Algorithm for Clustering Data Set Represented by Intuitionistic Fuzzy Estimates

Ludmila Todorova, Peter Vassilev*

Centre of Biomedical Engineering, Bulgarian Academy of Sciences
105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
E-mail: lpt@cbme.bas.bg, peter.vassilev@gmail.com

**Corresponding author*

Received: January 11, 2010 **Accepted: March 12, 2010**

Published: April 15, 2010

Abstract: *One of the main problems in the area of pattern recognition in biomedical research areas is to determine clusters of patterns with similar features. It is especially relevant in the case of intuitionistic fuzzy sets. In the present paper an iterative procedure for clustering of patterns represented by their intuitionistic fuzzy sets – degrees of membership, degrees of non-membership and indeterminacy. The procedure is open to selection and application of an appropriate to the data distribution similarity measure for intuitionistic fuzzy sets.*

Keywords: *Intuitionistic fuzzy sets, Pattern recognition, Distance, Similarity measures between IFS.*

Introduction

Pattern recognition plays an important part in human cognition. Humans are able to identify patterns that appear in many types of data, recognize instances of these patterns, and draw relevant conclusions [1].

There are fundamental problems in pattern recognition are (see e.g. [2]):

- *the identification of natural groups (clustering) composed by samples with similar patterns*;
- *the classification of each sample in one of k possible classes (labels)*.

In the present paper a clustering algorithm for patterns from intuitionistic fuzzy sets represented by their degrees of membership, non-membership and indeterminacy is proposed. These sets are especially suitable for the representation of data in the field of medicine and biomedicine, since they allow for greater flexibility in modeling the uncertain cases.

The intuitionistic fuzzy sets (*IFS*) introduced by Atanassov [3] are an extension of the theory of fuzzy sets created by Zadeh [4] as an adequate mathematical description of imprecision and uncertainty in nature.

Here we will briefly remind the basic notions of the theory of *IFS*. The set A^* is *IFS* if there exist:

$A^* = \{\langle x, \mu_A(x), \nu_A(x)\rangle / x \in E\}$

where the mappings $\mu_A: E \to [0, 1]$ and $\nu_A: E \to [0, 1]$ define the degree of membership and non-membership of the element $x \in E$ to the set A, which is a subset of E and for every $x \in E$:

$$0 \le \mu_A(x) + \nu_A(x) \le 1 \tag{1}$$

For the purposes of the present work it is assumed that $\mu_A(x)$ and $\nu_A(x)$ are obtained through expert evaluations.

The function π_A, which is defined by the formula:

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \tag{2}$$

corresponds to the degree of indeterminacy (uncertainty) of the element $x \in E$ regarding the set A.

The clustering of the patterns is done using the concept of distance. In the *IFS* the commonly defined metrics are the following:

- Hamming metrics. It is defined as:

$$h_A(x,y) = \frac{1}{2}\left(\left|\mu_A(x) - \mu_A(y)\right| + \left|\nu_A(x) - \nu_A(y)\right|\right) \tag{3}$$

- In [5] an analogue of the latter using all the degrees is introduced:

$$h_A(x,y) = \frac{1}{2}\left(\left|\mu_A(x) - \mu_A(y)\right| + \left|\nu_A(x) - \nu_A(y)\right| + \left|\pi_A(x) - \pi_A(x)\right|\right) \tag{4}$$

- Euclidean. It is defined as:

$$e_A(x,y) = \sqrt{\frac{1}{2}\left(\left(\mu_A(x) - \mu_A(y)\right)^2 + \left(\nu_A(x) - \nu_A(y)\right)^2\right)} \tag{5}$$

Also as in the previous case (see [5]) a modified version is:

$$e'_A(x,y) = \sqrt{\frac{1}{2}\left(\left(\mu_A(x) - \mu_A(y)\right)^2 + \left(\nu_A(x) - \nu_A(y)\right)^2 + \left(\pi_A(x) - \pi_A(y)\right)^2\right)} \tag{6}$$

The distances defined over two *IFS A* and *B* are:

- Hamming distance. For any two *IFS A* and *B* defined over a common universe set *E* the Hamming distance between *A* and *B* is given by:

$$H(A,B) = \frac{1}{2}\sum_{x \in E}\left|\mu_A(x) - \mu_B(x)\right| + \left|\nu_A(x) - \nu_B(x)\right| \tag{7}$$

