Novel Similarity Measure for Interval-Valued Data Based on Overlapping Ratio

Shaily Kabir*, Christian Wagner*†, Timothy C. Havens†, Derek T. Anderson‡, and Uwe Aickelin*

*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 and Computer Engineering, Mississippi State University, Mississippi State, USA

Email:{shaily.kabir, christian.wagner, uwe.aickelin}@nottingham.ac.uk, anderson@ece.msstate.edu, thavens@mtu.edu

***Abstract*—In computing the similarity of intervals, current similarity measures such as the commonly used Jaccard and Dice measures are at times not sensitive to changes in the width of intervals, producing equal similarities for substantially different pairs of intervals. To address this, we propose a new similarity measure that uses a bi-directional approach to determine interval similarity. For each direction, the overlapping ratio of the given interval in a pair with the other interval is used as a measure of uni-directional similarity. We show that the proposed measure satisfies all common properties of a similarity measure, while also being invariant in respect to multiplication of the interval endpoints and exhibiting linear growth in respect to linearly increasing overlap. Further, we compare the behavior of the proposed measure with the highly popular Jaccard and Dice similarity measures, highlighting that the proposed approach is more sensitive to changes in interval widths. Finally, we show that the proposed similarity is bounded by the Jaccard and the Dice similarity, thus providing a reliable alternative.**

I. Introduction

Similarity measures are widely utilized in a range of applications including decision making, data aggregation, approximate reasoning, and machine learning. Measuring the similarity between two objects captures the degree to which they are alike. Similarities are commonly expressed as non-negative real numbers, often between 0 (completely dissimilar) and 1 (identical) for simplicity. Similarity is typically assumed to be symmetrical; however, for certain stimuli, similarity may be better modeled by uni-directional or asymmetric functions [1]. Various similarity measures have been introduced in the literature to assess the likeness of data structures including numerals, intervals, and crisp and fuzzy sets. As individual similarity measures have their respective strengths and weaknesses, the selection of the most appropriate measure is widely considered to be application dependent.

Recently, interval-valued data and associated interval-similarity have gained much interest as they enable the efficient representation of imprecise and uncertain information [2]. Thus, intervals have been used in many applications, including the modeling of survey data [3], the clustering of symbolic data [4], and the capturing of natural language expressions [2].

For comparing intervals in terms of their similarity, the Jaccard [5] and Dice [6] similarity measures are the most commonly used. Both of these measures provide a symmetrical similarity which increases gradually from minimum similarity (0) to maximum similarity (1) in respect to increasing intersection between the intervals. Nevertheless, they are often subject to *aliasing*, i.e., they yield the same similarity for very different interval pairs. Fig. 1 shows an example of two interval pairs for which the Jaccard and the Dice similarity measures give the same similarity of 0.33 and 0.50 respectively—though intuitively, this is unexpected. Another way of viewing the example is that the measures, in effect, are sometimes not sensitive to (changes in) the relative width of the intervals, being instead driven by the size of their intersection and union.

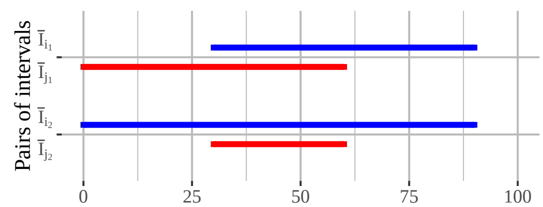


Fig. 1. Two different example pairs of intervals.

It is reasonable to consider that as the width of an interval varies, the similarity varies as well. Therefore, we propose a new similarity measure for pairs of intervals that focuses on the following similarity features:

- sensitivity to changes in the width of the intervals;
- sensitivity to the size of the intersection when one interval is a subset of another.

The proposed similarity measure uses the reciprocal overlapping ratios of the intervals to compute their asymmetric similarities which in turn are used to establish an overall symmetrical similarity, bounded by [0,1]. We compare the behavior of the new measure with the Jaccard and the Dice similarity measures using synthetic interval datasets. Along with the standard properties of similarity measures, we explore the properties of *invariance* and *linearity* for all three similarity measures.

The paper is structured as follows. Section II briefly reviews the Jaccard and the Dice similarity measures. Section III introduces the proposed similarity measure based on the overlapping ratio of intervals and discusses its properties. We demonstrate the behavior of the proposed similarity measure in comparison to both Jaccard and Dice in Section IV. Lastly, Section V concludes the paper and provides suggestions for future work.

II. Background

We now briefly review the concept of similarity measures as well as the specific similarity measures of Dice and Jaccard, the two commonly applied measures in the literature.

A. Similarity Measures

A similarity measure $S(A,B) \rightarrow [0,1]$ is a real-valued function that determines the similarity between two objects A and B. Generally, the similarity between two objects is bounded by 0 and 1, where 0 means that both objects are completely different and 1 means that they are identical. The four common properties of a similarity measure for sets A, B and C are as follows [7]:

- Boundedness: $0 \leq S(A,B) \leq 1$;
- Symmetry: $S(A,B) = S(B,A)$;
- Reflexivity: $S(A,B) = 1 \iff A = B$;
- Transitivity: If $A \subseteq B \subseteq C$ then $S(A,B) \geq S(A,C)$.

