CLICK – CLuster Identification via Connectivity Kernels

Wesley Nunes Gonçalves

Roteiro

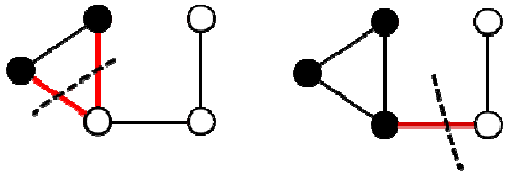
- Introdução
- Notações
- Algoritmo CLICK
- Complexidade Computacional
- Conclusões

Introdução

- Algoritmo baseado em grafos.
  - Cada objeto da base é representando por um vértice.
- Não requer o número de grupos ou sua estrutura.
- Algoritmo consiste em um processo principal que recursivamente particiona o grafo.
- Originalmente proposto para agrupamento de expressão génica.

Notações

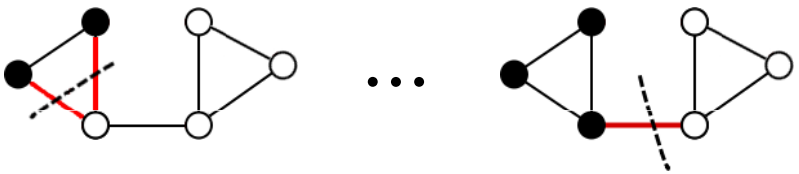
- $G = (V = \{v\}, E = \{w_{ij}\})$ – Grafo
  - V e E são os conjuntos de vértices e arestas, respectivamente.



- Corte no Grafo
  - $E' = cut(G) \mid E' \subset E$
  - $A, B \subset V \mid A \cup B = V,\ A \cap B = 0$
  - Peso do corte: $W(E') = \sum_{w_{ij} \in E'} w_{ij}$

Imagens adaptada do wikipedia

Notações

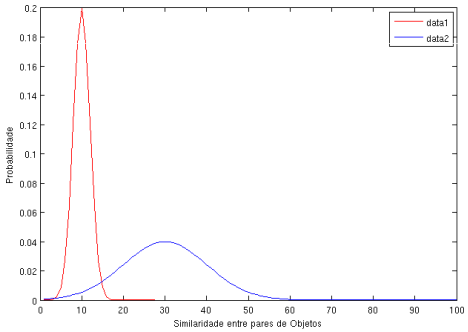


- Corte mínimo do Grafo $= \min_{E'} W(E')$

Imagens adaptada do wikipedia

CLICK

- Pressuposto: Similaridades S_{ij} entre pares de objetos podem ser modelados por gaussiana.



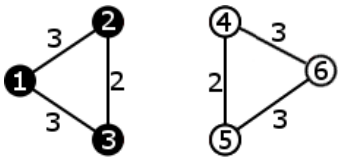
$$f(S_{ij}|\,\mu_M,\sigma_M) = \frac{1}{\sqrt{2\pi}\sigma_M}\, e^{\frac{(S_{ij}-\mu_M)}{2\sigma_M^2}}$$

$$f(S_{ij}|\,\mu_N,\sigma_N) = \frac{1}{\sqrt{2\pi}\sigma_N}\, e^{\frac{(S_{ij}-\mu_N)}{2\sigma_N^2}}$$

$$\mu_M > \mu_N$$

CLICK

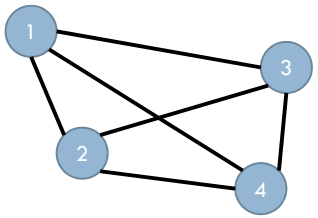
- Pressuposto é justificado empiricamente usando simulações.
- Os parâmetros das gaussianas podem ser estimados de duas formas:
  - Soluções parcialmente conhecidas.
  - Expectation Maximization.



$$\mu_M = (3+3+2+2+3+3)/6$$
$$\mu_N = (S_{14}+S_{15}+S_{16}+S_{24}+\ \ldots + S_{36})/9$$
$$p_{mates} = 6/15$$

