

Grupowanie spektralne: Podstawy i nowe kierunki

Sławomir T. Wierzchoń

<http://www.ipipan.waw.pl/~stw/sem2101/sem.pdf>

Instytut Podstaw Informatyki PAN



Seminarium

Zakładu Strukturalnych Metod Przetwarzania Wiedzy

Warszawa, 21 stycznia 2016

Treść

1 Wprowadzenie

- Dlaczego tak wiele algorytmów?
- Podstawowe grupy algorytmów

2 Grupowanie spektralne

- Krótka charakterystyka
- Podstawy
 - Konstrukcja grafu podobieństwa
 - Laplasjan grafu
 - Wartość, wektor, porządek Fiedlera
- Kryteria rozcinania grafu
- Aproksymacja Nyström'a

3 Nowe kierunki

- p -Laplacian
- 1-Laplasjan

Dlaczego tak wiele algorytmów?

Czym jest skupienie (grupa)?

„Skupienie to **maksymalnie koherentny** zbiór obserwacji”.

- V. Estivill-Castro. “Why so many clustering algorithms: A position paper”. *SIGKDD Explor. Newsl.*, 4(1):65-75, Jun. 2002.
- M. Pelillo. “What is a cluster?”. A NIPS 2009 Workshop “Clustering: Science or Art? Towards Principled Approaches”.

Podstawowe grupy algorytmów:

- Metody podziałowe (kombinatoryczne)
- Metody relacyjne (hierarchiczne)

Metody podziałowe

Założenie: $X = \{x_1, \dots, x_m\} \subset \mathbb{R}^n$. Znaleźć podział $C_1, \dots, C_k$ zbioru X .

Klasyczny przypadek: $\|x_i - x_j\|$ to odległość euklidesowa.
Znaleźć zbiór k prototypów $\mu_1, \dots, \mu_k \in \mathbb{R}^n$ minimalizujących

$$J(C_1, \dots, C_k) = \sum_{j=1}^k \sum_{i \in C_j} \|x_i - \mu_j\|^2$$

gdzie C_j to zbiór obiektów, których prototypem jest μ_j .

- A. Jain. "Data clustering: 50 years beyond k -means". *Pat. Rec. Lett.*, 31:651-666, 2010.

Ojcowie ...

- H. Steinhaus (1956): Sur la division des corp matériels en parties. *de l'Académie Polonaise des Sciences IV* (C1.III), 801–804
- S.P. Lloyd (1957): Least squares quantization in PCM. Bell Telephone Labs Memorandum, Murray Hill, NJ. Reprinted in: *IEEE Trans. Information Theory*, IT-28 (1982), vol. 2, 129-137
- E.W. Forgy (1965): Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometric Society Meeting, Riverside, California, 1965.
- J. MacQueen (1967): Some methods for classification and analysis of multivariate observations. In: L.M. LeCam, J. Neyman (eds.): *Proc. 5th Berkely Symp. Math. Statist. Probab.* 1965/66. Univ. of California Press, Berkely, vol. I,

... i synowie

- G. Ball, D., Hall (1965): ISODATA, a novel method of data analysis and pattern classification. Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA
- J.C. Bezdek (1981): *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York
- ...
- G. Cleuziou. A generalization of k -means for overlapping clustering. Rapport No RR-2007-15. Université d'Orléans, LIFO
- A. Banerjee, *et al.* Clustering with Bregman divergences. J. Mach. Learn. Res., 6:1705-1749, 1 Dec. 2005

Dywergencja Bregmana

Niech $\phi: \mathcal{S} \rightarrow \mathbb{R}$, $\mathcal{S} = \text{dom}(\phi) \subset \mathbb{R}^n$ będzie funkcją ściśle wypukłą i różniczkowalną w $\text{rint}(\mathcal{S})$. Dywergencja Bregmana to funkcja

$$d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^T \nabla \phi(\mathbf{y})$$

Nie musi spełniać warunku trójkąta, ani być symetryczna.

Example

- (1) Jeżeli $\phi(\mathbf{x}) = \|\mathbf{x}\|^2 = \mathbf{x}^T \mathbf{x}$, to $d_\phi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$
- (2) Jeżeli $\phi(\mathbf{p}) = \sum_{j=1}^n p_j \log_2 p_j$, gdzie $\sum_j p_j = 1$, to $d_\phi(\mathbf{p}, \mathbf{q}) = KL(\mathbf{p} \parallel \mathbf{q})$

Algorytm BHC (*Bregman Hard Clustering*)

$X = \{x_1, \dots, x_m\} \subset \mathcal{S} \subseteq \mathbb{R}^n$; π – rozkład p-stwa na $\mathcal{X}$. Znaleźć, dla danej dywergencji d_ϕ , podział $\mathcal{C}$ minimalizujący

$$B_m(\mathcal{C}, M) = \sum_{i=1}^k \sum_{j \in C_i} v_j d_\phi(\mathbf{x}_j, \mu_i)$$

Własności algorytmu BHC:

- Osiąga w skończonej liczbie kroków optimum lokalne.
- Prototypy mają postać $\mu_i = \frac{1}{|C_i|} \sum_{j \in C_i} \pi_j x_j$. Tak jest **tylko** dla dywergencji Bregmana.
- k -średnich, algorytm Linde-Buzo-Gray'a i teorio-informacyjny algorytm (Dhillon, Mallela & Kumar, 2003) to przykłady BHC.
- Granice między grupami wyznaczone są przez hiperpłaszczyzny postaci $\{x: d_\phi(x, \mu_i) = d_\phi(x, \mu_{i+1})\}$.

Wady algorytmów podziałowych

- Musi być $x_i \in \mathbb{R}^n$.
- Każda obserwacja **musi** być przydzielona do jakiejś grupy.
- Wrażliwość na inicjalizację
- Wrażliwość na liczbę wymiarów

Algorytmy relacyjne

$W: X \times X \rightarrow \mathbb{R}$ – podobieństwo/odmienność między parami obiektów. Zakładamy, że W jest macierzą wymiaru $m \times m$.