B. Jaccard Similarity Measure

The Jaccard similarity measure [5] is one of the most widely used similarity measures. It satisfies all of the above properties of a similarity measure [7]. Generally, the Jaccard similarity of two sets is defined as the ratio of the cardinality of their intersection and the cardinality of their union,

$$S_J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1}$$

Using the crisp set difference operation [8], (1) can be written as

$$S_J(A,B) = \frac{|A \cap B|}{|A \cap B| + |A \backslash B| + |B \backslash A|}, \tag{2}$$

where $A \backslash B$ is the set of items that are in A but not in B and $B \backslash A$ is the set of items that are in B but not in A. Note that this alternative form of the Jaccard similarity measure at (2) is relevant for showing its relationship with the Dice and our proposed similarity measures, detailed in Section III.

Beyond crisp sets, the Jaccard similarity measure is used to estimate the similarity for intervals or sets of intervals [9], [10]. A closed interval $\bar{I}_i$ is a set of real numbers characterized by two endpoints I_i^- and I_i^+ with $I_i^- \leq I_i^+$. The interval $\bar{I}_i$ is often represented as $[I_i^-, I_i^+]$. For comparing the intervals $\bar{I}_i$ and $\bar{I}_j$, the Jaccard similarity measure is expressed as

$$S_J(\bar{I}_i, \bar{I}_j) = \frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i \cup \bar{I}_j|}, \tag{3}$$

where $|\bar{I}_i \cap \bar{I}_j|$ is the size of the intersection between $\bar{I}_i$ and $\bar{I}_j$ and $|\bar{I}_i \cup \bar{I}_j|$ is the size of the entire interval segment(s) covering both $\bar{I}_i$ and $\bar{I}_j$. Hence, $S_J(\bar{I}_i, \bar{I}_j) = 1$ when $\bar{I}_i$ and $\bar{I}_j$ are completely overlapping and 0 when they are not overlapping at all. Similar to (2), we can rewrite (3) as

$$S_J(\bar{I}_i, \bar{I}_j) = \frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i \cap \bar{I}_j| + |\bar{I}_i \backslash \bar{I}_j| + |\bar{I}_j \backslash \bar{I}_i|}, \tag{4}$$

where $|\bar{I}_i \backslash \bar{I}_j|$ is the size of the interval segment of $\bar{I}_i$ that is not overlapping with $\bar{I}_j$ and $|\bar{I}_j \backslash \bar{I}_i|$ is the size of the interval segment of $\bar{I}_j$ that is not overlapping with $\bar{I}_i$.

Along with crisp sets and intervals, many use the Jaccard similarity measure for assessing the similarity between type-1 sets [11]. A fuzzy set [12] is defined as a set where the set's elements have membership ranging between 0 and 1. Formally, a type-1 fuzzy set F in the universe of discourse X is written as [13]

$$F = \{(x, \mu_F(x)) | x \in X\} \tag{5}$$

where $\mu_F(x) \in [0,1]$ is the membership grade of the element x in F. For two type-1 fuzzy sets F_1 and F_2, the Jaccard similarity can be written as [14]

$$S_J(F_1, F_2) = \frac{\sum_{i=1}^{N} \min(\mu_{F_1}(x_i), \mu_{F_2}(x_i))}{\sum_{i=1}^{N} \max(\mu_{F_1}(x_i), \mu_{F_2}(x_i))}, \tag{6}$$

where $\mu_{F_1}(x_i)$ and $\mu_{F_2}(x_i)$ are the membership grades of x_i in F_1 and F_2 respectively.

Equation (6) yields a value of 1 when the fuzzy sets are identical and 0 when they are completely disjoint. Note that the Jaccard similarity measure has been extended for interval type-2 [15], [16] and general type-2 fuzzy sets [17]; though, this is not discussed here.

C. Dice Similarity Measure

The Dice similarity measure [6] is closely related to Jaccard and is also a popular similarity measure. It considers the ratio of the size of the intersection of two sets and the average of their cardinality/size. Like the Jaccard similarity, it produces outputs in $[0,1]$. Specifically, for two crisp sets A and B, the Dice similarity is expressed as

$$S_D(A,B) = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)}, \tag{7}$$

where $|A|$ is the size of the set A. We can rewrite (7) by applying the crisp set difference operation [8]

$$S_D(A,B) = \frac{|A \cap B|}{|A \cap B| + \frac{1}{2}(|A \backslash B| + |B \backslash A|)}. \tag{8}$$

Note that the alternative expressions of Jaccard (2) and Dice (8) show clearly that the averaging operation in the denominator of (8) results in the Dice similarity always being equal (when sets are identical) to or larger than the Jaccard similarity. We expand on this in Section III.

