Managing the doubt in fuzzy clustering by means of interval-valued fuzzy sets

Aranzazu Jurio
Departamento de Estadística,
Informática y Matemáticas
Universidad Publica de Navarra
Pamplona, Spain
aranzazu.jurio@unavarra.es

Humberto Bustince
Departamento de Estadística,
Informática y Matemáticas
Universidad Publica de Navarra
Pamplona, Spain
bustince@unavarra.es

Vicenc Torra
Hamilton Institute
Maynooth University
Maynooth, Ireland
vtorra@ieee.org

***Abstract*—In this work we study how the outliers can distort a partitional clustering process. We present a new algorithm to avoid this distortion. It is based on the minimization of a new objective functions, which is an extension of the one of the Fuzzy Clusters Means algorithm. The main novelty is the use of interval values to calculate the membership degrees of each datum to each cluster. We show the performance of our proposal over different datasets and we present its advantages in image segmentation.**

***Index Terms*—clustering, interval membership degree, outliers**

I. Introduction

Clustering is an unsupervised classification problem where the purpose is to find the natural groups that exist in a dataset. It is based on the idea that the data belonging to the same group must have similar characteristics whereas the data that belong to different groups must be different in the same characteristics [5].

Clustering methods can be generally divided into two types: hierarchical methods and partitional ones. Hierarchical methods construct a tree based on the similarities between data [6] [12]. On the other side, partitional methods divide the whole dataset into a fixed number of clusters, each of them represented by a centroid. The centroid is the point whose sum of distances to all the data in the cluster is minimum [8] [9] [10]. In this work we focus on partitional clustering.

K-means [4] [10] is one of the most well known and used algorithms among the partitional methods. This algorithm divide all the existing data into c clusters and calculates the centroid of each cluster. The goal is to minimize the sum of the distances between every datum and its corresponding centroid.

$$J = \sum_{i=1}^{c} \sum_{x_k \in cluster_i} ||x_k - v_i||_A^2$$

where $||x||_A = \sqrt{x^t A x}$ is any norm associated with an inner product.

One of the main problems of the k-means algorithm, in the same way as it happens in most partitional algorithms, is that it is not able to deal with natural groups which are overlapped. In this situation, the data located in the overlapped area should belong to all of these clusters, but the characteristics of the algorithm restrict them to belong only to one of them.

In the literature, this problem is solved using fuzzy set theory [13]. With this tool, each datum can belong to more than one cluster, with different membership degrees. The membership degrees are real values between 0 and 1. The Fuzzy Cluster Means (FCM) algorithm [1] extends the idea of the k-means algorithm using fuzzy set theory.

However, if the data to classify include outliers, the FCM cannot detect them, so its results are distorted by them.

To solve this new problem, in this work we present a new clustering algorithm that extends the FCM. It is able to detect the outliers in the data in order to minimize their influence in the result.

In the same way as fuzzy sets allow to include new information in the clustering process, we use an extension of fuzzy sets to increase the amount of new information. This extension is the use of interval-valued fuzzy sets. In this work we use these sets to quantify the membership degrees. Therefore, every datum of the dataset belongs to all the existing clusters with different membership degrees made of intervals in [0,1]. We use the length of each interval to model the uncertainty we have that this datum belongs to the clusters we are detecting. For example, if we are completely sure that one datum belongs to one or more clusters, then its membership degrees will be intervals with length 0. On the contrary, if we are completely sure that one datum does not belong to any of the clusters, then its membership degrees will be intervals with length 1, which is the maximum.

The remaining of this work is organized as follows: in Section II we briefly summarize the Fuzzy Cluster Means algorithm and we show its inaccuracies when there are outliers; in Section III we show our new proposal and in Sections IV and V we apply it over several datasets and images to segment, respectively. Finally, in Section VI we finish with our conclusions.

This work is partially supported by the *Ministerio de Educación, Cultura y Deporte* from the Spanish Government under the program *José Castillejo para estancias de movilidad en el extranjero de jóvenes doctores*. It is also partially supported by the Public University of Navarra under the project PJUPNA13 and by the Spanish Government under the project TIN2016-77356-P (AEI/FEDER, UE).

