Globalne zależności w klastrowaniu hierarchicznym

Michał Woźniak

23 listopada 2008

Plan prezentacji

Praca źródłowa

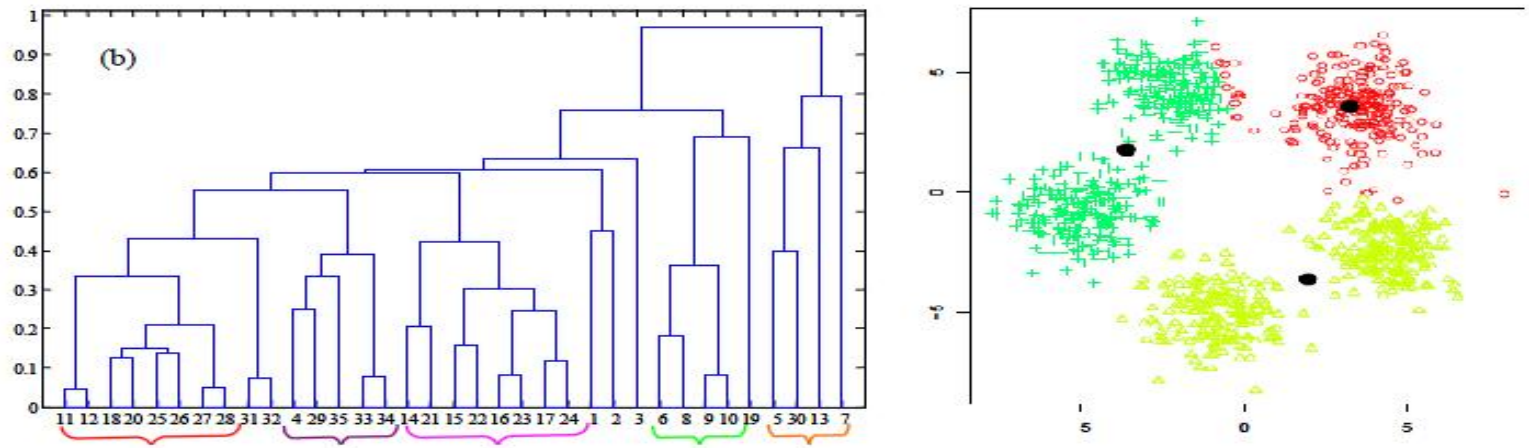
Global Considerations in Hierarchical Clustering Reveal Meaningful Patterns in Data

Roy Varshavsky (1), David Horn (2), Michal Linial (3)

1. *School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel*
2. *School of Physics and Astronomy, Tel Aviv University, Israel*
3. *Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Israel*

Dlaczego klastrować hierarchicznie?

- Dla wielu zagadnień ewolucyjnych podejście hierarchiczne jest naturalne
- Wynikiem klastrowania hierarchicznego jest drzewo, wygodne w analizie eksperckiej



Rysunek: Przykładowe wyniki klastrowania uzyskane metodą hierarchiczną i algorytmem k-średnich

Bottom-Up vs Top-Down

Metody klastrowania hierarchicznego możemy podzielić na 2 rodziny: *Bottom-Up* oraz *Top-Down*
Algorytmy z rodziny *Bottom-Up* są zdecydowanie bardziej popularne niż *Top-Down*: na 86 publikacji, których wykorzystywano klastrowanie hierarchiczne, tylko 3 wykorzystywały podejście *Top-Down*.
Prawdopodobne przyczyny:

- Większa intuicyjność podejścia *Bottom-Up*
- Większa dostępność wśród gotowych pakietów oprogramowania

Przykłady Top-Down I

Bisecting K-Means algorithm

Polega na iteracyjnym dzieleniu zbiorów przy pomocy algorytmu K-Średnich (k=2), ma typowe wady algorytmu k-średnich, takie jak tworzenie sferycznych klastów, oraz zależność od warunków początkowych.

