A Vector Similarity Measure for Type-1 Fuzzy Sets

Dongrui Wu and Jerry M. Mendel

Signal and Image Processing Institute, Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2564
dongruiw@usc.edu, mendel@sipi.usc.edu

Abstract. Comparing the similarity between two fuzzy sets (FSs) is needed in many applications. The focus herein is linguistic approximation using type-1 (T1) FSs, i.e. associating a T1 FS A with a linguistic label from a vocabulary. Because each label is represented by an T1 FS B_i, there is a need to compare the similarity of A and B_i to find the B_i most similar to A. In this paper, a vector similarity measure (VSM) is proposed for T1 FSs, whose two elements measure the similarity in shape and proximity, respectively. A comparative study shows that the VSM gives best results. Additionally, the VSM can be easily extended to interval type-2 FSs.

1 Introduction

Fuzzy sets (FSs), which handle uncertainties in a natural way, have been used in numerous applications. The application of particular interest in this paper is the linguistic approximation problem [1,2] using type-1 (T1) FSs [1], i.e. we have a system whose inputs are linguistic labels modeled by T1 FSs, and after some operations it outputs another T1 FS A, and, we want to map A to a linguistic label in a vocabulary so that it can be understood linguistically. Because each label in the vocabulary is represented by a T1 FS B_i, there is a need to compare the similarity of A and B_i to find the B_i most similar to A.

Many similarity measures for T1 FSs have been introduced. According to Cross and Sudkamp [4], they can be classified into four categories: (1) Set-Theoretic Measures, (2) Proximity-Based Measures, (3) Logic-Based Measures, and (4) Fuzzy-Valued Measures. Two similarity measures proposed particularly for the linguistic approximation problem are Bonissone's method [1,2] and Wenstøp's method [8]. In this paper, a vector similarity measure (VSM) for T1 FSs is proposed. It is simpler than either of these two methods, and has better performance on T1 FSs. Additionally, it can be easily extended to interval T2 FSs [9].

The rest of this paper is organized as follows: Section 2 reviews Bonissone's method and Wenstøp's method for linguistic approximation using T1 FSs.

[1] In this paper we call the original FSs introduced by Zadeh [10] in 1965 *T1 FSs* to distinguish them from their extension, *type-2 FSs*, which were also introduced by Zadeh [11] in 1975 to model more uncertainties.

P. Melin et al. (Eds.): IFSA 2007, LNAI 4529, pp. 575–583, 2007.

Section 3 proposes a VSM for the linguistic approximation problem. Section 4 compares the VSM with Bonissone's method and Wenstøp's method. Section 5 draws conclusions. Proofs of the theorems are given in Appendix A.

2 Existing Similarity Measures for Linguistic Approximation

The literature on similarity measures for T1 FSs is quite extensive [4]. Two similarity measures, Bonissone's method and Wenstøp's method, which are proposed particularly for linguistic approximation, will be reviewed in this section.

2.1 Bonissone's Linguistic Approximation Distance Measure

As mentioned in the Introduction, Bonissone's [1, 2] linguistic approximation distance measure was proposed to identify the linguistic label B_i which most closely resembles a given FS A.

The first step of Bonissone's method eliminates from further consideration those linguistic labels determined to be very far away from A. For a given T1 FS A, the distances between A and B_i, $d_1(A, B_i)$, are computed to identify M B_i that are close to A (according to some tolerance parameter). Bonissone [2] first computed four T1 FS features, *centroid*, *cardinality*, *fuzziness* and *skewness*, for A and B_i, and then defined $d_1(A, B_i)$ as the weighted Euclidean distance between the two four-dimensional points $[(p_A^1, p_A^2, p_A^3, p_A^4)^T$ and $(p_{B_i}^1, p_{B_i}^2, p_{B_i}^3, p_{B_i}^4)^T]$ represented by the values of the four features for each T1 FS, i.e.,

$$d_1(A, B_i) = \left[\sum_{j=1}^{4} w_j^2 (p_A^j - p_{B_i}^j)^2\right]^{1/2}. \tag{1}$$

The weights[2] w_j $(j = 1, 2, 3, 4)$ have to be pre-specified.