978-1-5386-1728-1/19/$31.00 ©2019 IEEE

II. Fuzzy c-means

Fuzzy Cluster Means (FCM) [1] is one of the most well known clustering algorithms. Thanks to the use of fuzzy sets, it allows each datum to belong to more than one cluster at the same time. Indeed, it is based on the idea that every datum has to belong to all the existing clusters with a specific membership degree. The membership degrees are values between 0 and 1, with the restriction that the sum of all membership values of each datum must be always 1.

Under this restriction, the goal of the FCM is to minimize the sum of the weighted sum of distances between every datum and all the centroids. The weights are proportional to the membership values.

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m ||x_k - v_i||_A^2$$

where x_k is the k-th datum to classify, v_i is the centroid of the cluster i, u_{ik} is the membership degree of the datum k to the cluster i and m is a real value greater than 1. Moreover, three constraints must be fulfilled:

- $u_{ik} \geq 0, \quad k = 1..n,\ i = 1..c$
- $\sum_{k=1}^{n} u_{ik} > 0, \quad i = 1..c$
- $\sum_{i=1}^{c} u_{ik} = 1, \quad k = 1..n$

The solution to this problem is an iterative process that starts with random centroids. Based on them, the algorithm calculates the membership values of every datum to every cluster:

$$u_{ik} = \left(\sum_{j=1}^{c} \left(\frac{||x_k - v_i||_A}{||x_k - v_j||_A} \right)^{2/(m-1)} \right)^{-1}$$

$k = 1..n$, $i = 1..c$. Based on the membership values, the algorithm updates the centroids:

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}$$

$i = 1..c$. The process finishes when the changes in the values are small enough.

The FCM is not able to deal properly with noisy datasets or datasets that include outliers. When working with this kind of datasets, all the data must be assigned to the different clusters, so the remote data affect the centroids in a wrong way.

In Figure 1 we show a dataset where each datum is represented by a black star. There are two clear overlapped clusters and one datum that does not belong to any of the groups. After applying the FCM for two clusters, we show in red circles the obtained centroids.

Both clusters are vertically centered at the point 2.5. However, as the outlier located at $(5, 20)$ belongs to both clusters, the centroids are moved upwards, exactly to the value 2.83.

Moreover, if we analyze the membership values, the datum located at (5, 2.5) belongs to both clusters with a value 0.5. In the same way, the datum located at (5, 20) also belongs to both clusters with a value of 0.5. Therefore, the algorithm considers both data equally to determine the clusters. But looking at the figure, it is clear that the first datum is located in the intersection of both clusters whereas the second datum is out of the clusters.

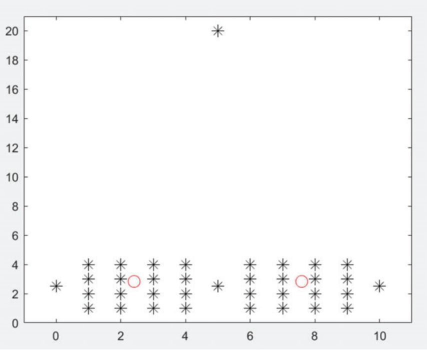

Fig. 1. Execution of the FCM over a dataset with one outlier. Original data in black stars. Final centroids in red circles.

From this example we can conclude that the Fuzzy Cluster Means does not lead to an appropriate representation of the clusters when the dataset to classify includes outliers, i.e. data that do not belong to any of the existing clusters. We present an alternative representation that permits us to distinguish between (5, 20) and (5, 2.5).

III. New algorithm for interval clustering

In this section we present our proposal for clustering. Its main novelty is the use of interval-valued fuzzy sets to represent the membership degree of every datum to every cluster.

We denote by $L([0, 1])$ the set of all closed subintervals in $[0, 1]$. That means

$$L([0,1]) = \{\mathbf{x} = [\underline{x}, \overline{x}] | (\underline{x}, \overline{x}) \in [0,1]^2 \ \text{ and } \ \underline{x} \leq \overline{x}\}$$

An interval-valued fuzzy set Z in the universe $U \neq \emptyset$ is a mapping $Z : U \rightarrow L([0, 1])$.

