DOIS ENFOQUES PARA A RESOLUÇÃO DO PROBLEMA DE AGRUPAMENTO

Marcella Braga de Assis Linhares

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientadores: Nelson Maculan Filho
Renan Vicente Pinto

Rio de Janeiro
Dezembro de 2021

DOIS ENFOQUES PARA A RESOLUÇÃO DO PROBLEMA DE AGRUPAMENTO

Marcella Braga de Assis Linhares

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientadores: Nelson Maculan Filho
Renan Vicente Pinto

Aprovada por: Prof. Nelson Maculan Filho
Prof. Renan Vicente Pinto
Prof. Adilson Elias Xavier
Prof. Pedro Henrique González Silva

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2021

Linhares, Marcella Braga de Assis

Dois Enfoques para a Resolução do Problema de Agrupamento/Marcella Braga de Assis Linhares. – Rio de Janeiro: UFRJ/COPPE, 2021.

XII, 55 p.: il.; 29, 7cm.

Orientadores: Nelson Maculan Filho
Renan Vicente Pinto

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2021.

Referências Bibliográficas: p. 51 – 55.

1. Clustering problem. 2. Nonlinear optimization models. 3. Modelagem matemática. 4. Problema de agrupamento. 5. Otimização. I. Maculan Filho, Nelson *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

"Não se pode falar de educação
sem amor."
(Paulo Freire)

Agradecimentos

Em primeiro lugar, eu gostaria de externar a minha imensa gratidão ao meu Pai Amado que me concedeu saúde, capacidade e condições necessárias para chegar até este momento de tamanha alegria. Se não fosse por Deus, eu, realmente, não estaria aqui. Então, agradeço ao Pai, Filho e Espírito Santo por me permitirem viver esta ocasião.

Gostaria de agradecer ao meu esposo, Jander Linhares, por acreditar tanto em mim, por me apoiar, ouvir e estar comigo em todos os momentos que vivi ao longo deste mestrado. Eu agradeço por todos os seus abraços que me consolaram, por todas as suas palavras de sabedoria que me motivaram e por todo seu carinho e apoio que me levantaram e me fizeram acreditar que seria possível. Obrigada por tudo que faz por mim, você é a pessoa mais incrível que eu conheço.

Agradeço aos meus pais e avó pelo apoio ao longo de toda a vida, por se esforçarem para me proporcionar o melhor e por todo ensinamento e direcionamento aos estudos. Obrigada por acreditarem em mim e me ensinarem valores que ficarão, no que depender de mim, como legado por gerações.

Agradeço ao meu orientador, professor Nelson Maculan, por me receber como aluna com tanto carinho, atenção e disponibilidade. Eu me sinto muito agraciada por ser orientada por alguém que eu admiro demais como profissional, que me trata com tanto respeito e que me constrange com tamanha humildade. Que honra tê-lo como meu orientador! Eu aprendo muito com o senhor, não somente ensinamentos sobre a pesquisa, mas ensinamentos valiosos sobre a vida que vou levar comigo por onde eu for. O senhor não faz ideia do quanto a sua orientação fez/ faz total diferença na minha trajetória. Obrigada por me inspirar tanto, motivar sempre e acreditar no meu trabalho.

Agradeço ao meu segundo orientador, professor Renan Pinto, por me receber de braços abertos com tanta disponibilidade. Se eu puder dizer isso, a minha sensação é que o senhor nasceu para ensinar. O que é uma das características mais incríveis, na minha humilde opinião. O senhor ensina com amor e isso me inspira tanto! Obrigada por estar sempre pronto para me ajudar, obrigada por toda paciência em sanar as minhas dúvidas (muitos áudios e mensagens rs!) e por acreditar no meu trabalho. A orientação do senhor com a do professor Maculan me proporcionou a leveza que

eu precisava para enfrentar as dificuldades que eu vivi e todos os demais desafios de uma pesquisa sendo realizada em plena pandemia global. Muito Obrigada!

Agradeço ao professor Abílio Lucena que me apresentou ao professor Maculan. E ao professor Daniel Ratton que me auxiliou neste importante momento, desempenhando a função de coordenador do programa com tanta maestria.

Agradeço a todos os meus amigos e amigas que me derem forças e me ajudaram no período deste trabalho. Em especial, agradeço aos amigos que fiz no PESC, principalmente, no LabOtim que era nosso ponto de encontro diário antes da pandemia. Agradeço à Amanda Azevedo que esteve comigo em todos os momentos, e se tornou uma grande amiga que vou levar para a vida. Obrigada por sua companhia, por chorar comigo, mas também por vibrar comigo, pela força, (incontáveis áudios rs) e todo incentivo que você me deu, minha amiga. Ao Victor Hugo que se tornou um amigo, parceiro em todas as horas de sufoco com o notebook e que sempre conseguia me surpreender e contagiar com sua tranquilidade nos momentos de mais desespero rs! À Jéssica Costa que se tornou minha parceira em todas as disciplinas desde o início do mestrado, a que dividia lugar no ônibus e na carona rs! Mas que agora eu tenho a honra de dividir o espaço na mesma equipe de trabalho no CapGov, obrigada amiga ! Agradeço ao comitê de representes discente do PESC, o qual eu tenho o privilégio de fazer parte e que costumo chamar de "equipe dos sonhos"rs!

Agradeço aos amigos que tive a felicidade de conhecer no grupo de apoio do Acolhe COPPE durante a pandemia. Com certeza, vocês tornaram meus dias de quarentena mais alegres e bem menos solitários. E, em especial, agradeço à psicóloga, Josiane Barros, que conduz este lindo projeto de acolhimento na COPPE . Obrigada, Josi, por me ouvir sempre atentamente, me compreender e me encorajar a tomar as atitudes necessárias que eu precisava para que eu pudesse mudar para melhor tantas situações nessa trajetória.

Agradeço ao Guty e ao Ricardo da secretaria do PESC que sempre me recepcionaram com um sorriso contagiante, sempre disponíveis a resolver nossos dilemas com sistema. E, à toda equipe PESC que faz o programa funcionar, especialmente ao pessoal da limpeza, que sempre chegavam muito cedo, assim como eu, e faziam questão de abrir o laboratório LUG bem cedinho para que eu não ficasse sozinha no bloco I até que outras pessoas começassem a chegar. Obrigada pelo cuidado comigo!

Agradeço aos professores Adilson Xavier e Pedro Henrique por aceitarem fazer parte deste momento compondo a nossa banca para a defesa.

Agradeço à CAPES pelo investimento a mim confiado para que eu pudesse ter condições de realizar esta pesquisa. Agradeço à COPPE e a UFRJ também.

Agradeço a todos que contribuíram diretamente ou indiretamente para que esta pesquisa fosse realizada. Isso inclui todos os profissionais que investiram na educação e na ciência e lutaram em prol dela, para que um dia eu também pudesse estar aqui.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

DOIS ENFOQUES PARA A RESOLUÇÃO DO PROBLEMA DE AGRUPAMENTO

Marcella Braga de Assis Linhares

Dezembro/2021

Orientadores: Nelson Maculan Filho
Renan Vicente Pinto

Programa: Engenharia de Sistemas e Computação

Neste trabalho, um estudo sobre duas novas abordagens para o problema de agrupamento é apresentado. Além de mostrar o desenvolvimento detalhado de dois modelos matemáticos propostos, esta pesquisa aponta para a importância de uma boa modelagem e para a diferença que ela pode causar na prática, contribuindo para o avanço na resolução do problema e, consequentemente, para o campo da programação matemática em geral. O diferencial dos novos modelos é que eles são desenvolvidos de maneira a evitar o problema da não diferenciabilidade e não convexidade em sua relaxação contínua. E a relevância destas novas abordagens são consolidadas através dos resultados computacionais desenvolvidos como experimentos comparativos para mostrar a força dos modelos propostos em contraste com outros modelos conhecidos e estudados na literatura.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

TWO APPROACHES TO SOLVING THE CLUSTERING PROBLEM

Marcella Braga de Assis Linhares

December/2021

Advisors: Nelson Maculan Filho
Renan Vicente Pinto

Department: Systems Engineering and Computer Science

In this paper, a study on two new approaches to the clustering problem is presented. Besides showing the detailed development of two proposed mathematical models, this research points to the importance of good modeling and the difference it can make in practice, contributing to the advancement in solving the problem and, consequently, to the field of mathematical programming in general. The distinguishing feature of the new models is that they are developed in a way that avoids the problem of non-differentiability and nonconvexity in their continuous relaxation. And the relevance of these new approaches are consolidated through the computational results developed as comparative experiments to show the strength of the proposed models in contrast to other known and studied models in the literature.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

COPPE	Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia, p. 1
CP1	Primeiro Modelo Proposto para o Problema de Cluster, p. 38
CP2-BIS	Nova Abordagem para o Segundo Modelo de Cluster Proposto, p. 40
CP2	Segundo Modelo Proposto para o Problema de Cluster, p. 39
MSSC	Minimum Sum of Squares Clustering Problem, p. 30
UWWP	Unitary Weighed Weber Problem, p. 28
WP	Weber Problem, p. 25

Capítulo 1

Introdução

1.1 O problema de Agrupamento

A Análise de Cluster ou Agrupamento é uma ferramenta utilizada para encontrar grupos semelhantes dentro de um determinado conjunto de entidades, isto é, trata-se de resolver o problema de encontrar subconjuntos homogêneos chamados clusters, em que os objetos que tenham maior similaridade entre si sejam agrupados juntos, enquanto que os objetos com maior diferença pertençam a clusters diferentes. A partir disso, dois princípios são definidos: o de homogeneidade e o de separação. O primeiro é equivalente à semelhança dos elementos de mesmo grupo e o segundo à separação ou diferença entre elementos de grupos diferentes. [2]

Além dos princípios descritos acima, ALOISE e HANSEN [3] descrevem *Clustering* como uma poderosa ferramenta para a automatização de dados e apresentam em seu artigo uma descrição mais formal para o problema. Também podemos encontrar nos trabalhos de HRUSCHKA e EBECKEN [4] e COLE [5] definições formais para o problema.

No livro de COLE [5], é possível encontrar a definição de um dos tipos de problema de cluster mais utilizado: o particionamento. Ele é descrito da seguinte forma:

Seja $X = \{X_1, X_2, X_3, ..., X_n\}$ um conjunto com n objetos , onde cada $X_i \in \mathbb{R}^p$ é um vetor com p características, isto é, p é a quantidade de coordenadas que dimensiona o objeto vetorial. Estes objetos devem ser particionados em k subconjuntos disjuntos que chamaremos de clusters $C = \{C_1, C_2, C_3, ..., C_k\}$, de maneira a respeitar as condições:

1. $C_1 \cup C_2 \cup C_3 \cup ... \cup C_k = X$;
2. $C_i \neq \emptyset, \forall\, 1 \leq i \leq k$;

3. $C_i \cap C_j = \emptyset, \forall\, 1 \leq i < j \leq k$;

As condições acima ressaltam que todos os objetos devem pertencer a um único cluster e que nenhum cluster pode ser vazio. Além disso, ele acrescenta que existem casos em que o k pode ser desconhecido. Já nos casos contrários a esse, o problema é chamado de k-Clusterização.

A dificuldade de resolução do problema é ilustrada de forma simplificada, porém muito instrutiva em [5]. No formato de um exemplo, é apresentado a quantidade de maneiras que se pode classificar n entidades em k grupos: se existirem 25 objetos e 5 clusters, de acordo com este livro citado, existem 2 436 684 974 110 751 formas de classificá-los dentro dos 5 grupos. Isso se dá pelo resultado de LIU [6] através da fórmula:

$$N(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^n.$$

E, no caso do número de cluster não ser conhecido, seriam

$$\sum_{i=1}^{n} N(n,k)$$

maneiras diferentes de agrupar n objetos em no máximo k clusters. Para este caso, com 25 objetos, o livro relata que seria $4 \cdot 10^{18}$ agrupamentos.

Esse exemplo revela de maneira simples o quanto estamos lidando com um problema difícil de ser resolvido. E isso revela que o fato da ideia do problema ser de fácil entendimento, não implica que ele seja um problema fácil de resolver. Em [3] ele é apresentado através do modelo MSSC - *Minimum Sum of Squares Clustering*, também conhecido como agrupamento pela soma mínima das distâncias quadráticas, que é um problema da classe NP-difícil conforme provado por ALOISE *et al.* [7].

O MSSC é um dos objetos de estudo desse trabalho e será abordado mais adiante. Sua modelagem é uma das mais comuns e frequentemente utilizada na literatura, como pode ser visto em [5, 8]. Entretanto, o problema de clusterização não se limita a uma única abordagem, ele pode ser formulado de diversas maneiras, além de poder ser direcionado por critérios variados. Ele também pode ser resolvido de diferentes formas, e esses critérios podem ser usados como um guia na escolha da metodologia mais eficiente em cada caso. Na Figura 1.1, ilustramos de maneira simples que é possível fazer diversos tipos de agrupamentos utilizado um único grupo

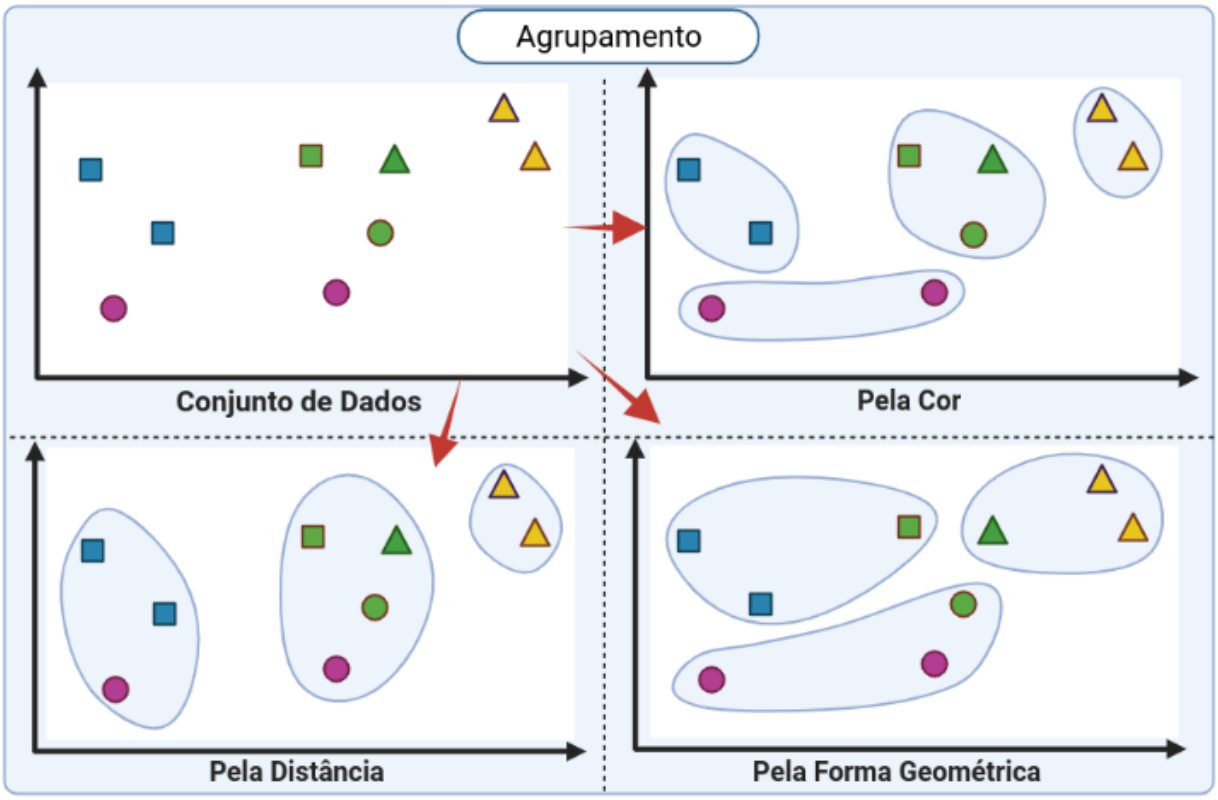

Figura 1.1: Exemplo da diversidade de agrupamentos ainda que sendo utilizado um único grupo de dados. Fonte da imagem: Criada pela autora no aplicativo BioRender.

de dados. E isso vai depender do objetivo do problema e das características dos dados usados. O problema de agrupamento é extremamente aplicável em diversas áreas, conforme vamos mostrar mais adiante, e por ser um problema interdisciplinar que busca modelar a realidade de cada caso, ele é também um problema bastante complexo de ser resolvido. Neste trabalho, damos enfoque ao estudo de duas novas abordagem com o objetivo de modelar o problema de agrupamento de forma a obtermos relaxações contínuas convexas e diferenciáveis, e investigamos o quanto elas são promissoras e relevantes diante de outras formulações que são comumente trabalhadas na literatura. Mas antes disso, após ter apresentado sua definição e ilustrado sua complexidade, ainda neste capítulo, apresentamos uma introdução geral sobre o problema de agrupamento com o objetivo de contextualizar o problema e mostrar sua grande utilidade.

1.2 Medidas de Proximidade

A forma de demonstrar a semelhança ou diferença entre os objetos do conjunto pode ser diferente em cada situação, isso vai depender do contexto em que o problema está enquadrado. Ademais, grande parte dos métodos de análise de agrupamento

estão diretamente associados a esta medida de proximidade ou função de distância [4]. Neste sentido, essas medidas são usadas para quantificar as dissimilaridades ou similaridades entre as entidades. Em seu trabalho, XU e WUNSCH [9] mostram a diferença entre as medidas de similaridade e de dissimilaridade e acrescentam que, normalmente, a primeira é usada quando as características dos objetos são binárias, enquanto que a segunda é comumente utilizada quando elas são descritas de maneira contínua. Essa informação apresentada por XU e WUNSCH [9] leva em consideração que os dados de cada objeto são disponibilizados no formato de um vetor multidimensional, em que os elementos do vetor são os atributos ou características, ou ainda, são chamados de recursos que podem ser qualitativos ou quantitativos, contínuos ou binários. Consequentemente, isso determina a abordagem da medida de proximidade a ser aplicada em cada ocasião.

