Quantitative Comparison of Similarity Measure and Entropy for Fuzzy Sets

Hongmei Wang, Sanghyuk Lee[*], and Jaehyung Kim

School of Mechatronics, Changwon National University
Sarim-dong, Changwon, Gyeongnam, Korea
iwanghongmei99@163.com, {leehyuk,hyung}@changwon.ac.kr

Abstract. Comparison and data analysis to the similarity measures and entropy for fuzzy sets are studied. The distance proportional value between the fuzzy set and the corresponding crisp set is represented as fuzzy entropy. We also verified that the sum of the similarity measure and the entropy between fuzzy set and the corresponding crisp set constitutes the total information. Finally, we derive a similarity measure from entropy with the help of total information property, and illustrate a simple example that the maximum similarity measure can be obtained using a minimum entropy formulation.

Keywords: Similarity measure, distance measure, fuzzy entropy.

1 Introduction

Analysis of data certainty and uncertainty is essential to process data mining, pattern classification or clustering, and discriminating data. Basically, well known distance measure such as Hamming distance can be used to design certainty and uncertainty measure commonly. To analyze the data, it is often useful to consider a data set as a fuzzy set with a degree of membership. Hence fuzzy entropy and similarity analyses have been emphasized for studying the uncertainty and certainty information of fuzzy sets [1-7].

The characterization and quantification of fuzziness needed in the management of uncertainty in the modeling and system designs. The entropy of a fuzzy set is called as the measure of its fuzziness by previous researchers [1-4]. The degree of similarity between two or more data sets can be used in the fields of decision making, pattern classification, etc., [5-7]. Thus far, numerous researchers have carried out research on deriving similarity measures [8,9]. Similarity measures based on the distance measure are applicable to general fuzzy membership functions, including nonconvex fuzzy membership functions [9]. Two measures, entropy and similarity, represent the uncertainty and similarity with respect to the corresponding crisp set, respectively. For data set, it is interesting to study the relation between entropy and similarity measure. The correlation between entropy and similarity for fuzzy sets has been presented as the physical view [10]. Liu also proposed a relation between distance and similarity measures; in his paper, the sum of distance and similarity constitutes the total

[*] Corresponding Author.

R. Huang et al. (Eds.): ADMA 2009, LNAI 5678, pp. 688–695, 2009.
© Springer-Verlag Berlin Heidelberg 2009

information [2]. In this paper, we analyze the relationship between the entropy and similarity measures for fuzzy sets, and compute the quantitative amount of corresponding measures. First, fuzzy entropy and similarity measures are derived by the distance measure. Discussion with entropy and similarity for data has been followed. With the proposed fuzzy entropy and similarity measure, the property that the total information comprises the similarity measure and entropy measure is verified. With the total information property, similarity measure can be obtained through fuzzy entropy. Two examples help to understand total information property between fuzzy entropy and similarity as the uncertainty and certainty measure of data.

In the following section, the relationship between entropy and similarity for a fuzzy set is discussed. In Section 3, the procedure for obtaining the similarity measure from the fuzzy entropy is derived. Furthermore, the computational examples are illustrated. The relation between entropy and similarity are clearly verified, and selection of reliable data is carried out by applying similarity measure and entropy. Discussions are followed at the end of Section 3. The conclusions are stated in Section 4.

2 Fuzzy Entropy and Similarity Measure

Every data set has uncertainty in its data group, and it is illustrated as the membership functions for fuzzy set. Data uncertainties are often measured by fuzzy entropy, explicit fuzzy entropy construction is also proposed by numerous researchers [9]. Fuzzy entropy of fuzzy set means that fuzzy set contains how much uncertainty with respect to the corresponding crisp set. Then, what is the data certainty with respect to the deterministic data? This data certainty can be obtained through similarity measure. We analyze the relation between fuzzy entropy and similarity as the quantitative comparison.

Two comparative sets are considered, one is a fuzzy set and the other is the corresponding crisp set. The fuzzy membership function pair is illustrated in Fig. 1, crisp set A_{near} represents the crisp set “near” to the fuzzy set A .

