UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

VICTOR VIANA DE ARAÚJO SILVA

Um Método Difuso Multivariado Baseado em Medoids

Recife

2022

VICTOR VIANA DE ARAÚJO SILVA

Um Método Difuso Multivariado Baseado em Medoids

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Renata Maria Cardoso Rodrigues de Souza

Coorientador (a): Bruno Almeida Pimentel

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S586m Silva, Victor Viana de Araújo
Um método difuso multivariado baseado em medoids / Victor Viana de Araújo Silva. – 2022.
63 f.: il., fig, tab.

Orientadora: Renata Maria Cardoso Rodrigues de Souza.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências.

1. Inteligência computacional. 2. Agrupamento difuso. I. Souza, Renata Maria Cardoso Rodrigues de (orientadora). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2023-80

Victor Viana de Araújo Silva

“Um Método Difuso Multivariado Baseado em Medoids”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 15/12/2022.

BANCA EXAMINADORA

_____

Prof. Dr. Adriano Lorena Inacio de Oliveira
Centro de Informática / UFPE

_____

Prof. Dr. Getulio José Amorim do Amaral
Departamento de Estatística/ UFPE

_____

Prof. Dr. Bruno Almeida Pimentel
Instituto de Computação/UFAL
(**Coorientador**)

Dedico este trabalho a Deus, minha família, meus amigos e aos professores que me apoiaram durante todo o tempo.

AGRADECIMENTOS

Em primeiro lugar, a Deus, que fez com que meus objetivos fossem alcançados durante todos os meus anos de estudos.

Aos meus orientadores, Dra. Renata Souza e Dr. Bruno Pimentel pelos ensinamentos e orientações que me permitiram construir este trabalho.

À minha família, pela amizade incondicional, pela paciência e pelo apoio durante todo o período em que me dediquei a este trabalho.

Aos meus amigos, pelas contribuições e troca de aprendizados que contribuíram de alguma forma para a realização deste trabalho.

RESUMO

A Ánalise de agrupamentos foi inicialmente utilizada por Tyron em 1939, em que visa organizar dados que possuam características similares dentro de um mesmo grupo e no caso contrário em que os dados possuem características distintas, eles serão alocados em grupos diferentes. Ou seja, se é levado em consideração a ideia de minimizar a distância intra-grupos e maximizar a distância inter-grupos. Com isso, dentre outros benefícios, podem ser visualizadas algumas vantagens da utilização desta técnica, como por exemplo a diminuição da dimensionalidade dos dados e a extração das características dos grupos. O principal método de agrupamento difuso é o *Fuzzy C-Means* (FCM), o qual possui algumas desvantagens tal como considerar que todos os grupos possuem formas esféricas e ser altamente influenciado em casos de conjuntos de dados ruidosos. O *Fuzzy C-medoid* (FCMdd) foi criado com o intuito de tentar mitigar esta problemática, porém não leva em consideração o impacto de cada variável no cálculo dos graus de pertinências. Diante desse cenário, o *Multivariate Fuzzy C-means* (MFCM) foi criado com o intuito de levar em consideração o efeito de cada variável no cálculo dos protótipos, porém, utiliza a média para o cálculo dos centróides podendo ser fortemente influenciada negativamente por dados ruidosos. Este trabalho introduz o método *Multivariate Fuzzy C-medoids* (MFCMdd), em que como o próprio nome já diz, os graus de pertinência são multivariados e utilizam observações do próprio conjunto de dados para serem os centróides, também conhecidos como *medoids*. Diante deste cenário, o método proposto MFCMdd, é comparado com os outros três métodos (FCM, FCMdd e MFCM) abordados de acordo com as métricas utilizadas para avaliação dos algoritmos, sendo elas o Índice de Rand Ajustado e o F-score. Com o objetivo de avaliar o desempenho dos métodos, um estudo comparativo em relação aos agrupamentos difusos usando o experimento Monte Carlo é realizado. Além disso, foram planejados experimentos com dados sintéticos e reais. Os resultados mostraram que o método proposto MFCMdd, perante o MFCM é preferível quando se há conjuntos de dados sem ruído ou também quando os conjuntos de dados possuem caráter esférico com dados ruidosos.

Palavras-chave: agrupamento difuso; grau multivariado; fuzzy c-means; fuzzy c-medoids; aplicação.

ABSTRACT

The cluster analysis was firstly used by Tyron in 1939, with aims to organize data with similar characteristics within the same group, while data with distinct aspects are assigned to different groups. With this, it's possible to see that some of the benefits of using this technique are the reduction of data dimensionality and the extraction of group characteristics. The most commonly used fuzzy clustering method is Fuzzy C-Means (FCM), but it has some drawbacks, such as considering all groups to be spherically shaped and in cases of noisy data sets, is considered being highly influenced by . To address this issue, the Fuzzy C-medoid (FCMdd) was developed, but it does not account for the impact of each variable when calculating membership degrees. Given this scenario, the Multivariate Fuzzy C-means (MFCM) was developed to account for the effect of each variable in the calculation of the prototypes, but it relies on the means to calculate the centroids, which can be heavily influenced by noisy data. This paper introduces the Multivariate Fuzzy C-medoids (MFCMdd) method, in which the membership degrees are multivariate and the centroids, also known as medoids, are observations from the data set itself. Given this scenario, the proposed method MFCMdd is compared to the other three methods (FCM, FDMdd, and MFCM) based on the metrics used to evaluate the algorithms, which are the Adjusted Rand Index and the F-score. A comparative study of fuzzy clustering using Monte Carlo experiment is performed to evaluate the performance of the methods. Experiments with both synthetic and real data were also carried out. The results showed that the proposed method MFCMdd, rather than MFCM, is preferable when the data sets are noiseless or have a spherical character with noisy data.

Keywords: fuzzy clustering; multivariate degree; fuzzy c-means; fuzzy c-medoids; application.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

γ	Letra grega Gama
$\in$	Pertence
δ	Delta
θ	Teta
σ	Sigma
μ	Mi

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Análise de Agrupamentos pode ser compreendido como uma técnica estatística que em que dado um conjunto de dados, se busca reunir em um mesmo grupo observações que possuam um maior grau de similaridade, enquanto que observações que não possuam um alto grau de dissimilaridades são alocadas em grupos distintos (FREI, 2006).

Áreas como análise de dados, pesquisa de mercado, reconhecimento de padrões, amostragem para formação de estratos, regiões geográficas para analisar suas peculiaridades, entre outras, vem utilizando do auxílio da Análise de Agrupamentos para extrair informações úteis.

Prass et al. (2004) comenta que esse uso em grande escala de técnicas acerca deste tema faz sentido devido à grande quantidade de vantagens que se pode extrair:

- auxilia no entendimento dos atributos do conjunto de dados;
- pode ser usada na geração de hipóteses;
- possibilita ao usuário encontrar grupos úteis;
- permite predição com base nos grupos formados, ao invés dos dados brutos;
- possibilita o desenvolvimento de um esquema de classificação para dados novos.

Diante deste cenário, algumas técnicas de agrupamento foram criadas, como é o caso do *K-means*, técnica mais famosa devido à sua facilidade de implementação e por ser uma das pioneiras do ramo. Seu principal intuito é alocar as observações de um conjunto de dados em um número de grupos pré-especificado de grupos, em que a qualidade do agrupamento pode ser mensurada através de uma função objetivo (MACQUEEN, 1967).

Outro algoritmo que motivou a construção dessa dissertação foi o PAM *(Partional Around Medoids)* ou *k-medoids* (KAUFMANN, 1987; KAUFMAN; ROUSSEEUW, 1990). Este algoritmo possui como principal contraste perante o *k-means* a escolha de observações do conjunto de dados para serem os centróides, denominados de *medoids*. Neste formato, há uma maior facilidade na interpretabilidade dos centróides dos clusters perante o *K-means* que utiliza a média das observações dos grupos para capturar os centróides.

Uma extensão do algoritmo k-medoid é o CLARA (*Clustering Large Applications*) em que ele é indicado quando se há o contexto de grandes conjuntos de dados (KAUFMAN; ROUSSEEUW, 1990), com o intuito de reduzir o tempo computacional nessas situações. Há estudos que comparam o k-medoid com o CLARA, como por exemplo em (LUCASIUS; DANE; KATEMAN, 1993). Este método considera uma pequena amostra do conjunto de dados e aplica o algoritmo *k-medoid* para gerar um conjunto ótimo de medoides para a amostra.

Como parte fundamental deste trabalho, os algoritmos podem utilizar a Lógica *Fuzzy*, também chamada de lógica nebulosa ou lógica difusa, na qual foi difundida em 1965 através do artigo Fuzzy Sets (ZADEH, 1965), escrita pelo pelo professor Lofti Zadeh. Entretanto, há controvérsias em relação à quem foi o pioneiro do tema. Cox (1994) afirma que em meados de 1920 um polonês chamado Jan Luasiewicz apresentou pela primeira vez uma lógica baseada nos princípios da incerteza em que era aceitável valores não precisos.

Cox (1994) também comenta que a principal distinção entre a Lógica *Fuzzy* perante a booleana é perspectiva de poder mensurar o grau de aproximação (pertinência) da solução exata, trazendo um contexto mais próximo do mundo real, ao invés de trabalhar com valores extremos.

Altrock (1997) compreende a Lógica *Fuzzy* como uma tentativa de se aproximar do raciocínio humano, ou seja, como os humanos relacionam as informações procurando respostas aproximadas perante os problemas. Por isso, o grande objetivo desta lógica é solucionar problemas que contenham incertezas nas informações.

Neste contexto, algumas técnicas foram bastante difundidas, como por exemplo, o *Fuzzy C-means*, proposto por Bezdek (1981), que é o método de agrupamento difuso mais conhecido e geralmente possui bons resultados na abordagem difusa, além de possuir uma relativa facilidade na implementação do algoritmo. Uma desvantagem desse método é o fato dele não possuir bons desempenhos quando o conjunto de dados possui dados aberrantes (*outliers*).

Há metodos que não utilizam médias para o cálculo dos protótipos, como é o caso do *Fuzzy C-medoids* (FCMdd). É um método bastante similar ao FCM, porém, ao invés da utilização da média, uma observação do próprio conjunto de dados minimiza a distância desse ponto para os demais do grupo, conhecida como *medoid*. Ou seja, a principal diferença perante o FCM é encontrada na lista de formação dos centróides (PINHEIRO; ALOISE; BLANCHARD, 2020).

Quanto à abordagem difusa multivariada, algumas vantagens são perceptíveis, como por exemplo, interpretar a relevância de cada observação para um determinado grupo de acordo com cada variável, a capacidade de obter mais informações dos conjuntos de dados ajudando

a melhorar a qualidade dos agrupamentos e por último uma nova alternativa de agrupamentos utilizando o contexto multivariado (PIMENTEL; SOUZA, 2016).

O *Multivariate Fuzzy C-means* (MFCM), criado por Pimentel e Souza (2013) possibilita que os graus de pertinência sejam diferentes para cada variável (desta forma, ele passa a ser chamado de multivariado), utilizando a média para o cálculo da lista dos centróides.

Entendendo que o aspecto multivariado traz virtudes antes não mapeadas, a seguinte dissertação visa elaborar o método *Multivariate Fuzzy C-medoids* (MFCMdd), cujo intuito é melhorar a qualidade do agrupamento perante o método MFCM quando se há o cenário com conjuntos de dados com classes sobrepostas e dados ruidosos.

1.2 OBJETIVOS

Este trabalho visa apresentar um método de agrupamento difuso em que os valores do grau de pertinência das observações para cada grupo são influenciados pelas variáveis. Ou seja, o nível de similaridade de uma observação com o protótipo é dado de acordo com cada variável do conjunto de dados. A principal proposta deste estudo é aumentar a quantidade de informações que o algoritmo pode receber utilizando observações do conjunto de dados (medoids) como centróides, além de conferir se aplicando esta técnica multivariada, ocorrerá melhora na qualidade dos agrupamentos comparado com outros já existentes na literatura. É importante destacar que os métodos explanados, podem ser divididos em duas categorias perante a capacidade de encontrar partições, sendo elas com a presença ou sem a presença de dados aberrantes.

O método a ser apresentado é baseado no método MFCM proposto por Pimentel e Souza (2013). Apesar do MFCM levar em consideração o fator da análise das variáveis por cada grupo, por fato de utilizar médias para o cálculo do centróide, se é esperado uma certa dificuldade quando se está em um cenário com dados aberrantes. Ademais, serão apresentados duas métricas bastante difundidas neste segmento com o intuito de avaliar a qualidade do agrupamento para grupo e variável a partir da partição difusa obtida através do método proposto comparado com outros já existentes na literatura. Com isso, a finalidade deste trabalho é responder duas perguntas (hipóteses) de pesquisa:

1. Dentre as técnicas expostas neste trabalho, o método proposto (MFCMdd) obtém vantagem na qualidade dos agrupamentos para conjuntos de dados com quais características?

2. Utilizando observações do conjunto de dados (medoids) para o cálculo do centróide ao invés da média, é possível melhorar a qualidade de agrupamento difuso multivariado?

Para avaliar o desempenho dos métodos propostos e os presentes na literatura, um experimento Monte Carlo é efetuado, cujo intuito é realizar um estudo comparativo desses métodos. Neste estudo, foram realizados experimentos com dados sintéticos e aplicações com dados reais (BACHE; LICHMAN, 2013). As métricas que foram levadas em consideração para fazer essa avaliação das performances dos métodos foram o Índice de Rand Ajustado (ARI) e o F-score.

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

A dissertação está estruturada de acordo com os seguintes capítulos:

Capitulo 2 - Revisão da Literatura

Neste capítulo, será exposto um pouco da história da Análise de Agrupamentos, algumas técnicas clássicas do tema, além de algumas aplicações realizadas na literatura.

Capitulo 3 - Metodologia

O método proposto será demonstrado neste capítulo da dissertação. O método utiliza uma abordagem multivariada difusa visando encontrar uma partição, em que o algoritmo *Multivariate Fuzzy C-means* foi tomado como base (PIMENTEL, 2017), porém, ao invés de utilizar a média para cálculo dos protótipos, uma observação do próprio conjunto de dados será levada em consideração para o cálculo dos centróides, conhecida como medoid.

Capitulo 4 - Resultados

Nesta seção, serão comparados os resultados do método proposto com os métodos difundidos na literatura para agrupamento difuso. Na primeira parte, são apresentadas as métricas que serão utilizadas neste estudo, sendo elas, o Índice de Rand Ajustado (ARI) e o *F-score*, fazendo uma comparação entre a partição a priori e a partição difusa obtida pelos métodos. Na segunda parte, as configurações dos dados sintéticos são apresentadas. Na terceira parte, conjuntos de dados reais são submetidos à esses métodos de agrupamentos e posteriormente avaliados. Por último, na quarta parte, o tópico de discussão aborda uma síntese dos resultados da simulação utilizando experimento de Monte Carlo, além das análises efetuadas para as

aplicações.

Capitulo 5 - Conclusões

No último capítulo dessa dissertação, serão apresentadas as conclusões obtidas acerca do estudo com dados simulados e com dados reais, comentando as vantagens e deficiências do método proposto perante os existentes na literatura, além de trabalhos futuros.

2 REVISÃO DA LITERATURA

2.1 ANÁLISE DE AGRUPAMENTOS

A Análise de Agrupamentos teve origem em 1939 (TRYON, 1939), em que a definiu como um processo lógico formulado como procedimento, por meio do qual as observações são reunidas em grupos com base em sua semelhança e dissimilaridade. Posteriormente, Aldenderfer e Blashfield (1984) trouxeram este tema de uma forma mais palpável para o leitor. Até aqueles com pouca ou nenhuma experiência no assunto conseguiam ter facilidade através de um guia programático, com técnicas estatísticas, métodos de validação e programas de software compatíveis que caminhavam em conjunto com o avanço dos computadores daquela época, facilitando os cálculos matriciais.

Percebe-se um grande crescimento no interesse em compreender, processar e reduzir dimensionalidade dos dados de uma forma otimizada (XU; WUNSCH, 2005). Portanto, um dos temas mais importantes no âmbito do reconhecimento de padrões é a Análise de Agrupamentos. Ele é considerado um processo de classificação não-supervisionada, ou seja, não se sabe previamente o rótulo que cada observação deve receber, de uma forma que os elementos semelhantes fiquem agrupados em um mesmo grupo, com a finalidade de otimizar a função objetivo.

Com base nisto, técnicas de Agrupamento são feitas com base nas similaridades, ângulo, curvatura, conectividade, distância, intensidade ou simetria entre as suas observações (BEZDEK, 2013). Roses e Leis (2002) comentam que os grupos obtidos devem exibir um homogeneidade interna (dentro de cada grupo), como uma grande heterogeneidade externa (entre os grupos). Em outras palavras, os objetos de um grupo devem ser mais similares entre si do que a objetos de outros grupos levando em consideração minimizar a distância intra-grupos e maximizar a distância inter-grupos.

