Applied Soft Computing 46 (2016) 284–295

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Generalized picture distance measure and applications to picture fuzzy clustering

Le Hoang Son

VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan, Hanoi, Viet Nam

ARTICLE INFO

Article history:
Received 14 January 2016
Received in revised form 4 May 2016
Accepted 6 May 2016
Available online 10 May 2016

Keywords:
Clustering quality
Hierarchical fuzzy clustering
Intuitionistic fuzzy sets
Picture distance measure
Picture fuzzy sets

ABSTRACT

Picture fuzzy set (PFS), which is a generalization of traditional fuzzy set and intuitionistic fuzzy set, shows great promises of better adaptation to many practical problems in pattern recognition, artificial life, robotic, expert and knowledge-based systems than existing types of fuzzy sets. An emerging research trend in PFS is development of clustering algorithms which can exploit and investigate hidden knowledge from a mass of datasets. Distance measure is one of the most important tools in clustering that determine the degree of relationship between two objects. In this paper, we propose a generalized picture distance measure and integrate it to a novel hierarchical picture fuzzy clustering method called Hierarchical Picture Clustering (HPC). Experimental results show that the clustering quality of the proposed algorithm is better than those of the relevant ones.

1. Introduction

Since fuzzy set (FS) [49] was firstly introduced by Zadeh in 1965, many extensions of FS have been proposed in the literature such as the type-2 fuzzy set (T2FS) [18], rough set (RS) [24], soft set, rough soft set and fuzzy soft set [15], intuitionistic fuzzy set (IFS) [3], intuitionistic fuzzy rough set (IFRS) [51], soft rough fuzzy set & soft fuzzy rough set [19], interval-valued intuitionistic fuzzy set (IVIFS) [38] and hesitant fuzzy set (HFS) [32]. The aim of those extensions is to overcome the limitations of FS regarding the degree of fuzziness, the uncertainty of membership degrees, and the existence of neutrality. Recently, a new generalized fuzzy set called picture fuzzy set (PFS) has been proposed by Cuong and Kreinovich in Ref. [6]. The word "picture" in PFS refers to generality as this set is the direct extension of FS and IFS. In the other words, PFS integrates information of neutral and negative into its definition so that when the value(s) of one (both) of those degrees is (are) equal to zero, it returns to IFS (FS) set. Comparing with IFS, PFS divides the hesitancy degree into two parts, i.e., refusal degree and neutral degree (see Definition 1 and Examples 1 and 2 for details). This set shows great promises of better adaptation to many practical problems in pattern recognition, artificial life, robotic, expert and knowledge-based systems than some existing types of fuzzy sets.

Definition 1. A picture fuzzy set (PFS) [6] in a non-empty set X is,

$$A = \left\{ \langle x, \mu_A(x), \eta_A(x), \gamma_A(x) \rangle | x \in X \right\},$$

where $\mu_A(x)$ is the positive degree of each element $x \in X$, $\eta_A(x)$ is the neutral degree and $\gamma_A(x)$ is the negative degree satisfying the constraints,

$$\mu_A(x), \eta_A(x), \gamma_A(x) \in [0, 1], \forall x \in X,$$

$$0 \le \mu_A(x) + \eta_A(x) + \gamma_A(x) \le 1, \forall x \in X.$$

The refusal degree of an element is calculated as $\xi_A(x) = 1 - (\mu_A(x) + \eta_A(x) + \gamma_A(x))$, $\forall x \in X$. In the case $\eta_A(x) = 0$ PFS returns to the IFS set, and when both $\eta_A(x) = \gamma_A(x) = 0$, PFS returns to the FS set. Some properties of PFS operations, the convex combination of PFS, etc. accompanied with proofs can be referenced in Ref. [6].

Example 1. In a democratic election station, the council issues 500 voting papers for a candidate. The voting results are divided into four groups accompanied with the number of papers namely "vote for" (300), "abstain" (64), "vote against" (115) and "refusal of voting" (21). Group "abstain" means that the voting paper is a white paper rejecting both "agree" and "disagree" for the candidate but still takes the vote. Group "refusal of voting" is either invalid voting papers or bypassing the vote. This example was happened in reality and IFS could not handle it since the neutral membership (group "abstain") does not exist.

