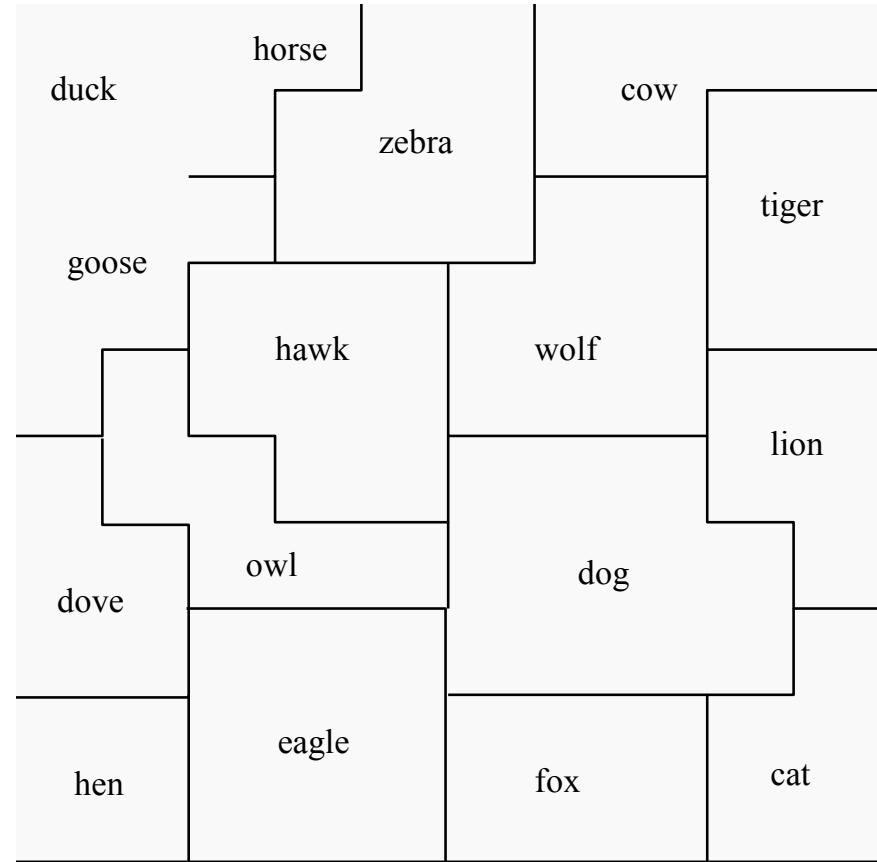
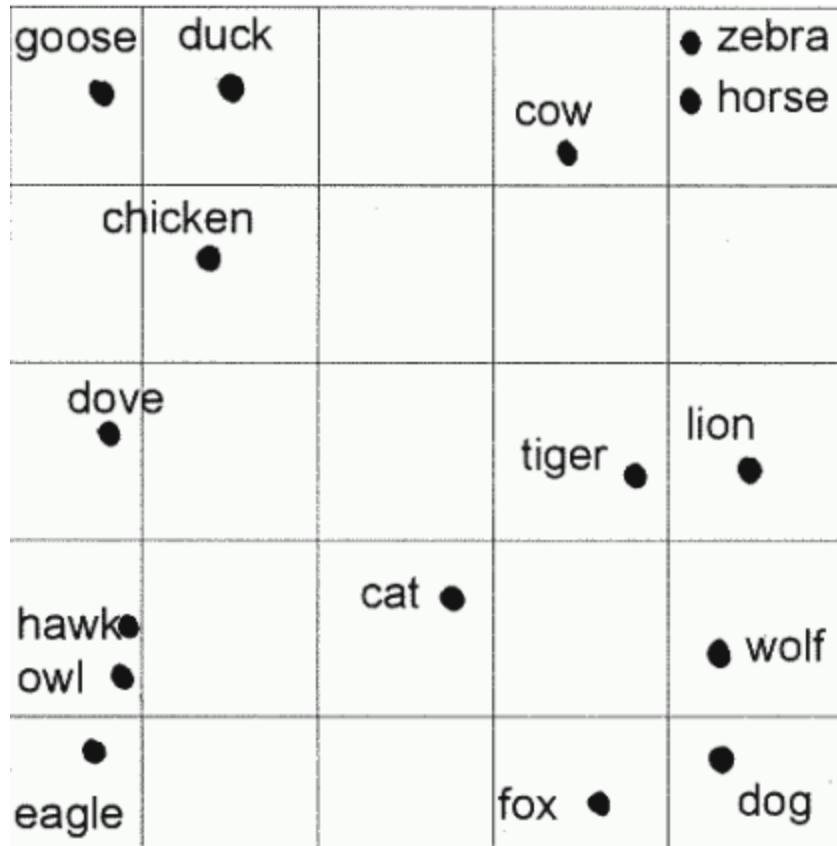


# Porównanie map MDS & SOM



MDS and SOM was used on data vectors from the previous page.

# **Uczenie nienadzorowane**

## **algorytmy grupowania wykład 11**

Jerzy Stefanowski

Instytut Informatyki PP

2021

Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI-TECH)  
projekt finansowany z środków Programu Operacyjnego Polska Cyfrowa  
POPC.03.02.00-00-0001/20



**Fundusze  
Europejskie**  
Polska Cyfrowa



**Rzeczpospolita  
Polska**

**Unia Europejska**  
Europejski Fundusz  
Rozwoju Regionalnego



# Plan wykładu

- Przypomnienie algorytmów grupowania
  - Algorytm k-średnich
  - Algorytmy hierarchiczne
- Grupowanie z wykorzystaniem sieci neuronowych
- Sieci Kohonena
  - Siec LVQ
  - Siec SOM
- Podsumowanie

# Cele algorytmów grupowania

- Obiekt – przykład uczący - opisany za pomocą  $n$  zmiennych  $X_1, X_2, \dots, X_n$  jest punktem  $x=(x_1, \dots, x_n)$  w  $n$ -wymiarowej przestrzeni  $\Omega$
- Cel podziału na grupy ( $S$ ) → obiekty podobne (reprezentowane przez punkty znajdujące się blisko siebie w przestrzeni) przydzielone do tej samej grupy, a obiekty niepodobne (reprezentowane przez punkty leżące w dużej odległości w przestrzeni) znajdują się w różnych grupach

# Czym jest skupienie?

1. Zbiorem najbardziej podobnych obiektów
2. Podzbiór obiektów, dla których odległość jest mniejsza niż ich odległość od obiektów z innych skupień.
3. Podobszar wielowymiarowej przestrzeni zawierający odpowiednio dużą gęstość obiektów, oddzielony od innych podobszarów o dużej gęstości strefą rzadkiego występowania obiektów

# Przykłady zastosowań analizy skupień

- Zastosowania ekonomiczne:
  - Identyfikacja grup klientów bankowych (np. właścicieli kart kredytowych wg. sposobu wykorzystania kart oraz stylu życia, danych osobowych, demograficznych) → cele marketingowe.
  - Systemy rekommendacji produktów i usług.
  - Rynek usług ubezpieczeniowych (podobne grupy klientów).
  - Analiza sieci sprzedaży (np. czy punkty sprzedaży podobne pod względem społecznego sąsiedztwa liczby personelu, itp., przynoszą podobne obroty).
  - Poszukiwanie wspólnych rynków dla produktów.
  - Planowanie przestrzene, np. analiza nieruchomości
- Badania naukowe (biologia, medycyna, nauki społeczne)
- Analiza zachowań użytkowników serwisów WWW
- Rozpoznawanie obrazów, dźwięku
- Wiele innych

# Podział znanych metod

- Podziałowo-optymalizacyjne: Znajdź podział na zadaną liczbę skupień wg. zadanego kryterium.
- Metody hierarchiczne: Zbuduj drzewiastą strukturę skupień.
- Gęstościowo (Density-based): Poszukuj obszarów o większej gęstości występowania obserwacji
- Grid-based: wykorzystujące wielowymiarowy podział przestrzeni siatką ograniczeń
- Model-based: hipoteza co do własności modelu pewnego skupienia i procedura jego estymacji.

# Inne kryteria podziału [Jain przegląd]

Rodzaj rozwiązań algorytmicznych

- Podziałowo-optymalizacyjne
  - Optymalizacja kryterium, np. k-średnich
  - Mieszaniny rozkładów prawd. (EM)
  - Grafowe
- Hierarchiczne
  - AHC (różne metody łączenia od mniejszych skupień do większych)
  - Deglomeracyjne (podział większych grup)
  - Dostosowane do masywnych danych (BIRCH)
  - Wspierające opis probabilistyczny (COBWEB)
- Inne

Skupienia: jednoznaczny przydział obiektu vs. rozmyty

Tryb przetwarzania danych (pełen dostęp vs. przyrostowy)

Więcej: A. Jain: Data Clustering: 50 Years Beyond K-Means

# Problemy do rozstrzygnięcia przed wyborem metody/algorytmu

- Jak odwzorować obiekty w przestrzeni?
  - Wybór zmiennych
  - Normalizacja zmiennych
- Jak mierzyć odległości między obiektami?
  - Przypomnienie wcześniejszego wykładu z kNN
- Jaką metodę grupowania zastosować?

# Różny zakres danych liczbowych

Normalizacja ma na celu doprowadzenie obiektów lub zmiennych do porównywalnych wielkości. Problem ten dotyczy zmiennych mierzonych w różnych jednostkach (np. sztuki, czas, waluta).

## Przykład

- Rozważmy 3 obiekty i dwie zmienne: wiek osoby mierzony w latach i jej dochód mierzony w złotych lub tys. zł.

Zmienna ->	X	Y1	Y2
Osoba	Wiek	Dochód	Dochód
	(w latach)	(w zł)	( w tys. zł)
A	35	12000	12,0
B	37	6700	6,7
C	45	7000	7,0

# Podobieństwo – podstawa analizy skupień

- Ciągle dyskusyjne – zwłaszcza w nietechnicznych zastosowaniach



- The real meaning of similarity is a philosophical question. We will take a more pragmatic approach
- Depends on representation and algorithm. For many rep./alg., easier to think in terms of a distance (rather than similarity) between vectors.
- Prościej mówić o odległościach między obserwacjami
  - Zwłaszcza jak są matematycznie dobrze zdefiniowane
  - Metryka odległości

# Algorytmy podziałowo – optymalizacyjne

- Zadanie: Podzielenie zbioru obserwacji na  $K$  zbiorów elementów (skupień  $C$ ), które są jak najbardziej jednorodne
- Jednorodność – funkcja oceny
- Intuicja → zmienność wewnętrzskupieniowa  $wc(C)$  i zmienność międzyskupieniowa  $bc(C)$

Mögliwe są różne sposoby zdefiniowania

- np. wybierzmy środki skupień  $\mathbf{r}_k$  (centroidy)  $\mathbf{r}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathbf{x}$
- Co prowadzi do

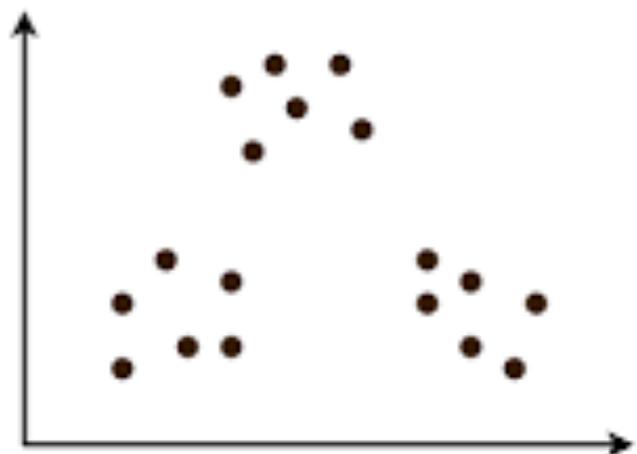
$$wc(C) = \sum_{k=1}^K \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \mathbf{r}_k)^2$$

$$bc(C) = \sum_{1 \leq j < k \leq K} d(\mathbf{r}_j, \mathbf{r}_k)^2$$

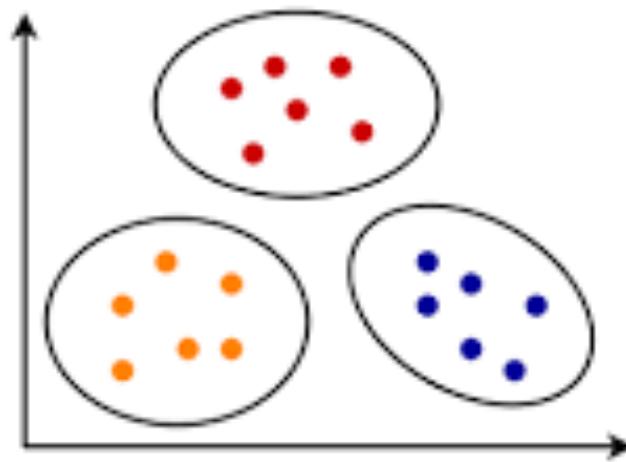
# Algorytm k średnich

- Cel:  $K$  - średnich  $\rightarrow$  minimalizacja  $wc(C)$
- Przeszukiwanie przestrzeni możliwych przypisań  $\rightarrow$  bardzo kosztowne (oszacowanie w ks. Koronackiego)
- Problem optymalizacji kombinatorycznej  $\rightarrow$  systematyczne przeszukiwanie metodą iteracyjnego udoskonalania:
  - Rozpocznij od rozwiązania początkowego (losowego).
  - Ponownie przypisz punkty do skupień tak, aby otrzymać największą zmianę w funkcji oceny.
  - Przelicz zaktualizowane środki skupień, ...
  - Postępuj aż do momentu, w którym nie ma już żadnych zmian w funkcji oceny lub w składzie grup.
- Zachłanne przeszukiwanie  $\rightarrow$  proste i prowadzi do co najmniej lokalnego minimum. Różne modyfikacje, np. rozpoczęnięcia od kilku rozwiązań startowych
- Złożoność algorytmu K - średnich  $\rightarrow O(Knl)$

