GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych

KLASTERYZACJA I SEGMENTACJA

Wojciech Waloszek

wowal@eti.pg.gda.pl

Teresa Zawadzka

tegra@eti.pg.gda.pl

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

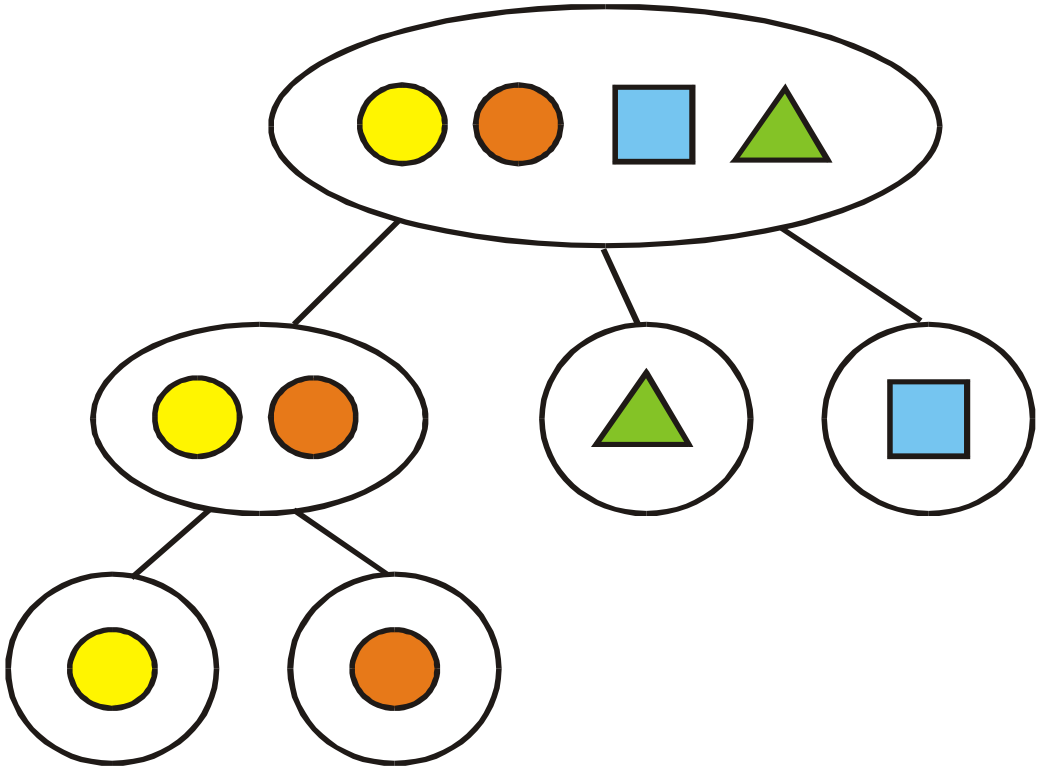
GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB

- Klasteryzacja polega na wygenerowaniu pewnych zasad podziału danych na zasadzie podobieństwa,
- Algorytm COBWEB jest przykładem algorytmu grupującego (tworzącego klastry),
- Algorytm COBWEB generuje hierarchiczny podział danych w postaci drzewa.

