Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases

Nikos Pelekis\(^1\), Dimitris K. Iakovidis\(^2\), Evangelos E. Kotsifakos\(^1\), Haralampos Karanikas\(^2\), Ioannis Kopanakis\(^3\)

\(^1\) Univ. of Piraeus, Piraeus, Greece, \{npelekis, ek\}@unipi.gr

\(^2\) National & Kapodistrian Univ. of Athens, Knowledge Management Lab., Athens, Greece
dimitris.iakovidis@ieee.org

\(^3\) Technological Educational Institute of Crete, Heraklion Crete, Greece
i.kopanakis@emark.teicrete.gr

Abstract: Content-based information retrieval (IR) involves low-level feature extraction and utilize similarity search methods applied either in the feature space or in derived higher-level semantic spaces. These methods assume that similarity is measured by accounting only the degree in which two entities are related, ignoring the hesitancy introduced by the degree in which they are unrelated. Aiming at semantically relevant IR from cultural databases, this paper proposes a novel intuitionistic fuzzy clustering scheme based on intuitive features and a similarity defined over higher-level patterns.

Keywords: Clustering, Content-based information retrieval, Intuitionistic fuzzy sets, Cultural databases

Reference:

Biographical notes:

1 Introduction

Recently, our capabilities of both generating and collecting data have increased rapidly. Consequently, data mining has become a research area with increasing importance. Among the unsupervised data mining tasks, clustering possesses a pivot position. Clustering is the organization of objects into separate collections, or differently, the partitioning of a data set into groups (clusters), so that the data vectors in each cluster are proximate according to some distance measure. Clustering has been applied to statistical data analysis, machine learning, pattern recognition, image analysis and bioinformatics.

Lately, clustering has been successfully utilized in retrieval tasks (Stehling, Falcao, and Nascimento, 2001; Zhang, R., Zhang, Z.(M), 2002; Carson, Belongie, Greenspan and Malik, 2002; Chen, Wang, Krovetz, 2005; Greenspan, Pinhas, 2007; Rahman,
Bhattacharya and Desai, 2007; Iakovidis et al., 2006; Iakovidis et al., 2007). The main idea behind this approach is that a cluster provides a content rich summarization of the included data vectors; as such it is reasonable to try to speed up and improve the search process by comparing the semantically higher level patterns represented by the clusters, instead of performing exhaustive comparisons over the data vectors. The key word in the previous discussion that makes the use of clustering an attractive solution in retrieval tasks is content. For non-computer researchers content usually refer to meaning. Considering the document retrieval case, the content of a manuscript is not the keywords associated with it, but the meaning of the text in the document. Similarly, for a painting the content for a computer scientist is a mixture of its shape, texture and colour features, while for an art critic might be the painting technique used.

In clustering approaches for such Content-Based Retrieval (CBR) applications, the semantic similarity is usually defined over a set of measurable features. The relevance between a query object and objects in the database is evaluated according to the similarity measure computed on these features. However, it is self-evident that there is a potential inherent ambiguity concerning the term “content” when this is quantified by the measurable characteristics. On the other hand, most clustering schemes usually try to avoid vague, imprecise or uncertain information, because it is considered as having a negative influence in the knowledge discovery process.

There are approaches like the Fuzzy C-Means (FCM) (Bezdek, Ehrlich and Full, 1984) and its variants (Thitimajshima, 2000; Yong, Chongxun, Pan, 2004; Chumsamrong, Thitimajshima and Rangsanseri, 2000; Leski, 2001) which they try to incorporate fuzzy information in the clustering process, however they usually do so by quantifying the degree of membership of each data vector to a cluster. This is typically performed using the distance measure between the data vector and a representative of a cluster (i.e. its centroid). We should note at this point that data vectors often are multi-dimensional objects, where each dimension (i.e. feature) may model an independent property, which may have been extracted by different techniques. This implies that each one of the features introduces fuzziness that we should also take into account when calculating the distance measure between data objects. As already stated, fuzzy clustering approaches work at the level of vector and they do not utilize qualitative information at the constitutional feature level. This paper accepts the challenge to deal with such kind of information, by introducing a methodology, which deliberately makes use of it. Our methodological framework can be applied to any CBR system, examples of which include Content-Based Image Retrieval (CBIR) systems, document retrieval or web search engines etc. In the sequel, we adopt as the case of our study the CBIR systems, so every paradigm trying to exemplify our approach will be taken from the image domain.

The computation of the feature vectors is a critical step in CBIR systems. However, this initial image processing phase is inevitably associated with the inclusion of fuzziness in each of the extracted features. This is due to that images themselves capture indeterminacy imposed by limitations of the imaging and acquisition mechanisms. This may be translated to existence of hesitancy regarding for example the intensity level or the grayness of the image. We argue that it is essential to quantify this hesitancy, which is useful semantic information, and utilize it appropriately in the retrieval process.

Fuzzy set theory has long been introduced to handle such vagueness by generalizing the notion of membership in a set. Motivated from Zadeh’s definition of fuzzy sets theory (Zadeh, 1965), different types of fuzzy sets (FS) have been proposed in the research
community. Among them, intuitionistic fuzzy sets (IFS) introduced by Atanassov (Atanassov 1986, 1989, 1994a, 1994b, 1999) proved to be an appropriate means to model the hesitancy coming from imprecise and/or imperfect information. Intuitionistic fuzzy sets use two characteristic functions, namely the membership and the non-membership, depicting the belongingness or non-belongingness of an object to the IFS, respectively.

The wide use of fuzzy logic in image processing is mostly due to the ability of fuzzy sets to cope with qualitative measures, such as the texture, the color intensity or the contrast, by modelling the ambiguity and vagueness regularly present in imagery databases. Conversely, IFS using membership and non-membership functions set up an indeterminacy index, which models the hesitancy of deciding the degree to which an element satisfies a property. Our intention is firstly to define an intuitionistic fuzzification process of the features extracted from images, and secondly to exploit the modelled hesitancy in a clustering scheme that appropriately makes use of it.

For clustering, we base on Fuzzy C-Means (FCM) introduced by Bezdek, Ehrlich and Full (1984) and propose a novel modification of the algorithm that takes into account data accompanied with intuitionistic membership and non-membership information. FCM partitions a dataset by only examining the data vectors and thus it overlooks the fact that vectors may have qualitative information defined per feature (i.e. dimension). For instance, a data point \( x_k \) may not be just a \( p \)-dimensional vector \( (x_{k1}, \ldots, x_{kp}) \), but a \( p \)-dimensional vector of triplets \( (\mu_{k1}, \gamma_{k1}), \ldots, (\mu_{kp}, \gamma_{kp}) \), where for each feature \( l \), we further have qualitative information given as the intuitionistic membership \( \mu_{kl} \) and non-membership \( \gamma_{kl} \) of the data point to the feature. Further to that FCM does not employ such qualitative information; it constrainedly defines its distance measure upon the data vectors. Having as goal to maximize the exploitation of the extracted content, we additionally define a novel distance metric operating solely on the intuitionistic fuzzy vectors and appropriately integrate it into our clustering scheme.

