UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ERIC CARDOSO SOARES

MÉTODOS DE AGRUPAMENTO COM PESOS PONDERADOS POR GRUPO PARA DADOS SIMBÓLICOS INTERVALARES

Recife

2024

ERIC CARDOSO SOARES

MÉTODOS DE AGRUPAMENTO COM PESOS PONDERADOS POR GRUPO PARA DADOS SIMBÓLICOS INTERVALARES

Trabalho apresentado ao Programa de Pós-graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador: Getúlio José Amorim do Amaral

Coorientadora: Renata Maria Cardoso Rodrigues de Souza

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Soares, Eric Cardoso.
Métodos de agrupamento com pesos ponderados por grupo para dados simbólicos intervalares / Eric Cardoso Soares. - Recife, 2024.
88f.: il.

Dissertação (Mestrado), Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Programa de Pós-Graduação em Estatística, 2024.
Orientação: Getúlio José Amorim do Amaral.
Coorientação: Renata Maria Cardoso Rodrigues de Souza.

1. Análise de Dados Simbólicos; 2. Dados Intervalares; 3. Métodos de Nuvens Dinâmicas; 4. Distâncias Adaptativas; 5. Ponderação por Grupo. I. Amaral, Getúlio José Amorim do. II. Souza, Renata Maria Cardoso Rodrigues de. III. Título.

UFPE-Biblioteca Central

ERIC CARDOSO SOARES

MÉTODOS DE AGRUPAMENTO COM PESOS PONDERADOS POR GRUPO PARA DADOS SIMBÓLICOS INTERVALARES

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovado em 25 de julho de 2024.

BANCA EXAMINADORA

Prof. Dr. Getúlio José Amorim do Amaral
Presidente, UFPE

Prof. Dr. Klaus Leite Pinto Vasconcellos
Examinador Interno

Prof. Dr. Adriano Lorena Inácio de Oliveira
Examinador Externo

AGRADECIMENTOS

A Deus, por ter-me permitido percorrer esta breve e intensa caminhada, mostrando que sou mais forte do que imagino. Agradeço por me manter firme diante de todas as circunstâncias, permitindo-me concluir mais uma realização pessoal e profissional durante minha carreira acadêmica.

À minha mãe, Maria da Guia, e ao meu pai, Clodoaldo, pelo suporte e por sempre estarem ao meu lado, recarregando-me com todo o amor necessário. Aos meus irmãos, Haleff e Lucas, por toda a irmandade. Amo vocês!

Meus agradecimentos especiais à Vanessa, que ao meu lado comemora mais uma conquista que é nossa. O seu suporte durante toda esta trajetória fez com que o processo se tornasse mais leve. Com você ao meu lado, sou mais forte. Te amo!

Aos meus familiares, avós, tios, tias, primos, primas; em especial, à minha tia Wanda, que, em vida, cuidou de mim como um filho. Tem sido difícil desde sua partida. Saudades eternas!

Ao meu orientador, Getúlio, e à coorientadora, Renata, pelas contribuições à nossa pesquisa. As contribuições de vocês fizeram com que o trabalho fosse desenvolvido com maestria.

A todos os professores e professoras do Departamento de Estatística da UFPE, em especial à Maria do Carmo, Gauss, Cribari e Roberto, que, com excelência, contribuíram para meu desenvolvimento acadêmico.

Por fim, à CAPES pelo suporte financeiro fornecido durante os anos de mestrado.

RESUMO

O desenvolvimento de métodos para a Análise de Dados Simbólicos é necessário para lidar com dados de elevado grau de complexidade. Diante disso, propomos novos métodos de nuvens dinâmicas utilizando a distância *City-Block* para dados simbólicos intervalares. Nestes métodos, que são adaptações de dados pontuais, introduzimos o peso do *cluster*, que busca minimizar problemas recorrentes de agrupamento, como péssima inicialização e obtenção de um mínimo local pobre. Para a validação dos métodos propostos, foram realizados experimentos com dados sintéticos balanceados, desbalanceados e dados reais, nos quais a qualidade do agrupamento foi avaliada por meio do Índice de *Rand* Ajustado e da Informação Mútua Normalizada. Para os dados sintéticos, foram necessárias a realização da simulação de Monte Carlo e testes estatísticos. Os experimentos evidenciaram que o desempenho dos métodos que utilizam o peso do *cluster* é superior aos métodos que não o utilizam, mostrando que essa ponderação tem potencial para corrigir os problemas de inicialização e de obtenção de um mínimo local pobre.

Palavras-chaves: Análise de Dados Simbólicos; Dados Interalares; Métodos de Nuvens Dinâmicas; Distâncias Adaptativas; Ponderação por Grupo.

ABSTRACT

The development of methods for Symbolic Data Analysis is necessary to deal with highly complex data. Given this, we propose new dynamic cloud methods using the City-Block distance for interval symbolic data. In these methods, which are adaptations of point data, we introduce the cluster weight, which seeks to minimize recurring clustering problems, such as poor initialization and obtaining a poor local minimum. To validate the proposed methods, experiments were carried out with balanced and unbalanced synthetic data and real data, in which the quality of the clustering was evaluated using the Adjusted Rand Index and Normalized Mutual Information. For synthetic data, Monte Carlo simulation and statistical tests were necessary. The experiments showed that the performance of methods that use the cluster weight is superior to methods that do not use it, showing that this weighting has the potential to correct initialization problems and obtaining a poor local minimum.

Keywords: Symbolic Data Analysis; Interval Data; Dynamic Cloud Methods; Adaptive Distances; Weighting by Group.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADA1	Distância Adaptativa por Atributo
ADA2	Distância Adaptativa por Atributo e Classe
ADAN	Distância Não Adaptativa
ADS	Análise de Dados Simbólicos
IMN	Informação Mútua Normalizada
IR	Índice de *Rand*
IRA	Índice de *Rand* Ajustado
PP	Peso do *Cluster* Ponderado por Meio do Produtório
PS	Peso do *Cluster* Ponderado por Meio do Somatório
SP	Sem o Peso do *Cluster*

SUMÁRIO

1 INTRODUÇÃO

Com o rápido avanço das tecnologias da informação e a crescente utilização da Internet, houve um aumento significativo no volume e na diversidade de dados disponíveis. A mineração de dados está se tornando cada vez mais importante devido ao seu papel fundamental na obtenção de informações e na realização de análises, especialmente em grandes volumes de dados. A precisão das informações é crucial para processos de tomada de decisão eficientes (AYESHA et al., 2010).

A mineração de dados visa obter conhecimento a partir dos dados. Suas principais tarefas incluem agrupamento, classificação, regressão, detecção de anomalias, aprendizado de regras de associação e sumarização. Por outro lado, o aprendizado de máquina se concentra em fazer previsões com base em treinamento e aprendizado, abrangendo métodos supervisionados e não supervisionados. A aplicabilidade e os desafios das técnicas de mineração de dados e aprendizado de máquina em grandes volumes de dados com alta complexidade têm sido amplamente estudados, apesar dos inúmeros obstáculos enfrentados (WANG, 2017).

Com o aumento das bases de dados, surge o desafio de extrair conhecimento útil para a análise e a tomada de decisões. Devido ao enorme volume de dados a ser processado, os métodos computacionais tradicionais se mostram ineficazes, pois não conseguem oferecer o desempenho necessário (DIDAY; NOIRHOMME-FRAITURE, 2008). A partir deste contexto, surgem os dados simbólicos.

Os dados simbólicos são uma forma de representar grandes conjuntos de dados de maneira mais compacta, embora mais complexa, como intervalos reais, distribuições de probabilidades e conjuntos de categorias, entre outros (DIDAY; NOIRHOMME-FRAITURE, 2008). Muitos conjuntos de dados simbólicos são originados pela agregação de conjuntos de dados pontuais, convertendo-os em conjuntos menores e mais manejáveis (DIDAY, 2016). Ao contrário dos dados clássicos, os dados simbólicos possuem variação e uma estrutura interna que devem ser consideradas durante a análise (BILLARD, 2006).

A análise de dados simbólicos é um campo que oferece diversos métodos para examinar conjuntos de dados. Os métodos estatísticos tradicionais carecem da potência e flexibilidade necessárias para interpretar conjuntos de dados extremamente grandes, e por isso foram desenvolvidas técnicas específicas de análise de dados simbólicos para extrair conhecimento desses dados (DIDAY; NOIRHOMME-FRAITURE, 2008). Dessa forma, a grande lacuna entre as enormes

bases de dados modernas e a extração de conhecimento está sendo novamente preenchida. Atualmente, uma das principais preocupações da análise de dados simbólicos é desenvolver novos métodos e adaptar métodos clássicos para trabalhar com esses novos tipos de dados (DIDAY; NOIRHOMME-FRAITURE, 2008).

No contexto de aprendizagem de máquina não supervisionado, os algoritmos de análise de agrupamento podem ser classificados em dois grandes grupos: aqueles que organizam um conjunto de dados em hierarquias de classes, utilizando uma árvore chamada dendrograma, e aqueles que estruturam os dados em uma partição com um número predefinido de classes. Em geral, os métodos que fornecem uma partição também oferecem um conjunto de representantes das classes, otimizando localmente um critério de adequação entre as classes e seus representantes.

Algoritmos de nuvem dinâmica englobam métodos de agrupamento particionais para a separação de um conjunto em um número pré-definido de grupos através da minimização de um critério. Este critério caracteriza o potencial de representatividade que os protótipos têm com relação a seus respectivos grupos. Algoritmos dessa natureza têm sido desenvolvidos e divulgados em várias áreas, incluindo medicina, biologia, economia, processamento de imagens, aprendizado de máquina, reconhecimento de padrões e mineração de dados, entre outras.

Na literatura, existem diversos métodos de nuvens dinâmicas para dados intervalares, bem como várias variações propostas ao longo dos anos. Esses métodos propõem melhorias no agrupamento para reconhecimento de grupos com diferentes formas, números variados de elementos em cada grupo, diferentes dimensões, entre outros aspectos. Para isso, alguns métodos introduzem pesos nas distâncias, denominadas distâncias adaptativas (CARVALHO; LECHEVALLIER, 2009; CARVALHO; BRITO; BOCK, 2006; SOUZA; CARVALHO, 2004), cujos pesos são estimados localmente ou globalmente. Esses métodos elevam os índices de avaliação e melhoram a qualidade do agrupamento, podendo ser extendidos a contextos reais.

O uso da distância nestes casos influencia na qualidade do agrupamento, a depender do conjunto de dados. Por exemplo, a distância Euclidiana é ideal para casos em que os dados estão organizados hiperesfericamente, o que dificilmente acontece em contextos reais. A distância *City-Block* mostra-se como uma alternativa mais robusta e menos sensíveis a *outliers* em comparação com a outra distância.

Acontece que, independentemente da distância escolhida, esses algoritmos de nuvens dinâmicas apresentam alguns problemas de inicialização e de convergência para mínimos locais inadequados. Isso significa que, dependendo da partição inicial do algoritmo de agrupamento,

o processo pode convergir para um mínimo local pobre. Portanto, é recomendável realizar várias execuções e selecionar a solução que apresenta o menor critério de adequação.

Na literatura, existem alguns métodos (OSKOUEI; BALAFAR; MOTAMED, 2021; LIU et al., 2019; TZORTZIS; LIKAS, 2014) que propõem soluções para esses problemas, em que são inseridos pesos para ponderar as distâncias entre os elementos e o representante da classe. Neste estudo apresentaremos uma extensão dessas soluções para métodos de nuvens dinâmicas destinados a dados simbólicos intervalares.

1.1 OBJETIVOS

O objetivo geral desta dissertação é desenvolver novos métodos de nuvens dinâmicas com o peso do *cluster* para dados simbólicos intervalares, utilizando a distância *City-Block*.

1. Propor novos métodos de nuvens dinâmicas para dados simbólicos intervalares que busquem solucionar problemas de inicialização e ótimo local pobre;
2. Estender métodos da literatura (OSKOUEI; BALAFAR; MOTAMED, 2021; LIU et al., 2019; TZORTZIS; LIKAS, 2014) para trabalhar com dados simbólicos intervalares;
3. Desenvolver novos métodos de agrupamento para dados simbólicos intervalares baseados nas abordagens desenvolvidas, com as ponderações de soma e produto;
4. Avaliar o desempenho dos métodos propostos, comparando-os com abordagens já existentes na literatura, através da execução em conjuntos de dados sintéticos e dados reais.

1.2 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está estruturada em seis capítulos, sendo este o primeiro. No Capítulo 2, apresenta-se uma revisão sobre diversos métodos de nuvens dinâmicas para dados intervalares propostos na literatura. No Capítulo 3, os métodos de nuvens dinâmicas para dados intervalares são estendidos e novos métodos de agrupamento com o peso do *cluster* são apresentados, considerando as ponderações por meio da soma e do produto. O Capítulo 4 detalha a metodologia utilizada, assim como apresenta os resultados das avaliações de desempenho dos métodos da literatura e dos métodos propostos com dados intervalares sintéticos. No Capítulo 5, são apresentados os resultados dos experimentos com dados reais. No Capítulo

6, são expostas as conclusões finais desta dissertação, sugestões para trabalhos futuros e as contribuições realizadas. Por fim, no Apêndice A são apresentados as provas de proposições e no Apêndice B, os testes estatísticos de comparação múltipla.

2 ANÁLISE DE DADOS SIMBÓLICOS INTERVALARES

Em contraste com os dados pontuais tradicionais, a Análise de Dados Simbólicos (ADS) (BOCK; DIDAY, 2012; BILLARD; DIDAY, 2003) é um campo de pesquisa e aplicação que descreve os dados por meio de variáveis mais complexas, denominadas dados simbólicos (BILLARD; DIDAY, 2019). Muitos conjuntos de dados simbólicos são criados a partir da agregação de grandes ou extremamente grandes conjuntos de dados clássicos, transformando-os em conjuntos menores e mais fáceis de manejar (DIDAY, 2016). Diferentemente dos dados clássicos, os dados simbólicos apresentam variação e uma estrutura interna que precisam ser consideradas durante a análise (BILLARD, 2006).

Os tipos de dados simbólicos são variados e incluem dados com múltiplos valores ou listas de dados categóricos, dados multimodais, dados intervalares, dados de histograma, dados de boxplot, dados poligonais (SILVA; SOUZA; CYSNEIROS, 2021), entre outros. Enquanto uma observação pontual clássica para dados quantitativos assume um valor pontual na reta real $\mathbb{R}$, os dados simbólicos do tipo intervalar são definidos como um subconjunto de $\mathbb{R}$. Neste estudo, serão considerados dados intervalares.

Seja uma variável aleatória $\mathbf{Y} = (Y_1, \ldots, Y_p)$ p-dimensional, tomando valores no espaço $\mathbb{R}^p$. Uma amostra aleatória de n observações leva a valores intervalares quando

$$\mathbf{Y}_i = ([a_i^1, b_i^1], \ldots, [a_i^j, b_i^j], \ldots, [a_i^p, b_i^p]),$$

em que $a_i^j \leq b_i^j$, com $1 \leq j \leq p$, $1 \leq i \leq n$, e os intervalos podem estar abertos ou fechados em quaisquer extremidades (por exemplo, $[a, b)$, $(a, b]$, $[a, b]$ ou (a, b)).

Dados simbólicos com valores intervalares podem ser encontrados em várias formas. Por exemplo, a temperatura de uma determinada cidade registrada ao longo dos meses do ano, ou o valor mínimo e máximo de uma ação específica registrado em um determinado período. Esses valores representam intervalos reais com limites inferiores e superiores, podendo, assim, ser considerados como variáveis simbólicas do tipo intervalo. Outra maneira de obter dados simbólicos é através da agregação de dados clássicos (CARVALHO; LECHEVALLIER, 2009; CARVALHO; LECHEVALLIER, 2009; CARVALHO; BRITO; BOCK, 2006; CARVALHO et al., 2006a; SOUZA; CARVALHO, 2004), conforme apresentada detalhadamente na Subseção 4.1.1.

A Análise de Agrupamentos (BLASHFIELD; ALDENDERFER, 1978) é uma técnica exploratória multivariada amplamente utilizada em diversas áreas de pesquisa (DING; LIU; WANG,

2024; HARDER et al., 2024; SEYALA; ABDULLAH, 2024; PANGESTU; SHAUFIAH; WIJAYA, 2024). A natureza multidisciplinar dos estudos sobre Análise de Agrupamentos tem impulsionado o desenvolvimento de novos métodos e aplicações para todos os tipos de dados simbólicos (SILVA et al., 2023). Os métodos de agrupamento, em geral, são divididos em dois grupos: métodos de agrupamento hierárquicos e métodos de agrupamento particionais (XU; WUNSCHII, 2005).

Os métodos hierárquicos (BILLARD; DIDAY, 2019) constroem hierarquias, que são clusters alinhados. Geralmente, estes são representados por árvores chamadas dendrogramas, mostrando as relações entre os clusters. As técnicas hierárquicas podem ser divididas em aglomerativas e divisivas.

Os métodos de partição dividem o conjunto de dados de observações $\mathbb{X}$ em grupos não hierárquicos, ou seja, não alinhados. Algoritmos de particionamento frequentemente utilizam os dados como unidades, como nos métodos do tipo K-médias, ou as distâncias, como nos métodos do tipo K-$medoids$, para realizar agrupamentos.

Uma partição P de um conjunto de observações $\mathbb{X}$ é composta por um conjunto de grupos $P = (C_1, \ldots, C_k, \ldots, C_K)$ que deve satisfazer duas condições: a interseção de qualquer par de grupos diferentes é vazia, ou seja, $C_{v_1} \cap C_{v_2} = \emptyset$ para todos $v_1 \neq v_2$, e a união de todos os K grupos é igual ao conjunto de dados original, ou seja, $\cup_v C_v = \mathbb{X}$. Essas condições garantem que cada observação pertence a exatamente um grupo e que todos os grupos juntos cobrem todas as observações do conjunto $\mathbb{X}$ (BILLARD; DIDAY, 2019).

Os métodos de partição podem ser classificados em partição rígida e partição difusa. Na partição rígida, cada elemento dos dados é associado a uma única classe. Já na partição difusa, cada elemento dos dados é associado a todas as classes da partição, ou seja, os elementos não pertencem apenas a um único cluster e as pertinências são calculadas dessa forma. Quando as pertinências têm valores entre 0 e 1, a soma de todas as possibilidades de um elemento pertencer a alguma classe é igual a 1, caracterizando um método probabilístico (PAL et al., 2005). Por outro lado, quando as pertinências assumem valores 0 ou 1, o método se torna rígido e o método difuso passa a ser equivalente a um método rígido (YANG; WU, 2006).

A ADS apresenta na literatura diversas abordagens introduzidas para realizar agrupamento com dados simbólicos intervalares. Carvalho (2007) apresentou métodos de agrupamento c-means difuso adaptativos e não adaptativos para particionar dados de intervalo simbólico. Os métodos propostos fornecem uma partição difusa e um protótipo para cada grupo, otimizando um critério de adequação baseado em distâncias euclidianas entre vetores de intervalos.

Também existem métodos de agrupamento baseados em *kernel*, os quais são capazes de

produzir hipersuperfícies de separação não lineares entre os grupos, mapeando implicitamente os dados em espaços de alta dimensão (GIROLAMI, 2002; ZHANG; CHEN, 2003; INOKUCHI; MIYAMOTO, 2004; FILIPPONE et al., 2008). Pimentel, Costa e Souza (2011) apresentaram um método baseado em funções *kernel* para particionar conjuntos de dados com valores intervalares, estendendo o algoritmo *fuzzy c-means* (ZHANG; CHEN, 2003). Vários métodos de agrupamento foram adaptados para usar funções *kernel*.

Rodríguez e Carvalho (2022) propuseram algoritmos de agrupamento para dados intervalares, introduzindo distâncias adaptativas locais e globais. Essas distâncias consideram a relevância conjunta das variáveis em cada fronteira dos intervalos, proporcionando uma análise mais precisa dos dados.

O foco deste estudo consiste na abordagem de um método de agrupamento particional rígido, chamado de métodos de nuvens dinâmicas, no qual os elementos são separados em classes e identifica-se um representante para cada classe. Estes métodos, que serão abordados na Seção 2.1, representam extensões do método de nuvens dinâmicas proposto por Diday (1971a) e Diday e Simon (1976a), adaptados para dados simbólicos do tipo intervalo (CARVALHO; LECHEVALLIER, 2009; CARVALHO; BRITO; BOCK, 2006; SOUZA; CARVALHO, 2004).

2.1 MÉTODOS DE NUVENS DINÂMICAS PARA DADOS INTERVALARES

Os algoritmos de nuvens dinâmicas são métodos de agrupamento não hierárquicos que visam, simultaneamente, particionar um conjunto de elementos em um número predefinido de classes e identificar protótipos ou representantes dessas classes, como médias, eixos, distribuições de probabilidade, entre outros. O objetivo é minimizar um critério que avalia a adequação entre as classes e seus protótipos (DIDAY; SIMON, 1976b; DIDAY, 1971b; CELEUX et al., 1989).

As vantagens desses métodos residem na formulação do problema de agrupamento como uma otimização de um critério de ajuste entre classes e seus representantes, além de criar uma estrutura na qual os algoritmos podem encontrar uma solução ótima local. Contudo, a convergência desses algoritmos pode ser problemática, dependendo tanto da configuração inicial dos pontos quanto da escolha da função de representação e da métrica de distância utilizada para medir o ajuste entre um grupo e sua representação.

Diversos avanços baseados em algoritmos de agrupamento dinâmico para conjuntos de dados descritos por dados intervalares simbólicos têm sido apresentados. Souza e Carvalho (2004) introduziram o agrupamento dinâmico com métodos não adaptativos e adaptativos

para intervalos, utilizando distâncias *City-Block*. Carvalho, Brito e Bock (2006) propuseram um método de agrupamento dinâmico com a distância L_2 para dados intervalares e apresentaram maneiras de padronizar esses dados. Chavent e Lechevallier (2002) desenvolveram um método de agrupamento dinâmico com distâncias adaptativas de *Hausdorff*, que superou a distância adaptativa *City-Block*. Carvalho e Lechevallier (2009) introduziram um método de agrupamento dinâmico com uma distância adaptativa única para dados intervalares, onde os pesos das distâncias mudam por variável a cada iteração do algoritmo. Carvalho et al. (2006b) apresentaram um método de agrupamento dinâmico particional para dados de intervalo baseado em distâncias adaptativas de *Hausdorff*.

Um dos métodos mais conhecidos de nuvens dinâmicas é o K-médias (JAIN, 2010; MACQUEEN et al., 1967). Este método proporciona uma partição rígida do conjunto de objetos em K grupos, onde cada objeto é atribuído ao grupo cuja média dos elementos do grupo está mais próxima. Embora seja comum o uso da distância Euclidiana em métodos de agrupamento, especialmente quando os grupos estão organizados hiperesfericamente, essa condição é rara na prática. No entanto, a distância Euclidiana pode ser sensível a ruídos e valores discrepantes. Algumas variações do K-médias podem ser consultadas em Reddy e Vinzamuri (2018) e Charu e Chandan (2013).

Uma variação interessante é o agrupamento K-medianas, que calcula a mediana para cada grupo em vez da média, tornando-se assim mais robusta em relação a *outliers* em comparação com o K-médias. O objetivo é encontrar subconjuntos de pontos medianos que minimizem o custo de atribuição dos objetos às medianas mais próximas. Diante desse cenário, optaremos por utilizar métodos de agrupamento para dados simbólicos intervalares, os quais são baseados na métrica de distância *City-Block*.