# Ilustracja k-średnich



Before K-Means



After K-Means

# Ustalanie liczby skupień i startowych centroidów

Liczبę skupień wybiera się na podstawie przesłanek merytorycznych albo szacuje się je metodami hierarchicznymi. Można dokonać obliczeń dla wszystkich wartości  $k$  z ustalonego przedziału:

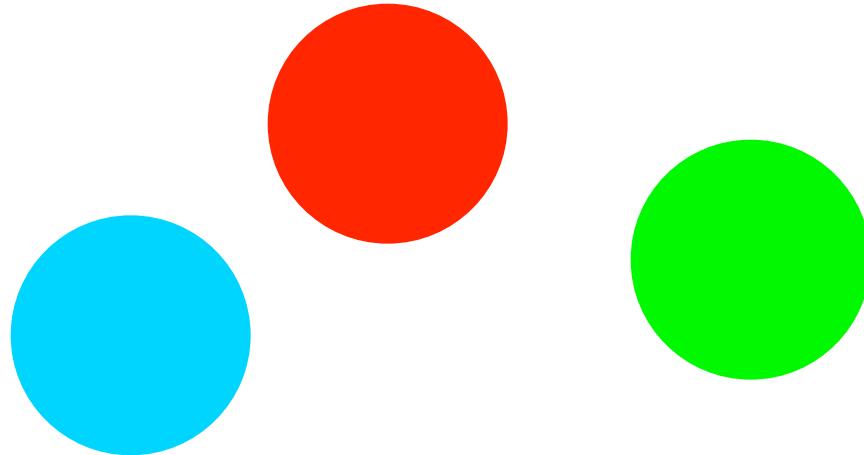
$$k_{\min} \leq k \leq k_{\max}$$

Możliwe są różne podejścia:

1. Arbitralny sposób np. przyjmuje się współrzędne pierwszych  $k$  obiektów jako załączki środków ciężkości
2. Losowy wybór środków ciężkości, przy czym może to być losowy wybór  $k$  obiektów ze zbioru danych albo losowy wybór  $k$  punktów przestrzeni niekoniecznie pokrywających się z położeniem obiektów
3. Wykorzystanie algorytmu optymalizującego w pewien sposób położenie początkowych środków ciężkości np. przez uwzględnianie  $k$  obiektów leżących daleko względem siebie
4. Przyjęcie jako początkowych środków ciężkości uzyskanych na podstawie podziału otrzymanego inną metodą, głównie jedną z metod hierarchicznych

# Pewne ukierunkowanie K-średnich

- Tworzy się „kuliste” kształty skupień



- Co z obserwacjami odstającymi i nieregularnymi kształtami skupień?

# K-means krótkie podsumowanie

## Zalety

- Proste i łatwe do zrozumienia
- Reprezentacja skupień jako centroidy

## Wady

- Jawne podanie liczby skupień
- Wszystkie przykłady muszą być przydzielone do skupień
- Problem z outliers (za duża wrażliwość)
- Ukierunkowanie na jednorodne „sferyczne” kształty skupień

# Dalsze rozszerzenia k-średnich

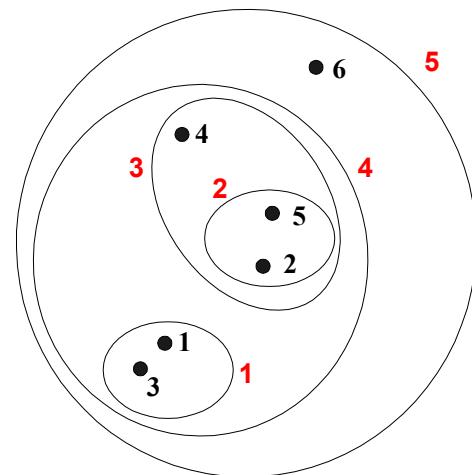
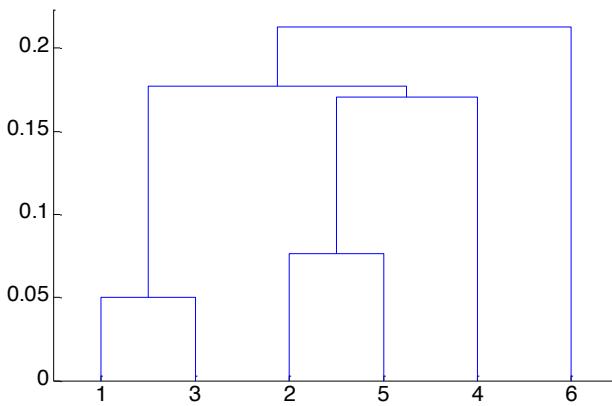
- Rozmyte k-means (Fuzzy ISODATA)
- Wersja k-medoids
- Rozszerzenia dla przetwarzania dużych wolumenów danych, np. PAM
- Inspiracje dla modeli statystycznych (EM)

Kolejny wykład omawia niektóre z nich

Warto zapoznać się z książką S.Wierzchoń, M.Kłopotek:  
Algorytmy analizy skupień. WNT 2015

# Grupowanie hierarchiczne

- Tworzy się stopniowo hierarchię zawierających się skupisk
  - Połączenie lub podział podzbiorów obiektów
- Wizualizacja – struktura drzewa nazwana **dendrogramem**



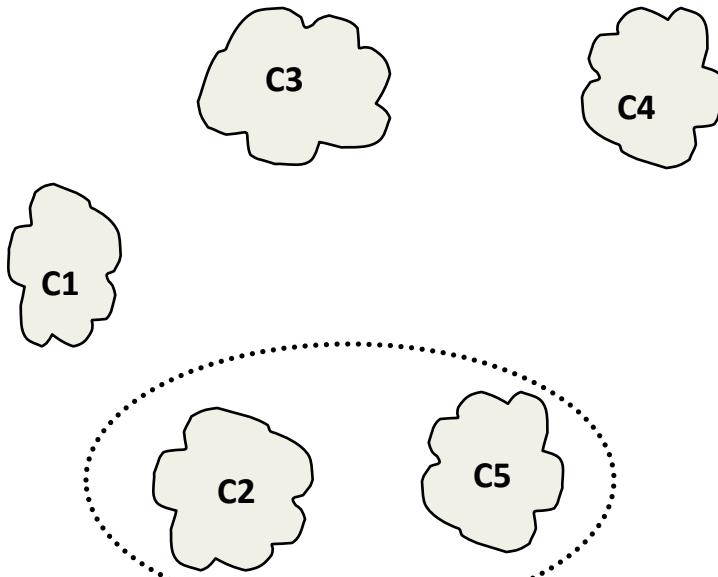
## Hierarchiczne metody aglomeracyjne - algorytm

1. W macierzy odległości znajduje się parę skupień najbliższych sobie.
2. Redukuje się liczbę klas łącząc znalezioną parę
3. Przekształca się macierz odległości metodą wybraną jako kryterium klasyfikacji
4. Powtarza się kroki 1- 3 dopóki nie powstanie jedna klasa zawierająca wszystkie skupienia.

# Jak przeliczać macierz odległości?

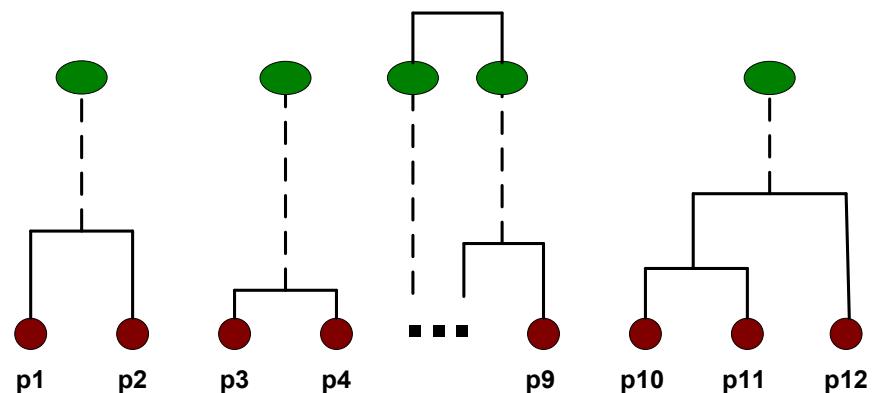
Łączymy dwa skupiska (C2 i C5) i aktualizujemy macierz odległości

Metody hierarchiczne różnią sposobem łączenia skupisk (ang. Linkage method)



	c1	c2	c3	c4	c5
c1					
c2					
c3					
c4					
c5					

Macierz odległości



# Hierarchiczne grupowanie wybór metody łączenia

1. Najbliższego sąsiedztwa (*Single linkage, Nearest neighbor*)
2. Najdalszego sąsiedztwa (*Complete linkage, Furthest neighbor*)
3. Mediany (*Median clustering*)
4. Środka ciężkości (*Centroid clustering*)
5. Średniej odległości wewnętrz skupień  
(*Average linkage within groups*)
6. Średniej odległości między skupieniami  
(*Average linkage between groups*)
7. Minimalnej wariancji Warda (*Ward's method*)

# Odległości między skupieniami

Single linkage  
minimum distance:

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \|p - p'\|$$

Complete linkage  
maximum distance:

$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \|p - p'\|$$

mean distance:

$$d_{mean}(C_i, C_j) = \|m_i - m_j\|$$

average distance:

$$d_{ave}(C_i, C_j) = 1 / (n_i n_j) \sum_{p \in C_i} \sum_{p' \in C_j} \|p - p'\|$$

$m_i$  Jest średnią obiektów z  $C_i$      $n_i$  Jest liczbą obiektów w skupisku  $C_i$

# Single Link Agglomerative Clustering

- Użyj maksymalnego podobieństwa dwóch obiektów:

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- Prowadzi do „(long and thin) clusters due to *chaining effect*” (efekt łańcuchowy); prowadzić do formowania grup niejednorodnych (heterogenicznych);
  - Dogodne w specyficznych zastosowaniach
- Pozwala na wykrycie **obserwacji odstających**, nie należących do żadnej z grup, i warto przeprowadzić klasyfikację za jej pomocą na samym początku, aby wyeliminować takie obserwacje i przejść bez nich do właściwej części analizy

# Complete Link Agglomerative Clustering

- Użyj maksymalnej odległości – minimalnego podobieństwa

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

- Ukierunkowana do “tight,” spherical clusters
- Metoda zalecana gdy, kiedy obiekty faktycznie formują naturalnie oddzielone "kępki". Metoda ta nie jest odpowiednia, jeśli skupienia są w jakiś sposób wydłużone lub mają naturę "łańcucha".

# Wrażliwość na dobór metod łączenia skupień

Diagram dla 22 przyp.

Pojedyncze wiązanie

Odległości euklidesowe

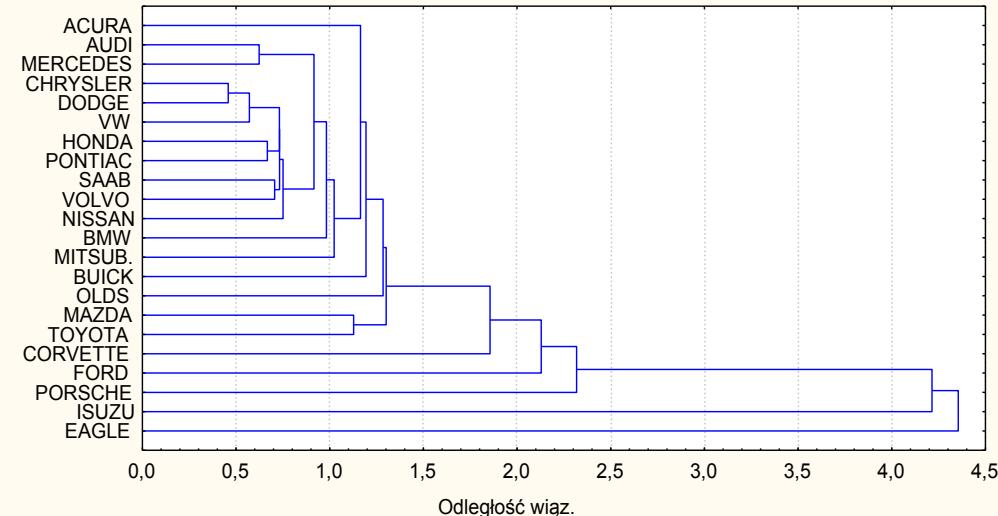
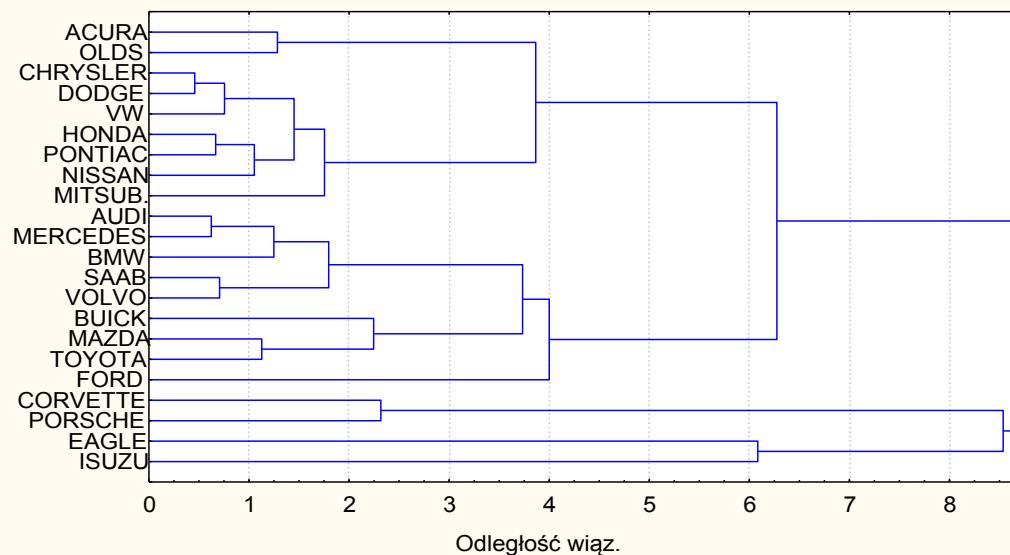


Diagram dla 22 przyp.

