Agrupamento fuzzy de dados de alta dimensionalidade e suas implicações

Fernanda Eustáquio

26 de Fevereiro de 2019

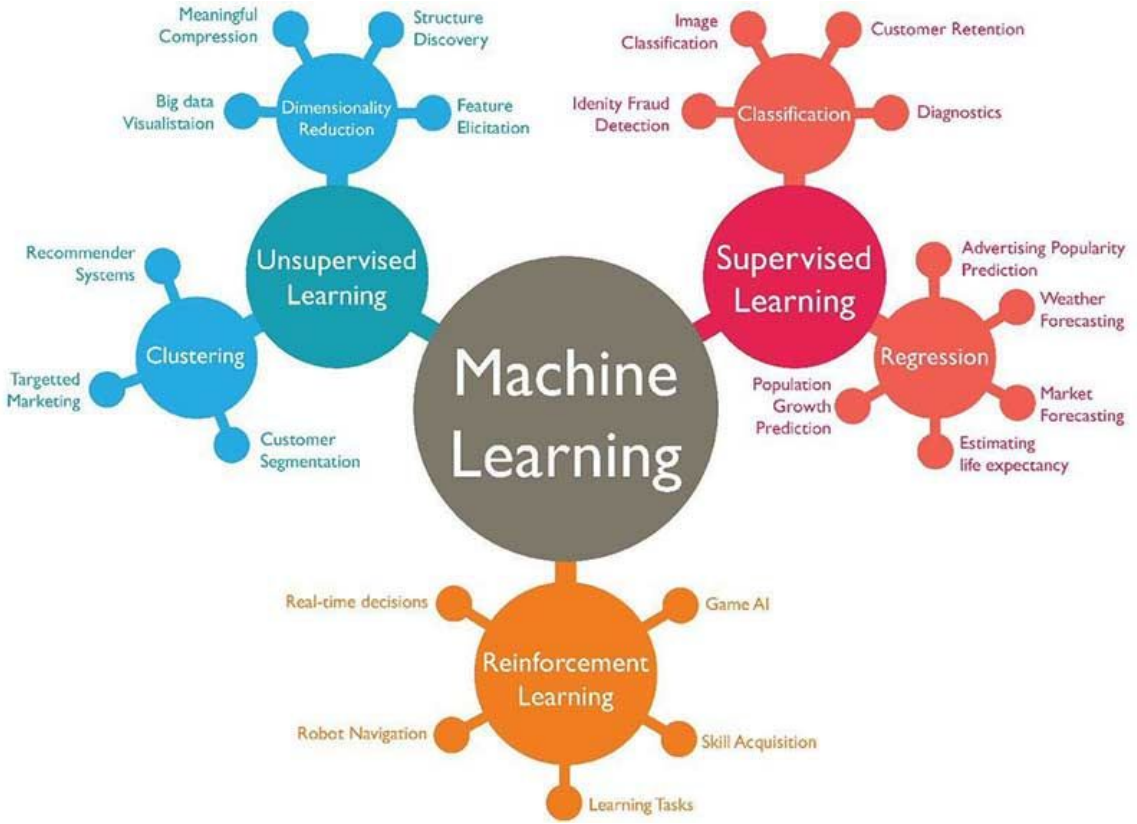
Universidade Federal da Bahia
Departamento de Ciência da Computação

Apresentação

- 2012.1 - 2017.1 Graduação em Ciência da Computação na Universidade Federal da Bahia (UFBA);
  - 2015.1 Iniciação Científica na área de Machine Learning;
    - F. Eustáquio, H. Camargo, S. Rezende, and T. Nogueira. **On fuzzy cluster validity indexes for high dimensional feature space**. In Advances in Fuzzy Logic and Technology 2017: Proceedings of The 10th Conference of the European Society for Fuzzy Logic and Technology, 2017, Warsaw, Poland, Volume 2, pages 12–23, 2018.
- 10/2017 - 09/2019 Mestrado em Ciência da Computação na área de Inteligência Computacional no Programa de Pós-graduação em Ciência da Computação (PGCOMP) da UFBA.
  - F. Eustáquio and T. Nogueira, **On monotonic tendency of some fuzzy cluster validity indices for high-dimensional data**, in 2018 7th Brazilian Conference on Intelligent Systems (BRACIS) (2018), pp. 558–563.

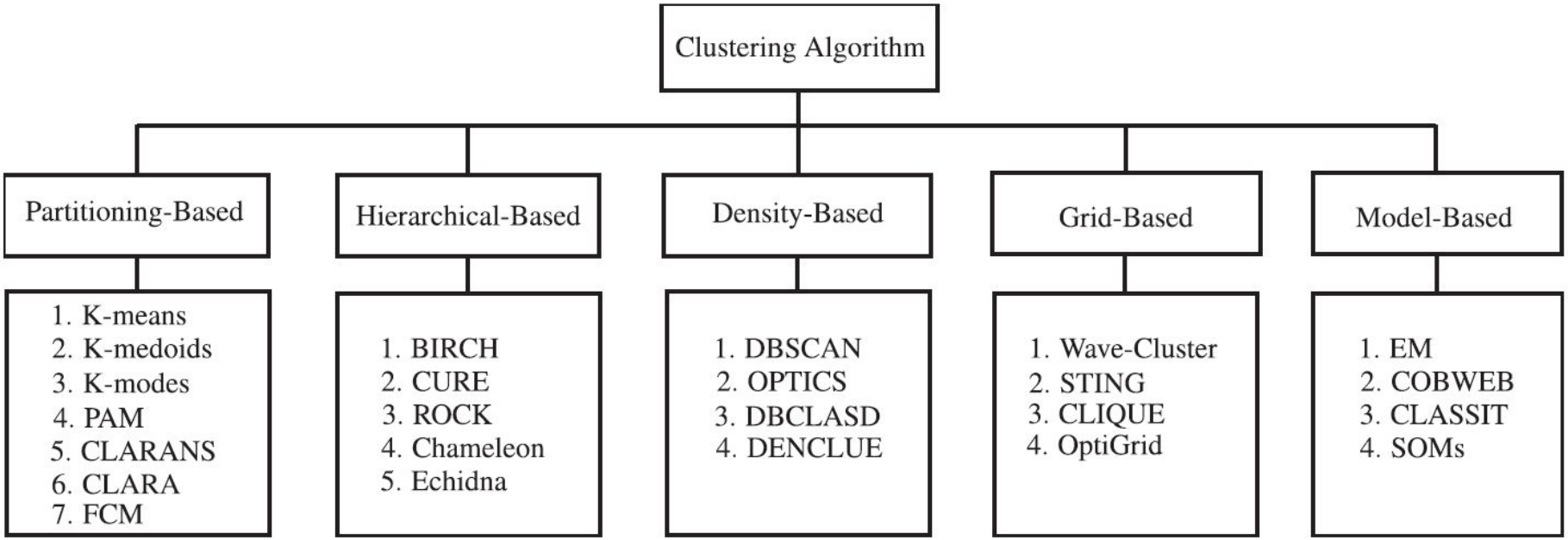
Agrupamento de dados

- Aprendizado Não-Supervisionado;
- Tarefa descritiva: identificação de comportamentos intrínsecos do conjunto de dados;
- Dados não rotulados.