In [9], [10], the Dice similarity is used along with the Jaccard similarity for interval-valued evidence. By following (4), the Dice similarity for two intervals $\bar{I}_i$ and $\bar{I}_j$ can be expressed as

$$S_D(\bar{I}_i, \bar{I}_j) = \frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i \cap \bar{I}_j| + \frac{1}{2}(|\bar{I}_i \backslash \bar{I}_j| + |\bar{I}_j \backslash \bar{I}_i|)}. \tag{9}$$

While less frequently used for fuzzy sets than Jaccard, the Dice similarity measure was used in [18], [19] for trapezoidal fuzzy numbers in the context of solving multicriteria decision-making problems.

III. Overlapping ratio based similarity measure

In this section, we introduce a new similarity measure for intervals based on their overlapping ratio. The proposed measure estimates the overall similarity of a pair of intervals by considering the reciprocal similarity of each of the intervals within the pair. We first define the concept of the overlapping ratio for a pair of intervals and, later, present the new proposed similarity measure and discuss its essential properties.

A. Overlapping Ratio of Intervals

Definition 1. The *overlapping ratio* (OR) of a given interval $\bar{I}_i$ within an interval pair $\{\bar{I}_i, \bar{I}_j\}$ captures the ratio of the size of the intersection of the pair and the size of the given interval. The OR is defined as

$$OR(\bar{I}_i, \bar{I}_j) = \frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i|}, \tag{10}$$

where $|\bar{I}_i \cap \bar{I}_j|$ is the size of the intersection between $\bar{I}_i$ and $\bar{I}_j$ and $|\bar{I}_i|$ is the size of $\bar{I}_i$. Note that for any interval $\bar{I}_i$ with a size of 0, i.e., $|\bar{I}_i| = 0$, $OR(\bar{I}_i, \bar{I}_j)$ is set to 0.

From (10), it is clear that the overlapping ratio for an interval in a pair will fall under one of the following cases:

1) $OR(\bar{I}_i, \bar{I}_j)$=1 when $\bar{I}_i$ is identical to $\bar{I}_j$;
2) $OR(\bar{I}_i, \bar{I}_j)$=0 when $\bar{I}_i$ is disjoint from $\bar{I}_j$;
3) otherwise, $0 < OR(\bar{I}_i, \bar{I}_j) < 1$.

B. Similarity Measure Based on the Overlapping Ratio

As noted, the motivation behind proposing a new similarity measure is to capture the potentially (very) different width of both intervals in the similarity calculation. Thus, the proposed overlapping ratio based similarity measure S_{OR}, defined next, takes into consideration the reciprocal similarity of both intervals within a pair in order to estimate their overall similarity.

Definition 2. The overlapping ratio based similarity measure S_{OR} for a pair of intervals, $\bar{I}_i$ and $\bar{I}_j$, is the *t-norm* of their reciprocal overlapping ratios, defined as

$$S_{OR}(\bar{I}_i, \bar{I}_j) = \bigstar(OR(\bar{I}_i, \bar{I}_j), OR(\bar{I}_j, \bar{I}_i)), \tag{11}$$

where $\bigstar$ is a *t-norm*.

In this paper, we use the minimum *t-norm* for $\bigstar$ throughout. We will discuss the product *t-norm* in future work.

Similar to (4) and (9), we can rewrite (11) as

$$S_{OR}(\bar{I}_i, \bar{I}_j) = \bigstar\left(\frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i \cap \bar{I}_j| + |\bar{I}_i \backslash \bar{I}_j|}, \frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i \cap \bar{I}_j| + |\bar{I}_j \backslash \bar{I}_i|}\right), \tag{12}$$

where $|\bar{I}_i \backslash \bar{I}_j|$ is the size of the non-overlapping segment of the interval $\bar{I}_i$ with respect to the interval $\bar{I}_j$ and vice-versa for $|\bar{I}_j \backslash \bar{I}_i|$.

Note that a distance measure $D_{OR}(\bar{I}_i, \bar{I}_j)$ can easily be derived from the S_{OR} similarity measure (11) by taking its complement—i.e., $\big(1 - S_{OR}(\bar{I}_i, \bar{I}_j)\big)$—thus capturing the dissimilarity between both intervals. We discuss this distance measure in more detail, including proving that it is a metric, in future work.

C. Properties of the Proposed Similarity Measure

This section explores the properties of the proposed overlapping ratio similarity measure $S_{OR}(\bar{I}_i, \bar{I}_j)$.