Também existem os métodos de nuvens dinâmicas com distâncias adaptativas, os quais definem diferentes distâncias a cada iteração. A proposta do agrupamento dinâmico com distâncias adaptativas, conforme apresentada por Diday e Govaert (1977), é associar uma distância a cada grupo, definida de acordo com sua estrutura dentro do grupo. A vantagem dessa abordagem é que o algoritmo de agrupamento é capaz de reconhecer diferentes formas e tamanhos de agrupamentos. Em uma abordagem, o peso é calculado globalmente, enquanto em outra abordagem, o peso é calculado localmente para cada grupo.

2.1.1 Distância Não Adaptativa

O método de nuvens dinâmicas com distância não adaptativa, conforme apresentado por Souza e Carvalho (2004), consiste em um algoritmo iterativo composto pelas etapas de inicialização, representação, alocação e critério de parada. Seu objetivo é obter uma partição com classes e cada classe com seus representantes, de modo que a função que mede a distância entre os grupos e seus representantes seja minimizada localmente.

Dado $\mathbb{X}$ um vetor de variáveis aleatórias $p-$dimensionais de observações simbólicas do tipo intervalo, de modo que $\mathbb{X} = \{x_1, \ldots, x_i, \ldots, x_n\}$, em que a observação $\mathbf{x}_i \in \mathbb{X}$ é representada como $\mathbf{x}_i = \Big([a_i^1, b_i^1], \ldots, [a_i^j, b_i^j], \ldots, [a_i^p, b_i^p]\Big)$. Seja uma partição $P = (C_1, \ldots, C_k, \ldots, C_K) \in \mathbb{X}$, com K classes e $\mathbf{G}^K = \mathbf{G} \times \ldots \times \mathbf{G}$ um conjunto de K-uplas $G = (\mathbf{g}_1, \ldots, \mathbf{g}_k, \ldots, \mathbf{g}_K)$ com $\mathbf{g}_k \in \mathbf{G}$. Seja $\mathbf{g}_k$ o protótipo intervalar p-dimensional da k-ésima classe, definido como $\mathbf{g}_k = \Big([\alpha_k^1, \beta_k^1], \ldots, [\alpha_k^j, \beta_k^j], \ldots, [\alpha_k^p, \beta_k^p]\Big)$.

O algoritmo de nuvens dinâmicas está relacionado à busca por uma partição $P^* \in \mathbf{P}_K$ em K classes e seus representantes, chamados de protótipo correspondente às classes $G^* \in \mathbf{G}^K$, de modo que

$$J_1(P^*, G^*) = \min\{J_1(P, G) \mid P \in \mathbf{P}_k,\ G \in \mathbf{G}^k\} \tag{2.1}$$

O critério $J_1(P, G)$ mede a adequação entre a partição P e o representante dessa partição G, conforme definido abaixo

$$J_1 = \sum_{k=1}^{K} \sum_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k) \tag{2.2}$$

em que $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ representa a distância entre o elemento $\mathbf{x}_i$ e o protótipo $\mathbf{g}_k$ A distância d_ϕ utilizada é a distância d_{L_1} de *Manhatan* ou *City-Block*, determinada por meio da Eq. 2.3.

$$d_\phi(\mathbf{x}_i, \mathbf{g}_k) = d_{L_1}(\mathbf{x}_i, \mathbf{g}_k) = \sum_{j=1}^{p} (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{2.3}$$

Com isso, o algoritmo de nuvens dinâmicas se tornará um K-medianas para dados simbólicos intervalares. Assim, substituindo a Eq. 2.3 em 2.2, obtemos a função objetivo J_1, com a distância *City-Block*, dada pela Eq. 2.4.

$$J_1 = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{2.4}$$

A inicialização do algoritmo é feita por uma partição obtida aleatoriamente $P^0 = (C_1^0, \ldots, C_k^0, \ldots, C_K^0)$ do conjunto de observações $\mathbb{X}$ ou por meio da escolha de K elementos diferentes do conjunto de observações para formar $G^0 = (\mathbf{g}_1^0, \ldots, \mathbf{g}_k^0, \ldots, \mathbf{g}_K^0)$, atribuindo cada elemento $\mathbf{x}_i$ ao protótipo $\mathbf{g}_{k^*}$ mais próximo, onde $k^* = \arg\min d_{L_1}(\mathbf{x}_i, \mathbf{g}_k)$.

As etapas de representação e alocação são repetidas durante algumas iterações, até que ocorra a convergência da função objetivo de modo que J_1 atinja um valor estacionário, geralmente representando o mínimo local. Os passos de representação nos novos protótipos $G^{\iota+1} = (\mathbf{g}_1^{\iota+1}, \ldots, \mathbf{g}_k^{\iota+1}, \ldots, \mathbf{g}_K^{\iota+1})$ são obtidos por meio da partição atual $P^\iota = (C_1^\iota, \ldots, C_k^\iota, \ldots, C_K^\iota)$, onde ι representa a iteração corrente. O cálculo do protótipo $\mathbf{g}_k$ que minimiza J_1 por meio da d_{L_1} é descrito na Seção 2.2. Mais detalhes sobre o algoritmo serão apresentados na Seção 2.3.

Além disso, dependendo da solução inicial (P^0, G^0), o algoritmo converge para uma solução ótima local para J_1. Isso significa que, dependendo das partições escolhidas inicialmente, diferentes valores de J_1 podem ser obtidos. Portanto, dependendo do objetivo da análise, é recomendável realizar um número pré-estabelecido de repetições do algoritmo com diferentes soluções iniciais obtidas aleatoriamente e selecionar como solução final o par (P^*, G^*) cujo valor da função objetivo J_1 seja mínimo.

2.1.2 Distância Adaptativa por Atributo

O método de nuvens dinâmicas com distâncias adaptativas por atributo (CARVALHO; LECHEVALLIER, 2009) define distâncias que variam dinamicamente a cada iteração do algoritmo. Essas distâncias são parametrizadas por um vetor de pesos, que pode ser estimado de maneira global para todas as classes, por atributo, ou localmente para cada classe. Essa flexibilidade permite ao algoritmo reconhecer uma variedade de formas entre os padrões, tornando possível a identificação de classes de tamanhos diferentes.

Considerando as definições de $\mathbb{X}$, $\mathbf{x}_i$, $\mathbf{g}_k$ e ι fornecidas na Subseção 2.1.1, podemos descrever o método de nuvens dinâmicas com distâncias adaptativas para dados simbólicos intervalares como um algoritmo composto por passos de inicialização, representação, alocação e critério de parada. O objetivo é obter uma partição P de $\mathbb{X}$ com K classes, um conjunto de

protótipos G_k, e um vetor de pesos $\boldsymbol{\lambda} = (\lambda^1, \ldots, \lambda^j, \ldots, \lambda^p)$, de modo que a função objetivo J_2, que mede o ajuste entre os grupos e seus protótipos, seja minimizada localmente.

Seja $\mathbf{P}_K$ um conjunto de partições $P = (C_1, \ldots, C_k, \ldots, C_K)$ de $\mathbb{X}$ com $\mathbf{g}_k \in \mathbf{G}$ e $\mathbf{D}^K = \mathbf{D} \times \ldots \times \mathbf{D}$ um conjunto de K distâncias $D = (\mathbf{d}_1, \ldots, \mathbf{d}_k, \ldots, \mathbf{d}_K)$ em que $\mathbf{d}_k \in D$. O algoritmo de nuvens dinâmicas com distância adaptativa por atributo (CARVALHO; LECHEVALLIER, 2009) está relacionado à busca por uma partição $P \in \mathbf{P}_K$ em K classes e seus representantes, chamados de protótipo correspondente às classes $G^* \in \mathbf{G}^K$ e um conjunto de distâncias $D^* \in \mathbf{D}^K$ tal que

$$J_2(P^*, G^*, D^*) = \min\{J_2(P, G, W) \mid P \in \mathbf{P}_k,\ G \in \mathbf{G}^k,\ D \in \mathbf{D}^K\} \tag{2.5}$$

onde $J_2(P, G, D)$ mede a adequação entre a partição P, o representante desta partição G, e o conjunto de distâncias D, definido abaixo.

$$J_2 = \sum_{k=1}^{K} \sum_{i \in C_k} d_{A_\phi}(\mathbf{x}_i, \mathbf{g}_k) \tag{2.6}$$

A distância adaptativa por atributo $d_{A_\phi}(\mathbf{x}_i, \mathbf{g}_k)$ do elemento $\mathbf{x}_i$ para o protótipo $\mathbf{g}_k$ é parametrizada por um vetor de pesos $\boldsymbol{\lambda} = (\lambda^1, \ldots, \lambda^j, \ldots, \lambda^p)$, conforme segue

$$d_{A_\phi}(\mathbf{x}_i, \mathbf{g}_k) = \sum_{j=1}^{p} \lambda^j d_\phi(\mathbf{x}_i, \mathbf{g}_k) \tag{2.7}$$

onde $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ é a distância *City-Block*, portanto, a distância adaptativa por atributo utilizando a distância L_1 é dada por:

$$d_{A^1_{L_1}}(\mathbf{x}_i, \mathbf{g}_k) = \sum_{j=1}^{p} \lambda^j (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{2.8}$$

Então, a função objetivo J_2, que utiliza a distância adaptativa por atributo com a distância *City-Block*, é dada por meio da Eq. 2.9.

$$J_2 = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \lambda^j (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{2.9}$$

A cada iteração, as distâncias mudam, mas para todas as classes C_k permanecem as mesmas, uma vez que os pesos λ^j são estimados globalmente.

De forma análoga ao método apresentado na Subseção 2.1.1, é escolhida uma partição aleatória $P^0 = (C_1^0, \ldots, C_k^0, \ldots, C_K^0)$ do conjunto de observações $\mathbb{X}$, ou podem ser escolhidos K elementos diferentes do conjunto de observações para formar $G^0 = (\mathbf{g}_1^0, \ldots, \mathbf{g}_k^0, \ldots, \mathbf{g}_K^0)$, atribuindo assim cada elemento $\mathbf{x}_i$ ao protótipo $\mathbf{g}_{k*}$ mais próximo, onde $k* = \arg\min d_{A_{L_1}^1}(\mathbf{x}_i, \mathbf{g}_k)$.

Com isso, ocorre a alternância entre os passos de representação e alocação até a função J_2 atingir um valor estacionário, indicando a convergência do critério, geralmente indicando um mínimo local.

Neste algoritmo, a etapa de representação passa a ter duas definições: (i) os melhores protótipos e (ii) os melhores pesos. Em (i), os novos protótipos $G^{\iota+1} = (\mathbf{g}_1^{\iota+1}, \ldots, \mathbf{g}_k^{\iota+1}, \ldots, \mathbf{g}_K^{\iota+1})$ são obtidos por meio da partição atual $P^\iota = (C_1^\iota, \ldots, C_k^\iota, \ldots, C_K^\iota)$. O cálculo do protótipo $\mathbf{g}_k$ que minimiza J_2 por meio de $d_{A_{L_1}^1}$ é dado na Seção 2.2. Mais detalhes sobre o algoritmo serão apresentados na Seção 2.3.

Na definição (ii), com a partição atual $P^\iota = (C_1^\iota, \ldots, C_k^\iota, \ldots, C_K^\iota)$ e os protótipos $\mathbf{g}_k$ das classes C_k fixos, os vetores de pesos λ^j que minimizam a função objetivo J_2 obedecem às seguintes restrições:

$$\lambda^j > 0 \qquad \text{e} \qquad \prod_{j=1}^{p} \lambda^j = 1 \tag{2.10}$$

Dessa forma, os pesos λ^j que seguem os critérios apresentados em 2.10 são determinados e atualizados por meio da Eq. 2.11.

$$\lambda^j = \frac{\left\{ \prod_{h=1}^{p} \left[\sum_{k=1}^{K} \left(\sum_{i \in C_k} d_\phi(\mathbf{x}_i^h, \mathbf{g}_k^h) \right) \right] \right\}^{\frac{1}{p}}}{\sum_{k=1}^{K} \left(\sum_{i \in C_k} d_\phi(\mathbf{x}_i^j, \mathbf{g}_k^j) \right)} \tag{2.11}$$

Durante a etapa de alocação, uma nova partição $P^{\iota+1}$ é definida atribuindo cada observação $\mathbf{x}_i$ à classe C_{k*}. Esses passos são repetidos até que não haja mudança nos elementos e nas classes, ou seja, $P^\iota = P^{\iota+1}$. A prova de que a Eq. 2.11 minimiza o critério J_2 pode ser vista em Carvalho e Lechevallier (2009).

2.1.3 Distância Adaptativa por Atributo e Classe

O método de nuvens dinâmicas com distâncias adaptativas por atributo e classe (SOUZA; CARVALHO, 2004) é uma extensão do algoritmo discutido na Subseção 2.1.2. Nele, os pesos são estimados localmente para cada classe k, ao contrário da abordagem global apresentada anteriormente. Com isso, as premissas e suposições propostas anteriormente são utilizadas para definir a função objetivo J_3, conforme a seguir:

$$J_3 = \sum_{k=1}^{K} \sum_{i \in C_k} d_{A_\phi}(\mathbf{x}_i, \mathbf{g}_k) \tag{2.12}$$

Em que d_{A_ϕ} é a distância adaptativa por atributo e por classe, parametrizada por um vetor de pesos $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^j, \ldots, \lambda_k^p)$ para cada classe C_k

$$d_{A_\phi}(\mathbf{x}_i, \mathbf{g}_k) = \sum_{j=1}^{p} \lambda_k^j d_\phi(\mathbf{x}_i, \mathbf{g}_k) \tag{2.13}$$

onde d_{A_ϕ} é a distância adaptativa por atributo e classe, utilizando a distância *City-Block*, denotada por $d_{A_{L_1}^2}$, conforme descrito a seguir:

$$d_{A_{L_1}^2}(\mathbf{x}_i, \mathbf{g}_k) = \sum_{j=1}^{p} \lambda_k^j (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{2.14}$$

Os pesos λ_k^j são estimados localmente para cada classe k, resultando em uma distância diferente para cada classe, que muda a cada iteração ι. Assim, a função objetivo J_3 é definida da seguinte forma:

$$J_3 = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \lambda_k^j (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{2.15}$$

O algoritmo opera de forma análoga ao mostrado na Subseção 2.1.2, com o passo de representação sendo dividido em duas partes. Para determinar os melhores pesos λ_k^j, com a partição atual $P^\iota = (C_1^\iota, \ldots, C_k^\iota, \ldots, C_K^\iota)$ e os protótipos $\mathbf{g}_k$ das classes C_k fixos, os vetores de pesos λ_k^j que minimizam a função objetivo J_3 devem obedecer às seguintes restrições:

$$\lambda_k^j > 0 \qquad \text{e} \qquad \prod_{j=1}^{p} \lambda_k^j = 1 \tag{2.16}$$

Diante das restrições apresentadas em 2.16, os pesos λ_k^j são determinados e atualizados por meio da Eq. 2.17.

$$\lambda_k^j = \frac{\left\{ \prod_{h=1}^{p} \left[\left(\sum_{i \in C_k} d_\phi(\mathbf{x}_i^h, \mathbf{g}_k^h) \right) \right] \right\}^{\frac{1}{p}}}{\sum_{i \in C_k} d_\phi(\mathbf{x}_i^j, \mathbf{g}_k^j)} \tag{2.17}$$

Por fim, durante a etapa de alocação, uma nova partição $P^{\iota+1}$ é definida, atribuindo cada observação $\mathbf{x}_i$ à classe C_{k*}. Esses passos serão repetidos até que não ocorram mudanças nos elementos e classes, ou seja, $P^{\iota} = P^{\iota+1}$.

Na Eq. 2.17, para cada classe k e variável j, o denominador da fração representa a medida da dispersão entre os grupos, ou coesão, da respectiva variável e classe em relação ao protótipo $\mathbf{g}_k$. Enquanto o numerador permanece constante, representando a média geométrica das dispersões intraclasse de todas as variáveis. Como o numerador é fixo, o comportamento do peso λ_j^k é determinado exclusivamente pelo denominador. Assim, o valor do peso aumenta à medida que a dispersão intraclasse correspondente diminui, indicando que os elementos da variável estão mais próximos do protótipo da classe. A prova de que a Eq. 2.17 minimiza o critério J_3 pode ser vista em Souza e Carvalho (2004).

2.2 PROTÓTIPO

Os protótipos $\mathbf{g}_k$ da classe C_k estão relacionados com a distância utilizada durante o agrupamento. O melhor protótipo é encontrado quando o somatório da distância $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ entre os elementos $\mathbf{x}_i$ e seu respectivo protótipo $\mathbf{g}_k$ é minimizado.

Considerando que a distância utilizada nos algoritmos apresentados nas Subseções 2.1.1, 2.1.2 e 2.1.3 é a d_{L_1}, ou *City-Block*; o melhor protótipo intervalar, mostrado em Souza e Carvalho (2004), $\mathbf{g}_k = \left([\alpha_k^1, \beta_k^1], \ldots, [\alpha_k^j, \beta_k^j], \ldots, [\alpha_k^p, \beta_k^p]\right)$, é determinado pela mediana de α_k^j e β_k^j da classe C_k, como segue:

$$\alpha_k^j = \underset{x_i \in C_k}{\text{Me}} \left\{ a_i^j \right\} \qquad \text{e} \qquad \beta_k^j = \underset{x_i \in C_k}{\text{Me}} \left\{ b_i^j \right\} \tag{2.18}$$

Em que Me $\left\{a_i^j\right\}$ e Me $\left\{b_i^j\right\}$ são, respectivamente, as medianas dos limites inferiores e superiores do intervalo que compõe o protótipo $\mathbf{g}_k$ da j-ésima variável das observações $\mathbf{x}_i \in C_k$.

2.3 ALGORITMO

O Algoritmo 1 descreve as etapas a serem executadas durante o processo de agrupamento por nuvens dinâmicas, empregando diferentes estratégias de distância: distância não adaptativa, distância adaptativa por atributo e distância adaptativa por atributo e classe, conforme discutido nas Subseções 2.1.1, 2.1.2 e 2.1.3, respectivamente.

Algoritmo 1 Algoritmo de agrupamento por nuvens dinâmicas para dados intervalares com o uso de distância não adaptativa e adaptativas

Entrada: Conjunto de dados $\mathbb{X}$, Número de clusters K, $2 \leq K < n$; Distância d_{L_1} (2.3), $d_{A^1_{L_1}}$ (2.8) ou $d_{A^2_{L_1}}$ (2.14).
Saída: Uma partição C_k que divide o conjunto de dados $\mathbb{X}$ com K classes

(1) Inicialização
Selecione uma partição $P^0 = (C^0_1, \ldots, C^0_k, \ldots, C^0_K)$ do conjunto de observações $\mathbb{X}$ ou determine K objetos diferentes $(\mathbf{g}_1, \ldots, \mathbf{g}_k, \ldots, \mathbf{g}_K)$ entre $\mathbb{X}$ e associe cada objeto $\mathbf{x}_i$ para uma classe C_{k*} tal que $k* = \arg\min d_{L_1}(\mathbf{x}_i, \mathbf{g}_k)$ para construir a partição inicial $P^0 = (C^0_1, \ldots, C^0_k, \ldots, C^0_K)$

(2) Representação
Para $k = 1$ **até** K **Faça**
(i) Determinar o protótipo $\mathbf{g}^j_k$ como na Eq. 2.18
Se a distância é adaptativa **Então**
(ii) Determinar o peso λ^j ou λ^j_k, de acordo com a Eq. 2.11 ou Eq. 2.17
Fim Se
Fim Para

(3) Alocação
$teste \leftarrow 0$
Para $i = 1$ **até** n **Faça**
defina a classe C_{k*} tal que

$$k* = \arg\min d(\mathbf{x}_i, \mathbf{g}_k), \text{ onde}$$
$$d(\mathbf{x}_i, \mathbf{g}_k) = d_{L_1}(\mathbf{x}_i, \mathbf{g}_k), \text{se usado a Eq. 2.3}$$
$$d(\mathbf{x}_i, \mathbf{g}_k) = d_{A^1_{L_1}}(\mathbf{x}_i, \mathbf{g}_k), \text{se usado a Eq. 2.8}$$
$$d(\mathbf{x}_i, \mathbf{g}_k) = d_{A^2_{L_1}}(\mathbf{x}_i, \mathbf{g}_k), \text{se usado a Eq. 2.14}$$

Se $i \in C_k$ **e** $k* \neq k$ **Então**
$teste \leftarrow 1$
$C_{k*} \leftarrow C_{k*} \cup \{i\}$
$C_k \leftarrow C_k \setminus \{i\}$
Fim Se
Fim Para

(4) Critério de Parada
Se $teste = 0$ **Então**
Pare
Senão
volte para ***(2)***
Fim Se

3 PESO DO *CLUSTER* EM MÉTODOS DE NUVENS DINÂMICAS

Neste capítulo, apresentamos o peso do *cluster* ponderado por meio do somatório e produtório. Em seguida, propomos novos método de nuvens dinâmicas com o peso do *cluster* ponderado para dados simbólicos intervalares. Estes métodos foram baseados em adaptações dos métodos de nuvens dinâmicas para dados clássicos, conforme descrito em Oskouei, Balafar e Motamed (2021), Liu et al. (2019) e Tzortzis e Likas (2014).

Dentre os algoritmos apresentados no Capítulo 2, existe a ponderação por meio da estimação adaptativa global e da estimação adaptativa local (LIU et al., 2019). O principal problema desses algoritmos é que eles são muito sensíveis à inicialização dos grupos primários, de modo que protótipos de grupos iniciais inadequados levam a mínimos locais ruins (OSKOUEI; BALAFAR; MOTAMED, 2021).

O peso do *cluster* resolve esse problema em grande parte, pois o cálculo do peso dos *clusters* é realizado com base nas distâncias apresentadas nas Eqs. 2.3, 2.8 e 2.14, equilibrando essas distâncias. Os pesos dos *clusters* são calculados automaticamente com base nas amostras atribuídas aos grupos durante a iteração. Sistematicamente, grupos de maior qualidade são obtidos independentemente dos protótipos iniciais (OSKOUEI; BALAFAR; MOTAMED, 2021). Entende-se por dispersão intra-*cluster* do grupo k, a soma das distâncias d_ϕ do vetor de observações $\mathbf{x}_i$ pertencentes à partição C_k para o protótipo $\mathbf{g}_k$, como $\sum_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k)$.

Os pesos do *cluster* predispõem nosso algoritmo a minimizar principalmente os grupos que atualmente exibem uma grande dispersão intra-*cluster*, essencialmente limitando a ocorrência de grupos de grande dispersão no resultado, e são aprendidos automaticamente juntamente com os pesos das distâncias adaptativas. Os pesos são aprendidos através de um procedimento iterativo e permitem que soluções de alta qualidade sejam sistematicamente descobertas, independentemente da inicialização (TZORTZIS; LIKAS, 2014; OSKOUEI; BALAFAR; MOTAMED, 2021).

Aparentemente (OSKOUEI; BALAFAR; MOTAMED, 2021; LIU et al., 2019; TZORTZIS; LIKAS, 2014), para grupos com pesos mais elevados, a distância ponderada pelo peso do *cluster* dos seus representantes das amostras aumenta. Consequentemente, um grupo com grande dispersão intra-*cluster* pode perder algumas de suas amostras atuais que estão longe de seu centro, e espera-se que sua dispersão intra-*cluster* diminua. Ao mesmo tempo, grupos com baixa dispersão intra-*cluster*, devido aos pequenos pesos, também podem adquirir amostras

que não estão próximas de seus centros, e sua dispersão intra-*cluster* aumentará (TZORTZIS; LIKAS, 2014; OSKOUEI; BALAFAR; MOTAMED, 2021).