GDAŃSK UNIVERSITY OF TECHNOLOGY

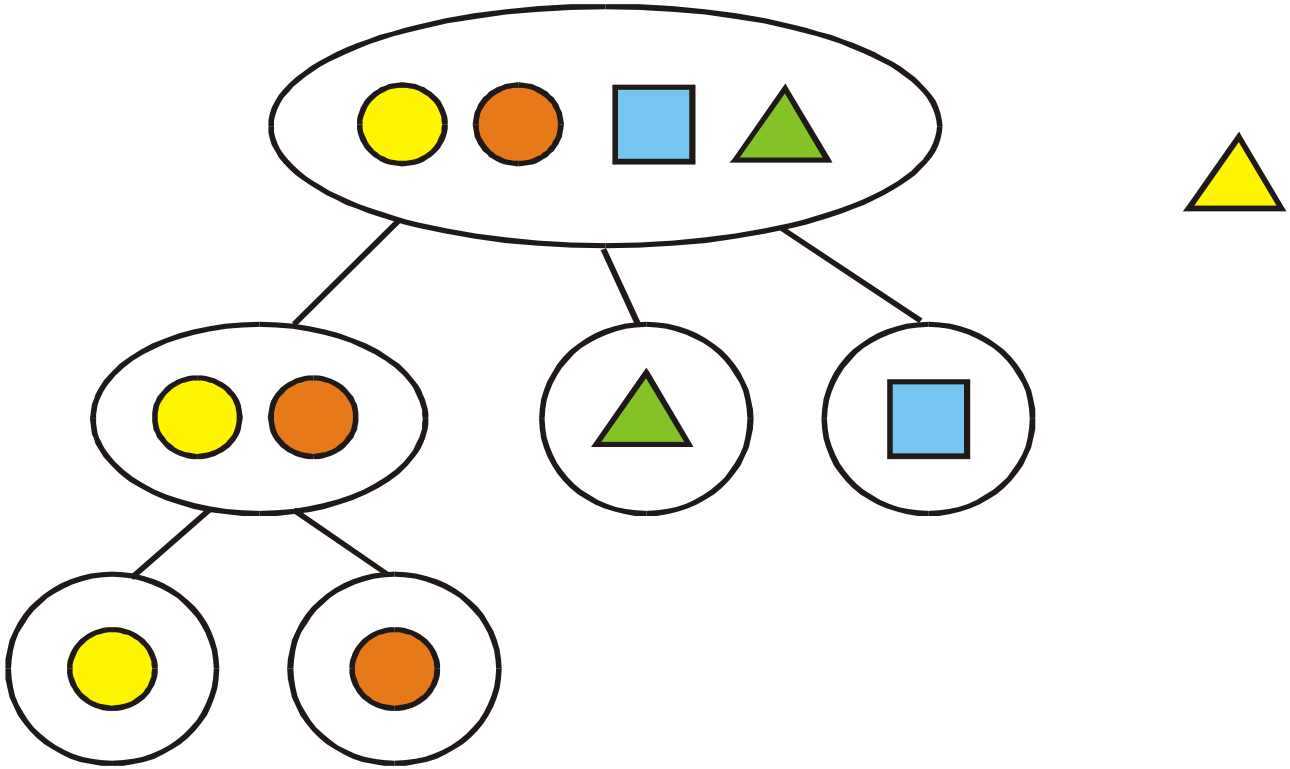
Drzewo COBWEB



GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – grupowanie

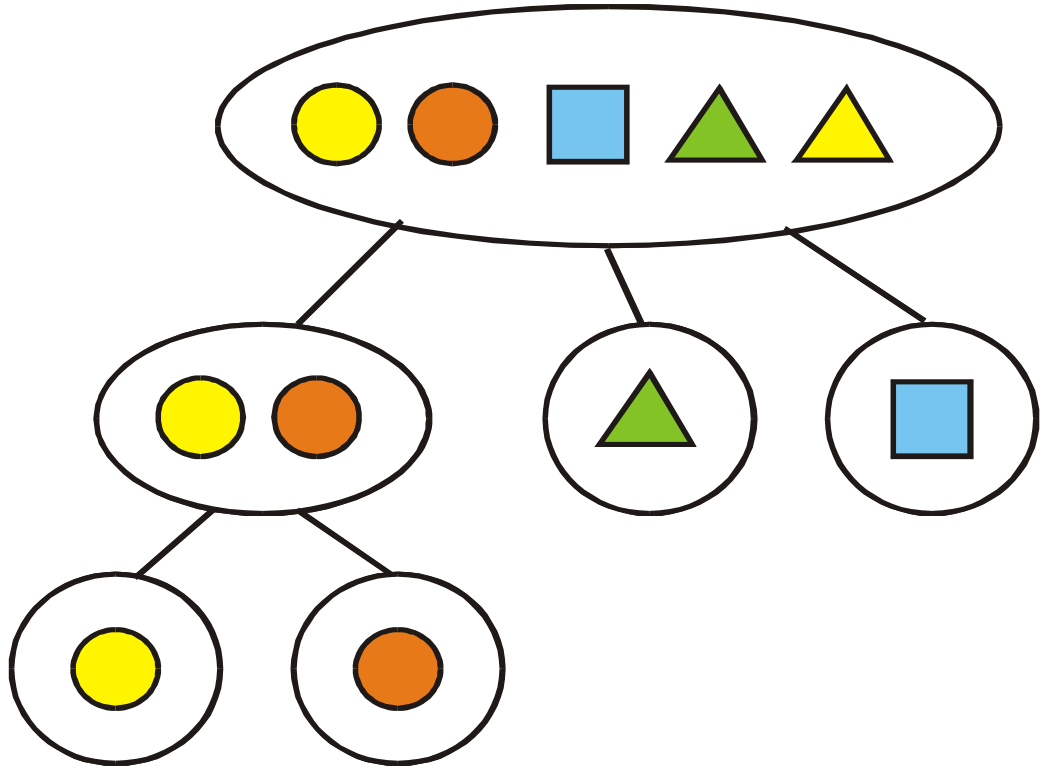
Załóżmy, że chcemy przeprowadzić proces grupowania dla kolejnego przykładu:



GDAŃSK UNIVERSITY OF TECHNOLOGY

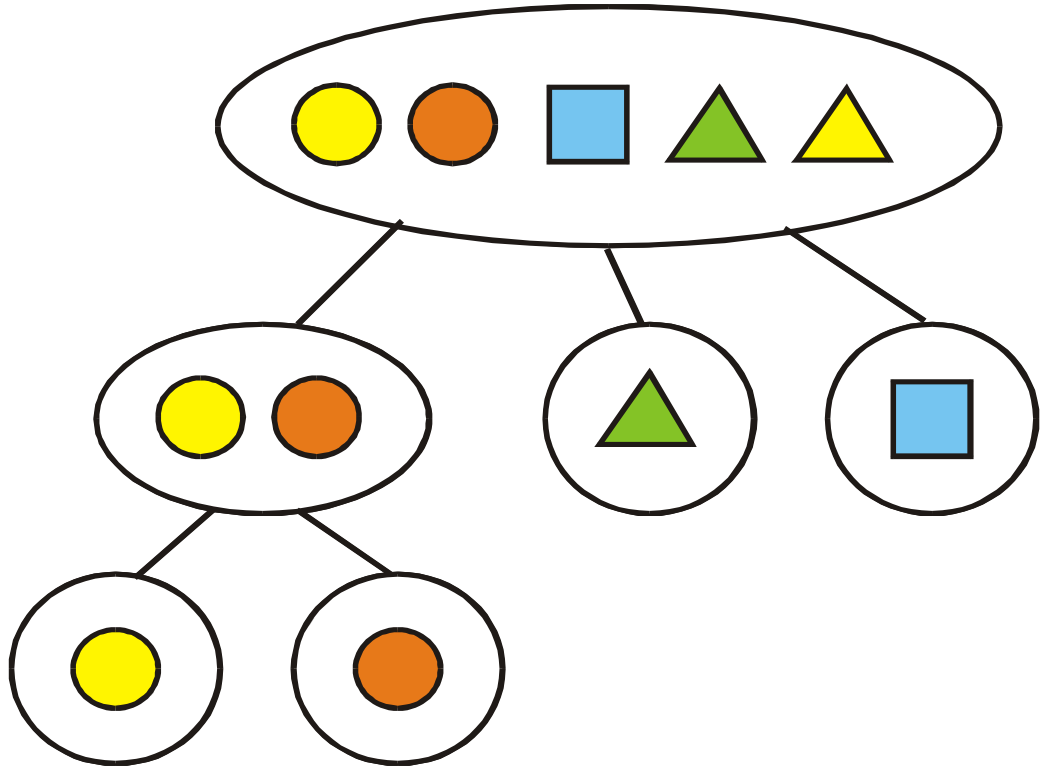
COBWEB – grupowanie

COBWEB dokłada przykład do głównej kategorii.



COBWEB – grupowanie

Następnie próbuje różnych możliwości dalszego grupowania.



GDAŃSK UNIVERSITY OF TECHNOLOGY

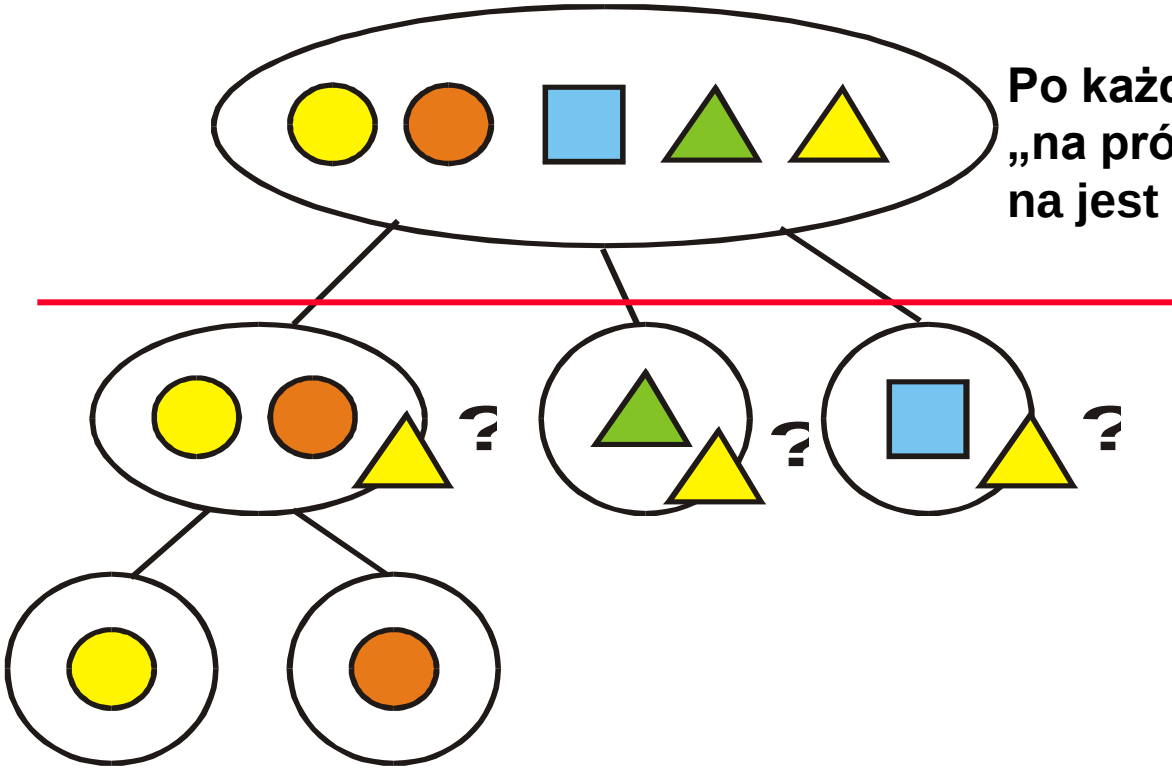
COBWEB – operacje

- Na każdym z poziomów drzewa algorytm może wykonać jedną z operacji:
  - zaliczyć przykład do istniejącej podkategorii,
  - utworzyć osobną podkategorię dla przykładu,
  - dokonać podziału podkategorii i zaliczyć przykład do jednej z jej kategorii potomnych,
  - połączyć dwie podkategorie i zaliczyć przykład do podkategorii połączonej,
- O tym, którą operację wykonać, decyduje wartość funkcji oceny jakości grupowania.

GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – grupowanie

Algorytm dokłada „na próbę” przykład do każdej podkategorii.



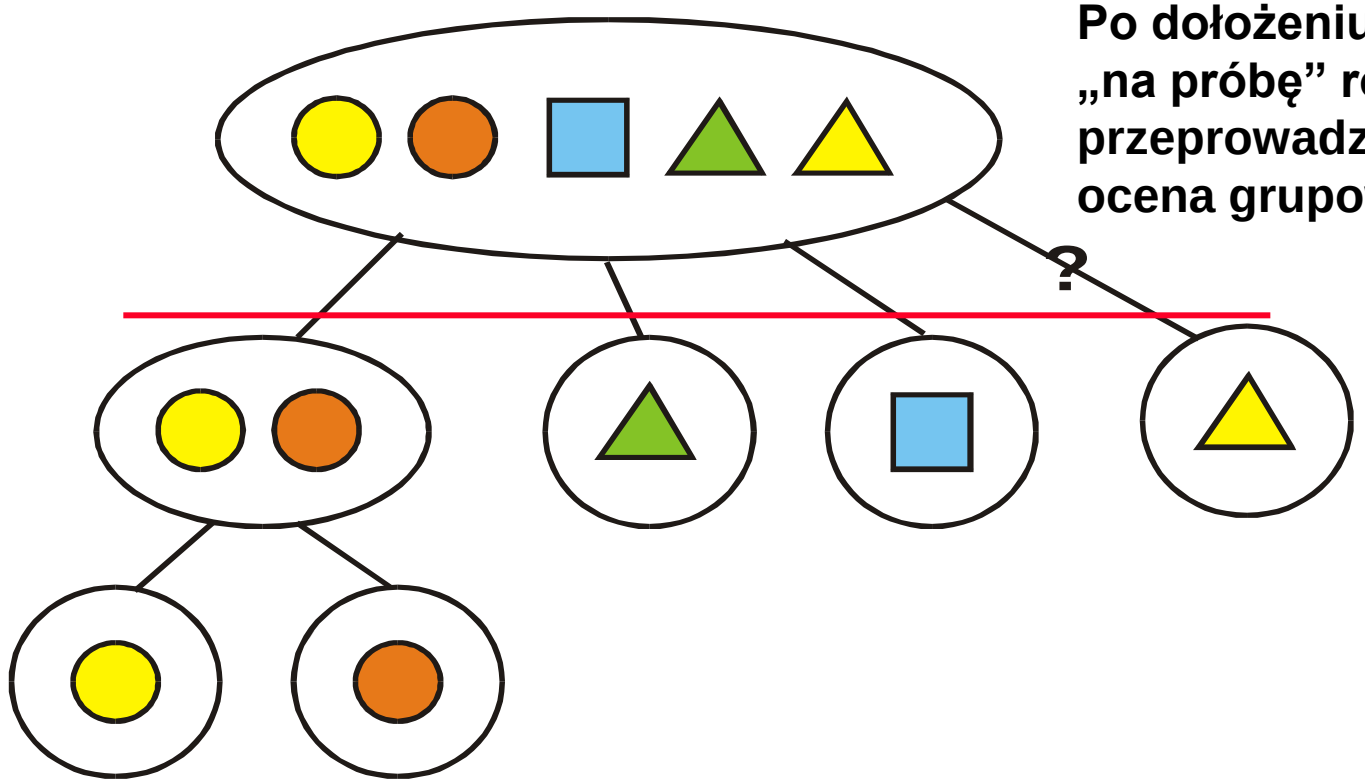
Po każdym dołożeniu „na próbę” przeprowadzana jest ocena grupowania

0,362

GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – grupowanie

Algorytm próbuje też stworzyć dla przykładu osobną podkategorię.

