Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 716753, 11 pages
http://dx.doi.org/10.1155/2013/716753

Research Article

Pattern Recognition in Numerical Data Sets and Color Images through the Typicality Based on the GKPFCM Clustering Algorithm

B. Ojeda-Magaña,[1,2] R. Ruelas,[2] M. A. Corona Nakamura,[2] D. W. Carr Finch,[2] and L. Gómez-Barba[1]

[1] *Systems Department, CUCEA, Guadalajara University, 45100 Zapopan, JAL, Mexico*
[2] *Department of Projects Engineering DIP-CUCEI, University of Guadalajara, 45101 Zapopan, JAL, Mexico*

Correspondence should be addressed to B. Ojeda-Magaña; benojed@hotmail.com

Received 19 July 2013; Accepted 25 October 2013

Academic Editor: Marco Perez-Cisneros

We take the concept of typicality from the field of cognitive psychology, and we apply the meaning to the interpretation of numerical data sets and color images through fuzzy clustering algorithms, particularly the GKPFCM, looking to get better information from the processed data. The Gustafson Kessel Possibilistic Fuzzy c-means (GKPFCM) is a hybrid algorithm that is based on a relative typicality (membership degree, Fuzzy c-means) and an absolute typicality (typicality value, Possibilistic c-means). Thus, using both typicalities makes it possible to learn and analyze data as well as to relate the results with the theory of prototypes. In order to demonstrate these results we use a synthetic data set and a digitized image of a glass, in a first example, and images from the Berkley database, in a second example. The results clearly demonstrate the advantages of the information obtained about numerical data sets, taking into account the different meaning of typicalities and the availability of both values with the clustering algorithm used. This approach allows the identification of small homogeneous regions, which are difficult to find.

1. Introduction

The objective of clustering algorithms is to find an internal structure in a numerical data set in order to separate it into n different groups or clusters, where the members of each group have a high similarity with its prototype (centroid, cluster center, signature, template, and code vector) and a high dissimilarity with the prototypes of the other groups. This justifies the existence of each one of the groups [1].

The clustering algorithms help us to get simplified representation of a numerical data set into n groups, as a way to get a better comprehension and knowledge of data. Clustering is one of the most popular unsupervised classification methods and has found many applications in pattern recognition, image segmentation, and data mining [2].

Also, the clustering algorithms that partition a given space in a hard, fuzzy, probabilistic, or possibilistic way, according to a data set and after a learning process, provide a set of prototypes as the most representative elements of each group. Classic k-means or hard c-means (HCM) [3] is probably one of the most used algorithms. This makes a hard partition of the objects in a predetermined number of clusters. On the other hand, the soft clustering techniques, as the fuzzy and the possibilistic, are characterized by the relaxation of the edge constraint allowing smoother transitions between the clusters. The soft borders address some particular challenges in many typical real-life applications where overlapping clusters, outliers, or uncertain cluster memberships can often be observed [4].

In this work we propose to take the interpretation of the typicality concept according to the cognitive psychology point of view, such that it is possible to obtain a more natural interpretation of data. Rosch and Mervis [5] have proposed a theory of prototypes for the classification of objects that belong to a semantic characteristic, taking into account their proximity to a prototype, according to a given criterion.

Thus, in each category there must be an internal resemblance among category members and an external dissimilarity, meaning that the similarity with the members of the other categories is low.

A prototype is based on the notion of typicality; all members belonging to the same category do not represent it in the same way; that is, some members are more typical than others. In [6] the authors were among those who first recognized this variation inside categories, when they discovered the concept of focal colors that underlie color identification. They show a similarity in color categorization among different languages, the so-called "basic colors"; that is, each color is represented by the best example (some nuances of colors are more representative than others), and inside a category of color there is no other basic color. For example, the scarlet color is not a basic color because it is in the red color category. Thus, colors are not uniform and they are represented by a center or better example and peripheral members. However, if we are interested in knowing the edge of each color category, for example, the threshold where the color stops being red but begins to be orange, we can use fuzzy sets. Zadeh [7] has proposed the theory of fuzzy sets in order to solve the problem of boundaries. With fuzzy sets, the members of a particular category have a membership degree in the $[0, 1]$ interval, where 1 is assigned to the most representative members and 0 to the elements that are not members of the category.

In this work we use the concepts of typicality and membership degrees in order to categorize linguistic concepts, looking for a better understanding of the information extracted from a numerical data set and digital images through segmentation using the Gustafson Kessel Possibilistic Fuzzy c-means (GKPFCM) clustering algorithm [8].

The remainder of the paper is organized as follows. In Section 2 we discuss the concepts of vagueness and typicality, as they are defined into the theory of prototypes, and take into account their differences. In Section 3 we present an analogy about the theory of prototypes and fuzzy clustering, putting in correspondence Rosch's typicality with relative and absolute typicalities provided by some fuzzy algorithms. At the end of this section, we present the GKPFCM clustering algorithm, its properties, and the possibility to get both typicalities with this algorithm. Section 4 contains a numerical data set and some color images, which we use as examples to show the advantages of better exploiting the concept of typicality. In Section 5 experimental results and discussion are presented. Finally in Section 6 we draw some conclusions.

2. Typicality and Vagueness

In the search of prototypes proposal by [9] typicality, generic-ity, and opacity are mentioned. In this work we use vagueness and typicality, both different but of interest as they help us to better define the characteristics of concepts. The typicality alludes to a degree to which the objects under study are considered good examples of the concept [10]. For example, in the category of birds, the dove is a typical case as it has the following characteristics: it can fly, it has feathers, it lays eggs, and it builds a nest in a tree. On the other side, an atypical case is the penguin because it satisfies only some characteristics, but not all. For this example, the prototype has the most common characteristics or the mean values of these characteristics.

The categories of some concepts can be vague or fuzzy; that is, there are objects whose membership to the category is uncertain, and this is not due to a lack of knowledge but to the lack of a clear rule defining the edges of the categories [9]. Some classical examples are the adjectives, *high* or *red*, or the nouns, *vegetable* or *chair*. Vagueness is mostly a question of truth (yes or no), and it represents a measure of correspondence of an object with a conceptual category. For some categories the edges are defined in an easier way, such as the category of birds. However, for other categories, as for the adjective *high*, the edges are not so easy to define, and therefore a membership degree in $[0, 1]$ is used.

