SYSTEMY UCZĄCE SIĘ

WYKŁAD 3. DRZEWA DECYZYJNE

Dr hab. inż. Grzegorz Dudek
Wydział Elektryczny
Politechnika Częstochowska

Częstochowa 2014

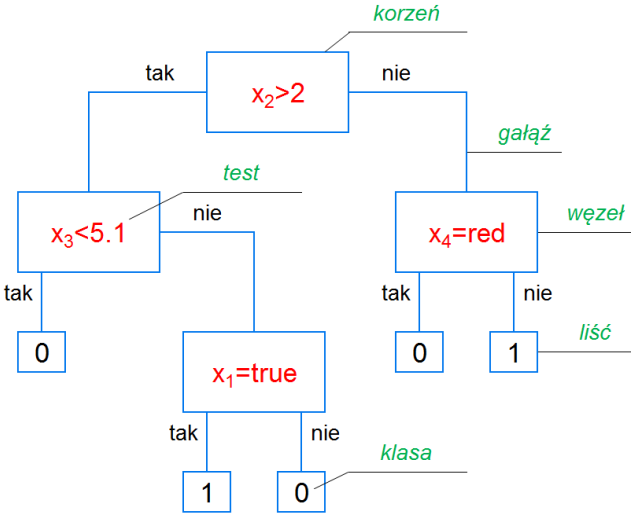
BUDOWA DRZEW DECYZYJNYCH

Drzewa decyzyjne są metodą indukcyjnego uczenia się pojęć i reprezentacji hipotez.

Drzewo decyzyjne jest strukturą złożona z węzłów, gałęzi i liści (węzłów terminalnych).

Węzły odpowiadają testom przeprowadzanym na wartościach atrybutów, gałęzie odpowiadają wynikom tych testów, a liście – etykietom kategorii (klas).

Przykładowe drzewo decyzyjne z testami binarnymi, dla dwóch kategorii i czterech atrybutów:

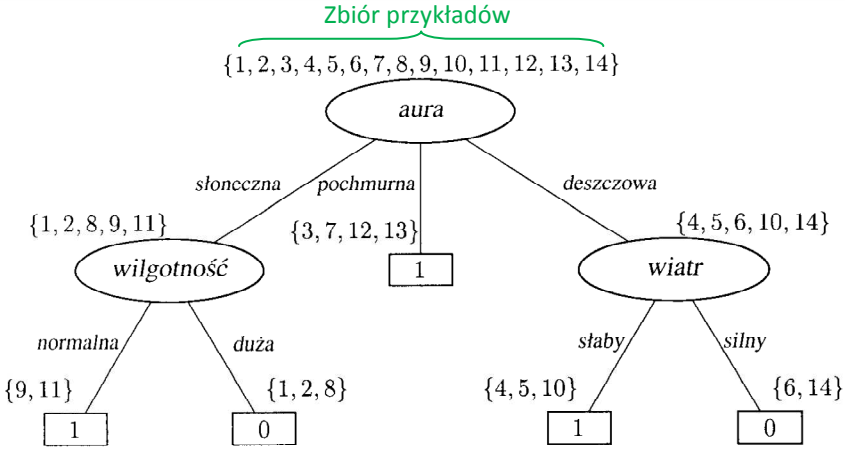


BUDOWA DRZEW DECYZYJNYCH

Testy *T* dzielą zbiór przykładów na podzbiory.

Jeśli przez *A* oznaczymy zbiór atrybutów, a przez R_T zbiór możliwych wyników testu (np. {tak, nie}), test można przedstawić jako funkcję

$$t : A \rightarrow R_T.$$



Każdemu z wyników testu odpowiada gałąź prowadząca z węzła do poddrzewa.

Określenie kategorii przykładu za pomocą drzewa decyzyjnego polega na przejściu od korzenia do jednego z liści, przez wykonanie w odwiedzanych węzłach skojarzonych z nimi testów i poruszanie się po gałęziach odpowiadających uzyskiwanym wynikom.

