AGRUPAMENTO EM SÉRIES TEMPORAIS UTILIZANDO O MODELO DE ALOCAÇÃO LATENTE DE DIRICHLET

Murilo C. O. Camargos Filho*

*Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil

Email: murilocamargos@ufmg.com

Abstract— In this paper the clustering problem was addressed with robustness to the initial choice of the number of clusters. The data was analyzed from a time series perspective in order to apply a technique known as Latent Dirichlet Allocation. Monte Carlo Markov chains simulations were done to estimate this model's distribuitions parameters. Two experiments were carried out, the first is the clustering of a time series with a mean change, the second is the clustering of the known Iris flowers dataset. The method behaved efficiently in identifying the correct number of clusters even when asked for find more clusters and in hitting **100%** accuracy in the Iris flower dataset.

Keywords— Clustering, Latent Dirichlet allocation, Markov chains, Time series.

Resumo— Neste trabalho foi considerado o problema de agrupamento de dados com robustez à estimativa inicial da quantidade de grupos existentes. Para isso, os dados foram analisados sob o a perspectiva de séries temporais visando a aplicação de uma técnica conhecida por Alocação Latente de Dirichlet. Simulações de Monte Carlo via cadeias de Markov foram utilizadas para estimação dos parâmetros das distribuições que compõem este modelo. Dois experimentos são realizados, o primeiro corresponde à uma série temporal univariada com mudança de média e o segundo à base de dados da flor de Iris. O método se comportou de forma eficiente ao identificar corretamente o número real de grupos mesmo quando solicitado um número maior e atingindo uma acurácia de **100%** na base de dados da flor de Iris.

Palavras-chave— Agrupamento, Alocação latente de Dirichlet, Cadeias de Markov, Séries temporais

1 Introdução

Em geral, os métodos de agrupamento visam subdividir um conjunto de dados $\boldsymbol{X}$ em k subconjuntos disjuntos de forma que $\boldsymbol{X}$ possa ser obtidos a partir da união dos c subconjuntos (Bezdek et al., 1984). O agrupamento pode ser feito de acordo com algum modelo paramétrico, como no algoritmo k-means (MacQueen et al., 1967) ou utilizando alguma medida de distância ou similaridade entre amostras, como no caso de algoritmos de agrupamento hierárquicos (Ben-Hur et al., 2001).

Além de métodos de agrupamento como o k-means, que atribuem cada amostra a um determinado grupo, os métodos *fuzzy* também são utilizados em diversas áreas como astronomia, geologia, imagens médicas, reconhecimento de alvos, segmentação de imagens, etc (Chuang et al., 2006). Nessa classe de algoritmos, as amostras pertencem à todos os grupos com diferentes graus de pertinência, permitindo capturar incertezas em aplicações reais (Coutinho e das Chagas, 2017).

Uma característica que tanto o FCM quanto o k-means compartilham é a necessidade de a quantidade de grupos ser conhecida *à priori*. Além disso, os dois algoritmos assumem que essa quantidade é fixa e imutável. De fato, a escolha de um número adequado de grupos é um problema difícil (Celeux e Soromenho, 1996). Um método de se resolver esse problema num contexto probabilístico é determinar o número de componentes numa mistura de distribuições, como em (Windham e Cutler, 1992) e (Wolfe, 1970).

Este trabalho tem como objetivo apresentar um método para agrupamento de séries temporais uni e multivariadas que seja robusta à estimativa inicial da quantidade de grupos, mantendo apenas os grupos identificados através dos dados. Para isso, uma variação do modelo de Alocação Latente de Dirichlet (LDA, do inglês, *Latent Dirichlet Allocation*) será usada no contexto de séries temporais.

O restante do trabalho está estruturado da seguinte forma: a Seção 2 descreve a variação do modelo LDA utilizada neste trabalho, o método para agrupamento de séries temporais é apresentado na Seção 3, os resultados da aplicação desse método num exemplo numérico e numa base de dados altamente conhecida são mostrados na Seção 4. A Seção 5 traz algumas discussões importantes para a interpretação dos resultados e a Seção 6 conclui o trabalho.

