Measuring Similarity Between Discontinuous Intervals - Challenges and Solutions

Shaily Kabir*, Christian Wagner*, Timothy C. Havens†, and Derek T. Anderson‡
*Intelligent Modelling and Analysis (IMA) Group and Lab for Uncertainty in Data and Decision Making (LUCID), School of Computer Science, University of Nottingham, Nottingham, UK
†ICC and Dept. Electrical and Computer Engineering, Michigan Technological University, Houghton, MI, USA
‡Dept. Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA
Email:{shaily.kabir, christian.wagner}@nottingham.ac.uk, thavens@mtu.edu, andersondt@missouri.edu

***Abstract*—Discontinuous intervals (DIs) arise in a wide range of contexts, from real world data capture of human opinion to α-cuts of non-convex fuzzy sets. Commonly, for assessing the similarity of DIs, the latter are converted into their continuous form, followed by the application of a continuous interval (CI) compatible similarity measure. While this conversion is efficient, it involves the loss of discontinuity information and thus limits the accuracy of similarity results. Further, most similarity measures including the most popular ones, such as Jaccard and Dice, suffer from *aliasing*, that is, they are liable to return the same similarity for very different pairs of CIs. To address both of these challenges, this paper proposes a generalized approach for calculating the similarity of DIs which leverages the recently introduced bidirectional subsethood based similarity measure (which avoids *aliasing*) while accounting for all pairs of the continuous subintervals within the DIs to be compared. We provide detail of the proposed approach and demonstrate its behaviour when applying bidirectional subsethood, Jaccard and Dice as similarity measures, using different pairs of synthetic DIs. The experimental results show that the similarity outputs of the new generalized approach follow intuition for all three similarity measures; however, it is only the proposed integration with the bidirectional subsethood similarity measure which also avoids *aliasing* for DIs.**

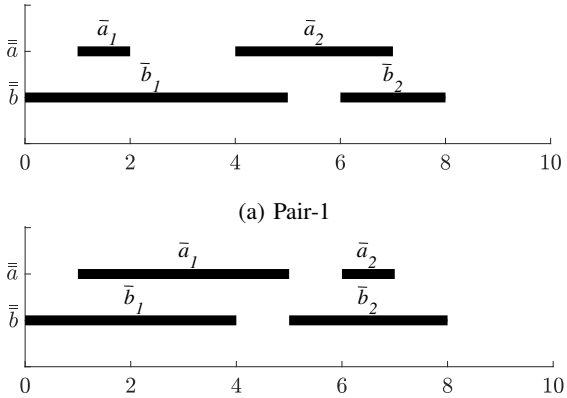
I. Introduction

Interval-valued data is used in many applications to model uncertain and imprecise data in a simple and efficient way. In particular, continuous intervals (CIs)—bounded by left and right endpoints [1]—are often used. Discontinuous intervals (DIs)—having a sequence of continuous subintervals [2]—can arise in many real-world situations, such as hazard detection [3], fusion of sensor data observed in a non-continuous space [4], temporal reasoning [5] [6], and expressing natural language with temporal repetition [7] where similarity between the DIs are often assessed and applied. Moreover, in fuzzy set (FS) theory, the α-cuts of non-convex FSs also result in the DIs [8]. In such cases, the similarity between non-convex FSs with the α-plane decomposition is dependent on the computation of similarity of DIs such as proposed in this paper.

Many similarity measures (SMs) have been proposed for CIs where Jaccard [9] and Dice [10] are the most popular ones. However, thus far, there is no specific SM for DIs that directly assesses their similarity. Instead, DIs are commonly converted into their continuous form (CIs) using some common approaches like interval addition [11], interval union [12], or a 'convexify' function [13] [14] and then the respective CI SM is applied to compute the similarity. However, the 'DI to CI' conversion involves the loss of discontinuity information of the DIs, changing the original meaning of the data, and thus affecting the accuracy of the similarity of the DIs. Figure 1 shows an example of this, where we consider two different pairs of DIs. The use of the 'convexify' function converts both cases into the same pair of CIs as shown in Fig. 2. As a result, we receive same similarity for both pairs of DIs by the Jaccard and Dice SMs, which goes against intuition in respect to the original DIs. Here, one way to avoid this type of information loss is to consider all possible combinations of the continuous subintervals within the DIs [15].

However, a further problem, particularly, the *aliasing* issue with common SMs, such as Jaccard and Dice has recently been identified [16], where the same similarity is returned for very different sets of intervals. A recently introduced SM for CIs using their overlapping ratios [16], also called bidirectional subsethood [17] has been shown to avoid *aliasing* for CIs.

In this paper, we propose a generalized SM for DIs which combines the bidirectional subsethood based SM [16] [17] with the idea of considering all pairs of continuous subintervals within the DIs. This generalized approach maintains discontinuity information in respect to DIs and uniformly



(a) Pair-1

(b) Pair-2

Fig. 1: Two different pairs of DIs.

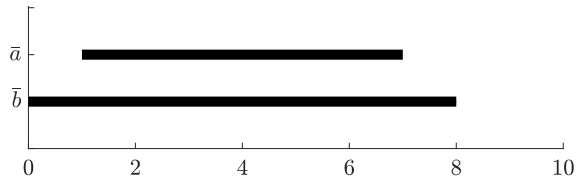


Fig. 2: Resulting CIs after applying the 'convexify' function [13] [14] to Pair-1 and Pair-2 of DIs in Fig. 1.