For a better understanding of the differences between typicality and vagueness, we take an example from the work of [10]. In order to appreciate the difference between typicality and vagueness, they note that. . .*many people believe that penguins are an atypical case of birds, only a few doubt that they are birds in reality; in this case the typicality is involved, but not the vagueness*.

The membership degree of the penguin and the dove to the category of birds is 1. However, the dove is more typical than the penguin. For the concept of *high* there is no example with maximum typicality, due to the fact that in some contexts height can be increased infinitely. On the other hand, the membership degrees of the concept *high* represent the variation in the certainty degrees using values in 0 and 1, both included, and they provide the edge to determine at which height something is "*high*" and at which height it is not.

When applying clustering algorithms to numerical data sets it is required that the members of each group have similar values to each one of the features; otherwise, some members would be identified as members of other groups. On the other hand, in the theory of prototypes each member of a group must have all of the features and no matter if some of them have atypical values for some features, they continue to belong to the group. This means that clustering algorithms identify disjoint subgroups, each one containing typical and atypical data, and the only common data to the subgroups is noise. Figure 1 illustrates this relation between prototypes and results of clustering algorithms.

3. Relative and Absolute Typicality in the Clustering Algorithms

The partitional clustering algorithms have a great similarity to the theory of prototypes, although the latter is related to building categories about concepts whereas the clustering algorithms focus on the unsupervised classification of numerical data; however both approaches have the same objective.

In this section we give an interpretation of the typicality of the fuzzy clustering algorithms, based on a psychological and cognitive interpretation, as presented in the previous

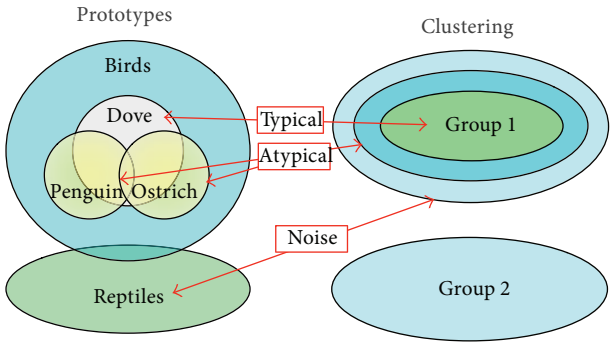

Figure 1: Illustration of the relation between prototype theory and fuzzy clustering.

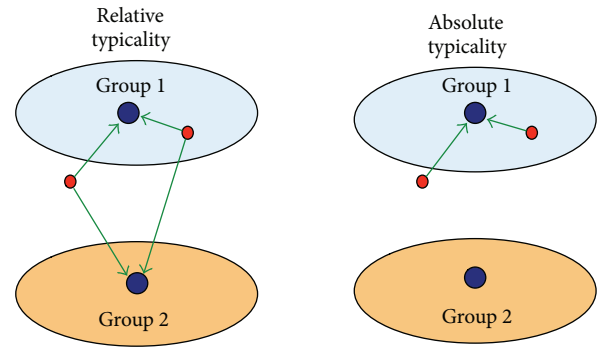

Figure 2: Relative typicality (FCM) and absolute typicality (PCM).

section and as a way to gain greater knowledge than usual from numerical data sets.

3.1. Fuzzy c-Means Clustering Algorithm (Relative Typicality). Ruspini [11] was the first to use fuzzy logic for clustering. After that, Dunn [12] proposed the first fuzzy clustering algorithm named Fuzzy c-means (FCM), with the parameter of fuzziness m equal to 2. Later on Bezdek [13] has generalized this algorithm. The FCM is an algorithm where the membership degree of each point to each fuzzy set A_i is calculated according to its prototype. The sum of all the membership degrees of each individual point must be equal to one. Therefore the degree of membership to a particular fuzzy set is influenced by the position of all the prototypes of the fuzzy sets, and that is the reason why Pal et al. [14] interpret the membership as a *relative typicality*.

With the FCM, the calculus of the membership degree of a point z_k to the fuzzy sets A_i is inversely proportional to the relative distance of this point to the prototypes (centers) of the fuzzy sets. Pal et al. [15] show a deficiency of the algorithm when there are several equidistant points from two prototypes, as the membership degrees to both fuzzy sets are the same, but the distance to the prototypes is different; one point is further than the other. These data must be handled with care as they do not represent both prototypes in the same way. Another disadvantage of the FCM algorithm is its sensitivity to noise or points far away from a concentration of prototypes.

3.2. Possibilistic c-Means Clustering Algorithm (Absolute Typicality). The Possibilistic c-means (PCM) clustering algorithm was proposed in [16], and its principal characteristic is the relaxation of the restriction that gives the relative typicality characteristic of the FCM. As a consequence, the PCM helps us to calculate a similarity degree between data points and each one of the prototypes, a value known as *absolute typicality* or simply typicality [14]. The nearest points to a prototype are identified as typical, whereas the furthest points are identified as atypical and noise if their typicality is zero or almost zero as mentioned in [17]. The PCM is very sensitive to the initial value of its parameters. Also, to avoid the coincidence of several prototypes, it is convenient to use the modified objective function proposed in [18], which contains a restriction resulting in repulsion of the prototypes and avoiding prototypes located at the same place.

3.3. Categories, Typicality, and Clustering Algorithms. The prototypes are selected as the best examples to represent categories or groups according to a given criterion, and they have the most important features. In the case of birds, for example, the dove is more typical than the ostrich and the penguin, because it has more features of a bird. However, ostriches and penguins are members of the category of birds. Therefore, there are an internal resemblance among the members of a group and an external dissimilarity to the members of other categories, even when several categories share some features, as it happens with birds and reptiles, as both kinds of these animals reproduce by eggs.

A similar situation happens with a numerical data set; that is, it is possible to take into account an external dissimilarity and an internal resemblance through distance measures, making it possible to quantify the similarity or the dissimilarity among patterns and prototypes of the different groups. Among the most used distance measures we have the Euclidean and the Mahalanobis distances. The former is used when the correlation of patterns is low. In this case the algorithm identifies groups with spherical forms. The last one is preferred for patterns with medium or high correlation. However, the correct decision on the selection of the distance measure depends on the available data and the statistical distribution of the features (attributes).

(i) *External Dissimilarity*. It results from fuzzy clustering algorithms, because the membership degree of a data point to a group depends on its membership degrees to the other groups. The data point is considered a member of the group to which it has the maximum value. In other words, it belongs to the nearest group.