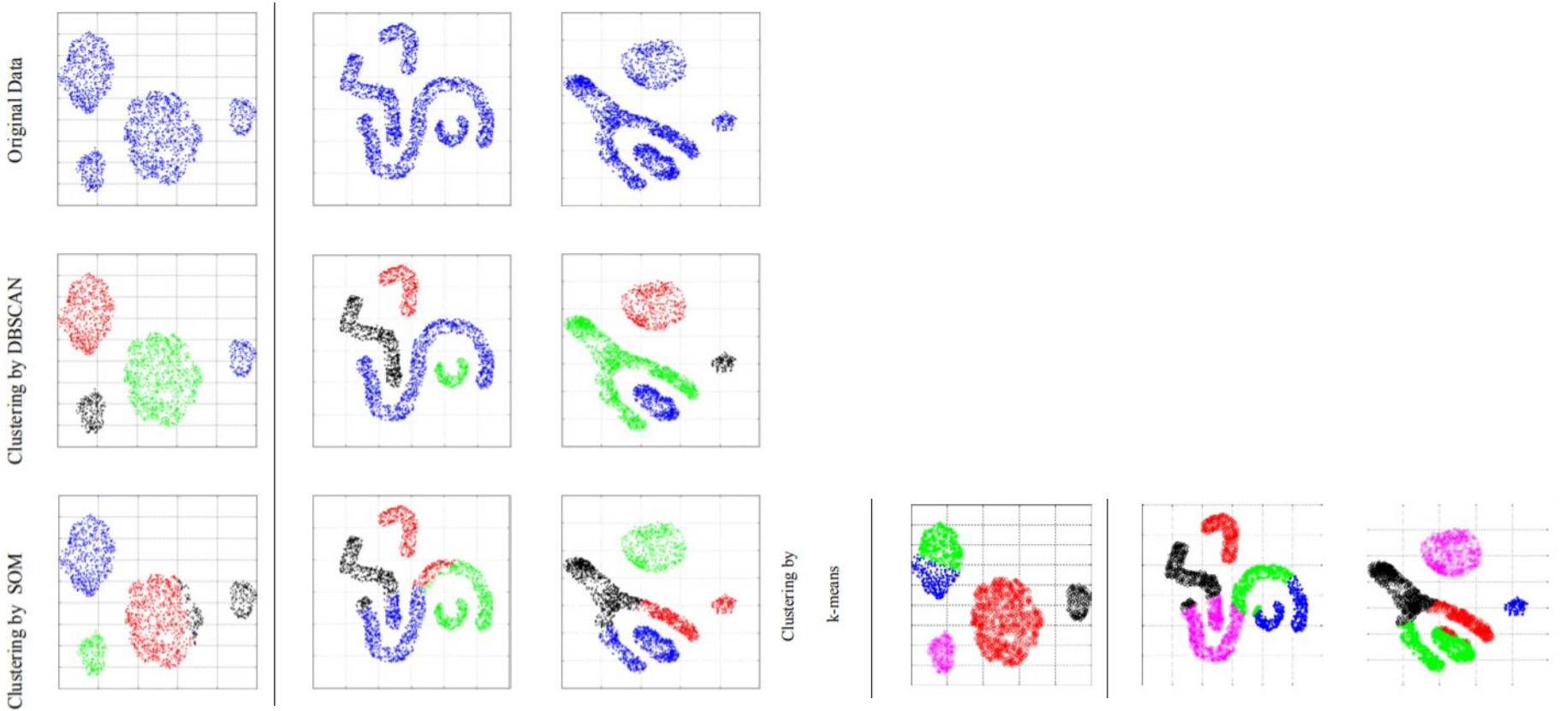
http://sites.ieee.org/futuredirections/2018/06/29/machine-human-integrated-learning-technologies-iii/

Uma visão geral da taxonomia de clustering



A. Fahad *et al*., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," in *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 267-279, Sept. 2014.

Características de agrupamento em dados espaciais



MUNTAZ, K & DURAISWAMY, K. (2010) An Analysis on Density Based Clustering of Multi Dimensional Spatial Data. Indian Journal of Computer Science and Engineering, Vol 1(1):8-12.

Categorização de algoritmos de agrupamento em relação a propriedades de big data

Categories	Abb. name	Volume			Variety		Velocity
		Size of Dataset	Handling High Dimensionality	Handling Noisy Data	Type of Dataset	Clusters Shape	complexity of Algorithm
Partitional algorithms	K-Means [25]	Large	No	No	Numerical	Non-convex	O(nkd)
	K-modes [19]	Large	Yes	No	Categorical	Non-convex	O(n)
	K-medoids [33]	Small	Yes	Yes	Categorical	Non-convex	$O(n^2dt)$
	PAM [31]	Small	No	No	Numerical	Non-convex	$O(k(n-k)^2)$
	CLARA [23]	Large	No	No	Numerical	Non-convex	$O(k(40+k)^2+k(n-k))$
	CLARANS [32]	Large	No	No	Numerical	Non-convex	$O(kn^2)$
	FCM [6]	Large	No	No	Numerical	Non-convex	O(n)
Hierarchical algorithms	BIRCH [40]	Large	No	No	Numerical	Non-convex	O(n)
	CURE [14]	Large	Yes	Yes	Numerical	Arbitrary	$O(n^2 \log n)$
	ROCK [15]	Large	No	No	Categorical and Numerical	Arbitrary	$O(n^2+nmmma+n^2logn)$
	Chameleon [22]	Large	Yes	No	All type of data	Arbitrary	$O(n^2)$
	ECHIDNA [26]	Large	No	No	Multivariate Data	Non-convex	$O(N * B(1 + \log_B m))$
Density-based algorithms	DBSCAN [9]	Large	No	No	Numerical	Arbitrary	O(nlogn) If a spatial index is used Otherwise, it is $O(n^2)$.
	OPTICS [5]	Large	No	Yes	Numerical	Arbitrary	O(nlogn)
	DBCLASD [39]	Large	No	Yes	Numerical	Arbitrary	$O(3n^2)$
	DENCLUE [17]	Large	Yes	Yes	Numerical	Arbitrary	$O(\log\|D\|)$
Grid- based algorithms	Wave-Cluster [34]	Large	No	Yes	Special data	Arbitrary	O(n)
	STING [37]	Large	No	Yes	Special data	Arbitrary	O(k)
	CLIQUE [21]	Large	Yes	No	Numerical	Arbitrary	O(Ck + mk)
	OptiGrid [18]	Large	Yes	Yes	Special data	Arbitrary	Between O(nd) and O(nd log n)
Model- based algorithms	EM [8]	Large	Yes	No	Special data	Non-convex	O(knp)
	COBWEB [12]	Small	No	No	Numerical	Non-convex	$O(n^2)$
	CLASSIT [13]	Small	No	No	Numerical	Non-convex	$O(n^2)$
	SOMs [24]	Small	Yes	No	Multivariate Data	Non-convex	$O(n^2m)$

