Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

DENCLUE – DENsity based CLUstEring

Ivani de O. N. Lopes

Seminário apresentado como parte das avaliações da disciplina Análise de Agrupamento de Dados, ministrada pelo prof. Dr. Ricardo Campello

São Carlos – SP
Novembro 2010

Resumo da apresentação

Introdução e contextualização

Fonte: Hinneburg and D. Keim (1998)

PROPOSTA

✓ Ser eficiente e efetivo em bases de dados multimídia, com forte presença de ruídos.

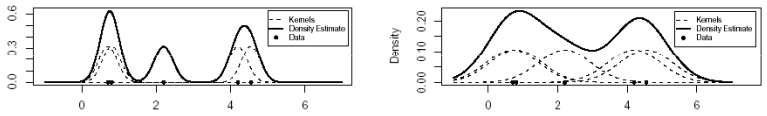
Pressuposições

✓ Clusters são regiões de alta densidade no espaço de dados,
✓ A influência de um objeto em sua vizinhança pode ser modelada matematicamente por uma função de influência,
✓ A densidade de um objeto é a soma das influências dele e de seus vizinhos.

Função de Influência e densidade Estimada por Kernels

Idéia

- A influência de um objeto é modelada por um kernel
- Densidade é a soma de todos os kernels



$$f(x,y) = \exp\left(\frac{-d(x,y)^2}{2\sigma^2}\right)$$

Função de influência: Kernel Gaussiano

$$f^D(x) = \sum_{i=1}^{N} \exp\left(\frac{-d(x,x_i)^2}{2\sigma^2}\right)$$

Estimativa de Densidade

Adaptado de Hinneburg e Hans-Henning: http://videolectures.net/ida07_hinneburg_dfc/

Ilustração do conceito de Kernel e densidade em duas dimensões

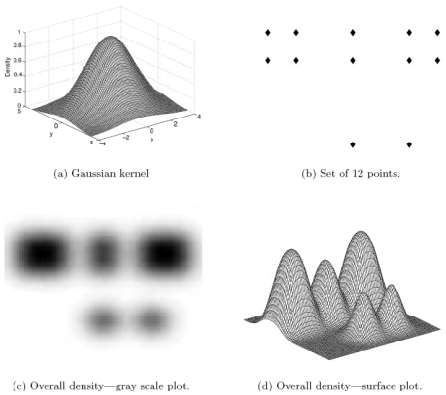


Figure 9.14. Example of the Gaussian influence (kernel) function and an overall density function.

DENCLUE – Ilustração unidimensional

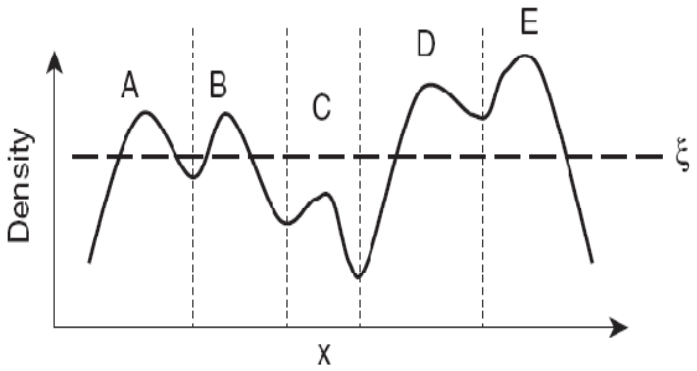


Figure 9.13. Illustration of DENCLUE density concepts in one dimension.

DENCLUE - Computacionalmente

1. Calcule a função de densidade para os objetos
2. Identifique os atratores de densidade
3. Encontre o atrator de densidade para cada objeto
4. Defina clusters que consistam de pontos associados a um atrator em particular
5. Descarte clusters cujo atrator possui densidade associada abaixo do limiar especificado.
6. Combine cluster que são conectados por um caminho de pontos, em que todos possuam densidade acima do limiar.

Computacionalmente cara - O(N^2).