Our proposal is based on one of the existing interpretations of interval-valued fuzzy sets: "*The membership degree of an element to the set corresponds to a value in the considered membership interval. We cannot say in a precise way what that value is; therefore, we just provide bounds for it*" [2].

Following this idea, we can assume that the length of the interval represents the uncertainty we have when determining the membership value of the element to the set.

Applying this interpretation to our problem, when the algorithm is completely sure that one datum belongs to a cluster, then the length of its membership intervals will be minimum. It does not matter if the membership value is $[1, 1]$ to one cluster and $[0, 0]$ to the other ones; or if the membership value is $[0.5, 0.5]$ to two clusters. On the contrary, if the algorithm is not sure whether one datum belongs to any of the natural clusters in the dataset, then the length of all its membership intervals will be bigger. In the limit case, one datum can belong with an interval of $[0, 1]$ to all of the existing clusters.

If there is no doubt about the membership degrees of one datum to the clusters, then all its membership lower bounds will be equal to its membership upper bounds. Following the constraints of the FCM, the sum of all the lower (or upper) memberships will be 1, so the total sum of all memberships must be 2. On the other hand, when the doubt about the memberships is maximum, the interval membership degrees to all clusters will be $[0, 1]$, so the total sum of all memberships of one datum will be equal to the number of clusters. Therefore, in our proposal the sum of all the bounds (lower and upper) of the memberships of one datum to all the clusters must be a value between 2 and c, being c the number of clusters.

In this proposal, we want to minimize the weighted sum of the distances between every datum and every centroid using the membership values as weights, like in the k-means and FCM algorithms. However, in this case, the membership values are intervals. When there is a small doubt, the lower and upper bounds of each interval are quite similar, and they represent the value of the weight. On the contrary, when there is a big doubt whether a datum belongs to a cluster, we do not want its information to modify the centroid of this cluster. We want this weight to be small. If the length of this interval is large, it means that the lower bound is small. Hence, in both cases we can use the lower bound of the membership interval as the weight for the weighted sum.

It is also necessary to restrict the sum of the lengths of the membership intervals. If we do not do this, the system would always be minimized with interval memberships of $[0, 1]$ for all data and clusters.

Therefore, the objective function we want to minimize in our proposal is the following:

$$J = \frac{1}{a}\sum_{k=1}^{n}\sum_{i=1}^{c}(\underline{u_{ik}})^m ||x_k - v_i||_A^2 + \sum_{k=1}^{n}\sum_{i=1}^{c}(\overline{u_{ik}} - \underline{u_{ik}})^m$$

where x_k is the k-th datum to classify, v_i is the centroid of the cluster i, $[\underline{u_{ik}}, \overline{u_{ik}}]$ is the interval membership of datum k to the cluster i and m is a real value greater than 1.

The parameter $^1/_a$ allows to adjust the relative importance of both terms of the equation. It is necessary to remark that both terms do not need to be in the same scale: the first term depends on the distances between data whereas the second term is always managing values between 0 and 1. By tuning this parameter, we can obtain a similar solution to the FCM if the importance of the second term is bigger, or we can obtain a solution with much more doubt on it, if the first term is more important.

This function must fulfil the following constraints:

- There should be proper intervals. $\overline{u_{ik}} \geq \underline{u_{ik}}, \quad k = 1..n, i = 1..c$
- All the clusters must have at least one datum with lower membership bound greater than 0. $\sum_{k=1}^{n} \underline{u_{ik}} > 0, \quad i = 1..c$
- The sum of all the bounds of the memberships of one datum to all clusters must be between 2 and c. $2 \leq \sum_{i=1}^{c}(\underline{u_{ik}} + \overline{u_{ik}}) \leq c, \quad k = 1..n$

When there are two clusters in the dataset, this function can be minimized by using Lagrangian multipliers. In this way, we get an iterative algorithm similar to the FCM. From a random initialization, we update the interval memberships based on the data from the centroids.

$$\underline{u_{ik}} = \frac{2(2a)^{1/m-1}}{||x_k - v_i||^{2/m-1}\left[c + 2(2a)^{1/m-1}\sum_{j=1}^{c}\frac{1}{||x_k - v_j||^{2/m-1}}\right]}$$

$$\overline{u_{ik}} = \frac{2\left[||x_k - v_i||^{2/m-1} + (2a)^{1/m-1}\right]}{||x_k - v_i||^{2/m-1}\left[c + 2(2a)^{1/m-1}\sum_{j=1}^{c}\frac{1}{||x_k - v_j||^{2/m-1}}\right]}$$

for $k = 1..n$, $i = 1..c$. From these interval memberships, we update the centroids.