Nas funções de dissimilaridade, as entidades do mesmo grupo têm menor distância, e as de grupos diferentes possuem uma distância maior. Uma forma de entender melhor essa relação entre a proximidade e as métricas de distância é, quando o valor da distância entre duas entidades é pequeno, dizemos que elas são próximas uma da outra e, de acordo com um critério adotado, tendem a pertencer ao mesmo agrupamento. Neste seguimento de estudo, pode-se encontrar diversas medidas de dissimilaridade nos trabalhos de COLE [5] , XU e WUNSCH [9], LINDEN [10].

De acordo com HRUSCHKA e EBECKEN [4], uma das mais utilizadas é a Distância Euclidiana, que definiremos a seguir:

- Distância Euclidiana

$$d_2(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2} = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2} \quad (1.1)$$

Onde d_2 é a distância euclidiana entre os objetos i e j, já os termos x_{il} e x_{jl} correspondem ao l-ésimo atributo dos objetos (vetores) em um espaço p-dimensional. Todas as próximas distâncias apresentadas a seguir obedecerão a estes mesmos significados de acordo com a simbologia apresentada acima, porém fazendo referência a outros tipos de métricas, não mais a $d_2(i,j)$ por exemplo.

Além da Euclidiana, existem outras distâncias que também podem ser utilizadas para medir a dissimilaridade dos objetos. A propósito, a próxima distância que será definida é uma generalização da euclidiana, é chamada de Distância de Minkowiski.

- Distância de Minkowski

$$d_m(i,j) = \sqrt[m]{|x_{i1} - x_{j1}|^m + |x_{i2} - x_{j2}|^m + ... + |x_{ip} - x_{jp}|^m} = \sqrt[m]{\sum_{l=1}^{p}(x_{il} - x_{jl})^m} \tag{1.2}$$

Um outro exemplo que pode ser usado para medir a proximidade entre as entidades é a Distância de Manhattan, que também é chamada de City-block ou L_1. Ela, assim como a euclidiana, é uma derivação da Minkowski, porém com $m = 1$, a euclidiana possui $m = 2$.

- Distância de Manhattan

$$d_1(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}| = \sum_{l=1}^{p} |x_{il} - x_{jl}| \tag{1.3}$$

Ainda como exemplo, a Distância Máximo pode ser citada, que também á conhecida por "Distância Sup" pois é definida como o valor máximo da distância entre os atributos dos objetos em questão.

- Distância Máximo

$$d_{max}(i,j) = \max_{1 \leq l \leq p} |x_{il} - x_{jl}| \tag{1.4}$$

Além destas apresentadas aqui, existem outros exemplos como a Distância de Mahalanobis, Distância de Hamming, entre outros que também podem ser vistas em XU e WUNSCH [9]

Funções de dissimilaridade são métricas definidas em um determinado conjunto. (GAN *et al.* [11]). Elas devem atender as seguintes características:

(ver também HAN *et al.* [12], DUONG *et al.* [13])

1. Simetria: a distância do objeto i para o j é a mesma se calculada do objeto j para o i.

$$d(i,j) = d(j,i)$$

2. Positividade: a distância entre os objetos precisa ser não-negativa.

$$d(i,j) \geqslant 0, \forall i,j$$

3. Reflexividade: a distância é nula se e somente se estiver sendo medida entre dois objetos iguais (ou com valores iguais em seus atributos) .

$$d(i,j) = 0 \Leftrightarrow i = j$$

4. Desigualdade Triangular: onde i, j e h são considerados objetos do conjunto.

$$d(i,j) \leqslant d(i,h) + d(h,j)$$

Conforme mencionado anteriormente, existe outro tipo de função que também cumpre o papel de mensurar o nível de proximidade entre as entidades, são as funções de similaridade. Estas podem ser representadas por $s(i,j)$, que de acordo com KAUFMAN e ROUSSEEUW [14] cumprem as propriedades a seguir:

1. Positividade: os valores podem variar entre 0 e 1.

$$0 \leqslant s(i,j) \leqslant 1$$

2. Simetria: a distância do objeto i para o j é a mesma se calculada do objeto j para o i.

$$s(i,j) = s(j,i)$$

3. A distância vale 1 se e somente se estiver sendo medida entre dois objetos iguais (ou com valores iguais em seus atributos).

$$s(i,j) = 1 \Leftrightarrow i = j$$

As medidas de similaridade normalmente variam entre 0 e 1, de forma que 0 representa nenhuma similaridade e 1 representa similaridade total entre os objetos, já os valores intermediários a esses, descrevem o grau ou o nível de proximidade entre as entidades do conjunto. Quanto mais próximos ou semelhantes eles forem, maior será o valor da função de similaridade, de acordo com a definição abordada por KAUFMAN e ROUSSEEUW [14].

Medidas de similaridade também podem ser encontradas em CUI *et al.* [15], GAN *et al.* [11].

Logo, as medidas de proximidade, as quais podem ser tanto de similaridade quanto de dissimilaridade, são de extrema importância para a resolução do problema em questão, uma vez que a escolha da métrica de distância tem grande impacto na forma como ocorrerá o agrupamento. Em seu trabalho, HANSEN e JAUMARD [2] relatam que a medida de proximidade, juntamente com restrições, como por exemplo, restrições do tipo de cluster a ser usado, cardinalidade, peso máximo, entre outras, compõem o problema de *clustering* como um problema de programação matemática. Neste caso, a proximidade é descrita até mesmo na função objeto do problema, o que demonstra seu grande impacto e relevância na modelagem, e consequentemente, nos resultados finais da clusterização.

1.3 Tipos de Algoritmos

Existem diversos algoritmos de agrupamento na literatura, em alguns trabalhos, como o de FAHAD *et al.* [16], podemos encontrá-los de maneira categorizada. Eles são setorizados por tipo, a partir de detalhes técnicos em seu procedimento interno no processo de agrupamento. As duas categorias mais frequentes na literatura são as do tipo hierárquico e de particionamento. [2, 5, 9, 14, 17]

- Agrupamento com abordagem hierárquica

  Os algoritmos de agrupamento do tipo hierárquico permitem resolver o problema de clusterização sem que seja necessário designar previamente a quantidade de grupos ou número de clusters em que os objetos deverão ser divididos, ele é baseado num processo de hierarquia que, por sua vez, está ligada à proximidade das observações no conjunto de dados fornecido. Essa hierarquia pode ser estruturada de diversas maneiras. As duas mais usadas são as aglorematívas e as divisivas. Estas compõem dois algoritmos hierárquicos conhecidos na literatura como: *Agglomerative Hierarchical Clustering Algorithms* (AGNES) e *Divisive Hierarchical Clustering Algorithm* (DIANA). A diferença entre eles é que os algoritmos de agrupamento hierárquicos aglomerativos têm uma abordagem conhecida por *bottom-up*, isto é, um comportamento ascendente (de baixo para cima), onde, inicialmente, é atribuído um objeto para cada cluster e, com base na similaridade, dois ou mais clusters são mesclados recursivamente. Já os algoritmos de agrupamento hierárquicos divisivos são considerados o oposto dos aglomerativos, eles tem uma abordagem *top-down* (de cima para baixo) em que todas as observações pertencem a um único clus-

ter inicialmente, e com base na dissimilaridade, este grande cluster é dividido em agrupamentos menores de forma recursiva. Essa diferença é ilustrada na Figura 1.2. Nos dois tipos de algoritmos, o processo recursivo acontece até que seja interrompido por um critério de parada definido, que pode ser o tempo ou até um limitante que fornecerá uma série de agrupamentos limite. Conforme mencionado, em ambos métodos, é preciso mensurar a proximidade dos agrupamentos, seja a similaridade (nos métodos aglomerativos) ou a dissimilaridade (nos métodos divisivos), isso é feito a partir do cálculo da distância entre os clusters. Essa distância pode ser obtida de várias maneiras, além da escolha de qual medida usar, também pode ser escolhido o formato de ligação dela, a saber: completa, única ou média. A ligação única - *single linkage* é caracterizada por unir dois agrupamentos baseada na menor distância entre os clusters, o que seria uma ligação no formato de "vizinho mais próximo", conforme a Figura 1.3 mostra. Já a ligação completa - *complete linkage* é o oposto da anterior. Esta é feita de maneira a ligar os clusters pela maior distância entre dois pontos, o chamado "vizinho mais distante", como pode ser visto na Figura 1.4. A ligação média - *average linkage* é caracterizada pela distância média, que é calculada de cada ponto de um cluster a todos os pontos do outro cluster, como mostra é ilustrado na Figura 1.5.

Além disso, uma outra característica marcante muito utilizada nos métodos hierárquicos é a utilização de dendrogramas como forma de visualização do agrupamento, o que ajuda a tornar os clusters visualmente mais compreensíveis com o auxílio de mais este gráfico. Um exemplo de dendrograma pode ser visto na Figura 1.6. Uma desvantagem nesta abordagem é o fato de que quando as decisões são tomadas, sejam em processos de fusão ou separação, elas não podem ser desfeitas, são carregadas ao longo do processo.

- Agrupamento com abordagem de particionamento

  Os algoritmos de particionamento, como o próprio nome já diz, têm o objetivo de particionar todos os objetos dados no problema em uma quantidade predeterminada de partições, isto é, a quantidade de clusters é definida previamente neste tipo de método, e este particionamento é feito sem nenhum tipo de estrutura hierárquica ou estágios aglomerativos e/ou divisivos. Pelo contrário, é feito de maneira direta, separando os dados entre as q partições predefinidas. Além disso, existem duas características específicas que funcionam como requisitos para serem cumpridos neste tipo de abordagem, elas são:

  1. Todo agrupamento precisa ter pelo menos um objeto;
  2. Cada objeto deve pertencer a um único agrupamento.

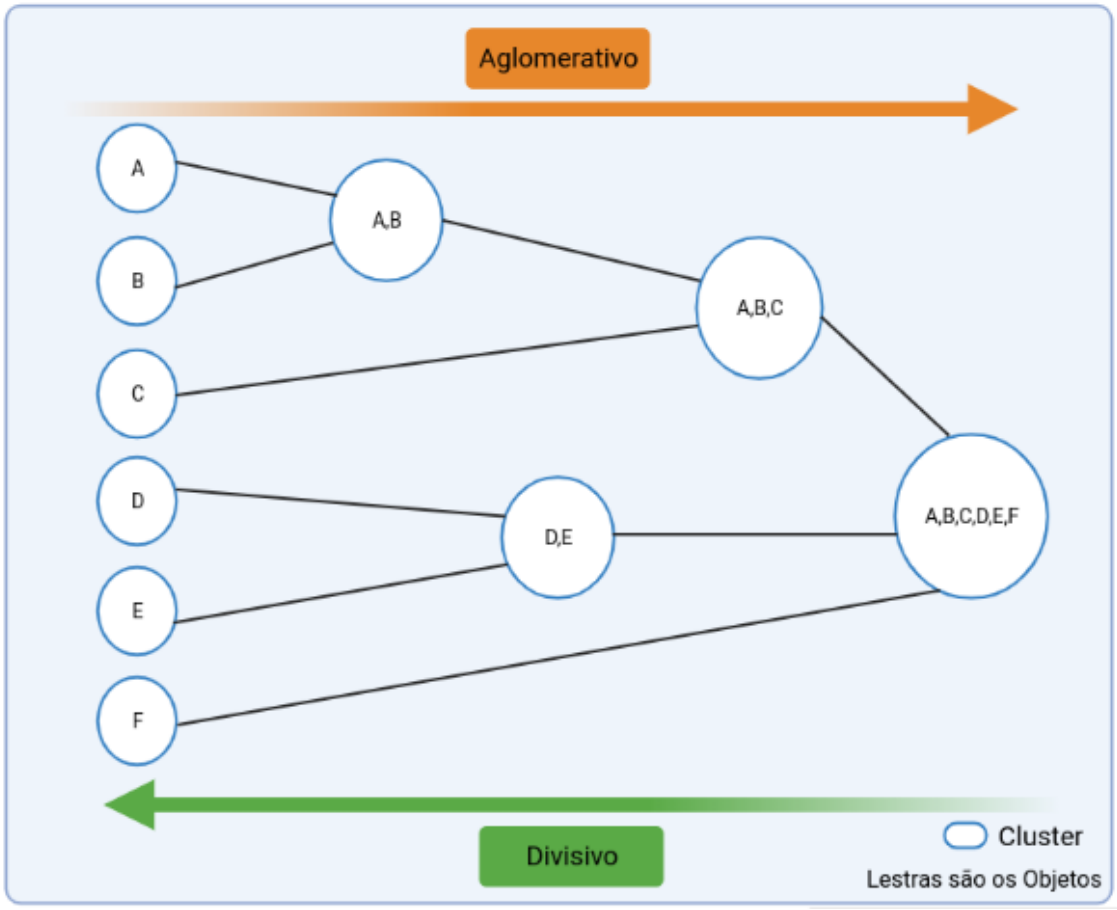

Figura 1.2: Diferença entre Abordagem Hierárquica Aglomerativa e Divisiva. Fonte da imagem: Criada pela autora no aplicativo BioRender.

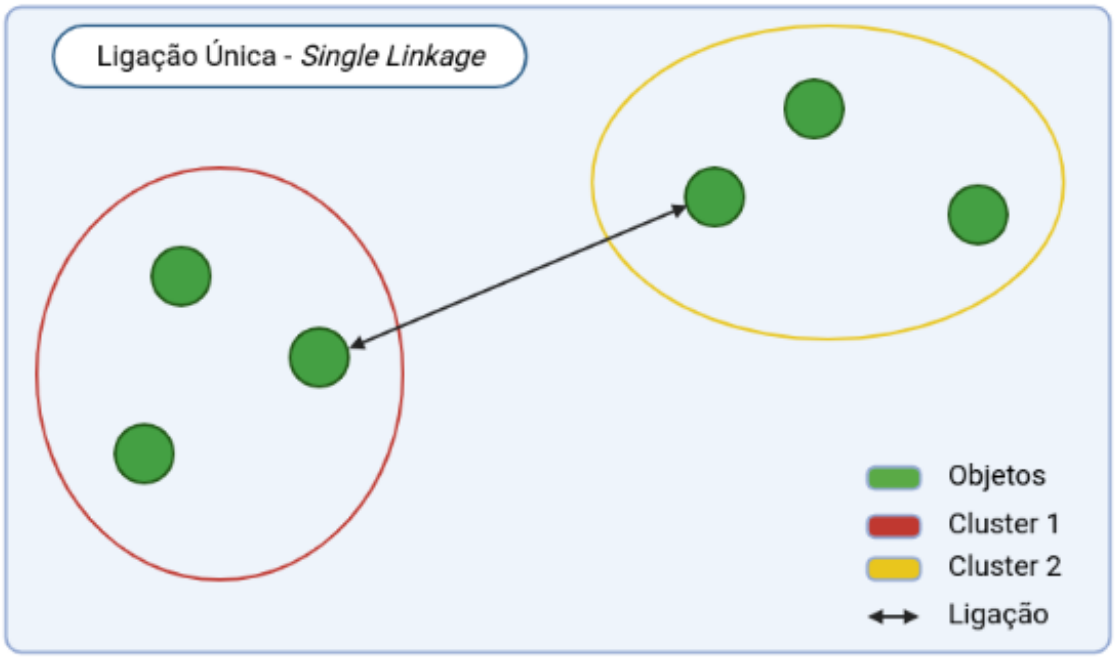

Figura 1.3: Estrutura de Ligação Única - *Single Linkage*. Fonte da imagem: Criada pela autora no aplicativo BioRender.

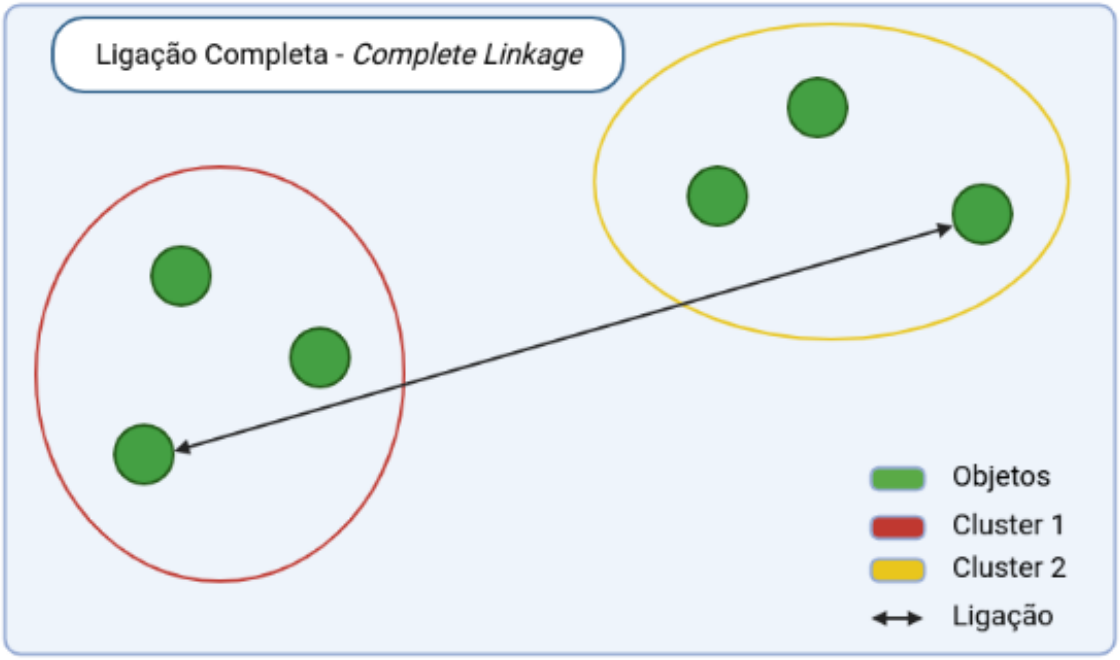

Figura 1.4: Estrutura de Ligação Completa - *Complete Linkage*. Fonte da imagem: Criada pela autora no aplicativo BioRender.