2 Preliminares

O modelo LDA é um modelo probabilístico generativo, proposto por (Blei et al., 2003), para coleções de dados discretos. O LDA é um modelo comumente utilizado para separação de documentos textuais em um número fixo de tópicos com base na ocorrência de determinadas palavras nesses documentos. Como mostrado pela Figura 1, o LDA é um modelo Bayesiano hierárquico de três níveis.

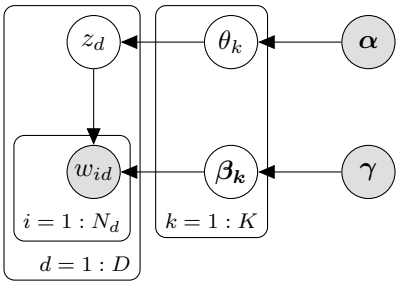

Figura 1: Modelo gráfico do LDA.

O modelo LDA faz parte do grupo de modelos de aprendizado não supervisionado, requerendo o mínimo de conhecimento *à priori* a respeito dos tópicos do corpo de documentos; apenas a quantidade de tópicos (K) é necessária ser informada nesta versão do modelo. Além disso, a coleção de documentos é inteiramente observável.

Dessa forma, temos que D é o número de documentos; N_d é o número de palavras do d-ésimo documento; $C = \{p_1, p_2, \ldots, p_m\}$ é o dicionário com as M palavas existentes nesses documentos; $w_{id} \in C$ é a i-ésima palavra do d-ésimo documento; $\boldsymbol{\gamma} \in \mathbb{R}^M$ é o parâmetro inicial da distribuição da proporção de palavras nos K tópicos; $\boldsymbol{\alpha} \in \mathbb{R}^K$ corresponde ao parâmetro inicial da distribuição da proporção dos tópicos entre os D documentos.

O modelo LDA busca estimar a probabilidade de um documento pertencer a um determinado tópico através da variável aleatória $\boldsymbol{\theta} \in \mathbb{R}^K$ e a probabilidade da ocorrência de cada palavra num determinado tópico através da variável aleatória $\boldsymbol{\beta_k} \in \mathbb{R}^M$. Ambas as variáveis aleatórias possuem distribuições de Dirichlet (1) com os parâmetros $\boldsymbol{\alpha}$ e $\boldsymbol{\gamma}$, respectivamente. Além disso, deseja-se atribuir a cada documento um tópico $z_d \in \{1, 2, \ldots, K\}$ através de uma distribuição multinomial com parâmetros $\boldsymbol{\theta}$.

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \qquad (1)$$

Dados os parâmetros $\boldsymbol{\alpha}$ e $\boldsymbol{\gamma}$, a distribuição conjunta da mistura de tópicos $\boldsymbol{\theta}$, do conjunto de tópicos $\boldsymbol{z}$, do conjunto de palavras $\boldsymbol{w}$ e da mistura de palavras é dada por:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{W}|\boldsymbol{\alpha}, \boldsymbol{\gamma}) \propto \prod_{k=1}^{K} p(\boldsymbol{\beta_k}|\boldsymbol{\gamma})\, p(\theta_k|\boldsymbol{\alpha}) \prod_{d=1}^{D} p(z_d|\boldsymbol{\theta}) \prod_{i=1}^{N_d} p(w_{id}|\boldsymbol{\beta}, z_d) \qquad (2)$$

O problema a ser resolvido é calcular a distribuição *à posteriori* das variáveis latentes, dado o corpo de documentos:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{W}|\boldsymbol{\alpha}, \boldsymbol{\gamma})}{p(\boldsymbol{W}|\boldsymbol{\alpha}, \boldsymbol{\gamma})} \qquad (3)$$

Segundo (Dickey, 1983), calcular o denominador de (3) é um problema intratável. Sendo assim, um método de aproximação precisa ser utilizado, um exemplo é o amostrador de Gibbs, uma forma do método de Monte Carlo via cadeias de Markov (Jordan et al., 1999).