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}$$

for $i = 1..c$. The process finishes when the changes in the values are small enough.

IV. Numerical examples

In this section we show the performance of our proposal over some illustrative examples. To visualize the results in an easy way, all of them are 2-dimensional examples.

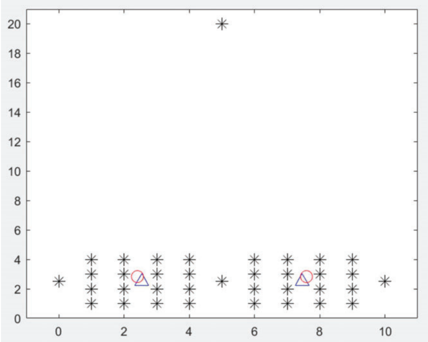

Fig. 2. Execution of our proposal and FCM over a dataset with one outlier. Dataset shown in black stars. Obtained centroids by our proposal in blue triangles. Obtained centroids by FCM in red circles.

This first example is the same dataset as the one in Figure 1. As we mentioned before, the FCM obtains distorted centroids because of the influence of the outlier. On the contrary, our proposal is able to avoid the bad influence and, therefore, to calculate the real centers of the clusters. In Figure 2 the final centroids of our proposal are in blue triangles and the ones obtained by the FCM are in red circles.

If we analyze the interval membership values, the datum located in the overlapped area (5, 2.5) has a membership value of [0.3488, 0.6512] to both clusters. I.e., the length is 0.3024. The outlier (5, 20) has a membership value of [0.0004, 0.9996] to both clusters and, thus, the length is 0.9991. We clearly observe that the algorithm is able to determine that these two data are different for the classification process. The influence of the outlier in the centroids is very small, due to the big length of the intervals. This is why the centroids are located in the real geometric centers of the clusters.

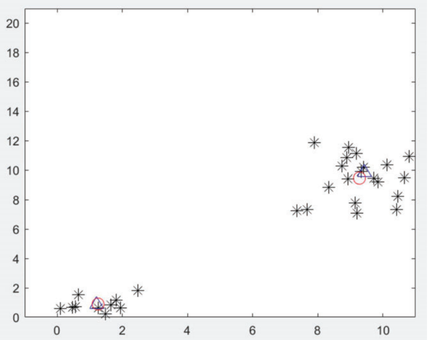

Fig. 3. Execution of the proposed algorithm and FCM over a dataset without any outlier. Original dataset in black stars. Centroids obtained by our proposal in blue triangles. Centroids obtained by FCM in red circles.

In the second example we start with a dataset made of two clusters without outliers (Figure 3). We observe that the centroids found by our algorithm (blue triangles) are very similar to the centroids found by the FCM (red circles).

If we add to the dataset three new data which are outliers, then the results of both algorithms are different (see Figure 4).

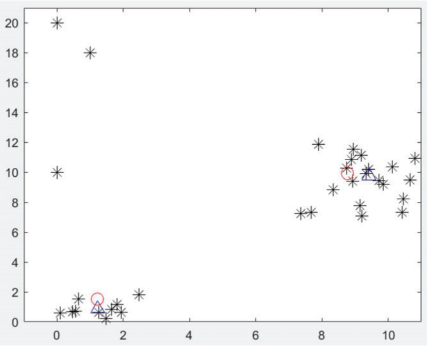

Fig. 4. Dataset made of three outliers and the original dataset from Figure 3. The three outliers are in positions (0,20), (1,18) and (0,10). Execution of the proposal and the FCM. Dataset to classify in black stars. Centroids obtained by our proposal in blue triangles. Centroids obtained by FCM in red circles.