- J. Czekanowski. *Zarys metod statystycznych w zastosowaniu do antropologii*. Towarzystwo Naukowe Warszawskie, 1913
- P.L. Graham, P. Hell. On the history of the minimum spanning tree problem, *IEEE Ann. of the History of Computing*, 7(1),43-57, 1985

Algorytm Czekanowskiego

- Wyznacz elementy macierzy W .
- Ustal dyskretyzację wartości w_{ij} , a przynależność w_{ij} do konkretnego przedziału oznacz innym symbolem graficznym (diagram Czekanowskiego).
- Uporządkuj wiersze i kolumny macierzy celem wizualizacji wewnętrznej struktury.

Example

	1	2	3	4	5	6
1	1.0	0.2	0.1	0.3	0.8	0.4
2	0.2	1.0	0.1	0.9	0.3	0.3
3	0.1	0.1	1.0	0.2	0.2	0.7
4	0.3	0.9	0.2	1.0	0.4	0.1
5	0.8	0.3	0.2	0.4	1.0	0.2
6	0.4	0.3	0.7	0.1	0.2	1.0

	1	2	3	4	5	6
1	■	□	□	□	■	□
2	□	■	□	□	■	□
3	□	□	■	□	□	■
4	□	■	□	□	□	□
5	■	□	□	□	□	□
6	□	□	■	□	□	■

	1	5	2	4	3	6
1	■	■	□	□	□	□
5	■	■	□	□	□	□
2	□	□	■	■	□	□
4	□	□	■	■	□	□
3	□	□	□	□	■	■
6	□	□	□	□	■	■

Co to jest grupowanie spektralne?

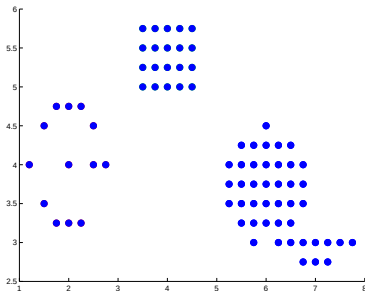
Problem: podzielić zbiór obserwacji $X = \{x_1, \dots, x_m\}$ na k rozłącznych grup.

- Utwórz symetryczną macierz wag $W: X \times X \rightarrow \mathbb{R}$.
- W oparciu o W utwórz graf podobieństwa $G(V, E)$, gdzie $V \equiv X$. Niech S będzie uogólnioną macierzą sąsiedztw.
- Oblicz **laplasjan**, $\mathbb{L}$, i wyznacz jego rozkład spektralny, $\mathbb{L} = Y\Lambda Y^T$.
- Utwórz **odwzorowanie spektralne** $V \ni v_i \mapsto \mathbf{y}_i^* = (y_{i1}, \dots, y_{ik})$.
- Podziel zbiór $\mathbf{Y}^* = \{\mathbf{y}_1^*, \dots, \mathbf{y}_i^*\}$ stosując dowolny algorytm grupowania.

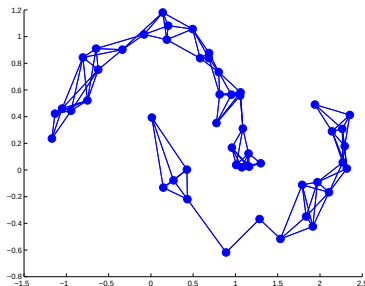


Krótka charakterystyka

Zwartość a spójność



(a)



(b)

Rysunek: Zwarte i dobrze separowalne skupienia (a); skupienia opisane lokalną informacją o powiązaniach między węzłami (b).

Metody grafowe a metody spektralne

- Metody rozcinania grafu, oparte na optymalizacji pewnej funkcji celu, są na ogół NP-trudne. Metody spektralne umożliwiają relaksację dyskretnego zadania optymalizacyjnego i w czasie wielomianowym zwracają rozwiązania, które należy binaryzować.
- von Luxburg (2007): “The reason why the spectral relaxation is so appealing is not that it leads to particularly good solutions. Its popularity is mainly due to the fact that it results in a standard linear algebra problem which is simple to solve.”
- Metody spektralne są stosowane w nielinowej redukcji wymiarowości, tzw. **manifold learning**: odkrywaniu niskowymiarowej rozmaitości, na której leżą wysokowymiarowe dane. Np: Laplacian eigenmaps, Diffusion maps, etc.

Literatura

- F. Chung. *Spectral Graph Theory*. AMS, Providence, RI, 1997
- strona: Applications of spectral graph theory
- U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395-416, 2007
- H. Jia, *et al.* The latest research progress on spectral clustering. *Neural Comput. & Applic.*, 24(7-8):1477-1486, 2014
- M. Filippone, *et al.* A survey on spectral and kernel methods for clustering. *Pattern Recognition*, 41(1):176-190, 2008

Zakres stosowalności

D.A. Spielman and S.-H. Teng. Spectral partitioning works: Planar graphs and finite element meshes. *Linear Algebra and its Applications*, 421(2-3), 284–305, 2007. DOI: 10.1016/j.laa.2006.07.020:

(...) spectral partitioning methods work well on bounded-degree planar graphs and finite element meshes – the classes of graphs to which they are usually applied.

Jak wyznaczać wagi?

■ Typowy wybór:

$$w_{ij} = \exp \left(- \frac{\|x_i - x_j\|^2}{\sigma^2} \right)$$

Uzasadnienie podają Belkin i Niyogi (2003). Ale: (a) podobieństwo zależy tylko od odległości, (b) jak wybrać σ ?