Um conceito bastante difundido é que a Análise de Agrupamentos pode ser compreendida como técnicas estatísticas multivariadas em que dada uma amostra de n objetos, com p características, também conhecidas como variáveis, procura-se agrupá-las em k grupos (BUSSAB; MIAZAKI; ANDRADE, 1990).

É importante salientar a diferença entre técnicas de classificação e de agrupamentos. Na primeira, os dados são alocados a partir de observações sob o qual já se conhece o rótulo verdadeiro, já na segunda que é o enfoque dessa pesquisa, não se conhece nenhuma classe

dos dados, se é definido quantos grupos serão utilizados e as observações são discriminadas dentre essa quantidade de grupos anteriormente definida.

Além disso, esse estudo apresentará os experimentos da simulação com dados ruidosos. Hu, An e Yu (2010) diz que conjuntos de dados em aplicações do mundo real geralmente são contaminados por ruído, por isso, efetuou aplicações com dados existentes acrescentando dados ruidosos com percentuais de 6%,12%,18%,24% e 30%.

2.2 ALGUMAS TÉCNICAS DE AGRUPAMENTO CLÁSSICAS

Analisando o formato no qual os algoritmos de agrupamentos encontram seus grupos, há duas possibilidades, sendo elas hierárquico e particional. No agrupamento hierárquico, o algoritmo gera uma série aninhada de partições com base em um critério que mescla ou divide clusters. No primeiro caso, ou seja na mescla dos clusters, temos um cenário de aglomeração, começando com cada observação em um único cluster cada e a partir daí, eles serão mesclados de acordo com as suas semelhanças. Ainda no contexto dos algoritmos hierárquicos, com base no critério de divisão começa com todos os dados em um único cluster e a partir daí as divisões são efetuadas até chegar no cenário final. Em relação ao algoritmo particional, se é obtido apenas uma única partição ao invés de uma série aninhada de partições. Um ponto negativo desta metodologia é que em alguns casos há uma certa dificuldade em encontrar o número ideal de clusters de uma forma automática (CRUZ; OCHI, 2011). Porém, em contrapartida, esse método é mais indicado quando se trabalha com grandes conjuntos de dados, pois o tempo desprendido para esse método é menor do que para métodos hierárquicos (JAIN; MURTY; FLYNN, 1999).

Os métodos particionais são classificados em duas categorias, sendo elas: rígido e não-rígido (JAIN; MURTY; FLYNN, 1999). No rígido, as observações pertencem exclusivamente a um grupo no intervalo $\{0,1\}$, sendo 1 se a observação pertencer a esta classe e 0 caso não pertença (PAL et al., 2005). No não-rígido, podem ser encontrados algoritmos difusos, probabilísticos ou possibilísticos (PAL et al., 2005). O grande benefício do não-rígido perante o rígido, é a sua habilidade em encontrar grupos sobrepostos, ou seja, graus de pertinência não discretos podem ser analisados.

O Fuzzy c-means (FCM), aprimorado por Bezdek, Ehrlich e Full (1984), é o método de agrupamento não-rígido mais conhecido. No FCM, o objetivo é minimizar a função critério baseada na lógica fuzzy, levando em consideração a similaridade de elementos e centros de

cluster. É mais proveitoso para conjuntos de dados que possuem grupos com perfis mais sobrepostos. Devido ao fato do FCM ser facilmente implementado e ter obtido resultados relevantes em muitas aplicações, tornou-se uma importante ferramenta para reconhecimento de padrões (JAIN; MURTY; FLYNN, 1999). A seguir, é apresentado o passo-a-passo do algoritmo fuzzy c-means (FCM):

Passo 1: Inicialmente, é definido o número de grupos *c*, o $m \in (1, \infty)$ que significa o parâmetro de fuzzificação, e o ϵ que significa um valor positivo arbitrário que serve como critério de parada, além da inicialização dos graus de pertinência de cada observação perante cada grupo;

Passo 2: Calcule os centróides iniciais, através de médias ponderadas;

$$c_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m},$$

em que u_{ik} é o grau de pertinência do grupo i na observação k, o x_k é a k-ésima observação e por último, o n é o n é o número total de observações.

Passo 3: Calcule a distância dos pontos para os centróides iniciais e aloque as observações para os grupos;

Passo 4: Atualize os graus de pertinência e gere a nova matriz de pertinências.

Repetir os dois últimos passos até que o critério de parada ϵ seja atingido.

Passo 5: Por último, desfuzzificar os valores de pertinência.

Porém, o FCM parte do pressuposto que todas as variáveis possuem o mesmo grau de pertinência no campo multivariado, podendo afetar negativamente o desempenho do algoritmo. Assim sendo, técnicas multivariadas podem ser uma saída para esta problemática. O método *Multivariate Fuzzy C-Means* (MFCM), criado por Pimentel e Souza (2013) permite que os graus de pertinência *fuzzy* sejam diferentes por variável. Isto acarreta em algumas vantagens, visto que dessa forma há a possibilidade de saber a relevância que cada observação traz para cada grupo de acordo com cada variável, acarretando em agrupamentos mais eficazes.

Passo 1: Inicialmente, é definido o número de grupos *c*, o parâmetro de fuzzyficação $m = 2$, o ϵ que significa um valor positivo arbitrário que serve como critério de parada, além da inicialização aleatória dos graus de pertinência u_{ijk} de cada observação *k* perante cada variável *j* e cada grupo *i*, para $i = 1, \ldots, c$;

Passo 2: Atualize os graus de pertinência da observação k pertencente ao grupo i perante cada variável *j*. Calcule o protótipo y_{ij} referente ao grupo i, sendo x_{ijk} a k-ésima observação alocada para o grupo *i* e variável *j*, através da equação abaixo:

$$y_{ij} = \frac{\sum_{k=1}^{n} (u_{ijk})^m x_{jk}}{\sum_{k=1}^{n} (u_{ijk})^m};$$

Passo 3: Atualize o protótipo y_i do grupo i. Ou seja, atualize o grau de pertinência *fuzzy* u_{ijk} da observação k pertencente ao grupo i na variável j, baseado na distância usando a equação abaixo, em que d_{ijk} é calculada pela distância euclidiana entre a observação e o protótipo perante cada variável:

$$u_{ijk} = \left[\sum_{h=1}^{c} \sum_{l=1}^{p} \left(\frac{d_{ijk}}{d_{hlk}} \right)^{(1/m-1)} \right]^{-1};$$

Passo 4: Repetir os dois passos anteriores até que o critério de parada seja atingido.

Outros algoritmos de agrupamento utilizam medoids ao invés de médias. Algumas vantagens a cerca desses métodos são que eles podem trabalhar com dados não numéricos, são mais resistentes a ruídos e além disso, facilita a interpretação dos dados, visto que os centróides de cada grupo passam a ser representados por observações do próprio conjunto de dados (KRISHNAPURAM; JOSHI; YI, 1999). Abaixo, podemos ver um resumo do algoritmo *fuzzy c-medoids* (FCMdd), que se assemelha bastante ao FCM, porém, a principal diferença é encontrada na lista de formação dos centróides dos grupos.

Passo 1: Inicialmente, é definido o número de grupos *c*, o parâmetro de fuzzyficação $m = 2$, os centróides iniciais dos grupos que neste caso são observações do próprio conjunto

de dados, e o ϵ que significa um valor positivo arbitrário que serve como critério de parada, além da inicialização dos graus de pertinência de cada observação perante cada grupo;

Passo 2: Calcule a distância dos pontos para os centróides e aloque as observações para os grupos;

Passo 3: Atualize os graus de pertinência e gere a nova matriz de pertinências, em que $d(x_k, v_i)$ é a distância euclidiana entre a observação x_k e o centróide/medoide do grupo em questão v_i e a $d(x_k, v_l)$ é a distância euclidiana das observações para cada um dos grupos v_l.

$$u_{ik} = \left[\sum_{l=1}^{c} \frac{d(x_k, v_i)}{d(x_k, v_l)}^{1/m-1} \right]^{-1};$$

Passo 4: Repetir os dois últimos passos até que o critério de parada ϵ seja atingido.

2.3 APLICAÇÕES ACERCA DA ANÁLISE DE AGRUPAMENTO

Diversos ramos como aprendizado de máquina, biologia computacional, mineração de dados, reconhecimento de padrão e visão computacional tem empregado algoritmos de agrupamento em suas simulações ou aplicações. Isso ocorre devido ao fato de nem sempre se ter de forma palpável os rótulos de dados e mesmo assim possuir o interesse na compreensão ou redução da dimensionalidade dos dados.

Dave (1991) introduz uma abordagem sobre dados ruidosos, em que eles possam ser atribuídos à classe de ruído e a capacidade que alguns algoritmos como o *k-means* e o *fuzzy c-means*, conseguem detectar realtivamente bem os clusters, mesmo com dados ruidosos. Essa abordagem traz gatilhos como por exemplo a aplicação desta técnica à algoritmos de análise de regressão.

Groenen e Jajuga (2001) apresentaram um modelo de agrupamento *fuzzy* baseado na raíz quadrada da distância de Minkowski que inclui distâncias euclidianas quadradas e não quadradas, além da distância L1. Com o intuito de encontrar um mínimo global, foi comparado uma estratégia especial chamada etapas *fuzzy* com redes de agrupamento de Kohonen fuzzy

(FKCN) e *multistart*. Para avaliar a qualidade dos métodos, são realizados dois experimentos numéricos e um estudo de simulação.

Labroche (2010) implementou dois algoritmos de agrupamento *fuzzy c-medoids* para conjuntos de dados muito grandes. Esses algoritmos fuzzy se propõem a trabalhar com dados contínuos, ou seja, dados que não estão disponíveis de uma vez só, mantendo a abordagem *fuzzy*. Esses dois algoritmos são aplicados e avaliados para conjuntos de dados artificiais e reais e como conclusão as novas abordagens têm um desempenho próximo, se não melhor, do que os algoritmos existentes.

Egrioglu et al. (2011) propôs uma abordagem de séries temporais utilizando a Lógica *Fuzzy*. Neste artigo, foi utilizado a técnica de agrupamento *fuzzy* conhecida como Gustafson–Kessel para a fuzzyficação de séries temporais. O autor enfatiza que para o cenário de séries temporais consiste em três estágios, sendo elas: fuzzyficação, determinação de relações *fuzzy* e por último, a desfuzzyficação.

No ramo de atendimento médico, visando explorar os impactos das falhas do processo de atendimento ambulatorial geriátrico em pacientes idosos de Taiwan, a teoria dos conjuntos *fuzzy* foi utilizada juntamente com a técnica de preferência de ordem por similaridade na tomada de decisão de múltiplos critérios para classificar os riscos de falha na Análise de Modo e Efeitos de Falha de Assistência Médica (HFMEA) (KUO; WU; HSU, 2012). Como resultados, a avaliação de risco do processo de atendimento ambulatorial geriátrico foi mais objetiva quando analisada com dados quantitativos.

Sabzekar e Naghibzadeh (2013) utilizaram *Support Vector Machines* (SVM) de forma híbrida com o método *fuzzy c-means*. Primeiramente o fuzzy c-means particiona os dados em grupos apropriados. Logo depois, as amostras com altos valores de associação em cada grupo são selecionadas para treinar um classificador de *Support Vector Machines* a restrições relaxadas multiclasse. Por fim, os rótulos de classe dos pontos de dados restantes são previstos pelo último classificador. Os resultados obtidos foram superiores ao método *fuzzy c-means* padrão diante dos experimentos propostos.

Na área de segmentação de imagens, Zhao, Fan e Liu (2014) usaram um algoritmo FCM suprimido baseado em seleção ótima com informações espaciais não locais de autoajuste para melhorar o desempenho da segmentação em imagens com o cenário de alto ruído. Este método é aplicado a Berkeley e outras imagens reais fortemente contaminadas por ruído. Os experimentos de segmentação de imagens demonstraram superioridade do método proposto sobre outros algoritmos *fuzzy*.

Zhang et al. (2014) propuseram um algoritmo FCM que usa uma estratégia de heurística genética para buscar pesos de intervalo mais apropriados para os atributos de dados, cujo objetivo é melhorar o desempenho do agrupamento. Como resultados, o agrupamento ponderado por intervalo pode atuar como um operador de otimização baseado no agrupamento ponderado numérico tradicional. Além disso, os efeitos da perturbação do peso do intervalo no desempenho do agrupamento podem ser diminuídos.

No intuito de identificar a importância de cada variável para cada cluster e melhorar a qualidade do clustering, dois algoritmos *fuzzy c-means* multivariados com ponderação foram criados. Os pesos visam representar a importância de cada variável diferente para cada grupo e melhorar a qualidade do agrupamento. Esses algoritmos foram aplicados a dados sintéticos e do repositório da UCI (PIMENTEL; SOUZA, 2016), mostrando a utilidade dos algoritmos com ponderação.

3 MÉTODO PROPOSTO

Apesar de o método Multivariate Fuzzy C-means utilizar uma abordagem em que as variáveis são levadas em consideração no ato do cálculo do grau de pertinência, a utilização das médias no cálculo dos centróides faz com que nos cenários em que há presença de dados aberrantes espera-se que não produzam bons resultados. Desta forma, uma abordagem multivariada que utiliza observações do próprio conjunto de dados para serem os centróides dos grupos é proposta.

O propósito do método Multivariado Fuzzy C-medoids (MFCMdd) é buscar uma matriz de protótipos $\mathbf{Y}^*$ e uma matriz de graus de pertinência $\mathbf{U}^*$, como demonstrado abaixo:

$$J^1(\mathbf{Y}^*, \mathbf{U}^*) = min\left\{J^1(\mathbf{Y}, \mathbf{U}) : \mathbf{Y} \in \mathbb{Y}^c, \mathbf{U} \in \mathbb{U}^n\right\},$$

em que, $\mathbb{Y}^c = \mathbb{Y}$ x ... x $\mathbb{Y}$ e $\mathbb{Y} = \mathbb{R}$ x ... x $\mathbb{R} = \mathbb{R}^p$, onde $\mathbb{Y}$ é o universo de representação dos protótipos, tal que $\mathbf{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{ip}) \in \mathbb{Y}$ e $\mathbf{Y} \in \mathbb{Y}^c$. Além disso, considere $\mathbf{U} = [\mathbf{u}_k]$ sendo uma matriz de matrizes de graus de pertinência multivariados, na qual, para cada observação k de Ω há uma matriz c x p de graus de pertinência multivariados $\mathbf{u}_k = [u_{ijk}]$, com k variando no intervalo (k - 1,...,n). O u_{ijk} é o grau de pertinência da observação k para o grupo i em relação à variável j e por último, $\mathbb{U}^n = \mathbb{U}$ x ... x $\mathbb{U}$, sendo $\mathbb{U}$ o universo de representação da matriz de graus de pertinência multivariados $\mathbf{u}_k$ em que $\mathbf{u}_k \in \mathbb{U}$ e $\mathbf{U} \in \mathbb{U}^n$.

O objetivo do algoritmo é minimizar a função objetivo apresentada abaixo:

$$J^1(\mathbf{Y}, \mathbf{U}) = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m d_{ijk},$$

em que d_{ijk} é a distância Euclidiana que mensura a dissimilaridade entre x_{kj} do objeto k e y_{ij} do protótipo do grupo C_i para uma dada variável j. Como pode ser vista adiante:

$$d_{ijk} = (x_{kj} - y_{ij})^2$$

No que se diz respeito a uma única variável j, os protótipos consideram o grau de pertinência e os valores quantitativos de cada observação de Ω, em que $\Omega = \{1, \ldots, k, \ldots, n\}$, ou seja, um conjunto de n objetos listados por k. Cada observação k é caracterizada por um vetor de variáveis quantitativas dado por $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kp})$ representado por p variáveis indexadas por j onde $x_{kj} \in \mathbb{R}$.