3 Método proposto

O método de agrupamento baseado no modelo LDA inicia-se no pré-processamento da série temporal $\boldsymbol{X}$. Os dados são transformados num conjunto discreto de variáveis linguísticas por meio da fuzzificação. A Figura 2 representa o processo de fuzzificação com 5 formas triangulares, ou seja, o sinal fuzzificado terá 5 possíveis valores linguísticos que serão escolhidos com base no valor da função pertinência $\mu(x)$ de cada uma das formas definidas.

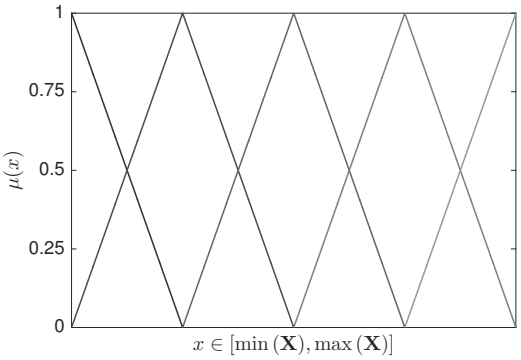

Figura 2: Formas fuzzy triangulares igualmente espaçadas.

Após cada valor da série temporal ter sido fuzzificada, o sinal será dividido em janelas de tamanho fixo W sem sobreposição, como mostrado na Figura 3.

Fazendo um paralelo com o modelo LDA, neste método, cada janela corresponderá a uma documento e cada dado fuzzy corresponderá a uma palavra de um dos documentos. O grupo do documento é, naturalmente, seu tópico. Como mencionado anteriormente, será utilizado o amostrador de Gibbs como método de aproximação para as distribuições das variáveis latentes.

3.1 Amostrador de Gibbs para o LDA

Para utilizar o amostrador de Gibbs, é necessário derivar as condicionais completas para cada variável latente. Dessa forma, usando (2), a condicional completa de $\boldsymbol{\beta_k}$ é dada por:

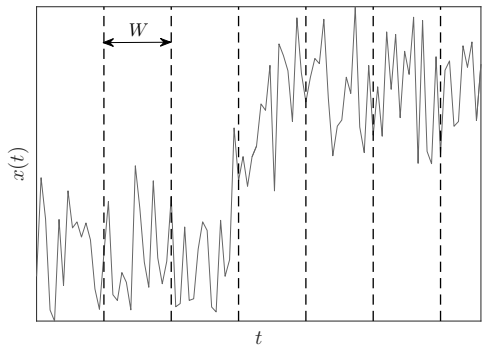

Figura 3: Sinal $x(t)$ com janelas de tamanho fixo W sem sobreposição.

$$p(\boldsymbol{\beta_k}|\boldsymbol{\beta_{-k}}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \propto p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{W}|\boldsymbol{\alpha}, \boldsymbol{\gamma})$$

$$\begin{aligned}
&\propto p(\boldsymbol{\beta_k}|\boldsymbol{\gamma}) \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(w_{id} = m|\boldsymbol{\beta_k}, z_d = k) \\
&\propto \text{Dir}(\boldsymbol{\beta_k}|\boldsymbol{\gamma}) \prod_{d=1}^{D} \prod_{i=1}^{N_d} \text{Mult}(w_{id} = m|\boldsymbol{\beta_{z_d=k}}) \\
&\propto \boldsymbol{\beta_k}^{\gamma-1} \prod_{d=1}^{D} \prod_{i=1}^{N_d} \boldsymbol{\beta}^{\mathbb{I}(w_{id}=m, z_d=k)} \\
&\propto \boldsymbol{\beta_k}^{\boldsymbol{\gamma}-1+\sum_d \sum_i \mathbb{I}(w_{id}=m, z_d=k)} \\
&\sim \text{Dir}\left(\boldsymbol{\gamma} + \sum_d \sum_i \mathbb{I}(w_{id} = m, z_d = k)\right), \quad (4)
\end{aligned}$$

em que $\boldsymbol{\beta_{-k}}$ indica todos os outros índices de $\boldsymbol{\beta}$ exceto k e $\mathbb{I}(\cdot)$ é a função indicadora. A condicional completa de θ_k para todo $k \in \{1, 2, \ldots, K\}$ é dada por:

$$p(\theta_k|\boldsymbol{\theta_{-k}}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \propto p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{W}|\boldsymbol{\alpha}, \boldsymbol{\gamma})$$

$$\begin{aligned}
&\propto p(\theta_k|\alpha_k) \prod_{d=1}^{D} p(z_d = k|\theta_k) \\
&\propto \text{Dir}(\theta_k|\alpha_k) \prod_{d=1}^{D} \text{Mult}(z_d = k|\theta_k) \\
&\propto \theta_k^{\alpha_k-1} \prod_{d=1}^{D} \theta_k^{\mathbb{I}(z_d=k)} \\
&\propto \theta_k^{\alpha_k-1+\sum_d \mathbb{I}(z_d=k)} \\
&\sim \text{Dir}\left(\alpha_k + \sum_d \mathbb{I}(z_d = k)\right). \quad (5)
\end{aligned}$$

A condicional completa de z_j é dada por:

$$\begin{aligned}
p(z_j|\boldsymbol{z_{-j}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &\propto p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{W}|\boldsymbol{\alpha}, \boldsymbol{\gamma}) \\
&\propto p(z_j = k|\boldsymbol{\theta}) \prod_{i=1}^{N_j} p(w_{ij} = m|\boldsymbol{\beta_k}, z_j = k) \\
&\propto \theta_k \prod_{i=1}^{N_j} \text{Mult}(w_{ij} = m|\boldsymbol{\beta_{z_j=k}}) \\
&\propto \theta_k \prod_{i=1}^{N_j} \boldsymbol{\beta_k}^{\mathbb{I}(w_{ij}=m)} \\
&\propto \theta_k \boldsymbol{\beta_k}^{\sum_i \mathbb{I}(w_{ij}=m)} = \rho_k. \quad (6)
\end{aligned}$$

Quando a quantidade de palavras no dicionário é grande, o calculo de (6) pode se tornar numericamente instável. Uma solução para isso é aplicar a função logaritmo:

$$\begin{aligned}
\rho_k &= \ln \left\{ \theta_k \boldsymbol{\beta_k}^{\sum_i \mathbb{I}(w_{ij}=m)} \right\} \\
&= \ln \theta_k + \left[\sum_i \mathbb{I}(w_{ij} = m)\right] \ln \boldsymbol{\beta_k} \\
&= \exp\left\{ \ln \theta_k + \left[\sum_i \mathbb{I}(w_{ij} = m)\right] \ln \boldsymbol{\beta_k} \right\}, \quad (7)
\end{aligned}$$

em seguida deve-se normalizar esses valores para que tenha soma unitária dividindo cada ρ_k por $\sum_{k=1}^{K} \rho_k$. Dessa forma, a partir de (4), (5) e (7), podemos utilizar o amostrador de Gibbs para obter uma estimativa das distribuições envolvidas no processo.

4 Resultados

4.1 Série temporal unidimensional

Considere a série temporal representada na Figura 4 e dada por:

$$x(t) = \begin{cases} 1 + \epsilon_t & 0 \leq t < 100 \\ 5 + \epsilon_t & 100 \leq t < 200, \end{cases} \quad (8)$$

em que $\epsilon_t \sim \mathcal{N}(0, 1)$. A série foi proposta de forma que houvessem dois grupos.

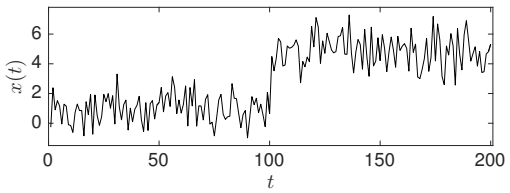

Figura 4: Série temporal $x(t)$ caracterizada por (8).

A técnica proposta é comparada com o algoritmo Fuzzy C-means (FCM) (Bezdek et al., 1984)

para agrupamento, pois ambas requerem que o número de protótipos de grupo (ou tópicos) seja informado. A Figura 5 mostra o resultado do agrupamento informando-se a existência de dois grupos. O resultado encontrado pelo FCM é idêntico ao encontrado pelo LDA.