After pre-screening linguistic labels far away from A, Bonissone's second step uses the *modified Bhattacharya distance* [6] to discriminate between the M linguistic labels close to A, i.e.,

$$d_2(A, B_k) = \left[1 - \int_X \left(\frac{\mu_A(x)\mu_{B_k}(x)}{card(A) \cdot card(B_k)}\right)^{1/2} dx\right]^{1/2} \qquad k = 1, \ldots, M \tag{2}$$

The linguistic label corresponding to the smallest $d_2(A, B_k)$ is considered most similar to A.

2.2 Wenstøp's Linguistic Approximation Method

Wenstøp [8], who considered the same problem as Bonissone, states: "*a linguistic approximation routine is a function from the set of fuzzy subsets to a set of*

[2] We show w_j^2 in (1), because this is the way the equation is stated in [2].

linguistic values." Wenstøp used two parameters of a T1 FS, its *imprecision* (cardinality) and its *location* (centroid). The imprecision (p^1) was defined as the sum of membership values, whereas the location (p^2) was defined as the center of gravity. He then computed

$$d_W(A, B_i) = \left[(p_A^1 - p_{B_i}^1)^2 + (p_A^2 - p_{B_i}^2)^2\right]^{1/2} \qquad i = 1, \ldots, N \tag{3}$$

and chose B_i with the smallest $d_W(A, B_i)$ as the one most similar to A. Observe that Wenstøp's method is a simplified version of Bonissone's first step.

3 The VSM for T1 FSs

In this section a VSM for T1 FSs is proposed. Four desirable properties a similarity measure should possess are introduced first.

3.1 Four Desirable Properties of a Similarity Measure

The following four properties are proposed for a reasonable similarity measure for T1 FSs.

1) The similarity between two T1 FSs is 1 if and only if they are exactly the same.
2) If two T1 FSs intersect, there should be some similarity between them.
3) If two T1 FSs become more distant from each other, similarity between them should decrease.
4) The similarity between two T1 FSs should be a constant regardless of the order in which they are compared, i.e. $s(A, B) = s(B, A)$.

Next a VSM which possesses these properties is proposed.

3.2 The VSM for T1 FSs

When the similarity of two T1 FSs A and B are compared, it is necessary to compare their shapes as well as proximity; hence, a VSM, $\mathbf{s}_v(A, B)$, with two components is proposed,

$$\mathbf{s}_v(A, B) = (s_1(A, B), s_2(A, B))^T \,, \tag{4}$$

where $s_1(A, B) \in [0, 1]$ is a similarity measure on the shapes of A and B, and $s_2(A, B) \in [0, 1]$ is a similarity measure on the proximity of A and B. To define $\mathbf{s}_v(A, B)$, $s_1(A, B)$ and $s_2(A, B)$ must first be defined.

3.3 Definition of $s_1(A, B)$

Because the proximity of A and B is considered in $s_2(A, B)$, when computing $s_1(A, B)$ A and B are "aligned" so that their shapes can be compared. A reasonable alignment method is to move one or both of A and B so that their centroids, $c(A)$ and $c(B)$, coincide (see Fig. 1). The two T1 FSs can be moved to any location as long as $c(A)$ and $c(B)$ coincide; this will not affect the value of $s_1(A,B)$. In this paper B is moved to A and called B', as shown in Fig. 1.

Once the two T1 FSs are "aligned," $s_1(A,B)$ is computed by Jaccard's *unparameterized ratio model of similarity* [3] [5]:

$$s_1(A,B) = \frac{card(A \cap B')}{card(A \cup B')} = \frac{\int_X \min(\mu_A(x), \mu_{B'}(x))dx}{\int_X \max(\mu_A(x), \mu_{B'}(x))dx}. \tag{5}$$

Observe that $s_1(A,B)$ is a set-theoretic measure [4].

Theorem 1. *(a)* $0 \le s_1(A,B) \le 1$*; (b)* $s_1(A,B) = 1 \Leftrightarrow A = B'$*; and, (c)* $s_1(A,B) = s_1(B,A)$*.*

Proof: See Appendix A.1. ■

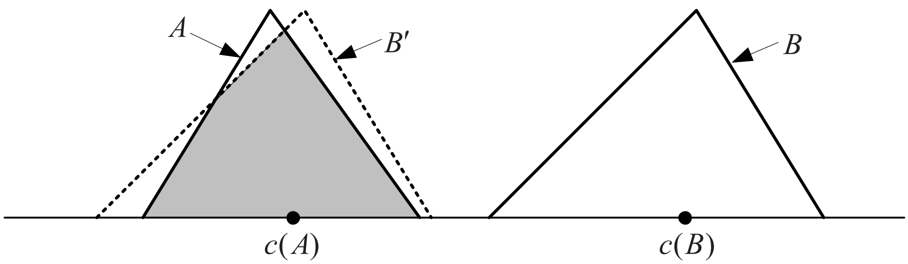

Fig. 1. An example of the VSM for T1 FSs. $c(A)$ and $c(B)$ are the centroids of A and B, respectively. B' is obtained by moving B so that $c(B)$ coincides with $c(A)$. Note that the shaded region can also be obtained by moving $c(A)$ to $c(B)$.

