Pós-Graduação em Ciência da Computação

Elaine Cristina de Assis

Algoritmos de Particionamento Aplicados à Análise Estatística de Formas

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2018

Elaine Cristina de Assis

Algoritmos de Particionamento Aplicados à Análise Estatística de Formas

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

Orientadora: **Renata Maria Cardoso Rodrigues de Souza**
Coorientador: **Getúlio José Amorim do Amaral**

Recife
2018

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

A848a Assis, Elaine Cristina de
Algoritmos de particionamento aplicados à análise estatística de formas / Elaine Cristina de Assis. – 2018.
88 f.: il., fig., tab.

Orientadora: Renata Maria Cardoso Rodrigues de Souza.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2018.
Inclui referências.

1. Inteligência computacional. 2. Métodos de agrupamento. I. Souza, Renata Maria Cardoso Rodrigues de (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2018-099

Elaine Cristina de Assis

Algoritmos de Partição Aplicados à Análise Estatística de Formas

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovado em: 08/02/2018.

_____

Orientadora: Profa. Dra. Renata Maria Cardoso Rodrigues de Souza

BANCA EXAMINADORA

_____

Prof. Dr. Cleber Zanchettin
Centro de Informática / UFPE

_____

Prof. Dr. Paulo José Duarte Neto
Departamento de Estatística e Informática/ UFRPE

_____

Prof. Dr. Leandro Carlos de Souza
Departamento de Computação/UFERSA

_____

Prof. Dr. Carlos Wilson Dantas de Almeida
Departamento de Sistemas e Computação/UFCG

_____

Prof. Dr. Bruno Almeida Pimentel
Instituto de Ciências Matemáticas e de Computação/ USP

Decido este trabalho à minha mãe e ao meu noivo que foram minha fortaleza perante as dificuldades enfrentadas.

AGRADECIMENTOS

Gostaria de agradecer muito aos meus orientadores Renata Souza e Getulio Amorim pelos ensinamentos e paciência que tiveram comigo. Ao meu noivo Marcilio Martins pelo incentivo e compreensão nessa jornada árdua. Aos meus amigos que direta ou indiretamente contribuíram para essa conquista. A minha mãe que é a minha maior fortaleza (mas Mainha ainda acha que todo dia eu saio de casa pra ir pra escola. Ao meu professor/orientador triatleta aventureiro Jones Albuquerque que acreditou em mim desde o início dessa caminhada. Enfim agradeço a todos que caminharam comigo até aqui.

“Conhece-te a ti mesmo, torna-te consciente de tua ignorância e será sábio."
(Sócrates)

RESUMO

A análise estatística de forma é usada para tomar decisões observando a forma de objetos. A forma de um objeto é a informação restante quando os efeitos de locação, escala e rotação são removidos com a utilização de operações matemáticas adequadas. Algoritmos de agrupamento por particionamento encontram uma partição que maximiza ou minimiza algum critério numérico. Esta pesquisa apresenta algoritmos de agrupamento baseados em protótipos e em busca, adaptados para os dados da área de análise estatística de formas. Esses algoritmos são novas versões dos algoritmos de agrupamento K-médias e KI-médias, Subida da Encosta, Busca Tabu, apropriados para o tratamento de dados de formas bidimensionais. Para avaliar a formação dos agrupamentos para os algoritmos propostos, novos critérios de agrupamento foram gerados para dados de formas a partir de estatísticas de testes de hipóteses e critérios usuais já existentes na literatura. A fim de melhorar a qualidade dos resultados de agrupamento, os algoritmos foram também analisados quando foram utilizados em conjunto com o método *Bagging* que faz uma reamostragem com repetição para os dados de entrada e gerou grupos através de uma votação majoritária. Estudos de simulação foram realizados para validar esses métodos propostos e três conjuntos de dados reais disponíveis na literatura também foram considerados. A qualidade dos experimentos foi avaliada pelo índice *Rand* corrigido e os resultados mostraram que os algoritmos propostos para formas bidimensionais são eficientes para os conjuntos de dados analisados.

Palavras-chaves: Análise Estatística de Formas. Métodos de Agrupamento Particionais. Procedimento *Bagging*.

ABSTRACT

Partitioning clustering algorithms find a partition that maximizes or minimizes some numeric criteria. Statistical shape analysis is used to make decisions by observing the shape of objects. The shape of an object is the remaining information when the location, scale and rotation effects are removed with the use of appropriate mathematical operations. This research presents clustering algorithms based on prototypes and in search adapted to the data of the area of statistical shape analysis. These algorithms are new versions of the K-means and KI-means, Hill Climbing and Tabu Search clustering algorithms, suitable for the treatment of two-dimensional data. In order to evaluate the formation of the clusters for the proposed algorithms, new clustering criteria were generated for form data from test statistics of hypotheses and usual criteria already existent in the literature. In order to improve the quality of clustering results, the algorithms were also analyzed when they were used in conjunction with the Bagging method, which makes a resampling with repetition for the input data and generated groups through a majority vote. Simulation studies were performed to validate these proposed methods and three real data sets available in the literature were also considered. The quality of the experiments was evaluated by the corrected Rand index and the results showed that the algorithms proposed for planar shapes are efficient for the data sets analyzed.

Key-words: Statistical Shape Analysis. Partitional Clustering Methods. Bagging Procedure.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A análise de agrupamento é uma ferramenta de análise exploratória de dados, cujo objetivo é organizar um conjunto de itens em grupos tais que os itens dentro de um determinado grupo têm um elevado grau de semelhança, enquanto itens pertencentes a grupos diferentes têm um alto grau de dissimilaridade. Os métodos de agrupamento podem ser divididos em métodos hierárquicos e métodos de particionamento ((EVERITT et al., 2011)). Algoritmos de agrupamentos baseados no método hierárquico organizam um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos. Métodos de particionamento de dados são baseados na transferência de objetos entre os grupos em que a qualidade das soluções é medido por um critério de agrupamento ((XU; WUNSCH, 2005)). Em cada iteração, os algoritmos realocam os dados para otimizar o valor da função de critério, até a convergência.

Em geral esses métodos de agrupamento por particionamento são divididos em dois grupos de paradigmas: rígido *(hard)* e difuso/nebuloso *(fuzzy)*. Os algoritmos rígidos associam um item a apenas um grupo, enquanto os algoritmos difusos/nebulosos associam um item a todos os grupos através de um grau de pertinência do item em cada grupo ((JAIN; MURTY; FLYNN, 1999)). Um dos mais conhecidos métodos de agrupamento rígidos baseado em médias é o algoritmo K-médias (*K-means*) proposto por (MACQUEEN, 1967) que iterativamente encontra k elementos cuja soma das distâncias para todos os outros indivíduos do conjunto de dados seja mínima, estes elementos são chamados centróides, e atribui cada objeto ao centróide mais próximo, onde a coordenada de cada centróide é a média das coordenadas dos objetos no grupo. O algoritmo de agrupamento difuso mais conhecido é o c-médias difuso (*fuzzy c-means* (FCM)) ((DUNN, 1974); (BEZDEK, 1981); (YANG, 1993)) que encontra c centróides e associa um objeto a todos os grupos através de um grau de pertinência do objeto em cada grupo também baseado na média das coordenadas.

Imagens de objetos estão disponíveis em muitas áreas, tais como biologia, medicina e arqueologia. A forma dos objetos pode ser útil para a tomada de importantes decisões, como a de um médico que precisa decidir se um paciente tem ou não esquizofrenia baseado em sua ressonância magnética digitalizada do cérebro. A Análise Estatística de Formas (*Statistical Shape Analysis*) é uma área de conhecimento que pode ser usada para tomar este tipo de decisão, pois ela lida com estudos sobre informações contidas nas formas dos objetos. Essa é uma área relevante para vários campos de pesquisa como biologia, medicina, análises de imagem, arqueologia, geografia, agricultura, reconhecimento de alvos militares, como descrito nos trabalhos de (DRYDEN; MARDIA, 1998), (SMALL, 1996),

(BOOKSTEIN, 1991) e (KENDALL et al., 1999).

A forma de um objeto, definida na área de análise estatística de formas, é a informação que permanece quando os efeitos de locação, escala e rotação são removidos através de operações matemáticas, como descrito em (DRYDEN; MARDIA, 1998). Quando são removidos os efeitos de locação e escala, os dados são chamados de dados de pré-formas e são descritos em uma hiperesfera complexa, ou seja, no espaço não Euclidiano. Quando a informação de rotação é removida do espaço de pré-formas obtem-se as formas dos objetos no espaço de formas que também é não Euclidiano. Os primeiros trabalhos na área da análise estatística de formas foram relatados por (KENDALL, 1977), (BOOKSTEIN, 1986) e (KENDALL, 1984) que traziam os conceitos fundamentais da área.

Agrupamento é um dos relevantes tópicos em análise estatística de formas e alguns estudos mostram sua relevância, mas uma limitação fundamental dos algoritmos de agrupamento é que eles só são projetados para espaços Euclidianos ((EVERITT et al., 2011)). Assim, o desenvolvimento de algoritmos para o espaço não Euclidiano é necessário quando forem utilizados dados no espaço de pré-formas. Observando a relevância de agrupamentos para analisar formas de objetos, alguns trabalhos têm sido propostos. (GEORGESCU, 2009) propôs uma generalização do algoritmo K-médias a fim de integrar métricas Procrustes e estimação da forma média completa, de uma maneira a ser capaz de agrupar objetos com múltiplos ou difusos contornos. (AMARAL et al., 2010) adapta o método K-médias proposto por (HARTIGAN; WONG, 1979) para um problema real de oceanografia onde é necessário obter a classificação não supervisionada de espécies similares de peixes.

Deixando o espaço das formas com duas dimensões, (VINUE; SIMO; ALEMANY, 2014) adaptaram o algoritmo K-médias original de (LLOYD, 1982) para análise estatística de formas aplicado às formas 3D. A fim de encontar métodos de agrupamento razoáveis para objetos geométricos, (NABIL; GOLALIZADEH, 2016) consideraram quatro medidas de distância não Euclidianas combinadas com diferentes métodos de agrupamento aglomerativo diretamente no espaço de forma. Esta abordagem, que é principalmente aplicada às formas 2D, é de alguma forma semelhante ao trabalho de (AMARAL et al., 2010), mas os trabalhos diferem em seus resultados para as medidas de distância.

As decisões tomadas em conjunto são na maioria das vezes mais eficientes do que as decisões individuais. Os métodos *ensemble* são técnicas utilizadas para se trabalhar em conjunto que combinam informações para produzir um algoritmo mais eficiente. Os métodos *Bagging* e *Boosting* são exemplos de métodos *ensemble* ((CHERNICK; LABUDDE, 2011)). O método *Bagging*, proposto por (BREIMAN, 1996), tem por objetivo reduzir a variabilidade nos resultados de particionamentos de dados através de média. O método utiliza várias versões de um conjunto de treinamento utilizando a amostragem *bootstrap*, ou seja, com reposição. Cada um destes conjuntos de dados é usado para treinar um modelo diferente. Em geral, este método aplicados aos algoritmos de agrupamento pode convergir para um mínimo local. A inicialização pode ser utilizada para estabelecer a

reprodutibilidade dos grupos e a variabilidade global do procedimento seguinte. Portanto, a reamostragem *bootstrap* pode melhorar as chances de encontrar o mínimo global ou encontrar um mínimo local mais próximo do mínimo global desejado.

A análise de agrupamento pode se beneficiar da combinação dos pontos fortes de outros algoritmos que são usados individualmente. Os métodos *clustering ensembles* tem como foco combinar múltiplas partições que ofereçam um agrupamento melhorado dos conjuntos de dados que possam atingir um melhor resultado que um método de agrupamento individual. ((GHAEMI R.; MUSTAPHA, 2009)). O método *Bagging* tem sido aplicado em conjunto com métodos de agrupamento como um procedimento de *clustering ensembles* para melhorar o desempenho destes algoritmos, incorporando à estrutura da análise de agrupamento novos conjuntos de treinamento formados por amostras *bootstrap*. (LEISCH, 1999) introduziu um novo método de análise de dados chamado de *bagged clustering* que combinava métodos hierárquicos e métodos de particionamento da análise de agrupamento. (DUDOIT; FRIDLYAND, 2003) tem usado o método *Bagging* para melhorar a classificação com os algoritmos de particionamento baseados em *medoids* ((KAUFMAN; ROUSSEEUW, 1990)) e apresentaram dois tipos de reamostragem *Bagging* para analisar os dados.

Na área da análise de agrupamentos, os métodos de partição de interesse desta pesquisa são os baseados em protótipos e baseados em buscas. Os algoritmos baseados em protótipos assumem um ponto representativo para representar um agrupamento e tentam atingir um valor ótimo maximizando ou minimizando uma dada função critério de agrupamento. Os algoritmos baseados em buscas tentam encontrar uma partição que atinja o ótimo global ou um ótimo global aproximado baseados em um conjunto de ações realizadas nos dados. Os métodos de agrupamento podem ser considerados como uma categoria de problemas de otimização.

Esta tese utiliza os métodos de agrupamento baseados em protótipos: K-médias proposto por (MACQUEEN, 1967), KI-médias apresentado por (BANFIELD; BASSIL, 1977); e os métodos baseados em buscas: Subida da encosta descrito por (FRIEDMAN; RUBIN, 1967) e Busca Tabu proposto por (GLOVER, 1989). Esses algoritmos foram adaptados para trabalhar com conjuntos de dados da análise estatística de formas no espaço de pré-formas onde o espaço é não Euclidiano. Os algoritmos também foram aplicados para dados no espaço tangente onde é feita uma aproximação do espaço de pré-formas para que possam ser usados métodos da análise multivariada usual. O algoritmo K-médias já é comumente usado em aplicações de agrupamento. O algoritmo KI-médias é uma abordagem híbrida de agrupamento baseado no K-médias e em uma heurística. O método Subida da Encosta é um algoritmo usado amplamente em inteligência artificial ((RUSSELL; NORVIG, 2014)) que faz trocas de objetos observando a evolução do critério de agrupamento. O algoritmo Busca Tabu guarda os lugares já visitados pelo método para evitar que ele volte a um ótimo local visitado anteriormente para superar a otimalidade local. O uso de algoritmos

baseados em buscas para agrupar dados de formas é um tema em aberto.

Para atender as características dos dados no espaço de pré-formas e no espaço tangente, novos critérios de agrupamentos foram analisados. Tais critérios foram baseados em diferentes estatísticas de testes de hipóteses existentes na literatura e que já tinham sido aplicados para dados da análise estatística de formas. Além destas estatísticas, um critério de agrupamento baseado em soma dos erros quadráticos, com devidas adaptações, também foi utilizado. As estatísticas de testes ainda não tinham sido aplicadas aos algoritmos propostos. O método de reamostragem *Bagging* foi utilizado, em conjunto com os algoritmos de agrupamneto, para tentar melhorar a eficácia dos resultados obtidos por esses algoritmos e este método ainda não tinha métodos de *clustering ensembles* com os algoritmos *KI*-médias, Subida da Encosta e Busca Tabu. Para validar os métodos, foram realizados experimentos com quatro conjuntos de dados simulados de formas planas e três conjuntos de dados reais (Crânios de Gorilas, Vértebra de Ratos e Peixes) disponíveis na literatura. A avaliação do desempenho dos algoritmos de agrupamentos para os conjuntos de dados foi medido através do Índice de Rand Corrigido (*Correct Rand (CR)*) ((HUBERT; ARABIE, 1985)).

1.2 OBJETIVOS

a) Propor e analisar novas versões de algoritmos de agrupamento baseados em protótipos e busca para dados no espaço de pré-formas e no espaço tangente. Essas versões são baseadas nos algoritmos *KI*-médias, Subida da Encosta e Busca Tabu. A comparação destes métodos de agrupamento, a fim de identificar uma melhor alternativa para aplicação em dados da área de análise estatística de formas, teve como motivação o artigo de (BRUSCO; STEINLEY, 2007) que fez uma comparação de diferentes heurísticas para avaliar a performance dos agrupamentos gerados. A partir desta análise busca-se saber: Quais dos algoritmos de agrupamentos analisados seriam uma boa alternativa para trabalhar com dados da análise estatística de formas? Qual o melhor espaço de dados para identificar as formas dos objetos?

b) Avaliar a utilização de três diferentes estatísticas de testes de hipóteses e um critério baseado em soma dos erros quadráticos, adaptado para o espaço de pré-formas, como critérios de agrupamentos para os algoritmos propostos. A utilização foi baseada em (AMARAL; DRYDEN; WOOD, 2007) que aplicaram essas estatísticas para dados da análise estatística de formas. Já o critério de soma dos quadrados também foi motivado pelo artigo (BRUSCO; STEINLEY, 2007) que o utilizou na comparação de suas heurísticas. Baseado nessa avaliação: Qual o comportamento desenvolvido pelos critérios de agrupamento no espaço de pré-formas e espaço tangente?

c) Introduzir o contexto do método de reamostragem *Bagging*, trabalhando em conjunto com os algoritmos propostos e com o K-médias, para dados da análise esta-

tística de formas a fim de melhorar a eficiência dos resultados dos agrupamentos. O artigo (DUDOIT; FRIDLYAND, 2003) usou o método de reamostragem *Bagging* para melhorar o desempenho dos algoritmos de agrupamentos baseados em *medoids* (PAM) e serviu de motivação para o uso do *Bagging* para os algoritmos de agrupamentos aqui analisados. Com a utilização do método de reamostragem *Bagging* pretende-se responder a seguinte questão: Seria possível a melhoria de desempenho dos agrupamentos com a utilização deste método?

1.3 ORGANIZAÇÃO DA TESE

Os próximos capítulos desta tese de doutorado estão organizados da seguinte forma:

Capítulo 2: **Fundamentação Teórica**
Este capítulo apresentará um breve resumo com a descrição da área de análise estatística de formas, análise de agrupamentos e métodos de reamostragem.

Capítulo 3: Algoritmos de Agrupamentos Propostos
Neste capítulo serão apresentados os métodos de agrupamento K-médias, KI-médias, Subida da Encosta e Busca Tabu em suas versões adaptadas para trabalhar no espaço dos dados de pré-formas e no espaço tangente. Aqui também foram mostrados os critérios de agrupamentos utilizados pelos algoritmos. O método *clustering ensembles*, utilizando o algoritmo de reamostragem *Bagging* aliado aos métodos de agrupamento, também foi descrito neste capítulo.

Capítulo 4: Experimentos e Resultados
A descrição dos conjuntos de dados simulados e de dados reais, levando em consideração o espaço de pré-formas e o espaço tangente, utilizados para validar os métodos é realizada neste capítulo, assim como, a análise dos resultados obtidos.

Capítulo 5: Conclusões
E finalmente, neste capítulo são apresentadas as conclusões referentes à pesquisa realizada. Aqui, também são descritos os trabalhos futuros referentes à continuidade desta pesquisa.

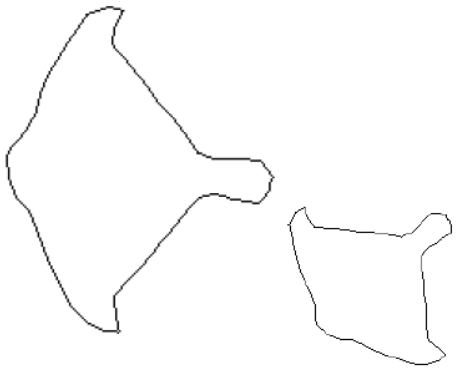
2 FUNDAMENTAÇÃO TEÓRICA

2.1 ANÁLISE ESTATÍSTICA DE FORMAS

Formas de objetos estão disponíveis em diversas áreas de conhecimento, tais como visão computacional, biologia, medicina e arqueologia e a análise de formas é útil na tomada de decisões. A análise estatística de formas é um tópico relativamente recente e está relacionada com características e comparações de formas de objetos. (DRYDEN; MARDIA, 1998) descrevem em seu livro os principais conceitos sobre análise de formas de objetos.

As informações sobre a forma de um objeto é o que resta depois de remover os efeitos da locação, escala e rotação ((KENDALL, 1984)). A Figura 1 ilustra duas vértebras de ratos com diferentes locações, escalas e rotações que precisam ser removidas para gerar a forma destas imagens.

Figura 1 – Duas vértebras de ratos com diferentes locações, escalas e rotações.



Fonte: (DRYDEN; MARDIA, 1998).

O primeiro passo para se obter a forma de um objeto, dentre outros, é adicionar pontos no contorno da imagem, chamados marcos anatômicos. Quando, através de transformações adequadas, são removidos os efeitos de escala e locação a partir das coordenadas de um objeto no espaço dos marcos anatômicos (*landmarks*), um novo conjunto de coordenadas de um objeto pode ser obtido. Este conjunto é chamado de coordenadas de pré-formas (*preshapes*). O novo sistema de coordenadas recebe o nome de espaço das pré-formas. A forma é finalmente obtida removendo a informação de rotação das coordenadas pré-formas do objeto. A informação de rotação é eliminada rotacionando um objeto para que ele fique tão próximo quanto possível de um molde. O novo conjunto de coordenadas do objeto está dentro de um novo espaço, que é chamado espaço de formas.

O espaço de pré-formas e o espaço de formas são espaços não-euclidianos e não é, portanto, possível aplicar uma análise estatística padrão nesses espaços. Para evitar as dificuldades de espaços não-euclidianos, é possível definir uma aproximação linear a esses

espaços. O espaço tangente é uma aproximação linear local ao espaço em um ponto particular. Para uma dada amostra aleatória de objetos, as coordenadas pré-formas desses objetos podem ser projetadas no espaço tangente da forma média amostral. As novas coordenadas são chamadas coordenadas tangentes. No espaço tangente podem ser utilizados muitos procedimentos comumente usados em análise multivariada padrão.

2.1.1 Representação Matemática de Formas

Existem diversas maneiras de extrair dados de formas de objetos, uma abordagem muito utilizada em análise estatística de formas é adicionar um número finito de pontos no contorno de um objeto. Esses pontos são chamados de marcos anatômicos. Marcos anatômicos são pontos para a identificação de locais especiais sobre o objeto e suas coordenadas numéricas são então utilizadas para representar um objeto. A posição dos pontos adicionados aos objetos estão associadas à coordenadas cartesianas e estas coordenadas pertencem a um espaço que é chamado de espaço dos marcos anatômicos. Os marcos anatômicos podem ser fixados pelo conhecimento do pesquisador e/ou questões geométricas. Os marcos podem ser divididos em três tipos:

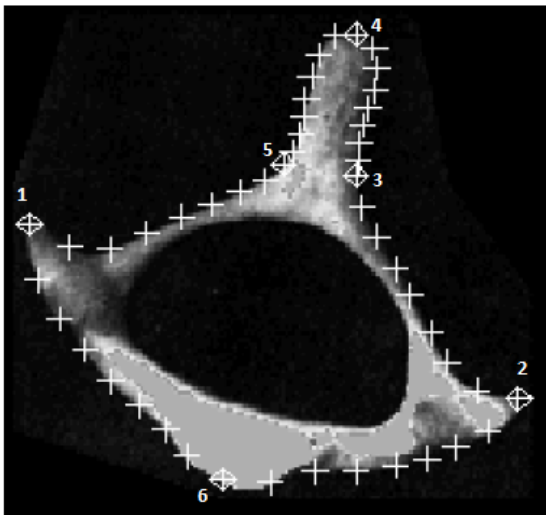
1. **Marcos Anatômicos:** são pontos escolhidos em um objeto, por um especialista, que tem algum significado biológico;
2. **Marcos Matemáticos:** são pontos escolhidos de acordo com alguma propriedade matemática ou geométrica da imagem;
3. **Pseudo-Marcos Anatômicos:** são pontos localizados nas extremidades dos objetos entre os marcos anatômicos ou matemáticos e precisam ser adicionados para a definição da forma.

A Figura 17 mostra uma vértebra de um rato com 6 marcos anatômicos matemáticos, juntos com 54 pseudo-marcos anatômicos aproximadamente equidistantes entre pares de marcos. Estes marcos foram fixados para um experimento de avaliação dos efeitos da seleção para peso corporal na forma das vértebras dos ratos.

Ao submeter os marcos anatômicos a um equipamento de medição, é observado que cada objeto possui tamanho, posição e rotação diferente em relação aos outros objetos. Assim, para se obter uma padronização dos objetos é necessário que sejam removidos os efeitos de locação escala e rotação antes de iniciar a análise dos conjuntos de dados.

Para representar matematicamente a forma de um objeto, inicialmente k marcos anatômicos são escolhidos para o contorno deste objeto. Seja Y uma matriz de coordenadas

Figura 2 – Marcos anatômicos na segunda vértebra torácica *T2* de um rato.



