UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE MATEMÁTICA

Bruna da Silva Donadel

ANÁLISE DE FUNÇÕES DE MEDIDA PARA O MÉTODO K-MEANS

Florianópolis

2018

Bruna da Silva Donadel

ANÁLISE DE FUNÇÕES DE MEDIDA PARA O MÉTODO K-MEANS

Trabalho de Conclusão de Curso apresentado ao Curso de Matemática, do Departamento de Matemática - Centro de Ciências Físicas e Matemáticas da Universidade Federal de Santa Catarina, para obtenção de grau de Licenciada em Matemática.
Orientador: Prof. Dr. João Artur de Souza

Florianópolis

2018

Bruna da Silva Donadel

ANÁLISE DE FUNÇÕES DE MEDIDA PARA O MÉTODO K-MEANS

Este Trabalho de Conclusão de Curso foi julgado adequado para a obtenção do Título de "Licenciada em Matemática", e aprovado em sua forma final pela Matemática - Licenciatura.

Florianópolis, 13 de outubro 2018.

Profa. Dra. Sonia Elena Palomino Castro
Coordenadora do Curso

Banca Examinadora:

Prof. Dr. João Artur de Souza
Orientador

Profa. Dra. Silvia Martini de Holanda

Prof. Me. Alessandro Costa Ribeiro

Aos meus pais,
Ivone da Silva Donadel e Joelso Donadel.

AGRADECIMENTOS

Primeiramente agradeço a Deus por fazer em minha vida muito mais do que pedi ou pensei.

Agradeço também a minha família, que é minha base e meu porto seguro. Tudo que sou também é mérito deles. Obrigada mãe, pelas ligações diárias e por cada gesto de cuidado; obrigada pai, pelo apoio incondicional; e obrigada manos, sem vocês a vida não teria mesma graça.

Gratidão a OBMEP pela parceria de longas datas e, juntamente com o Instituto TIM, pela bolsa que me sustentou durante esses anos de graduação longe de casa.

Sou grata também àquele que acompanha de perto meus dramas e conquistas: meu namorado, meu guardião. Obrigada por sempre ter me apoiando, ajudando e por estar fazendo meus dias mais leves e felizes.

Meu muito obrigada ao professor João Artur de Souza, que me abriu novos horizontes e, mesmo sem tanto tempo, aceitou o compromisso de orientar nesse trabalho.

Gratidão ao Alessandro Costa Ribeiro, que me acompanhou e ajudou no projeto inicial desse trabalho, como também na versão final dele.

Meu carinho a cada um dos professores que fizeram parte da minha história na UFSC, especialmente ao professores José Luiz Rosas Pinho pelos anos no PET, Leandro Batista Morgado pelos quatro semestres maravilhosos, Carmem Suzane Comitre Gimenez por provocar as reflexões certas, Alda Dayana Mattos Mortari pelo exemplo de profissional, Silvia Martini de Holanda pela disponibilidade e por seus conselhos, e Leonardo Silveira Borges por não medir esforços para me ajudar quando foi preciso.

Agradeço também aos parceiros de graduação André Borges (que esteve ao meu lado nos melhores e piores momentos), Letícia Carvalho, Victória Gittens, Ana Carolina Altomani, Gabriela Rodrigues, Carlos Leal, Gabriel Schafaschek, Sabrina Vigano e Lara Lopes de Miranda pela companhia nas aulas, no PET e também fora do ambiente acadêmico.

Minha gratidão e carinho por minha química favorita, Bruna Inácio Trajano, que pacientemente responde às minhas perguntas bobas e que dividiu o teto comigo por esses anos com muito bom humor. Gratidão também a família dela pelo carinho, me tratando quase como filha.

Gratidão aos professores do Programa de Iniciação Científica Jr. da OBMEP, Lucas Spillere Barchinski e Elizete Maria Possamai Ribeiro, pelo exemplo e incentivo a seguir por esse caminho.

E por último, mas não menos importante, agradeço aos professores e amigos da minha terra natal, Timbé do Sul, e a cada um que de alguma forma fez e faz parte da minha história.

"Nenhum problema pode ser resolvido pelo mesmo estado de consciência que o criou."
Albert Einstein

RESUMO

Na era do conhecimento, obter conhecimento de dados das mais variadas fontes pode representar o diferencial que um prestador de serviços ou pesquisador precisa para se destacar ou dar continuidade em suas pesquisas. Um importante procedimento para mineração de dados é a clusterização, que por sua vez necessita de funções de medida para que possa ser executada. Neste trabalho, avaliou-se a influência que a escolha de uma medida de distância, similaridade ou dissimilaridade tem sobre o resultado da clusterização obtida através do método *k-means*. As funções de medida testadas foram a distância euclidiana ao quadrado, a distância Manhattan, a dissimilaridade pelo cosseno e a dissimilaridade por correlação e concluiu-se que diferentes conjuntos de dados, em suas particularidades, se adequam melhor a diferentes medidas de distância quando clusterizados pelo método *k-means*.

Palavras-chave: Clusterização, K-means, Medida de Distância, Medida de Dissimilaridade.

ABSTRACT

In the knowledge age, obtaining knowledge of data from a wide variety of sources can represent the differential that a service provider or researcher needs to excel or to continue with their research. An important procedure for data mining is clustering, which in turn requires measurement functions to be performed. In this work, the influence that the choice of a measure of distance, similarity or dissimilarity has on the result of the clustering obtained through the k-means method was evaluated. The measurement functions tested were the square Euclidean distance, the Manhattan distance, the dissimilarity by the cosine and the dissimilarity by correlation and it was concluded that different data sets, in their particularities, are better adapted to different distance measures when clustered by the k-means method.

Keywords: Clustering, K-means, Distance Measure, Dissimilarity Measure.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

Vivemos na era do conhecimento, onde a competitividade e globalização exigem qualidade em informação e conhecimento útil para que empresas e pessoas possam se destacar e sobreviver no mercado (SANTOS; BASTOS, 2017). Nesse aspecto, usufruir do grande volume de dados que vem sendo armazenado, graças ao barateamento de *hardwares* nas últimas décadas e do avanço tecnológico no desenvolvimento de estruturas de armazenamento (CAMILO; SILVA, 2009), pode ser a estratégia decisiva para o sucesso.

Apesar de benéfica, a revolução da informação é desafiadora. Um dos desafios é como extrair informação e conhecimento útil desta vasta gama de dados. Por isso são desenvolvidas diversas ferramentas para tratá-los e analisá-los. Essas ferramentas basicamente são algoritmos automatizados por softwares computacionais.

Uma maneira natural para tentar entender um aglomerado de informações é dividi-lo em partes menores e capturar a ideia geral de cada uma (JAIN; DUBES, 1988). Tratando-se de Mineração de dados, esse processo de dividir os dados em grupos menores para análise é conhecido por Clusterização (CASSIANO, 2014), mas também pode ser encontrado na literatura como Análise de Agrupamento, *Clustering, Classification Analysis* ou *Numerical Taxonomy* (CASSIANO, 2014).

Os métodos de clusterização têm sido usados em diversos problemas, tais como: reconhecimento de padrões de compra para detectar perfis de clientes, processamento de imagens médicas para fornecer uma identificação genética de doenças, análise de dados, análise de sintomas de doenças, funcionalidades de genes, aspectos da personalidade de indivíduos, *marketing*, segmentação de imagens, entre outras (CASSIANO, 2014; GRABUSTS, 2011).

Basicamente, a clusterização é um estudo formal de algoritmos e métodos de agrupamento (JAIN; DUBES, 1988), sendo o objetivo desses algoritmos identificar subconjuntos da base inicial, chamados de *clusters*, de tal forma que os elementos (também chamados de observações, registros ou objetos) de um mesmo *cluster* sejam mais similares ou próximos entre si do que dos elementos de outros *clusters*, ou seja, que possuam maior homogeneidade *intracluster*, e simultaneamente que os elementos de diferentes *clusters* sejam o mais dissimilares ou distantes possível, ou seja, que possuam máxima heterogeneidade *intercluster* (CASSIANO, 2014).

Então é natural a dúvida: como medir a dissimilaridade ou distância entre os registros? A resposta dessa questão é parte fundamental para aplicação de um dado algoritmo de agrupamento e influencia diretamente o resultado obtido.

Por trás da ideia de medir a distância ou a dissimilaridade entre dados de um conjunto existe uma função, mas considerando-se que há diversas formas possíveis de definir uma função de medida. Este trabalho tem o intuito de analisar comparativamente o desempenho da distância euclidiana ao quadrado, distância Manhattan, dissimilaridade pelo cosseno e dissimilaridade por correlação no método de clusterização *k-means*. E levando em conta que há uma infinidade de métodos e que outros têm sido criados, seria inviável testar todos, por esse motivo o estudo será feito em um método específico, a saber, o método *k-means*, escolhido por conta da sua eficiência computacional e facilidade de implementação (JAIN, 2009).

1.1 DELIMITAÇÃO DO TEMA

Neste trabalho pretende-se analisar a influência que diferentes funções de medidas causam no desempenho do algoritmo de clusterização *k-means*, as funções escolhidas para o estudo são a distância euclidiana ao quadrado, a distância Manhattan e a dissimilaridade pelo cosseno e a dissimilaridade por correlação de Pearson. E os conjuntos de dados ou "*datasets*" utilizados como referência de comparação são o tradicional Iris (disponível no repositório da Universidade da Califórnia em Irvine – *UCI Machine Learning Repository*), o grupo de dados Vinho (também disponível no repositório da UCI, registrado com o nome *Wine data set*) e um conjunto de dados de cromossomos humanos.

O trabalho será desenvolvido com auxílio dos softwares MatLab e RStudio, que já têm algumas funções necessárias para o processamento dos dados pré-programadas.

1.2 PROBLEMA

O primeiro trabalho publicado sobre método de Clusterização data de 1948 (CASSIANO, 2014), desde então inúmeros algoritmos e variações deles têm sido criados, porém nenhum é universalmente apropriado. Por esse motivo diversos métodos continuam em constante estudo na possibilidade de melhorar. Nesse sentido, essa pesquisa busca responder à seguinte pergunta: qual a relevância da escolha da função de medida para o método de clusterização *k-means*?

1.3 OBJETIVOS

1.3.1 Objetivo geral

Este trabalho pretende analisar comparativamente o desempenho da distância euclidiana ao quadrado, distância Manhattan, dissimilaridade pelo cosseno e dissimilaridade por correlação no método de clusterização *k-means*.

1.3.2 Objetivos específicos

1. Apresentar os conceitos gerais sobre Clusterização encontrados na literatura;
2. Apresentar o método *k-means*;
3. Definir os conceitos de medida de distância e medida de dissimilaridade e descrever quais serão utilizadas neste trabalho;
4. Comparar os agrupamentos com cada uma das medidas de distância propostas no método *k-means*;

1.4 JUSTIFICATIVA

Organizar dados em grupos é uma das formas mais naturais de compreensão e aprendizagem (JAIN; DUBES, 1988). Entretanto, esse problema torna-se difícil à medida que a quantidade de objetos a ser agrupados aumenta, por isso é necessário recorrer a ferramentas computacionais.

Diz-se que métodos de clusterização sem sobreposição são algoritmos que particionam um conjunto M em k subconjuntos $U_1, U_2, \cdots, U_k$, tais que $\cup_{i=1}^{k} U_i = M$, $U_i \cap U_j = \varnothing$ sempre que $i \neq j$ e $U_i \neq \varnothing$, para $i, j \in \{1, 2, \cdots, k\}$. Nesse caso, segundo Hruschka e Ebecken (2003), a quantidade de possíveis clusterizações distintas é dada por:

$$NW(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \cdot C_i^k \cdot (k-i)^n \tag{1.4.1}$$

onde n é o número de elementos do conjunto M, k a quantidade de *clusters* que se deseja obter e C_i^k é o número total de combinações de k elementos tomados i a i, dado por:

$$C_i^k = \frac{k!}{i!(k-i)!}. \tag{1.4.2}$$

Por exemplo, um conjunto de 25 elementos pode ser dividido em 3 *clusters* de

$NW(25, 3) = 141.197.991.025$ maneiras diferentes. A quantidade de partições de n dados em k grupos aumenta com uma razão de aproximadamente $\frac{k^n}{k!}$, então comparar todas para identificar a melhor partição torna-se inviável para *datasets* "volumosos" (HRUSCHKA; EBECKEN, 2003). Tratando-se de complexidade computacional, Hruschka e Ebecken (2003) afirma que esse problema é NP-completo. Basicamente, problemas computacionais podem ser classificados como P, NP, NP-difícil e NP-completo, esta última categoria abrange os problemas que estão em NP e que admitem que qualquer outro problema em NP possa ser reduzidos a eles em tempo polinomial, para maiores esclarecimentos e definições recomenda-se a obra de Sipser (2005).

Enfim, ao mesmo tempo em que a clusterização é útil para identificar perfis, nichos ou grupos e revelar suas características, propiciando um aprendizado sobre os dados em questão, há dificuldade em manipular grandes *datasets* para este fim, por esse motivo diversos estudos continuam sendo produzidos nessa área (GRABUSTS, 2011).

Considerando que a clusterização, enquanto estudo, nos fornece diversos métodos para fazer agrupamentos, mas independente do método que for escolhido é necessário estipular uma medida de similaridade, dissimilaridade ou distância entre pares de elementos para verificar quão próximos estão dois objetos. É de interesse a busca pela influência que funções de medida causam em determinado método de clusterização, tanto é que esse assunto tem recebido a atenção no processamento inteligente de dados pelo fato de que pode ser aplicado em diversos problemas, ainda assim parece pouco explorado.

1.5 METODOLOGIA DA PESQUISA

Este trabalho apresenta uma pesquisa quantitativa, pois emprega técnicas numéricas/estatísticas a fim de responder a pergunta de pesquisa proposta na seção 1.2 e alcançar os objetivos traçados (RICHARDSON et al., 2007).

É realizada a revisão da literatura para introduzir o leitor a temática da pesquisa e aprofundar os conhecimentos a respeito de Clusterização de dados, do método *k-means*, das distâncias propostas e da validação dos resultados.

Perante o conhecimento adquirido, é realizada a análise do desempenho de método *k-means* com variações na medida de distância utilizada. Por fim, apresenta-se as conclusões e perspectivas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentadas informações, conceitos e definições relacionadas à Clusterização de dados, ao método de clusterização *k-means*, à funções de medida e à validação dos resultados.

2.1 INTRODUÇÃO

Levando em conta que atualmente muito tem se investido para extração e armazenamento de dados, é necessário ter ferramentas capazes de manipulá-los a fim de gerar informação útil. Por essa perspectiva surge a Mineração de Dados, unindo técnicas de estatística, computação, reconhecimento de padrões e visualização para esse fim (CAMILO; SILVA, 2009).