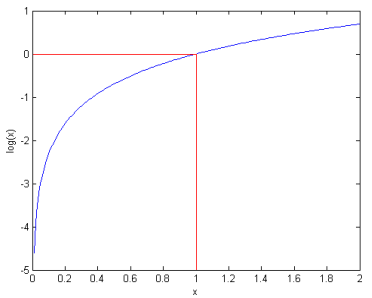
CLICK

- Representa os objetos por um grafo G = (V,E)



$$w_{ij} = \log\frac{p_{mates}\, f(S_{ij}\,|\,\mu_M,\sigma_M)}{(1-p_{mates})\, f(S_{ij}|\,\mu_N,\sigma_N)}$$

$$f(S_{ij}\,|\,\mu_M,\sigma_M) > f(S_{ij}\,|\,\mu_N,\sigma_N)\ \text{então}\ w_{ij} > 0$$
$$f(S_{ij}\,|\,\mu_M,\sigma_M) < f(S_{ij}\,|\,\mu_N,\sigma_N)\ \text{então}\ w_{ij} < 0$$



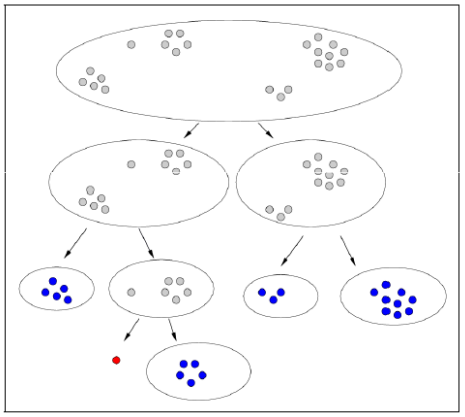
CLICK

```
Basic-CLICK(G(V, E))
    if (V(G) = {v}) then
        move v to the singleton set R
    elseif (G is a kernel) then
        Output V(G)
    else
        (H, H̄, cut) ← MinWeightCut(G)
        Basic-CLICK(H)
        Basic-CLICK(H̄)
    end if
end
```

Figure 4.1: The Basic-CLICK algorithm

Imagens retiradas de [1]

CLICK



Imagens retiradas de Sharan R, Shamir R. (2000)

CLICK

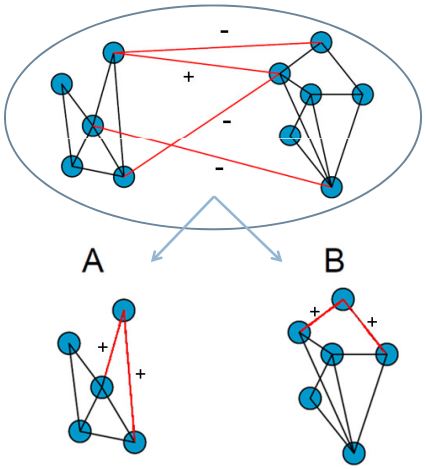
- Testar para **todos** cortes E' em G duas hipóteses:
  - $H_0^{E'}$: E' contém somente arestas entre 'non-mates'.
  - $H_1^{E'}$: E' contém somente arestas entre 'mates'.
- G é um kernel se e somente se $H_1^{E'}$ é valida para todos cortes E'.
- G **não** é um kernel se existe pelo menos um corte E' para qual a hipótese $H_0^{E'}$ é valida.

CLICK

- H_1 é aceita para um corte se $\mathrm{W}(E') > 0 \,|\mathrm{W}(E') = \sum_{w_{ij} \in E'} w_{ij}$.
- Mas temos que testar se $\mathrm{W}(E') > 0$ para todos os cortes E'.
- Para isso, vamos tomar C como o corte mínimo, ou seja, $\mathrm{W}(C) < \mathrm{W}(E')$ para todos cortes E'.
- Se $\mathrm{W}(C) > 0$, então $\mathrm{W}(E') > 0$ para todos cortes E'.

CLICK

- Verificando se G é um kernel



CLICK

- Algoritmo completo:

$R \leftarrow N$.
While some change occurs **do:**
  Execute Basic-CLICK(G_R).
  Let $\mathcal{L}$ be the list of kernels produced.
  Let R be the set of singletons produced.
  Adoption($\mathcal{L}, R$).
Merge($\mathcal{L}$).
Adoption($\mathcal{L}, R$).