We can visually check that the addition of three outliers has just slightly modified the centroids obtained by our algorithm. Numerically, the original centroids were

- Cluster 1→ (1.2107, 0.8496)
- Cluster 2→ (9.4268, 9.8496)

and now they are

- Cluster 1→ (1.2223, 0.8670)
- Cluster 2→ (9.4149, 9.7710)

However, in the new execution of the FCM the centroids are clearly influenced by the outliers. The centroid of the cluster 1 is moved upwards whereas the centroid of the cluster 2 is moved to the left. Numerically, the original centroids were

- Cluster 1→ (1.2531, 0.8985)
- Cluster 2→ (9.2796, 9.4647)

and now they are

- Cluster 1→ (1.2181, 1.5217)
- Cluster 2→ (8.7661, 9.9312)

One of the important aspects to bear in mind when applying our clustering proposal is the adjustment of the parameter $1/a$. This parameter is related to the length of the obtained membership intervals. If $1/a$ represents big values, then the length of the intervals obtained by the algorithm is also big. On the contrary, if the parameter $1/a$ represents small values, then the length of the obtained intervals are also small. It is needed to select a suitable value for every dataset.

Based on several examples, a proper value is when a takes the value of the percentile 10 or 15 of the distances of all the data.

V. Examples on image segmentation

In this section we apply our algorithm for the segmentation of images. The goal of image segmentation is to divide all the pixels in an image into several regions, each one representing one object in the image. In particular, it has to assign a label to every pixel in such a way that pixels with common characteristics share the same label.

The simplest method to segment an image is thresholding, which consists in selecting a value in such a way that all pixels whose intensity is greater than this value are labelled as object and all pixels whose intensity is lower than this value are labelled as background, or vice versa [7]. However, when we work with colour images or we consider more characteristics of a pixel than its intensity, we cannot simply apply thresholding. In these cases, one of the most used techniques to segment an image is clustering [5] [11].

When segmenting an image with the FCM, it is commonly accepted that each pixel should belong to the region with the biggest membership value. Following this convention, in Figure 5 we show an image from the Berkeley dataset (https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/) and in Figure 6 its segmentation using the FCM. We have used the average colour of each region for each cluster.

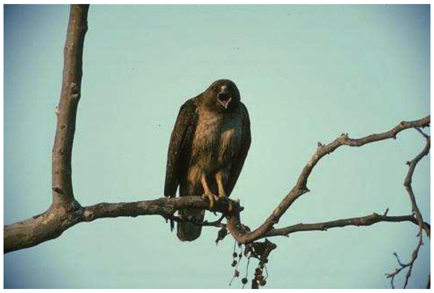

Fig. 5. Original image 42049 from Berkeley dataset.

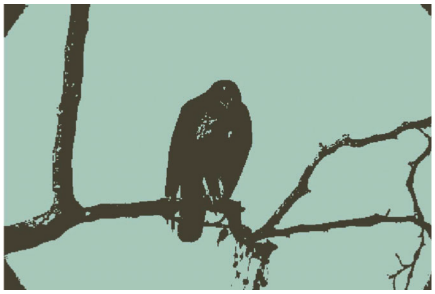

Fig. 6. Image 42049 segmented using FCM. Each clusters is coloured with the average colour of its pixels.

It is easily observable that the segmentation is not perfect. There are some pixels inside the bird and the branches which are assigned to the background. Moreover, the pixels in the four squares should belong to the background instead of to the object.

We have explained that one of the advantages of our proposed algorithm is that the clusters are not influenced by the outliers. But in this application we focus on the other advantage of our proposal: it is able to identify the doubt in the assignment of each datum. In our example, it is able to identify the doubt when assigning each pixel to a region of the image. Based on this, we can distinguish between the pixels that we are sure the region they belong to and the pixels that are hardly identifiable.

The pixels that the algorithm is not sure enough to classify can suffer an extra post process, taking into account also the information of their surrounding pixels, in order to get more compact regions.

