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

JERFSON BRUNO DO NASCIMENTO HONÓRIO

CLASSIFICAÇÃO NÃO SUPERVISIONADA NO CONTEXTO DE TAMANHO E FORMA

Recife

2022

JERFSON BRUNO DO NASCIMENTO HONÓRIO

CLASSIFICAÇÃO NÃO SUPERVISIONADA NO CONTEXTO DE TAMANHO E FORMA

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador (a): Getúlio José Amorim do Amaral

Recife

2022

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

H774c Honório, Jerfson Bruno do Nascimento
Classificação não supervisionada no contexto de tamanho e forma / Jerfson Bruno do Nascimento Honório. – 2022.
57 f.: il., fig., tab.

Orientador: Getúlio José Amorim do Amaral.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2022.
Inclui referências.

1. Estatística aplicada. 2. Bagging. 3. Boosting. 4. Hill climbing. 5. k-médias. I. Amaral, Getúlio José Amorim do (orientador). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2022 - 48

JERFSON BRUNO DO NASCIMENTO HONÓRIO

"CLASSIFICAÇÃO NÃO SUPERVISIONADA NO CONTEXTO DE TAMANHO E FORMA"

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

Aprovada em: 17 de fevereiro de 2022.

BANCA EXAMINADORA

Professor Getúlio José Amorim do Amaral

DE/UFPE

Professora Fernanda De Bastiani

DE/UFPE

Professora Lúcia Pereira Barroso

IME/USP

Este trabalho é dedicado às crianças adultas que, quando pequenas, sonharam em se tornar cientistas.

"Um general nunca demonstra desespero. Ele inspira confiança em suas tropas. Ele os leva adiante, mesmo que seja para a morte." (RIORDAN, 2011)

RESUMO

Esta dissertação tem como objetivo propor métodos não supervisionados de classificação para dados de tamanho e forma, considerando imagens bidimensionais (formas planas). Os novos métodos são baseados em testes de hipóteses, algoritmo K-médias e o algoritmo *hill climbing*. Também propomos as combinações dos algoritmos, com os métodos de ensemble: *bagging* e *boosting*. Para os dados simulados, gerados a partir da distribuição normal complexa, propomos três possíveis cenários para avaliar o desempenho dos métodos propostos. Neles, as combinações dos algoritmos foram superiores às suas versões base, sendo o algoritmo *bagging hill climbing*, o mais poderoso em dois cenários. Ainda pelos resultados numéricos, concluímos que quando os tamanhos dos centroides se diferenciam, o desempenho dos algoritmos melhora. Para os conjuntos de dados reais (vértebras torácicas T2 de camundongos,ressonância magnética de pessoas com esquizofrenia e crânio de grandes macacos), os métodos ensembles (*bagging* e *boosting*) novamente foram o destaque, sendo sempre superiores às versões base. Finalmente, considerando os dados sintéticos e reais, o *bagging hill climbing* é escolhido como o melhor método.

Palavras-chaves: *bagging; boosting; hill climbing;* k-médias; marcos anatômicos; teste de hipóteses.

ABSTRACT

This master thesis aims to propose unsupervised classification methods for size and shape data, considering two-dimensional images (planar shape). The new methods are based on hypothesis testing, K-means algorithm and hill climbing algorithm. We also propose combinations of algorithms, with ensemble methods: bagging and boosting. For the simulated data, generated from the complex normal distribution, we propose three possible scenarios to evaluate the performance of the proposed methods. In them, the combinations of the algorithms were superior to their base versions, with the bagging hill climbing algorithm being the most powerful in two scenarios. Also from the numerical results, we conclude that when the centroid sizes are different, the performance of the algorithms improves. For the real data sets (T2 thoracic vertebrae of mice, magnetic resonance imaging of people with schizophrenia and skull of great apes), the ensembles methods were again the highlight, being always superior to the base versions, the bagging and boosting methods achieve the best performance in data sets. Finally, considering the synthetic and real data, bagging hill climbing is chosen as the best method.

Keywords: boosting; bagging; hill climbing; hypothesis test; k-means; landmarks.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com os avanços da tecnologia, a captura de imagens bidimensionais e tridimensionais tem se tornado cada vez mais comum no nosso cotidiano. Essas imagens fornecem diversas informações para estudos estatísticos, sendo essa área chamada de morfometria. A morfometria é uma das maneiras de estudar estas imagens que se encontram bem consolidadas com diversas aplicações, tais como: Medicina, Zoologia, Biologia e outros. Nesse contexto, existem estudos que tratam da forma e estudos que tratam o tamanho e forma dos objetos capturados nas imagens. No caso de forma, os efeitos de locação, escala e rotação são removidos. No caso de tamanho e forma, o efeito de escala não é removido.

Uma das mudanças mais importantes, embora relativamente menos aclamados dos avanços da tecnologia no século atual, foi o amadurecimento da análise estatística de tamanho e forma como uma área teórica e aplicada, uma vez que no atual século a maioria das tecnologias usam reconhecimento facial, ou seja, propriedades geométricas de tamanho e forma. As aplicações da análise estatística de tamanho e forma se estendem por quase todas as áreas científicas e tecnológicas aplicadas, das menores às maiores escalas.

A análise de tamanho e forma dos objetos pode ser útil para a tomada de importantes decisões, como a de um médico que precisa decidir se um câncer é maligno ou benigno, baseado em uma ressonância magnética digitalizada. Este tipo de decisão pode ser tomada, por essa área trabalhar com as informações contidas nos objetos.

O primeiro trabalho nesta área foi feito por Kendall 1977. No entanto, foi no trabalho de Kendall 1984, que a análise de formas e tamanho e forma foram formalizadas, assim, foram definidos os conceitos básicos das áreas. Um dos pontos mais importantes do trabalho de Kendall, foi a proposta dos sistemas de coordenadas, que tinham a finalidade de obter a forma de um objeto através de pontos dispostos, chamados de marcos anatômicos. Essas coordenadas propostas estão dispostas em um espaço não euclidiano e, portanto, são necessárias ferramentas próprias para lidar com as análises.

O tamanho e a forma de um objeto, são as informações que permanece quando os efeitos de locação e rotação são removidos através de operações matemáticas, como descrito em Dryden e Mardia 2016. Quando são removidos esses efeitos, os dados são chamados de dados de tamanho e forma, e são descritos em um espaço não Euclidiano.

Em diversas ocasiões, em análise estatística de tamanho e forma, é necessário agrupar um

conjunto de dados em grupos, de tal maneira, que se tenha grupos com características mais homogêneas.

O agrupamento é um dos relevantes tópicos em análise estatística de tamanho e forma, pois, os algoritmos de agrupamento existentes são projetados para o espaço euclidiano, os tornando algoritmos limitados para análise estatística de tamanho e forma. Assim, o desenvolvimento de algoritmos para o espaço não euclidiano é necessário quando forem utilizados dados nesse espaço. Observando a relevância de agrupamentos para analisar o tamanho e a forma de objetos, propomos uma generalização do algoritmo K-médias e *hill climbing*, a fim de integrar métricas para que se possa usar dados de tamanho e forma.

Amaral et al. 2010 adaptou o método K-médias proposto por Macqueen 1967 para um problema real de oceanografia, em que é necessário obter a classificação não supervisionada de espécies similares de peixes. Nesse cenário, ele fez adaptações usando dados de pré-formas.

Esta dissertação utiliza os métodos de agrupamento baseada em K-médias, proposto por Macqueen 1967, *hill climbing*, proposto por Landau et al. 2011, e uma modificação do algoritmo *hill climbing* baseados em teste de hipóteses. Esses algoritmos foram adaptados para trabalhar com conjuntos de dados de tamanho e forma, em que o espaço é não euclidiano. Os algoritmos também foram usados como classificadores bases para os métodos *ensembles*; os métodos *ensembles* (*boosting* e *bagging*) são baseados na noção de combinar vários classificadores básicos de forma que o classificador final ensemble, obtenha um desempenho melhor do que cada classificador base individual.

Para validar os métodos propostos, foram realizados experimentos com três conjuntos de dados simulados e três conjuntos de dados reais (Vértebra de Ratos, Ressonância magnética de pessoas com esquizofrenia e Crânios de macacos). As métricas para os agrupamentos foram usadas a partir dos métodos de validação interna e externa, além da acurácia. Validação externa e validação interna são as duas principais categorias de validação de agrupamentos. A principal diferença é se informação externa, conhecimento a priori dos objetos, é usada ou não para validar os agrupamentos. [Landau et al. 2011].

1.1 OBJETIVOS

Segue alguns dos objetivos dessa dissertação:

- Realizar um estudo sobre análise estatística de tamanho e forma;

- Propor e analisar novas versões de algoritmos de agrupamento (K-médias, *hill climbing* e *hill climbing*: baseado em testes de hipóteses) no espaço de tamanho e forma;
- Avaliar a utilização dos métodos de validação (índice de Rand, índice de Procrustes e acurácia) no espaço de tamanho e forma;
- Introduzir os métodos *ensembles* (*boosting* e *bagging*) para classificação não supervisionada no espaço de tamanho e forma.

1.2 ORGANIZAÇÃO DA DISSERTAÇÃO

Os próximos capítulos desta dissertação de mestrado estão organizados da seguinte forma:

Capítulo 2: Análise Estatística de Tamanho e Forma: este capítulo apresentará uma introdução sobre análise estatística de tamanho e forma e de algumas métricas, utilizadas nesse espaço.
Capítulo 3: Classificação Não Supervisionada: neste capítulo serão apresentados uma introdução sobre classificação não supervisionada e seus mais famosos algoritmos. Além das suas versões modificadas para dados de tamanho e forma.
Capítulo 4: Métodos *Ensembles*: neste capitulo serão apresentados uma introdução sobre os métodos *ensembles* (*boosting* e *bagging*) e como eles foram modificados para classificação não supervisionada e dados de tamanho e forma.
Capítulo 5: Métodos de Validação: neste capitulo serão apresentados alguns métodos de validação que propomos para avaliar os agrupamentos.
Capítulo 6: Experimentos e Resultados: neste capitulo serão descritos os conjuntos de dados simulados e de dados reais.
Capítulo 7: Conclusões e Trabalhos Futuros: e finalmente, neste capítulo são apresentadas as conclusões referentes à pesquisa realizada. Aqui, também são descritos os trabalhos futuros referentes à possíveis continuidade desta pesquisa.

1.3 *FRAMEWORK* COMPUTACIONAL

Segue toda a configuração utilizada para realizar as simulações desta dissertação.

Sistema Operacional: macOs Monterey 12.1 (RyzenTosh)
Tipo de sistema operacional: 64 bits
Processadores: 8 x AMD Ryzen 7 2700x CPU @ 3,8GH
Memória: 16 GiB de RAM
Processador gráfico: AMD Radeon RX Vega64

Das informações relacionadas a versão do software para as análises são: R Core Team 2021 versão 4.1.2. Pacote usado: shapes versão 1.2.6.

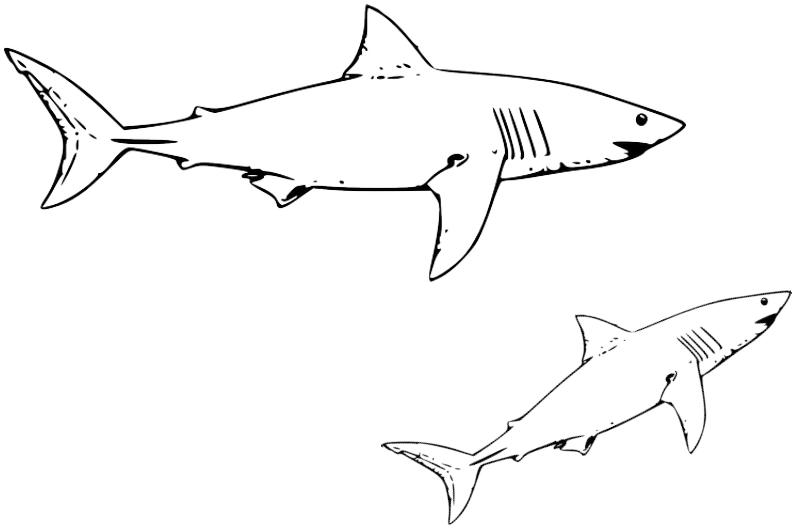
2 ANÁLISE ESTATÍSTICA DE TAMANHO E FORMA - *SIZE-AND-SHAPE ANALYSIS*)

A análise de tamanho e forma é uma área bastante útil para lidar com formas de objetos e informações geométricas. É um tópico relevante para vários campos de pesquisa e aplicações, uma vez que no atual século a maioria das tecnologias usam reconhecimento facial, ou seja, propriedades geométricas de tamanho e forma. As aplicações da análise estatística de tamanho e forma se estendem por quase todas as áreas científicas e tecnológicas aplicadas, das menores às maiores escalas, dentre elas temos: biologia, medicina, análises de imagem, arqueologia, geografia, geologia, agricultura, genética, reconhecimento de alvos militares etc. Dryden e Mardia 2016 descrevem em seu livro os principais conceitos sobre análise de tamanho e forma, além de apresentar várias aplicações no mundo real.

Tamanho e Forma (*size-and-shape*) é definido como "toda informação geométrica que permanece quando os efeitos de locação e rotação são removidos de um objeto". [Kendall 1977].

"Dizemos que dois objetos têm o mesmo tamanho e forma, se eles puderem ser transladados e girados um ao outro, de modo que correspondam exatamente, isto é, se os objetos forem transformações de corpo rígido um do outro". [Dryden e Mardia 2016].

Figura 1 – Duas representações de tubarões com diferentes locações, escalas e rotações.



Fonte: Autor

Nas próximas seções serão apresentadas algumas abordagens feitas em análise estatística

de tamanho e forma.

2.1 MARCOS ANATÔMICOS - *LANDMARKS*

Dentre as diversas maneiras de extrair as informações de tamanho e formas de objetos, uma abordagem muito utilizada em morfometria, 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 ou sobre o objeto, e suas coordenadas numéricas são utilizadas para representar esse objeto. A posição dos pontos adicionados aos objetos, está associada a normalmente às coordenadas cartesianas e estas coordenadas pertencem a um espaço que é chamado de espaço dos marcos anatômicos. A definição de marcos, segundo Dryden e Mardia 2016, é "um ponto correspondente em cada objeto que corresponde entre e dentro da população", ou seja, todos os indivíduos possuem os mesmos marcos e as mesmas quantidades (homólogos). Esses pontos podem ser no espaço bi ou tridimensional, e os mesmos correspondem exclusivamente a uma posição de uma característica particular do objeto de interesse. Segundo Kendall et al. 2009, o número de pontos não segue uma restrição teórica.

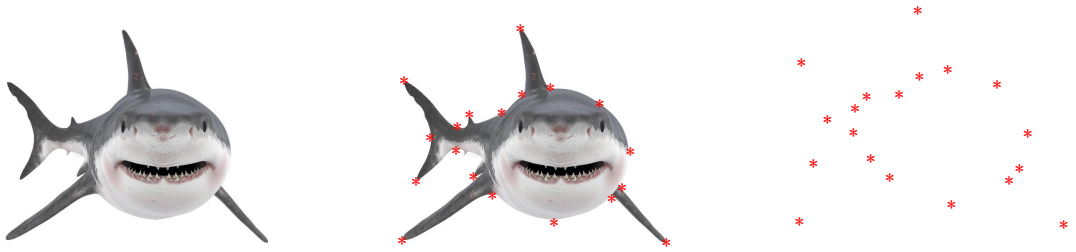
Dryden e Mardia 2016 define que os marcos são divididos em três sub-grupos:

1. **Marcos anatômicos:** são pontos atribuídos por um especialista na área de estudo que correspondem a alguma característica biologicamente significativa.

2. **Marcos matemáticos:** são pontos alocados em um objeto de acordo com algumas propriedades matemáticas, por exemplo: pontos de alta curvatura, inflexão, máximos, mínimos etc.

3. **Pseudo-Marcos:** são pontos construídos em um objeto, localizados no contorno ou entre os marcos anatômicos e/ou matemáticos. Sua utilização é mais para conceituar a forma do objeto.

Figura 2 – Representação da forma baseada em marcos.



(a) Contorno e forma do objeto.

(b) Marcos anatômicos.

(c) 19 marcos para configuração da forma do objeto.

Fonte: Autor

Na Figura 2, pode-se notar a representação do contorno de um tubarão, os marcos anatômicos e a configuração da forma. Quando digitalizados os marcos, os objetos geralmente possuem tamanhos, posições e rotações diferentes, dentro do equipamento de medição. Por isso, é necessário retirar esses efeitos. Em análise de tamanho e forma, apenas os efeitos de locação e rotação do conjunto de dados são retirados. Dessa forma, nas próximas seções, mostram-se os passos para filtrar esses efeitos.

2.2 FILTRANDO EFEITOS

Inicialmente, devemos especificar um sistema de coordenadas para definir como serão realizadas as remoções dos efeitos, e também, para descrever o tamanho e a forma de um objeto. Em análise estatística de tamanho e forma existem diversos sistemas de coordenadas, sendo algum deles: coordenadas de Bookstein, propostas por Bookstein 1984 e Bookstein 1986; coordenadas polares de Kent proposta por Kent 1994; coordenadas de forma Goodall-Mardia QR desenvolvida por Goodall e Mardia 1992 e Goodall e Mardia 1993 e as coordenadas de Kendall, proposta por Kendall 1984.

As transformações mencionadas anteriormente podem ser feitas de diversas maneiras dependendo do sistema de coordenadas, mas nesta dissertação, serão usadas as coordenadas de Kendall 1984.

Para representar matematicamente a forma de um objeto, inicialmente k marcos anatômicos são escolhidos para o contorno deste objeto. Seja $\mathbf{Y}$ a matriz de coordenadas cartesianas $k \times m$ de k marcos anatômico e m dimensões. As coordenadas desses marcos anatômicos são

representadas pela seguinte matriz de configuração:

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

Ao obtermos a matriz de configuração do objeto, os dados de tamanho e forma são obtidos removendo os efeitos de locação e rotação. Para eliminar estes efeitos, algumas transformações precisam ser realizadas na matriz de configuração $\mathbf{Y}$. Vale destacar que nesta dissertação serão considerados os casos onde $k \geq 3$ e $m = 2$, o que corresponde às formas planas. Assim, a matriz de configuração $\mathbf{Y}$ da expressão 2.1 resume-se a

$$\mathbf{Y} = [\mathbf{y}_1 \quad \mathbf{y}_2]^\top = \begin{bmatrix} y_{1,1} & y_{1,2} \\ y_{2,1} & y_{2,2} \\ \vdots & \vdots \\ y_{k,1} & y_{k,2} \end{bmatrix}.$$

Para iniciar as transformações em busca do tamanho e da forma do objeto, o primeiro passo é remover o efeito de locação. Um dos pontos mais importantes do trabalho de Kendall 1984 foi a proposta dos sistemas de coordenadas. Estes sistemas possuem finalidade de obter a forma de um objeto por meio dos pontos dispostos através dos marcos. Dessa forma, para remover a locação devemos definir a sub-matriz de Helmert $(\mathbf{H})$. Essa sub-matriz $\mathbf{H}$ é a matriz de Helmert $(k-1) \times k$ sem a primeira linha. A matriz Helmert completa $\mathbf{H}^F$ é comumente usada em estatística, sendo a mesma uma matriz ortogonal quadrada $k \times k$ com sua primeira linha de elementos igual a $1/\sqrt{k}$. Já as j-ésima linha possui $j+1$ elementos, $j \geq 1$, iguais a:

$$(h_j, ..., h_j, -jh_j, 0, ..., 0), \quad h_j = -\{j(j+1)\}^{-1/2},$$

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

Um exemplo da matriz de Helmert, dado $k = 4$, é definida como:

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

e sua sub-matriz é definida por:

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

Para remover o efeito de locação da matriz de configuração $\mathbf{Y}$, basta multiplicar a matriz de configuração pela sub-matriz de Helmert $\mathbf{H}$ de dimensão $(k-1) \times k$. Com isso, a configuração Hermertizada é dada por:

$$\mathbf{Y}_{\mathbf{H(k-1 \times 2)}} = \mathbf{H}_{\mathbf{(k-1 \times k)}} \mathbf{Y}_{\mathbf{(k \times 2)}}. \tag{2.2}$$

Vale notar que a configuração Helmertizada, Equação 2.2 e Figura 3b, perde a dimensão original dos dados. Este problema é corrigido com a multiplicação da matriz $\mathbf{H}^\top$. Por $\mathbf{H}^F$ ser ortogonal, as configurações Helmertizadas são conectadas às configurações centradas pela seguinte propriedade da matriz de Helmert:

$$\mathbf{H}^\top_{\mathbf{(k \times k-1)}} \mathbf{H}_{\mathbf{(k-1 \times k)}} = (\mathbf{I}_l - \frac{1}{l} 1_l 1_l^\top) = \mathbf{C}_{\mathbf{(k \times k)}},$$

em que $\mathbf{I}_l$ é uma matriz identidade $l \times l$ e $\mathbf{C}$, uma matriz de centralização $k \times k$.

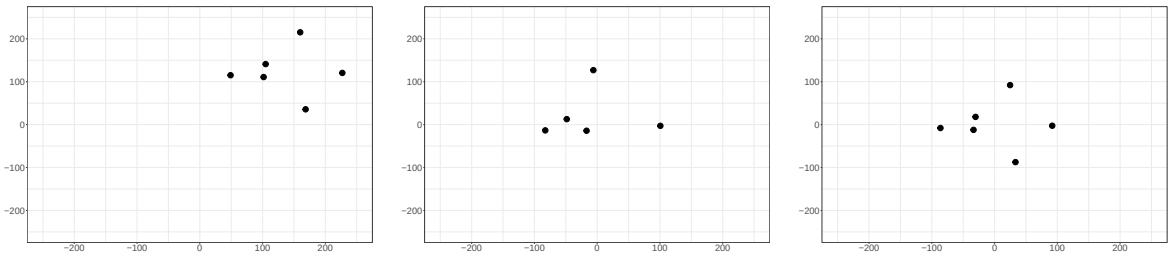
Dessa forma, a matriz de configuração centrada é dada pela Equação 2.3.

$$\mathbf{Y}_{\mathbf{C(k \times 2)}} = \mathbf{H}^\top \mathbf{H} \mathbf{Y} = \mathbf{C}_{\mathbf{(k \times k)}} \mathbf{Y}_{\mathbf{(k \times 2)}}. \tag{2.3}$$

[Dryden e Mardia 2016]

Um exemplo do que acontece com a configuração original, configuração Helmertizada e configuração centrada segue na Figura 3.

Figura 3 – Representação das configurações das vértebras de camundongos baseado em marcos.



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

Fonte: Autor

Depois da remoção do efeito de locação, o efeito de rotação deve ser removido para gerar os dados de tamanho e forma.

A rotação de uma configuração sobre a origem é dada pela pós-multiplicação da matriz de configuração $\mathbf{Y}$ ou $\mathbf{Y_C}$, ambas com dimensões $k \times m$, por uma matriz de rotação $\mathbf{\Gamma}$, com dimensão $m \times m$.

Definição 1 *Uma matriz de rotação* $\mathbf{\Gamma}$ *satisfaz* $\mathbf{\Gamma}^\top\mathbf{\Gamma} = \mathbf{\Gamma}\mathbf{\Gamma}^\top = \mathbf{I}_m$ *e* $\det(\mathbf{\Gamma}) = +1$. *Em outras palavras, uma matriz de rotação é matriz ortogonal especial, que é uma matriz ortogonal com determinante* $+1$. *O conjunto de todas as matrizes de rotação* $\mathbf{\Gamma}$ *é conhecido como grupo ortogonal especial* $SO(m)$.

Uma matriz de rotação tem $\frac{1}{2}m(m-1)$ graus de liberdade. Para $m = 2$ dimensões, a matriz de rotação pode ser parametrizada por um único ângulo θ, $-\pi \leq \theta \leq \pi$ em radianos, para girar no sentido horário sobre a origem.

$$\mathbf{\Gamma} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Como falamos anteriormente, os dados de tamanho e forma de uma matriz de configuração $\mathbf{Y}$, são todas as informações geométricas sobre $\mathbf{Y}$ que são invariáveis sob locação e rotação, e isso pode ser representado pelo conjunto $[\mathbf{Y_C}]$ dado por:

$$[\mathbf{Y_C}] = \{\mathbf{CY\Gamma} : \mathbf{\Gamma} \in SO(m)\}$$

em que $SO(m)$ é apresentado na Definição 1. [Dryden e Mardia 2016]

2.3 DISTÂNCIA NO ESPAÇO DE TAMANHO E FORMA

Nesta seção, mostra-se um dos conceitos básicos fundamentais para a análise estatística de tamanho e forma, o cálculo da distância. Em seu livro, Dryden e Mardia 2016 mostram a medida de distância para o espaço de tamanho e forma, conhecida como distância de Procrustes no espaço de tamanho e forma ou distância intrínseca no espaço de tamanho e forma.

Considere duas configurações de k-pontos em m dimensões, $\mathbf{Y}_1, \mathbf{Y}_2 \in R^{km}(k \times m)$. A distância de Procrustes no espaço de tamanho e forma é dada por:

$$d_S(\mathbf{Y}_1, \mathbf{Y}_2) = \sqrt{S_1^2 + S_2^2 - 2S_1S_2\cos\rho(\mathbf{Y}_1, \mathbf{Y}_2)}, \tag{2.4}$$

em que S_1, S_2 são os tamanhos dos centroides das configurações $\mathbf{Y}_1$, $\mathbf{Y}_2$, e ρ é a distância Riemanniana no espaço de forma, dada pela Equação 2.5:

$$\rho = \rho(\mathbf{Y_1}, \mathbf{Y_2}) = \arccos\left(|\mathbf{Z}^*_{\mathbf{Y_1}}\mathbf{Z}_{\mathbf{Y_2}}|\right), \quad 0 \leq \rho \leq \pi/2, \tag{2.5}$$

em que $\mathbf{Z}_{\mathbf{Y_1}} = \mathbf{HY_1}/||\mathbf{HY_1}||$ e $\mathbf{Z}_{\mathbf{Y_2}} = \mathbf{HY_2}/||\mathbf{HY_2}||$ representam as pré-formas de $\mathbf{Y}_1$ e $\mathbf{Y}_2$, respectivamente. O espaço de pré-formas é uma esfera complexa de raio unitário. Dessa forma, ρ representa o angulo entre as pré-formas.

Para um melhor entendimento sobre essa distância no espaço de tamanho e forma, uma prova é demonstrada por Dryden e Mardia 2016. Porém, a distância no espaço de tamanho e forma foi derivada por Le 1988.

Prova 1 *Inicialmente, considere duas configurações de k-pontos em m dimensões, $\mathbf{Y}_1^0, \mathbf{Y}_2^0 \in R^{km}(k \times m)$. Remove-se a locação pré-multiplicando a submatriz de Helmert pelas configurações para se obter as coordenadas Helmertizadas, $\mathbf{Y}_1 = \mathbf{HY}_1^0$, $\mathbf{Y}_2 = \mathbf{HY}_2^0$.*

Para os tamanhos dos centroides temos: $S_1 = ||\mathbf{Y}_1|| = ||\mathbf{CY}_1^0||$ e $S_2 = ||\mathbf{Y}_2|| = ||\mathbf{CY}_2^0||$. A distância entre o tamanho e a forma das configurações é encontrada minimizando a distância euclidiana sobre as rotações, ou seja:

$$\begin{aligned} d_S^2(\mathbf{Y}_1^0, \mathbf{Y}_2^0) &= \inf_{\mathbf{\Gamma} \in SO(m)} ||\mathbf{Y}_2 - \mathbf{Y}_1\mathbf{\Gamma}||^2 \\ &= tr(\mathbf{Y}_1^\top \mathbf{Y}_1) + tr(\mathbf{Y}_2^\top \mathbf{Y}_2) - 2 \sup_{\mathbf{\Gamma} \in SO(m)} tr(\mathbf{Y}_2^\top \mathbf{Y}_1 \mathbf{\Gamma}) \\ &= S_1^2 + S_2^2 - 2S_1S_2 \sup_{\mathbf{\Gamma} \in SO(m)} tr\left(\frac{\mathbf{Y}_2^\top}{S_2}\frac{\mathbf{Y}_1\mathbf{\Gamma}}{S_1}\right). \end{aligned}$$

em que $\inf$, *denota o ínfimo,* $\sup$, *o supremo e $tr(\cdot)$ o traço da matriz.*

Usando um dos resultados do espaço das pré-formas, consideraremos como verdadeira a Equação 2.6, portanto, o resultado segue.

$$\sup_{\mathbf{\Gamma} \in SO(m)} tr\left(\frac{\mathbf{Y}_2^\top}{S_2}\frac{\mathbf{Y}_1\mathbf{\Gamma}}{S_1}\right) = \cos\rho(\mathbf{Y}_1, \mathbf{Y}_2). \tag{2.6}$$

[Le 1995]

2.4 MÉDIA NO ESPAÇO DE TAMANHO E FORMA

Nesta seção, mostra-se um dos conceitos básicos para amostras aleatórias de objetos de tamanho e forma. A obtenção da estimativa de tamanho e forma média dos objetos é um

importante conceito relacionado com a análise dos dados em análise estatística de tamanho e forma.

"A estrutura de variabilidade de formato e tamanho é a principal preocupação na maioria das aplicações. No entanto, o conceito de tamanho e forma médios também tem um papel fundamental para tal análise". [Bookstein 1986].

Consideramos uma situação em que uma população de tamanho e forma pode ser modelada por uma distribuição de probabilidade. As primeiras definições de média para dados de tamanho e forma foram dadas por Fréchet 1948 e Grove e Karcher 1973.

Definição 2 *Dada uma amostra aleatória de tamanho e forma,* $\mathbf{Y_1}, \mathbf{Y_2}, ..., \mathbf{Y_n}$*, a média, conhecida também como média de Fréchet ou Karcher é definida como:*

$$\hat{\mu}_s = \arg\inf_{\mu} \frac{1}{n} \sum_{i=1}^{n} d_S^2(\mathbf{Y_i}, \mu) \tag{2.7}$$

em que d_S *é a distância de tamanho e forma dada na Equação 2.4, e* μ*, nesta situação, é uma configuração* $\mathbf{Y}$ *escolhida e proposta com média inicial,* $E[\mathbf{Y}] = \mu$*.*

2.5 MODELANDO DADOS NO ESPAÇO DE TAMANHO E FORMA

A distribuição normal complexa circularmente simétrica (central) ou simplesmente, distribuição normal complexa central, é uma das principais distribuições para análise estatística de tamanho e forma para marcos bidimensional (formas planas) por ser invariante sob rotações.

A distribuição normal complexa central corresponde ao caso em que a média é zero, sendo totalmente especificada pela matriz de covariância. A distribuição normal complexa central é denotada como:

$$\mathbf{Z} \sim CN(0, \mathbf{\Sigma}).$$

Se, $\mathbf{Z} = \mathbf{X} + i\mathbf{Y}$ tem uma distribuição normal complexa central, então o vetor $[\mathbf{X}, \mathbf{Y}]$ tem uma distribuição normal multivariada com estrutura de covariância

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N\left(\begin{bmatrix} \text{Re } \mu \\ \text{Im } \mu \end{bmatrix}, \frac{1}{2} \begin{bmatrix} \text{Re } \Sigma & -\text{Im } \Sigma \\ \text{Im } \Sigma & \text{Re } \Sigma \end{bmatrix} \right),$$

em que, $\mu = E[\mathbf{Z}] = 0$. A função densidade de probabilidade é dada por:

$$f_{\mathbf{Z}}(z) = \frac{1}{\pi^n \det(\mathbf{\Sigma})} \exp\{-(z-\mu)^\top \mathbf{\Sigma}^{-1}(z-\mu)\}.$$

[Andersen et al. 1995]

Para simular a distribuição normal complexa central, usamos o método proposto por Nascimento et al. 2016. O algoritmo 1 mostra o passo a passo de como são gerados os dados de tamanho e forma baseados na distribuição normal complexa central.

Algoritmo 1: Simulação da distribuição normal complexa central.

1 Considere uma matriz definida positiva Hermitiana de ordem p, sendo ela, definida como $\mathbf{\Sigma} = \mathbf{R} + i\mathbf{I}$ em que $(i = \sqrt{-1})$ e sejam $\mathbf{R}$ e $\mathbf{I}$ partes reais e imaginárias, respectivamente;

2 Gere k normais multivariadas de ordem 6, $[y_1, ..., y_6]^\top \sim N_6(\mathbf{0}, \mathbf{\Sigma})$, em que

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{R} & -\mathbf{I} \\ \mathbf{I} & \mathbf{R} \end{bmatrix}.$$

3 Os k dados de tamanho e forma, são gerados a partir da equação $W_k = [y_1\, y_2\, y_3]^\top + i\,[y_4\, y_5\, y_6]^\top$

3 CLASSIFICAÇÃO NÃO SUPERVISIONADA

Nos dias atuais, a ampla quantidade de informações que são disponibilizadas devem ser classificadas e agrupadas para que se tenha transformações em nosso cotidiano. Na história da humanidade, temos diversas situações em que houve a necessidade de classificar e agrupar, seja para sobrevivência ou compreensão de fenômenos.

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. A análise de agrupamento é uma das técnicas mais conhecidas e populares, assim, existem muitos modelos de agrupamento que analisam distribuições, densidades e possíveis pontos centrais em um conjunto de dados.

O conceito classificar e agrupar, também são conhecidos como sistemas de classificação, e os mesmos podem ser divididos em supervisionados ou não-supervisionados. Na classificação supervisionada, o mapeamento de um conjunto de dados de entrada é feito para um conjunto finito de classes rotuladas discretas. Por outro lado, na classificação não supervisionada, chamada de agrupamento, não há dados rotulados disponíveis. Sob certo ponto de vista, os 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. A ideia do agrupamento é de separar um conjunto de dados em um número de grupos (*clusters*), com base em alguma medida de similaridade, tal que, os itens em um dado grupo tenham um alto grau de similaridade (semelhança), enquanto itens pertencentes a grupos diferentes, tenham um alto grau de dissimilaridade (diferenças) em relação aos outros grupos.

Algoritmos de agrupamento por particionamento encontram uma partição que maximiza ou minimiza algum critério numérico. Esta dissertação apresentará alguns algoritmos de agrupamento adaptados para os dados de tamanho e forma. Antes de definir os algoritmos, precisamos mostrar como uma partição inicial é definida. O algoritmo 2, mostra o passo a passo de como se obter uma partição inicial, a mesma é necessária para montar protótipos dos grupos para que possamos iniciar os algoritmos propostos.

Algoritmo 2: Partição inicial

1 Seja $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ amostras de dados de tamanho e forma, selecione um número K de grupos.

2 Escolha K centroides aleatórios para geração inicial dos grupos

3 Aloque as amostras $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ aos centroides mais próximos de acordo com a distância da Equação 2.4.

4 Calcule a medida de dissimilaridade dos grupos gerados e guarde.

5 Repita $T = 10.000$ vezes o passo 2, 3 e 4.

6 Escolha os centroides com menor valor na medida de dissimilaridade.

3.1 *HILL CLIMBING*

O *hill climbing* é uma técnica de otimização matemática que pertence à família dos algoritmos de busca local. O algoritmo *hill climbing* é processo iterativo que começa com uma solução arbitrária para um problema, e em seguida, tenta encontrar uma solução melhor, fazendo uma mudança incremental na solução. Se a mudança produzir uma solução melhor, outra mudança incremental será feita na nova solução e assim por diante, até que nenhuma outra melhoria possa ser encontrada. Em outras palavras, o método *hill climbing* usa uma partição inicial e depois calcula a variação nos valores dos critérios de agrupamento no movimento do objeto i, do seu atual grupo k, para um novo grupo l (l diferente de k). [Landau et al. 2011]

Uma descrição do algoritmo *hill climbing* e do critério de agrupamento utilizado, foi introduzido por Landau et al. 2011 em seu livro. Nele, o algoritmo *hill climbing* é baseado na soma dos quadrados totais (medida de dissimilaridade). Esse critério, escolhe a partição correspondente ao valor mínimo da soma dos quadrados dentro do grupo, ou, equivalentemente, ao valor máximo da soma dos quadrados entre os grupos. Essencialmente, determina-se a medida de dissimilaridade baseado na soma dos quadrados totais pela Equação 3.1:

$$J = \sum_{m=1}^{\pi_m} \sum_{n=1}^{n_m} \mathsf{dist}(Y_{mn}, \mu_m) \tag{3.1}$$

em que dist é uma distância qualquer, Y_{mn} é a n-ésima observação do grupo m e μ_m é a média do grupo m.

Ao trabalharmos com dados de tamanho e forma, a distância, normalmente Euclidiana, é

substituída pela distância de Procrustes no espaço de tamanho e forma, assim, resultando na Equação 3.2:

$$J = \sum_{m=1}^{\pi_m} \sum_{n=1}^{n_m} d_S(Y_{mn}, \mu_m) \tag{3.2}$$

em que, d_S é definido pela Equação 2.4.

Após definida a partição inicial e medida de dissimilaridade, o algoritmo *hill climbing* modificado para análise estatística de tamanho e forma é dado por meio do algoritmo 3:

Algoritmo 3: *Hill climbing* modificado para análise estatística de tamanho e forma.

Passo 1: Dados de Entrada

1.1 Seja $\Omega = 1, ..., n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}$.

1.2 Calcule os dados de tamanho e forma $\mathbf{Y_C}$ definido na Equação 2.3.

1.3 Escolha K centroides e gere uma partição inicial π_K

1.4 Fixe o número de iterações $T = 100$ e um erro $\varepsilon > 0$; Faça $t = 1$.

1.5 Calcule a medida de dissimilaridade J_t dos grupos π_K, definido na Equação 3.2.

Passo 2: Atribuindo os objetos aos grupos

2.1 Calcule a medida de dissimilaridade J_{t+1} em relação a média μ_K, no movimento do objeto i do seu atual grupo π_K para um novo grupo π_L (L diferente K).

Passo 3: Critério de Parada

3.1 Se $||J_{t+1} - J_t|| < \varepsilon$ ou $t > T$, então **Pare**. Caso contrário, faça $t = t + 1$ e volte para o passo 2.

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

3.2 K-MÉDIAS

Na mineração de dados, o algoritmo ou agrupamento K-médias é um método de segregar diversos dados em torno de centroides ($cluster$), particionando n observações dentre K grupos, em que cada observação pertence ao grupo mais próximo. Em outras palavras, o al-

goritmo K-médias tenta minimizar as variâncias dentro do agrupamento, usando como base uma distância da observação para a média.

Uma descrição do algoritmo K-médias foi introduzido por Macqueen 1967. Nele, o algoritmo K-médias usa o valor médio dos objetos dos grupos criados através da partição inicial, como protótipo para a geração da partição final dos grupo. Como foi mencionado anteriormente, nossas adaptações são para dados de tamanho e forma, com isso, usaremos a medida de distância da Equação 2.4 e o critério de agrupamento da Equação 3.2.

Dessa forma, o algoritmo K-médias modificado para análise estatística de tamanho e forma é dado por meio do algoritmo 4:

Algoritmo 4: K-médias modificado para análise estatística de tamanho e forma.

Passo 1: Dados de Entrada

1.1 Seja $\Omega = 1, ..., n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}$.

1.2 Calcule os dados de tamanho e forma $\mathbf{Y_C}$ definido na Equação 2.3.

1.3 Escolha K centroides e gere uma partição inicial π_K

1.4 Fixe o número de iterações $T = 100$ e um erro $\varepsilon > 0$; Faça $t = 1$.

1.5 Calcule as médias μ_{K_t} dos grupos π_{K_t}, definido na Equação 2.7.

Passo 2: Atribuindo os objetos aos grupos

2.1 Calcule a distância de cada objeto em relação a média μ_{K_t} e atribua esse objeto ao novo grupo $\pi_{K_{t+1}}$ mais próximo.

2.1 Calcule as médias dos novos grupos $\pi_{K_{t+1}}$ formados.

Passo 3: Critério de Parada

3.1 Se $||\pi_{K_{t+1}} - \pi_{K_t}|| < \varepsilon$ ou $t > T$, então **Pare**. Caso contrário, faça $t = t + 1$ e volte para o passo 2.

O algoritmo K-médias converge quando não há mais mudanças significativas na média dos grupos 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 TESTES DE HIPÓTESES PARA COMPARAR O TAMANHO E FORMA MÉDIA DE DUAS POPULAÇÕES

A fim de melhorar a qualidade do agrupamento, propomos uma adaptação no algoritmo *hill climbing*, baseando-se em testes de hipóteses.

Nós consideramos dois testes de hipóteses que é mostrado por Dryden e Mardia 2016 em seu livro. O primeiro é o teste T^2 de Hotelling, e o outro, o teste de Goodall. O primeiro é menos restrito que o segundo, porém mais complexo. O teste de Goodall supõe que a distribuição conjunta no espaço de marcos anatômicos é uma normal complexa e isotrópica, o que significa que a variância para cada marco é a mesma. Por outro lado, o teste T^2 de Hotelling supõe normalidade para as observações e a isotropia não é assumida. Nos dois testes, as hipóteses são apresentadas pela Equação 3.3:

$$H_0 : [\mu_1] = [\mu_2] \qquad vs. \qquad H_1 : [\mu_1] \neq [\mu_2] \tag{3.3}$$

em que μ_i representa a média do grupo i.

Vale salientar que os testes abaixo são apresentados na literatura para o espaço de formas, no entanto, Dryden e Mardia 2016 menciona que os mesmos podem ser usados no espaço de tamanho e forma, com uma pequena mudança nos graus de liberdade, $q = (k-1)m - m(m-1)/2$.

3.3.1 Teste de Hotelling

Antes de explicarmos o teste de Hotelling, precisamos entender um pouco sobre espaço tangente.

O espaço tangente é a versão linearizada do espaço de formas na proximidade de um ponto particular no espaço trabalhado. Uma das vantagens do espaço tangente, é que podemos utilizar as técnicas padrões de multivariada diretamente. Existem vários tipos diferentes de coordenadas no espaço tangente, aqui, usaremos as coordenadas tangente Procrustes parcial.

Seja $\mathbf{Y} = \{\mathbf{Y_1}, ..., \mathbf{Y_{n_1}}\}$ uma amostra de configurações complexas, as coordenadas tangentes para essa amostra poderá ser calculada como

$$\mathbf{t}_i = e^{i\hat{\theta}}[\mathbf{I}_{k-1\times k-1} - \hat{\mu}\hat{\mu}^*]\mathbf{z}_i \tag{3.4}$$

em que $\mathbf{z}_i = \mathbf{HY_i}/||\mathbf{HY_i}||$ são as pré-forma da amostra aleatória de $\mathbf{Y}$, $\mathbf{I}$ é uma matriz identidade, $\hat{\theta}$ minimiza $||\hat{\mu} - \mathbf{z}e^{i\hat{\theta}}||^2$, e a forma média $\hat{\mu}$ é o autovetor associado ao maior autovalor da matriz produto $\mathbf{z}$.

Suponha agora que $\mathbf{z}_1, ..., \mathbf{z}_n$ é uma amostra aleatória de pré-formas e $\mathbf{t}_1, ..., \mathbf{t}_n$ suas coordenadas tangentes. Seja, $\mathbf{v}_i$ um vetor $2k-2$ no qual, o que é obtido por empilhar as coordenadas real e imaginaria de $\mathbf{t}_i$. Essa operação é definida por $cvec$, em que:

$$\mathbf{v}_i = cvec(\mathbf{ti}) = (\mathrm{Re}(\mathbf{t}_i)^\top, \mathrm{Im}(\mathbf{t}_i)^\top)^\top. \tag{3.5}$$

A partir disso, o teste de Hotelling segue.

Seja $\mathbf{X} = \{\mathbf{X_1}, ..., \mathbf{X_{n_1}}\}$ e $\mathbf{Y} = \{\mathbf{Y_1}, ..., \mathbf{Y_{n_2}}\}$ amostras aleatórias de configurações complexas independentes, e seja, $\{\mathbf{z}_1^X, ..., \mathbf{z}_{n_1}^X\}$ e $\{\mathbf{z}_1^Y, ..., \mathbf{z}_{n_2}^Y\}$ suas pré-formas, respectivamente.

As coordenadas tangentes $\{\mathbf{v}_1, ..., \mathbf{v}_{n_1}\}$ e $\{\mathbf{w}_1, ..., \mathbf{w}_{n_2}\}$ são obtidas a partir da pré-forma média $\hat{\mu}$, calculada a partir das duas amostras combinadas, isto é, $\{\mathbf{z}_1^X, ..., \mathbf{z}_{n_1}^X, \mathbf{z}_1^Y, ..., \mathbf{z}_{n_2}^Y\}$ com $n_1 + n_2$ observações.

Com isso, é proposto um modelo normal multivariado no espaço tangente, em que

$$\mathbf{v}_i \sim N(\xi_1, \hat{\Sigma}_1), \quad \mathbf{w}_l \sim N(\xi_2, \hat{\Sigma}_2), \;\; i = 1, ..., n_1 \;\; \text{e} \;\; l = 1, ..., n_2. \tag{3.6}$$

Supõe-se que as matrizes de covariância sejam iguais para $\mathbf{v}_i$ e $\mathbf{w}_l$ ($\hat{\Sigma}_1 = \hat{\Sigma}_2$). Além das médias $\bar{\mathbf{v}}$ e $\bar{\mathbf{w}}$ serem definidas por:

$$\bar{\mathbf{v}} = \frac{1}{n}\sum_{i=1}^{n_1} \mathbf{v}_i \qquad \text{e} \qquad \bar{\mathbf{w}} = \frac{1}{n}\sum_{l=1}^{n_2} \mathbf{w}_i.$$

Com isso, a distância de Mahalanobis entre as médias $\bar{\mathbf{v}}$ e $\bar{\mathbf{w}}$, é dada por:

$$D^2 = D(\bar{\mathbf{v}}, \bar{\mathbf{w}}) = (\bar{\mathbf{v}} - \bar{\mathbf{w}})^\top \hat{\Sigma}^{-1} (\bar{\mathbf{v}}, \bar{\mathbf{w}}), \tag{3.7}$$

em que $\hat{\Sigma}^{-1}$ é a inversa generalizada de Moore-Penrose de $\hat{\Sigma} = \frac{n_1\hat{\Sigma}_1 + n_2\hat{\Sigma}_2}{n_1+n_2-2}$.

Para que seja dito que uma matriz seja inversa generalizada de Moore-Penrose, a mesma deve satisfazer as seguintes condições: Seja $\mathbf{U}$ uma matriz qualquer, temos:

$$\mathbf{UKU} = \mathbf{U}$$

$$\mathbf{KUK} = \mathbf{K}$$

$$(\mathbf{KU})^\top = \mathbf{KU}$$

$$(\mathbf{UK})^\top = \mathbf{UK}$$

Estas são as chamadas condições de Penrose e $\mathbf{K}$ é denominada inversa generalizada de Moore-Penrose. [Searle 1984].

Sob H_0, tem-se que $\mu_1 = \mu_2$. Com isso, a estatística de teste de Hotelling é definida:

$$F = \frac{n_1 n_2 (n_1 + n_2 - q - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)q} D^2 \sim F_{q,n_1+n_2-q-1}$$

em que o parâmetro q no espaço de tamanho e forma é definido por $q = (k-1)m - m(m-1)/2$. Logo, rejeitamos a hipótese H_0 se $P(F_{q,n_1+n_2-q-1} \geq F) \leq \alpha$. [Hotelling 1992]

3.3.2 Teste de Goodall

Considere duas amostras aleatórias de configurações $\mathbf{X} = \{\mathbf{X_1}, ..., \mathbf{X_{n_1}}\}$ e $\mathbf{Y} = \{\mathbf{Y_1}, ..., \mathbf{Y_{n_2}}\}$. Supõe-se que ambas as populações tem variância comum σ^2 para cada coordenada (isotropia). Sob H_0, e com σ pequeno, as distâncias de Procrustes completas são aproximadamente distribuídas por:

$$\begin{aligned}
\sum_{i=1}^{n_1} d_F^2(X_i, \hat{\mu}_1) &\sim \tau_0^2 \chi^2_{(n_1-1)q} \\
\sum_{i=1}^{n_2} d_F^2(Y_i, \hat{\mu}_2) &\sim \tau_0^2 \chi^2_{(n_2-1)q} \\
d_F^2(\hat{\mu}_1, \hat{\mu}_2) &\sim \tau_0^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \chi^2_q
\end{aligned}$$

em que $\tau_0 = \sigma/||\mu_1||$, ρ é a distância Riemanniana definida na Equação 2.5 e $d_F(\mathbf{X}, \mathbf{Y})$ é a distância de Procrustes completa definida por:

$$d_F^2(\mathbf{X}, \mathbf{Y}) = \sin \rho.$$

Com isso, a estatística do teste de Goodall é dado por:

$$F = \frac{n_1 + n_2 - 2}{n_1^{-1} + n_2^{-1}} \frac{d_F^2(\hat{\mu}_1, \hat{\mu}_2)}{\sum_{i=1}^{n_1} d_F^2(X_i, \hat{\mu}_1) + \sum_{i=1}^{n_2} d_F^2(Y_i, \hat{\mu}_2)} \sim F_{q,(n_1+n_2-2)q}$$

em que, novamente, o parâmetro q no espaço de tamanho e forma é definido por $q = (k-1)m - m(m-1)/2$. Logo, rejeitamos a hipótese H_0 se $P(F_{q,(n_1+n_2-2)q} \geq F) \leq \alpha$. [Goodall 1991].

Após a explicação dos dois testes, propomos um novo algoritmo baseado no algoritmo *hill climbing*. Nele, tentamos minimizar o valor-p dos testes no movimento do objeto i do seu atual grupo π_K, para um novo grupo π_L (L diferente K). Dessa forma, as medidas de

dissimilaridade usadas, serão baseadas nos valores-p dos testes de Hotteling e Goodall. Essas duas novas medidas de dissimilaridade são definidas como:

$$J_1 = \text{Hotelling}_{\text{pvalor}},$$
$$J_2 = \text{Goodall}_{\text{pvalor}}.$$

Logo, o algoritmo baseado em teste de hipóteses é dado pelo algoritmo 5:

Algoritmo 5: *Hill climbing* baseado em testes de hipóteses.

Passo 1: Dados de Entrada

1.1 Seja $\Omega = 1, ..., n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}$.

1.2 Calcule os dados de tamanho e forma $\mathbf{Y_C}$ de acordo com a Equação 2.3.

1.3 Escolha aleatoriamente K centroides distintos e gere uma partição inicial π_K

1.4 Fixe o número de iterações T e um erro $\varepsilon > 0$; Faça $t = 1$.

1.5 Calcule as médias μ_k dos grupos π_K.

Passo 2: Atribuindo os objetos aos grupos

2.1 Calcule o valor-p do teste J_1 ou J_2 em relação a média μ, no movimento do objeto i do seu atual grupo π_K para um novo grupo π_L (L diferente K).

Passo 3: Critério de Parada

3.1 Se $||p\text{valor}_{t+1} - p\text{valor}_t|| < \varepsilon$ ou $t > T$, então **Pare**. Caso contrário, faça $t = t + 1$ e volte para o passo 2.

Assim como o algoritmo *hill climbing*, o algoritmo baseado em testes de hipóteses, converge quando não há mais mudanças significativas nos valores-p ou quando o número máximo de iterações predefinido é atingido.

4 MÉTODOS *ENSEMBLES*

"Métodos de aprendizagem de *ensemble* ou sistemas de classificadores múltiplos, têm sido cada vez mais importantes, porque têm mostrado repetidamente a capacidade de melhorar a precisão dos algoritmos". [Chawla et al. 2002].

Os *ensembles* podem ser divididos em duas categorias: 1) homogêneos: quando cada classificador básico usa o mesmo algoritmo de classificação e 2) heterogêneos: quando classificadores básicos usam algoritmos de classificação diferentes. Nesta dissertação, todos os algoritmos de classificação por ensemble se enquadram na categoria homogênea. Além disso, as técnicas de aprendizagem por *ensemble* podem ser caracterizadas como métodos de *ensemble* paralelos e sequenciais. As técnicas de *ensemble* sequencial, como *boosting*, geram classificadores de maneira sequencial, enquanto os *ensembles* paralelos, como *bagging*, gera classificadores em paralelo. [Liu e Zhou 2013, Tahir et al. 2010]

Os *ensembles* são baseados na noção de combinar vários classificadores básicos, de forma que o classificador final *ensemble*, obtenha um desempenho melhor do que cada classificador individual.

Para conhecermos mais profundamente os algoritmos *boosting* e *bagging*, devemos entender o conceito 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 [Gareth et al. 2013].

4.1 *BOOSTING*

Boosting é uma técnica de conjunto popular, que melhora o desempenho de classificação de classificadores base de forma iterativa, atribuindo um peso inicial igual a cada instância $(1/n)$, antes de construir um modelo de classificação, e posteriormente, ajustando os pesos de cada instância para a próxima iteração. A ideia geral do *boosting*, é atribuir pesos às instâncias de

modo que esses pesos sejam alterados de acordo com a classificação correta dos dados. Assim, as instâncias com pesos mais altos (instâncias previamente classificadas incorretamente) são mais prováveis de estarem presentes, enquanto as instâncias com pesos mais baixos (instâncias classificadas corretamente na iteração anterior) são menos prováveis de estarem presentes nos novos dados de treinamento. Além disso, alguns casos podem aparecer várias vezes, enquanto alguns podem não aparecer com base nas distribuições de peso. Em seguida, um classificador base é treinado no conjunto de dados de treinamento recém-derivado e o erro associado ao modelo é avaliado. Finalmente, os modelos de classificação de cada iteração são combinados em uma votação ponderada para montar o classificador final. Por atribuir pesos as observações a cada iteração, o *boosting* na maioria das vezes se sai melhor com a presença de *outliers* no dados [Freund, Schapire et al. 1996, Seiffert et al. 2008, Guile e Wang 2010].

Ao longo da literatura, vários algoritmos de *boosting* foram propostos e desenvolvidos por pesquisadores, sendo a maioria deles, para classificação supervisionada. Frossyniotis, Likas e Stafylopatis 2004 desenvolveu um algoritmo *boosting* para classificação não supervisionada, chamado de *boost-clustering*. Com isso, o algoritmo *boost-clustering* gera replicas *bootstrap* de mesmo tamanho dos dados originais, e os pesos para cada instância são modificados a partir da qualidade do agrupamento. Modificamos tal algoritmo para ser usado para dados de tamanho e forma. O algoritmo *boost-clutering* para dados de tamanho e forma é dado pelo algoritmo 6.

Algoritmo 6: *boost-clustering* adaptado para dados de tamanho e forma.

Passo 1: Dados de Entrada

1.1 Seja $\Omega = 1, ..., n$ um conjunto de dados descrito pela matriz de configuração Y.

1.2 Calcule os dados de tamanho e forma $\mathbf{Y_S}$ de acordo com a Equação 2.3.

1.4 Fixe o número de iterações T, um erro $\varepsilon_{\max} = 0$, e o peso $p_i^1 = 1/n$, com $i = 1, ..., n$; Faça $t = 1$

Passo 2: Atribuindo os objetos aos grupos

2.1 Forme uma réplica *bootstrap* da matriz de configuração $\mathbf{Y_S}$ de acordo com o peso p_i^t.

2.2 Aplique o algoritmo de agrupamento para agrupar a amostra e forme a partição $H_i^t = (h_{i,1}^t, h_{i,2}^t, ..., h_{i,C}^t)$, em que $h_{i,C}$ é o grau de associação da instância i com o *cluster* j.

2.3 Se $t > 1$, selecione ou substitua o *cluster* de H^t de acordo com a pontuação correspondente mais alta, fornecidos pelo *boost-clustering* até agora, usando H_{aux}^{t-1}

2.4 Calcule um pseudo-erro ε_t:

$$\varepsilon_t = \frac{1}{2}\sum_{i=1}^{N} p_i^t(1 - h_{i,\max}^t - h_{i,\min}^t)$$

2.5 Se $\varepsilon_t < \varepsilon_{\max}$, faça $ct = ct + 1$. Se não, faça $ct = 0$ e $\varepsilon_{\max} = \varepsilon_t$.

2.5 Seja $\beta_t = \frac{1-\varepsilon_t}{\varepsilon_t}$.

2.6 Atualize os pesos:

$$p_i^{t+1} = \frac{p_i^t \beta_t^{1-h_{i,\max}^t - h_{i,\min}^t}}{Y}$$

em que Y é uma constante de normalização de modo que $\sum_{i=1}^{N} p_i^{t+1} = 1$

2.7 Calcule a partição auxiliar

$$H_{aux}^t = \arg\max_{k=1,...,C} = \sum_{\tau=1}^{t}\left[\frac{\log(\beta_t)}{\sum_{j=1}^{t}\log(\beta_j)} h_{i,k}^{\tau}\right]$$

Passo 3: Critério de Parada

3.1 Se $ct = 3$ ou $t >= T$ ou $\varepsilon_t > 0.5$, Pare. E a partição final é dada por $H^f = H^t$. Caso contrário, faça $t = t + 1$ e volte para o passo 2

4.2 *BAGGING*

Bagging, ou agregação de *bootstrap*, é um algoritmo de aprendizagem de conjunto, que agrega resultados de vários classificadores base, para melhorar o resultado da classificação final. O *bagging* desenvolvido por Breiman 1996, começa tomando uma amostra aleatória com substituição de mesmo tamanho (N instâncias) do conjunto de dados original. Em seguida, um modelo é treinado em cada conjunto de dados das amostras *bootstrap* geradas. Este processo de geração de replicações *bootstrap* e geração de modelos é repetido várias vezes, resultando em vários modelos de classificação. Finalmente, os resultados de todos esses modelos de classificação são agregados em uma decisão final.

O algoritmo *bagging* baseado em classificação não supervisionada para dados de tamanho e forma, é dado pelo algoritmo 7.

Algoritmo 7: *Bagging* adaptado para dados de tamanho e forma.

Passo 1: Dados de Entrada

1.1 Seja $\Omega = 1, ..., n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}$.

1.2 Calcule os dados de tamanho e forma $\mathbf{Y_C}$ de acordo com a Equação 2.3.

1.4 Fixe o número de iterações T e b além de um erro $\varepsilon > 0$ e fixe $t = 1$.

Passo 2: Atribuindo os objetos aos grupos

2.1 Forme b-ésimas amostras bootstrap para dados de tamanho e forma.

2.2 Aplique o algoritmo de agrupamento para agrupar amostras e guarde os rótulos dos grupos gerados.

Passo 3: Critério de Parada

3.1 Repita o passo 2 T vezes, e o rótulo final de cada observação é dado pela maioria dos rótulos gerados para essa observação.

5 MÉTODOS DE VALIDAÇÕES

A validade de um estudo refere-se a qução bem os resultados encontrados representam resultados verdadeiros para indivíduos semelhantes fora do estudo. Este conceito de validade, se aplica a todos os tipos de estudos, sejam eles clínicos, sobre prevalência, associações, intervenções e diagnóstico. A validade de um estudo de pesquisa inclui dois domínios: a validade interna e a validade externa.

Nesta seção, dois índices de validação adequados para análise de agrupamento são descritos. Esses índices são medidas para avaliar a qualidade dos resultados das partições encontradas para algum algoritmo de agrupamento. Além disso, é apresentado o método da acurácia, o mesmo verifica se as observações do agrupamento foram para o grupo correto, pois muitas vezes, as partições geradas pelos algoritmos são bem definidas, mas, o objeto agrupado não pertencia aquela partição.

5.1 VALIDAÇÃO INTERNA

A validade interna é definida como a extensão em que os resultados observados representam uma possível verdade para a população. Nesta dissertação, usaremos o índice residual de Procrustes, já que o mesmo é muito utilizado em análise estatística de forma. Porém, esse método é baseado no espaço de formas. Com isso, iremos verificar como o mesmo se comporta nos espaço de tamanho e forma, sendo assim, mais um objetivo dessa dissertação, verificar se para dados de tamanho e forma, o índice residual de Procrustes é um bom validador de agrupamento.

O Índice Residual Procrustes é útil para encontrar o número ideal de grupos, e também, é usando por muitos autores para avaliar a qualidade do agrupamento em um conjunto de dados.

Após obter a alocação dos indivíduos do conjunto de dados usando um dos algoritmos, calcula-se a norma quadrática dos resíduos de cada indivíduo dentro do seu grupo (r_{in}) e fora do seu grupo (r_{out}). Ou seja,

$$r_{in}(i) = \left[\mathbf{Y_{C_i}} - \left(\sum_{\mathbf{Y_C} \in C_r}^{n} \frac{\mathbf{Y_C}}{tam(C_r)} \right) \right]^2$$

$$r_{out}(i) = \min_{1,\ldots,K\ s\neq r} \left[\mathbf{Y_{C_i}} - \left(\sum_{\mathbf{Y_C} \in C_s}^{n} \frac{\mathbf{Y_C}}{tam(C_s)} \right) \right]^2$$

em que tam representa a quantidade de observações dentro do *cluster*, C_r é o *cluster* a qual o objeto i pertence, e C_s, é o *cluster* ao qual o objeto i não pertence.

Com isso, o índice residual Procrustes $pr(i)$ para cada indivíduo do conjunto de dados é dado como:

$$pr(i) = \frac{r_{out}(i) - r_{int}(i)}{\max(r_{out}(i), r_{int}(i))},$$

em que, $-1 \leq pr(i) \leq 1$.

Valores próximos de 1 indicam que o indivíduo i possui dissimilaridade menor dentro do grupo comparando com outro grupo, logo o indivíduo está agrupado apropriadamente. Valores negativos ou próximos de -1 indicam que o indivíduo i pode ter sido alocado no grupo errado. [Dryden e Mardia 2016].

Para se obter um índice de validação geral de Procrutes, é calculada a média aritmética de todos os $pr(i)$:

$$PR = \frac{1}{n} \sum_{i=1}^{n} pr(i).$$

A Tabela 1, da uma interpretação para os resultados do PR, baseada na tabela de interpretação do índice silhueta do livro de Izenman 2008.

Tabela 1 – Tabela de interpretação do índice de Procrustes

PR	**Interpretação**
$0.70 - 1$	Um agrupamento forte foi encontrado
$0.51 - 0.70$	Um agrupamento razoável foi encontrado
$0.26 - 0.50$	O agrupamento é fraco e pode ser artificial
≤ 0.25	Nenhum agrupamento foi encontrado

Fonte: [Izenman 2008]

5.2 VALIDAÇÃO EXTERNA

Feito por Rand 1971, o índice de Rand ou medida Rand em estatística, é uma medida da similaridade entre agrupamentos de dados. Do ponto de vista matemático, o índice de Rand está relacionado à precisão, mas é aplicável mesmo quando os rótulos de classe não são usados.

Denotado por R, o índice de Rand é calculado como:

$$R = \frac{a+b}{\begin{pmatrix} n \\ 2 \end{pmatrix}},$$

em que $0 \leq R \leq 1$ e :

- a: O número de vezes que um par de elementos pertence ao mesmo cluster em um método de agrupamento.
- b: O número de vezes que um par de elementos pertence a grupos de diferença em um método de agrupamento.
- $\begin{pmatrix} n \\ 2 \end{pmatrix}$: O número de pares não ordenados em um conjunto de n elementos.

Valores próximos de 1 indicam que o agrupamento está bem definido. Valores próximos de 0, indicam que o agrupamento pode ter sido definido de forma errada.

5.3 ACURÁCIA

Quando se fala em tecnologia atualmente, um termo que tem se tornado cada vez mais comum é o da acurácia. Usada muitas vezes quase como um sinônimo de precisão e eficiência. Em outras palavras, a acurácia serve para ver a porcentagem de acertos do algoritmo de classificação. O mesmo é definida como

$$Acc = \frac{acerto}{total = acerto + erro} \times 100,$$

em que o $acerto$ é definido como a alocação correta do objeto no agrupamento, e o $erro$, a alocação incorreta do objeto ao agrupamento.

6 EXPERIMENTOS E RESULTADOS

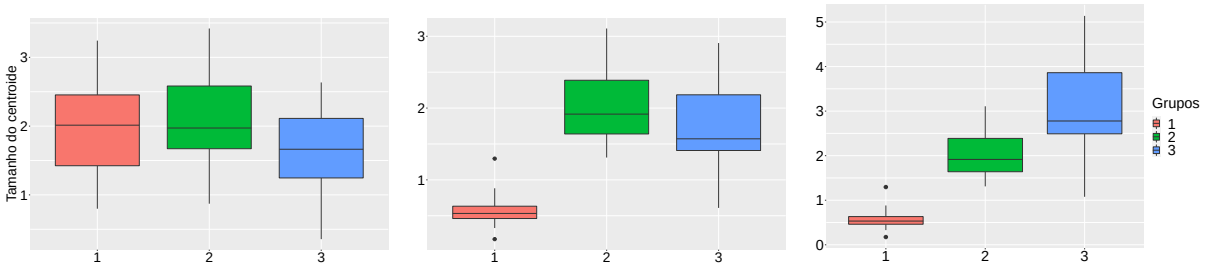
Após toda fundamentação teórica, aplicaremos os algoritmos modificados propostos em uma base de dados artificial, e com ela, iremos verificar como os algoritmos se comportam em cada um dos cenário propostos. Depois disso, aplicaremos os algoritmos a três bases de dados reais; a primeira para dados de vértebras torácicas T2 de camundongos; a segunda para dados de ressonância magnética de pessoas com esquizofrenia, e a terceira, para dados de referência de crânio de grandes macacos. Essas bases de dados podem ser encontradas no pacote shapes do *software* R Core Team 2021. Os nomes das bases no pacote são `mice`, `schizophrenia` e apes, respectivamente.

6.1 DADOS SIMULADOS

Para verificar o comportamento dos algoritmos, foram criados dados artificiais de tamanho e forma contendo três grupos. Nesses dados simulados, gerados a partir da distribuição normal complexa central, propomos três cenários possíveis: O primeiro para grupos bastante homogêneos[1]; o segundo para dois grupos homogêneos e um heterogêneo[2]; e o terceiro, para os três grupos heterogêneos. Para mais detalhes de como esses dados foram gerados, ver a seção 2.5.

São apresentados na Figura 4, os boxplots dos tamanhos dos centroides para cada cenário.

Figura 4 – Boxplots dos tamanhos dos centroides para os dados de cada cenário proposto.



(a) cenário 1. (b) cenário 2. (c) cenário 3.

Fonte: Autor

[1] Apresenta uma grande semelhanças na estrutura dos dados.
[2] Apresenta uma grande diferença na estrutura dos dados.

6.1.1 Cenário 1: grupos homogêneos.

No primeiro cenário, ao construirmos o tamanho dos centroides de cada grupo, Figura 4a, podemos ver o quanto eles são semelhantes. Neste cenário, todos os algoritmos tiveram bastantes dificuldades para realizar o agrupamento, sendo o algoritmo *boosting* K-médias o mais poderoso, segundo a validação interna e externa. Porém, quando tratamos da acurácia, o teste *bagging* K-médias se torna o mais poderoso, acertando 48,3% da classificação dos objetos em seus grupos.

Tabela 2 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados simulados do cenário 1.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,575	0,573	38,3%
Boosting: K-médias	0,595	0,609	46,7%
Bagging: K-médias	0,572	0,489	48,3%
Hill climbing	0,561	0,575	35,0%
Boosting: *hill climbing*	0,564	0,452	35,0%
Bagging: *hill climbing*	0,564	0,325	43,3%
Teste de Hipóteses: Hoteling	0,553	0,510	40,0%
Boosting: Hoteling	0,555	0,452	36,7%
Bagging: Hoteling	0,565	0,449	46,7%
Teste de Hipóteses: Goodall	0,555	0,552	40,0%
Boosting: Goodall	0,563	0,245	36,7%
Bagging: Goodall	0,558	0,423	36,7%

Fonte: Autor

Vale ressaltar, que neste cenário, não é possível afirmar se os métodos *ensembles* melhoraram os classificadores base. Todas as informações dos índices deste cenário, se encontram na Tabela 2.

6.1.2 Cenário 2: dois grupos homogêneos e um heterogêneo.

No segundo cenário, ao construirmos o tamanho dos centroides de cada grupo, Figura 4b, podemos ver que apenas um grupo difere dos demais. Nesse cenário, os algoritmos tiveram melhor desempenho se comparado ao primeiro cenário. Para o segundo cenário, o teste *bagging hill climbing* foi o mais poderoso, segundo seu índice externo e a acurácia. Para o índice interno,

o algoritmo mais poderoso foi o K-médias.

Tabela 3 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados simulados do cenário 2.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,683	0,631	63,3%
Boosting: K-médias	0,672	0,526	66,7%
Bagging: K-médias	0,704	0,596	65,0%
Hill climbing	0,724	0,611	68,3%
Boosting: *hill climbing*	0,703	0,588	66,7%
Bagging: *hill climbing*	0,737	0,560	73,3%
Teste de Hipótese: Hoteling	0,563	0,325	41,7%
Boosting: Hoteling	0,703	0,588	63,3%
Bagging: Hoteling	0,671	0,610	63,3%
Teste de Hipótese: Goodall	0,655	0,602	61,7%
Boosting: Goodall	0,654	0,475	56,7%
Bagging: Goodall	0,715	0,569	70,7%

Fonte: Autor

Neste cenário, podemos ver com mais clareza que pelo menos um dos métodos *ensembles* melhoraram os classificadores base, isso seguindo seus índices internos e sua acurácia. Todas as informações dos índices deste cenário, se encontram na Tabela 3.

6.1.3 Cenário 3: grupos heterogêneos.

No terceiro cenário, ao construirmos o tamanho dos centroides de cada grupo, Figura 4c, podemos ver que todos os grupos são diferentes. Nesse cenário, os algoritmos tiveram bom desempenho para realizar o agrupamento, sendo mais uma vez, o algoritmo *bagging hill climbing* o mais poderoso nesse cenário, segundo seus índices externos e acurácia. Para o índice interno, novamente o algoritmo mais poderoso foi o K-médias.

Tabela 4 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados simulados do cenário 3.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,699	0,597	50,0%
Boosting: K-médias	0,672	0,523	56,7%
Bagging: K-médias	0,745	0,537	70,0%
Hill climbing	0,725	0,592	53,3%
Boosting: *hill climbing*	0,735	0,431	66,7%
Bagging: *hill climbing*	0,781	0,478	78,3%
Teste de Hipótese: Hoteling	0,584	0,406	41,7%
Boosting: Hoteling	0,662	0,431	63,3%
Bagging: Hoteling	0,723	0,484	70,0%
Teste de Hipótese: Goodall	0,706	0,573	51,7%
Boosting: Goodall	0,692	0,490	58,3%
Bagging: Goodall	0,723	0,535	68,3%

Fonte: Autor

Neste cenário, podemos ver com mais clareza, que os métodos *ensembles* melhoraram os classificadores base, isso seguindo seus índices internos e sua acurácia. Todas as informações dos índices deste cenário, se encontram na Tabela 4.

Com isso, concluímos que à medida que o tamanho dos centroides dos grupos diferem, os algoritmos vão tendo mais facilidade em realizar o agrupamento. Além disso, os algoritmos base combinados com os métodos *ensembles*, sempre conseguem ter desempenho superior, isso seguindo os seus índices externos e acurácia.

6.2 DADOS REAIS

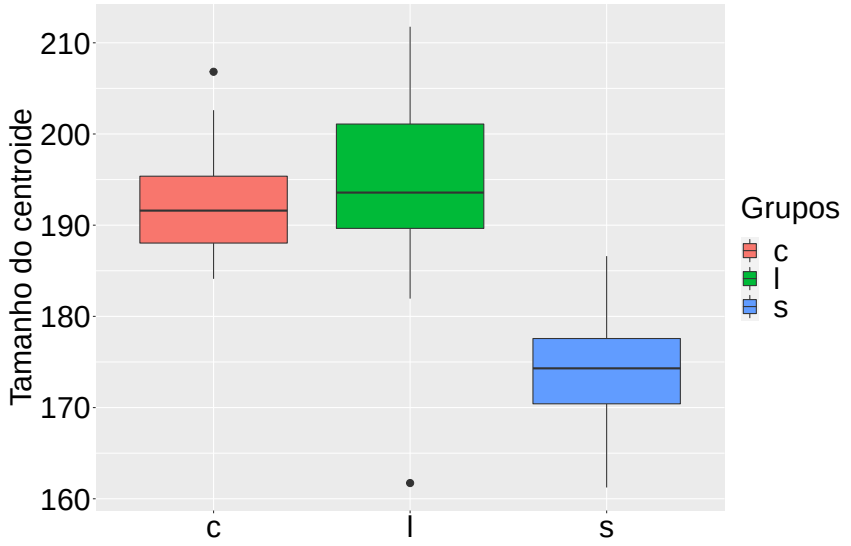
6.2.1 Vértebras torácicas T2 de camundongos.

A base de dados das vértebras torácicas T2 de camundongos possui 6 marcos em 2 dimensões, e possui um total de 76 indivíduos. Esses indivíduos foram classificados em 3 grupos: controle(**c**)=30, grandes(**l**)=23 e pequenos(**s**)=23. Os 6 pontos de referência foram obtidos usando um método semi-automático em pontos de alta curvatura, em que, é mostrado com mais detalhes em Dryden e Mardia 2016.

Para ter uma ideia inicial de como os algoritmos se comportarão, foi feito o boxplot do tamanho dos centroides de cada grupo. Pela Figura 5, podemos ver que apenas um grupo

difere dos demais, além da presença de *outliers* em dois grupos. Com isso, podemos supor a partir dos resultados dos dados simulados, que os algoritmos terão uma boa performance para realizar os agrupamentos.

Figura 5 – Boxplots dos tamanhos dos centroides para os dados das vértebras de camundongos.



Fonte: Autor

Tabela 5 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados das vértebras de camundongos.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,737	0,592	75,0%
Boosting: K-médias	0,746	0,599	75,0%
Bagging: K-médias	0,761	0,588	78,9%
Hill climbing	0,728	0,593	73,7%
Boosting: *hill climbing*	0,739	0,582	76,3%
Bagging: *hill climbing*	0,790	0,585	81,6%
Teste de Hipótese: Hoteling	0,730	0,576	73,7%
Boosting: Hoteling	0,648	0,582	61,8%
Bagging: Hoteling	0,756	0,581	77,6%
Teste de Hipótese: Goodall	0,666	0,523	64,5%
Boosting: Goodall	0,735	0,566	73,7%
Bagging: Goodall	0,752	0,592	77,6%

Fonte: Autor

A Tabela 5 mostra os resultados dos algoritmos, com ela, podemos ver que os destaques são por conta das combinações geradas a partir dos métodos *ensembles*. Em apenas uma

ocasião, o mesmo se manteve abaixo da sua versão base, isso levando em conta o índice externo e a acurácia. A combinação dos classificadores base com o algoritmo *bagging* foi superior em todas as suas versões, sendo o algoritmo *hill climbing*, o mais assertivo e o mais poderoso entre eles. Já na combinação com o *boosting*, apenas uma das versões ficou abaixo do esperado (*boosting*: Hoteling), não melhorando sua versão base.

De modo geral, os índices internos e externos de validação tiveram bons resultados, indicando assim que os agrupamentos foram bem definidos e posteriormente, validando os agrupamentos.

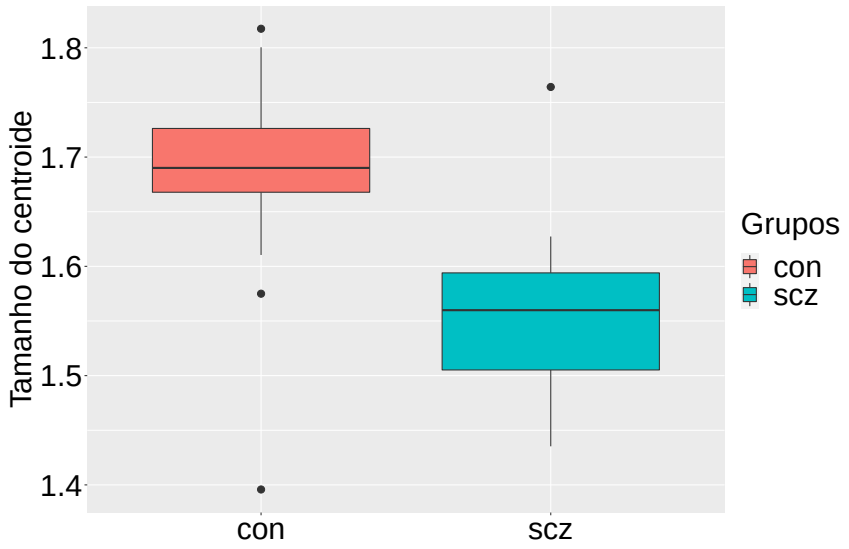
6.2.2 Ressonância magnética de pessoas com esquizofrenia

A esquizofrenia é um transtorno mental grave que muda o modo como a pessoa pensa, sente e se comporta socialmente. Ou seja, essa desestruturação psíquica tem sintomas como alucinações, delírios, dificuldades no raciocínio e alterações no comportamento como indiferença afetiva e isolamento social.

A base de dados das ressonâncias magnéticas de pessoas com esquizofrenia possui 13 marcos em 2 dimensões, e possui um total de 28 indivíduos. Esses indivíduos foram classificados em 2 grupos: controle(**con**)=14 e esquizofrênico(**scz**)=14. Para mais detalhes sobre a base de dados, ver Dryden e Mardia 2016.

Novamente, para ter uma ideia inicial de como os algoritmos se comportarão, foi feito o boxplot do tamanho dos centroides de cada grupo. Pela Figura 6, podemos ver que os dois grupos estão bem separados, porém, existe a presença de outliers nos dois grupos. Com isso, podemos supor a partir dos resultados dos dados simulados, que os algoritmos terão uma boa performance para realizar os agrupamentos.

Figura 6 – Boxplots dos tamanhos dos centróides para os dados das ressonâncias magnéticas de pessoas com esquizofrenia.



Fonte: Autor

Tabela 6 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados das ressonâncias magnéticas de pessoas com esquizofrenia.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,696	0,419	82,1%
Boosting: K-médias	0,746	0,418	85,7%
Bagging: K-médias	0,696	0,419	82,1%
Hill climbing	0,746	0,418	85,7%
Boosting: *hill climbing*	0,756	0,418	87,7%
Bagging: *hill climbing*	0,696	0,419	82,1%
Teste de Hipótese: Hoteling	0,651	0,394	78,6%
Boosting: Hoteling	0,651	0,418	78,6%
Bagging: Hoteling	0,696	0,419	82,1%
Teste de Hipótese: Goodall	0,696	0,419	82,1%
Boosting: Goodall	0,696	0,419	82,1%
Bagging: Goodall	0,746	0,400	85,7%

Fonte: Autor

A Tabela 6 mostra os resultados dos algoritmos, com ela, podemos ver que novamente os destaques são por conta das combinações geradas a partir dos métodos *ensembles*. Em todas as ocasiões, os índices dos métodos *ensembles* tiveram um ganho de performance comparado com a versão base, isso segundo o índice externo e a acurácia. Um destaque aqui é para o

índice interno, o mesmo se manteve estável desde a versão base. Isso significa que o mesmo não conseguiu captar a mudança nas partições geradas, dando a entender, que para dados de tamanho e forma, o mesmo não é tão informativo.

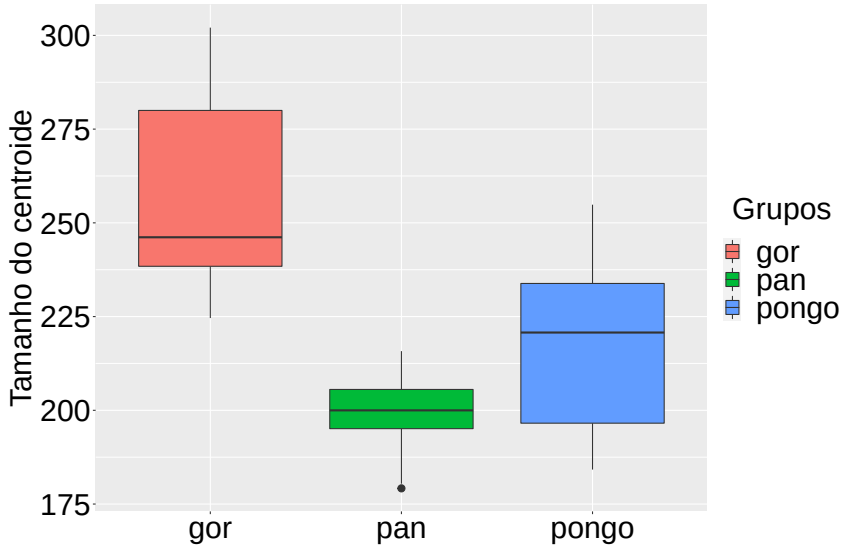
O algoritmo *boosting hill climbing*, obteve melhor performance, isso levando em conta seu índice externo e sua acurácia. Como foi descrito na seção do *boosting*, o método tem vantagem sobre dados com a presença de outliers, por trabalhar com modelos sequenciados e pesos na hora do agrupamento. De modo geral, os índices internos e externos de validação tiveram bons resultados, indicando assim que os agrupamentos foram bem definidos e posteriormente validando os agrupamentos.

6.2.3 Crânio de grandes macacos.

Nessa aplicação, temos os dados de referência dos crânios de grandes macacos. Eles são divididos em 3 espécies/grupos: Gorilas(**gor**)=59, Chimpanzés(**pan**)=54 e Orangotangos(**pongo**)=54. Os dados possuem 8 marcos em 2 dimensões, totalizando 167 indivíduos. Os dados são descritos em detalhes por O'Higgins 1989 e por Dryden e Mardia 2016.

Para ter uma ideia inicial de como os algoritmos se comportarão, foi feito o boxplot do tamanho dos centroides de cada especie, e o mesmo é mostrado na Figura 7.

Figura 7 – Boxplots dos tamanhos dos centroides para os dados de referência dos crânios de grandes macacos.



Fonte: Autor

Como podemos ver, os tamanhos dos centroides das espécies são bem parecidos de maneira

geral, a diferença mais significativa, ficou por conta das espécies dos gorilas e dos orangotangos. Ainda na Figura 7, a espécie dos chimpanzés possui alguns *outliers*, além de ter um certa semelhança com a especie orangotango, fazendo assim, termos suposições através dos resultados dos dados simulados, que os algoritmos terão uma certa dificuldade no agrupamento.

Tabela 7 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados de referência dos crânios de grandes macacos.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0.729	0.762	31,7%
boosting K-médias	0.703	0.702	67,7%
bagging K-médias	0.721	0.762	34,1%
hill climbing	0.708	0.761	32.3%
boosting: *hill climbing*	0.664	0.437	57.5%
bagging: *hill climbing*	0.728	0.762	34.1%
Teste de Hipotese: Hoteling	0.721	0.762	34.1%
boosting: Hoteling	0.637	0.437	47.3%
bagging: Hoteling	0.715	0.762	34.1%
Teste de Hipotese: Goodall	0.719	0.762	34.7%
boosting: Goodall	0.721	0.762	34.1%
bagging: Goodall	0.721	0.762	34.1%

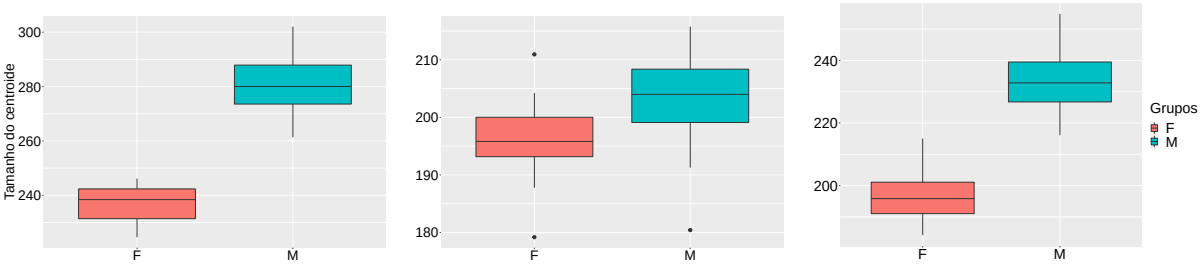
Fonte: Autor

Pela Tabela 7, as suposições feitas anteriormente foram confirmadas. A maioria dos algoritmos tiveram uma taxa de acerto abaixo dos 40% mostrando que os algoritmos tiveram problemas em classificar corretamente as espécies de macacos em seus grupos. Em contra partida, os índices internos e externos tiveram bons resultados, indicando assim que os agrupamentos foram bem definidos. Ainda na Tabela 7, podemos ver o destaque da combinação dos algoritmos com o *boosting*, os mesmos obtiveram as maiores taxas de acerto, sendo o algoritmo *bagging hill climbing* o mais poderoso segundo o índice externo.

Após a baixa taxa de acerto, propormos considerar dos dados acima, o sexo de cada espécie. Essa nova investigação tentará agrupar as diferenças cranianas entre os sexos de cada espécie.

Dos 59 gorilas estudados, sabemos que foram 29 gorilas machos e 30 fêmeas. Dos 54 chimpanzés, sabemos que foram 28 machos e 26 fêmeas. Dos 54 orangotangos, sabemos que foram 30 machos e 24 fêmeas. É interessante avaliar se há uma diferença de tamanho e forma entre as raças, para verificar se há alguma diferença entre os sexos nas regiões do rosto e da caixa craniana.

Figura 8 – Boxplots dos tamanhos dos centroides para os dados de referência dos crânios de grandes macacos baseada no sexo de cada espécie.



(a) Gorilas (b) Chipanzés (c) Orangotangos

Fonte: Autor

Novamente, para ter uma ideia inicial de como os algoritmos se comportarão, foi feito o boxplot do tamanho dos centroides de cada espécie, baseando-se agora no sexo. A Figura 8a e a Figura 8c, mostram uma diferença significativa nos tamanhos dos centroides. Com isso, supomos que os algoritmos nesses dois casos mencionados terão bons resultados no agrupamento. Já na Figura 8b, podemos ver a presença de outliers e a semelhança nos boxplot dos centroides por sexo. Com isso, supomos que os algoritmos nesse caso, terão uma certa dificuldade no agrupamento.

Gorilas

Para a espécie dos gorilas, vimos pela Figura 8a, que os centroides estão bem separados, e supomos que os algoritmos teriam bons resultados no agrupamento.

Tabela 8 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados de referência dos crânios de Gorilas

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	1,000	0,699	100%
Boosting: K-médias	1,000	0,699	100%
Bagging: K-médias	1,000	0,699	100%
Hill climbing	1,000	0,699	100%
Boosting: *hill climbing*	1,000	0,699	100%
Bagging: *hill climbing*	1,000	0,699	100%
Teste de Hipotese: Hoteling	1,000	0,699	100%
Boosting: Hoteling	1,000	0,699	100%
Bagging: Hoteling	1,000	0,699	100%
Teste de Hipotese: Goodall	1,000	0,699	100%
Boosting: Goodall	1,000	0,699	100%
Bagging: Goodall	1,000	0,699	100%

Fonte: Autor

Pela Tabela 8, essas suposições foram confirmadas, mostrando que os algoritmos conseguiram ter um acerto de 100%. Além disso, os índices internos e externos tiveram bom resultados, implicando que os agrupamentos foram bem definidos. Novamente, o destaque é por conta do índice interno, o mesmo, obteve um valor máximo de 0,699, ficando muito distante do valor máximo que o índice pode atingir. Dessa forma, fica evidente que o índice interno não é uma boa escolha quando tratamos de dados de tamanho e forma.

Chimpanzés

Para a espécie dos chimpanzés, vimos pela Figura 8b, que os centroides estavam bem parecidos, além da presença de outliers, supomos então, que os algoritmos teriam uma certa dificuldade no agrupamento.

Tabela 9 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados de referência dos crânios de Chipanzés

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,591	0,394	72,2%
Boosting K-médias	0,628	0,376	75,9%
Bagging K-médias	0,591	0,394	72,2%
Hill climbing	0,591	0,401	72,2%
Boosting: *hill climbing*	0,609	0,396	79,1%
Bagging: *hill climbing*	0,591	0,394	72,2%
Teste de Hipotese: Hoteling	0,560	0,340	68,5%
Boosting: Hoteling	0,618	0,396	75,9%
Bagging: Hoteling	0,575	0,386	70,4%
Teste de Hipotese: Goodall	0,609	0,380	74,1%
Boosting: Goodall	0,591	0,385	72,2%
Bagging: Goodall	0,609	0,396	74,1%

Fonte: Autor

Pela Tabela 9, essas suposições foram confirmadas, mostrando que os algoritmos tiveram dificuldades, mesmo a maioria das taxas de acerto sendo superior aos 70%. O destaque aqui se leva pela combinação dos algoritmos com o *boosting*, como vimos pela Figura 8b, os dados continham a presença de outliers, e como foi descrito na seção do *boosting*, o método tem vantagem sobre dados com a presença de outliers, por trabalhar com modelos sequenciados e pesos na hora do agrupamento. O algoritmo mais poderoso para essa base de dados segundo o índice externo foi *boosting*: K-médias, porém, o mais assertivo foi *boosting* hill-climbing.

Orangotangos

Para a espécie dos orangotangos, vimos pela Figura 8c que os centroides estão bem separados, e supomos que os algoritmos teriam bons resultados no agrupamento.

Tabela 10 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados de referência dos crânios de Orangotangos

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,963	0,612	98,1%
Boosting K-médias	0,927	0,609	96,3%
Bagging K-médias	1,000	0,621	100%
Hill climbing	0,963	0,612	98,1%
Boosting: *hill climbing*	0,963	0,621	98,1%
Bagging: *hill climbing*	0,927	0,613	96,3%
Teste de Hipotese: Hoteling	0,963	0,612	98,1%
Boosting: Hoteling	0,893	0,621	94,4%
Bagging: Hoteling	0,963	0,612	98,1%
Teste de Hipotese: Goodall	0,963	0,612	98,1%
Boosting:Goodall	0,963	0,621	98,1%
Bagging: Goodall	0,963	0,612	98,1%

Fonte: Autor

Pela Tabela 10, essas suposições foram confirmadas, mostrando que os algoritmos conseguiram ter um acerto superior aos 95%. Além disso, os índices internos e externos tiveram bom resultados, implicando que os agrupamentos foram bem definidos. O algoritmo mais poderoso nessa base de dados foi *bagging* K-médias, segundo seus índices externos e acurácia.

7 CONCLUSÕES E TRABALHOS FUTUROS

7.1 CONCLUSÕES

Nesta dissertação, são apresentados algoritmos de agrupamento baseados em teste de hipóteses, K-médias e *hill climbing*, além de suas versões combinadas com os métodos *ensembles* (*boosting* e *bagging*). Esses algoritmos foram adaptados para trabalhar com dados de tamanho e forma.

No geral, os resultados obtidos mostraram que os algoritmos propostos obtiveram bom desempenho, para realizar os agrupamentos, usando dados de tamanho e forma.

Se tratando dos dados simulados, os algoritmos e suas respectivas combinações, foram eficientes para todos os cenários propostos. Para o cenário em que os dados eram bem homogêneos, a combinação do algoritmo K-médias com o método *boosting*, foi a mais poderosa. Já nos dois outros cenários, em que os dados eram mais heterogêneos, o algoritmo *hill climbing* com a combinação do método *bagging*, foi o mais poderoso.

Na avaliação dos resultados dos dados reais, a versão combinada do algoritmo *hill climbing* com o método *bagging*, foi a mais poderosa. Apenas nas bases de dados que continham outliers, os algoritmos baseados em *boosting* se saiam melhor que os algoritmos baseados em *bagging*. Além disso, todas as combinações com os métodos emsenbles causaram influência na melhoria de desempenho dos algoritmos em versões individuais.

Assim, este trabalho contribuiu para a literatura teórica dos métodos de agrupamento para dados de análise estatística de tamanho e forma. Colaborou também, com a proposta de utilização das estatísticas de teste de hipóteses como novos critérios de agrupamento e com a incorporação dos métodos *ensembles*, usando o *bagging* e o *boosting* com os algoritmos de agrupamento, a fim de melhorar a eficiência dos algoritmos reduzindo a variabilidade dos dados.

7.2 TRABALHOS FUTUROS

Segue algumas propostas de trabalhos futuros.

1. Propor e analisar agrupamentos baseados em medoides;
2. Propor e analisar novos métodos de validação;

3. Utilizar outros métodos de reamostragem;

4. Utilizar outros métodos baseados em ensembles;

5. Utilizar outros métodos baseados em redes neurais;

6. Utilizar outros tipos de coordenadas; e

7. Fazer uma comparação entre coordenadas.

REFERÊNCIAS

AMARAL, G. J.; DORE, L. H.; LESSA, R. P.; STOSIC, B. K-means algorithm in statistical shape analysis. *Communications in Statistics—Simulation and Computation®*, Taylor & Francis, v. 39, n. 5, p. 1016–1026, 2010.

ANDERSEN, H. H.; HOJBJERRE, M.; SORENSEN, D.; ERIKSEN, P. S. *Linear and graphical models: for the multivariate complex normal distribution*. [S.l.]: Springer Science & Business Media, 1995. v. 101.

BOOKSTEIN, F. L. A statistical method for biological shape comparisons. *Journal of theoretical biology*, Elsevier, v. 107, n. 3, p. 475–520, 1984.

BOOKSTEIN, F. L. Size and shape spaces for landmark data in two dimensions. *Statistical science*, Institute of Mathematical Statistics, v. 1, n. 2, p. 181–222, 1986.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, 1996.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

DRYDEN, I. L.; MARDIA, K. V. *Statistical Shape Analysis, with Applications in R. Second Edition.* Chichester: John Wiley and Sons, 2016.

FRÉCHET, M. Les éléments aléatoires de nature quelconque dans un espace distancié. In: *Annales de l'institut Henri Poincaré*. [S.l.: s.n.], 1948. v. 10, n. 4, p. 215–310.

FREUND, Y.; SCHAPIRE, R. E. et al. Experiments with a new boosting algorithm. In: CITESEER. *icml*. [S.l.], 1996. v. 96, p. 148–156.

FROSSYNIOTIS, D.; LIKAS, A.; STAFYLOPATIS, A. A clustering method based on boosting. *Pattern Recognition Letters*, Elsevier, v. 25, n. 6, p. 641–654, 2004.

GARETH, J.; DANIELA, W.; TREVOR, H.; ROBERT, T. *An introduction to statistical learning: with applications in R*. [S.l.]: Spinger, 2013.

GOODALL, C. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 53, n. 2, p. 285–321, 1991.

GOODALL, C.; MARDIA, K. V. The noncentral bartlett decompositions and shape densities. *Journal of Multivariate Analysis*, Elsevier, v. 40, n. 1, p. 94–108, 1992.

GOODALL, C. R.; MARDIA, K. V. Multivariate aspects of shape theory. *The Annals of Statistics*, JSTOR, p. 848–866, 1993.

GROVE, K.; KARCHER, H. How to conjugate c 1-close group actions. *Mathematische Zeitschrift*, Springer, v. 132, n. 1, p. 11–20, 1973.

GUILE, G. R.; WANG, W. Factors affecting boosting ensemble performance on dna microarray data. In: IEEE. *The 2010 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2010. p. 1–7.

HOTELLING, H. The generalization of student's ratio. In: *Breakthroughs in statistics*. [S.l.]: Springer, 1992. p. 54–65.

IZENMAN, A. J. *Modern Multivariate Statistical Techniques*. [S.l.]: Springer New York, 2008.

KENDALL, D. G. The diffusion of shape. *Advances in Applied Probability*, Cambridge University Press, v. 9, n. 3, p. 428–430, 1977.

KENDALL, D. G. Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society*, v. 16, n. 2, p. 81–121, 03 1984. ISSN 0024-6093.

KENDALL, D. G.; BARDEN, D.; CARNE, T. K.; LE, H. *Shape and shape theory*. [S.l.]: John Wiley & Sons, 2009. v. 500.

KENT, J. T. The complex bingham distribution and shape analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 56, n. 2, p. 285–299, 1994.

LANDAU, S.; LEESE, M.; STAHL, D.; EVERITT, B. *Cluster Analysis*. [S.l.]: Wiley, 2011. (Wiley Series in Probability and Statistics). ISBN 9780470978443.

LE, H. *Shape theory in flat and curved spaces and shape densities with uniform generators*. Tese (Doutorado) — University of Cambridge, 1988.

LE, H. Mean size-and-shapes and mean shapes: a geometric point of view. *Advances in applied probability*, Cambridge University Press, v. 27, n. 1, p. 44–55, 1995.

LIU, X.-Y.; ZHOU, Z.-H. Ensemble methods for class imbalance learning. *Imbalanced Learn. Found. Algorithms, Appl*, Wiley Online Library, p. 61–82, 2013.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.: s.n.], 1967. p. 281–297.

NASCIMENTO, A.; AMARAL, G.; ACHIC, B.; CRUZ, J. Influential observation in complex normal data for problems in allometry. *Communications in Statistics - Theory and Methods*, Taylor and Francis, v. 45, n. 9, p. 2714–2729, 2016.

O'HIGGINS, P. A morphometric study of cranial shape in the hominoidea. 1989.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, [American Statistical Association, Taylor and Francis, Ltd.], v. 66, n. 336, p. 846–850, 1971. ISSN 01621459.

RIORDAN, R. *O trono de fogo*. Intrínseca, 2011. (AS CRÔNICAS DOS KANE). ISBN 9788580571202. Disponível em: <https://books.google.com.br/books?id=i919pzJOJ3IC>.

SEARLE, S. Restrictions and generalized inverses in linear models. *The American Statistician*, Taylor & Francis, v. 38, n. 1, p. 53–54, 1984.

SEIFFERT, C.; KHOSHGOFTAAR, T. M.; HULSE, J. V.; NAPOLITANO, A. Resampling or reweighting: A comparison of boosting implementations. In: IEEE. *2008 20th IEEE International Conference on Tools with Artificial Intelligence*. [S.l.], 2008. v. 1, p. 445–451.

TAHIR, M. A.; KITTLER, J.; MIKOLAJCZYK, K.; YAN, F. Improving multilabel classification performance by using ensemble of multi-label classifiers. In: SPRINGER. *International Workshop on Multiple Classifier Systems*. [S.l.], 2010. p. 11–21.