A. Fahad *et al*., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," in *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 267-279, Sept. 2014.

Suponha que um usuário busca por notícias em categorias específicas de um portal

https://esporte.uol.com.br/ultimas-noticias/ag-fight/2017/08/30/max-holloway-desafia-conor-mcgregor-para-revanche-nos-pesos-penas.htm

https://noticias.uol.com.br/ultimas-noticias/agencia-estado/2017/08/28/para-moro-gilmar-mendes-deveria-manter-prisao-em-2-grau.htm

L HOST PAGSEGURO CURSOS UOL RESOLVE

UOL

BUSCA E-MAIL CONTA UOL

sine Bate-papo Notícias Carros Economia Folha Esporte Entretenimento TV e famosos Jogos Estilo Educação Víde

Televisão

"Baby boom" dos Abravanel! Silvio Santos ganhará três netos em 2018

https://www.bol.uol.com.br/entretenimento/2017/07/14/baby-boom-dos-abravanel-silvio-santos-ganhara-tres-netos-em-2018.htm

Deputado Fábio Faria e a mulher, Patrícia Abravanel, tentam anular parte da delação de executivo da J&F

Casal questiona trecho da delação de Ricardo Saud em que ele relata propina para o deputado. Defesa vai usar como prova mensagem deixada no telefone de Patricia pela mulher de Joesley Batista.

Filha de Silvio Santos participou de jantar para negociar propina, diz delator 119

Do UOL, em São Paulo 19/05/2017 | 19h45 > Atualizada 21/05/2017 | 15h47

Ouvir texto Imprimir Comunicar erro

https://g1.globo.com/politica/noticia/deputado-fabio-faria-e-a-mulher-patricia-abravanel-tentam-anular-parte-da-delacao-da-jf.ghtml
https://noticias.uol.com.br/politica/ultimas-noticias/2017/05/19/patricia-abravanel-participou-de-jantar-para-negociar-propina-diz-delator.htm

POLÍTICA

Deputado Fábio Faria e a mulher, Patrícia Abravanel, tentam anular parte da delação de executivo da J&F

Casal questiona trecho da delação de Ricardo Saud em que ele relata propina para o deputado. Defesa vai usar como prova mensagem deixada no telefone de Patricia pela mulher de Joesley Batista.

Filha de Silvio Santos participou de jantar para negociar propina, diz delator 119

Do UOL, em São Paulo 19/05/2017 | 19h45 > Atualizada 21/05/2017 | 13h47

Ouvir texto Imprimir Comunicar erro

Notícias praticamente duplicadas!

ESPORTE

Fernanda Gentil termina casamento com empresário carioca

172

UOL Esporte 06/04/2016 | 17:21

Compartilhe | Imprimir | Comunicar erro

gentilfernanda SEGUIR

12.4k curtidas

gentilfernanda Aê como meu post aqui gerou duvida, vou esclarecer apenas (e somente) por aqui: postei só porque gostei da frase e da mensagem que ela passa e apaguei porque Momó achou

https://uolesportevetv.blogosfera.uol.com.br/2016/04/06/fernanda-gentil-termina-casamento-com-empresario-carioca/

Fernanda Gentil se separa do marido, após cinco anos de casamento

10

Do UOL, no Rio 06/04/2016 | 16h54

Ouvir texto | Imprimir | Comunicar erro

https://tvefamosos.uol.com.br/noticias/redacao/2016/04/06/fernanda-gentil-se-separa-do-marido-apos-cinco-anos-de-casamento.htm

FAMOSOS

Deputado Fábio Faria e a mulher, Patrícia Abravanel, tentam anular parte da delação de executivo da J&F

0.55 POLÍTICA FAMOSOS 0.45

Filha de Silvio Santos participou de jantar para negociar propina, diz delator 119

Do UOL, em São Paulo 19/05/2017 | 19h45 > Atualizada 21/05/2017 | 13h47

Ouvir texto Imprimir Comunicar erro

0.2

ESPORTE FAMOSOS

0.8

Fernanda Gentil termina casamento com empresário carioca 172

UOL Esporte 06/04/2016 17:21

Comunicar erro

SEGUIR

12.4k curtidas 28 min

gentilfernanda Aê como meu post aqui gerou duvida, vou esclarecer apenas (e somente) por aqui: postei só porque gostei da frase e da mensagem que ela passa e apaguei porque Momô achou

Fernanda Gentil se separa do marido, após cinco anos de casamento 10

Do UOL, no Rio 06/04/2016 16h54

Ouvir texto Imprimir Comunicar erro

Agrupamento fuzzy de dados de alta dimensionalidade

- Para organização automática e flexível de documentos;
- Por acreditar que as fronteiras entre os grupos de dados de alta-dimensionalidade são ambíguas e não bem separadas;
- Permite que um mesmo objeto pertença a mais de um grupo com diferente grau de pertinência;
- Fuzzy c-Means (FCM) é o algoritmo de agrupamento fuzzy mais bem conhecido e utilizado.

Bases de alta dimensionalidade

- Comumente utilizadas em técnicas de agrupamento;
- Agrupamento não necessita de informação prévia para extrair a estrutura dos dados;
- Bases de centenas a milhares de dimensões;
  - Textuais;
  - Expressão Gênica (microarray);
  - Imagem.

