Uczenie nienadzorowane

algorytmy grupowania wykład 12 cz II

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

- Rozszerzenia klasycznych algorytmów grupowania
  - Algorytm k-średnich
    - K-medoid, PAM, …
  - Algorytmy hierarchiczne
    - Podstawowe AHC
    - BIRCH
- Algorytymy gęstościowe
  - DBSCAN
- Podejścia wykorzystujące modele statystyczne
  - Algorytm mieszanin rozkładów (EM)
- Inne algorytmy grupowania dla trudnych danych
- Ocena jakości grupowania
- Podsumowanie

Grupowanie z wykorzystaniem modeli prawdopodobieństwa

- Podejścia oparte na założeniu, że danych są generowanie w wyniku realizacji pewnego procesu statystycznego
- Zakłada się pewien model rozkładu prawdopodobieństwa występowanie obserwacji
- Każdemu potencjalnemu **skupisku** odpowiada **model**, w postępowaniu (algorytmie) weryfikuje się stopień dobrego dopasowania oryginalnych danych do przyjętego modelu
- Celem grupowania jest znalezienie zbioru (mieszaniny) modeli (rozkładów) opisujących skupiska oraz estymacja parametrów tych modeli
- Obiekty przydziela się do skupisk zgodnie ze sparametryzowanymi modelami i zasadą klasyfikacji Bayesowskiej

Mieszaniny rozkładów (1)

- Typowe podejście do grupowania wykorzystującego modele statystyczne – przyjęcie założenia mieszaniny wielowymiarowych rozkładów prawdopodobieństwa (przykład algorytm EM)
- Założenia: Podział danych X={x_1,…,x_m} na *K* skupisk jest równoznaczny z łącznym rozkładem prawdopodobieństwa zbudowanym z K składowych rozkładów o parametrach θj. Łączny rozkład ze zbiorem parametrów θ={θ_1,…,θ_K} :

$$P(x \mid \theta) = \sum_{j=1}^{K} p(j) \cdot p_j(x \mid \theta_j)$$

- gdzie p(j) jest prawdopobieństwem przydziału obiektu x do j-tego skupiska (modelu); $\Sigma p(j)=1$

Mieszaniny rozkładów (2)

- Łączny rozkład prawdopodobieństwa dla obiektu x

$$P(x \mid \theta) = \sum_{j=1}^{K} p(j) \cdot p_j(x \mid \theta_j)$$

- Interpretacja statystyczna (modele generatywne) - przykłady (obiekty) uczące otrzymywane są dwustopniowo:
  - Losowanie jednego z *K* źródeł − które generuje przykłady z swojej grupy : *p*(j) prawdopodobieństwo wylosowania j-tego źródła
  - Sam przykład jest generowany zgodnie z funkcją gęstości prawdopodobieństwa $f_j(\mathbf{x}|\theta j)$ wynikającą z przyjętego modelu

Mieszaniny rozkładów (3)

- Mając rozkład prawdopodobieństwa dla obiektu x

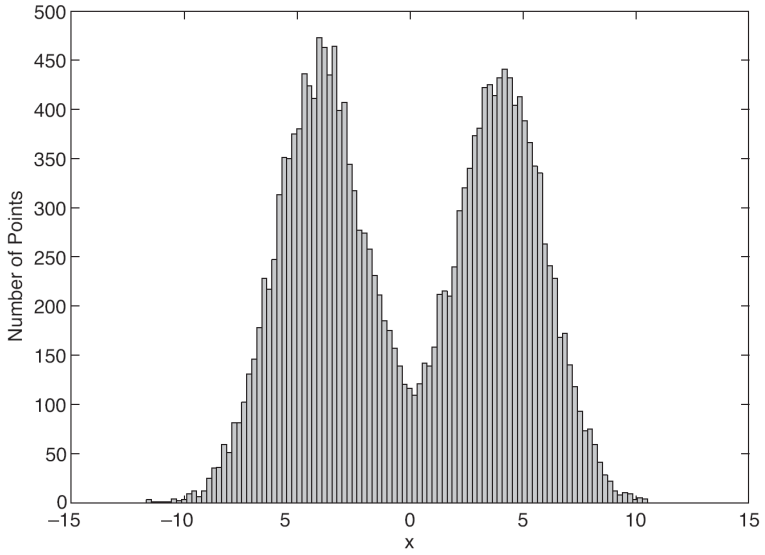
$$P(x \mid \theta) = \sum_{j=1}^{K} p(j) \cdot p_j(x \mid \theta_j)$$

- Jeśli wszystkie obiekty w X są generowanie niezależnie, to łączne prawdopodobieństwo otrzymania / wygenerowania obiektów $X=\{x_1,\ldots,x_m\}$ jest iloczynem prawdopodobieństw dla indywidualnych obiektów

$$P(X \mid \theta) = \prod_{i=1}^{m} P(x_i \mid \theta) = \prod_{i=1}^{m} \sum_{j=1}^{K} p(j) \cdot p_j(x \mid \theta_j)$$

- Najczęściej zakłada się rozkłady normalne. Nazywa się to mieszaniną rozkładów Gaussowskich (ang. mixture of Gaussians)

Przykład analizy mieszaniny rozkładów normalnych

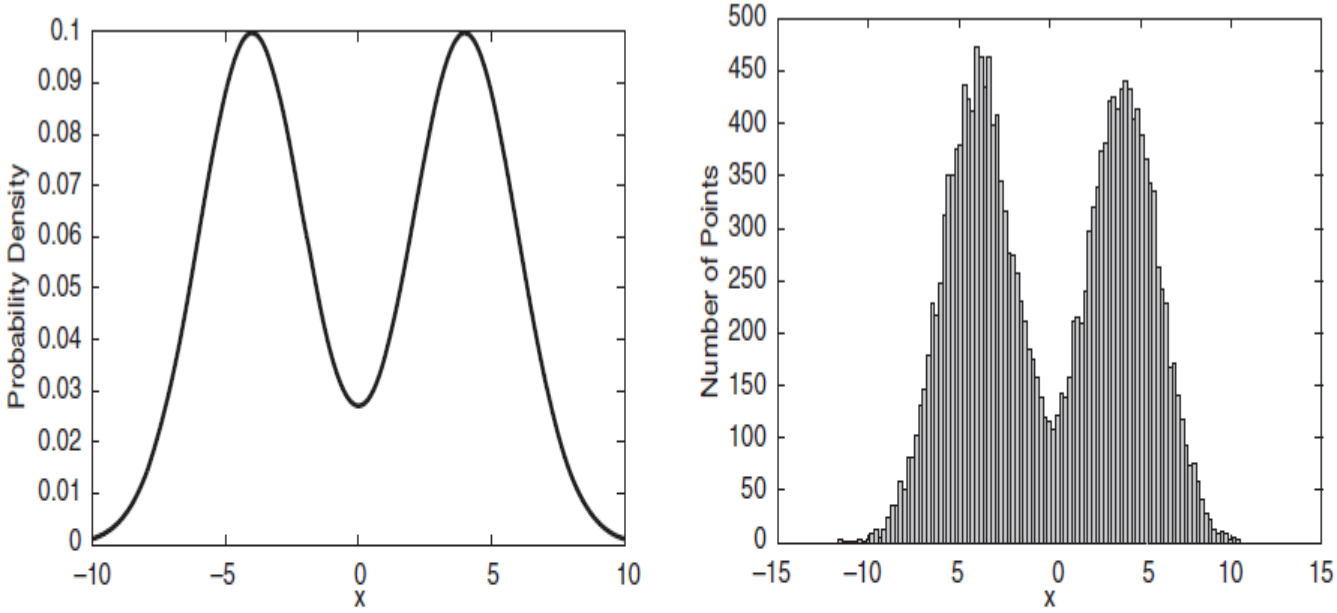


- **Prosty przykład** [Kumar et al]: rozważ modelowanie obiektów tworzących histogram − patrz rys.
- Model może być mieszaniną dwóch rozkładów normalnych (każdy sparametryzowany wartością oczekiwaną oraz odchyleniem standardowym σ - patrz wzór

$$prob(x_i|\Theta) = \frac{1}{\sqrt{2\pi}\sigma}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Jeżeli estymuje oba parametry – to przy założeniu równych prawdopodobieństw komponentów p(1)=p(2)=0.5:
  - Można w pełni opisać oba skupiska
  - Można obliczyć prawdopodobieństwo przydziału dowolnego obiektu do skupiska 1 oraz skupiska 2
  - Przypisać obiekt do bardziej prawdopodobnego skupiska

Model – mieszanina gausowska



(a) Probability density function for the mixture model.

(b) 20,000 points generated from the mixture model.

Figure 8.2. Mixture model consisting of two normal distributions with means of -4 and 4, respectively. Both distributions have a standard deviation of 2.

Rysunek za książka Tan, Kumar Data mining

Złożenie rozkładów normalnych

W przykładzie

Prob = $p(1)N(\mu_1,\sigma_1)+p(2)N(\mu_2,\sigma_2)$

Lecz ogólnie rozważamy rozkłady wielowymiarowe zależne od wektora μ oraz macierzy kowariancji Σ

Funkcja gęstości n-wymiarowego rozkładu normalnego wektora losowego X c
jest wzorem:

$$f_{\mu,\Sigma}(X) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right).$$

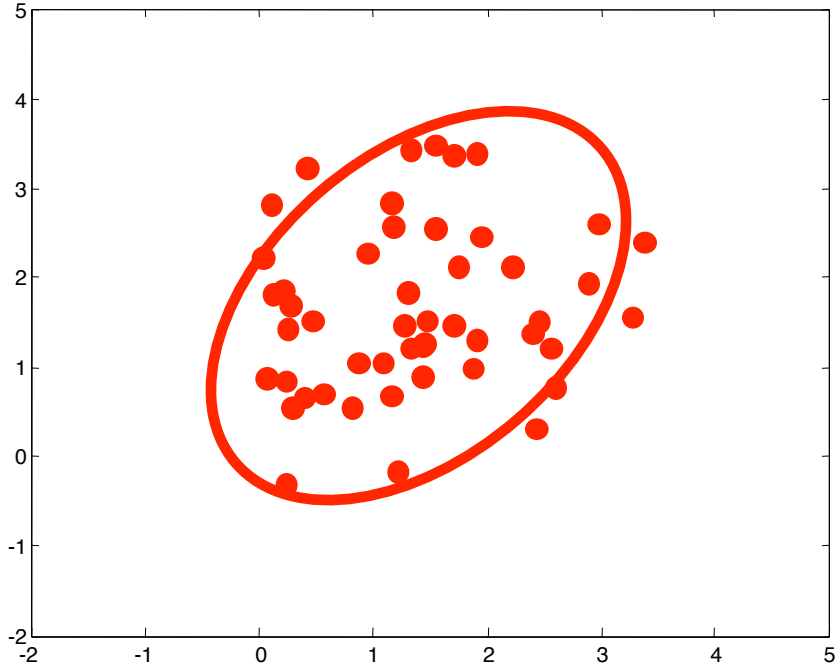
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Estymator macierzy kowariancji o największej wiarygodności:

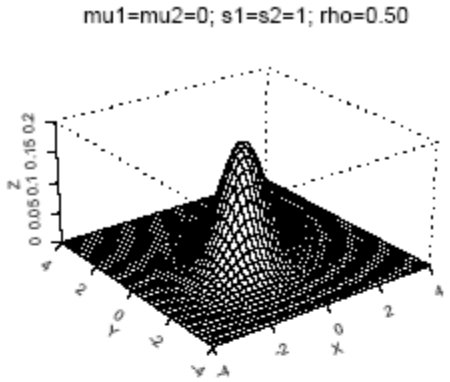
$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(X_i-\hat{\mu})(X_i-\hat{\mu})^T.$$

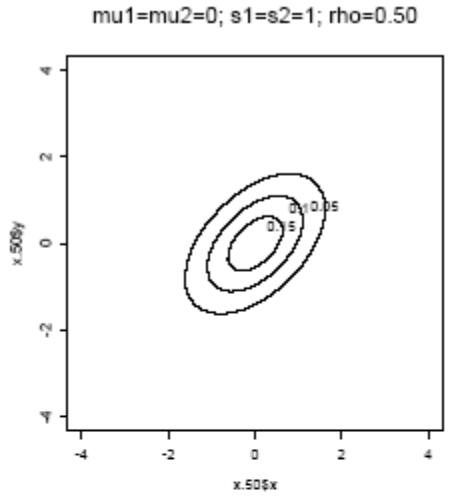
Modelowanie dwuwymiarowego rozkładu normalnego (d=2)

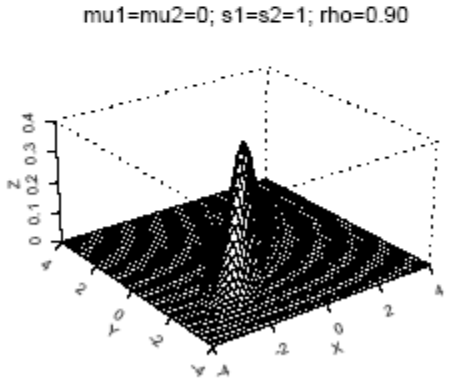
$$\mathcal{N}(\underline{x}\ ;\ \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$

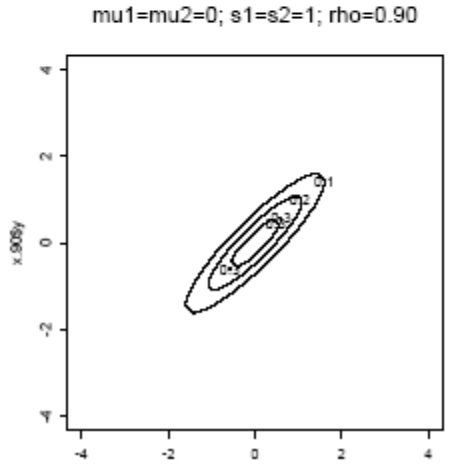


Dwuwymiarowe rozkłady prawdopodobieństwa









Więcej w literaturze nt. rozkładów prawdopodobieństwa

Funkcja wiarygodności i MLE

- Dla wybranego modelu statystycznych danych – należy oszacować jego parametry Θ na podstawie *m* prób (tutaj przykładów uczących)
- **Metoda największej wiarygodności** (ang. Maximum likelihood method)
  - Na podstawie wybranego rozkładu określamy prawdopodobieństwo a posteriori obiektu x_i
  - Parametry rozkładu dobiera się tak, aby maksymalizować prawdopodobieństwa a posteriori rozkładu dla obiektów z danych uczących $X=\{x_1,\ldots,x_m\}$
- Definiujemy funkcję wiarygodności (L – z ang. **Likelihood function**)

Funkcja wiarygodności i MLE

- Prawd. a posteriori obiektów z X − iloczyn indywidulanych prawd. dla obiektów
- Definiujemy funkcję wiarygodności (L – z ang. Likelihood function)

$$L(X;\theta) = \prod_{i=1}^{m} p(x_i \mid \theta) = \prod_{i=1}^{m} \prod_{j=1}^{k} p(x_i \mid \theta) \cdot p(j)$$

- Cel − wybierz parametry Θ maksymalizujące powyższą funkcję wiarygodności
- Najczęściej wykorzystuje się logarytmiczne przekształcenie funkcji wiarygodności LogL(X;θ)

$$LogL(X;\theta) = \sum_{i=1}^{m} \log p(x_i \mid \theta)$$

Funkcja wiarygodności dla N(μ,Σ)

W algorytmie EM wykorzystuje się mieszany rozkładów Gaussowskich – rozważmy przykład jednowymiarowy

$$L(X;\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

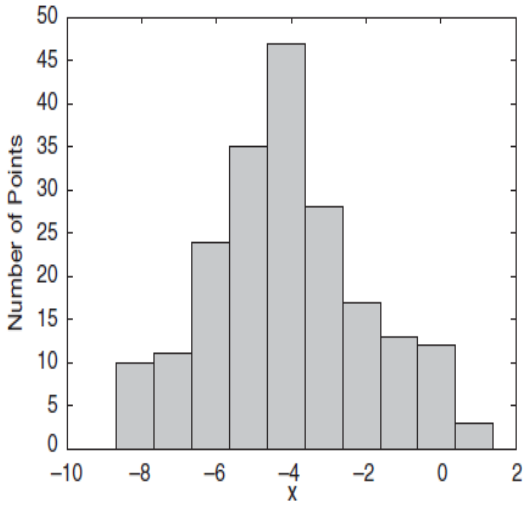
Logarytmiczna funkcja wiarygodności

$$LogL(X;\theta) = -\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - 0.5m\log 2\pi - m\log\sigma$$

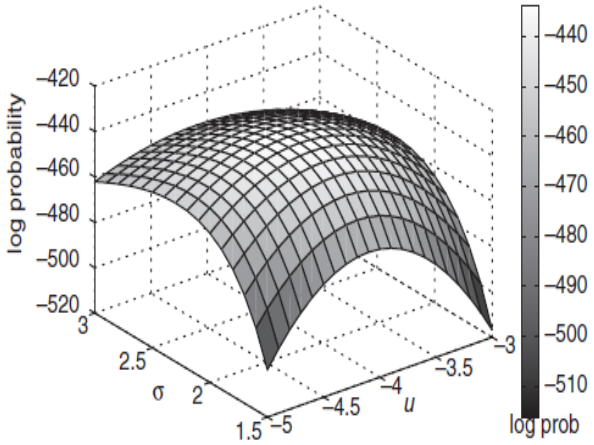
Przykład ilustracyjny kolejny slajd − prosty wybór

Procedura wyznaczania estymatorów dla rozkładów ciągłych − pochodne cząstkowe logarytmicznej funkcji L względem nieznanych parametrów

MLE – poszukiwanie parametrów maksymalizujących LogL(X;Θ)



(a) Histogram of 200 points from a Gaussian distribution.