The major contributions of this paper are the following:

- We propose a novel methodological framework for content-based retrieval, which can be applied in three successive steps. The first step lies on the transformation of a dataset in terms of IFS theory, the second step requires a mean to assess the similarity between IFS, while the third step involves a clustering algorithm that is able to work with and exploit the qualitative advantages of the IFS.

Selecting as case study for the application of our framework the imagery domain:

- We introduce an intuitionistic fuzzy representation of color digital images.
- We define a novel similarity measure between IFS and prove its superiority over other proposed metrics.
- We modify the well-known FCM clustering algorithm and come up with a clustering scheme that may partition intuitionistic fuzzy datasets in a content-based and semantic-rich manner.
- In order to thoroughly evaluate our approach we conduct a comprehensive set of experiments using cultural heritage images derived from the repository of the Foundation of Hellenic World (FHW).

The rest of the paper is structured as follows: The related work of the fields previously mentioned is discussed in Section 2. In Section 3, some basic notions and the definition of IFS are reviewed. In Section 4, we present our suggestion for intuitionistic fuzzification of images, while in Section 5 we introduce and assess our similarity measure between intuitionistic fuzzy sets. In Section 6 we present the results of our
Nikos Pelekis, Dimitris K. Iakovidis, Evangelos E. Kotsifakos, Ioannis Kopanakis

experimental study on the proposed clustering scheme in cultural imagery databases. Section 7 concludes the paper and provides ideas for future work.

2 Related work

As the amount of pictorial information stored in both local and public medical databases is growing, efficient image indexing and retrieval becomes a necessity.

In the last decade the advances in information technology allowed the development of Content-Based Image Retrieval (CBIR) systems, capable of retrieving images based on their similarity with one or more query images. Some of these systems are QBIC (Faloutsos et al., 1994), VisualSEEK (Smith and Chang, 1996), Virage (Hampapur et al., 1997), Netra (Ma and Manjunath, 1999), PicSOM (Laaksonen, Koskela, Laakso and Oja, 2000), SIMPLicity (Wang, Li and Wiederhold, 2001), CIRES (Iqbal and Aggarwal, 2002) and FIRE (Deselaers, Keysers and Ney, 2004). More than fifty CBIR systems are surveyed in (Veltcamp and Tanase, 2000).

Common ground for most of the systems cited above is that image retrieval is based on similarity measures estimated directly from low-level image features. This approach is likely to result in the retrieval of images with significant perceived differences from the query image, as the low-level features lack of semantic interpretation. This motivated researchers to focus on the utilization of higher-level semantic representations of image contents for content-based medical image retrieval. Recent approaches to content-based medical image retrieval include semantic mapping via hybrid Bayesian networks (Lin, Yin, Gao, Chen and Qin, 2006), Semantic Error-Correcting output Codes (SECC) based on individual classifiers combination (Yaoa, Antani, Longb, Thoma and Zhanga, 2006) and a framework that uses machine learning and statistical similarity matching techniques with relevance feedback (Rahman, Bhattacharya and Desai, 2007). However, these approaches involve supervised methodologies that require prior knowledge about the dataset and introduce constraints to the semantics required for the image retrieval task.

In contrast with other proposed clustering-based systems we give a brief description of their characteristics in comparison to our approach. The Blobworld system (Carson, Belongie, Greenspan and Malik, 2002) cluster pixels in a joint color-texture-position eight-dimensional space while the proposed system cluster windows extracted from the images and not pixels reducing thus the space needed to store the feature vectors (we need one feature vector for each window and not for each pixel). Furthermore, the query in Blobworld is done over regions while in our system it is done over the whole image. The major difference though is in the image comparison process. The blobworld cluster pixels to find regions, but in the query process it uses the original features of the pixels to compare two regions. Our system uses the patterns, the clustering output, to compare the images. The advantage is that we do not have to store the original feature vectors and the time complexity to compare the patterns is significantly smaller, especially compared to Blobworld system, where a lot of preprocessing and post processing is made.

FUZZYCLUB (Zhang, R., Zhang, Z.(M), 2002) is another cluster based image retrieval system. It uses a modified K-means algorithm to cluster pixels into regions, and it performs a secondary clustering to group all regions in the database. While the way that the comparison between two regions is similar in concept (compare every region from one image with everyone from the other and find an average of the best matchings), the
Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases

Comparison is again made over the features (color, texture, shape) extracted from the image. Our approach uses the extracted patterns to perform the comparison.

CLUE scheme (Cluster-based retrieval of images) (Chen, Wang, Krovetz, 2005) on the other hand uses clustering not to cluster the pixels of images in the database but to group the results of a query (the retrieved images) into categories in order to improve the user interaction with the image retrieval system.

Another image retrieval system, proposes the Color-based Clustering (CBC) (Stehling, Falcao and Nascimento, 2001). An agglomerative clustering algorithm is used to cluster the pixels and to find regions in the image that are smaller than a given threshold. This approach has the disadvantage that the regions cannot be larger than the threshold value (which is a percentage of the image area), and thus some large regions might end to be splitted. In the query process, the distance between two images is a weighted composition of the distances between the regions that compose each image.

The proposed approach combines the advantages of the clustering-based CBIR methodologies (Stehling, Falcao and Nascimento, 2001; Zhang, R., Zhang, Z. (M), 2002; Carson, Belongie, Greenspan and Malik, 2002; Chen, Wang, Krovetz, 2005) with a semantically rich representation of medical images.

The plethora and importance of the potential applications of Intuitionistic Fuzzy Sets have drawn the attention of many researchers that have proposed various kinds of IFS similarity measures. Example applications include identification of functional dependency relationships between concepts in data mining systems, approximate reasoning, pattern recognition and others. Similarity measures between IFSs have been proposed by Chen (1995, 1997) with $S_C$ measure, by Hong and Kim (1999) with $S_H$, by Fan and Zhangyan with $S_L$, and Li, Zhongxian and Degin (2002) who proposed the $S_O$ measure. Dengfeng and Chuntian (2002) proposed the $S_{DC}$ measure, Mitchel (2003) proposed a modification of the $S_{DC}$ measure, the $S_{HB}$ measure, Zhizhen and Pengfei (2003) proposed three measures $S^R$, $S^L$, and $S^O_h$, and three more measures have been proposed by Hung and Yang (2004), the $S^{2}_{HR}$, $S^{2}_{HY}$, and $S^{2}_{HY}$. Li, Olson and Qin (2007) provide a detailed comparison of these measures, pointing out the weaknesses of each one.