Metoda Warda

Odległości euklidesowe



Rysunki – z własnego  
uruchomienia Statsoft Statistica

# Metoda średnich połączeń [Unweighted pair-group average]

- W metodzie tej odległość między dwoma skupieniami oblicza się jako średnią odległość między wszystkimi parami obiektów należących do dwóch różnych skupień
- Metoda ta jest efektywna, gdy obiekty formują naturalnie oddzielone "kępki", ale zdaje także egzamin w przypadku skupień wydłużonych, mających charakter "łańcucha"

# Metoda ważonych środków ciężkości (mediany) [Weighted pair-group centroid]

- Jest to metoda podobna jak poprzednia, z tym wyjątkiem, że w obliczeniach wprowadza się „ważenie”, aby uwzględnić różnice między wielkościami skupień (tzn. liczbą zawartych w nich obiektów).
- Zatem, metoda ta jest lepsza od poprzedniej w sytuacji, gdy istnieją (lub podejrzewamy, że istnieją) znaczne różnice w rozmiarach (liczności) skupień

# Metody łączenia – Ward method

- Gdy powiększamy jedno ze skupień  $C_k$ , wariancja wewnętrzgrupowa (liczona przez kwadraty odchyлеń od średnich w zbiorach  $C_k$ ) rośnie.
- Metoda polega na takim powiększaniu zbiorów  $C_k$ , która zapewnia **najmniejszy przyrost tej wariancji** dla danej iteracji.
- Kryterium grupowania jednostek: minimum zróżnicowania wektorów cech  $x_j$  tworzących zbiór  $C_k$  ( $k = 1, \dots, K$ ) względem wartości średnich w tych zbiorach.
- Ogólnie, metoda ta jest traktowana jako bardzo efektywna, chociaż zmierza do tworzenia skupień o małej wielkości → zrównoważone drzewa o wielu elementach
- Ważne – powiązanie z miarą odległości między obiektami (Pearson vs. inne)

# Przykłady użycia metody Warda

## Cars data

Diagram dla 22 przyp.

Metoda Warda

Odległości euklidesowe

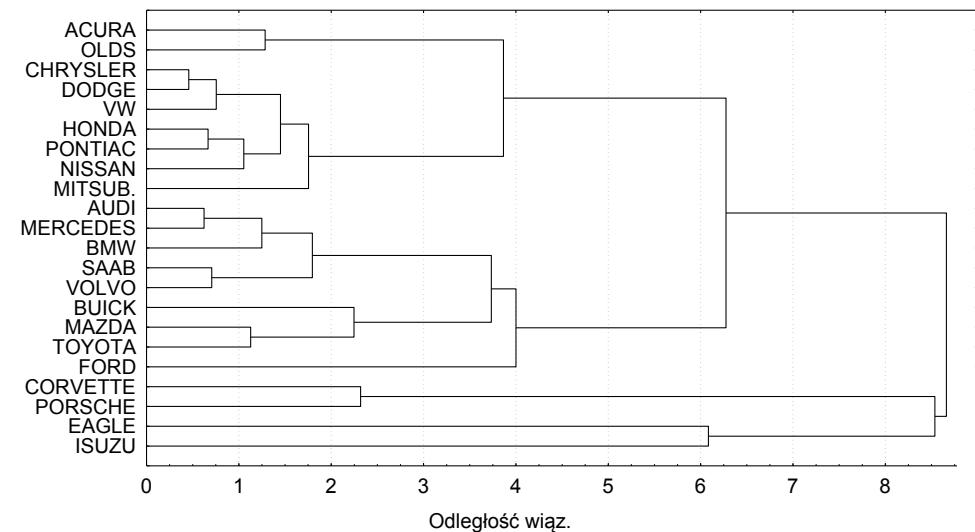
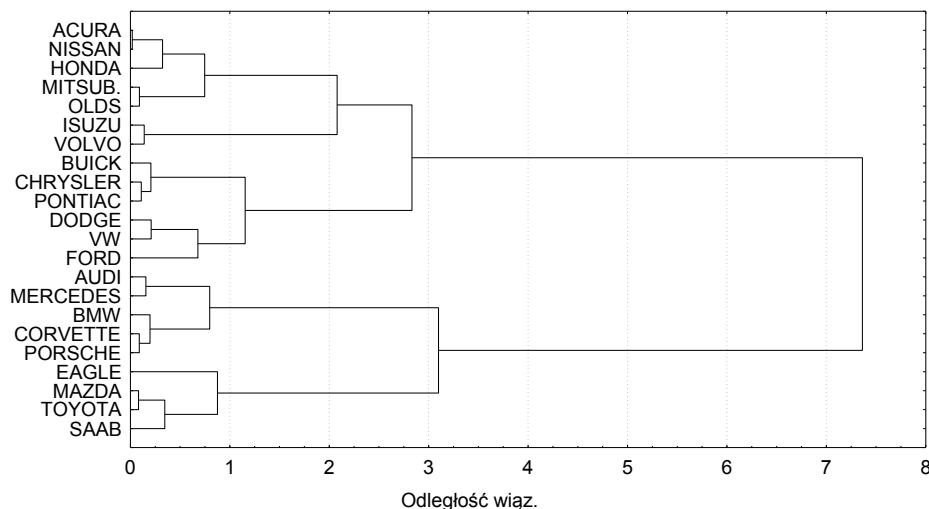


Diagram dla 22 przyp.

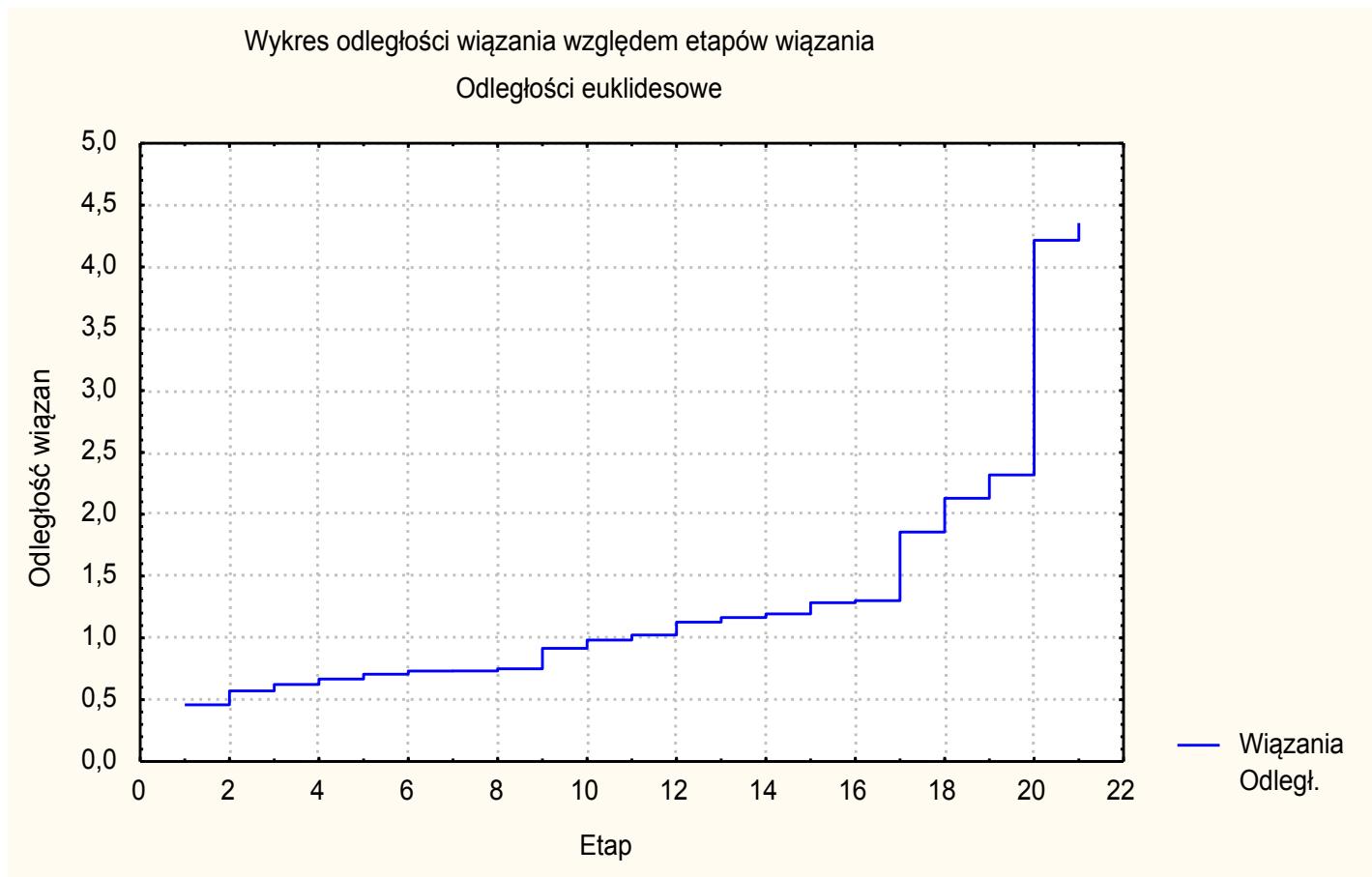
Metoda Warda

1-r Pearsona



# AHC – jak odnaleźć liczbę skupień?

Znajdź punkt przegięcia („kolanko”) wykresu



# Sieci neuronowe - samoorganizujące

Propozycja T.Kohonena

# Typowe zadania dla sieci

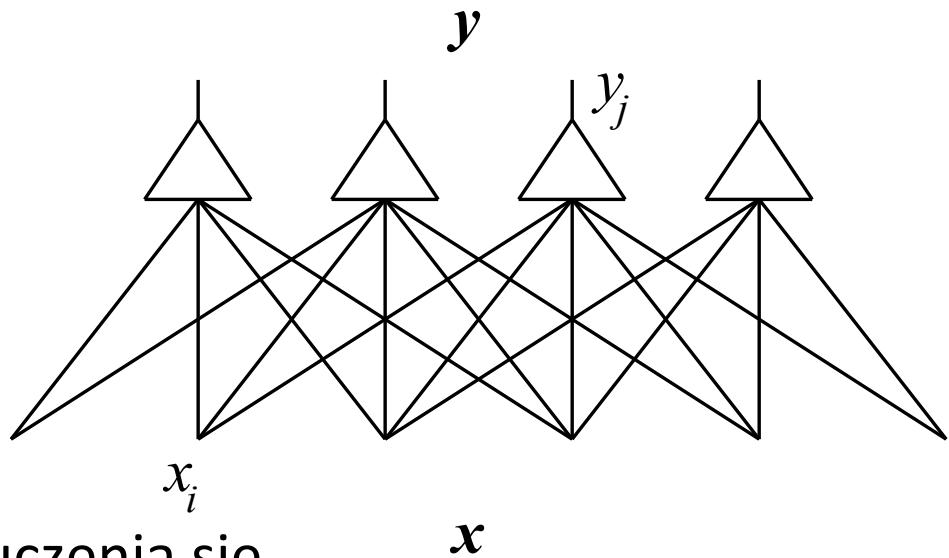
- **Grupowanie obserwacji:** Sieć w wyniku procesu uczenia dokonuje podziału przykładów uczących na klasy (grupy) przykładów podobnych do siebie i przyporządkowuje każdej klasie różne elementy wyjściowe sieci – **sieci LVQ**.
- **Tworzenie mapy cech:** Dane wejściowe transformowane są z wielowymiarowej przestrzeni przykładów w „małowymiarową” przestrzeń ich cech charakterystycznych. Elementy warstwy wyjściowej są geometrycznie uporządkowane. Wymaga się, aby podobne przykłady wejściowe generowały aktywność bliskich geometrycznie elementów wyjściowych - **Sieci SOM Kohonena**.
- **Analiza czynników głównych:** Sieć posiada wyjście wieloelementowe, a każdy z elementów odpowiada za jeden z tzw. czynników głównych, według których określone jest podobieństwo sygnałów wejściowych.

