



Pozyskiwanie wiedzy

ANDRZEJ MACIOŁ

Metody pozyskiwania wiedzy

bezpośrednie
zapisanie
wiedzy

pozyskiwanie
wiedzy
na podstawie
instrukcji

pozyskiwanie
wiedzy
na podstawie
analogii

pozyskiwanie
wiedzy
na podstawie
przykładów

pozyskiwanie
wiedzy
na podstawie
obserwacji

Bezpośrednie zapisanie wiedzy

- ▶ uczenie na pamięć (ang. rote learning)
- ▶ system uczony (uczeń) otrzymuje gotową wiedzę w postaci kompletnych i spójnych zbiorów reguł zapisanych zgodnie z obowiązującymi w systemie zasadami zapisu wiedzy

Pozyskiwanie wiedzy na podstawie instrukcji

- ▶ uczenie przez przekazanie informacji (ang. learning by being told)
- ▶ istotną rolę w tej metodzie odgrywa nauczyciel, który tworzy wiedzę w postaci akceptowalnej przez system ekspertowy
- ▶ uczeń dokonuje integracji nowej wiedzy z pewną wiedzą aprioryczną
- ▶ nauczyciel narzuca natomiast strukturę i charakter zapisywanej wiedzy

Pozyskiwanie wiedzy na podstawie analogii

- ▶ polega na takiej transformacji istniejącej wiedzy, by mogła być użyteczna do opisów faktów podobnych (choć nie identycznych)
- ▶ proces ten może odbywać się bez nauczyciela choć wymaga aktywnego zaangażowania ucznia do wyszukiwania i „tłumaczenia” analogii

Pozyskiwanie wiedzy na podstawie przykładów

- ▶ generuje się ogólny opis pojęć (klas) na podstawie zbioru przykładów i/lub kontrprzykładów obiektów reprezentujących te pojęcia (klasy) – metoda indukcyjna
- ▶ przykłady są dostarczane przez nauczyciela

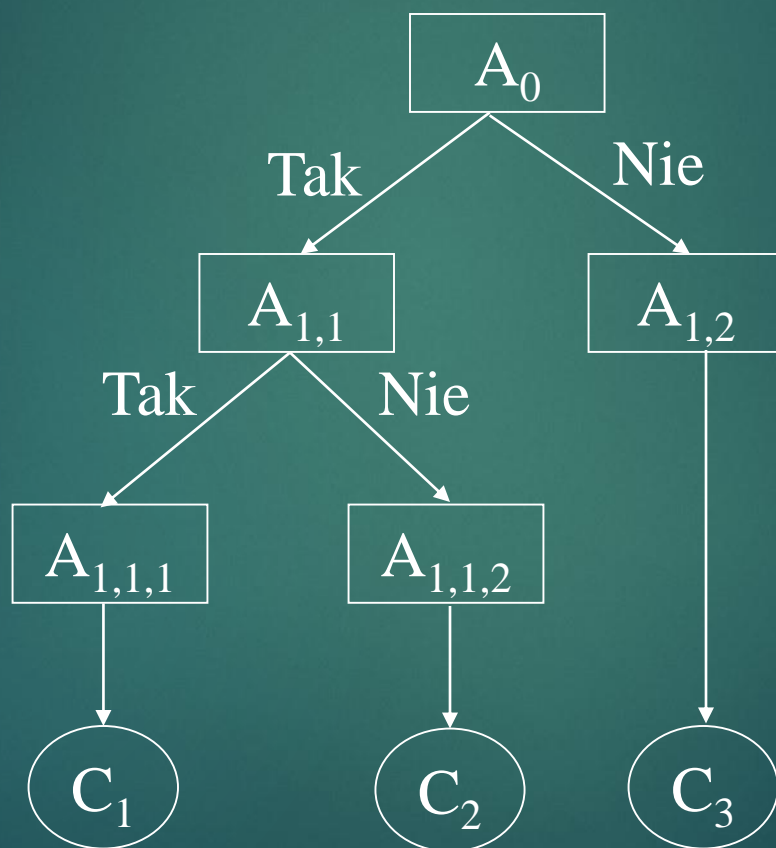
Pozyskiwanie wiedzy na podstawie obserwacji

