

Agrupamento de formas planas utilizando distâncias baseadas na distribuição Bingham complexa

Adenice V. Ferreira,¹ Getúlio J. A. do Amaral,² Abraão D. C. do Nascimento³
PPGEST-DE/UFPE, Recife, PE

Técnicas de agrupamento na área de análises de forma são utilizadas com a finalidade de categorizar as formas de maneira não supervisionada [9]. Assis et al. [1] obtiveram bons resultados em agrupamentos, utilizando marcos anatômicos (*landmarks*), ao implementar quatro diferentes critérios de convergência em um algoritmo k-means, adaptado para análise de formas planas. Entretanto, no cenário em que há baixa concentração dos dados, tais algoritmos não atuaram bem. Considerando esse cenário, achamos coerente fazer o uso de divergências com abordagens diferentes na adaptação do k-means, na tentativa de melhorar o agrupamento. A abordagem considerada foi a distância entre amostras, tomando como base os K-vizinhos mais próximos de cada observação e centróides. Para tanto, considerou-se o estudo de Félix [4], que apresenta em sua tese as divergências de Rényi, Bhattacharyya e Hellinger derivadas para a distribuição Bingham, aplicadas ao contexto de *landmarks*.

Tomamos o algoritmo de Lloyd [7] e a estatística proposta por Goodall [5] como base. Basicamente dispomos de três distâncias (Hellinger, Rényi e Bhattacharyya) e três critérios de convergência/parada diferentes (1-baseados na estatística de Goodall ou Goodall adaptada; 2-número de elementos do grupo na iteração anterior igual ao da atual). Dessa forma, foi possível derivar nove novos algoritmos de agrupamento: BB1 J_B (k-means com distância de Bhattacharyya para Bingham e estatística de Goodall adaptada), BB1 J_G (k-means com distância de Bhattacharyya para Bingham e estatística de Goodall), BB2 (k-means com distância de Bhattacharyya para Bingham e critério de convergência 2), BH1 J_H (k-means com distância de Hellinger para Bingham e estatística de Goodall adaptada), BH1 J_G (k-means com distância de Hellinger para Bingham e estatística de Goodall), BH2 (k-means com distância de Hellinger para Bingham e critério de convergência 2), BR $^{\beta}$ 1 J_R^{β} (k-means com distância de Rényi para Bingham, com parâmetro β , e estatística de Goodall adaptada), BR $^{\beta}$ 1 J_G (k-means com distância de Rényi para Bingham, com parâmetro β , e estatística de Goodall) e BR $^{\beta}$ 2 (k-means com distância de Rényi para Bingham, com parâmetro β , e critério de convergência 2).

A nível de comparação, considerou-se o método de agrupamento k-means para formas planas (KMPS), proposto por Assis et al. [1]. Para tanto foram utilizados como medidas comparativas, o índice de Rand [6], coeficiente Kappa [2] e Acurácia [8] (considerando que os grupos reais eram conhecidos). Na aplicação em banco de dados reais, foi utilizado o banco *DrivFace* [3], disponível para *download* no repositório da UCI *Machine Learning* por meio do sítio eletrônico: <<https://archive.ics.uci.edu/ml/datasets/DrivFace>>.

A base de dados *DrivFace* contém marcos anatômicos (nariz, canto direito e esquerdo da boca, posição dos olhos e do rosto) de quatro sujeitos, ao longo de dias diferentes, durante a condução de um veículo em cenários reais. A câmera fixa na parte frontal e no centro do carro captou três posições para onde se dirigiam as cabeças dos motoristas: frente, direita e esquerda.

¹agdof1@de.ufpe.br

²gjaa@de.ufpe.br

³abraao@de.ufpe.br

O objetivo do agrupamento foi discriminar entre tais posições, comparando-as duas a duas, e para isso foram definidos 4 cenários. No cenário 1, tem-se motoristas olhando para a direita (grupo 1) e motoristas olhando para frente (grupo 2), e o algoritmo com melhor desempenho foi o $BR^{\beta}2$ (acurácia=0,8691, coeficiente Kappa=0,2971 e índice de Rand=0,2491). Tanto no cenário 2, indivíduos olhando para a esquerda (grupo 1) e indivíduos olhando para frente (grupo 2), quanto no cenário 3, motoristas olhando para a direita (grupo 1) e motoristas olhando para a esquerda (grupo 2), destaca-se o algoritmo KMPS.

O cenário 4, sujeitos olhando para frente (grupo 1) e sujeitos olhando para a direita ou para a esquerda (grupo 2), foi escolhido propositalmente por conta da alegação mencionada por Diaz-Chito et al. [3], de que os maiores números de acidentes ocorrem quando os motoristas não estão atentos ao trânsito, ou seja, olhando para a direita ou para a esquerda. Neste cenário, o melhor desempenho foi do algoritmo $BB1J_B$, com acurácia igual a 0,7690, coeficiente Kappa de 0,3255 e índice de Rand igual a 0,2107.

Agradecimentos

Os autores gostariam de registrar agradecimentos à CAPES e CNPq pelo suporte financeiro.

Referências

- [1] E. C. Assis, R. M. C. R. Souza e G. J. A. Amaral. “Using bagging to enhance clustering procedures for planar shapes”. Em: **International Journal of Business Intelligence and Data Mining** 18.1 (2021), pp. 30–48. DOI: 10.1504/IJBIDM.2021.111740.
- [2] J. Cohen. “A coefficient of agreement for nominal scales”. Em: **Educational and psychological measurement** 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.
- [3] K. Diaz-Chito, A. Hernández-Sabaté e A. M. López. “A reduced feature set for driver head pose estimation”. Em: **Applied Soft Computing** 45 (2016), pp. 98–107. DOI: <https://doi.org/10.1016/j.asoc.2016.04.027>.
- [4] W. V. Félix. “Statistical inference based on information theory for pre-shape data”. Tese de doutorado. Universidade Federal de Pernambuco-UFPE, 2019.
- [5] C. Goodall. “Procrustes methods in the statistical analysis of shape”. Em: **Journal of the Royal Statistical Society: Series B (Methodological)** 53.2 (1991), pp. 285–321. DOI: 10.1111/j.2517-6161.1991.tb01825.x.
- [6] L. Hubert e P. Arabie. “Comparing partitions”. Em: **Journal of classification** 2.1 (1985), pp. 193–218. DOI: 10.1007/BF01908075.
- [7] R. Izbicki e T. M. dos Santos. **Aprendizado de máquina: uma abordagem estatística**. 1a. ed. São Carlos, SP: Rafael Izbicki, 2020. ISBN: 9786500024104.
- [8] D. R. Velez et al. “A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction”. Em: **Genetic Epidemiology: the Official Publication of the International Genetic Epidemiology Society** 31.4 (2008), pp. 306–315. DOI: 10.1002/gepi.20211.
- [9] G. Vinué, A. Simó e S. Alemany. “The k-means algorithm for 3D shapes with an application to apparel design”. Em: **Advances in Data Analysis and Classification** 10.1 (2014), pp. 103–132. DOI: 10.1007/s11634-014-0187-1.