TABLE I: Acronyms and Notation

SM	Similarity Measure	CI	Continuous Interval
DI	Discontinuous Interval	FS	Fuzzy Set
S_J	Jaccard SM	S_D	Dice SM
S_h	Subsethood	X	Universe of discourse
S_{S_h}	Bidirectional subsethood based SM for CIs		
a	Crisp set		
$\overline{a}$	Continuous interval $\{\overline{a} \subseteq \mathbb{R} : \overline{a} = [a^-, a^+],\ a^- < a^+\}$		
$\overline{\overline{a}}$	Discontinuous interval $\{\overline{\overline{a}} \subseteq \mathbb{R} : \overline{\overline{a}} = \bigcup_{i=1}^{m} [a_i^-, a_i^+]\}$		
A	Fuzzy set $\{(x, \mu_A(x)) \,\vert x \in X$ and $\mu_a(x) \in [0,1]\}$		

handles the similarity computation of both CIs and DIs. We explore and contrast the behaviour of the resulting SM in respect to employing the well-known Jaccard and Dice SMs as part of the same framework, highlighting that such approaches, while avoiding information loss, still suffer from *aliasing*.

The rest of this paper is organized as follows. In Section II, we present some background facts of CIs and DIs, subsethood, two common SMs for the CIs along with the bidirectional subsethood based SM [16] [17]. Section III introduces the proposed generalized SM for the DIs and discusses its properties. We demonstrate this generalized SM using a set of synthetic examples of DIs and discuss the results in Section IV. Section V concludes the paper along with future work. Table I presents a list of acronyms and notation used in this paper.

II. Background

In this section, we first define CIs and DIs, followed by a review of subsethood, as well as the Jaccard and Dice SMs. Finally, we briefly review the bidirectional subsethood based SM for CIs [16], [17].

A. Continuous Intervals

A CI is a set of real numbers characterized by a left and a right endpoints [1]. Mathematically, it is represented as $\overline{a} = [a^-, a^+]$ with $a^- < a^+$ [11][1]. The cardinality, or equivalently, the size or width of a CI $\overline{a}$ is $|\overline{a}| = |a^+ - a^-|$ [18]. Three common approaches for representing multiple disjoint CIs with a single CI are:

- Interval Addition: If $\overline{a} = [a^-, a^+]$ and $\overline{b} = [b^-, b^+]$ are two bounded, non-empty CIs, then their addition, $\overline{a} + \overline{b} = [a^- + b^-, a^+ + b^+]$ is also a bounded, non-empty CI [11].
- Interval Union: If $\overline{a} \cap \overline{b} = \emptyset$, their union $\overline{a} \cup \overline{b} = [\min\{a^-, b^-\}, \max\{a^+, b^+\}]$ results in a single bounded, non-empty CI [12].
- A 'Convexify' function: It takes two CIs $\overline{a}$ and $\overline{b}$ as inputs and returns the smallest CI that covers both $\overline{a}$ and $\overline{b}$ [13].

B. Discontinuous Intervals

A DI consists of a number of continuous subintervals (i.e., CIs) [2][2]. Mathematically, it is represented as [4]

$$\overline{\overline{a}} = \bigcup_{i=1}^{m} \overline{a}_i = \bigcup_{i=1}^{m} [a_i^-, a_i^+],$$

where $\overline{\overline{a}}$ is the DI and m is the number of its CIs such that $\overline{a}_1 < ... < \overline{a}_i < ... < \overline{a}_m$, and $\overline{a}_i$ is the ith CI of $\overline{\overline{a}}$ such that $a_i^- < a_i^+$. Alternatively, $\overline{\overline{a}}$ can be presented as $\langle [a_1^-, a_1^+], ..., [a_i^-, a_i^+], ..., [a_m^-, a_m^+] \rangle$ [6].

C. Subsethood

Subsethood is a relation that expresses the degree to which one object is a subset of the other object. For two crisp sets, a and b, the subsethood is [19]

$$S_h(a, b) = \frac{|a \cap b|}{|a|}, \tag{1}$$

where $|a \cap b|$ is the cardinality of the intersection of a and b, and $|a|$ is the cardinality of a. S_h is in between 0 and 1 where $S_h(a, b) = 1$ when a is a proper subset of b ($a \subseteq b$), and $S_h(a, b) = 0$ when a is not a subset of b ($a \not\subset b$)—they are disjoint sets.[3]

Equivalently, the subsethood between two CIs $\overline{a}$ and $\overline{b}$ can be defined as

$$S_h(\overline{a}, \overline{b}) = \frac{|\overline{a} \cap \overline{b}|}{|\overline{a}|}, \tag{2}$$

where $|\overline{a} \cap \overline{b}|$ is the size of the intersection between $\overline{a}$ and $\overline{b}$ and $|\overline{a}|$ is the size of $\overline{a}$.

For the FSs[4] A and B, the degree of subsethood is [23]

$$S_h(A, B) = \frac{\sum_{i=1}^{n} \min(\mu_A(x_i), \mu_B(x_i))}{\sum_{i=1}^{n} \mu_A(x_i)}, \tag{3}$$

where $\sum_{i=1}^{n} \min(\mu_A(x_i), \mu_B(x_i))$ is a measure of the cardinality of the intersection of membership functions of A and B, and $\sum_{i=1}^{n} \mu_A(x_i)$ is a measure of the cardinality of A.

D. Jaccard Similarity Measure

The Jaccard SM [9] between sets a and b is defined as the ratio of the cardinality of their intersection and the cardinality of their union,

$$S_J(a, b) = \frac{|a \cap b|}{|a \cup b|}. \tag{4}$$

Beyond sets, the Jaccard SM is used to estimate the similarity for CIs or sets of CIs such as employed for example in data

[1]Note that $\overline{a}$ is also defined as a convex [4] or closed interval [16].

[2]Note that $\overline{\overline{a}}$ is also known as a non-convex interval [2].

[3]Note that subsethood is also known as set-inclusion [20] and overlapping ratio [16] as it captures the overlapping ratio between intervals.