(b) Log likelihood plot of the 200 points for different values of the mean and standard deviation.

Figure 8.3. 200 points from a Gaussian distribution and their log probability for different parameter values.

Analiza wykresu : μ=-4.1 oraz θ=2.1

Algorytm EM

EM nazwa ang. **Expectation-Maximization**

- Inicjalizacja początkowych wartości parametrów rozkładów

Repeat

1. (*Expectation step*) Dla każdego obiektu z X oblicz jego przynależność do skupiska (rozkładu)
2. (*Maximization step*) Użyj tych prawdopodobieństw do iteracyjnej aktualizacji parametrów rozkładu

Until (zmiany parametrów nie są znaczące)

Zapis algorytmu EM

Algorithm 9.2 EM algorithm.

1: Select an initial set of model parameters.
(As with K-means, this can be done randomly or in a variety of ways.)
2: **repeat**
3: **Expectation Step** For each object, calculate the probability that each object belongs to each distribution, i.e., calculate $prob(distribution\ j|\mathbf{x}_i, \Theta)$.
4: **Maximization Step** Given the probabilities from the expectation step, find the new estimates of the parameters that maximize the expected likelihood.
5: **until** The parameters do not change.
(Alternatively, stop if the change in the parameters is below a specified threshold.)

Krok oczekiwanej przynależności do skupiska (expectation)

Mając oszacowanie parametrów j-tego rozkładu normalnego μ_j oraz Σ_j (σ_j) oraz wstępne p(j) oblicz przynależności każdego obiektu z X do odpowiedniego skupiska (j=1,..,K)

$$t_{ij}^{(h)} = \frac{p_j^{(h)} \cdot p(x_i \mid \mu_j^{(h)}, \Sigma_j^{(h)})}{\sum_{l=1}^{K} p_l^{(h)} \cdot p(x_i \mid \mu_l^{(h)}, \Sigma_l^{(h)})}$$

gdzie t_{ij}^{h} stopień przynależności obiektu xi do j-tego skupiska w h-tej iteracji (prawdopodobieństwa z reguły Bayesowskiej)

W przypadku rozkładu normalnego używamy funkcji gęstości prawdopodobieństwa f jako p(x|θ)

Krok estymacji parametrów maksymalizujących logL

Na podstawie wyliczonych przynależności t_{ij}^{h} poszuje się nowych estymatorów parametrów rozkładów (MLE – maksymalizujących log funkcji wiarygodności)

$$p_j^{(h+1)} = \frac{1}{m} \sum_{i=1}^{m} t_{ij}^{(h)}$$

$$\mu_j^{(h+1)} = \frac{\sum_{i=1}^{m} t_{ij}^{(h)} \cdot x_i}{m \cdot p_j^{(h+1)}}$$

$$\Sigma_j^{(h+1)} = \frac{\sum_{i=1}^{m} t_{ij}^{(h)} \cdot (x_i - \mu_j^{(h)}) \cdot (x_i - \mu_j^{(h)})^T}{m \cdot p_j^{(h+1)}}$$

Expectation-Maximization

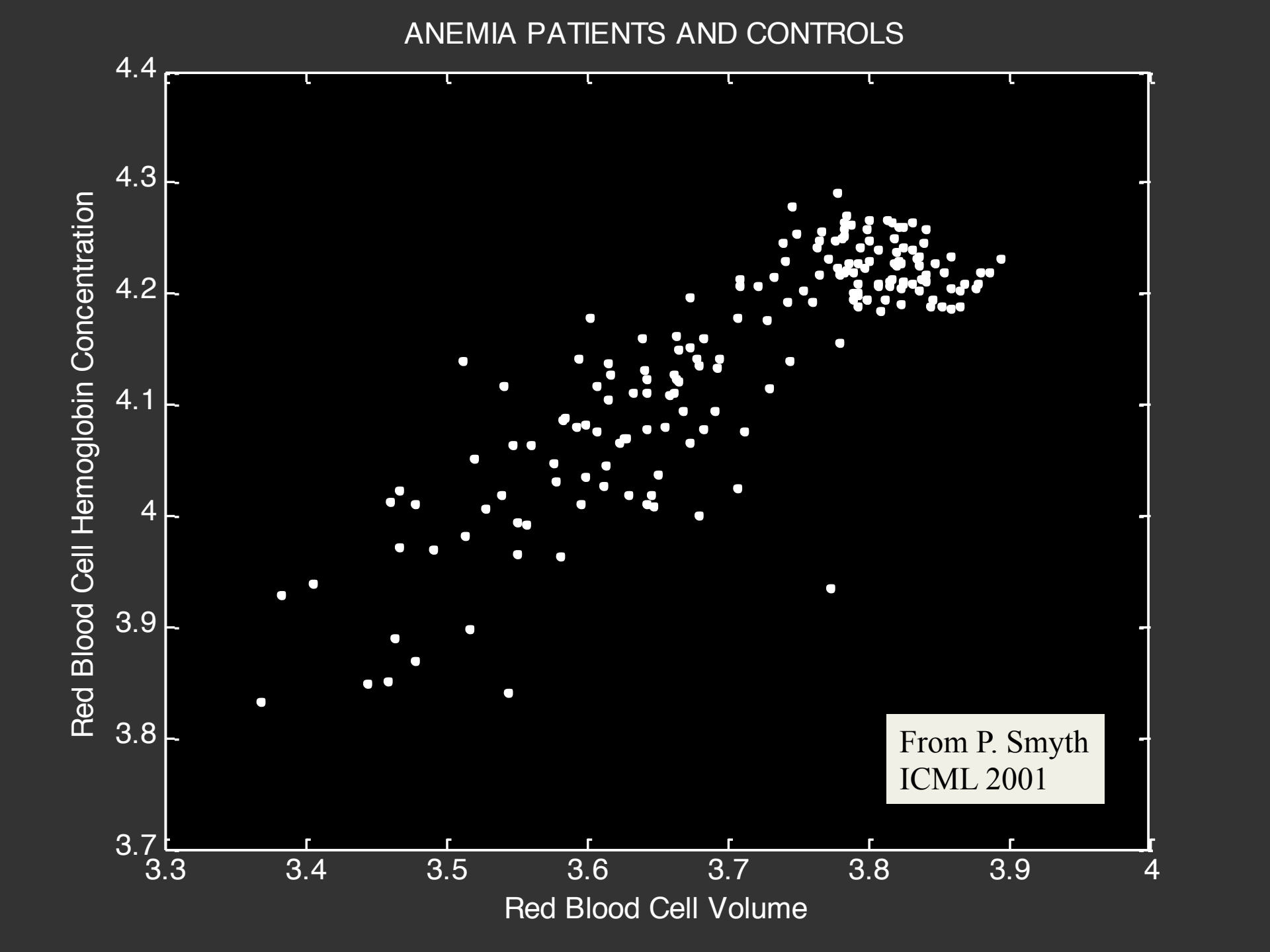
- Kolejne iteracje algorytmu powinny polepszać (maksymalizować) oszacowanie log funkcji wiarygodności (the log-likelihood L) modeli
- Iteruj procedurę aż do zbieżności (warunku zatrzymania)

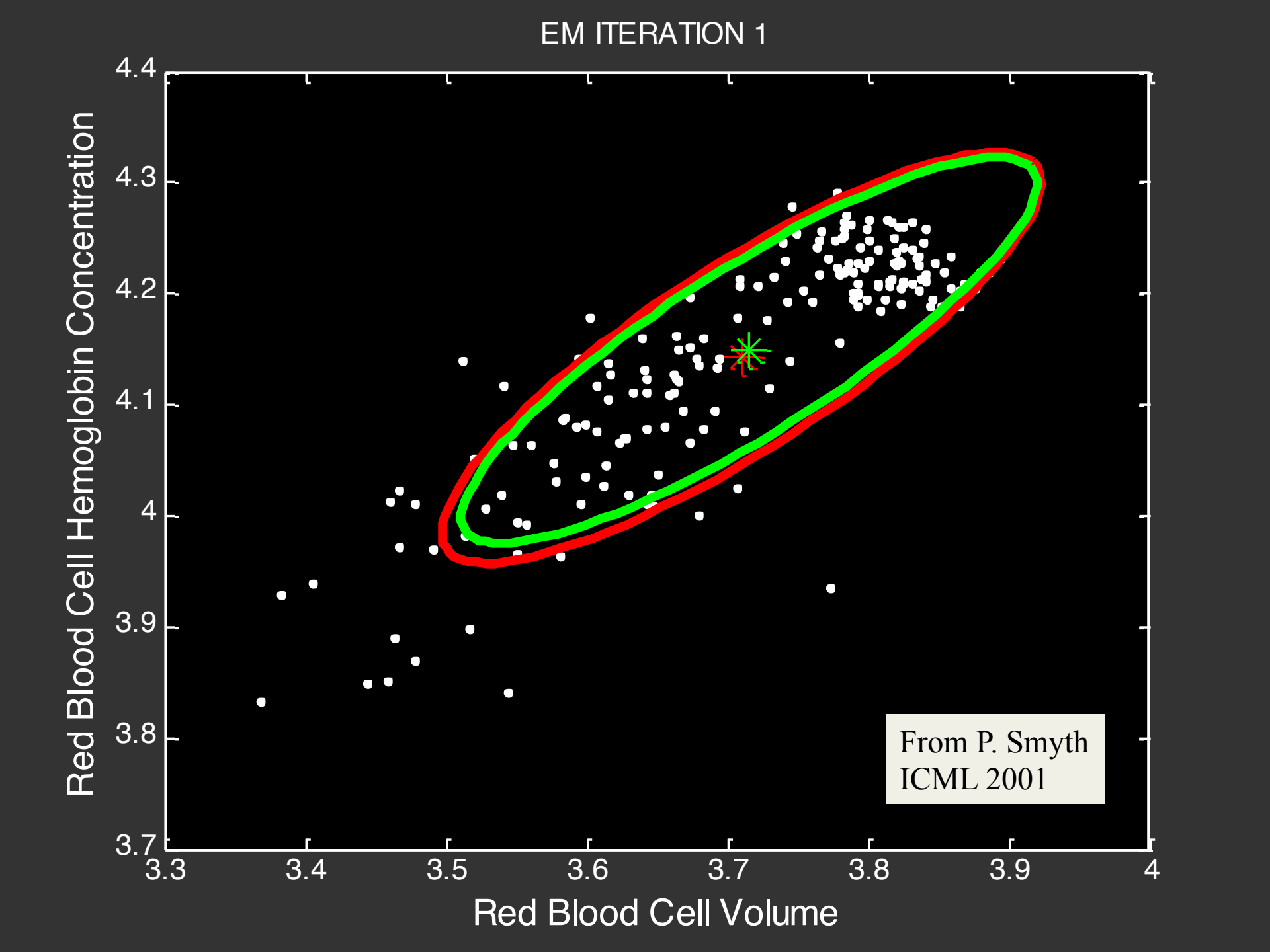
Możliwość uwzględniania niekompletnych danych

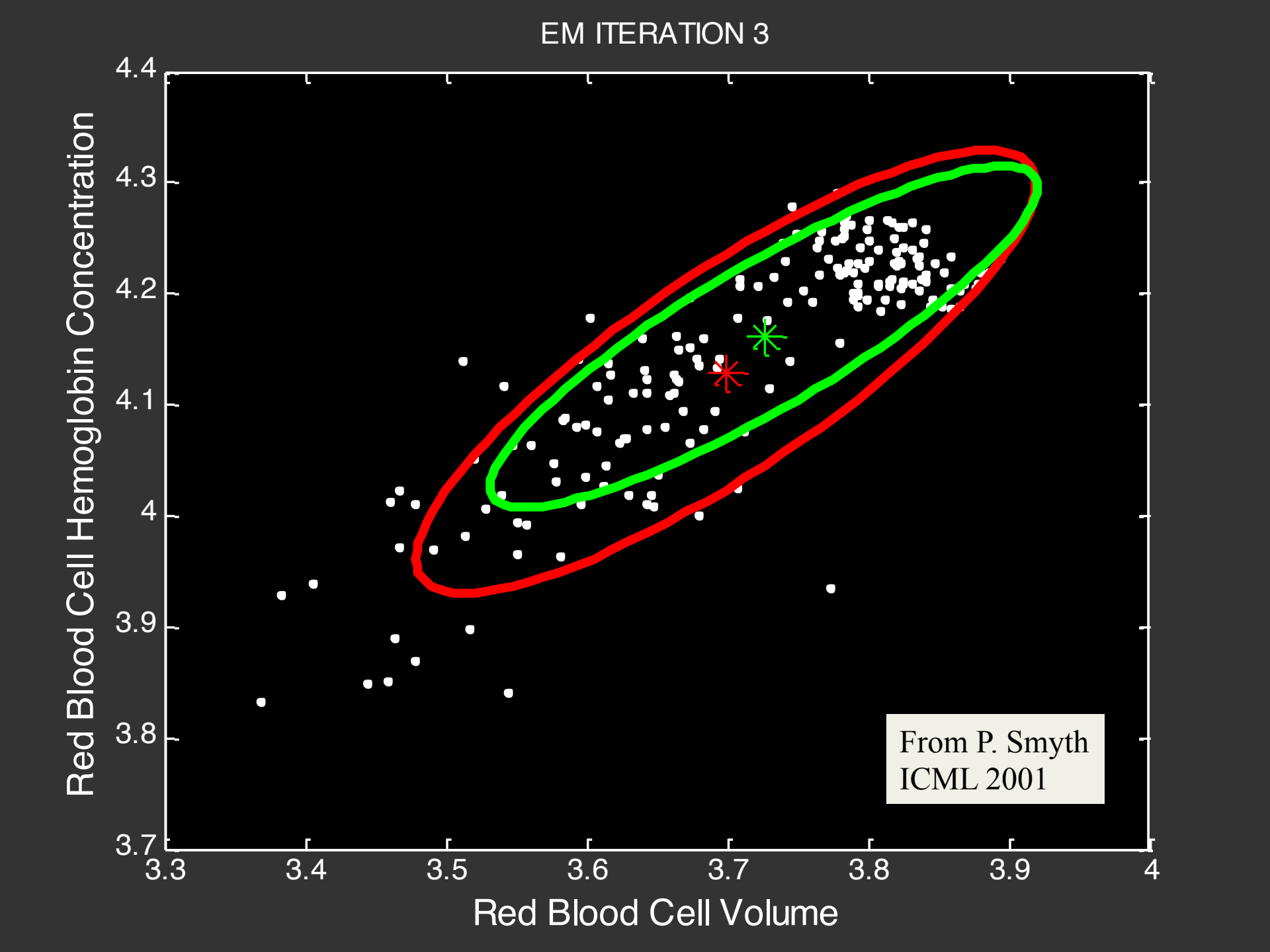
Autorzy:

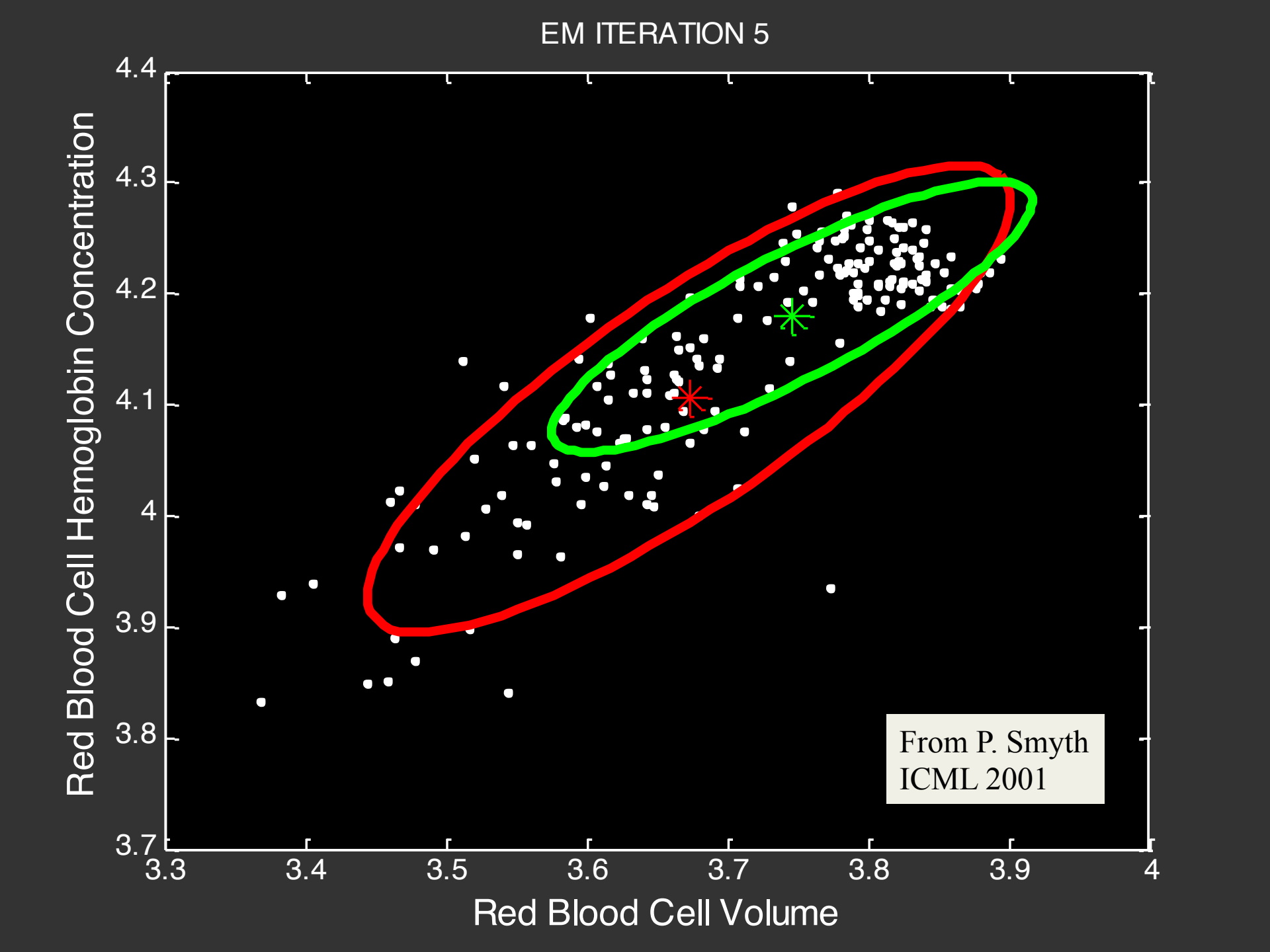
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1–38.