O algoritmo de agrupamento, desta forma, não sofre as consequências causadas pela inicialização aleatória. Ao aplicar este mecanismo de ponderação, os resultados tornam-se menos afetados pela inicialização e soluções de alta qualidade podem ser descobertas de forma mais consistente, mesmo partindo de um conjunto inicial de centros ruim. Além disso, os grupos obtidos são balanceados em relação à sua dispersão.

A seguir, apresentamos a equação para determinar o peso do *cluster* com a ponderação por meio da soma na Seção 3.1 e o peso do *cluster* com a ponderação por meio do produto na Seção 3.2. Na Seção 3.3, mostramos a inserção desses pesos nos métodos apresentados no Capítulo 2, gerando novos métodos de algoritmos de nuvens dinâmicas com o peso do *cluster* ponderado por meio do somatório e com o peso do *cluster* ponderado por meio do produtório para dados intervalares.

3.1 PESO DO *CLUSTER* (PONDERAÇÃO SOMA)

O peso λ_k do *cluster* k ponderado por meio do somatório é determinado e atualizado por meio da Eq. 3.1.

$$\lambda_k = \left[\sum_{h=1}^{K} \left[\frac{\left(\sum\limits_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k) \right)}{\left(\sum\limits_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_h) \right)} \right]^{\frac{1}{t-1}} \right]^{-1} \tag{3.1}$$

onde $\boldsymbol{\lambda} = [\lambda_k]$ representa um vetor com K pesos do *cluster*, em que λ_k representa o peso do k-ésimo *cluster*. Além disso, seguindo a regra em Celebi, Kingravi e Vela (2013), também é adicionado um coeficiente coesivo t que suaviza a transição do peso do *cluster* na t-ésima iteração. O valor deste expoente converge para um ponto mínimo quando $0 \leq t < 1$. O peso do *cluster* ponderado segue as seguintes restrições.

$$\lambda_k > 0 \qquad \text{e} \qquad \sum_{k=1}^{K} \lambda_k = 1 \tag{3.2}$$

Com isso, podemos notar que o peso λ_k do grupo k sempre será positivo e a soma dos K pesos sempre resultará em 1. Observa-se na Eq. 3.1 que, para o grupo k, o numerador do

peso é a dispersão intra-*cluster* da partição C_k em relação ao protótipo $\mathbf{g}_k$. O denominador é sempre constante e representa a soma intra-*cluster* de todos os K grupos. Com isso, verifica-se que o comportamento do peso λ_k depende apenas do numerador. Assim, observar que quanto maior for a dispersão do grupo k, maior será o peso λ_k. A prova de que a Eq. 3.1 minimiza o critério de adequação pode ser vista no Apêndice A.

3.2 PESO DO *CLUSTER* (PONDERAÇÃO PRODUTO)

O peso λ_k do *cluster* k ponderado por meio do produtório é determinado e atualizado por meio da Eq. 3.3

$$\lambda_k = \frac{\left\{\left[\prod_{h=1}^{K}\left(\sum_{i\in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_h)\right)\right]\right\}^{\frac{1}{K}}}{\sum_{i\in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k)} \tag{3.3}$$

onde $\boldsymbol{\lambda} = [\lambda_k]$ representa um vetor com K pesos adaptativos do *cluster*, em que λ_k representa o peso do k-ésimo grupo. O peso do *cluster* adaptativo segue as seguintes restrições:

$$\lambda_k > 0 \qquad \text{e} \qquad \prod_{k=1}^{K} \lambda_k = 1 \tag{3.4}$$

O peso λ_k do grupo k sempre será positivo e o produto dos K pesos sempre resultará em 1. Observa-se que, para o grupo k, o denominador do peso é a dispersão intra-*cluster* da partição C_k em relação ao protótipo $\mathbf{g}_k$. O numerador é sempre constante e equivale à média geométrica das dispersão intra-*cluster* de todos os K grupos. Com isso, verifica-se que o comportamento do peso λ_k depende apenas do denominador. Assim, o valor do peso aumenta quando a respectiva dispersão intra-*cluster* diminui, ou seja, os elementos da respectiva variável estão muito próximos do protótipo do grupo. A prova de que a Eq. 3.3 minimiza o critério de adequação pode ser vista no Apêndice A.

3.3 MÉTODOS DE NUVENS DINÂMICAS COM O PESO DO *CLUSTER*

Os algoritmos de nuvens dinâmicas com o peso do *cluster* consistem em métodos de partição que visam separar as observações $\mathbf{x}_i$ em K grupos, em que cada partição C_k possua

seu peso λ_k e seu representante $\mathbf{g}_k$. O objetivo desses métodos, juntamente com seus pesos, é minimizar o critério de adequação J entre os elementos dos grupos e seus protótipos.

Nas subseções seguintes, serão apresentados os métodos de nuvens dinâmicas com o peso do *cluster* com a ponderação por meio do somatório e produtório de maneira análoga aos métodos apresentados na Seção 2.1. A diferença consiste na inserção dos pesos λ_k, que ajudam na ponderação da dispersão intra-*cluster*, como apresentado anteriormente. Esses pesos buscam solucionar o problema de uma péssima inicialização e de um mínimo local pobre.

3.3.1 Distância Não Adaptativa com o Peso do *Cluster* (Ponderação Soma)

O método de nuvens dinâmicas com distância não adaptativa e o peso do *cluster* ponderado por meio do somatório é um algoritmo iterativo que traz os seguintes passos: inicialização, representação, alocação e critério de parada. O objetivo é obter uma partição C_k com K grupos, cada grupo com seus representantes $\mathbf{g}_k$ e pesos λ_k, a fim de que a função J_4, que mede a distância entre os grupos e seus representantes, seja localmente minimizada. A função objetivo J_4 é dada a seguir:

$$J_4 = \sum_{k=1}^{K} \lambda_k^t \sum_{i \in C_k} \sum_{j=1}^{p} (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{3.5}$$

A distância $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ utilizada entre os elementos $\mathbf{x}_i \in C_k$ e o protótipo $\mathbf{g}_k$ é a *City-Block*, definida na Eq. 2.3. Neste método, a etapa de representação passa a ter duas definições: (i) os melhores protótipos e (ii) os melhores pesos do *cluster*. Para (i), a obtenção é dada por meio da Eq. 2.18, e para (ii), é dada por meio da Eq. 3.1, em que os pesos seguem a restrição dada na Eq. 3.2. Alternam-se os passos de representação e alocação até que a função J_4 atinja um valor estacionário, indicando a convergência do critério, geralmente sugerindo um mínimo local.

3.3.2 Distância Não Adaptativa com o Peso do *Cluster* (Ponderação Produto)

O método de nuvens dinâmicas com distância não adaptativa e peso de *cluster* ponderado por meio do produtório é um algoritmo iterativo composto pelas seguintes etapas: inicialização, representação, alocação e critério de parada. Seu objetivo é obter uma partição C_k com K

grupos, onde cada grupo possui seus representantes $\mathbf{g}_k$ e pesos λ_k, de modo que a função J_5, que mede a distância entre os grupos e seus representantes, seja minimizada localmente. A função objetivo J_5 é definida pela equação a seguir:

$$J_5 = \sum_{k=1}^{K} \lambda_k \sum_{i \in C_k} \sum_{j=1}^{p} (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{3.6}$$

A distância utilizada $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ entre os elementos $\mathbf{x}_i \in C_k$ e o protótipo $\mathbf{g}_k$ é a *City-Block*, definida em 2.3. Nesse algoritmo, a etapa de representação passa a ter duas definições, assim como em 3.3.1; a distinção é que os pesos são dados pela Eq. 3.3, onde os pesos seguirão a restrição apresentada em 3.4. De forma análoga, alternam-se os passos de representação e alocação até que a função J_5 atinja um valor estacionário, indicando a convergência do critério, correspondendo geralmente a um mínimo local.

3.3.3 Distância Adaptativa por Atributo com o Peso do *Cluster* (Ponderação Soma)

O método de nuvens dinâmicas com distância adaptativa por atributo com o peso do *cluster* ponderado é um algoritmo iterativo que inclui os seguintes passos: inicialização, representação, alocação e critério de parada. Seu objetivo é obter uma partição C_k com K grupos e p pesos do atributo λ_j. Cada grupo possui seus representantes $\mathbf{g}_k$ e pesos λ_k, de modo que a função J_6, que mede a distância entre os grupos e seus representantes, seja localmente minimizada. A função objetivo J_6 é dada na equação a seguir:

$$J_6 = \sum_{k=1}^{K} \lambda_k^t \sum_{i \in C_k} \sum_{j=1}^{p} \lambda^j (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{3.7}$$

A distância utilizada $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ entre os elementos $\mathbf{x}_i \in C_k$ e o protótipo $\mathbf{g}_k$ é a *City-Block* adaptativa, definida em 2.8, em que os pesos λ^j são estimados globalmente de acordo com a Eq. 2.10, seguindo as restrições dadas em 2.11. Neste algoritmo, a etapa de representação passa a ter três definições: (i) os melhores protótipos, (ii) os melhores pesos do atributo e (iii) os melhores pesos do *cluster*. Para (i), a obtenção é dada por meio da Eq. 2.18; para (ii), os melhores pesos do atributo são dados por meio da Eq. 2.11, seguindo as restrições 2.10; e para (iii), é dado por meio da Eq. 3.1, em que os pesos seguirão a restrição dada em

3.2. Alternam-se os passos de representação e alocação até que a função J_6 atinja um valor estacionário, indicando a convergência do critério, geralmente correspondendo a um mínimo local.

3.3.4 Distância Adaptativa por Atributo com o Peso do *Cluster* (Ponderação Produto)

O método de nuvens dinâmicas com distância adaptativa por atributo e peso de *cluster* ponderado por meio do produtório é um algoritmo iterativo composto pelas seguintes etapas: inicialização, representação, alocação e critério de parada. O objetivo é obter uma partição C_k com K grupos e p pesos do atributo λ_j. Cada grupo tem seus representantes $\mathbf{g}_k$ e pesos λ_k, de forma que a função J_7, que mede a distância entre os grupos e seus representantes, seja minimizada localmente. A função objetivo J_7 é definida a seguir:

$$J_7 = \sum_{k=1}^{K} \lambda_k \sum_{i \in C_k} \sum_{j=1}^{p} \lambda^j (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{3.8}$$

A distância utilizada $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ entre os elementos $\mathbf{x}_i \in C_k$ e o protótipo $\mathbf{g}_k$ é a *City-Block*, definida em 2.8. Nesse algoritmo, a etapa de representação passa a ter três definições, assim como em 3.3.1. A distinção é que os pesos são dados pela Eq. 3.3, onde os pesos seguirão a restrição apresentada em 3.4. De forma análoga, alternam-se os passos de representação e alocação até que a função J_7 atinja um valor estacionário, indicando a convergência do critério, geralmente correspondendo a um mínimo local.

3.3.5 Distância Adaptativa por Atributo e Classe com o Peso do *Cluster* (Ponderação Soma)

O método de nuvens dinâmicas com distância adaptativa por atributo e classe, com o peso do *cluster* ponderado por meio do somatório, é um algoritmo iterativo que traz os seguintes passos: inicialização, representação, alocação e critério de parada. O objetivo é obter uma partição C_k com K grupos e pesos do atributo λ_k^j, e cada grupo com seus representantes $\mathbf{g}_k$ e pesos λ_k, a fim de que a função J_8, que mede a distância entre os grupos e seus representantes, seja localmente minimizada. A função objetivo J_8 é dada na equação a seguir:

$$J_8 = \sum_{k=1}^{K} \lambda_k^t \sum_{i \in C_k} \sum_{j=1}^{p} \lambda_k^j (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{3.9}$$

A distância utilizada $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ entre os elementos $\mathbf{x}_i \in C_k$ e o protótipo $\mathbf{g}_k$ é a *City-Block* adaptativa, definida em 2.8, em que os pesos λ_k^j são estimados localmente de acordo com a Eq. 2.17, seguindo as restrições dadas em 2.16. Neste algoritmo, a etapa de representação passa a ter três definições: (i) os melhores protótipos, (ii) os melhores pesos do atributo por *cluster* e (iii) os melhores pesos do *cluster*. Para (i), a obtenção é dada por meio da Eq. 2.18; para (ii), os melhores pesos do atributo são dados por meio da Eq. 2.17, seguindo as restrições 2.16; e para (iii), é dado por meio da Eq. 3.1, em que os pesos seguirão a restrição dada em 3.2. Alternam-se os passos de representação e alocação até que a função J_8 atinja um valor estacionário, indicando a convergência do critério, geralmente correspondendo a um mínimo local.

3.3.6 Distância Adaptativa por Atributo e Classe com o Peso do *Cluster* (Ponderação Produto)

O método de nuvens dinâmicas com distância adaptativa por atributo e classe com peso do *cluster* ponderado por meio do produtório é um algoritmo iterativo composto pelas seguintes etapas: inicialização, representação, alocação e critério de parada. O objetivo é obter uma partição C_k com K grupos e pesos do atributo λ_k^j. Cada grupo tem seus representantes $\mathbf{g}_k$ e pesos λ_k, de forma que a função J_9, que mede a distância entre os grupos e seus representantes, seja minimizada localmente. A função objetivo J_9 é definida na equação a seguir:

$$J_9 = \sum_{k=1}^{K} \lambda_k \sum_{i \in C_k} \sum_{j=1}^{p} \lambda_k^j (|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|) \tag{3.10}$$

A distância utilizada $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ entre os elementos $\mathbf{x}_i \in C_k$ e o protótipo $\mathbf{g}_k$ é a *City-Block*, definida em 2.14. Nesse algoritmo, a etapa de representação passa a ter três definições, assim como em 3.3.5. A distinção é que os pesos são dados pela Eq. 3.3, onde os pesos seguirão a restrição apresentada em 3.4. De forma análoga, alternam-se os passos de representação e alocação até que a função J_9 atinja um valor estacionário, indicando a convergência do critério, geralmente correspondendo a um mínimo local.

3.4 ALGORITMO

O Algoritmo 2 apresenta as etapas que devem ser realizadas durante o agrupamento por nuvens dinâmicas com o peso do *cluster* adaptativo e ponderado, utilizando a distância não adaptativa, distância adaptativa por atributo e distância adaptativa por atributo e classe. Além disso, ajustamos o parâmetro t da Eq. 3.1 com o valor inicial de $t_{ini} = 0.01$, sendo acrescido de 0.01 a cada nova iteração, não ultrapassando um $t_{max} = 0.05$.

Algoritmo 2 Algoritmo de agrupamento por nuvens dinâmicas para dados intervalares com o uso de distância não adaptativa e adaptativas e peso do *cluster* adaptativo e ponderado

Entrada: Conjunto de dados $\mathbb{X}$, Número de clusters K, $2 \leq K < n$; Distância d_{L_1} (2.3), $d_{A^1_{L_1}}$ (2.8) ou $d_{A^2_{L_1}}$ (2.14).
Saída: Uma partição C_k que divide o conjunto de dados $\mathbb{X}$ com K classes

(1) Inicialização
Selecione uma partição $P^0 = (C^0_1, \ldots, C^0_k, \ldots, C^0_K)$ do conjunto de observações $\mathbb{X}$ ou determine K objetos diferentes $(\mathbf{g}_1, \ldots, \mathbf{g}_k, \ldots, \mathbf{g}_K)$ entre $\mathbb{X}$ e associe cada objeto $\mathbf{x}_i$ para uma classe C_{k*} tal que $k* = \arg\min d_{L_1}(\mathbf{x}_i, \mathbf{g}_k)$ para construir a partição inicial $P^0 = (C^0_1, \ldots, C^0_k, \ldots, C^0_K)$

(2) Representação
Para $k = 1$ **até** K **Faça**
(i) Determinar o protótipo $\mathbf{g}^j_k$ como na Eq. 2.18
Se a distância é adaptativa **Então**
(ii) Determinar o peso λ^j ou λ^j_k, de acordo com a Eq. 2.11 ou Eq. 2.17
Fim Se
(iii) Determinar o peso do *cluster* λ_k ponderado por meio do somatório (Eq. 3.1) ou por meio do produtório (Eq. 3.3)
Fim Para

(3) Alocação
$teste \leftarrow 0$
Para $i = 1$ **para** n **Faça**
defina a classe C_{k*} tal que

$$k* = \arg\min d(\mathbf{x}_i, \mathbf{g}_k), \text{ onde}$$
$$d(\mathbf{x}_i, \mathbf{g}_k) = d_{L_1}(\mathbf{x}_i, \mathbf{g}_k), \text{se usado a Eq. 2.3}$$
$$d(\mathbf{x}_i, \mathbf{g}_k) = d_{A^1_{L_1}}(\mathbf{x}_i, \mathbf{g}_k), \text{se usado a Eq. 2.8}$$
$$d(\mathbf{x}_i, \mathbf{g}_k) = d_{A^2_{L_1}}(\mathbf{x}_i, \mathbf{g}_k), \text{se usado a Eq. 2.14}$$

Se $i \in C_k$ **e** $k* \neq k$ **Então**
$teste \leftarrow 1$
$C_{k*} \leftarrow C_{k*} \cup \{i\}$
$C_k \leftarrow C_k \setminus \{i\}$
Fim Se
Fim Para

(4) Critério de Parada
Se $teste = 0$ **Então**
Pare
Senão
volte para ***(2)***
Fim Se

4 AVALIAÇÃO EXPERIMENTAL

Nesta fase experimental, realizamos a avaliação dos algoritmos descritos nos Capítulos 2 e 3. Na Tabela 1, pode ser observada a relação dos nove algoritmos a serem empregados, além das subseções onde cada um deles é exposto, acompanhados das respectivas abreviações para referência no decorrer do texto.

As linhas representam as distâncias que serão utilizadas: Distância Não Adaptativa (ADAN), Distância Adaptativa por Atributo (ADA1) e Distância Adaptativa por Atributo e Classe (ADA2), respectivamente. A primeira coluna representa as distâncias tradicionais da literatura Sem o Peso do *Cluster* (SP); a segunda, as distâncias combinadas com o Peso do *Cluster* Ponderado por Meio do Somatório (PS) da Seção 3.1; e a terceira, as distâncias com o Peso do *Cluster* Ponderado por Meio do Produtório (PP) da Seção 3.2.

Tabela 1 – Abreviação (subseção) dos algoritmos

	SP	**PS**	**PP**
FIXA	KM (2.1.1)	KMPS (3.3.1)	KMPP (3.3.2)
ADA1	KMS (2.1.2)	KMSPS (3.3.3)	KMSPP (3.3.4)
ADA2	KMA (2.1.3)	KMAPS (3.3.5)	KMAPP (3.3.6)

Fonte: Autor (2024)

Para isso, foram realizados experimentos com dados sintéticos (4.1) e dados reais (5). Os dados sintéticos foram divididos em dados com centros de distribuição simétrica (4.1.1) e dados com centros de distribuição assimétrica (4.1.2), que proporcionam diferentes níveis de dificuldade para a avaliação dos métodos apresentados neste estudo. As principais variações nessas configurações estão relacionadas à dispersão dos dados e à quantidade de elementos em cada classe.

Os experimentos com dados sintéticos foram realizados utilizando simulação de Monte Carlo. A simulação consistiu na realização de 50 replicações dos conjuntos de dados sintéticos, em que cada algoritmo realizou uma repetição até a convergência para um valor estacionário do critério J. A saída foi dada por um vetor com 50 índices de validação, que foram utilizados para interpretar o desempenho dos algoritmos.

A medida de qualidade do resultado do agrupamento será dada por meio do Índice de *Rand* Ajustado (IRA) e Informação Mútua Normalizada (IMN). O IRA (HUBERT; ARABIE, 1985) é

uma versão ajustada do Índice de *Rand* (IR) (RAND, 1971). Utilizado para avaliar a qualidade do agrupamento rígido, o IRA mensura a similaridade entre uma partição conhecida *a priori* e outra obtida do método de agrupamento *a posteriori*. Dadas duas partições, o IRA é calculado conforme a Eq. 4.1.

$$IRA = \frac{a + b - f}{a + b + c + d - f} \tag{4.1}$$

Em que a = número de elementos da mesma classe e mesmo grupo; b = número de elementos de diferentes classes e diferentes grupos, c = número de elementos da mesma classe, mas diferentes grupos, d = número de elementos de diferentes classes, mas do mesmo grupo; e f é dado por meio da Eq. 4.2.

$$f = \frac{(a + c)(a + d) + (b + c)(b + d)}{a + b + c + d} \tag{4.2}$$

O IRA assume valores no intervalo de $[-1, 1]$, onde 1 indica que as partições *a priori* e *a posteriori* são idênticas, enquanto valores próximos de 0 (ou negativos) indicam diferenças entre as partições. Assim, um agrupamento com bom desempenho resulta em um IRA próximo ou exatamente igual a 1.

A segunda medida de avaliação que foi utilizada é a IMN (STREHL; GHOSH, 2003), que assume valores no intervalo $[0, 1]$. Aqui, o valor 1 indica uma correlação perfeita entre as partições *a priori* e *a posteriori*. A IMN é calculada por meio da Eq. 4.3.

$$IMN = \frac{2\sum_{i=1}^{K}\sum_{j=1}^{C} P_{ij} \log\left(\frac{P_{ij}}{P_{Ki}P_{Cj}}\right)}{\left(-\sum_{i=1}^{K} P_{Ki}\log(P_{Ki})\right) + \left(-\sum_{j=1}^{C} P_{Cj}\log(P_{Cj})\right)} \tag{4.3}$$

Onde P_{ij} é a probabilidade conjunta de um objeto pertencer simultaneamente à classe a priori i e à classe de agrupamento j; P_{Ki} é a probabilidade marginal de um objeto pertencer à classe a priori i; P_{Cj} é a probabilidade marginal de um objeto pertencer à classe de agrupamento j; K é o número de classes a priori; e C é o número de grupos resultantes após o método de agrupamento.

Com os vetores resultantes da simulação de Monte Carlo, foram apresentados a média e o desvio padrão para análise estatística. As análises estatísticas foram realizadas por meio dos testes de *Friedman* e *Nemenyi*.

O teste de *Friedman* (FRIEDMAN, 1937; FRIEDMAN, 1940) é um teste não paramétrico utilizado para comparar três ou mais grupos relacionados. Seja r_i^j a classificação do j-ésimo de k algoritmos no i-ésimo de N conjuntos de dados. O teste de *Friedman* compara as classificações médias dos algoritmos, $R_j = \frac{1}{N}\sum_i r_i^j$. Sob a hipótese nula, que afirma que todos os algoritmos são equivalentes e, portanto, suas classificações R_j devem ser iguais, a estatística de Friedman

$$S = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{4.4}$$

é distribuída de acordo com χ^2 com $k-1$ graus de liberdade, quando N e k são suficientemente grandes (como regra prática, $N > 10$ e $k > 5$). Se a hipótese nula for rejeitada, podemos proceder com um teste *post-hoc*. O teste de *Nemenyi* (NEMENYI, 1963), é um teste não paramétrico, realizado após a aplicação de um teste de *Friedman* e executa o teste de comparações múltiplas. O desempenho dos métodos é significativamente diferente se as suas classificações médias diferirem em pelo menos a Diferença Crítica (DC).

$$DC = q\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{4.5}$$

Adotou-se um nível de significância $\alpha = 0.05$ para ambos os testes. Para este estudo, nos casos em que foram necessário a realização deste pós-teste, os resultados estão dispostos no Apêndice B.