[4]A fuzzy set (FS) is a set whose elements have membership within [0,1] [21]. A type-1 FS A on a universe of discourse X is $A = \{(x, \mu_A(x)) \,\vert x \in X\}$ where $\mu_A(x) \in [0, 1]$ is the membership grade of x in A [22].

fusion [24], [25] and that of fuzzy sets [26]. For comparing two CIs $\overline{a}$ and $\overline{b}$, the Jaccard SM is expressed as

$$S_J\left(\overline{a},\overline{b}\right)=\frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cup\overline{b}\right|}, \tag{5}$$

where $\left|\overline{a}\cap\overline{b}\right|$ is the size of the intersection between $\overline{a}$ and $\overline{b}$ and $\left|\overline{a}\cup\overline{b}\right|$ is the size of the interval segment(s) covering them. When $\overline{a}$ and $\overline{b}$ are completely overlapped, $S_J\left(\overline{a},\overline{b}\right)=1$ and when they are non-overlapped, $S_J\left(\overline{a},\overline{b}\right)=0$. Again, for the FSs A and B on the discrete universe of discourse X, the Jaccard similarity is extended as [27]

$$S_J\left(A,B\right)=\frac{\sum_{i=1}^{N}\min\left(\mu_A\left(x_i\right),\mu_B\left(x_i\right)\right)}{\sum_{i=1}^{N}\max\left(\mu_A\left(x_i\right),\mu_B\left(x_i\right)\right)}, \tag{6}$$

where $\mu_A\left(x_i\right)$ and $\mu_B\left(x_i\right)$ are the membership grades of x_i in A and B respectively. Equation (6) gives 1 for identical FSs and 0 for disjoint FSs. Note that the Jaccard SM has been further extended for interval-valued [28] and type-2 fuzzy sets [29]; though, this is not discussed further here.

E. Dice Similarity Measure

The Dice SM [10] between sets a and b is the ratio of the cardinality of their intersection and the average of their cardinality, expressed as

$$S_D\left(a,b\right)=\frac{|a\cap b|}{\frac{1}{2}(|a|+|b|)}. \tag{7}$$

In [24], [25], the Dice similarity is used along with the Jaccard similarity for the CIs. As for sets, the Dice similarity for CIs $\overline{a}$ and $\overline{b}$ is

$$S_D\left(\overline{a},\overline{b}\right)=\frac{\left|\overline{a}\cap\overline{b}\right|}{\frac{1}{2}\left(\left|\overline{a}\right|+\left|\overline{b}\right|\right)}. \tag{8}$$

While less frequently used for FSs than Jaccard, the Dice SM is for example used in [30], [31] for trapezoidal FSs in the context of solving multi-criteria decision-making problems.

F. Bidirectional Subsethood Based Similarity Measure for Continuous Intervals

A new SM for the CIs was introduced in [16] which uses the reciprocal subsethoods [17] or overlapping ratios [16] of a pair of CIs for capturing their similarity. This measure for two CIs $\overline{a}$ and $\overline{b}$ [16] [17] is

$$S_{S_h}\left(\overline{a},\overline{b}\right)=\bigstar\left(S_h(\overline{a},\overline{b}),S_h(\overline{b},\overline{a})\right), \tag{9}$$

where $\bigstar$ is a *t-norm*[5]. We can rewrite (9) using (2) as,

$$S_{S_h}\left(\overline{a},\overline{b}\right)=\bigstar\left(\frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\right|},\frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{b}\right|}\right). \tag{10}$$

This SM directly captures any changes in the size of CIs and is sensitive to the size of their intersection when one CI is a subset of another in a pair. Further, it is always within [0,1], and is bounded above and below by the Jaccard and Dice SMs respectively for the minimum *t-norm*.

In the next section, we introduce a generalized measure where we can apply any of the S_J, S_D or S_{S_h} SMs for estimating the similarity between CIs or DIs by meeting their continuity or discontinuity property.

III. Proposed Generalized Similarity Measure for Discontinuous Intervals

In this section, we propose a generalized SM for computing the similarity between two DIs by comparing all possible pairs of their continuous subintervals. As stated, this generalized SM is equally applicable for the CIs. First, we present the proposed generalization and then demonstrate its major properties. We note that while the proposed approach is computationally expensive, we focus on the quality of the resulting similarity assessment only in this paper. We have already made progress on making the approach computationally more efficient, but considering the constraints on manuscript size, we will focus on this in our future publication.

A. Proposed Generalized Similarity Measure

In the proposed generalization, we use the basic notion that as a DI contains one or more continuous subintervals, comparing two DIs is analogous to systematically comparing their subintervals. With this intent, we first determine all possible pairs of subintervals within the DIs and compute their similarity. Later, we aggregate all these similarities to determine overall similarity between the DIs. Equation (11) presents the proposed generalized SM for DIs $\overline{\overline{a}}=\langle\overline{a}_1,...,\overline{a}_m\rangle$ and $\overline{\overline{b}}=\langle\overline{b}_1,...,\overline{b}_n\rangle$,

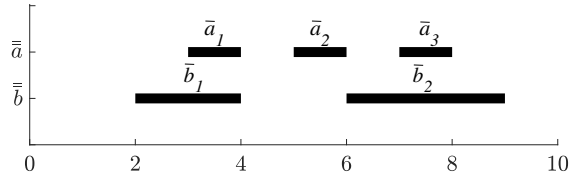
$$S\left(\overline{\overline{a}},\overline{\overline{b}}\right)=\frac{1}{\max(m,n)}\sum_{i=1}^{m}\sum_{j=1}^{n}S\left(\overline{a}_i,\overline{b}_j\right), \tag{11}$$

where $S\left(\overline{a}_i,\overline{b}_j\right)$ computes the similarity between subintervals $\overline{a}_i\in\overline{\overline{a}}$ and $\overline{b}_j\in\overline{\overline{b}}$ of each pair $\{\overline{a}_i,\overline{b}_j\}$ using any of the three SMs (S_J, S_D, and S_{S_h}). $\max(m,n)$ is the maximum number of pairs that can arise from the comparison of a single subinterval. The $\max(m,n)$ operator in the normalization step guarantees a maximum similarity of 1 – achieved when two DIs are identical. While other operators could potentially be explored, the $\max(m,n)$ operator provides intuitive behaviour for the similarity measure.