Document-term matrix

- Object: document (d_k);
- Feature: term (t_p);
- Cell: Term Frequency-Inverse Document Frequency ($tf - idf(t_p, d_k)$);
- Sparse.

	t_1	t_2	$\cdots$	t_p
d_1	$tf - idf(t_1, d_1)$	$tf - idf(t_2, d_1)$	$\cdots$	$tf - idf(t_p, d_1)$
d_2	$tf - idf(t_1, d_2)$	$tf - idf(t_2, d_2)$	$\cdots$	$tf - idf(t_p, d_2)$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
d_n	$tf - idf(t_1, d_n)$	$tf - idf(t_2, d_n)$	$\cdots$	$tf - idf(t_p, d_n)$

Gene expression matrix

- Object: gene (g_k);
- Feature: condition under a gene is developed (c_p);
- Cell: gene expression ($expr.(c_p, g_k)$);
- Not sparse.

	c_1	c_2	$\cdots$	c_p
g_1	$expr.(c_1, g_1)$	$expr.(c_2, g_1)$	$\cdots$	$expr.(c_p, g_1)$
g_2	$expr.(c_1, g_2)$	$expr.(c_2, g_2)$	$\cdots$	$expr.(c_p, g_2)$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
g_n	$expr.(c_1, g_n)$	$expr.(c_2, g_n)$	$\cdots$	$expr.(c_p, g_n)$

Bases de alta dimensionalidade

Data set	p	n	#classes	Domain	Matrix Sparsity (%)
Sorlie	456	85	5	Breast Cancer	1.19
Christensen	1,413	217	3	N/A	0.46
CSTR	1,725	299	4	Scientific	96.86
Alon	2,000	62	2	Colon Cancer	1.61
Khan	2,308	63	4	Small Round Blue Cell Tumors	1.59
NSF	2,743	1,600	16	Scientific	99.75
Gravier	2,905	168	2	Breast Cancer	0.6
SyskillWebert	4,339	333	4	Web pages	97.85
Su	5,565	100	4	N/A	13.63
Hitech	6,593	600	6	Newspaper	97.92
Shipp	6,817	58	2	Lymphoma	1.31
Pomeroy	7,128	60	2	CNS Tumor	1.67
Golub	7,129	72	3	Leukemia	1.4
West	7,129	49	2	Breast Cancer	2.04
WAP	8,049	1,560	20	Web pages	98.52
Irish-Sentiment	8,658	1,660	3	Sentiment Analysis	98.7
Subramanian	10,100	50	2	N/A	2.01
Opinosis	10,784	51	3	Sentiment Analysis	95.73
20Newsgroups	11,015	2,000	4	E-mails	99.11
Gordon	12,533	181	2	Lung Cancer	0.55
Singh	12,600	102	2	Prostate Cancer	1.09
Chiaretti	12,625	111	2	Leukemia	0.78
Tian	12,625	173	2	Myeloma	0.58
Yeoh	12,625	248	6	Leukemia	0.4
La1s	13,195	3,204	6	News articles	98.9
Chin	22,215	118	2	Breast Cancer	0.85
Borovecki	22,283	31	2	Huntington's Disease	3.31
Burczynski	22,283	127	3	Crohn's Disease	0.79
Chowdary	22,283	104	2	Breast Cancer	1.31
Reviews	22,926	4,069	5	News articles	99.2

- Número de dimensões é bem maior do que o número de objetos

  $p >> n$

- O número de dimensões da base Borovecki chega a ser 719 vezes maior do que o n

Fuzzy c-Means

- Versão fuzzy do k-Means;
- Quando m ➔ 1 (1 < m < ∞), FCM converge para o k-Means;
- Atribui um grau de pertinência de um objeto para cada grupo;
- FCM é um processo iterativo que inicializa aleatoriamente sua pseudo-partição *U* ou os centróides *V* = {v1, v2, ..., vc} dos *c* grupos;
- Nos processos seguintes, FCM atualiza *U* e *V*.
- Se ||U(t-1) - U(t)|| < *E* ou o número máximo de iterações foi alcançado, para. Caso contrário atualiza *U* e *V*.
- As atualizações de *U* e *V* devem minimizar a dissimilaridade entre um objeto e o centróide pela função objetivo

$$J_m(U, V; X) = \sum_{i=1}^{c} \sum_{k=1}^{n} A_i^m(x_k) \|x_k - v_i\|^2.$$

Fuzzy c-Means - ***fclust*** (Fuzzy clustering) package

Algorithms for fuzzy clustering, cluster validity indices and plots for cluster validity and visualizing fuzzy clustering results.

```
FKM (X, k, m, RS, stand, startU, conv, maxit)
```

Arguments

`X`	Matrix or data.frame
`k`	Number of clusters (default: 2)
`m`	Parameter of fuzziness (default: 2)
`RS`	Number of (random) starts (default: 1)
`stand`	Standardization: if `stand=1`, the clustering algorithm is run using standardized data (default: no standardization)
`startU`	Rational starting point for the membership degree matrix `U` (default: no rational start)
`conv`	Convergence criterion (default: 1e-9)
`maxit`	Maximum number of iterations (default: 1e+6)

Fuzzy c-Means - ***ppclust*** (Probabilistic and Possibilistic Cluster Analysis) package

```
fcm(x, centers, memberships, m=2, dmetric="sqeuclidean", pw = 2,
    alginitv="kmpp", alginitu="imembrand",
    nstart=1, iter.max=1000, con.val=1e-09,
    fixcent=FALSE, fixmemb=FALSE, stand=FALSE, numseed)
```

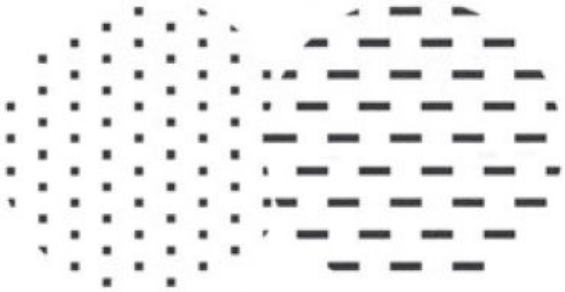
Arguments

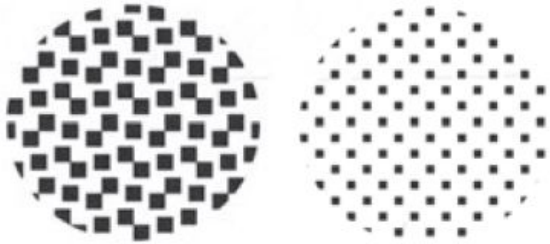
x	a numeric vector, data frame or matrix.
centers	an integer specifying the number of clusters or a numeric matrix containing the initial cluster centers.
memberships	a numeric matrix containing the initial membership degrees. If missing, it is internally generated.
m	a number greater than 1 to be used as the fuzziness exponent or fuzzifier. The default is 2.
dmetric	a string for the distance metric. The default is 'sqeuclidean' for the squared Euclidean distances. See get.dmetrics for the alternative options.
pw	a number for the power of Minkowski distance calculation. The default is 2 if the dmetric is 'minkowski'.