Proposição 3.1.: Deixando o grau de pertinência u_{ijk} fixo, o protótipo q que minimiza o critério J^1 é atualizado, utilizando a seguinte expressão:

$$q = \underset{1 \leq i \leq c}{\arg\min} \sum_{j=1}^{p} \sum_{k=1}^{n} (u_{ijk})^m d_{ijk}$$

$$y_i = x_q$$

Proposição 3.2.: Deixando o protótipo y_i fixo, o grau de pertinência u_{ijk} que minimiza o critério J^1 é atualizado usando a equação abaixo, respeitando a restrição $\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ijk} = 1$:

$$u_{ijk} = \left[\sum_{h=1}^{c} \sum_{l=1}^{p} \left(\frac{d_{ijk}}{d_{hlk}} \right)^{(1/m-1)} \right]^{-1};$$

Diante do contexto em que o grau de pertinência multivariado u_{ijk} de uma dada observação x_k pertencente ao grupo C_i para uma dada variável j, é importante ressaltar as premissas que devem ser levadas em consideração, sendo elas:

1. $u_{ijk} \in [0, 1]$ para todo i, j e k;
2. $0 < \sum_{j=1}^{p} \sum_{k=1}^{n} u_{ijk} < n$,para todo i;
3. $\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ijk} = 1$

No método proposto MFCMdd, assim como no MFCM, cada observação tem p graus de pertinência para um dado grupo. Porém, ao término da execução do algoritmo, caso o usuário deseje saber a qual grupo cada observação foi alocado, é necessário encontrar a partição final, isto é, uma partição rígida $P = \{C_1, \ldots, C_i, \ldots, C_c\}$ em que $\cup_{i=1}^{c} C_i = \Omega$ e $\cap_{i=1}^{c} C_i = \emptyset$. Cada observação $\mathbf{x_k}$ deve ser alocada ao grupo C_{i*}, respeitando a relação abaixo:

$$i* = \underset{1<i<c}{\operatorname{argmax}}\, \delta_{ik},$$

em que δ_{ik} representa o grau de pertinência para cada grupo de um dado objeto k. A matriz de graus de pertinência por cada grupo $\Delta = [\delta_{ik}]$ é obtida através da equação abaixo:

$$\delta_{ik} = \sum_{j=1}^{p} u_{ijk},$$

respeitando as seguintes condições:

1. $\delta_{ik} \in [0,1]$ para todo i e k;
2. $0 < \sum_{k=1}^{n} \delta_{ik} < n$,para todo i;
3. $\sum_{i=1}^{c} \delta_{ik} = 1$, para todo k.

Algoritmo MFCMdd (Ω, c)

Sendo o conjunto de dados Ω e o número de grupos c, o algoritmo MFCMdd segue o passo-a-passo;

1. Sendo $\epsilon > 0$. Fixe c, $2 < c < n$; fixe m, $1 < m < \infty$; fixe T e estabeleça $\epsilon > 0$; Inicialize aleatoriamente $u_{ijk} (i = 1, \ldots, c; j = 1, \ldots, p$ e $k = 1, \ldots, n)$ da observação k pertencente ao grupo C_i para a variável j tal que $u_{ijk} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ijk} < n$ e $\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ijk} = 1$, para todo $k \in \Omega$;
2. $t \leftarrow 0$;
3. $J^1(t) \leftarrow 0$;
4. $J^1(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \sum_{k=1}^{n} (u_{ijk})^m d_{ijk}$;
5. Durante o tempo em que $|J^1(t) - J^1(t+1)| > \epsilon$ e $t < T$. O próximo passo deve ser realizado, sendo ele a principal diferença entre os métodos MFCM e MFCMdd.
6. **Atualize a matriz de protótipos** Y**:** Fixe o grau de pertinência u_{ijk} e atualize os protótipos y_i, explanado na Proposição 3.1.
7. **Atualize a matriz de graus de pertinênia** U**:** Fixe o protótipo $y_i (i = 1, \ldots, c; j = 1, \ldots, p$ e $k = 1, \ldots, n)$, a seguir, atualize o grau de pertinência u_{ijk} usando a Proposição 3.2;
8. $J^1(t) \leftarrow J^1(t+1)$;
9. $J^1(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \sum_{k=1}^{n} (u_{ijk})^m d_{ijk}$;
10. $t \leftarrow t + 1$;

11. **Fim. Enquanto,**

12. **Calcule a matriz de grau de pertinência por cada grupo** Δ: Junte os graus de pertinência multivariados usando a Equação $\delta_{ik} = \sum_{j=1}^{p} u_{ijk}$.

Retorne as matrizes $\mathbf{Y}$ **e** $\mathbf{U}$.

Segundo Diday e Simon (1976), as características de convergência deste tipo de algoritmo podem ser investigadas através de duas séries, sendo elas: $v_t = (\mathbf{Y}^t, \mathbf{U}^t) \in \mathbb{Y}^c$ x $\mathbb{U}^n$ e $\omega_t = J^1(\mathbf{Y}^t, \mathbf{U}^t)$, $t = 0, 1, 2, \ldots, T$. O algoritmo começa da série inicial $v_0 = (\mathbf{Y}^0, \mathbf{U}^0)$ e calcula os próximos termos da série v_t até convergir (atingir um valor estacionário) segundo o critério J^1.

Proposição 3.3.: A série $w_t = J^1(v_t)$ decai a cada interação e converge.

Proposição 3.4.: A série $v_t = (\mathbf{Y}^t, \mathbf{U}^t)$ converge.

Portanto, é possível perceber a semelhança entre o passo-a-passo do método MFCM e MFCMdd, em que a principal diferença está na escolha dos protótipos, em que no primeiro método é utilizada uma média para esta finalidade, já o método proposto visa utilizar uma observação do conjunto de dados para ser o novo protótipo.

4 EXPERIMENTOS E RESULTADOS

Este capítulo apresentará as métricas levadas em consideração nesta dissertação, os resultados das análises de experimentos através de simulações feitas com os métodos do *Fuzzy c-means* (FCM), *Fuzzy c-medoids* (FCMdd), *Multivariate Fuzzy c-means* (MFCM) e o *Multivariate Fuzzy c-medoids* (MFCMdd). Todas essas simulações foram analisadas com 0%,10% e 20% de ruído. Além disso, serão apresentados resultados desses métodos de agrupamento para bases de dados reais extraídos da UCI.

4.1 MÉTRICAS

Nesta seção serão abordadas as métricas que foram utilizadas para este estudo na área da análise de agrupamento. A primeira métrica foi o Índice de Rand Ajustado, porém, antes de se aprofundar nela é importante comentar sobre a versão pioneira do Índice de Rand, criada em 1971.

4.1.1 Índice de Rand

Dado um conjunto de n objetos $S = \{O_1, ..., O_n\}$, suponha que $U = \{u_i, ..., u_R\}$ e $V = \{v_i, ..., v_C\}$ representam duas partições das observações que serão comparadas, sendo U um critério externo e V um resultado do agrupamento, considere:

- a é o número de pares de elementos que são colocados na mesma classe em U e no mesmo grupo em V;
- b é o número de pares de elementos que são colocados na mesma classe em U, mas não no mesmo grupo em V;
- c é o número de pares de elementos que são colocados no mesmo grupo em V, porém, não são colocados na mesma classe em U;
- d é o número de pares de elementos que são colocados em diferentes classes em U e diferentes grupos em V;

É importante destacar que a e d podem ser compreendidos como acordos e b e c como desacordos. Dito isto, o Índice de Rand (RAND, 1971) é exposto da seguinte forma:

$$IR = \frac{a+d}{a+b+c+d};$$

A interpretação desta métrica pode ser dada em que ela varia entre 0 e 1 e quando temos o cenário em que duas partições concordam perfeitamente, o resultado do Índice de Rand é 1, da mesma forma que se as partições dscordam completamente, o resultado deste índice é 0. Porém, esse Índice de Rand clássico não considera informações de partições e consequentemente ele não é apropriado para avaliação de agrupamentos difusos.

O Índice de Rand Ajustado proposto por Hubert e Arabie (1985) utiliza a distribuição hipergeométrica generalizada como o modelo de aleatoriedade, ou seja, as partições U e V são escolhidas aleatoriamente da forma que o número de objetos nas classes e clusters sejam fixos. Considere:

- n_{ij} o número de objetos que estão na classe u_i e no cluster v_j
- $n_{i.}$ e $n_{.j}$ os números de objetos na classe u_i e no cluster v_j, respectivamente.

Pode-se construir uma tabela para melhor ilustração do que foi comentado acima (YEUNG; RUZZO, 2001).

Tabela 1 – Notação para tabela de contingência para comparar duas partições

Classes/Grupos	v_1	v_2	...	v_C	**Soma**
u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	...	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R.}$
Soma	$n_{.1}$	$n_{.2}$	...	$n_{.C}$	$n_{..} = n$

Fonte: Yeung e Ruzzo (2001)

Através desta tabela e de alguns cálculos algébricos o Índice de Rand Ajustado pode ser expresso da seguinte forma:

$$IRA = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \left[\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}\right]/\binom{n}{2}};$$

Através dessa fórmula, tem-se que $-1 \leq IRA \leq 1$, ou seja, o índice recebe o valor 1 quando as duas partições são equivalentes, o valor 0 quando o acordo verificado entre as duas

partições se deve ao acaso e valores negativos quando o grau de semelhança entre as duas partições é menor que o valor esperado por uma atribuição ao acaso (ALBUQUERQUE; BARROS, 2020).

4.1.2 F-score

A segunda métrica utilizada nesta dissertação foi o *F-score*, em que ela é derivada de duas outras métricas, a precisão e o *recall*. A precisão pode ser compreendida como de todos os dados classificados como positivos, quantos são realmente positivos. O *recall* pode ser entendido como qual a porcentagem de dados classificados como positivos comparado com a quantidade real de positivos que existem em nossa amostra (DERCZYNSKI, 2016).

$$Precisão = \frac{VerdadeirosPositivos(VP)}{VerdadeirosPositivos(VP) + FalsosPositivos(FP)}$$

$$Recall = \frac{VerdadeirosPositivos(VP)}{VerdadeirosPositivos(VP) + FalsosNegativos(FN)}$$

Diante dessas duas métricas apresentadas, o F-score ou F-measure pode ser interpretado como uma média harmônica, através da seguinte relação:

$$F - score = \frac{2 * Precisão * Recall}{Precisão + Recall}$$

Através dessa fórmula, tem-se que $0 \leq F - score \leq 1$, ou seja, o maior valor que ess métrica pode receber é 1 indicando precisão e recall perfeitos e o menor valor é 0, no caso de se ao menos uma das duas métricas anteriormente apresentadas (Precisão ou Recall) forem 0.

4.2 CONJUNTOS DE DADOS SINTÉTICOS

Nesta seção, quatro conjuntos de dados sintéticos foram considerados para comparar o método proposto (MFCMdd) com os métodos explanados no tópico 2.2, sendo eles o *Fuzzy c-means* (FCM), *Fuzzy c-medoids* (FCMdd) e o *Multivariate Fuzzy c-means* (MFCM). É importante dar ênfase no comparativo entre as duas técnicas multivariadas, visto que como o próprio nome diz, ambos estarão no cenário multivariado em que as variáveis do conjunto de dados também passam a ter uma maior relevância.

A seguir, serão apresentados as configurações dos conjuntos de dados sintéticos seguindo uma distribuição normal bivariada de componentes independentes para uma melhor visualização do leitor, porém, este trabalho utilizou uma abordagem tridimensional durante as simulações. Dois conjuntos de dados possuem 350 observações espalhadas em três classes, enquanto que outros dois conjuntos de dados possuem 300 observações espalhadas em três classes. Quatro modelos de configurações diferentes são considerados e serão analisados cenários em que há sobreposição entre classes.

Seja $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, um vetor de médias e uma matriz de covariância diagonal, respectivamente, denotado como:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$$

4.2.1 Configuração do modelo 1

O conjunto de dados 1 consiste em 350 pontos distribuídos em três classes de tamanhos: 200 observações (classe 1), 100 observações (classe 2) e 50 observações (classe 3). Os pontos de cada classe foram gerados de acordo com os seguintes parãmetros:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5 \\ 0 \\ 15 \end{pmatrix} \qquad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 81 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 16 \end{pmatrix}$$

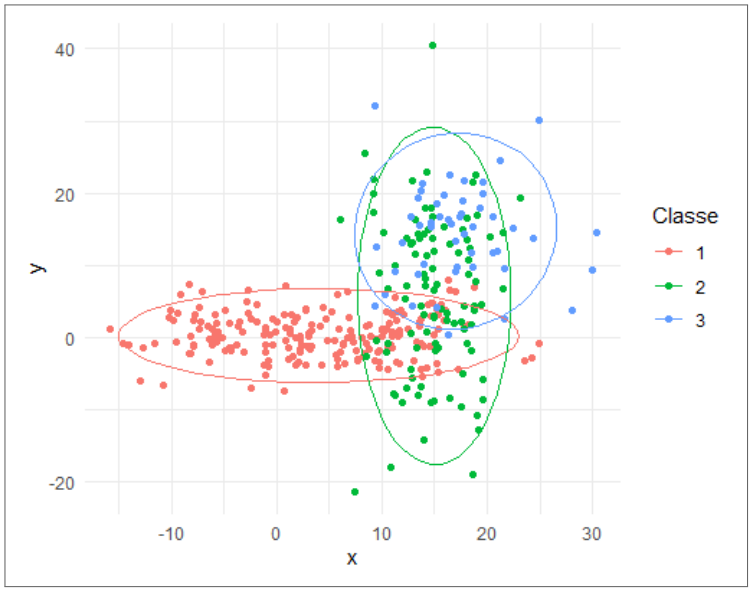
$$\boldsymbol{\mu}_2 = \begin{pmatrix} 15 \\ 5 \\ 4 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 64 \end{pmatrix}$$

$$\boldsymbol{\mu}_3 = \begin{pmatrix} 18 \\ 14 \\ 9 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 25 & 0 & 0 \\ 0 & 36 & 0 \\ 0 & 0 & 100 \end{pmatrix}$$

Através da configuração dos parâmetros acima, o primeiro conjunto de dados sintéticos traz

uma configuração onde as partições são formadas por classes elípticas de tamanhos diferentes. Abaixo, pode ser visto uma ilustração com apenas duas dimensões para uma maior facilidade de visualização do leitor (Figura 1).

Figura 1 – Conjunto de dados 1



Fonte: O autor (2022)

4.2.2 Configuração do modelo 2

O conjunto de dados 2 também consiste em 350 pontos distribuídos em três classes de tamanhos: 200 observações (classe 1), 100 observações (classe 2) e 50 observações (classe 3). Os pontos de cada classe foram gerados de acordo com os seguintes parãmetros:

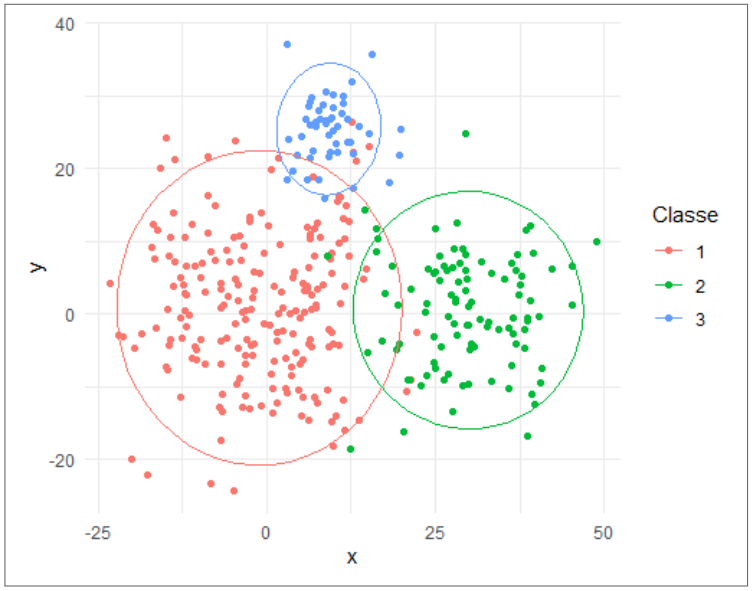
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 30 \\ 0 \\ 30 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 49 & 0 & 0 \\ 0 & 49 & 0 \\ 0 & 0 & 49 \end{pmatrix}$$

$$\boldsymbol{\mu}_3 = \begin{pmatrix} 10 \\ 25 \\ 10 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}$$

Através da configuração dos parâmetros acima, o segundo conjunto de dados sintéticos traz uma configuração onde as partições são formadas por classes esféricas de tamanhos diferentes. Abaixo, pode ser visto uma ilustração com apenas duas dimensões para uma maior facilidade de visualização do leitor (Figura 2).