Depending on the purpose of the segmentation, we can fix the threshold to classify one pixel as correctly segmented or as a doubtful one. This threshold is always based on the length of the membership intervals. For example, in Figure 7 we see several segmentations with several thresholds in the length of the intervals to be considered doubtful pixels.

The last one is very similar to the segmentation from the FCM. However, in the previous ones we see that our algorithm is able to identify as doubtful pixels the ones that will be wrongly classified if we do not consider the doubt. We think that this result adds a really useful information to the segmentation process.

We finally show another example that displays similar results. In Figure 8 we see another image from the same dataset and its segmentation using the FCM. In Figure 9 we show several segmentations using our algorithm and different levels of doubt. It is clear that most of the considered doubtful pixels are the ones that can be classified in a wrong way, i.e., the ones in the clouds and the ones in the left-down corner. Therefore, we can infer that it is important to take into account the extra information that our algorithm is able to provide, when segmenting images.

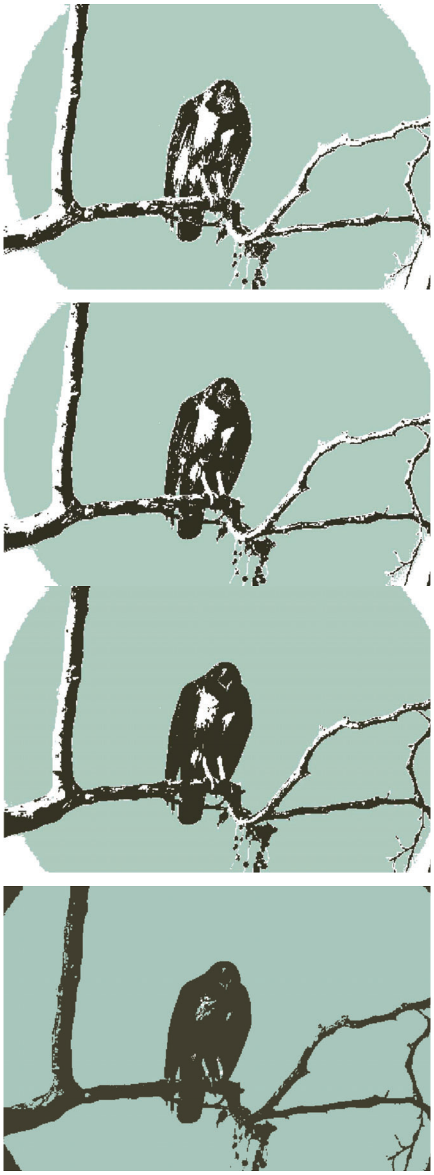

Fig. 7. Image 42049 segmented using our algorithm with different levels of doubt. Each clusters is coloured with the average colour of its pixels. The doubtful pixels are coloured in white.

VI. Conclusions

In this work we have presented a new clustering algorithm based on the Fuzzy Cluster Means. In our proposal we use an extension of fuzzy sets, the interval-valued fuzzy sets, in order to avoid the influence of the detected outliers. Moreover, we also provide a level of doubt in the assignment of every datum based on the length of the membership degrees. We have shown the correct performance of the algorithm over different datasets, with and without outliers. Finally, we have

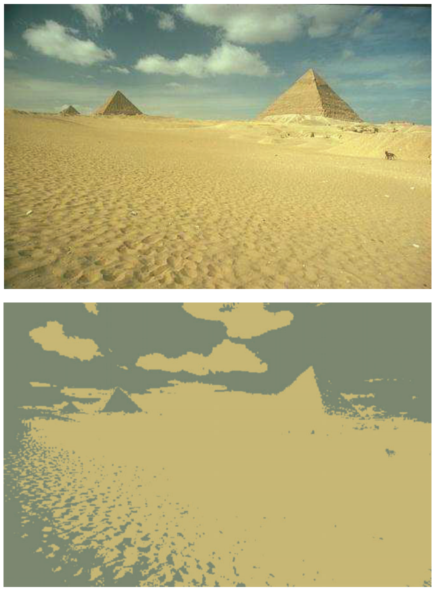

Fig. 8. Image 260058 and its segmentation with the FCM. Each clusters is coloured with the average colour of its pixels.

seen the application of the new algorithm to segment images. In this case, our algorithm is able to find similar solutions to the ones of the FCM, but it also gives extra information that is very relevant to improve the quality of the solutions.