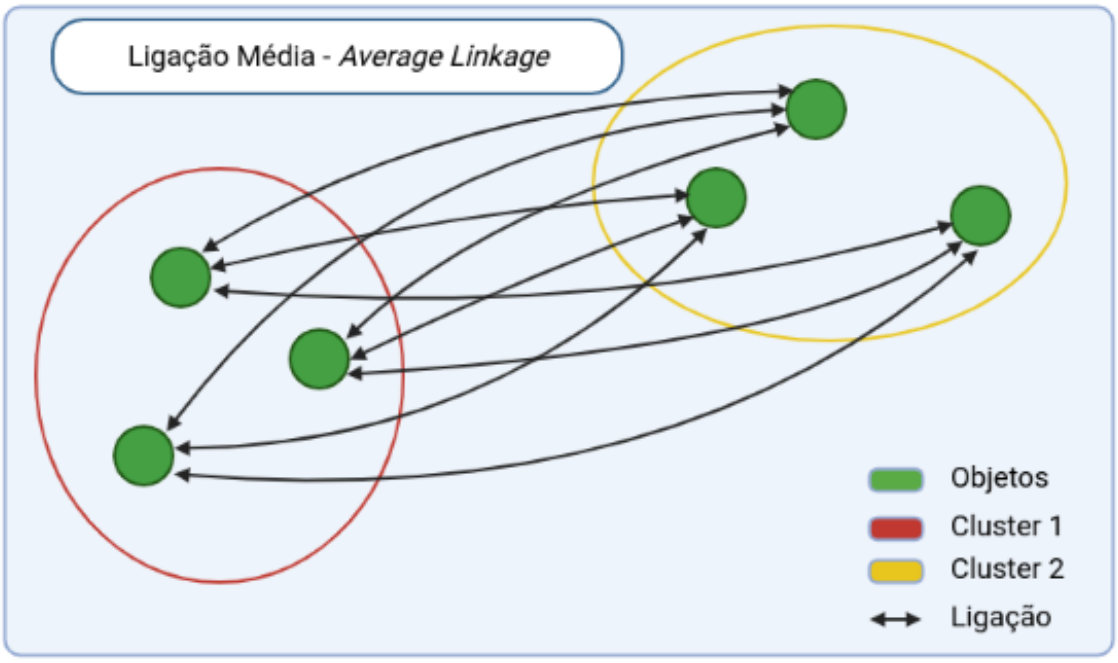

Figura 1.5: Estrutura de Ligação Média - *Average Linkage*. Fonte da imagem: Criada pela autora no aplicativo BioRender.

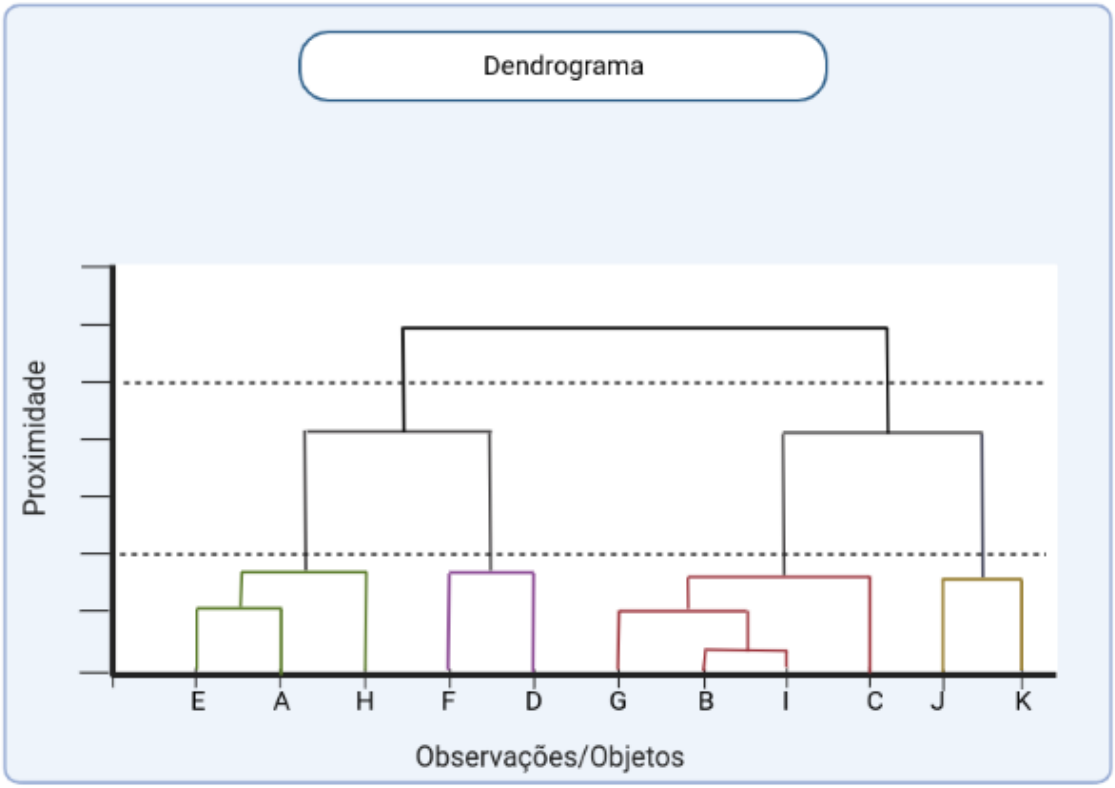

Figura 1.6: Ilustração de um Gráfico Dendrograma. As linhas pontilhadas indicam dois exemplos de cortes/limites diferentes que, neste caso, pode resultar em 4 clusters coloridos ou somente 2 clusters. Fonte da imagem: Criada pela autora no aplicativo BioRender.

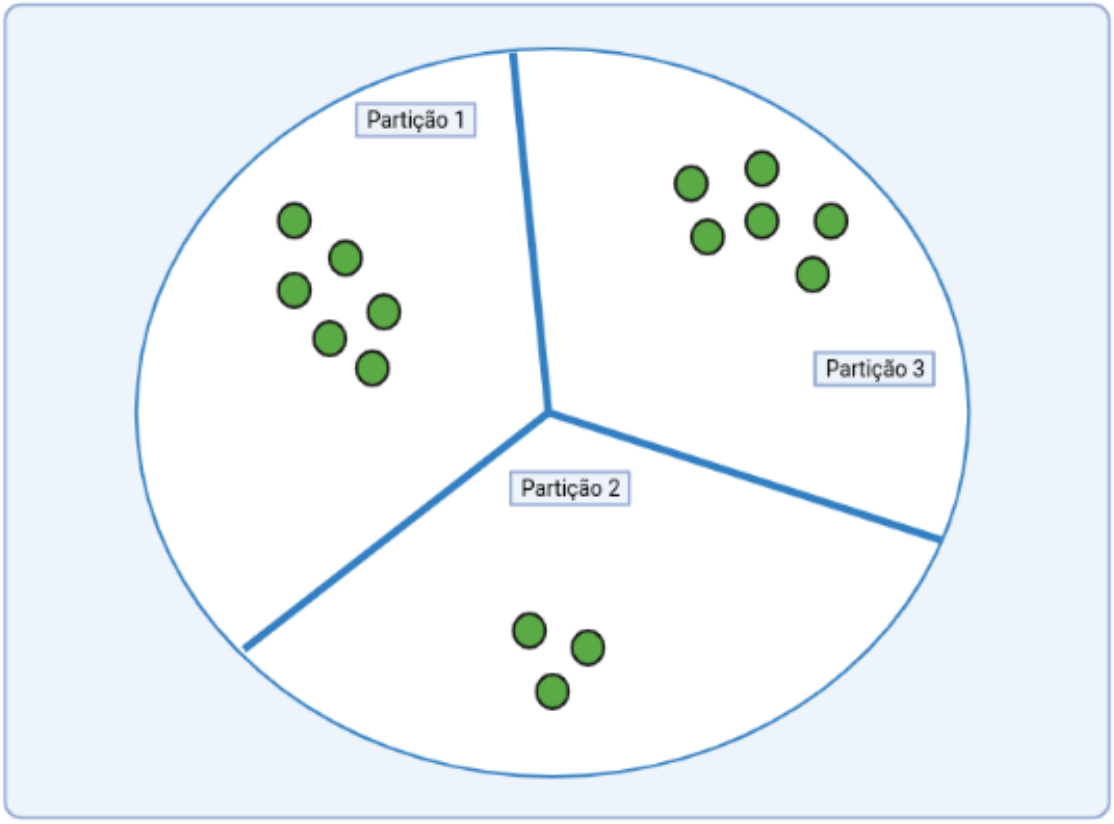

Figura 1.7: Abordagem de Particionamento. Exemplo com 15 objetos e 3 partições. Fonte da imagem: Criada pela autora no aplicativo BioRender.

Isto implica que os q grupos pré-fixados devem possuir todos os objetos do conjunto de dados e que cada um destes objetos devem estar contidos em um único cluster, além de não haver agrupamento vazio na solução. A Figura 1.7 mostra um exemplo de como funciona o processo em algoritmos de particionamento em um exemplo com 15 objetos para serem particionados em 3 clusters.

1.4 Aplicações

No trabalho de KAUFMAN e ROUSSEEUW [14], podemos ler que desde a infância o ser humano é ensinado a identificar padrões, ou ainda, fazer a separação ou classificação do que há ao seu redor. Essa classificação se dá no ato de identificar objetos, pessoas, animais ou qualquer outro ser existente a sua volta baseado nas características dos mesmos, por exemplo, aprender a classificar homens e mulheres, aves e peixes, objetos escolares e objetos de cozinha, roupas de frio e de banho, entre outros. O livro relata que separar, agrupar ou classificar é essencial, e este aprendizado acompanha as pessoas desde a infância, mesmo que inconscientemente. E, com o passar do tempo, houve a necessidade de agrupamentos ainda mais robustos, isto é, aqueles que fogem da percepção e julgamento subjetivo do "sistema olho-cérebro" humano, que, por sua vez, atende à demandas de até três dimensões na classificação. Consequentemente, surgindo problemas maiores, também surgiram estudos de como poderiam ser feitos esses agrupamentos com maiores dimensões, de forma automatizada. Ainda em [14], é descrito que uma única definição e geral de cluster é inexistente, o que exitem são variadas definições que dependem das características da aplicação em questão. Por sua vez, existem diversas aplicações que se diferem totalmente e uma das principais diferenças está no tipo de dados utilizados, assim como na necessidade do uso de variáveis contínuas e/ou discretas, no uso de dissimilaridades e similaridades. Essas peculiaridades de cada aplicação refletem bem o motivo pelo qual existem tantos métodos de solução para o problema de clusterização e a dificuldade (se não, impossibilidade) de se encontrar uma única forma para resolver de maneira eficiente todas as aplicações para o problema de agrupamento existentes, além da dificuldade para escolher qual a melhor forma de solução em cada caso nas diversas aplicações.

Podemos encontrar na literatura diversas aplicações, inclusive em áreas de estudo variadas, como por exemplo, biologia, marketing, medicina, estatística, geografia, química, história, telecomunicações, astronomia, ciências sociais, psicologia, entre outros. O problema de clusterização é, de fato, um problema extremamente aplicável. Por esse motivo, podemos encontrar uma gama extensiva de estudos passíveis de aplicações relacionadas ao problema em suas variadas versões. [14]

Aqui vamos citar algumas dessas aplicações existentes:

- Na medicina, o problema de agrupamento pode ser bastante útil no auxílio de diagnósticos e prognósticos. Os profissionais da saúde, além de utilizarem exames laboratoriais e suas próprias experiências na área, também podem ser amparados com o auxílio desta ferramenta podendo obter maior êxito nas

previsões e, consequentemente, na tomada de decisão, uma vez que ações preventivas são de extrema importância quando se trata de saúde. O problema de agrupamento pode auxiliar na classificação de pacientes com pré-disposição a uma determinada doença, assim como numa possível previsão com mais rapidez para os diagnósticos médicos e prognósticos . Esse tipo de aplicação ocorre na pesquisa de BAGIROV *et al.* [18] e trata-se da utilização de técnicas baseadas em otimização global para aumentar a precisão e auxiliar na tomada de decisão de diagnóstico médico, além de relatar um método que calcula centros de clusters de maneira a auxiliar na previsão de um possível retorno do câncer de mama em pacientes cujo tumor tenha sido removido. WANG e GARIBALDI [19] também retratam em seu trabalho a utilização do problema de agrupamento no auxílio de identificação do câncer.

- Na psicologia, que também é uma área ligada à medicina, a forma com que o problema de agrupamento pode ser abordado está ligado à análise comportamental dos indivíduos, fazendo com que seja possível identificar padrões de transtornos psicológicos de maneira a auxiliar na classificação deles. WOLKIND e EVERITT [20] realizam a aplicação do problema de cluster em um grupo de crianças com três anos de idade e mostram que é possível desenvolver um sistema de diagnóstico para distúrbios psicológicos mesmo em crianças tão pequeninas.

- No marketing, o problema de agrupamento pode ser útil para uma divulgação direcionada a clientes potenciais. Isso pode ocorrer quando, por exemplo, a base de dados utilizada obtém características dos clientes e de possíveis compradores, buscando similaridades entre eles e explorando ao máximo quais seriam os futuros "clientes potenciais", fazendo com que esse agrupamento permita que o marketing seja mais assertivo, alcançando o maior número de pessoas de maneira eficiente. No trabalho feito por CHIANG *et al.* [21], podemos encontrar esse tipo de estudo em que o problema de agrupamento é aplicado à analise de mercado com o objetivo de fornecer uma estratégia de divulgação eficiente a uma empresa de transporte aéreo.

- Na biologia, o problema de cluster pode auxiliar no agrupamento de dados de expressão gênica, sendo útil para identificar grupos de genes e amostras biologicamente relevantes. É com base nesse tipo de agrupamento de dados em expressão gênica que JIANG *et al.* [22] apresenta seu trabalho. Além dele, esse tipo de aplicação em importantes processos biológicos envolvendo genes também é tema de um capítulo do livro abordado por GAN *et al.* [11].

- Na logística empresarial, o problema de clusterização pode ser utilizado como

ferramenta para propor o layout das estações de trabalho, posicionamento de equipamentos e unidades de apoio com objetivo de maximizar o funcionamento de processos, melhorando o desenvolvimento de tarefas e otimizando o ambiente de trabalho. Um exemplo seria o problema de agrupar uma grande quantidade de funcionários que possuam algum padrão característico de similaridade proporcionando maior produtividade à equipe de trabalho. Isso é relatado por NEGREIROS *et al.* [23], em que o problema de agrupamento capacitado é utilizado como ferramenta para auxiliar na tomada de decisão de como seria o melhor layout para alocar grandes equipes de tecnologia da informação em fábricas de softwares.

- Na geofísica, o problema de agrupamento pode ser uma ferramenta muito favorável para análise de dados sísmicos. No trabalho apresentado por DZWINEL *et al.* [24], o problema de cluster é usado para analisar dados reais e sintéticos de terremotos para prever tendências a abalos sísmicos.

Além das aplicações aqui citadas, existem outras grandes áreas que, da mesma forma, utilizam a análise de cluster como abordagem para prever, otimizar, prevenir, decidir e resolver da melhor maneira possível seus respectivos problemas. Entre elas, podemos referenciar ainda os trabalhos de GUAN *et al.* [25] na área de sistemas de informação para detectar invasores cibernéticos, LIU *et al.* [26] no sistema de busca de imagens, LIU *et al.* [27] para classificação de emissor de radar que se torna muito útil na área militar, AKKAYA *et al.* [28] para o agrupamento de sensores sem fio em redes de telecomunicações, entre outros.

1.5 Objetivos

Além de apresentarmos um estudo sobre problemas de agrupamento em geral, o principal motivo deste trabalho é fazer um estudo mais aprofundado sobre as novas modelagens propostas para o problema. Nesta pesquisa, o desenvolvimento destes modelos é apresentado, além da forma como eles podem ser significativos na área de *clustering*. Comparamos as novas abordagens com modelagens já apresentadas na literatura e mostramos, através de resultados computacionais, sua relevância de maneira efetiva para instâncias de diferentes formatos.

No Capítulo 1, apresentamos o problema de forma geral, mostramos as possíveis medidas de proximidade que podem ser utilizadas para descrever a similaridade ou dissimilaridade do problema, os tipos de algoritmos que podem ser usados dependendo da abordagem utilizada e seus detalhes técnicos no procedimento de

agrupamento e suas aplicabilidades.

No Capítulo 2, apresentamos uma revisão bibliográfica mostrando como o problema de agrupamento já foi abordado na literatura como um problema de otimização. Além disso, mostramos de que maneira ele já foi resolvido, sendo por algoritmos exatos ou heurísticos conhecidos na literatura.

No Capítulo 3, escrevemos a metodologia do trabalho expondo modelos já utilizados nas bibliografias e apresentando o desenvolvimento dos modelos propostos.

No Capítulo 4, descrevemos como foram realizados todos os experimentos computacionais, mostramos as tabelas com os resultados das comparações entre os modelos apresentados, da literatura e os propostos, e, além disso, discutimos sobre todas as comparações, critérios utilizados e resultados em cada tabela apresentada.

No Capítulo 5, concluímos com as considerações finais sobre o projeto e possíveis trabalhos que podem ser realizados futuramente.

