How to Compare Uncertain Data Types - Towards Robust Similarity Measures



Shaily Kabir

School of Computer Science

University of Nottingham

This dissertation is submitted for the degree of

Doctor of Philosophy

January 2020

To my parents for their love, sacrifice, and support.

Declaration

The works in this dissertation are based on research carried out at the Intelligent Modelling and Analysis (IMA), and Lab for Uncertainty in Data and Decision Making (LUCID) Research Groups, School of Compute Science, University of Nottingham, UK. I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Shaily Kabir

January 2020

Acknowledgements

At the beginning, I express my utmost gratitude to the Almighty Allah for bestowing His blessings on me to accomplish this arduous PhD study with completion of the dissertation.

I would like to express deepest gratitude and appreciation to my supervisor, Professor Christian Wagner for his continuous support, patience, encouragement, and guidance throughout my PhD study. My sincere thankfulness goes to him for his insightful directions, necessary revisions, and constructive suggestions on the preparation of a number of conference and journal papers as well as my dissertation. He also had been very kind and helpful in many personal matters. I could not have imagined having such a cooperative supervisor during for my PhD study! I am also thankful to Dr Josie McCulloch, Research Fellow at the School of Computer Science, University of Nottingham for her kind assistance in finding useful research materials and giving suggestions at many occasions.

Besides my advisors, I would like to thank Dr Timothy C. Havens, Michigan Technological University, USA and Dr Derek Anderson, University of Missouri, USA for their useful suggestions and comments while we teamed up to write several conference papers and journals.

My sincere thanks also goes to my student colleagues and friends for lending their hands and sparing time. Particularly, I owe to my fellow labmates Utkarsh Agrawal, Elissa Maadi, Tajul Razak, Chao Chen, Direnc Pekaslan, and Francisco Navarro

Barron. As I recall, we all worked together, spent hours and sleepless nights in the lab, had fun, stimulated discussions, and always were supportive to each other. Surely, I would remember the happiness, sorrows, and cooperation during my PhD study at Nottingham. I also thankful to my friend Mosarrat Jahan for encouraging me throughout my PhD journey.

I am also thankful to the Commonwealth Scholarship Commission (CSC), UK for providing scholarship and living expenses during my PhD program. Needless to say, without their financial support, it would not have been possible for me to continue and complete this program. I also would like to extend my appreciation to Vice Chancellor, Dhaka University and Chairman, Computer Science and Engineering Department, Dhaka University for granting study leave aiming to finish my PhD study. Last but not the least, it would have been difficult on my part to accomplish this journey without continuous support and sacrifice from my beloved parents, my husband, my daughter 'Autri', and my son 'Aumiyo'. They have missed my company and affection a lot as I had to be always busy in the lab doing my research works. Their unending love and uncompromising faith always inspired me and strengthened my confidence throughout my PhD journey.

Abstract

In view of the importance of effective comparison of uncertain data in real-world applications, this thesis focuses on developing new similarity measures with high accuracy. As it identifies and articulates an inherent limitation of the popular set-theoretic similarity measures for continuous intervals where they return the same similarity value for very different sets of intervals (termed as *aliasing*), this thesis first underpins a new axiomatic definition of a *robust* similarity measure and then proposes a new similarity measure for continuous intervals based on their bidirectional subsethood. Beyond establishing theoretical foundation of the new measure, the thesis also demonstrates its robust results vis-a-vis existing measures and suitability for real world applications. In the next stage, it develops a generalized framework to assess similarity between discontinuous intervals as current approaches involve loss of discontinuity information and are also affected by *aliasing* of the popular measures—these weaknesses impact the accuracy of similarity results. This thesis further integrates Allen's theory with the new generalized framework to make the latter more efficient.

Moving beyond intervals, this thesis extends the new similarity measure both vertically and horizontally (α-cut based) for comparing type-1 (T1) fuzzy sets as the shortcoming of popular similarity measures persists with their extension to T1 fuzzy sets. The empirical evaluation of the extended new measures with respect to key existing fuzzy set-theoretic similarity measures shows that the vertically extended new measure behaves intuitively for various types of fuzzy sets, except for non-normal fuzzy

sets; however, the α-cut based extended new measure meets expectation in all cases. At the final stage, the utility of the new similarity measure is explored to improve the robustness of fuzzy integral (FI) based uncertain-data aggregation. As existing approaches to generate fuzzy measures (FMs) rely on popular similarity measures to capture the degree of similarity among individual sources (and their combinations), they are also impacted by their aforesaid limitation. Therefore, this thesis develops a new FM based on the new similarity measure which can generate intuitive aggregation outcome when used in combination with an FI.

Table of contents

List of figures **xvii**

List of tables **xxiii**

1 Introduction **1**

1.1 Background and Motivation 2

1.2 Research Aim & Objectives 4

1.3 Contribution to Knowledge 5

1.4 Organization of the Thesis 8

2 Research Background **11**

2.1 Introduction 11

2.2 Modelling of Uncertainty 12

2.3 Intervals 13

2.3.1 Continuous Intervals 13

2.3.2 Discontinuous Intervals 17

2.4 Fuzzy Sets 19

2.4.1 Type-1 Fuzzy Sets 20

2.4.2 Set-Theoretic Operations on Type-1 Fuzzy Sets 24

2.5 Similarity Measures for Uncertain Data Types 25

2.5.1 Set-Theoretic Similarity Measures for Intervals 27

2.5.2 Set-Theoretic Similarity Measures for Type-1 Fuzzy Sets 31
2.6 Subsethood 35
2.6.1 For Crisp Sets 35
2.6.2 For Fuzzy Sets 36
2.7 Aggregation of Uncertain Data Types 38
2.7.1 Fuzzy Measures 39
2.7.2 Fuzzy Integrals 43
2.8 Summary 44

3 Measuring Similarity Between Continuous Intervals 47
3.1 Introduction 47
3.2 Motivation 48
3.3 Proposed Similarity Measure 50
3.3.1 Definition of Subsethood for Continuous and Closed Intervals . 50
3.3.2 An Expanded Axiomatic Definition for Robust Similarity Measures on Continuous and Closed Intervals 52
3.3.3 Proposed Similarity Measure based on Bidirectional Subsethood 53
3.3.4 Properties of the Proposed Similarity Measure 54
3.4 Proposed Distance Measure 62
3.5 Demonstration and Analysis 67
3.5.1 Synthetic Dataset Based Demonstration 68
3.5.2 A Real World Example 75
3.6 Summary 83

4 Measuring Similarity Between Discontinuous Intervals 85
4.1 Introduction 85
4.2 Motivation 86

4.3 Proposed Generalized Framework 89
4.4 Refinement of the Proposed Generalized Framework with Allen's Theory 98
4.4.1 Derivation of Similarity Values from Allen's Theory 99
4.4.2 Extension of Allen's Theory 99
4.4.3 Integration of Proposed Generalized Framework with Extended Binary Relations of Allen's Theory 100
4.5 Demonstration and Analysis 102
4.5.1 Datasets and Experimental Setup 102
4.5.2 Demonstration with Synthetic Dataset-1 103
4.5.3 Demonstration with Synthetic Dataset-2 104
4.5.4 Demonstration with Synthetic Dataset-3 109
4.6 Summary 113

5 Measuring Similarity Between Type-1 Fuzzy Sets 115
5.1 Introduction 115
5.2 Motivation 116
5.3 Proposed Vertical-Slice Based Extension 118
5.3.1 Proposed Vertical-Slice Based Extension of the Bidirectional Subsethood Based Similarity Measure 118
5.3.2 Properties of the Vertical-Slice Based Extension 119
5.3.3 Comparison with Current Literature 123
5.4 Proposed α-cut Based Extension 130
5.4.1 Proposed α-cut Based Extension of the Bidirectional Subsethood Based Similarity Measure 131
5.4.2 Properties of the α-cut Based Extension 133
5.4.3 Dealing with Non-normal Fuzzy Sets 134
5.4.4 Dealing with Non-convex Fuzzy Sets 136

5.4.5 Comparison with Current Literature 139
5.5 Summary 145

6 Aggregation of Multi-Source Interval-Valued Data 149
6.1 Introduction 149
6.2 Motivation 151
6.3 A New Fuzzy Measure Using Inter-Source Agreement 154
6.3.1 Defining Subsethood for a Set of Intervals 154
6.3.2 Defining Bidirectional Subsethood Based Similarity Measure for a Set of Intervals 157
6.3.3 Proposed Agreement-Based Fuzzy Measure 158
6.4 Demonstration and Analysis 165
6.4.1 Demonstration with Synthetic Dataset-1 166
6.4.2 Demonstration with Synthetic Dataset-2 169
6.4.3 A Real-World Example 171
6.5 Conclusion 175

7 Conclusions 177
7.1 Thesis Summary 177
7.2 Contributions 180
7.3 Limitations 183
7.4 Future Work 184

References 187

Appendix A Similarity Results for Summer and Winter Temperature Data 203

Appendix B Survey on Assessing Security Risk of a Proposed Computing System 207

List of figures

1.1 Jaccard and Dice measures showing *aliasing* for Pair-1 and Pair-2. . . . 3

2.1 A continuous and closed interval, $\overline{a} = [2, 8]$. 14
2.2 Seven basic relations of Allen's theory on continuous intervals $\overline{a}$ and $\overline{b}$. 16
2.3 A discontinuous interval, $\overline{\overline{a}} = \{[2, 4], [6, 8]\}$. 17
2.4 Conversion of a discontinuous interval into a continuous interval. 18
2.5 Three different membership functions for T1 fuzzy sets. 21
2.6 Normal and non-normal T1 fuzzy sets. 22
2.7 Two different presentations of T1 fuzzy set. 23
2.8 Convex and non-convex T1 fuzzy sets. 23
2.9 The FM lattice for three data sources. 40

3.1 Two different pairs of intervals showing *aliasing* with the Jaccard and Dice similarity measures. 49
3.2 Example that shows to satisfy the triangle inequality by the D_{S_h} measure. 66
3.3 Interval pairs used to demonstrate the *aliasing* of similarity measures. . 68
3.4 Interval pairs where one interval as a complete subset of the other. . . . 70
3.5 Interval pairs with equal size and equal ratio of overlap. 72
3.6 Interval pairs used to demonstrate the invariance of similarity measures. 73

3.7 Displaying linearity by the S_D and $S_{S_h\wedge}$ measures, and convexity by the S_J and $S_{S_h\Pi}$ measures. 74
3.8 Standard areas (districts) of the UK [139]. 76
3.9 Popular similarity measures show identical similarity for two pairs of areas having different ranges of temperature in Autumn. 82

4.1 Two different pairs of discontinuous intervals. 87
4.2 Continuous forms of discontinuous intervals in Pair-1 and Pair-2 of Fig. 4.1 after applying the 'convexify' function (see [44]). 87
4.3 Another two different pairs of discontinuous intervals. 88
4.4 Resulting continuous intervals after applying the interval addition (see [34]) to Pair-3 and Pair-4 of discontinuous intervals in Fig. 4.3. 88
4.5 Results of averaging the similarity of all subinterval pairs to determine the overall similarity between discontinuous intervals. 91
4.6 Using S_J, S_D, and S_{S_h} similarity measures, the similarity results of all combinations of subintervals within a pair of discontinuous intervals—one with three and the other with two subintervals. 92
4.7 Three approaches to compute similarity between two different pairs of discontinuous intervals. 95
4.8 Three approaches to compute similarity between two different pairs of discontinuous intervals. 97
4.9 Comparison between the generalized framework (S_G) and its combination with the extended Allen's theory (S_GA) in respect to the S_{S_h} measure. 102
4.10 Dataset-1 with varying degree of overlap between subintervals, $\overline{\overline{a}}$ and $\overline{\overline{b}}$ to examine how smoothly similarity results alter. 105

4.11 Similarity results showing proportionate decline by the generalized framework (S_G) with both S_{S_h} and S_D similarity measures, but not with the S_J similarity measure for different scenarios in Fig. 4.10. 106

4.12 Dataset-2 comprising pairs of discontinuous intervals with an increased number of subintervals and varying degree of overlap along with their similarity results. Note: The generalized framework (S_G) with the S_J and S_D measures return identical results for scenarios 2.3 and 2.4, i.e., they are subject to *aliasing*, but with only the S_{S_h} measure captures the change in the respective discontinuous intervals and thus avoids *aliasing*. 107

5.1 Two different example pairs of T1 fuzzy sets. 117

5.2 Interval representations for the pairs of T1 fuzzy sets in Fig. 5.1. 117

5.3 Trapezoidal and Gaussian fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the degree of overlap between fuzzy sets is increasing from right to left. 124

5.4 Trapezoidal and Gaussian T1 fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the degree of subsethood between fuzzy sets is increasing. 125

5.5 Trapezoidal and Gaussian fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the height of the non-normal fuzzy set is increasing while the other set is normal. 126

5.6 Trapezoidal and Gaussian T1 fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the height of the non-normal fuzzy set is increasing while the other set is also non-normal. 127

5.7 Trapezoidal and Gaussian T1 fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the 'degree of non-convexity' of the fuzzy set is increasing. 128

5.8 Two different pairs of T1 fuzzy sets where both non-normal fuzzy sets are in Pair-1, and one normal and one non-normal fuzzy sets are in Pair-2. 136

5.9 A pair of convex and non-convex fuzzy sets. 137

6.1 Example highlighting the behaviour of the AG FM [22], where $\overline{U}_2$ and $\overline{U}_3$ capture the union of the intersections of all of two, respectively three source combinations as per Eq. (2.29). 152

6.2 Two different interval-valued sets $\overline{h} = [\overline{h}_1, \overline{h}_2]$ and $\overline{r} = [\overline{r}_1, \overline{r}_2]$ with equal Jaccard similarity (S_J) of 0.33 and Dice similarity (S_D) of 0.50 respectively. 153

6.3 Comparison of aggregation results from the CFI [129] with the AS_h FM and the three other existing FMs (AG, $GenA$ and AA FMs) for four different synthetic sets of intervals. 167

6.4 (a) Five sets of interval-valued data with degrading interval-overlap (b) The aggregation results of the AG, $GenA$, AA and AS_h FMs in combination with the CFI [129] for Fig. 6.4(a). 170

6.5 Comparison of aggregation results from the CFI in respect to the AS_h FM and three other existing FMs (AG, $GenA$ and AA FMs) for experts ratings (Group A and C) on the difficulty to *attack* component-5. . . . 173

6.6 Comparison of aggregation results from the CFI in respect to the AS_h FM and three other existing FMs (AG, $GenA$ and AA FMs) for experts ratings (Group D and E) on the difficulty to *attack* component-11. . . . 174

List of tables

3.1 Similarity and distance between the intervals of Fig. 3.2. 66

3.2 The S_J and S_D measures exhibit *aliasing* for the interval pairs in Fig. 3.3. 68

3.3 Similarity results for the interval pairs as shown in Fig. 3.4. 71

3.4 Similarity results for the interval pairs as shown in Fig. 3.5. 72

3.5 All three measures display invariant similarity for the interval pairs in Fig. 3.6. 73

3.6 Seasonal Temperature in Standard Areas of the UK in 2016. 76

3.7 Similarity results for temperature data using the S_J measure. 78

3.8 Similarity results for temperature data using the S_D measure. 79

3.9 Similarity results for temperature data using the $S_{S_h\wedge}$ measure. 80

3.10 Similarity results for temperature data using the $S_{S_h\Pi}$ measure. 81

4.1 Derivation of Similarity Value from Allen's Theory. 99

4.2 Performances of the proposed generalized framework (S_G) against its combination with the extended Allen's theory (S_GA) in respect to total similarity calculation and execution time in milliseconds (ms). 110

5.1 Summary of Experiment Results (✓- expected behaviour and × - unexpected behaviour). 129

5.2 Similarity results for the pairs of fuzzy sets (in Fig. 5.8) using the $S^{T1:NN}_{S_h\alpha}$ measure. 136

5.3 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where their overlap are increased from right to left. 140

5.4 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where the degree of subsethood between fuzzy sets is increasing. 141

5.5 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where the height of the non-normal fuzzy set is increasing while the other set remains normal. . 141

5.6 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where the height of the non-normal fuzzy set is increasing while the other set is also non-normal. 142

5.7 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where the 'degree of non-convexity' of fuzzy set is increasing. 143

5.8 Summary of Experiment Results (✓- expected behaviour and × - unexpected behaviour). 144

5.9 Execution time (ms) for different discretization levels (250, 500, 750, 1000) along the whole support of Pair-1 of the Gaussian fuzzy sets (A', B'_1) in Fig. 5.4(b). 144

6.1 The worth of subsets of intervals using the $\tilde{g}^{AS_h}$ FM prior to normalization. 161

6.2 The worth of subsets of intervals using the g^{AS_h} FM after normalization. 162

6.3 The normalized worth of subsets of intervals using the AS_h FM (g^{AS_h}) 168

A.1 Similarity results for Summer temperature data using the Jaccard similarity measure (S_J) 203

A.2 Similarity results for Winter temperature data using the Jaccard similarity measure (S_J) 204

A.3 Similarity results for Summer temperature data using the Dice similarity measure (S_D) 204

A.4 Similarity results for Winter temperature data using the Dice similarity measure (S_D) 205

A.5 Similarity results for Summer temperature data using the proposed similarity measure (S_{S_h}) 205

A.6 Similarity results for Winter temperature data using the proposed similarity measure (S_{S_h}) 206

Acronyms and Notation

FM	Fuzzy Measure
FI	Fuzzy Integral
CFI	Choquet Fuzzy Integral
SFI	Sugeno Fuzzy Integral
AG	Fuzzy Measure of Agreement
$GenA$	Fuzzy Measure of Generalized Accord
AA	Additive Measure of Agreement
AS_h	Proposed Bidirectional Subsethood Based Fuzzy Measure
S	Similarity Measure
D	Distance Measure
S_J	Jaccard Similarity Measure
S_D	Dice Similarity Measure
S_T	Tversky's Ratio Model
S_h	Subsethood
S_{S_h}	Proposed Bidirectional Subsethood Based Similarity Measure
D_{S_h}	Proposed Bidirectional Subsethood Based Distance Measure
S_G	Proposed Generalized Framework for Discontinuous Intervals
S_GA	Proposed Generalized Framework Combined with Extended Allen's Theory
X	Universe of Discourse
a	Crisp Set
$\overline{a}$	Continuous Interval $\{\overline{a} \subseteq \mathbb{R} : \overline{a} = [a^-, a^+],\ a^- \leq a^+\}$
$\overline{\overline{a}}$	Discontinuous Interval $\{\overline{\overline{a}} = \bigcup_{i=1}^{m} \overline{a}_i = \bigcup_{i=1}^{m} [a_i^-, a_i^+]\}$
A	Fuzzy set $\{(x, \mu_A(x)) \,\vert x \in X \text{ and } \mu_A(x) \in [0,1]\}$

$T1$	Type-1 fuzzy set
$T2$	Type-2 fuzzy set
$\overline{A}_{\alpha}$	An α-cut of A
$S^{T1}_{S_h v}$	Proposed Vertical-slice Based Extended Bidirectional subsethood based Similarity Measure for T1 fuzzy sets
$S^{T1}_{S_h \alpha}$	Proposed α-cut Based Extended Bidirectional subsethood based Similarity Measure for T1 fuzzy sets
$\overline{h}$	A Set of Continuous Intervals
$\tilde{g}^{AG}$	Fuzzy Measure of Agreement (Prior normalization)
g^{AG}	Fuzzy Measure of Agreement (After normalization)
g^{GenA}	Fuzzy Measure of Generalized Accord
g^{AA}	Additive Measure of Agreement
$\tilde{g}^{AS_h}$	Proposed Fuzzy Measure Based on Bidirectional Subsethood Based Similarity Measure (Prior normalization)
g^{AS_h}	Proposed Fuzzy Measure Based on Bidirectional Subsethood Based Similarity Measure (After normalization)

Chapter 1

Introduction

In the real world, uncertainty greatly influences human reasoning and decision making [1]. Efficient capturing and modelling of uncertain data is, therefore, prerequisite for achieving better decision making. Classical set theory, based on two-valued logic 0 and 1, can capture certain objects with clear and precise definitions, such as, 'all negative numbers'; however, it fails to present objects which are vaguely and imprecisely defined, such as, 'all good students' [2]. To overcome this difficulty, the concept of fuzzy set theory—a generalization of classical set theory—was developed [2]. Fuzzy set theory provides a mathematical framework to capture and model the uncertainty in data. A fuzzy set allows an object to be a partial member of the set where the degree of membership of the object varies from 0 to 1 [3]. Although fuzzy sets were originally designed as type-1 (T1), different extensions of the T1 fuzzy sets, such as type-2 (T2) and higher number fuzzy sets, have been later introduced with a view to providing more effective models of uncertainty [4].

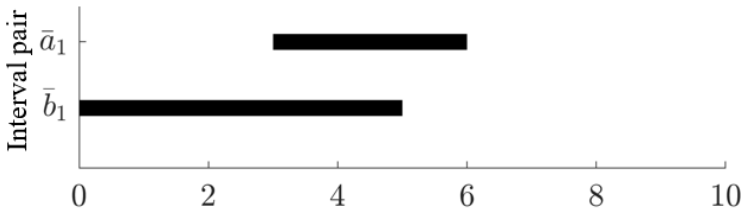
Apart from the fuzzy sets, interval-valued data have recently gained interest for the modelling of uncertainty and vagueness, particularly in the modelling of survey data [5], the representation of symbolic data [6], and the capture of natural language expressions [7]. An interval offers a simple and efficient representation of uncertain data where it

is defined as a set of real numbers, bounded by the left and right endpoints [8]. Here, the width of an interval is important as to capturing the degree of uncertainty in data where a narrow interval reflects greater certainty against less certain wider interval [5].

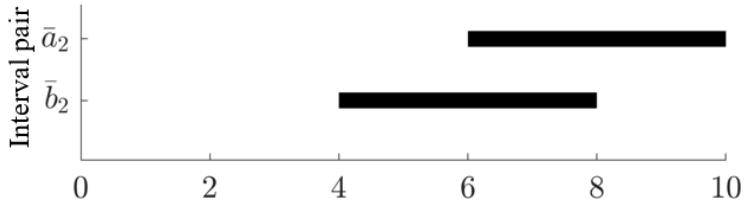
Many real-world applications, such as pattern recognition, decision making, data aggregation, and approximate reasoning often involve uncertain information as inputs, and the outcomes greatly depend on their effective comparison [9]. Considering the significance of effective comparison, this thesis focuses on developing novel similarity measures for comparing the uncertain information—captured by intervals and T1 fuzzy sets (common and advanced features)—with high accuracy.

1.1 Background and Motivation

Many similarity measures have been proposed in the literature where each measure has its own strengths and weaknesses [10]. Typically, a similarity measure determines how similar two objects are, where their degree of similarity is often bounded by 0 (for dissimilar objects) and 1 (for identical objects). In general, similarity measures are categorized as distance-based and set-theoretic measures [11]. Among them, set-theoretic measures are widely used due to the limitation of the other category in comparing objects with qualitative features [12]. In the context of intervals, classical set-theoretic similarity measures are used for their comparison where the Jaccard [13] and Dice [14] similarity measures are the most popular ones. However, these popular measures, along with most set-theoretic measures have the tendency to be subject to *aliasing* where they offer the same similarity for very different pairs of sets/intervals with different cardinalities or sizes [15–17] (discussed in detail in Chapter 3). This *aliasing* poses problems as it has the potential to give counter-intuitive results and poor automated reasoning in many real-world applications where the intervals need to be systematically compared. Consider an example in Fig. 1.1 with two different pairs



(a) Pair-1



(b) Pair-2

Fig. 1.1 Jaccard and Dice measures showing *aliasing* for Pair-1 and Pair-2.

of intervals for which the Jaccard and Dice measures are facing *aliasing* by generating the same similarity of 0.33 and 0.50 respectively.

So far, the comparison of discontinuous intervals has received very little attention. In reality, there is no specific similarity measure which can be directly applicable for comparing discontinuous intervals. At present, the common approach to assess their similarity is to first convert the discontinuous intervals into their continuous forms [18] and then apply any similarity measure compatible to continuous intervals. This approach is rather simple but involves loss of discontinuity information and thus affects the accuracy of the similarity results [19].

For fuzzy sets, any set-theoretic similarity measure on intervals can be leveraged for assessing their similarity, as both vertical-slice and α-cut based representations of fuzzy sets are, in fact, intervals. In practice, both the vertical-slice based and α-cut based extensions of similarity measures on intervals are used. But, it is important to evaluate their performances in generating expected similarity results for fuzzy sets with various features, such as normality/non-normality and convexity/non-convexity [20].

Similarity measures are often used to aggregate 'similar' data from multiple uncertain sources using aggregation functions where FIs are used to combine data from multiple sources. FIs are defined in respect to FMs which capture a notion of 'worth' of both the individual data sources and all their possible combinations. As existing approaches to generate these FMs rely on popular set-theoretic similarity measures to capture the degree of similarity of individual sources (and their combinations), they are also affected by the limitations (i.e., *aliasing*) of those similarity measures. Hence, developing and applying a *robust* similarity measure, not subject to *aliasing*, potentially has a great utility to avoid the shortcomings in inter-source agreement calculation and thus improves robustness of FI based data aggregation/fusion of uncertain data.

1.2 Research Aim & Objectives

The primary research aim of this thesis is to develop novel, effective and efficient set-theoretic similarity measures for comparing uncertain information, modelled by intervals and fuzzy sets. Specifically, the new measures are intended to capture the similarity between intervals and fuzzy sets with normal/non-normal, convex/non-convex properties with high accuracy. To achieve the research aim, the following objectives are identified:

1. Investigate popular set-theoretic similarity measures on intervals to identify and articulate their potential limitations.

2. Introduce a new axiomatic definition of a *robust* similarity measure to address the identified limitations of existing similarity measures.

3. Put forward a novel set-theoretic similarity measure for intervals that satisfies the full set of axioms for a *robust* similarity measure, thus minimizes the identified limitations and provides similarity with high accuracy.

4. Derive a new dissimilarity measure, ideally a metric for intervals from the new similarity measure so as to directly apply it in real-world applications, such as pattern recognition.

5. Extend the generality of new interval similarity measure to discontinuous intervals and T1 fuzzy sets.

6. Investigate existing agreement-based FMs to identify whether they are also affected by the shortcomings (i.e., *aliasing*) of common similarity measures.

7. Develop a new FM which can avoid the identified limitations of inter-source agreement (based on similarity) and for demonstration and evaluation purposes, employ it in respect to a FI for providing robust aggregation of uncertain multi-source data.

1.3 Contribution to Knowledge

The research in this thesis develops a new *robust* similarity measure for intervals by addressing the limitation of popular set-theoretic measures. The new measure is extended to deal with the challenges in assessing similarity for other complex data-types – discontinuous intervals and T1 fuzzy sets with non-normal and non-convex properties. The new measure is further expanded to generate a new FM, which in combination with FI brings robustness in aggregation of uncertain multi-source data.

More precisely, the contributions resulting from this thesis are as follows:

- An inherent limitation of popular set-theoretic similarity measures on continuous intervals is identified and articulated, particularly, their tendency to be subject to *aliasing* where they return same similarity for very different sets of intervals having different cardinalities or sizes.

- New axioms are introduced to establish desirable properties of *robust* similarity measures in order to address, in particular, *aliasing*.
- An axiomatic definition of subsethood for continuous intervals is introduced.
- A new set-theoretic similarity measure is proposed for continuous intervals which minimizes *aliasing* and satisfies both traditional and new axioms of a similarity measure.
- A new dissimilarity measure, which satisfies the properties of a metric, is derived from the new similarity measure, enabling its direct use in classification and clustering applications.
- A generalized framework, retaining discontinuity information (thus minimizing information loss), is developed to assess similarity between discontinuous intervals.
- Three binary relations of Allen's theory [21] are extended and integrated with the generalized framework to enable it to be more efficient in assessing similarity among discontinuous intervals.
- A vertical-slice based and also an α-cut based extensions of the new interval similarity measure are proposed to assess similarity for T1 fuzzy sets.
- A new FM based on the new similarity measure is developed which is shown to avoid the limitations of inter-source agreement (based on similarity), and in combination with the FI, it also provides intuitive aggregation results.

This research has contributed to four peer-reviewed conference papers and three journal papers where one is accepted for publication and other two papers are pending submission. These publications are listed below.

- Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. (2019). Aggregation of Crowd-Sourced Intervals Using Fuzzy Integral and Bidirectional Subsethood Based Fuzzy Measure. In submission preparation for IEEE Transactions on Fuzzy Systems.

- Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. (2019). Efficient Comparison of Discontinuous Intervals using a Combination of Bidirectional Subsethood Based Similarity Measure and Extended Allen's Theory. In submission preparation for IEEE Transactions on Fuzzy Systems.

- Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. (2019). A Similarity Measure Based on Bidirectional Subsethood for Intervals. Accepted for publication in IEEE Transactions on Fuzzy Systems, 2019 (doi: 10.1109/TFUZZ.2019.2945249).

- Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. (2019). Measuring Similarity Between Discontinuous Intervals - Challenges and Solutions. In International Conference on Fuzzy Systems (Fuzz-IEEE 2019), New Orleans, Louisiana, USA, 2019.

- Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. (2018). A Bidirectional Subsethood Based Similarity Measure for Fuzzy Sets. In IEEE World Congress on Computational Intelligence (WCCI 2018), Rio de Janeiro, Brazil, 2018.

- Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; Anderson, Derek T.; and Aickelin, Uwe (2017). Novel Similarity Measure for Interval-Valued Data Based on Overlapping Ratio. In International Conference on Fuzzy Systems (Fuzz-IEEE 2017), Naples, Italy, 2017.

- Pekaslan, Direnc; Kabir, Shaily, Garibaldi, Jonathan M.; and Wagner, Christian (2017). Determining Firing Strengths Through A Novel Similarity Measure to Enhance Uncertainty Handling in Non-Singleton Fuzzy Logic Systems. In International Joint Conference on Computational Intelligence (IJCCI 2017), Funchal, Madeira-Portugal, 2017.

1.4 Organization of the Thesis

The rest of this thesis is organized as follows.

Chapter 2 presents background on uncertainty modelling with intervals and T1 fuzzy sets along with a discussion on their different properties, such as continuity/discontinuity, normality/non-normality and different arithmetic and set-theoretic operations on them. After this, a brief review on different triangular norms and conorms is given. Following this, an overview on similarity measures for intervals and T1 fuzzy sets is provided, specifically focusing on set-theoretic similarity measures where their potential limitations are highlighted. A brief overview of subsethood for intervals and T1 fuzzy set is discussed followed by a short review on the relation between fuzzy subsethood and fuzzy similarity measures. Lastly, a discussion on multi-source data aggregation and different aggregation operators is given, specifically focusing on the aggregation by FIs in respect to a FM. Here, different FMs are discussed and the potential limitation of data-driven FMs is explored.

Chapter 3 identifies, articulates, and addresses an inherent limitation of popular set-theoretic similarity measures on continuous intervals – their tendency to be subject to *aliasing* – where they return the same similarity value for very different sets of intervals with different cardinalities. To address *aliasing*, new axioms are introduced to establish desirable properties for *robust* similarity measures. Subsequently, a new set-theoretic similarity measure is developed for continuous intervals using the concept

of bidirectional subsethood which satisfies both traditional and new axioms of a similarity measure. Further, the definition of subsethood for continuous intervals is introduced, followed by proofs of its mathematical properties as a part of developing the new similarity measure. In addition, a new dissimilarity measure, which satisfies the properties of a metric, is derived from the new similarity measure, enabling its direct application in classification and clustering problems.

Chapter 4 develops a generalized framework to assess the degree of similarity between discontinuous intervals together with a discussion of its essential properties. The new generalized framework utilizes the new similarity measure from Chapter 3 to measure similarity between the pairs of continuous subintervals within the discontinuous intervals. Moreover, a refinement of this framework is provided where it is combined with the extended Allen's theory [21] to make the latter more efficient.

Chapter 5 extends the new similarity measure of continuous intervals from Chapter 3 to assess similarity for T1 fuzzy sets, initially with a vertical-slice based extension and later with an α-cut based extension as for certain cases of T1 fuzzy sets the vertical-slice based extended measure does not produce expected outcomes. A further refinement of the α-cut based extended measure is developed to deal with non-normal and non-convex T1 fuzzy sets.

Chapter 6 identifies that existing agreement-based FMs are impacted by the limitations of intersection operation and *aliasing* of the common similarity measures, thus generates same agreement and worth for different sets of sources. Hence, Chapter 6 develops methods to extract a new FM based on the concept of agreement. In this case, the new FM is built on the new interval similarity measure from Chapter 3 to avoid the limitation of generating the same worth information for very different sets of sources. The new FM is then used in connection with an FI to aggregate multi-source data, particularly interval-valued data. Importantly, the new FM is designed in such

way that it can carry out aggregation under the situation of total disagreement among sources. The new FM is evaluated against the existing FMs in [22], [23], and [24].

Lastly, Chapter 7 summarizes and concludes the thesis by highlighting the main contributions and limitations of the research work. The scope for future work is also discussed, addressing the need for methods to deal with similarity assessment for higher level of uncertain data-types.

Chapter 2

Research Background

2.1 Introduction

This chapter discusses pertinent concepts and research works from the existing literature, which underpin this thesis. Section 2.2 initially provides a brief discussion on efficient modelling of uncertain, imprecise, and vague data. After that, Section 2.3 presents a review on intervals together with some basic arithmetic and logical operations often used on them. Section 2.4 provides a review on fuzzy sets and different fuzzy set operations. Following these revisions, Section 2.5 presents a detailed discussion on measuring similarity between intervals and also between fuzzy sets, including different measures/approaches to calculate their similarity. After this, Section 2.6 defines subsethood for crisp sets and T1 fuzzy sets. Finally, Section 2.7 presents a brief overview on different aggregation operators, particularly focusing on the FI defined in respect to an FM. Here, review FMs which are based on the concept of modelling agreement among the data (evidence) arising from multiple sources. Pages *xxiv* and *xxv* provide tables of acronyms and abbreviations for the ease of the readers as various notations and mathematical functions are used throughout this thesis.

2.2 Modelling of Uncertainty

The presence of uncertainty in human decision making is common. Uncertainty is undesirable but it is hard to avoid in real-world problems. Various methods and strategies are designed to minimize uncertainty and ease decision making. Generally, uncertainty occurs when there is a lack of information to decide whether a statement is true or false [25]. Two terms – vagueness and imprecision – are closely related to uncertainty [3] where the first term expresses the lack of exactness of a statement, and the latter term expresses the fuzziness of a statement. Consider a simple example of uncertainty that can arise from a statement 'the food is hot'. Here , one can easily be in a state of uncertainty as the food can either be 'spicy' or 'warm'.

Generally, the quality of human decision making under uncertainty largely depends on how successfully uncertainty is captured and processed. There are many approaches that attempt to capture uncertainty in data. Well-known approaches are probability distributions [26], possibility distributions [1], belief functions [27], intervals [5][28], fuzzy sets [4][29]. This thesis is particularly focusing on intervals and T1 fuzzy sets as they can capture uncertainty at the fundamental level and be extended subsequently to higher-order fuzzy sets.

The simplest, yet efficient approach to capture uncertainty in data is based on intervals where an interval is bounded by the left and right endpoints [8]. The size of an interval plays a critical role to represent the degree of uncertainty which exists in data where a narrow interval represents more certainty, and a wider one shows more uncertainty [30]. Beyond intervals, fuzzy sets have already proven to be useful and powerful model of representing uncertainty. While T1 fuzzy sets [2] have been the most common form of modelling uncertainty, recent years have shown a significant increase in research toward more complex forms of fuzzy sets, in particular, T2 fuzzy sets [4][31].

As this thesis focuses on intervals and T1 fuzzy sets in capturing uncertainty in data, next two sections give an overview on intervals, some basic operations on them, T1 fuzzy sets and different fuzzy set operations.

2.3 Intervals

Interval-valued data have gained much interest for the modelling of uncertainty and vagueness, particularly in the modelling of survey data [5], the representation of symbolic data [6], and the capture of natural language expressions [7], as they offer simple and efficient representation of uncertain, vague, and imprecise information. Generally, intervals are continuous and closed where they are bounded by the left and right endpoints. However, in many real-world applications, discontinuous intervals are utilized [32][33]. The following two subsections illustrate these two categories of intervals (i.e., continuous and discontinuous).

2.3.1 Continuous Intervals

A continuous interval is a set of real numbers characterized by left and right endpoints [8]. Mathematically, a continuous interval $\overline{a}$ is represented as [34]

$$\overline{a} = [a^-, a^+] \text{ with } a^- \leq a^+. \tag{2.1}$$

A continuous interval $\overline{a}$ is often known as degenerate if $a^- = a^+$; in such case, $\overline{a}$ contains a single real number a [35]. The cardinality, or equivalently, the size or width of a continuous interval $\overline{a}$ is defined as $|\overline{a}| = |a^+ - a^-|$ [36]. It is noted that $\overline{a}$ is also known as a convex [33] or closed interval [16][37]. Figure 2.1 presents a continuous and closed interval $\overline{a} = [2, 8]$.

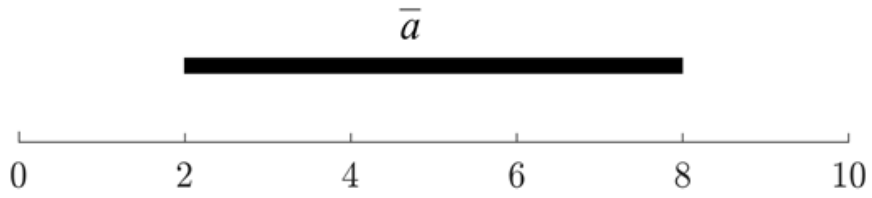


Fig. 2.1 A continuous and closed interval, $\overline{a} = [2, 8]$.

Two continuous intervals $\overline{a} = [a^-, a^+]$ and $\overline{b} = [b^-, b^+]$ are said to be equal if they represent the same set, i.e., $\overline{a} = \overline{b}$, if $a^- = b^-$ and $a^+ = b^+$ [35]. Again, two continuous intervals $\overline{a} = [a^-, a^+]$ and $\overline{b} = [b^-, b^+]$ are said to be intersected if $a^- \leq b^+ \wedge b^- \leq a^+$ or $b^- \leq a^+ \wedge a^- \leq b^+$ [37].

2.3.1.1 Basic Operations on Continuous Intervals

The basic arithmetic and set-theoretic operations on continuous intervals are discussed below.

(1) Interval Arithmetic Operations

The basic interval arithmetic operations for continuous intervals $\overline{a} = [a^-, a^+]$ and $\overline{b} = [b^-, b^+]$ are as follows [34] –

- *Addition*:

$$\overline{a} + \overline{b} = [a^- + b^-, a^+ + b^+];$$

- *Subtraction*:

$$\overline{a} + \overline{b} = [a^- - b^+, a^+ - b^-];$$

- *Multiplication*:

$$\overline{a} \times \overline{b} = [\min(a^- b^-, a^- b^+, a^+ b^-, a^+ b^+), \max(a^- b^-, a^- b^+, a^+ b^-, a^+ b^+)];$$

- *Division*:

$$\overline{a} \div \overline{b} = [a^-, a^+] \times [1/b^-, 1/b^+], \text{ where } 0 \notin [b^-, b^+].$$

(2) Set-Theoretic Operations

Two basic set-theoretic operations on continuous intervals $\overline{a} = [a^-, a^+]$ and $\overline{b} = [b^-, b^+]$ (where $\overline{a} \cap \overline{b} \neq \emptyset$) are as follows [38][39] –

- *Union*:

$$\overline{a} \cup \overline{b} = [\min(a^-, b^-), \max(a^+, b^+)];$$

- *Intersection*:

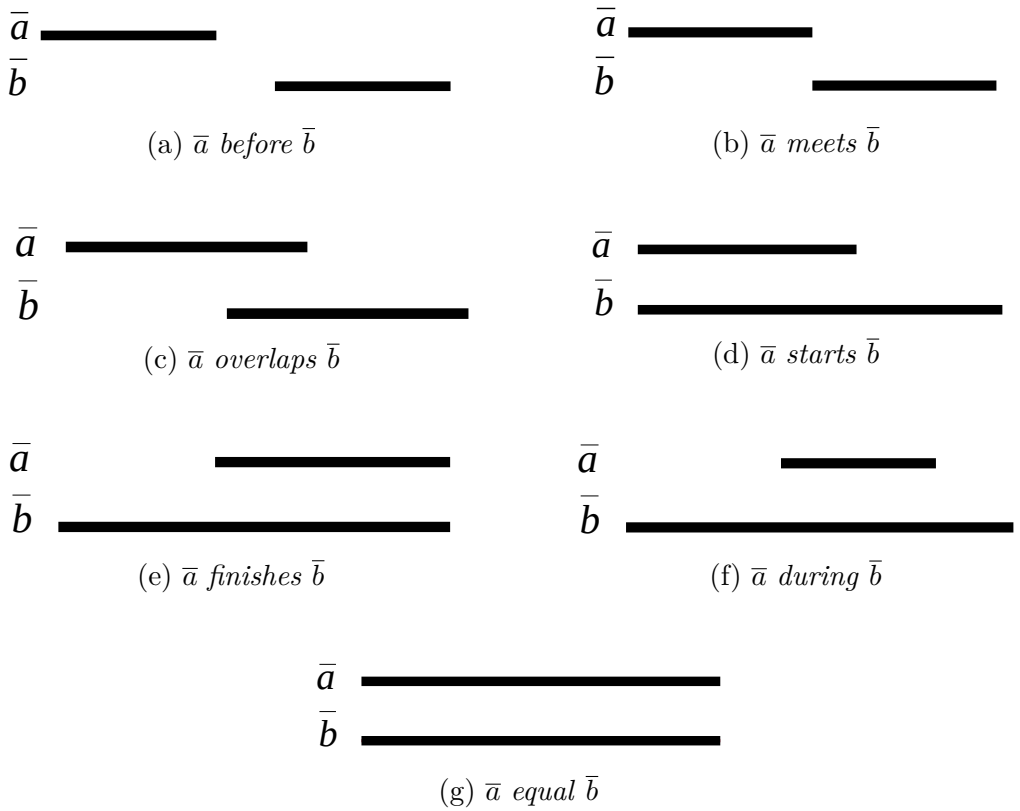
$$\overline{a} \cap \overline{b} = [\max(a^-, b^-), \min(a^+, b^+)].$$

2.3.1.2 Allen's Theory on Continuous Intervals

Allen [21] introduced a set of thirteen relations between continuous intervals for interval-based temporal reasoning. They are commonly known as Allen's theory [40] or Allen's interval algebra [41]. In Allen's theory, the first seven relations are defined as basic relations, and the remaining six are defined as inverse relations [42]. The seven basic relations are – *before*, *meets*, *overlaps*, *starts*, *finishes*, *during*, and *equal* [43].

Figure 2.2 pictorially shows the basic relations between two continuous intervals $\overline{a}$ and $\overline{b}$. They are described in detail in the following –

(1) $\overline{a}$ *before* $\overline{b}$ — the end of $\overline{a}$ is before the start of $\overline{b}$;



(a) $\overline{a}$ *before* $\overline{b}$

(b) $\overline{a}$ *meets* $\overline{b}$

(c) $\overline{a}$ *overlaps* $\overline{b}$

(d) $\overline{a}$ *starts* $\overline{b}$

(e) $\overline{a}$ *finishes* $\overline{b}$

(f) $\overline{a}$ *during* $\overline{b}$

(g) $\overline{a}$ *equal* $\overline{b}$

Fig. 2.2 Seven basic relations of Allen's theory on continuous intervals $\overline{a}$ and $\overline{b}$.

(2) $\overline{a}$ *meets* $\overline{b}$ — the end of $\overline{a}$ is same as the start of $\overline{b}$;

(3) $\overline{a}$ *overlaps* $\overline{b}$ — the start of $\overline{a}$ is before the start of $\overline{b}$ and the end of $\overline{a}$ is between the start and the end of $\overline{b}$;

(4) $\overline{a}$ *starts* $\overline{b}$ — both $\overline{a}$ and $\overline{b}$ start together but $\overline{a}$ ends first;

(5) $\overline{a}$ *finishes* $\overline{b}$ — $\overline{a}$ starts after $\overline{b}$ but both end together;

(6) $\overline{a}$ *during* $\overline{b}$ — $\overline{a}$ starts after $\overline{a}$ but ends before $\overline{b}$;

(7) $\overline{a}$ *equal* $\overline{b}$ — both $\overline{a}$ and $\overline{b}$ start and end together.

The six other relations are the inverse of (1) to (6) basic relations. The inverse relations are not discussed in detail as this thesis only focuses on Allen's basic relations; for further details please see [21][44].

Remark 2.1:

Allen's binary relations on continuous intervals [21] are utilized in Chapter 4 (in an extended form) in order to compare pairs of discontinuous intervals.

2.3.2 Discontinuous Intervals

A discontinuous interval consists of a number of continuous subintervals [45]. Mathematically, a discontinuous interval $\overline{\overline{a}}$ is defined as follows

$$\overline{\overline{a}} = \bigcup_{i=1}^{m} \overline{a}_i = \bigcup_{i=1}^{m} [a_i^-, a_i^+], \tag{2.2}$$

where m is the number of continuous subintervals within $\overline{\overline{a}}$ such that $\overline{a}_1 < ... < \overline{a}_i < ... < \overline{a}_m$ [33]. $\overline{a}_i$ is the ith subinterval with $a_i^- \leq a_i^+$. $\overline{\overline{a}}$ can alternatively be presented as $\langle [a_1^-, a_1^+], ..., [a_i^-, a_i^+], ..., [a_m^-, a_m^+] \rangle$ [46]. It is noted that discontinuous intervals are also known as non-convex intervals [40]. Figure 2.3 presents a discontinuous intervals $\overline{\overline{a}}$ with two continuous subintervals $\overline{a}_1 = [2, 4]$ and $\overline{a}_2 = [6, 8]$.

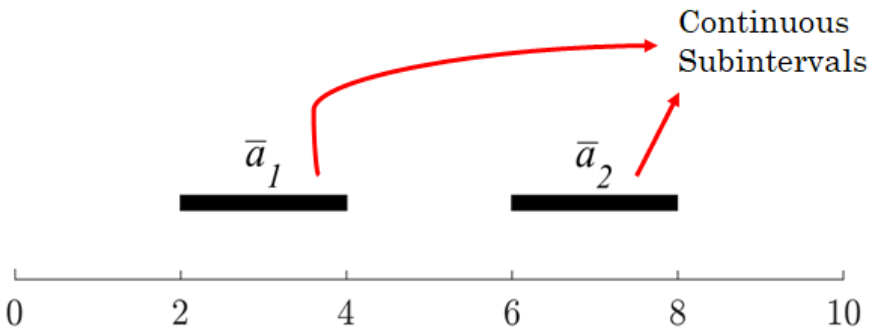
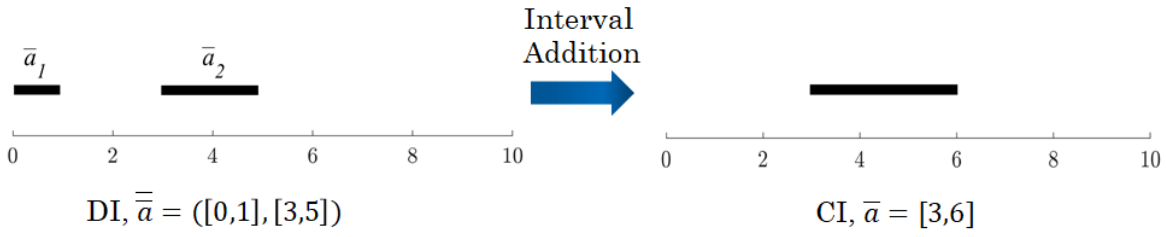


Fig. 2.3 A discontinuous interval, $\overline{\overline{a}} = \{[2, 4], [6, 8]\}$.

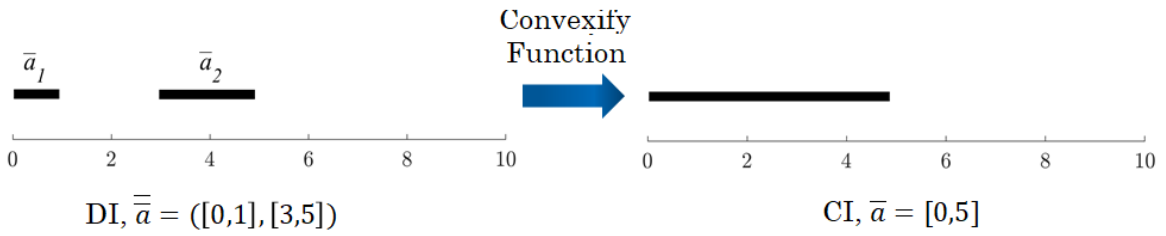
2.3.2.1 Conversion of Discontinuous Interval into Continuous Form

Unlike continuous intervals, there is no straightforward method to perform reasoning with discontinuous intervals. Generally, they are converted or transformed into a continuous form before applying any operation on them [32]. There are two common approaches which can be applied to transform the discontinuous intervals to their continuous forms [19]. These two approaches are discussed below:

- Interval Addition: If $\overline{a} = [a^-, a^+]$ and $\overline{b} = [b^-, b^+]$ are two bounded, non-empty continuous intervals, then their addition, $\overline{a} + \overline{b} = [a^- + b^-, a^+ + b^+]$ is also a bounded, non-empty continuous interval [34].
- The 'Convexify' function: It takes two continuous intervals $\overline{a}$ and $\overline{b}$ as inputs and returns the smallest continuous interval that covers both $\overline{a}$ and $\overline{b}$ [44].



(a) Interval addition



(b) Convexify function

Fig. 2.4 Conversion of a discontinuous interval into a continuous interval.

Example 2.1:

Consider a discontinuous interval $\overline{\overline{a}} = \{[0, 1], [3, 5]\}$. The interval addition [34] transforms the subintervals into the continuous interval, $\overline{a} = [3, 6]$, whereas the 'Convexify' function [44] converts them into the continuous interval, $\overline{a} = [0, 5]$. Figure 2.4 (a) and (b) present the conversion of $\overline{\overline{a}}$ to its continuous form $\overline{a}$ using the interval addition and convexify function respectively.

In the following section, definition and properties of T1 fuzzy sets are discussed.

2.4 Fuzzy Sets

Fuzzy sets are a popular and efficient method to capture and model uncertain and imprecise data [5]. They were first introduced by Zadeh [2] as an extension of classical or crisp sets to model data uncertainty. In a crisp set, the membership of an element in a set is described as either belonging (with a membership of 1) or not belonging (with a membership of 0) to the set. Contrarily, a fuzzy set allows its elements to have some degree of membership between 0 and 1 to represent their belongingness to the set. Formally, a fuzzy set A in the universe of discourse X is characterized by its membership function μ_A that takes on values in the interval between 0 and 1. It is worth mentioning that the membership grades of elements within the fuzzy set are not probabilities as probabilities on a finite set must add to the value of 1 while there is no such restriction for the membership grades of a fuzzy set. To illustrate their difference, consider a scenario from [47] where *Autri* is in a house having two adjacent rooms – one is kitchen and the other is dining room. In respect to crisp sets, the status of *Autri* is either 'in the kitchen' or 'not in the kitchen'. However, if *Autri* stands at the doorway, she can be considered as 'partially in the kitchen' and quantifying this state

produces a fuzzy set membership. With her little toe only in the dining room, it might be said that *Autri* is 0.99 'in the kitchen'.

Fuzzy sets were originally designed as T1 fuzzy sets, and later extended to higher order fuzzy sets (e.g., T2 fuzzy sets [48][49]) to express higher levels of fuzziness which cannot be captured by T1 fuzzy sets [50]. Within a T1 fuzzy set, the membership of elements is represented by a single value in the range 0 and 1 [51]. In the following two subsections, T1 fuzzy sets and basic set-theoretic operations on them are discussed.

2.4.1 Type-1 Fuzzy Sets

Formally, a T1 fuzzy set, A in the universe of discourse X is characterized by a membership function $\mu_A(x)$ that determines the degree of membership of each element $x \in X$ and is defined as

$$A = \{(x, \mu_A(x)) | x \in X\}, \tag{2.3}$$

where $\mu_A(x) \in [0, 1]$ [2].[1] In a continuous universe of discourse X, the T1 fuzzy set is often presented as [4]

$$A = \int_{x \in X} \mu_A(x)/x, \tag{2.4}$$

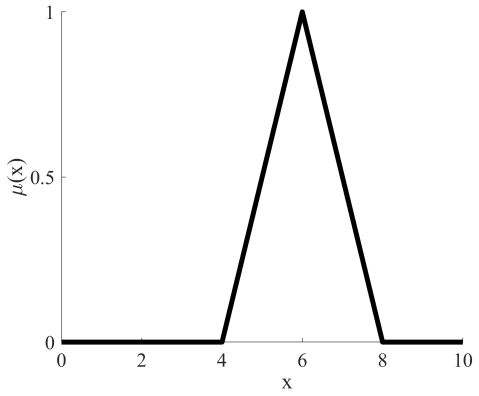
whereas, in a discrete universe of discourse X, it is defined as [4]

$$A = \sum_{x \in X} \mu_A(x_i)/x_i, \tag{2.5}$$

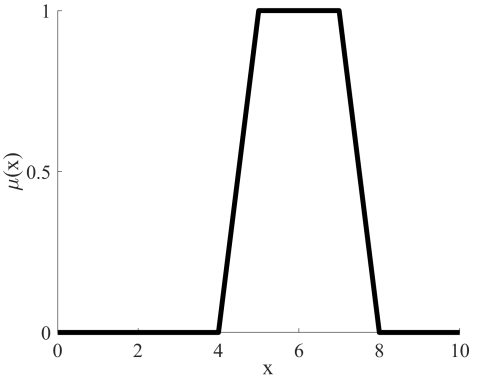
where $\sum$ is the collection of all points $x \in X$ with associated membership $\mu_A(x)$.