REFERENCES

[1] J.C. Bezdek, “Pattern Recognition with Fuzzy Objective Function Algorithms” Plenum Press, 1981.

[2] H. Bustince, E. Barrenechea, M. Pagola, J. Fernandez, Z. Xu, B. Bedregal, J. Montero, H. Hagras, F. Herrera, B. De Baets, “A historical account of types of fuzzy sets and their relationships” IEEE Transactions on Fuzzy Systems 24, 179-194, 2016.

[3] K. Fu, J. Mui, “A survey on image segmentation” Pattern Recognition 13(1), 3–16, 1981.

[4] E.W. Forgy, “Cluster analysis of multivariate data: efficiency versus interpretability of classifications”, Biometrics 21, 768–769, 1965.

[5] A.K. Jain, M.N. Murty, P.J. Flynn, “Data clustering: A review” ACM Computing Surveys 31, 264-323, 1999.

[6] S.C. Johnson, “Hierarchical clustering schemes” Psychometrika 32, 241-254, 1967.

[7] A. Jurio, H. Bustince, M. Pagola, P. Couto, W. Pedrycz, “New measures of homogeneity for iamge processing: an application to fingerprint segmentation” Soft Computing 18, 1055-1066, 2013.

[8] L. Faufman, P.J. Rousseeuw, “CLustering by means of Medoids”, in Statistical Data Analysis Based on the L_1–Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416, 1987.

[9] S.P. Lloyd, “Least square quantization in PCM”, IEEE Transactions on Information Theory 28, 129-137, 1982.

[10] J.B. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations”, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1, 281-297, 1967.

[11] N.R. Pal, S.K. Pal, “A review on image segmentation techniques” Pattern Recognition 26, 1277-1294, 1993.

[12] J.H. Ward Jr., “Hierarchical Grouping to Optimize an Objective Function”, Journal of the American Statistical Association, 58, 236–244, 1963.

[13] L.A. Zadeh, “Fuzzy sets”, Information and Control 8, 338–353, 1965.

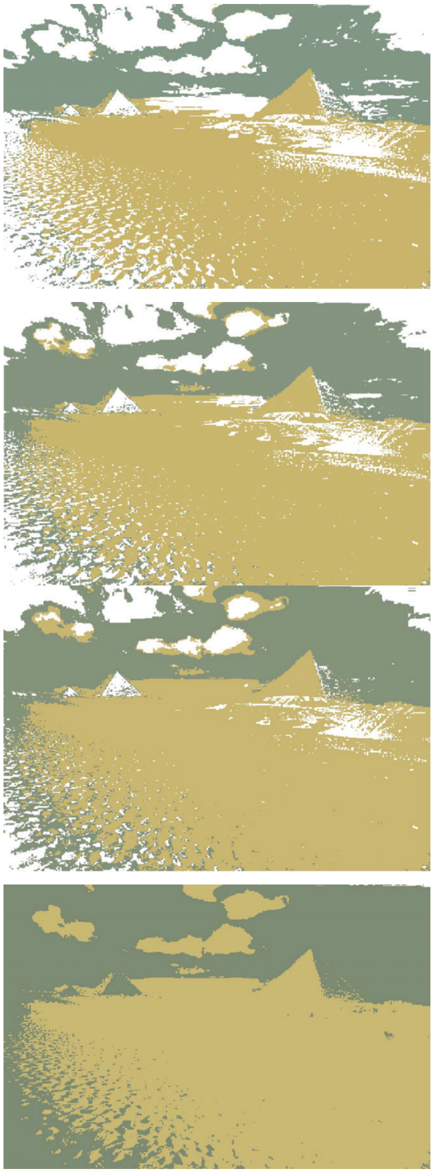

Fig. 9. Image 260058 segmented using our algorithm with different levels of doubt. Each clusters is coloured with the average colour of its pixels. The doubtful pixels are coloured in white.