Documentos mais antigos que tratam da Clusterização de Dados geralmente não diferenciam Clusterização de Classificação, tratando a Clusterização até mesmo como um tipo de Classificação. Entretanto, o conceito atual faz uma separação nítida entre essas duas ferramentas da Mineração de Dados. A Análise de Agrupamento é tida como o método que divide conjuntos de dados em subconjuntos com características internas semelhantes sem ser necessário qualquer informação prévia sobre como eles devem ser, enquanto a Classificação é o processo em que se determina a que grupo um determinado novo dado melhor se adequa, nesse caso é obrigatório conhecer previamente os grupos possíveis ou saber suas características. Portanto, a Clusterização pode ser vista como o processo anterior a Classificação (CASSIANO, 2014).

Considerando que na Análise de Agrupamento, em tese, não há classes previamente conhecidas ou informações sobre, o usuário não manipula os resultados do método, por isso ele é dito não supervisionado. Resumidamente, a Análise de Agrupamento consiste em técnicas numéricas que agrupam automaticamente, de tal forma que elementos de um mesmo *cluster* sejam mais similares ou próximos entre si do que dos elementos de outros *cluster*, ou seja, que possuam maior homogeneidade *intracluster*, e simultaneamente que os elementos de diferentes *clusters* sejam o mais dissimilares ou distantes possível, isto é, que apresentem máxima heterogeneidade *intercluster* (CASSIANO, 2014)

Há diversos métodos de Clusterização, a saber, métodos hierárquicos, particionais, baseados em densidade, baseados em grafos, baseados em redes neurais, baseados na lógica Fuzzy e outros. Uma das categorias mais conhecidas é a dos métodos particionais, onde o usuário define a quantidade de agrupamentos que se deseja obter e então o método

converge iterativamente para a partição "ótima" dentro das condições iniciais (CASSIANO, 2014). O método *k-means* é o mais comum dos métodos particionais, como já dito, sua implementação é fácil, eficiente e tem bom desempenho (JAIN, 2009), além de possibilitar variações, por exemplo, mudando a função de medida usada no método.

Cabe aqui salientar que o *k-means* possui duas versões: a versão clássica proposta por Lloyd e a versão proposta posteriormente por Hartigan (SLONIM; AHARONI; CRAMMER, 2013). Sobre esses métodos, autores como Slonim, Aharoni e Crammer (2013) e Elsheikha et al. (2016) argumentam em seus trabalhos que o método de Hartigan é melhor, pois seu conjunto de mínimos locais é um subconjunto dos mínimos locais de Lloyd, portanto ele geralmente converge para resultados melhores.

Na literatura encontram-se trabalhos que têm investigado o desempenho de diferentes métodos de clusterização, como por exemplo, o artigo de Shi, Wang e Zhang (2017), onde é feita uma comparação de 7 métodos, incluído o *k-means*. Nesse estudo conclui-se que para seus testes 3 métodos (*affinity propagation* (AP), *density peak based clustering* (DP) e DBSCAN) geralmente apresentam melhores resultados em relação aos outros e a escolha entre eles depende do *dataset* que se deseja clusterizar. Entretanto, é destacado que, a depender da função de validação dos resultados, o *k-means* pode ser considerado tão bom quanto AP e DP, além de ser mais eficiente por ser mais rápido que os anteriores.

De modo geral, não há um consenso a respeito de como deve-se proceder quando o assunto é clusterização, apesar de haver essa gama de pesquisas em relação ao assunto.

Nos tópicos seguintes são abordados com mais rigor o método *k-means*, as funções de medida utilizadas e o critério de comparação dos resultados que serão base dessa pesquisa.

2.2 MÉTODO K-MEANS

Tendo um conjunto de dados disposto em uma matriz $n \times d$, onde cada um dos n vetores horizontais representa uma observação contendo d características ou atributos, pode-se interpretar estes vetores como pontos num espaço d-dimensional. O algoritmo na versão de Lloyd, originalmente com a distância euclidiana, procede da seguinte maneira:

(i) o usuário ou um algoritmo especialista define a quantidade k de *clusters* em que sua base deve ser dividida (CARDOSO et al., 2008); então escolhe-se k pontos para serem os centros iniciais dos *clusters*, nessa etapa pode-se escolher os centros manualmente, aleatoriamente ou optar por algum algoritmo de escolha;

(ii) calcula-se a distância de cada observação até cada um dos centros, então cada um dos pontos é associado ao centro mais próximo, formando os clusters;

(iii) reposicionam-se os centros para a média dos pontos de cada agrupamento (não sendo necessário que esse ponto médio coincida com algum já existente na base de dados);

(iv) calcula-se a função objetivo, que geralmente é o erro quadrático E dado pela soma de erros locais E_j, para $j = 1, 2, \cdots, k$ (CASSIANO, 2014):

$$E = \sum_{j=1}^{k} E_j = \sum_{j=1}^{k} \sum_{x \in C_j} ||x - m_j||^2, \tag{2.2.1}$$

onde C_j é o j-ésimo *cluster*, x é um vetor de C_j e m_j é o centro dele;

(v) retorna-se ao segundo passo e reinicia o processo. O algoritmo acaba quando se minimiza o erro E ou quando se atinge um limite de iterações, caso o usuário o defina.

Perceba que, por construção, o erro não pode aumentar com o decorrer das iterações, assim o método converge para uma partição de ponto fixo após uma quantidade finita de iterações (SLONIM; AHARONI; CRAMMER, 2013), ou seja, o algoritmo termina quando os centros dos *clusters* param de sofrer alterações. Porém uma partição inicial "ruim" pode levar o método a convergir para um ponto de mínimo local (JAIN; DUBES, 1988; SLONIM; AHARONI; CRAMMER, 2013), assim, para se obter melhor resultado é indicado que se aplique o algoritmo várias vezes com centros iniciais distintos, pois assim há maior chance de encontrar uma partição de mínimo global.

As versões desse algoritmo em que há mudança na medida de distância podem sofrer alterações, como por exemplo redefinição do item (iii). Neste trabalho as mudanças adotadas serão apresentadas posteriormente.

Para ilustrar, aplicou-se o método no conjunto de dados bidimensional, obtido a partir das colunas 3 e 4 do *dataset* Iris do repositório da UCI (2018), e $k = 3$. Abaixo é representado o espaço de atributos e na sequência a convergência do método:

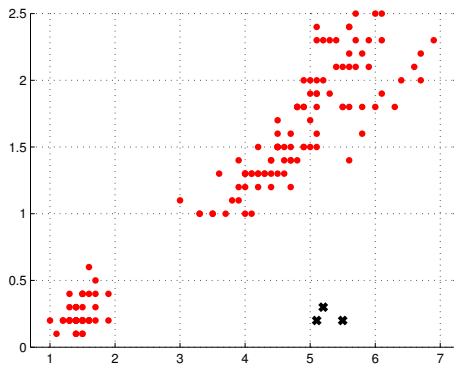
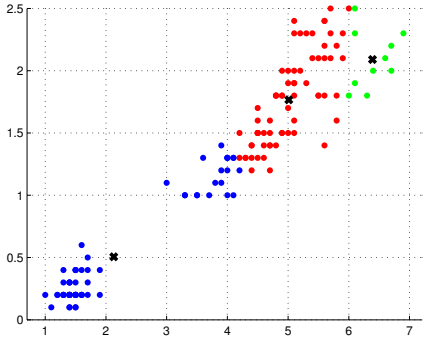
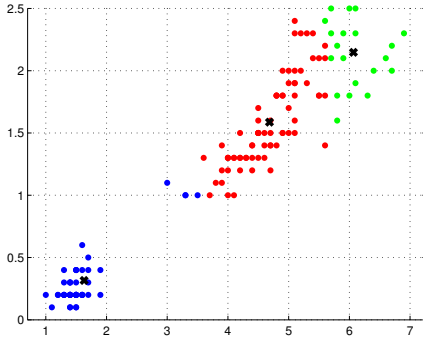


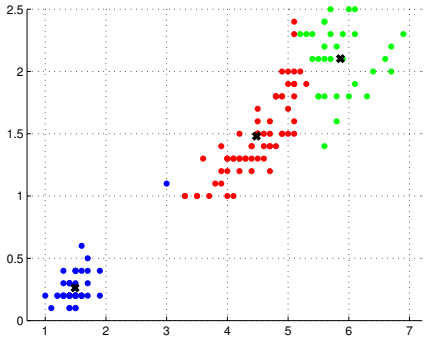
Figura 1. Espaço de atributos antes da aplicação do método



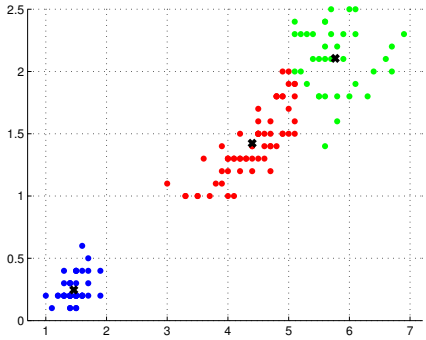
a) 1ª iteração



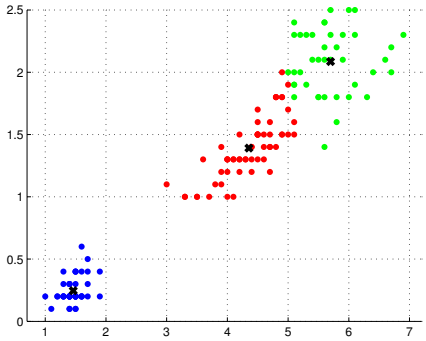
b) 2ª iteração



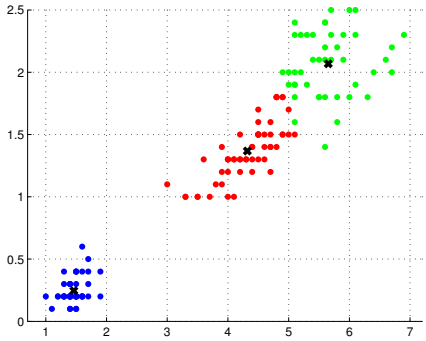
c) 3ª iteração



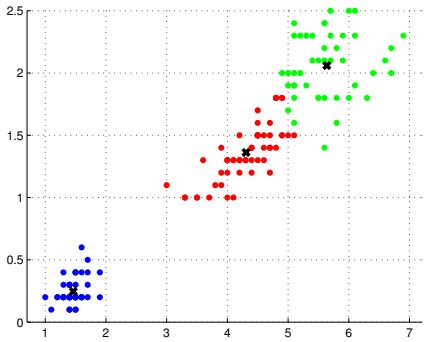
d) 4ª iteração



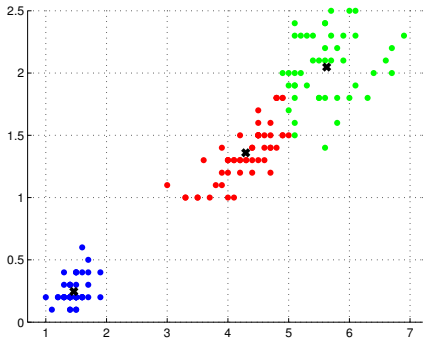
e) 5ª iteração



f) 6ª iteração



g) 7ª iteração



h) 8ª iteração

Figura 2. Iterações do método *k-means* com distância euclidiana ao quadrado

Nesse caso, a matriz de dados é 150×2 e usou-se $k = 3$, com os centros iniciais de coordenadas (5,2; 0,3), (5,5; 0,2) e (5,1; 0,2), indicados como "x" preto na figura 1.

Após a primeira iteração (figura 2.a), temos uma divisão inicial dos grupos que é representada pelas diferentes cores dos pontos e os centros são reposicionados para a posição média dos pontos de seus respectivos grupos, nesse caso suas coordenadas serão (5,01; 1,76), (6,39; 2,09) e (2,13; 0,50) e estão representadas pelos "x" pretos.

A cada iteração esse processo é refeito: os grupos vão sendo reorganizados e os centros recalculados. Após a oitava iteração não há mais mudanças nos grupos, ou seja, o algoritmo acaba.

2.3 MEDIDAS DE DISTÂNCIA OU DISSIMILARIDADE

Em um conjunto não vazio M qualquer, a noção de distância, similaridade ou dissimilaridade entre objetos desse espaço é generalizada por uma função d que, sob algumas condições, associa a cada par ordenado $(x, y) \in M$ um número real. Mais comumente, a noção de distância é associada à *métrica* (SANTANA, 2012), que por sua vez gera apenas valores não negativos, se anula estritamente quando os pontos em questão são equivalentes, é simétrica e respeita a desigualdade triangular. Formalmente, Lima (2009) define:

Definição 2.3.1. *Uma métrica num conjunto M é uma função*

$$d : M \times M \to \mathbb{R}$$

que satisfaz as seguintes condições para quaisquer $x, y, z \in M$:
Condição 1: $d(x, x) = 0$;
Condição 2: se $x \neq y$, então $d(x, y) > 0$;
Condição 3: $d(x, y) = d(y, x)$;
Condição 4: $d(x, z) \leq d(x, y) + d(y, z)$.

Há outras formas de definir distância, por exemplo, "enfraquecendo" a condição 2 da definição acima para "*se $x \neq y$, então $d(x, y) \geq 0$*", neste caso tem-se uma *pseudo-métrica* (LIMA, 2009; SANTANA, 2012; GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012). Também existem outros autores, como Stasiu (2007), que preferem chamar de medida similaridade, ao invés de medida de distância, as função em dado espaço M não vazio que retorna valores dentro do intervalo fechado, por exemplo o intervalo $[0, 1]$ e que "tem por finalidade atribuir uma medida do grau de semelhança entre dois objetos", enquanto função de dissimilaridade tem por objetivo indicar o grau de diferença entre pares.

Na sequência são apresentadas algumas funções de distância e dissimilaridade aplicadas aos pontos $x = (x_1, x_2, \cdots, x_n)$ e $y = (y_1, y_2, \cdots, y_n)$ de um espaço com n dimensões.