Example 2. Personnel selection is a very important activity in the human resource management of an organization. The process of selection follows a methodology to collect information about

E-mail addresses: sonlh@vnu.edu.vn, chinhson2002@gmail.com

http://dx.doi.org/10.1016/j.asoc.2016.05.009
1568-4946/

an individual in order to determine if that individual should be employed. The selection results could be classified into 4 classes: true positive, true negative, false negative, and false positive which are somehow equivalent to the positive, neutral, negative and refusal degrees of PFS. Each candidate is ranked according to 4 classes by his ability and suitability for the job, and the final decision is made based on results of the classes. For example, if two candidates are ranked A-(50%, 20%, 20%, 10%) and B-(40%, 10%, 30%, 20%), the final decision can be made through the union operator and maximum of the positive degree in PFS which returns the value of 50% (A is selected).

An emerging trend in PFS and other advanced fuzzy sets is the development of soft computing methods especially *clustering algorithms* on these sets, which could produce better quality of results than that on FS. For instance, clustering algorithms on interval T2FS focusing on uncertainty associated with the fuzzifier were investigated in Refs. [14,52]. Regarding the IFS set, Pelekis et al. [23] proposed a clustering approach utilizing a similarity-metric defined over IFS. Xu and Wu [45] developed the IFCM algorithm to classify IFS and interval-valued IFS. Son et al. [26] proposed an intuitionistic fuzzy clustering algorithm for geo-demographic analysis. Xu and his group developed a number of intuitionistic fuzzy clustering methods in various contexts [36,37,39,42]. Fuzzy clustering algorithms on other sets namely HFS and PFS were found in Refs. [4,27]. It is clear from the literature that *distance measure* is the most important factor for an efficient clustering algorithm. The most widely used distance measures for two FSs A and B on $X = \{X_1, .., X_N\}$ is the Hamming, Euclidean and Hausdorff metrics [6]. Because of the FS's drawbacks, distance measures on other sets mostly IFS have been proposed. Atanassov [3], Chen [5], Dengfeng and Chuntian [7], Grzegorzewski [10], Hatzimichailidis et al. [11], Hung and Yang [12,13], Li et al. [16], Liang and Shi [17], Mitchell [21], Papakostas et al. [22], Szmidt and Kacprzyk [28–30], Wang and Xin [35], Xu and Chen [41], Xu and Xia [46], Yang and Chiclana [47] and Xu [44] presented some distance measures in IFS namely the (normalized) intuitionistic Hamming and Euclidean distances, and the (normalized) Hausdorff intuitionistic Hamming and Euclidean distances. A basic distance measure on PFS has been given by Cuong and Kreinovich [6] as follows.

$$d_P(A,B) = \left(\frac{1}{N} \sum_{i=1}^{N} \left((\mu_A(x_i) - \mu_B(x_i))^p + (\eta_A(x_i) - \eta_B(x_i))^p + (\gamma_A(x_i) - \gamma_B(x_i))^p \right) \right)^{1/p}.$$

We recognize that $d_P(A, B)$ is a generalization of those in IFS and FS when $\eta_A(x) = 0$ and both $\eta_A(x) = \gamma_A(x) = 0$, respectively. As explained above, the integration of neutral degree $\eta_A(x)$ would measure information of objects more accurately and increase quality and accuracy of achieved results. Yet again, to help improving the performance as motivated by the previous researches on IFS that tended to combine some basic distance measures into a complex one to improve the generality and accuracy, in this paper we propose a novel generalized picture distance measure and use it in a new clustering method on PFS called Hierarchical Picture Clustering (HPC). The reason for designing a new measure can be illustrated by an example as follows. Consider that we would like to measure the truth-value of the proposition G = "through a point exterior to a line one can draw only one parallel to the given line". The proposition is incomplete, since it does not specify the type of geometrical space it belongs to. In an Euclidean geometric space the proposition G is true; in a Riemannian geometric space the proposition G is false (since there is no parallel passing through an exterior point to a given line); in a geometric space covering the PFS set (constructed from mixed spaces, for example from a part of Euclidean subspace together with another part of Riemannian space) the proposition G is indeterminate (true and false in the same time) [48]. It is obvious that objects, notions, ideas, etc. can be better measured in PFS than in other types of fuzzy sets.