4.1 DADOS INTERVALARES SINTÉTICOS

Foram consideradas oito configurações para a geração dos dados sintéticos. As configurações de 1 a 4 foram geradas a partir de centros de uma distribuição normal bivariada (4.1.1), enquanto as configurações de 5 a 8 foram geradas por meio de centros de uma distribuição exponencial (4.1.2).

Para definir os centros dos dados intervalares, foram adotadas quatro abordagens, variando a distribuição da variância entre as variáveis e entre os grupos. As características de cada configuração são as seguintes: as Configurações 1 (4.1.1.1) e 5 (4.1.2.1) possuem variâncias diferentes entre variáveis e classes; as Configurações 2 (4.1.1.2) e 6 (4.1.2.2) possuem variâncias diferentes entre as variáveis e iguais entre as classes; as Configurações 3 (4.1.1.3) e 7 (4.1.2.3) possuem variâncias iguais entre as variáveis e as classes; e, por fim, as Configurações 4 (4.1.1.4) e 8 (4.1.2.4) possuem variâncias iguais entre as variáveis e diferentes entre as classes.

Cada configuração possui uma versão balanceada, na qual os elementos dos grupos possuem a mesma quantidade, e uma versão desbalanceada, na qual há uma quantidade diferente de elementos para cada grupo. Assim, os cenários balanceados contêm um total de 300 elementos, divididos igualmente entre as três classes. O cenário desbalanceado também possui 300 elementos e três classes, porém as classes 1, 2 e 3 possuem 30, 70 e 200 elementos, respectivamente.

4.1.1 Dados com Centros de Distribuição Simétrica

Os centros foram gerados por meio de abordagens comuns na literatura (CARVALHO; LECHEVALLIER, 2009; CARVALHO; LECHEVALLIER, 2009; CARVALHO; BRITO; BOCK, 2006; CARVALHO et al., 2006a; SOUZA; CARVALHO, 2004), nos quais possuem coordenadas $z_i = (z_{x_i}, z_{y_i})$ e foram distribuídos de acordo com uma distribuição normal multivariada, com parâmetros μ e Σ, conforme apresentado abaixo:

$$\mu = \begin{bmatrix} \mu_x \\ \\ \mu_y \end{bmatrix} \quad \text{e} \quad \Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ \\ 0 & \sigma_y^2 \end{bmatrix}$$

Os parâmetros μ_x, μ_y, σ_x^2 e σ_y^2 das configurações simétricas são apresentados em cada subseção. Após a geração dos centros, para obter os intervalos, foi necessário gerar as amplitudes por meio do intervalo $[v, u]$, $u, v \in \mathrm{R}$, representado por $\gamma_1 \sim \mathcal{U}(v, u)$ e $\gamma_2 \sim \mathcal{U}(v, u)$, conforme mostrado na Eq. 4.6:

$$\left[z_{x_i} - \frac{\gamma_1}{2}, z_{x_i} + \frac{\gamma_1}{2}\right], \left[z_{y_i} - \frac{\gamma_2}{2}, z_{y_i} + \frac{\gamma_2}{2}\right] \tag{4.6}$$

Aqui, γ_1 e γ_2 representam a altura e a largura dos retângulos, respectivamente. Esses parâmetros foram selecionados aleatoriamente a partir de um mesmo intervalo predefinido. Para este experimento, os valores escolhidos foram $\gamma_1 \sim \mathcal{U}(1, 5)$ e $\gamma_2 \sim \mathcal{U}(1, 5)$. Isso significa que os dados intervalares possuem distribuições uniformes iguais para todas as classes e dimensões.

4.1.1.1 Configuração 1

Os dados da Configuração 1 possuem centros gerados a partir da distribuição normal e apresentam variâncias diferentes entre variáveis e classes. Os parâmetros para cada uma das três classes podem ser vistos abaixo:

- Classe 1: $\mu_x = 20$, $\mu_y = 30$, $\sigma_x^2 = 4$ e $\sigma_y^2 = 100$;
- Classe 2: $\mu_x = 25$, $\mu_y = 65$, $\sigma_x^2 = 121$ e $\sigma_y^2 = 9$;
- Classe 3: $\mu_x = 35$, $\mu_y = 40$, $\sigma_x^2 = 16$ e $\sigma_y^2 = 144$.

As classes estão representadas por cores distintas, facilitando a identificação dos grupos. A classe 1, 2 e 3 são definidas por centros que estão representados pelas cores azul, vermelha e verde, respectivamente. Os centros das classes balanceadas e desbalanceadas estão dispostos nas Figuras 1a e 1b. Por possuírem os parâmetros σ_x^2 e σ_y^2 diferentes, as classes possuem formato elipsoidal com volumes distintos, como ilustrado na Figura 1.

Após a etapa de geração de intervalo mostrada na Eq. 4.6, os dados intervalares balanceados e desbalanceados foram gerados e apresentados nas Figuras 1c e 1d. Neste caso, em que as variâncias são diferentes entre as variáveis e as classes, percebe-se uma semelhança com contextos mais próximos da realidade, uma vez que não é comum encontrar dados reais com classes e variáveis que possuem características semelhantes à esta configuração.

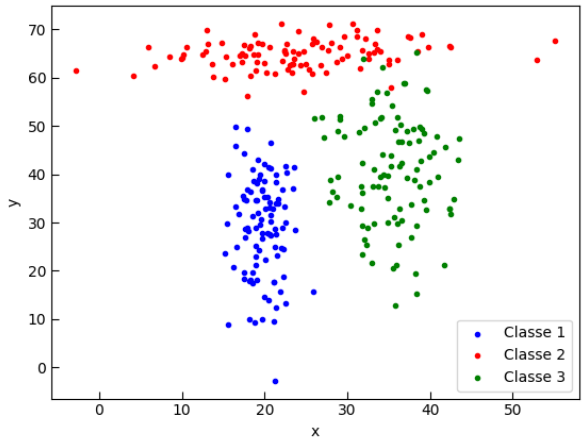
Após a geração dos dados intervalares da Configuração 1, os nove algoritmos foram avaliados utilizando simulação de Monte Carlo. Na Tabela 2 estão dispostos os valores médios e o desvio padrão dos 50 IRA e IMN, tanto para as classes balanceadas quanto para as classes desbalanceadas.

Tabela 2 – Média e desvio padrão dos vetores de IRA e IMN da Configuração 1

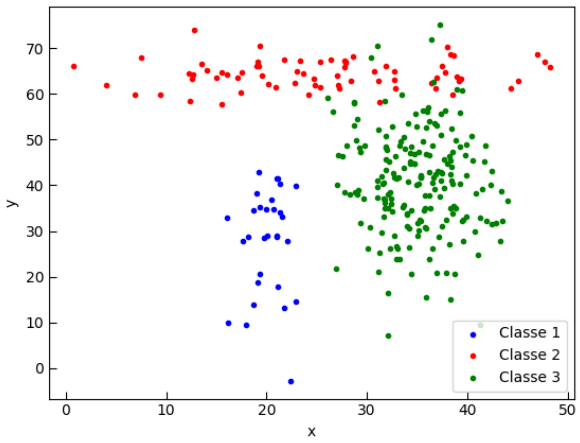
	classes balanceadas						**classes desbalaceadas**					
	SP		**PS**		**PP**		**SP**		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.7485	0.7197	**0.7530**	**0.7254**	0.7443	0.7213	0.3447	0.3997	**0.3581**	**0.4062**	0.2600	0.3545
	± 0.0947	± 0.0825	± 0.1077	± 0.0844	± 0.0689	± 0.0603	± 0.0535	± 0.0392	± 0.0819	± 0.0531	± 0.0374	± 0.0267
ADA1	0.6873	0.6873	0.6144	0.6138	**0.7160**	**0.6978**	**0.3638**	0.3638	0.3351	**0.3924**	0.2715	0.3543
	± 0.1346	± 0.1183	± 0.1156	± 0.0957	± 0.1189	± 0.1070	± 0.1000	± 0.0683	± 0.0616	± 0.0441	± 0.0618	± 0.0332
ADA2	**0.9110**	**0.8824**	0.8918	0.8640	0.8661	0.8475	0.4502	**0.4750**	**0.4553**	0.4654	0.2736	0.3615
	± 0.0290	± 0.0354	± 0.0807	± 0.0649	± 0.0261	± 0.0272	± 0.1589	± 0.1157	± 0.1844	± 0.1419	± 0.0584	± 0.0316

Fonte: Autor (2024)

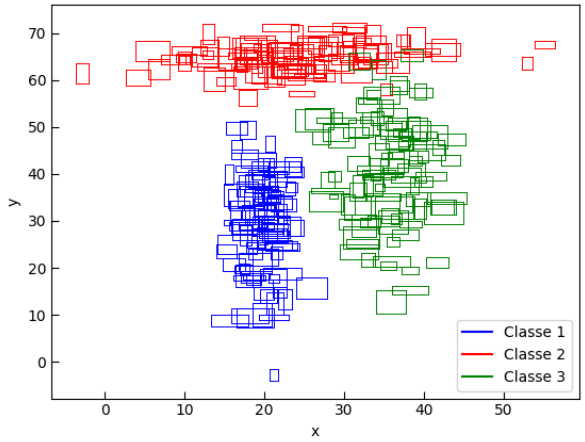
Figura 1 – Centros e Dados Intervalares da Configuração 1



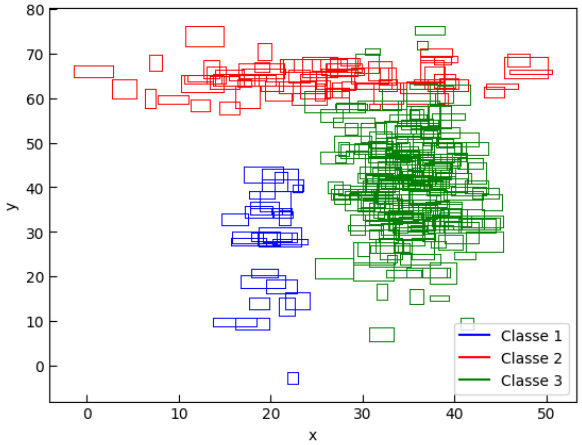
(a) Centros da Configuração 1 com classes balanceadas



(b) Centros da Configuração 1 com classes desbalanceadas



(c) Dados intervalares da Configuração 1 com classes balanceadas



(d) Dados intervalares da Configuração 1 com classes desbalanceadas

Fonte: Autor (2024)

Após realizar o teste de *Friedman*, verificou-se que todos os valores-p obtidos nos testes estatísticos para o IRA e o IMN foram menores que α. Isso fornece uma forte evidência para rejeitar H_0 e aceitar a hipótese alternativa H_1, indicando que os métodos são estatisticamente diferentes. No Apêndice B, são apresentados os testes de comparação múltipla realizados por meio do teste de *Nemenyi*.

Os resultados das classes balanceadas mostram que a ADA2 apresentou o melhor desempenho geral em termos de IRA e IMN, com valores médios significativamente mais altos em comparação com a ADAN e ADA1. Dentro dessa distância, a ponderação SP se destacou ligeiramente sobre as outras ponderações, sugerindo que a ADA2 combinada com SP é a configuração mais eficaz para classes balanceadas. No entanto, a diferença de desempenho entre

as ponderações foi menor na ADAN, onde PS teve um desempenho ligeiramente melhor.

Para as classes desbalanceadas, os resultados mostram uma queda no desempenho geral das métricas IRA e IMN em comparação com as classes balanceadas. Novamente, a ADA2 foi a mais eficaz, com as ponderações SP e PS apresentando desempenhos quase equivalentes, sendo superiores à PP. A diferença de desempenho entre a ADAN e ADA1 foi mais pronunciada, com ADA1 mostrando melhores resultados com a ponderação SP, embora ainda inferior à ADA2.

Em suma, a análise sugere que a combinação de ADA2 e ponderação SP é a mais recomendada para a Configuração 1, especialmente para conjuntos de dados balanceados. Para classes desbalanceadas, a ADA2 ainda se mantém como a melhor escolha, com uma leve preferência por SP ou PS dependendo das especificidades do conjunto de dados.

4.1.1.2 Configuração 2

Os dados da Configuração 2 são provenientes de centros gerados por uma distribuição normal, com variâncias distintas entre variáveis, mas iguais entre classes. Os parâmetros de cada classe são:

- Classe 1: $\mu_x = 20$, $\mu_y = 30$, $\sigma_x^2 = 4$ e $\sigma_y^2 = 100$;
- Classe 2: $\mu_x = 25$, $\mu_y = 65$, $\sigma_x^2 = 4$ e $\sigma_y^2 = 100$;
- Classe 3: $\mu_x = 30$, $\mu_y = 30$, $\sigma_x^2 = 4$ e $\sigma_y^2 = 100$;

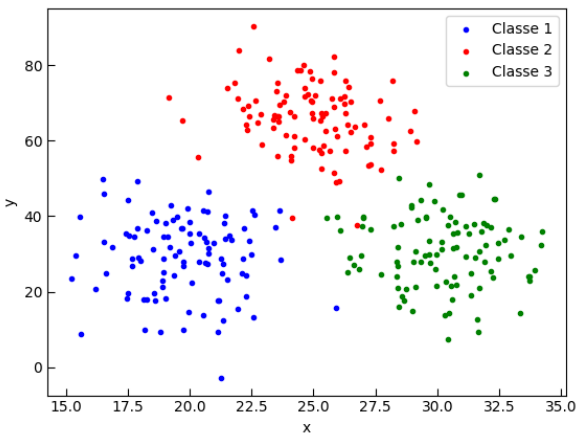
Os centros das classes balanceadas e desbalanceadas são apresentados nas Figuras 2a e 2b. Nesta configuração, as classes têm formato elipsoidal com volumes iguais, diferenciando-se apenas pela posição no plano cartesiano, como ilustrado na Figura 2.

As Figuras 2c e 2d mostram os dados intervalares balanceados e desbalanceados após a geração dos intervalos, conforme a Eq. 4.6. Esta configuração, com variâncias iguais entre classes, permite inferir sobre a performance dos algoritmos a serem avaliados.

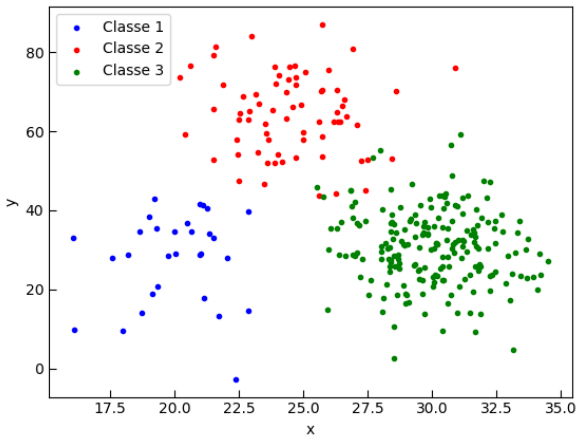
Os algoritmos foram testados com os dados intervalares da Configuração 2 utilizando simulação de Monte Carlo. A Tabela 3 apresenta as médias e desvios padrão dos 50 IRA e IMN calculados.

Após a aplicação do teste de *Friedman*, três situações para classes balanceadas apresentaram valor-p maior que α indicando que, para esses métodos, não há diferença estatística: (i) IMN com KM, KMPS e KMPP, (ii) IRA com KMA, KMAPS e KMAPP, e (iii) IMN com KMA, KMAPS e

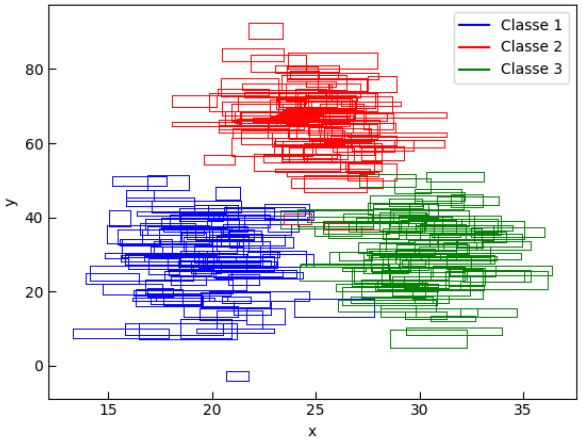
Figura 2 – Centros e Dados Intervalares da Configuração 2



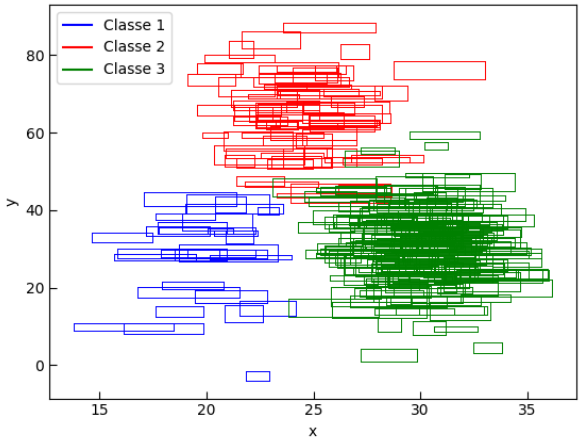
(a) Centros da Configuração 2 com classes balanceadas



(b) Centros da Configuração 2 com classes desbalanceadas



(c) Dados intervalares da Configuração 2 com classes balanceadas



(d) Dados intervalares da Configuração 2 com classes desbalanceadas

Fonte: Autor (2024)

Tabela 3 – Média e desvio padrão dos vetores de IRA e IMN da Configuração 2

	classes balanceadas						classes desbalanceadas					
	SP		PS		PP		SP		PS		PP	
	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN
FIXA	0.5220	0.5489	0.5267	0.5500	**0.5487**	**0.5505**	0.4231	0.4548	**0.4233**	**0.4565**	0.3461	0.3789
	± 0.1891	± 0.1580	± 0.1882	± 0.1577	± 0.1591	± 0.1424	± 0.1577	± 0.1237	± 0.1647	± 0.1290	± 0.0497	± 0.0424
ADA1	0.8585	0.8585	0.8644	0.8373	**0.9255**	**0.8880**	0.4590	0.4590	**0.4781**	**0.5054**	0.4412	0.4871
	± 0.1819	± 0.1489	± 0.1540	± 0.1294	± 0.0265	± 0.0358	± 0.1850	± 0.1459	± 0.1857	± 0.1435	± 0.1348	± 0.0844
ADA2	0.8065	0.7860	0.8159	0.7948	**0.8429**	**0.8170**	0.4613	0.4906	**0.4712**	**0.4939**	0.4048	0.4428
	± 0.2021	± 0.1746	± 0.1906	± 0.1654	± 0.1790	± 0.1557	± 0.1846	± 0.1488	± 0.1961	± 0.1550	± 0.0987	± 0.0687

Fonte: Autor (2024)

KMAPP. Em todos os outros testes estatísticos para a Configuração 2, a hipótese nula H_0 foi rejeitada. No Apêndice B, estão dispostos os testes de comparação múltipla obtidos por meio do teste de *Nemenyi*.

Os resultados das classes balanceadas na Configuração 2 mostram que a ADA1 apresentou os melhores desempenhos em termos de IRA e IMN, especialmente com o PP. Os valores médios de IRA e IMN para ADA1 com PP foram significativamente mais altos que os demais métodos, destacando-se como a combinação mais eficaz. A ADA2 também mostrou um bom desempenho, embora ligeiramente inferior a ADA1, com uma leve vantagem para o PP em ambas as métricas.

Para as classes desbalanceadas, a ADA1 novamente se destacou, especialmente com o PP, apresentando os melhores valores médios de IRA e IMN. Entretanto, a diferença entre ADA1 e ADA2 foi menor nesse cenário, com ADA2 mostrando desempenho competitivo, especialmente com a ponderação SP. A ADAN teve um desempenho consistentemente inferior em comparação com ADA1 e ADA2, com diferenças mais acentuadas no PP.

Em resumo, a análise para a Configuração 2 sugere que a combinação da ADA1 com o PP é a mais recomendada, independentemente do balanceamento das classes. Esta combinação mostrou-se superior nas métricas de avaliação tanto para classes balanceadas quanto desbalanceadas. A ADA2, embora não tão eficaz quanto a ADA1, também apresentou bons resultados, sendo uma opção viável especialmente para casos desbalanceados.

4.1.1.3 Configuração 3

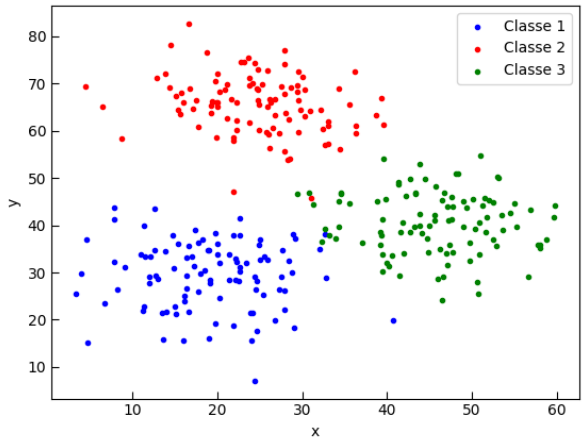
A Configuração 3 possui centros gerados através de uma distribuição normal, caracterizando-se por dados com variâncias iguais tanto a nível de variáveis quanto de classes. Os parâmetros utilizados para gerar as classes desta configuração são:

- Classe 1: $\mu_x = 20$, $\mu_y = 30$, $\sigma_x^2 = 49$ e $\sigma_y^2 = 49$;
- Classe 2: $\mu_x = 25$, $\mu_y = 65$, $\sigma_x^2 = 49$ e $\sigma_y^2 = 49$;
- Classe 3: $\mu_x = 45$, $\mu_y = 40$, $\sigma_x^2 = 49$ e $\sigma_y^2 = 49$;

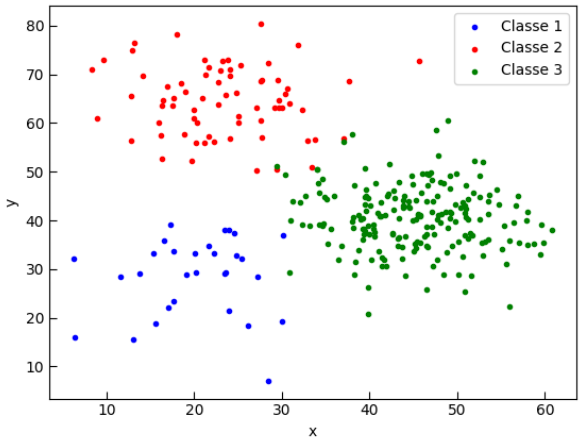
As Figuras 3a e 3b mostram os centros da Configuração 3 em suas versões balanceadas e desbalanceadas, respectivamente. Para todas as classes desta configuração, os parâmetros

σ_x^2 e σ_y^2 são iguais, indicando que esta configuração é composta por três classes de formato esférico e mesmo volume, conforme ilustrado na Figura 3.