Some of the measures, like $S_C$, $S_H$, $S_L$, $S_{HB}$ and $S^{1}_{HR}$, $S^{2}_{HR}$, and $S^{2}_{HY}$ focus on the aggregation of the differences between membership values and those of non-membership values while others apply distances such as Minkowski, for $S_{DC}$, or Hausdorff, for $S^{2}_{HR}$, $S^{2}_{HR}$, and $S^{2}_{HY}$ in order to calculate the degree of similarity of the fuzzy sets. $S_{DC}$, $S^R$, and $S^L$ focus on the difference between median values of intervals and $S^O_h$ focus on the difference between length of interval.

Considering the effectiveness of these measures, some of them like the $S_C$ of Chen (1995, 1997) and $S^{1}_{HR}$, $S^{2}_{HR}$, and $S^{2}_{HY}$ disobey certain properties of IFS similarity measures, while all of the above mentioned measures fail in specific cases that Li, Olson and Qin (2007) mention with counter-intuitive examples. A comparison of these measures with the proposed one is provided in the section 5, evaluating our measure with the counter-intuitive examples given in (Li, Olson and Qin, 2007).
3 Intuitionistic fuzzy sets

The theoretical foundations of fuzzy and intuitionistic fuzzy sets are described in (Zadeh, 1965; Atanassov, 1986). This section briefly outlines the related notions used in this paper.

Definition 1 (Zadeh, 1965). Let a set E be fixed. A fuzzy set on E is an object \( \tilde{A} \) of the form
\[
\tilde{A} = \left\{ (x, \mu_A(x)) \middle| x \in E \right\}
\]
where \( \mu_A : E \rightarrow [0,1] \) defines the degree of membership of the element \( x \in E \) to the set \( \tilde{A} \subseteq E \). For every element \( x \in E \), \( 0 \leq \mu_A(x) \leq 1 \).

Definition 2 (Atanassov, 1986; Atanassov, 1994). An intuitionistic fuzzy set \( A \) is an object of the form
\[
A = \left\{ (x, \mu_A(x), \gamma_A(x)) \middle| x \in E \right\}
\]
where \( \mu_A : E \rightarrow [0,1] \) and \( \gamma_A : E \rightarrow [0,1] \) define the degree of membership and non-membership, respectively, of the element \( x \in E \) to the set \( A \subseteq E \). For every element \( x \in E \), it holds that \( 0 \leq \mu_A(x) \leq 1 \), \( 0 \leq \gamma_A(x) \leq 1 \) and
\[
0 \leq \mu_A(x) + \gamma_A(x) \leq 1
\]
For every \( x \in E \), if \( \gamma_A(x) = 1 - \mu_A(x) \), \( A \) represents a fuzzy set. The function
\[
\pi_A(x) = 1 - \mu_A(x) - \gamma_A(x)
\]
represents the degree of hesitancy of the element \( x \in E \) to the set \( A \subseteq E \).

For every two intuitionistic fuzzy sets \( A \) and \( B \) the following operations and relations are valid (Atanassov, 1986; Atanassov 1994)
\[
A \subseteq B \iff \forall x \in E, \mu_A(x) \leq \mu_B(x) \text{ and } \gamma_A(x) \geq \gamma_B(x)
\]
\[
A = B \iff A \subseteq B \text{ and } B \subseteq A
\]
\[
A^c = \left\{ (x, \gamma_A(x), \mu_A(x)) \middle| x \in E \right\}
\]
\[
A \cap B = \left\{ (x, \min(\mu_A(x), \mu_B(x)), \max(\gamma_A(x), \gamma_B(x))) \middle| x \in E \right\}
\]
\[
A \cup B = \left\{ (x, \max(\mu_A(x), \mu_B(x)), \min(\gamma_A(x), \gamma_B(x))) \middle| x \in E \right\}
\]
\[
A @ B = \left\{ (x, \frac{1}{2}(\mu_A(x) + \mu_B(x)), \frac{1}{2}(\gamma_A(x) + \gamma_B(x))) \middle| x \in E \right\}
\]
Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases

\[
\hat{A}_i \equiv \left\{ x, \frac{1}{n} \sum_{i=1}^{n} \mu_{A_i}(x), \frac{1}{n} \sum_{i=1}^{n} \gamma_{A_i}(x) \right\} \quad x \in E
\]

(11)

**Definition 3** (Dengfeng and Chuntian, 2002). Let \( S \) be a mapping \( \text{IFSs}(E) \times \text{IFSs}(E) \rightarrow [0,1] \), where \( \text{IFSs}(E) \) denotes the set of all intuitionistic fuzzy sets in \( E \). \( S(A, B) \) is said to be the degree of similarity between \( A \in \text{IFSs}(E) \) and \( B \in \text{IFSs}(E) \), if \( S(A, B) \) satisfies the following conditions:

a. \( S(A, B) \in [0,1] \)
b. \( S(A, B) = 1 \Leftrightarrow A = B \)
c. \( S(A, B) = S(B, A) \)
d. \( S(A, C) \leq S(A, B) \) and \( S(A, C) \leq S(B, C) \) if \( A \subseteq B \subseteq C \), \( C \in \text{IFSs}(E) \)

Representing the data of a real-world clustering problem by means of intuitionistic fuzzy sets, is a challenging issue providing the opportunity to investigate the effectiveness of the intuitionistic fuzzy theory in practice.

4 Intuitionistic fuzzy representation of data

The proposed intuitionistic fuzzy clustering requires that each data element \( x \) of a universe \( E \), belongs to an intuitionistic fuzzy set \( A \subset E \) by a degree \( \mu_{A_i}(x) \) and does not belong to \( A \) by a degree \( \gamma_{A_i}(x) \). The data elements can be of any kind. For the purposes of this study, which focuses to the clustering of image data we extend the definition of the intuitionistic fuzzy representation of a greyscale digital image (Vlachos and Sergiadis, 2005), for the representation of a color digital image.