- Niech x_i^K – K -ty najbliższy sąsiad obserwacji x_i i niech $\sigma_i = \|x_i - x_i^K\|$. Wtedy

$$w_{ij} = \exp \left(- \frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j} \right)$$

gdzie $K \approx \lfloor \ln m \rfloor$ – por. (Zelnik-Manor & Perona, 2000).

Graf podobieństwa

$W \in \mathbb{R}^{m \times m}$ – symetryczna macierz podobieństwa/odmienności, przy czym $w_{ii} = 0$, $i = 1, \dots, m$, $w_{ij} \geq 0$.

$S = [s_{ij}]$ – uogólniona (ważona) macierz sąsiedztw reprezentująca graf $G = (V, E)$, $V \equiv X$.

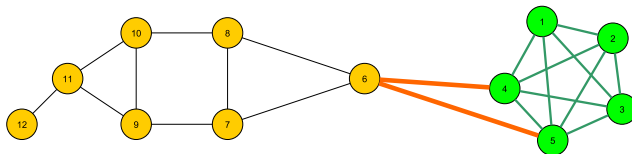
- Graf **ϵ -sąsiedztw**: $\{v_i, v_j\} \in E \Leftrightarrow s_{ij} > \epsilon$. Ponadto: $s_{ij} = 1$ gdy $\{v_i, v_j\} \in E$, $s_{ij} = 0$ w p.p.
- Graf **K -najbliższych sąsiadów** (K -NN). Niech $N_K(v_i)$ – zbiór K najbliższych (najbardziej podobnych) węzłów do v_i . Wtedy

$$\{v_i, v_j\} \in E \Leftrightarrow v_i \in N_K(v_j) \wedge v_j \in N_K(v_i)$$

Ponadto $s_{ij} = w_{ij}$ gdy $\{v_i, v_j\} \in E$, $s_{ij} = 0$ w p.p.

Koszt rozcięcia grafu

$$\text{cut}(A, B) = \sum_{i \in A, j \in B} s_{ij}, \quad \text{assoc}(A, A) = \sum_{i, j \in A} s_{ij}$$



Rysunek: Kolor pomarańczowy – *cut*, zielony – *assoc*

Szukamy podziału (A, B) o minimalnym koszcie *cut* lub maksymalnym *assoc*.

Laplasjan

Oznaczenia:

- $d_i = \sum_{j=1}^m s_{ij}$ – „typowość” wężła v_i ; $D = \text{diag}(d_1, \dots, d_m)$.
- χ – wektor charakterystyczny podziału (A, B) , tzn.: $\chi_j = +1$ gdy $v_j \in A$ oraz $\chi_j = -1$ gdy $v_j \in B$.

$$\chi^T D \chi = \sum_{j=1}^m d_j \chi_j^2 = \left(\sum_{\substack{i \in A \\ j \in A}} s_{ij} \right) + \left(\sum_{\substack{i \in A \\ j \in B}} s_{ij} \right) + \left(\sum_{\substack{i \in B \\ j \in B}} s_{ij} \right) + \left(\sum_{\substack{i \in B \\ j \in A}} s_{ij} \right)$$

$$= \text{assoc}(A, A) + \text{assoc}(B, B) + 2\text{cut}(A, B)$$

$$\chi^T S \chi = \sum_{i=1}^m \sum_{j=1}^m s_{ij} \chi_i \chi_j = \text{assoc}(A, A) + \text{assoc}(B, B) - 2\text{cut}(A, B)$$

$$\chi^T (D - S) \chi = \chi^T L \chi = 4\text{cut}(A, B)$$

Własności laplasjanu $L = (D - S)$ (1)

- dla dowolnego wektora $x \in \mathbf{R}^m$

$$x^T L x = x^T (D - S) x = \sum_{i,j} s_{ij} (x_i - x_j)^2 \geq 0$$

- L jest macierzą symetryczną i dodatnio-półokreśloną o elementach

$$l_{ij} = \begin{cases} d_i - s_{ij} & \text{gdy } i = j \\ -s_{ij} & \text{gdy } i \neq j \text{ oraz } \{v_i, v_j\} \in E \\ 0 & \text{w p.p.} \end{cases}$$

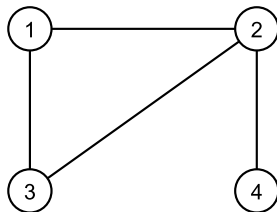
czyli wartości własne $0 = \lambda_1 \leq \dots \leq \lambda_m$.

- $(x = \text{const}) \Rightarrow Lx = \mathbf{0} \Rightarrow \lambda_1 = 0 \wedge y_1 = \mathbf{e}/\sqrt{m}$.

Graf i jego laplasjan

$$S = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \mathbf{d} = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

$$L = D - S = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

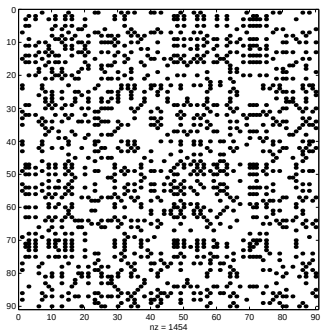


Wartość i wektor Fiedlera (λ_2, y_2)

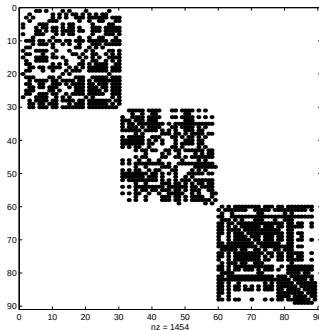
$$0 \leq \lambda_2 \leq m;$$

Fiedler (1973,1975) udowodnił, że

- $\lambda_2 > 0$ w.t.w. $G = (V, E)$ jest grafem spójnym.
- Jeżeli $G' = (V, E')$, $E' \subset E$, to $\lambda_2(G') \leq \lambda_2(G)$. Niskie wartości λ_2 przemawiają za **istnieniem** wewnętrznej struktury w zbiorze X . Np. gdy graf **planarny**, to $\lambda_2 \leq \frac{8d_{max}}{m}$ (Spielman & Teng, 2007).
- Wektory własne y_i odpowiadające wartościom własnym λ_i są ortogonalne. Zatem $y_1^T y_2 = 0 \Rightarrow \sum_j y_{j2} = 0$.
- Jeżeli G jest spójnym i nieskierowanym grafem oraz $A = \{v_i: y_{i2} \geq 0\}$, $B = V \setminus A = \{v_i: y_{i2} < 0\}$, to podgraf $G_B = (B, E_B)$ jest spójny.