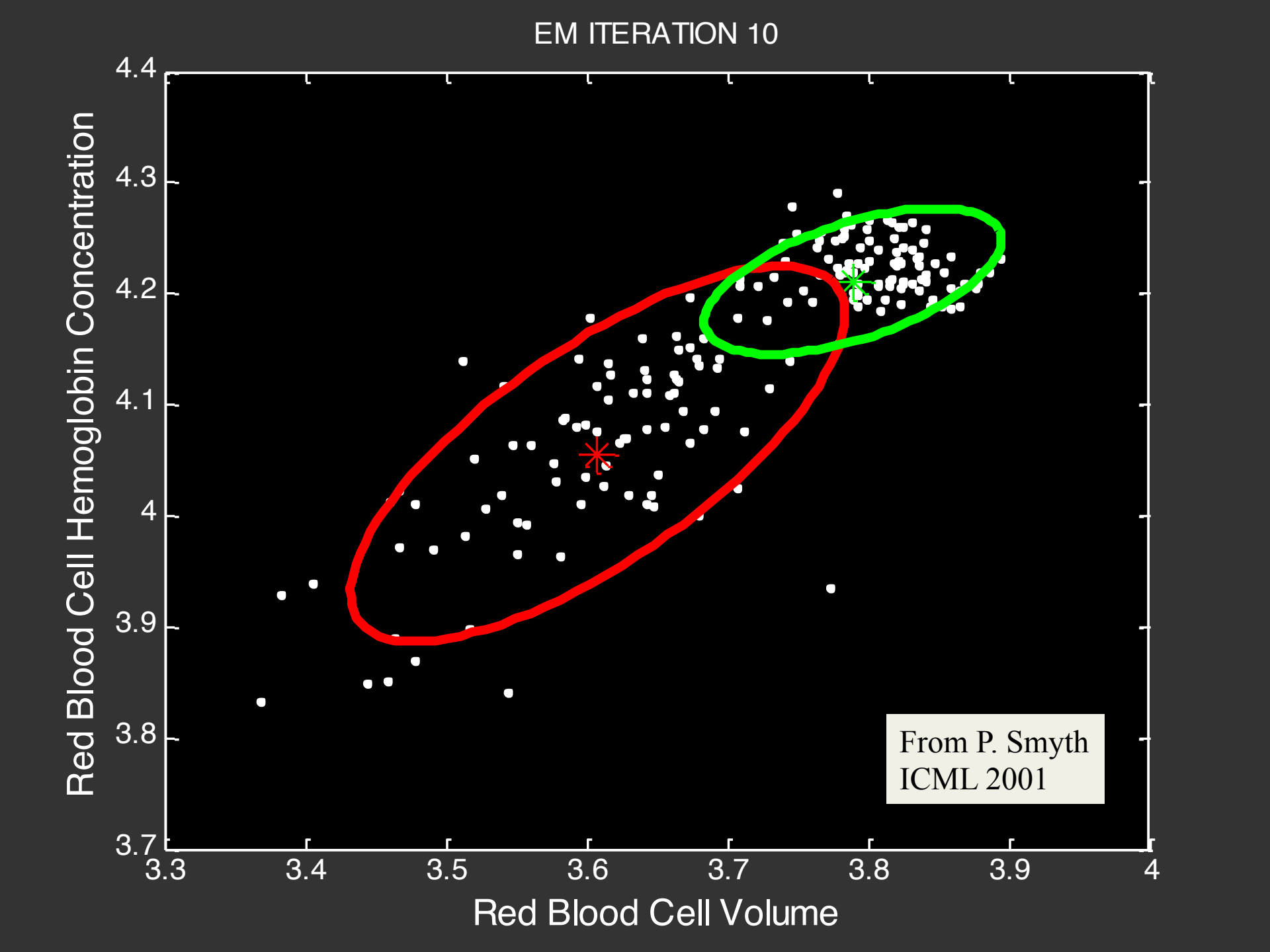
Przykład działania algorytmu EM

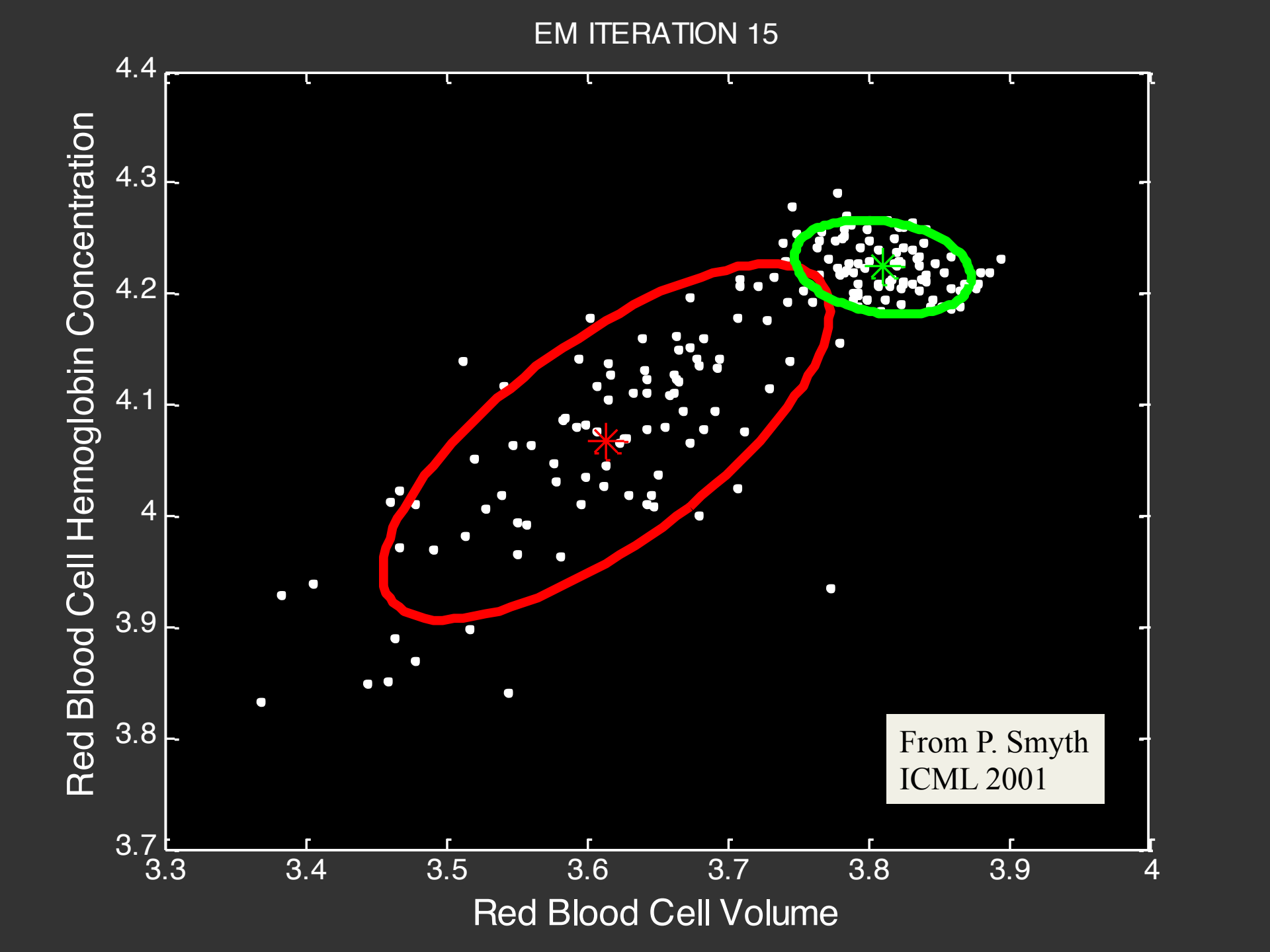

ANEMIA PATIENTS AND CONTROLS
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001