Fonte: (DRYDEN; MARDIA, 1998).

cartesianas $k \times m$ de k marcos anatômicos e m dimensões. As coordenadas desses marcos anatômicos são matematicamente representadas pela seguinte matriz de configuração:

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,m} \\ y_{2,1} & \cdots & y_{2,m} \\ \vdots & \ddots & \vdots \\ y_{k,1} & \cdots & y_{k,m} \end{bmatrix}. \tag{2.1}$$

Nesta pesquisa serão consideradas as formas dos objetos com duas dimensões, isto é, quando $m = 2$ que representa as formas planas dos objetos. Assim, a matriz de configuração passa a ser definida como

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} \\ y_{2,1} & y_{2,2} \\ \vdots & \vdots \\ y_{k,1} & y_{k,2} \end{bmatrix}. \tag{2.2}$$

As coordenadas cartesianas de cada marco anatômico são representadas no espaço real R^m. Quando $m = 2$, formas planas dos objetos são obtidas e o espaço dos marcos anatômicos R^2.

2.1.2 Espaço das Pré-Formas

A forma de uma matriz de configuração é obtida pela informação que permanece depois da eliminação dos efeitos de locação, escala e rotação. Para eliminar estes efeitos, algumas

transformações precisam ser realizadas na matrix Y. O espaço de formas é o conjunto de todas as possíveis formas. Quando $m = 2$ a matriz de configuração pode ser escrita como um vetor complexo (k x 1), definido por:

$$\mathbf{z}^0 = (y_{1,1} + i * y_{1,2}, \ldots, y_{k,1} + i * y_{k,2})^T = (\mathbf{z}^0_{(1)}, \ldots, \mathbf{z}^0_{(k)})^T \tag{2.3}$$

onde os elementos correspondem às coordenadas complexas dos marcos anatômicos. O símbolo ^{0} é usado para indicar que a configuração conserva os efeitos de locação, escala e rotação.

Para iniciar as transformações em busca da forma do objeto, o primeiro passo é remover o efeito de locação. Esta transformação pode ser feita de diversas maneiras dependendo do sistema de coordenadas, mas aqui serão usadas as coordenadas de Kendall ((KENDALL, 1984)) que usa a sub-matriz de Helmert $\mathbf{H}$ para remover a locação (ver (DRYDEN; MARDIA, 1998), p. 34 e ver (SMALL, 1996), p. 130). A matriz de Helmert completa $\mathbf{H}^F$ é uma matriz ortogonal k x k em que a primeira linha tem a todos os elementos iguais a $1/\sqrt{k}$, e tem linhas $j + 1$ para $j \geq 1$ dadas por:

$$(h_j, \ldots, h_j, -jh_j, 0, \ldots, 0), \;\; h_j = -\{j(j+1)\}^{-1/2}, \tag{2.4}$$

com $j = 1, \ldots, k-1$, onde o número de elementos zeros na linha $j+1$ é igual a $k - j - 1$.

Para $k = 4$ a matriz de Helmert completa é dada por:

$$\mathbf{H}^F = \begin{bmatrix} 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} \end{bmatrix}. \tag{2.5}$$

e a sub-matriz de Helmert é

$$\mathbf{H} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} \end{bmatrix}. \tag{2.6}$$

O efeito de locação é removido multiplicando a configuração complexa z^0 pela sub-matriz de Helmert H com dimensão $(k-1)$ x k, que é a matriz de Helmert H^F com a primeira linha removida. A configuração helmertizada é dada por:

$$\mathbf{w} = \mathbf{H}\mathbf{z}^0, \tag{2.7}$$

onde $\mathbf{w}$ é o vetor complexo $\mathbf{z}^0$ com o efeito de locação removido.

Por $\mathbf{H}^F$ ser ortogonal, as configurações helmertizadas são conectadas às configurações centradas pela seguinte propriedade da matriz de Helmert ((DRYDEN; MARDIA, 1998)):

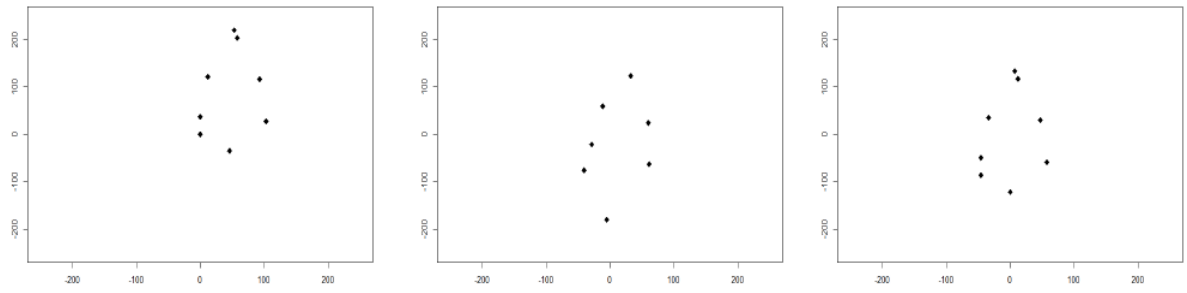
$$\mathbf{H}^T\mathbf{H} = (\mathbf{I}_l - \frac{1}{l}1_l 1_l^T), \tag{2.8}$$

onde $\mathbf{I}_l$ é uma matriz identidade l x l. Assim, se o vetor $\mathbf{z}^0$ (l x 1) é uma configuração complexa, então a matrix

$$\mathbf{z}^C = \mathbf{H}^T\mathbf{w} = \mathbf{H}^T\mathbf{H}\mathbf{z}^0 = (\mathbf{I}_l - \frac{1}{l}\mathbf{1}_l\mathbf{1}_l^T)\mathbf{z}^0 = \mathbf{z}^0 - \frac{1}{l}\sum_{j=1}^{l}\mathbf{z}^0_{(j)}\mathbf{1}_l \tag{2.9}$$

é uma configuração centrada, onde $||\mathbf{z}^C|| = 1$. A configuração centrada é igual a configuração helmetizada multiplicada por $\mathbf{H}^T$. Um exmplo de configuração centrada segue na Figura 3. A Figura 3(a) representa a configuração inicial para um crânio de gorila macho, na Figura 3(b) tem-se a sua configuração helmertizada e na Figura 3(c) é mostrada a configuração centrada deste objeto.

Figura 3 – Configuração centrada.



(a) Configuração original. (b) Configuração Helmertizada. (c) Configuração Centrada.

Fonte: (OLIVEIRA, 2016).

Depois de remoção do efeito de locação, o efeito de escala deve ser removido para gerar os dados das pré-formas $\mathbf{z}$. Então, a pré-forma $\mathbf{z}$ não tem a informação sobre locação e escala. A escala pode ser removida da configuração helmertizada $\mathbf{w}$ dividindo-a pela sua norma:

$$\mathbf{z} = \frac{\mathbf{w}}{||\mathbf{w}||} = \frac{\mathbf{w}}{\sqrt{\mathbf{w}^*\mathbf{w}}} = \frac{\mathbf{H}\mathbf{z}^0}{\sqrt{(\mathbf{H}\mathbf{z}^0)^*\mathbf{H}\mathbf{z}^0}}, \tag{2.10}$$

onde $\mathbf{w}^*$ é transposto conjugado complexo de $\mathbf{w}$ e $||.||$ é a norma complexa de $\mathbf{w}$. O vetor $\mathbf{z}$ é chamado de vetor de pré-forma da configuração complexa $\mathbf{z}^0$ que ainda permanece com a informação de rotação.

O espaço de pré-formas é o espaço de todos os possíveis $k-1$ vetores complexos com as informações de locação e escala removidas. Assim, o espaço de pré-formas planas é uma hiperesfera complexa em $(k-1)$ dimensões complexas definida por:

$$CS^{k-2} = \{\mathbf{z} : \mathbf{z}^*\mathbf{z} = 1, \mathbf{z} \in C^{k-1}\}, \tag{2.11}$$

onde C^{k-1} é o espaço complexo com dimensão $k-1$.

A forma de um objeto é toda a informação geométrica sobre a matriz de configuração $\mathbf{Y}$ que é invariante sobre os efeitos de locação, rotação e escala. O espaço de formas é o espaço de pré-formas com informações de rotação removidas. A forma de um objeto é representada por:

$$[z] = \{e^{i\theta}z : \theta \in [0, 2\pi)\}, \tag{2.12}$$

onde $[z]$ se identifica com qualquer uma de suas versões rotacionadas. O espaço de formas pode ser definido como o conjunto de todas as formas possíveis. O espaço de formas quando $m = 2$ é o espaço complexo projetivo CP^{k-2}, o espaço de linhas complexas passando pela origem, como descrito por ((KENDALL, 1984)).

2.1.2.1 Forma Média

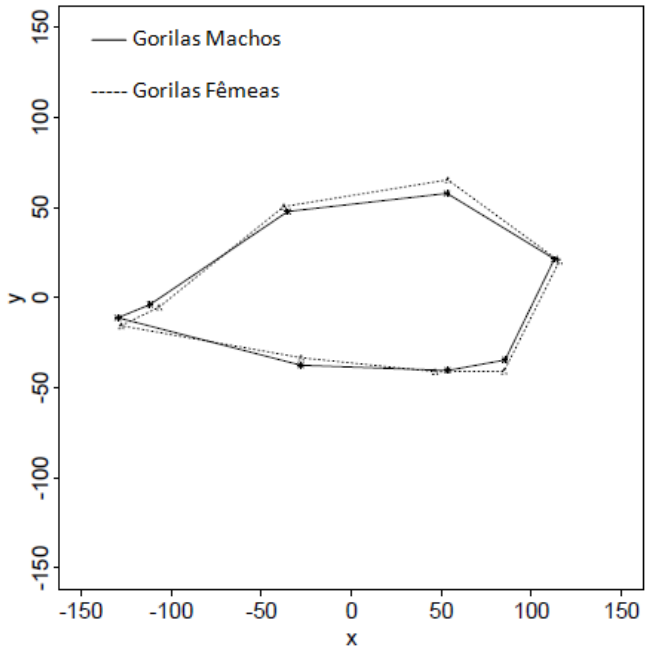
A obtenção da estimativa da forma média das formas dos objetos é um importante conceito relacionado com a análise dos dados em análise estatística de formas. A Análise de Procrustes é uma técnica que tem o objetivo de *casar* dois objetos, ou seja, ajustar a pré-forma de cada objeto à forma média dos objetos. Quando dois ou mais objetos são considerados, eles podem ter diferentes rotações, locações e escalas. Este ajuste de objetos é feito usando as pré-formas desses objetos porque as pré-formas têm a mesma locação e escala.

Considere $z_1^0, \ldots, z_n^0$, em que z_i^0 foi definido na Equação 2.3, como amostras aleatórias de configurações complexas de uma população de n objetos. Sejam $z_1, \ldots, z_n$ as pré-formas de $z_1^0, \ldots, z_n^0$, em que z_i está definida na Equação 2.10. A forma média Procrustes completa $\hat{\mu}$, definida por (KENT, 1994), pode ser encontrada como o autovetor correspondente ao maior autovalor da soma de quadrados complexa da matriz produto

$$S = \sum_{i=1}^{n} w_i w_i^* / (w_i^* w_i) = \sum_{i=1}^{n} z_i z_i^*, \tag{2.13}$$

Deste modo, $\hat{\mu}$ é dado pelo autovetor complexo correspondente ao maior autovalor de S. Todas rotações de $\hat{\mu}$ são também soluções, mas todos estas correspondem a mesma forma $[\hat{\mu}]$. A Figura 4 ilustra a forma média para o conjunto de dados de crânios de gorilas que tem 29 gorilas machos e 30 gorilas fêmeas ((DRYDEN; MARDIA, 1998)).

Figura 4 – Forma média Procrustes completa de crânios de gorilas machos e fêmeas.



Fonte: (DRYDEN; MARDIA, 1998).

As novas coordenadas resultantes do ajustamento das pré-formas dos objetos à sua forma média, para uma dada amostra de pré-formas $\mathbf{z}_i$, são chamadas ajustes Procrustes ou coordenadas Procrustes.

Seja $\mathbf{z}_1, \ldots, \mathbf{z}_n$ uma amostra aleatória de pré-formas e sejam $\mathbf{w}_1, \ldots, \mathbf{w}_n$ as correspondentes configurações helmertizadas. As configurações têm uma rotação arbitrária ((DRYDEN; MARDIA, 1998)). Antes de continuar com a análise estatística de formas, é necessário rotacionar todas as configurações de tal maneira que estejam o mais próximo possível da forma média amostral e isto é feito pela seguinte equação:

$$\mathbf{w}_i^P = \frac{\mathbf{w}_i^* \hat{\mu} \mathbf{w}_i}{(\mathbf{w}_i^* \mathbf{w}_i)}, \quad i = 1, \ldots, n, \tag{2.14}$$

onde $\mathbf{w}_1^P, \ldots, \mathbf{w}_n^P$ são os ajustes Procrustes completos ou coordenadas Procrustes completas. Uma vez que as pré-formas podem ser escritas como $\mathbf{z}_i = \mathbf{w}_i / ||\mathbf{w}_i||$ e $||\mathbf{w}_i|| = \sqrt{\mathbf{w}_i^* \mathbf{w}_i}$, as coordenadas Procrustes também podem ser calculadas como

$$\mathbf{w}_i^P = \mathbf{z}_i^* \hat{\mu} \mathbf{z}_i, \quad i = 1, \ldots, n. \tag{2.15}$$

onde cada $\mathbf{w}_i^P$ é o ajuste Procrustes completo de $\mathbf{w}_i$ em $\hat{\mu}$.

2.1.2.2 Distâncias para Formas Planas

Um outro conceito fundamental para a análise estatística de formas é o paradigma de distância que complementa a definição do espaço de forma não Euclidiana. Em seu livro, (DRYDEN; MARDIA, 1998) mostram as medidas de distâncias de Procrustes que podem ser utilizadas para o caso de formas planas de objetos ($m = 2$).

Considere duas configurações de pré-formas centradas $\mathbf{z}_y^C = (\mathbf{z}_{y1}^C, \ldots, \mathbf{z}_{yk}^C)^T$ e $\mathbf{z}_w^C = (\mathbf{z}_{w1}^C, \ldots, \mathbf{z}_{wk}^C)^T$, em que $||\mathbf{z}_y^C|| = 1 = ||\mathbf{z}_w^C||$ e $\mathbf{z}_y^{C*}1_k = 0 = \mathbf{z}_w^{C*}1_k$. A distância de Procrustes completa d_F é definida por

$$d_F^{(2)} = \sqrt{1 - |\mathbf{z}_y^{C*}\mathbf{z}_w^C|^2}. \tag{2.16}$$

Esta distância é invariante aos efeitos de locação, escala e rotação. Assim, pode-se considerar

$$cos\rho = (1 - d_F^{(2)})^{1/2}. \tag{2.17}$$

Para dados no plano o espaço de pré-formas é uma esfera complexa CS^{k-2} de raio unitário em dimensão complexa $k - 1$. A distância de Procrustes ρ é o ângulo entre as pré-formas complexas z_y^C e z_w^C definida por

$$\rho = arccos(|\mathbf{z}_y^{C*}\mathbf{z}_w^C|), \tag{2.18}$$

o valor de ρ não é afetada pelo efeito de rotação das pré-formas. Como o raio da esfera da pré-forma é 1 pode-se considerar ρ como sendo a distância ótima no círculo na esfera da pré-forma. Essa distância também é conhecida como geodésica.

A distância Procrustes parcial entre $\mathbf{z}_y^C$ e $\mathbf{z}_w^C$ e é dada por

$$d_P^{(2)} = \sqrt{2(1 - cos\rho)}. \tag{2.19}$$

a distância d_P também é invariante quanto a rotação.

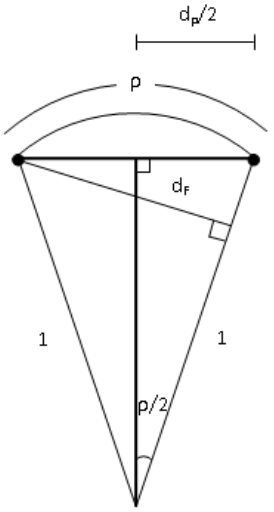
A Figura 5 ilustra as distâncias de Procrustes 2.16, 2.18 e 2.19 na hiperesfera de pré-forma.

2.1.2.3 Modelando Dados de Forma: Distribuição *Bingham* Complexa

A distribuição de *Bingham* complexa ((KENT, 1994)) é uma das mais importantes distribuições para conjuntos de dados de marcos anatômicos com duas dimensões. O espaço de pré-formas consiste de uma esfera unitária complexa em $k - 1$ dimensões CS^{k-1} definida em 2.11. Se $\mathbf{z}$ é um vetor unitário complexo aleatório com distribuição de *Bingham* complexa (com notação $CB_{k-1}(A)$), a função densidade de probabilidade $\mathbf{z}$ é dada por

$$f(\mathbf{z}) = c(A)^{-1}exp(\mathbf{z}^*A\mathbf{z}), \quad \mathbf{z} \in CS^{k-1} \tag{2.20}$$

Figura 5 – Relação entre as distâncias de Procrustes $d_F^{(2)}$, ρ e $d_P^{(2)}$.



Fonte: (DRYDEN; MARDIA, 1998).

onde A é uma matrix Hermitiana $(k-1)$ x $(k-1)$ e $c(A)$ uma constante normalizadora. Se $A = I$, $f(\mathbf{z})$ torna-se uma distribuição uniforme sobre CS^{k-1}; devido a restrição $z^*z = 1$.

A distribuição de *Bingham* complexa tem a propriedade de simetria complexa, em que $\mathbf{z}$ e $e^{i\theta}\mathbf{z}$, onde $\theta \in [0, 2\pi)$, tem a mesma distribuição (ver (KENT, 1994), p. 290):

$$f(e^{i\theta}\mathbf{z}) = f(\mathbf{z}), \tag{2.21}$$

sendo, assim, não influenciada pelas rotações dos dados de pré-formas $\mathbf{z}$. Esta propriedade faz com que a distribuição Bingham complexa seja adequada para análise de dados de formas em duas dimensões, em que locação e escala foram previamente removidos para $\mathbf{z}$, e uma vez que uma distribuição de forma deve respeitar a definição da forma dada em 2.12.

A simulação da distribuição de Bingham complexa foi proposta por vários métodos em (KENT; CONSTABLE; ER, 2004). Nesta pesquisa foi utilizado o método Truncação pelo simplex. Inicialmente, $(k-1)$ exponenciais truncadas $Texp(\lambda_j)$ são geradas, e, em seguida, estas variáveis aleatórias são expressas em coordenadas polares para proporcionar uma distribuição de Bingham complexa. O Algoritmo 1 define a geração da distribuição

exponencial truncada para a simulação da distribuição de Bingham complexa.

Algoritmo 1: Simulação da distribuição exponencial truncada

1. Simule uma variável aleatória $U_j \sim U[0,1] \quad j = 1, \ldots, k-1$.
2. Seja $S'_j = -(1/\lambda_j)log(1 - U_j(1 - e^{-\lambda_j}))$, $S' = (S'_1, \ldots, S'_{k-1})^T$ de modo que S'_j são amostras aleatórias independentes Texp(λ_j).
3. Se $\sum_{j=1}^{k-1} S'_j < 1$, estabelece $S = S'$. Caso contrário, rejeita-se S' e retorne ao passo 1.

O método para simular uma distribuição Bingham complexa usa $(k-1)$ exponenciais truncadas para gerar um vetor k de uma distribuição Bingham complexa. Seja $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k = 0$ os autovalores de $-A$. Com $\lambda = (\lambda_1, \ldots, \lambda_{k-1})$ sendo o vetor dos primeiros $k-1$ autovalores. O Algoritmo 2 retorna um vetor k, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_k)$ o qual tem um distribuição Bingham complexa.

Algoritmo 2: Simulação da distribuição de Bingham complexa

1. Gere $S = (S_1, \ldots, S_{k-1})$ onde $S_j \sim Texp(\lambda_j)$ são variáveis aleatórias independentes usando o Algoritmo 1.
2. Se $\sum_{j=1}^{k-1} S'_j < 1$, faça $S_k = 1 - \sum_{j=1}^{k-1} S'_j$. Caso contrário, volte ao passo 1.
3. Gere ângulos independentes $\theta_j \sim U[0, 2\pi), \quad j = 1, \ldots, k$.
4. Calcule $z_j = S_j^{1/2}\exp(i\theta_j), \quad j = 1, \ldots, k$.

A Tabela 1 ilustra a formação de dados simulados para formas planas gerados pela distribuição de Bingham complexa contendo 4 marcos anatômicos e dimensão $m = 2$.

Tabela 1 – Dados Simulados com 4 Marcos Anatômicos Modelados pela Distribuição de Bingham

	Objeto 1		**Objeto 2**		**Objeto 3**	
Marcos/Dimensão	1	2	1	2	1	2
1	-0.03611885	0.2811598	0.1535699	0.2621857	-0.1715638	0.2217322
2	-0.03192184	0.2902611	0.1088777	0.2500708	-0.1903721	0.2343956
3	-0.04447181	0.2872104	0.1248484	0.2617987	-0.1691181	0.2277718
4	0.11251250	-0.8586312	-0.3872960	-0.7740552	0.5310540	-0.6838996

Fonte: Própria.

2.1.3 Espaço Tangente

O espaço linearizado do espaço de formas é o espaço de coordenadas tangentes que se aproxima de um ponto particular do espaço de formas. Uma das maiores vantagens do

espaço tangente é que técnicas padrões de análise multivariada podem ser usadas diretamente. A variabilidade da análise de forma pode ser tratada no espaço tangente. O espaço tangente do espaço de formas CP^{k-2} no ponto z é o espaço vetorial de todos os vetores tangentes a CP^{k-2} no ponto $\mathbf{z}$. Dentre os diferentes tipos de coordenadas tangentes existentes, esta pesquisa usará as coordenadas tangente Procrustes parciais, que são dadas por

$$\mathbf{t}_i = e^{i\hat{\theta}}[I_{k-1} - \hat{\mu}\hat{\mu}*]\mathbf{z}_i, i = 1, \ldots, n, \tag{2.22}$$

onde $\mathbf{z}_i$ é o vetor de pré-formas definido pela Equação 2.10 e $\hat{\theta}$ minimiza $||\hat{\mu} - \mathbf{z}e^{i\hat{\theta}}||$. Considere $\mathbf{z}_1, \ldots, \mathbf{z}_n$ uma amostra aleatória de pré-formas e $\mathbf{t}_1, \ldots, \mathbf{t}_n$ suas coordenadas tangentes. Seja $\mathbf{v}_i$ um vetor de tamanho $2k-2$ obtido empilhando as partes real e imaginária das coordenadas de cada $\mathbf{t}_i$, essa operação é representada por

$$\mathbf{v}_i = cvec(\mathbf{t}_i) = (\mathbf{Re}(\mathbf{t}_i)^T, \mathbf{Im}(\mathbf{t}_i)^T)^T, \tag{2.23}$$

onde $\mathbf{Re}(\mathbf{t}_i)$ é a parte real e $\mathbf{Im}(\mathbf{t}_i)$ é a parte imaginária de $\mathbf{t}_i$. Se o número de marcos anatômicos é k, um vetor pré-forma $\mathbf{z}_i$ tem dimensão $(k-1)$ e seu correspondente vetor de coordenadas tangentes $\mathbf{v}_i$, tem dimensão $(2k-2)$.

A Tabela 2 ilustra a conversão dos dados simulados gerados na Tabela 1 em dados do espaço tangete.

Tabela 2 – Dados Simulados com 4 Marcos Anatômicos Convertidos para o Espaço Tangente

Marcos	Objeto 1	Objeto 2	Objeto 3
1	-1.168858e-02	2.680412e-02	-8.901039e-03
2	3.453824e-03	2.367168e-03	-5.745177e-03
3	9.823714e-06	-5.041838e-05	-1.078913e-05
4	-3.195904e-03	6.119755e-03	-1.563755e-02
5	-3.323071e-03	-2.074764e-03	1.204759e-02
6	4.587780e-05	-1.513549e-04	1.200385e-04

Fonte: Própria.

Em situações de alta concentração dos dados, a distância Euclidiana no espaço tangente é considerada uma boa aproximação para as distâncias de Procrustes, definidas na seção 2.1.2.2, para o espaço de formas ((DRYDEN; MARDIA, 1998)). Considere duas configurações $\mathbf{Y}_1$ e $\mathbf{Y}_2$ com $\mathbf{v}_1$ e $\mathbf{v}_2$ sendo dois vetores de coordenadas tangentes reais definidas como na Equação 2.23, respectivamente. As medidas de dissimilaridade mais comuns modeladas de acordo com o espaço tangente entre $\mathbf{v}_1$ e $\mathbf{v}_2$ são mostradas nas próximas quatro equações.

1. **Distância de Manhattan**

$$d(\mathbf{v}_1, \mathbf{v}_2) = \sum_{j=1}^{n} |v_{1j} - v_{2j}|. \tag{2.24}$$

2. **Distância Euclidiana**

$$d(\mathbf{v}_1, \mathbf{v}_2) = \sqrt{\sum_{j=1}^{n} (v_{1j} - v_{2j})^2}. \tag{2.25}$$

3. **Distância de Minkowski**

$$d(\mathbf{v}_1, \mathbf{v}_2) = \sqrt[p]{\sum_{j=1}^{n} |v_{1j} - v_{2j}|^p}. \tag{2.26}$$

4. **Distância de Mahalanobis**

$$d(\mathbf{v}_1, \mathbf{v}_2) = \sqrt{(v_1 - v_2)^T S^{-1} (v_1 - v_2)}, \tag{2.27}$$

em que S^{-1} é a matriz de covariância.

