Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, Vol. 3, N. 1, 2015.
Trabalho apresentado no XXXV CNMAC, Natal-RN, 2014.

Using ckMeans algorithm in image segmentation process: Preliminary results on mammography analysis

Rogério R. de Vargas, **Graçaliz P. Dimuro**,
Center of Computational Sciences (C3), Federal University of Rio Grande (FURG),
96201-900, Rio Grande, RS, Brazil
homepage: http://rogerio.in, http://www.gracalizdimuro.com,

Benjamín R. C. Bedregal
Logic, Language, Information, Theory and Applications (LoLITA),
Federal University of Rio Grande do Norte (UFRN),
59078-970, Natal, RN, Brazil
E-mail: bedregal@dimap.ufrn.br.

Abstract: *Clustering algorithms aim at modelling fuzzy (i.e., ambiguous) unlabelled pat- terns efficiently. Our goal is to propose a ckMeans algorithm to the image segmentation process. To validate the proposed methodology we applied the algorithm in mammography images. We present the initial results considering just one image.*

Keywords: *ckMeans, Clustering, Image Segmentation.*

1 Introduction

Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both. If meaningful groups are the goal, then the clusters should capture the natural structure of some data. In some cases, however, cluster analysis is only a useful starting point for other purposes, such as data summarization. Cluster analysis has long played an important role in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining [1].

According to Fayyad, Han and Agrawal [2, 3, 4], clustering techniques seeks to identify a set of categories or classes to describe the data. One starts from a situation where there are no classes, only elements of a universe. From these elements, clustering techniques are responsible for defining classes and framing elements.

In many applications of image processing, the gray levels of pixels belonging to the object are substantially different from the gray levels of the pixels belonging to the background. Thresholding is a simple but effective tool for classifying, according to the level of gray, the pixels into two types: those that belong to the background and those that belong to the object. The process of dividing an image into disjoint regions (or classes) such that each one of them has very concrete attributes or properties is called segmentation. Each one of these regions represents an object of the image.

In this paper we work on the case in which there is only one object in the image, therefore we have two regions: the object and the background. Usually an image is composed of many objects, so in practice it is necessary to choose more than one threshold in order to segment the image [5]. Hence, in this paper we propose to use ckMeans algorithm in the process of clustering in image segmentation, considering an application to mammography analysis.

Section 2 describes the ckMeans algorithm. Section 3 discusses experiments of the ckMeans algorithms. Finally, Section 4 describes the final conclusions and future works.

DOI: 10.5540/03.2015.003.01.0386

2 ckMeans Algorithm

The idea is basically to share the fuzzy set $X = \{x_1, x_2, \ldots, x_n\}$ in p clusters where μ_{ij} is the membership degree of the sample x_i that belongs to the j-th cluster and the result of clustering is expressed by membership degrees on matrix μ.

The ckMeans algorithm [6] attempts to find sets of data by minimizing an objective function shown in Equation (1):

$$J = \sum_{i=1}^{n} \sum_{j=1}^{p} \mu_{ij}^{m} d_{ij} \left(x_i; c_j\right)^2 \tag{1}$$

where:

- n is the number of data;
- p is the number of clusters considered in the algorithm, which must be decided before execution;
- m is a fuzzification parameter[1] in the range $(1; w)$, indicating the width of n dimensional cluster perimeter. Usually, m is the range $[1.25; 2]$ [7];
- x_i a vector of training data, where $i = 1, 2, \ldots, n$. These are the cluster attributes selected from the source data elements (such as columns in a database table);
- c_j the centroid (or centrer) of a fuzzy cluster $(j = 1, 2, \ldots, p)$;
- $d_{ij} \left(x_i; c_j\right)$ is the distance[2] between x_i and c_j;

The input of the algorithm is n data, the number of clusters p and value m. Its steps are:

1. Starts μ (membership degree) with a continuous random value between zero (no relevance) and 1 (total relevance) where the sum of pertinence must be one.

2. Calculate the centroid of the cluster j as follows: We stabilize the centroid of each cluster as in the K-Means algorithm. However, in our algorithm we first create a new matrix, which is called μCrisp, containing values 1 or zero. Each line of this new matrix has 1 in position of the greatest value of this line in the μ and zero in the other positions of the line. When a column of the matrix μCrisp, after this step, has only zeros in it, it is assigned the value 1 in the position that corresponds to the largest value of the same column in the matrix μ.