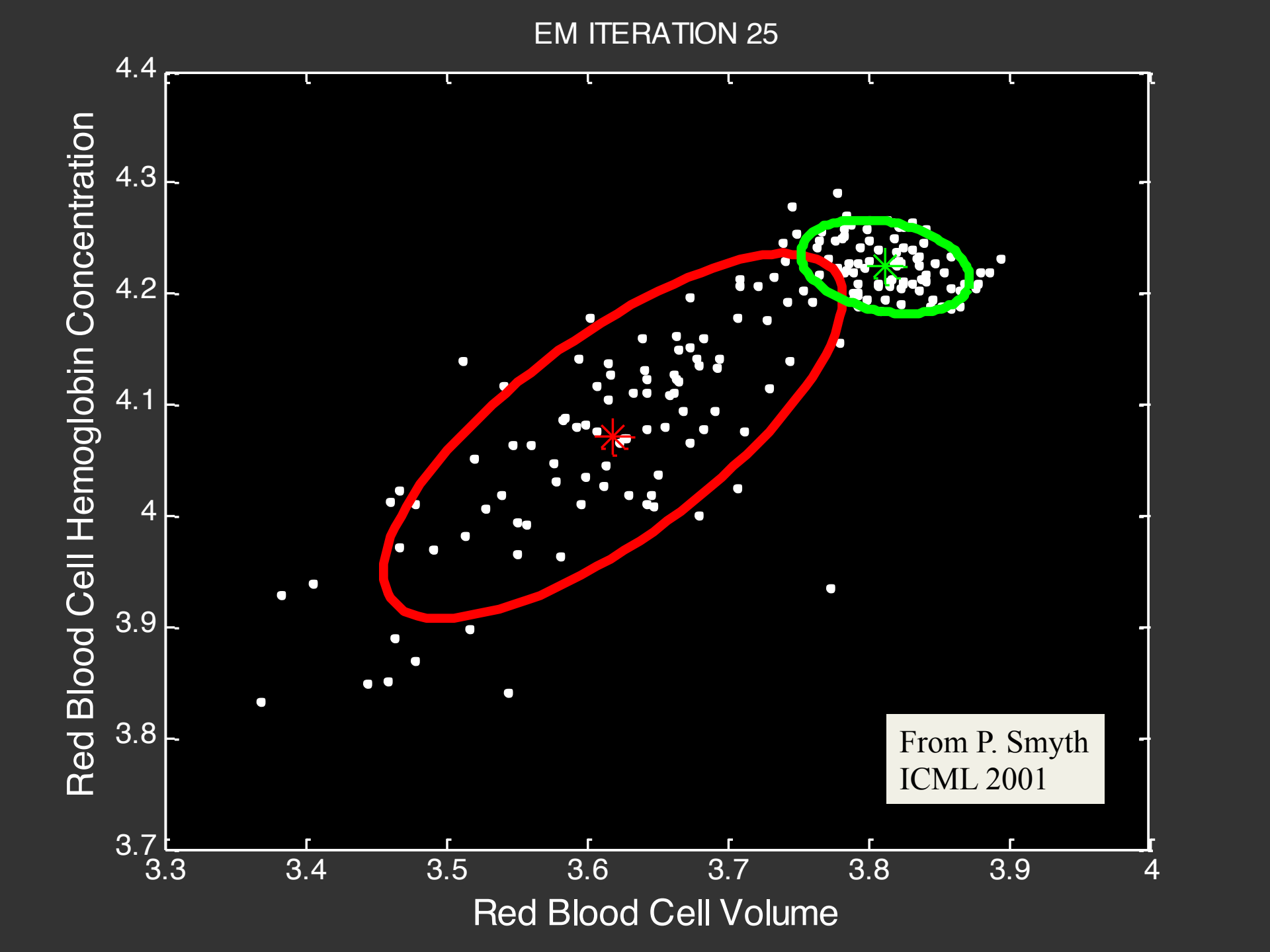

EM ITERATION 1
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001

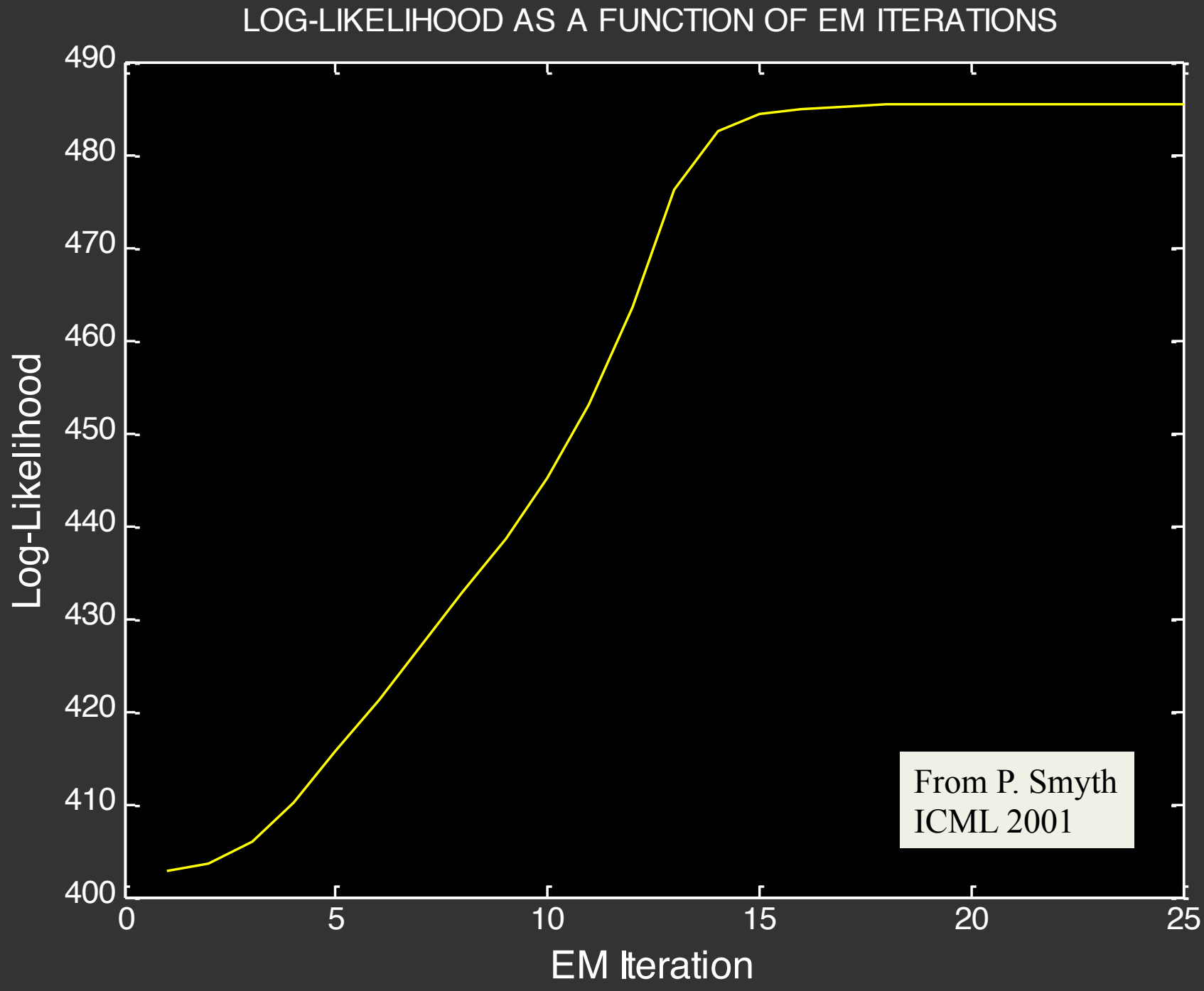

EM ITERATION 3
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001


EM ITERATION 5
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001

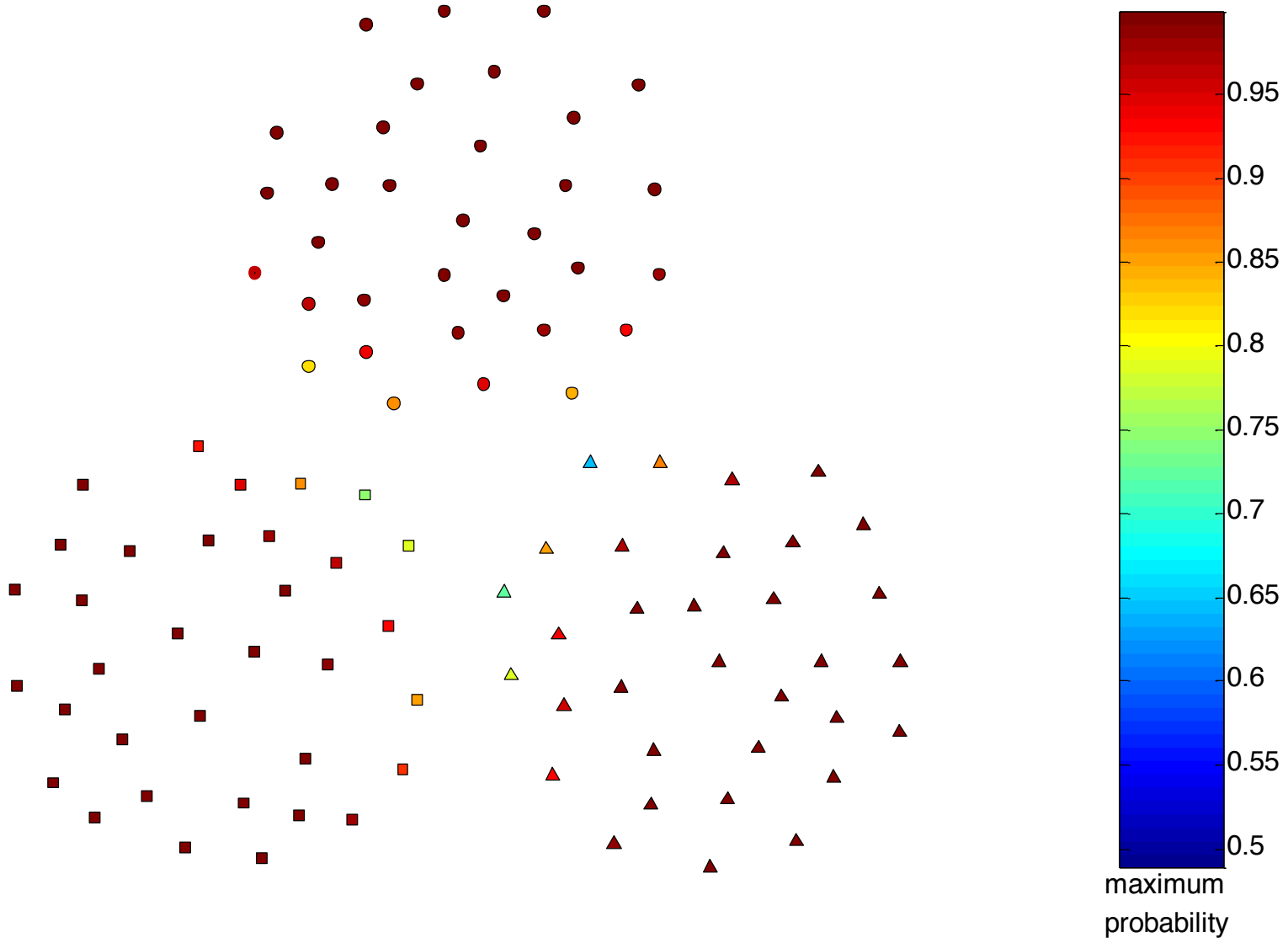

EM ITERATION 10
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001


EM ITERATION 15
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001

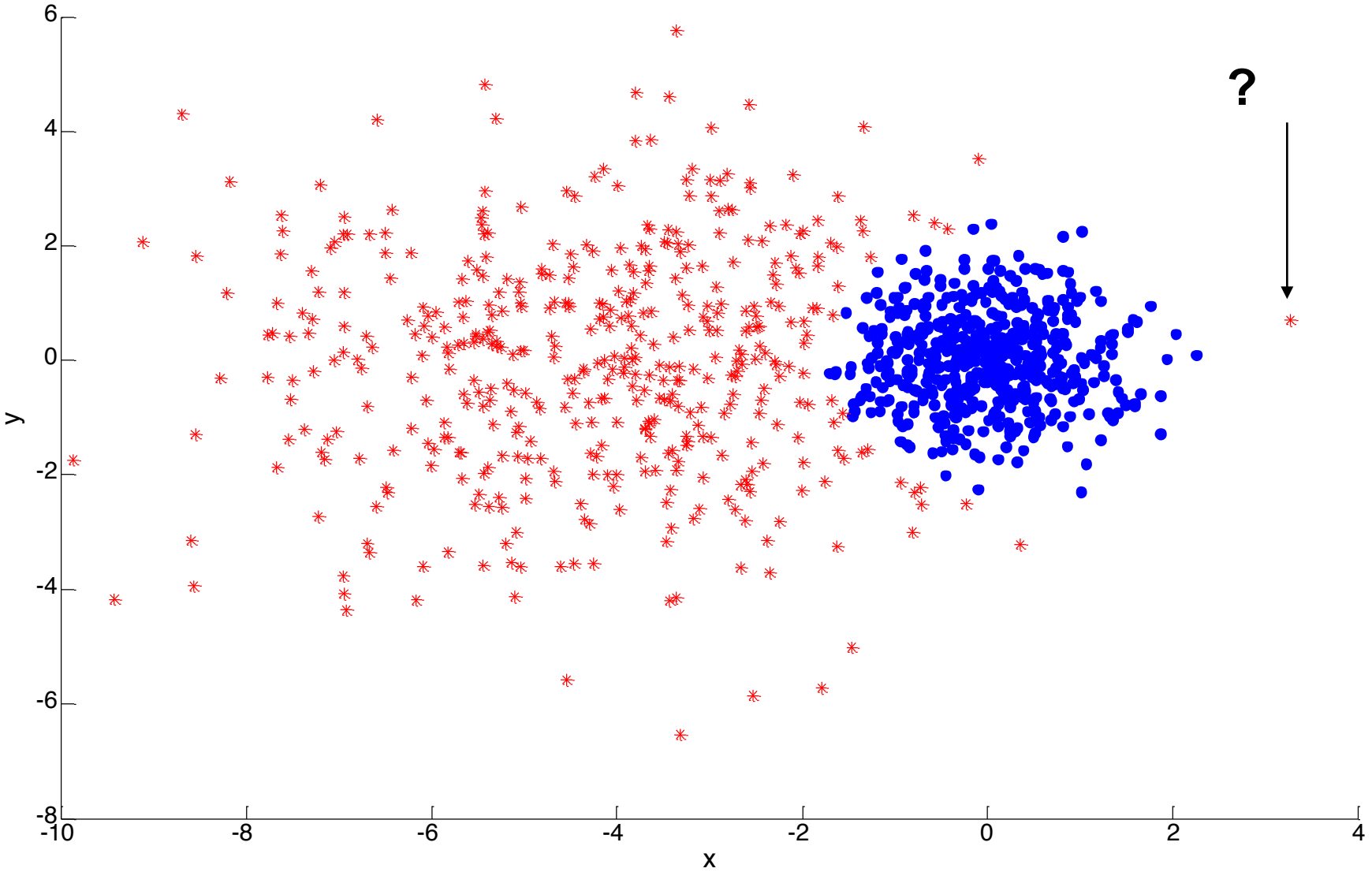

EM ITERATION 25
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
From P. Smyth
ICML 2001


LOG-LIKELIHOOD AS A FUNCTION OF EM ITERATIONS
Log-Likelihood
EM Iteration
490
480
470
460
450
440
430
420
410
400
0
5
10
15
20
25
From P. Smyth
ICML 2001

Trzy skupiska odkryte algorytmem EM



Probabilistic Clustering: Dense and Sparse Clusters



EM w porównaniu do algorytmu k-średnich

- EM łączy podejście probabilistyczne oraz zasadę oszacowanie MLE z paradygmatem algorytmów iteracyjno-optymalizacyjnych
- Podobieństwa do algorytmu k-średnich, gdyż
  - podział obiektów także do k-skupisk (choć może być „miękki - soft” przydział)
  - Krok oczekiwania (E) odpowiednik przydziału odległościowego do najbliższego centroidu
  - Krok maksymalizacji (M) aktualizacja oszacowań parametrów odpowiednik przeliczania położenia centroidów, lecz z wykorzystaniem maksymalizacji funkcji wiarygodności LogL

EM w porównaniu do algorytmu k-średnich

- EM jest bardziej ogólny niż k-średnich, z uwagi na różne rozkłady można modelować skupiska o innych kształtach niż sferyczne (np. eliptyczne)
- Silniejsze założenia i podstawy statystyczne
- W literaturze dalsze rozszerzanie lub wykorzystanie modeli probabilistycznych (np. CEM, SNOB, AUTOCLASS), także w wersji hierarchicznej − przegląd książka T.Morzy Eksploracja danych

Ograniczenia algorytmu EM

- Zbieżność może być powolna
- Poszukuje tzw. lokalne minimum
- Dobór liczby skupisk *k* – nie jest łatwy (są propozycje automatyzacji − patrz książka K.Stąpor)
- Trudność dopasowania do potencjalnych skupisk będących b. rzadkie (z małą ilością obiektów)
  - Nieodporny na obecność obserwacji samotniczych (outliers) lub szumu (noise points)
- Liczba parametrów modeli wzrasta $O(d^2)$, gdzie d jest liczbą cech
  - Zwłaszcza l. parametrów dla macierzy kowariancji

Grupowanie pojęciowe

- Tworzenie skupisk, które modeluje potencjalne pojęcia ukryte w danych – oraz automatycznie wspiera opisy skupisk w języku potencjalnie interpretowalnym przez człowieka
- Pierwsze algorytmy symboliczne (dla atrybutów jakościowych, oraz wykorzystanie zmodyfikowanych algorytmów odkrywania reguł)
- R.Michalski, R. Stepp: Learning from observation. Conceptual Clustering (1983)

Cluster (Michalski, Stepp)

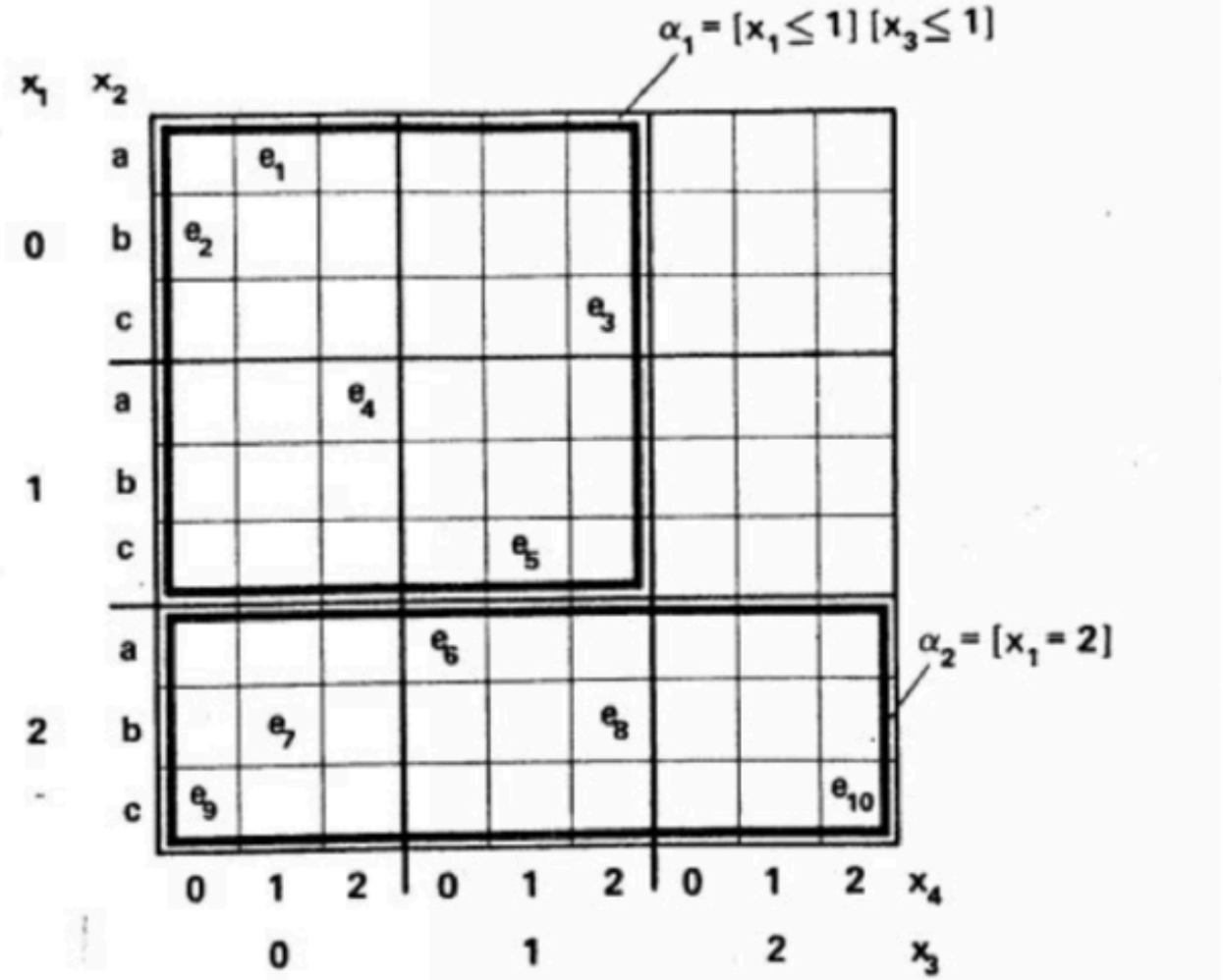


Figure 11-11: A diagrammatic representation of the clustering $\{\alpha_1, \alpha_2\}$.

Cluster (Michalski, Stepp)

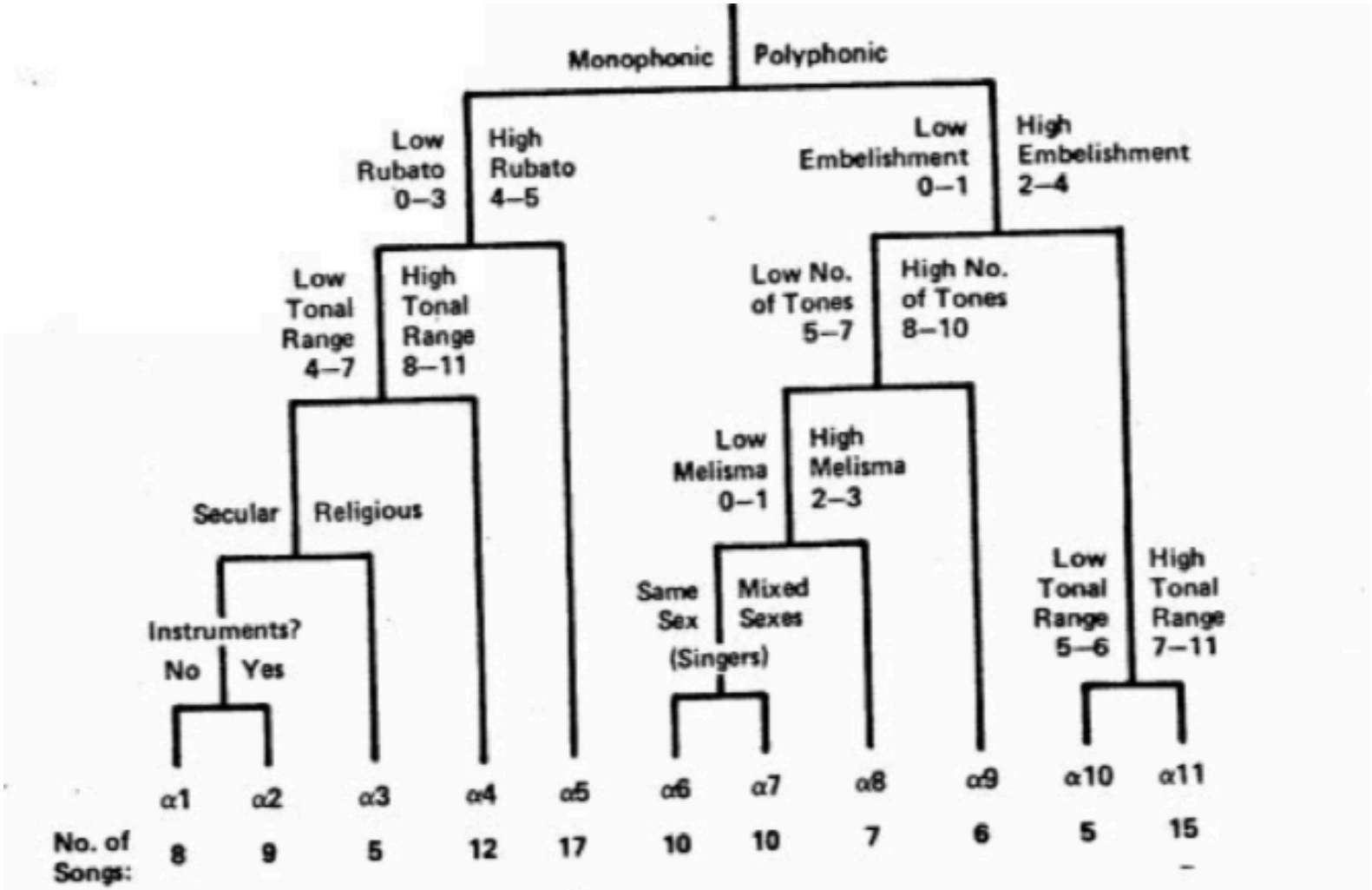


Figure 11-13: A classification hierarchy of Spanish folk songs produced by CLUSTER/2.