(ii) *Internal Resemblance*. It results from algorithms such as the PCM, and represents the resemblance between a data point and a prototype. For this reason it is possible to establish thresholds in order to identify typical, atypical, and noisy data.

Thus, the FCM and the PCM algorithms provide information about a numerical data set (see Figure 2) and, taking into

account their difference, they could be useful to better understand the structure of data sets, as typical and atypical data can be differentiated. In fact, combining both algorithms makes it possible to estimate the prototypes of groups as a function of the internal resemblance and external dissimilarity. Reference [19] has proposed a hybrid algorithm, the Possibilistic Fuzzy c-means (PFCM), an improvement of the Fuzzy Possibilistic c-means (FPCM) [14], which provides all the advantages of the FCM and PCM algorithms and avoids some problems of these algorithms used separately. Other proposal is that of typicality degrees by a framework of prototype construction in a way that both categories common and discriminative features are characterized, by Lesot [20].

3.4. The GKPFCM Clustering Algorithm. Numerical data sets are complex because they generally have a lot of data, and they could be incomplete, or data could be imprecise, uncertain, and even vague in certain cases. Thus, the extracted knowledge from them must agree with the features and the data set, and this information must be easily understood by users. The first step is then to identify the prototypes, afterwards the groups, and finally Rosch's typicality of data.

The PFCM algorithm is based on the Euclidian distance and the identified clusters are constrained to spherical shapes, as described previously. So, in order to give more flexibility to the algorithm, such that the identified clusters are better adapted to the distribution of groups in the data set, we decided to use the Gustafson Kessel Possibilistic Fuzzy c-means (GKPFCM) algorithm proposed by Ojeda-Magana et al. [8] which is a combination of the improvement done by Babuška et al. [21] to the GK [22] algorithm, which we called GK-B, and the PFCM proposed by Pal et al. [15]. This algorithm is based on the Mahalanobis distance, according to the Gustafson and Kessel method [22], and clusters could also have ellipsoidal shapes.

The GKPFCM Algorithm. Given the data set $\mathbf{Z} = \{z_1, z_2, \ldots, z_n\}$ choose the number of cluster $1 < c < n$ and the termination tolerance $\varepsilon > 0$ (usually 10^{-3}).

(I) Provide an initial value for the prototypes (center) v_i, $i = 1, \ldots, c$. These are regularly obtained on a random basis.

(II) Find the value of the parameter δ_i. To do this we must run the FCM clustering algorithm [1], weighting exponent $m > 1$ (usually 2), and then use the following equation as proposed in [16, 23]:

$$\delta_i = K \frac{\sum_{k=1}^{N} \mu_{ik}^m \left\| z_k - v_i \right\|_A^2}{\sum_{k=1}^{N} \mu_{ik}^m}, \tag{1}$$

$K > 0$, although the most common option is $K = 1$.

(III) Choose standard parameters by GK-B [21] ρ_i (cluster volumes), β (condition number threshold), and γ (weighting parameter); regularly the first two parameters remain constant $\rho_i = 1$, $i = 1, \ldots, c$, and $\beta = 10^{15}$; the only parameter that changes is γ which takes values between 0 and 1.

(IV) Choose standard parameters by PFCM (a, b, m, and η). These parameters play an important role in the calculation of membership degrees, typicality values, and prototypes [15].

(V) Calculate the covariance matrices for each group according to

$$F_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m (z_k - v_i)(z_k - v_i)^T}{\sum_{k=1}^{N} (\mu_{ik})^m}, \quad 1 \le i \le c, \tag{2}$$

and estimate the covariance as suggested in [21]:

$$F_i = (1 - \gamma) F_i + \gamma \det(F_0)^{1/n} I, \tag{3}$$

where n represents the number of characteristics. Extract the eigenvalues λ_{ij} and eigenvectors ϕ_{ij}, of matrix F_i, and find $\lambda_{i,\max} = \max_j \lambda_{ij}$ and $\lambda_{i,\max} = \lambda_{ij}/\beta, \forall_j$ which satisfy $\lambda_{i,\max}/\lambda_{i,j} \ge \beta$.

Finally F_i is rebuilt using the following equation:

$$F_i = [\phi_{i,1}, \ldots, \phi_{i,n}] \operatorname{diag}(\lambda_{i,1}, \ldots, \lambda_{i,n}) [\phi_{i,1}, \ldots, \phi_{i,n}]^{-1}, \quad 1 \le i \le c. \tag{4}$$

Finally F_i is rebuilt using the following equation:

$$F_i = [\phi_{i,1}, \ldots, \phi_{i,n}] \operatorname{diag}(\lambda_{i,1}, \ldots, \lambda_{i,n}) [\phi_{i,1}, \ldots, \phi_{i,n}]^{-1}, \quad 1 \le i \le c. \tag{5}$$

(VI) Calculate the distance among data and prototypes:

$$D_{ikA_i}^2 = (z_k - v_i)^T \left[\rho_i \det(F_i)^{-1/n} F_i^{-1} \right] (z_k - v_i). \tag{6}$$

(VII) Determine the membership matrix $U = [\mu_{ik}]$ using the following equation:

$$\mu_{ik} = \left(\sum_{j=1}^{c} \left(\frac{D_{ikA_i}}{D_{jkA_i}} \right)^{2/(m-1)} \right)^{-1}, \quad 1 \le i \le c;\ 1 \le k \le n. \tag{7}$$

(VIII) Determine the typicality matrix $T = [t_{ik}]$ using the following equation:

$$t_{ik} = \frac{1}{1 + \left((b/\gamma_i) D_{ik_A}^2 \right)^{1/(\eta - 1)}}, \quad 1 \le i \le c;\ 1 \le k \le n. \tag{8}$$

(IX) Modify the prototypes v_i according to the following equation:

$$v_i = \frac{\sum_{k=1}^{N} (a\mu_{ik}^m + bt_{ik}^\eta) z_k}{\sum_{k=1}^{N} (a\mu_{ik}^m + bt_{ik}^\eta)}, \quad 1 \le i \le c. \tag{9}$$

(X) Verify that the error is within the proposal tolerance ε:

$$\left\| V_{\text{new}} - V_{\text{old}} \right\|_{\text{error}} \le \varepsilon. \tag{10}$$

(XI) If the error is greater than ε, return to step (V) or else stop.