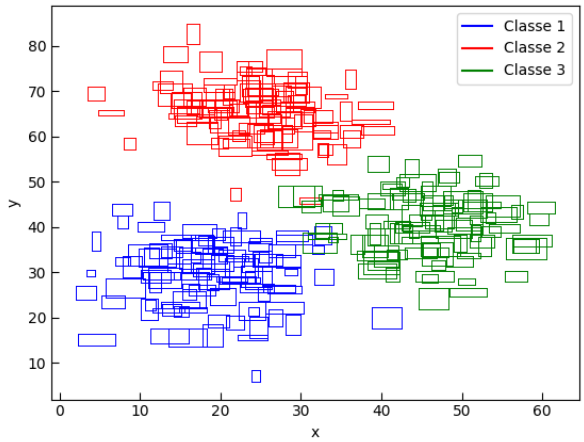
Figura 3 – Centros e Dados Intervalares da Configuração 3



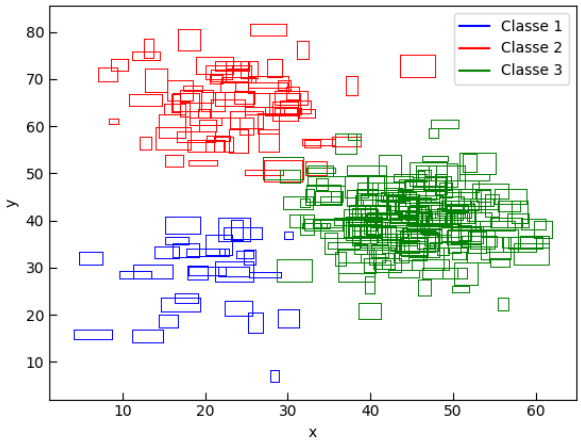
(a) Centros da Configuração 3 com classes balanceadas



(b) Centros da Configuração 3 com classes desbalanceadas



(c) Dados intervalares da Configuração 3 com classes balanceadas



(d) Dados intervalares da Configuração 3 com classes desbalanceadas

Fonte: Autor (2024)

Após o processo de geração dos intervalos por meio dos centros, conforme a Eq. 4.6, as Figuras 3c e 3d apresentam os dados intervalares da Configuração 3. Este cenário é raro em dados reais, pois raramente grupos possuem formato esférico e variâncias iguais a nível de variável e de grupo.

Os algoritmos apresentados neste estudo foram testados com os dados intervalares da Configuração 3 utilizando simulação de Monte Carlo. A Tabela 4 apresenta as médias e desvios padrão dos 50 IRA e IMN calculados.

Após a aplicação do teste de *Friedman*, houve três situações com classes balanceados em

Tabela 4 – Média e desvio padrão dos vetores de IRA e IMN da Configuração 3

	classes balanceadas						classes desbalanceadas					
	SP		PS		PP		SP		PS		PP	
	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN
FIXA	0.8881	0.8610	0.8963	0.8610	**0.9070**	**0.8682**	0.6673	0.6576	**0.6911**	**0.6754**	0.4551	0.4912
	± 0.0999	± 0.0777	± 0.0746	± 0.0624	± 0.0323	± 0.0393	± 0.1853	± 0.1247	± 0.1956	± 0.1340	± 0.1010	± 0.0641
ADA1	0.8885	0.8885	0.8954	0.8605	**0.9083**	**0.8698**	0.5865	0.5865	**0.6084**	**0.6133**	0.4277	0.4871
	± 0.1043	± 0.0851	± 0.0752	± 0.0637	± 0.0317	± 0.0394	± 0.1938	± 0.1361	± 0.2001	± 0.1429	± 0.0799	± 0.0491
ADA2	0.8866	0.8544	0.8919	0.8573	**0.9069**	**0.8686**	0.5821	0.5980	**0.6334**	**0.6278**	0.3946	0.4717
	± 0.1007	± 0.0812	± 0.0750	± 0.0652	± 0.0325	± 0.0397	± 0.1816	± 0.1212	± 0.1940	± 0.1362	± 0.0276	± 0.0293

Fonte: Autor (2024)

que a hipótese nula H_0 não foi rejeitada, indicando que não houve diferença estatística entre os métodos: (i) IRA com KM, KMPS e KMPP, (ii) IMN com KM, KMPS e KMPP, (iii) IRA com KMS, KMSPS e KMSPP, e (iv) IMN com KMS, KMSPS e KMSPP. Para os outros testes estatísticos, a hipótese nula H_0 foi rejeitada, evidenciando diferença estatística entre os métodos. Os testes de de comparação múltipla obtidos por meio do teste de *Nemenyi* são apresentados no Apêndice B.

Na Configuração 3, para classes balanceadas, os resultados mostram que todas as distâncias apresentaram desempenhos bastante próximos em termos de IRA e IMN. As ADAN, ADA1 e ADA2, combinadas com qualquer tipo de ponderação, tiveram valores médios de IRA e IMN elevados e semelhantes, indicando que qualquer uma dessas combinações pode ser eficaz. No entanto, o PP com a ADA1 apresentou uma ligeira vantagem, especialmente em termos de estabilidade dos resultados, apresentando desvio padrão menor.

Para classes desbalanceadas, os resultados mostram uma tendência similar, mas com desempenho inferior em comparação com classes balanceadas. A ADAN com o PS teve o melhor desempenho em termos de IRA, enquanto o SP com ADA1 e ADA2 apresentou um desempenho competitivo. No entanto, as variações no desempenho foram maiores, conforme indicado pelos desvios padrão mais altos, especialmente para a combinação ADA1-SP e ADA2-SP. Isso sugere que a robustez dos algoritmos é mais desafiada em cenários desbalanceados.

Em resumo, a Configuração 3 evidencia que, para classes balanceadas, a combinação de qualquer das distâncias com o PP é recomendada devido à alta consistência e desempenho. Para classes desbalanceadas, a ADAN com o PS parece ser a escolha mais robusta, mas a variabilidade nos resultados indica a necessidade de um cuidado maior na escolha do algoritmo e ponderação. A análise sugere que, embora os algoritmos se comportem de forma semelhante em cenários balanceados, os desbalanceados requerem uma avaliação mais detalhada para

assegurar a eficácia.

4.1.1.4 Configuração 4

A principal característica da Configuração 4 é ter variâncias iguais entre as variáveis, mas diferentes entre as classes. Assim como nas configurações anteriores, os centros são gerados por meio de uma distribuição normal bivariada, com os seguintes parâmetros:

- Classe 1: $\mu_x = 20$, $\mu_y = 30$, $\sigma_x^2 = 49$ e $\sigma_y^2 = 49$;
- Classe 2: $\mu_x = 25$, $\mu_y = 65$, $\sigma_x^2 = 64$ e $\sigma_y^2 = 64$;
- Classe 3: $\mu_x = 45$, $\mu_y = 40$, $\sigma_x^2 = 81$ e $\sigma_y^2 = 81$;

Os centros da Configuração 4, em suas versões balanceada e desbalanceada, são mostrados nas Figuras 4a e 4b, respectivamente. Como na configuração anterior, os parâmetros σ_x^2 e σ_y^2 são iguais entre as variáveis, resultando em configurações elipsoidais, mas com volumes diferentes, conforme ilustrado na Figura 4, devido às variâncias distintas entre as classes.

As Figuras 4c e 4d mostram os dados intervalares após o processo de geração, conforme a Eq. 4.6, nas versões balanceada e desbalanceada, respectivamente. Tal como na configuração anterior, este cenário é raro em dados reais, mas será útil para discutir a qualidade do agrupamento quando há variação nas variâncias a nível de variáveis e classes.

Os nove algoritmos apresentados neste estudo foram testados para esta configuração, e os resultados da simulação de Monte Carlo podem ser vistos na Tabela 5, que apresenta os valores médios e os desvios padrão do IRA e IMN.

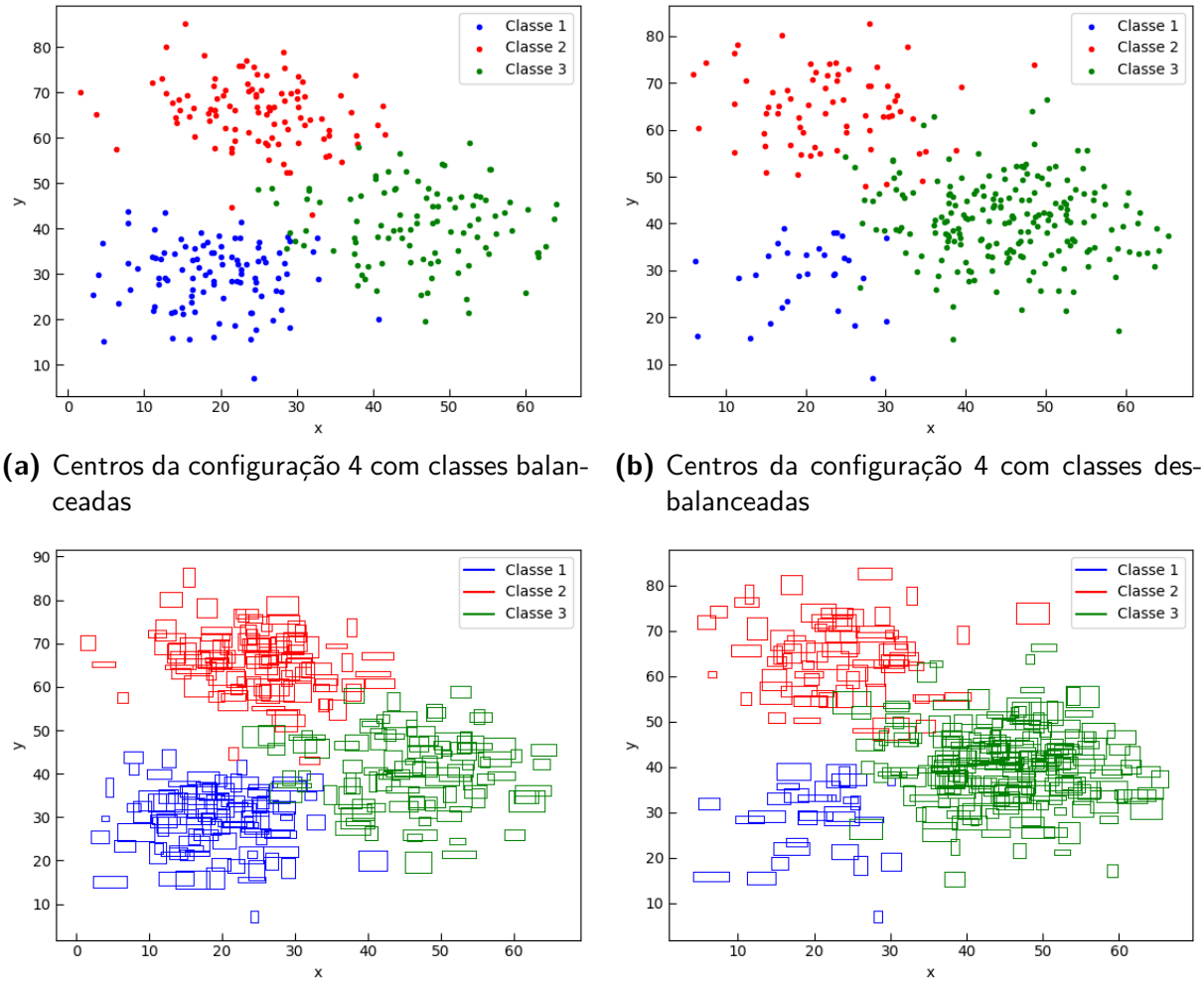
Tabela 5 – Média e desvio padrão dos vetores de IRA e IMN da configuração 4

	classes balanceadas						**classes desbalanceadas**					
	SP		**PS**		**PP**		**SP**		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.8324 ± 0.0405	0.7853 ± 0.0445	0.8275 ± 0.0440	0.7804 ± 0.0479	**0.8367** ± 0.0401	**0.7859** ± 0.0439	0.5322 ± 0.1289	0.5340 ± 0.0787	**0.5462** ± 0.1542	**0.5456** ± 0.0947	0.4023 ± 0.0799	0.4420 ± 0.0484
ADA1	**0.8321** ± 0.0406	**0.8321** ± 0.0447	0.8285 ± 0.0466	0.7811 ± 0.0497	0.8253 ± 0.0883	0.7770 ± 0.0772	0.4795 ± 0.1938	0.4795 ± 0.0899	**0.4784** ± 0.1486	**0.4943** ± 0.1000	0.3606 ± 0.0604	0.4227 ± 0.0349
ADA2	0.8279 ± 0.0402	0.7805 ± 0.0434	0.8259 ± 0.0477	0.7794 ± 0.0496	**0.8326** ± 0.0401	**0.7816** ± 0.0440	0.4699 ± 0.1157	0.4904 ± 0.0756	**0.4923** ± 0.1458	**0.5031** ± 0.0967	0.3328 ± 0.0373	0.4135 ± 0.0325

Fonte: Autor (2024)

Após a análise estatística com o teste de *Friedman*, apenas o IRA com `KM`, `KMPS` e `KMPP` com classes balanceadas apresentou valor-p abaixo de α. Para a versão desbalanceada, em

Figura 4 – Centros e Dados Intervalares da Configuração 4



(a) Centros da configuração 4 com classes balanceadas

(b) Centros da configuração 4 com classes desbalanceadas

(c) Dados intervalares da configuração 4 com classes balanceadas

(d) Dados intervalares da configuração 4 com classes desbalanceadas

Fonte: Autor (2024)

todos os testes estatísticos, a hipótese nula H_0 foi rejeitada, indicando diferenças estatísticas entre os métodos. Os testes de comparação múltipla realizados por meio do teste de *Nemenyi* são apresentados no Apêndice B.

Na Configuração 4, os resultados para classes balanceadas indicam que a ADAN com o PP obteve os valores mais altos tanto para o IRA quanto para a IMN. Isso sugere que essa combinação particular pode ser mais eficaz na formação de grupos em conjuntos de dados com classes balanceadas. Além disso, todas as combinações de distância e ponderação apresentaram desvios padrão relativamente baixos, indicando consistência nos resultados.

No entanto, para classes desbalanceadas, os resultados foram menos conclusivos. Embora a ADAN com o PS ainda tenha se destacado em termos de IRA, outras combinações,

como ADA1-PS e ADA2-PS, também demonstraram desempenho competitivo. No entanto, os desvios padrão foram significativamente maiores em comparação com classes balanceadas, sugerindo maior variabilidade nos resultados e destacando a sensibilidade dos algoritmos a conjuntos de dados desbalanceados.

Em suma, a Configuração 4 destaca a importância da escolha adequada da distância e ponderação, especialmente em conjuntos de dados desbalanceados. Enquanto a ADAN com o PP mostra consistência e desempenho robusto, outras combinações também podem ser eficazes, embora com maior variabilidade nos resultados. Isso destaca a necessidade de uma avaliação cuidadosa das características dos dados ao selecionar algoritmos de agrupamento para diferentes cenários.

4.1.2 Dados com Centros de Distribuição Assimétrica

Para os dados assimétricos, a geração dos centros é semelhante ao método utilizado para os dados simétricos. Considere um ponto p_i, onde $p_i = (p_{x_i}, p_{y_i})$ com $p_{x_i} \sim \text{Exp}(\lambda_x)$ e $p_{y_i} \sim \text{Exp}(\lambda_y)$. Aqui, λ_x e λ_y são parâmetros de taxa que serão descritos nas configurações assimétricas.

Após gerar os centros p_i, aplica-se o método descrito na Subseção 4.1.1, escolhendo γ_1 e γ_2 a partir do intervalo $[v, u]$, com $u, v \in \mathrm{R}$. Para gerar os dados intervalares nas configurações assimétricas, os parâmetros são os mesmos das configurações anteriores, a saber: $\gamma_1 \sim \mathcal{U}(1, 5)$ e $\gamma_2 \sim \mathcal{U}(1, 5)$. A geração foi realizada conforme a Eq. 4.7.

$$\left[p_{x_i} - \frac{\gamma_1}{2}, p_{x_i} + \frac{\gamma_1}{2}\right], \left[p_{y_i} - \frac{\gamma_2}{2}, p_{y_i} + \frac{\gamma_2}{2}\right] \tag{4.7}$$

Importante notar que, após a geração dos centros, as classes 1 e 3 foram deslocadas em todas as configurações assimétricas para evitar a sobreposição das três classes. Esse realocamento é feito somando-se (ou subtraindo-se) um valor ao conjunto de dados após sua geração. Para notações posteriores, dado $\lambda_j = \lambda + \omega$, λ é o parâmetro da distribuição e ω é o deslocamento utilizado para evitar a sobreposição entre classes.

4.1.2.1 Configuração 5

Esta configuração possui centros gerados a partir da distribuição exponencial. A característica principal da Configuração 5 consiste em possuir variâncias diferentes entre as variáveis e as classes. Os parâmetros para cada classe e seus respectivos deslocamentos estão dispostos a seguir.

- Classe 1: $\lambda_x = 0.5 + 3$ e $\lambda_y = 0.3 + 5$
- Classe 2: $\lambda_x = 0.7$ e $\lambda_y = 0.4$
- Classe 3: $\lambda_x = 0.6 - 3$ e $\lambda_y = 0.2 - 5$

Os centros estão dispostos nas Figuras 5a e 5b para classes balanceadas e desbalanceadas, respectivamente. Após o processo de geração de intervalo apresentado na Eq. 4.7, os dados intervalares balanceados e desbalanceados para a Configuração 5 foram gerados e apresentados nas Figuras 5c e 5d, conforme pode ser visto na Figura 5.

Os algoritmos foram avaliados nesta configuração com o auxílio da simulação de Monte Carlo. Os resultados dos agrupamentos são apresentados na Tabela 6, que dispõe dos valores médios e o desvio padrão das medidas de avaliação para cada algoritmo em suas versões balanceada e desbalanceada.

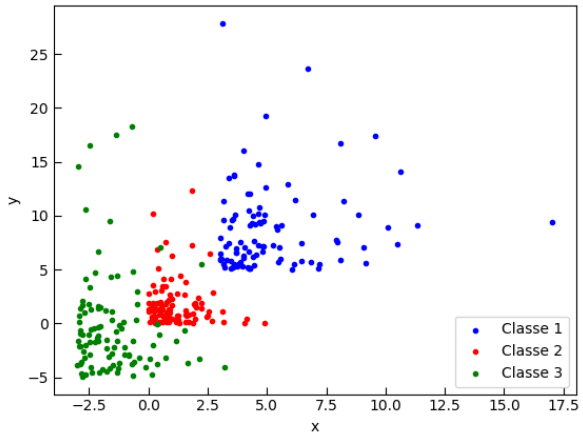
Tabela 6 – Média e desvio padrão dos vetores de IRA e IMN da Configuração 5

	classes balanceadas						**classes desbalanceadas**					
	SP		**PS**		**PP**		**SP**		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.4935	0.5048	0.4861	0.5007	**0.5541**	**0.5565**	**0.2766**	0.3324	0.2750	**0.3353**	0.2725	0.3017
	± 0.1118	± 0.1019	± 0.1096	± 0.0987	± 0.1434	± 0.1273	± 0.0524	± 0.0420	± 0.0654	± 0.0531	± 0.0465	± 0.0338
ADA1	0.5287	0.5287	0.5320	0.5472	**0.5968**	**0.5982**	**0.3226**	0.3226	0.3081	**0.3622**	0.2850	0.3311
	± 0.1320	± 0.1225	± 0.1334	± 0.1254	± 0.1636	± 0.1484	± 0.0796	± 0.0637	± 0.0913	± 0.0666	± 0.0545	± 0.0515
ADA2	0.5435	0.5591	0.5475	0.5646	**0.5855**	**0.5869**	0.3270	0.3877	**0.3306**	**0.3901**	0.2930	0.3503
	± 0.1436	± 0.1318	± 0.1495	± 0.1369	± 0.1787	± 0.1607	± 0.0816	± 0.0622	± 0.1028	± 0.0760	± 0.0570	± 0.0681

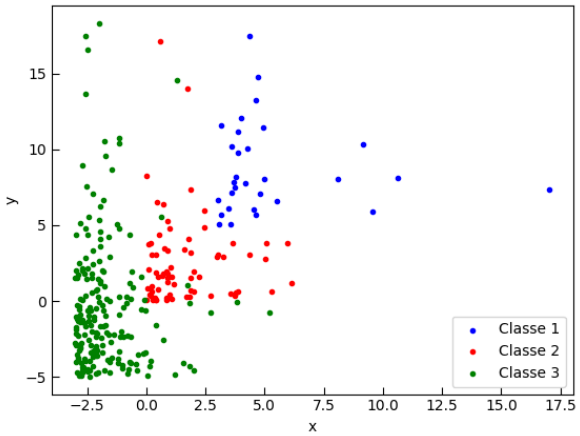
Fonte: Autor (2024)

Após a realização do teste de *Friedman*, por meio do valor-p, podemos destacar as seguintes situações em que a hipótese nula H_0 foi rejeitada: (i) o IRA das classes balanceadas para os métodos KM, KMPS e KMPP; (ii) a IMN das classes balanceadas para os métodos KM, KMPS e KMPP; (iii) a IMN das classes desbalanceadas para os métodos KM, KMPS e KMPP; (iv) o IRA das classes desbalanceadas para os métodos KMS, KMSPS e KMSPP; (v) a IMN das classes

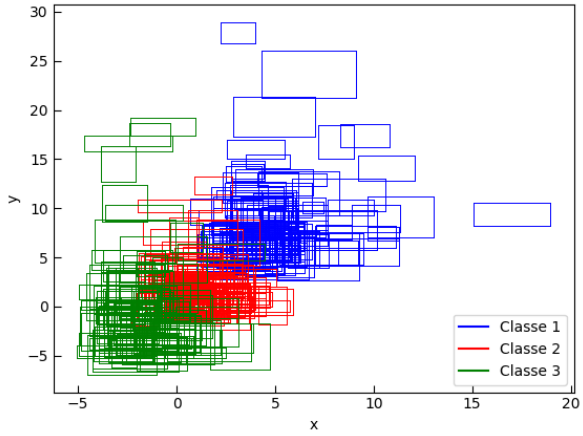
Figura 5 – Centros e Dados Intervalares da Configuração 5



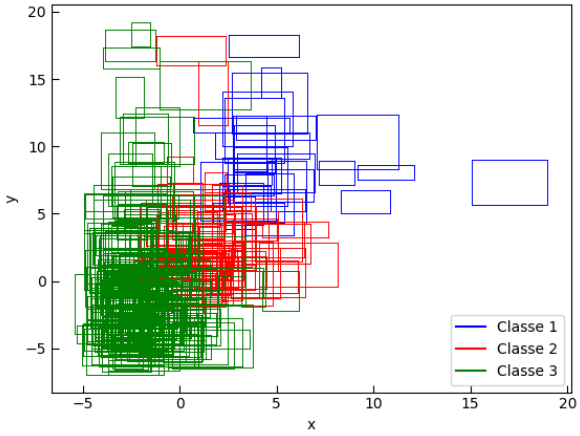
(a) Centros da Configuração 5 com classes balanceadas



(b) Centros da Configuração 5 com classes desbalanceadas



(c) Dados intervalares da Configuração 5 com classes balanceadas



(d) Dados intervalares da Configuração 5 com classes desbalanceadas

Fonte: Autor (2024)

desbalanceadas para os métodos KMS, KMSPS e KMSPP. No Apêndice B, são apresentados os testes de comparação múltipla obtidos por meio do teste de *Nemenyi*.

Na Configuração 5, os resultados para classes balanceadas mostram que a ADA1 com o PP obteve os valores mais altos tanto para o IRA e o IMN. Isso sugere que essa combinação particular pode ser mais eficaz na formação de grupos em conjuntos de dados com classes balanceadas. Além disso, o PS e o PP não apresentaram diferença significativa quando combinado com a ADA2, diferentemente das combinações ADAN-PP e ADA1-PP.

No entanto, para classes desbalanceadas, embora a ADA2 com o PS ainda tenha se destacado em termos de IRA, outras combinações, como ADA1-SP e ADA2-SP, também demonstraram desempenho competitivo. Contudo, os desvios padrão foram significativamente

menores em comparação com classes balanceadas, indicando menor variabilidade nos resultados e destacando a robustez dos algoritmos a conjuntos de dados desbalanceados.