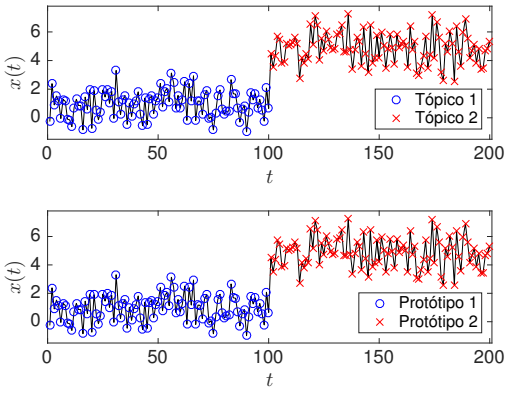

Figura 5: Resultado do agrupamento realizado pelo LDA e FCM respectivamente para um número de dois grupos.

A Figura 6 mostra o resultado do agrupamento informando-se a existência de três grupos. Nessa situação, a diferença entre uma técnica e outra se evidencia: enquanto o FCM encontra três grupos, o modelo LDA encontra dois grupos, os tópicos 1 e 3. Vale salientar que a mistura de tópicos nos documentos é uma variável aleatória e, por isso, o tópico 2 existe com uma probabilidade muito baixa.

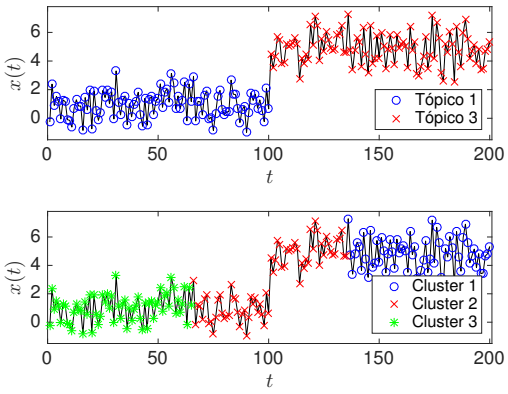

Figura 6: Resultado do agrupamento realizado pelo LDA e FCM respectivamente para um número de três grupos.

4.2 Base de dados da flor de Iris

A base de dados da flor de Iris é um conjunto multivariado introduzido pelo estatístico e biólogo britânico Ronald Fisher (Fisher, 1936). O conjunto possui 50 amostras de cada uma de três espécies de Iris: Setosa, Virgínica e Versicolor, totalizando 150 amostras. Para cada amostra, foram coletadas 4 características: o comprimento e largura das sépalas e pétalas.

No caso de uma base de dados multivariada, o método proposto será aplicado em cada dimensão separadamente, resultando em 4 grupos independentes para cada amostra. Em seguida, cada combinação desses quatro elementos será considerado um grupo.

Foram utilizadas as amostras sem nenhum tipo de pré-processamento, como redução de dimensionalidade. Testou-se a acurácia de três métodos de agrupamento: FCM, k-means e o LDA. O resultado da aplicação da técnica proposta nessa base de dados teve uma acurácia de **100%** e é mostrado na Figura 7. Tanto o FCM quanto o k-means obtiveram uma acurácia de de **89.33%**.

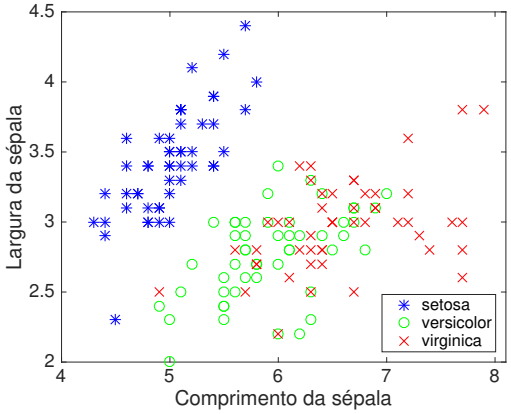

Figura 7: Resultado do agrupamento realizado pelo LDA na base de dados da flor de Iris.