Theorem 1. *(Boundedness). $S_{OR}(\bar{I}_i, \bar{I}_j)$ is bounded by $[0, 1]$.*

Proof: Two essential boundary conditions of the *t-norm* ($\bigstar$) are $\bigstar(a, 1) = \bigstar(1, a) = a$ and $\bigstar(a, 0) = \bigstar(0, a) = 0$, $\forall a \in [0, 1]$ [20]. If a is considered as the overlapping ratio of an interval, it is always within the interval $[0, 1]$. Thus, $S_{OR}(\bar{I}_i, \bar{I}_j)$ is also bounded by $[0, 1]$. ■

Theorem 2. *(Symmetry). $S_{OR}(\bar{I}_i, \bar{I}_j)$ follows the property of symmetry. That is, $S_{OR}(\bar{I}_i, \bar{I}_j) = S_{OR}(\bar{I}_j, \bar{I}_i)$.*

Proof: The *t-norm* ($\bigstar$) is symmetric [20]. Therefore, $S_{OR}(\bar{I}_i, \bar{I}_j)$ is also symmetric. ■

Theorem 3. *(Reflexivity). $S_{OR}(\bar{I}_i, \bar{I}_j)$ follows the property of reflexivity. That is, $S_{OR}(\bar{I}_i, \bar{I}_j) = 1 \iff \bar{I}_i = \bar{I}_j$.*

Proof: If $\bar{I}_i = \bar{I}_j$, then $OR(\bar{I}_i, \bar{I}_j) = OR(\bar{I}_j, \bar{I}_i) = 1$. From the boundary conditions of the *t-norm*($\bigstar$) [20], $\bigstar(1, 1) = 1$, thus making $S_{OR}(\bar{I}_i, \bar{I}_j) = 1$. Alternatively, $S_{OR}(\bar{I}_i, \bar{I}_j) = 1$ means that both $OR(\bar{I}_i, \bar{I}_j)$ and $OR(\bar{I}_j, \bar{I}_i)$ are equal to 1. This only happens when $\bar{I}_i$ and $\bar{I}_j$ are identical intervals. ■

Theorem 4. *(Transitivity). $S_{OR}(\bar{I}_i, \bar{I}_j)$ follows the property of transitivity. That is, $S_{OR}(\bar{I}_i, \bar{I}_j) \geq S_{OR}(\bar{I}_i, \bar{I}_k)$ when $\bar{I}_i \subseteq \bar{I}_j \subseteq \bar{I}_k$.*

Proof: if $\bar{I}_i \subseteq \bar{I}_j \subseteq \bar{I}_k$, then

$$S_{OR}(\bar{I}_i, \bar{I}_j) = \bigstar\left(\frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i|}, \frac{|\bar{I}_i \cap \bar{I}_j|}{\bar{I}_j|}\right) = \bigstar\left(\frac{|\bar{I}_i|}{|\bar{I}_i|}, \frac{|\bar{I}_i|}{|\bar{I}_j|}\right) = \frac{|\bar{I}_i|}{|\bar{I}_j|},$$

$$S_{OR}(\bar{I}_i, \bar{I}_k) = \bigstar\left(\frac{|\bar{I}_i \cap \bar{I}_k|}{|\bar{I}_i|}, \frac{|\bar{I}_i \cap \bar{I}_k|}{\bar{I}_k|}\right) = \bigstar\left(\frac{|\bar{I}_i|}{|\bar{I}_i|}, \frac{|\bar{I}_i|}{|\bar{I}_k|}\right) = \frac{|\bar{I}_i|}{|\bar{I}_k|}.$$

As $\bar{I}_j \subseteq \bar{I}_k$, it follows that $|\bar{I}_j| \leq |\bar{I}_k|$. Therefore, $\frac{|\bar{I}_i|}{|\bar{I}_j|} \geq \frac{|\bar{I}_i|}{|\bar{I}_k|}$ and hence, $S_{OR}(\bar{I}_i, \bar{I}_j) \geq S_{OR}(\bar{I}_i, \bar{I}_k)$. ■

Theorem 5. *$S_{OR}(\bar{I}_i, \bar{I}_j)$ is bounded by the Jaccard and the Dice similarity measures when $\bigstar$ is the minimum t-norm. That is, $S_J(\bar{I}_i, \bar{I}_j) \leq S_{OR}(\bar{I}_i, \bar{I}_j) \leq S_D(\bar{I}_i, \bar{I}_j)$.*

Proof: For the interval pair $\{\bar{I}_i, \bar{I}_j\}$, consider the formulations of the similarity measures at (4), (9), and (12). To prove this theorem, we consider four cases: 1) $\bar{I}_i = \bar{I}_j$, 2) $\bar{I}_i \cap \bar{I}_j = \emptyset$, 3) $\bar{I}_i \subset \bar{I}_j$, and 4) $\bar{I}_i \cap \bar{I}_j \neq \emptyset$ and $\bar{I}_i \not\subset \bar{I}_j$.

Case 1: If $\bar{I}_i = \bar{I}_j$, then all three measures yield a similarity of 1. That is, $S_J(\bar{I}_i, \bar{I}_j) = S_D(\bar{I}_i, \bar{I}_j) = S_{OR}(\bar{I}_i, \bar{I}_j) = 1$.