(a)



(b)

Rysunek: macierz L : (a) oryginalna, (b) po posortowaniu względem wartości y_2

Prosty algorytm grupowania (MATLAB)

```
X = load('data.dat');
utwórz macierz wag W;
przekształć W w macierz powiązań S;
L = diag(sum(S)) - S;
[Y, Lambda]=eigs(L,'SM',2);
binaryzuj wektor y = Y(:,2)
```

Ciekawy wariant binaryzacji przedstawili Tolliver, & Miller (2006).
Ewentualnie: powtarzaj algorytm (wybierając podgraf o minimalnej wartości λ_2) do chwili uzyskania podziału na k grup.

Prosty algorytm grupowania (MATLAB)

```
X = load('data.dat');
utwórz macierz wag W;
przekształć W w macierz powiązań S;
L = diag(sum(S)) - S;
[Y, Lambda]=eigs(L,'SM',2);
binaryzuj wektor y = Y(:,2)
```

Ciekawy wariant binaryzacji przedstawili Tolliver, & Miller (2006).
 Ewentualnie: powtarzaj algorytm (wybierając podgraf o minimalnej wartości λ_2) do chwili uzyskania podziału na k grup.

Prosty algorytm grupowania (MATLAB)

```
X = load('data.dat');
utwórz macierz wag W;
przekształć W w macierz powiązań S;
L = diag(sum(S)) - S;
[Y, Lambda]=eigs(L,'SM',2);
binaryzuj wektor y = Y(:,2)
```

Ciekawy wariant binaryzacji przedstawili Tolliver, & Miller (2006).
Ewentualnie: powtarzaj algorytm (wybierając podgraf o minimalnej wartości λ_2) do chwili uzyskania podziału na k grup.

Inne kryteria rozcinania

$$Rcut(C_1, \dots, C_k) = \sum_{j=1}^k \frac{cut(C_j, \overline{C_j})}{n_j}, \quad (\text{Hagen \& Kahng, 1992})$$

$$Ncut(C_1, \dots, C_k) = \sum_{j=1}^k \frac{cut(C_j, \overline{C_j})}{\text{vol } C_j}, \quad (\text{Shi \& Malik, 2000})$$

$$Mcut(C_1, \dots, C_k) = \sum_{j=1}^k \frac{cut(C_j, \overline{C_j})}{\text{assoc}(C_j, C_j)}, \quad (\text{Ding et al., 2001})$$

gdzie $n_j = |C_j|$, $\text{vol } C_j = \sum_{v_i \in C_j} d_i$.

Relaksacja $Rcut$ (1)

Niech $h_i = (h_{1i}, \dots, h_{ji}, \dots, h_{m_i})^T$ wektor o składowych

$$h_{ji} = \begin{cases} 1/\sqrt{n_i} & \text{gdy } j \in C_i \\ 0 & \text{w p.p.} \end{cases}, \quad H = (h_1, \dots, h_k)$$

$\|h_i\| = h_i^T h_i = 1$ oraz $h_i^T h_j = 0$ gdy $i \neq j$, czyli macierz H spełnia warunek $H^T H = \mathbb{I}$. Ponadto,

$$h_i^T L h_i = \frac{cut(C_i, \overline{C_i})}{n_i} = (H^T L H)_{ii}$$

Zatem

$$Rcut(C_1, \dots, C_k) = \sum_{i=1}^k h_i^T L h_i = \sum_{i=1}^k (H^T L H)_{ii} = \text{tr}(H^T L H)$$

Relaksacja R_{cut} (2)

$$\min_{(C_1, \dots, C_k) \in \mathcal{C}^k} \text{tr}(H^T L H), \text{ gdzie } H \text{ jak wyżej}$$

Relaksacja:

$$\min_{\substack{H \in \mathbb{R}^{n \times k} \\ H^T H = \mathbb{I}}} \text{tr}(H^T L H)$$

Na mocy tw. Fan'a (1947), por. (Mirzal & Furukawa, 2010)

$$\min_{\substack{H \in \mathbb{R}^{n \times k} \\ H^T H = \mathbb{I}}} \text{tr}(H^T L H) = \sum_{j=1}^k \lambda_j \quad \text{jeżeli } H = (y_1, \dots, y_k) R$$

gdzie y_i - wektor własny odpowiadający i -tej minimalnej wartości własnej λ_i laplasjanu L , oraz $R \in \mathbb{C}^{k \times k}$ – macierz unitarna.