DENCLUE – Implementação Eficiente

Pré-agrupamento

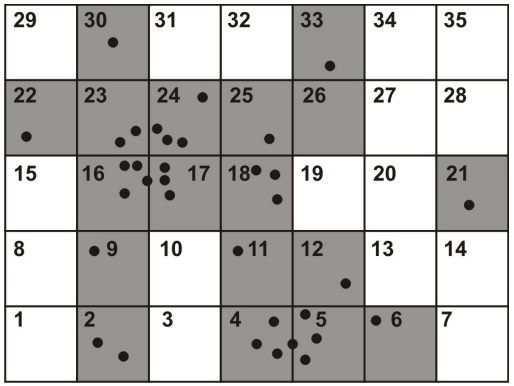


Figura. Mapeamento do espaço de dados em 2D

Resumo da Etapa 1 do algoritmo DENCLUE

Pré-agrupamento

✓Divida o hiper-retângulo delimitador mínimo em hiper-cubos d-dimensionais, com aresta de comprimento 2σ;

Exemplo: Considere o conjunto de objetos

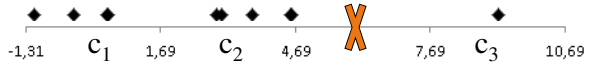
$$X = \{-1,31;\ -0,43;\ 0,34;\ 3,57;\ 2,76;\ 0,30;\ 9,06;\ 4,45;\ 2,87;\ 4,42\}$$

Dados slide 47 de Campello (2010): http://wiki.icmc.usp.br/images/2/21/Algoritmos_Particionais_II.pdf

Resumo da Etapa 1 do algoritmo DENCLUE

Pré-agrupamento

✓ Descarte os hiper-cubos vazios;

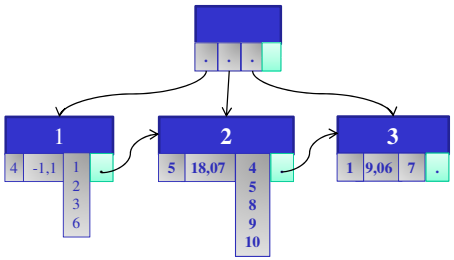


Descartar terceiro intervalo

Resumo da Etapa 1 do algoritmo DENCLUE

Pré-agrupamento

✓Armazene os dados dos hiper-cubos em uma árvore $B+$



Resumo da Etapa 1 do algoritmo DENCLUE

Pré-agrupamento

✓Encontre os hiper-cubos altamente densos, isto é:

$C_{sp}=\{c \in C_P \mid N_c \geq \xi_c\}, \xi_c = \xi / 2d$

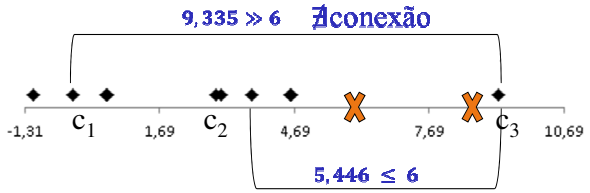
Resumo da Etapa 1 do algoritmo DENCLUE

Pré-agrupamento

✓Encontre os hiper-cubos, que possuem conexão com algum hiper-cubo denso.

$$C_{conexão} = \{c \in C_P | \exists c_s \in C_{SP} \; e \; \exists conexão \; (c_s, c)\}.$$

Existe conexão entre $c_1, c_2 \in CP$, se $d(centróide(c_1), centróide(c_2)) \leq 4\sigma$



Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento

✓Determine $C_r = C_{SP} \cup C_{conexão}$

$$C_r = \{c_1, c_2\} \cup \{c_3\} = \{c_1, c_2, c_3\}$$

Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento

✓Para $x \in c$ e $c, c_1 \in C_r$, tome

$$near(x) = \{x_1 \in c_1 \mid d(mean(c_1), x) \leq k\sigma \; e \; \exists conexão(c_1, c)\}$$

$C_r = \{c_1, c_2, c_3\}$; $k = 4$; $\sigma = 1{,}5$; $x = x_1$

$near(x_1)$ em $c_1 = c_1$

$near(x_1)$ em $c_2 = c_2$

- $d(3{,}614; -1{,}31) = 4{,}924 \leq 6$

$near(x_1)$ em $c_3 = \{\}$

- $d(9{,}06; -1{,}31) = 10{,}37 > 6$

Portanto, $near(x_1) = X - \{c_3\}$

Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento

✓Calcule:

- a função de densidade local $\hat{f}^D(x) = \sum_{x_1 \in near(x)} \exp(-\frac{d(x, x_1)^2}{2\sigma^2})$
- o gradiente $\nabla \hat{f}^D(x) = \sum_{x_1 \in near(x)} (x_1 - x) \exp(-\frac{d(x, x_1)^2}{2\sigma^2})$

$$\hat{f}^D(x_1) = \sum_{x \in near(x_1)} exp(-\frac{d(x, x_1)^2}{2\sigma^2})$$
$$= exp[-\frac{(-1{,}31-(-1{,}31))^2}{2.(1{,}5)^2}] + exp[-\frac{(-0{,}43-(-1{,}31))^2}{2.(1{,}5)^2}] + \cdots + exp[-\frac{(4{,}42-(-1{,}31))^2}{2.(1{,}5)^2}]$$
$$= 3{,}002$$

$$\nabla \hat{f}^D(x_1) = \sum_{x \in near(x_1)} (x - x_1) exp(-\frac{d(x_1, x)^2}{2\sigma^2})$$
$$= (-1{,}31 - (-1{,}31)) exp[-\frac{(-1{,}31-(-1{,}31))^2}{2.(1{,}5)^2}] + \cdots + (4{,}42 - (-1{,}31)) exp[-\frac{(4{,}42-(-1{,}31))^2}{2.(1{,}5)^2}]$$
$$= 2{,}768$$

Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento

✓ Calcule: a função de densidade local e o gradiente.
continuação...

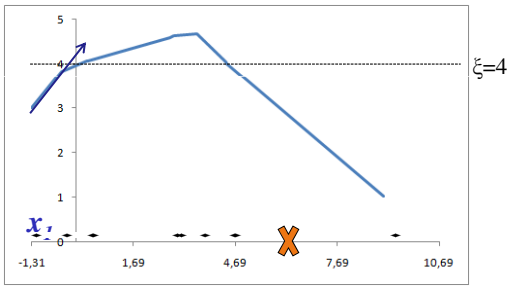


Figura. Curva de densidade do espaço X

Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento

✓ Encontre os atratores de densidade (máximos locais de $\hat{f}^D$), por uma busca hill-climbing, guiada pelo gradiente de $\hat{f}^D$ e una os objetos aos atratores mais próximos.

$$x = x^0, x^{l+1} = x^l + \delta \frac{\nabla \hat{f}^D(x^l)}{\|\nabla \hat{f}^D(x^l)\|}$$

Tomando δ=1 e ξ=4

Iteração 0: $x^0 = x_1$

Iteração 1: $x^{0+1} = x^0 + 1 * \frac{\nabla \hat{f}^D(x^0)}{\|\nabla \hat{f}^D(x^0)\|}$=-1,31+ 1 * $\frac{2,76}{2,76}$= - 0,31

Como, $\hat{f}^D(x^1)$=3,002 < $\hat{f}^D(x^0)$=3,906 .... Continua

Converge na 5ª iterãção, com x^*=3,69 e $\hat{f}^D(x^*)$=4,638 > 4, então define-se o atrator de densidade $x^* = x^0$=3,69.

Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento

✓ Encontre os atratores de densidade
continuação...

x^0	$\hat{f}^D(x^0)$	x^1	$\hat{f}^D(x^1)$	x^2	$\hat{f}^D(x^2)$	x^3	$\hat{f}^D(x^3)$	x^4	$\hat{f}^D(x^4)$	x^5	$\hat{f}^D(x^5$	x^6	$\hat{f}^D(x^6)$
-1,31	3,002	-0,31	3,906	0,69	4,088	1,69	4,162	2,69	4,568	**3,69**	4,638	4,69	3,69
-0,43	3,839	Unir objeto 2 ao atrator de densidade x* 3,69, pois d(-0,43; -0,31) 0,12 < 0,75											
0,34	4,082	Unir objeto 3 ao atrator de densidade x*= 3,69, pois d(0,34; 0,69)=0,35 < 0,75											
3,57	4,681	Unir objeto 4 ao atrator de densidade x* 3,69, pois d(3,57; 3,69) 0,12 < 0,75											
2,76	4,596	Unir objeto 5 ao atrator de densidade x*= 3,69, pois d(2,76; 3,69)=0,07 < 0,75											
0,30	4,079	Unir objeto 6 ao atrator de densidade x*= 3,69, pois d(0,30; 0,69)=0,39 < 0,75											
9,06	1,018	8,06*	0,920	x^0 não definiu um atrator de densidade, pois $\hat{f}^D(x^0)$ 1,018 < 4									
4,45	4,006	**3,45**	4,706	2,45	4,462								
2,87	4,636	Unir objeto 9 ao atrator de densidade x*= 3,69, pois d(2,87; 2,69)=0,18 < 0,75											
4,42	4,042	Unir objeto 10 ao atrator de densidade x*= 3,45, pois d(4,45; 4,42)=0,03 < 0,75											

Discussão dos parâmetros

DENCLUE depende de dois parâmetros σ e ξ.