Figura 2 – Conjunto de dados 2



Fonte: O autor (2022)

4.2.3 Configuração do modelo 3

O conjunto de dados 3 consiste em 300 pontos distribuídos em três classes cada uma com 100 observações. Os pontos de cada classe foram gerados de acordo com os seguintes parãmetros:

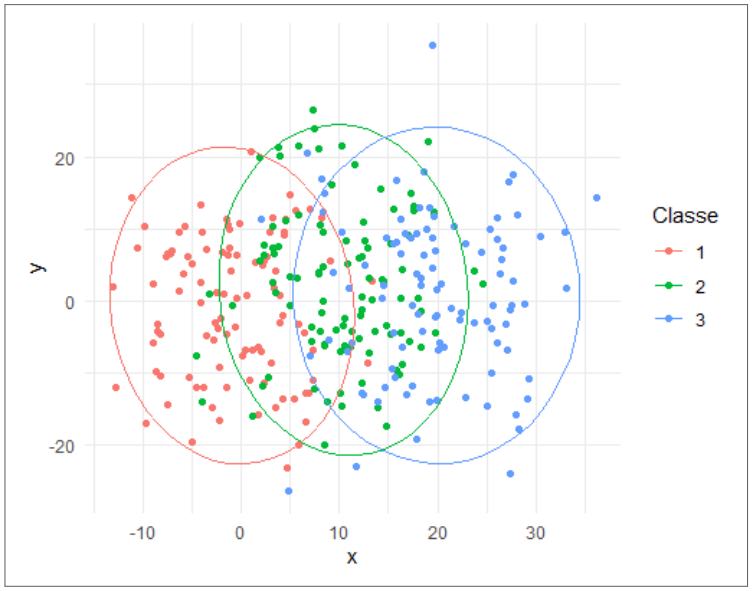
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \qquad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 36 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 10 \\ 0 \\ 7 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 36 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_3 = \begin{pmatrix} 20 \\ 0 \\ 12 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 36 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Através da configuração dos parâmetros acima, o terceiro conjunto de dados sintéticos traz uma configuração onde as partições são formadas por classes elípticas de tamanhos iguais. Abaixo, pode ser visto uma ilustração com apenas duas dimensões para uma maior facilidade de visualização do leitor (Figura 3).

Figura 3 – Conjunto de dados 3



Fonte: O autor (2022)

4.2.4 Configuração do modelo 4

O conjunto de dados 4 também consiste em 300 pontos distribuídos em três classes de tamanhos: 150 observações (classe 1), 100 observações (classe 2) e 50 observações (classe 3). Os pontos de cada classe foram gerados de acordo com os seguintes parãmetros:

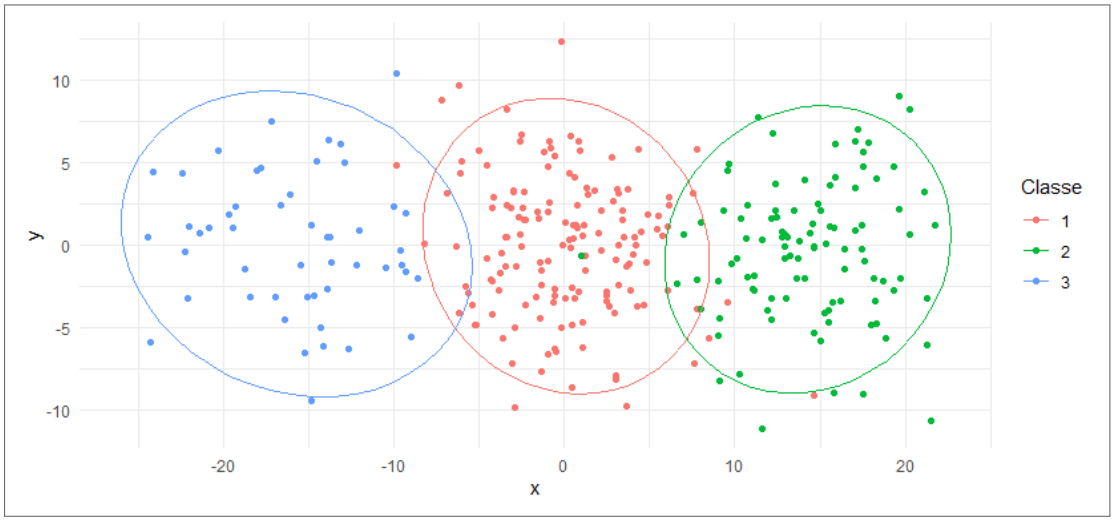
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix} \qquad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 15 \\ 0 \\ 15 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}$$

$$\boldsymbol{\mu}_3 = \begin{pmatrix} -15 \\ 0 \\ -15 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}$$

Através da configuração dos parâmetros acima, o quarto conjunto de dados sintéticos traz uma configuração onde as partições são formadas por classes esféricas de tamanhos iguais. Abaixo, pode ser visto uma ilustração com apenas duas dimensões para uma maior facilidade de visualização do leitor (Figura 4).

Figura 4 – Conjunto de dados 4



Fonte: O autor (2022)

A grande diferença entre uma configuração para outra é o formato e o volume das classes. As variáveis não são correlacionadas, exatamente pelo fato de que os métodos multivariados e os da literatura adotados nesta dissertação supõem a priori que as variáveis são independentes.

De uma forma geral, os conjuntos de dados podem ser compreendidos da seguinte forma:

- Conjunto de dados 1: classes elípticas de diferentes tamanhos;
- Conjunto de dados 2: classes esféricas de diferentes tamanhos;
- Conjunto de dados 3: classes elípticas de tamanhos iguais;
- Conjunto de dados 4: classes esféricas de tamanhos iguais;

A avaliação dos resultados de agrupamento fornecidos pelos métodos é baseada nos cálculos das duas métricas apresentadas na subseção 4.1. Em nossos experimentos, usamos o parâmetro de fuzzyficação m igual a 2, além de 5 réplicas de Monte Carlo, em que para cada réplica, um método de agrupamento foi aleatoriamente executado 50 vezes. O melhor resultado de acordo com a função objetivo foi selecionado e as métricas Índice de Rand Ajustado e F-score foram calculados comparando a partição conhecida à priori com a obtida por cada método de agrupamento.

Os conjuntos de dados sintéticos foram utilizados sem ruído, ou seja, as configurações dos conjuntos de dados não foram alteradas e também foram acrescentados ruídos dispostos de acordo com uma distribuição uniforme univariada, no intervalo $[-100, 50]$, em que para cada variável é gerado e adicionado um ruído com a distribuição citada anteriormente. Neste material, foram elaborados experimentos com os seguintes percentuais de dados ruidosos: 0%, 10% e 20%. As análises desses experimentos serão expostas na seção 4.4.

4.3 CONJUNTOS DE DADOS REAIS

Nesta seção, conjuntos de dados do repositório de aprendizado de máquina da UCI ((BACHE; LICHMAN, 2013) foram utilizados para comparar os métodos em estudo. Abaixo na Tabela 2, podemos visualizar as bases de dados utilizadas com seus respectivos números de grupos, total de observações, além da quantidade de observações por grupo.

4.3.1 Análise Descritiva dos dados e Análise de Agrupamento

4.3.1.1 Haberman

Esse conjunto de dados contém informações a cerca de um estudo realizado entre 1958 e 1970 no Hospital Billings da Universidade de Chicago sobre a sobrevivência ou não de pacientes

Tabela 2 – Conjuntos de dados reais da UCI

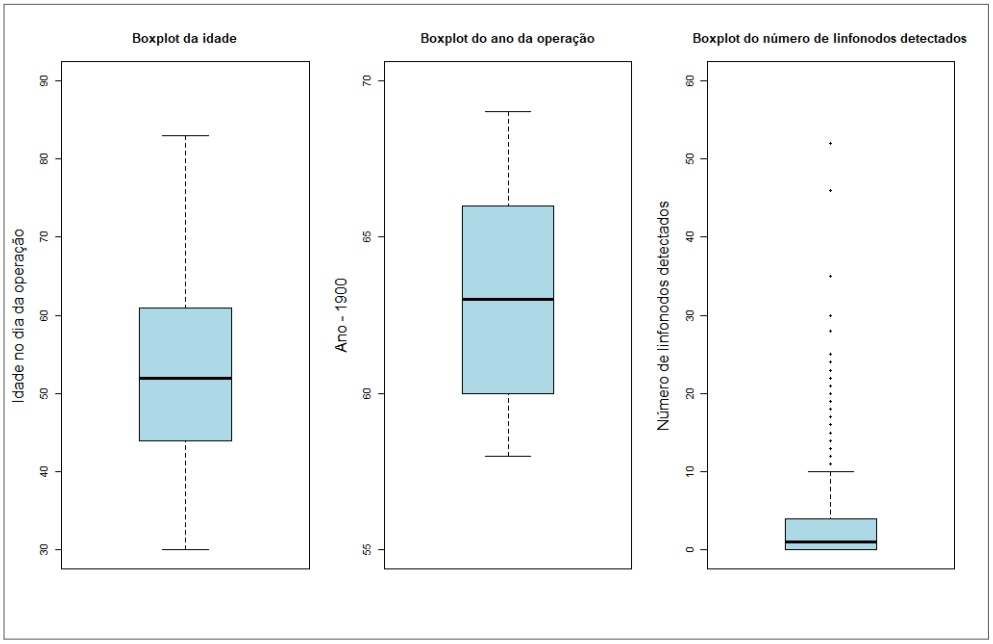
Dataset	Grupos	Observações	Grupo 1	Grupo 2	Grupo 3
Haberman	2	306	225	81	-
Abalone	3	4177	1528	1307	1342
Seeds	3	210	70	70	70
Iris	3	150	50	50	50
Wine	3	178	59	71	48

Fonte: O autor (2022)

que foram submetidas a cirurgias de câncer de mama. Esse conjunto de dados contém 306 observações, 3 variáveis independentes e a variável dependente possui 2 classes em estudo, sendo a primeira se o paciente sobreviveu 5 anos ou mais após a cirurgia de câncer de mama (225 observações) e a segunda classe faz referência às pessoas que morreram dentro de 5 anos pós cirurgia (81 observações).

Abaixo, podemos verificar como as 3 variáveis da base de dados se comportam através do gráfico de *Boxplot*. A Primeira exprime informações da idade em que o paciente se encontrava no momento da cirurgia, a segunda imagem exprime o valor do (ano da operação - 1900), ou seja, quanto maior seu valor, mais recente a cirurgia foi executada e por último o quantitativo de linfonodos axilares positivos detectados. Analisando os *boxplots*, verifica-se que apenas a variável Linfonodos apresenta *outliers*, ou valores discrepantes.

Figura 5 – Boxplots do Haberman



Fonte: O autor (2022)

Outra análise descritiva bastante interessante a ser feita é o estudo sobre as variâncias das variáveis por grupos de acordo com as informações cedidas pelo conjunto de dados. Podemos notar que há uma diferença considerável entre as variâncias dos dois grupos analisados para a variável Linfonodos, como pode ser visto na tabela abaixo, exprimindo a ideia que as observações para o grupo 2 estão mais espalhadas, ou em outras palavras, mais distantes da média, perante as observações do grupo 1 para esta variável. As demais variáveis em estudo aparentemente não possuem uma grande diferença entre as variâncias dos dois grupos.

Tabela 3 – Variâncias das variáveis por grupo - Haberman

Grupos	**Idade**	**Ano de operação - 1900**	**Linfonodos**
Grupo 1	121.27	10.39	34.46
Grupo 2	103.37	11.17	84.38

Fonte: O autor (2022)

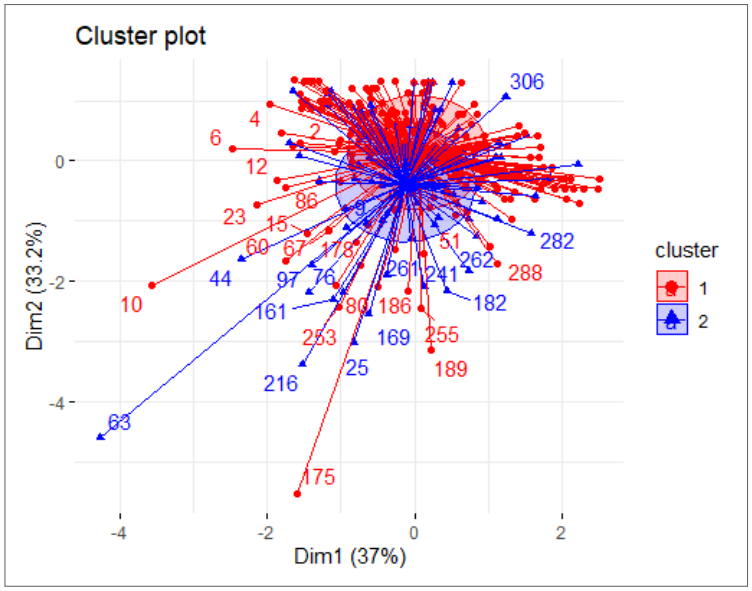
Tabela 4 – Médias dos índices ARI e F-score - Haberman

Métricas	**FCM**	**FCMdd**	**MFCM**	**MFCMdd**
ARI	0.0002	0.0286	**0.037**	0.013
F-score	0.5479	0.5613	**0.6071**	0.5545

Fonte: O autor (2022)

Diante da Tabela 4, observa-se que o método MFCM obteve uma melhor performance perante os demais metodos analisando as médias das duas métricas previamente explanadas. Abaixo, será demonstrado um gráfico de Análise de Componentes Principais (PCA) dos dados originais. Essa é uma abordagem estatística cujo um dos seus intuitos é condensar um vasto número de variáveis em um conjunto reduzido de variáveis estatísticas conseguindo analisar as inter-relações entre elas perdendo o mínimo de informações (ARAÚJO, 2010). Além disso, o mesmo tipo de gráfico será gerado utilizando os quatro métodos em estudo.

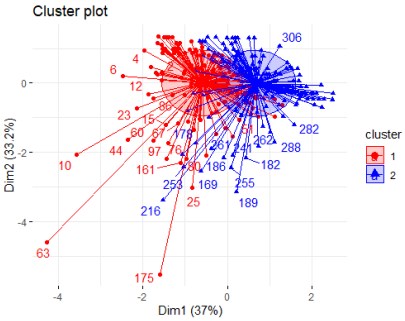
Figura 6 – PCA dos Dados Originais - Haberman



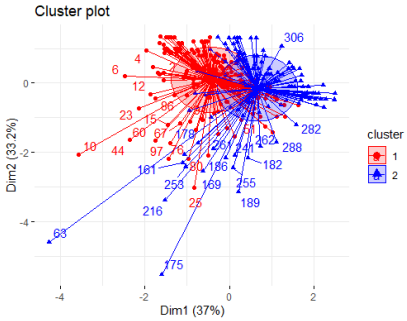
Fonte: O autor (2022)

Figura 7 – Gráfico de PCA dos métodos de agrupamento para o conjunto Haberman

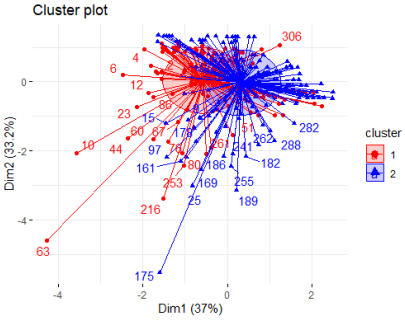
(a) PCA do FCM



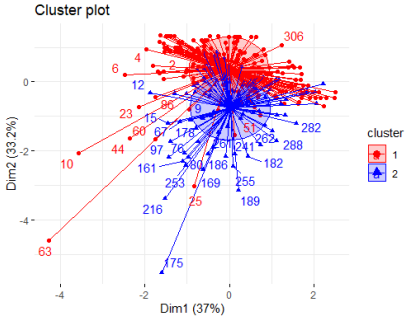
(b) PCA do FCMdd



(c) PCA do MFCM



(d) PCA do MFCMdd



Fonte: O autor (2022)

Com o auxílio da Análise Descritiva dos dados conjuntamente com o gráfico de Análise dos Componentes Principais (PCA) dos dados originais, consegue-se extrair em qual dos quatro

cenários dos conjuntos de dados apresentados na seção 4.2 o Haberman mais se assemelha.

De acordo com a Tabela 3, verifica-se que a variância intra grupos da maioria das variáveis não é similar, ou seja, as variâncias das variáveis dentro de um mesmo grupo possuem diferenças significativas. Além disso, em relação ao contexto inter grupos, percebe-se que as variâncias das variáveis são próximas, exceto para a variável linfonodo. No que se refere aos *boxplots* (Figura 5), das três variáveis do conjunto de dados, apenas a Linfonodo possui *outliers*, contudo, sem ser em grande quantidade. Portanto, devido à configuração encontrada na Tabela 3, conclui-se que devido ao fato das variâncias intra grupos não serem similares, o conjunto de dados Haberman possui características elípticas. Também pode-se afirmar que como as variâncias inter grupos apresentam variâncias similares perante os dois grupos para a maioria das variáveis, o conjunto de dados pode ser considerado de tamanhos iguais e por último, pelo gráfico de *boxplot* percebe-se que há um pouco de ruído exclusivamente na variável linfonodo.

Conclusão 1: O conjunto de dados Haberman possui características de um elipsóide de tamanhos iguais com ruído, comprovando o que o gráfico de PCA para os dados originais sinalizava.

