

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

BACHARELADO EM CIÊNCIAS EXATAS E TECNOLÓGICAS

AGRUPAMENTO DE DADOS: K- MÉDIAS

LUANN FARIAS PALMA

CRUZ DAS ALMAS, 2018

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

BACHARELADO EM CIÊNCIAS EXATAS E TECNOLÓGICAS

AGRUPAMENTO DE DADOS: K-MÉDIAS

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Recôncavo da Bahia como parte dos requisitos para obtenção do título de Bacharel em Ciências Exatas e Tecnológicas.

Orientadora: Profa. Dra. Julianna Pinele Santos Porto.

LUANN FARIAS PALMA

CRUZ DAS ALMAS, 2018

RESUMO

Agrupamento de dados é uma técnica que consiste em particionar um conjunto de dados em grupos segundo alguma função de dissimilaridade. Existem diversas técnicas de agrupamento e elas podem ser classificadas como hierárquicas e não-hierárquicas. Neste trabalho, abordamos algumas delas dando ênfase a um algoritmo relacionado à técnica não-hierárquica denominada de k-médias. Para isso, decorrer do texto, apresentam-se algumas das métricas e algumas aplicações na área de agrupamento de dados. Além disso, foi desenvolvido um algoritmo para realizar a segmentação de imagens através do algoritmo k-médias. De modo a avaliar a metodologia proposta, realizamos um experimento no qual comparamos as distâncias utilizadas na segmentação.

Palavras- chave: Agrupamento de dados, K- médias, Distâncias, Segmentação de Imagens.

ABSTRACT

Clustering is a technique that consists of partitioning a data set into groups according to some dissimilarity function. There are several grouping techniques and they can be classified as hierarchical and non-hierarchical. In this work, we address some of them, emphasizing an algorithm related to the non-hierarchical technique called k-means. For this, in the course of the text, some of the metrics and some applications in the area of data grouping are presented. In addition, an algorithm was developed to perform the segmentation of images through the k-means algorithm. In order to evaluate the proposed methodology, we performed an experiment in which we compared the distances used in the segmentation.

Key- words: Clustering, K- means, Distances, Images Segmentation.

SUMÁRIO

Introdução	6
1. Algoritmos de Agrupamento de Dados.....	8
1.1. Métodos Hierárquicos	9
1.2. Métodos Não- Hierárquicos.....	10
1.2.1. K-médias	11
1.2.2. K-Medoides	13
1.3. Outros Métodos.....	14
1.3.1. Agrupamento fuzzy	14
1.3.2. Redes Neurais Artificiais	14
2. Distâncias.....	16
2.1. Distância Euclidiana	16
2.2 Distância da Soma	19
2.3 Distância do Máximo	21
3. Segmentação de Imagens.....	24
3.1 Espaço de cores.....	25
3.1.1 RGB.....	25
3.1.2 HSV	26
3.1.3 CIELAB.....	27
3.2 Segmentação de imagens e algoritmo k-médias.....	28
4. Experimentos	30
5. Conclusão	40
6. Referências	41
Anexos	43

Introdução

O agrupamento de dados é uma técnica que consiste em caracterizar um grupo ou conjunto de dados de acordo com alguma característica de semelhança. Sua análise consiste em resumir as informações que são coletadas, usando uma função para diferenciá-las em elementos semelhantes ou distintos.

De forma intuitiva, o agrupamento de dados consiste em encontrar grupos de objetos de forma que os objetos em um grupo sejam similares (ou relacionados) um ao outro e diferentes (ou não relacionados à) os objetos em outros grupos (TAN ET AT., 2006).

Figura 1: Agrupamento

O que é agrupamento entre os seguintes objetos?



Grupo é um conceito subjetivo.



Fonte: Koegh, 2003.

A literatura sobre análise de agrupamento de dados é bem vasta. Segundo A. K. Jain (2010), o agrupamento de dados apareceu pela primeira vez no título de um artigo de 1954 sobre dados antropológicos.

Agrupar um conjunto de dados é de suma importância no sentido exploratório, possuindo aplicações diretas em áreas como engenharia, biologia, psicologia, medicina, administração e em ciências da computação. Por exemplo, no marketing a técnica pode ser aplicada a fim de descobrir grupos de clientes, na astronomia para encontrar possíveis grupos de estrelas e galáxias, no mineração de texto,

caracterizando documentos, técnicas de controle, detecção de anomalias, dentre outros.

De forma mais específica, em Duarte (2015) aplicou-se a técnica do k-médias com o objetivo de propor uma diferenciação das Regiões da Saúde Brasileira. Foi realizado um agrupamento em cinco grupos, sendo alguns deles a longevidade, riqueza, escolaridade. Na engenharia, para analisar a fitossociologia de comunidades florestais ao longo da sub-bacia hidrográfica do Rio Passo Fundo, Lounghi (2013) utilizou agrupamentos com o objetivo de estratificar a vegetação arbórea encontrada na região. Além disso, Guidini (2008) também aplicou técnicas de agrupamento de dados para realização de análises estatísticas em organizações empresarias com o objetivo de identificar os estilos de gestão.

Neste trabalho, apresentamos algumas das técnicas de agrupamento de dados, dando ênfase ao algoritmo k-médias (*k-means*) em que foram utilizadas as distâncias Euclidiana, da soma e do máximo com o propósito de realizar a segmentação de imagens.

No Capítulo 1, serão encontradas as definições de algumas técnicas de agrupamento de dados, além dos detalhes de funcionamento do algoritmo k-médias.

No Capítulo 2, apresentamos a definição de distância, no qual será demonstrado que as distâncias Euclidiana, da soma e do máximo satisfazem tal definição.

No Capítulo 3, falaremos um pouco sobre a segmentação de imagens, dando ênfase a forma *RGB* de segmentação, que será utilizada no trabalho.

No Capítulo 4, será realizado um experimento, mostrando como ficaram as imagens segmentadas, qual distância obteve melhores resultados, além do gráfico comparativo entre elas. No Capítulo 5, encontram-se as conclusões finais do trabalho. O algoritmo utilizado neste trabalho encontra-se em anexo.

Capítulo 1

1. Algoritmos de Agrupamento de Dados

Análise de agrupamento, também conhecida como *clustering*, é um conjunto de técnicas computacionais que consiste em separar objetos em grupos (*clusters*) baseados nas suas características (LINDEN, 2009). Esta técnica tem o objetivo de separar os grupos de acordo com alguma função de dissimilaridade, visando possuir características parecidas dentro de seus grupos e características distintas entre eles, ao mesmo tempo.

Em termos de definições gerais temos (Everitt, 1974):

- *Um grupo é uma aglomeração de pontos no espaço tal que a distância entre quaisquer dois pontos no grupo é menor do que a distância entre qualquer ponto no grupo e qualquer ponto fora deste;*

Com o passar dos anos, vários métodos e algoritmos vem sendo desenvolvidos com o intuito de dinamizar a aplicação das técnicas de agrupamento de dados em âmbitos sociais e econômicos. Dentre elas, podemos citar como as principais os métodos hierárquicos, o k-médias, o k-medoides, *fuzzy* e redes neurais. Todos os citados partem do princípio da análise de grupo, que segue um processo de partição de um banco de dados heterogêneos para subgrupos homogêneos. No agrupamento, não existem classes pré-definidas, sendo assim, os elementos são agrupados tendo como base alguma forma de semelhança, o que a diferencia da técnica de classificação, que se baseia em critérios pré-definido.

De acordo com Zaiane (2003), uma análise de grupo criteriosa exige métodos que apresentem as seguintes características:

- Ser capaz de lidar com dados com alta dimensionalidade;
- Ser “compatível” com o número de dimensões e com a quantidade de elementos a serem agrupados;
- Habilidade para lidar com diferentes tipos de dados;
- Capacidade de definir agrupamentos de diferentes tamanhos e formas;
- Exigir o mínimo de conhecimento para determinação dos parâmetros de entrada;

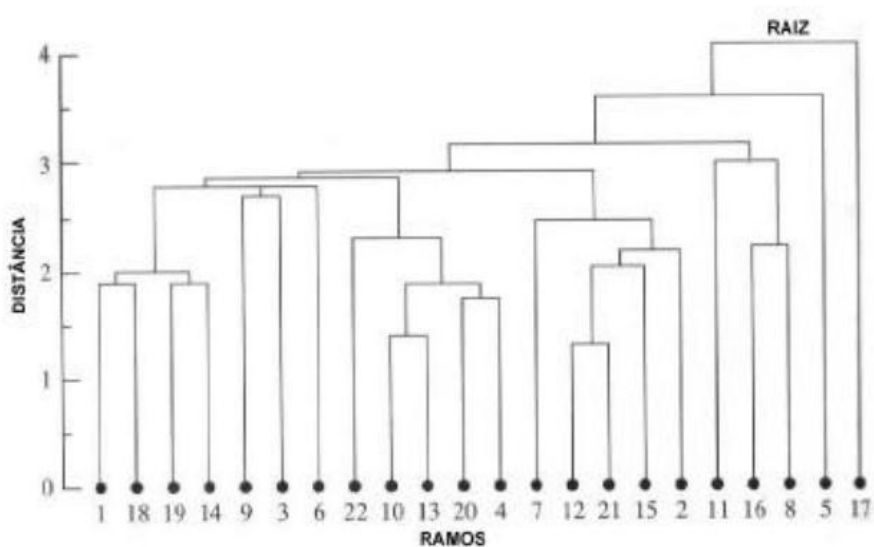
- Ser sensível à presença de ruído, uma vez que um elemento com um valor extremamente alto pode distorcer a distribuição dos dados;
- Apresentar resultado consistente independente da ordem em que os dados são apresentados.

Em geral, algoritmo algum atende a todos esses requisitos e, por isso, é importante entender as características de cada algoritmo para a escolha de um método adequado a cada tipo de dado ou problema (HALDIKI, 2001).

1.1. Métodos Hierárquicos

Algoritmos de agrupamento podem ser baseados em métodos hierárquicos, capazes de realizar análise de *clusters*. Eles tem como principal característica a possibilidade de, em determinado passo do algoritmo, mesclar um *cluster* com o outro, fazendo assim vários agrupamentos. Eles organizam os dados baseados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos, o que resulta em uma árvore binária (dendograma), ver Figura 2, onde a raiz da mesma representa o conjunto de dados inteiros e as folhas representam os indivíduos finais (DONI, 2004).