- Euclidean distance. For any two *IFS A* and *B* defined over a common universe set *E* the Euclidean distance between *A* and *B* is given by:

$$E(A,B) = \sqrt{\frac{1}{2}\sum_{x \in E}\left(\mu_A(x) - \mu_B(x)\right)^2 + \left(\nu_A(x) - \nu_B(x)\right)^2} \tag{8}$$

During the years various similarity measures between *IFS* have been defined. Of those that have been used in pattern recognition the most notable are the following:

- Chen [6, 7] proposed the concept of similarity measures between *IFS* sets and defined it as follows:

$$S_C(A,B)=1-\frac{\sum_{i=1}^{n}\left|S_A(x_i)-S_B(x_i)\right|}{2n} \tag{9}$$

where

$$S_A(x_i)=\mu_A(x_i)-\nu_A(x_i) \tag{10}$$

$$S_B(x_i)=\mu_B(x_i)-\nu_B(x_i) \tag{11}$$

- Hong and Kim [8] and Fan and Zhangyan [9] proposed new similarity measures S_H and S_L as given below.

$$S_H(A,B)=1-\frac{\sum_{i=1}^{n}\left|\mu_A(x_i)-\mu_B(x_i)\right|+\left|\nu_A(x_i)-\nu_B(x_i)\right|}{2n} \tag{12}$$

$$S_L(A,B)=1-\frac{\sum_{i=1}^{n}\left|S_A(x_i)-S_B(x_i)\right|}{4n}-\frac{\sum_{i=1}^{n}\left|\mu_A(x_i)-\mu_B(x_i)\right|+\left|\nu_A(x_i)-\nu_B(x_i)\right|}{4n} \tag{13}$$

- Yanhong *et al.* (2002) proposed a different similarity measure S_0 as follows:

$$S_0(A,B)=1-\sqrt{\frac{1}{2n}\left(\sum_{i=1}^{n}(\mu_A(x_i)-\mu_B(x_i))^2+(\nu_A(x_i)-\nu_B(x_i))^2\right)} \tag{14}$$

- Dengfeng and Chuntian [10] proposed their similarity measure of *IFS*s, which we will denote as S_{DC}. They applied this measure to pattern recognition. This measure was originally presented as a form of weighted similarity measure. But it can also be written as

$$S_{DC}(A,B)=1-\sqrt[p]{\frac{\sum_{i=1}^{n}\left|\psi_A(x_i)-\psi_B(x_i)\right|^p}{n}} \tag{15}$$

where p is a parameter,

$$\psi_A(x_i)=\frac{\mu_A(x_i)+1-\nu_A(x_i)}{2} \tag{16}$$

$$\psi_B(x_i)=\frac{\mu_B(x_i)+1-\nu_B(x_i)}{2} \tag{17}$$

- Mitchell [11] gave a simple modification of S_{DC} and corrected a problem occurring in S_{DC}'s. He adopted a statistical viewpoint interpreting A and B as ensembles of ordered membership functions filling the space between $\mu_A(x_i)$ and $1-\nu_A(x_i)$ as well as between $\mu_B(x_i)$ and $1-\nu_B(x_i)$. Let $p_\mu(A,B)$ and $p_\nu(A,B)$ denote the similarity measures between the membership function $\mu_A(x_i)$ and $\mu_B(x_i)$ as well as between the "high" membership function $1-\nu_A(x_i)$ and $1-\nu_B(x_i)$, respectively:

$$p_{\mu}(A,B)=S_{DC}\left(\mu_A(x_i),\ \mu_B(x_i)\right) \tag{18}$$

$$p_{\nu}(A,B)=S_{DC}\left(1-\nu_A(x_i),1-\nu_B(x_i)\right) \tag{19}$$

Then, the modified S_{DC}, called S_{HB}, would be:

$$S_{HB}(A,B)=\frac{1}{2}\left(p_{\mu}(A,B)+p_{\nu}(A,B)\right) \tag{20}$$

- To overcome the weakness of S_{DC}, Zhizhen and Pengfei [12] proposed $S_e^p(A,B)$, $S_s^p(A,B)$ and $S_h^p(A,B)$ as follows:

$$S_e^p(A,B)=1-\sqrt[p]{\frac{\sum_{i=1}^{n}\left(\phi_{\mu}(x_i)+\phi_{\nu}(x_i)\right)^p}{n}} \tag{21}$$