Remark 1. When DIs possess only a single CI, the formulation for S at (11) will return the original formulation for CIs.

Example 1. We consider an example in Fig. 3 for determining the similarity for a pair of DIs, $\overline{\overline{a}}$ and $\overline{\overline{b}}$. Here, $\overline{\overline{a}}$ contains three subintervals $\langle\overline{a}_1=[3,4],\overline{a}_2=[5,5.9],\overline{a}_3=[7,8]\rangle$, and $\overline{\overline{b}}$ has two subintervals $\langle\overline{b}_1=[2,4],\overline{b}_2=[6,9]\rangle$. Therefore, $m=3$ and $n=2$. There are six pairs of subintervals—$\{\overline{a}_1,\overline{b}_1\}$, $\{\overline{a}_1,\overline{b}_2\}$, $\{\overline{a}_2,\overline{b}_1\}$, $\{\overline{a}_2,\overline{b}_2\}$, $\{\overline{a}_3,\overline{b}_1\}$, and $\{\overline{a}_3,\overline{b}_2\}$ where the first and last pairs are overlapped, and the rest are

[5]A triangular norm (*t-norm*) is an associative, symmetric, and increasing function $\bigstar:[0,1]^2\rightarrow[0,1]$ such that $\bigstar(1,x)=x$ for all $x\in[0,1]$ [32].



SM	Similarity between subintervals $\{\bar{a}_1, \bar{b}_1\}$	$\{\bar{a}_1, \bar{b}_2\}$	$\{\bar{a}_2, \bar{b}_1\}$	$\{\bar{a}_2, \bar{b}_2\}$	$\{\bar{a}_3, \bar{b}_1\}$	$\{\bar{a}_3, \bar{b}_2\}$
S_J	0.50	0.0	0.0	0.0	0.0	0.3333
S_D	0.6667	0.0	0.0	0.0	0.0	0.50
S_{S_h}	0.50	0.0	0.0	0.0	0.0	0.3333

Fig. 3: Using S_J, S_D, and S_{S_h} SMs, the similarity results of all combinations of subintervals within a pair of DIs—one with three and the other with two subintervals.

disjoint. Figure 3 shows the similarity for all pairs using S_J, S_D and S_{S_h} (with the minimum *t-norm*) SMs. Hence, the overall similarity between $\bar{\bar{a}}$ and $\bar{\bar{b}}$ using (11) along with S_{S_h} SM is, $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) = \frac{1}{\max(3,2)} \times 0.8333 = \frac{1}{3} \times 0.8333 = 0.2778$. In a similar manner, the total similarity between $\bar{\bar{a}}$ and $\bar{\bar{b}}$ with S_J and S_D SMs are 0.2778 and 0.3889, respectively.

Theorem 1. *The proposed generalized approach with S_J, S_D and S_{S_h} SMs satisfies all common properties of a SM for the DIs $\bar{\bar{a}}$, $\bar{\bar{b}}$, and $\bar{\bar{c}}$ such that:*

(a) $0 \leq S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) \leq 1$ *(boundedness);*

(b) $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) = S\left(\bar{\bar{b}}, \bar{\bar{a}}\right)$ *(symmetry);*

(c) $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) = 1 \iff \bar{\bar{a}} = \bar{\bar{b}}$ *(reflexivity);*

(d) $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) = 0 \iff \bar{\bar{a}}$ *and* $\bar{\bar{b}}$ *are disjoint (disjointness);*

(e) $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) \geq S\left(\bar{\bar{a}}, \bar{\bar{c}}\right)$ *when* $\bar{\bar{a}} \subseteq \bar{\bar{b}} \subseteq \bar{\bar{c}}$. *(transitivity).*

Proof: Consider $\bar{\bar{a}} = \langle \bar{a}_1, ..., \bar{a}_m \rangle$, $\bar{\bar{b}} = \langle \bar{b}_1, ..., \bar{b}_n \rangle$, and $\bar{\bar{c}} = \langle \bar{c}_1, ..., \bar{c}_p \rangle$.

(a) $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right)$ involves the S_J, S_D or S_{S_h} SMs to compute similarity for all pairs of the subintervals $\bar{a}_i \in \bar{\bar{a}}$ and $\bar{b}_j \in \bar{\bar{b}}$. All S_J, S_D and S_{S_h} SMs are bounded by 0 and 1 [16], i.e., $0 \leq S\left(\bar{a}_i, \bar{b}_j\right) \leq 1$, $\forall \bar{a}_i, \bar{b}_j$. Hence, the mean of all such similarities is again within 0 and 1, implying $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) \in [0, 1]$.

(b) All of S_J, S_D and S_{S_h} measures are symmetric [16], thus making the S measure symmetric too.

(c) If $\bar{\bar{a}} = \bar{\bar{b}}$, it means that both $\bar{\bar{a}}$ and $\bar{\bar{b}}$ have an equal number of m subintervals and each $\bar{a}_i$ is identical to each $\bar{b}_i$, i.e., $\bar{a}_i = \bar{b}_i$, $1 \leq i \leq m$. Among all subinterval pairs, m pairs have identical subintervals and the rest have disjoint subintervals. It implies that m pairs receive a similarity of 1 and the rest have a similarity of 0. Hence, the similarity between $\bar{\bar{a}}$ and $\bar{\bar{b}}$ is, $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) = \frac{1}{\max(m,m)} \times m = \frac{m}{m} = 1$. Thus, $S(\bar{\bar{a}}, \bar{\bar{b}}) = 1$ means that $\bar{\bar{a}}$ and $\bar{\bar{b}}$ are identical DIs.