Case 2: If $\bar{I}_i \cap \bar{I}_j = \emptyset$ (do not intersect), then all three measures give a similarity of 0. Thus, $S_J(\bar{I}_i, \bar{I}_j) = S_D(\bar{I}_i, \bar{I}_j) = S_{OR}(\bar{I}_i, \bar{I}_j) = 0$.

Case 3: If $\bar{I}_i \subset \bar{I}_j$ (complete subset), then $|\bar{I}_i \cap \bar{I}_j| = |\bar{I}_i|$. With respect to $\bar{I}_i$, there is no non-overlap segment of $\bar{I}_j$; hence, $|\bar{I}_i \backslash \bar{I}_j| = 0$. Inversely, there is a non-overlap segment of $\bar{I}_j$ in

$\bar{I}_i$; thus, $|\bar{I}_j \backslash \bar{I}_i| \neq 0$. In this case, the three similarity measures can be simplified to

$$S_J(\bar{I}_i, \bar{I}_j) = \frac{|\bar{I}_i|}{|\bar{I}_i| + |\bar{I}_j \backslash \bar{I}_i|},$$
$$S_D(\bar{I}_i, \bar{I}_j) = \frac{|\bar{I}_i|}{|\bar{I}_i| + \frac{1}{2}|\bar{I}_j \backslash \bar{I}_i|},$$
$$S_{OR}(\bar{I}_i, \bar{I}_j) = \bigstar\left(\frac{|\bar{I}_i|}{|\bar{I}_i|}, \frac{|\bar{I}_i|}{|\bar{I}_i| + |\bar{I}_j \backslash \bar{I}_i|}\right) = \frac{|\bar{I}_i|}{|\bar{I}_i| + |\bar{I}_j \backslash \bar{I}_i|},$$

which implies that

$$S_J(\bar{I}_i, \bar{I}_j) = S_{OR}(\bar{I}_i, \bar{I}_j) < S_D(\bar{I}_i, \bar{I}_j).$$

Case 4: If $\bar{I}_i \cap \bar{I}_j \neq \emptyset$ and $\bar{I}_i \not\subset \bar{I}_j$ (intersect but not complete subset), then assume $|\bar{I}_i| = w_i$, $|\bar{I}_j| = w_j$ and $|\bar{I}_i \cap \bar{I}_j| = w_{ij}$. Considering the case $w_i \leq w_j$, the three similarity measures can be rewritten as

$$S_J(\bar{I}_i, \bar{I}_j) = \frac{w_{ij}}{w_{ij} + (w_i - w_{ij}) + (w_j - w_{ij})},$$
$$S_D(\bar{I}_i, \bar{I}_j) = \frac{w_{ij}}{w_{ij} + \frac{1}{2}((w_i - w_{ij}) + (w_j - w_{ij}))},$$
$$S_{OR}(\bar{I}_i, \bar{I}_j) = \bigstar\left(\frac{w_{ij}}{w_{ij} + (w_i - w_{ij})}, \frac{w_{ij}}{w_{ij} + (w_j - w_{ij})}\right)$$
$$= \frac{w_{ij}}{w_{ij} + (w_j - w_{ij})}, \because w_i \leq w_j.$$

It is true that

$$\frac{w_{ij}}{w_{ij} + (w_i - w_{ij}) + (w_j - w_{ij})} < \frac{w_{ij}}{w_{ij} + \frac{1}{2}((w_i - w_{ij}) + (w_j - w_{ij}))},$$

thus $S_J(\bar{I}_i, \bar{I}_j) < S_D(\bar{I}_i, \bar{I}_j)$. Again, it is clear that

$$\frac{w_{ij}}{w_{ij} + (w_i - w_{ij}) + (w_j - w_{ij})} < \frac{w_{ij}}{w_{ij} + (w_j - w_{ij})},$$

implying that $S_J(\bar{I}_i, \bar{I}_j) < S_{OR}(\bar{I}_i, \bar{I}_j)$. Also,

$$\frac{w_{ij}}{w_{ij} + (w_j - w_{ij})} = \frac{w_{ij}}{w_{ij} + \frac{1}{2}(w_j - w_{ij}) + \frac{1}{2}(w_j - w_{ij})}$$
$$\leq \frac{w_{ij}}{w_{ij} + \frac{1}{2}(w_i - w_{ij}) + \frac{1}{2}(w_j - w_{ij}))}, \because w_i \leq w_j,$$

indicating that $S_{OR}(\bar{I}_i, \bar{I}_j) \leq S_D(\bar{I}_i, \bar{I}_j)$. Hence, $S_J(\bar{I}_i, \bar{I}_j) < S_{OR}(\bar{I}_i, \bar{I}_j) \leq S_D(\bar{I}_i, \bar{I}_j)$. Note that for the case $w_j \leq w_i$, the same procedure can be used to prove the above relation. ■

IV. Demonstration and Analysis

In this section, we demonstrate and analyze the behavior of the proposed S_{OR} similarity measure by comparing its output to those of both Jaccard and Dice for a set of key synthetic examples. We specifically focus on exploring the following key aspects:

- Aliasing, i.e., similarity measures producing the same output for different input intervals.
- Behavior when one interval is a complete subset of another.
- Behavior for intervals of equal size and equal overlapping ratios.
- Invariance to scaling/multiplication of interval endpoints.
- Linearity in measure output in respect to linearly increasing interval overlap.