Here

$$\phi_{\mu}(x_i)=\left|\mu_A(x_i)-\mu_B(x_i)\right|/2,\ \phi_{\nu}(x_i)=\left|(1-\nu_A(x_i))/2-(1-\nu_B(x_i))/2\right| \tag{22}$$

$$S_s^p(A,B)=1-\sqrt[p]{\frac{\sum_{i=1}^{n}\left(\varphi_{s1}(x_i)+\varphi_{s2}(x_i)\right)^p}{n}} \tag{23}$$

$$\varphi_{s1}(x_i)=\left|m_{A1}(x_i)-m_{B1}(x_i)\right|/2 \tag{24}$$

$$\varphi_{s2}(x_i)=\left|m_{A2}(x_i)-m_{B2}(x_i)\right|/2 \tag{25}$$

$$m_{A1}(x_i)=\left(\mu_A(x_i)-m_A(x_i)\right)/2 \tag{26}$$

$$m_{B1}(x_i)=\left(\mu_B(x_i)-m_B(x_i)\right)/2 \tag{27}$$

$$m_{A2}(x_i)=\left(m_A(x_i)+1-\nu_A(x_i)\right)/2 \tag{28}$$

$$m_{B2}(x_i)=\left(m_B(x_i)+1-\nu_B(x_i)\right)/2 \tag{29}$$

$$m_A(x_i)=\left(\mu_A(x_i)+1-\nu_A(x_i)\right)/2 \tag{30}$$

$$m_B(x_i)=\left(\mu_B(x_i)+1-\nu_B(x_i)\right)/2 \tag{31}$$

$$S_h^p(A,B)=1-\sqrt[p]{\frac{\sum_{i=1}^{n}\left(\eta_1(i)+\eta_2(i)+\eta_3(i)\right)^p}{3n}} \tag{32}$$

$\eta_1(i)=\phi_{\mu}(x_i)+\phi_{\nu}(x_i)$ (occurring in $S_e^p(A,B)$) or $\eta_1(i)=\varphi_{s1}(x_i)+\varphi_{s2}(x_i)$)
(occurring in $S_s^p(A,B)$) (33)

$\eta_2(i)=\psi_A(x_i)-\psi_B(x_i)$ (occurring in S_{DC}) (34)

$$\eta_3(i)=\max\left(l_A(i),l_B(i)\right)-\min\left(l_A(i),l_B(i)\right) \tag{35}$$

$$l_A(i)=\left(1-\mu_A(x_i)-\nu_A(x_i)\right)/2,\ l_B(i)=\left(1-\mu_B(x_i)-\nu_B(x_i)\right)/2 \tag{36}$$

- Hung and Yang [13] presented three new similarity measures between *IFS*s based on Hausdorff distance:

$$S_{HY}^{1}(A,B)=1-d_H(A,B) \tag{37}$$

$$S_{HY}^{2}(A,B)=\left(e^{-d_H(A,B)}-e^{-1}\right)/\left(1-e^{-1}\right) \tag{38}$$

$$S_{HY}^{3}(A,B)=\left(1-d_H(A,B)\right)/\left(1+d_H(A,B)\right) \tag{39}$$

Here

$$d_H(A,B)=\frac{1}{n}\sum_{i=1}^{n}\max\left(\left|\mu_A(x_i)-\mu_B(x_i)\right|,\left|\nu_A(x_i)-\nu_B(x_i)\right|\right) \tag{40}$$

Algorithm

In the present paper an algorithm based on an iterative procedure. Before the start of the algorithm, on the basis of the data, a similarity measure is chosen to be implemented in the procedure.

Step 1: All the values of the degrees of membership $\mu(x_i)$ of the considered set are arranged in order of diminishing value.

The pattern(s) x_A^e, for which $\mu(x_A^e)=\max(\mu(x_i))$ is taken as etalon of the first class.

The pattern(s) x_B^e, for which $\mu(x_A^e)=\min(\mu(x_i))$ is taken as etalon of the second class.