(d) If $\bar{\bar{a}}$ and $\bar{\bar{b}}$ are disjoint, it means that no subinterval of $\bar{\bar{a}}$ is overlapping with any of subintervals of $\bar{\bar{b}}$, i.e., $\left|\bar{a}_i \cap \bar{b}_j\right| = 0$, $1 \leq i \leq m$ and $1 \leq j \leq n$. In other words, all subinterval pairs consist of disjoint subintervals, thus receiving a similarity of 0. Hence, the similarity between $\bar{\bar{a}}$ and $\bar{\bar{b}}$ is, $S\left(\bar{\bar{a}}, \bar{\bar{b}}\right) = \frac{1}{\max(m,n)} \times 0 = 0$.

(e) All of S_J, S_D and S_{S_h} measures are transitive [16], which implies that the S measure is also transitive. ■

IV. Demonstration

This section presents the behaviour of the proposed generalized approach based on the bidirectional subsethood based SM (S_{S_h}) along with the Jaccard (S_J) and Dice (S_D) SMs for the DIs. Herein, we conduct two separate sets of experiments with different synthetic examples, each designed to facilitate intuitive understanding of the behaviour of the approaches.

With the first synthetic dataset, we gradually decrease overlapping between the subintervals for a pair of DIs to see how smoothly the similarity alters from 1 to 0. In particular, the change in similarity results is investigated for a gradual change in the overlapping of subintervals. With the second synthetic dataset, we change the number of subintervals and their degree of overlapping. In particular, we expect to see changes in the similarity due to a rise in the number of subintervals and their potential (lack of) overlap. In all experiments, we use the minimum *t-norm* for the S_{S_h} SM as it is the most common in practice. Further, all of these experiments are implemented using Java on an Intel(R) Core(TM) i3-4005U series based machine running at 1.70 GHz with 8GB RAM.

A. Synthetic Dataset-1

We consider a number of scenarios for a pair of DIs ($\bar{\bar{a}}$ and $\bar{\bar{b}}$) where each of them includes two subintervals. In each scenario, we vary the degree of overlap between subintervals of $\bar{\bar{a}}$ and $\bar{\bar{b}}$ to explore how the generalized approach (S) responds with the respective SMs S_J, S_D, and S_{S_h} and how smoothly the similarity results change from 1 to 0. We keep $\bar{\bar{a}}$ unchanged in all scenarios but shift the subintervals of $\bar{\bar{b}}$ consecutively by a factor of 25%. In Fig. 4(b)-(e), we gradually shift the rightmost subinterval $[6, 8]$ of $\bar{\bar{b}}$ by a factor of 25% till its only intersection is the right-end point of the subinterval $[6, 8]$ of $\bar{\bar{a}}$. In Fig. 4(f)-(i), we further shift the leftmost subinterval $[1, 3]$ of $\bar{\bar{b}}$ by a factor of 25% until its only intersection is the right-end point of the subinterval $[1, 3]$ of $\bar{\bar{a}}$. Figure (5)(a) presents in detail the shifting of subintervals of $\bar{\bar{b}}$ for all scenarios, and Fig. (5)(b) graphically exhibits the similarity results using all three SMs.

The results in Fig. 5(b) show that the initial similarity between $\bar{\bar{a}}$ and $\bar{\bar{b}}$ is 1 from the S measure with all three SMs (as they are identical in Fig. 4(a)). Their similarity gradually decreases to 0.50 when the second subinterval $[6, 8]$ of $\bar{\bar{b}}$ is gradually shifted by the factor of 25%. The similarity drops further and gradually reaches to 0 when the first subinterval $[1, 3]$ of $\bar{\bar{b}}$ is also repeatedly shifted. Although one would intuitively expect that the similarity between the DIs should decrease proportionately as to the rate of change in their overlapping, Fig. 5(b) shows a proportionate decline in similarity results by both S_{S_h} and S_D SMs, while the S_J SM exhibits

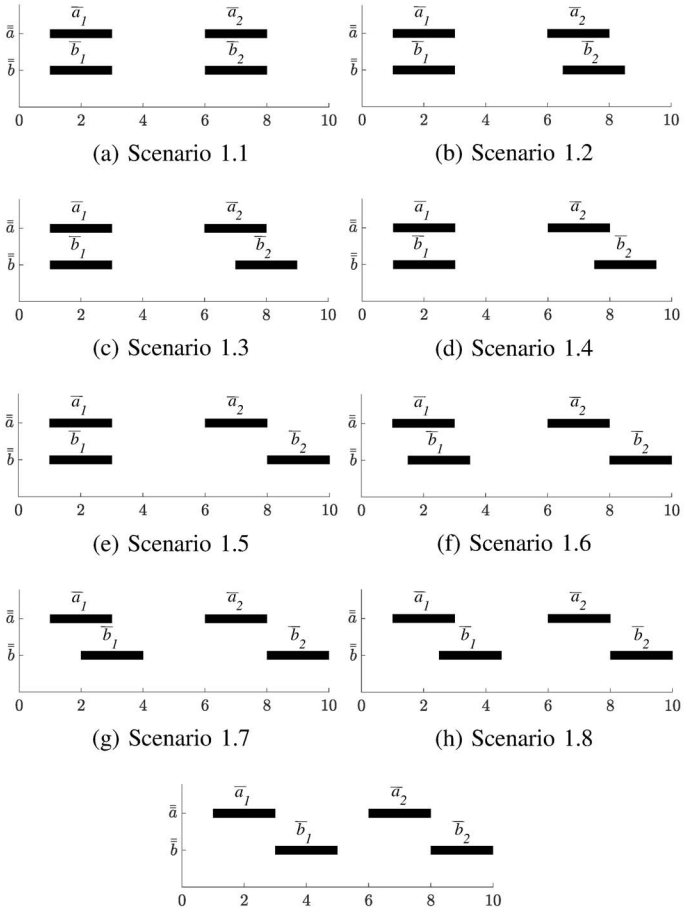
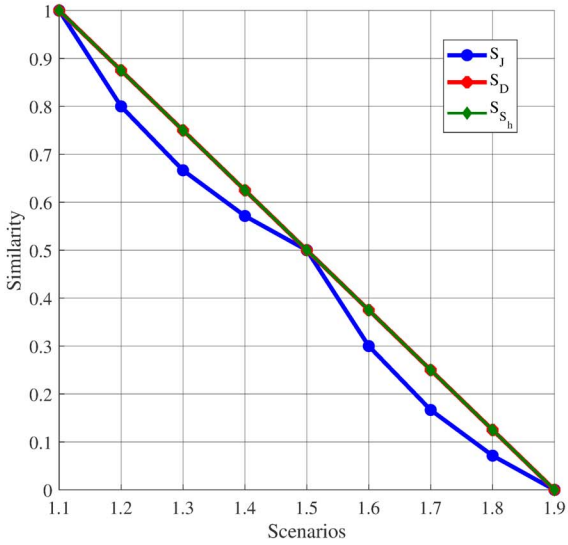