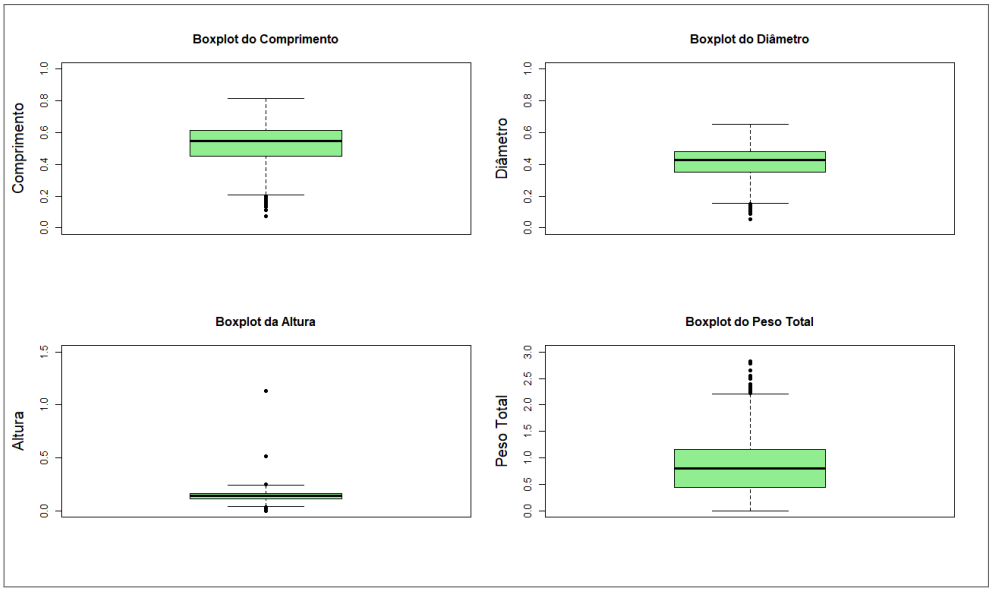
Conclusão 2: Para o conjunto de dados Haberman, o método MFCM obteve vantagem perante as métricas ARI e *F-score* (Tabela 4), isso devido ao fato desse método possuir vantagem quando há um cenário elíptico com tamanhos iguais com ruído, correspondendo ao Conjunto de Dados 3.

4.3.1.2 Abalone

Esse conjunto de dados visa prever a idade do abalone a partir de medições físicas. A idade do abalone é determinada cortando a casca através do cone, manchando-a e contando o número de anéis através de um microscópio. Esse conjunto de dados contém 4177 obervações, 8 variáveis independentes e a variável dependente possui 3 classes em estudo, sendo Adulto do sexo masculino (1528 observações), adulto do sexo feminino (1307 observações) e infantil (1342 observações).

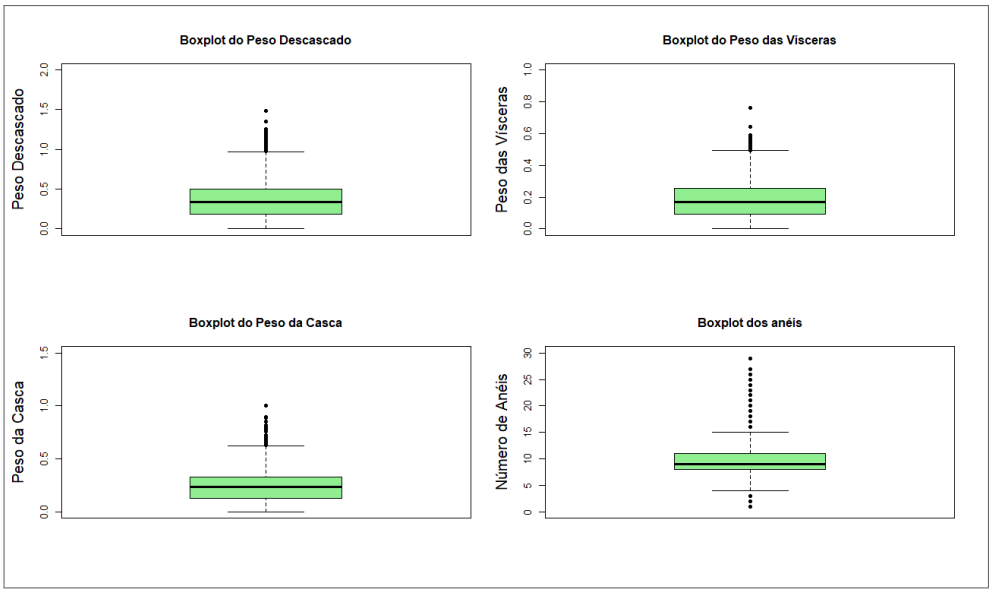
Abaixo, podemos verificar como as 8 variáveis da base de dados se comportam através dos gráficos de *Boxplot*. As variáveis desse conjunto de dados fazem referência ao comprimento, diâmetro, altura, peso total, peso descascado, peso das vísceras, peso da casca e número de anéis. Percebe-se que em todas essas variáveis independentes, há presença de *outliers*.

Figura 8 – Boxplots do Abalone - Parte 1



Fonte: O autor (2022)

Figura 9 – Boxplots do Abalone - Parte 2



Fonte: O autor (2022)

Analisando as variâncias das variáveis por grupos de acordo com as informações cedidas pelo conjunto de dados, as variáveis "peso total"e "número de anéis"possuem variâncias consideráveis diferentes do grupo 3 perante os demais grupos em estudo.

Em relação aos agrupamentos, serão apresentados os resultados das métricas em estudo, perante os métodos em análise FCM, FCMdd, MFCM e MFCMdd para o conjunto de dados Abalone.

Tabela 5 – Variância das variáveis por grupo - Abalone

Grupos	Compr.	Diâm.	Altura	Peso Tot.	Peso Desc.	Peso das Vísc.	Peso Casca	Anéis
Grupo 1	0.01	0.01	0.00	0.22	0.05	0.01	0.02	9.16
Grupo 2	0.01	0.01	0.00	0.19	0.04	0.01	0.02	9.64
Grupo 3	0.01	0.01	0.00	0.08	0.02	0.00	0.01	6.31

Fonte: O autor (2022)

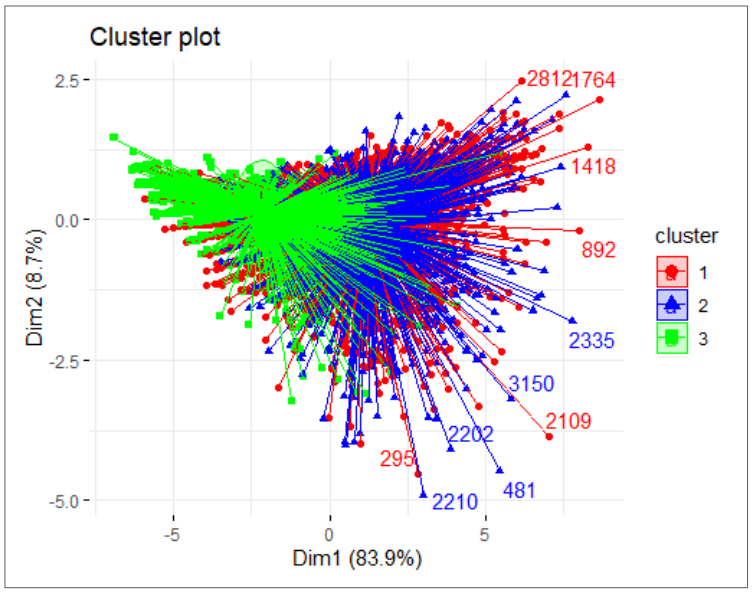
Tabela 6 – Médias dos índices ARI e F-score - Abalone

Métricas	FCM	FCMdd	MFCM	MFCMdd
ARI	0.1331	0.0541	0.1324	**0.1501**
F-score	0.4578	0.4384	0.4224	**0.4623**

Fonte: O autor (2022)

Diante da Tabela 6, observa-se que o método MFCMdd obteve uma melhor performance perante os demais metodos analisando as médias das duas métricas previamente explanadas. Abaixo, será demonstrado um gráfico de Análise de Componentes Principais (PCA) dos dados originais. Além disso, o mesmo tipo de gráfico será gerado utilizando os quatro métodos em estudo.

Figura 10 – PCA dos Dados Originais - Abalone

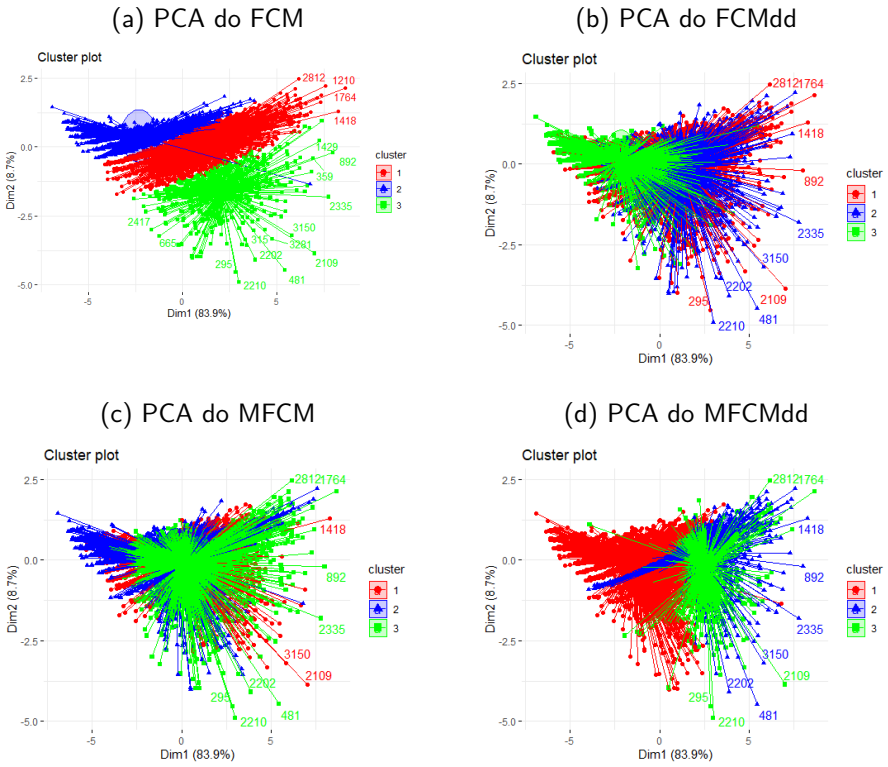


Fonte: O autor (2022)

Através da Análise Descritiva dos dados conjuntamente com o gráfico de Análise dos Componentes Principais (PCA) dos dados originais, consegue-se extrair em qual dos quatro cenários dos conjuntos de dados apresentados na seção 4.2 o Abalone mais se assemelha.

De acordo com a Tabela 5, verifica-se que a variância intra grupos da maioria das variáveis é similar, ou seja, apenas a variável Anéis possui uma diferença significativa perante as outras

Figura 11 – Gráfico de PCA dos métodos de agrupamento para o conjunto Abalone



Fonte: O autor (2022)

variâncias das variáveis do mesmo grupo. Além disso, em relação ao contexto inter grupos, percebe-se a maioria das variâncias das variáveis são similares, destoando apenas em duas variáveis, sendo elas, Peso Total e Anéis no grupo 3, perante os demais grupos. No que se refere aos *boxplots* (Figuras 8-9), todas as variáveis do conjunto de dados possuem *outliers*. Portanto, devido à configuração encontrada na Tabela 5, conclui-se que devido ao fato das variâncias intra grupos serem similares, o conjunto de dados Abalone possui características esféricas. Também pode-se afirmar que como as variâncias inter grupos apresentam variâncias similares perante os três grupos para a maioria das variáveis, o conjunto de dados pode ser considerado de tamanhos iguais e por último, pelo gráfico de boxplot percebe-se que há ruído em todas as variáveis.

Conclusão 1: O abalone possui características de uma esfera de tamanhos iguais com ruído, apesar do gráfico de PCA para os dados originais não demonstrar claramente isso por causa do grande quantitativo de observações deste conjunto de dados (4177 observações).

Conclusão 2: Para o conjunto de dados Abalone, o método MFCMed obteve

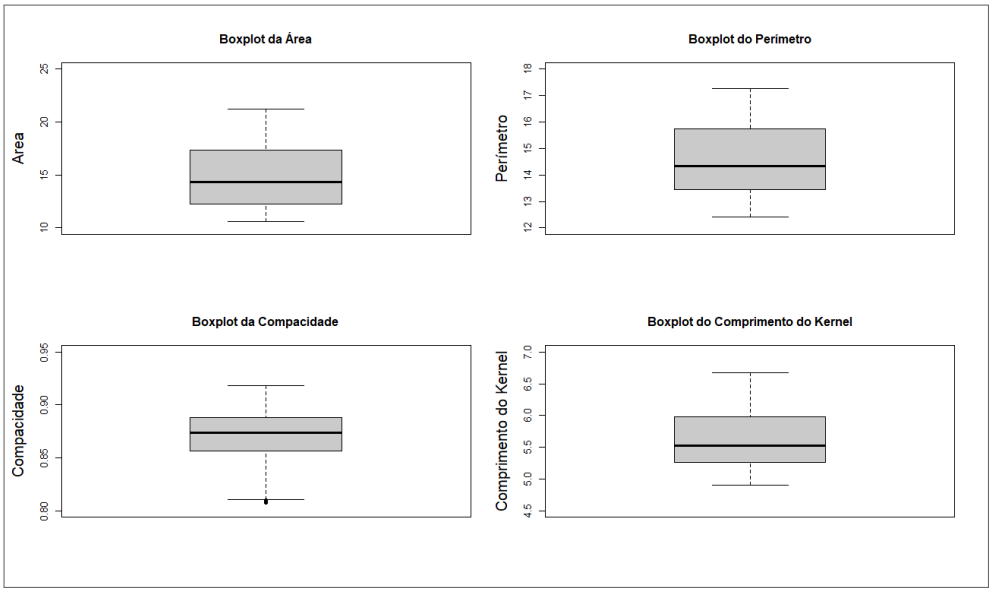
vantagem perante a média das métricas ARI e *F-score* (Tabela 6).

4.3.1.3 Seeds

Esse conjunto de dados lida com grãos de trigo de espécies variadas colhidos em colheitas combinadas provenientes de campos experimentais, explorados no Instituto de Agrofísica da Academia Polonesa de Ciências em Lublin. Esse conjunto de dados contém 210 obervações, 7 variáveis independentes e a variável dependente possui 3 classes em estudo, sendo os tipos de trigo Kama (70 observações), Rosa (70 observações) e Canadian (70 observações).

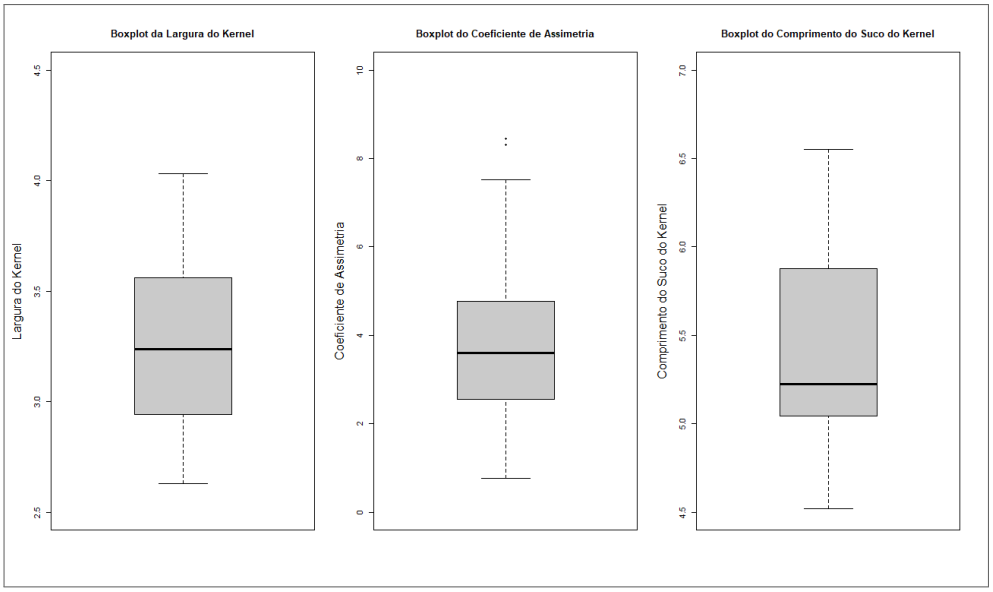
Abaixo, podemos verificar como as 7 variáveis da base de dados se comportam através dos gráficos de *Boxplot*. As variáveis desse conjunto de dados fazem referência à Área do trigo, perímetro, compacidade, comprimento do kernel, largura do kernel, coeficiente de assimetria e comprimento do suco do kernel. Percebe-se que as variáveis Compacidade e Coeficiente de Assimetria possuem outliers.