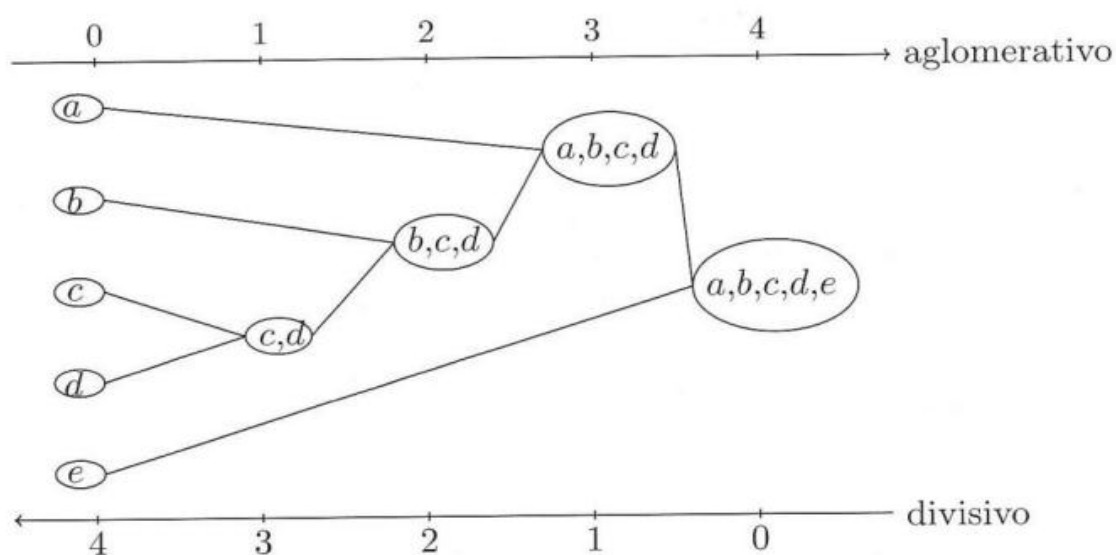
Figura 2: Árvore binária (dendograma)



Fonte: Doni, 2004.

Os métodos hierárquicos podem ser subdivididos em métodos aglomerativos e divisivos. Na forma aglomerativa, cada elemento se inicia representando um grupo e com o passar do tempo o elemento ou grupo se liga a outros por forma de similaridade, até o último passo, onde se forma um grupo único com todos os elementos. A forma divisiva acontece de forma oposta a aglomerativa, onde o grupo inicial contém todos os elementos do grupo de dados e com o decorrer do tempo, é dividido em subgrupos, de forma que os elementos de um subgrupo possuam certa distância dos elementos de outro.

Figura 3: Método hierárquico aglomerativo e divisivo



Fonte: profloresta.agro.ufg.br

Existem vários relatos da aplicação dos métodos hierárquicos nas áreas da medicina e biologia. Em Totti (2001), foram organizados grupos de espécies baseados em sua similaridade, a partir de características reprodutivas, vegetais e agrônômicas. A partir da construção de dendogramas foi possível, através das médias, obter uma análise sobre a correlação das mesmas.

1.2. Métodos Não- Hierárquicos

Agrupamentos de dados também podem ser baseados em métodos não-hierárquicos ou de particionamento. Estes métodos são desenvolvidos para agrupar

elementos em k grupos, em que k é a quantidade de grupos estabelecida previamente.

1.2.1. K-médias

Um dos métodos não-hierárquicos mais conhecidos e utilizados atualmente é o k-médias, que será descrito a seguir, e será o foco deste trabalho.

K-médias é um algoritmo de agrupamento de dados não-hierárquico que utiliza uma técnica iterativa para particionar um conjunto de dados. Ele foi proposto num trabalho pioneiro de S. Lloyd em 1957, contudo, só foi publicado no ano de 1982. Esse algoritmo busca minimizar a distância dos elementos de um conjunto de dados com k centros de forma iterativa.

Considere um conjunto de dados $X = \{p_1, p_2, \dots, p_n\}$, e uma distância $d: (\cdot, \cdot)$ nesse conjunto. O algoritmo k-médias inicia-se com a escolha de k centroides (centros) $\{c_1, c_2, \dots, c_k\}$ para o agrupamento depois, associa cada ponto do conjunto X ao seu centro mais próximo segundo a distância d formando assim k grupos X_i . O próximo passo do algoritmo é atualizar os centroides. Em Lloyd (1957), o centroide foi escolhido como sendo o ponto que minimiza a soma do quadrado da distância Euclidiana, d_E , entre ele mesmo e cada ponto do conjunto.

$$c_i = \operatorname{argmin} \sum_{p_j \in X_i} d_E^2(c_i, p_j),$$

Esse ponto é justamente o centro de massa do grupo X_i e é dado por:

$$c_i = \frac{p_{i1} + p_{i2} + \dots + p_{im}}{|X_i|} \quad (1)$$

em que $|X_i|$ corresponde a cardinalidade de X_i e $p_{ij} \in X_i$ com $j = 1, \dots, m$ (PINELE, 2017).

Resumindo, o algoritmo consiste nas seguintes etapas:

Dado um conjunto de amostras com N elementos, uma quantidade k e uma distância d ,

Escolher pontos $\{c_1, c_2, \dots, c_k\}$ para serem os centroides de cada grupo, que pode ser realizado de forma aleatória ou não.

a) Relacionar cada dado do conjunto com os centroides escolhidos:

Dizemos que $p_j \in X_j$ se

$$d(p_j, c_j) < d(p_j, c_i), \text{ com } i = 1, \dots, k \text{ e } i \neq j.$$

b) Atualizar o centroide de cada grupo:

$$c_i = \frac{p_{i1} + p_{i2} + \dots + p_{im}}{|X_i|}$$

c) Repetimos os passos *b)* e *c)* até o centroide não mudar em duas iterações sucessivas.

Na Figura 3, pode-se observar cada etapa que foi citada acima. Observa-se que na Figura 3a) o conjunto está disperso e têm-se como objetivo dividi-los em k grupos. Em seguida, na Figura 3b) são escolhidos dois centroides. Nas Figuras 3c) e 3d) os dados são atribuídos aos seus respectivos grupos de acordo com a menor proximidade entre ele e os centroides. Após isso, um novo centroide é calculado, e os dados se movem até que atinja a convergência.

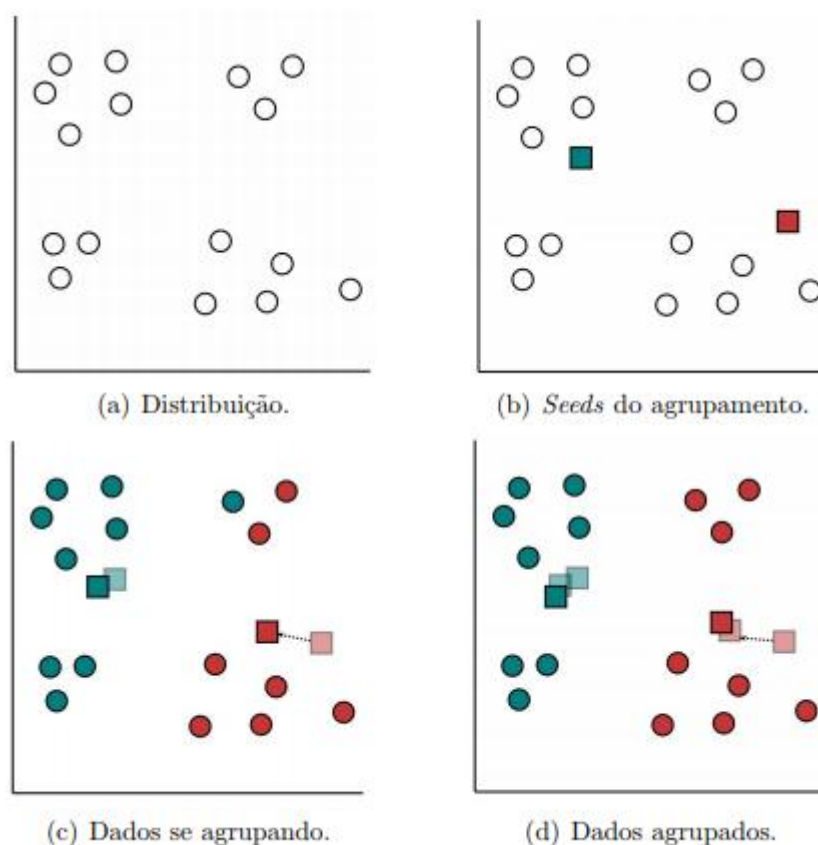
O objetivo do algoritmo k -médias é minimizar a soma do erro quadrático sobre todos os k grupos.

$$\sum_{j=1}^k \sum_{p_i \in X_i} d(p_i, c_j)^2.$$

Minimizar essa função é um problema muito difícil. É possível mostrar que o algoritmo k -médias converge para um mínimo local (A. K. JAIN, 2010).

Um dos problemas associados ao método em análise é que nem todos os valores de k apresentam grupos satisfatórios, sendo assim, aplica-se o método várias vezes para diferentes valores de k , escolhendo os resultados que apresentem melhor interpretação dos grupos ou uma melhor representação gráfica (BUSSAB, 1990).

Como exemplos de áreas em que é aplicado o agrupamento de dados utilizando o k -médias, pode-se citar: mineração de dados, estatística, engenharia, aprendizado de máquina, medicina, marketing, administração e biologia. Um exemplo da sua aplicação trata-se do artigo de Guidini (2008), que trata-se da aplicação da técnica citada utilizando um questionário e analisando as respostas obtidas a fim de realizar uma análise estatística e classificar os estilos de gestão e organizações empresariais.

Figura 4: Agrupamento de dados utilizando o *k-means*

Fonte: Prado 2008.

1.2.2. K-Medoides

Outro método de agrupamento não-hierárquico conhecido é o k-medoides (*k-medoids*). Sua estrutura e seu funcionamento são bem similares ao k-médias, o que os diferenciam é que enquanto no k-médias o centroide não precisa pertencer ao conjunto de dados, no k-medoides o centro é um dos pontos do conjunto, denominado medóide (DONI,2004).