Przykłady Top-Down II

Principal Direction Divisive Partitioning (PDDP)

Bazuje na iteracyjnym podziale zbioru przykładów w zależności od znaku ich rzutu na pierwszą składową główną. Pierwszą składową główną możemy definiować jako:

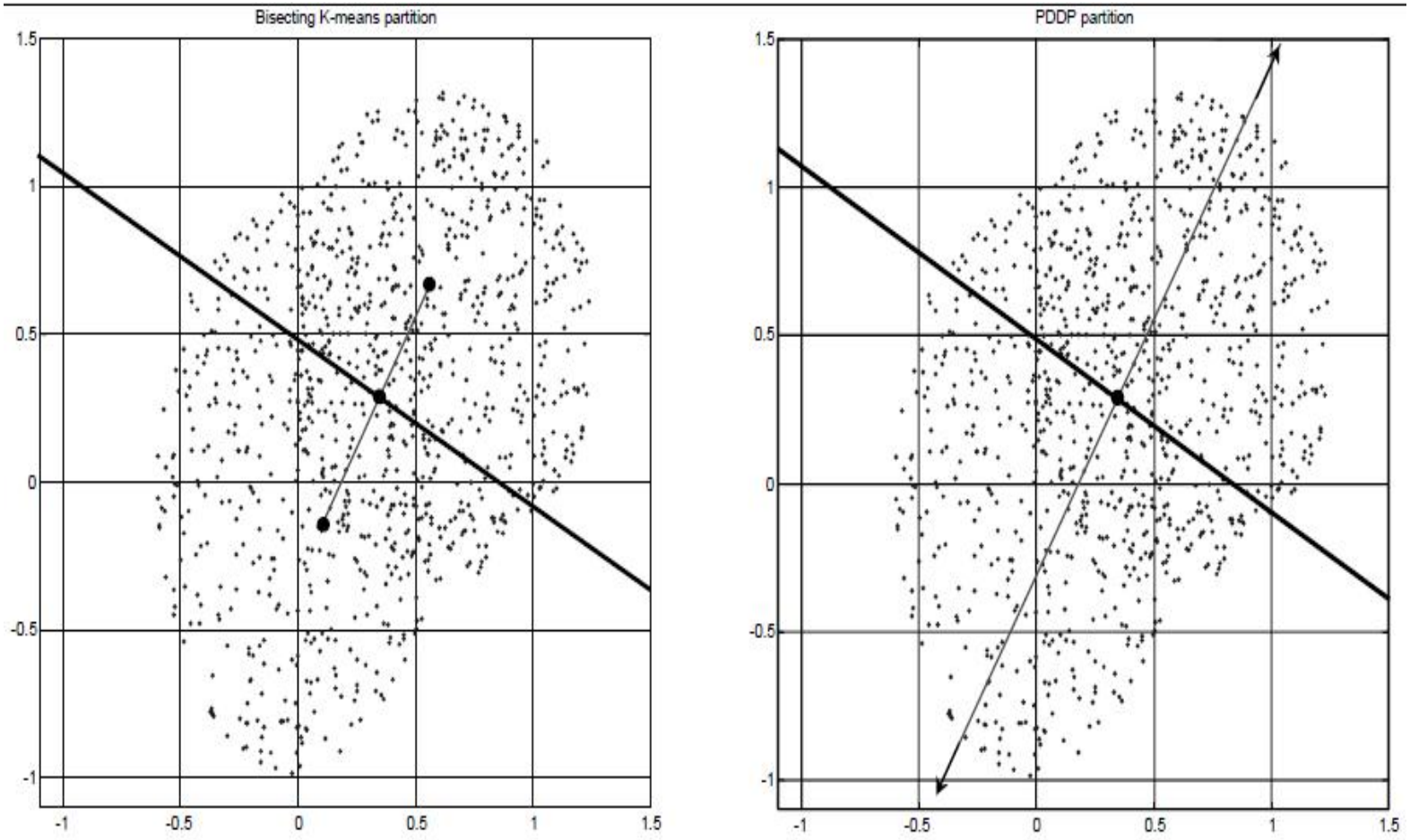
$$w_1 = \underset{\|w\|=1}{\operatorname{argmax}}\, var(w^T x)$$

Niech c - centroid, gdzie $c = \frac{1}{N}\sum_{i=1}^{N} x_i$ Definiujemy podział zbioru na 2 klastry M_L i M_R:

$$x \in \begin{cases} M_L & w_1^T(x-c) \geqslant 0 \\ M_R & w_1^T(x-c) < 0 \end{cases}$$

Algorytm PDDP zazwyczaj osiąga lepsze wyniki oraz jest stabilniejszy, w pracy posłużył do porównywania.