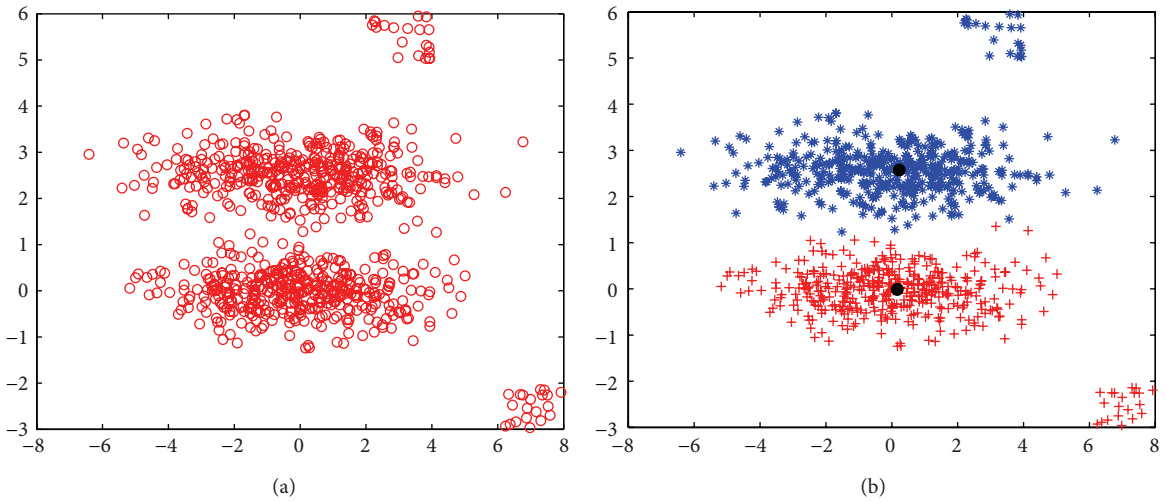

FIGURE 3: (a) Synthetic data set Z_{880}. (b) Partition of the data set with the GKPFCM algorithm.

4. Application of the Typicality Concept to Numerical Data Sets

As previously discussed, an example from the theory of prototypes is, for example, the category of birds, which include birds that can fly and birds that cannot. Obviously, the first kind of birds is more typical than the second one. If we try to qualify the characteristic of "flying," the less typical birds would have a very low score and this causes them to be further from the prototype. In the same way, when working with typicality through fuzzy clustering algorithms, the typicality can be used after groups are identified and data further away from the prototypes can be qualified as atypical. However, as we do not have constraints in a numerical data set, if data points are extremely far away from the prototypes, they could be considered as noise, in case they do not belong to another group.

Using a hybrid algorithm, such as the GKPFCM, it is possible to identify the groups and to use the typicality values of the algorithm in a way that different subsets of data in each group can be defined depending on how typical they are. In order to do this, it is necessary to establish thresholds dividing each group into typical, atypical, and noise data [24]. The number of thresholds to define is equal to the number of subsets we want to differentiate minus one. The division of the groups in this way is particularly interesting when we try to get information about a particular subset. For example, in the work of [25] they show how important the identification of the atypical data cloud is, in order to arrive at a medical diagnosis.

Applying the GKPFCM clustering algorithm we can identify groups and their prototypes. In this particular work and in order to maintain a more direct relationship with the theory of prototypes, we will only divide each group into typical and atypical data points, using the typicality values (matrix T of the algorithm). The approach to do this is as follows.

(I) Propose the parameters a, b, m, and η and the number of c clusters before the execution of the GKPFCM algorithm.

(II) Run the GKPFCM algorithm to estimate the relative typicality provided by the U matrix and the absolute typicality from the T matrix.

(III) From the U values, providing the external dissimilarity, find the boundaries among the clusters.

(IV) From the T values providing an internal dissimilarity and using thresholds, it is possible to differentiate among typical, atypical, or noise data.

In the next two subsections we apply this approach to a synthetic numerical data set and to a digitized image of a glass in a first example and four color images from the Berkeley database in a second example.

4.1. Application to a Synthetic Numerical Data Set. For the synthetic numerical data set we use similar data as that presented by Lesot and Kruse [26]; however we have added more noise. The data set consists of two Gaussian distribution clouds with 400 data points each and identical covariance matrices for both subsets. Forty points were added as noise located at the corners of the space where all data, including noise, is defined (see Figure 3(a)). For this work, this data set will be referred to as Z_{880}:

$$\sum_1 = \sum_2 = \begin{bmatrix} 4.47 & 0 \\ 0 & 0.22 \end{bmatrix}, \qquad \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 2.5 \end{bmatrix}. \tag{11}$$

To reduce the effects of noise in the prototype it is necessary to make a good choice of parameters a, b, m, and η for the GKPFCM algorithm. The parameters a and b have a great influence on the calculation of the prototype. In his work [19] recommended a value of b greater than the value of a, such that the prototypes are more influenced by

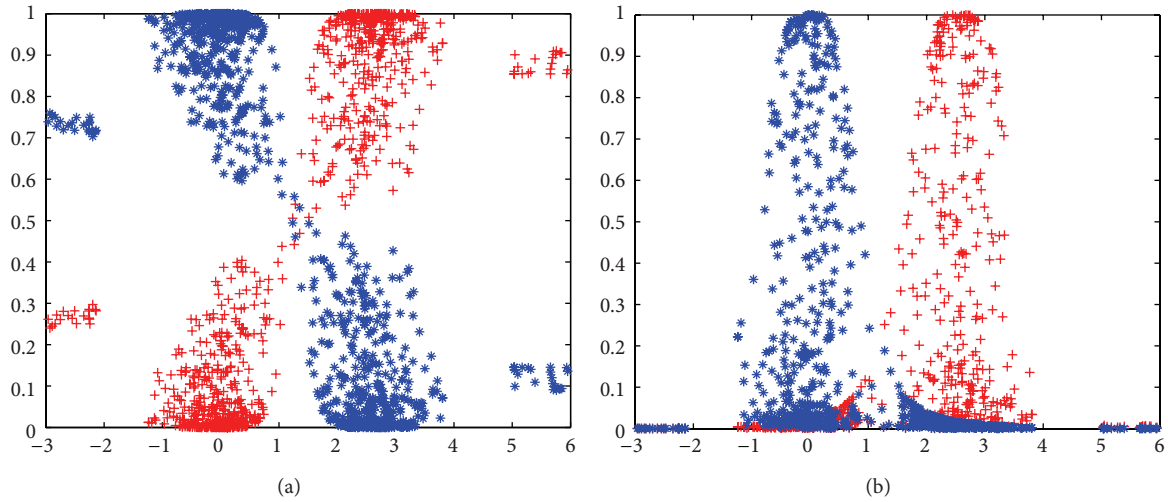

FIGURE 4: (a) Orthogonal projection of U matrix on the membership degree axis and (b) orthogonal projection of T matrix on the typicality value axis.