3.4 Definition of $s_2(A,B)$

$s_2(A,B)$ measures the proximity of A and B, and is defined as

$$s_2(A,B) \equiv h(d(A,B)) \tag{6}$$

where $d(A,B) = |c(A) - c(B)|$ is the Euclidean distance between the centers of the centroids of A and B (see Fig. 1), and h can be any function satisfying: (1) $\lim_{x\to\infty} h(x) = 0$; (2) $h(x) = 1$ if and only if $x = 0$; and, (3) $h(x)$ decreases monotonically as x increases.

Theorem 2. $s_2(A,B) \in [0,1]$*, and* $s_2(A,B) = 1$ *if and only if* $c(A) = c(B)$*.*

Proof: Theorem 2 is obvious from (6) and the above constraints on $h(x)$. ■

An example of $s_2(A,B)$ is

$$s_2(A,B) = e^{-rd(A,B)}, \tag{7}$$

[3] It is called *coefficient of similarity* by Sneath in [7]. The term *index of communality* has also been used [4].

where r is a positive constant. $s_2(A,B)$ is chosen as an exponential function because we believe the similarity between two FSs should decrease rapidly as the distance between them increases.

3.5 On Converting $\mathbf{s}_v(A,B)$ to a Scalar Similarity Measure $s_s(A,B)$

$\mathbf{s}_v(A,B)$ enables us to separately quantify the similarity of two features, shape and proximity. In linguistic approximation $\mathbf{s}_v(A,B_i)$ $(i=1,2,\ldots,N)$ need to be ranked to find the B_i most similar to A. This can be achieved by first converting the vector $\mathbf{s}_v(A,B_i)$ to a scalar similarity measure $s_s(A,B_i)$ and then ranking $s_s(A,B_i)$ $(i=1,2,\ldots,N)$.

In this paper, the scalar similarity between two T1 FSs A and B is computed as the product of their similarities in shape and proximity [4], i.e.

$$s_s(A,B) = s_1(A,B) \times s_2(A,B) \tag{8}$$

Properties of $s_s(A,B)$ include:

Theorem 3. *(a) $A = B \Leftrightarrow s_s(A,B) = 1$; (b) $s_s(A,B) > 0$; (c) $s_s(A,B) > s_s(A,C)$ if B and C have the same shape and C is further away from A than B is; and, (d) $s_s(A,B) = s_s(B,A)$.*

Proof: See Appendix A.2. ■

Theorem 3 shows that $s_s(A,B)$ satisfies the four properties stated in Section 3.1.

4 Comparisons

4.1 Comparison with Bonissone's Linguistic Approximation Distance Measure

Both $\mathbf{s}_v(A,B)$ and Bonissone's method consider the shapes and proximity of A and B. The main differences between them are:

(1) $\mathbf{s}_v(A,B)$ is a one-step method, whereas Bonissone's method is a two-step method.
(2) $\mathbf{s}_v(A,B)$ considers two features of A and B (shape and proximity). In Bonissone's first step, four features (centroid, cardinality, fuzziness and skewness) are considered, and in his second step, only one feature is considered (the modified Bhattacharya distance).
(3) $\mathbf{s}_v(A,B)$ measures the similarity between A and B, i.e. a larger $\mathbf{s}_v(A,B)$ means A and B are more similar. On the other hand, Bonissone's method measures the distance (or difference) between A and B, i.e. a larger $d_2(A,B)$ means A and B are less similar.

[4] Recently, Bonissone, et al. [3] defined a similarity measure as a weighted minimum of several sub-similarity measures. Although similar to our idea, their objective is quite different from our objective; hence, their similarity measure is not used in this paper.