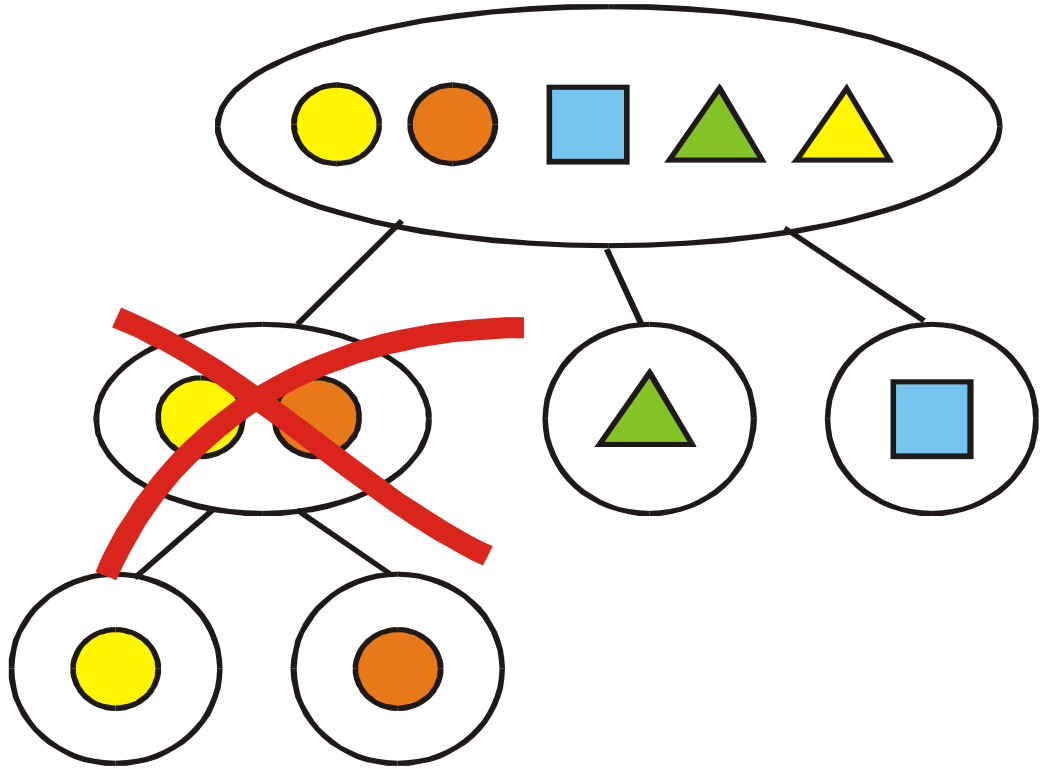


Po dołożeniu kategorii „na próbę” również przeprowadzana jest ocena grupowania

GDAŃSK UNIVERSITY OF TECHNOLOGY

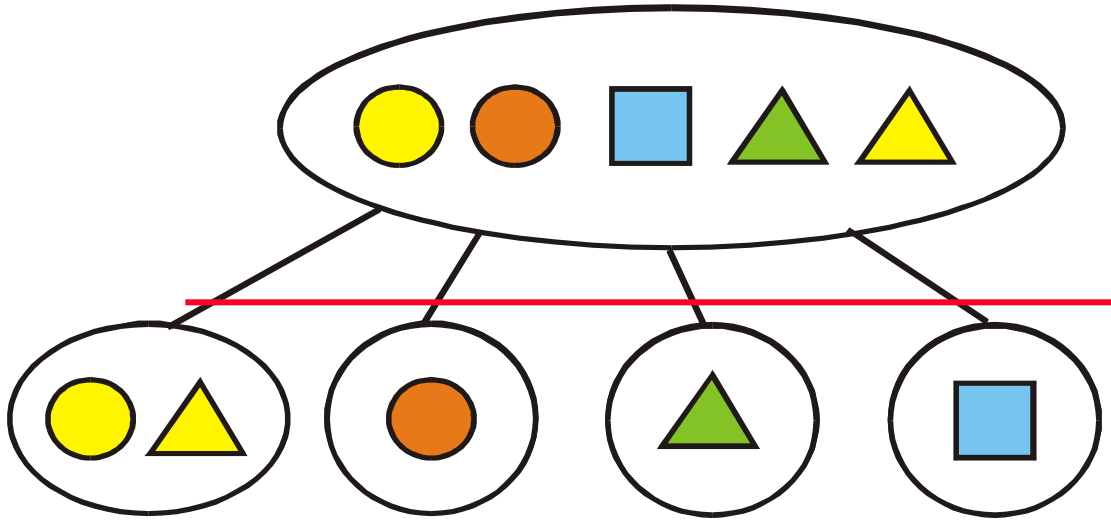
COBWEB – podział

COBWEB próbuje też usunąć jedną z podkategorii
i dodać przykład do jednej z jej kategorii potomnych.



COBWEB – podział

COBWEB próbuje też usunąć jedną z podkategorii
i dodać przykład do jednej z jej kategorii potomnych.

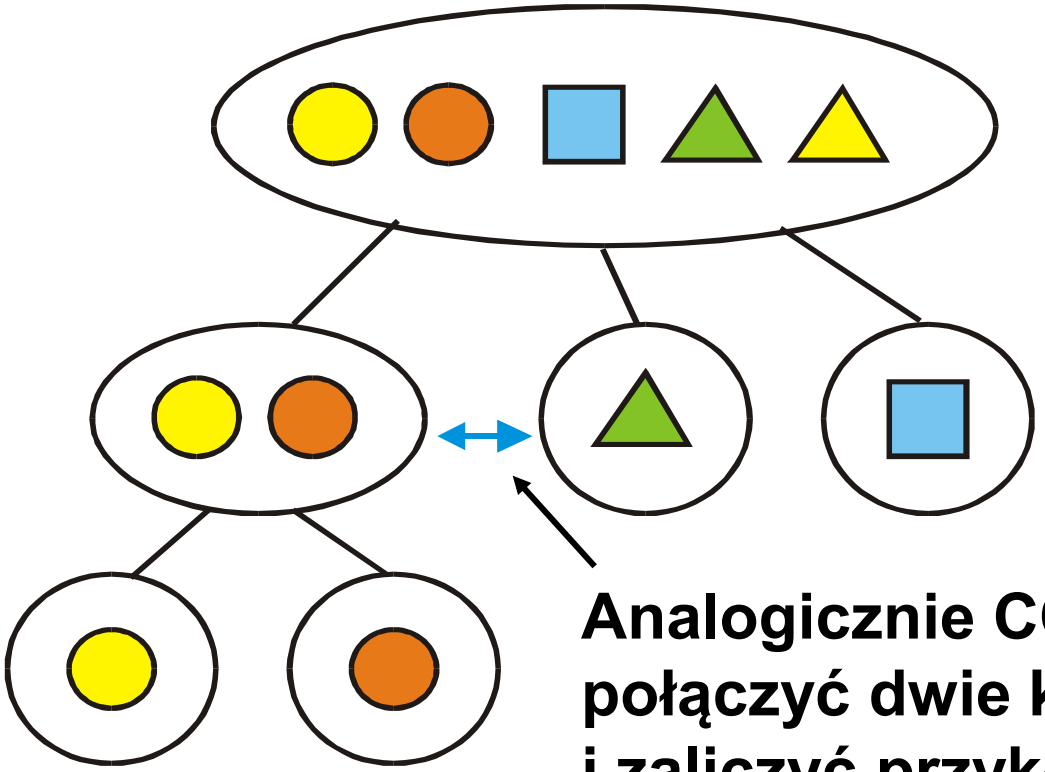


Oczywiście po tych posunięciach (nadal wykonywanych „na próbę”) obliczamy ocenę grupowania

0,326

GDAŃSK UNIVERSITY OF TECHNOLOGY

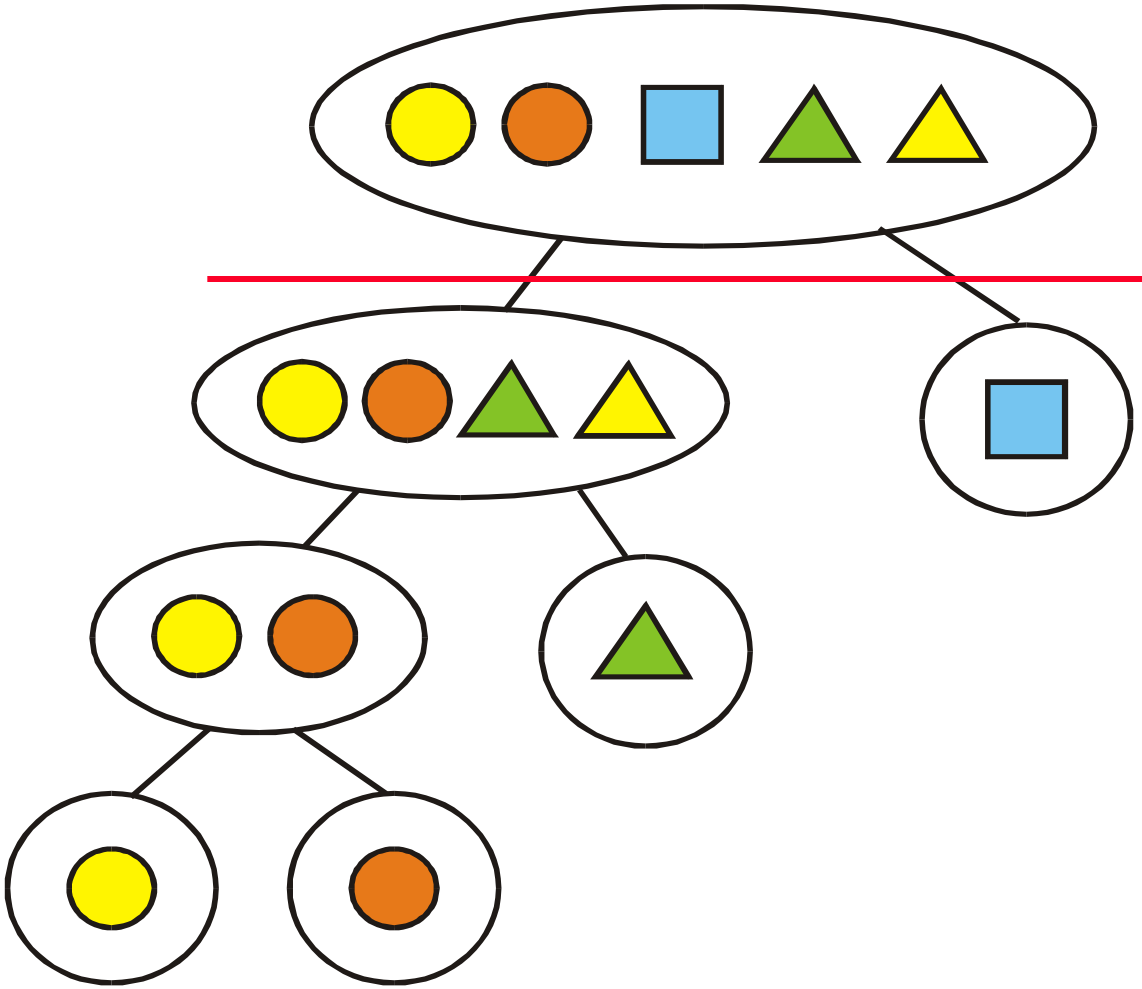
COBWEB – scalenie



Analogicznie COBWEB może połączyć dwie kategorie i zaliczyć przykład do kategorii połączonej

GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – scalenie

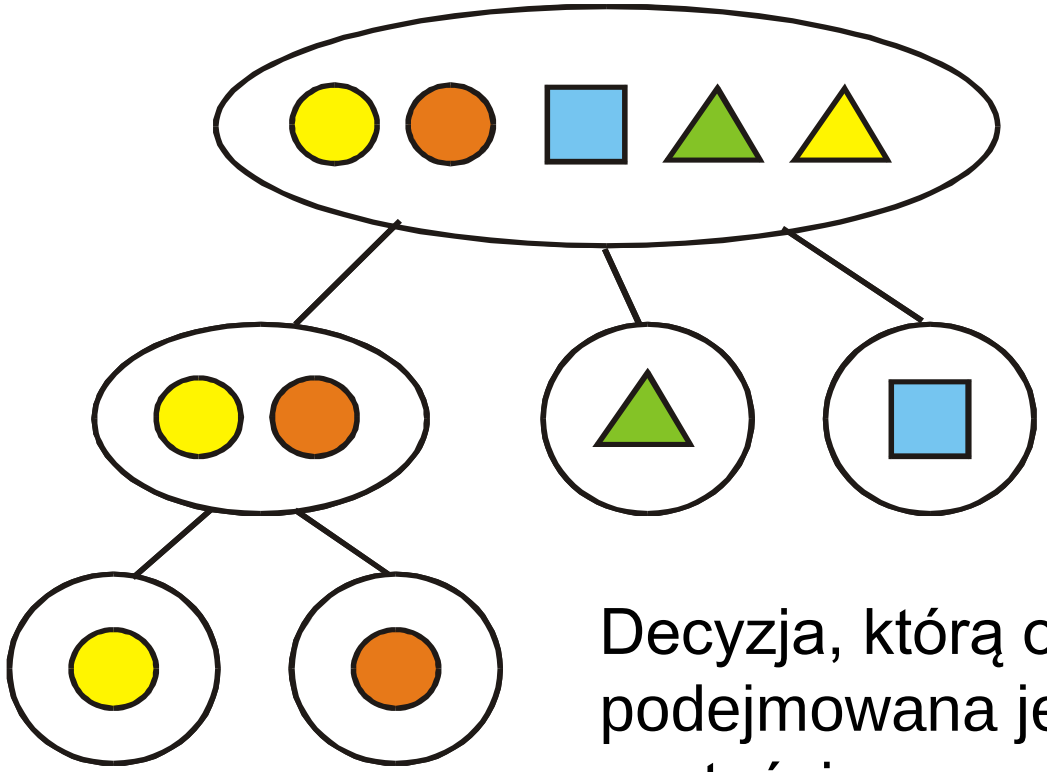


0,428

Po tym, ostatnim już, posunięciu „na próbę” obliczamy ocenę grupowania

GDAŃSK UNIVERSITY OF TECHNOLOGY

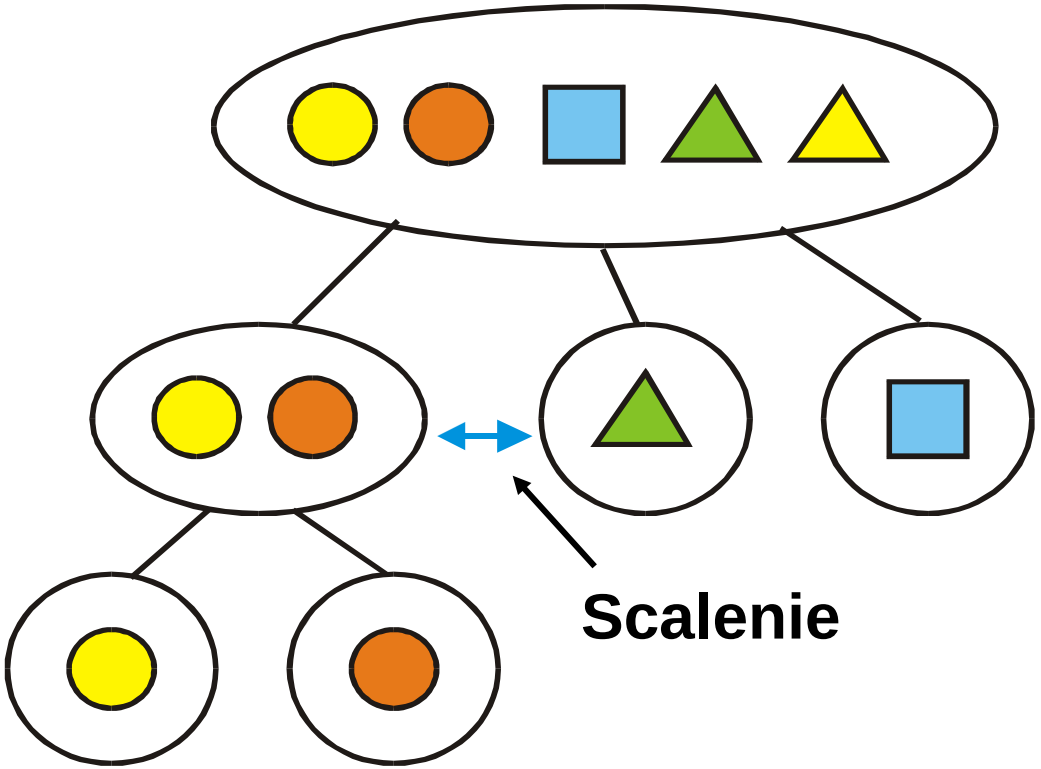
COBWEB – przykład



Decyzja, którą operację wykonać podejmowana jest na podstawie wartości oceny grupowania na każdym poziomie drzewa.