# Wprowadzenie do sieci Kohonena

- Inny tryb uczenia się:
  - Bez nadzoru (brak informacji  $y$  o zadanym wyjściu; tylko opis przykładów  $x$ )
  - Sieć sama powinna wykrywać istotne zależności w danych wejściowych, badać podobieństwo wektorów  $x$ , rozpoznawania cech istotnych czy regularności bez „nadzoru”
  - Typowe zastosowanie → grupowanie, kodowanie i kompresja, projekcja wielowymiarowa, wykrywanie cech istotnych.
  - Kluczowe jest badanie podobieństwa wektorów (wejścia, wag),
    - Miara iloczynu skalarnego wektora wag i wektora wejściowego
- **Zasady uczenia się konkurencyjnego** (przez współzawodnictwo)
- Tylko zwycięskie neurony lub ich sąsiedzi są nauczani (modyfikacja wag)
- Na ogólną prostszą topologię sieci

# Podstawowa sieć Kohonena LVQ

- Celem jest grupowanie wektorów wejściowych  $\mathbf{x}$
- Istota działania → podobieństwo wektorów
- Podobne wektory powinny pobudzać te same neurony
- Prosta topologia



- gdzie  $y_j = \mathbf{w}_j \mathbf{x} = \sum_i w_{ij} x_i$
- Reguła konkurencyjnego uczenia się

# Wektory i miary podobieństwa

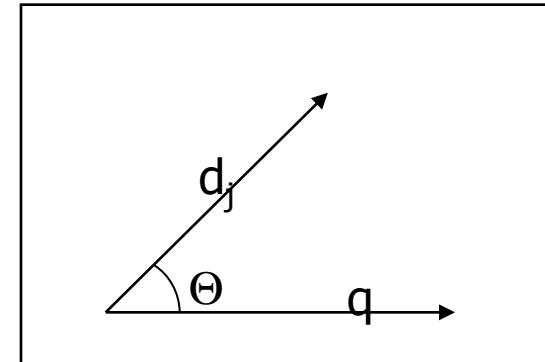
- Dany jest zbiór uczący  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$
- Podobieństwo dwóch wektorów – odległość Euklidesowa

$$\|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \cdot (\mathbf{x}_i - \mathbf{x}_j)}$$

- Równoważna miara cosinusowa (kątowa)

$$\cos(\theta) = \frac{\mathbf{x}_i^T \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}$$

$$\cos \theta_j = \frac{\mathbf{a}_j^T \mathbf{q}}{\|\mathbf{a}_j\|_2 \|\mathbf{q}\|_2} = \frac{\sum_{i=1}^t a_{ij} q_i}{\sqrt{\sum_{i=1}^t a_{ij}^2} \sqrt{\sum_{i=1}^t q_i^2}}$$



# Sieć Kohonena - LVQ

- Przetwarza się kolejne wektory  $x$  poszukując  $p$  grup – odpowiadając im wybrane neurony
- Przed rozpoczęciem uczenia wektory wag są inicjowane losowo (małe liczby z przedziału  $-0.5;0.5$ )
- Wektory wag są normalizowane dla kolejnych  $p$  neuronów

$$\hat{\mathbf{w}}_i = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|}$$

- Stosuję się regułą „zwycięzca bierze wszystko” w celu identyfikacji neuronu zwycięzcy.
- Wagi TYLKO neuronu zwycięskiego podlegają modyfikacji

# Reguła „zwycięzca bierze wszystko”

- Określenie zwycięzcy:

$$\| \mathbf{x} - \hat{\mathbf{w}}_m \| = \min_{i=1,\dots,p} \| \mathbf{x} - \hat{\mathbf{w}}_i \|$$

- Alternatywnie iloczyn skalarny

$$\hat{\mathbf{w}}_m^T \mathbf{x} = \max_{i=1,\dots,p} \hat{\mathbf{w}}_i^T \mathbf{x}$$

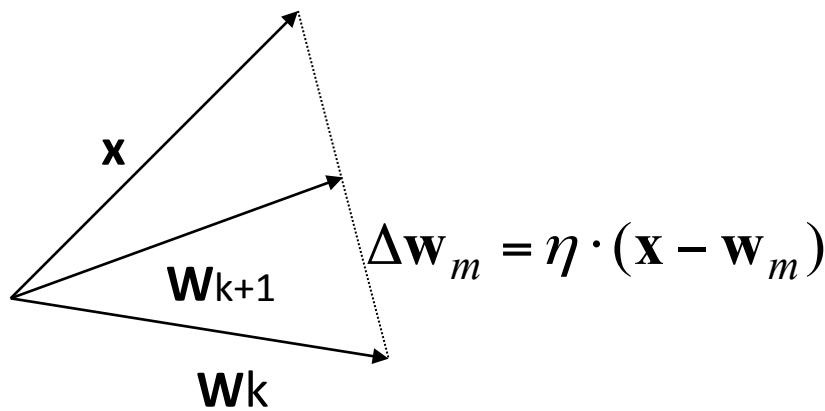
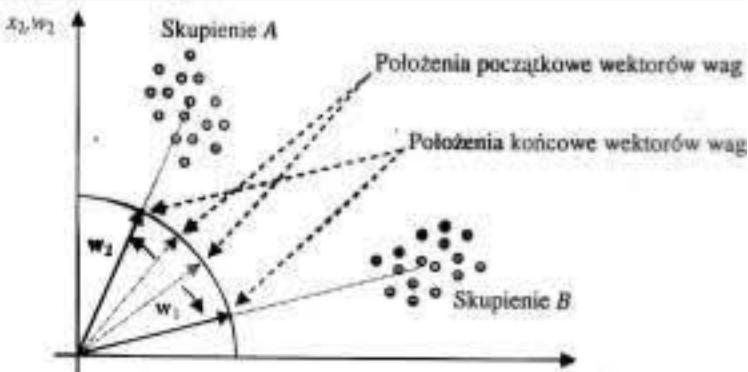
- Zwycięzcą jest jeden neuron  $m$ . Korekcja wag  $\mathbf{w}_m$  odbywa się wyłącznie dla neuronu zwycięzcy według reguły:
- $\eta$  - stała uczenia (na ogólny między 0.1 a 0.7)  $\Delta \mathbf{w}_m = \eta \cdot (\mathbf{x} - \mathbf{w}_m)$
- Przykład – interpretacja geometryczna

# Uczenie zwycięskiego neuronu

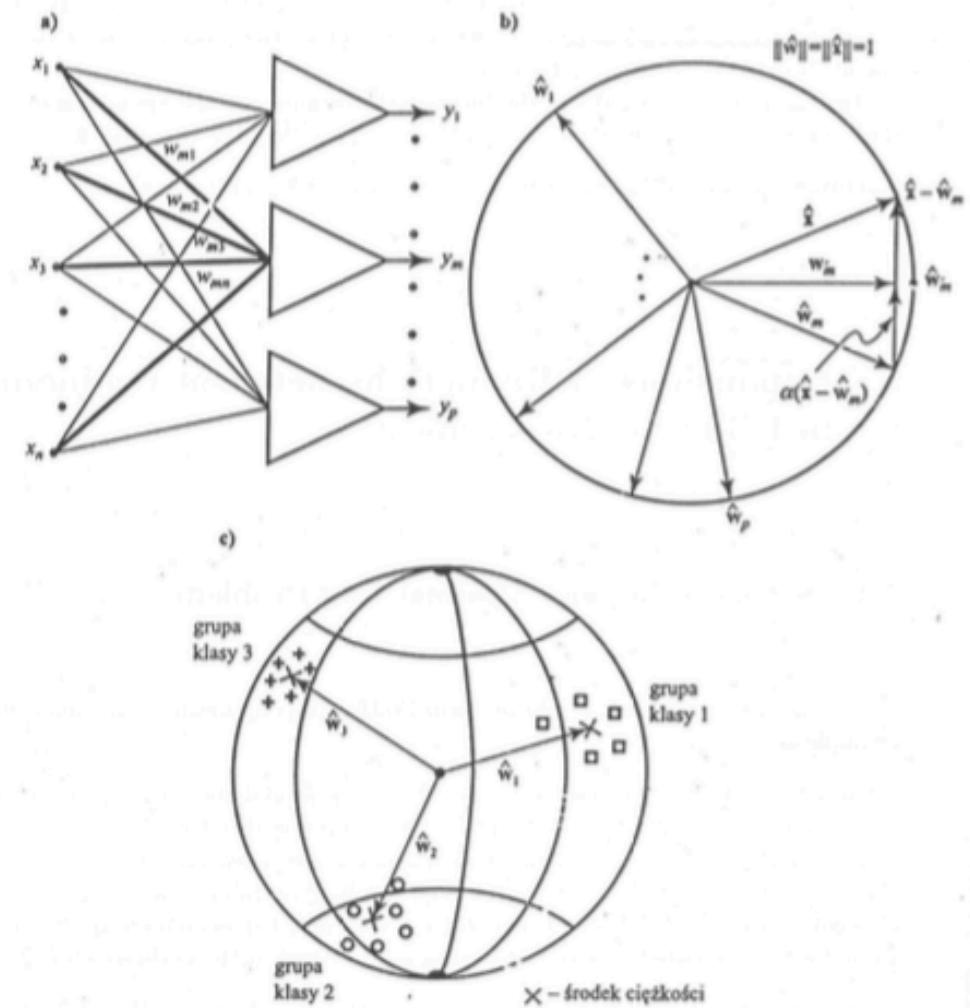
Po odnalezieniu zwycięskiego neuronu dokonuje się aktualizacji wag (k numer kolejnego kroku)

$$\mathbf{w}_m^{k+1} = \mathbf{w}_m^k + \eta \cdot (\mathbf{x} - \hat{\mathbf{w}}_m^k)$$

$$\hat{\mathbf{w}}_m^{k+1} = \frac{\mathbf{w}_m^{k+1}}{\|\mathbf{w}_m^{k+1}\|}$$

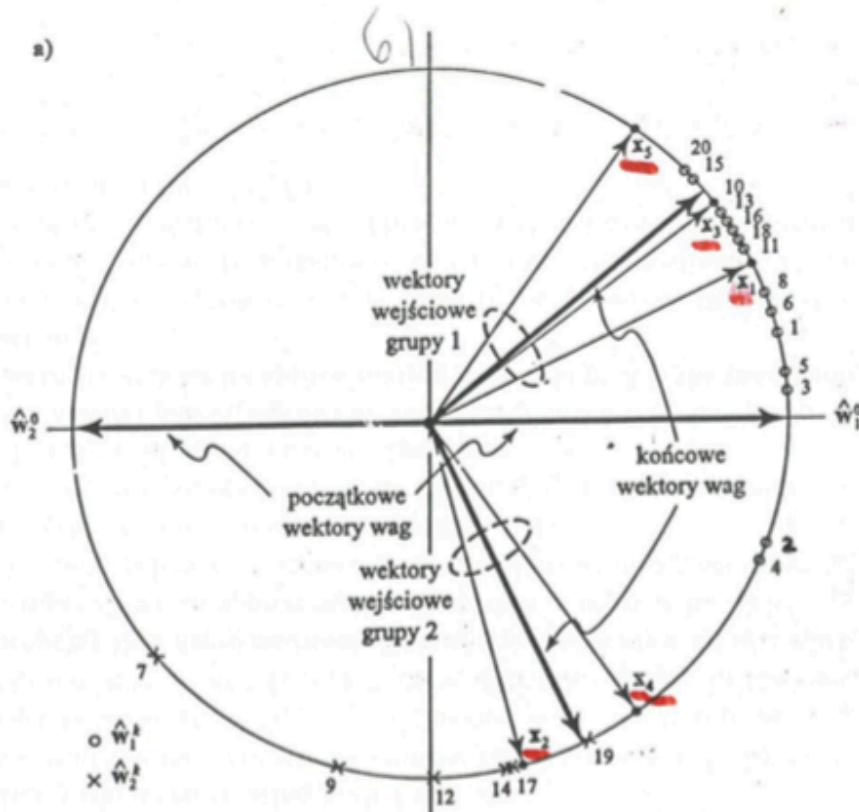


# Ilustracja uczenia sieci LVQ



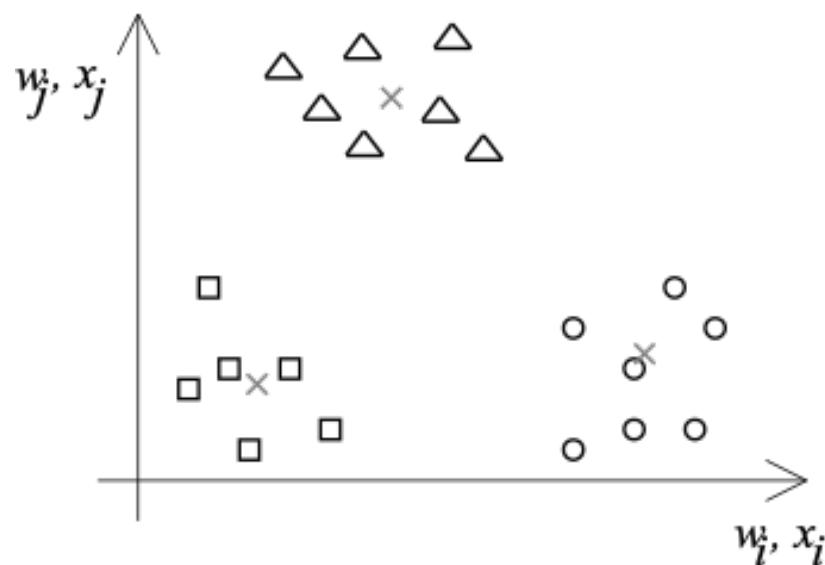
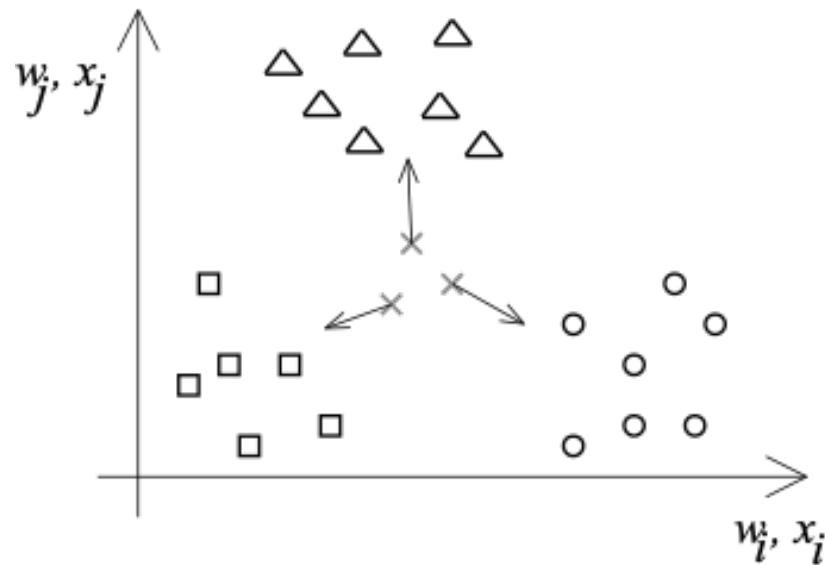
Rys. 7.6. Uczenie się z rywalizacją: a) warstwa ucząca się, b) interpretacja geometryczna kroku uczenia, c) wektory wag na kuli jednostkowej ( $p = n = 3$ ) (grubszego linie odpowiadają modyfikowanym wagom)