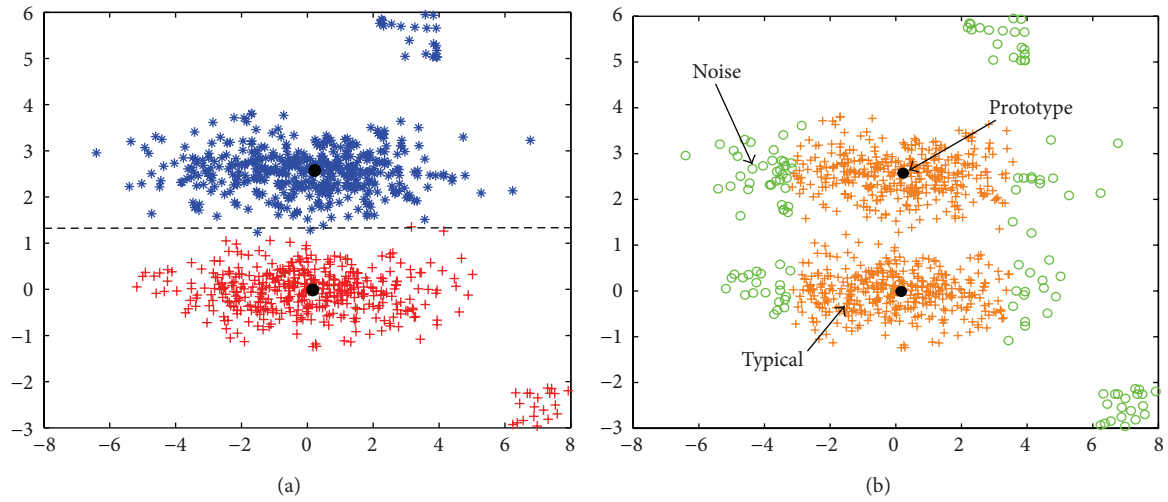

FIGURE 5: (a) Partition based on the external dissimilarity. (b) Partition based on the internal resemblance.

the membership values. On the other hand, it is recommended a small value for η and a value greater than 1 for m. Nevertheless, a too high value of m reduces the effect of membership of data to the clusters, and the algorithm behaves as a simple PCM.

For this work, we use the values $a = 1$, $b = 5$, $m = 2$, and $\eta = 2$ for the GKPFCM algorithm. The estimated values for the prototypes are

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0.1659 & -0.0112 \\ 0.2273 & 2.5721 \end{bmatrix} \quad \text{(see Figure 3(b))}. \tag{12}$$

Figure 3(a) shows the data set Z_{880} and Figure 3(b) the groups identified with the GKPFCM, which were identified correctly even in the presence of noise. In Figure 4 we can see the orthogonal projection of the membership degrees (U matrix) and typicality values (T matrix). Importantly, noise is included as an element of either group in the matrix of membership degrees, whereas noise is disregarded in the typicality values.

The presence of noise in the U matrix can be explained as follows: as it represents an *external dissimilarity*, some elements of the noise are more dissimilar to the further prototype and they are considered members of the group with the nearest prototype. In Figure 4(a) we can see that the membership degrees of outliers or noise data points are very high, as a consequence of the relation of data points with the prototypes. Figure 5(a) shows the division of clusters by a straight line based on the *external dissimilarity*. As the algorithm has the constraint that the sum of all the membership degrees of each point must be one, this line is defined by a membership degree of 0.5.

On the other hand, the evaluation of the *internal resemblance*, through T matrix of the algorithm, assigns a very low typicality value to the noisy points and, as can be seen in Figure 4(b), a threshold can be established such that noise

can be completely eliminated from the group. In this case the groups G_1 and G_2 would be divided in two subgroups as

$$G_1 = G_{1\text{typical}} \cup G_{1\text{noise}}, \qquad G_2 = G_{2\text{typical}} \cup G_{2\text{noise}}. \quad (13)$$

For example, selecting a threshold of $\alpha = 0.01$, the subgroup of typical data in each group G1 and G2 corresponds to the points with typicality values greater than or equal to α, and points with lower typicality values are considered members of the noise subgroup. The proposed threshold for this example has been selected empirically and by observing Figure 4(b). However, when only attempting to eliminate noise from the groups, the value of this parameter will be very low. Figure 5(a) shows a typical partition of the space according to the groups identified, whereas Figure 5(b) shows the partition of the groups according to the typicality of data. In this particular case the value of the threshold can be decreased a little more in order to include some points that could be considered typical or data to both sides of the subgroups of typical data.

As it is shown in Figure 5(b), this is similar to the theory of prototypes where each object having values in one or several characteristics, which are very different from the mean values, is considered atypical.

4.2. Application in Color Image Segmentation. One of the most challenging problems in computer vision is that of the segmentation of images, as these ones must be divided in regions of interest, and the objects in the scene must be clearly identified. The aim of segmentation is to divide an image into nonoverlapping regions that are homogeneous in some features such as gray levels of pixels, color, texture, and depth or even a combination of some of these. The level to which the subdivision is carried out depends on the problem being solved.

Partitional clustering algorithms are considered for image segmentation, because of the great similarity between segmentation and clustering, although clustering was developed for feature space, whereas segmentation was developed for the spatial domain of an image.

From a general point of view, the segmentation of images can be divided into two types: region based or edge based. In this work we focus on the former approach, where the objects in the image result from homogeneous regions inside the RGB color space. We use two examples to show these results, the first one concerning the analysis of the differences with the typicality of Rosch, whereas the second one is based on the absolute typicality in order to improve the results of segmentation.

As a first test we have used an image of a glass where we identify this object and separate it from the background. As can be seen in Figure 6, the glass is one color though strongly affected by illumination producing a not very homogeneous image with some pixels very similar to those of the background.

The clustering algorithms work in the features space; the RGB color space for this work. For the final result we use the mean value for each cluster. So, the quality of the segmentation can be evaluated through the similarity between the real colors of the image and the identified colors based on the concept of typicality.