As mentioned, the membership function of the fuzzy set defines how each x in X is mapped to a membership value (or degree of membership) between 0 and 1. The most commonly used membership functions for the T1 fuzzy set are the triangular,

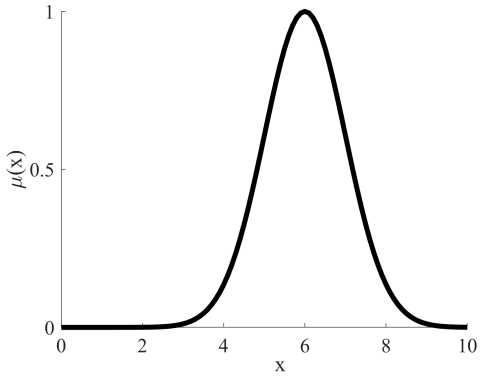
[1] For alternate mathematical notations of fuzzy sets, please see [52–54].



(a) Triangular membership function



(b) Trapezoidal membership function



(c) Gaussian membership function

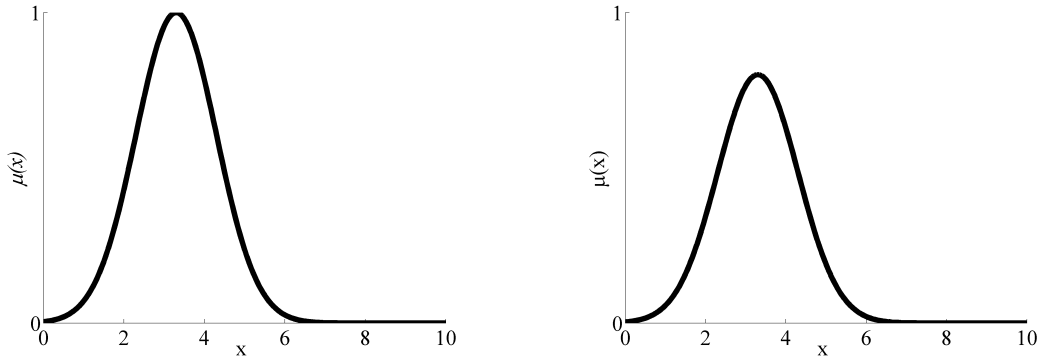
Fig. 2.5 Three different membership functions for T1 fuzzy sets.

trapezoidal, and Gaussian membership functions. Figure 2.5 graphically shows these membership functions.

Definition 2.1: The height of a T1 fuzzy set is the largest membership grade associated with any element in the set [2]. It is defined as

$$h(A) = \sup_{x \in X} \mu_A(x). \tag{2.6}$$

Definition 2.2: A T1 fuzzy set is normal when at least one of its elements has the maximum membership grade, that is, $h(A) = 1$ [49]. It is non-normal when $h(A) < 1$.



(a) Normal fuzzy set (b) Non-normal fuzzy set

Fig. 2.6 Normal and non-normal T1 fuzzy sets.

Figure 2.6 shows an example of T1 fuzzy sets with normal and non-normal properties. There are two common representations of T1 fuzzy sets. One representation is based on their vertical-slices and the other one in based on their α-cuts or horizontal-slices [20]. The vertical-slice representation involves drawing a vertical line at a given x to get its associated membership, whereas the α-cut representation is doing the opposite, drawing a horizontal line at a given membership value (denoted α) to get all xs having the membership of α or higher. Definition 2.3 explains the α-cuts of a T1 fuzzy set. Figure 2.7 shows an example of these two representations for a Gaussian T1 fuzzy set.

Definition 2.3: An α-cut of a T1 fuzzy set A is a crisp set $\overline{A}_\alpha$ that comprises all elements whose membership grade within A is greater than or equal to α [31]. Mathematically, it is written as

$$\overline{A}_\alpha = \{x | \mu_A(x) \geq \alpha, \alpha \in [0,1]\}. \tag{2.7}$$

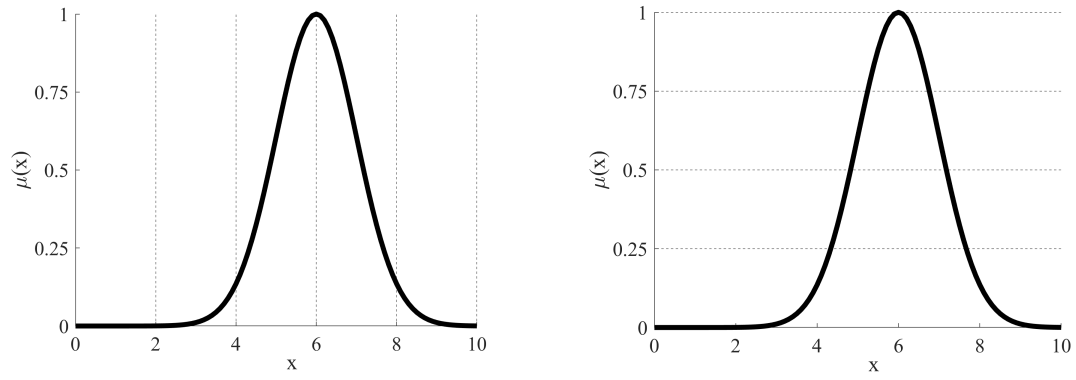
In particular, the T1 fuzzy set A can be represented by its α-cuts as

$$A = \int_0^1 \alpha \overline{A}_\alpha, \tag{2.8}$$

where $\int_0^1$ is the union of all A_α within [0,1] and $\alpha\overline{A}_\alpha$ shows the mapping of pairs α and $\overline{A}_\alpha$ [31]. Contrarily, within a discrete range of α-cuts, it can be rewritten as

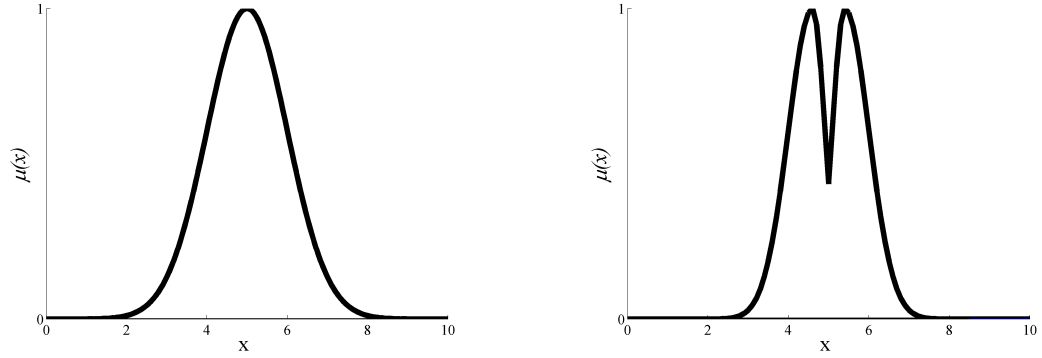
$$A = \sum_{i=0}^{N} \alpha_i \overline{A}_{\alpha_i}, \tag{2.9}$$

where N denotes the total number of discretizations on the membership axis (total number of α-cuts).



(a) Vertical slices (b) Horizontal slices (α-cuts)

Fig. 2.7 Two different presentations of T1 fuzzy set.



(a) Convex fuzzy set (b) Non-convex fuzzy set

Fig. 2.8 Convex and non-convex T1 fuzzy sets.

Definition 2.4: A T1 fuzzy set A is convex if and only if all α-cuts of A are convex subsets of X [31]. It is defined as -

$$\mu_A\left(\lambda x_1 + (1-\lambda)x_2\right) \geq \min[\mu_A(x_1), \mu_A(x_2)], \tag{2.10}$$

for all $x_1, x_2 \in X$ and $\lambda \in [0, 1]$. Otherwise, A is a non-convex fuzzy set.

Figure 2.8 presents an example of T1 fuzzy sets having convex and non-convex properties.

Definition 2.5: A T1 fuzzy set A is defined as a fuzzy number if [55]:

- A is normal;
- A_α is a closed interval for all $\alpha \in [0, 1]$;
- The support of A, $\{x|A(x) > 0\}$, is bounded.

As all α-cuts are closed intervals, thus every fuzzy number is also a convex fuzzy set.

2.4.2 Set-Theoretic Operations on Type-1 Fuzzy Sets

Let μ_A and μ_B be the membership functions for the T1 fuzzy sets A and B. The following set-theoretic operations can be performed on these sets [2].

1. Fuzzy Union – the union of $A, B \in X$ is

$$A \cup B = \{x, S\left(\mu_A(x), \mu_B(x)\right) | x \in X\} \tag{2.11}$$

where S is any *t-conorm* [56]. For any value of x, Eq. (2.11) can be written as

$$\mu_{(A \cup B)}(x) = S\left(\mu_A(x), \mu_B(x)\right)$$

2. Fuzzy Intersection – the intersection of $A, B \in X$ is

$$A \cap B = \{x, \bigstar (\mu_A(x), \mu_B(x)) \,|x \in X\} \tag{2.12}$$

where $\bigstar$ is any *t-norm* [57]. For any value of x, Eq. (2.12) can be written as

$$\mu_{(A \cap B)}(x) = \bigstar (\mu_A(x), \mu_B(x))$$

3. Fuzzy Complement – the complement of $A \in X$ is

$$A^c = \{(x, 1 - \mu_A(x))|x \in X\} \tag{2.13}$$

For any value of x, Eq. (2.13) can be written as

$$\mu_{A^c}(x) = 1 - \mu_A(x)$$

For details of *t-norm* ($\bigstar$) and *t-conorm* (S), please see [58] and [59].

In the next section, different existing similarity measures in respect to intervals and T1 fuzzy sets are discussed.

2.5 Similarity Measures for Uncertain Data Types

A similarity measure is a real-valued function that determines how alike two objects are. Similarity measures are widely utilized in a range of applications including decision making [60], data aggregation [23], approximate reasoning [61], machine learning [62],

and so on. Mathematically, for two sets a and b, it is expressed as [10]

$$S(a,b) \to \mathbb{R}. \tag{2.14}$$

Similarity is commonly represented by a non-negative real number ($\mathbb{R}$), often between 0 and 1, where 0 means objects are not similar at all, and 1 means they are identical. Similarity measures are generally symmetrical, though, for certain objects, similarity can be better modelled by unidirectional or asymmetric functions [12]. The four common properties of a similarity measure for sets a, b, and c are as follows [63] –

[A1] Boundedness: $0 \leq S(a,b) \leq 1$;

[A2] Symmetry: $S(a,b) = S(b,a)$;

[A3] Reflexivity: $S(a,b) = 1 \iff a = b$;

[A4] Transitivity: If $a \subseteq b \subseteq c$ then $S(a,b) \geq S(a,c)$.

It is worth mentioning that the above properties are mirrored in (and in many cases arise from) the axiomatic definitions associated with a wide variety of established similarity measures (e.g., please see [64–66]). In addition, similarity measures can often be transformed to capture distance using various functions [67], showing dissimilarity among objects (e.g., color images as in [68]). Generally, a dissimilarity measure is a real-valued function that determines how far apart two objects are. Like before, for two sets a and b, it is expressed as $D(a,b) \to \mathbb{R}^+$ [63]. A dissimilarity measure is defined as a metric when it satisfies the following properties for sets a, b, and c [69][70] –

[B1] Non-negativity: $D(a,b) \geq 0$;

[B2] Symmetry: $D(a,b) = D(b,a)$;

[B3] Identity: $D(a,b) = 0 \iff a = b$;

[B4] Triangle Inequality: $D(a,c) \leq D(a,b) + D(b,c)$.

It is worth noting that a dissimilarity measure is not always a distance metric, but is often referred to as a semi-metric if it satisfies all of the preceding properties except the triangle inequality [71]. A dissimilarity measure is often within the range $[0,1]$ where 0 is used for identical objects and 1 for dissimilar objects [63] — when it is derived from a similarity measure bounded by $[0,1]$.

Various similarity measures have come into use to capture similarity among crisp sets and fuzzy sets, though each of the measures has its own strengths and weaknesses. It is worth noting that most of the similarity measures introduced for crisp sets are also applicable for continuous intervals [16]. Generally, similarity measures are mainly classified as set-theoretic measures [72] and distance-based measures [73]. For fuzzy sets, two additional categories of similarity measures are logic-based measures and fuzzy-valued measures [10]. Among these categories, set-theoretic similarity measures are well-known and widely used [11]. On the other hand, distance-based similarity measures are very popular for sets/fuzzy sets [11] but often facing limitation in the comparison of some objects having qualitative features (e.g., comparing a pear and an apple with respect to taste), not having quantitative features [12].

In the next two subsections, existing set-theoretic similarity measures in respect to intervals and T1 fuzzy sets are discussed more.

2.5.1 Set-Theoretic Similarity Measures for Intervals

Similarity assessment between two crisp sets using set-theoretic measures typically involves their common and distinctive features. The set-theoretic measures typically utilize basic set operations including union, intersection, and cardinality[2] to compare the sets. It is worth mentioning that Tversky's parameterized ratio model of similarity

[2]Note that the cardinality of a set a is defined as the number of elements within a [74]. In this thesis, the terms 'cardinality' and 'size' of a set are interchangeably used.

[12] generalizes many classical set-theoretic similarity measures where most of them are equally usable for continuous intervals. These similarity measures are further extended to assess similarity between fuzzy sets. The ratio model expresses similarity between two sets a and b as a ratio of the common and distinct features of a and b

$$S_T(a,b) = \frac{f(a \cap b)}{f(a \cap b) + \alpha f(a-b) + \beta f(b-a)}, \tag{2.15}$$

where $f(a \cap b)$ presents the common features, $f(a-b)$ is the features that a has but b does not, and vice versa [12]. The factors α and β are non-negative (≥ 0). $S_T(a,b)$ turns into the Jaccard measure [13] when $\alpha = \beta = 1$, whereas $S_T(a,b)$ becomes the Dice measure [14] with $\alpha = \beta = 0.5$. For $\alpha = 1$ and $\beta = 0$, $S_T(a,b)$ behaves like the inclusion/subsethood for a (the degree to which a is a subset of b) [10].

The most widely used set-theoretic similarity measure is the Jaccard measure [13] which satisfies all of the above properties [A1] - [A4]. Generally, the Jaccard similarity of two crisp sets a and b is defined as the ratio of the cardinality of their intersection and the cardinality of their union,

$$S_J(a,b) = \frac{|a \cap b|}{|a \cup b|}. \tag{2.16}$$

Using the crisp set difference operation [74], Eq. (2.16) can be written as

$$S_J(a,b) = \frac{|a \cap b|}{|a \cap b| + |a \backslash b| + |b \backslash a|}, \tag{2.17}$$

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. Equation (2.17) can also be derived from Tversky's parameterized ratio model of similarity [12] by setting the α and β to 1 and letting f be a cardinality

function. Beyond crisp sets, the Jaccard similarity measure is used to estimate the similarity for intervals or sets of intervals, such as employed for example in data fusion [23][24] and that of fuzzy sets [75].

As mentioned earlier, a continuous/closed interval $\overline{a}$ is a set of real numbers bounded by two endpoints a^- and a^+ with $a^- \leq a^+$. For comparing two intervals $\overline{a}$ and $\overline{b}$, the Jaccard similarity measure is expressed as

$$S_J\left(\overline{a}, \overline{b}\right) = \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cup \overline{b}\right|}, \tag{2.18}$$

where $|\overline{a}| \neq 0$ and $\left|\overline{b}\right| \neq 0$. $\left|\overline{a} \cap \overline{b}\right|$ 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 both $\overline{a}$ and $\overline{b}$. Hence, $S_J\left(\overline{a}, \overline{b}\right) = 1$ when $\overline{a}$ and $\overline{b}$ are completely overlapping and $S_J\left(\overline{a}, \overline{b}\right) = 0$ when they are not overlapping at all. Similar to Eq. (2.17), Eq. (2.18) can be rewritten as

$$S_J\left(\overline{a}, \overline{b}\right) = \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{a} \backslash \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|}, \tag{2.19}$$

where $\left|\overline{a} \backslash \overline{b}\right|$ is the size of the interval segment of $\overline{a}$ that is not overlapping with $\overline{b}$ and $\left|\overline{b} \backslash \overline{a}\right|$ is the size of the interval segment of $\overline{b}$ that is not overlapping with $\overline{a}$.

The Sørensen-Dice similarity measure—henceforth referred to as Dice measure for convenience [14] is closely related to the Jaccard similarity measure. To assess the similarity between two sets, it considers the ratio of the cardinality of their intersection and the average of their cardinality. alike the Jaccard similarity, it produces outputs in $[0, 1]$. Specifically, for two crisp sets a and b, the Dice similarity is expressed as -

$$S_D\left(a, b\right) = \frac{|a \cap b|}{\frac{1}{2}(|a| + |b|)}, \tag{2.20}$$

where $|a|$ is the cardinality of the set a. Eq. (2.20) can be rewritten by applying the crisp set difference operation [74]

$$S_D(a,b) = \frac{|a \cap b|}{|a \cap b| + \frac{1}{2}(|a \backslash b| + |b \backslash a|)}. \tag{2.21}$$

Note that (2.21) can also be obtained from Tversky's ratio model [12] when both α and β are 0.5. The alternative expressions of Jaccard at Eq. (2.17) and Dice at Eq. (2.21) show clearly that the averaging operation in the denominator of Eq. (2.21) results in the Dice similarity always being equal to—when sets are identical—or larger than the Jaccard similarity.

The Dice similarity is used along with the Jaccard similarity for intervals in [23] [24]. The Dice similarity for two intervals $\overline{a}$ and $\overline{b}$ can be expressed as -

$$S_D\left(\overline{a}, \overline{b}\right) = \frac{\left|\overline{a} \cap \overline{b}\right|}{\frac{1}{2}(\left|\overline{a}\right| + \left|\overline{b}\right|)}, \tag{2.22}$$

where $|\overline{a}| \neq 0$ and $\left|\overline{b}\right| \neq 0$. By following Eq. (2.19), Eq. (2.22) can be rewritten as follows.

$$S_D\left(\overline{a}, \overline{b}\right) = \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \frac{1}{2}(\left|\overline{a} \backslash \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|)}. \tag{2.23}$$

Other Set-theoretic Similarity Measures

Besides Jaccard and Dice measures, other set-theoretic similarity measures such as the Szymkiewicz-Simpson coefficient [76], the Otsuka coefficient [77], the Sokal-Sneath Coefficient [78], the Simple-Matching coefficient [79] are also common and used to assess similarity for crisp sets as well as for intervals.

It is worth mentioning that most of these set-theoretic similarity measures including Jaccard and Dice measures often suffer from a so-far not discussed pitfall, best described

as *aliasing*. It occurs where the same similarity value is generated by a similarity measure for very different pairs of intervals with dissimilar cardinalities or sizes, at worst it leads to incorrect inference in real-world applications. The reason why these similarity measures are subject to this *aliasing* is discussed in detail in Chapter 3 and also to mitigate this issue, a new set-theoretic similarity measure has been introduced for continuous intervals in the same chapter. It is also important to mention that the classical set-theoretic similarity measures are not directly applicable for discontinuous intervals [19]. Generally, the current approach is to convert the discontinuous intervals into their continuous forms and then a set-theoretic similarity measure compatible to the continuous intervals is applied to assess their similarity, details are provided in Chapter 4. To deal with this, a generalized framework, retaining discontinuity property, is proposed in the same Chapter 4 to assess similarity between discontinuous intervals.

2.5.2 Set-Theoretic Similarity Measures for Type-1 Fuzzy Sets

In respect to T1 fuzzy sets, a similarity measure captures how similar two fuzzy sets are regarding the degree of membership of their elements. Zadeh [80] first introduced the notion of similarity relations for T1 fuzzy sets. Typically, if two fuzzy sets are identical, their similarity is 1, whereas if they are fully disjoint, their similarity is 0. Comparison between fuzzy sets has been successfully used in a wide range of applications, including linguistic approximation [81], pattern recognition [82], computing with words [83], and clustering [84][85]. From an axiomatic point of view, different axiomatic definitions of similarity measures for fuzzy sets are available in the literature [86] where most of these definitions depend on the contexts in which they are constructed. Formally, common properties of a fuzzy similarity measure for fuzzy sets A, B, and C are the followings [63] -

[C1] Symmetry: $S(A, B) = S(B, A)$;

[C2] Reflexivity: $S(A,B) = 1 \iff A = B$;

[C3] Transitivity: If $A \subset B \subset C$ then $S(A,B) \geq S(A,C)$;

[C4] Overlapping: If $A \cap B \neq \emptyset$ then $S(A,B) > 0$. Otherwise, $S(A,B) = 0$;

where the first three properties [C1] - [C3] are equivalent to the properties [A2] - [A4].

In practice, classical set-theoretic similarity measures are extended to assess similarity between fuzzy sets [75], and a comprehensive overview of these measures can be seen in [11] and [87]. Two common approaches for extension are based on the vertical-slice and the α-cut representation of fuzzy sets [20] (discussed in subsection 2.4). Under the vertical-slice based extension, the fuzzy sets are compared with respect to the corresponding membership values for each element in the universe, whereas with the α-cut based extension, the fuzzy sets are compared with respect to their α-level sets. It is worth noting that both extensions show high similarity for larger overlap between fuzzy sets, and low similarity in the reverse situation [88]. Although many fuzzy set-theoretic similarity measures exist, the selection of a particular fuzzy similarity measure largely depends on the applications.

Among the fuzzy set-theoretic similarity measures, the extended Jaccard measure [13] proposed by Pappis and Karacapilidis [65] is the most common and widely used for T1 fuzzy sets. This measure satisfies all common properties of a fuzzy set similarity measure [C1] - [C4]. For two T1 fuzzy sets A and B on a discrete and finite universe of discourse X, the extended Jaccard measure is defined as [65]:

$$S_J^{T1}(A,B) = \frac{\sum_{i=1}^{N} \min\left(\mu_A(x_i), \mu_B(x_i)\right)}{\sum_{i=1}^{N} \max\left(\mu_A(x_i), \mu_B(x_i)\right)}, \tag{2.24}$$

where $\mu_A(x_i)$ and $\mu_B(x_i)$ are the membership grades of x_i in A and B respectively. The Jaccard similarity measure is further extended for higher-level of fuzzy sets [89][90].

Moreover, [10] has introduced a vertical-slice based extension of the Jaccard measure for T1 fuzzy sets, however, in this case both T1 fuzzy sets are required to be convex.

Apart from the extended Jaccard measure, other fuzzy set-theoretic similarity measures introduced by Chen et al. [91], Wang [92] and Raha et al. [93] are also well-known and used in many applications. Table 2.1 summarizes these similarity measures for the T1 fuzzy sets along with their essential properties. A fuzzy measure based on matching functions [91] is also included in Table 2.1 as it uses the maximum *t-conorm*. A detailed discussion of set-theoretic similarity measures for fuzzy sets can be found in [10] and [94].

It is worth mentioning that most of the vertical-slice based extended set-theoretic similarity measures perform in line with a variety of T1 fuzzy sets comprising different properties, such as normality, convexity, and non-convexity. However, for comparing non-normal T1 fuzzy sets, most of these similarity measures behave non-intuitively which leads to explore the α-cut based extended set-theoretic similarity measures for addressing the above case with the non-normal fuzzy sets (details in Chapter 5).

Table 2.1 Set-theoretic similarity measures and their properties for T1 fuzzy sets, A and B.

Similarity Measure		Properties			
Equation	Proposed by	Reflexivity	Symmetry	Transitivity	Overlapping
$S_J(A,B) = \frac{\sum_{i=1}^{n} \min(\mu_A(x_i),\mu_B(x_i))}{\sum_{i=1}^{n} \max(\mu_A(x_i),\mu_B(x_i))}$	Pappis and Karacapilidis [65]	Yes	Yes	Yes	>0
$S_C(A,B) = \max_i(\min(\mu_A(x_i), \mu_B(x_i))$	Chen et al. [91]	No	Yes	No	> 0
$S_P(A,B) = \frac{\sum_{i=1}^{n} \mu_A(x_i) \times \mu_B(x_i)}{\max\left(\sum_{i=1}^{n} (\mu_A(x_i))^2, \sum_{i=1}^{n} (\mu_B(x_i))^2\right)}$	Chen et al. [91]	Yes	Yes	No	> 0
$S_W(A,B) = \frac{1}{n} \sum_{i=1}^{n} \frac{\min(\mu_A(x_i),\mu_B(x_i))}{\max(\mu_A(x_i),\mu_B(x_i))}$	Wang [92]	Yes	Yes	Yes	> 0
$S_R(A,B) = \frac{\max_x(\min(\mu_A(x),\mu_B(x)))}{\max(\max_x(\mu_A(x)),\max_x(\mu_B(x)))}$	Raha et al. [93]	Yes	Yes	Yes	>0

To bring more accuracy in measuring similarity for uncertain data (captured by intervals and T1 fuzzy sets), this thesis aims to develop robust similarity measures involving the concept of subsethood. Therefore, in the following section, as an underlying concept, axiomatic definitions of subsethood in respect to crisp sets and fuzzy sets are discussed.

2.6 Subsethood

Subsethood is defined as a relation that indicates the degree to which one object is a subset of other object [95].

2.6.1 For Crisp Sets

For two crisp sets a and b, the subsethood of a with respect to b is defined as follows.

$$S_h(a,b) = \frac{|a \cap b|}{|a|}, \tag{2.25}$$

where $|a \cap b|$ is the cardinality of the intersection of sets, a and b, and $|a|$ is the cardinality of set a. From Eq. (2.25), it is clear that subsethood is smaller when more elements of set a are not part of set b and is larger when more elements of set a are part of set b. In general, Eq. (2.25) is bounded by the interval $[0, 1]$, where 1 means that a is a subset of b ($a \subseteq b$) and 0 means a is not a subset of b ($a \not\subset b$) – they are disjoint. The following two axioms are used to define the subsethood for crisp sets [96]:

[D1] $S_h(a,b) = 1$ if and only if $a \subseteq b$;

[D2] $S_h(a,b) = 0$ if and only if $a \not\subset b$.

Example 2.2:

Consider two sets $a = \{1, 2, 3\}$ and $b = \{1, 2, 3, 4, 5, 6\}$. Here, a is a subset of b as every

element from a are also in b. In opposite case, b is not a subset of a as elements 4, 5, and 6 from b are not in a. Thus, $S_h(a,b) = 1$ and $S_h(b,a) = 0$.

2.6.2 For Fuzzy Sets

A binary notion of subsethood for T1 fuzzy sets was first defined in [2] where for two fuzzy sets A and B on the universe X, A is a subset of B ($A \subseteq B$) if and only if $\mu_A(x) \leq \mu_B, \forall x \in X$. This definition is inherently crisp—A is, or is not a subset of B – which is incoherent in respect to the fuzzy set theory. Many alternatives for fuzzy subsethood have been introduced. Using the set-theoretic approach, the degree to which A is a subset of B on a finite X is defined as [97]

$$S_h(A,B) = \begin{cases} 1, & A = \emptyset \\ \dfrac{\sum_{i=1}^{n} \min(\mu_A(x_i), \mu_B(x_i))}{\sum_{i=1}^{n} \mu_A(x_i)}, & A \neq \emptyset, \end{cases} \tag{2.26}$$

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. Using the fuzzy implication operator $\mathscr{I}$, the fuzzy subsethood is expressed as [98]

$$S_h(A,B) = \inf_{x \in X} \mathscr{I}(\mu_A(x), \mu_B(x)) \tag{2.27}$$

where $\mathscr{I} : [0,1] \times [0,1] \to [0,1]$ is such that $\mathscr{I}(0,0) = \mathscr{I}(0,1) = \mathscr{I}(1,1) = 1$ and $\mathscr{I}(1,0) = 0$.

In the literature, fuzzy subsethood has been characterized by different axiomatizations. Initially, four axioms were proposed to define it as a binary fuzzy relation [99]. Subsequently, [100] developed another set of axioms for fuzzy subsethood. We note that these axioms (except the first two axioms) are equivalent to those in [99]. [101]

argued that some fuzzy subsethoods do not maintain all these axioms and proposed new ones. For fuzzy sets $A, B, C \in X$, these axioms are –

[E1] $S_h(A, B) = 1$ if and only if $A \subseteq B$;

[E2] If $P \leq A$, then $S_h(A, A^c) = 0$ if and only if $A = X$;

[E3] if $A \subseteq B \subseteq C$, then $S_h(C, A) \leq S_h(B, A)$ and if $A \subseteq B$, then $S_h(C, A) \leq S_h(C, B)$;

where P is the fuzzy set with $\mu_P(x) = \frac{1}{2}$ and A^c is the complement of fuzzy set A, i.e., $\mu_{A^c}(x) = 1 - \mu_A(x)$.

Later, axiom [E3] was modified as "if $A \subseteq B \subseteq C,$ then $S_h(C, A) \leq S_h(B, A)$ and $S_h(C, A) \leq S_h(C, B)$ " in [102] for defining new fuzzy subsethoods and their application to cluster validity. Further, axioms [E2] and [E3] were altered in [103] as "$S_h(A, A^c) = 0$ if and only if $A = X$", and "if $A \subseteq B,$ then $S_h(A, C) \geq S_h(B, C)$ and $S_h(C, A) \leq S_h(C, B)$" respectively to introduce a new fuzzy DI-subsethood by aggregating implication functions.

In many cases, fuzzy subsethood axioms are extended for higher level of fuzzy sets, e.g., T2 fuzzy sets [104]. These extensions are not discussed further as this thesis only focuses on the subsethood for crisp and T1 fuzzy sets (for more details please see [105–107]).

Remark 2.2: Building on the existing literature on subsethood, particularly its key axioms, Chapter 3 introduces an axiomatic definition of subsethood for continuous and closed intervals which is later involved in proposing a new set-theoretic similarity measure for intervals in the same chapter.

The next section presents different aggregation functions, specially focusing on FIs in respect to FMs. Here, we give importance on how to aggregate similar data

from multiple uncertain sources by involving similarity measures as it is an important application area of similarity measure.

2.7 Aggregation of Uncertain Data Types

Data aggregation (fusion) is defined as a process of combining data from multiple sources so as to answer question of interests and to make appropriate decisions [108]. The data aggregation problem often deals with uncertain and imprecise multi-source data and generally involves various mathematical functions, algorithms, and methodologies to combine them [26][109]. To explain this, consider a situation where two sources (e.g., thermometers) provide two different room temperatures. Under such inconsistency and uncertainty, the data aggregation may involve additional information such as reliability or worth of sources together with the available temperature data and fuse them to find the best (range of) temperature for the room.

According to [110], the success of a data aggregation largely depends on the quality of knowledge produced by the aggregation process which in turn relies on the representation of multi-source data, the quality of uncertainty model, and the accuracy and applicability of prior knowledge. In the literature, numerous data aggregation strategies and functions have been proposed [111][112]. Aggregation functions such as weighted average [113], weighted minimum and maximum [114], and ordered weighted average [115] are commonly used [116]. Further, FIs with respect to an FM, which are the generalizations of aforementioned functions are very popular alternatives for aggregation [117–119].

This thesis pays attention to commonly used FIs in combination with FMs which are explained in the next two subsections.

2.7.1 Fuzzy Measures

The concept of a measure is one of the most important concepts in mathematics [120]. Classical measures share the additive property which can be restrictive, making the measure inapplicable to many real-world problems such as uncertainty modelling and data aggregation [121]. The FM does not maintain additivity, but follows the less strict property of monotonicity.

An FM is defined as a hierarchical weighting structure (a lattice) that captures the worth of all data sources along with all of their possible subsets [122]. Mathematically, an FM, g defined on a discrete and finite set of sources, $X = \{x_1, ..., x_n\}$ is a set function $g : 2^X \rightarrow [0, 1]$ satisfying the following properties [123]:

(1) $g(\emptyset) = 0$ and $g(X) = 1$ (Bounded)

(2) If $a \subseteq b \subseteq X$ then $g(a) \subseteq g(b) \subseteq g(X)$ (Monotonic and non-decreasing)

In the context of multi-source data aggregation, $g(a)$ presents the weight or worth of each subset a of X. Property (1) states that the worth of empty set ($\emptyset$) is 0 and the worth of universal set (X) is 1. Property (2) shows the monotonicity of g, i.e., if a is a subset of b ($a \subseteq b$), then the worth of a is smaller or equal to the worth of b. There is a third property of continuous FMs, which is not applicable to discrete FMs, as used in this thesis and most practical applications. Figure 2.9 presents an FM lattice for three data sources $\{x_1, x_2, x_3\}$.

There are three major approaches for generating FMs [122], which discussed below.

- Expert-defined approach – it is a manual specification of FM by expert. It can be challenging, since for even a small number of sources, the manual task of generating the FM lattice becomes vast and difficult. For example, the worth of 31 subsets needs to be specified for only five sources.[3]

[3]For n sources, there are $(2^n - 1)$ non-empty subsets of sources.

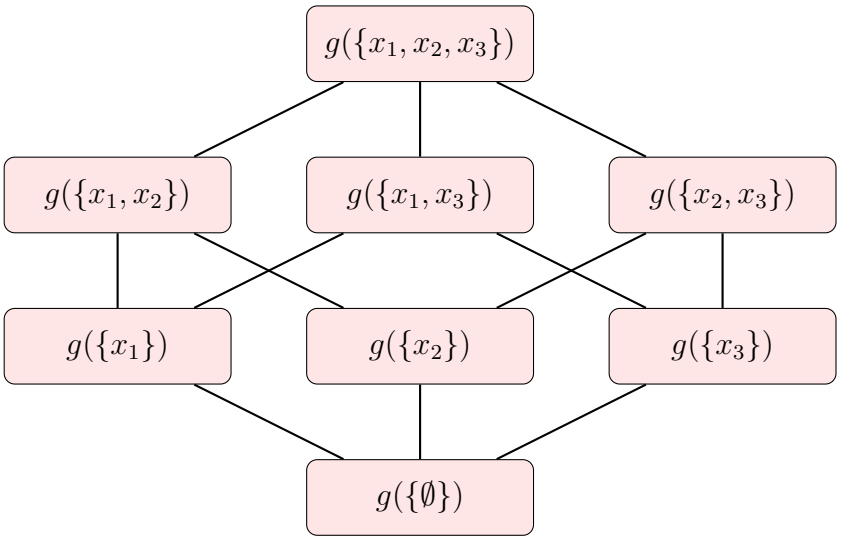


Fig. 2.9 The FM lattice for three data sources.

- Algorithmic approach – various algorithm-specified FMs, such as the Sugeno λ-measure [124] and decomposable measure [75] are used in practice where they derive the FM from the worth of single sources (densities) considering the FM's mathematical constraints. These algorithm-driven measures need the weights (densities) of base sources to be defined in advance and derive the entire FM lattice from the source densities, constrained by the properties (1) and (2) of FMs. Besides, more recently introduced data-driven FMs [22–24] derive the entire FM lattice directly from the available multi-source data based on various criteria such as the level of agreement over given subsets of sources.

- Optimization – Several learning methods, e.g., evolutionary algorithms [125], quadratic programming [126] are used to derive FMs while relying on training data. These methods define the FM as a set of independent weights which are optimized to produce minimum output error in conjunction with the FI.

This thesis particularly focuses on data-driven FMs (based on the concept of modelling source-agreement) which are used in conjunction with FIs (see subsection 2.7.2) for aggregating multi-source data where sources are anonymous and no external

information is available. The data-driven FMs analyze various properties of given data (e.g., agreement among sources) and subsequently extract the worth of all subsets of sources from the data. Currently, there are three existing data-driven FMs based on source-agreement – *FM of agreement* (*AG*) [22], *FM of generalized accord* (*GenA*) [23], and *additive measure of agreement* (*AA*) [24]. The discussion on these agreement-based FMs with respect to a set of intervals $\overline{h} = \{\overline{h}_1, \overline{h}_2, ..., \overline{h}_n\}$ arising from n individual sources is given below.

(1) Fuzzy Measure of Agreement

Wagner and Anderson [22] proposed the original *FM of Agreement* (*AG*) that is extracted from a pool of interval-valued data with no prior knowledge about sources. The *AG* FM is defined as -

$$\tilde{g}^{AG}(\overline{A}_i) = \begin{cases} 0 & \text{for i = 0,1} \quad (2.28a) \\ \sum_{K=2}^{i} \left| U_K(\overline{A}_i) \right| z_K & \text{for i = [2 : n]} \quad (2.28b) \end{cases}$$

where $\overline{A}_i = \{\overline{h}_{\pi(1)}, \overline{h}_{\pi(2)}..., \overline{h}_{\pi(i)}\}$ is the permuted set of intervals with $\overline{A}_0 = \emptyset$, $z_i = \frac{i}{n}$ and $|.|$ refers to the cardinality/size of the interval. Here, $\overline{U}_K(\overline{A}_i)$ unites the intersections of the K-tuples in $\overline{A}_i \subseteq \overline{h}$ as defined in Eq. (2.29) [23][24].

$$\overline{U}_K(\overline{A}_i) = \bigcup_{k_1=1}^{i-K+1} \bigcup_{k_2=k_1+1}^{i-K+2} ... \bigcup_{k_K=k_{K-1}+1}^{i} (\overline{h}_{\pi(k_1)} \cap \overline{h}_{\pi(k_2)} \cap ... \cap \overline{h}_{\pi(k_K)}) \quad (2.29)$$

Further, $\tilde{g}^{AG}(\overline{A}_i)$ is normalized by $\tilde{g}^{AG}(\overline{h})$ in Eq. (2.30) .

$$g^{AG}(\overline{A}_i) = \frac{\tilde{g}^{AG}(\overline{A}_i)}{\tilde{g}^{AG}(\overline{h})} \quad (2.30)$$

It is worth mentioning that the *AG* FM was originally designed as a component of a meta-measure, i.e., for use in combination with other data-driven FMs, such as *FM of Specificity* [22].

(2) Fuzzy Measure of Generalized Accord

Havens et al. [23] proposed the *FM of Generalized Accord* (*GenA*) which uses the similarity of subsets of sources (estimated for example by Jaccard or Dice similarity measures) for establishing their agreement/accord. The *GenA* FM is defined as -

$$g^{GenA}(\overline{A}_i) = \begin{cases} 0 & \text{for i = 0 , 1} \quad (2.31a) \\ \alpha_{\overline{h}} \sum_{K=2}^{i} S_K(\overline{A}_i) & \text{for i = [2 : n]} \quad (2.31b) \end{cases}$$

where $\overline{A}_i = \{\overline{h}_{\pi(1)}, \overline{h}_{\pi(2)}...., \overline{h}_{\pi(i)}\}$ is the permuted set of intervals with $\overline{A}_0 = \emptyset$, and $S_K(\overline{A}_i)$ is defined in Eq. (2.32).

$$S_K(\overline{A}_i) = \binom{n}{K}^{-1} \sum_{k_1=1}^{i-K} \sum_{k_2=k_1+1}^{i-K+1} ... \sum_{k_K=k_{K-1}+1}^{i} s(\{\overline{h}_{\pi(k_1)}, \overline{h}_{\pi(k_2)}, ..., \overline{h}_{\pi(k_K)}\}) \quad (2.32)$$

Here, $\binom{n}{K}$ is the number of possible K-tuples in $\overline{h}$. The quantity $S_K(\overline{A}_i)$ is the sum of similarities of the K-tuples in $\overline{A}_i \subseteq \overline{h}$, weighted by $\binom{n}{K}^{-1}$. Further, the constant $\alpha_{\overline{h}}$ is defined in Eq. (2.33) so that $g^{GenA}(\overline{h}) = 1$.

$$\alpha_{\overline{h}} = \left(\sum_{K=2}^{n} S_K(\overline{A}_n) \right)^{-1} \quad (2.33)$$

(3) Additive Measure of Agreement

Havens et al. [24] proposed the *Additive Measure of Agreement* (*AA*) in order to alleviate the asymmetry issue. This FM also involves similarity measure for determining the

agreement for a set of sources. The AA FM is expressed in Eq. (2.34).

$$\tilde{g}^{AA}(\overline{A}_i) = \tilde{g}^{AA}(\overline{A}_{i-1}) + \sum_{\substack{j=1 \\ j\neq i}}^{n} s^p(\overline{h}_j, \overline{h}_{\pi(i)}), \;\; i = [n], p \geq 0 \tag{2.34}$$

where p is a tuning parameter. Further, $\tilde{g}^{AA}(\bar{A}_i)$ is normalized by $\tilde{g}^{AA}(\bar{A}_n)$ in Eq. (2.35) to satisfy the property of an FM.

$$g^{AA}(A_i) = \frac{\tilde{g}^{AA}(\overline{A}_i)}{\tilde{g}^{AA}(\overline{A}_n)} \tag{2.35}$$

One advantage of the AA FM is that it minimizes the asymmetry issue related to the positive- and negative-valued intervals in respect to FM when used in conjunction with the FI.

2.7.2 Fuzzy Integrals

FIs have been efficiently used as powerful non-linear aggregation operators [127], commonly used in evidence fusion [123][128]. They aggregate multi-source data (evidence) by combining them with respect to the worth of all possible subsets of sources (captured by an FM). Two well-known FIs are the Sugeno fuzzy integral (SFI) [124] and the Choquet fuzzy integral (CFI) [129]. In practice, discrete SFI and CFI are used [129][130], and this thesis particularly uses the discrete CFI as it is more popular for data fusion.

Let $h : X \rightarrow [0, \infty)$ be a real-valued function that represents the data or evidence from a source. The discrete CFI is defined in Eq. (2.36).

$$\int_{CFI} h \circ g = CFI_g(h) = \sum_{i=1}^{n} h(x_{\pi(i)})[g(A_i) - g(A_{i-1})] \tag{2.36}$$

where π is a permutation of X arranged like $h(x_{\pi(1)}) \geq h(x_{\pi(2)}) \geq ... \geq h(x_{\pi(n)})$. $A_i = \{x_{\pi(1)}, x_{\pi(2)}, ..., x_{\pi(i)}\}$ is a subset of sources. g is the FM where $g(A_i)$ is the worth of the subset A_i with $g(A_0) = 0$. A detailed discussion on the properties of FIs and the CFI can be found in [131].

In most cases, the multi-source data $\overline{h}$ is represented in a numeric form. However, in some applications $\overline{h}$ is better represented by interval-valued or fuzzy set-valued forms [22–24]. For example, in response to a survey question, "What is the temperature today?", many people may prefer to answer with an interval, e.g., "between 6^oC to 10^oC", or a fuzzy set, e.g., "about 9^oC". A number of research works have focused on extending the FIs for both interval-valued and fuzzy set-valued data [132–135].

Let $\overline{h} : X \rightarrow I(\mathbb{R})$ be a set of interval-valued data where $I(\mathbb{R})$ is the set of all closed intervals over the real numbers and $\overline{h}_i = h(x_i) = [h_i^-, h_i^+]$ be the ith interval (where h_i^- and h_i^+ are the left and right endpoints respectively). Following the notation in [23], the CFI on $\overline{h}$ is defined as

$$\int_{CFI} \overline{h} \circ g = CFI_g(\overline{h}) = [CFI_g(h^-), CFI_g(h^+)], \tag{2.37}$$

where the output $CFI_g(\overline{h})$ is itself interval-valued [75]. In other words, the CFI for interval-valued data is computed by separately computing the CFI for the numeric case of the left and right interval endpoints respectively [23].

2.8 Summary

This chapter has presented an overview of modelling uncertainty using intervals and fuzzy sets. In particular, it has discussed their definitions, various properties, their different categories and how they are mathematically modelled using vertical-slice and α-cut (horizontal-slice) representations (in case of fuzzy sets). This chapter has also

described similarity measures and how dissimilarity measures are derived from them. Then, a discussion has been given on how to assess similarity between intervals and also between T1 fuzzy sets by applying different set-theoretic similarity measures.

This chapter has also presented discussions on subsethood for crisp and fuzzy sets, and provided different axiomatizations to characterize subsethood. Further, as a background for an application of the similarity measures in uncertain data aggregation, this chapter has given an overview of data aggregation and various aggregation operators, especially focusing on FIs defined in respect to FMs.

With this research background, in the upcoming chapters, this thesis will investigate existing set-theoretic measures to identify their limitations in comparing uncertain data types covering intervals and T1 fuzzy sets, and propose robust similarity measures to gain more accuracy in their comparison. Further, as an application of the new similarity measures, this thesis will explore their potential benefits in aggregation of 'similar data' from uncertain sources involving FIs in respect to FMs.

The next chapter will identify and articulate a limitation (i.e., *aliasing*) of the popular set-theoretic similarity measures and hence introduce a *robust* similarity measure for continuous (closed) intervals to avoid *aliasing* and thus providing similarity with high accuracy. This will fulfill objectives 1 to 4 of the thesis.

Chapter 3

Measuring Similarity Between Continuous Intervals

3.1 Introduction

In view of the growing interest in interval-valued data for modelling uncertain data and the significance of their effective comparison, this chapter focuses on comparing similarity between continuous intervals with high accuracy. In practice, well-established set-theoretic similarity measures, such as the Jaccard [13] and Dice [14] measures are commonly used for comparing continuous intervals, while axiomatically a wide breadth of possible measures have been theoretically explored. This chapter identifies, articulates, and addresses an inherent and so far not discussed limitation of popular measures – their tendency to be subject to *aliasing* – where they return the same similarity value for very different sets of intervals having dissimilar sizes. This limitation risks counter-intuitive results and poor automated reasoning in real-world applications dependent on systematically comparing interval-valued system variables or states.

In this context, this chapter introduces new axioms to establish desirable properties for *robust* similarity measures, followed by putting forward a novel set-theoretic

similarity measure based on the concept of bidirectional subsethood which satisfies both traditional and new axioms. The proposed measure is designed to be sensitive to the variation in the size of intervals, thus avoiding *aliasing*.

In this chapter, the motivation behind this new similarity measure is discussed first and then, the underpinnings of *robust* similarity measures are developed by putting forward five new axioms which complement the traditional axioms of common similarity measures (section 2.5 in Chapter 2). Building on this set of nine axioms, the chapter then proposes a new set-theoretic similarity measure for intervals based on their bidirectional subsethood, followed by a discussion of its major properties. As part of developing the new similarity measure, this chapter also introduces the definition of subsethood for intervals and proofs of its mathematical properties. After that, a new dissimilarity or distance measure, which satisfies the properties of a metric, is derived. Lastly, the new measure is compared against popular Jaccard and Dice similarity measures to show that it avoids *aliasing* while delivering intuitive results (e.g., being bounded above and below by Jaccard and Dice measures), thus supporting its suitability in real world applications.

3.2 Motivation

Intervals are often compared through various similarity measures. Among them, the Jaccard [13] and Dice [14] similarity measures are the most commonly used set-theoretic measures in practice. These two measures provide symmetrical similarity, growing slowly from a minimum (0) to a maximum value (1) in response to an increasing degree of overlap between two intervals. Nevertheless, both of these, and indeed most/all set-theoretic similarity measures often suffer from a potential shortcoming, which we call *aliasing*. *Aliasing* is where the same similarity value appears for very different interval pairs having different cardinalities or sizes. Figure 3.1 shows an example of

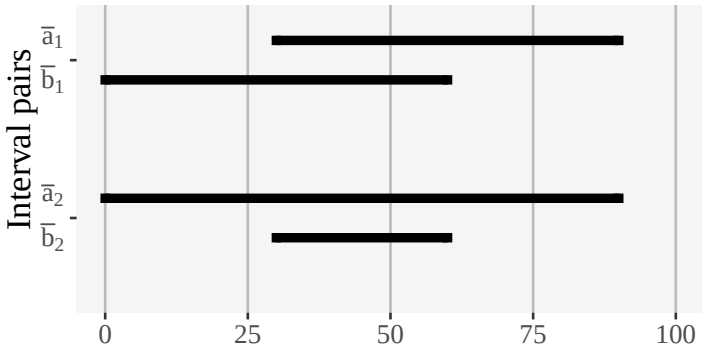


Fig. 3.1 Two different pairs of intervals showing *aliasing* with the Jaccard and Dice similarity measures.

such interval pairs for which the Jaccard and Dice measures give the same similarity of 0.33 and 0.50 respectively. While similarity measures returning such identical results for different sets of intervals is at least counter-intuitive, at worst it leads to incorrect inference in real-world applications. The reason why these measures exhibit *aliasing* is because their sensitivity in respect to changes in the relative size of the intervals is limited, i.e., they are largely driven by the size of intersection and union. However, it is reasonable to expect that similarity would vary *both* in respect to the overlap and to the mutual similarity-in-size of the intervals. This expectation leads us to explore the axioms required for a robust similarity measure. We then proceed to develop a new set-theoretic similarity measure for intervals in respect to these axioms, which is more responsive towards the variation in the size of intervals than the current measures, thus avoiding *aliasing* and giving similarity with high accuracy.

In the next section, we first introduce subsethood for continuous and closed intervals, and its properties. Then, an expanded axiomatic definition for robust similarity measures has been presented, followed by a new similarity measure for intervals which fulfills the said definition and provides robust comparison of intervals.

3.3 Proposed Similarity Measure

As outlined in section 3.1, *aliasing* poses a so far not discussed challenge to the robust comparison of intervals. In order to address this and move towards robust similarity measures, an axiomatic definition for *robust* similarity measures for intervals is first put forward in subsection 3.3.2. Then, building on this, in subsection 3.3.3, a novel set-theoretic similarity measure is proposed for intervals based on their reciprocal overlapping ratios or subsethoods [16], showing that it follows the axiomatic definition of a *robust* similarity measure. It is worth mentioning that the properties of the subsethood for intervals are also introduced in subsection 3.3.1 in respect to the wider literature (discussed in Chapter 2).

Remark 3.1:
The terms 'overlapping ratio' and 'subsethood' are equivalent in practice, but 'subsethood' is commonly used and well understood in the literature. We use the later term throughout the thesis.

The next subsection presents the definition of subsethood for a pair of continuous and closed intervals, and also its key properties.

3.3.1 Definition of Subsethood for Continuous and Closed Intervals

Following the definition of subsethood for crisp sets (in Chapter 2), the degree of subsethood for a pair of two continuous and closed intervals, $\overline{a}$ and $\overline{b}$ can be defined as

$$S_h\left(\overline{a}, \overline{b}\right) = \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a}\right|}, \tag{3.1}$$

where $\left|\overline{a} \cap \overline{b}\right|$ is the size of the intersection between $\overline{a}$ and $\overline{b}$ and $|\overline{a}| \neq 0$. It is noted that subsethood is also known as inclusion [136] and overlapping ratio [16] as it captures the overlapping ratio between intervals. Building on the axiomatic definitions of subsethood in the literature (discussed in Chapter 2), the key properties of subsethood for intervals are captured in Theorem 3.1 as follows.

Theorem 3.1:

A real-valued function $S_h : \overline{a} \times \overline{b} \rightarrow [0, 1]$ captures the subsethood for a pair of continuous and closed intervals, if it follows the properties,

(1) $S_h(\overline{a}, \overline{b}) = 1$ if and only if $\overline{a} \subseteq \overline{b}$,

(2) $S_h(\overline{a}, \overline{b}) = 0$ if and only if $\overline{a}$ and $\overline{b}$ are disjoint,

(3) if $\overline{a} \subseteq \overline{b} \subseteq \overline{c}$, then $S_h(\overline{c}, \overline{a}) \leq S_h(\overline{b}, \overline{a})$ and $S_h(\overline{c}, \overline{a}) \leq S_h(\overline{c}, \overline{b})$.

Proof: Below we give proofs for all the properties (1–3).

(1) If $\overline{a} \subseteq \overline{b}$, then $|\overline{a} \cap \overline{b}| = |\overline{a}|$. Thus, $S_h(\overline{a}, \overline{b}) = \frac{|\overline{a} \cap \overline{b}|}{|\overline{a}|} = \frac{|\overline{a}|}{|\overline{a}|} = 1$. Again, $S_h(\overline{a}, \overline{b}) = 1$ only when $\overline{a}$ is included in $\overline{b}$.

(2) If $\overline{a}$ and $\overline{b}$ are disjoint, then $|\overline{a} \cap \overline{b}| = 0$. Thus, $S_h(\overline{a}, \overline{b}) = 0$. Contrarily, $S_h(\overline{a}, \overline{b}) = 0$ when $\overline{a}$ and $\overline{b}$ are not overlapping.

(3) If $\overline{a} \subseteq \overline{b} \subseteq \overline{c}$, then $|\overline{a}| \leq |\overline{b}| \leq |\overline{c}|$. Thus, $S_h(\overline{c}, \overline{a}) = \frac{|\overline{c} \cap \overline{a}|}{|\overline{c}|} = \frac{|\overline{a}|}{|\overline{c}|}$ and $S_h(\overline{b}, \overline{a}) = \frac{|\overline{b} \cap \overline{a}|}{|\overline{b}|} = \frac{|\overline{a}|}{|\overline{b}|}$. Since $|\overline{b}| \leq |\overline{c}|$, it follows that $\frac{1}{|\overline{c}|} \leq \frac{1}{|\overline{b}|}$. Hence, $\frac{|\overline{a}|}{|\overline{c}|} \leq \frac{|\overline{a}|}{|\overline{b}|}$, implying $S_h(\overline{c}, \overline{a}) \leq S_h(\overline{b}, \overline{a})$. Moreover, $S_h(\overline{c}, \overline{b}) = \frac{|\overline{c} \cap \overline{b}|}{|\overline{c}|} = \frac{|\overline{b}|}{|\overline{c}|}$. Since $|\overline{a}| \leq |\overline{b}|$, it follows that $\frac{|\overline{a}|}{|\overline{c}|} \leq \frac{|\overline{b}|}{|\overline{c}|}$, proving that $S_h(\overline{c}, \overline{a}) \leq S_h(\overline{c}, \overline{b})$. ■

In the next subsection, an expanded axiomatic definition for a *robust* similarity measure for continuous and closed intervals in introduced.