2.3.1 Distância Euclidiana

Provém da ideia natural de medida de distância entre dois pontos. No espaço bidimensional é obtida através do Teorema de Pitágoras, que indica comprimento do segmento de reta que os liga, nesse caso a distância entre os pontos $x = (x_1, x_2)$ e $y = (y_1, y_2)$ é

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}.$$

A generalização dessa medida para o espaço n-dimensional é dada por:

$$\begin{aligned} d_1(x, y) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \\ &= \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2} \end{aligned} \tag{2.3.1}$$

Para monstrar que essa função d_1 atende a quarta condição de métrica precisa-se do seguinte resultado:

Proposição 2.3.1. *(Desigualdade de Cauchy-Shwartz) Sejam $x, y \in \mathbb{R}^n$, então*

$$\left(\sum_{i=1}^{n} x_i \cdot y_i\right)^2 \leq \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2. \tag{2.3.2}$$

Demonstração: Se y é nulo, o resultado é imediato. Caso contrário, então para $t \in \mathbb{R}$, defini-se

$$f(t) =: \sum_{i=1}^{n}(x_i + ty_i)^2. \tag{2.3.3}$$

Logo, $f(t) \geq 0$, agora note que

$$f(t) = \sum_{i=1}^{n}(x_i + ty_i)^2 = \sum_{i=1}^{n} x_i^2 + 2t\sum_{i=1}^{n} x_i y_i + t^2\sum_{i=1}^{n} y_i^2,$$

portanto, $f(t)$ é um polinômio de segundo grau com coeficientes $a = \sum_{i=1}^{n} y_i^2 > 0$, $b = 2 \cdot \sum_{i=1}^{n} x_i y_i$ e $c = \sum_{i=1}^{n} x_i^2$, então a desigualdade $f(t) \geq 0$ para todo $t \in \mathbb{R}$ só é satisfeita quando

$$b^2 - 4ac \leq 0 \Rightarrow 4 \cdot \left(\sum_{i=1}^{n} x_i \cdot y_i\right)^2 - 4 \cdot \sum_{i=1}^{n} y_i^2 \cdot \sum_{i=1}^{n} x_i^2 \leq 0,$$

logo,

$$\left(\sum_{i=1}^{n} x_i \cdot y_i\right)^2 \leq \sum_{i=1}^{n} y_i^2 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2 \ .$$

■

Finalmente, verifica-se abaixo que a função d_1 obedece às condições de métrica:

Condição 1:

$$\begin{aligned} d_1(x,x) &= \sqrt{(x_1-x_1)^2+(x_2-x_2)^2+\cdots+(x_n-x_n)^2} \\ &= \sqrt{0+0+\cdots+0} \\ &= 0 \end{aligned}$$

Condição 2:

Supondo que $x \neq y$, então existe pelo menos um $i \in \{1,2,\cdots,n\}$ tal que $x_i \neq y_i$, logo $(x_i-y_i)^2 > 0$. Portanto

$$\begin{aligned} d_1(x,y) &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \\ &\geq \sqrt{(x_i-y_i)^2} \\ &> 0 \end{aligned}$$

Condição 3:

$$\begin{aligned} d_1(x,y) &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \\ &= \sqrt{(y_1-x_1)^2+(y_2-x_2)^2+\cdots+(y_n-x_n)^2} \\ &= d_1(y,x) \end{aligned}$$

Condição 4: Note que $d_1(x,z)$, $d_1(x,y)$ e $d_1(y,z)$ são reais não negativos, então $d_1(x,z) \leq d_1(x,y)+d_1(y,z) \Leftrightarrow \big(d_1(x,z)\big)^2 \leq \big(d_1(x,y)+d_1(y,z)\big)^2$. Veja a seguir a demonstração desta segunda desigualdade:

$$\begin{aligned} \big(d_1(x,z)\big)^2 &= \sum_{i=1}^{n}(x_i-z_i)^2 \\ &= \sum_{i=1}^{n}\big((x_i-y_i)+(y_i-z_i)\big)^2 \\ &= \sum_{i=1}^{n}\big((x_i-y_i)^2+2(x_i-y_i)(y_i-z_i)+(y_i-z_i)^2\big) \\ &= \sum_{i=1}^{n}(x_i-y_i)^2+2\cdot\sum_{i=1}^{n}(x_i-y_i)(y_i-z_i)+\sum_{i=1}^{n}(y_i-z_i)^2 \\ &\leq \sum_{i=1}^{n}(x_i-y_i)^2+2\sqrt{\big(\Sigma_{i=1}^{n}(x_i-y_i)\cdot(y_i-z_i)\big)^2}+\sum_{i=1}^{n}(y_i-z_i)^2 \\ &\leq \sum_{i=1}^{n}(x_i-y_i)^2+2\sqrt{\big(\Sigma_{i=1}^{n}(x_i-y_i)^2\big)\cdot\big(\Sigma_{i=1}^{n}(y_i-z_i)^2\big)}+\sum_{i=1}^{n}(y_i-z_i)^2 \\ &= \left(\sqrt{\Sigma_{i=1}^{n}(x_i-y_i)^2}+\sqrt{\Sigma_{i=1}^{n}(y_i-z_i)^2}\right)^2 \\ &= \big(d_1(x,y)+d_1(y,z)\big)^2, \end{aligned}$$

Portanto, vale que

$$d_1(x,y) \leq d_1(x,y)+d_1(y,z)\ .$$

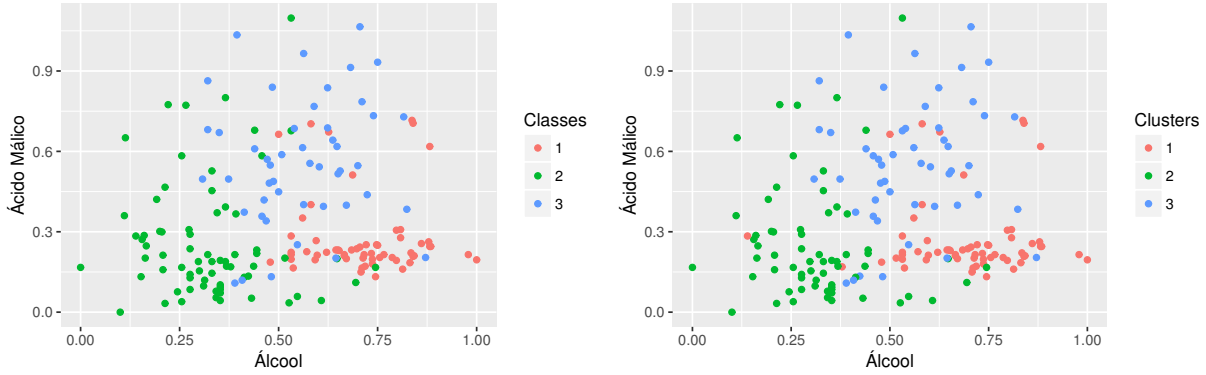
Como todas as condições foram satisfeitas, conclui-se que a função d_1 é uma métrica.

Outra distância, derivada dessa, é a distância euclidiana ao quadrado dada por

$$\begin{aligned} d_{1.1}(x,y) &= (x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2 \\ &= \sum_{i=1}^{n}(x_i-y_i)^2. \end{aligned} \tag{2.3.4}$$

Essa última é mais simples que a primeira, pois não exige a extração de uma raiz quadrada, assim torna-se uma alternativa de menor custo e é comumente usada em métodos de clusterização. Porém, essa função não satisfaz a desigualdade triangular, um contra exemplo em $\mathbb{R}$ é o seguinte: $d_{1.1}(1,3) = 4 > 2 = d_{1.1}(1,2) + d_{1.1}(2,3)$. Portanto, a função $d_{1.1}$ não é métrica.

Em um exemplo para teste, aplicou-se o método *k-means* com distância euclidiana ao quadrado para clusterizar o conjunto dos dados normalizado do conjunto Vinho – disponível no repositório da UCI. O resultado obtido teve 96,63% de acerto, veja a seguir uma representação das categorias originais dos vinhos e os *clusters* obtidos com o método:



a) Grupos originais

b) *Clusters* obtidos

Figura 3. Comparação das categorias originais de vinho e dos *clusters* obtidos pelo método *k-means* com distância euclidiana ao quadrado

O conjunto de dados em questão possui 178 observações (pontos) com 13 atributos cada, mas no gráfico apresenta-se as coordenadas relativas a apenas dois atributos, a saber, quantidade de Álcool e Ácido Málico presente nos vinhos. Originalmente ele apresenta 3 classes de vinho que possuem, respectivamente, 59, 71 e 48 exemplares. Após a clusterização dos dados, o grupo 1 obteve 62 elementos, o grupo 2 obteve 65 e o grupo 3 obteve 51, sendo que 6 desses pontos aparecem em algum grupo diferente do que corresponde a classe original deles.

2.3.2 Distância Manhattan

Também conhecida como "distância *city block*", porque "calcula a distância ao longo de um caminho semelhante ao modo como nos movemos em uma cidade em vez da menor distância" (QIN; YI; ZHANG, 2018), é dada pela soma dos módulos das diferenças das coordenadas:

$$\begin{aligned} d_2(x,y) &= \mid x_1 - y_1 \mid + \mid x_2 - y_2 \mid + \cdots + \mid x_n - y_n \mid \\ &= \sum_{i=1}^{n} \mid x_i - y_i \mid \end{aligned} \tag{2.3.5}$$

Essa é uma função que pode ser considerada parecida com a euclidiana pelo seu apelo geométrico, porém é mais "barata", pois se dispensa o cálculo da raiz quadrada e os expoentes.

Verifica-se abaixo que essa função de distância obedece às condições de métrica:

Condição 1:

$$\begin{aligned} d_2(x,x) &= \mid x_1 - x_1 \mid + \mid x_2 - x_2 \mid + \cdots + \mid x_n - x_n \mid \\ &= 0 + 0 + \cdots + 0 \\ &= 0 \end{aligned}$$

Condição 2:

Supondo que $x \neq y$, então existe pelo menos um $i \in \{1, 2, \cdots, n\}$ tal que $x_i \neq y_i$, logo $\mid x_i - y_i \mid > 0$. Portanto

$$\begin{aligned} d_2(x,y) &= \mid x_1 - y_1 \mid + \mid x_2 - y_2 \mid + \cdots + \mid x_n - y_n \mid \\ &\geq \mid x_i - y_i \mid \\ &> 0 \end{aligned}$$

Condição 3:

$$\begin{aligned} d_2(x,y) &= \mid x_1 - y_1 \mid + \mid x_2 - y_2 \mid + \cdots + \mid x_n - y_n \mid \\ &= \mid y_1 - x_1 \mid + \mid y_2 - x_2 \mid + \cdots + \mid y_n - x_n \mid \\ &= d_2(y,x) \end{aligned}$$

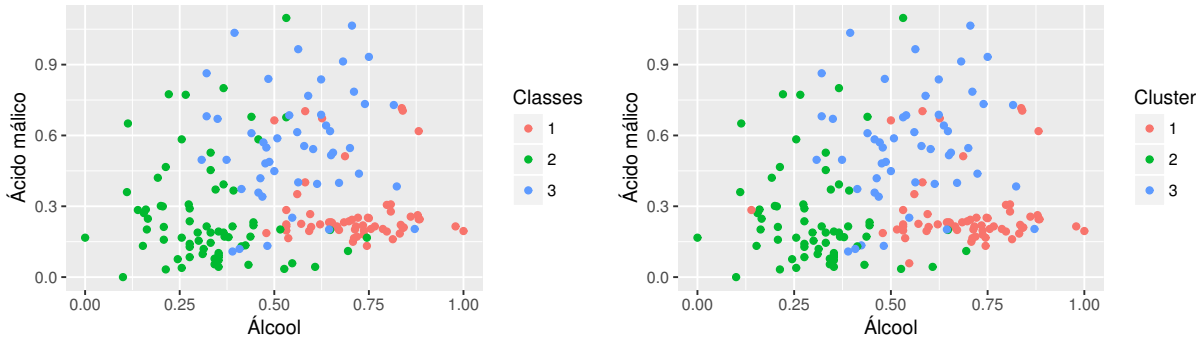
Condição 4:

$$\begin{aligned} d_2(x,z) &= \sum_{i=1}^{n} \mid x_i - z_i \mid \\ &= \sum_{i=1}^{n} \mid x_i - y_i + y_i - z_i \mid \\ &\leq \sum_{i=1}^{n} \mid x_i - y_i \mid + \sum_{i=1}^{n} \mid y_i - z_i \mid \\ &= d_2(x,y) + d_2(y,z) \end{aligned}$$

Disso, conclui-se que essa distância é de fato uma métrica.

Quando essa distância é usada no método *k-means* o algoritmo sofre uma mudança: os centros dos *clusters* (item (iii) da seção 2.2) são reposicionados para a mediana dos pontos de cada agrupamento, não para a média.

Fazendo um teste análogo ao anterior, aplicou-se no mesmo *dataset* o método *k-means* com distância Manhattan. O resultado obtido teve 96,06% de acerto, veja a seguir uma representação das categorias originais dos vinhos e os *clusters* obtidos com o método:



a) Grupos originais

b) *Clusters* obtidos

Figura 4. Comparação das categorias originais de vinho e dos *clusters* obtidos pelo método *k-means* com métrica Manhattan

Das 178 observações, o método acerta o grupo de 171. As 3 classes de vinho que possuem, respectivamente, 59, 71 e 48 exemplares, na clusterização aparecem com 63, 64 e 51, respectivamente.

2.3.3 Dissimilaridade pelo Cosseno

Sua função é proveniente do cosseno do ângulo α formado entre os vetores x e y e é definida da seguinte forma:

$$d_3(x,y) = 1 - cos(\alpha) = 1 - \frac{\langle x, y\rangle}{\mid x \mid \cdot \mid y \mid} = 1 - \frac{\Sigma_{i=1}^{n} x_i \cdot y_i}{\sqrt{\Sigma_{i=1}^{n} x_i^2} \cdot \sqrt{\Sigma_{i=1}^{n} y_1^2}}. \tag{2.3.6}$$

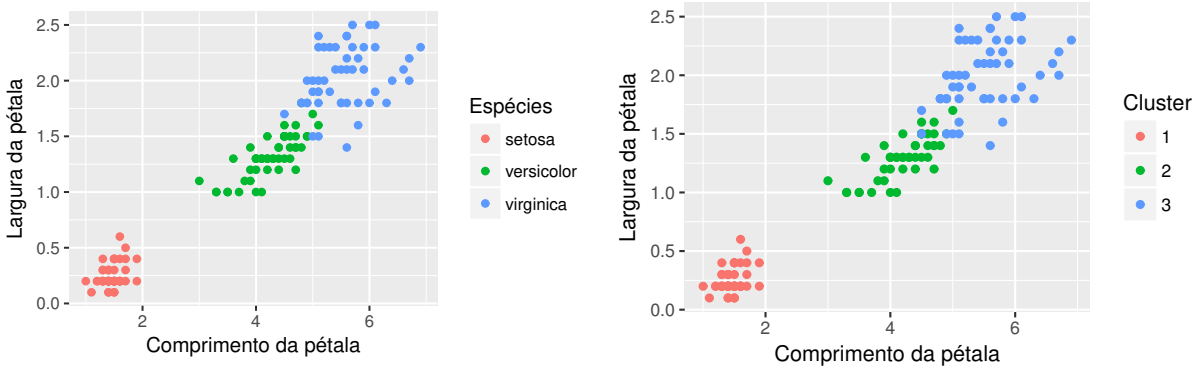
Leydesdorff e Rafols (2011) afirmam que essa função baseada no cosseno mede a dissimilaridade entre objetos, mas pode ser assumida como distância. Entretanto, essa função não cabe no conceito de métrica ou pseudo-métrica, pois não atende a desigualdade triangular, veja um contra exemplo: em $\mathbb{R}^2$, tomando $x = (1,0)$, $y = (\sqrt{2}/2, \sqrt{2}/2)$ e $z = (0,1)$, tem-se que:

$$d_3(x,z) = 1 - 0 = 1 \geq 2 - \sqrt{2} = 1 - \frac{\sqrt{2}}{2} + 1 - \frac{\sqrt{2}}{2} = d_3(x,y) + d_3(y,z)$$
$$\therefore d_3(x,z) \geq d_3(x,y) + d_3(y,z)\ .$$

Considera-se neste trabalho essa função como uma medida de dissimilaridade, com contradomínio restrito ao intervalo fechado $[0, 2]$, em que resultado 0 indica que os vetores são perfeitamente alinhados no mesmo sentido e resultado 2 mostra que eles são perfeitamente opostos. Esta função e a função cosseno têm sido usadas em diversos estudos, inclusive em processos de mineração de texto.