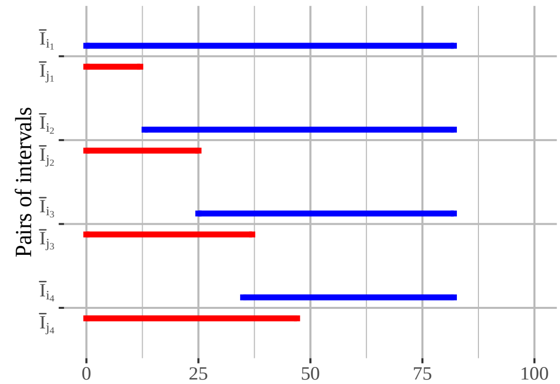


Fig. 2. Interval pairs used to demonstrate the results of similarity measures for changes in the width of the intervals.

TABLE I
Similarity results for the interval pairs as shown in Fig. 2.

Interval Pair	S_J	S_D	S_{OR}
I	0.15	0.26	0.15
II	0.15	0.26	0.17
III	0.15	0.26	0.21
IV	0.15	0.26	0.26

1) Experiment on aliasing: In Fig. 2, four different pairs of intervals $\{\bar{I}_i, \bar{I}_j\}$ are considered, where all pairs have an intersection of equal size. The similarity results for the pairs using the three similarity measures are shown in Table I. The S_J and the S_D measures give a similarity of 0.15 and 0.26 respectively for all pairs. Indeed, both measures provide identical similarities for pairs of intervals when the size of the union of their non-overlapping segments remains constant. On the contrary, the proposed S_{OR} measure yields different similarity for all cases. Note, that as shown in Theorem 5 the results of the S_{OR} measure are bounded by the similarity produced by the S_J and S_D measures. The reason that the S_{OR} measure produces different results for each case is that it captures changes in the width of both input intervals which affects their reciprocal similarity and the overall similarity.

2) Experiment with interval pairs when one interval is a complete subset of the other: Five interval pairs are shown in Fig. 3, where $\bar{I}_j$ is a complete subset of $\bar{I}_i$ in all pairs and overlapping by 10%, 20%, 30%, 40%, and 50%, respectively of $\bar{I}_i$. Table II presents the similarity for all pairs with all three measures. Note that for all pairs, the overlapping ratio of $\bar{I}_j$ is 1 while the overlapping ratio of $\bar{I}_i$ depends on the size of $\bar{I}_i$ and $\bar{I}_j$, i.e., $\frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i|}$. Therefore, intuitively their mutual similarity can be at most $\frac{|\bar{I}_i \cap \bar{I}_j|}{|\bar{I}_i|}$ for each pair. The S_J and the S_{OR} measures perform accordingly while the S_D measure exceeds this limit.

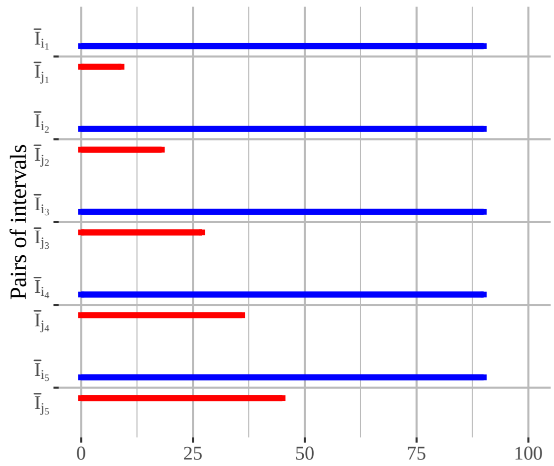


Fig. 3. One interval as a complete subset of the other interval.

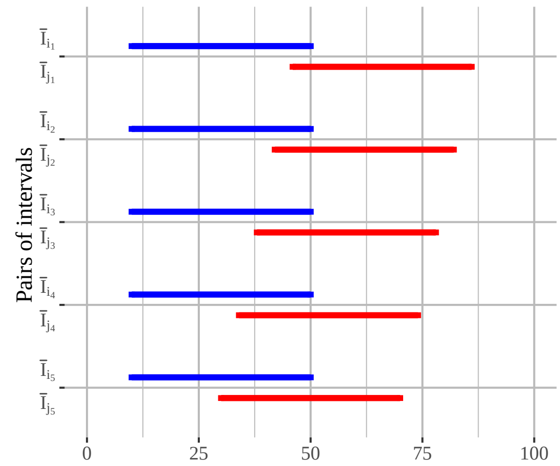


Fig. 4. Interval pairs with equal width and equal ratio of intersection.