Fig. 4: Varying the degree of overlap between subintervals of $\overline{\overline{a}}$ and $\overline{\overline{b}}$ to examine how smoothly similarity results alter.

DI		Scenarios								
$\overline{\overline{b}} = \{\overline{b}_1, \overline{b}_2\}$		1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
Shift $\overline{b}_1$		0%	0%	0%	0%	0%	25%	50%	75%	100%
Shift $\overline{b}_2$		0%	25%	50%	75%	100%	100%	100%	100%	100%
SM Results	S_J	1.0	0.80	0.6667	0.5714	0.50	0.30	0.1667	0.0714	0.0
	S_D	1.0	0.875	0.75	0.625	0.50	0.375	0.25	0.125	0.0
	S_{S_h}	1.0	0.875	0.75	0.625	0.50	0.375	0.25	0.125	0.0

(a) % of Shifting of subintervals $\overline{b}_1$ and $\overline{b}_2$ along with the similarity results from S_J, S_D, and S_{S_h} SMs (Note: $\overline{\overline{a}}$ (DI) remains unchanged in all scenarios)



(b) Similarity results from the S with S_J, S_D, and S_{S_h} SMs

Fig. 5: Similarity results showing proportionate decline by both S_{S_h} and S_D SMs while steeper decline by S_J SM for different scenarios in Fig. 4.

steeper decline. It is noted that in this experiment, though both SMs S_D and S_{S_h} have shown the same expected results, the next set of synthetic examples exhibit their different behaviour.

B. Synthetic Dataset-2

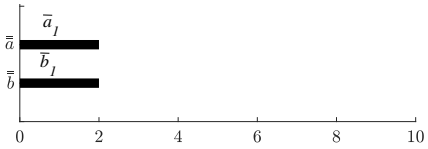
We consider six scenarios of a pair of DIs ($\overline{\overline{a}}$ and $\overline{\overline{b}}$), where the number of subintervals within the DIs and their degree of overlap are varied.

Scenario 2.1 – In Fig. 6(a), both $\overline{\overline{a}}$ and $\overline{\overline{b}}$ have an identical subinterval, [0,2]. The S with all three SMs returns the same similarity of 1–as expected.

Scenario 2.2 – In Fig. 6(b), a new subinterval $[3, 5]$ is added to $\overline{\overline{a}}$; therefore it now turns into $\langle[0, 2], [3, 5]\rangle$ whereas $\overline{\overline{b}}$ remains unchanged. Since the new subinterval $[3, 5]$ of $\overline{\overline{a}}$ is disjoint to $\overline{\overline{b}}$, adding it is likely to reduce overall similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ as compared to the *Scenario 2.1* (Fig. 6(a)). The S with all three SMs performs as to this expectation.

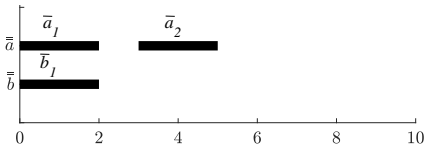
Scenario 2.3 – In Fig. 6(c), $\overline{\overline{a}}$ is kept the same as the *Scenario 2.2* (Fig. 6(b)), but we add one more subinterval $[3, 9]$ to $\overline{\overline{b}}$, which now becomes $\langle[0, 2], [3, 9]\rangle$. From one end, subinterval $[3, 9]$ of $\overline{\overline{b}}$ has about 33% overlap with subinterval $[3, 5]$ of $\overline{\overline{a}}$. On the other side, the subinterval $[3, 5]$ of $\overline{\overline{a}}$ exhibits complete 100% overlap with $[3, 9]$ of $\overline{\overline{b}}$. Hence, the overall similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ is expected to increase as compared to the *Scenario 2.2* (Fig. 6(b)). This expected increment in the similarity result is seen for the S with all three SMs.

Scenario 2.4 – In Fig. 6(d), for $\overline{\overline{a}}$, we extend one edge of subinterval $[3, 5]$ to $[3, 7]$, while its other subinterval remains the same as the *Scenario 2.3* (Fig. 6(c)). Hence, $\overline{\overline{a}}$ now becomes $\langle[0, 2], [3, 7]\rangle$. On the other hand, for $\overline{\overline{b}}$, we reduce the size of its one subinterval $[3, 9]$ and set it as $[5, 9]$. Therefore, $\overline{\overline{b}}$ is now $\langle[0, 2], [5, 9]\rangle$. Here, the subinterval $[3, 7]$ of $\overline{\overline{a}}$ and the subinterval $[5, 9]$ of $\overline{\overline{b}}$ have a 50% overlap. In this case, we expect that the overall similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ should be different from *Scenario 2.3* (Fig. 6(c)). The results show that the S with S_{S_h} SM can capture the change in the degree of overlap and provide a different similarity as expected whereas the S with both S_J and S_D SMs provides the same similarity of 0.6667 and 0.75 respectively (Fig. 6(d)) as found in the *Scenario 2.3* (Fig. 6(c)). This suggests that the *aliasing* issue persists for both S_J and S_D SMs, but is addressed when employing the S_{S_h} SM as detailed in [17].