ZSTĘPUJĄCY SCHEMAT KONSTRUKCJI DRZEW DECYZYJNYCH

Drzewa decyzyjne najczęściej buduje się według zstępującego schematu konstrukcji [Cic, s. 146].

Zaczynając od korzenia (poziom 0) przemieszczamy się do kolejnych poziomów podejmując decyzje o utworzeniu liścia bądź węzła. Jeśli kryterium stopu jest spełnione tworzony jest liść i na podstawie podzbioru przykładów, które do niego dotarły ustalana jest jego etykieta. W przeciwnym przypadku tworzony jest węzeł i wybierany jest dla niego test.

Konstrukcja drzewa decyzyjnego – algorytm rekurencyjny

funkcja *buduj-drzewo*(P, d, S)
argumenty wejściowe:

- P — zbiór przykładów etykietowanych pojęcia c,
- d — domyślna etykieta kategorii,
- S — zbiór możliwych testów;

zwraca: drzewo decyzyjne reprezentujące hipotezę przybliżającą c na zbiorze P;

1: **jeśli** *kryterium-stopu*(P, S) **to**
2: utwórz liść $\mathbf{l}$;
3: $d_{\mathbf{l}} := kategoria(P, d)$;
4: **zwróć** $\mathbf{l}$;
5: **koniec jeśli**
6: utwórz węzeł $\mathbf{n}$;
7: $t_{\mathbf{n}} := wybierz\text{-}test(P, S)$;
8: $d := kategoria(P, d)$;
9: **dla wszystkich** $r \in R_{t_{\mathbf{n}}}$ **wykonaj**
10: $\mathbf{n}[r] := buduj\text{-}drzewo(P_{t_{\mathbf{n}}r}, d, S - \{t_{\mathbf{n}}\})$;
11: **koniec dla**
12: **zwróć** $\mathbf{n}$.

d jest etykietą przypisywaną liściowi, jeśli właściwej etykiety nie można ustalić na podstawie *P* i *S*,
r jest wynikiem testu

KRYTERIUM STOPU I USTALENIE ETYKIETY

Liść tworzony jest w przypadkach gdy:

- aktualny zbiór przykładów zawiera wyłącznie przykłady jednej kategorii; wtedy do etykiety liścia wpisuje się symbol tej kategorii,
- aktualny zbiór przykładów jest pusty; wtedy do etykiety liścia wpisuje się symbol kategorii domyślnej (np. kategorii najliczniej reprezentowanej w zbiorze *P* na wcześniejszym poziomie rekursji),
- wyraźna większość przykładów z aktualnego zbioru przykładów ma tę samą kategorię; wtedy do etykiety liścia wpisuje się symbol tej kategorii (ma to związek z generalizacją),
- wyczerpał się zbiór testów *S*; wtedy do etykiety liścia wpisuje się symbol kategorii domyślnej.

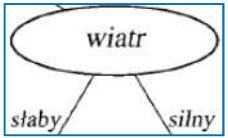
Ostateczne kryterium stopu *kryterium_stopu*(*P*,*S*):

$$P = \emptyset \vee \left|\{d' \in C \mid (\exists x \in P)\ c(x) = d'\}\right| = 1 \vee S = \emptyset$$

Wybór kategorii liścia:

$$kategoria(P, d) = \begin{cases} d & \text{jeśli } P = \emptyset, \\ \arg\max_{d'} \left|P^{d'}\right| & \text{w przeciwnym przypadku.} \end{cases}$$

RODZAJE TESTÓW



- testy tożsamościowe – sprawdzamy wartość atrybutu; wynikiem testu jest wartość atrybutu, np. dla atrybutu *wiatr* R_t = {słaby, silny} (N, P)
- testy równościowe (N, P)

$$t(x) = \begin{cases} 1 & \text{jeśli } a(x) = v, \\ 0 & \text{jeśli } a(x) \neq v, \end{cases}$$

gdzie *a* jest atrybutem przykładu **x**, *v* jest jedną z możliwych wartości atrybutu, R_t = {1, 0}