Figure 6: Original image of a glass.

In order to compare the results, two clustering algorithms have been used. The first algorithm is the FCM, used to identify two clusters in the image. The algorithm provides the U matrix of membership degrees or the relative typicalities, wherewith the RGB space is partitioned.

As can be seen in Figure 7, the corresponding results are relatively good, as the glass is clearly identified. However, some regions of the glass are affected by the illumination and are related to the background. This result is a consequence of the relative typicality and the Euclidean measure on which the algorithm is based. In this case, the clusters are constrained to spherical shapes, and the case of study has shapes more closely related to hyperellipsoidal forms.

As we are interested in the typicality concept for knowledge discovery in numerical data sets, we have used the GKPFCM algorithm such that the absolute and the relative typicalities are available, and the results can be improved using them. Figure 7 shows the results, those concerning the relative typicality and those concerning the absolute typicality.

Comparing the results of the previous images, the FCM gives comparable or even better results than the GKPFCM based on the relative typicality. However, this last algorithm uses a measure that allows better adaptation of the clusters to the distribution of data, and hence the results provide more homogeneous regions. This can be seen more easily through the absolute typicality whose results are shown in Figure 7. In this case there is a better discrimination between the glass and the background, and the pixels affected by the illumination, though atypical, are associated here with the glass, and not with the background as was the case regarding the relative typicality, no matter if it was obtained from the FCM or the GKPFCM clustering algorithms.

Regarding the results of the absolute typicality, this provides additional information that can be used to get more homogeneous segmented regions and to identify the atypical pixels inside them. This process has been called *subsegmentation* [17], where a threshold α must be established

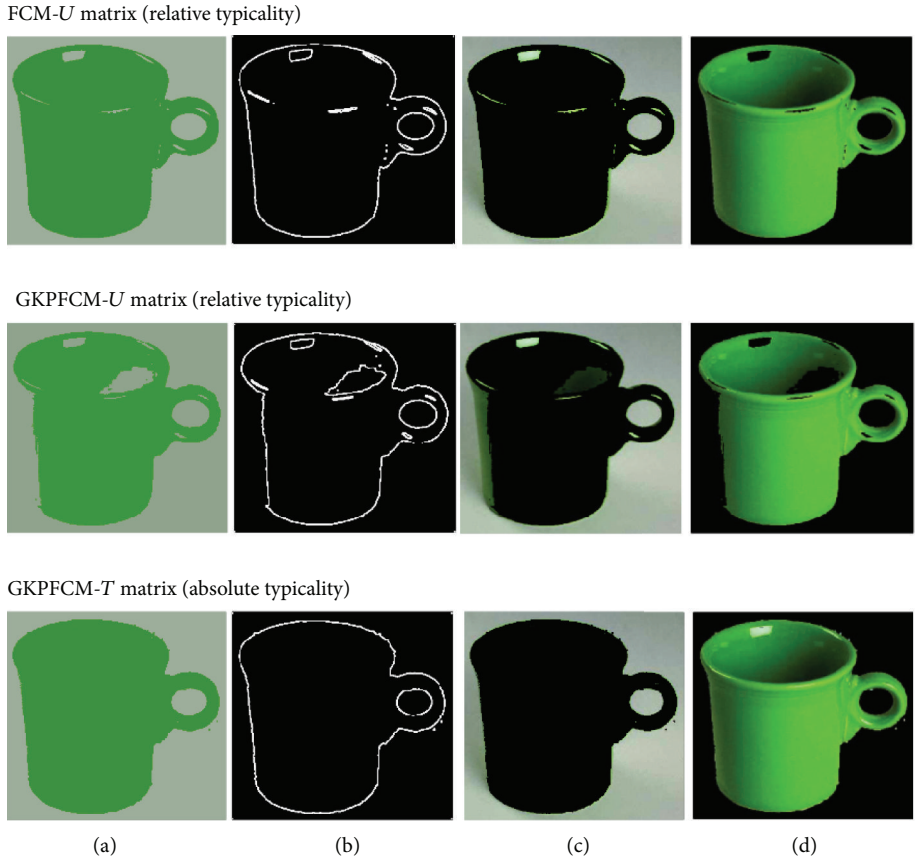

FIGURE 7: Segmentation results of the clustering algorithm: FCM at the upper image, GKPFCM (U matrix) at the middle, and GKPFCM (T matrix) at the bottom. (a) Segmentation in two regions. (b) Identified edges, (c) representation of the background, and (d) representation of the glass by the corresponding identified pixels.

in order to separate the typical and the atypical pixels. This parameter takes a value in the interval $(0, 1)$.

Four images from the Berkeley database (see Figure 8) have been selected, as well as four clustering algorithms, the k-means, the FCM, the Gustafson-Kessel algorithm with an improvement proposed by Babuška (GK-B), and the GKPFCM.

The first image from the Berkeley database concerns an airplane where we can find three objects: the airplane (region_1), the clouds (region_2), and the sky (region_3). The segmentation with the k-means and the FCM results in an airplane where almost half of its pixels, region_1, belong to the sky, region_3.

On the other hand, the GK-B gives better results than in the previous cases, as the segmented regions represent in a more approximated way the objects in the image. The GKPFCM algorithm was also applied, and the corresponding results are analyzed according to the relative and absolute typicalities. The results are not so good in the first case as the region_1 is totally associated with region_2. However, the results are much better when the absolute typicality is used, and the threshold is established at $\alpha = 0.05$. The atypical pixels in each region, that is, subregion_2, subregion_3, and subregion_4, help enhance the results of identification of the objects, as there are more details associated to their corresponding objects, particularly the airplane. So, in this case we have 3 regions and 3 subregions.

The only drawback of the previous results is that the atypical pixels, in the RGB features space, are located at both extremes of the ellipsoids. This depends on the particular distribution of pixels in the features space and the value assigned to the threshold. For the airplane identification, the atypical pixels allow much better identification of this object. However, in this case we also have the lighter atypical pixels of the sky. These can be seen at the left lower side of the resulting image.

The image of the field was also segmented in three regions, the stones (region_1), the bushes (region_2), and the earth (region_3). No method among the k-means, the FCM, and the GK-B was able to detect region_2, except the GKPFCM which gives better results when a threshold $\alpha = 0.06$ was used. In this particular case, the atypical pixels of region_2 correspond to the bushes and are represented by sub-region_4.