**Definition 4**. A color digital image \( P \) of \( a \times b \) pixels size, composed of \( \xi \) channels \( P_k \), \( k=1,2,...,\xi \), digitized in \( q \) quantization levels per channel, is represented as the intuitionistic fuzzy set

\[
\Phi = \left\{ \left( c_{\theta}^i, \mu_{\theta}(c_{\theta}^i), \gamma_{\theta}(c_{\theta}^i) \right) \right\}_{\theta \in \theta^k}, \ i=1,2,...,a, \ j=1,2,...,b, \ k=1,2,...,\xi
\]

(12)

where \( \theta^k \) is the value of \( P_k \) at the position \( (i, j) \), and \( \mu_{\theta}(\theta^k) \) and \( \gamma_{\theta}(\theta^k) \) define the membership and the non-membership of \( \theta^k \) to \( P_k \), respectively.

As a membership function \( \mu_{\theta}(\theta) \), we consider the probability of occurrence of \( \theta \in [0,q-1] \) in an image channel

\[
\mu_{\theta}(\theta) = \frac{h(\theta)}{m \cdot n}, \quad \forall \theta \in [0,q-1]
\]

(13)

where

\[
h(\theta) = \left\{ (i,j) \in P_k | \theta^k = \theta, \ i=1,...,a; \ j=1,...,b; \ k=1,2,...,\xi \right\}
\]

(14)
is the crisp histogram of the pixel values in the channel, and \[ \mu_b(\theta) \] represents the cardinality of the enclosed set. The probability distribution described by Eq. 13 comprises a first-order statistical representation of the image channel that is easy to compute, and it is invariant to the rotation and translation.

Considering that real-world digital images usually contain noise of various origins, and imprecision in the channel values, the degree of belongingness of a value \( \theta \) in an image channel as expressed by \( \mu(\theta) \) is subject to uncertainty. In order to model this situation, we introduce a penalty factor \( p(\theta) \) so that \( \theta \) belongs less to the image channel if \( h(\theta) \) diverges more from the fuzzy histogram \( \hat{h}(\theta) \). The fuzzy histogram, originally proposed by Jawahar and Ray (1996), is defined as

\[
\hat{h}(\theta) = \left\| (i,j) \in P \mu_b(\theta^a); i = 1, \ldots, a, j = 1, \ldots, b, k = 1, 2, \ldots \right\|
\]

with

\[
\mu_b(\theta) = \max\left\{ 0, 1 - \frac{|\theta - \theta|}{\psi} \right\}
\]

where parameter \( \psi \) controls the span of the fuzzy number \( \theta : R \rightarrow [0,1] \) representing a fuzzy intensity level \( \theta \).

Therefore, the non-membership of \( \theta \) to an image channel can be expressed by

\[
\gamma_b(\theta) = 1 - \mu_b(\theta) \cdot p(\theta)
\]

The penalty factor \( p(\theta) \) is chosen to be proportional to the distance between the crisp \( h(\theta) \) and the fuzzy histogram \( \hat{h}(\theta) \), so that Eq. (3) is satisfied

\[
p(\theta) = \lambda \frac{|\hat{h}(\theta) - h(\theta)|}{\max_\theta|\hat{h}(\theta) - h(\theta)|}
\]

where \( \lambda \in [0,1] \) is constant. The denominator facilitates normalization purposes.

The membership and the non-membership defined by equations (13) and (17) over the values of the image channels, will be considered to form feature vectors. In order to evaluate the similarity between these vectors a novel similarity measure is proposed.

5 Proposed Similarity Measure

In this section we propose a novel similarity measure between intuitionistic fuzzy sets, based on the membership and non-membership values of their elements. Given an intuitionistic fuzzy set \( A \) we define two fuzzy sets, namely \( M_A, \Gamma_A \in F(E) \) where \( F(E) \) is the set of all fuzzy subsets of an element \( x \in E \). The membership and non-membership of these sets is defined as \( M_A = \{ \mu_A(x) \}, \Gamma_A = \{ \gamma_A(x) \} \forall x \in E \). In this connection, \( A \) can be described by the pair \( (M_A, \Gamma_A) \).

**Definition 5.** Considering two intuitionistic fuzzy sets \( A=(M_A, \Gamma_A), B=(M_B, \Gamma_B) \), where \( M_A, M_B, \Gamma_A, \Gamma_B \in F(E) \), and considering \( E \) as a finite universe \( E = \{ x_1, x_2, \ldots, x_n \} \),
we define the similarity measure \( Z_1 \) between the intuitionistic fuzzy sets \( A \) and \( B \) by the following equation:

\[
Z_1(A, B) = \frac{z_1(M_A, M_B) + z_1(\Gamma_A, \Gamma_B)}{2}
\]

where

\[
z_1(A', B') = \begin{cases} 
\sum_{i=1}^{n} \min(A'(x_i), B'(x_i)), & A' \cup B' \neq \emptyset \\
\sum_{i=1}^{n} \max(A'(x_i), B'(x_i)), & A' \cup B' = \emptyset
\end{cases}
\]

with \( A', B' \in F(E) \).

In order to accept \( Z_1 \) as a similarity metric we need to prove that \( z_1 \) satisfies the properties defined in Definition 3. It is straightforward to prove that properties P1, P2 and P3 are satisfied by \( z_1 \). We supply the proof for the 4th property.

**Lemma.** For all \( A', B', C' \in F(E) \), where \( F(E) \) is the set of all fuzzy subsets of an element \( x \in E \) and considering \( E \) as a finite universe \( E = \{x_1, x_2, ..., x_n\} \), if \( A' \subseteq B' \subseteq C' \) then \( z_1(A', C') \leq z_1(A', B') \) and \( z_1(A', C') \leq z_1(B', C') \).

**Proof:** By \( A' \subseteq B' \subseteq C' \) it implies that \( A'(x_i) \leq B'(x_i) \leq C'(x_i) \forall x_i \in E \) and

\[
z_1(A', C') = \sum_{i=1}^{n} \min(A'(x_i), C'(x_i)) = \sum_{i=1}^{n} A'(x_i) - \sum_{i=1}^{n} \max(A'(x_i), C'(x_i)) = \sum_{i=1}^{n} A'(x_i) - \sum_{i=1}^{n} C'(x_i),
\]

\[
z_1(A', B') = \sum_{i=1}^{n} \min(A'(x_i), B'(x_i)) = \sum_{i=1}^{n} A'(x_i) - \sum_{i=1}^{n} \max(A'(x_i), B'(x_i)) = \sum_{i=1}^{n} A'(x_i) - \sum_{i=1}^{n} B'(x_i),
\]

\[
z_1(B', C') = \sum_{i=1}^{n} \min(B'(x_i), C'(x_i)) = \sum_{i=1}^{n} B'(x_i) - \sum_{i=1}^{n} \max(B'(x_i), C'(x_i)) = \sum_{i=1}^{n} B'(x_i) - \sum_{i=1}^{n} C'(x_i),
\]

hence, \( z_1(A', C') \leq z_1(A', B') \) and \( z_1(A', C') \leq z_1(B', C') \).

