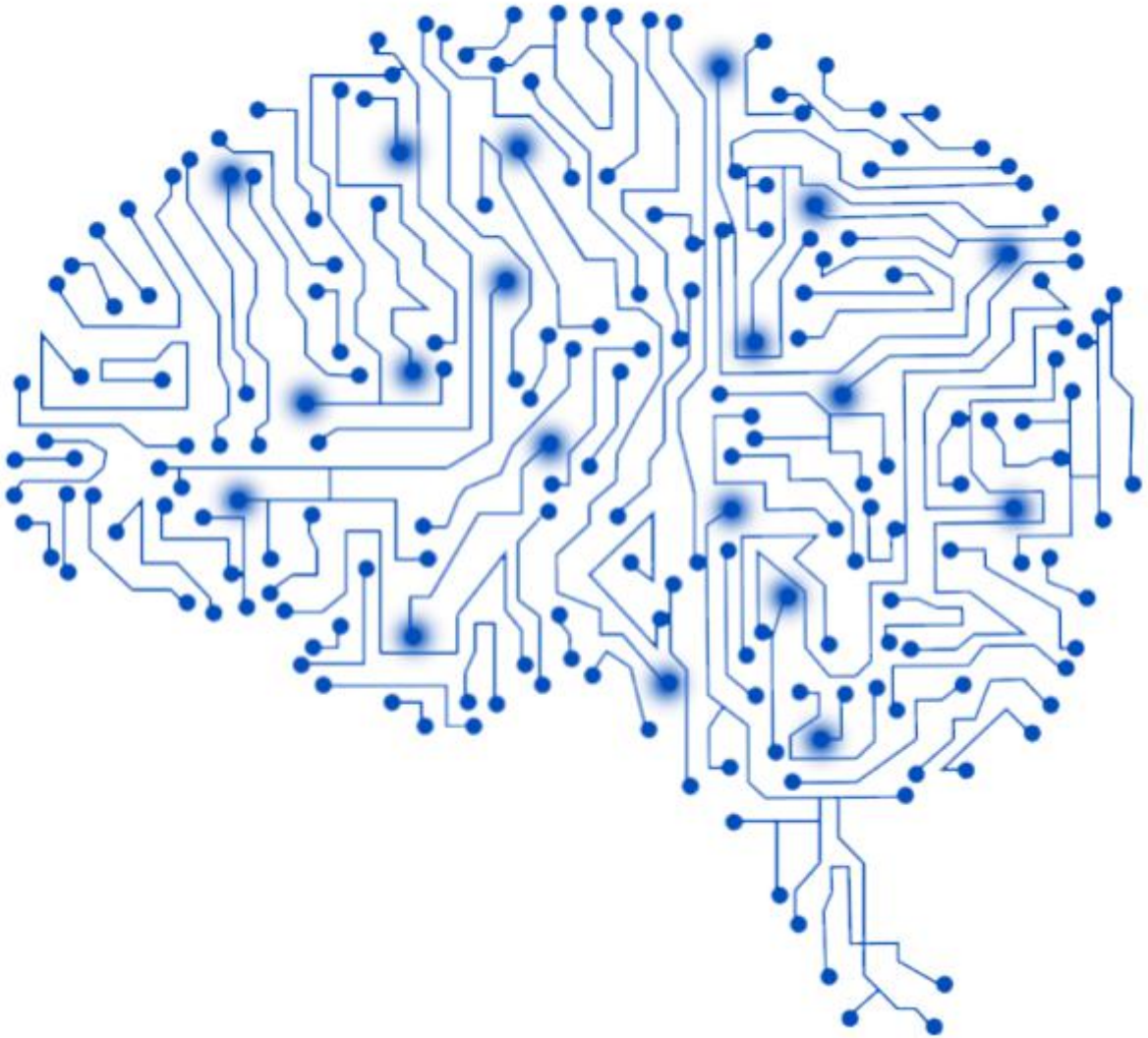
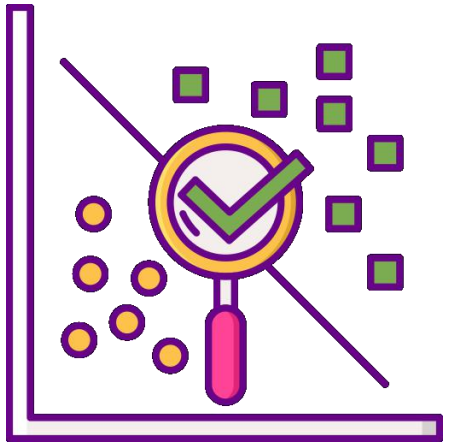
Cultura de Dados