5 Discussão

No primeiro exemplo, foi mostrada a diferença entre o LDA e o FCM em termos da quantidade de grupos criados. Tanto uma técnica quanto a outra requerem que o usuário informe a quantidade de grupos *à priori*. Enquanto o FCM resolve um problema de otimização que obrigatoriamente envolva o número de grupos solicitado, o LDA parte do princípio de que a distribuição dos grupos segue uma Multinomial cujos parâmetros são estimados por uma simulação de Monte Carlo via cadeias de Markov. Por esse motivo, caso existam apenas dois grupos, o terceiro terá uma probabilidade muito baixa de ser amostrado, fazendo com que apenas os grupos que realmente existem sejam atribuídos aos dados, como mostrado na Figura 6.

No segundo exemplo, a acurácia de **100%** do LDA se deve ao fato de que a técnica apresentada processa as características como se fossem séries temporais; como mostrado na Figura 8, nas amostras 50 e 100, que são exatamente os pontos em que há mudança de espécie na base, pode-se notar

uma mudança na média das 4 séries. As outras técnicas processam os dados no espaço através de medidas de distância entre as características de uma amostra e outra. Por esse motivo, mesmo que os dados viessem de forma desordenada, os algoritmos FCM e k-means manteriam sua acurácia enquanto a técnica proposta (LDA) encontraria apenas um grupo.

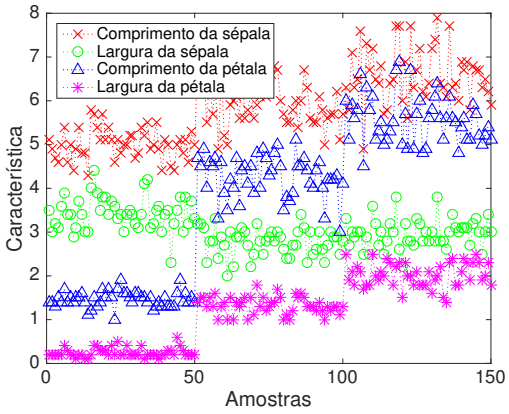

Figura 8: Amostras da base de dados da flor de Iris vistas sob a perspectiva de séries temporais.

6 Conclusões

Neste trabalho foi considerado o problema de agrupamento de dados com robustez à estimativa inicial da quantidade de grupos existentes. Para isso, os dados foram analisados sob o a perspectiva de séries temporais visando a aplicação de uma técnica conhecida por Alocação Latente de Dirichlet. Simulações de Monte Carlo via cadeias de Markov foram utilizadas para estimação dos parâmetros das distribuições que compõem este modelo. Para as situações descritas, o método se comportou de forma eficiente atingindo uma acurácia de **100%** na base de dados da flor de Iris e corretamente identificando o número real de grupos mesmo quando solicitado um número maior.

Referências

Ben-Hur, A., Horn, D., Siegelmann, H. T. e Vapnik, V. (2001). Support vector clustering, *Journal of machine learning research* **2**(Dec): 125–137.

Bezdek, J. C., Ehrlich, R. e Full, W. (1984). FCM: The fuzzy c-means clustering algorithm, *Computers & Geosciences* **10**(2-3): 191–203.

Blei, D. M., Ng, A. Y. e Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of machine Learning research* **3**(Jan): 993–1022.

Celeux, G. e Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model, *Journal of Classification* **13**(2): 195–212.

Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J. e Chen, T.-J. (2006). Fuzzy c-means clustering with spatial information for image segmentation, *Computerized Medical Imaging and Graphics* **30**(1): 9–15.

Coutinho, P. H. S. e das Chagas, T. P. (2017). Proposal of new hybrid fuzzy clustering algorithms — application to breast cancer dataset, *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, IEEE.

Dickey, J. M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses, *Journal of the American Statistical Association* **78**(383): 628–637.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**(2): 179–188.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. e Saul, L. K. (1999). *Machine Learning* **37**(2): 183–233.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, Oakland, CA, USA, pp. 281–297.

Windham, M. P. e Cutler, A. (1992). Information ratios for validating mixture analyses, *Journal of the American Statistical Association* **87**(420): 1188–1192.

Wolfe, J. H. (1970). PATTERN CLUSTERING BY MULTIVARIATE MIXTURE ANALYSIS, *Multivariate Behavioral Research* **5**(3): 329–350.