Since \( A, B, C \in IF_{FS}(E) \) and \( A \subseteq B \subseteq C \) we have

\[
\mu_i(x) \leq \mu_{i'}(x) \leq \mu_c(x) \quad \text{and} \quad \gamma_i(x) \geq \gamma_i(x) \geq \gamma_c(x) \quad \forall x_i \in E, i = 1, 2, ..., n.
\]

therefore, \( z_1(M_A, M_B) \) and \( z_1(\Gamma_A, \Gamma_B) \) satisfy all properties a-d and so \( Z_1 \) also satisfies these properties. Thus, \( Z_1 \) is a similarity metric. To demonstrate the proposed measure a simple numeric example is given below.
Example. Assuming three sets \( A, B, C \) \( \in \text{IFS}s(E) \) with 
\[ A = \{ x, 0.4, 0.2 \}, \quad B = \{ x, 0.5, 0.3 \}, \quad C = \{ x, 0.5, 0.2 \} \]
we want to find whether \( B \) or \( C \) is more similar to \( A \). Using the equations (19) and (20) we compute the similarity of \( B \) and \( C \) to set \( A \).

\[
Z_1(A, B) = \frac{0.4 \cdot 0.2}{0.5 \cdot 0.3} = 0.733 \quad \text{and} \quad Z_2(A, C) = \frac{0.4 \cdot 0.2}{0.5 \cdot 0.2} = 0.8
\]

So, we conclude that \( C \) is more similar to \( A \) than \( B \).

The proposed intuitionistic similarity measure uses the aggregation of the minimum and maximum membership values in combination with those of the non-membership values. Although it is very simple to calculate, it is sensitive to small value changes and it deals well with all the counter-intuitive cases in which other measures fail. Most of the similarity measures reviewed in Section Error! Reference source not found., fail to evaluate to a valid intuitionistic value for specific cases. Some of them evaluate to 0 or 1 suggesting that the compared sets are either totally irrelevant or identical, while it is obvious that this is not true, and some other measures result in a high similarity value for obviously different sets. More specifically, in Table 1 we present all the counter-intuitive cases that Li, Olson and Qin (2007) have defined and the other measures fail, along with the calculation of the proposed measure for those cases.

In case (I) of Table 1 measure values \( S_1(A, B) \) and \( S_{10}(A, B) \) imply that \( A \) and \( B \) are totally similar. In cases (II) and (IV) other measures result in a rather big similarity value—our measure is not that optimistic. Moreover, in case (IV) it is obvious that sets \( A \) is more similar to \( C \) than to \( B \) (\( A \) and \( C \) have the same non-membership value), something that other measures do not take into account. In (III), while \( B \) and \( C \) are totally different, measures \( S_1, S_{10}, S'_3 \) give a similarity value of 0.5. On the contrary in (V) measures \( S_{10}, S_{1}, S_3 \) give a similarity value of 0 even if the non-membership value of both \( A \) and \( B \) is the same, suggesting a level of similarity between the two sets. In (VI) and (VII) measures \( S_{10}, S_1, S_3 \) result in a rather high similarity value and in (VII) they do not recognize that \( A \) is more similar to \( C \) than to \( B \), due to the same non-membership value of \( A \) and \( C \).

The above indicate the intuitiveness of the proposed measure, which satisfies all the properties of a similarity metric and does not fail in cases that other measures fail. Furthermore, the proposed measure is easy to calculate and does not use exponents or other functions that significantly slow down the calculations.
Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases

Table 1 Proposed and other similarity measures with counter-intuitive cases

<table>
<thead>
<tr>
<th>No</th>
<th>Measure</th>
<th>Counter-intuitive cases</th>
<th>Measure Values</th>
<th>Proposed measure value</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$S_C, S_{DC}$</td>
<td>$A = (x,0.0)$, $B = (x,0.5,0.5)$</td>
<td>$S_C(A,B)=S_{DC}(A,B)=1$</td>
<td>$Z_1 = 0$</td>
</tr>
<tr>
<td>II</td>
<td>$S_{IH}, S_{IH}, S'_s$</td>
<td>$A = (x,0.3,0.3)$, $B = (x,0.4,0.4)$, $C = (x,0.3,0.4)$, $D = (x,0.4,0.3)$</td>
<td>$S_{IH}(A,B)=S_{IH}(A,B)=S'<em>s(A,B)=0.9$, $S</em>{IH}(C,D)=S_{IH}(C,D)=S'_s(C,D)=0.9$</td>
<td>$Z_1(A,B)=Z_1(C,D)=0.75$</td>
</tr>
<tr>
<td>III</td>
<td>$S_{IH}, S_{IH}, S'_s$</td>
<td>$A = (x,1.0)$, $B = (x,0.0)$, $C = (x,0.5,0.5)$</td>
<td>$S_{IH}(A,B)=S_{IH}(A,B)=S'<em>s(A,B)=0.5$, $S</em>{IH}(B,C)=S_{IH}(B,C)=S'_s(B,C)=0.5$</td>
<td>$Z_1(A,B)=0.5$, $Z_1(B,C)=0$</td>
</tr>
<tr>
<td>IV</td>
<td>$S_L$ and $S'_s$</td>
<td>$A = (x,0.4,0.2)$, $B = (x,0.5,0.3)$, $C = (x,0.5,0.2)$</td>
<td>$S_L(A,B)=S'_s(A,B)=0.95$, $S_L(A,C)=S'_s(C,D)=0.95$</td>
<td>$Z_1(A,B)=0.73$, $Z_1(A,C)=0.9$</td>
</tr>
<tr>
<td>V</td>
<td>$S^1_{sp}, S^2_{sp}, S^3_{sp}$</td>
<td>$A = (x,1.0)$, $B = (x,0.0)$</td>
<td>$S^1_{sp}(A,B)=S^2_{sp}(A,B)=S^3_{sp}(A,B)=0$</td>
<td>$Z_1(A,B)=0.5$</td>
</tr>
<tr>
<td>VI</td>
<td>$S^1_{sp}, S^2_{sp}, S^3_{sp}$</td>
<td>$A = (x,0.3,0.3)$, $B = (x,0.4,0.4)$, $C = (x,0.3,0.4)$, $D = (x,0.4,0.3)$</td>
<td>$S^1_{sp}(A,B)=S^1_{sp}(C,D)=0.9$, $S^2_{sp}(A,B)=S^2_{sp}(C,D)=0.85$, $S^3_{sp}(A,B)=S^3_{sp}(C,D)=0.82$</td>
<td>$Z_1(A,B)=Z_1(C,D)=0.75$</td>
</tr>
<tr>
<td>VII</td>
<td>$S^1_{sp}, S^2_{sp}, S^3_{sp}$</td>
<td>$A = (x,0.4,0.2)$, $B = (x,0.5,0.3)$, $C = (x,0.5,0.2)$</td>
<td>$S^1_{sp}(A,B)=S^1_{sp}(A,C)=0.9$, $S^2_{sp}(A,B)=S^2_{sp}(A,C)=0.85$, $S^3_{sp}(A,B)=S^3_{sp}(A,C)=0.82$</td>
<td>$Z_1(A,B)=0.73$, $Z_1(A,C)=0.9$</td>
</tr>
</tbody>
</table>