Test dla atrybutów:
N – nominalnych,
P – porządkowych,
C - ciągłych

- testy przynależnościowe (N, P, C)

$$t(x) = \begin{cases} 1 & \text{jeśli } a(x) \in V, \\ 0 & \text{jeśli } a(x) \notin V, \end{cases}$$

gdzie *V* jest podzbiorem wartości atrybutu, R_t = {1, 0}

- testy podziałowe (N, P, C)

$$t(x) = \begin{cases} 1 & \text{jeśli } a(x) \in V_1 \\ 2 & \text{jeśli } a(x) \in V_2 \\ \ldots & \\ m & \text{jeśli } a(x) \in V_m \end{cases}$$

gdzie parami rozłączne zbiory V_1, V_2, …, V_m stanowią wyczerpujący podział przeciwdziedziny atrybutu, R_t = {1, 2, …, *m*}

- testy nierównościowe (P, C)

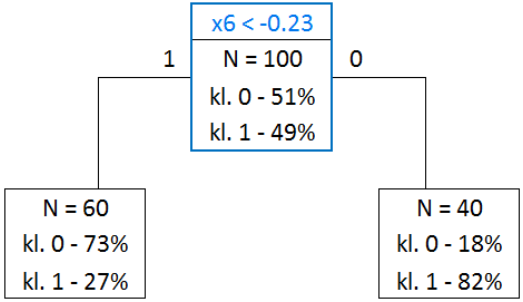
$$t(x) = \begin{cases} 1 & \text{jeśli } a(x) \leqslant \theta, \\ 0 & \text{jeśli } a(x) > \theta, \end{cases}$$

gdzie θ jest wartością progową z przeciwdziedziny atrybutu, R_t = {1, 0}

KRYTERIUM WYBORU TESTU

Dla tworzonego węzła należy wybrać najlepszy test ze zbioru możliwych testów *wybierz_test*(*P,S*).

Test powinien dzielić zbiór przykładów tak, aby po podziale w podzbiorach przykładów ich kategorie były silnie zróżnicowane.



W teorii informacji informację zawartą w zbiorze przykładów wyraża wzór:

$$I(P) = \sum_{d \in C} -\frac{|P^d|}{|P|} \log \frac{|P^d|}{|P|}$$

Informacja jest duża, gdy liczba przykładów poszczególnych kategorii jest zbliżona.

Entropia zbioru przykładów *P* ze względu na wynik *r* testu *t*:

$$E_{tr}(P) = \sum_{d \in C} -\frac{|P_{tr}^d|}{|P_{tr}|} \log \frac{|P_{tr}^d|}{|P_{tr}|}$$

Entropia zbioru przykładów *P* ze względu na test *t*:

$$E_t(P) = \sum_{r \in R_t} \frac{|P_{tr}|}{|P|} E_{tr}(P)$$

Przykład. Dany jest zbiór 100 przykładów, spośród których 51 należy do klasy 0, a 49 do klasy 1. W wyniku testu zbiór ten został podzielony na dwa podzbiory. Pierwszy z nich zawiera 60 przykładów: 44 z klasy 0 i 16 z klasy 1. Drugi podzbiór zawiera 40 przykładów: 7 z klasy 0 i 33 z klasy 1. Oblicz entropie przed i po podziale.

Przed podziałem: $I(P) = -\frac{51}{100}\log\frac{51}{100} - \frac{49}{100}\log\frac{49}{100} = 0.6929$

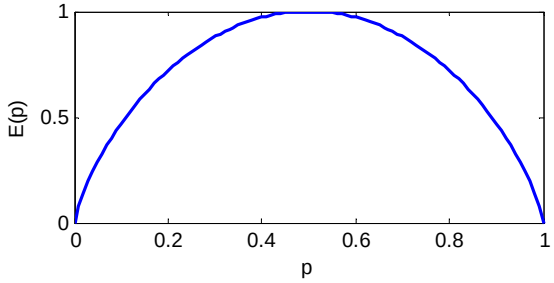
Po podziale: $E_{t1}(P) = -\frac{44}{60}\log\frac{44}{60} - \frac{16}{60}\log\frac{16}{60} = 0.5799$, $E_{t0}(P) = -\frac{7}{40}\log\frac{7}{40} - \frac{33}{40}\log\frac{33}{40} = 0.4637$

$$E_t(P) = \frac{60}{100}0.5799 + \frac{40}{100}0.4671 = 0.5334$$

Osiągnięto redukcję entropii z 0.6929 do 0.5334.