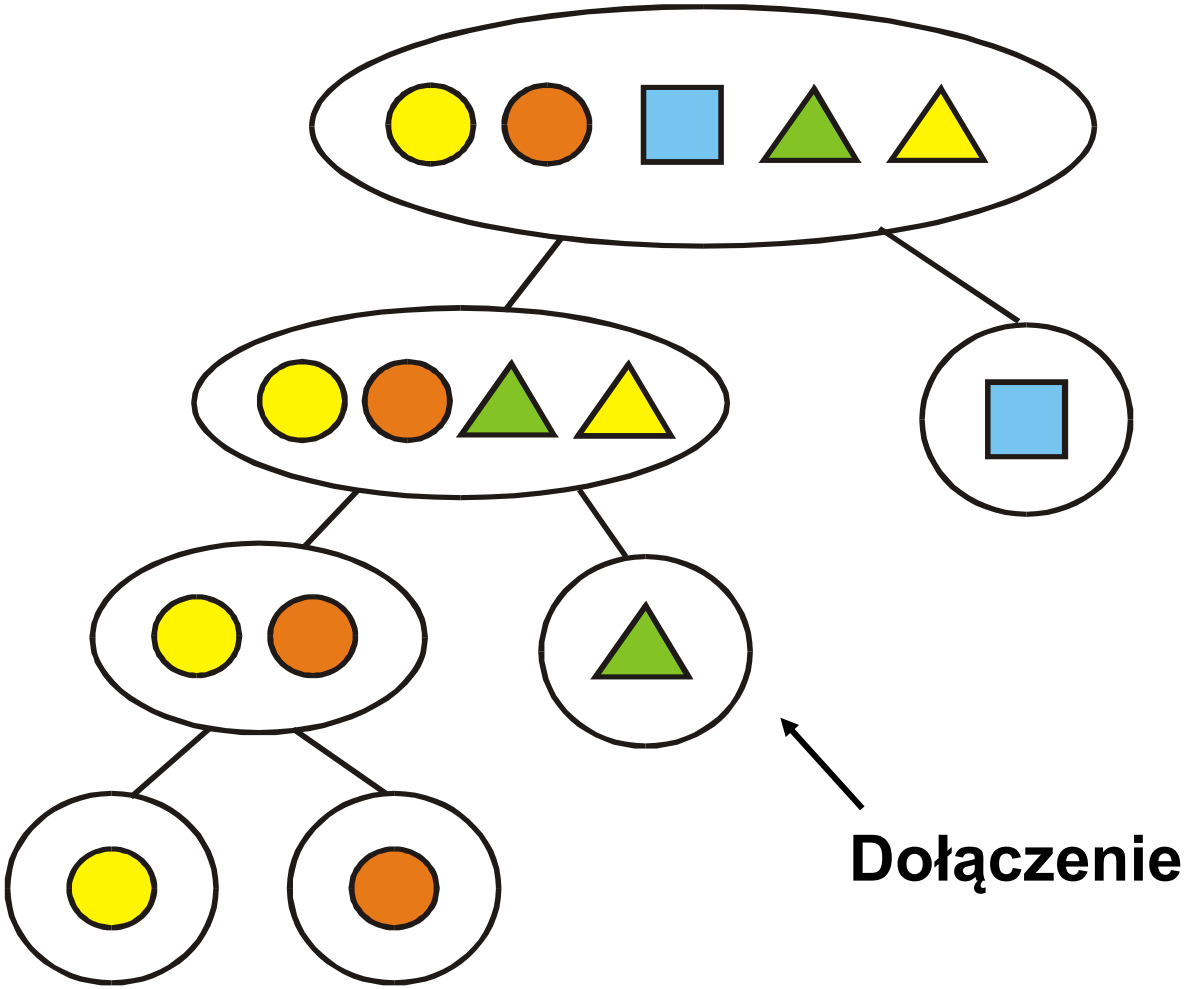
GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – przykład



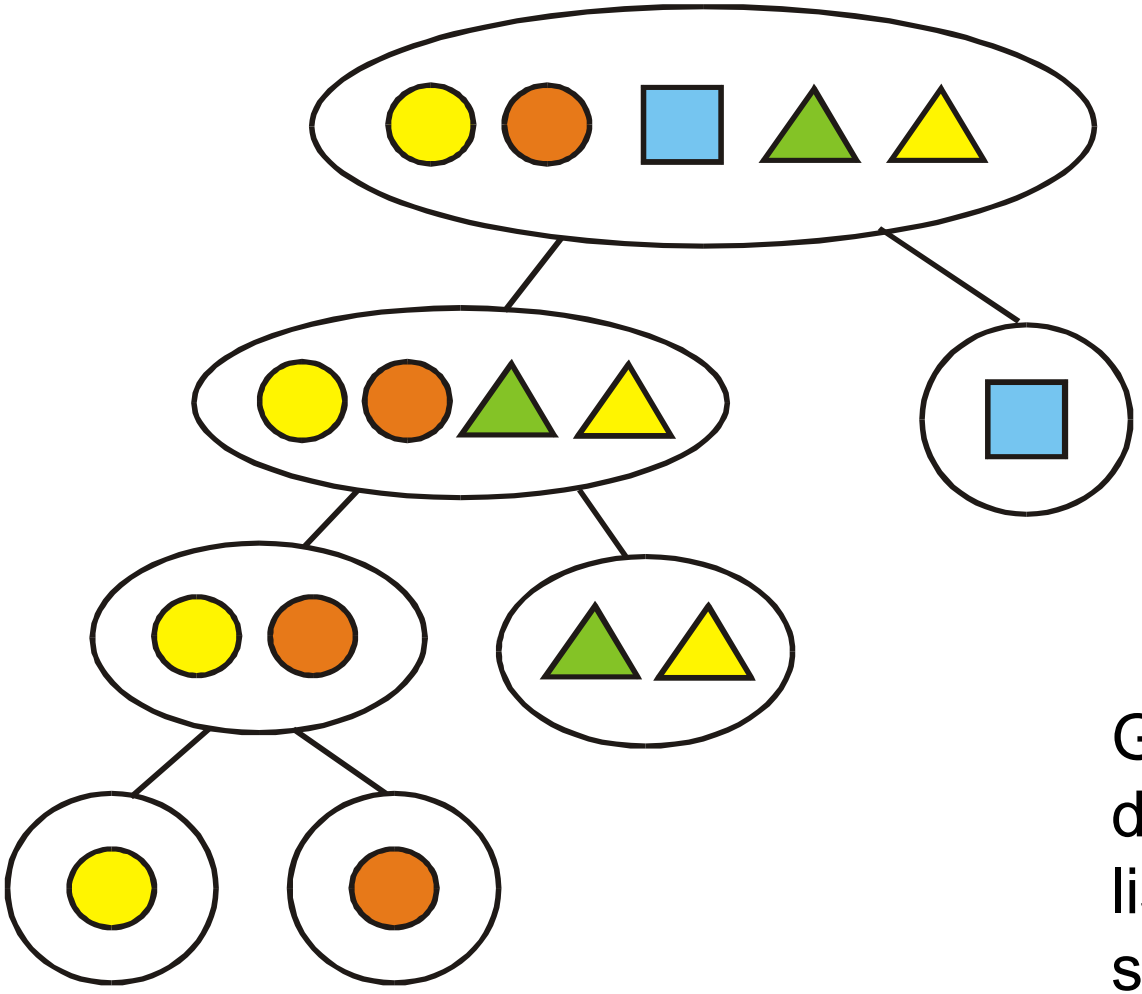
GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – przykład