Quando o método *k-means* é aplicado com a distância do cosseno ele sofre uma mudança no reposicionamento dos centros dos *clusters* que passam a ser obtidos pela média das componentes dos versores dos seus vetores, sendo que defini-se o versor de um vetor como um vetor com comprimento de 1 unidade e que apresenta mesma direção e sentido que o vetor original.

Em um exemplo para teste, aplicou-se o método *k-means* com dissimilaridade pelo cosseno para clusterizar o conjunto dos dados Iris – disponível no repositório da UCI. O resultado obtido teve 96,67% de acerto (5 flores foram colocadas em grupos errados), veja a seguir uma representação das espécies da flor Iris e os *clusters* obtidos com o método:



a) Espécies originais

b) *Clusters* obtidos

Figura 5. Comparação das espécies da Iris e dos *clusters* obtidos pelo método *k-means* com dissimilaridade pelo cosseno

O conjunto de dados em questão possui 150 observações com 4 atributos cada, mas o gráfico indica as coordenadas relativas a apenas dois atributos das flores. Originalmente ele apresenta 3 espécies da Iris que possuem 50 exemplares cada. Após a clusterização dos dados, o grupo 1 obteve 50 elementos, o grupo 2 obteve 45 e o grupo 3 obteve 55, neste caso houveram 5 pontos que apareceram no grupo errado.

2.3.4 Dissimilaridade por Correlação

Proveniente da estatística, essa função usa o coeficiente de correlação de Pearson entre dois pontos ($\rho_{x,y}$) na definição da medida de dissimilaridade, onde

$$\rho_{x,y} = \frac{cov(x,y)}{\sqrt{var(x) \cdot var(y)}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

sendo $\overline{x}$ e $\overline{y}$ as médias aritméticas das entradas desses vetores, obtidas assim:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{e} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Tendo isso em mente, define-se a distância de correlação como:

$$d_4(x,y) = 1 - \rho_{x,y} = 1 - \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}. \tag{2.3.7}$$

Agora note que centralizando os pontos para a origem da seguinte forma $X = (x_1 - \overline{x}, x_2 - \overline{x}, \cdots, x_n - \overline{x})$ e $Y = (y_1 - \overline{y}, y_2 - \overline{y}, \cdots, y_n - \overline{y})$, obtemos então que:

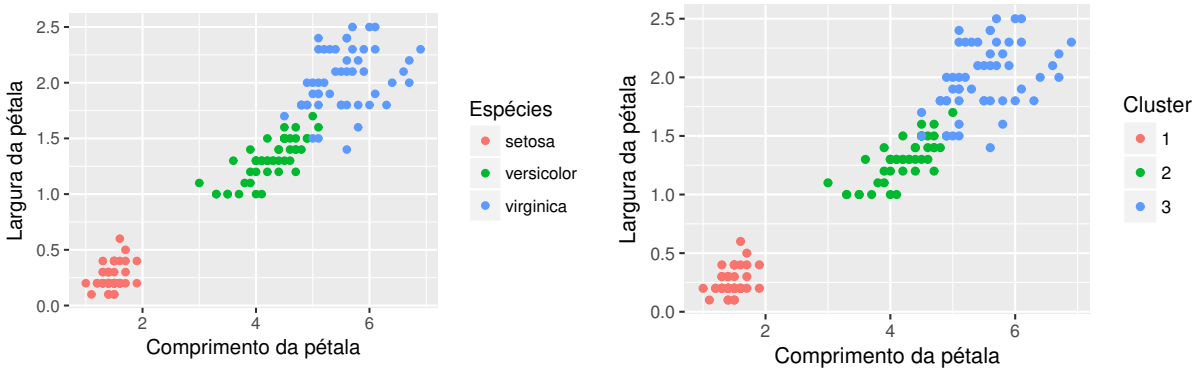
$$\begin{aligned} d_4(x,y) &= 1 - \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}} \\ &= 1 - \frac{\Sigma_{i=1}^{n} X_i \cdot Y_i}{\sqrt{\Sigma_{i=1}^{n} X_i^2 \cdot \Sigma_{i=1}^{n} Y_i^2}} \\ &= d_3(X,Y) \end{aligned}$$

Portanto, as duas últimas medidas de dissimilaridade são muito parecidas, porém a última centraliza os pontos sobre a origem (LEYDESDORFF; RAFOLS, 2011).

Quando essa função é usada no método *k-means*, o algoritmo reposiciona os centros para a média dos vetores centralizados na origem e normalizados pelo desvio padrão.

Novamente tomando o conjunto Iris para exemplificar, aplicou-se o método *k-means* com dissimilaridade por correlação. O resultado obtido teve 96,00% de acertos (6 flores colocadas em grupos errados).

Veja a seguir uma representação das espécies da flor Iris e os *clusters* obtidos com o método:



a) Espécies originais

b) *Clusters* obtidos

Figura 6. Comparação das espécies da flor Iris com dos *clusters* obtidos pelo método *k-means* com a dissimilaridade por correlação

Das 150 observações, o método clusteriza corretamente 144 delas. As 3 espécies da Iris que possuem 50 exemplares cada, na clusterização também aparecem com 50 cada, porém trocando 3 flores versicolor e 3 flores virginica.

2.4 VALIDAÇÃO DOS RESULTADOS

Uma etapa bastante importante da clusterização de dados é a validação dos resultados. Mais do que apenas criar grupos, é necessário que esses grupos satisfaçam as condições de máxima homogeneidade *intracluster* e máxima heterogeneidade *intercluster*, ou seja, é preciso ter qualidade no particionamento.

> A validação de estruturas de agrupamento é a parte mais difícil e frustrante da análise de *cluster*. Sem um esforço forte nessa direção, a análise de *cluster* permanecerá uma caixa preta acessível apenas para aqueles verdadeiros crentes que tenham experiência e grande coragem (tradução livre) (JAIN; DUBES, 1988).

Tendo em vista que os *datasets* usados neste trabalho já tem os grupos pré-definidos, para avaliar quantitativamente o quão boas são as partições obtidas pelo método *k-means* com diferentes funções de distâncias optou-se pelo uso de um índice externo de validação, isto é, um índice comparativo entre a partição obtida e as classes originais para medir o grau de semelhanças entre elas – novamente surge a noção de medida: o quanto determinada partição é semelhante da partição original?

Portanto, tentando identificar qual distância gera resultados mais fiéis à realidade, optou-se pelo uso de um índice de comparação externo. Neste caso, o Índice de Rand Ajustado (do inglês, *Adjusted Rand Index* ARI) foi considerado o que melhor se adequa

aos objetivos previstos para validação do desempenho do método *k-means* com diferentes medidas e é também um dos mais comumente usados (VINH; EPPS; BAILEY, 2010).

Esse índice comparativo entre a partição obtida e as classes originais varia entre 0 e 1, sendo que quanto mais próximo de 1 é o índice, melhor é a partição em relação às classes originais (VINH; EPPS; BAILEY, 2010).

Antes de definir o ARI, é preciso entender seus coeficientes de entrada. Em um determinado conjunto de dados M qualquer contendo n elementos, seja U uma clusterização, portanto $U = \{U_1, U_2, \cdots, U_r\}$, de tal forma que $U_i \subset M, \forall i \in \{1, 2, \cdots, r\}$, $\cup_{i=1}^{r} U_i = M$ e $U_i \cap U_j = \varnothing$ para $i \neq j$. Para comparar uma segunda partição $V = \{V_1, V_2, \cdots, V_s\}$ com U, o ARI usa contagem de pares de dados nos quais dois agrupamentos concordam ou discordam (VINH; EPPS; BAILEY, 2010). Os C_2^n (total de combinações de 2 elementos tomados entre os n disponíveis) pares de itens podem ser divididos em quatro grupos:

i) pares que pertencem ao mesmo *cluster* em U e em V, o total desses pares será denotado por N_{11};

ii) pares que pertencem a *clusters* distintos em U e em V, total denotado por N_{00};

iii) pares que pertencem ao mesmo *cluster* em U e pertencem a *clusters* distintos em V, total denotado por N_{10};

iv) pares que pertencem a *clusters* distintos em U e pertencem ao mesmo *cluster* em V, total denotado por N_{01};

Então o cálculo do ARI, é dado por:

$$ARI(U, V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}. \tag{2.4.1}$$

3 APLICAÇÃO DO MÉTODO K-MEANS COM DIFERENTES FUNÇÕES DE MEDIDA

Para manipular os dados e realizar a clusterização optou-se pelo uso do software MatLab, que é "uma plataforma de programação projetada especificamente para engenheiros e cientistas. O coração do MATLAB é a linguagem MATLAB, uma linguagem baseada em matriz que permite a expressão mais natural da matemática computacional" (tradução livre) (MATHWORKS, 1994-2018). No MatLab a função *k-means* padrão é programada com o método de convergência de Lloyd e quando o usuário não indica as posições dos centros iniciais, eles são escolhidos por meio de um algoritmo que promete melhorar o tempo de execução e a qualidade da partição. E para plotar os gráficos e calcular o ARI foi utilizado o RStudio, que é um software livre de ambiente de desenvolvimento para a linguagem *R* (WIKIPÉDIA, 2018).

Os *datasets* usados como referência de comparação são o Iris de Fisher, tradicionalmente usado, disponível no repositório UCI, que apresenta 4 características, a saber: as medidas de comprimento e largura de pétalas e de sépalas de 150 flores Iris classificadas por Fisher em 3 espécies; o Vinho, também disponível no repositório UCI, traz 13 atributos químicos sobre 178 vinhos oriundos da mesma região da Itália, mas de três diferentes cultivares; e, por último, um conjunto referente a cromossomos humanos, que tem 4060 observações e informa 3 características sobre eles: o comprimento e posição do centrômero dos cromossomos e o padrão de banda ao longo do eixo longitudinal, tais informações permitem identificar 7 grupos, os grupos de Denver (SOUZA, 1999).

Os *datasets* em estudo indicam a classe de cada observação em alguma de suas colunas, mas considerando que métodos de clusterização não utilizam classes pré-definidas, os *datasets* usados no estudo terão o vetor de classes omitido para a implementação no método *k-means*. Além disso, o foco deste estudo é a influência causada pela escolha função de medida sobre o algoritmo de clusterização, não sendo objeto de estudo a determinação do número de *clusters* k em que cada uma das bases idealmente deve ser dividida, assim cada *dataset* foi particionado com a mesma quantidade de grupos em que está dividido originalmente.

Apresentam-se a seguir os *clusters* obtidos na aplicação do método *k-means* com cada uma das métricas indicadas.

3.1 CLUSTERIZAÇÃO DO CONJUNTO DE DADOS IRIS

Na tentativa de garantir que o algoritmo não convirja para um mínimo local, o método foi reproduzido 100 vezes com centros iniciais diferentes. A seguir são apresentados gráficos de dispersão de pontos do *dataset* Iris para representar os agrupamentos obtidos mediante o método *k-means* com as funções de medida previstas. A figura é bidimensional, mas que contempla uma terceira característica dos dados no tamanho dos pontos (apenas a característica "largura da sépala" não está representada) e os pontos tem transparência para que possam ser percebidas as sobreposições.

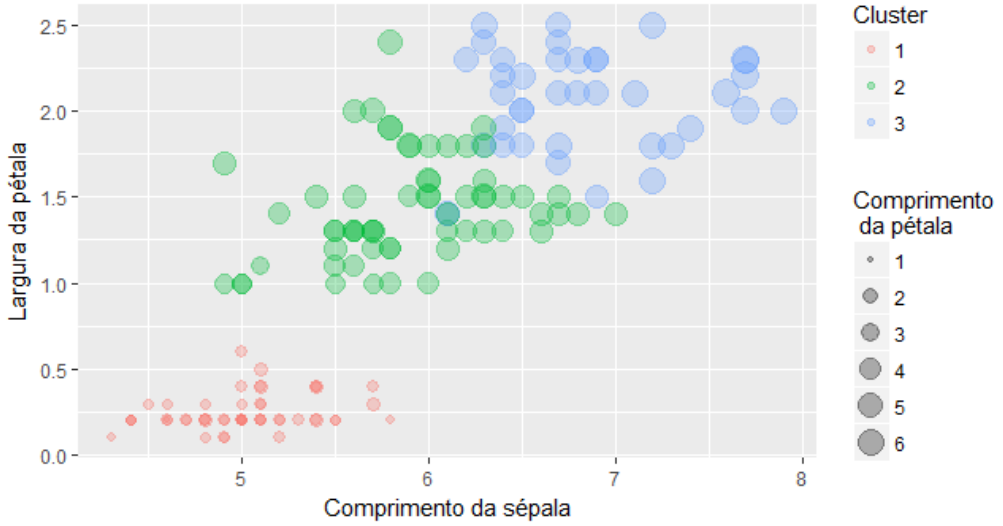


Figura 7. Representação dos *clusters* da Iris obtidos com distância euclidiana ao quadrado

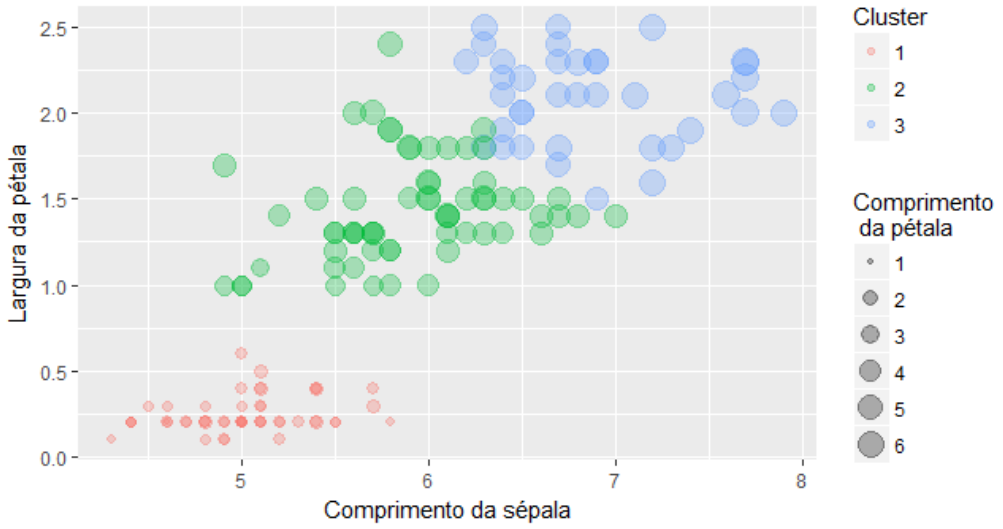


Figura 8. Representação dos *clusters* da Iris obtidos com métrica Manhattan

Note que a diferença entre as partições dos gráficos acima é de apenas um ponto (nº

135) que pertence ao *cluster* 3 segundo a distância euclidiana ao quadrado e pertence ao 2 segundo na distância Manhattan.

