Machine learning e análise estatística de formas: uma aplicação no espaço tangente

Ariane Hayana Thomé de Farias[1], Jhonnata Bezerra de Carvalho[1] (PQ)

* *ariane.hayana@outlook.com*

[1] *Universidade Federal do Amazonas, Departamento de Estatística (DE-UFAM).*

Palavras-Chave: Pré-formas, *Kendall*, Formas.

Introdução

A identificação de padrões em imagens tem sido amplamente utilizada nas mais diversas áreas do conhecimento e isso se deve especialmente ao avanço tecnológico que potencializou a necessidade de aplicações cada vez mais eficientes para o processamento digital de imagens. Com o avanço das ferramentas computacionais, a análise estatística de formas (*Statistical Shape Analysis*) tem possibilitado o estudo das formas geométricas de objetos, com aplicações em áreas como medicina, biologia, arqueologia, geografia, bem como em procedimentos de reconhecimento facial e diagnóstico de doenças por meio de ressonâncias magnéticas, entre outras áreas, conforme citam Dryden e Mardia (2016). Ainda segundo os autores, a análise estatística de formas é uma área da estatística responsável pelo estudo das formas geométricas de objetos. Dentre os pilares conceituais da análise estatística de formas cabe destacar o trabalho realizado por Kendall (1977), que foi pioneiro nesta temática e apresentou as primeiras formalizações e conceitos em seu trabalho intitulado "*Shape manifolds, procrustean metrics, and complex projective spaces*" (KENDALL, 1984).

Para tanto, alguns conceitos formais sobre o assunto são fundamentais para a compreensão da análise estatística de formas, a saber: a definição de *forma*, como um conjunto finito de pontos que são denominados de marcos (*landmarks*), que por sua vez, são fundamentais para a captura de pontos de correspondência em cada objeto que são utilizados para comparar o tamanho e a forma geométrica em estudo.

Desta forma, este trabalho apresenta aplicações no espaço tangente de métodos da análise estatística de formas utilizando aprendizado de máquina para a classificação de objetos bidimensionais, com objetivos de: (1) identificar padrões intrínsecos através da classificação de objetos no contexto da análise estatística de formas; (2) utilizar algoritmos de classificação em dados simulados, bem como aplicá-los em quatro bancos de dados reais; e (3) avaliar a performance dos métodos de classificação.

Material e Métodos

Este estudo foi dividido em duas etapas: a primeira consiste em realizar experimentos de simulação com diversas configurações para analisar o desempenho dos classificadores e na segunda etapa utilizou-se quatro bases de dados reais nas quais as observações foram projetadas no espaço tangente, tendo a forma média como polo de projeção. Em ambas as etapas foram utilizados os classificadores SVM (*support vector machine*), ADF (análise discriminante de Fisher), Árvore de decisão, Floresta aleatória e *XGBoost*, os quais foram aplicadas diversas configurações. Na simulação utilizou-se o método de validação cruzada *k-fold*, com $n = 100$ amostras em $k = 5$ partições. Para a classificação no espaço tangente, realizou-se o procedimento de separação e projeção dos dados simulados no plano tangente tendo como polo a forma média. Os testes de classificação foram realizados em uma distribuição normal multivariada, a qual o vetor de médias corresponde às coordenadas de uma figura geométrica triangular. As configurações para as variâncias (σ^2) foram: 0,001, 0,005, 0,009, 0,01, 0,014, 0,018 e 0,02. Nos classificadores foram aplicadas configurações distintas, conforme os parâmetros utilizados em cada algoritmo.

Por outro lado, nas aplicações em dados reais a validação cruzada utilizada foi o *leave-one-out*, para avaliar o desempenho dos classificadores. Além da aplicação desta, utilizou-se também a acurácia como métrica de validação para identificar quanto que o modelo está classificando corretamente. Verificou-se os resultados obtidos na matriz de confusão, e, a partir desta, foi possível avaliar a performance geral dos modelos.

O suporte computacional utilizado neste estudo foi o *software* estatístico R (R Core Team, 2022), e os dados utilizados nas aplicações são oriundos do pacote *shapes* (DRYDEN, 2021), que contém diversos conjuntos de dados e ferramentas fundamentais para análise estatística de formas (DRYDEN; MARDIA, 2016). Assim, alguns bancos de dados bidimensionais foram selecionados e suas características estão descritas na Tabela 1.

Tabela 1. Resumo das características dos dados em estudo.

Dados	Formato	Grupos
Vértebras de camundongos	6 × 2 × 46	Grande e pequeno
Esquizofrenia	13 × 2 × 28	Controle e paciente
Crânios de Gorilas	8 × 2 × 59	Fêmea e macho
Crânios de ratos	8 × 2 × 36	Dias 7 e 150

Fonte: Elaboração própria.

Para ilustrar, o banco de dados Esquizofrenia corresponde a dados cerebrais coletados de ressonâncias magnéticas de voluntários e pacientes esquizofrênicos com intuito de identificar quaisquer diferenças entre o grupo de controle e o grupo de pacientes. Trata-se de coordenadas

bidimensionais com 13 marcos para cada um dos 28 indivíduos. Para ilustrar, a Figura 1 representa a disposição dos marcos em uma ressonância magnética de um paciente. Outros resultados deste estudo estão dispostos em um *dashboard* detalhadamente[1].

Figura 1. Treze marcos anatômicos (marcados em cruz "+") do cérebro de um paciente com esquizofrenia.