GDAŃSK UNIVERSITY OF TECHNOLOGY

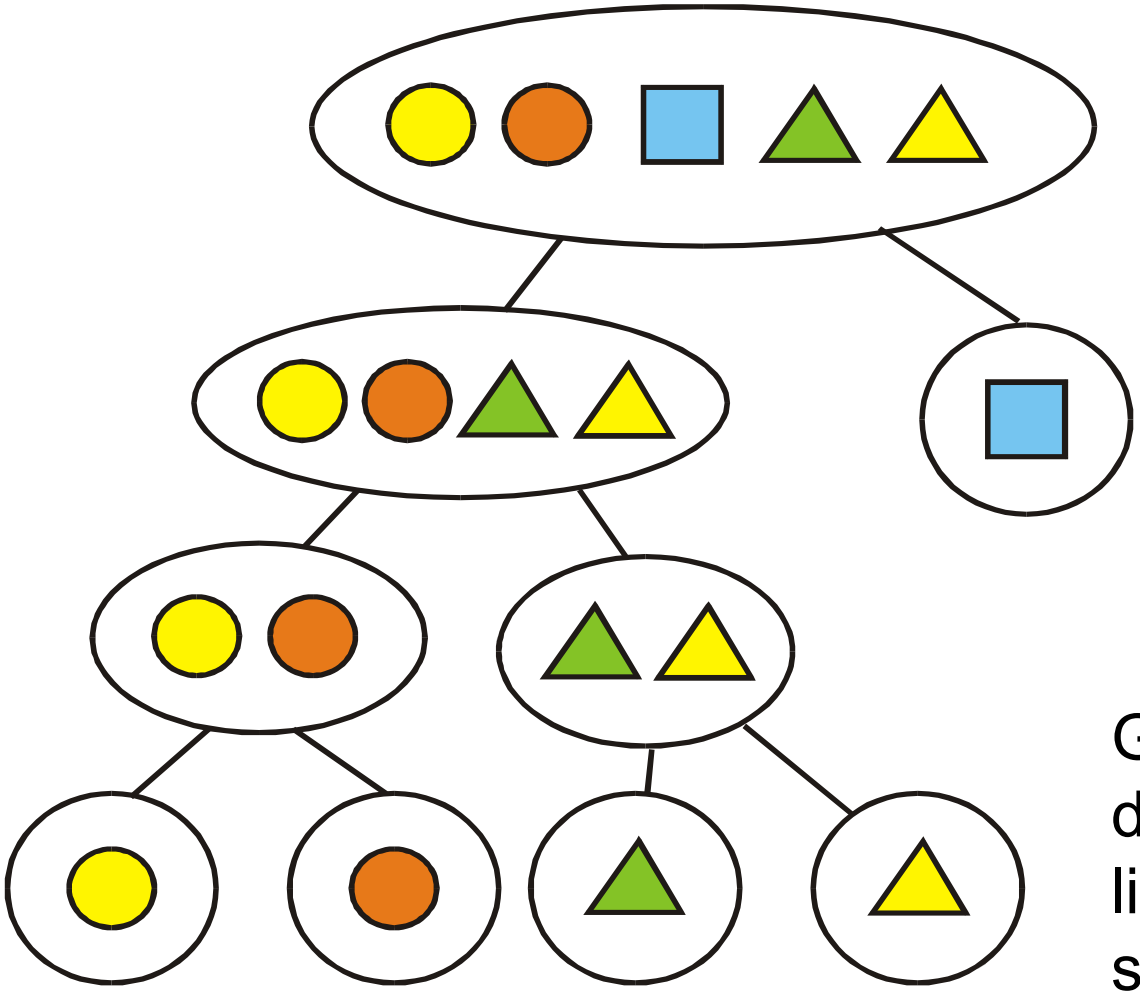
COBWEB – przykład



Gdy przykład dołączany jest do liścia, generowane są dwa nowe liście

GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – przykład



Gdy przykład dołączany jest do liścia, generowane są dwa nowe liście

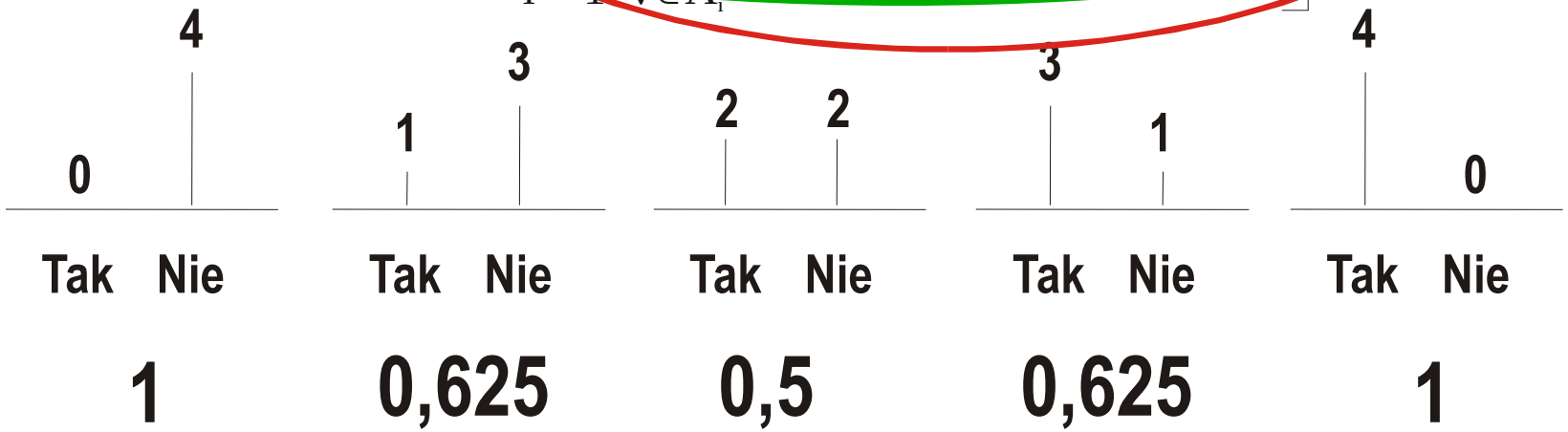
COBWEB – ocena grupowania

$$f(h) = \frac{1}{|C_h|} \sum_{d \in C_h} Pr_{x \in \Omega}(h(x)=d) \cdot$$

zwięzłość

$$\cdot \left[\sum_{i=1}^{n} \sum_{v \in A_i} Pr_{x \in \Omega}(a_i(x)=v \mid h(x)=d)^2 - \sum_{i=1}^{n} \sum_{v \in A_i} Pr_{x \in \Omega}(a_i(x)=v)^2 \right]$$

precyzja



GDAŃSK UNIVERSITY OF TECHNOLOGY

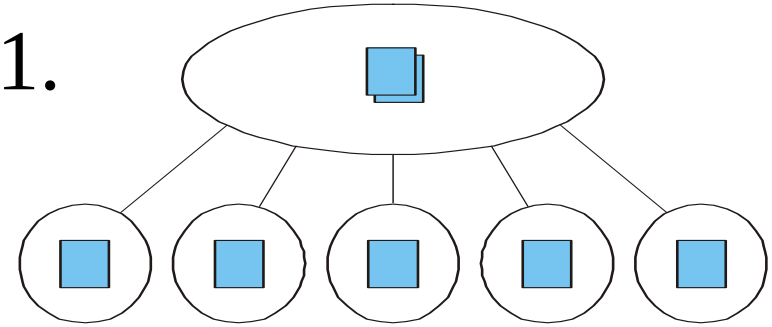
COBWEB – zalety

- COBWEB może być używany w trybie inkrementacyjnym,
- Produkowany jest hierarchiczny podział przykładów,
- Funkcja oceny jakości grupowania może być dostosowana do grupowania niestandardowych typów atrybutów lub nawet niestandardowych przykładów

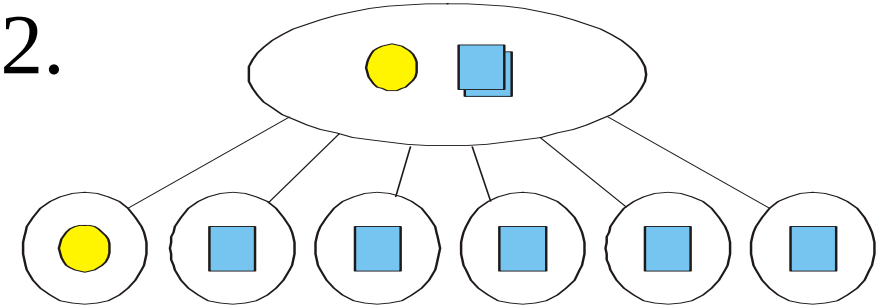
GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – wada

