Salomon
System przetwarzania wiedzy

---

1 Drzewa decyzyjne w teorii decyzji

W teorii decyzji drzewo decyzyjne jest drzewem decyzji i ich mozliwych konsekwencji (stanów natury). Zadaniem drzew decyzyjnych może być zarówno stworzenie planu, jak i rozwiazanie problemu decyzyjnego.
Metoda drzew decyzyjnych jest szczególnie przydatna w problemach decyzyjnych z licznymi, rozgałeziajacymi sie wariantami oraz w przypadku podejmowania decyzji w warunkach ryzyka.

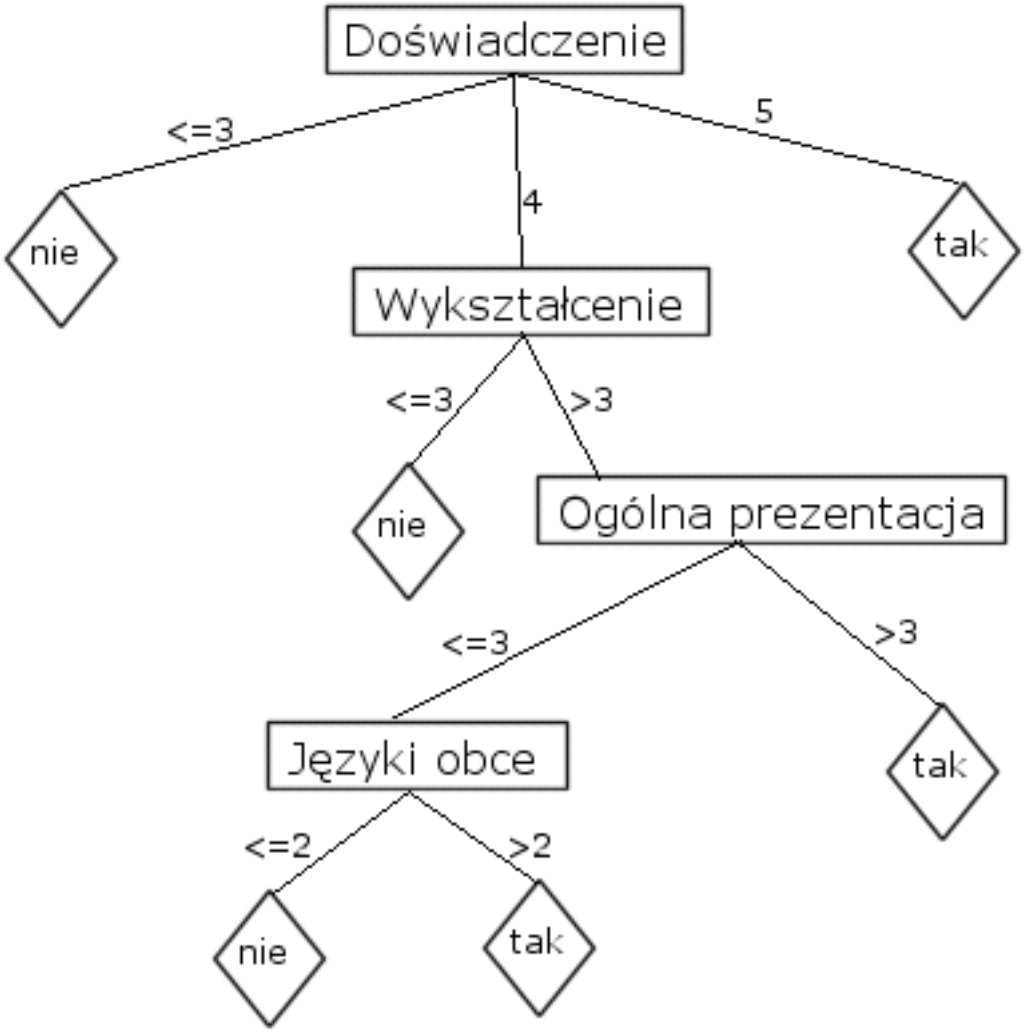
2 Drzewa decyzyjne w uczeniu maszynowym

Drzewa decyzyjne w uczeniu maszynowym słuzy do wyodrebniania wiedzy z zestawu przykładów (patrz eksploracja danych). Zakładamy, ze posiadamy zestaw przykładów: obiektów opisanych przy pomocy atrybutów, którym przyporzadkowujemy jakas decyzje (patrz tabela decyzyjna).
Przykład: Chcemy zautomatyzować proces przyjmowania kandydatów na praktyki w duzej firmie. Posiadamy setki przykładów z przeszłosci, chcemy wydobyć z nich reguły decyzyjne. Atrybuty Wykształcenie, Jezyki obce, Doswiadczenie i Ogólne wrazenie sa kodowane skala od 1 do 5.