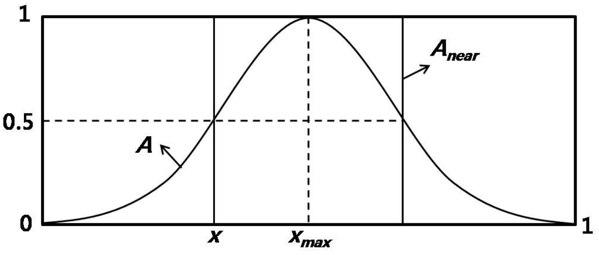

Fig. 1. Membership functions of fuzzy set A and crisp set $A_{near} = A_{0.5}$

A_{near} can be assigned by various variable. For example, the value of crisp set $A_{0.5}$ is one when $\mu_A(x) \geq 0.5$, and is zero otherwise. Here, A_{far} is the complement of A_{near} , i.e., $A^C{}_{near} = A_{far}$. In our previous result, the fuzzy entropy of fuzzy set A with respect to A_{near} is represented as follows [9]:

$$e(A, A_{near}) = d(A \cap A_{near}, [1]_X) + d(A \cup A_{near}, [0]_X) - 1 \tag{1}$$

where $d(A \cap A_{near}, [1]_X) = \frac{1}{n}\sum_{i=1}^{n} | \mu_{A \cap A_{near}}(x_i) - 1 |$ is satisfied. $A \cap A_{near}$ and $A \cup A_{near}$ are the minimum and maximum value between A and A_{near}, respectively. $[0]_X$ and $[1]_X$ are the fuzzy sets in which the value of the membership functions are zero and one, respectively, for the universe of discourse. d satisfies Hamming distance measure. Eq. (1) is not the normal fuzzy entropy. The normal fuzzy entropy can be obtained by multiplying the right-hand side of Eq. (1) by two, which satisfies maximal fuzzy entropy is one. Numerous fuzzy entropies can be presented by the fuzzy entropy definition [2]. Here we introduce same result of (1) as follows.

$$e(A, A_{near}) = d(A, A \cap A_{near}) + d(A_{near}, A \cap A_{near}) \tag{2}$$

The fuzzy entropies in Eqs. (1) and (2) satisfy for all value of crisp set A_{near}. Hence, $A_{0.1}$ and $A_{0.5}$ or some other $A_{0.X}$ can be satisfied. Now, it is interesting to search for what value of $A_{0.X}$ make maximum or minimum value of entropy.

Eqs. (1) and (2) are rewritten as follows:

$$e(A, A_{near}) = 2\int_0^x \mu_A(x)dx + 2\int_x^{x_{\max}} 1 - \mu_A(x)dx\,. \tag{3}$$

Let $\frac{d}{dx}M_A(x) = \mu_A(x)$; $e(A, A_{near})$ has been shown to be

$$e(A, A_{near}) = 2M_A(x)\mid_0^x +2(x_{\max} - x) - 2M_A(x)\mid_x^{x_{\max}} \cdot$$

The maxima or minima are obtained by differentiation:

$$\frac{d}{dx}e(A, A_{near}) = 2\mu_A(x) - 2 + 2\mu_A(x)\,.$$

Hence, it is clear that the point x satisfying $\frac{d}{dx}e(A, A_{near}) = 0$ is the critical point for the crisp set. This point is given by $\mu_A(x) = 1/2$, i.e., $A_{near} = A_{0.5}$. The fuzzy entropy between A and $A_{0.5}$ has a minimum value because $e(A)$ attains maxima when the corresponding crisp sets are $A_{0.0}$ and $A_{x_{\max}}$. Hence, for a convex and symmetric fuzzy set, the minimum entropy of the fuzzy set is equal to that of the crisp set $A_{0.5}$. This indicates that the corresponding crisp set that has the least uncertainty or the greatest similarity with the fuzzy set is $A_{0.5}$.