Podział K-means i PDDP



Rysunek: Podział zbioru na 2 klastry metodami K-means i PDDP

Algorytm

1. Zdefinuj oryginalny zbiór danych (Liczba zbiorów =1)
2. [*Opcjonalnie*] Przeprowadź preprocessing dla każdego zbioru
3. Uruchom algorytm QC (Quantum Clustering) dla każdego zbioru
4. Podziel każdy zbiór na 2 zawierające:
   - Przykłady należące do klastru zawierającego globalne minimum funkcji V (o tym za chwilę)
   - Pozostałe
5. Rekurencyjnie skocz do 2 dla każdego zbioru zawierającego więcej niz 2 przykłady

Metoda SVD

Twierdzenie

Każdą macierz rzeczywistą M o wymiarach m na n można przedstawić w postaci rozkładu SVD:

$$M = U\Sigma V^T$$

gdzie:

- U i V - macierze ortonormalne, tzn $U^{-1} = U^T$
- Σ - macierz diagonalna, taka że $\Sigma = diag(\sigma_i)$, gdzie σ_i to wartości nieujemne uporządkowane malejąco

Metoda SVD

Zastosowanie

Metodę SVD stosuje sie w celu redukcji liczby wymiarów, oraz uzyskania danych "wolnych od hałasu".

Macierz danych $X = U\Sigma V^T$ redukujemy poprzez wyzerowanie diagonali macierzy Σ dla indeksów $i > r$: $X_r = U\Sigma_r V^T$

Trundność polega na rozsądnym wybraniu r, na ogół szuka się "kolanka" na wykresie sumarycznego udziału kolejnych wartosci własnych $f(r) = \sum_{i=1}^{r} \sigma_i$.

Teoria

Gęstość rozkładu prawdopodobieństwa

Konstruujemy gęstość prawdopodobieństwa pojawienia się "cząstki" y w d-wymiarowej przestrzeni:

$$\psi(y) = \sum_{i=1}^{n} e^{-\frac{\|y-x_i\|^2}{2\sigma^2}}$$

Maksima funkcji ψ odpowiadają centrom klastrów.

Równanie Schrödingera

Równanie Schrödingera w postaci niezależnej od czasu:

$$H\psi = (-\frac{\sigma^2}{2}\Delta + V(y))\psi(y) = E\psi(y)$$

gdzie Δ to operator Laplace'a

Teoria II

Funkcja potencjału V

Wyznaczamy funkcję V, będącą potencjałem Schrödingera.

$$V(y) = E + \frac{\frac{\sigma^2}{2}\Delta\psi(y)}{\psi(y)} = E - \frac{d}{2} + \frac{1}{2\sigma^2\psi(y)}\sum_{i=1}^{n}(y - x_i)^2 e^{-\frac{\|y-x_i\|^2}{2\sigma^2}}$$

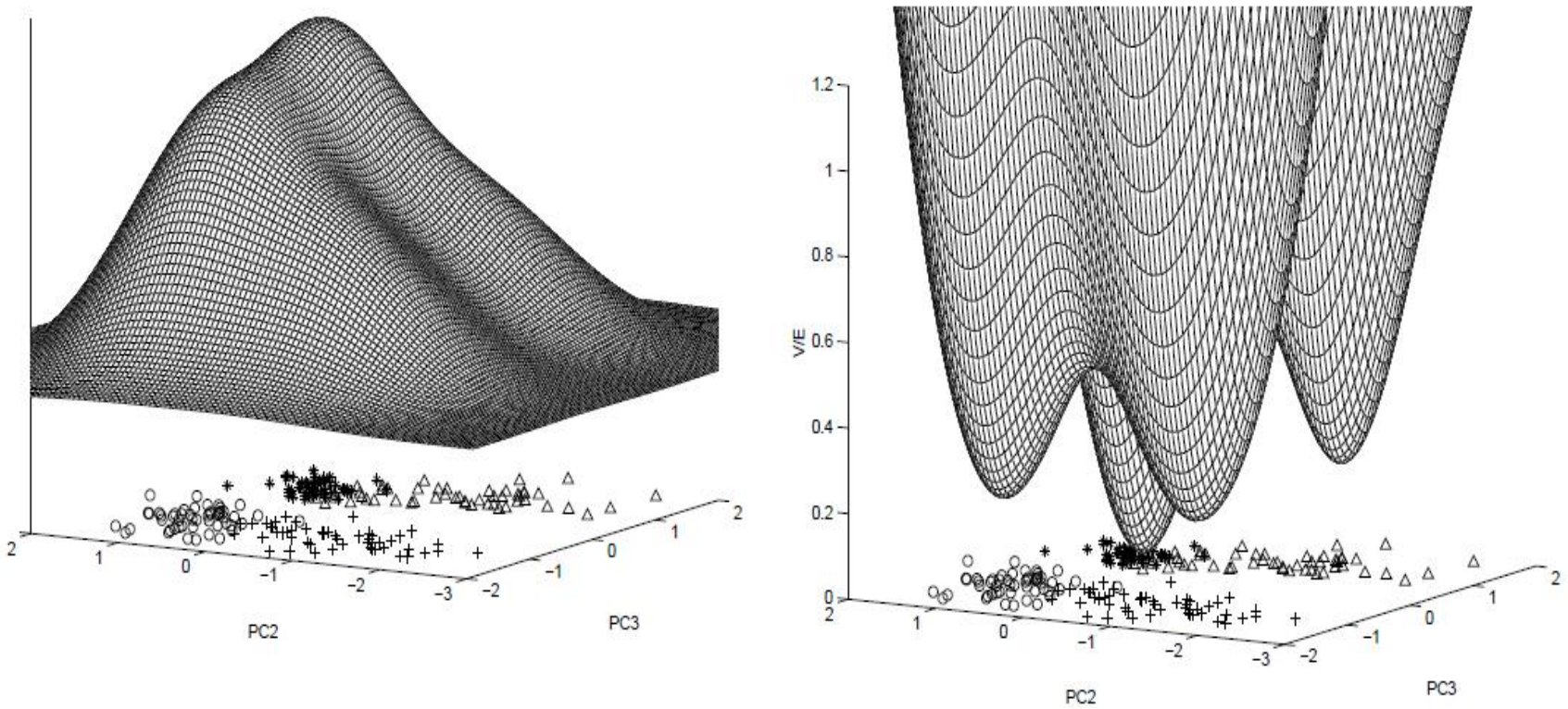
Definiujemy V jednoznacznie zadając $E = -\min_y \frac{\frac{\sigma^2}{2}\Delta\psi(y)}{\psi(y)}$.