TABLE II
SIMILARITY RESULTS FOR THE INTERVAL PAIRS AS SHOWN IN FIG. 3.

Interval Pair	S_J	S_D	S_{OR}	subset by
I	0.10	0.18	0.10	10%
II	0.20	0.33	0.20	20%
III	0.30	0.46	0.30	30%
IV	0.40	0.57	0.40	40%
V	0.50	0.67	0.50	50%

TABLE III
SIMILARITY RESULTS FOR THE INTERVAL PAIRS AS SHOWN IN FIG. 4.

Interval Pair	S_J	S_D	S_{OR}	intersected by
I	0.05	0.10	0.10	10%
II	0.11	0.20	0.20	20%
III	0.18	0.30	0.30	30%
IV	0.25	0.40	0.40	40%
V	0.33	0.50	0.50	50%

3) Experiment with interval pairs of equal size and equal overlapping ratio: In Fig. 4, five interval pairs are shown, where the intervals are of equal size and their intersection is varied to be 10%, 20%, 30%, 40%, and 50% of their size. Table III provides the results for all pairs using the three similarity measures. In all pairs, the overlapping ratio is equal, and it is intuitive to expect the similarity to be the extent of this overlapping ratio. In this case, the S_D and the S_{OR} measures satisfy the expectation whereas the S_J measure yields lower similarity.

4) Experiment on invariance: Five pairs of intervals are shown in Fig. 5 where both endpoints of I_{i_1} and I_{j_1} are gradually multiplied by a factor, $n = \{2, 3, 4, 5\}$ to produce new interval pairs. Yet, the overlapping ratio is maintained for individual intervals in all the pairs. Adapting the definition from [21], a similarity measure is *invariant* if its similarity output remains constant regardless of multiplying the endpoints of interval pairs by a factor. Table IV shows the similarity for all pairs using the three measures where n refers to the factor applied to the interval endpoints. The results shows that all three measures satisfy the *invariance* property for the given pairs of intervals.

5) Experiment on linearity: Adapting the definition from [21], a similarity measure on intervals is *linear* if its similarity output varies linearly in respect to a linear change in the size of the intersection of the intervals. In Fig. 6(a), the intersection between two intervals of equal size is gradually increased in 10% steps. The corresponding similarity outputs for the pairs and all three measures are shown graphically in Fig. 6(b). In this case, the S_D and the S_{OR} measures exhibit *linearity* while the S_J measure exhibits *convexity* (differences rise with the increase of intersection).

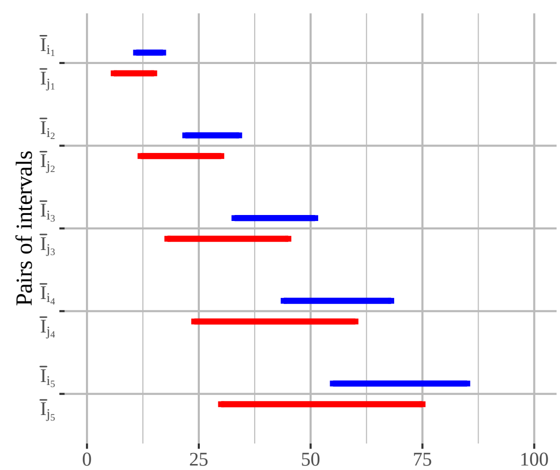


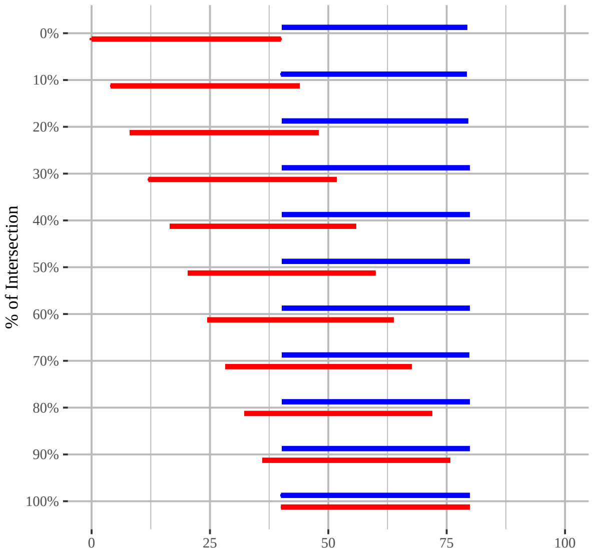
Fig. 5. Interval pairs used to show the invariance of similarity measures.

TABLE IV
SIMILARITY RESULTS FOR THE INTERVAL PAIRS AS SHOWN IN FIG. 5.