R topics documented:

Validação do agrupamento

- Processo de validação é feito por índices de validação de agrupamento fuzzy (CVIs);
- CVIs quantificam a qualidade da partição obtida;
- Dois critérios são utilizados para avaliar a qualidade dos grupos:
  - **Compacidade**: mede a proximidade dos objetos do grupo, como, por exemplo, a variância. Baixo valor de variância indica que os elementos de um grupo são próximos (intra-cluster);
  - **Separação**: mede quão distintos dois grupos são. Computa a distância inter-cluster.

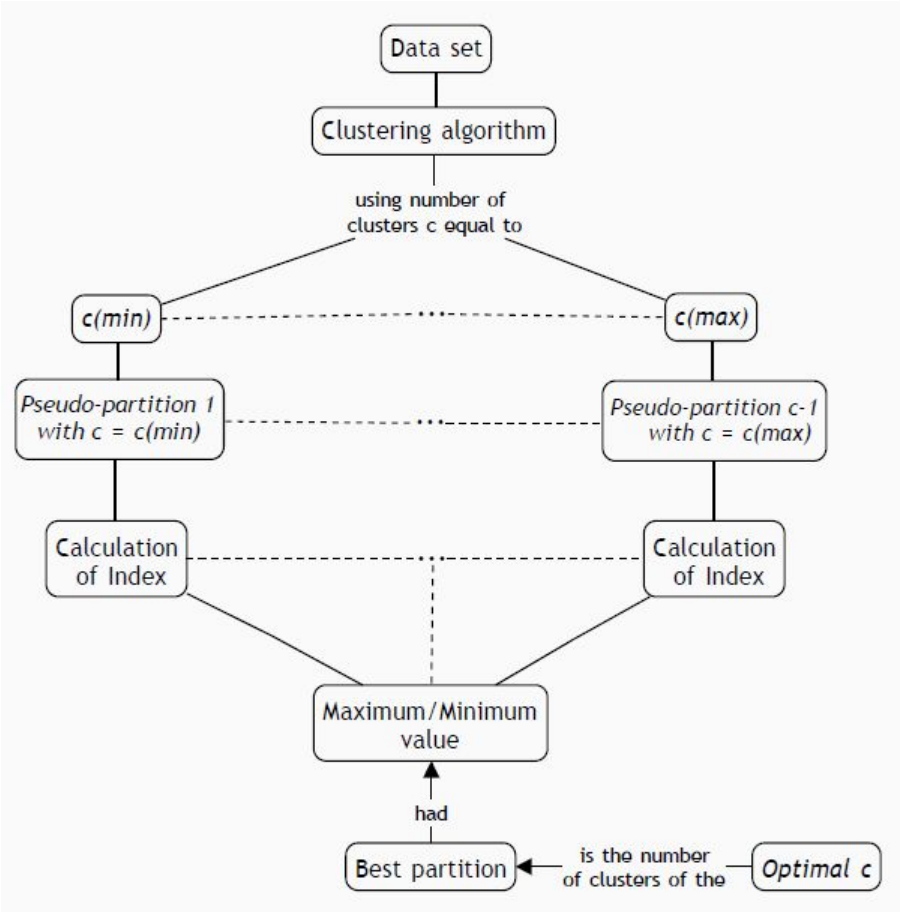




Steinbach M., Ertöz L., Kumar V. (2004) The Challenges of Clustering High Dimensional Data. In: Wille L.T. (eds) New Directions in Statistical Physics. Springer, Berlin, Heidelberg

Validação do agrupamento

- Na maioria das aplicações, o número de grupos *c* é desconhecido e os CVIs são utilizados com o critério relativo de avaliação para encontrar o melhor número *c;*
- Regra de ouro: **c(max) ⩽ √n**;
- Utiliza o número de classes de uma base rotulada como referência de número esperado de clusters;
- **Cluster é diferente de classe!**



Índices de validação

Index	Definition	Optimal value
PC [3]	$\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}A_i^2(x_k)$	Maximum
MPC [9]	$1-\frac{c}{c-1}(1-PC)$	Maximum
IPC [37]	$100\left[\frac{PC(c-1)-PC(c)}{PC(c-1)}-\frac{PC(c)-PC(c+1)}{PC(c)}\right]$	Maximum
PE [2]	$-\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}A_i(x_k)\log_a(A_i(x_k))$	Minimum
NPE [14]	$\frac{nPE}{n-c}$	Minimum
KYI [27]	$\frac{2}{c(c-1)}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\sum_{k=1}^{n}[c\times min(A_i(x_k),A_j(x_k))\times h(x_k)^1]$	Minimum
P [8]	$\frac{1}{n}\sum_{k=1}^{n}max_i(A_i(x_k))-\frac{1}{K}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\left[\frac{1}{n}\sum_{k=1}^{n}min(A_i(x_k),A_j(x_k))\right]$	Maximum
MPO [22]	$\left(\frac{c+1}{c-1}\right)^{1/2}\frac{\sum_{k=1}^{n}\sum_{i=1}^{c}A_i{}^2(x_k)}{min_{1\le i\le c}\left\{\sum_{k=1}^{n}A_i{}^2(x_k)\right\}}-\frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}O_{ijk}(c,U)^2\right)$	Maximum
GD [25]	$\frac{\sum_{k=1}^{n}max_{1\le i\le c}(A_i(x_k))-max_{1\le j\le c,j\neq i}(A_j(x_k))}{n}-\left(\frac{c}{n}\right)$	Maximum