KRYTERIUM WYBORU TESTU

Zależność $E(P) = -p\log_2 p - (1-p)\log_2(1-p), \quad p = |P^0|/|P|$:



Zbiór przykładów o równomiernym rozkładzie kategorii ma maksymalna entropię. Przy przewadze jednej kategorii nad drugą entropia się zmniejsza.

Przyrost informacji wynikający z zastosowaniu testu *t* do zbioru przykładów *P*:

$$g_t(P) = I(P) - E_t(P)$$

Kryterium wyboru testu – wybierz test maksymalizujący przyrost informacji $\arg\max_t g_t(P)$ lub minimalizujący entropię $E_t(P)$ (*I*(*P*) jest niezależne od testu).

Spotyka się kryteria oparte na innych miarach zróżnicowania klas, np. na indeksie Giniego:

$$G_{tr}(P) = \sum_{d \in C} \frac{|P_{tr}^d|}{|P_{tr}|}\left(1 - \frac{|P_{tr}^d|}{|P_{tr}|}\right); \quad G_t(P) = \sum_{r \in R_t} \frac{|P_{tr}|}{|P|} G_{tr}(P)$$

USTALANIE ZBIORU TESTÓW KANDYDUJĄCYCH *S*

Korzystając z kryterium wyboru testu *wybierz_test*(*P,S*) wybieramy optymalny test ze zbioru *S*. Ale jak tworzony jest zbiór *S*? Zależy to od typu testu.

- Testy tożsamościowe – możliwych testów jest tyle ile atrybutów.
- Testy równościowe – możliwych testów dla jednego atrybutu (nominalnego lub porządkowego) jest tyle ile ten atrybut posiada wartości.
- Testy przynależnościowe i podziałowe – możliwych testów dla jednego atrybutu jest tyle ile podzbiorów *V* zdefiniujemy.
- Testy nierównościowe – możliwych testów dla jednego atrybutu (ciągłego lub porządkowego) jest tyle ile przyjmuje on różnych wartości w zbiorze trenującym minus 1.

  Sposób ustalania zbioru testów kandydujących dla testów nierównościowych:

  1. Posortuj wartości atrybutu w zbiorze *P*.
  2. Przyjmij środki przedziałów sąsiednich wartości atrybutu po posortowaniu jako progi θ.
  3. Progi te definiują zbiór testów kandydujących.

USTALANIE ZBIORU TESTÓW KANDYDUJĄCYCH *S*

Przykład. Dany jest zbiór przykładów opisanych atrybutem ciągłym *masa* i nominalnym *kolor*.

Nr przykładu	1	2	3	4	5	6	7	8	9	10
Masa	2	6	4	8	7	6	1	6	8	8
Kolor	R	B	R	G	G	R	B	B	B	G
Klasa	0	0	1	1	1	0	0	1	1	0

Ustal zbiór testów kandydujących i wybierz najlepszy test.

Testy dla *kolor*: *kolor* = R, *kolor* = G, *kolor* = B.

Testy dla *masa*:

1. Sortujemy wartości *masy*: 1 | 2 | 4 | 6 6 6 | 7 | 8 8 8
2. Wyznaczamy progi θ: 1.5, 3, 5, 6.5, 7.5
3. Testy kandydujące: *masa* $\leq$ 1.5, *masa* $\leq$ 3, *masa* $\leq$ 5, *masa* $\leq$ 6.5, *masa* $\leq$ 7.5

Zbiór S	*kolor* = R	*kolor* = G	*kolor* = B	*masa* $\leq$ 1.5	*masa* $\leq$ 3	*masa* $\leq$ 5	*masa* $\leq$ 6.5	*masa* $\leq$ 7.5
Przykłady zaliczające	1 3 6	4 5 10	2 7-9	7	1 7	1 3 7	1-3 6-8	1-3 5-8
Przykłady niezaliczające	2 4 5 7-10	1-3 6-9	1 3-6	1-6 8-10	2-6 8-10	2 4 5 6 8-10	4 5 9 10	4 9 10
$g_t(P)$								

PRZYCINANIE DRZEWA

Przycinanie drzewa redukuje efekt przeuczenia (mały błąd uczący, duży testowy).