3.3.2 An Expanded Axiomatic Definition for Robust Similarity Measures on Continuous and Closed Intervals

A real-valued function $S : \bar{a} \times \bar{b} \rightarrow [0, 1]$ is defined as a similarity measure for continuous and closed intervals if it maintains the following axioms:

[F1] $0 \leq S(\bar{a}, \bar{b}) \leq 1$ (boundedness);

[F2] $S(\bar{a}, \bar{b}) = S(\bar{b}, \bar{a})$ (symmetry);

[F3] $S(\bar{a}, \bar{b}) = 1$ if and only if $\bar{a} = \bar{b}$ (reflexivity);

[F4] If $\bar{a} \subseteq \bar{b} \subseteq \bar{c}$, then $S(\bar{a}, \bar{b}) \geq S(\bar{a}, \bar{c})$ and $S(\bar{b}, \bar{c}) \geq S(\bar{b}, \bar{a})$ (transitivity);

[F5] $S(\bar{a}, \bar{b}) = 0$ if and only if $|\bar{a} \cap \bar{b}| = 0$ (disjointness);

[F6] For two pairs $\{\bar{a}_1, \bar{b}_1\}$ and $\{\bar{a}_2, \bar{b}_2\}$ where $|\bar{a}_1 \cap \bar{b}_1| > 0$, $|\bar{a}_2 \cap \bar{b}_2| > 0$, $\left|\bar{a}_1 \cup \bar{b}_1\right| = \left|\bar{a}_2 \cup \bar{b}_2\right|$ and $\left|\bar{a}_1 \cap \bar{b}_1\right| = \left|\bar{a}_2 \cap \bar{b}_2\right|$, $S\left(\bar{a}_1, \bar{b}_1\right) \neq S\left(\bar{a}_2, \bar{b}_2\right)$ if $\left|\bar{a}_1\right| \neq \left|\bar{a}_2\right|$ and $\left|\bar{a}_1\right| \neq \left|\bar{b}_2\right|$, and $\left|\bar{b}_1\right| \neq \left|\bar{a}_2\right|$ and $\left|\bar{b}_1\right| \neq \left|\bar{b}_2\right|$ (non-aliasing);

[F7] $S(\bar{a}, \bar{b}) < 1$ if $\bar{a} \subset \bar{b}$ (subsethood-limit);

[F8] For two pairs $\{\bar{a}_1, \bar{b}_1\}$ and $\{\bar{a}_2, \bar{b}_2\}$ where $|\bar{a}_1 \cap \bar{b}_1| > 0$ and $|\bar{a}_2 \cap \bar{b}_2| > 0$, $S\left(\bar{a}_1, \bar{b}_1\right) = S\left(\bar{a}_2, \bar{b}_2\right)$ if $\bar{a}_2 = n \times \bar{a}_1 = [n \times a_1^-, n \times a_1^+]$ and $\bar{b}_2 = n \times \bar{b}_1 = [n \times b_1^-, n \times b_1^+]$ where $n > 0$ is a scaling factor (scaling-invariance);

[F9] For two pairs $\{\bar{a}_1, \bar{b}_1\}$ and $\{\bar{a}_2, \bar{b}_2\}$ where $\left|\bar{a}_1\right| = \left|\bar{b}_1\right| = \left|\bar{a}_2\right| = \left|\bar{b}_2\right|$, $S\left(\bar{a}_1, \bar{b}_1\right) < S\left(\bar{a}_2, \bar{b}_2\right)$ if $\left|\bar{a}_1 \cap \bar{b}_1\right| < \left|\bar{a}_2 \cap \bar{b}_2\right|$ (increased-overlap).

Justification of the axioms: The axioms [F1] - [F4] are derived following the common axioms [A1] - [A4] stated in Chapter 2. Axiom [F5] is put forward for disjoint intervals where their similarity should be 0 as they are not overlapping. The *aliasing* example discussed in Section 3.2 calls for the axiom [F6]. Axiom [F7] is presented for intervals

when one interval is a proper subset of the other. Here, the similarity should be less than 1 as both intervals are not equal. Axiom [F8] is introduced for intervals where they are scaled up by a factor. In this case, the ratio of overlapping and non-overlapping segments stays the same while scaling, thus, the similarity between intervals should remain constant. Lastly, axiom [F9] suggests higher similarity for equal-sized intervals when their overlap increases.

Based on the expanded axiomatic definition of robust similarity measures, the next subsection proposes a novel set-theoretic similarity measure for continuous and closed intervals that satisfies all the axioms [F1] - [F9].

3.3.3 Proposed Similarity Measure based on Bidirectional Subsethood

As noted in section 3.2, the motivation behind the new similarity measure is to establish a similarity measure which is sensitive to the potentially (very) different cardinalities of the sets being compared, in particular when one is a subset of the other. Thus, the proposed similarity measure S_{S_h} takes into consideration the reciprocal subsethood (i.e., the overlapping ratio) of both sets/intervals being compared, in order to estimate their overall similarity.

Definition 3.1:

The bidirectional subsethood based similarity measure S_{S_h} for a pair of intervals, $\overline{a}$ and $\overline{b}$, is the *t-norm* of their reciprocal subsethoods $S_h(\overline{a}, \overline{b})$ and $S_h(\overline{b}, \overline{a})$, i.e.,

$$S_{S_h}(\overline{a}, \overline{b}) = \bigstar(S_h(\overline{a}, \overline{b}), S_h(\overline{b}, \overline{a})), \tag{3.2}$$

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

Similar to Eqs. (2.19) and (2.23), Eq. (3.2) can be written as

$$S_{S_h}\left(\overline{a},\overline{b}\right) = \bigstar\left(\frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right|+\left|\overline{a}\backslash\overline{b}\right|}, \frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right|+\left|\overline{b}\backslash\overline{a}\right|}\right), \tag{3.3}$$

where $\left|\overline{a}\backslash\overline{b}\right|$ is the size of non-overlapping segment(s) of $\overline{a}$ as to $\overline{b}$ and vice-versa for $\left|\overline{b}\backslash\overline{a}\right|$. Also, $|\overline{a}| \neq 0$ and $\left|\overline{b}\right| \neq 0$.

Remark 3.2:

Throughout of this thesis, we use $\bigstar$ to refer to either minimum *t-norm* or product *t-norm* while using '$\wedge$' and 'Π' to indicate the minimum and product *t-norm*s respectively.

Theorem 3.2:

Consider a minimum ($\wedge$) or product (Π) *t-norm* as $\bigstar$ and the subsethood for intervals S_h. Then, $S_{S_h}(\overline{a},\overline{b}) = \bigstar(S_h(\overline{a},\overline{b}), S_h(\overline{b},\overline{a}))$ is a similarity measure for intervals $\overline{a}$ and $\overline{b}$ satisfying all axioms [F1] - [F9].

Proof: The proofs are given per axioms in the next subsection 3.3.4. ■

3.3.4 Properties of the Proposed Similarity Measure

In this subsection, we explore the properties of the proposed S_{S_h} similarity measure. Note that all properties discussed in this chapter are maintained by the S_{S_h} measure with both minimum ($\wedge$) and product (Π) *t-norm*s, unless otherwise specified.

Theorem 3.3 (Boundedness): $0 \leq S_{S_h}\left(\overline{a},\overline{b}\right) \leq 1$.

Proof: All *t-norm*s are bounded by 0 and 1[137]. This is also true for the subsethood S_h (see subsection 3.3.1). This eventually means that $S_{S_h}\left(\overline{a},\overline{b}\right)$ measure always within $[0, 1]$, thus addressing the axiom [F1]. ■

Theorem 3.4 (Symmetry): $S_{S_h}\left(\overline{a},\overline{b}\right) = S_{S_h}\left(\overline{b},\overline{a}\right)$.

Proof: Any *t-norm* ($\bigstar$) is symmetric [137]. Therefore, $S_{S_h}\left(\overline{a},\overline{b}\right)$ is also symmetric. The S_{S_h} measure, thus, satisfies the axiom [F2]. ■

Theorem 3.5 (Reflexivity): $S_{S_h}\left(\overline{a},\overline{b}\right) = 1 \iff \overline{a} = \overline{b}$.

Proof: If $\overline{a} = \overline{b}$, then $S_h\left(\overline{a},\overline{b}\right) = S_h\left(\overline{b},\overline{a}\right) = 1$. From the boundary conditions of the *t-norm* ($\bigstar$) [137], $\bigstar(1,1) = 1$, thus making $S_{S_h}\left(\overline{a},\overline{b}\right) = 1$. Alternatively, $S_{S_h}\left(\overline{a},\overline{b}\right) = 1$ means that both $S_h\left(\overline{a},\overline{b}\right)$ and $S_h\left(\overline{b},\overline{a}\right)$ are equal to 1. This only happens when $\overline{a}$ and $\overline{b}$ are identical intervals. Here, the S_{S_h} measure maintains the axiom [F3]. ■

Theorem 3.6 (Transitivity): If $\overline{a} \subseteq \overline{b} \subseteq \overline{c}$, then $S_{S_h}\left(\overline{a},\overline{b}\right) \geq S_{S_h}\left(\overline{a},\overline{c}\right)$ and $S_{S_h}\left(\overline{b},\overline{c}\right) \geq S_{S_h}\left(\overline{a},\overline{c}\right)$.

Proof: if $\overline{a} \subseteq \overline{b} \subseteq \overline{c}$, then

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

As $\overline{b} \subseteq \overline{c}$, it follows that $\left|\overline{b}\right| \leq \left|\overline{c}\right|$. Therefore, $\frac{\left|\overline{a}\right|}{\left|\overline{b}\right|} \geq \frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}$ and hence, $S_{S_h}\left(\overline{a},\overline{b}\right) \geq S_{S_h}\left(\overline{a},\overline{c}\right)$. Again,

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

As $\overline{a} \subseteq \overline{b}$, it follows that $\left|\overline{b}\right| \geq \left|\overline{a}\right|$. Hence, $\frac{\left|\overline{b}\right|}{\left|\overline{c}\right|} \geq \frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}$ and it implies that $S_{S_h}\left(\overline{b},\overline{c}\right) \geq S_{S_h}\left(\overline{a},\overline{c}\right)$. In this case, the new similarity measure satisfies the axiom [F4]. ■

Theorem 3.7 (Disjointness): $S_{S_h}\left(\overline{a},\overline{b}\right)=0$, if and only if $\overline{a}$ and $\overline{b}$ are disjoint.

Proof: If $\overline{a}$ and $\overline{b}$ are disjoint, then $\left|\overline{a}\cap\overline{b}\right|=0$. Hence, $S_h\left(\overline{a},\overline{b}\right)=0$ and $S_h\left(\overline{b},\overline{a}\right)=0$, implying that $S_{S_h}\left(\overline{a},\overline{b}\right)=\bigstar(0,0)=0$. Again, $S_{S_h}\left(\overline{a},\overline{b}\right)=0$ only when both $S_h\left(\overline{a},\overline{b}\right)$ and $S_h\left(\overline{b},\overline{a}\right)$ are equal to 0. This only happens when $\overline{a}$ and $\overline{b}$ are not overlapping. Here, the proof of this theorem shows that the new measure follows the axiom [F5]. ■

Theorem 3.8 (Non-aliasing): For two interval pairs $\{\overline{a}_1,\overline{b}_1\}$ and $\{\overline{a}_2,\overline{b}_2\}$ where $\left|\overline{a}_1\cap\overline{b}_1\right|>0$, $\left|\overline{a}_2\cap\overline{b}_2\right|>0$, $\left|\overline{a}_1\cup\overline{b}_1\right|=\left|\overline{a}_2\cup\overline{b}_2\right|$, and $\left|\overline{a}_1\cap\overline{b}_1\right|=\left|\overline{a}_2\cap\overline{b}_2\right|$, $S_{S_h}\left(\overline{a}_1,\overline{b}_1\right)\neq S_{S_h}\left(\overline{a}_2,\overline{b}_2\right)$ if $\left|\overline{a}_1\right|\neq\left|\overline{a}_2\right|$ and $\left|\overline{a}_1|\neq|\overline{b}_2\right|$, and $\left|\overline{b}_1|\neq|\overline{a}_2\right|$ and $\left|\overline{b}_1|\neq|\overline{b}_2\right|$.

Proof: For the pair $\{\overline{a}_1,\overline{b}_1\}$, $S_h\left(\overline{a}_1,\overline{b}_1\right)=\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{a}_1\right|}$ and $S_h\left(\overline{b}_1,\overline{a}_1\right)=\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{b}_1\right|}$. Thus,

$$S_{S_h}\left(\overline{a}_1,\overline{b}_1\right)=\bigstar\left(\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{a}_1\right|},\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{b}_1\right|}\right).$$

For the pair $\{\overline{a}_2,\overline{b}_2\}$, $S_h\left(\overline{a}_2,\overline{b}_2\right)=\frac{\left|\overline{a}_2\cap\overline{b}_2\right|}{\left|\overline{a}_2\right|}$ and $S_h\left(\overline{b}_2,\overline{a}_2\right)=\frac{\left|\overline{a}_2\cap\overline{b}_2\right|}{\left|\overline{b}_2\right|}$. Thus,

$$S_{S_h}\left(\overline{a}_2,\overline{b}_2\right)=\bigstar\left(\frac{\left|\overline{a}_2\cap\overline{b}_2\right|}{\left|\overline{a}_2\right|},\frac{\left|\overline{a}_2\cap\overline{b}_2\right|}{\left|\overline{b}_2\right|}\right).$$

As $\left|\overline{a}_1\cap\overline{b}_1\right|=\left|\overline{a}_2\cap\overline{b}_2\right|$, therefore,

$$S_{S_h}\left(\overline{a}_2,\overline{b}_2\right)\equiv\bigstar\left(\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{a}_2\right|},\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{b}_2\right|}\right).$$

Given that $\left|\overline{a}_1\right|\neq\left|\overline{a}_2\right|$ and $\left|\overline{a}_1\right|\neq\left|\overline{b}_2\right|$, it implies that

$$\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{a}_1\right|}\neq\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{a}_2\right|}\text{ and }\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{a}_1\right|}\neq\frac{\left|\overline{a}_1\cap\overline{b}_1\right|}{\left|\overline{b}_2\right|}.$$

Similarly, $\left|\overline{b}_1\right| \neq \left|\overline{a}_2\right|$ and $\left|\overline{b}_1\right| \neq \left|\overline{b}_2\right|$ meaning that

$$\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|} \neq \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a}_2\right|} \text{ and } \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|} \neq \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_2\right|}.$$

Therefore,

$$\bigstar\left(\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a}_1\right|}, \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|}\right) \neq \bigstar\left(\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a}_2\right|}, \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_2\right|}\right).$$

Hence, $S_{S_h}\left(\overline{a}_1, \overline{b}_1\right) \neq S_{S_h}\left(\overline{a}_2, \overline{b}_2\right)$, thus satisfying the axiom [F6]. ■

Theorem 3.9 (Subsethood-limit): $S_{S_h}(\overline{a}, \overline{b}) < 1$ if $\overline{a} \subset \overline{b}$.

Proof: If $\overline{a} \subset \overline{b}$, then $\left|\overline{a}\right| < \left|\overline{b}\right|$ and $\left|\overline{a} \cap \overline{b}\right| = \left|\overline{a}\right|$. Thus, $S_h\left(\overline{a}, \overline{b}\right) = \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a}\right|} = \frac{\left|\overline{a}\right|}{\left|\overline{a}\right|} = 1$ and $S_h\left(\overline{b}, \overline{a}\right) = \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{b}\right|} = \frac{\left|\overline{a}\right|}{\left|\overline{b}\right|}$. It means that $S_{S_h}(\overline{a}, \overline{b}) = \bigstar\left(1, \frac{\left|\overline{a}\right|}{\left|\overline{b}\right|}\right) = \frac{\left|\overline{a}\right|}{\left|\overline{b}\right|}$. As $\left|\overline{a}\right| < \left|\overline{b}\right|$, then $\frac{\left|\overline{a}\right|}{\left|\overline{b}\right|} < 1$. Hence, $S_{S_h}(\overline{a}, \overline{b}) < 1$. Here, the new similarity measure, thus, addresses the axiom [F7]. ■

Theorem 3.10 (Scaling-invariance): For two interval pairs $\{\overline{a}_1, \overline{b}_1\}$ and $\{\overline{a}_2, \overline{b}_2\}$ with $\left|\overline{a}_1 \cap \overline{b}_1\right| > 0$ and $\left|\overline{a}_2 \cap \overline{b}_2\right| > 0$, $S_{S_h}\left(\overline{a}_1, \overline{b}_1\right) = S_{S_h}(\overline{a}_2, \overline{b}_2)$ if $\overline{a}_2 = n \times \overline{a}_1 = [n \times a_1^-, n \times a_1^+]$ and $\overline{b}_2 = n \times \overline{b}_1 = [n \times b_1^-, n \times b_1^+]$ where $n > 0$ is a scaling factor.

Proof: For the pair $\{\overline{a}_1, \overline{b}_1\}$, $S_h\left(\overline{a}_1, \overline{b}_1\right) = \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a}_1\right|}$ and $S_h\left(\overline{b}_1, \overline{a}_1\right) = \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|}$. Thus,

$$S_{S_h}\left(\overline{a}_1, \overline{b}_1\right) = \bigstar\left(\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a_1}\right|}, \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|}\right).$$

For the pair $\{\overline{a}_2, \overline{b}_2\}$, $S_h\left(\overline{a}_2, \overline{b}_2\right) = \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{a}_2\right|}$ and $S_h\left(\overline{b}_2, \overline{a}_2\right) = \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{b}_2\right|}$. Thus,

$$S_{S_h}\left(\overline{a}_2, \overline{b}_2\right) = \bigstar\left(\frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{a}_2\right|}, \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{b}_2\right|}\right).$$

Given that $\overline{a}_2 = n \times \overline{a}_1$ and $\overline{b}_2 = n \times \overline{b}_1$ which implies that $\left|\overline{a}_2 \cap \overline{b}_2\right| = \left|(n \times \overline{a}_1) \cap (n \times \overline{b}_1)\right| = n \times \left|\overline{a}_1 \cap \overline{b}_1\right|$. Therefore,

$$\begin{aligned} S_{S_h}\left(\overline{a}_2, \overline{b}_2\right) &= \bigstar\left(\frac{n \times \left|\overline{a}_1 \cap \overline{b}_1\right|}{n \times \left|\overline{a}_1\right|}, \frac{n \times \left|\overline{a}_1 \cap \overline{b}_1\right|}{n \times \left|\overline{b}_1\right|}\right) \\ &= \bigstar\left(\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a}_1\right|}, \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|}\right). \end{aligned}$$

Hence, $S_{S_h}\left(\overline{a}_1, \overline{b}_1\right) = S_{S_h}\left(\overline{a}_2, \overline{b}_2\right)$. Here, the new similarity measure follows the axiom [F8]. ■

Theorem 3.11 (Increased-overlap): For two interval pairs $\{\overline{a}_1, \overline{b}_1\}$ and $\{\overline{a}_2, \overline{b}_2\}$ where $\left|\overline{a}_1\right| = \left|\overline{b}_1\right| = \left|\overline{a}_2\right| = \left|\overline{b}_2\right|$, $S_{S_h}\left(\overline{a}_1, \overline{b}_1\right) < S_{S_h}\left(\overline{a}_2, \overline{b}_2\right)$ if $\left|\overline{a}_1 \cap \overline{b}_1\right| < \left|\overline{a}_2 \cap \overline{b}_2\right|$.

Proof: For the pair $\{\overline{a}_1, \overline{b}_1\}$, $S_h\left(\overline{a}_1, \overline{b}_1\right) = \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a}_1\right|}$ and $S_h\left(\overline{b}_1, \overline{a}_1\right) = \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|}$. Thus,

$$S_{S_h}\left(\overline{a}_1, \overline{b}_1\right) = \bigstar\left(\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a_1}\right|}, \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|}\right).$$

For the pair $\{\overline{a}_2, \overline{b}_2\}$, $S_h\left(\overline{a}_2, \overline{b}_2\right) = \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{a}_2\right|}$ and $S_h\left(\overline{b}_2, \overline{a}_2\right) = \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{b}_2\right|}$. Thus,

$$S_{S_h}\left(\overline{a}_2, \overline{b}_2\right) = \bigstar\left(\frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{a}_2\right|}, \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{b}_2\right|}\right).$$

As $\left|\overline{a}_1\right| = \left|\overline{b}_1\right| = \left|\overline{a}_2\right| = \left|\overline{b}_2\right|$, therefore,

$$S_{S_h}\left(\overline{a}_2, \overline{b}_2\right) \equiv \bigstar \left(\frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{a}_1\right|}, \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{b}_1\right|} \right).$$

Given that $\left|\overline{a}_1 \cap \overline{b}_1\right| < \left|\overline{a}_2 \cap \overline{b}_2\right|$, it implies that $\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a}_1\right|} < \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{a}_1\right|}$ and $\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|} < \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{b}_1\right|}$.

Therefore,

$$\bigstar \left(\frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{a}_1\right|}, \frac{\left|\overline{a}_1 \cap \overline{b}_1\right|}{\left|\overline{b}_1\right|} \right) < \bigstar \left(\frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{a}_1\right|}, \frac{\left|\overline{a}_2 \cap \overline{b}_2\right|}{\left|\overline{b}_1\right|} \right).$$

Hence, $S_{S_h}\left(\overline{a}_1, \overline{b}_1\right) < S_{S_h}\left(\overline{a}_2, \overline{b}_2\right)$. Here, the S_{S_h} similarity measure addresses the axiom [F9]. ■

Theorem 3.12:

$S_{S_h}\left(\overline{a}, \overline{b}\right)$ is bounded by the Jaccard and Dice similarity measures when $\bigstar$ is the minimum ($\wedge$) t-norm. That is,

$$S_J\left(\overline{a}, \overline{b}\right) \leq S_{S_h \wedge}\left(\overline{a}, \overline{b}\right) \leq S_D\left(\overline{a}, \overline{b}\right).$$

Proof: For the interval pair $\{\overline{a}, \overline{b}\}$, consider the formulations of the similarity measures at Eqs. (2.19), (2.23), and (3.3). To prove this theorem, we consider four cases: 1. $\overline{a} = \overline{b}$, 2. $\overline{a} \cap \overline{b} = \emptyset$, 3. $\overline{a} \subset \overline{b}$, and 4. $\overline{a} \cap \overline{b} \neq \emptyset$, and $\overline{a} \not\subset \overline{b}$ and $\overline{b} \not\subset \overline{a}$.

Case 1: If $\overline{a} = \overline{b}$, then all three measures yield a similarity of 1. That is, $S_J\left(\overline{a}, \overline{b}\right) = S_D\left(\overline{a}, \overline{b}\right) = S_{S_h \wedge}\left(\overline{a}, \overline{b}\right) = 1$.

Case 2: If $\overline{a} \cap \overline{b} = \emptyset$, all three measures result in a similarity of 0. Thus, $S_J\left(\overline{a}, \overline{b}\right) = S_D\left(\overline{a}, \overline{b}\right) = S_{S_h \wedge\left(\overline{a}, \overline{b}\right)} = 0$.

Case 3: If $\overline{a} \subset \overline{b}$ (subset), then $\left|\overline{a} \cap \overline{b}\right| = \left|\overline{a}\right|$. With respect to $\overline{b}$, there is no non-overlap segment of $\overline{a}$; hence, $\left|\overline{a} \backslash \overline{b}\right| = 0$. Inversely, there is a non-overlap segment of $\overline{b}$ as to $\overline{a}$; thus, $\left|\overline{b} \backslash \overline{a}\right| \neq 0$. In this case, the three SMs can be simplified to

$$
\begin{aligned}
S_J\left(\overline{a}, \overline{b}\right) &= \frac{\left|\overline{a}\right|}{\left|\overline{a}\right| + \left|\overline{b} \backslash \overline{a}\right|}, \\
S_D\left(\overline{a}, \overline{b}\right) &= \frac{\left|\overline{a}\right|}{\left|\overline{a}\right| + \frac{1}{2}\left|\overline{b} \backslash \overline{a}\right|}, \\
S_{S_h \wedge}\left(\overline{a}, \overline{b}\right) &= \wedge\left(\frac{\left|\overline{a}\right|}{\left|\overline{a}\right|}, \frac{\left|\overline{a}\right|}{\left|\overline{a}\right| + \left|\overline{b} \backslash \overline{a}\right|}\right) = \frac{\left|\overline{a}\right|}{\left|\overline{a}\right| + \left|\overline{b} \backslash \overline{a}\right|},
\end{aligned}
$$

which implies that

$$
S_J\left(\overline{a}, \overline{b}\right) = S_{S_h \wedge}\left(\overline{a}, \overline{b}\right) < S_D\left(\overline{a}, \overline{b}\right).
$$

Case 4: If $\overline{a} \cap \overline{b} \neq \emptyset$, and $\overline{a} \not\subset \overline{b}$ and $\overline{b} \not\subset \overline{a}$ ($\overline{a}$ and $\overline{b}$ are partially overlapping), then assume the case $\left|\overline{a}\right| \leq \left|\overline{b}\right|$. It implies that $\left|\overline{a}\right| - |\overline{a} \cap \overline{b}| \leq \left|\overline{b}\right| - |\overline{a} \cap \overline{b}| \implies \left|\overline{a} \backslash \overline{b}\right| \leq \left|\overline{b} \backslash \overline{a}\right|$. The three similarity measures are –

$$
\begin{aligned}
S_J\left(\overline{a}, \overline{b}\right) &= \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{a} \backslash \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|}, \\
S_D\left(\overline{a}, \overline{b}\right) &= \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \frac{1}{2}\left(\left|\overline{a} \backslash \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|\right)}, \\
S_{S_h \wedge}(\overline{a}, \overline{b}) &= \wedge\left(\frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{a} \backslash \overline{b}\right|}, \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|}\right) \\
&= \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|}, \text{ as } \left|\overline{a} \backslash \overline{b}\right| \leq \left|\overline{b} \backslash \overline{a}\right|.
\end{aligned}
$$

It is true that

$$\frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{a} \backslash \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|} < \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \frac{1}{2}\left(\left|\overline{a} \backslash \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|\right)},$$

thus $S_J\left(\overline{a}, \overline{b}\right) < S_D\left(\overline{a}, \overline{b}\right)$.

Again, it is clear that

$$\frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{a} \backslash \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|} < \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|},$$

implying that $S_J\left(\overline{a}, \overline{b}\right) < S_{S_{h\wedge}}\left(\overline{a}, \overline{b}\right)$. Also,

$$\frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \left|\overline{b} \backslash \overline{a}\right|} = \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \frac{1}{2}\left|\overline{b} \backslash \overline{a}\right| + \frac{1}{2}\left|\overline{b} \backslash \overline{a}\right|} \leq \frac{\left|\overline{a} \cap \overline{b}\right|}{\left|\overline{a} \cap \overline{b}\right| + \frac{1}{2}\left|\overline{a} \backslash \overline{b}\right| + \frac{1}{2}\left|\overline{b} \backslash \overline{a}\right|},$$

$$\text{as } \left|\overline{a} \backslash \overline{b}\right| \leq \left|\overline{b} \backslash \overline{a}\right|,$$

indicating that $S_{S_{h\wedge}}\left(\overline{a}, \overline{b}\right) \leq S_D\left(\overline{a}, \overline{b}\right)$. Hence, $S_J\left(\overline{a}, \overline{b}\right) < S_{S_{h\wedge}}\left(\overline{a}, \overline{b}\right) \leq S_D\left(\overline{a}, \overline{b}\right)$. Note that for $\left|\overline{b}\right| \leq \left|\overline{a}\right|$, the same procedure can be used to prove this relation. ■

The next section introduces a new distance measure which is derived from the proposed S_{S_h} similarity measure, enabling its direct use in classification and clustering applications. Then, it discusses essential properties of the new distance measure in terms of it being a metric. In addition, the transitivity property [63] of this new distance metric is also explored.

3.4 Proposed Distance Measure

A new distance measure, $D_{S_h}(\overline{a},\overline{b})$, can easily be derived from the proposed similarity measure (S_{S_h}) at Eq. (3.2) by taking its complement, capturing the dissimilarity between both intervals,

$$D_{S_h}\left(\overline{a},\overline{b}\right) = 1 - S_{S_h}\left(\overline{a},\overline{b}\right). \tag{3.4}$$

Alternatively, Eq. (3.4) can be written as

$$\begin{aligned} D_{S_h}(\overline{a},\overline{b}) &= 1 - \bigstar\left(\frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{a}\backslash\overline{b}\right|}, \frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{b}\backslash\overline{a}\right|}\right) \\ &= \bigstar\left(1 - \frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{a}\backslash\overline{b}\right|}, 1 - \frac{\left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{b}\backslash\overline{a}\right|}\right) \\ &= \bigstar\left(\frac{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{a}\backslash\overline{b}\right| - \left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{a}\backslash\overline{b}\right|}, \frac{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{b}\backslash\overline{a}\right| - \left|\overline{a}\cap\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{b}\backslash\overline{a}\right|}\right) \\ &= \bigstar\left(\frac{\left|\overline{a}\backslash\overline{b}\right|}{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{a}\backslash\overline{b}\right|}, \frac{\left|\overline{b}\backslash\overline{a}\right|}{\left|\overline{a}\cap\overline{b}\right| + \left|\overline{b}\backslash\overline{a}\right|}\right). \end{aligned}$$

It is noted that this alternative form of the proposed $D_{S_h}\left(\overline{a},\overline{b}\right)$ measure can now directly be used in pattern recognition problems with sets/intervals such as classification and clustering. Next, the essential properties of the $D_{S_h}\left(\overline{a},\overline{b}\right)$ measure are discussed in terms of it being a metric.

Theorem 3.13:
$D_{S_h}\left(\overline{a},\overline{b}\right) = 1 - S_{S_h}\left(\overline{a},\overline{b}\right)$ is a metric.

Proof: To prove that the proposed distance measure, $D_{S_h}\left(\overline{a},\overline{b}\right)$ is a metric, we need to show that it satisfies all four properties [B1] - [B4] of a distance metric (discussed in

Chapter 2) for the intervals $\overline{a}$, $\overline{b}$, and $\overline{c}$ where $\bigstar$ is the minimum ($\wedge$) or product (Π) *t-norm*. That is,

(1) $D_{S_h}\left(\overline{a},\overline{b}\right) \geq 0$ (non-negativity);

(2) $D_{S_h}\left(\overline{a},\overline{b}\right) = D_{S_h}\left(\overline{b},\overline{a}\right)$ (symmetry);

(3) $D_{S_h}\left(\overline{a},\overline{b}\right) = 0 \iff \overline{a} = \overline{b}$ (identity);

(4) $D_{S_h}\left(\overline{a},\overline{c}\right) \leq D_{S_h}\left(\overline{a},\overline{b}\right) + D_{S_h}\left(\overline{b},\overline{c}\right)$ (triangle inequality).

Below the necessary proofs for all the above properties (a–d) are provided.

(1) By Theorem 3.3, $S_{S_h}\left(\overline{a},\overline{b}\right)$ is always bounded by $[0,1]$. As $D_{S_h}\left(\overline{a},\overline{b}\right) = 1 - S_{S_h}\left(\overline{a},\overline{b}\right)$, it implies that $0 \leq D_{S_h}\left(\overline{a},\overline{b}\right) \leq 1$, thus meeting the non-negativity property.

(2) By Theorem 3.4, $S_{S_h}\left(\overline{a},\overline{b}\right)$ is always symmetric; thus, $D_{S_h}\left(\overline{a},\overline{b}\right)$ is also symmetric.

(3) Theorem 3.5 shows that if $\overline{a} = \overline{b}$, then $S_{S_h}\left(\overline{a},\overline{b}\right)$ is always 1. Hence, $1 - S_{S_h}\left(\overline{a},\overline{b}\right)$ makes $D_{S_h}\left(\overline{a},\overline{b}\right) = 0$. Alternatively, If $\overline{a} \neq \overline{b}$, then $S_{S_h}\left(\overline{a},\overline{b}\right) \neq 1$, which implies that $D_{S_h}\left(\overline{a},\overline{b}\right) \neq 0$.

(4) For the intervals $\overline{a}$, $\overline{b}$, and $\overline{c}$, we consider $\overline{a} \subseteq \overline{b} \subseteq \overline{c}$. It implies that $\left|\overline{a}\right| \leq \left|\overline{b}\right| \leq \left|\overline{c}\right|$. To prove this theorem, the formulation of the S_{S_h} measure at Eq. (3.3).

Case 1: When all three intervals are equal $\left(\overline{a} = \overline{b} = \overline{c}\right)$, $S_{S_h}\left(\overline{a},\overline{b}\right) = S_{S_h}\left(\overline{b},\overline{c}\right) = S_{S_h}\left(\overline{a},\overline{c}\right) = 1$. Accordingly, $D_{S_h}\left(\overline{a},\overline{b}\right) = D_{S_h}\left(\overline{b},\overline{c}\right) = D_{S_h}\left(\overline{a},\overline{c}\right) = 0$, thus satisfying the triangle inequality.

Case 2: When $\overline{a} \subset \overline{b} \subset \overline{c}$, it implies that $\left|\overline{a}\right| < \left|\overline{b}\right| < \left|\overline{c}\right|$. Now, the similarity and distance for each pair of intervals can be calculated by applying Eq. (3.3):

(i) For $\overline{a} \subset \overline{b}$,

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

$$\therefore\ D_{S_h}\left(\overline{a},\overline{b}\right) = 1 - S_{S_h}\left(\overline{a},\overline{b}\right) = 1 - \frac{\left|\overline{a}\right|}{\left|\overline{b}\right|}.$$

(ii) For $\overline{b} \subset \overline{c}$,

$$S_{S_h}\left(\overline{b},\overline{c}\right) = \frac{\left|\overline{b}\right|}{\left|\overline{c}\right|}$$

$$\therefore\ D_{S_h}\left(\overline{b},\overline{c}\right) = 1 - S_{S_h}\left(\overline{b},\overline{c}\right) = 1 - \frac{\left|\overline{b}\right|}{\left|\overline{c}\right|}.$$

(iii) For $\overline{a} \subset \overline{c}$ (by the transitive property of subsets [74]),

$$S_{S_h}\left(\overline{a},\overline{c}\right) = \frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}$$

$$\therefore\ D_{S_h}\left(\overline{a},\overline{c}\right) = 1 - S_{S_h}\left(\overline{a},\overline{c}\right) = 1 - \frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}.$$

Now, it is to be proved that $D_{S_h}\left(\overline{a},\overline{c}\right) < D_{S_h}\left(\overline{a},\overline{b}\right) + D_{S_h}\left(\overline{b},\overline{c}\right)$.

By placing the distance of each pair of intervals in the above equation, the following relation is received.

$$
\begin{aligned}
1-\frac{\left|\overline{a}\right|}{\left|\overline{c}\right|} &< 1-\frac{\left|\overline{a}\right|}{\left|\overline{b}\right|}+1-\frac{\left|\overline{b}\right|}{\left|\overline{c}\right|} \\
\iff 1-\frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}-1+\frac{\left|\overline{b}\right|}{\left|\overline{c}\right|} &< 1-\frac{\left|\overline{a}\right|}{\left|\overline{b}\right|} \\
\iff -\frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}+\frac{\left|\overline{b}\right|}{\left|\overline{c}\right|} &< 1-\frac{\left|\overline{a}\right|}{\left|\overline{b}\right|} \\
\iff \frac{\left|\overline{b}\right|-\left|\overline{a}\right|}{\left|\overline{c}\right|} &< \frac{\left|\overline{b}\right|-\left|\overline{a}\right|}{\left|\overline{b}\right|} \\
\iff \frac{1}{\left|\overline{c}\right|} &< \frac{1}{\left|\overline{b}\right|}.
\end{aligned}
$$

It is true as $\left|\overline{b}\right| < \left|\overline{c}\right|$, thereby satisfying the triangle inequality. Thus, $D_{S_h}\left(\overline{a},\overline{c}\right) \leq D_{S_h}\left(\overline{a},\overline{b}\right)+D_{S_h}\left(\overline{b},\overline{c}\right)$. ■

Example 3.1 (Example Satisfying Triangle Inequality):

Consider three intervals $\overline{a}$, $\overline{b}$, and $\overline{c}$ in Fig. 3.2 where $\overline{b},\overline{c} \subset \overline{a}$ and $\overline{b} \cap \overline{c} = \emptyset$. Table 3.1 presents the similarity and distance between each pair of the three intervals using the S_{S_h} and D_{S_h} measures respectively with the minimum ($\wedge$) and product (Π) *t-norm*s. Note that the same similarity and distance results are received using both *t-norm*s. From the distance results, it is observed that for the interval pair $\{\overline{a},\overline{c}\}$, their distance—which is 0.58—is less than the summed distances of the pairs $\{\overline{a},\overline{b}\}$ and $\{\overline{b},\overline{c}\}$—which is 1.50—implying $D_{S_h}\left(\overline{a},\overline{c}\right) < D_{S_h}\left(\overline{a},\overline{b}\right)+D_{S_h}\left(\overline{b},\overline{c}\right)$. This relation is also maintained for the interval pairs $\{\overline{a},\overline{b}\}$ and $\{\overline{b},\overline{c}\}$, thereby demonstrating the D_{S_h} measure meeting the triangle inequality.

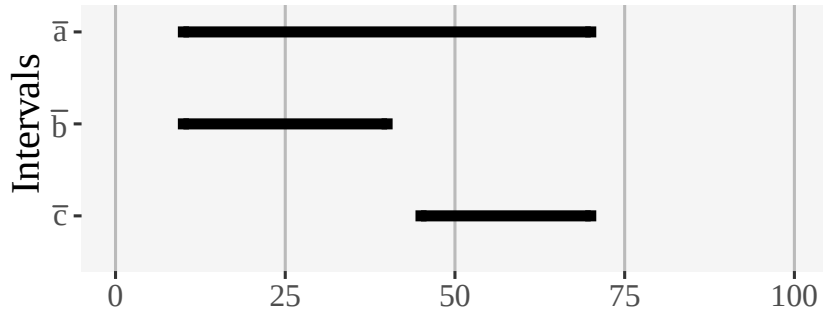


Fig. 3.2 Example that shows to satisfy the triangle inequality by the D_{S_h} measure.

Table 3.1 Similarity and distance between the intervals of Fig. 3.2.

Interval Pair	$\mathbf{S_{S_h\wedge}}$	$\mathbf{S_{S_h\Pi}}$	$\mathbf{D_{S_h\wedge}}$	$\mathbf{D_{S_h\Pi}}$
$\{\overline{a}, \overline{b}\}$	0.50	0.50	0.50	0.50
$\{\overline{a}, \overline{c}\}$	0.42	0.42	0.58	0.58
$\{\overline{b}, \overline{c}\}$	0.0	0.0	1.0	1.0

Theorem 3.14:

$D_{S_h}\left(\overline{a}, \overline{b}\right)$ follows the property of transitivity [63]. That is, $D_{S_h}\left(\overline{a}, \overline{b}\right) \leq D_{S_h}\left(\overline{a}, \overline{c}\right)$ when $\overline{a} \subseteq \overline{b} \subseteq \overline{c}$.

Proof: When $\overline{a} \subseteq \overline{b} \subseteq \overline{c}$, then

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

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

Thus,

$$D_{S_h}\left(\overline{a}, \overline{b}\right) = 1 - S_{S_h}\left(\overline{a}, \overline{b}\right) = 1 - \frac{\left|\overline{a}\right|}{\left|\overline{b}\right|},$$

$$D_{S_h}\left(\overline{a},\overline{c}\right) = 1 - S_{S_h}\left(\overline{a},\overline{c}\right) = 1 - \frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}.$$

As $\overline{b} \subseteq \overline{c}$, it follows that $\left|\overline{b}\right| \leq \left|\overline{c}\right|$. Therefore, $\frac{\left|\overline{a}\right|}{\left|\overline{b}\right|} \geq \frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}$ which implies that

$$1 - \frac{\left|\overline{a}\right|}{\left|\overline{b}\right|} \leq 1 - \frac{\left|\overline{a}\right|}{\left|\overline{c}\right|}.$$

Hence $D_{S_h}\left(\overline{a},\overline{b}\right) \leq D_{S_h}\left(\overline{a},\overline{c}\right)$. ■

The next section demonstrates the performances of the new similarity measure vis-a-vis popular set-theoretic similarity measures using both synthetic and real world cases.

3.5 Demonstration and Analysis

This section demonstrates and analyses the behaviour of the proposed similarity measure (S_{S_h}) with the minimum ($\wedge$) and product (Π) *t-norm*s in the context of the Jaccard (S_J) and Dice (S_D) measures for a set of synthetic examples in the first part and with a real dataset in the last part. Here, different features of interval-pairs are altered to explore how well these similarity measures perform or follow intuitive results. In particular, their performance are analyzed by focusing on the following aspects:

1. Propensity to exhibiting *aliasing* in response to variations in interval sizes.
2. Behaviour in respect to intervals where one is a complete subset of the other.
3. Behaviour in respect to intervals of equal sizes and overlapping ratio.
4. Response to variations in interval size, while maintaining the same level of subsethood.

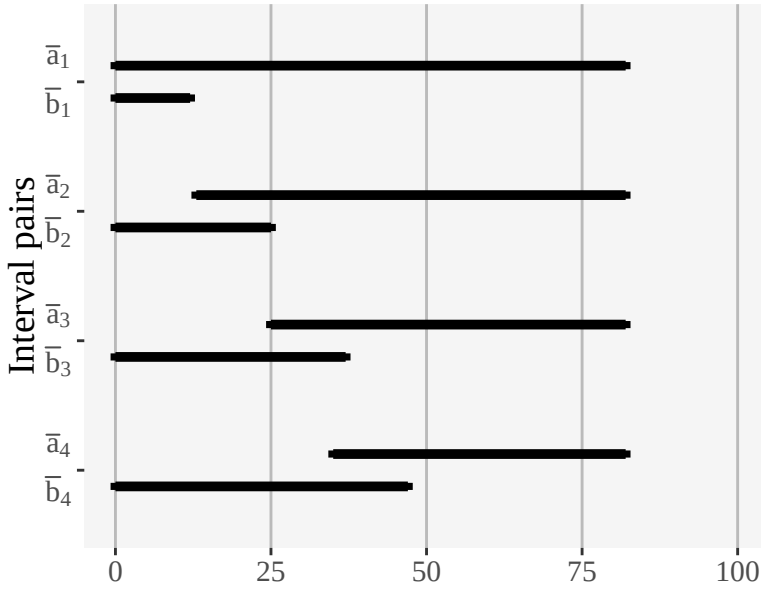


Fig. 3.3 Interval pairs used to demonstrate the *aliasing* of similarity measures.

Table 3.2 The S_J and S_D measures exhibit *aliasing* for the interval pairs in Fig. 3.3.

Interval Pair	$\mathbf{S_J}$	$\mathbf{S_D}$	$\mathbf{S_{S_h\wedge}}$	$\mathbf{S_{S_h\Pi}}$
$\{\bar{a}_1, \bar{b}_1\}$	0.15	0.26	0.15	0.15
$\{\bar{a}_2, \bar{b}_2\}$	0.15	0.26	0.17	0.08
$\{\bar{a}_3, \bar{b}_3\}$	0.15	0.26	0.21	0.07
$\{\bar{a}_4, \bar{b}_4\}$	0.15	0.26	0.26	0.06

5. Response to linear increase in the overlap of intervals.

In addition, all experiments are implemented using Java on an Intel(R) Core(TM) i3-4005U series based machine running at 1.70 GHz with 8GB RAM.

3.5.1 Synthetic Dataset Based Demonstration

For each of the above cases, a series of synthetic intervals is proposed and visualized.

3.5.1.1 Experiment on aliasing propensity

In Fig. 3.3, four different pairs of intervals $\{\bar{a}, \bar{b}\}$ are considered where all pairs have an intersection of equal size. The similarity results for the pairs using the three similarity

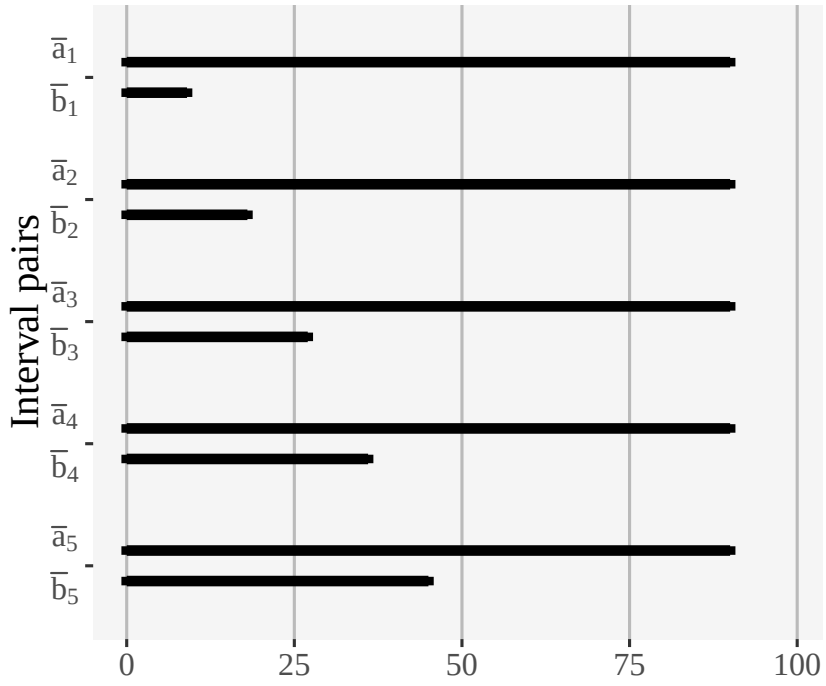
measures are shown in Table 3.2. The S_J and S_D measures are subject to the *aliasing* by providing the same similarity of 0.15 and 0.26 respectively for all pairs.

Indeed, both measures provide—perhaps unexpectedly—identical similarities for pairs of intervals when the size of the union of their non-overlapping segments remains constant. On the contrary, the S_{S_h} measure with both minimum and product *t-norms* ($S_{S_h\wedge}$ and $S_{S_h\Pi}$) yields a different degree of similarity for all cases, thereby exhibiting its *aliasing*-free behaviour. The reason is that the S_{S_h} measure better captures the changes in the size of both input intervals as compared to S_J and S_D measures, which eventually affects to their reciprocal subsethood and the overall similarity. Note that as shown in Theorem 3.12, the results of the $S_{S_h\wedge}$ measure are bounded by the S_J and S_D measures.

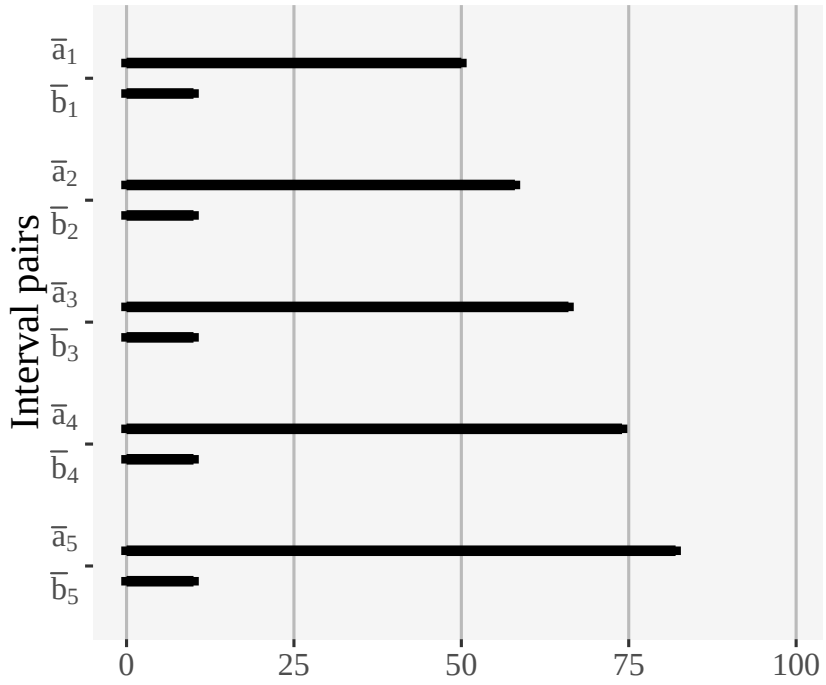
3.5.1.2 Experiment with interval pairs when one interval is a complete subset of the other

Two separate cases are considered with different sets of interval pairs, where in both cases one interval is a complete subset of the other: (a) an increasing degree of overlap between both intervals by increasing the size of the smaller interval, and (b) decreasing the degree of overlap by increasing the size of the larger interval.

Figure 3.4(a) shows five interval pairs with $\overline{b} \subset \overline{a}$, where $\overline{b}$ covers 10%, 20%, 30%, 40%, and 50% of $\overline{a}$. On the contrary, in Fig. 3.4(b) another five interval pairs are presented where the degree of overlap (in %) is gradually decreased by increasing the size of $\overline{a}$. Tables 3.3(a) and 3.3(b) show the similarity for all pairs with all three similarity measures. Note that for all pairs, the subsethood of $\overline{b}$ is 1 while the subsethood of $\overline{a}$ depends on the size of $\overline{a}$ and $\overline{b}$, i.e., $\frac{|\overline{a}\cap\overline{b}|}{|\overline{a}|}$. Intuitively, their mutual similarity should be at most $\frac{|\overline{a}\cap\overline{b}|}{|\overline{a}|}$ for each pair. From the results, it is observed that both S_{S_h} (with



(a) Increasing interval overlapping



(b) Increasing interval non-overlapping

Fig. 3.4 Interval pairs where one interval as a complete subset of the other.

Table 3.3 Similarity results for the interval pairs as shown in Fig. 3.4.

Interval Pair	$\mathbf{S_J}$	$\mathbf{S_D}$	$\mathbf{S_{S_h\wedge}}$	$\mathbf{S_{S_h\Pi}}$	**subset by**
$\{\overline{a}_1, \overline{b}_1\}$	0.10	0.18	0.10	0.10	10%
$\{\overline{a}_2, \overline{b}_2\}$	0.20	0.33	0.20	0.20	20%
$\{\overline{a}_3, \overline{b}_3\}$	0.30	0.46	0.30	0.30	30%
$\{\overline{a}_4, \overline{b}_4\}$	0.40	0.57	0.40	0.40	40%
$\{\overline{a}_5, \overline{b}_5\}$	0.50	0.67	0.50	0.50	50%

(a) Increasing interval overlapping

Interval Pair	$\mathbf{S_J}$	$\mathbf{S_D}$	$\mathbf{S_{S_h\wedge}}$	$\mathbf{S_{S_h\Pi}}$	**subset by**
$\{\overline{a}_1, \overline{b}_1\}$	0.20	0.33	0.20	0.20	20%
$\{\overline{a}_2, \overline{b}_2\}$	0.172	0.294	0.172	0.172	17.2%
$\{\overline{a}_3, \overline{b}_3\}$	0.152	0.263	0.152	0.152	15.2%
$\{\overline{a}_4, \overline{b}_4\}$	0.135	0.238	0.135	0.135	13.5%
$\{\overline{a}_5, \overline{b}_5\}$	0.122	0.217	0.122	0.122	12.2%

(b) Increasing interval non-overlapping

minimum and product *t-norm*s) and S_J measures perform according to the intuition while the S_D measure exceeds this expected limit.

3.5.1.3 Experiment with interval pairs of equal size and equal overlapping ratio

In Fig. 3.5, five interval pairs are shown where the intervals are of equal size and the size of their intersection is varied to 10%, 20%, 30%, 40%, and 50% of their size. Table 3.4 provides the results for all pairs using the three similarity measures. In all pairs, the subsethood is equal, and it is intuitive to expect the similarity to be the same as this subsethood (as the intervals are both of equal size). In this case, the S_{S_h} with minimum *t-norm* ($S_{S_h\wedge}$) and S_D measures follow the intuition whereas the S_{S_h} with product *t-norm* ($S_{S_h\Pi}$) and S_J measures yield a lower degree of similarity.

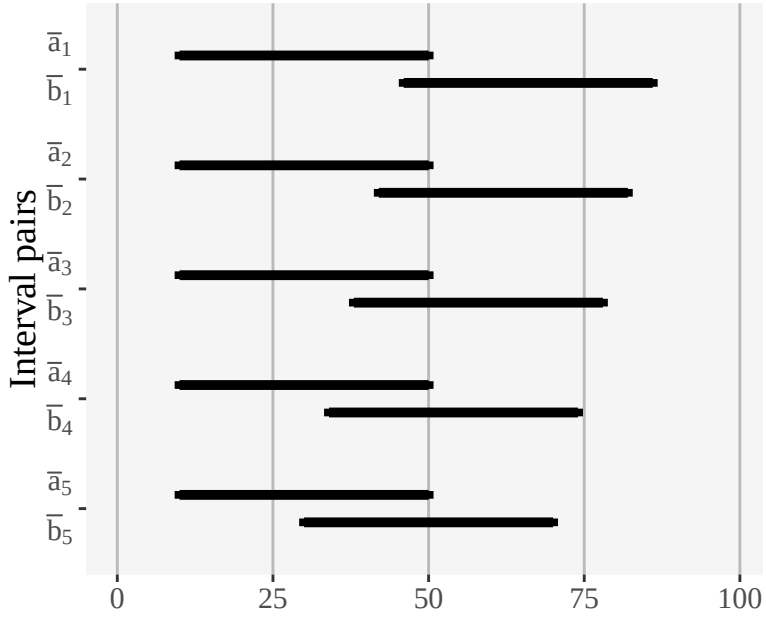


Fig. 3.5 Interval pairs with equal size and equal ratio of overlap.

Table 3.4 Similarity results for the interval pairs as shown in Fig. 3.5.

Interval Pair	$\mathbf{S_J}$	$\mathbf{S_D}$	$\mathbf{S_{S_h\wedge}}$	$\mathbf{S_{S_h\Pi}}$	**intersecting by**
$\{\overline{a}_1, \overline{b}_1\}$	0.05	0.10	0.10	0.01	10%
$\{\overline{a}_2, \overline{b}_2\}$	0.11	0.20	0.20	0.04	20%
$\{\overline{a}_3, \overline{b}_3\}$	0.18	0.30	0.30	0.09	30%
$\{\overline{a}_4, \overline{b}_4\}$	0.25	0.40	0.40	0.16	40%
$\{\overline{a}_5, \overline{b}_5\}$	0.33	0.50	0.50	0.25	50%

3.5.1.4 Experiment on scaling invariance

Five pairs of intervals are shown in Fig. 3.6, where both endpoints of $\overline{a}$ and $\overline{b}$ are gradually multiplied by a factor $n \in \{2, 3, 4, 5\}$ to generate new interval pairs. The degree of subsethood stays the same across the pairs. Adapting the definition from [15], a similarity measure is *invariant* if its similarity output remains constant regardless of scaling the interval endpoints by a factor. Table 3.5 shows the similarity for all pairs using the three similarity measures where n is the factor applied to the interval endpoints. The results demonstrate the *scaling invariance* property for the given pairs of intervals for all similarity measures.