Índices de validação

Index	Definition	Optimal value
FS [19]	$\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^m(x_k)(\|\|x_k - v_i\|\|^2 - \|\|v_i - \bar{v}\|\|^2)$	Minimum
XB [46]	$\frac{\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^m(x_k)\|\|x_k - v_i\|\|^2}{n \times min_{k\neq i}\|\|v_i - v_j\|\|^2}$	Minimum
K [28]	$\frac{\sum_{k=1}^{n}\sum_{i=1}^{c} A_i^2(x_k)\|\|x_k - v_i\|\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|\|v_i - \bar{v}\|\|^2}{min_{i\neq j}\|\|v_i - v_j\|\|^2}$	Minimum
PBMF [32]	$\left(\frac{1}{c} \times \frac{\sum_{k=1}^{n}\|\|x_k - \bar{v}\|\|}{J'_m} \times max_{i,j=1}^{c}\|\|v_i - v_j\|\|\right)^2$	Maximum
PCAES [45]	$\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^2(x_k)/A_M - \sum_{i=1}^{c} exp\left(-min_{j\neq i}\|\|v_i - v_j\|\|^2/B_T\right)$	Maximum
T [39]	$\frac{\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^2(x_k)\|\|x_k - v_i\|\|^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{\substack{j=1 \\ j\neq i}}^{c}\|\|v_i - v_j\|\|^2}{min_{i\neq j}\|\|v_i - v_j\|\|^2 + 1/c}$	Minimum
SF [5]	$\frac{\sum_{k=1}^{n}(max_{1\leq i\leq c}(A_i(x_k)) - max_{1\leq j\leq c, j\neq i}(A_j(x_k)))^{\alpha} S_{x_k}(V;X)^3}{\sum_{k=1}^{n}(max_{1\leq i\leq c}(A_i(x_k)) - max_{1\leq j\leq c, j\neq i}(A_j(x_k)))^{\alpha}}$	Maximum
WLI [44]	$\frac{\sum_{i=1}^{c}\left(\frac{\sum_{k=1}^{n} A_i^2(x_k)\|\|x_k - v_i\|\|^2}{\sum_{k=1}^{n} A_i(x_k)}\right)}{(min_{i\neq j}\|\|v_i - v_j\|\|^2 + median_{i\neq j}\{\|\|v_i - v_j\|\|^2\})}$	Minimum

Índices de validação

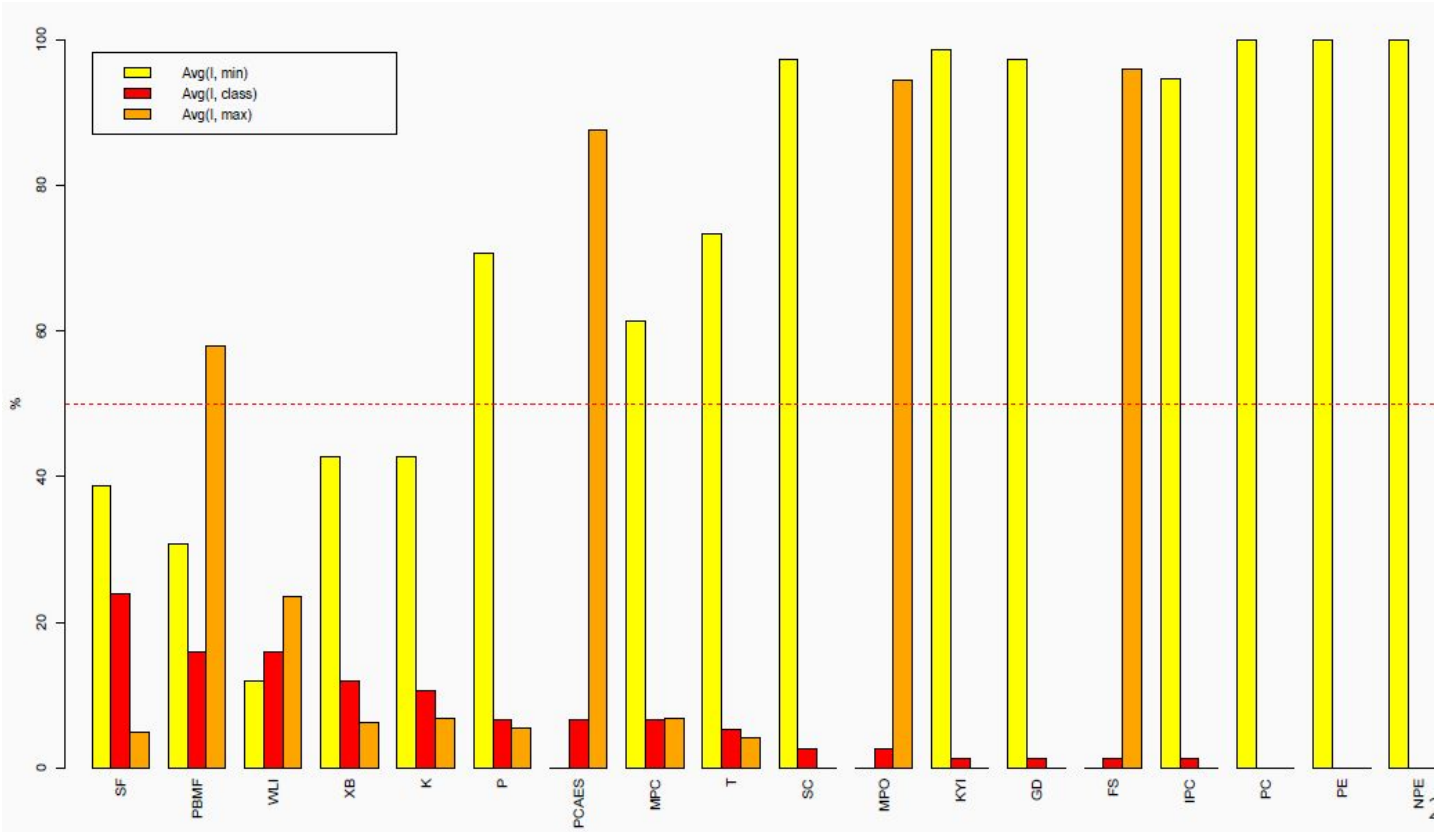
	Indices that use only the properties of fuzzy membership degrees	Indices that combine fuzzy membership degrees and the structure of data
Input	U	U, V, X
Compactness	$max_i A_i(x_k)\uparrow$ (Fuzzy Union)	$\|\|x_k - v_i\|\|\downarrow$
Separation	$min(A_i(x_k), A_j(x_k))\downarrow$ (Fuzzy Intersection)	$\|\|v_i - v_j\|\|\uparrow$
Indices	PC, MPC, IPC PE, NPE, KYI P, MPO, GD	FS, XB, K SC, PBMF, PCAES T, SF, WLI

Resultados

- Experimentos:
  - 30 bases de dados;
  - 5 valores de fator de fuzzificação *m* ([1.5; 2.5]);
  - Cosseno;
- FCM
  - Índice WLI selecionou o número esperado de grupos em respectivamente **40% das bases de dados**, com m = 1.8;
- FPCM
  - Índices K e XB selecionaram o número esperado de grupos em respectivamente **33.3% das bases**, com m = 2.2;
- PFCM
  - Índices K e XB selecionaram o número esperado de grupos em respectivamente **33.3% das bases**, com m = 2.5.

Resultados do FCM

SF reconheceu corretamente o número de grupos em **24% das pseudo-partições do FCM**.