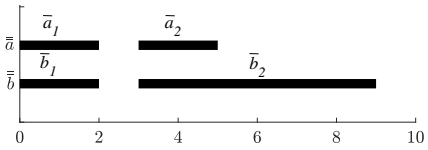
SM	Similarity
S_J	1.0
S_D	1.0
S_{S_h}	1.0

(a) Scenario 2.1



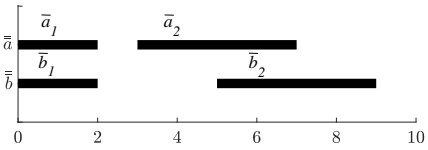
SM	Similarity
S_J	0.50
S_D	0.50
S_{S_h}	0.50

(b) Scenario 2.2



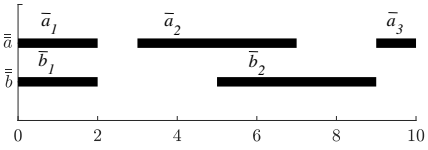
SM	Similarity
S_J	0.6667
S_D	0.75
S_{S_h}	0.6667

(c) Scenario 2.3



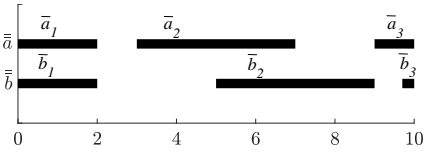
SM	Similarity
S_J	0.6667
S_D	0.75
S_{S_h}	0.75

(d) Scenario 2.4



SM	Similarity
S_J	0.4444
S_D	0.50
S_{S_h}	0.50

(e) Scenario 2.5



SM	Similarity
S_J	0.5445
S_D	0.6538
S_{S_h}	0.60

(f) Scenario 2.6

Fig. 6: Similarity results for the pairs of DIs with an increasing number of subintervals and varying degree of overlap. Note: SMs S_J and S_D return identical results for scenarios 2.3 and 2.4, i.e., they are subject to *aliasing* - only S_{S_h} captures the change in the respective DIs and thus avoids *aliasing*.

Scenario 2.5 – In Fig. 6(e), we add one more subinterval $[9, 10]$ to $\overline{\overline{a}}$ as designed in *Scenario 2.4* (Fig. 6(d)), thus setting $\overline{\overline{a}}$ as $\langle[0, 2], [3, 7], [9, 10]\rangle$, while $\overline{\overline{b}}$ remains the same. As this new subinterval $[9, 10]$ of $\overline{\overline{a}}$ is disjoint from all subintervals of $\overline{\overline{b}}$, adding it should decrease the similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ as compared to the *Scenario 2.4* (Fig. 6(d)). The results show that the S with all three SMs yielded expected similarity.

Scenario 2.6 – In Fig. 6(f), $\overline{\overline{a}}$ remains the same but $\overline{\overline{b}}$ is changed by adding one more subinterval $[9.7, 10]$. Thus, $\overline{\overline{b}}$ is now $\langle[0, 2], [5, 9], [9.7, 10]\rangle$. The new subinterval of $\overline{\overline{b}}$ has 30% overlap with the subinterval $[9, 10]$ of $\overline{\overline{a}}$. Therefore, the overall similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ is expected to be higher than that of the *Scenario 2.5* (Fig. 6(e)). Again, we receive higher similarity results from the S with all three SMs.

In summary, the S with S_{S_h} and S_D SMs follow a proportionate decline in similarity results as we gradually move the subintervals of a pair of DIs from a complete overlap to disjoint positions. Contrarily, the S with S_J SM yields slightly higher than proportionate decline in the similarity results. Importantly, the S with S_J and S_D SMs still exhibit *aliasing*, whereas the S with S_{S_h} SM is sensitive to changes in overlap and thus avoid it.

V. Conclusion

In this paper, we have proposed a generalized approach to computing the similarity of DIs by integrating the bidirectional subsethood based SM [16] [17] with the strategy of considering the similarity of all continuous subinterval-combinations within the DIs. The new generalized SM is equally suitable for CIs and DIs. It does not require conversion/approximation of DIs to CIs, thus avoiding changes to the original data. We have compared the performance of the generalized approach using the bidirectional subsethood SM along with the Jaccard and Dice SMs for different synthetic pairs of DIs. The results show intuitive behaviour of the resulting generalized approach while highlighting that only by using the recently developed bidirectional subsethood similarity as part of the generalized approach, can avoid the *aliasing* issue.

In our generalized approach, we always consider all possible pairs of subintervals. As a result, an increase in the number of subintervals within the DIs leads to the increase in the number of similarity calculations. In particular where DIs have many/all disjoint subinterval pairs, such a 'brute force' approach results in substantial execution time. To mitigate this, in the future, we will integrate this generalized SM with Allen's theory [5] for reducing the number of similarity calculations and overall execution time. Further, we plan to use it for assessing similarity of non-convex FSs. We also aim to apply it in generating data-driven fuzzy measures from DI-valued data [33] and use it with fuzzy integrals for aggregation.

Acknowledgment

This work was supported by the UK EPSRCs Leveraging the Multi-Stakeholder Nature of Cyber Security grant, EP/P011918/1.

References

[1] B. Leban, D. McDonald, and D. Forster, "A representation for collections of temporal intervals." in *Proc. 5th National Conf. Artificial Intelligence*, 1986, pp. 367–371.

[2] P. Ladkin, "Time representation : a taxonomy of interval relations," in *Proc. 6th National Conf. Artificial Intelligence*, Philadelphia, USA, 1986, pp. 360–366.

[3] C. Wagner, D. T. Anderson, and T. C., "Generalization of the fuzzy integral for discontinuous interval-and non-convex interval fuzzy set-valued inputs," in *Proc. IEEE Int. Conf. Fuzzy Systems*, Hyderabad, India, 2013, pp. 1–8.

