Chapter 33
Similarity Measure Design for Non-Overlapped Data

Sanghyuk Lee

Abstract Study on similarity measure on fuzzy sets (FSs) for the case of non-overlapped data was proposed, and analyzed. Comparison with similarity measure on overlapped case was done. Different approach to similarity measure was analyzed, and adequate similarity measure on non-overlapped data was designed by considering neighbor information. With artificial data rational calculation results were obtained.

Keywords Similarity measure · Non-overlapped data · Intuitionistic data

33.1 Introduction

Analysis on fuzzy data provides useful information background to data analysis by way of heuristic point of view, it has been carried out by numerous researchers [1–4]. Specially, study on evaluation of uncertainty and certainty with respect to the corresponding fact was carried out by way of fuzzy entropy and similarity measure design [4–11]. Similarity measure guarantees the similarity degree between comparing data sets. Obtained results have been applied to solve pattern recognition and clustering problem or etc. [12]. Basically, measure is defined on non-empty class $\mathcal{C}$ over whole non-empty set X [14]. That μ is called a measure on $\mathcal{C}$ iff it is countably additive and there exists $E \in \mathcal{C}$ such that $\mu(E) < \infty$.

Designing similarity measure was based on its definition. Its characteristics are commutativity, complementary feature, overlapped characteristics, and triangular inequality feature. Based on these properties, similarity measure was derived by two approaches [7–11], fuzzy number [7, 8] and distance measure [9–11].

S. Lee (✉)
Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
e-mail: Sanghyuk.Lee@xjtlu.edu.cn

H.-K. Jung et al. (eds.), *Future Information Communication Technology and Applications*, Lecture Notes in Electrical Engineering 235, DOI: 10.1007/978-94-007-6516-0_33,

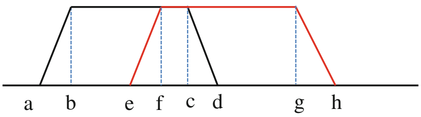


Fig. 33.1 Fuzzy membership function with fuzzy number

Similarity measure design with fuzzy number was seemed to be easier than with distance measure, because only finite number of membership values were constituted to construct similarity measure [7, 8]. However, this approach has fatal drawback, measure design should be restricted only triangular or trapezoidal fuzzy membership function. Whereas similarity measure with distance measure can be applied to unlimited membership function even its design procedure is rather complex and tedious.

Whether the similarity measures are proposed by fuzzy number or distance measure, it provided the degree of similarity between data sets. Similarity with fuzzy number was used with finite number of data. It means there was no needs to be overlapped, because it is depend on combination with $\{a,b,c,d\}$ and $\{e,f,g,h\}$ Fig. 33.1.

Whereas similarity measure design with distance measure was considered distance information between two membership functions. It means that the obtained similarity measure cannot guarantee the similarity calculation of non-overlapped singleton distributed data. Hence, in order to design similarity measure for singleton data it needs different approach. In the similarity measure design with distance measure, measure structure should be related with the same support of universe of discourse. Hence, to consider non-overlapped data, we deleted the assumption about same support of universe of discourse. In this literature, similarity measure for non-overlapped data would be derived by considering neighbor data. By comparing each data with whole neighbor data information, similarity measure was completed. The obtained measure was proved and applied to artificial example. Computation result was also compared with conventional similarity measure.

In the following chapter, preliminary results on similarity were proposed. Similarity measure based on fuzzy number and distance measure were introduced, and it was applied to discrete date. Non-consistency was shown by calculation results. Similarity measure on non-overlapped data was proposed and proved in Chap. 3. Similarity measure calculation for non-overlapped data was also proposed in the same chapter, and it was analyzed. Calculation results were seemed to be rationale. Finally, conclusions are followed in Chap. 4. Notations of this literature are used from reference of Liu [4].

33.2 Preliminaries

Similarity measure was proposed by Liu [4]. It was designed by using distance measure. It satisfies four properties of similarity measure.