Os Métodos Fuzzy e ROCK para Algoritmos de Agrupamento

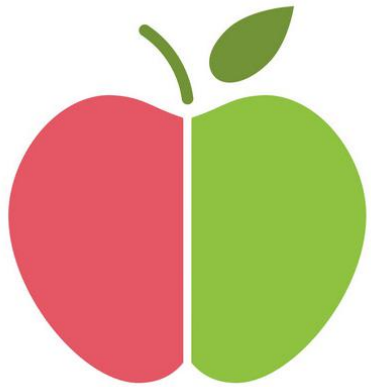
Kelver Campanha - Set/2020

Introdução



Clustering ou análise de cluster é uma técnica de aprendizado de máquina que envolve a atribuição de dados a clusters/agrupamentos de forma que os itens no mesmo cluster são os mais semelhantes possíveis, enquanto que os itens pertencentes a clusters diferentes são os mais diferentes quanto possível. Os clusters são formados por meio de medidas de similaridade.

Diferente de algoritmos como o k-Means, onde cada observação pertence exclusivamente a um único cluster. No agrupamento fuzzy, as observações **podem pertencer a vários clusters**. Por exemplo, uma maçã pode ser vermelha **ou** verde, mas uma maçã também pode ser vermelha **e** verde (agrupamento difuso). Aqui, a maçã pode ser vermelha até certo ponto, assim como verde até certo ponto, mas em geral, é comum alocar-se o elemento a aquele conglomerado para o qual sua probabilidade de pertencer é maior.



Medotologia

Assim como o método k-Means, o Método Fuzzy (Bezdek, 1981), é um método iterativo que requer a especificação do número de grupos *c* a ser utilizado. Suponha que haja *n* elementos amostrais e para cada elemento tenham sido medidas p-variáveis aleatórias. O Método Fuzzy procura a partição que minimiza a função objetivo:

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m d(X_j, V_i)$$

onde V_i é o protótipo (semente ou centróide ponderado) do conglemerado *i*, *i=1,...,c*;
m>1 é o parâmetro Fuzzy e quanto mais alto for, mais difuso será o cluster no final;
u_{ij} é a probabilidade de que o elemento X_j pertença ao conglomerado cujo protótipo é V_i;
d(.) é a distância escolhida pelo pesquisador (em geral é a distância euclidiana).

Medotologia

A função *J* é minimizada quando as probabilidades u_{ij} são escolhidas pela expressão:

$$u_{ij} = \left[\sum_{i=1}^{c} \left(\frac{d(X_j, V_i)}{d(X_j, V_k)} \right)^{2/(m-1)} \right]^{-1}$$

onde,

$$V_i = \frac{\sum_{j=1}^{n} (u_{ij})^m X_j}{\sum_{j=1}^{n} (u_{ij})^m}$$

para todo *i=1,2,...,c* e *j=1,2,...,n*. Os protótipos e as probabilidades iniciais são gerados de uma distribuição *U(0,1)*. Os protótipos vão se modificando a cada iteração, e o algoritmo é interrompido quando a distância entre os protótipos de uma determinada iteração em relação a anterior é menor ou igual a um certo valor de erro preestabelecido.

Exemplo

A partir de dados de desenvolvimento de 21 países obtidos do banco de dados da ONU (2002), e disponível em http://hdr.undp.org/, deseja-se agrupá-los segundo o perfil dos mesmos. As variáveis utilizadas (segundo índices) são: Expectativa de vida, Educação, Renda (PIB) e Estabilidade política.