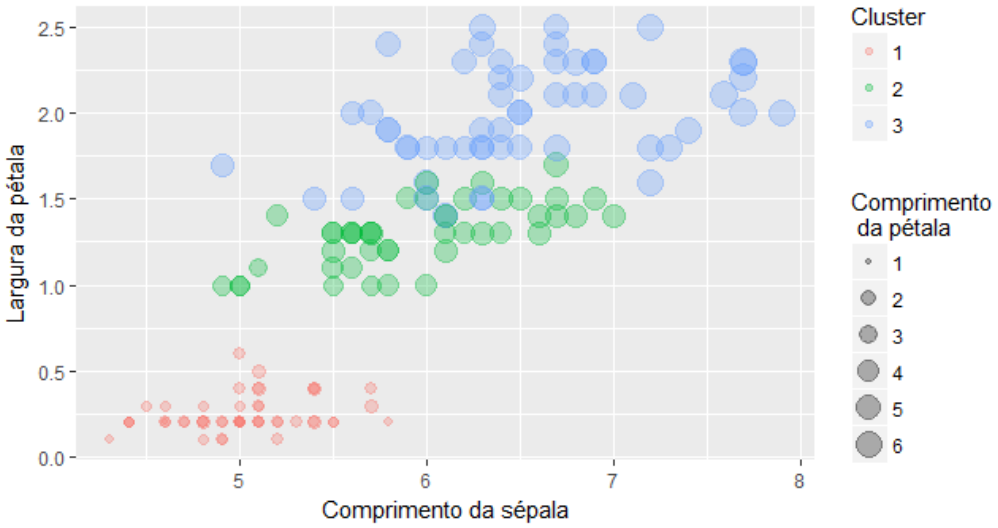


Figura 9. Representação dos *clusters* da Iris obtidos com dissimilaridade pelo cosseno

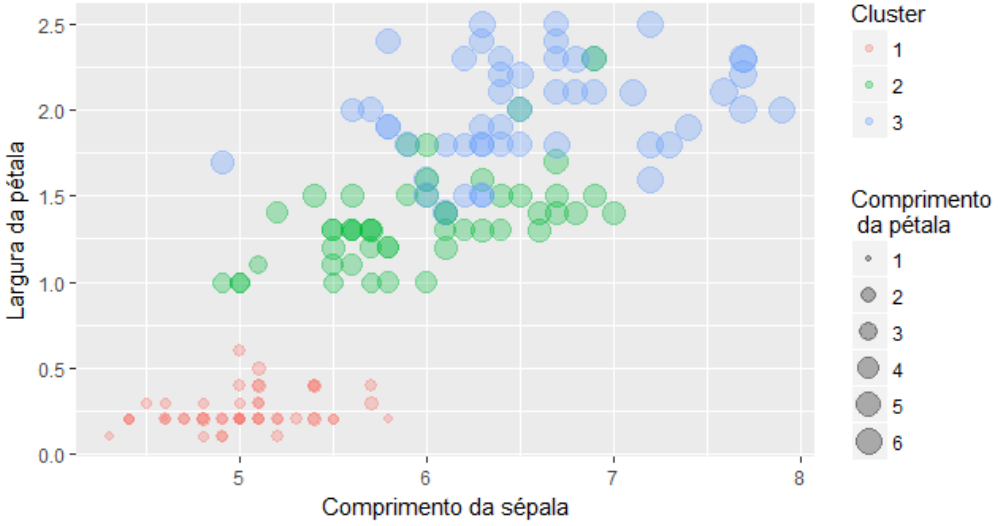


Figura 10. Representação dos *clusters* da Iris obtidos com dissimilaridade por correlação

Nesse experimento com "poucos" dados, evidencia-se que há mudanças provocadas pela escolha da função de distância. Segue abaixo a dispersão dos dados com indicação das espécies de Denver.

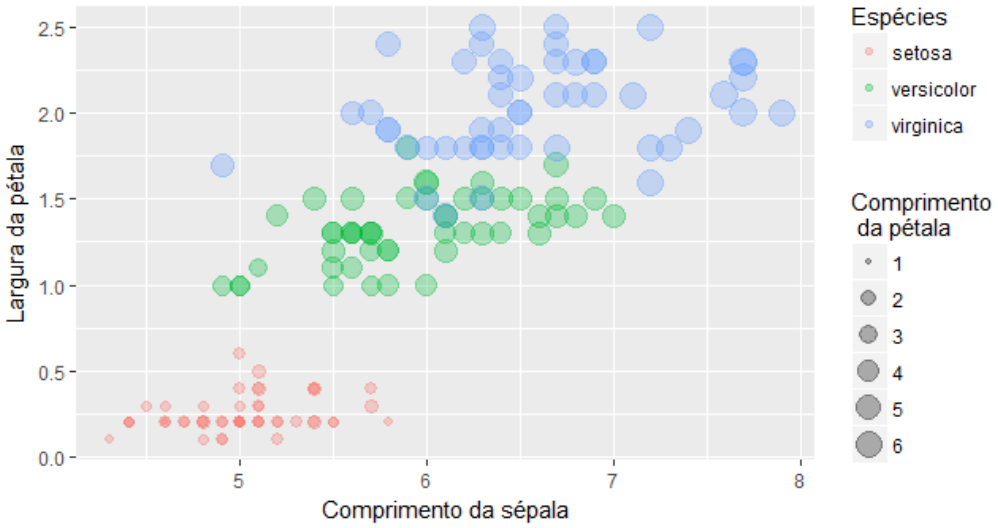
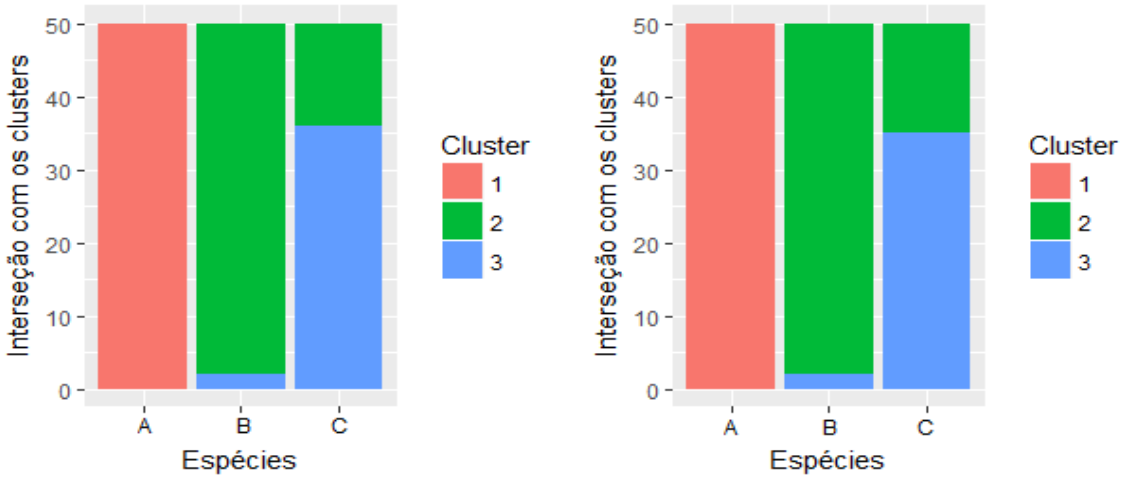


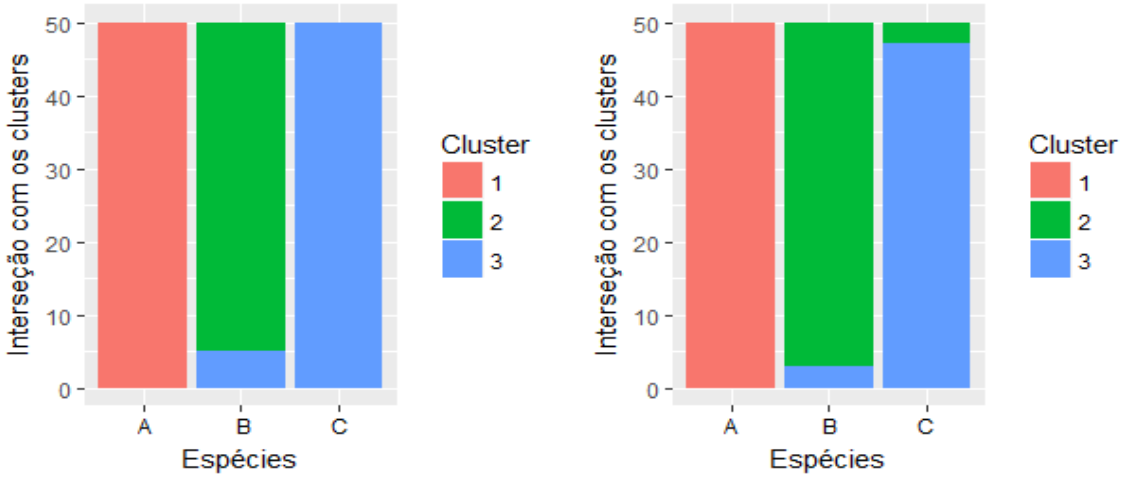
Figura 11. Representação das espécies da Iris segundo a classificação de Fisher

Originalmente cada espécie tem 50 exemplares, que comparadas com os *clusters* resultam nas seguintes interseções:



(a) Interseção das espécies da Iris e os *clusters* obtidos com distância euclidiana ao quadrado

(b) Interseção das espécies da Iris e os *clusters* obtidos com distância Manhattan



(c) Interseção das espécies da Iris e os *clusters* obtidos com dissimilaridade pelo cosseno

(d) Interseção das espécies da Iris e os *clusters* obtidos com dissimilaridade por correlação

Figura 12. Interseções dos *clusters* provenientes do método *k-means* com diferentes funções de medida e as espécies da flor Iris.

Por questões de espaço no eixo das abscissas dos gráficos, os nomes das espécies Setosa, Versicolor e Virginica foram substituídos por A, B e C, respectivamente.

Esses gráficos indicam os acertos e erros de agrupamento, por exemplo, a figura 12.(a) indica que 2 flores da espécie Versicolor foram inseridas mediante a distância euclidiana ao quadrado no *cluster* 3, que corresponde a espécie Virginica, e 14 flores da Virginica foram tidas como pertencentes ao grupo 2, correspondente a espécie Versicolor. Então conclui-se que em relação a espécie Setosa todas as medidas se mostram eficientes com 100% de acerto; já na espécie Versicolor ocorreu que 4%, 4%, 10% e 6% das flores foram atribuídas a outro grupo pelas distâncias euclidiana ao quadrado, Manhattan, do cosseno e de correlação, respectivamente; e na espécie Virginica os erros foram de 28%,

30%, 0 e 3%, respectivamente. Portanto, os melhores agrupamentos foram obtidos através da dissimilaridade pelo cosseno e por correlação, com uma pequena vantagem na primeira. Essa conclusão é reforçada pelo índice de Rand Ajustado (ARI) que recebe valor 0.9039 para o agrupamento da dissimilaridade pelo cosseno e 0.8857 no obtido com correlação, 0.7302 na distância euclidiana ao quadrado e 0.7173 na métrica Manhattan.

3.2 CLUSTERIZAÇÃO DO CONJUNTO DE DADOS VINHO

Por conta desse conjunto ter atributos em escalas muito diferentes (por exemplo, o atributo "matiz" varia entre 0.48 e 1.71, enquanto a "prolina" varia entre 278 e 1680), a análise pode ser afetada (CASSIANO, 2014), então cada atributo foi normalizado para o intervalo $[0, 1]$. Além disso, método foi reproduzido 100 vezes com centros iniciais diferentes.

A seguir são apresentados gráficos de dispersão de pontos para representar os agrupamentos obtidos com o método *k-means* com diferentes medidas de distância e dissimilaridade. A figura é bidimensional, mas contempla uma terceira característica dos dados no tamanho dos pontos (neste caso, 10 atributos ficaram omissos) e os pontos tem transparência para que possam ser percebidas as sobreposições.