Em resumo, na Configuração 5, para as classes balanceadas, o uso do PP proporciona um leve aumento nos valores de IRA e IMN. Em contrapartida, para as classes desbalanceadas, o PP apresenta o efeito contrário, fazendo com que, nesses casos, o uso do PS eleve levemente os valores de IRA e IMN. No entanto, a maior variabilidade nos resultados para classes balanceadas destaca a importância de uma análise cuidadosa e da escolha adequada de algoritmos de agrupamento para diferentes cenários de dados.

4.1.2.2 Configuração 6

A Configuração 6 possui centros gerados a partir da distribuição exponencial, com variâncias diferentes entre variáveis, mas iguais entre as classes. Os parâmetros das classes e seus deslocamentos são os seguintes:

- Classe 1: $\lambda_x = 0.7 + 3$ e $\lambda_y = 0.5 + 5$
- Classe 2: $\lambda_x = 0.7$ e $\lambda_y = 0.5$
- Classe 3: $\lambda_x = 0.7 - 3$ e $\lambda_y = 0.5 - 5$

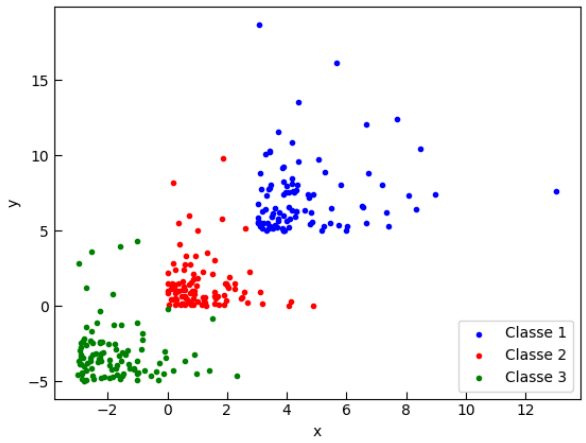
Os centros dos dados para a Configuração 6, nas versões balanceada e desbalanceada, estão nas Figuras 6a e 6b, respectivamente. Após o processo de geração de intervalo detalhado na Eq. 4.7, os dados intervalares balanceados e desbalanceados são mostrados nas Figuras 6c e 6d.

Com os dados intervalares gerados, os algoritmos foram avaliados utilizando simulação de Monte Carlo. A Tabela 7 apresenta a média e o desvio padrão do vetor de IRA e IMN para esta configuração, considerando classes balanceadas e desbalanceadas.

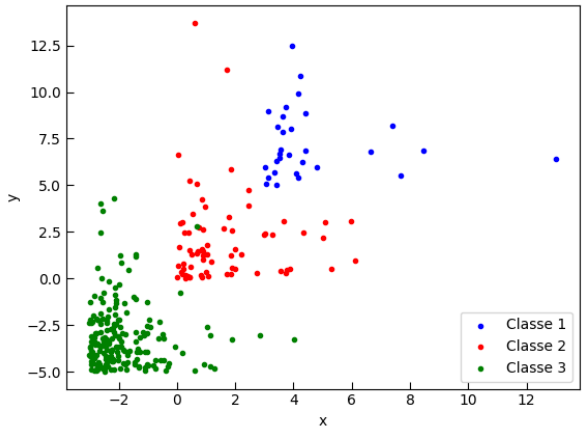
As situações em que não houve diferença estatística, conforme o teste de *Friedman*, indicam que a hipótese nula H_0 não deve ser rejeitada. Essas situações são: (i) o IRA das classes balanceadas para os métodos KMS, KMSPS e KMSPP; e (ii) o IRA das classes balanceadas para os métodos KMA, KMAPS e KMAPP. Os resultados dos testes de comparação múltipla obtidos por meio do teste de *Nemenyi*, estão dispostos no Apêndice B.

Na Configuração 6, os resultados para classes balanceadas revelam que o uso do PP resulta em valores mais altos tanto para o IRA quanto para a IMN em comparação com

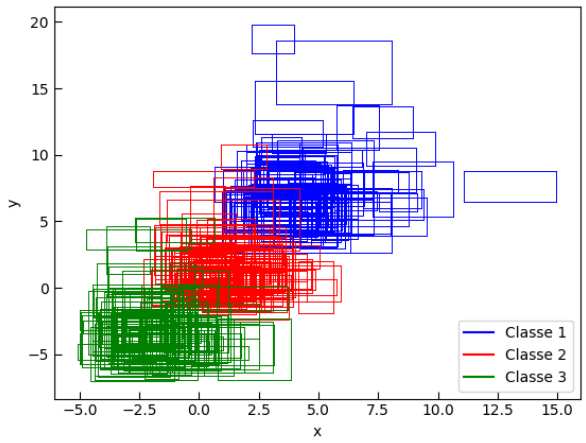
Figura 6 – Centros e Dados Intervalares da Configuração 6



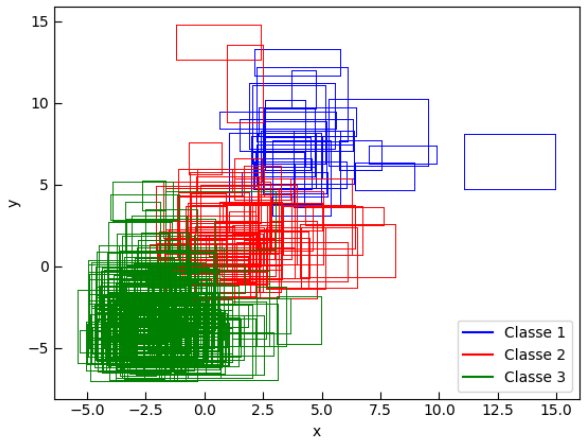
(a) Centros da Configuração 6 com classes balanceadas



(b) Centros da Configuração 6 com classes desbalanceadas



(c) Dados intervalares da Configuração 6 com classes balanceadas



(d) Dados intervalares da Configuração 6 com classes desbalanceadas

Fonte: Autor (2024)

Tabela 7 – Média e desvio padrão dos vetores de IRA e IMN da configuração 6

	classes balanceadas						classes desbalanceadas					
	SP		PS		PP		SP		PS		PP	
	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN
FIXA	0.7597	0.7532	0.7590	0.7526	**0.7926**	**0.7854**	0.5502	0.5480	**0.5782**	**0.5656**	0.4587	0.4689
	± 0.1033	± 0.0842	± 0.1059	± 0.0869	± 0.0330	± 0.0273	± 0.1467	± 0.0764	± 0.1607	± 0.0929	± 0.0968	± 0.0284
ADA1	0.7688	0.7688	0.7637	0.7568	**0.7840**	**0.7819**	0.5659	0.5659	**0.6180**	**0.5905**	0.4572	0.4686
	± 0.0896	± 0.0741	± 0.1082	± 0.0894	± 0.0694	± 0.0582	± 0.1426	± 0.0753	± 0.1445	± 0.0874	± 0.0937	± 0.0250
ADA2	0.7657	0.7583	0.7563	0.7507	**0.7868**	**0.7827**	0.5367	0.5498	**0.5649**	**0.5685**	0.4635	0.4774
	± 0.0896	± 0.0749	± 0.1083	± 0.0908	± 0.0736	± 0.0651	± 0.1315	± 0.0609	± 0.1440	± 0.0800	± 0.0852	± 0.0316

Fonte: Autor (2024)

outras combinações de distância e peso. Esse padrão sugere que o PP pode ser mais eficaz na formação de grupos em conjuntos de dados com classes balanceadas. Além disso, os desvios padrão associados a esses valores são relativamente baixos, indicando uma consistência nos resultados.

Para classes desbalanceadas, observamos que, a tendência geral ainda mostra que o uso do PS tende a fornecer os maiores valores médios para IRA quanto para a IMN. Isso sugere que, mesmo em conjuntos de dados desbalanceados, o PS pode ser uma estratégia eficaz para o agrupamento neste cenário.

Em síntese, os resultados da Configuração 6 evidenciam a relevância do peso do *cluster* na performance do agrupamento. A utilização do PP mostra como uma escolha sólida, especialmente para classes balanceadas, enquanto o uso do PS se destaca em cenários desbalanceados, proporcionando consistência nos resultados e potencialmente aprimorando a qualidade do agrupamento. No entanto, a sensibilidade aos desbalanceamento nos dados permanece uma consideração crucial ao interpretar os desempenhos obtidos no agrupamento.

4.1.2.3 Configuração 7

A Configuração 7 possui centros gerados através da distribuição exponencial, com variâncias iguais entre variáveis e classes. Os parâmetros das distribuições que geram os centros dessa configuração são os seguintes:

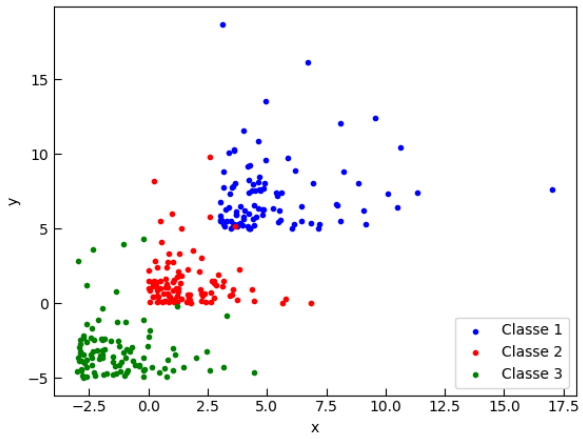
- Classe 1: $\lambda_x = 0.5 + 3$ e $\lambda_y = 0.5 + 5$
- Classe 2: $\lambda_x = 0.5$ e $\lambda_y = 0.5$
- Classe 3: $\lambda_x = 0.5 - 3$ e $\lambda_y = 0.5 - 5$

Os centros gerados podem ser vistos nas Figuras 7a e 7b para classes balanceadas e desbalanceadas, respectivamente. Após o processo descrito na Eq. 4.7 para gerar os dados intervalares, estes estão apresentados nas Figuras 7c e 7d, conforme mostrado na Figura 7.

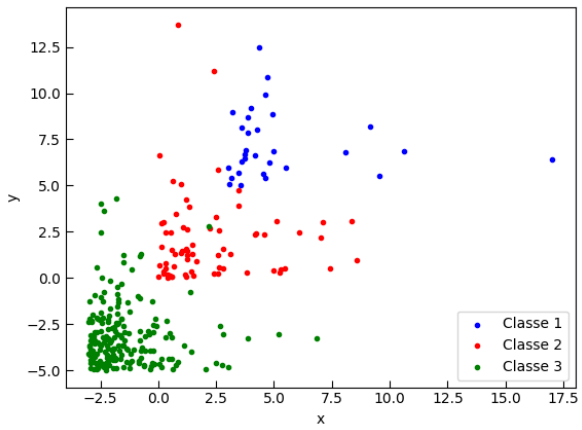
Os algoritmos foram avaliados através da simulação de Monte Carlo. A Tabela 8 apresenta a média e o desvio padrão dos índices de avaliação desses experimentos, tanto para classes balanceadas quanto para classes desbalanceadas.

O teste de *Friedman* indicou que a hipótese nula H_0 não foi rejeitada para o IRA das classes balanceadas entre os métodos KM, KMPS e KMPP. Para as demais situações, a hipótese

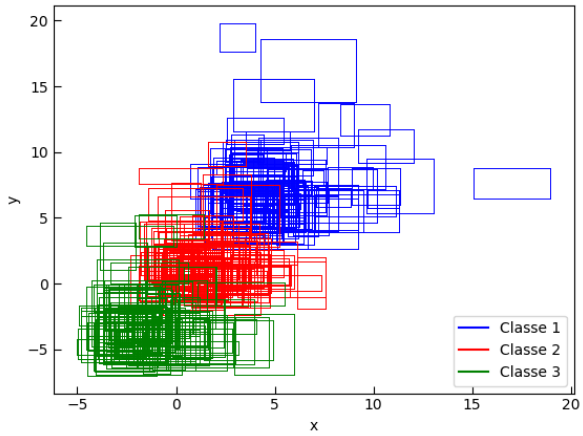
Figura 7 – Centros e Dados Intervalares da Configuração 7



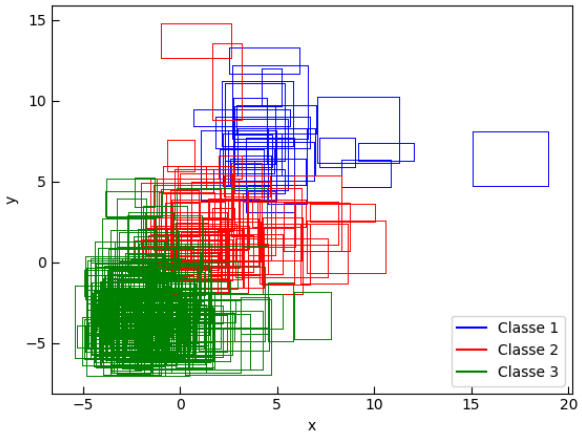
(a) Centros da Configuração 7 com classes balanceadas



(b) Centros da Configuração 7 com classes desbalanceadas



(c) Dados intervalares da Configuração 7 com classes balanceadas



(d) Dados intervalares da Configuração 7 com classes desbalanceadas

Fonte: Autor (2024)

Tabela 8 – Média e desvio padrão dos vetores de IRA e IMN da Configuração 7

	classes balanceadas						classes desbalanceadas					
	SP		PS		PP		SP		PS		PP	
	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN
FIXA	0.7256	0.7243	0.7333	0.7317	**0.7544**	**0.7539**	0.5401	0.5294	**0.5811**	**0.5593**	0.4277	0.4444
	± 0.1115	± 0.0962	± 0.0927	± 0.0774	± 0.0676	± 0.0580	± 0.1377	± 0.0762	± 0.1429	± 0.0872	± 0.0811	± 0.0228
ADA1	0.7156	0.7156	0.7266	0.7252	**0.7457**	**0.7446**	0.5553	0.5553	**0.5701**	**0.5487**	0.4215	0.4406
	± 0.1244	± 0.1084	± 0.1009	± 0.0885	± 0.0921	± 0.0797	± 0.1262	± 0.0762	± 0.1261	± 0.0728	± 0.0719	± 0.0293
ADA2	0.7259	0.7243	0.7236	0.7234	**0.7514**	**0.7488**	0.5332	0.5295	**0.5539**	**0.5490**	0.4353	0.4535
	± 0.1008	± 0.0881	± 0.0996	± 0.0861	± 0.0708	± 0.0611	± 0.1277	± 0.0601	± 0.1457	± 0.0806	± 0.0677	± 0.0354

Fonte: Autor (2024)

nula H_0 foi rejeitada, evidenciando diferença estatística entre os métodos. No Apêndice B, são apresentados os testes de comparação múltipla realizados por meio do teste de *Nemenyi*.

Na Configuração 7, os resultados para classes balanceadas mostram que o uso do PP e PS demonstraram consistentemente valores mais altos de IRA e IMN em comparação com o SP. Essa tendência segue quando usado as ADAN, ADA1 e ADA2. A consideração do peso do *cluster* pode contribuir significativamente para a qualidade dos grupos formados neste cenário.

No caso das classes desbalanceadas, observa-se uma tendência semelhante com o PS, e uma piora da ponderação PP em relação ao SP. Isso sugere que a consideração do PS pode ser especialmente benéfica para conjuntos de dados desbalanceados, ajudando a melhorar a qualidade dos grupos formados.

Portanto, na Configuração 7, tanto a ADAN quanto as ADA1 e ADA2 mostram-se adequadas para a tarefa de agrupamento, tanto para conjuntos de dados balanceados quanto desbalanceados. Esse resultado indica que, embora o PP possa ser eficaz para conjuntos de dados balanceados, pode não ser a melhor escolha para dados desbalanceados.

4.1.2.4 Configuração 8

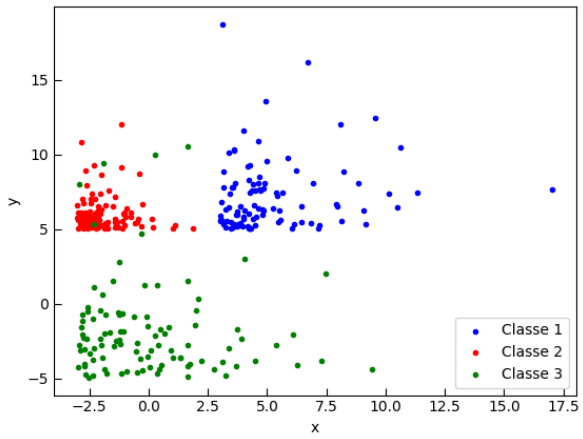
A Configuração 8 possui centros gerados a partir da distribuição exponencial. A característica dessa configuração consiste em ter variâncias iguais entre variáveis e classes. Os parâmetros para cada classe e seus respectivos deslocamentos são:

- Classe 1: $\lambda_x = 0.5 + 3$ e $\lambda_y = 0.5 + 5$
- Classe 2: $\lambda_x = 0.7 - 3$ e $\lambda_y = 0.7 + 5$
- Classe 3: $\lambda_x = 0.6 - 3$ e $\lambda_y = 0.6 - 5$

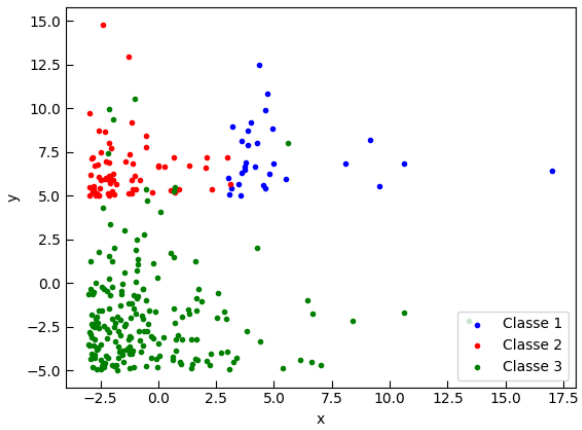
Os centros dos dados intervalares estão mostrados nas Figuras 8a e 8b para classes balanceadas e desbalanceadas, respectivamente. Após o processo de geração de intervalos descrito na Eq. 4.7, os dados intervalares balanceados e desbalanceados para esta configuração foram gerados e estão apresentados nas Figuras 8c e 8d, conforme mostrado na Figura 8.

Os algoritmos foram avaliados utilizando simulação de Monte Carlo, e os resultados dos agrupamentos são apresentados na Tabela 9, mostrando os valores médios e o desvio padrão das medidas de avaliação para cada algoritmo em suas versões balanceada e desbalanceada.

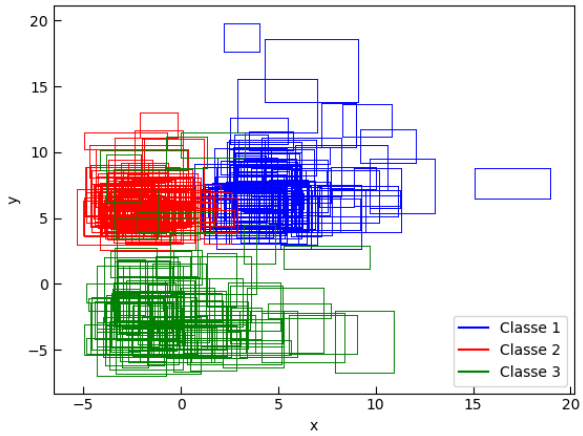
Figura 8 – Centros e Dados Intervalares da Configuração 8



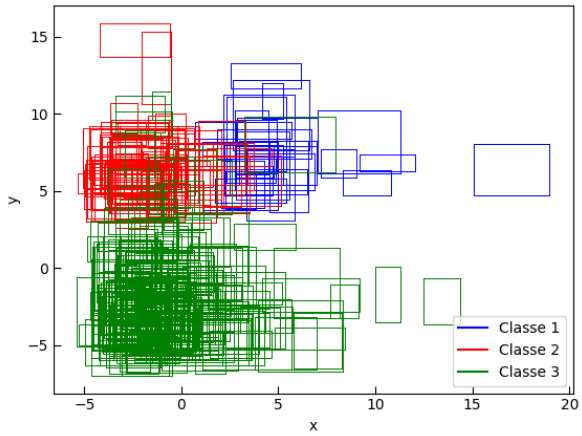
(a) Centros da Configuração 8 com classes balanceadas



(b) Centros da Configuração 8 com classes desbalanceadas



(c) Dados intervalares da Configuração 8 com classes balanceadas



(d) Dados intervalares da Configuração 8 com classes desbalanceadas

Fonte: Autor (2024)

Tabela 9 – Média e desvio padrão dos vetores de IRA e IMN da Configuração 8

	classes balanceadas						classes desbalanceadas					
	SP		PS		PP		SP		PS		PP	
	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN	IRA	IMN
FIXA	0.6824	0.6811	0.6759	0.6764	**0.8126**	**0.7833**	0.4714	0.4880	**0.4782**	**0.4923**	0.3944	0.4387
	± 0.1986	± 0.1512	± 0.2023	± 0.1541	± 0.1056	± 0.0917	± 0.1488	± 0.0710	± 0.1580	± 0.0815	± 0.1023	± 0.0394
ADA1	0.6730	0.6730	0.6706	0.6712	**0.7430**	**0.7331**	0.4761	0.4761	**0.4803**	**0.4917**	0.3642	0.4235
	± 0.2053	± 0.1580	± 0.2052	± 0.1582	± 0.1872	± 0.1435	± 0.1565	± 0.0833	± 0.1567	± 0.0850	± 0.0768	± 0.0327
ADA2	0.6831	0.6757	0.6743	0.6679	**0.7724**	**0.7489**	**0.4555**	**0.4706**	0.4079	0.4464	0.4122	0.4420
	± 0.1638	± 0.1300	± 0.1540	± 0.1234	± 0.1547	± 0.1313	± 0.1106	± 0.0480	± 0.1150	± 0.0614	± 0.1097	± 0.0414

Fonte: Autor (2024)

A análise estatística por meio do teste de *Friedman* mostrou que a hipótese nula H_0 não foi rejeitada para o IRA das classes desbalanceadas entre os métodos KMA, KMAPS e KMAPP. Para as outras situações, houve rejeição da hipótese nula H_0, evidenciando diferença estatística entre os métodos. Os testes de comparação múltipla, realizados por meio do teste de *Nemenyi*, estão apresentados no Apêndice B.

Na Configuração 8, os resultados para classes balanceadas indicam que as ADAN, ADA1 e ADA2 apresentam desempenho semelhante em relação ao IRA e à IMN quando comparadas com as abordagens SP e PS. Os valores médios de IRA e IMN são superiores para essas três distâncias quando utilizado o PP, sugerindo que a combinação das distâncias com este peso é superior às abordagens SP e PS.

Já para as classes desbalanceadas, observa-se uma tendência semelhante, com os métodos ADAN e ADA1 apresentando resultados comparáveis às abordagens SP e PS, mas com uma queda de desempenho ao utilizar o ponderamento PP. No caso da ADA2, a inserção dos pesos PS e PP apresentou piores resultados, tanto para os valores de IRA quanto para a IMN.

Assim, na Configuração 8, tanto a ADAN quanto as distâncias ADA1 e ADA2 utilizando o PP mostram-se adequadas para a tarefa de agrupamento para as classes balanceadas. Diferentemente dos dados desbalanceados, em que o uso do PP provoca uma queda de desempenho, mostram-se melhores resultados com o uso do PS. Isso sugere que a ponderação PP pode ser menos adequada para lidar com desbalanceamento nesta configuração.