Relaksacja $Ncut$ (1)

Niech $h_i = (h_{1i}, \dots, h_{ji}, \dots, h_{mi})^T$ wektor o składowych

$$h_{ji} = \begin{cases} 1/\sqrt{\text{vol } C_i} & \text{gdy } j \in C_i \\ 0 & \text{w p.p.} \end{cases}, \quad H = (h_1, \dots, h_k)$$

$h_j^T h_j = 1/d_j$, oraz $h_i^T h_j = 0$ gdy $i \neq j$, czyli macierz H spełnia warunek $H^T D H = \mathbb{I}$. Ponadto,

$$h_i^T L h_i = \frac{\text{cut}(C_i, \overline{C_i})}{\text{vol } C_i} = (H^T L H)_{ii}$$

Zatem

$$Ncut(C_1, \dots, C_k) = \sum_{i=1}^k h_i^T L h_i = \sum_{i=1}^k (H^T L H)_{ii} = \text{tr}(H^T L H)$$

Relaksacja $Ncut$ (2)

Relaksacja zadania NCut przyjmuje postać

$$\min_{\substack{H \in \mathbb{R}^{n \times k} \\ H^T D H = I}} \text{tr}(H^T L H)$$

Niech $Q = D^{1/2} H$. Wówczas

$$\min_{\substack{Q \in \mathbb{R}^{n \times k} \\ Q^T Q = I}} \text{tr}[Q^T (D^{-1/2} L D^{-1/2}) Q]$$

i można skorzystać z tw. Fan'a (1949), tzn. rozwiązaniem jest $Q = (y_1, \dots, y_k) R$, gdzie y_i to i -ty wektor własny macierzy $\mathcal{L} = D^{-1/2} L D^{-1/2}$.

Symetryczny normalizowany laplasjan $\mathcal{L} = D^{-1/2}LD^{-1/2}$ (1)

- dla dowolnego wektora $x \in \mathbb{R}^m$

$$x^T \mathcal{L} x = \sum_{i=1}^m \sum_{j=1}^m s_{ij} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2$$

- $\mathcal{L} = D^{-1/2}LD^{-1/2}$ jest macierzą symetryczną i dodatnio-półokreśloną o elementach

$$\mathcal{L}(i, j) = \begin{cases} (d_i - s_{ii})/d_i & \text{gdy } i = j \\ -s_{ij}/(\sqrt{d_i d_j}) & \text{gdy } i \neq j \text{ oraz } \{v_i, v_j\} \in E \\ 0 & \text{w p.p.} \end{cases}$$

- jeżeli $y_1 = D^{1/2}\mathbf{e}$, to

$$\mathcal{L} y_1 = D^{-1/2} L D^{-1/2} (D^{1/2} \mathbf{e}) = 0 \Rightarrow (0, v_1) \text{ para własna } \mathcal{L}.$$

Symetryczny normalizowany laplasjan $\mathcal{L} = D^{-1/2}LD^{-1/2}$ (2)

$$\mathcal{L} = D^{-1/2}LD^{-1/2} = \mathbb{I} - D^{-1/2}SD^{-1/2} = \mathbb{I} - \bar{\mathcal{L}}$$

Jeżeli (λ_i, y_i) para własna $\mathcal{L}$, to $(1 - \lambda_i, y_i)$ – para własna macierzy $\bar{\mathcal{L}}$. Wystarczy więc wyznaczyć k wektorów własnych odpowiadających **maksymalnym** wartościom własnym macierzy $\bar{\mathcal{L}}$:
 $1 = \mu_1 \geq \dots \geq \mu_m.$

Algorytm NJW (Ng, Jordan, & Weiss, 2002)

WE: symetryczna macierz S taka, że $s_{ii} = 0$

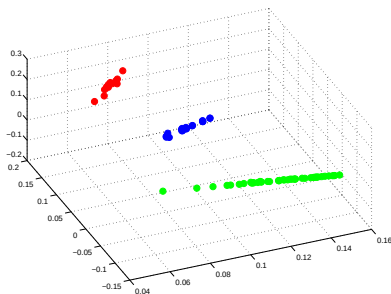
- 1 wyznaczyć $\bar{\mathcal{L}} = D^{-1/2} S D^{-1/2}$
- 2 wyznaczyć $Y = (y_1, \dots, y_k)$; y_j - j -ty dominujący wektor
- 3 rzutuj wiersze macierzy Y na sferę jednostkową, tzn. utwórz macierz $Z = [z_{ij}]$, $i = 1, \dots, m$, $j = 1, \dots, k$ o elementach

$$z_{ij} = y_{i,j} / \left(\sum_{\ell=1}^k y_{i\ell}^2 \right)^{-1/2}$$

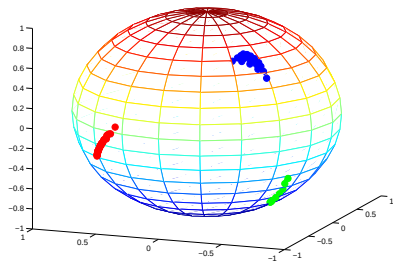
- 4 wyznaczyć finalny podział danych reprezentowanych przez macierz Z .



Algorytm NJW - ilustracja (1)

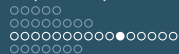


(a)

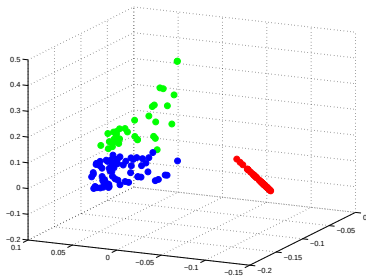


(b)

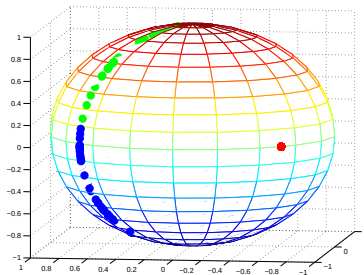
Rysunek: Algorytm NJW zastosowany do zbioru złożonego z trzech rozłącznych grup: (a) dane we współrzędnych spektralnych (wiersze



Algorytm NJW - ilustracja (2)



(a)



(b)

Rysunek: 4-wymiarowe dane iris: (a) we współrzędnych spektralnych, (b) po rzutowaniu na sferę jednostkową

Rotacja spektralna, (Huang, Nie, Huang, 2013)

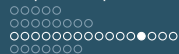
Obracamy wiersze macierzy Y tak, aby stały się najbliższe (w sensie normy Frobeniusa) do charakterystycznych wektorów postaci

$$h_{ji} = \begin{cases} 1 & \text{gdzie } v_j \in C_i \\ 0 & \text{w p.p.} \end{cases}, \quad i = 1, \dots, k, j = 1, \dots, n$$