Przykład: $n = 1$

$$V(y) = E - d/2 + \frac{1}{2\sigma^2}(y - x_1)^2$$

Funkcja kwadratowa, która minimum osiąga dla $y = x_1$.

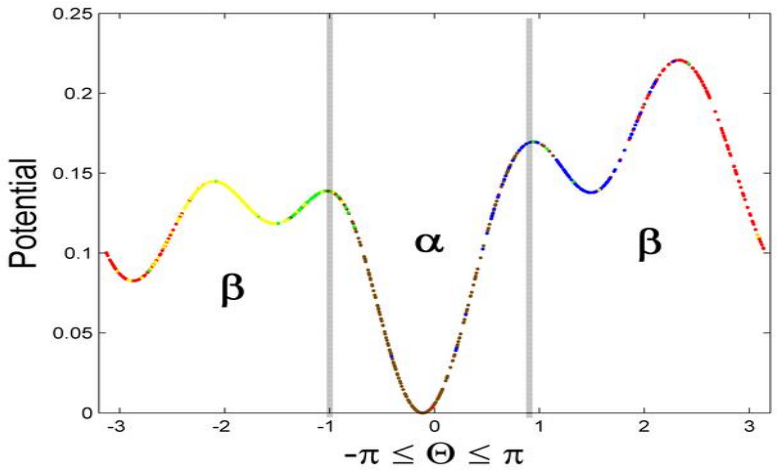
Obrazki



Rysunek: Wykresy funkcji ψ oraz odpowiadającej jej funkcji potencjału Schrödingera V

Wykorzystanie

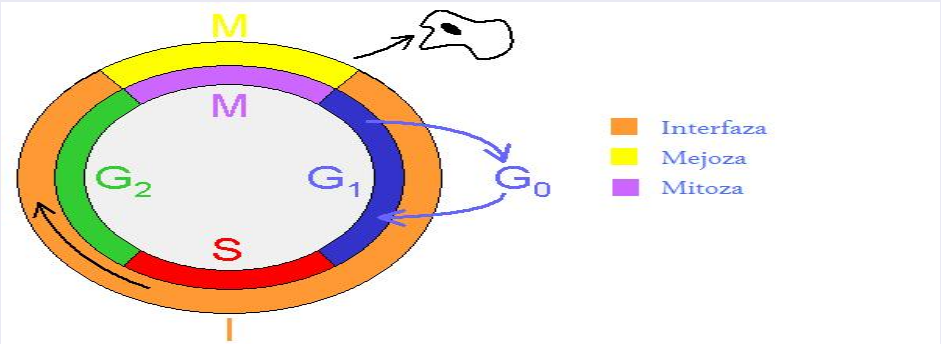
1. Wejście: Dane po zastosowaniu SVD (r=2)
2. Rzutujemy dane na przestrzeń jednowymiarową, licząc odległości między obserwacjami przy pomocy funkcji *cos*: $d(x_1, x_2) = 1 - cos(\phi_1, \phi_2)$.
3. Wyznaczmy funkcję potencjału Schrödingera algorytmem QC.
4. Wyjście: Podział zbioru na 2 na klastry, jak na rysunku poniżej:



Zbiory danych

Cell Cycle genes

798 genów zidentyfikowanych jako regulujące cykl komórkowy drożdży piekarniczych (*yeast saccharomyces cerevisiae*), podzielone na 5 klas w zależnosci od fazy cyklu, która jest przez nie regulowana(oznaczone przez M/G1, G1, S, G2 i M), ekspresję genów zbadano w 72 momentach cyklu.

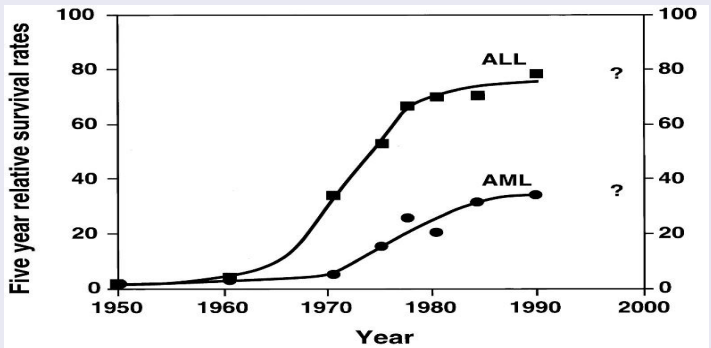


Rysunek: Cykl komórkowy

Zbiory danych II

Leukemia patients

Eksperyment wskazał 72 pacjentów z 2 typami białaczki ALL i AML, ALL podzielono na 2 rodziaje: T-cell i B-cell, zbiór AML został podzielony na pacjentów, ktorzy przeszli leczenie i nie. Dla każdego pacjenta została zarejestrowana ekspresja 7129 genów. Zadanie klastrowania polega na znalezieniu 4 rodzajów raka w zbiorze 72 pacjentów.



Rysunek: Współczynnik przeżywalności dla białaczki

Zbiory danych III

Standard and Poor

Zbiór danych firmy Standard and Poor (S&P), która zawiera dzień po dniu zmiany w cenie akcji w czasie 273 dni targowych. 487 akcji zostało podzielone na 10 klas.

Ion Channel proteins

Dane zostały pobrane z bazy SwissProt. Dla 614 białek uruchomiono BLAST każdy z każdym, otrzymano macierz odległości. 518 z tych białek zostało odnalezione w bazie InterPro. Na podstawie informacji z InterPro ze zbiorem bialek zostały skojarzone 3 poziomy szczegółowości: 3, 11, 19 klastrów.

Kryteria I

Zakładamy, że posiadamy wiedzę ekspercką na temat wszystkich obserwacji.

Node Score

P-wartość zaobserowanego przypisania k obiektów do klasy zawierającej n elementów dana jest wzorem:

$$P = \sum_{x=k}^{n} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

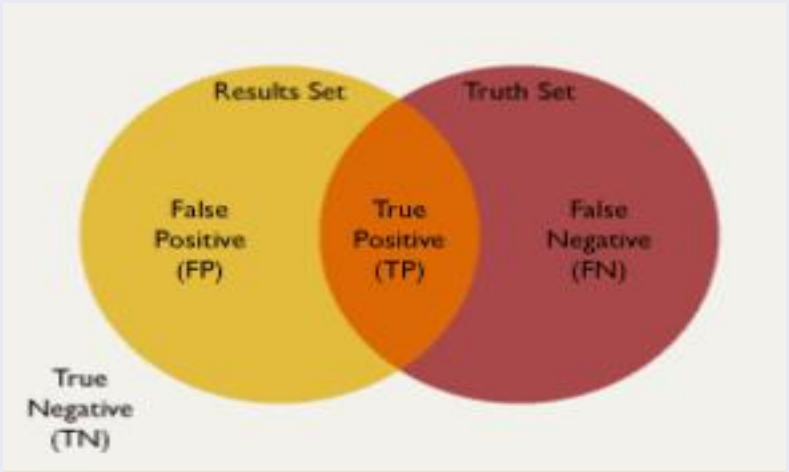
gdzie: N - ilość wszystkich obserwacji, K - ilość obserwacji w rozpatrywanym zbiorze.