Figura 12 – Boxplots do Seeds - Parte 1



Fonte: O autor (2022)

Figura 13 – Boxplots do Seeds - Parte 2



Fonte: O autor (2022)

Analisando as variâncias das variáveis por grupos de acordo com as informações cedidas pelo conjunto de dados, as variáveis "Área", "Perímetro"e "Coef. de Assimetria"possuem variâncias consideráveis diferentes do grupo 3 perante os demais grupos em estudo.

Tabela 7 – Variância das variáveis por grupo - Seeds

Grupos	Área	Perímetro	Compacidade	Comp. do Kernel	Larg. do Kernel	Coef. de Assimetria	Comp. do Suco do Kernel
Grupo 1	1.48	0.33	0.00	0.05	0.03	1.38	0.07
Grupo 2	2.07	0.38	0.00	0.07	0.03	1.4	0.06
Grupo 3	0.52	0.12	0.00	0.02	0.02	1.79	0.03

Fonte: O autor (2022)

Em relação aos agrupamentos, abaixo serão apresentados os resultados das métricas em estudo, perante os métodos em análise FCM, FCMdd, MFCM e MFCMdd para o conjunto de dados Seeds.

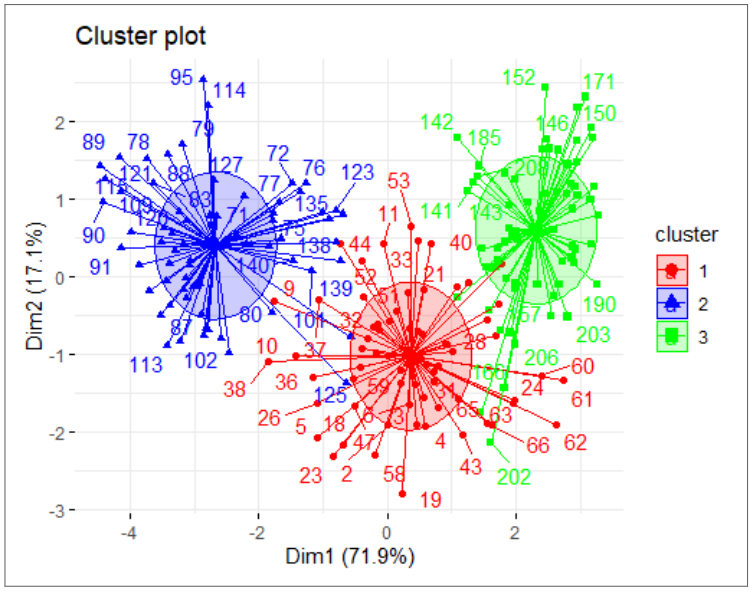
Tabela 8 – Médias dos índices ARI e F-score - Seeds

Métricas	FCM	FCMdd	MFCM	MFCMdd
ARI	**0.7166**	0.6232	0.1715	0.2388
F-score	**0.8106**	0.7494	0.4544	0.4905

Fonte: O autor (2022)

Diante da Tabela 8, observa-se que o método FCM obteve uma melhor performance perante os demais metodos analisando as médias das duas métricas previamente explanadas. Abaixo, será demonstrado um gráfico de Análise de Componentes Principais (PCA) dos dados originais. Além disso, o mesmo tipo de gráfico será gerado utilizando os quatro métodos em estudo.

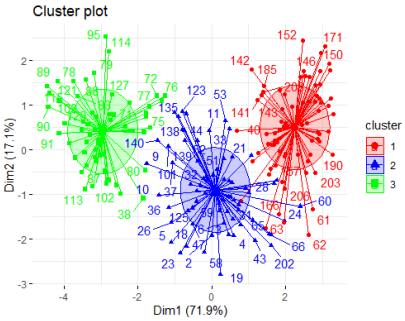
Figura 14 – PCA dos Dados Originais - Seeds



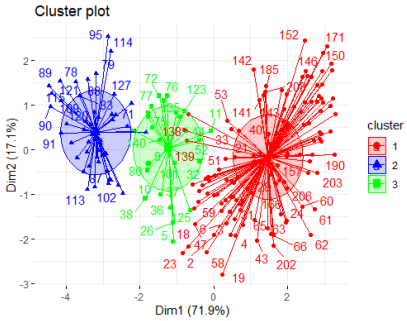
Fonte: O autor (2022)

Figura 15 – Gráfico de PCA dos métodos de agrupamento para o conjunto Seeds

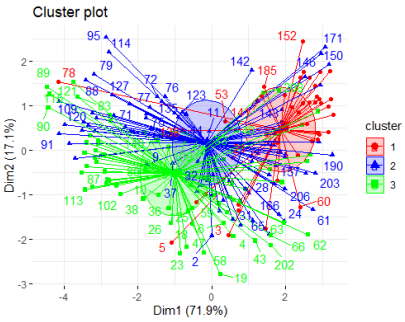
(a) PCA do FCM



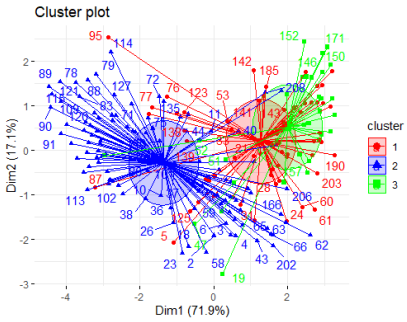
(b) PCA do FCMdd



(c) PCA do MFCM



(d) PCA do MFCMdd



Fonte: O autor (2022)

De acordo com a Tabela 7, verifica-se que a variância intra grupos da maioria das variáveis é similar, ou seja, apenas as variáveis Área e Coeficiente de Assimetria possuem uma dife-

rença significativa nas variâncias. Além disso, em relação ao contexto inter grupos, percebe-se que a maioria das variâncias das variáveis são similares entre os grupos, destoando um pouco nas variáveis Área e Perímetro no grupo 3, perante os demais grupos. No que se refere aos *boxplots* (Figuras 12-13), apenas duas variáveis possuem *outliers* com um quantitativo quase nulo, portanto é uma característica que não impactará significativamente no cálculo dos protótipos. Logo, devido à configuração encontrada na Tabela 7, conclui-se que devido ao fato das variâncias intra grupos serem similares na sua maioria, o conjunto de dados Seeds possui características esféricas. Também pode-se afirmar que como as variâncias inter grupos apresentam variâncias similares perante os três grupos para a maioria das variáveis, o conjunto de dados pode ser considerado de tamanhos iguais e por último, através do gráfico de *boxplot*, percebe-se que não há ruído significativo.

Conclusão 1: O conjunto de dados Seeds possui características de uma esfera de tamanhos iguais sem ruído, comprovando o que o gráfico de PCA para os dados originais sinalizava.

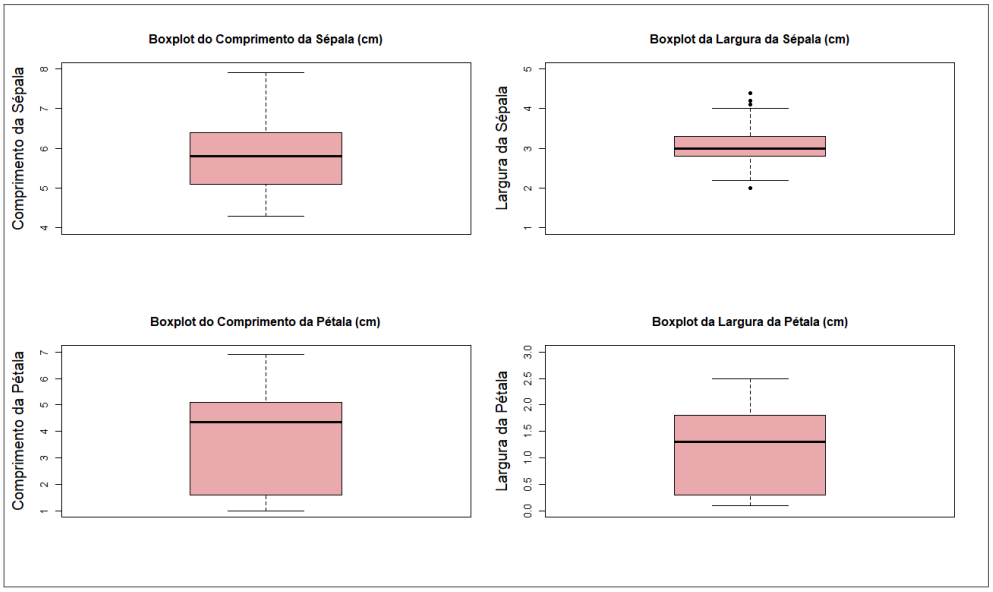
Conclusão 2: Para o conjunto de dados Seeds, o método FCM obteve vantagem perante as métricas ARI e F-score (Tabela 8), devido ao fato desse método possuir vantagem quando há um cenário esférico com tamanhos iguais e sem ruído, correspondendo ao Conjunto de Dados 4.

4.3.1.4 Íris

Uma das bases de dados mais conhecidas da literatura. O conjunto de dados contém 3 classes de 50 observações cada, onde cada classe se refere a um tipo de planta de íris, sendo elas setosa, virgínica e versicolour. Como comentado, esse conjunto de dados contém 150 obervações, 4 variáveis independentes e a variável dependente.

Abaixo, podemos verificar como as 4 variáveis da base de dados se comportam através dos gráficos de Boxplot. As variáveis desse conjunto de dados fazem referência à comprimento da sépala em cm, largura da sépala em cm, comprimento da pétala em cm e largura da pétala em cm. Percebe-se que apenas na variável Largura da Sépala há presença de outliers.

Figura 16 – Boxplots do Iris



Fonte: O autor (2022)

Analisando as variâncias das variáveis por grupos de acordo com as informações cedidas pelo conjunto de dados, percebe-se que o Grupo 1 possui variâncias menores que nos demais grupos em estudo, exceto para a variável Largura da Sépala.

Tabela 9 – Variância das variáveis por grupo - Íris

Grupos	Comp. da Sépala	Larg. da Sépala	Comp. da Pétala	Larg. da Pétala
Grupo 1	0.1242	0.1452	0.0301	0.0115
Grupo 2	0.2664	0.0985	0.2208	0.0391
Grupo 3	0.4043	0.1040	0.3046	0.0754

Fonte: O autor (2022)

No que se refere aos agrupamentos, na Tabela 10 serão apresentados os resultados das métricas em estudo, perante os métodos em análise FCM, FCMdd, MFCM e MFCMdd para o conjunto de dados Seeds.

Tabela 10 – Médias dos índices ARI e F-score - Íris

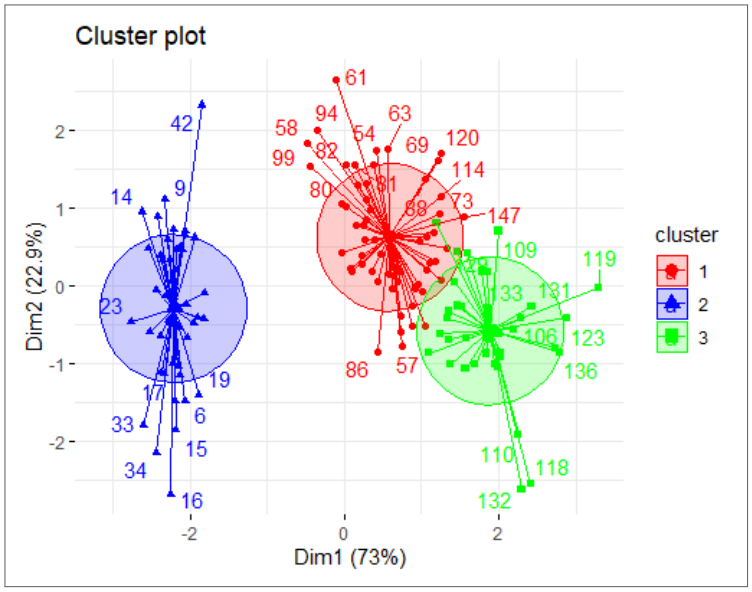
Métricas	FCM	FCMdd	MFCM	MFCMdd
ARI	**0.7294**	0.3589	0.1786	0.2303
F-score	**0.8196**	0.5901	0.4715	0.4896

Fonte: O autor (2022)

Diante da Tabela 10, observa-se que para este conjunto de dados, o método FCM obteve novamente uma melhor performance perante os demais metodos analisando as médias das duas métricas previamente explanadas. Abaixo, será demonstrado um gráfico de Análise de

Componentes Principais (PCA) dos dados originais. Além disso, o mesmo tipo de gráfico será gerado utilizando os quatro métodos em estudo.

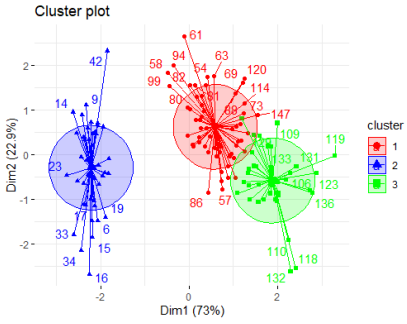
Figura 17 – PCA dos Dados Originais - Íris



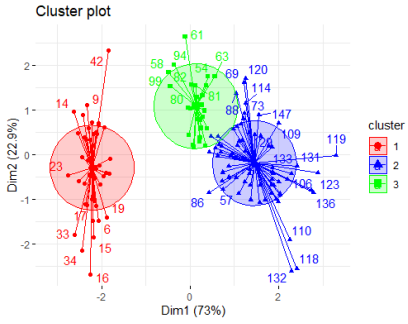
Fonte: O autor (2022)

Figura 18 – Gráfico de PCA dos métodos de agrupamento para o conjunto Íris

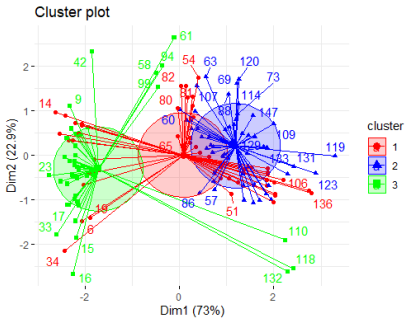
(a) PCA do FCM



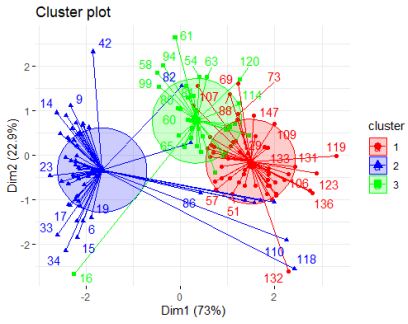
(b) PCA do FCMdd



(c) PCA do MFCM



(d) PCA do MFCMdd



Fonte: O autor (2022)

De acordo com a Tabela 9, verifica-se que a variância intra grupos da maioria das variáveis é similar. Além disso, em relação ao contexto inter grupos, percebe-se que a maioria das variâncias das variáveis são similares, destoando um pouco nas variáveis Comprimento da Sépala e Comprimento da Pétala. No que se refere aos *boxplots* (Figura 16), apenas duas variáveis possuem outliers com um quantitativo quase nulo, portanto é uma característica que não impactará significativamente no cálculo dos protótipos. Logo, devido à configuração encontrada na Tabela 9, conclui-se que devido ao fato das variâncias intra grupos serem similares na sua maioria, o conjunto de dados Íris possui características esféricas. Também pode-se afirmar que como as variâncias inter grupos apresentam variâncias similares perante os três grupos para a maioria das variáveis, o conjunto de dados pode ser considerado de tamanhos iguais e por último, através do gráfico de *boxplot*, percebe-se que não há ruído significativo.

Conclusão 1: O conjunto de dados Íris também possui características de uma esfera de tamanhos iguais sem ruído, comprovando o que o gráfico de PCA para os dados originais sinalizava.

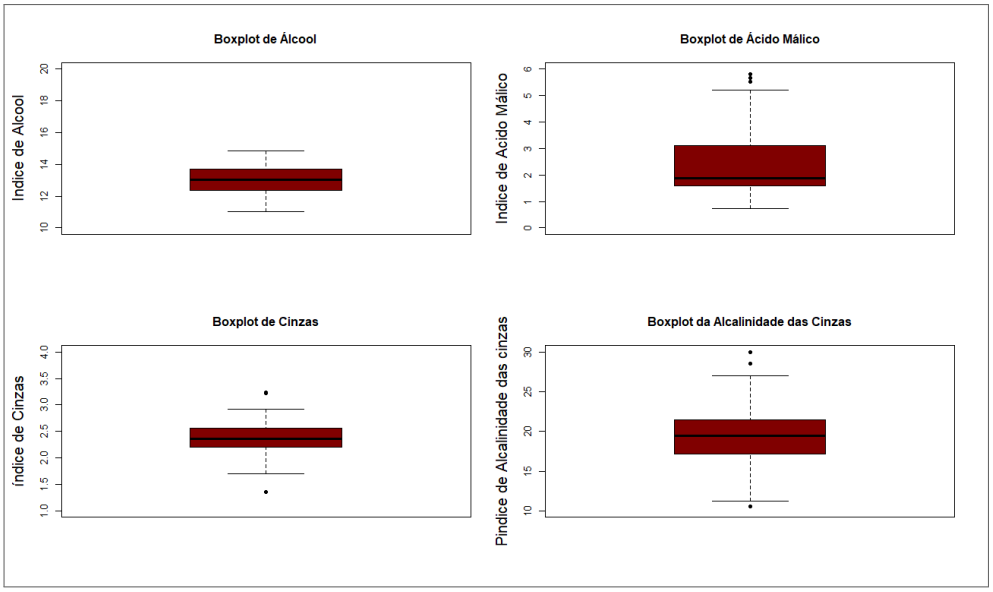
Conclusão 2: Para o conjunto de dados Íris, o método FCM obteve vantagem perante as métricas ARI e *F-score* (Tabela 10), devido ao fato desse método possuir vantagem quando há um cenário esférico com tamanhos iguais e sem ruído, correspondendo ao Conjunto de Dados 4.