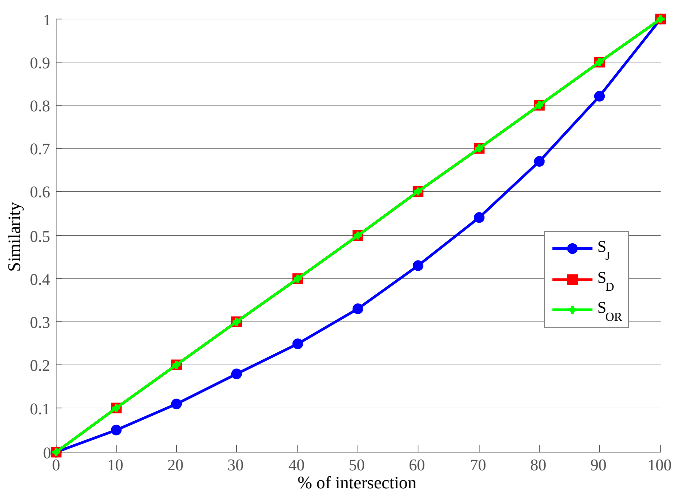
Interval Pair	S_J	S_D	S_{OR}	multiplied by n
I	0.36	0.53	0.44	1
II	0.36	0.53	0.44	2
III	0.36	0.53	0.44	3
IV	0.36	0.53	0.44	4
V	0.36	0.53	0.44	5

V. CONCLUSION AND FUTURE WORK

In this paper, we have introduced a new similarity measure that considers the reciprocal similarity of a pair of intervals



(a) Interval pairs



(b) Similarity versus overlap

Fig. 6. Interval pairs used to show the linearity of similarity measures.

in computing their overall similarity. We have used the overlapping ratio of the intervals within the pair for capturing the asymmetric similarity. We have also demonstrated that the new measure satisfies essential properties of a similarity measure. Lastly, we have compared the behavior of the proposed measure with the two popular Jaccard and Dice similarity measures using synthetic datasets. The results have shown that the proposed similarity measure is more sensitive to the changes in the width of intervals, and further it is *invariant* and *linear*. We have also proved that the proposed similarity is bounded by the Jaccard and the Dice similarity.

In future, we will use the proposed similarity measure for capturing the mutual agreement of interval-valued evidence for aggregation. As each α-cut of a normal and convex fuzzy set is a closed interval [22], we aim to extend the proposed similarity measure for comparing fuzzy sets.

Acknowledgment

Shaily Kabir acknowledges the financial support of the Commonwealth Scholarship Commission in the UK.

References

[1] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.

[2] Y. Ren, Y.-H. Liu, J. Rong, and R. Dew, "Clustering interval-valued data using an overlapped interval divergence," in *Proc. 8th Australasian Data Mining Conf. (AusDM'09)*. Melbourne, Australia: Australian Computer Society, Inc., 2009, pp. 35–42.

[3] C. Wagner, S. Miller, J. M. Garibaldi, D. T. Anderson, and T. C. Havens, "From interval-valued data to general type-2 fuzzy sets," *IEEE Trans. Fuzzy Systems*, vol. 23, no. 2, pp. 248–269, 2015.

[4] D. S. Guru, B. B. Kiranagi, and P. Nagabhushan, "Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns," *Pattern Recognition Letters*, vol. 25, no. 10, pp. 1203–1213, 2004.

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

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

[7] J. McCulloch, C. Wagner, and U. Aickelin, "Analysing fuzzy sets through combining measures of similarity and distance," in *Proc. IEEE Int. Conf. Fuzzy Systems*, Beijing, China, 2014, pp. 155–162.

[8] K. H. Rosen, *Discrete mathematics and its applications*, 7th ed. McGraw-Hill, 2012.

[9] 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.

[10] 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.

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

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

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

[14] C. Wagner, S. Miller, and J. M. Garibaldi, "Similarity based applications for data-driven concept and word models based on type-1 and type-2 fuzzy sets," in *Proc. IEEE Int. Conf. Fuzzy Systems*, Hyderabad, India, 2013, pp. 1–9.

[15] 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 (InTech'08)*, Samui, Thailand, 2008, pp. 47–55.

[16] D. Wu and J. M. Mendel, "A comparative study of ranking methods, similarity measures and uncertainty measures for interval type-2 fuzzy sets," *Information Sciences*, vol. 179, no. 8, pp. 1169–1192, 2009.

[17] 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.

[18] 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.

[19] ——, "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.

[20] D. M. Gabbay and J. Woods, *Handbook of the History of Logic: The Many Valued and Nonmonotonic Turn in Logic*. Elsevier, 2007.

[21] R. E. Tulloss, "Assessment of similarity indices for undesirable properties and a new tripartite similarity index based on cost functions," *Mycology in sustainable development: expanding concepts, vanishing borders*, pp. 122–143, 1997.

[22] J. Navarro, C. Wagner, U. Aickelin, L. Green, and R. Ashford, "Measuring agreement on linguistic expressions in medical treatment scenarios," in *Proc. IEEE Symp. Computational Intelligence*, Athens, Greece, 2016, pp. 1–8.