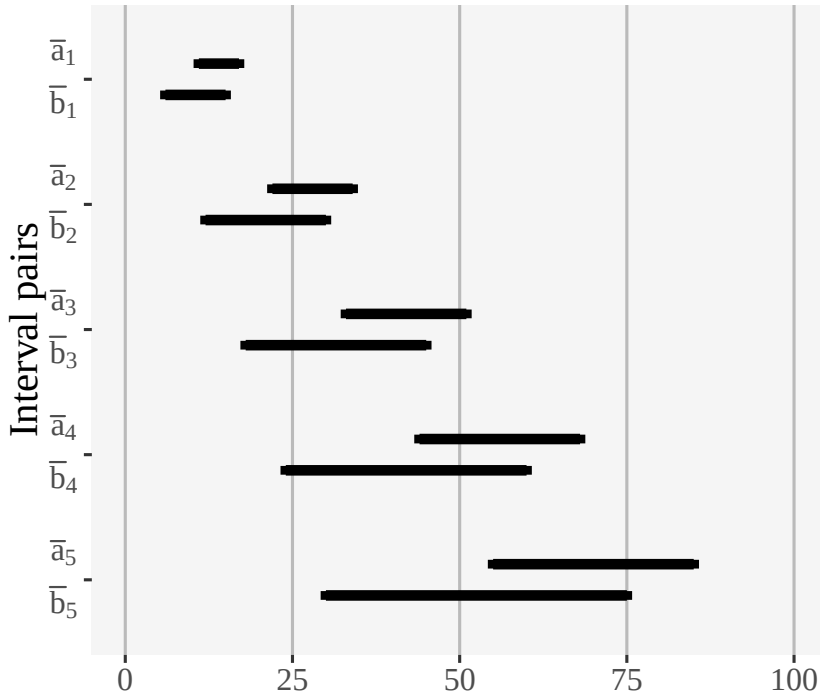


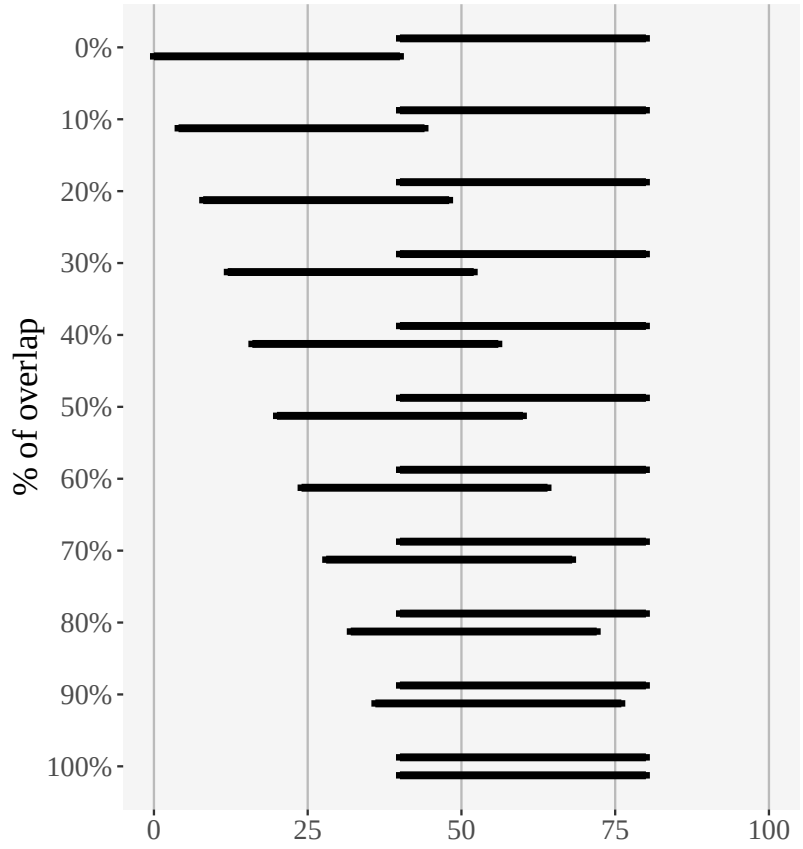
Fig. 3.6 Interval pairs used to demonstrate the invariance of similarity measures.

Table 3.5 All three measures display invariant similarity for the interval pairs in Fig. 3.6.

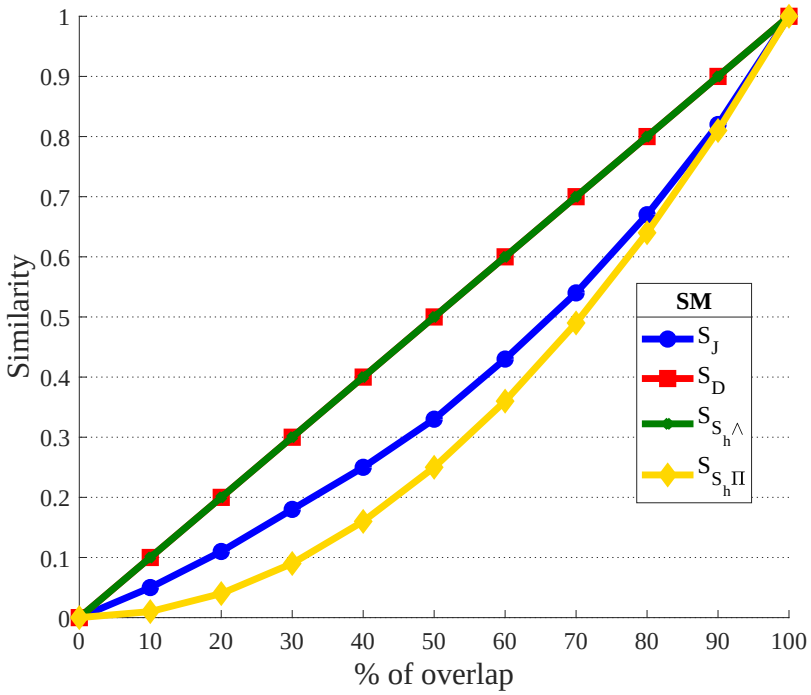
Interval Pair	$\mathbf{S_J}$	$\mathbf{S_D}$	$\mathbf{S_{S_h\wedge}}$	$\mathbf{S_{S_h\Pi}}$	**multiplied by** n
$\{\bar{a}_1, \bar{b}_1\}$	0.36	0.53	0.44	0.29	1
$\{\bar{a}_2, \bar{b}_2\}$	0.36	0.53	0.44	0.29	2
$\{\bar{a}_3, \bar{b}_3\}$	0.36	0.53	0.44	0.29	3
$\{\bar{a}_4, \bar{b}_4\}$	0.36	0.53	0.44	0.29	4
$\{\bar{a}_5, \bar{b}_5\}$	0.36	0.53	0.44	0.29	5

3.5.1.5 Experiment on increased overlap linearly

Adapting the definition from [15], a similarity measure on intervals is *linear* if its similarity output varies linearly as to a linear change in the size of the overlap/intersection of the intervals. In Fig. 3.7(a), the overlap between two intervals of equal size is gradually increased in 10% steps. The corresponding similarity outputs for the pairs and all three similarity measures are shown graphically in Fig. 3.7(b). Results show that all three similarity measures provide higher similarity for increased overlap. Particularly, the S_{S_h} with minimum *t-norm* ($S_{S_h\wedge}$) and S_D measures exhibit *linearity* in the similarity



(a) Rising overlap while intervals are of equal size



(b) Similarity versus overlap

Fig. 3.7 Displaying linearity by the S_D and $S_{S_h\wedge}$ measures, and convexity by the S_J and $S_{S_h\Pi}$ measures.

results while the S_{S_h} with product *t-norm* ($S_{S_h\Pi}$) and S_J measure display *convexity* (similarity increases rapidly with an increase in the size of the intersection).

In summary, case 1) shows that S_J and S_D measures are subject to the *aliasing* and return the same similarity for very different interval-pairs whenever the union of their non-overlapping remains constant (regardless of any changes in their size). In case 2), when one interval is a complete subset of the other, proportionately increasing or decreasing their degree of overlap results in the S_{S_h} and S_J measures returning sensible results whereas the S_D measure overestimates the similarity. In case 3), the S_{S_h} with minimum *t-norm* ($S_{S_h\wedge}$) and S_D measures meet the expectation when intervals are of equal size and their intersection is changed proportionately, whereas the S_J measure underestimates the similarity. In case 4), we multiply interval end-points by a factor, maintaining the same degree of subsethood. Here, all three measures show *scaling invariance* in the results as expected. Lastly, in case 5), for linear increases in the size of overlap between the intervals of equal size, the S_{S_h} with minimum *t-norm* ($S_{S_h\wedge}$) and S_D measures exhibit *linearity* in the similarity results while the S_{S_h} with product *t-norm* ($S_{S_h\Pi}$) and S_J measure display *convexity*. These experiments demonstrate that the proposed similarity measure shows behaviour in line with expectation, which is not the case for the other, commonly used measures.

3.5.2 A Real World Example

In this part, a real world dataset has been used to review the behaviour of the three similarity measures (S_{S_h} with minimum ($\wedge$) and product (Π) *t-norm*s, S_J, and S_D), particularly in respect to the *aliasing* issue. The dataset used for this demonstration is the temperature data of different standard areas (districts) of the UK for the year 2016 from the UK Met Office. The entire dataset is available at [138]. The map of the standard areas used by the UK Met office is shown in Fig. 3.8. We have used

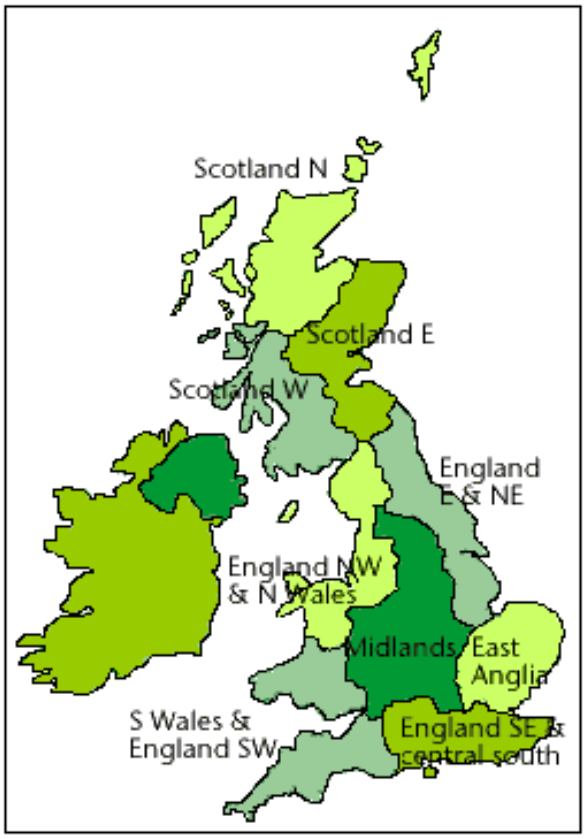


Fig. 3.8 Standard areas (districts) of the UK [139].

Table 3.6 Seasonal Temperature in Standard Areas of the UK in 2016.

Area	**Acronym**	**Seasons**							
		Winter		**Spring**		**Summer**		**Autumn**	
		Min	Max	Min	Max	Min	Max	Min	Max
Scotland N	SN	0.5	6.0	2.7	9.8	9.1	15.8	4.8	11.2
Scotland E	SE	0.4	6.2	2.4	10.3	9.1	16.9	4.4	11.5
Scotland W	SW	1.9	7.2	3.2	11.1	10.0	17.3	5.6	12.0
England E & NE	EEN	2.4	8.5	3.6	11.6	10.9	19.5	6.6	13.5
England NW & Wales N	ENW	3.1	8.6	3.9	11.8	11.4	18.7	6.5	13.1
Midlands	ML	3.3	9.4	3.9	12.6	11.7	20.6	6.9	14.1
East Anglia	EA	3.8	10.0	4.5	13.2	12.3	22.1	7.7	15.5
England SW & Wales S	ESW	4.5	9.8	4.4	12.7	11.9	19.5	7.3	14.2
England SE & Central S	ESC	4.3	10.2	4.7	13.5	12.4	21.6	7.5	15.4

temperature data of nine standard areas: Scotland North (SN), Scotland East (SE), Scotland West (SW), England East & North East (EEN), England Northwest & Wales North (ENW), Midlands (ML), East Anglia (EA), England South-west & Wales South (ESW), and England South-east & Central South (ESC) for four different seasons in 2016.

The intervals are constructed by taking the minimum and the maximum temperature of each area for every season, as shown in Table 3.6. We have applied the S_{S_h} (with minimum ($\wedge$) and product (Π) *t-norm*s), S_J and S_D measures to estimate similarities between all possible pairs of the nine areas for all seasons; here the results for Autumn and Spring seasons are reported which are shown in Tables 3.7, 3.8, 3.9 and 3.10. The results for other two seasons—Winter and Summer—are included in Appendix A.

Regarding similarity based on seasonal temperature, one can expect higher similarity for pairs of areas geographically close to each other (i.e., located in the same latitude), while expecting lower similarity for remote areas as it is generally recognized that latitude (i.e., distance from the equator) is an important factor affecting variations in temperature. Hence, in this real world example, the districts in the Northern UK (e.g., SE, SW) are likely to be more similar as they are in the South (e.g., ML, EA). By the same token, a district in the North will be highly dissimilar to a district in the South. Generally, all three similarity measures appear to meet up with such expectations. For instance, considering Autumn-temperatures between SN and the other eight areas from geographically close to far off, three similarity measures appear to generate higher to lower similarity as temperature-range differs more.

However, in several cases, the S_J and S_D measures suffer from the *aliasing*—returning the same similarity for different ranges of temperature. As mentioned earlier, both measures provide identical similarity for pairs of areas as long as the size of the intersection and union of their temperature-range remain constant. Table 3.7 and 3.8

Table 3.7 Similarity results for temperature data using the S_J measure.

	Autumn								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9014	0.7778	0.5287	0.5663	0.4624	0.3271	0.4149	0.3491
SE	0.9014	1	0.7763	0.5385	0.5747	0.4742	0.3423	0.4286	0.3636
SW	0.7778	0.7763	1	0.6835	0.7333	0.6	0.4343	0.5465	0.4592
EEN	0.5287	0.5385	0.6835	1	0.9286	0.88	0.6517	0.8158	0.6818
ENW	0.5663	0.5747	0.7333	0.9286	1	0.8158	0.6	0.7532	0.6292
ML	0.4624	0.4742	0.6	0.88	0.8158	1	0.7442	0.9315	0.7765
EA	0.3271	0.3423	0.4343	0.6517	0.6	0.7442	1	0.7927	0.9625
ESW	0.4149	0.4286	0.5465	0.8158	0.7532	0.9315	0.7927	1	0.8272
ESC	0.3491	0.3636	0.4592	0.6818	0.6292	0.7765	0.9625	0.8272	1

(a) For autumn season

	Spring								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.8987	0.7857	0.6966	0.6484	0.596	0.5048	0.54	0.4722
SE	0.8987	1	0.8161	0.7283	0.6809	0.6275	0.537	0.5728	0.5045
SW	0.7857	0.8161	1	0.8929	0.8372	0.766	0.66	0.7053	0.6214
EEN	0.6966	0.7283	0.8929	1	0.939	0.8556	0.7396	0.7912	0.697
ENW	0.6484	0.6809	0.8372	0.939	1	0.908	0.7849	0.8409	0.7396
ML	0.596	0.6275	0.766	0.8556	0.908	1	0.871	0.9318	0.8229
EA	0.5048	0.537	0.66	0.7396	0.7849	0.871	1	0.9318	0.9444
ESW	0.54	0.5728	0.7053	0.7912	0.8409	0.9318	0.9318	1	0.8791
ESC	0.4722	0.5045	0.6214	0.697	0.7396	0.8229	0.9444	0.8791	1

(b) For spring season

Table 3.8 Similarity results for temperature data using the S_D measure.

	Autumn								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9481	0.875	0.6917	0.7231	0.6324	0.493	0.5865	0.5175
SE	0.9481	1	0.8741	0.7	0.7299	0.6434	0.5101	0.6	0.5333
SW	0.875	0.8741	1	0.812	0.8462	0.75	0.6056	0.7068	0.6294
EEN	0.6917	0.7	0.812	1	0.963	0.9362	0.7891	0.8986	0.8108
ENW	0.7231	0.7299	0.8462	0.963	1	0.8986	0.75	0.8593	0.7724
ML	0.6324	0.6434	0.75	0.9362	0.8986	1	0.8533	0.9645	0.8742
EA	0.493	0.5101	0.6056	0.7891	0.75	0.8533	1	0.8844	0.9809
ESW	0.5865	0.6	0.7068	0.8986	0.8593	0.9645	0.8844	1	0.9054
ESC	0.5175	0.5333	0.6294	0.8108	0.7724	0.8742	0.9809	0.9054	1

(a) For autumn season

	Spring								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9467	0.88	0.8212	0.7867	0.7468	0.6709	0.7013	0.6415
SE	0.9467	1	0.8987	0.8428	0.8101	0.7711	0.6988	0.7284	6707
SW	0.88	0.8987	1	0.9434	0.9114	0.8675	0.7952	0.8272	0.7665
EEN	0.8212	0.8428	0.9434	1	0.9686	0.9222	0.8503	0.8834	0.8214
ENW	0.7867	0.8101	0.9114	0.9686	1	0.9518	0.8795	0.9136	0.8503
ML	0.7468	0.7711	0.8675	0.9222	0.9518	1	0.931	0.9647	0.9029
EA	0.6709	0.6988	0.7952	0.8503	0.8795	0.931	1	0.9647	0.9714
ESW	0.7013	0.7284	0.8272	0.8834	0.9136	0.9647	0.9647	1	0.9357
ESC	0.6415	0.6707	0.7665	0.8214	0.8503	0.9029	0.9714	0.9357	1

(b) For spring season

Table 3.9 Similarity results for temperature data using the $S_{S_h\wedge}$ measure.

	Autumn								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9014	0.875	0.6667	0.7121	0.5972	0.4487	0.5652	0.4684
SE	0.9014	1	0.831	0.6901	0.7042	0.6389	0.4872	0.5915	0.5063
SW	0.875	0.831	1	0.7826	0.8333	0.7083	0.5513	0.6812	0.5696
EEN	0.6667	0.6901	07826	1	0.942	0.9167	0.7436	0.8986	0.7595
ENW	0.7121	0.7042	0.8333	0.942	1	0.8611	0.6923	0.8406	0.7089
ML	0.5972	0.6389	0.7083	0.9167	0.8611	1	0.8205	0.9444	0.8354
EA	0.4487	0.4872	0.5513	0.7436	0.6923	0.8205	1	0.8333	0.9747
ESW	0.5652	0.5915	0.6812	0.8986	0.8406	0.9444	0.8333	1	0.8481
ESC	0.4684	0.5063	0.5696	0.7595	0.7089	0.8354	0.9747	0.8481	1

(a) For autumn season

	Spring								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.8987	0.8354	0.775	0.7468	0.6782	0.6092	0.6506	0.5795
SE	0.8987	1	0.8987	0.8375	0.8101	0.7356	0.6667	0.7108	0.6364
SW	0.8354	0.8987	1	0.9375	0.9114	0.8276	0.7586	0.8072	0.7273
EEN	0.775	0.8375	0.9375	1	0.9625	0.8851	0.8161	0.8675	0.7841
ENW	0.7468	0.8101	0.9114	0.9625	1	0.908	0.8391	0.8916	0.8068
ML	0.6782	0.7356	0.8276	0.8851	0.908	1	0.931	0.9425	0.8977
EA	0.6092	0.6667	0.7586	0.8161	0.8391	0.931	1	0.9425	0.9659
ESW	0.6506	0.7108	0.8072	0.8675	0.8916	0.9425	0.9425	1	0.9091
ESC	0.5795	0.6364	0.7273	0.7841	0.8086	0.8977	0.9659	0.9091	1

(b) For spring season

Table 3.10 Similarity results for temperature data using the $S_{S_h\Pi}$ measure.

	Autumn								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9014	0.7656	0.4792	0.523	0.4013	0.2454	0.3444	0.2708
SE	0.9014	1	0.7661	0.4901	0.5335	0.4139	0.2607	0.3601	0.2853
SW	0.7656	0.7661	1	0.6603	0.7161	0.5645	0.3704	0.5002	0.4005
EEN	0.4792	0.4901	0.6603	1	0.9278	0.8768	0.625	0.8074	0.6604
ENW	0.523	0.5335	0.7161	0.9278	1	0.8089	0.5664	0.7387	0.6015
ML	0.4013	0.4139	0.5645	0.8768	0.8089	1	0.7293	0.9308	0.7658
EA	0.2454	0.2607	0.3704	0.625	0.5664	0.7293	1	0.785	0.9622
ESW	0.3444	0.3601	0.5002	0.8074	0.7387	0.9308	0.785	1	0.8235
ESC	0.2708	0.2853	0.4005	0.6604	0.6015	0.7658	0.9622	0.8235	1

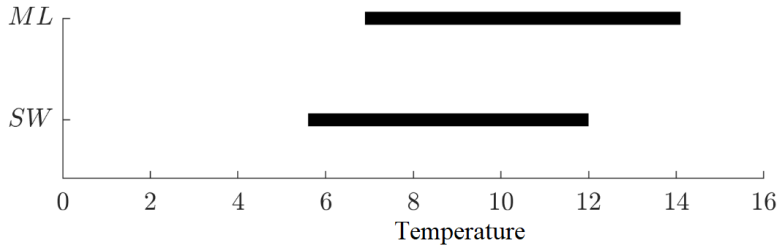
(a) For autumn season

	Spring								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.8987	0.7766	0.6768	0.6206	0.5635	0.4548	0.4948	0.4163
SE	0.8987	1	0.8077	0.7103	0.6563	0.596	0.4895	0.5309	0.4511
SW	0.7766	0.8077	1	0.89	0.8306	0.7543	0.6338	0.6846	0.5892
EEN	0.6768	0.7103	0.89	1	0.9381	0.8519	0.7243	0.7807	0.6763
ENW	0.6206	0.6563	0.8306	0.9381	1	0.908	0.7754	0.8351	0.7251
ML	0.5635	0.596	0.7543	0.8519	0.908	1	0.8668	0.9312	0.8152
EA	0.4548	0.4895	0.6338	0.7243	0.7754	0.8668	1	0.9312	0.9437
ESW	0.4948	0.5309	0.6846	0.7807	0.8351	0.9312	0.9312	1	0.8762
ESC	0.4163	0.4511	0.5892	0.6763	0.7251	0.8152	0.9437	0.8762	1

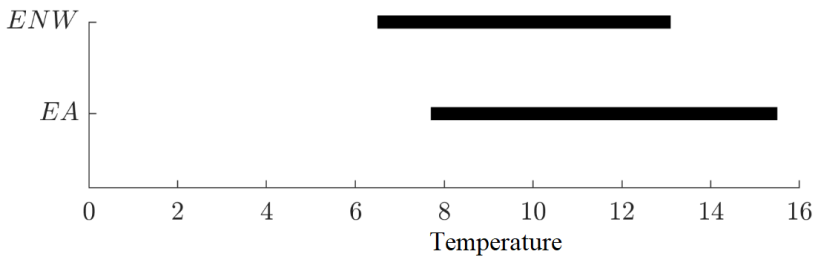
(b) For spring season

show that they produce the same similarity (i.e., 0.7396 and 0.8503 respectively) for pairs of areas, (EEN versus EA) and (ENW versus ESC) in Spring, even though they have separate temperature-ranges. Analogously, it is also noticed that identical similarities for the pairs, (SW vs ML) and (EA vs ENW), and the pairs (ML vs ENW) and (ESW vs EEN) in Autumn. Figure 3.9 presents the temperature ranges for the pairs of areas $\{SW, ML\}$ and $\{EA, ENW\}$ for which the Jaccard and Dice similarity measures face aliasing. In contrast, the S_{S_h} measure with minimum and product *t-norm*s (i.e., $S_{S_h\wedge}$ and $S_{S_h\Pi}$), responding to variation in temperature-range, yields a distinct outcome for all of these pairs as shown in Table 3.9 and 3.10.

This real world example showing *aliasing* inherent to the S_J and S_D measures highlights the potential for misleading inference in practice. In particular, while clustering the areas with respect to temperature, some areas may be placed in the same group albeit having different ranges of temperature. This can be an issue when



(a) SW versus ML



(b) EA versus ENW

Fig. 3.9 Popular similarity measures show identical similarity for two pairs of areas having different ranges of temperature in Autumn.

the real-world interval data is compared for the purpose of grouping, ranking, and decision making. Therefore, a similarity measure which avoids this *aliasing*—such as the proposed, is desirable.

3.6 Summary

This chapter has identified and articulated a potential limitation– specifically *aliasing* of popular set-theoretic similarity measures. *Aliasing* may lead to misleading results, in turn affecting real-world applications resting on robust comparison of intervals. To address this, this chapter has introduced an expanded axiomatic definition of a *robust* similarity measure by putting forward five new axioms which complement the traditional axioms (section 2.5 in Chapter 2) associated with most common similarity measures. In respect to the full set of axioms of a *robust* similarity measure, this chapter has introduced a novel set-theoretic similarity measure for continuous and closed intervals which utilized their bidirectional subsethood to compare them. The new similarity measure has shown to avoid shortcomings, such as *aliasing* while producing intuitive similarity results (e.g., being bounded above and below by Jaccard and Dice measures), thus justifying its applicability in real world applications. As part of the development of the new similarity measure, the chapter has also put forward the definition of subsethood for intervals and proofs of its mathematical properties. Finally, a corresponding dissimilarity or distance measure has been derived from the new similarity measure, which is also proven to be a metric.

At an experimental level, the chapter has provided a detailed investigation contrasting the behaviour of the proposed similarity measure vis-a-vis the Jaccard and Dice similarity measures – using both synthetic and real-world interval-valued data. The exhaustive analyses have confirmed that the new measure exhibits desirable behaviour while maintaining all essential features of a similarity measure. In particular, the new

similarity measure has been resilient to *aliasing* and has provided desirable results in respect to all key features (e.g., *linearity*, *scaling invariance*), whereas popular similarity measures have been shown to produce counter-intuitive results in some cases. Therefore, the works that have been done in this chapter fulfill objectives 1 to 4 of the thesis.

The contributions to this chapter have been published in the following international conference and journal.

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

  Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. Novel Similarity Measure for Interval-Valued Data Based on Overlapping Ratio. In International Conference on Fuzzy Systems (Fuzz-IEEE 2017), Naples, Italy, 2017.

  Status: *Published*

- **A Similarity Measure Based on Bidirectional Subsethood for Intervals**

  Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. A Similarity Measure Based on Bidirectional Subsethood for Intervals. In IEEE Transactions on Fuzzy Systems, 2019 (doi:10.1109/TFUZZ.2019.2945249).

  Status: *Accepted*

The next chapter focuses on comparing discontinuous intervals and present a generalized framework to assess their similarity as current approaches involve loss of discontinuity information and are also affected by *aliasing* of the popular measures, which ultimately affect the accuracy of similarity results. Further, this framework is made computationally more efficient by integrating it with an extension of Allen's theory [21]. Altogether, it will satisfy objective 5 of the thesis.

Chapter 4

Measuring Similarity Between Discontinuous Intervals

4.1 Introduction

Chapter 3 introduced a novel set-theoretic similarity measure for continuous and closed intervals. However, in many real-world applications, such as hazard detection [32], fusion of sensor data observed in a non-continuous space [33], temporal reasoning [21][46], and expressing natural language with temporal repetition [139] where discontinuous intervals often appear and their similarity are also assessed. Besides, in the fuzzy set theory, the α-cuts of non-convex fuzzy sets result in discontinuous intervals [20]. In such case, similarity between non-convex fuzzy sets with the α-plane decomposition relies on computation of similarity between discontinuous intervals. In this context, this chapter focuses on similarity estimation of discontinuous intervals.

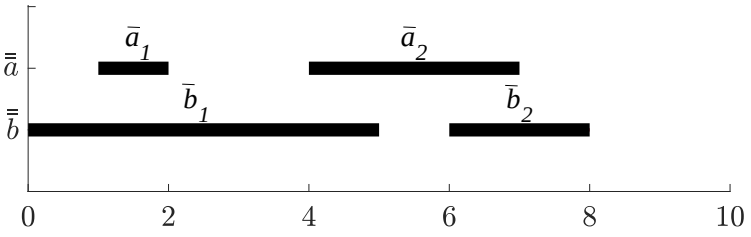
A discontinuous interval comprises a sequence of continuous subintervals [33]. For measuring similarity between discontinuous intervals, similarity measures suitable for continuous intervals can not be used directly, rather the current approach is to first convert them into their continuous form and then a similarity measure compatible

to continuous intervals is applied. This approach often involves loss of discontinuity information of intervals, thereby affecting the accuracy of similarity results. Taking this limitation into consideration, this chapter presents a generalized framework to measure degree of similarity between discontinuous intervals while maintaining their discontinuity property.

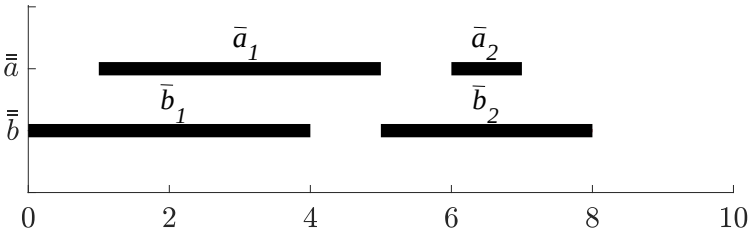
In this chapter, the motivation behind this generalization is discussed first. Then, a generalized framework is introduced for assessing the similarity between discontinuous intervals, followed by a discussion on essential properties of this framework. Subsequently, this framework is refined by combining it with an extension of Allen's theory [21] which makes the proposed framework more efficient in respect to computational effort and execution time. Lastly, the behaviour of the generalized framework is analyzed in respect to the new bidirectional subsethood based similarity measure (Eq. (3.2) in Chapter 3) together with the Jaccard and Dice measures to assess its usefulness for computing the similarity between discontinuous intervals.

4.2 Motivation

Many similarity measures have been proposed for continuous intervals where Jaccard [13] and Dice [14] are the most popular ones (discussed in Chapter 3). So far, there is no specific similarity measure for discontinuous intervals that can be directly applied to assess their similarity. Instead, discontinuous intervals are commonly converted into their continuous form by means of interval addition [34] or a 'convexify' function [44][140] (see Chapter 2), and then a respective similarity measure for continuous intervals is applied to compute their similarity. However, this 'discontinuous to continuous' conversion involves loss of discontinuity information and changes the original meaning of the data, thus affecting the accuracy of similarity result.



(a) Pair-1



(b) Pair-2

Fig. 4.1 Two different pairs of discontinuous intervals.

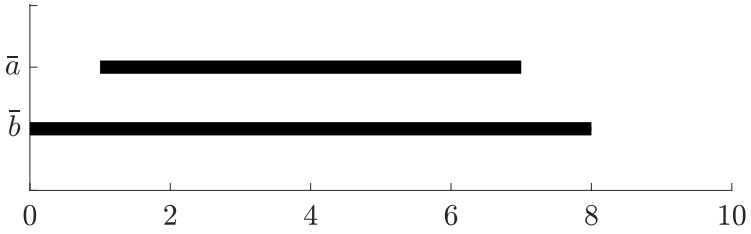
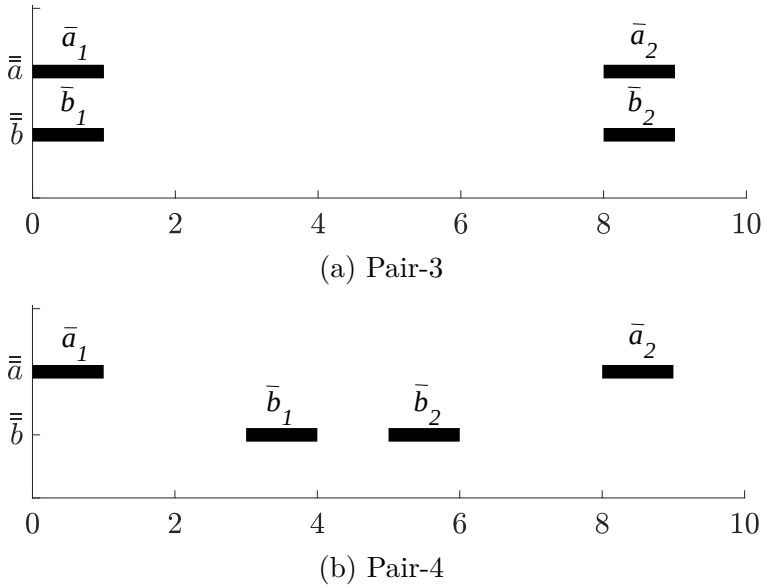


Fig. 4.2 Continuous forms of discontinuous intervals in Pair-1 and Pair-2 of Fig. 4.1 after applying the 'convexify' function (see [44]).

Figure 4.1 presents an example of two different pairs of discontinuous intervals. While using the 'convexify' function [44][140] for conversion, these two pairs of discontinuous intervals are changed into the same pair of continuous intervals, as shown in Fig. 4.2. This outcome leads to the same similarity for both pairs when the Jaccard and Dice measures are applied, which goes against intuition. For another example in Fig. 4.3, the use of interval addition [34] turns two different pairs of discontinuous intervals into the same pair of identical intervals (see Fig. 4.4), thus generating a similarity of 1 by the Jaccard and Dice measures, which also is not expected. One



(a) Pair-3

(b) Pair-4

Fig. 4.3 Another two different pairs of discontinuous intervals.

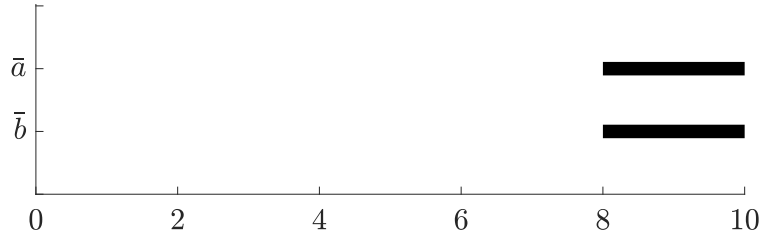


Fig. 4.4 Resulting continuous intervals after applying the interval addition (see [34]) to Pair-3 and Pair-4 of discontinuous intervals in Fig. 4.3.

way to avoid such situations is to retain discontinuity information while assessing the similarity. Therefore, this chapter puts forward a generalized framework which compares discontinuous intervals by considering all pairs of their subintervals and does not require any conversion, thus avoiding information loss.

Beyond this, there still remains a possibility to face *aliasing* with common similarity measures as identified in Chapter 3, even after all pairs of continuous subintervals within the discontinuous intervals are considered. As Chapter 3 proposed a new similarity measure for continuous intervals and demonstrated its performance in avoiding *aliasing*,

this chapter incorporates this new measure into the proposed generalized framework to assess the similarity between discontinuous intervals.

Next section presents the proposed generalized framework to compare discontinuous intervals and discusses its essential properties.

4.3 Proposed Generalized Framework

In this section, a generalized framework is introduced for computing the similarity between discontinuous intervals by considering all possible pairs of their continuous subintervals. First, the proposed framework is presented and then its major properties are demonstrated. It is worth noting that initial focus is given to the accuracy of the similarity assessment, and later this framework is made computationally more efficient.

The proposed framework uses the basic notion that as a discontinuous interval contains one or more continuous subintervals, comparing two such intervals is analogous to systematically comparing their subintervals. With this intent, all possible pairs of subintervals within the discontinuous intervals are first determined and then the similarity for each pair is computed. Later, all these similarities are aggregated to determine overall similarity between these discontinuous intervals. Equation (4.1) presents the proposed generalized framework for the discontinuous intervals $\overline{\overline{a}} = \langle \overline{a}_1, ..., \overline{a}_m \rangle$ and $\overline{\overline{b}} = \langle \overline{b}_1, ..., \overline{b}_n \rangle$,

$$S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = \frac{1}{\max(m,n)} \sum_{i=1}^{m} \sum_{j=1}^{n} S\left(\overline{a}_i, \overline{b}_j\right), \tag{4.1}$$

where $S\left(\overline{a}_i, \overline{b}_j\right)$ computes the similarity between continuous subintervals $\overline{a}_i \in \overline{\overline{a}}$ and $\overline{b}_j \in \overline{\overline{b}}$ of each pair $\{\overline{a}_i, \overline{b}_j\}$. It is noted that S can be any similarity measure defined for continuous intervals (see Chapter 3). This chapter explores the following choices for S: Jaccard, Dice, and Bidirectional subsethood based similarity measures. $\max(m,n)$ 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 discontinuous intervals are identical.

Remark 4.1:

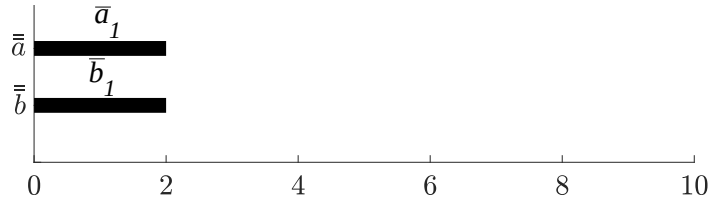
When both discontinuous intervals consist of only single continuous subinterval, the formulation for S_G at Eq. (4.1) will return the original formulation of the similarity measure (S) for continuous intervals. It implies that Eq. (4.1) turns into Eq. (3.2) when S is defined as the bidirectional subsethood based similarity measure (S_{S_h}). Again, Eq. (4.1) turns into Eqs. (2.18) and (2.20) when S is the Jaccard (S_J) and Dice (S_D) similarity measures respectively.

Remark 4.2:

An average function is not applied to aggregate similarity of all pairs of subintervals as in the normalization step it does not guarantee an overall similarity of 1 for identical discontinuous intervals. Two discontinuous intervals are said to be identical if their subintervals are also identical [40]. Figure 4.5 shows an example of three identical discontinuous interval pairs where their subintervals are gradually increased (i.e., Pair-1 has one pair of identical subintervals whereas Pair-3 contains three such pairs). Figure 4.5 also demonstrates overall similarity results for discontinuous intervals shown in Pair-1, Pair-2, and Pair-3 by using the average function. As the discontinuous intervals are identical in each pair, their similarity should be 1. However, when the average function is used to combine similarity results, it returns decreasing overall similarity for an increased number of subintervals within discontinuous intervals, which is counter-intuitive.

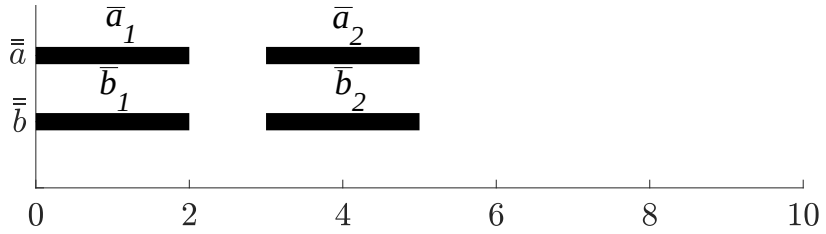
Example 4.1:

Consider the example in Fig. 4.6 for determining the similarity for a pair of discontinuous



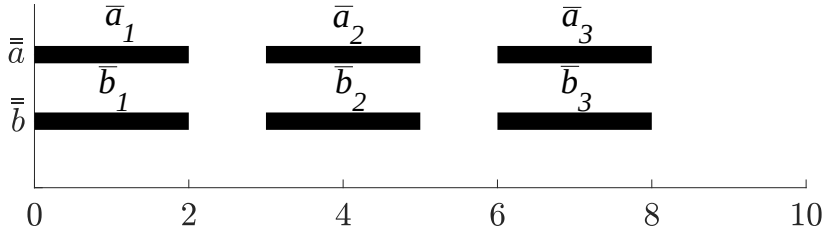
Pair of subintervals	Similarity for each subinterval pair	Total number of subinterval pairs	Overall similarity between discontinuous intervals
$\{\bar{a}_1, \bar{b}_1\}$	1.0	1	1.0

(a) Pair – 1



Pair of subintervals	Similarity for each subinterval pair	Total number of subinterval pairs	Overall similarity between discontinuous intervals
$\{\bar{a}_1, \bar{b}_1\}$	1.0	4	0.50
$\{\bar{a}_1, \bar{b}_2\}$	0.0		
$\{\bar{a}_2, \bar{b}_1\}$	0.0		
$\{\bar{a}_2, \bar{b}_2\}$	1.0		

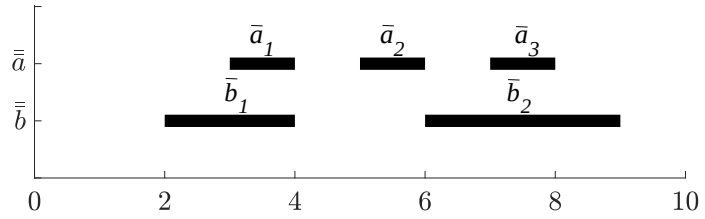
(b) Pair – 2



Pair of subintervals	Similarity for each subinterval pair	Total number of subinterval pairs	Overall similarity between discontinuous intervals
$\{\bar{a}_1, \bar{b}_1\}$	1.0	9	0.3333
$\{\bar{a}_1, \bar{b}_2\}$	0.0		
$\{\bar{a}_1, \bar{b}_3\}$	0.0		
$\{\bar{a}_2, \bar{b}_1\}$	0.0		
$\{\bar{a}_2, \bar{b}_2\}$	1.0		
$\{\bar{a}_2, \bar{b}_3\}$	0.0		
$\{\bar{a}_3, \bar{b}_1\}$	0.0		
$\{\bar{a}_3, \bar{b}_2\}$	0.0		
$\{\bar{a}_3, \bar{b}_3\}$	1.0		

(c) Pair – 3

Fig. 4.5 Results of averaging the similarity of all subinterval pairs to determine the overall similarity between discontinuous intervals.



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. 4.6 Using S_J, S_D, and S_{S_h} similarity measures, the similarity results of all combinations of subintervals within a pair of discontinuous intervals—one with three and the other with two subintervals.

intervals, $\bar{\bar{a}}$ and $\bar{\bar{b}}$. Here, $\bar{\bar{a}}$ contains three subintervals $\langle \bar{a}_1 = [3,4], \bar{a}_2 = [5,5.9], \bar{a}_3 = [7,8] \rangle$, and $\bar{\bar{b}}$ has two subintervals $\langle \bar{b}_1 = [2,4], \bar{b}_2 = [6,9] \rangle$. Therefore, $m = 3$ and $n = 2$. There are six pairs of subintervals—$\{\bar{a}_1, \bar{b}_1\}$, $\{\bar{a}_1, \bar{b}_2\}$, $\{\bar{a}_2, \bar{b}_1\}$, $\{\bar{a}_2, \bar{b}_2\}$, $\{\bar{a}_3, \bar{b}_1\}$, and $\{\bar{a}_3, \bar{b}_2\}$ where the first and last pairs are overlapping, and the rest are disjoint. Figure 4.6 shows the similarity for all pairs using Jaccard (S_J at Eq. (2.18)) and Dice (S_D at Eq. (2.22)) measures together with the bidirectional subsethood based similarity measure (S_{S_h} at Eq. (3.2)) (with the minimum *t-norm*). Hence, the overall similarity between $\bar{\bar{a}}$ and $\bar{\bar{b}}$ using Eq. (4.1) along with S_{S_h} measure is, $S_G\left(\bar{\bar{a}}, \bar{\bar{b}}\right) = \frac{1}{\max(3,2)} \times 0.8333 = \frac{1}{3} \times 0.8333 = 0.2778$. In a similar manner, the total similarity between $\bar{\bar{a}}$ and $\bar{\bar{b}}$ with S_J and S_D measures are 0.2778 and 0.3889, respectively.

Theorem 4.1:

The proposed generalized framework (S_G) in respect to the S_J, S_D and S_{S_h} similarity measures satisfies all common properties [A1] - [A4] of a similarity measure (discussed in Chapter 2) for the discontinuous intervals $\bar{\bar{a}}$, $\bar{\bar{b}}$, and $\bar{\bar{c}}$ such that

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

(2) $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = S_G\left(\overline{\overline{b}}, \overline{\overline{a}}\right)$ (symmetry);

(3) $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = 1 \iff \overline{\overline{a}} = \overline{\overline{b}}$ (reflexivity);

(4) $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) \geq S_G\left(\overline{\overline{a}}, \overline{\overline{c}}\right)$ when $\overline{\overline{a}} \subseteq \overline{\overline{b}} \subseteq \overline{\overline{c}}$. (transitivity).

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

(1) $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right)$ involves the S_J, S_D or S_{S_h} similarity measures to compute similarity for all pairs of the subintervals $\overline{a}_i \in \overline{\overline{a}}$ and $\overline{b}_j \in \overline{\overline{b}}$. All these similarity measures are bounded by 0 and 1 [16] (see Chapter 2 and 3). When $\overline{\overline{a}}$ and $\overline{\overline{b}}$ are disjoint, all pairs consist of disjoint subintervals, thereby, receiving a similarity of 0 from all these measures. Thus, the overall similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ is, $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = \frac{1}{\max(m,n)} \times 0 = 0$. When $\overline{\overline{a}}$ and $\overline{\overline{b}}$ are identical, they have equal number of m identical subintervals. Among all subinterval pairs, there are m pairs with identical subintervals and the remaining pairs have disjoint subintervals. Thus, m pairs receive a similarity of 1 from all three similarity measures, whereas other pairs receive a similarity of 0. Therefore, in this case, the overall similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ is, $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = \frac{1}{\max(m,m)} \times m = 1$. These two cases imply that $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right)$ is always within the range 0 and 1.

(2) All of S_J, S_D and S_{S_h} similarity measures are symmetric [16] (see Chapter 2 and 3), thus making $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right)$ symmetric too.

(3) If $\overline{\overline{a}} = \overline{\overline{b}}$, it means that both $\overline{\overline{a}}$ and $\overline{\overline{b}}$ have an equal number of m subintervals and each $\overline{a}_i$ is identical to each $\overline{b}_i$, i.e., $\overline{a}_i = \overline{b}_i$, $1 \leq i \leq m$. 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 $\overline{\overline{a}}$ and $\overline{\overline{b}}$ is, $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = \frac{1}{\max(m,m)} \times m = \frac{m}{m} = 1$. Again, $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = 1$ only when $\overline{\overline{a}}$ and $\overline{\overline{b}}$ are identical discontinuous intervals.

(4) All of S_J, S_D and S_{S_h} similarity measures are transitive [16] (see Chapter 2 and 3), which implies that $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right)$ is also transitive. ■

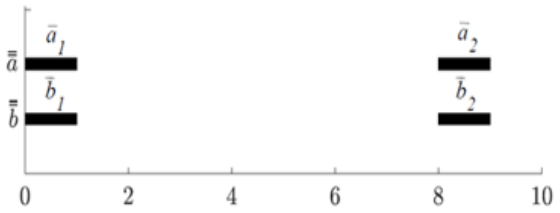
Theorem 4.2:

The proposed generalized framework (S_G) with S_J, S_D and S_{S_h} similarity measures follows the property of disjointness. That is, $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = 0$ when $\overline{\overline{a}}$ and $\overline{\overline{b}}$ are disjoint discontinuous intervals.

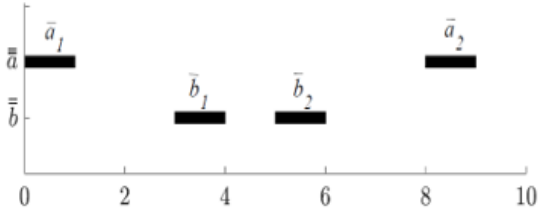
Proof: Consider $\overline{\overline{a}} = \langle \overline{a}_1, ..., \overline{a}_m \rangle$ and $\overline{\overline{b}} = \langle \overline{b}_1, ..., \overline{b}_n \rangle$. If $\overline{\overline{a}}$ and $\overline{\overline{b}}$ are disjoint, it means that no subinterval of $\overline{\overline{a}}$ is overlapping with any of subintervals of $\overline{\overline{b}}$, i.e., $\left|\overline{a}_i \cap \overline{b}_j\right| = 0$, $1 \leq i \leq m$ and $1 \leq j \leq n$. That is, all subinterval pairs consist of disjoint subintervals, thus receiving a similarity of 0. Hence, the similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ is, $S_G\left(\overline{\overline{a}}, \overline{\overline{b}}\right) = \frac{1}{\max(m,n)} \times 0 = 0$.

Example 4.2:

Consider two pairs of discontinuous intervals in Fig. 4.7 where in the first pair (Pair-1), the discontinuous intervals are completely overlapping, and in other pair (Pair-2), the discontinuous intervals are non-overlapping (disjoint). Intuitively, Pair-1 should receive a similarity of 1, whereas Pair-2 should receive 0. Figure 4.7 also presents similarity values for both pairs using three approaches – current approach with interval addition [34], current approach with convexify function [44], and the proposed generalized framework (S_G) at Eq. (4.1) in respect to S_J, S_D, and S_{S_h} similarity measures. The results show that the interval addition [34] transforms them into the same pair of continuous intervals where they are totally overlapping and getting a similarity of 1 from all three measures. This means that both discontinuous interval pairs receive a similarity of 1, which in non-intuitive. Again. using the convexify function [44], Pair-1 converts into a pair of identical continuous intervals, and Pair-2 converts into a pair of partially overlapped continuous intervals. After conversion, Pair-1 receives a similarity



Pair-1



Pair-2

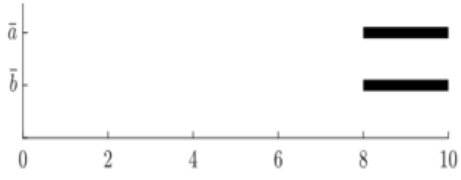
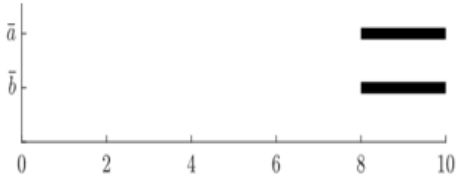
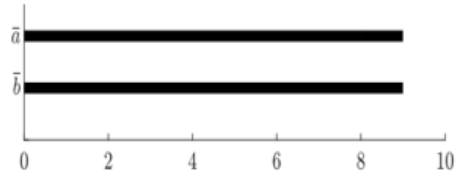
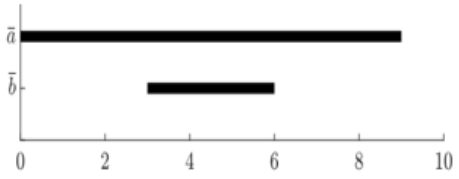
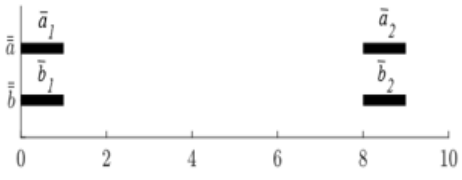
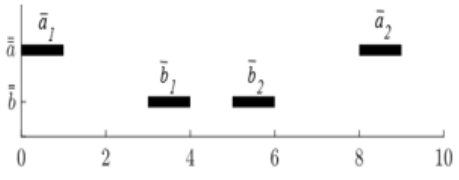
Approaches	Discontinuous Intervals		Similarity Measures					
	Pair-1	Pair-2	S_J		S_D		S_{S_h}	
			Pair-1	Pair-2	Pair-1	Pair-2	Pair-1	Pair-2
Current approach with interval addition [34]			1.0	1.0	1.0	1.0	1.0	1.0
Current approach with Convexify function [44]			1.0	0.33	1.0	0.50	1.0	0.33
Proposed generalized framework			1.0	0	1.0	0	1.0	0

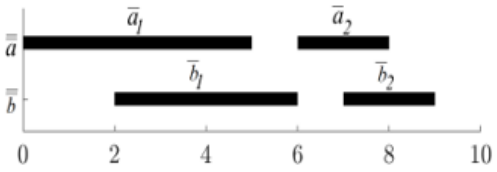
Fig. 4.7 Three approaches to compute similarity between two different pairs of discontinuous intervals.

of 1 from these measures, but Pair-2 receives a similarity of 0.33, 0.50, and 0.33 by S_J, S_D and S_{S_h} measures, respectively. These are also non-intuitive similarity for the pair of disjoint discontinuous intervals. Contrarily, the generalized framework (S_G) with all three measures gives intuitive similarity of 1 and 0 for Pair-1 and Pair-2, respectively.

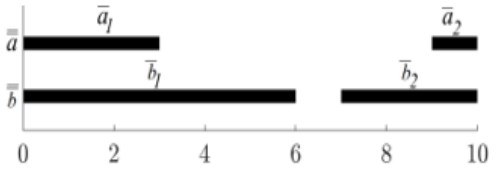
Example 4.3:

Consider another two pairs of discontinuous intervals in Fig. 4.8 where they have different degree of overlapped continuous subintervals. Intuitively, they should receive non-identical similarity as the subintervals pairs within Pair-1 and Pair-2 are completely different. Figure 4.8 presents similarity values for both pairs using three approaches – the interval addition [34], the convexify function [44], and the proposed generalized framework (S_G) at Eq. (4.1) in respect to S_J, S_D and S_{S_h} similarity measures. Results show that both interval addition [34] and the proposed framework (S_G) using S_J and S_D measures are subject to *aliasing* (same similarity) for Pair-1 and Pair-2, but provide different similarity for the same pairs using the S_{S_h} measure. Further, the convexify function [44] provides distinct similarity for those pairs using all three measures.

Though the proposed generalized framework rightly captures similarity between discontinuous intervals and provides *aliasing*-free similarity in respect to the new S_{S_h} measure (Eq. (3.2) in Chapter 3), it is often computationally expensive when discontinuous intervals comprise many subinterval pairs. Next section presents a refinement of this generalized framework to make it computationally more efficient by integrating it with an extension of Allen's theory [21].