- ▶ metoda indukcyjna oparty o przykłady (obserwacje) pochodzące ze świata zewnętrznego lub z eksperymentów – uczenie bez nauczyciela
- ▶ do indukcji można wykorzystywać techniki eksploracji danych (ang. data mining), grupowania, metody statystyki, sztuczne sieci neuronowe, algorytmy genetyczne

Przykłady metod pozyskiwania wiedzy

- ▶ Algorytm ID3
- ▶ Metoda generowania pokryć

Binarne drzewo decyzyjne



Algorytm ID3 - Quinlana

Entropia:

$$I = \sum_{i=1}^n (-p_i \log_2 p_i)$$

p_i – prawdopodobieństwo wystąpienia stanu i

$$\sum_{i=1}^n p_i = 1$$

Entropia - przykład

- ▶ Entropia eksperymentu polegającego na losowaniu w oparciu o „rzut monetą”

$$I = \sum_{i=1}^n (-p_i \log_2 p_i)$$

$$I = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}$$

$$I = -\frac{1}{2} (-1) - \frac{1}{2} (-1) = 1$$

Entropia - przykład

- ▶ Załóżmy, że moneta jest „oszukana” i prawdopodobieństwo wylosowania orła wynosi $1/4$ a reszki $3/4$

$$I = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4}$$

$$I = 0,81$$

- ▶ Ponieważ wiemy o oszustwie nasza niepewność jest mniejsza

Algorytm ID3 - Quinlana

Entropia w przypadku wielu przykładów i wielu rezultatów:

$$I = \sum_{k=1}^n \left(-\frac{n_k}{n} \log_2 \frac{n_k}{n} \right)$$

n_k — liczba przykładów należących do klasy k

n — liczba wszystkich przykładów

Entropia w ujęciu częstościowym - przykład

- ▶ Załóżmy, że znamy 10 przykładów wniosków kredytowych ocenionych pozytywnie
- ▶ Spośród nich 6 przypadków dotyczy kredytów spłaconych, 2 spłaconych po terminie i 2 niespłaconych w ogóle

$$I = -\frac{6}{10} \log_2 \frac{6}{10} - \frac{2}{10} \log_2 \frac{2}{10} - \frac{2}{10} \log_2 \frac{2}{10}$$

$$I = 1,37$$

Entropia w ujęciu częstościowym - przykład

- ▶ Załóżmy, że znamy 10 przykładów wniosków kredytowych ocenionych pozytywnie
- ▶ Obliczmy entropię, w przypadku gdy 3 przykłady dotyczyły kredytów spłaconych, 4 spłaconych z opóźnieniem i 3 niespłaconych

$$I = -\frac{3}{10} \log_2 \frac{3}{10} - \frac{4}{10} \log_2 \frac{4}{10} - \frac{3}{10} \log_2 \frac{3}{10}$$

$$I = 1,57$$

Entropia w ujęciu częstościowym - wnioski

- ▶ Jeżeli wiemy, że pozytywnie ocenione wnioski kredytowe są „raczej” spłacane to mamy większą wiedzę (a raczej mniejszą niewiedzę) niż w przypadku gdy wszystkie konsekwencje pozytywnej oceny wniosku dają podobny wynik

Znaczenie entropii

- ▶ Im wyższa jest miara entropii tym mniej wiemy o ocenianej sytuacji
- ▶ W pierwszym przypadku nie mieliśmy żadnych przesłanek by oceniać możliwość niespłacenia kredytu w oparciu o zewnętrzne informacje
- ▶ Wiemy jednak, że możliwość spłacenia kredytu jest znacznie wyższa niż niespłacenia czy opóźnienia