Apesar das coordenadas tangentes serem úteis na análise de formas, por tornarem aplicáveis os métodos padrões de análise multivariada, foram observadas suas limitações por serem uma aproximação do espaço de formas. Nos trabalhos de (AMARAL; DRYDEN; WOOD, 2007) e (AMARAL et al., 2010) foram relatadas limitações das coordenadas tangentes quando utilizadas em dados de baixa concentração.

2.2 ANÁLISE DE AGRUPAMENTOS

A ampla quantidade de informações disponibilizada no dias atuais, que são representadas como dados, precisa ser analisada e gerenciada. Um dos meios essenciais para lidar com esses dados é classificá-los ou agrupá-los em um conjunto de categorias ou grupos. Na verdade, como uma das atividades mais primitivas dos seres humanos, a classificação tem um papel importante e indispensável na longa história do desenvolvimento humano ((XU; WUNSCH, 2005)). O conceito de agrupamento está relacionado a diversos ramos do conhecimento, fazendo parte das pesquisas de muitas áreas, tais como Reconhecimento de Padrões, Estatística, Matemática, Engenharia e Física, sendo usado em várias aplicações, como medicina, psiquiatria, serviços sociais, pesquisa de mercado, educação e arqueologia.

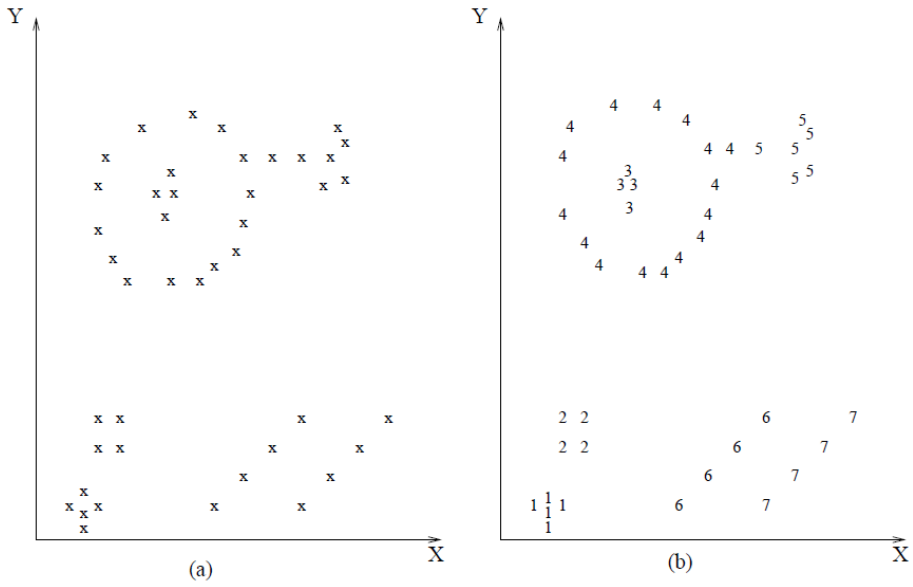
Os sistemas de classificação podem ser divididos em supervisionados ou não supervisionados. Na classificação supervisionada, o mapeamento de um conjunto de vetores de dados de entrada é feito para um conjunto finito de classes rotuladas discretas. Por outro lado, na classificação não supervisionada, chamada de agrupamento ou análise exploratória de dados, não há dados rotulados disponíveis. Sob certo ponto de vista, rótulos

estão presentes na atividade de agrupamento, cada grupo formado poderia ser entendido como um rótulo, mas estes rótulos são obtidos a partir dos próprios dados. O problema de agrupamento é o de separar um conjunto de dados em um número de grupos (*clusters*) com base em alguma medida de similaridade.

A Análise de Agrupamento *(Cluster Analysis)* é o termo usado para os procedimentos que buscam desvendar grupos em dados e tem por objetivo organizar um conjunto inicial de itens em grupos tal que os itens em um dado grupo têm alto grau de similaridade (semelhança), enquanto itens pertencentes a grupos diferentes têm um alto grau de dissimilaridade (diferenças) em relação aos outros grupos. Normalmente, um protótipo local também é produzido, ele caracteriza os membros de um grupo. A estrutura do dado é inferida através da análise dos grupos resultantes (e/ou seus protótipos) ((CHERKASSKY; MULIER, 2007)).

Um importante conceito utilizado em Análise de Agrupamento é partição, a qual é caracterizada por conter um definido número de grupos, tais que um elemento em cada grupo não pode pertencer a qualquer um outro, isto é, são disjuntos. Em algumas circunstâncias, no entanto, grupos sobrepostos podem fornecer uma solução mais aceitável ((EVERITT et al., 2011)). Um exemplo de agrupamento é mostrado na Figura 6. Os padrões de entrada são mostrados na Figura 6(a) e os grupos desejados são mostrados na Figura 6(b). Pontos pertencendo ao mesmo grupo receberam o mesmo rótulo.

Figura 6 – Exemplo de Agrupamento de Dados.



Fonte: (JAIN; MURTY; FLYNN, 1999).

Um fundamental conceito para Análise de Agrupamento é a medida de distância a ser usada para formas os grupos. A distância entre dois conjuntos **A** e **B** pode ser calculada de diversas maneiras, dependendo da aplicação do problema, e é fundamental para definir para qual grupo será alocado cada objeto e, conseqüentemente, irá determinar a forma

dos grupos. Considere que cada objeto $\boldsymbol{a} \in \mathbf{A}$ é descrito por variáveis contínuas. A seguir são mostradas algumas das medidas de dissimilaridade mais comuns.

1. **Distância de Manhattan**

$$d(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{n} |a_j - b_j|. \tag{2.28}$$

2. **Distância Euclidiana**

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{j=1}^{n} (a_j - b_j)^2}. \tag{2.29}$$

3. **Distância de Minkowski**

$$d(\mathbf{a}, \mathbf{b}) = \sqrt[p]{\sum_{j=1}^{n} |a_j - b_j|^p}. \tag{2.30}$$

4. **Distância de Mahalanobis**

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{(a-b)^T S^{-1} (a-b)}, \tag{2.31}$$

onde S^{-1} é a matriz de covariância.

A distância de *Manhattan*, também conhecida como distância *City-block*, é equivalente ao espaço L_1, a distância Euclidiana corresponde ao espaço L_2 e *Minkowski* ao espaço L_p. Esta última pode ser considerada uma generalização das duas medidas anteriores. Quando $p \to \infty$, a distância é a conhecida como *Chebyshev* e pode ser calculada como $d(a, b) = max(|a_1 - b_1|, |a_2 - b_2|, \ldots, |a_m - b_m|)$ ((SOUZA; CARVALHO, 2004)).

Na aprendizagem não-supervisionada, como ocorre nos métodos de agrupamento, não há a necessidade de informações *a priori* sobre o domínio avaliado, levando-se em consideração, apenas, a disposição dos dados e suas propriedades internas. Existem diversas técnicas para a estimação do número ideal de conjuntos finais que devem ser criados de maneira a tornar a divisão dos dados mais representativa para o problema estudado, como apresentado em (SERGIOS; KOUTROMBAS, 2006). Além disso, podem ser classificados de acordo com a forma que interpretam os dados e a maneira com que esses objetos se organizam em agrupamentos.

Diferentes abordagens têm sido propostas para agrupar dados. Em análise de dados, distingui-se dois grandes grupos de métodos: hierárquicos e particionais ((EVERITT et al., 2011); (GORDON, 1999)). Métodos hierárquicos formam sequências de partições dos dados de entrada gerando assim hierarquias completas, enquanto métodos de particionamento procuram obter uma simples partição dos dados de entrada em um número fixo de grupos.

2.2.1 Métodos Hierárquicos

Algoritmos de agrupamentos baseados no método hierárquico organizam um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos. Os resultados de um algoritmo hierárquico são normalmente mostrados como uma árvore binária ou um dendograma. A raiz do dendograma representa o conjunto de dados inteiro e os nós folha representam os indivíduos. Os nós intermediários representam a magnitude da proximidade entre os indivíduos. A altura do dendograma expressa a distância entre um par de indivíduos ou entre um par de grupos ou ainda entre um indivíduo e um grupo. O resultado do agrupamento pode ser obtido cortando-se o dendograma em diferentes níveis. Esta forma de representação fornece descrições informativas e visualização para as estruturas de grupos em potencial, especialmente quando há realmente relações hierárquicas nos dados como, por exemplo, em dados de pesquisa sobre evolução de espécies.

Algoritmos hierárquicos são divididos em aglomerativos ou divisivos. Algoritmos aglomerativos começam com cada indivíduo no conjunto de dados sendo representado por um único grupo. Uma série de combinações entre os grupos iniciais ocorrem resultando depois de um número de passos em que todos os indivíduos estão em um mesmo grupo. No início da execução do algoritmo aglomerativo, como todos os grupos são constituídos de 1 elemento, a matriz de dissimilaridade entre os grupos é a mesma obtida quando calcula-se essa métrica por pares de elementos.

Algoritmos divisivos operam na direção oposta ao método aglomerativo. No começo, o conjunto de dados inteiro está em um só grupo e um procedimento que divide os grupos vai ocorrendo sucessivamente até que cada indivíduo seja um único grupo. Para um conjunto com N indivíduos há $2^{N-1} - 1$ diferentes possíveis divisões do conjunto em dois subconjuntos. Os algoritmos divisivos podem ser divididos em dois principais grupos: monotéticos (*monothetics*), que dividem o conjunto de dados levando-se em conta apenas um único atributo, e politéticos (*polythetics*), cujas divisões são realizadas a partir dos diversos atributos que um padrão pode ter.

No contexto de métodos de agrupamento, uma potencial vantagem dos métodos divisivos sobre os métodos aglomerativos pode ocorrer quando o objetivo é encontrar uma partição dos dados em um relativamente pequeno número de grupos ((HASTIE; TIBSHIRANI; FRIEDMAN, 2009)). Estes métodos podem funcionar em conjunto com métodos de particionamento tais como K-médias ou K-Medoids aplicados recursivamente na função de dividir o grupo anterior em dois grupos distintos tendo o número de grupos $k = 2$. Entretanto, essa modificação irá depender da configuração inicial em cada etapa do método hierárquico divisivo.

2.2.2 Métodos de Particionamento

Ao contrário de métodos hierárquicos, métodos particionais associam um conjunto de indivíduos a K grupos sem criar uma estrutura hierárquica. Partindo de partições e protótipos iniciais, os elementos mudam de um grupo para outro formando novos grupos e a forma destes grupos vai se modificando até chegarem a partição final. Em princípio, a partição ótima, baseada em algum critério específico, pode ser encontrada enumerando-se todas as possibilidades, mas esta abordagem exaustiva é inviável pelo tempo computacional que seria necessário para fazer uma busca por todas as combinações possíveis.

Em geral esses métodos são divididos em dois tipos de paradigmas: os do tipo rígidos (*hard*) e os do tipo difusos (*fuzzy*). Os rígidos se caracterizam por considerarem as classes disjuntas, ou seja, não existem elementos em comum em dois diferentes grupos, desta forma, cada indivíduo pertence a somente um grupo. Por outro lado, os algoritmos do tipo difuso estendem esse conceito de associação de cada elemento em uma classe: um indivíduo pode pertencer a diversas classes de acordo com uma função de pertinência capaz de associar cada padrão a cada um dos grupos assumindo valores no intervalo $[0, 1]$.

2.2.2.1 Métodos Rígidos

Consistem em obter uma partição a partir de um determinado conjunto de n elementos agrupados em um número pré-definido de k grupos, onde $k \leq n$, de forma que cada grupo possua pelo menos um elemento e cada elemento deve pertencer unicamente a um grupo. Estas funções se caracterizam por analisar os grupos em cada iteração e usar métricas para fornecer informações sobre o particionamento. Um dos fatores importantes em métodos particionais é a função objetivo ou critério. A soma de erros quadráticos é uma das funções mais amplamente utilizadas como critério ((XU; WUNSCH, 2005); (JAIN; MURTY; FLYNN, 1999)). O problema da otimização de funções de custo consiste em fazer uso da função e encontrar a melhor solução dentre todas as possíveis soluções objetivando minimizar ou maximizar um critério antes estabelecido.

O mais clássico algoritmo de particionamento rígido é o K-médias (K-*means*) introduzido por (MACQUEEN, 1967). Os grupos homogêneos são gerados minimizando o erro do agrupamento definido, geralmente, como a soma das distâncias euclidianas quadradas entre cada conjunto de dados e os correspondentes representantes (centróides). Estes representantes são definidos através das médias dos elementos em cada grupo e a minimização do critério de agrupamento se dará pela formação das partições que apresentem maior similaridade dentro dos grupos e menor similaridade entre os grupos obtidos pelos valores das distâncias dos objetos aos centróides. O K-médias reconhece apenas regiões esféricas devido a sua distância fixa.

Diversos pesquisadores modificaram o algoritmo K-médias em diferentes aspectos. Dentre eles, (HANSEN; MLADENOVIC, 2001) propuseram o algoritmo J-médias (J-*means*)

que é uma nova heurística de busca local para resolver o problema de agrupamento minimizando a soma dos quadrados. A versão do K-médias de (HONG; KWONG, 2009) propõe agrupar os elementos do conjunto de dados de acordo com suas incertezas de agrupamento usando múltiplos algoritmos de agrupamento. (ZONG Y.; PAN, 2013) propuseram um agrupamento K-médias melhorado com base no "conhecimento acumulado local". A ideia dessa abordagem é capturar a informação comum dos resultados de agrupamento e utilizar como conhecimento aprendido do conjunto de dados e, em seguida, acumular as informações de múltiplas execuções do agrupamento com inicializações diferentes como a solução final ideal do agrupamento. Para evitar o problema com dados ruidosos, (WANG; SU, 2011) propuseram um algoritmo K-médias aprimorado, no qual uma filtragem é feita com os dados de entrada para remover os dados ruidosos, de modo que esses dados de ruídos não possam participar no cálculo dos protótipos dos grupos iniciais. Ao remover esses dados ruidosos, o resultado do agrupamento é melhorado significativamente e o impacto dos dados de ruído no algoritmo K-médias diminui de forma eficaz e os resultados do agrupamento são mais precisos.

Outro conhecido algoritmo de particionamento é o baseado em *medoids* ((KAUFMAN; ROUSSEEUW, 1987)). O algoritmo K-médias é sensível a observações aberrantes já que um objeto com valor extremamente grande pode substancialmente distorcer a distribuição dos dados. Ao invés de utilizar o valor médio dos objetos no grupo como um ponto de referência, um *medoid* pode ser usado, que é o objeto mais centralmente localizado no grupo. Deste modo, o método de particionamento pode ainda ser fornecido baseado no princípio da minimização da soma das dissimilaridades entre cada objeto e seu correspondente ponto de referência. Isto forma a base do método *K-medoids* ((HAN; KAMBER; PEI, 2006); (VELMURUGAN; SANTHANAM, 2011)).

O Particionamento em Torno de *Medoids* (*Partitioning Around Medoids - Program PAM*) foi um dos primeiros algoritmos *K-medoids* introduzidos. O algoritmo usado no programa *PAM* é baseado na busca de k objetos representativos entre os objetos do conjunto de dados. Na literatura de análise de agrupamentos tais objetos representativos são frequentemente chamados centróides. No algoritmo *PAM* os objetos representativos são os chamados *medoids* dos grupos. Após encontrar um conjunto de k objetos representativos, os k grupos são construídos atribuindo cada objeto do conjunto de dados para o objeto representativo mais próximo. Alternativamente o programa pode ser usado com a entrada por uma matriz de dissimilaridades entre objetos.

2.2.2.2 Métodos Difusos

Agrupamento difuso permite associar um indivíduo com todos os grupos usando uma função de pertinência ((ZADEH, 1965)). Tem sido aplicado com sucesso em vários campos, incluindo finanças e marketing. Teoria dos conjuntos difusos foi inicialmente aplicada para o agrupamento em (RUSPINI, 1969). O livro de (BEZDEK, 1981) é uma boa fonte de

material sobre agrupamento difuso. Mais recentemente, diversos trabalhos exploraram o conceito difuso/nebuloso, utilizando variadas abordagens, têm sido relatados na literatura, como (YANG, 1993), (PIMENTEL; SOUZA, 2013), (WU, 2012), (CARVALHO F. A. T.; LECHEVALLIER, 2015) e (MACIEL D. B. M.; PIMENTEL, 2017).

Na prática a separação entre grupos é uma noção difusa. Nos algoritmos do tipo difuso um indivíduo pode pertencer a diversas classes de acordo com uma função de pertinência capaz de associar cada padrão a cada um dos grupos assumindo valores no intervalo $[0, 1]$. Neste caso, cada classe é um conjunto nebuloso de todos os objetos. Cada elemento x possui um grau de pertinência para um grupo k, de forma que a soma de todos os graus relativos a esse elemento x deve ser igual a 1. Uma desvantagem desse tipo de algoritmo é a definição da função de pertinência. Diferentes funções são usadas, entre elas estão as baseadas em centróides de grupos. Outra desvantagem é a dependência da escolha inicial dos protótipos das classes, como ocorre no algoritmo rígido K-médias.

Seja Ω um conjunto de dados com n objetos, que é descrito por p atributos, e seja c um inteiro entre 1 e n que representa o número de grupos. Uma partição difusa $U = (u_{ki})$ é definida por uma matriz $c \times n$ que satisfaz

$$\begin{aligned} u_{ki} \in [0,1], \quad 1 \leq k \leq c; 1 \leq i \leq n \\ \sum_{k=1}^{c} u_{ki} = 1, \quad 1 \leq i \leq n, \\ 0 < \sum_{i=1}^{n} u_{ki} < n, \quad 1 \leq k \leq c, \end{aligned} \tag{2.32}$$

onde u_{ki} representa o grau de pertinência do i-ésimo objeto ao k-ésimo grupo.
A partição rígida de u_{ki} é definida como

$$w_{ki} = \begin{cases} 1 & \text{se } k = \text{argmax}_{1 \leq k \leq c}\, u_{ki}, \\ 0 & \text{caso contrário} \end{cases} \tag{2.33}$$

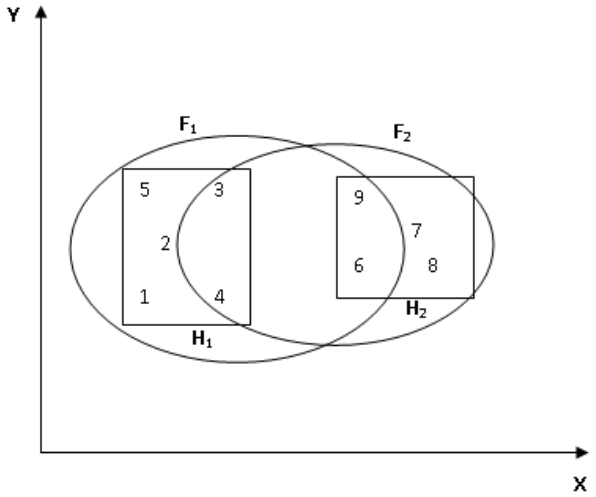
onde o objeto é alocado ao grupo com maior grau de pertinência.

O algoritmo de agrupamento difuso mais popular é o algoritmo c-médias difuso (*fuzzy c-means*). No algoritmo c-médias difuso os elementos mais distantes do centróide possuem um menor grau de pertinência, enquanto aqueles mais próximos ao centróide tem uma pertinência maior e o centróide é obtido fazendo-se uma média ponderada de todos os indivíduos daquele grupo. O projeto de funções de pertinência é o problema mais importante na agregação difusa ((VANISRI; LOGANATHAN, 2010)). Semelhante ao K-médias, o c-médias difuso tem problemas para lidar com ruídos e anomalias nos dados e também pode levar a um mínimo local. Uma das vantagens deste método em relação ao método rígido K-médias é a característica de encontrar valores de resultados mais eficazes quando

os dados são dispostos com sobreposição, pois objetos de diferentes classes possuem características similares.

A Figura 7 ilustra a ideia de grupos rígidos e difusos. Os retângulos dividem o conjunto de dados em dois grupos rígidos: $H_1 = 1, 2, 3, 4, 5$ e $H_2 = 6, 7, 8, 9$. Um algoritmo de agrupamento difuso poderia produzir os dois grupos difusos F_1 e F_2 representados por elipses.

Figura 7 – Exemplo de agrupamento rígido x difuso. O agrupamento rígido é representado pelos grupos H^1 e H^2; e o agrupamento difuso é representado pelos grupos F^1 e F^2.



Fonte: (JAIN; MURTY; FLYNN, 1999).

2.2.2.3 Métodos Baseados em Núcleo

Um dos grandes desafios em análise de agrupamentos está em realizar a separação dos grupos quando os dados não são linearmente separáveis. A distância Euclidiana é a medida de dissimilaridade mais comum para algoritmos de agrupamento particionais (rígidos e difusos). A distância Euclidiana funciona bem com conjuntos de dados em que os grupos são hiperesféricos e linearmente separáveis. Quando ocorre da estrutura dos dados possuir uma complexidade maior com grupos não hiperesféricos e/ou padrões não linearmente separáveis, isso pode prejudicar o desempenho dos algoritmos de agrupamento. Para tentar escapar dessa limitação, vários métodos que são capazes de lidar com esses tipos de dados foram propostos, entre eles, métodos de agrupamento baseados em funções de núcleo (*Kernel functions*) ((CRISTIANINI; SHAEW-TAYLOR, 2000); (FERREIRA M. R.P.; SIMõES, 2016)).

A essência dos métodos baseados em núcleo envolve um mapeamento não-linear do espaço original para um espaço de características de alta dimensão, com o objetivo de aumentar o poder computacional dos métodos lineares. O raciocínio é que em espaços de alta dimensão pode ser possível obter grupos bem definidos e linearmente separáveis. Esta

técnica é conhecida como *Mercer kernel*, uma alternativa muito utilizada em problemas de caráter não linear ((GIROLAMI, 2002)).

Uma vez que a escolha do núcleo e da medida de similaridade/dissimilaridade é crucial nestes métodos, muitas técnicas foram propostas para aprender automaticamente a forma dos núcleos a partir dos dados. As modificações dos métodos são regidas por duas abordagens principais: os protótipos dos agrupamentos são obtidos no espaço original e as distâncias entre objetos e protótipos são calculadas por meio de núcleos; e agrupamento em espaço de características, no qual os protótipos são obtidos no espaço de características ((FILIPPONEA M.; ROVETTAA, 2008)).

Dentre os diversos autores que abordam métodos baseados em núcleo em suas pesquisas, (MULLER K.-R.; SCHOLKOPF, 2001) forneceram uma introdução para as técnicas de *support vector machines (SVMs)*, *kernel Fisher discriminant analysis* e *kernel principal component analysis (PCA)*, como exemplos para métodos de aprendizado com base em núcleo. O método *Kernel* K-médias foi desenvolvido por (GIROLAMI, 2002) e tem a capacidade de gerar hiper-superfícies não-lineares que podem separar grupos transformando o espaço de entrada para um espaço de características de mais alta dimensão e depois realizar a agregação dentro deste espaço de características. (FILIPPONEA M.; ROVETTAA, 2008) apresentaram um levantamento de métodos de agrupamento baseados em núcleo. Os métodos de agrupamento apresentados são versões baseadas em núcleo de muitos algoritmos clássicos de agrupamento, por exemplo, K-médias, mapas auto-organizáveis e gás neural. Os autores também mostraram os métodos de agrupamento difusos utilizando núcleo como extensões do algoritmo de agrupamento rígido *kernel k-means*.

2.2.2.4 Métodos Probabilísticos

Outro tipo de método de agrupamento é o método probabilístico. Neste tipo de agrupamento, o método pode ser visto como um procedimento para identificar regiões densas no conjunto de dados usando uma função de probabilidade. Um popular algoritmo utilizado como técnica de agrupamento probabilístico é o algoritmo *Expectation-Maximization (EM)* ((DEMPSTER A. P.; RUBIN, 1977)) que é usado na análise estatística para estimação de parâmetros por máxima verossimilhança na presença de dados incompletos ((AITKIN; WILSON, 1980)).

O *EM* tem a vantagem de ser um algoritmo simples, estável e robusto para dados ruidosos e, ainda, admite dados contínuos e categóricos com cada indivíduo podendo pertencer a diferentes grupos com diferentes probabilidades. O *EM* é considerado uma versão probabilística do método K-médias, pois ambos admitem a existência de k grupos e tentam encontrar os agrupamentos baseados em protótipos. Usando o método de misturas de distribuições (*EM* de Misturas Gaussianas (EM-MMG)) ((BILMES, 1998)) tem-se uma representação de uma função de probabilidades consistindo de diversos componentes, cada componente gerando um grupo. Dessa forma, o conjunto de dados passa a ser uma mistura

de grupos e o problema consiste em identificar os elementos que constituem cada grupo e inferir sobre os parâmetros da distribuição de probabilidade assumida para cada um deles.

Em um trabalho mais recente, (YU Z.; YU, 2014) criaram uma nova estrutura de conjunto de grupo probabilístico, denominada *Gaussian Mixture Model based cluster Structure Ensemble(GMMSE)*, para identificar a estrutura de grupo mais representativa do conjunto de dados. O *GMMSE* aplica a abordagem de ensacamento para produzir um conjunto de dados variante. Então, um conjunto de modelos de mistura gaussiana é usado para capturar as estruturas de grupos subjacentes dos conjuntos de dados. *GMMSE* aplica K-médias para inicializar os valores dos parâmetros do modelo de mistura gaussiana e adota a abordagem de (EM) para estimar os valores dos parâmetros do modelo.

2.2.2.5 Distâncias Adaptativas

As medidas de distância utilizadas pelos métodos de análise de agrupamento podem deixar de ter apenas a característica fixa e assumir um caráter adaptativo ((DIDAY; GOVAERT, 1977); (DIDAY; SIMON, 1976)). A versão adaptativa desses algoritmos é capaz de associar uma distância diferente para cada classe a cada nova iteração, aumentando a qualidade da partição. A principal diferença entre os métodos tradicionais e os algoritmos com distância adaptativa é que, estes, usam pesos para cada variável de forma que o cálculo final da distância é balanceada pela dispersão de cada variável.

Os vetores de pesos das distâncias adaptativas ainda podem ter duas abordagens: por atributo ou único ou por classe e por atributo. A principal ideia da abordagem de distâncias adaptativas por atributo ou única é que há uma distância para comparar objetos e seus protótipos que muda a cada iteração, mas que é a mesma para todos os grupos. E na abordagem de distâncias adaptativas por classe e por atributo, há uma distância para comparar grupos e os seus respectivos protótipos que muda a cada iteração, como também varia de um atributo de uma classe para o mesmo atributo em outra classe. Desta forma, em geral, os protótipos podem melhor representar os grupos e a partição final encontrada é melhor que a do método que usa distância não-adaptativa. A vantagem, sobre os métodos de distância fixa, desse tipo de distância adaptativa é que o algoritmo de agrupamento torna-se capaz de achar grupos de diferentes formas e tamanhos.

Dentre alguns trabalhos que envolvem métodos de agrupamento com distâncias adaptativas, (CARVALHO F. A. T.; JUNIOR, 2006) apresentaram métodos de agrupamentos difusos baseados em diferentes distâncias adaptativas quadráticas definidas por matrizes de covariância difusas. (CARVALHO; SOUZA, 2010) introduzem métodos de agrupamentos dinâmicos para dados simbólicos com características mistas baseados em uma adequada distância Euclidiana adaptativa quadrática com pré-processamento dos dados de entrada transformando-os em dados do tipo histograma. Utilizando distâncias adaptativas com núcleo, (FERREIRA; CARVALHO, 2014) propuseram métodos de agrupamento *kernel c-*

médias difusos variável em que as medidas de dissimilaridade são obtidas como somas de distâncias Euclidianas entre objetos e centróides computados individualmente para cada variável por meio de funções de núcleo. (HAJJAR; HAMDAN, 2013) apresentam um mapa auto-organizado, uma espécie de rede neural artificial utilizada para mapear dados de alta dimensão em um espaço baixa dimensão, para dados de valor de intervalo com base em distâncias adaptativas de *Mahalanobis* para fazer o agrupamento de dados de intervalo com preservação da topologia.

2.2.3 Métodos Baseados em Busca

A solução para um problema pode ser definido como uma sequência fixa de ações e um objetivo pode ser considerado como um conjunto de estados do mundo. Para resolver um problema e atingir um objetivo específico deve-se decidir que tipos de ações e estados devem ser considerados. O processo de procurar por uma sequência de ações que alcançam um determinado objetivo para resolver um problema é chamado de busca. Um algoritmo de busca recebe um problema como entrada e devolve uma solução sob a forma de uma sequência de ações. Depois que uma solução é encontrada, as ações recomendadas podem ser executadas ((RUSSELL; NORVIG, 2014)).

Para solucionar um dado problema encontrando uma sequência de ações, os algoritmos de busca consideram várias sequências de ações possíveis. As sequências de ações possíveis que começam a partir do estado inicial formam uma árvore de busca com o estado inicial fixado na raiz; os ramos são as ações, e os nós correspondem aos estados no espaço de estados do problema. Essa é a essência da busca que segue, escolhe uma opção e deixa as outras escolhas reservadas para depois, no caso da primeira escolha não ser a melhor opção para a solução do problema.

Os algoritmos que esquecem os caminhos que já percorreram em busca de uma solução estão mais aptos a repeti-los. A maneira de evitar a exploração de caminhos redundantes é trabalhar com uma estrutura de dados chamada busca em grafo que lembra de todo o nó que foi expandido. Nós que foram gerados recentemente e que já fizeram parte de caminhos anteriores podem ser descartados. Os algoritmos que não se preocupam com o caminho percorrido, até o objetivo a ser alcançado, formam uma classe diferente de algoritmos: os algoritmos de busca local. Como esse tipo de busca usa apenas um único estado atual e se move apenas para os vizinhos desse estado, ela possibilita a vantagem de usar pouca memória e pode encontrar soluções razoáveis em grandes ou infinitos (contínuos) espaços de estados.

Os algoritmos de busca local exploram uma topologia de espaço de estados (Figura 8). Essa topologia tem, ao mesmo tempo, uma posição que é definida pelos estados e uma elevação que pode ser definida por uma função objetivo. Se a elevação determinar um custo, o objetivo será encontrar o vale mais baixo, um mínimo global; se a elevação corresponder a uma função objetivo, então o objetivo será encontrar o pico mais alto, um

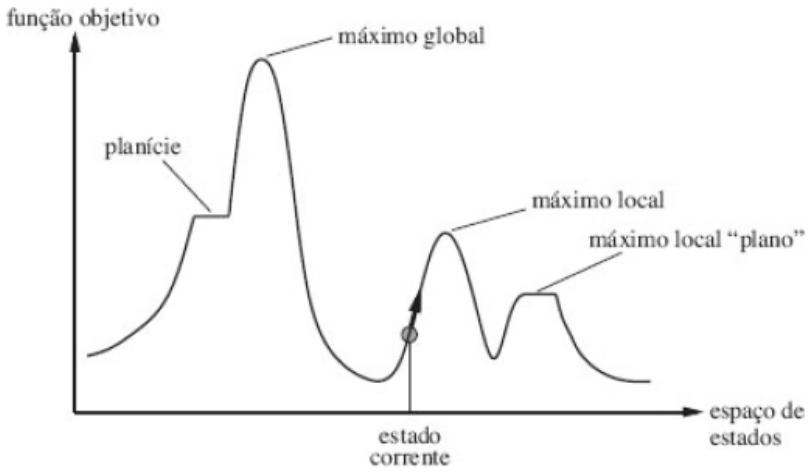
máximo global.

Os algoritmos de busca local também são úteis para resolver problemas de otimização, nos quais o objetivo é encontrar o melhor estado de acordo com uma função objetivo. Um algoritmo de busca local completo sempre encontra um objetivo, caso ele exista, e um algoritmo ótimo sempre encontra um mínimo/máximo global.

2.2.3.1 Algoritmo Subida da Encosta

O algoritmo de busca da subida de encosta é um laço repetitivo que se move de forma contínua no sentido do valor crescente, isto é, encosta acima. Esse algoritmo de busca é a técnica de busca local mais básica. Em cada passo, o nó atual é substituído pelo melhor vizinho. O algoritmo termina quando alcança um valor em que nenhum vizinho tem valor mais alto que ele ou valor mais baixo, se for usada uma estimativa de custo. E o algoritmo também finaliza sua execução se atingir o número de iterações definido anteriormente. Como o algoritmo não mantém uma árvore de busca, a estrutura de dados do nó atual só precisa registrar o estado e o valor de sua função objetivo. A busca subida da encosta só examina valores de estados dos vizinhos imediatos do estado atual ((RUSSELL; NORVIG, 2014)).

Figura 8 – Topologia Algoritmo Subida da Encosta



Fonte: (RUSSELL; NORVIG, 2014).

O algoritmo Subida de Encosta modifica o estado atual para tentar melhorá-lo, como mostra a seta da Figura 8.

Teoricamente, para se ter uma partição ideal para o agrupamento seria necessário calcular o valor do critério para cada partição possível e escolher uma partição que dá um valor ideal para esse critério. Na prática, no entanto, a tarefa não é tão fácil de ser executada, pois o número de cálculos é tão grande que a enumeração completa de cada partição se torna impraticável, mas é possível de ser realizada a depender dos recursos computacionais disponíveis. A fim de não precisar enumerar todas as partições para encontrar a

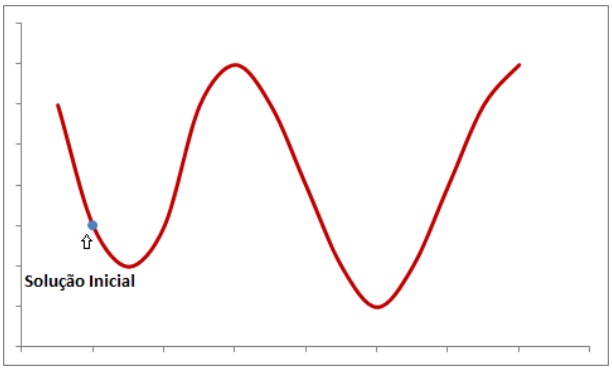
que otimizasse o critério de agrupamento (FRIEDMAN; RUBIN, 1967) desenvolveram algoritmos para procurar o valor ótimo de um critério de agrupamento, reordenando as partições existentes e mantendo a nova partição apenas se proporcionar uma melhoria, são os chamados algoritmos Subida da Encosta (*Hill-Climbing*).

2.2.3.2 Algoritmo Busca Tabu

O método Busca Tabu (*Tabu Search*) é uma estratégia que explora o espaço de soluções de um problema de otimização combinatória envolvendo aplicações que podem variar entre teoria de grafos para problemas de programação inteira ((GLOVER, 1989)). Ele é um mecanismo de memória adaptativa auxiliar que evita que o método volte a um ótimo local visitado anteriormente para superar a otimalidade local e atingir um resultado ótimo ou próximo ao ótimo global. Durante o processo de busca, dependendo do conteúdo da memória adaptativa, determinados movimentos são definidos como proibidos (ou tabus). Esses movimentos tabus forçam a busca a passar por regiões do espaço de soluções ainda inexploradas, orientando a busca em direção à solução ótima do problema. Portanto, na Busca Tabu, a utilização de memória adaptativa permite a implementação de procedimentos capazes de explorar o espaço de soluções de forma econômica e efetiva, realizando escolhas locais guiadas pela informação coletada durante o próprio processo de busca.

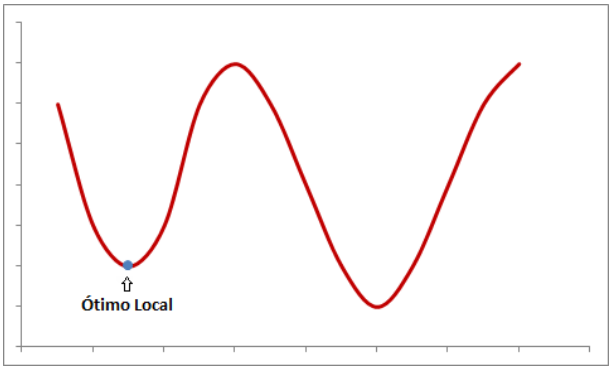
A busca tabu começa com a determinação de uma solução inicial entre as possíveis opções de inicialização existentes na literatura (Figura 9), move-se, a cada iteração, em busca das melhores soluções no espaço de dados (Figura 10) sem permitir que movimentos a levem a uma soluções já alcançadas (chamadas de tabu) armazenando-as em uma lista tabu (Figura 11). Essa lista permanece guardando as soluções já visitadas durante um determinado número de iterações (iterações tabu) e determinado espaço de tempo (prazo tabu). Ao final da busca é esperado que o método chegue ao melhor resultado (ótimo global) (Figura 12) ou que alcance o ponto mais próximo dele.

Figura 9 – Busca Tabu - Solução Inicial



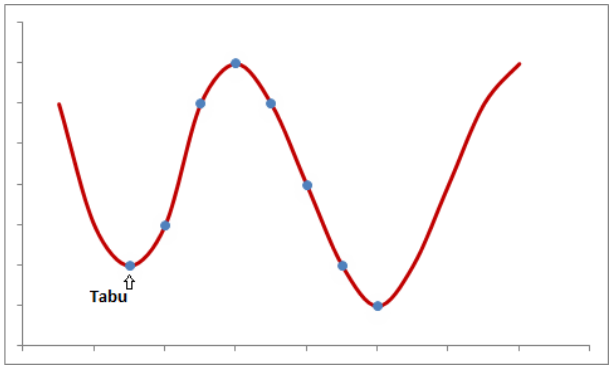
Fonte: Própria.

Figura 10 – Busca Tabu - Ótimo Local



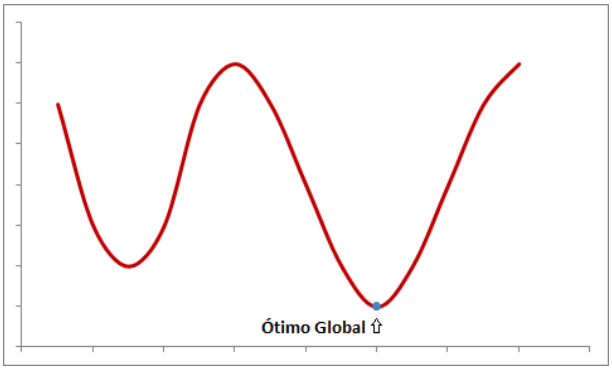
Fonte: Própria.

Figura 11 – Busca Tabu - Elementos Tabu



Fonte: Própria.

Figura 12 – Busca Tabu - Ótimo Global



Fonte: Própria.

A Busca Tabu tem também como característica importante de seu método, o fato de a solução final ter pouca ou nenhuma dependência da solução inicial. Essa característica tem fundamento na forma como a busca é feita no espaço de dados onde não é permitida a volta aos ótimos locais.

Nos últimos anos, novos algoritmos de busca local foram apresentados na esperança de

escapar de ótimos locais, como algoritmos *Simulated annealing* ((KIRKPATRICK; GELATT; ANDVECCHI, 1983); (KLEIN; DUBES, 1989)) e algoritmos Genéticos ((MAULIK; BANDYOPADHYAY, 2000)) têm sido sugeridos para tentar encontrar o ótimo global.

2.3 MÉTODOS DE REAMOSTRAGEM

Os métodos de reamostragem são ferramentas indispensáveis na estatística moderna. Eles funcionam formando repetidamente novas amostras de um conjunto de treinamento e, assim, montam modelos de interesse em cada amostra para obter informações adicionais sobre o modelo. As abordagens de reamostragem podem ser computacionalmente custosas porque envolvem várias repetições do mesmo método estatístico sobre os conjuntos de treinamento, mas com o avanço do poder computacional, tais métodos podem ser usados sem grandes preocupações com o custo computacional ((JAMES G.; TIBSHIRANI, 2017)).

Dentre várias técnicas, o *bootstrap* é uma das que fazem parte da classe de estatísticas não paramétricas que são comumente chamadas de métodos de reamostragem. O *bootstrap* é usado em vários contextos e é mais utilizado como técnica para fornecer uma medida de precisão de uma estimativa de parâmetro ou de um determinado método de seleção de aprendizagem estatística.

Com a publicação do seu artigo nos Anais de Estatística em 1979, o autor Brad Efron ((EFRON, 1979)) criou o método *bootstrap* para ser uma aproximação simples ao método de reamostragem *jackknife* ((QUENOUILLE, 1949); (TUKEY, 1958)). A motivação original do autor foi derivar propriedades do *bootstrap* para entender melhor o *jackknife*, no entanto, em muitas situações, o bootstrap é tão bom quanto ou melhor como um procedimento de reamostragem. Em pequenas amostras, o *jackknife* se mostrava mais útil, mas para amostras com grande quantidade de dados tornava-se computacionalmente ineficiente ((CHERNICK; LABUDDE, 2011)).

Embora seja intuitivo que as decisões tomadas em conjunto são, muitas vezes, melhores que as decisões individuais, ainda são recentes os investimentos nesse tipo de ideia aplicada à área de Aprendizagem de Máquina. Os métodos *ensemble* são técnicas utilizadas para se trabalhar em conjunto. Esses métodos empregam múltiplos modelos e combinam suas previsões para resolver um determinado problema em conjunto, com essa abordagem eles podem obter melhores desempenhos do que qualquer um dos modelos agindo isoladamente. Os procedimentos de *Bagging* e *Boosting* são exemplos de métodos *ensemble* que combinaram informações para produzir um algoritmo mais eficiente ((CHERNICK; LABUDDE, 2011)). O *boosting* é considerado mais forte do que o *Bagging* em conjunto de dados sem ruído com estruturas de classes complexas, enquanto o *Bagging* é mais robusto do que o *boosting* nos casos em que os dados de ruído estão presentes.

Segundo (BREIMAN, 1996), *Bagging*, um nome derivado de "agregação de *bootstrap*", foi o primeiro método eficaz de geração de novos conjuntos de aprendizagem e é um dos métodos mais simples de selecionar dados. É um método de aprendizagem para melho-

rar os modelos em termos de estabilidade e precisão, reduzir a variabilidade e ajudar a evitar a sobreposição. Este algoritmo foi originalmente concebido para a classificação e é normalmente aplicado aos modelos de árvore de decisão, mas pode ser utilizado com qualquer tipo de modelo de classificação ou regressão.

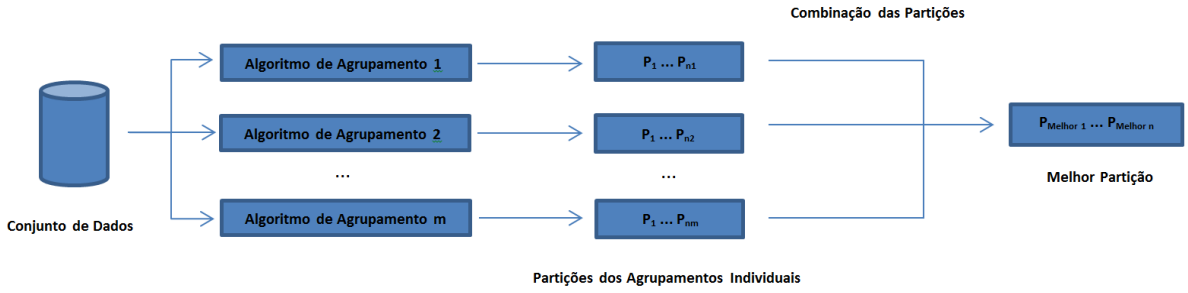
O método *Bagging* funciona utilizando várias versões de um conjunto de treinamento utilizando a amostragem *bootstrap*, ou seja, com reposição, que são usadas como novos conjuntos de aprendizagem. Cada um destes conjuntos de dados é usado para treinar um modelo diferente. As saídas dos modelos são combinados pela média (em caso de regressão) ou votação (em caso de classificação) para criar uma única saída. Para (BREIMAN, 1996), em métodos de regressão ou classificação, era possível melhorar as regras de predição através da redução da variabilidade da estimativa utilizando o método *Bagging*.

2.3.1 Bagging para Análise de Agrupamento

A análise de agrupamento pode construir métodos mais eficientes quando se beneficia da combinação dos pontos fortes de outros algoritmos usados individualmente. Dentre as várias abordagens existentes para melhorar o desempenho da análise de agrupamento, está o método *clustering ensembles*. O foco da pesquisa em *clustering ensembles* é tentar combinar múltiplas partições que ofereçam um agrupamento melhorado dos conjuntos de dados que possa atingir um resultado global. Os métodos de *clustering ensembles* podem ter melhor performance que os algoritmos de agrupamento individuais em aspectos como robustez por alcançar melhor desempenho médio, inovação ao encontrar uma solução combinada inalcançável por outro algoritmo de agrupamento e estabilidade em soluções de agrupamento que contenham menor sensibilidade ao ruído, *outliers* ou variações de amostragem ((GHAEMI R.; MUSTAPHA, 2009)).

Os *clustering ensembles* geralmente são algoritmos de dois estágios. No primeiro estágio, são armazenados os resultados independentes gerados por algum algoritmo de agrupamento, como o K-médias. Em seguida, é usada uma função de consenso específica para encontrar uma partição final a partir de resultados armazenados. A Figura 13 mostra uma arquitetura do método *clustering ensembles*.

Figura 13 – Arquitetura *Clustering Ensembles*.



Fonte: Própria.

Dado vários agrupamentos do conjunto de dados, o método *clustering ensembles* tem a função de encontrar uma combinação de agrupamento com melhor qualidade do que os algoritmos individuais e, assim, encontrar uma partição que seja a melhor entre as partições de saída dos vários algoritmos de agrupamento. Para encontrar a melhor combinação entre as partições extraídas dos algoritmos de agrupamento são utilizadas funções de consenso.

Diversos métodos são conhecidos para a geração de partições para os *clustering ensembles*, dentre eles: usar diferentes algoritmos de agrupamento, alterar a inicialização ou outros parâmetros do algoritmo ou particionar diferentes subconjuntos dos dados originais. Algumas funções de consenso foram desenvolvidas para *clustering ensembles*, tais como, particionamento de hipergrafia, abordagem de votação, algoritmo de informação mútua, funções baseadas em co-associação e modelo de mistura finita ((GHAEMI R.; MUSTAPHA, 2009)).

Nesta pesquisa, a técnica de *clustering ensembles* foi formada pelo método *Bagging* e os algoritmos de agrupamento propostos. O *Bagging* foi aplicado para melhorar o agrupamento dos dados da área de Análise estatística de formas particionando diferentes subconjuntos dos dados originais e usando a abordagem de votação como função de consenso para encontrar a melhor partição gerada pelos algoritmos de agrupamento.

O método *Bagging* tem sido aplicado para melhorar o desempenho dos algoritmos de agrupamento, incorporando à estrutura da análise de agrupamento novos conjuntos de treinamento formados por amostras *bootstrap*. (LEISCH, 1999) introduziu um novo método para análise de agrupamento utilizando *Bagging*. Este novo método, chamado de *bagged clustering*, combinava métodos hierárquicos e métodos de particionamento: com um estágio de pré-processamento para redução da complexidade do agrupamento hierárquico e com procedimento de combinação para vários resultados do método de particionamento.

A ideia central de Leisch era estabilizar métodos de particionamento como K-médias executando repetidamente o algoritmo de agrupamento e combinando os resultados. No algoritmo K-médias, as inicializações e as pequenas mudanças no conjunto de treinamento podem ter grande influência sobre o mínimo local atingido pelo algoritmo. Assim, os efeitos aleatórios do conjunto de treinamento eram reduzidos pela utilização de repetições com reamostragem para formar novos conjuntos de dados.

(DUDOIT; FRIDLYAND, 2003) usaram o método *Bagging* para melhorar a classificação com os algoritmos de particionamento baseados em *medoids* (PAM) ((KAUFMAN; ROUSSEEUW, 1987)) que tem dois principais argumentos: a matriz de dissimilaridade e o número de grupos k. Os autores propuseram dois tipos de *Bagging* para métodos de agrupamento denotados por *BagClust1* e *BagClust2* para usar com o algoritmo PAM. No *BagClust1* o agrupamento é repetidamente aplicado para cada amostra *bootstrap* e a partição final é obtida levando em consideração o rótulo com maior número de votos para cada observação. No abordagem *BagClust2*, é formada uma nova matriz de dissimilaridade que é usada

na entrada para realizar o agrupamento. Tal matriz registra para cada par de observações a proporção de tempo em que foram agrupadas juntas em grupos *bootstrap*.

(HAGEMAN J. A.; SMILDE, 2007) mostraram que a utilização de *Bagging* para agrupamento de dados metabolômicos (dados para estudo de todos os metabólitos - metaboloma - de uma célula, tecido ou organismo. o objetivo da metabolômica é identificar e quantificar o meteboloma de um sitema biológico) usando o algoritmo K-médias era mais eficiente do que usar o mesmo K-médias isoladamente. O agrupamento de metabólitos é uma ferramenta frequentemente usada para analisar os dados da metabolômica. Foram utilizados dados gerados por computador com o modelo de glóbulos vermelhos humanos. Os dados gerados eram dados metabolômicos ruidosos e a aplicação do método *Bagging* ajudou a melhorar a precisão do agrupamento dos dados.

Diferentemente do que usualmente é feito para os métodos de *clustering ensembles*, (JIA J.; JIAO, 2011) propuseram combinar apenas alguns resultados disponíveis dos agrupamentos visando conseguir melhores resultados. Os agrupamentos dos componentes do sistema *ensemble* foram gerados por agrupamento espectral (SC), que são métodos baseados na teoria dos grafos, e o método *Bagging* é usado para avaliar o agrupamento dos componentes. Depois uma parte dos agrupamentos é selecionada aleatoriamente para obter um resultado de consenso e são calculados a informação mútua normalizada (*normalized mutual information* (NMI)) ou o índice de Rand ajustado (*adjusted Rand index (ARI)*) entre os resultados. Os dados foram classificados pela agregação de múltiplos valores de NMI ou ARI. Os resultados experimentais mostraram que o algoritmo proposto pelos autores pode obter um resultado melhor do que os métodos tradicionais de *clustering ensembles*.

Para combinar os pontos fortes dos métodos *ensembles Boosting* e *Bagging*, (YANG; JIANG, 2016) propuseram um novo tipo de reamostragem híbrida para gerar iterativamente as partições iniciais do algoritmo de agrupamento que usa tanto o *boosting* quanto o *Bagging* nessa geração, aumentando significativamente a confiabilidade dessa inicialização. Além disso, eles também propuseram uma nova função de consenso para combinar partições de entrada em uma partição de consenso robusta considerando a informação estrutural global e local das partições iniciais. Um único algoritmo de agrupamento é aplicado à esses dados para obter a partição de consenso consolidada. Para avaliar a abordagem foram usados dados simulados 2D, coleção de *benchmarks* e conjuntos de dados de reconhecimento facial do mundo real, e os resultados evidenciaram que a abordagem supera os *benchmarks* existentes.

3 ALGORITMOS DE AGRUPAMENTOS PARA FORMAS PLANAS

3.1 ADAPTAÇÃO DOS ALGORITMOS

Nesta pesquisa foram propostos algoritmos de agrupamento adaptados para particularidade dos dados da área de análise estatística de formas de objetos. Tais algoritmos foram escolhidos por serem mais comuns na literatura da análise de agrupamentos. Também foram utilizados três critérios de agrupamento que anteriormente já eram utilizados como estatísticas de testes para o contexto de formas de objetos, e aqui foram usados no contexto de análise agrupamentos para validar os grupos. Além destes, um critério baseado na soma dos erros quadráticos, adaptado para os dados de formas de objetos, foi utilizado.

O algoritmo de agrupamento *KI*-médias ((BANFIELD; BASSIL, 1977)) e os algoritmos de agrupamento baseados em busca Subida da Encosta ((FRIEDMAN; RUBIN, 1967)) e Tabu Search ((GLOVER, 1989)) ganharam novas versões utilizando uma nova medida de distância e novos critérios de agrupamentos para dados de formas planas de objetos. Além disso, os três algoritmos anteriores e o algoritmo K-médias tiveram a formação de seus agrupamentos analisada com base na introdução ao contexto do método reamostragem *Bagging* (BREIMAN, 1996) formando um *clustering ensemble* para os dados. A medida de distância utilizada para atribuição dos objetos aos grupos foi a distância de *Procrustes* completa definida pela Equação (2.16).

De acordo (FRIEDMAN; RUBIN, 1967) e (BLASHFIELD, 1976) os resultados de um método de otimização podem ser bastante influenciados pela escolha da partição inicial. Diferentes partições iniciais podem levar a diferentes ótimos locais do critério de agrupamento. Uma partição inicial pode ser definida com base em um conhecimento prévio, ou pode ser o resultado de um outro método de agrupamento ou, ainda, uma partição inicial pode ser escolhida ao acaso. (MARRIOTT, 1982) sugere que a convergência lenta e grupos muito diferentes dado por diferentes partições iniciais geralmente indicam que os protótipos foram erroneamente escolhidos, em particular, que não há evidência clara de agrupamento.

Para tentar escapar da influência da partição inicial, protótipos aleatórios foram gerados para formar as partições iniciais aqui utilizadas para os algoritmos propostos. Inicialmente foram escolhidos, aleatoriamente, centróides para formar as partições e, em seguida, cada objeto foi atribuído ao centróide mais proximo. Após terem sido feitas as alocações, foi avaliado o critério de agrupamento para cada partição gerada. Esse processo foi repetido 100 vezes e a partição que apresentou o melhor resultado para os critérios de agrupamento foi escolhida como partição inicial dos algoritmo propostos.

Inicialmente, considere $\Omega = \{1, \ldots, i, \ldots, n\}$ um conjunto de n objetos indexados por i com o correspondente conjunto de índices $C = \{1, \ldots, n\}$. Cada objeto i é representado

por uma matriz de configuração do conjunto $\mathbf{Y}^* = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ definida em pela Equação (2.2). Seja $\mathbf{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_K\}$ o conjunto de K protótipos relacionados a K grupos de uma partição $\Pi_K = \{S_1, \ldots, S_K\}$. Cada conjunto $S_k (1 \leq k \leq K)$ contém os objetos atribuídos ao cluster k da partição, Π_K. Os algoritmos particionais baseados em protótipos buscam uma partição Π_K de Ω em K grupos e um conjunto correspondente de protótipos $\mathbf{G}$ que minimiza ou maximiza localmente um critério de adequação. Já os algoritmos particionais baseados em busca encontram uma partição Π_K de Ω obtendo uma sequência de ações que minimiza ou maximiza localmente um critério de adequação.

3.2 CRITÉRIOS DE AGRUPAMENTO

Como em análise estatística de formas os critérios de agrupamento precisam ser substituídos ou adaptados para atender as características dos dados, esta pesquisa apresenta novas opções para estes critérios. Para encontrar a melhor solução, dentre todas as possíveis partições geradas pelos algoritmos de agrupamento, são utilizados quatro critérios de agrupamento para trabalhar com formas planas de objetos. Dentre estes critérios, três são frequentemente usados como estatísticas de teste de hipóteses e foram adaptados para a análise de agrupamentos. Foram consideradas nesta pesquisa as estatísticas de teste *Goodall* ((GOODALL, 1991)), *Hotelling* ((JOHNSON; WICHERN, 2008)) e *James* ((JAMES, 1954)). Estas estatísticas são projetadas para dados altamente concentrados e são motivadas por aproximações adequadas para a esfera de pré-formas ou para dados do espaço tangente.

Os critérios baseados nas estatísticas de testes *Hotelling* e *James* atuam no espaço das coordenadas tangentes fazendo aproximação dos dados na esfera. O critério baseado na estatística de teste *Goodall* está situada no espaço das pré-formas. O último critério utilizado é baseado na soma dos erros quadráticos ((XU; WUNSCH, 2005)), que é uma das funções mais amplamente utilizadas na literatura de análise de agrupamento como critério para formação dos grupos. Tal critério, nesta pesquisa, também atua no espaço das pré-formas depois de ter sido adaptado para tal espaço. As partições geradas pelos algoritmos devem ter como objetivo maximizar os critérios *Goodall*, *Hotelling* e *James* e minimizar o critério baseado em soma de quadrados.

3.2.1 Critérios de Agrupamento para o Espaço de Pré-formas

Seja $\mathbf{z}_1, \ldots, \mathbf{z}_n$ uma amostra aleatória de pré-formas e μ a forma média que representa os protótipos do conjunto G para essas pré-formas. Os algoritmos de agrupamento propostos para o espaço pré-forma buscam uma partição e um conjunto correspondente de protótipos que maximizam um critério baseado na estatística de teste proposta por (GOODALL, 1991) ou minimizam um critério baseado em somas dentro dos grupos.

3.2.1.1 Critério Baseado na Estatística de *Goodall*

Seja $\hat{\mu}_j$ a média de Procrustes completa e $\hat{\mu}$ é a forma média de Procrustes completa agrupada. A estatística proposta por (GOODALL, 1991) assume simetria rotacional e igual estrutura de dispersão entre populações ((DRYDEN; MARDIA, 1998)). Um critério adequado baseado nesta estatística é dado por

$$J^1(\Pi_K, \mathbf{G}) = n(n-1)K \frac{\sum_{j=1}^{K} d_F^{(2)}(\hat{\mu}_j, \hat{\mu})}{(K-1)\sum_{j=1}^{K}\sum_{i\in C_j} d_F^{(2)}(\mathbf{z}_{ji}, \hat{\mu}_j)}, \tag{3.1}$$

onde $d_F^{(2)}$ é a distância Procrustes completa entre pré-formas para $m = 2$ definido pela Equação (2.16).

3.2.1.2 Critério Baseado em Soma de Quadrados

Na literatura de agrupamento, é usual considerar um critério baseado em somas de quadrados dentro de um cluster. No entanto, não é adequado para o espaço pré-formas não-Euclidiano. Assim, foi gerada uma medida - baseada na função usual de soma de quadrados - adequada para atender as características dos dados neste tipo de espaço, definida como ((OLIVEIRA, 2016))

$$J^2(\Pi_K, \mathbf{G}) = \sum_{j=1}^{K}\sum_{i\in C_j} D\left(\hat{\mu}_j, \mathbf{z}_i\right), \tag{3.2}$$

onde a função D é uma das medidas de distância definidas pelas Equações (2.16), (2.17) ou (2.18).

3.2.2 Critérios de Agrupamento para o Espaço Tangente

Seja $\mathbf{v}_1, \ldots, \mathbf{v}_n$ vetores de coordenadas tangentes reais definidas como na Equação (2.23). Os algoritmos de agrupamento propostos para o espaço tangente buscam uma partição e um conjunto correspondente de protótipos que maximizam os critérios com base na estatística de teste *Hotelling* definida por (JOHNSON; WICHERN, 2008) e estatística de teste *James* proposta por (JAMES, 1954).

3.2.2.1 Critério Baseado na Estatística de *Hotelling*

A estatística T^2 de *Hotelling* permite estruturas gerais de dispersão, mas assume que estas sejam iguais entre as populações. Um critério adequado baseado na estatística *Hotelling* definida por (JOHNSON; WICHERN, 2008) é dado por

$$J^3(\Pi_K, \mathbf{G}) = \frac{|\sum_{j=1}^{K}\sum_{i\in C_j}(\mathbf{v}_{ji} - \bar{\mathbf{v}}_j)(\mathbf{v}_{ji} - \bar{\mathbf{v}}_j)'|}{|\sum_{j=1}^{K}\sum_{i\in C_j}(\mathbf{v}_{ji} - \bar{\mathbf{v}})(\mathbf{v}_{ji} - \bar{\mathbf{v}})'|}, \tag{3.3}$$

onde $\bar{\mathbf{v}}$ é o vetor global das médias das coordenadas tangentes.

3.2.2.2 Critério Baseado na Estatística de *James*

A estatística de *James* é uma versão modificada da estatística de *Hotelling* que permite diferentes estruturas de dispersão entre populações quando as covariâncias não são assumidas como sendo iguais ((SEBER, 1984)). Um critério adequado baseado na estatística definida por (JAMES, 1954) para o espaço tangente é definido como

$$J^4(\Pi_K, \mathbf{G}) = \sum_{j=1}^{K} \sum_{i \in C_j} (\mathbf{v}_i - \bar{\mathbf{v}_j})' \mathbf{\Psi}_j (\mathbf{v}_i - \bar{\mathbf{v}_j}), \tag{3.4}$$

onde as matrizes de pesos $\mathbf{\Psi}_j = \left(\frac{S_j}{n_j}\right)^{-1}$ são o inverso da matriz de variância-covariância S_j dividida pelo tamanho da amostra n_j. Se $\mathbf{\Psi}_j = \mathbf{I}$, o critério J^4 corresponde à distância Euclidiana entre $\mathbf{v}_i$ and $\mathbf{v}_j$.

3.3 ALGORITMOS BASEADOS EM PROTÓTIPOS

Esta subseção traz a descrição dos algoritmos baseados em protótipos K-médias e KI-médias adaptados para atuarem na área se análise estatística de formas de objetos. Tais métodos foram analisados tanto no espaço de pré-formas quanto no espaço tangente.

3.3.1 Algoritmo *K*-Médias

Uma descrição do algoritmo K-médias introduzido por (MACQUEEN, 1967) em sua versão adaptada para trabalhar com dados de análise estatística de formas é dada a seguir ((BRUSCO; STEINLEY, 2007); (OLIVEIRA, 2016)). A usual distância Euclidiana foi substituída pela distância Procrustes mostrada na Equação (2.16) para os dados no espaço de pré-formas. O algoritmo K-médias usa o valor médio dos objetos em um grupo como o protótipo do grupo. O cálculo da variação nos valores dos critérios de agrupamento resulta do movimento do objeto i do seu atual grupo k para um novo grupo l ($l \neq k$).

Em análise estatística de formas, o K-médias encontra uma partição verificando a medida de distância entre os objetos e a forma média dos objetos. O algoritmo de agrupamento K-médias adaptado para formas planas, no espaço de pré-formas, segue os passos

definidos no Algoritmo 3.

Algoritmo 3: K-médias para Dados de Pré-Formas

1. Dados de Entrada

- Seja $\Omega = 1, \ldots, n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}^*$; Fixe o número de iterações T e fixe o erro $\varepsilon > 0$.
- Calcule os dados de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$ de acordo com a Equação (2.10).
- Escolha um dos critérios de agrupamento definidos pelas Equações (3.1) ou (3.2).
- Escolha aleatoriamente diferentes K centróides e gere uma partição inicial π_K a partir destes protótipos.

↓

2. Atribuindo os objetos aos grupos

- Calcule a forma média μ dos grupos, onde: $\mu = \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^*$. Faça $t = 1$.
- Atribua cada objeto ao grupo em que a medida de dissimilaridade é menor em relação a forma média μ.
- Calcule o critério de agrupamento.
- Calcule a forma média μ dos novos grupos.

↓

3. Critério de Parada

- Se $||J_{t+1} - J_t|| \leq \varepsilon$ ou $t > T$, então Pare. Caso contrário, faça $t = t + 1$ e volte ao Passo 2.

1

Para o espaço tangente, o algoritmo K-médias adaptado para formas planas segue os

passos definidos no Algoritmo 4.

Algoritmo 4: K-médias para Dados no Espaço Tangente

1. Dados de Entrada

- Seja $\Omega = 1, \ldots, n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}^*$; Fixe o número de iterações T e fixe o erro $\varepsilon > 0$.
- Calcule os dados de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$ de acordo com a Equação (2.10).
- Calcule o vetor de coordenadas tangentes $\mathbf{v}_i$ utilizando a Equação (2.23)
- Escolha um dos critérios de agrupamento definidos pelas Equações (3.3) ou (3.4).
- Escolha aleatoriamente diferentes K centróides e gere uma partição inicial π_K a partir destes protótipos.

↓

2. Atribuindo os objetos aos grupos

- Calcule a forma média dos grupos $\bar{v}$.
- Atribua cada objeto ao grupo em que a medida de dissimilaridade é menor em relação a forma média $\bar{v}$.
- Calcule o critério de agrupamento.
- Calcule a média $\bar{v}$ dos objetos dos novos grupos.

↓

3. Critério de Parada

- Se $||J_{t+1} - J_t|| \leq \varepsilon$ ou $t > T$, então Pare. Caso contrário, faça $t = t + 1$ e volte ao Passo 2.

1

O algoritmos K-médias converge quando não há mais mudanças significativas no critério de agrupamento ou quando o número máximo de iterações predefinido é atingido. A solução do K-médias é dependente da partição inicial, podendo ele convergir para um ótimo local se essa partição for mal definida.

3.3.2 Algoritmo KI-Médias

O algoritmo KI-médias (KI-means) é um método de agrupamento híbrido que utiliza duas etapas iterativas para formar partições dos conjuntos de dados. O algoritmo tem duas fases distintas - uma transfere um objeto de um grupo para o outro e outro troca dois objetos entre grupos. Dada uma partição inicial, cada transferência possível é testada por sua vez, para ver se ela melhora o valor do critério. Transferências de grupos com apenas um membro não são permitidas. Quando não há mais transferências que melhorem o valor critério, cada possível troca de dois objetos entre os diferentes grupos é testada e, novamente, estes são apenas executados se um benefício é encontrado. Quando não há

mais trocas que proporcionam melhoria a fase de transferência é reintroduzida, e assim por diante até que não mais transferências ou trocas possam melhorar o valor do critério ((BANFIELD; BASSIL, 1977)).

Nesta pesquisa o algoritmo *KI*-médias para formas planas é baseado no algoritmo descrito por (BRUSCO; STEINLEY, 2007) em seu artigo de comparação de heurísticas. Na primeira etapa, este algoritmo utiliza o algoritmo K-médias para agrupar os dados e na segunda etapa é realizada troca de pares de objetos entre os grupos (esta etapa é chamada de "*I-Means*" pelos autores). Essas duas etapas são iterativamente implementadas até que nenhuma operação possa melhorar os critérios de agrupamento.

O algoritmo de agrupamento *KI*-médias para o espaço de pré-formas é apresentado no Algoritmo 5.

Algoritmo 5: Algoritmo *KI*-médias Dados de Pré-Formas

1. Dados de Entrada

- Gere as partições utilizando o algoritmo K-médias descrito no Algoritmo 3.

↓

2. Atribuindo os objetos aos grupos

- Seja $\phi = 0$.
- Para $(i < j) \in C | i \in C_k, j \in C_l$ e $l \neq k$:
  Calcule δ_{kl}^{ij}, que representa a diferença inicial e final dos critérios de agrupamento definidos pelas Equações (3.1) ou (3.2) das trocas de pares feitas movendo o objeto o_i de S_k para S_l e o objeto o_j de S_l para S_k.
- Se $\delta_{kl}^{ij} < 0$, então:
  Mova o_i de S_k para S_l, e mova o_j de S_l para S_k, e defina $\phi = 1$.
- Se $\phi = 0$, então Pare. Caso contrário, defina $\phi = 0$ e volte à Etapa 2.

↓

3. Critério de Parada

- Se objetos trocaram de grupo durante a mais recente execução da Etapa 1 ou 2, então volte para a Etapa 1; caso contrário, Pare.

1

O algoritmo de agrupamento *KI*-médias para os objetos no espaço tangente é apre-

sentado no Algoritmo 6.

Algoritmo 6: Algoritmo KI-médias para Dados no Espaço Tangente

1. Dados de Entrada

- Gere as partições utilizando o algoritmo K-médias descrito no Algoritmo 4.

2. Atribuindo os objetos aos grupos

- Seja $\phi = 0$.
- Para $(i < j) \in C | i \in C_k, j \in C_l$ e $l \neq k$:
  Calcule δ_{kl}^{ij}, que representa a diferença inicial e final dos critérios de agrupamento definidos pelas Equações (3.3) ou (3.4) das trocas de pares feitas movendo o objeto o_i de S_k para S_l e o objeto o_j de S_l para S_k.
- Se $\delta_{kl}^{ij} < 0$, então:
  Mova o_i de S_k para S_l, e mova o_j de S_l para S_k, e defina $\phi = 1$.
- Se $\phi = 0$, então Pare. Caso contrário, defina $\phi = 0$ e volte à Etapa 2.

3. Critério de Parada

- Se objetos trocaram de grupo durante a mais recente execução da Etapa 1 ou 2, então volte para a Etapa 1; caso contrário, Pare.

1

O desempenho do algoritmo KI-médias também poderá variar de acordo com o método de seleção das partições iniciais. O agrupamento ótimo encontrado pode não ser único como outros agrupamentos podem dar o mesmo valor de critério. Também, o algoritmo pode não alcançar um ótimo global, de modo que diferentes partições iniciais devem ser testadas.

3.4 ALGORITMOS BASEADOS EM BUSCA

Aqui são descritos dos algoritmos baseados em busca Subida da Encosta e Busca Tabu adaptados para atuarem na área se análise estatística de formas de objetos. Tais métodos foram analisados tanto para o espaço de pré-formas quanto para o espaço tangente.

3.4.1 Algoritmo Subida da Encosta

Além dos algoritmos baseados em protótipos como o K-médias e o KI-médias, existem também os algoritmos de particionamento baseados em busca que também otimizam um critério de agrupamento. Dentre estes métodos, o algoritmo de busca Subida da Encosta (*Hill Climbing*) (FRIEDMAN; RUBIN, 1967) procura o valor ótimo de um critério de agrupamento gerando partições a partir de melhorias iterativas.

Assim como nos demais algoritmos, o método Subida da Encosta para formas planas também não pode utilizar métodos da estatística multivariada para gerar agrupamentos. Assim, este algoritmo também foi modificado para trabalhar com esses tipos de dados. O Algoritmo 7 descreve o método Subida da Encosta para os dados no espaço de pré-formas.

Algoritmo 7: Algoritmo Subida da Encosta para Dados de Pré-Formas

1. Dados de Entrada

- Seja $\Omega = 1, \ldots, n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}^*$.
- Calcule os dados de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$ de acordo com a Equação (2.10).
- Escolha um dos critérios de agrupamento definidos pelas Equações (3.1) ou (3.2).
- Escolha aleatoriamente diferentes K centróides e gere uma partição inicial π_K a partir destes protótipos.

↓

2. Atribuindo os objetos aos grupos

- Mova cada objeto de um grupo para outro produzindo a alteração do critério de agrupamento.
- Faça a alteração que conduz a maior melhoria no valor do critério de agrupamento.

↓

3. Critério de Parada

- Repita o passo 2 até não haver mais melhorias no critério de agrupamento.

1

O Algoritmo 8 descreve o método Subida da Encosta para os dados no espaço tangente.

Algoritmo 8: Algoritmo Subida da Encosta para Dados no Espaço Tangente

1. Dados de Entrada

- Seja $\Omega = 1, \ldots, n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}^*$.
- Calcule os dados de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$ de acordo com a Equação (2.10).
- Calcule o vetor de coordenadas tangentes $\mathbf{v}_i$ utilizando a Equação (2.23)
- Escolha um dos critérios de agrupamento definidos pelas Equações (3.3) ou (3.4).
- Escolha aleatoriamente diferentes K centróides e gere uma partição inicial π_K a partir destes protótipos.

↓

2. Atribuindo os objetos aos grupos

- Mova cada objeto de um grupo para outro produzindo a alteração do critério de agrupamento.
- Faça a alteração que conduz a maior melhoria no valor do critério de agrupamento.

↓

3. Critério de Parada

- Repita o passo 2 até não haver mais melhorias no critério de agrupamento.

1

3.4.2 Algoritmo Busca Tabu

O algoritmo Busca Tabu adaptado para formas planas aqui descrito é baseado nos algoritmos descritos pelos autores (BRUSCO; STEINLEY, 2007) e (PACHECO; VALENCIA, 2003). Esses autores definiram os parâmetros para a implementação do algoritmo.

O parâmetro $tabu-tenure = K$ define o número de iterações *tabu* durante as quais um objeto não é permitido retornar para um grupo e o $max - noimprovement = 50$ guarda o número máximo de iterações que não houve nenhuma melhoria e define o critério de parada para o algoritmo. A matriz $N \times K$, $\mathbf{Y}= [y_{ik}]$ contém o registro *tabu* que é o número da iteração para a atribuição de cada objeto i a cada cluster k (para $1 \leq i \leq N$ e $1 \leq k \leq K$). As constantes $niter$ e $iter - better$ representam o atual número de iterações do algoritmo e a última iteração que produziu uma solução com um valor da função objetivo melhorado, respectivamente.

O Algoritmo 9 descreve o método Busca Tabu para os dados no espaço de pré-formas.

Algoritmo 9: Algoritmo Busca Tabu para Dados de Pré-Formas

1. Dados de Entrada

- Seja $\Omega = 1, \ldots, n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}^*$.
- Calcule os dados de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$ de acordo com a Equação (2.10).
- Escolha um dos critérios de agrupamento definidos pelas Equações (3.1) ou (3.2).
- Escolha aleatoriamente diferentes K centróides e gere uma partição inicial π_K a partir destes protótipos.
- Defina $\pi_K^* = \pi_K$, $niter = 0$, $iter - better = 0$, $max - noimprovement = 50$ e $y_{ik} =$ - $tabu - tenure$.

↓

2. Atribuindo os objetos aos grupos

- Faça $niter = niter + 1$.
- Mova cada objeto de um grupo para outro e calcule o critério de agrupamento.
- Escolha o melhor critério tal que $niter > y_{ik} + tabu - tenure$.
- Faça a atualização de π_K que conduziu a maior melhoria no valor do critério de agrupamento e defina $y_{ik} = niter$.
- Se o critério de agrupamento atual for melhor que o critério anterior, então $\pi_K^* = \pi_K$ e $iter - better = niter$.

↓

3. Critério de Parada

- Se $niter$ - $iter - better > max - noimprovement$, então Pare; caso contrário, volte para a Etapa 2.

1

O Algoritmo 10 descreve o método Busca Tabu para os dados no espaço tangente.

Algoritmo 10: Algoritmo Busca Tabu para Formas Planas

1. Dados de Entrada

- Seja $\Omega = 1, \ldots, n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}^*$.
- Calcule os dados de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$ de acordo com a Equação (2.10).
- Calcule o vetor de coordenadas tangentes $\mathbf{v}_i$ utilizando a Equação (2.23).
- Escolha um dos critérios de agrupamento definidos pelas Equações (3.3) ou (3.4).
- Escolha aleatoriamente diferentes K centróides e gere uma partição inicial π_K a partir destes protótipos.
- Defina $\pi_K^* = \pi_K$, $niter = 0$, $iter - better = 0$, $max - noimprovement = 50$ e $y_{ik} =$ - $tabu - tenure$.

↓

2. Atribuindo os objetos aos grupos

- Faça $niter = niter + 1$.
- Mova cada objeto de um grupo para outro e calcule o critério de agrupamento.
- Escolha o melhor critério tal que $niter > y_{ik} + tabu - tenure$.
- Faça a atualização de π_K que conduziu a maior melhoria no valor do critério de agrupamento e defina $y_{ik} = niter$.
- Se o critério de agrupamento atual for melhor que o critério anterior, então $\pi_K^* = \pi_K$ e $iter - better = niter$.

↓

3. Critério de Parada

- Se $niter$ - $iter - better > max - noimprovement$, então Pare; caso contrário, volte para a Etapa 2.

1

3.5 ALGORITMO *BAGGING* PARA FORMAS PLANAS

Clustering ensembles combinam diferentes partições para fornecer um agrupamento melhorado dos conjuntos de dados. Esses algoritmos geralmente tem dois estágios: primeiro são armazenados os resultados independentes gerados por algum algoritmo de agrupamento e depois é utilizada uma função de consenso específica para encontrar uma partição final, a partir dos resultados obtidos nos agrupamentos individuais.

Nesta pesquisa, a técnica de *ensembles* utilizada foi o método *Bagging* formando *clustering ensembles* com os algritmos de agrupamento propostos. O método *Bagging* foi aplicado para melhorar o agrupamento dos dados da análise estatística de formas parti-

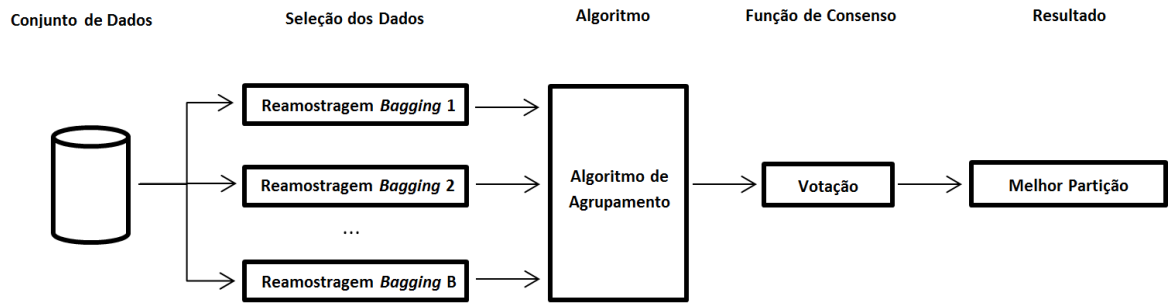
cionando diferentes subconjuntos dos dados originais e usando a abordagem de votação como função de consenso para encontrar a melhor partição gerada pelos algoritmos de agrupamento. Em caso de empate na abordagem de votação, a escolha era feita aleatoriamente. O procedimento *Bagging* utilizado, neste trabalho, foi inspirado no algoritmo *BagClust1*, introduzido por (DUDOIT; FRIDLYAND, 2003), que foi adaptado para ser utilizado com os algoritmos propostos das seções anteriores. Os conjuntos de dados são perturbados e agregados por voto majoritário para a escolha de um determinado grupo para os objetos. A perturbação do conjunto de aprendizagem pode causar mudanças significativas no preditor construído, assim, o método *Bagging* também pode levar a uma elevação da precisão dos resultados obtidos.

Os autores (DUDOIT; FRIDLYAND, 2003) usaram o método *Bagging* para melhorar a classificação com os algoritmos de particionamento baseados em medóides (PAM) ((KAUFMAN; ROUSSEEUW, 1987)) que tem dois principais argumentos: a matriz de dissimilaridade (por exemplo, a matriz de distância Euclidiana que foi usada pelos autores) e o número de grupos k. Os autores propuseram dois tipos de *Bagging* para métodos de agrupamento denotados por *BagClust1* e *BagClust2* para usar com o algoritmo PAM. No *BagClust1* o agrupamento é repetidamente aplicado para cada amostra *bootstrap* e a partição final é obtida levando em consideração o rótulo com maior número de votos para cada observação. Na abordagem *BagClust2*, é formada uma nova matriz de dissimilaridade que é usada na entrada para realizar o agrupamento. Tal matriz registra para cada par de observações a proporção de tempo em que foram agrupadas juntas em grupos *bootstrap*. A partição final é a resultante do método de agrupamento que tem esta matriz como entrada.

A motivação por trás da aplicação do método *Bagging* para análise de agrupamento é reduzir a variabilidade nos resultados de particionamentos via média ((DUDOIT; FRIDLYAND, 2003)). Neste pesquisa, o algoritmo *Bagging* gera b-ésimas amostras com repetição para selecionar os dados de entrada dos algoritmos de agrupamento. Esta amostragem é repetida 100 vezes ($B = 100$) para gerar diferentes conjuntos de rótulos para os dados de formas planas. No final das repetições é observado qual rótulo foi mais atribuído para cada objeto e o grupo que recebeu a maioria dos votos é dito vencedor. O agrupamento final é formado pelos rótulos vencedores para cada objeto. A Figura 14 mostra a sequência de passos utilizados para o agrupamento de formas planas em conjunto com o método *Bagging*.

O Algoritmo 11 mostra os passos do método *Bagging* adaptado para formas planas

Figura 14 – Etapas do Agrupamento com *Bagging*.



Fonte: Própria.

para um número fixo de grupos K no espaço de pré-formas:

Algoritmo 11: Método *Bagging* para os Algoritmos no Espaço de Pré-Formas

1. Dados de Entrada

- Seja $\Omega = 1, \ldots, n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}^*$.
- Calcule os dados de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$ de acordo com a Equação (2.10).

2. Obtendo os Rótulos

- Forme as b-ésimas amostras *bootstrap* de pré-formas $\mathcal{L}^b = (\mathbf{z}_1^b, \ldots, \mathbf{z}_n^b)$.
- Aplique o algoritmo de agrupamento $\mathcal{P}$, entre os algoritmos definidos nas Seções 3.3 e 3.4, para o conjunto de aprendizagem *bootstrap* $\mathcal{L}^b$ e obtenha os rótulos dos grupos $\mathcal{P}(\mathbf{z}_i^b; \mathcal{L}^b$ para cada observação em $\mathcal{L}^b$.

3. Função de Consenso

- Repita o passo 2 B vezes e atribua o rótulo final do grupo obtido para cada observação i por maioria dos votos.

1

O Algoritmo 12 mostra os passos do método *Bagging* adaptado para formas planas

para um número fixo de grupos K no espaço tangente:

Algoritmo 12: Método *Bagging* para os Algoritmos no Espaço Tangente

1. Dados de Entrada

- Seja $\Omega = 1, \ldots, n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}^*$.
- Calcule os dados de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$ de acordo com a Equação (2.10).
- Calcule o vetor de coordenadas tangentes $\mathbf{v}_i$ utilizando a Equação (2.23).

↓

2. Obtendo os Rótulos

- Forme as b-ésimas amostras *bootstrap* de pré-formas $\mathcal{L}^b = (\mathbf{v}_1^b, \ldots, \mathbf{v}_n^b)$.
- Aplique o algoritmo de agrupamento $\mathcal{P}$, entre os algoritmos definidos nas Seções 3.3 e 3.4, para o conjunto de aprendizagem *bootstrap* $\mathcal{L}^b$ e obtenha os rótulos dos grupos $\mathcal{P}(\mathbf{v}_i^b; \mathcal{L}^b)$ para cada observação em $\mathcal{L}^b$.

↓

3. Função de Consenso

- Repita o passo 2 B vezes e atribua o rótulo final do grupo obtido para cada observação i por maioria dos votos.

1

O método *Bagging* ajuda a melhorar os métodos de agrupamento propostos reduzindo a variabilidade dos conjuntos de dados. Apesar das vantagens, o método quando provoca uma perturbação nos dados de entrada pode mover um agrupamento que tenha atingido um ótimo global para um ótimo local dependendo da configuração gerada pela pertubação inicial. Essa é uma deficiência do método *Bagging*.

4 EXPERIMENTOS E RESULTADOS

4.1 PARÂMETROS UTILIZADOS PELOS ALGORITMOS

Para validar os métodos de agrupamento baseados em protótipos e em busca para análise estatística de formas planas, três aplicações com os conjuntos de dados reais de crânios gorilas, vértebras de ratos e peixes foram considerados. Além dos dados reais, foram utilizados conjuntos de dados simulados com diferentes concentrações, utilizando experimento Monte Carlo. Os algoritmos foram analisados em dois tipos espaços que foram o espaço de pré-formas e o espaço tangente. Nos experimentos para dados no espaço de pré-formas, a distância Procustes completa substituiu as medidas de distância comumente usadas em algoritmos de agrupamento. Os critérios de agrupamentos utilizados para avaliar a formação dos grupos foram J^1 (*Goodall*), J^2 (Soma de Quadrados), J^3 (*Hotelling*) e J^4 (*James*). Para aumentar a eficácia dos experimentos e diminuir a variabilidade dos dados, uma técnica de *clustering ensemble* foi utilizada onde os algoritmos foram implementados em conjunto com o método de reamostragem *Bagging* utilizando como função de consenso a abordagem de votação.

Para que a escolha da partição inicial não afetasse os resultados dos métodos, k centróides são obtidos aleatoriamente a partir dos dados de formas de objetos para gerar as partições iniciais dos algoritmos levando em consideração os valores dos critérios de agrupamento. Esse processo de geração da partição inicial foi repetido 100 vezes e foi escolhida a partição que melhor otimizou os critérios de agrupamento. Os métodos de agrupamento foram executados individualmente ou em conjunto com o método de reamostragem *Bagging* para obter o melhor resultado que otimizava os valores dos critérios de agrupamento selecionados para os dois espaços de dados.

Os algoritmos utilizados nos experimentos foram implementados na linguagem de programação *R* (ver http://cran.r-project.org) utilizando o ambiente de programação *RStudio* na sua versão 1.0.143 disponível em https://www.rstudio.com/. Da linguagem de programação *R*, foi utilizado o pacote *shapes.R* desenvolvido por (DRYDEN; MARDIA, 1998) para facilitar a utilização de alguns conceitos e conjuntos de dados da análise estatística de formas.

A avaliação do desempenho dos algoritmos nos conjuntos de dados foi realizada através do Índice de *Rand* Corrigido (*Correct Rand - CR*) ((HUBERT; ARABIE, 1985)). O índice *CR* é calculado para o melhor resultado obtido pela otimização dos critérios de agrupamento. O objetivo é mostrar a eficácia dos métodos baseados em protótipos e busca para diferentes conjuntos de dados na área de análise estatística formas.

4.1.1 Índice de Rand Corrigido

Em (HUBERT; ARABIE, 1985), os autores definiram índices através da comparação de duas partições. Quando duas partições são independentes, no sentido que uma depende dos valores no conjunto de dados e a outra não, a distribuição de alguns desses índices pode ser estabelecida a partir de uma perspectiva teórica.

O índice de Rand corrigido é um índice externo de avaliação que mede a similaridade entre uma partição rígida *a priori* e uma partição obtida fornecida por um algoritmo de agrupamento. O índice *CR* tem seus valores contidos no intervalo [-1,1], tal que quanto mais próximos de 1, mais semelhante a partição encontrada pelo método é à partição definida *a priori*; enquanto valores perto de 0 (ou mesmo negativos) correspondem a partições obtidas por acaso. A definição do índice *CR* é dada a seguir.

Seja $U = \{u_1, \ldots, u_i, \ldots, u_R\}$ e $V = \{v_1, \ldots, v_j, \ldots, v_C\}$ duas partições do mesmo conjunto de dados tendo respectivamente R e C grupos. O índice de Rand corrigido é dada pela Equação 4.1:

$$CR = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \tag{4.1}$$

onde $\binom{n}{2} = \frac{n(n-1)}{2}$ e n_{ij} representa o número de objetos em comum que estão nos grupos u_i e v_i; $n_{i.}$ indica o número de objetos no grupo u_i; $n_{.j}$ indica o número de objetos no grupo v_j; e n é o total número de objetos no conjunto de dados.

4.1.2 Ganho Relativo ao uso do Método *Bagging*

Uma medida de ganho relativo ao uso do *Bagging* é considerada nesse pequisa. Tal medida é usada para avaliar os valores do índice de *Rand* corrigido (*CR*), para os métodos que atuaram isoladamente em comparação com os métodos que trabalharam em conjunto com o método *Bagging*), para cada critério de agrupamento relacionado a um específico conjunto de dados. O ganho relativo ao uso do *Bagging* é dados por:

$$GR = 100\frac{CRBagging - CR}{CR}, \tag{4.2}$$

onde *CRBagging* é o valor do índice de *Rand* corrigido quando o algoritmo faz uso do método *Bagging* e o *CR* é o valor obtido quando o algoritmo atua isoladamente. Os valores de ganho relativo para as partições que obtiveram índice de *Rand* negativo não serão observados por serem consideradas partições obtidas ao acaso. O objetivo desta medida é verificar se houve um ganho ao se utilizar os algoritmos propostos em conjunto com o método de reamostragem *Bagging*.

4.2 DADOS SIMULADOS

O conjunto de dados simulados no espaço de pré-formas foi gerado com distribuição *Bingham* complexa. Foram utilizados quatro marcos anatômicos ($k = 4$) para amostras com número total de objetos igual a 100 ($n = 100$). Diferentes valores para a concentração $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ foram utilizados. Os valores para as concentrações λ foram definidos para representar altas e baixas concentrações dos dados. Os valores de $\lambda = (50, 100, 900)$ e $\lambda = (398, 499, 900)$ representam uma tendência à alta concentração nos dados e os valores $\lambda = (798, 899, 900)$ e $\lambda = (898, 899, 900)$ representam baixa concentração nos dados simulados de formas de objetos. Os 100 elementos que formaram este conjunto de dados foram divididos em dois grupos de tamanhos diferentes, onde o primeiro grupo continha 40 objetos e o segundo continha 60 objetos.

As Tabelas 3 e 4 apresentam os resultados obtidos pelos algoritmos no espaço de pré-formas e as Tabelas 5 e 6 mostram os resultados para o espaço tangente. Os valores das concentrações $\lambda = (50, 100, 900)$, $\lambda = (398, 499, 900)$, $\lambda = (798, 899, 900)$ e $\lambda = (898, 899, 900)$ foram definidos como concentrações 1, 2, 3 e 4, respectivamente. As médias dos índices de Rand corrigidos (com o respectivo desvio padrão entre parênteses) são mostrados para os algoritmos de agrupamento baseados em protótipos e busca quando estão sendo utilizados isoladamente. Foram realizados experimentos de Monte Carlo com 100 replicações para as amostras de dados simulados para os métodos K-médias, KI-médias, Subida da Encosta e Busca Tabu para os quatro diferentes critérios de agrupamento considerados.

A técnica de *clustering ensembles* também foi aplicada para os dois espaços de dados. Tal técnica, que foi aplicada aos dados simulados em conjunto com os algoritmos de agrupamento, usou o método de reamostragem *Bagging* para selecionar os dados de entrada dos algoritmos. Foram geradas 100 replicações de *bootstrap* que que foram agrupadas pelos algoritmos para formar os rótulos que passaram pelo processo de votação para gerar a melhor partição.

4.2.1 Resultado das Simulações no Espaço das Pré-formas

A Tabela 3 mostra os resultados do índice *CR*, com o respectivo desvio padrão entre parênteses, para as concentrações de dados simulados com os algoritmos baseados em protótipos.

Já a Tabela 4 exibe os índices *CR*, com o respectivo desvio padrão entre parênteses, para os algoritmos baseados em busca.

A partir dos resultados dispostos nas Tabelas 3 e 4, podemos observar que o uso do método *Bagging* combinado com os dois paradigmas de métodos de agrupamento, protótipos e busca, permitiu melhorar, significativamente, o desempenho medido pelo índice de *Rand* para todas as configurações de concentrações, métodos e critérios. Não houve diferença entre os paradigmas de métodos considerando o uso do *Bagging* e as duas

Tabela 3 – Índices *CR* para Métodos baseados em Protótipos

		Métodos			
Concentração	**Critério**	K-médias	K-médias *Bagging*	KI-médias	KI-médias *Bagging*
1	J^1	0.7023	**1.0000**	0.7023	**1.0000**
		(0.3241)	-	(0.3241)	-
	J^2	0.6376	**1.0000**	0.6569	**1.0000**
		(0.3228)	-	(0.3246)	-
2	J^1	0.6893	**1.0000**	0.6893	**1.0000**
		(0.3249)	-	(0.3249)	-
	J^2	0.6440	**1.0000**	0.6569	**1.0000**
		(0.3235)	-	(0.3246)	-
3	J^1	0.6003	0.9207	0.6583	0.9207
		(0.2654)	-	(0.2990)	-
	J^2	0.5476	0.9599	0.6291	0.9599
		(0.2911)	-	(0.3143)	-
4	J^1	0.1406	0.6357	0.2097	0.6359
		(0.1118)	-	(0.1531)	-
	J^2	0.1160	0.7022	0.1808	0.7022
		(0.1128)	-	(0.1638)	-

Fonte: Própria.

primeiras configurações de concentração. Era esperado que o uso do *Bagging* aplicado às configurações de alta concentração permitisse aos algoritmos alcançarem o ótimo global para ambos os critérios.

Observando os resultados para as demais concentrações, foi possível notar que o método *Bagging*, atuando em conjunto com os algoritmos, também proporcionou uma melhoria considerável na qualidade dos agrupamentos. Cabe ressaltar que essa melhoria de qualidade foi melhor observada pelo acréscimo alcançado pelo valor do índice de *Rand* promovido pela combinação dos métodos para a configuração 4.

Com base nos métodos baseados em protótipos, sem o uso do *Bagging*, era esperado que o método KI-médias fosse superior ao método K-médias para todas as concentrações, pois o KI-médias possui duas etapas iterativas em que a segunda etapa, que executa troca de pares de objetos, tenta melhorar o resultado obtido pelo agrupamento da primeira etapa, onde é usado o método K-médias. O critério J^1 obteve melhor desempenho, na maioria dos casos, para os métodos baseados em protótipos. Vale salientar que quando se faz uso do *Bagging* para as configurações 3 e 4, o critério J^2 se sobressai em relação ao critério J^1.

Com relação aos métodos baseados em busca executados individualmente, foi possível verificar que, no geral, os métodos obtiveram desempenhos similares. Um destaque pode ser dado ao desempenho do método Subida da Encosta que obteve melhor performance para a configuração 4 com o critério J^1. No geral, o uso do critério J^1 permitiu aos

Tabela 4 – Índices *CR* para Métodos baseados em Busca

		Métodos			
Concentração	**Critério**	Subida da Encosta	S. E. *Bagging*	Busca Tabu	B. T. *Bagging*
1	J^1	0.7346	**1.0000**	0.6764	**1.0000**
		(0.3199)	-	(0.3252)	-
	J^2	0.6828	**1.0000**	0.7217	**1.0000**
		(0.3251)	-	(0.3220)	-
2	J^1	0.7022	**1.0000**	0.6699	**1.0000**
		(0.3241)	-	(0.3251)	-
	J^2	0.6246	**1.0000**	0.6440	**1.0000**
		(0.3210)	-	(0.3235)	-
3	J^1	0.5850	0.9207	0.6272	0.9207
		(0.2795)	-	(0.2839)	-
	J^2	0.5911	0.9207	0.6164	0.9207
		(0.2957)	-	(0.3109)	-
4	J^1	0.2996	0.6361	0.2055	0.6687
		(0.1801)	-	(0.1428)	-
	J^2	0.1557	0.6686	0.1575	0.7022
		(0.1430)	-	(0.1519)	-

Fonte: Própria.

algoritmos alcançarem melhor desempenho comparado ao uso do critério J^2 para todas as configurações. Quando é observada a combinação dos métodos, o critério J^2 para a configuração 4 supera o critério J^1 para os métodos baseados em busca.

4.2.2 Resultado das Simulações no Espaço Tangente

A Tabela 5 mostra os resultados do índice *CR*, com o respectivo desvio padrão entre parênteses, para as concentrações de dados simulados com os algoritmos baseados em protótipos para o espaço tangente.

Os resultados para o espaço tangente dos métodos baseados em busca são exibidos na Tabela 6.

Nos resultados das Tabelas 5 e 6 foi percebido que o uso do método *Bagging* combinado aos algoritmos de agrupamneto é fundamental para melhorar o desempenho dos algoritmos, para todas as configurações, neste espaço de dados. Era esperado que os valores do índice de *Rand* fossem, no geral, inferiores àqueles das Tabelas 3 e 4, pois o uso do espaço tangente produz uma redução na qualidade dos agrupamentos devido a aproximação aplicada aos dados para este espaço.

Foi possível constatar que, no geral, o critério J^3 apresentou melhor desempenho para os paradigmas de métodos, critérios e concentrações. Valores de ótimo global foram alcançados apenas na configuração 2 para os métodos: *KI*-médias *Bagging* usando o critério J^4 e, Subida da Encosta *Bagging* com os critérios J^3 e J^4.

Tabela 5 – Índices *CR* para Métodos baseados em Protótipos

		Métodos			
Concentração	**Critério**	*K*-médias	*K*-médias *Bagging*	*KI*-médias	*KI*-médias *Bagging*
1	J^3	0.2047	0.9599	0.1586	0.9599
		(0.2772)	-	(0.2179)	-
	J^4	0.1748	0.9599	0.1586	0.9599
		(0.2832)	-	(0.2179)	-
2	J^3	0.2064	0.9599	0.1509	0.8447
		(0.2883)	-	(0.2099)	-
	J^4	0.1917	0.9599	0.1406	**1.0000**
		(0.3048)	-	(0.2063)	-
3	J^3	0.1840	0.7721	0.0963	0.7369
		(0.1965)	-	(0.1370)	-
	J^4	0.1713	0.7369	0.0855	0.6040
		(0.1830)	-	(0.1209)	-
4	J^3	0.0206	0.4293	0.0079	0.4847
		(0.0498)	-	(0.0261)	-
	J^4	0.0194	0.4563	0.0079	0.4032
		(0.0493)	-	(0.0261)	-

Fonte: Própria.

Sem empregar o método *Bagging*, os algoritmos baseados em protótipos obtiveram melhores performances com o critério J^3 para todas as concentrações de dados. Levando em consideração os algoritmos baseados em busca, também trabalhando isoladamente, o critério J^3 se destacou em relação ao J^4, no geral. A exceção ficou com o algoritmo Subida da Encosta sobre a configuração 4 onde o critério J^4 teve seu valor de índice de *Rand* superior ao critério J^3.

4.2.3 Ganho Relativo ao Uso do *Bagging* para os Conjuntos de Dados Simulados

A Tabela 7 mostra o ganho relativo ao uso do método *Bagging* entre os algoritmos em execuções idividuais e usando o método *Bagging* para concentração 1.

Já a Tabela 8 mostra o ganho relativo para os algoritmos sendo auxiliados pelo método *Bagging* em comparação aos resultados obtidos pelos algoritmos quando trabalharam sozinhos, para a concentração 2.

A Tabela 9 mostra o ganho relativo entre os algoritmos sem o auxílio e usando o método *Bagging* para a concentração 3.

E, por fim, a Tabela 10 mostra o ganho relativo entre os algoritmos sem o auxílio e usando o método *Bagging* para 4.

Os ganhos relativos ao uso do *Bagging*, observados nas tabelas anteriores, revelam que a técnica de *clustering ensembles*, aplicada a estes dados simulados, foi eficiente para todos os algoritmos, independente da concentração, do espaço de dados e do critério agrupa-

Tabela 6 – Índices *CR* para Métodos baseados em Busca

Concentração	**Critério**	**Métodos**			
		Subida da Encosta	S. E. *Bagging*	Busca Tabu	B. T. *Bagging*
1	J^3	0.1750	0.9599	0.1901	0.9599
		(0.1559)	-	(0.2229)	-
	J^4	0.1495	0.9599	0.1748	0.9599
		(0.1394)	-	(0.2832)	-
2	J^3	0.1654	**1.0000**	0.1694	0.9207
		(0.1492)	-	(0.1684)	-
	J^4	0.1417	**1.0000**	0.1301	0.8823
		(0.1379)	-	(0.1350)	-
3	J^3	0.0867	0.7369	0.0814	0.7369
		(0.1134)	-	(0.1223)	-
	J^4	0.0782	0.7721	0.0718	0.7025
		(0.1016)	-	(0.0898)	-
4	J^3	0.0162	0.6359	0.0135	0.4034
		(0.0410)	-	(0.0335)	-
	J^4	0.0172	0.4566	0.0093	0.3296
		(0.0377)	-	(0.0286)	-

Fonte: Própria.

Tabela 7 – Ganho Relativo (%) do índice *CR* para a Concentração 1

Critério	**Métodos**			
	K-médias x *K*-médias *Bagging*	*KI*-médias x *KI*-médias *Bagging*	Subida da Encosta x S. E. *Bagging*	Busca Tabu x B. T. *Bagging*
	Espaço de Pré-formas			
J^1	42.38	42.38	36.12	47.84
J^2	56.83	52.23	46.45	38.56
	Espaço Tangente			
J^3	368.93	505.23	448.51	404.94
J^4	449.14	505.23	542.07	449.14

Fonte: Própria.

mento utilizados. Vale observar, também, que os métodos no espaço tangente obtiveram os maiores ganhos ao trabalhar em conjunto com o método *Bagging*, o que demonstra a dependêcia que esse espaço de dados tem em relação ao método de reamostragem. A eficiência causada pelo método *Bagging* ao espaço tangente comprova o que foi mencionado em (AMARAL; DRYDEN; WOOD, 2007), que a reamostragem dos dados para formar os objetos de entrada dos algoritmos foi capaz de corrigir, significativamente, a aproximação dos dados na formação dos agrupamentos.

Apesar da relevância da atuação do método *Bagging* com os algoritmos de agrupamento, é necessário registrar que o custo computacional dessa combinação pode ser bas-

Tabela 8 – Ganho Relativo (%) do índice *CR* para a Concentração 2

Critério	**Métodos**			
	K-médias x *K*-médias *Bagging*	*KI*-médias x *KI*-médias *Bagging*	Subida da Encosta x S. E. *Bagging*	Busca Tabu x B. T. *Bagging*
	Espaço de Pré-formas			
J^1	45.07	45.07	42.40	49.27
J^2	55.27	52.23	60.10	55.27
	Espaço Tangente			
J^3	365.06	459.77	504.59	443.50
J^4	400.73	611.23	605.71	578.17

Fonte: Própria.

Tabela 9 – Ganho Relativo (%) do índice *CR* para a Concentração 3

Critério	**Métodos**			
	K-médias x *K*-médias *Bagging*	*KI*-médias x *KI*-médias *Bagging*	Subida da Encosta x S. E. *Bagging*	Busca Tabu x B. T. *Bagging*
	Espaço de Pré-formas			
J^1	53.37	39.86	57.38	46.79
J^2	75.29	52.58	55.76	49.36
	Espaço Tangente			
J^3	319.61	665.21	749.94	805.28
J^4	330.18	606.43	887.34	878.41

Fonte: Própria.

Tabela 10 – Ganho Relativo (%) do índice *CR* para a Concentração 4

Critério	**Métodos**			
	K-médias x *K*-médias *Bagging*	*KI*-médias x *KI*-médias *Bagging*	Subida da Encosta x S. E. *Bagging*	Busca Tabu x B. T. *Bagging*
	Espaço de Pré-formas			
J^1	352.13	203.24	112.31	225.40
J^2	505.34	288.38	329.41	345.84
	Espaço Tangente			
J^3	1983.98	5935.44	3825.30	2888.14
J^4	2252.06	5003.79	2554.65	3444.08

Fonte: Própria.

tante elevado. Esse custo pode ser observado no tempo execução dos algoritmos, medido em segundos, descrito na Tabela 11, onde foram analisados os tempos de execução do algoritmo baseado em protótipos *KI*-médias e do algoritmo baseado em busca Busca Tabu para a configuração 4.

Os tempos registrados na tabela 11 revelam o quão custoso é se trabalhar com o mé-

Tabela 11 – Comparação do Tempo de Execução dos Algoritimos de Agrupamento em Segundos

	Métodos			
Critério	*KI*-médias	*KI*-médias *Bagging*	Busca Tabu	Busca Tabu *Bagging*
	Espaço de Pré-formas			
J^2	65,54	5079,87	5,01	1131,52
	Espaço Tangente			
J^3	2538,07	89529,49	676,75	49738,65

Fonte: Própria.

todo *Bagging*, apesar do aumento do desempenho proporcionado ao algoritmos. Para o critério J^2, utilizar a combinação dos métodos gera um custo 77,5079 vezes maior para o *KI*-médias e 225,8522 para o Busca Tabu. Quando observado o critério J^3, a versão combinada do *KI*-médias necessita de um tempo que é 35,2746 vezes maior que a sua versão isolada. Já o Busca Tabu *Bagging* chega a precisar de um tempo 73,4963 vezes maior que o Busca Tabu puro.

Essa avaliação com os conjuntos de dados simulados abordou quatro diferentes configurações de dados simulados que foram escolhidas para evidenciar caractéristicas de altas e baixas concentrações dos dados. No geral, os critérios dos algoritmos no espaço de pré-formas foram, consideravelmente, mais eficientes que os critérios utilizados pelo espaço tangente. Estes resultados mostraram que a perda de informação provocada pela aproximação nos dados do espaço tangente afetou a qualidade dos agrupamentos.

Foi possível notar, também, que como a queda mais brusca na qualidade dos agrupamentos foi notada quando é feita a troca do espaço de pré-formas para o espaço tangente, a combinação dos algoritmos com o método de reamostragem *Bagging* é fundamental, principalmente, para os algoritmos no espaço tangente. Apesar dessa conclusão, deve-se levar em conta que o aumento do desempenho causado pela colaboração do método *Bagging*, aliado aos algoritmos de agrupamento, traz consigo o ônus do aumento do tempo para a execução dos algoritmos.

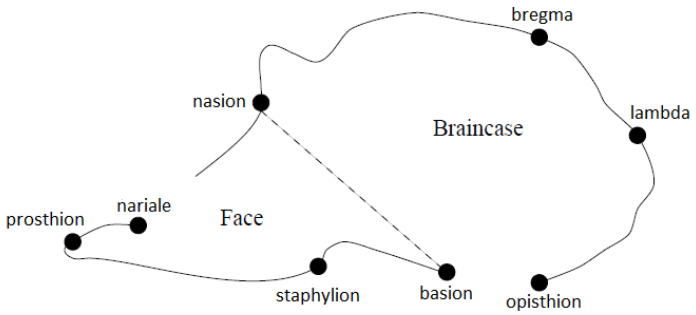
4.3 CONJUNTOS DE DADOS REAIS

Além dos conjuntos de dados simulados, três conjuntos de dados reais de formas de objetos foram utilizados nesta pesquisa. Os conjuntos de dados reais Crânio de Gorilas e Vértebras de Ratos são conjuntos clássicos da literatura da análise estatística de formas. O conjunto de Peixes foi formado pela união de duas classes da base de imagens *Core Experiment* (CE-1B).

4.3.1 Crânios de Gorilas

O conjunto de dados crânios gorilas contém 59 objetos, sendo 29 de gorilas machos e 30 de gorilas fêmeas, que foram obtidos em uma investigação para avaliar as diferenças entre os sexos cranianos de gorilas adultos ((DRYDEN; MARDIA, 1998)). Esse conjunto de dados foi produzido por O'Higgins em sua Tese de doutorado datada de 1989 (O'HIGGINS, 1989) e foi melhor detalhado por (O'HIGGINS; DRYDEN, 1993). Oito marcos anatômicos localizados por um biólogo especialista foram escolhidos para cada crânio. A Figura 15 representa oito marcos anatômicos escolhidos por um biólogo especialista para cada crânio do gorilas.

Figura 15 – Oito marcos anatômicos em um crânio de gorila.



Fonte: (DRYDEN; MARDIA, 1998).

Os dados dos marcos anatômicos dos gorilas machos e fêmeas são difíceis de analisar diretamente, porque as espécies são de tamanhos diferentes e têm sido posicionado e orientado diferentemente pelo biólogo experiente, como pode ser observado na Figura 16. Dessa forma, para analisar o agrupamento do conjunto de dados foi utilizado as pré-formas, que posteriormente passou por uma aproximação característica do espaço tangente.

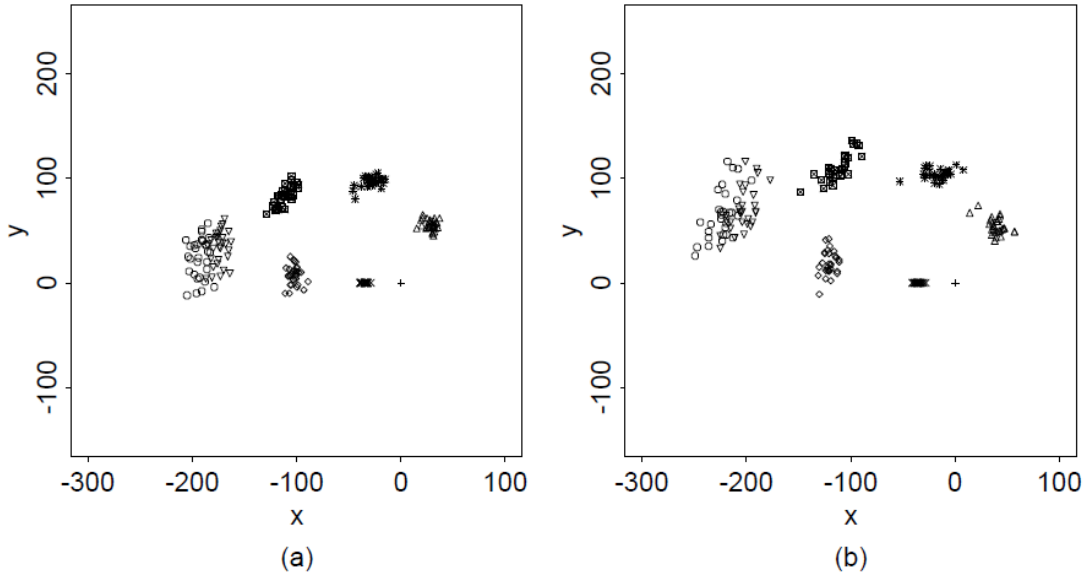
4.3.1.1 Resultados Crânios de Gorilas

A Tabela 12 apresenta uma comparação entre os resultados dos agrupamentos realizados pelos métodos K-médias, KI-médias, Subida da Encosta e Busca Tabu quando atuam individualmente e em conjunto com o método *Bagging* no espaço de pré-formas.

A Tabela 13 apresenta uma comparação entre os resultados dos agrupamentos realizados pelos métodos K-médias, KI-médias, Subida da Encosta e Busca Tabu quando atuam individualmente e combinados com o método *Bagging* no espaço das coordenadas tagentes.

Observando os resultados apresentados nas Tabelas 12 e 13, pode-se notar que os métodos baseados em busca Subida da Encosta e Busca Tabu, quando atuaram indi-

Figura 16 – (a) Os 30 marcos anatômicos dos crânios de gorilas fêmeas e (b) 29 marcos anatômicos dos crânios de gorilas machos. Cada símbolo representa cada marco anatômico definido para os crânios de gorilas.



Fonte: (DRYDEN; MARDIA, 1998).

Tabela 12 – Índices *CR* Crânios de Gorilas para o Espaço de Pré-formas

	Métodos baseados em Protótipos			
Critério	*K*-médias	*K*-médias *Bagging*	*KI*-médias	*KI*-médias *Bagging*
J^1	0.6841	0.6841	0.6841	0.6841
J^2	0.6281	0.7436	0.6281	0.7436
	Métodos baseados em Busca			
	Subida da Encosta	S. E. *Bagging*	Busca Tabu	B. T. *Bagging*
J^1	0.6841	0.5743	0.6279	0.7426
J^2	0.6841	0.7426	0.5741	0.7426

Fonte: Própria.

vidualmente, atingiram o valor de ótimo global, superando os resultados dos métodos baseados em protótipos. Esse resultado foi alcançado pelos métodos baseados em busca quando utilizaram o critério de agrupamento J^3. No geral, o critério J^3 apresentou as melhores performances para os dois espaços de dados, em relação aos demais critérios.

Analisando a combinação dos métodos, o espaço de pré-formas foi o espaço mais beneficiado pelo método de reamostragem *Bagging*. Para os métodos em que essa contribuição não foi tão efetiva, a reamostragem dos dados pode ter provocado um perturbação que prejudicou um estado de ótimo já alcançado pelos algoritmos. Essa perturbação moveu o agrupamento para um ótimo local inferior ao atingido anteriormente.

Tabela 13 – Índices *CR* para Crânios de Gorilas para o Espaço Tangente

	Métodos baseados em Protótipos			
Critério	*K*-médias	*K*-médias *Bagging*	*KI*-médias	*KI*-médias *Bagging*
J^3	0.8666	0.8034	0.9321	0.8034
J^4	-0.0108	0.5741	0.0874	0.4735
	Métodos baseados em Busca			
	Subida da Encosta	S. E. *Bagging*	Busca Tabu	B. T. *Bagging*
J^3	**1.0000**	0.7426	**1.0000**	0.0176
J^4	-0.0041	-0.0116	0.0313	-0.0177

Fonte: Própria.

4.3.1.2 Ganho Relativo ao Uso do *Bagging* para o Conjunto de Dados Crânios de Gorilas

A Tabela 14 mostra o ganho relativo entre os algoritmos sem o auxílio e auxiliados pelo método *Bagging* para este conjunto de dados.

Tabela 14 – Comparação entre Algoritimos de Agrupamento de acordo com o Ganho Relativo (%) do índice CR

	Métodos			
Critério	*K*-médias X *K*-médias *Bagging*	*KI*-médias X *KI*-médias *Bagging*	Subida da Encosta X S. E. *Bagging*	Busca Tabu X B. T. *Bagging*
	Espaço de Pré-formas			
J^1	0.00	0.00	-16.05	18.26
J^2	18.38	18.38	8.55	29.35
	Espaço Tangente			
J^3	-7.29	-13.80	-25.74	-98.24
J^4	-	441.76	-	-

Fonte: Própria.

Os ganhos apresentados na Tabela 14 evidenciaram que a atuação em conjunto dos algoritmos com o método *Bagging* proporcionou mais eficiência na qualidade dos agrupamentos, principalmente para o espaço de pré-formas. Nesse espaço de dados, a melhoria no desempenho foi observada tanto quando se fez uso do critério J^1 quanto para o critério J^2, na maioria dos casos. Analisando os resultados para o espaço tangente, o uso do *Bagging* foi apenas relevante para a combinação com o algoritmo *KI*-médias *Bagging*, utilizando o critério J^4.

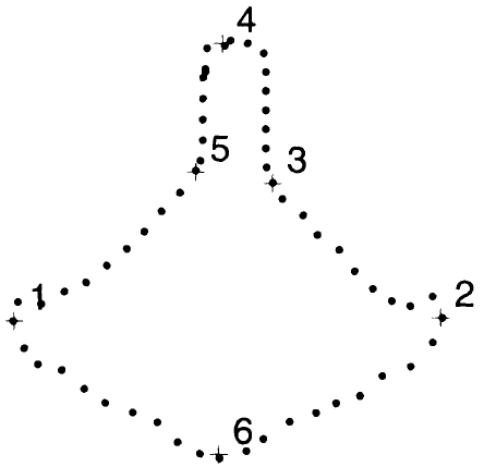
4.3.2 Vértebras de Ratos

Com o objetivo de avaliar se existe uma diferença no tamanho e na forma das vértebras de ratos, foram obtidos três grupos que caracterizavam essas vértebras: Controle com 30, Grande com 23 e Pequeno com 23 ossos ((DRYDEN; MARDIA, 1998)). O grupo Controle

contém ratos não selecionados, o grupo Grande contém ratos selecionados de cada geração de acordo ao grande peso corporal e o grupo Pequeno são selecionados pelo pequeno peso corporal. Foi considerada a segunda vértebra torácica T2 dos ratos, em que cada vértebra dos grupos de ratos foram colocadas sob um microscópico e digitalizada usando uma câmera de vídeo para dar um nível de cinza na imagem. Os dados das vértebras de ratos definidos foram obtidos em um experimento para avaliar os efeitos da seleção para peso corporal.

Seis marcos anatômicos foram fixados em cada vértebra de ratos, os quais são obtidos usando um método semi automático em pontos de alta curvatura, como descrito por (MARDIA, 1989) e por (DRYDEN, 1989). A Figura 17 ilustra uma vértebra de um rato com 6 marcos anatômicos matemáticos e 54 pseudo-marcos anatômicos.

Figura 17 – Segunda vértebra torácica *T2* de um rato.



Fonte: (DRYDEN; MARDIA, 1998).

Os marcos 1 e 2 estão em pontos máximos de função curvatura aproximada (geralmente na parte mais larga da vértebra, em vez de nas pontas), 3 e 5 estão nos pontos extremos de curvatura negativa na base da apófise espinhosa, 4 está na ponta da apófise espinhosa, 6 está no ponto de curvatura máxima no lado oposto ao marco 4. Neste trabalho, foram utilizados apenas os grupos que contemplavam as vértebras Grandes e Pequenas do ratos selecionados para avaliar os algoritmos.

4.3.2.1 Resultados Vértebras de Ratos

Na Tabela 15 foram apresentados os resultados dos agrupamentos para os métodos K-médias, KI-médias, Subida da Encosta e Busca Tabu quando atuaram sozinhos ou em conjunto com o método de reamostragem. Tais resultados foram observados para o espaço de pré-formas.

Na Tabela 16 foram apresentados os resultados dos agrupamentos para os métodos K-médias, KI-médias, Subida da Encosta e Busca Tabu quando atuaram sozinhos ou em

Tabela 15 – Índices *CR* Vértebras de Ratos para o Espaço de Pré-formas

	Métodos baseados em Protótipos			
Critério	*K*-médias	*K*-médias *Bagging*	*KI*-médias	*KI*-médias *Bagging*
J^1	0.2108	0.5354	0.2105	0.6033
J^2	0.5354	0.6750	0.5354	0.6033
	Métodos baseados em Busca			
	Subida da Encosta	S. E. *Bagging*	Busca Tabu	B. T. *Bagging*
J^1	0.1334	0.6033	0.1330	0.6033
J^2	0.6748	0.4126	0.6748	0.6748

Fonte: Própria.

conjunto com o método de reamostragem. Tais resultados foram observados para o espaço tangente.

Tabela 16 – Índices *CR* Vértebras de Ratos para o Espaço Tangente

	Métodos baseados em Protótipos			
Critério	*K*-médias	*K*-médias *Bagging*	*KI*-médias	*KI*-médias *Bagging*
J^3	0.5355	**0.6751**	0.4117	0.6034
J^4	-0.0062	0.3558	-0.0217	0.6748
	Métodos baseados em Busca			
	Subida da Encosta	S. E. *Bagging*	Busca Tabu	B. T. *Bagging*
J^3	0.4117	0.4117	0.5355	0.6750
J^4	-0.0217	-0.0124	0.0246	0.3558

Fonte: Própria.

De acordo com os resultados das Tabelas 15 e 16, os algoritmos que incorporaram o método *Bagging* tiverem melhores desempenhos, em sua maioria, do que os algoritmos puros, levando em consideração ambos espaços de dados e todos os critérios de agrupamento utilizados. A exceção ficou por conta do algoritmo Subida da Encosta que ao usar o critério J^2. Essa constatação sobre o uso do *Bagging* demonstra a relevância do uso desse método de reamostragem para este conjunto de dados.

Atentando para os algoritmos quando agem de forma individual, cabe destacar o critério J^2 que superou os resultados do índice *CR* em realação ao critério J^1 para o espaço de pré-formas. No caso do espaço tangente, o critério que se sobressaiu foi o J^3 com resultados bem mais significativos em relação ao critério J^4.

O resultado de maior relevância para esse conjunto de dados foi alcançado pelo método *K*-médias *Bagging*, quando fez uso do critério J^3 para o espaço tangente. Cabe ressaltar que outros métodos, do mesmo espaço tangente e do espaço de pré-formas, também atingiram bons resulatdos, se aproximando do valor deste resultado de maior relevância.

4.3.2.2 Ganho Relativo ao Uso do *Bagging* para o Conjunto de Dados Vértebras de Ratos

A Tabela 17 mostra o ganho relativo entre os algoritmos sem o auxílio e trabalhando em conjunto com o método *Bagging* para este conjunto de dados.

Tabela 17 – Comparação entre Algoritimos de Agrupamento de acordo com o Ganho Relativo (%) do índice CR

Critério	**Métodos**			
	K-médias X *K*-médias *Bagging*	*KI*-médias X *KI*-médias *Bagging*	Subida da Encosta X S. E. *Bagging*	Busca Tabu X B. T. *Bagging*
	Espaço de Pré-formas			
J^1	153.98	186.60	352.24	353.60
J^2	26.07	12.68	-38.85	0.00
	Espaço Tangente			
J^3	26.06	46.56	0.00	26.05
J^4	-	-	-	1346.34

Fonte: Própria.

É possível observar na Tabela 17 que a técnica de *clustering ensembles*, onde é realizada a combinação de métodos, favoreceu o crescimento da qualidade dos agrupamentos, tanto para o espaço de pré-formas quanto para o espaço tangente, mesmo que os critérios de agrupamento fossem modificados. Vale observar que quando se fez uso do critério J^2, para a comparação entre os métodos Subida da Encosta e Subida da Encosta *Bagging*, a qualidade do agrupamento ser reduzida.

4.3.3 *Core Experiment* (CE-1B) - Peixes

Este conjunto de dados foi gerado a partir de imagens binárias da base de imagens *Core Experiment* ou simplesmente *CE-1* ((JEANNIN; BOBER, 1999)), criada pelo grupo *Moving Picture Experts Group* (*MPEG*) da *International Standards Organization* (*ISO*). Essa base é dividida em três partes: A, B e C. Cada uma dessas partes representa o que cada algoritmo que utiliza essa base de dados deve analisar. A parte A avalia o desempenho do algoritmo com relação às mudanças de escala e rotação. A parte B já avalia a capacidade de reconhecer formas semelhantes nos dados. A parte C se encarrega por analisar a capacidade do algoritmo trabalhar com mudanças de perspectivas ou deformações.

O objetivo desta base de dados é avaliar o desempenho dos algoritmos que trabalham com formas de objetos, descrevendo e procurando formas conexas, na presença de transformações geométricas, tais como mudança de escala, rotação e deformações. A parte da base de dados CE-1 utilizada, nesta pesquisa, foi a parte B (CE-1B) que é composta por 1400 imagens: são 70 classes de várias imagens e cada classe é composta por 20 imagens. A base de dados CE-1B apresenta certa dificuldade em sua análise por apresentar objetos

que são do tipo *outliers* e também por apresentar classes em que existe a presença de sobreposição nos dados.

Das 70 classes da base CE-1B, foram escolhidas 2 classes para compor os dados de Peixes deste trabalho: a classe *LMFish* e a classe *Fish*. A Figura 18 (a) mostra um exemplo da classe *LMFish* e a Figura 18 (b) mostra um exemplo de objeto da classe *Fish*.

Figura 18 – (a) Exemplo classe LMFish e (b) Exemplo classe Fish.

Fonte: (JEANNIN; BOBER, 1999).

Essas duas classes foram unidas para formar este conjunto de dados com 40 formas de objetos. Nos objetos deste novo conjunto de dados foram posicionados 15 marcos anatômicos, que melhor definiam a estrutura dos objetos, para descrever a forma dos dados. Tais marcos foram colocados levando em consideração os pontos de maior curvatura das formas. A extração das informações geradas pelas posições dos marcos anatômicos foi realizada utilizando os programas *tpsUtil* (versão 1.74) ((ROHLF, 1997)) que é utilizado para criar arquivos *thin-plate spline* (TPS) com as imagens selecionadas e *tpsDig2* (versão 2.30) ((ROHLF, 1998)), desenvolvidos F. James Rohlf. O *tpsDig2* é um programa para digitalizar marcos anatômicos e contornos para análises morfométricas. Este programa inclui operações simples de aprimoramento de imagem, fatores de escala, perfil de brilho da imagem e suporte para arquivos AVI e MOV.

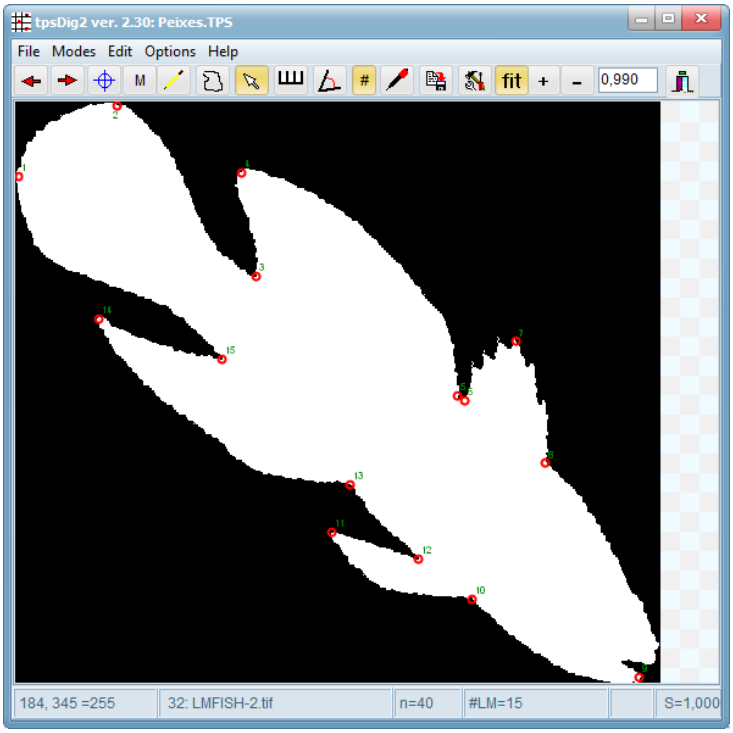
4.3.3.1 Resultados Peixes

Os resultados para os agrupamentos gerados pelos algoritmos, para os dados de Peixes no espaço de pré-formas, foram apresentados na Tabela 18. Os métodos K-médias, KI-médias, Subida da Encosta e Busca Tabu foram avaliados quando agiram sozinhos ou em conjunto com o método de reamostragem *Bagging*.

Os resultados para os agrupamentos gerados pelos algoritmos, para os dados de Peixes no espaço tangente, foram apresentados na Tabela 19. Os métodos K-médias, KI-médias, Subida da Encosta e Busca Tabu foram avaliados quando agiram sozinhos ou em conjunto com o método de reamostragem *Bagging*.

Considerando os resultados mostrados nas Tabelas 18 e 19, foi possível visualizar que os algoritmos baseados em protótipos ou em busca obtiveram desempenhos, da geração

Figura 19 – Quinze marcos anatômicos em um objeto do conjunto de dados Peixes no *software tpsDig2*.



Fonte: Própria.

Tabela 18 – Índices *CR Core Experiment* (CE-1B) - Peixes para o Espaço de Pré-formas

	Métodos baseados em Protótipos			
Critério	*K*-médias	*K*-médias *Bagging*	*KI*-médias	*KI*-médias *Bagging*
J^1	0.8047	0.0983	0.8047	**0.8998**
J^2	0.8047	0.0983	0.8047	**0.8998**
	Métodos baseados em Busca			
	Subida da Encosta	S. E. *Bagging*	Busca Tabu	B. T. *Bagging*
J^1	0.8047	**0.8998**	0.8047	**0.8998**
J^2	0.8047	**0.8998**	0.8047	**0.8998**

Fonte: Própria.

dos agrupamentos, semelhantes, para cada espaço de dados utilizado. Com relação aos métodos quando agiram de forma isoladada, a escolha do espaço de pré-formas é preferível ao espaço tangente por ter apresentado resultados consideravelmente mais elevados. Mais uma vez, a aproximação dos dados, realizada pelo espaço tangente, prejudicou a qualidade dos agrupamentos dos algoritmos avaliados.

Avaliando a ação do método de reamostragem *Bagging*, tanto o espaço de pré-formas quanto o espaço tangente obtiveram melhorias nos resultados em relação aos algoritmos puros. Um destaque se dá aos resultados para espaço tangente pela diferença significativa dada pela combinação dos métodos, evidenciando a necessidade do uso do método *Bagging*

Tabela 19 – Índices *CR Core Experiment* (CE-1B) - Peixes para o Espaço tangente

	Métodos baseados em Protótipos			
Critério	K-médias	K-médias *Bagging*	KI-médias	KI-médias *Bagging*
J^3	0.0648	0.0366	0.0648	**0.8998**
J^4	0.0648	0.0366	0.0648	**0.8998**
	Métodos baseados em Busca			
	Subida da Encosta	S. E. *Bagging*	Busca Tabu	B. T. *Bagging*
J^3	0.0648	**0.8998**	0.2292	**0.8998**
J^4	0.0648	**0.8998**	0.1367	**0.8998**

Fonte: Própria.

para este espaço de dados. A combinação dos métodos não conseguiu ser eficiente para o método K-médias nos dois espaços de dados, provocando uma perda na qualidade dos agrupamentos para este algoritmo.

4.3.3.2 Ganho Relativo ao Uso do *Bagging* para o Conjunto de Dados Peixes

A Tabela 20 mostra o ganho relativo entre os algoritmos sem o auxílio e usando o método *Bagging* para este conjunto de dados.

Tabela 20 – Comparação entre Algoritimos de Agrupamento de acordo com o Ganho Relativo (%) do índice CR

	Métodos			
Critério	K-médias X K-médias *Bagging*	KI-médias X KI-médias *Bagging*	Subida da Encosta X S. E. *Bagging*	Busca Tabu X B. T. *Bagging*
	Espaço de Pré-formas			
J^1	-87.78	11.81	11.81	11.81
J^2	-87.78	11.81	11.81	11.81
	Espaço Tangente			
J^3	-43.51	1288.58	1288.58	292.58
J^4	-43.51	1288.58	1288.58	558.22

Fonte: Própria.

Os ganhos relativos vistos na Tabela 20 mostraram, claramente, a dependência apresentada pelo espaço tangente em relação ao método de reamostragem. Os melhores ganhos foram obtidos pelos algoritmos KI-médias, Subida da Encosta e Busca Tabu atuando no espaço tangente, o que foi uma evolução para este espaço tendo em vista que os seus resultados atuando isoladamente tinham sido bem inferiores aos resultados obtidos pelos mesmos algoritmos no espaço de pré-formas. O uso do *Bagging* é necessário para corrigir a perda da qualidade do agrupamento provocada pela aproximação realizada nos dados.

Nesse contexto, quando se faz uso do método *Bagging*, com exceção do algoritmo K-médias, os resultados do índice de *Rand* corrigido são semelhantes, então, não existe

preferência entre os métodos e critérios de agrupamento para este conjunto de dados. O algoritmo K-médias não proporcionou ganhos relativos ao uso do método *Bagging* nem no espaço de pré-formas nem no espaço tangente.

4.4 CONSIDERAÇÕES FINAIS

Observando os experimentos realizados para todos os conjuntos de dados, é possível considerar que em relação aos dados simulados houve aumento no desempenho dos agrupamentos. Isso ocorreu quando se fez uso da combinação dos métodos isolados com os métodos de reamostragem, para todas as concentrações de dados escolhidas, para todos os critérios e para ambos espaços de dados. A observação dos resultados mostrou, também, que o uso do *Bagging* é essencial para o espaço tangente em comparação ao espaço de pré-formas. Ao trabalhar individualmente, o espaço de pré-formas obteve resultados consideravelmente mais relevantes que o espaço tangente, como era esperado.

Em relação aos dados reais, os resultados dependeram das características inerentes de cada conjunto de dados. No geral, a incorporação do método de reamostragem *Bagging* aos algoritmos contribuiu para um avanço na qualidade dos agrupamentos. Essa constatação não impediu que alguns algoritmos atingissem um ótimo global sem necessitar da reamostragem nos dados de entrada.

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 CONCLUSÕES

Nesta pesquisa foram apresentados métodos de agrupamentos baseados em protótipos e busca adaptados para trabalhar com dados da área de análise estatística de formas. A motivação para a utilização destes métodos está no fato de proporcionar novas opções de algoritmos de agrupamento para analisar formas planas de objetos e, assim, contribuir com alternativas de agrupamento de dados para a literatura de área análise estatística de formas. Foram apresentados os métodos de agrupamentos baseados em protótipos e busca, trabalhando individualmente e combinados com o método de reamostragem *Bagging*. O objetivo era avaliar o desempenho dos métodos em relação a espaço de dados, ao paradigma em que são baseados, aos critérios de agrupamento e a vantagem ou desvantagem da união com o método de reamostragem.

Foram realizados experimentos com dados simulados, onde foram geradas quatro configurações com características distintas, e com três conjuntos de dados reais de formas de objetos compostos pelos conjuntos de Crânios de Gorilas, Vértebras de Ratos e Peixes. Estes conjuntos de dados foram utilizados pelos métodos K-médias, KI-médias, Subida da Encosta e Busca Tabu. Estatísticas de teste de hipóteses, já existentes na literatura para análises estatística de formas, foram empregadas como critérios de agrupamento. Um critério baseado em soma de quadrados, adaptado para o espaço de pré-formas, também foi utilizado. O algoritmo K-médias foi proposto para ser usado com os critérios de agrupamento baseados na estatísticas de testes de hipóteses, pois já havia sido implementado para o critério de soma de quadrados para dados da análise de formas.

No geral, os resultados obtidos mostraram que os métodos baseados em protótipos e busca obtiveram melhores desempenhos para a qualidade dos agrupamentos quando atuaram em conjunto com a reamostragem, fornecida pelo método *Bagging*, do que quando tentavam formar grupos de forma isolada. Para os dados simulados, as combinações dos algoritmos foram eficientes para todas as configurações de espaços de dados, concentrações e critérios de agrupamento. Para estes conjuntos de dados, o espaço de pré-formas é o mais recomendado para realizar o agrupamento dos dados de formas planas, em especial quando se faz uso do critério J^1. O estudo realizado sobre as quatro concentrações dos dados simulados evidenciou, também, como é fundamental o uso do método *Bagging* para corrigir os efeitos causados pela aproximação realizada nos dados realizada pelo espaço tangente.

Na avaliação dos resultados dos dados reais, a versão combinada dos métodos também causou influência na melhoria de desempenho dos algoritmos. Quando foram usados em versões individuais, os algorimos baseados em busca, com o critério J^3, são os mais in-

dicados para os dados de Crânios de Gorilas porque atingiram o ótimo global para esses dados. Os algorimos baseados em busca no espaço de pré-formas são mais adequados para os dados de Vértebras de Ratos, preferencialmente utilizando o critério J^2. Os subconjuntos de dados da base CE1-B de Peixes tem como melhor opção os algoritmos do espaço de pré-formas.

Já para as versões com o uso do *Bagging*, os dados de Crânios de Gorilas obtiveram a formação mais eficiente dos agrupamentos com os algoritmos K-médias *Bagging* e KI-médias *Bagging* usando o critério J^3. Para os dados de Vértebras de Ratos, o algoritmo mais indicado também foi o K-médias *Bagging*, pertencendo ao espaço tangente, para o critério J^3. Observando o conjunto de dados de Peixes, os resultados foram semelhantes para a maioria dos métodos, com exceção do K-médias.

Assim, este trabalho contribuiu para a literatura teórica dos métodos de agrupamento para dados de análise estatística de formas com o desenvolvimento de métodos de agrupamento adaptados para a estrutura dos tipos dados de formas planas. Colaborou também com a proposta de utilização das estatísticas de teste de hipóteses como novos critérios de agrupamento e com a incorporação da técnica de *cluster ensembles*, usando o método de reamostragem *Bagging* em cooperação com os algoritmos de agrupamento, a fim de melhorar a eficiência dos algoritmos reduzindo a variabilidade dos dados.

5.2 TRABALHOS FUTUROS

Para dar continuidade a esta pesquisa, a proposta é utilizar o algoritmo de agrupamento baseados em medóides, que é conhecido por ser menos sensível na presença de observações aberrantes/ruídos, para os dados da análise estatística de formas. Também serão analisadas as versões difusas para os métodos propostos, em que os dados serão associados aos grupos usando uma função de pertinência. Para validação dos métodos, novos conjuntos de dados reais serão observados. Também será analisada uma nova função de consenso para os algoritmos que usam o método *Bagging* e, além disso, utilizar outros métodos de reamostragem em conjunto com os algoritmos. Cada etapa dos trabalhos futuros é listada abaixo:

1. Propor e analisar o método de agrupamento baseados em medóides (o algoritmo *K-medoids*) que é menos sensível à *outliers* para formas planas;
2. Introduzir as versões difusas para os algoritmos baseados em protótipos a fim de verificar seu comportamento em comparação às versões rígidas;
3. Analisar uma nova função de consenso para o método *Bagging*;
4. Utilizar outros métodos de reamostragem;
5. Acrescentar outros conjuntos de dados reais para validar os métodos.

REFERÊNCIAS

AITKIN, M.; WILSON, G. T. Mixture models, outliers, and the em algorithm. *Technometrics*, v. 22, p. 325–331, 1980.

AMARAL, G. J. A.; DORE, L. H.; LESSA, R. P.; STOSIC, B. k-means algorithm in statistical shape analysis. *Communications in Statistics - Simulation and Computation*, v. 39, p. 1016–1026, 2010.

AMARAL, G. J. A.; DRYDEN, I. L.; WOOD, A. T. A. Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. *Journal of the American Statistical Association*, v. 102, p. 695–707, 2007.

BANFIELD, C.; BASSIL, L. A transfer algorithm for nonhierarchical classification. *Applied Statistics*, v. 26, p. 206–210, 1977.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1st. ed. New York: Plenum Press, 1981.

BILMES, J. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *Report, International Computer Science Institute, Berkeley, CA*, 1998.

BLASHFIELD, R. K. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, v. 83, p. 377–385, 1976.

BOOKSTEIN, F. *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge: Cambridge University Press, 1991.

BOOKSTEIN, F. L. Size and shape spaces for landmark data in two dimensions (with discussion). *Statistical Science*, v. 1, p. 181–222, 1986.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, p. 123–140, 1996.

BRUSCO, M. J.; STEINLEY, D. A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. *Psychometrika*, v. 72, p. 583–600, 2007.

CARVALHO, F. A. T. D.; SOUZA, R. M. C. R. Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recognition Letters*, v. 31, p. 430–443, 2010.

CARVALHO F. A. T., D. M. F. M. D.; LECHEVALLIER, Y. A multi-view relational fuzzy c-medoid vectors clustering algorithm. *Neurocomputing (Amsterdam)*, v. 163, p. 115–123, 2015.

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

CHERKASSKY, V.; MULIER, F. *Learning from data: concepts, theory, and methods*. 2nd. ed. New Jersey: John Wiley and Sons, 2007.

CHERNICK, M. R.; LABUDDE, R. A. *An Introduction to Bootstrap Methods with Applications to R*. 1st. ed. New Jersey: John Wiley and Sons, 2011.

CRISTIANINI, N.; SHAEW-TAYLOR, J. *An introduction to Support Vector Machines and other kernel-based learning methods*. United States of America: Cambridge University Press, 2000.

DEMPSTER A. P., L. N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, v. 39, p. 1–38, 1977.

DIDAY, E.; GOVAERT, G. Classification automatique avec distances adaptatives. *Informatique Computer Science*, v. 11, p. 329–349, 1977.

DIDAY, E.; SIMON, J. J. Clustering analisys. *Digital Pattern Recognition*, p. 47–94, 1976.

DRYDEN, I. L. *The Statistical Analysis of Shape analysis. Ph. D. thesis.* United Kingdom: University of Leeds, 1989.

DRYDEN, I. L.; MARDIA, K. V. *Statistical Shape Analysis*. 1st. ed. New York: John Wiley and Sons, 1998.

DUDOIT, S.; FRIDLYAND, J. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, v. 19, p. 1090–1099, 2003.

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact, well-separated clusters. *Journal of Cybernetics*, v. 3, p. 32–57, 1974.

EFRON, B. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, v. 7, p. 1–26, 1979.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. *Cluster Analysis*. 5rd. ed. London: John Wiley and Sons, 2011.

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

FERREIRA M. R.P., D. C. F. d. A. T.; SIMõES, E. C. Kernel-based hard clustering methods with kernelization of the metric and automatic weighting of the variables. *Pattern Recognition)*, v. 51, p. 310–321, 2016.

FILIPPONEA M., C. F. M. F.; ROVETTAA, S. A survey of kernel and spectral methods for clustering. *Pattern Recognition*, v. 41, p. 176–190, 2008.

FRIEDMAN, H. P.; RUBIN, J. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, v. 62, p. 1159–1178, 1967.

GEORGESCU, V. Clustering of fuzzy shapes by integrating procrustean metrics and full mean shape estimation into k-means algorithm. *In: The 13th IFSA (International Fuzzy Systems Association) World Congress and The 6th Conference of EUSFLAT (European Society for Fuzzy Logic and Technology)*, p. 1679–1684, 2009.

GHAEMI R., S. M. N. I. H.; MUSTAPHA, N. A survey: Clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, v. 26, p. 636–645, 2009.

GIROLAMI, M. Mercer kernel based clustering in feature space. *IEEE Transactions on Neural Networks*, v. 13, p. 780–784, 2002.

GLOVER, F. Tabu search - part i. *ORSA Journal on Computing*, v. 1, p. 190–206, 1989.

GOODALL, C. R. Procrustes methods in the statistical analysis of shape (with discussion). *Journal of the Royal Statistical Society: Series B*, v. 53, p. 285–339, 1991.

GORDON, A. D. *Classification*. 2nd. ed. United States: Chapman Hall/CRC, 1999.

HAGEMAN J. A., v. d. B. R. A. W. J. A. H. H. C. J.; SMILDE, A. K. Bagged k-means clustering of metabolome data. *Critical Reviews in Analytical Chemistry*, v. 36, p. 211–220, 2007.

HAJJAR, C.; HAMDAN, H. Interval data clustering using self-organizing maps based on adaptive mahalanobis distances. *Neural Networks*, v. 46, p. 124–132, 2013.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco: Morgan Kaufmann Publisher, 2006.

HANSEN, P.; MLADENOVIC, N. J -means: A new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, v. 34, p. 405–413, 2001.

HARTIGAN, J. A.; WONG, M. A. A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, v. 28, p. 100–108, 1979.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. 2nd. ed. California: Springer, 2009.

HONG, Y.; KWONG, S. Learning assignment order of instances for the constrained k-means clustering algorithm. *IEEE Trans. Syst., Man, Cybern. B*, v. 39, p. 568–574, 2009.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, v. 2, p. 193–218, 1985.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys*, v. 31, p. 264–323, 1999.

JAMES, G. S. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of thepopulation variances are unknown. *Biometrika*, v. 41, p. 19–43, 1954.

JAMES G., W. D. H. T.; TIBSHIRANI, R. *An Introduction to Statistical Learning with Applications in R*. 7th. ed. New York: Springer, 2017.

JEANNIN, S.; BOBER, M. *Description of ce for mpeg-7 motion/shape*. Seoul: Technical Report ISO/IEC JTC1/SC29/WG11/N2690, 1999.

JIA J., X. X. L. B.; JIAO, L. Bagging-based spectral clustering ensemble selection. *Pattern Recognition Letters*, v. 32, p. 1456–1467, 2011.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. 6th. ed. New Jersey: Pearson, 2008.

KAUFMAN, L.; ROUSSEEUW, P. J. Clustering by means of medoids. *Statistical Data Analysis Based on the L1–Norm and Related Methods*, p. 405–416, 1987.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data, An Itroduction to Cluster Analysis*. Belgium: John Wiley and Sons, 1990.

KENDALL, D. The diffusion of shape. *Statistical Science*, v. 9, p. 428–430, 1977.

KENDALL, D. G. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, v. 16, p. 81–121, 1984.

KENDALL, D. G.; BARDEN, D.; CARNE, T. K.; LE, H. *Shape and Shape Theory*. 1st. ed. New York: Wiley Series in Probability and Statistics, 1999.

KENT, J. T. The complex bingham distribution and shape analysis. *Journal of the Royal Statistical Society, Series B*, v. 56, p. 285–299, 1994.

KENT, J. T.; CONSTABLE, P. D. L.; ER, F. Simulation for de complex bingham distribution. *Statistics and Computing*, v. 14, p. 53–57, 2004.

KIRKPATRICK, S.; GELATT, C. D.; ANDVECCHI, M. P. Optimization by simulated annealing. *Science*, v. 220, p. 671–680, 1983.

KLEIN, R. W.; DUBES, R. C. Experiments in projection and clustering by simulated annealing. *Pattern Recognition*, v. 22, p. 213–220, 1989.

LEISCH, F. Bagged clustering. *Working paper 51 SFB adaptive information systems and modelling in economics and management science WU Vienna University of Economics and Business*, 1999.

LLOYD, S. Least squares quantization in pcm. *IEEE Transactions Information Theory*, v. 28, p. 129–137, 1982.

MACIEL D. B. M., A. G. J. A. S. R. M. C. R.; PIMENTEL, B. A. Multivariate fuzzy k-modes algorithm. pattern analysis and applications. *Neurocomputing (Amsterdam)*, v. 20, p. 59–71, 2017.

MACQUEEN, J. Some methods for classifcation and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, v. 1, p. 281–297, 1967.

MARDIA, K. V. Discussion to 'a survey of the statistical theory of shape' by d. g. kendall. *Statistical Science*, v. 4, p. 108–111, 1989.

MARRIOTT, F. H. C. Optimization methods of cluster analysis. *Biometrika*, v. 69, p. 417–421, 1982.

MAULIK, U.; BANDYOPADHYAY, S. Genetic algorithm-based clustering technique. *Pattern Recognition*, v. 3, p. 1455–1465, 2000.

MULLER K.-R., M. S. R. G. T. K.; SCHOLKOPF, B. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, v. 2, p. 181–201, 2001.

NABIL, M.; GOLALIZADEH, M. On clustering shape data. *Journal of Statistical Computation and Simulation*, v. 86, p. 2995–3008, 2016.

O'HIGGINS, P. *A morphometric study of cranial shape in the Hominoidea. Ph. D. thesis*. United Kingdom: University of Leeds, 1989.

O'HIGGINS, P.; DRYDEN, I. L. Sexual dimorphism in hominoids: further studies of craniofacial shape differences in pan, gorilla, pongo. *Journal of Human Evolution*, v. 24, p. 183–205, 1993.

OLIVEIRA, R. A. D. *Algoritmos para Determinação do Número de Grupos em Estudos de Formas Planas*. Brasil: Universidade Federal de Pernambuco, 2016.

PACHECO, J.; VALENCIA, O. Design of hybrids for the minimum sum-of-squares clustering problem. *Computational Statistics and Data Analysis*, v. 43, p. 235–248, 2003.

PIMENTEL, B. A.; SOUZA, R. M. C. R. A multivariate fuzzy c-means method. *Applied Soft Computing (Print)*, v. 13, p. 1592–1607, 2013.

QUENOUILLE, M. H. Approximate tests of correlation in time series. *J. R. Stat. Soc. B*, v. 11, p. 18–84, 1949.

ROHLF, F. J. *tpsUtil*. Version 1.74. New York: http://life.bio.sunysb.edu/morph/, 1997.

ROHLF, F. J. *tpsDig2*. Version 2.30. New York: http://life.bio.sunysb.edu/morph/, 1998.

RUSPINI, E. H. A new approach to clustering. *Information and Control*, v. 15, p. 22–32, 1969.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. New Jersey: Pearson Education Limited, 2014.

SEBER, G. A. F. *Multivariate Observations*. 5rd. ed. New York: John Wiley and Sons, 1984.

SERGIOS, T.; KOUTROMBAS, K. *Pattern Recognition*. 4th. ed. New York: Academic Press, 2006.

SMALL, C. G. *The Statistical Theory of Shape*. 1st. ed. New York: Springer-Verlag, 1996.

SOUZA, R.; CARVALHO, F. A. D. Dynamical clustering of interval data based on adaptive chebyshev distances. *IEE Electronics Letters*, v. 40, p. 658–659, 2004.

TUKEY, J. W. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics.*, v. 29, p. 614–623, 1958.

VANISRI, D.; LOGANATHAN, D. Survey on fuzzy clustering and rule mining. *International Journal of Computer Science and Information Security*, v. 8, p. 183–187, 2010.

VELMURUGAN, T.; SANTHANAM, T. A survey of partition based clustering algorithms in data mining: An experimental approach. *Information Technology Journal*, v. 10, p. 478–484, 2011.

VINUE, G.; SIMO, A.; ALEMANY, S. The k-means algorithm for 3d shapes with an application to apparel design. *Advances in Data Analysis and Classification*, v. 10, p. 1–30, 2014.

WANG, J.; SU, X. An improve k-means clustering algorithm. *Communication Software and Networks (ICCSN)*, p. 44–46, 2011.

WU, K.-L. Analysis of parameter selections for fuzzy c-means. *Pattern Recognition*, v. 45, p. 407–415, 2012.

XU, R.; WUNSCH, D. I. I. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, v. 16, p. 645–678, 2005.

YANG, M. S. A survey of fuzzy clustering. *Mathematical and Computer Modelling*, v. 18, p. 1–16, 1993.

YANG, Y.; JIANG, J. Hybrid sampling-based clustering ensemble. *IEEE Transactions on Neural Networks and Learning Systems*, v. 27, p. 952–965, 2016.

YU Z., L. L. W. H. Y. J. H. G. G. Y.; YU, G. Probabilistic cluster structure ensemble. *Information Sciences*, v. 267, p. 16–34, 2014.

ZADEH, L. Fuzzy sets. *Information and Control*, v. 3, p. 338–353, 1965.

ZONG Y., J. P. X. G.; PAN, R. A clustering algorithm based on local accumulative knowledge. *Journal Of Computers*, v. 8, p. 365–371, 2013.