Pair-1



Pair-2

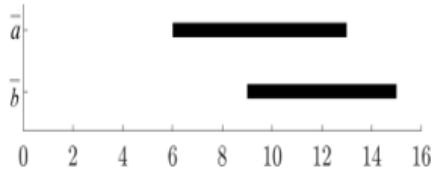
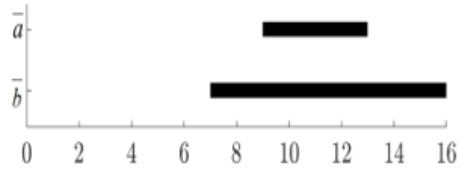
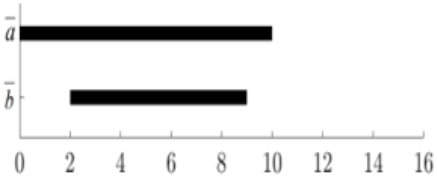
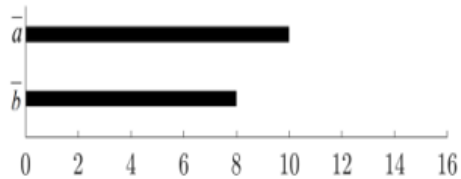
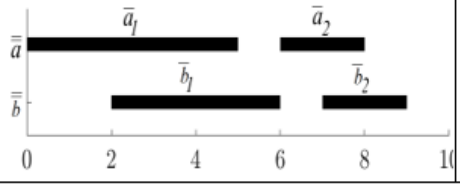
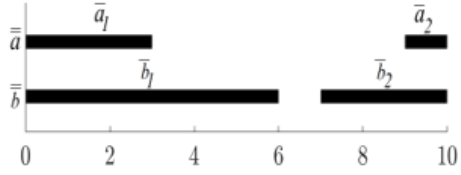
Approaches	Discontinuous Intervals		Similarity Measures					
	Pair-1	Pair-2	S_J		S_D		S_{S_h}	
			Pair-1	Pair-2	Pair-1	Pair-2	Pair-1	Pair-2
Current approach with interval addition [34]			0.444	0.444	0.615	0.615	0.571	0.444
Current approach with Convexify function [44]			0.70	0.80	0.824	0.889	0.70	0.80
Proposed generalized framework			0.417	0.417	0.584	0.584	0.55	0.417

Fig. 4.8 Three approaches to compute similarity between two different pairs of discontinuous intervals.

4.4 Refinement of the Proposed Generalized Framework with Allen's Theory

The proposed generalized framework (S_G) always consider all possible pairs of subintervals between discontinuous intervals to compute their similarity. More precisely, it always involves $\Theta(m \times n)$ similarity calculations in determining similarity for a pair of discontinuous intervals having m and n subintervals respectively. As a result, an increase in the number of subintervals within the pair of discontinuous intervals leads to an increase in the number of similarity calculations. In particular, where discontinuous intervals have many/all disjoint subinterval pairs, such a 'brute force' approach results in substantial execution time. To mitigate this, in this section, the proposed generalized framework (S_G) is integrated with an extension of Allen's theory [21] (see Chapter 2) for reducing the number of similarity calculations and overall execution time. The proposed refinement involves $O(m \times n)$ similarity calculations to determine the similarity between discontinuous intervals, thereby minimizing the overall execution time in comparison with the proposed generalized framework (S_G). Next three steps are followed for the integration of the proposed generalized framework (S_G) with an extension of Allen's theory.

1. Derive similarity value between the pair of continuous intervals directly from Allen's binary relations [21] (discussed in Chapter 2).

2. Extend three of the basic binary relations of Allen's theory [21] in order to use them for the pairs of continuous subintervals within two discontinuous intervals.

3. Develop algorithm by integrating the proposed generalized framework (S_G) with the extended binary relations of Allen's theory [21] with a view to reducing similarity calculation and execution time.

Table 4.1 Derivation of Similarity Value from Allen's Theory.

Basic Binary Relation	Similarity Value
$\overline{a}$ *before* $\overline{b}$	0
$\overline{a}$ *meets* $\overline{b}$	0
$\overline{a}$ *overlap* $\overline{b}$	$(0, 1)$
$\overline{a}$ *starts* $\overline{b}$	$(0, 1)$
$\overline{a}$ *finishes* $\overline{b}$	$(0, 1)$
$\overline{a}$ *during* $\overline{b}$	$(0, 1)$
$\overline{a}$ *equal* $\overline{b}$	1

4.4.1 Derivation of Similarity Values from Allen's Theory

From the definition of binary relations proposed by Allen [21] for continuous intervals (details are discussed in Chapter 2), it is possible to directly derive similarity between such intervals. Table 4.1 presents the derived similarity for a pair of continuous intervals $\overline{a}$ and $\overline{b}$ based on Allen's seven basic relations [21]. It is important to note that the inverse binary relations of Allen's theory [21] are not taken into consideration.

4.4.2 Extension of Allen's Theory

Allen's theory [21] is developed for defining the relation between continuous intervals (see Chapter 2); thus, this theory can not be directly used for discontinuous intervals. As discontinuous intervals contain numerous continuous subintervals, Allen's theory [21] (with an extension) can intuitively be used for defining relation between their continuous subintervals within each pair. To do this, three basic binary relations of Allen's theory [21] – *before*, *meet*, and *equal* are extended. Consider two discontinuous intervals $\overline{\overline{a}} = \langle \overline{a}_1, ..., \overline{a}_m \rangle$ and $\overline{\overline{b}} = \langle \overline{b}_1, ..., \overline{b}_n \rangle$. The extended basic binary relations are described below.

[R1] If a subinterval $\overline{a}_i$ of $\overline{\overline{a}}$ is *before* a subinterval $\overline{b}_j$ of $\overline{\overline{b}}$, then $\overline{a}_i$ is also *before* all subintervals of $\overline{\overline{b}}$ that occur after $\overline{b}_j$, i.e., $\overline{b}_{j+1}$, ..., $\overline{b}_n$.

[R2] If a subinterval $\overline{a}_i$ of $\overline{\overline{a}}$ *meets* a subinterval $\overline{b}_j$ of $\overline{\overline{b}}$, then $\overline{a}_i$ is *before* all subintervals of $\overline{\overline{b}}$ that occur after $\overline{b}_j$, i.e., $\overline{b}_{j+1}$, ..., $\overline{b}_n$.

[R3] If a subinterval $\overline{a}_i$ of $\overline{\overline{a}}$ *equals* a subinterval $\overline{b}_j$ of $\overline{\overline{b}}$, then $\overline{a}_i$ is *before* all subintervals of $\overline{\overline{b}}$ that occur after $\overline{b}_j$, i.e., $\overline{b}_{j+1}$, ..., $\overline{b}_n$, and $\overline{b}_j$ is *before* all subintervals of $\overline{\overline{a}}$ that occur after $\overline{a}_i$, i.e., $\overline{a}_{i+1}$, ..., $\overline{a}_m$.

4.4.3 Integration of Proposed Generalized Framework with Extended Binary Relations of Allen's Theory

Algorithm 1 presents an algorithm to compute the similarity between discontinuous intervals using the proposed generalized framework (S_G) combined with the extended binary relations of Allen's theory [21]. The extended binary relations are applied in this algorithm to compute the similarity between continuous subintervals within the pairs. This algorithm is developed in such a way that it can include any of S_{S_h}, S_J, and S_D similarity measures for computing similarity for subinterval pairs.

Example 4.4:

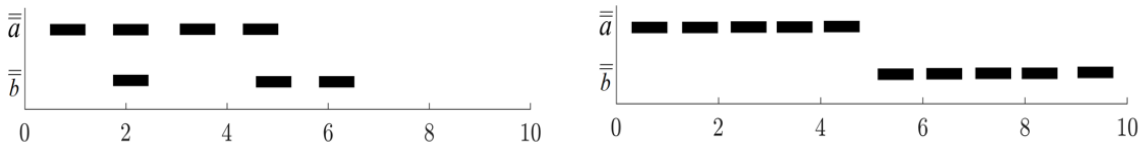
Consider two pairs of discontinuous intervals in Fig. 4.9 where their similarity calculations and overall execution time in milliseconds (ms) are shown separately for the proposed generalized framework (S_G) at Eq. (4.1) and the same framework integrated with the extended binary relations of Allen's theory (S_GA). Results show that though both approaches provide same similarity for the pairs, substantial reduction in similarity calculations and execution time (ms) can be achieved from the refined approach (i.e., combining the generalized framework with the extended Allen's theory).

The next section demonstrates the behaviour of the proposed generalized framework in respect to the S_J, S_D, and S_{S_h} similarity measures using a series of synthetic pairs of discontinuous intervals.

Algorithm 1 Similarity between discontinuous intervals with reduced similarity calculations and execution time

Input: $\overline{\overline{a}} = \langle \overline{a}_1, ..., \overline{a}_m \rangle$ and $\overline{\overline{b}} = \langle \overline{b}_1, ..., \overline{b}_n \rangle$ be a pair of discontinuous intervals. $Sim[m][n]$ be the similarity matrix for all pairs of subintervals $\{\overline{a}_i, \overline{b}_j\}$, $1 \leq i \leq m$ and $1 \leq j \leq n$. Sum_{mn} be the summation of similarity of all subinterval pairs.
Output: Similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$.

if $\overline{a}_m$ before/meet $\overline{b}_0$ or $\overline{b}_n$ before/meet $\overline{a}_0$ **then**
    return 0;
end if
for $i \rightarrow 1$ to m **do**
    for $j \rightarrow 1$ to n **do**
        if $\overline{a}_i$ *equal* $\overline{b}_j$ **then**
            Set $Sim[i][j] = 1$
            Apply [R3] and use Table 4.1 to compute similarity for
            the pairs $\{\overline{a}_i, \overline{b}_k\}$ and $\{\overline{a}_l, \overline{b}_j\}$, where $j + 1 \leq k \leq n$ and $i + 1 \leq l \leq m$
            break;
        end if
        if $\overline{a}_i$ *before/meet* $\overline{b}_j$ **then**
            Set $Sim[i][j] = 0$
            Apply [R1]/[R2] and use Table 4.1 to compute similarity for
            the pairs $\{\overline{a}_i, \overline{b}_k\}$, where $j + 1 \leq k \leq n$
            break;
        end if
        if $\overline{b}_j$ *before/meet* $\overline{a}_i$ **then**
            Set $Sim[i][j] = 0$
            Apply [R1]/[R2] and use Table 4.1 to compute similarity for
            the pairs $\{\overline{a}_l, \overline{b}_j\}$, where $i + 1 \leq l \leq m$
            continue;
        end if
        Compute $Sim[i][j]$ using any of S_{S_h}, S_J, and S_D
    end for
end for
for $i \rightarrow 1$ to m **do**
    for $j \rightarrow 1$ to n **do**
        Set $Sum_{mn} = Sum_{mn} + Sim[i][j]$
    end for
end for
return $\frac{1}{\max(m,n)} \times Sum_{mn}$



(a) Pair-1 (b) Pair-2

Pair	No. of Subintervals		S_G			S_GA		
	$\lvert\overline{\overline{a}}\rvert$	$\lvert\overline{\overline{b}}\rvert$	No. of Similarity Calculation	Execution Time (ms)	Similarity Result	No. of Similarity Calculation	Execution Time (ms)	Similarity Result
Pair-1	4	3	12	0.051	0.375	5	0.038	0.375
Pair-2	5	5	25	0.072	0.0	1	0.004	0.0

Fig. 4.9 Comparison between the generalized framework (S_G) and its combination with the extended Allen's theory (S_GA) in respect to the S_{S_h} measure.

4.5 Demonstration and Analysis

In this section, a series of experiments with synthetic discontinuous intervals are carried out to explore the behaviour of the proposed generalized framework (S_G) at Eq. (4.1) in respect to Jaccard (S_J) at Eq. (2.18) and Dice (S_D) at Eq. (2.22), and the new bidirectional subsethood based similarity measure (S_{S_h}) at Eq. (3.2). In all experiments, the minimum *t-norm* is used for the S_{S_h} measure as it is the most common in practice.

4.5.1 Datasets and Experimental Setup

For demonstration, three separate synthetic datasets are used. The datasets are:

- In the first synthetic dataset (Dataset-1), the overlapping between subintervals for a pair of discontinuous intervals is gradually decreased to see how smoothly similarity alters from 1 to 0. In particular, the changes in similarity results are investigated for a gradual change in the overlapping of subintervals.
- In the second synthetic dataset (Dataset-2), the number of subintervals and their degree of overlapping are changed. In particular, the changes in the similarity

results are explored due to a rise in the number of subintervals and their varying degree of overlap.

- Third synthetic dataset (Dataset-3) comprises pairs of discontinuous intervals which are randomly generated. In particular, in the initial pair, each discontinuous interval contains 10 subintervals. While generating other pairs, their subintervals are gradually increased to 20, 30, 40, 50, and 60. Using this dataset, the performance of the proposed generalized framework is investigated against its combination with the extended binary relations of Allen's theory. .

All 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.

4.5.2 Demonstration with Synthetic Dataset-1

Figure 4.10 presents (synthetic) Dataset-1 that considers a number of scenarios[1] for a pair of discontinuous intervals $\overline{\overline{a}}$ and $\overline{\overline{b}}$, where each of them includes two subintervals. In each scenario, the degree of overlap between subintervals of $\overline{\overline{a}}$ and $\overline{\overline{b}}$ are varied to explore how the generalized framework (S_G at Eq. (4.1)) responds with the respective similarity measures S_J, S_D, and S_{S_h} and how smoothly the similarity results change from 1 to 0. $\overline{\overline{a}}$ is keeping unchanged in all scenarios but the subintervals of $\overline{\overline{b}}$ are shifted consecutively by a factor of 25%.

Initially, both $\overline{\overline{a}}$ and $\overline{\overline{b}}$ have identical subintervals $[1, 3]$ and $[6, 8]$ (Scenario 1.1 in Fig. 4.10(a)). In Fig. 4.10(b)-(e) (Scenario 1.2 to 1.5), the rightmost subinterval $[6, 8]$ of $\overline{\overline{b}}$ is gradually shifted by a factor of 25% till its only intersection is the right-end point of the subinterval $[6, 8]$ of $\overline{\overline{a}}$. In Fig. 4.10(f)-(i) (Scenario 1.6 to 1.9), the leftmost subinterval $[1, 3]$ of $\overline{\overline{b}}$ is further shifted by a factor of 25% until its only intersection is the right-end point of the subinterval $[1, 3]$ of $\overline{\overline{a}}$. Figure 4.11(a) presents in detail the

[1]In Fig. 4.10, Dataset-1 with Scenario x is presented as Scenario 1.x where $x = 1$ to 9.

shifting of subintervals of $\overline{\overline{b}}$ for all scenarios, and Fig. 4.11(b) graphically shows the similarity results for all three measures.

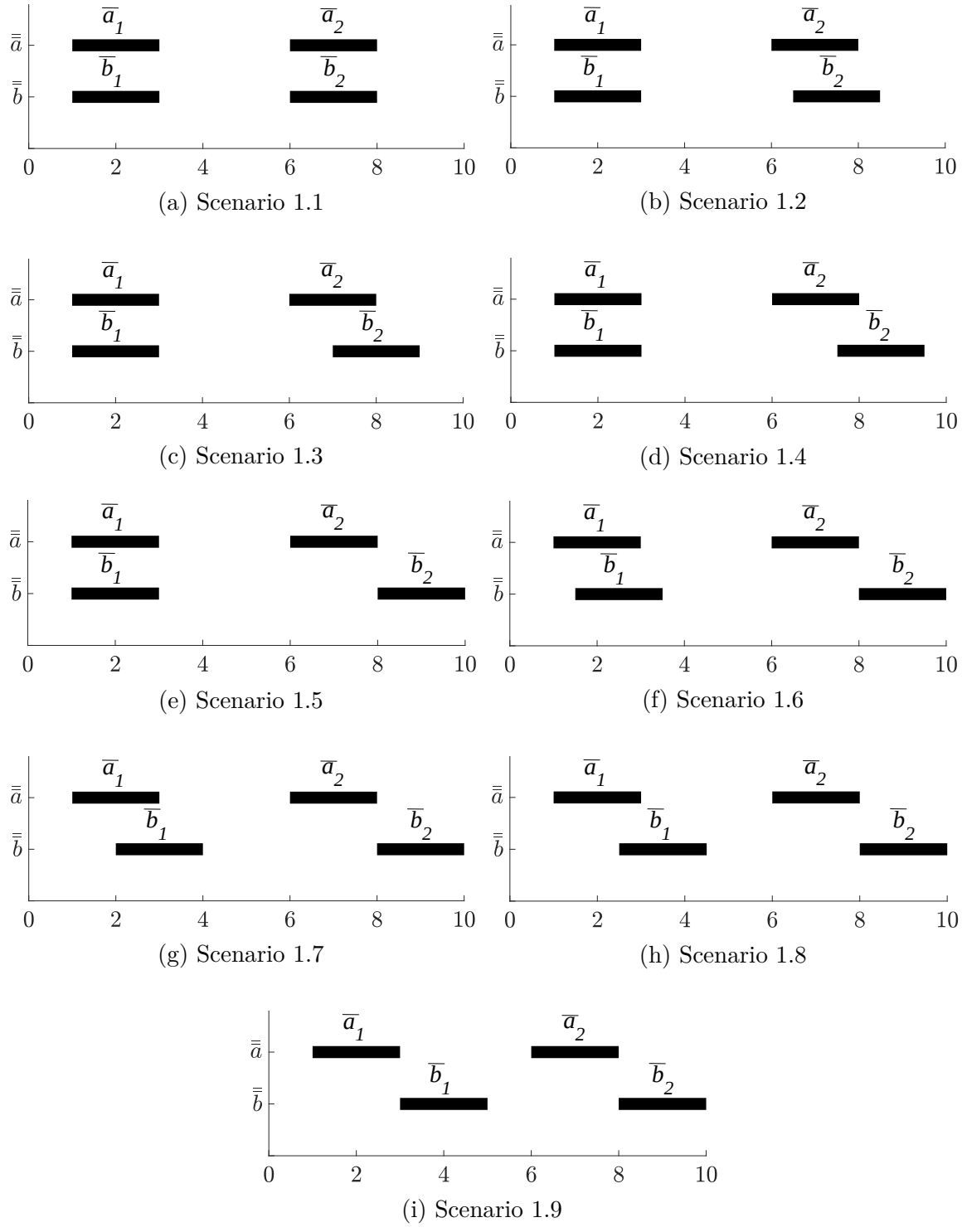
From Fig. 4.11(b), the initial similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ is 1 from the generalized framework (S_G) in respect to all three similarity measures as they are identical in Scenario 1.1 (see Fig. 4.10(a)). Their similarity gradually decreases to 0.50 when the second subinterval $[6, 8]$ of $\overline{\overline{b}}$ is gradually shifted by the factor of 25% (Scenario 1.2 to 1.5). The similarity drops further and gradually reaches to 0 when the first subinterval $[1, 3]$ of $\overline{\overline{b}}$ is also repeatedly shifted (Scenario 1.6 to 1.9). Although one would intuitively expect that similarity between the discontinuous intervals should decrease proportionately as to the rate of change in their overlapping, Fig. 4.11(b) shows a proportionate decline in similarity results by both S_{S_h} and S_D measures whereas it is not observed for the S_J measure. Here, it is worth noting that though both similarity measures, S_D and S_{S_h} have produced the same expected results, the next experiment with (synthetic) Dataset-2 would reveal their different behaviour.

4.5.3 Demonstration with Synthetic Dataset-2

Figure 4.12 presents (synthetic) Dataset-2 that considers six different scenarios[2] for a pair of discontinuous intervals, $\overline{\overline{a}}$ and $\overline{\overline{b}}$, where the number of subintervals within the discontinuous intervals and their degree of overlap are varied. Herein, it is aimed to examine how the generalized framework (S_G) with these three similarity measures behaves in respect to changes in the number of subintervals along with the degree of their overlap within the pairs.

Scenario 2.1 – In Fig. 4.12(a), both $\overline{\overline{a}}$ and $\overline{\overline{b}}$ have an identical subinterval, [0,2]. The generalized framework (S_G) with all three similarity measures returns the same similarity of 1 – as expected.

[2]In Fig. 4.12, Dataset-2 with Scenario x is presented as Scenario 2.x where $x = 1$ to 6.



(a) Scenario 1.1 (b) Scenario 1.2

(c) Scenario 1.3 (d) Scenario 1.4

(e) Scenario 1.5 (f) Scenario 1.6

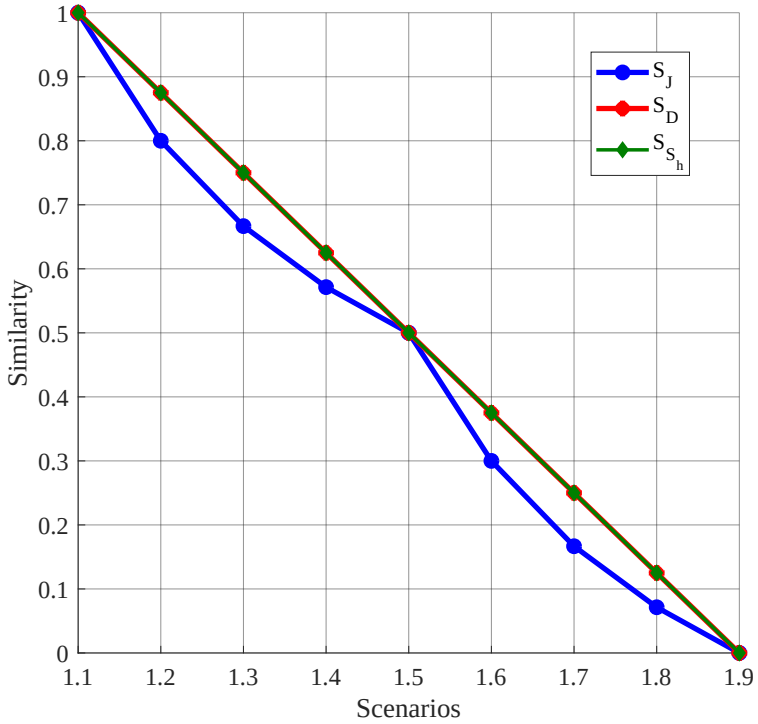
(g) Scenario 1.7 (h) Scenario 1.8

(i) Scenario 1.9

Fig. 4.10 Dataset-1 with varying degree of overlap between subintervals, $\overline{\overline{a}}$ and $\overline{\overline{b}}$ to examine how smoothly similarity results alter.

Discontinuous Interval		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

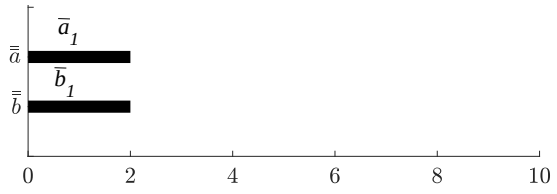
(a) % of Shifting of subintervals $\overline{b}_1$ and $\overline{b}_2$ along with the similarity results from the generalized framework (S_G) in respect to the S_J, S_D, and S_{S_h} measures (Note: $\overline{\overline{a}}$ remains unchanged in all scenarios)



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

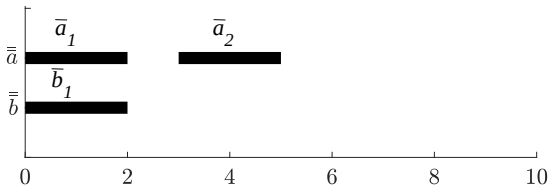
Fig. 4.11 Similarity results showing proportionate decline by the generalized framework (S_G) with both S_{S_h} and S_D similarity measures, but not with the S_J similarity measure for different scenarios in Fig. 4.10.

Scenario 2.2 – In Fig. 4.12(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 $\overline{\overline{b}}$



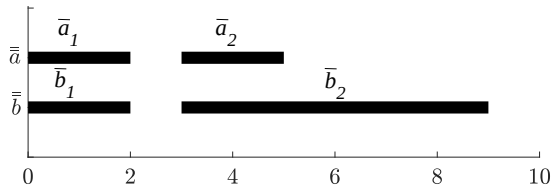
SM	Similarity
S_J	1.0
S_D	1.0
S_{S_h}	1.0

(a) Scenario 2.1



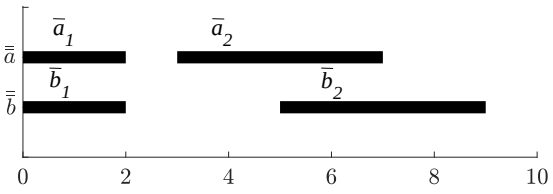
SM	Similarity
S_J	0.50
S_D	0.50
S_{S_h}	0.50

(b) Scenario 2.2



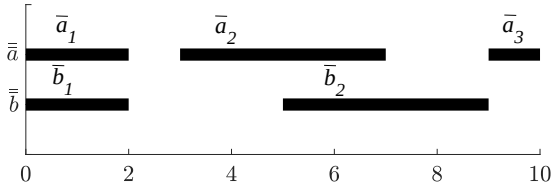
SM	Similarity
S_J	0.6667
S_D	0.75
S_{S_h}	0.6667

(c) Scenario 2.3



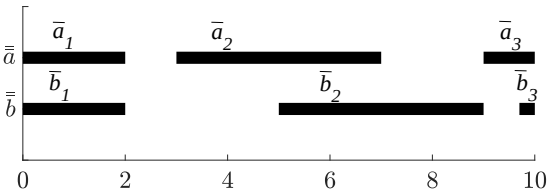
SM	Similarity
S_J	0.6667
S_D	0.75
S_{S_h}	0.75

(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. 4.12 Dataset-2 comprising pairs of discontinuous intervals with an increased number of subintervals and varying degree of overlap along with their similarity results. Note: The generalized framework (S_G) with the S_J and S_D measures return identical results for scenarios 2.3 and 2.4, i.e., they are subject to *aliasing*, but with only the S_{S_h} measure captures the change in the respective discontinuous intervals and thus avoids *aliasing*.

as compared to the *Scenario 2.1* (Fig. 4.12(a)). The generalized framework (S_G) with all three similarity measures performs as to this expectation.

Scenario 2.3 – In Fig. 4.12(c), $\overline{\overline{a}}$ is kept the same as the *Scenario 2.2* (Fig. 4.12(b)), but one more subinterval $[3, 9]$ is added 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. 4.12(b)). This expected increment in the similarity result is seen for the generalized framework (S_G) with all three similarity measures.

Scenario 2.4 – In Fig. 4.12(d), for $\overline{\overline{a}}$, one edge of subinterval $[3, 5]$ is extended to $[3, 7]$, while its other subinterval remains the same as the *Scenario 2.3* (Fig. 4.12(c)). Hence, $\overline{\overline{a}}$ now becomes $\langle[0, 2], [3, 7]\rangle$. On the other hand, for $\overline{\overline{b}}$, the size of its one subinterval $[3, 9]$ is reduced 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, it is expected that the overall similarity between $\overline{\overline{a}}$ and $\overline{\overline{b}}$ should be different from *Scenario 2.3* (Fig. 4.12(c)). The results show that the generalized framework (S_G) with S_{S_h} similarity measure can capture the change in the degree of overlap and provide a different similarity as expected, whereas the generalized framework (S_G) with both S_J and S_D measures provides the same similarity of 0.6667 and 0.75 respectively (Fig. 4.12(d)) as found in the *Scenario 2.3* (Fig. 4.12(c)). This suggests that *aliasing* persists for the generalized framework (S_G) with both S_J and S_D similarity measures, but not with the S_{S_h} similarity measure as detailed in [16] (see Chapter 3).

Scenario 2.5 – In Fig. 4.12(e), one more subinterval $[9, 10]$ to $\overline{\overline{a}}$ is added as designed in *Scenario 2.4* (Fig. 4.12(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. 4.12(d)). The results show that the generalized framework (S_G) with all three similarity measures yielded expected similarity.

Scenario 2.6 – In Fig. 4.12(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. 4.12(e)). Again, higher similarity results are received from the generalized framework (S_G) with all three similarity measures.

In summary, the generalized framework (S_G) with S_{S_h} and S_D similarity measures follow a proportionate decline in similarity results as the subintervals of a pair of discontinuous intervals are gradually moved from a complete overlap to disjoint positions. Contrarily, the generalized framework (S_G) with S_J similarity measure yields slightly higher than proportionate decline in similarity results. Importantly, the generalized framework (S_G) with S_J and S_D similarity measures still exhibit *aliasing*, whereas the generalized framework (S_G) with S_{S_h} similarity measure is sensitive to changes in overlap and thus avoid *aliasing*.

4.5.4 Demonstration with Synthetic Dataset-3

The (synthetic) Dataset-3 consists of six different pairs of discontinuous intervals $\overline{\overline{a}}$ and $\overline{\overline{b}}$, which are randomly generated. Specifically, in each scenario, discontinuous intervals contain equal number of subintervals. Initially, both discontinuous intervals have 10 randomly generated subintervals which are gradually increased to 20, 30, 40, 50, and 60 subintervals. Table 4.2 presents the discontinuous interval pairs with equal number of subintervals. This experiment investigates the performance of the proposed generalized framework (S_G) against its combination with the extended binary relations of Allen's theory [21] (S_GA), in respect to total similarity calculation and execution

Table 4.2 Performances of the proposed generalized framework (S_G) against its combination with the extended Allen's theory (S_GA) in respect to total similarity calculation and execution time in milliseconds (ms).

No of Subintervals in each pair of discontinuous intervals	S_G			S_GA		
	No. of Similarity Calculation	Execution Time (ms)	Similarity Result	No. of Similarity Calculation	Execution Time (ms)	Similarity Result
10	100	2.41	0.2428	24	0.10	0.2428
20	400	2.93	0.2331	52	0.18	0.2331
30	900	3.44	0.4089	96	0.28	0.4089
40	1600	4.66	0.2824	116	0.41	0.2824
50	2500	5.18	0.3133	154	0.88	0.3133
60	3600	4.13	0.0	1	0.0008	0.0

time (in milliseconds) in computing the similarity for pairs of discontinuous intervals. In this experiment, the new S_{S_h} similarity measure at Eq. (3.2) is used to compute similarity for subinterval pairs.

Results from Table 4.2 reveal that the generalized framework combined the extended Allen's theory [21] (S_GA) minimizes total number of similarity computation and execution time as compared to the proposed generalized framework (S_G). In particular, the generalized framework combined with the extended Allen's theory [21] (S_GA) takes only 1 similarity calculation as well as reduces execution time substantially for the last pair of disjoint discontinuous intervals having 60 subintervals whereas the generalized framework (S_G) involves 3600 similarity calculations with prolonged execution time.

From the demonstrations with different synthetic datasets, Table 4.3 presents an overall summary of the four different approaches — current approach with interval addition [34], current approach with 'convexify' function [44], the proposed generalized framework (S_G) at Eq.(4.1), and its combination with the extended binary relations of Allen's theory [21] (S_GA) in respect to the Jaccard (S_J), Dice (S_D), and bidirectional subsethood based (S_{S_h}) similarity measures. Table 4.3 shows that current approaches with S_J and S_D measures involve information loss and *aliasing*, thus affecting accuracy of similarity results, though they are computationally inexpensive. The proposed generalized framework (S_G) with S_{S_h} measure does not face information loss and

aliasing, hence provides intuitive similarity results, but with prolonged execution time. In contrast, the proposed framework in combination with the extended Allen's theory [21]) (S_GA) provides the same intuitive results with substantially less execution time compared to the use of the generalized framework alone. It is worth mentioning that both the proposed framework (S_G) and its refinement (S_GA) with the S_J and S_D measures still suffer from *aliasing* in computing similarity for discontinuous intervals.

Table 4.3 An overall summary of four approaches in respect to the S_J, S_D, and S_{S_h} similarity measures.

Approaches	Similarity Measure		
	S_J	S_D	S_{S_h}
1. Current approach with interval addition [34]	• Information loss • Non-intuitive similarity • Subject to *aliasing* • Computationally less expensive	• Information loss • Non-intuitive similarity • Subject to *aliasing* • Computationally less expensive	• Information loss • Non-intuitive similarity • Minimizes *aliasing* • Computationally less expensive
2. Current approach with 'Convexify' function [44]	• Information loss • Non-intuitive similarity • Subject to *aliasing* • Computationally less expensive	• Information loss • Non-intuitive similarity • Subject to *aliasing* • Computationally less expensive	• Information loss • Non-intuitive similarity • Minimizes *aliasing* • Computationally less expensive
3. Proposed generalized framework	• No information loss • Intuitive similarity • Subject to *aliasing* • Computationally expensive	• No information loss • Intuitive similarity • Subject to *aliasing* • Computationally expensive	• No information loss • Intuitive similarity • Minimizes *aliasing* • Computationally expensive
4. Proposed generalized framework with extended Allen's Theory	• No information loss • Intuitive similarity • Subject to *aliasing* • Computationally less expensive than the proposed generalized framework	• No information loss • Intuitive similarity • Subject to *aliasing* • Computationally less expensive than the proposed generalized framework	• No information loss • Intuitive similarity • Minimizes *aliasing* • Computationally less expensive than the proposed generalized framework

4.6 Summary

This chapter has proposed a generalized framework to compute similarity of discontinuous intervals by considering similarity of all continuous subinterval-pairs. The new generalized framework is equally suitable for continuous as well as discontinuous intervals. It does not require any conversion from discontinuous into continuous intervals, thus avoiding associated information loss. The performances of the generalized framework using the bidirectional subsethood similarity measure along with the Jaccard and Dice measures have been compared using different synthetic pairs of discontinuous intervals. The results have shown intuitive behaviour of the generalized framework while highlighting that only by using the bidirectional subsethood based similarity as part of the generalized framework, *aliasing* can continue to be avoided.

At the computational level, in the generalized framework, all possible pairs of subintervals have been considered to compute similarity which results in a potentially vast number of similarity calculations, in particular, where discontinuous intervals have many disjoint subinterval pairs. To mitigate this and to make the proposed generalized framework computationally more efficient, it has been further integrated with an extension of Allen's theory [21]. Experiment results have shown that both similarity calculation and execution time have been substantially reduced with this integration as compared to the use of generalized framework alone. The works that have been done in this chapter satisfy objective 5 of the thesis.

A part of the contribution to this chapter has been published in the following international conference.

- **Measuring Similarity Between Discontinuous Intervals - Challenges and Solutions**

  Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. Measuring Similarity Between Discontinuous Intervals - Challenges and Solutions.

In International Conference on Fuzzy Systems (Fuzz-IEEE 2019), New Orleans, Louisiana, USA, 2019.

Status: *Published*

A journal paper, built on our initial work, is now being prepared for submission to IEEE Transactions on Fuzzy Systems.

- **Efficient Comparison of Discontinuous Intervals using a Combination of Bidirectional Subsethood Based Similarity Measure and Extended Allen's Theory**

  Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. Efficient Comparison between Discontinuous Intervals using a Combination of Bidirectional Subsethood Based Similarity Measure and Extended Allen's Theory.

  Status: *To be submitted*

Chapter 3 has introduced a novel set-theoretic similarity measure for continuous (closed) intervals, whereas Chapter 4 has put forward a generalized framework to assess the similarity between discontinuous intervals. In these chapters, different experimental results have shown that the new similarity measure as well as its inclusion within the generalized framework can avoid *aliasing* of the popular set-theoretic measures and provide similarity with high accuracy. Taking advantages of these proposed measures, the next chapter presents their extensions (see Chapter 2) for robust comparison between T1 fuzzy sets. These extensions will also meet up objective 5 of the thesis.

Chapter 5

Measuring Similarity Between Type-1 Fuzzy Sets

5.1 Introduction

Chapter 3 introduced a *robust* set-theoretic similarity measure for continuous (closed) intervals to deal with an inherent limitation (i.e., *aliasing*) of the popular similarity measures where they return the same similarity for substantially different pairs of intervals. Afterwards, Chapter 4 proposed a generalized framework to address information loss with the current approaches while assessing the similarity between discontinuous intervals. It also has highlighted that *aliasing* can continue to be avoided for discontinuous intervals only by using the new similarity measure (Eq. (3.2) in Chapter 3) as part of the generalized framework. Therefore, this chapter aims to exploit these new measures and extend them to T1 fuzzy sets with the intention to capture their similarity more accurately.

There are two common approaches to perform such extensions – one based on vertical-slices and the other on α-cuts of fuzzy sets (discussed in Chapter 2). As both vertical and α-cuts of normal and convex T1 fuzzy sets are closed intervals, the

new similarity measure for continuous intervals are initially expanded using the two approaches to compare such fuzzy sets. Later, the proposed generalized framework involving the new measure is extended to deal with the comparison between non-convex T1 fuzzy sets.

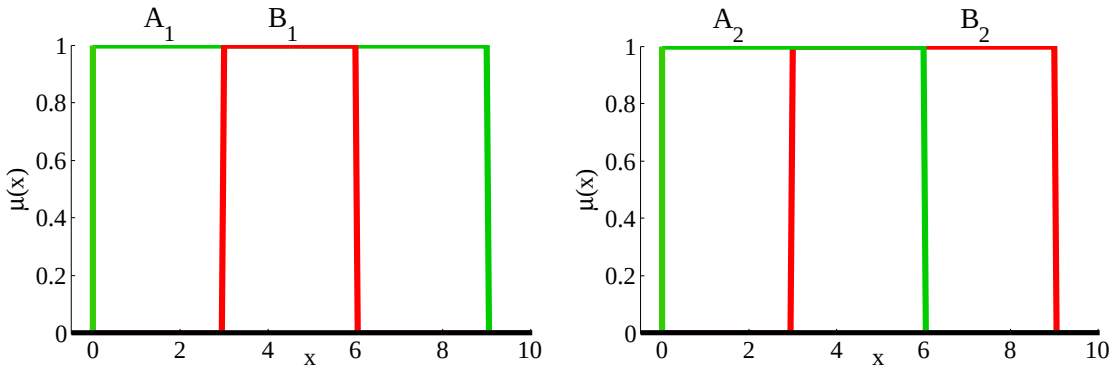
This chapter first presents the motivation behind the extension of the new interval similarity measure (Eq. (3.2) in Chapter 3) to T1 fuzzy sets. Following this, this new measure is extended based on the vertical-slice representation along with a discussion of the essential properties that the extended measure supports. After that, the behaviour of the proposed similarity measure (extended vertically) against key existing fuzzy set-theoretic measures (also depending on the vertical-slice representation) is demonstrated using a series of synthetic fuzzy sets, followed by a discussion of their limitation to compare non-normal T1 fuzzy sets. To avoid this limitation, an α-cut based extension of the new measure is explored for T1 fuzzy sets together with its major properties. Lastly, the behaviour of the proposed extended measure based on the α-cut representation is compared with the vertically extended measure (and other key existing measures) using the same set of synthetic fuzzy sets.

5.2 Motivation

Classical set-theoretic similarity measures are often extended to assess similarity between fuzzy sets [75], and an overview of such extended measures is discussed in [11] and [87]. However, Chapter 3 has identified and articulated a pitfall (i.e., *aliasing*) of these similarity measures and proposed a new similarity measure for intervals based on the concept of subsethood, which has shown to avoid *aliasing* and produce intuitive results. These findings inspire us to leverage this new measure to assess similarity of T1 fuzzy sets as T1 fuzzy sets can be represented as a series of vertical-slices or

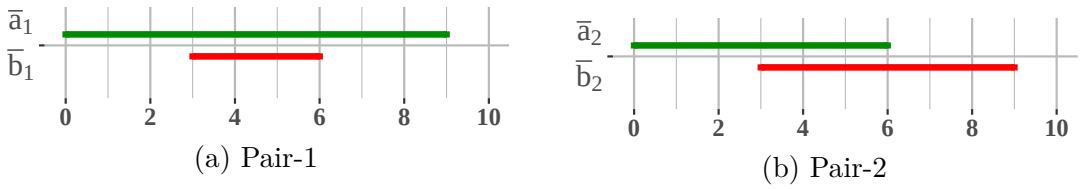
α-cuts, and both of these slices are generally intervals. The following example further substantiates the rationale behind extending the new measure for T1 fuzzy sets.

Consider two different pairs of T1 fuzzy sets in Fig. 5.1 and their corresponding interval representations in Fig. 5.2. Here, both fuzzy set pairs and their interval representations are dissimilar, therefore, intuitively they should have a different degree of similarity. Chapter 3 has revealed that both Jaccard and Dice measures produce equal similarity for these interval pairs as they are subject to *aliasing* (see Fig. 3.1). When their extensions (see [65] and [141]) are applied to compare the fuzzy set pairs in Fig. 5.1, they still provide the same similarity for them. Contrarily, Chapter 3 has demonstrated that the new similarity measure can generate the expected difference in similarity for the interval pairs of Fig. 5.2. Hence, its extension is also expected to capture the similarity between fuzzy sets more appropriately.



(a) Pair-1 (b) Pair-2

Fig. 5.1 Two different example pairs of T1 fuzzy sets.



(a) Pair-1 (b) Pair-2

Fig. 5.2 Interval representations for the pairs of T1 fuzzy sets in Fig. 5.1.

The next section presents a vertical-slice based extension of the bidirectional subsethood based similarity measure (Eq. (3.2) in Chapter 3) for T1 fuzzy sets and discusses its major properties.

5.3 Proposed Vertical-Slice Based Extension

This section introduces a vertical-slice based extension of the new similarity measure (S_{S_h}) introduced in Chapter 3 to assess the similarity between T1 fuzzy sets. The proposed extension of the S_{S_h} measure is presented first and then its major properties are explored and demonstrated.

5.3.1 Proposed Vertical-Slice Based Extension of the Bidirectional Subsethood Based Similarity Measure

In the vertical-slice based extension, the reciprocal subsethood for a pair of T1 fuzzy sets is first determined by considering the membership values of their elements and then a *t-norm* of the subsethood outcomes are taken to measure overall similarity between the fuzzy sets. The proposed extension of the S_{S_h} measure for fuzzy sets, A and B, is thus defined as,

$$S^{T1}_{S_h v}(A, B) = \bigstar \left(S_h(A, B), S_h(B, A) \right), \tag{5.1}$$

where $S_h(A, B)$ is the subsethood of A with respect to B and $S_h(B, A)$ is the subsethood of B with respect to A. As stated before, $\bigstar$ is the *t-norm*.

Using the definition of subsethood (S_h) for T1 fuzzy sets (see subsection 2.6.2 in Chapter 2), Eq. (5.1) can be expressed in the following manner,

$$S^{T1}_{S_h v}(A,B) = \bigstar \left(\frac{\sum_{i=1}^{n} \min\left(\mu_A(x_i), \mu_B(x_i)\right)}{\sum_{i=1}^{n} \mu_A(x_i)}, \frac{\sum_{i=1}^{n} \min\left(\mu_A(x_i), \mu_B(x_i)\right)}{\sum_{i=1}^{n} \mu_B(x_i)} \right), \tag{5.2}$$

where n is the number of discretizations of the whole support of A and B.[12]

In the next subsection, the properties of this vertically extended $S^{T1}_{S_h v}$ measure are explored and proved.

5.3.2 Properties of the Vertical-Slice Based Extension

This subsection proves that the vertically extended $S^{T1}_{S_h v}$ measure (in Eq. (5.1)) maintains all major properties [C1] - [C4] of a fuzzy set similarity measure (discussed in Chapter 2) for T1 fuzzy sets, A and B. It also shows that this extended $S^{T1}_{S_h v}$ measure is always bounded by 0 and 1.

Theorem 5.1 (Boundedness): $S^{T1}_{S_h v}(A,B) \in [0,1]$.

Proof: The subsethood $S_h(A,B)$ is bounded by $[0,1]$ [97], which follows that the extended similarity measure $S^{T1}_{S_h v}(A,B)$ is also bounded by $[0,1]$. ■

Theorem 5.3 (Symmetry): $S^{T1}_{S_h v}(A,B) = S^{T1}_{S_h v}(B,A)$.

Proof: The *t-norm* ($\bigstar$) is symmetric [137]. Hence, $S^{T1}_{S_h v}(A,B)$ is also symmetric. ■

Theorem 5.2 (Reflexivity): $S^{T1}_{S_h v}(A,B) = 1 \iff A = B$.

[1]It is noted that in case of discrete fuzzy sets A and B (where they have the same number of elements), n denotes the number of elements in A and B.

[2]Note that Eq. (5.2) behaves like a fuzzy measure mentioned in [94] for the minimum *t-norm*, though the origin of that measure is unknown.

Proof: When $A = B$, $S_h(A,B) = S_h(B,A) = 1$. From the boundary conditions of the *t-norm* ($\bigstar$) [137], $\bigstar(1,1) = 1$, thus making $S_{S_h v}(A,B) = 1$. Alternatively, $S_{S_h v}(A,B) = 1$ means that both $S_h(A,B)$ and $S_h(B,A)$ are equal to 1. This only happens if A and B are identical. ■

Theorem 5.4 (Transitivity): $S^{T1}_{S_h v}(A,B) \geq S^{T1}_{S_h v}(A,C)$ when $A \subseteq B \subseteq C$.

Proof: When $A \subseteq B \subseteq C$, it follows that

$$\sum_x \mu_A(x) \leq \sum_x \mu_B(x) \leq \sum_x \mu_C(x). \tag{5.3}$$

Case 1: when $\bigstar$ is the minimum *t-norm*. From Eq. (5.2),

$$\begin{aligned} S^{T1}_{S_h v}(A,B) &= \bigstar \left(\frac{\sum_x \min(\mu_A(x), \mu_B(x))}{\sum_x \mu_A(x)}, \frac{\sum_x \min(\mu_A(x), \mu_B(x))}{\sum_x \mu_B(x)} \right) \\ &= \frac{\sum_x \min(\mu_A(x), \mu_B(x))}{\sum_x \mu_B(x)} \qquad \text{[using Eq. (5.3)]} \end{aligned}$$

$$\begin{aligned} S^{T1}_{S_h v}(A,C) &= \bigstar \left(\frac{\sum_x \min(\mu_A(x), \mu_C(x))}{\sum_x \mu_A(x)}, \frac{\sum_x \min(\mu_A(x), \mu_C(x))}{\sum_x \mu_C(x)} \right) \\ &= \frac{\sum_x \min(\mu_A(x), \mu_C(x))}{\sum_x \mu_C(x)} \qquad \text{[using Eq. (5.3)]} \end{aligned}$$

From Eq. (5.3), $\sum_x \mu_B(x) \leq \sum_x \mu_C(x)$, thus making

$$\frac{\sum_x \min(\mu_A(x), \mu_B(x))}{\sum_x \mu_B(x)} \geq \frac{\sum_x \min(\mu_A(x), \mu_C(x))}{\sum_x \mu_C(x)}.$$

Hence $S^{T1}_{S_h v}(A,B) \geq S^{T1}_{S_h v}(A,C)$.

Case 2: when $\bigstar$ is the product *t-norm*. From Eq. (5.2),

$$S^{T1}_{S_h v}(A,B) = \bigstar \left(\frac{\sum\limits_x \min\left(\mu_A(x), \mu_B(x)\right)}{\sum\limits_x \mu_A(x)}, \frac{\sum\limits_x \min\left(\mu_A(x), \mu_B(x)\right)}{\sum\limits_x \mu_B(x)} \right)$$
$$= \frac{\sum\limits_x \min\left(\mu_A(x), \mu_B(x)\right)^2}{\sum\limits_x \mu_A(x) \times \sum\limits_x \mu_B(x)} \qquad \text{[using Eq. (5.3)]}$$

$$S^{T1}_{S_h v}(A,C) = \bigstar \left(\frac{\sum\limits_x \min\left(\mu_A(x), \mu_C(x)\right)}{\sum\limits_x \mu_A(x)}, \frac{\sum\limits_x \min\left(\mu_A(x), \mu_C(x)\right)}{\sum\limits_x \mu_C(x)} \right)$$
$$= \frac{\sum\limits_x \min\left(\mu_A(x), \mu_C(x)\right)^2}{\sum\limits_x \mu_A(x) \times \sum\limits_x \mu_C(x)} \qquad \text{[using Eq. (5.3)]}$$

From Eq. (5.3), $\sum\limits_x \mu_B(x) \leq \sum\limits_x \mu_C(x)$. Hence,

$$\frac{\sum\limits_x \min\left(\mu_A(x), \mu_B(x)\right)}{\sum\limits_x \mu_B(x)} \geq \frac{\sum\limits_x \min\left(\mu_A(x), \mu_C(x)\right)}{\sum\limits_x \mu_C(x)},$$

It follows that

$$\frac{\sum\limits_x \min\left(\mu_A(x), \mu_B(x)\right)^2}{\sum\limits_x \mu_B(x)} \geq \frac{\sum\limits_x \min\left(\mu_A(x), \mu_C(x)\right)^2}{\sum\limits_x \mu_C(x)},$$

$$\implies \frac{\sum\limits_x \min\left(\mu_A(x), \mu_B(x)\right)^2}{\sum\limits_x \mu_A(x) \times \sum\limits_x \mu_B(x)} \geq \frac{\sum\limits_x \min\left(\mu_A(x), \mu_C(x)\right)^2}{\sum\limits_x \mu_A(x) \times \sum\limits_x \mu_C(x)}.$$

Thus $S^{T1}_{S_h v}(A,B) \geq S^{T1}_{S_h v}(A,C)$. ■

Theorem 5.5 (Overlapping): $S^{T1}_{S_h v}(A, B) > 0$ when $A \cap B \neq \emptyset$; otherwise, $S^{T1}_{S_h v}(A, B) = 0$.

Proof: If $A \cap B \neq \emptyset$, $\exists x$ such that $\min(\mu_A(x), \mu_B(x)) > 0$, which follows that the numerator of Eq. (5.2) is

$$\sum_x \min(\mu_A(x), \mu_B(x)) > 0.$$

For the denominators in Eq. (5.2),

$$\sum_x \mu_A(x) \geq \sum_x \min(\mu_A(x), \mu_B(x)) \quad \text{and} \quad \sum_x \mu_B(x) \geq \sum_x \min(\mu_A(x), \mu_B(x)).$$

Therefore $S^{T1}_{S_h v}(A, B) > 0$.

Conversely, if $A \cap B = \emptyset$, then $\min(\mu_A(x), \mu_B(x)) = 0 \ \forall x$, thus making the numerator of (5.2) as

$$\sum_x \min(\mu_A(x), \mu_B(x)) = 0.$$

Therefore,

$$\frac{\sum_x \min(\mu_A(x), \mu_B(x))}{\sum_x \mu_A(x)} = 0 \quad \text{and} \quad \frac{\sum_x \min(\mu_A(x), \mu_B(x))}{\sum_x \mu_B(x)} = 0.$$

Hence $S^{T1}_{S_h v}(A, B) = 0$. ■

The next subsection demonstrates the behaviour of the vertically extended $S^{T1}_{S_h v}$ measure in respect to existing fuzzy set-theoretic similarity measures (also depending on the vertical-slice representation) using synthetic datasets on T1 fuzzy sets.

5.3.3 Comparison with Current Literature

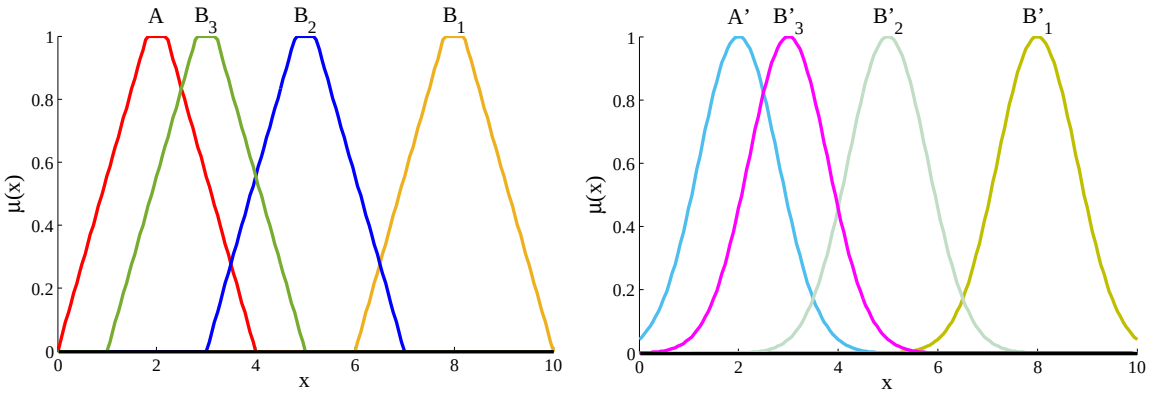
This section compares the proposed vertical-slice Based extended measure ($S^{T1}_{S_hv}$) at Eq. (5.1) with well-known fuzzy set-theoretic similarity measures using T1 fuzzy sets with different features (i.e., normal/non-normal and convex/non-convex). Comparisons are given against the following fuzzy set-theoretic similarity measures:

- S_J – proposed by Pappis and Karacapilidis [65]
- S_C – proposed by Chen et al. [91]
- S_P – proposed by Chen et al. [91]
- S_W – proposed by Wang [92]
- S_R – proposed by Raha et al. [93]

Table 2.1 (in Chapter 2) provides details of S_J, S_C, S_P, S_W, and S_R measures. It is worth noting that all these existing fuzzy similarity measures also depend on the vertical-slice representation of fuzzy sets. For the proposed $S^{T1}_{S_hv}$ measure, both minimum and product *t-norms* are used. Further, for comparison, both trapezoidal and Gaussian T1 fuzzy sets are used as the trapezoidal membership function is simple, yet captures a great deal of flexibility and can model non-symmetric concepts. On the other hand, the Gaussian membership function is a favourite statistical representation.

All of these similarity measures are compared by focusing on the changes in similarity results while the properties of the fuzzy sets are gradually altered, i.e., by varying the degree of overlap, the subsethood, the height of (non-normal) fuzzy sets, and their 'degree of non-convexity'. Figures 5.3 to 5.7 graphically show different synthetic fuzzy sets that are used in this comparison accompanied with similarity results from all measures while considering 1000 discretizations (vertical-slices) along the whole support of the pairs of fuzzy sets.

(1) Convex-normal fuzzy sets: In Fig. 5.3, three pairs of trapezoidal ($\{A, B_1\}$, $\{A, B_2\}$, and $\{A, B_3\}$) and Gaussian ($\{A', B'_1\}$, $\{A', B'_2\}$, and $\{A', B'_3\}$) fuzzy sets are used, where the degree of overlap between the pairs is increased gradually (e.g., the overlap between A and B_1 in $\{A, B_1\}$ is less than the overlap between A and B_3 in $\{A, B_3\}$). Here, all pairs are normal and convex. Specifically, pairs are shifted from completely disjoint to a gradually increasing overlap along the x-axis. Intuitively, as the overlapping between pairs increases, similarity is expected to rise accordingly. Figure 5.3 shows expected behaviour of all similarity measures.