[4] G. Beliakov and S. James, "A penalty-based aggregation operator for non-convex intervals," *Knowledge-based Systems*, vol. 70, no. 1, pp. 335–344, 2014.

[5] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, 1983.

[6] R. A. Morris, W. D. Shoaff, and L. Khatib, "Path consistency in a network of non-convex intervals," in *Proc. 13 Int. Conf. Artificial Intelligence*, Chambery, France, 1993, pp. 655–661.

[7] D. Cukierman and J. Delgrande, "Characterizing temporal repetition," in *Proc. 3rd Int. Workshop Temporal Representation and Reasoning*, Florida, USA, 1996, pp. 80–87.

[8] J. McCulloch, C. Wagner, and U. Aickelin, "Measuring the directional distance between fuzzy sets," in *13th UK Workshop on Computational Intelligence*, 2013, pp. 38–45.

[9] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bulletin de la Socit vaudoise des Sciences Naturelles*, vol. 44, pp. 223–270, 1908.

[10] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[11] T. Hickey, Q. Ju, and M. H. Van Emden, "Interval arithmetic: From principles to implementation," *J. of the ACM*, vol. 48, no. 5, pp. 1038–1068, 2001.

[12] H. Schichl, F. Domes, T. Montanher, and K. Kofler, "Interval unions," *BIT Numerical Mathematics*, vol. 57, pp. 531–556, 2017.

[13] P. B. Ladkin, "The completeness of a natural system for reasoning with time intervals." in *IJCAI*, 1987, pp. 462–465.

[14] R. Bleisinger and B. Kröll, "Representation of non-convex time intervals and propagation of non-convex relations," 1994.

[15] C. Wagner, D. T. Anderson, and T. C. Havens, "Generalization of the fuzzy integral for discontinuous interval-and non-convex interval fuzzy set-valued inputs," in *Proc. IEEE Int. Conf. Fuzzy Systems*, Hyderabad, India, 2013, pp. 1–8.

[16] S. Kabir, C. Wagner, T. C. Havens, D. T. Anderson, and U. Aickelin, "Novel similarity measure for interval-valued data based on overlapping ratio," in *Proc. IEEE Int. Conf. Fuzzy Systems*, Naples, Italy, 2017, pp. 1–6.

[17] S. Kabir, C. Wagner, T. C. Havens, and D. T. Anderson, "A bidirectional subsethood based similarity measure for fuzzy sets," in *IEEE World Congress on Computational Intelligence*, Rio de Janeiro, Brazil, 2018, pp. 1–6.

[18] I. Araya, G. Trombettoni, and B. Neveu, "A contractor based on convex interval taylor," in *Int. Conf. Integration of Artificial Intelligence and Operations Research Techniques in Constraint Programming*, 2012, pp. 1–16.

[19] H. T. Nguyen and K. Vladik, "Computing degrees of subsethood and similarity for interval-valued fuzzy sets: fast algorithms," in *Proc. 9th Int. Conf. Intelligent Technologies*, Thailand, 2008, pp. 47–55.

[20] Y. Li, K. Qin, and X. He, "Inclusion and subsethood measure for interval-valued fuzzy sets and for continuous type-2 fuzzy sets," *Int. J. of Computational Intelligence Systems*, vol. 6, no. 3, pp. 411–422, 2013.

[21] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[22] J. M. Mendel, *Uncertain rule-based fuzzy logic systems: introduction and new directions*. Prentice Hall PTR Upper Saddle River, 2001.

[23] B. Kosko, "Fuzzy entropy and conditioning," *Information Sciences*, vol. 40, no. 2, pp. 165–174, 1986.

[24] T. C. Havens, D. T. Anderson, C. Wagner, H. Deilamsalehy, and D. Wonnacott, "Fuzzy integrals of crowd-sourced intervals using a measure of generalized accord," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2013, pp. 1–8.

[25] T. C. Havens, D. T. Anderson, and C. Wagner, "Data-informed fuzzy measures for fuzzy integration of intervals and fuzzy numbers," *IEEE Trans. Fuzzy Systems*, vol. 23, no. 5, pp. 1861–1875, 2015.

[26] D. Dubois and H. Prade, *Fuzzy sets and systems: theory and applications*. Academic press, Newyork, 1980.

[27] C. P. Pappis and N. I. Karacapilidis, "A comparative assessment of measures of similarity of fuzzy values," *Fuzzy Sets and Systems*, vol. 56, no. 2, pp. 171–174, 1993.

[28] H. T. Nguyen and V. Kreinovich, "Computing degrees of subsethood and similarity for interval-valued fuzzy sets: fast algorithms," in *Proc. 9th Int. Conf. Intelligent Technologies*, Samui, Thailand, 2008, pp. 47–55.

[29] J. McCulloch, C. Wagner, and U. Aickelin, "Extending similarity measures of interval type-2 fuzzy sets to general type-2 fuzzy sets," in *Proc. IEEE Int. Conf. Fuzzy Systems*, Hyderabad, India, 2013, pp. 1–8.

[30] J. Ye, "Multicriteria decision-making method using the dice similarity measure between expected intervals of trapezoidal fuzzy numbers," *J. Decision Systems*, vol. 21, no. 4, pp. 307–317, 2012.

[31] ——, "The dice similarity measure between generalized trapezoidal fuzzy numbers based on the expected interval and its multicriteria group decision-making method," *J. Chinese Institute of Industrial Engineers*, vol. 29, no. 6, pp. 375–382, 2012.

[32] B. Bedregal, H. Bustince, E. Palmeira, G. Dimuro, and J. Fernandez, "Generalized interval-valued owa operators with interval weights derived from interval-valued overlap functions," *Int. J. of Approximate Reasoning*, vol. 90, pp. 1–16, 2017.

[33] D. T. Anderson, P. Elmore, F. Petry, and T. C. Havens, "Fuzzy choquet integration of homogeneous possibility and probability distributions," *Information Sciences*, vol. 363, pp. 24–39, 2016.