Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Novos métodos de agrupamento para formas planas

Jerfson B. N. Honório[1]
Getúlio J. A. do Amaral[2]
CCEN, Recife, PE

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 tamanho e a forma de um objeto, são as informações que permanecem quando os efeitos de locação e rotação são removidos através de operações matemáticas, como descrito em [1]. 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.

O presente trabalho utiliza os métodos de agrupamento baseados em K-médias, proposto por [3], *hill climbing*, proposto por [2], e uma modificação do algoritmo *hill climbing* baseada 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

[1]jerfson.bruno@ufpe.br
[2]gjaa@de.ufpe.br

© 2022 SBMAC

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. [2]

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.

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.

Referências

[1] I. L. Dryden e K. V. Mardia. **Statistical Shape Analysis, with Applications in R. 2a. ed.** Chichester: John Wiley e Sons, 2016.

[2] B. S. Everitt, S. Landau e M. Leese. **Cluster Analysis**. 4a. ed. Wiley Publishing, 2009. ISBN: 0340761199.

[3] J. Macqueen. "Some methods for classification and analysis of multivariate observations". Em: **I5-th Berkeley Symposium on Mathematical Statistics and Probability**. 1967, pp. 281–297.

© 2022 SBMAC