Clustering de dados de taxa da fotossíntese usando k-means clustering e clustering Hierárquico

Reconhecimento de Padrões

Leissi Castañeda León, Jorge Guevara Díaz

Instituto de Matemática e Estatística
Universidade de São Paulo,São Paulo-Brasil
leissicl@vision.ime.usp.br, jorjasso@vision.ime.usp.br

Abstract

Este trabalho descreve o clustering usando o algoritmo de k-means e o algoritmo de linkage aplicados a dois conjuntos de dados da taxa de fotossíntese de cana-de-açúcar ao longo de 14 instantes de tempo em duas situações: ambiente e saturada com excesso de CO2.

General Terms

1. Introdução

Este artigo descreve experimentos sobre dois conjuntos de dados. O primer conjunto de dados [1] é un conjunto que não é fácil de fazer clustering, chamaremos a este conjunto de dados *CI*. Para este conjunto foram usadas alguns algoritmos para transformar os dados como PCA, PCA-nolinear e 08 normalizacoes diferentes dos dados. O segundo conjunto de dados [2] é un conjunto onde os resultados obtidos por os algoritmos de clustering deram bons resultados, o qual indica que foi mais fácil o fazer clustering, chamaremos a este conjunto *CII*. As técnicas de clustering usadas foram o algoritmo k-means, e cluster hierárquico.

O projecto foi iniciado pelo professor pelo professor Marcos Buckeridge da USP. O trabalho é baseado nos dados proporcionados por ele e sua equipe de trabalho.

O informe esta organizado da seguinte maneira: A Secção 2 descreve os dados a usar e a representação que usaremos neste artigo, descreve também a etapa de normalização feita sobre os dois conjuntos de dados. A Secção 3 descreve o análisis de componentes principais dos conjuntos de dados CI e CII usando os algoritmos PCA e algoritmo PCA-nolinear . A Secção 4 descreve os algoritmos de clustering k-means e clustering hierárquico usados. A Secção 5 descreve os resultados obtidos. A Secção 6 descreve a discussão dos resultados e as conclusões obtidas.

[1] "O primer conjunto de dados que foi disponibilizado no PACA o 19 de Novembro do 2010"

[2] "O novo conjunto de dados disponibilizado no PACA o 6 de Dezembro do 2010"

2. Os dados

Os dados representam a taxa de fotossíntese de cana-de-açúcar ao longo de 14 instantes de tempo em duas situações: ambiente (classe 1) e saturada (classe 2, com excesso de CO2).

Seja $X_i = x_1, ..., x_t, ...x_{14}$, a amostra i que contém os dados nos $0 \leq t \leq 14$ instantes de tempo. Logo X_{it} representa a informação associada a amostra número i no instante de tempo t (valor da variável t para a observação i).

A visualização (scatterplot) dos dados do conjunto Ci e do conjunto CII é mostrada nas figuras as gráficas das variáveis X_{it} vs X_{iq} para todas as amostras mostrando a relação entre duas variáveis a maneira de um gráfico bidimensional . As figuras mostram como o conjunto de dados CI é más difícil de fazer clustering que o conjunto de dados CII.

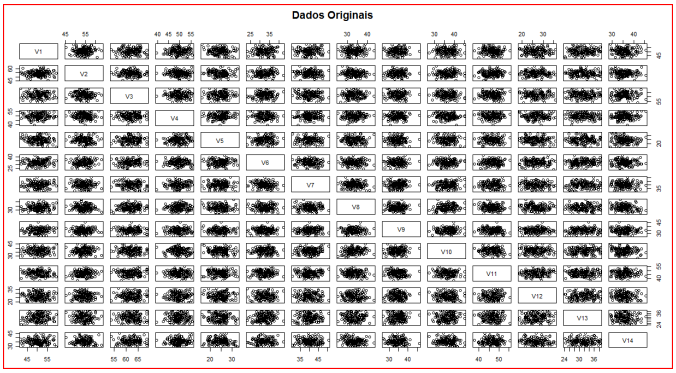

Figure 1. Scatterplot do conjunto de dados CI. O gráfico mostra pouca separabilidade das variáveis.