Capítulo 2

Revisão Bibliográfica

2.1 Agrupamento como um problema de otimização

O trabalho de FENG [29] propõe que algoritmos de *clustering* funcionem como um problema de otimização de forma a minimizar ou maximizar uma determinada função global, que depende de cada caso. Ele apresenta algoritmos hierárquico e particional, mais especificamente, *Hierarchical Agglomerative Clustering* (HAC) para o primeiro caso e *K-means* para o segundo, como exemplos de métodos de solução para o problema e, em seguida, mostra suas respectivas funções globais que, para o autor, representam a função de otimização que compõe o algoritmo. Além disso, mostra que existem funções globais de otimização por trás dos métodos de agrupamento e que, por este motivo, eles podem ser representados como um problema de otimização. A ideia abordada pelo autor retrata as vantagens dessa representação na estrutura de otimização com relação à modelagens mais complexas dos problemas, contextos com multiplicidade de objetivos que refletem numa formulação mais fidedigna e na representação do termo de similaridade de agrupamento.

Um outro artigo que aborda a análise de cluster visando a programação matemática é o de HANSEN e JAUMARD [2]. Eles apresentam uma revisão das principais classes dos problemas de agrupamento e apontam variados métodos de solução usados para resolver o problema de forma eficiente. Logo na introdução do paper, os autores relatam uma série de perguntas as quais julgam importante serem respondidas dado um problema de agrupamento, elas são:

- *Qual é o objetivo do agrupamento? - a questão do critério (ou critérios);*
- *Temos justificativa para perseguir esse objetivo? - a questão da axiomática;*
- *Quais restrições devem ser consideradas? - a questão da escolha do tipo de agrupamento;*

- *Quão difícil é realizar o agrupamento? - a questão da complexidade;*
- *Como deve ser feito o agrupamento? - a questão do projeto do algoritmo;*
- *O agrupamento obtido é significativo? - a questão da interpretação.* [2]

Essas questões apresentadas pelos autores são de importante impacto no processo, pois as respostas elaboram atributos essenciais para conduzir o estudo de um problema de agrupamento, desde seus passos iniciais até as análises e interpretações finais.

No trabalho de KABADI *et al.* [30] podemos ver a importância da modelagem matemática como ferramenta para enunciar o problema de agrupamento, além do valor que uma modelagem de boa qualidade agrega ao processo de resolução do problema tratado, ela influencia de tal forma que pode ser crucial para os resultados deste processo. O autor se reporta ao agrupamento como um dos principais componentes para muitos modelos de otimização combinatória, inclusive ele apresenta a forma como a clusterização está presente em modelos de otimização combinatória. KABADI *et al.* [30] utiliza problemas como o de alocação de tarefas, de localização em diferentes aplicações e de roteamento, que são enunciados na literatura por modelos de otimização combinatória, para exemplificar esta familiaridade com os objetivos a seres alcançados quando se trata de agrupamento. O autor utiliza como exemplo o problema de determinar o custo mínimo para alocação de tarefas a processadores em um projeto de arquitetura automobilística cumprindo as restrições de custo e de capacidade dos processadores. Ele faz a seguinte alusão: atribuir n tarefas a m processadores, satisfazendo as respectivas restrições e sendo $n >> m$. De fato, é o processo de particionar n tarefas em grupos ou clusters, atribuindo cada grupo à um processador de maneira a satisfazer as restrições do problema, o que não deixa de ser um problema de clusterização. Um outro exemplo que o autor utiliza é o de empresas que possuem diversos escritórios em localidades diferentes e que precisam instalar alguns centros de treinamento que atendam a todos esses escritórios, o problema é caracterizado por encontrar a localização ideal de m centros de treinamento em que n escritórios possam enviar seus funcionários para serem treinados de forma a minimizar o custo de instalação e o custo total com o transporte dessas viagens, sendo $n >> m$. Este tipo de problema é considerado um problema de p-medinas na otimização inteira, porém o autor reforça que o fato de particionar n localidades onde estão posicionados os escritórios em m grupos onde serão instalados os centros de treinamento é, justamente, um problema de clusterização. Além desses, outros exemplos reais e práticos são apresentados para identificar os problemas de agrupamento presentes em modelos de otimização combinatória e

ele também exemplifica o motivo pelo qual é importante ser cuidadoso com relação à modelagem na programação matemática especificando os modelos de clusterização.

BUDKA [31] retrata em seu trabalho a importância da escolha de forma apropriada da função objetivo fazendo referência ao agrupamento como um problema de otimização. De acordo com ele, a função objetivo do problema, que é modelada pelo pesquisador, não deve ser eleita de maneira arbitrária a partir de inúmeras medidas existentes na literatura, mas que, em cada aplicação, ela deve ser escolhida ou projetada de forma a compreender as propriedades e características do problema aplicado em questão. O autor propõe esta reflexão com o propósito de avaliar como a escolha consciente deste critério influencia na melhora do processo de análise qualitativa dos resultados na clusterização vista como um problema de otimização.

OLIVEIRA e STEWART [32] que fazem uma aplicação do problema de agrupamento na bioinformática, também utilizam a otimização como abordagem. Segundo eles, a vantagem de enunciar um problema de cluster como um problema de otimização é a facilidade de entendimento quando se tem sempre uma estrutura de função objetivo que vai medir a "qualidade" dos resultados e procura-se aplicar algum método que encontre um minimizador para esta função objetivo. Por outro lado, eles também relatam que a grande dificuldade desse processo e, principalmente, dessa abordagem na otimização, está no fato da maioria dos problemas de otimização discreta serem muito difíceis de resolver [7].

Em seu trabalho, XAVIER [8] também aborda o problema de cluster como um problema de otimização e, além de apresentar um novo método para a resolução do problema, que é chamado de Suavização Hiperbólica, ele também estabelece uma reformulação para o problema que faz parte desta técnica. Nós apresentamos esta reformulação com mais detalhes na Seção 3.1.3. O método apresentado por ele foi gerado através de uma adaptação do método de Penalidade Hiperbólica também introduzido pelo autor XAVIER [33]. Neste primeiro artigo apresentado por XAVIER [8], ele toma como base para sua reformulação o problema de agrupamento pela soma mínima das distâncias ao quadrado ou *Minimum Sum of Squares Clustering* (MSSC), utilizando, neste primeiro momento, a distância euclidiana como medida de proximidade, mas ele deixa estabelecido que a técnica poderia ser aplicada em outras formulações do problema de cluster. Mais adiante, ele apresenta uma segunda versão do seu método em que faz uma adaptação na metodologia particionando os conjuntos de observações em duas partes não sobrepostas, uma caracterizada por pontos de frenteira os quais estão próximos de dois ou mais centróides e a outra caracterizada por pontos gravitacionais em que eles estão, consideravelmente, mais próximos de um único centróide em comparação aos demais. Segundo o autor, esta

alternativa mostra-se mais rápida em termos de tempo computacional comparada com a primeira e em [34] esta aplicação também é feita no problema de agrupamento pela soma mínima das distâncias euclidianas ao quadrado. Já em [35] ele utiliza o mesmo procedimento para resolver o problema, porém usando a métrica de Manhattan ou L1 como medida de proximidade. Mais tarde, os autores XAVIER e XAVIER [36] nomeiam esta segunda versão da técnica como método de Suavização Hiperbólica Acelerado e, neste trabalho, a técnica é utilizada para resolver o problema de Fermat-Weber com múltiplas origens, que é uma outra modelagem matemática para o problema de agrupamento. O mesmo autor participou de um outro trabalho, ainda nesta linha de estudo, propondo uma abordagem incremental ao seu algoritmo, além de, neste caso, gerar pontos de partida através de uma função de cluster auxiliar. Essa abordagem é feita para o problema da soma mínima das distâncias euclidianas ao quadrado por BAGIROV *et al.* [37].

Em todos esses trabalhos citados, o problema de clusterização foi abordado como um problema de otimização, sendo resolvidos por diferentes modos, seja por métodos exatos ou heurísticos, mas em todos eles foram utilizados conceitos e modelagens da programação matemática, isto é, foram abordados via otimização.

2.2 Algoritmos Utilizados

Nesta seção, nós iremos citar algumas propostas de soluções já utilizadas na literatura, com o objetivo de abordar, de maneira breve, de que forma o problema em questão foi resolvido ao longo do tempo por trabalhos realizados. Para facilitar o entendimento, iremos dividir os algoritmos em Hierárquicos e Particionais.

A escolha do algoritmo a ser utilizado é tão importante quanto a forma como o problema vai ser modelado. Em cada aplicação, é preciso observar suas características para que as escolhas possam ser tomadas de maneira consciente ou, pelo menos, da melhor maneira possível. Existem diversas decisões a serem tomadas quando se trata de um problema de agrupamento e todas elas influenciarão, de alguma forma, nos resultados finais os quais deverão ser interpretados pelo pesquisador. São elas: o critério, as restrições, as características dos objetos a serem consideradas, a medida de proximidade e o algoritmo a ser aplicado [2].

Os algoritmos do tipo hierárquico que são citados de maneira recorrente são: algoritmos hierárquicos aglomerativos e algoritmos hierárquicos divisivos.

Os algoritmos hierárquicos Aglomerativos - *Agglomerative Hierarchical Clustering Algorithms* (AGNES) têm início com uma matriz de distâncias, que pode ser escolhida entre as medidas de proximidade citadas aqui no trabalho, por exemplo.

Nesta matriz contém a distância entre todos os objetos dados. Neste método, inicialmente, todos os objetos pertencem a um cluster, isto é, temos o mesmo número de clusters e objetos no início do algoritmo e, ao longo do processo, a ideia é mesclá-los iterativamente até obter um único cluster com todos os objetos ou até um determinado critério de parada designado pelo pesquisador. Seus principais passos consistem em identificar a menor distância entre dois clusters na matriz e juntá-los formando um novo cluster que os contém, para isso é necessário atualizar a matriz de distâncias a cada iteração ou junção feita. Para atualizar a matriz de distâncias, é utilizada uma das 3 estratégias de ligações já mencionadas anteriormente, ligação única, completa ou média, uma vez que a matriz atualizada precisa conter a distância entre o novo cluster (gerado pela junção corrente) e todos os demais do problema. Os algoritmos hierárquicos divisivos *Divisive Hierarchical Clustering Algorithm* (DIANA) são considerados o oposto, eles iniciam com todos os objetos pertencendo a um único cluster e fazem o processo de identificar as maiores diferenças entre os objetos para que seja possível separá-los em clusters diferentes até que cada objeto esteja disposto em um único cluster ou até que seja designado um critério de parada fixado pelo pesquisador. Estes métodos resolvem problemas de agrupamento de maneira aproximada e eles podem ser encontrados nos trabalhos de HANSEN e JAUMARD [2] e XU e WUNSCH [9].

Na literatura, podemos encontrar diversos tipos de algoritmos de particionamento, sendo eles heurísticos, meta-heurísticos ou exatos. Os dois primeiros resolvem o problema de maneira aproximada, eles buscam mínimos locais de boa qualidade. Já o terceiro tem o objetivo de resolver o problema de maneira exata, buscando o mínimo global do problema. Quase sempre encontrar o mínimo (ou máximo) global de um problema de otimização com dados reais é uma tarefa muito complexa, principalmente quando se trata de problemas NP-difíceis. O estudo de novas modelagens para o problema, que por sua vez é o objeto de pesquisa deste trabalho, é uma alternativa para tentar contornar ou auxiliar na dificuldade de encontrar respostas exatas no processo de resolução para problemas NP-difíceis como este [7].

O algoritmo *Branch and Bound* (B&B) é um exemplo de método exato de particionamento que resolve o problema através de uma enumeração implícita. Uma vez que para problemas com instâncias muito grandes torna-se improvável a resolução aplicando a enumeração explícita, o método apresenta a ideia de enumerar de forma inteligente, particionando o conjunto solução do problema em subconjuntos usando limitantes (bounds) superiores e inferiores para "podar" os nós da árvore de enumeração de maneira a não explorar todas as combinações

possíveis.

Outro exemplo de método de particionamento exato é o algoritmo de *Branch and Cut*, que é uma extensão do método de B&B. Ele gera planos de corte que consistem em inequações ou desigualdades válidas para o problema reduzindo o espaço viável de soluções da relaxação contínua com o objetivo de se aproximar da envoltória convexa dos pontos inteiros viáveis. Esse processo reduz o tamanho da árvore de enumeração do *branch and bound*.

Uma outra alternativa utilizada na literatura e que também é um algoritmo de particionamento exato é o algoritmo de Geração de Colunas, muito utilizado em problemas de programação inteira que dispõem de uma região viável como um conjunto que pode ser escrito como a interseção de dois ou mais conjuntos, sendo introduzida a ideia de decomposição para a resolução de um problema mestre e seu subproblema. Na programação linear, o algoritmo de geração de colunas é, geralmente, usado quando o problema possui um número exponencial de variáveis. A combinação da técnica de geração de colunas com o algoritmo de B&B resulta no algoritmo chamado *Branch and Price*.

Estas abordagens de algoritmos exatos utilizados no contexto de clusterização podem ser encontradas nos trabalhos de HANSEN e JAUMARD [2], ALOISE e HANSEN [3], DUONG *et al.* [13], GRÖTSCHEL e WAKABAYASHI [38] e no contexto geral dos algoritmos no WOLSEY [39].

Algoritmos particionais que resolvem de forma aproximada também foram bastante utilizados para resolver o problema de agrupamento. Um exemplo deles é o *Particle Swarm Optimization* (PSO) que é uma meta-heurística populacional inspirada na natureza que se baseia na movimentação de um enxame de partículas (bando ou cardume) que caminham (voam) juntas em direção ao objetivo (alimento), que seria a solução. O algoritmo descreve a evolução de um conjunto de pontos no espaço de busca fazendo um balanço entre o comportamento individual e coletivo das partículas e o que rege esse movimento é uma equação de velocidade. A solução do problema é mensurada por uma função chamada de função *fitness*. Esse tipo de abordagem podemos encontrar no trabalho de CUI *et al.* [15].

Um exemplo de heurística muito utilizada na literatura é o *K-means*. Nele, a variável k indica o número de agrupamentos pré-definido e seu processo inicia a partir da escolha de k centroides, que inicialmente é aleatória. Porém, após o cálculo da distância de todos os pontos dados a cada um dos k centroides e

da atribuição desses pontos ao centroide mais próximo, há uma atualização da posição dos centroides tomando a média de todos os pontos que foram atribuídos a eles. Esse processo se repete até que, após a atualização do posicionamento dos centroides, nenhum ponto dado seja redirecionado a outro agrupamento, isto é, a algum centroide diferente. Por conseguinte, um ponto de mínimo local é encontrado, o que é chamado de ponto de convergência. Nos trabalhos de CUI *et al.* [15], XU e WUNSCH [9], HANSEN e JAUMARD [2], DUONG *et al.* [13] podemos encontrar esta abordagem e também existem variações dele na literatura, como por exemplo a que é apresentada no artigo de BAGIROV [40] .

Um outro algoritmo utilizado para resolver problemas de agrupamento e que retorna soluções aproximadas ou soluções primais ou ainda ótimos locais é o método de Suavização Hiperbólica desenvolvido no trabalho de XAVIER [8] e que mencionamos anteriormente. Na Subseção 3.1.3, apresentamos com mais detalhes o modelo Min-Sum-Min que é vinculado a este método.

Um outro método utilizado e abordado nos trabalhos de COLE [5] e HRUSCHKA e EBECKEN [4] é o algoritmo genético ou *Genetic Algorithm* (GA). É uma meta-heurística populacional baseada na recombinação genética e mutação em que para uma população inicial são aplicados operadores de cruzamento, responsáveis por contribuir na convergência do algoritmo para soluções boas, e operadores de mutação que contribuem para a diversificação de soluções, fazendo com o que o espaço de soluções seja mais explorado. Esses operadores de cruzamento e mutação são responsáveis por explorar a população expandindo o espaço de busca e são aplicados a cada iteração.

O artigo de SANTI *et al.* [41] aplica o método *Variable neighborhood search* (VNS) no problema de agrupamento. Ele é um método de descida (para problemas de minimização) com vizinhança variável, isto é, uma meta-heurística que explora o espaço de soluções através de trocas de vizinhança e sempre que há uma melhora na solução, volta à primeira estrutura de vizinhança. O método inicia com uma solução inicial gerada por um método construtivo e, a partir disso, seleciona uma nova solução vizinha ou, simplesmente, vizinho e faz uma busca local. Caso não haja melhora no valor da função aplicada à nova solução vizinha, o processo é repetido para uma vizinhança "maior" ou que explore ainda mais o espaço de busca até que o critério de parada seja satisfeito.

Nós podemos encontrar no trabalho de AL-SULTAN [42] uma abordagem sobre a Busca Tabu ou *Tabu Search* aplicada ao problema de agrupamento. Essa

meta-heurística equivale a uma busca local dinâmica com um processo adaptativo que visa contornar o problema de não haver soluções vizinhas aprimorantes e evitar que a busca retorne a um ótimo local já visitado durante o processo. Para que isso ocorra, o método utiliza estruturas flexíveis de memória que armazenam informações sobre o espaço de busca. Neste algoritmo, uma lista de movimentos proibidos é criada com a intenção de evitar que as soluções entrem em um ciclo. Ela é chamada de lista tabu e possui um tamanho determinado, mas esta lista acaba tornando o processo bem restritivo. E, por esse motivo, ela é balanceada por uma função de aspiração que libera somente alguns movimentos para que o espaço de busca seja bem explorado.