   The ckMeans algorithm returns a matrix μCrisp, which the values of the elements belong to the set $\{0,\ 1\}$ as shown in Equation (2). In other words, μCrisp is the matrix while μCrisp$_{ij}$ is the content of matrix at the position (ij):

$$\mu Crisp_{ij} = \max \left(\left\lfloor \frac{\mu_{ij}}{\max\limits_{l=1}^{p} \mu_{il}} \right\rfloor , \left\lfloor \frac{\mu_{ij}}{\max\limits_{l=1}^{n} \mu_{lj}} \right\rfloor \right) \tag{2}$$

   The first argument of the right side of Equation (2) is for any datum whose value is 1 for the cluster it belongs with the greatest degree, and 0 for the others. The second argument

[1] We only consider rational values to simplify the calculation of Equations (1), (2) and (4). Actually, it is used rational m's.

[2] When the values are numbers, it is usually used the Euclidean distance.

DOI: 10.5540/03.2015.003.01.0386

is for the greatest degree of each column (cluster) is 1, so as to ensure that all clusters have at least one element. Thus, on rare occasions may happen that a line has more than one value 1 (which does not occur in the algorithm K-Means), but as this matrix is only auxiliary, this does not bring any inconvenience.

The steps of the algorithm to calculate μCrisp_{ij}[3] are performed as follows:

(a) Read μ;

(b) Find the larger value at the first line of the matrix μ. After that, assign, on μCrisp matrix, the value 1 to the position corresponding to the larger value position on matrix μ and 0 to the others. To complete the process, repeat the same procedure to the other lines;

(c) Store in a vector the number of 1's that each column μCrisp has.

If a column in μCrisp has no 1's, assign 1 in the position of the largest value of that column of the matrix μ.

After calculating the matrix μCrisp calculate the new centroids of clusters as in Equation (3):

$$c_j = \frac{\Sigma_{i=1}^{n} x_i \mu Crisp_{ij}}{\Sigma_{i=1}^{n} \mu Crisp_{ij}} \tag{3}$$

c_j is calculated by adding the data belonging to cluster (in crisp form) and dividing it by the number of classified objects as 1 in the matrix μCrisp for this cluster.

3. Calculate an initial value (a data) for J using the Equation (1);

4. Calculate the table of the fuzzy membership function as shown in Equation (4):

$$\mu_{ij} = \frac{\left(\frac{1}{d_{ij}(x_i;c_j)}\right)^{\frac{2}{m-1}}}{\displaystyle\sum_{k=1}^{p} \left(\frac{1}{d_{ik}\left(x_i;c_k\right)}\right)^{\frac{2}{m-1}}} \tag{4}$$

5. Return to step 2 until a convergence condition is reached.

Some possible stopping conditions are:

- A fixed number of iterations is executed;
- The user reports a value $\epsilon > 0$ of convergence, and if

$$d_{ij}\left(J_U; J_A\right) \leq \epsilon$$

then the algorithm stops, where J_A is the objective function (Equation (1)) calculated in the previous iteration and J_U is the objective function of the last iteration.

3 Preliminary Results

The steps used for the to fuzzy image segmentation process are showed in Figure 1:

[3] There may be a situation where the result of μCrisp_{ij} is not completely accurate Equation (2). In this case, the greatest value of the column μ_{ij} have 1 in μCrisp_{ij}.

DOI: 10.5540/03.2015.003.01.0386

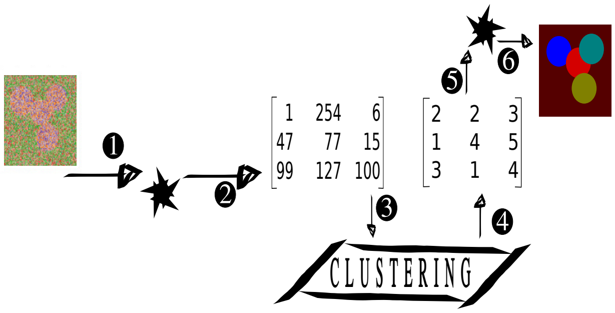

Figure 1: Steps of Image Segmentation Process

1. Load image to clustering;
2. Convert image to a data matrix;
3. Clustering matrix;
4. Return a matrix cluster;
5. Load matrix cluster;
6. Convert the matrix cluster in a image.

The first tests using the ckMeans algorithm just converting an image to data, applying the clustering process and then returning an image, is shown below. The simulated image is a mammography obtained from MIAS MiniMammographic Database [8].

This file (Figure 2) lists the films in the MIAS database and provides appropriate details as follows: mdb005 (filename), Fatty (character of background tissue), CIRC (class of abnormality, well-defined/circumscribed masses), Benign (severity of abnormality), at coordinate (477, 133) (x, y image-coordinates of centre of abnormality; coordinate system origin is the bottom-left corner) and 30 (approximate radius (in pixels) of a circle enclosing the abnormality).

The best configuration presented in ckMeans algorithm after several simulations are: $m = 2$ (Fuzziness), $\epsilon = 0.001$ and $p = 5$ (clusters number).