Algorytm ID3 - Quinlana


Entropia po ocenie warunku j na temat całego problemu:

$$E_j = \frac{n_j^+}{n} I_j^+ + \frac{n_j^-}{n} I_j^-$$

n_j^+ — liczba przykładów potwierdzonych przez warunek j

n_j^- — liczba przykładów zaprzeczonych przez warunek j

n — liczba wszystkich przykładów

- 
- ▶ Załóżmy, że uzyskujemy następującą dodatkową informację:
 - ▶ spośród 10 przykładów w 6 przykładach kredyty były zabezpieczone hipoteką a w 4 nie
 - ▶ wówczas

$$n_1^+ = 6$$

$$n_1^- = 4$$

$$n = 10$$

Algorytm ID3 - Quinlana

Entropia po ocenie warunku j :

$$I_j^+ = \sum_{k=1}^m (X_{j,k}^+)$$

$$I_j^- = \sum_{k=1}^m (X_{j,k}^-)$$

Algorytm ID3 - Quinlana

$$X_{j,k}^+ = \begin{cases} 0 & dla \ n_{j,k}^+ = 0 \\ -\frac{n_{j,k}^+}{n_j^+} \log_2 \frac{n_{j,k}^+}{n_j^+} & dla \ n_{j,k}^+ > 0 \end{cases}$$

$$X_{j,k}^- = \begin{cases} 0 & dla \ n_{j,k}^- = 0 \\ -\frac{n_{j,k}^-}{n_j^-} \log_2 \frac{n_{j,k}^-}{n_j^-} & dla \ n_{j,k}^- > 0 \end{cases}$$

Algorytm ID3 - Quinlana

- $n_{j,k}^+$ — liczba przykładów potwierdzających, że jeżeli warunek j jest spełniony to przykład należy do klasy k
- $n_{j,k}^-$ — liczba przykładów potwierdzających, że jeżeli warunek j nie jest spełniony to przykład należy do klasy k

Dodatkowe informacje

- ▶ Załóżmy, że w przypadku wniosków zabezpieczonych hipoteką (razem 6) 5 zostało spłaconych w terminie i jeden z opóźnieniem; natomiast kredyty nie zabezpieczone (razem 4) w jednym przykładzie został spłacony w terminie, w jednym spłacony z opóźnieniem a w dwóch niespłacony

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego zabezpieczonych hipoteką kredytów – spłaconych w terminie

$$j = 1, \quad k = 1$$

$$n_{1,1}^+ = 5, \quad n_j^+ = 6$$

$$X_{1,1}^+ = -\frac{n_{j,k}^+}{n_j^+} \log_2 \frac{n_{j,k}^+}{n_j^+} = -\frac{5}{6} \log_2 \frac{5}{6}$$

$$X_{1,1}^+ = 0,22$$

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego zabezpieczonych hipoteką kredytów – spłaconych z opóźnieniem

$$j = 1, \quad k = 2$$

$$n_{1,2}^+ = 1, \quad n_1^+ = 6$$

$$X_{1,2}^+ = -\frac{n_{j,k}^+}{n_j^+} \log_2 \frac{n_{j,k}^+}{n_j^+} = -\frac{1}{6} \log_2 \frac{1}{6}$$

$$X_{1,2}^+ = 0,43$$

Obliczenia

- ▶ Ponieważ żaden zabezpieczony kredyt nie pozostał niespłacony obliczamy łączną entropię po ocenie warunku „zabezpieczone kredyty”

$$I_1^+ = 0,22 + 0,43 + 0 = 0,65$$

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego nie zabezpieczonych hipoteką kredytów – spłaconych w terminie

$$j = 1, \quad k = 1$$

$$n_{1,1}^- = 1, \quad n_1^- = 4$$

$$X_{1,1}^- = -\frac{n_{j,k}^-}{n_j^-} \log_2 \frac{n_{j,k}^-}{n_j^-} = -\frac{1}{4} \log_2 \frac{1}{4}$$

$$X_{1,1}^- = 0,5$$

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego nie zabezpieczonych hipoteką kredytów – spłaconych z opóźnieniem

$$j = 1, \quad k = 2$$

$$n_{1,2}^- = 1, \quad n_1^- = 4$$

$$X_{1,2}^- = -\frac{n_{j,k}^-}{n_j^-} \log_2 \frac{n_{j,k}^-}{n_j^-} = -\frac{1}{4} \log_2 \frac{1}{4}$$