The image of the horses was segmented in four classes, and we got good results with all the algorithms. Nevertheless, if the images are observed carefully, there are a lot of details that are lost. Take for example the white line in front of the horse's head. The algorithms k-means, FCM, and GK-B

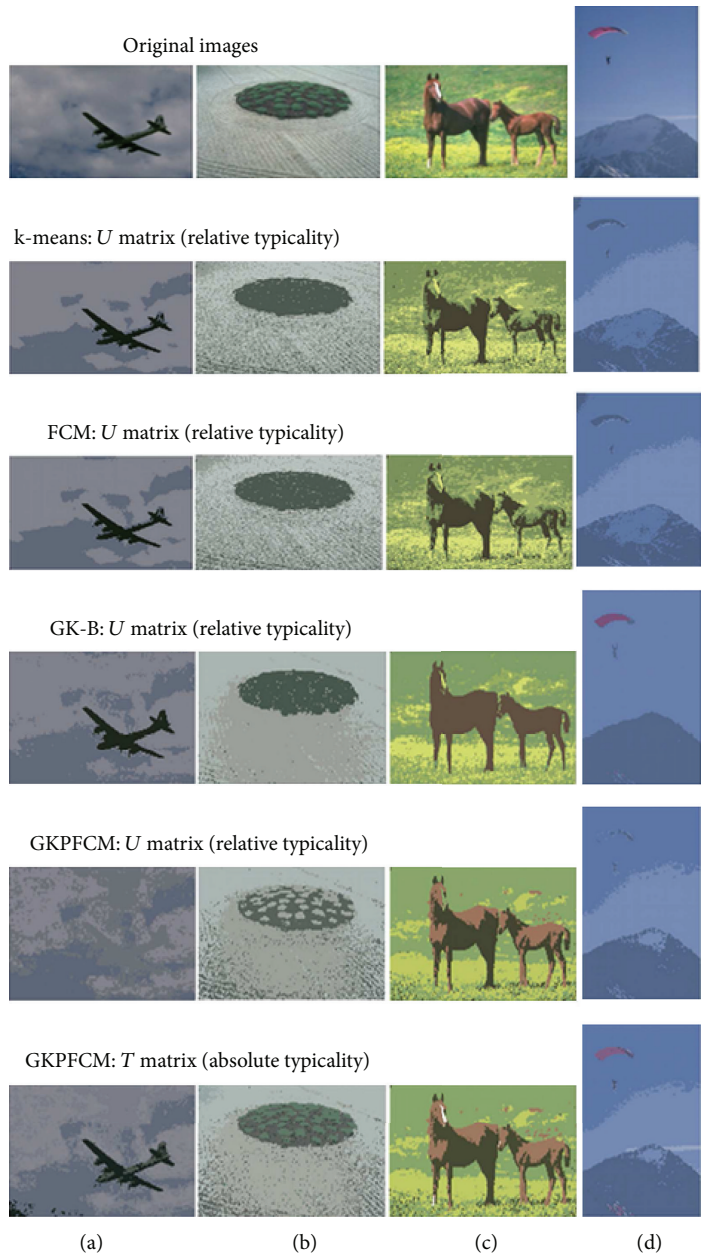

Figure 8: Segmentation results of the four images from the Berkeley database, with four partitional clustering algorithms.

associate this object to other regions and, even if the number of classes is increased, this region is not detected. On the other hand, the GKPFCM with a very low threshold $\alpha = 0.001$ identifies this object, as well as the atypical pixels of the horse's legs.

The last image was that of a parachute. This image was also segmented in three regions. Here we find that the k-means, the FCM, and the GKPFCM-U are being incapable of correctly identifying the parachute. An exception must be made here, as was done for the GK-B and the GKPFCM. For the last algorithm a threshold $\alpha = 0.02$ has been experimentally found, which gives the better results. Besides, this algorithm also detects the cloud that is behind the mountain. This object is very difficult to identify by the partitional clustering algorithms.

From the results of the previous examples we can find that the absolute typicality is clear, especially for the segmentation of color images. In this case the typicality value could be viewed as a means to apply a homogenization procedure, as the division of each region in typical and atypical pixels gives results with the most uniform pixels.

5. Result and Discussion

One of the major challenges when looking for patterns in data sets and images is to find the most homogeneous groups

in a feature space. In this work, we have used the GKPFCM clustering algorithm which meets the following features.

(i) It adapts better to the natural shape of the data, that is, convex hyperellipsoids.

(ii) It provides *the external dissimilarity* (U matrix) and *internal resemblance* (T matrix), for every prototype.

(iii) It has parameters, whose particular function has been explained, to give more importance to any dissimilarities or resemblances.

(iv) With the internal resemblance provided by the algorithm, we can identify the atypical data of each class and more homogeneous regions can be formed without the need to increase the number of groups.

The first point is a result of using the Mahalanobis distance in the GKPFCM algorithm. This leads to the achievement of a better partition of the feature space. However, the drawback of the algorithm is its limitation to the identification of nonconvex groups; that is, as the nonconvexity becomes more severe, the quality of results diminishes.

The second point follows from the availability of the relative and absolute typicalities, which are directly associated, through the relationship between fuzzy clustering and the theory of prototypes, with an external dissimilarity and an internal resemblance. This allows for a better categorization, knowing, additionally, the degree of typicality of each object to each category. The simplest case is the k-means, which provides a discrete relative typicality, and the FCM that provides a continuum relative typicality in the interval $[0, 1]$.

Through the four parameters of the GKPFCM algorithm (a, b, m, and η), we have the possibility to establish a compromise between the typicalities in absolute (m and η) and relative (a and b) ways. The values of the parameters used in this work are $a = 1$, $b = 3$, $m = 2$, and $\eta = 3$, giving more importance to the internal resemblance.

With the absolute typicality, or the internal resemblance, it has been possible to identify the atypical data inside each cluster. The result is a set of typical data and a set of atypical data, the former representing a more homogeneous region in this case. Nevertheless, the atypical set could not necessarily be homogeneous as the data could be located at both extremes of the corresponding ellipsoid in the RGB color space.