- Escolha σ no intervalo que estabiliza m(σ), como na figura abaixo.
- Escolha ξ dentro do intervalo: $\|D_N\|.\sqrt{2\pi\sigma^2}^d \leq \xi \leq \min_{x^* \in X}\{f^{D_C}(x^*)\}$

D=$D_N \cup D_C$, D_N é conjunto de objetos ruído (noise) e D_C é o conjunto de objetos sem ruídos.

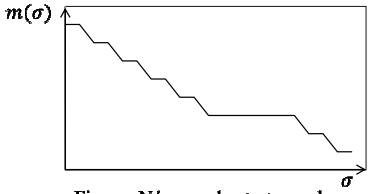


Figura. Número de atratores de densidade em função de σ.

Considerações Finais

- Forte fundamentação matemática. É possível que o DENCLUE estime densidades com maior acurácia do que outros algoritmos baseados em grid.
- Generaliza outros algoritmos particionais, hierárquicos e baseados em condições locais.
- Invariante a ruídos. Encontra os mesmos atratores mesmo quando o número de objetos ruído tende a infinito.
- Significativamente mais rápido que algoritmos baseados em densidade existentes.
- Permite uma descrição matemática compacta de clusters de forma arbitrária em espaços de altas dimensões.
- Requer cuidado na determinação dos parâmetros
- Encontra clusters de diferentes formas e tamanhos, mas pode ter problemas com conjuntos de alta dimensionalidade e dados que contenham clusters de densidades muito diferentes (Tan, et al., 2006).
- Complexidade no pior caso $O(N\log_2 N)$ (Theodoridis e Koutroumbas 1999)

Referências Bibliográficas

Hinneburg, A. and Keim, D. An efficient approach to clustering in large multimedia databases with noise. *In Proceedings KDD'98*, pages 58-65. AAAI Press, 1998.

Tan, P.-N., Steinbach, M., and Kumar, V., *Introduction to Data Mining*, Addison-Wesley, 2006.

Theodoridis, S. and Koutroumbas, K. *Pattern Recognition*, Academic Press, New York, 1999.

ANEXOS

Resumo da Etapa 1 do algoritmo DENCLUE

Pré-agrupamento – (mapear e armazenar a porção relevante da base de dados)

✓ Divida o (hiper-)retângulo delimitador mínimo em hiper-cubos d-dimensionais, com aresta de comprimento 2σ;

✓ Descarte os hiper-cubos vazios;

✓ Armazene em uma árvore de busca aleatória ou em uma árvore $B+$:

- as chaves dos cubos não vazios (*de acordo com suas posições relativas a alguma origem*)
- ponteiros para os objetos que pertencem a cada cubo
- o número de objetos que pertencem a cada cubo (N_c)
- e os vetores de soma dos objetos que pertencem ao cubo c ($\Sigma_{xec} X$)

✓Encontre os hiper-cubos altamente densos, isto é:

$C_{SP} = \{c \in CP \mid N_c \geq \xi_c\}$
(C_P = conjunto de cubos não vazios. Em geral $||C_{SP}|| \ll ||C_P||$)
($\xi_c = \xi / s t$, em que ξ é o limiar de densidade que define um objeto como atrator de densidade.
(esta etapa reduz o número de testes na etapa seguinte)

✓ Encontre os hiper-cubos, que possuem conexão com algum hiper-cubo denso.

$$C_{conexão} = \{c \in C_P \mid \exists c_s \in C_{SP} \, e \exists conexão\ (c_s, c)\}.$$

Existe conexão entre $c_1, c_2 \in CP$, se $d(centróide(c_1), centróide(c_2)) \leq 4\sigma$

Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento

✓Determine $C_r = C_{SP} \cup C_{conexão}$

(C_r é o conjunto dos hiper-cubos densos ou conectados a hiper-cubos densos)