$$X_{1,2}^- = 0,5$$

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego nie zabezpieczonych hipoteką kredytów – niespłaconych

$$j = 1, \quad k = 3$$

$$n_{1,3}^- = 2, \quad n_1^- = 4$$

$$X_{1,3}^- = -\frac{n_{j,k}^-}{n_j^-} \log_2 \frac{n_{j,k}^-}{n_j^-} = -\frac{2}{4} \log_2 \frac{2}{4}$$

$$X_{1,2}^- = 0,5$$

Obliczenia

- Obliczamy entropię po ocenie warunku dotyczącego nie zabezpieczonych kredytów oraz ogółem przez informację o zabezpieczeniu

$$I_1^- = 0,5 + 0,5 + 0,5 = 1,5$$

$$E_j = \frac{n_j^+}{n} I_j^+ + \frac{n_j^-}{n} I_j^-$$

$$E_1 = \frac{6}{10} 0,65 + \frac{4}{10} 1,5 = 0,99$$

Dodatkowe informacje

- ▶ Załóżmy, że uzyskaliśmy dodatkową informację o przeznaczeniu kredytu; wśród 10 przykładów 5 były to kredyty konsumpcyjne a 5 na zakup samochodu
- ▶ Spośród kredytów konsumpcyjnych 3 zostały spłacone, 1 spłacony w terminie i jeden niespłacony
- ▶ Obliczmy entropię po uzyskaniu informacji, że kredyt był konsumpcyjny

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego kredytów konsumpcyjnych – spłaconych w terminie

$$j = 2, \quad k = 1$$

$$n_{1,1}^+ = 3, \quad n_j^+ = 5$$

$$X_{1,1}^+ = -\frac{n_{j,k}^+}{n_j^+} \log_2 \frac{n_{j,k}^+}{n_j^+} = -\frac{3}{5} \log_2 \frac{3}{5}$$

$$X_{1,1}^+ = 0,44$$

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego kredytów konsumpcyjnych – spłaconych z opóźnieniem

$$j = 2, \quad k = 2$$

$$n_{2,2}^+ = 1, \quad n_2^+ = 5$$

$$X_{2,2}^+ = -\frac{n_{j,k}^+}{n_j^+} \log_2 \frac{n_{j,k}^+}{n_j^+} = -\frac{1}{5} \log_2 \frac{1}{5}$$

$$X_{2,2}^+ = 0,46$$

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego kredytów konsumpcyjnych – niespłaconych

$$j = 2, \quad k = 3$$

$$n_{2,3}^+ = 1, \quad n_2^+ = 5$$

$$X_{2,3}^+ = -\frac{n_{j,k}^+}{n_j^+} \log_2 \frac{n_{j,k}^+}{n_j^+} = -\frac{1}{5} \log_2 \frac{1}{5}$$

$$X_{2,3}^+ = 0,46$$

Obliczenia

- Obliczamy łączny poziom entropii po ocenie przykładów dotyczących kredytów konsumpcyjnych

$$I_1^+ = 0,44 + 0,46 + 0,46 = 1,37$$

Dodatkowe informacje

- ▶ Załóżmy, że spośród kredytów innych niż konsumpcyjne 3 zostały spłacone w terminie, 1 z opóźnieniem i 1 w ogóle niespłacony

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego kredytów innych niż konsumpcyjne – spłaconych w terminie

$$j = 2, \quad k = 1$$

$$n_{1,1}^- = 3, \quad n_j^- = 5$$

$$X_{1,1}^- = -\frac{n_{j,k}^-}{n_j^-} \log_2 \frac{n_{j,k}^-}{n_j^-} = -\frac{3}{5} \log_2 \frac{3}{5}$$

$$X_{1,1}^- = 0,44$$

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego kredytów innych niż konsumpcyjne – spłaconych z opóźnieniem

$$j = 2, \quad k = 2$$

$$n_{2,2}^- = 1, \quad n_2^- = 5$$

$$X_{2,2}^- = -\frac{n_{j,k}^-}{n_j^-} \log_2 \frac{n_{j,k}^-}{n_j^-} = -\frac{1}{5} \log_2 \frac{1}{5}$$