4.2 Comparison with Wenstøp's Linguistic Approximation Method

Wenstøp's linguistic approximation method is quite similar to the VSM method in that both of them use the centroid and cardinality. The differences are:

(1) The VSM computes the similarity between two T1 FSs, whereas Wenstøp's method computes the difference between two T1 FSs.
(2) The VSM first aligns A and B and then computes the cardinalities of $A \cap B$ and $A \cup B$, whereas Wenstøp's method computes cardinalities of A and B directly.
(3) The VSM can be used for T1 FSs of any shapes, whereas, as shown in [8], the two parameters in Wenstøp's method are insufficient criteria for satisfactory linguistic approximation. As a further refinement, he includes other characteristics of FSs, e.g. non-normality, multi-modality, fuzziness and dilation [8].

4.3 Examples

For T1 FSs shown in Fig. 2, the results of Bonissone's linguistic approximation distance measure, Wenstøp's linguistic approximation measure and the VSM are shown in Table 1. The domain of x was discretized into 201 equally-spaced points in all three methods, and $r \equiv 4/|X|$ ($|X|$ is the length of the support of $A \cup B$) in the VSM [see (7)]. Note that all B_k ($k = 1, 2, 3, 4$) are assumed to survive Bonissone's first step, hence (2) was used to compute Bonissone's distance measure. Observe that all methods indicate B_2 is more similar to A than is B_1, which seems reasonable. When B_3 and B_4 are considered, Bonissone's measure indicates that they have the same similarity to A [5], and Wenstøp's measure indicates that B_4 is more similar to A than B_3 is. Both results seem counter-intuitive, because B_3 should be more similar to A than B_4 is, as indicated by the VSMs.

Table 1. Comparisons of similarity measures for T1 FSs A and B_k ($k = 1, \ldots, 4$) shown in Fig. 2

Measure	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$d_2(A, B_k)$	0.2472	0.1617	1	1
$d_W(A, B_k)$	28.5679	16.6650	38.6805	37.5736
$s_s(A, B_k)$	0.6368	0.7208	0.0086	0.0013

[5] If one FS must be chosen from B_k ($k = 1, 2, 3, 4$) so that it is most similar to A, then B_3 and B_4 may be removed during Bonissone's first step because they are too far away from A; however, if only B_3 and B_4 are available and one of them must be chosen so that it is more similar to A, Bonissone's method will have a problem because both B_3 and B_4 survive in the first step, and in the second step $d_2(A, B_3) = d_2(A, B_4)$.

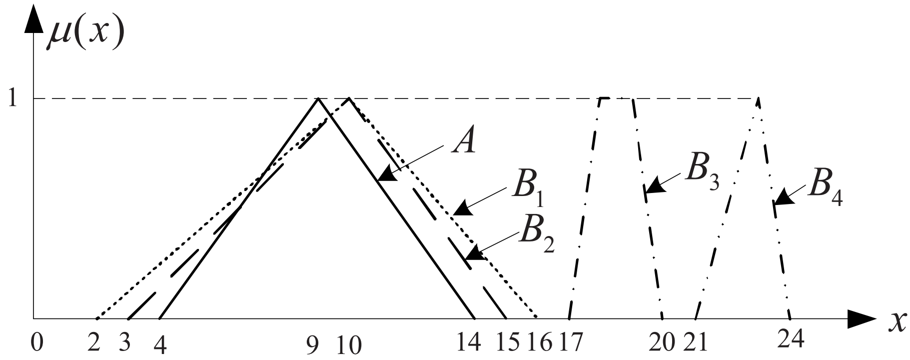

Fig. 2. T1 FSs used in the comparative study

5 Conclusions

A vector similarity measure for T1 FSs has been proposed in this paper. It is easy to understand, and its two components enable us to consider the similarity between shapes and proximity separately and explicitly. The VSM is simpler than two existing linguistic approximation methods, and yet a comparative study showed that it has better performance. Additionally, the VSM can be easily extended to interval T2 FSs [9].