COBWEB – różne elementy w jednym

- Podejście hierarchicznego grupowania
- Wsparcie dla tworzenia opisów pojęć z wykorzystaniem prawdopodobieństw
- Elementy probabilistycznego modelowania
- Uczenie przyrostowe z danych

Ogólny schemat postępowania:

Start:

tree consists of empty root node

Then:

add instances one by one

update tree appropriately at each stage

to update, find the right leaf for an instance

May involve restructuring the tree (split leaf or join to cluster)

Base update decisions on category utility

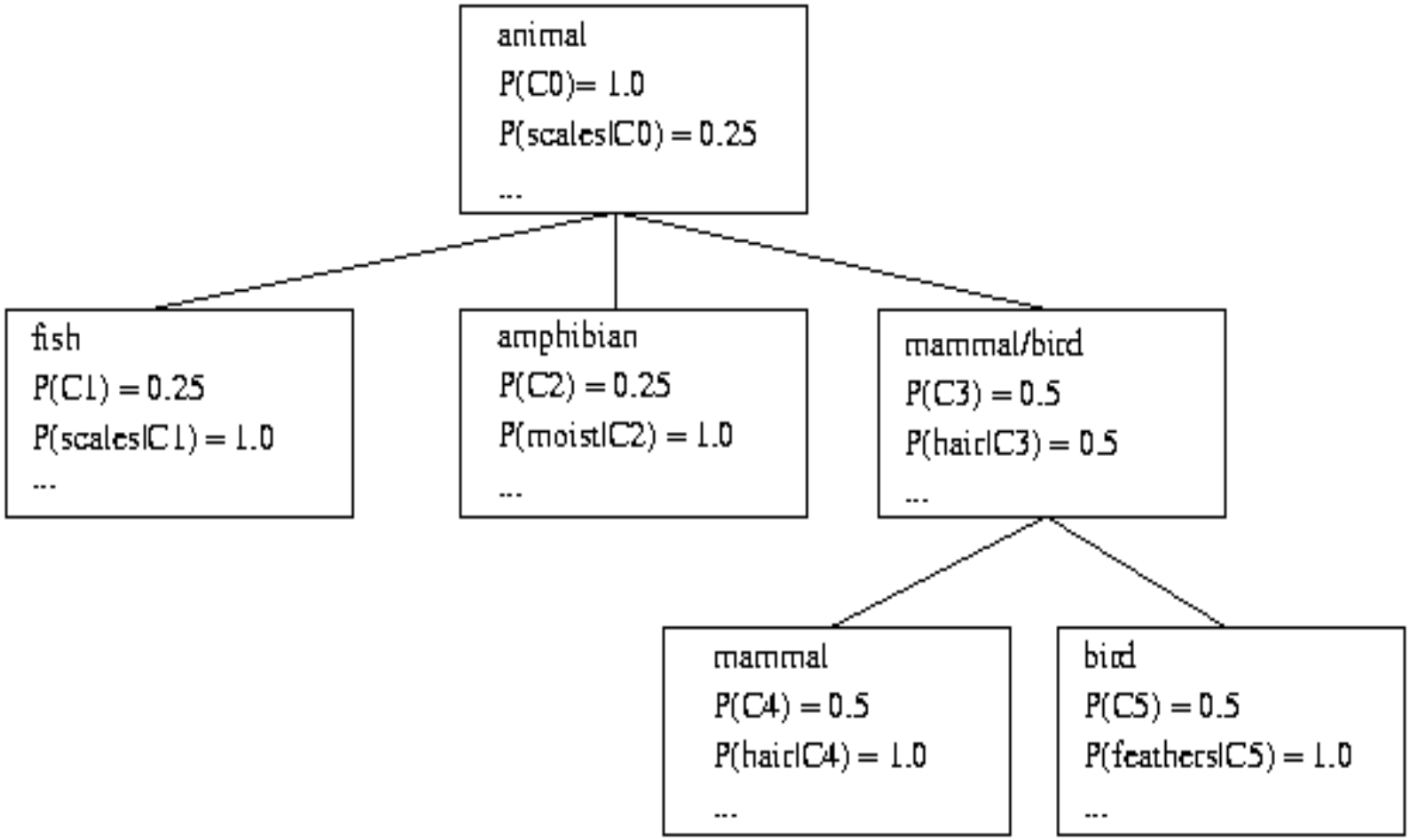
COBWEB - D.Fisher 1986

- Podział zbioru obiektów tak, aby znaleźć taką strukturę kategorii (klas), która prowadzi do maksymalizacji informacji, jaką można przewidzieć znając kategorie przykładu (pot. klasyfikacja)
- heurystyczna funkcja oceny grupowania
  - Inspiracja wnioskowaniem Bayesowskim

$$\frac{1}{|C|}\sum_{d\in C} P(c(x)=d)\left[\sum_{a_i}\sum_{v_{ij}} P(a_i(x)=v_{ij}\mid c(x)=d)^2 - \sum_{a_i}\sum_{v_{ij}} P(a_i(x)=v_{ij})^2\right]$$

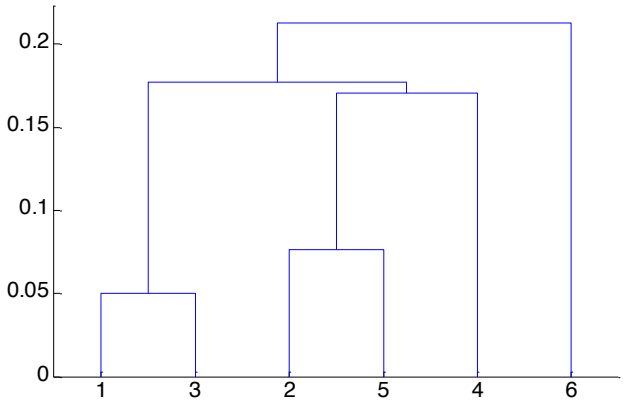
COBWEB – dane zoo

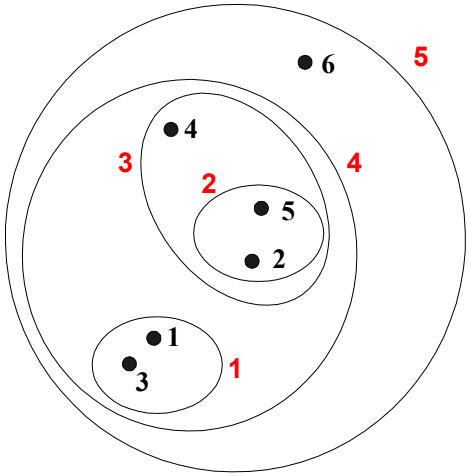
Dynamiczna struktura drzewa + opis skupisk zestaw prawdop.



Grupowanie hierarchiczne

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





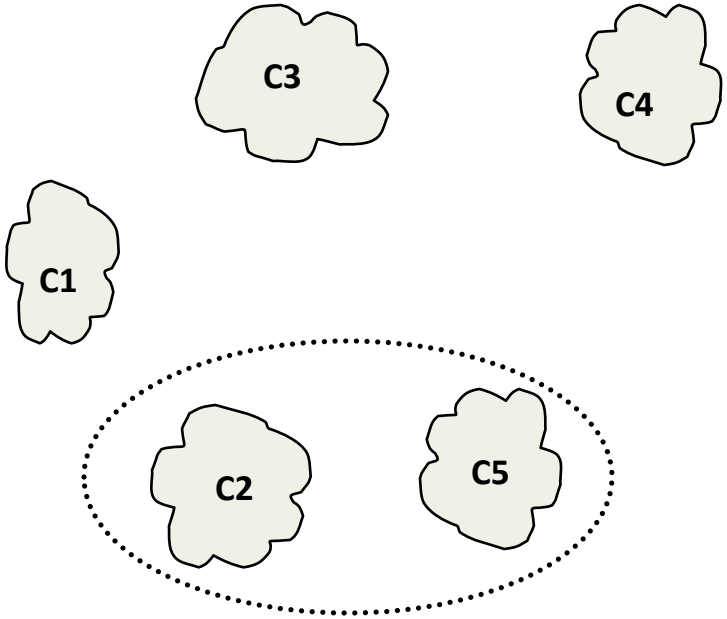
Hierarchiczne metody aglomeracyjne - algorytm

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?

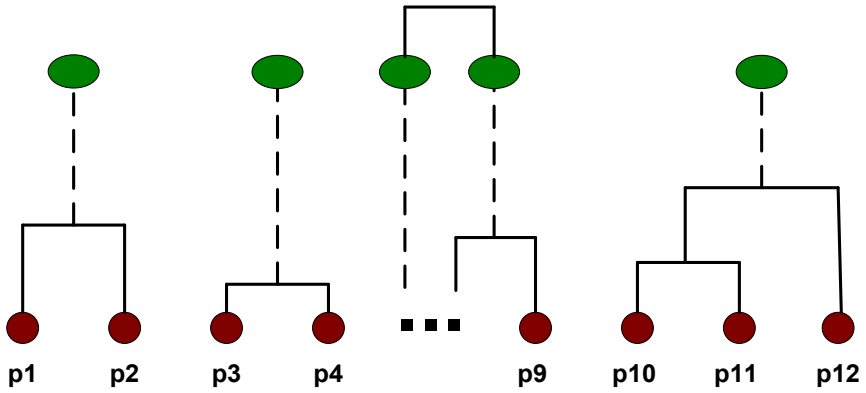
Łączymy dwa skupiska (C2 i C5) i aktulizujemy macierz odległości

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 wewnątrz 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:

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

Complete linkage
maximum distance:

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

mean distance:

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

average distance:

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

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

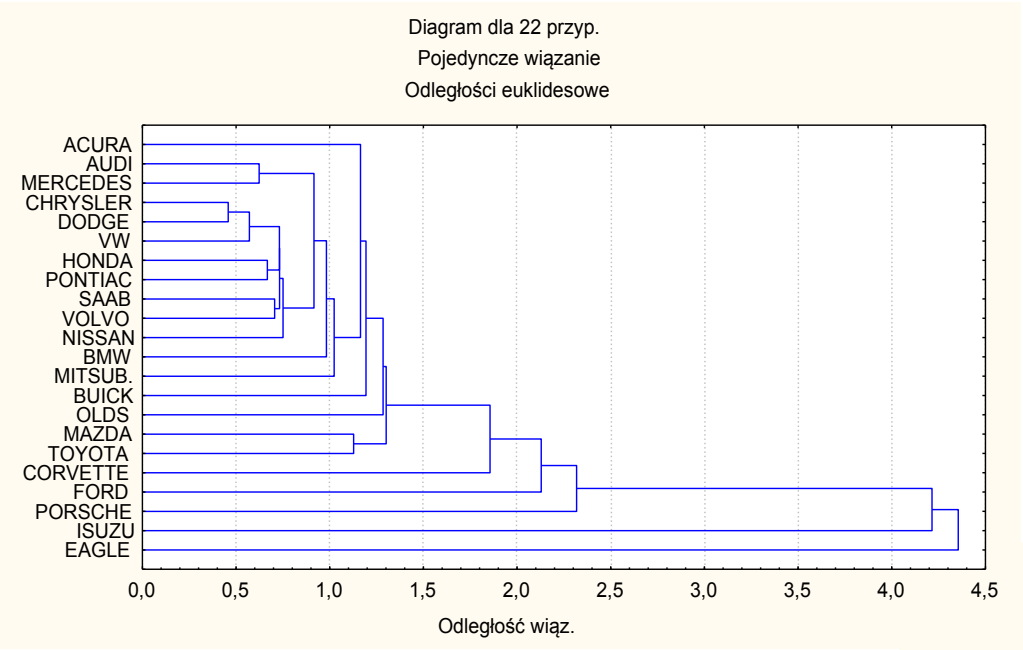
Complete Link Agglomerative Clustering

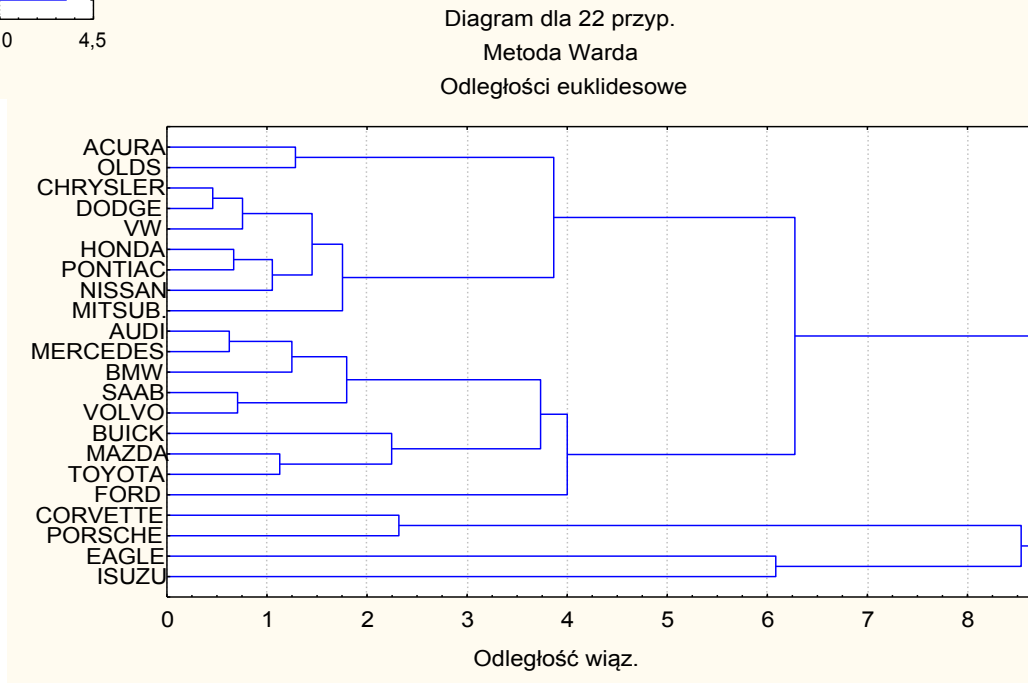
- Użyj maksymalnej odległości – minimalnego podobieństwa

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

- 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





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]

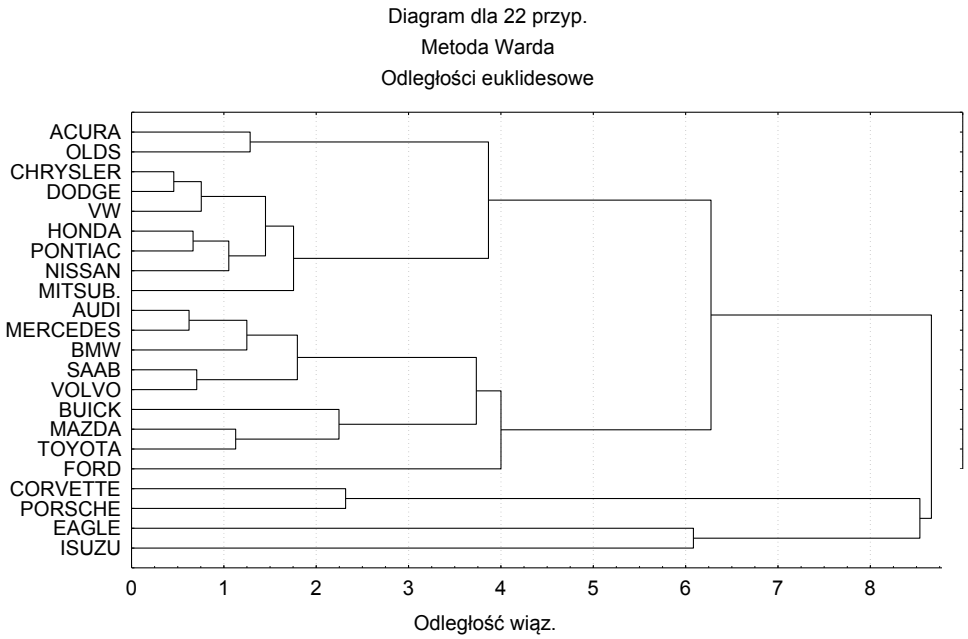
- Jest to metoda podobna jak poprzednia, z tym wyjątkiem, że w obliczeniach wprowadza się „ważenie”, aby uwzględnić różnice między wielkościami skupień (tzn. liczbą zawartych w nich obiektów).
- Zatem, metoda ta jest lepsza od poprzedniej w sytuacji, gdy istnieją (lub podejrzewamy, że istnieją) znaczne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

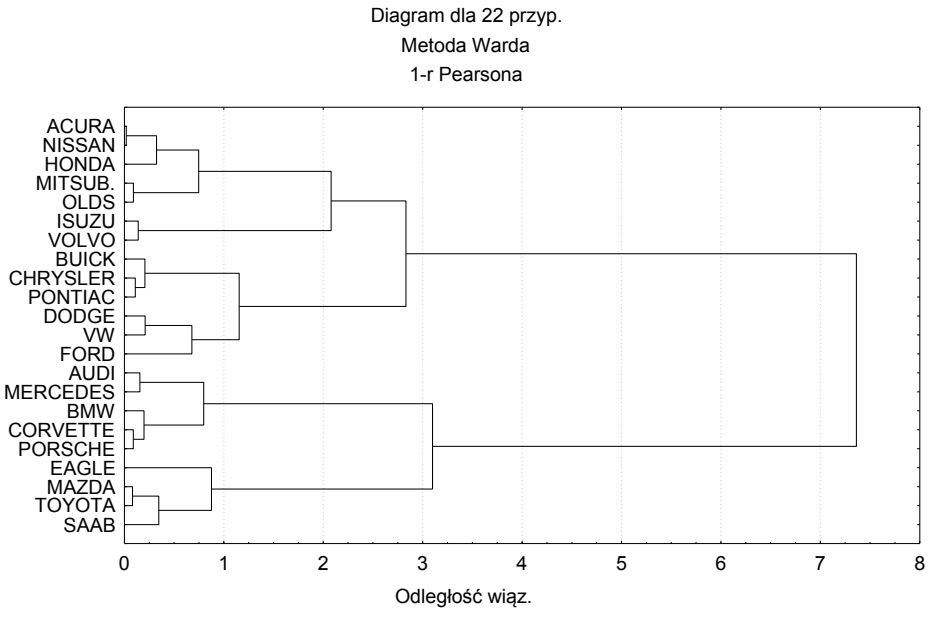
Metody łączenia – Ward method

- Gdy powiększamy jedno ze skupień *C*k, wariancja wewnątrzgrupowa (liczona przez kwadraty odchyleń 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,....., *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)