✓Para $x \in c \in c, c_1 \in C_r$, tome

$near(x) = \{x_1 \in c_1 \mid d(mean(c_1), x) \leq k\sigma \; e \; \exists conexão(c_1, c)\}$

($k = 4$ é adequado para fins práticos)
(o conjunto $near(x)$ é usado para calcular a função de densidade local)

✓ Calcule a função de densidade local e o gradiente:

$\hat{f}^D(x) = \sum_{x_1 \in near(x)} \exp(-\frac{d(x,x_1)^2}{2\sigma^2})$; $\nabla \hat{f}^D(x) = \sum_{x_1 \in near(x)} (x_1 - x) \exp(-\frac{d(x,x_1)^2}{2\sigma^2})$

✓ Encontre os atratores de densidade (máximos locais de $\hat{f}^D$), por uma busca hill-climbing, guiada pelo gradiente de $\hat{f}^D$: $x = x^0, x^{i+1} = x^i + \delta \frac{\nabla \hat{f}^D(x^i)}{\|\nabla \hat{f}^D(x^i)\|}$

- O cálculo pára em $k \in \mathbb{N}$ se $\hat{f}^D(x^{k+1}) < \hat{f}^D(x^k)$ e toma $x^* = x^k$ como um novo atrator de densidade.
- Após a determinação do atrator de densidade x^* para o objeto x e se $\hat{f}^D(x^*) \geq \xi$ o objeto x é unido a x^*
- Por questão de eficiência, o algoritmo armazena todos os pontos x' com $d(x^i, x') \leq \sigma/2$, para qualquer iteração $0 \leq i \leq k$ durante a busca hill-climbing e une estes objetos ao cluster definido por x^*. (usando esta heuristica, todos os pontos que estão próximos do caminho de x ao seu atrator de densidade x^* podem ser classificados sem a necessidade da busca hill-cibing ser aplicada a eles.)

Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento

✓Encontre $near(x)$. **continuação...**

Objeto	*x*	D=near(x)
1	-1,31	X-{c_3}
2	-0,43	X-{c_3}
3	0,34	X-{c_3}
4	3,57	X
5	2,76	X-{c_3}
6	0,30	X-{c_3}
7	9,06	X-{c_1, {5}, {9}}
8	4,45	X
9	2,87	X-{c_3}
10	4,42	X

Resumo da Etapa 2 do algoritmo DENCLUE

Agrupamento – (obter as densidades e agrupamentos a partir da estrutura gerada da Etapa 1)

✓ Encontre os atratores de densidade (máximos locais de $\hat{f}^D$), por uma busca hill-climbing, guiada pelo gradiente de $\hat{f}^D$ e una os objetos aos atratores mais próximos.

$$x = x^0, x^{i+1} = x^i + \delta \frac{\nabla \hat{f}^D(x^i)}{\|\nabla \hat{f}^D(x^i)\|}$$

Objeto	*x*	$\hat{f}^D(x)$	$\nabla \hat{f}^D(x)$
1	-1,31	3,002	2,768
2	-0,43	3,839	1,373
3	0,34	4,082	0,168
4	3,57	4,681	-0,617
5	2,76	4,596	0,871
6	0,30	4,079	0,209
7	9,06	1,018	-0,087
8	4,45	4,006	-2,748
9	2,87	4,636	0,752
10	4,42	4,042	-2,689

Complexidade

PASSOS DENCLUE:

1. MBR ←Determinar MBR(D)
   Minimal Bounding (hyper-)Rectangle.
2. C_P ← Determinar conjunto de cubos não vazios(D, MBR, σ)
   C_{sp} ← Determinar cubos altamente densos(C_P, ξ_c)
3. map, C_r ← determinar cubos altamente densos ou que possuem conexão com algum(C_P, C_{sp}, σ)
4. Agrupamento ← determinar os atratores de densidade (map, C_r, σ, ξ)

Complexidade em cada passo, em que $\|.\|$ é a cardinalidade do conjunto:

1. $O(\|D\|)$,
2. $O(\|D\|+\|C_P\|\log(\|C_P\|))$
3. $O(\|C_{sp}\|.\|C_P\|)$
4. $O(\|D_r\|.\log(\|C_r\|))$, D_r=D-{outliers}

Segundo Theodoridis e Koutroumbas, o pior caso do DENCLUE é executado com $O(N\log_2 N)$, se igualando ao DBSCAN.