2.1 Normalização dos dados

A normalização dos dados multivariados é uma etapa importante para a determinação das distancias en clustering [3]. O objectivo da normalização é ajustar o tamanho (magnitude) e o peso relativo das variáveis de entrada. Foi considerado fazer uma normalização dos dados antes do aplicar os algoritmos de clustering, considerando as

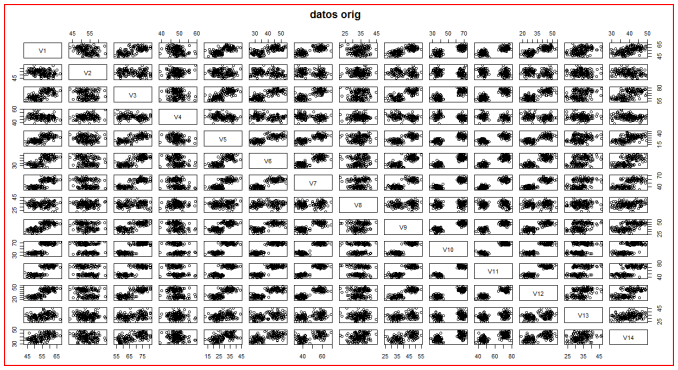

Figure 2. Scatterplot do conjunto de dados CII. O gráfico mostra que as varáveis tem boa separabilidade

normalizações descritas em [3]:

$$
\begin{aligned}
\text{Sim normalizar } (n_0),\ z_{it} &= x_{it} \\
\text{Normalização } (n_1),\ z_{it} &= \frac{x_{it}-u_t}{\sigma_t} \\
\text{Unificação } (n_2),\ z_{it} &= \frac{x_{it}-u_t}{\max_i(X_{it})-\min_i(X_{it})} \\
\text{Unificação mínimo}(n_3),\ z_{it} &= \frac{x_{it}-\min_i(X_{it})}{\max_i(X_{it})-\min_i(X_{it})} \\
\text{Normalização intervalo } (n_4),\ z_{it} &= \frac{x_{it}-u_t}{\max_i|x_{it}-u_t|} \\
\text{Transformação cociente } (n_5),\ z_{it} &= \frac{x_{it}}{\sigma_t} \\
\text{Transformação cociente } (n_6),\ z_{it} &= \frac{x_{it}}{\max_i(X_{it})-\min_i(X_{it})} \\
\text{Transformação cociente } (n_7),\ z_{it} &= \frac{x_{it}}{\max_i(X_{it})} \\
\text{Transformação cociente } (n_8),\ z_{it} &= \frac{x_{it}}{u_t}
\end{aligned}
$$

Onde z_{it} é a observação normalizada de x_{it}, μ_t e σ_t é a media e a desviação estándar da variavel t.

Foi feita também uma normalizacao Hellinguer HT [8].

3. Transformação a componentes Principais

A normalização dos dados anteriores produce invariância a translação e escalamento mais não a rotação. para obter invariância a rotação deve-se rotar os eixos de coordenadas de tal maneira que eles coincidam com os autovetores da matriz de covariância [1]. É dizer uma transformação aos componentes principais. Neste artigo consideramos fazer experimentos com analise de componentes principais (PCA) e análisis de componentes no linear (NLPCA).

3.1 Analise de Componentes Principais-PCA

Conhecido como Karhunen-Loéve transform, o algoritmo a continuação tem como dados de entrada os dados X, a media dos dados μ e a matriz de covarianza dos dados Σ :

ALGORITMO-PCA(X,μ,Σ)

1 subtrair a media dos dados $X-\mu$
2 calcular os autovetores e autovalores λ_i, e_i
3 ordenar e_i pelo autovalor λ_i e construir a matriz A
4 projetar os dados transformados usando a matriz A, mediante
5 $x' = A^t(X-\mu)$

Previamente foi feito uma normalização dos dados mediante $X_i = X_i/||X_i||$.

Depois de aplicar o algoritmo PCA no conjunto de dados CI a importância dos componente principais são as seguintes:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
σ	1.26	1.18	1.12	1.11	1.07	1.04	1.01
σ^2	0.11	0.10	0.09	0.09	0.08	0.08	0.07
$\sum\sigma^2$	0.11	0.21	0.30	0.39	0.47	0.55	0.62
	PC8	PC9	PC10	PC11	PC12	PC13	PC14
σ	0.98	0.94	0.91	0.88	0.83	0.80	0.73
σ^2	0.07	0.06	0.06	0.06	0.05	0.05	0.04
$\sum\sigma^2$	0.69	0.75	0.81	0.87	0.92	0.96	1.00

Onde σ é a desviação estándar, σ^2 é proporção da variância e $\sum\sigma^2$ é proporção de acumulação da variância.