Assim, cada elemento presente do banco de dados é agrupado conforme a sua similaridade com o medóide predefinido, e, de forma iterativa, ele é atualizado por um representante do mesmo grupo, até que o algoritmo atinja a convergência.

Um exemplo da sua aplicação pode ser encontrado no artigo de Amaral (2011). Estudo realizado na área de agrárias pela EMBRAPA que pesquisaram sobre os impactos causados por eventos climáticos extremos, utilizando o algoritmo k-medoides para analisar as séries temporais e indicar as áreas com melhores padrões de desenvolvimento simples da cana-de-açúcar.

1.3. Outros Métodos

Além das técnicas de agrupamento de dados utilizando métodos hierárquicos e não-hierárquicos, existem outras técnicas de agrupamento denominadas *fuzzy*, redes neurais (mapas de Kohonen), algoritmos evolutivos, que podem ser aplicados na área. Neste tópico mostraremos, brevemente, como as técnicas de agrupamento *fuzzy* e redes neurais funcionam.

1.3.1. Agrupamento fuzzy

Os métodos de agrupamento *fuzzy* podem ser comparados aos métodos de agrupamento não-hierárquicos, pois assim como eles, é necessário a indicação da quantidade de grupos desejados ao início do algoritmo. A diferença entre eles é que os agrupamentos *fuzzy* permitem visualizar a associação de cada elemento do banco de dados com cada grupo, no qual é possível verificar quando um elemento participa de dois ou mais grupos diferentes e analisar o seu grau de associação (DONI, 2004).

A principal vantagem dos agrupamentos *fuzzy* em relação aos outros métodos por particionamento, é que ele fornece informações mais detalhadas sobre a estrutura dos dados, pois são apresentados os graus de associação de cada elemento a cada grupo, não tendo, portanto, a formação de agrupamentos rígidos. A desvantagem desse método é que a quantidade de coeficientes de associação cresce rapidamente com o aumento do número de elementos e de grupos. Entretanto, trata-se de uma técnica válida, pois ela associa graus de incerteza aos elementos nos grupos e, essa situação, em geral, se aproxima mais das características reais dos dados (KAUFMANN, 1990).

1.3.2. Redes Neurais Artificiais

Redes neurais artificiais (RNAs) são modelos matemáticos que se assemelham às estruturas biológicas e que têm capacidade computacional adquirida por meio de aprendizado e generalização (HAYKIN, 1994).

O agrupamento por redes neurais artificiais consistem em um modo de resolver os problemas relacionados à inteligência artificial. Tem como um dos objetivos realizar o aprendizado de uma máquina baseado nos circuitos cerebrais,

buscando um comportamento de conhecimento gradativo, em que o computador realizaria tarefas e evoluiria ao fim das mesmas, assim como cometer erros, fazer descobertas, entre outros.

Uma das classes de redes neurais mais conhecidos são os Mapas Auto-Organizáveis de Kohonen. Essas classes são caracterizadas por serem baseados em uma forma de aprendizado competitivo, onde estruturas denominadas de neurônios tendem a aprender a distribuição estatística dos dados de entrada. O seu campo de aplicação é designado principalmente para reconhecimento de padrões (DONI, 2004).

Capítulo 2

2. Distâncias

A maioria dos métodos citados no Capítulo 1 são executados baseando-se em uma medida de dissimilaridade. Para realizar a análise de grupos são utilizadas, normalmente, funções de distâncias (LIMA 2015).

Definição:

Considere o espaço \mathbb{R}^n .

Dados $x, y, z \in \mathbb{R}^n$, uma distância nesse espaço é a função:

$$d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^+$$

$$(x, y) \rightarrow d(x, y)$$

Que satisfaz as seguintes propriedades:

- | | | |
|------|-------------------------------------|---------------------------|
| i) | $d(x, y) \geq 0$ | (Positividade) |
| ii) | $d(x, y) = 0 \Leftrightarrow x = y$ | (Nulidade) |
| iii) | $d(x, y) = d(y, x)$ | (Simetria) |
| iv) | $d(x, y) \leq d(x, z) + d(y, z)$ | (Desigualdade Triangular) |

As funções definidas nas seções a seguir são exemplos de funções de distâncias chamadas Euclidiana, da soma e do máximo, respectivamente.

2.1. Distância Euclidiana

A distância Euclidiana é a distância mais conhecida dentre as métricas. Essa distância é a menor distância entre dois pontos no \mathbb{R}^n , que pode ser representada pela hipotenusa, observada no Teorema de Pitágoras.

Sejam $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$ e $z = (z_1, \dots, z_n) \in \mathbb{R}^n$.

A distância Euclidiana definida por:

$$d_E(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}.$$

Vamos mostrar que a distância Euclidiana satisfaz as propriedades de distância:

i) $d_E(x, y) \geq 0$

Observe que:

$$(x_i - y_i)^2 \geq 0, \forall i = 1, \dots, n$$

logo,

$$\sum_{i=1}^n (x_i - y_i)^2 \geq 0.$$

Dessa forma,

$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \geq 0.$$

ii) $d_E(x, y) = 0 \Leftrightarrow x = y$

Observe que,

$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} = 0$$

$$\Leftrightarrow \sum_{i=1}^n (x_i - y_i)^2 = 0,$$

$$\Leftrightarrow (x_i - y_i)^2 = 0, \forall i = 1, \dots, n$$

$$\Leftrightarrow (x_i - y_i) = 0, \forall i = 1, \dots, n$$

$$\Leftrightarrow x_i = y_i, \forall i = 1, \dots, n$$

$$\Leftrightarrow x = y.$$

$$\text{iii)} \quad d_E(x, y) = d_E(y, x)$$

Sabe-se que,

$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} = d_E(y, x).$$

$$\text{iv)} \quad d_E(x, y) \leq d_E(x, z) + d_E(y, z)$$

Para mostrar essa propriedade, vamos utilizar a desigualdade de Cauchy- Schwarz, mostraremos que:

$$d_E(x, y)^2 \leq (d_E(x, z) + d_E(y, z))^2.$$

Temos,

$$d_E(x, y)^2 = \sum_{i=1}^n (x_i - y_i)^2.$$

Como,

$$\begin{aligned} (x_i - y_i)^2 &= (x_i - z_i + z_i - y_i)^2 \\ &= ((x_i - z_i) + (z_i - y_i))^2 \\ &= (x_i - z_i)^2 + 2(x_i - z_i)(z_i - y_i) + (z_i - y_i)^2 \end{aligned}$$

segue que,

$$\begin{aligned} d_E(x, y)^2 &= \sum_{i=1}^n (x_i - y_i)^2 \\ &= \sum_{i=1}^n (x_i - z_i)^2 + 2(x_i - z_i)(z_i - y_i) + (z_i - y_i)^2 \\ &= \sum_{i=1}^n (x_i - z_i)^2 + 2 \sum_{i=1}^n (x_i - z_i)(z_i - y_i) + \sum_{i=1}^n (z_i - y_i)^2. \quad (2) \end{aligned}$$

Para finalizar esta demonstração, utilizaremos a desigualdade de Cauchy-Schwarz¹.

$$\sum_{i=1}^n \alpha_i \beta_i \leq \sqrt{\sum_{i=1}^n \alpha_i^2} \sqrt{\sum_{i=1}^n \beta_i^2}$$

¹ A demonstração da desigualdade de Cauchy-Schwarz pode ser encontrada em (LIMA, 2015).

em que, $\alpha = (\alpha_1, \dots, \alpha_n)$ e $\beta = (\beta_1, \dots, \beta_n)$ são vetores do \mathbb{R}^n . Chamando $\alpha_i = (x_i - z_i)$ e $\beta_i = (z_i - y_i)$, temos que,

$$\sum_{i=1}^n (x_i - z_i)(z_i - y_i) \leq \sqrt{\sum_{i=1}^n (x_i - z_i)^2} \sqrt{\sum_{i=1}^n (z_i - y_i)^2}.$$

Substituindo a desigualdade acima na Equação (2), segue que,

$$\begin{aligned} d_E(x, y)^2 &= \sum_{i=1}^n (x_i - y_i)^2 \\ &= \sum_{i=1}^n (x_i - z_i)^2 + 2 \sum_{i=1}^n (x_i - z_i)(z_i - y_i) + \sum_{i=1}^n (z_i - y_i)^2 \\ &\leq \sum_{i=1}^n (x_i - z_i)^2 + 2 \sqrt{\sum_{i=1}^n (x_i - z_i)^2} \sqrt{\sum_{i=1}^n (z_i - y_i)^2} + \sum_{i=1}^n (z_i - y_i)^2 \\ &= d_E(x, z)^2 + 2 d_E(x, z) d_E(z, y) + d_E(z, y)^2 \\ &= (d_E(x, z) + d_E(y, z))^2. \end{aligned}$$

Portanto,

$$d_E(x, y)^2 \leq (d_E(x, z) + d_E(y, z))^2.$$

Como todas as propriedades de distâncias são válidas, provamos que a distância Euclidiana é de fato uma distância.

2.2 Distância da Soma

Utilizaremos também a distância da Soma, conhecida também como distância de Manhattan ou geometria do taxista, representa a distância entre dois pontos como a soma das diferenças absolutas de suas coordenadas. Seu nome é associado ao formato quadriculado da maior parte das ruas da cidade de Manhattan que os taxistas tendem a tomar como rota (LIMA, 1977).