Em AL-SULTANA e KHAN [43], além da heurística *K-means* e das meta-heurísticas Busca Tabu e Algoritmo Genético já mencionadas anteriormente, os autores também aplicam o algoritmo *Simulated Annealing* (SA), que é um método inspirado no recozimento físico, em que um sólido é aquecido além do seu ponto de fusão e depois resfriado de forma lenta e cuidadosa para que se atinja o estado de baixa energia. Neste algoritmo, a função custo (objetivo ou *fitness*) desempenha o papel da energia e deseja-se melhorá-la a cada iteração até chegar ao ponto de energia mínima, isto é, o valor mínimo da função objetivo (para problemas de minimização). Porém, ao longo do processo, são aceitos rearranjos de piora à função custo para que a "temperatura", que é um parâmetro com valor inicial alto do algoritmo, não atinja o valor zero tão rapidamente. Isso é feito para contornar a possibilidade do SA ficar preso em um ótimo local de baixa qualidade. Esses rearranjos de piora provêm do algoritmo de metrópolis, em que as pioras são aceitas por uma função de aceitação probabilística. O parâmetro "temperatura" inicia com um valor alto, isto implica numa maior probabilidade de aceitar movimentos de piora, porém ele tende a zero para chegar no estado de baixa energia e, conforme diminui, a probabilidade de aceitar soluções de piora também diminui. No final do algoritmo, os movimento de piora quase não são mais aceitos e ele tem fim quando a temperatura chega bem próximo de zero e nenhuma solução de piora é aceita, então chega-se a um ótimo local.

Capítulo 3

Metodologia

O presente trabalho busca fazer um estudo sobre o problema de agrupamento, que é um tema bastante explorado, conhecido e com uma extensa variedade de assuntos tratados na literatura. O nosso objetivo, principalmente neste capítulo, é apresentar algumas modelagens matemáticas utilizadas na formulação do problema de clusterização já descritas, além de enunciar o processo de duas novas modelagens para o problema apresentadas recentemente em [44]. Essas novas formulações são os principais objetos de estudo deste trabalho. O processo de desenvolvimento desses novos modelos tem seu ponto de partida em modelos clássicos já descritos na literatura, são eles: o agrupamento pela soma mínima do quadrado das distâncias euclidianas ou *Min-Sum-of-Squares Clustering* que é uma abreviação de *Minimum Sum of Squares Clustering Problem* (MSSC) encontrado nos trabalhos de ALOISE e HANSEN [3], DU MERLE *et al.* [17], BAGIROV [40] e o Problema de Weber ou ainda *Weber Problem* (WP) que pode ser visto no artigo de CHEN *et al.* [45] e que foi proposto incluindo pesos na função objetivo. Em particular, é usado o Problema de Weber com Peso Unitário, este também é conhecido como *Unitary Weighted Weber Problem* (UWWP) que apesar de não ter sido tão explorado, é utilizado no processo e será apresentado a seguir, assim como as demais formulações.

3.1 Modelos Utilizados

A principal motivação destas novas modelagens é obter formulações que contenham relaxação contínua convexa e diferenciável para o problema, uma vez que o problema e, consequentemente suas respectivas formulações matemáticas já descritas na literatura e que serão apresentadas aqui, obtêm relaxações contínuas não convexas e não diferenciáveis. Nesta seção, apresentamos alguns aspectos das áreas de convexidade e diferenciabilidade e, de maneira simples e objetiva, mostramos algumas motivações pelas quais buscamos trabalhar com relaxações contínuas convexas e diferenciáveis no campo da otimização, e em especial, neste

trabalho.

O estudo sobre convexidade é muito latente no campo da otimização, uma vez que suas propriedades são muito úteis e favoráveis quando estamos lidando com modelos convexos, isto é, compostos por função objetivo e região viável convexas. Uma dessas propriedades menciona que, na otimização convexa, qualquer ótimo local encontrado é também o ótimo global para o problema, e isso, de fato, é uma característica extremamente favorável quando estamos buscando a solução de algum problema. Por outro lado, um modelo não convexo não possui esta mesma facilidade, o que pode fazer com que seja bem mais difícil encontrar a solução ótima global, principalmente quando a região viável deste problema possui uma grande quantidade de mínimos locais. A Figura 3.1 ilustra bem uma região factível não convexa composta por diversos mínimos locais. Através dela podemos ter a ideia da dificuldade que é percorrer um espaço de busca com estas características de forma a chegar no mínimo global passando por diversos mínimos locais e/ou tentando contorná-los no trajeto. A definição de conjunto convexo, que é a mesma de um espaço viável convexo para este contexto, pode ser explicada através de uma interpretação geométrica da seguinte forma: se traçarmos uma reta entre dois pontos pertencentes a um conjunto convexo, todo ponto pertencente a este segmento de reta deve necessariamente pertencer ao mesmo conjunto como podemos observar na Figura 3.2. Nos trabalhos de LUENBERGER *et al.* [46] e PERESSINI *et al.* [47] é possível encontrar ainda mais detalhes sobre este assunto.

Além da convexidade, a diferenciabilidade também pode ser considerada uma característica bastante favorável em modelos de otimização, principalmente pelo fato de ser uma base de resultados para o desenvolvimento e execução de algoritmos que se apoiam no cálculo diferencial e utilizam artifícios como derivadas primeira e segunda (gradiente e hessiana) no processo de resolução desses problemas de otimização não lineares. O trabalho de FLETCHER [48] aborda as áreas da otimização diferenciável e não diferenciável. Logo, quando se trata de um modelo não diferenciável, isso significa que nem todos os pontos do domínio de função possuem derivada, e que, consequentemente, existem pontos na região viável do problema que não são considerados nos cálculos para esse tipo de "algoritmo clássico que utiliza derivada no ponto". A dificuldade é que esses pontos não "cobertos" ou não considerados podem ser exatamente os pontos de mínimo ou de máximo da função, o que torna a resolução deste tipo de problema mais complexa e restrita aos algoritmos que não se baseiam no cálculo diferencial. Para ilustrar essa não diferenciabilidade no ponto de mínimo, podemos usar como exemplo a função módulo: $f(x) = |x|$, com $x \in R$. Ela não é diferenciável na origem e este ponto é,

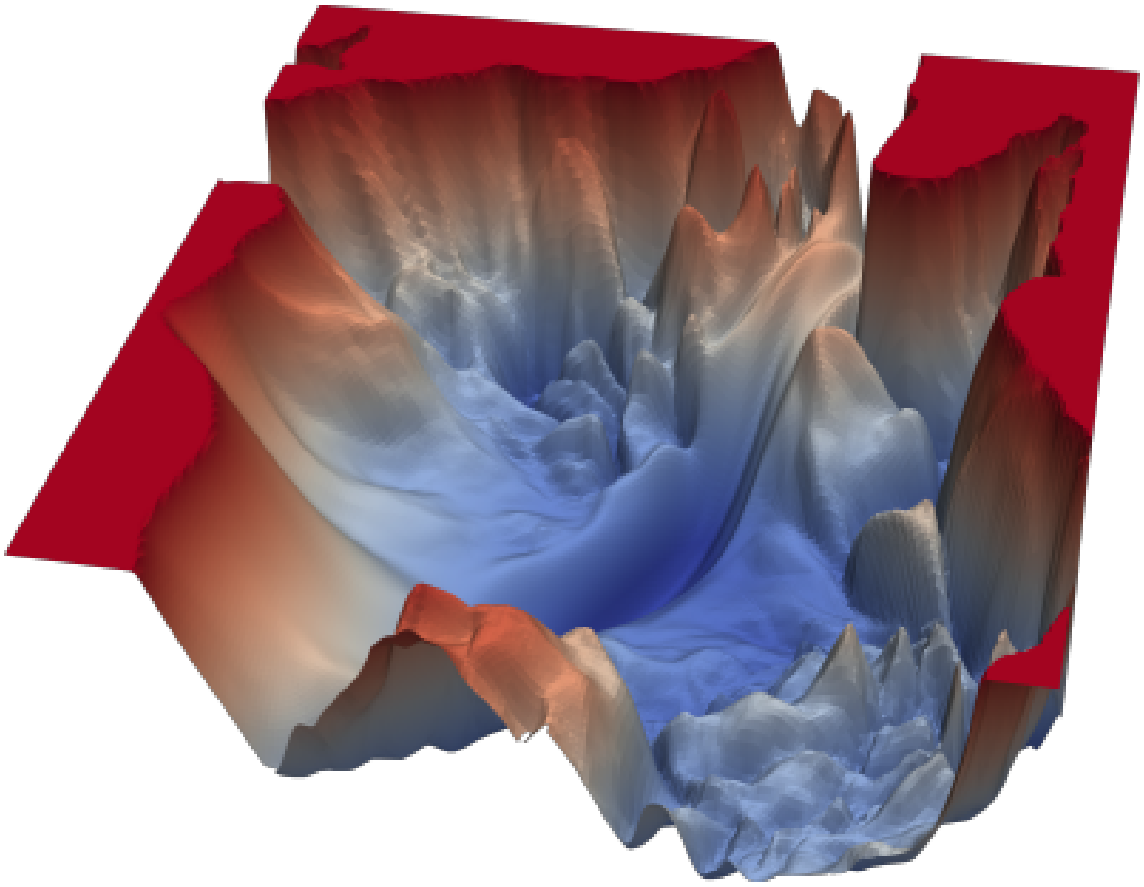

Figura 3.1: Ilustração de um Espaço de Busca em 3D Não Convexo com Diversos Mínimos Locais. Fonte da imagem: LI *et al.* [1] .

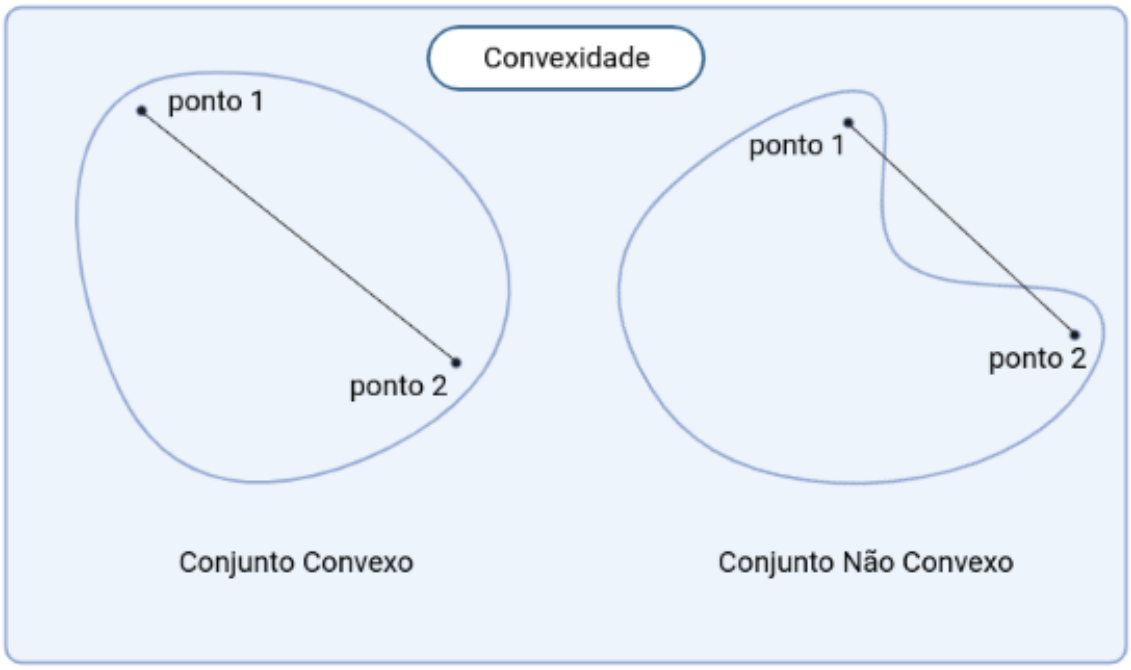

Figura 3.2: Ilustração de Convexidade. Fonte da imagem: Criada pela autora no aplicativo BioRender.

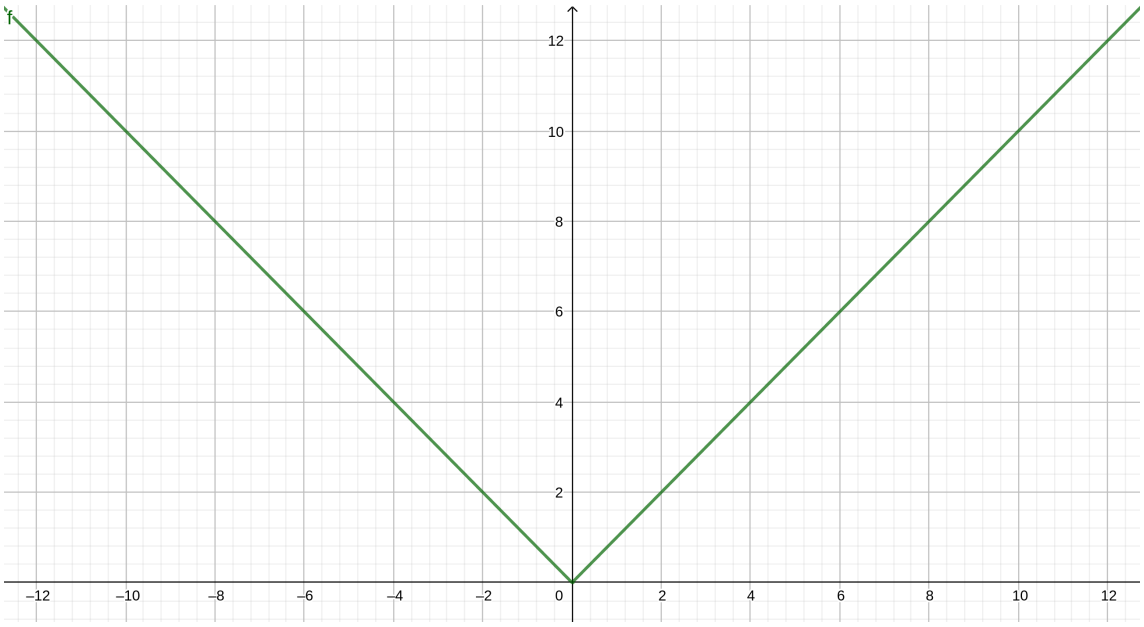

Figura 3.3: Ilustração da Função Módulo. Fonte da imagem: Criada pela autora no aplicativo GeoGebra.

justamente, o mínimo da função, conforme podemos visualizar na Figura 3.3 .

Conseguir contornar a não diferenciabilidade e não convexidade em modelos de otimização estudados, por vezes, é uma tarefa bastante complexa, porém muito requisitada, uma vez que se for possível contorná-las, também é possível ter acesso aos resultados que as áreas da convexidade e da diferenciabilidade proporcionam ao campo da otimização. Lembrando que o fato de conseguir trabalhar com relaxações contínuas convexas e diferenciáveis do problema, não faz com que ele se torne um problema fácil, mas nos permite explorar resultados e métodos já fornecidos pela literatura para problemas com essas características.

3.1.1 Problema de Weber com Peso Unitário

O problema de Weber tem como objetivo minimizar a soma total das distâncias entre os pontos e seus respectivos centros, uma vez que cada um dos pontos é atribuído a um único cluster, por consequência cada um deles também é atribuído à um único centroide.

Considere um conjunto de p pontos, sendo $a^i \in \mathbb{R}^n$ a localização dada de cada ponto, com $i \in P = \{1, 2, 3, ..., p\}$. Em outras palavras, a^i é a localização do i-ésimo ponto dentre os p pontos do problema. Definimos as variáveis $x^j \in \mathbb{R}^n$ como a localização do centro do cluster C_j a ser encontrada, com $j \in K = \{1, 2, 3, ..., k\}$, ou seja, x^j é a localização do centro (centroide) do j-ésimo cluster a ser encontrada

dentre os k clusters existentes no problema. Considerando p,$k \in \mathbb{N}$ e $p \gg k$.

O intuito é encontrar a localização (as coordenadas) de x^j e quais pontos a^i pertencem a cada cluster C_j. Para isso, usamos as variáveis binárias y_{ij} que nos indicarão este pertencimento da seguinte forma: $y_{ij} = 1$, se o ponto i pertence ao cluster j, ou $y_{ij} = 0$, caso contrário. Logo, o modelo matemático pode ser escrito desta forma:

$$(UWWP): \qquad \text{Minimizar} \quad \sum_{i=1}^{p} \sum_{j=1}^{k} \|a^i - x^j\| \, y_{ij} \tag{3.1}$$

$$\text{sujeito a :} \quad \sum_{j \in K} y_{ij} = 1, \quad \forall i \in P \tag{3.2}$$

$$x^j \in \mathbb{R}^n, \quad \text{para todo } j \in K \tag{3.3}$$

$$y_{ij} \in \{0, 1\}, \quad \forall\, i \in P, \forall\, j \in K \tag{3.4}$$

A Equação 3.1 é a função objetivo do problema que minimiza a distância entre os pontos e seus respectivos centros. A norma encontra-se multiplicada pela variável y_{ij} pois somente será contabilizada a distância válida referente ao ponto i que pertencer ao cluster j, isto é, as variaveis y_{ij} iguais a 1. A Restrição 3.2 indica que todo ponto i deverá pertencer a um único cluster j e as demais Restrições 3.3 e 3.4 denotam o domínio das variáveis em questão.

Neste caso, a distância do ponto a^i ao seu respectivo centro x^j do cluster C_j é dada pela norma como pode ser visto em 3.1, que pode ser escrita também desta maneira:

$$d_{ij} = ||a^i - x^j||_2 = \sqrt{\sum_{l=1}^{n} (a_l^i - x_l^j)^2}, \quad \forall\, i \in P, \forall\, j \in K,$$

de acordo com a definição de norma 2.

3.1.2 Critério de Mínima Soma de Quadrados

O problema de agrupamento pela soma mínima das distâncias ao quadrado é bastante utilizado na literatura e, apesar de ser muito parecido com o (UWWP)

abordado na subseção anterior, eles se diferem na função objetivo. Essa diferença pode parecer mínima quando visualizamos a modelagem de ambos, porém acarreta uma grande diferença na prática. Essa diferença será mostrada mais adiante na seção de resultados. Esta formulação utilizando como critério a soma mínima das distâncias quadráticas apresenta um problema matemático diferenciável e não convexo com um extensivo número de mínimos locais, conforme abordado por BAGIROV e YEARWOOD [49], além de ser um problema NP-difícil, como também é descrito por BRUCKER [50].

