Indukcja reguł

- Kompleks k składa się z selektorów.
- $k_1 = \{<$ słoneczna $\vee$ deszczowa, zimna $\vee$ ciepła, $?, ? >\}$
  $k_2 = \{<$ słoneczna, ciepła, $?, ? >\}$
  $k_2 \prec k_1$
  k_2 jest bardziej szczegółowe od k_1, k_1 jest bardziej ogólne od k_2
- $S \rhd k$ to dokładniej $(\exists k \in S) k \rhd x$ - zbiór wszystkich x pokrywanych przez $k \in S$
- $\{k_1 \rhd x\} = \{1, 2, 5, 6, 9\}$
- $\{k_2 \rhd x\} = \{1, 2\}$

Indukcja reguł - sekwencyjne pokrywanie

funkcja *sekwencyjne-pokrywanie(T)*

argumenty wejściowe:

- T - zbiór trenujący dla pojęcia c

zwraca: zbiór reguł reprezentujący hipotezę przybliżającą c

$R := 0; P := T;$

jak długo $P \neq 0$ wykonaj

$k :=$ znajdź-kompleks(T, P);

$d :=$ kategoria(k, T, P);

$R := R \cup \{k \rightarrow d\}$;

$P := P - P_k$;

koniec jak długo

zwróć R

Indukcja reguł - algorytm CN2

funkcja *znajdź-kompleks-cn2(T, P)*
argumenty wejściowe:

- T - zbiór trenujący dla pojęcia c,
- P - podzbiór zbioru T zawierający przykłady nie pokryte przez wygenerowane wcześniej reguły

zwraca: statystycznie istotny kompleks pokrywający pewną liczbę przykładów z P z dużą dokładnością;

$S := \{<?>\}$;$k_* :=<?>$;
jak długo $S \neq \phi$ wykonaj
$S' := S \cap \mathbb{S}$;
$S' := S' - S - \{<\phi>\}$;
dla wszystkich kompleksów $k \in S'$ wykonaj
jeśli $\psi_k(P) > \theta \wedge \vartheta_k(P) > \vartheta_{k_*}(P)$ to $k_* := k$
koniec jeśli
koniec dla
$S :=$ Arg $\max_{k \in S'}^{m} v_k(P)$
koniec jak długo
zwróć k_*

Algorytm CN2 - funkcja oceniająca kompleksy

Entropię zbioru P ze względu na kompleks k określa się następująco:

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

Entropia ma tę cechę, że największą wartość przyjmuje dla zrównoważonych rozkładów częstości kategorii. Funkcja oceniająca kompleksy musi być zanegowaną entropią:

$$\vartheta_k(P) = -E_k(P)$$

Algorytm CN2 - statystyka χ

Niech f_i oznacza *zaobserwowaną częstość* (liczbę wystąpień) i-tej wartości atrybutu y_i dla $i = 1, 2, 3, \ldots, v_1$ i odpowiednio f_j dla y_j dla $j = 1, 2, 3, \ldots, v_2$, f_{ij} liczbę (częstość) jednoczesnych wystąpień i-tej i j-tej wartości atrybutów y_i i y_j, a e_{ij} to wartość oczekiwana jednoczesnego wystąpienia przy założeniu niezależności y_1 i y_2 i $(v_1 - 1)(v_2 - 1)$ stopniach swobody.

$$\chi^2 = \sum_{i=1}^{v_1} \sum_{j=1}^{v_2} \frac{(f_{ij} - e_{ij})^2}{e_{ij}},$$

$$\text{gdzie } e_{ij} = \frac{f_i^1 f_j^2}{n}$$

Im większa wartość statystyki tym bardziej atrybuty są zależne od siebie.

Algorytm CN2 - statystyka χ

$$\chi_k^2(P) = \sum_{d \in C} \frac{(|P_k^d| - e_k^d(P))^2}{e_k^d(P)},$$

gdzie $e_k^d(P) = |P_k| \frac{|P^d|}{|P|}$

x	*aura*	*temperatura*	*wilgotność*	*wiatr*	$c(x)$
1	*słoneczna*	*ciepła*	*duża*	*słaby*	*0*
2	*słoneczna*	*ciepła*	*duża*	*silny*	*0*
3	*pochmurna*	*ciepła*	*duża*	*słaby*	*1*
4	*deszczowa*	*umiarkowana*	*duża*	*słaby*	*1*
5	*deszczowa*	*zimna*	*normalna*	*słaby*	*1*
6	*deszczowa*	*zimna*	*normalna*	*silny*	*0*
7	*pochmurna*	*zimna*	*normalna*	*silny*	*1*
8	*słoneczna*	*umiarkowana*	*duża*	*słaby*	*0*
9	*słoneczna*	*zimna*	*normalna*	*słaby*	*1*
10	*deszczowa*	*umiarkowana*	*normalna*	*słaby*	*1*
11	*słoneczna*	*umiarkowana*	*normalna*	*silny*	*1*
12	*pochmurna*	*umiarkowana*	*duża*	*silny*	*1*
13	*pochmurna*	*ciepła*	*normalna*	*słaby*	*1*
14	*deszczowa*	*umiarkowana*	*duża*	*silny*	*0*