# Ilustracja procesu uczenia dwóch skupisk



Krok $k$	$\hat{w}_1^k$	$\hat{w}_2^k$
1	18,46	-180,00
2	-30,77	
3	7,11	
4	-31,45	
5	7,11	
6	31,45	
7		-130,22
8	34,43	
9		-100,00
10	43,78	
11	40,33	
12		-90,00
13	42,67	
14		-80,02
15	47,90	
16	42,39	
17		-80,01
18	43,69	
19		-75,01
20	48,42	

# Wspólna wizualizacja wag i przykładów

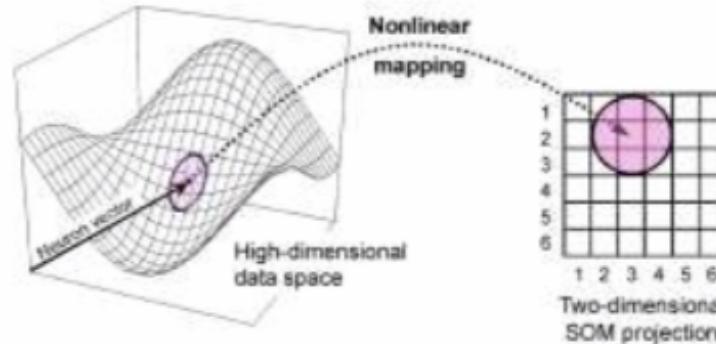


# Kilka uwag o uczeniu sieci

- Po zakończeniu uczenia znormalizowane wektory wag (neuronów) wskazują środki ciężkości wykrytych grup obserwacji → analogia do k-srednich.
- Dobór wag początkowych – rozrzucenie po przestrzeni (hiperkuli)
- Problem doboru liczby neuronów (martwe neurony)
- Tzw. techniki sumienia → „sumienie” ten sam neuron nie zwycięża zbyt często
- Stopniowe zmniejszanie prędkości uczenia
- Iteracyjne powtarzanie prezentacji przykładów
- W ostatnim kroku – „kalibracja” sieci

# Odzwierowanie cech istotnych

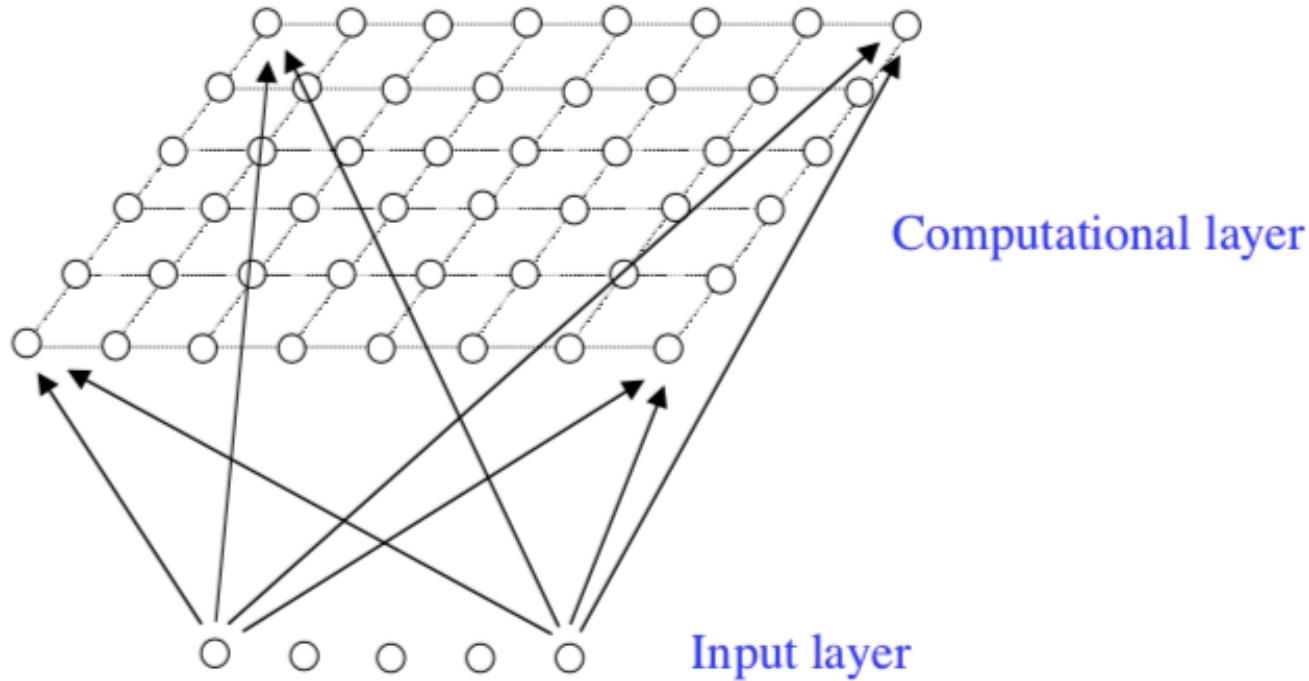
- W eksploracji danych duże znaczenie ma transformacja wysoce-wielowymiarowych danych wejściowych w małowymiarowe przestrzenie tak aby zauważać pewną harmonijną strukturę danych
- Podejścia analityczne (projekcje w statystyce), np. PCA, Skalowanie wielowymiarowe MDS -> patrz niezależny wykład dr Susmagi
- SOM – projekcja nieliniowa z zachowaniem bliskości położenia przykładów na warstwie wizualizacyjnej



# Sieci SOM

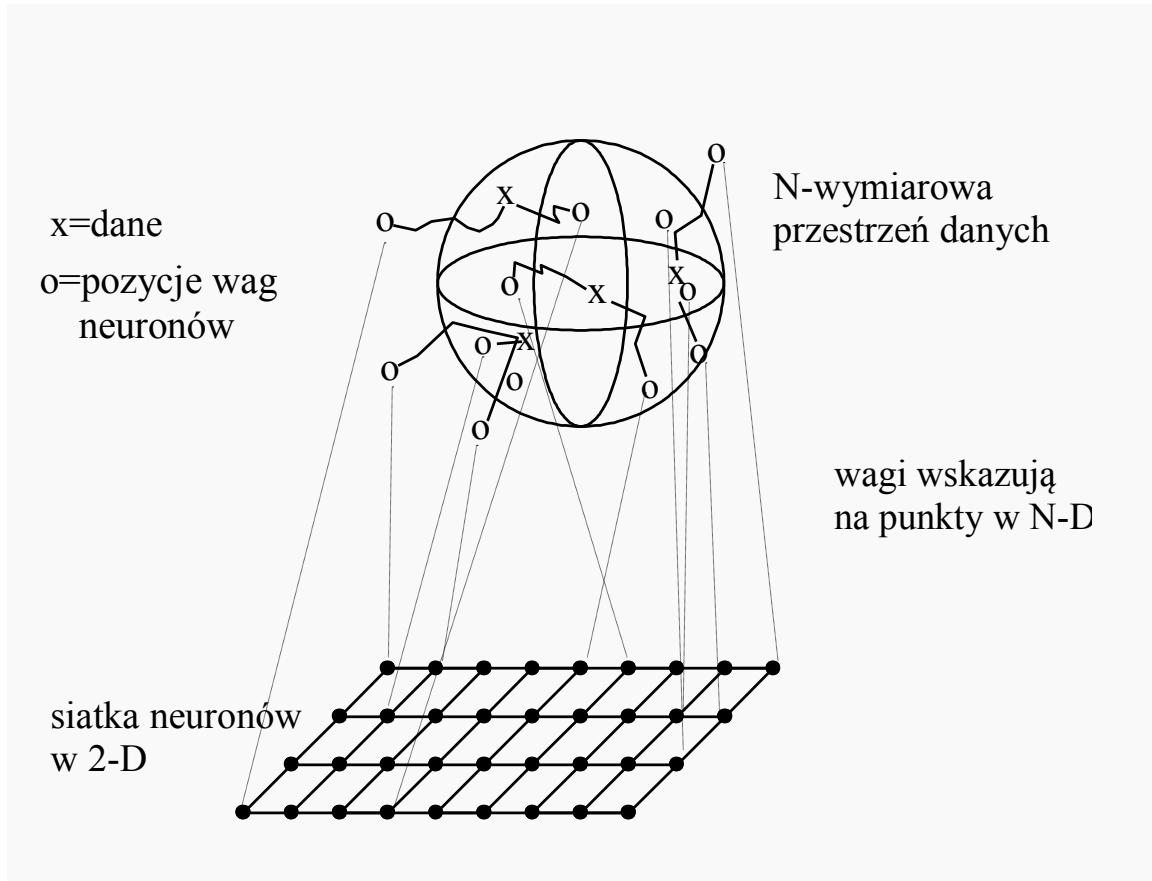
- Podstawą odwzorowania takie uporządkowanie neuronów, takie że położenie zwycięskich neuronów niesie informacje
- Topologia → relacja sąsiedztwa
- Podobne przykłady wejściowe  $\times$  powinny aktywizować sąsiednie neurony
- „Gęstość” wzorców w zredukowanej przestrzeni musi odpowiadać gęstości przykładów w oryginalnej przestrzeni

# Typowa architektura sieci SOM



Wielowymiarowe wejścia  $x$  podane na neurony w warstwie wyjściowej, która może być specjalnie wizualizowana

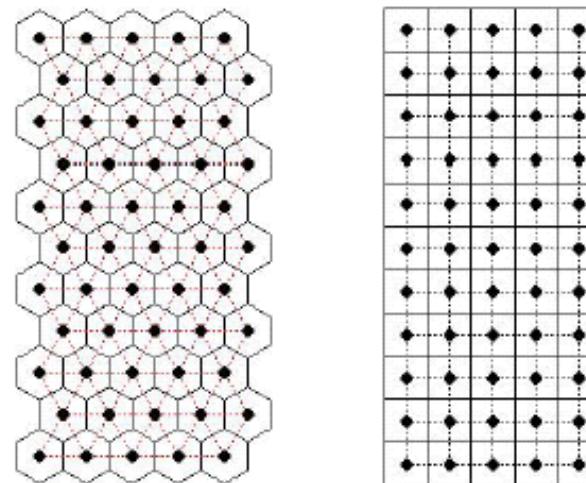
# Idea uczenia sieci SOM



Rysunek za książką A.Żurada

# Typowe topologie sieci SOM

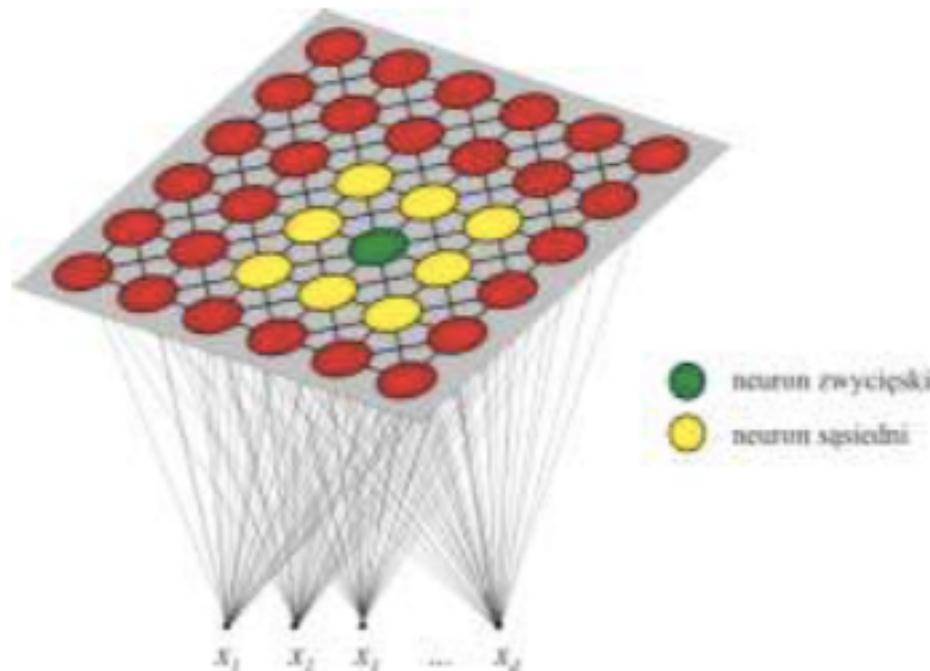
Dwie typowe topologie /odmiana ułożenia neuronów w plaster miodu lub macierz kwadratowa / i sąsiedztwo najbliższych neuronów