References

1. Bonissone, P.P.: A pattern recognition approach to the problem of linguistic approximation. In: Proc. IEEE Int'l Conf. on Cybernetics and Society, Denver, CO (1979) 793–798
2. Bonissone, P.P.: A fuzzy sets based linguistic approach: Theory and applications. In: Proc. 12th Winter Simulation Conference, Orlando, FL (1980) 99–111
3. Bonissone, P.P., Varma, A., Aggour, K.S., Xue, F.: Design of local fuzzy models using evolutionary algorithms. *to be published in* Journal of Computational Statistics and Decision Analysis (2006)
4. Cross, V.V., Sudkamp, T.A.: Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications. Physica-Verlag, Heidelberg, NY (2002)
5. Jaccard, P.: Nouvelles recherches sur la distribution florale. Bulletin de la Societe de Vaud des Sciences Naturelles **44** (1908) 223
6. Kailath, T.: The divergence and bhattacharyaa distance measure in signal detection. IEEE Trans. on Communication Technology **15** (1967) 609–637
7. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy. W.H. Freeman and Company, San Francisco, CA (1973)
8. Wenstøp, F.: Quantitative analysis with linguistic values. Fuzzy Sets and Systems **4** (1980) 99–115

9. Wu, D., Mendel, J.M.: A vector similarity measure for interval type-2 fuzzy sets and type-1 fuzzy sets. *submitted to* Information Sciences (2006)
10. Zadeh, L.A.: Fuzzy sets. Information and Control **8** (1965) 338–353
11. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning-1. Information Sciences **8** (1975) 199–249

A Proof of Theorems

A.1 Proof of Theorem 1

Proof of (a). Because

$$0 \leq \min(\mu_A(x), \mu_{B'}(x)) \leq \max(\mu_A(x), \mu_{B'}(x)) \tag{9}$$

it follows that

$$0 \leq \int_X \min(\mu_A(x), \mu_{B'}(x))dx \leq \int_X \max(\mu_A(x), \mu_{B'}(x))dx \tag{10}$$

Consequently,

$$s_1(A, B) = \frac{\int_X \min(\mu_A(x), \mu_{B'}(x))dx}{\int_X \max(\mu_A(x), \mu_{B'}(x))dx} \in [0, 1]. \tag{11}$$

Proof of (b). $A = B'$ means $\mu_A(x) = \mu_{B'}(x)$. Substituting these two equations into (5),

$$s_1(A, B) = \frac{\int_X \mu_A(x)dx}{\int_X \mu_A(x)dx} = 1, \tag{12}$$

which proves the necessity of Theorem 1(b).

To prove the sufficiency of the result, observe that $s_1(A, B) = 1$ means

$$\int_X \min(\mu_A(x), \mu_{B'}(x))dx = \int_X \max(\mu_A(x), \mu_{B'}(x))dx \tag{13}$$

(13) holds if and only if

$$\mu_A(x) = \mu_{B'}(x) \qquad \forall x \in X. \tag{14}$$

(14) means $A = B'$.

Proof of (c). $s_1(A, B) = s_1(B, A)$ is obvious because the min and max operators in (5) do not concern the order of $\mu_A(x)$ and $\mu_{B'}(x)$, i.e. $\min(\mu_A(x), \mu_{B'}(x))$ $= \min(\mu_{B'}(x), \mu_A(x))$ and $\max(\mu_A(x), \mu_{B'}(x)) = \max(\mu_{B'}(x), \mu_A(x))$. ■

A.2 Proof of Theorem 3

Proof of (a). Sufficiency: $A = B$ means $s_1(A, B) = 1$ and $s_2(A, B) = 1$; hence, $s_s(A, B) = 1$.

Necessity: $s_s(A, B) = 1$ if and only if $s_1(A, B) = 1$ and $s_2(A, B) = 1$. $s_1(A, B) = 1$ means the shapes of A and B are the same, and $s_2(A, B) = 1$ means the distance between A and B is zero. Consequently, $A = B$.

Proof of (b). Observe that $s_1(A, B) > 0$ and $s_2(A, B) > 0$. Consequently, $s_s(A, B) > 0$.

Proof of (c). B and C have the same shape means

$$s_1(A, B) = s_1(A, C). \tag{15}$$

C is further away from A than B means

$$s_2(A, B) > s_2(A, C). \tag{16}$$

Hence,

$$s_1(A, B) \times s_2(A, B) > s_1(A, C) \times s_2(A, C), \tag{17}$$

i.e. $s_s(A, B) > s_s(A, C)$.

Proof of (d). Because neither $s_1(A, B)$ nor $s_2(A, B)$ concern the order of A and B, i.e. $s_1(A, B) = s_1(B, A)$ and $s_2(A, B) = s_2(B, A)$, it follows that $s_s(A, B) = s_s(B, A)$. ∎