6 Clustering intuitionistic fuzzy data

Most clustering methods assume that each data vector belongs only to one cluster. This is rational if the feature vectors reside in compact and well-separated clusters. However, in real-world applications clusters overlap, meaning that a data vector may belong partially to more than one clusters. In such a case and in terms of fuzzy set theory (Zadeh, 1965), the degree of membership of a vector $x_i$ to the $i$-th cluster $u_k$ is a value in the interval $[0,1]$. Ruspini (1969) introduced this idea which was later applied by Dunn (1973) to propose a clustering methodology based on the minimization of an objective function. In
Nikos Pelekis, Dimitris K. Iakovidis, Evangelos E. Kotsifakos, Ioannis Kopanakis


FCM is an iterative algorithm and its aim is to find cluster centroids that minimize a criterion function, which measures the quality of a fuzzy partition. A fuzzy partition is denoted by a \((c \times N)\)-dimensional matrix \(U\) of reals \(u_{ik} \in [0,1]\), \(\forall 1 \leq i \leq c\) and \(1 \leq k \leq N\), where \(c\) and \(N\) is the number of clusters and the cardinality of the feature vectors, correspondingly. The following constraint is imposed upon \(u_{ik}\):

\[
\sum_{i=1}^{c} u_{ik} = 1.0 < \sum_{i=1}^{c} u_{ik} \leq N \tag{21}
\]

Given this, the FCM objective function has the form:

\[
J_m(U,V) = \sum_{i=1}^{N} \sum_{k=1}^{c} (u_{ik})^m d_{ik}^2 \tag{22}
\]

where \(V\) is a \((c \times p)\)-dimensional matrix storing the \(c\) centroids, \(p\) is the dimensionality of the data, \(d_{ik}\) is an \(A\)-norm measuring the distance between data vector \(x_k\) and cluster centroid \(v_i\), and \(m \in [1, \infty)\) is a weighting exponent. The parameter \(m\) controls the fuzziness of the clusters. When \(m\) approximates 1, FCM performs a hard partitioning as the \(k\)-means algorithm does, while as \(m\) converges to infinity the partitioning is as fuzzy as possible. There is no analytical methodology for the optimal choice of \(m\).

Bezdek, Ehrlich and Full (1984) proved that if \(m\) and \(c\) are fixed parameters and \(I_1, I_c\) are sets defined as:

\[
\forall 1 \leq k \leq N, \begin{cases} I_k = \{i | 1 \leq i \leq c, d_{ik} = 0\} , \\ I^c = \{1,2,\ldots,c\} \setminus I_k , \end{cases} \tag{23}
\]

then \(J_m(U,V)\) may be minimized only if:

\[
u_{ik} = \begin{cases} \frac{(d_{ik})^2}{\sum_{j=1}^{c} (d_{jk})^2} , & I_k = \emptyset , \\ 0 , & i \not\in I_k , \\ \sum_{i \in I_k} u_{ik} = 1 , & I_k \neq \emptyset , \end{cases} \tag{24}
\]

and

\[
v_i = \frac{\sum_{k=1}^{N} (u_{ik})^m x_k}{\sum_{k=1}^{N} (u_{ik})^m}. \tag{25}
\]

By iteratively updating the cluster centroids and the membership degrees for each feature vectors, FCM iteratively moves the cluster centroids to the "right" location within the data set. In detail, the algorithm that results in the optimal partition is the Picard algorithm which is described below:
**Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases**

Step 1: Determine \( c (1 < c < N), \ m \in [1, \infty) \) and initialize \( V^{(0)} \), \( j \leftarrow 1 \).

Step 2: Calculate the membership matrix \( U^{(j)} \), using equation (24),

Step 3: Update the centroids’ matrix \( V^{(j)} \), using equation (25) and \( U^{(j)} \).

Step 4: If \( \| U^{(j)} - U^{(j)} \| > \varepsilon \) then \( j \leftarrow j+1 \) and go to Step 2.

**Figure 1 Fuzzy C-Means algorithm**

The parameter \( \varepsilon \) makes the algorithm to converge when the improvement of the fuzzy partition over the previous iteration is below a threshold, while \( \| U \|_F \) denotes the Frobenious norm.

The FCM algorithm minimizes intra-cluster variance, but shares the same problems with k-means (MacQueen, 1967). It does not ensure that it converges to an optimal solution, while the identified minimum is local and the results depend on the initial choice of the centroids.

FCM tries to partition the dataset by just looking at the feature vectors and as such it ignores the fact that these vectors may be accompanied by qualitative information which may be given per feature. For example, following the idea of intuitionistic fuzzy set theory, a data point \( x_k \) is not just a \( p \)-dimensional vector \( (x_{k1}, \ldots, x_{kp}) \) of quantitative information, but instead it is a \( p \)-dimensional vector of triplets \( [(\mu_k, \gamma_k), \ldots, (\mu_k, \gamma_k)] \), where for each \( x_k \) measurement there exists qualitative information which is provided via the intuitionistic membership \( \mu_k \) and non-membership \( \gamma_k \) of the current data point to the feature \( l \). It is evident that the FCM algorithm does not utilize intrinsically such qualitative information. In the application scenario of clustering images, a feature \( l \) may correspond to color information. Obviously, it would be of advantage if the clustering methodology could take into account the degree of membership and the degree of non-membership, regarding (for instance) how much red the image is, and how sure we are about our belief.