Resultado Fuzzy c-Means

P1	P2	P3	P4	País	Cluster
0.87	0.08	0.02	0.03	Reino Unido	1
0.90	0.05	0.02	0.02	Austrália	1
0.92	0.04	0.01	0.02	Canadá	1
0.91	0.05	0.02	0.02	Estados Unidos	1
0.91	0.05	0.02	0.02	Japão	1
0.80	0.12	0.03	0.04	França	1
0.74	0.14	0.05	0.07	Cingapura	1
0.19	0.66	0.06	0.09	Argentina	2
0.68	0.20	0.05	0.07	Uruguai	1
0.13	0.59	0.11	0.17	Cuba	2
0.07	0.12	0.56	0.25	Colômbia	3
0.09	0.81	0.04	0.06	Brasil	2
0.09	0.16	0.36	0.39	Paraguai	4
0.11	0.66	0.08	0.15	Egito	2
0.03	0.05	0.78	0.14	Nigéria	3
0.03	0.06	0.11	0.79	Senegal	4
0.05	0.09	0.58	0.28	Serra Leoa	3
0.08	0.11	0.59	0.22	Angola	3
0.04	0.08	0.12	0.75	Etiópia	4
0.16	0.36	0.16	0.32	Moçambique	2
0.07	0.84	0.04	0.06	China	2

Média das variáveis por grupo

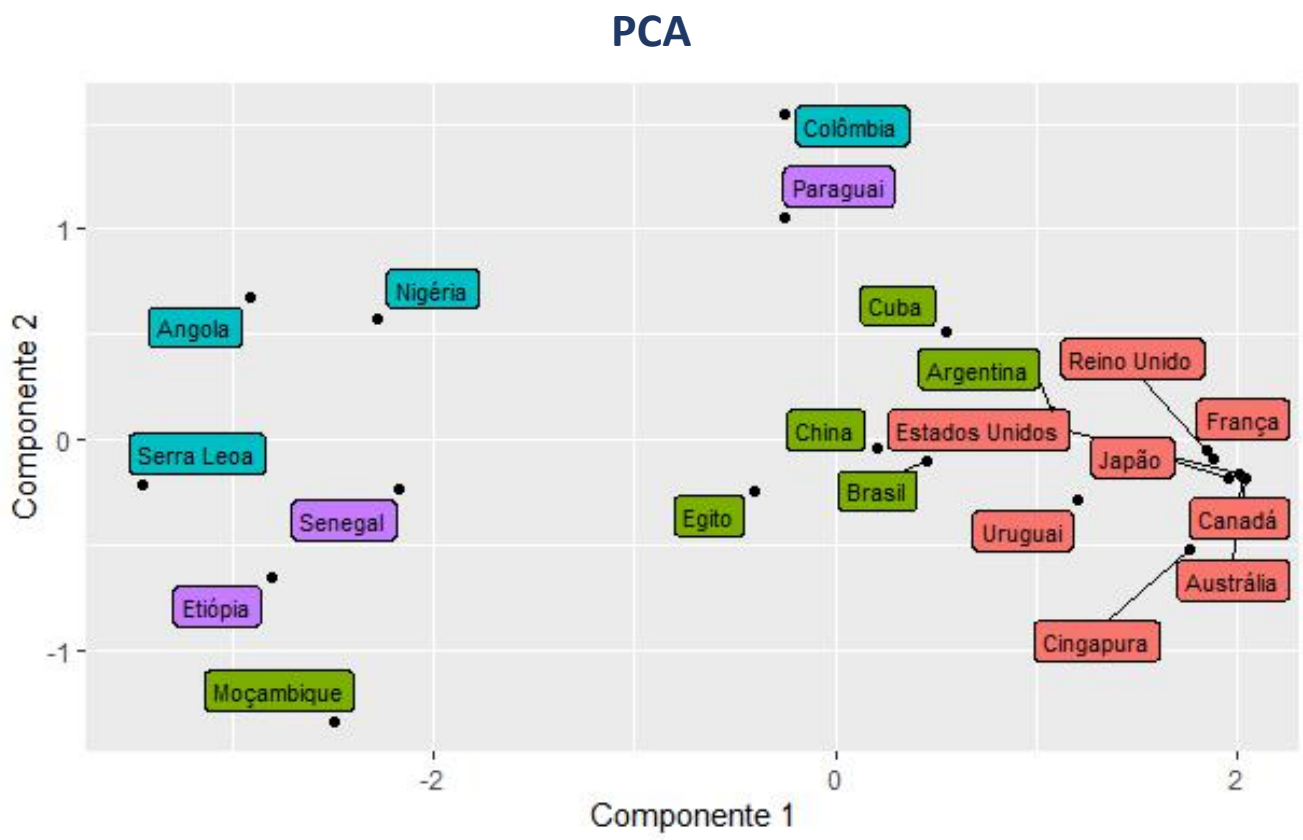
Índices/Grupos	1	2	3	4
Índice de Expectativa de Vida	0.884	0.678	0.445	0.510
Índice de Educação	0.954	0.740	0.530	0.517
Índice PIB	0.908	0.622	0.460	0.467
Estabilidade Política e Violência	1.185	0.315	-1.490	-0.700

A técnica de componentes principais foi utilizada em complemento ao agrupamento obtido de forma a visualizarmos os clusters formados.

