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.

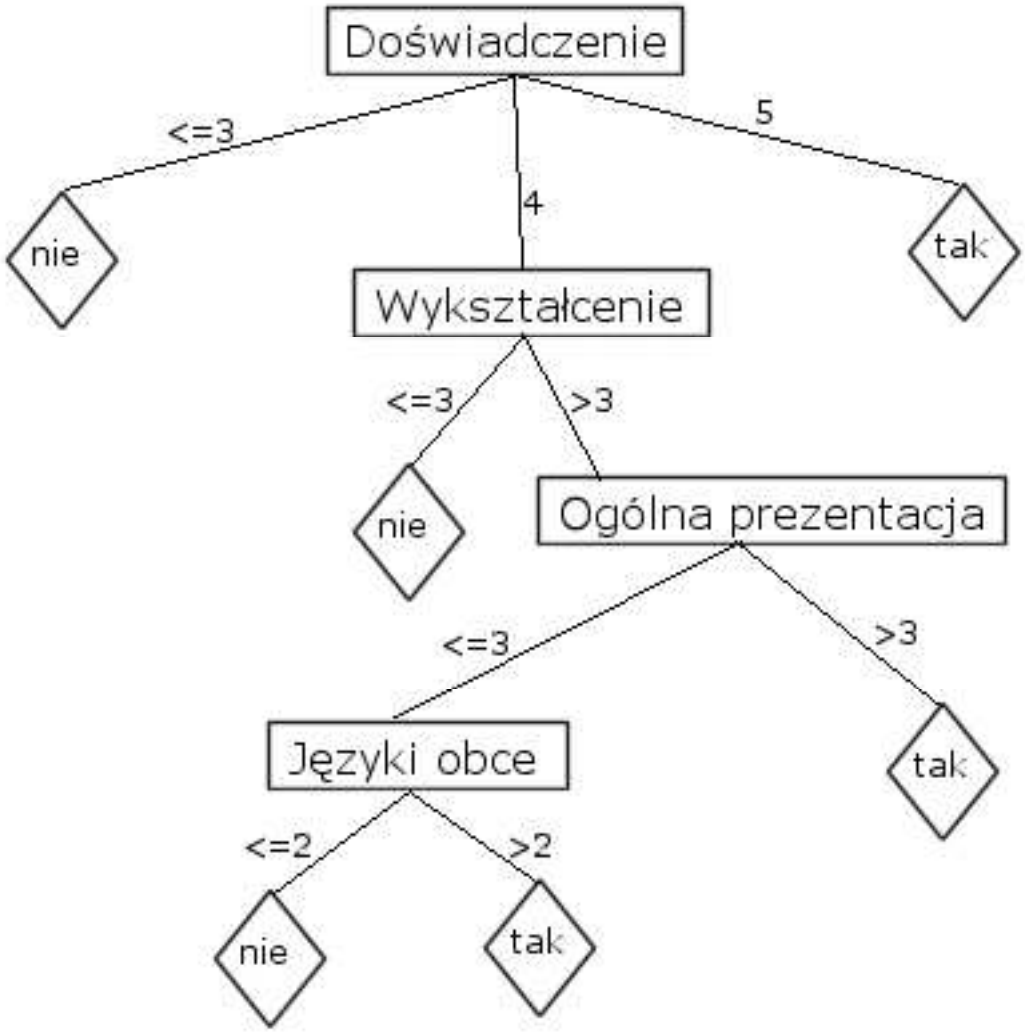
2 Drzewa decyzyjne w uczeniu maszynowym

Drzewa decyzyjne w uczeniu maszynowym słuza 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:



Drzewo w takiej postaci odzwierciedla w jaki sposób 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 Wstep

Algorytm tworzenia drzew decyzyjnych ID3 jest jednym z prostszych algorytmow ale zarazem daje on dosc dobre wyniki. Celem jest oczywiscie stworzenie drzewa, ktore

za pomoca wartosci atrybutow przyjmowanych przez elementy podzieli nam dziedzine na klasy rownowaznosci (w domysle podgrupy majace taka sama wartosc jednej ze zmiennych). Oczywiscie w typowym przypadku mozliwosci stworzenia takiego drzewa bedzie wiele. Wiec ustalamy dodatkowy cel, jakim bedzie minimalna wysokosc drzewa, liczona jako najwieksza odleglosc od korzenia do liscia. Algorytm ID3 zawsze (jezeli to mozliwe) stworzy nam drzewo decyzyjne. Natomiast nie zawsze jest to drzewo optymalnej wielkosci. Algorytm ID3 jest algorytmem zachłannym, decyzje o rozbudowie drzewa sa podejmowane na podstawie przyblizonej oceny kazdego z wariantow jakie mozemy przyjac w danym kroku. Raz podjeta decyzja nie jest juz zmieniana - nie jest to algorytm adaptatywny. Przyjrzyjmy sie jego dzialaniu.