Wiek	Płeć	Wykształcenie	Jezyki	Doswiadczenie	Prezentacja	Przyjety
25	m	2	4	1	4	nie
22	k	4	3	4	2	nie
21	m	4	5	5	4	tak
29	m	1	3	2	3	nie

Table 1:

Na podstawie tabeli decyzyjnej tworzymy drzewo, którego wezłami sa poszczególne atrybuty, gałeziami wartosci odpowiadajace tym atrybutom, a liscie tworza poszczególne decyzje. Na podstawie przykładowych danych wygenerowano nastepujace drzewo:


Doświadczenie
<=3
4
5
nie
Wykształcenie
tak
<=3
>3
nie
Ogólna prezentacja
<=3
>3
Języki obce
tak
<=2
>2
nie
tak

Drzewo w takiej postaci odzwierciedla w jaki sposób były na podstawie atrybutów były podejmowane decyzje klasyfikujace (dla uproszczenia połaczono niektóre gałezie). Zaleta tej reprezentacji jest jej czytelnosć dla człowieka. W prosty sposób mozna przekształcić ja do reprezentacji regułowej.

3 Algorytm tworzenia drzewa ID3

3.1 W duzym skrocie

Dopuki kazdy z lisci drzewa nie nalezy do tej samej klasy rownowaznosci (nie jest homogeniczny - jego zmienne decyzyjne nie sa jednakowe) powtarzaj:

- Wybierz niehomogeniczny lisc
- Zamien ten lisc na wezel testowy dzielacy ten podzbior na tak niehomogeniczne podzbiory jak to mozliwe, zgodnie z wyliczeniem entropii

3.2 Troche dokladniej

- Oblicz entropie dla kazdego z atrybutow, ktore chcemy wykorzystac do tworzenia nastepnego poziomu drzewa item Wybierz najlepszy z nich (ten ktory minimalizuje entropie)

3.3 Minimalizacja entropii?

Generalnie zalozenie jest aby tworzyc drzewa decyzyjne optymalnej wielkosci ale z praktycznego punktu widzenia nie jest to uzasadnione ze wzgledu na duzy koszt obliczeniowy. W zastepstwie korzystamy z przyblizonych procedur tworzenia malych, ale niekoniecznie najmniejszych drzew decyzyjnych.

3.4 Wybieranie atrybutu do obliczania entropii

Najwazniejszym punktem algorytmu ID3 jest wybor atrybutu do testowania na kazdym lisciu drzewa.
Procedura:

- Nalezy sprawdzic jak atrybut dzieli elementy nalezace do liscia
- Minimalizuj srednia entropie (oblicz entropie dla kazdego z wezlow, dla kazdego z z atrybutow i wybierz ten z najmniejsza entropia

3.5 Formula Entropii

Entropia jest miara z teorii informacji, charakteryzujaca czystosc i homogenicznosc zbioru atrybutow.

3.5.1 Dane

- nb, liczba instancji w lisciu b
- nbc, liczba instancji w lisciu b nalezacych do klasy c. nbc <= nb
- nt, calkowita liczna instancji we wszystkich lisciach

3.5.2 Prawdopodobienstwo

$P_b = \frac{n_{bc}}{n_b}$

- Jezeli wszystkie instancje w grupie sa klasyfikowane pozytywnie, wtedy Pb = 1 (lisc homogeniczny pozytywnie)
- Jezeli wszystkie instancje w grupie sa klasyfikowane negatywnie, wtedy Pb = 0 (lisc homogeniczny negatywnie)

3.5.3 Entropia

$Entropia = Sum(c) - (\frac{n_{bc}}{n_b})log_2(\frac{n_{bc}}{n_b})$

- Entropia jest zerowa jezeli zbior jest idealnie homogeniczny
- Entropie wynosi 1 jezeli zbior jest idealnie niehomogeniczny ze wzgledu na atrybut (tzn. nie jest on dzielony na zadne podgrupy przez ten atrybut)

3.5.4 Srednia entropia

$Sredniaentropia = Sum(b) * (\frac{n_b}{n_t}) * [Sum(c) - (\frac{n_{bc}}{n_b})log_2(\frac{n_{bc}}{n_b})]$