4.3.1.5 Wine

Este conjunto de dados é origindo de um estudo sobre análise química de vinhos cultivados na mesma região da Itália, porém de 3 cultivares diferentes. Esse conjunto de dados contém 178 obervações, 13 variáveis independentes e a variável dependente que possui três classes de cultivares diferentes.

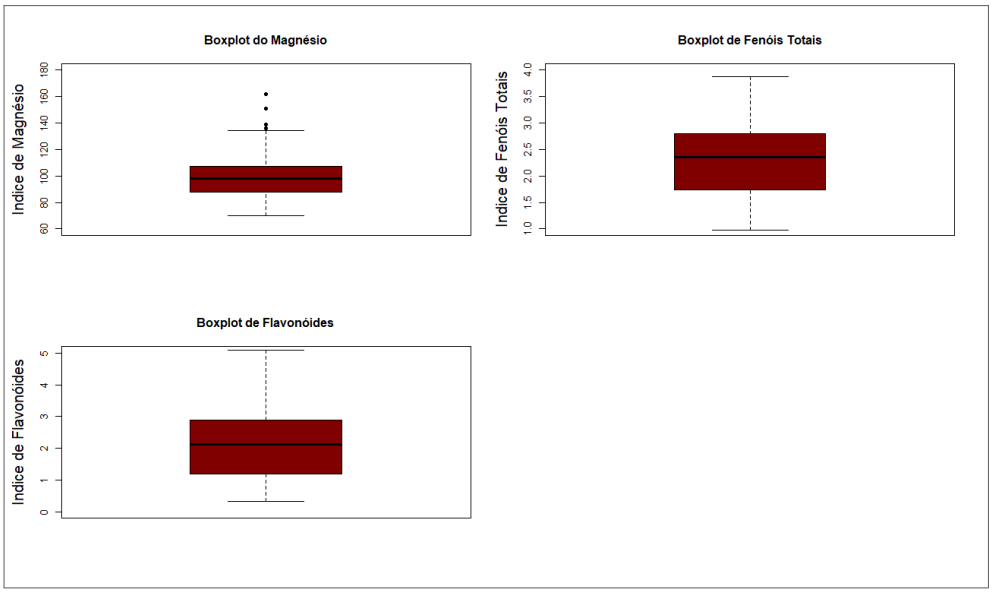
Abaixo, podemos verificar como as 13 variáveis da base de dados se comportam através dos gráficos de *Boxplot*. As variáveis desse conjunto de dados fazem referência ao álcool, ácido málico, cinzas, alcalinidade das cinzas, magnésio, fenóis totais, flavonóides, fenóis não flavonóides, proantocianinas, intensidade da cor, matiz, OD280/OD315 de vinhos diluídos e Prolina. Percebe-se que há *outliers* em 7 variáveis das 13 em estudo.

Figura 19 – Boxplots do Wine (Parte 1)



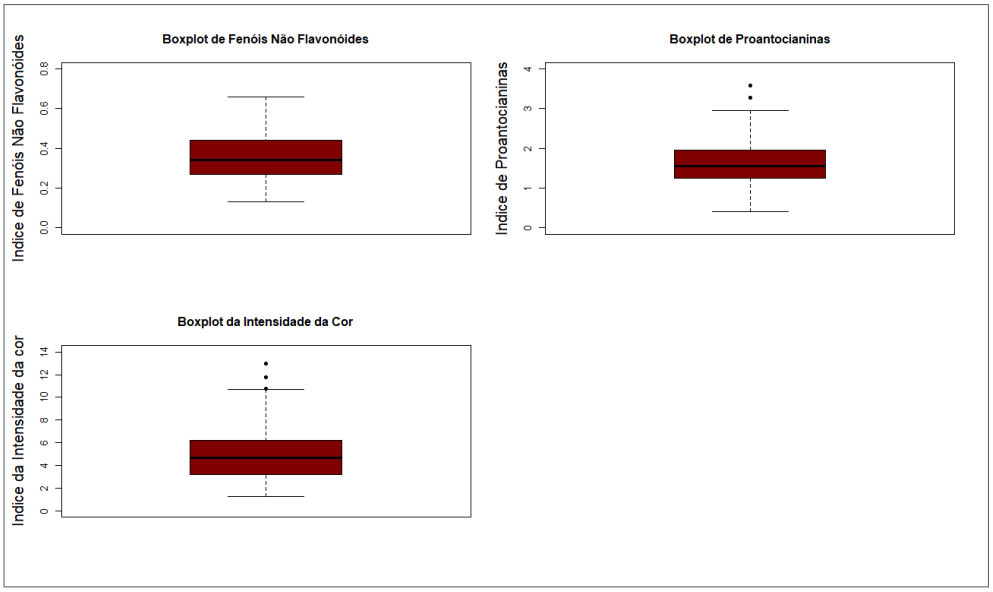
Fonte: O autor (2022)

Figura 20 – Boxplots do Wine (Parte 2)



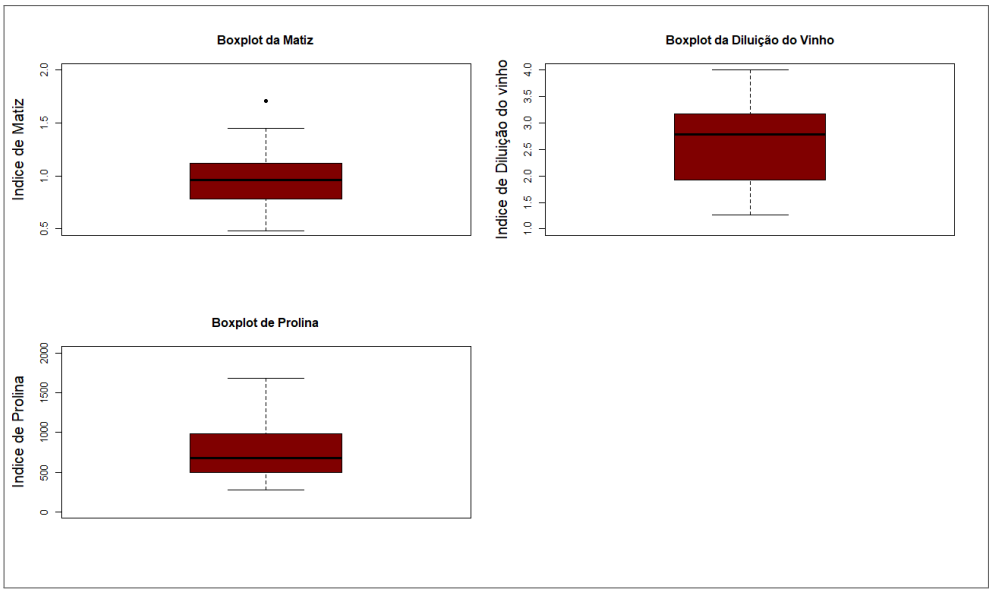
Fonte: O autor (2022)

Figura 21 – Boxplots do Wine (Parte 3)



Fonte: O autor (2022)

Figura 22 – Boxplots do Wine (Parte 4)



Fonte: O autor (2022)

Analisando as variâncias das variáveis por grupos de acordo com as informações cedidas pelo conjunto de dados, percebe-se que o Grupo 2 obteve variância diferente dos demais grupos em algumas variáveis, como por exemplo em "Magnésio"e "Prolina".

Com relação aos agrupamentos, abaixo serão apresentados os resultados das métricas em estudo, perante os métodos em análise FCM, FCMdd, MFCM e MFCMdd para o conjunto de dados Wine.

Tabela 11 – Variância das variáveis por grupo - Wine Parte 1

Grupos	Álcool	Ácido Málico	Cinzas	Alc. Cinzas	Magnésio	Fenóis tot.	Flavonóides
Grupo 1	0.21	0.47	0.05	6.48	110.23	0.11	0.16
Grupo 2	0.29	1.03	0.10	11.22	280.68	0.30	0.50
Grupo 3	0.28	1.18	0.03	5.10	118.60	0.13	0.09

Fonte: O autor (2022)

Tabela 12 – Variância das variáveis por grupo - Wine Parte 2

Grupos	Fenóis não Flavonóides	Proantoc.	Intens. da cor	Matiz	Diluição	Prolina
Grupo 1	0.00	0.17	1.53	0.01	0.13	49071.45
Grupo 2	0.02	0.00	0.86	0.04	0.25	24715.37
Grupo 3	0.02	0.17	5.34	0.01	0.07	13247.33

Fonte: O autor (2022)

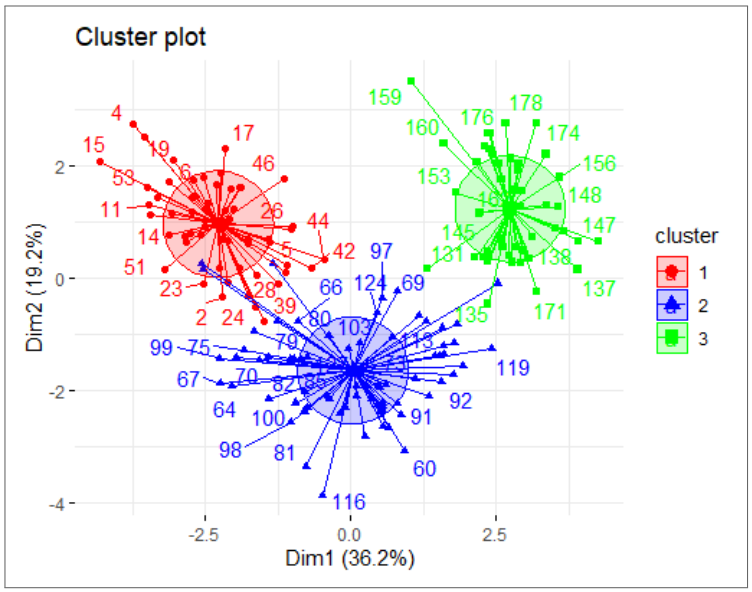
Tabela 13 – Médias dos índices ARI e F-score - Wine

Métricas	FCM	FCMdd	MFCM	MFCMdd
ARI	0.3539	**0.4011**	0.0439	0.1462
F-score	0.5728	**0.6065**	0.3667	0.4339

Fonte: O autor (2022)

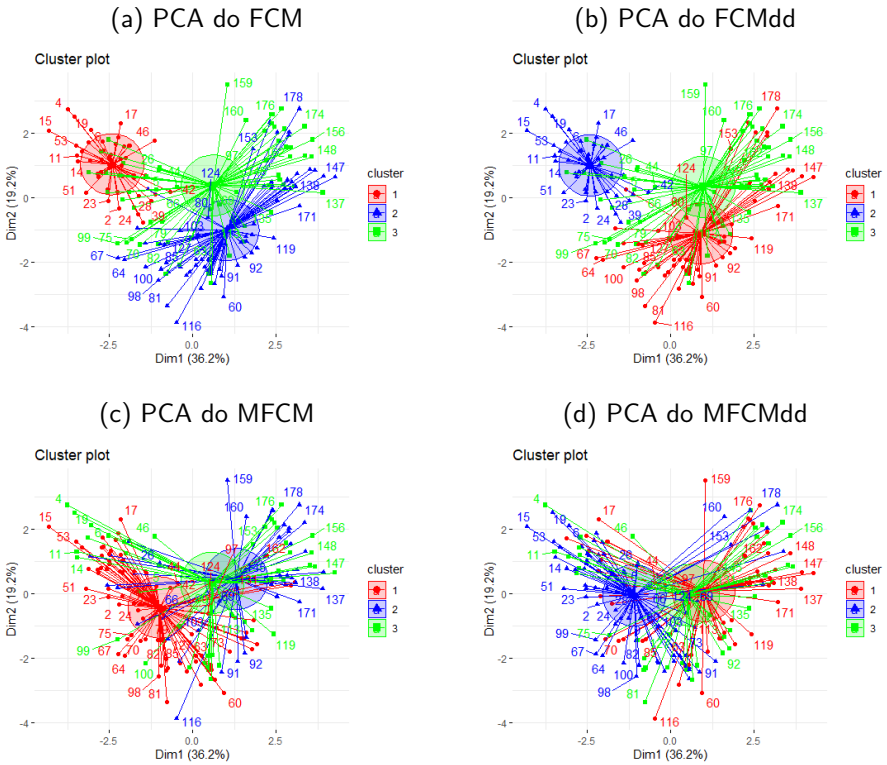
Diante da Tabela 13, observa-se que para este conjunto de dados, o método FCMdd obteve uma melhor performance perante os demais metodos analisando as médias das duas métricas previamente explanadas. Abaixo, será demonstrado um gráfico de Análise de Componentes Principais (PCA) dos dados originais. Além disso, o mesmo tipo de gráfico será gerado utilizando os quatro métodos em estudo.

Figura 23 – PCA dos Dados Originais - Wine



Fonte: O autor (2022)

Figura 24 – Gráfico de PCA dos métodos de agrupamento para o conjunto Wine



Fonte: O autor (2022)

De acordo com as Tabelas 11 e 12, verifica-se que a variância intra grupos da maioria das variáveis é similar, destoando um pouco mais nas variáveis Mangnésio e Prolina. Além disso, em relação ao contexto inter grupos, percebe-se que a maioria das variâncias das variáveis são similares, destoando um pouco nas mesmas variáveis citadas acima, além da alcalinidade das cinzas e intensidade da cor, perante os grupos. No que se refere aos *boxplots* (Figuras 19-22), metade das variáveis possuem outliers com um pouco de ruído. Logo, devido à configuração encontrada nas Tabelas 11 e 12, conclui-se que devido ao fato das variâncias intra grupos serem similares na sua maioria, o conjunto de dados Wine possui características esféricas. Também pode-se afirmar que como as variâncias inter grupos apresentam variâncias similares perante os três grupos para a maioria das variáveis, o conjunto de dados pode ser considerado de tamanhos iguais e por último, através do gráfico de *boxplot*, percebe-se que há pouco ruído significativo.

Conclusão 1: O conjunto de dados Wine também possui características de uma esfera de tamanhos iguais, porém, nesse caso com um pouco de ruído, comprovando o que o gráfico de PCA para os dados originais sinalizava.

Conclusão 2: Para o conjunto de dados Wine, o método FCMdd obteve vantagem perante as métricas ARI e *F-score* (Tabela 13), devido ao fato desse método possuir vantagem quando há um cenário esférico com tamanhos iguais e com pouco ruído, correspondendo ao Conjunto de Dados 4.

4.4 DISCUSSÃO

Como resposta aos objetivos expostos em 1.2, podemos verificar que:

4.4.1 Conjuntos de Dados Simulados

A Tabela 14 demonstra os valores das médias do Índice de Rand Ajustado (ARI) e do F-score pelos métodos FCM, FCMdd, MFCM e MFCMdd para os conjuntos de dados 1-4. Os maiores valores de média dos índices estão em negrito para cada conjunto de dados e para cada variação de ruído. A discussão sobre esses resultados podem ser vistas adiante.