3.2 Informacje wstepne

Na wejsciu algorytm dostaje zestaw danych, zwanych "test cases". Przy pomocy tych danych bedzie budowane drzewo decyzyjne. Danymi sa rekordy posiadajace wiele atrybutow. Konieczne jest okreslenie ktory z atrybutow jest artybutem wzgledem ktorego bedziemy tworzyc drzewo decyzyjne. W zamieszczonym wczesniej przykladzie jest to fakt przyjecia lub nieprzyjecia kandydata do pracy. Milczacym zalozeniem jest ze pozostale atrybuty maja byc brane pod uwage przy tworzeniu drzewa decyzyjnego. Drzewo sklada sie z zestawu wezlow i lisci. Wezly w stworzonym drzewie beda reprezentowac testy wartosci atrybutu a liscie beda podjetymi decyzjami. Odnoszac to do opisanego wczesniej przypadku mozna powiedziec ze np. test wartosci wspolczynnika okreslajacego znajomosc jezykow obcych (podzial na rekordy <=2 i >2) jest wezlem drzewa a decyzja ze kandydat zostal przyjety jest jego lisciem. Poczatkowo drzewo sklada sie z jednego tylko liscia do ktorego przypiete sa wszystkie "test cases".
Oznaczmy:

- 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

Teraz przy pomocy tych oznaczen mozemy zapisac podstawowe wzory prawdopodobiensta: $P_b = \frac{n_{bc}}{n_b}$

- 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)

Bazujac na tym zapisie prawdopodobienstwa zdefiniujejmy sobie formule entropii - bedzie nam ona potrzebna pozniej, przy tworzeniu drzewa decyzyjnego.

3.2.1 Entropia

Entropia jest miara z teorii informacji, charakteryzujaca czystosc i homogenicznosc zbioru atrybutow. $Entropia = Sum(c)(-\frac{n_{bc}}{n_b})log_2(\frac{n_{bc}}{n_b})$

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

3.2.2 Srednia entropia

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

3.3 Algorytm tworzenia drzewa decyzyjnego ID3 w duzym skrocie

Dopoki kazdy z lisci drzewa nie jest homogeniczny (zmienne decyzyjne jego elementow nie sa jednakowe) powtarzaj:

- Wybierz sposrod nieuzytych jeszcze atrybutow, ten ktory minimalizuje srednia entropie
- Rozwin niehomogeniczne liscie wzgledem wybranego atrybutu.

3.4 Minimalizacja entropii = Minimalizacja wysokosci drzewa ???

Ogolne 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.

3.5 Algorytm w pseudokodzie

1. Zainicjuj drzewo (wszystkie elementy przypisane do korzenia, ktory jest lisciem)
2. Dopoki nie wszystkie liscie sa homogeniczne powtarzaj

   - Jezeli nie ma nieuzytych atrybutow -> **koniec z bledem**

   - Oblicz srednia entropie dla nieuzytych jeszcze atrybutow

   - Wybierz ten atrybut, ktory minimalizuje srednia entropie (dla ktorego wyliczony w poprzednim punkcie wskaznik jest najmniejszy)

- Rozwijaj niehomogeniczne liscie wzgledem wybranego atrybutu (lisc staje sie wezlem, do ktorego sa "przyczepione" liscie powstale z podzialu tego liscia na liscie zawierajace kazda z przyjmowanych przez wybrany atrybut wartosci)

3. Wypisz drzewo -> **poprawny koniec**

3.6 Podsumowanie

Algorytm ID3 jest prostym algorytmem pozwalajacym na generowanie poprawnych drzew decyzyjnych. Jego podstawowa forma, opisana powyzej umozliwia tworzenie poprawnych drzew, niekoniecznie posiadajacych minimalna wysokosc. Dodatkowo podstawowa forma algorytmu nie uwzglednia "laczenia" wartosci w grupy, czyli np. dla opisanego na poczatku przykladu dla testu wartosci "jezyki obce" nie otrzymalibysmy dwoch lisci o wartosciach <=2 i >2 tylko 4 liscie o wartosciach 1,2,4,5. Usuwanie nadmiarowosci z drzewa jest juz elementem rozszerzen tego algorytmu.