Alternatywą przycinania jest zapobieganie nadmiernemu wzrostowi w trakcie konstrukcji drzewa poprzez tworzenie liści w miejsce węzłów (z etykietą kategorii większościowej).

Przycinanie drzewa – zastępowanie wybranych węzłów (a tym samym całych poddrzew) liśćmi, którym przypisuje się kategorię większości przykładów trenujących, związanych z eliminowanym węzłem.

Zazwyczaj przycinanie realizowane jest w sposób wstępujący, tzn. w pierwszej kolejności rozważa się przycięcie węzłów położonych najniżej w drzewie. Podstawowe znaczenie ma kryterium przycinania, które decyduje, czy węzeł będzie zastąpiony liściem.

Najczęściej kryterium przycinania jest błąd na oddzielnym zbiorze przykładów (niezależnym od zbioru trenującego). Przycięcie następuje, jeśli na tym oddzielnym zbiorze liść uzyskuje błąd nie większy, niż błąd poddrzewa.

PRZYCINANIE DRZEWA

funkcja *przytnij-drzewo*$(\mathbb{T}, P)$
argumenty wejściowe:

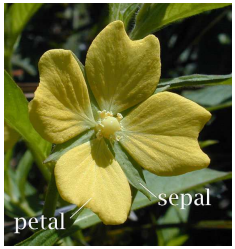
- $\mathbb{T}$ — drzewo do przycięcia,
- P — zbiór przycinania;

zwraca: drzewo $\mathbb{T}$ po przycięciu;

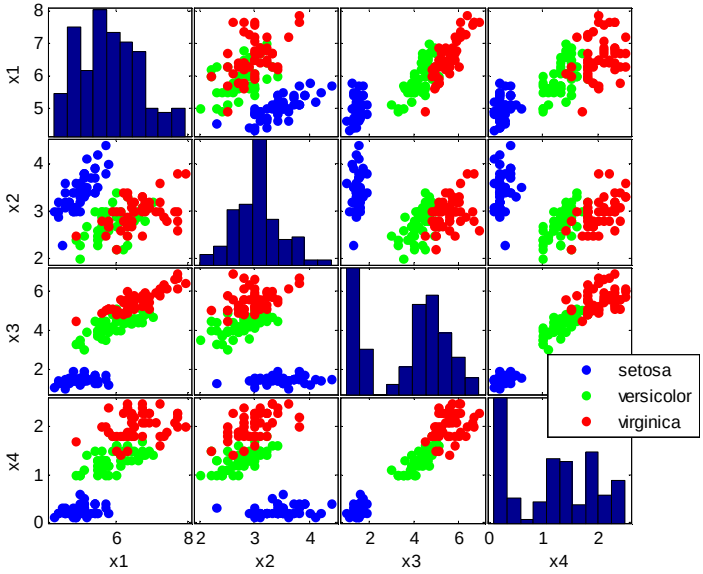
1: **dla wszystkich** węzłów $\mathbf{n}$ drzewa $\mathbb{T}$ **wykonaj**
2: zastąp $\mathbf{n}$ liściem $\mathbf{l}$ z etykietą większościowej kategorii w zbiorze $P_{\mathbb{T},\mathbf{n}}$ jeśli nie powiększy to szacowanego na podstawie P błędu rzeczywistego drzewa $\mathbb{T}$;
3: **koniec dla**
4: **zwróć** $\mathbb{T}$.