A formulação matemática e definição de variáveis usadas no modelo de *Minimum Sum of Squares Clustering Problem* é apresentada a seguir. Ela também foi abordada por ALOISE e HANSEN [3], DU MERLE *et al.* [17], BAGIROV [40].

Dado um conjunto de p pontos e k clusters, sendo $a^i \in \mathbb{R}^n, \forall i \in P = \{1, 2, 3, .., p\}$ a localização dada de cada ponto, $x^j \in \mathbb{R}^n, \forall j \in K = \{1, 2, 3, .., k\}$ representa a localização do centro do cluster C_j a ser encontrada e y_{ij} é a variável de decisão binária que indica o pertencimento de um ponto i a um cluster j. O propósito é encontrar a melhor localização dos centroides e descobrir quais pontos pertencem a cada cluster de forma a minimizar a distância entre o ponto e o centro do cluster atribuído a ele.

$$(MSSC): \qquad \text{Minimizar} \quad \sum_{i=1}^{p} \sum_{j=1}^{k} \|a^i - x^j\|^2 \, y_{ij} \tag{3.5}$$

$$\text{sujeito a} : \quad \sum_{j \in K} y_{ij} = 1, \quad \forall i \in P \tag{3.6}$$

$$x^j \in \mathbb{R}^n, \quad \text{para todo } j \in K \tag{3.7}$$

$$y_{ij} \in \{0, 1\}, \quad \forall\, i \in P, \forall\, j \in K \tag{3.8}$$

A Equação 3.5 faz referência à função objetivo do problema que, como o próprio nome sugere, tem o objetivo de minimizar a soma das distâncias euclidianas ao quadrado. Nesta equação, a norma quadrática é multiplicada pela variável binária y_{ij} com o intuito de somente serem contabilizadas as distâncias em que um ponto i pertença à um determinado cluster j, isto é, quando y_{ij} for igual a 1. A Restrição 3.6 implica que todo ponto i somente pode pertencer a um único cluster j e as Restrições

3.7 e 3.8 indicam o domínio das variáveis do problema, sendo x^j pertencente ao espaço n-dimensional de maneira geral, (uma forma de visualizar de maneira mais didática é se, eventualmente, o problema estiver no R^2 por exemplo, a variável x^j será um vetor com duas coordenadas, ou seja o tamanho de x^j vai depender do espaço indicado em cada aplicação, no entanto ele está generalizado como R^n no modelo) e y_{ij} binário.

3.1.3 Min-Sum-Min

Nesta subseção, nós iremos mostrar a reformulação matemática para o problema de agrupamento utilizada no método de Suavização Hiperbólica apresentado inicialmente no trabalho de XAVIER [8], mas que instituiu diversos outros trabalhos voltados para a aplicação desta técnica de resolução em problemas com características não diferenciáveis, como os trabalhos de XAVIER e XAVIER [34], XAVIER *et al.* [35], XAVIER e XAVIER [36], BAGIROV *et al.* [37].

Essa técnica foi desenvolvida através de uma adaptação do método de penalização hiperbólica que foi introduzido por XAVIER [33]. A ideia é fazer uma aproximação de problemas não diferenciáveis para problemas completamente diferenciáveis de classe C^∞. A solução é dada por uma sequência de subproblemas diferenciáveis restritos que se aproximam da solução do problema original. Estes subproblemas restritos são reduzidos à problemas de otimização irrestritos através do Teorema da Função Implícita que é apresentado no trabalho de XAVIER [8] o qual tem por objetivo demonstrar esta reformulação de forma a validá-la. E por se tratar de subproblemas completamente diferenciáveis, eles podem ser resolvidos de maneira eficiente por métodos preponderantes na literatura como o de Newton, Quase-Newton ou Gradiente Conjugado por exemplo.

A reformulação que iremos descrever aqui tem como base o critério da soma mínima das distâncias ao quadrado, que, neste caso, foi utilizado um modelo não diferenciável e não convexo. Porém, esta técnica pode ser aplicada em diversas outras formulações matemáticas não diferenciáveis existente na literatura que modelam o problema de agrupamento, conforme pode ser encontrado em [35] [36].

Dado um conjunto $P = \{a^1, a^2, ..., a^p\}$ com p pontos num espaço euclidiano n-dimensional, esse conjunto deve ser dividido em k agrupamentos, sendo k um número pré-fixado de grupos disjuntos. Os centroides desses agrupamentos são representados por x^j, $j = \{1, 2, ..., k\}$ com $x^j \in \mathbb{R}^n$, e o conjunto de coordenadas deles será representado por $X \in \mathbb{R}^{nk}$. O pontos a^i são dados e a menor distância entre cada ponto e o centroide x^j mais próximo é calculada e dada por :

$$z_i = \min_{x^j \in X} ||a^i - x^j||_2. \tag{3.9}$$

Uma medida que é intitulada pelo trabalho de XAVIER [8] por *medida de qualidade* é fornecida pela soma dos quadrados das distâncias entre os pontos e os centros mais próximos dos agrupamentos, os quais eles pertencem, é dada por:

$$D(X) = \sum_{i=1}^{p} z_i^2. \tag{3.10}$$

O problema busca encontrar a localização ótima dos centroides e, dessa maneira, também encontrará a melhor *medida de qualidade* para o problema, isto é, o melhor valor para a "função objetivo". A localização ótima para o conjunto de coordenadas dos centroides é dado por:

$$X^* = \arg \min_{X \in \mathbb{R}^{nk}} \; D(X) \tag{3.11}$$

onde X é o conjunto das localizações (coordenadas) de todos os centroides dos k agrupamentos. Logo, usando as Equações 3.9 e 3.11, obtêm-se a expressão que descreve a min-sum-min em que o argumento dessa expressão é o resultado ideal para o problema a ser resolvido.

$$X^* = \arg \min_{X \in \mathbb{R}^{nk}} \sum_{i=1}^{p} \min_{x^j \in X} ||a^i - x^j||_2^2 \tag{3.12}$$

Em que a Equação 3.12 pode ser substituída pelo problema de programação matemática equivalente a seguir:

$$\text{Minimizar} \quad \sum_{i=1}^{p} z_i^2 \tag{3.13}$$

$$\text{sujeito a :} \quad z_i = \min_{j=1,..,k} ||a^i - x^j||_2, \quad i = 1, ..., p$$

onde o objetivo é minimizar z_i^2 sendo sujeito a definição de z_i, que por sua vez, é a distância mínima entre cada ponto a^i com $i = \{1, 2, ..., p\}$ e centroide do grupo mais próximo x^j com $j = \{1, 2, ..., k\}$. Levando em consideração a definição de z_i, que é a mínima distância conforme descrito na Equação 3.9, então temos as seguintes desigualdades satisfeitas:

$$z_i - ||a^i - x^j||_2 \leq 0, \qquad j = 1, 2, ..., k. \tag{3.14}$$

Logo, substituindo as igualdades pelas desigualdades válidas 3.14 no modelo 3.13, teremos a formulação relaxada a seguir:

$$\text{Minimizar} \quad \sum_{i=1}^{p} z_i^2 \tag{3.15}$$

$$\text{sujeito a :} \quad z_i - ||a^i - x^j||_2 \leq 0, \quad i = 1, ..., p, \ j = 1, ..., k.$$

A partir desta formulação, podemos analisar que, com esta substituição, a variável z_i não foi limitada inferiormente, o que ocasiona $z_i = 0$ solução possível e ótima para o problema e, consequentemente, a solução do problema ou medida de qualidade também valerá zero, o que não contribui na resolução do mesmo. Isso mostra que o Modelo 3.15 não equivale ao Modelo 3.13, portanto serão feitas mudanças no Modelo 3.15 para que seja obtida uma equivalência entre as formulações.

A primeira modificação é feita introduzindo a função $\varphi(y) = \max\{0, y\}$, e tendo em vista que a Equação 3.14 é válida, verificamos o que ocorre somente aplicando a função:

$$\sum_{j=1}^{k} \varphi(z_i - ||a^i - x^j||_2) = 0, \qquad i = 1, ..., p. \tag{3.16}$$

Uma vez que z_i continua não limitado (livre) inferiormente, ocorre que $z_i = 0$ continua viável e, por consequência disso, o valor máximo da função introduzida recairá sempre no valor igual a zero. Para contornar esta situação, a igualdade na Equação 3.16 é substituída pelo sinal de maior "$>$" fazendo com que z_i seja limitado inferiormente não podendo valer zero, mas sim o menor valor possível positivo. Para isso também é levado em conta que a função objetivo faz com que z_i obtenha o mínimo valor plausível, pois é uma função de minimização. Isso resulta no problema a seguir:

$$\text{Minimizar} \quad \sum_{i=1}^{p} z_i^2 \tag{3.17}$$

$$\text{sujeito a :} \quad \sum_{j=1}^{k} \varphi(z_i - ||a^i - x^j||_2) > 0, \quad i = 1, ..., p$$

Para a obtenção de um problema com desigualdades não estritas, a Formulação 3.17 é modificada de maneira que a expressão restritiva do modelo é perturbada, fazendo com que o modelo modificado resulte em:

$$\text{Minimizar} \quad \sum_{i=1}^{p} z_i^2 \tag{3.18}$$

$$\text{sujeito a :} \quad \sum_{j=1}^{k} \varphi(z_i - ||a^i - x^j||_2) \geq \varepsilon, \quad i = 1, ..., p$$

onde $\varepsilon > 0$. Fazendo esta aproximação, tem-se que quando $\varepsilon \to 0_+$ (tende a zero), o conjunto viável da Formulação 3.17 é o limite do conjunto viável do problema modificado 3.18. Logo, é possível solucionar o Modelo 3.17 resolvendo uma sequência de problemas no formato de 3.18 com uma sequência valores decrescentes para ε que se aproximam de 0, conforme apresentado em [8].

3.2 Modelos Propostos

Os principais objetos de estudo deste trabalho serão apresentados nesta seção, são eles: duas reformulações para o problema de agrupamento, mais especificamente, para os modelos do Problema de Weber com Peso Unitário e o Problema da Soma Mínima das Distâncias Quadráticas. O modelo (MSSC) possui relaxação contínua não convexa e o modelo (UWWP) possui relaxação contínua não convexa e não diferenciável. A proposta das duas novas abordagens é reformulá-los de maneira que os modelos passem a ter relaxação contínua convexa e diferenciável.

Iniciaremos com a reformulação do Problema de Weber com Peso Unitário apresentado na Seção 3.1.1. De acordo com a definição do modelo já apresentado anteriormente, temos que a norma na função objetivo equivale a

$$||a^i - x^j||_2 = \sqrt{\sum_{l=1}^{n}(a_l^i - x_l^j)^2}, \quad \forall\, i \in P,\ \forall\, j \in K,$$

e que a variável y_{ij} é binária, ou seja $y_{ij} \in \{0, 1\}$, o que nos permite dizer que $y_{ij}^2 = y_{ij}$ e ainda que $\sqrt{y_{ij}} = y_{ij}$. Logo, é possível fazer a seguinte manipulação algébrica: substituímos a variável y_{ij} pela raiz da mesma ao quadrado sabendo que isso não irá interferir no valor dela, isso ocorre da seguinte forma:

$$\begin{aligned}
\|a^i - x^j\|\, y_{ij} &= y_{ij} \sqrt{\sum_{l=1}^{n}(a_l^i - x_l^j)^2} \\
&= \sqrt{y_{ij}^2} \sqrt{\sum_{l=1}^{n}(a_l^i - x_l^j)^2} \\
&= \sqrt{\sum_{l=1}^{n} y_{ij}^2\,(a_l^i - x_l^j)^2} \\
&= \sqrt{\sum_{l=1}^{n} m_{ijl}^2}
\end{aligned}$$

No último passo, temos uma mudança de variável e, com base nesta mudança, podemos adaptar a função objetivo de (UWWP) 3.1 em função de m_{ijl}:

$$\text{Minimizar} \quad \sum_{i=1}^{p}\sum_{j=1}^{k} \|a^i - x^j\|\, y_{ij} \quad = \quad \sum_{i=1}^{p}\sum_{j=1}^{k} \sqrt{\sum_{l=1}^{n} m_{lij}^2}$$

definindo a varável m_{ijl} como:

$$m_{ijl} = y_{ij}\ (a_l^i - x_l^j), \quad \forall i \in P,\ \forall j \in K,\ \forall l \in N,$$

onde $N = \{1, 2, ..., n\}$.

É possível verificar que a igualdade acima, que define m_{ijl}, é uma equação não-linear e não convexa, então ao invés de acrescentá-la como restrição ao modelo adaptado em função da nova variável m_{ijl}, acrescentaremos as seguintes famílias de restrições mostradas a seguir pelo fato de serem lineares:

$$-(1 - y_{ij})M + (a_l^i - x_l^j) \ \leq\ m_{lij} \ \leq\ (a_l^i - x_l^j) + (1 - y_{ij})M \tag{3.19}$$

$$-M\, y_{ij} \ \leq\ m_{lij} \ \leq\ M\, y_{ij} \tag{3.20}$$

Em que M tem valor constante dado por:

$$M = \max_{1 \leq s < q \leq p} ||a^s - a^q||,$$

que é o valor da maior distância entre dois pontos pertencentes ao problema. De acordo com um resultado conhecido, os centóides pertencem à envoltória convexa dos pontos dados. Logo, a distância entre um ponto e seu respectivo centroide não vai ser maior do que a distância máxima entre dois pontos do problema.

Na Restrição 3.19, se $y_{ij} = 1$ então m_{ijl} assumirá o valor de $(a_l^i - x_l^j)$, porém se $y_{ij} = 0$, m_{ijl} estará limitada ao valor de M, estará livre.

Já na Restrição 3.20, se $y_{ij} = 0$ então $m_{ijl} = 0$, mas se $y_{ij} = 1$, o valor da distância entre o ponto e seu centroide está limitada ao valor de M, que é a maior distância entre dois pontos do problema.

Em função da variável m_{ijl} o modelo matemático é escrito da seguinte forma:

$$\text{Minimizar} \quad \sum_{i=1}^{p} \sum_{j=1}^{k} \sqrt{\sum_{l=1}^{n} m_{lij}^2}$$

$$\text{sujeito a:} \quad \sum_{j \in K} y_{ij} = 1, \qquad \text{para todo } i \in P$$

$$-(1 - y_{ij})M + (a_l^i - x_l^j) \leq \ m_{lij} \leq \ (a_l^i - x_l^j) + (1 - y_{ij})M$$

$$-M\, y_{ij} \leq \ m_{lij} \leq \ M\, y_{ij}$$

$$x^j \in \mathbb{R}^n, \qquad \text{para todo } j \in K$$

$$m_{lij} \in \mathbb{R}, \qquad \forall\, l \in N, \forall\, i \in P, \forall\, j \in K$$

$$y_{ij} \in \{0, 1\}, \qquad \forall\, i \in P, \forall\, j \in K$$

É possível ver que este novo modelo escrito em função das variáveis m_{ijl} possui as duas novas famílias de restrições lineares incluídas limitando seu espaço de soluções e definindo m_{ijl}. Todavia, a função objetivo continua sendo não diferenciável, logo será utilizada a variável z_{ij} com a intenção de contornar esta situação.

Para esta segunda mudança de variável acontecer é necessário que z_{ij} seja definida e que esta definição seja acrescentada ao modelo como uma nova restrição, da mesma forma ocorreu com m_{ijl} na primeira mudança de variável deste processo. Sendo assim, teremos:

$$z_{ij} \ = \ \sqrt{\sum_{l=1}^{n} m_{ijl}^2} \quad \geq 0$$

Como estamos lidando com distâncias e m_{ijl}^2 (ao quadrado), logo os valores serão não-negativos, então $z_{ij} \geq 0$. Além disso, se usarmos "$\geq$" no lugar da igualdade, ainda assim recairemos na igualdade pelo fato de estar minimizando z_{ij}. Ademais por serem valores não-negativos, ambos lados podem ser elevados ao quadrado, o que não interfere na desigualdade e que induz as restrições do tipo cone de segunda ordem, que é uma restrição explorada em XUE e YE [51].

$$z_{ij} \geq \sqrt{\sum_{l=1}^{n} m_{ijl}^2} \qquad \Longleftrightarrow$$

$$z_{ij}^2 \geq \sum_{l=1}^{n} m_{ijl}^2, \quad z_{ij} \geq 0$$

Logo, a primeira reformulação para o Problema de Clusterização (CP1) deste trabalho, tendo como base o Problema de Weber com Peso Unitário, pode ser escrita desta forma:

$$(CP1): \qquad \text{Minimizar} \sum_{j=1}^{k} \sum_{i=1}^{p} z_{ij}, \tag{3.21}$$

$$\text{sujeito a}: \quad z_{ij}^2 \geq \sum_{l=1}^{n} m_{lij}^2, \quad i \in P,\ j \in K, \tag{3.22}$$

$$-(1-y_{ij})M + (a_l^i - x_l^j) \ \leq \ m_{lij} \ \leq \ (a_l^i - x_l^j) + (1-y_{ij})M, \tag{3.23}$$

$$-My_{ij} \ \leq \ m_{lij} \ \leq \ My_{ij}, \tag{3.24}$$

$$z_{ij} \geq 0, \quad i \in P, \quad j \in K, \tag{3.25}$$

$$\sum_{j=1}^{k} y_{ij} = 1, \quad i \in P, \tag{3.26}$$

$$y_{ij} \in \{0\ 1\}, \quad i \in P, \quad j \in K, \tag{3.27}$$

$$x^j \in \mathbb{R}^n, \quad j \in K, \tag{3.28}$$

$$m_{lij} \in \mathbb{R}, \quad l \in N, \quad i \in P, \quad j \in K. \tag{3.29}$$

onde, as Restrições 3.23 e 3.24 estão definidas para todo $i \in P$, para todo $j \in K$ e com $l \in N$, onde $N = \{1, 2, 3, ..., n\}$, sendo P o conjunto de pontos dados, K o de clusters e l os atributos de cada ponto, a dimensão do espaço em que encontra-se o problema. As Restrições 3.22 e 3.25 são conhecidas como restrições de cone de segunda ordem e podem ser vistas no trabalho de XUE e YE [51]. A Restrição 3.26 já mencionada anteriormente, exige que cada ponto pertença a um único agrupamento e as Restrições 3.27, 3.28 e 3.29 definem o domínio das variáveis contidas no problema.