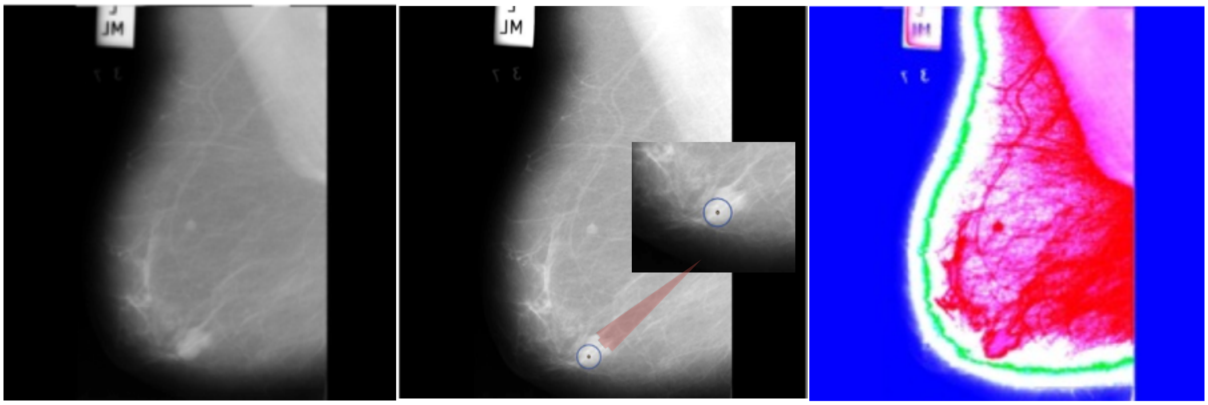

Figure 2: Image Segmentation Process in Mammography with ckMeans Algorithm.

Note that an image in grayscale was converted to 5 colors. These colors mean the number of clusters. The result of this clustering may help a physician to better observe the mammography and make better decisions.

4 Conclusion

Cluster analysis is a procedure performed in only one run. In many circumstances, you need a series of trials and repetitions. Still, there is an effective and universal criteria to guide the selection of attributes and clustering algorithms. Validation criteria derived impressions about the quality of the clusters, but how to choose the same criterion is still a problem that requires more effort [9].

In this work we proposed use the ckMeans algorithm in the image process segmentation, with an application to mammography images. The results showed that the algorithm clustering the colors by number of cluster defined a priori.

Future works will focus on the application of the ckMeans algorithm to other images and compare it with the results and variants of clustering algorithm. We also intend to use Overlap functions [10] and some edge detection algorithm.

References

[1] P. Tan; M. Steinbach; V. Kumar, *Introduction to Data Mining.* Us ed. [S.l.]: Addison Wesley, 2005. Hardcover. ISBN 0321321367.

[2] U. Fayyad, *Advances in Knowledge Discovery and Data Mining.* [S.l.]: AAAI/MIT Press, 1996.

[3] J. HAN et al., Intelligent Query Answering by Knowledge Discovery Techniques. *IEEE Transactions on Knowledge and Data Engineering*, v. 8, p. 373–390, 1996.

[4] R. Agrawal, Data Mining: The Quest Perspective. *Australian Computer Science Comm. — Proc. 7th Australian Database Conf., ADC*, v. 18, n. 2, p. 119–120, 1996.

[5] P. Melo-Pinto et al., Image segmentation using atanassov's intuitionistic fuzzy sets. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 40, n. 1, p. 15–26, jan. 2013. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2012.05.055>.

[6] R. de Vargas, B. Bedregal, A comparative study between fuzzy c-means and ckmeans algorithms. In: *Proc. Conf. North American Fuzzy Information Processing Society (NAFIPS 2010).* Toronto, Canada: [s.n.], 2010.

[7] E. Cox, *Fuzzy modeling and genetic algorithms for data mining and exploration.* [S.l.]: Elsevier/Morgan Kaufmann, 2005. Hardcover. (Morgan Kaufmann series in data management systems).

[8] J. Suckling, The mammographic image analysis society digital mammogram database exerpta medica. *International Congress Series*, v. 1069, n. 0, p. 375 – 378, 1994.

[9] J. Cavalcanti, *Clusterização Baseada em Algoritmos Fuzzy.* Dissertação (Mestrado) — Universidade Federal de Pernambuco, Recife, Brasil, 2006.

[10] A. Jurio et al., Some properties of overlap and grouping functions and their application to image thresholding. *Fuzzy Sets and Systems*, v. 229, n. 0, p. 69 – 90, 2013. Theme: Computer Science.

DOI: 10.5540/03.2015.003.01.0386