All the studies on similarity measures deal with derivations of similarity measures and applications in the distance-measure-based computation of the degree of similarity. Liu has also proposed an axiomatic definition of the similarity measure [2]. The similarity measure $\forall A, B \in F(X)$ and $\forall D \in P(X)$ has the following four properties:

(S1) $s(A,B)=s(B,A)$, $\forall A,B\in F(X)$

(S2) $s(D,D^c)=0$, $\forall D\in P(X)$

(S3) $s(C,C)=\max_{A,B\in F} s(A,B)$, $\forall C\in F(X)$

(S4) $\forall A,B,C\in F(X)$, if $A\subset B\subset C$, then $s(A,B)\geq s(A,C)$ and $s(B,C)\geq s(A,C)$

where $F(X)$ denotes a fuzzy set, and $P(X)$ is a crisp set.

In our previous studies, other similarity measures between two arbitrary fuzzy sets are proposed as follows [9]:

For any two sets $A,B\in F\left(X\right)$,

$$s(A,B)=1-d(A\cap B^C,[0]_X)-d(A\cup B^C,[1]_X) \tag{4}$$

and

$$s(A,B)=2-d(A\cap B,[1]_X)-d\left(A\cup B,[0]_X\right) \tag{5}$$

are similarity measures between set A and set B .

The proposed similarity measure between A and A_{near} is presented in Theorem 2.1The usefulness of this measure is verified through a proof of this theorem.

Theorem 2.1. $\forall A\in F(X)$ and the crisp set A_{near} in Fig. 1,

$$s(A,A_{near})=\ d(A\cap A_{near},[0]_X)\ +\ d(A\cup A_{near},[1]_X) \tag{6}$$

is a similarity measure.

Proof. (S1) follows from Eq. (6), and for crisp set D , it is clear that $s(D,D^C)=0$. Hence, (S2) is satisfied. (S3) indicates that the similarity measure of two identical fuzzy sets $s(C,C)$ attains the maximum value among various similarity measures with different fuzzy sets A and B since $d(C\cap C,[0]_X)+d(C\cup C,[1]_X)$ represents the entire region in Fig. 1. Finally, from $d(A\cap A_{1near},[0]_X)\geq d(A\cap A_{2near},[0]_X)$ and $d(A\cup A_{1near},[1]_X)\geq d(A\cup A_{2near},[1]_X)$, $A\subset A_{1near}\subset A_{2near}$; it follows that

$$\begin{aligned} s(A,A_{1near}) &= d(A\cap A_{1near},[0]_X)+d(A\cup A_{1near},[1]_X)\\ &\geq d(A\cap A_{2near},[0]_X)+d(A\cup A_{2near},[1]_X)\\ &= s(A,A_{2near}) \end{aligned}.$$

Similarly, $s(A_{1near},A_{2near})\geq s(A,A_{2near})$ is satisfied by the inclusion properties $d(A_{1near}\cap A_{2near},[0]_X)\geq d(A\cap A_{2near},[0]_X)$ and $d(A_{1near}\cup A_{2near},[1]_X)\geq d(A\cup A_{2near},[1]_X)$. ■

The similarity in Eq. (6) represents the areas shared by two membership functions. In Eqs. (4) and (5), fuzzy set B can be replaced by A_{near} . In addition to those in Eqs. (4) and (5), numerous similarity measures that satisfy the definition of a similarity measure can be derived. From Fig. 1, the relationship between data similarity and entropy for fuzzy set A with respect to A_{near} can be determined on the basis of the total area. The total area is one (universe of discourse × maximum membership value = $1\times 1=1$); it represents the total amount of information. Hence, the total information comprises the similarity measure and entropy measure, as shown in the following equation:

$$s(A,A_{near})+e(A,A_{near})\ =1 \tag{7}$$

With the similarity measure in Eq. (5) and the total information expression in Eq. (7), we obtain the following proposition:

Proposition 2.1. In Eq. (7), $e(A,A_{near})$ follows from the similarity measure in Eq. (5):

$$e(A,A_{near})=1-s(A,A_{near})\ =d(A\cap A_{near},[1]_X)+d(A\cup A_{near},[0]_X)-1$$

The above fuzzy entropy is identical to that in Eq. (1). The property given by Eq. (7) is also formulated as follows:

Theorem 2.2. The total information about fuzzy set A and the corresponding crisp set A_{near},

$$\begin{aligned}&s(A,A_{near})+e(A,A_{near})\\&=d(A\cap A_{near},[0]_X)+d(A\cup A_{near},[1]_X)\\&\quad+d(A\cap A_{near},[1]_X)+d(A\cup A_{near},[0]_X)-1\end{aligned}$$

equals one.