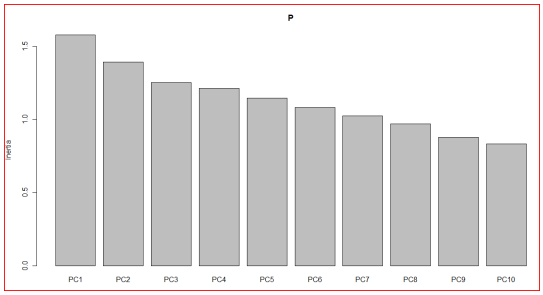

Figure 3. Gráfico da importância dos dez primeiros componentes principais para o conjunto de dados CI.

O analise das componentes principais mostram que a acumulação da variância usando as 12 primeiras componentes pode-se explicar o 92% da variação dos dados, por isso o clustering será feito nos dados CI usando todas as componentes principais.

No caso do conjunto de dados CII a importância dos componente principais são as seguintes:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
σ	2.90	1.08	0.98	0.93	0.90	0.69	0.57
σ^2	0.60	0.09	0.07	0.06	0.06	0.03	0.02
$\sum\sigma^2$	0.60	0.68	0.75	0.81	0.87	0.90	0.93
	PC8	PC9	PC10	PC11	PC12	PC13	PC14
σ	0.53	0.43	0.43	0.37	0.34	0.26	0.22232
σ^2	0.02	0.01	0.01	0.01	0.008	0.005	0.004
$\sum\sigma^2$	0.95	0.96	0.97	0.98	0.99	0.10	1.0

Onde σ é a desviação estándar, σ^2 é proporção da variância e $\sum\sigma^2$ é proporção de acumulação da variância.

O analise das componentes principais mostram que a acumulação da variância usando as 5 primeiras componentes pode-se explicar o 87% da variação dos dados, por isso o clustering será feito nos dados CII usando só 5 componentes principais isso faz uma redução da dimensão dos dados do 64.2%.

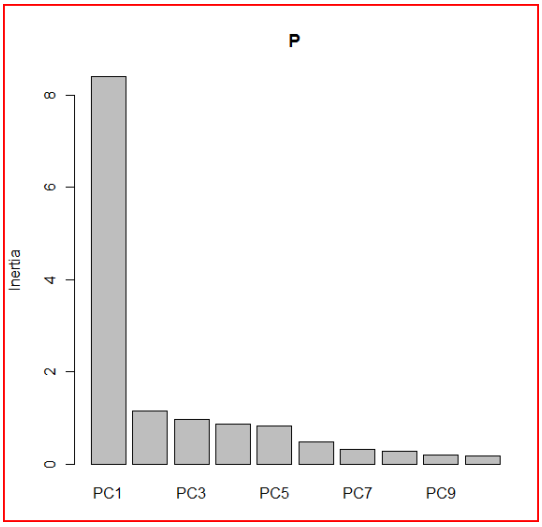

Figure 4. Gráfico da importância dos dez primeiros componentes principais para o conjunto de dados CII.

3.2 Analise de Componentes Não Lineares-NLCA

Se os dados representam complicadas iterações nas características, então o espaço linear pode ser uma pobre representação e é necessário trabalhar com componentes não lineares [1]. Para fazer um PCA-não linear é usado uma red neural do tipo perceptrón multicapa (MLP) com algoritmo de aprendizado Backpropagation. A idea é encontrar uma curva no espacio multidimensional que descreva a maior variância possível. Neste caso a matriz de transformação dos dados é oculta nos pesos da rede neural. Uma desvantagem do método é que é lento.

Este algoritmo foi usado só no conjunto de dados CI

4. Clustering

4.1 Kmeans

O algoritmo do k-means e um algoritmo de análise de clustering, faz a partição dos dados en k grupos de tal maneira que a suma dos quadrados dos pontos a os centros dos clusters é minimizada.