Zbiór $\mathbb{S}$ kompleksów atomowych

$\mathbb{S} = \{<$ deszczowa, ?, ?, ? $>$,

$<$ deszczowa $\vee$ słoneczna, ?, ?, ? $>$,

$<$ deszczowa $\vee$ pochmurna, ?, ?, ? $>$,

$<$ pochmurna, ?, ?, ? $>$,

$<$ pochmurna $\vee$ słoneczna, ?, ?, ? $>$,

$<$ słoneczna, ?, ?, ? $>$,

$<$?, ciepła, ?, ? $>$,

$<$?, ciepła $\vee$ zimna, ?, ? $>$,

$<$?, ciepła $\vee$ umiarkowana, ?, ? $>$,

$<$?, umiarkowana, ?, ? $>$,

$<$?, umiarkowana $\vee$ zimna, ?, ? $>$,

$<$?, zimna, ?, ? $>$,

$<$?, ?, duża, ? $>$, $<$?, ?, normalna, ? $>$, $<$?, ?, ?, silny $>$, $<$?, ?, ?, słaby $>\}$

Kolejne kroki algorytmu CN2 1/3

1. Początkowo $R = \phi, P = T = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}, \mathbb{S}$
2. Następuje wywołanie *znajdź-kompleks*(T, P).
   - $S = \{<?>\} \neq \phi, k_* =<?>$ i $\vartheta_{k_*}(P) = -E_{k_*}(P) = -0.940$,
   - $S' = \mathbb{S} = S \cap \mathbb{S}$,
   - $k =<$ pochmurna, $?, ?, ?>$ ma największą wartość $\vartheta_k = 0$ w zbiorze $\mathbb{S}$; $S = \{k\}, k_* = k$,
3. $R = \{<$ pochmurna, $?, ?, ?> \rightarrow 1\}$, $P = \{1, 2, 4, 5, 6, 8, 9, 10, 11, 14\}$,
4. $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),
   - $S = \{<?>\} \neq \phi, k_* =<?>$ i $\vartheta_{k_*}(P) = -1$,
   - $S' = \mathbb{S} = S \cap \mathbb{S}$,
   - $k =<?,$ ciepła, $?, ?>$ ma największą wartość $\vartheta_k = 0$ w zbiorze $\mathbb{S}$; $S = \{k\} \neq \phi, k_* = k$,
5. $R = \{<$ pochmurna, $?, ?, ?> \rightarrow 1, <?,$ ciepła, $?, ?> \rightarrow 0\}$, $P = \{4, 5, 6, 8, 9, 10, 11, 14\}$,

Kolejne kroki algorytmu CN2 2/3

1. $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),
   - $S' = \mathbb{S} = S \cap \mathbb{S}$,
   - $k =<?, ?, \text{normalna}, ? >$ zostaje wybrane z najwyższą wartością $\vartheta_k = -0,721$ w zbiorze $\mathbb{S}$; $S = \{k\} \neq \phi, k_* = k$,
   - k_* nie ma wartości 0 (pętla jak długo się nie kończy),
   - w następnym cyklu dla $S' = S \cap \mathbb{S}$ największą wartość $\vartheta_k = 0$ ma kompleks $k =<?, ?, \text{normalna}, \text{słaby} >$, $k_* = k$
2. $R = \{<$ pochmurna$, ?, ?, ? > \rightarrow 1, <?,$ ciepła$, ?, ? > \rightarrow 0, <?, ?,$ normalna$,$ słaby $> \rightarrow 1\}$, $P = \{4, 6, 8, 11, 14\}$,
3. po kilku dalszych wywołaniach funkcji *znajdź-kompleks*(T, P) otrzymujemy $R = \{<$ pochmurna$, ?, ?, ? > \rightarrow 1, <?,$ ciepła$, ?, ? > \rightarrow 0, <?, ?,$ normalna$,$ słaby $> \rightarrow 1, <?,$ zimna$, ?, ? > \rightarrow 0, <?, ?,$ normalna$, ? > \rightarrow 1, <?, ?, ?,$ silny $> \rightarrow 0, <$ słoneczna$, ?, ?, ? > \rightarrow 0\}, P = \{4\}$

Kolejne kroki algorytmu CN2 3/3

1. $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),
   - $S = \{<?>\} \neq \phi, k_* = <?>$ i $\vartheta_{k_*}(P) = -E_{k_*}(P) = 0$,
2. Ostatecznie

   $R = \{<$ pochmurna, $?, ?, ?> \rightarrow 1,$

   $<?,$ ciepła, $?, ?> \rightarrow 0,$

   $<?, ?,$ normalna, słaby $> \rightarrow 1,$

   $<?,$ zimna, $?, ?> \rightarrow 0,$

   $<?, ?,$ normalna, $?> \rightarrow 1,$

   $<?, ?, ?,$ silny $> \rightarrow 0,$

   $<$ słoneczna, $?, ?, ?> \rightarrow 0,$

   $<?> \rightarrow 1\}$