***Proof*.** It is clear that the sum of the similarity measure and fuzzy entropy equals one, which is the total area in Fig. 1. Furthermore, it is also satisfied by computation,

$$d(A\cap A_{near},[0]_X)\ +d(A\cap A_{near},[1]_X)=1 \text{ and}$$
$$d(A\cup A_{near},[1]_X)\ +d(A\cup A_{near},[0]_X)=1\,.$$

Hence, $s(A,A_{near})+e(A,A_{near})=1+1-1=1$ is satisfied. ■

Now, it is clear that the total information about fuzzy set A comprises similarity and entropy measures with respect to the corresponding crisp set.

3 Similarity Measure Derivation from Entropy

In this section, with the property of Theorem 2.2 similarity measure derivation with entropy is carried out. Entropy derivation from similarity is also possible. This conversion makes possible measure formulation from complementary measure. With consideration of previous similarity measure (5), fuzzy entropy (1) has been obtained. It is also possible to obtain another similarity measure using fuzzy entropy different from that in Eq. (2). The proposed fuzzy entropy is developed by using the Hamming distances between a fuzzy set and the corresponding crisp set. The following result clearly follows from Fig. 1. Eq. (2) represents the difference between A and the corresponding crisp set A_{near}. From Theorem 2.2, the following similarity measure that satisfies Eq. (7) follows:

$$s(A, A_{near}) = 1 - d(A, A \cap A_{near}) - d(A_{near}, A \cap A_{near}) \tag{8}$$

Here, it is interesting to determine whether Eq. (8) satisfies the conditions for a similarity measure.

Proof. (S1) follows from Eq. (8). Furthermore, $s(D, D^C) = 1 - d(D, D \cap D^C) - d(D^C, D \cap D^C)$ is zero because $d(D, D \cap D^C) + d(D^C, D \cap D^C)$ satisfies $d(D,[0]_X) + d(D^C,[0]_X) = 1$. Hence, (S2) is satisfied. (S3) is also satisfied since $d(C, C \cap C) + d(C, C \cap C) = 0$; hence, it follows that $s(C, C)$ is a maximum. Finally,

$$1 - d(A, A \cap B) - d(B, A \cap B) \geq 1 - d(A, A \cap C) - d(C, A \cap C)$$

because $d(A, A \cap B) = d(A, A \cap C)$ and $d(B, A \cap B) \leq d(C, A \cap C)$ are satisfied for $A \subset B \subset C$. The inequality $s(B, C) \geq s(A, C)$ is also satisfied in a similar manner. ■

Similarity based on fuzzy entropy has the same structure designed from similarity definition. Following corollary insist that two similarity measures has the same structure even though they are derived from different ways.

Corollary 3.1. Proposed similarity measures (6) and (8) are equal.

$$d(A \cap A_{near},[0]_X) + d(A \cup A_{near},[1]_X) = 1 - d(A, A \cap A_{near}) - d(A_{near}, A \cap A_{near}).$$

This equality can be verified easily by analyzing Fig 1.

Now, by using Eq. (8), we obtain the maximum similarity measure for the fuzzy set. In our previous result, the minimum fuzzy entropy could be obtained when we considered the entropy between the fuzzy sets A and $A_{0.5}$. Hence, it is obvious that the obtained similarity

$$s(A, A_{0.5}) = 1 - d(A, A \cap A_{0.5}) - d(A_{0.5}, A \cap A_{0.5}) \tag{9}$$

represents the maximum similarity measure.