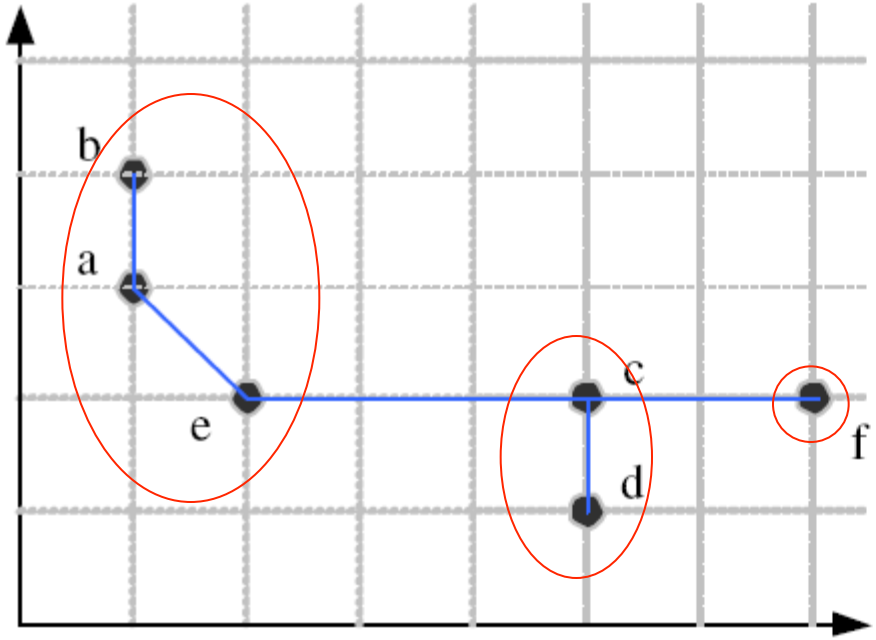
Przykłady użycia metody Warda

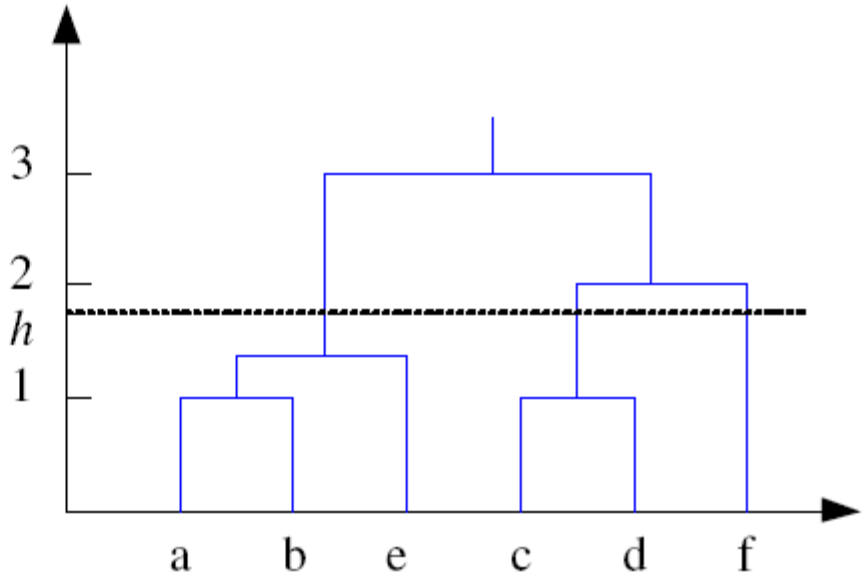
Cars data





Jak wykorzystać AHC do oszacowania potencjalnej liczby skupisk w danych?

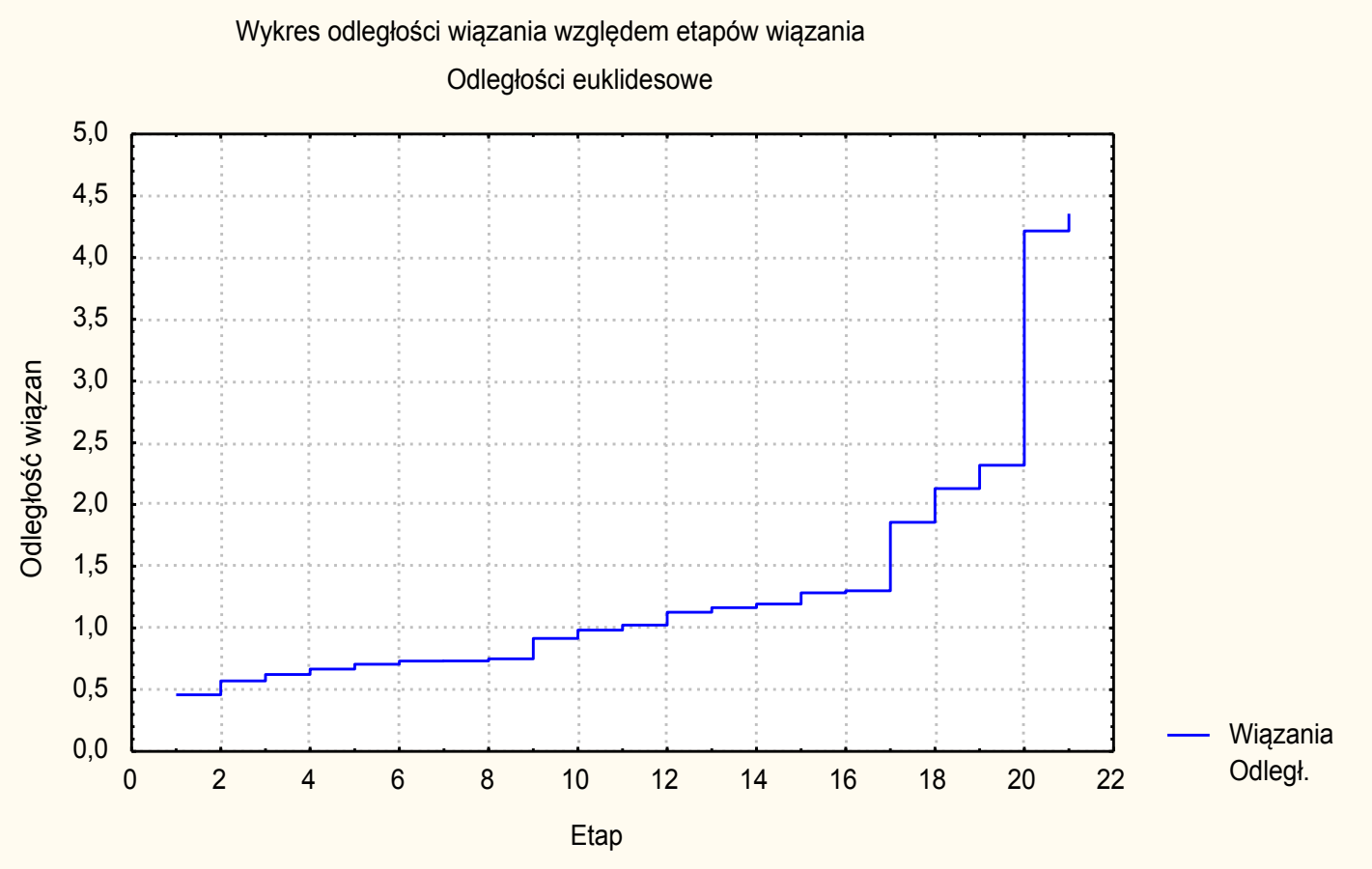




Dendrogram

AHC – jak odnaleźć liczbę skupień?

Znajdź punkt przegięcia („kolanko”) wykresu odległości wiązania względem kolejnych etapów łączenia obiektów / skupisk



Grupowanie Dużych Repozytoriów

- Skalowalność (względem liczby przykładów lecz także wysokiej wymiarowości atrybutów)
- Uwzględnianie złożonych typów danych
- Konstruowanie dowolnych kształtów skupień
- Wspomaganie parametryzacji (jak dobrać k, parametry DBSCAN, itp.) = postulaty tzw. AUTOML
- Odporność na szum i obserwacje nietypowe
- Grupowanie przyrostowe
- Przetwarzanie strumieni danych
- Zmiana położenia definicji pojęć (data shift / w przypadku strumieni tzw. concept drift)

Problemy i wyzwania

- Od lat 90tych widoczny postęp w zakresie skalowalnych algorytmów (zwłaszcza ja dane nie mieszczą się w PAO):
  - Partitioning: *k*-means, *k*-medoids, PAM, CLARANS
  - Hierarchical: BIRCH, CURE
  - Density-based: CLIQUE, OPTICS
  - Grid-based: STING, WaveCluster.
  - Model-based: Autoclass, Denclue, Cobweb.

Lecz

- Obecne techniki ciągle nie spełniają wystarczająco dobrze stawianych wymagań
- Otwarte problemy i wyzwania badawcze; zwłaszcza dla nietypowych i złożonych danych

Metody hierarchiczne dla dużych zbiorów danych

- Niektóre z ograniczeń metod aglomeracyjnych:
  - słaba skalowalność: złożoność czasowa przynajmniej $O(n^2)$, gdzie n jest liczbą obiektów,
  - „krytyczne” znaczenie decyzji o wyborze punktu połączenia kolejnych skupień w trakcie budowania drzewa hierarchii,
  - algorytmy nie zmieniają, ani nie poprawiają, wcześniej podjętych decyzji.
- Rozwinięcia algorytmów hierarchicznych oraz ich integracja z metodami gęstościowymi:
  - BIRCH (1996): użycie drzew o strukturze„CF-tree”, uczenie przyrostowe i stopniowa poprawa jakości pod-skupień.
  - CURE (1998): wybór losowy odpowiednio rozproszonych punktów, wstępne grupowanie z określeniem ich punktów reprezentatywnych, łączenie grup w nowe skupienia wraz z przesuwaniem punktów reprezentatywnych w stronę środków tworzonego skupienia zgodnie z „shrinking factor α” ; eliminacja wpływu „outliers”.

BIRCH – efektywne rozszerzenie grupowanie hierarchicznego

- Działa efektywnie: decyzja dla jednej grupy (dzielenie czy połączenie z inną grupą) nie wymaga przeglądania całego zbioru danych
- I/O koszt jest liniowy względem rozmiaru danych: przeglądanie zbioru danych raz
- Ukierunkowany na tworzenie zrównoważonego drzewa hierarchii skupisk

BIRCH – ang. Balanced Iterative Reducing and Clustering using Hierarchies – Zhang et al. (1996)

- Wykorzystuje hierarchiczne drzewo CF (ang. Clustering Feature)
- Działanie algorytmu:
  - Faza 1: przyrostowo przeczytaj raz DB (zew. Baza danych) w celu zbudowania w pamięci początkowej struktury drzewa CF (rodzaj wielopoziomowej kompresji danych zachowującej wewnętrzną strukturę zgrupowań danych).
  - Faza 2: zastosuj wybrany (inny) algorytm skupień dla lepszego pogrupowania obiektów w liściach drzewa CF.
- *Dobra skalowalność*: znajduje zadawalające grupowanie po jednokrotnym przeczytaniu bazy danych i ulepsza je wykorzystując niedużo dodatkowych operacji odczytu DB.
- *Ograniczenia:* zaproponowany dla danych liczbowych, wrażliwość wyników na kolejność prezentacji przykładów.

Informacje o danych skupiskach

CF - struktura wykorzystywana w konstrukcji drzewa

Podstawowe parametry BIRCH:

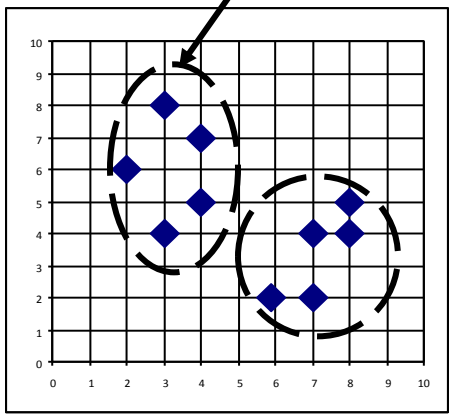
- *B* – maksymalna liczba rozgałęzień w drzewie
- *L* – maksymalna liczba obiektów w liściu
- *T* – maksymalny promień (grup) w liściu

Clustering Feature: $CF = (N, \overrightarrow{LS}, SS)$

N: **Number of data points**

LS: $\sum_{i=1}^{N} \overrightarrow{X_i}$

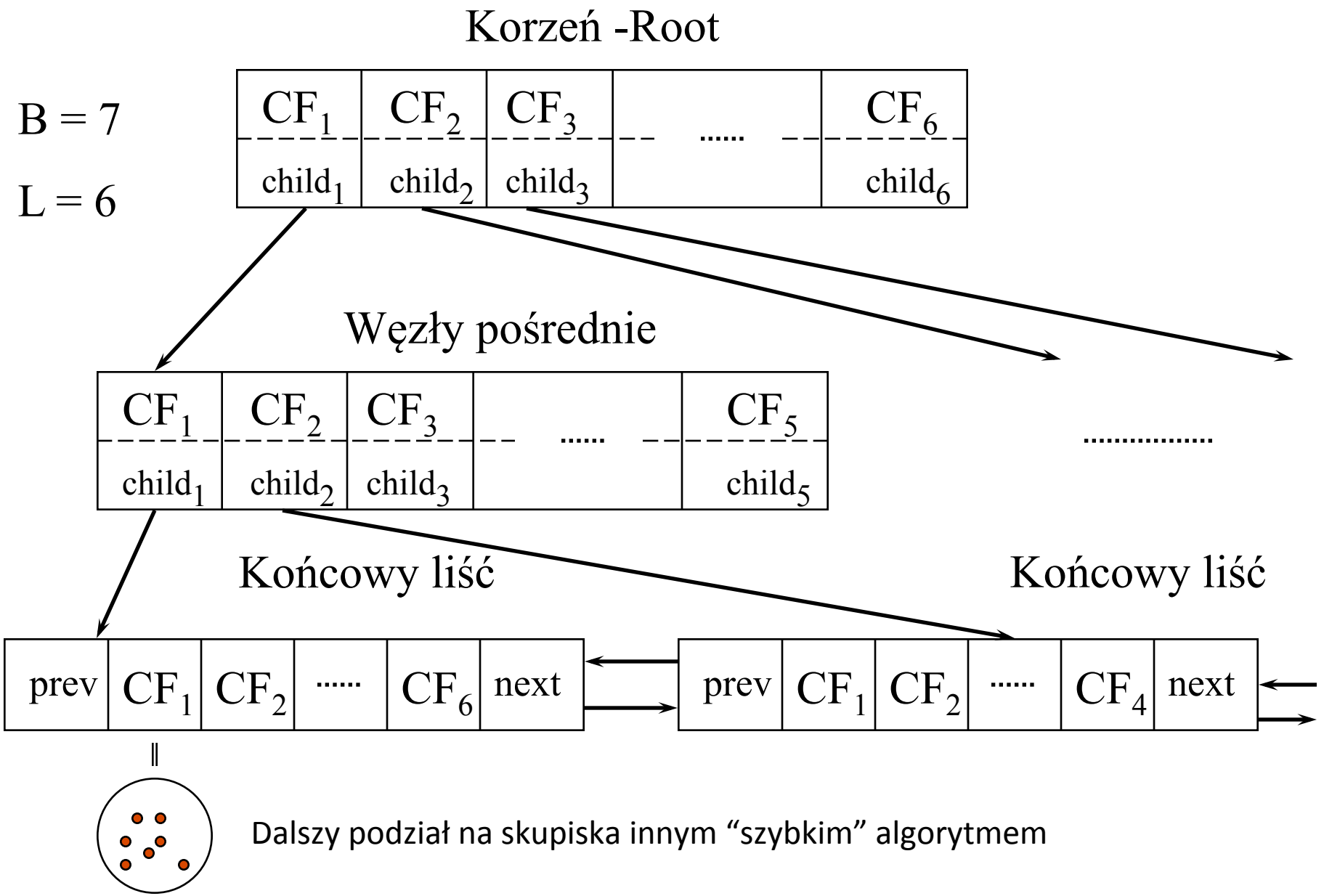
SS: $\sum_{i=1}^{N} \overrightarrow{X_i^2}$



CF = (5, (16,30),(54,190))

(3, 4)
(2, 6)
(4, 5)
(4, 7)
(3, 8)

Przykładowa struktura CF Tree



Przetwarzenie kolejnego przykładu

- Wstawienie przykładu do struktury drzewa

Krok 1. Wybierz liść l do wstawiania. Użyj jednej z funkcji odległości do wyznaczenia najbliższej grupy do badanego punktu

Krok 2. Jeśli w liściu l jest miejsce to wstaw x,

Jeśli nie Podziel liść l na dwa liście i popraw ścieżkę od l do korzenia.