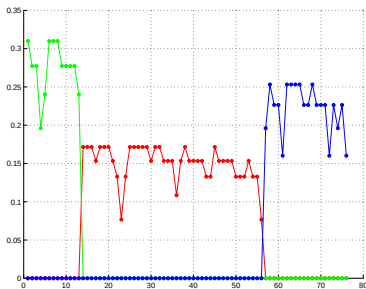
tnz. szukamy unitarnej macierzy R i macierzy $H = (h_1, \dots, h_k)$ (jej kolumnami są wektory charakterystyczne), dla których indeks

$$J(H, R) = \|YR - H\|_F^2$$

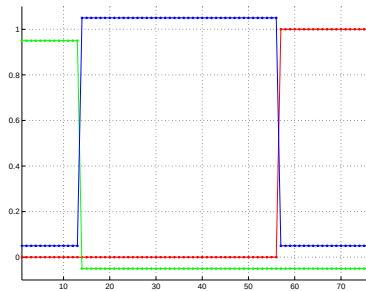
osiąga minimum.



Rotacja spektralna w działaniu



(a)



(b)

Rysunek: Zbiór złożony z trzech separowalnych grup (a) dominujące wektory własne macierzy $\bar{\mathcal{L}}$, (b) kolumny macierzy H .

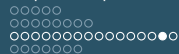
Liczba spójnych komponentów a spektrum macierzy $\mathcal{L}$

$$\mathcal{L} = \begin{pmatrix} \mathcal{L}^1 & & & \\ & \mathcal{L}^2 & & \\ & & \ddots & \\ & & & \mathcal{L}^k \end{pmatrix}, \quad \sigma(\mathcal{L}) = \bigcup_{i=1}^k \sigma(\mathcal{L}^i)$$

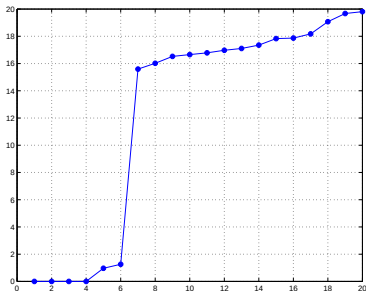
gdzie $\sigma(\mathcal{L}^i)$ – spektrum $\mathcal{L}^i$; ponadto: $\lambda_1^i = 0, \lambda_2^i > 0$. Zatem

$$\lambda_1(\mathcal{L}) = \dots = \lambda_k(\mathcal{L}) = 0, \lambda_{k+1}(\mathcal{L}) = \min_{1 \leq j \leq k} \lambda_2(\mathcal{L}^j) > 0$$

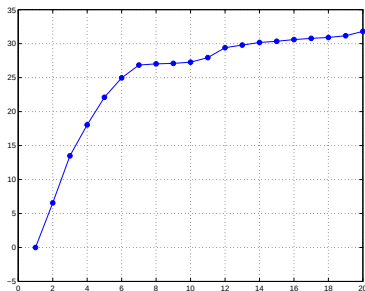
$$\lambda_{k+1}(\mathcal{L}) - \lambda_k(\mathcal{L}) \leftarrow \text{rozstęp spektralny}$$



Liczba spójnych komponentów a spektrum macierzy $\mathcal{L}$



(a)



(b)

Rysunek: Przerwa spektralna w zbiorze: (a) złożonym z 6 rozłącznych grup, (b) dwóch nachodzących na siebie półkieszyców

Normalizować czy nie?

von Luxburg, Belkin, & Bousquet (2008):

(...) our analysis shows that while one type of spectral clustering (“normalized”) is consistent under very general conditions, another popular version of spectral clustering (“unnormalized”) is only consistent under some very specific conditions which do not have to be satisfied in practice.

(...) from a statistical point of view, it is clear that normalized rather than unnormalized spectral clustering should be used whenever possible.

Aproksymacja Nyström'a (1)

Złożoność pamięciowa algorytmu grupowania spektralnego:
 $O(m^2)$, czasowa – $O(m^3)$.

Macierz $\mathbb{L}$ można zastąpić jej aproksymacją $\mathbb{L} \approx ZZ^T$, gdzie
 $Z \in \mathbb{R}^{m \times \ell}$, $\ell \ll m$. (Williams, & Seeger, 2001).

Wówczas macierz Y , której kolumnami jest ℓ dominujących
wektorów własnych $\mathbb{L}$ aproksymuje się $\tilde{Y} = ZU\Lambda^{-1/2}$ przy czym
 $[U, \Lambda] = \text{eig}(Z^T Z)$, (Fowlkes, *et al*, 2004)

Aproksymacja Nyström'a (2)

Niech (Fowlkes, *et al*, 2004)

$$S = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}, \quad A \in \mathbb{R}^{\ell \times \ell}, B \in \mathbb{R}^{\ell \times m-\ell}, C \in \mathbb{R}^{m-\ell \times m-\ell},$$

gdzie $\ell \ll m$, i niech $S = Y\Lambda Y^T$, $A = U\Sigma U^T$. Wtedy

$$\tilde{Y} = \begin{bmatrix} U \\ B^T U \Sigma^{-1} \end{bmatrix}_{m \times \ell} \approx Y(:, 1 : \ell)$$

przy czym $\tilde{S} = \tilde{Y}\Sigma\tilde{Y}^T$, oraz $\|S - \tilde{S}\|_F = \|C - B^T U \Sigma^{-1}\|_F$.

Ale: $\tilde{Y}^T \tilde{Y} \neq \mathbb{I}$.