PRZYKŁAD APLIKACYJNY – DANE

Klasyfikacja zbioru danych Iris (Fisher's Iris)

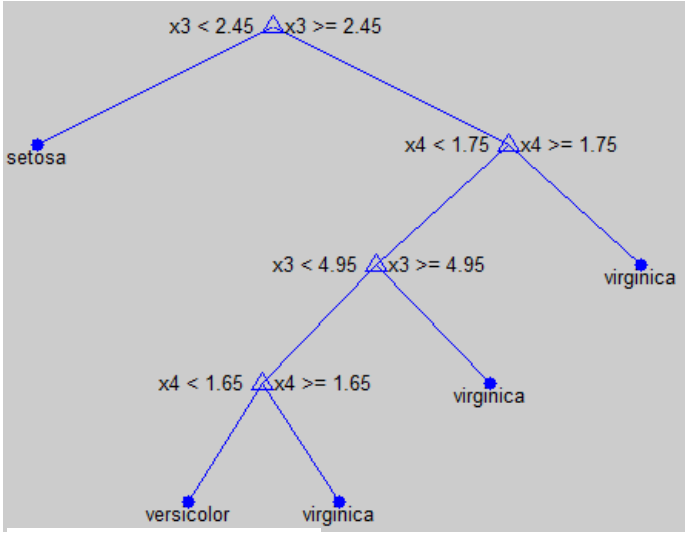


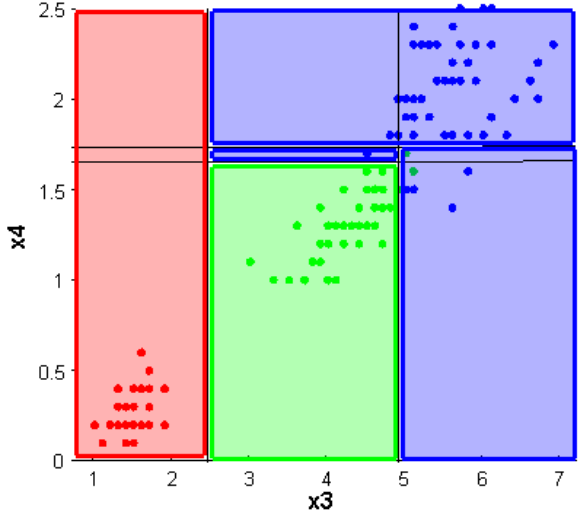
Zbiór zawiera po 50 przykładów każdego z trzech gatunków kwiatu: sentosa, virginica oraz versicolor. Przykłady zbudowane są z czterech atrybutów: szerokość i długość płatków sepal i petal



X1 (Sepal length)	X2 (Sepal width)	X3 (Petal length)	X4 (Petal width)	Klasa (Species)
5,1	3,5	1,4	0,2	1 (setosa)
4,9	3	1,4	0,2	1 (setosa)
…	…	…	…	…
6,1	2,8	4	1,3	2 (versicolor)
…	…	…	…	…
6,2	3,4	5,4	2,3	3 (virginica)
5,9	3	5,1	1,8	3 (virginica)

PRZYKŁAD APLIKACYJNY – WYNIKI





9

Zapis drzewa w postaci reguł:

1. if x3<2.45 then node 2 elseif x3>=2.45 then node 3 else setosa
2. class = setosa
3. if x4<1.75 then node 4 elseif x4>=1.75 then node 5 else versicolor
4. if x3<4.95 then node 6 elseif x3>=4.95 then node 7 else versicolor
5. class = virginica
6. if x4<1.65 then node 8 elseif x4>=1.65 then node 9 else versicolor
7. class = virginica
8. class = versicolor
9. class = virginica

Macierz błędów klasyfikacji (przekłamań)

W każdej komórce jest liczba przykładów należących do klasy *i*, a rozpoznanych jako należące do klasy *j* (oszacowane w procedurze leave-one-out)

	Klasa prawdziwa		
Klasa przewidywana ↓	1	2	3
1	**50**	**0**	**0**
2	**0**	**46**	**3**
3	**0**	**4**	**47**
nierozpoznana	**0**	**0**	**0**

Odsetek poprawnie sklasyfikowanych przykładów: **95,33%**