The main differences of the proposed distance measure with $d_P(A, B)$ and those on IFS such as in Xu [44] are highlighted as follows.

Firstly, as being shown above, $d_P(A, B)$ is a natural expansion of the well-known Minkowski distance of order $p \geq 1$ between two points under fuzzy environments. When $p = 1$ or $p = 2$, we have the Manhattan and Euclidean distances, respectively. In the limiting case of p reaching infinity, we obtain the Chebyshev distance. The Minkowski distance has the best performance for numerical data but works ineffectively with asymmetric binary variables, non-metric vector objects, etc. [20]. For example, the similarity between two vectors can be denoted as a cosine measure which is further used to define a distance [48]. For asymmetric binary variables, the contingency table, which reflects the matching states between two objects, is used to compute the distance between asymmetric binary variables [25]. It is very often that a non-linear function is adopted as the distance metric for processing non-spherical data [9]. One of the most common ways to create such the function is combining the basic distance measures into a complex one so that the deficiencies of the standalone metrics are settled. This intuition leads to debut of the proposed measure which may enhance performance and accuracy of results.

Secondly, the proposed measure is a combination of the Hamming, Euclidean and Hausdorff distances. It is different to $d_P(A, B)$ which in essence is the normalized form of well-known Minkowski distance of order $p \geq 1$. In the next section, we will explain why the hybridization should be made and emphasize on the advantages and disadvantages of using the proposed measure. However, it is noted that the proposed distance measure is a generalization version of $d_P(A, B)$.

Thirdly, the proposed distance measure is different to those on IFS such as in Xu [44] in many aspects. Let us take some examples. In Ref. [44], Xu generalized the intuitionistic Hamming and Euclidean distances of Szmidt and Kacprzyk [28] as below.

$$dd(A,B) = \left(\frac{1}{2N} \sum_{i=1}^{N} \left(|\mu_A(x_i) - \mu_B(x_i)|^\alpha + |\eta_A(x_i) - \eta_B(x_i)|^\alpha + |\gamma_A(x_i) - \gamma_B(x_i)|^\alpha \right) \right)^{1/\alpha}.$$

He then defined several similarity measures from the above distance function, for instance:

$$s(A,B) = 1 - \left(\frac{1}{2N} \sum_{i=1}^{N} \left(|\mu_A(x_i) - \mu_B(x_i)|^\alpha + |\eta_A(x_i) - \eta_B(x_i)|^\alpha + |\gamma_A(x_i) - \gamma_B(x_i)|^\alpha \right) \right)^{1/\alpha},$$

$$s(A,B) = 1 - \left(\frac{\sum_{i=1}^{N} \left(|\mu_A(x_i) - \mu_B(x_i)|^\alpha + |\eta_A(x_i) - \eta_B(x_i)|^\alpha + |\gamma_A(x_i) - \gamma_B(x_i)|^\alpha \right)}{\sum_{i=1}^{N} \left(|\mu_A(x_i) + \mu_B(x_i)|^\alpha + |\eta_A(x_i) + \eta_B(x_i)|^\alpha + |\gamma_A(x_i) + \gamma_B(x_i)|^\alpha \right)} \right)^{1/\alpha}$$

Even though $d(A, B)$ is quite similar to $d_P(A, B)$, we recognize that $d(A, B)$ is designed on the basis of IFS which means $\mu_A(x) + \eta_A(x) + \gamma_A(x) = 1$ while $d_P(A, B)$ is the distance on PFS satisfying $0 \leq \mu_A(x) + \eta_A(x) + \gamma_A(x) \leq 1$. Indeed, it is not intuitive and logical when taking the difference between $\eta_A(x)$ and $\eta_B(x)$ since