Rysunek 7.2: Sąsiedztwo na mapach Kohonena: Neurony ułożone w siatkę (strona lewa) hexagonalną i (strona prawa) prostokątną. {gridhr.JPG}

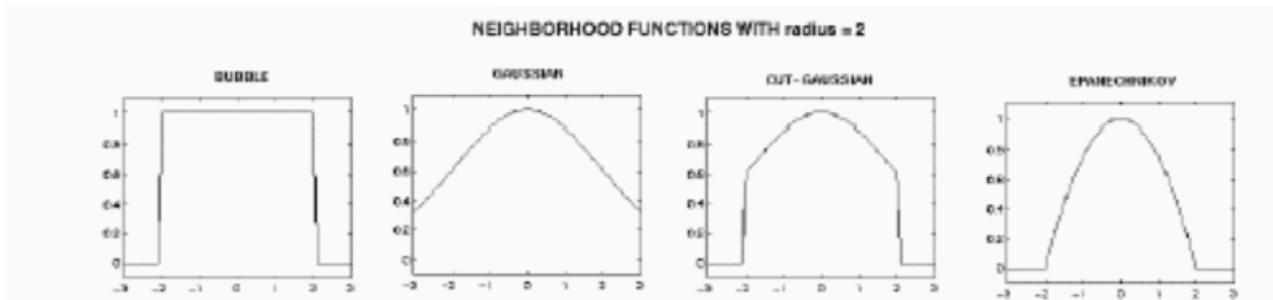
# Identyfikacja neuronu zwycięzcy

Uczymy neuron zwycięzcy  $x_m$  oraz jego najbliższych sąsiadów zgodnie z wybraną funkcją sąsiedztwa

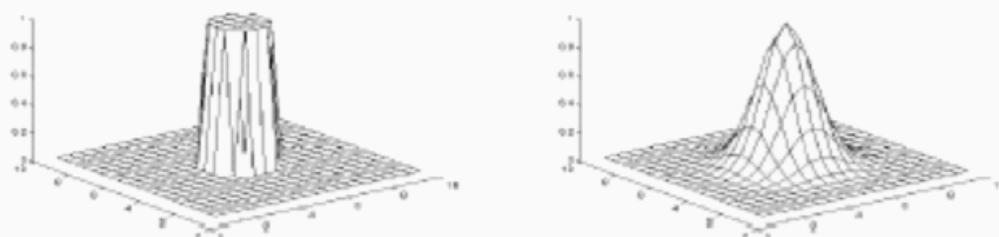


# Funkcje lokalnego sąsiedztwa

Uczymy neuron zwycięzcy  $x_m$  oraz jego najbliższych sąsiadów zgodnie z wybraną funkcją sąsiedztwa



Rysunek 7.3: Jednowymiarowe funkcje sąsiedztwa dla promienia  $R=2$ : bubble, gaussian, cut-gaussian, Epanechnikov {figs7/neigh4.ps}.



Rysunek 7.4: Funkcje sąsiedztwa: bubble i gaussian określone na płaszczyźnie. Funkcja bubble wyróżnia sąsiedztwo w sposób ostry: 1 - tak, 0 - nie; funkcja gaussian w sposób łagodny jako liczbę z przedziału  $(0, 1]$ . {bubble2.ps, gauss2.ps}

# Algorytm SOM

Siatka neuronów  $i = 1 \dots K$  w 1D-3D, każdy neuron z  $N$  wagami.

Neurony z wagami  $\mathbf{W}_i(t) = \{W_{i1} W_{i2} \dots W_{iN}\}$ , wektory  $\mathbf{X} = \{X_1, X_2 \dots X_N\}$ .

$t$  - dyskretny czas; nie ma połączeń pomiędzy neuronami!

1. Inicjalizacja: przypadkowe  $\mathbf{W}_i(0)$  dla wszystkich  $i=1..K$ .  
Funkcja sąsiedztwa  $h(|r-r_c|/\sigma(t), t)$  definiuje wokół neuronu położonego w miejscu  $r_c$  siatki obszar  $O_s(r_c)$ .
2. Oblicz odległości  $d(\mathbf{X}, \mathbf{W})$ , znajdź neuron z wagami  $W_c$  najbardziej podobnymi do  $\mathbf{X}$  (neuron-zwycięzcę).
3. Zmień wagi wszystkich neuronów w sąsiedztwie  $O_s(r_c)$
4. Powoli zmniejszaj siłę  $h_o(t)$  i promień  $\sigma(t)$ .
5. Iteruj aż ustaną zmiany lub wyczerpiesz liczbę epok.

Efekt: podział na wieloboki Voronoia.

# Zastosowanie SOM do analizy alfabetu

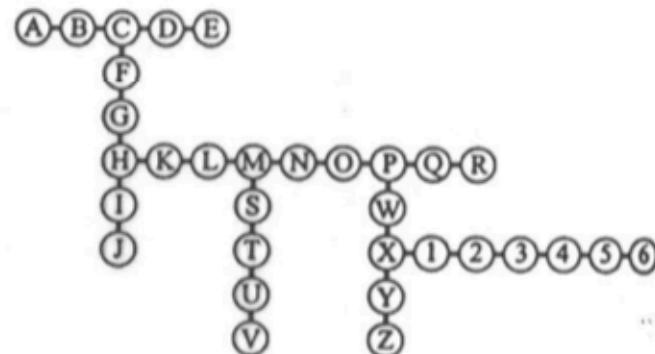
a)

	symbol wektora																															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	1	2	3	4	5	6
składowe																																
$x_1$	1	2	3	4	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3			
$x_2$	0	0	0	0	0	1	2	3	4	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3			
$x_3$	0	0	0	0	0	0	0	0	0	0	1	2	3	4	5	6	7	8	3	3	3	3	6	6	6	6	6	6	6			
$x_4$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	4	1	2	3	4	2	2			
$x_5$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	4		

b)

B	C	D	E	*	Q	R	*	Y	Z
A	*	*	*	*	P	*	*	X	*
*	F	*	N	O	*	W	*	*	1
*	G	*	M	*	*	*	*	2	*
H	K	L	*	T	U	*	3	*	*
*	I	*	*	*	*	*	*	4	*
*	J	*	S	*	*	V	*	5	6

c)



Rys. 7.15. Przykład samoorganizującej się mapy cech: a) zbiór obrazów uczących  
 b) mapa powstała po cyklu uczenia, c) drzewo o minimalnej rozpiętości (z pracy Kohonen (1984), za zgodą ©Springer Verlag)  
 Lecz: właściwe kodowanie symboli – "sąsiedztwo"

# Analiza mapy fonemów języka fińskiego

SOM Toolbox: Intro to SOM by Teuvo Kohonen - Windows Internet Explorer - [Praca w trybie offline]

E:\nowki\RBF\SOM Toolbox Intro to SOM by Teuvo Kohonen.htm

Live Search

Plik Edycja Widok Ulubione Narzędzia Pomoc

SOM Toolbox: Intro to SOM by Teuvo Kohonen

LABORATORY OF COMPUTER AND INFORMATION SCIENCE ADAPTIVE INFORMATICS RESEARCH CENTRE CIS

CIS Toolbox Home About Docs Download Links

## The Self-Organizing Map (SOM)

by Teuvo Kohonen

### Introduction

The SOM is a new, effective software tool for the visualization of high-dimensional data. It converts complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display. As it thereby compresses information while preserving the most important topological and metric relationships of the primary data items on the display, it may also be thought to produce some kind of abstractions. These two aspects, visualization and abstraction, can be utilized in a number of ways in complex tasks such as process analysis, machine perception, control, and communication.

The SOM usually consists of a two-dimensional regular grid of nodes. A model of some observation is associated with each node (cf. Fig. 1).

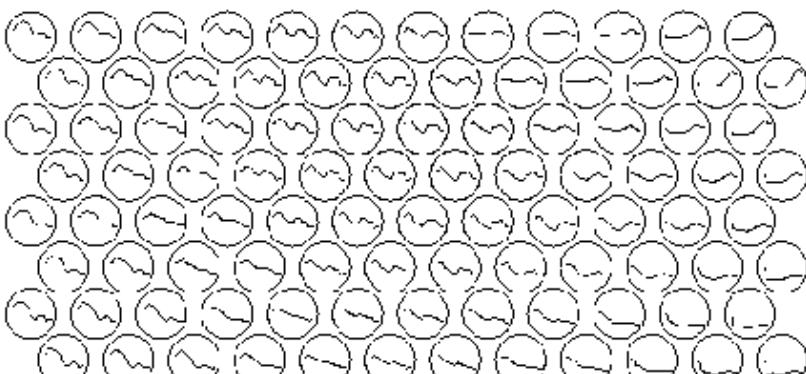
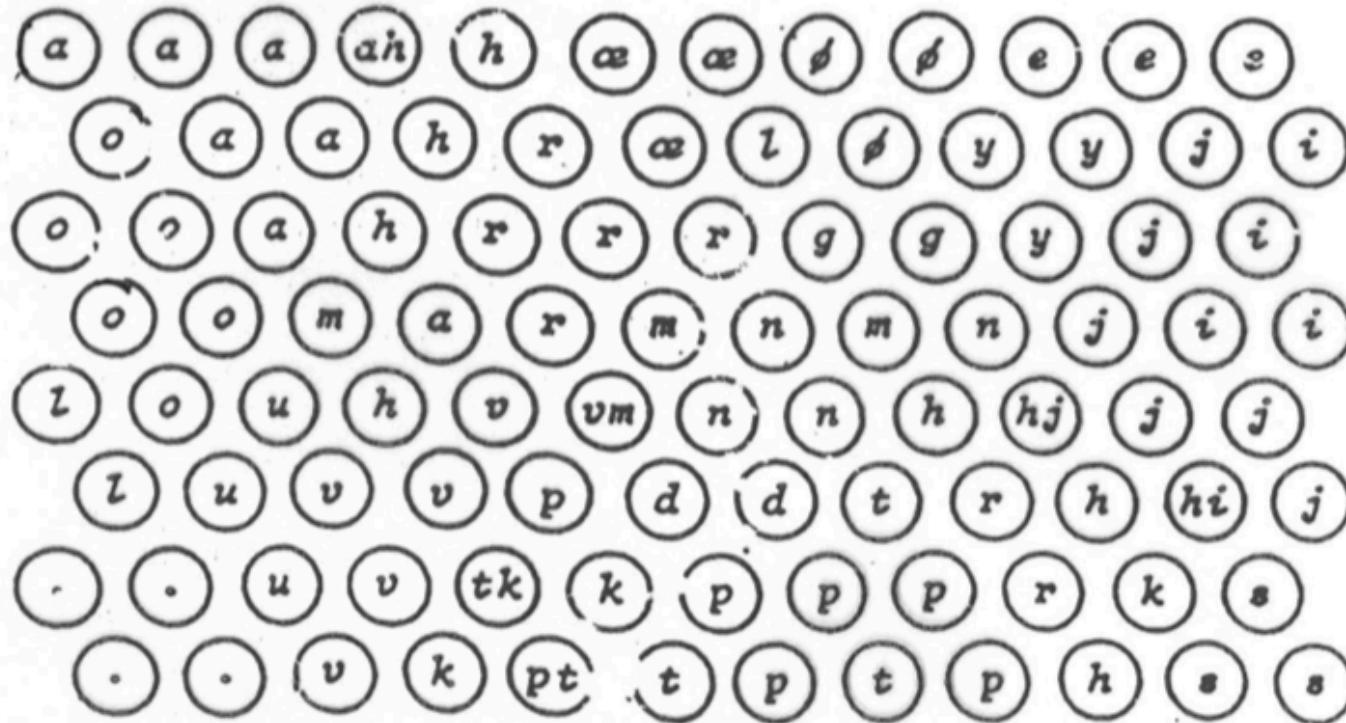


Figure 1: In this exemplary application, each processing element in the hexagonal grid holds a model of a short-time spectrum of natural speech

# Neuronowa mapa - typewriter

- Dane wejściowe: Sygnał mowy (pojedyncze słowa)
- Przetwarzanie danych: 1) 12-bitowy przetwornik analogowo-cyfrowy próbuje sygnał wejściowy co ~10ms; 2) 256-punktowa szybka analiza Fouriera (FFT); 3) wyjście: 15-punktowe spektrum sygnału wejściowego.
- Sieć: SOM: 15 wejść, warstwa wyjściowa:  $8 \times 12 = 96$  neuronów.
- Uczenie: Sygnał mowy (pojedyncze słowa).
- Kalibracja: Po nauczeniu sieci podawano na jej wejścia "wzorcowe" fonemy, opatrując zwycięzców odpowiednimi etykietami. Prawie każdej klasie fonemów odpowiada jeden zwycięzca lub grupa blisko położonych zwycięzców.