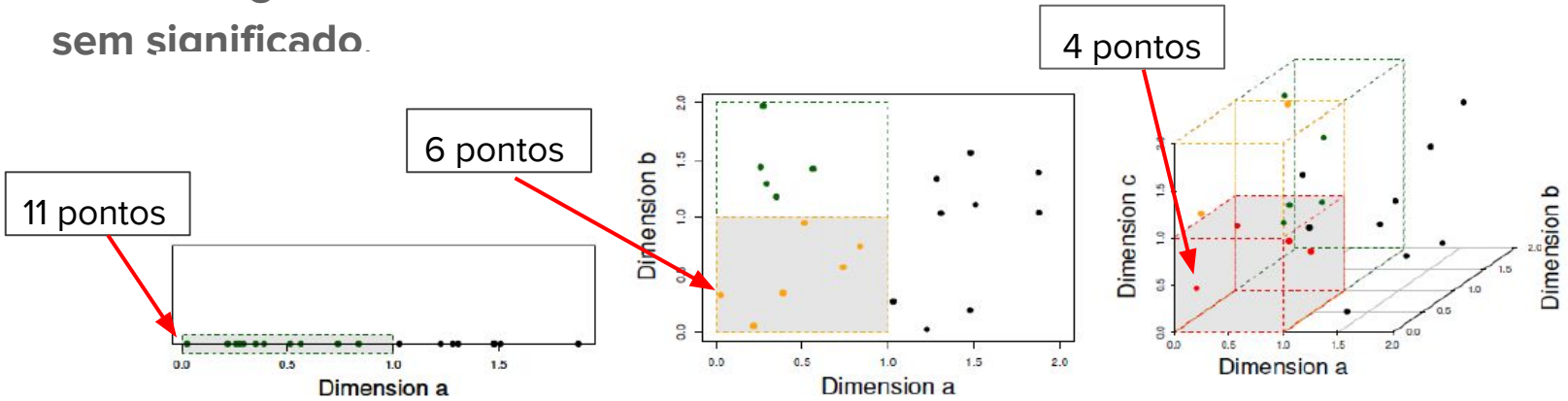


Desempenho deficiente dos CVIs para bases de alta dimensionalidade

1. Representação numérica dos objetos pode não possuir informação adequada para discriminar os grupos;
2. **O algoritmo utilizado pode não extrair a estrutura da base de dados;**
    a. FCM, FPCM, PFCM
3. Os **valores apropriados dos parâmetros** de um algoritmo que produz uma interpretação correta dos dados **nunca são utilizados**;
    a. *m* [1.5; 2.5]
    b. *m* [1.1; 10.0]
    c. c(max) = {c(class) + (min), 10}
4. Os CVIs **podem falhar** em indicar que bons grupos definidos por especialistas são realmente muito bons.
    a. 10 bases textuais?
    b. 20 bases de expressão gênica?
    c. 18 CVIs?

The *curse of dimensionality*

- ❏ Algoritmos de agrupamento fuzzy convencionais têm sido desenvolvidos para agrupar dados de baixa dimensionalidade;
- ❏ Muitas dimensões são irrelevantes e elas podem gerar muito ruído e esconder a estrutura dos grupos;
- ❏ Os pontos são provavelmente localizados em diferentes subespaços de dimensões;
- ❏ Adicionando mais dimensões, os pontos se tornam amplamente difusos até se tornarem **igualmente distantes** e suas **dissimilaridades** ou **similaridades** se tornam **sem significado**.



Lance Parsons, Ehtesham Haque, and Huan Liu. 2004. Subspace clustering for high dimensional data: a review. SIGKDD Explor. Newsl. 6, 1 (June 2004), 90-105. DOI: https://doi.org/10.1145/1007730.1007731

The *curse of dimensionality*

- **Muitas técnicas de agrupamento** que organizam objetos similares em um mesmo grupo e objetos dissimilares em diferentes grupos são **criticamente dependentes da medida de distância**;
- Este problema em medir a distância ou similaridade entre vetores de alta dimensão e consequentemente agrupar e validar bases de dados de alta dimensionalidade é bem conhecida como **a maldição da dimensionalidade**

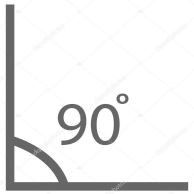
Medidas de Proximidade - Cosseno

- Mede a similaridade entre dois objetos calculando o cosseno do ângulo entre os vetores dada pela equação:

$$sim(x_k, v_i) = cos\theta = \frac{x_k \cdot v_i}{|x_k|\,|v_i|}$$

$$1 - sim(x_k, v_i).$$

- Cosseno = 0: o ângulo entre os objetos é de 90°, ou seja, os objetos não são similares



- Cosseno = 1: o ângulo entre os objetos é de 0°, indicando que estes são similares

Medidas de Proximidade

- Minkowski distance, para f ⩾ 1
  - f = 1: Manhattan
  - f = 2: Euclidiana

$$\left[\sum_{l=1}^{p}(x_{kl} - v_{il})^f\right]^{1/f}$$

- Métrica de Distância Fracionada: f ∈ (0; 1);
- **Para bases de alta dimensionalidade é preferível utilizar medidas com valor de expoente f < 2**.

Lidando com a *maldição da dimensionalidade*

Transformação de atributos (feature transformation):

- Modifica o espaço original de atributos através de combinação linear dos atributos para gerar poucas dimensões;
- As **novas dimensões não têm o mesmo significado** dos atributos originais ⟹ os resultados do agrupamento são usualmente difíceis de interpretar;
- Este processo **não descarta atributos irrelevantes** mas usa todos eles em suas combinações;
- É **bem indicado** para bases de dados onde a **maioria das dimensões** são **relevantes;**
- PCA (Principal Component Analysis), ICA (Independent Component Analysis), MDS (Multidimensional Scaling).

Lidando com a *maldição da dimensionalidade*

Feature weighting:

- Atribui um peso entre [0.0; 1.0] para cada atributo;
- O número de dimensões continua o mesmo;
- Atributos relevantes: peso próximo de 1.0;
- Atributos irrelevantes: peso próximo de 0.0;
- Extensão da seleção de atributos.

Feature name	Feature-weight
(SL, SW, PL, PW)	(1, 1, 1, 1)
(SL, PW)	(1, 0, 0, 1)
(SL, SW, PL, PW)	(0.8, 0.1, 0.3, 1.0)

Lidando com a *maldição da dimensionalidade*

- As técnicas anteriores só podem revelar **grupos de objetos que são similares em somente um subconjunto de seus atributos**;
- Pode ser muito difícil encontrar **somente um subconjunto de atributos** que seja **aceitável para todos os grupos**;
- Subspace clustering (SC) algorithms:
  - Encontram grupos que são similares em múltiplos, possivelmente subespaços sobrepostos de atributos;
  - Múltiplos subespaços de atributos podem conter diferentes grupos;
  - SC busca por grupos dentro de diferentes subespaços de atributos;
  - Estes subespaços de atributos consistem em várias combinações de atributos em uma mesma base de dados.