Ortogonalizacja $\tilde{Y}$ (1)

A jest dodatnio półokreślona

- 1 Niech F będzie **pewną** macierzą ortogonalną, a Ω – macierzą diagonalną. Definiujemy macierz V w następujący sposób

$$V = \underbrace{\begin{bmatrix} A \\ B^T \end{bmatrix}}_Z A^{-1/2} F \Omega^{-1/2}$$

Sprawdzamy, że $V \Omega V^T = Z A^{-1} Z^T = \tilde{S}$

- 2 Z warunku $V^T V = \mathbb{I}$ wyznaczamy macierz F : jej kolumnami są wektory własne macierzy $A + A^{-1/2} B B^T A^{-1/2}$.

Ortogonalizacja $\tilde{Y}$ (2)

- 1 przedstaw macierz A w postaci $A = U\Lambda U^T$ i wyznacz rozszerzenie $\tilde{Y}$;
- 2 wyznacz macierz $Z = \tilde{Y}\Lambda^{1/2}$;
- 3 wyznacz rozkład macierzy $Z^T Z$ postaci $Z^T Z = F\Sigma F^T$;
- 4 wyznacz macierz $\tilde{V} = ZF\Sigma^{-1/2}$.

Dalsze przekształcenia

$$D^{-1/2}LD^{-1/2}y = \lambda y \quad \text{Problem własny (PW)}$$

$$\textcolor{red}{D}^{1/2}D^{-1/2}LD^{-1/2}y = \lambda \textcolor{red}{D}^{1/2}y$$

$$L(D^{-1/2}y) = \lambda D^{1/2}\textcolor{red}{D}^{1/2}D^{-1/2}y$$

$$Lz = \lambda Dz \quad \text{uogólniony PW}$$

$$z = D^{-1/2}y$$

Asymetryczna normalizacja laplasjanu

$$Lz = \lambda Dz \Rightarrow D^{-1}Lz = \lambda z \equiv \underbrace{(\mathbb{I} - D^{-1}S)}_{L_{rw}} z = \lambda z$$

$P = D^{-1}S$ – charakteryzuje błędzenie losowe po grafie, tzn. odwracalny łańcuch Markowa.

Interpretacje probabilistyczne

Jeżeli G jest spójnym i nie-dwudzielnym grafem, to dla P istnieje rozkład stacjonarny $\pi = (\pi_1, \dots, \pi_m)^T$, gdzie $\pi_i = d_i / (\text{vol } V)$. Niech $A \subset V$ i niech $P_A(i) = d_i / (\text{vol } A)$, gdy $v_i \in A$. Wtedy (Meila, & Shi, 2001):

$$P_{A\bar{A}} = \text{cut}(A, \bar{A}) / (\text{vol } A)$$

Stąd

$$N\text{cut}(A, \bar{A}) = P_{A\bar{A}} + P_{\bar{A},A}$$

tnz. jeżeli $\{A, \bar{A}\}$ jest podziałem o niskim $N\text{cut}$ to $P_{A\bar{A}}$ oraz $P_{\bar{A},A}$ są niskie. Czyli: jeżeli wędrowiec rozpoczyna błądzenie w zbiorze A (lub $\bar{A}$) to p-stwo opuszczenia tego zbioru jest niskie.

Nierówności izoperymetryczne (1)

Koszt cięcia Cheegera

$$RCcut(A, \bar{A}) = \frac{cut(A, \bar{A})}{\min(|A|, |\bar{A}|)}$$

$$NCcut(A, \bar{A}) = \frac{cut(A, \bar{A})}{\min(\text{vol } A, \text{vol } \bar{A})} \leftarrow \text{konduktancja}$$

$$P_{A\bar{A}} = NCcut(A, \bar{A}), \text{ (Spielman)}$$

Ponadto (Bühler, & Hein, 2009):

$$RCcut(A, \bar{A}) \leq Rcut(A, \bar{A}) \leq 2RCcut(A, \bar{A})$$

$$NCcut(A, \bar{A}) \leq Ncut(A, \bar{A}) \leq 2NCcut(A, \bar{A})$$

Nierówności izoperymetryczne (2)

Niech $h_R = \min_{A \subset A} RCut(A, \bar{A})$, $h_N = \min_{A \subset A} NCut(A, \bar{A})$.

Wtedy:

$$\frac{h_R^2}{2 \max_i d_i} \leq \lambda_2(L) \leq 2h_R, \text{ nierówność Cheegera}$$

$$\frac{1}{2} h_N^2 \leq \lambda_2(\mathcal{L}) \leq 2h_N$$

p -laplasjan

$$x^T L^{(p)} x = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m s_{ij} (x_i - x_j)^p, \quad p > 1, x \in \mathbb{R}^m$$

Ponadto

$$(L^{(p)} x)_i = \sum_{j=1}^m s_{ij} \phi_p(x_i - x_j)$$

gdzie $\phi_p(v) = |v|^{p-1} \text{sign}(v)$. Gdy $p = 2$, to $\phi_2(v) = v$, a $L^{(2)}$ to „zwykły” laplasjan L .

Normalizowany p -laplasjan:

$$(L^{(p)} x)_i = \frac{1}{d_i} \sum_{j=1}^m s_{ij} \phi_p(x_i - x_j)$$

Wektory i wartości własne

Liczba $\lambda^{(p)}$ jest wartością własną p -laplasjanu $L^{(p)}$ jeżeli

$$(\exists y \in \mathbb{R}^m): (L^{(p)}y)_i = \lambda^{(p)}\phi_p(y_i), \quad i = 1, \dots, m$$

y to p -wektor własny odpowiadający $\lambda^{(p)}$.