$$X_{2,2}^- = 0,46$$

Obliczenia

- Obliczmy entropię po ocenie warunku dotyczącego kredytów innych niż konsumpcyjne – niespłaconych

$$j = 2, \quad k = 3$$

$$n_{2,3}^- = 1, \quad n_2^- = 5$$

$$X_{2,3}^- = -\frac{n_{j,k}^-}{n_j^-} \log_2 \frac{n_{j,k}^-}{n_j^-} = -\frac{1}{5} \log_2 \frac{1}{5}$$

$$X_{2,3}^- = 0,46$$

Obliczenia

- Obliczamy łączną entropię po ocenie przykładów dotyczących kredytów innych niż konsumpcyjne oraz ogółem przez informację o rodzaju kredytu

$$I_1^- = 0,44 + 0,46 + 0,46 = 1,37$$

$$E_j = \frac{n_j^+}{n} I_j^+ + \frac{n_j^-}{n} I_j^-$$

$$E_1 = \frac{5}{10} 1,37 + \frac{5}{10} 1,37 = 1,37$$

Algorytm ID3 - Quinlana

Warunek wyboru warunku k :

$$\max_j (I - E_j)$$

Porównanie dwóch informacji

- ▶ Informacja o zabezpieczeniu

$$I - E_1 = 1,37 - 0,99 = 0,38$$

- ▶ Informacja o rodzaju kredytu

$$I - E_2 = 1,37 - 1,37 = 0,0$$

Przykład – dobór formy reklamy



wiek	<20	1	1	0	0	0	1	1	0	0	0	0	0
	20-30	0	0	1	0	1	0	0	1	1	0	0	0
	>30	0	0	0	1	0	0	0	0	0	1	1	1
płeć	K	1	0	0	0	1	0	1	1	0	1	1	0
	M	0	1	1	1	0	1	0	0	1	0	0	1
mieszka	wieś	0	0	0	0	0	1	1	1	1	1	0	1
	miasto	1	1	1	1	1	0	0	0	0	0	1	0
reklama	Internet	1	1										
	Prasa			1	1	1							
	Telewizja						1	1	1	1	1	1	1

$$n = 12$$

$$n_1 = 2$$

$$n_2 = 3$$

$$n_3 = 7$$

Przykład – dobór formy reklamy

$$I = -\frac{n_1}{n} \log_2 \frac{n_1}{n} - \frac{n_2}{n} \log_2 \frac{n_2}{n} - \frac{n_3}{n} \log_2 \frac{n_3}{n}$$

$$I = -\frac{2}{12} \log_2 \frac{2}{12} - \frac{3}{12} \log_2 \frac{3}{12} - \frac{7}{12} \log_2 \frac{7}{12}$$

$$I = 1,3844$$

Przykład – dobór formy reklamy



Entropia po potwierdzeniu warunku 1

$$I_1^+ = -\frac{n_{1,1}^+}{n_1^+} \log_2\left(\frac{n_{1,1}^+}{n_1^+}\right) - 0 - \frac{n_{1,3}^+}{n_1^+} \log_2\left(\frac{n_{1,3}^+}{n_1^+}\right)$$

$$I_1^+ = -\frac{2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right)$$

$$I_1^+ = 1$$

Przykład – dobór formy reklamy

Entropia po zaprzeczeniu warunku 1

$$I_1^- = 0 - \frac{n_{1,2}^-}{n_1^-} \log_2 \left(\frac{n_{1,2}^-}{n_1^-} \right) - \frac{n_{1,3}^-}{n_1^-} \log_2 \left(\frac{n_{1,3}^-}{n_1^-} \right)$$

$$I_1^- = -\frac{3}{8} \log_2 \left(\frac{3}{8} \right) - \frac{5}{8} \log_2 \left(\frac{5}{8} \right)$$

$$I_1^- = 0,9544$$

Przykład – dobór formy reklamy

Łączna entropia po ocenie warunku j :