Na Tabela 10, pode-se ver um resumo de todos os algoritmos e configurações sintéticas. O asterisco * indica qual método obteve o melhor desempenho, definido pela distância e comparado ao SP, ao PS e ao PP. De acordo com a Tabela 1, é possível rever a nomenclatura dos algoritmos para melhor entendimento destes resultados.

Há fortes evidências de que, para configurações balanceadas, a ADA2 juntamente com o PP é o mais indicado, pois fornece as melhores métricas de acordo com os experimentos. Para configurações desbalanceadas, as evidências indicam que a ADA1 se destaca juntamente com o PS, apresentando os melhores resultados.

Cabe destacar também que, em grande parte dos casos, a inserção do PP para configurações desbalanceadas, independentemente da distância utilizada, causa uma queda nas medidas de avaliação do agrupamento, não sendo recomendado para uso. Por outro lado, em grande parte das configurações, as distâncias combinadas SP e PS mostram resultados aproximados, com destaque, como mencionado, para o uso do PS.

Tabela 10 – Síntese dos Resultados do Agrupamento para os Dados Sintéticos

	configurações balanceadas								**configurações desbalanceadas**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KM													*			
KMPS	*								*	*	*	*		*	*	*
KMPP		*	*	*	*	*	*	*								
KMS				*					*			*	*			
KMSPS										*	*			*	*	*
KMSPP	*	*	*		*	*	*	*								
KMA	*															*
KMAPS									*	*	*	*	*	*	*	
KMAPP		*	*	*	*	*	*	*								

Fonte: Autor (2024)

5 EXPERIMENTOS COM DADOS INTERVALARES REAIS

A avaliação experimental dos algoritmos apresentados neste estudo foi ampliada para incluir conjuntos de dados reais. Os métodos de agrupamento por nuvens dinâmicas foram aplicados a esses dados com apenas uma repetição, até a convergência para um valor estacionário do critério J. Em seguida, as medidas de avaliação IRA e IMN foram calculadas e apresentadas nas tabelas ao longo do texto.

Foram utilizados nove conjuntos de dados reais: cogumelos do gênero Amanita (**AMA**), Carros (**CAR**), Clima da Europa Ocidental (**CEO**), Climas mistos (**CLM**), Fases de Gesto (**GES**), Peixes (**PEI**), Reconhecimento de atividade Humana por Dados de Aceleração (**ACE**), Sementes (**SEM**) e Temperatura (**TEM**).

Tabela 11 – Informações dos Conjuntos de Dados Reais

Conjunto de dados	n	p	K	B/D
AMA	23	3	3	D
CAR	33	8	4	D
CEO	324	4	2	D
CLM	280	4	3	D
GES	200	9	5	B
PEI	12	13	4	D
ACE	60	3	4	B
SEM	45	2	3	B
TEM	37	12	4	D

Fonte: Autor (2024)

A Tabela 11 trás detalhadamente informações relacionadas aos conjuntos de dados, como: número de observações, número de variáveis intervalares, números de grupos em que as observações estão divididas, bem como se os conjunto de dado é balanceado ou desbalanceado. Nas seções seguintes pode ser visto com mais detalhe sobre os conjuntos de dados, bem como os resultados dos experimentos.

5.1 CONJUNTO AMANITA

O conjunto de dados Amanita (DESJARDIN; WOOD; STEVENS, 2015) contém informações extraídas do Índice de Espécies de Fungos da Califórnia de espécies de cogumelos que compõem

o gênero Amanita. As informações contidas são: o diâmetro do píleo, o comprimento e a espessura da estipe, totalizando três variáveis intervalares. A título de curiosidade, o píleo é conhecido como o chapéu do cogumelo e a estipe seria o caule do cogumelo.

A rotulação *a priori* destes dados estão relacionados com sua comestibilidade com 1, 2 e 3 para as espécies desconhecidas, comestíveis e não comestíveis, respectivamente. A Tabela 12 apresenta as métricas obtidas por meio dos algoritmos apresentados neste estudo.

Tabela 12 – Resultados do IRA e a IMN para o conjunto Amanita

	SP		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	-0.0028	0.1230	**0.0434**	**0.1603**	-0.0110	0.0790
ADA1	**0.1441**	0.2368	0.0612	0.1685	0.1284	**0.2752**
ADA2	0.0612	0.1685	0.0612	0.1685	**0.2258**	**0.3351**

Fonte: Autor (2024)

Os resultados da tabela revelam que, para o conjunto de dados Amanita, o uso do PS resultou em um desempenho superior em comparação com o método SP e PP. Isso sugere que, para este conjunto de dados específico, a inserção de peso do *cluster* resultou em melhorias significativas no desempenho dos algoritmos de nuvens dinâmicas em relação ao método SP.

5.2 CONJUNTO CARROS

O conjunto de dados Carros (DIDAY; NOIRHOMME-FRAITURE, 2008) apresenta 33 observações que listam modelos de carros com 8 variáveis simbólicas intervalares, 2 variáveis categóricas e uma variável nominal. Para estes experimentos serão considerados apenas as variáveis simbólicas intervalares, que são: Preço, Cilindrada, Velocidade Máxima, Aceleração, Distância entre os Eixos, Comprimento, Largura e Altura.

Para a rotulação *a priori* a variável nominal Categoria foi utilizada. Com isto, cada observação foi rotulada como Utilitário, Bernila, Luxuoso e Esportivo. A Tabela 13 mostra os valores obtidos para o IRA e a IMN.

Os resultados para o conjunto de dados Carro mostram que os métodos de nuvens dinâmicas tiveram um desempenho consistente em termos de IRA e IMN. As medidas de avaliação foram semelhantes entre o SP e PS para todas as configurações de distância. No entanto,

Tabela 13 – Resultados do IRA e a IMN para o conjunto Carro

	SP		PS		PP	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.3921	0.5086	**0.3921**	**0.5086**	0.3782	0.4718
ADA1	0.6142	0.7173	**0.6142**	**0.7173**	0.4815	0.6105
ADA2	0.6142	0.7173	0.6142	**0.7173**	**0.6333**	0.6974

Fonte: Autor (2024)

para a ADA2, o PP teve um desempenho ligeiramente superior, sugerindo que a inserção do PP pode levar a resultados melhores para esse conjunto.

5.3 CONJUNTO CLIMA EUROPA OCIDENTAL

O conjunto Clima Europa Ocidental (FILHO; SOUZA, 2013) é composto por 324 observações do clima de cidade em vários países. O conjunto possui 16 variáveis simbólicas intervalares que são as temperaturas mínimas e máximas ao longo dos 12 meses do ano, juntamente com os mínimos e máximos de precipitação das 4 estações do ano. A rotulação *a priori* é separada em climas mediterrâneo e oceânico. Na Tabela 14 é possível verificar os valores obtidos para o IRA e a IMN.

Tabela 14 – Resultados do IRA e a IMN para o conjunto Clima Europa Ocidental

	SP		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	**0.7430**	**0.6152**	0.7326	0.6036	0.7026	0.6079
ADA1	0.7538	0.6279	**0.7538**	**0.6279**	0.7018	0.5803
ADA2	0.7531	0.6227	**0.7531**	**0.6227**	0.6518	0.5664

Fonte: Autor (2024)

Os resultados para o conjunto de dados Clima Europa Ocidental apontam uma tendência consistente no desempenho dos métodos de nuvens dinâmicas em diferentes configurações de distância. Em geral, o PS e o PP em comparação SP apresentam resultados semelhantes em termos de IRA e IMN. No entanto, para todas as configurações de distância, o método PP tende a ter valores ligeiramente inferiores em comparação com SP e PS. Essa discrepância sugere que, para o conjunto de dados Clima Europa Ocidental, a adição de PP pode não fornecer benefícios significativos em relação às distancias SP ou com o PS.

5.4 CONJUNTO CLIMA MISTOS

O conjunto Clima Mistos (FILHO; SOUZA, 2013) assim como o conjunto anterior possui observações do clima de cidade em alguns países. Possuindo 280 observações e as mesmas 16 variáveis do conjunto de dados clima europa ocidental, este conjunto de dados possui suas rotulações *a priori* em 3 climas: savana, equatorial e sub-ártico. Através da Tabela 15 podemos verificar os valores obtidos para os índices de avaliação após a aplicação dos métodos de nuvens dinâmicas.

Tabela 15 – Resultados do IRA e a IMN para o conjunto Clima Mistos

	SP		PS		PP	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.4012	0.5039	0.3804	0.4592	**0.4150**	**0.5242**
ADA1	0.3993	0.5029	0.3917	0.4740	**0.4133**	**0.5232**
ADA2	0.3949	0.4951	0.3809	0.4640	**0.4162**	**0.5278**

Fonte: Autor (2024)

Para o conjunto de dados Clima Mistos, os resultados revelam padrões semelhantes aos observados em conjuntos anteriores. Os métodos de nuvens dinâmicas, tanto com o PS e o PP, quanto SP, mostram desempenhos comparáveis em termos IRA e IMN. No entanto, o PP teve um desempenho ligeiramente superior em comparação SP e o PS. Isso sugere que, para o conjunto de dados Clima Mistos, a inserção do PP pode melhorar a qualidade dos agrupamentos em comparação com outros métodos.

5.5 CONJUNTO FASE DE GESTOS

O conjunto de dados de fase de gestos (MADEO; LIMA; PERES, 2013) foi obtido por fases (escanso; brusco; retração; preparação; e, no aguardo) extaídas de vídeos em que as pessoas gesticulam. Após o pré-processamento de dados para redução e geração de dados simbólicos intervalares foram selecionadas as variáveis para o conjunto de dados.

Para a representação das fases as 9 variáveis foram: posição da mão esquerda em x (lhx); posição da mão esquerda em y (lhy); posição da mão esquerda em z (lhz); posição da mão direita em x (rhx); posição da mão direita em y (rhy); posição da mão direita em z (rhz); posição da cabeça em x (hx); posição da cabeça em y (hy); e, posição da cabeça em z (hz).

Este conjunto de dados está rotulado *a priori* de acordo com as 5 fases extraídas dos vídeos. A Tabela 16 apresenta os valores do IRA e da IMN para o conjunto de fases de gestos.

Os resultados para o conjunto de dados Fase de Gestos revelam que os métodos de nuvens dinâmicas com o PS e PP em comparação com algoritmos SP demonstram desempenhos relativamente similares em termos de IRA e IMN. No entanto, uma tendência interessante é observada na configuração ADAN, onde métodos SP e com o PS apresentam resultados idênticos para ambas as métricas, enquanto o método PP registra valores ligeiramente inferiores. Essa discrepância sugere que, para o conjunto de dados Fase de Gestos, o uso de PP pode

Tabela 16 – Resultados do IRA e a IMN para o conjunto Fase de Gestos

	SP		PS		PP	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.0295	0.0663	**0.0295**	**0.0663**	0.0221	0.0582
ADA1	0.0263	0.0649	**0.0263**	**0.0649**	0.0257	0.0640
ADA2	0.0248	0.0591	0.0169	0.0581	**0.0347**	**0.0850**

Fonte: Autor (2024)

não ser tão eficaz quanto outras abordagens de atribuição de peso, como o PS.

5.6 CONJUNTO PEIXES

O conjunto de dados simbólicos intervalares Peixes provém de estudos ecotoxicológicos que foi descoberto altos níveis de contaminação de mercúrio em algumas regiões devido ao consumo de peixe de água doce contaminado (BOUDOU; RIBEYRE, 1997). Para compreender tais fenômenos, pesquisadores do *Laboratoire d'Ecophysiologie et d'Ecotoxicologie des Systèmes Aquatiques* (LEESA) coletaram informações para compor este conjunto de dados.

Possuindo 12 observações compreendidas por espécies de peixes e descritos por 13 variáveis simbólicas intervalares, tais como: Comprimento, Peso, Músculo, Intestino, Estômago, Guelras, Fígado, Rins, Fígado/Músculo, Rins/Músculo, Guelras/Músculo, Intestino/Músculo, Estômago/Músculo e uma variável categórica Dieta, este conjunto de dado possui sua rotulação *a priori* dada por meio desta variável categórica. Com isso, estas espécies estão classificadas em 4 classes, que são carnívoros, detritívoros, omnívoros e herbívoros.

Na Tabela 17 podemos verificar os valores obtidos para o IRA e a IMN após a aplicação dos métodos de nuvens dinâmicas descritos neste estudo.

Tabela 17 – Resultados do IRA e a IMN para o conjunto Peixes

	SP		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.2087	0.5746	**0.2087**	**0.5746**	-0.0161	0.3760
ADA1	0.4880	0.7239	**0.4880**	**0.7239**	0.3018	0.6068
ADA2	**0.4880**	**0.7239**	0.2490	0.5825	0.2747	0.5984

Fonte: Autor (2024)

Os resultados dos experimentos indicam que a utilização do PS e PP não melhorou a qualidade das medidas de avaliação em comparação com os métodos SP. Em particular, a adição do PS manteve as mesmas avaliações que SP para a ADAN e ADA1, mas piorou significativamente para ADA2. Já o PP mostrou um impacto negativo nas avaliações em todos os casos analisados, com reduções mais acentuadas nos valores de IRA e IMN. Portanto, com base nos dados apresentados, o uso do peso do *cluster* não proporcionou melhorias e, em alguns casos, degradou a qualidade do agrupamento.

5.7 CONJUNTO DADOS DE ACELERAÇÃO

Para conjunto de Reconhecimento de Atividade Humana por Dados de Aceleração (CASALE; PUJOL; RADEVA, 2011) foram coletados dados por meio de um acelerômetro posto no tórax dos participantes. Após isto houve a identificação de padrões de locomoção de participantes que realizaram algumas atividades em que para cada participante e atividade os valores mínimos e máximos de aceleração foram determinados, formando assim os dados intervalares.

Na observação de duração de cada atividade é obtida uma sequência de valores relativos às acelerações nos eixos x, y e z, que determinam as variáveis do conjunto. Aos 60 indivíduos, para os dados *a priori* foram atribuidos os seguintes rótulos: parado; andando; andando e falando com alguém; e, parado e falando com alguém. A Tabela 18 apresenta as métricas obtidas por meio dos algoritmos apresentados neste estudo.

Tabela 18 – Resultados do IRA e a IMN para o conjunto Aceleração

	SP		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.1003	0.2098	0.1058	0.2150	**0.1236**	**0.2093**
ADA1	0.0993	0.2086	0.0993	**0.2086**	**0.1238**	0.2058
ADA2	0.0838	0.1747	0.0838	0.1747	**0.1004**	**0.1902**

Fonte: Autor (2024)

Para o conjunto de dados Aceleração, tanto algoritmos SP quanto com o PS e PP têm resultados semelhantes em todas as configurações de distância, sugerindo que o peso do *cluster* pode não ser crucial para melhorar a precisão dos agrupamentos neste conjunto de dados. No entanto, o PP tende a ter resultados ligeiramente melhores do que PS em algumas configurações, especialmente em ADAN. Isso sugere que, para este conjunto de dados específico, o uso do PP pode levar a resultados ligeiramente superiores em comparação com o PS.

5.8 CONJUNTO SEMENTES

O conjunto Sementes (LICHMAN, 2013) provém de dados pontuais das propriedades métricas dos grãos pertecentes a três diferentes variedades de trigo, que são Kama, Rosa e Canadian. As informações disponíveis são: área; perímetro; compacticidade; comprimento do trigo; largura do trigo; coeficiente de assimetria; e, tamanho do encaixe do trigo.

Para a construção dos dados intervalares foram escolhidas as variáveis comprimento e largura do trigo e realizado um pré-processamento dos dados (SOUZA, 2016) em que o conjunto resultante há 45 observações divididas em 3 classes *a priori*, que são Kama, Rosa e Canadian. A Tabela 19 mostra os valores obtidos para o IRA e a IMN após a aplicação dos métodos de nuvens dinâmicas.

Tabela 19 – Resultados do IRA e a IMN para o conjunto Sementes

	SP		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.4045	0.5255	**0.6016**	**0.6551**	0.5631	0.5941
ADA1	0.4045	0.5255	0.4045	0.5255	**0.5631**	**0.5941**
ADA2	0.5624	0.6305	**0.6016**	**0.6551**	0.5631	0.5941

Fonte: Autor (2024)

Os resultados para o conjunto de dados Sementes mostram que o uso do PS e PP teve um impacto positivo na maioria dos casos analisados, em comparação com os métodos SP. Com a ADAN, tanto o PS quanto o PP resultaram em melhorias significativas nas medidas de avaliação. Para a ADA1, o uso do PP também apresentou melhorias em ambas as métricas, enquanto o PS não teve impacto. No caso da ADA2, o PS melhorou ambas as métricas, enquanto o PP melhorou o IRA, mas teve uma pequena queda no IMN. Portanto, os PS e PP mostraram uma tendência geral de melhorar a qualidade do agrupamento para o conjunto de dados Sementes, especialmente com a ADAN e ADA2.

5.9 CONJUNTO TEMPERATURA

O conjunto de dados Temperatura das Cidades (GURU; KIRANAGI; NAGABHUSHAN, 2004) mostra os mínimos e máximos das temperaturas de 37 cidades em que a classificação *a priori* foi dada por um grupo de pesquisadores por meio de 4 classes. A Tabela 20 apresenta o IRA e o IMN, ambas obtidas por meio dos resultados dos métodos de agupamento por nuvens dinâmicas apresentados neste estudo.

Os resultados para o conjunto de dados Temperatura indicam que a utilização do peso do *cluster* teve um impacto variado dependendo do tipo de ponderação e da distância utilizada. Para a ADAN, o uso do PP melhorou significativamente ambas as medidas de avaliação em comparação com o uso SP. No entanto, o PS não teve nenhum impacto, mantendo os

Tabela 20 – Resultados do IRA e a IMN para o conjunto Temperatura

	SP		**PS**		**PP**	
	IRA	**IMN**	**IRA**	**IMN**	**IRA**	**IMN**
FIXA	0.4001	0.4635	0.4001	0.4635	**0.4665**	**0.5305**
ADA1	0.4635	0.5293	**0.4635**	**0.5293**	0.4635	0.5293
ADA2	0.4001	0.4635	0.4001	0.4635	**0.4635**	**0.5293**

Fonte: Autor (2024)

mesmos valores das métricas observadas no cenário SP. Para as distâncias ADA1 e ADA2, o uso do PS não apresentou melhoria significativa, mantiveram os mesmos valores observados no cenário SP, apenas o uso PP em ADA2 resultou em aumento das medidas de avaliação. Em resumo, o PP demonstrou um benefício específico na ADAN, enquanto os outros cenários não apresentaram mudanças significativas com a aplicação de pesos.

Na Tabela 21, para os dados reais balanceados, assim como nos sintéticos o peso do *cluster* ponderado por meio do produtório mostra um desempenho satisfatório, como no conjunto **ACE** em que os três melhores índices são por meio desta ponderação. Para os dados desbalaçeados a ponderação por meio do somatório mostra-se como destaque para alcance de melhores índices de agrupamento.

Tabela 21 – Síntese dos Resultados do Agrupamento para os Dados Reais

	balanceados			**desbalanceados**					
	GES	ACE	SEM	AMA	CAR	CEO	CLM	PEI	TEM
KM						*			
KMPS	*		*	*	*			*	
KMPP		*					*		*
KMS				*					
KMSPS	*				*	*		*	*
KMSPP		*	*				*		
KMA								*	
KMAPS			*			*			
KMAPP	*	*		*	*		*		*

Fonte: Autor (2024)

Por meio dos experimentos é possíveel inferir que o peso do *cluster* indica melhora na

inicialização de dados reais, e como mostrado nas tabelas anteriores, casos em que as ponderações não são superiores aos métodos já existentes na literatura, não causa uma queda. Estas observações evidenciam que os métodos aqui apresentados são satisfatórios em praticamente todos os casos.

6 CONSIDERAÇÕES FINAIS

Neste estudo, foram propostos novos métodos de nuvens dinâmicas com o peso do *cluster* para dados simbólicos intervalares, utilizando a distância *City-Block*. Este peso é calculado iterativamente durante as etapas do agrupamento. O peso do *cluster* é calculado por meio da ponderação do somatório e do produtório. O peso do *cluster* ponderado pelo somatório é diretamente proporcional à sua dispersão intra-*cluster*, enquanto o peso do *cluster* ponderado pelo produtório é inversamente proporcional à sua dispersão intra-*cluster*.

Os métodos propostos e desenvolvidos neste estudo, baseados em extensões da literatura, evidenciaram que o uso deste peso minimiza problemas recorrentes de agrupamento, como mínimos locais pobres e inicialização ruim. Essas evidências foram verificadas através de testes com dados sintéticos balanceados, desbalanceados e dados reais, comparando os resultados com métodos existentes na literatura e os novos métodos propostos.

Como sugestão para trabalhos futuros, os métodos aqui apresentados podem ser estendidos com a utilização da distância Euclidiana ou *Hausdorff*, que estão entre as mais comuns na literatura. Há também a possibilidade de estender esses métodos utilizando a abordagem difusa.

Por fim, as principais contribuições deste estudo foram propor métodos para minimizar problemas recorrentes de agrupamento, a extensão e desenvolvimento de novos métodos para dados simbólicos intervalares considerando a ponderação por grupo, e comprovar a eficácia dessa ponderação através de experimentos.

REFERÊNCIAS

AYESHA, S.; MUSTAFA, T.; SATTAR, A. R.; KHAN, M. I. Data mining model for higher education system. *European Journal of Scientific Research*, v. 43, n. 1, p. 24–29, 2010.

BILLARD, L. Symbolic data analysis: what is it? In: SPRINGER. *Compstat 2006-Proceedings in Computational Statistics: 17th Symposium Held in Rome, Italy, 2006*. [S.l.], 2006. p. 261–269.

BILLARD, L.; DIDAY, E. From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association*, Informa UK Limited, v. 98, n. 462, p. 470–487, jun. 2003. ISSN 1537-274X. Disponível em: <http://dx.doi.org/10.1198/016214503000242>.

BILLARD, L.; DIDAY, E. *Clustering Methodology for Symbolic Data*. Wiley, 2019. ISBN 9781119010401. Disponível em: <http://dx.doi.org/10.1002/9781119010401>.

BLASHFIELD, R. K.; ALDENDERFER, M. S. The literature on cluster analysis. *Multivariate Behavioral Research*, Informa UK Limited, v. 13, n. 3, p. 271–295, jul. 1978. ISSN 1532-7906. Disponível em: <http://dx.doi.org/10.1207/s15327906mbr1303_2>.

BOCK, H.-H.; DIDAY, E. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. [S.l.]: Springer Science & Business Media, 2012.

BOUDOU, A.; RIBEYRE, F. Mercury in the food web: accumulation and transfer mechanisms. *Metal ions in biological systems*, MARCEL DEKKER AG, v. 34, p. 289–320, 1997.

CARVALHO, F. d. A. D.; LECHEVALLIER, Y. Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition*, Elsevier BV, v. 42, n. 7, p. 1223–1236, jul. 2009. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2008.11.016>.

CARVALHO, F. d. A. de. Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recognition Letters*, Elsevier, v. 28, n. 4, p. 423–437, 2007.

CARVALHO, F. d. A. de; SOUZA, R. M. de; CHAVENT, M.; LECHEVALLIER, Y. Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, Elsevier BV, v. 27, n. 3, p. 167–179, fev. 2006. ISSN 0167-8655. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2005.08.014>.

CARVALHO, F. d. A. de; SOUZA, R. M. de; CHAVENT, M.; LECHEVALLIER, Y. Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, Elsevier BV, v. 27, n. 3, p. 167–179, fev. 2006. ISSN 0167-8655. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2005.08.014>.