Figure 2: The CLICK algorithm.

Basic-CLICK($G(V, E)$)
  if ($V(G) = \{v\}$) **then**
    move v to the singleton set R
  elseif (G is a kernel) **then**
    Output $V(G)$
  else
    (H,$\bar{H}$, cut) ← MinWeightCut(G)
    Basic-CLICK(H)
    Basic-CLICK($\bar{H}$)
  end if
end

Figure 4.1: The Basic-CLICK algorithm

Imagens retiradas de Sharan R, Shamir R. (2000)

Refinamentos

- Limpar Grafo G
  - $t = \frac{1}{|G|}\sum_{v \in V} degree(v)$
  - Todos vértices com grau menor que t são removidos e inseridos no conjunto de $singletons$.
- Cacular somente corte mínimo para s-t, com s e t sendo os dois vértices mais distantes do grafo (diâmetro).

Desempenho

- Complexidade
  - Encontrar cortes mínimos = $O(n^2\sqrt{m})$, para n vértices e m arestas.
  - Usando estratégia s-t = $O(nm^{2/3})$.

#Elements	#Edges	Density	Time(min)
517	22,084	0.17	0.5
826	10,978	0.03	0.2
2,329	134,352	0.05	0.8
20,275	303,492	0.001	32.5
72,623	1,043,937	0.0004	53
117,835	4,614,038	0.0007	126.3

Table 7: A summary of the time performance of CLICK on the above mentioned datasets. The second column specifies the number of edges in the similarity graph for each instance. The third column specifies the fraction of edges with respect to the total number of element pairs.

Imagens retiradas de Sharan R, Shamir R. (2000)

Conclusões

- Algoritmo não requer o número de grupos.
- Exclusão de possíveis ruídos.

- Pesado computacionalmente (parte do algoritmo já possui ordem quadrática).
- Estimação incorreta das distribuições gaussinas pode influenciar diretamente nos resultados do agrupamento.

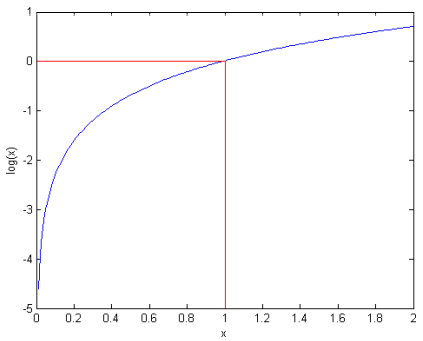
Referências

- Sharan R, Shamir R. (2000) CLICK: a clustering algorithm with applications to gene expression analysis. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB). AAAI Press, Vienna, pp. 307-316.
- Sharan R, Maron-Katz A, Shamir R. (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data. Bioinformatics 19(14):1787-1799.

- Obrigado

Material Complementar

- Provar que $W(E') > 0$
- Vamos assumir que H_1 é aceita se a $\Pr(H_1 | E') > \Pr(H_0 | E')$.
- $\log \frac{\Pr(H_1 | E')}{\Pr(H_0 | E')} > 0$



CLICK

- Vamos assumir que H_1 é aceita se a $\Pr(H_1 | E') > \Pr(H_0 | E')$.
- $\log \frac{\Pr(H_1 | E')}{\Pr(H_0 | E')} > 0 \rightarrow \log \frac{\Pr(H_1)\Pr(E' | H_1)}{\Pr(H_0)\Pr(E' | H_0)} > 0$

CLICK

- Vamos assumir que H_1 é aceita se a $\Pr(H_1 | E') > \Pr(H_0 | E')$.
- $\log \frac{\Pr(H_1 | E')}{\Pr(H_0 | E')} > 0 \longrightarrow \log \frac{\Pr(H_1)\Pr(E' | H_1)}{\Pr(H_0)\Pr(E' | H_0)} > 0$
- $$\log \frac{\prod_{i,j \in E'} p_{mates} \prod_{i,j \in E'} f(S_{ij} | \mu_M, \sigma_M)}{\prod_{i,j \in E'} (1 - p_{mates}) \prod_{i,j \in E'} f(S_{ij} | \mu_N, \sigma_N)} > 0$$