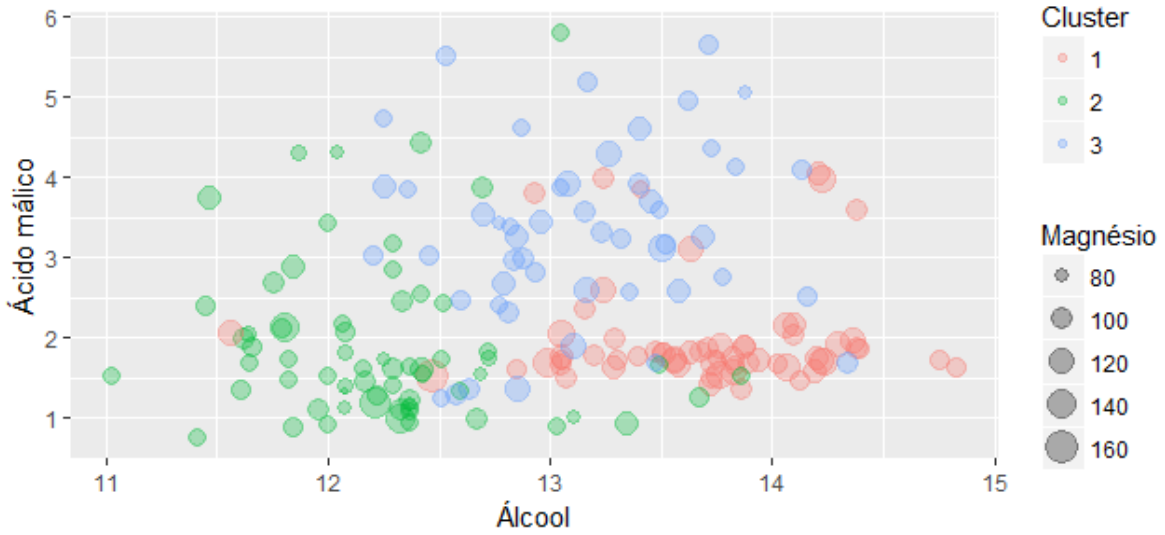


Figura 13. Representação dos *clusters* dos vinhos obtidos com distância euclidiana ao quadrado

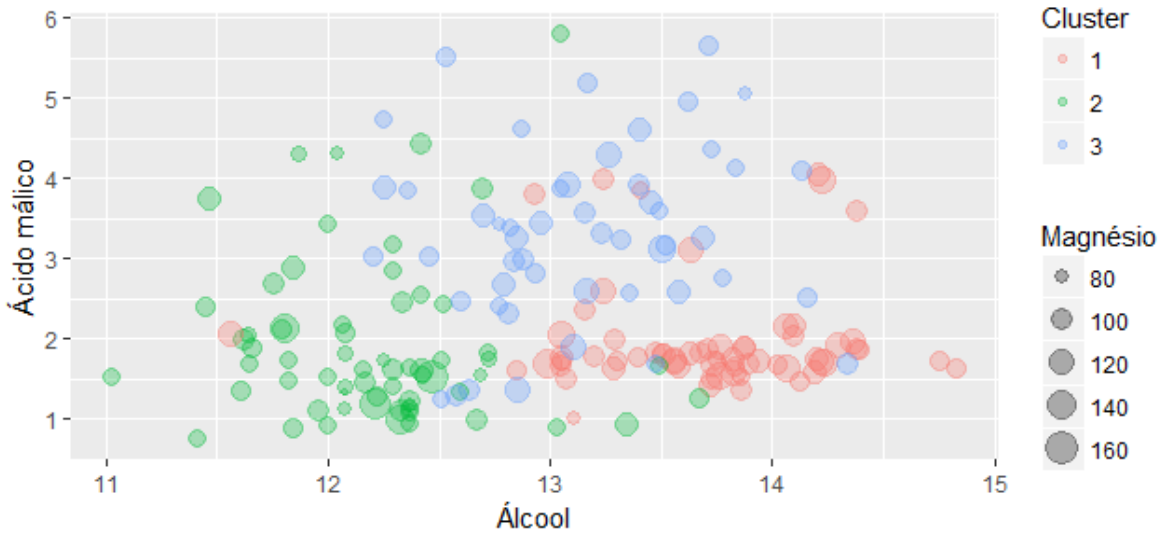


Figura 14. Representação dos *clusters* dos vinhos obtidos com distância Manhattan

Novamente as distâncias euclidiana ao quadrado e Manhattan produzem partições muito parecidas, divergindo na classificação de apenas 3 pontos.

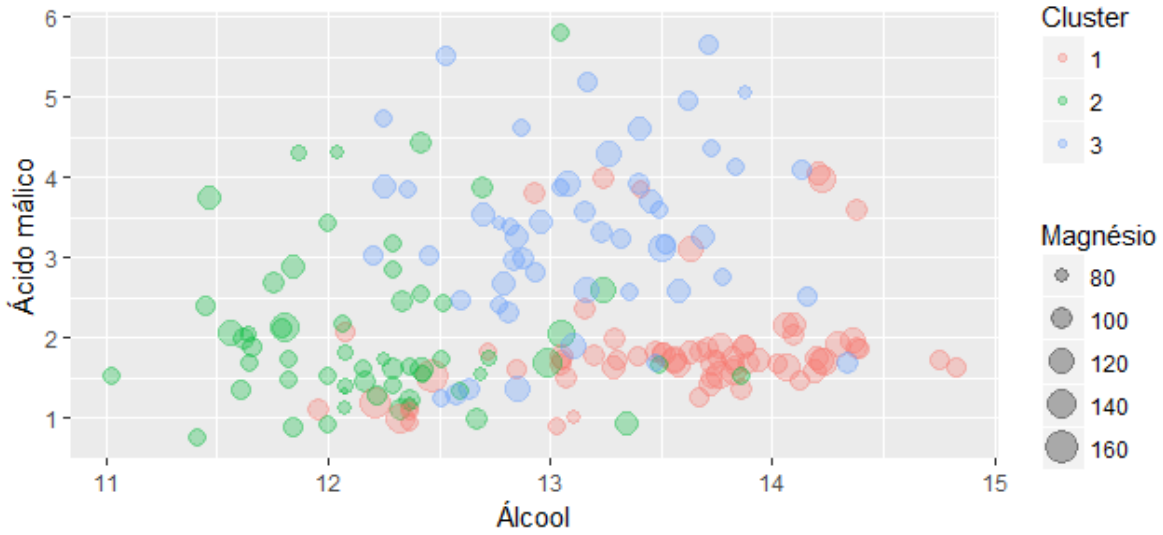


Figura 15. Representação dos *clusters* dos vinhos obtidos com dissimilaridade pelo cosseno

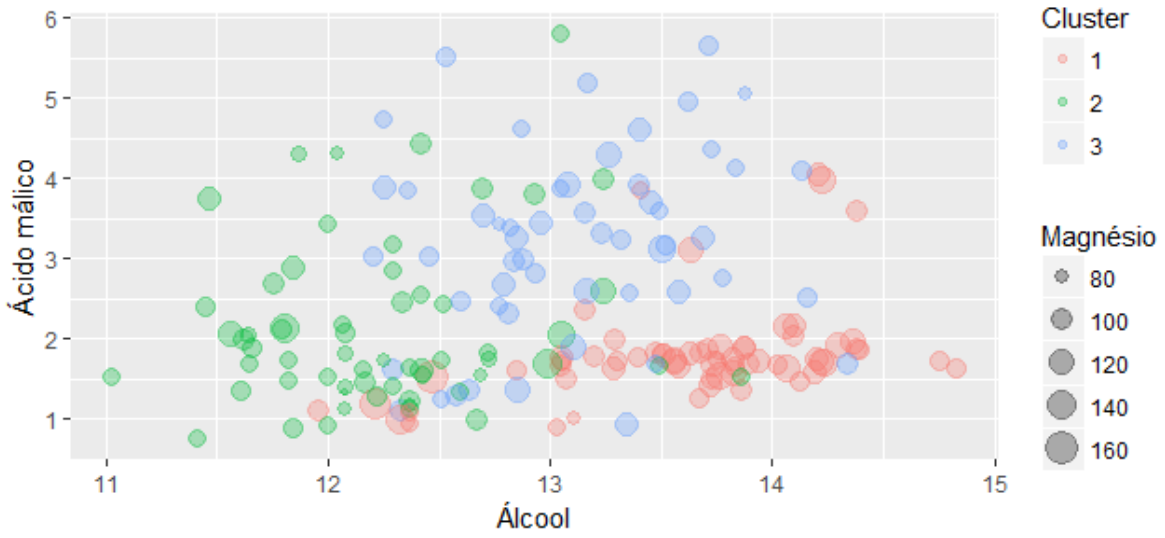


Figura 16. Representação dos *clusters* dos vinhos obtidos com dissimilaridade por correlação

Ao que tudo indica, a quantidade de atributos não gerou muitas confusões na definição dos grupos, afinal, todos os resultados obtidos ficam razoavelmente semelhantes. A seguir é apresentada a dispersão dos dados com indicação das classes originais dos vinhos.

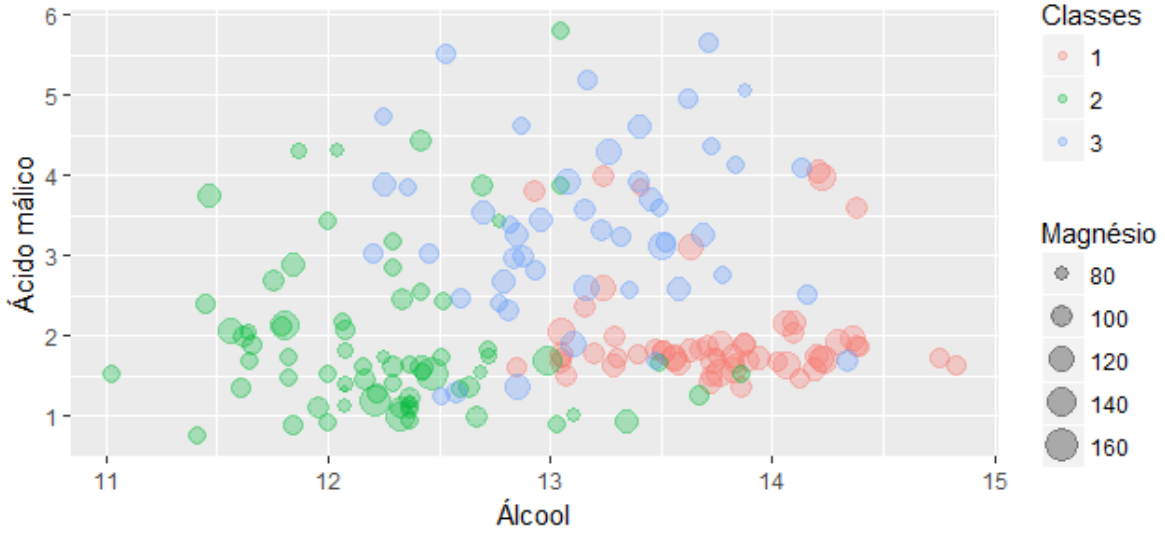
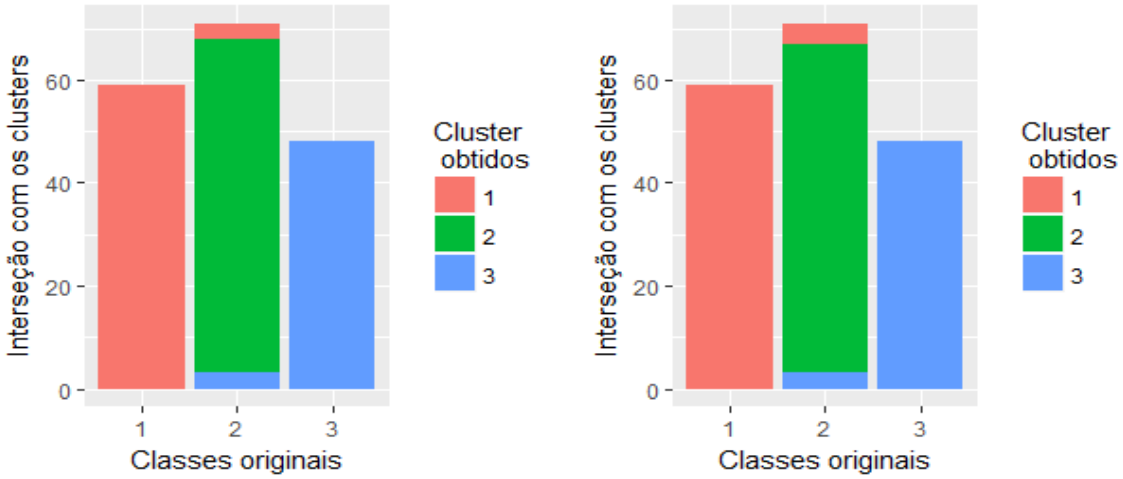


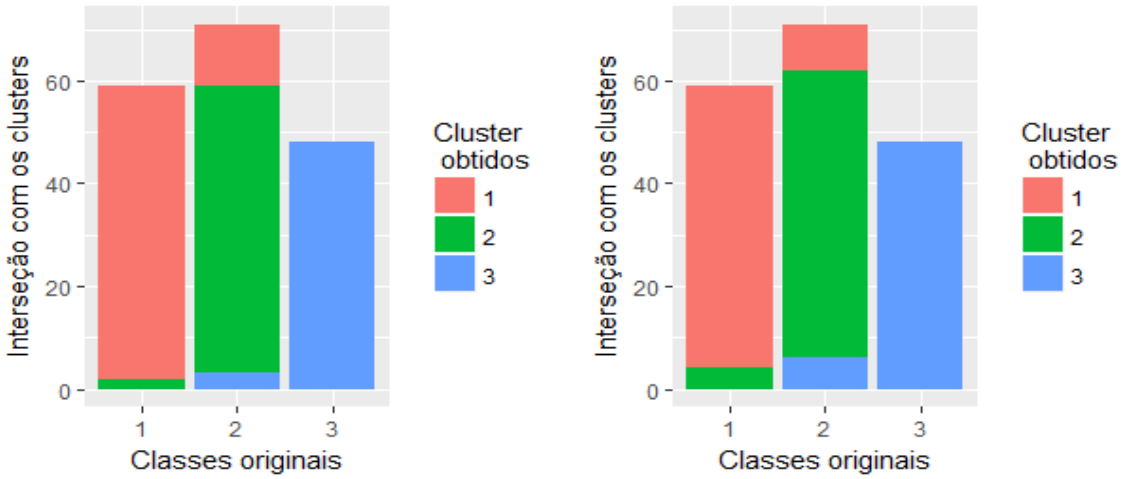
Figura 17. Representação das classes originais dos vinhos

Comparando as classes de vinho originais com os *clusters* de cada medida na ordem em que foram apresentadas há divergência de 6 (3.37%), 7 (3.93%), 17 (9.55%) e 19 vinhos (10.67%), respectivamente. Para uma melhor compreensão das diferenças entre classes e grupos, observe o cruzamento das informações:



(a) Interseção das classes de vinho e os *clusters* obtidos com distância euclidiana ao quadrado

(b) Interseção das classes de vinho e os *clusters* obtidos com distância Manhattan



(c) Interseção das classes de vinho e os *clusters* obtidos com dissimilaridade pelo cosseno

(d) Interseção das classes de vinho e os *clusters* obtidos com dissimilaridade por correlação

Figura 18. Interseções dos *clusters* provenientes do método *k-means* com diferentes funções de medida e as classes originais dos vinhos.