Obs: A definição do número de clusters e a qualidade do agrupamento foram omitidas, podendo serem avaliadas da maneira usual.

Exemplo

97% da variabilidade dos dados pode ser explicada pelas duas primeiras componentes. As probabilidades associadas principalmente a Moçambique e ao Paraguai, sugerem que esses países compartilham de caraterísticas dos outros grupos formados.



Medotologia

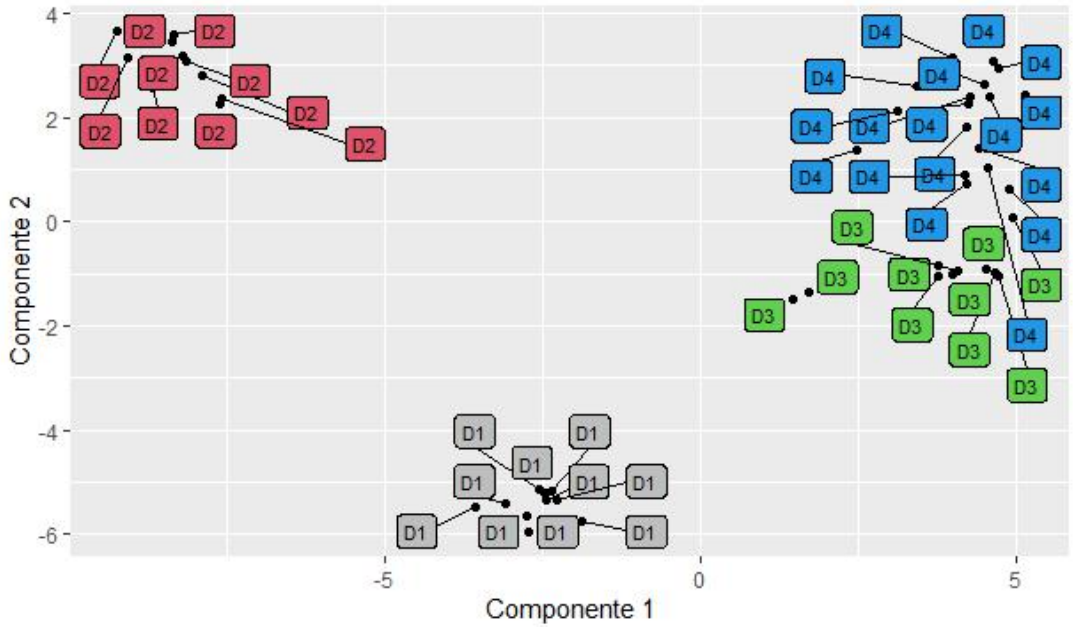
No caso de dados categóricos, há o Fuzzy k-Modes (Huang e Ng, 1999), muito similar ao caso de variáveis contínuas (troca-se médias por modas e distância por similaridade). Outra alternativa, seria de ao invés de passar para o algoritmo a matriz de dados originais, já repassar uma matriz de similaridade/dissimilaridade.

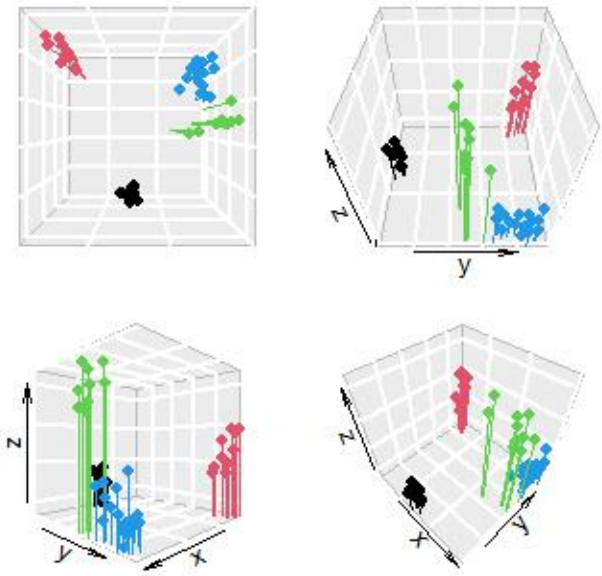
Exemplo: O banco de dados deste exemplo dados encontra-se disponível em http://archive.ics.uci.edu/ml/index.php e foi alterado para fins didáticos. Os dados correspondem a 21 variáveis, e cada atributo pode conter de 2 a 7 categorias. Os atributos representam características das plantas e suas condições ambientais. As plantas de grão de soja são naturalmente divididas em 4 grupos segundo o tipo de doença que apresentam. A distribuição dos dados pode ser vista em Matos (2007). Deseja-se avaliar se o tipo de doença está associada aos perfis de plantas.