Definition 2.1 [4] A real function $s : F^2 \rightarrow R^+$ is called a similarity measure, if s has the following properties:

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

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

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

(S4) $A,B,C \in F(X)$, if $A \subset B \subset C$, then $s(A,B) \geq s(A,C)$ and $s(B,C) \geq s(A,C)$. where $R^+ = [0,\infty), X$ is total set, $F(X)$ is the class of all fuzzy sets of X, $P(X)$ is the class of ordinary sets of X, and D^C is the complement set of D. By this definition, numerous similarity measures could be derived.

In order to design the similarity measure via distance, it is needed to introduce the distance measure [4]. Similarity measure can be represented as explicit structure with help of distance measure.

Definition 2.2 A real function $d : F^2 \rightarrow R^+$ is called a distance measure on F if d satisfies the following properties:

(D1) $d(A,B) = d(B,A), \quad A,B \in F(X)$

(D2) $d(A,A) = 0, \quad A \in F(X)$

(D3) $d(D,D^C) = max_{A,B \in F} d(A,B),\ D \in F(X)$

(D4) $A,B,C \in F(X)$, if $A \subset B \subset C$, then $d(A,B) \leq d(A,C)$ and$d(B,C) \leq d(A,C)$.

Hamming distance was commonly used as distance measure between fuzzy sets A and B,

$$d(A,B) = \frac{1}{n}\sum_{i=1}^{n} |\mu_A(x_i) - \mu_B(x_i)|$$

where $X = \{x_1,\ x_2,\ \ldots,\ x_n\}, |k|$ was the absolute value of k. $\mu A(x)$ was the membership function of $A \in F(X)$. Following theorem satisfied similarity measure.

Theorem 2.1 For any set $A,B \in F(X)$, if d satisfies Hamming distance measure, then

$$s(A,B) = 1 - d((A \cap B),(A \cup B)) \tag{33.1}$$

is the similarity measure between set A and B.

Proof Commutativity of (S1) is clear from (33.1) itself, that is

$$s(A,B) = s(B,A).$$

For (S2),

$$\begin{aligned} s\big(D,D^C\big) &= 1 - d\big(\big(D \cap D^C\big),\big(D \cup D^C\big)\big) \\ &= 1 - d\big([0]_X,[1]_X\big) = 0 \end{aligned}$$

is obtained because of $(D \cap D^C) = [0]_X$ and $(D \cup D^C) = [1]_X$. Where, $[0]_X$ and $[1]_X$ denote zero and one over whole universe of discourse of X.

(S3) is also easy to prove,

$$\begin{aligned} s(C,C) &= 1 - d((C \cap C),(C \cup C)) \\ &= 1 - d(C,C) = 1 \end{aligned}$$

It is natural that $d(C, C) = [0]_X$, satisfied maximal value.

Finally,

$$d((A \cap B),(A \cup B)) = d(A,B) < d(A,C) = d((A \cap C),(A \cup C))$$

guarantees $s(A, C) < s(A, B)$, and

$$d((B \cap C),(B \cup C)) = d(B,C) < d(A,C) = d((A \cap C),(A \cup C))$$

also provides $s(A, C) < s(A, B)$ therefore triangular inequality is obvious by the definition, hence (S4) is also satisfied.

Besides Theorem 2.1, numerous similarity measures are possible. Other similarity measure is also illustrated in Theorem 2.2, and its proof is also found in previous results [9–11].

Theorem 2.2 For any set $A, B \in F(X)$, if d satisfies Hamming distance measure, then

$$s(A,B) = 1 - d(A, A \cap B) - d(B, A \cap B) \tag{33.2}$$

$$s(A,B) = 2 - d\big((A \cap B),[1]_X\big) - d\big((A \cup B),[0]_X\big) \tag{33.3}$$