ALGORITMO-K-MEANS

1 **Begin** Inicializar n, k, μ_1,...μ_k
2 **do** Classificar n amostras de acordo com o mais cercano μ_i
3 recalcular μ_i
4 **Until** no existam cambios em μ_i
5 **return** μ_1,...μ_k

Foram usados os algoritmos Hartigan-Wong [5], Lloyd [6], Forgy [4] e MacQueen [7].

4.2 Clustering hierárquico

O algoritmo procede da seguinte maneira: Inicialmente cada amostra e assinada a um cluster, logo o algoritmo procede de maneira iterativa en cada etapa juntando os dois clusters mais similares, continuando até que exista um só cluster

ALGORITMO-CLUSTERING-HIERÁRQUICO

1 **Begin** Inicializar c, $\hat{c} \leftarrow n$, $D_i \leftarrow X_i$,$i = 1, ..., n$
2 **do** $\hat{c} \leftarrow \hat{c} - 1$
3 Procurar os clusters mais cercanos D_i e D_j
4 Mesclar D_i e D_j
5 **Until** $c = \hat{c}$
6 **return** c clusters

Onde c e o número de clusters finales que se quer obter, $\hat{c}$ é o número de clusters iniciales. As distancias entre clusters são computadas usando a fórmula de Lance-Williams. Foram usadas três métodos: ligação simples, ligação completa e ligação ward

Sejam n_r o número do objetos no D_r e x_{ri} o objeto i do D_r, Então:

1. ligação simples: Usa a menor distancia entre os objetos en dois clusters

$$distancia(D_r, D_r) = \min(d(x_{ri}, x_{sj})) \tag{1}$$

para $i \in i, ..., n_r$, $j \in 1, ..., n_s$

2. ligação completa: Usa a maior distancia entre os objetos en dois clusters

$$distancia(D_r, D_r) = \max(d(x_{ri}, x_{sj})) \tag{2}$$

para $i \in i, ..., n_r$, $j \in 1, ..., n_s$

3. ligação ward: Usa a suma incremental dos quadrados, incrementa a suma total interna dos quadrados do cluster como resultado de unir dos clusters e é definida como a suma dos quadrados das distancias entre todos os objetos do cluster e o centroide do cluster e é apropriado só para distancias euclideanas.

$$distancia(D_r, D_r) = \sqrt{\frac{2n_r n_s}{n_r + n_s}} \|\bar{x}_r - \bar{x}_s\|_2 \tag{3}$$

Onde $\|\|_2$ é a distancia euclideana, $\bar{x}_r$ e $\bar{x}_s$ são os centroides dos clusters r e s e n_r e n_s sao os número de elementos nos clusters r e s.

As medidas de distancia exploradas foram:

- Euclideana

$$d(X_i, X_j) = \sqrt{\sum_{t=1}^{T} (x_{it} - x_{jt})^2} \tag{4}$$

- Maximo

$$d(X_i, X_j) = \max_{t=1}^{T} |x_{it} - x_{jt}| \tag{5}$$

- Manhatan

$$d(X_i, X_j) = \sum_{t=1}^{T} |x_{it} - x_{jt}| \tag{6}$$

- Canberra

$$d(X_i, X_j) = \sum_{t=1}^{T} \frac{|x_{it} - x_{jt}|}{|x_{it}| + |x_{jt}|} \tag{7}$$

5. Resultados

Foram feitos quatro experimentos no conjunto de dados CI e CII, onde foram exploradas todas as técnicas de normalização ou de transformação a componentes principais para depois tentar clusterizar os dados com o algoritmo k-means e com clustering hierárquico.

Experimento I

A seguinte tabela mostra os resultados do experimento I: Aplicar clustering hierárquico variando as funções distancia: Euclidiana, Máximo, Manhatan e Canberra e com as técnicas de normalização $n_0 ... n_{n_8}$, da mesma maneira o análise de componentes principais PCA e NLCA no conjunto de dados CI. No caso do PCA foram consideradas: PCA-5 , PCA-10 e PCA-14 análise com 5,10 e 14 componentes principais respectivamente. A tabela mostra de cor amarelo os menores valores obtidos por função distancia.