Computation between Similarity measure and Fuzzy entropy : Let us consider the next fuzzy set with membership function $A = \{x, \mu_A(x)\}$:

$$\{(0.1,0.2), (0.2,0.4), (0.3,0.7), (0.4,0.9), (0.5,1), (0.6,0.9), (0.7, 0.7), (0.8,0.4), (0.9,0.2), (1,0)\}.$$

The fuzzy entropy and similarity measures are calculated using Eqs. (2) and (9) are given in Table 1.

Table 1. Similarity and Entropy value between fuzzy set and corresponding crisp set

Similarity measure	Measure value	Fuzzy entropy	Entropy value
$s(A, A_{0.1})$	0.64	$e(A, A_{0.1})$	0.36
$s(A, A_{0.3})$	0.76	$e(A, A_{0.3})$	0.24
$s(A, A_{0.5})$	0.80	$e(A, A_{0.5})$	0.20
$s(A, A_{0.8})$	0.72	$e(A, A_{0.8})$	0.28
$s(A, A_{0.95})$	0.56	$e(A, A_{0.95})$	0.44

The similarity measure for $s(A, A_{0.5})$ is calculated by using the following equation:

$$s(A, A_{0.5}) = 1 - 1/10(0.2 + 0.4 + 0.4 + 0.2) \ - 1/10(0.3 + 0.1 + 0.1 + 0.3) = 0.8.$$

Fuzzy entropy $e(A, A_{0.5})$ is also calculated by

$$e(A, A_{0.5}) = 1/10(0.2 + 0.4 + 0.4 + 0.2) + 1/10(0.3 + 0.1 + 0.1 + 0.3) = 0.2.$$

The remaining similarity measures and fuzzy entropies are calculated in a similar manner.

4 Conclusions

Quantification of fuzzy entropy and similarity for fuzzy sets were studied. Fuzzy entropies for fuzzy sets were developed by considering the crisp set “near” the fuzzy set. The minimum entropy can be obtained by calculating area, and it satisfies when the crisp set is $A_{near} = A_{0.5}$. The similarity measure between the fuzzy set and the corresponding crisp set is also derived using the distance measure. The property that the sum of fuzzy entropy and the similarity measure between fuzzy set and corresponding crisp set is derived as a constant value. It is proved that the fuzzy entropy and similarity measure values constitute whole area of information.

Acknowledgments. The authors of this paper were partly supported by the Second Stage of Brain Korea 21 Projects.

References

1. Pal, N.R., Pal, S.K.: Object-background segmentation using new definitions of entropy. In: IEEE Proc., vol. 36, pp. 284–295 (1989)
2. Xuecheng, L.: Entropy, distance measure and similarity measure of fuzzy sets and their relations. Fuzzy Sets and Systems 52, 305–318 (1992)
3. Bhandari, D., Pal, N.R.: Some new information measure of fuzzy sets. Inform. Sci. 67, 209–228 (1993)

4. Ghosh, A.: Use of fuzziness measure in layered networks for object extraction: a generalization. Fuzzy Sets and Systems 72, 331–348 (1995)
5. Rébillé, Y.: Decision making over necessity measures through the Choquet integral criterion. Fuzzy Sets and Systems 157(23), 3025–3039 (2006)
6. Kang, W.S., Choi, J.Y.: Domain density description for multiclass pattern classification with reduced computational load. Pattern Recognition 41(6), 1997–2009 (2008)
7. Shih, F.Y., Zhang, K.: A distance-based separator representation for pattern classification. Image and Vision Computing 26(5), 667–672 (2008)
8. Chen, S.J., Chen, S.M.: Fuzzy risk analysis based on similarity measures of generalized fuzzy numbers. IEEE Trans. on Fuzzy Systems 11(1), 45–56 (2003)
9. Lee, S.H., Kim, J.M., Choi, Y.K.: Similarity measure construction using fuzzy entropy and distance measure. In: Huang, D.-S., Li, K., Irwin, G.W. (eds.) ICIC 2006. LNCS (LNAI), vol. 4114, pp. 952–958. Springer, Heidelberg (2006)
10. Lin, S.K.: Gibbs Paradox and the Concepts of Information, Symmetry, Similarity and Their Relationship. Entropy 10, 15 (2008)