CARVALHO, F. d. A. T. de; BRITO, P.; BOCK, H.-H. Dynamic clustering for interval data based on l 2 distance. *Computational Statistics*, Springer Science and Business Media LLC, v. 21, n. 2, p. 231–250, jun. 2006. ISSN 1613-9658. Disponível em: <http://dx.doi.org/10.1007/s00180-006-0261-z>.

CARVALHO, F. d. A. T. de; LECHEVALLIER, Y. Dynamic clustering of interval-valued data based on adaptive quadratic distances. *IEEE Transactions on Systems, Man, and*

Cybernetics - Part A: Systems and Humans, Institute of Electrical and Electronics Engineers (IEEE), v. 39, n. 6, p. 1295–1306, nov. 2009. ISSN 1558-2426. Disponível em: <http://dx.doi.org/10.1109/TSMCA.2009.2030167>.

CASALE, P.; PUJOL, O.; RADEVA, P. Human activity recognition from accelerometer data using a wearable device. In: _____. *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2011. p. 289–296. ISBN 9783642212574. Disponível em: <http://dx.doi.org/10.1007/978-3-642-21257-4_36>.

CELEBI, M. E.; KINGRAVI, H. A.; VELA, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, Elsevier BV, v. 40, n. 1, p. 200–210, jan. 2013. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2012.07.021>.

CELEUX, G.; DIDAY, E.; GOVAERT, G.; LECHEVALLIER, Y.; RALAMBONDRAINY, H. *Classification automatique des données*. [S.l.]: Dunod Paris, 1989.

CHARU, C. A.; CHANDAN, K. R. Data clustering: algorithms and applications. *An Introduction to Toxicogenomics*, 2013.

CHAVENT, M.; LECHEVALLIER, Y. Dynamical clustering of interval data: Optimization of an adequacy criterion based on hausdorff distance. *Jajuga, K., Sokołowski, A., Bock, HH. (eds) Classification, Clustering, and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization.*, 2002.

DESJARDIN, D. E.; WOOD, M. G.; STEVENS, F. A. *California mushrooms: The comprehensive identification guide*. [S.l.]: Timber Press, 2015.

DIDAY, E. Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de statistique appliquée*, v. 19, n. 2, p. 19–33, 1971.

DIDAY, E. Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de Statistique Appliquée*, Société de Statistique de France, v. 19, n. 2, p. 19–33, 1971. Disponível em: <http://www.numdam.org/item/RSA_1971__19_2_19_0/>.

DIDAY, E. Thinking by classes in data science: the symbolic data analysis paradigm. *WIREs Computational Statistics*, Wiley, v. 8, n. 5, p. 172–205, ago. 2016. ISSN 1939-0068. Disponível em: <http://dx.doi.org/10.1002/wics.1384>.

DIDAY, E.; GOVAERT, G. Automatic classification with adaptive intervals.[classification automatique avec distances adaptatives.]. *RAIRO Inf Comput Sci*, v. 11, p. 329–349, 01 1977.

DIDAY, E.; NOIRHOMME-FRAITURE, M. *Symbolic Data Analysis and the SODAS Software*. Wiley, 2008. ISBN 9780470723562. Disponível em: <http://dx.doi.org/10.1002/9780470723562>.

DIDAY, E.; SIMON, J. Clustering analysis. *Digital Pattern Recognition*, Springer Berlin Heidelberg, p. 47–94, 1976.

DIDAY, E.; SIMON, J. C. Clustering analysis. In: _____. *Digital Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1976. p. 47–94. ISBN 978-3-642-96303-2. Disponível em: <https://doi.org/10.1007/978-3-642-96303-2_3>.

DING, K.; LIU, Z.; WANG, M. Cluster analysis: application of k-means algorithm to explore the sea surface temperature and salinity patterns. In: SPIE. *International Conference on Computer Application and Information Security (ICCAIS 2023)*. [S.l.], 2024. v. 13090, p. 643–654.

FILHO, T. d. M. e S.; SOUZA, R. M. Fuzzy learning vector quantization approaches for interval data. In: *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2013. Disponível em: <http://dx.doi.org/10.1109/FUZZ-IEEE.2013.6622424>.

FILIPPONE, M.; CAMASTRA, F.; MASULLI, F.; ROVETTA, S. A survey of kernel and spectral methods for clustering. *Pattern Recognition*, Elsevier BV, v. 41, n. 1, p. 176–190, jan. 2008. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2007.05.018>.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, Informa UK Limited, v. 32, n. 200, p. 675–701, dez. 1937. ISSN 1537-274X. Disponível em: <http://dx.doi.org/10.1080/01621459.1937.10503522>.

FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 11, n. 1, p. 86–92, mar. 1940. ISSN 0003-4851. Disponível em: <http://dx.doi.org/10.1214/aoms/1177731944>.

GIROLAMI, M. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, Institute of Electrical and Electronics Engineers (IEEE), v. 13, n. 3, p. 780–784, maio 2002. ISSN 1045-9227. Disponível em: <http://dx.doi.org/10.1109/TNN.2002.1000150>.

GURU, D.; KIRANAGI, B. B.; NAGABHUSHAN, P. Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. *Pattern Recognition Letters*, Elsevier BV, v. 25, n. 10, p. 1203–1213, jul. 2004. ISSN 0167-8655. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2004.03.016>.

HARDER, A. A.; OLBRICHT, G. R.; EKUMA, G.; HIER, D. B.; OBAFEMI-AJAYI, T. Multiple imputation for robust cluster analysis to address missingness in medical data. *IEEE Access*, IEEE, v. 12, p. 42974–42991, 2024.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, Springer Science and Business Media LLC, v. 2, n. 1, p. 193–218, dez. 1985. ISSN 1432-1343. Disponível em: <http://dx.doi.org/10.1007/BF01908075>.

INOKUCHI, R.; MIYAMOTO, S. Lvq clustering and som using a kernel function. In: IEEE. *2004 IEEE international conference on fuzzy systems (IEEE cat. no. 04CH37542)*. [S.l.], 2004. v. 3, p. 1497–1500.

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, Elsevier, v. 31, n. 8, p. 651–666, 2010.

LICHMAN, M. *UCI machine learning repository*. 2013.

LIU, J.; GUO, Y.; LI, D.; WANG, Z.; XU, Y. Kernel-based minmax clustering methods with kernelization of the metric and auto-tuning hyper-parameters. *Neurocomputing*, Elsevier BV, v. 359, p. 173–184, set. 2019. ISSN 0925-2312. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2019.05.056>.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MADEO, R. C. B.; LIMA, C. A. M.; PERES, S. M. Gesture unit segmentation using support vector machines: segmenting gestures from rest positions. In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. ACM, 2013. (SAC '13). Disponível em: <http://dx.doi.org/10.1145/2480362.2480373>.

NEMENYI, P. *Distribution-free Multiple Comparisons*. Princeton University, 1963. Disponível em: <https://books.google.com.br/books?id=nhDMtgAACAAJ>.

OSKOUEI, A. G.; BALAFAR, M. A.; MOTAMED, C. Fkmawcw: Categorical fuzzy k-modes clustering with automated attribute-weight and cluster-weight learning. *Chaos, Solitons amp; Fractals*, Elsevier BV, v. 153, p. 111494, dez. 2021. ISSN 0960-0779. Disponível em: <http://dx.doi.org/10.1016/j.chaos.2021.111494>.

PAL, N.; PAL, K.; KELLER, J.; BEZDEK, J. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, Institute of Electrical and Electronics Engineers (IEEE), v. 13, n. 4, p. 517–530, ago. 2005. ISSN 1063-6706. Disponível em: <http://dx.doi.org/10.1109/TFUZZ.2004.840099>.

PANGESTU, C.; SHAUFIAH, S.; WIJAYA, R. X spotify cares clustering analysis using k-means and k-medoids. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, v. 8, n. 1, p. 497–507, 2024.

PIMENTEL, B. A.; COSTA, A. F. B. F. da; SOUZA, R. M. C. R. de. Kernel-based fuzzy clustering of interval data. In: *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*. IEEE, 2011. Disponível em: <http://dx.doi.org/10.1109/FUZZY.2011.6007336>.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, Informa UK Limited, v. 66, n. 336, p. 846–850, dez. 1971. ISSN 1537-274X. Disponível em: <http://dx.doi.org/10.1080/01621459.1971.10482356>.

REDDY, C. K.; VINZAMURI, B. A survey of partitional and hierarchical clustering algorithms. In: _____. *Data Clustering*. Chapman and Hall/CRC, 2018. p. 87–110. ISBN 9781315373515. Disponível em: <http://dx.doi.org/10.1201/9781315373515-4>.

RODRíGUEZ, S. I. R.; CARVALHO, F. d. A. Tenório de. Clustering interval-valued data with adaptive euclidean and city-block distances. *Expert Systems with Applications*, Elsevier BV, v. 198, p. 116774, jul. 2022. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2022.116774>.

SEYALA, N.; ABDULLAH, S. N. Cluster analysis on longitudinal data of patients with kidney dialysis using a smoothing cubic b-spline model. *International Journal of Mathematics, Statistics, and Computer Science*, v. 2, p. 85–95, 2024.

SILVA, W. J.; SOUZA, P. J.; SOUZA, R. M.; CYSNEIROS, F. J. A. A clustering algorithm for polygonal data applied to scientific journal profiles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2023.

SILVA, W. J. F.; SOUZA, R. M. C. R.; CYSNEIROS, F. J. A. psda: A tool for extracting knowledge from symbolic data with an application in brazilian educational data. *Soft Computing*, Springer-Verlag, Berlin, Heidelberg, v. 25, n. 3, p. 1803–1819, fev. 2021. ISSN 1432-7643. Disponível em: <https://doi.org/10.1007/s00500-020-05252-5>.

SOUZA, L. C. d. Agrupamento e regressão linear de dados simbólicos intervalares baseados em novas representações. Universidade Federal de Pernambuco, 2016.

SOUZA, R. M. de; CARVALHO, F. d. A. de. Clustering of interval data based on city–block distances. *Pattern Recognition Letters*, Elsevier BV, v. 25, n. 3, p. 353–365, fev. 2004. ISSN 0167-8655. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2003.10.016>.

STREHL, A.; GHOSH, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, v. 3, n. Dec, p. 583–617, 2003.

TZORTZIS, G.; LIKAS, A. The minmax k-means clustering algorithm. *Pattern Recognition*, Elsevier BV, v. 47, n. 7, p. 2505–2516, jul. 2014. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2014.01.015>.

WANG, L. Data mining, machine learning and big data analytics. *International Transaction of Electrical and Computer Engineers System*, Science and Education Publishing Co., Ltd., v. 4, n. 2, p. 55–61, ago. 2017. ISSN 2373-1273. Disponível em: <http://dx.doi.org/10.12691/iteces-4-2-2>.

XU, R.; WUNSCHII, D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, Institute of Electrical and Electronics Engineers (IEEE), v. 16, n. 3, p. 645–678, maio 2005. ISSN 1045-9227. Disponível em: <http://dx.doi.org/10.1109/TNN.2005.845141>.

YANG, M.-S.; WU, K.-L. Unsupervised possibilistic clustering. *Pattern Recognition*, Elsevier BV, v. 39, n. 1, p. 5–21, jan. 2006. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2005.07.005>.

ZHANG, D.-Q.; CHEN, S.-C. Kernel-based fuzzy and possibilistic c-means clustering. In: *Proceedings of the International Conference Artificial Neural Network*. [S.l.: s.n.], 2003. v. 122, p. 122–125.

APÊNDICE A – PROPOSIÇÕES

Proposição A.1. *Seja G fixo, $\boldsymbol{\lambda}$ é um mínimo local estrito de $J(\boldsymbol{\lambda})$ se e somente se $\boldsymbol{\lambda}$ for calculado por meio da Eq. 3.1.*

Prova. Aplicamos a técnica dos multiplicadores de Lagrange, semelhante à Oskouei, Balafar e Motamed (2021), para resolver o seguinte problema de minimização com restrição (ver Eq. A.1):

$$\mathcal{L}(\boldsymbol{\lambda}, \psi) = \sum_{k=1}^{K} \lambda_k^t \sum_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k) - \psi \left(\sum_{k=1}^{K} \lambda_k - 1 \right), \tag{A.1}$$

onde ψ é o multiplicador de Lagrange associado à restrição $\sum_{k=1}^{K} \lambda_k = 1$, e $d_\phi(\mathbf{x}_i, \mathbf{g}_k)$ é uma métrica de distância entre $\mathbf{x}_i$ e $\mathbf{g}_k$.

Ao definir o gradiente $\nabla \mathcal{L}(\boldsymbol{\lambda}, \psi) = 0$ em respeito a ψ e λ_k, obtemos as Eqs. A.2 e A.3.

$$\frac{\partial \mathcal{L}}{\partial \psi} = - \left(\sum_{k=1}^{K} \lambda_k - 1 \right) = 0, \tag{A.2}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k} = t \lambda_k^{t-1} \sum_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k) - \psi = 0, \tag{A.3}$$

Por meio da Eq. A.2 é dada a condição de restrição, em que a soma dos pesos λ_k devem ser igual a 1.

$$\sum_{k=1}^{K} \lambda_k = 1. \tag{A.4}$$

A Eq. A.3 implica na Eq. A.5

$$\psi = t \lambda_k^{t-1} \sum_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k). \tag{A.5}$$

Da Eq. A.5, obtemos a Eq. A.6

$$\lambda_k = \left(\frac{\psi}{t \sum_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k)} \right)^{\frac{1}{t-1}} \tag{A.6}$$

Substituindo a Eq. A.6 na Eq. A.4, temos a Eq. A.7.

$$\sum_{k=1}^{K} \left(\frac{\psi}{t \sum_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k)} \right)^{\frac{1}{t-1}} = 1. \tag{A.7}$$

Por meio da Eq. A.7, obtemos a Eq. A.8.

$$\psi = \frac{t}{\left[\sum_{k=1}^{K} \left(\frac{1}{\sum_{i \in C_k} d_\phi(\mathbf{x}_i, \mathbf{g}_k)} \right)^{\frac{1}{t-1}} \right]^{t-1}}. \tag{A.8}$$

Substituindo a Eq. A.8 na Eq. A.6 obtemos a Eq. 3.1. Isto completa a prova.

□

Proposição A.2. *Seja G fixo, $\boldsymbol{\lambda}$ é um mínimo local estrito de $J(\boldsymbol{\lambda})$ se e somente se $\boldsymbol{\lambda}$ for calculado por meio da Eq. 3.3.*

Prova. A prova da Proposição A.2 é o mesmo que a prova da Proposição A.1. □

APÊNDICE B – VALOR-P DO TESTE DE *NEMENYI*

Tabela 22 – Valor-p do teste de *Nemenyi* para a Configuração 1

IRA - ADAN		
	KM	KMPS
KMPS	0.7030	-
KMPP	0.0065	0.0003

(a) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados balanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.0033	-
KMSPP	0.9347	0.0104

(b) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados balanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.0190	-
KMAPP	0.0000	0.0000

(c) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados balanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.3952	-
KMPP	0.3153	0.0164

(d) Valor-p do teste de *Nemenyi* para o IMN utilizando ADAN para dados balanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.0089	-
KMSPP	0.9347	0.0253

(e) Valor-p do teste de *Nemenyi* para o IMN utilizando ADA1 para dados balanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.0104	-
KMAPP	0.0000	0.0023

(f) Valor-p do teste de *Nemenyi* para o IMN utilizando ADA2 para dados desbalanceados

IRA - ADAN		
	KM	KMPS
KMPS	0.6403	-
KMPP	0.0000	0.0000

(g) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados desbalanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.8201	-
KMSPP	0.0000	0.0000

(h) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados desbalanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.6403	-
KMAPP	0.0000	0.0000

(i) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados desbalanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.9156	-
KMPP	0.0000	0.0000

(j) Valor-p do teste de *Nemenyi* para o IMN utilizando ADAN para dados desbalanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.9516	-
KMSPP	0.0000	0.0000

(k) Valor-p do teste de *Nemenyi* para o IMN utilizando ADA1 para dados desbalanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.8713	-
KMAPP	0.0000	0.0000

(l) Valor-p do teste de *Nemenyi* para o IMN utilizando ADA2 para dados desbalanceados

Fonte: Autor (2024)

Tabela 23 – Valor-p do teste de *Nemenyi* para a Configuração 2

IRA - ADAN		
	KM	KMPS
KMPS	0.7030	-
KMPP	0.1122	0.0141

(a) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados balanceado

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.2456	-
KMSPP	0.1386	0.0014

(b) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados balanceado

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.5141	-
KMSPP	0.1122	0.0055

(c) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados balanceado

IRA - ADAN		
	KM	KMPS
KMPS	0.9782	-
KMPP	0.0000	0.0000

(d) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados desbalanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.0332	-
KMSPP	0.9782	0.0558

(e) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados desbalanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.6403	-
KMAPP	0.0005	0.0000

(f) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados desbalanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.7030	-
KMPP	0.0000	0.0000

(g) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados desbalanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.0003	-
KMSPP	0.0898	0.1696

(h) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados desbalanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.9156	-
KMAPP	0.0219	0.0065

(i) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados desbalanceados

Fonte: Autor (2024)

Tabela 24 – Valor-p do teste de *Nemenyi* para a Configuração 3

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.1122	-
KMAPP	0.8713	0.0332

(a) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados balanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.0337	-
KMAPP	0.8376	0.1308

(b) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados balanceados

IRA - ADAN		
	KM	KMPS
KMPS	0.4531	-
KMPP	0.0000	0.0000

(c) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados desbalanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.9516	-
KMSPP	0.0000	0.0000

(d) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados desbalanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.9156	-
KMAPP	0.0000	0.0000

(e) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados desbalanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.7635	-
KMPP	0.0000	0.0000

(f) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados desbalanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.9782	-
KMSPP	0.0000	0.0000

(g) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados desbalanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.9516	-
KMAPP	0.0000	0.0000

(h) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados desbalanceados

Fonte: Autor (2024)

Tabela 25 – Valor-p do teste de *Nemenyi* para a Configuração 4

IRA - ADAN		
	KM	KMPS
KMPS	0.0504	-
KMPP	0.9984	0.0439

(a) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados balanceados

IRA - ADAN		
	KM	KMPS
KMPS	0.9782	-
KMPP	0.0000	0.0000

(b) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados desbalanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.7635	-
KMSPP	0.0000	0.0000

(c) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados desbalanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.9156	-
KMAPP	0.0000	0.0000

(d) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados desbalanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.7635	-
KMPP	0.0000	0.0000

(e) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados desbalanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.8713	-
KMSPP	0.0000	0.0000

(f) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados desbalanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.5768	-
KMAPP	0.0000	0.0000

(g) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados desbalanceados

Fonte: Autor (2024)

Tabela 26 – Valor-p do teste de *Nemenyi* para a Configuração 5

IRA - ADAN		
	KM	KMPS
KMPS	0.2677	-
KMPP	0.0801	0.0006

(a) Valor-p do teste de *Nemenyi* para o IRA utilizando a ADAN para dados balanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.5453	-
KMPP	0.0047	0.0001

(b) Valor-p do teste de *Nemenyi* para a IMN utilizando a ADAN para dados balanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.0712	-
KMSPP	0.0065	0.6718

(c) Valor-p do teste de *Nemenyi* para o IRA utilizando a ADA1 para dados desbalanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.9945	-
KMPP	0.0006	0.0004

(d) Valor-p do teste de *Nemenyi* para a IMN utilizando a ADAN para dados desbalanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.3952	-
KMSPP	0.0023	0.1005

(e) Valor-p do teste de *Nemenyi* para a IMN utilizando a ADA1 para dados desbalanceados

Fonte: Autor (2024)

Tabela 27 – Valor-p do teste de *Nemenyi* para a Configuração 6

IRA - ADAN		
	KM	KMPS
KMPS	0.9033	-
KMPP	0.1026	0.0360

(a) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados balanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.7635	-
KMPP	0.0033	0.0002

(b) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados balanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.9782	-
KMSPP	0.0801	0.0492

(c) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados balanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.6403	-
KMAPP	0.1249	0.0122

(d) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados balanceados

IRA - ADAN		
	KM	KMPS
KMPS	0.9945	-
KMPP	0.0000	0.0000

(e) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados desbalanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.9156	-
KMSPP	0.0000	0.0000

(f) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados desbalanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.9156	-
KMAPP	0.0000	0.0000

(g) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados desbalanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.7635	-
KMPP	0.0000	0.0000

(h) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados desbalanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.3952	-
KMSPP	0.0000	0.0000

(i) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados desbalanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.9945	-
KMAPP	0.0000	0.0000

(j) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados desbalanceados

Fonte: Autor (2024)

Tabela 28 – Valor-p do teste de *Nemenyi* para a Configuração 7

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.9634	-
KMSPP	0.0442	0.0835

(a) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados balanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.7924	-
KMAPP	0.0011	0.0104

(b) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados balanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.9945	-
KMPP	0.0219	0.0164

(c) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados balanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.8713	-
KMSPP	0.0033	0.0164

(d) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados balanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.8944	-
KMAPP	0.0005	0.0028

(e) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados balanceados

IRA - ADAN		
	KM	KMPS
KMPS	0.8713	-
KMPP	0.0000	0.0000

(f) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados desbalanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.9782	-
KMSPP	0.0000	0.0000

(g) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados desbalanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.8713	-
KMAPP	0.0000	0.0000

(h) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados desbalanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.7030	-
KMPP	0.0000	0.0000

(i) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados desbalanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.5141	-
KMSPP	0.0000	0.0000

(j) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados desbalanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.2909	-
KMAPP	0.0001	0.0000

(k) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados desbalanceados

Fonte: Autor (2024)

Tabela 29 – Valor-p do teste de *Nemenyi* para a Configuração 8

IRA - ADAN		
	KM	KMPS
KMPS	0.9156	-
KMPP	0.0000	0.0000

(a) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados balanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.9782	-
KMSPP	0.0000	0.0000

(b) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados balanceados

IRA - ADA2		
	KMA	KMAPS
KMAPS	0.3952	-
KMAPP	0.0000	0.0000

(c) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA2 para dados balanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.9156	-
KMPP	0.0000	0.0000

(d) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados balanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.9516	-
KMSPP	0.0000	0.0000

(e) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados balanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.2909	-
KMAPP	0.0000	0.0000

(f) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados balanceados

IRA - ADAN		
	KM	KMPS
KMPS	0.7030	-
KMPP	0.0000	0.0000

(g) Valor-p do teste de *Nemenyi* para o IRA utilizando ADAN para dados desbalanceados

IRA - ADA1		
	KMS	KMSPS
KMSPS	0.9782	-
KMSPP	0.0000	0.0000

(h) Valor-p do teste de *Nemenyi* para o IRA utilizando ADA1 para dados desbalanceados

IMN - ADAN		
	KM	KMPS
KMPS	0.9516	-
KMPP	0.0000	0.0000

(i) Valor-p do teste de *Nemenyi* para a IMN utilizando ADAN para dados desbalanceados

IMN - ADA1		
	KMS	KMSPS
KMSPS	0.9156	-
KMSPP	0.0000	0.0000

(j) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA1 para dados desbalanceados

IMN - ADA2		
	KMA	KMAPS
KMAPS	0.1386	-
KMAPP	0.0008	0.1868

(k) Valor-p do teste de *Nemenyi* para a IMN utilizando ADA2 para dados desbalanceados

Fonte: Autor (2024)