Step 2: We find:

– $x_A^{e'}$ – the closest neighbor of x_A^e (in the sense of *IFS* distances) and:

$$x_A^{e'}\in\omega_A,\ \text{if}\ \left|\mu(x_A^{e'})-\mu(x_A^e)\right|<\left|\mu(x_A^{e'})-\mu(x_B^e)\right| \tag{41}$$

$$x_A^{e'}\in\omega_B,\ \text{if}\ \left|\mu(x_A^{e'})-\mu(x_A^e)\right|>\left|\mu(x_A^{e'})-\mu(x_B^e)\right| \tag{42}$$

– $x_B^{e'}$ – the closest neighbor of x_B^e and:

$$x_B^{e'}\in\omega_B,\ \text{if}\ \left|\mu(x_B^{e'})-\mu(x_B^e)\right|<\left|\mu(x_B^{e'})-\mu(x_A^e)\right| \tag{43}$$

$$x_A^{e'}\in\omega_B,\ \text{if}\ \left|\mu(x_B^{e'})-\mu(x_B^e)\right|>\left|\mu(x_B^{e'})-\mu(x_A^e)\right| \tag{44}$$

Step 3: Each etalon of cluster is localized and corrected according to the formula:

$$x_A^{e'}=\frac{1}{n_A}\sum_{i=1}^{n_A}\mu_i(x) \tag{45}$$

$$x_B^{e'}=\frac{1}{n_B}\sum_{i=1}^{n_B}\mu_i(x) \tag{46}$$

where: n_A and n_B is the number of patterns in the respective cluster in the current moment.

Step 4:
If $n_A+n_B<N$ – return to Step 2.
If $n_A+n_B=N$ – return to Step 5.

Step 5: Arrange all the values of the degrees of non-membership $\nu(x_i)$ of the patterns in order of diminishing value.

The pattern(s) x_A^e, for which $\nu(x_A^e)=\min(\nu(x_i))$ is taken as an etalon of the first class.

The pattern(s) x_B^e, for which $\nu(x_A^e)=\max(\nu(x_i))$ is taken as an etalon of the second class.

Step 6: We find:

– $x_A^{e'}$ – the closest neighbor of x_A^e and:

$$x_A^{e'} \in \omega_A\text{, if } \left|\nu\left(x_A^{e'}\right)-\nu\left(x_A^e\right)\right| < \left|\nu\left(x_A^{e'}\right)-\nu\left(x_B^e\right)\right| \tag{47}$$

$$x_A^{e'} \in \omega_B\text{, if } \left|\nu\left(x_A^{e'}\right)-\nu\left(x_A^e\right)\right| > \left|\nu\left(x_A^{e'}\right)-\nu\left(x_B^e\right)\right| \tag{48}$$

– $x_B^{e'}$ – the closest neighbor of x_B^e and:

$$x_B^{e'} \in \omega_B\text{, if } \left|\nu\left(x_B^{e'}\right)-\nu\left(x_B^e\right)\right| < \left|\nu\left(x_B^{e'}\right)-\nu\left(x_A^e\right)\right| \tag{49}$$

$$x_A^{e'} \in \omega_B\text{, if } \left|\nu\left(x_B^{e'}\right)-\nu\left(x_B^e\right)\right| > \left|\nu\left(x_B^{e'}\right)-\nu\left(x_A^e\right)\right| \tag{50}$$

Step 7: Each etalon of cluster is localized and corrected according to the formula:

$$x_A^{e'} = \frac{1}{n_A}\sum_{i=1}^{n_A}\nu_i(x) \tag{51}$$

$$x_B^{e'} = \frac{1}{n_B}\sum_{i=1}^{n_B}\nu_i(x) \tag{52}$$

where: n_A and n_B is the number of the patterns in the respective cluster to the current moment.

Step 8:
If $n_A + n_B < N$ – return to Step 6.
If $n_A + n_B = N$ – return to Step 9.

Step 9: The etalons of the clusters are with coordinates $\left(\mu\left(x_A^{e'}\right), \nu\left(x_A^{e'}\right)\right)$ and $\left(\mu\left(x_B^{e'}\right), \nu\left(x_B^{e'}\right)\right)$, determined in Steps 3 and 7.

If the pattern x_i remains in one and the same cluster according to Steps 1 - 4 of the procedure and according to Steps 5 - 8 of the procedure it remains in this cluster. Otherwise we go to Step 10.