Dados $x, y \in \mathbb{R}^n$,

$$d_S(x, y) = |x_1 - y_1| + \dots + |x_n - y_n| = \sum_{i=1}^n |x_i - y_i|.$$

Vamos mostrar que d_S satisfaz as propriedades de distância.

$$i) \quad d_S(x, y) \geq 0$$

Observe que $|x_i - y_i| \geq 0, \forall i = 1, \dots, n$ logo,

$$d_S(x, y) = \sum_{i=1}^n |x_i - y_i| \geq 0.$$

$$ii) \quad d_S(x, y) = 0 \Leftrightarrow x = y$$

$$d_S(x, y) = \sum_{i=1}^n |x_i - y_i| = 0$$

$$\Leftrightarrow |x_i - y_i| = 0, \forall i = 1, \dots, n$$

$$\Leftrightarrow x_i = y_i, \forall i = 1, \dots, n$$

$$\Leftrightarrow x = y.$$

$$iii) \quad d_S(x, y) = d_S(y, x)$$

Como $|x_i - y_i| = |y_i - x_i|, \forall i = 1, \dots, n$ logo,

$$d_S(x, y) = \sum_{i=1}^n |x_i - y_i| = \sum_{i=1}^n |y_i - x_i| = d_S(y, x).$$

$$iv) \quad d_S(x, y) \leq d_S(x, z) + d_S(y, z)$$

Da desigualdade triangular, segue que $|a + b| \leq |a| + |b|$, assim:

$$|x_i - y_i| = |x_i + z_i - z_i - y_i|, \forall i = 1, \dots, n$$

$$\Rightarrow |x_i - y_i| \leq |x_i + z_i| + |z_i - y_i|, \forall i = 1, \dots, n$$

logo,

$$\begin{aligned} d_S(x, y) &= \sum_{i=1}^n |x_i - y_i| \\ &\leq \sum_{i=1}^n |x_i + z_i| + \sum_{i=1}^n |z_i - y_i| \\ &= \sum_{i=1}^n |x_i - z_i| + \sum_{i=1}^n |y_i - z_i| \\ &= d_S(x, z) + d_S(y, z). \end{aligned}$$

Portanto, a distância da soma é de fato uma distância.

2.3 Distância do Máximo

A distância do Máximo, também conhecida como distância de Chebyshev, entre dois vetores é dada pela maior diferença de suas coordenadas. Ela também é conhecida como a distância do tabuleiro de xadrez, de forma que o rei cumpre exatamente ao movimento citado anteriormente (LIMA, 1977).

Dados $x, y \in \mathbb{R}^n$, a distância do máximo é definida por:

$$d_M(x, y) = \max\{|x_1 - y_1|, \dots, |x_n - y_n|\} = \max_{1 \leq i \leq n}\{|x_i - y_i|\}.$$

Vamos mostrar que a mesma satisfaz as propriedades de distância.

$$\text{i) } d_M(x, y) \geq 0$$

Observe que $|x_i - y_i| \geq 0, \forall i = (1, \dots, n)$ logo,

$$d_M(x, y) = \max_{1 \leq i \leq n}\{|x_i - y_i|\} \geq 0.$$

$$\text{ii) } d_M(x, y) = 0 \Leftrightarrow x = y$$

$$d_M(x, y) = \max_{1 \leq i \leq n}\{|x_i - y_i|\} = 0$$

$$\Leftrightarrow |x_i - y_i| = 0, \forall i = (1, \dots, n)$$

$$\Leftrightarrow (x_i - y_i) = 0 \text{ ou } -(x_i - y_i) = 0, \forall i = 1, \dots, n$$

$$\Leftrightarrow x_i = y_i, \forall i = 1, \dots, n$$

$$\Leftrightarrow x = y.$$

$$\text{iii) } d_M(x, y) = d_M(y, x)$$

Como $|x_i - y_i| = |y_i - x_i|, \forall i = 1, \dots, n$, segue que,

$$d_M(x, y) = \max_{1 \leq i \leq n}\{|x_i - y_i|\}$$

$$= \max_{1 \leq i \leq n}\{|y_i - x_i|\}$$

$$= d_M(y, x).$$

$$\text{iv)} \quad d_M(x, y) \leq d_M(x, z) + d_M(z, y)$$

Observe que:

$$|x_i - y_i| = |x_i + z_i - z_i - y_i|, \forall i = 1, \dots, n,$$

Segue a desigualdade triangular que,

$$|x_i - y_i| \leq |x_i + z_i| + |z_i - y_i|, \forall i = 1, \dots, n.$$

Como,

$$|x_i - z_i| \leq \max_{1 \leq i \leq n} \{|x_i - z_i|\} = d_M(x, z), \forall i = 1, \dots, n$$

e,

$$|z_i - y_i| \leq \max_{1 \leq i \leq n} \{|z_i - y_i|\} = d_M(z, y), \forall i = 1, \dots, n,$$

Temos,

$$\begin{aligned} |x_i - y_i| &\leq |x_i + z_i| + |z_i - y_i|, \\ &\leq d_M(x, z) + d_M(z, y) \end{aligned}$$

para todo $i = 1, \dots, n$, logo

$$\begin{aligned} d_M(x, y) &= \max_{1 \leq i \leq n} \{|x_i - y_i|\} \\ &\leq d_M(x, z) + d_M(z, y). \end{aligned}$$

Dessa forma, temos que a distância do máximo é de fato uma distância.

Todas as distâncias demonstradas acima podem ser generalizadas pela distância de Minkowski, dada por:

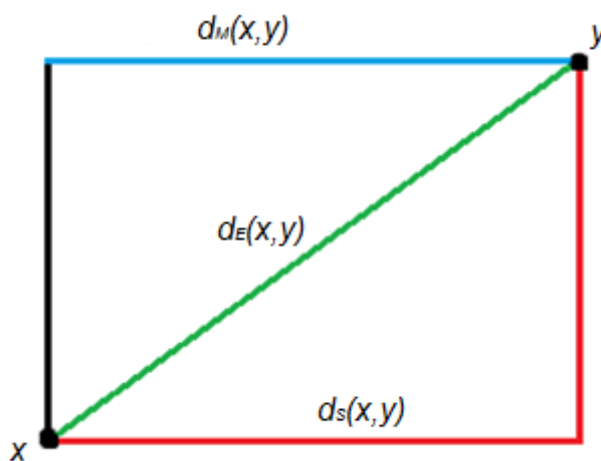
$$d(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^{\frac{1}{p}} \right)^p.$$

Observe que quando $p=1$ obtêm-se a fórmula da distância da soma. Quando $p=2$ temos a fórmula da distância Euclidiana e quando p tende a infinito, obtemos a distância do máximo.

Dados dois pontos x e y a Figura 5 ilustra as condições citadas acima. Nela, pode-se observar um retângulo qualquer onde duas das suas extremidades são representadas pelos pontos x e y . Nota-se, em azul, a distância do máximo, descrita pela maior distancia absoluta entre os pontos x e y . Assim como a distância Euclidiana, em verde na figura, retrata-se como a menor distância conhecida entre

dois pontos. E em vermelho, a distância da soma, onde a distância entre dois pontos é dada pela soma das diferenças absolutas de suas coordenadas.

Figura 5: Interpretação geométrica das distâncias d_E , d_S , d_M .



Capítulo 3

3. Segmentação de Imagens

Com o avanço tecnológico, o processo digital tem um papel de grande relevância. Uma área de grande significância no processamento digital são os métodos de dividir uma imagem em um conjunto de *pixels* ou objetos, sendo assim, a união de todas as regiões formam uma imagem. Segmentação implica na divisão de uma imagem em diferentes objetos ou regiões conectadas que não se sobrepõem (YERPUDE, 2012).

O objetivo da segmentação é facilitar a utilização das informações fundamentadas nas formas e texturas de objetos que constitui a imagem, dividindo-se a imagem em partes expressivas. Vários tipos de métodos de segmentação de imagem têm sido propostos no decorrer dos últimos anos, e estes podem ser classificados em duas categorias, o método baseado no contorno e o método baseado na região.

Toda imagem digital é formada por pontos de forma quadrada chamados *pixel*. O termo *pixel* é uma abreviatura do inglês *Picture element* que significa elemento da figura. Corresponde a menor unidade de uma imagem digital, onde são descritos a cor e o brilho específico de uma célula da imagem. Quanto maior a quantidade de *pixels*, maior será o tamanho da imagem e melhor a qualidade de detalhe visível da mesma (LAUREGA, 2006). Em imagens coloridas o *pixel* é formado por três cores, como mostraremos na seção a seguir.

Para realizar o tratamento computacional de uma imagem, é necessário representar suas informações em um formato adequado. Assim, elas são representadas por matrizes de pixels, que correspondem aos valores de *pixels* da imagem, definindo seu mapeamento para que possa ser representada pelo computador (MORGAN, 2008).

3.1 Espaço de cores

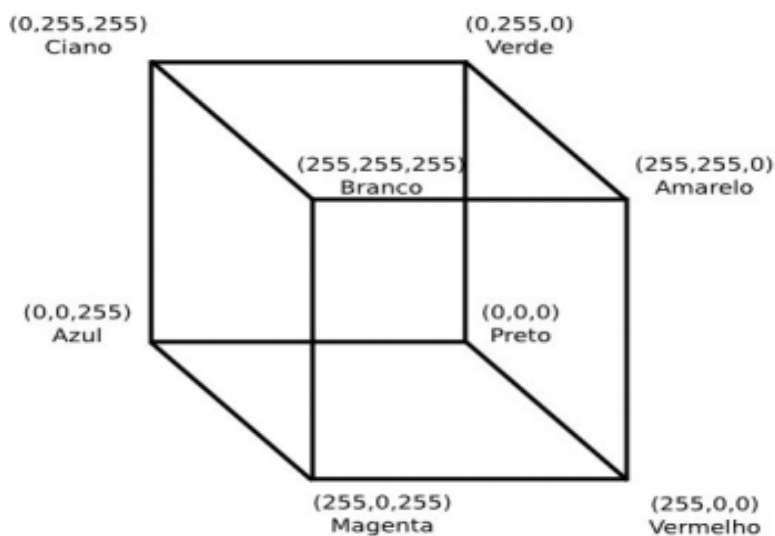
O método baseado no contorno consiste em usar as discontinuidades presentes em uma imagem a fim de detectar as suas bordas, e assim particioná-las. Por outro lado, o método baseado na região tem o objetivo de dividir os pixels presentes em uma imagem dividindo-os em grupos correspondentes a uma característica específica, como a cor, por exemplo.

Vale ressaltar que dentro da segmentação de imagens existe uma seção conhecida como espaço de cores. Que nada mais é que uma forma de representar as cores de uma forma específica, sendo as mais comuns a *RGB*, a *HSV* e a *CIELab*.

3.1.1 RGB

Podemos descrever o espaço *RGB* como um sistema de coordenadas que possui um valor mínimo igual a 0 e um valor máximo igual a 255. Isto é, cada uma das cores, vermelho, verde e azul, possuem 256 tonalidades.