O problema visa encontrar x^j que é a localização de centro do cluster C_j juntamente com a identificação de quais pontos pertencem a cada cluster dado

pela variável de decisão y_{ij} de maneira a minimizar a distância entre eles. Se x_*^j para algum $j \in K$ é uma solução ótima para (CP1), então x_*^j é a localização ótima do centro ou controide para C_j. A relaxação contínua de (CP1) é um problema de otimização suave e convexo podendo ser resolvido por *solvers* disponíveis no mercado como o Xpress, que por sua vez funciona bem para problemas que contenham restrições do tipo cone de segunda ordem.

Uma consideração importante é que neste caso, isto é, utilizando esta reformulação tendo como base o problema de Weber, podemos fazer uma mudança de escala pré-processando os pontos dados a^i de maneira que todo ponto passe a ser $a^i := \frac{1}{M}\, a^i$, $\forall i \in P$. Então teremos:

$$|a_l^i - a_l^j| \leq 1, \quad \forall\, i, j \in P,\ i \neq j,$$

Com isso, podemos usar $M = 1$ nas Restrições 3.23 e 3.24.

Essa mudança de escala funciona somente para este caso usando (UWWP).

Dados dois pontos (vetores) tais que a distância entre eles é D, se dividirmos esses dois pontos por M, a nova distância entre eles será D/M, o que é proporcional e, de fato, considerando M a distância máxima entre dois pontos do problema, os valores de D/M (distâncias na nova escala) serão menores ou, no máximo, iguais a 1. Logo, podemos tomar M igual a 1 pois será o maior valor. Por outro lado, se fizermos esse processo utilizando (MSSC), isso não acontece e mostramos que esta mudança de variável não funciona no próximo caso que será exposto.

De maneira análoga é feita a reformulação utilizando como base o Problema da Soma Mínima das Distâncias ao Quadrado, porém sua função objetivo é elevada ao quadrado e, consequentemente, a raiz quadrada passa a não mais fazer parte da equação, o que nos permite deixá-la em função de m_{ij}, não precisando seguir para o próximo passo colocando-a em função de z_{ij}.

O Segundo Problema de Cluster (CP2) que é uma reformulção de (MSSC) é dado por:

$$(CP2): \qquad \text{Minimizar} \quad \sum_{i=1}^{p}\sum_{j=1}^{k}\sum_{l=1}^{n} m_{ijl}^2 \tag{3.30}$$

$$\text{sujeito a:} \quad \sum_{j\in K} y_{ij} = 1, \quad \forall\, i \in P, \tag{3.31}$$

$$-(1-y_{ij})M + (a_l^i - x_l^j) \;\leq\; m_{lij} \;\leq\; (a_l^i - x_l^j) + (1-y_{ij})M, \tag{3.32}$$

$$-My_{ij} \;\leq\; m_{lij} \;\leq\; My_{ij}, \tag{3.33}$$

$$y_{ij} \in \{0\ 1\}, \quad i \in P, \quad j \in K. \tag{3.34}$$

$$x^j \in \mathbb{R}^n, \quad j \in K, \tag{3.35}$$

$$m_{lij} \in \mathbb{R}, \quad l \in N, \quad i \in P, \quad j \in K. \tag{3.36}$$

onde a função objetivo 3.30 se encontra em função de m_{ij} e as Restrições 3.32 e 3.33 definem esta variável. A Restrição 3.31 indica que todo ponto do problema deve pertencer a um único agrupamento e as demais Restrições 3.34, 3.35 e 3.36 definem o domínio das variáveis y_{ij}, x^j e m_{lij} a serem encontradas.

Uma outra abordagem para o modelo (CP2) será mostrada a seguir. Trata-se de uma segunda formulação exata na qual sua relaxação contínua também é convexa, chamamos este novo modelo de (CP2-BIS), por representar o mesmo problema modelado por (CP2).

Para esta nova modelagem consideramos:

$$M = \max_{1\leq s<q\leq p} ||a^s - a^q||_2^2,$$

onde M é considerada a distância máxima entre dois pontos dados no problema. O modelo (CP2-BIS) é dado por:

$$(CP2-BIS): \qquad \text{Minimizar } \sum_{i=1}^{p} z_i, \tag{3.37}$$

sujeito a:

$$z_i \geq ||a^i - x^j||_2^2 - M(1-y_{ij}), \quad i=1,2,...,p,\ j=1,2,...,k. \tag{3.38}$$

$$\sum_{j=1}^{k} y_{ij} = 1, \quad i=1,2,...,p. \tag{3.39}$$

$$z_i \geq 0, \quad i=1,2,...,p. \tag{3.40}$$

$$x^j \in R^n, \quad j=1,2,...,k. \tag{3.41}$$

$$y_{ij} \in \{\ 0,1\ \},\ i=1,2,...p,\ j=1,2,...,k. \tag{3.42}$$

$$m_{lij} \in \mathbb{R}, \quad l=1,2,..,n, \quad i=1,2,..,p, \quad j=1,2,..,k. \tag{3.43}$$

A função objetivo 3.37 minimiza z_i que está definido na primeira Restrição 3.38, de maneira que: sendo $y_{ij}=1$, o valor de z_i corresponde ao valor da norma e, no caso contrário em que $y_{ij}=0$, ele será negativo e desconsiderado, uma vez que z_i está limitado inferiormente por zero de acordo com a terceira Restrição 3.40. A segunda restrição do problema impõe que cada ponto pertence a um único agrupamento e as Restrições 3.41 e 3.42 fazem referência ao domínio das variáveis x_j e y_{ij} que equivalem a localização dos centroides e pertencimento dos pontos aos clusters do problema, respectivamente. Além da ultima Restrição 3.43 que equivale ao domínio da variável m_{lij}.

Capítulo 4

Resultados e Discussões

4.1 Resultados

Neste capítulo, temos por objetivo mostrar os resultados encontrados, os quais utilizamos um conjunto de instâncias que propusemos, além de outras que podem ser encontrados na literatura, de forma a fazer uma análise que nos possibilita discorrer sobre o quão fortes são os objetos de estudo deste trabalho, a saber, os modelos apresentados: CP1 e CP2. Logo, tendo em mente este propósito, apresentamos uma análise comparativa entre os modelos apresentados e, frequentemente, utilizados na literatura, UWWP e MSSC, e os modelos propostos neste trabalho.

Todos os experimentos computacionais realizados neste trabalho são descritos neste capítulo. Os modelos foram implementados na linguagem AMPL e os *solvers* utilizados foram : o Xpress versão 8.11, o Cplex versão 20.1 e o Baron versão 21.1. Os testes foram realizados em um processador Intel Core i7-2600, 12GB de memória (RAM) e sistema operacional Ubuntu 16.04 64 bits.

Resolvemos os modelos CP1 e CP2 usando dois tipos de *solvers*: o Xpress e o Cplex e resolvemos os modelos UWWP e MSSC com o *solver* Baron pelo fato do primeiro ser não convexo e não diferenciável e o segundo não convexo. Comparamos a resolução, tempo de solução e gap do modelo CP1 com o UWWP e do modelo CP2 com o MSSC, pelo fato dos modelos propostos terem sido desenvolvidos com base nestes modelos citados na literatura, respectivamente.

Utilizamos um conjunto de instâncias para problemas de clusterização o qual propusemos, em que pontos aleatórios foram gerados no interior de um quadrado de lado igual a 1, no espaço 2-dimensional, a serem agrupados em 4 clusters. Geramos instâncias neste formato com o número de pontos que variam de 10 à 40 e estas

instâncias foram usadas como dado de entrada para gerar os resultados das Tabelas 4.1 e 4.2. Além deste, também usamos os conjuntos de dados a seguir:

- *GermanTowns* : é uma instância que descreve as coordenadas cartesianas de 59 cidades alemãs, originalmente proposta por SPATH [52]. São 59 pontos 2-dimensionais a serem agrupados em 4 clusters.
- *Ruspini* : este exemplo foi proposto por RUSPINI [53] e é composto por 74 pontos artificiais de dimensão 2 a serem agrupados em 6 clusters.
- *Iris* : é um conjunto de dados extraído da variação de plantas Iris contendo a observação de 4 características, foi proposto por ANDERSON [54] e abordado por FISHER [55] . Instância é composta por 150 pontos 4-dimensionais a serem divididos por 4 grupos.
- *Speath* : contém 96 pontos em um espaço 5-dimensional para serem agrupados em 6 clusters por SPATH [56].
- *Wine* : contém 178 pontos no espaço 13-dimensional para serem agrupados em 11 clusters. Esse conjunto de dados é derivado de uma análise de vinhos com 13 componentes na região da Itália por FORINA *et al.* [57].
- *Yeast* : contém 1484 pontos em um espaço 8-dimensional a serem clusterizados em 4 grupos. Conjunto de dados apresentado por HORTON e NAKAI [58] que descreve proteínas.

As Tabelas 4.1 e 4.2, representam a comparação entre os modelos CP1, proposto neste trabalho, e o *Unitary Weighted Weber Problem* - UWWP mencionado na Subseção 3.1.1, além da comparação entre os modelos CP2 e o *Minimum Sum of Squares Clustering Problem* - MSSC mencionado na Subseção 3.1.2, respectivamente, utilizando como dados de entrada o conjunto de instâncias gerados neste trabalho.

Na Tabela 4.1 o modelo CP1 foi resolvido pelos *solvers* Cplex e Xpress e ambos encontraram a solução ótima $f^{\star}$ para as instâncias geradas com p pontos como mostra a primeira e segunda coluna. A coluna "t" representa o tempo em segundos que os *solvers* Cplex e Xpress levaram para chegar na solução ótima e a coluna "n" representa o número do nó o qual foi encontrada a melhor solução. As ultimas três colunas da tabela são os resultados do modelo UWWP resolvido pelo *solver* Baron em que f representa a melhor solução encontrada, a coluna "n" descreve o nó em que foi encontrada a solução e para mensurar o "gap" utilizamos o seguinte cálculo

$gap = \left(\frac{limite\ superior - limite\ inferior}{limite\ superior}\right) * 100$, gap disponibilizado em percentual. O tempo de solução para todas as instâncias do modelo UWWP resolvido pelo Baron foi limitado em 10800 segundos, o equivalente a 3 horas, e todas elas utilizaram o tempo total de processamento.

Para facilitar o entendimento, vamos interpretar a primeira linha com os resultados na Tabela 4.1 : temos que a instância contendo 10 pontos p=10 foi usada como dado de entrada para o modelo CP1 que foi resolvido com os *solvers* Cplex e Xpress. Ambos encontraram o ótimo global $f^\star$, porém cada um em seu tempo de processamento "t" em segundos e no nó "n" respectivamente. Já o modelo UWWP foi resolvido com o *solver* Baron o qual encontrou a solução f que não é a ótima no tempo fixado em 3 horas. Ele encontrou a solução no nó "n" com o "gap" um pouco maior que 88%.

CP1 x UWWP								
	CP1					UWWP		
		Cplex		Xpress		Baron		
p	$f^\star$	t	n	t	n	f	n	gap
10	**1.4268**	3	38570	9	37051	1.7000	254241	88.11
12	**1.8625**	22	192773	50	193185	2.3000	446945	95.65
13	**2.1201**	44	447958	120	449647	2.5000	49828	96
15	**2.4807**	256	2087813	476	1906841	3.0000	297499	100

Tabela 4.1: Tabela com resultados da comparação entre os modelos CP1 e UWWP utilizando dados gerados.

Na Tabela 4.2 podemos seguir a mesma interpretação que a anterior, porém com resultados referente aos modelos CP2 e MSSC. Além disso, ela contém uma coluna extra "t(*)" que expressa o tempo em que o Baron encontrou a melhor solução para o modelo MSSC (apesar de ter rodado por todo o tempo fixado, o equivalente a 3 horas, no tempo "t(*)" ele encontrou a melhor solução com valor "f" para o problema, e ainda assim, não provou a otimalidade). Uma observação importante com relação ao tempo é que todos os exemplos dos modelos UWWP e MSSC resolvidos pelo Baron nas Tabelas 4.1 e 4.2 foram rodados utilizando o tempo limite igual a 3 horas e quase todos utilizaram todo esse tempo para rodar, exceto um único exemplo: a instância de 10 pontos no modelo MSSC que levou 2850 segundos, o equivalente a 47,50 minutos.

CP2 x MSSC									
	CP2					MSSC			
		Cplex		Xpress		Baron			
p	$f^\star$	t	n	t	n	f	t(*)	n	gap
10	**0.2566**	0.25	3930	0.43	481	0.2566	1.31	3954	0
12	**0.3873**	1	22442	0.95	1495	0.3873	14.68	2511	76.76
13	**0.4040**	1	29390	1	1789	0.4040	3.26	111	74.88
15	**0.5105**	4	68597	2	3853	0.5105	0.07	10527	91.95
20	**0.6701**	20	270283	13	17305	0.6701	2.04	66	99.25
30	**0.9594**	370	2441578	325	309125	0.9594	13.88	16183	100
40	**1.2342**	1738	6682434	1012	850351	1.2342	0.31	-1	99.99

Tabela 4.2: Tabela com resultados da comparação entre os modelos CP2 e MSSC utilizando dados gerados.

Os resultados exibidos nas Tabelas: 4.3, 4.4, 4.5 e 4.6 são disponibilizados de forma a compararmos os modelos CP1 e UWWP e os modelos CP2 e MSSC, mas para este caso são utilizadas instâncias da literatura como dado de entrada. As Tabelas 4.3 e 4.4 exibem os resultados dos modelos CP1 e UWWP e o que as diferencia é que na Tabela 4.3 o modelo CP1 é resolvido pelo *solver* Cplex e na 4.4 ele é resolvido pelo Xpress, da mesma forma ocorre nas Tabelas 4.5 e 4.6 exibindo os resultados dos *solvers* Cplex e Xpress respectivamente, porém neste caso, se tratando do modelo CP2, uma vez que estas duas últimas tabelas apontam os resultados dos modelos CP2 e MSSC. Todos os experimentos computacionais destas quatro tabelas citadas tiveram seu tempo de processamento fixado em 10800 segundos (3 horas) e eles utilizaram este tempo no processo de solução. A primeira coluna das tabelas faz referência à instância utilizada. Na sequência, a coluna "f" indica o valor da melhor solução encontrada. As colunas denotadas por "t(*)" indicam o tempo em que a melhor solução para o problema foi encontrada durante o processamento, isso não quer dizer que o experimento foi interrompido neste tempo, e, como já dito anteriormente, cada experimento rodou por 3 horas. A coluna "n" retrata em que nó foi encontrada a melhor solução e onde lemos o resultado "-1" nesta coluna, isso significa que a solução disponibilizada foi encontrada por heurísticas ainda no pré-processamento do *solver* a que se refere. A coluna "gap" é calculada da mesma maneira como foi citado anteriormente para as duas primeiras tabelas e estas quatro colunas mencionadas neste parágrafo se repetem nas tabelas fazendo referência aos resultados do *solver* o qual elas se tratam.

CP1 x UWWP								
	Cplex				Baron			
Instância	f	t(*)	n	gap	f	t(*)	n	gap
german_towns	1801.01	2596.34	2105185	100	1578.24	4.95	27	100
ruspini	**916.15**	6296.55	4179997	99.87	934.32	1.66	-1	100
iris	109.19	8426.06	402702	98.90	161.83	18	-1	100
spaeth	92826.80	8893.76	1069605	100	76282.53	15.93	10	100
wine	30094.35	1903.16	8419	100	77220.53	2015.26	3009	100
yeast	370.32	4261.91	-1	100	2645.80	2645.80	93	100

Tabela 4.3: Tabela com resultados da comparação entre os modelos CP1 e UWWP resolvidos por Cplex e Baron respectivamente e utilizando instâncias da literatura.

CP1 x UWWP								
	Xpress				Baron			
Instância	f	t(*)	n	gap	f	t(*)	n	gap
german_towns	**1528.41**	159	51259	100	1578.24	4.95	27	100
ruspini	1028.60	4166	1165699	100	934.32	1.66	-1	100
iris	**101.28**	1472	100477	99.41	161.83	18	-1	100
spaeth	**71863.72**	2961	159931	100	76282.53	15.93	10	100
wine	**10039.43**	1404	6608	100	77220.53	2015.26	3009	100
yeast	**354.33**	41	-1	100	2645.80	2645.80	93	100

Tabela 4.4: Tabela com resultados da comparação entre os modelos CP1 e UWWP resolvidos por Xpress e Baron respectivamente e utilizando instâncias da literatura.

CP2 x MSSC								
	Cplex				Baron			
Instância	f	t(*)	n	gap	f	t(*)	n	gap
german_towns	32531.50	10027.97	14928977	94.47	**32333.49**	0.53	1	100
ruspini	**8150.53**	686.71	523813	93.84	44264.27	66.73	30	100
iris	**57.22**	8598.78	1855269	85.99	57.35	1.31	1	100
spaeth	**38614856.84**	5804.07	410001	100	62449313.49	1.86	1	100
wine	4272319.63	10800	12858	100	**589719.66**	5394.02	5004	100
yeast	104.37	10800	11032	100	**71.29**	213.12	37	100

Tabela 4.5: Tabela com resultados da comparação entre os modelos CP2 e MSSC resolvidos por Cplex e Baron respectivamente e utilizando instâncias da literatura.