CLICK

- Vamos assumir que H_1 é aceita se a $\Pr(H_1 | E') > \Pr(H_0 | E')$.
- $\log \frac{\Pr(H_1 | E')}{\Pr(H_0 | E')} > 0 \longrightarrow \log \frac{\Pr(H_1)\Pr(E' | H_1)}{\Pr(H_0)\Pr(E' | H_0)} > 0$
- $$\log \frac{\prod_{i,j \in E'} p_{mates} \prod_{i,j \in E'} f(S_{ij} | \mu_M, \sigma_M)}{\prod_{i,j \in E'} (1 - p_{mates}) \prod_{i,j \in E'} f(S_{ij} | \mu_N, \sigma_N)} > 0$$
- $$\log \prod_{i,j \in E'} \frac{p_{mates} f(S_{ij} | \mu_M, \sigma_M)}{(1 - p_{mates}) f(S_{ij} | \mu_N, \sigma_N)} > 0$$

CLICK

- Vamos assumir que H_1 é aceita se a $\Pr(H_1 | E') > \Pr(H_0 | E')$.
- $\log \frac{\Pr(H_1 | E')}{\Pr(H_0 | E')} > 0 \longrightarrow \log \frac{\Pr(H_1)\Pr(E' | H_1)}{\Pr(H_0)\Pr(E' | H_0)} > 0$
- $$\log \frac{\prod_{i,j \in E'} p_{mates} \prod_{i,j \in E'} f(S_{ij} | \mu_M, \sigma_M)}{\prod_{i,j \in E'} (1 - p_{mates}) \prod_{i,j \in E'} f(S_{ij} | \mu_N, \sigma_N)} > 0$$
- $$\sum_{i,j \in E'} \log \frac{p_{mates} f(S_{ij} | \mu_M, \sigma_M)}{(1 - p_{mates}) f(S_{ij} | \mu_N, \sigma_N)} > 0$$

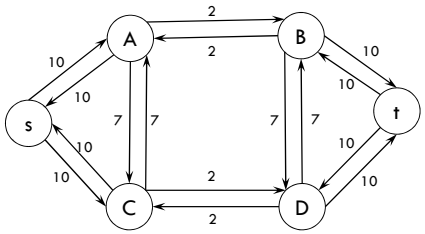
CLICK

- Vamos assumir que H_1 é aceita se a $\Pr(H_1 | E') > \Pr(H_0 | E')$.
- $\log \frac{\Pr(H_1 | E')}{\Pr(H_0 | E')} > 0 \longrightarrow \log \frac{\Pr(H_1)\Pr(E' | H_1)}{\Pr(H_0)\Pr(E' | H_0)} > 0$
- $$\log \frac{\prod_{i,j \in E'} p_{mates} \prod_{i,j \in E'} f(S_{ij} | \mu_M, \sigma_M)}{\prod_{i,j \in E'} (1 - p_{mates}) \prod_{i,j \in E'} f(S_{ij} | \mu_N, \sigma_N)} > 0$$
- $$\sum_{i,j \in E'} \log \frac{p_{mates} f(S_{ij} | \mu_M, \sigma_M)}{(1 - p_{mates}) f(S_{ij} | \mu_M, \sigma_M)} > 0$$
- $$\sum_{i,j \in E'} w_{ij} = \boldsymbol{W(E') > 0}$$