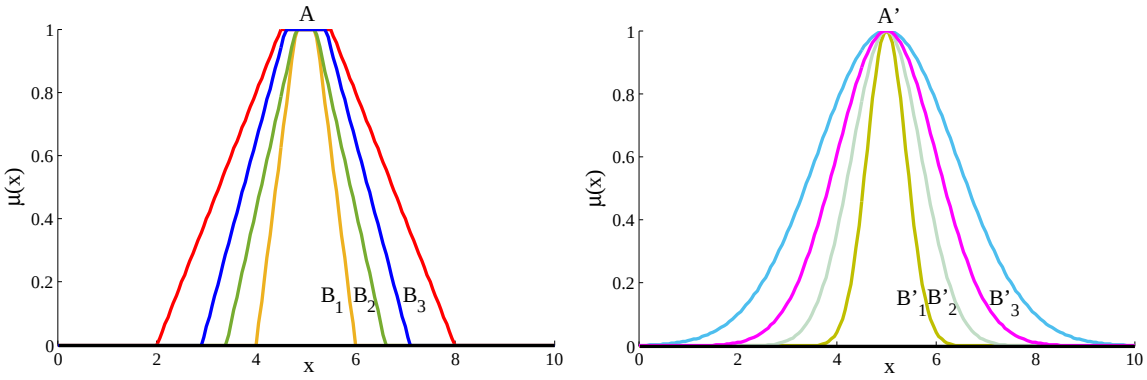


(a) Trapezoidal Fuzzy Sets (b) Gaussian Fuzzy Sets

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.0	0.0326	0.3968	0.0	0.0321	0.373
S_C	0.0	0.2778	0.8333	0.0	0.1724	0.8226
S_P	0.0	0.0322	0.724	0.0	0.0301	0.6766
S_W	0.0	0.0552	0.2728	0.0	0.0689	0.2929
S_R	0.0	0.2778	0.8333	0.0	0.1724	0.8226
$S^{T1}_{S_h v:min}$	0.0	0.0631	0.5682	0.0	0.0625	0.5426
$S^{T1}_{S_h v:prod}$	0.0	0.004	0.3228	0.0	0.0039	0.2949

Fig. 5.3 Trapezoidal and Gaussian fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the degree of overlap between fuzzy sets is increasing from right to left.

(2) Convex-normal fuzzy sets (with proper subsets): In Fig. 5.4, three pairs of (normal and convex) trapezoidal ($\{A, B_1\}$, $\{A, B_2\}$, and $\{A, B_3\}$) and Gaussian ($\{A', B'_1\}$, $\{A', B'_2\}$, and $\{A', B'_3\}$) fuzzy sets are considered, where all B_1, B_2, and B_3 are proper subsets of A (similarly, all B'_1, B'_2, and B'_3 are proper subsets of A'). Here, the degree of subsethood in the pairs is gradually increased (e.g., the degree of subsethood of B_1 in respect to A is less than that of B_3 in respect to A). In this case, it is intuitive to expect that as the degree of subsethood gradually increases, similarity will also rise. Figure 5.4 shows that both S_C and S_R measures yield a similarity of 1



(a) Trapezoidal Fuzzy Sets

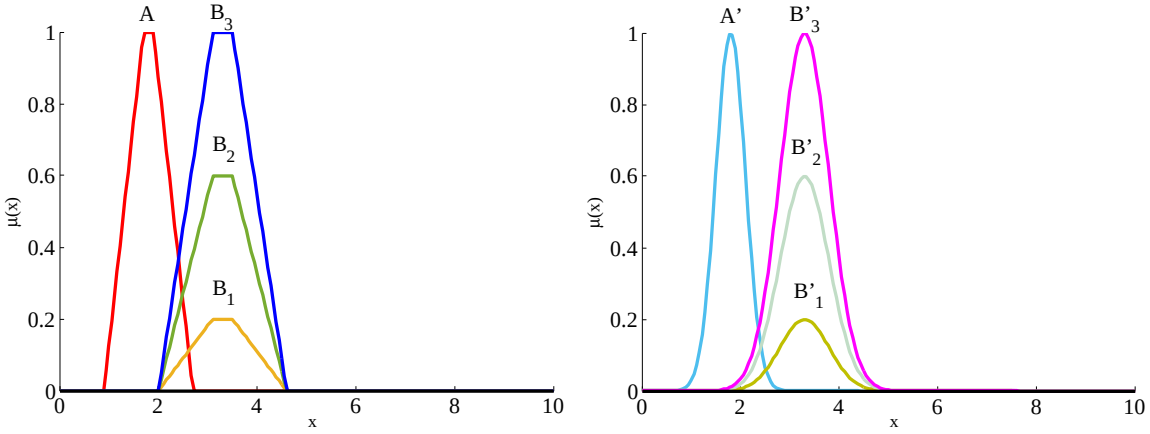
(b) Gaussian Fuzzy Sets

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.3429	0.5143	0.7143	0.2858	0.5002	0.7145
S_C	1.0	1.0	1.0	1.0	1.0	1.0
S_P	0.4422	0.6275	0.817	0.3885	0.6325	0.822
S_W	0.2042	0.3281	0.5001	0.1045	0.2024	0.3576
S_R	1.0	1.0	1.0	1.0	1.0	1.0
$S^{T1}_{S_h v:min}$	0.3429	0.5143	0.7143	0.2858	0.5002	0.7145
$S^{T1}_{S_h v:prod}$	0.3429	0.5143	0.7143	0.2858	0.5002	0.7145

Fig. 5.4 Trapezoidal and Gaussian T1 fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the degree of subsethood between fuzzy sets is increasing.

for all pairs despite being non-identical while the other measures return an increasing degree of similarity in line with expectation.

(3) Convex fuzzy sets (with one normal and one non-normal fuzzy set): trapezoidal ($\{A, B_1\}$, $\{A, B_2\}$, and $\{A, B_3\}$) and Gaussian ($\{A', B'_1\}$, $\{A', B'_2\}$, and $\{A', B'_3\}$) fuzzy sets are considered in Fig. 5.5 where A and A' are normal sets while B_1 and B'_1 are initially non-normal, but their heights are gradually increased and turned into normal sets, B_3 and B'_3 respectively. As the height of the non-normal fuzzy set



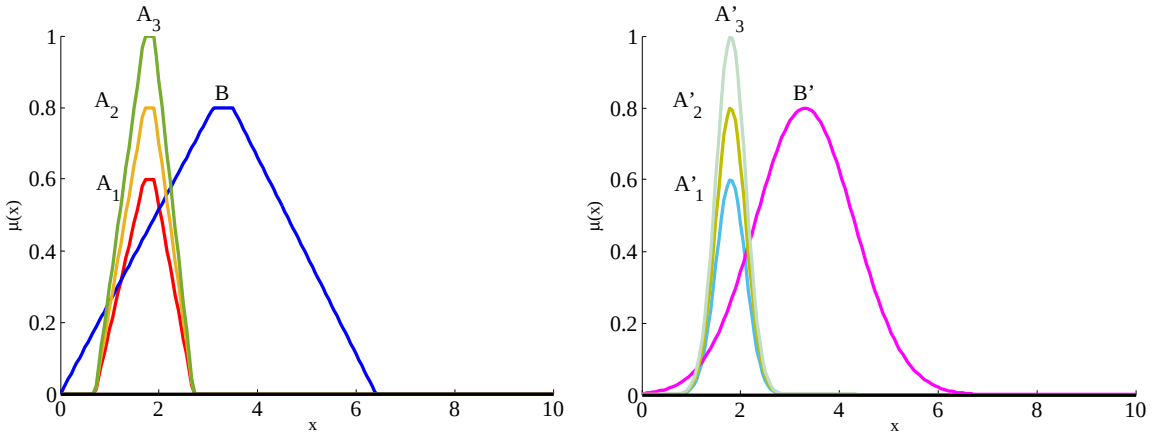
(a) Trapezoidal Fuzzy Sets (b) Gaussian Fuzzy Sets

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.0313	0.0515	0.0541	0.0216	0.0304	0.0316
S_C	0.1108	0.2658	0.3676	0.0587	0.1249	0.1724
S_P	0.0177	0.0532	0.0573	0.0089	0.0268	0.0279
S_W	0.0443	0.0662	0.0721	0.0389	0.0418	0.0434
S_R	0.1108	0.2658	0.3676	0.0587	0.1249	0.1724
$S^{T1}_{S_h v:min}$	0.0389	0.093	0.0961	0.0281	0.0588	0.0625
$S^{T1}_{S_h v:prod}$	0.0051	0.0096	0.0111	0.0024	0.0035	0.0042

Fig. 5.5 Trapezoidal and Gaussian fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the height of the non-normal fuzzy set is increasing while the other set is normal.

increases, the similarity within the pairs should intuitively rise, and their similarity reaches its maximum value when both sets in a pair share the same height. Here, all measures maintain the expected rising trend in the similarity results.

(4) Convex fuzzy sets (with both fuzzy sets non-normal): Here, in Fig. 5.6, three pairs of trapezoidal ($\{A_1, B\}$, $\{A_2, B\}$, and $\{A_3, B\}$) and Gaussian ($\{A'_1, B'\}$, $\{A'_2, B'\}$, and $\{A'_3, B'\}$) fuzzy sets are considered where initially both fuzzy sets are convex and non-normal (i.e., $\{A_1, B\}$ and $\{A'_1, B'\}$). In this experiment, B and B' are



(a) Trapezoidal Fuzzy Sets

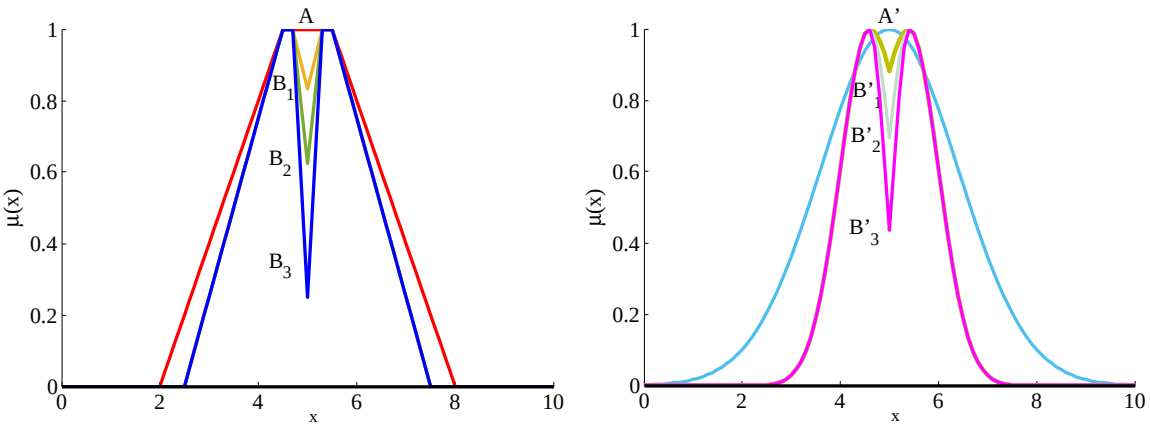
(b) Gaussian Fuzzy Sets

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A_1, B\}$	$\{A_2, B\}$	$\{A_3, B\}$	$\{A'_1, B'\}$	$\{A'_2, B'\}$	$\{A'_3, B'\}$
S_J	0.2112	0.2219	0.224	0.1282	0.1311	0.1325
S_C	0.5178	0.5533	0.5764	0.3835	0.4119	0.4322
S_P	0.1925	0.2567	0.3209	0.1099	0.1466	0.1832
S_W	0.1991	0.1902	0.1767	0.1231	0.1167	0.1116
S_R	0.6473	0.6916	0.5764	0.4794	0.5148	0.4322
$S^{T1}_{S_h v:min}$	0.2165	0.2406	0.2553	0.1395	0.1521	0.1614
$S^{T1}_{S_h v:prod}$	0.1932	0.1789	0.1612	0.0855	0.0762	0.0687

Fig. 5.6 Trapezoidal and Gaussian T1 fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the height of the non-normal fuzzy set is increasing while the other set is also non-normal.

kept non-normal, while A_1 and A'_1 gradually turn into normal fuzzy sets, A_3 and A'_3. Here again, the similarity within the pairs should rise as the height of the non-normal fuzzy set increases, and its value reaches the maximum when both sets in a pair share the same height. The results show that all measures except S_R behave non-intuitively by giving less similarity for the pairs, $\{A_2, B\}$ and $\{A'_2, B'\}$.

(5) Non-convex, normal fuzzy sets: Finally, three pairs of fuzzy sets – trapezoidal ($\{A, B_1\}$, $\{A, B_2\}$, and $\{A, B_3\}$) and Gaussian ($\{A', B'_1\}$, $\{A', B'_2\}$, and $\{A', B'_3\}$) are considered in Fig. 5.7 where B_1, B_2, and B_3 are both non-convex fuzzy sets and



(a) Trapezoidal Fuzzy Sets (b) Gaussian Fuzzy Sets

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.8429	0.8245	0.7929	0.4058	0.3647	0.3220
S_C	1.0	1.0	1.0	0.9829	0.9773	0.9726
S_P	0.906	0.883	0.841	0.5199	0.4619	0.4016
S_W	0.6575	0.647	0.6282	0.1639	0.1493	0.1342
S_R	1.0	1.0	1.0	0.9829	0.9773	0.9726
$S^{T1}_{S_h v:min}$	0.8429	0.8245	0.7929	0.4062	0.3650	0.3222
$S^{T1}_{S_h v:prod}$	0.60	0.4571	0.40	0.4022	0.3613	0.3186

Fig. 5.7 Trapezoidal and Gaussian T1 fuzzy sets and their comparison according to the proposed vertically extended similarity measure $S^{T1}_{S_h v}$ and five other fuzzy set-theoretic similarity measures (S_J, S_C, S_P, S_W, and S_R) where the 'degree of non-convexity' of the fuzzy set is increasing.

proper subsets of A (similarly, B'_1, B'_2, and B'_3 are non-convex sets and proper subsets of A'). Here, the 'degree of non-convexity' is increased in a gradual manner from B_1 to B_3 (also, from B'_1 to B'_3). As the non-convexity 'increases', the degree of overlap, or, more formally, the cardinality of the intersection between the pairs declines, intuitively resulting in a decreasing degree of similarity. Here, both S_C and S_R measures show a perhaps unexpected behaviour for trapezoidal fuzzy sets – yielding a similarity of 1 for all pairs, while following the expected decreasing trend for the Gaussian fuzzy sets. All other measures show a decreasing trend for all pairs in respect to the similarity results.

Table 5.1 presents the summary of the experiment results, which show that all measures depict expected behaviour for varying degrees of overlap. Both S_C and S_R measures deviate from intuition while changing the degree of subsethood. For variations in the height of pairs of non-normal fuzzy sets, only S_R measure follows intuition. Finally, with a rising degree of non-convexity, both S_C and S_R measures show unexpected results for trapezoidal fuzzy sets while meeting the expectation for the Gaussian fuzzy sets. Overall, S_J, S_P, S_W and the proposed extended $S^{T1}_{S_h v}$ (with

Table 5.1 Summary of Experiment Results (✓- expected behaviour and × - unexpected behaviour).

Similarity Measure	Changing Features between T1 Fuzzy Sets				
				Increasing height of non-normal fuzzy set	
	Increase Overlap	Raising Subsethood	Increase Non-convexity	One non-normal, other normal fuzzy sets	Both non-normal fuzzy sets
S_J	✓	✓	✓	✓	×
S_C	✓	×	×	✓	×
S_P	✓	✓	✓	✓	×
S_W	✓	✓	✓	✓	×
S_R	✓	×	×	✓	✓
$S^{T1}_{S_h v:min}$	✓	✓	✓	✓	×
$S^{T1}_{S_h v:prod}$	✓	✓	✓	✓	×

minimum and product *t-norms*) measures largely follow intuition for a variety of fuzzy sets—except for pairs of fuzzy sets where both are non-normal.

Remark 5.1:
The results from Table 5.1 show that most of the vertical-slice based extended measures behave non-intuitively while comparing non-normal fuzzy sets. This finding motivates us to explore the utility of extending the new similarity measure at Eq. (3.2) based on the α-cuts of fuzzy sets.

The next section presents an α-cut based extension of the bidirectional subsethood based similarity measure (Eq. (3.2) in Chapter 3) for T1 fuzzy sets and discusses its major properties.

5.4 Proposed α-cut Based Extension

In this section, an α-cut (or horizontal-slice) based extension of the new similarity measure (S_{S_h}) is proposed to compute similarity between T1 fuzzy sets as α-cuts of a normal and convex T1 fuzzy set are closed intervals. However, an α-cut can result in an empty set or discontinuous interval in case of a non-normal and non-convex fuzzy set. In general, the α-cut based extended similarity measure can not be directly applied to compare non-normal and non-convex fuzzy sets, and thus requires further expansion to deal with these properties [20][142]. Hence, the proposed α-cut based extended measure is further refined in sections 5.4.3 and 5.4.4 to allow the comparison of non-normal and non-convex fuzzy sets.

5.4.1 Proposed α-cut Based Extension of the Bidirectional Subsethood Based Similarity Measure

The proposed α-cut based extended similarity measure determines the similarity between fuzzy sets A and B by taking into account how much they are similar over their α-cuts. Thus, in this extension, the similarity between A and B is first determined in respect to each α-cut using the new similarity measure, S_{S_h}. Then, all similarity results along the α-cuts are combined to make an overall similarity between A and B.

5.4.1.1 Similarity between Fuzzy Sets at a Given α-cut

Suppose, $\overline{A}_{\alpha_i} = \{x|\mu_A(x) \geq \alpha_i\}$ and $\overline{B}_{\alpha_i} = \{x|\mu_B(x) \geq \alpha_i\}$ are the α-cuts for normal and convex T1 fuzzy sets A and B respectively at a given α_i. As $\overline{A}_{\alpha_i}$ and $\overline{B}_{\alpha_i}$ are closed intervals, the new similarity measure, S_{S_h} at Eq. (3.2) is directly applied to estimate their similarity. For doing so, the subsethood for $\overline{A}_{\alpha_i}$ and $\overline{B}_{\alpha_i}$ are first determined using the definition of subsethood (S_h) for intervals at Eq. (3.1). Thus, the subsethood of $\overline{A}_{\alpha_i}$ in respect to $\overline{B}_{\alpha_i}$ is defined as

$$S_h(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}) = \frac{\left|\overline{A}_{\alpha_i} \cap \overline{B}_{\alpha_i}\right|}{\left|\overline{A}_{\alpha_i}\right|}, \tag{5.4}$$

where $S_h(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i})$ is calculated by taking the ratio of the cardinality of the intersection of $\overline{A}_{\alpha_i}$ and $\overline{B}_{\alpha_i}$, and the cardinality of $\overline{A}_{\alpha_i}$ (see Chapter 3). In a similar manner, the subsethood of $\overline{B}_{\alpha_i}$ in respect to $\overline{A}_{\alpha_i}$ is defined as

$$S_h(\overline{B}_{\alpha_i}, \overline{A}_{\alpha_i}) = \frac{\left|\overline{A}_{\alpha_i} \cap \overline{B}_{\alpha_i}\right|}{\left|\overline{B}_{\alpha_i}\right|}. \tag{5.5}$$

Now, using the S_{S_h} measure at Eq. (3.2), the similarity between $\overline{A}_{\alpha_i}$ and $\overline{B}_{\alpha_i}$ at the given α_i is calculated as follows,

$$S_{S_h}\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right) = \bigstar\left(S_h\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right), S_h\left(\overline{B}_{\alpha_i}, \overline{A}_{\alpha_i}\right)\right). \tag{5.6}$$

5.4.1.2 Overall Similarity between Fuzzy Sets Based on the α-cuts

Subsection 5.4.1.1 has presented the similarity calculation between α-cuts, $\overline{A}_{\alpha_i}$ and $\overline{B}_{\alpha_i}$ of fuzzy sets A and B at a given α-cut (α_i). When all such similarities between A and B are determined in respect to all α-cuts, they are aggregated using an average function to determine an overall similarity for them. Thus, the proposed α-cut based extended similarity measure for the normal and convex T1 fuzzy sets A and B is defined as,

$$S^{T1}_{S_h\alpha}(A, B) = \frac{1}{|\alpha_i|}\sum_{\alpha_i=0}^{1} S_{S_h}\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right), \tag{5.7}$$

where $S_{S_h}\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right)$ is the similarity between A and B at each α_i and $|\alpha_i|$ is the total number of α-cuts used for A and B. Using Eq. (5.6), Eq. (5.7) can be alternatively expressed as follows,

$$S^{T1}_{S_h\alpha}(A, B) = \frac{1}{|\alpha_i|}\sum_{\alpha_i=0}^{1} \bigstar\left(S_h(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}), S_h(\overline{B}_{\alpha_i}, \overline{A}_{\alpha_i})\right). \tag{5.8}$$

It is worth mentioning that for aggregating similarities between A and B along all α-cuts, a suitable option is to use an average function to estimate their overall similarity.

The next subsection discusses and proves the major properties of the α-cut based extended $S^{T1}_{S_h\alpha}$ measure.

5.4.2 Properties of the α-cut Based Extension

In this subsection, the major properties of the α-cut based extended $S^{T1}_{S_h\alpha}$ measure (Eq. (5.7)) are discussed and proved.

Theorem 5.7 (Boundedness): $S^{T1}_{S_h\alpha}(A,B)$ at Eq. (5.7) is always bounded by 0 and 1. That is, $0 \leq S^{T1}_{S_h\alpha}(A,B) \leq 1$ for fuzzy sets A and B.

Proof: $S_{S_h}\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right)$ is always within the range between 0 and 1 at each α_i (see Chapter 3), which follows that $S^{T1}_{S_h\alpha}(A,B)$ is also bounded by $[0,1]$. ■

Theorem 5.6:

The α-cut based extended $S^{T1}_{S_h\alpha}$ measure in Eq. (5.7) satisfies all common properties [C1] - [C4] of a fuzzy set similarity measure (discussed in Chapter 2) for T1 fuzzy sets, A, B and C. That is,

(1) $S^{T1}_{S_h\alpha}(A,B) = S^{T1}_{S_h\alpha}(B,A)$ (symmetry);

(2) $S^{T1}_{S_h\alpha}(A,B) = 1 \iff A = B$ (reflexivity);

(3) $S^{T1}_{S_h\alpha}(A,B) \geq S^{T1}_{S_h\alpha}(A,C)$ when $A \subseteq B \subseteq C$ (transitivity);

(4) $S^{T1}_{S_h\alpha}(A,B) > 0$ when $A \cap B \neq \emptyset$; otherwise, $S^{T1}_{S_h\alpha}(A,B) = 0$ (overlapping).

Proof: Below the proofs for all the properties (1–4) are given.

(1) $S_{S_h}\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right)$ is always symmetric (see Chapter 3) at each α_i, thus making $S^{T1}_{S_h\alpha}(A,B)$ symmetric too.

(2) $S_{S_h}\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right)$ is always reflexive (see Chapter 3) at each α_i, implying that $S^{T1}_{S_h\alpha}(A,B)$ is reflexive.

(3) if $A \subseteq B \subseteq C$, then at each α_i, $\overline{A}_{\alpha_i} \subseteq \overline{B}_{\alpha_i} \subseteq \overline{C}_{\alpha_i}$. As S_{S_h} always maintains the transitivity property (see Chapter 3), so at any α_i, $S_{S_h}(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}) \geq S_{S_h}(\overline{A}_{\alpha_i}, \overline{C}_{\alpha_i})$. It implies that $S^{T1}_{S_h\alpha}(A, B) \geq S^{T1}_{S_h\alpha}(A, C)$.

(4) When A and B are overlapping, then $|\overline{A}_{\alpha_i} \cap \overline{B}_{\alpha_i}| \neq 0$ for some/all α_i. In this case, $S_{S_h}(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}) = \bigstar \Big(S_h(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}), S_h(\overline{B}_{\alpha_i}, \overline{A}_{\alpha_i})\Big) > 0$ (see Chapter 3), hence making $S^{T1}_{S_h\alpha}(A, B) > 0$. In other case, for disjoint sets A and B, $|\overline{A}_{\alpha_i} \cap \overline{B}_{\alpha_i}| = 0$ for all α_i, which makes $S_{S_h}(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}) = 0$ for all α_i. As a result, $S^{T1}_{S_h\alpha}(A, B) = 0$. ■

The α-cut based extended $S^{T1}_{S_h\alpha}$ measure in Eq. (5.7) is developed for normal and convex T1 fuzzy sets, not suitable for non-normal T1 fuzzy sets. Therefore, the next subsection presents a refinement of this extended $S^{T1}_{S_h\alpha}$ measure with a view to comparing non-normal T1 fuzzy sets.

5.4.3 Dealing with Non-normal Fuzzy Sets

The proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) at Eq. (5.7) is particularly designed to compare normal and convex T1 fuzzy sets. As mentioned, a fuzzy set can also be non-normal and any α-cut exceeding the height of such fuzzy set results in an empty set ($\emptyset$). The proposed extended $S_{S_h\alpha}$ measure thus needs further expansion to deal with the similarity computation of non-normal fuzzy sets. The refinement of the proposed $S_{S_h\alpha}$ measure for non-normal and convex T1 fuzzy sets is discussed below.

Let μ^* denote the maximum membership values (or heights) of T1 fuzzy sets A and B such that $\forall x_1, x_2 \in X$

$$\mu^* = \max[h(A), h(B)] = max[max_{x_1}(\mu_A(x_1)), max_{x_2}(\mu_B(x_2))]. \tag{5.9}$$

where $h(A)$ and $h(B)$ are the height of A and B respectively.

The proposed refinement of the $S^{T1}_{S_h\alpha}$ measure at Eq. (5.7) for non-normal and convex T1 fuzzy sets is as follows,

$$S^{T1:NN}_{S_h\alpha}(A,B) = \frac{1}{|\alpha_i|}\sum_{\alpha_i=0}^{\mu^*} S_{S_h}\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right), \tag{5.10}$$

where $\overline{A}_{\alpha_i} \neq \emptyset$ and $\overline{B}_{\alpha_i} \neq \emptyset$. $|\alpha_i|$ is the total number of α-cuts within 0 and μ^*. It is noted that the name $S^{T1:NN}_{S_h\alpha}$ stands for the α-cut based extended similarity measure for non-normal and convex T1 fuzzy sets.

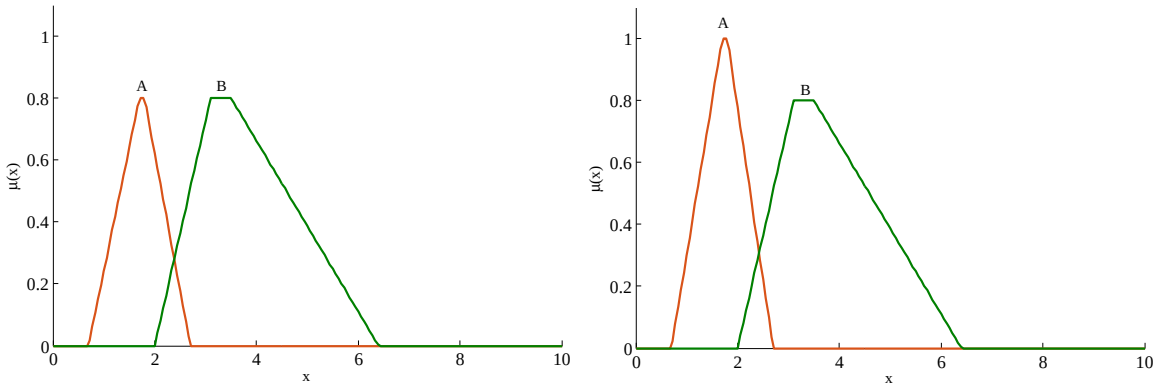
Example 5.1:

Figure 5.8 presents two different pairs of T1 fuzzy sets where in the first pair (Pair-1), both sets are non-normal with the same maximum membership value of 0.8; whereas in the second pair (Pair-2), one set is normal and the other set is non-normal with maximum membership value of 0.8. As mentioned, two fuzzy sets with the same maximum membership value (or the same height) are intuitively more similar than two fuzzy sets with different maximum membership values (or different heights). Table 5.2 presents similarity results for Pair-1 and Pair-2 using the extended $S^{T1:NN}_{S_h\alpha}$ measure (with minimum *t-norm*). It shows that the $S^{T1:NN}_{S_h\alpha}$ measure generates higher similarity result for Pair-1 as compared to Pair-2. Thus, the proposed extended $S^{T1:NN}_{S_h\alpha}$ measure behaves intuitively for non-normal fuzzy sets.

Remark 5.2:

When both A and B are normal and convex fuzzy sets, then $\mu^* = 1$, which makes the $S^{T1:NN}_{S_h\alpha}(A,B)$ measure in Eq. (5.10) identical to the $S^{T1}_{S_h\alpha}(A,B)$ measure in Eq. (5.7). Hence,

$$S^{T1:NN}_{S_h\alpha}(A,B) \equiv S^{T1}_{S_h\alpha}(A,B).$$



(a) Pair-1 (b) Pair-2

Fig. 5.8 Two different pairs of T1 fuzzy sets where both non-normal fuzzy sets are in Pair-1, and one normal and one non-normal fuzzy sets are in Pair-2.

Table 5.2 Similarity results for the pairs of fuzzy sets (in Fig. 5.8) using the $S^{T1:NN}_{S_h\alpha}$ measure.

Pair of Fuzzy Sets	Similarity Result
Pair-1	0.1771
Pair-2	0.1549

As mentioned, the α-cut based extended $S^{T1}_{S_h\alpha}$ measure in Eq. (5.7) is developed only for normal and convex T1 fuzzy sets. Therefore, the next subsection presents a refinement of the $S^{T1}_{S_h\alpha}$ measure for comparing non-convex T1 fuzzy sets.

5.4.4 Dealing with Non-convex Fuzzy Sets

The proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) at Eq. (5.7) was introduced for comparing normal and convex fuzzy sets. However as mentioned, a fuzzy set can be non-convex, and an α-cut of a non-convex fuzzy set may result in a discontinuous interval. Figure 5.9 shows a pair of T1 fuzzy sets A and B where the α-cut of the non-convex set B at $\alpha = 0.3$ results in a discontinuous interval. The proposed $S^{T1}_{S_h\alpha}$ measure can not be directly used to compare non-convex fuzzy sets and

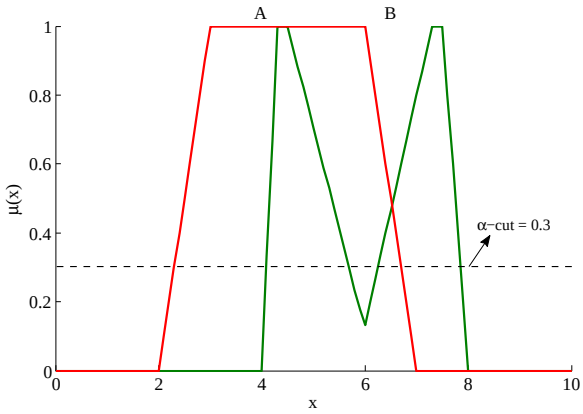


Fig. 5.9 A pair of convex and non-convex fuzzy sets.

requires further refinement. The refinement of the proposed $S_{S_h\alpha^{T1}}$ measure for the normal and non-convex fuzzy sets is discussed below.

In Chapter 4, a generalized framework (Eq. (4.1)) has been introduced in respect to the bidirectional subsethood based similarity measure (Eq. (3.2) in Chapter 3), which provides an efficient comparison between discontinuous intervals. In this refinement, the generalized framework involving the new measure (Eq. (3.2) in Chapter 3) is used to measure similarity between discontinuous intervals resulting from the α-cuts of non-convex fuzzy sets.

Let $\overline{\overline{A}}_{\alpha_i} = \{\overline{A}_{\alpha_{i1}}, ..., \overline{A}_{\alpha_{im}}\}$ and $\overline{\overline{B}}_{\alpha_i} = \{\overline{B}_{\alpha_{i1}}, ..., \overline{B}_{\alpha_{in}}\}$ be discontinuous intervals arising from normal and non-convex fuzzy sets, A and B at α_i. Using Eq. (4.1) in respect to the S_{S_h} measure in Eq. (3.2), the similarity between $\overline{\overline{A}}_{\alpha_i}$ and $\overline{\overline{B}}_{\alpha_i}$ is calculated as follows,

$$S_{S_h}\left(\overline{\overline{A}}_{\alpha_i}, \overline{\overline{B}}_{\alpha_i}\right) = \frac{1}{\max(m,n)} \sum_{p=1}^{m} \sum_{q=1}^{n} S_{S_h}\left(\overline{A}_{\alpha_{ip}}, \overline{B}_{\alpha_{iq}}\right), \tag{5.11}$$

where m and n are the total number of continuous subintervals in $\overline{\overline{A}}_{\alpha_i}$ and $\overline{\overline{B}}_{\alpha_i}$ respectively. $\overline{A}_{\alpha_{ip}} \neq \emptyset$ and $\overline{B}_{\alpha_{iq}} \neq \emptyset$ are the p-th and q-th continuous subintervals of $\overline{\overline{A}}_{\alpha_i}$ and $\overline{\overline{B}}_{\alpha_i}$ respectively, $1 \leq p \leq m$ and $1 \leq q \leq n$.

Now, the proposed extension of the $S^{T1}_{S_h\alpha}$ measure in Eq. (5.7) for normal and non-convex fuzzy sets is given below,

$$S^{T1:NC}_{S_h\alpha}(A,B) = \frac{1}{|\alpha_i|}\sum_{\alpha_i=0}^{1} S_{S_h}\left(\overline{\overline{A}}_{\alpha_i}, \overline{\overline{B}}_{\alpha_i}\right). \tag{5.12}$$

It is noted that $S^{T1:NC}_{S_h\alpha}$ stands for the α-cut based extended similarity measure for normal and non-convex T1 fuzzy sets.

Remark 5.3:

When both A and B are normal and convex fuzzy sets, then $S_{S_h}\left(\overline{\overline{A}}_{\alpha_i}, \overline{\overline{B}}_{\alpha_i}\right) \equiv S_{S_h}\left(\overline{A}_{\alpha_i}, \overline{B}_{\alpha_i}\right)$ which makes the $S^{T1:NC}_{S_h\alpha}$ measure in Eq. (5.12) identical to the $S^{T1}_{S_h\alpha}$ measure in Eq. (5.7). Hence,

$$S^{T1:NC}_{S_h\alpha}(A,B) \equiv S^{T1}_{S_h\alpha}(A,B).$$

Remark 5.4:

For comparing non-normal and convex T1 fuzzy sets, Eq. (5.10) is applied; whereas Eq. (5.12) is used for comparing normal and non-convex T1 fuzzy sets. Therefore, by combining Eqs. (5.10) and (5.12) results in an extended measure which can assess the similarity between any types (normal, non-normal, convex and non-convex) of T1 fuzzy sets. The combined α-cut based extended S_{S_h} measure for T1 fuzzy sets, A and B stands as follows,

$$S^{T1}_{S_h\alpha}(A,B) = \frac{1}{|\alpha_i|}\sum_{\alpha_i=0}^{\mu^*} S_{S_h}(\overline{\overline{A}}_{\alpha_i}, \overline{\overline{B}}_{\alpha_i}), \tag{5.13}$$

where μ^* is measured using Eq. (5.9). Note that $\overline{\overline{A}}_{\alpha_i}$ and $\overline{\overline{B}}_{\alpha_i}$ are the α-cuts of A and B respectively which are generally continuous, but can also be discontinuous. Again, A and B are generally normal, but can also be non-normal (in this case $\mu^* \leq 1$).

The next subsection demonstrates the behaviour of the proposed α-cut based extended $S^{T1}_{S_h\alpha}$ measure at Eq. (5.13) in respect to the proposed vertical-slice based extended $S^{T1}_{S_hv}$ measure at Eq. (5.7) and key existing fuzzy set-theoretic similarity measures using the same synthetic T1 fuzzy sets as followed in subsection 5.3.3.

5.4.5 Comparison with Current Literature

This section compares the performance of the proposed α-cut based extended $S^{T1}_{S_h\alpha}$ measure at Eq. (5.13) with the proposed vertical-slice based extended $S^{T1}_{S_hv}$ measure Eq. (5.7) and key existing fuzzy set-theoretic similarity measures (i.e., S_J, S_C, S_P, S_W, and S_R measures) using the same synthetic T1 fuzzy sets as shown in Figs. 5.3 to 5.7 in subsection 5.3.3. Here, Eq. (5.13) is applied for the comparison as it can be used for any types of T1 fuzzy sets, let alone non-normal and non-convex fuzzy sets. Again, both minimum and product *t-norm*s are applied for both extended measures. Further, for the similarity calculation, 1000 discretizations (vertical-slices) along the whole support of the fuzzy sets and the same number of discretizations (α-cuts) along the maximum membership values within the fuzzy sets are considered. As the same set of synthetic T1 fuzzy sets are being used in this experiment, here we do not redraw the same figures (Figs. 5.3 to 5.7), rather only the similarity results of different measures are shown.

(1) Convex-normal fuzzy sets: Table 5.3 presents the similarity results of different measures for three pairs of trapezoidal ($\{A, B_1\}$, $\{A, B_2\}$, and $\{A, B_3\}$) and Gaussian ($\{A', B'_1\}$, $\{A', B'_2\}$, and $\{A', B'_3\}$) T1 fuzzy sets (see Fig. 5.3) where the degree of overlap between the pairs have been increased gradually. All pairs are normal and convex, and they are shifted from completely disjoint to an increasing overlap along the x-axis. Here, similarity is expected to be higher with an increased overlapping between the pairs. Table 5.3 shows that all similarity measures behave intuitively.

Table 5.3 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where their overlap are increased from right to left.

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.0	0.0326	0.3968	0.0	0.0321	0.373
S_C	0.0	0.2778	0.8333	0.0	0.1724	0.8226
S_P	0.0	0.0322	0.724	0.0	0.0301	0.6766
S_W	0.0	0.0552	0.2728	0.0	0.0689	0.2929
S_R	0.0	0.2778	0.8333	0.0	0.1724	0.8226
$S^{T1}_{S_hv:min}$	0.0	0.0631	0.5682	0.0	0.0625	0.5426
$S^{T1}_{S_hv:prod}$	0.0	0.004	0.3228	0.0	0.0039	0.2949
$S^{T1}_{S_h\alpha:min}$	0.0	0.0456	0.4495	0.0	0.023	0.3589
$S^{T1}_{S_h\alpha:prod}$	0.0	0.007	0.2716	0.0	0.002	0.1881

(2) Convex-normal fuzzy sets (with proper subsets): Here, three pairs of (normal and convex) trapezoidal ($\{A, B_1\}$, $\{A, B_2\}$, and $\{A, B_3\}$) and Gaussian ($\{A', B'_1\}$, $\{A', B'_2\}$, and $\{A', B'_3\}$) fuzzy sets are considered (see Fig. 5.4), where all B_1, B_2, and B_3 are proper subsets of A (as B'_1, B'_2 and B'_3 are of A'). In this case, the degree of subsethood between the pairs is gradually increased. So, we expect higher similarity as the degree of subsethood gradually increases. Table 5.4 shows that both S_C and S_R measures give a similarity of 1 for all pairs despite being non-identical, while all other measures including the proposed α-cut based extended measure return an increasing degree of similarity in line with the expectation.

(3) Convex fuzzy sets (with one normal and one non-normal fuzzy set): Table 5.5 presents similarity results for three pairs of trapezoidal ($\{A, B_1\}$, $\{A, B_2\}$, and $\{A, B_3\}$) and Gaussian ($\{A', B'_1\}$, $\{A', B'_2\}$, and $\{A', B'_3\}$) fuzzy sets (see Fig. 5.5) where A and A' are normal sets while B_1 and B'_1 are initially non-normal, but their heights are gradually increased and turned into normal sets, B_3 and B'_3 respectively.

Table 5.4 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where the degree of subsethood between fuzzy sets is increasing.

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.3429	0.5143	0.7143	0.2858	0.5002	0.7145
S_C	1.0	1.0	1.0	1.0	1.0	1.0
S_P	0.4422	0.6275	0.817	0.3885	0.6325	0.822
S_W	0.2042	0.3281	0.5001	0.1045	0.2024	0.3576
S_R	1.0	1.0	1.0	1.0	1.0	1.0
$S^{T1}_{S_hv:min}$	0.3429	0.5143	0.7143	0.2858	0.5002	0.7145
$S^{T1}_{S_hv:prod}$	0.3429	0.5143	0.7143	0.2858	0.5002	0.7145
$S^{T1}_{S_h\alpha:min}$	0.3488	0.5023	0.7233	0.2733	0.4715	0.6696
$S^{T1}_{S_h\alpha:prod}$	0.3488	0.5023	0.7233	0.2733	0.4715	0.6696

Table 5.5 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where the height of the non-normal fuzzy set is increasing while the other set remains normal.

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.0313	0.0515	0.0541	0.0216	0.0304	0.0316
S_C	0.1108	0.2658	0.3676	0.0587	0.1249	0.1724
S_P	0.0177	0.0532	0.0573	0.0089	0.0268	0.0279
S_W	0.0443	0.0662	0.0721	0.0389	0.0418	0.0434
S_R	0.1108	0.2658	0.3676	0.0587	0.1249	0.1724
$S^{T1}_{S_hv:min}$	0.0389	0.093	0.0961	0.0281	0.0588	0.0625
$S^{T1}_{S_hv:prod}$	0.0051	0.0096	0.0111	0.0024	0.0035	0.0042
$S^{T1}_{S_h\alpha:min}$	0.0195	0.0454	0.0621	0.0132	0.0197	0.0244
$S^{T1}_{S_h\alpha:prod}$	0.0051	0.0115	0.0159	0.002	0.005	0.008

As the height of the non-normal fuzzy set increases, the similarity within the pairs should intuitively rise, and their similarity reaches the maximum value when both sets

in a pair share the same height. Here, all measures maintain expected increasing trend in similarity results.

(4) Convex fuzzy sets (with both fuzzy sets non-normal): Table 5.6 shows similarity results of different measures for the three pairs of trapezoidal ($\{A_1, B\}$, $\{A_2, B\}$, and $\{A_3, B\}$) and Gaussian ($\{A'_1, B'\}$, $\{A'_2, B'\}$, and $\{A'_3, B'\}$) fuzzy sets (see Fig. 5.6) where initially both fuzzy sets are convex and non-normal (i.e., $\{A_1, B\}$ and $\{A'_1, B'\}$). In this experiment, B and B' are kept non-normal, while A_1 and A'_1 gradually turn into normal fuzzy sets, A_3 and A'_3. Here again, the similarity within the pairs should rise as the height of the non-normal fuzzy set increases, and its value reaches the maximum when both sets in a pair share the same height. The results show that only the proposed $S^{T1}_{S_h\alpha}$ and S_R measures behave intuitively by showing the pairs, $\{A_2, B\}$ and $\{A'_2, B'\}$ to be more similar than all other pairs.

Table 5.6 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where the height of the non-normal fuzzy set is increasing while the other set is also non-normal.

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.2112	0.2219	0.224	0.1269	0.1311	0.1325
S_C	0.5178	0.5533	0.5764	0.3835	0.4119	0.4322
S_P	0.1925	0.2567	0.3209	0.1099	0.1466	0.1832
S_W	0.1991	0.1902	0.1767	0.1231	0.1167	0.1116
S_R	0.6473	0.6916	0.5764	0.4794	0.5148	0.4322
$S^{T1}_{S_hv:min}$	0.2165	0.2406	0.2553	0.1395	0.1521	0.1614
$S^{T1}_{S_hv:prod}$	0.1932	0.1789	0.1612	0.0855	0.0762	0.0687
$S^{T1}_{S_h\alpha:min}$	0.1553	0.1771	0.1549	0.0849	0.0941	0.0812
$S^{T1}_{S_h\alpha:prod}$	0.1449	0.1563	0.1302	0.0681	0.0723	0.0603

(5) Non-convex, normal fuzzy sets: Table 5.7 shows similarity results of different measures for three pairs of fuzzy sets – trapezoidal ($\{A, B_1\}$, $\{A, B_2\}$, and $\{A, B_3\}$) and Gaussian ($\{A', B'_1\}$, $\{A', B'_2\}$, and $\{A', B'_3\}$) (see Fig. 5.7) where B_1, B_2, and B_3 are both non-convex fuzzy sets and proper subsets of A (similarly, B'_1, B'_2, and B'_3 are non-convex sets and proper subsets of A'). Here, the 'degree of non-convexity' from B_1 to B_3 (also, B'_1 to B'_3) is increased in a gradual manner. As the non-convexity 'increases', the degree of overlap between the pairs declines, resulting in a decreasing degree of similarity. Here, both S_C and S_R measures show perhaps an unexpected behaviour for trapezoidal fuzzy sets, yielding a similarity of 1 for all pairs, while following the expected decreasing trend for the Gaussian fuzzy sets. All other measures including the proposed extended $S^{T1}_{S_h\alpha}$ measure show a decreasing trend for all pairs in respect to similarity results.

Table 5.7 Comparison of the proposed α-cut based extended similarity measure ($S^{T1}_{S_h\alpha}$) with the proposed vertical-slice based extended similarity measure ($S^{T1}_{S_hv}$) and with key fuzzy set-theoretic measures (S_J, S_C, S_P, S_W, and S_R) for the trapezoidal and Gaussian fuzzy sets where the 'degree of non-convexity' of fuzzy set is increasing.

Similarity Measure	Pair of Trapezoidal Fuzzy Sets			Pair of Gaussian Fuzzy Sets		
	$\{A, B_1\}$	$\{A, B_2\}$	$\{A, B_3\}$	$\{A', B'_1\}$	$\{A, B'_2\}$	$\{A', B'_3\}$
S_J	0.8429	0.8245	0.7929	0.4058	0.3647	0.3220
S_C	1.0	1.0	1.0	0.9829	0.9773	0.9726
S_P	0.906	0.883	0.841	0.5199	0.4619	0.4016
S_W	0.6575	0.647	0.6282	0.1639	0.1493	0.1342
S_R	1.0	1.0	1.0	0.9829	0.9773	0.9726
$S^{T1}_{S_hv:min}$	0.8429	0.8245	0.7929	0.4062	0.3650	0.3222
$S^{T1}_{S_hv:prod}$	0.60	0.4571	0.40	0.4022	0.3613	0.3186
$S^{T1}_{S_h\alpha:min}$	0.8295	0.7955	0.7534	0.7219	0.5999	0.5329
$S^{T1}_{S_h\alpha:prod}$	0.6735	0.5997	0.5376	0.6324	0.5682	0.5052

In summary, Table 5.8 presents a glimpse of the behaviour of all similarity measures for various features of T1 fuzzy sets. The overall results reveal that only the α-cut

based extended similarity measure in Eq. (5.13) provides intuitive results for all cases, whereas the other measures give non-intuitive results in at least some of cases.

Table 5.8 Summary of Experiment Results (✓- expected behaviour and × - unexpected behaviour).

Similarity Measure	Changing Features between Fuzzy Sets				
	Increase Overlap	Raising Subsethood	Increase Non-convexity	Increasing height of non-normal fuzzy set	
				One non-normal, other normal fuzzy sets	Both non-normal fuzzy sets
S_J	✓	✓	✓	✓	×
S_C	✓	×	×	✓	×
S_P	✓	✓	✓	✓	×
S_W	✓	✓	✓	✓	×
S_R	✓	×	×	✓	✓
$S^{T1}_{S_h v:min}$	✓	✓	✓	✓	×
$S^{T1}_{S_h v:prod}$	✓	✓	✓	✓	×
$S^{T1}_{S_h \alpha:min}$	✓	✓	✓	✓	✓
$S^{T1}_{S_h \alpha:prod}$	✓	✓	✓	✓	✓

Table 5.9 Execution time (ms) for different discretization levels (250, 500, 750, 1000) along the whole support of Pair-1 of the Gaussian fuzzy sets (A', B'_1) in Fig. 5.4(b).

Similarity Measure	No. of Discretization			
	250	500	750	1000
$S_J(A,B)$	17.43	29.62	36.28	41.86
$S_C(A,B)$	13.42	25.74	33.20	35.75
$S_P(A,B)$	15.40	27.40	34.08	41.17
$S_W(A,B)$	15.65	27.90	34.33	41.27
$S_R(A,B)$	14.50	26.48	33.25	38.40
$S^{T1}_{S_h v:min}(A,B)$	15.74	27.97	35.18	41.34
$S^{T1}_{S_h v:prod}(A,B)$	14.83	26.55	33.72	40.89
$S^{T1}_{S_h \alpha:min}(A,B)$	17.45	29.81	37.27	43.01
$S^{T1}_{S_h \alpha:prod}(A,B)$	16.91	28.87	35.83	41.39

Further, Table 5.9 presents a comparative analysis regarding the respective execution times in milliseconds (ms) for all measures for various levels of vertical-slices and α-cuts

(250, 500, 750, 1000). For this investigation, only the Gaussian fuzzy sets (A', B'_1) in Fig. 5.4(b) has been taken into consideration. The results show that all measures take approximately similar execution time for various levels of discretization. Here, it is important to note that the programming language (e.g., Java) to implement these similarity measures may influence their execution time.

5.5 Summary

In view of the advantages of the new bidirectional subsethood based similarity measure (Eq. (3.2) in Chapter 3) in comparing continuous and discontinuous intervals, this chapter has extended the new measure to compare various features of T1 fuzzy sets. Initially, the new measure has been extended using the vertical-slice representation ($S^{T1}_{S_h v}$ at Eq. (5.7)) for comparing T1 fuzzy sets (i.e., convex, non-convex, normal, and non-normal), which maintains essential properties [C1] - [C4] of a fuzzy set similarity measure (section 2.5 in Chapter 2). After establishing the theoretical foundation, the behavior of this extended measure, $S^{T1}_{S_h v}$ has been demonstrated and compared in respect to key existing fuzzy set-theoretic similarity measures using a series of synthetic fuzzy sets. The overall experimental results have revealed that the vertical-slice based extension of the new similarity measure, $S^{T1}_{S_h v}$ along with key S_J, S_P, and S_W measures behave intuitively for changing features (i.e., degree of overlap, subsethood, or change in the 'degree of non-convexity') of various types of T1 fuzzy sets, except for a pair of non-normal fuzzy-sets.

To avoid this limitation, the new similarity measure (Eq. (3.2) in Chapter 3) has been further extended based on the α-cuts of fuzzy sets and established its major properties. Using the same set of synthetic T1 fuzzy sets, the performance of the new α-cut based extended measure ($S^{T1}_{S_h \alpha}$ at Eq. (5.13)) has been compared with the vertical-slice based extended measure ($S^{T1}_{S_h v}$) along with key existing fuzzy set-theoretic

similarity measures. It has been revealed that only the α-cut based extended measure, ($S^{T1}_{S_h\alpha}$ at Eq. (5.13)) can avoid the counter-intuitive behaviour for non-normal fuzzy-set pair and provide similarity results in line with the expectation for all cases. Further, the investigation on the execution time relating to different similarity measures at various levels of discritization has shown that all of these measures take approximately similar execution time (in milliseconds) for the similarity calculation. These works in this chapter fulfills objective 5 of the thesis.

A part of the contribution to this chapter, particularly extending the new similarity measure vertically and its usefulness in two different applications have been published in the following two international conferences.

- **A Bidirectional Subsethood Based Similarity Measure for Fuzzy Sets**
  Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. A Bidirectional Subsethood Based Similarity Measure for Fuzzy Sets. In IEEE World Congress on Computational Intelligence (WCCI 2018), Rio de Janeiro, Brazil, 2018.
  Status: *Published*

- **Determining Firing Strengths Through A Novel Similarity Measure to Enhance Uncertainty Handling in Non-Singleton Fuzzy Logic Systems**
  Pekaslan, Direnc; Kabir, Shaily, Garibaldi, Jonathan M.; and Wagner, Christian. Determining Firing Strengths Through A Novel Similarity Measure to Enhance Uncertainty Handling in Non-Singleton Fuzzy Logic Systems. In International Joint Conference on Computational Intelligence (IJCCI 2017), Funchal, Madeira-Portugal, 2017.
  Status: *Published*

Chapter 3, 4, and 5 have shown the usefulness of the new set-theoretic similarity measure for assessing the similarity between continuous (closed)/discontinuous intervals and also between T1 fuzzy sets.

To reap further gains, the next Chapter 6 explores the utility of using the new similarity measure to extract an FM for improving robustness in FI based data aggregation. This will meet up objectives 6 and 7 of the thesis.

Chapter 6

Aggregation of Multi-Source Interval-Valued Data

6.1 Introduction

Chapter 3 identified and articulated a shortcoming (i.e., *aliasing*) of the popular set-theoretic similarity measures where they produce the same similarity for very different pairs of intervals. Hence, it established a new *robust* similarity measure (Eq. (3.2) in Chapter 3) for continuous intervals and comprehensively demonstrated its intuitive behaviour by avoiding *aliasing*. The inclusion of this new measure within the proposed generalized framework in Chapter 4 also yielded good results in similarity calculation of discontinuous intervals. Moving beyond intervals, the α-cut based extension of this new measure in Chapter 5 also provided desirable similarity results for all types of T1 fuzzy sets. To reap further gains, this chapter focuses on one of the applications of the similarity measures—the aggregation of 'similar data' from uncertain sources by using a powerful aggregation function, that is, an FI specified in respect to an FM. As the existing FMs based on popular similarity measures are often affected by *aliasing* and hence the aggregation outcome; this chapter takes advantage of the new similarity

measure (Eq. (3.2) in Chapter 3) to extract a new FM for bringing more accuracy in uncertain data aggregation when used in combination with an FI. The FI aggregates data arising from a set of sources by considering both the data and the worth of all possible subsets of sources, which is captured by a FM [122]. Generally, the FM is defined by experts or algorithms [23] or often generated from training data through various learning methods [125] (details in Chapter 2). However, constructing a FM for datasets without any external information, such as, ground-truth or expert insight about sources is challenging. To deal with this, [22] first proposed to extract different FMs directly from raw multi-source data by analyzing its different properties, such as, agreement and specificity. These FMs are commonly known as data-driven FMs [24]. At present, there are three data-driven FMs, which are built on source-agreement. They are – *FM of Agreement* (*AG*) [22], *FM of Generalized Accord* (*GenA*) [23], and *Additive Measure of Agreement* (*AA*) [24] (details in Chapter 2).