Krok 3. Rekonstrukcja drzewo przez połączenie dwa najbliższe węzły i podzielić na dwa (w razie potrzeby): merge i resplite

Inne algorytmy grupowania

Środowisko data mining:

- Algorytmy “gridowe”
  - STING, CLIQUE, WaveCluster
- Grafowe algorytmy
  - Chameleon
  - Jarvis-Patrick
  - Based on Nearest Neighbor (SNN)
- Tzw. Sub-space clustering
- Online stream clustering

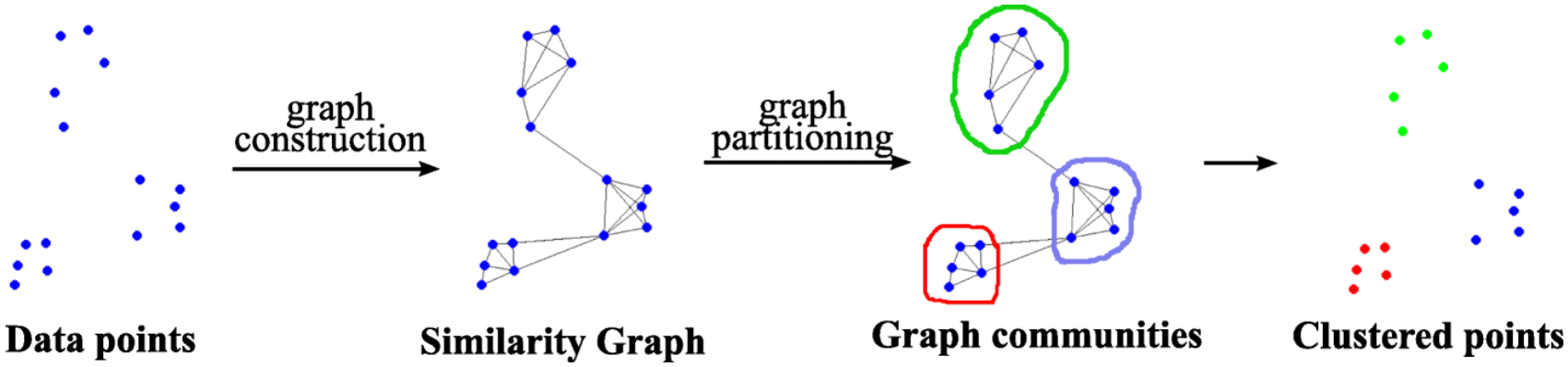
Grafowe algorytmy

Przeznaczone do przetwarzania danych jako grafów (często dużych, np. w odniesieniu do Internetu lub sieci społecznych,

lub dotyczących naturalnie występujących powiązań, np. związki chemiczne)

Także grupujące standardowe reprezentacje danych, wykorzystując ich strukturę wewnętrzną modelowaną jako graf.

- Grafy najbliższych sąsiadów (np. ε-ball, kNN and CkNN graphs)
- Drzewa rozpinające w grafach



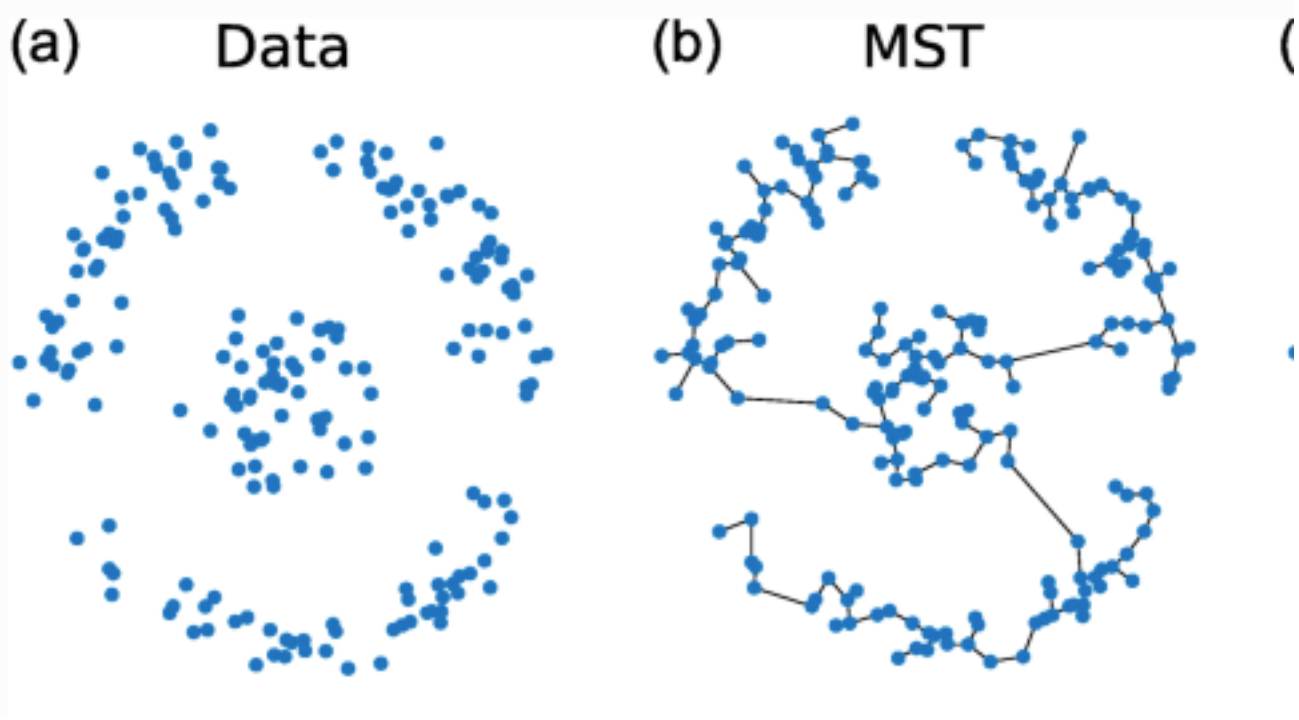
Minimalne drzewa rozpinające (MST)

Inne podejścia do grupowania wykorzystujące globalne własności geometrii połączeń pomiędzy punktami

Minimalne drzewo rozpinające może to modelować. Później połączone z analizą k-sąsiadów lub tzw. dendrytów (metoda wrocławska) można zidentyfikować skupiska

Definicja formalna: Niech G =(V,E,f) będzie nieskierowanym grafem spójnym nieskierowanym ważonym (Wagi odległości lub podobieństwa między wierzchołkami). Minimalnym drzewem rozpinającym (minimum spanning tree, MST) w G nazwiemy takie drzewo rozpinające (graf bez cykli, który łączy wszystkie rozważane wierzchołki), w którym suma wag krawędzi jest najmniejsza możliwa.

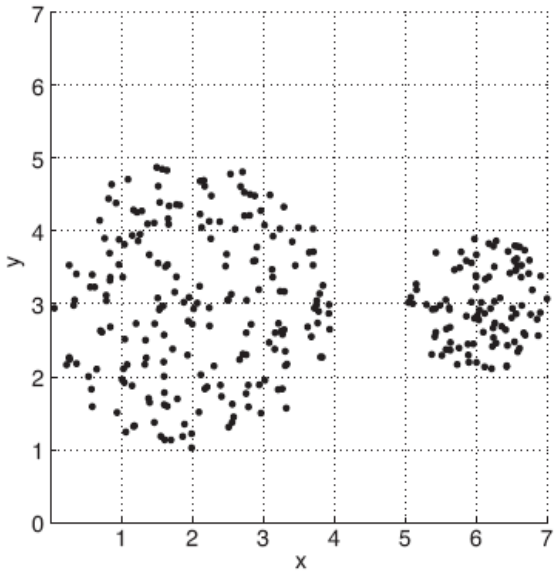
Minimalne drzewa rozpinające (MST)



Grid-based Clustering

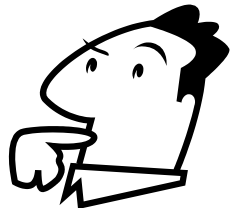
Algorithm 9.4 Basic grid-based clustering algorithm.

1: Define a set of grid cells.
2: Assign objects to the appropriate cells and compute the density of each cell.
3: Eliminate cells having a density below a specified threshold, τ.
4: Form clusters from contiguous (adjacent) groups of dense cells.



0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	**17**	**18**	**6**	0	0	0
14	**14**	**13**	**13**	0	**18**	**27**
11	**18**	**10**	**21**	0	**24**	**31**
3	**20**	**14**	**4**	0	0	0
0	0	0	0	0	0	0

Ocena jakości skupień



Czy można poszukiwać pojedynczej miary?

- Pewne „trudne” rady

"The problem of how to judge the quality of a clustering is difficult and there seems to be no universal answer to it."

"The nature of processes leading to useful classifications remains little understood, despite considerable effort in this direction."
— R. Michalski, R. Stepp [MS83]

"How do you know the resulting classifications are any *good?"*
— D. Fisher [Fis87]

Ocena wyników algorytmów grupowania – wielokryterialna / brak pojedynczej dominującej miary takiej jak w uczeniu nadzorowanym oraz zależna od rodzaju grupowania (płaska, hierarchiczna, …)

Różne spojrzenia na ocenę grupowania

- Wewnętrzne (ocena tylko charakterystyki skupień i rozkładem przykładów)
  - Brak dodatkowych źródeł informacji, np. zbioru odniesienia etykiet
  - Miary oceny oparte na danych (internal measures)
- Zewnętrzne
  - „Benchmarking on existing labels”
  - Porównanie skupień z tzw. ground-truth categories / zadanym podziałem
- Ocena ekspercka

Różne spojrzenia na ocenę grupowania

Ponadto miary są przeznaczone do oktreślonego rodzaju grupowania:

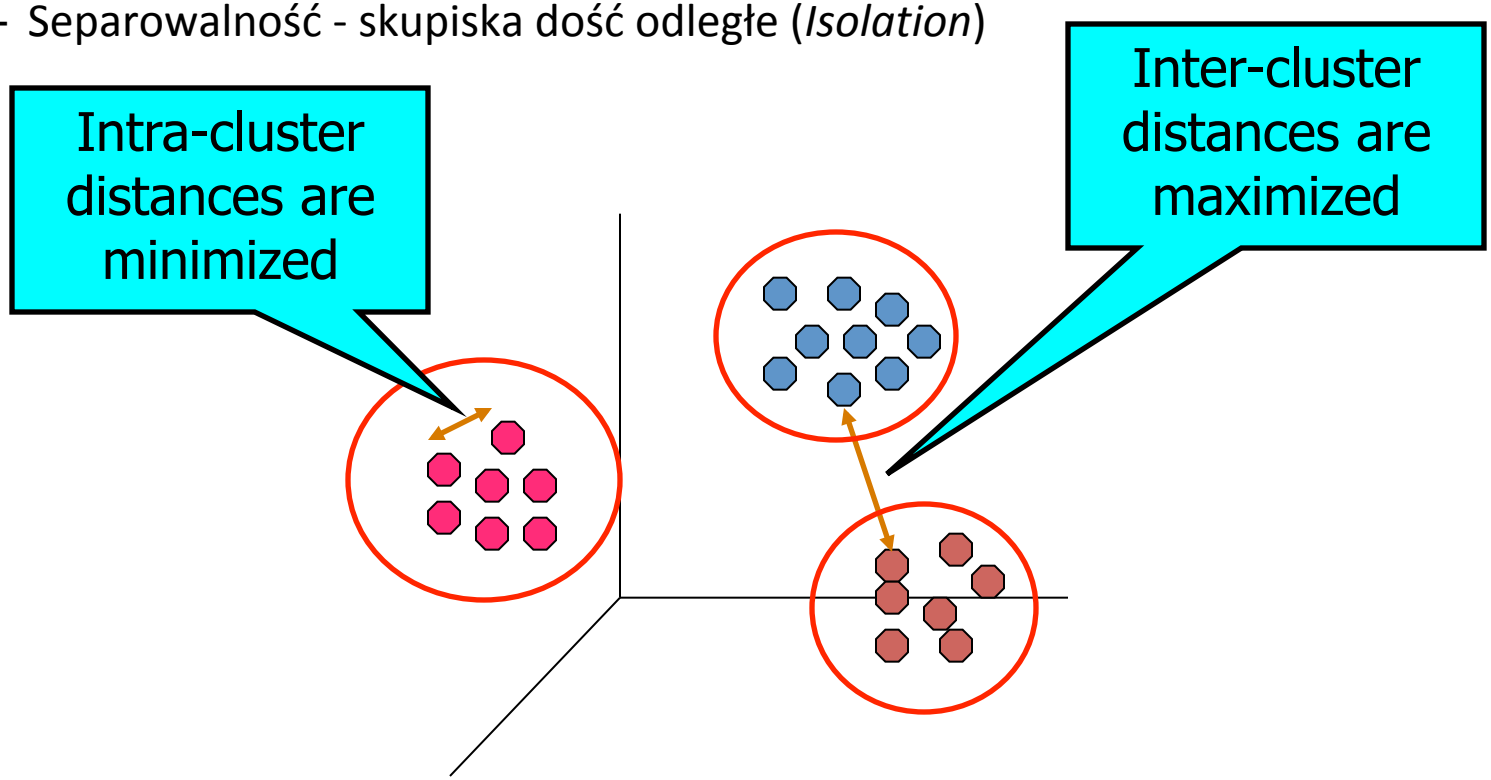
- Płaska struktura skupisk vs. hierarchiczna
- Jednoznaczne (ang. Crisp) przydziały do skupisk vs. rozmyte przynależności

Rozpocznijmy do płaskiej struktury (jak k-means) z jednoznaczymi przydziałami = literatura ponad 30 różnych propozycji

Ocena wewnętrzna jakości skupień

Miary oceny oparte na danych (internal measures)

- Oparte na odległościach
- Duże podobieństwo obiektów wewnątrz skupienia (*Compactness*) / zwartość skupiska
- Separowalność - skupiska dość odległe (*Isolation*)



Podstawowe miary wewnętrzne

- Zwartość skupisk (możliwe bliskie obserwacje – mała średnica skupisk lub odległość od centroidu)
- Separowalność skupisk (powinny być maksymalnie odróżnialne od siebie; średnie odległości pomiędzy parami punktów lub środków skupień)

Najprostsze miary

K – skupisk C_k, każde o liczności n_k

Skupiska C_k charakteryzowane przez centroidy - średnie obiekty w skupieniu

$$\mathbf{r}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

- Błąd zmienności wewnątrz skupieniowej

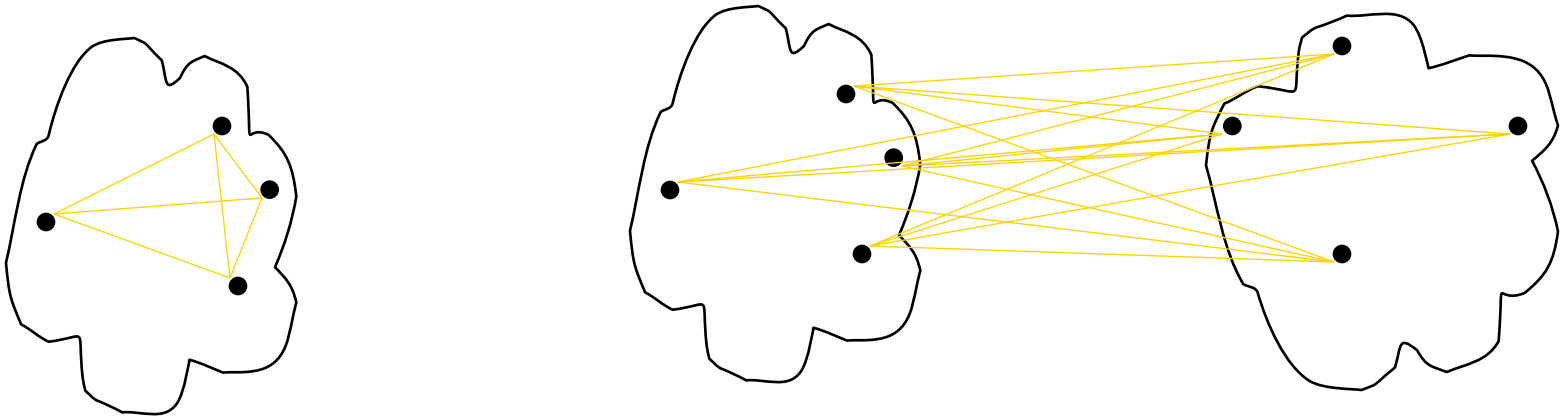
$$wc(C) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \mathbf{r}_k)$$

- Separowalność (odległości między centroidami)

$$bc(C) = \frac{1}{K(K-1)/2} \sum_{1 \le j < k \le K} d(\mathbf{r}_j, \mathbf{r}_k)$$

Inne miary: Cohesion and Separation

- A proximity graph − graf podobnieństwa można wykorzystać
  - Cluster cohesion is the sum of the weight (odległości) of all links within a cluster.
  - Cluster separation is the sum of the weights (odległości) between nodes in the cluster and nodes outside the cluster.



cohesion

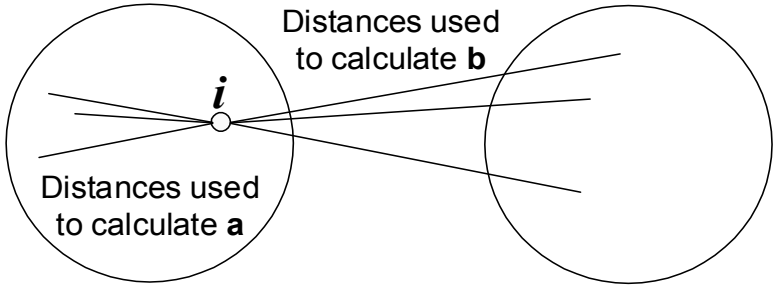
separation

Silhouette Coefficient / współczynnik sylwetkowy lub zarysu

- Współczynnik Silhouette wykorzystuje podobieństwo obiektu do innych obiektów w tym samym skupisku z odniesieniem do podobieństwa do obiektów z innych skupisk
- Dla pojedynczego obiektu ***i***
  - Oblicz ***a*** = średnia odległość obiektu ***i*** do innych obiektów w tym samym skupisku
  - Oblicz ***b*** = min (średnia odległość ***i*** punktów z innego skupiska)
  - The silhouette coefficient jest zdefiniowany jako