Reafirmado pelo ARI, a distância euclidiana ao quadrado tem melhor desempenho recebendo índice de 0.8975, a segunda melhor é a distância Manhattan com índice 0.8804, a dissimilaridade pelo cosseno recebe 0.7256 no ARI e a medida de dissimilaridade por correlação tem o pior desempenho com índice 0.6976.

3.3 CLUSTERIZAÇÃO DO CONJUNTO DE DADOS CROMOSSOMO

Em seguida apresenta-se a análise dos dados dos cromossomos, analogamente as apresentações anteriores da Iris e do vinho. Como esses dados possuem somente 3 atri-

butos, todas as informações disponíveis foram contempladas no gráfico de dispersão dos pontos, veja os resultados de acordo com as clusterizações feitas com cada uma das funções de medida:

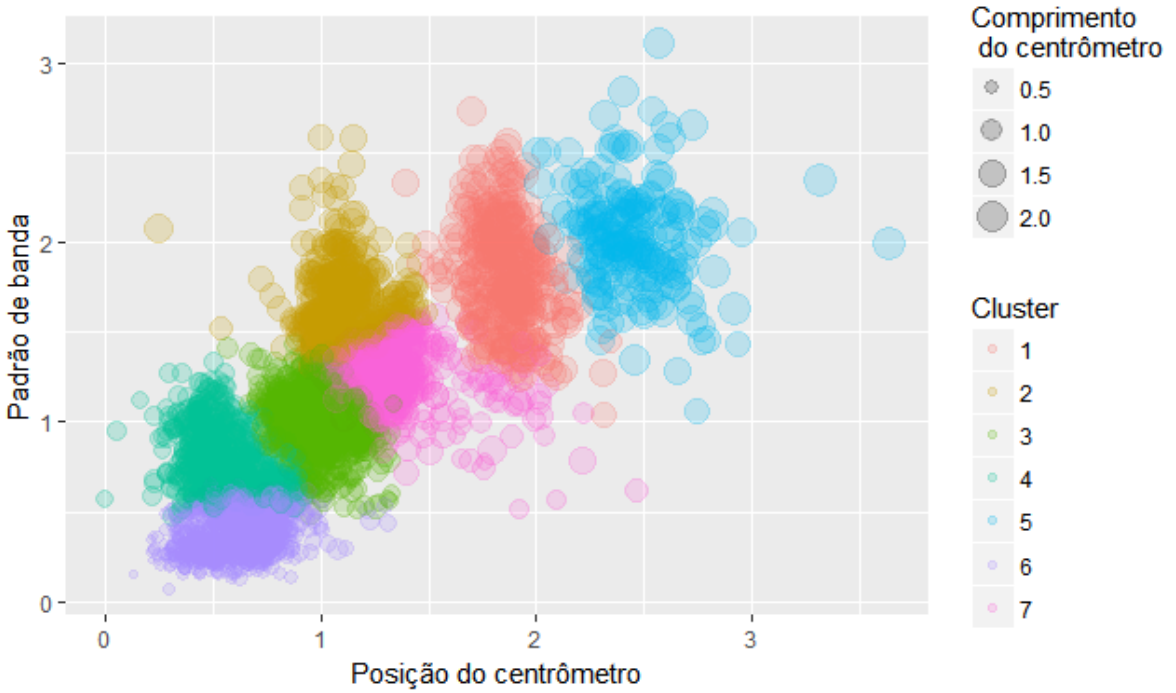


Figura 19. Representação dos *clusters* dos cromossomos obtidos com distância euclidiana ao quadrado

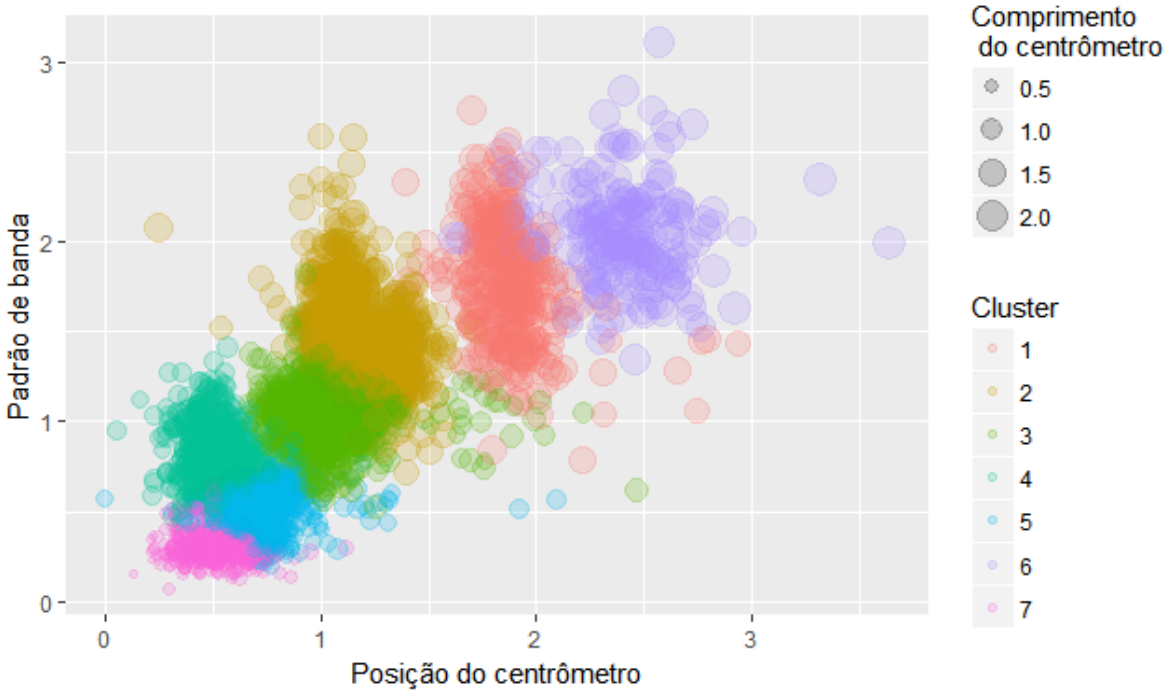


Figura 20. Representação dos *clusters* dos cromossomos obtidos com métrica Manhattan

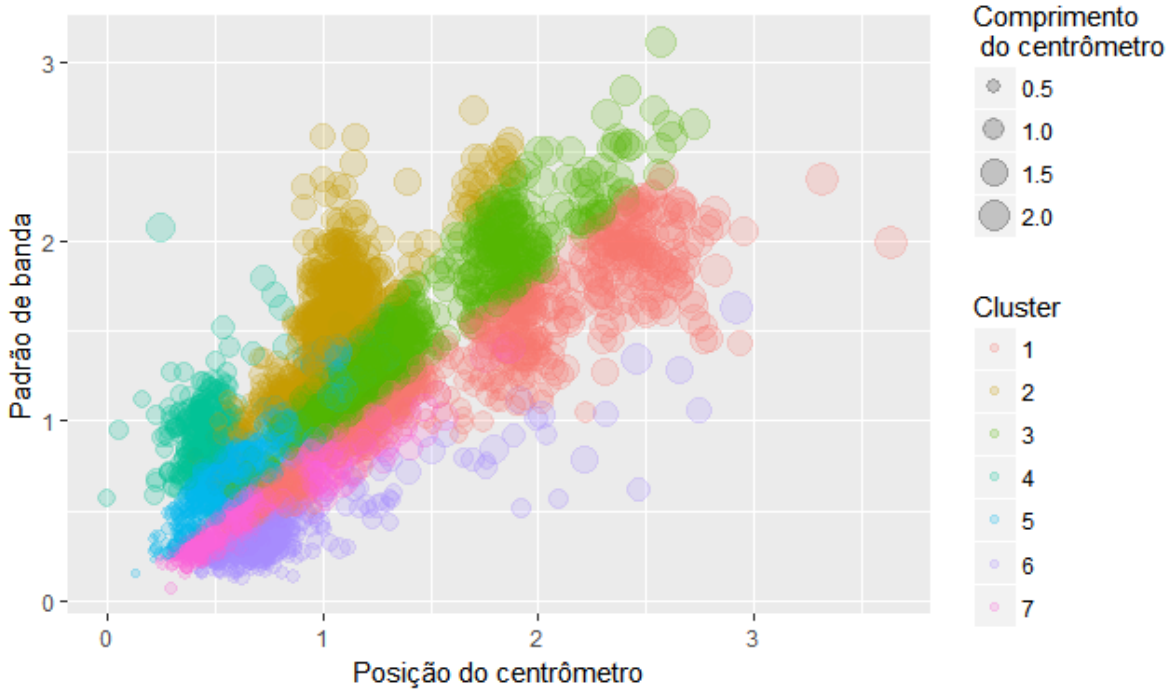


Figura 21. Representação dos *clusters* dos cromossomos obtidos com dissimilaridade pelo cosseno

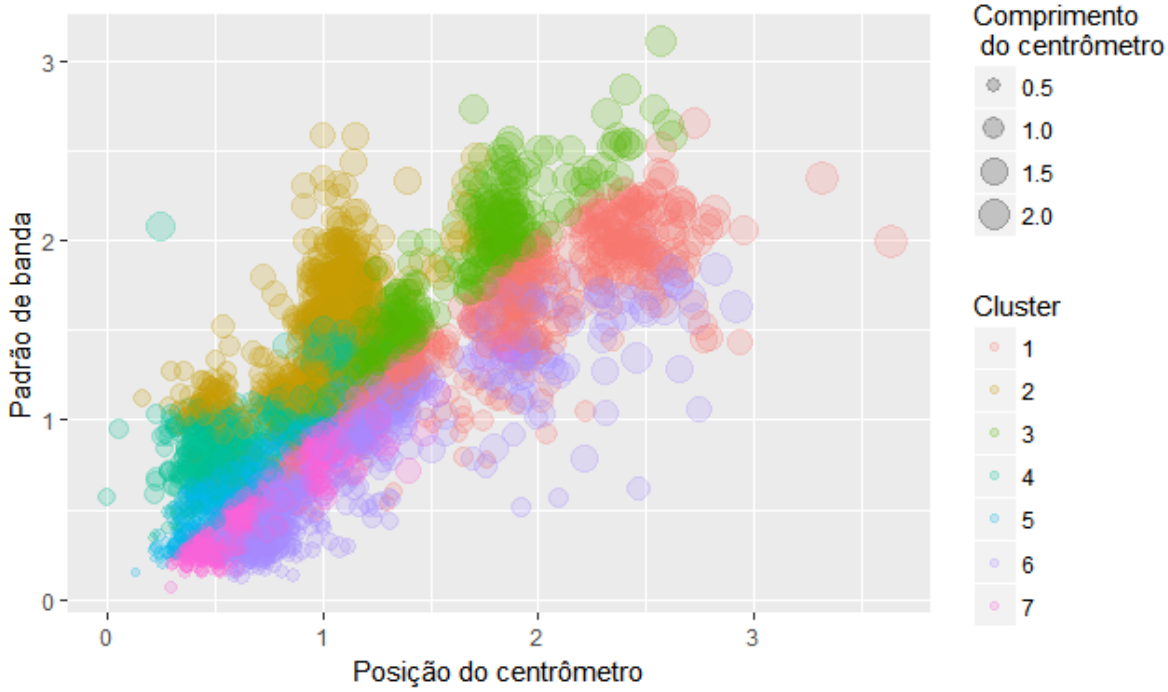


Figura 22. Representação dos *clusters* dos cromossomos obtidos com dissimilaridade por correlação

Percebe-se que por ser um conjunto maior de observações, sem separação nítida entre agrupamentos, todas as funções de medida convergiram de formas distintas, proporcionando resultados bem diferentes entre si.

Veja a seguir a dispersão original para os cromossomos segundo os grupos de Denver.

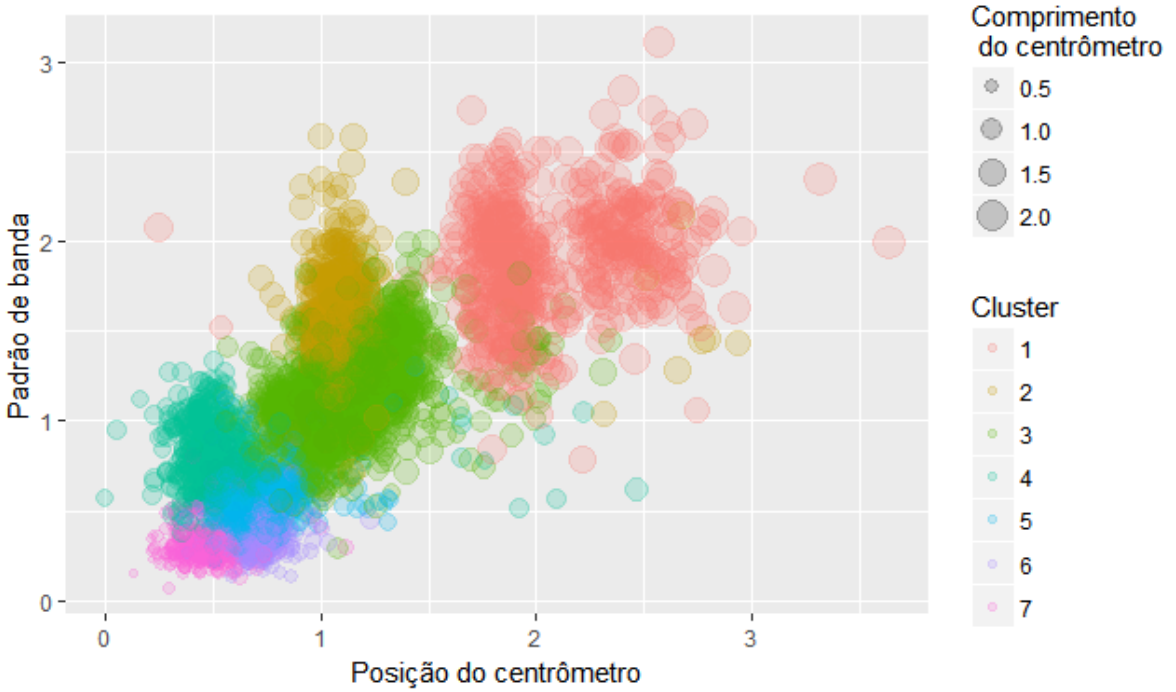


Figura 23. Representação do grupos de Denver dos cromossomos

Os *clusters* obtidos em todas as medidas não foram muito precisos, por exemplo, nenhuma delas foi sensível para identificar aproximadamente os grupos 5, 6 e 7.

A seguir, apresentam-se os gráficos de interseção entre os grupos de Denver e os *clusters* segundo sua medida. Os tamanhos das barras são diferentes correspondendo às quantidades de dados que cada grupo possui originalmente, neste caso, os grupos de 1 a 7 têm 515, 346, 1359, 535, 539, 360 e 406, respectivamente.