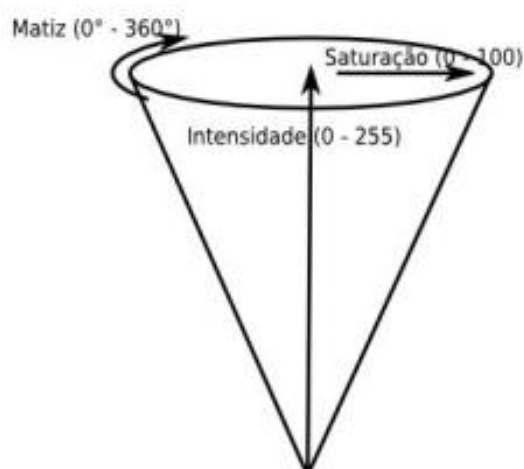
Quando cada canal (eixo do sistema de coordenadas) assumem os valores iguais ao seu valor máximo, a cor branca é representada e, quando todos os canais assumem seu o valor mínimo, a cor preta será representada. Combinando cada possibilidade de variações do canal, obtêm-se um resultado de mais de 16 milhões de cores diferentes. Tal representação pode ser observada na Figura 6, a seguir.

Figura 6: Espaço de cor *RGB*.

Fonte: Prado, 2008.

3.1.2 HSV

O espaço *HSV* (*Hue*, *Saturation*, *Value*) o canal *H* representa uma matriz e seus valores estão entre 0° e 360° . Como podemos observar na Figura 7 a seguir, ao movimentar-se pelo círculo pode-se obter cores diferentes, de 0° a 120° é representada a tonalidade de vermelho, de 120° a 240° a tonalidade de verde e de 240° a 360° a tonalidade de azul. O canal *S* representa a saturação com sua faixa de variação de 0 a 100 e representa a intensidade da cor, por exemplo, quanto menor for a saturação, menor será a intensidade da tonalidade de uma cor. E o canal *V* condiz ao valor da cor, representado num intervalo de 0 a 255. Neste canal, quanto menor for seu valor, mais próximo será a cor preta.

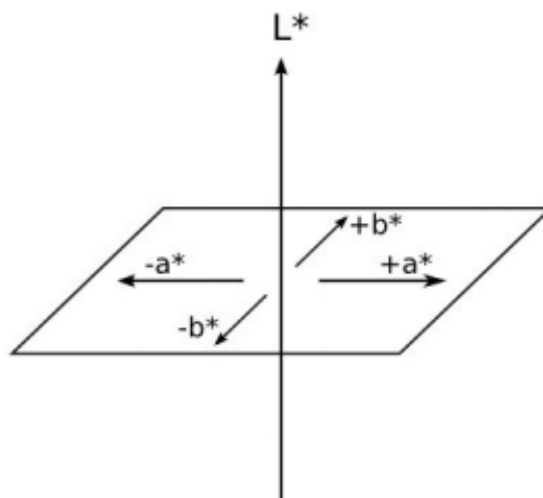
Figura 7: Espaço de cor *HSV*.

Fonte: Prado, 2008.

3.1.3 CIELAB

O espaço de cor *CIE Lab* tem a intenção de padronizar a escala de cores. Como pode-se ver na imagem a seguir, sua representação pode ser vista em que $L^* \in [0,100]$ condiz à intensidade de luz, onde 100 representa a cor branca e 0 representa a cor preta. Os valores de a^* e b^* não possuem limites, onde a^* positivo representa a cor vermelha, a^* negativo representa a cor verde, b^* positivo a cor amarela e b^* negativo a cor azul.

Atualmente vários pesquisadores tem tratado a segmentação de imagem como um problema de classificação não supervisionada ou um problema de agrupamento. Em alguns dos seus métodos são utilizadas técnicas de agrupamento como a *fuzzy*, e o *k*-médias, realizando uma associação entre o protótipo e os seus pixels.

Figura 8: Espaço de cor *CIE Lab*.

Fonte: Prado, 2008.

Neste trabalho, realizaremos a segmentação de imagem, baseado no espaço de cores *RGB*, aplicando a técnica de agrupamento de dados k-médias.

3.2 Segmentação de imagens e algoritmo k-médias

O método k-médias, apresentado na seção 2.2.1, pode ser aplicado na área de segmentação de imagens para subdividir uma imagem em k grupos de cores, escolhidos pelo usuário, diminuindo assim a quantidade de cores presentes na imagem.

Para fazer a segmentação de imagens utilizando o k-médias, primeiro consideramos os *pixels* da imagem como um conjunto de dados. Depois aplicamos o algoritmo k-médias para agrupar os *pixels* de acordo com alguma distância, no qual os representantes de cada grupo, ou seja, os centroides foram calculados como na Equação (1). A segmentação é dada quando trocamos cada *pixel* da matriz pelo representante do grupo ao qual ele pertence.

Por exemplo, considere a imagem da Lena (Figura 9a) que possui 256x256 *pixels* em *RGB*. Vamos agrupar os *pixels* dessa imagem de forma a representá-la apenas com 8 cores (ver Figura 9b). Esse processo foi realizado utilizando a

distância Euclidiana como função de dissimilaridade. Ao escolher um valor de $k = 8$, foram realizadas as comparações de todos os *pixels* da imagem com todos os centroides e cada um deles foi atribuído a um grupo. Ao fim deste processo é escolhido um novo centroide e o processo se repete até que atinja a convergência. Dessa forma, a todos os dados presentes em seus respectivos grupos finais são designadas cores que o grupo represente.

Fazendo essa comparação temos:

Figura 9a: Lena

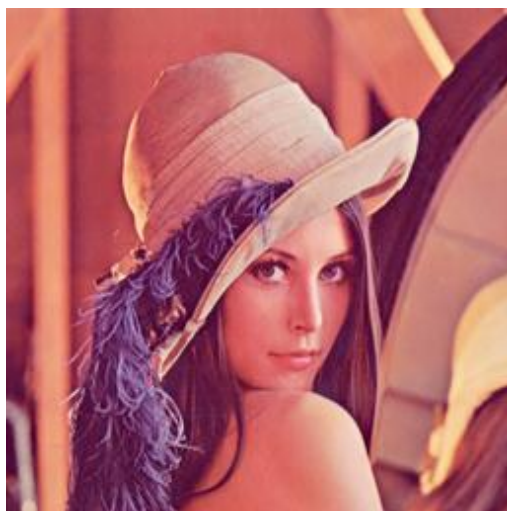
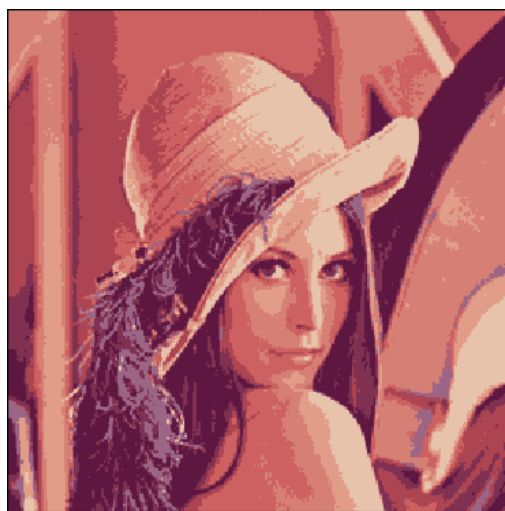


Figura 9b: Lena segmentada em 8 cores



No próximo capítulo, realizaremos a segmentação de imagens para outras figuras e distâncias e analisaremos a qualidade da segmentação.

Capítulo 4

4. Experimentos

Os experimentos realizados neste capítulo consistem na aplicação do k -médias operando seguindo as três distâncias citadas no Capítulo 3 (Euclidiana, da soma e do máximo), para realizar a segmentação de imagens. Para cada uma das distâncias analisamos a qualidade da segmentação.

O PSNR ou *Peak Signal Noise Ratio* consiste em uma relação de um sinal com um ruído de pico. Como o próprio nome já diz, o PSNR é uma relação entre o máximo possível de potência de um sinal, pela potência do ruído, quando comparamos um sinal antes e depois de um processo de degradação, sendo que a unidade utilizada para representa-lo é o dB (decibéis). Aplicando este conceito em vídeos e imagens, temos que o PSNR é a relação entre a entrada e a saída de um processo de compressão com perdas, que avalia o quanto a compressão introduziu ruídos na imagem ou frame original (PEREIRA, 2008).

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right) = 20 \log_{10} \left(\frac{MAX}{\sqrt{MSE}} \right),$$

onde MAX é o valor máximo possíveis de um $pixels$ e MSE (*Mean Square Error*):

$$MSE = \frac{1}{256^3} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} ||I(i,j) - K(i,j)||^2,$$

de forma que I representa a matriz de pixels correspondente a imagem original e K a matriz de pixels da imagem pós-segmentação.

As imagens utilizadas neste experimento foram escolhidas pelo autor para serem aplicadas apenas no trabalho em questão e elas foram Lena, Babbon e Peppers, ver Figuras 9, 10 e 11. As imagens originais possuem uma resolução de 256x256 pixels e o agrupamento foi realizando para cada uma delas tomando $k = 2, 4, 8, 16, e 32$ grupos.

Figura 10. Babbon.

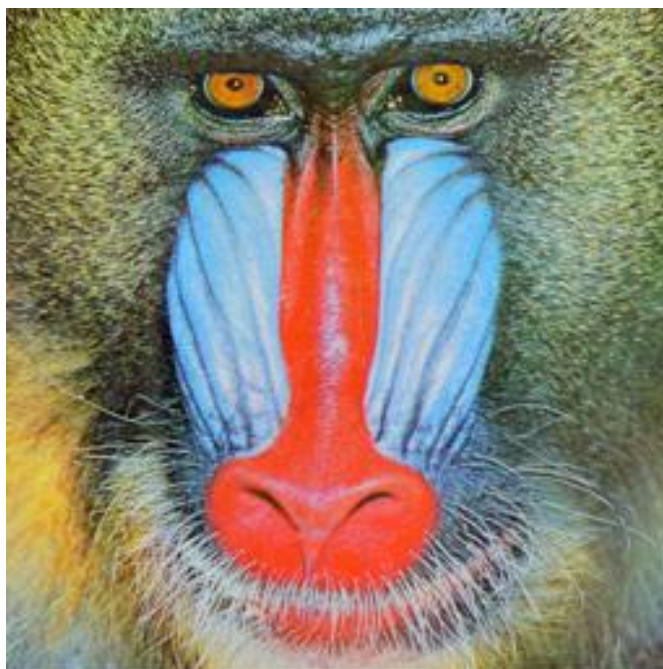


Figura 11. Peppers



As Tabelas 1, 2 e 3 mostram a segmentação das imagens utilizando as distâncias Euclidiana, da soma e do máximo respectivamente.