$$E_1 = \frac{n_1^+}{n} I_1^+ + \frac{n_1^-}{n} I_1^-$$

$$E_1 = \frac{4}{12} * 1 + \frac{8}{12} * 0,9544$$

$$E_1 = 0,9696 \quad I - E_1 = 0,4147$$

Przykład – dobór formy reklamy

j	$I-E_j$
1	0,4147
2	0,1852
3	0,1140
4	0,0290
5	0,0290
6	0,6548
7	0,6548

Przykład – dobór formy reklamy

wiek	<20	1	1	0	0	0	0
	20-30	0	0	1	0	1	0
	>30	0	0	0	1	0	1
płeć	K	1	0	0	0	1	1
	M	0	1	1	1	0	0
mieszka	wieś	0	0	0	0	0	0
	miasto	1	1	1	1	1	1
reklama	Internet	1	1				
	Prasa			1	1	1	
	Telewizja						1

wiek	<20	1	1	0	0	0	0
	20-30	0	0	1	1	0	0
	>30	0	0	0	0	1	1
płeć	K	0	1	1	0	1	0
	M	1	0	0	1	0	1
mieszka	wieś	1	1	1	1	1	1
	miasto	0	0	0	0	0	0
reklama	Internet						
	Prasa						
	Telewizja	1	1	1	1	1	1

Przykład – dobór formy reklamy

j	$I-E_j$
1	0,8249
2	0,3658
3	0,3658
4	0,1142
5	0,1142

Drzewo decyzyjne



Algorytm ID3 przy ciągłych wartościach cech

- ▶ Załóżmy, że cechy obiektów przyjmują wartości z pewnych ciągłych przedziałów, wówczas

$$O = \{o_1, \dots, o_i, \dots, o_n\}$$

$$A = \{a_1, \dots, a_i, \dots, a_m\}$$

$$K = \{k_1, \dots, k_i, \dots, k_l\}$$

gdzie

O — zbiór obiektów (przykładów)

A — zbiór atrybutów identyczny dla wszystkich przykładów

K — zbiór klas, do których kwalifikujemy przykłady

Algorytm ID3 przy ciągłych wartościach cech

$$(d_{i,1} = w_{i,1}) \wedge \dots \wedge (d_{i,j} = w_{i,j}) \wedge \dots \wedge (d_{i,m} = w_{i,m})$$

$$\Rightarrow (c_i = k_l) \quad i$$

$$w_{i,j} \in (w \min_j; w \max_j)$$

gdzie

$d_{i,j}$ — wartość atrybutu j w przykładzie i

c_i — numer klasy do której należy i -ty przykład

Algorytm ID3 przy ciągłych wartościach cech

- ▶ Ponieważ dla takiego zapisu nie można wprost wykorzystać metody ID3 należy wprowadzić dodatkowo dla kolejnych atrybutów wartości w^* dzielące dziedzinę na dwa rozłączne podzbiory
- ▶ należy tak przekształcać warunki i przykłady by możliwy był następujący zapis:

$$(d_{i,1} < w_1^*) \wedge \dots \wedge (d_{i,j} < w_j^*) \wedge \dots \wedge (d_{i,m} < w_m^*) \\ \Rightarrow (c_i = k_l)$$

Przykład

- ▶ Należy określić zależność wielkości sprzedaży od wieku klienta, poziomu wykształcenia oraz odległości od sklepu

wiek	wykształcenie	odległość	wartość zakupów	
lata	poziom	m	zł/m-c	klasa
18	2	200	100	1
20	3	500	50	1
35	3	100	400	3
21	4	600	200	2
40	5	400	550	3

Wybór punktu podziału

- ▶ Badamy jaki jest poziom entropii po uwzględnieniu wieku klienta. W tym celu wybieramy taki punkt podziału w^* ze zbioru $\{20, 21, 35, 40\}$, który wprowadza najwięcej informacji. Uzyskujemy następujące tabele:

Wybór punktu podziału wg wieku

wiek	<20	1	0	0	0	0
	>=20	0	1	1	1	1
klasa	1	1	1	0	0	0
	2	0	0	1	0	0
	3	0	0	0	1	1

wiek	<21	1	1	0	0	0
	>=21	0	0	1	1	1
klasa	1	1	1	0	0	0
	2	0	0	1	0	0
	3	0	0	0	1	1

wiek	<35	1	1	1	0	0
	>=35	0	0	0	1	1
klasa	1	1	1	0	0	0
	2	0	0	1	0	0
	3	0	0	0	1	1

wiek	<40	1	1	1	1	0
	>=40	0	0	0	0	1
klasa	1	1	1	0	0	0
	2	0	0	1	0	0
	3	0	0	0	1	1

Wybór punktu podziału wg wieku

w^*	$I-E_j$
20	0,5219
21	1,063
35	1,063
40	0,5219

Wybór punktu podziału wg wykształcenia

w^*	$I-E_j$
3	0,5219
4	0,7294
5	0,5219

Wybór punktu podziału wg odległości

w^*	$I-E_j$
200	0,5219
400	0,3961
500	0,7294
600	0,8553

Koniec pierwszego etapu

- ▶ Jeżeli za czynnik decydujący w pierwszym etapie o podziale przypadków wiek < 35 wówczas uzyskamy następujące podzbiory

wiek	wykształcenie	odległość	wartość zakupów	
lata	poziom	m	zł/m-c	klasa
18	2	200	100	1
20	3	500	50	1
21	4	600	200	2

wiek	wykształcenie	odległość	wartość zakupów	
lata	poziom	m	zł/m-c	klasa
35	3	100	400	3
40	5	400	550	3

Kontynuacja

- ▶ Przedstawione procedury powtarzamy dla każdego z podzbiorów aż do pełnego rozjaśnienia problemu
- ▶ UWAGA: metoda nie dopuszcza przykładów sprzecznych

Generowanie pokryć - przykład

- ▶ Wybieramy podzbiór P obiektów należących do klasy k i podzbiór M obiektów nie należących do danej klasy
- ▶ Z podzbioru P wybieramy dowolny przykład x^r
- ▶ Ustalamy różnice w warunkach pomiędzy wybranym przykładem a wszystkimi przykładami z podzbioru M

Przykład – dobór formy reklamy

wiek	<20	1	1	0	0	0	1	1	0	0	0	0	0
	20-30	0	0	1	0	1	0	0	1	1	0	0	0
	>30	0	0	0	1	0	0	0	0	0	1	1	1
płeć	K	1	0	0	0	1	0	1	1	0	1	1	0
	M	0	1	1	1	0	1	0	0	1	0	0	1
mieszka	wieś	0	0	0	0	0	1	1	1	1	1	0	1
	miasto	1	1	1	1	1	0	0	0	0	0	1	0
reklama	Internet	1	1										
	Prasa			1	1	1							
	Telewizja						1	1	1	1	1	1	1

M

P

Generowanie pokryć - przykład

- ▶ Podzbiór **P** – wszystkie przykłady należące do klasy **telewizja**

$$x^r = [\text{wiek} < 20] \wedge [\text{płeć} = M] \\ \wedge [\text{mieszka} = \text{wieś}]$$

- ▶ Ustalamy różnice

$$dr_1 = [\text{płeć} = M] \wedge [\text{mieszka} = \text{wieś}]$$

$$dr_2 = [\text{mieszka} = \text{wieś}]$$

$$dr_3 = [\text{wiek} < 20] \wedge [\text{mieszka} = \text{wieś}]$$

$$dr_4 = [\text{wiek} < 20] \wedge [\text{mieszka} = \text{wieś}]$$

$$dr_5 = [\text{wiek} < 20] \wedge [\text{płeć} = M]$$

Generowanie pokryć - przykład

- ▶ Generujemy pokrycia wybierając po jednym warunku z każdej różnicy łącząc je każdy z każdym:

z dr_1 wybieramy $[płeć=M]$

z dr_2 - $[mieszka=wieś]$

z dr_3 - $[wiek<20]$

- ▶ ponieważ w kolejnych różnicach nie ma już różnych warunków uzyskujemy pokrycie:

$$C_1 = [płeć=M] \wedge [mieszka=wieś] \wedge [wiek<20]$$

Generowanie pokryć - przykład

- ▶ Koniunkcja warunków w pokryciu wskazuje na przykład, który na pewno nie należy do zbioru **M**, a należy do zbioru **P**.
- ▶ Dalej:
 - z dr_1 wybieramy $[mieszka=wieś]$
 - z dr_3 - $[wiek < 20]$
 - z dr_5 – nie wybieramy $[płeć=M]$ bo uzyskalibyśmy pokrycie C_1
- ▶ ponieważ w kolejnych różnicach nie ma już różnych warunków uzyskujemy pokrycie:
 $C_2 = [mieszka=wieś] \wedge [wiek < 20]$

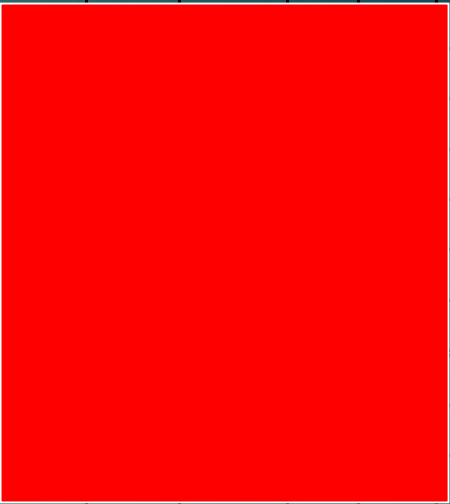

Generowanie pokryć - przykład

- ▶ Pokrycie C_2 informuje nas o tym, że wszystkie osoby mieszkające na wsi młodsze niż 20 lat na pewno nie preferują telewizji jako medium reklamowego.
- ▶ Zauważmy, że z każdej różnicy możemy wybrać warunek $[mieszka=wieś]$
- ▶ Uzyskujemy pokrycie:
 $C_3 = [mieszka=wieś]$
- ▶ co oznacza, że spełnienie tego warunku wyklucza przynależność do badanej klasy.

Generowanie pokryć - przykład

- ▶ Wybieramy najkorzystniejsze pokrycie - C_3 do lewej strony reguły dopisując warunki C_3
- ▶ uzyskujemy regułę
- ▶ $[mieszka=wieś] \Rightarrow [reklama =Telewizja]$
- ▶ Ze zbioru P usuwamy przykłady zgodne z regułą C_3

Przykład – dobór formy reklamy

wiek	<20	1	1	0	0	0		0	
	20-30	0	0	1	0	1		0	
	>30	0	0	0	1	0		1	
płeć	K	1	0	0	0	1		1	
	M	0	1	1	1	0		0	
mieszka	wieś	0	0	0	0	0		0	
	miasto	1	1	1	1	1		1	
reklama	Internet	1	1						
	Prasa			1	1	1			
	Telewizja							1	

Generowanie pokryć - przykład

- ▶ Podzbiór **P** – wszystkie przykłady należące do klasy **telewizja**

$x^r = [\text{wiek} > 30] \wedge [\text{płeć} = K] \wedge [\text{mieszka} = \text{miasto}]$

- ▶ Ustalamy różnice

$dr_1 = [\text{wiek} > 30]$

$dr_2 = [\text{wiek} > 30] \wedge [\text{płeć} = K]$

$dr_3 = [\text{wiek} > 30] \wedge [\text{płeć} = K]$

$dr_4 = [\text{płeć} = K]$

$dr_5 = [\text{wiek} > 30]$

Generowanie pokryć - przykład

- ▶ Generujemy pokrycia
- ▶ $C_1 = [\text{wiek} > 30] \wedge [\text{płeć} = K]$
- ▶ Wybieramy pokrycie – C_1
- ▶ Ze zbioru P usuwamy przykłady zgodne z regułą C_1
- ▶ uzyskujemy regułę
- ▶ $[\text{mieszka} = \text{wieś}] \wedge [\text{wiek} > 30] \wedge [\text{płeć} = K] \Rightarrow [\text{reklama} = \text{Telewizja}]$
- ▶ Kontynuujemy działania aż do rozróżnienia wszystkich przykładów