The main reason that FCM is unable to effectively utilize such intuitionistic vectors is that its distance function operates only on the feature vectors and not on the qualitative information which may be given per feature. In this paper, we propose a different perspective by substituting the distance function with the intuitionistic fuzzy set distance metric introduced in Section 5. Using the proposed distance function the fuzzy c-means objective function takes the form:

\[
J_m^{IFS}(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (a_{ik})^m |x_k - v_i|_{IFS}^{m}
\]  

The minimization of (26) can be achieved term by term:

\[
J_m^{IFS}(U, V) = \sum_{k=1}^{N} \phi_k(U)
\]

where

\[
\phi_k(U) = \left( \sum_{i=1}^{c} (a_{ik})^m |x_k - v_i|_{IFS}^{m} \right)
\]
The Lagrangian of (28) with constraints from (21) is:

\[
\forall \, i, k \in \mathcal{N}, \quad \Phi_k(U, \lambda) = \sum_{r=1}^{m} \left( u_{r,k} \right)^{\frac{1}{m}} \left| v_i - v_j \right|^{\frac{m}{2}} - \lambda \left( \sum_{r=1}^{m} u_{r,k} - 1 \right)
\]

(29)

where \( \lambda \) is the Lagrange multiplier. Setting the partial derivatives of \( \Phi_k(U, \lambda) \) to zero we obtain:

\[
\forall \, i, k \in \mathcal{N}, \quad \frac{\partial \Phi_k(U, \lambda)}{\partial u_{i,k}} = \sum_{r=1}^{m} u_{r,k}^{\frac{1}{m}} - 1 = 0
\]

(30)

and

\[
\forall \, i, k \in \mathcal{N}, \quad \frac{\partial \Phi_k(U, \lambda)}{\partial \lambda} = m(u_{i,k})^{m-1} \left| v_i - v_j \right|^{\frac{m}{2}} - \lambda = 0
\]

(31)

Solving (31) for \( u_{i,k} \) we get:

\[
u_{i,k} = \left( \frac{\lambda}{m} \right)^{\frac{1}{m}} \left| v_i - v_j \right|^{\frac{1}{m}}
\]

(32)

From (30) and (32) we obtain:

\[
\left( \frac{\lambda}{m} \right)^{\frac{1}{m-1}} = \frac{1}{\sum_{r=1}^{m} \left| v_i - v_j \right|^{\frac{m}{2}}}
\]

(33)

The combination of (32) and (33) yields:

\[
\forall \, i, k \in \mathcal{N}, \quad u_{i,k} = \left( \frac{\lambda}{m} \right)^{\frac{1}{m}} \left| v_i - v_j \right|^{\frac{1}{m}}
\]

(34)

Similarly with \( J_m(U, V) \), \( J_m^{\text{IFS}}(U, V) \) may be minimized only if:

\[
\forall \, i, k \in \mathcal{N}, \quad u_{i,k} = \begin{cases} 
\left| v_i - v_j \right|^{\frac{1}{m}} & \text{if } i \in I_k \\
\left( \frac{\lambda}{m} \right)^{\frac{1}{m}} \left| v_i - v_j \right|^{\frac{1}{m}} & \text{if } i \notin I_k \end{cases}
\]

(35)

while the centroids are computed by (25). It should be clarified that \( u_{i,k} \) corresponds to the membership of the \( k \)-th intuitionistic fuzzy vector to the \( i \)-th cluster and has nothing to do with the internal intuitionistic fuzzy memberships of the vector. Furthermore, as our distance function between two vectors is computed solely upon the intuitionistic fuzzy memberships and non-memberships of the vectors, after the computation of the centroids by equation (25) and before the next
iteration, where the \( u_k \) memberships to the new clusters are updated, there is a need for an additional step which estimates the intuitionistic fuzzy memberships and non-memberships of the new (virtual) centroids. In other words, it is necessary to deduce the membership \( \mu_i \) and non-membership \( \gamma_i \) values of each feature \( l \) that corresponds to the \( l \)-th dimension of the \( i \)-th centroid. At each iteration and for every centroid we extract the membership degree \( \mu_i \) of centroid \( v_i \) as the average of the membership degrees of all the intuitionistic fuzzy vectors that belong to cluster \( i \). Similarly, we extract the non-membership degrees \( \gamma_i \). More formally, if \( P_i \) is a set defined as:

\[
\forall_{b \in \text{vec}} P_i = \left\{ k \mid 1 \leq k \leq N; d_{ak} < d_{ai}, \forall 1 \leq r \leq N \land r \neq i \right\}
\]

then the intuitionistic fuzzy set \( IFS_{v_i} \) for centroid \( v_i \) is defined as:

\[
\forall_{b \in \text{vec}} IFS_{v_i} = \@_{v \in P_i} IFS_v
\]

From (Atanassov, 1994) we obtain:

\[
\forall_{b \in \text{vec}} \mu_i = \frac{\sum_{v \in P_i} \mu_v}{|P_i|}, \quad \gamma_i = \frac{\sum_{v \in P_i} \gamma_v}{|P_i|}
\]

Given the above discussion, the modified FCM algorithm that clusters intuitionistic fuzzy data is subsequently described:

**Step 1** Determine \( c \) \((1 < c < N)\), \( m \in [1, \infty) \) and initialize \( U^{(0)} \) by selecting \( c \) random intuitionistic fuzzy vectors, \( j \leftarrow 1 \).

**Step 2:** Calculate the membership matrix \( U^{(j)} \), using (35).

**Step 3:** Update the centroids’ matrix \( V^{(j)} \), using (25) and \( U^{(j)} \), and compute membership and non-membership degrees of \( V^{(j)} \) using (38).

**Step 4:** If \( \left| U^{(j)} - U^{(j-1)} \right| > \varepsilon \) then \( j \leftarrow j + 1 \) and go to Step 2.

**Figure 2** Modified Fuzzy C-Means algorithm for clustering intuitionistic fuzzy data

In comparison to the literal FCM algorithm the clustering scheme presented in Figure 2 introduces (a) a different initialization tactic of the \( V \) matrix as in our case centroid vectors are intuitionistic fuzzy vectors (step 1), (b) a new way of the calculation of the membership degrees of a vector to a cluster, taking into account both membership and non-membership values of the intuitionistic fuzzy vectors (step 2) and (c) a method to update the \( V \) matrix at each iteration based solely on the theory of the intuitionistic fuzzy sets (step 3).

### 7 Results

In Section 2, we have pinpointed that a number of content-based information retrieval approaches proposed in the recent literature has been based on clustering. Therefore, the performance of the clustering algorithm used is essential to the information retrieval task. In the following we present the results of the experiments conducted to evaluate the
performance of the proposed clustering algorithm, in comparison with the well established FCM.