Tabela 1. Agrupamento realizado com a distância Euclidiana.



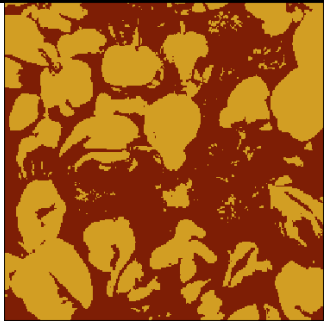


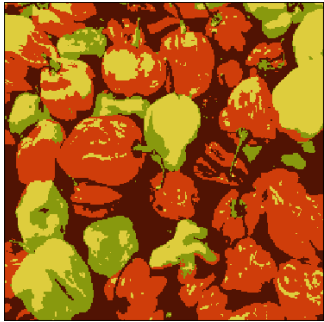



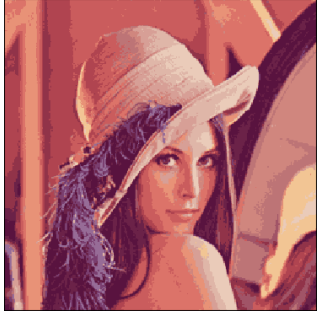
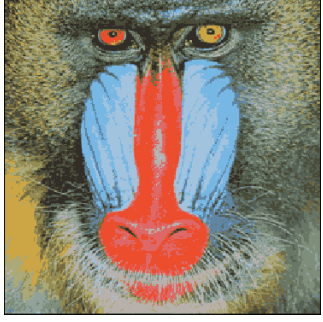
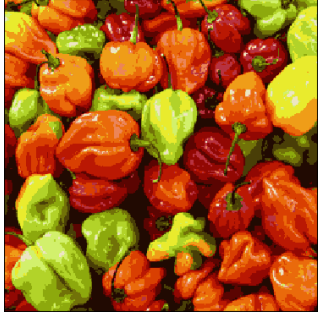
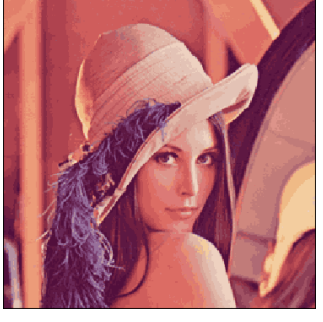
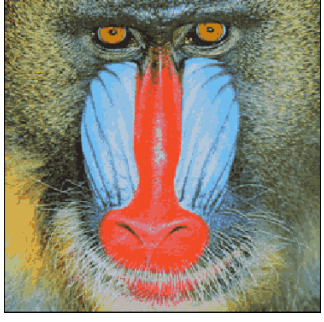

Gr	Lena	Babbon	Peppers
2			
4			
8			
16			
32			

Tabela 2. Agrupamento realizado com a distância da Soma.



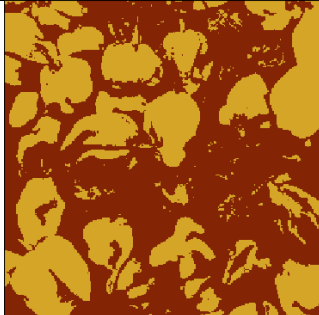


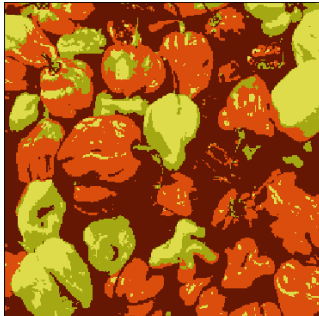
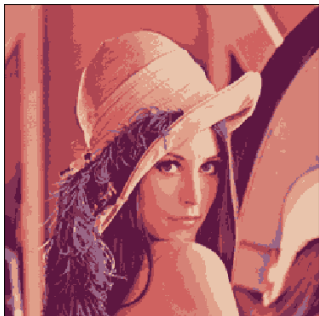

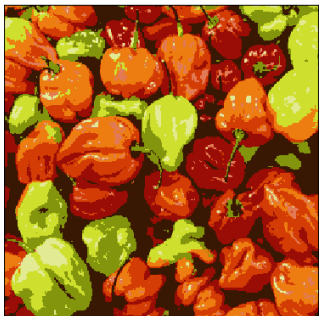








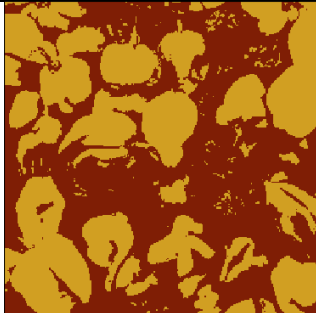
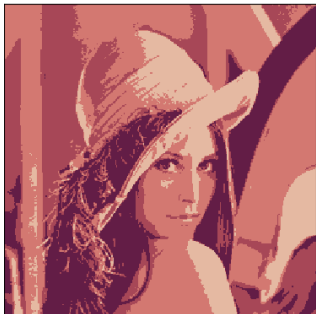

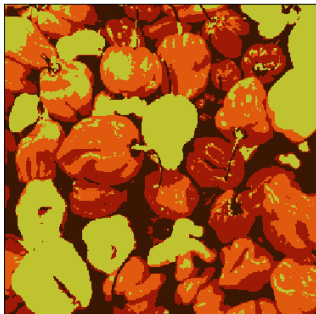
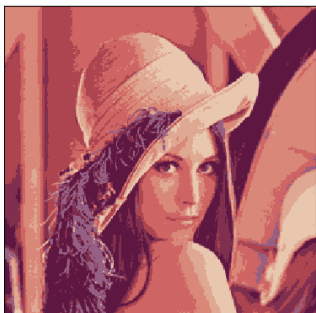








Gr	Lena	Babbon	Peppers
2			
4			
8			
16			
32			

Tabela 3. Agrupamento realizado com a distância do Máximo.

Gr	Lena	Babbon	Peppers
2			
4			
8			
16			
32			

Calculamos o PSNR de cada segmentação. Nas Tabelas 4, 5 e 6, estão representados os valores do PSNR resultantes da segmentação utilizando, respectivamente as distâncias Euclidiana, da soma e do máximo.

Tabela 4: Dados de PSNR resultantes da distância Euclidiana

Grupos	Lena	Babbon	Peppers
2	38.6573	35.5263	34.5865
4	43.2099	39.4635	37.5618
8	46.4226	42.3499	40.8663
16	49.3447	44.9803	43.7813
32	52.1167	47.4277	46.2736

Tabela 5: Dados de PSNR resultantes da distância da Soma

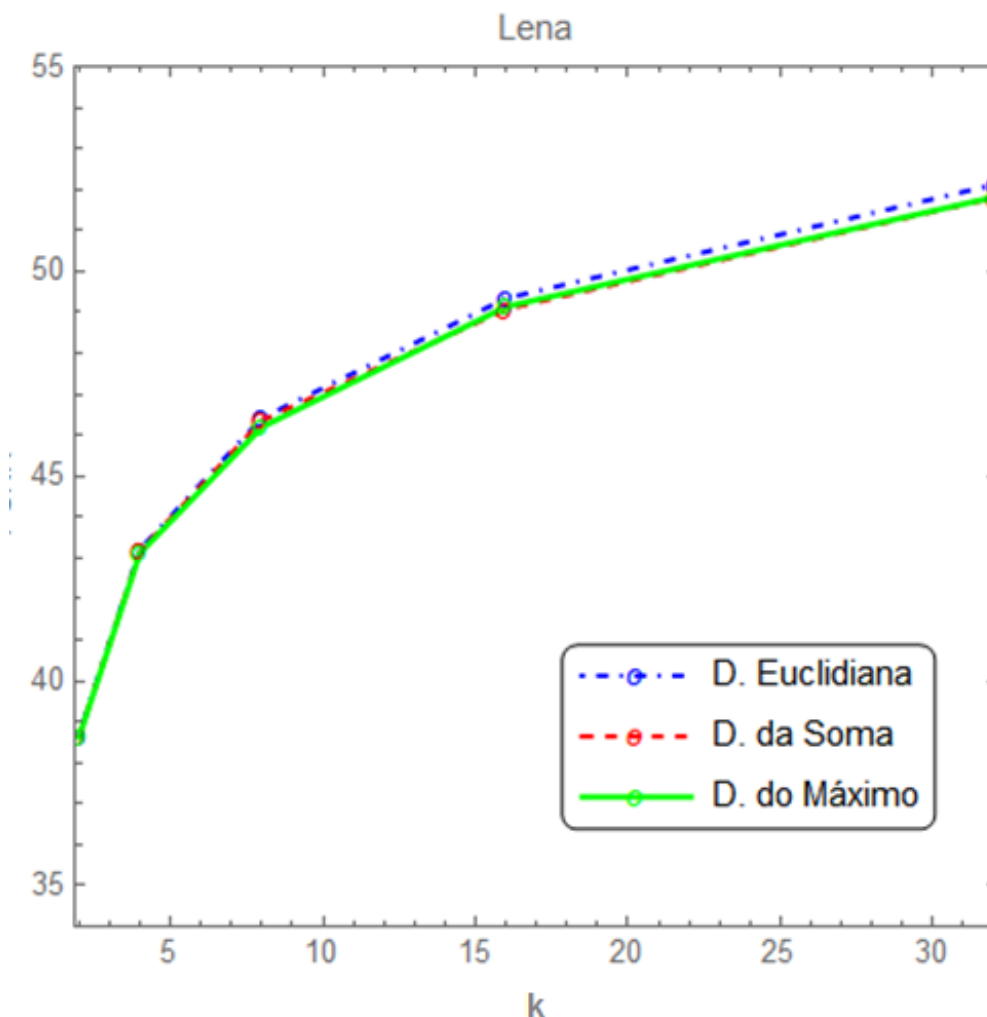
Grupos	Lena	Babbon	Peppers
2	38.6448	35.5021	34.5197
4	43.1675	39.3999	37.2194
8	46.3594	42.2431	40.6836
16	49.0782	44.8199	43.4866
32	51.8023	47.1634	45.8846