Exemplo

77% da variabilidade dos dados é explicada com 2 compontes e 86% da variabilidade dos dados é explicada com até 3 compontes. 100% das sementes foram alocadas corretamente aos seus respectivos clusters. Os grupos 3 e 4 apresentam alguma incerteza quanto a classificação nesses dois grupos.

Visualização dos Clusters





Medotologia

O método ROCK (Robust Clustering Using Links), Guha et. al. (2000), é um método hierárquico aglomerativo e fundamentado na noção de “links”. O número de links entre dois elementos representa o número de vizinhos que eles tem em comum, com base em um nível de similaridade pré-especificado. Quanto maior o número de links, maior a chance de que 2 elementos pertençam ao mesmo cluster. O Propósito é maximizar a soma dos links entre as observações do mesmo grupo e minimizar a soma dos links entre observações de grupos diferentes.

Inicialmente cada elemento constitui um grupo isolado. O número de links entre todos os elementos são calculados e os mesmos vão sendo unidos até que o número desejado de grupos é obtido ou não haja mais links entre os grupos. Esse processo é iterativo e em cada passo do algoritmo; a medida de adequabilidade do ajuste (“bondade de ajuste”) é calculada e os clusters que maximizam esse medida são unidos.

$$g(G_j, G_l) = \frac{link(G_j, G_l)}{(n_j + n_l)^{1+2f(\theta)} - n_j^{1+2f(\theta)} - n_l^{1+2f(\theta)}}$$

Medotologia

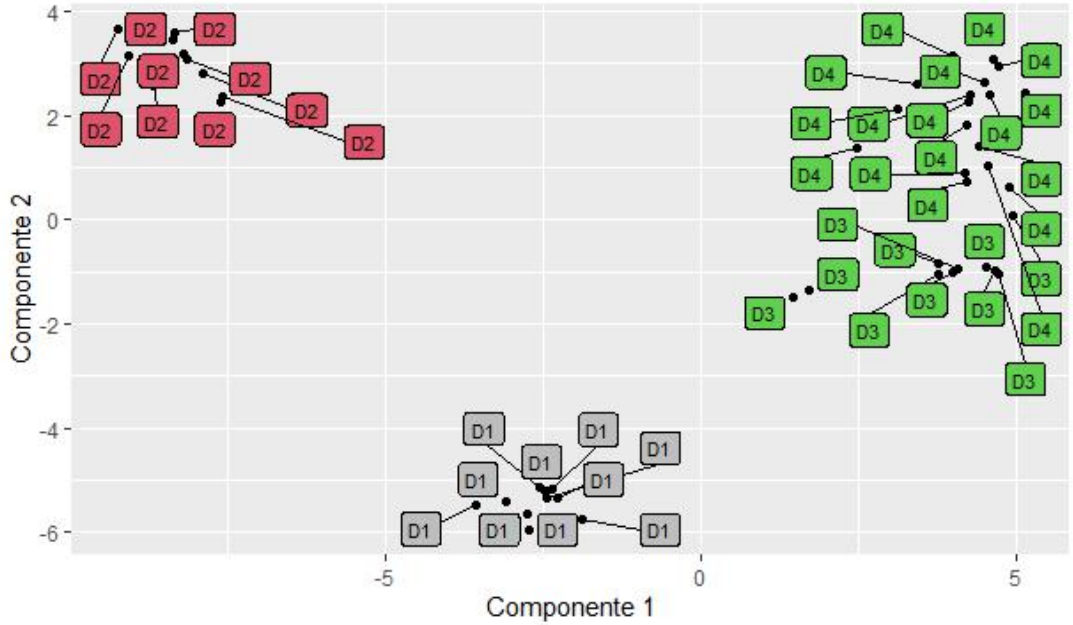
$$link(G_j, G_l) = \sum_{X_q \in G_j, X_r \in G_l} link(X_q, X_r)$$