		Euclid.	Maximo	Manhatan	Canberra
n_0	Complete	52.14	51.43	44.29	43.57
	Single	49.29	49.29	49.29	49.29
	Ward	47.14	-	-	-
n_1	Complete	39.29	42.14	45	46.42
	Single	49.29	49.29	49.29	49.29
	Ward	50	-	-	-
n_2	Complete	57.14	45	46.43	46.43
	Single	49.29	49.29	49.29	49.29
	Ward	54.29	-	-	-
n_3	Complete	57.14	45	46.43	44.29
	Single	49.29	49.29	49.29	49.29
	Ward—	54.29	-	-	-
n_4	Complete	52.14	47.86	47.86	46.43
	Single	49.29	49.29	49.29	49.29
	Ward	47.14	-	-	-
n_5	Complete	39.29	42.14	45	43.57
	Single	49.29	49.29	49.29	49.29
	Ward	50	-	-	-
n_6	Complete	57.14	53.57	46.43	43.57
	Single	49.29	49.29	49.29	49.29
	Ward	54.29	-	-	-
n_7	Complete	45.71	47.14	53.57	43.57
	Single	49.29	49.29	49.29	49.29
	Ward	56.43	-	-	-
n_8	Complete	47.14	46.43	42.86	43.57
	Single	49.29	49.29	49.29	49.29
	Ward	50.71	-	-	-
PCA 5	Complete	46.43	49.29	47.86	52.86
	Single	49.29	49.29	49.29	49.29
	Ward	52.86	-	-	-
PCA 10	Complete	48.57	49.29	50	60.71
	Single	49.29	49.29	49.29	49.29
	Ward	51.43	-	-	-
PCA 14	Complete	39.29	49.29	52.86	52.14
	Single	49.29	49.29	50.71	50.71
	Ward	50	-	-	-
NLCA	Complete	47.86	44.29	48.57	52.86
	Single	49.29	49.29	49.29	49.29
	Ward	50	-	-	-
HT	Complete	49.29	46.43	50.71	45.71
	Single	50.71	50.71	49.29	49.29
	Ward	44.29	-	-	-

Experimento II

A seguinte tabela mostra os resultados do experimento II: Aplicar clustering k-means nos 4 variantes do algoritmo : Hartigan-Wong, Lloyd, Forgy e MacQueen no conjunto de dados CI. A tabela mostra de cor amarelo os menores valores obtidos por cada variante do algoritmo.

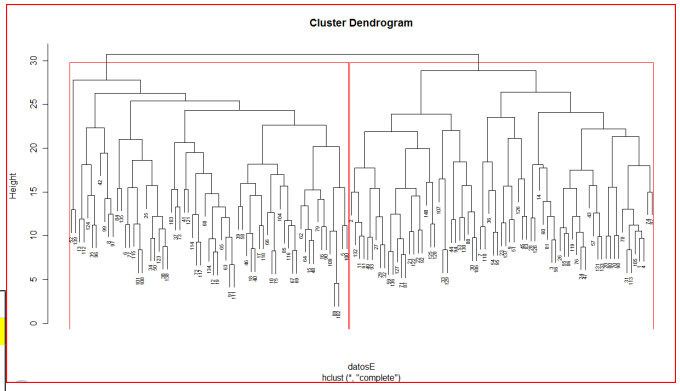

Figure 5. Clustering hierárquico con linkage complete para o conjunto de dados CII com normalizacao n_0.

	Hartigan-Wong	Lloyd	Forgy	MacQueen
n_0	44.29	53.57	47.86	48.57
n_1	45	49.29	55	54.29
n_2	51.43	50.71	50.71	47.14
n_3	51.43	50.71	50.71	50
n_4	51.43	49.29	48.57	50.71
n_5	55	53.57	48.57	52.14
n_6	51.43	50	48.57	47.86
n_7	53.57	52.86	52.86	48.57
n_8	47.14	52.14	51.43	46.43
PCA 5	47.86	49.29	53.57	49.29
PCA 10	47.14	51.43	54.29	51.43
PCA 14	55	52.86	55	47.86
NLCA	56.43	50.71	50	50.71
HT	57.14	55.71	55.71	60

Experimento III

A seguinte tabela mostra os resultados do experimento III : Aplicar clustering hierárquico variando as funções distancia: Euclidiana, Máximo, Manhatan e Canberra e com as técnicas de normalização $n_0 ... n_{n_8}$, da mesma maneira o análise de componentes principais PCA e NLCA no conjunto de dados CII. No caso do PCA foram consideradas: PCA-5 análise com 5 componentes principais pois como as 5 componentes descrevem o 87% da variação dos dados respectivamente, fazendo uma redução da dimensão dos dados do 64.2%.