Among these FMs, the *AG* FM captures the source agreement using the intersection operation which determines overlap among multi-source data without tracking changes in their cardinality/size [23]. On the other hand, the two other FMs apply any of the similarity measures, such as the Jaccard or Dice measure to estimate similarity among data from sources, which have already been identified as subject to *aliasing* (discussed in Chapter 3). Due to the limitations of both intersection and popular similarity measures, the existing agreement-based FMs often generate same agreement and worth for very different subsets of sources which eventually affects the aggregation outcome. Further, all these FMs are not designed to deal with the situation when sources are disagreeing with each other in respect to their data.

To address the above issues, this chapter leverages the new similarity measure (Eq. (3.2) in Chapter 3) to extract a new agreement-based FM, which can avoid *aliasing* in the calculation of inter-source agreement. The new FM is also designed to capture

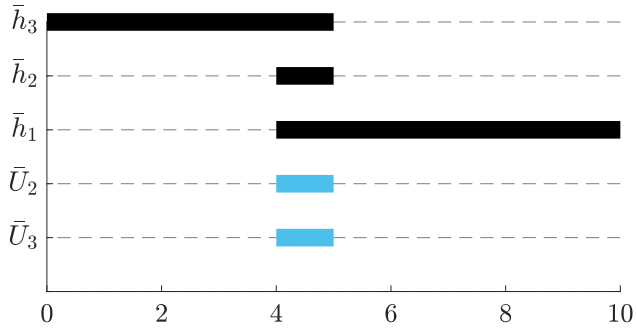
the worth information of individual sources even if they are in complete disagreement. Therefore, by avoiding *aliasing* and capturing worth for disagreeing sources, the new FM in combination with an FI brings robustness in uncertain data fusion.

In this chapter, the motivation behind the new agreement-based FM is discussed first. Then, the subsethood for a set of intervals is defined for a set of intervals and established its properties, followed by its inclusion in the new similarity measure (Eq. (3.2) in Chapter 3). Afterwards, a new FM is developed using the new similarity measure for a set of intervals. Lastly, the behaviour of the new FM in combination with an FI is demonstrated and compared against existing three other agreement-based FMs using two sets of synthetic interval data-sets and a real-world example.

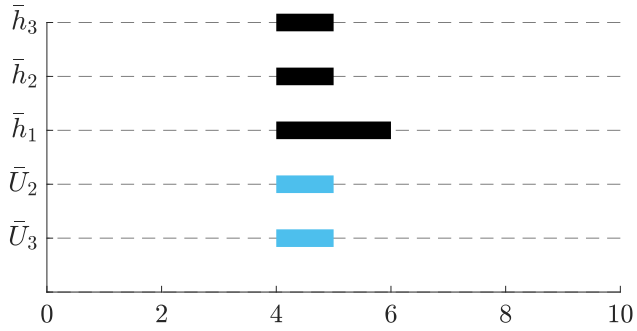
6.2 Motivation

Among the existing agreement-based FMs, the *AG* FM [22] is strongly influenced by the cardinality/size of overlap among multi-source data, which is computed by the intersection operation. Specifically, in computing agreement for a set of sources, it does not consider the overlap in relation to the cardinality/size of original data, rather relies only on their overlap [23]. As a result, it does not perform intuitively for some cases. Figure 6.1 presents an example of such cases with Scenarios 1 and 2, each having three intervals, $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$. Figure 6.1 shows that even though human observers would indicate Scenario 2 exhibits higher agreement among sources and thus should have higher worth, the *AG* FM behaves differently by assigning the same agreement and worth for the two scenarios. This is also apparent from Eq. (2.28b) (in Chapter 2) as $\tilde{g}^{AG}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}) = |\overline{U}_2(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\})|z_2 + |\overline{U}_3(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\})|z_3$ and each of $|\overline{U}_2|$ and $|\overline{U}_3|$ is the same in Scenarios 1 and 2.

Likewise, the *GenA* FM [23] also faces the same limitation, i.e., same agreement (and worth) for very different sets of sources, when it applies the Jaccard or Dice



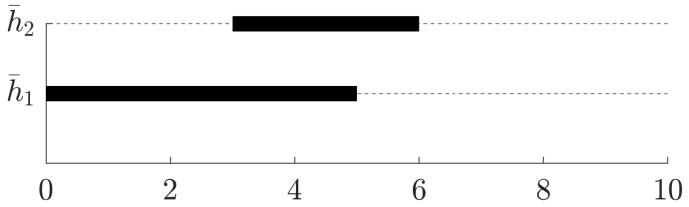
(a) Scenario 1



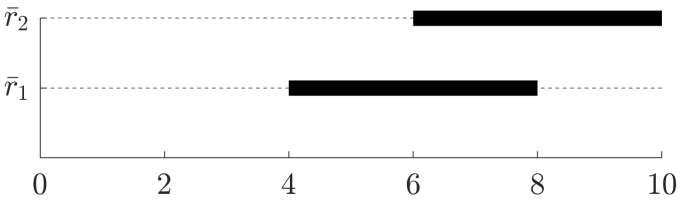
(b) Scenario 2

Fig. 6.1 Example highlighting the behaviour of the *AG* FM [22], where $\overline{U}_2$ and $\overline{U}_3$ capture the union of the intersections of all of two, respectively three source combinations as per Eq. (2.29).

similarity measure (subject to *aliasing*) to compute the agreement. For example, consider two different interval-valued datasets $\overline{h}$ and $\overline{r}$ respectively in Fig. 6.2 where $\overline{h}$ consists of two intervals $\{\overline{h}_1 = [0,5], \overline{h}_2 = [3,6]\}$ and $\overline{r}$ also has two intervals $\{\overline{r}_1 = [4,8], \overline{r}_2 = [6,10]\}$. For $\overline{h}$ (Fig. 6.2(a)), the Jaccard similarity between $\overline{h}_1$ and $\overline{h}_2$ is, $S_J(\overline{h}_1, \overline{h}_2) = \frac{|\overline{h}_1 \cap \overline{h}_2|}{|\overline{h}_1 \cup \overline{h}_2|} = \frac{2}{6} = 0.33$ and the Dice similarity between $\overline{h}_1$ and $\overline{h}_2$ is, $S_D(\overline{h}_1, \overline{h}_2) = \frac{2\times|\overline{h}_1 \cap \overline{h}_2|}{|\overline{h}_1|+|\overline{h}_2|} = \frac{4}{8} = 0.50$. In a similar manner, for $\overline{r}$ (Fig. 6.2(b)), the Jaccard similarity between $\overline{r}_1$ and $\overline{r}_2$ is, $S_J(\overline{r}_1, \overline{r}_2) = \frac{|\overline{r}_1 \cap \overline{r}_2|}{|\overline{r}_1 \cup \overline{r}_2|} = \frac{2}{6} = 0.33$ and the Dice similarity between $\overline{r}_1$ and $\overline{r}_2$ is, $S_D(\overline{r}_1, \overline{r}_2) = \frac{2\times|\overline{r}_1 \cap \overline{r}_2|}{|\overline{r}_1|+|\overline{r}_2|} = \frac{4}{8} = 0.50$. As both $\overline{h}$ and $\overline{r}$ have the same Jaccard and Dice similarity, the *GenA* FM provides the same agreement (i.e., a weighted similarity in respect to $\binom{n}{K}^{-1}$ [23]) and thus assigns the same worth to them (using Eq. (2.31b) in Chapter 2).



(a) Interval-valued set $\overline{h}$



(b) Interval-valued set $\overline{r}$

Fig. 6.2 Two different interval-valued sets $\overline{h} = [\overline{h}_1, \overline{h}_2]$ and $\overline{r} = [\overline{r}_1, \overline{r}_2]$ with equal Jaccard similarity (S_J) of 0.33 and Dice similarity (S_D) of 0.50 respectively.

However, the intervals in the sets ($\overline{h}$ and $\overline{r}$) do not appear to be equally agreeing to each other. Figure 6.2(a) shows that a large part of $\overline{h}_2$ overlaps with $\overline{h}_1$ while at the same time, only a small part of $\overline{h}_1$ overlaps with $\overline{h}_2$. It suggests that $\overline{h}_2$ is agreeing more towards $\overline{h}_1$ than $\overline{h}_1$ towards $\overline{h}_2$. On the other hand, both $\overline{r}_1$ and $\overline{r}_2$ are equally agreeing to each other (Fig. 6.2(b)). Therefore, the overall agreement and worth for $\overline{h}$ and $\overline{r}$ are expected to be different. Similar to the *GenA* FM, the *AA* FM [24] also faces the same limitation when it utilizes the Jaccard and Dice measures for computing the source-agreement.

To address this limitation, the next section introduces a new agreement-based FM involving the new similarity measure (Eq. (3.2) in Chapter 3) which shows intuitive behaviour while calculating the agreement among a set of intervals and generating the worth accordingly.

6.3 A New Fuzzy Measure Using Inter-Source Agreement

This section put forwards a new FM that utilizes the new bidirectional subsethood based similarity measure (Eq. (3.2) in Chapter 3), not subject to *aliasing*, for computing the source-agreement. For using the new similarity measure for a set of intervals, this section first defines subsethood for a set of intervals as the original definition of subsethood was introduced for only a pair of intervals (subsection 3.3.1 in Chapter 3). Based on this definition of subsethood, the new similarity measure (Eq. (3.2) in Chapter 3) is also revisited to enable it to compute similarity for a set of intervals. Finally, a new FM involving the new similarity measure is introduced to estimate the agreement for a set of intervals. In short, this section maintains the following three steps to extract a new agreement-based FM for a set of interval-valued data.

1. Define the subsethood for a set of intervals arising from a set of sources;
2. Define the new similarity measure in Eq. (3.2) based on the definition of subsethood for a set of intervals;
3. Propose a new data-driven FM by involving the new similarity measure Eq. (3.2) to estimate the inter-source agreement for a set of intervals.

The next subsection defines the subsethood for a set of intervals and discusses its major properties.

6.3.1 Defining Subsethood for a Set of Intervals

Chapter 3 introduced an axiomatic definition of subsethood for a pair of intervals. In this subsection, the original definition of subsethood (see subsection 3.3.1 in Chapter 3)

is defined for a set of intervals where the subsethood of an interval in respect to a set of intervals is estimated as a mean of its subsethood to each interval within the set.

Consider a set of intervals $\overline{h} = \{\overline{h}_1, \overline{h}_2, ..., \overline{h}_n\}$ arising from n sources. The subsethood of an interval, $\overline{h}_r$ as regards to a set of intervals $\overline{A}_i \subseteq \overline{h}$ is defined as a mean of its subsethood to each interval $\overline{h}_t$ in $\overline{A}_i$. It is expressed as

$$\begin{aligned} S_h(\overline{h}_r, \overline{A}_i) &= \frac{1}{|\overline{A}_i|} \sum_{\overline{h}_t \in \overline{A}_i} S_h(\overline{h}_r, \overline{h}_t) \\ &= \frac{1}{|\overline{A}_i|} \sum_{\overline{h}_t \in \overline{A}_i} \frac{|\overline{h}_r \cap \overline{h}_t|}{|\overline{h}_r|}, \end{aligned} \tag{6.1}$$

where $S_h(\overline{h}_r, \overline{A}_i) \rightarrow [0, 1]$ such that $S_h(\overline{h}_r, \overline{A}_i) = 1$ when $\overline{h}_r$ is a subset of all of $\overline{h}_t \in \overline{A}_i$ ($\overline{h}_r \subseteq \overline{h}_t$) and $S_h(\overline{h}_r, \overline{A}_i) = 0$ when $\overline{h}_r$ is not a subset of any of $\overline{h}_t \in \overline{A}_i$ ($\overline{h}_r \not\subset \overline{h}_t$).

The following example shows how the subsethood of an interval in respect to a set of intervals is calculated.

Example 6.1:

Consider a set of three intervals $\overline{h} = \{\overline{h}_1 = [0, 5], \overline{h}_2 = [3, 6], \overline{h}_3 = [1, 7]\}$. From Eq. (6.1), $S_h(\overline{h}_1, \{\overline{h}_2, \overline{h}_3\}) = \frac{1}{2}\left(S_h(\overline{h}_1, \overline{h}_2) + S_h(\overline{h}_1, \overline{h}_3)\right) = \frac{1}{2}(0.40 + 0.80) = 0.60$. Similarly, $S_h(\overline{h}_2, \{\overline{h}_1, \overline{h}_3\}) = 0.84$ and $S_h(\overline{h}_3, \{\overline{h}_1, \overline{h}_2\}) = 0.59$.

The next theorem discusses the key properties of subsethood for a set of intervals and their proofs.

Theorem 6.1:

The function $S_h(\overline{h}_r, \overline{A}_i)$ in Eq. (6.1) captures a measure of subsethood for one interval $\overline{h}_r$ in respect to a nonempty subset of intervals $\overline{A}_i$. As a measure of subsethood, it satisfies the following properties (see subsection 3.3.1 in Chapter 3).

(1) $S_h(\overline{h}_r, \overline{A}_i) = 1$ if and only if $\overline{h}_r \subseteq \overline{h}_t$ with all $\overline{h}_t \in \overline{A}_i$,

(2) $S_h(\overline{h}_r, \overline{A}_i) = 0$ if and only if $\overline{h}_r$ is disjoint from all $\overline{h}_t \in \overline{A}_i$,

(3) if $\overline{A}_i = \{\overline{h}_r, \overline{h}_p, \overline{h}_q\}$ and $\overline{h}_r \subseteq \overline{h}_p \subseteq \overline{h}_q$, then $S_h(\overline{h}_q, \overline{A}_i \backslash \overline{h}_q) \leq S_h(\overline{h}_p, \overline{A}_i \backslash \overline{h}_p)$ and $S_h(\overline{h}_q, \overline{A}_i \backslash \overline{h}_q) \leq S_h(\overline{h}_r, \overline{A}_i \backslash \overline{h}_r)$ where $\overline{A}_i \backslash \overline{h}_i$ is the nonempty subset of intervals without having $\overline{h}_i$, $i \in \{r, p, q\}$.

Proof: Proofs for all the properties (1–3) are given below.

(1) If $\overline{h}_r \subseteq \overline{h}_t$, then $S_h(\overline{h}_r, \overline{h}_t) = 1$ (see subsection 3.3.1 in Chapter 3). Hence, if $\overline{h}_r$ is included in all $\overline{h}_t \in \overline{A}_i$, then $\sum S_h(\overline{h}_r, \overline{h}_t) = |A_i| \times 1 = |A_i|$, implying that $S_h(\overline{h}_r, \overline{A}_i) = \frac{1}{|A_i|} \times |A_i| = 1$. Again, $S_h(\overline{h}_r, \overline{A}_i) = 1$ only when $\overline{h}_r$ is included in all $\overline{h}_t \in \overline{A}_i$.

(2) If $\overline{h}_r$ and $\overline{h}_t$ are disjoint, then $S_h(\overline{h}_r, \overline{h}_t) = 0$ (see subsection 3.3.1 in Chapter 3). Hence, if $\overline{h}_r$ is disjoint from all $\overline{h}_t \in \overline{A}_i$, then $\sum S_h(\overline{h}_r, \overline{h}_t) = |A_i| \times 0 = 0$, meaning that $S_h(\overline{h}_r, \overline{A}_i) = \frac{1}{|A_i|} \times 0 = 0$. Again, $S_h(\overline{h}_r, \overline{A}_i) = 0$ when $\overline{h}_r$ is disjoint from all $\overline{h}_t \in \overline{A}_i$.

(3) If $\overline{h}_r \subseteq \overline{h}_p \subseteq \overline{h}_q$, then $|\overline{h}_r| \leq |\overline{h}_p| \leq |\overline{h}_q|$. Thus, using Eq. (6.1)

$$\begin{aligned} S_h(\overline{h}_q, \overline{A}_i \backslash \overline{h}_q) &= S_h(\overline{h}_q, \{\overline{h}_r, \overline{h}_p\}) \\ &= \frac{1}{2}\left\{\frac{|\overline{h}_q \cap \overline{h}_r|}{|\overline{h}_q|} + \frac{|\overline{h}_q \cap \overline{h}_p|}{|\overline{h}_q|}\right\} \\ &= \frac{1}{2}\left\{\frac{|\overline{h}_r|}{|\overline{h}_q|} + \frac{|\overline{h}_p|}{|\overline{h}_q|}\right\}, \text{as } \overline{h}_r \subseteq \overline{h}_q \text{ and } \overline{h}_p \subseteq \overline{h}_q, \end{aligned}$$

$$
\begin{aligned}
S_h(\overline{h}_p, \overline{A}_i \backslash \overline{h}_p) &= S_h(\overline{h}_p, \{\overline{h}_r, \overline{h}_q\}) \\
&= \frac{1}{2}\left\{\frac{|\overline{h}_p \cap \overline{h}_r|}{|\overline{h}_p|} + \frac{|\overline{h}_p \cap \overline{h}_q|}{|\overline{h}_p|}\right\} \\
&= \frac{1}{2}\left\{\frac{|\overline{h}_r|}{|\overline{h}_p|} + \frac{|\overline{h}_p|}{|\overline{h}_p|}\right\}, \text{as } \overline{h}_r \subseteq \overline{h}_p \text{ and } \overline{h}_p \subseteq \overline{h}_q \\
&= \frac{1}{2}\left\{\frac{|\overline{h}_r|}{|\overline{h}_p|} + 1\right\}.
\end{aligned}
$$

As $|\overline{h}_p| \leq |\overline{h}_q|$, it implies that $\frac{|\overline{h}_p|}{|\overline{h}_q|} \leq 1$. Thus, $S_h(\overline{h}_q, \overline{A}_i \backslash \overline{h}_q) \leq S_h(\overline{h}_p, \overline{A}_i \backslash \overline{h}_p)$.

Again, using Eq. (6.1)

$$
\begin{aligned}
S_h(\overline{h}_r, \overline{A}_i \backslash \overline{h}_r) &= S_h(\overline{h}_r, \{\overline{h}_p, \overline{h}_q\}) \\
&= \frac{1}{2}\left\{\frac{|\overline{h}_r \cap \overline{h}_p|}{|\overline{h}_r|} + \frac{|\overline{h}_r \cap \overline{h}_q|}{|\overline{h}_r|}\right\} \\
&= \frac{1}{2}\left\{\frac{|\overline{h}_r|}{|\overline{h}_r|} + \frac{|\overline{h}_r|}{|\overline{h}_r|}\right\}, \text{as } \overline{h}_r \subseteq \overline{h}_p \text{ and } \overline{h}_r \subseteq \overline{h}_q \\
&= \frac{1}{2}\{1 + 1\}.
\end{aligned}
$$

As $|\overline{h}_r| \leq |\overline{h}_p|$ and $|\overline{h}_r| \leq |\overline{h}_q|$, it implies that $\frac{|\overline{h}_r|}{|\overline{h}_q|} \leq 1$ and $\frac{|\overline{h}_p|}{|\overline{h}_q|} \leq 1$. Hence, $S_h(\overline{h}_q, \overline{A}_i \backslash \overline{h}_q) \leq S_h(\overline{h}_r, \overline{A}_i \backslash \overline{h}_r)$. ■

The next subsection defines the bidirectional subsethood based similarity measure using the definition of subsethood for a set of intervals (subsection 6.3.1) as originally this new similarity measure was developed on the concept of subsethood for only a pair of intervals (Eq. (3.2) in Chapter 3) .

6.3.2 Defining Bidirectional Subsethood Based Similarity Measure for a Set of Intervals

Chapter 3 introduced the new similarity measure based on the concept of bidirectional subsethood for a pair of intervals to assess their similarity. In this subsection, based on

the definition of subsethood as discussed in the previous subsection (subsection 6.3.1), this new similarity measure is revisited to enable it to compute similarity for a set of intervals.

Consider again a set of intervals $\overline{h} = \{\overline{h}_1, \overline{h}_2, ..., \overline{h}_n\}$ arising from n sources. The bidirectional subsethood based similarity measure, S_{S_h} for $\overline{h} = \{\overline{h}_1, \overline{h}_2, ..., \overline{h}_n\}$ is the *t-norm* ($\bigstar$) of their reciprocal subsethoods, i.e.,

$$\begin{aligned} S_{S_h}\left(\overline{h}\right) = S_{S_h}\left(\overline{h}_1, \overline{h}_2, ..., \overline{h}_n\right) &= \bigstar\left(S_h(\overline{h}_1, \{\overline{h}_2, ..., \overline{h}_n\}), ..., S_h(\overline{h}_n, \{\overline{h}_1, ..., \overline{h}_{n-1}\})\right) \\ &= \bigstar\left(S_h(\overline{h}_1, \overline{h}\backslash\overline{h}_1), ..., S_h(\overline{h}_n, \overline{h}\backslash\overline{h}_n)\right) \end{aligned} \tag{6.2}$$

where $\overline{h}\backslash\overline{h}_i$ is the nonempty subset of intervals without having $\overline{h}_i$, $i \in \{1, ..., n\}$.

Remark 6.1:
As the $S_{S_h}\left(\overline{h}\right)$ measure at Eq. (6.2) carries the original definition of the S_{S_h} measure (Eq. (3.2) in Chapter 3), it also meets up all the axioms [F1] - [F9] as defined for the S_{S_h} measure in Chapter 3.

The next subsection proposes a new FM using the bidirectional subsethood based similarity measure defined in the previous subsection 6.3.2 to estimate the inter-source agreement for a set of intervals.

6.3.3 Proposed Agreement-Based Fuzzy Measure

This subsection develops a new data-driven FM involving the bidirectional subsethood based similarity measure (Eq. (6.2)) to compute the agreement among a set of sources providing intervals. As mentioned, the reason behind developing this new FM is to avoid the limitation of the existing agreement-based FMs (i.e., generating same agreement and worth for different sets of sources) when they use popular similarity measures (subject to *aliasing*) and the intersection operation.

Consider again a set of n intervals, $\overline{h} = \{\overline{h}_1, \overline{h}_2, ..., \overline{h}_n\}$. For any nonempty subset $\overline{A}_i \in \overline{h}$, $1 \leq i \leq n$, the new FM, $\tilde{g}^{AS_h}$ using the new similarity measure (Eq. (6.2)) is defined as follows (which is later normalized to a proper FM, g^{AS_h}):

$$\tilde{g}^{AS_h}(\overline{A}_0) = 0, \tag{6.3a}$$

$$\tilde{g}^{AS_h}(\overline{A}_1) = \binom{n}{1}^{-1} \times \sum_{k_1=1}^{1} S_{S_h}\left(\overline{h}_{k_1}, \overline{h}_{k_1}\right) = \frac{1}{n}, \tag{6.3b}$$

$$\begin{aligned}\tilde{g}^{AS_h}(\overline{A}_i) = i \times \tilde{g}^{AS_h}(\overline{A}_1) + \binom{n}{2}^{-1} \sum_{k_1=1}^{i-1} \sum_{k_2=k_1+1}^{i} S_{S_h}\left(\overline{h}_{k_1}, \overline{h}_{k_2}\right) \\ + ... + \binom{n}{i}^{-1} S_{S_h}\left(\overline{h}_1, ..., \overline{h}_i\right),\end{aligned} \tag{6.3c}$$

where $\overline{A}_0 = \emptyset$, $\overline{A}_1$ is a singleton subset, and $\overline{A}_i$ is a non-singleton subset with i sources, $1 < i \leq n$. $\binom{n}{K}$ is total number of K-tuples in the set, $\overline{h}$, where $1 \leq K \leq n$. Equation (6.3a) represents the worth of an empty subset, $\overline{A}_0$, which is always 0. Equation (6.3b) shows the worth of a singleton subset, $\overline{A}_1$, which is its *self* similarity, weighted by a value $\binom{n}{1}^{-1}$. Equation (6.3c) shows the worth of a non-singleton subset, $\overline{A}_i$, which is the sum of the similarities of all K-tuples in $\overline{A}_i$, $1 \leq K \leq i$, weighted by an overall value of $\binom{n}{K}^{-1}$. Equation (6.3c) can also be written as,

$$\begin{aligned}\tilde{g}^{AS_h}(\overline{A}_i) = \frac{i}{n} + \binom{n}{2}^{-1} \sum_{k_1=1}^{i-1} \sum_{k_2=k_1+1}^{i} \bigstar\left(S_h(\overline{h}_{k_1}, \{\overline{h}_{k_2}\}), S_h(\overline{h}_{k_2}, \{\overline{h}_{k_1}\})\right) \\ + ... + \binom{n}{i}^{-1} \bigstar\left(S_h(\overline{h}_1, \overline{A}_i \backslash \overline{h}_1), ..., S_h(\overline{h}_i, \overline{A}_i \backslash \overline{h}_i)\right).\end{aligned} \tag{6.4}$$

Remark 6.2:

Equation (6.3c) shows how the new agreement-based FM ($\tilde{g}^{AS_h}$) using the new similarity measure (Eq. (6.2)) computes the worth for a non-singleton subset of sources. For a non-singleton subset ($|\overline{A}_i| > 1$), all possible K-tuples ($1 \leq K \leq i$) within $\overline{A}_i$ are

determined first (e.g., 1-tuples (with one source), 2-tuples (with two sources)). Then, for all K-tuples, similarity is calculated by the new similarity measure (Eq. (6.2)). Finally, the worth for the non-singleton subset $(\overline{A}_i)$ is determined by summing up the similarities of all K-tuples in $\overline{A}_i$, $1 \leq K \leq i$, weighted by an overall value of $\binom{n}{K}^{-1}$.

Remark 6.3:

Equation (6.3b) presents that the worth of a singleton subset $(\overline{A}_1)$ is, $\tilde{g}^{AS_h}(\overline{A}_1) = \frac{1}{n}$, where $n = |\overline{h}|$. For the disagreeing sources, the inclusion of the worth of singleton subsets aids to generate the new FM and carry out the aggregation when it is combined with a FI.

Following the similar approach in [23][24], Eq. (6.3c) is rewritten as follows,

$$\tilde{g}^{AS_h}(\overline{A}_i) = \frac{i}{n} + \sum_{K=2}^{i} \left[\binom{n}{K}^{-1} Z_K(\overline{A}_i) \right], \; i \geq 1, \tag{6.5}$$

where the first part of Eq. (6.5) is the sum of the similarities of all 1-tuples in $\overline{A}_i$, weighted by a value, $\binom{n}{1}^{-1}$. The other part gives summation of the similarities of all K-tuples in $\overline{A}_i$ $(K \geq 2)$, weighted by a value, $\binom{n}{K}^{-1}$. $Z_K(\overline{A}_i)$ captures the cumulative similarity for all K-tuples in A_i $(K \geq 2)$ using Eq. (6.2) and is defined in Eq. (6.6):

$$Z_K\left(\overline{A}_i\right) = \sum_{k_1=1}^{i-K+1} \sum_{\substack{k_2= \\ k_1+1}}^{i-K+2} \ldots \sum_{\substack{k_K= \\ k_{K-1}+1}}^{i} \bigstar \left(S_h(\overline{h}_{k_1}, \overline{A}_i \backslash \overline{h}_{k_1}), S_h(\overline{h}_{k_2}, \overline{A}_i \backslash \overline{h}_{k_2}), \right.$$
$$\left. \ldots, S_h(\overline{h}_{k_K}, \overline{A}_i \backslash \overline{h}_{k_K}) \right). \tag{6.6}$$

Finally, $\tilde{g}^{AS_h}(\overline{A}_i)$ is normalized by $\tilde{g}^{AS_h}(\overline{h})$ in Eq. (6.7) so that $g^{AS_h}(\overline{A}_i) \leq 1$ and $g^{AS_h}(\overline{h}) = 1$, which maintains the properties of an FM (discussed in Chapter 2).

$$g^{AS_h}(\overline{A}_i) = \frac{\tilde{g}^{AS_h}(\overline{A}_i)}{\tilde{g}^{AS_h}(\overline{h})}, \; 1 \leq i \leq n. \tag{6.7}$$

Table 6.1 The worth of subsets of intervals using the $\tilde{g}^{AS_h}$ FM prior to normalization.

$\lvert\{\overline{A}_i\}\rvert$=1	$\lvert\{\overline{A}_i\}\rvert$=2	$\lvert\{\overline{A}_i\}\rvert = 3$
$\tilde{g}^{AS_h}(\{\overline{h}_1\})$=0.333	$\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\})$=0.723	$\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\})$=1.352
$\tilde{g}^{AS_h}(\{\overline{h}_2\})$=0.333	$\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_3\})$=0.723	
$\tilde{g}^{AS_h}(\{\overline{h}_3\})$=0.333	$\tilde{g}^{AS_h}(\{\overline{h}_2, \overline{h}_3\})$=0.734	

The next example shows how the new FM calculates the worth of all subsets of sources providing intervals.

Example 6.2:

Consider the set of intervals $\overline{h} = \{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$ (see Fig. 6.1(a), Scenario 1). A manual calculation for $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\})$ is described below, while the complete $\tilde{g}^{AS_h}$ FM is shown in Table 6.1. From Eq. (6.5), $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\}) = \frac{2}{3} + \binom{3}{2}^{-1} Z_2(\{\overline{h}_1, \overline{h}_2\})$. Using Eq. (6.6),

$$\begin{aligned} Z_2(\{\overline{h}_1, \overline{h}_2\}) &= \bigstar\left(S_h(\overline{h}_1, \{\overline{h}_2\}), S_h(\overline{h}_2, \{\overline{h}_1\})\right) \\ &= \bigstar(\frac{1.0}{6.0}, \frac{1.0}{1.0}) = \bigstar(0.167, 1.0) = 0.167. \end{aligned}$$

Therefore,

$$\begin{aligned} \tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\}) &= \frac{2}{3} + \binom{3}{2}^{-1} \times 0.167 \\ &= \frac{2}{3} + \frac{1}{3} \times 0.167 = 0.723. \end{aligned}$$

The remaining results for the complete $\tilde{g}^{AS_h}$ FM (non-normalized FM) are given in Table 6.1. The $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\})$ is further normalized by using the maximum value of $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\})$, which is 1.352 (from Table 6.1). That is, $g^{AS_h}(\{\overline{h}_1, \overline{h}_2\}) = \frac{0.723}{1.352} = 0.535$. The final normalized g^{AS_h} FM is shown in Table 6.2.

Table 6.2 The worth of subsets of intervals using the g^{AS_h} FM after normalization.

$\lvert\{\overline{A}_i\}\rvert$=1	$\lvert\{\overline{A}_i\}\rvert$=2	$\lvert\{\overline{A}_i\}\rvert = 3$
$g^{AS_h}(\{\overline{h}_1\})$=0.246	$g^{AS_h}(\{\overline{h}_1, \overline{h}_2\})$=0.535	$g^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\})$=1.0
$g^{AS_h}(\{\overline{h}_2\})$=0.246	$g^{AS_h}(\{\overline{h}_1, \overline{h}_3\})$=0.535	
$g^{AS_h}(\{\overline{h}_3\})$=0.246	$g^{AS_h}(\{\overline{h}_2, \overline{h}_3\})$=0.543	

Generally, a FM is monotonic and non-decreasing (subsection 2.7.1 in Chapter 2). The following theorem shows that the new agreement-based FM also satisfies the monotonicity and non-decreasing properties.

Theorem 6.2:

The new agreement-based FM using the bidirectional subsethood based similarity measure at Eq. (6.3) is monotonic and non-decreasing.

At first, we prove the following corollary.

Corollary 6.1: $Z_K\left(\overline{A}_{i-1}\right) \leq Z_K\left(\overline{A}_i\right)$ for $2 \leq K \leq i$, where $i = \left|\overline{A}_i\right|$, $1 \leq i \leq n$.

Proof: Let us consider $K = 2$. From Eq. (6.6),

$$
\begin{aligned}
Z_2\left(\overline{A}_i\right) &= \sum_{k_1=1}^{i-1} \sum_{k_2=2}^{i} \bigstar \left(S_h(\overline{h}_{k_1}, \{\overline{h}_{k_2}\}), S_h(\overline{h}_{k_2}, \{\overline{h}_{k_1}\})\right) \\
&= \sum_{k_1=1}^{i-2} \sum_{k_2=2}^{i-1} \bigstar \left(S_h(\overline{h}_{k_1}, \{\overline{h}_{k_2}\}), S_h(\overline{h}_{k_2}, \{\overline{h}_{k_1}\})\right) + \\
&\quad \sum_{k_1=1}^{i-1} \bigstar \left(S_h(\overline{h}_{k_1}, \{\overline{h}_i\}), S_h(\overline{h}_i, \{\overline{h}_{k_1}\})\right) \\
&= Z_2\left(\overline{A}_{i-1}\right) + \sum_{k_1=1}^{i-1} \bigstar \left(S_h(\overline{h}_{k_1}, \{\overline{h}_i\}), S_h(\overline{h}_i, \{\overline{h}_{k_1}\})\right)
\end{aligned}
$$

It follows that $Z_K\left(\overline{A}_{i-1}\right) \leq Z_K\left(\overline{A}_i\right)$ for $K = 2$. Eq. (6.6) can be rewritten as Eq. (6.8).

$$Z_K\left(\overline{A}_i\right) = Z_K\left(\overline{A}_{i-1}\right) + \sum_{k_1=1}^{(i-1)-(K-1)+1} \sum_{\substack{k_2= \\ k_1+1}}^{(i-1)-(K-1)+2} ... \sum_{\substack{k_{K-1}= \\ k_{K-2}+1}}^{i-1} \bigstar\Big(S_h(\overline{h}_{k_1}, \{\overline{h}_{k_2}, ..., \overline{h}_{k_{K-1}}, \overline{h}_i\}), ..., S_h(\overline{h}_i, \{\overline{h}_{k_1}, ..., \overline{h}_{k_{K-1}}\})\Big) \quad (6.8)$$

By the distributive property, Eq. (6.8) can be reduced to Eq. (6.9) for $K = 2$ and $1 \leq i \leq n$.

$$Z_K\left(\overline{A}_i\right) = Z_K\left(\overline{A}_{i-1}\right) + \bigstar\left(Z_{K-1}\left(\overline{A}_{i-1}\right), S_h\left(\overline{h}_i, \overline{A}_i \backslash \overline{h}_i\right)\right). \quad (6.9)$$

The last part of Eq. (6.9) is always ≥ 0, which implies that $Z_K\left(\overline{A}_{i-1}\right) \leq Z_K\left(\overline{A}_i\right)$ for $2 \leq K \leq i$. Hence, the proof of the corollary is complete. ■

The proof of Theorem 2 is given below:

Proof: From Corollary 6.1, we see that for any set $\overline{A}_{i-1} \subset \overline{A_i}$,

$$\tilde{g}^{AS_h}\left(\overline{A}_{i-1}\right) = \frac{(i-1)}{n} + \sum_{K=2}^{i-1} \binom{n}{K}^{-1} Z_K\left(\overline{A}_{i-1}\right) < \frac{i}{n} + \sum_{K=2}^{i} \binom{n}{K}^{-1} Z_K\left(\overline{A}_i\right) = \tilde{g}^{AS_h}\left(\overline{A}_i\right)$$

Thus, $\tilde{g}^{AS_h}\left(\overline{A}_{i-1}\right) < \tilde{g}^{AS_h}\left(\overline{A}_i\right)$, proving the theorem. ■

The following algorithm presents the computation of the g^{AS_h} FM, $g^{AS_h}\left(\overline{A}_i\right)$, $1 \leq i \leq n$.

Algorithm 2 Computation of the lattice of g^{AS_h} FM.

Input: A set of intervals $\overline{h} = \{\overline{h}_1, \overline{h}_2, ..., \overline{h}_n\}$
Output: The lattice of g^{AS_h} FM

Set $\tilde{g}^{AS_h}(\overline{A}_0) = 0$
for each $\overline{A}_i$, $i \rightarrow 1$ to n **do**
    $Self_AG = i/n$
    for each integer $K \rightarrow 2$ to i **do**
        Compute and store $Z_K(\overline{A}_i)$ using (6.6)
    end for
    Compute $\tilde{g}^{AS_h}(\overline{A}_i)$ using (6.5)
end for
for each $\overline{A}_i$, $i \rightarrow 1$ to n **do**
    Compute $g^{AS_h}(\overline{A}_i)$ using (6.7)
end for

The following examples are used to show that the new FM g^{AS_h} based on the new similarity measure at Eq. (6.2) avoids the limitation of existing agreement-based FMs, g^{AG}, g^{GenA}, and g^{AA} (i.e., generating same agreement and worth for different sets of sources).

Example 6.3:

Consider again the two cases shown as Scenarios 1 and 2 in Fig. 6.1 where each of them has three intervals, $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$. Example 6.2 has already shown that the $\tilde{g}^{AS_h}$ FM assigns 1.352 as worth to the set $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$ of Scenario 1 (see Table 6.1). In other words, $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}) = 1.352$ for the Scenario 1. In a similar manner, $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}) = 2.17$ for the Scenario 2. Thus, the worth generated by the $\tilde{g}^{AS_h}$ FM for the set $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$ in the Scenario 1 and 2 are different, unlike the $\tilde{g}^{AG}$ FM with the intersection operation, which provides the same worth for them (discussed in section 6.2).

Example 6.4:

Consider two different sets of interval-valued data, $\overline{h}$ and $\overline{r}$ as shown in Fig. 6.2 where each set has two intervals. The steps that are followed in calculating $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\})$

and $\tilde{g}^{AS_h}(\{\overline{r}_1, \overline{r}_2\})$, are given below. In both cases, all K-tuples ($1 \leq K \leq 2$) within the set are determined first.

Calculation of $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\})$:

The weighted summation of similarities for all 1-tuples within $\{\overline{h}_1, \overline{h}_2\}$ is $\frac{i}{n} = \frac{2}{2} = 1$, which is the first part of Eq. (6.5). The weighted similarities for the 2-tuple in $\{\overline{h}_1, \overline{h}_2\}$ is 0.40, representing the last part of Eq. (6.5). In this case, the new measure at Eq. (6.2) with the minimum *t-norm* is used to compute the similarity. Thus, $\tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\}) = 1 + 0.40 = 1.40$.

Calculation of $\tilde{g}^{AS_h}(\{\overline{r}_1, \overline{r}_2\})$:

The weighted summation of similarities for all 1-tuples within $\{\overline{r}_1, \overline{r}_2\}$ is $\frac{i}{n} = \frac{2}{2} = 1$, and the weighted similarities for the 2-tuple in $\{\overline{r}_1, \overline{r}_2\}$ is 0.50 (using Eq. (6.2) with the minimum *t-norm*). Thus, $\tilde{g}^{AS_h}(\{\overline{r}_1, \overline{r}_2\}) = 1 + 0.50 = 1.50$.

In this example, $\tilde{g}^{AS_h}(\{\overline{r}_1, \overline{r}_2\}) > \tilde{g}^{AS_h}(\{\overline{h}_1, \overline{h}_2\})$, which implies that the new FM provides different agreement and worth for the sets, $\overline{h}$ and $\overline{r}$, whereas the g^{GenA} FM with the Jaccard or Dice similarity measure gives same agreement and worth for them (discussed in section 6.2).

In the next section, the behaviour of the new agreement-based FM is demonstrated and compared with respect to the three existing agreement-based FMs using two sets of synthetic datasets and a real world example.

6.4 Demonstration and Analysis

This section demonstrates and analyses the behaviour of the new agreement-based FM (g^{AS_h}) in the context of the AG, $GenA$, and AA FMs for two sets of synthetic examples and later with a real dataset. For convenience, the new FM is denoted as AS_h and the CFI [129] is used as the FI throughout. In all experiments, the minimum *t-norm* is

used for the AS_h FM (as it is the most common in practice) and the Jaccard similarity measure is applied for the $GenA$ and AA FMs. All experiments are implemented using Java on an Intel(R) Core(TM) i3-4005U series based machine running at 1.70 GHz with 8GB RAM.

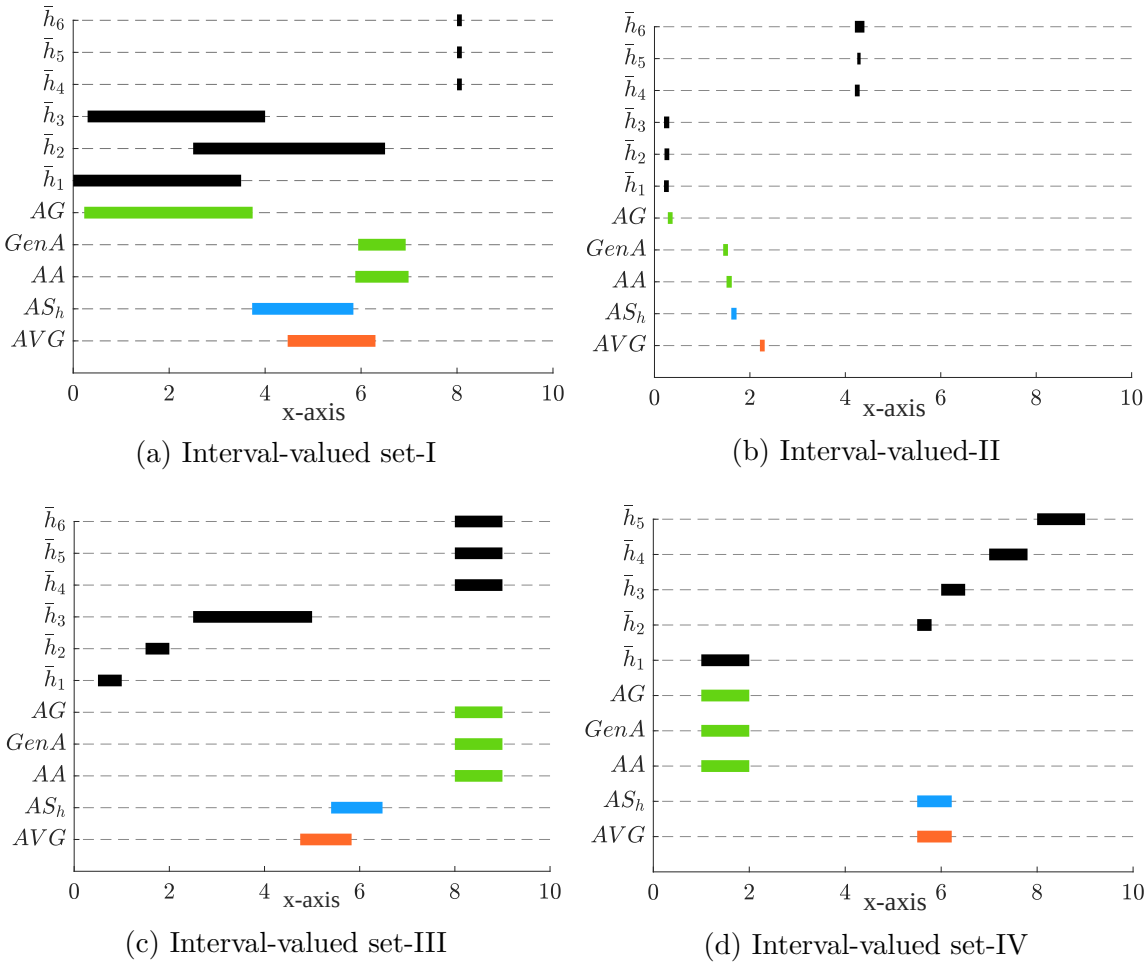
6.4.1 Demonstration with Synthetic Dataset-1

This section presents a series of numerical examples using interval-valued synthetic datasets to analyse and illustrate the behaviour of the AS_h FM with respect to the three other comparable FMs. Fig. 6.3 shows four examples of synthetic datasets together with aggregated results based on the CFI [129] using these FMs.

(1) Interval-valued set-I shown in Fig. 6.3(a) consists of three smaller intervals $\overline{h}_4$, $\overline{h}_5$, and $\overline{h}_6$ that agree completely and three larger intervals $\overline{h}_1$, $\overline{h}_2$ and $\overline{h}_3$ agreeing to a certain degree. The aggregation results from the FMs in combination with CFI [129] (Fig. 6.3(a)) show that the AG FM gives importance only to the subset of larger intervals, whereas the $GenA$ and AA FMs are strongly influenced by the subset of smaller intervals as they agree totally. However, the AS_h FM not only gives more importance to the subset of smaller intervals having a complete agreement, but also considers other subsets, $\{\overline{h}_1, \overline{h}_3\}$ and $\{\overline{h}_2, \overline{h}_3\}$ with agreement to a certain degree.

(2) For interval-valued set-II shown in Fig. 6.3(b), there are three intervals $\overline{h}_1$, $\overline{h}_2$ and $\overline{h}_3$ having better agreement than three other intervals $\overline{h}_4$, $\overline{h}_5$ and $\overline{h}_6$. In this case, the AG FM is greatly influenced by the subset $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$, whereas the $GenA$, AA, and AS_h FMs appear to show better aggregation results considering two subsets of sources ($\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$ and $\{\overline{h}_4, \overline{h}_5, \overline{h}_6\}$) when used in combination with the CFI [129].

(3) Interval-valued set-III shown in Fig. 6.3(c) again has six intervals where half of the intervals agree to each other completely and the rest of them disagree totally.



(a) Interval-valued set-I

(b) Interval-valued-II

(c) Interval-valued set-III

(d) Interval-valued set-IV

Fig. 6.3 Comparison of aggregation results from the CFI [129] with the AS_h FM and the three other existing FMs (AG, $GenA$ and AA FMs) for four different synthetic sets of intervals.

Here, the AG, $GenA$ and AA FMs are completely influenced by the subset of agreed intervals, i.e., $\{\overline{h}_4, \overline{h}_5, \overline{h}_6\}$. Like other FMs, the AS_h FM shows the influence of the subset $\{\overline{h}_4, \overline{h}_5, \overline{h}_6\}$, but, at the same time, it also shows to consider the disagreed singletons, $\{\overline{h}_1, \overline{h}_2, \overline{h}_3\}$.

(4) Interval-valued set-IV shown in Fig. 6.3(d) consists of five intervals where all intervals are in complete disagreement. In this case, the AG, $GenA$, and AA FMs give $\overline{h}_1$ as the aggregation result when used with the CFI [129]. As there is no overlap

among the intervals, for all three FMs, the worth of all subsets of intervals are equal to 0, except the subset $\overline{h} = \{\overline{h}_1, \overline{h}_2, \overline{h}_3, \overline{h}_4, \overline{h}_5\}$, for which $g^{AG}(\overline{h}) = g^{GenA}(\overline{h}) = g^{AA}(\overline{h}) = 1$ to satisfy the property of a FM (subsection 2.7.1 in Chapter 2). This causes existing FMs to produce the aggregation result, $\overline{h}_1$, when they are combined with CFI [129]. Contrarily, the AS_h FM, by its construction, assigns worth to all singletons, which is later normalized by $\tilde{g}^{AS_h}(\overline{h})$. Even though there is no agreement among the intervals, the AS_h FM still can estimate the worth of other subsets by utilizing the worth of singletons. Table 6.3 shows the normalized worth of all subsets of intervals for the dataset-IV (in Fig. 6.3(d)) using the AS_h FM, where all intervals are in complete disagreement.

Intuitively, when there is no overlap between intervals and all intervals are unique, then all sources should be treated with an equal worth and the aggregation should be equal to the mean of their endpoints. In Fig. 6.3(d), only the AS_h FM in combination

Table 6.3 The normalized worth of subsets of intervals using the AS_h FM (g^{AS_h})

$\lvert\{\overline{A}_i\}\rvert = 1$	$\lvert\{\overline{A}_i\}\rvert = 2$	$\lvert\{\overline{A}_i\}\rvert = 3$
$g^{AS_h}(\{\overline{h}_1\})$=0.2	$g^{AS_h}(\{\overline{h}_1, \overline{h}_2\})$=0.4	$g^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3\})$=0.6
$g^{AS_h}(\{\overline{h}_2\})$=0.2	$g^{AS_h}(\{\overline{h}_1, \overline{h}_3\})$=0.4	$g^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_4\})$=0.6
$g^{AS_h}(\{\overline{h}_3\})$=0.2	$g^{AS_h}(\{\overline{h}_1, \overline{h}_4\})$=0.4	$g^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_5\})$=0.6
$g^{AS_h}(\{\overline{h}_4\})$=0.2	$g^{AS_h}(\{\overline{h}_1, \overline{h}_5\})$=0.4	$g^{AS_h}(\{\overline{h}_1, \overline{h}_3, \overline{h}_4\})$=0.6
$g^{AS_h}(\{\overline{h}_5\})$=0.2	$g^{AS_h}(\{\overline{h}_2, \overline{h}_3\})$=0.4	$g^{AS_h}(\{\overline{h}_1, \overline{h}_3, \overline{h}_5\})$=0.6
	$g^{AS_h}(\{\overline{h}_2, \overline{h}_4\})$=0.4	$g^{AS_h}(\{\overline{h}_1, \overline{h}_4, \overline{h}_5\})$=0.6
	$g^{AS_h}(\{\overline{h}_2, \overline{h}_5\})$=0.4	$g^{AS_h}(\{\overline{h}_2, \overline{h}_3, \overline{h}_4\})$=0.6
	$g^{AS_h}(\{\overline{h}_3, \overline{h}_4\})$=0.4	$g^{AS_h}(\{\overline{h}_2, \overline{h}_3, \overline{h}_5\})$=0.6
	$g^{AS_h}(\{\overline{h}_3, \overline{h}_5\})$=0.4	$g^{AS_h}(\{\overline{h}_2, \overline{h}_4, \overline{h}_5\})$=0.6
	$g^{AS_h}(\{\overline{h}_4, \overline{h}_5\})$=0.4	$g^{AS_h}(\{\overline{h}_3, \overline{h}_4, \overline{h}_5\})$=0.6
$\lvert\{\overline{A}_i\}\rvert = 4$	$\lvert\{\overline{A}_i\}\rvert = 5$	
$g^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3, \overline{h}_4\})$=0.8	$g^{AS_h}(\{\overline{h}_2, \overline{h}_3, \overline{h}_4, \overline{h}_5\})$=1.0	
$g^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_3, \overline{h}_5\})$=0.8		
$g^{AS_h}(\{\overline{h}_1, \overline{h}_2, \overline{h}_4, \overline{h}_5\})$=0.8		
$g^{AS_h}(\{\overline{h}_1, \overline{h}_3, \overline{h}_4, \overline{h}_5\})$=0.8		
$g^{AS_h}(\{\overline{h}_2, \overline{h}_3, \overline{h}_4, \overline{h}_5\})$=0.8		

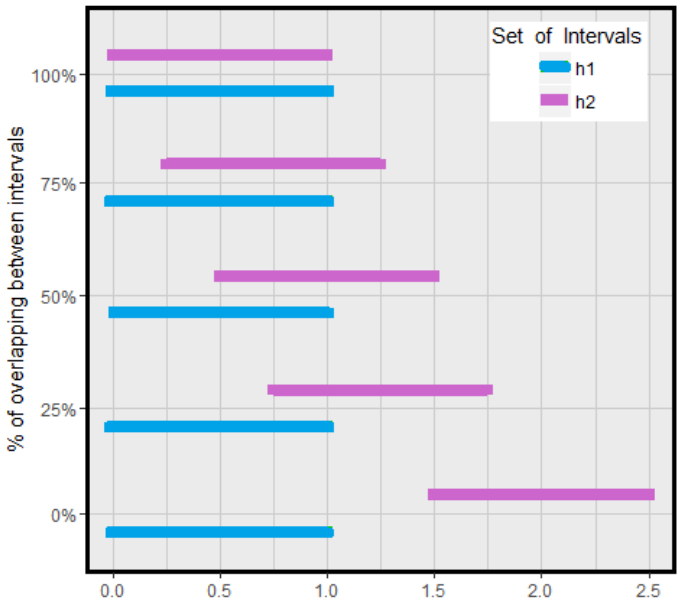
with the CFI [129] generates the aggregation results accordingly (i.e., performs like an AVG operator), whereas other FMs produce the aggregation results identical to the input interval with the lowest endpoints (i.e., perform like a MIN operator).

To sum up, in respect to various degrees of overlap among a set of intervals, the AS_h FM appears to generate more intuitive results than other existing FMs, AG, $GenA$, and AA when used with the CFI [129]. Particularly, the AG FM [129] gives more importance to intervals with larger overlap, whereas the $GenA$, and AA FMs are to the intervals with larger similarity. Besides, all three FMs pay attention to only the agreed intervals, ignoring the intervals with complete disagreement. In both cases, the AS_h FM gives importance to the intervals with larger similarity, but simultaneously considers other intervals. Finally, when the sources are in complete disagreement, the AS_h FM performs like an average (AVG) operator, whereas the AG, $GenA$, and AA FMs behave like a minimum (MIN) operator when used with the CFI [129].

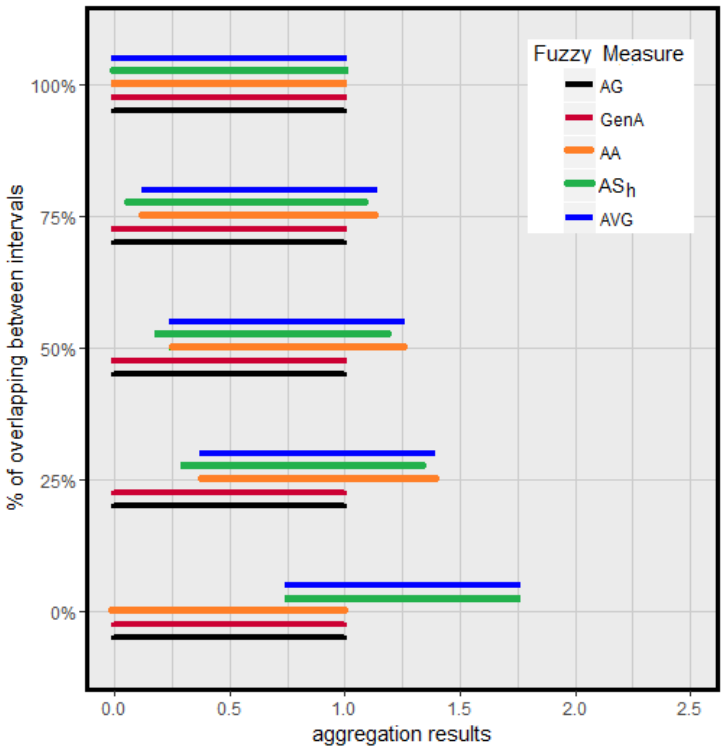
6.4.2 Demonstration with Synthetic Dataset-2

In this subsection, using another set of synthetic dataset, the performance of the new FM is demonstrated and evaluated in respect to the existing agreement-based FMs, AG, $GenA$, and AA. In particular, here we investigate how the FMs in combination with the CFI [129] behave in producing the aggregation result when the overlap between intervals are gradually decreased. Here, five different sets of two interval-valued data $\overline{h}_1$ and $\overline{h}_2$ are considered in Fig. 6.4(a) with 100%, 75%, 50%, 25%, and 0% interval-overlap respectively. Note that $\overline{h}_1$ is set to $[0, 1]$ in all five sets, while $\overline{h}_2$ is altered depending on the % of overlap.