Fonte: Dryden e Mardia (2016) adaptado de Bookstein (1996).

Resultados e Discussão

Na primeira etapa deste estudo, os resultados das acurácias por classificador em cada valor de σ^2 foram simulados e dispostos na Tabela 2.

Tabela 2. Acurácia por classificador pelo método *k-fold*.

Classificador	σ^2						
	0,001	0,005	0,009	0,01	0,014	0,018	0,02
SVM Linear	97,0	83,5	77,0	76,0	71,0	70,0	68,5
SVM Polinomial	97,0	80,5	74,0	70,0	66,0	61,0	58,0
SVM Gaussiano	98,5	84,0	78,5	77,5	73,5	73,0	71,5
ADF	70,5	68,0	65,0	65,0	63,5	63,5	63,0
Árvore de decisão	96,5	83,5	78,0	81,5	74,5	71,5	70,5
Floresta aleatória	97,0	82,0	80,0	80,0	74,5	72,0	72,0
XGBoost	98,0	86,5	81,5	81,0	77,0	74,5	75,5

Fonte: Elaboração própria.

Os resultados obtidos mostram que, de modo geral, três classificadores se destacaram em desempenho: *XGBoost*, SVM Gaussiano e Floresta Aleatória com os maiores resultados de acurácia. Nota-se também que conforme σ^2 aumenta, a acurácia diminui, e, além disso, verifica-se que a ADF apresentou o pior desempenho, com acurácia máxima de 70,5% quando $\sigma^2 = 0{,}001$, resultado inferior se comparado com o classificador SVM Gaussiano (98,5%). Outro ponto relevante diz respeito aos resultados da Árvore de decisão e da Floresta Aleatória serem idênticos quando σ^2 é igual a 0,014. Por outro lado, na Tabela 3 pode-se observar os resultados obtidos em dados reais:

Tabela 3. Acurácia dos bancos de dados por classificador.

Classificador	Acurácia (%)			
	Camundongos	Esquizofrenia	Gorilas	Ratos
SVM Linear	93,5	67,9	98,3	100,0
SVM Polinomial	87,0	64,3	93,2	100,0
SVM Gaussiano	91,3	71,4	100,0	100,0
ADF	76,1	46,4	72,9	80,6
Árvore de decisão	95,7	78,6	83,1	100,0
Floresta aleatória	95,7	71,4	94,9	100,0
XGBoost	93,5	67,9	93,2	100,0

Fonte: Elaboração própria.

[1] Disponível em: <https://bit.ly/3PGk2MR>.

Na Tabela 3 é notório que os classificadores, em sua maioria, apresentaram o melhor desempenho para o banco de dados de Crânios de ratos. Os dados de Esquizofrenia foram os que apresentaram os menores percentuais de acurácia, tendo a Árvore de decisão como o melhor classificador dentre os utilizados (com 78,57% de acurácia). Observa-se também que a SVM Gaussiano e a Floresta Aleatória alcançaram os mesmos resultados com 71,43% e os demais classificadores produziram resultados inferiores a estes. No conjunto de dados Vértebras de camundongos observam-se resultados superiores em comparação ao de Esquizofrenia. As melhores acurácias foram obtidas nos classificadores Árvore de decisão e Floresta Aleatória. Outro ponto relevante diz respeito ao resultado produzido pela ADF de 76,09%, bem abaixo dos demais classificadores.

Por fim, apresentam-se os resultados para o conjunto Crânio de gorilas, que dentre as diferentes configurações utilizadas, atingiu 100% de acurácia no classificador SVM Gaussiano, com resultados exitosos também com a SVM Linear (98,31%) e Floresta Aleatória (94,92%) e assim como nos conjuntos mencionados anteriormente, a ADF apresentou o menor resultado (72,88%).

Conclusões

Neste trabalho realizou-se um estudo sobre análise estatística de formas no espaço tangente, com aplicações de técnicas de *machine learning* na classificação de objetos. As contribuições da aplicação de algoritmos de classificação na análise estatística de formas foram diversas, especialmente por ser uma ferramenta poderosa na tomada de decisões em várias áreas.

No que tange às classificações, os resultados obtidos, de modo geral, apresentaram bom desempenho tanto nas simulações, quanto nas aplicações, e, embora os dados de Esquizofrenia tenham apresentado uma acurácia máxima de 78,57%, é importante destacar que existem diversos fatores que impactam no resultado final das classificações (a identificação precisa dos marcos atribuídos por especialistas, a quantidade de dados de entrada do algoritmo, entre outros).

[1] BOOKSTEIN, F. L. Biometrics, biomathematics and the morphometric synthesis. *Bulletin of mathematical biology*, Springer, v. 58, n. 2, p. 313, 1996.

[2] DRYDEN, I. L. *Shapes package*. Vienna, Austria, 2021. Contributed package, Version 1.2.6. Disponível em: <http://www.R-project.org>.

[3] DRYDEN, I. L.; MARDIA, K. V. *Statistical Shape Analysis with applications in R*. [S.l.]: John Wiley & Sons, 2016.

[4] KENDALL, D. G. The diffusion of shape. *Advances in applied probability*, JSTOR, v. 9, n. 3, p. 428–430, 1977.

[5] KENDALL, D. G. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London mathematical society*, Wiley Online Library, v. 16, n. 2, p. 81–121, 1984.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.