$$s(A,B) = d\big((A \cap B),[0]_X\big) + d((A \cup B),[1]_X) \tag{33.4}$$

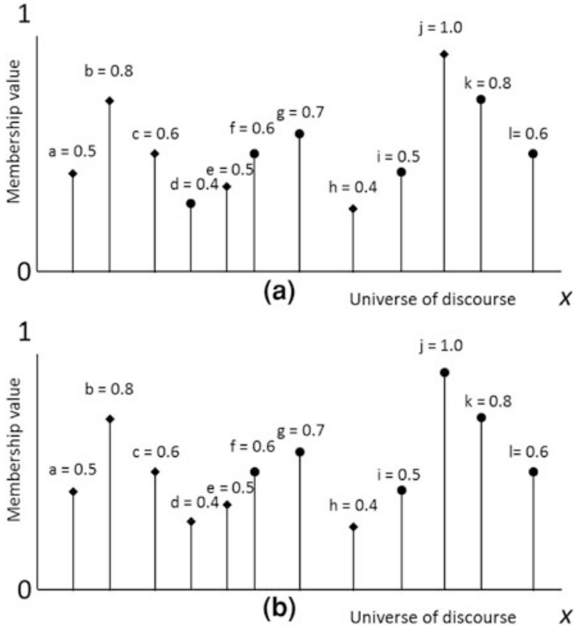
are the similarity measure between set A and set B.

Mentioned similarity measure was verified its usefulness by proof of its definition. Consider the following example, data are distributed discrete singletons and non-overlapped. Two data pairs that constitute different distributions are considered in Fig. 33.2. Twelve data with six diamonds (♦) and six circles (•) are illustrated with different combination in Fig. 33.2a and b. Similarity degree between circles and diamonds must be different between Fig. 33.2a and Fig. 33.2b because of different distribution. Two different data in Fig. 33.2a are less discriminate than Fig. 33.2b. It means that similarity measure of Fig. 33.2a has higher value than Fig. 33.2b. Next, similarity calculations are carried out with conventional similarity measure such as (33.1), (33.2), (33.3), and (33.4) at first.

First, In order to compare different distributions, six diamonds (♦) and six circles (•) show same magnitude. Not only analytical but also heuristic point of views, two distribution pair must show different similarity measure. By calculation of similarity measures (33.2), it is clear that

$$\begin{aligned} s(\blacklozenge,\bullet) &= 1 - d(\blacklozenge,\blacklozenge \cap \bullet) - d(\bullet,\blacklozenge \cap \bullet) \\ &= 1 - d(\blacklozenge,[0]_X) - d(\bullet,[0]_X) \end{aligned}$$

Fig. 33.2 **a** Data distribution between circle and diamond, **b** Data distribution between circle and diamond



However, calculation of $d(\blacklozenge,[0]_X) + d(\bullet,[0]_X)$ represents the summation of total magnitude of distribution. Then, similarity results of Fig. 33.2 (s) and (b) are equivalent by the assumption of distribution. It means that similarity measure calculation with (33.3) for Fig. 33.2a and b are same. Similarity measure with (33.3) represents as follows

$$\begin{aligned} s(\blacklozenge,\bullet) &= 2 - d\left(\left(\blacklozenge\bigcap\bullet\right),[1]_X\right) - d\left(\left(\blacklozenge\bigcup\bullet\right),[0]_X\right) \\ &= 2 - d([0]_X,[1]_X) - d\left(\left(\blacklozenge\bigcup\bullet\right),[0]_X\right) \\ &= 2 - 1 - d\left(\left(\blacklozenge\bigcup\bullet\right),[0]_X\right) \\ &= 1 - d\left(\left(\blacklozenge\bigcup\bullet\right),[0]_X\right) \end{aligned} \tag{33.5}$$

Calculation outputs show the same result, because there is no intersection, it is always satisfied

$$\blacklozenge \cap \bullet = [0]_X$$

Same results are also obtained even similarity measure (33.5) is used. Due to

$$d(\blacklozenge,[0]_X) + d(\bullet,[0]_X) = d\left(\left(\blacklozenge\bigcup\bullet\right),[0]_X\right)$$

is satisfied. Next, different magnitude distribution pair was considered. With slight change of Fig. 33.2b, following distribution was illustrated. Only slight