    $s = (b - a) / \max(a,b)$

  - Zakres od -1 do 1
  - Na ogół pomiędzy 0 i 1.
  - Im bliższe wartości 1 tym lepsze skupisko



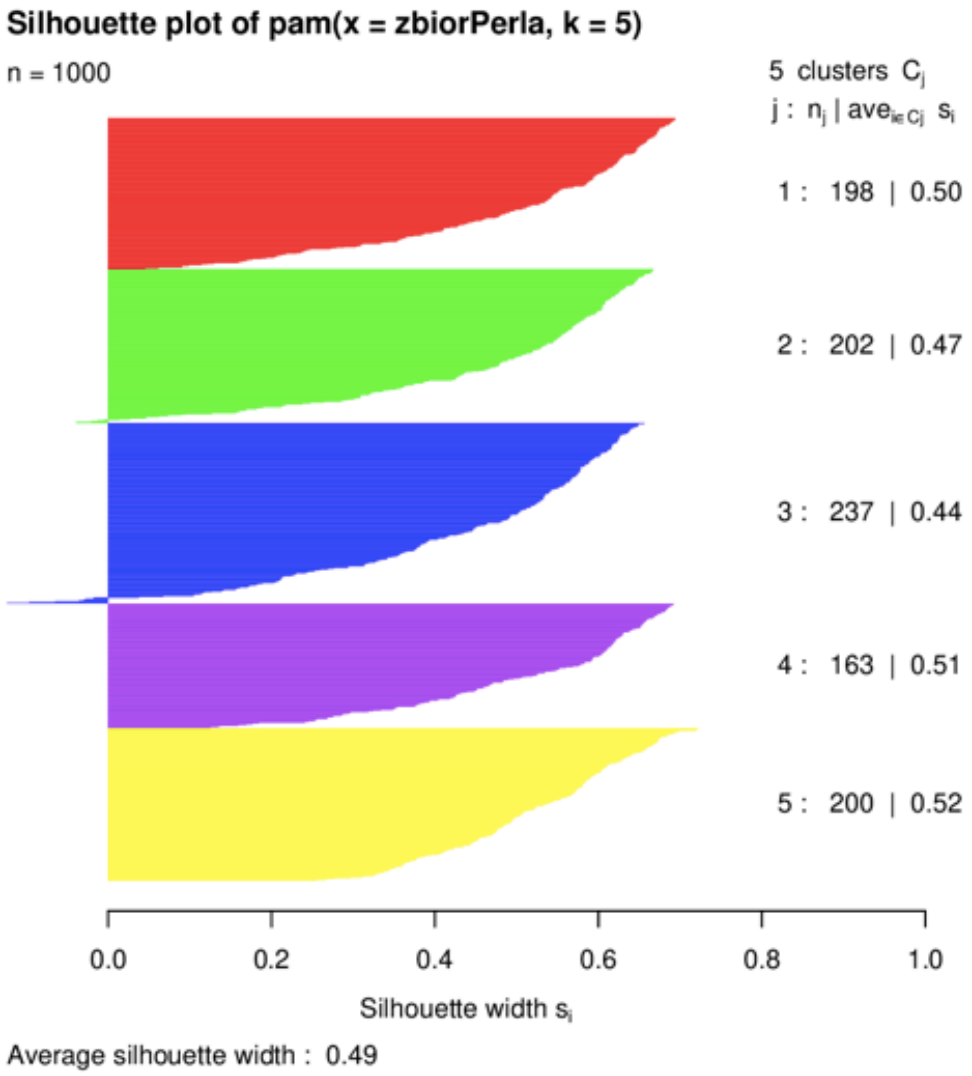
- Można dalej obliczać średnie wartości dla całego skupiska lub zbioru skupisk

Średnia wartość zarysu $\bar{s}_k$, dla każdego skupienia mówi o tym jak dobrze dane są przydzielone do tego skupienia. Z tego względu średni zarys dla całego zbioru danych może służyć jako miara jakości podziału. Współczynnik zarysu ma postać $SC = \max_k \bar{s}_k$, i jego interpretacja została zawarta w poniższej tabeli.

SC	Interpretacja
0,71-1,00	Silna struktura
0,51-0,70	Istotna struktura
0,26-0,50	Słaba struktura
$\leq$ 0,25	Brak struktury

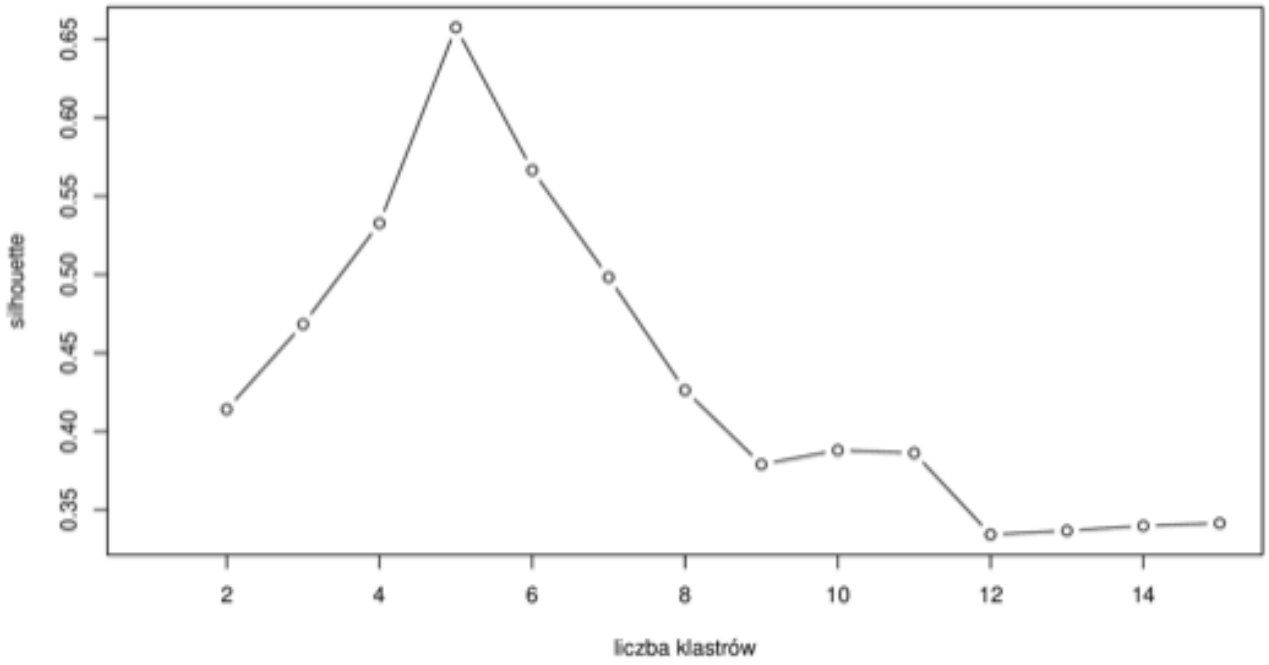
Zarys podlega wizualizacji za pomocą wykresu zarysu.

Przykład użycia współczynnika Silhouette



Rysunek 3.6: Wykres dopasowania punktów do poszczególnych klastrów z użyciem miary silhouette.

Wykorzystanie Silhouette do wyboru liczby skupisk w algorytmie PAM



Inne kryteria wewnętrznej jakości skupień

- Compactness → determining the weakest connection within the cluster, i.e., the largest distance between two objects *Ri* and *Rk* within the cluster.
- **Isolation** → determining the strongest connection of a cluster to another cluster, i.e., the smallest distance between a cluster centroid and another cluster centroid

$$\left(\sum_{C_j} \left(\frac{max(D(R_i,R_k)) \text{ where } (R_i,R_k) \in C_j}{min(D(C_{jc},C_{mc})) \text{ where } C_m \neq C_j} \right) \right)^{-1}$$

- Object positioning → the quality of clustering is determined by the extent to which each object Rj has been correctly positioned in given clusters

$$\sum_{R_i} (max(D((R_i,R_k))) - min(D(R_i,R_m)))$$

$$\text{where } (R_i,R_k) \in C_j \text{ and } R_m \notin C_j .$$

Inne współczynniki

- Współczynnik Dunna
- Wskaźnik Daviera-Bouldina
- Indeks Calińskiego i Harabasza / także b. przydatny do wyboru liczby skupień w k-średnich

- Indeks DAVIESA-BOULDINA (*ang. Davies-Bouldin index*):

$$DB = \frac{1}{n} \sum_{i=1}^{K} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)},$$

gdzie σ_i jest średnią odległością wszystkich punktów ze skupienia i do jego środka, a $d(c_i, c_j)$ jest odległością pomiędzy środkami skupień i oraz j.

- Indeks DUNNA (*ang. Dunn index*):

$$D = \frac{\min\limits_{1 \leq i < j \leq K} d(i, j)}{\max\limits_{1 \leq k \leq K} d'(k)},$$

gdzie $d(i, j)$ jest odległością pomiędzy skupieniami i oraz j, a $d'(k)$ odległością wewnątrz skupienia k.

Caliński i Harabasz (1974) zaproponowali aby końcową liczbę skupień wybierać w oparciu o wartości indeksu postaci:

$$CH(K) = \frac{\mathrm{tr}(B(C_K))/(K-1)}{\mathrm{tr}(W(C_K))/(n-K)}.$$

Optymalną wartość K dobieramy tak, aby ją zmaksymalizować.

Literatura

Caliński, T., Harabasz, J. (1974). *A dendrite method for cluster analysis*. Communications in Statistics 3(1):1–27.

tr –ślad macierzy :

Niech A będzie macierzą kwadratową stopnia n. **Śladem** macierzy A nazywamy wielkość

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}.$$

Rozkład całkowitej sumy kwadratów

Załóżmy, że dokonaliśmy podziału na K skupień.

$C(i) = k$ – gdy x_i należy do $k-$tego skupienia.

T – suma kwadratów odległości między elementami tego samego skupienia i różnych skupień

$$T = W + B$$

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d_{ii'}$$

$$B = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'}$$

(T –Total, W – Within, B – Between)

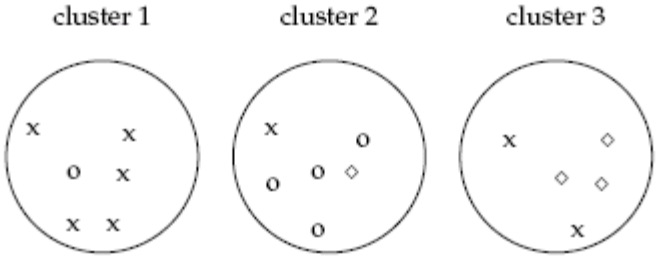
Minimalizacja W, czyli minimalizacja rozrzutu punktów wewnątrz skupień $\equiv$ maksymalizacji rozrzutu punktów między skupieniami.

Zewnętrzne miary

- Porównuje się strukturę skupisk vs. znany podział z zbioru testowego
- Część miar wykorzystuje specjalną tablicę dwudzielczą
- Na ogół stosowany do oceny algorytmu i porównania go do innych
- Popularne miary
  - Wskaźnik Randa
  - Purity
  - Odmiany miary F
- Przegląd miar − dostępny na https://en.wikipedia.org/wiki/Cluster_analysis

Ocena jakości algorytmu gdy znany jest właściwy przydział do klas (ang. ground truth)

Jain's example



► **Figure 16.4** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ◇, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

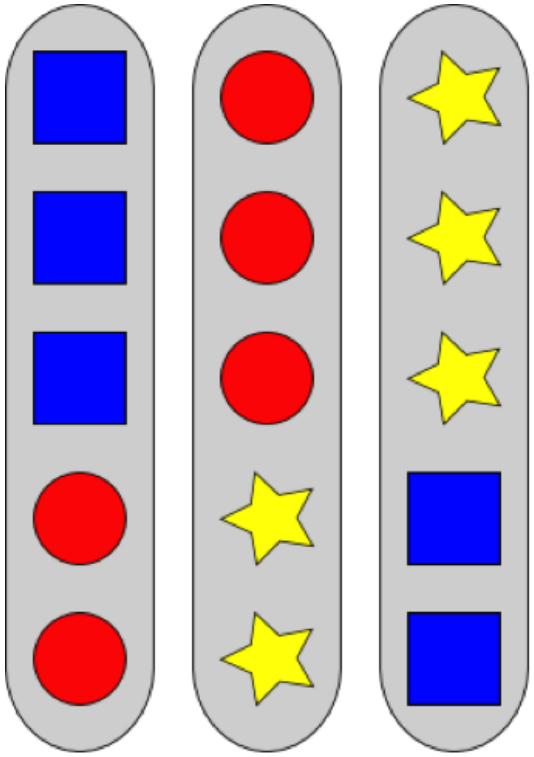
	purity	NMI	RI	F_5
minimum	0.0	0.0	0.0	0.0
maximum	1	1	1	1
value for Figure 16.4	0.71	0.36	0.68	0.46

► **Table 16.2** The four external evaluation measures applied to the clustering in Figure 16.4.

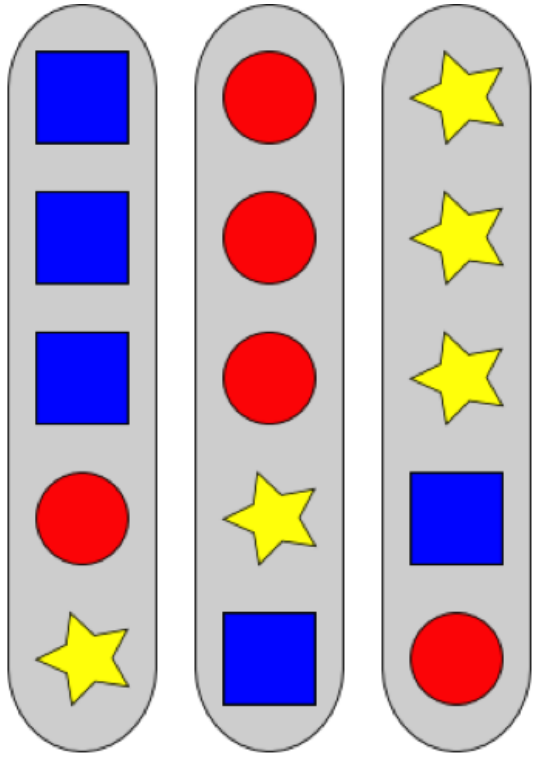
Odniesienie do zewnętrznego podziału

- Dostępne referencyjne etykiety (manually labeled data)
  - Ekspert etykietuje w zależności od własności danych
  - Istnieją benchmarki TREC, Reuters, itp..
  - Tryb sem-supervised
- Różne podejścia:
  - „Accuracy of clustering: Percentage of pairs of tuples in the same cluster that share common label” (etykiety w skupieniach)
    - Faworyzujemy małe „czyste” skupienia
    - Czy powinniśmy mieć zgodność liczby skupień i etykiet

Pewne problemy w ocenie odwzorowań



F-measure: 0.6



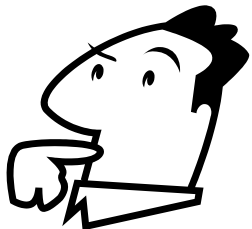
F-measure: 0.6

Ogólne zasady

- Homogeneity - Jednorodności
  - Każde skupienie zawiera przykłady z jak najmniejszej liczby etykiet klas
  - Idealnie – tylko jedna klasa
- „Completeness”
  - Każda klasy reprezentowana w możliwe najmniejszej liczbie skupień
- Typowe miary
  - Purity
  - F-miara

Ocena grupowania

- Inna niż w przypadku uczenia nadzorowanego (predykcji wartości)
- Poprawność grupowania zależna od oceny obserwatora / analityka
- Różne metody AS są skuteczne przy różnych rodzajach skupień i założeniach, co do danych:
  - Co rozumie się przez skupienie, jaki ma kształt, dobór miary odległości → sferyczne vs. inne
- Dla pewnych metod i zastosowań:
  - Miary zmienności wewnątrz i między – skupieniowych
  - Idea zbiorów kategorii odniesienia (np. TREC)



Analiza skupień - podsumowanie

- Liczne i ważne zastosowanie praktyczne analizy skupień (AS).
- AS używana „samodzielnie” w zgłębianiu danych, lub jako jedno z narzędzi podczas wstępnego przetwarzania w procesie KDD.
- Jakość skupień i działanie wielu algorytmów związane są określeniem miary odległości obiektów.
- Podstawowe klasy metod:
  - hierarchiczne,
  - podziałowo/optymalizacyjne,
  - gęstościowe,
  - „grid-based”,
  - wykorzystujące modele matematyczne (np. probabilistyczne lub neuronowe).
- Ważne zagadnienie to także wykrywanie obiektów nietypowych (outliers discovery).

Więcej w książkach, artykułach

Szukaj też samodzialnie

Wybrane źródła literaturowe

- A. D. Gordon: Classification. Chapman & Hall 1999
- B. S. Everitt, S. Landau, M. Leese, Cluster analysis, Oxford University Press, 2001

Może pytanie lub komentarze?

?

Time is going on …

Pytanie i komentarze?

Dalszy kontakt:

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