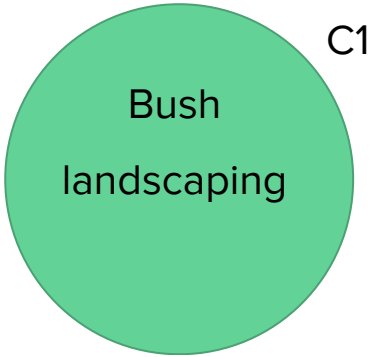
Subspace clustering

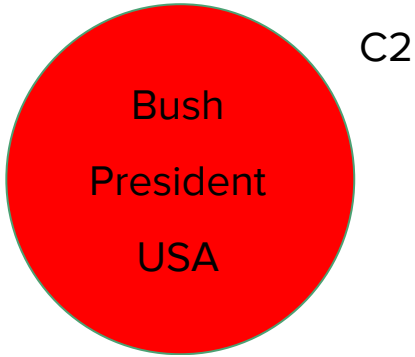
"A query for the term **Bush** could return documents on the **president** of the **United States** as well as information on **landscaping**"

- Os grupos de documentos provavelmente seriam relacionados em diferentes conjuntos de atributos

Atributos	At1	At2	At3	At4
	Bush	President	Landscaping	USA

Os termos **President** and **USA** se tornam atributos ruidosos que podem afetar a identificação de C1.





O termo **Landscaping** é o atributo ruidoso que pode afetar a identificação de C2.

Lance Parsons, Ehtesham Haque, and Huan Liu. 2004. Subspace clustering for high dimensional data: a review. SIGKDD Explor. Newsl. 6, 1 (June 2004), 90-105. DOI: https://doi.org/10.1145/1007730.1007731

Subspace clustering

Hard Subspace Clustering (HSC):

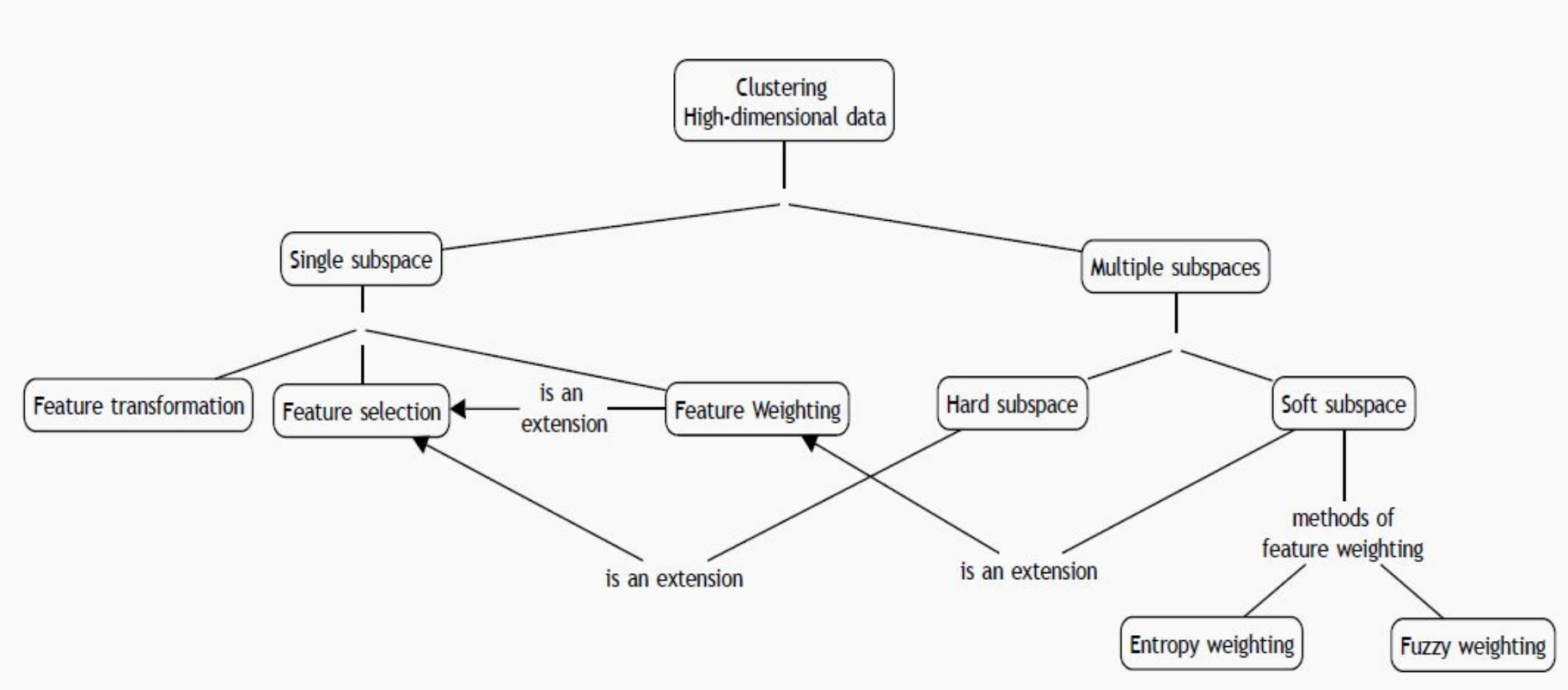
- Identifica subespaços exatos de atributos para diferentes grupos;
- Bottom-up: CLIQUE, ENCLUS and MAFIA;
- Top-down: PROCLUS, ORCLUS, FINDIT and -Clusters.

Soft Subspace Clustering (SSC):

- Atribui pesos diferentes para um mesmo atributo em diferentes grupos baseado na relevância desse atributo nos correspondentes grupos;

Clusters	Feature Terms			
	Bush (l_1)	*President* (l_2)	*Landscaping* (l_3)	*USA* (l_4)
C_1	0.4	0.02	0.5	0.08
C_2	0.4	0.22	0.08	0.3

Abordagens para agrupamento de dados de alta dimensionalidade



8th Brazilian Conference on Intelligent Systems

http://www.bracis2019.ufba.br/

Agradecimento

I would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES) for granting the scholarship during the period of this master’s degree at Federal University of Bahia.

Obrigada!

fernanda.eustaquio@ufba.br