Kryteria II

Level Score

Poziom nr *l* w drzewie zawiera wszystkie wierzchołki oddzielone *l* krawędziami od korzenia. Dla każdego wierzchołka na danym poziomie wyznaczono punkty Jaccarda:

$$J = \frac{TP}{TP + FN + FP}$$



Punkty *level score*, zostały zdefiniowane jako średnia punktów Jaccarda ze wszystkich wierzchołków na danym poziomie.

Kryteria III

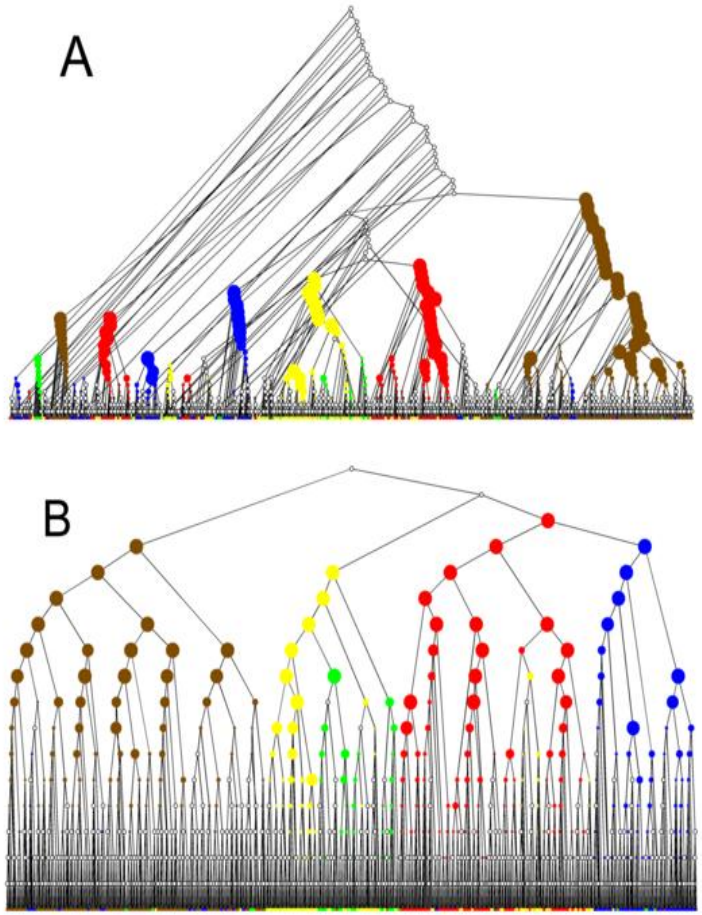
Tree Score

Zdefiniowano miarę ważoną *best-J-Score*:

$$J^* = \frac{1}{N} \sum_{i}^{C} n_i J_i^*$$

gdzie: J_i^* to najlepsza wartość punktów Jaccarda, dla klasy i, C-ilość klas, n_i ilość przykładów należących do klasy i, oraz N - ilość wszystkich klas.

Eksperyment 1



Rysunek: Porównanie drzew dla zestawu "Cell cycle genes", dla BU-Average(A) oraz TDQC(B)

Eksperyment 2

	Elements	Features	Classes	Non-hierarchical		Hierarchical				
				QC	K-Means	BU			TD	
						BU-S	BU-A	BU-C	PDDP	TDQC
Cell cycle	798	72	5	0.613	0.537 (0.06)	0.265	0.472	0.409	0.542	**0.646**
Leukemia	72	7129	4	0.758	0.519 (0.1)	0.465	0.522	0.53	0.545	**0.804**
S&P	487	273	10	0.400	0.306 (0.05)	0.2	0.261	0.445	0.441	**0.504**

K-Mean was performed 10 times and averaged (and std is in parenthesis), Hierarchical algorithms are BU (S, A and C marks the Single, Average, Complete, respectively) and TD (PDDP, TDQC) algorithms. Best scores are bold faced.

Rysunek: Porównanie algorytmów metodą *tree-score*

Eksperyment 3

- Porównanie "zrównoważenia" drzew uzyskiwanych przez różne algorytmy zastosowane do 4 zbiorów danych.
- Miara "zrównoważenia" drzewa: $\frac{D-log_2 N}{N-log_2 N}$