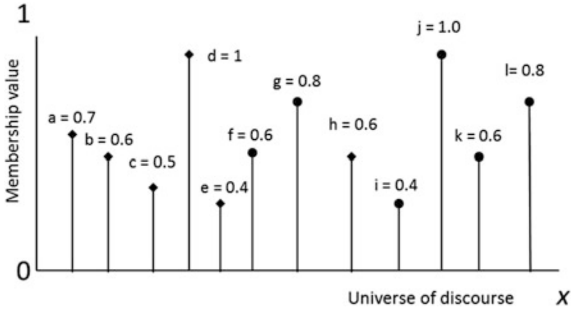


Fig. 33.3 Data distribution with different magnitude

modification of magnitude was done in Fig. 33.3. By same procedure, calculation of (33.1) to (33.4) shows that similarity measure is different from Fig. 33.2b. However, it is proportional to magnitude.

However, calculation results of Fig 33.2 and 33.3 are different each other. It means similarity measure based on distance measure was not consistency. It just provides difference between maximal value, mainly one, and singletons average. Therefore in order to analyze the degree of similarity between distributed singleton data, another similarity measure design should be needed.

By comparison with Fig. 33.2a and Fig. 33.2b, former shows more similar than b. However, the calculation output was not consistency. It means that (33.2), and (33.3) were only efficient for overlapped data distribution. Due to every operation was based overlapped data such as $A \cap B$, $A \cup B$, and Hamming distance, similarity measures (33.2) and (33.3) were only applicable to overlapped type membership function. Similarity measure structure for non-overlapped discrete data distribution is derived in next chapter.

33.3 Non-overlapped Data Analysis

Assume every data are distributed without overlapping. Then, the degree of similarity must be determined from neighbor data information. Hence, consideration of neighbor information is necessary. In the next theorem, similarity measure on non-overlapped data was derived.

Theorem 3.1 For singletons or discrete data $a, b \in P(X)$, if d satisfied Hamming distance measure, then

$$s(a,b) = 1 - |s_a - s_b| \tag{33.6}$$

is similarity measure between singleton a and b. In (33.6), s_a and s_b satisfy $d\big((a \bigcap \mathbf{R}), [1]_X\big)$ and $d\big((b \bigcap \mathbf{R}), [1]_X\big)$, respectively. Where R is whole data distribution including a and b.

Proof (S1) is clear by the definition since

$$|s_a - s_b| = |s_b - s_a|$$

For (S2),

$$\begin{aligned} s(D, D^C) &= 1 - |s_D - s_{D^c}| \\ &= 1 - \left| d\left(\left(D \bigcap \mathbf{R}\right), [1]_X\right) - d\left(\left(D^C \bigcap \mathbf{R}\right), [1]_X\right) \right| = 0 \end{aligned}$$

For D satisfies one, $d\left((D \bigcap \mathbf{R}), [1]_X\right) = 0$ and $d\left((D^C \bigcap \mathbf{R}), [1]_X\right) = 1$, hence following result is obtained. Whereas D satisfies zero, opposite results are obtained.

(S3) is clear from definition,

$$\begin{aligned} s(C, C) &= 1 - |s_C - s_C| \\ &= 1 - |d\left(\left(C \bigcap \mathbf{R}\right), [1]_X\right) - d\left(\left(C \bigcap \mathbf{R}\right), [1]_X\right)| = 1 \end{aligned}$$

Finally, (S4) $A, B, C \in F(X)$, if $A < B < C$, then

$$\begin{aligned} \mathrm{s}(A, B) &= 1 - |s_A - s_B| \\ &= 1 - |d\left(\left(A \bigcap \mathbf{R}\right), [1]_X\right) - d\left(\left(B \bigcap \mathbf{R}\right), [1]_X\right)| \\ &\geq 1 - |d\left(\left(A \bigcap \mathbf{R}\right), [1]_X\right) - d\left(\left(C \bigcap \mathbf{R}\right), [1]_X\right)| = s(A, C) \end{aligned}$$

because $d\left((B \bigcap \mathbf{R}), [1]_X\right) > d\left((C \bigcap \mathbf{R}), [0]_X\right)$ is satisfied. Similarly $\mathrm{s}(B, C) \geq s(A, C)$ is also satisfied. Hence, (33.6) also satisfies the similarity measure definition 2.1.