Tabela 6: Dados de PSNR resultantes da distância do Máximo

Grupos	Lena	Babbon	Peppers
2	38.6278	35.1023	34.5732
4	43.1235	39.3357	37.5014
8	46.2010	42.1709	40.6780
16	49.1411	44.7507	43.3416
32	51.8137	47.3162	46.1051

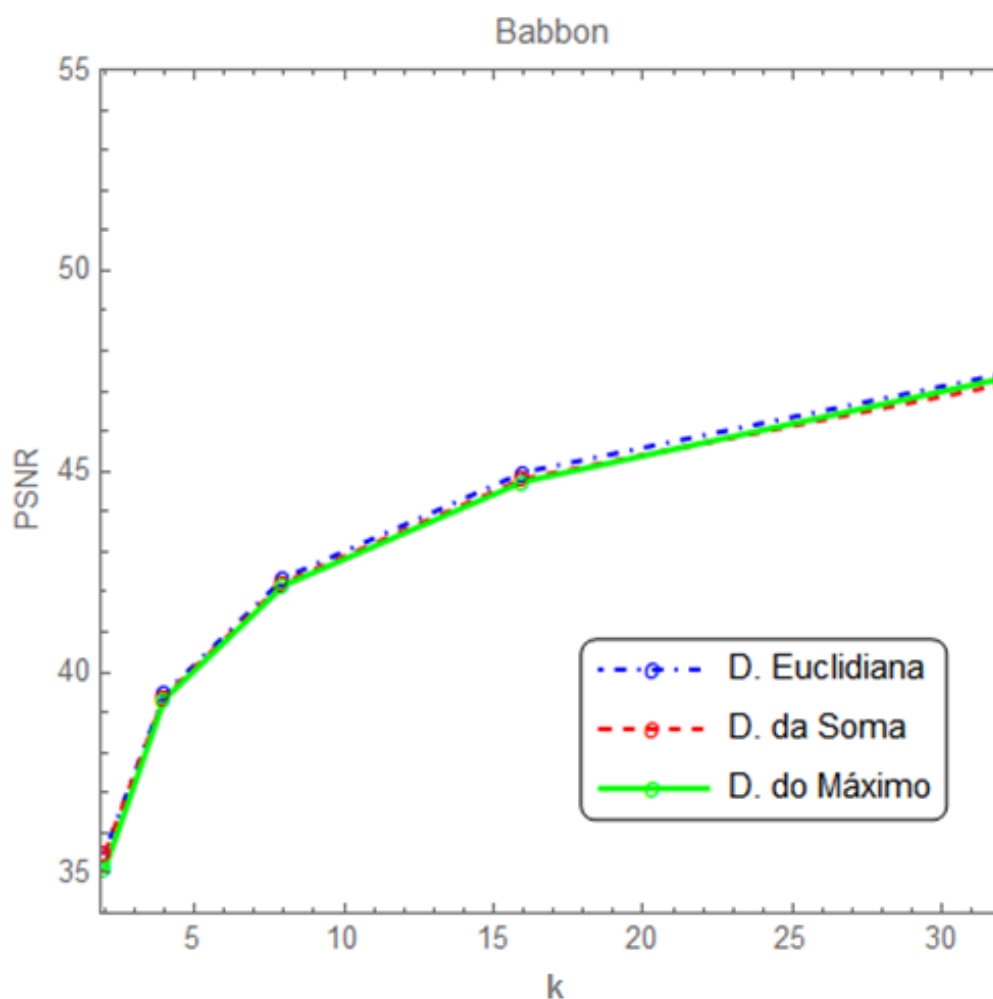
Apresentamos os dados das tabelas acima em forma de gráfico, para melhor visualizá-los.

Gráfico 1: Dados resultantes da Lena.



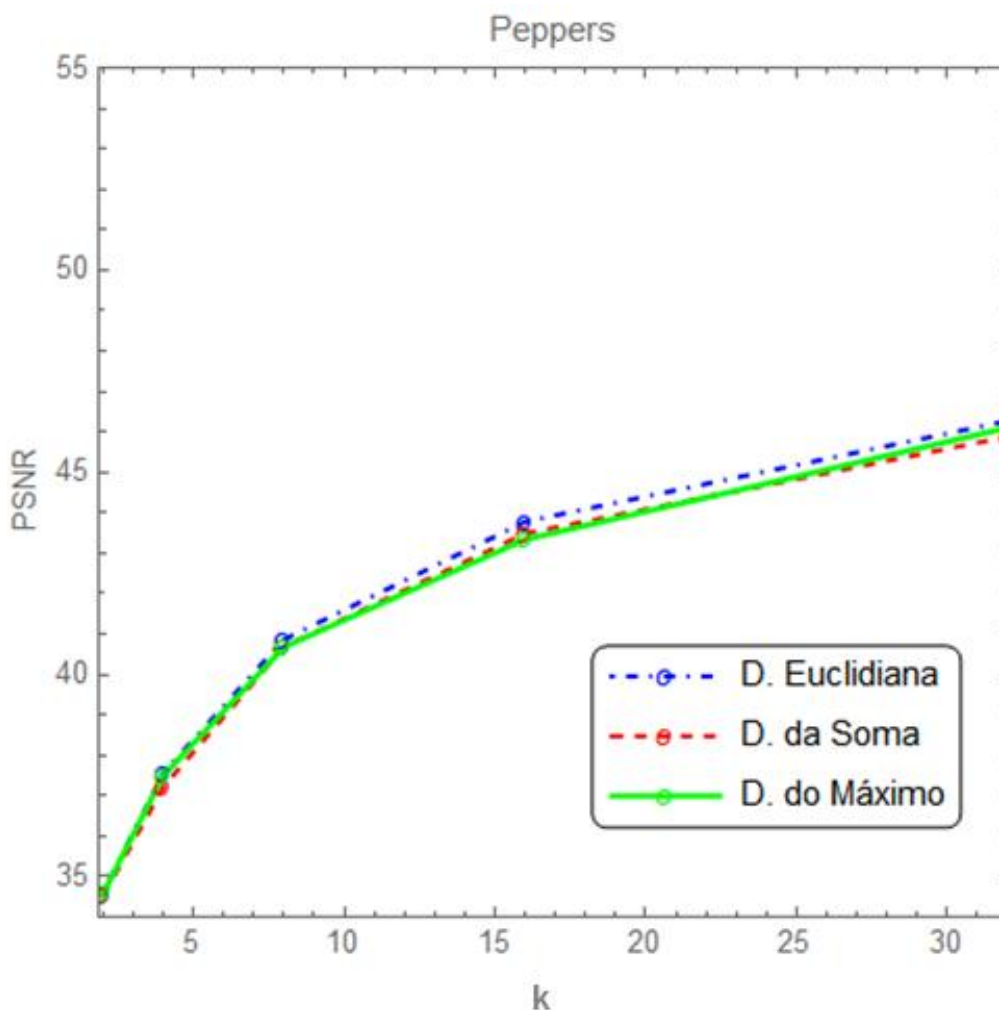
O Gráfico 1, representa os dados obtidos da Lena, pode-se observar que a distância Euclidiana obteve melhores resultados ao realizar segmentação possuindo melhores PSNR's em todos os grupos. Dado curioso foi que nas distâncias da soma e do máximo houve certa variância entre os resultados, onde o agrupamento para uma quantidade de grupos menores (dois, quatro e oito), obtiveram melhores resultados utilizando a distância da soma enquanto o agrupamento para dezesseis e trinta e dois possuíram melhores resultados utilizando a distância do máximo.

Gráfico 2: Dados resultantes do Babbon.



O Gráfico 2 , nos mostra os dados que representam a segmentação do Babbon, observa-se que, quase todos os dados de PSNR da distância Euclidiana foram superiores novamente, com exceção do agrupamento realizado para quatro grupos da distância da soma com dado de 39.3999. Vale ressaltar que, as distâncias da soma e do máximo, novamente, possuíram certa variação, desta vez com certa dominância da distância da soma, uma vez que o agrupamento para dois, quatro, oito e dezesseis grupos realizados foram superiores, porém, a imagem que deveria possuir melhor qualidade, por ser agrupada com uma quantidade de grupos maiores pertenceu à distância do máximo, com valor de PSNR de 47.3262.

Gráfico 3: Dados resultantes da Peppers



Do Gráfico 3, foi constatado a predominância do agrupamento utilizando a distância Euclidiana, uma vez que, todos os seus grupos possuíram melhores PSNR's que as demais distâncias. Diferentemente das outras duas imagens, a distância do máximo possuiu melhores resultados no agrupamento de dois e quatro grupos em relação a distância da soma, porém, ela possuiu melhores resultados nos agrupamentos para oito e dezesseis grupos. Todavia, a distância do máximo voltou a possuir melhores resultados que a da soma no agrupamento para trinta e dois grupos.

Analisando as tabelas e gráficos presentes, pode-se constatar a excelência do agrupamento de dados utilizando a distância Euclidiana em todas as imagens, exceto por uma rara exceção. Em relação às distâncias da soma e do máximo houve bastante variação em seus dados, porém em sua maior parte, foi possível observar que para os agrupamentos utilizando grupos de dados menores (dois, quatro e oito)

a distância da soma obteve melhores resultados no geral, enquanto para os demais grupos (dezesseis e trinta e dois), que devem possuir melhores imagens finais, os melhores resultados foram obtidos pela distância do máximo.

5. Conclusão

Neste trabalho, realizamos um estudo mais aprofundado sobre algoritmos de agrupamento de dados. Descrevemos alguns destes métodos como a técnica de agrupamento hierárquico e o k-médias, assim como algumas de suas aplicações e as definições de centroide. Demonstramos que as distâncias Euclidiana, da soma e do máximo, satisfazem as propriedades para serem uma distância. Além disso, estudamos sobre as definições de segmentação de imagens, sobre seus tipos e aplicações e escolhemos uma delas para abordar no decorrer do trabalho.