In this work the images are in the RGB color space, even though there are other proposed color spaces in order to improve the results of image processing [27, 28]. This is not a problem here as the GKPFCM and the GK-B algorithms achieve very good identification of groups. On the other hand, the k-means and the FCM give less satisfactory results due mainly to the inadequacy of the shapes produced by the Euclidean distance, the measure used to determine the membership of each datum. By contrast, the other algorithms use a measure that generates forms better adapted to the shapes of the objects to be identified. The hybrid GKPFCM algorithm, however, provides the best results as we can get extra information, which helps to obtain a better partition of the feature space. This can be clearly observed in the different results of the image segmentation.

Taking advantage of the typicality values, the absolute typicality, or the internal resemblance, we are able to enhance the segmentation process and to get more homogeneous regions, at least for the typical data. This is a promising result, as we are able to find very small homogeneous regions, which are very difficult to identify, even if the number of regions to segment is increased to a very large value.

6. Conclusions

Categorizing data into concepts, analogous to the theory of prototypes, allows us to understand the problem of unsupervised classification (also known as clustering) and to propose an approach to look for particular points inside each of the categories. This was shown using a synthetic numerical data set, a digitized image of a glass, and four images from the Berkeley database. In these cases, the external dissimilarity and internal resemblance are used in a better way and more information can be obtained from the same data compared to a classical approach. The existence of hybrid algorithms, as the GKPFCM or the PFCM, allows us to get both values at the same time, providing us with more information about the internal structure of data sets. In this work we have related classifications made by human beings to those made by automatic algorithms. This approach is very interesting when we try to look for special cases inside an image. For this reason we have attempted to join the theory of prototypes and the partitional clustering algorithms.

Acknowledgments

The authors wish to thank The National Council for Science and Technology (CONACyT) in Mexico and the Departamento de Sistemas de Información (CUCEA) and the Departamento de Ingeniería de Proyectos (CUCEI) at the Universidad de Guadalajara for the help provided to complete this study.

References

[1] J. C. Bezdek, J. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Springer, Boston, Mass, USA, 1999.

[2] R. Xu and D. Wunsch II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[3] J. MacQueen, "Some method for classification and analysis of multivariate observation," in *Proceedings of the 5th Berkley Symposium on Mathematical Statistic and Probability*, L. LeCam and J. Neyman, Eds., vol. 1, pp. 281–297, University of California Press, Berkeley, Calif, USA, 1967.

[4] P. Georg, F. Crespo, P. Lingras, and R. Weber, "Soft clustering—fuzzy and rough approach and their extension and derivates," *International Journal of Approximate Reasoning*, vol. 54, no. 2, pp. 307–322, 2013.

[5] E. Rosch and C. B. Mervis, "Family resemblances: studies in the internal structure of categories," *Cognitive Psychology*, vol. 7, no. 4, pp. 573–605, 1975.

[6] B. Berlin and P. Kay, *Basic Color Terms: Their Universality and Evolution*, University of California Press, 1968.

[7] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[8] B. Ojeda-Magana, R. Ruelas, M. A. Corona-Nakamura, and D. Andina, "An improvement to the possibilistic fuzzy c-means clustering algorithm," in *Image Procesing and Biomedicine*, vol. 20 of *Intelligent Automation and Soft Computing*, pp. 585–592, TSI Press, 2006.

[9] J. A. Hampton, "Typicality, graded membership, and vagueness," *Cognitive Science*, vol. 31, no. 3, pp. 355–384, 2007.

[10] D. Osherson and E. E. Smith, "On typicality and vagueness," *Cognition*, vol. 64, no. 2, pp. 189–206, 1997.

[11] E. H. Ruspini, "Numerical methods for fuzzy clustering," *Information Sciences*, vol. 2, no. 3, pp. 319–350, 1970.

[12] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[13] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, NY, USA, 1981.

[14] N. R. Pal, K. Pal, and J. C. Bezdek, "Mixed c-means clustering model," in *Proceedings of the 6th IEEE International Conference on Fussy Systems*, pp. 11–21, July 1997.

[15] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A new hybrid c-means clustering model," in *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 179–184, July 2004.

[16] R. Krishnapuram and J. M. Keller, "Possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.

[17] B. Ojeda-Magaña, J. Quintanilla-Domínguez, R. Ruelas, and D. Andina, "Images sub-segmentation with the PFCM clustering algorithm," in *Proceedings of the 7th IEEE International Conference on Industrial Informatics (INDIN '09)*, pp. 499–503, June 2009.

[18] H. Timm, C. Borgelt, C. Döring, and R. Kruse, "An extension to possibilistic fuzzy cluster analysis," *Fuzzy Sets and Systems*, vol. 147, no. 1, pp. 3–16, 2004.

[19] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517–530, 2005.

[20] M. J. Lesot, "Typicality-based clustering," *International Journal of Information Technology and Intelligent Computing*, vol. 1, no. 2, pp. 279–292, 2006.

[21] R. Babuška, P. J. van der Veen, and U. Kaymak, "Improved covariance estimation for Gustafson-Kessel clustering," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 1081–1085, May 2002.

[22] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proceedings of the IEEE Conference on Decision and Control Including the 17th Symposium on Adaptive Processes*, pp. 761–766, January 1979.

[23] R. Krishnapuram and J. M. Keller, "The possibilistic C-means algorithm: insights and recommendations," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 3, pp. 385–393, 1996.

[24] B. Ojeda-Magaña, R. Ruelas, F. S. Buendía-Buendía, and D. Andina, "A greater knowledge extraction coded as fuzzy rules and based on the fuzzy and typicality degrees of the GKPFCM clustering algorithm," *Intelligent Automation and Soft Computing*, vol. 15, no. 4, pp. 555–571, 2009.

[25] J. Quintanilla-Domínguez, B. Ojeda-Magaña, A. Marcano-Cedeño, M. G. Cortina-Januchs, A. Vega-Corona, and D. Andina, "Improvement for detection of microcalcifications through clustering algorithms and artificial neural networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, article 91, 11 pages, 2011.

[26] M. J. Lesot and R. Kruse, "Gustafson-Kessel-like clustering algorithm based on typicality degree," in *International Conference on Information Processing and Management of Uncertainty in Knowledge Based Systems*, pp. 117–130, World Scientific, London, UK, 2006.

[27] O. Lézoray and C. Charrier, "Color image segmentation using morphological clustering and fusion with automatic scale selection," *Pattern Recognition Letters*, vol. 30, no. 4, pp. 397–406, 2009.

[28] T. D. Nguyen and G. Lee, "Color image segmentation using tensor voting based color clustering," *Pattern Recognition Letters*, vol. 33, no. 5, pp. 605–614, 2012.