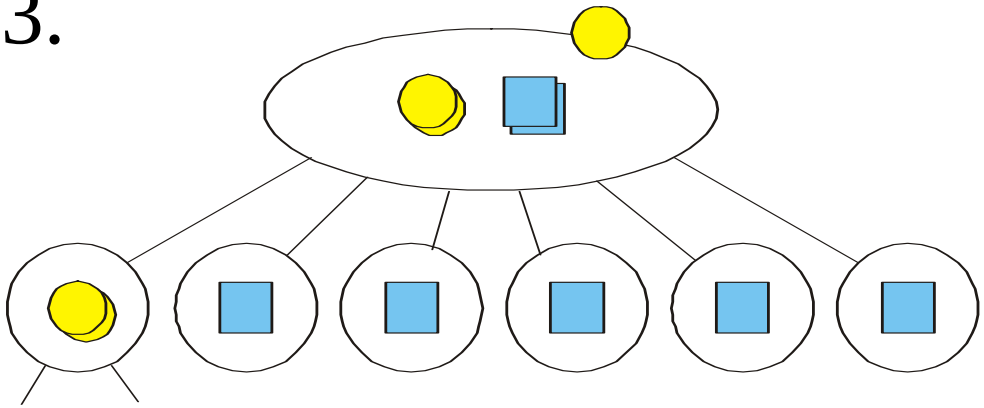
1.



2.

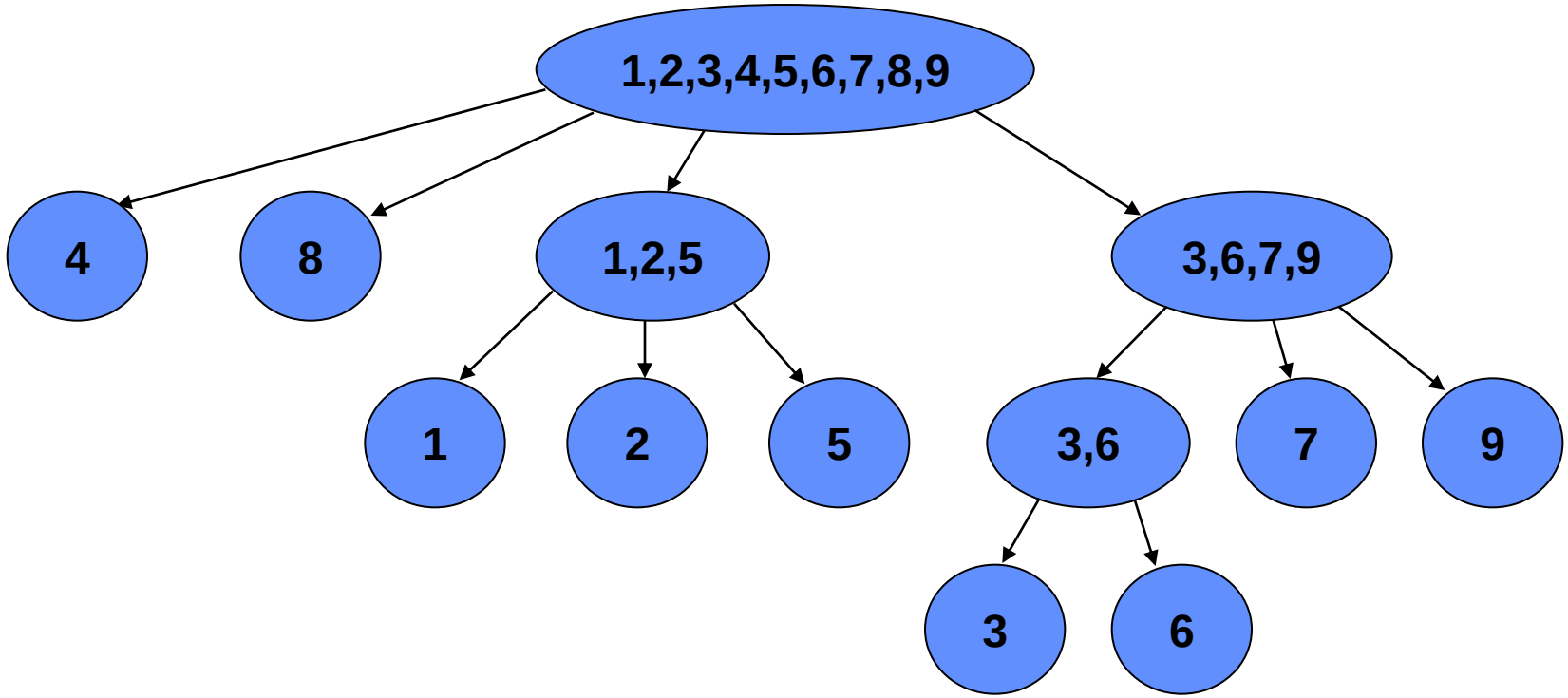


3.



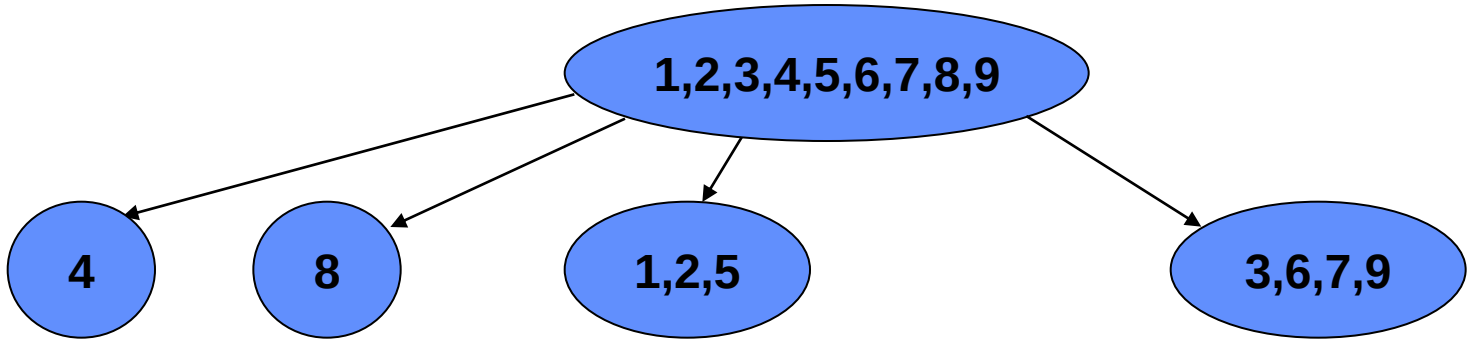
GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – przykład „Z.K.”



GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – klasyfikacja



W praktyce jako wynik grupowania traktuje się drugi poziom drzewa COBWEB licząc od korzenia

Takie zredukowane drzewo może być traktowane również jako klasyfikator

GDAŃSK UNIVERSITY OF TECHNOLOGY

Grupowanie - przykład

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

GDAŃSK UNIVERSITY OF TECHNOLOGY

COBWEB – podsumowanie

- COBWEB dołącza każdy przykład do drzewa kategorii, wykonując proste operacje,
- O tym, którą operację wykonać, decyduje funkcja oceny jakości grupowania według kryteriów precyzji i zwięzłości,
- Działa w trybie inkrementacyjnym, można dostosowywać sposób jego działania,
- Algorytm zachłanny, efekt jego działania mocno zależy od kolejności podawania przykładów na wejście

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dziękujemy za uwagę

Zapraszamy na wykład:

KLASTERYZACJA I SEGMENTACJA cz. 2