No experimento podemos observar que o algoritmo operou de forma satisfatória. Ao realizar comparação de qual seria a melhor distância para o k-médias realizar a segmentação de imagens, constatou-se, assim como em Singh (2013), a distância Euclidiana obteve-se melhores valores de PSNR em relação às distâncias da soma e do máximo.

O programa utilizado para realizar este trabalho (ver anexo) foi o Python 3.6, onde foram desenvolvidas funções que calcula as distâncias, assim como o k-médias e a segmentação de imagens através do cálculo do PSNR. Vale ressaltar que, a depender da imagem e da quantidade de grupos que se deseja realizar a segmentação, o programa demorará certo tempo para ser executado. Este tempo pode variar entre segundos para segmentações de pequenos grupos e 40 minutos a uma hora para segmentações de grupos maiores.

6. Referências

- Archana, S.; Yadav, A.; Rana, A. **K-means with Three different Distance Metrics**. International Journal of Computer Applications (0975 – 8887) Volume 67– No.10, April 2013.
- Bussab, W. O.; Miazaki, E. S.; Andrade, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 105p.
- Doni, M. V. **Análise de cluster: métodos hierárquicos e de particionamento**. São Paulo, 2004.
- Duarte, C. M. R.; Pedroso, M. M.; Bellido, J. G.; Moreira, R. S. **Regionalização e desenvolvimento humano: uma proposta de tipologia de Regiões de Saúde no Brasil**. Rio de Janeiro, 2015.
- Everitt, B. S., Landau, S., and Leese, M., **Cluster 55 Analysis**, Arnold, 4th Edition, 2001.
- Guidini, M. N.; Nascimento, A. M.; Bone, R. B.; Alves, T. W. **Aplicação do k-means cluster para a classificação de estilos gerenciais**. Rio de Janeiro, 2008.
- Haykin, S., **Neural Networks, a Comprehensive Foundation**, Macmillan. New York, NY, 1994
- Halkini, M.; Batistakis, Y.; Vazirgiannis, M. **On Clustering Validation Techniques**, 2001.
- K. Jain, **Data Clustering: 50 Years Beyond K-Means**, Pattern Recognition Letters, 2010.
- Kaufmann, Leonard; Rousseeuw, Peter J. **Finding groups in data: an introduction to cluster analysis**. New York: Wiley, 1990.
- Koegh, E. A. **Gentle Introduction to Machine. Learning and Data Mining for the Database Community**. Manaus, SBBB 2003
- Laurega, Luciane. MeSegHi: **Um Método de Segmentação para o Processamento Linear e Não-Linear de imagens**. **Dissertação de Mestrado** – Programa de Pós-Graduação em Engenharia de Produção (PPGEP) – Universidade Federal de Santa Maria. 2006.
- Lima, E. L. **Espaços Métricos**, 5ª Edição, 2015.

Linden, Ricardo. **Técnicas de Agrupamento** - Revista de Sistemas de Informação da FSMA n. 4 (2009) pp. 18-36

Lloyd, S., **Least squares quantization in PCM**, IEEE transactions on information theory 28.2: 129-137, 1982.

Longhi, Solon Jonas. **Agrupamento e análise fitossociológica de comunidades florestais na sub-bacia hidrográfica do Rio Passo Fundo-RS**. Rio Grande do Sul, 1997.

Morgan, J. **Técnicas de segmentação de imagens na geração de programas para máquinas de comando numérico**. Santa Maria, Rio Grande do Sul, 2008).

Pereira, R. **Archive for the 'PSNR' Tag: Analisando objetivamente a qualidade de um vídeo**, 2008.

Pinele, J.. **Geometria do Modelo Estatístico das Distribuições Normais Multivariadas**. Tese de doutorado. Campinas, 2017.

Prado, T. C. **Segmentação de Imagens Coloridas Utilizando Técnicas de Agrupamento de Dados**. Florianópolis, Santa Catarina, 2008.

Profloresta Agro. **Agrupamento de dados**. Disponível em: <
https://profloresta.agro.ufg.br/up/417/o/Aula_3_An%C3%A1lise_de_agrupamento.pdf?1458264500r > Acesso em: 18 de dezembro de 2018.

Romani, L. A. S.; Gonçalves, R. R. do V.; Amaral, B. F. do; Zullo Junior, J.; Traina Junior, C.; Sousa, E. P. M. de; Traina, A. J. M. **Acompanhamento de safras de cana-de-açúcar por meio de técnicas de agrupamento em séries temporais de NDVI**. Brasília, DF, 2011

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

Totti, R.; Vencovsky, R.; Batista, L. A. R. **Utilização de métodos de agrupamentos hierárquicos em acessos de Paspalum (Graminea (Poaceae))**. São Paulo, 2001.

Yerpude, A.; Dr. Dubey, S. **Colour image segmentation using K – Medoids Clustering**. Rungta College of Engg. & Tech. Bhilai, Chhattishgarh, India, 2012.

Zaiane, O.; Oliveira, S. **Geometric data transformation for privacy preserving Clustering**. Edmonton, Alberta, Canada, 2003.

Anexos

Python 3.6

Algoritmo k-médias.

```

from math import sqrt
import math
import numpy as np
import matplotlib.pyplot as plt
from PIL import Image

#CARREGAMENTO DA IMAGEM E LEITURA DOS PIXELS
imagem1 = Image.open("Lena.png")
pixels = imagem1.load()
(width,height) = imagem1.size
all_pixels = []
for x in range(width):
    for y in range(height):
        cpixel = [pixels[x, y],x,y]
        all_pixels.append(cpixel)

lista_pontos = all_pixels

k = 32

lista_centroides = []

for i in range(k):
    lista_centroides.append(lista_pontos[i][0])
    #print('s=', lista_centroides)

#CRIAÇÃO DA LISTA DE DADOS
lista_grupos = []

for i in range(k):
    lista_grupos.append([])

#FUNÇÃO ARGUMENTO MÍNIMO
def argmin(lista):
    minimo = lista[0]
    n = 0
    for i in range(len(lista)):
        if lista[i] < minimo:
            minimo = lista [i]
            n = i
    return n

```

#FUNÇÃO SUBTRAÇÃO

```

v1= []
v2= []
def subtracao(v1,v2):
    s = []
    for x in range(len(v1)):
        s.append((v1[x] - v2[x]))
    return s

```

#FUNÇÃO DISTÂNCIA MÁXIMO

```

v1=[]
v2=[]
def maximo(v1,v2):
    dist=[]
    for x in range(len(v1)):
        dist.append(abs(v1[x]-v2[x]))

    maximo=max(dist)
    return(maximo)

```

#FUNÇÃO DISTÂNCIA EUCLIDIANA

```

v1=[]
v2=[]
def euclidiana(v1,v2):
    dist = 0.0
    for x in range(len(v1)):
        dist += pow((v1[x] - v2[x]),2)

    eucli = sqrt(dist)
    return eucli
euclidiana(v1,v2)

```

#FUNÇÃO DISTÂNCIA DA SOMA

```

v1=[]
v2=[]
def soma(v1,v2):
    dist1 = 0.0
    for x in range(len(v1)):
        dist1 += abs(v1[x] - v2[x])

    soma= dist1
    return soma

```

#FUNÇÃO CENTRÓIDE

```

def centroide(lista):
    tamanho_lista = max(len(lista),1)
    soma_x = 0
    soma_y = 0
    soma_z = 0

    for i in range(tamanho_lista):
        if lista != []:
            soma_x = soma_x + lista[i][0][0]
            soma_y = soma_y + lista[i][0][1]
            soma_z = soma_z + lista[i][0][2]

```

```

        centroide =
[soma_x/tamanho_lista,soma_y/tamanho_lista,soma_z/tamanho_lista]
        return centroide

lista_distancias = []
centroide_anterior = []
centroide_final = lista_centroides

#FUNÇÃO RGB
def vetor_rgb(lista):
    new_lista = lista
    for i in range(len(lista)):
        new_lista[i] = lista[i][0]
    return new_lista

#EXECUÇÃO DO "K-MEANS"
while centroide_anterior != centroide_final:

    lista_grupos = []

    for i in range(k):
        lista_grupos.append([])

    for i in range(len(lista_pontos)):
        lista_distancias = []
        for j in range(k):

lista_distancias.append(maximo(lista_centroides[j],lista_pontos[i][0]))
        #print("dist=",lista_distancias)
        lista_grupos[argmin(lista_distancias)].append(lista_pontos[i])
        #print("grupos=",lista_grupos)

    centroide_anterior = centroide_final
    lista_centroides = []
    lista_grupos_final = lista_grupos

    for l in range(k):
        lista_centroides.append(centroide(lista_grupos[l]))
    centroide_final = lista_centroides
    print("c_tipol=",lista_centroides)

#print("gruposfinal=",lista_grupos_final)

#print(len(lista_grupos))

#FUNÇÃO PARA TIRAR AS COORDENADAS DA LISTA DE PIXELS;
def troca_rgb(lista,rgb):
    for i in range(k):
        for j in range(len(lista[i])):
            lista[i][j][0] = rgb[i]
    return lista

new_lista_grupos = troca_rgb(lista_grupos,lista_centroides)

```

```

#print("new=",new_lista_grupos)

#MATRIZEZ DE ZEROS (PARA "A" E "B")
m = width

A = [0]*m
B = [0]*m

for i in range(m):
    A[i] = [0]*m
    B[i] = [0]*m

lista=[]

#TRANSFORMAR A LISTA "new_lista_grupos" EM UMA ÚNICA LISTA;
for i in new_lista_grupos:
    lista.extend(i)

for i in range(width*height):
    A[lista[i][1]][lista[i][2]] = lista[i][0]

#print('A= ',A)

#TRANSPOSTA DA MATRIZ A;
for i in range(len(A)):
    for j in range(len(A)):
        B[i][j] = A[j][i]

#MOSTRAR A IMAGEM
image= np.asarray(B,dtype=np.uint8)

plt.imshow(image)
plt.show()

#CÁLCULO DO PSNR
mse= 0.0

for i in range(width):
    for j in range(width):
        mse = mse +
pow(euclidiana(subtracao(pixels[i,j],A[i][j]), [0,0,0]),2)

mse = mse / (256 * 256 * 256)

PSNR = 10*math.log10(pow(255,2)/mse)

print('PSNR=' , PSNR)

```