The application scenario for the experimental evaluation involves clustering of a 400 image collection spanning four equally distributed classes of different color themes including amphorae, ancient monuments, coins, and statues (Figure 3). The images have been provided by the Foundation of Hellenic World, which maintains a publicly available repository of texts, images and multimedia data collections of Greek historical items and art (FHW). The images are of different sizes and have been inconsistently acquired from different sources. They have been digitized in 256 quantization levels per RGB channel and have been downscaled to fit into a 256×256 bounding box.

![Example images from the four classes used in the experiments](image)

Figure 3 Example images from the four classes used in the experiments, (a) amphorae, (b) ancient monuments, (c) coins, and (d) statues.

Each image was represented by an intuitionistic fuzzy set according to (12), using only chromatic information so as to be approximately independent from intensity variations. In order to decorrelate the intensity from the chromatic image components, the images have been transformed to the $I_1I_2I_3$ color space according to the following equation (Ohta et al., 1980)

$$
\begin{bmatrix}
I_1 \\
I_2 \\
I_3
\end{bmatrix} = \begin{bmatrix} 0.333 & 0.333 & 0.333 \\ 0.500 & 0.000 & -0.500 \\ -0.500 & 1.000 & -0.500 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}
$$

(39)

In this color space, the $I_1$ component explains the highest proportion of the total variance and represents intensity, whereas $I_2$ and $I_3$ correspond to the second and the third highest proportion respectively and carry chromatic information. A very useful property of this space is that image regions of different colors can be easily discriminated by simple thresholding operations. In other words, the histograms produced by the values of its color components exhibit peaks corresponding to regions of different colors in the image.

Among the chromatic components of $I_1I_2I_3$, we selected $I_2$ as the most discriminating for the color regions comprising the available images. This is in agreement with (Ohta et al., 1980) which suggests that the discrimination power of $I_2$ could only marginally increase with the contribution of $I_3$. Moreover, we observed that the image channel corresponding to the $I_3$ component exhibit a low dynamic range of values, having a single-peak histogram that varies slightly between images belonging to different classes.

Examples of membership and non-membership functions used for the intuitionistic fuzzy representation of color images are illustrated in Figure 4. The values of the parameters used in equations (16)-(18) for the estimation of the membership and of the
Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases

non-membership functions are $\lambda = 1$ and $a = 5$. In Figs. 5a, 5c and 5d, the higher of the two peaks correspond to the white background regions of the images, whereas the lower peaks correspond to the depicted objects. Similarly, in Fig. 5b the higher peak corresponds to the marble of the ancient monument and the lower peaks correspond to the sky region. As regards the non-membership functions, an intuitive interpretation could be given by considering their correlation with the corresponding membership functions. The correlation is usually less around the peaks that correspond to less homogenous image regions. For example in Fig. 5b, the absolute correlation between the membership and the non-membership function estimated for the region of the ancient monument is 70%, whereas for the region of sky is 82%. Similarly, the absolute correlation between the membership and the non-membership functions in Figs. 5a, 5b and 5c, for the homogenous white background regions reaches 96.5%.

Figure 4 Membership and non-membership functions corresponding to the images of Fig. 3. The horizontal axes represent the values of $I_2$ normalized within the range $[0, 255]$, whereas the vertical axes have been rescaled in order to improve the visibility of the graphs. The graphs focus on the regions of the membership and non-membership functions for which the variance is higher. The lines that intersect the frame of the graphs extending beyond the visible area join to peak membership and non-membership values.
Clustering experiments were conducted with all possible class combinations, using a) the proposed clustering algorithm with the intuitionistic fuzzy data, b) FCM with crisp \( l_2 \)-histogram data, and c) FCM with fuzzy \( l_2 \)-histogram data. In all the experiments, the same parameters (\( \epsilon = 0.0001 \) \( m = 2.0 \)) and initialization conditions were used. The clustering performance was evaluated in terms of classification accuracy, algorithm iterations and absolute execution time. The experiments were executed on a PC with Intel Pentium M at 1.86 GHz, 512 MB RAM and 60 GB hard disk. The results are summarized in Figure 5.
Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases

![Graph showing comparative results of clustering algorithms](image)

Figure 5 Comparative results of using the proposed clustering algorithm with the intuitionistic fuzzy data, and of using the FCM with the crisp and with the fuzzy data as input: (a) classification accuracy, (b) number of iterations required for the clustering algorithms to converge, and (c) execution time required in seconds.

Figure 5a shows that in all the experiments the accuracy achieved by the proposed algorithm was higher than the accuracy obtained by FCM for four or three classes. The maximum accuracies achieved with the proposed algorithm are 74.4% and 93.3% for four and three classes respectively. These percentages reduce to 64.4% and 79.2%, in the case of FCM clustering with fuzzy data. The results of the clustering experiments performed with data from two classes show that the accuracy of the proposed algorithm can be considered comparable with, and only in some cases higher than, the accuracy obtained by FCM. The maximum accuracy obtained by both algorithms reached 100%.

Comparing the two algorithms in terms of efficiency, Figs 5b and 5c, show that the proposed algorithm has a considerable advantage over FCM, as it requires less algorithm iterations and in most cases less time to reach convergence. The average improvement in absolute execution time is 63±27%.

8 Conclusions – Future work

We presented a novel approach to clustering of data sets produced in the context of intuitionistic fuzzy set theory. More specifically,

- we proposed a novel variant of the FCM clustering algorithm that copes with uncertainty in the localization of feature vectors due to imprecise measurements and noise;
- we introduced an intuitionistic fuzzy representation of color digital images as a paradigm of intuitionistic fuzzification of data;
- we defined a novel similarity measure between intuitionistic fuzzy sets and proved its superiority over other metrics. This measure was incorporated in the proposed FCM variant.

We have conducted a comprehensive set of experiments and based on the results, it can be concluded that the proposed clustering algorithm can be more efficient and more
Nikos Pelekis, Dimitris K. Iakovidis, Evangelos E. Kotsifakos, Ioannis Kopanakis

effective than the well established FCM algorithm, especially as the number of clusters increases.

Future perspectives of this work include:
• systematic evaluation of the proposed scheme for the clustering of various kinds of datasets after appropriately representing them in terms of intuitionistic fuzzy sets theory;
• enhancement of the proposed clustering scheme so as to take into account not only the membership, but also the non-membership of each data vector to a cluster.

Acknowledgment

This research work is supported by the project MetA-On, founded by the Operational Programme “Information Society” of the Greek Ministry of Development, General Secretariat for Research and Technology, co-funded by the European Union. The collection of images used in this study is courtesy of the Foundation of the Hellenic World (FHW), http://www.fhw.gr.

References

Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases


Nikos Pelekis, Dimitris K. Iakovidis, Evangelos E. Kotsifakos, Ioannis Kopanakis