CP2 x MSSC								
	Xpress				Baron			
Instância	f	t(*)	n	gap	f	t(*)	n	gap
german_towns	38820.39	1387	309888	90.87	**32333.49**	0.53	1	100
ruspini	8350.01	4014	696433	66.99	44264.27	66.73	30	100
iris	69.72	2590	164466	93.78	57.35	1.31	1	100
spaeth	39890394.89	8497	466470	100	62449313.49	1.86	1	100
wine	1120083.29	5215	16830	100	**589719.66**	5394.02	5004	100
yeast	97.67	8020	8009	99.99	**71.29**	213.12	37	100

Tabela 4.6: Tabela com resultados da comparação entre os modelos CP2 e MSSC resolvidos por Xpress e Baron respectivamente e utilizando instâncias da literatura.

4.2 Discussões dos Resultados

Analisando a Tabela 4.1 podemos comparar os resultados do modelo CP1 resolvido pelos *solvers* Cplex e Xpress e o modelo UWWP resolvido pelo *solver* Baron. Conseguimos notar uma grande diferença entre as soluções, uma vez que o modelo CP1 pôde ser resolvido de maneira ótima pelos dois *solvers* utilizados, enquanto que o modelo UWWP não encontrou a solução ótima, mas sim soluções com gap entre 88 e 100 por cento. Além disso, comparando o tempo de processamento de cada experimento, podemos observar uma enorme diferença entre os modelos. Todos os dados de entrada utilizados no modelo CP1 resolvidos tanto pelo Cplex como pelo Xpress foram resolvidos de forma bem mais rápidas do que no modelo UWWP, em que o tempo fixado foi de 3 horas e todos os experimentos resolvidos pelo Baron desta tabela utilizaram este tempo por completo. Logo, para estes experimentos, podemos concluir que o modelo proposto CP1 obteve uma performance superior ao outro modelo, tanto no quesito temporal quanto na qualidade das soluções apresentadas na tabela.

Para os experimentos apresentados na Tabela 4.1, além da comparação entre os modelos, podemos perceber também uma diferença no tempo de solução entre os *solvers* Cplex e Xpress, ainda que os dois tenham encontrado a solução ótima para o problema, identificamos que o Cplex obteve uma performance temporal superior ao outro, em todos os casos.

Observando a Tabela 4.2 podemos comparar os resultados entre os modelos CP2 e MSSC. Todas as instâncias disponibilizadas nesta tabela foram resolvidas de maneira ótima pelos *solvers* Cplex e Xpress utilizados para resolução do primeiro modelo. Elas foram resolvidas nos tempos constatados pela coluna "t" indicada por cada *solver* e contabilizados em segundos. Por outro lado, as soluções encontradas pelo *solver* Baron, o qual resolve o segundo modelo apresentado na tabela, só podem ser consideradas ótimas pelo fato de já termos resolvido o modelo CP2 e ter tido a constatação desta otimalidade através da resolução dele. Exceto pela solução da primeira instância resolvida, todos os demais "gaps" obtidos pelo modelo MSSC somente reforçam que o modelo proposto CP2, de fato, para estas instâncias, melhorou/aumentou de maneira significativa os limitantes inferiores em sua resolução, fazendo com que pudéssemos provar a otimalidade destes casos e constatando a força que a nova modelagem representa para os resultados analisados.

As Tabelas 4.3 e 4.4 apresentam a comparação entre os modelos CP1 e UWWP utilizando as instâncias encontradas na literatura referenciadas nas tabelas. E por mais que estes resultados estejam disponibilizados em duas tabelas, por uma

questão de melhor visualização e disposição na página, elas se complementam com relação à análise geral destes resultados. Nossa análise comparativa é feita com o respeito aos modelos, então comparamos os resultados que os *solvers* Xpress e Cplex apresentam com os do Baron. Logo, verificando as soluções "f" disponibilizadas nas duas tabelas pertencentes aos três *solvers*, podemos perceber que, para todas as instâncias, as melhores soluções são encontradas pelo modelo CP1 e distribuídas entre o Xpress com o maior número delas e o Cplex. As melhores soluções encontradas nesta comparação entre os três *solvers* estão em negrito nas tabelas. Todos estes experimentos rodaram num tempo fixado de 3 horas (tanto para resolução com Cplex, Xpress e Baron), e mesmo sem ter conseguido resolvê-los dentro deste tempo de maneira ótima, conseguimos mostrar que o modelo proposto encontrou soluções de melhor qualidade em todos os casos e percebemos uma leve alteração com relação ao "gap" em alguns casos do modelo CP1.

Com relação às Tabelas 4.5 e 4.6, mostramos a comparação entre os modelos CP2 e MSSC usando como dados de entrada instâncias da literatura. Estas tabelas também são analisadas de maneira conjunta. Verificando as soluções "f" dos *solvers* Cplex e Xpress em comparação com as do Baron, conseguimos constatar que na metade dos casos o modelo CP2 nos fornece as melhores soluções através do *solver* Cplex. Logo, ainda que todos os experimentos apresentados nas duas tabelas tenham rodado por 3 horas e não tenha sido possível resolvê-los de maneira ótima, parte das soluções de melhor qualidade foram apresentadas pelo modelo proposto neste trabalho. Além disso, percebemos que em alguns deste casos o "gap" diminui de maneira a observarmos uma melhora nos limites inferiores destes problemas resolvidos pelo modelo CP2, o que reforça a qualidade da nova modelagem.

Capítulo 5

Conclusões

Apresentamos um estudo sobre dois novos modelos para o problema de agrupamento e, nesta abordagem, fizemos uma comparação entre os modelos propostos e modelos já conhecidos na literatura com o objetivo de, não somente contribuir com a modelagem matemática, mas de mostrar a sua força de forma quantitativa por meio dos resultados computacionais. Analisando todos os experimentos com respeito às comparações realizadas, tanto entre os modelos CP1 e UWWP como entre os modelos CP2 e MSSC, e utilizando como dados de entrada as instâncias geradas neste trabalho e as instâncias utilizadas que estão disponíveis na literatura, conseguimos averiguar a relevância que o objeto de estudo desta pesquisa apresenta. Através dos resultados computacionais, verificamos a superioridade dos modelos propostos neste trabalho em diversos quesitos apresentados e, consequentemente, conseguimos constatar sua significativa contribuição no âmbito da pesquisa na área de modelagem matemática para problemas de agrupamento.

O estudo deste trabalho relativo às novas abordagens para o problema de agrupamento nos permite fazer um panorama e abrir diversas perspectivas visando novas pesquisas. Uma das vertentes seria, por exemplo, a utilização de outros métodos de otimização para resolver os modelos aqui apresentados, uma outra vertente seria a utilização de heurísticas e meta-heurísticas, alguns exemplos de métodos podem ser vistos na Seção 2.2. Além de novas pesquisas na área de resolução dos problemas apresentados, um estudo dos modelos utilizando outras medidas de proximidade também podem ser feito, algumas medidas são apresentadas na Seção 1.2.

Referências Bibliográficas

[1] LI, H., XU, Z., TAYLOR, G., et al. "Visualizing the loss landscape of neural nets", *arXiv preprint arXiv:1712.09913*, 2017.

[2] HANSEN, P., JAUMARD, B. "Cluster analysis and mathematical programming", *Mathematical programming*, v. 79, n. 1-3, pp. 191–215, 1997.

[3] ALOISE, D., HANSEN, P. "A branch-and-cut SDP-based algorithm for minimum sum-of-squares clustering", *Pesquisa Operacional*, v. 29, n. 3, pp. 503–516, 2009.

[4] HRUSCHKA, E. R., EBECKEN, N. F. "A genetic algorithm for cluster analysis", *Intelligent Data Analysis*, v. 7, n. 1, pp. 15–25, 2003.

[5] COLE, R. M. *Clustering with genetic algorithms*. 1998.

[6] LIU, C. L. "Introduction to combinatorial mathematics", 1968.

[7] ALOISE, D., DESHPANDE, A., HANSEN, P., et al. "NP-hardness of Euclidean sum-of-squares clustering", *Machine learning*, v. 75, n. 2, pp. 245–248, 2009.

[8] XAVIER, A. E. "The hyperbolic smoothing clustering method", *Pattern Recognition*, v. 43, n. 3, pp. 731–737, 2010.

[9] XU, R., WUNSCH, D. "Survey of clustering algorithms", *IEEE Transactions on neural networks*, v. 16, n. 3, pp. 645–678, 2005.

[10] LINDEN, R. "Técnicas de agrupamento", *Revista de Sistemas de Informação da FSMA*, v. 4, n. 4, pp. 18–36, 2009.

[11] GAN, G., MA, C., WU, J. *Data clustering: theory, algorithms, and applications*. SIAM, 2020.

[12] HAN, J., PEI, J., KAMBER, M. *Data mining: concepts and techniques*. Elsevier, 2011.

[13] DUONG, K.-C., VRAIN, C., OTHERS. "Constrained clustering by constraint programming", *Artificial Intelligence*, v. 244, pp. 70–94, 2017.

[14] KAUFMAN, L., ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*, v. 344. John Wiley & Sons, 2009.

[15] CUI, X., POTOK, T. E., PALATHINGAL, P. "Document clustering using particle swarm optimization". In: *Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005.*, pp. 185–191. IEEE, 2005.

[16] FAHAD, A., ALSHATRI, N., TARI, Z., et al. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis", *IEEE transactions on emerging topics in computing*, v. 2, n. 3, pp. 267–279, 2014.

[17] DU MERLE, O., HANSEN, P., JAUMARD, B., et al. "An interior point algorithm for minimum sum-of-squares clustering", *SIAM Journal on Scientific Computing*, v. 21, n. 4, pp. 1485–1505, 1999.

[18] BAGIROV, A., RUBINOV, A., YEARWOOD, J. "Using global optimization to improve classification for medical diagnosis and prognosis." *Topics in Health Information Management*, v. 22, n. 1, pp. 65–74, 2001.

[19] WANG, X., GARIBALDI, J. M. "A comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis". In: *Proceedings of the 2nd International Conference in Computational Intelligence in Medicine and Healthcare, BIOPATTERN Conference, Costa da Caparica, Lisbon, Portugal*, v. 28. Citeseer, 2005.

[20] WOLKIND, S., EVERITT, B. "A cluster analysis of the behavioural items in the pre-school child1", *Psychological Medicine*, v. 4, n. 4, pp. 422–427, 1974.

[21] CHIANG, I. W.-Y., LIANG, G.-S., YAHALOM, S. "The fuzzy clustering method: Applications in the air transport market in Taiwan", *Journal of Database Marketing & Customer Strategy Management*, v. 11, n. 2, pp. 149–158, 2003.

[22] JIANG, D., TANG, C., ZHANG, A. "Cluster analysis for gene expression data: a survey", *IEEE Transactions on knowledge and data engineering*, v. 16, n. 11, pp. 1370–1386, 2004.

[23] NEGREIROS, M. J., MACULAN, N., BATISTA, P. L., et al. "Capacitated clustering problems applied to the layout of IT-teams in software factories", *Annals of Operations Research*, pp. 1–29, 2020.

[24] DZWINEL, W., YUEN, D. A., BORYCZKO, K., et al. "Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space", *Nonlinear Processes in Geophysics*, v. 12, n. 1, pp. 117–128, 2005.

[25] GUAN, Y., GHORBANI, A. A., BELACEL, N. "Y-means: A clustering method for intrusion detection". In: *CCECE 2003-Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No. 03CH37436)*, v. 2, pp. 1083–1086. IEEE, 2003.

[26] LIU, T., ROSENBERG, C., ROWLEY, H. A. "Clustering billions of images with large scale nearest neighbor search". In: *2007 IEEE workshop on applications of computer vision (WACV'07)*, pp. 28–28. IEEE, 2007.

[27] LIU, J., LEE, J. P., LI, L., et al. "Online clustering algorithms for radar emitter classification", *IEEE transactions on pattern analysis and machine intelligence*, v. 27, n. 8, pp. 1185–1196, 2005.

[28] AKKAYA, K., SENEL, F., MCLAUGHLAN, B. "Clustering of wireless sensor and actor networks based on sensor distribution and connectivity", *Journal of Parallel and Distributed Computing*, v. 69, n. 6, pp. 573–587, 2009.

[29] FENG, A. "Document clustering: an optimization problem". In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 819–820, 2007.

[30] KABADI, S., MURTY, K. G., SPERA, C. "Clustering problems in optimization models", *Computational Economics*, v. 9, n. 3, pp. 229–239, 1996.

[31] BUDKA, M. "Clustering as an example of optimizing arbitrarily chosen objective functions". In: *Advanced Methods for Computational Collective Intelligence*, Springer, pp. 177–186, 2013.

[32] OLIVEIRA, S., STEWART, D. E. "Clustering for bioinformatics via matrix optimization". In: *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 559–563, 2011.

[33] XAVIER, A. E. *Penalização hiperbólica: Um novo método para resolução de problemas de otimização*. Tese de Doutorado, M. Sc. Thesis-COPPE-UFRJ, Rio de Janeiro, 1982.

[34] XAVIER, A. E., XAVIER, V. L. "Solving the minimum sum-of-squares clustering problem by hyperbolic smoothing and partition into boundary and gravitational regions", *Pattern Recognition*, v. 44, n. 1, pp. 70–77, 2011.

[35] XAVIER, A. E., XAVIER, V. L., VILLAS-BOAS, S. B. "Solving the Minimum Sum of L1 Distances Clustering Problem by Hyperbolic Smoothing and Partition into Boundary and Gravitational Regions". In: *Algorithms from and for Nature and Life*, Springer, pp. 33–46, 2013.

[36] XAVIER, V. L., XAVIER, A. E. "Accelerated hyperbolic smoothing method for solving the multisource Fermat–Weber and k-Median problems", *Knowledge-Based Systems*, v. 191, pp. 105226, 2020.

[37] BAGIROV, A. M., ORDIN, B., OZTURK, G., et al. "An incremental clustering algorithm based on hyperbolic smoothing", *Computational Optimization and Applications*, v. 61, n. 1, pp. 219–241, 2015.

[38] GRÖTSCHEL, M., WAKABAYASHI, Y. "A cutting plane algorithm for a clustering problem", *Mathematical Programming*, v. 45, n. 1, pp. 59–96, 1989.

[39] WOLSEY, L. A. *Integer programming*. John Wiley & Sons, 2020.

[40] BAGIROV, A. M. "Modified global k-means algorithm for minimum sum-of-squares clustering problems", *Pattern Recognition*, v. 41, n. 10, pp. 3192–3199, 2008.

[41] SANTI, É., ALOISE, D., BLANCHARD, S. J. "A model for clustering data from heterogeneous dissimilarities", *European Journal of Operational Research*, v. 253, n. 3, pp. 659–672, 2016.

[42] AL-SULTAN, K. S. "A tabu search approach to the clustering problem", *Pattern recognition*, v. 28, n. 9, pp. 1443–1451, 1995.

[43] AL-SULTANA, K. S., KHAN, M. M. "Computational experience on four algorithms for the hard clustering problem", *Pattern recognition letters*, v. 17, n. 3, pp. 295–308, 1996.

[44] MACULAN, N., NEGREIROS, M., PINTO, R. "Two Optimization Models For Clustering Problems", *LII Simpósio Brasileiro de Pesquisa Operacional*, 2020.

[45] CHEN, P.-C., HANSEN, P., JAUMARD, B., et al. "Weber's problem with attraction and repulsion", *Journal of Regional Science*, v. 32, n. 4, pp. 467–486, 1992.

[46] LUENBERGER, D. G., YE, Y., OTHERS. *Linear and nonlinear programming*, v. 2. Springer, 1984.

[47] PERESSINI, A. L., SULLIVAN, F. E., UHL JR, J. J. *The mathematics of nonlinear programming*. Springer-Verlag, 1988.

[48] FLETCHER, R. *Practical methods of optimization*. John Wiley & Sons, 2013.

[49] BAGIROV, A. M., YEARWOOD, J. "A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems", *European journal of operational research*, v. 170, n. 2, pp. 578–596, 2006.

[50] BRUCKER, P. "On the complexity of clustering problems". In: *Optimization and operations research*, Springer, pp. 45–54, 1978.

[51] XUE, G., YE, Y. "An efficient algorithm for minimizing a sum of Euclidean norms with applications", *SIAM Journal on Optimization*, v. 7, n. 4, pp. 1017–1036, 1997.

[52] SPATH, H. *Cluster analysis algorithms for data reduction and classification of objects*. Ellis Horwood Chichester, 1980.

[53] RUSPINI, E. H. "Numerical methods for fuzzy clustering", *Information Sciences*, v. 2, n. 3, pp. 319–350, 1970.

[54] ANDERSON, E. "The irises of the Gaspe Peninsula", *Bull. Am. Iris Soc.*, v. 59, pp. 2–5, 1935.

[55] FISHER, R. A. "The use of multiple measurements in taxonomic problems", *Annals of eugenics*, v. 7, n. 2, pp. 179–188, 1936.

[56] SPATH, H. *The cluster dissection and analysis theory fortran programs examples*. Prentice-Hall, Inc., 1985.

[57] FORINA, M., LANTERI, S., ARMANINO, C., et al. "Parvus-an extendible package for data exploration, classification and correlation, institute of pharmaceutical and food analysis and technologies, via brigata salerno, 16147 genoa, italy (1988)", *Av. Loss Av. O set Av. Hit-Rate*, 1991.

[58] HORTON, P., NAKAI, K. "A probabilistic classification system for predicting the cellular localization sites of proteins." In: *Ismb*, v. 4, pp. 109–115, 1996.