exemple' sau 'date de antrenament'.

Totusi, cunostintele nu pot fi folosite direct pentru predictie, ci este nevoie ca acestea sa fie analizate si prelucrate pentru a gasi sabloanele prin care contextul (atributele) este legate de rezultat (decizie), pentru ca orice combinatie de valori de atribute sa aibe parte de o predictie cat mai precisa a deciziei, chiar daca aceasta nu a fost intalnita in baza de cunostinte.

Aceasta este nevoia pe care algoritmul descris in continuare o acopera, nevoia de a structura atributele intr-o maniera in care acestea sa procure informatia necesara predictiei rezultatului noului context.

## Solutia

Un mod de a clasifica datele dintr-o baza de cunsotinte pentru a extrage alte date prin care sa se poata prezice rezultatul unui eveniment, avand la dispozitie contextul (la nivel de baza de cunostinte - atribute) in care are loc acesta, este crearea Arborilor de Decizie prin algoritmul ID3.

Pe scurt, ID3 construieste un Arbore de Decizie dintr-un set fix de exemple, arbore care este folosit pentru clasificarea probelor viitoare. Exemplele au cateva atribute si apartin unei clase precum 'da' sau 'nu', numita si 'decizie'.

Algoritmul ID3 se baseaza pe algoritmul Concept Learning System.

Descrierea algortmului CLS pentru un set de instante de antrenament C:

**Pas 1**: Daca toate instantele lui C sunt pozitive, creaza nodul 'Da' si opreste;

Daca toate instantele lui C sunt negative, creaza nodul 'Nu' si opreste;

Altfel selecteaza un attribut, F cu valorile v1, ..., vn si creaza un nod de decizie;

**Pas 2**: Partitioneaza instantele de antrenament in C in subseturi C1, C2, ..., Cn conform valorilor lui F;

**Pas 3**: Aplica algoritmul recursiv la toate subseturile Ci.

Nota: Expertul decide care atribut va fi selectat.

ID3 aduce o imbunatatire substantiala algoritmului CLS printr-o euristica de selectie a atributelor. Euristica ordoneaza atributele in functie de cantitatea de informatie pe care o furnizeaza.

## Rezultate

Rezultatele sunt referitoare la un set mic de date: 10 exemple; 3 noi probe; 4 atribute; 2 clase (True/ False). Presupunem ca se cere ca ID3 sa decida daca intr-o zi cu o anumita lista de valori ale atributelor este potrivita pentru a juca football.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Humidity | Wind | Degrees | Decision |
| 1 | sunny | high | week | 05-10 | True |
| 2 | overcast | normal | strong | 10-15 | False |
| 3 | overcast | high | week | 20-25 | True |
| 4 | rain | high | strong | 20-25 | False |
| 5 | rain | normal | week | 20-25 | False |
| 6 | rain | normal | week | 15-20 | False |
| 7 | overcast | normal | week | 10-15 | False |
| 8 | overcast | normal | week | 15-20 | True |
| 9 | sunny | high | strong | 05-10 | True |
| 10 | rain | normal | week | 20-25 | False |

Probe de test:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Humidity | Wind | Degrees | Decision |
| 11 | sunny | high | strong | 05-10 | True |
| 12 | sunny | normal | strong | 10-15 | True |
| 13 | overcast | high | week | 15-20 | 50% |

## Concluzie

Algoritmul ID3 este un algoritm relativ rapid de implementat si care produce rezultate satisfacatoare in multe cazuri, insa dupa cum se poate observa si in rezultate, nu in toate cazurile. Structura arborelui rezultate face ca in unele cazuri, predictia sa nu fie una clara, poate nici chiar preponderent falsa sau preponderent adevarata, ci efectiv 1/2.

Acest neajuns se datoreaza, probabil, faptului ca algoritmul ID3 nu este gandit ca, pentru o cat mai precisa predictie, sa foloseasca atat de multe atribute cat e necesar, ci este limitat la numarul de nivele ale arborelui care, evident, nu este egal cu numarul de atribute disponibile.

## Bibliografie

http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm