XXXII Congresso de
Iniciação Científica
Unicamp

UNICAMP

PRP
pró-reitoria de pesquisa
unicamp

Study on the *k*-means and Threshold Methods for Image Segmentation: Application in Melanoma Tumor

Beatriz Borges[1] and Vinícius Wasques[1]

[1] Ilum School of Science/Brazilian Center for Research in Energy and Materials, Campinas, Brazil

Abstract— **This paper presents a study of recognition image of cancer tumors. Specifically, the set of images considered in this article is related to the melanoma, which is a particular type and the most serious skin cancer. Two methods of image segmentation are considered in this work, the threshold and *k*-means methods. A discussion about the best value of the threshold and the ideal number of clusters for the *k*-means method is presented, as well as future works. All the simulations were implemented in Python.**

Keywords— **Image Recognition, Melanoma Skin Cancer, Threshold method, *k*-means method, Python.**

I. Introduction

Image processing plays an important role in various fields of science, for example in medical problems, where image recognition is necessary for the classification and analysis of tumors. Among the techniques found in the literature, clustering algorithms stand out as powerful tools for grouping similar elements within large datasets. One such clustering algorithm is the *k*-means method [1], which offers a versatile approach to partitioning data into distinct clusters.

The *k*-means algorithm is a powerful unsupervised learning technique that aims to partition a dataset into *k* clusters based on similarity criteria. This method can be used in several problems, for example in data clustering.

Also, this method can be applied in image segmentation problems, that is, the process of partitioning an image into multiple regions or segments. Among the numerous segmentation techniques, the *k*-means clustering algorithm stands out as a widely-used and effective method for this purpose [2, 3, 4].

In the context of image segmentation, the *k*-means method operates by treating each image pixel as a data point in a high-dimensional space, where the dimensions correspond to the color channels. The algorithm iteratively assigns each pixel to the cluster whose centroid is closest in terms of a similarity measure, chosen by the modeler. Consequently, the algorithm gathers images based on color similarity, and with this, the process provides a meaningful segmentation, allowing the extraction of objects or regions of interest from complex visual data.

Although its effectiveness, the *k*-means method has its limitations, such as sensitivity to the initial choice of cluster centroids and its assumption of clusters having a spherical shape in the feature space. Nonetheless, with proper parameter tuning and preprocessing, the *k*-means method remains a valuable tool in the image segmentation techniques.

Another technique that is widely used in image processing problems is the threshold method. This approach involves setting a specific boundary or threshold value to distinguish among different states or classes within a dataset. By defining this threshold, it is possible to make informed decisions, classify data points, or detect anomalies based on specific criteria [5, 6, 7].

This work is dedicated to study the applications of the *k*-means and threshold methods for image segmentation, in order to enable the analysis of cancerous tumors. For this purpose, images found in the literature were used, and the *k*-means method was applied to segment these images, in order to show the advantages of the method and helping the analyze of the tumor.

The paper is structured as follows. In Section II we present some basic definitions for a better understanding of the threshold and *k*-means methods. In Section III we present the application of the methods to a set of images found in the literature. In Section IV we provide a brief analysis of the results and in Section V we present the final remarks of the papers, as well as future works.

II. Preliminaries

This section provides the mathematical definition of *k*-means and threshold methods, for the better understanding of the paper.

A. *The threshold method*

The threshold method can be viewed as a binary classification, where the goal is to separate data points into two classes based on a threshold value.

Denoting the features of each data point as x, and the corresponding class labels as y, where $y = 0$ represents one class and $y = 1$ represents the other, the threshold method compares a certain feature, or combination of features, to a threshold value θ, whose decision rule is given by $y = 0$, if $x < \theta$ or $y = 1$, if $x \geq \theta$.

If the feature x is less than the threshold value θ, then the data point is assigned to class 0, otherwise, it is assigned to the class 1.

For a more general case, the decision rule may involve a more complex comparison involving multiple features or a combination of features. For instance, if we are dealing with multiple features $x_1, \ldots, x_n$, then decision rule is given by

$$y = \begin{cases} 0, & if\, f(x_1, \ldots, x_n) < \theta \\ 1, & if\, f(x_1, \ldots, x_n) \geq \theta \end{cases}, \tag{1}$$

where $f(x_1, \ldots, x_n)$ is some function associated with the features.

The threshold value is typically determined during the model training process, where the objective is to find the optimal value for θ that minimizes a certain cost function or maximizes a certain performance metric, such as accuracy or precision [8].

B. The k-means method

The k-means algorithm works by iteratively assigning data points to clusters and updating the cluster centroids based on the mean of the data points assigned to each cluster. This algorithm can be organized in the following steps.

First the algorithm starts by randomly initializing k cluster centroids. Recall that a centroid of n data points $\{x_1, \ldots, x_n\}$ in a p-dimensional space, where $x_i = (x_{1i}, \ldots, x_{pi})$ with $i = \{1, \ldots, n\}$, is defined by

$$C = (\mu_1, \ldots, \mu_p), \tag{2}$$

where

$$\mu_k = \frac{1}{n} \sum_{i=1}^{n} x_{ki}, \tag{3}$$

with $k = \{1, \ldots, p\}$.

These centroids can be randomly chosen from the data points themselves or from a uniform distribution within the range of the data. Next, each data point is assigned to the nearest cluster centroid based on a distance metric. In general, the Euclidean distance is considered, that is,

$$d(x,y) = ||x_i - y_i||_e = \left(\sum_{i=1}^{p} (x_i - y_i)^2 \right)^{\frac{1}{2}}, \tag{4}$$

where $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_p)$.

Mathematically, for each data point x_i, the algorithm calculates the distance between x_i and each centroid μ_j, and assigns x_i to the cluster with the closest centroid. This is known as the assignment Step and can be summarized in the following task:

$$cluster(x_i) = arg\min_j ||x_i - c_j||_e^2 \tag{5}$$

where $|| \cdot ||_e$ is the euclidian metric, $cluster(x_i)$ represents the cluster in which x_i is assigned and c_j stands for the centroid of cluster j.

After assigning all data points to clusters, the centroids are updated based on the mean of the data points assigned to each cluster. The new centroid c_j for cluster j is computed as the mean of all data points x_i assigned to cluster j, that is,

$$c_j = \frac{1}{N_j} \sum_{x_i \in cluster(j)} x_j \tag{6}$$

where N_j is the number of data points assigned to cluster j, $cluster(j)$ represents the set of data points assigned to cluster j.

These steps are repeated until the convergence criteria are met, that is, when either the centroids no longer change significantly between iterations or a maximum number of iterations is reached. Once the convergence is achieved, the algorithm provides the final cluster assignments and centroids [8].

In summary, the k-means algorithm focuses on the minimization of the sum of squared distances between each data point and its assigned centroid. In this paper, both threshold and k-means methods will be implemented based on adaptations of Python libraries found in the literature.

III. The k-means and Threshold methods applied to cancerous tumors images

In this section it will be presented the study of cancer tumor given by a set of images. These images can be found in [9].

In order to analyze these images, the threshold and k-means methods will by applied, considering as input the image given in Figure 1.

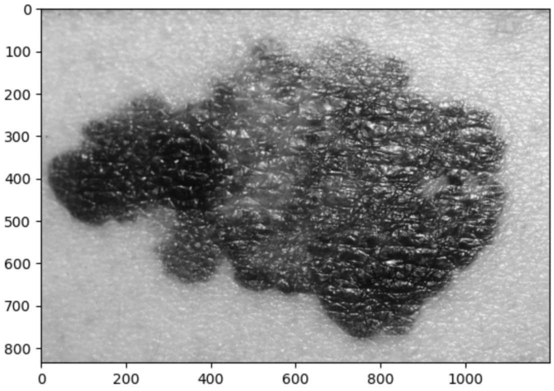

Fig. 1: Melanoma, image as input. Source: The authors

The results of the simulations are presented in Subsections III-A and III-B, and the discussion of the obtained results are presented in Section IV.

A. *Threshold method*

For the following project, the threshold method was used for segmenting melanoma images, in order to compare it with other available methods.

The simple threshold operates by binarizing images based on their brightness, using a threshold defined by the user. Pixel values range from 0 to 255, where 0 represents the absence of brightness (black) and 255 represents maximum brightness intensity (white). When setting a threshold, pixels with above values are assigned the maximum value (255), while pixels with below values are assigned the minimum value (0).

Figure 2 depicts the threshold method applied to image of the cancer tumor presented in Figure 1, considering the threshold value as $\theta = 155$.

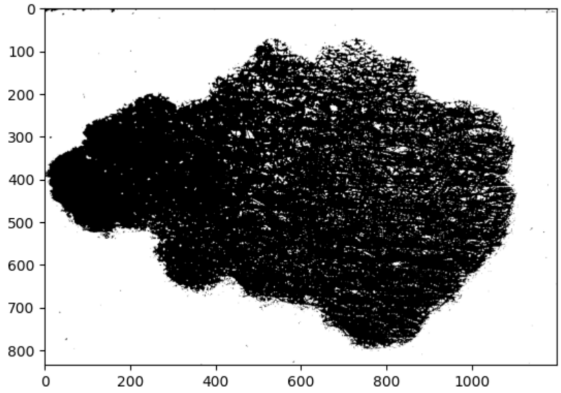

Fig. 2: Melanoma threshold 155. Source: The authors

It is possible to observe in Figure 2 that the cancer tumor was identify by the threshold method. Note that the image border was not well identified by this method. In fact, this same issue appears in all threshold values that was considered in this study. Recall that the $\theta = 155$, which gives raise to the image given in Figure 2, was the best value for this particular image of the cancer tumor.

B. *The k-means method*

The *k*-means method was also used to segment images, following the same set of images used in the previous method. This method differs from thresholding mainly by working with clusters instead of performing direct binarization.

Clusters are groups or sets of points or objects in a space that share similar characteristics among themselves and are distinct from the characteristics of points or objects in other clusters. This means that the pixels in the input image are grouped based on some common feature. In this study, the Euclidean distance between pixels was used as a common feature. By choosing only 3 clusters, it is possible to determine to which cluster a certain pixel belongs based on Euclidean distance, assigning the pixel to the cluster with the smallest distance.

Figure 3 depicts the *k*-means method applied to the image presented in Figure 1. It is possible to observe that the cancer tumor was also identify by the *k*-means method and also the border was better identify than the threshold method.

It is important to observe that the results, in terms of the recognition of the border, was not entirely satisfactory.

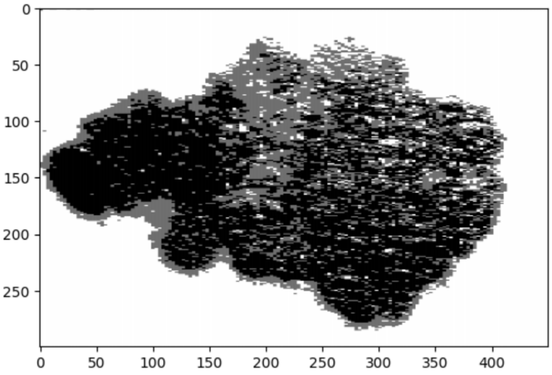

Fig. 3: Melanoma *K*-means 3 clusters. Source: The authors

Both methods use the same image as input for comparative purposes. This image is converted to grayscale, meaning that only the brightness intensity of the pixels is stored in arrays.

An array is a data structure that stores a set of elements, usually of the same type, in sequential order. Each element in an array is identified by a unique index representing its position in the sequence. According to the segmentation method applied, the images are treated differently.

IV. Analysis of the results

For the first approach, it was considered different threshold values. The value $\theta = 155$ was considered the most appropriated for this particular image. It is important to highlight that for higher values of threshold θ the image obtained from the simulations contained a lot of noise. On the other hand, for lower values of θ the image obtained lacks in the information inside the tumor.

For the second approach, the most appropriated number of cluster was 3, where similar issues observed in the threshold method also occur for the k-means method. It is worth to mention that as more clusters were considered, more noise was observed, mainly at the image border.

The analysis of the results obtained with the two methods presented, using different thresholds and numbers of clusters, reveals that both methods face difficulties in determining the edges, especially when they exhibit a gradient.

This difficulty can be attributed to the nature of the segmentation methods used. In the case of thresholding, determining a precise boundary can be challenging, especially in areas of gradual transition between regions of different intensity. Similarly, the k-means method may not adequately capture intensity variations when grouping pixels into clusters, resulting in inaccurate segmentations in gradient regions.

V. Finals Remarks

This paper provided a study of image recognition considering two important and widely approaches found in the literature, the threshold and k-means methods.

These methods were applied to a recognition image problem, in order to identify cancer tumors. The input data is provided in Figure 1 and the result of the simulations via threshold and k-means are illustrated in Figures 2 and 3, respectively.

Both approaches produced good results in terms of the recognition of the cancer tumor, but they lack in a satisfactory recognition of the border.

For future works, we intend to study and apply the c-means method, which is also a clustering technique where each data point is grouped into different clusters and assigned a probability score, combining with the fuzzy sets theory, where each data point can belong to more than one cluster [10]. This future work aims to better identify the border of these images.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

The second author thanks the support of FAPESP under grant nº 2023/03927-0.

References

1. MacQueen J. B.. *Some Methods for classification and Analysis of Multivariate Observations*. California: University of California Press 1967.
2. Dehariya Vinod Kumar, Shrivastava Shailendra Kumar, Jain R.C.. Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms in *2010 International Conference on Computational Intelligence and Communication Networks*:386-391 2010.
3. Zheng X., Lei Q., Yao R., Gong Y., Yin Q.. Image segmentation based on adaptive K-means algorithm *J Image Video Proc.*. 2018;68:1–10.
4. Malathi R., Nadirabanu K. A. R.. Brain Tumor Detection and Identification Using K-Means Clustering Technique in *Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications*:14–18 2015.
5. Kittler J, Illingworth J, Föglein J. Threshold selection based on a simple image statistic *Computer Vision, Graphics, and Image Processing.* 1985;30:125-147.
6. Niu Zuodong, Li Handong. Research and analysis of threshold segmentation algorithms in image processing *Journal of Physics: Conference Series.* 2019;1237:022122.
7. Zhu Shiping, Xia Xi, Zhang Qingrong, Belloulata Kamel. An Image Segmentation Algorithm in Image Processing Based on Threshold Segmentation in *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*:673-678 2007.
8. Gonzalez R. C., Woods R. E.. *Digital Image Processing*. Pearson 2007.
9. Adobe Stock. ”Melanoma” at https://stock.adobe.com/br/search?k=melanoma
10. Coelho S., Fernandes M. A. R., Miot H. A., Yoriyaz H.. Use of the c-means fuzzy method to skin lesion dermatoscopic image segmentation *Revista Brasileira de Física Médica.* 2012;6:99–102.

Author: Beatriz Borges
Institute: Ilum School of Science/Brazilian Center for Research in Energy and Materials (CNPEM)
Street: Lauro Vannucci, 1020
City: Campinas
Country: Brazil
Email: beatrizborgesribeiro2@gmail.com