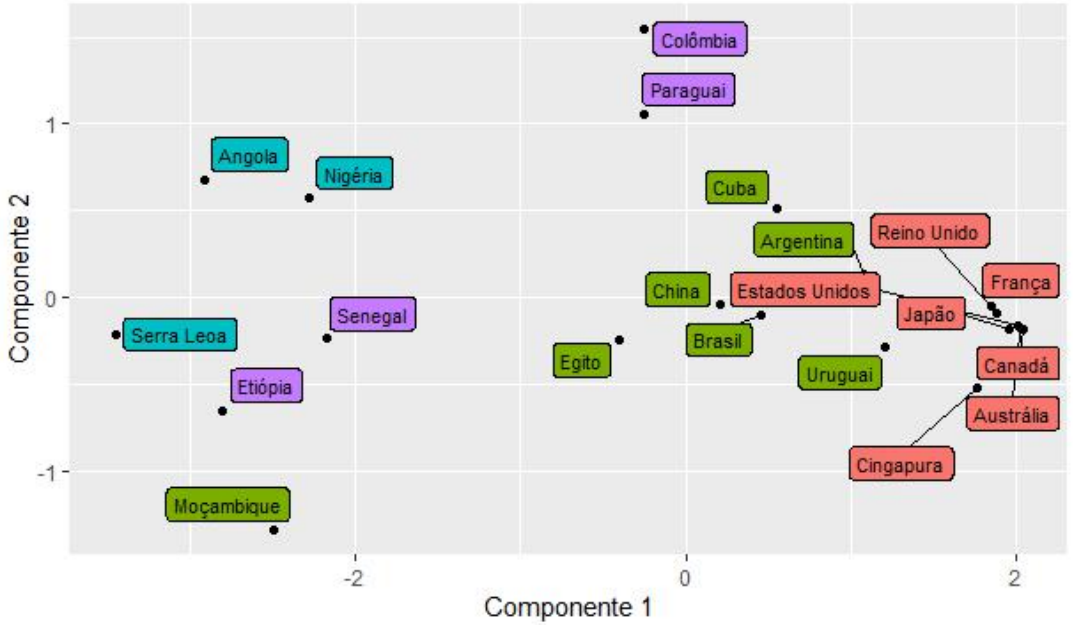
X_q *e* X_r são duas observações, θ é o valor mínimo de similaridade para que duas observações sejam consideradas vizinhas, G_j e G_l denotam os clusters comparados, n_j e n_l os tamanhos amostrais desses clusters e $f(\theta)$ é uma função pré-especificada e uma sugestão de Guha et. al. (2000).

O método ROCK pode ser utilizada tanto para dados de natureza numérica ou categóricos.

Exemplo

Visualização dos Clusters Formados pelo Método ROCK nos Exemplos Anteriores





Medotologia

Os métodos anteriormente mencionados são não paramétricos já que independem da distribuição de probabilidade dos dados. Sendo conhecida a distribuição de probabilidades dos elementos dos k grupos, pode-se obter um distribuição de probabilidade conjunta de forma a determinar-se a melhor partição dos dados. Esses métodos são mais complexos e não necessariamente melhores, pois a violação da suposição inicial pode comprometer totalmente qualquer inferência.

Conclusão

Embora os resultados de métodos como o K-means e K-modes tenham sido omitidos, o material da professora Sueli Mingoti (DEST-UFMG), fazia uma comparação de desempenho entre esses métodos e os algoritmos aqui apresentados.

- Para o exemplo com dados dos índices de desenvolvimento dos países, o K-means e o método Fuzzy produziram a mesma partição, e o método ROCK produziu uma participação ligeiramente diferente.

- Para os dados do exemplo do grão de soja, o método ROCK não foi capaz de separar os clusters 3 e 4, no entanto eles realmente são muito semelhantes e, se não fosse conhecido antecipamente o número de clusters, análises sugerem que adotar 3 clusters seria uma solução possível. O método Fuzzy teve desenpenho tão bom quanto o K-modes.

- Os algoritmos apresentados obtiveram desempenho satisfatório e são opções viáveis para problemas de agrupamento.

Referências

[1] Notas de aula da professora Sueli Aparecida Mingoti (DEST-UFMG).
[2] Mingoti, S. A - Análise de Dados Através de Métodos de Estatística Multivariada, Editora UFMG, 2013.