Fig. 6.4(b) shows that all the FMs (used with the CFI [129]) aggregates the intervals equally (i.e., $[0, 1]$) when 100% interval-overlap exists. However, despite degrading overlap, the AG and $GenA$ FMs continue to show the same aggregation output (i.e.,



(a) Sets of interval-valued data



(b) Aggregation of sets of interval-valued data

Fig. 6.4 (a) Five sets of interval-valued data with degrading interval-overlap (b) The aggregation results of the AG, $GenA$, AA and AS_h FMs in combination with the CFI [129] for Fig. 6.4(a).

$[0, 1]$), whereas the AA and AS_h FMs follow the overlap degradation and aggregate the intervals accordingly. Finally, when intervals are in complete disagreement (i.e., 0% overlap), the AS_h FM in combination with the CFI [129] performs like an average operator, whereas the aggregation outcome with the AA FM suddenly drops and goes back to the initial aggregation output $[0, 1]$. This investigation reveals the shortcoming of the existing FMs in data aggregation when they are used with the CFI.

6.4.3 A Real-World Example

In this part, a real-world dataset is used to review the behaviour of the new and existing agreement-based FMs (AS_h, AG, $GenA$, and AA). For this demonstration, a cybersecurity dataset is used which is involved in the security risk assessment on a proposed computing system [143, 144]. The dataset was collected through a decision making exercise at GCHQ, Cheltenham, UK, the UK government's signals intelligence and information assurance agency where thirty-nine GCHQ selected cybersecurity experts with different levels of expertise took part. The experts were divided into seven different groups (A to G) and provided data anonymously.

The experts were well-informed about the architecture of the system and its components. They were given a list of ten technical attacks (such as, via voice over internet protocol (VoIP) client, email) on the system, and the twenty six different security components (such as, anti-virus software, firewalls) that make up the attacks. The experts were asked to complete two survey exercises where they gave their answers as an interval, on a scale of 0 to 100. In Survey-1, they were asked to rank ten technical attacks according to the difficulty to complete the attack without being noticed. In Survey-2, they were asked to rate each of these twenty-six security components in terms of overall difficulty to attack/evade the component and the characteristics/factors contributed to the difficulty of attacking/evading the component. The security compo-

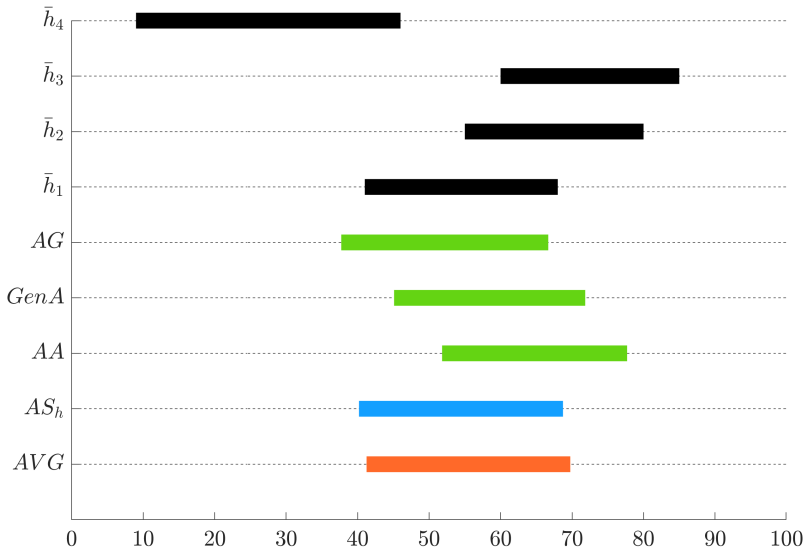
nents were categorized as *attack* (a component to be attacked) and *evade* (a component to be evaded). The set of questions asked for *attack* and *evade* components are given in Appendix B.

To examine the performance of all agreement-based FMs including the new one, we focus on a specific question among a set of questions for the *attack* components. Particularly, we pay attention to Question-8 (see Appendix 2) for the *attack* components where the experts were asked to provide overall ratings on the difficulty to *attack* a component. On this question, for each group, expert's overall ratings on an *attack* component are aggregated using the CFI [129] in respect to all agreement-based FMs. As all aggregation results of experts opinion belonging to different groups tend to show a similar trend, here we only present the aggregation results of the experts ratings of Group A and C for the *attack* component-5. Further, the aggregation results of the experts ratings (belonging to Group D and E) on the *attack* component-11 are also displayed.

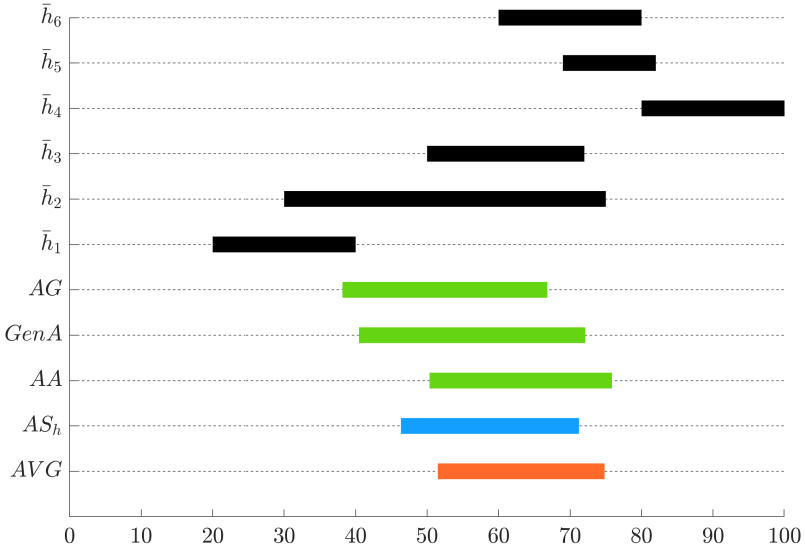
Fig. 6.5(a) presents four intervals $\overline{h}_1$, $\overline{h}_2$, $\overline{h}_3$ and $\overline{h}_4$ capturing experts ratings (Group A) on the attack component-5 as well as their aggregation results by the CFI [129] in respect to the new and existing agreement-based FMs (AS_h, AG, $GenA$, and AA). As it appears, the $GenA$ and AA FMs are more influenced by the subset of intervals with larger similarity, whereas the AG and AS_h FMs give nearly similar aggregation results.

Fig. 6.5(b) presents six intervals $\overline{h}_1$, $\overline{h}_2$, $\overline{h}_3$, $\overline{h}_4$, $\overline{h}_5$, and $\overline{h}_6$ capturing experts ratings (Group C) on the *attack* component-5 and their aggregation results by the CFI [129] in respect to the new and existing agreement-based FMs (AS_h, AG, $GenA$, and AA). In this case, the AG and $GenA$ FMs appear to be more influenced by the subset $\{\overline{h}_1$, $\overline{h}_2$, $\overline{h}_3\}$.

Similarly, Fig. 6.6(a) and (b) present four and five intervals respectively capturing experts ratings (Group D and E) on the *attack* component-11. It also gives aggregation



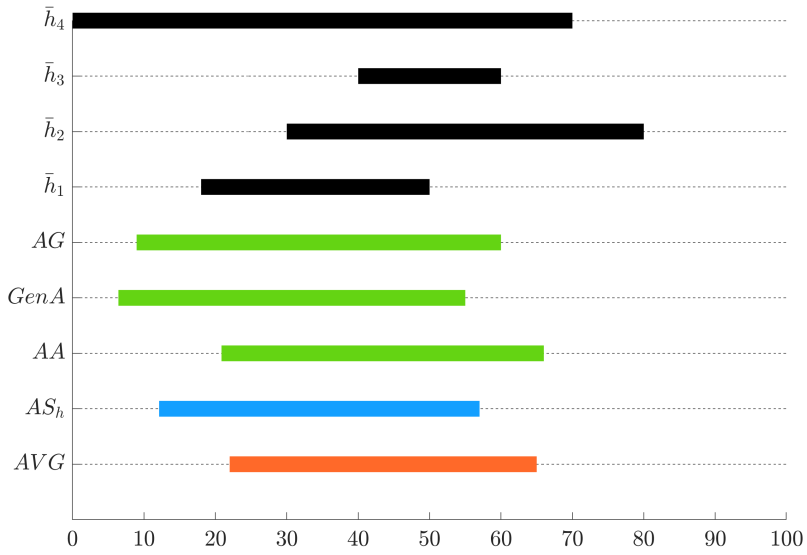
(a) Interval-valued dataset from Group-A



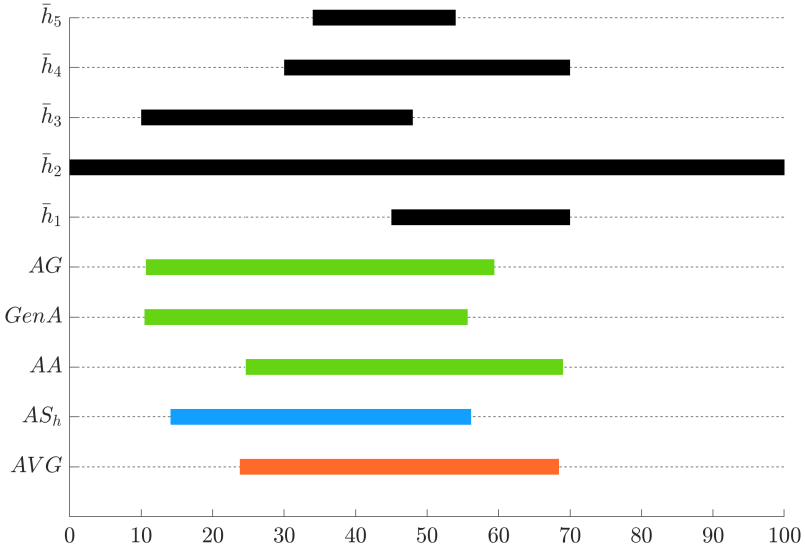
(b) Interval-valued dataset from Group-C

Fig. 6.5 Comparison of aggregation results from the CFI in respect to the AS_h FM and three other existing FMs (AG, $GenA$ and AA FMs) for experts ratings (Group A and C) on the difficulty to *attack* component-5.

results by the CFI [129] in respect to the new and existing agreement-based FMs (AS_h, AG, $GenA$, and AA). In both cases, the AA FM acts like an average operator whereas



(a) Interval-valued dataset from Group-D



(b) Interval-valued dataset from Group-E

Fig. 6.6 Comparison of aggregation results from the CFI in respect to the AS_h FM and three other existing FMs (AG, $GenA$ and AA FMs) for experts ratings (Group D and E) on the difficulty to *attack* component-11.

the other three FMs (AS_h, AG, and $GenA$) show nearly similar aggregation results when used with the CFI [129].

In summary, the real-data analysis shows that in most of the cases, the AG, $GenA$, and AS_h produce nearly similar aggregation results of experts ratings whereas the AA FM behaves like an average operator. However, the demonstration with two different sets of synthetic data has presented some interesting differences in aggregation results among the FMs when used with the CFI [129].

6.5 Conclusion

Focusing on the application of the similarity measures in FI based uncertain data aggregation, this chapter has identified that the existing agreement-based FMs are affected by the limitation of intersection operation and *aliasing* of the common similarity measures when they use them and thus provides the same agreement and worth for very different sets of sources, which eventually affects the aggregation outcome. As the new bidirectional subsethood based similarity measure (Eq. (3.2) in Chapter 2) has shown to avoid *aliasing* and thus provides intuitive results for continuous (closed), discontinuous, and T1 fuzzy sets, this chapter has leveraged the new similarity measure (Eq. (3.2) in Chapter 2) to extract a new agreement-based FM which can avoid the limitation of existing agreement-based FMs. The new FM has also been designed to capture the worth of individual sources when they are in complete disagreement.

The new FM has been compared as regards to its aggregation using the CFI [129] in respect to three other existing agreement-based FMs using both synthetic datasets and a real-world example. The aggregation results have shown that the new FM not only puts more weight to the subsets having high and complete agreement but also pays attention to other intervals with agreement to a certain degree or complete disagreement. It also results in intuitive aggregation even when all sources are in complete disagreement where the aggregation result is the same as the mean of the endpoints of intervals. The exhaustive investigation in this chapter, particularly with

the synthetic datasets have also exhibited the suitability of the new FM in combination with an FI for interval-valued data aggregation. The works that have been done in this chapter satisfy objective 6 and 7 of the thesis.

The contribution to this chapter is now being prepared as an article for submission to IEEE Transactions on Fuzzy Systems.

- **Aggregation of Crowd-Sourced Intervals Using Fuzzy Integral and Bidirectional Subsethood Based Fuzzy Measure**

  Kabir, Shaily; Wagner, Christian; Havens, Timothy C.; and Anderson, Derek T. Aggregation of Crowd-Sourced Intervals Using Fuzzy Integral and Bidirectional Subsethood Based Fuzzy Measure.

  Status: *To be submitted*

The next chapter summarizes and concludes the works of this thesis by highlighting its contributions and limitations. Further, it discusses the scope for future work.

Chapter 7

Conclusions

7.1 Thesis Summary

Many real-world applications often involve uncertain data as inputs, and the outcomes largely rely on their effective comparison. Hence, a wide range of similarity measures have been theoretically established for comparing uncertain data, but, in practice, how accurately they capture similarity is still contentious. In view of growing number of areas leveraging uncertain information and the importance of their effective comparison, the aim of this thesis is to establish methods for precisely assessing similarity between uncertain data, captured by intervals and T1 fuzzy sets.

Having this aim in mind, Chapter 2 has discussed relevant concepts and previous research works to identify the limitations of existing similarity measures and to develop new methods. Particularly, it has given an overview of popular set-theoretic similarity measures [11] and their limitations for intervals and T1 fuzzy sets. Further towards robust similarity measures involving the concept of subsethood, Chapter 2 has presented discussions on subsethood for crisp and fuzzy sets, and provided different axiomatizations for their characterization. Finally, it has reviewed data aggregation

and various aggregation operators including fuzzy integrals to explore the usefulness of applying new similarity measures in data aggregation from multiple uncertain sources.

A variety of set-theoretic similarity measures has come into use to capture similarity among intervals, though each of these measures has its own strengths and weaknesses [11][104][88]. For intervals, the Jaccard and Dice similarity measures are the most frequently used [23] as they follow all common properties of a 'standard' similarity measure [145]. However, both these measures, and indeed most set-theoretic similarity measures are often subject to an inherent limitation, termed as *aliasing*, where they generate the same similarity for very different pairs of intervals with different sizes. To address *aliasing*, Chapter 3 has introduced new axioms for underpinning a *robust* similarity measure for continuous intervals and then proposed a new set-theoretic similarity measure which employs the bidirectional subsethood of intervals to assess similarity. The new measure is sensitive towards the variation in the size of intervals, enabling it to minimize *aliasing*. Importantly, this new measure satisfies both traditional [70] and new axioms to make itself a *robust* similarity measure for the continuous intervals. The behaviour of the proposed measure in respect to existing popular similarity measures has been demonstrated using both synthetic and real interval-valued datasets.

Though continuous intervals are the common form of intervals, discontinuous intervals, consisting of a series of continuous subintervals [45], are used in many applications. Measures to assess similarity for the continuous intervals are not directly applicable for the discontinuous intervals. Chapter 4 has proposed a generalized framework to measure similarity between discontinuous intervals by considering similarities of all possible pairs of their continuous subintervals. Importantly, the new framework assesses similarity between intervals while maintaining their original information (continuous or discontinuous). Further, this chapter has developed a new algorithm to make

the new generalized framework computationally more efficient by combining it with the extended binary relations of Allen's theory [21]. The performance of this new generalized framework in respect to Jaccard, Dice and the bidirectional subsethood based new similarity measure (in Chapter 3) has also been explored and discussed against its integration with the extended Allen's theory.

Beyond intervals, in Chapter 5, this thesis has extended the new similarity measure for intervals vertically and horizontally (α-cut based) with a view to comparing T1 fuzzy sets as their vertical-slices and α-cuts often give rise to continuous intervals [20]. Initially, the new interval similarity measure has been extended vertically; however, this extended measure alike key existing fuzzy set-theoretic similarity measures (also depending on vertical-slice representation) is found to be non-intuitive for comparing non-normal T1 fuzzy sets. Hence, the new similarity measure has been extended using the α-cut based extension. As this extension is not directly applicable for non-normal and non-convex fuzzy sets, it is further refined for comparing T1 fuzzy sets having non-normal and non-convex features. Both of these extensions have been compared against popular fuzzy set-theoretic similarity measures, and a comparative analysis of these measures has also been provided.

In the final stage, Chapter 6 has explored the utility of the new robust similarity measures in FI based data aggregation from uncertain sources where FIs are specified in respect to FMs capturing 'worth' of both the individual data sources and their combinations. As current approaches to generate the FMs rely on existing similarity measures to capture agreement among sources, they are also affected by *aliasing* of those common similarity measures. Therefore, Chapter 6 has taken advantages of the new *robust* similarity measure to estimate inter-source agreement, which is later used to extract a new data-driven FM from interval-valued data. Different experimentation has shown that the new fuzzy measure as compared to the existing agreement-based FMs

[22][23][24] can avoid limitation of inter-source agreement calculation as well as worth estimation and thus provides intuitive aggregation outcome when used in combination with an FI.

7.2 Contributions

A summary of key contributions of this thesis is discussed below.

Identification of an inherent limitation of popular set-theoretic similarity measures. A new definition of robust similarity measures with new and traditional axioms to address this limitation.

Set-theoretic similarity measures are well-known to assess similarity between intervals, though other methods also exist. Among them, Jaccard and Dice measures are the most common. Besides, Szymkiewicz-Simpson coefficient [76], Otsuka coefficient [77], Sokal-Sneath Coefficient [78], and Simple-Matching coefficient [79] are also applied. This thesis has identified, articulated, and addressed an inherent limitation of common similarity measures (including Jaccard and Dice) where they return the same similarity for substantially different pairs of intervals, referred to as *aliasing*. To address *aliasing*, this thesis has put forward a new definition of *robust* similarity measures for intervals by introducing new axioms in addition to the traditional ones for a similarity measure.

A bidirectional subsethood based similarity measure for continuous intervals and a new distance metric

This thesis has proposed a new set-theoretic similarity measure for continuous intervals considering the application of the bidirectional subsethood of intervals. The new measure has been designed to be more responsive towards the size of intervals and the size of their intersection. Thus, the new measure not only minimizes *aliasing* but also satisfies

all new and traditional axioms of a *robust* similarity measure. Additionally, this thesis has axiomatically defined subsethood for intervals along with its essential properties as part of developing the new similarity measure. A new distance metric has also been derived from the new similarity measure, enabling its direct use in classification and clustering problems. A comprehensive investigation has shown that the proposed similarity measure produces intuitive results for continuous intervals. Moreover, the results are shown to always be bounded by the Jaccard and Dice measures when applying minimum *t-norm*, making the new similarity measure a suitable alternative in real world applications.

A computationally efficient generalized framework for discontinuous intervals

There is no similarity measure, which can be directly applicable to compare discontinuous intervals - intervals with many continuous subintervals. Current approaches are subject to loss of discontinuity information as discontinuous intervals are converted into their continuous forms before assessing similarity. This information loss eventually affects the accuracy of similarity result. This thesis developed a generalized framework to compute similarity between discontinuous intervals by counting similarities of all pairs of their subintervals. The proposed framework retains the original information of the discontinuous intervals while computing their similarity and is equally applicable for both continuous and discontinuous intervals. Comparison of the proposed generalized framework in respect to the new similarity measure vis-a-vis popular similarity measures has shown its intuitive performance in similarity results, though entails longer execution time. Hence, the proposed generalized framework is further refined by integrating it with the extended Allen's theory [21] to make the later computationally more efficient. Experimental results have shown that this integration substantially reduces similarity calculation as well as overall execution time.

A bidirectional subsethood based similarity measure for T1 fuzzy sets

Classical set-theoretic similarity measures are generally extended for the T1 fuzzy sets. Both vertical-slice and horizontal-slice (α-cut) based extensions are possible for their comparison, though the first approach is rather common. As the new bidirectional subsethood based similarity measure has been found to be a *robust* similarity measure providing intuitive similarity results, this thesis has initially expanded this new interval measure vertically to capture similarity between T1 fuzzy sets. However, this extended measure has behaved non-intuitively like other existing fuzzy set-theoretic similarity measures (also relying on vertical-slice based representation) when both fuzzy sets are non-normal. To address this limitation, this thesis has extended the new interval similarity measure using α-cut based extension. The new α-cut based extension has been further refined so that it can be applied to compare non-normal and non-convex T1 fuzzy sets. In comparing T1 fuzzy sets with different properties, the new extension has avoided the unexpected behaviour with non-normal fuzzy-sets and produced intuitive results in all cases.

Efficient data aggregation using a bidirectional subsethood based fuzzy measure and fuzzy integral

An area of application of the similarity measures is to aggregate similar data arising from uncertain sources using FIs in respect to FMs. Current approaches to develop the FM often apply popular similarity measures, which are subject to *aliasing*. This pitfall of the existing measures often brings about the same inter-source agreement for very different subsets of sources, thus affecting their worth calculation in the FMs and finally to aggregation outcome when used in combination with an FI. To avoid this limitation, this thesis has introduced a new agreement-based FM using the new robust similarity measure, which is not subject to *aliasing*. The new FM can also generate worth information for subsets of sources having no agreement among them.

The demonstrations have also revealed that the new FM in combination with an FI gives intuitive aggregation outcome as compared to other existing agreement-based FMs, suggesting its suitability for real-world applications.

7.3 Limitations

This section presents some of the limitations within the work presented in this thesis.

- Chapter 3 has proposed a new set-theoretic similarity measure for continuous intervals considering their bidirectional subsethoods. The new similarity measure has been designed using the minimum and product *t-norm*s as they are the most common *t-norm*s and frequently used. However, this thesis has not explored all other *t-norm*s such as Łukasiewicz *t-norm* as each *t-norm* has its own properties and including them 'under the same roof' is a complex task. Further work is necessary to cover all common *t-norm*s to make the new measure more robust.

- The new bidirectional subsethood based similarity measure has been expanded for T1 fuzzy sets which comprehensively assesses their similarity. Earlier, it has been mentioned that T2 fuzzy-sets better capture higher-level of uncertainty in data as compared to T1 fuzzy sets. This measure has not been extended to interval and general T2 fuzzy sets. Further work is necessary to upgrade the extended measure to cover similarity assessment between T2 fuzzy sets.

- This thesis has developed methods to assess similarity for intervals and T1 fuzzy sets. Moreover, they have been applied to aggregate uncertain multi-source data efficiently. In this thesis, for simplicity, some demonstrations of the new measures have been shown using synthetic datasets as they can generate interesting cases by altering different properties of intervals/fuzzy-sets and can do a bit more rigorous study.

7.4 Future Work

Some potential new directions and future work based on this thesis are discussed below.

- We aim to make the new bidirectional subsethood based similarity measure to be more flexible and robust by considering all common *t-norm*s. Some preliminary work has already been made as part of this PhD for assessing similarity between intervals. Initial results have been demonstrated in Appendix B and this idea will be applied further for other *t-norm*s in future work.

- We already have extended the new bidirectional subsethood based similarity measure for comparing T1 fuzzy sets where the α-cut based extension has shown intuitive similarity results as compared to other existing fuzzy set-theoretic measures. As both interval and general T2 fuzzy sets better capture data uncertainty over T1 fuzzy sets; in future, we intend to expand this measure to assess similarity between T2 fuzzy sets.

- We have involved the new bidirectional subsethood based similarity measure to extract a new agreement-based FM and later used it in combination with the CFI [129] for interval-valued data aggregation. The new FM in combination with the CFI [129] has performed intuitively in aggregating interval datasets and also carried out aggregation under the situation when sources are in complete disagreement. In future, we will expand the new similarity measure to extract a new FM directly from the pool of fuzzy set-valued datasets so that it can be used with an FI to aggregate them.

- We have derived a distance metric from the new bidirectional subsethood based similarity measure. This distance metric has not been applied in any application to assess its suitability. In future, we will use this metric in classification and

clustering applications with interval-valued datasets and explore its potentiality for these applications.

References

[1] D. Dubois and H. Prade, *Possibility theory: an approach to computerized processing of uncertainty*. Springer Science & Business Media, 2012.

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

[3] H.-J. Zimmermann, *Fuzzy set theory—and its applications*. Springer Science & Business Media, 2011.

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

[5] 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 Transaction on Fuzzy Systems*, vol. 23, no. 2, pp. 248–269, 2015.

[6] 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.

[7] Y. Ren, Y.-H. Liu, J. Rong, and R. Dew, "Clustering interval-valued data using an overlapped interval divergence," in *Proceedings of 8th Australasian Data Mining Conference*, pp. 35–42, 2009.

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

[9] E. Vicente, A. Mateos, and A. Jiménez, "A new similarity function for generalized trapezoidal fuzzy numbers," in *Proceedings of International Conference on Artificial Intelligence and Soft Computing*, pp. 400–411, Springer, 2013.

[10] V. V. Cross and T. A. Sudkamp, *Similarity and compatibility in fuzzy set theory: assessment and applications*, vol. 93. Springer Science & Business Media, 2002.

[11] R. Zwick, E. Carlstein, and D. V. Budescu, "Measures of similarity among fuzzy concepts: A comparative analysis," *International Journal of Approximate Reasoning*, vol. 1, no. 2, pp. 221–242, 1987.

[12] A. Tversky, "Features of similarity.," *Psychological review*, vol. 84, no. 4, p. 327, 1977.

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

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

[15] 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.

[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 *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–6, 2017.

[17] D. A. Dos Santos and R. Deutsch, "The positive matching index: A new similarity measure with optimal characteristics," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1570–1576, 2010.

[18] 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 *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–8, 2013.

[19] S. Kabir, C. Wagner, T. C. Havens, and D. T. Anderson, "Measuring similarity between discontinuous intervals - challenges and solutions," in *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–7, 2019.

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

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

[22] C. Wagner and D. T. Anderson, "Extracting meta-measures from data for fuzzy aggregation of crowd sourced information," in *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–8, 2012.

[23] 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 *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–8, 2013.

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

[25] P. Bosc and H. Prade, "An introduction to the fuzzy set and possibility theory-based treatment of flexible queries and uncertain or imprecise databases," in *Uncertainty Management in Information Systems*, pp. 285–324, Springer, 1997.

[26] R. R. Yager, "A measure based approach to the fusion of possibilistic and probabilistic uncertainty," *Fuzzy Optimization and Decision Making*, vol. 10, no. 2, pp. 91–113, 2011.

[27] E. Lefevre, O. Colot, and P. Vannoorenberghe, "Belief function combination and conflict management," *Information Fusion*, vol. 3, no. 2, pp. 149–162, 2002.

[28] D. Dubois, H. Prade, and R. Yager, "Merging fuzzy information," in *Fuzzy Sets in Approximate Reasoning and Information Systems*, pp. 335–401, Springer, 1999.

[29] G. J. Klir and T. A. Folger, "Fuzzy sets, uncertainty, and information," 1988.

[30] S. Miller, C. Wagner, J. M. Garibaldi, and S. Appleby, "Constructing general type-2 fuzzy sets from interval-valued data," in *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–8, 2012.

[31] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Information sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[32] 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 *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–8, 2013.

[33] 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.

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

[35] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to interval analysis*. Society for Industrial and Applied Mathematics, 2009.

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

[37] W. Cowley and D. Plexousakis, "An interval algebra for indeterminate time," in *Innovative Applications of Artificial Intelligence Conference*, pp. 470–475, 2000.

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

[39] Y. Yao, "Interval-set algebra for qualitative knowledge representation," in *Proceedings of 5th International Conference on Computing and Information*, pp. 370–374, IEEE, 1993.

[40] M. J. Almeida, "A system for reasoning with nonconvex intervals," in *Proceedings of 6th International Workshop on Temporal Representation and Reasoning*, pp. 8–16, 1999.

[41] M. Grüninger and Z. Li, "The time ontology of allen's interval algebra," in *Proceedings of 24th International Symposium on Temporal Representation and Reasoning*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[42] S. Schockaert, M. De Cock, and E. E. Kerre, "Fuzzifying allen's temporal interval relations," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 2, pp. 517–533, 2008.

[43] L. Khatib, "Reasoning with sequences of intervals for efficient constraint propagation," 1996.

[44] P. B. Ladkin, "The completeness of a natural system for reasoning with time intervals," in *Proceedings of 10th International Joint Conference on Artificial Intelligence*, pp. 462–465, 1987.

[45] P. Ladkin, "Time representation : a taxonomy of interval relations," in *Proceedings of 6th National Conference on Artificial Intelligence*, pp. 360–366, 1986.

[46] R. A. Morris, W. D. Shoaff, and L. Khatib, "Path consistency in a network of non-convex intervals," in *Proceedings of 13 International Conference on Artificial Intelligence*, pp. 655–661, 1993.

[47] "Fuzzy logic." (Date last accessed 12-August-2017).

[48] J. M. Mendel, "On km algorithms for solving type-2 fuzzy set problems," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 3, pp. 426–446, 2012.

[49] G. J. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: theory and applications*. Prentice Hall PTR, 1995.

[50] J. M. Mendel and R. B. John, "Type-2 fuzzy sets made simple," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 2, pp. 117–127, 2002.

[51] J. M. Mendel, R. I. John, and F. Liu, "Interval type-2 fuzzy logic systems made simple," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 6, pp. 808–821, 2006.

[52] H. B. Sola, J. Fernandez, H. Hagras, F. Herrera, M. Pagola, and E. Barrenechea, "Interval type-2 fuzzy sets are generalization of interval-valued fuzzy sets: Toward a wider view on their relationship," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 5, pp. 1876–1882, 2014.

[53] J. Aisbett, J. T. Rickard, and D. G. Morgenthaler, "Type-2 fuzzy sets as functions on spaces," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 4, pp. 841–844, 2010.

[54] J. T. Rickard, J. Aisbett, and G. Gibbon, "Fuzzy subsethood for fuzzy sets of type-2 and generalized type-n," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 1, pp. 50–60, 2008.

[55] J. J. Buckley and E. Eslami, *An introduction to fuzzy logic and fuzzy sets*, vol. 13. Springer, 2002.

[56] E. P. Klement, R. Mesiar, and E. Pap, "Triangular norms. position paper i: basic analytical and algebraic properties," *Fuzzy Sets and Systems*, vol. 143, no. 1, pp. 5–26, 2004.

[57] B. Schweizer and A. Sklar, "Statistical metric spaces," *Pacific J. Math*, vol. 10, no. 1, pp. 313–334, 1960.

[58] M. M. Gupta and J. Qi, "Theory of t-norms and fuzzy inference methods," *Fuzzy Sets and Systems*, vol. 40, no. 3, pp. 431–450, 1991.

[59] À. Garcı, E. Armengol, F. Esteva, *et al.*, "Fuzzy description logics and t-norm based fuzzy logics," *International Journal of Approximate Reasoning*, vol. 51, no. 6, pp. 632–655, 2010.

[60] X. Zhang, "A novel approach based on similarity measure for pythagorean fuzzy multiple criteria group decision making," *International Journal of Intelligent Systems*, vol. 31, no. 6, pp. 593–611, 2016.

[61] I. Turksen and Z. Zhong, "An approximate analogical reasoning approach based on similarity measures," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, no. 6, pp. 1049–1056, 1988.

[62] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, "Similarity-based machine learning methods for predicting drug–target interactions: a brief review," *Briefings in bioinformatics*, vol. 15, no. 5, pp. 734–747, 2013.

[63] J. McCulloch, C. Wagner, and U. Aickelin, "Analysing fuzzy sets through combining measures of similarity and distance," in *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 155–162, 2014.

[64] L. Xuecheng, "Entropy, distance measure and similarity measure of fuzzy sets and their relations," *Fuzzy Sets and Systems*, vol. 52, no. 3, pp. 305–318, 1992.

[65] 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.

[66] H. Bustince, E. Barrenechea, and M. Pagola, "Restricted equivalence functions," *Fuzzy Sets and Systems*, vol. 157, no. 17, pp. 2333–2346, 2006.

[67] I. Beg and S. Ashraf, "Similarity measures for fuzzy sets," *Applications and Computational Mathematics*, vol. 8, no. 2, pp. 192–202, 2009.

[68] M. Galar, J. Fernández, G. Beliakov, and H. Bustince, "Interval-valued fuzzy sets applied to stereo matching of color images," *IEEE Transaction on Image Processing*, vol. 20, no. 7, pp. 1949–1961, 2011.

[69] D. J. Weller-Fahy, B. J. Borghetti, and A. A. Sodemann, "A survey of distance and similarity measures used within network intrusion anomaly detection," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 70–91, 2014.

[70] L. Baccour and R. I. John, "Experimental analysis of crisp similarity and distance measures," in *Proceedings of 6th International Conference on Soft Computing and Pattern Recognition*, pp. 96–100, 2014.

[71] A. Gardner, J. Kanno, C. A. Duncan, and R. Selmic, "Measuring distance between unordered sets of different sizes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 137–143, 2014.

[72] L.-A. C. Hayek, "Analysis of amphibian biodiversity data," *Measuring and monitoring biological diversity: standard methods for amphibians*, pp. 207–273, 1994.

[73] Z. C. Johanyák and S. Kovács, "Distance based similarity measures of fuzzy sets," in *Proceedings of 3rd Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence*, 2005.

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

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

[76] D. Szymkiewicz, "Une conlribution statistique à la géographie floristique," *Acta Societatis Botanicorum Poloniae*, vol. 11, no. 3, pp. 249–265, 1934.

[77] Y. Otsuka, "The faunal character of the japanese pleistocene marine mollusca, as evidence of climate having become colder during the pleistocene in japan," *Bulletin of the Biogeographical Society of Japan*, vol. 6, no. 16, pp. 165–170, 1936.

[78] R. R. Sokal and P. H. A. Sneath, "Principles of numerical taxonomy.," *Principles of numerical taxonomy*, 1963.

[79] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationship," *University of Kansas science bulletin*, vol. 28, pp. 1409–1438, 1958.

[80] L. A. Zadeh, "Similarity relations and fuzzy orderings," *Information sciences*, vol. 3, no. 2, pp. 177–200, 1971.

[81] D. Wu and J. M. Mendel, "A vector similarity measure for linguistic approximation: Interval type-2 and type-1 fuzzy sets," *Information Sciences*, vol. 178, no. 2, pp. 381–402, 2008.

[82] H. B. Mitchell, "Pattern recognition using type-ii fuzzy sets," *Information Sciences*, vol. 170, no. 2-4, pp. 409–418, 2005.

[83] D. Wu and J. M. Mendel, "Perceptual reasoning for perceptual computing: A similarity-based approach," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 6, pp. 1397–1411, 2009.

[84] L. A. Zadeh, "Fuzzy sets and their application to pattern classification and clustering analysis," in *Classification and clustering*, pp. 251–299, Elsevier, 1977.

[85] S. Miyamoto and K. Nakayama, "Similarity measures based on a fuzzy set model and application to hierarchical clustering," *IEEE transactions on Systems, Man, and Cybernetics*, vol. 16, no. 3, pp. 479–482, 1986.

[86] Y. Li, K. Qin, and X. He, "Some new approaches to constructing similarity measures," *Fuzzy Sets and Systems*, vol. 234, pp. 46–60, 2014.

[87] L. Baccour, A. M. Alimi, and R. I. John, "Some notes on fuzzy similarity measures and application to classification of shapes recognition of arabic sentences and mosaic," *IAENG International Journal of Computer Science*, vol. 41, no. 2, pp. 81–90, 2014.

[88] 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*, pp. 1–6, 2018.

[89] 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.

[90] D. Wu and J. M. Mendel, "Similarity measures for closed general type-2 fuzzy sets: Overview, comparisons, and a geometric approach," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 3, pp. 515–526, 2018.

[91] S.-M. Chen, M.-S. Yeh, and P.-Y. Hsiao, "A comparison of similarity measures of fuzzy values," *Fuzzy Sets and Systems*, vol. 72, no. 1, pp. 79–89, 1995.

[92] W.-J. Wang, "New similarity measures on fuzzy sets and on elements," *Fuzzy Sets and Systems*, vol. 85, no. 3, pp. 305–309, 1997.

[93] S. Raha, A. Hossain, and S. Ghosh, "Similarity based approximate reasoning: fuzzy contro," *Journal of Applied Logic*, vol. 6, no. 1, pp. 47–71, 2008.

[94] A. Pal, B. Mondal, N. Bhattacharyya, and R. Swapan, "Similarity in fuzzy systems," *Journal of Uncertainty Analysis and Applications*, vol. 2, no. 18, pp. 1–28, 2014.

[95] H. T. Nguyen and K. Vladik, "Computing degrees of subsethood and similarity for interval-valued fuzzy sets: fast algorithms," in *Proceedings of 9th International Conference on Intelligent Technologies*, (Thailand), pp. 47–55, 2008.

[96] Z. Takáč, "Inclusion and subsethood measure for interval-valued fuzzy sets and for continuous type-2 fuzzy sets," *Fuzzy Sets and Systems*, vol. 244, no. 1, pp. 106–120, 2013.

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

[98] W. Bandler and L. Kohout, "Fuzzy power sets and fuzzy implication operators," *Fuzzy Sets and Systems*, vol. 4, no. 1, pp. 13–30, 1980.

[99] L. Kitainik, "Fuzzy inclusions and fuzzy dichotomous decision procedures," in *Optimization Models Using Fuzzy Sets and Possibility Theory*, pp. 154–170, Springer, 1987.

[100] D. Sinha and E. R. Dougherty, "Fuzzification of set inclusion: theory and applications," *Fuzzy Sets and Systems*, vol. 55, no. 1, pp. 15–42, 1993.

[101] V. R. Young, "Fuzzy subsethood," *Fuzzy Sets and Systems*, vol. 77, no. 3, pp. 371–384, 1996.

[102] J. Fan, W. Xie, and J. Pei, "Subsethood measure: new definitions," *Fuzzy Sets and Systems*, vol. 106, no. 2, pp. 201–209, 1999.

[103] H. Bustince, V. Mohedano, E. Barrenechea, and M. Pagola, "Definition and construction of fuzzy di-subsethood measures," *Information Sciences*, vol. 176, no. 21, pp. 3190–3231, 2006.

[104] Z. Takáč, M. Minárová, J. Montero, E. Barrenechea, J. Fernandez, and H. Bustince, "Interval-valued fuzzy strong s-subsethood measures, interval-entropy and p-interval-entropy," *Information Sciences*, vol. 432, pp. 97–115, 2018.

[105] H. Bustince, "Indicator of inclusion grade for interval-valued fuzzy sets. application to approximate reasoning based on interval-valued fuzzy sets," *International Journal of Approximate Reasoning*, vol. 23, no. 3, pp. 137–209, 2000.

[106] W. Zeng and P. Guo, "Normalized distance, similarity measure, inclusion measure and entropy of interval-valued fuzzy sets and their relationship," *Information Sciences*, vol. 178, no. 5, pp. 1334–1342, 2008.

[107] I. K. Vlachos and G. D. Sergiadis, "Subsethood, entropy, and cardinality for interval-valued fuzzy sets—an algebraic derivation," *Fuzzy Sets and Systems*, vol. 158, no. 12, pp. 1384–1396, 2007.

[108] D. Dubois and H. Prade, "On the use of aggregation operations in information fusion processes," *Fuzzy Sets and Systems*, vol. 142, no. 1, pp. 143–161, 2004.

[109] V. Torra and Y. Narukawa, *Modeling decisions: information fusion and aggregation operators*. Springer, 2007.

[110] G. L. Rogova and V. Nimier, "Reliability in information fusion: literature survey," in *Proceedings of 7th International Conference on Information fusion*, vol. 2, pp. 1158–1165, 2004.

[111] V. Torra and Y. Narukawa, *Modeling decisions: information fusion and aggregation operators*. Springer Science & Business Media, 2007.

[112] D. Dubois and H. Prade, "On the use of aggregation operations in information fusion processes," *Fuzzy Sets and Systems*, vol. 142, no. 1, pp. 143–161, 2004.

[113] Y.-Y. Guh, R.-W. Po, and E. S. Lee, "The fuzzy weighted average within a generalized means function," *Computers & Mathematics with Applications*, vol. 55, no. 12, pp. 2699–2706, 2008.

[114] F. Herrera and E. Herrera-Viedma, "Aggregation operators for linguistic weighted information," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 27, no. 5, pp. 646–656, 1997.

[115] S.-M. Zhou, R. I. John, F. Chiclana, and J. M. Garibaldi, "On aggregating uncertain information by type-2 owa operators for soft decision making," *International Journal of Intelligent Systems*, vol. 25, no. 6, pp. 540–558, 2010.

[116] R. Boukezzoula, S. Galichet, and L. Foulloy, "Min and max operators for fuzzy intervals and their potential use in aggregation operators," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 6, pp. 1135–1144, 2007.

[117] M. Grabisch and C. Labreuche, "Bi-capacitiesâ€"ii: the choquet integral," *Fuzzy Sets and Systems*, vol. 151, no. 2, pp. 237–259, 2005.

[118] J.-L. Marichal, "On sugeno integral as an aggregation function," *Fuzzy Sets and Systems*, vol. 114, no. 3, pp. 347–365, 2000.

[119] S. Ovchinnikov, "Piecewise linear aggregation functions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 01, pp. 17–24, 2002.

[120] L. Garmendia, "The evolution of the concept of fuzzy measure," in *Intelligent Data Mining: Techniques and Applications*, pp. 185–200, Springer, 2005.

[121] R. Mesiar, "Fuzzy measures and integrals," *Fuzzy Sets and Systems*, vol. 156, no. 3, pp. 365–370, 2005.

[122] C. Wagner, T. C. Havens, and D. T. Anderson, "The arithmetic recursive average as an instance of the recursive weighted power mean," in *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–6, 2017.

[123] M. Grabisch, T. Murofushi, and M. Sugeno, *Fuzzy measures and integrals-Theory and applications*, vol. 40. Physica Verlag, 2000.

[124] M. Sugeno, *Theory of fuzzy integrals and its applications.* Tokyo Institute of Technology, 1974.

[125] D. T. Anderson, J. M. Keller, and T. C. Havens, "Learning fuzzy-valued fuzzy measures for the fuzzy-valued sugeno fuzzy integral," in *Inernational Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 502–511, 2010.

[126] A. J. Pinar, J. Rice, L. Hu, D. T. Anderson, and T. C. Havens, "Efficient multiple kernel classification using feature and decision level fusion," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 6, pp. 1403–1416, 2016.

[127] A. Temko, D. Macho, and C. Nadeu, "Fuzzy integral based information fusion for classification of highly confusable non-speech sounds," *Pattern Recognition*, vol. 41, no. 5, pp. 1814–1823, 2008.

[128] M. F. Anderson, D. T. Anderson, and D. J. Wescott, "Estimation of adult skeletal age-at-death using the sugeno fuzzy integral," *American Journal of Physical Anthropology*, vol. 142, no. 1, pp. 30–41, 2010.

[129] G. Choquet, "Theory of capacities," in *Annales de l'institut Fourier*, vol. 5, pp. 131–295, 1954.

[130] T. Murofushi and M. Sugeno, "An interpretation of fuzzy measures and the choquet integral as an integral with respect to a fuzzy measure," *Fuzzy sets and Systems*, vol. 29, no. 2, pp. 201–227, 1989.

[131] M. Grabisch, "The application of fuzzy integrals in multicriteria decision making," *European Journal of Operational Research*, vol. 89, no. 3, pp. 445–456, 1996.

[132] D. T. Anderson, T. C. Havens, C. Wagner, J. M. Keller, M. F. Anderson, and D. J. Wescott, "Sugeno fuzzy integral generalizations for sub-normal fuzzy set-valued inputs," in *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–8, 2012.

[133] T. C. Havens, D. T. Anderson, and J. M. Keller, "A fuzzy choquet integral with an interval type-2 fuzzy number-valued integrand," in *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 1–8, 2010.

[134] R. Yang, Z. Wang, P. Heng, and K. Leung, "Fuzzified choquet integral with a fuzzy-valued integrand and its application on temperature prediction," *IEEE*

Transactions on Systems, Man, and Cybernetics, vol. 38, no. 2, pp. 367–380, 2008.

[135] P. Meyer and M. Roubens, "On the use of the choquet integral with fuzzy numbers in multiple criteria decision support," *Fuzzy Sets and Systems*, vol. 157, no. 7, pp. 927–938, 2006.

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

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

[138] U. K. Met Office, "2016 Weather Summaries." accessed 2017-12-16.

[139] D. Cukierman and J. Delgrande, "Characterizing temporal repetition," in *Proceedings of 3rd International Workshop on Temporal Representation and Reasoning*, pp. 80–87, 1996.

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

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

[142] B. B. Chaudhur and A. Rosenfeld, "On a metric distance between fuzzy sets," *Pattern Recognition Letters*, vol. 17, no. 11, pp. 1157–1160, 1996.

[143] S. Miller, S. Appleby, J. M. Garibaldi, and U. Aickelin, "Towards a more systematic approach to secure systems design and analysis," *International Journal of Secure Software Engineering*, vol. 4, no. 1, pp. 11–30, 2013.

[144] S. Miller, C. Wagner, U. Aickelin, and J. M. Garibaldi, "Modelling cyber-security experts' decision making processes using aggregation operators," *Computers & Security*, vol. 62, pp. 229–245, 2016.

[145] J. Ye, "Multicriteria group decision-making method using vector similarity measures for trapezoidal intuitionistic fuzzy numbers," *Group Decision and Negotiation*, vol. 21, no. 4, pp. 519–530, 2012.

Appendix A

Similarity Results for Summer and Winter Temperature Data

Table A.1 Similarity results for Summer temperature data using the Jaccard similarity measure (S_J)

	Summer								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.859	0.7073	0.4712	0.4583	0.3565	0.2692	0.375	0.272
SE	0.859	1	0.4712	0.5769	0.5729	0.4522	0.3538	0.4808	0.36
SW	0.7073	0.4712	1	0.6737	0.6782	0.5283	0.4132	0.5684	0.4224
EEN	0.4712	0.5769	0.6737	1	0.8488	0.8041	0.6429	0.8837	0.6636
ENW	0.4583	0.5729	0.6782	0.8488	1	0.7609	0.5981	0.8395	0.6176
ML	0.3565	0.4522	0.5283	0.8041	0.7609	1	0.7981	0.8539	0.8283
EA	0.2692	0.3538	0.4132	0.6429	0.5981	0.7981	1	0.7059	0.9388
ESW	0.375	0.4808	0.5684	0.8837	0.8395	0.8539	0.7059	1	0.732
ESC	0.272	0.36	0.4224	0.6636	0.6176	0.8283	0.9388	0.732	1

Table A.2 Similarity results for Winter temperature data using the Jaccard similarity measure (S_J)

Area	**Winter**								
	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9014	0.7778	0.5287	0.5663	0.4624	0.3271	0.4149	0.3491
SE	0.9014	1	0.7763	0.5385	0.5747	0.4742	0.3423	0.4286	0.3636
SW	0.7778	0.7763	1	0.6835	0.7333	0.60	0.4343	0.5465	0.4592
EEN	0.5287	0.5385	0.6835	1	0.9286	0.88	0.6517	0.8158	0.6818
ENW	0.5663	0.5747	0.7333	0.9286	1	0.8158	0.60	0.7532	0.6292
ML	0.4624	0.4742	0.60	0.88	0.8158	1	0.7442	0.9315	0.7765
EA	0.3271	0.3423	0.4343	0.6517	0.60	0.7442	1	0.7927	0.9625
ESW	0.4149	0.4286	0.5465	0.8158	0.7532	0.9315	0.7927	1	0.8272
ESC	0.3491	0.3636	0.4592	0.6818	0.6292	0.7765	0.9625	0.8272	1

Table A.3 Similarity results for Summer temperature data using the Dice similarity measure (S_D)

Area	**Summer**								
	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9241	0.8286	0.6405	0.6286	0.5256	0.4242	0.5455	0.4277
SE	0.9241	1	0.9139	0.7317	0.7285	0.6228	0.5227	0.6494	0.5294
SW	0.8286	0.9139	1	0.805	0.8082	0.6914	0.5848	0.7248	0.5939
EEN	0.6405	0.7317	0.805	1	0.9182	0.8914	0.7826	0.9383	0.7978
ENW	0.6286	0.7285	0.8082	0.9182	1	0.8642	0.7485	0.9128	0.7636
ML	0.5256	0.6228	0.6914	0.8914	0.8642	1	0.8877	0.9212	0.9061
EA	0.4242	0.5227	0.5848	0.7826	0.7485	0.8877	1	0.8276	0.9684
ESW	0.5455	0.6494	0.7248	0.9383	0.9128	0.9212	0.8276	1	0.8452
ESC	0.4277	0.5294	0.5939	0.7978	0.7636	0.9061	0.9684	0.8452	1

Table A.4 Similarity results for Winter temperature data using the Dice similarity measure (S_D)

	Winter								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9481	0.875	0.6917	0.7231	0.6324	0.493	0.5865	0.5175
SE	0.9481	1	0.8741	0.70	0.7299	0.6434	0.5101	0.60	0.5333
SW	0.875	0.8741	1	0.812	0.8462	0.75	0.6056	0.7068	0.6294
EEN	0.6917	0.70	0.812	1	0.963	0.9362	0.7891	0.8986	0.8108
ENW	0.7231	0.7299	0.8462	0.963	1	0.8986	0.75	0.8593	0.7724
ML	0.6324	0.6434	0.75	0.9362	0.8986	1	0.8533	0.9645	0.8742
EA	0.493	0.5101	0.6056	0.7891	0.75	0.8533	1	0.8844	0.9054
ESW	0.5865	0.60	0.7068	0.8986	0.8593	0.9645	0.8844	1	0.9054
ESC	0.5175	0.5333	0.6294	0.8108	0.7724	0.8742	0.9054	0.9054	1

Table A.5 Similarity results for Summer temperature data using the proposed similarity measure (S_{S_h})

	Summer								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.859	0.7945	0.5698	0.6027	0.4607	0.3571	0.5132	0.3696
SE	0.859	1	0.8846	0.5698	0.7051	0.5843	0.4694	0.641	0.3696
SW	0.7945	0.8846	1	0.5698	0.8082	0.6292	0.5102	0.7105	0.5326
EEN	0.5698	0.5698	0.5698	1	0.8488	0.8764	0.7347	0.8837	0.7717
ENW	0.6027	0.7051	0.8082	0.8488	1	0.7865	0.6531	0.8947	0.6848
ML	0.4607	0.5843	0.6292	0.8764	0.7865	1	0.8469	0.8539	0.8913
EA	0.3571	0.4694	0.5102	0.7347	0.6531	0.8469	1	0.7347	0.9388
ESW	0.5132	0.641	0.7105	0.8837	0.8947	0.8539	0.7347	1	0.7717
ESC	0.3696	0.3696	0.5326	0.7717	0.6848	0.8913	0.9388	0.7717	1

Table A.6 Similarity results for Winter temperature data using the proposed similarity measure (S_{S_h})

	Winter								
Area	SN	SE	SW	EEN	ENW	ML	EA	ESW	ESC
SN	1	0.9014	0.875	0.6667	0.7121	0.5972	0.4487	0.5652	0.4684
SE	0.9014	1	0.875	0.6901	0.7042	0.6389	0.4872	0.5915	0.5063
SW	0.875	0.875	1	0.7826	0.8333	0.7083	0.5513	0.6812	0.5696
EEN	0.6667	0.6901	0.7826	1	0.942	0.9167	0.7436	0.8986	0.7595
ENW	0.7121	0.7042	0.8333	0.942	1	0.8611	0.6923	0.8406	0.7089
ML	0.5972	0.6389	0.7083	0.9167	0.8611	1	0.6923	0.9444	0.8354
EA	0.4487	0.4872	0.5513	0.7436	0.6923	0.6923	1	0.8333	0.9747
ESW	0.5652	0.5915	0.6812	0.8986	0.8406	0.9444	0.8333	1	0.8481
ESC	0.4684	0.5063	0.5696	0.7595	0.7089	0.8354	0.9747	0.8481	1

Appendix B

Survey on Assessing Security Risk of a Proposed Computing System

For assessing the security risk of a proposed computing system, thirty-nine experts were asked to complete two survey exercises. In Survey-1, they were asked to rank ten technical attacks according to the difficulty to complete the attack without being noticed. In Survey-2, the experts were asked to rate each of the twenty-six security components in terms of overall difficulty to *attack/evade* the component and the characteristics/factors contributed to the difficulty of *attacking/evading* the component.

The set of questions asked for an *attack* component:

(1) How complex is the target component (e.g., in terms of size of code, number of sub-components)?

(2) How much does the target component process/interact with any data input?

(3) How often would you say this type of attack is reported in the public domain?

(4) How likely is it that there will be a publicly available tool that could help with this attack?

(5) How inherently difficult is this type of attack (i.e., how technically demanding would it be to do from scratch, with no tool-set to help)?

(6) How mature is this type of technology?

(7) How easy is to carry out this attack without being noticed?

(8) Overall, how difficult would it be for an attacker to do this?

The set of questions asked for an *evade* component:

(1) How complex is the job of providing this kind of defence?

(2) How likely is it that there will be publicly available information that could help with evading this defence?

(3) How mature is this type of technology?

(4) Overall, how difficult would it be for an attacker to do this?