# Analiza mapy fonemów języka fińskiego



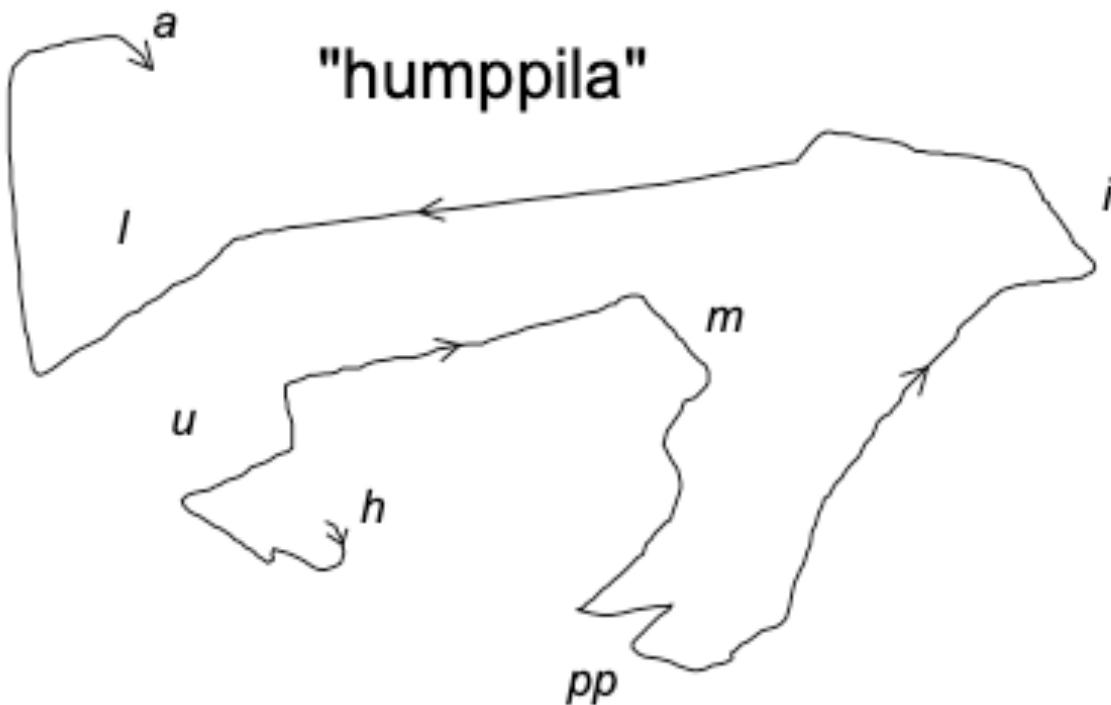
Rys. 7.18. Mapa fonemów języka fińskiego [z pracy Kohonena (1988), za zgodą ©IEEE]

Podobne fonemy – zachowują bliskość  
Możliwość analizy online wypowiedzi i tworzenie obrazów

# SOM neural typewriter

- Podczas wypowiadania słowa notowane jest położenie zwycięzcy, które zmieniając się tworzy obraz - marszrutę. Otrzymana w ten sposób transkrypcja fonetyczna słowa może być dalej przetwarzane w innym systemie
- Rodzaj automatycznej maszyny do pisania, rozpoznaje słowa ich “nie rozumiejąc”
- trafność rozpoznawania fonemów: 92-97%,

"humppila"

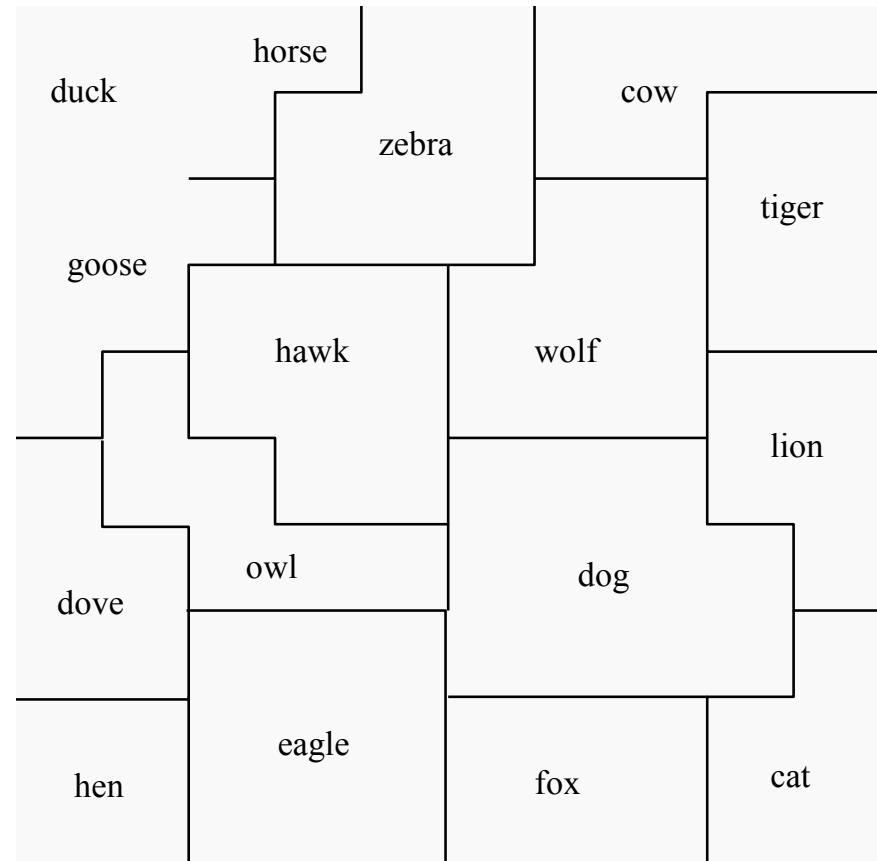
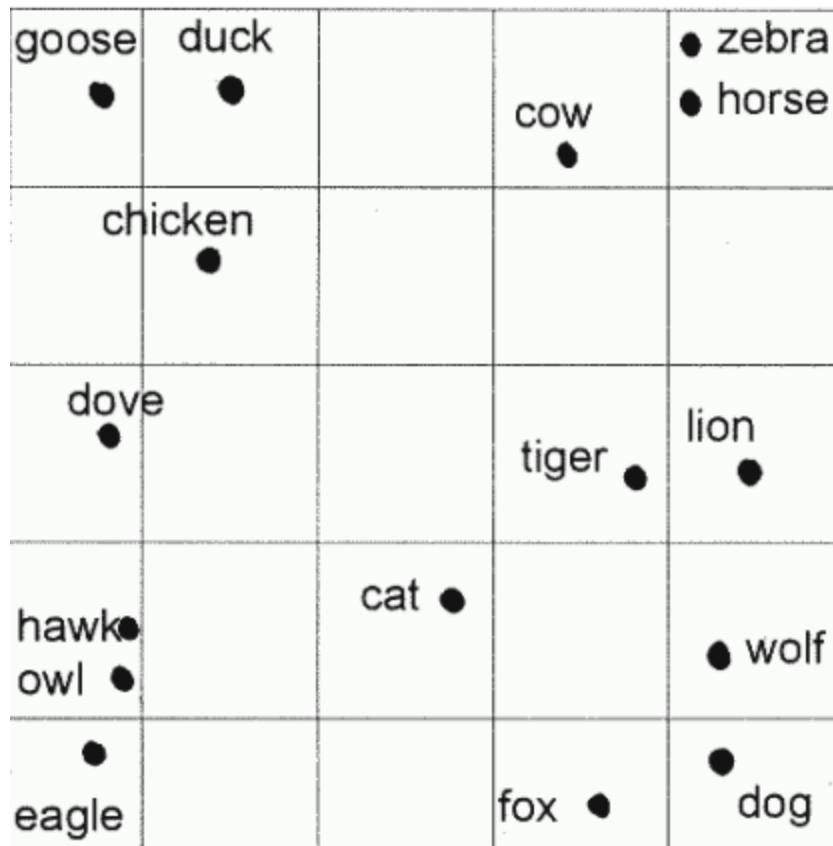


# Przykład mapy semantycznej

		d	d	g		e		w	t	h	z		
animal		o	h	u	o	a	g	f	d	o	c	g	i
		v	e	c	s	w	w	l	o	o	l	a	e
		e	n	k	e	l	k	e	x	g	f	t	r
is	small	1	1	1	1	1	1	0	0	0	1	0	0
	medium	0	0	0	0	0	0	1	1	1	0	0	0
	big	0	0	0	0	0	0	0	0	0	1	1	1
has	2 legs	1	1	1	1	1	1	1	0	0	0	0	0
	4 legs	0	0	0	0	0	0	0	1	1	1	1	1
	hair	0	0	0	0	0	0	0	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	1
	mane	0	0	0	0	0	0	0	0	1	0	0	1
	feathers	1	1	1	1	1	1	1	0	0	0	0	0
likes	hunt	0	0	0	0	1	1	1	1	0	1	1	1
	run	0	0	0	0	0	0	0	1	1	0	1	1
	to fly	1	0	0	1	1	1	1	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0

SOM was used on such data by Ritter and Kohonen 1989,  
MDS by Naud & Duch (1996) = dalsza analiza w wykładzie W.Ducha nt.  
Wizualizacji wielowymiarowych -> patrz strona WWW

# Porównanie map MDS & SOM



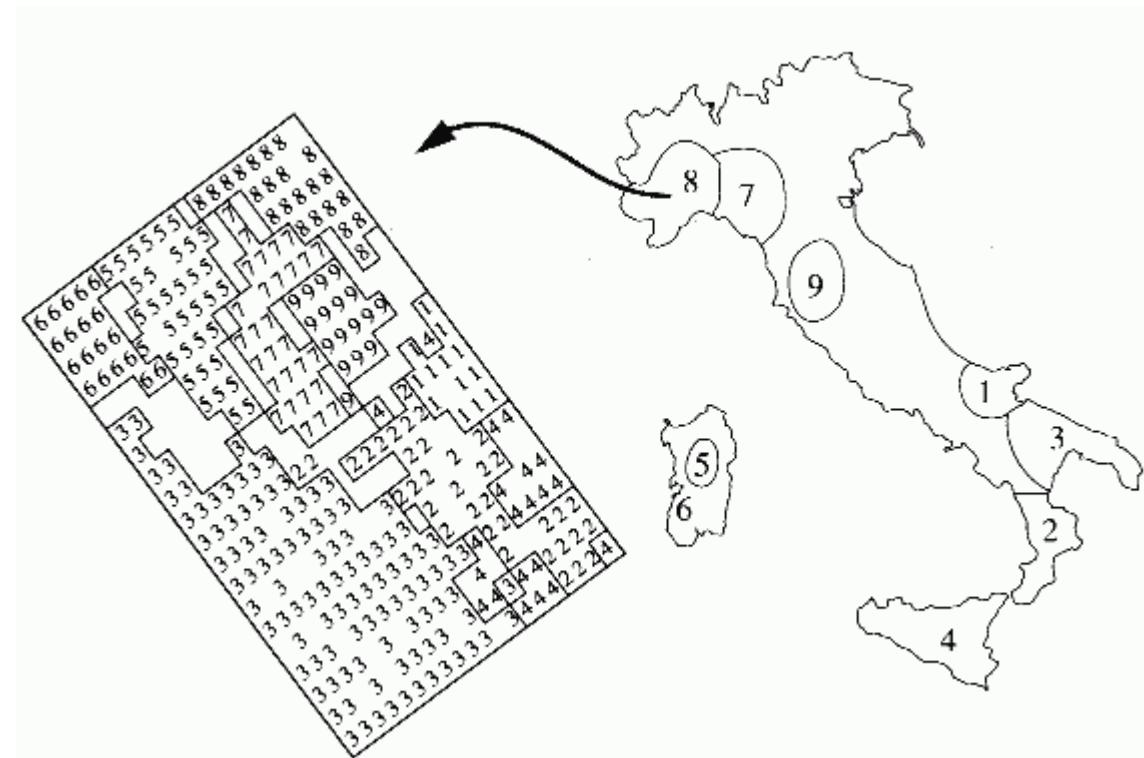
Użycie SOM i MDS na danych opisujących zwierzęta – zwróć uwagę na bliskość semantyczną rodzin zwierząc, np. kotowate po prawej stronie mapy; ptaki tworzą także podobne grupy

# Analiza jakości oliwy w różnych rejonach Włoch

572 próbek oliwy pochodzącej z 9 prowincji Włoch

Badania laboratoryjne – 8 podstawowych składników  
SOM 20 x 20 neuronów,  
Odwzorowanie 8D => 2D.  
Kalibracja zwycięskich neuronów z etykietą prowincji

Analiza sąsiedztwa i bliskości innych rejonów



Zauważ że przekształcenie zachowało nieznane uczeniu sąsiedztwo topograficzne , tylko prowincja #3 jest lekko rozproszona