Jeżeli $y_1^{(p)}$ jest stałym wektorem, to $(L^{(p)}y_1^{(p)})_i = 0$ for $i = 1, \dots, m$, skąd wynika, że $\lambda_1^{(p)} = 0$. Krotność $\lambda_1^{(p)} = 0$ jest równa liczbie spójnych składowych.

p -wartość Fiedlera

Niech

$$F_2^{(p)}(v) = \frac{v^T L^{(p)} v}{\min_{c \in \mathbb{R}} \|v - c\mathbf{e}\|_p^p}, \quad v \in \mathbb{R}^m$$

Wtedy (Bühler, & Hein, 2009):

$$\lambda_2^p = \min_{v \in \mathbb{R}^m} F_2^{(p)}(v)$$

$\lambda_2^p > 0$ jeżeli G jest grafem spójnym. Ponadto

$$(a) \quad \lim_{p \rightarrow 1+} \lambda_2^{(p)} = h_R$$

$$(b) \quad h_R \leq RCcut(C_p, \overline{C_p}) \leq p \left(\max_{i \in V} d_i \right)^{\frac{p-1}{p}} h_R^{\frac{1}{p}}$$

gdzie C_p zbiór otrzymany przez optymalne progowanie wektora Fiedlera odpowiadającego $\lambda_2^{(p)}$.

1-Laplasjan (Szlam, & Bresson 2010)

Dla dowolnego wektora charakterystycznego $x = \chi_A$ mamy:

$$1 \quad \|x - m(x)\|_1 = \min(A, \bar{A})$$

$$2 \quad x^T L^{(1)} x = \sum_{i,j} s_{ij} |x_i - x_j| = \sum_{i \in A, j \in \bar{A}} s_{ij} = 2 \text{cut}(A, \bar{A})$$

$$\min_{x \in \mathbb{R}^m} \frac{x^T L^{(1)} x}{\|x - m(x)\|_1} \leq h_R$$

Uogólnienie na przypadek $k > 2$ podają (Bresson, *et al.* 2013).

Modularność, (Newman, 2006)

$$B = A - \frac{\mathbf{d}\mathbf{d}^T}{\text{vol } V} = A - P$$

gdzie A – binarna macierz powiązań, $\mathbf{d} = (d_1, \dots, d_m)^T$,
 $p_{ij} = d_i d_j / (\text{vol } V)$ – p-stwo tego, że losowo wybrane węzły v_i, v_j są sąsiadami w referencyjnym grafie losowym.

Dziękuję za uwagę



[Belkin & Niyogi, 2003]

M. Belkin & P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation *Neural Computation*, 6(15):1373-1396, Jun. 2003.



[Bresson, et al. 2013]

X. Bresson, et al. Multiclass total variation clustering. *Advances in Neural Information Processing Systems, NIPS'2013*, pp. 1421-1429, 2013



[Bühler, & Hein, 2009]] T. Bühler, & M. Hein. Spectral clustering based on the graph p -Laplacian. In: *Proc. of the 26th Annual Intl Conf. on Machine Learning, ICML '09*, pp. 81–88, 2009. ACM



[Ding, et al., 2001]

C. Ding *et al.* A Min-max cut algorithm for graph partitioning and data clustering. In: *Proc. IEEE Int. Conf. on Data Mining, ICDM 2001*, pp. 2001, pp. 107-114



[Fan, 1949]

K. Fan. On a theorem of Weyl concerning eigenvalues of linear transformations. I. *Proc. Natl Acad. Sci. USA*, 35(11), 652-655, 1949



[Fiedler, 1973]

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Math. J.*, 23(98):298–305, 1973



[Fiedler, 1975]

M. Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Math. J.*, 25:619–672, 1975



C. Fowlkes. Spectral grouping using the Nyström method *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(2), 214-225, 2004



L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. on Computer-Aided Design*, 11(9), 1074-1085, 1992



J. Huang, F. Nie, & H. Huang. Spectral Rotation versus k -Means in Spectral Clustering. In: *Proc. 27-th AAAI Conf. on Artificial Intelligence*, pp. 431-437, 2013



[Meila, & Shi, 2001]

M. Meila, & J. Shi. A random walks view on spectral segmentation. In: *8th Int. Workshop on AI and Statistics, AISTATS 2001*. Hyatt Hotel, Key West, Florida, USA, 4-7 Jan. 2001.



[Mirzal & Furukawa, 2010]

A. Mirzal & M. Furukawa. Eigenvectors for clustering: Unipartite, bipartite, and directed graph cases. *arXiv:1005.2603 [cs.LG]*, 2010



[Newman, 2006]

M.E.J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582, 6 Jun. 2006.



[Ng, Jordan, & Weiss, 2002]

A. Ng, M. Jordan, & Y. Weiss. On spectral clustering: analysis and an algorithm. In: T. Dietterich, S. Becker, & Z. Ghahramani (eds.). *Advances in Neural Information Processing Systems 15*, pp. 849–856, MIT Press 2002



[Shi & Malik, 2000]

J. Shi & J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8), 888-905, 2000



[Szlam, & Bresson, 2010]

A.D. Szlam, & X. Bresson. Total variation and Cheeger cuts. *Proc. of the 27th Intl Conf. on Machine Learning (ICML-10)*, pp. 1039-1046, 2010.



[Tolliver, & Miller, 2006]

D.A. Tolliver, & G.L. Miller. Graph partitioning by spectral rounding: Applications in image segmentation and clustering. In: A. Fitzgibbon, C.J. Taylor, & Y. LeCun (eds.) *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006, pp. 1053-1060



[von Luxburg, Belkin, & Bousquet, 2008]

U. von Luxburg, M. Belkin, & O. Bousquet. Consistency of spectral clustering. *Ann. Statist.* 36(2), 555-586, 2008



[Williams, & Seeger, 2001]

C. Williams, & M. Seeger. Using the Nyström method to speed up kernel machines. In: T.K. Leen, T.G. Dietterich, and V. Tresp, eds., *Proc. 2000 Conf. Advances in Neural Information Processing Systems 13*, pp. 682-688, 2001



[Zelnik-Menor & Perona, 2000]

L. Zelnik-Menor & P. Perona. Self-tuning spectral clustering.
In: *Advances in Neural Information Processing Systems* 17, pp.
1601-1608. MIT Press, 2000