Similarity measure (33.6) is also designed with distance measure such as Hamming distance. As noted in before, conventional measures were not proper for non-overlapping continuous data distribution, this property is verified by the similarity measure calculation of Fig. 33.2a and b.

Next, calculate the similarity measure between circle and diamond with (33.6). For Fig. 33.2a,

$$\begin{aligned} s(\blacklozenge, \bullet) &= 1 - |d\left(\left(\blacklozenge \bigcap \mathbf{R}\right), [1]_X\right) - d\left(\left(\bullet \bigcap \mathbf{R}\right), [1]_X\right)| \\ &= 1 - 1/6|2.3 - 2.4| = 0.983 \end{aligned}$$

is satisfied.

For calculation of Fig. 33.2(b),

$$\begin{aligned} s(\blacklozenge, \bullet) &= 1 - \left| d\left(\left(\blacklozenge \bigcap \mathbf{R}\right), [1]_X\right) - d\left(\left(\bullet \bigcap \mathbf{R}\right), [1]_X\right) \right| \\ &= 1 - 1/6|2.8 - 1.8| = 0.833 \end{aligned}$$

is also obtained.

Calculation result shows that;

- Proposed similarity measure is possible to evaluate degree of similarity for non-overlapped distributions.
- First distribution pair shows better similarity.

33.4 Conclusions

Similarity measure on non-overlapped and overlapped data was designed. Two approaches to design similarity measure were introduced. Similarity measure design based on fuzzy number showed rather easy to formulate. However it has drawback for limitation of membership function. Whereas distance measure was not easy to design, however there was no limitation for the membership function. With the conventional similarity measure, calculation of similarity on non-overlapped data was carried out. Calculation results were not acceptable because conventional similarity measure was designed based on overlapped data characteristic. Hence similarity measure calculation was not compatible.

With the help of neighbor information, similarity measure on non-overlapped data was obtained. Comparison result on overlapped and non-overlapped data showed that conventional similarity measure is not useful to calculate non-overlapped discrete data. With conventional similarity calculation it was only guaranteed the distance between maximal value and data average. Hence, if non-overlapped data shows dame magnitude, always same similarity measure was given.

References

1. Zadeh LA (1965) Fuzzy sets and systems. In: Proceedings of a symposium on systems theory, Polytechnic Institute of Brooklyn, New York, pp 29–37
2. Pal NR, Pal SK (1989) Object-background segmentation using new definitions of entropy. IEEE Proc 36:284–295
3. Kosko B (1992) Neural networks and fuzzy systems. Prentice-Hall, Englewood Cliffs
4. Liu X (1992) Entropy, distance measure and similarity measure of fuzzy sets and their relations. Fuzzy Sets Syst 52:305–318
5. Bhandari D, Pal NR (1993) Some new information measure of fuzzy sets. Inform Sci 67:209–228
6. De L, Termini S (1972) A definition of non-probabilistic entropy in the setting of fuzzy entropy. J Gen Syst 5:301–312
7. Hsieh CH, Chen SH (1999) Similarity of generalized fuzzy numbers with graded mean integration representation, Proc 8th Int Fuzzy Syst Associ World Congr 2:551–555
8. Chen SJ, Chen SM (2003) Fuzzy risk analysis based on similarity measures of generalized fuzzy numbers. IEEE Trans on Fuzzy Syst 11(1):45–56
9. Lee SH, Pedrycz W, Sohn G (2009) Design of similarity and dissimilarity measures for fuzzy sets on the basis of distance measure. Int J Fuzzy Syst 11:67–72

10. Lee SH, Ryu KH, Sohn GY(2009) Study on entropy and similarity measure for fuzzy set. IEICE Trans Inf Syst E92-D:1783–1786
11. Lee SH, Kim SJ, Jang NY (2008) Design of fuzzy entropy for non convex membership function. CCIS 15:55–60
12. Wang Z, Klir GJ (1992) Fuzzy measure theory. Plenum Press, New York