# Ocena wpływu cech na grupy

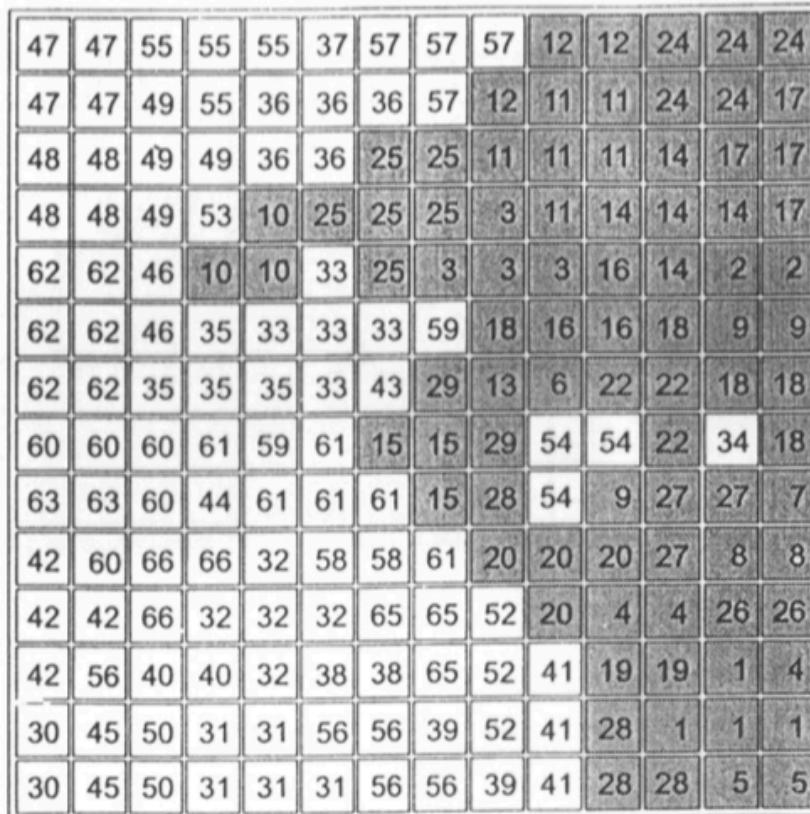
Analizie poddano dane dotyczące 66 banków hiszpańskich, przy czym blisko połowa z nich (29) zbankrutowała w czasie kryzysu. Do reprezentacji profilu każdego z banków, na podstawie analiz statystycznych, wybrano dziewięć wskaźników finansowych (tablica 5.7). Ich wartości zostały, podobnie jak w poprzednim przypadku, znormalizowane do zerowej średniej i jednostkowej wariancji.

Tablica 5.7. Wskaźniki finansowe wykorzystane w badaniu banków [37]

Symbol	Opis
R1	Aktywa bieżące do aktywów całkowitych
R2	(Aktywa bieżące - środki pieniężne) do aktywów całkowitych
R3	Aktywa bieżące do zobowiązań
R4	Rezerwy do zobowiązań
R5	Przychody netto do aktywów
R6	Przychody netto do kapitału obrotowego
R7	Przychody netto do zobowiązań
R8	Koszty sprzedaży do sprzedaży
R9	Przepływy pieniężne do zobowiązań

Do analizy danych wykorzystano sieć 196 neuronów, uformowanych w macierz

# Analiza banków hiszpańskich



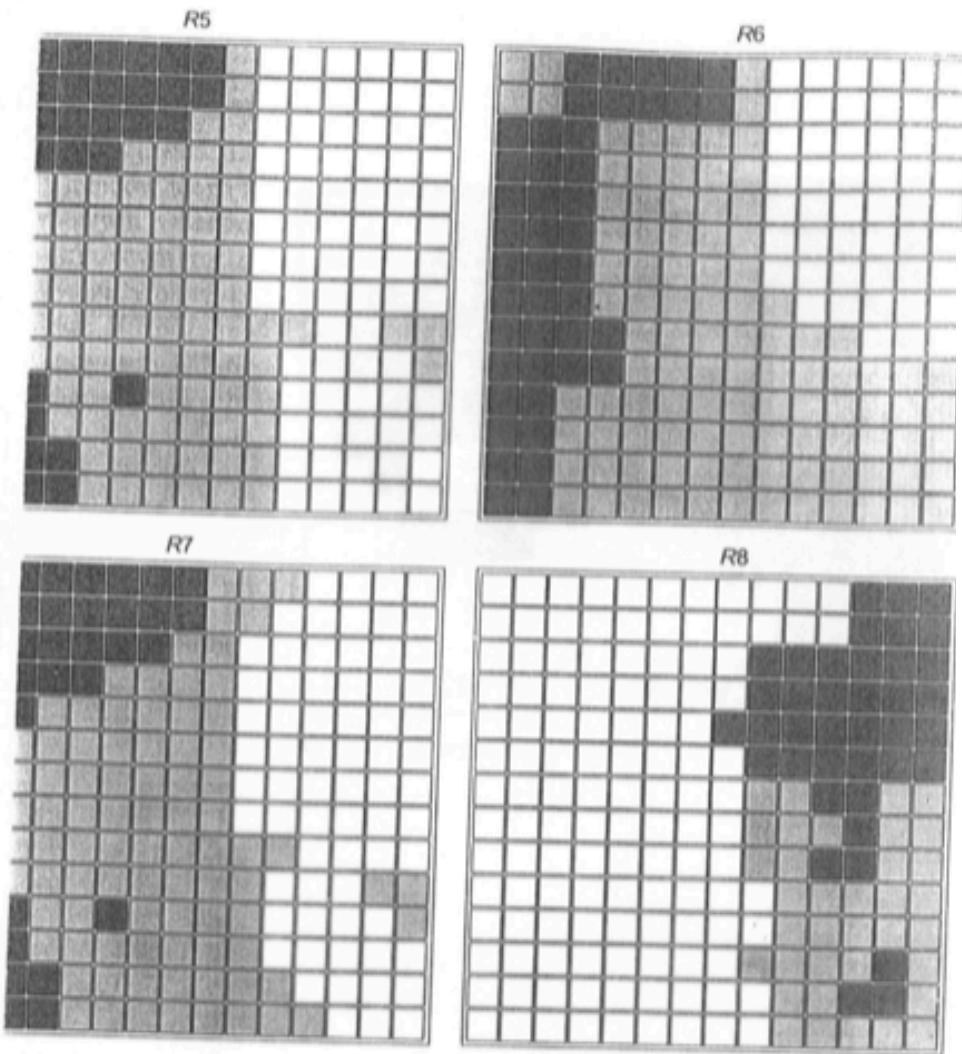
obszar bezpieczny



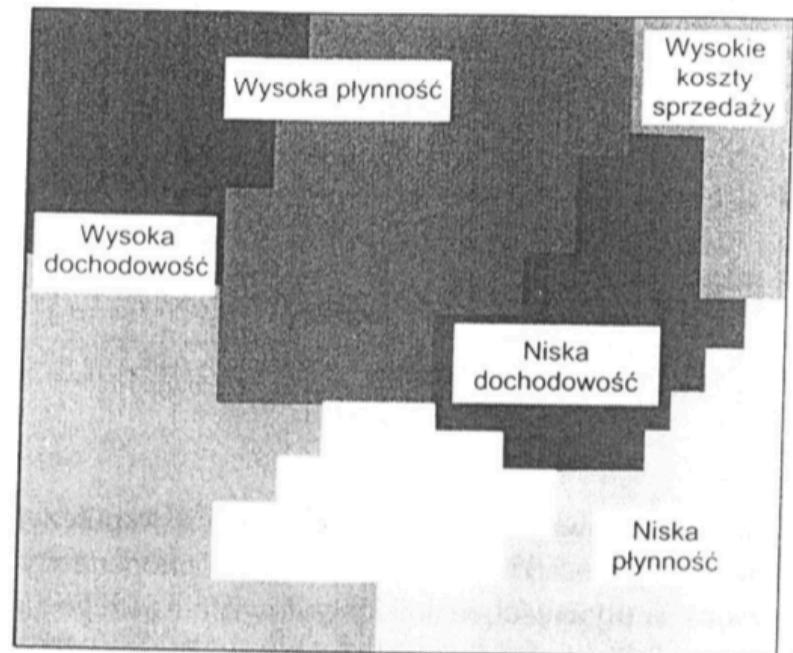
obszar bankructwa

Rys. 5.34. Mapa cech analizowanych banków [37]

# Indywidualna analiza znaczenia cech



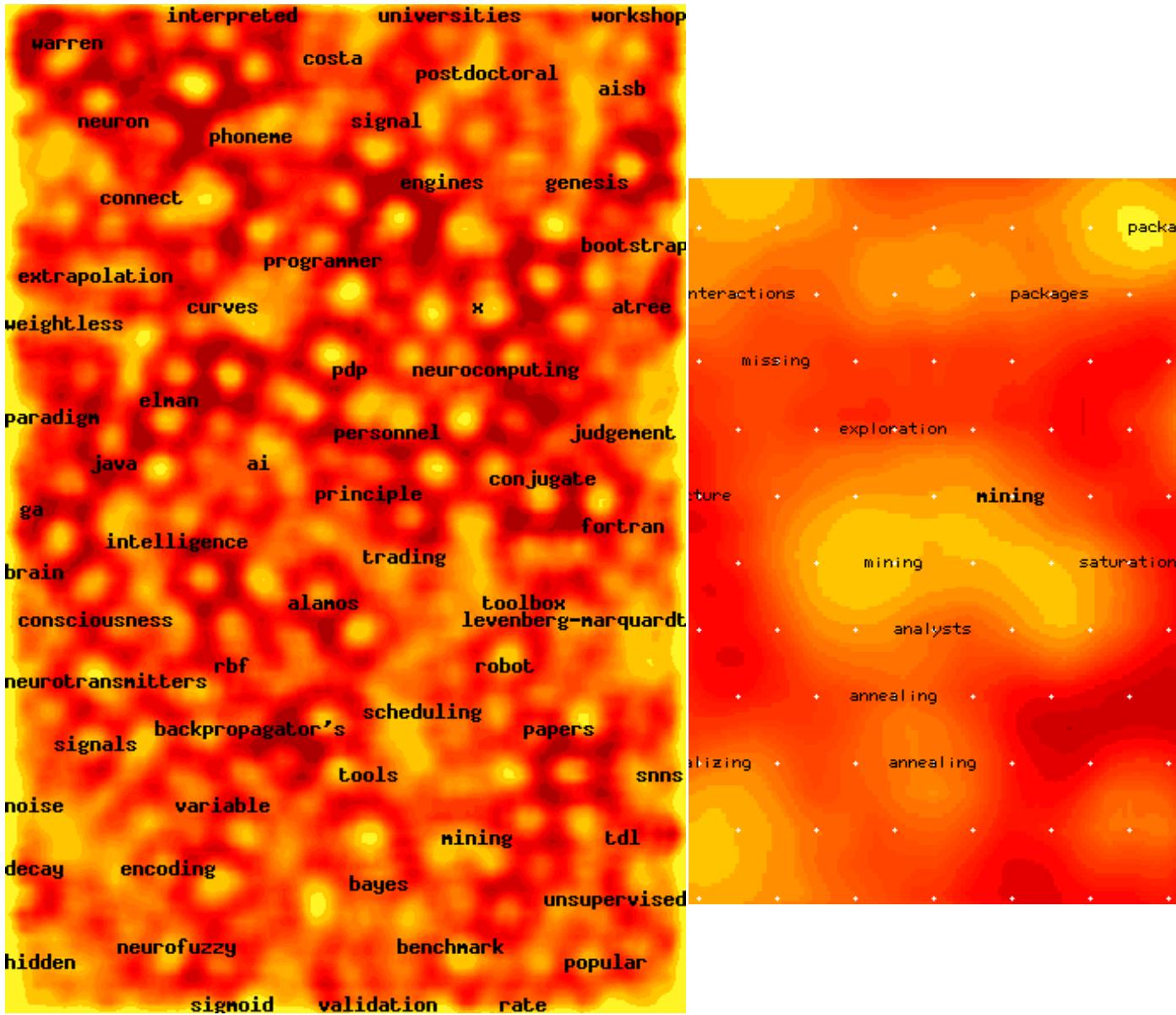
Rys. 5.36. Mapy wag sieci Kohonena dla wskaźników od R5 do R8 [37]



Rys. 5.38. Regiony na mapie cech wyznaczone przez mapy wag [37]

# Web SOM

- SOM do pogrupowania 12088 artykułów z internetu
- Aplikacja do stopniowego (zoom) przeglądania zawartości skupisk
- Spójrz na [websom.hut.fi](http://websom.hut.fi) Web page



# Odnośniki do literatury

SOM intensywny rozwój od lat 80 poprzedniego wieku

- Wiele różnych zastosowań

Analiza skupień – obszerna literatura:

- Kohonen, Teuvo; Honkela, Timo (2007). "Kohonen Network". Scholarpedia WWW
- Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps"
- T. Kohonen, Self-Organization and Associative Memory. Springer, Berlin, 1984
- Żurada J., Barski M., Jędruch W.: Sztuczne sieci neuronowe. PWN 1996.
- Stapor K. Automatyczna klasyfikacja obiektów. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2005

# **Pytanie i komentarze?**

Dalszy kontakt:

[jerzy.stefanowski@cs.put.poznan.pl](mailto:jerzy.stefanowski@cs.put.poznan.pl)

<http://www.cs.put.poznan.pl/jstefanowski/>



**Fundusze  
Europejskie**  
Polska Cyfrowa



**Rzeczpospolita  
Polska**

**Unia Europejska**  
Europejski Fundusz  
Rozwoju Regionalnego