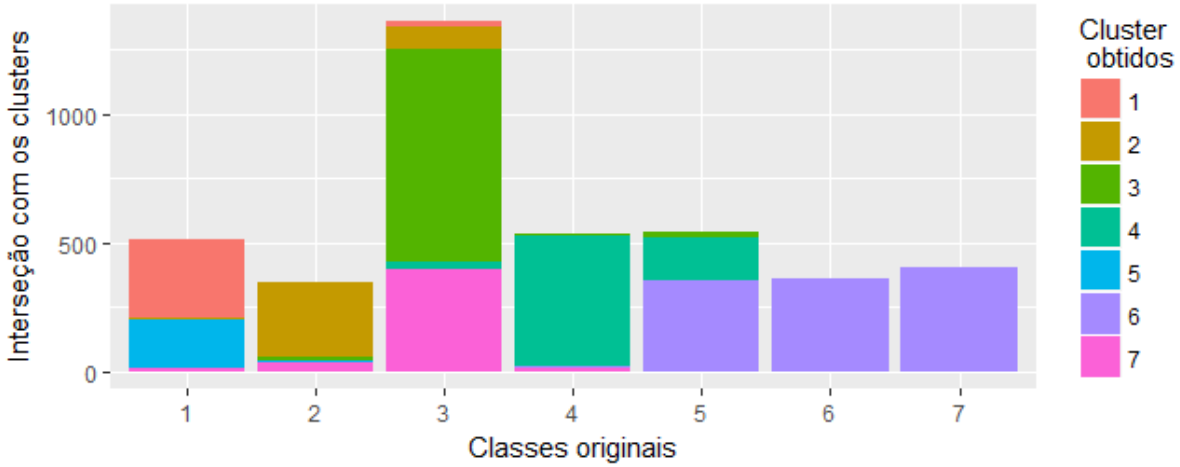


Figura 24. Interseção dos grupos de Denver e os *clusters* obtidos com distância euclidiana ao quadrado

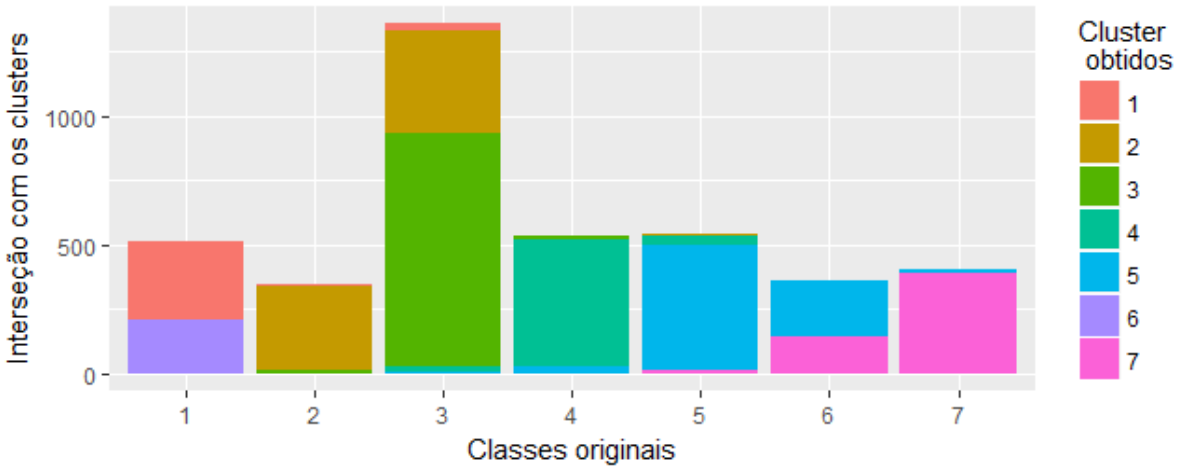


Figura 25. Interseção dos grupos de Denver e os *clusters* obtidos com distância Manhattan

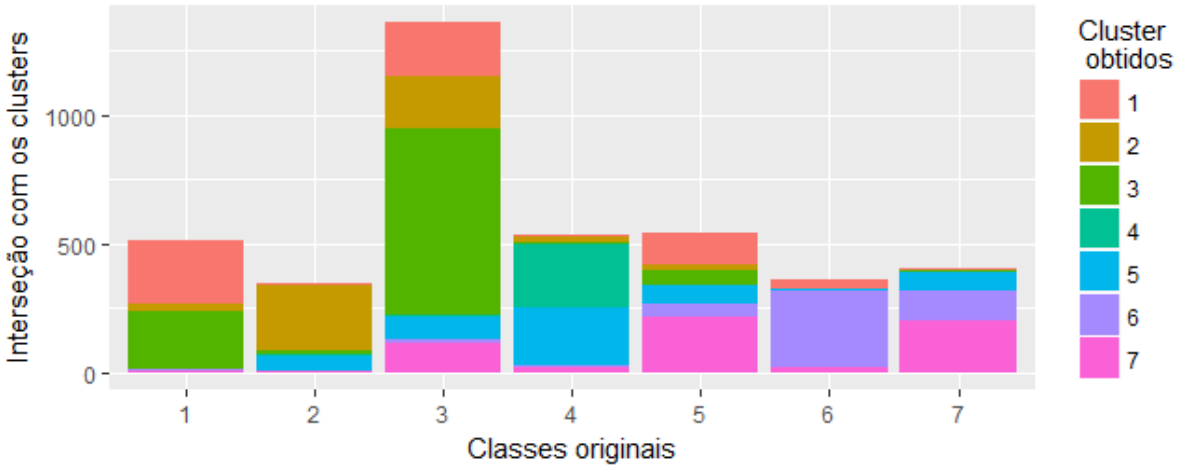


Figura 26. Interseção dos grupos de Denver e os *clusters* obtidos com dissimilaridade pelo cosseno

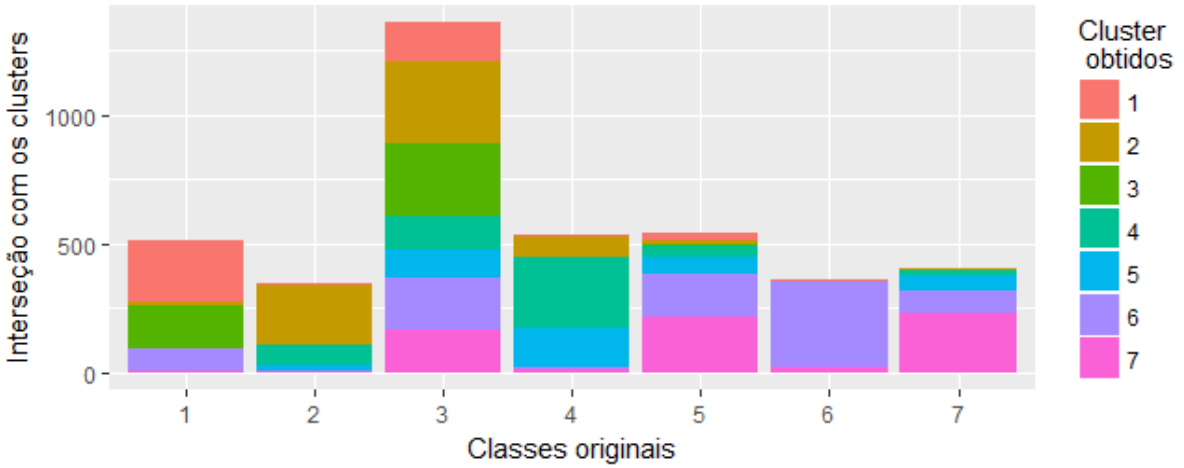


Figura 27. Interseção dos grupos de Denver e os *clusters* obtidos com dissimilaridade por correlação

Nesse teste, pode-se perceber o padrão geométrico de que as medidas de distância euclidiana ao quadrado e Manhattan tendem a formar grupos sem muita variação nas

dimensões, enquanto as medidas de dissimilaridade pelo cosseno e por correlação identificam grupos longilíneos que tendem a manter próxima a razão entre as coordenadas de cada ponto do mesmo grupo.

Neste caso, as medidas de distância se aproximam mais do resultado desejado, mesmo ambas tendo repartido o grupo 1 de Denver.

Segundo a escala do Índice de Rand Ajustado, a distância euclidiana ao quadrado recebe valor 0.4819 (tendo agrupado certo 2286 pontos dos 4060 totais, ou seja, aproximadamente 56,3%), a distância Manhattan 0.5828 (acertando 2894 pontos, 71,28%), a dissimilaridade pelo cosseno 0.2653 (2033 pontos, 50,07%) e a dissimilaridade por correlação 0.1461 (1663 pontos, 40,96%), portanto a melhor distância nesse caso é a Manhattan.

4 CONCLUSÃO

Nesse trabalho comparou-se, por meio de simulação, o desempenho do método *k-means* com variações na medida de dissimilaridade e distância entre objetos. As funções de medida testadas foram a distância euclidiana ao quadrado, a distância Manhattan, a dissimilaridade pelo cosseno e a dissimilaridade por correlação de Pearson. E para validação dos resultados fez-se uso do Índice de Rand Ajustado.

A tabela abaixo apresenta resumidamente os resultados obtidos e também as características de cada um dos conjuntos de teste.

Datasets			ARI referenta a função de medida			
Nome	Quantidade atributos	Quantidade observações	Euclidiana ao quadrado	Manhattan	Dissimilaridade pelo cosseno	Dissimilaridade por correlação
Iris	4	150	0.7302	0.7173	**0.9039**	0.8857
Vinho	13	178	**0.8975**	0.8804	0.7256	0.6976
Cromossomo	3	4060	0.481	**0.5828**	0.2653	0.1461

Por esse experimento concluiu-se cada conjunto de dados, em sua particularidade, se adapta melhor a uma função de medida diferente. Em uma análise superficial parece plausível crer que medidas de similaridade pelo cosseno e por correlação tendem a se adequar melhor com dados em que a proporção entre as características é mais relevante que o tamanho delas em si, isso quer dizer que a suposta tendência é manter no mesmo *cluster* objetos que possuem aproximadamente a mesma razão entre pares de atributos, como é o caso da Iris em que espera-se que cada espécie da flor mantenha seu formato apesar das pequenas variações de tamanho que exemplares diferentes podem ter; enquanto medidas de distância euclidiana ao quadrado e Manhattan são mais sensíveis ao valor que cada característica recebe, neste caso os *clusters* são limitados aproximadamente por intervalos em que cada característica deve pertencer.

Por conseguinte, a relevância da escolha de função de medida está diretamente relacionada a qualidade da partição final, podendo ela se adequar bem a realidade ou não.

Entretanto, não foi objetivo deste trabalho explorar as propriedades das funções de medida para analisar a relação do método *k-means* com a performance na clusterização, esse mérito é dispensado para trabalhos futuros, como também o estudo funções de medida para dados não numéricos.

REFERÊNCIAS

CAMILO, C. O.; SILVA, J. C. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. [S.l.], 2009. 28 p.

CARDOSO, G. S. et al. Clusterização k-means: Uma proposta de melhoria. In: *Workshop de Trabalhos de Iniciação Científica e Graduação Bahia, Alagoas e Sergipe*. Salvador: Anais do ERBASE 2008, 2008.

CASSIANO, K. M. *Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade*. Tese (Doutorado) — PUC-RIO, Rio de Janeiro, 2014.

ELSHEIKHA, S. et al. Cluster analysis of diffusion tensor fields with application to the segmentation of the corpus callosum. *Procedia Computer Science*, 2016.

GASPAR-CUNHA, A.; TAKAHASHI, R.; ANTUNES, C. *Manual de computação evolutiva e metaheurística*. [s.n.], 2012. (Ensino). ISBN 9789892601502. <https://books.google.com.br/books?id=9Di5CwAAQBAJ>.

GRABUSTS, P. The choice of metrics for clustering algorithms. In: *Proceedings of the 8th International Scientific and Practical Conference*. Rezekne: Ra Izdevniecība, 2011. p. 70–76.

HRUSCHKA, E. R.; EBECKEN, N. F. A genetic algorithm for cluster analysis. *IOS Press*, 2003.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. Michigan State University, New Jersey: Prentice Hall, 1988. 320 p.

JAIN, P. *Large Scale Optimization Methods for Metric and Kernel Learning*. Dissertação (Mestrado) — University Of Texas At Austin, Texas, 2009.

LEYDESDORFF, L.; RAFOLS, I. Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, v. 5, p. 87–100, Janeiro 2011.

LIMA, E. L. *Espaços Métricos*. Rio de Janeiro, RJ: IMPA, 2009. 299 p.

MATHWORKS. *What is MATLAB?* 1994–2018. <https://www.mathworks.com/discovery/what-is-matlab.html>.

QIN, L.; YI, Z.; ZHANG, Y. Unsupervised surface roughness discrimination based on bio-inspired artificial fingertip. *Sensors and Actuators A: Physical*, p. 483–490, 2018.

RICHARDSON, R. J. et al. *Pesquisa social: métodos e técnicas*. São Paulo: Editora Atlas S.A., 2007. 334 p.

SANTANA, F. L. *Generalizações do Conceito de Distâncias, i-Distâncias, Distâncias intervalares e Topologia*. Tese (Doutorado) — UFRN, Natal, 2012.

SANTOS, F. dos; BASTOS, L. C. Casa da qualidade e qualidade da informação: revisão sistemática. *Perspectivas em Ciência da Informação*, v. 22, n. 1, p. 100–111, mar. 2017.

SHI, X.; WANG, W.; ZHANG, C. An empirical comparison of latest data clustering algorithms with state-of-the-art. *Indonesian Journal of Electrical Engineering and Computer Science*, 2017.

SIPSER, M. *Uma Introdução á Teoria da Computação*. [S.l.]: Cengage Learning, 2005. 488 p.

SLONIM, N.; AHARONI, E.; CRAMMER, K. Hartigan's k-means versus lloyd's k-means – is it time for a change? In: *International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 2013. p. 1677–1684.

SOUZA, J. A. *Reconhecimento de Padrões Usando Indexação Recursiva*. Tese (Doutorado) — UFSC, Florianópolis, 1999.

STASIU, R. K. *Avaliação da Qualidade de Funções de Similaridade no Contexto de Consultas por Abrangência*. Tese (Doutorado) — UFRGS, Porto Alegre, 2007.

UCI. *Machine Learning Repository*. 2018. Online, acesso em 16/10/2018. <https://archive.ics.uci.edu/ml/>.

VINH, N. X.; EPPS, J.; BAILEY, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. 2010.

WIKIPÉDIA. *RStudio*. 2018. <https://pt.wikipedia.org/wiki/RStudio>.