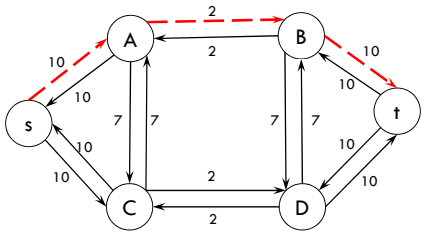
Material Complementar

Algoritmo para cálculo do corte mínimo s-t – Fluxo máximo

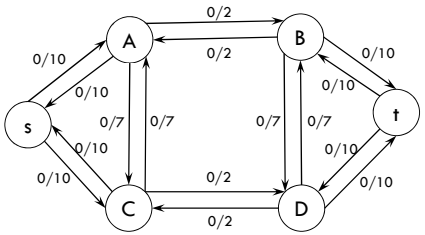
1. $f(i,j) = w_{ij}$
2. Enquanto existir caminho p entre s e t em G_f
    1. Encontre $c_f = \min\{f(i,j) | (i,j) \in p\}$
    2. Para cada aresta $(i,j) \in p$
        1. $f(i,j) = f(i,j) - c_f$
        2. $f(i,j) = f(i,j) + c_f$



Material Complementar

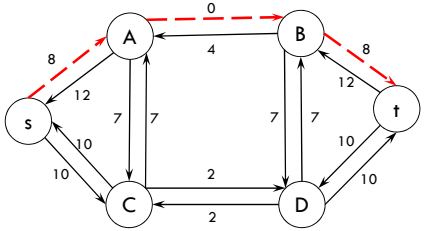


Rede Residual

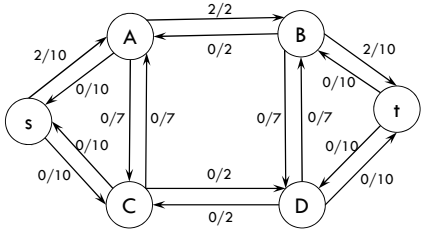


Rede Fluxo

Material Complementar

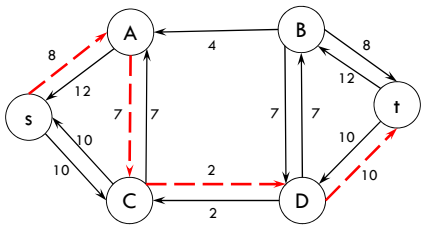


Rede Residual

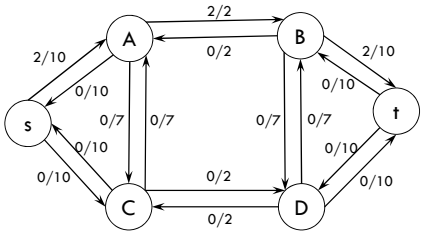


Rede Fluxo

Material Complementar

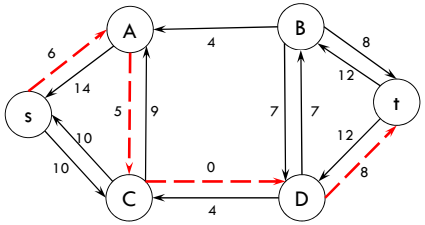


Rede Residual

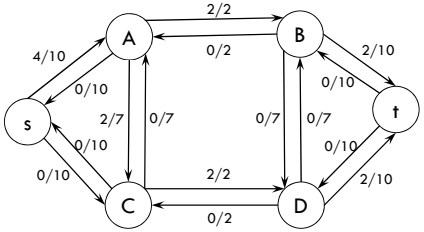


Rede Fluxo

Material Complementar

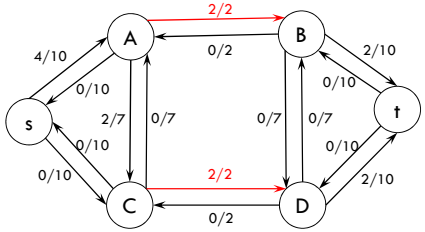


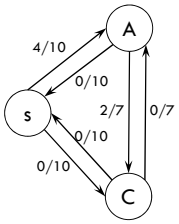
Rede Residual

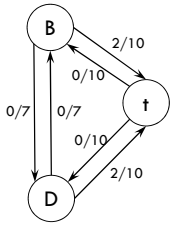


Rede Fluxo

Material Complementar







Material Complementar

- Corte mínimo para s-t.
  - Computar cortes para todos s, t ∈ $V \mid s \neq t$
  - Escolher corte com menor peso.