		Euclid.	Maximo	Manhatan	Canberra
n_0	Complete	0.00	0.00	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-
n_1	Complete	0.00	77.14	0.00	0.00
	Single	0.00	50.71	0.00	0.00
	Ward	0.00	-	-	-
n_2	Complete	0.00	34.29	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-
n_3	Complete	0.00	34.29	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-
n_4	Complete	0.00	0.00	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-
n_5	Complete	0.00	77.14	0.00	0.00
	Single	0.00	50.71	0.00	0.00
	Ward	0.00	-	-	-
n_6	Complete	0.00	12.14	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-
n_7	Complete	0.00	0.00	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-
n_8	Complete	0.00	0.00	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-
PCA 5	Complete	0.00	0.00	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-
NLCA	Complete	40	52.14	60.71	49.29
	Single	49.29	49.29	50.71	49.29
	Ward	13.57	-	-	-
HT	Complete	0.00	0.00	0.00	0.00
	Single	0.00	0.00	0.00	0.00
	Ward	0.00	-	-	-

Experimento IV

A seguinte tabela mostra os resultados do experimento III: Aplicar clustering k-means nos 4 variantes do algoritmo : Hartigan-Wong, Lloyd, Forgy e MacQueen no conjunto de dados CII.

	Hartigan-Wong	Lloyd	Forgy	MacQueen
n_0	0	0	0	0
n_1	0	0	0	0
n_2	0	0	0	0
n_3	0	0	0	0
n_4	0	0	0	0
n_5	0	0	0	0
n_6	0	0	0	0
n_7	0	0	0	0
n_8	0	0	0	0
PCA 5	0	0	0	0
NLCA	0	0	0	0
HT	0	0	0	0

6. Discussão

No caso do experimento I a menor erro e de 39.29% nos casos de fazer a normalização n_1, n_5 e usar PCA con todas as componentes principais. Isto é uma melhora considerable pois considerando as taxas de erro obtidas que são aproximadamente do 50%.

No caso do experimento II a menor taxa de erro que se tem é de 44.29% para n_0, pode-se notar também algo de melhora no caso do PCA.

No caso do experimento III, o clustering foi satisfatório com 0% de erro no caso de clsutering herárquico. Só a distancia máximo tem un comportamento ruim. No caso do uso do NLCA não se tem melhora em nenhum dos experimentos, no caso do experimento III o NLCA tem resultados muito ruins.

Note que no se tem resultados para as distancias Máximo, Manhatan e Camberra para o linkage ward, pois como foi visto na Secção 4.2 ward tem sentido quando se usa a distancia euclidiana. No experimento IV se tem resultados satisfactorios en todos os casos. Uma conclusão importante é que o uso de técnicas de normalização tanto a translação e escala como as mencionadas na Secção 2.1 e invariantes a rotação como o análise de componentes principais 3 são importantes antes de fazer o clustering pois permitem diminuir a taxa de erro. No caso das componentes principais é possível diminuir a dimensão da data de una maneira considerável . Outra conclusão é que é importante decidir qual distancia se vai usar no caso de clustering hierárquico, pois ten dependência na taxa de erro do clustering.

Como adjunto tem o arquivo *Detallhe dos resultados.xlsx* onde se descreve os experimentos feitos assim como o número de dados obtidos por cluster.

References

[1] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern classification. Wiley, 2 edition, November 2001.

[2] Broughton, S., Allen;Bryan, Kurt, M. Discrete Fourier analysis and wavelets. Applications to signal and image processing. (2009)

[3] Jajuga, K., Walesiak, M. (2000), Standardisation of data set under different measurement scales, In: R. Decker, W. Gaul (Eds.), Classification and information processing at the turn of the mil-lennium, Springer-Verlag, Berlin, Heidelberg, 105-112.

[4] Forgy, E. Cluster analysis of multivariate data: efficiency versus interpretability of classifications Biometrics, 1965, 21, 768-780

[5] Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100-108

[6] Lloyd, S. P. (1957, 1982) Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory 28, 128-137

[7] Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam J. Neyman, 1, pp. 281-297. Berkeley, CA

[8] Ecologically Meaningful Transformations For Ordination Of Species Data