Step 10: According to the preliminary chosen similarity measure the pattern x_i is assigned to a cluster with which it has better similarity.

Conclusion

The proposed algorithm reflects in equal measure the expert evaluations of the degrees of membership and non-membership. In this way a given pattern would be assigned to a respective cluster only when it has high values for the degree of membership and low for the degree of non-membership.

Patterns for which both conditions are not simultaneously fulfilled are assigned to a cluster with which it has the greatest similarity with regards to the chosen similarity measure for the *IFS*s. The application of the proposed iterative procedure with different similarity measures permits an evaluation of the results provided by these measures for certain biomedical problems. In this manner the usefulness of the different similarity measures for *IFS* may be compared and analyzed in various cases.

Acknowledgements

This work is partially supported by the European Social Fund and Bulgarian Ministry of Education, Youth and Science under Operative Program "Human Resources Development" grant BG051PO001-3.3.04/40 and by the National Science Fund of Bulgaria grant № DTK 02/48 – "System for Computer Aid of the Decisions for Weaning Critically Ill Ventilator Dependent from Mechanical Ventilation".

References

1. Brody S. (2005). Cluster-based Pattern Recognition in Natural Language Text, Master's thesis, University Jerusalem, Israel.
2. Papa J. (2008). Supervised Pattern Classification using Optimum-path Forest, Ph.D. Thesis, Institute of Computing, University of Campinas.
3. Atanassov K. (1999). Intuitionistic Fuzzy Sets, Heidelberg, Physica-Verlag.
4. Zadeh L. (1965). Fuzzy Sets, Information and Control, 8, 338-353.
5. Szmidt E., J. Kacprzyk (2000). Distances between Intuitionistic Fuzzy Sets, Fuzzy Sets and Systems, 114(3), 505-518.
6. Chen S. (1995). Measures of Similarity between Vague Sets, Fuzzy Sets Systems, 74(2), 217-223.
7. Chen S. (1997). Similarity Measures between Vague Sets and between Elements, IEEE Trans. Syst. Man Cybernet., 27(1), 153-158.
8. Hong D., C. Kim (1999). A Note on Similarity Measures between Vague Sets and between Elements, Inform. Science, 115, 83-96.
9. Fan L., X. Zhangyan (2001). Similarity Measures between Vague Sets, J. Software, 12(6), 922-927.
10. Yanhong L., D. Olson, Z. Qin (2007). Similarity Measures between Intuitionistic Fuzzy (Vague) Sets: A Comparative Analysis, Pattern Recognition Letters, 28(2), 278-285.
11. Mitchell H. (2003). On the Dengfeng-Chuntian Similarity Measure and its Application to Pattern Recognition, Pattern Recognition Letters, 24, 3101-3104.
12. Zhizhen L., S. Pengfei (2003). Similarity Measures on Intuitionistic Fuzzy Sets, Pattern Recognition Letters, 24, 2687-2693.
13. Hung W. L., M. S. Yang (2004). Similarity Measures of Intuitionistic Fuzzy Sets based on Hausdorff Distance, Pattern Recognition Letters, 25, 1603-1611.

Res. Assoc. Lyudmila Todorova, Ph.D.
E-mail: lpt@clbme.bas.bg

Lyudmila Todorova was born in 1961. She received a M. Sc. Degree in Department of Automation (1984) and Ph.D. Degree (2007) at Technical University – Sofia. At present she is a Research Associate at the Centre of Biomedical Engineering – Bulgarian Academy of Sciences. Her scientific interests are in the fields of pattern recognition, statistical methods, fuzzy methods, decision making, decision making in medicine. She has about 30 scientific publications with more than 40 known citations.

Peter Vassilev, Ph.D. Student
E-mails: ilywrin@clbme.bas.bg, peter.vassilev@gmail.com

Peter Vassilev is a Ph.D. student in the Centre of Biomedical Engineering, Sofia. He is currently working on his Ph.D. thesis under the supervision of Prof. Krassimir Atanassov dedicated to application of intuitionistic fuzzy sets and their modifications to the area of artificial intelligence. He holds M.Sc. degree in applied mathematics and his research interests are in the area of bioinformatics, artificial intelligence and algorithms.