Tabela 14 – Médias do ARI e F-score para os métodos de agrupamento e conjunto de dados 1-4

Conjunto	Métrica	0%				10%				20%			
		FCM	FCMdd	MFCM	MFCMdd	FCM	FCMdd	MFCM	MFCMdd	FCM	FCMdd	MFCM	MFCMdd
1	ARI	**0.3152**	0.2994	0.0297	0.1011	**0.3927**	0.2746	0.1133	0.0887	**0.3309**	0.2213	0.1753	0.1536
	F-score	**0.5759**	0.5756	0.3971	0.4403	**0.6724**	0.572	0.5063	0.4377	**0.6411**	0.5553	0.53	0.4856
2	ARI	**0.6889**	0.5434	0.0766	0.1182	**0.4871**	0.4485	0.0907	0.1779	**0.6144**	0.4065	0.0921	0.2125
	F-score	**0.8091**	0.719	0.4241	0.455	**0.6851**	0.6582	0.4417	0.488	**0.799**	0.6548	0.453	0.5117
3	ARI	0.2963	0.2636	0.4998	**0.5606**	0.3360	0.2345	**0.7344**	0.4889	0.3901	0.2147	**0.6758**	0.5655
	F-score	0.5308	0.5114	0.666	**0.7087**	0.5955	0.4976	**0.823**	0.6619	0.6346	0.5037	**0.7899**	0.7101
4	ARI	**0.9481**	0.9045	0.1069	0.1421	0.584	**0.9427**	0.0308	0.2027	0.3576	**0.689**	0.0213	0.163
	F-score	**0.9682**	0.9419	0.4288	0.4533	0.7869	**0.9649**	0.4012	0.4927	0.676	**0.8378**	0.3871	0.4664

Fonte: O autor (2022)

Diante dos resultados apresentados na Tabela 14, referentes aos dados simulados, o método FCM obteve vantagem em todos os cenários (sem ruído e com ruído) para os conjuntos de dados 1 e 2, sendo eles, elipses de tamanhos diferentes e esferas de tamanhos diferentes, respectivamente. Além disso, o mesmo método FCM, obteve vantagem para o contexto de esfera de tamanhos iguais sem ruído.

O FCMdd obteve vantagem no cenário de Esfera de tamanhos iguais com ruído, já o MFCM obteve vantagem no cenário elíptico de tamanhos iguais com ruído. O MFCMdd, método proposto nesta dissertação obteve vantagem perante todos os outros métodos quando se há o cenário de conjuntos de dados elipsóides de tamanhos iguais sem ruído.

Como citado em um dos objetivos deste estudo, comparando apenas os dois cenários multivariados (MFCM x MFCMdd), observa-se que o método proposto (MFCMdd) obteve vantagem perante o MFCM nos seguintes cenários:

- **0% de ruído:** Vantagem perante as médias do ARI e F-score para todos os quatro conjuntos de dados em estudo;
- **10% e 20% de ruído:** Nos conjuntos de dados 2 (Esfera de tamanhos diferentes) e 4 (Esfera de tamanhos iguais), com $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, as médias do ARI e F-score para o MFCMdd foram superiores ao do MFCM;

4.4.2 Conjuntos de Dados Reais

Em relação aos cinco conjuntos de dados reais, sendo eles o Haberman, Abalone, Seeds, Íris e o Wine. Pode-se verificar que:

O conjunto de dados Haberman possuiu uma ligeira vantagem para o MFCM. Este resultado é esperado visto que as caracaterísticas são de um cenário elíptico com tamanhos iguais com ruído com classes sobrespostas. Ou seja, pode-se concluir que o impacto das características de cada variável influencia no cálculo dos graus de pertinência para este conjunto de dados e como não se tem ruídos relevantes, o MFCM foi o método que obteve maior qualidade no agrupamento.

Para o conjunto de dados Abalone, há uma ligeira vantagem para o método proposto MFCMdd. As características deste conjunto de dados podem ser compreendidas como de caráter esférico de tamanhos iguais com ruído com classes sobrepostas. Ou seja, pode-se concluir que o impacto das características de cada variável influencia no cálculo dos graus de pertinência para este conjunto de dados, logo um cenário multivariado é mais indicado, além do quesito ruído em que se trabalhar com *medoids* tem as suas vantagens.

Para os conjuntos de dados Seeds e Íris, há uma vantagem expressiva para o método FCM. Este resultado é esperado visto que as caracaterísticas são de um cenário esférico com tamanhos iguais sem ruído sem classes sobrespostas. Ou seja, pode-se concluir que o impacto das características de cada variável não influenciam diretamente no cálculo dos graus de pertinência para este conjunto de dados e como não se foi constatado ruídos relevantes, o FCM obteve vantagem perante os outros métodos para esses dois conjuntos de dados.

Por fim, o conjunto de dados Wine, possuiu uma melhor performance para o método FCMdd. Este resultado é esperado visto que as caracaterísticas são de um cenário esférico com tamanhos iguais com ruído sem classes sobrespostas. Ou seja, pode-se concluir que o impacto das características de cada variável não influenciam diretamente no cálculo dos graus de pertinência para este conjunto de dados, porém, como foi constatado ruídos, se é mais indicado utilizar o cálculo dos centróides através de *medoids*, ao invés das médias como nos conjuntos de dados Seeds e Íris.

5 CONCLUSÕES

5.1 CONCLUSÕES

Nesta dissertação foi apresentado uma nova abordagem de agrupamento fuzzy que lida com os graus de pertinência baseados nas informações de cada variável utilizando *medoids* para o cálculo dos centróides. O método *Multivariate Fuzzy C-means* difundido na literatura leva em consideração o impacto de cada variável no cálculo dos graus de pertinência, porém, utiliza a média para o cálculo dos centróides o que traz margem para prejuízos quando se há um cenário com dados ruidosos, visto que a média é fortemente influenciada por valores aberrantes.

Com isso, o método de agrupamento apresentado neste trabalho leva em consideração não somente as características de cada variável nos cálculos dos graus de pertinência, mas também a utilização de observações do próprio conjunto de dados para serem os centróides, conhecidos como *medoids*.

Para apresentar a utilidade do método proposto (MFCMdd), foram realizados experimentos com dados simulados e aplicações para cinco conjuntos de dados. As métricas que serviram de avaliação para o comparativo do método de agrupamento proposto com outros três métodos disseminados na literatura (FCM, FCMdd e MFCM) foram o Índice de Rand Ajustado (ARI) e o F-score.

Quatro configurações de dados sintéticos com diferentes particularidades perante as matrizes de covariância e cinco conjuntos de dados reais do Repositório de Aprendizado de Máquina da UCI foram levados em consideração neste trabalho e avaliados perante as métricas citadas no parágrafo anterior.

Para o conjunto de dados sintéticos, o Índice de Rand Ajustado e o F-score foram utilizados como medidas de avaliação dos métodos de agrupamento. Geralmente, na literatura se é utilizado o método de Monte Carlo com 100 réplicas (SOUZA; CARVALHO, 2004; CARVALHO et al., 2006; CARVALHO; TENÓRIO; JUNIOR, 2006; FERREIRA; CARVALHO, 2014), porém, neste estudo foram utilizadas 50 réplicas para cada conjunto de dados, devido ao fato de que se foi verificado que acima desta quantidade de iterações, os resultados se mantiveram estáveis e assim para otimização da geração de resultados, foi adotado essa quantidade ao longo da dissertação. Como destaque o método proposto é mais adequado perante o Multivariate C-means, quando os conjuntos de dados não possuem ruídos independente de serem esféricos

ou elípticos ou quando se há o cenário de dados esféricos com ruído.

Em relação aos cinco conjuntos de dados reais, através das métricas citadas neste trabalho, o método proposto foi mais adequado para o conjunto de dados Abalone. Este resultado advém da presença de algumas características, sendo elas um caráter esférico de tamanhos iguais com ruído com classes sobrepostas, em que cada variável gerou um impacto significativo no cálculo dos graus de pertinência neste conjunto de dados e como há um percentual de ruído relevante, a utilização de medoids para o cálculo dos centróides é mais indicado, portante o MFCMdd obteve vantagem perante os demais métodos.

5.2 TRABALHOS FUTUROS

Durante o desenvolvimento deste trabalho, foram identificados algumas ideias para trabalhos futuros. A primeira ideia seria trabalhar com um algoritmo multivariado *fuzzy c-medoids* com ponderação, em que os pesos visam representar a importância de cada variável diferente para cada grupo e melhorar a qualidade do agrupamento, utilizando o *medoid* para o cálculo dos centróides.

Dados reais geralmente não seguem uma distribuição Normal. Logo, um possível tópico de trabalho futuro é considerar a geração de dados sintéticos através de outras distribuições de probabilidade, como por exemplo, as distribuições Exponencial, Gama, Beta, entre outras e verificar a eficácia dos métodos nesses casos.

Outra ideia que surgiu foi a utilização do Índice de Rand Fuzzyficado (FR Index) como métrica de avaliação dos agrupamentos. Isto faz sentido devido ao fato do trabalho estar lidando com agrupamentos difusos, logo, talvez essa métrica pudesse agregar para o presente estudo.

Também como uma possível ideia de futuros trabalhos, seria utilizar dados intervalares ao invés de dados pontuais, visto que é uma área de estudo que ainda não foi amplamente investigada, trazendo margem para novas descobertas.

Por fim, este trabalho utilizou conjuntos de dados com no máximo 3 classes, pretendendo possuir uma maior interpretabilidade do agrupamento. Portanto, como trabalhos futuro, poderia ser interessante a utilização de conjuntos de dados com um número maior de classes, buscando analisar como seriam os comportamentos dos métodos de agrupamento abordados.

REFERÊNCIAS

ALBUQUERQUE, M. A. de; BARROS, K. N. N. de O. Determinação do número de grupos em análise de agrupamento via de raio de influência. *Brazilian Journal of Development*, v. 6, n. 6, p. 38342–38355, 2020.

ALDENDERFER, M. S.; BLASHFIELD, R. K. *A review of clustering methods*. [S.l.]: SAGE Publications Ltd: London, 1984. 33–61 p.

ALTROCK, C. V. Fuzzy logic and neurofuzzy applications in business and finance. 1997.

ARAÚJO, R. M. M. de. Análise de componentes principais e análise de agrupamentos–aplicação em variáveis de educação e renda no estado de pernambuco. *e-xacta*, v. 3, n. 1, 2010.

BACHE, K.; LICHMAN, M. Uci machine learning repository. Irvine, CA, USA, 2013.

BEZDEK, J. C. Objective function clustering. In: *Pattern recognition with fuzzy objective function algorithms*. [S.l.]: Springer, 1981. p. 43–93.

BEZDEK, J. C. *Pattern recognition with fuzzy objective function algorithms*. [S.l.]: Springer Science & Business Media, 2013.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, Elsevier, v. 10, n. 2-3, p. 191–203, 1984.

BUSSAB, W. d. O.; MIAZAKI, É. S.; ANDRADE, D. d. Introdução à análise de agrupamentos. 1990.

CARVALHO, F. d. A. D.; SOUZA, R. M. D.; CHAVENT, M.; LECHEVALLIER, Y. Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, Elsevier, v. 27, n. 3, p. 167–179, 2006.

CARVALHO, F. d. A. de; TENÓRIO, C. P.; JUNIOR, N. L. C. Partitional fuzzy clustering methods based on adaptive quadratic distances. *Fuzzy Sets and Systems*, Elsevier, v. 157, n. 21, p. 2833–2857, 2006.

COX, E. *The fuzzy systems handbook: a practitioner's guide to building, using, and maintaining fuzzy systems*. [S.l.]: Academic Press Professional, Inc., 1994.

CRUZ, M. D.; OCHI, L. S. O problema de clusterização automática: Um novo método utilizando ils. In: *Anais do X Congresso Brasileiro de Inteligência Computacional (X CBIC), Fortaleza-CE*. [S.l.: s.n.], 2011.

DAVE, R. N. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, Elsevier, v. 12, n. 11, p. 657–664, 1991.

DERCZYNSKI, L. Complementarity, f-score, and nlp evaluation. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. [S.l.: s.n.], 2016. p. 261–266.

DIDAY, E.; SIMON, J. Clustering analysis. In: *Digital pattern recognition*. [S.l.]: Springer, 1976. p. 47–94.

EGRIOGLU, E.; ALADAG, C.; YOLCU, U.; USLU, V. R.; ERILLI, N. A. Fuzzy time series forecasting method based on gustafson–kessel fuzzy clustering. *Expert Systems with Applications*, Elsevier, v. 38, n. 8, p. 10355–10357, 2011.

FERREIRA, M. R.; CARVALHO, F. D. A. D. Kernel fuzzy c-means with automatic variable weighting. *Fuzzy Sets and Systems*, Elsevier, v. 237, p. 1–46, 2014.

FREI, F. *Introdução à análise de agrupamentos*. [S.l.]: Unesp, 2006.

GROENEN, P. J.; JAJUGA, K. Fuzzy clustering with squared minkowski distances. *Fuzzy Sets and Systems*, Elsevier, v. 120, n. 2, p. 227–237, 2001.

HU, Q.; AN, S.; YU, D. Soft fuzzy rough sets for robust feature evaluation and selection. *Information Sciences*, Elsevier, v. 180, n. 22, p. 4384–4400, 2010.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, Acm New York, NY, USA, v. 31, n. 3, p. 264–323, 1999.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data: An introduction to cluster analysis–john wiley & sons. *Inc., New York*, 1990.

KAUFMANN, L. Clustering by means of medoids. In: *Proc. Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, 1987*. [S.l.: s.n.], 1987. p. 405–416.

KRISHNAPURAM, R.; JOSHI, A.; YI, L. A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In: IEEE. *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No. 99CH36315)*. [S.l.], 1999. v. 3, p. 1281–1286.

KUO, R.-J.; WU, Y.-H.; HSU, T.-S. Integration of fuzzy set theory and topsis into hfmea to improve outpatient service for elderly patients in taiwan. *Journal of the Chinese Medical Association*, Elsevier, v. 75, n. 7, p. 341–348, 2012.

LABROCHE, N. New incremental fuzzy c medoids clustering algorithms. In: IEEE. *2010 Annual Meeting of the North American Fuzzy Information Processing Society*. [S.l.], 2010. p. 1–6.

LUCASIUS, C. B.; DANE, A. D.; KATEMAN, G. On k-medoid clustering of large data sets with the aid of a genetic algorithm: background, feasiblity and comparison. *Analytica Chimica Acta*, Elsevier, v. 282, n. 3, p. 647–669, 1993.

MACQUEEN, J. Classification and analysis of multivariate observations. In: *5th Berkeley Symp. Math. Statist. Probability*. [S.l.: s.n.], 1967. p. 281–297.

PAL, N. R.; PAL, K.; KELLER, J. M.; BEZDEK, J. C. A possibilistic fuzzy c-means clustering algorithm. *IEEE transactions on fuzzy systems*, IEEE, v. 13, n. 4, p. 517–530, 2005.

PIMENTEL, B. A. Métodos de agrupamento difuso multivariado baseados no fuzzy c-means. Universidade Federal de Pernambuco, 2017.

PIMENTEL, B. A.; SOUZA, R. M. D. A multivariate fuzzy c-means method. *Applied Soft Computing*, Elsevier, v. 13, n. 4, p. 1592–1607, 2013.

PIMENTEL, B. A.; SOUZA, R. M. de. Multivariate fuzzy c-means algorithms with weighting. *Neurocomputing*, Elsevier, v. 174, p. 946–965, 2016.

PINHEIRO, D. N.; ALOISE, D.; BLANCHARD, S. J. Convex fuzzy k-medoids clustering. *Fuzzy Sets and Systems*, Elsevier, v. 389, p. 66–92, 2020.

PRASS, F. S. et al. Estudo comparativo entre algoritmos de análise de agrupamentos em data mining. Florianópolis, SC, 2004.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, Taylor & Francis, v. 66, n. 336, p. 846–850, 1971.

ROSES, C. F.; LEIS, R. P. Um estudo das condições sócio-econômicas de municípios gaúchos através da análise de cluster. *Revista Administração On Line*, v. 3, n. 3, 2002.

SABZEKAR, M.; NAGHIBZADEH, M. Fuzzy c-means improvement using relaxed constraints support vector machines. *Applied Soft Computing*, Elsevier, v. 13, n. 2, p. 881–890, 2013.

SOUZA, R. M. de; CARVALHO, F. d. A. D. Clustering of interval data based on city–block distances. *Pattern Recognition Letters*, Elsevier, v. 25, n. 3, p. 353–365, 2004.

TRYON, R. C. Cluster analysis. edwards brothers. *Ann Arbor, Michigan*, v. 122, 1939.

XU, R.; WUNSCH, D. Survey of clustering algorithms. *IEEE Transactions on neural networks*, Ieee, v. 16, n. 3, p. 645–678, 2005.

YEUNG, K. Y.; RUZZO, W. L. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. *Bioinformatics*, v. 17, n. 9, p. 763–774, 2001.

ZADEH, L. A. Fuzzy sets. *Information and control*, Elsevier, v. 8, n. 3, p. 338–353, 1965.

ZHANG, L.; PEDRYCZ, W.; LU, W.; LIU, X.; ZHANG, L. An interval weighed fuzzy c-means clustering by genetically guided alternating optimization. *Expert Systems with Applications*, Elsevier, v. 41, n. 13, p. 5960–5971, 2014.

ZHAO, F.; FAN, J.; LIU, H. Optimal-selection-based suppressed fuzzy c-means clustering algorithm with self-tuning non local spatial information for image segmentation. *Expert systems with applications*, Elsevier, v. 41, n. 9, p. 4083–4093, 2014.